IBM Software Group IBM Watson Solutions
Whitepaper

THE ERA OF COGNITIVE SYSTEMS

An inside look at IBM Watson and how it works

What is Language and Why is it Hard for Computers to Understand?

Language is the expression of ideas. It is the medium by which we communicate an understanding of things between people. It is how we convey fear, hope, history and directions for the future. Some say it is what enables us to think, speculate, and imagine. It is at the base of our cognition – our ability to understand the world around us; or at least at the base of our ability to manipulate and exchange that understanding.

And it is incredibly imprecise.

Our language is full of innuendos, idiosyncrasies, idioms, and ambiguity. We have noses that run, and feet that smell. How can a slim chance and a fat chance be the same, while a wise man and a wise guy are opposites? How can a house burn up as it burns down? Why do we fill in a form by filling it out?

And yet, it can be amazingly accurate.

We convey so much meaning, and accomplish so much collaboration even in the midst of all the difficulties with language. Somehow, we can see through the gaps, the inconsistencies and contradictions, the irregularity, and the lack of clarity and still understand each other with a great deal of accuracy.

This difference between precision and accuracy is important. Precision is the mechanical or scientific exactness that can be found in a passage of text. We can determine whether a specific word exists within a passage with a high degree of precision. Accuracy is the degree to which one passage infers another passage would be considered by reasonable people to be true.

The answer to “2+2” is precisely 4. Mathematics teaches us this, and that no matter how many zeros you place after the decimal to represent greater precision it will still always derive to 4. But what if when we said “2+2” we didn’t mean it to be taken literally as a mathematical formula, but rather as an idiom for a car configuration – as in two front seats and two back seats. Or a psychologist might use it to refer to a family with two parents and 2 children? In those other contexts, the answer “4” would not be an accurate interpretation of what we’re trying to convey in the language.

In fact, to accurately answer a question you often have to consider the available context for the question. Without enough evidentiary information, it is hard to accurately respond to a question – even though you can precisely answer elements in the question literally.

What if when we said “2+2” we meant a car configuration – as in two front seats and two back seats?

Many natural language systems have attempted to emphasize precision within the confines of specific well-formed rules. For example, sentiment analysis will often look for a set of very specific words and their synonyms within a social media site. These systems will then, without further assessment of the context in which those words are being use, tally the number of times those words are co-located with some brand in the same phrase. It will take the phrase, "… stopped by Dunkin Donuts for a coffee this morning, it was great …" and assert the co-location of the brand and the term 'great' is an indication of positive sentiment. However, if the rest of the phrase was, "…, it was great to hear that Starbucks is opening soon so I'm not tempted to eat donuts every morning" then the system might miss that the sentiment is really not about Dunkin Donuts. We call this *shallow Natural Language Processing (NLP)* because while it may be fairly precise within its more narrow focus, it is not very accurate.

However, it's also important to realize that shallow NLP actually has an important role in many systems. If your intent is to create a statistically relevant assessment of sentiment trends, over huge quantities of information, the lack of accuracy for each individual example is likely not an issue. Assuming there are approximately as many false-positives as false-negatives over a sufficiently large sample set, they will cancel each other out. And as long as the pool of cancelled tallies remains relatively constant across different sample sets over time, then the remaining un-cancelled data will yield statistically relevant trending information. Thus, the additional processing costs required for the additional accuracy for any given instance is unwarranted.

> Shallow Natural Language Processing (NLP) may be fairly precise within its more narrow focus but is not very accurate.

On the other hand, when the individual instances do matter, then systems that are designed to be precise without focusing on high levels of accuracy tend to be brittle – that is, they perform well within the narrow parameters of their intended design, but don't perform well when those parameters change. We liken these systems to using brick-laying construction techniques. Bricks are strong, and fairly easy to construct with. For decades and centuries we refined the brick-laying construction technique to be fairly precise. We were able to build relatively large, ornate, and long-lasting structures. However, while brick buildings have great load strength, they have poor tensile strength. They fall down easily in earthquakes and don't support large spans. And after a certain point their load strength will fail too.

You can observe these same limitations in some of our consumer products today. Pull out your favorite voice-activated personal assistant and say, "Find me pizza." What you will get back is a local listing of Pizza restaurants – exactly what you wanted. Now say, "Don't find me pizza." You will still get back a local listing of Pizza restaurants. Not exactly what we asked for. Likewise, say "Find me Pizza nearby" or "Find me Pizza faraway", and once again you get exactly the same local listings returned. The point is, these systems are designed to a specific set of rules – looking for specific keyword combinations to decide what kind of answer to produce. They don't know how to distinguish between things for which there is no rule. They are precise, but not very accurate.

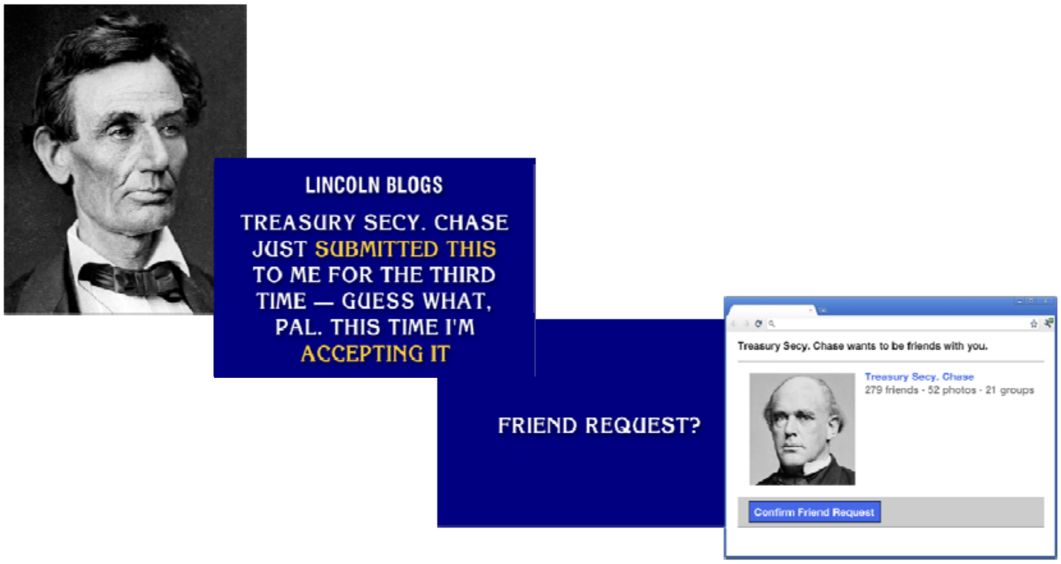
To overcome the limitations of brick building we have shifted to using steel and reinforced concrete for our largest buildings. And likewise we are seeing a shift in construction techniques for natural language processing when accuracy is needed over narrow precision. These techniques incorporate much more context into the evaluation of the question. We refer to this as *Deep NLP*, or sometimes Deep Question-Answering (DeepQA) when the problem is about answering natural language questions.

> We are seeing a shift in construction techniques for natural language processing when accuracy is needed

IBM Watson is a Deep NLP system. It achieves accuracy by attempting to assess as much context as possible. It gets that context both within the passage of the question as well as from the knowledge base (called a *corpus*) available to it for finding answers.

While preparing for the Jeopardy! game Watson was asked the question (clue), “Treasury Secy. Chase just submitted this to me for the third time – guess what pal, this time I’m accepting it” from the category “Lincoln Blogs”. First, notice the abbreviation, “Secy.”, which had to be taken to mean “Secretary”. Further notice that Secretary is not meant here to be someone that takes dictation and manages the appointment book. The combined terms “Treasury Secretary” is significant here as a noun and role. So, to answer this question Watson had to find a passage that involved submitting and accepting something between Treasury Secretary Chase and Lincoln (given the category of the clue, although also notice the category does not say “President Lincoln” necessarily). And the correct answer turned out to be “What is a resignation?”

When describing this example at an Elementary school sometime after the broadcast of Watson playing Jeopardy!, one 5th grader offered “What is a Friend Request?” as a possible answer.



TREASURY SECY. CHASE JUST SUBMITTED THIS TO ME FOR THE THIRD TIME — GUESS WHAT, PAL. THIS TIME I'M ACCEPTING IT

Treasury Secy. Chase wants to be friends with you.

This is interesting in part because it says a lot about the degree to which social media has permeated deeply into the fabric of our next generation society. However, it is also instructional because it could also be taken as fairly reasonable answer to the clue. But, we know this response is inaccurate because we have historical context – we know that Facebook was not available in the late nineteenth century. Notice, it was context that enabled us to increase the accuracy of the system in producing this answer. Without that context we would be lost.

It's worth emphasizing the point that we, as humans, have very little difficulty processing our language. That's not to say that we don't get mixed up and confused some times. But generally we do much better at resolving the meaning of things we've written than computers typically do. We have an innate quality about how we disambiguate language that we want to capture and harness in computing systems. That has been a key goal of the Artificial Intelligence community for the past four decades. And to a large extent we've been able to increase the *precision* of language processing. But, it was only with IBM Watson that we have finally been able to break through the level of accuracy that is needed for information systems to function well in the real world of broad natural language.

Without context, we would be lost

And there is a huge driving force to solve this problem. We are experiencing an explosion of data production. Ninety percent of all the data in the world was produced in the last 2 years. This trend is expected to grow as we interconnect and instrument more and more of our world. And 80% of all the information in the world is text – literature, reports, articles, research papers, theses, e-mails, blogs, tweets, forums, chat and text messages, and so forth. We need computers that can understand this flood of information to get more out of it.

IBM Watson Understands Language

To effectively navigate through the current flood of unstructured information requires a new era of computing we call Cognitive Systems. Watson is an example of a Cognitive System. It is able to tease apart the human language to identify inferences between text passages with human-like high accuracy, and at speeds and scale far faster and far bigger than any person could do on their own. It manages a very high level of accuracy when it comes to understanding the correct answer to a question.

However, Watson doesn't really understand the individual words in the language. What it does understand are the features of language as used by people and from that is able to determine whether one text passage (call it the 'question') infers another text passage (call it the 'answer'), with incredible accuracy under changing circumstances. In the Jeopardy! quiz show, Watson had to determine whether the question, "Jody Foster took this home for her role in 'Little Man'" inferred the answer "Jody Foster won an Emmy for her role in 'Little Man'". In this case, taking something home inferred winning an Emmy. But it doesn't always. Sometimes taking something home infers a cold, or groceries, or any number of things. Context matters. Temporal and spatial constraints matter. All of that adds to enabling a Cognitive System to behave with human-like characteristics. And, to go back to an earlier point, a rules-based approach would have to have a near infinite number of rules to capture every case we might encounter in language.

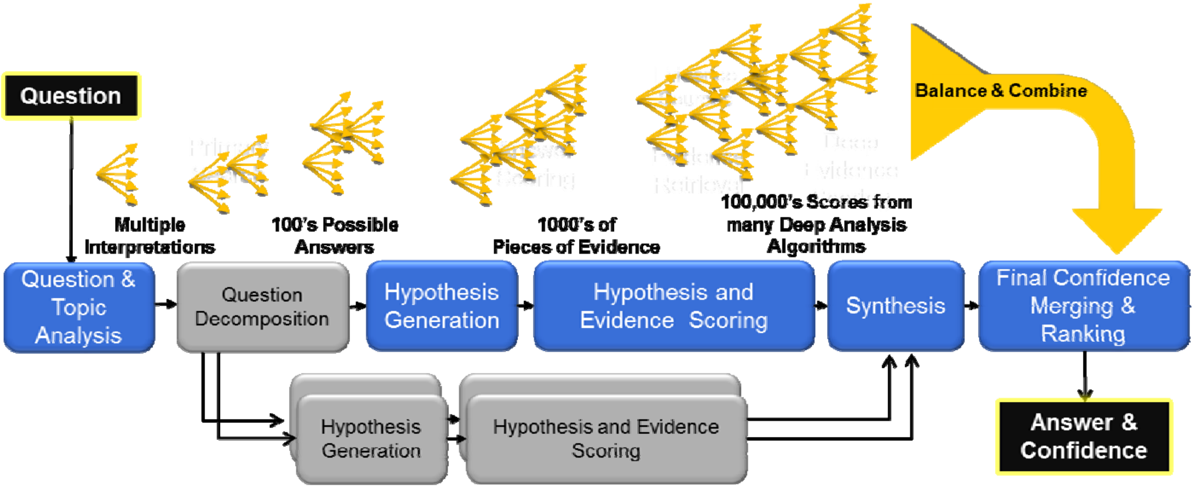
> To effectively navigate through the current flood of unstructured information requires a new era of computing we call Cognitive Systems

Watson does this by tearing apart the question and potential answers in the corpus, and examining it and the context of the statement in literally hundreds of different ways and uses that to gain a degree of confidence in its interpretation of the question and potential answers.

But let’s back up a bit. How does Watson derive an answer to a question? Let’s walk through the process.

When a question is first presented to Watson, it parses the question to pull out the major features of the question. It then generates a set of hypotheses by looking across the corpus for passages that have some potential for containing the correct answer. It then performs a deep comparison of the language of the question and the language of each potential answer using a variety of reasoning algorithms.

This is the hard part. There are hundreds of reasoning algorithms – each of which does a different kind of comparison. Some look at the matching of terms and synonyms. Some look at the temporal and spatial features. Some look at relevant sources of


Question
Multiple Interpretations
100’s Possible Answers
1000’s of Pieces of Evidence
100,000’s Scores from many Deep Analysis Algorithms
Balance & Combine
Question & Topic Analysis
Question Decomposition
Hypothesis Generation
Hypothesis and Evidence Scoring
Synthesis
Final Confidence Merging & Ranking
Hypothesis Generation
Hypothesis and Evidence Scoring
Answer & Confidence

contextual information. And so forth.

Each of these reasoning algorithms will produce one or more scores indicating the extent to which the potential answer is inferred by the question based on that algorithm's specific area of focus. Each of the resulting scores is then weighted against a statistical model that captures how well that algorithm has done at establishing the inferences between two similar passages for that domain during Watson's "training period." That statistical model can then be used to summarize a level of confidence Watson has about the evidence that the candidate answer is inferred by the question.

And it does this for each of the candidate answers until it is able to find one that surfaces as being a stronger candidate than the others.

Of paramount importance to the operation of Watson is a knowledge corpus. This corpus consists of all kinds of unstructured knowledge. Example of knowledge that gets ingested into the corpus are text books, guidelines, how-to manuals, FAQs, benefit plans, and news feeds to mention only a few. Prior to any of this, Watson will have *ingested* the corpus – going through the entire body of content to get it into a form that is easier to work with. The ingestion process will also *curate* the content – that is, making sure the corpus contains appropriate content; sifting out the articles or pages that are out of date or irrelevant, or that come from potentially unreliable sources.

Example of knowledge include text books, guidelines, how-to manuals, FAQs, benefit plans, and news feeds

As we indicated, some of the reasoning algorithms focus on spatial and temporal features of the passage. This turns out to be critical to disambiguating a tremendous amount of what humans say and write. When we say, "Find me pizza" it is just taken for granted that we mean something nearby. But what is nearby is always a relative thing. In other cases, spatial relationships show up relative to geographic markers – neighborhood in a city, or a state in a country. Likewise, temporal features are also present in the context of much of what we write. When we say, “Get cheese from the store on your way home” there is an inferred time frame. Presumably the writer and the recipient have a shared, contextual understanding of when they will be on their way home.

The statement, “In May 1898 Portugal celebrated the 400th anniversary of this explorer’s arrival in India” demonstrates both spatial and temporal dimensions. The celebration occurred in Portugal, but the thing they were celebrating was the explorer’s arrival in India. Does that suggest the explorer went from Portugal to India? Was he ever in Portugal? Notice that the celebration occurred in 1898, but the event occurred 400 years earlier. So the event actually occurred in 1498. The passage that provided the answer to the question said, “On the 27th of May 1498, Vasco da Gama landed in Kappad Beach.” The spatial and temporal evaluation had to be performed on both the question ***and*** the candidate answer passages.

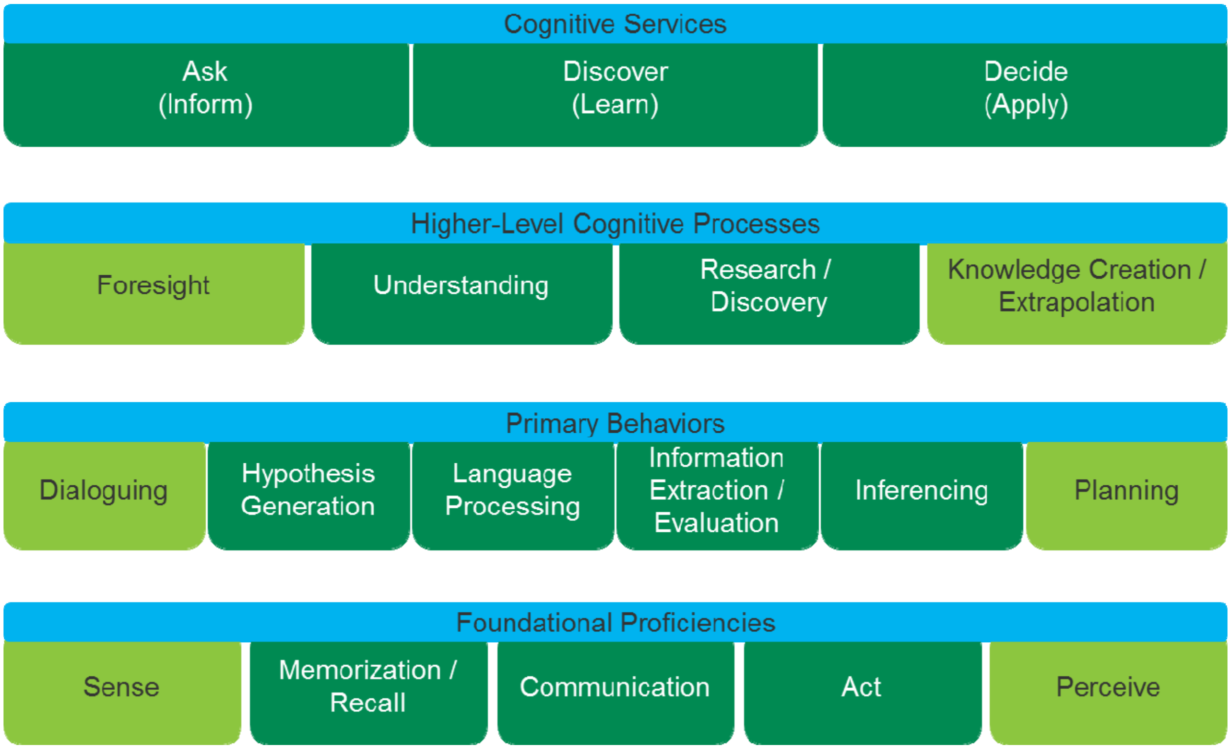
Context is derived from both immediate information as well as knowledge available more broadly. Watson can derive immediate information from the title of the document, other passages in the document, or the source database from where it came. Context can also come more broadly from a shared history. Remember that we knew that “What is a Friend Request?” was probably an incorrect answer to the clue in the Lincoln Blogs. That is because we share a common historical context – one that tells us about when certain things happened relative to each other. We know that Facebook was created fairly recently. Whereas we know that Abraham Lincoln lived some 150 years ago – well before Facebook became popular. Context and reasoning help us create a cognitive basis for processing language.

Spatial and temporal evaluation had to be performed on both the question and the candidate answer

Understanding Language is Just the Beginning

We define Cognitive Systems as applying human-like characteristics to conveying and manipulating ideas, that when combined with the inherent strengths of digital computing can solve problems with higher accuracy, more resilience, and on a massive scale over very large bodies of information.

We can decompose a Cognitive System as having several key elements. Like humans, Cognitive Systems have a way of gathering, memorizing and recalling information – the equivalent of Human memories. They also have a basic ability to communicate and act. These are organized by certain behavioral constructs – such as the ability to create



Cognitive Services

Ask (Inform) | Discover (Learn) | Decide (Apply)

Higher-Level Cognitive Processes

Foresight | Understanding | Research / Discovery | Knowledge Creation / Extrapolation

Primary Behaviors

Dialoguing | Hypothesis Generation | Language Processing | Information Extraction / Evaluation | Inferencing | Planning

Foundational Proficiencies

Sense | Memorization / Recall | Communication | Act | Perceive and test hypotheses; the ability to tease apart and create inferences about language; and the ability to extract and evaluate useful information (such as dates, places, and values). These are foundational skills – without which neither computers nor humans can determine the correct correlation between questions and answers.

Higher order cognitive processes can leverage those fundamental behaviors to achieve a level of understanding. To understand something requires that we be able to break it apart down to finer and finer elements that behave in well-ordered ways within a given scale. Just as how things work in physics on human scales is not how things work at cosmic or sub-atomic scales, likewise cognitive systems are designed to work at human scales – albeit over extremely large collections of humans. As such, understanding language starts with understanding the finer rules of language. Not just formal grammar, but the informal grammatical conventions of everyday use.

However, just as we do as humans, cognitive systems are driven to understand things by decomposing expressions of an idea, and then combining that with context and the probability that certain terms in the passage are being used in a certain way. And, as it is with humans, our confidence in that being the case is proportional to the evidence that we can locate that supports those probabilities and the number of reasoning algorithms we have available to test our hypotheses.

Just as we do as humans, cognitive systems are driven to understand things by decomposing expressions of an idea, and then combining that with context

Once we’ve established a certain level of understanding, decomposing the problem against its probable intent, cognitive systems can recompose the elements in various ways, each of which can be tested to *imagine* new concepts. These combinatorics can then be used to drive new discovery and insight – helping us to not only find answers to questions, but to also help us realize the questions that we never thought of to ask.

We can then exploit these capabilities to solve problems that fit certain common patterns – cognitive services. We can *Ask* questions that yield answers. We can use the system to *Discover* new insights – realize things we hadn't recognized previously. And we can use these systems to make sound *Decisions* – or at least assist people in the decisions they need to make.

In the future, as Cognitive Systems grow richer we expect them to gain the ability to sense – not just read text, but to see, hear, feel; to have a basic awareness of their environment. And we expect these systems to be able to perceive things – to recognize shapes and changing conditions that will further inform their context and ability to infer and reason. And we expect them to adopt higher order behaviors and cognitive processes, such as to carrying on a dialog, to plan different strategies for solving problems, and to gain foresight and extrapolate that into new knowledge.

In essence, Cognitive Systems will internalize many of the behaviors that humans find "natural", and apply them in massive scale to help people solve the problems that today often fall outside their grasp. We are beginning a new Era – an era where computers go beyond just performing routine procedural tasks more efficiently, to employing human-like cognition to make people more intelligent about what they do.

As Cognitive Systems grow richer, we expect them to gain the ability to sense

Problems Come in Different Shapes

As we move forward with Watson, we are discovering other uses for Watson. The classic "Ask" Watson – that is, where a user asks a question (or provides a clue, or a patient record, etc.), and from which Watson derives an answer, it's confidence that the question infers that answer, and the evidence that supports that answer – has found a home in Oncology Diagnosis, Utilization Management (that is, pre-approval of insurance coverage for scheduled medical procedures), Credit Analysis, and basic Research – wherever a professional needs assistance in getting the most relevant information to their problem space. One client recently remarked that one of the greatest revelations for them was that by using Watson to answer questions they realized that they were fundamentally asking the wrong questions – that while Watson was answering the questions they always ask 'correctly', there are other, better, and more important questions they needed to be asking that allowed them to think about their business problem in a whole new way; in ways that helped them understand the competitive threats and opportunities in their market place that had never occurred to them before.

Recently we have found utility in Watson with *Discover* and *Decision* applications. When Watson is transparent about its inference processes – allowing the user to assess and provide feedback to Watson's findings – we have found that Watson can be an effective teaching tool for practitioners. The practitioner can work alongside Watson to reveal new insights that were not previously understood. Together we can find new treatments that improve healthcare; we can resolve better investment options and value; we can discover new customer preferences that strengthen relationships with your business.

> There are other more important questions that allowed them to think about their business problem in a whole new way

These 'Discovery' type applications are being further improved with work we're doing right now in IBM Research and Software Development labs. Recent breakthroughs in inference-chaining – that is, determining that this infers that, which infers something else, and so on – is creating even deeper insight. Knowing that diabetes causes high blood sugar is important. However, taking the next step to infer that high blood sugar causes blindness is even more critical to caring for the "whole patient." These types of multi-level inferences can be captured as an inference graph from which we can observe a broad spectrum of downstream considerations. More importantly, convergence in the graph is a powerful way of deriving even more significant inferences – answers that can reveal even deeper insights and hidden consequences. By coalescing preceding confidence values, we can aggregate and establish even higher confidence in an answer as being the preferred answer to the question.

And further, we can produce reverse-inferences – in effect, discovering the questions to answers that were never asked. Determining that a patient that has a history of resting tremors and an 'unexpressive face' might infer they have Parkinson's disease. However, determining that the patient also has difficulty walking might further reveal damage to the Substantia Nigra nervous system. This might have been missed without prompting for the previously un-asked questions.

We are investing now in these kinds of major improvements to Watson that we believe are going to lead to further breakthroughs in healthcare, finance, contact centers, government, chemical industries and a Smarter Planet. It is these types of advances that are going propel us into an era of *Cognitive Systems*.

> We can produce reverse-inferences – in effect, discovering the questions to answers that were never asked

In many solutions, we are tying Watson into other more traditional forms of computing, such as statistical analytics, rules and business processing, collaboration, reporting, etc. to solve business problems. For example, if we can combine Watson's ability to answer questions about potential events that could signal a risk for an investment with other statistical analysis then we can improve risk and valuation processes for financial institutions. Likewise, insight that we gain about customer responses through deep natural language processing can suggest changes to buying and usage behavior that may not be otherwise evident in the structured data. In the Healthcare industry, Watson is being used to assist insurance companies on whether to pre-approve treatments as part of their Utilization Management processes.

Accuracy is Improved Through Generalization

As we continue to evolve and develop these types of Cognitive Systems we have to exercise some care. We are at a classical juncture – one that we humans face all the time. That is, whether to specialize or generalize. We can specialize NLP technology to one very specific domain – concentrating, for example, on just the linguistic features of that domain. Doing so is very tempting – maybe even necessary in the early phases of evolution to ensure the viability of the technology – survival of the species, if you will. However, doing so is likely to take us back to the era of building with bricks. If the thing that makes Cognitive Systems special is its ability to adapt with human-like dexterity, then that instructs us that we need to generalize. We need to be able to recognize and draw inferences from a broader set of linguistic variation, under a broader set of circumstances – as our knowledge changes, as the context changes, and as contemporary linguistics change.

Doing so will allow us to more readily adapt to new and larger problems. We are already applying Watson to both the Health Care and the Financial Services industries. This has two benefits: it lets us bring the advantages of Watson to different domains with high value problems, but it also allows us to evolve the language processing algorithms of Watson to handle a broader set of linguistic variation. This both enables easier adaptation to other domains, but also improves the utility of Watson to our existing domain applications.

> We are at a classic juncture – whether to specialize or generalize

The Health Care applications are interesting because they often need both precision *and* accuracy. That is, accuracy is needed to properly interpret the text in the patient's health description to infer the patient's disease. On the other hand, the National

Comprehensive Cancer Network (NCCN) guideline for breast cancer must be justified by the presence of very precise terms in the patient's health record. And then further accuracy is required to find evidence that supports that treatment.

Whenever we encounter a linguistic anomaly (that is, something in the language that we've never encountered before) we make a decision about whether the problem is truly unique to the domain or something that is characteristic of a broader set of linguistic problems. Wherever possible we go back to our core algorithms to determine whether we can generalize the algorithm to recognize and evaluate the new situation. Just as with humans, this approach enables us to map our understanding on to new experiences, and thus grow the contextual base for the system.

The result is that we increase accuracy, scope and scale – accuracy of linguistic inferencing (that is, getting the right answer, for the right reason in a timely manner); scope of the problem space; and the scaling to massive amounts of data and questions, in many more domains.

We expect to see even greater value in "next best action" solutions, social sentiment analysis, petroleum and chemical refinement, and many other applications. We are just at the beginning of a major new era of computing – one that is less precise, but much more accurate – an era of applying human-like behavior to large scale computing problems. This is the era of Cognitive Systems.

> We are just at the beginning of a major new era of computing – one that is less precise, but much more accurate

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
September 2012

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

Please Recycle