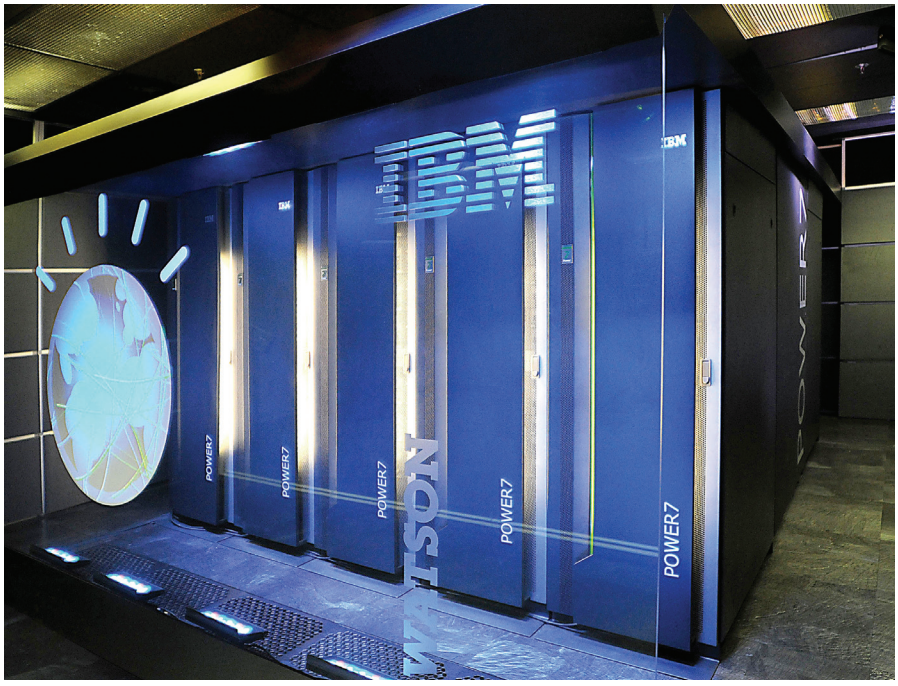
Science | DOI:10.1145/1965724.1965730 Kirk L. Kroeker

Weighing Watson's Impact

Does IBM's Watson represent a distinct breakthrough in machine learning and natural language processing or is the 2,880-core wunderkind merely a solid feat of engineering?

In the history of speculative fiction, from the golden age of science fiction to the present, there are many examples of artificial intelligences engaging their interlocutors in dialogue that exhibits self-awareness, personality, and even empathy. Several fields in computer science, including machine learning and natural language processing, have been steadily approaching the point at which real-world systems will be able to approximate this kind of interaction. IBM's Watson computer, the latest example in a long series of efforts in this area, made a television appearance earlier this year in a widely promoted human-versus-machine "Jeopardy!" game show contest. To many observers, Watson's appearance on "Jeopardy!" marked a milestone on the path toward achieving the kind of sophisticated, knowledge-based interaction that has traditionally been relegated to the realm of fiction.

The "Jeopardy!" event, in which Watson competed against Ken Jennings and Brad Rutter, the two most successful contestants in the game show's history, created a wave of coverage across mainstream and social media. During the three-day contest in February, hints of what might be called Watson's quirky personality shone through, with the machine wagering oddly precise amounts, guessing at answers after wildly misinterpreting clues, but ultimately prevailing against its formidable human opponents.



IBM's Watson soundly defeated the two most successful contestants in the history of the game show "Jeopardy!," Ken Jennings and Brad Rutter, in a three-day competition in February.

PHOTO COURTESY IBM

Leading up to the million-dollar challenge, Watson played more than 50 practice matches against former "Jeopardy!" contestants, and was required to pass the same tests that humans must take to qualify for the show and compete against Jennings, who broke the "Jeopardy!" record for the most consecutive games played, resulting in winnings of more than $2.5 mil-

lion, and Rutter, whose total winnings amounted to $3.25 million, the most money ever won by a single "Jeopardy!" player. At the end of the three-day event, Watson finished with $77,147, beating Jennings, who had $24,000, and Rutter, who had $21,600. The million-dollar prize money awarded to Watson went to charity.

Named after IBM founder Thomas J. Watson, the Watson system was built by a team of IBM scientists whose goal was to create a standalone platform that could rival a human's ability to answer questions posed in natural language. During the "Jeopardy!" challenge, Watson was not connected to the Internet or any external data sources. Instead, Watson operated as an independent system contained in several large floor units housing 90 IBM Power 750 servers with a total of 2,880 processing cores and 15 terabytes of memory. Watson's technology, developed by IBM and several contributing universities, was guided by principles described in the Open Advancement of Question-Answering (OAQA) framework, which is still operating today and facilitating ongoing input from outside institutions.

Judging by the sizeable coverage of the event, Watson piqued the interest of technology enthusiasts and the general public alike, earning "Jeopardy!" the highest viewer numbers it had achieved in several years and leading to analysts and other industry observers speculating about whether Watson represents a fundamental new idea in computer science or merely a solid feat of engineering. Richard Doherty, the research director at Envisioneering Group, a technology consulting firm based in Seaford, NY, was quoted in an Associated Press story as saying that Watson is "the most significant breakthrough of this century."

Doherty was not alone in making such claims, although the researchers on the IBM team responsible for designing Watson have been far more modest in their assessment of the technology they created. "Watson is a novel approach and a powerful architecture," says David Ferrucci, director of the IBM DeepQA research team that created Watson. Ferrucci does characterize Watson as a breakthrough in artificial intelligence, but he is careful to qualify this assertion by saying that the breakthrough is in the development of artificial-intelligence systems.

"The breakthrough is how we pulled everything together, how we integrated natural language processing, information retrieval, knowledge representation, machine learning, and a general reasoning paradigm," says Ferrucci. "I think this represents a breakthrough. We would have failed had we not invested in a rigorous scientific method and systems engineering. Both were needed to succeed."

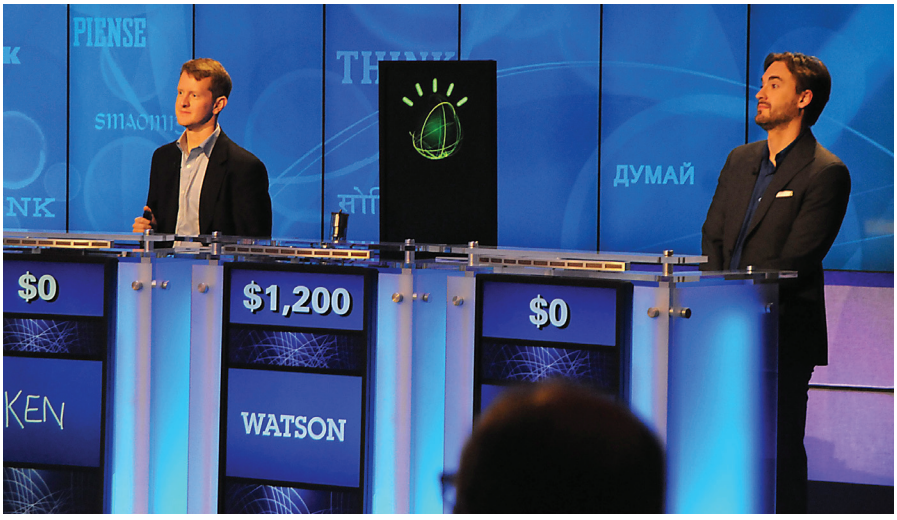
Contextual Evidence

The DeepQA team was inspired by several overarching design principles, with the core idea being that no single algorithm or formula would accurately understand or answer all questions, says Ferrucci. Rather, the idea was to build Watson's intelligence from a broad collection of algorithms that would probabilistically and imperfectly interpret language and score evidence from different perspectives. Watson's candidate answers, those answers in which Watson has the most confidence, are produced from hundreds of parallel hypotheses collected and scored from contextual evidence.

Ferrucci says this approach required innovation at the systems level so individual algorithms could be developed independently, then evaluated for their contribution to the system's overall performance. The approach allowed for loosely coupled interaction between algorithm components, which Ferrucci says ultimately reduced the need for team-wide agreement. "If every algorithm developer had to agree with every other or reach some sort of consensus, progress would have been slowed," he says. "The key was to let different members of the team develop diverse algorithms independently, but regularly perform rigorous integration testing to evaluate relative impact in the context of the whole system."

Ferrucci and the DeepQA team are expected to release more details later this year in a series of papers that will outline how they dealt with specific aspects of the Watson design. For now, only bits and pieces of the complete picture are being disclosed. Ferrucci says that, looking ahead, his team's research agenda is to focus on how Watson can understand, learn, and interact more effectively. "Natural language understanding remains a tremendously difficult challenge, and while Watson demonstrated a powerful approach, we have only scratched the surface," he says. "The challenge continues to be about how you build systems to accurately connect language to some representation, so the system can automatically learn from text and then reason to discover evidence and answers."

Lillian Lee, a professor in the computer science department at Cornell University, says the reactions about Watson's victory echo the reactions following Deep Blue's 1997 victory over chess champion Garry Kasparov, but with several important differences. Lee, whose research focus is natural



Watson's on-stage persona simulates the system's processing activity and relative answer confidence through moving lines and colors. Watson is shown here in a practice match with Ken Jennings, left, and Brad Rutter at IBM's Watson Research Center in January.

PHOTO COURTESY IBM

language processing, points out that some observers were dismissive about Deep Blue's victory, suggesting that the system's capability was due largely to brute-force reasoning rather than machine learning. The same criticism, she says, cannot be leveled at Watson because the overall system needed to determine how to assess and integrate diverse responses.

"Watson incorporates machine learning in several crucial stages of its processing pipeline," Lee says. "For example, reinforcement learning was used to enable Watson to engage in strategic game play, and the key problem of determining how confident to be in an answer was approached using machine-learning techniques, too."

Lee says that while there has been substantial research on the particular problems the "Jeopardy!" challenge involved for Watson, that prior work should not diminish the team's accomplishment in advancing the state of the art to Watson's championship performance. "The contest really showcased real-time, broad-domain question-answering, and provided as comparison points two extremely formidable contestants," she says. "Watson represents an absolutely extraordinary achievement."

Lee suggests that with language-processing technologies now maturing, with the most recent example of such maturation being Watson, the field appears to have passed through an important early stage. It now faces an unprecedented opportunity in helping sift through the massive amounts of user-generated content online, such as opinion-oriented information in product reviews or political analysis, according to Lee.

While natural-language processing is already used, with varying degrees of success, in search engines and other applications, it might be some time before Watson's unique question-answering capabilities will help sift through online reviews and other user-generated content. Even so, that day might not be too far off, as IBM has already begun work with Nuance Communications to commercialize the technology for medical applications. The idea is for Watson to assist physicians and nurses in finding information buried in medical tomes, prior cases, and the latest science journals. The first commercial offerings from the collaboration are expected to be available within two years.

"Natural language understanding remains a tremendously difficult challenge, and while Watson demonstrated a powerful approach, we have only scratched the surface," says David Ferrucci.

Beyond medicine, likely application areas for Watson's technology would be in law, education, or the financial industry. Of course, as with any technology, glitches and inconsistencies will have to be worked out for each new domain. Glitches notwithstanding, technology analysts say that Watson-like technologies will have a significant impact on computing in particular and human life in general. Ferrucci, for his part, says these new technologies likely will mean a demand for higher-density hardware and for tools to help developers understand and debug machine-learning systems more effectively. Ferrucci also says it's likely that user expectations will be raised, leading to systems that do a better job at interacting in natural language and sifting through unstructured content.

To this end, explains Ferrucci, the DeepQA team is moving away from attempting to squeeze ever-diminishing performance improvements out of Watson in terms of parsers and local components. Instead, they are focusing on how to use context and information to evaluate competing interpretations more effectively. "What we learned is that, for this approach to extend beyond one domain, you need to implement a positive feedback loop of extracting basic syntax and local semantics from language, learning from context, and then interacting with users and a broader community to acquire knowledge that is otherwise difficult to extract," he says. "The system must be able to bootstrap and learn from its own failing with the help of this loop."

In an ideal future, says Ferrucci, Watson will operate much like the ship computer on "Star Trek," where the input can be expressed in human terms and the output is accurate and understandable. Of course, the "Star Trek" ship computer was largely humorless and devoid of personality, responding to queries and commands with a consistently even tone. If the "Jeopardy!" challenge serves as a small glimpse of things to come for Watson—in particular, Watson's precise wagers, which produced laughter in the audience, and Watson's visualization component, which appeared to express the state of a contemplative mind through moving lines and colors—the DeepQA team's focus on active learning might also include a personality loop so Watson can accommodate subtle emotional cues and engage in dialogue with the kind of good humor reminiscent of the most personable artificial intelligences in fiction. C

Further Reading

Baker, S.
Final Jeopardy: Man vs. Machine and the Quest to Know Everything. **Houghton Mifflin Harcourt, New York, NY, 2011.**

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., Schlaefer, N., and Welty, C.
Building Watson: An overview of the DeepQA project, *AI Magazine 59*, Fall 2010.

Ferrucci, D., et al.
Towards the Open Advancement of Question Answering Systems. *IBM Research Report RC24789 (W0904-093)*, April 2009.

Simmons, R.F.
Natural language question-answering systems, *Communications of the ACM 13*, 1, Jan. 1970.

Strzalkowski, T., and Harabagiu, S. (Eds.)
Advances in Open Domain Question Answering. **Springer-Verlag, Secaucus, NJ, 2006.**

Kirk L. Kroeker works in communications and has written extensively about the impact of emerging technologies.