PERSPECTIVES FROM THE FRONT

Data quality, AI performance and trust

Data & Trust Alliance

+ IBM®

With respect to AI, **data quality** generally refers to the accuracy, completeness, consistency, timeliness, uniqueness and validity of a given dataset, as well as the data's fitness for the purpose for which it is being used. But the practical meaning of "high-quality data" in any given context will depend on the needs of the organization and the specific use case involved.

IN TODAY'S ECONOMY, data informs every decision enterprises make, from product development to supply-chain management to customer billing. A failure to understand and manage the quality of your data can create significant legal, financial and reputational risk, and will ultimately limit the ability of an enterprise to innovate and thrive.

"Data is the food for AI. If you don't know the quality of the data, you don't know the quality of the AI results you're getting."

—**JOANN STONIER,** fellow of data and AI at Mastercard

These concerns are especially urgent as business leaders race to embrace AI for both the automation and the innovation it enables. AI models and the business tools they provide are only as effective, and as safe, as the data on which they are trained and from which they continue to learn as they interact with customers and other sources of information. "Data is the food for AI," says JoAnn Stonier, fellow of data and AI at Mastercard. "If you don't know the quality of the data, you don't know the quality of the AI results you're getting."

Simply put, algorithms trained on bad data equate to bad business outcomes. Because most enterprises today depend on data — their own as well as data they access from third parties — business leaders need to take a measured, value-driven approach to data quality. There are three critical steps in this process.

STEP ONE

Determine what data quality means for your company

Defining "quality" is highly contextual for any organization, as it depends on each company's business goals and use cases. "At different companies, there are different focus areas and different aspects of data quality that matter," says Ioana Mazare, vice president of enterprise data strategy at UPS.

Regardless of contextual differences, however, there are baseline considerations that are important to every organization when it comes to the data used in AI applications. World Economic Forum identifies six core elements of data quality that should be assessed when training and vetting AI — whether built in-house or bought from a vendor:

- **ACCURACY:** Can you confirm — and document — that the data you're using represents actual subjects in the world?
- **COMPLETENESS:** Can you ensure the data is comprehensive and that all its values, or fields, are complete?

- **CONSISTENCY:** Can you ensure that data stored in multiple places, across networks and applications, is stable and consistent in its formatting and the values it represents?

- **TIMELINESS:** Can you measure the delay between when data is generated and when it is used, and ensure that the delay does not compromise the accuracy of the data?

- **UNIQUENESS:** Can you identify any duplication or overlaps across the data sets you are using to train your AI models?

- **VALIDITY:** Can you ensure that your data is captured in the correct format or syntax for its intended purpose or use case, including metadata such as valid data types, ranges and patterns?

> "At different companies, there are different focus areas and different aspects of data quality that matter."
>
> —**IOANA MAZARE,** vice president of enterprise data strategy at UPS

While there are other aspects of data quality that may be important for organizations' individual use cases, these six considerations provide a general picture of what constitutes "good data."

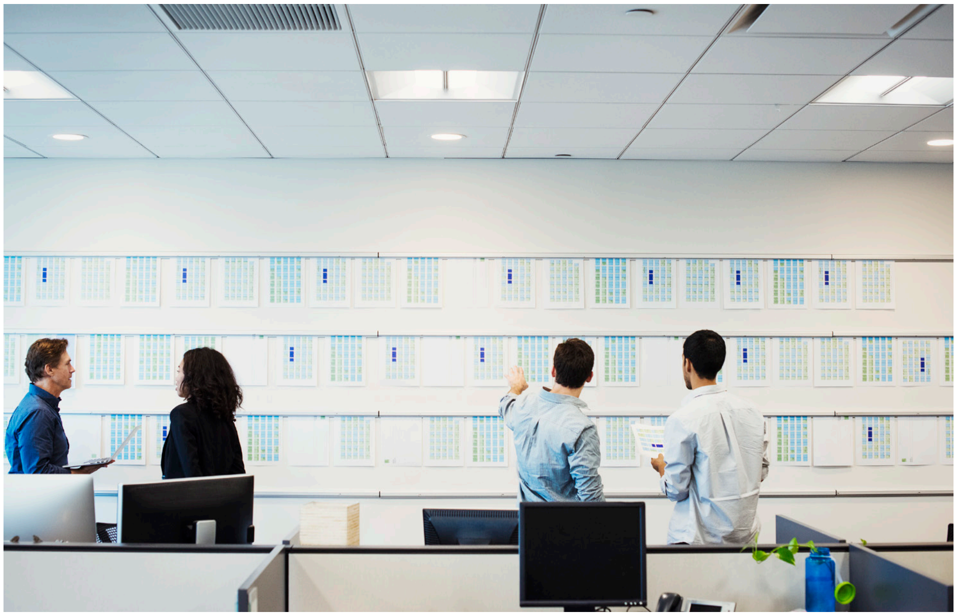
Two additional dimensions are key to ensuring data quality for AI:

Data provenance is defined by the World Wide Web Consortium as "information about entities, activities and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness." Provenance encompasses the details of how your enterprise accessed the data and any relevant legal rights, which is particularly important with third-party data. Provenance is partially captured in metadata that details a data packet's origin and any changes made to it, along with timestamps. By automating metadata creation, companies can help detect and weed out irrelevant or duplicated data, as well as data that may have been tampered with, "weaponized" or "poisoned."

Fitness for purpose is about whether the data represents the appropriate population, market or factors for the use case of the AI tools being trained. If the data you choose for training your AI is not appropriate for your purpose, the AI will fail. For example, if a financial services algorithm trained on European market data is used to analyze markets in Africa, the model won't work effectively.

Importantly, historical datasets may reflect historical biases, and AI trained on those datasets will perpetuate those biases. "An important part of data responsibility is ensuring we have fulsome, complete datasets that accurately reflect all communities we serve," says Manav Misra, chief data and analytics officer at Regions Bank. "When we train AI tools and models with a truly holistic view, we are adhering to the ethical and responsible use of data, and the result is greater convenience and financial inclusion for all."

Issues of provenance and appropriateness are a constant concern for Dena Mendelsohn, privacy officer and compliance manager at digital healthcare startup Transcarent. "We need to not get overly excited about an extremely large dataset that apparently could solve all of our problems," Mendelsohn says. "We need to find out the source of the data, who controls the data and what individual or corporate rights apply, and whether it has relevance to our intended use and is going to get us to our intended outcome."



Data provenance and fitness for purpose are particularly important in the context of the new breed of publicly available generative AI, which has been trained on enormous datasets from the internet. "Having the ability to understand the provenance and lineage of your data is going to become more and more important to firms that want to use generative AI for commercial purposes," Stonier says.

David Cox, vice president for AI models at IBM research, says provenance in particular is key. "Every business leader should be asking where their AI data comes from, and whether it's been cleaned and prepared for training," Cox says. "If tomorrow there's a problem within a certain dataset, enterprises need a commitment that the data will be removed and the model will be retrained."

Major providers of data storage, management and analytics systems, including AWS, Azure, Google Cloud and IBM, as well as many specialized firms, offer tools to manage data provenance.

STEP TWO

Create enterprise-wide standards and governance

Creating consistent standards for data quality is essential to the ethical and responsible use of AI. There are no universal standards for data quality in business, however, as data quality means something different for every company. Therefore, it's critical for organizations to define their own data quality standards — and do so with rigor.

What does "rigor" look like? Bernardo Tavares, chief technology and data officer at Kenvue (formerly Johnson & Johnson Consumer Health), created a Data Quality Index (DQI) metric, which is expressed as a percentage. "In the case of product data quality, DQI is measured as the percentage of records that are available, accurate and connected across all of our systems," Tavares says. "We do that by applying automated business rules. It took us over a year to define that metric." His team started with a task force that focused on the most profitable products and brought the data to a 95% DQI.

Today, the company can compare products, from Listerine to Neutrogena, by their DQI percentages. "It's a number that we track at the C-suite level that is easy to remember and report on," Tavares says. Having shown the efficacy of the DQI for high-value parts of the business, Tavares' team was able to get funding to build open-source tools for "golden record" creation and cleansing, and to begin expanding their initiative. "Slowly but surely, we're creating this notion of the DQI for all of our core entities," Tavares says.

Data quality standards are the foundation for data governance. An organization's chief data officer (CDO) most often is responsible for creating those standards and implementing a governance framework for enforcing them, as well as working with data analysts, data scientists, software developers, lawyers, compliance teams and anyone using AI tools across the enterprise.

> “Don’t boil the ocean. Pick an area where the greatest pain point is. Start with a key question that you’re not able to answer because your data is not good.”
>
> — **BERNARDO TAVARES,** chief technology and data officer at Kenvue

According to a recent Gartner report, increased investment in data and analytics, and the demands placed on organizations’ data teams, “reflect a growing confidence in chief data and analytics officers’ abilities and recognition of the data office as an indispensable business function. However, this leads to more work as pressure grows for D&A to achieve tangible business results.”

If your organization doesn’t have a CDO, don’t let that stop you from acting — as Stonier says, “chances are you have somebody in your firm who has been looking at data quality for a while.” She advises CEOs to “look to any existing roles or functions with the expertise. It could be a CIO, someone in your IT department or someone in business intelligence.”

Beyond a company’s specific data-quality management processes, the increasingly pervasive impact of these new capabilities will require fostering a data-and-AI-focused culture.

STEP THREE

Create a companywide culture around data management

“We had to create a culture around data,” Tavares says. “This started with the declaration that data is a real asset, and an understanding that we need to have people who are passionate in that space and the tools and structures to manage data through life cycles.”

Building a data culture is essential because the job of data governance never ends. AI models will continually change, as will regulatory requirements — and continuous governance cannot be achieved by simply creating policies. The C-suite must drive ongoing education efforts around data quality, and the CEO must lead. The tone from the top, from the CEO on down, matters because quality standards need to be integrated into every function.

Although building a company-wide culture is a long-term — indeed, never-ending — project, it likely needs to start small. “Don’t boil the ocean,” Tavares says. “Pick an area where the greatest pain point is. Start with a key question that you’re not able to answer because your data is not good.”

For Tavares, this led to a focus on the company’s products, where data assets created the most value for the company.

Even that was challenging. “We got experts and stakeholders in the room, 70 people, and they couldn’t agree on a common data definition for a product,” Tavares says. “If you’re in R&D, if you’re in sales, etc., you look at it differently.” The team came up with a model that defines the basic elements and relationships that truly represent a Kenvue product. Only then did they start working on data quality, by “hydrating” this model with clean, connected data. Helped by AI, automation and a company-wide data quality objective, this iterative approach was effective and later applied to several “Data360” domains.

Finally, fundamental to building a healthy culture around data is remembering that each piece of data represents a human being: your customer, your employee, your partner, and their presence and activities in the world. Focusing on the people in your data quality ecosystem — the people who create, use and are serviced by the AI tools that data underpins — will allow your business to better govern its decisions around the appropriate use of data in AI. “If you design with the human in mind,” Stonier says, “the guardrails become clear.” ●

Key Questions for the C-Suite

- What do we define as **mission-critical data?**
- How are we currently measuring the **health of our data** for AI?
- Are we continuously evaluating our **data standards** as business needs change?
- What **education and training** do we have in place to guide employees on appropriate use of data?

→ Learn more about the Data & Trust Alliance

Get expert insights on AI for business