O'REILLY®

The AI Ladder

Accelerate Your Journey to AI

Rob Thomas &
Paul Zikopoulos

O'REILLY®

The AI Ladder

AI may be the greatest opportunity of our time, with the potential to add nearly $16 trillion to the global economy over the next decade. But so far, adoption has been much slower than anticipated, or so headlines may lead you to believe. With this practical guide, business leaders will discover where they are in their AI journey and learn the steps necessary to successfully scale AI throughout their organization.

Authors Rob Thomas and Paul Zikopoulos from IBM introduce C-suite executives and business professionals to the AI Ladder—a unified, prescriptive approach to help them understand and accelerate the AI journey. Complete with real-world examples and real-life experiences, this book explores AI drivers, value, and opportunity, as well as the adoption challenges organizations face.

- Understand why you can't have AI without an information architecture (IA)
- Appreciate how AI is as much a cultural change as it is a technological one
- Collect data and make it simple and accessible, regardless of where it lives
- Organize data to create a business-ready analytics foundation
- Analyze data, and build and scale AI with trust and transparency
- Infuse AI throughout your entire business and create intelligent workflows

Robert D. Thomas (@robdthomas) is senior vice president of IBM Cloud and Data Platform. He directs IBM's product design and investment strategy, expert labs, global software product development, marketing, and field operations across the company's vast software portfolio.

Paul C. Zikopoulos (@BigData_paulz) is vice president of cognitive big data systems at IBM. Named a 2019 Top 100 Big Data AI thought leader by Analytics Insights with 20 books and over 350 articles published, he's considered a global expert in big data, analytics, and AI technologies.

Contributors include Daniel Hernandez, Beth Smith, Bill Higgins, Ruchir Puri, Ritika Gunnar, Seth Dobrin, and Armand Ruiz.

AI

9 781492 073444

Twitter: @oreillymedia
facebook.com/oreilly

The AI Ladder

Accelerate Your Journey to AI

Rob Thomas and Paul Zikopoulos

*Contributors: Daniel Hernandez, Beth Smith, Bill Higgins,
Ruchir Puri, Ritika Gunnar, Seth Dobrin, and Armand Ruiz*

Beijing · Boston · Farnham · Sebastopol · Tokyo O'REILLY®

The AI Ladder

by Rob Thomas and Paul Zikopoulos

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Acquisitions Editor: Rachel Roumeliotis
Developmental Editor: Nicole Taché
Production Editor: Kristen Brown
Copyeditor: Rachel Head
Proofreader: James Fraleigh

Indexer: WordCo Indexing Services, Inc.
Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Rebecca Demarest

May 2020: First Edition

Revision History for the First Edition

2020-04-29: First Release
2020-05-15: Second Release

See *http://oreilly.com/catalog/errata.csp?isbn=9781492073437* for release details.

The views expressed in this work are those of the authors, and do not represent the publisher's views. While the publisher and the authors have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the authors disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

This work is part of a collaboration between O'Reilly and IBM. See our statement of editorial independence.

978-1-492-07344-4

[LSI]

Table of Contents

Preface

Everybody's talking about AI. Why? Well, we believe AI presents a tremendous opportunity for businesses of every size, across any industry. We believe AI is shaping (or will shape) future outcomes, automating decisions, processes, and workflows, enabling people to do higher value work, and reimagining new business models. We believe AI is a powerful tool for rethinking and reinventing businesses, not just squeezing a few extra percent out of tired processes. In fact, we don't believe we'll be calling it AI in the future in the same way that we don't cite using a simple messaging service (SMS) to quickly communicate with someone anymore (we just say we "texted" them). Quite simply, artificial intelligence will become ambient intelligence… because AI will be doing its thing, in the background, unobtrusively, everywhere. AI will lift and shift some companies into heights never imagined and will serve as a rift or cliff for others. AI will be the lifeblood of a company, helping it reimagine tired business processes, create new models, and super-charge reinvention; and in the uncertainty of these modern times, what business doesn't need reinvention?

The reality is, AI isn't magic. It's hard work. It requires the right tools, methodologies, and mindset to overcome the challenges companies face such as data complexity, talent scarcity, and a lack of trust in AI systems. So, as critical as AI is to any organization's long-term success, it's also likely you'll run into some of these challenges and fail. We've seen plenty of failures; perhaps you have, too. And that's a good thing, because with AI we believe you should embrace fast but safe failure. AI is about a different culture. It's about a culture of iteration and experimentation, with small agile groups working for three to five weeks at a time. Your goal should be to do a hundred AI projects a year. If you do that, half of them are probably going to fail. And that's okay. Because while half of your AI projects might fail, the half that succeed are going to drive huge outcomes for your organization.

We've had privileged access to observe and analyze a corpus of over 30,000 IBM AI engagements around the world, and what we've concluded (via personal experiences or research around them) from those engagements is that organizations need a thoughtful and well-architected approach to AI, particularly in today's hybrid multicloud world.

That's why we developed the AI Ladder. It's our proven methodology to help businesses succeed with AI—not just by updating a few old processes, but by re-creating themselves and doing business in a new way. It's proven; it's solid; it's effective; and make no mistake about it…it works.

We wrote most of this book before COVID-19 originated. As we finished, it was looming, and now that we're done, it's here in full force. Businesses are being forced to reinvent themselves, whether they like it or not. In today's current environment, reinventing your business with AI is no longer an option. It's a requirement.

You want to succeed. We want you to succeed. And that's why we wrote this book.

Who This Book Is For

As we mentioned, everyone is talking about AI, so we wrote this book for business leaders who are (or are hearing others) talking about AI. You might be pretty informed on all aspects of AI or you might just be getting started and have little understanding of how it works, why it works, or how to implement and scale it successfully. We wrote it for all business leaders. Naturally it'll really help those leaders who are not sure where to start, but we think our personal experiences will greatly benefit those who have started and are unsure where to go next or are not making the progress they'd hoped for. After all, why not learn from the mistakes we've seen and the mistakes we made so you don't make the same ones?

This book requires no previous knowledge of AI. The topics we introduce will be helpful for a vast audience, ranging from business leaders, lines of business, as well as data scientists.

We've included success stories, but we also outline failures and challenges, and provide solutions to avoid or overcome them. The concepts you need to know to successfully scale AI across your business can be found in the following pages.

O'Reilly Online Learning

For more than 40 years, *O'Reilly Media* has provided technology and business training, knowledge, and insight to help companies succeed.

Our unique network of experts and innovators share their knowledge and expertise through books, articles, and our online learning platform. O'Reilly's online learning platform gives you on-demand access to live training courses, in-depth learning paths, interactive coding environments, and a vast collection of text and video from O'Reilly and 200+ other publishers. For more information, please visit *http://oreilly.com*.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at *https://oreil.ly/The_AI_Ladder_1e*.

Email *bookquestions@oreilly.com* to comment or ask technical questions about this book.

For news and more information about our books and courses, see our website at *http://www.oreilly.com*.

Find us on Facebook: *http://facebook.com/oreilly*

Follow us on Twitter: *http://twitter.com/oreillymedia*

Watch us on YouTube: *http://www.youtube.com/oreillymedia*

Acknowledgments from the Authors

This book would not have been possible without the insights and input from some of the most brilliant minds we know; an impressive "murderers' row" of data thought leadership if you will. Sincerely, thank you all: Ruchir Puri, Beth Smith, Daniel Hernandez, Ritika Gunnar, Seth Dobrin, Armand Ruiz, and Bill Higgins.

Several people have provided helpful feedback on the draft versions of this book, including our editors at O'Reilly, Mike Loukides (a guy who we noticed does a heck of lot of research behind every suggestion…and we really appreciated it) and Nicole Taché.

Finally, we want to heartfully thank (although at times we cursed their deadlines) Elizabeth Schaefer, Caitlin Leddy, and Katie Schafer for their personal efforts that went into getting this book from an idea we had in a coffee shop to what's in your hands today. Some people show up at their job every day to earn a paycheck, but we're privileged to collaborate with ladies like these who instead come to work to build something they believe in.

Acknowledgments from Paul Zikopoulos

The most beautiful wrist instruments (though I just call them watches) are powered by finely tuned gears that present a set of complications and time to their owners. When done to perfection, they are timepieces that last generations and carry with them stories that transcend the years. It's only when you open such a keepsake that you can truly appreciate how their multiple inner workings drive the beauty on the outside. While I'm no beauty, my inner workings are driven by friends, family, and colleagues that likely don't get appreciated (or told) how my public timekeeping hands only turn because of them.

Personally, my mom and dad programmed me from childhood to be curious, to question everything, and never stop learning. My wife (Kelly) and daughter (Chloë) are the other reason I drive myself like I do. To all of them—love.

Professionally, there are people who have inspired me (find them in past book dedications), some new inspirational sources, and those that forever inspire me. In the forever group, thanks to Bob Picciano for just always guiding me and to John Teltsch for his inspirational viewpoint on teams and companies: "They are nothing without their people." And to Rob, my coauthor, much respect. Rob has celebrated me, challenged me, laughed at and with me, and forever earned my gratitude; it was Rob that really got in my face and pushed my pivot to big data and AI years ago. This guy simply makes good people great…and for whatever greatness I have in my field, you've been a part of it for a very long time. I've also spent my last years in a new division at IBM where I've found new people who inspire me and don't even realize they're doing it...so thanks to Chris Konarski, Art Beller, Ellen Smith, Terry Bird, Drew Valentine, Ed Walsh, Kelly Robinson, and Sumit Gupta. Liz, Catilin, and Katie from above—I can't say enough about you three, so I'll use one word: grateful. They are incredible people who give me a reason to come to work every day and try to bring my best. I'd also be remiss without a shoutout to Steve Rodriguez for all his help with a golf AI project (and the positive energy he provides); buried in the book is a special thanks.

Finally, to my Tokata Farms horse family: Rodney and Karen Reid, Dean and Lisa Baker (whose house in the Caribbean served as a great location for writing some of this book), Shannon Sawyer, Ryan and Zoë Frech, and Mike and Shannon Pickard. You guys are like blood to me. In the time we've been together, none of you seem to be able to figure out what I blog about or what AI is…so I partially wrote this book for you. Lots of love to you all, and giddy up.

CHAPTER 1

What in the AI? How Did We Get Here?

The idea for this book came to us on a crisp fall day as we pondered on lunch in the hustle-and-bustle streets of New York City. Rob pulled up Foursquare to find a good place to eat (a restaurant recommendation validated by thousands of people we don't know) and we settled on a spot that was a bit of a walk, tucked away on a city block we'd never heard of before. Paul checked the weather forecast (bespoke to about a mile and updated every 15 minutes) to see if we needed jackets, while Rob pulled up the destination using Google Maps in augmented reality mode. As we trekked the city streets, Paul used the Starbucks app to order teas that were piping hot, ready the moment we arrived at the halfway mark. We had devices that monitored our steps and heart rates, and just before we arrived at the restaurant, Rob Shazam'ed a song we heard blasting out of a café that we thought was cool. And just like that, we generated a heck of a lot of decorated data. By *decorated*, we mean that it's not just data, it's contextualized data. It's data superimposed on a map of Manhattan, data that can be correlated with other data, data that gives a picture of what we were doing, and what other visitors to New York might want to do, are doing, or have done. Where we walked, what we listened to, what we ate and drank—that's all decorated data, data *in context*.

Over lunch we reflected on all that data and declared ourselves digital city cartographers, mapping our lives, events, and interactions with the city in an everlasting digital footprint. And oh, what a digital footprint we made. After all, we generated more data in our hour-long walk than most people in ancient civilizations would have come across in a lifetime!

Toward the end of our lunch, we agreed on this simple fact: today, most organizations' data acumen isn't where it needs to be. What's more, most organizations don't have an information architecture (IA) to accelerate AI outcomes. You'll often hear us declare, “You just can’t have AI without an IA!” When it comes to data, it’s simply not a level playing field and this book was written to give you a leg up.

On the walk back to the office, we talked about the struggles we’ve seen organizations face with AI, and the struggles we will likely see in the future (this AI thing is just getting started). We witnessed the Hadoop craze and its once-dominating hype; for the last few years we’ve been hearing about inconsistent business outcomes (as defined by the business) but those same projects are also called successful IT projects (as defined by IT). We know lots of organizations have “big data projects,” but big data without analytics is, well, just a bunch of data. The key takeaway is that data is at your doorstep, but most organizations aren’t ready to welcome it and turn it into insights. The playing field might not be level, due to differences in resources like access to data, funding, and staff, but we think these handicaps are largely self-imposed—and we know how to help organizations remove those handicaps.

With this in mind, we decided that we needed to write a book. This book would give organizations some guidance about how to use their data productively. It also gives organizations a process to get started with AI. That guidance, that process, is what we call the *AI Ladder*. We introduce the ladder formally in Chapter 4, and we discuss each of the rungs in detail in Chapters 5 through 9.

Here’s a very quick summary of the AI ladder:

Collect data
Find the data that your organization has access to, regardless of where it is or how it is stored. This includes data from external sources and data that’s currently “falling on the floor.”

Organize the data
Data is just a “seething mass of bits” if it isn’t organized. Data needs to be trustworthy if you’re going to have trustworthy results. It needs to be cataloged so others can use it; it needs to be governed, and access needs to be controlled, for regulatory compliance; and it needs to be cleaned so you know it is accurate.

Analyze the data using machine learning
This is the fun part; it’s where you build and deploy AI models developed from your data.

Infuse AI throughout the organization
AI can transform your organization—but it won’t if it’s limited to a few projects in a few departments. The most exciting part of the AI Ladder isn’t the first few successes, it’s finding out how to make your entire organization more effective.

The foundation on which the AI Ladder rests is a modern information architecture (IA): it's the utopian lift of democratizing AI across an enterprise. We say repeatedly in this book that "there is no AI without IA." At the same time, if you try to create a modern information architecture, collect your data, and organize your data before you can start any analysis, you're not likely to get anywhere, at least not this decade. That's a problem we address specifically. Although it's a ladder, there are ways to take shortcuts, start with some successful projects, and get on the road to AI without going rung by rung. Indeed, those first successes will help you get the buy-in and support you need for everything else.

Collecting Data in Real Time, but Understanding It in Stale Time

We have yet to meet an organization that has told us it has a serious data collection problem, but we've heard from countless organizations that they can't understand the mountains of data they collect. If you listen closely to most organizations' data challenges, they will admit they are data-rich and information-poor. In other words, they have a real-time data collection strategy, but they only understand their data in stale time.

With this in mind, we propose this simple equation as a guiding principle when reading this book:

```
data collection
– data understanding
----------------------------
= the price of not knowing
```

Take a look at Figure 1-1. A typical organization's ability to collect data is illustrated by the steeply sloped thick line. Over time, lots and lots of data is collected, and the speed at which data is generated increases, often exponentially. Meanwhile, the organization's data understanding capabilities grow more slowly, as illustrated by the thinner, flatter line.

Now take all the space between the thick line (*data collection*) and the thin line (*data understanding*), and you have what we call *the price of not knowing*. In this gap, the organization is guilty of not knowing what it could already know (or may have known in the past). The consequence? You name it: money lost, opportunities squandered, vacation flights missed, cars damaged, fraud enabled, lives lost, and more. We sometimes jokingly refer to this as *Enterprise Amnesia*.

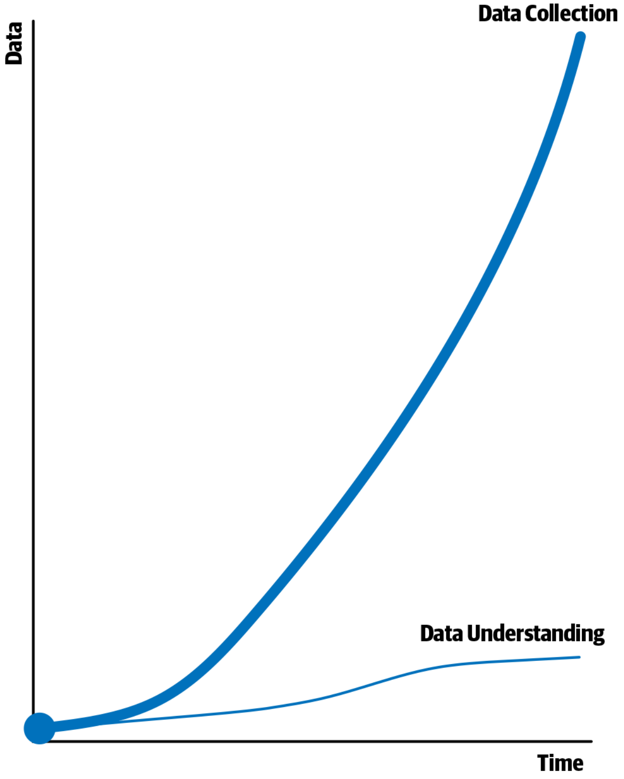


Figure 1-1. Graphical representation of a typical organization's data collection and data understanding capabilities

While you can argue that organizations simply don't have the ability to understand the impact of the data coming in (this book will change that), what about the "things" they used to know but have forgotten? That's context. We've all experienced this in our personal lives. For example, when your favorite airline delays your plane for the third time in a month, but the customer service agent who is rebooking your flight has no idea what you've been through as they work on your case, that's a lack of context. Imagine if the agent proactively apologized and upgraded you on your return flight. Here is a missed opportunity for a great client interaction because the airline, and the agent, lacks context. They fail to provide empathy, because they literally have no idea this is the third time their operations have altered your flight plans...but they do know it...but it's not well known enough for anyone to act on it (the context is forgotten).

Consider all the events any company already knows about, or could know about. Do they apply this knowledge to a 24/7 decisioning environment? For example, an electrical power company knows what a compromised power tower looks like and understands its characteristics: perhaps a blown transformer, rusting bolts that support the infrastructure, or encroaching brush and trees that might catch fire. This same provider has also likely recorded the impact that rainfall and salinity have on its infrastructure. But has it turned that recorded knowledge into simulations, to predict when a tower needs proactive maintenance? Does this company have a static time-based maintenance protocol or is it using conditioned-based monitoring to trigger maintenance routines? Does it use drones with computer vision to rapidly and more safely inspect those power towers? If the answer is "no," these are signs of Enterprise Amnesia.

If organizations start applying data acumen (a term we'll use to include AI, machine learning, and deep learning, as well as other approaches we discuss in the book), they can generate a new data collection curve (the dashed line in Figure 1-2). This curve will capture some of the value hidden in the amnesia abyss (the areas between the thick and thin lines) and open up new opportunities for top-line growth, better service, and better outcomes.

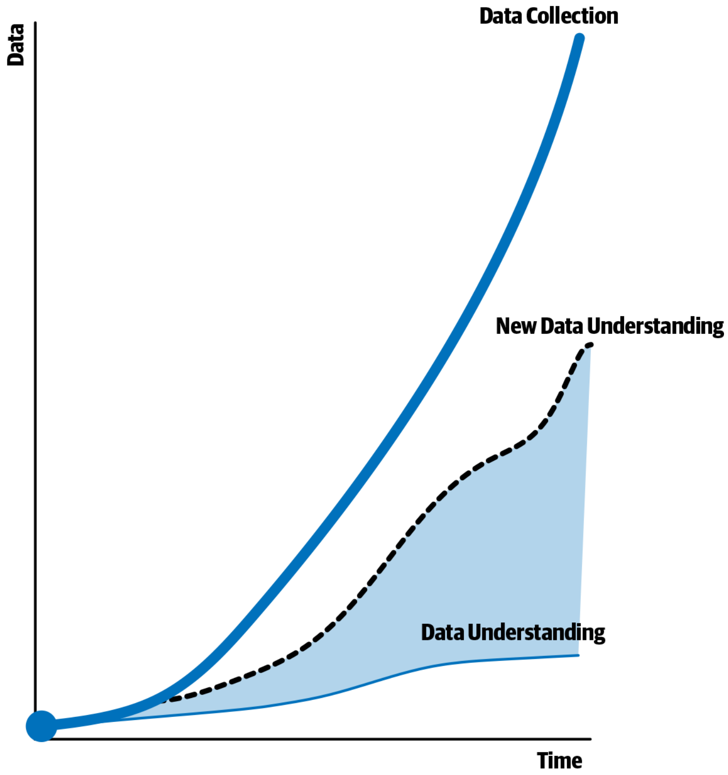


Figure 1-2. Organizations that apply data acumen will notice a greater opportunity to correlate data collection and data understanding

Notice that while this new dashed data collection curve is sloped more steeply, it's not a straight line: the curve has humps, lumps, and bumps. Modernizing your approach to data, and thus generating a new data collection curve, is a highly agile process that will encounter failures, success, and restarts. Culture matters here too (more on that later in this book).

The Modality of Everything and the Data Collection Curve

It's important to understand that the ways we interact with our environment, both physical and virtual—the "modality of everything"—are changing. These changes set new expectations for the talent you will be recruiting and the way you engage the value chain (from material sourcing through to your customers). Most of all, these changes to how we interact with the world make the data collection curve steeper—and that makes it all the more important for the data understanding curve to keep pace.

We have daughters the same age. Neither of them had any idea what a 3.5-inch disk was when we first showed it to them. The modality of storage has changed so much in the past few decades that all our kids truly know is "the cloud" (if they see a thumb drive in our hands, they remark "OK Boomer…"). The future of storage feels like "anything you want and as much as you want."

The modality of expressions has changed, too. Expressions that used to be textual are now visual. Our kids don't communicate via email; they use visual-first expression platforms like TikTok, Instagram, and Snapchat, fully equipped with virtual bunny ears and alien eyes, along with other filters that we "Boomers" (neither of us are Boomers, but it's Generation Z talk for anyone over 30) see no reason for.

Watch how people interact with their technology today. How many mouse clicks and scroll wheels do you hear in an office these days? The modality of interaction has changed from scrolls and clicks to touchpad swipes and gestures, where the breadth of a pinch or intensity of a touch means something. We call this "digital body language." The amount of digital body language that companies have collected has exploded in recent years. What hasn't exploded is the ability to make sense of it and turn it into actionable insights.

Today, we live in a world where everything can be measured. As the Internet of Things becomes the Internet of Everything, edge devices bring more data to an organization's doorstep than ever before. The ultimate goal would be to morph data from the Internet of Everything into the Intelligence of Everything. But that isn't going to happen with the current slope of the data understanding curve.

Soon our primary interaction with technology will be through voice. Anyone who uses a voice-driven modality knows we're not there yet, but that is changing quickly. Newer deep learning techniques will make voice interaction more trustworthy in the coming years, and that means more data. As you can likely deduce, the modality of everything will further steepen the data collection curve.

Even Steeper: The Future of the Data Collection Curve

Here's the very real and unfortunate-for-many news: the data collection curve is about to get much steeper. Let's discuss some examples:

Blockchain

First, stop thinking about Bitcoin and cryptocurrency, and force yourself to think of blockchain as a distributed trust protocol with the potential to redefine business models. Blockchains are certainly used in supporting cryptocurrencies, but there are many other applications where trust is costly and needed. From income-share agreements (which we think will disrupt the student loan market) to food supply chains (still struggling with traceability, and now being pressured to move toward transparency), remittance payments, the settling of financial transactions, getting control of the opioid crisis, document exchange, portable medical records, trusting a stranger to drive you somewhere or deliver you what you ordered, mitigating fraud for fast-moving aid payments during a crisis, and everything in between, blockchain technology has the potential to create some of the biggest data sets we've ever seen.

Bots and assistants

Sophisticated bots generate enormous amounts of interaction data. That data will enable developers to train algorithms to determine the next best action, analyze tone and intent, cross-sell, up-sell, substitute-sell, spot identity fraud, and more. Those algorithms will also be run on new interactions, and so the "learn-from→apply-to" cycle begins.

Bots are set to redefine baseline business-to-consumer interactions. For example, chatbots allow brands to personalize their marketing. They can sit natively in messaging apps (where people "hang" and interact), they can be iterated and deployed quickly, and more. But one study in particular speaks volumes to us: it shows that using intelligent bots results in a spike in customer engagement. Servion's study estimates that by 2025, AI-powered bots will sit behind 95% of all customer service interactions. How much will these agents steepen the data collection curve?

Weather

We're not suggesting that businesses will collect their own weather data, but they will decorate their data with it and apply it to their business to create what we like to call "the moment of Wow!" For example, as 2017's Hurricane Irma bore down on Florida, Tesla pushed a software update to some of its models, giving extra mileage to help owners get further away from danger and closer to safety. Wow!

Consider how weather relates to insurance. Auto insurance is one heck of a tension-filled business. We all know it well: you pay premiums and have a great record, and then some event happens that isn't your fault; you fight over the impact and cost of that event and watch your rates climb. You know the irony of this situation? Both parties want the same thing: not to have experienced this event. A single hailstorm in Phoenix once caused $10 million of claim damage. What would happen if your insurance company provided you with an asset registration app that warned you of an incoming weather event, allowed you to temporarily register your asset's location, and gave you tips and suggestions for avoiding (or minimizing) a claim? This isn't the stuff of fantasy. One insurance company built such an app, and found that its registered policyholders in a certain region acted on 50% of its alerts. Of those that acted on alerts, only 6.1% of them filed a claim. What's more, weather events gave this company an average of 10 opportunities to communicate with their clients as "partners," outside of the regular premium renewal process.

5G

5G cell technology won't just mean the ability to download a movie in seconds or charge your phone once a month. It will catapult the world into an augmented reality (AR) modality, for everything from insurance inspections to changing the oil in your lawn mower to the experience of buying a boxed item (check out the LEGO store in New York City). This bigger "pipe" means more data. 5G is the battle for data supremacy, and it needs to be considered in any AI or data strategy.

Wearables

Wearables and other smart technologies will bring unfathomable amounts of data to our doorsteps, all increasing the cost of not knowing: imagine toilets that screen urine for diseases, or floors that can not only identify the individual walking on them with more accuracy than a fingerprint, but can also predict hip or knee deterioration. (You don't need to imagine any of this, it's here and happening today.)

Where We Are Now—Haystacks, Needles, and More Data

Faced with the challenge of catching up, are you feeling that overwhelming desire to give up right now? Fear not. Everyone has had it (or will get it) before they start climbing the AI Ladder.

Together, we have about 50 years of data experience across thousands of client interactions. We were around when data was about finding a "needle in the haystack" (data warehousing). We saw many declare the end of warehouses and watched organizations go head over heels for Hadoop, expecting to glean unlimited insights from data lakes that morphed into data swamps. Essentially, many thought they could find the needle if they added more hay to the haystack! (Don't get us wrong—there are successful Hadoop projects, mostly in data preparation and online query archives, but widespread analytical insights never materialized, for reasons outside the scope of this book.) We've seen mistakes made, and we've made our own mistakes (we have the scar tissue to prove it), but we've learned quite a lot in our collective half-century of data strategy observations. That's where the information architecture we talk about throughout this book comes from.

In today's reality, we are looking for needles in stacks of needles. That data collection curve is becoming too steep to keep up. We as humans are going to need some help, and this help comes in the form of AI—AI that can watch and observe, feel, listen, understand, annotate, categorize, transcribe, sense, translate, compose, perhaps even smell! This AI isn't meant to replace humans; it's meant to help us because we simply can't keep pace.

AI will change every job out there today, and if you bury your head in the sand on AI, you're likely to miss out on the data understanding curve altogether. But stop and consider the potential for collaboration between humans and AI. Humans are capable of compassion, intuition, design, value judgment, and common sense (we're tempted to toss in a joke here...but we did say capable). When we think of computers, we think instant recall, discovery, large-scale math fact checking, immunity to mind-numbing work, never taking a break. We're good at things computers will never be good at, and computers are good at things we'll never be good at. Collaboration—joining our strengths—is just common sense. And humans are supposed to be good at that.

There's no question that Robotic Process Automation (RPA) is coming. It will help to eliminate mundane tasks associated with many business processes: the tasks most employees would gladly relinquish. The goal is for employees to view RPA technologies "teammates" of a sort, willing and able to perform repetitive tasks without complaint, and unmatched in terms of speed and accuracy.

AI will certainly cause some displacement. That's what happens with any technological revolution. After all, the invention of the automobile put many hostlers (people who take care of horses) out of work. When the Romans invented aqueducts, they probably put thousands of water carriers out of work. How many people wanted to spend their lives carrying water? We firmly believe that most people will end up declaring, "I can't do my job without this AI technology!"

We are in the early days of a promising new technology, and of the new era to which it is giving birth. This technology is as radically different from the programmable systems that the IT industry has produced for half a century as those systems were from the tabulators that preceded them. World-changing technology carries major implications and responsibilities, but that's outside the focus of this book (though we do touch on bias and ethics).

How to Displace Today's Disruptors

Think about the different ways your business can wrangle competitive advantage. Economies of scale? A time-tested classic for sure! If you're FedEx, P&G, or Walmart, you certainly enjoy cost advantages obtained from scale. These companies are optimized for economies of scale. Because of that, they can innovate and take business from competitors that don't have the same economies. Now stop to consider how many organizations get to experience this benefit.

Network effects are another form of competitive advantage. Be it Facebook (the world's largest media producer, yet it produces no content), Alibaba, or some other, these companies have all built business models from network effects. Facebook works because nearly everyone is on Facebook. Let's face it, most organizations will never see network effects at this scale.

The problem with these two forms of competitive advantage—economies of scale and network effects—is that there are nearly insurmountable barriers to taking advantage of them. They simply aren't available to everyone! Not every company can be the lowest-cost producer, and not every company can build network effects at scale.

Is there a third way to gain competitive advantage, and if so, what is it? Let's start with some observations. Uber isn't really a taxi company, though it is the largest taxi company in the world. Airbnb isn't an accommodation provider, though it offers the most places to stay in the world. Consider the last decade's disruptors—we didn't list them all here, but you know who they are. How did they change things up? There is a new basis for competitive advantage: *data*. That's what's behind Uber, and that's what's behind Airbnb, among others.

What else do all of the last decade's disruptors have in common? They all seized a moment with data. They went to conduct business (sometimes new business) in places they didn't belong. (This is something you can do with data: for example, Best Buy's most profitable division is home healthcare.) They created network effects, took business away from incumbents large and small, and created new business models not imaginable a few years ago.

These disruptors know your searching preferences, your viewing preferences, when you post, where you go, who you talk to, how you change your behavior when your phone's battery drops below 10%, what you buy, and more. They know these things because they understand a lot of the data they collect. Yesterday's disruptors are pushing even harder to collect more data so they can disrupt more broadly and deeply in this decade than the last.

But here's the thing about data: most of the data in this world can't even be "Googled," which means it's not readily available to the last decade's disruptors. That's why they are offering so much "free" stuff: so they can collect more data about you.

Think about it. A bank can recognize someone who is likely to default on their loan; it has seen hundreds of thousands of those cases. An experienced insurance adjuster knows the cost of damages associated with a low-speed collision just by looking at it; they have adjudicated thousands of car claims across all levels of damage. AI is an excellent opportunity for companies to capture this kind of knowledge, which is especially important as the most experienced part of many organizations' workforce moves into retirement.

We've mentioned the lack of a level playing field. Data is the one area where every company has an equal opportunity to be great. Knowing your own data is a competitive advantage available to everyone. Climbing to the top of the AI Ladder and understanding your data will be your competitive differentiator. But you need the acumen (and a plan to put to good use the superpowers that come with it) to become the disruptor and not the disruptee.

You may have noticed that we use the word "acumen" a lot when we talk about data. We want you to think of data in terms of your acumen. This is because even if you have lots of data (hint: you do), it's not much use to you unless you know how to put it to work (that's the acumen part). That's exactly what this book is here to help you do. Be forewarned: the landscape of data acumen is ever changing. We like to tell people to think of their data and data acumen like a gym membership, for two reasons. First, if you don't use it, you'll get nothing out of it. Second, if you stop using it, you start to lose whatever gains you built while using it.

Let's Get Ready for a Climb!

We wrote this book not just to advance your AI business skills, but also to give you acumen on how to start and complete effective AI projects.

Scope of This Book

We think it's worth a moment to level set and tell you what we chose not to write about in this book too (mostly because it's already been done). If you're looking to understand backpropagation and what momentum does to a training run, that's not in this book. Want to know what a convolution is for computer vision? Not here either.

This book is about strategy, consideration, and will serve as a playbook for the things we think will make for successful AI. Bottom line: this is not a how AI works book…it's a book on how to get AI to work. And because this is a book about getting AI to work, we're using a broad definition of AI that some purists might not like. But for the purposes of this book, AI will include its "children," machine learning and deep learning—cognition and learning are key components to the definition.

Furthermore, while you can read this book cover to cover (it's best that way) you can jump to certain sections to help your journey's troublespots.

AI might seem overhyped right now. People are going to overrepresent AI's potential impact on the next year or so. But we're certain they will underestimate its effect on the next 5 or 10. AI is here to stay.

The challenges and hype aside, we're at a key inflection point as a society. Make no mistake about this. Our world is rapidly moving from one where most processes are curated and run by *humans supported by technology* to one where processes are run by *AI technologies supported by humans*.

Apple aficionados may recall Steve Jobs's famous quote: to him, a computer is "a bicycle for our minds." During the brief time Jobs spent in college, he was fascinated by two things: fonts (which is why Apple is so focused on beautiful design and emotional attachment) and the locomotive efficiency of living organisms. As he learned, humans are about a third of the way down the list in terms of the most mechanically efficient moving creatures. Number one is the condor (it caught us by surprise too). However, when a human is on a bike, their locomotive efficiency blows the condor off the charts. It's like a superpower.

If a computer is a bicycle for our minds, then AI is a bicycle for analytics. It's a solution to the quandary we're living in: we store everything but have no way to apply intelligence to it all. The analytics world has always been about programming for insights. Computers are programmed with rules and instructions to perform various tasks very quickly. But the idea behind AI is that it learns on its own, through observations. After all, computers are excellent at finding patterns, so let them do that.

AI (including subdisciplines like machine learning and deep learning) will do for the 21st century what the industrial revolution did for the 18th century. As we said, we are at an inflection point. Data acumen will give some companies (and individuals) a lift and shift, while it will hand others a rift and cliff.

Today, different companies are at different stages in their AI journey. Despite all the coverage and claims, most companies are just starting the journey. Companies still brag about their customer service, their merchandising strategies, their mass marketing campaigns, and more. Underpinning those business "brags" are perhaps some algorithms here and there. As these companies venture further into their journeys, they will privately (and perhaps publicly) brag about how they use AI. But the companies that are really transformed by AI won't be limited to a few algorithms assisting the business strategy and value propositions. They will have thousands of algorithms. And they won't be updated monthly, quarterly, or yearly, they will be automatically updated as needed: daily, hourly, and perhaps even faster. There won't be algorithms designed for cohorts and segments, but a plethora of hyperpersonalized ones for individuals. You'll often hear us say "death of the average and personalized to the one."

We'll make a bold declaration: in the years ahead, we'll stop talking about artificial intelligence; instead, we'll talk about *ambient* intelligence. Why ambient? The definition of the word includes phrases such as "existing or present on all sides" or "an encompassing atmosphere." Much like unobtrusive lighting or background music, it's there but it fades into the unnoticed as you go about your day. AI will become so intertwined in our day-to-day personal and professional routines, it will become ambient intelligence.

As a final suggestion before you start reading this book, spend a moment or two asking yourself, "What is the cost of my company's not knowing?" Now get on your bike and let's start the ride.

CHAPTER 2
The Journey to AI

For the past few years, it's been difficult to escape the constant barrage of news about AI. We've seen everything: skepticism, excitement, analyst pronouncements, news reports about successes and failures, new tools, new platforms, and new problems. We've certainly seen lots of press telling businesses how they can take advantage of AI, and it would be hard to miss how every vendor is telling you there's AI in its product or service. Hype? Certainly. But beneath that hype there's also a reality. From the smallest startups to the largest enterprises, companies are reaching out to AI to improve everything from customer service to fundamental research. What's easy to miss in all the news about deepfakes, Go champions, and the rest is that companies are starting to use AI to reinvent their businesses.

Systems like Project Debater from IBM, along with DeepMind's success in playing Go at a world championship level, have illustrated the potential of AI. However, there is a big difference in the data acumen required for AI demonstrations, consumer applications, and enterprise applications. Digital assistants such as Apple HomePod and Amazon Alexa use AI to serve basic consumer functions. But in many respects, building successful enterprise applications is more difficult than selling a voice-enabled consumer product, or even winning a complex game like Go (an incredible feat). In the enterprise, companies need to use AI to reinvent their businesses. They're exploring a new, unknown territory: new (or redefined) business processes, new ways to satisfy customers, new ways to improve operations. Those companies are becoming more efficient, more effective, more profitable—and they're doing that now, not in some hypothetical future. There's a clear imperative: filter out the hype and seize the opportunity.

What's the scale of the opportunity? Gartner suggests that in 2020, AI was set (no one knows the effect of the COVID-19 pandemic on this prediction) to become a net job creator, eliminating 1.8 million jobs while creating 2.3 million new ones. Of course, net is net; it doesn't account for individuals. When it comes to the effects of AI on employment, our inclination is to agree that AI will create more jobs than it eliminates, but we believe the jobs it creates will be different from anything we've seen before. We're used to talking about "blue" and "white" collar jobs, but we believe we will see a whole class of "new" collar jobs that involve deep interaction and collaboration between humans and AI.

While no one can be sure what the net job effect will be, we can assure you that people who are comfortable using technology and AI in their day-to-day jobs will replace those who aren't. Why? Because AI gives superpowers to frontend staff, inspectors, oil rig workers, medical practitioners, bankers, risk calculators, and more. Bottom line: AI won't replace workers as much as it will make them more effective. It will automate routine tasks and assist in more complex situations so employees can focus on higher-value work. According to industry analysts, AI is poised to add $15 trillion to the global economy by 2030. And it is going to transform organizations and change the way people work across every industry.

It's easy to think of AI as futuristic science fiction, but it isn't; it's here now. AI applications are everywhere; we all use them many times each day. When you search for a website or buy something from an online retailer, you're using AI. The recommendations you see when you visit your favorite online store are almost certainly generated by AI, as are the logistics for getting your purchases to you (inventory management, shipping, routing, etc.). AI is behind generational advances in speech recognition, speech translation, image recognition, and autonomous vehicles. It has enabled the creation of voice-activated assistants for your phone and your home, and it plays an important role in customer care, social media, and cybersecurity. Let's consider a few examples.

Humana, one of the largest insurance providers in the United States, reduced costly preservice calls and improved the provider experience with conversational AI. It handles over a million calls per month from healthcare providers to confirm medical benefits and perform other routine tasks. Most of these calls were previously handled by human operators because the providers preferred not to deal with Humana's old interactive voice response (IVR) system. We've all been there: we've navigated voice prompt systems using strings of digits that, by the time we get to the information or person we need, look like hashed encryption keys. To provide a faster, friendlier, and more consistent way for healthcare staff to access preservice medical eligibility, verification, authorization, and referral information, Humana chose Watson Assistant for Voice Interaction. This new solution relies on AI to understand the intent of a provider's call, verify they are permitted to access the system and member information, and then determine how best to provide the information requested. The solution handles most calls without requiring human intervention, reducing cost and workload and increasing client satisfaction. When it comes to designing self-service interactive features such as chatbots, the standout projects we've seen go well beyond coding up simple rules for common questions. Designing a great interactive bot is more about understanding a user's intent (especially long tail interactions) and guiding them through what they're trying to do, and if that requires taking some action (for example, verifying insurance coverage or filling a prescription), taking the action as part of the process.

The Royal Bank of Scotland (RBS) has paved the way for banking innovation ever since it was established in 1727. From the world's first overdraft and the first home purchase loan by a UK bank, to the first fully fledged internet banking service and mobile banking app, this pioneering bank has a history of making life easier for its customers. That's why RBS recently welcomed a new team member, Marge, a virtual agent built with IBM Watson Assistant. Marge helps employees help customers. With United Kingdom regulations for home mortgages becoming challenging, keeping up with regulation was taking the focus away from the customer experience. Marge provides RBS agents with the answers to most questions they may be asked by a customer. The results (Figure 2-1) have been a 20% growth in net promoter score—NPS, a widely used metric for the loyalty of customer relationships—an increase in employee engagement, and a boost in employee confidence when speaking to a customer.

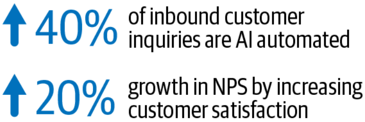


Figure 2-1. The Royal Bank of Scotland saw notable returns on its AI investment

The benefits of AI cannot be measured in money alone. AI offers hope for solving environmental challenges, finding cures for diseases (we were proud to see IBM put AI to work during the COVID-19 outbreak for everything from using its world's fastest supercomputer to uncover compounds that help to understand the disease to supporting a digital Q&A system via IBM Watson Assistant for Citizens), deepening our understanding of elementary particles, elucidating the structure of the cosmos, and greatly improving the health and well-being of some of the most disadvantaged and financially underserved people on earth.

Using AI, people are solving problems that were previously considered unsolvable, such as predicting how proteins will fold by studying the DNA that codes them. Humans are made of more than 20,000 genes and billions of genetic letter codes. These complex arrays are the building blocks of life. They hold secrets that can illuminate many genetic diseases that people struggle with today. This opens the door to whole new fields of research and gives hope that therapies for incurable diseases will be developed soon.

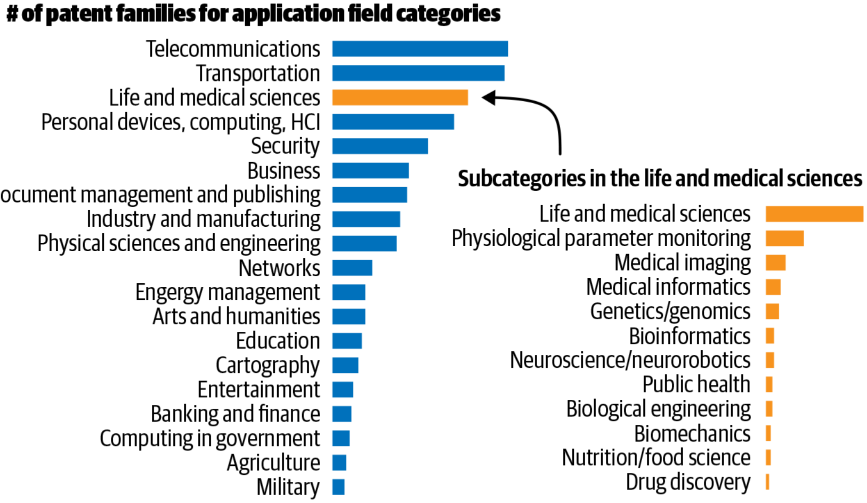
Consider the genetic disease sickle cell anemia. Sickle cell anemia kills hundreds of thousands of people around the world every year (50% of babies born with it die before the age of 5) and afflicts millions more with excruciating pain—mostly those of African descent. Pain from a sickle cell can occur anywhere blood circulates. Red blood cells (which are usually donut-shaped) bend into an inflexible sickle shape, causing them to pile up inside blood vessels. The result is a "traffic jam" that stops the proper delivery of oxygen throughout the body, leading to problems that include pain, bone deterioration, strokes, and organ failure.

Sickle cell anemia comes from a simple DNA "spelling error." Out of the billions of pieces of genetic information in a double-helix DNA strand; to keep it simple, a "T" that should have been an "A" is the reason for this suffering. Fix that error and you cure the disease. And we are on the brink of finding a cure.

The genetic cause for sickle cell anemia was discovered with the use of data acumen. Stop to consider that there are more than 10,000 diseases for which the underlying genetic cause is known or suspected. There are many more diseases that are known to have genetic components, but no causative mutation has been found or implicated. How do you find patterns and fixes in billions of pieces of data? Once you've found the misspelling that causes the disease, how do you diagnose patient DNA and determine an appropriate treatment? Some diseases involve misspellings in multiple genes, and these misspellings are much harder to understand. Tracing the cause requires comparing the DNA sequences of many patients, all of whom have their own unique DNA—pattern recognition is a task well suited for AI. There are many more root causes and treatments to be discovered, and there are many diseases that AI can already help to diagnose, including glaucoma. In the next few years we will certainly see rapid growth in this field driven by collaboration between AI and researchers, because AI will boost data acumen (remember, it's the bicycle for analytics).

Consider the task of monitoring patients with acute illnesses. We've long had the telemetry needed to monitor vital signs. But what do we do with all that data? At the Neonatal Intensive Care Unit at Toronto's Hospital for Sick Children, Dr. Carolyn McGregor partnered with IBM to use data acumen to diagnose life-threatening infections 24 hours before any doctor or nurse could notice any symptoms. What's particularly important is how this class of infections were detected. The problem wasn't that the babies suddenly spiked a fever, which then disappeared before a nurse could notice. It was the opposite: the babies' vital signs didn't go through their normal daily cycles. Their vitals were always normal—too normal—until it was too late, and the infection had taken hold. It's difficult, if not impossible, to imagine a doctor or nurse saying "Something is wrong, that baby is too normal." That's just not how we think. Data acumen discovered and used a pattern that would almost certainly escape a human, and in doing so, saved babies' lives.

According to a 2019 World Intellectual Patent Office (WIPO) report, "Life and Medical Sciences" is the third-highest field for AI-related patent applications (below "Telecommunications" and "Transportation"), with roughly 40,000 patent applications. Figure 2-2 shows the breakdown (a patent may refer to more than one category or subcategory).



Source: WIPO Technology Trends 2019—Artificial Intelligence

Figure 2-2. A 2019 WIPO study shows the number of AI-related patent families by application field (image by Ben Lorica)

We could write an entire book of use cases where AI can help, giving examples in every industry we can think of. Instead we'll give you this guidance for thinking about how AI can help: use AI for *prediction*, *optimization*, and *automation*.

Now, given all the glowing predictions (our own included) of what AI will achieve for us, it's high time to take a step back and define it more precisely.

What Is Artificial Intelligence, Anyway?

History hasn't been kind here. Science fiction about humanoid robots dates back to the 19th century, and arguably to ancient Greek literature: Pandora (she of the famous box) was a humanoid machine, an AI. That tradition doesn't help us, though; if anything, it encourages us to indulge our fears (and in some cases fantasies). But AI is both more and less limited than our fears and hopes suggest. While AI can do harm, it's not going to cause an apocalypse—but no one, not even the ancient Greeks, imagined an artificial intelligence that could avert an apocalypse, whether medical, environmental, or something else.

AI is not magic. It is not the work of sorcerers or fiction; it's computer science. Here's a definition that we think reflects the current state of our technology:

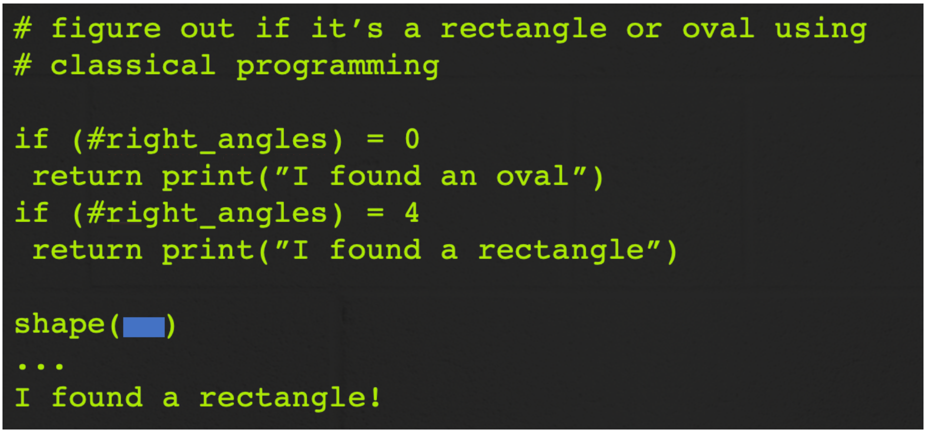
> AI makes it possible for machines to learn from experience, adjust to new inputs, and perform human-like tasks through the combination of math and computer science.

It's important to understand how this differs from traditional computer programming. Programs describe what a computer should do, step by step, in excruciating detail. If the programmer makes a mistake, the program doesn't work correctly; it has a bug, and the instructions need to be revised. Once written, a program doesn't change its behavior unless someone comes along and rewrites it.

For example, let's say a shape recognition program is well coded and seems to generalize to the real world data it encounters just fine, until it doesn't. For example, triangles are suddenly showing up and they are getting classified as ovals; a triangle may or may not have a right angle, but it's certainly not an oval. The programmer needs to develop new rules and because triangles started showing up in the real world data, that not only means running through the coding lifecycle again, but trying to think of every possible combination.

Rather than simply following rote instructions created by a programmer, AI applications "learn" through trial and error, and modify themselves to get increasingly better results. Yes, traditional programming is involved, but that programming is general and has more to do with the handling and preprocessing of large data sets than with familiar business tasks like billing. The "program" is relatively unimportant; it's the training process that determines what the AI will (and can) do. Essentially, AI can surface what the programmer missed more quickly, and handle it, as long as it has data examples.

Quite simply, instead of writing code, you feed data to an AI system and it builds its own logic; you train the AI system with examples instead of instructions. Figure 2-3 shows some code that might be used to determine if a shape is an oval or a rectangle. This is "traditional" programming. A programmer writes the rules, and when you pass a rectangle to the built-in `shape()` function, it goes through its programmer-written, rule-based logic and returns (in this case) the shape…which is indeed a rectangle.



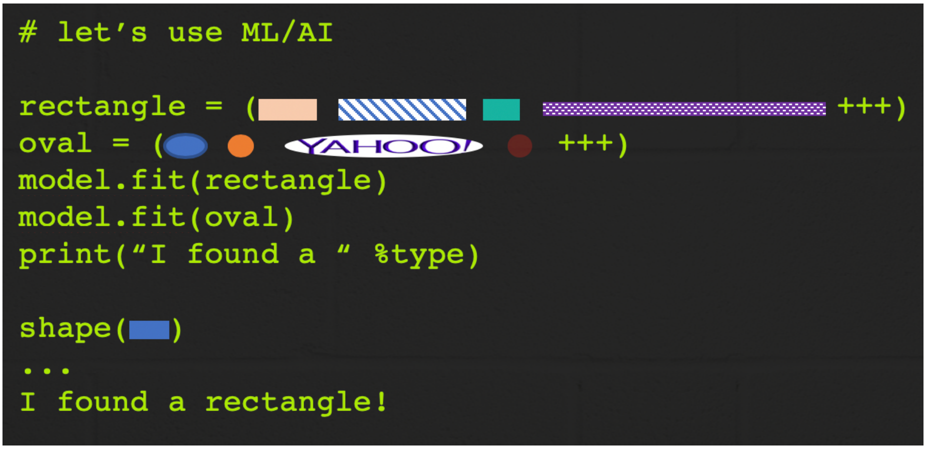
```
# figure out if it’s a rectangle or oval using
# classical programming

if (#right_angles) = 0
 return print(”I found an oval”)
if (#right_angles) = 4
 return print(”I found a rectangle”)

shape(▬)
...
I found a rectangle!
```

Figure 2-3. Simplified code showing how to determine the shape of an object (source: IBM)

What does this example look like in an AI world, where the algorithm determines and writes the rules? In this case, the programmer is involved in getting and preprocessing the data, writing the code that exposes the `shape()` function, and even writing the code that invokes the algorithm training itself, but they don’t write any of the logic. In Figure 2-4, the programmer feeds examples of shapes to the algorithm and the algorithm figures out what an oval or a rectangle actually is.



```
# let’s use ML/AI

rectangle = (▬ ▨ ■ ▨ +++)
oval = (● ● YAHOO! ● +++)
model.fit(rectangle)
model.fit(oval)
print(“I found a “ %type)

shape(▬)
...
I found a rectangle!
```

Figure 2-4. Simplified code showing how a programmer feeds examples to an algorithm and the algorithm makes a classification based on rules it learns through observations (source: IBM)

In our simplified AI example, triangles may (because of their straight lines) or may not get classified as rectangles when you start out. But if you find triangles are getting classified as as rectangles (or classified as unknown), you can simply add examples of triangles (labeled as such) to the training set, and the rules will evolve to distinguish triangles.

If you were building a face identifier, as opposed to a shape identifier, nothing would change. A traditional programmer would write logic that would outline a set of rules that determine who is in the picture, using the features they determine to best identify the person. In contrast, the AI solution would look at all the pictures of the people you want to identify and come up with the distinguishing feature set itself! It might use eyebrows, lips, ears—who knows (it's all mathematically represented by something called *landmark estimation*). How about a classic object classification problem: dog or cat? Same thing. The classic programmer has to write rules about what makes a cat a cat and a dog a dog, while the AI solution writes the rules on its own.

Even the programming needed to create the AI model may disappear in the future. A lot of research in academia and in startups aims to create platforms that can generate AI models without any programming. If an application can "learn" whether a photo is of a cat or a dog, can it also learn how to write the program that decides whether a photo is of a cat or a dog? The answer appears to be "yes"!

Types of AI

Many discussions of AI distinguish between *strong AI* and *narrow AI*. Strong AI (sometimes referred to as "general" AI) is the type of stuff you see in movies, such as Skynet in the movie *Terminator*. It's not something we can do today, or possibly ever, and it's not what we talk about in this book. Narrow AI (sometimes referred to as "weak" AI) is where all the potential and benefit of AI are being realized today. It focuses on single, narrowly defined tasks, though at times it appears to be capable of much more than that. Pretty much everything around us today is narrow AI, from Siri and Alexa to AI beating the best humans in games of chess or Go. The word "narrow" implies the limitations of this type of AI: it can learn how to do one thing and do it really well. You don't take an AI solution that is great at chess and expect it to be great at *Atari Breakout*. That said, you can pipeline multiple narrow AI solutions to expand their capabilities and benefit from previous work via transfer learning—a topic we cover later in this book. Most AI solutions combine several models, each solving one part of a larger problem, and by combining models you can build very powerful (sometimes frighteningly so) applications. But the really scary, futuristic, "science fiction" side of AI? As Andrew Ng once remarked, fearing killer robots is like worrying about overpopulation on Mars. Maybe a time will come when that's necessary—but we have urgent, immediate problems to deal with on earth, and that's where AI can help us.

AI systems are built on *data* and *models*. Let's explore each a bit more.

Data

In order to be useful to AI, data must be in digital form. It can come from anywhere, but it has to be digital. Histories of financial transactions, telemetry from sensors at the bottom of the ocean, science journals in any of a few dozen languages, X-rays from thousands of hospitals around the world, customer purchase records and demographics; in the last decade, we've seen an explosion in the way data can be collected and used. Data can be "structured" or "unstructured"—though the trend for AI is toward unstructured data—and can be stored in any number of digital formats. Today, the easiest place to start your AI applications is with labeled data (data whose outcomes are known).

For example, if you have pictures of cats and other animals that you want to collectively be known as not-cats, label them as such and feed them to an algorithm so it can generate features that determine whether something is a cat or not. If you have tens of thousands of X-ray images of the most common pathogens that make up pneumonia, you need to know the pathogen that each image shows, along with what the image doesn't show (of course, you've collected this data from the radiologists making diagnoses in a hospital for the last decade). Then you feed labeled images to the algorithm to build a pneumonia detector. Dealing with data can easily be the biggest, most challenging part of an AI project. Dealing with labeled data you already have makes it easier to get started.

While it's outside the scope of this book to delve into the machine learning methods that are in use today, we think it's worthwhile to share some of the lingo with you. You'll definitely see it—and you'll need to understand it, at least at a high level. At the highest level, machine learning is divided into three types:

- *Supervised learning* is what we've just discussed (or oval and rectangle example). You have data—say, thousands of pictures of animals—and it's all labeled. The algorithm tunes itself so that it can recognize animal pictures it hasn't yet seen. The same is true for medical images, faces, loan defaults, high-risk insurance policies, and more. Classic examples of supervised learning techniques are linear regression, support vector machines (SVM), *k*-nearest neighbor (*k*-NN), decision trees, and classification.
- *Unsupervised learning* is done with data that isn't labeled. It's a popular technique to use when you don't know what to expect but want to have AI shape the data into clusters based on its intuition. In other words, unsupervised learning is great at finding patterns in the data—things that "look similar"—and letting you decide what the patterns mean; a clustering algorithm like *k*-means is a great example.

- *Semi-supervised learning* is a growing area of interest because it takes a lot of time to curate good, clean labeled data. Using semi-supervised learning, you can augment the labeled data you have with new data or mix labeled and unlabeled data together.

Deep learning, also known as "deep neural networks," is a specific approach to machine learning that uses all of the aforementioned techniques; it has been very successful in recent years, and it's the basis for much of the current boom in AI. Deep learning requires much more data and compute power than simpler machine learning algorithms, but its performance ("accuracy" in AI speak) reaches superpower level for certain use cases. In turn, deep learning can be used to implement:

- *Reinforcement learning*, which works on a reward system to optimize the outcome. Think of a blind rat trying to get through a maze where a piece of cheese awaits as a prize. As the rat hits a dead end, it makes a note of this in its mental model. It keeps adjusting its steps as it tries again, moving on until it finds the cheese. You're most likely to come across video games today showcasing the benefits of reinforcement learning, but its application will completely change the Robotic Process Automation (RPA) space.
- *Generative adversarial networks* (GANs), in which AI systems compete with each other (hence "adversarial"). GANs are applicable in many areas, including gaming and strategizing. We've even seen them in applications where AI algorithms learn to create artworks! This is an exciting area for many reasons, but using GANs to generate more labeled data is a really cool emerging strategy for making better trained algorithms. GANs can also be put to use by bad actors: "deep fakes" are a serious problem today and are only going to get more concerning as this technology evolves.
- *Transfer learning*, in which knowledge gained from one problem is applied to another. For example, can a system developed to recognize cars be used to accelerate the building of a system used to recognize cranes? Can parts of a system developed to recognize diseased tomatoes be used to recognize bruised apples? This is a hot area and we're starting to see the fruits of it now—meaning that we don't have to go back to square one when facing a new problem that's similar to an older problem. Transfer learning makes developing working models easier and, in addition, greatly reduce the computing power needed to develop those models. In fact, no one building a new computer vision project starts from scratch anymore—they leverage transfer learning for baseline detection of curves, lines, and so on, then evolve the model to classify their business domain. (After all, a circle used to detect a wheel to classify a car from a boat is no different than a circle used to classify an apple from a banana—it all starts with a circle.)

Models

Models are programs that process data and learn to recognize patterns in it. There are many AI techniques used to do this. Some of them employ what are known as "neural networks" and others don't, but what all AI techniques have in common is that models are "trained" on a subset of the available data.

In our example (Figure 2-5), an AI model was trained to recognize pictures of cats and dogs by showing it pictures of cats and dogs and telling it "that's a cat" and "that's a dog."

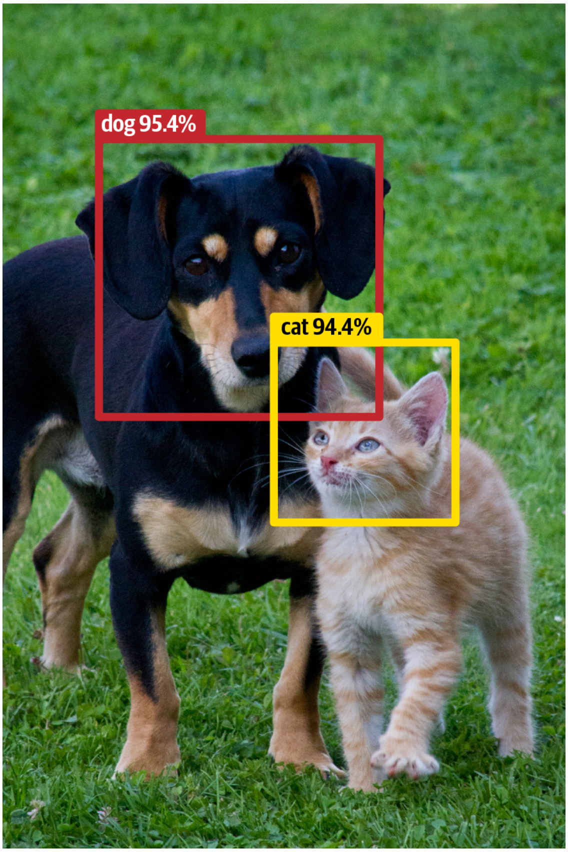


Figure 2-5. AI models are trained to recognize pictures of animals by repeatedly exposing them to labeled pictures of the animals it is trying to detect

Over time, the AI model determined the patterns in the data that make some images a "cat" and others a "dog." Once the model has been trained, it can "look at" pictures it hasn't seen before and determine whether the picture shows a cat or a dog. That decision is typically accompanied by a confidence score, which is very important in enabling AI to work alongside humans. Perhaps the AI model comes across an animal that is neither. It might classify that animal as a "cat" or a "dog" with a very low confidence score, or it might classify it as "unknown." The programmer decides what the threshold for the confidence score should be; for example, they might specify that any prediction with less than a 70% confidence score should result in a classification of "unknown." As the AI model is trained, it gets feedback on whether its pattern matching was right or wrong. It adjusts itself accordingly, continually updating its algorithms. When the AI model has been trained, it should be able to recognize cat images that it has never seen before. It has "learned" which pictures are cats. Good AI systems continue to update their models automatically even after they are deployed in production.

The same kind of process is used by email systems to determine which messages are spam. Again, you start with a collection of messages that need to be labeled either "spam" (like that note from the African general looking to give you $10 million if you help him move the rest of his fortune off the continent) or "not spam" (like a note from your boss). Training the model involves processing these sample messages (our model's training data set), determining which are spam, and then testing on a set of known messages (the test data set) to see whether the model has been trained correctly—that is, if the error rate (spam messages missed, good messages incorrectly labeled as spam) is acceptably low. When the error rate is low enough, it's time to try the model on real-world messages—data it's never seen before.

A couple of things are worth noting at this point. First, it's convenient to say that the model learns the characteristics of spam or of cats, and at some level, that's true. However, it's often very difficult to point out exactly why a sophisticated model decides that a message is spam, or that a picture of a kitten is a cat. Ensemble modeling (stacking together many different algorithms) and dimensionality reduction (reducing the number of features used as input to the algorithm by combining them, dropping some, or engineering new hybrid features) further compound this problem. The most competitive algorithms today aren't simple, straightforward ones; they are complex mishmashes that involve multiple models, often of different kinds. ("Boosted" is a popular term that refers to a class of these ensemble algorithms.) This leads to the domain of *AI explainability*. Some audiences don't care whether you can explain what the model is doing. But for some applications, such as determining whether or not to offer someone a loan, that explanation is crucial and, in many cases, legally required (the European Union's General Data Protection Regulation requires that anyone whose request for credit is denied by an automated process request an explanation).

Second, not all errors are the same. The cost of misclassifying spam email is usually low, but the cost of misclassifying a cancer diagnosis might be quite high! Both false positives and false negatives matter. While it's certainly more important to ensure you don't miss cancer, giving someone a false cancer diagnosis can have devastating consequences. If you're building an autonomous vehicle, the cost of avoiding a pedestrian who isn't there is low (as long as you don't run into something else), while failing to detect a pedestrian can be catastrophic, as we've seen in the news. These are all decisions that the engineers building AI applications will need to understand and think about—and to be honest, not enough people have been doing that, until now.

This isn't a history book, and we don't intend to focus on theory, but it's helpful to look a bit at the history of AI to find clues for avoiding mistakes and identify misconceptions that have held us back. We'll do that next.

Where AI Has Been

According to Ruchir Puri, chief scientist and IBM Fellow at IBM Research, in the period from the 1960s until now, AI research has gone through three phases. Each phase began with high hopes and great promise, and each ended with a "winter" (a period of disinterest and hibernation) when results fell short of predicted outcomes and promises of AI breakthroughs were not kept. Some might look at the current (fourth) phase and think of Jon Snow's *Game of Thrones* "Winter is coming" warning. But we don't see any winter on the horizon, despite some over-hyping of AI. To put it simply: AI is not magic. Glass slippers turning people into princesses is magic. Instead, what AI can deliver is *superpowers*.

The first phase was what Puri calls "symbolic" or "logical" AI. The approach during this phase was based on "rules-based" systems, in which rules were formulated that attempted to mimic human intuition. For example, to tell the AI how we humans determine whether or not something is a cat we might come up with a set of rules like:

- Rule 1: It has two ears.
- Rule 2: It's furry.
- Rule 3: It has whiskers.
- Rule 4: It has four legs.

And so on. These rules can be implemented by a complex series of if-then-else statements. The rules-based approach led to the development of so-called *expert systems*. Expert systems were very useful in some contexts, and were particularly common in the 1980s, but were very limited. They tended to be extremely fragile (back to our classic programmer and their shape-classifying dilemma): the if-then logic breaks down as soon as it encounters something unexpected. If anything changes (the environment, the data, the pattern), rules-based systems have to be manually updated and redeployed, which takes a long time and makes them more fragile and prone to error.

The second phase is what Puri calls “connectionist” AI. During this phase, programs were written to create systems based on the way the brain works in biological systems, with neurons that interconnect with each other and create feedback loops in order to learn. This was the inception of what today are called “neural networks,” and it’s the basis for much of modern AI. But for a long time, this approach too failed to meet its initial promise. We now know that this approach was on the right track, but the memory capacity, processor speed, and the amount of labeled data necessary for neural networks to succeed were far beyond what was available to researchers at the time. The connectionist approach thus fell out of favor and experienced its own “winter,” perhaps better termed a “slumber,” until the world had the compute capacity and the quantities of labeled data that were needed.

The third, more recent, phase of AI was the so-called “big data” movement. This approach was an early prototype of today’s AI, even though early data scientists were careful to dissociate themselves from AI (which many viewed as discredited). It used lots and lots of data, and it created models that learned from the data. But the algorithms used to look for patterns were often not sophisticated enough to find the patterns humans were interested in, and they had to be explicitly programmed to look for them. Big data had some successes, but it had many disappointments too, and the consequences of those disappointments are still being felt by many high-investment, low-yield projects today.

So what has changed? Why are we saying that AI is finally here, that this time the educated promises are being kept (ignoring the fantasy ones)? Underlying each of these phases were four common themes:

- Not enough high-quality, trustworthy labeled data
- Not enough computational horsepower or adequate processing techniques
- Algorithms that were not up to the job or not sophisticated enough to activate the data
- Insufficient investment in people and resources (skills)

All of these problems have been addressed over the last 10 years or so. How?

First, we’ve learned that training models to recognize patterns in data requires lots and lots of data. And it must be good-quality data. In the early days of AI, researchers either underestimated the amount of data they would need, or didn’t fully understand what “high-quality data” implied. If they did understand these things, they were unable to obtain such data for their particular projects. Over the last decade, the field of data analytics has greatly matured, and good-quality data now exists in abundance. The rise of social media was revolutionary: rather than having to take thousands of photographs yourself to train your algorithm, you could grab millions of photographs from Flickr—a practice that has since become controversial, but was crucial to the resurgence of AI for computer vision. Similarly, instead of scanning thousands of books to train language models, you could download the entire corpus of Wikipedia (15 GB in English), in dozens of different languages. (The irony is the amount of available labeled data is becoming a problem yet again; not because there isn't a lot of data out there, but because the world is trying to build bigger and more powerful models. This has given rise to areas of research around generating labeled data, using those GANs we covered earlier in this chapter.)

Second, we've learned that training a model to recognize patterns in lots and lots of data requires lots and lots of computational resources. Using computers that were available in the 1960s, '70s, '80s, and even as recently as 10 years ago, modern AI techniques would be at best impractical and more likely impossible. But once it was realized that graphical processing units (GPUs) could be adapted to AI tasks (because they are so good at matrix multiplication, a cornerstone of AI math), and that it was possible to harness thousands of GPUs together, modern AI became within reach.

Third, the evolution of algorithms represents a major breakthrough for AI. In Puri's words, we have entered the "neuro-symbolic" era of AI. Modern AI algorithms and activation functions use both connectionist and symbolic/logical approaches, depending on the problem being tackled. That is to say, AI is now being effectively harnessed to improve AI. These algorithmic advances will be key as we continue to see an increase in the quality and quantity of data, as well as in the power and efficiency of computational resources.

Fourth, we are finally making the investment in people and resources necessary for AI projects. While there is still a shortage of AI developers, AI is no longer confined to research institutions like universities and the R&D centers of large corporations. Until recently, it was often seen as a hermetic and even a somewhat fringe discipline that didn't produce a lot of significant results. The occasional breakthroughs that AI did deliver were often inscrutable even to other computer scientists. Today, those companies that will differentiate themselves and "win" won't be those that leave AI to the domain of the privileged few, but rather those that democratize AI for the many across their entire business.

Colleges, universities, and even boot camp hackathons are training AI developers by the thousands. Beyond academia, thousands of working developers are training themselves on the job by taking on AI-related projects, working on their own projects in their spare time, collaborating on open source libraries, and learning from massive open online courses (MOOCs). The proliferation of open source frameworks for AI development has made it much easier for developers to get started. State-of-the-art (SOTA) algorithms are freely distributed in open source libraries, and massive computing power is available by the minute (even by the second) from cloud providers. Bottom line: you can experiment outside of academia, and without buying into an expensive commercial software package.

So that's where AI has been. Now, what can it do, specifically for businesses, today?

What Does AI Mean for Business?

For businesses, AI can best be understood in terms of the kinds of problems it can help you solve. As a manager, you don't care what kind of neural network is embedded in your AI application. For your purposes, AI is simply a set of techniques that can radically improve three things: predictions, automation, and optimization.

First, AI is about *predictions*—organizations want to be able to forecast what's going to happen in their business, at both the macro and micro level. You've almost certainly experienced e-commerce sites that recommended purchases. That's prediction: AI is predicting what you will want to buy. AI applications can also predict customer demand, financial performance, and the onset of chronic disease, among many other things. Related to this is the ability of AI to classify data, be it text or images. A lot of work on AI has started with image classification, in part because classifying images is relatively easy (and candidly, we the authors have had a lot of fun with it). We believe that natural language processing (NLP) and natural language understanding (NLU) will become the "nervous system" for many AI applications, with text classification becoming foundational for many use cases. Today, companies use text classification to flag inappropriate comments on social media, understand sentiment in customer reviews, determine whether email is sent to the inbox or filtered into the spam folder, understand counterparty risk in large complex loan agreements, and more.

Today's ever-connected and ever-social world makes AI opportunities seemingly endless—it's truly a "use your imagination" exercise when it comes to use cases. Consider the rare and beautiful whale shark and its spot patterns. Much like a giraffe or cheetah, each has its own distinct identifying pattern (like a human fingerprint). Wild Me, a conservation society interested in using technology to "tell the amazing story of animals' lives," built an AI solution that combs through thousands of hours of tourists' YouTube videos to locate and extract still images of whale sharks. It then identifies, analyzes, locates, and virtually tags them. Using this data, this conservation society can predict whale shark populations, migration patterns, and more, and use this information to affect policy; they were able to elevate the whale shark's status to endangered because of this AI project.

Wildlife conservation using social media is a huge (and not too well known) use case for AI-based computer vision. The takeaway (aside from making the planet a better place) is that social media provides a flood of free data that can be used to create wildlife population models, assess whether protected areas are boosting population numbers, find new hotspots for conservation, and more.

Next, AI is about *automation*. There's tremendous value in automating critical yet time-consuming business processes that are often done manually. When AI takes over this kind of work, it frees employees to focus on higher-value, more creative work. With Humana, we saw that AI was capable of handling many of the incoming calls on its own, without human intervention; it passed the problems it couldn't handle on to human agents.

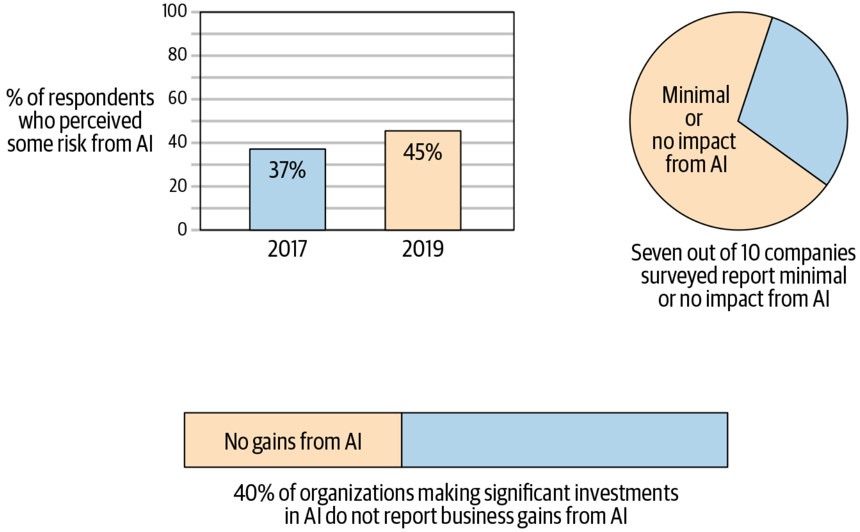
Finally, AI is about *optimization*, whether of routing and logistics, marketing expenditures, the configuration of your cloud installation, or any other elements of the business. AI has been successfully used to minimize HVAC costs. AI can also be used to manage supply chains, maintain adequate levels of inventory without overstocking, and make decisions about pricing. All of these applications require tracking huge numbers of variables; in some cases, this can add up to a dizzying array of "features," much more than a human could possibly think about. That's no problem for a computer.

Despite many success stories, practical adoption of AI is still slow, compared to what it could be. Many companies have started pilot projects, but relatively few of those projects make it to production, let alone systematically change the way the business thinks and operates. Despite the promising forecasts, techniques that have been developed and validated by computer scientists and researchers in the field of machine learning are not being employed by industries and organizations that would clearly benefit from them. Why?

The Journey to AI

We think the journey to AI will represent a lift, shift, rift, or cliff for individuals and the organizations they work for. While AI is the biggest opportunity of our lifetime, it can also be seen as a risk, a far-off utopia with no clear path for organizations to follow (Figure 2-6). When should organizations pursue their journey to AI? What do they need to get started? What can they expect along the way? And how do they know when they've arrived?

In a fireside chat with Tim O'Reilly, we spoke about some of the cultural changes that we think organizations will need to make before they can succeed with AI. First, they need to overcome the fear factor. It has been argued many times that, although AI may threaten the jobs of some workers, the biggest threat is likely to middle management. As we implied earlier, managers who use AI will replace managers who do not. So it's not a surprise that some people feel AI is a threat. And if they see AI as a threat, they are likely to resist it, rather than take advantage of it—even if resistance isn't in their best interest. It's difficult for them to see it as an enabler of new practices and ways of working that can increase their productivity.



Seven out of 10 companies surveyed report minimal or no impact from AI

40% of organizations making significant investments in AI do not report business gains from AI

Figure 2-6. AI can be seen as a risk for some organizations, as recent statistics show; statistics from 2019 MIT Sloan Management Review "Artificial Intelligence Global Executive Study and Research Report"

Given how familiar enterprise resource planning (ERP) projects are to existing IT staff, it's important to understand why AI projects are different. Treating them like the projects you've done in the past is a mistake that will lead to failure. ERP projects are large projects with big budgets and significant staff; a company may only have one or two ERP projects in progress at a time, but each project could easily be staffed by dozens of developers and managers. What's more, these are often "can never fail" projects. The entire company is "IT-frozen" and fingers crossed when they are implemented, migrated, or switched over.

Unlike ERP, which is well established and understood, AI is fundamentally experimental. You need to have tens, if not hundreds, of experiments in place, and you need to expect many of them to fail and be restarted. Unlike with traditional software, where we understand how to measure progress reasonably well, we don't have those kinds of metrics for AI development. A practitioner can spend weeks or months trying to get a model to work, only to decide that it needs to be abandoned—or to find something else that works beautifully and that they could have tried on the first day. That's not a symptom of bad planning; that's life in the AI world. And given that, it makes sense to start a number of experiments, expecting failures along the way. It also makes sense to start with the right experiments: experiments for which you have the data, that will move the needle on some metric, and that are well defined and not overly complex. What is that metric? Don't walk into the boardroom and promise to completely change the business; start on something small that shows business value.

Start with the data you have, with some tasks that are mundane and that humans can do with very little think time. Prove it out. Experiment—do lots of experiments—but don't bet the company (or your reputation) on AI until you've established a track record that shows you can win. Simple, well-defined, important, and data-rich: that's the short recipe for AI success. Fail fast and fail safe, we always say.

AI will not eliminate your staff, but their roles will change. AI, and machine learning specifically, requires an entirely different way of thinking about software:

- Less time will be spent doing traditional programming. More time will be spent defining problems very precisely. Humans are great at working with vague problem definitions; computers, especially artificially intelligent ones, are not. It's easy to marvel at feats of AI, like beating world champions at chess and Go, and to forget that the rules of these games are extremely precise definitions of the problems to be solved (that's why it's called narrow AI).
- AI is all about data and statistics. You will need staff—data scientists and data engineers—who understand statistics, and who understand how to work with data. Since the use of data will increasingly be subject to regulation, you will need staff who understand concepts like data governance and data provenance. Putting the slogan "Move fast and break things" into practice without safety measures, explainability, and data understanding will be a ticket to a lawsuit or investigation.
- Data and statistics are just a pile of numbers unless you know what they mean. A key part of your AI team will be people who know your business backward and forward. They are the business domain or subject matter experts (SMEs). They will come from other parts of the company, and may have limited technical background, but their role won't be technical. They will help you to define relevant problems to solve (avoiding data science projects that techies are renowned for building; you want business problems solved with AI) and understand whether your results make sense in the business context. For example, if you want to optimize your sales pipeline, they will tell you what has to happen to close a deal.
- AI specialists need to work with SMEs who understand the business, the business context, the regulatory environment, and so on. In turn, the SMEs will have to adapt from their current roles (which may involve areas like strategic planning) to work with the AI specialists. It's a symbiotic relationship: SMEs help AI developers create AI applications that, in turn, further the AI team's understanding of how the business works (and presumably the SME will better understand AI and how it can help the business). It's all too easy for AI on its own to optimize the wrong thing.
- Operations staff are already familiar with concepts like version control and monitoring. AI will stretch these concepts; in addition to version control for code, AI will need version control for training data and models. In addition to monitoring

for performance (when discussing AI, *performance* means the accuracy of the model against data it has never seen before), load, and other issues, they will need to monitor AI applications for fairness, equity, and relevance. Unlike traditional software, AI models tend to grow stale over time, and need to be retrained.

It has been reported that upwards of 80% of business leaders do not understand the data and infrastructure required for successful adoption of AI technology (Figure 2-7). So if you feel a bit bewildered, don't worry, you have plenty of company. That's one of the reasons we wrote this book—to help you! But if you're in a position of leadership, you need to start, right now, to learn how to transform your organization to make it AI-ready. Remember: lift, shift, rift, or cliff.

Will you encounter challenges during your transformation? Definitely. Will some of these challenges be due to internal resistance within your organization? Certainly. But all of them can be overcome, and the companies that rise to meet them will prosper.

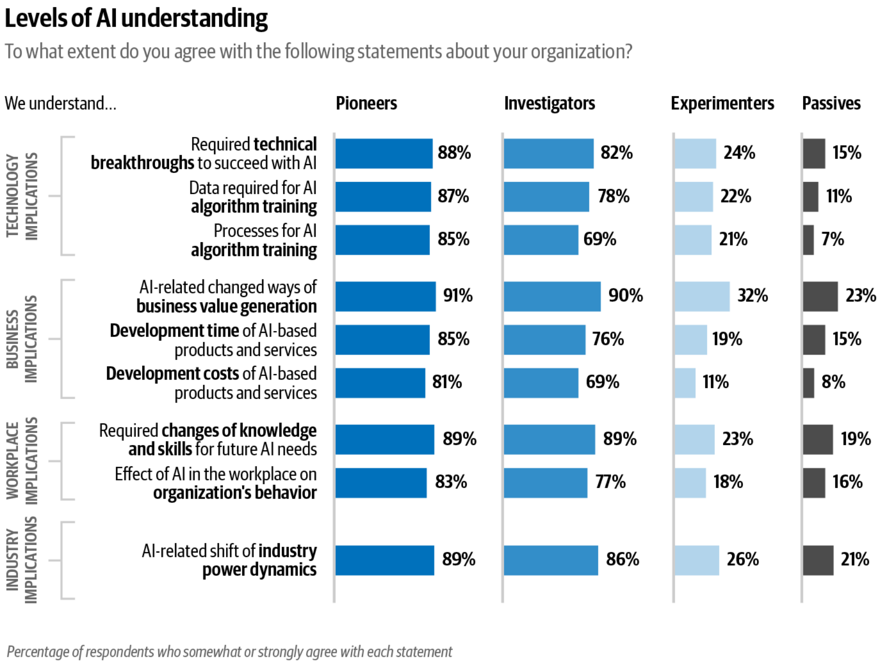


Figure 2-7. Varying levels of AI understanding; statistics from: 2017 MIT Sloan Research Report "Reshaping Business with Artificial Intelligence"

All Radically New Technologies Face Resistance

AI initiatives help organizations predict and shape future outcomes; allow people to do higher-value work; automate decisions, processes, and experiences; and reimagine new business models. Ultimately, this means increased value for organizations, however your organization defines "value"—for example, in terms of revenue, innovation, or opportunities for individual human growth and satisfaction.

AI, however, is often seen (and implemented) as a "black box" whose logic is beyond the comprehension of mere mortals. This is the interpretability part of AI we talked about earlier—integrating AI into our daily lives will require interpretability to increase its social acceptance. Part of this mystical aura can be attributed to the name "AI" itself. The term "artificial intelligence," when applied to modern machine learning technologies, is a (perhaps unfortunate) artifact of its ancestry in the efforts of computer scientists as far back as the 1960s who were attempting to create computer programs that could simulate human intelligence. Going back to HAL 9000, the AI that runs amok in the 1968 science fiction movie *2001: A Space Odyssey,* AIs have been seen as hyperintelligent beings pursuing their own agendas, and not always on the side of humanity. In this mischaracterized environment, people are understandably wary of AI. People don't naturally gravitate to things that they neither understand nor trust.

As Andrew Ng, cofounder of Coursera (and AI celebrity), has said, "AI is the new electricity." Think about it...when electricity was first harnessed by humans, it was considered to be the domain of sorcerers: a magic power that left audiences puzzled about where it came from and how it was generated. Eventually, of course, electricity was demystified and it is now accepted by humans everywhere as just a natural part of the world—but it undeniably changed how things were done, eliminated some jobs, and ultimately created new ones (some never before imagined). All transformative innovations go through a similar evolution: discovery, exploration, application, and eventual ubiquity.

When we consider AI, we find ourselves in a state similar to people from the 19th century who were starting to see electricity for the first time. We understand that there is great power in AI, but we haven't fully discovered how to unleash its potential. And many of us regard it as some kind of magic. Of course, this isn't the case: it's computer science and statistics. Applying it is work. There is no wand to be waved at enterprise inefficiencies, and having the technology alone is not enough. AI requires cultural shifts and organizational change just as much as the arrival of electricity did. If you're in a position of leadership, guiding that organizational change is going to be your most important job.

It's important to understand that AI is here to stay. It is not a fad, any more than electricity or the internet—or, for that matter, the wheel—were fads. The world's two largest economies, the US and China, are investing heavily in this technology and the rate of breakthroughs is in overdrive. AI confers enormous advantages to organizations that employ it effectively. Those organizations that do not adapt to this new reality are ultimately going to wither away. In fact, the lifespan of large (successful) companies is falling fast. The proof? In 1965, the average tenure of an S&P 500 listing was 33 years. By 1990 it had dropped to 20 years, and it hit 15 years in 2010. Estimates for 2027 show tenure on the S&P 500 falling further to just 12 years, and it's predicted that half of the S&P 500 will be replaced over the next 10 years. Like we keep saying: lift, shift, rift, or cliff.

As we've mentioned, it is important to understand that the primary applications of AI will be *augmenting* human capabilities, not *replacing* them. Although AI programs are designed to act upon what they've learned, in most cases they don't act autonomously. Rather, they're tools that can be employed by people in a kind of symbiotic partnership. AI is a tool for improving human decision making at speed and scale, and has the potential to augment the work of every employee. As we've said, we think AI will be a net *creator* and changer of jobs, rather than a destroyer; it will free humans to be more creative and remove the burden of rote, repetitive drudgery while lowering the skill barrier for higher-order thinking and creativity.

By understanding what AI technology actually is (and isn't), and by looking at the history of AI in academic and business worlds, we can better understand why it has not yet been fully embraced and explore ways to hasten its adoption.

Where Are We Now? And Where Are We Going?

AI technology is already being used to solve problems and make discoveries that seem almost magical. But we'll remind you it's vitally important to keep things in perspective. Just as AI can do things beyond human ability, like recognizing the license plate numbers of cars driving by at 50 miles an hour, there are many things that humans can do, even as infants, that AI experts believe will remain beyond AI for a long time to come. Humans will, and must, remain "in the loop." It's important to realize that AI is probabilistic, meaning there is a definite probability that what AI says is likely to happen will not happen (expressed in its confidence level, like our example system telling you the AI's confidence that it's looking at a picture of a cat). This probability is different for every AI.

More profoundly, most everyone would say that AI can beat humans at chess. What isn't as well known is that a team of humans and a computer can usually beat the best AI chess programs. Something special happens when humans and machines collaborate that goes beyond what a machine can do on its own. Intelligence can't be reduced to a single dimension. We can't march into an AI future thinking that our own insight and creativity don't have an important place. Worse yet, AI can be trained with bad or biased data. We don't see an imminent risk of AI-powered robots taking over the world, but we do see a danger in AI making automated decisions based on untrustworthy learnings that are not curated or explainable. "Fairness" is a profoundly human concept that doesn't translate well into computing. If we want systems that are fair, we will need to keep humans in the loop and understand the data that's being used to train them; after all, just because it's AI doesn't mean it's free of bias or even fair.

Machines still have no understanding of psychology, ethics, or cultural norms. While they can determine whether or not a particular solution can work, they can't determine whether it's a *good* decision—to AI, the ends will always justify the means. There are some amusing examples of what happens when game-playing AI systems discover they can "win" by exploiting loopholes in game mechanics. Nor can AI decide what problems to solve, what models to build. There's been a lot of research on AI and creativity focusing on computer-generated music, artworks, and the like (the domain of those GANs we discussed earlier in the chapter), but an AI can't yet decide that it wants to be creative. It does what it's programmed to do, and if that's paint like Picasso, Picasso is what you will get.

Still, it's nice to dream about what kinds of transformations might await us with the help of AI. Consider what AI might do to bring healthcare to the billions of people on this planet who currently have none. Imagine healthcare practitioners in remote areas of the globe—increasingly accessible by telecommunications networks—being able to use apps on handheld devices to access AI running anywhere on earth that can analyze blood samples, interpret vital signs, recognize signs of illness invisible to humans, assist in diagnosis, and recommend courses of treatment. The opportunities for improvement are staggering, and right before our eyes. (We're quite certain AI will be further called upon to better humankind's "hand" after the COVID-19 pandemic experience.)

Or consider how AI might be used to monitor the health of the earth itself. New sensors are being deployed in unthinkable numbers all over the planet, every single day of the year. They're generating petabytes of data by the minute. Who knows what insights AI might be able to find hidden in all that data? We may discover entirely new approaches to deal with the climate crisis and with water and air pollution, and AI may lead us to new opportunities for agriculture and the supply of safe water.

Similar opportunities abound for industry. Supply chains can be optimized to reduce cost and improve performance. Call centers can be radically improved, simultaneously reducing costs, increasing customer satisfaction, and providing more rewarding career opportunities for employees. Buildings and operations can be made more energy-efficient and less polluting. AI plays a dominant role in understanding risk and detecting and preventing fraud, leading to improved financial performance; automation is the only way to detect fraud at the scale of a large financial institution.

Some wins will be much smaller, but no less important. Listerine has developed a smartphone app that uses AI to detect when people are smiling. That might seem trivial, but it's revolutionary for blind people, some of whom have never seen a smile. Imagine knowing for the first time when someone—a baby, a loved one, a friend—is smiling at you. This book focuses on enterprise applications of AI, but it's important to remember that there's also a human side, in which AI will undoubtedly improve people's lives in real ways.

Advances in AI will continue at an ever-increasing pace. These advances will take place in the context of a partnership among three main participants: academia, business, and (increasingly) regulatory groups (whether governments or other bodies). Academia and related entities like IBM Research will continue to explore the frontiers of AI, devising new hardware, software, algorithms, and ways of thinking. Businesses and other organizations, in conjunction with academia and abiding by regulatory frameworks, will use AI to solve real-world problems. Regulatory agencies will look to safeguard privacy, defend against cyberthreats, and in general promote use of AI for the good of humankind while minimizing its use for causing harm.

Moving Forward

All radically new technologies face resistance; the world will always be divided between people who fear change, people who want to adopt change for its own sake, and people who are skeptical. AI is no different from steam engines, automobiles, or, for that matter, the internet. As you start the journey to AI in your organization, remember:

- AI isn't magic; neither were steam engines or automobiles. AI is computer science. It's hard work, and acumen is built over time (much like a flawless Steve Rodriguez golf swing). AI is a family of techniques for building intelligent software that learns from data and acts on what it has learned. It's about building tools to assist the people in your company, helping them to become more effective.
- Focus on the dull, repetitive tasks, such as a bus driver counting the number of people getting on the bus to ensure it's within its safety limits. AI could free up that driver to work on more fulfilling and interesting tasks, such as helping older passengers with baggage or in finding a seat. In aggregate, organizations will have more capability to analyze a business situation so they understand the true problems and opportunities that lie within. Businesses (and people) stagnate when they keep on doing the same thing over and over. AI will help you and the organization you lead or work for grow.

- You'll only grow by starting projects. But don't treat your AI projects like big IT projects, with significant staff and budgets. It's better to start a lot of small projects, each with a limited staff, limited scope, and a relatively short time scale. Some of these projects will succeed, and several small successes are better than watching one giant project fail. And the giant project almost certainly will fail—for a company just starting out on AI, it's bound to be too complex.

AI will revolutionize the business world. The only real question is whether you're going to be a part of that revolution. You're reading this book, so we can only assume you've made your decision to go forth—so let's start by looking at how AI projects have succeeded and how they've failed, with an eye toward making sure your projects are among the successes.

CHAPTER 3

How to Overcome AI Failures and Challenges

In Chapter 2, we talked about why AI is the greatest opportunity of our time—namely, its potential to add almost $16 trillion to the global economy by 2030. We also learned that adoption is slow. But what are the reasons behind that slow adoption? Why would such a tremendous opportunity to expand the economy in unparalleled ways not inspire a rush to production?

As it turns out, implementing AI in organizations is hard, and it's only in the past few years that technology, price, and funding have met in the middle. Some existing AI experiments have failed, and in a very public way—but like with most headlines, there's usually more to the story.

In this chapter, we'll talk about why the time for AI is now, some early examples of success in AI, and some very familiar failures of AI in business and what went wrong. We'll then dive into the challenges to find out why adoption is slow today, even when the issues of data availability and horsepower are solved. And finally, we'll look at some of the tools and services available to organizations that dramatically lessen the impact of those challenges.

Let's dig in.

AI's Emergence in Business Today

As O'Reilly detailed in its 2019 report *AI Adoption in the Enterprise*, the maturity of an organization's AI adoption varies by industry (Figure 3-1). For example, 30% of those who work in finance describe their organization as having a mature AI practice, compared to just 16% from the public sector.

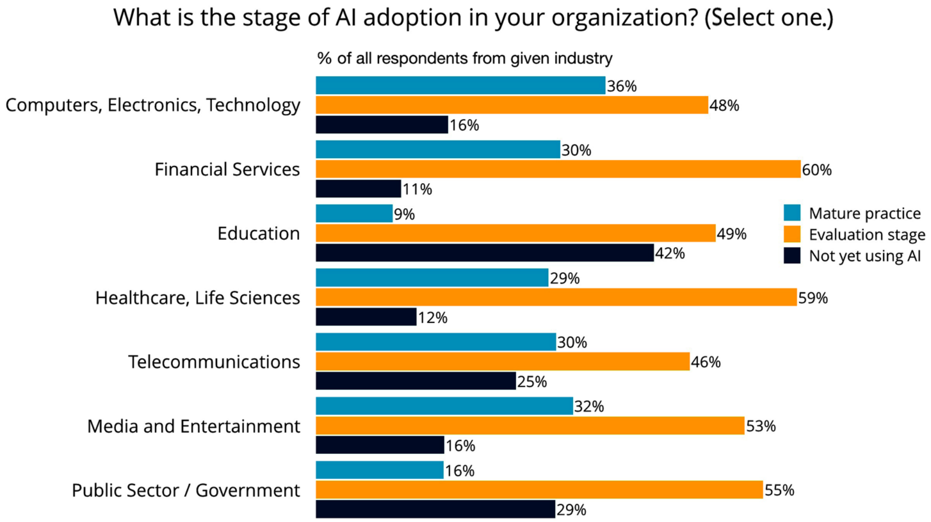


Figure 3-1. Stage of AI adoption maturity by industry (some column percentages don't total 100%, due to rounding)

Three key factors play a role in the emergence of AI in business: the ubiquitous availability of data; the recent decrease in cost and increase in performance of computers, processors, and memory for faster AI computations; and the corresponding reduction in the investment required to start new AI projects.

Let's take a closer look at each of these factors.

Data

Our ability to generate and retain data today is exponentially greater than in the past, and has profound effects on how businesses grow. During the 20th century, scale effects in business were largely driven by breadth and distribution. A company with manufacturing operations around the world had an inherent cost and distribution advantage, leading to more competitive products. In the internet era, economies of scale are still important, but they're different. The 20th century's economies of manufacturing operations and distribution are still important, but we also see important scale effects from networks (for example, the user networks of Facebook and Google) and from data. A small to mid-sized company will have trouble competing with Facebook's scale effects, but data changes the rules. Anyone can assemble large, powerful data sets; the real differentiator is the ability to apply data acumen and analytical expertise to derive value from high-value data. That's a competitive arena that's wide open, where anyone can be a winner. Data acumen is what enables new companies to create new networks, enjoy new economies of scale, and disrupt established players.

Why is data acumen the new differentiator? Here are a few reasons:

- First, we have many more devices, sensors, applications, and services that can generate and send data about their activities. We have smart roadways embedded with small sensors connected to the internet that measure traffic. We have cell phones that send enormous amounts of metadata and location information to the cloud. We have security appliances that generate thousands of log entries every minute. Medical implants like pacemakers can stream data wirelessly back to doctors and electronic medical records systems. Jet engines generate terabytes of data over long intercontinental flights. And even "old-fashioned" web applications generate logs by the gigabyte. The sources of data are seemingly endless, which is why we introduced you to a rather steep data collection curve at the start of this book.
- Second, our ability to receive data from any number of places—over the internet, over a local network, over a virtual private network (VPN), through varying bands of cell services, and more—means data is more available to us. In the past, even when data was generated, it was locked in the place it was recorded and not easily transported. As devices become more connected, and as next-generation wireless and satellite connectivity gains penetration in the marketplace, there remain few (if any real) barriers to moving data to and from diverse locations and destinations. It is nearly impossible to overstate the positive impact that wireless broadband, pervasive fiber optic, and wideband cabled internet access have had on data's renaissance. While we noted 5G as a cell service, it's important not to overlook the impact it will have on future data collection rates. 5G can be up to 100 times faster than 4G networks, not to mention more reliable. What's more, 5G latency rates are about 50% less than 4G, and this is going to shapeshift the livestreaming industry. In other words, it's going to be even easier to move and collect more data than ever before.
- Third, the storage media we use are capable of retaining significantly more data than in the past, at a significantly reduced cost. For example, storing a terabyte of data in 1999 cost several thousand dollars, whereas today solid-state drives with 1 TB capacity can be had for under $100. As a point of nostalgia, we remember working together on a "Terabyte Club" spreadsheet that listed all the clients (ours and competitors') whose data warehouses were over a terabyte. It was quite a prestigious list back then. Today, one of us has a 1 TB Apple Music library, and it could sit in our back pocket.
- Fourth, keeping data used to require decisions and tradeoffs: what data is critical to business and what can be safely and immediately discarded? Now, with huge capacity and low costs, an organization doesn't have to worry about capacity and cost constraints when deciding whether to keep data or not. (There may be other reasons why they choose not to keep data, such as regulatory rules that dictate

what data you can keep or how long you can keep it, or the cost of pre-processing the data, but it's likely not going to be a capacity discussion.)

- And finally, the speed at which we can read and write all of this data has increased several fold over the past decades. Getting historical data out of traditional data warehouses meant waiting for extended periods of time for slow queries to execute. Collecting information in real time was not typically feasible, and batch processes generally stitched up collections of data at regular intervals; the data was then archived in data warehouses, where it was locked away from most folks in an organization. But now data can be streamed in real time, queries run in near real time, and insights derived at the speed that data is coming in (if you have the data acumen), because storage and database technologies have become much more capable and democratized.

While there are many different kinds of data sources available to organizations today, and we're starting to see organizations look to unstructured data to train their AI systems, O'Reilly's *AI Adoption in the Enterprise* survey found that 78% of survey respondents are currently using structured data (logs, time series, geospatial data) to train their AI systems (Figure 3-2). This really shows the potential for AI because 80% of the world's data is unstructured!

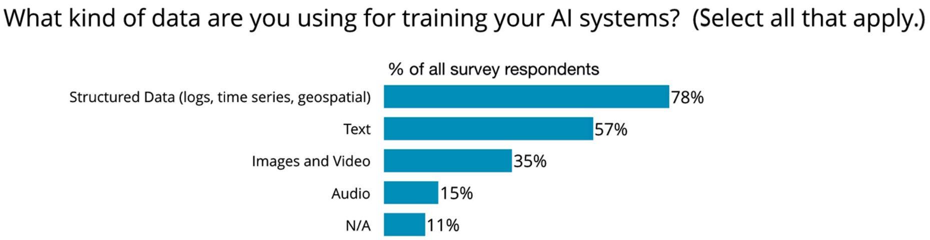


Figure 3-2. Data sources used for training AI systems

It all comes together as a perfect storm: our ability to generate data, keep it, read it, and analyze it have all matured to the point where there is enough data to be kept, enough space to keep it in, and enough speed to be able to read it in volume in a reasonable time.

Computing Power

Even if you have cheap storage and tons of data, your bounty is of little value without the ability to analyze it, tease out inferences and next actions, and run agile experiments based on those results. Just as a pantry full of ingredients does you no good if your dinner party starts in an hour, you can't wave a magic wand and immediately produce value from data. You need the ability to roast the meat and bake the cake quickly in order to unlock the raw ingredients' value and deliver real results.

AI, at a fundamental level, is essentially the computer's ability to run a set of complex math problems quickly and repetitively. (There is more to it than that, but just as all computing at its base level is a series of ones and zeros, AI is complex mathematics. Even natural language processing often represents words in number-based vectors.) The deep learning and machine learning techniques that enable today's AI applications rely on being able to run billions, even trillions, of calculations in a very short time (read: a second or less), understand the result, feed that result into a new calculation, and repeat the process many times over. While it might be irresistibly sexy to think of AI as a computer developing a brain, with neurons and self-awareness, at the end of the day our AI today is just a computer doing really advanced math problems—algorithms—over and over and over again.

You can see, then, that the faster a processor can complete those calculations, the more capable the AI application will be. Only recently, however, have processors with multiple cores—and in particular, GPUs with dedicated circuitry designed for these types of calculations (IBM has developed software to optimize or exploit them)—been developed as mass-market items with prices that mere mortals can afford. Additionally, processors are being built with fast flash memory directly on the chip, reducing the need to return to the system bus to write and retrieve data from system RAM (although the speeds of the system bus and the RAM itself have also dramatically improved over time) and eliminating that round-trip travel time.

Put simply, the overall experience AI will deliver for your business and your users and customers, now and in the future, will largely depend on the computing power available.

Investment

In the past, even organizations with large budgets had a hard time justifying the latest and greatest in computing expense, with the prices of systems and components being so high—particularly when analyzed on the basis of input/output operations per second (IOPS) or floating-point operations per second (FLOPS). Staying on the cutting edge of computing was very expensive, and even in the recent past, the cutting edge was the only place where performance was good enough to really start working with AI. Now that the prices of the compute resources needed for AI have come down, it is much easier to justify investing in AI, not only in a lab but also in production at scale. What's more, you don't have to invest in leading edge technology to get started or to get pretty good performance; don't get us wrong—it will help you build bigger and more powerful models, but as long as the compute is backed by a GPU, you can get started.

Cloud computing is reducing the hardware investment necessary for AI in another way. Instead of incurring massive capital expenditures to procure expensive specialized hardware at scale (even at the commodity prices we see these components being offered at today), with a credit card and a set of credentials you can essentially rent the processing power and storage you need from many public cloud providers or spread costs in an efficient manner across business units participating in a private cloud environment. AI projects that might previously have required the investment of tens or hundreds of thousands of dollars in workstations, graphics cards, and attached storage can now be set up in the cloud in a matter of minutes for a few dollars per hour, depending on your configuration. That computing power is available in seconds; there's no need for shipping, installing, or configuring hardware (to say nothing of the time it takes for a request to get through purchasing). And you only pay for what you use, by the minute.

In addition, rapid innovation in specialized AI hardware chips that enable data to be analyzed closer to its source will accelerate the massive democratization of AI compute power, both in the cloud and at the edge. These chips are starting to deliver performance that previously required massive investment in data centers, while consuming only a fraction of the energy.

It's not just the hardware and software investment that powers AI, however: it's the people behind it. Historically, the skills required to manage models, hyperparameter tuning, feature engineering, and model explainability—and to move AI applications into production—have been scarce. Even if the technology had been capable and affordable, the people to run that tech were few and far between. In fact, data scientists and AI practitioners are an expensive bunch, commanding high salaries in a marketplace where demand has exceeded supply for some time now. That market reality has led AI vendors to look to automate as much of the process as possible, democratizing the practice of AI and making it more accessible to the "citizen data scientist," the everyday business worker with some data familiarity looking to run experiments without becoming beholden to the IT team or professional, PhD-holding data scientists.

Make no mistake, the experts and practitioners still have their place; but today, the investment in experts can be reduced and spent more strategically, in favor of creating a data-driven organization that supports citizen data scientists.

Early Examples of AI Success

Let's now take a look at some early examples of how businesses successfully infused AI into their applications, processes, or workflows.

Example: Vodafone's TOBi Transforms the Customer Experience

By building an AI-based chatbot named TOBi, UK-based Vodafone was able to dramatically reduce costs while improving sales. Today, TOBi is the telecom industry's first chatbot able to complete a transaction for a customer from beginning to end.

After a modest start answering simple FAQs through web chat, TOBi graduated to being able to manage sales of basic cellular plans. From there, it was able to help customers buy new phones—a matter of greater complexity because payment terms on phones can vary. Today, in Phase II of the project, Vodafone is enabling TOBi to access backend customer data to answer questions about customer accounts as well as a multitude of other personalized questions. For example, TOBi can tell callers how much it will cost under their plans to use their phones in different countries—questions that customers would have otherwise had to wait on hold for a live person to respond to. In Phase III, Vodafone plans to roll out kiosks in stores that enable customers to perform transactions such as pay-as-you-go top-ups.

Naturally, this has eliminated a number of customer call center jobs. But instead of laying off workers, Vodafone is running training sessions and up-leveling employee skills so they can become "conversational analysts" who help program TOBi.

Vodafone has released some numbers on TOBi, and they are impressive. For starters, TOBi achieves double the conversion rate of the website. It has dramatically lower abandonment rates, and Vodafone spends 50% less time answering customer queries. TOBi is continually being monitored and improved to evolve the customer experience, with plans in the pipeline to enable it to carry out upgrades and perform tariff migration. Not only has TOBi had a business impact, but it has helped improve brand scores. The net promoter score (NPS) for those who have used it is in the 60s (for those of you not familiar with the NPS system, anything above 20 is considered favorable, and 50 is pretty darn good), and Vodafone reports that TOBi is able to resolve 70% of customer queries without human interaction; difficult questions are transferred to live agents.

Example: How a French Bank Built on Its Strength of Quality Customer Service

Crédit Mutuel, one of France's leading banks, has over 5,000 branches that receive more than 350,000 online inquiries a day, and volume is increasing by 23% a year. Maintaining the quality of client relationships while dealing with an ever-growing stream of customers and client requests meant reinventing the role of client advisor or losing the bank's competitive edge.

After running a diagnosis of how client advisors were spending their time, Crédit Mutuel found that a significant part of their work involved answering simple and repetitive questions. With this in mind, the bank turned to IBM to find a solution that could speed up everyday processes and allow client advisors time to address more complicated and nuanced problems.

Crédit Mutuel has trained Watson to help its client advisors provide customers with quick and comprehensive information on an array of offerings, from car and housing insurance to a range of savings and investment products. Thanks to its Watson-infused email analyzer and four virtual assistants, Crédit Mutuel is able to enrich interactions between client advisors and customers. Watson has made it possible to find the right answers 60% faster, and it can help deflect and address 50% of the 350,000 daily emails received by the bank's client advisors.

Early AI Failures

As we've said all along, AI is difficult. You don't have to look very hard to find examples of why. According to a July 2019 IDC study, a quarter of organizations experience failure rates of as much as 50% in their AI initiatives. Let's touch on a few examples of early AI projects that failed.

In our first example, the reason for failure was data. We were working with a company that wanted to create a customer service chatbot. The company built the model using synthetic data. As you might expect, when they put the model into production, it wasn't an authentic experience. Incorrect data led to unreliable insights and inconsistent user experiences, which resulted in a failed project. The company needed help to surface different, real data sources throughout the organization. And IBM ultimately helped them do just that.

In another failed project, the reasons for failure were cultural misalignment and overly ambitious (or perhaps unclear) experiments. This financial services company was eager to start AI engagements, but it lacked a clear strategy for when and how to use AI. The company's initial attempt was too broad—it needed to start small. (To use a baseball analogy, we always tell clients just to try and get on base with their first AI project. When you "step up" and try to hit a grand slam on your first AI project, you usually strike out quite spectacularly.) Other attempts from this company struggled to get off the ground because the team was not aligned. Eventually, they created an AI center of excellence to bridge developers, data scientists, and key business stakeholders. This center enabled the various teams to agree on which organizational goals were worthy of AI projects.

In yet another project, the reason for failure was a biased model. An online retailer was receiving tens of thousands of resumes a day. It had tried various methods to cull the best of the best, including the industry-standard method of automatic screening by keywords, but it hadn't managed to solve this time-consuming recruiting challenge. The company built an AI algorithm designed to vet resumes automatically, but there was a problem: the system seemed to be disregarding women for engineering or managerial positions. It downgraded resumes that had the word *women* (like "Women's Chess Club") in them, and blocked candidates from all-women's colleges. The cause? The system had been trained on 10 years' worth of resumes and the resulting hiring decisions. Since the vast majority of engineers and managers hired had been male (75%), the system simply "learned," erroneously and unintentionally, that the retailer thought men were better qualified for those positions than women. At first the company didn't understand how this could be happening, as there wasn't even a field for gender on its job application form. It turned out the algorithms were using other information as proxies for gender: names, university attendance, membership in women's organizations, etc. When the company realized this, it immediately stopped using the system and modified the algorithms to avoid (and continually monitor for) gender bias.

This is what we were talking about earlier when we said we were more scared of AI making automated decisions based on untrustworthy or biased data than we were of AI taking over the world. You always have to be aware of the nuances of the data you are using to train your AI. Lots of face recognition algorithms have been trained on Caucasians (specifically, white movie stars, since their pictures are so readily available); the unintentional consequence is that these algorithms perform poorly (remember, performance in AI means accuracy) on candidates that aren't Caucasians. Why? Because there aren't enough non-Caucasian faces in the training set, the resulting model is inevitably biased. Now stop and consider the impact such a bias could have on the computer vision system in an autonomous vehicle.

AI Challenges: Data, Talent, Trust

Ultimately, AI adoption—especially given the size of the opportunity and the overall stakes—can only be characterized as *slow*. We understand the power of AI, but we haven't fully discovered how to unleash its potential. The reality is, as we've said several times, AI is not magic. Applying it is hard work. There is no wand to be waved at enterprise inefficiencies, and having the technology alone is not enough.

There are three major inhibitors to AI adoption at scale. They are:

- Data
- Talent
- Trust

Let's explore each of these challenges.

AI Challenge: Data

As we will repeat throughout this book, data isn't just a key ingredient for any successful AI project—it's *the* key ingredient. One problem organizations face when first trying to use AI is knowing their data. Data is the foundation and fuel for AI. Good data, and lots of it, is needed to train models, and then a steady stream of new good data is needed for the resulting AI-infused business processes to do their work. There are five general classes of data problems:

Siloed data

We always tell customers that organizational silos result in data silos. You may have heard the term "data silos" used to describe different parts of an organization that have data stored in their own particular fashion and not easily shared with other parts of the organization (be it for technical reasons or political). The Sales division has its data, Customer Service has its data, Finance has its data, and so on. These separate databases are sometimes guarded quite fervently as an asset of each division.

AI doesn't work well in a data-siloed environment. When all relevant data is made available, even from parts of the company that may have never worked directly together, AI can often discover unexpected patterns, problems, and opportunities.

Lack of data

AI's appetite for data is voracious, and companies sometimes simply do not have enough data readily available to make an AI project feasible. (Truth be told, more often than not, companies have the data; they just don't realize it because it's stuck in silos, or dropped on the floor as soon as it's generated.) To address the problem of insufficient data, companies need to start by collecting and organizing their data, acquiring additional data from third parties, and making all of this data easily accessible (in a governed manner) to employees throughout the organization. Many products and services are available to begin the process of building a data platform that can manage the intake of existing data and integration of new data, as well as helping to publish it within the organization in a controlled way so that it can be used by AI projects and for other data-driven tasks.

Too much disorganized data

Some companies have the opposite problem: there is so much data that it's disorganized and spread too widely. This phenomenon is generally a data engineering problem, and while these types of organizations have done a good job at collecting data, they need to organize it to make it ready for AI—and this is why so many data lakes got unintentionally turned into data swamps. Again, there are data platform products and services that can help with managing the entire data lifecycle, including cataloging it and organizing so that it can be published from a single place. Adopting one of these platforms is sometimes a necessary but frustrating obstacle for companies wanting to dive deeply into AI.

Bad data

"Garbage in, garbage out" is as true in the current days of AI as it was in the early days of data warehousing. Data may be old, conflicting (as when different silos contain different addresses for the same customer, to cite a trivial example), incomplete, and so forth. The problem, according to Forrester, is that even though business leaders list "improving the use of data" as a top priority, 60% are challenged by managing data quality.

Much of the most important work in AI involves preparing and cleaning data—doing the laborious work (which we'll describe in later chapters) of turning bad data into good data.

Problems with data quality or volume

Once a prototype model is in place and working, the later stages of a project involve rolling it out to a wider audience. At this point many projects run into issues with either the availability of data, its volume, or its quality. Quality problems have obvious consequences, but volume and availability are both concerns as well: for example, to be accurate across all classes, all classes have to be represented equally in the data used to train your models, even if that isn't necessarily the case in the general population. Think back to our face recognition example—the model couldn't perform well on non-Caucasians because they were not part of the training set, but they are found in the general population the model was designed to work on. As we will explore later in this chapter and throughout this book, the quality of output from AI projects is only as good as the quality of the data that goes in, so ensuring that the data you are using is sufficiently labeled, validated, non-biased, and sourced is critical. The wrong data input into an experimental model can have disastrous results and can kill many AI experiments; after all, if you give a human bad data and they can't solve the problem with it, AI won't be able to solve the problem either. Why? Because it's not magic!

Putting proper data governance in place is a key step toward preventing and solving data quality problems. Governance means controlling access to the data, tracking changes to the data, understanding the source of the data, understanding regulations that control how the data can be used, and more.

AI Challenge: Talent

It's relatively easy to imagine that in a cutting-edge, fast-paced, constantly changing area like AI, potential employees with the required knowledge, skill sets, and experience are rare—and with AI popularity on the rise, these folks are in high demand. But this is just one of the talent-related factors at play; company culture and organizational silos can also pose problems when it comes to AI adoption.

High demand, low supply for potential employees

The largest hurdle for getting AI off the ground in any company is getting a team of experts in place. Unfortunately, lots of companies are competing for a small set of employees that already possess the right skills and experience.

Just a few metrics will shine a light on this issue:

- IT Chronicles reports that “a recent poll confirmed that 56% of senior AI professionals believed that a lack of additional, qualified AI workers was the single biggest hurdle to be overcome in terms of achieving the necessary level of AI implementation across business operations. As an example, Element AI last year estimated that there were fewer than 10,000 people around the world who have the necessary skills to create fully functional machine learning systems. In short, though huge demand already exists for AI skills, the shortage of talent is slowing down hiring—and without new AI hires, organizations simply cannot press forward with their AI strategies.”
- CIO.com reports, “Over the past three years alone the number of AI-related job postings on Indeed [a worldwide employment-related search engine for job listings] has increased by 119 percent, according to the platform’s latest AI talent report. The machine learning engineer role was cited as the third most in-demand AI job of the moment with machine learning ranked as the most in-demand AI skill. … AI [is] projected to create 2.3 million jobs by 2020, according to a Gartner report.”

O’Reilly’s *AI Adoption in the Enterprise* survey confirms this strong demand for data scientists and other AI-related professionals (Figure 3-3): more than half of all respondents signaled their organizations were in need of machine learning experts and data scientists.

Without the right people on your AI and data team, adoption of AI in your business is bound to be slower and more difficult than you may like.

Using AI to glean insights from the enormous amounts of data available in an organization is a departure from the previous model of old-school business analysts poring over pivot tables and spreadsheets to make decisions. AI provides you with new automated workers that can complete tasks reliably—but that doesn’t mean the business analysts become obsolete. Instead, they take on the role of citizen data scientists and become the managers of these new AI “workers,” evaluating their work and intervening when things go wrong. That may mean overriding the AI’s decisions at times, but more importantly it involves going back and helping to improve the data used to train the models, so that they can do the task successfully the next time.

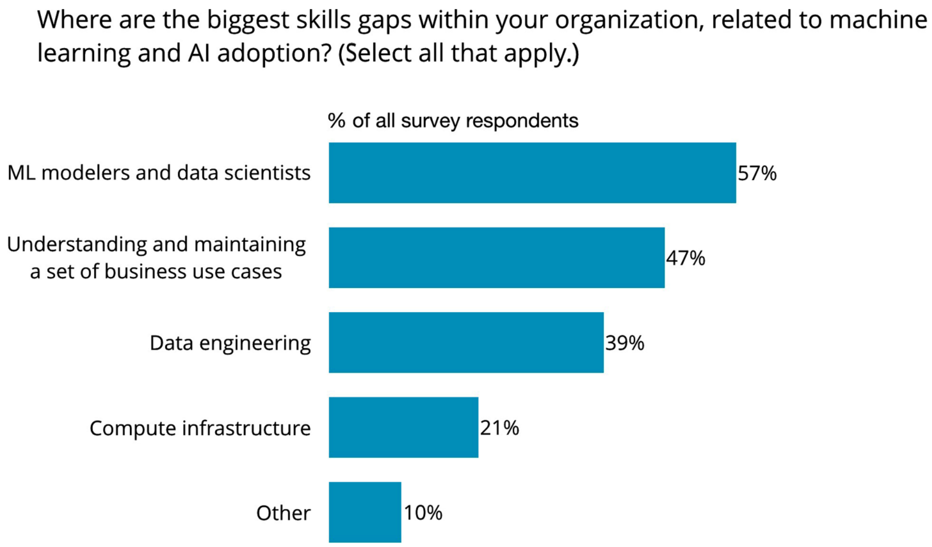


Figure 3-3. Skills gap related to machine learning and AI adoption

During this cultural and business model change, it is important to pair the AI with domain knowledge workers or subject matter experts: drawing on the skills of lawyers, customer care representatives, and doctors, for example, to help prepare the data to train your AI models. For instance, you might use insurance adjusters to label pictures of car damage, showing what was damaged and the estimated price of the repair. When the model is in production, AI will facilitate a cultural change for these individuals by automating routine tasks.

All industries will be affected by AI. Even lawyers are entering a transformed world: the world of legal cognition. For example, LegalMation has developed an AI platform that automates routine litigation tasks. It helps legal teams draft high-quality litigation work in minutes, saving time, enabling a shift in strategic focus, and driving down costs by as much as 80%.

Fennemore Craig, an Arizona-based corporate law firm, is all aboard the AI train. The firm used an AI system built by ROSS Intelligence and IBM Watson to comb through millions of pages of case law, with AI writing up its findings in a draft memo for current work. AI is quickening a process that would normally require almost a full week of work, completing it in just one day. How? AI doesn't get tired, and it doesn't need coffee or lunch breaks. Fennemore Craig believes its AI system is producing materials equal to those from a first-year law student. A human lawyer then adds deeper analysis and punches up the language, making it more like a lawyer's written document. How effective is Fennemore Craig's AI system? US News ranked them in its "Best Law Firms" rankings.

Let's talk more about the cultural challenges of adopting AI.

Culture inhibitors

To facilitate AI adoption, the people within your organization don't only need to possess the relevant skill set; they need to buy into the promise of AI as well. These employees have to fit into a culture that both understands the potential that AI offers and embraces its presence in as many facets of the business workflow as possible. According to O'Reilly's *AI Adoption in the Enterprise* report, survey respondents often cited "company culture" and "difficulties identifying use cases" as serious challenges (Figure 3-4).

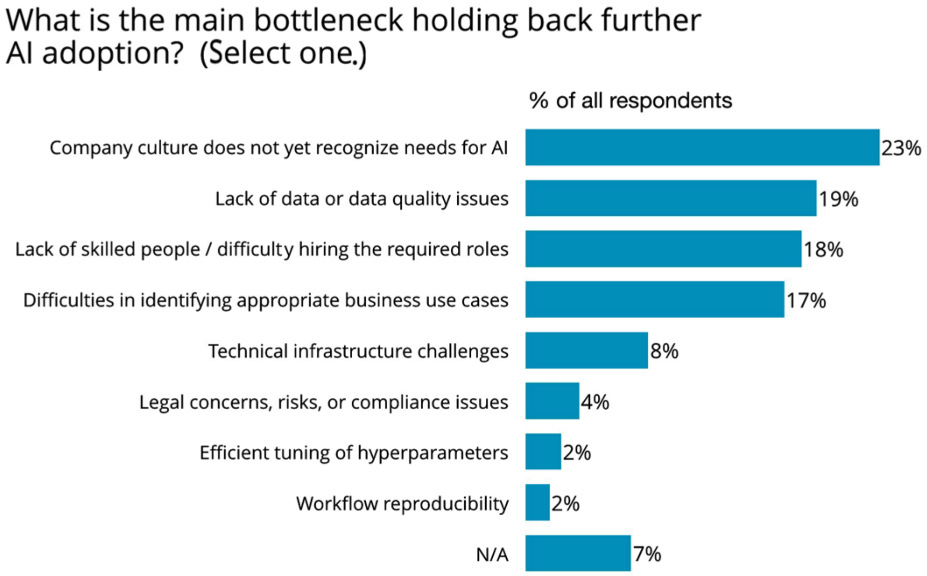


Figure 3-4. Challenges for AI adoption in the enterprise

Just as many existing companies failed to embrace digital transformation (think taxis or food delivery as prime examples) and consequently were late to join the mobile revolution, many organizations today are unwilling to do the deep rethinking of their business models and workflows that will allow them to fully embrace the opportunities of AI. You'll miss out on the full gambit of potential benefits if you try to put AI on top of existing business models and workflows; you should expect the models and workflows to change because of AI, and need to be willing to redefine your business from the ground up. We can't stress enough how important this is. As you build your AI, reimagine your business processes from the ground up with your new superpowers; we guarantee you the impact will be greater.

The organizations that will enjoy the most success from their AI projects will become AI-centric organizations. Put simply, at some companies, this transition is easier to manage and complete than at others. This is because becoming AI-centric involves more than just successfully executing a single business project. It's about transforming an entire business culture, creating a culture of iteration, experimentation, adaptability, and imagination. And it can take time, energy, and effort.

If you're an experienced senior manager, you know that managing organizational change is never easy. Although AI is going to open up new vistas for many employees, others are certainly going to feel threatened by it. Territoriality and resistance can arise. That resistance, if not properly anticipated and managed, can sink even the most promising AI undertaking, and exceptional leaders too.

In fact, while helping thousands of organizations incorporate data acumen into their business practices, we've found that the most common cause of failure in AI projects is organizational unreadiness. Trust what we're about to tell you: *people make the difference in any successful AI project, so it's critical that they work together*. Let's cover some examples of what we mean by "organizational unreadiness."

Siloed people and departments

Because AI projects require data from across the company to perform effectively, often leaders and staff from a variety of departments need to come together in order for an experiment to work. This requires establishing and maintaining relationships with key people across the company. Remember, organizational silos result in data silos!

People in different departments have different skill sets and expertise that may be relevant to an AI project. The IT team may have significant experience estimating licensing costs for a data platform and the overall expense of and timeframe for implementing it. That knowledge needs to be included in any project plan. What's more, IT has experience in enterprise-hardening solutions for availability, scale, and cyber resiliency (a very big deal, and AI is definitely not safe from its reach). Your business team may know how certain data is stored and how it's retrieved. They may also be aware of key caveats that could affect the scope or quality of data. That type of knowledge and institutional memory is important.

Planning for turnover within teams is also crucial, especially for long-range AI projects that span fiscal years. If a supportive leader transitions into a different role and their replacement has a dim view of AI experiments, you will need a plan to win them over, and selling a project's current status and any successes along the way will only help. Developing relationships not only with supporters but also skeptics can help when the time comes to prove the value of an in-progress AI project.

If you aren't able to partner successfully across your business with the right people in the right divisions, it can cause an otherwise successful project to stall out (or at minimum, go slower)—and often this can happen right in the middle of an implementation, which is the most frustrating place of all to have a problem.

AI Challenge: Trust

Next, there's the issue of trust. It is critical to ensure that AI recommendations or decisions are fully traceable (have provenance). Enterprises must be able to audit the lineage of the models and the associated training data, along with the inputs and outputs for each AI recommendation. As AI advances, and humans and AI systems increasingly work together, it is essential that we trust the output of these systems to inform our decisions. One way to instill trust is to be able to explain one's actions. We depend on explainability as a basis for interpersonal relations; AI will be no different.

Some implications of both trust and explainability:

- If models have inherent bias, the decisions and recommendations they issue will naturally be marked by that bias. Depending on your industry, bias may be more than just bad practice; it may run afoul of laws and regulations.
- If models are "black boxes," it becomes very difficult to ascertain why certain decisions or recommendations were issued, and equally as troublesome to pinpoint the exact data and model in the workflow that's responsible (most AI projects require multiple models). The quandary is that typically the more accurate a model becomes, the more opaque it becomes as well because of ensembling and other techniques. You will almost certainly need this information to prevent problems from recurring. For example, you may need more or better data, you may need to adjust the neural network's hyperparameters, and so on. In a black box system it can be challenging to determine what is causing the problem.

Alongside policy considerations and business efforts, science has a central role to play in developing and applying tools to wire AI systems for trust. It is imperative for people to know why AI is making the decisions it is making. It is equally imperative for people to have confidence in their AI. As IBM Research has written about extensively, the only way people are going to trust their AI is to know it well enough to explain it, to be able to trace it and audit it, and to be able to see the inputs and outputs for each recommendation.

Fairness

AI today is based on our computers' ability to learn, detect patterns, and inform decisions. Yet, whether we want to acknowledge it or not, we live in a world that contains biases and prejudices, and these biases can easily enter AI systems through the data that we create, collect, train, or process. The systems pick them up, encode them, and have the potential to scale them. Even as we build models, we as humans, even with the best of intentions, are susceptible to unintentionally injecting latent biases—and with AI, it's dangerously easy to do so. Even the smallest miscue in this process can have serious consequences down the road when data at scale is fed to a poisoned model.

For example, what happens when a man and a woman with similar financial histories apply for a credit card? It's been common for a man to get a higher credit limit, because the AI system that computes the credit limits relies on historical data showing that women typically are paid lower wages. It doesn't really matter whether or not the system uses gender as an input—it probably doesn't. AI models are very good at inferring gender from other characteristics (for example, purchase histories). It doesn't "know" that one person is female and the other male (unless you give it data to specifically tell it so); it does know that one is more likely to buy hair products and the other more likely to buy tickets to sporting events. Systemic bias and unfairness can be very subtle.

This is why instrumenting AI for fairness is essential. Fairness is aspirational; data is historical, and history is neither kind nor fair. We believe that AI can be engineered to be fair, but that's a choice we have to make. Properly calibrated and programmed, AI can assist humans in making more informed choices, processing and evaluating facts faster and better, or allocating resources more fairly, giving us an opportunity to break the chain of human biases.

Explainability

In many applications, trust in an AI system will come from its ability to "explain itself." But when it comes to understanding and explaining the inner workings of an algorithm, one size does not fit all. As more applications make use of AI, businesses need visibility into the recommendations issued by those applications. There are certain sectors—finance, healthcare, government—in which the need for adherence to strict regulations may present barriers to AI adoption, and where the system needs explainability on multiple levels. Different stakeholders require explanations for different purposes and objectives, and the explanations provided must be tailored to their needs. While a regulator will aim to understand the system as a whole and probe into its logic, consumers affected by a specific decision will be interested only in factors impacting their case—for example, in a loan processing application, they will expect an explanation for why the request was denied and want to understand what changes could lead to approval.

Take for example the General Data Protection Regulation (GDPR), which makes vast and sweeping changes to the laws surrounding data collection and use. Under GDPR Article 14, AI applications must explain their outcomes (for example, when customers are applying for credit) in order to be used in production situations. In addition, Article 14 requires an appeal process where humans review cases in which AI rejected a credit application. That's a full stop if your AI solution doesn't allow for human review. While you may downplay the relevance of the GDPR if your business is not based in the EU, the effects of the regulation end up applying to most large corporations that do business internationally. And this regulation is just one example. California has also passed a privacy law (the California Consumer Privacy Act, or CCPA), and other states are expected to follow with similar legislation. Your company culture directly affects how agile you and your ecosystem can be, because new laws could be enacted at any time. You need the ability to address these new laws and policies.

We predict a future where AI algorithms come with their own "nutrition facts" labels that explain the algorithm, how the data was collected, what data was used to train the model, when the model was last updated, and more. (Margaret Mitchell et al.'s paper "Model Cards for Model Reporting" goes a long way toward defining what these labels might look like.)

Robustness

While AI-powered systems can augment human decision making and improve outcomes, they are not infallible and may be vulnerable to adversarial attacks. This raises security concerns, potentially compromising people's confidence in the systems. While the technical community exposes and fixes vulnerabilities in software systems on an ongoing basis, attacks on AI-powered systems pose new challenges. For example, attackers may poison training data by injecting carefully designed samples to compromise system security. They can steal an AI model by studying its outputs or fool the algorithm by introducing noise or adversarial perturbations. Hackers can even perform "black box" attacks on deployed AI models, where they try to infer the neural network's weights and parameter settings by querying the model. There's the famous case of a group hacking a popular neural network to convince it that a cat was actually a toaster—this is a great example of a black box adversarial attack.

Transparency and accountability

Safety, fairness, explainability, and robustness are characteristics we need to demand from an AI system. But to achieve trust in AI, engineering these traits into a final solution will not be enough; we must also be able to measure and communicate the performance levels of a system on each of these dimensions. IBM Research has introduced the concept of factsheets for AI services, which outlines how these metrics can be communicated to end users as a way of informing their understanding of how the service works, evaluating its functionality, and comprehending its strengths and limitations. The performance metrics and other events of an AI system could be captured on a blockchain (distributed ledger) fabric for immediate distribution.

When companies put AI models into production, it is essential to break open the "black box" of AI and develop a strategy for continuous monitoring of the models' output after deployment. Without this monitoring step, organizations have no visibility into what their AI is doing, how often it is being used, what the outcomes are, and what biases might have been revealed in the data used to train the model. Author Cathy O'Neil refers to the ways that models can encode bias in the title of her book *Weapons of Math Destruction*. Unless you have built mechanisms that can address those concerns, your organization will not gain the trust of your customers, employees, or the public.

In later chapters we'll explain in detail how to build in trust by opening the black box of AI.

Value alignment

As AI systems advance and we deploy them in complex environments and decision-making scenarios, we need to provide them with the ability to reason through different outcomes, discriminate between "good" and "bad" decisions, and ensure the outcomes we truly want. Sometimes what we want extends beyond organizational goals. For example, we want to see AI being used in a socially acceptable way (consider the current discussion of face recognition in public places). The quality of a decision should be evaluated in terms of accuracy, ability to satisfy users' preferences and optimization criteria, as well as other properties related to the impact of the decision, such as whether it is ethical or whether it complies with feasibility constraints or safety regulations.

Overcoming Challenges with Advanced Research and Products

Sobering as those challenges may be, visionary leaders can certainly see the value of AI—and they're already starting to commit dollars to the cause. The current problem is, those dollars may not be effectively spent. According to IDC's Worldwide Artificial Intelligence Systems Spending Guide, AI software spending will grow at an annual rate of 28.4% to reach $97.9 billion by 2023. That seems like a lot of money, and it is, especially since *only 20% of analytic insights will deliver business outcomes*. Data science and machine learning surveys often indicate that less than 50% of data science and machine learning analytic assets are deployed in production.

How can you raise the number of applicable AI-generated business insights? And how can you significantly increase the number of data science and machine learning assets deployed into production? What types of research and products are there that can help make this type of spending more effective? We'll cover some of these products in later chapters.

Overcoming Challenges with the Right Partner

Companies need to be able to connect their data to their AI in order to deliver good business outcomes. In the past few years, AI has seen some highly publicized successes—and some highly publicized failures. You know which list you want to be on. How do you make sure you are among the successes?

- First, look at your data. Siloed data, insufficient data, and unreliable data are all common causes of failure.
- Second, look at your talent. We all know that AI experts are hard to find and expensive to hire, but as we'll see, there are ways to work around these issues. Siloed people and departments are a much bigger problem. For AI to succeed, everyone needs to work effectively as a team, and they need to keep their skills up and possibly develop new ones. The best team members are the ones who are always learning.
- Finally, work on building trust. AI needs to be fair, explainable, robust, and transparent. It should reflect your corporate values. These things don't just happen. They require work; they need to be recognized as part of the problem you're solving from the outset.

If you're just starting on the journey to AI, that's a lot to keep track of. Choosing a technology partner with the right methodology, experience, products, and services to help modernize the enterprise is a crucial step toward bringing the power of data and AI into your organization.

The AI Ladder, which we'll dive into in the next few chapters, can help orient your journey and place the right stakes in the ground for a solid, trusted vertical climb.

CHAPTER 4

The AI Ladder: A Path to Organizational Transformation

Despite AI's enormous potential, it is still underutilized by many industries and organizations. There are several reasons for this, some having to do with data (in one way or another), some with the lack of people with the right skill sets, and some with issues of trust or misunderstanding. Most of them are connected in some way to the broader problem that many organizations are simply not ready—at a process level, a personnel level, or an infrastructure level—or don't know how to make the changes necessary for a successful transition to becoming an AI-focused entity. One thing we've seen: if you start your AI journey without being ready, you're bound to fail. But if you don't start because you're not ready and are unwilling to take the steps to become ready, you're also bound to fail. So the big question is: how do you become ready? That's the question managers need to answer.

We all need to remember we're in a builder's market. Organizations have to try things out. Pick a problem you've always wanted to solve—whether it's making better predictions, automating something, or optimizing a process—and use AI to solve it. That first project will get you some experience and force you to start collecting (or preparing) data seriously and building the data infrastructure you will need for larger projects. Perhaps more importantly, you will learn a lot about your team, its culture, and yourself, and ultimately your organization's readiness for AI. So much the better if solving the problem moves the needle on some corporate metric: at this point, you don't need to demonstrate that AI will revolutionize your business, but you should be able to demonstrate that it can help and is worth further investment. You don't even need to use advanced AI techniques. Define the problem as precisely as you can, and don't make it more complex than it has to be. AI rewards people who try to solve specific, well-bounded problems ("supreme clarity" is a favorite phrase we use daily) that are aligned to business unit objectives. That business unit objective might be "Delight customers," with a key result of "Net promoter score of 50." This needs to be further broken down into its component parts: for example, "Make pricing simpler," "Predict lifetime value," and "Automate first-line customer service."

The vast majority of AI failures are due to failures in data preparation and organizational science, not the AI models themselves. The failures in data preparation and organization, in turn, can ultimately be traced to the root cause that we have called "organizational unreadiness." To experience significant success with AI, virtually every aspect of your company's operations is going to have to change in one way or another. Some companies are not willing to commit to this scale of change, so they attempt to fence their AI projects to limit their impact. Although we recommend starting with small, simple AI projects (the first-base hits we talked about in Chapter 3), you must understand that you will need to go further to score runs. Limiting your use of AI because you're not willing to change is another way to guarantee failure.

Suitability of AI

Organizations need to understand what AI is suitable (and not suitable) for. By now, everyone in a position of senior leadership has heard about AI, knows that it's a crucial technology, and is anxious to get started with it. They may have some apprehension, but they also don't want to miss the boat. Fear of missing out (FOMO) drives many organizations to jump to implement an "AI solution." Not fully understanding the technology, they assume that it will fix any business problem. This is the "magical thinking" approach that we discussed in the previous chapter. It doesn't work and it's bound to fail.

A key component of suitability is measurability. Every AI project must have clear metrics by which to judge whether or not the project was a success. Without clear and agreed-upon measurable goals, projects can devolve into acrimony as different parties argue about vague notions of success or failure. Across our many years in the industry, we continue to marvel at how teams in the same company can have completely different interpretations of the same goal; it's like one team's American football touchdown is another's field goal, and yet another team's first down. *Get a quorum on what success looks like from the very beginning.* Trust us on this one—we learned the hard way and have the scar tissue to prove it.

Remember, at its core, AI represents a powerful new set of software and data engineering techniques for making sense out of vast amounts of data of varying complexity. It isn't a magic wand that can do anything; it must be applied to problems that it is well suited to solve. You therefore must have a general understanding of what kinds of problems AI is good at, and what kinds of problems are best handled in other ways. In other words, as organizations embark on their AI journey, they need to identify the business problems they are trying to solve, ask the right questions, and identify whether AI is a suitable approach to achieve their business goals.

Determining the Right Business Problems to Solve with AI

Not every problem can be solved with AI—just the right ones. Choosing the right project is a key element of success. As we saw in Chapter 3, many AI projects fail because they attempt too much before the organization is ready. Overly ambitious large-scale projects, or projects that entail high risk to the organization if they do not work out, are not good places to start.

To determine the right business problems to solve with AI, consider these best practices:

- Specify the business unit objective to which you would like to apply AI.
- Determine which key result of that objective will make a noticeable contribution when it succeeds.
- Use design thinking to identify and prioritize potential applications of AI.
- Understand that failure is not failure (so long as it's done safely), but a learning experience, when it is identified early and acted on immediately.
- Look for parts of the business that have leaders who will require implementation from their team. AI has a negative ROI if it's never implemented.
- Choose metrics that are tied to a quantifiable cost savings, increased revenue, or net new revenue.
- Understand that most problems worth solving will likely require multiple AI models.
- Use agile methodologies and break each component part into two- to three-week sprints. The goal is to deliver something tangible after two sprints, then build off of that.
- Start with data you already have (and can actually be used to solve the problem) that is either not being used or being used to support a mundane workflow by humans (like a quick visual inspection).

Ask your team to propose a project that can show preliminary results in about six weeks. Based on our own experience and the experience of IBM consultants who have done thousands of such projects, we've concluded this is about the right length of time for an initial sanity check. Anything much shorter and you won't have time to get the work organized and going; anything much longer and you run the risk of going off into the weeds.

Building a Data Team

Now let's consider your team. Depending on the size of your organization, you probably don't need an army of data scientists. You need a leader who understands the technology, who has experience, and who has a network they can draw upon when the team needs to grow. Again, having well-defined projects of reasonable scale, with well-defined and reasonable criteria for judging success or failure, will help you attract the kind of employees you're going to need.

You also need—sooner rather than later—specialists in data operations (DataOps) to build and maintain your data infrastructure. They're responsible for selecting, building, and maintaining tools for data organization (we cover this in Chapter 7), as well as interfacing with the teams responsible for analyzing, deploying, and maintaining applications (detailed in Chapter 8).

It's worth pointing out that the AI lifecycle often exhibits resource requirements that are the opposite of traditional software development. It is not unusual for a relatively small team of software engineers to develop an application that requires an enormous amount of server "horsepower" to run when it's deployed to the world. In the AI lifecycle, the training process usually requires more computing resources than running the model in production.

Consider, for example, a massive online retailer. They might have a thousand software developers and quality engineers, but their traditional application might be used by a hundred million or more customers at the same time. The size of the production environment dwarfs that of the development environment. In AI, however, much of the computation is done up front, in the preprocessing of data and the training of models and algorithms that iterate over that data. Whether that computation is done on hardware owned by the corporation, on rented virtual servers in the cloud, or in some kind of hybrid architecture, it is not going to look like a traditional development architecture. With AI, the development environment often dwarfs the production environment because building and training the model requires enormous CPU- and GPU-intensive number crunching and iteration.

Putting the Budget in Place

Management must understand that budgets for AI projects won't look like those for traditional projects and must fund them accordingly. For one, managers should generally treat AI projects as operating expenses, because for almost all organizations, the best place to run your AI infrastructure is in the cloud (hybrid cloud is perfectly acceptable). This enables you to harness massive amounts of compute power and storage for just a small per-hour charge (or departmental charge-back in the case of a private cloud), so you can spin up tremendous capacity when necessary and shut it all down (and thus avoid paying for it) when you don't need it. "Cloud is a capability, not a destination" is our mantra; if you're only thinking of it as another place to store your data or do your computation, you're not going to get the advantages (we'll say more about this soon). Managers used to looking for approval for giant capital expenses that got written down over 3, 5, 7, or even 10 years now need to shift into understanding how monthly bills from cloud deployments will hit a profit and loss statement and how those justifications might need to be made to the powers that be in an organization.

Secondly, the people cost for an AI project may be low, at least compared with the headcounts used for, say, an ERP system upgrade or a heavy lift to a new version of a very specific line of business software. AI and machine learning teams need to be smaller and nimbler. They don't need to build out architectural diagrams and recommend integration plans, and they generally use open source software to experiment.

Developing an Approach

Data scientists have the most fun in the build phase, because it lets them exercise their freedom and explore a set of data to understand patterns, select and engineer features, build and train their models, and optimize hyperparameters. This is where a host of tools and frameworks come together:

Open languages
: Python, R, Scala, etc.

Open frameworks and models
: TensorFlow/Keras, PyTorch, XGBoost, Scikit-Learn, etc.

Approaches and techniques
: Generative adversarial networks (GANs), reinforcement learning (RL), supervised learning (regression, classification), unsupervised learning (clustering), etc.

Productivity-enhancing capabilities
: Automated AI, visual modeling, feature engineering, principal component analysis (PCA), algorithm selection, activation function selection, and hyperparameter optimization, etc.

Model development tools
: DataRobot, H2O, IBM Cloud Pak for Data with Watson Studio, Azure ML Studio, Sagemaker, Anaconda, etc.

Deployment tools
: Cortex, TFX, etc.

Our pro tip: your data science team is best positioned to choose the right tools—the ones they are familiar with, the ones they work quickly in and around, and the ones that will integrate best into your organization's existing software development infrastructure.

There Is No AI Without IA

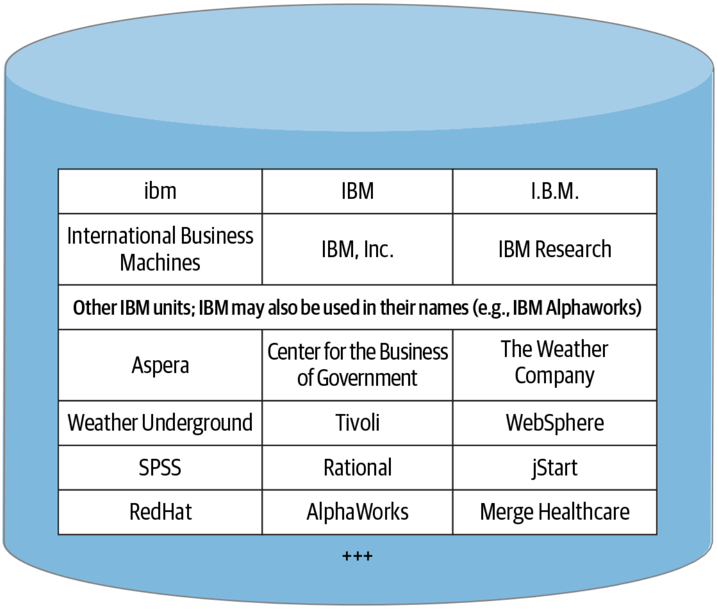
But how do you start even a simple AI project? Fortunately, you don't have to do everything at once. Becoming an AI-centric organization may be a journey, but you can do it in stages. Those stages make up the AI Ladder.

One way to think about starting this process is to remember our tagline, "You can't have AI without IA." AI relies on an information architecture (the IA part), a concept we'll explore in the following chapters. An information architecture is the foundation on which data is organized and structured across a company. For now, think of it as a universal methodology, tailored to your organization's unique circumstances, that guarantees a steady supply of data that is:

- Accessible
- Accurate
- Secure
- Traceable and verifiable

Accessible means that you can get to the data: it isn't tied to a particular vendor or proprietary tools, nor is it locked up in some silo within the organization. It's always been true that combining data sources gives results that are much more powerful than results from the data sources taken individually; the whole is greater than the sum of the parts. That's why we stated earlier that organizational silos result in data silos.

Accurate means that your data is, well, accurate. It's a mistake to assume that your incoming data is correct, usable, or free of bias. If you look at your data, even casually, you will see plenty of problems: missing values, values that are obviously incorrect, misspelled words in textual data, and plain old ambiguity. Someone once told us how many different ways they've seen the corporate name IBM appear in a list from a vendor database: it was well over 100 (Figure 4-1). Spelled out, with capitals, with periods, names of companies that IBM has acquired, "International Business Machines" in various languages—they all map to IBM. It's been said many times that 80% of a data science project is getting, preparing, and cleaning the data.



ibm	IBM	I.B.M.
International Business Machines	IBM, Inc.	IBM Research
Other IBM units; IBM may also be used in their names (e.g., IBM Alphaworks)		
Aspera	Center for the Business of Government	The Weather Company
Weather Underground	Tivoli	WebSphere
SPSS	Rational	jStart
RedHat	AlphaWorks	Merge Healthcare

+++

Figure 4-1. Various spellings of IBM in a database

Secure means that data is protected from inappropriate access. Many IT executives have left their companies in disgrace after a data theft was discovered. All too often, these thefts could have been prevented by taking minimal security precautions. This has become a huge storyline recently, with the US government declaring cyberattacks as the number one threat to the United States, above terrorism and nuclear attacks (though we suspect in the near future, pandemics will make an appearance for obvious reasons). Our advice: start with the principle of least privilege and apply a defense in depth strategy.

Traceable and *verifiable* means that you know where your data came from and have confidence that it's accurate. This is called "data lineage." Again, it's something that developers didn't worry about a few years ago (and many still don't today because it's not in their culture), but it's increasingly important to be able to trace your data to its source, and to know that the source is reliable. Knowing your data's source is particularly important for AI projects. If your application is trained on bad data, it will give you bad results. If it's trained with biased data, you'll get biased results. And if someone can corrupt your data before you use it, they—not you—control what your AI application can do.

Organizations need an information architecture that is modern and open by design: flexible, not tied to any one vendor, and capable of working in public clouds, private clouds, and on your own on-premises investment.

Once you've internalized the importance of the phrase "You can't have AI without IA" you'll understand that in order to become ready for AI, your organization is going to have to do a wholesale reinvention of itself. If that sounds to you like a pretty tall order, you're right. It is. That's where the AI Ladder comes in.

The AI Ladder

We often hear how clients struggle with their skills, and struggle with how to get a quick win with AI. According to Ritika Gunnar, VP of IBM Data and AI, the AI Ladder is a framework to help organizations build an information architecture, and ultimately determine where they are in their AI journey. It's a model for how we talk to clients about their data maturity, looking at how they collect, organize, and analyze data, and ultimately infuse AI throughout their organization.

The AI Ladder is a unified, proven methodology that leaders can use to overcome the challenges we've discussed and accelerate their journey. The idea is that there are "rungs" along the way to a complete transformation, where AI has been scaled throughout every part of the organization.

> We acknowledge that our metaphor is not exact. A person with two legs can only stand on one or two rungs of a ladder at the same time, and needs to climb the rungs of a ladder in sequential order. The AI Ladder isn't exactly like that. Although there is a logical order to the "rungs," adopting this approach doesn't necessarily have to be a linear journey. You can start anywhere and build your ladder incrementally, assembling it as you might put together a jigsaw puzzle. The important thing to grasp is that this is a framework by which you can understand the steps involved in reinventing your organization. The best part about this framework is that once it's in place and practiced, subsequent AI projects get easier and easier, which allows you to expand the impact of AI across your enterprise.

The AI Ladder provides guidance in four key areas, illustrated in Figure 4-2:

- How to collect data
- How to organize data
- How to analyze data
- How to infuse AI throughout the organization

How does information architecture relate to the ladder? It's something you'll build along the way. It doesn't make sense to collect data unless you can store it effectively, and you'll certainly need to clean your data before you can do anything truly useful with it. But if you think this means putting together a substantial infrastructure full of complex tools for data governance and provenance before you can get started, you aren't likely to get started. Again, begin with a small project. As that project unfolds, you'll start to see what's needed and can begin building your data infrastructure. Then, when you tackle a big project, you'll find that you've already done much of the work, and in addition you've practiced and refined your skills.

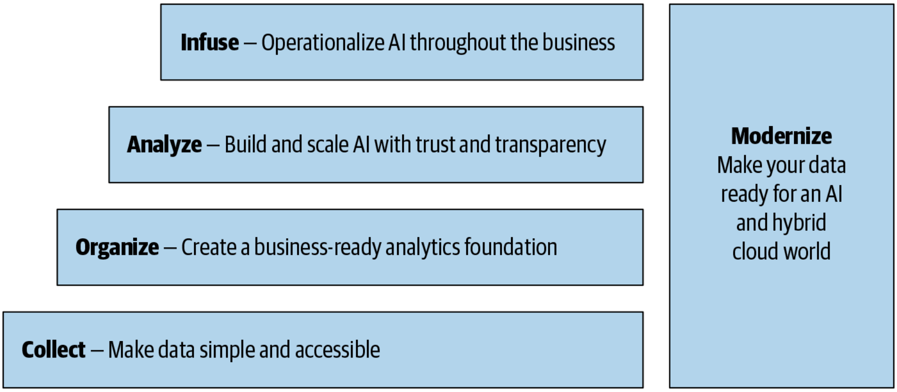

Infuse – Operationalize AI throughout the business
Analyze – Build and scale AI with trust and transparency
Organize – Create a business-ready analytics foundation
Collect – Make data simple and accessible
Modernize
Make your data ready for an AI and hybrid cloud world

Figure 4-2. The AI Ladder, a guiding strategy for organizations to transform their business by connecting data and AI

Collect

Collecting data simply means gathering and storing it. Collect data of every type, regardless of where it resides, enabling flexibility in the face of ever-changing data sources. Make accessing the data simple. Get data out of corporate silos and be aware of public data sources that may be useful. Find out what data you need but don't have, and make plans to get it; your company may be discarding important data because nobody has thought about how to use it (like log files from servers). Make the data as broadly accessible as possible, subject to regulation and your own internal policies (which may need to change); democratizing access to data often leads to new insights. Enable people to work with data; don't put up barriers or leave preexisting barriers standing.

Be prepared to use data of all types. Don't limit yourself to transactional data that's stored in a traditional relational database, just because that's easier to deal with. If you can combine structured transactional data with unstructured clickstream data from customers (like a key/value store or a document database), you'll be able to build applications that are far more valuable than anything you could construct from either data type on its own.

You may need to think about ongoing data collection, and you may need to build pipelines to manage incoming data. Don't let data fall on the floor just because collecting it is difficult; the best AI projects sometimes need the data that they need to improve themselves.

Organize

Organizing data means creating a trusted, business-ready foundation with built-in governance, protection, and compliance. The first step is making sure your data is organized and cataloged. It's surprising how few companies know what data they have, or where to find it. It's also surprising that many companies don't really know what their data means; it's common for different organizations within the same company to use the same terms in slightly different ways. Accounting and Sales could very likely have business unit definitions of a "sale" that mean slightly different things. As part of organizing your data and developing a catalog, you will have to agree on common definitions (an enterprise-wide glossary if you will), how the data is to be accessed, where it's stored, and how it's indexed. Accounting might only care about account numbers, while Sales cares about customer names. You have to make sure the data can be accessed both ways.

Next, you need to clean your data to ensure it's accurate, compliant, and business-ready. If it isn't, you'll constantly be stumbling over data problems when you're trying to do your analysis.

Finally, access to the data needs to be controlled. Only users who have permission should be able to access the data, and they should only be able to access the subset of the data that they need. You need to comply with current regulations, and with your own policies (which may extend beyond regulation). To see what this might entail, think back to our earlier discussion of the GDPR, which has stringent rules about how data can be used. Consider also what might happen if an account manager is tempted to "adjust" some figures in their favor. Data governance protects the entire company.

We've seen many companies take a "least effort to comply" approach to data governance, in effect asking "How do we do as little as possible to stay on the right side of the law?" But that's shortsighted. Sure, it may save some legal bills and keep you from getting fined, but good governance has the potential to create regulatory dividends by repurposing that same data effort for other uses that accelerate your AI journey. If you know what your data is, and can manage the metadata that describes it, you can govern it. And if you can govern it, data scientists will be able to find it and use it more efficiently. The average think "governance for compliance." The heroes think "governance for insights," and compliance comes along for the ride.

Analyze

Analyzing data is where you build and deploy your AI models. Start simple; you don't always need to use complex neural networks where a simple regression will do. We've seen this many teams jump to building a "sexy" convolutional neural network (CNN) for a computer vision project when other approaches are simpler and work just fine. Always remember to KISS (keep it simple silly) your AI solutions where you can. Building models often consumes more compute power than using the models in production. While many models seem to be built on a developer's laptop, it's usually the case that behind that laptop is a cloud—so think about the kind of cloud (public, private, or hybrid) and what cloud providers you use. There are many platforms and toolkits for building models; most cloud providers support most of these platforms, and even provide hardware acceleration.

It's sad, but many great AI projects never make it into production; unfortunately, what works on a developer's laptop frequently doesn't work at scale. Operations teams are still learning how to manage AI applications. The problem can be as simple as a matter of languages and toolkits ("You built this with R, but to run it at scale we think we need it in Python,") or as complex as difficult scaling requirements ("You need to retrain the model nightly, but training takes 10 hours.") To avoid problems with production, it's wise to integrate developers with your business SMEs and operations (DevOps, DataOps, etc.) teams as soon as possible.

Your teams will inevitably be responsible for monitoring your AI, and in particular monitoring for compliance and fairness. They also need to be concerned about models that become stale over time. Think about it: your model's accuracy is at its best at the moment you put it in production. From that point on, the forces of nature (the data it was trained on to make predictions in the subject area you're focused on) are at work to deteriorate its performance (accuracy) because the world is forever changing.

Unlike traditional software, the behavior of AI changes with time. As the environment changes—and as the AI changes the environment—models tend to become less accurate. You have to watch for this gradual degradation in performance (algorithms whose performance worsens over time are said to go "stale"), and see that the model is retrained periodically (how often will depend on the data and the use case of the AI —it could need to be retrained weekly or monthly, or perhaps daily or hourly).

Infuse

Infusing AI is where things get really exciting. Can you use AI to eliminate boring, rote tasks? (And since staff like some easy wins, can you preserve just enough boring, simple tasks to keep employees from thinking AI just made their jobs harder?) Can you build AI systems that give knowledge workers the information they need to make important insights? Can you build tools to optimize your supply chains? There's no limit to what you do when you reach the top of the AI Ladder.

Breaking an AI strategy down into pieces—or the rungs of a ladder—serves as a guiding principle for organizations. First, the AI Ladder helps organizations understand where they are in their journey: Are they still wrestling with the problems of collecting and organizing data? Or have they had some wins, and are now poised to push AI through the company? Second, it helps them determine where they need to focus. It's not uncommon for an organization to adopt AI solutions at the Infuse rung, while still building out their Collect and Organize strategies across the organization. It's important to understand that you never leave any rung behind. You may be building AI tools for every group in the company, but you still have to make sure that you're collecting the data you need, cleaning it, cataloging it, and controlling access to it. Your processes for those tasks may (and should) be simple at first, but as your use of AI grows, so will your requirements. For example, when you start building tools for finance, you will certainly be met with a different set of of explanation and regulatory requirements that you didn't face when you were building an intelligent customer service application.

Simplify, Automate, and Transform

The AI Ladder ultimately allows organizations to simplify and automate how they turn data into insights by unifying the collection, organization, and analysis of data regardless of where it lives. By employing the AI Ladder, enterprises can build the foundation for a governed, efficient, and agile approach to AI.

No matter where you are in your AI journey, transforming your organization is still going to take a lot of work. Having the data alone is not enough. Having a team of skilled engineers and data scientists is not enough. (We've seen some companies buy boutique data science firms and are still nowhere on their corporate transformation. Why? No AI without IA.) Having the technology and skills is simply not enough. Every rung of the AI Ladder is critical. But when you put everything together, you really can change your company in radical ways. You'll start realizing that there are possibilities for new ways to engage with your customers, and that will make you more profitable and leave them more satisfied.

CHAPTER 5

Modernize Your Information Architecture

Many companies have a problem with *legacy software*: mission-critical software that was written decades ago. However, legacy software is only the tip of the iceberg; the more dangerous part lies below, and that's *legacy data*. An AI transformation can't be completed unless you modernize your information architecture. That modernization is the foundation on which the AI Ladder rests (Figure 5-1). As you will come to realize (and already know from your personal lives), when climbing a ladder, foundation matters. A solid foundation lets you climb higher, faster, and with more confidence.

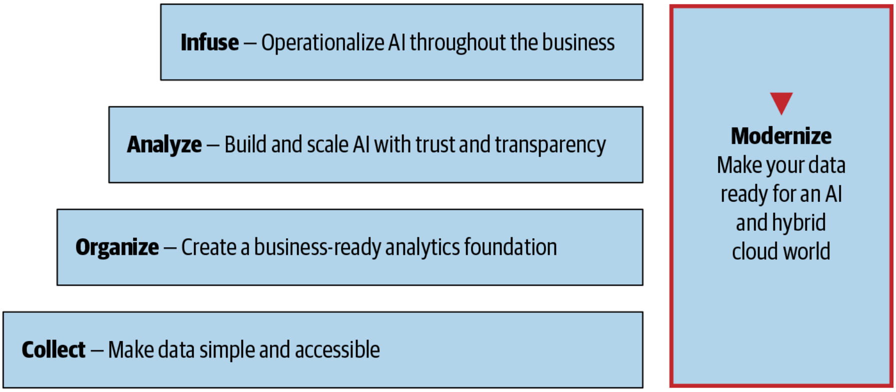

Infuse – Operationalize AI throughout the business
Analyze – Build and scale AI with trust and transparency
Organize – Create a business-ready analytics foundation
Collect – Make data simple and accessible
▼
Modernize
Make your data ready for an AI and hybrid cloud world

Figure 5-1. A modern information architecture is the foundation on which the AI Ladder sits

What's the problem that needs to be solved? Why is legacy data an issue? Too often, data is stored in departmental silos: every department has its own database (and these databases often contain the same data at varying degrees of staleness), its own rules for who can access it, and its own definitions for what's in the database. Working with siloed legacy data is obviously a problem.

Fragmentation isn't limited to departmental database silos either. Many companies have rightly taken a leap into the cloud, but without addressing the silo issue. That means that their data lives in many clouds: some in IBM's, some in Amazon's, some in Microsoft's or Google's, and some in a private cloud. Every department has its own solution—often a solution that was crafted primarily as a way to get around the IT organization. Silos are silos—be they in the cloud or on premises. How do we create some kind of unified access to this mess?

Data is less likely to be lost if it's managed in a central repository than if it's stored in some department's private database. Some companies have centralized data in an enterprise data warehouse, which takes care of some standardization issues and provides an attempt at a "single version of the truth" for corporate data. These are important advantages over fragmented, ad hoc databases. But data warehouses can also have their disadvantages. Accessing data in a data warehouse tends to be complex and is rarely democratized (hence the departmental data silos). If getting data requires making requests to a database administrator (DBA), then waiting hours (often overnight) to get the data you need, your AI experiments are going to take much longer. What's more, the open standards process model for data science—the cross-industry standard process, CRISP-DM—includes a continuous cycle of experimentation, iteration, and learning. Continuous experimentation and iteration can certainly include schema changes to the data model, and data warehouses have been notoriously slow to implement schema changes.

Modernizing your infrastructure means unifying the collection, organization, and analysis of your data, regardless of where it lives, within a hybrid multicloud data and AI platform. As Daniel Hernandez, IBM's VP of Data and AI, says, this architecture needs to be available wherever data is. "You should bring AI to the data," Hernandez says, rather than hunting the data down in departmental silos and bringing the data to AI. His advice is a cornerstone of any modern-day analytics architecture that takes advantage of massively parallel processing (MPP) and high-performance computing: ship function to data, not data to function.

Hybrid clouds are inevitable solutions to some of the real problems facing modern enterprises, which include:

- Data that must remain on premises for regulatory reasons
- Data that is too bulky to move to a cloud provider (for example, sensor data from a large, highly automated factory)
- Data that ended up in different public clouds because of different cloud efforts within the organization or through merger and acquisition activities

Should you rationalize this picture with a single cloud provider? That may sound like the best of all possible worlds, but in practice it's a huge, expensive project and a solution that may take years to implement—we generally don't recommend it, except for specific circumstances. Sending terabytes of data from an instrumented factory to a central cloud (and getting it out again) on the daily can be slow, expensive, and ineffective.

Modernizing your infrastructure means investing in tools and processes that can provide a single view of your data across multiple clouds and databases. Getting data out of separate departmental clouds is a battle not easily won. A hybrid multicloud data infrastructure, composed of private clouds, public clouds, and on-premises systems, allows you to access the data wherever it is. The best of all possible worlds isn't wrestling all of your data into one cloud; it's building data infrastructure in which you don't have to care where your data is located.

This kind of infrastructure can provide greater operational agility and the power to identify, analyze, and respond quickly to changes. Such platforms not only address challenges with data quality, but also provide access to more data to fuel smarter AI. And these platforms are flexible enough to accommodate new kinds and sources of data as they become available. Taking advantage of "the cloud" means realizing that the cloud isn't just another place, another destination for your data; it's a capability that, when properly implemented, means that you no longer have to think about "place."

In this chapter, we'll discuss, at a high level, how to modernize your data infrastructure. We'll briefly allude to the Collect rung and survey various methods of data collection, and how they have evolved over time. Then, we'll look at data virtualization, which is a relatively new way of storing, accessing, and thinking about disparate data sources. Then we'll explore why data governance is the key to modernizing the Organize rung. Next, we'll discuss how modernizing the Analyze rung means automating the building, running, and managing of AI models in your organization. Finally, we'll cover how to extend your platform with additional AI applications—a prelude to the Infuse rung.

A Modern Infrastructure for AI

Figure 5-2 shows the features that a modern data infrastructure provides.

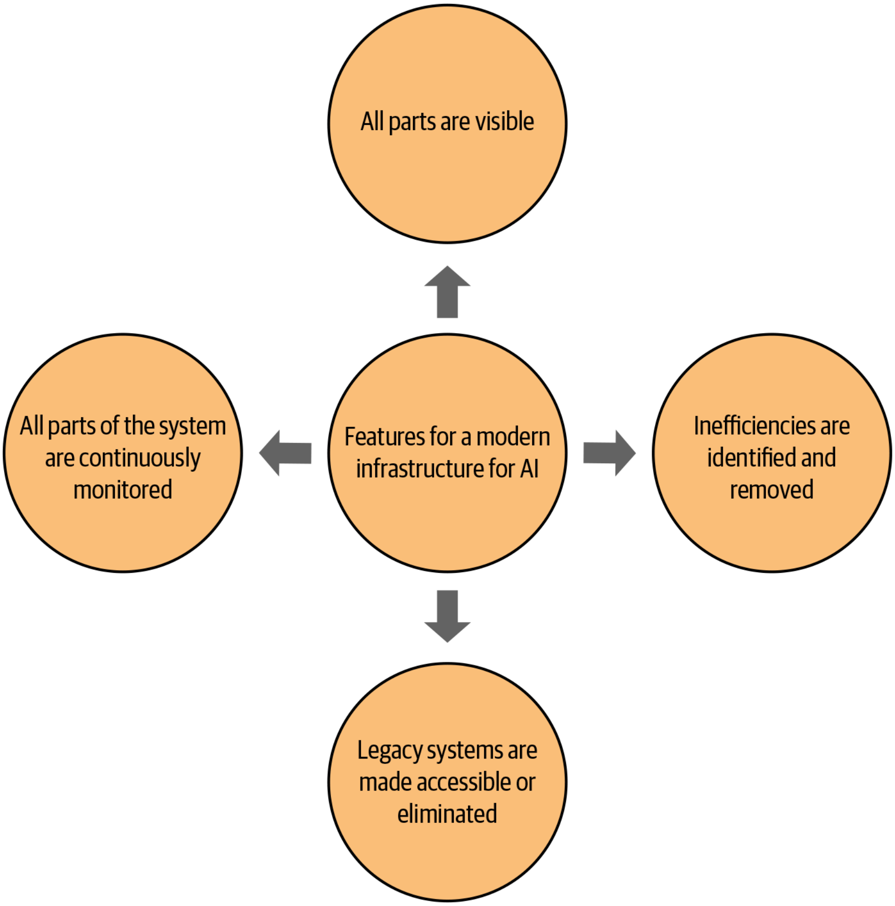

All parts are visible
All parts of the system are continuously monitored
Features for a modern infrastructure for AI
Inefficiencies are identified and removed
Legacy systems are made accessible or eliminated

Figure 5-2. Features of a modern infrastructure for AI

Let's explore each of these features.

All Parts Are Visible

Most organizations today already have an infrastructure that uses public and private clouds, on-premises systems, and the rest. For any reasonably managed enterprise, identifying most of these resources should not be too much of a challenge. So, aren't you already modernized? In our experience with customers, the answer is "Maybe, but probably not."

Many organizations do not have a comprehensive understanding of where their data is stored and analyzed. This problem goes beyond the notion of "silos," where data is isolated and useful only to local groups that know about its existence, have permission to use it, and in many cases (as was a problem with Hadoop) possess the programmatic skills to access it. Some organizations don't even know how many silos they have, or what silo technology they're using—or what it's costing them. (To prove our point, spend a few weeks trying to figure out how many Microsoft Access databases or Microsoft Excel spreadsheets are running critical business processes. How many did you know about before you started counting? We know how the story ends. Do you?)

So, the first step on the road to modernization is understanding all of your data assets and the infrastructure they rely on. This includes your physical resources as well as your actual data. This may mean adopting a data platform, or investing in cloud management software, or a custom build of an ideal console for your multicloud scenario. In any case, you cannot have fully modernized infrastructure without first understanding exactly what resources you have, and ensuring that you have the ability to control them and interact with them in a unified way.

Legacy Systems Are Made Accessible or Eliminated

This is a tough one, because legacy systems are often the heart of an organization, and doing a heart transplant is a scary prospect. But this is one place where you confront the reality of adopting AI. If you are not ready to make this step, your organization is not ready for AI.

This is not to say you have to rip out that legacy system in order to do AI. But the legacy system holds data, and it needs to at least be updated so that data can be cataloged, gathered when needed, updated when necessary, and continuously available in some cases. This may require middleware that understands how to talk to legacy applications while also providing a services-based API engine that responds in real time to programmatic calls made by your data platform. Or it might require a streaming solution so that events that happen on your legacy platform are written out to a more modern data lake, or some other accessible location, in real or near real time, so that they're available for use outside of the legacy system. It's also important to note that many legacy systems are evolving and becoming "less legacy." Many vendors have realized their customers aren't comfortable replatforming enterprise-hardened processes that the business relies on. They've focused on making those practices open with access APIs, so they can operate as first-class citizens in a modernization agenda. If your legacy systems aren't open, push hard on the vendor to make them open, or plan to get off them. Data virtualization, which we discuss later in this chapter, is one solution to the problem of legacy data.

One thing is for certain, however: a legacy system that is unaware of how data is consumed today, and that is not updated, is a blocking issue on the road to an AI-driven organization. It's not only possible to build a data architecture that makes legacy data accessible, usable, and protected as part of an overall data governance policy; it's your job to do it. As an example, let's take a look at how Network Rail, a British railroad company, made the transformation.

Example: Network Rail uses AI to modernize its infrastructure

Unlocking data from legacy enterprise systems can deliver terrific value, but it is also challenging. At Network Rail, which is responsible for the infrastructure of the UK's railway system, data is generated at incredibly high velocity and in incredibly large volumes. Network Rail spends £130 million ($168 million) every week on improvements for passengers, accounting for 22% of the UK's entire infrastructure spend. It is also Britain's biggest builder, supporting 117,000 jobs. The company has thus amassed a huge amount of data, and it's very good at using this data to describe what it did last month. The problem is, it wasn't so good at looking into the future.

To change that, Network Rail's infrastructure projects are on course to become more data-centric. A first goal was to develop a "single version of the truth" that would impact three key areas: business operations (using data to make business work better), reports (allowing project managers to look at the relevant data and make decisions), and analytics (to help them do their jobs better).

Siloed reporting was a major roadblock, but even more serious was inconsistent understanding and use of data between different departments, regions, and executives. Consolidating Network Rail's existing data to enable consistent analyses was a start. But the firm wanted to unlock the value of data through machine learning and prescriptive analytics to predict the future: to understand where projects might succeed or fail and what could be done to ensure the latter didn't happen.

As with any IT project, Network Rail realized it was important to start small, fail fast (and safe), and add value quickly. So it created a machine learning minimum viable product (MVP) targeting parts of the data environment where it knew it had the most significant data quality problems. Network Rail's machine learning model focuses on the interrelationships between source systems—schedule, costs, risks, funding, past performance, clients, and suppliers. It automates data feeds and identifies which of them to use to predict success probabilities for different projects, such as the construction of a new station, electrification, a new tunnel, improving bridges, land management, or repairing railway lines after a catastrophic event.

This led to the identification of a number of challenges, such as making sure the solution was grounded in the real world, cleaning up noisy data, automating manual processes, and managing the changing technology and vendor landscape. As a result, Network Rail has improved its year-upon-year forecasting, improved process compliance, enhanced data quality, and unlocked reporting for users across the organization.

All Parts of the System Are Continuously Monitored

A modern information architecture has the ability to describe itself in real time. You must be able to interrogate your infrastructure to learn what activities are going on, and where they are taking place, at any point in time. Information about your information infrastructure should surface in a single place, and constant alerting and reporting should be integrated into existing enterprise IT monitoring systems. If you've had experience with IT infrastructure where you had to look in one place for one piece of information, and somewhere else for another—a different application, or maybe even a different physical console in another location—you know that's an antipattern to success. Uniform, consistent monitoring is one key to a modern data infrastructure.

Inefficiencies Are Identified and Removed

The ability to monitor capabilities and operations in real time allows you to aggressively control costs and optimize performance. The same goes for computational resources (real or virtual) and software as a service (down to the level of functions as a service). Next-generation databases and a data platform layer can assist in finding these inefficiencies and locking them down so that you can continually reinvest the savings; for example, the ability to separate compute and storage costs for cloud-based projects is a major cost savings benefit, as is autonomic compression, and so on.

New Architectures for IT

IT continues to evolve. Over the last few decades, new phenomena have become attractive to IT departments for one reason or another. These phenomena include internet-based data storage and processing ("the cloud"), entirely new database architectures designed to handle unstructured data, virtual machines and clusters of virtual machines, and even new pricing models for software. In a cloud model, you pay (directly or via a departmental charge-back) for the storage you use for the actual computing time used, and optionally even for the execution of individual functions. If you look at a typical on-premises computer facility, you'll see a lot of excess capacity, and that excess capacity is money spent. There's no way around this wasted capacity; if you have a sudden surge in traffic, you need the ability to handle it. For example, if you're a retailer and you offer a sale, you might see a 100× traffic spike. If you can't handle that, your customers will be angry and may not return. A cloud model allows you to only pay for the computing power you actually use—no more—and the computing power you use dynamically grows or shrinks according to your needs. (Note: despite what many believe, cost savings isn't the number one reason to use the cloud...it's agility.)

But as organizations have dipped their toes into these technological waters, many have lost sight of their overall information architecture. As a result, they often don't have a clear understanding of where their data is actually stored, what data they actually have available, and what it's costing them to retain and service that data.

AI is bringing this problem into clearer view—or rather, organizations' desire to embrace AI is highlighting data preparation and governance problems—and this is driving new ideas about data architecture and data infrastructure to the forefront.

Data: The Fuel; Cloud: The Means

Building a modern data infrastructure isn't a one-size-fits-all operation. Do you want a data warehouse? A data lake? What about NoSQL databases or object stores? It depends on your needs. You have to consider your organization's assets and challenges. It also isn't a quick, simple project; you will need to get started on some AI projects *while* you're building the infrastructure—and you can use the problems you encounter in those initial projects to focus your attention on what infrastructure to build for the long run.

To start your thinking about data infrastructure, we'll look at why it's a necessity, what problems it solves, and what options you have for building a solution. Data infrastructure impacts the Collect and Organize rungs of the AI Ladder most profoundly, but there is no rung on the AI Ladder it doesn't touch. Fortunately, modernizing your data infrastructure will improve everything it touches.

To the Cloud, and Beyond: Cloud as Capability

The first law of data is that it expands to fill the storage available—and that's why you need to think seriously about infrastructure. When IBM created the first commercial disk drive in the 1950s, it had a storage capacity of 5 MB (the approximate size of a single song on your phone) and was the size of a refrigerator. Now we have multiple-terabyte drives the size of a paperback novel (in the consumer market, they're the size of credit cards). You almost certainly have more digital data on your laptop than any large company had in the late 1950s.

Where does that data come from? The web, mobile devices, digital media, the Internet of Things, and other recent technologies have certainly accelerated the pace of data creation—though it may be more accurate to say that, in many respects, the data was there all along; we just didn't have a way to digitize it or anywhere to put it. Given today's storage capabilities, that's almost frightening. This era's large corporations have data by the exabyte (1 exabyte equals approximately 1 million terabytes).

Even that's not the entire story. For over a decade now we've been talking about storing data in the cloud, where the storage available is, for all practical purposes, unlimited. What comes after exabytes? Zettabytes and yottabytes—and yes, the cloud providers will eventually offer storage at zettabyte scale. If data expands to fill the space available, that's a *lot* of data. The cloud doesn't enable the creation of more data, but it does give you someplace to more easily put the data you have created.

The cloud also solves some other important problems for data storage. You could assemble a personal petabyte—just buy 125 consumer-grade 8 TB drives and put them in a large bookcase (with a rats' nest of cabling in back). It will cost you substantially less than a mid-range car (as we write, a single 8 TB disk drive costs under $150). But you won't be able to effectively back up your data, you probably won't be able to move data into your disk array fast enough to fill it, and whenever one of those drives fails (and you can count on the fact that drives fail), you'll be dead in the water. These are problems cloud providers solve for you; staff at cloud data centers are constantly replacing failing drives without loss of data, and several cloud providers have services where they (literally) back up a truck full of drives to your loading dock to haul your data to their facility. You've heard of sneakernet (the informal term for moving data around by physically transporting it, such as on a USB key)? This is forklift-net.

The cloud is also an important challenge for modernization. It's common to talk about "the cloud" as if it were a place: IBM's cloud, Amazon's cloud, Microsoft's Azure cloud, Google's cloud, or even a "private cloud." As the title of this section implies, we see the cloud as a *capability* (not necessarily as a destination): it gives you the ability to access data regardless of where it is, whether it's in your own data center, out on the "edge" of the network, or held by a cloud service. Location independence is important because you're likely not going to keep all your data in a single cloud or work with just a single cloud provider. Many companies have started moving to the cloud without a coordinated plan, which means that several groups within the company are likely to have their own cloud projects, using whatever vendor happened to seem right. You probably have data that needs to remain in your own data center. And if there's ever a merger, the other company will probably have its own data and its own cloud vendors, and they won't be the same as yours.

That's why it's important to see the cloud as a capability, and to build infrastructure that doesn't lock you in to a single vendor. Multiple clouds that won't interoperate, that get in the way of data access rather than simplifying it? That may look like modern infrastructure, but it isn't. You need to build a multicloud platform.

Fuel for the Fire

To AI, data is like fuel for the fire. Without data, algorithms have nothing to process. But the quality, accessibility, and governance of that data play important roles in how much you can trust the results of your AI and machine learning projects. One thing we have learned in our engagements helping companies through their AI transformation journeys is that "data" in most companies is not unified, not easily accessible, and not governed. When you think about it, it's quite simple: if you can't access data, you can't use it, and accessing more and more data is fundamental when it comes to using AI to detect patterns and intuitions about a business from its data. If it's difficult to access, you will have trouble building AI applications on top of it. And if you don't know where it came from, or how it has been managed, you won't be able to trust the results it produces. Remember this handy phrase: *data is the fuel, cloud is the means, and AI is the accelerator*.

But also remember what that handy phrase doesn't mean: make no mistake about it, a bunch of big disk drives and a lot of data stored in several different clouds isn't a data architecture, it's a mess. And if you want to build a good fire, you don't just throw logs on randomly. You're intentional; you start with the kindling, move to the smaller branches, then add the big logs, taking care to ensure that there's room for air to get into the center.

Most companies already have plenty of information. They have been collecting it for years, in the form of transaction receipts, inventory statistics, website analytics, event logs on servers, or third-party data that is used to assist in business processes. But they don't have a modern data infrastructure; they have some version of a mess. The problem is this: for AI, you need to be able to tap this information quickly and easily, and also know where the information comes from and how it was processed and shaped. Traditional ETL (extract/transform/load) approaches simply won't work, for these reasons:

Latency

Accessing legacy data might be slow. It doesn't matter whether it's slow because of an ancient database, ancient networking technology, or working through the DBA team. It still might take anywhere from minutes to hours to get data (and more likely days to weeks—even months—for data structure changes) back after you request it. Combining lots of data from different sources is fraught with latency.

Risk

There may be problems with data integrity or quality, which we'll learn more about as we get into this chapter.

Data islands

There may be crucial data sitting in a system that uses proprietary formats and isn't fronted with a service-oriented architecture or APIs that can easily access it. What do you do when your data is on one of these islands and can't afford to get voted off it? Even more concerning are data islands that started as copies of other data, then the data on the island was changed so that it meant something else, but the metadata never changed. If your data has changed and no longer matches the metadata, you're in trouble. This is a big problem, especially in the area of financial reporting.

Couple all of that with the fact that your organization may not even be collecting certain kinds of data, like high-volume event-based streaming data, because your existing data technology is not able to handle it, or it may be dumping this data to storage with no real plans to derive value from it. How do you begin understanding such data and integrating it within your organization's overall data plan?

Let's reconsider the problem of data silos. Data silos are as much an organizational problem as an infrastructure problem. The job of a data infrastructure is to give you responsive, performant, and trustworthy company-wide access to data. If data is fragmented into a lot of departmental databases, you're not going to achieve that—especially if those databases aren't in sync with each other. That's a mess. If data is "in the cloud," but "in the cloud" means that developers in different departments have pulled out their credit cards to get around IT department procedures and rules, that's a different kind of mess (we call it *Shadow IT*). And all of that mess is out there in abundance. In fact, it's more the rule, rather than the exception. A modern data infrastructure will need to address it, in all its forms.

From Databases to Data Warehouses, Data Marts, and Data Lakes

Data infrastructure entered the pre-modern era with relational databases. Relational databases were invented at IBM in 1970, though the first commercial relational database management system (RDBMS), Oracle, wasn't released until 1979. Since then, they've become the standard for storing data in bulk. Relational databases are fast, can store huge volumes of data, can be distributed across many computers, and support features that businesses care about, like transactions. However, the most important reason for the success of relational databases is probably the structured query language, SQL, that is used to access them. SQL was invented (also by IBM) in the mid-1970s, and soon became a common language supported by most databases and essential knowledge for data engineers. It has become so ubiquitous that when new databases and ways to access data were developed, adoption was slow until a SQL interface was developed (when Hadoop was in its prime, it ironically became a battle-ground of Hadoop-enabled SQL engines). That's no less true for building a data infrastructure today: a common language is a key enabler.

Over time, models built on top of the RDBMS model began to emerge. Data warehouses and data marts (Figure 5-3) were solutions to the problem of how to make data more accessible for business intelligence (BI) applications. An enterprise data warehouse is essentially a central, structured, company-wide repository for all of the organization's data (at least that's the intent—how it ends up varies). Data (almost always structured) is first cleaned, then sent to the warehouse from other databases throughout the company so the entire organization can have a high-performance, trusted, single-version-of-the-truth data repository (or so the intent was—but it often doesn't turn out that way). The data warehouse architecture enables efficient storage and retrieval for data in bulk, but the downside is that accessing the data calls for upfront schema modeling. That is, you need to describe the data—you need to know what will be in each column of a data table—before you can start storing it. Changes to the schema are difficult and expensive. As we'll see, that's a significant disadvantage for AI projects.

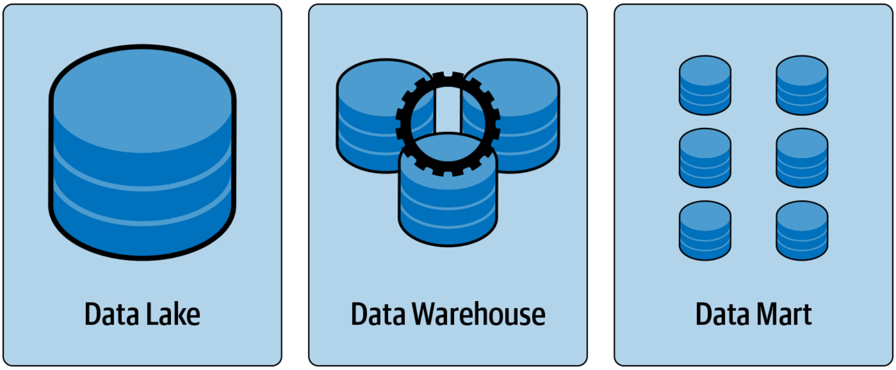


Figure 5-3. Data lakes, data warehouses, and data marts

Data warehouses can be difficult to access; even what seem like simple queries (like bill of materials processing or best routings) can become very complex. Operating groups, such as Sales, still end up with their own databases because data warehouses aren't designed to turn around individual operations, like making a sale, quickly. The data mart evolved as a solution to this problem. The data mart contains a subset of the data in the warehouse particular to a given department, and prewritten queries (typically performance-optimized using proprietary technologies) particular to that department's needs. For example, the data warehouse might contain information useful to both the Sales department and the Manufacturing department. For ease of access and faster performance, each department would have its own customized view. Those views of the data would be the data mart. This approach provides ways of narrowing scope and reducing complexity to address well-defined problems (not to mention processing times). The cost of this optimization is a reduction in flexibility and adaptability, but these benefits are taxed with trust issues and data sprawl. A data warehouse takes data out of silos; a data mart rebuilds the silos within the warehouse (a virtual data mart), but often outside of it, with different kinds of technologies that require different kinds of access protocols.

Data lakes go in the other direction. A data lake combines data from a variety of databases, and by definition is not optimized for any particular application (Figure 5-4). Rather, it's designed to hold "all the data" (structured and unstructured) so that it can be made available to whatever application may require it somewhere down the line. The cost of this generality is typically a reduction in performance and understandability. The process of building a data lake is often reduced to "Collect the data, then forget it." Data lakes tend to fall into the schema-later category we talk about a bit later (see "Next-Generation Databases" on page 93), because it's easy to put data in and hard to get it out—hence the "then forget it" bit.

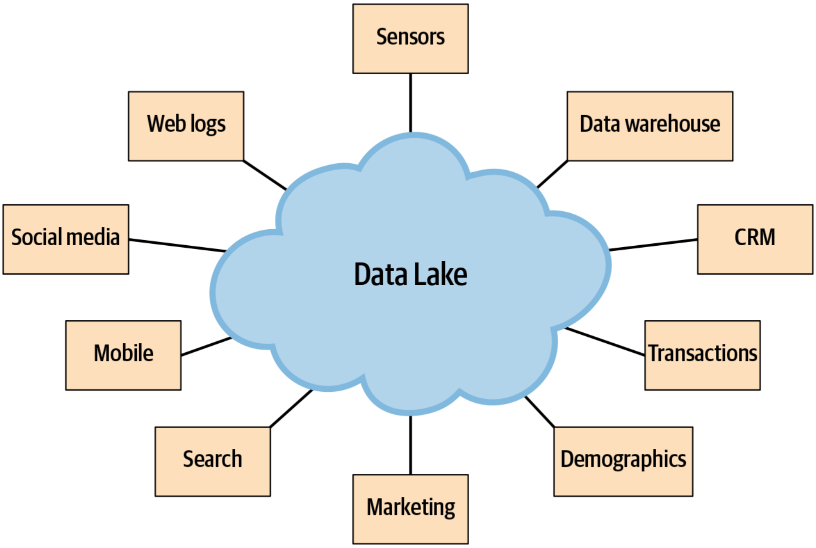


Figure 5-4. Sources that feed a data lake

Gold mining provides an appropriate analogy for thinking about data lakes. During the Gold Rush, early prospectors sometimes saw gold nuggets the size of a fist (think of this as high value per byte data) just sitting in a stream waiting to be found. Towns sprung up and investment went in because the value was obvious; that's a data warehouse in our analogy. But those large, easy nuggets quickly became harder to find. Gold mining today is a different matter. It uses different capital equipment (open source) never available in the prospecting era and involves sifting through tons of dirt (low value per byte data) for specks of gold (new insights) nearly too small to be seen by the naked eye. The process of finding what you're looking for in a data lake can be worse than looking for a needle in a haystack. It's like looking for one particular needle in a haystack-sized pile of needles. It can be virtually impossible.

Data lakes can become so vast and lack so much governance that not even expert administrators understand everything that's in them or how best to access the information they contain. This situation is often referred to as a "data swamp."

But when done well, data lakes solve a lot of problems: they resolve inconsistencies and duplicates in data, and act as a single source of truth for the company (not to mention being especially useful for data preparation and queryable cold data archives). There's nothing more frustrating than the game of "dueling databases": when two databases have the same data, but they disagree. Even when the disagreement is small (especially when the data is small—big differences tend to be easier to resolve), it quickly becomes hard to trust the data. Employees have to start second-guessing everything, and do their own data validation—and that's not a good way to work. The results are neither reliable nor trustworthy. Data lakes make it easier for staff to find and trust the data that they need—but without governance (as was the case with most initial data lakes), the data swamp makes it harder to find the data.

Don't misunderstand us...we're not saying you aren't going to have data lakes, data warehouses, or data marts. But we want you to be on the lookout for some of the pitfalls of each solution as you plan out a modernization strategy. Knowing what to look for is as important as knowing what to do with a problem when you find it!

Example: Wireless Carrier Architects a Solution Using Both a Data Lake and a Data Warehouse

A top-tier wireless carrier needed to improve its text-based marketing effectiveness. Most of the more than four million messages a day it was sending to its subscribers contained unique URLs linking to mobile web content, and the wireless carrier was looking for a better way to measure how subscribers interacted with these text messages and links.

To be able to measure all of these interactions, the firm created a data architecture that consisted of both a data lake and a data warehouse. This would allow the company to log each outbound message, the unique URL included in each message, and click behavior metrics on a massive scale.

The architecture was built using:

- A PostgreSQL database for transaction processing
- An object store data lake for storing log files
- A data warehouse for reporting and analytics

Why both a data lake and a data warehouse? For starters, the object store data lake is a low-cost, flexible, and easily accessible repository to store customer data, regardless of when or how it will be analyzed.

Secondly, a data warehouse significantly improved the carrier's previous ETL process, which had taken several hours to complete and was done every night. Instead of doing nightly exports from the database to the data warehouse, transaction log files are created by the PostgreSQL database and sent to the object store every 5 minutes. That data then gets loaded from the object store into the data warehouse, and SQL scripts are run to aggregate or transform the data for specific reports. (In this case the team had deep SQL skills, so they opted to perform their aggregations and transformations within the database—which makes a lot of sense for the right use case.) These reports include aggregated messaging statuses, delivery statistics, click-through rates, and many more metrics, by various campaign types across date ranges. So instead of loading 24 hours' worth of data once a day, the wireless carrier can now load data every 5 minutes, enabling it to access a near-real-time view of its customers' behavior.

The PostgreSQL database ensures that the customer behavior data is valid; the object store is the data lake that stores the log files used for analysis; and the data warehouse provides fast data reporting at scale. This shows how you can design an architecture that includes a database, data warehouse, and data lake to store and analyze data at massive scale. This architecture allows the carrier to better understand how its subscribers are interacting with their messages, measure the success of text-based campaigns, and gain insight into planning future campaigns.

Data Virtualization

If data warehouses are often slow and hard to use effectively, and if data lakes frequently turn into data swamps, what other options do we have for building a modern data infrastructure?

"Today, customer environments typically have tens of thousands of databases," says IBM's Daniel Hernandez. "If you want to deploy AI, you need to be able to tap into that data." Traditional techniques, like ETL, are often expensive and brittle in modern environments. Data virtualization lets organizations tap into that data—within those tens of thousands of databases—without moving it.

You may be familiar with virtualization as it pertains to virtual PCs and servers: consolidating many workloads onto a single physical host by abstracting away the physical host's components and carving up their resource capacity into individual virtualized computers that operate independently and run their own operating systems. Each virtual PC thinks it's a distinct PC, with the hypervisor layer on the host coordinating how physical resources are shared, while addressing security boundaries and keeping memory siloed.

The key idea behind virtualization is abstraction: the complexities and peculiarities of the host's physical components are abstracted away and presented to the guest operating system as standardized devices. It doesn't matter if you have an LSI SCSI adapter with an AMD processor on one system and an NVIDIA graphics card with POWER9 processors on another host. As long as they're running a hypervisor, the same standardized virtualized hardware will be surfaced to guest operating systems. This lets functions like live migration and virtualized clusters work correctly from host to host.

Take that same idea of abstraction and apply it to data instead of physical hardware. Now you have a good basis for understanding data virtualization. As we've said, most organizations have data spread across a variety of environments: public clouds (with many providers), private clouds, and traditional on-premises deployments. This results in the data proliferation challenge that many organizations are facing today: information is distributed across multiple silos, databases, and clouds. That information may be in multiple formats and structures, and consumed or generated by different applications.

Data virtualization masks these differences and makes the data available to users in a single place. Data looks as if it resides in a single database and is accessible via a single program or application, but in reality the complexities of how that data is stored in the enterprise IT infrastructure are abstracted away.

Data virtualization has several advantages:

- Data can reside in one spot or several spots, but wherever it resides it's always easily available. Data virtualization technology, combined with a robust query engine, allows you to access data in traditional SQL databases, Hadoop, NoSQL databases, and legacy systems as if it were all stored in the same "vault," while in actuality you are querying several distinct systems. This technology lets you modernize your access to data without disrupting your current infrastructure, changing the underlying storage mechanisms, and in many cases changing your applications. You create agility and flexibility by wrapping your legacy technologies in a virtualization layer.

- Your investments in legacy technology can be preserved, but you can still take advantage of next-generation query engines. Most data virtualization applications have connectors for legacy technologies or have layers that act as "middleware," so that data locked away in legacy systems can still be exposed and you don't have to throw away investments you've already paid for.
- You can have a single copy of originating data while performing transformations and edits on virtual copies of the data. This helps protect the integrity of the underlying data while also providing the freedom to edit, modify, and process as necessary.
- You remain agile and able to change your infrastructure and the underlying resources without affecting how data is delivered. You can undertake months- or years-long infrastructure shifts that end up transparent to the end users because their virtualized access to data is never disrupted. You can even migrate to public or private clouds, or some mix thereof, depending on your organization's roadmap. We implied this, but it deserves an explicit mention: data virtualization can drive downstream changes (and testing) for legacy apps that rely on underlying infrastructure knowledge.
- You can enforce central data governance and security policies. Even if your governance can't reach the underlying data, it can certainly control access to that data and how it is used—so even legacy products that won't integrate well with access and lifecycle policies can now benefit from the protection of a good governance policy. For example, data virtualization layers could apply enterprise-wide masking policies to personally identifiable information (PII) that aren't a feature in the underlying system.

Data virtualization is an essential piece of the AI puzzle for large organizations, making it possible to unify access to data for analytics and governance purposes. It's a compelling way to turn a collection of fragmented data sources into a hybrid multicloud infrastructure. Having a solution to surface data in this way is a significant step in your transition to becoming an AI-centric organization. IBM's Cloud Pak for Data provides a toolset that is cloud-independent (it supports private clouds and not just IBM's cloud, but all major public clouds) and enables you to use and manage data regardless of where it is located.

Let's take a look at some other ways to unify access to data through a single query engine: Big SQL, object storage, open data stores and formats, and NoSQL-style databases.

Unifying Access to Data Through Big SQL

IBM makes tools and technologies that can help you query data no matter where it lives. IBM Big SQL is a high-performance massively parallel SQL engine that seeks to make querying enterprise data from across the organization easier and more secure. It allows you to access a variety of data sources—including Hadoop's HDFS and WebHDFS, relational databases (not just IBM relational databases), NoSQL databases, object stores, and more—using a single database connection or single query.

Big SQL's engine can execute complex queries for relational data and Hadoop data. Big SQL also has an advanced SQL compiler and a cost-based optimizer to help make query execution more efficient. For example, perhaps a certain data type doesn't exist in the underlying data system. Big SQL can "function compensate" for this and map references to nonexisting data types in the code to the legacy system (even if it must do some "magic" in between). Big SQL can accelerate performance through strategies such as branch-tree elimination and predicate pushdown on its query execution plans to mitigate tradeoffs between performance and accessibility. Combining these optimizations with an MPP engine helps distribute query execution across nodes in a cluster.

Why would you choose to use a tool like Big SQL? To achieve greater scalability and performance. Consider the following two use cases:

Enterprise data warehouse (EDW) offloading
: Big SQL understands commonly used SQL syntax from other vendors and producers. You can quickly and easily offload (for example, to Hadoop, among other options) and consolidate old data from existing enterprise data warehouses or data marts, while preserving most of the SQL from those platforms. This is an excellent opportunity to create an online query archive and lower the cost of storing data that has "cooled" (is not accessed as much), while still making it easily available for AI.

Federated access to relational data
: For data that can't be moved, Big SQL provides federated access to many data sources (with IBM Fluid Query technology) as well as SQL and NoSQL databases (with the use of Spark connectors). For example, you can use a single database connection to access data across Hadoop and dozens of relational/NoSQL database types, whether they are in the cloud, on local systems, or both.

Big SQL and similar analytical query engines make data that lives in disparate places in your organization accessible within a single tool—a key facet of preparing for an AI transformation.

Object Storage as the Preferred Fabric

Instead of thinking about data as records, or as files stored in kilobytes and megabytes, object storage thinks about data as distinct units—objects. These objects comprise the files and folders that make up data, with attached metadata and a custom unique identifier that uniquely labels a piece of data. In contrast to block storage (which stores data based on its physical address, sector, or cell on a disk or drive), or file storage (which uses a folder metaphor), object storage is essentially an abstraction. You have data as objects, integrated with the object's description, and you're off to the races, just scaling out widely as your overall corpus of data grows.

This architecture adds comprehensive metadata that describes the data's lineage and integrity, and eliminates the tiered structures used in flat file storage (hot versus cold items and the subsequent placement of those items in optimized tiers or areas of a disk closer to a spindle for performance reasons). Everything is just stored in a space with a single layer, known as a storage pool (Figure 5-5). An object store is essentially limitless. You can continue adding data infinitely as long as you describe it properly; you no longer have to store data hierarchically in a filesystem and know the path to it, and you can scale out your storage across hundreds or even thousands of nodes to drastically reduce the actual cost of storing data.

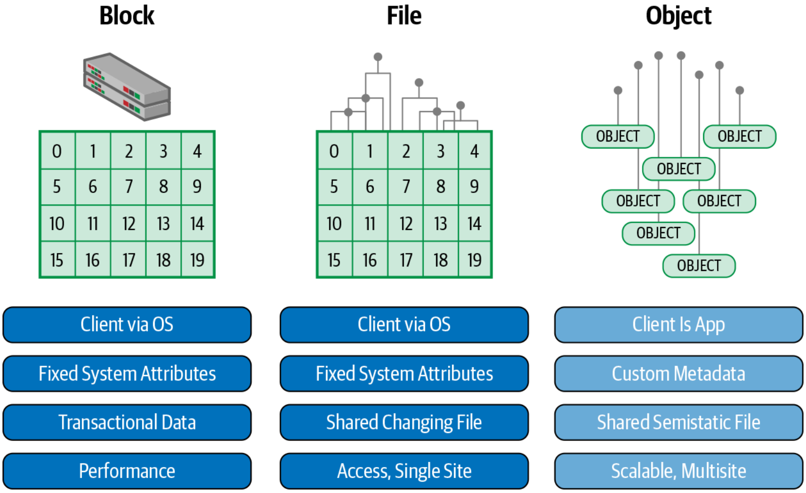


Figure 5-5. Object storage versus file or block storage

Another benefit of object storage is that as new forms of data emerge and new formats come along to describe those new forms of data, they can simply be added and retained within your existing storage architecture in an object storage fabric. There is little need to change or reinvent anything as the future draws nearer.

Object storage abstracts away the physical ones and zeros on a disk. It makes classifying and using data much easier and improves an organization's ability to keep data at a minimal cost. It's a strategic architectural step in the journey to AI.

We felt it important to comment that you will still use different storage formats depending on the task at hand. High performance (as in throughput) requirements typically don't use object storage (as Figure 5-5 denotes). Just as your modern architecture will include a polyglot of database technology (SQL, noSQL, and so on), so too will storage—but object storage will be the main part of the fabric.

Open Data Stores and Open Data Formats

When in doubt, the new preference is for data to be stored in open stores and open formats for maximum compatibility and lifespan. Data stored in proprietary formats is harder to catalog, access, consume, and share, and in some cases it is nearly impossible to use anywhere but within the originating application. Open formats like JavaScript Object Notation (JSON), comma-separated values (CSV), and open tabular formats ensure data is long-lived because it can be used easily in contexts other than the originating program.

Almost everyone knows about open data formats for tabular data: CSV files can store just about anything you can dump into Excel. What isn't so widely known is that open data formats exist for other kinds of data too—including geospatial data, with all of its descriptive attributes about counties, countries, states, roads, and more, which is well served by geoJSON. Many cloud tools can easily import open formats such as these, making data more portable and extensible. For example, JSON has become a ubiquitous data transport layer in the cloud, in NoSQL databases, across web services, and more.

While JSON is very widely used, you may want a data format that provides more. For one thing, JSON is a plain-text format that doesn't address security. And while JSON itself is standardized, there are few standards for representing specific kinds of objects, or for representing the metadata that's essential for issues like access control. Going beyond JSON, we see more and more clients demanding open and standardized data protocols. The Open Data Platform Initiative (ODPi), a nonprofit organization, has responded to these demands by making its sole mission to work with vendors across the data ecosystem in standardizing formats in the big data industry, "so that data can be easily and securely shared across products, platforms, and systems." These standards are designed to be open source and vendor neutral so that data governance, connectivity, and analytics, among other things, work the same across vendors.

The more open your data is in terms of format, connection, and storage at rest location, the simpler your transition to an AI-driven culture will be.

Next-Generation Databases

Relational databases store and represent data in the ubiquitous rows-and-column format and require a schema (built up front) that defines the attributes of data to be stored in each field. This dependence on a schema provides many benefits, such as consistency of data and fast access, but also has drawbacks: a schema is rigid and not only can limit the kind of data you can natively store and the ways you can analyze that data, but is notoriously difficult to change. Data scientists often look for ways to get rid of data, combine columns, and invent new data columns to represent data derived from other columns. Data stores that require schema work up front add long delays to this process.

Perhaps most importantly, relational databases ensure ACID transactions (guaranteeing the properties of atomicity, consistency, isolation, and durability). A transaction consists of several data items: for example, an account number, an amount, whether the amount is a deposit or a withdrawal, and an update to the account's balance. Transactional integrity means that everything happens, or nothing does: transactions happen in their entirety or not at all (this is atomicity). You can't have a situation where the amount is stored but the balance isn't updated, leaving the database in an inconsistent state—not even if the software crashes partway through the transaction. If that happens, the transaction is "rolled back." Transactional integrity guards against corrupt or inconsistent data, and is a key component of data integrity—the hallmark of the relational model. (The other characteristics of ACID compliance are outside the scope of this book, but they all work together to ensure trust in the data operation being performed by the database.)

Because data in an RDBMS is highly structured as it goes in, and its attributes are constrained by a well-defined schema, this makes its data easy to find and fast to get out. But this model can be very restrictive and inflexible when it comes to putting data in. Consider a field called "gender." In an older database, the schema might constrain values to be either male or female. But today you need other categories, such as non-binary. Accommodating this change requires a change to the schema, which is a labor-intensive process prone to introducing complicating side effects. For example, you may be able to change the schema, but what about legacy software that is built around the old schema? Retrofitting old database structures to meet new requirements is a complicated proposition; we've seen some schema changes take months to find their way to production. Sometimes, it's just not practical.

Next-generation databases purport to solve many of those problems—in particular, NoSQL-style databases are designed with the understanding that today's data is not nice, temporally correct, neat, structured, or static. The word "NoSQL" never represented any specific technology; what it really means is that there are architectural options for storing data that go beyond relational databases (NoSQL really means "not only SQL"). And many of these options are particularly appropriate for modern data flows. You may have real-time data streaming feeds with millions of records per second, a 12-hour doorbuster event hosted on your e-commerce cloud that drops 10 million transactions on you and then stops completely, and you want to capture new data points that relate to the campaign (including patron selfies during the event). Data doesn't always come evenly or with a plan. And it goes without saying that your organization needs to be able to consume that data in a performant way.

Relational databases were never meant to work with data silos, and they weren't originally designed to handle various types of data or to be flexible and scalable. Your AI-driven organization needs to include next-generation NoSQL-style databases, which offer performance improvements of up to 100× over previous-generation database products for the right applications. When you think about modernizing your IT infrastructure and information architecture for AI, you will need to consider new architectures coming about from a new preferred fabric, open data stores and formats, unified access to data through a single query engine and data virtualization, and new NoSQL-style performant databases.

We aren't suggesting that relational databases go away; not at all. We're saying that your business has different business domains that are best modeled and historically retained using different techniques. Think of a database's schema as the upfront work required to get the data out. In a traditional RDBMS, you do that work all up front; that's why it's so hard to change the schema when you need the data for a new application. In a NoSQL database you do the work to read the data (build the schema) when you want to get the data out; data is easy to put in and its "shape" is easy to change on the way out. You'll often hear us refer to RDBMSs as schema-first and NoSQL databases as schema-later (synonyms include schema-on-read or schema-on-need) or, depending on the use case, schema-never.

The Power of an AI Database

In this new era of data, traditional databases as systems of record no longer cut it. Today, expectations are higher. We expect intelligence on every level of a modern technology stack, from frontend user interfaces (with natural language and speech recognition capabilities) through the full range of applications, and all the way back to our data management layer. Databases must be smarter. They should understand what is being searched and find the most relevant information while using the most optimized way to locate the data. As organizations try to become more nimble, they want faster and simpler ways to do analytics across these hybrid data stores without the expensive and time-consuming efforts to copy, replicate, transform, or move data—they also must self-manage and self-heal, which reduces maintenance overhead.

In other words, what's needed is an AI-infused database.

Such databases offer built-in support for data science development. They support multiple programming languages and open source frameworks and formats (like JSON), enabling developers to more easily analyze and build machine learning models into applications. AI databases, such as Db2, make it easier for developers to write applications that require less management, are more resilient to outages, and help improve productivity.

Streaming Data

Data isn't just static. We've mentioned real-time data sources already—but how do you capture that data? Real-time data isn't data conveniently entered into a form; it arrives at high speed, asynchronously, and often without standards or meaningful structure. Think customer clickstreams; think Internet of Things; think telemedicine. You could capture a patient's every heartbeat if you wanted to, or vibration data from a machine on a shop floor!

If your data infrastructure needs to incorporate streaming data (and it probably does), you must plan for it. Tools like Apache Kafka and Apache Pulsar are designed to capture high-speed, low-latency data streams. They handle external events like a messaging system; they are fault-tolerant, highly scalable, and extremely fast. Their job is to make sure that the real-time data AI applications need isn't lost along the way. You may not need these tools, but as you modernize your data infrastructure, you do need to consider how to handle real-time data.

Get the Right Tools

No matter how you decide to modernize your data infrastructure, you will need to invest in enterprise-scale data preparation tools. In addition to unifying data access, the proper tools can cut the time practitioners spend preparing and cleaning data in half, if not more! Today's tools are capable of:

- Extracting data visually from a variety of sources and then profiling that data by searching it and grabbing samples of it to understand its quality and composition.
- Handling all facets of metadata—the properties that describe data—as well as tracking the sources of data and their lineage even after combinations and transformations are applied, creating a sort of catalog of both raw source data and transformed data along with a ledger of everything that happened to the data along the way. For example, IBM Spectrum Discover monitors storage repositories to learn about the data that is streaming in. Its services can detect numbers that look like credit card numbers and apply corporate-mandated masking policies, and so on. You will see more and more AI around metadata in the coming years.

- Integrating with a larger data platform that your organization has installed, so that the data that is read and created by these preparation tools follows the same governance and restriction policies that all other data in your enterprise must follow. This integration also allows for easier discovery of data sets across sources.

The Importance of Open Source Technologies

When researching technologies for AI, you may have noticed that many machine learning and AI products are based on open source tools. A modern infrastructure optimized for AI uses open source software for many key functions. This is not an accident; it is one of the reasons why AI continues to grow at such a fast pace today.

Let's explore why.

> There are many definitions of "open source," and sometimes heated debate over small distinctions; we're not going to go into that. We're going to use the common-sense definition of open source to mean code that is readable and shareable, not proprietary. Open source projects are developed and sustained by ad hoc groups of developers who may be located anywhere around the globe. Many open source projects are founded and funded by commercial companies, including IBM, and many commercial companies contribute code to existing open source projects.

Community Thinking and Culture

Open source software has a rich history in IT and at IBM. Watson is approximately 90% open source. We have long considered open source software a friend of IT and its community of contributors and experts to be second to none in terms of quality and breadth. The open source community fosters a sort of "maker" or "hacker" culture—not hacker in terms of a malicious actor, but one who tinkers and improves on a thing to make it work better or more efficiently. This type of continuous tinkering creates virtuous cycles of feedback that improve overall software quality as well as making possible new features and even opening up new frontiers in technology. Open source communities have been responsible for a deep, broad stream of innovation, especially around programming languages, tools, and AI frameworks.

Because open source projects are sustained by contributors who may be volunteers or may be employed by competing companies, a spirit of community is essential. Community-building tools such as meetups (which are also great places to find and recruit talent), conferences, and online forums help spread knowledge. The evolution of a project and its interfaces is defined by the community. Further, the community tends to come up with ideas for even more projects that are related to or enhance the value of existing projects. Successful open source projects, like Apache Spark and Linux, accrue a surrounding ecosystem of compatible or complementary products.

Another benefit of the community is that using open source software allows you to leverage a much larger development group than you could reasonably hire to develop a proprietary solution. Although some individual companies (including IBM) have made substantial contributions to open source, it's important to understand that open source projects are inherently distributed, and rarely tied to a specific organization—even when there is a commercial sponsor, or a company selling support or services. For example, Red Hat Linux could never exist if participation were limited to Red Hat employees and customers. Another example is the Hadoop project, which was started at Yahoo! and became the basis for several companies. It's at the center of a large ecosystem of open source data tools—but Yahoo! as an independent corporate entity no longer exists, nor do some of those startups. In the open source world, domain knowledge becomes decentralized, and the benefits of the community's collective talent, skill, and experience accrue over a wider base of code. Your organization reaps the benefit of that broad skill base.

High Code and Component Quality

Open source projects are developed out in the open today, typically on shared version and source control services like GitHub. This means that the source code—the very bowels of the program—is available for you and your development team to scour and understand, unlike with a proprietary program where you are provided with a precompiled executable that you simply run.

Having source code available in the open means that it's not a black box, and developers can read, understand, expand, or modify it as necessary. There's no need for you to go out and invent a new solution if an open source one already exists (and you have the option to "fork" the code as well if you need to take the project in a different direction). This is particularly important in the age of AI, where explainability and traceability are key to removing bias, ensuring accountability, and establishing trust. Open source code is also arguably more secure, because many more developers have the chance to review the code and find security holes or other potential vulnerabilities. As the saying goes, "With enough eyes, all bugs are shallow"—and with open source, there are many more eyes on the code than even the largest corporation could possibly afford.

Successful open source projects offer supporting materials and communities with vibrant forums that can answer questions, provide help, and the like. They allow people to quickly get up to speed, which reduces the costs of "onboarding" new team members.

Real Examples of Modernizing IT Infrastructure

Modernizing your information infrastructure isn't for the faint of heart. Unifying access to data that sprawls across many databases and clouds is a big undertaking. Don't let that stop you from starting AI experiments; but do realize that, in the long run, it's a problem you'll need to solve to take full advantage of AI.

Siemens and Fannie Mae are two organizations that have tamed the dragon of modernizing IT infrastructure—let's see how.

Example: Siemens Looks to the Cloud to Unify Its Data Processes

Siemens, the 170-year-old global technology behemoth, must keep vigilant about cybercrime. Indeed, the mandate of its Cyber Defense Center is to protect Siemens and its customers from viruses, malware, intellectual property theft, and other forms of malicious attacks. Because the level of threat is more than any human could deal with—Siemens won't reveal the intensity of the threats, but it handles many thousands of attacks per second—Siemens used Amazon Web Services (AWS) to build a next-generation data analytics platform to address the problem.

Its goal was to use cloud-based AI to process the huge amounts of data, and make immediate decisions about how to counter any detected threats. Siemens uses Amazon SageMaker to label and prepare data, choose and train machine learning algorithms, make predictions, and act. The solution also uses AWS Glue, a fully managed ETL service, and AWS Lambda, a serverless service that runs code in response to events.

With a data lake based on Amazon S3 that can collect as much as 6 TB of log data per day, Siemens' security staff can perform forensic analysis on years' worth of data without compromising the performance or availability of its security incident and event management (SIEM) solution. The serverless AWS cyberthreat analytics platform handles 60,000 potentially critical events per second but is managed by a team of fewer than a dozen people.

Example: Fannie Mae Transforms with a Governed and Centralized Data Environment

Fannie Mae is a leading financier for homebuyers and renters across the United States. Founded in 1938, the $10 billion organization deals with the same challenges as any corporation, including being saddled with a broad range of legacy environments and data silos. But as the company became increasingly data-focused, it needed to transform those legacy silos into a more responsive, cloud-based data lake, in order to create a modern data environment capable of ensuring that the right data got to the right person at the right time.

Fannie Mae first established a governance standard mandating that every data set and field in its data lake was documented. Each data set had to undergo a design process where it got curated and assigned a unique identifier that stayed with it regardless of how often or where it was copied to (in the same way an individual's US Social Security number is the same regardless of their employment status). Each data set also had a long list of properties that had to be completed before the identifier could be issued.

While this made the data more accessible, it was too slow a process. It could take months from the time the design was approved to when the data was actually transferred into the data lake. In the meantime, multiple Fannie Mae apps each continued to generate more than 10 million new files every day—files that would ultimately need to be integrated into the data lake as well.

In a world of increasingly data-driven competitors, Fannie Mae needed to either transform itself into a firm with an agile and well-governed data environment, or risk being disrupted by other organizations. So, the data team at Fannie Mae began its own journey: a complete transformation that would create a modern data infrastructure for the organization. The first step was to ask for—and receive—sponsorship by the CEO (board sponsorship is critical for any governance program). The second step was to build a centralized data operation where the team could manage data to a high level of quality, consistency, and timeliness across the enterprise.

Imitating the Development Operations (DevOps) framework, the team is using a "DataOps" mentality to reduce data preparation cycle time. Fannie Mae's new data platform is thus now holding data to the same level of criticality as developing mission-critical applications.

Today, Fannie Mae has a modern data environment characterized by near-real-time updates, which results in a richer and more granular real-time customer experience. Data is now viewed as a business asset, and the firm is encouraging experimentation at scale with data.

Don't Neglect the Foundation!

In this chapter, we set the stage for how a modern information architecture acts as the foundation for the AI Ladder. Regardless of how you approach your journey to AI, you will eventually need to rethink your company's data structure. This doesn't mean that you can't have some successes first—but it does mean that an antiquated foundation will eventually get in the way. You don't want to build a state of the art, modern office on a leaky 18th-century stone foundation. So, remember:

- You have legacy systems. You don't have to rip them out; you probably can't. But you can (and must) make the data in those systems accessible throughout the organization, whether through data virtualization or some other technology.

- If you're like most companies, you've made a partial migration to "the cloud." In practice, this means migration to many clouds, scattered across several providers. Again, you need to unify access to data (notice we did not say cloud providers) no matter where it resides, so that you can manage and govern it appropriately. After all, moving to the cloud doesn't eliminate silos. You can have silos in the cloud just as you can in your on-premises infrastructure; silos ultimately arise from the organization, not the technology.
- You have IT staff that are responsible for deploying and maintaining business-critical software. Ensure that they're up to date on modern practices like continuous integration and deployment (CI/CD), infrastructure as code, and monitoring (collectively, these practices are often grouped under the name "DevOps"); creating a DevOps mentality and culture will put you in great shape for moving your AI projects into production.

Yes, the AI Ladder has a fourth rung—infusing AI throughout the organization. And if there's only one thing to take away from this chapter, it's that you won't succeed at infusing AI if you don't have a solid foundation to build upon. Back to the ladder analogy we started this chapter with: say you're at home doing some roof maintenance and need to climb high. Contrast climbing up to your roof on a creaky wooden ladder with no one around to having someone you trust hold a sturdy aluminum ladder in place for you; in the latter scenario you climb higher, faster, and with more confidence. Like we said, your AI needs IA.

Don't neglect the foundation!

CHAPTER 6

Collect Your Data

In the preceding chapter we talked about the process of modernizing your entire data infrastructure to make it one integrated, efficient platform. A consistently audited, asset-optimized, and instrumented-in-IT infrastructure is essential for flexible and cost-efficient operation in the AI-centric world.

You may have noticed that there were a few important things we didn't talk about in that chapter. For example, we haven't talked about how you acquire data. Nor have we talked about the quality of the data that resides in that platform, or how to make it available to the AI processes and programs that may require it or benefit from it. We certainly haven't talked about how to get the data you already have under control. Many companies think "Of course we have data," only to find that there are many reasons why this data isn't really accessible. Those reasons may be technical, political, regulatory, or some combination of the three—but they're real. We always tell people that big data without analytics is…well…just a bunch of data. Never forget: data may be an asset, but it's a valueless asset if you can't use it.

In this chapter we will talk about getting access to all relevant data and evaluating its utility. We'll look at ways to consolidate data sources, because at far too many companies data resides in departmental silos that prevent it from being used effectively. We'll discuss the pros and cons of combining data sources into "data lakes" and how your team can determine whether building (or maintaining) a data lake is worth doing, given your needs. We'll talk about how data lakes frequently turn into data swamps, and alternatives to data lakes.

If you have lots of data but can't do anything with it, you're information-rich and insight-poor. After all, what's the value of data you can't use? If this feels like you, take heart: the Collect rung of the AI Ladder (Figure 6-1) is where we start to change that. Let's get started!

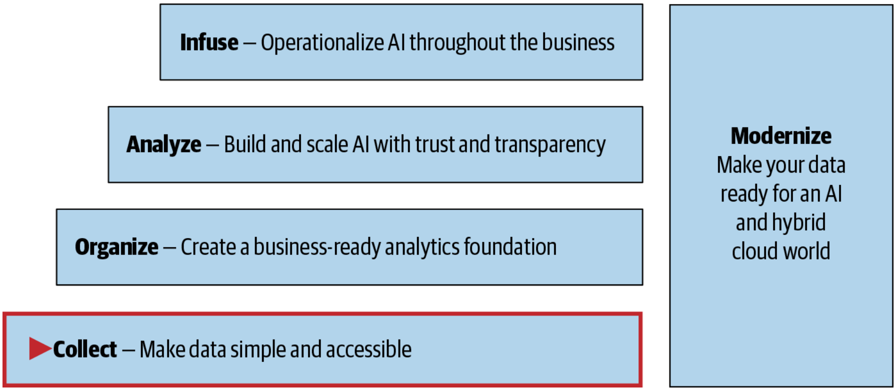

Infuse – Operationalize AI throughout the business
Analyze – Build and scale AI with trust and transparency
Organize – Create a business-ready analytics foundation
▶Collect – Make data simple and accessible
Modernize
Make your data ready for an AI and hybrid cloud world

Figure 6-1. The Collect rung of the AI Ladder

What Needs to Happen on the Collect Rung

The Collect rung is all about making data simple and accessible. That means collecting data of all different types and sources, whether it be from edge devices and IoT sensors, bank statements, traffic cameras, or anything else, regardless of where it lives, and making this data accessible throughout the organization. As businesses evolve, their data continues to build up and take on different forms—but for too long now, data has been held captive. To embrace and start doing AI, you need to be able to access your most valuable commodity, which is, you guessed it…data.

The data your organization collects—both data streaming in from ongoing transactions and data that it already has stored—needs to be available to any group in the organization, not trapped ("held captive") in corporate silos. It doesn't matter whether those silos are political or technical; incompatible data formats and data stores can be just as big a barrier as departmental politics. If different departments use different databases, and those databases can't interoperate (we'll say it over and over again in this book: organizational silos become data silos), you're looking at the worst kind of vendor lock-in—your own! You can't just get one part of the organization to change; you have to work with all of them.

We've even heard of cases where departments (particularly in medicine) hold on to their data by making it accessible only by fax. You can ask for it, and you can get it, but you won't get database records; you'll get a big pile of paper spilling out of your fax machine (if you still have one). There are many reasons for a response like that. Particularly in medicine, even in these days of electronic medical (or health) records (EMRs or EHRs), a surprising number of records are still on paper; staff may be overworked and not have the time to look up an appropriate answer, or it may just be a passive-aggressive response ("It's my data and I don't want you to have it, so I'll send it in the least useful form possible"—we've seen it and heard it all). Regardless of the cause, that's what we're fighting against. It doesn't matter whether the problem is a personal grudge or a deep-seated technical incompatibility between legacy databases. If everyone's going to be on one team, everyone needs access to the data.

Let's quickly consider an example. Nedbank, one of the largest banks in South Africa, has a wide range of banking services, ranging from corporate and investment banking to retail and business banking, insurance, and asset management. It's clear Nedbank is working with a lot of data. The firm needed a single holistic view of all its data assets, as well as the ability to analyze those assets in context. Nedbank turned to IBM, and data virtualization in particular, to craft a solution that would help the bank consolidate all of its assets and provide a unified view. Not only did Nedbank make this big-picture view accessible to its analysts and users, but it essentially built the data architecture it needed to expand into larger projects involving data science and machine learning.

That's why we say there's no AI without IA: you aren't going to make progress unless you create an architecture, a structure that makes all your data accessible regardless of where it's coming from, whether it's real-time transactional data or data stored in departmental databases. Furthermore, data users need uniform access to that data, regardless of its source or type. They shouldn't be responsible for solving data integration problems. If one department stores numbers as integers and another stores them as text strings, converting between the different formats shouldn't be the data user's problem. In reality, you'll see much more complex examples of conflicting, and even incompatible, data formats; your organization's information architecture needs to solve those problems.

And finally, access to data and insights derived from the data must be available across the entire organization (subject, of course, to regulatory constraints). Insights aren't very useful if they are limited to a privileged few. Too many organizations are hampered by the inability to share data, or the insights derived from the data. For example, understanding risk in one part of the business can help in underwriting it in another part. It's important to realize that everyone is on the same team and are fully aligned to the same goals, and that combining data sets creates more value than using the data sets individually. With data and AI, the whole is always greater than the sum of its parts. And if data is a crucial asset, then it's your responsibility to maximize the value of that asset—not to limit it by letting data remain in silos.

Example: EMC Develops a Data Collection Strategy

EMC, a major market player in storage and analytics technologies, wanted to develop a more targeted marketing outreach approach so it could identify new business opportunities as soon as they arose. It had all the data it needed to perform this type of analysis, but lacked a holistic, strategic approach to data collection and management that would allow it to extract insights from the data quickly.

Over the past decade, EMC has acquired more than 80 companies. This has quite naturally resulted in a fragmented data infrastructure with dozens of different data silos, and a rise of "shadow IT" operations with little to no data governance. Technically, IT was acting as the "gatekeeper" when business users needed to run a report using any of this data. This considerably slowed the company's ability to derive value from its data assets.

To address this problem, EMC combined all its data into a single data lake and imposed governance on it to enable sharing among various business groups. The improvement was astounding. The time to get an answer to a data query was reduced from 4 hours to less than a minute. The firm also achieved 80% accuracy in predicting customer behavior and was able to move to near-real-time decision making for marketing programs. And by creating analytics "sandboxes" for business units, EMC was able to offer business users self-serve access to a rich array of comprehensive data sources for analysis by machine learning algorithms and AI, dramatically reducing dependence on IT.

Start with a Data Census: Learn What's Out There

One of the first steps in collecting data is determining what data you need to collect. Remember that multiple linked sources of data are always more powerful than individual data sets. Think carefully and creatively about what's available to you—you will probably find that more data is available than meets the eye:

- Start with the data you already have: your company's internal data. Track down what you have and where it lives, department by department. (Think about it: what's the point in collecting more data when you're not using the data you already have?) Most companies have many internal databases that don't communicate with each other. Find these, and take notes on where and how the data is stored. You're likely to find many competing vendors; some data probably lives in the cloud already, and some may still be handwritten. (As mentioned earlier, handwritten medical records such as doctor notes are still quite common. If you're working with local governments, you're also likely to find critical data in filing cabinets.) Think creatively. Have you ever heard, "This call may be monitored for quality or training purposes"? If your company has a call center, these recordings are invaluable for training an AI application or model to do customer service. The task of "database archaeology" may be a bigger project than you expect, but it's a key to success.
- What public data sources are available to you? You can look at weather data (one of the most underappreciated data sources around), real estate records, tax records, statistics on crime, trash collection, tourism—almost anything. A lot of data is available as a matter of public record. Getting access to it may be hard (although that's changing more and more), and getting it into usable form may be

harder, but it's there. For example, a number of years ago we worked on a project that involved New York City taxi data, and while the data was freely available, to get a copy of it we had to give a municipal department sealed, never-before-opened drives. Some cities (such as San Francisco) openly publish data on crime, trash collection, and more, and make it easily available. Others don't. Consider another example: one of us once wanted to look up a small town's final election tally (not just who won). They weren't in the newspaper. It wasn't on the town's website. They were finally found on the state's registrar of voters website, which had a PDF of a scanned, typewritten document from the town hall. It appeared to have been typed on a manual typewriter, and we bet they had to walk down the street to the library to scan it. Data collection isn't for the faint of heart.

- What's available in social media? Social media consists of vast amounts of data, and much of it may be relevant to your applications. But be careful: there are legal restrictions on how social data can be used, and regardless of regulation, users are increasingly wary of how their data is used. Take care not to create a PR disaster by being creepy, but do realize that social data can be an invaluable asset.
- What private data is available? We keep saying that data is an asset, so it shouldn't surprise anyone that there are companies buying and selling those assets. You may be able to buy the data you need.

> Don't neglect to record what this data is, where it resides, and how it's structured. If the data is unstructured, what does it represent? A database of job listings might have an "unstructured" component—a set of text objects collected from job postings, with no internal form—but you at least need to know that they're job listings, and possibly what kinds of job listings. (We call this semistructured data, which more accurately describes most data.) Knowing what you have is the first step toward organizing your data and building a data catalog, which we'll discuss in more detail in the next chapter.

But you don't always know where your most valuable data might come from—or where you might need to go to acquire it. A boiler company needed to reduce the time it took to make repairs at customers' homes. In the midst of an especially cold winter, backed-up demand for service meant that it was taking an average of three days to get engineers onsite to fix older models that didn't have electronic indicators. This backup naturally resulted in a lot of angry customers during frigid weather. Because it didn't have access to data from the boilers themselves, the company turned to public streams of information—weather patterns and forecasts—as well as transactional and demographic client data it had collected over the years, and combined that with social media data. By using AI and machine learning to analyze all this data, the company was able to predict with 98% accuracy when and where its customers' boilers were most likely to break down. The result? Better (and preemptive) service, satisfied customers, and lower expenses, all by finding and using data intelligently.

Understand Data in a Business Context, and Partner with SMEs

It's important to understand that you're acquiring and using data in a context that is specific to your organization's goals. Any organization that has specific goals is a business. If you're a manufacturer, you need to look at the data in the context of supply chains, labor, and so forth. If you're doing biological research, you need to look at data in the context of scientific research, and biology in particular. And if you're a charity that promotes social welfare, it's important to look at your data in the context of the changes you want to bring about. For our purposes, these are all different kinds of businesses, with different kinds of goals, operating in different settings. As we talk about locating and getting access to your organization's data assets for use in AI applications, remember that data only has value in terms of its *business use*. The businesses, their goals, and the ways they use data will vary. (Many regulations, such as the GDPR, actually require that the data you collect has to be used for the purpose it was intended for—this is called "purpose limitation.")

Regardless of what kind of business you're building, the AI Ladder is the same; it isn't tied to any one context or set of goals.

It's easy to think of data in the abstract, but data is always associated with a context. To get the most out of data, you need to be in touch with subject matter experts (SMEs) who are familiar with the context. SMEs are at least as important to the "collection" process as data scientists; they will be able to tell you a lot about what data is available, what that data really means, what's missing, and how it can be used. They will also be able to tell you what the data can show you, and what kinds of results you should be looking for. That advice will be important as you start building AI projects.

Think of your organization's data as a library of millions of books written in your company's language. An AI application may be capable of finding unexpected patterns in those books, but that doesn't mean it can tell you what patterns are *important* or *worth focusing on*. Your team needs people who understand the context in which data was recorded, and for what purpose. These people—the SMEs—can "speak the language" of that data from a business context. All companies, no matter the industry, have business SMEs that can accelerate the AI journey—you just need to invite them to join you.

Consider an insurance company settling an auto accident claim. They should bring in an adjuster, likely with 30 years of experience and who's seen literally thousands of car accidents. This person can look at a photo and accurately estimate the cost to fix the damage without surveying the car! (If the damage is to a specific area, they might note that the car needs to be examined.) This kind of institutional knowledge is critical to AI. That's why SMEs need to be involved in your company's AI transformation, to label data (taking the work out of a data scientist's hands so that they can do data science work) and more. Finally, consider this: for the first time in history, we're seeing a workplace where there are five generations of workers. We are at the brink of a massive institutional knowledge leak as our economy's oldest workers move into retirement (or retirement jobs). That is a risk for any business—AI is about capturing that knowledge.

Getting Beyond Transactional Data

If a retail company only has access to its historical, transactional data, that means its AI models are being trained on a single data source. There's a lot you can do with that data—but think about how a modern online retailer works. What if that retail company, in addition to its structured transactional data, could also take advantage of data from social media, weather data, and real-time clickstreams? Now, this retail company can build AI models that will tell it what its customers bought last week, how they feel about it, what they're shopping for at this very second, and how they're shopping for these things. The company thus acquires a three-dimensional view of its business.

This multi-dimensional view isn't without cost. Clickstream data and social media data can easily dwarf the volume of traditional data, and that can quickly become an infrastructure problem. If you decide to store the data on premises, you'll need to invest heavily in storage and backup capabilities. (Unfortunately, many businesses skimp on backup, to their own peril. Backup is not just about recovery from incidents; it's about protection from human error and, even more importantly, adding a strong element of cyber resilience by air-gapping recovery protocols from malware attacks that can go six months or more before they are discovered.) Storing this data in the cloud is another option; cloud providers are capable of dealing with petabyte and even exabyte data sets without trouble (though depending on your business, there may be regulatory restrictions on where you store your data).

There's also reputational cost to consider. Customers are used to being tracked while they're online, so using their social and clickstream data is neither new or unexpected, but they're increasingly unhappy about it. (And no, nobody reads those 6-point-font policies that tell you how your data could be used.) What will your customers accept? At what point will they consider your data collection "creepy"? And what, if anything, will they do about it? Those are questions only you, and your subject matter experts, can answer.

The Challenges of Collecting New Sources of High-Volume Unstructured Data

Managing data that arrives in real time requires new kinds of tools, such as Apache Kafka, which facilitates streaming data into databases. And unstructured data frequently requires new kinds of databases. As we saw in the previous chapter, RDBMSs can only store data effectively if that data's structure is specified in a schema. If your incoming data is unstructured (or its shape is unknown), there's no schema to work with. This is where NoSQL databases come in. The term NoSQL is starting to become out of favor, but it's important to realize that the technologies referenced by this term are not; they provide other ways to manage data—document databases, graph databases, column stores, key value stores—and those tools will more than likely become part of your information agenda. Many of them are optimized for high-speed write access, which is exactly what you need for real-time data.

Getting access to all kinds of data means being "polyglot," not in terms of programming language, but in terms of data storage and models. One of the toughest tasks on the first rung of the AI Ladder is designing uniform ways for everyone across the company to access the data. SQL has become a common language across all kinds of databases, including many NoSQL databases. Build web services that allow online access to databases from any point in the organization, using technology that's already available at everyone's desk (or, for that matter, a lunchroom, conference center, or airliner). And cloud providers can create robust data collections that can be accessed worldwide and are resistant (though not immune) to outages.

The technical aspects of data access are solvable. What about the organizational aspects?

Organizational Aspects of Data Access

Consider the situation where the Sales division of a company maintains one database containing information about customers while the Customer Service division maintains a separate database, and where each division guards its information and it's not easy to share it with anyone else. This situation is common—if anything, it's the rule rather than the exception. Often there are perfectly good reasons for separate databases, and potentially for restrictions on disseminating the data they contain. Everyone has horror stories about what happened when the "wrong people" accessed their database. But just as often, there isn't any reason for "dueling databases" other than tradition, territoriality, or institutional sclerosis. Whatever the underlying cause, dueling databases is a problem you're going to have to resolve if you want to have a truly AI-ready infrastructure. To be truly useful, machine learning models are going to require access to all relevant data.

If your project only has access to some of the data, or if data from different sources is contradictory, it will be impossible to build reliable models or to gain true analytical insight. The challenges become worse as businesses evolve, because the volume of data being generated continues to grow at an accelerating rate (this is data velocity). All of this limits the possibilities of AI to transform an organization, because insights are only as good as the data.

In many organizations that employ some variant of the "data mart" concept, individual groups within the organization are considered "owners" of that mart—everything from soup to nuts: hardware, software, and data. This "ownership" idea can foster a proprietary attitude and encourages the creation of arbitrary restrictions on people outside the group. This siloed model is still common, but it's no longer effective. And it hints at fear and corporate dysfunction: if someone sees my data, they'll have better insight than me; if someone sees my data, they may want to change the way I do business. Those fears are common, but they're not healthy.

In the words of Daniel Hernandez, IBM's VP of Data and AI, who has worked solutions with upward of one hundred customers: "The reality is that people who 'own' the data are being circumvented all the time. That is, in the absence of a workable information architecture, people do what they think they have to do in order to get their jobs done. This means that data state and origins are no longer known, therefore its integrity cannot be trusted. Most of the problems caused by this way of operating are invisible." In other words, if data isn't accessible through regular and appropriate channels, people will find back doors; they'll do what they need to do. But when people get data through a back door, nobody knows what they'll find. They may not know whether the data has been cleaned and verified; they might not even know what format it's in, and they may spend lots of time reverse-engineering the data's intuitions—possibly misunderstanding it in the process. They certainly won't know what regulatory restrictions apply to the data's use. Dan concludes, "This is a crisis, man." Indeed it is.

Example: Procter & Gamble Avoid Data Silos Using a Central Data Warehouse

Procter & Gamble (P&G) has invested significant amounts of money over an extended period of time to ensure it doesn't have data silos among its many brands, which must speak to tens of millions of consumers every day. This dedication to a "single source of truth" has involved establishing global standards for data type and quality, and storing the data across all product lines and regions in a standard format in a central data warehouse.

P&G's Global Business Services (GBS) organization has developed tools, systems, and processes to provide managers throughout P&G with direct access to up-to-date data and advanced analytics. But because data *context* is so important, P&G GBS has also embedded data analysts—data SMEs—within each of its business units to work alongside leaders and managers (the business SMEs) in driving real-time insight-based decision making.

Over time, P&G's managers and business users have found that aggregating and comparing data across product lines and regions has become significantly less complicated.

The company's IT organization also automated the creation of reports that could be used across multiple business units. This not only made it simpler and easier to retrieve and analyze data, but also standardized the way data was visualized across the global operations. Establishing a visual "language" for data enabled analysts and managers from one business unit to work effectively with their counterparts from other product divisions or regions.

Example: eBay Eliminates Data Silos by Publishing Business Processes as APIs

According to the 2019 Connectivity Benchmark Report, disconnected data silos are creating business challenges for more than 83% of companies. eBay, surprisingly, was one of them. eBay's mission is to facilitate a global online marketplace where people can trade virtually anything. The company has about 180 million active users and employs 10,000 workers around the world.

eBay's entire business model depends on transactions. Every day, it must support communication between buyers and sellers, customer support, third-party fulfillment centers, and shipping services. All its data needs to be accessible and easily analyzed using machine learning and AI algorithms.

But eBay's data was stored using different formats and in different locations, and not easily accessible. IT had built one-off integrations between applications, but this custom code created tightly coupled connections between components, resulting in redundant work and making it difficult to replace or introduce new infrastructure components.

To eliminate the data silos and allow it to continue to innovate without a complete ripping and replacing of its data architecture, eBay used MuleSoft's Anypoint platform to package and publish business processes as APIs that business users could leverage without depending on IT.

In the first stage of the project, eBay's IT team pulled data from systems like Workday, Concur, and Sailpoint to manage the data requests from business users. It then reused the APIs from the first project to further enable line of business teams, enabling self-service and a move to the cloud. As it continues to refine this project, eBay will unlock more and more data for analysis.

Ownership, Stewardship, Regulatory Compliance, and Discipline

We've said that all of your organization's data must be available to your AI programs. When we say that data silos must be abolished, and that arbitrary restrictions must be relaxed, that does not mean that it's now a field day and all restrictions are gone. It is vitally important, not only for regulatory compliance but also to ensure the integrity of the data itself, that access to data be properly controlled. Rules about privacy must be strictly observed, and you must be able to guarantee that the data has not been improperly used or altered—the viability of the business and its reputation depend on it.

So this is a paradox: data must be made more freely available, but at the same time it must be more rigorously controlled.

Some of the technical aspects of resolving this paradox are covered in the next chapter, which includes topics like data lineage, provenance, and governance. Here we are stressing the organizational solution to this challenge: you must ensure that every part of your organization works cooperatively to implement corporate-wide policies of data integrity and privacy. And the groundwork that enables you to create and implement these policies starts when you're collecting data. We again emphasize that the widespread adoption of AI is going to be disruptive. And that means that a big part of the job of management—from mid-level to the very top—is going to be managing that change, and fostering the cooperation among groups that will make this transition possible.

From the top down, the creation of a corporation-wide information architecture policy must be a top priority. Your policy needs to take into account regulations that cover data use (HIPAA, the GDPR, California's CCPA, and others), ethical standards, and your company's values. How do you want to treat your customers? Do you value their trust? What are the consequences of violating that trust, whether by misusing their data or losing control of it in a cyberattack? SMEs should know about regulations that apply to their specializations.

While you're on the Collect rung of the AI Ladder, start collecting the metadata that will make compliance and other issues easier. As you're discovering what data is available, both within your organization and outside it, keep track of where the data comes from. Make sure you understand the rules under which the data can be used. If there's any material that describes the data set, keep track of that. The paper "Datasheets for Datasets" by Timnit Gebru et al. makes proposals for information that should be collected to describe a data set. These ideas apply to any data source, whether public, proprietary, or private. Collect all this information now, and you won't have to spend time looking for it later.

Example: Owens-Illinois

Owens-Illinois, a large manufacturing company, was undergoing a digital transformation with the goal of reducing the cost of serving its customers and its total cost of ownership (TCO) and IT expenses. The cost of its legacy database management software had become unsustainable. The company was generating petabytes of data, ranging from product performance metrics to metrics covering revenue and profitability. The relational database it was using was no longer suited to the task, and trying to make it do something it wasn't designed to do was a waste of money.

So Owens-Illinois decided to migrate its database systems to a more flexible and cost-effective multicloud platform. The company was able to move operations to a cloud provider where appropriate, keeping some components in-house as needed. By using a hybrid data management system to simplify data collection and access, it was able to lower storage usage costs while improving performance and transaction response times. Owens-Illinois now not only has a reliable database that can support its data volume but, more importantly, has benefited from the cost reductions it expected. These reductions came from many areas, including database, platform, and managed services costs.

With this firm data foundation, the company's IT department can produce business reports its clients can actually use, as opposed to somebody pulling the data out of the system and then trying to figure out how to manipulate it for a valued report. This foundation also sets the groundwork for the company to look ahead to AI and machine learning solutions to continue on its path of digital transformation.

Collecting Data: You Can Win This Battle!

This rung of the AI Ladder is about collecting data, including discovering the data your company has and making that data accessible to everyone who needs to use it. That means breaking down corporate silos, understanding regulations and policies surrounding data use, and controlling access to data to remain in compliance with applicable regulations and policies.

Yes, there's a lot to be done. You have a head start if you already have well-established data practices, but it's definitely possible to build a modern data collection process and make data accessible even in heavily siloed and regulated organizations. And the rewards will be considerable: you're charting your company's direction in the 21st century, and preparing to take advantage of the data that you've had all along but never used effectively. Remember that there's no AI without IA; AI starts with building an information architecture. And remember that, with data, the whole is always more than the sum of the parts. The more data sets you have access to—internal, external, public, private—the better your insights will ultimately be.

Once you've done battle with the corporate silos and defined your data policies, you're in control of your own destiny. Now it's time to discuss data organization.

CHAPTER 7
Organize Your Data

In Chapter 5 we discussed the use of hybrid multicloud architectures, composed of private clouds, public clouds, and on-premises systems, to host databases of all kinds. We looked at ways to combine data sources into "data lakes" and explored how to determine whether that's worth doing. Modernizing entails building an infrastructure that allows you the flexibility to choose the right platform to host each data source or application.

In Chapter 6 we detailed the Collect rung, a rung dedicated to removing barriers to accessing your data and evaluating its utility. We looked at sources of data that you can use to augment your own. We also discussed getting a head start on building data catalogs (a prelude to this chapter): recording where your data comes from and uncovering exactly what that data means.

Now it's time to look at how to improve the processes that track, protect, catalog, characterize, and govern data, organizing it to make sure that it's suitable for use in AI applications (Figure 7-1).

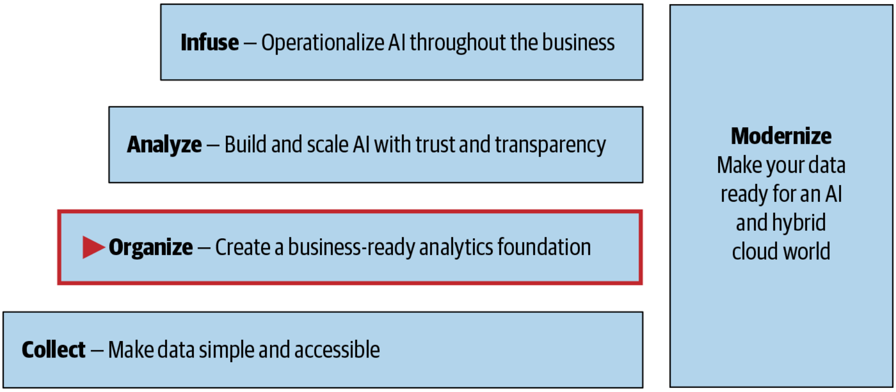

Infuse – Operationalize AI throughout the business
Analyze – Build and scale AI with trust and transparency
▶ Organize – Create a business-ready analytics foundation
Collect – Make data simple and accessible
Modernize
Make your data ready for an AI and hybrid cloud world

Figure 7-1. The Organize rung of the AI Ladder

This is important because the standards for data used in AI applications are higher than most organizations are used to. There are two reasons for these higher standards: poor data leads to poor AI, and regulation demands quality data.

Poor Data Leads to Poor AI

While inconsistent, incomplete, and otherwise "low-quality" data may still provide value when analyzed by traditional methods, it simply won't work as input to machine learning or deep learning models. The old maxim "garbage in, garbage out" is especially true for AI. Poor-quality data results in useless models or no models at all. It's very important to understand why.

If you're just computing an average—say, the average amount that a given customer spends in a year—and some of the numbers are off, that isn't a big deal. A $10,000 bill for delivering cement entered incorrectly as $20,000? Averaged over all your customers, that's just not a big deal; if it only happens once, it's not going to make or break a large company's numbers for the year. But when you build a model, you're not just computing a single average. You're building intelligence that is going to be used to make predictions, or optimize your performance, many times. Any errors (or bias for that matter) in the data used to create the model are going to be exacerbated when you use that model in an automated system—say, to create quotes for future deliveries. If that error is repeated thousands of times, it can become devastating (now think about this scenario in healthcare).

It's worth dwelling on this point, because people are increasingly concerned with AI that's fair and unbiased. While there's a lot of discussion of whether algorithms are fair or unfair, algorithms aren't really the point. A neural network with 600 nodes and 7 hidden layers doesn't care a bit whether it's charging the right amount for cement, or (for that matter) whether a customer is Black, White, Hispanic, or Asian. Truth be told, the algorithm doesn't even know this, just like it doesn't really know what a cat is; it's just doing a lot of matrix multiplication and building weights. All it knows are numbers, even when working with text. If the data on which that model is trained is incorrect, biased, or unfair, those effects tend to be magnified. If nothing else, they're repeated. A few "small" errors in your input data, and you may find that you're consistently undercharging (or overcharging) on every transaction. Or you may find out that you're violating some sort of law. In many cases, the problem with your data might not be a few errors in data entry, but systematic bias in the way the data is generated. And these problems of systematic bias, compounded by the training process and built into a model that's used to make decisions in bulk, are the problems that lead to lawsuits and public relations disasters.

Regulation Demands Quality Data

The second reason for higher standards is that laws and regulations demand them. Organizing data means not only eliminating redundancies and contradictions and working around bad or missing data, but also taking the steps necessary to see that provisions have been made for ensuring privacy and security, and maintaining metadata to guarantee that you can trace the evolution of your data (data lineage). Good business practice, not to mention laws and regulatory norms, requires great diligence.

Let's think about what tracking data provenance means. Like everything else in life, data evolves over time. Over time, new data is added, data is imported from other databases, new fields are created or inferred and added to the database, and erroneous items are corrected. First, you need to know where the data comes from. What's its source? How was the data collected? And what are the terms under which it can (or cannot) be used? That's data provenance. Just as provenance in the art world means understanding the history of a piece of artwork to verify that it's genuine, *data provenance* means understanding the history and origins of a data set, so you understand what it contains. (In Chapter 6, we discussed partnering with SMEs to understand your data.) This is all metadata that needs to be collected and maintained along with the data itself. You can't trust data that you just "find" and don't know anything about. You wouldn't want a fake Rembrandt; why would you trust data from an unknown and possibly unreliable source?

Second, you need to know what happened to the data after it was collected and before it was used. That's *data lineage*. How was the data modified? What happened to it during the transformation and preparation process (for example, how the data was shaped or aggregated, or from what inputs the new data type was feature-engineered)? Does the data set combine data from multiple other databases? Does it contain data that is the result of some computation on the original data? Who touched the data? When was the data touched? Who is responsible for this data's stewardship? Data lineage is all about collecting and maintaining that metadata.

Tracking data lineage is difficult, and there aren't many tools (or companies) that can do this well. The marketplace has a great set of tools for tracking the lineage of source code, including GitHub and its many antecedents, but few tools are available for tracking the evolution of data sets.

Many of the tasks described in the previous two chapters can be performed by people with traditional IT backgrounds: networking, DevOps, database administration, business analytics, and the like. On the Organize rung of the AI Ladder your data scientists will be getting more involved, because this is the place where data gets scrubbed, labeled, and in general made suitable for use in an AI context.

What Needs to Happen on the Organize Rung

We've said a lot about how important it is to ensure that your data is accurate and to understand its lineage. That all has to be done at scale.

On this rung of the AI Ladder, there are three key issues your organization must consider:

- Making data "business-ready"—clean, complete, and compliant
- Documenting and cataloging data sources
- Ensuring that appropriate data governance is in place

When data is not business-ready, finding, understanding, and putting it to productive use is a constant challenge for everyone, including data scientists, analysts, and line of business users. Your data also needs to be trustworthy—if it isn't, there's no way your results will be trustworthy, and nobody wants to invest in developing AI systems that can't give trusted results. Building a single authoritative source of truth can be a huge help. That source might be a data lake. If constructed carefully, a data lake can go a long way toward unifying data access, resolving discrepancies between different data sources, and solving many other data problems.

That doesn't mean you should build a data lake if you don't already have one. There are other technologies, including data virtualization (we touched on this earlier and discuss it further later in this chapter) that can help you address data quality issues.

Cleaning Data

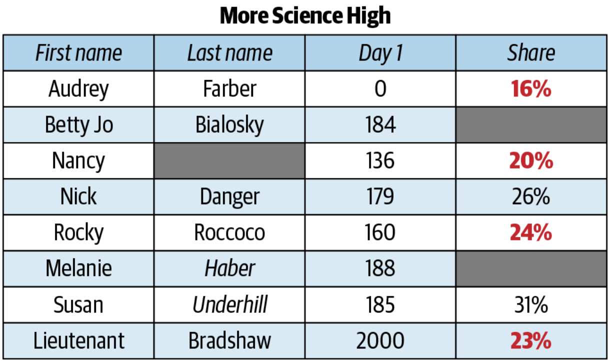
You need to address data quality and determine whether your data is business ready. This involves cleaning the data and ensuring it is "business ready": complete and compliant with all applicable regulations and policies.

By "cleaning," we mean removing the kinds of inconsistencies and inaccuracies we mentioned earlier. Most data scientists agree that 80% of their work is data cleaning and preparation. In Forrester's "Predictions 2019: Artificial Intelligence" report, 60% of decision makers at firms adopting AI cited data quality as their number one challenge. Ensuring the quality of data is laborious, time-consuming, and certainly not as much fun as experimenting with different kinds of models. Is data cleaning a bottleneck, a problem for business productivity? That's the wrong way to look at it. Thinking that you'd be more profitable or competitive if your data scientists didn't waste all that time cleaning data is like thinking you can ship products to the field without testing them. That might work—until it doesn't. Cleaning your data is absolutely essential. It's the foundation of good data practice.

Let's look at the kinds of problems that data cleaning addresses. For the most part, the problems seem simple—until you start thinking about them:

Missing data

You have a stream of sensor-collected temperature readings, possibly from machinery in a manufacturing plant dropped into a data lake. Suddenly, those readings disappear. How do you handle this? You can't just replace the missing readings with a 0, as that would yield incorrect averages; depending on the application you might be able to ignore them or replace them with some kind of average value, but it's still important to investigate the cause. Was the device turned off? Did the thermometer fail? You might find that all the missing data points come at night, and you have to consider what that means—otherwise, you'll have an incomplete picture of what's happening in your plant. (Is someone accidentally turning the sensor network off before the night shift starts? True story: we once worked with a client that was having nightly outages. We watched as the IBM service team scoured logs and ran tests, only to find out that the cleaning staff were unplugging the servers to clean behind them. When the machines were plugged back in, their recovery protocols kicked in and all was fine.) Figure 7-2 shows an example of missing data in a different hypothetical scenario.



More Science High

First name	*Last name*	*Day 1*	*Share*
Audrey	Farber	0	**16%**
Betty Jo	Bialosky	184	
Nancy		136	**20%**
Nick	Danger	179	26%
Rocky	Roccoco	160	**24%**
Melanie	*Haber*	188	
Susan	*Underhill*	185	31%
Lieutenant	Bradshaw	2000	**23%**

Figure 7-2. Spreadsheet with missing and incorrect data

Incorrect data

Again, consider a data set of temperature readings. Suddenly, the readings go haywire: they're all significantly above normal. Does that indicate that the machine is failing? Or that the thermometer sensors went bad? It can be even harder in this case to "repair" the data, because you don't know that it's actually incorrect.

Contradictory data

What happens if, in a customer database, one customer has two different addresses? One record might say they live in Chicago; another might say they live in Lincoln Park (a neighborhood in Chicago). One record might have the zip code 60634; another might have 60643. Which one is right? Is it a simple data entry error? Or could the records correspond to two different people? This is a master data management problem that needs to be handled on the Organize rung.

Different names for the same thing

As we mentioned in Chapter 4, an employee once counted over 100 names in a single database that were used to refer to contracts with IBM. When you have different names for the same thing, treating each version as a distinct entity is certain to wreak havoc on your results (if IBM and I.B.M. show up in your analyses as different companies, that will lead to results that are inaccurate and unreliable). When we attend conferences and our badges read "International Business Machines," many people have no idea where we work! Solving this problem is called entity resolution. Once you're settled on a canonical name (for example, IBM), your data cleaning process must be able to detect and disambiguate entities that look different, but are actually the same. (This is particularly important for fraud detection.)

Ambiguous data

O'Reilly Media was once working with a database of job listings, and looking for positions requiring expertise in Apple computer products. That may sound easy, but in an unstructured data set, it's very difficult to tell the difference between Apple (the computer company) and apple (the fruit). Candidly, we were surprised when the folks at O'Reilly told us this was a problem for them (something we didn't believe at first), but we were even more surprised when they showed us how many job listings there were for apple pickers, apple processors, and the like.

Data in different formats

Sometimes you're combining data from two or more sources, and the same types of data are formatted (and stored) differently by the two originating systems. You may need to perform some operations on this data to normalize it before using it with machine learning algorithms. Here's a trivial example: what if you're working with accounting data, and one source stores amounts as regular integers (not a great practice, but it happens), while another uses a special AMOUNT user-defined datatype? You've got to reconcile the two.

Insufficient features
Even if your data has multiple components, you may wish to introduce additional variables—known as features—to improve the accuracy and results of the algorithm. This may involve finding external data sets that relate to your data, so you can expand the pool of available correlations and causes from which to draw upon.

And that's just a start. One thing should be obvious, though: detecting problems with data is the easy part of the job. Deciding what to do about those issues can be challenging, and doing it incorrectly can cause serious problems.

Believe it or not, most data cleaning is done by hand: data scientists write scripts in some language like Python to detect problems, and fix them in whatever way they decide is appropriate. But this challenging and labor-intensive process is itself a problem. As Arvind Krishna, IBM's CEO, commented in an interview with the *Wall Street Journal* in May 2019: "About 80% of the work with an AI project is collecting and preparing data. Some companies aren't prepared for the cost and work associated with that going in. And you say: 'Hey, wait a moment, where's the AI? I'm not getting the benefit.' And you kind of bail on it."

If 80% of a data scientist's work is data preparation and cleaning, cutting that time in half is a huge win. It triples the time they have available for other work, while provisioning processing power and databases so fast that they cut the training time to practically zero only gets back 20%! The biggest wins come from optimizing the parts of the process where you're spending the most time—this is called Amdahl's law. Here are a few best practices for speeding up data preparation:

- Disqualifying a source of data early in the process is not a bad thing. An unfortunate number of data scientists think they've found the data they need, only to discover that it's not usable—it may be out of date, biased, or have other flaws. There is no reason to waste time with a marginal source that adds little to the accuracy or explainability of your results. (The same goes for features in a data set, like a column.) Here's a clue: who else has used this data set? How often has it been used? Data that's used frequently is likely to be up to date; data that's been sitting in "cold storage" for a few years is probably there for a reason. Finding data that nobody is using is an exciting part of data collection, but don't jump to conclusions when you get to the Organize rung.
- Experiment with different subsets of the data to find which is easiest to clean. Cleaning an entire data set can take hours, but cleaning small subsets of that data can take a fraction of the time and be just as effective given different scope requirements.

- Play with different record filters. Filtering out certain variables or features may not have harmful effects on an AI experiment. Then again, it may. Understand how this is done.

The solution to the problem of data preparation and cleaning will inevitably be to use AI to build AI. Speaking at O'Reilly's AI Conference, Stanford's Chris Ré said that we've democratized model building and data collection, but democratizing data cleaning is a harder problem. Yet thanks to work by Ré, IBM, and others, we're beginning to see tools to automate data preparation.

Documenting and Cataloging Data

As part of organizing its data, your organization must also document and catalog that data. We're sure you wouldn't be surprised at how little effort most companies put into documenting their software assets—but if nobody can find the data they need, for all practical purposes it doesn't exist (you don't want the success for your data investments to be based on lucky finds). A catalog is nothing more than a record of the data that you have, where it is, how it is accessed, and how it can be used.

To understand why it's important to document and catalog data, it's helpful to once again invoke a library analogy. If a library were simply a room filled with thousands of books, it would have little value to the average reader. Libraries are useful because they are organized and provide a catalog to help readers find information in various ways—you can find all the books by a specific author, in a specific genre, or on a specific topic. No matter what attribute you use to find the book you're looking for (genre, topic, etc.), you end up with a unique locator where you can find it. This catalog must be maintained and updated every time a book is added to or removed from the library.

The same goes for data. Organizations need to have a catalog of their data to provide information about its source, who owns it, metadata mapped to its business context, and so on. There's a long list of information that should go in the catalog. Questions to consider include:

- Where is the data located and how is it accessed?
- How is the data stored (for example, the data type used, as a key-value pair, and so on)?
- What are the data set's contents?
- How are business terms defined?
- What policies control how the data is used?
- What regulations control how the data is used?

- Who is allowed to access the data?
- Who is charged with the responsibility of stewardship over this data?

A data catalog can also contain additional information, as discussed in the paper "Datasheets for Datasets". Additional things to think about are:

- How was the data collected? This can have a significant influence on its accuracy. Was data collected in a survey, where participation was voluntary? Is this a sample from a larger data set, and if so how was that sample constructed? Where did data collection take place? For example, for retail point-of-sale data it's obviously important to know what store the data was collected at. Customer shopping habit data from Neiman Marcus and data from Walmart will lead to very different results.
- Why was the data collected? What was the motivation? Data is rarely collected without a purpose, and that purpose can color the results; it can easily influence how data is collected or what potential data sources are selected.
- What biases are inherent in the data? Until we can collect data from the future, data will always be historical, and subject to historical biases. Real estate data, for example, will show the effects of redlining, and if you don't account for that it will bias your results.

This additional information can go a long way toward helping you understand and address biases. Bias and unfairness in AI almost always comes from the training data, and is amplified by the process of training the model.

Here's one sneaky way in which bias can enter into your models. Let's say you're building a medical application that recommends treatments for patients. This kind of model can never be 100% accurate (nor can a human doctor), but it needs to be appropriately accurate for all groups of patients. If your model is more accurate for some groups of patients than for others, you have a problem: the model is biased.

In India, iKure was working on an application for predicting cardiovascular disease, using AI and wearable devices. Using a dashboard within IBM Cloud Pak for Data, the organization realized that its results were biased, predicting disease more often in male subjects. Examining the data sources revealed that males were overrepresented in the sample population relative to females. Each data item on its own was correct, but because females were underrepresented in the training data, the application wasn't able to make accurate predictions for them. This kind of bias is called "disparate error," and it arises directly from your training data. The average success rate may in fact be the same for both groups, but you are more likely to overtreat or undertreat a subgroup. Problems like this are very difficult to solve if you're just looking at your data—but they become easier when you've built a data catalog, particularly if that catalog is automated. It's important to be aware of problems of bias, and to include relevant information about possible sources of bias in your data catalogs.

It's worth noting that your goal in AI isn't to *exceed* human performance (remember, in the AI world performance means accuracy). Start out with the goal of *matching* or getting pretty darn close to human performance. Sure, you hear lots of stories about how AI can spot health issues in medical imaging better than radiologists—that's fine. But simply matching human performance will go a long way toward AI's acceptance, as well as providing a way for those professionals to free their time from mundane cases, and in the end save more lives. There's a ton of value there.

Understanding Data: The "Seller" Gong Show

An important part of documenting data is understanding exactly what the data means. We all know that language is ambiguous, but that human language is at the root of our data. You, in combination with your subject matter experts, need to be aware of this ambiguity and work to resolve it. We experienced this ourselves when, as part of collecting data for a machine learning project, we tried to locate and standardize all the data from all the possible sources within IBM that had to do with sellers or products. It turned out that within our large corporation, there were many individuals and groups that were each certain they knew what a "seller" was. The problem was that no two individuals or groups agreed on that definition. In fact, we watched helplessly as one group reported on sellers based on how they were funded in their respective geographies, while another group would roll up seller reports by division through job titles. Ultimately this impasse resulted in what we affectionately dubbed "The 'Seller' Gong Show."

In retrospect, this isn't surprising. Practitioners of domain-driven design (DDD) often find that different parts of a business define common terms in different ways. Because everyone knows what a "seller" is, nobody questions whether their understanding is shared. A few eyebrows may be raised occasionally, but this is quickly forgotten. A seller is someone who sells, right? Not so fast—let's take a moment to think carefully about the possibilities. A seller could be:

- Someone who gets paid on the deal or assists in some manner, but doesn't actually do the final selling
- A person who sells things (for example, an individual on eBay or a corporate salesperson)
- Someone who enables sellers to sell and helps on occasional deals
- A corporation that sells specific products (for example, Weyerhaeuser)
- A supplier that your business buys from

- An entity that buys your product and resells it (for example, a bookstore resells books bought from a wholesaler or publisher)
- An account number (to the finance group)
- An address or a shipper number (to the warehouse)
- A name and an employee number (to HR)
- An individual product (for example, a best seller)

That list could go on almost indefinitely. What's important to understand is that within a single company all of these meanings can be in play, and they're reflected in each group's data. The seller in a real estate transaction isn't the same as the vendor who provides parts for the manufacturing process, and both are different from the company that sells you your internet connection.

That may be OK for day-to-day operations between people; we're good at understanding contexts and knowing which kind of "seller" is appropriate in any given conversation. But data applications don't understand context, and they really don't like ambiguity. If you're going to build AI applications—and specifically, AI applications that span different parts of your business—you will have to track down and resolve these ambiguities. You're going to have to define what a "seller" means, and stick to it, at least as far as your data is concerned. (We'd bet that your humans will also perform better if they can agree on their terms, but that's another issue.) Your data catalog should make it clear what these terms mean and how the data is stored.

Metadata for Models

We've said a lot about what goes into a data catalog: all your company's data sources, along with the data describing those sources (where they are, how they're accessed, what kind of access is allowed, how they've been modified, and more). But metadata about data isn't the only kind of information that goes into a data catalog. What else?

Your data catalog also needs to show what models are deployed and the metadata associated with those models: their purpose, how they were constructed, what types of data sets are used for ingestion, and so on. Data should be a browse-and-shop paradigm, just like how you would shop online for any product, and only metadata can make that happen. The data science team should not have to reinvent the wheel every time a new request comes down the pike. For example, perhaps someone in the Marketing department has already built a sentiment analyzer for industry-specific manufacturing terms; Customer Support could use that ontology to apply machine learning to automatically tag and route emails based on a faulty part in question or use transfer learning to capitalize on work already done such that the AI understands the product set from the get-go, and expand that to suit another use case.

The data catalog should document how to access the models you use, in addition to documenting how to access the data. What's the API? Is the model accessed through REST or some other interface? Does some other hook execute the model when some triggering event happens (like a credit card application being submitted, or an online book's free preview chapter being read)? Some models are even designed to run in real time as more and more inputs are submitted, such as sentiment and intent analysis models that analyze interactions between customer support agents and their clients. No plans to wrap programmatic APIs and services around that corporate solution? Go back and read Chapter 5 again.

Maintaining the Catalog

Maintaining a data catalog sounds like a lot of work—and it can be, but it can also be liberating. You'll know what your data is, where it is, and what its properties are. You'll know exactly how far you can trust it. You'll know who can and can't access it—and a data catalog like IBM Watson Knowledge Catalog doesn't just record data access rules, it also enforces them. The metadata can describe if specific data needs to be masked when surfaced to an application, what positions are authorized to access different data sets (enabling access to be allowed or denied automatically as appropriate), and more. Automating access control is much easier, less frustrating, and more consistent and accurate than working through system or database administrators. It's too easy to make mistakes when you're telling a DBA who should or shouldn't have access to a data set; a catalog can manage this automatically.

Just like when a library acquires new books, when new data sources become available they should immediately be added to the catalog (even if it's to the 'restricted use' section). This rule applies to external and internal data sources, data sources that are replicated, and data sources that are integrated with other sources; if the catalog is not complete, it won't be useful. Since the catalog also enforces data access rules, catalog maintainers should remain abreast of regulations.

Governing Data

Finally, your organization must govern the data to ensure that only permissioned users have access. Say the phrase "data governance," and watch your data scientists and AppDev teams roll their eyes. Controlling access not only sounds boring, it sounds like it will prevent you from doing what you want.

That attitude is both dated and dangerous. A decade or so ago, we lived in the "wild west" of data: if you had data, you could do what you wanted with it. Data is increasingly subject to regulation; the GDPR and CCPA are just the beginning of what promises to be a significant trend. This means that you need tools to govern how your data is used, where it comes from, and who is allowed to use it. "Data governance" doesn't mean rebuilding the silos, but rather ensuring that data is used correctly, legally, and ethically. (Many data governance platforms, including IBM's, have knowledge of GDPR, CCPA, and other regulations built in, and are equipped to help you enforce them.) Our pro tip: get in front of this. You don't want to lead an organization that looks at governance from a "least effort to comply" approach. If you can flip this culture on its head, getting your head around good governance for reasons other than regulatory requirements not only keeps you out of trouble, but there are serious regulatory dividends to be had that will accelerate your AI journey.

Controlling data access also protects you. Many businesses underestimate the pitfalls that poor data governance can create. A major bank learned this the hard way when it discovered that a sales manager and their team had adjusted numbers that ended up costing the bank millions—not to mention the effect on the brand's reputation. Because the company didn't have appropriate data governance policies, it fell victim to the implications of broad access control: too many people were allowed to access the data, and not all of them were trustworthy. (We talk a lot in this book about making all data available for access—but we say this implying it is being done in a governed manner; that may be restricting some data on a need-to-know basis, or auditing any change to a certain data field, and so on.) Take our advice: it doesn't matter why the data was adjusted, whether for criminal purposes or with the best of intentions. Your data can't be reliable and trustworthy if you don't know who can access it. Just as organizations must govern access to data, they must also govern the data itself. Changes to data need to be recorded (data lineage), so that when you discover a problem you can back the changes out—and this extends all the way to air-gapped tape archives, because you'll want information on those backups cataloged too. Data governance is even a defense against cyberattacks that corrupt data; if you know where your data came from and how it has been modified, you're in better shape when you need to recover.

We've seen many examples of failures in data governance. Facebook's Cambridge Analytica scandal was, in essence, such a failure; it turned into a major public relations disaster that put Facebook executives on a never-ending "apology tour" including appearances before the US Congress, and at the time it cost the company over $100 billion in market capitalization. To this day, this event has eroded people's trust in the organization, and several other companies as well. Facebook wasn't in control of who was using its data. When the problem was reported, it appeared to ignore it; when it couldn't ignore it any longer, it told Cambridge Analytica to stop using the data but did nothing to enforce this. It's a textbook case of how to fail at data governance.

So how do you *not* fail? Establishing a consistent policy on data access is a big job. It starts on the Collect rung of the AI Ladder, where you should be aware of (and collect) the terms under which data can be used. When we talk about data governance, we're referring to the management of data availability, relevance, usability, integrity, traceability, and security. Governance helps organizations manage their information knowledge and answer questions such as:

- What kinds of information do we have?
- What does that information mean? How are data fields defined?
- How did we get this data?
- Are we managing this data like we said we would when it comes to its lifecycle and security and regulatory needs?

While these questions seem reasonably straightforward, putting data governance into practice can be a huge obstacle for many companies. Why? Because in most organizations, no one has owned "data" across the entire business, at any point. Individual departments are each collecting some data, retaining it for who knows how long, and using and disposing of it in a completely uncontrolled way. In order to implement an effective enterprise data governance strategy, you generally have to circumvent the departments and leaders that own the very business data at stake—and that has political and technical pitfalls. You have to define a policy about how data should and shouldn't be used. Remember the corporate-wide information architecture policy we stressed in Chapter 6? That's what we're talking about here. This policy must be informed by the metadata you collected with the data, and then you have to enforce the policy; otherwise, it's meaningless.

Devising a methodology to enforce the policy is a separate job. Your company's information architecture must accomplish both of these goals. If you approach data governance by trying to do the minimum necessary, you will get the least out of it. For example, it may be possible to mount a cron job to "clean" your data so that it conforms to all current regulations. But ask yourself, "What if the regulations change?" A well-designed information architecture anticipates change and allows your company to effortlessly adapt to new requirements.

As Seth Dobrin, IBM's VP and Chief Data Officer of Cloud and Cognitive Software, says: "Data governance isn't a retardant; it can be an accelerant, if done right." Proper data governance ensures the following:

Known data
: You know what your data is, and where it is. You can't govern data if you don't know what data you have or who owns the policies and protocols around that data.

Conformed data
: You know that your data is trustworthy, complete, and consistent.

Secure data

You know that your data is safe: those who should have access can access what they need, and those who shouldn't are blocked.

Compliant data

You know that your data use enables the enterprise to run, build, and manage AI within the confines of internal and external requirements.

Known data, conformed data, secure data, compliant data: that's the positive side. Here's what you'll be fighting if you don't develop a data governance strategy:

Stale data

Without comprehensive governance, you have an incomplete picture of when data was captured, and you also have no assurance that data is not being continually captured. More recent data may be available somewhere else. Or there may be valuable data in some obscure database, but because that database is inaccessible outside of a small group, and that group doesn't understand its value, the data becomes stale; its value then deteriorates into nothing.

Data of unknown origin

Without governance, you will have no way to trace how data came to be in any given database (data provenance). Was it from a real-time system? Is it an excerpt from some other system? Has the data been modified or transformed in any way? These are questions you cannot answer without a sound data governance strategy.

Data that's untrustworthy because its lineage cannot be verified

If you don't know whether you have all of the data, whether it's original or has been modified, and whether it reflects the current reality or is stale, you can't trust it. (This hits this point home: think of a traveler without a passport trying to gain entry into a foreign country. Data looking to enter your AI should have the same rigor.)

Ungoverned, untrustworthy data is an AI anti-pattern we have seen over and over, from medium-sized businesses to the largest enterprises. And it is a huge roadblock to becoming an AI-driven organization. A company will never become AI-oriented if leaders and employees aren't confident in the results produced by the AI. Data is the bread and butter of AI, and the quality of that data directly affects its results. Your AI is only as good as your data. If you have bad data, you're not going to have trusted, transparent AI.

Again, data cataloging tools can help to enforce data governance, and the payoff from enforcing these rules can be huge. But it's cultural, too: better linear and data governance can save you from PR disasters, as well as millions of dollars in losses. It's fundamental to your AI strategy.

Enterprise Performance Management

We see three common approaches for enterprise-wide data initiatives and data governance:

Centralized
: Set up a separate division for data projects, hire a new team, and make new plans. The biggest pitfall with this approach is that it separates the data projects from the enterprise's core. We have not seen this approach work. You need more than new people and new tools; you need a point of view, a roadmap to achieve it, and a relentless focus on training. If your data expertise is centralized in a single team, it will be very difficult to infuse AI throughout the organization.

Decentralized
: Every division does its own thing. The issue with this approach is that hiring becomes very difficult to scale, and there are multiple strategies and priorities. You are bound to end up with competing solutions to the same problems. You're almost asking for data silos, and all the problems they entail.

Hub/spoke
: This is a hybrid model, with a centralized strategy and priorities and decentralized execution. In our experience (including IBM's own transformation), it's where we've seen the most success.

IBM's efforts to make critical data available across the enterprise are a great example of implementing data governance and enterprise performance management. IBM took the "hub/spoke" approach to building its data practice. We're confident that this is the right model, and that it's most likely to lead to successes for other organizations. Four years ago, IBM hired its first global chief data officer (CDO), Inderpal Bhandari, who set out on a mission to ensure that the organization led the way in exceeding requirements for data privacy, breaking down silos of data, and making data readily accessible. This effort was the basis of IBM's own business transformation around the use of data and AI.

IBM's data privacy approach was first focused on tracking the complete lineage of all personal data (PD). IBM went beyond the requirements of the GDPR by defining "personal data" to mean any data that was related to an individual. It could have carried out this task by attempting to track these connections manually—a process that would have been expensive and inefficient, requiring hundreds of people to implement. Instead, AI was used to do the vast majority of the work required to build data lineage across IBM. Even when automated by traditional rule-based approaches, identifying personal data can be extremely difficult and error prone. Since we are firm believers in Occam's razor (a problem-solving approach that essentially guides practitioners to remove the unnecessary stuff, on the basis that the simplest solution is most likely the right solution), we were excited when IBM extended the capabilities of these rules by applying AI to fill in the gaps and reduce the number of false positives and false negatives often generated by rule-based platforms.

IBM's approach to integrating siloed data was to create a set of enterprise-wide data standards. This is an ongoing process that started with the most critical of data and continues to this day, with new standards proposed every month. The process looks like this:

- The global CDO proposes a set of standards that are usually based on the largest producer or consumer of that data type (for example, finance data standards are based on the CFO's definitions).
- These standards are then vetted by a business unit data office (BUDO) council. The BUDOs are either business unit CDOs or senior business unit representatives who have purview for much of the data.
- Once the proposed standards are accepted, there is an exercise to assess the impact of those standards on the business units. Based on that assessment, a timeline for implementation is set.

This three-step process, coupled with IBM's approach to making data accessible across the enterprise, has enabled the creation of trusted views of data for running the enterprise. The two-pronged data accessibility approach combines centralizing data and virtualizing data across a hybrid cloud. Both the centralized and the virtualized data are made available via the Cognitive Enterprise Data Platform, or CEDP. (It is not called a "data lake" because that term often implies a Hadoop-based platform.) CEDP is a polyglot, object store–centric data platform based on IBM's Cloud Pak ecosystem.

The most mature application is the Enterprise Performance Management (EPM) tool, which takes the data from CEDP and conforms it to a more rigorous set of standards. Today, this is mostly a finance-centered set of reports and dashboards used by everyone from the product teams to the CEO. By using the same data source from the lowest to the highest level, and ensuring that the data conforms to standards, we can be confident that the numbers add up. This leads to more productive decision making, instead of disagreements about the quality or trust in the data.

Example: ANZ Banking Group Embeds Sound Data Management and Governance Policies

The multinational banking giant Australia and New Zealand Banking Group (ANZ) has been using data analytics to extract more personalized insights about its institutional customers. The goal is to make better strategic business decisions on such issues as liquidity, risk, cash management, new store locations, or inventory. Given its size—its institutional banking division operates across 34 markets globally—ANZ faces significant regulatory control. Plus, market trust in banking firms in general still suffers from the 2008 financial crisis. As a result, the firm puts a top priority on security, governance, and compliance, especially with regard to data.

By using a managed cloud provider service that serves as a repository for its data pipelines, ANZ has managed to ensure the security of its customer data and the reliability of its applications. To support this approach, ANZ uses analytics and AI to analyze aggregated, deidentified data sets, including credit card data—a process that previously took its bank professionals days, but now takes just seconds.

By embedding good data management and governance into data analytics and AI programs from the get-go, ANZ has been able to solve one of the biggest challenges emerging in AI: it is able to quickly ensure that it addresses potential bias in its data sets and rising concerns about customer privacy. It is also able to ascertain that its systems are transparent, explainable, and compliant, and ultimately instill trust with customers and auditors alike.

Internal ANZ surveys show that a lack of trust in AI is the biggest roadblock to its implementation, and that proving to customers that data security, privacy, and transparency will be prioritized has become the most important requirement for investment in machine learning tools.

DataOps

Data governance, data catalogs, data provenance, data pipelines: all this sounds like a lot of infrastructure to build and maintain. It is. That's why a new specialty has emerged, called Data Operations, or DataOps. The rationale for DataOps is simple: your data scientists and AI experts want to spend their time building models and working with data, not maintaining the infrastructure that supports their work. And becoming an expert in that infrastructure is easily a full-time job.

In addition to maintaining the infrastructure that keeps data organized, DataOps people are key players in ensuring a smooth transition to production, making sure production keeps operating smoothly as the project undergoes new releases, and helping build the operational infrastructure that the project requires. This means that DataOps staff must also work with the team that's responsible for operations and production—often called DevOps, though many DevOps advocates argue that DevOps is about culture, not a "team" or a job title (we'll have more to say about this in the next chapter). Table 7-1 shows the general distinction between DevOps and DataOps.

Table 7-1. Differentiating DevOps and DataOps

DevOps	DataOps
Automated deployment	Data infrastructure
Continuous integration and deployment	Data organization
Monitoring	Data pipelines
Observability	Data governance
Testing	Data catalogs
Source maintenance	Data access
Agile process	BI and analytics tools
(all in cooperation with developers)	(all in cooperation with data practitioners)

We've seen points of view arguing that you should hire your first DataOps person even before you hire a data scientist—and they make some valid points. The first thing data scientists will do is start building the infrastructure needed to support their projects, and that's not a task for which they're particularly well suited; data scientists are typically from academic backgrounds, where they're more used to running experiments than deploying code or managing data infrastructure. Whether or not you're willing to go that far, you do need to take the complexity of managing your information architecture into account, along with the challenge of deploying AI at enterprise scale. If deployment fails, or never happens, the road to AI becomes an exercise in futility.

Now That Your Data Is Trustworthy, on to Analysis!

If there's one key takeaway from this chapter, it's that *data isn't valuable unless it's trustworthy*. If your company doesn't trust its data, it has no business trusting the results derived from the data—especially since the process of training a model tends to compound errors. If you're going to succeed with AI, you need trustworthy data. Period.

And if there's a second key takeaway from this chapter, it's that *data governance isn't a roadblock to success; it's an enabler*. We'll be so bold as to suggest that in our experience, it's an accelerant! Yes, we'll admit, in some ways it looks like cybersecurity: "We spent some millions of dollars, and nothing bad happened last year. Guess that was a success." But we can guarantee that you're bound for failure in AI (as well as security) if you take that attitude. "Move fast and break things" was the last decade's slogan, and it led straight to Cambridge Analytica and an unprecedented number of cyberattacks and data thefts (that's why we keep saying "fail fast but fail safe," which is a more appropriate take on this common catchphrase). Governance enables you to develop AI responsibly. And when done properly, an approach to governance that's coordinated by your AI platform is more streamlined and less error-prone than working through system administrators and DBAs. Organize your data. Get it under control, get it cataloged, get it governed. Then, you're ready to move on.

Over the preceding three chapters we've covered the transformations you'll have to make in order to ensure a steady supply of business-ready data for your machine learning projects. In the next chapter, we'll discuss how to use that data to find answers to real business problems, employing machine learning techniques. We'll also explain how to take those models from R&D to the production environment—the "real world."

As you get ready to read the next chapter and move to the Analyze rung of the AI Ladder, remember our mantra: big data without analytics is just a bunch of data.

CHAPTER 8
Analyze Your Data

After an organization has collected its data and organized it in a trusted, unified view, it can tap into that data to build and scale AI models across the business. This allows companies to glean insights from all of their data, no matter where it resides, and engage with AI to transform their business—putting themselves at a clear competitive advantage.

The earlier rungs on the AI Ladder were about getting ready to make usable models based on reliable data. Now, on the Analyze rung (Figure 8-1), we can actually put that data to work and bring AI and machine learning into the picture. This is where we move from theory to practice; where the rubber meets the road, so to speak.

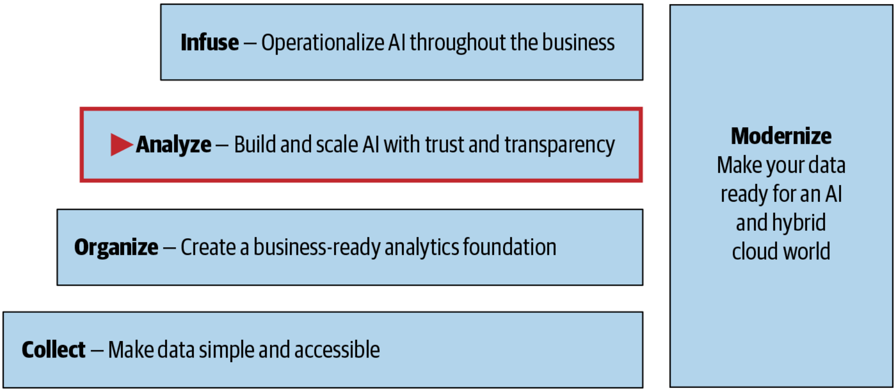

Infuse – Operationalize AI throughout the business
▶ Analyze – Build and scale AI with trust and transparency
Organize – Create a business-ready analytics foundation
Collect – Make data simple and accessible
Modernize
Make your data ready for an AI and hybrid cloud world

Figure 8-1. The Analyze rung of the AI Ladder

This rung of the AI Ladder presents its own unique challenges. Hasty investments at this stage can create some serious issues with regard to tools, people, and processes. Point solutions create complexity in integration, maintenance, and support, leading to increased technical challenges and costs. And there is the whole issue of ensuring trust and explainability for the analytical models built on this rung. A failure on this front can make your entire AI investment worth nothing, because nobody believes the results are worth using.

In this chapter, we'll discuss the process of analyzing your data and building your analytics capability. Many organizations build great AI models, but it ends there; the models never get to production, or they get to production and die. To succeed, you need to build and scale your AI in a systematic manner, based on the AI lifecycle. This lifecycle is conceptually similar to the development lifecycle for traditional software projects. The AI lifecycle consists of three phases: building a model (*Build*), deploying it (*Run*), and managing (*Manage*) its execution across all iterations. The Analyze rung of the AI Ladder is about installing the pieces of that lifecycle in your enterprise.

Why Organizations Need an End-to-End AI Lifecycle

The main reason to put an AI lifecycle in place is to create a sustainable, repeatable, consistent framework by which AI operations can be infused through the enterprise. Instead of tasting a little joy and success with a small experiment and then wildly scaling it out with no structure, smart organizations take a cue from the operations world and put together a structure within the company so that AI is "operationalized": run in production to certain standards that ensure quality is high, uptime is high, errors are low, and feedback loops are used and consumed. Many AI projects are started, and even finished, successfully, but never deployed to production. Understanding the AI lifecycle and putting one in place will help your projects make the transition to production, and ultimately allow you to scale AI across the organization through your various workflows.

This lifecycle is fairly simple on the surface, but like in an onion, there are layers of complexity to peel back and explore.

Build

The first step of an end-to-end AI lifecycle is building models. This process is like what happens in an artist's studio or on a carpenter's workbench; it's where companies create and train models they can then use for predictions. At this phase in the AI lifecycle, it is critical to ensure that your company is using the right algorithms to build its models for making predictions. In this stage, you should be considering the tools and techniques that you will use not only for preparing data and engineering features, but also for training your machine learning models.

In the typical software development lifecycle, this is the Development phase. In AI your developers are the data engineers and data scientists, and perhaps a software engineer to help with integration and APIs. They're engaged in exploring and preparing data, building models and training them with code, evaluating each model's performance (accuracy), and then, when it's ready, publishing that model to a global catalog for increased availability.

We'll assume for the purposes of this discussion that you have good data that has been properly vetted and access-controlled. To build an AI application that will solve your business problem, you now need to find the model that makes the best use of that data.

Example: Using Machine Learning, an Insurer Cuts Costs and Boosts Productivity

A North American health insurance company wanted to improve its claims-processing process. Specifically, Medicaid claims with errors or missing information were being denied and sent back to the healthcare provider to be corrected and resubmitted. Because the insurer receives thousands of claims daily, this was causing long delays in processing the resubmissions. But with too much of a delay, the insurer could be subject to fines in the form of interest and penalties. In addition, its claims-processing team was taking a productivity hit.

With IBM's help, the company built a machine learning model that analyzed claims as they came in and automated the manual reviewing of resubmitted claims by human workers. Its ultimate goals were to predict duplicate claims, identify the claims most likely to be resubmitted, predict possible issues with claims up front, and educate providers on how to avoid having to resubmit claims. It achieved all this and more, and was able to realize both cost savings and productivity enhancements.

Run

After a model has been built and verified, it needs to be put into production within an application or a business process. You should be aware that a model is typically a relatively small part of an AI application; before you can deploy the model, the entire application needs to be built. Once a model is deployed, it is running in the organization—perhaps making claims decisions, pricing decisions, and so on—and can be retrained (which will have to be done to avoid drift or staleness) as needed. Depending on the size of the application, deploying it "in the real world" may be as simple as hosting a website, or it may entail a complex logistical operation requiring physical resources and specialized talent. "Hidden Technical Debt in Machine Learning Systems" by D. Sculley et al. is an excellent discussion of AI's impact on operations (Figure 8-2).

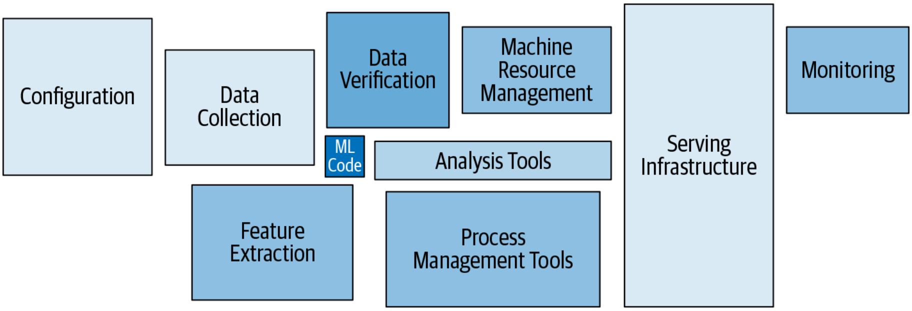


Figure 8-2. Code is only a fraction of real-world machine learning systems (source: D. Sculley et al.)

Manage

The final step of the AI lifecycle is to keep things running: to manage the model and its results over time. During this step of the AI lifecycle, two things need to be happening:

- Alignment of model output with key performance indicators (KPIs)
- Iterative learning, retraining, and redeployment of the model

Putting AI in place for experimentation is one thing, but deploying AI in production is an entirely different animal, requiring different skill sets and techniques. Research has shown that most organizations need help turning production AI into a consistent, repeatable, constantly improving virtuous cycle that has a real impact on business results without running afoul of the pitfalls of automated modeling and decisioning.

IBM Watson OpenScale can help businesses to monitor AI in production for consistency, performance, drift, bias, staleness, and explainability. Watson OpenScale can create more explainable outcomes, more accurately assess risk and predict results, and help to trace how decisions or recommendations are made to satisfy regulatory requirements. It can also prevent or correct model drift and alert the right people before the consequences are too severe. In early 2020, IBM Watson OpenScale won an AI Excellence Award; with it, it was named one of 31 products, people, or companies making AI a reality in 2020!

Aligning Model Output with Business Metrics

You'll need to synchronize and align the results you are getting from your model with the KPIs and business metrics that drove the whole process from the start. If your model is spitting out results that are difficult or impossible to translate back into some quantifiable impact on a business indicator, management will have no way of telling whether the AI experiment was valuable or not. Is profit up? Is fraud down? Are claims adjudicated faster? Ensuring that you are transparent about your results is very important. That's the key to being trustworthy—if the AI team isn't trusted, it doesn't matter whether or not your data is. Don't be afraid of failures; acknowledge them, and make sure you start enough projects that you also have some successes.

Learning, Iterating, Learning

Imagine you've just hired a new job applicant who graduated at the top of their class and has a stellar résumé. They know everything there is to know about the job, and have the skills that your business needs. There's just one catch: from the moment they join your team, they act in a manner that makes you think they've vowed never to learn anything new again. Given the choice you probably wouldn't (we're telling you not to!) make that hire, because you know that lifelong learning is vital if someone is going to add long-term value to your team. Put it this way—between the two of us, we have overseen the careers of thousands and managed organizations with tens of thousands of employees. One thing separates the spinners from the winners: lifelong learning. We've seen incredible talents who learned their way into high-profile IBM Distinguished Engineer positions, then sputtered out because they failed to keep learning. At the same time, we've seen incredible talents rise because they never stopped learning and stepping into new domains (neither of us had any formal computer training, and we figured out a way to hack some code together to build models). You'll often hear us say, "Tech years are like dog years." Yup.

So, if people need to strive to be lifelong learners, and we encapsulate people's knowledge into our AI models, then why should our models be any different? Yet when we turn to the field of AI, we see companies making a similar mistake all the time with their models. Data scientists work hard to develop, train, and test new models. However, once they get deployed, those models don't learn anything new. After a few weeks or months, the models become static and stale, and their utility as a predictive tool deteriorates—always remember, the moment you put your model into production is the moment it starts to get out of date and drift.

Data scientists are well aware of this problem, and would love to find a way to enable their models to participate in the equivalent of lifelong learning. However, moving a model into production is typically a tough task, and deployment requires help from busy IT specialists. Automating deployment makes it much faster and more reliable (companies like Facebook and Amazon deploy certain application components dozens, if not hundreds, of times a day), but it's still no wonder that most data scientists prefer to hand over their latest model and move on to the next project, rather than persist with continually retraining and redeploying their existing models.

Deployment isn't just painful for data scientists; it can be a headache for IT teams, too. Data scientists might have used any one of a wide variety of languages, frameworks, and tools to build their models, and there is no guarantee that those choices will make the model easy to integrate into production systems. For example, we frequently see models that are built on the latest open source packages, but can't be deployed in production because older versions of those packages are the only ones approved for deployment. Data scientists and other developers working with open source technologies are always pushing the boundaries of their capabilities with the "latest and greatest" packages, and that's a good thing, but doing so can have strange pitfalls—and this is one of them. What happens if your data science team is using version 1.9 of some great new AI library, but IT has only approved the use of version 1.7? As an AI leader you should know that many open source projects don't prioritize compatibility with prior versions, so you have a problem. Rebuilding the project with version 1.7 might work—or it might not. Fighting IT policy might work—or it might not, and you need that team on your side for future projects. You'll definitely have problems with IT if you fight for version 1.9, win, and then that version causes problems for them down the road.

Here's a real-life example: we built a model that used a Python scikit-image library with a HOG descriptor, and our script suddenly broke when we did a `pip upgrade` operation. Why? The latest version would not accept the `visualize=true` option in our code. Suddenly we had to spell it as `visualise=true` (note the `s`). We're not sure what Commonwealth English teacher/open source committer was behind this, but it illustrates the point perfectly. If you had working AI code that supported business decisions, and that code was simply set free across the enterprise, you might find yourself in a world of pain if you don't comply with your IT team's protocols for version support. Back-level code compatibility is just one of the many pitfalls awaiting you in the deployment phase.

In a worst-case scenario, the model may need to be substantially refactored or even rebuilt from scratch before it can be deployed. And if data scientists ask for their models to be redeployed too frequently, they may be met with significant resistance from the IT department, so they move on. You can't allow that to happen.

Example: Risk Management Company Gets Creative to Offset the Expense of Training Models

A risk-management company based in San Francisco was using drones to take photographs of building roofs to detect damage. The results were then analyzed manually by highly experienced employees to determine if they had any flaws, faults, or damage. But the firm was unable to scale to meet demand in the market—the image analysis was extremely time-consuming, and it didn't have sufficient skilled personnel to inspect all the roofs the firm was being requested to examine. It decided to build a deep-learning model to help it analyze images automatically, speeding up the analysis process and freeing up its skilled workers for especially difficult examinations.

Given the lack of a preexisting data set to train the AI model, the first step was to build a custom data set. But creating a big enough data set to train a deep neural network from scratch is very expensive. To accelerate the training process, the model was pretrained with a state-of-the-art (SOTA) model.

The model was first trained to capture relevant information from ImageNet, then fine-tuned on a small data set to learn specific information unique to the firm's specific problem (identifying roof damage). To facilitate the tagging process, the firm approached the challenge as a classification task. Instead of classifying the images pixel by pixel, which would have been very time-consuming, it classified small image patches.

This process is referred to as *transfer learning*: you take the lower levels of a model and use them as the basis to customize a model for your needs. In the computer vision space (where convolutional neural networks, or CNNs, are ubiquitous) this is how work is done. No one starts from scratch anymore; it's too time-consuming. Think about it: if you're trying to figure out if something is a cat, or a tile on a roof, the lowest level of the model is starting to identify shapes (edges, curves, and so on). Finding a half-torn roof shingle is no different from looking for a bent line. You're looking for a rectangle; an edge is an edge. At the lower levels, these models are all similar, so modeling doesn't have to start from scratch.

Today, this company can analyze thousands of images simultaneously. It can automatically identify the severity of roof damage, and evaluate other issues including water pooling, loose cables, and rust. The company was able to scale to meet demand, and dramatically increased its orders and thus revenues. The insurance adjuster example we described earlier in the book mimics this same scenario. There are many high-value AI use cases that transcend industry; you just change the data, and start with a pretrained model that you fine-tune for your specific application.

Automating the AI Lifecycle

Over time, any machine learning model becomes stale and needs to be retrained. Training models is a complex, computationally intensive task. It requires a highly tuned system with the right combination of software, drivers, compute power, memory, network, and storage resources. Retraining models that are already in production is also taxing for developers and data scientists, who would prefer to focus on doing what they do best—concentrating on data and its refinements, training new machine learning and deep learning models over these large data sets, and creating cutting-edge models.

Training a model the first couple of times to meet your performance objectives can certainly be exciting, but training the model 1,000 times to keep its performance intact is not. That's where automation comes in; it frees teams from managing the repetitive training process by hand, so they can do more productive and interesting work. New tools exist to manage the training lifecycle automatically. With these tools, each training run is automatically started, monitored, and stopped upon completion, saving users time and money as they only pay for the resources they use. (The IBM platform can even run experiments with various hyperparameter settings to find optimal settings, and if a training run starts with a nonoptimal setting and the model starts to overfit or underfit, the platform will exit the run early to avoid wasting computer resources and building a model that would be useless in production.) This is often called "deep learning as a service."

As we've seen, training doesn't just happen in the Build phase of the AI lifecycle; models grow stale and periodically need to be retrained. With Watson Machine Learning (WML), for example, when your model is ready to move into production, you can specify how frequently you would like to retrain it and automate the redeployment process. You can also monitor and validate the results of the retrained model to ensure that the new version is an improvement—and with integrated version control, you can easily roll back to a previous release if necessary. IBM also offers Watson Machine Learning Accelerator (WML-A), which is a combination of hardware and software designed to turbo-boost AI productivity and training.

These automation capabilities help reduce the need for IT teams to act as intermediaries in the deployment process, eliminating the biggest bottleneck for continuous improvement of machine learning models. They also place more power in the hands of data scientists, empowering them to focus on building and maintaining the most accurate models possible, instead of being forced to sacrifice quality for practicality. They enable collaboration between developers and operations staff—which is what DevOps is all about. Most importantly, these tools give your models the chance to do what they were always meant to do: learn. By continuously retraining your models against the latest data, you can ensure that they continue to reflect today's business realities, giving your organization the insight it needs to make smarter decisions and seize competitive advantage.

Let's look at a few other automation tools.

AutoAI

One of the best ways to simplify the on-ramp to AI proficiency is to remove some of the steps. Getting started quickly matters to both the data scientist looking at data sets and wanting to run experiments and build models, and to the AI practitioner who wants to reduce model development time from months or weeks to days or hours. The speed at which AI projects can progress depends on how quickly your organization can move through the AI lifecycle.

A variety of products and services have emerged to help with this, including H2O Driverless AI, AutoML (both Google and Microsoft have products with the same name), and IBM's AutoAI. These tools automate the tedious parts of the AI lifecycle, including preparing data, developing models, performing feature engineering, and tuning the hyperparameters of those models.

IBM's AutoAI focuses on data preparation, hyperparameter tuning, and feature engineering as well, but it also includes features such as transfer learning, neural architecture search, explainability, debiasing, and support for model deployment and management. AutoAI helps to prep the data coming in to an analysis problem, select the right model to use for machine learning and experimentation, and ensure consistency and repeatability are maintained throughout the entire process. AI projects typically require many experiments; keeping track of the models built in those experiments and ensuring that the results are repeatable and consistent is a difficult problem even for professional developers.

AutoAI works across clouds (for example, it can run on the IBM Cloud as well as private or competitors' clouds via the containerized IBM Cloud Pak for Data offering) and is available at the click of a button—another example of how the required investment in both computing power and expertise has diminished significantly. As IBM's Dinesh Nirmal says, "AutoAI helps you generate models with minimal knowledge of data science. You don't have to know anything about the models to get started. You can generate models in hours, rather than days or weeks." That's the real promise of AutoAI, regardless of the vendor: that people who aren't highly trained AI developers will be able to do experiments and create initial models.

These models may not be final, production-ready versions; they still need data scientists to tune them for performance (accuracy) and speed. That said, it's important to appreciate what this democratization accomplishes. It allows everyone in the company to have an idea, collect data, and build a model. That's how companies revolutionize their workplaces and build new business processes: they allow innovation to take place where the innovation is needed. Additionally, AutoAI provides Python code as a product, thus eliminating a black box tool and vendor lock-in. Automated AI allows everyone to experiment with new ideas and new processes, and when that happens you're well on your way to building the AI-enabled company of the future.

Example: Wunderman Thompson delivers new prospective customers through AutoAI

Wunderman Thompson is a data and technology agency that delivers insights into customer behavior based on an extensive data set that covers about 270 million individuals, with roughly 17,000 features.

For more than 20 years, Wunderman Thompson has been developing customer-prospecting models for its clients, but has found it difficult to take full advantage of its massive data sets. The company spent the last year building a data lake that gave a unified view into all of its data, but it didn't have the capability to feed this much signal through its existing machine learning infrastructure and practices. The data scientists knew that feature engineering would be beneficial to their model, but they also knew it would require a huge amount of manual effort and domain knowledge.

To get over this hurdle, Wunderman Thompson partnered with IBM to create a machine learning practice capable of discovering new customer insights and delivering increased ROI for its clients. Using Watson Studio, the company was able to amplify its Modeling-as-a-Service capabilities and leverage its entire portfolio of data assets, building models based on all of the available data sources. Bringing cloud-scale computing to the problem enabled the data scientists to train their AI applications against full populations and allowed far more features to survive into training. In turn, this led to models that performed better than anything they had ever seen.

Then, using a combination of AutoAI within Watson Studio and open source tools, they built a multicloud machine learning pipeline to generate predictions at a high volume. Armed with these improved models, Wunderman Thompson is enabling brands to find new customers at unprecedented scale. The company is already seeing incredible results from the live rollout of this project.

Example: Bank uses humans, machine learning, and deep learning to measure model risk

A leading US bank's quantitative research (QR) team was building models for traders across the institution to make critical trading decisions. The bank wanted to evaluate the risks associated with these models. To do this, the QR team manually evaluated the usage restrictions for each model. They flagged potentially risky ones, which were manually reviewed and resolved in a multistep process. This process was slow, tedious, and ultimately costly, as the persons doing the review were highly qualified and highly paid "quants." By using Watson Studio, the bank was able to leverage powerful GPUs and optimized software and storage to accelerate training of deep learning models for time series data to predict risk exposure. This allowed the firm to accelerate the review process for alerts using machine learning–based autoclassification, and to monitor and contain risk by detecting outliers with deep learning–based risk exposure predictions.

NeuNetS

In 2018 IBM released NeuNetS, a new capability that addressed the skills gap for developing AI models for a wide range of business domains. NeuNetS uses AI to automatically synthesize deep neural network models, quickly scaling up the adoption of AI by companies and SMEs. By fully automating AI model development and deployment, NeuNetS allows nonexpert users to build neural networks for specific tasks and data sets in a fraction of the time it would normally take, without sacrificing accuracy.

NeuNetS algorithms are designed to create new neural network models without reusing pretrained models. This allows for the exploration of a wide space of network architecture configurations while fine-tuning the model for the specific user-provided data set.

Since the beta release of NeuNetS in 2018, IBM has received a lot of positive feedback from the Watson Studio user community. This inspired the evolution of NeuNetS into its next phase: merging NeuNetS into AutoAI. This integration allows for the automation and enhancement of a new suite of neural network models, and it addresses much broader data science use cases, including tabular data prediction and classification, time series forecasting, text mining, and image recognition. In 2020, we expect that AutoAI users will start to benefit from NeuNetS in a growing number of use cases.

Incorporating AI into DevOps Processes

The AI lifecycle doesn't take place in a vacuum. Your company no doubt has myriad applications running: all the software that it takes to run any modern business, from customer-facing web applications to payroll. And it no doubt has teams of people who manage and support these applications, and manage their own complex lifecycles. It's a mistake to think that AI projects can (or should) make it to production without their support.

In the previous chapter, we introduced DataOps—the people responsible for managing data infrastructure and managing AI applications as they move to production. DataOps interface with the teams that support the company's existing software applications. In many cases, that means DevOps. Whether DevOps is a process, a culture, a group, or a movement is controversial (DevOps has resisted definition)—and, ultimately, unimportant. What is important is that it's how production works at many companies, and that it has a strong focus on automating software lifecycles.

The origins of DevOps are in the rise of the web, which forced many organizations to rethink IT operations. Processes that worked in the 1980s and early 1990s, when software was shipped on physical tapes or disk drives, no longer made sense when value was delivered online and could change at a moment's notice. The result was a movement, or paradigm, that grew out of the early web powerhouses like Flickr and Yahoo! and spread through the rest of the IT world. The goal of DevOps was to remove the barriers between software developers and operations staff; it wasn't acceptable for developers to throw a piece of software "over the wall" and expect IT to take it from there. Whether or not DevOps has succeeded is subject to question (as is the nature of DevOps itself); we still have developers, we still have operations teams, and they aren't the same. But development and operations teams certainly spend a lot more time communicating than they used to, and that's progress.

Although DevOps has resisted definition, there are some very clear principles behind it. Let's take a look at those principles, how AI challenges them, and how those challenges might be resolved:

Automation of the build, integration, and deployment processes

This is frequently called "continuous integration/continuous deployment," or CI/CD for short. The slogan "Infrastructure as code" says a lot. If you are deploying applications by hand, even simple ones, you will get it wrong—perhaps even frequently. Manual deployment is fertile ground for mistakes that will bring your infrastructure down at one point or another. And what's more, these mistakes have a habit of not just happening once. To deploy reliably, you need to automate the process. As we said earlier in this chapter, that's exactly what we want: an automated lifecycle for AI.

So where's the problem? In continuous deployment, the cycle is kicked off when a developer checks code into a source code repository such as GitHub. That triggers an automatic build, automatic execution of test suites, and possibly (if all the tests pass) automated deployment to production. Companies that do this successfully, including all of the major web enterprises, release mission-critical application components many times—potentially hundreds, maybe even thousands—a day. But the releases are tied to code changes, and with AI, the product can change without a change to the code. For example, if you change the training data (perhaps to augment it, perhaps to rectify biases) and retrain the model, you change the model without touching the code. So how do you trigger the deployment process?

Testing

In a DevOps world, testing is integrated into the deployment process. You can't deploy without running through the entire test suite first. That kind of testing discipline is needed in AI, and it's not something AI developers are familiar with. They're more familiar with test and validation *data sets*, which are critical for AI success, but different from the kind of testing that operations groups routinely practice for traditional applications.

But there's a deeper problem. The tests you find in a CI/CD pipeline are deterministic. If you have inputs A and B, you expect output C. If you don't get output C, the test fails. Once you add machine learning to the application, some of these outputs will be statistical, rather than deterministic; you won't always get C, and that's not an indication of failure. Rather than pass/fail tests, you should select an accuracy metric (such as an F1 score), define a target for that metric, and evolve the product to meet that target. This way, you aren't asking your machine learning components to produce deterministic output, but you are defining acceptable levels of performance and asking the models to deliver that performance repeatedly.

Rapid iteration

Companies that are successful with DevOps deploy hundreds of times a day—and that is not a joke. But how will you do that if your test criteria are statistical, meaning that you have to run some of the tests many times to get a statistically valid result? And what exactly constitutes the build process, which is part of the cycle? Compiling software takes long enough, but training and retraining models can take hours or days. You may need to reengineer the process in a way that limits how often models are retrained, or that does training in a separate build cycle.

The point of rapid iteration is to deliver products to the customer that are continuously and consistently better. The development team usually defines improvement metrics for each "sprint," which typically lasts one or two weeks. Machine learning models complicate this process because they are statistical. Assume that a model starts with 80% accuracy, and the target is to improve by 1 percentage point per week for 10 weeks. What if other components (say, the web interface) meet this goal, but the AI components don't? Can the user accept improvement to parts of the application, without receiving improvements to the AI? If the improvement targets are too strict for the models, and too many builds are "failing," you may be missing opportunities to deploy features that the user wants (and needs). But if the user really needs improved AI, delivering a better web interface isn't the point. Should the continuous deployment process wait for machine learning models to "catch up" and attain this week's accuracy threshold? Or should it release new updates without improved machine learning? Which alternative better serves the customer? This is an important decision, and it's not a simple choice to make.

Monitoring

Application monitoring is a cornerstone of DevOps, and application monitoring is essential to AI. However, AI requires different kinds of metrics, and these are tied to its statistical nature. You can let a traditional application run for years, and while it might not have some features you like, it will still do what it's supposed to do. That's not true of AI. How do you detect staleness, fairness, bias, and drift, among other issues? And of course, AI applications need traditional monitoring too—for availability, response, server load, and everything that any other application needs.

Monitoring is evolving to "observability": rather than simply logging key parameters, we want to be able to observe the application's state. Just remember that the state of an AI application may include millions and millions of parameters, and be a significant challenge to current notions of observability.

Movement to the cloud

While DevOps isn't inherently cloud-based, in any organization it will be the operations staff that understands (and is affected by) moving applications to the

> cloud. AI applications will be developed, trained, and deployed in the cloud; their data will live in the cloud. Remember that the cloud is a capability, not a destination, and that most companies will be using hybrid multicloud solutions spread across several public cloud providers and their own data centers. AI doesn't present any specific challenges to cloud operations, but as your AI projects move toward deployment, be sure to take advantage of the people who are already working with cloud infrastructure.

So, what are we saying? DevOps is a nice idea, but forget integrating AI into the process? Not at all. We are saying that the goals of DevOps and the goals of the AI lifecycle are similar. AI simply challenges the implementation of DevOps, and will require some rethinking—but these are solvable problems.

Operations staff need to become your allies. In most cases, they'll be interested in AI and willing to learn. They may even be using some AI tools of their own, for tasks like capacity planning, intrusion detection, and log file analysis—and if they aren't, those are possible options for your initial AI projects. But you will need to meet them halfway. Although the DataOps team members will probably be the ones to work most closely with the operations groups, remember that DevOps is predicated on reducing the barrier between "developers" and "operations." Learn from them, understand their concerns, and be creative in solving the problems you will face together. A great AI product that never makes it into production doesn't do anyone any good.

Throughout this book, we've been arguing for starting a large number of AI projects, each of which can succeed (or fail) quickly. That's a mindset that fits well with DevOps. Take advantage of it.

Emphasizing Trust and Transparency

Because AI models function as black boxes, producing predictions that human experts may not be able to explain and perceiving patterns that humans typically can't detect, people are understandably wary—the technology seems so powerful, yet so mysterious. People need to believe that models and algorithms do not have built-in biases or faulty logic. As a result, organizations must ensure their AI models and systems are designed to produce results that can be explained and documented. For example, a bank needs to be able to tell a consumer what the factors were behind their loan application being denied (it's the law in the European Union, as per Article 14 in the GDPR), and what the consumer would need to do to change that decision.

An organization will never truly be able to scale its AI across its workflows without first establishing trust and transparency. Here are some examples from different areas showcasing why companies need visibility into what their AI is doing:

Insurance

Insurance underwriters can use AI to make more consistent and accurate underwriting or risk assessments around claims, ensure fairer outcomes for customers, and explain AI model recommendations for regulatory and business intelligence purposes. Inability to explain risk assessment and other decisions exposes the underwriter to legal liability, particularly if assessments for different subgroups vary. And whether management trusts the AI or not, inaccurate recommendations could lead to huge losses—in which case, the AI won't be trusted for long.

Telecommunications

Data scientists can build AI models and work with their IT operations teams to confidently recommend proactive asset maintenance for telecommunications infrastructure. Predictive maintenance is a great application for AI—but it's a game changer if the AI can tell you not just that something is going to fail, but why. And again, there are liability issues: telecommunications often plays a critical role in disaster management and recovery, and outages aren't taken lightly. (The COVID-19 crisis has put tremendous pressure on these companies, who are often providers of Internet connectivity as well. Using AI to understand and manage load will certainly be something to look at. For example, during the height of the COVID-19 crisis, the third season of *Ozark* debuted on Netflix, forcing Netflix to occasionally downgrade the quality of the video stream due to demand and reduced bandwidth.)

Healthcare

Organizations in the healthcare industry must maintain regulatory compliance by tracing and explaining AI model decisions across workflows, and intelligently detect and correct bias to improve outcomes. It's vital for healthcare AI to be able to explain why it is making decisions, whether they're about treating patients or managing medical staff.

Fraud

Fraud activities have increased at a rapid pace over the years. Fraud is difficult to detect. That's because the data used in fraud detection tasks is complex and often unstructured. And if data is siloed, it can be difficult for businesses to get a 360-degree view of it, leading to false positives or missed alerts and potentially costing companies hundreds of millions of dollars. But you won't make friends by accusing innocent people of fraud; management must trust AI to predict fraud correctly, with few false positives, and to explain exactly why it considers any transaction fraudulent.

Consider these use cases, the use cases we didn't have space to mention, and the use cases you come up with on your own. Now imagine how much better our world could become with AI that is trustworthy and transparent. Better visibility into what your AI is doing could translate to longer, healthier lives, better support during national crisis management, better revenues (more jobs), and more. Fraud, for example, creates enormous costs that are ultimately borne by consumers; what's more, the saddening spike in human trafficking is made possible because of fraud (money laundering). With AI, our world can indeed become a whole lot better.

Organizations can engage with AI to help them handle issues like these with predictive insights, real-time analysis, more sophisticated modeling techniques, and automation technologies—and by investing in a hybrid multicloud platform, they can accomplish all of this in an agile, governed, and secure environment.

Example: By Shining Light on Data Attributes, a Bank's AI System Demonstrates Integrity, Fairness, Explainability, and Resiliency

"AI is all about trust," says Kelly Combs, Director of Emerging Technology Risk at global consulting firm KPMG. Her expertise encompasses robotic process automation (RPA), AI, and how to responsibly govern, scale, and manage technologies under the umbrella of intelligent automation. Combs, who was recently recognized as an "IBM Woman Leader in AI," believes the reason that only 5% of KPMG's clients are heavily adopting AI is because AI is still viewed as a black box. She aims to vanquish that view, by shifting to systems that are transparent and explainable.

Combs believes that AI needs to possess four trust attributes: integrity, fairness, explainability, and resilience. When working with a banking client, KPMG had to put governance first, as well as satisfying these four imperatives. The bank used AI models that depended on data attributes such as FICO scores (FICO is a US data analytics company focused on credit scoring), income levels, and age to help make predictions about whether to approve or deny loans or lines of credit. But attempting to apply governance to these models manually was not working. It was simply not possible to understand when models began to "drift," or favor certain outcomes or individuals, thereby violating the imperative of fairness. To solve this problem, KPMG turned to Watson OpenScale to build and scale a transparent and explainable AI system. By providing insight into which data attributes were contributing to the decision making of an AI model, KPMG was able to help the client better understand unconscious bias, be more proactive in response to regulatory change, and provide the trust needed to drive AI adoption.

A similar scenario with a different approach and outcome played out when Apple released its first credit card in 2019. Some high-profile users of the card—the creator of Ruby on Rails and, ironically, a cofounder of Apple—noticed that their spouses were getting up to 10× less credit despite the significant amount of jointly titled assets (in one case all the assets were jointly held). As negative publicity surfaced, Apple couldn't describe how the algorithms that determined credit lines even worked, let alone justify the output. Goldman Sachs (the issuing bank for the credit card) insisted there was no bias, yet couldn't offer proof. Weeks later the bank declared the algorithm couldn't be biased because a third party had vetted it and noted it didn't take gender as an input. It asked publicly, "How can the algorithm be biased if we don't ask for your gender?" And therein lie several problems. First, it's certainly possible for an algorithm to demonstrate bias because of the training data; dominate the training and validation test sets with men, and you've just created unintentional bias! Second, features can serve as proxies for other features (like address for race), so whether the algorithm takes gender (which could be inferred by first name with relative accuracy) as an input isn't relevant (hence why we are spending so much time on sources of bias in this book). While the features used to make credit decisions remain proprietary, it's also a good bet that they include purchase histories and many other kinds of data that can be used as a proxy for gender—and AI is incredibly good at finding proxies.

Here's our pro tip: just because you remove a feature like gender doesn't mean you've removed the bias. Apple and Goldman Sachs eventually found themselves the subject of a New York Department of Financial Services investigation citing "Any discrimination, intentional or not, violates New York law." More damaging was the publicity. If transparency, fairness, and explainability had been priorities from the start, Apple and Goldman Sachs wouldn't have been looking at bad publicity or a compliance investigation. It's important to us that you don't misunderstand our point. Apple or Goldman Sachs aren't in the business of purposely building biased algorithms, and perhaps their credit-scoring algorithm isn't biased at all. But the publicity was fast and furious, as was the PR trust issue. That's our point: make it explainable from the start so you can put the right models into place, but defend them by understanding their inner workings as well.

Example: Avoiding the "Black Box" Dilemma

Forecasting (predicting) is one of the top applications of machine learning. By using statistical models and AI algorithms, machine learning is able to help companies predict possible outcomes and trends. Oursky, a Hong Kong–based AI app developer, was commissioned to apply machine learning to predict the results of NBA games. The firm's work provides a great example of how to build an explainable model that makes accurate predictions.

Oursky had two goals: first, to develop a machine learning algorithm to predict the winning and losing probabilities for each NBA team, and second, to ensure it wasn't just a black box algorithm. Humans had to be able to interpret the reasons for the model's predictions. The training data came from historical NBA records, with players' logs of games dating back to 1983 (the first year for which complete box scores were available) downloaded from *basketball-reference.com*. The data was cleaned and the model was tested with three sets of initialization parameters on a subset of the data from 2013 to 2017. Instead of just providing predictions of wins and losses and scores, Oursky used the SHapley Additive exPlanations (SHAP) framework, which avoided the "black box" problem by giving explanations for the model's predictions.

The result? The model was more accurate than predictions published on FiveThirtyEight (a website that focuses on opinion poll analysis), both when testing against previous NBA seasons (with a success rate over 67%) and in production against the current season (success rate over 80%). The firm's client also was pleased that it hadn't just produced another "black box" oracle; the model provided satisfactory explanations and reasoning for its predictions.

Avoiding the Piecemeal Approach

It's possible, and indeed easy, to add AI capabilities to existing operations and software applications in an ad hoc, piecemeal fashion. You can, for example, use simple APIs to access cloud-based services and build voice recognition or language translation capabilities into your existing programs. But harvesting the full power of AI requires more than an ad hoc approach; you need to be able to analyze your own AI projects and establish clear metrics for success and failure, so that your organization can learn from its failures and continually improve. As we've seen, AI systems have their own lifecycle, not unlike the well-understood software development lifecycle. But since AI is a new discipline, some of the necessary steps will be unfamiliar and may be resisted by parts of your organization.

Ad hoc integration of AI tools into existing processes by grassroots teams is natural, and some measure of it is probably unavoidable. But successful organizations put in place an end-to-end AI lifecycle to build, run, and manage their AI applications. Adherence to an orderly process ensures high-quality data, efficient development of models and algorithms, and robust, stable deployment environments. And most importantly, only a well-run AI lifecycle can provide trust, explainability, and compliance with corporate and regulatory policy.

Example: SaaS Company Gleans New Insights by Applying AI to Historical Data

Xactly is a leading provider of enterprise-class, cloud-based, incentive compensation solutions. Its services help companies design the right incentives to motivate sales employees and managers to excel and drive excellence in their businesses. But this was a crowded field, and Xactly was having trouble differentiating itself. Over the years, the firm had collected data on hundreds of companies, detailing their incentive practices and related outcomes. It knew this data was valuable, but wasn't sure how to go about exploring and analyzing it.

With the help of TIBCO, Xactly was able to convert the data it had collected on compensation plans for more than 300,000 sales professionals into actionable information, with a machine learning model helping surface patterns in the data. The company found that women perform better in sales overall, but their overall compensation is lower than that received by men in the same roles. This finding made quite a splash in the benefits industry, and was used to support an industry report on gender pay inequality that went viral. Moreover, after identifying this fact, the company did the same analysis on data from its own sales organization and found that some of its female employees were being underpaid. It promptly changed that, and the machine learning model will continue to be used internally to make sure the firm achieves its gender pay equality goals.

The AI project also shot Xactly into the ranks of industry thought leadership. Today, Xactly partners with some of the biggest companies in the sales compensation space, such as Gartner, MHI, and Sales Globe, and is turned to by the media as an expert in its field for comments and analyses on benefits and compensation trends.

Ready to Infuse...

You might have the most amazing model in the world, one that yields spectacular results—but the number of great models that never make it to production is nothing short of shocking. Don't fail at this stage. To get your models into production, you need to:

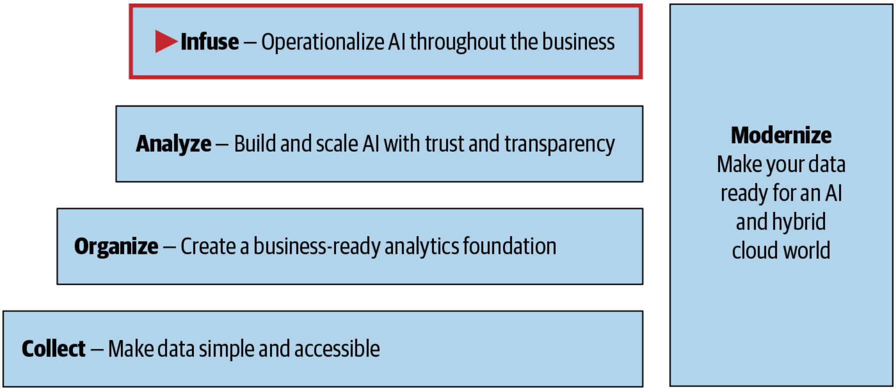
- Ensure that your results can be trusted and explained. We've repeated this so many times throughout the book, you may be beginning to wonder if our copyeditor missed something—she didn't. Never forget: if your data isn't trustworthy, why should anyone trust the results computed from that data?
- Take advantage of tools that automate model building, data cleaning, and other key processes. They will reduce your need to hire AI specialists.

Your AI transformation will go nowhere if you can't deploy what you produce. But you can deploy, if you're properly prepared. Understanding the AI lifecycle will prepare you to infuse AI projects throughout your business.

CHAPTER 9

Infuse AI Throughout the Business

You should be thinking about how you're going to infuse AI throughout your organization from day one. That's one reason to start with a large number of small projects, rather than a single big one: many small successes will mean much more to stakeholders than one big success, and certainly much more than one big failure. We're telling you right now, some of your new projects will fail—that's true in any area, but particularly so when you're taking on a new technology. If you're going to take advantage of AI, you need to use it everywhere; you need to push it into every department, every business process, every activity, and make your workflows intelligent. We've arrived at the Infuse rung of the AI Ladder (Figure 9-1).



▶ **Infuse** – Operationalize AI throughout the business
Analyze – Build and scale AI with trust and transparency
Organize – Create a business-ready analytics foundation
Collect – Make data simple and accessible
Modernize Make your data ready for an AI and hybrid cloud world

Figure 9-1. The Infuse rung of the AI Ladder

C-suite executives are turning to AI because in order to compete, they need to be able to innovate at speed. Their goals include:

- Predicting and shaping future outcomes
- Humanizing customer experiences
- Empowering people to focus on higher-value work
- Supporting human capital in their efforts to reimagine new business models by infusing intelligence into their workflows

For many organizations, the best strategy to achieve these goals is to start at the top of the AI Ladder. That sounds counterintuitive, but remember we noted earlier that the ladder isn't perfectly linear; it's possible to start at the top (or on any rung, really), pushing AI through the organization, and then go back to work on collecting, organizing, and analyzing your data. Here's the trick: you don't have to start with nothing. You can start with prebuilt AI applications that can easily be adapted to your business situation.

For an AI transformation to succeed, you have to infuse intelligence across all of your workflows. And prebuilt applications are less likely to be trapped in the limbo between the lab and production.

The work we've done with clients has identified five key business domains to which C-level executives are applying AI to transform their workflows. They are:

1. Customer service
2. Financial operations
3. Risk and compliance
4. IT operations
5. Business operations

That being said, innovation is rapid, and whether you go beyond these five domains is up to you. For now, let's dive into this selection.

Customer Service

Business leaders are looking to increase customer satisfaction and reduce attrition, while cutting costs. They want to be able to deliver a quick and consistent customer service experience across all touchpoints, while also helping automate customer support operations. You're probably familiar with the 80/20 law, also called the Pareto principle: 80% of your company's revenue comes from 20% of your existing customer base. If you work out the statistics, this phenomenon gives you a lot of leverage. A small increase in customer retention can have a much larger impact on your business, particularly as the cost of acquiring new customers grows; an article in Harvard Business School's Working Knowledge series argues that a 5% increase in retention can lead to a 25% to 95% increase in profitability. Improving customer service to increase retention is a priority.

How can AI impact the customer service domain? Conversational assistants are a good place to start. Many products, including Watson Assistant, allow you to build a customer service agent with minimal expertise. According to IBM Distinguished Engineer Bill Higgins, "They're already pretrained with common phrases for different industries, and require little customization to adapt them for a specific business's needs."

One of the great successes of AI has been in the automating of call center activities. In Chapter 2, we saw how Humana improved customer satisfaction and its operations by implementing a conversational assistant to handle and route calls from medical care providers. Rather than forcing customers to use the old-style interactive voice response systems that we've all come to hate ("Press * if you are about to scream"), Humana created an intelligent assistant that was able to handle most of its clients' problems successfully, and without being annoying. Over 60% of the calls are routine questions about insurance coverage that can now be handled automatically. Problems that the assistant can't solve are routed to humans. The result? Clients get the information they need more quickly, simple rote question answering is reduced, and the organization has experienced both higher customer satisfaction and lower costs.

Creating an intelligent assistant didn't require an in-house staff of AI experts; it required Watson Assistant for Voice Interaction, plus the expertise of in-house business SMEs who knew what kinds of questions the automated assistant would receive and how they should be handled. Your organization will also have SMEs who know the details of your business, the competitive environment, and everything else. Pairing the expertise you already have with a prebuilt AI solution gives you the benefits of AI even before you've built your own AI development capability.

There's an exceptional (and eye-opening) book around customer service called *The Effortless Experience*, by Matthew Dixon, Nick Toman, and Rick Delisi. We consider it a must-read. Not only will it completely change the way you think about customer service (the authors assert that delighting customers does very little for customer loyalty and repeat business), it will make you understand just how important this section (and this book) truly are. Without spoiling it, the book asserts that if you make life easy for your customers (think assistants who can resolve most of a customer's reasons for contacting customer service in the first place and save them from having to switch channels—like having to call in because the bot can't solve the issue, or getting transferred), your clients are more likely to stay with you and buy again. Like we said...a must-read.

Financial Operations

Financial operations are essential to any business. But how do you know where the puck is heading without good planning and forecasting capabilities? As hockey great Wayne Gretzky famously said (and we know it's one of the most overused corporate cliches ever, but it's really applicable here), "Skate to where the puck is going to be, not where it has been." If you don't know where the puck is heading—if your business's ability to make plans and forecast results isn't up to par—you won't know where to skate. The better your forecasting ability, the more quickly you'll be able to adapt to changing markets and business conditions. Agility is everything, and AI gives you the ability to be agile.

Your finance team can't be agile with a manual, error-prone, and fragmented financial planning and analysis (FP&A) process. To ensure you're prepared to react to market volatility in real time, you need the ability to synthesize information from all available data sources, uncover trends, and deliver insights faster than ever. An automated, flexible planning solution helps finance leaders focus on driving business forward, instead of being stuck looking back. Those who utilize more flexible planning solutions can attain 50% faster forecasting and are 50% more likely to have the ability to conduct "What if?" analyses, critical for testing different scenarios before implementing changes to their plans.

AI can provide sound insights for financial planning, which needs to understand both trends in financial markets and customer behavior. Prebuilt planning tools can be used throughout the company, not just in finance: in HR for planning staffing levels and analyzing compensation, in operations for demand and inventory management, in marketing for analyzing customer profitability and planning successful promotions, and in sales for everything from planning quotas to building forecasts.

As an example, Continental Foods is a leading manufacturer of soups and sauces in several European countries. This business is highly subject to local taste—and making products that fit local tastes leads to significant planning and stocking problems, because each region needs its own set of products. Continental was able to use data and AI to generate sales data and predictions for every product in every one of its markets—literally every country and region in Europe. With these forecasts, the firm was able to analyze which products were selling and why, and adjust its production and marketing decisions accordingly. Continental was already in a strong position; AI allowed the organization to use its corporate advantage to drive even higher returns and greater efficiencies.

Risk and Compliance

Keeping pace with regulatory requirements and managing risk is a top priority for all business leaders. With the assistance of AI, you can create intelligent workflows that are optimized for your organization, improve data quality through assisted data classification and categorization, and save your risk and compliance team time.

AI can be a big help for companies that provide advice on compliance with tax and banking regulations. It's been estimated that over 300 million pages of banking regulations exist worldwide, with $100 billion spent every year on compliance and about $150 billion on fines and penalties. In the US, there are 74,000 pages of tax law, with thousands of them being changed yearly. This makes the tax code incredibly complex for even the savviest tax preparer. Humans will always need to understand the tax code, but what better tool to help us with that than an AI that can tirelessly read thousands of pages of regulations without forgetting or ignoring any detail? KPMG used IBM Watson, trained on over 10,000 tax documents, to build an application that makes the correct recommendations in 3 out of 4 cases. More importantly, the people who trained the model were KPMG's "tax professionals," or SMEs. The result? Employees are more productive, they can analyze cases more rapidly, and they are more confident in their results because Watson provides an audit trail.

Prebuilt AI products also exist for tasks like fraud detection, ranging from credit card fraud to money laundering. Fraud detection involves tracking many data sources: customer data, real-time transaction data, IDs of devices processing transactions, alerts, and more. Using AI, Infinity Property and Casualty was able to reduce the time needed to detect fraudulent claims from 14 days to 24 hours, and another global financial firm improved its ability to detect and interrupt fraudulent transactions by 40%. That's important because the sooner you spot a fraudulent transaction, the easier it is to recover. Fraud detection in real time is the holy grail, and that's only conceivable with AI. You won't get there by hiring more auditors; human auditing will always be behind the curve because it's reactive rather than proactive.

So where are human auditors needed in this process? This is where collaboration between humans and AI becomes important. Fraud is a rapidly changing world; criminals are getting more creative all the time (and you can be sure that they're using AI to get better at what they do, too). To address this, IBM's Paul Govoni stresses the importance of model governance, and of challenging the models used to detect fraud. The assumptions made in the design process, and whether the models are adequate for real-life situations, need to be tested. The result will be a fraud detection system that can be proactive, rather than simply reactive. Detecting and preventing fraud isn't quite like finding money, but it's close—and if you're a financial institution, the amounts involved can be large.

Any company can become the victim of fraud; AI can keep you from becoming one of the victims.

IT Operations

IT groups will inevitably be responsible for running your AI applications. But IT can also use AI to make the IT team itself more effective. (We call these the "AI for IT" use cases.)

IT can use prebuilt applications to monitor logs, detect intrusions, and even predict or detect component failure. One of the biggest problems in IT is monitoring logs effectively. Intrusions and data breaches frequently aren't discovered until months (or in some cases years) after they've taken place—and those breaches often have severe consequences. But with some server farms spitting out terabytes of log files per day, and hundreds if not thousands of servers to watch, how is a human to detect a breach? Even with sophisticated home-grown scripts written in Python or some other language to analyze the logs, are you willing to bet that you'll detect a breach in a timely way? Just as AI can look through thousands of legal documents or tax regulations, it can watch log files for patterns that suggest intrusions, failures, or other events, and raise alarms, long before a human could. In IT, time is rarely on your side. AI can detect problems before they become big problems.

IT can also use AI for capacity planning. Online retailers can experience 100× traffic surges for special events like "Black Friday" (the Friday following Thanksgiving Day in the United States—the biggest retail shopping day of the year). If your IT group can't handle the traffic, you're sunk; and if you can't predict the traffic that needs to be handled, you will either fail to handle it or overspend on excess capacity. Either way, you lose, and you're likely to lose big. On the other hand, if you can predict the capacity you need for such an event accurately, and handle that traffic with a cloud or multicloud solution, you win. That's what AI enables: it can monitor many dimensions of incoming information (current customer demand, historical data, economic trends, internet memes) to deliver the predictions you need to win.

Business Operations

There are many opportunities to profit from AI in business operations. To give some examples, by infusing AI into its business operations your organization can:

- Transform how it maintains industrial assets.
- Optimize how it operates real estate assets.
- Build intelligent, self-correcting supply chains.

- Streamline how it engineers industrial products.
- Inform decisions with weather-infused insights.

This translates to benefits like reduced operating costs and increased uptime and availability, and gives organizations the ability to deliver outstanding client experiences with every order. We'll stick with two examples: buildings and supply chains.

Operating a building is an expensive proposition. Operations account for 71% of a building's cost over its lifespan, and buildings consume 42% of all electricity. Heating, air conditioning, utilities, and other expenses add up to large numbers in most real estate site operations (RESO) budgets. Many companies have used AI to minimize HVAC (heating, ventilation, and air conditioning) costs, including companies operating some of the largest data centers (huge consumers of HVAC and water) in the world. IBM's Cognitive Buildings strategy has been able to reduce electricity use by up to 50% in office buildings, in addition to helping managers think about how employees use alternate workspaces (like coworking facilities or working from home). These are real savings—even for a small company, optimizing the cost of running a building can add up to millions of dollars, and for a large enterprise it can save tens or hundreds of millions. And that doesn't take the environmental benefits into account; every dollar not spent on heating is a dollar that isn't generating greenhouse gases.

Supply chains are notoriously complex; they're not unlike massive multiplayer online games, except that they're played in the real world, with real products, supplies, containers, ships, and money. It's not just a matter of ordering the right supplies and hoping that they come; you have to ensure that they come on time, that what arrives is what you actually ordered, and that you comply with import/export regulations (including customs duties). Almost anything you can do to improve supply chain management (SCM) will have a big impact on productivity.

IBM Sterling Supply Chain combines AI, blockchain, and cloud technologies, enabling it to do far more than simple planning; it can also track the products you need to guarantee their provenance, detect disruptions in the supply, dynamically update plans to work around problems, and trace supplies as they move through the value chain.

Anheuser-Busch uses Sterling Supply Chain to simplify its supply chains, standardizing and reducing complexity wherever possible and finding solutions where it isn't. As we said, supply chains can get incredibly complex. If you make beer, you can predict big spikes in consumption around major athletic events; are you boosting those predictions with weather data, local buying demographics, and trending sentiment about players on the teams involved, the product they sponsor, and the crowd's sentiment toward the players themselves? Of course to fulfill predicted demand, you need to have the product in stores beforehand. That means you have to have the beer in stock, preferably fresh. But beer doesn't just appear on a packaging line: it has to spend time fermenting, so you need to ramp up your breweries months in advance. And in turn, those breweries rely on grains and other agricultural products that are subject to growing seasons and fickle supply variability because of factors like drought, disease, and regulation. Keeping everything flowing is a huge task. It's all too easy to slip up—to miss news of some plant disease, or of shipping delays that might force you to go to other suppliers, or changing economics that send your plans out of whack.

These are all problems that Anheuser-Busch faces daily, but the company has found that using AI puts power back into the hands of the business users. AI takes the routine work of monitoring spreadsheets, projections, and invoices out of supply chain managers' hands. It alerts them when something is turning into a problem—possibly a problem that won't hit until sometime in the future—and lets them spend their valuable time making the decisions that matter. This is known as "managing by exception," and it has a real effect on the bottom line.

Themes Across All Intelligent Workflows

To be successful in creating intelligent workflows, you need to give your team the right tools so that they can:

Visualize their data
: BI tools like Tableau and Cognos Analytics help with visualization and can enable you and your staff to build self-service and governed dashboards that show status at a glance.

Discover, distill, and understand that data
: Executives live in a sea of data and documents; the ability to find, digest, and understand those documents is a superpower. IBM's Watson Discovery service uses natural language processing techniques to find relevant information, whatever its format or location.

Manage risk and keep pace with laws and regulatory compliances
: Every workflow has risk; you need to maintain compliance with laws, protect against fraud, and react quickly to changes in the market. OpenPages with Watson provides services for monitoring operational risk, compliance, financial controls, and even performing internal audits.

Adding AI to your workflows allows you to predict and shape future outcomes, visualize decision making, and automate processes. You can move away from siloed, static functions and unify people, data, and technology.

There's an old saying in business: "Don't build what you can buy." Especially given the shortage of AI experts and data scientists, going with a prebuilt application can make a lot of sense. It isn't completely pain-free (for example, you'll still need to train an AI chatbot for the conversations you expect), but it's a relatively effortless way to jump-start your journey to AI. Save the effort of building your own AI systems for areas that are so closely tied to your specific business needs that there's no other solution.

When you purchase prebuilt applications, questions you need to ask your vendor include "Who owns the data?" and "What are you going to do with this data beyond what I'm asking you to do for me with it?" Get ready for some potentially shocking replies! IBM never uses its clients' data to train its algorithms, but that practice isn't common in the industry. Last we looked, some well known competitors follow a very different approach; they use your data to better train their own algorithms, then sell the models trained on your data back to you and to your competitors. Many companies would consider this highly undesirable, as your biggest differentiator is probably your data.

Building the Next-Generation C-Suite

Give your teams the fuel to fulfill your organization's ambition. The next generation of successful C-suite executives will be infusing AI across the businesses they lead to enable a more predictive, digitally automated enterprise.

Infusing AI is all about pushing AI through your entire organization, not just a few groups. You're not just changing a process here and there; you're reinventing the company, and building tools for the next-generation C-suite. These are the tools that will enable your company to take on its challenges and prosper in the future. Here are a few things to keep in mind as you work to infuse AI throughout your organization:

- Start with a business problem you have, not an AI project you want to do. Whether you're just starting on your AI journey or you're well underway, always come back to the core business problems you're trying to solve. Focus on your company's needs and pain points.
- Think about how AI can help build intelligent workflows around customer service, risk and compliance, planning and forecasting, IT, and operations. In doing so, you're building the next generation C-suite: a set of tools for your company that will enable it to meet future challenges.
- Remember that you can't have AI without IA. Organizations need a modern information architecture (the IA part) to connect data to all available sources, to make it accessible to users and teams, to build and deploy AI models dynamically, and to simplify and unify data and AI services across cloud environments. Understanding and using the AI Ladder will help your organization build an information architecture and ultimately reach your goals.

- Realize that AI is not magic. It's work. It's science. It requires proper tools, methodologies, and mindsets. If you have these, you'll be able to overcome the gaps in data, skills, and trust that prevent many companies from embracing AI.

At this stage, you've won most of your battles. Now it's time to finish the job. A small cadre of AI enthusiasts isn't going to change your company. Careful here: we're not advising AI for AI's sake, or even for your career's sake. But if your company misses out on AI, it's likely to miss out in much more important ways, as it gradually loses its ability to compete and remain relevant in the marketplace. AI doesn't exist so others can marvel at its technological profundity. AI's purpose is to solve real-world business problems; to predict, optimize, and automate your processes and make your business and your people more effective. The hitch? You have to develop the ability to realize these advantages. Remember, AI isn't going to replace managers. Rather, the managers who use AI are going to replace those who don't.

We'll be blunt yet again: AI is the biggest opportunity of our time. That's why we called it a lift, shift, rift, or cliff. If you're on the road to AI, congratulations… but it's not time to rest yet. Finish the job. Build tools that will serve your executives into the future. Create the next-generation C-suite. Infuse AI through your entire company. And remember it has been shown time and time again that companies that have data acumen outperform those that do not. The same is true within companies: groups and divisions that have data acumen outperform those that don't. If you climb the AI Ladder, you'll have it in abundance; if you simply walk by the Ladder…not so much.

CHAPTER 10

Tips and Best Practices on How to Get Started

You've learned about what needs to take place on the four rungs of the AI Ladder—Collect, Organize, Analyze, and Infuse—and how modernizing your information architecture will happen organically along the way. If you're feeling like it's potentially too big of a climb or are unsure about exactly where to begin, this chapter will help put you at ease. Here, we've compiled some proven tips and best practices for you as your organization begins its AI journey.

Manage Organization-Wide Change

As we've seen, AI is becoming the most powerful technology of our time, with the potential to increase productivity not just tenfold but a hundredfold, or in some cases a thousandfold. Change of this magnitude is inevitably disruptive—and in this respect, AI promises to be the most disruptive technology in history, possibly even more than the web. Just as the automobile put buggy-whip makers out of business, just as digital cameras put makers of photographic film mostly out of business, AI is going to turn whole industries upside down. Your job is to anticipate and manage this change so that AI can be used to your advantage, and not be seen as a threat. In other words, you have to be the one turning your competitors upside down, not the other way around. Bottom line: you must ensure you're using AI as a lift or shift...and it's not pushing you into a rift or cliff.

What kinds of changes are we talking about? In short, everything. We've discussed the modifications to corporate culture that must be made on the journey to AI. Your company will need to become more data-oriented, and more experimental. Humans will need to accept AI applications as assistants. Everyone will need to embrace change, rather than seeking to preserve things the way they are. And they will need to start looking for opportunities to effect change and the data they'll need to make those changes.

Change in Daily Tasks

In daily standups, instead of asking "What's blocking our progress?" as you would in most traditional DevOps and agile environments, ask, "What new inputs can we give the model? What hyperparameters can we tune for better performance (accuracy)? What policy can we give the model to further refine its results?" A model's accuracy will start to drop the moment it's deployed. Even if you don't pull the model and recode it every single day (while models need to be rebuilt periodically, rebuilding obsessively can be a road to madness, but it all depends on the type of function the model supports), that doesn't mean you shouldn't be looking to improve and planning for the next revision. Asking the important questions every day prepares you for the next round of revision. If for some reason your data team isn't having short regular standup meetings, this is a good time to start. Also, ensure you have strong collaboration skills using tools such as Slack (other favorites we use include Trello and Mural). The approaches we outlined in this section helped us cut the number of emails and meetings by almost 50% in the organizations we lead, and staff are more engaged and more informed than ever before.

Change in Overall Business Processes

All businesses have business processes. Whether formal or informal, large or small, local or universal, the processes exist—and AI's value is in its ability to harvest all the wealth, all the knowledge, all the potential benefit currently untapped within those processes. And as you climb the AI Ladder, AI will become the core of your business processes, as they are reimagined with automation, optimization, and prediction.

Infusing AI through your organization will change your business processes. In most cases, you're likely going to rebuild established processes from the ground up (that's our general advice), as opposed to layering AI on top of them. Again, humans will need to accept machines as assistants. In some cases, machines will handle the easy tasks, leaving the more complex (and hopefully more interesting) ones to humans. In other cases, humans will have AIs working beside them to aid in decision making, offering suggestions and validating human insights. One of the most impressive things about Watson's win on *Jeopardy!* wasn't its ability to come up with answers; it was the process that led to the answers. Watson showed the potential answers it was considering, including the answer it ultimately gave, with its confidence factor for each response—a key component when scaling AI, because it helps to explain the decisions the AI took. The list of potential answers is interesting because there are many situations where the obvious answer is wrong, and seeing the list of alternate possibilities might lead to a creative solution. One of AI's greatest benefits may be helping humans to have insights they wouldn't have on their own (think back to the chess insights we shared in Chapter 2).

Processes can change in more radical ways, too. Optimization across the company may well uncover processes that are duplicated or just plain unnecessary. In any company that's been around for a while, there are legacy processes that exist because "That's how we've always done things" (an answer neither of us have ever tolerated in our standups), as well as extraordinarily complex processes that have metastasized. Some advice? Whenever you hear this phrase, you've likely found a great place to look for an opportunity to make things better—even without AI. Another pro tip: when you encounter resistance and hear words like "They won't support it" or the name of a group ("Marketing says they can't..."), don't accept it. When folks take a position, you need names so that you can resolve or better understand their concerns.

Other processes can be revisited and simplified, for example by implementing predictive maintenance. Rather than shutting down key machinery for regular maintenance, AI can use sensors to detect vibration or audio signatures that indicate the machine needs service or is about to fail. Sensors are inexpensive, and an AI application can easily detect problem patterns well before humans notice them.

Let's step back for a moment. Try not to think of AI in the language of data and models. Think of it, rather, in terms of processes that can be improved. If you're running a hospital, all of your processes are related to your mission of improving the health of your patients. Sure, AI can be used to help predict readmissions, read X-rays, or predict the number of emergency room visits a patient with a certain kind of issue will make. But as you embark on the process of infusing AI throughout your organization, keep your eyes on the prize: your mission to provide better patient outcomes. This will be your touchstone as you transform your organization. Your goal (in this example) isn't to use AI; it is to improve patients' health. Learn to rethink that goal in ways that help you understand your AI transformation. What metrics show whether you're achieving your desired results? Individual stories and anecdotes don't count. And don't trust your metrics blindly; realize that every metric can be gamed, and that every organization has some people who will game a metric for their personal advantage. Make sure the metrics you land on measure the right things, too—a major research hospital that gets the worst cases may well have a higher fatality rate than a small local hospital that only does routine medicine.

As business processes evolve, you must manage the organizational culture shocks that are certain to happen. Some people will feel threatened by change; you will have to address their concerns. Ignoring them won't make the problem go away, but if you can get them on your side, if you can show them how AI will benefit them, they can become powerful and effective allies. As we saw in Chapter 3, one of the biggest causes for failure in AI projects is cultural mismatch. Some organizations simply are not ready.

That's where the notion of the AI Ladder came from: the observation that organizations that had not come to terms with the need to change how their data is collected, stored, organized, and governed were simply unprepared to use AI to analyze that data and get any benefit from it (even if they wanted to). We'll repeat: the AI Ladder isn't necessarily linear; many companies start on the Infuse rung, and then "climb down" to the lower rungs as they progress. It's possible to get onto the AI Ladder with minimal effort by introducing prebuilt applications; you'll eventually want to move on, but leveraging preexisting products is a good way to get started with some wins.

Example: Pharmacy uses automation to modernize laborious and error-prone processes

A pharmacy was searching for ways to better serve its patients. But most processes in pharmacies are manual and time-consuming. Take entering prescriptions into the pharmacy's main system. The doctor's directions need to be converted into a code from a dictionary, and the drug names matched with the pharmacy database. Standardization of this data is critical, as the pharmacists use the application to accurately and safely fill prescriptions for customers.

For this pharmacy, this process proved to be rife with human error. There was no way to be sure that two workers would interpret the information on the prescription in the same way. Turnaround time also varied widely. Using robotic process automation combined with an AI algorithm, this pharmacy was able to determine from historical data how to standardize drug names and doctor directions.

At first, the algorithm was less than trustworthy, and the pharmacy was hesitant to use it. But by refining it and feeding it more data, the pharmacy was finally able to create an algorithm that was accepted by employees, ultimately achieving 95% automation with 99.1% accuracy. Today, the pharmacy is processing nearly 400,000 prescriptions each month, with pharmacists now applying their brain power and experience to tasks that require critical thinking.

Change in Thinking About Data

Everyone in the company will have to think about data. What data does your company have? As we've seen, many companies don't know what they have, particularly if that data is walled up in departmental silos. Many companies discard data that could revolutionize their operations (we see this with log data a lot); these organizations just aren't thinking about data, and don't see how it could be important. The "save everything" mantra of a decade ago has largely been discredited (though as you learned at the start of this book, that was more because the data understanding curve was so flat—and AI will change that), but everyone on staff should think constantly about what data should be saved. What data don't you have? What data would you like to have? What data could you get? What data do you have that you could do more with (like data exhaust)? And what data would help you to solve your biggest problems? If you think about the data you need, rather than the data you have, you might find that that data already exists, or can easily be acquired.

Make Data a Team Sport (And Some Cool History About Car Racing)

In 2013, Ron Howard directed and released the movie *Rush*, a film that captured the rivalry between James Hunt and Niki Lauda during the 1976 Formula One racing season. It's a vivid portrait of the drivers and their personalities, with a pretty typical (and captivating) focus on the drivers as the heroes of the race. But it does something deeper and more interesting as well. The film looks into the essence of Formula One —a true team sport.

The "Formula" in Formula One refers to the set of rules to which all participants' cars must conform. Formula One rules were agreed upon in 1946, on the heels of World War II. Modern Formula One cars are open-cockpit, open-wheeled single-seat vehicles. Their cornering speed comes from "wings" mounted at the front and rear of the vehicle. The tires also play a major role in cornering speed. Carbon disc brakes are used to increase performance. Engines have evolved to turbocharged V6s. All these components are integrated to provide precision and performance, and to win the race. However, the precision and design of the vehicle are useless without the right team.

In Formula One, an "entrant" is the person who registers a car and driver for the race, and who maintains the vehicle. The "constructor" is the person who builds the engine or chassis and owns the intellectual rights to the design. The "pit crew" is the team that prepares and maintains the vehicle before, during, and after the race. While TV cameras focus on the driver, with a couple of obligatory shots of the pit crew scrambling to change tires, the real story is the collaboration of the complete team. In some cases, more than a hundred experts are working together to make the difference between success and failure.

What's the point of this car racing story? The best way to succeed with AI isn't by assembling a group of superstars. The companies that get the most out of AI won't be those with the most and the best data scientists. That simply doesn't scale. As Bill Joy, cofounder of Sun Microsystems, once said, "No matter who you are, most of the smartest people work for someone else." If you think you can corner the market on good data scientists and data experts, think again. Instead, you need to build teams that always want to learn, try new things, are curious, and can collaborate with each other to get the job done. Remember, team effort almost always beats individual efforts. To succeed, you have to build the entire team; like Formula One, data is truly a team sport!

Subject Matter Experts

As we've noted throughout the book, SMEs are a key part of any data team—maybe the most important part. Data scientists may be scarce, but there are lots of ways to build software and algorithms: we've discussed using prebuilt applications and automated model-building. Expertise in the details of your business, on the other hand, is really scarce. Fortunately, your business already has its own experts; if it didn't, it wouldn't have survived. You need someone who understands the whole picture: what's working, what isn't, the finances, the current challenges, the hiring situation, and the regulatory environment. A good SME doesn't just know your business; they must be familiar with what's happening outside of the business. An SME needs to understand the overall business environment, the competition, the customers, the ways the market is changing and adapting, and the challenges you will face in the future. Their wisdom and understanding will be a crucial component of your team.

Data is just a bunch of numbers. The question you need to ask yourself is whether your data team includes people who understand what those numbers mean.

Data Scientists

You need data scientists to guide your adoption of AI. That's a given. But keep things in perspective. Even corporations whose entire existence is built on analyzing nearly incomprehensible amounts of data—companies like IBM, Microsoft, Google, and Facebook—have fewer data scientists on their staff than you'd think, when you consider them as a proportion of the workforce. If you run a chain of hotels, or a fleet of trucks, or a snowboard and apparel company, you're not suddenly going to have to grow a whole new data science division.

Remember what so many studies have shown: data scientists spend up to 80% of their time locating, getting access to, and preparing and cleaning data. None of these tasks require advanced training in data science. Your IT staff, your software engineering staff, and other members of your team can learn to do these things. Reserve highly skilled data scientists for difficult tasks, like hyperparameter tuning in new models; investigating dimensionality reduction so that the models have fewer parameters and run leaner, making them more suited for edge device deployment and the like; or validating a model's performance with mathematical methodologies like F1 scores, *p*-values, Mann–Whitney U tests, confusion matrices, and more.

You can get junior data scientists started with Watson AI, Amazon Sagemaker, H20.ai, and other cloud-based tools. This will allow them to gain experience, learn new concepts, and grow into senior data scientists. But there is more value here than most realize. Take a moment to think about how tools such as IBM's AutoAI work. They look at data, apply proven best practices (like feature engineering, bagging and boosting, and so on), and spit out an explainable model. Who is behind that "magic"? Let your junior data scientists explore competitions like Kaggle (an ongoing competition that is essentially the Olympics of data science and whose grand master champions are its "hall of fame" Olympians), which will expose them to some of the most creative practitioners in the world, and give them insights into the best-of-the-best approaches to data science (the world's most accomplished data scientists are almost always on staff for any vendor's auto AI offering). Let them learn by doing, just as the organization itself is learning by doing. If you're investing in prebuilt AI applications, make sure the junior data scientists are involved in training them. Success in AI is less about writing good code than it is about training the model.

Standardizing the definition of a data scientist requires a recognized assessment platform. This is why IBM and the OpenGroup created the Open Certified Data Scientist (Open CDS) program. This program defines three levels of certification for data scientists:

- Level 1: Certified Data Scientist
- Level 2: Master Certified Data Scientist
- Level 3: Distinguished Certified Data Scientist

The certification process consists of acquiring milestone badges, completing an experience application, and having a subject matter expert review.

Data Operations (DataOps) Specialists

Data scientists are great at understanding and working with data, but they often have little experience running real-world applications. And you can't blame them; they want to do data analysis and build new AI models, not keep a bunch of servers up and running or argue with managers about getting access to siloed data. As you progress in your journey to AI, you will need to nurture people who specialize in data operations. DataOps, as we discussed earlier, encompasses most of the activities on the second rung of the ladder (the Organize rung). DataOps is responsible for creating and maintaining data catalogs, creating and maintaining the data pipelines and other infrastructure that feeds your AI applications, and managing data access. If you start with prebuilt AI applications, you might be able to postpone an investment in DataOps—but you will eventually have to make it.

DataOps may also include packaging, deploying, and monitoring your application: getting the application out into the real world, at scale. Several years ago, IBM made the leap to infrastructure as code by automating tasks like application deployment. With AI, automation is the only way to go. (Is that surprising?) AI applications are too complex to be hand-deployed by traditional admin staff. People responsible for deploying AI into production will use open source tools like Ansible, Terraform, Chef, and Puppet to automate software deployment. The ability to create and deploy containers using tools like Docker, and to orchestrate container runtimes in private or public clouds using Kubernetes, will also be important. Find the people in your organization who are ready for this challenge and turn them loose.

Data Engineers

Consider building a separate data engineering team. A dedicated data engineering team could handle storing, retrieving, extracting, and serving the data used by your AI team. This team would have the necessary expertise in requisite technologies like Spark, Hadoop, Ray, and traditional SQL, as well as the hardware experience and technical skills to build a data architecture that is scalable and cost-effective. They could also coordinate responsibility for the cloud services that AI projects in your organization will consume. Having specialists handle the data, and leaving the model building and analysis to the data scientists who have experience with that, is a great way to avoid a trap that we've fallen into in our own journeys and seen many others fall into. Give jobs to those who can do them best.

Training for Career Development

Send your traditional software engineers to machine learning and deep learning training events and camps. (As we write, travel restrictions are causing many companies to move from in-person training to online training.) Frameworks like TensorFlow/Keras, PyTorch, and more are geared toward traditional coders and are making it easier for regular developers to start building machine learning models. Additionally, many of these services are available from a multitude of cloud providers, packaged in ways that software developers are familiar with. It may make sense to expand your data and AI team by enrolling other developers in training for these frameworks so that they can be active contributors to your projects when additional staffing is needed. In our experience, Python programmers have a leg up versus other candidates in evolving as standouts on your AI team: it's not just them, but a good place to cull from some high-potential talent. Why? Most AI tools are deeply integrated with Python, and it's such a vibrant programming language that the programmers who use it are always learning—another fundamental trait for successful teams and careers.

Now add the endless opportunities for self-learning. Make training opportunities available to your staff—even if it costs money. Some of our favorites include massively open online courses (MOOCs) such as those offered by Coursera and CognitiveClass.ai, anything penned by Andrew Ng, and training by other individuals who are standouts in their field, like Adrian Rosebrock and Adam Geitgey—we've used them to train some of our own staff and the results have been outstanding. There are many other great options, but you can't go wrong starting with these.

Data science is at the center of the transformation we're talking about, but it's not the essence of it. The essence of the transformation lies in changing the mindset of every person and every aspect of the company to embrace AI.

Embrace AI Centers of Excellence

One idea that has worked at IBM and several of our customer sites is creating an AI Center of Excellence (COE). COEs are simply resources where the concepts we've covered in this book can be taught and explored. In a COE, you can develop pilot projects, document lessons learned, and share them. This approach has the beneficial effect of spreading the message that AI is a gateway to thrilling new worlds. It's not a mysterious, incomprehensible threat dominated by an arcane priesthood, but something that anyone can learn about and explore.

When you have a COE, people on your staff can understand why certain policies must be put in place, and why certain processes must be followed for AI projects to succeed. These policies and processes don't arise by arbitrary decree from on high, but because they're necessary for success and trust. COEs are essentially about openness. If your AI projects are dominated by an insular group of specialists, everyone else on the staff will inevitably become resentful. If you're open about what you're doing, and give others the opportunity to take part (and, in turn, to build their own careers—possibly even taking them in a new direction), you're much more likely to have a smooth transition. AI is an opportunity for everyone; spread the opportunity around! Our advice when staffing a COE? Don't stack it up with project managers who can't do the work—we see this a lot. A COE needs terrific project management, so get the best, but you have to do actual work. Quite simply, you'll want many more cooks than restaurant managers if you're going to change things.

Example: Honda Sets Up Knowledge Hubs to Build Minimum Viable Products, Organize Training, Share Data

Global car manufacturer Honda aspires to become the leading maker of cars that are designed to allow consumers to feel the joy of driving. It invests billions in R&D to explore new ways of making this happen. New sources of data from the Internet of Things—vehicle diagnostics and telematics, smartphones, biometric sensors, and more—as well as the massive amounts of unstructured data from social media and customer surveys have enabled Honda engineers to gain a better understanding of how cars and drivers interact in the real world.

But what was the best way to introduce big data analytics technologies into Honda's R&D arm? By partnering with IBM, Honda was able to establish a knowledge hub for AI and machine learning technologies, helping to set up minimum viable products (MVPs), organize training courses, and encourage engineers to share their knowledge, experience, and data. IBM SPSS Modeler in particular quickly became a popular tool. More than 100 Honda engineers have now completed the training, and many of them use SPSS regularly in their work.

Honda R&D also uses IBM Watson Discovery for text mining, giving researchers near-instant insight into vast stores of documents and other data. For example, in the US, the National Highway Traffic Safety Authority (NHTSA) provides a rich source of insight into consumers' problems and safety concerns. When a Honda executive asked a question during a strategic meeting about how customers liked a particular feature in a particular car, the engineers were able to log into Watson Content Analytics, analyze more than a million records in the NHTSA data set, and within 10 minutes find examples of relevant feedback from customers.

Honda R&D's big data training programs have helped foster an open and collaborative culture among the company's engineering teams. Learning about big data analytics has also helped engineers think outside the box. Instead of simply analyzing the parameters that they think are important, they can use data mining techniques to uncover patterns and clues that they might never have thought about otherwise.

Build Ethics Into Your Process

The public is understandably wary of abusive uses of data in general, and AI in particular. There may be no one left on earth whose data hasn't been stolen in some major data breach. (If that's not true, it's slowly becoming that way. The impact of stolen data on personal lives and corporations is moving from an "if" condition to a "when.") And there's almost certainly no one left who doesn't believe that their data is being bought and sold, for activities over which they have no control. Add to that the fear of AI that arises from science fiction, and centuries (even millennia) of speculation over humanoid machines (which we mentioned briefly in Chapter 2), and you have an atmosphere of profound distrust.

One approach to the problem of distrust is regulation. IBM has proposed a framework for "precision regulation" that will hold companies accountable without becoming overly broad and hindering innovation; it stresses accountability, transparency, explainability, and fairness. But rebuilding trust isn't ultimately about regulation; regulation signals the absence of trust and has never made anyone trust an abusive industry. We have this saying we've both been telling our kids since they could speak: you lose trust in buckets and you gain it back in droplets. Quite simply, trust is hard to build and easy to lose—but building trust isn't an impossible task. Trust starts with ethical behavior. Make sure you have corporate standards for what you will and won't do with data, and make sure your developers and your clients follow them. Straight up: decide what kind of AI organization you want to be. Do you want to be a good actor or a bad actor? Put that in a charter, establishing a common understanding of how you will work with AI, and ensure everyone on your teams understands this social contract. Remember that at the start, Facebook's Cambridge Analytica fiasco was, in essence, a problem about controlling access to data—a problem that could have been prevented by good data catalogs tied to metadata about allowed and forbidden access and usage, and automated tools implementing access controls.

How do you ensure that you're a good actor? It's time to think carefully about your company's policies and values. Ethics covers a lot of territory, but it's worth giving some thought to three things in particular: privacy, safety, and fairness.

Privacy

In *Privacy in Context*, Helen Nissenbaum argues that the essence of privacy isn't keeping data hidden. Data has to be used, and people understand that; an online retailer can't ship to you if you won't give them your address. Privacy becomes a problem when contexts change: shipping you your Christmas presents is fine, selling your address to spammers isn't. Selling your phone number to a telemarketer is definitely out of bounds.

Therefore, think carefully about what you will do with the data you collect. It can be tempting to sell data—particularly for a small company, because income from selling data looks like "free money"—but it's a poor way to retain your customers' trust. And realize that what we've told you many times in this book is true for bad actors, too: data sources become much more powerful when they're combined. Several pieces of data from different sources, innocuous in themselves, can give a bad actor a very detailed picture of your customers' activities. You don't want to be the source that helps someone spread rumors that one of your customers is filing for bankruptcy. It won't be as simple as them checking your database for "filing for bankruptcy" records. Corporate espionage is sophisticated, and bad actors have also discovered AI; correlating patterns on late credit card payments with people applying for new jobs can enable them to draw some surprisingly accurate conclusions. Don't become an unwitting accomplice. Regulations like the GDPR are helpful—if you stay in compliance, you won't go too far astray—but compliance isn't the entire story.

Safety

Safety is another important concern. You need to protect your data from unauthorized access. That goes without saying—companies have been concerned about cybersecurity for years. But realize that safety is about protecting your customers as much as it is about protecting your business plan and balance sheet. Safety is an ethical concern. Frankly, way too many data breaches have resulted from inattention and bad practices, pure and simple. Yes, there are brilliant data criminals out there. But they're a very small minority. In most data breaches (including some of the most highly publicized and politicized cases), the criminals just walked through the (virtual) front door. Nobody bothered to lock it, or even to encrypt all the sensitive data inside. And in some cases, the criminals were stealing data for months before anyone noticed.

If you're building smart devices, ranging from light bulbs to cars, safety has more dimensions. Autonomous vehicles can, and have, killed people. (So do human drivers, but you don't want your company to be at fault.) Smart light bulbs and thermostats have been used in harassment cases. Smart locks and surveillance cameras have been commandeered and used to spy on their owners. Many of these situations don't involve AI to any significant extent, but that doesn't mean you can ignore them. As you embark on your journey to AI, please realize that "safety" can be about much more than credit card and Social Security numbers. Ask yourself this simple question: "When was the last time I applied a patch to all of my internet-connected devices?" All of the intelligent "AI" devices making their way into our homes (even a connected toaster is available for purchase) are potential targets for DDOS attacks and need to be made and kept safe—the connected-home market hasn't fully embraced this concept yet.

Fairness

Finally, let's take on fairness. We all think we know what fairness means—at least, we recognize when someone is unfair to us. It's much harder to understand what fairness means to others.

There's a reason we use the terms "fairness" and "bias" separately. Bias is a technical term in statistics that has to do with the relationship between the source data and the conclusions you draw from that data. In AI, unfairness almost always arises from problems in the training data. The results the AI produces may well be statistically unbiased, in the sense that they properly reflect the training data, but not fair. Data frequently reflects historical biases. For example, in dense urban areas, many neighborhoods are segregated by race and national origin. Would a real estate AI application that's trying to match home buyers to houses engage in "redlining," preferentially recommended houses according to race? If you're not very careful, that's likely to happen.

Likewise, AI often reflects bias in how the data was collected. For example, if your data was created by a survey, where was the survey done? In a business district? In wealthy neighborhoods? Poor neighborhoods? Did you survey your customers, and if so, do you expect your existing customers to be similar to the ones you want to steal from your competitors? This is precisely why data provenance is so important: you need to know where your data came from and how it was collected to understand whether your AI is generating trustworthy and fair results.

It's particularly important to look at how different classes are represented in your training data. Are minorities represented well enough that your system will be accurate across all groups? If you're building a medical application, and black people only represent 12% of your training data, your application will probably be less accurate for this subgroup—it doesn't matter that they only make up 12% of the population.

This is one reason face recognition applications are significantly less accurate for minorities, and natural language voice-to-text applications have trouble understanding people with accents. If there isn't enough diversity in the training set, whether it's diversity of skin color or of accents and regional dialects, AI won't be able to learn enough to treat those differences correctly—even if its overall accuracy across all groups is adequate.

Building Trust in AI

For AI to be used to its full potential, it must rest on a foundation of trust. All stakeholders must trust the technology and its results. A bank, for example, has many processes that can be transformed with AI: processing loans and mortgages, opening new accounts, managing overdrafts, dealing with fraud, and more. While these aren't issues of life and death, they're very important to the financial health of the bank and its customers. If AI is making recommendations or taking actions that directly impact a customer's finances, obviously the customer, the banking staff, and anyone else involved must trust it. But that's only the start. Banks deal with enormously complicated issues of privacy, security, and legal responsibility. Does the bank leak information about its customers to the outside world? That has huge legal implications. Is the bank's data protected from attack? Financial institutions are frequently the targets of criminals who steal data and resell it online. Can the bank explain the reasons for the decisions it makes? If someone is denied a loan, the bank should be able to explain why, and "the AI said so" doesn't cut it; an AI application must be able to explain its decisions. Infusing AI through an organization must happen in an orderly manner that guarantees the results are secure, reliable, explainable, and trustworthy.

There's much more to say about ethics and trust, but this was a good start. The costs of repairing your reputation, once it has been damaged, can be astronomical. Once the public labels you as a bad actor, it will be hard (perhaps impossible) to change their mind. So be a good actor. Make that a core value.

Choose Projects Selectively, and Embrace Failure

In the age of AI, every organization must become a learning organization. Those that cannot learn from failure will not survive. Whenever a technology as transformative as machine learning and AI arrives on the scene, there are going to be failures. There are so many new concepts to be mastered, so many moving parts to coordinate; sometimes things are just not going to work out.

Successful organizations are resilient. They can tolerate the occasional setback without focusing on blame, and without abandoning the entire process. That's why we recommend starting with small projects, on the scale of five to seven workers and spanning four to six weeks. In our experience, it's much better to launch 100 such projects and see 50 of them fail than to launch only a few big projects, even if some of them succeed. The more people participate, and the more people who see your successes, the better your chances will be.

Back to our baseball analogy, don't go for a home run, which comes with a higher risk of striking out. Just get on base any way you can. Walk, single, get hit by a pitch, whatever it takes to advance. Just get a little bit of success and build on that. Choose a project with meaningful business impact. Whether it's an AI journey or your personal goals, we've concluded that people often underestimate the importance of a small win, a small bit of progress, a single moment. For many, what starts as a minor win accumulates into something much more. We make this promise in life and in AI: it's really too easy to underestimate the value of making small improvements on a daily basis.

To ensure meaningful business impact, pick metrics that allow for good results and have reasonable expectations of change. As good as some data scientists are, they can't predict how their models will behave once they are released into the wild and fed with real-world data that they have never seen before. (When you hear data scientists talk about models working in the real world, they'll use the word "generalizes.") But who decides when a model isn't working well enough? Since models often return confidence scores to reflect how accurate their estimates are, your QA team along with the data scientists need to decide what a proper quality threshold is. Is 90% confidence enough? When model performance deteriorates over time, as is inevitable, how low do you let that confidence ranking slip before taking action? Is a drop from 90 to 89% confidence significant? For a healthcare application it might be, but for a consumer recommendation engine the chances that anyone will notice a 1% drop in accuracy are negligible. Is a 5% degradation over 10 business days worth a review that results in retraining the model? Is a 1% drop in a month too much? These are the types of questions and answers you need to iron out within the team before deploying the model.

Example: Insurer Tracks Metrics to Communicate Success of Its Model

A leading Latin American insurer with more than $1 billion in direct written premiums was seeing customer churn increase, as well as more and more lapsed policies over time. Turning to Relativity6, a consulting firm specializing in developing AI-based algorithms for financial services firms, the insurer handed over an internal customer data set from its personal auto, health, and life insurance lines that included policies, claims, transactions, customer information, and product description data. Relativity6 performed an exploratory data analysis, then preprocessed the data sources, created data models, and applied feature-learning precepts to the data. It then applied proprietary machine learning algorithms to train win-back and retention models.

The insurer used the results to embark on a telephone campaign to win back customers who hadn't renewed their policies, focusing on those who were predicted to be the most likely to return. The company was also able to use these results to predict (and prevent) churn, and to upsell customers on new products. Knowing that there would be some skeptics to convince, the team tracked the success metrics carefully. The AI algorithms performed significantly better than previous processes used by the insurer to win back customers, with a 98% success rate. This high rate of success convinced the firm to continue with its telephone campaign to both win back old customers and cross-sell existing customers on new products.

Beware of False Negatives

We've seen too many cases where AI projects didn't work out as planned (or in some cases, as hyped), only to have senior management draw the conclusion that "AI doesn't work for us" or "AI is a fad." In many of these cases it wasn't an AI problem but a data problem, so it's important to conduct proper postmortem analyses on projects that fail or aren't going as well as expected. In short, there's no question that you will learn from your successes, but make sure you take the time to learn from failures too. "Fail fast and fail safe" doesn't mean fail and move on without learning anything about that failure. In fact, it's the opposite; "fail fast and fail safe" implies constant learning and application of that learning to the next iteration to mitigate future failures. These need not be complicated post mortems, but it's important to understand what went wrong if you're going to learn from your failures. You'll often find that the problem was in the data, not the code. (Hopefully "fail safe" needs no explanation; we're not talking about bruised egos here.)

Also beware of reinventing the wheel. Remember, to a large extent, the value of your AI project is going to come from your data, your business processes, and your team, not from proprietary algorithms. Oddly enough, AI is making code and algorithms less important, rather than more. Data is everything. Your programmers are unlikely to create algorithms that compete with what's coming out of research centers and universities—and these state-of-the-art solutions are quickly built into open source machine learning libraries and platforms (this is why transfer learning is so important). So embrace open source solutions for building your models, and direct your energies to where they're most needed.

Finally, be patient. When we look back on our early experiences training models, the picture looks something like Figure 10-1.

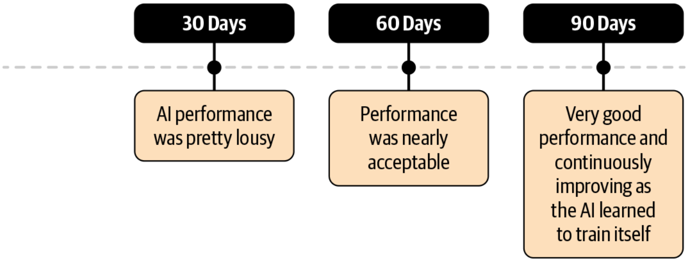

30 Days
60 Days
90 Days
AI performance was pretty lousy
Performance was nearly acceptable
Very good performance and continuously improving as the AI learned to train itself

Figure 10-1. Historical timeline for training AI models

If a project gets canceled at the 30-day mark, none of that value will be realized. The training process can be frustrating: it's easy to go for weeks without making any apparent progress, then suddenly hit on the parameter settings that make everything work. There's no reason you couldn't have done this on day one, but you didn't get lucky. That's how it goes. Patience is rewarded.

CHAPTER 11

The Future of AI

When you work in a nascent technology field like AI, people are forever asking you what the future holds. It often seems as though they're expecting to hear something that would come from the pages of a science fiction novel. But although it's fun to let our imaginations run wild occasionally, we usually try to stay grounded.

IBM employs futurists to "blue sky" the things that may await us 15 or 20 years down the road, and we share some of their visions here. But we're not novelists, and the kinds of things we like to talk about aren't based on unfettered imagination so much as extrapolation of actual current trends, plus the research going on in IBM's R&D centers.

Along those lines, the future of AI is mostly about making existing products better and, ultimately, helping clients put those products into production. But we have also identified a few major trends, or themes, that we feel will shape the way AI will evolve over the next five or so years. We'll cover those themes in this chapter. We'll also talk about some of the biggest business use cases we think we'll see AI being applied to in the future. Finally, we'll wrap up the chapter by taking a look at what the future of human work will look like in an AI-driven world and how the shift to edge computing might change things.

AI Themes to Take Us Through the Next Five Years

Where is AI going to take us by 2025? We've identified five different themes about the way in which AI will continue to mature, spread, and influence all facets of our world—business, technological, and social—over the next half-decade (Figure 11-1).

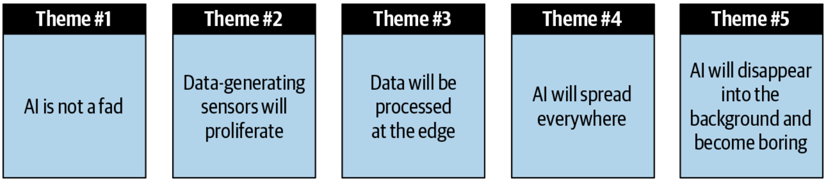

Theme #1
AI is not a fad
Theme #2
Data-generating sensors will proliferate
Theme #3
Data will be processed at the edge
Theme #4
AI will spread everywhere
Theme #5
AI will disappear into the background and become boring

Figure 11-1. Five themes for AI in the next five years

Theme #1: AI Is Not a Fad

First and foremost, AI is not a fad; but if you're this far into the book, you've clearly figured that out by now. It's not going away; there are no more AI "winters" coming, even though the reality still doesn't match the hype. AI is firmly established as an engineering discipline that is generating enormous and real value. It's as established as electricity—so much so that at the 2019 NeurIPS conference (the leading annual gathering of AI academics) participants were reminiscing about how AI "deep learning" was seen as a fringe technology as recently as 2006. Now AI practitioners are searching to generalize algorithms rather than simply solve niche problems. At the conference, Jeff Clune, an AI researcher at Uber, talked about an emerging AI field called "meta learning," in which AI algorithms can create their own learning algorithms to constantly evolve and thrive in continually changing environments. Although AI has experienced a lot of recent, narrow successes, Clune said, an increasingly generalized and flexible AI will help systems become safer and more reliable.

Over the next several years we will see AI appear in hundreds of new and unexpected guises.

Theme #2: Data-Generating Sensors Will Proliferate

A second theme is that data-generating sensors will proliferate at an astronomical rate, and the amount of raw data generated and processed by them will be staggering. A recent forecast from IDC estimates that there will be 41.6 billion connected IoT devices, or "things," generating 79.4 zettabytes of data by 2025 (Figure 11-2). Upon reflecting on the amount of data that will be generated each year from IoT technology, we're boldly renaming this popular acronym to the *Internet of Everything* (IoE).

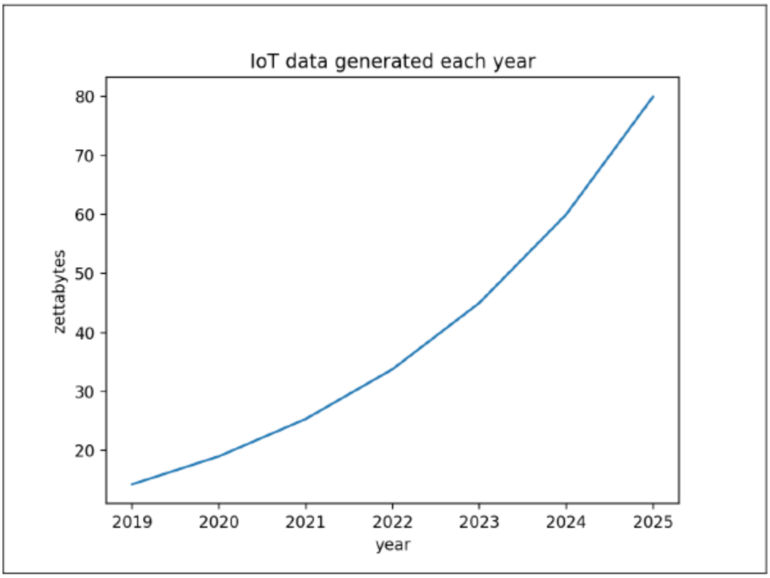


Figure 11-2. IoT data generated each year, in zettabytes (based on data from IDC)

These sensors will ubiquitously monitor the atmosphere, the oceans and rivers, the air inside buildings, the traffic in cities, and on and on. Corporations will have greater access, and be able to augment their insights using data that they do not generate themselves. Instead, it will come from private suppliers, government agencies, financial markets, and all sorts of new and unexpected sources—some of which have yet to be invented! While the amount of structured data contained in relational databases will continue to grow, it will constitute a shrinking proportion of collected and subsequently understood data. Most data will be in the form of unstructured objects.

> It's important to understand that all data has some sort of structure to it, so technically you'd consider this kind of data semistructured. For example, we'd all consider a picture to be unstructured data, but did you know there's hidden EXIF metadata embedded in every digital photograph? This metadata includes details such as the date and time the picture was taken and a range of camera settings like shutter speed, aperture, and more. (You can view the EXIF metadata for any picture at a number of websites for free; for example, *http://exif-viewer.com*). Similarly, a text document is likely wrapped in JSON or XML; even a Microsoft PowerPoint document has XML structure to it. However, the untapped value of text and pictures lies in the unstructured parts (the text within the encapsulating tags or the image itself), and that's why most people—including us—just refer to this kind of data as unstructured, even though it does have some structure.

And don't forget the comparative growth of metadata, or data about data, that will accompany the explosion of the IoT. Metadata will add to the possibilities of AI systems, as it provides valuable context and insights that can drive more intelligence and make sense of seemingly random scenarios, data sets, or environments.

Theme #3: Data Will Be Processed at the Edge

A third theme, closely related to the second, is that much of this sensor-generated data will be processed at the site where it is measured or collected. This is the new data processing architecture called "edge computing," where the computation is done at the outer edge of the network.

When Gartner named its top technology trends for the next 12 months in December 2019, edge computing was at the top of the list. Edge computing solves an important problem: the latency involved in data's round-trip journey from the edge of the network to the place where it is processed is too much for today's AI applications. Especially as the IoT (or our newly dubbed IoE) grows, the edge model will enable a multitude of use cases. Take autonomous driving. Sub-submicrosecond response times are needed for autonomous vehicles to absorb local data, analyze it, and make a decision that can literally mean life or death. While cloud computing often requires a round trip from the device to the cloud and back again, localizing data at the edge can minimize traffic latency and empower users with actionable data without delays. As 5G connectivity reaches the edge, it will allow for even greater volume and speed. We take a deeper dive into edge computing at the end of this chapter.

Generally speaking, the models you build through training are still quite computationally expensive to execute (inference). To get maximum value from your edge models, they need to be run against every transaction that occurs and that can't have a significant impact to the overall time it takes the transaction to run (or, for example, people may decide to cancel their purchase or use a different payment method). Essentially, speed of execution and throughput are the magnified considerations when AI is on the edge.

One way of reducing the inference cost of a neural network is to remove parts of the network that were used for training but aren't needed for inferencing. This process is referred to as "pruning." Another way is to combine multiple layers into a single computational step, which optimizes the model further. Also, depending on the accuracy, training and inference may not require the same level of mathematical precision to achieve the same or relatively similar level of accuracy. Training involves the processing of large amounts of data within a complex model and this typically requires the use of high-precision 32-bit (FP32) or 64-bit (FP64) floating-point numbers. Inferencing, on the other hand, has less of a reliance on high-precision numbers. The use of 8-bit or 16-bit integers or smaller floating-point numbers can improve the speed of the calculations, while still maintaining a high level of accuracy.

These are things you have to consider when AI will be processed on the edge. You have to expand your aperture to not just focus on the model's performance (accuracy), but how can you deliver the most efficient and fast-running version of that model on the edge; you might even decide to give up a touch of accuracy because the model runs so much faster without that specific layer.

Theme #4: AI Will Spread Everywhere

A fourth theme is that AI will continue to bring about significant changes not only in business, medicine, and science, but throughout society. It is going to continue to change the nature of work, the organization of cities, and education from preschool through lifetime learning. AI will also be harnessed to solve large-scale problems confronting humanity, such as the climate crisis and pandemics (among other social causes), not just challenges faced by businesses.

What of the change from the ubiquitous Internet of Things to the Internet of Everything? We're going to change our terms yet again (even in a book, we fail fast and are agile). Just as the Internet of Things became the Internet of Everything, we'll see another change. The proliferation of sensors that collect data, along with the ability to generate and process data at the edge (theme #3), will transform the Internet of Everything into the Intelligence of Everything. AI will be pushed into edge computing devices, which for the first time will become truly "smart." As Mary Poppendieck said in her keynote at O'Reilly's 2020 Software Architecture conference, "My robotic vacuum cleaner wouldn't work during an internet outage—its map of the house was stored on some server. We have to rethink software so that doesn't happen."

AI on its own is poised to change the world as we know it—but when you consider the Intelligence of Everything, there are an incredible number of ways that AI for social good can change society for the better! Here are just a few examples:

The environment

Climate change and its various subcategories—pollution, depletion of natural resources, sustaining biodiversity—is probably the most critical challenge humanity faces today. AI is already helping, and the range of possibilities is enormous.

One nonprofit wildlife conservation organization, The Rainforest Connection, uses TensorFlow to detect illegal logging and wildlife poaching in vulnerable areas. The group has hidden modified smartphones powered with solar panels that act as acoustic monitoring devices—called "Guardians"—in the trees of threatened areas. The Guardians continuously monitor forest sounds, sending all the audio up to cloud-based servers over the local cell phone network. Once the audio is in the cloud, The Rainforest Connection uses AI to analyze all the auditory data in real time, listening for chainsaws, logging trucks, and other sounds of illegal activity that can help pinpoint problems in the forest. As the stakes of missed detections are high, the analysis of the huge volumes of audio data that come in constantly from every phone, 24 hours a day, must be done quickly and accurately.

Other efforts are using AI to identify and track bird species. Another group is working to count bird populations on remote islands based on birdcalls. If you think about it, this is an astoundingly difficult problem, since most birds travel in groups. If you walked into a crowded bar, could you tell how many people were there just by listening? We don't think so, but AI might be able to do it. This sort of intelligent monitoring can be adapted to many environmental risk scenarios around the globe.

First response to crises

AI is also transforming first responders' ability to handle both natural and man-made disasters, aiding in the response to events such as earthquakes, wildfires, and tsunamis, disease outbreaks, and search-and-rescue missions. Today, drones, robots, and sensors can already provide intelligent and accurate information about affected landscapes, damaged buildings at disaster sites, or car accidents; they can stream video of a car pileup and provide information on casualties while first responders are en route to the accident. AI can help rescuers understand the environment and provide rapid analysis of the extent of damage so that any trapped victims can be retrieved more quickly and safely. These AI-based tools will only grow more accurate and useful in the next few years.

For example, 1CONCERN uses AI to predict and quantify the impact of natural disasters on buildings and landscapes, using machine learning to build more resilient businesses, infrastructure, and global communities. It has already mapped out more than 150,000 square miles, covering 11 million structures and 39 million people, and has modeled almost 15,000 fault lines. Its platform brings together all this data and uses machine learning tools to determine which areas and edifices will suffer most from a natural disaster, producing a comprehensive picture that can be used during emergency operations. 1CONCERN will continue to map out terrain in the coming years, with ambitious plans to cover the planet.

For another example, look at BlueLine Grid, created and developed by Bill Bratton, the former commissioner of the New York Police Department, and colleagues. BlueLine Grid (acquired by WorldAware in 2018) provides a mobile communication and mass notification platform built for interagency communication between civil service employees and private sector security teams. The platform is effective because it uses AI to rapidly locate necessary public employees by geographic proximity, area, or agency. This represents just the beginning of AI being used to foster efficient connectivity, collaboration, and communication among first responders.

Medicine and Pharma

AI is already finding applications in many medical and healthcare organizations and will continue to do so, supporting areas such as disease prevention and diagnosis, precision medicine, treatment plan optimization, medication management, and drug creation and compound discovery. AI-powered software will also be used to optimize both administrative and clinical workflows to reduce costs, accelerate claim processing, improve supply cost management, and train physicians.

Of course, despite these promising uses of AI, a number of issues need to be addressed before the technology becomes an utterly reliable solution. Most importantly, there's that problem of bias we keep telling you about. Bias can creep into algorithms designed by humans, and medicine is no exception; the effect of bias could literally be a matter of life or death. For medical applications, it will also be important to focus on the transparency of AI tools, ensuring that the decision-making process isn't enclosed in the proverbial AI black box.

At the time of writing this book, AI is playing a significant role in the COVID-19 pandemic. IBM Watson is being used for intelligent search, enabling doctors, researchers, and other medical professionals to make sense of the massive (and rapidly growing) volume of research papers about the disease.

With COVID-19 affecting hundreds of countries, areas, and territories, IBM deployed Watson to help government agencies and healthcare organizations throughout the world use AI to put critical data and information into the hands of their citizens. Think about it: at a time when it's imperative to get critical health information out to its citizens, two hour wait times to get COVID-19 questions answered could literally become a question of life or death. IBM Watson Assistant for Citizens helps governments use AI to understand and respond to common questions about COVID-19.

Theme #5: AI Will Disappear into the Background and Become Boring

Our final theme is that AI will cease to be a separate domain altogether by 2025. Just as today every company has turned into a technology company, in five years every company will be an AI company. By that we mean AI will be embedded into the very way businesses operate. If your organization is not completely invested in and infused with AI by 2025, there's a good chance it will have lost its competitive edge.

This situation is similar to the one that one of us described in a 2016 blog post that ended up becoming a book: *The End of Tech Companies*. It created quite a stir, describing a fundamental shift happening in the world. The argument was that it didn't matter if an organization was a retailer, a manufacturer, a healthcare provider, an agricultural producer, or a pharma company; traditional distribution models, operational mechanics, and methods of value creation were going to radically change by 2020, and companies that didn't embrace digital technology would be left far, far behind.

Now that we've reached that date, we want to take note that what was predicted has mostly come to pass. With some minor exceptions, every company, large and small, is now in the tech business. They all have to be. Without technology driving everything from supply chain management to marketing to sales to distribution, companies would flounder.

It's been said (repeatedly) that we are experiencing a fourth Industrial Revolution: the internet, data, the IoT, and software are replacing industrialization as the driving force of productivity and change. This observation wouldn't be accurate anymore without adding AI to the list. AI is already becoming mainstream; in five years it will be so ubiquitous that companies without it will incur a huge disadvantage (if they haven't already been driven out of business).

So, here's our bold new prediction for 2020: *This is the beginning of the end of "AI companies."* By 2025 there will only be "companies," all of them steeped in AI technology. Embracing AI will be key to survival.

Future AI Use Cases for Business

The most noticeable trend in business applications of AI technology is that the rate of adoption is going to rapidly increase as more and more companies, large and small, navigate and master the principles of the AI Ladder.

Already, AI permeates the lives of consumers and businesses alike. In the future, we won't even make connections between AI and the amazing new products and services that are broadly available on the market, or that power our organizations. It will be that ubiquitous.

AI applications will increasingly feature natural language processing (NLP) and natural language understanding (NLU). By this we don't mean just the familiar "smart assistant" applications like Apple's Siri or Amazon's Alexa. Those are impressive tools, but we consider them essentially speech recognition systems with minimal task understanding. (Try talking to Siri for about two minutes straight and see what happens.) With NLP and NLU we're talking about the ability to parse human language, whether written, spoken (longtail interactions), or signed, to derive meaning and intent. AI will read everything from medical textbooks, scientific papers, and historical novels to juicy tax law and legal briefs. The smart assistant of the near future will be able to schedule appointments (as Google showcased), organize meetings, and plan agendas based on simple directives from which it will infer meaning. AI will be trained to analyze documents and communicate their content in ways most useful to humans. Consider Project Debater, developed by IBM Research. You can speak—debate—with this AI for minutes at a time, and it keeps up. This an AI-powered, computational argumentation tool that absorbs massive and diverse sets of information and perspectives to help people build more persuasive arguments. Language translation capabilities will continue to improve as well. Human language is the nervous system that allows communities, corporations, societies, and all forms of social organization to function. AI will become part of this system.

AI-assisted analytics will evolve from "searching" to "exploring." Consider the difference between planning a drive from New York City to Billings, Montana, and planning the first-ever trek to the South Pole. In the first case you generally know what to expect, but you appreciate some help searching for the best route or good places to stay along the way. In the second case you have very little idea what to expect, other than cold and danger. Similarly, language models have frequently been used for hypothesis testing. In the future, these models will be used for unsupervised exploration of the data space, to experiment with previously unseen patterns, through AI systems based on reinforcement learning (where training is based on rewards or penalties for successful or unsuccessful outcomes).

This evolution from hypothesis-driven to data-driven research is already being seen in the search for new compounds and molecules that could be used to cure diseases, curtail pandemics, address ailments, and more. There are uncountable billions of possible therapeutic molecules. What are their properties? How might they be used as medicines? What's the best way to synthesize them? AI will assist human scientists in finding the answers to questions like these. In countless other fields, AI will allow knowledge workers to "let the data guide them."

Let's take a look at some of the many, many business use cases that will be empowered by AI in the next five years.

Cybersecurity

In cybersecurity, we're going to see AI systems used to both prevent and mitigate threats, as security solutions will increasingly use machine learning to recognize threat attack vectors. AI will be like steroids for security incident and event management (SIEM) solutions. (We expect criminals to up their "attack" game using AI as well.) By bringing together data from nontraditional sources to augment classic intrusion detection and intelligence analysis—say, from social media—such solutions will become more adept at "listening" to what the cyberworld is telling it about pending or potential cyberattacks or phishing schemes. Thanks to NLP, NLU, and pattern detection, finally security professionals will have intuitive natural language interfaces from which they can more easily gather security intelligence—and share it with business executives—to accelerate mitigation efforts. Additionally, an AI-enabled IoT can easily flag potential breaches or anomalies in both virtual and physical defenses.

Among the best use cases of AI for cybersecurity will be malware detection, fraud detection, and network and intrusion detection—these classic cyberdefense actions will only get better and faster as AI continues to evolve. For example, AI will be able to learn how authorized users behave, and will be able to detect even the smallest deviation from that behavior. AI algorithms will be able to collect data such as behavioral deviations typical of hackers, and stop attacks that are in process.

In fact, AI will be an imperative in cyberdefense because the bad guys are using it, too! We will need AI in order to take on their AI.

Autonomous Driving, Autonomous Everything

Autonomous driving is one current and popular application of AI that will continue to advance over the next five years. Today we're seeing widespread manufacturing of semiautonomous vehicles that still require drivers but incorporate AI through the use of advanced driver-assistance systems (ADAS); for example, using computer vision to interpret speed limit signs and display the limit on the dashboard. As this technology continues to improve, we will eventually see fully automated driving. According to McKinsey, AI will transform the automotive industry; the next decade will see amazing advances that will reduce travel time as cars automatically analyze traffic on the fly and enable more efficient usage of ride-sharing apps. Tokyo and other cities like Phoenix, Arizona already have self-driving taxis. Shipping and delivery costs will likely plummet too as more and more aspects of driving (like trucking, last mile delivery, and so on) become automated.

Other benefits of autonomous vehicles may include:

- Fewer car accidents caused by drivers or weather conditions
- Reduction of aggressive driving or "road rage"
- Enabling transportation of disabled and older persons
- More efficient use of transportation infrastructure

Why "autonomous everything"? Autonomous vehicles are actually the tail end of the trend, because they involve human life and need to earn our full trust before being unleashed freely in our world. Autonomous machines that will cut your lawn and vacuum your home are commonplace, all with on-board AI. 2019 saw the release of an autonomous robot whose sole purpose is to deliver toilet paper in times of need—another example of AI that needs to be trustworthy!

Conversational Digital Agents and Personal Assistants

Another business use case for AI is conversational agents. This is actually a $2.5 billion market, so it's getting people's attention. We predict that in three years, any company on the path to a digital transformation will have virtual agents incorporating AI. Virtual agents are already used widely in customer call centers and at home ("Alexa, play some jazz"), but the abilities of these agents are currently quite limited. As AI research progresses, conversational agents will improve to handle all sorts of tasks.

For example, Mastercard's KAI, launched in 2017, is an always-available "virtual bank teller." Consumers can ask the bot questions about their accounts, review purchase history, monitor spending levels, learn about MasterCard cardholder benefits, and receive contextual offers. These kinds of bots are now beginning to show up everywhere, and within five years their language skills will be so sophisticated that they will be impossible to distinguish from human agents. (Our ethical sense strongly suggests that whenever a client is interacting with AI in an environment where it's impossible to distinguish between the AI and a human agent, the client should be informed they are interacting with AI. If it's helping them, they won't mind.)

Real Estate

AI will revolutionize real estate. It's already commonly used for managing information: collecting data on properties, analyzing documents, verifying property parameters, and doing real-time translations for international buyers. Housing valuations are also being enhanced by AI, which can even incorporate pictures into assessments automatically.

AI has the potential to enhance the workflows of all real estate market participants (buyers, sellers, brokers, asset managers, and investors). It can provide prospective buyers with tailored recommendations and predict whether they will be able to keep up with loan payments, and it can help renters who are searching for a property connect with like-minded roommates. In the coming years, AI will infuse this industry at scale.

Soon, AI will automate energy, fire protection, and security by providing 24/7 monitoring and control of commercial and domestic facilities. In businesses, this means that smart buildings will turn lights and HVAC services on and off as needed (and get it right, we may add—if you've experienced this technology today, we're quite certain there is still some room for improvement), enhancing the value of properties and minimizing the cost of ownership. In private households "smart home" devices will automate everything, from the living room to the kitchen to the garage. Smart homes will be easier to sell and command higher prices, just like today's homes with luxury kitchens.

Retail

A recent global study found that two in five retailers are already using AI. That number is expected to double within the next year.

Indeed, the retail industry has come a long way in the last decade. From high-end fashion brands like Dior to general merchandisers like Walmart, retailers are amassing mountains of data on customers' preferences from both conventional and unconventional sources, and trying to understand it to do everything from product development (identifying which styles to manufacture, in what quantities) to business operations such as forecasting sales and streamlining inventory.

Today, retailers are starting to use AI and advanced analytics to take these activities a step further. For example, Asos's Style Match app (Figure 11-3) allows consumers to upload photos of clothing—something a friend wore, an item from another store, or even a garment they saw on television—and uses image recognition to search its catalog and return Asos products to the shopper that resemble those in the photo.

Sports apparel manufacturer Lululemon uses AI to collect and sort feedback by design and then to optimize its supply chain and put new designs into production that resonate with consumers.

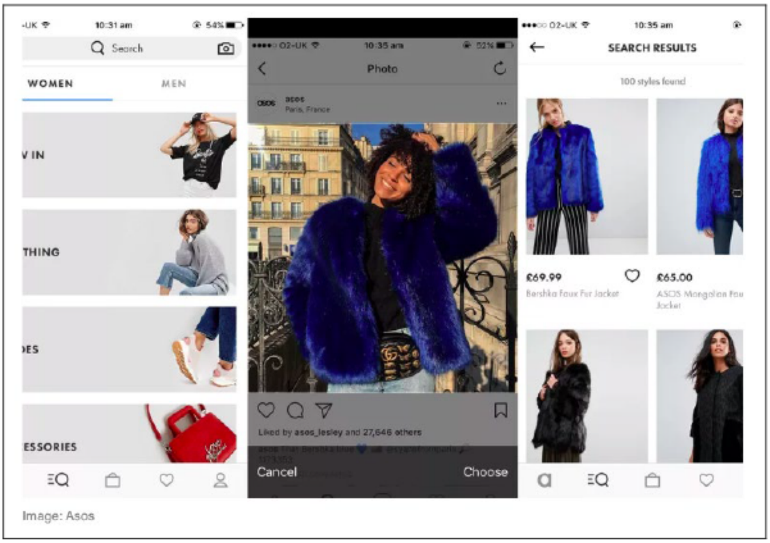


Image: Asos

Figure 11-3. Screenshot from the Asos Style Match app

Insurance

According to McKinsey, the insurance industry is on the verge of a "seismic, tech-driven shift."

AI is already transforming insurance distribution and underwriting. With the help of AI algorithms, policies are being priced, purchased, and bound in near real time. But within 5 or 10 years, AI systems will be capable of so much more.

Take the process of purchasing insurance. It will be vastly accelerated, with much less input and paperwork required by both the insurer and the customer. Sufficient data will have been collected about both individuals and businesses for AI algorithms to create risk profiles, so that cycle times for completing the purchase of an auto, commercial, or life policy will not only be reduced to minutes or even seconds, but hyper-personalized too. As AI enters the underwriting domain, carriers will be able to assess risk in a much more granular way, ushering in a new wave of mass-market instant products.

Insurance will also be transformed from an annual purchase/renewal cycle to a continuous model, with products being offered to adapt to an individual's activities and behaviors. We'll also see micro coverage at scale—for example, phone battery insurance and flight delay insurance paying out different amounts based on the delay—weather, maintenance, or congestion. We'll see product warranties that offer different coverage for specific refrigerator or dryer components, leading to "a la carte" risk pricing models that consumers can customize to their particular needs.

Claims processing in 2030 will depend on advanced AI algorithms to handle initial claims routing, reducing processing times from days to hours or even minutes. IoT sensors and other methods of capturing data—such as drones—will replace manual efforts previously carried out by humans. Claims triage and repair services will be triggered automatically. For example, if a policyholder gets in a car accident, they can simply take streaming video of the damage; that video will be instantly translated into loss descriptions and estimated repair amounts, and a claim filed with minimal involvement from the policyholder.

Inside the home, IoT devices will be used to proactively monitor water levels, furnace performance (and safety), and other risks, and will proactively alert both policyholders and insurers of issues before they become serious. We've all known someone who has experienced a refrigerator water leak that ruins a kitchen floor, or a furnace that fails in the middle of the night. With AI, all of that will be avoidable.

Customer Service

According to Gartner, up to 80% of customer service interactions will be handled by AI this year (2020). More—and more naturally speaking—AI-powered agents are backed up by powerful backend databases of product information, customer profiles and activities, troubleshooting scripts, and other automated tools.

Many, if not most, businesses have already decided that it is not cost-effective for human staff members to answer simple questions. And transferring customers to multiple agents just to get a query answered is hardly a good way to retain them. Instead, these assistants with NLP or NLU superpowers will be able to handle most calls, passing only the most difficult ones over to humans.

In the coming years AI will increasingly perform more advanced tasks that belong to the tedious and routine part of every business, such as monitoring and analyzing feedback, reviews, support tickets, and other forms of communication. AI algorithms will be able to confidently derive meaning from what is written or said, interpret it, and even generate a proper reply. AI will be able to identify sentiment and underlying implications in both written and spoken language, sensing what emotions are being expressed: sorrow, happiness, rejection, or displeasure.

In short, customer service is set to change forever. All those calls that were "monitored for quality assurance"? Only about 2% of them were ever proactively monitored to ensure you had a good experience; they were mostly used to investigate complaints, long after the damage was done. Now the other 98% will come into play, improving customer service in real time.

The Future of Work in an AI-Driven World

The nature of work will continue to change rapidly in the coming years. Some jobs—notably those that involve simple, repetitive tasks—will simply go away (by "go away," we want to be clear that we expect AI will augment human intelligence, freeing human actors up for other tasks, not replace humans as professionals). Other jobs—those that require knowledge, flexibility, interpersonal skills, and creativity—will evolve. The role of insurance agents, for example, will change substantially in the next decade, according to McKinsey. As AI-infused technology becomes the norm, agents will become primarily educators of consumers, helping them manage "portfolios of risk coverage." Entirely new jobs will emerge too, as the landscape of work evolves.

MIT and IBM Watson AI Lab conducted a study in October 2019 on "The Future of Work". In it, they proclaim that AI is likely to change how every job is performed. The study showed that AI is beginning to redefine the nature of tasks performed in certain jobs as automation gains ground.

"As new technologies continue to scale within businesses and across industries, it is our responsibility as innovators to understand not only the business process implications, but also the societal impact," said Martin Fleming, IBM's Vice President and Chief Economist, in a statement. "To that end, this empirical research from the MIT-IBM Watson AI Lab sheds new light on how tasks are reorganizing between people and machines as a result of AI and new technologies. All of these changes will put enormous strains on organizations, and managers with the skills to guide them through these changes will be in great demand. Training for the AI revolution will become a significant business opportunity."

The news from IBM and MIT is actually quite reassuring—especially compared to some of the more dour predictions that have declared that as many as 40% of jobs will disappear within the next 15 years. While most jobs will evolve as AI is put into production, the new research demonstrates that few jobs will disappear altogether. What will change is the way we work.

The researchers noted that for each occupation, on average, across more than 18,500 tasks, workers performed just 3.7 fewer tasks overall in 2017 than in 2010. Indeed, when looking at the impact of AI on tasks across the seven years of the research data set, IBM and MIT found that of the tasks that are more suited to being done by AI—for example, scheduling or validating credentials—workers were asked to perform 4.3 fewer tasks. Of the tasks that are less suitable for machine learning (design work, or tasks that require specialized industry knowledge), workers were asked to perform 2.9 fewer tasks.

A Deeper Dive into AI and Edge Computing

We've mentioned the Internet of Things as a fertile producer of data for AI systems. And we think it's worthwhile to take a deeper dive into the single most important new technology trend we'll be seeing over the next five years: how the Internet of Things (IoT) will evolve into the Internet of Edge Computing (no acronym yet!). IDC estimates that the edge computing market—which barely existed a couple of years ago—will be worth $34 billion by 2023. This includes the hardware and software and components required to complement the processing done at larger data centers located far away.

Back when we were discussing the need to modernize data infrastructure in Chapter 5, we noted a piece of advice from Daniel Hernandez, IBM's VP of Data and AI: "You should bring AI to the data." That's what edge computing is all about. Instead of sending data in raw form over networks to be processed in a data center, cloud, or other centralized site, information is analyzed and acted on where it's gathered. Coupled with faster 5G networks and artificial intelligence, edge computing will give rise to new forms of analyses. The idea is to leave the data on the devices that collect or generate it, and *bring intelligence to them*—to the edges of the network—rather than bringing the data back to some central site. Just as the premise behind Hadoop was to ship function to data, edge computing ships intelligence to data. As Ruchir Puri, Chief Scientist and IBM Fellow at IBM Research, says, edge computing is "placing enterprise applications and [their] components closer to where the data is created, and where the action needs to be taken." To help clients journey into this emerging area, IBM released its Visual Inspector platform in 1Q2020—a device management-to-model deployment and inferencing platform designed for edge computing environments.

This definition of edge computing is enterprise-centric, and that's why it's important to us. To us, edge computing isn't primarily about cell phone apps or home automation. It's about monitoring thousands of machines on manufacturing shop floors; it's about monitoring and reacting to environmental conditions in data centers; it's about managing anything that generates data, and moving the AI to the data so that decisions can be made locally, and more quickly.

Edge computing will require its own infrastructure. Ethernet and WiFi can't handle the huge number of devices that need to be tracked. A factory floor might have 1,000 machines; each machine may have hundreds or even thousands of sensors, looking for signatures in vibrations or temperatures that signal impending issues. Those sensors themselves may have very limited capabilities—for example, they might be peel-off stickers with built-in batteries, and just enough intelligence so that they know when to wake up and send a packet of data—or they may be full-fledged computers. (In practice, the power required to operate a radio is an order of magnitude greater than the power needed to run a processor. Keeping these devices "quiet" has a big effect on battery life, and hence on maintenance. We've written a lot about how AI eliminates boring, repetitive tasks; we will have failed if you're hiring people to replace thousands of batteries a day.) These devices will connect to an edge server through a local edge network that keeps the traffic off of the regular corporate network. The edge servers collect data, take what action is necessary, and then send a summary back to the central server.

That summary could be as little as an alarm when something goes wrong at the edge; or it could be a digest of the data; or the edge systems could use machine learning to build a partial, local model, which is sent back to a central server, where it is used to assemble a global model from all the local models (resulting in a centralized edge-crowdsourced boosted algorithm). The global model can then be sent back to the edge, where it represents a more complete view of the data than the local model. This is called federated learning, and you may be using it already: Android phones use federated learning to build a local model of the user's actions and ship that model back to Google when the phone is inactive.

Why is edge computing so important? First, edge devices will generate a lot of data—as mentioned earlier, the volume could approach 80 zettabytes in the next five years. And none of that data will be of any use unless it can be processed, analyzed, and understood on the spot, using edge computing. These edge computing systems will need to be able to filter out data, decide what to keep locally and what to send into the cloud, and analyze information in milliseconds to deliver the promise of true, real-time automation.

Moving computation to the data also minimizes latency. Although 5G networks promise significant reductions in latency, we expect that it will still be a problem for sensitive applications such as autonomous vehicles, smart grids, industrial automation, remote surgeries, and drone management. Even if 5G "solves" latency temporarily, expect it to rear its head again. If you have a jet engine that's showing signs of impending failure, you can't afford to wait for a 1,000-mile round trip to a central server to take action.

A bigger communications stream means more opportunity to fill the pipe with data. Just as data expands to fill the storage you have available, data also expands to fill the bandwidth you have. Today you hear about how 5G will allow you to download a movie in seconds as opposed to minutes. But with a bigger pipe to a device, demands for more fidelity arise. We're already seeing this on the consumer side of things: 4K and 8K video streaming doubles and quadruples the data rate, and pretty soon all your photos will be 30 MB because you want that high resolution even though you don't likely need it.

Security and privacy are also important for edge networks. Data that's in transit is data that's vulnerable. There are many regulatory domains where sending data between companies raises legal issues. For example, a smart MRI machine might not be allowed to send patient images to a cloud server for analysis; analyzing that data might have to be done locally. That's edge computing. If you have a smart building, you probably want the central application to send weather predictions and other data out to the edge servers, but you don't want to send gigabytes of HVAC data back to central servers in the cloud; you want to make those decisions locally, even on a room-by-room basis, and send summaries. (There are many things that data could reveal to those engaging in corporate espionage.) The bottom line is this: you want to move the AI to the data, not the data to the AI.

Now, what does the future of edge computing look like?

In the future, the Internet of Things, AI, and edge computing will merge to create what we've dubbed the Intelligence of Everything. AI can be used in IoT platforms for root cause analysis, predictive maintenance of machinery, or outlier detection. Devices like cameras, microphones, and other sensors will constantly collect data that is used to train AI applications in the cloud and inferenced on the edge. Do you want smart tomatoes? We can instrument greenhouses with cameras that can detect diseases. The next step beyond autonomous vehicles is networked autonomous vehicles that can negotiate with other cars for merges, lane changes, and hazard avoidance. These technologies are coming fast, and we'll see them pushed much further. But this vision will fail if every device has its own API. We already see that problem with home automation: which devices work with which apps? If you're an Apple user, are you locked into Apple-partner–manufactured door locks, dishwashers, thermostats, and televisions? That's a manageable problem for a home, but an untenable problem for an enterprise. A "smart car" that can only negotiate lane changes with other Fords is next to useless. As we learned during the evolution of the internet, interoperability is essential, and interoperability doesn't happen without standards.

The Linux Foundation announced recently that it had created a new open source project called LF Edge that counts 60 of the biggest names in technology as members. LF Edge is "an umbrella organization with the goal of creating an open, interoperable framework for edge computing independent of hardware, silicon, cloud, or operating system." By bringing together industry leaders, it aims to foster collaboration across the many areas that stand to be transformed by edge computing, including the industrial manufacturing, energy, transportation, retail, home and building automation, automotive, and healthcare industries.

And, of course, a number of commercial companies are moving into this exciting new tech arena. Hailo, an Israeli startup, is developing a specialized AI deep learning processor that it hopes will empower intelligent devices on the edge with the performance of a data center–class computer. Because they will operate in real time at reduced power consumption, size, and cost, Hailo processors promise to enable edge devices to go beyond handling sensors and streaming volumes of data to actually do remote processing themselves.

Swim, a Silicon Valley edge computing startup, has developed DataFabric, an edge computing solution that uses AI to allow businesses to connect real-time data streams from distributed devices and systems. Its mission is to enable businesses to transform, analyze, and act upon data generated at the edge, at the speed at which it's generated.

Using the Edge and AI for Good

One of the world's premier accessibility researchers gives us some insight into what edge computing will mean for the disabled.

IBM Fellow Dr. Chieko Asakawa has dedicated her career to developing technology that makes the world more accessible for people with disabilities. Blind since the age of 14, she has helped develop several pioneering accessibility technologies, including the earliest practical voice browser in the 1990s, which further opened the internet to the visually impaired. As a visiting faculty member at Carnegie Mellon University, she is now leading an effort to develop an AI-powered navigation system for the blind and other disabled populations.

According to Dr. Asakawa, AI is going to help blind people "see" the world—and explore it. "Right now, we are working on the Cognitive Assistance Project for Visual Impairment," she said in a Q&A with IBM. "People with vision see the things around them so they always have some context. For us [the blind], we don't have contextual information. We can get it from AI, but only when technology like computer vision is connected to knowledge and when the Internet of Things provides location information."

A critical point is that vision and knowledge need to be communicated to humans very quickly—for example, when an AI system is giving directions while a visually impaired person is walking in a city, context needs to be provided as that person is moving. Dr. Asakawa's team is working on this in its open source NavCog app; see Figure 11-4. "[These kinds of tools can help] the elderly, too. They need elevators and maybe help finding shops or recognizing items in a store if their vision isn't good," she says. But GPS doesn't help when someone is indoors.

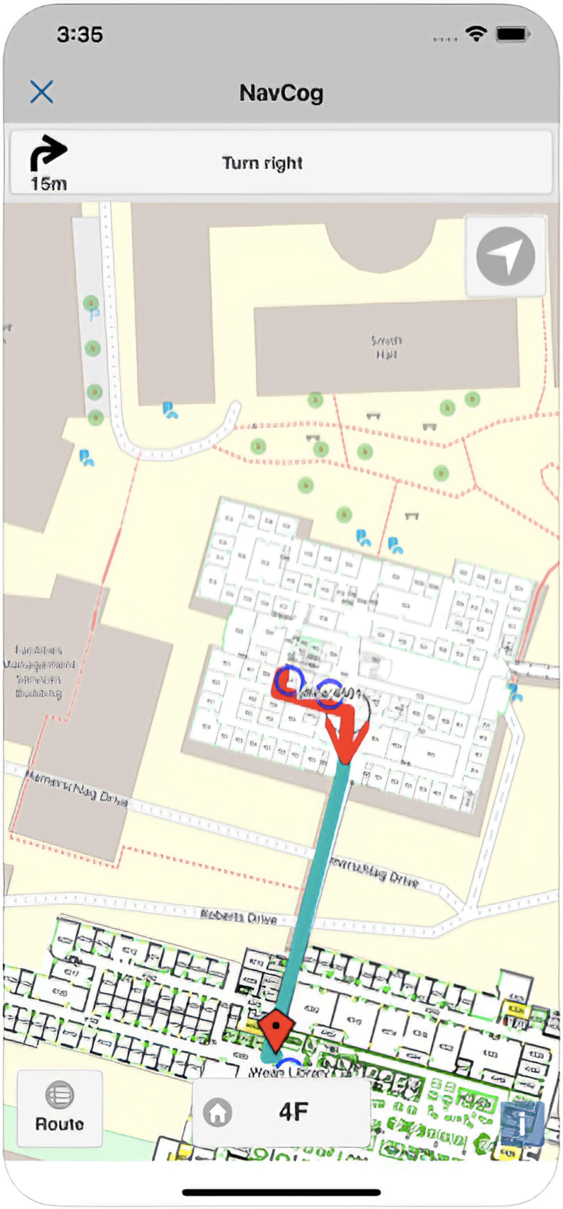


Figure 11-4. Screenshot from NavCog app

The new tools can benefit even those people without accessibility issues. “Think about when you travel to a foreign country and you can’t read the signs or food labels or find a type of shop because you don’t know the language,” she says. At such times, these tools (which exist today) would be useful to a broad population of potential users.

Infrastructure work needs to be done, of course. To provide the detailed information that the visually impaired need to explore the real world, beacons have to be placed every 5 to 10 meters. "These can be built into building structures pretty easily today," says Dr. Asakawa.

But a complication is that radio waves aren't always the same—they move. "So we have to use machine learning to help the system calculate your most likely location," she says. Thanks to machine learning, her team can now achieve accuracy of location within one to two meters. "We are continuing to work on the machine learning algorithms to improve accuracy and reduce the number of beacons that need to be installed."

Thirty years from now, Dr. Asakawa says the world can expect a broad range of disabilities to be augmented by AI, because AI-based cognitive assistants will be able to supplement any of the five senses. "So imagine in the future, you will be able to access information any time without vision or without hearing," she says. "With AI, many disabilities will no longer be as big of an issue."

Conclusion

We have been urging a shift in perspective on AI. We realize that people are anxious about losing their jobs because of AI. There are even some existential fears about what AI will do to humanity.

Rather than imagining what will happen if AI replaces humans, we encourage you to think about how AI will augment human intelligence. It's not a competitor; it's an assistant. A well-designed AI program should make people more productive, as they allow machines to take on the more time-consuming (and tedious) repetitive tasks that previously weighed them down.

Indeed, the real value of AI is in freeing humans up to do what humans do best—to create and intuit and make complex decisions, to come up with ideas about new lines of business and new opportunities for growth—rather than wasting time on tasks that probably don't add much value.

AI promises to do all this, and more. We're confident that you're steeped with AI acumen, have a plan, and are ready to climb the AI Ladder. Bring on the next five years of AI innovation!

Index

About the Authors

Robert D. Thomas is senior vice president of IBM Cloud and Data Platform. He directs IBM's product design and investment strategy, expert labs, global software product development, marketing and field operations across the company's vast software portfolio.

Rob was previously the General Manager of IBM Data and Watson AI. Under his leadership, IBM has emerged as a leader in data and AI, spanning databases, data integration and governance, business intelligence, financial planning, data science and AI tools, and AI applications. Since joining IBM's software unit, Rob has held roles of increasing responsibility and overseen 4 acquisitions by the firm, representing over $2.5 billion in transaction value.

Born in Florida, Rob earned his BA in economics at Vanderbilt University and earned his MBA at the University of Florida, Rob serves on the board of Domus (Stamford, CT), which assists underprivileged children in Fairfield County. Follow him on Twitter: *@robdthomas*.

Paul C. Zikopoulos is the vice president of cognitive big data systems at IBM. Paul is an award-winning professional writer and speaker who has been consulted on the topic of big data by the popular TV show *60 Minutes*. Most recently, Paul was named to the Analytics Insights list of "2019 Top 100 AI & Big Data Influencers," and has appeared on over a dozen other global "Experts to Follow" and "Influencers" lists. You'll find Paul taking a very active role around Women in Technology (he's an advisory board member for Women 2.0), as well as other causes where he sits in an advisory board role (Coding for Veterans and the Masters of Management Analytics & AI programs at Canada's prestigious Queen's University). Paul has written 20 books (including four *For Dummies* titles and over 360 articles) during the 25 years he's been focused on data. He doesn't think NoSQL is something you put on a resume if you don't have SQL skills, and he knows JSON is a technology, not a person in his department.

Paul's always keeping with his grass roots...a newbie with no computer courses before coming to IBM. He knows on his dumbest days that he's never as dumb as he feels—and on his smartest days, he's never as smart as he feels either. Ultimately, Paul is trying to figure out the world according to Chloë—his daughter, whom he notes didn't come with a handbook and is more complex than the topics of big data and AI, but more fun too. Follow him on Twitter: *@BigData_paulz*.

Colophon

The animal on the cover of *The AI Ladder* is a douroucouli, or grey-legged night monkey (*Aotus lemurinus*). This primate can be found in Central and South America. Though most commonly seen in viny, dense forests and rainforests, this monkey occupies a variety of habitats.

The grey-legged night monkey, also sometimes known as the owl monkey, has large, round brown eyes set into a distinctively small face. There is a large black spot between its eyes, and the tip of its tail is black as well. The night monkey's fur is wooly and most often grey in coloration; however, its underparts can range from a pale yellow to a bright orange. It has long, slender fingers with padded tips.

Night monkeys are nocturnal and most active well after dusk. During the day, they sleep in tree hollows, under the cover of vines, or else in dense brush. They live in small family groups (often adult pairs and their offspring) and produce a wide range of calls, including owl-like hoots. They get their food largely from the canopies they inhabit, and eat everything from fruit and flower nectar to foliage and insects. They have also been known to occasionally eat small birds and mammals.

Night monkey populations have suffered as a result of collection for biomedical research and some hunting. Their conservation status is Vulnerable at the time of this writing. Many of the animals on O'Reilly covers are endangered; all of them are important to the world.

The cover image is by Karen Montgomery, based on a black and white engraving from *Natural History of Animals* by Vogt & Specht. The cover fonts are URW Typewriter and Guardian Sans. The text font is Adobe Minion Pro; the heading font is Adobe Myriad Condensed; and the code font is Dalton Maag's Ubuntu Mono.