CHAPTER 6

Watson and the *Jeopardy!* Challenge

How does Watson—IBM's Jeopardy!*-playing computer—work? Why does it need predictive modeling in order to answer questions, and what secret sauce empowers its high performance? How does the iPhone's Siri compare? Why is human language such a challenge for computers? Is artificial intelligence possible?*

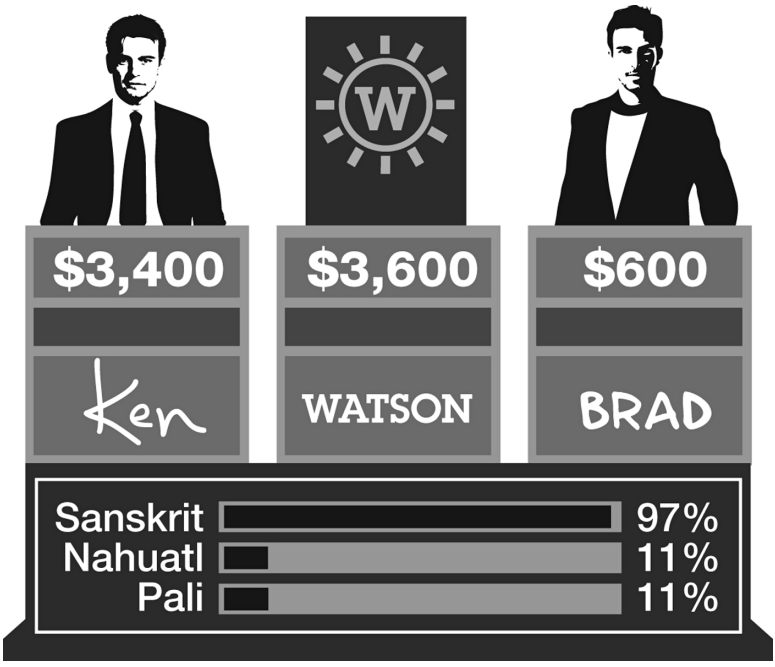
January 14, 2011. The big day had come. David Gondek struggled to sit still, battling the butterflies of performance anxiety, even though he was not the one onstage. Instead, the spotlights shone down upon a machine he had helped build at IBM Research for the past four years. Before his eyes, it was launched into a battle of intellect, competing against humans in this country's most popular televised celebration of human knowledge and cultural literacy, the quiz show *Jeopardy!*

Celebrity host Alex Trebek read off a clue, under the category "Dialing for Dialects:"

VEDIC, DATING BACK AT LEAST 4,000 YEARS, IS THE EARLIEST DIALECT OF THIS CLASSICAL LANGUAGE OF INDIA *

* *Jeopardy!* questions stamped with an asterisk were posed during Watson's televised match.

Watson,[1] the electronic progeny of David and his colleagues, was competing against the two all-time champions across the game show's entire 26-year televised history. These two formidable opponents were of a different ilk, holding certain advantages over the machine, but also certain disadvantages. They were human.



Watson competes against two humans on *Jeopardy!*

[1] In this chapter, *Watson* refers to the highly specialized IBM computer that competed on *Jeopardy!* in 2011. Although the name *Watson* referred only to that specific system at that time, IBM has subsequently broadened its use of the word in its corporate branding strategy. *Watson* now also refers to at least three other loosely related initiatives: 1) IBM's promising research efforts applying some of the same analytical approaches developed for *Jeopardy!* within healthcare and other application areas; 2) *Watson Analytics*, a cloud-based business tool for predictive analytics and data visualization available in wide release; and 3) the technology one IBM partner credits for the function of its product, the CogniToys Dino, a toy dinosaur designed to conduct educational dialogues with children.

Watson buzzed in ahead of its opponents. Deaf and unable to hear Trebek's professional, confident voice, it had received the *Jeopardy!* clue as a transmission of typed text. The audience heard Watson's synthesized voice respond, phrasing it according to the show's stylistic convention of posing each answer in the form of a question. "What is Sanskrit?"[2]

For a computer, questions like this might as well be written in Sanskrit. Human languages like English are far more complex than the casual speaker realizes, with extremely subtle nuance and a pervasive vagueness we nonmachines seem completely comfortable with. Programming a computer to work adeptly with human language is often considered the ultimate challenge of artificial intelligence (AI).

TEXT ANALYTICS

It was Greek to me.

—William Shakespeare

I'm completely operational, and all my circuits are functioning perfectly.

—HAL, the intelligent computer from *2001: A Space Odyssey* (1968)

Science fiction almost always endows AI with the capacity to understand human tongues. Hollywood glamorizes a future in which we chat freely with the computer like a well-informed friend. In *Star Trek IV: The Voyage Home* (1986), our heroes travel back in time to a contemporary Earth and are confounded by its primitive technology. Our brilliant space engineer Scotty, attempting to make use of a Macintosh computer, is so accustomed to computers understanding the spoken word that he assumes its mouse must be a microphone. Patiently picking up the mouse as if it were a quaint artifact, he jovially beckons, "Hello, computer!"

[2] In this chapter, I refer to each *Jeopardy!* clue as a *question* and each contestant response as an *answer*. It is a game of question answering, despite its stylistic convention of phrasing each contestant response in the form of a question beginning "what is" or "who is."

2001: A Space Odyssey's smart and talkative computer, HAL, bears a legendary, disputed connection in nomenclature to IBM (just take each letter back one position in the alphabet); however, author Arthur C. Clarke has strenuously denied that this was intentional. Ask IBM researchers whether their question-answering Watson system is anything like HAL, which goes famously rogue in the film, and they'll quickly reroute your comparison toward the obedient computers of *Star Trek*.

The field of research that develops technology to work with human language is *natural language processing* (NLP, aka *computational linguistics*). In commercial application, it's known as *text analytics*. These fields develop analytical methods especially designed to operate across the written word.

If data is all Earth's water, textual data is the part known as "the ocean." Often said to compose 80 percent of all data, it's everything we the human race know that we've bothered to write down. It's potent stuff—content-rich because it was generated with the intent to convey not just facts and figures, but human knowledge.

But text, data's biggest opportunity, presents the greatest challenge.

OUR MOTHER TONGUE'S TRIALS AND TRIBULATIONS

It is difficult to answer, when one does not understand the question.

—Sarek, Spock's father, in *Star Trek IV: The Voyage Home*

Let's begin with the relatively modest goal of grammatically deconstructing the Sanskrit question, repeated here:

VEDIC, DATING BACK AT LEAST 4,000 YEARS, IS THE EARLIEST DIALECT OF THIS CLASSICAL LANGUAGE OF INDIA *

For example, consider how "of India" fits in. It's a prepositional phrase that modifies "this classical language." That may seem obvious to you, human reader, but if the final two words had been "of course," that phrase would

instead modify the main verb, "is" (or the entire phrase, depending on how you look at it).

Determining how each component such as "of India" fits in relies on a real understanding of words and the things in the world that they represent. Take the classic linguistic conundrum, "Time flies like an arrow." Which is the main verb of the sentence? It is *flies* if you interpret the sentence as: "Time moves quickly, just as an arrow does." But it could be *time* if you read it as the imperative, ordering you to "Measure the speed of flies as you would measure that of an arrow."

The preferred retort to this aphorism, often attributed to Groucho Marx, is: "Fruit flies like a banana." It's funny and grammatically revealing. Suddenly *like* is now the verb, instead of a preposition.

"I had a car." If the duration of time for which this held true was one year, I would say, "I had a car *for* a year." But change one word and everything changes. "I had a baby." If the duration of labor was five hours, you would say, "I had a baby *in* five hours," not "*for* five hours." The word choice depends on whether you're describing a situation or an event, and the very meaning of the object—*car* or *baby*—makes the difference.

"I ate spaghetti with meatballs." Meatballs were part of the spaghetti dish.

"I ate spaghetti with a fork." The fork was instrumental to eating, not part of the spaghetti.

"I ate spaghetti with my friend Bill." Bill wasn't part of the spaghetti, nor was he instrumental to eating, although he was party to the eating event.

"I had a ball." Great, you had fun.

"I had a ball but I lost it." Not so much fun! But in a certain context, the same phrase goes back to being about having a blast:

Q: *"How was your vacation and where is my video camera?"*

A: *"I had a ball but I lost it."*

In language, even the most basic grammatical structure that determines which words directly connect depends on our particularly human view of

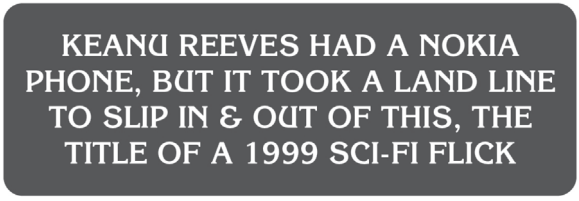
and extensive knowledge about the world. The rules are fluid, and the categorical shades of meaning are informal.[3]

ONCE YOU UNDERSTAND THE QUESTION, ANSWER IT

> *How can a slim chance and a fat chance be the same, while a wise man and wise guy are opposites?*
>
> —Anonymous

> *Why does your nose run, and your feet smell?*
>
> —George Carlin

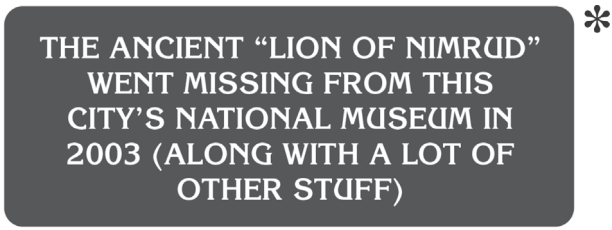
Beyond processing a question in the English language, a whole other universe of challenge lurks: *answering it*. Assume for a moment the language challenges have been miraculously met and the computer has gained the ability to "understand" a *Jeopardy!* question, to grammatically break it down, and to assess the "meaning" of its main verb and how this meaning fuses with the "meanings" of the other words such as the subject, object, and prepositional phrases to form the question's overall meaning. Consider the following question, under the category "Movie Phone:"



[3] We face yet another "Mission Impossible" trying to get the computer to write instead of read. Generating human language trips up the naïve machine. I once received a voice-synthesized call from Blockbuster (a video rental chain of its day) reminding me of my rented movie's due date. "This is a message for Eric the Fifth Siegel," it said. My middle initial is V. Translation between languages also faces hazards. An often-cited example is that "The spirit is willing, but the flesh is weak," if translated into Russian and back, could end up as "The vodka is good, but the meat is rotten."

A perfect language-understanding machine could invoke a routine to search a database of movies for one starring Keanu Reeves in which a plot element involves using a land-line telephone to "get out of" something—that something also being the title of the movie (*The Matrix*). Even if the reliable transformation of question to database lookup were possible, how could any database be sure to include coverage of these kinds of abstract movie plot elements, which are subjective and open ended?

As another example that would challenge any database, consider this *Jeopardy!* question under the category "The Art of the Steal:"



First, to succeed, the system must include the right information about each art piece, just as movie plot elements were needed for the *Matrix* question. IBM would have needed the foresight to include in a database of artworks whether, when, and where each item was stolen (for this item, the answer is Baghdad). Second, the system would also need to equate "went missing" with being stolen. That may be a reasonable interpretation regarding artwork, but if I said that my car keys went missing, we wouldn't reach the same conclusion. How endlessly involved would a mechanical incarnation of human reason need to be in order to automatically make such distinctions? Written sources such as newspaper articles did in fact use a diverse collection of words to report this art carving's *disappearance*, *looting*, *theft*, or *being stolen*.

Movies and artworks represent only the tip-top of a vast iceberg. *Jeopardy!* questions could fall within any domain, from the history of wine to philosophy to literature to biochemistry, and the answer required could be a person, place, animal, thing, year, or abstract concept. This unbounded challenge is called *open question answering*. Anything goes.

The old-school AI researcher succumbs to temptation and fantasizes about building a Complete Database of Human Knowledge. That researcher is fun

to chat with. He holds a grandiose view regarding our ability to reach for the stars by digging deep, examining our own inner cognitions, and expressing them with computer programs that mimic human reason and encode human knowledge. But someone has to break it to the poor fellow: This just isn't possible. As more pragmatic researchers concluded in the 1980s and 1990s, it's too large and too ill defined.

In reality, given these challenges, IBM concluded only 2 percent of *Jeopardy!* questions could be answered with a database lookup. The demands of open question answering reach far beyond the computer's traditional arena of storing and accessing data for flight reservations and bank records. We're going to need a smarter robot.

THE ULTIMATE KNOWLEDGE SOURCE

> *We are not scanning all those books to be read by people. We are scanning them to be read by an AI.*
>
> —A Google employee regarding Google's book scanning, as quoted by George Dyson in *Turing's Cathedral: The Origins of the Digital Universe*

A bit of good news: IBM didn't need to create comprehensive databases for the *Jeopardy!* challenge because the ultimate knowledge source already exists: *the written word*. I am pleased to report that people like to report; we write down what we know in books, Web pages, Wikipedia entries, blogs, and newspaper articles. All this *textual data* composes an unparalleled gold mine of human knowledge.

The problem is that these things are all encoded in human language, just like those confounding *Jeopardy!* questions. So the question-answering machine must overcome not only the intricacies and impossibilities of the question itself, but the same aspects of all the millions of written documents that may hold the question's answer.

Googling the question won't work. Although it's a human's primary means of seeking information from the Internet's sea of documents, Google doesn't hone down to an answer. It returns a long list of Web pages, each with hundreds or thousands of possible answers within. It is not designed for

the task at hand: identifying the singular answer to a question. Trying to use Google or other Internet search solutions to play *Jeopardy!*—for example, by doing a search on words from a question and answering with the document topic of the top search result—does not cut it. If only question answering were that easy to solve! This kind of solution answers only 30 percent of the questions correctly.

APPLE'S SIRI VERSUS WATSON

How does the iPhone personal assistant Siri compare with Watson? First introduced as the main selling point to distinguish the iPhone 4S from the preceding model, Siri responds to a broad, expanding range of voice commands and inquiries directed toward your iPhone.

Siri handles simpler language than Watson does: Users tailor requests for Siri knowing that they're speaking to a computer, whereas Watson fields *Jeopardy!*'s clever, wordy, information-packed questions that have been written with only humans in mind, without regard or consideration for the possibility that a machine might be answering. Because of this, Siri's underlying technology is designed to solve a different, simpler variant of the human language problem.

Although Siri responds to an impressively wide range of language usage, such that users can address the device in a casual manner with little or no prior instruction, people know that computers are rigid and will constrain their inquiries accordingly. Someone might request, "Set an appointment for tomorrow at 2 o'clock for coffee with Bill," but will probably not say, "Set an appointment with that guy I ate lunch with a lot last month who has a Yahoo! e-mail address," and will definitely not say, "I want to find out when my tall, handsome friend from Wyoming feels like discussing our start-up idea in the next couple weeks."

Siri flexibly handles relatively simple phrases that pertain to smartphone tasks such as placing calls, text messaging, performing Internet

(*continued*)

APPLE'S SIRI VERSUS WATSON (CONTINUED)

searches, and employing map and calendar functions (she's your *social techretary*).

Siri also fields general questions, but it does not attempt full open question answering. Invoking a system called WolframAlpha (accessible for free online), it answers simply phrased, fact-based questions via database lookup; the system can only provide answers calculated from facts that appear explicitly within its impressive, curated collection of structured, tabular database, such as:

The birthdates of famous people—How old was Elton John in 1976?

Astronomical facts—How long does it take light to go to the moon?

Geography—What is the biggest city in Texas?

Healthcare—What country has the highest average life expectancy?

One must phrase questions in a simple form, since WolframAlpha is designed first to compute answers from tables of data, and only secondarily to attempt to handle complicated grammar.

Siri processes spoken inquiries, whereas Watson processes transcribed questions. Researchers generally approach processing speech (*speech recognition*) as a separate problem from processing text. There is more room for error when a system attempts to transcribe spoken language before also interpreting it, as Siri does.

Siri includes a dictionary of humorous canned responses. If you ask Siri about its origin with, "Who's your daddy?" it will respond, "I know this must mean something . . . everybody keeps asking me this question." This should not be taken to imply adept human language processing. You might also ask, "What does the fox say?"

Siri and WolframAlpha's question answering performance is continually improved by ongoing research and development efforts, guided in part by the constant flow of incoming user queries.

ARTIFICIAL IMPOSSIBILITY

I'm wondering how to automate my wonderful self—
a wond'rous thought that presupposes my own mental health.
Maybe it's crazy to think thought's so tangible, or that I can sing.
Either way, if I succeed, my machine will attempt the very same thing.
—What artificial intelligence researchers sing in the shower

It is irresistible to pursue this because, as we pursue understanding natural language, we pursue the heart of what we think of when we think of human intelligence.
—David Ferrucci, Watson Principal Investigator, IBM Research

There's a fine line between genius and insanity.
—Oscar Levant

Were these IBM researchers certifiably nuts to take on this grand challenge, attempting to programmatically answer any *Jeopardy!* question? They were tackling the breadth of human language that stretches beyond the phrasing of each question to include a sea of textual sources, from which the answer to each question must be extracted. With this ambition, IBM had truly doubled down.

I would have thought success impossible. After witnessing the world's best researchers attempting to tackle the task through the 1990s (during which I spent six years in natural language processing research, as well as a summer at the same IBM Research center that bore Watson), I was ready to throw up my hands. Language is so tough that it seemed virtually impossible even to program a computer to answer questions within a limited domain of knowledge such as movies or wines. Yet IBM had taken on the unconstrained, open field of questions across any domain.

Meeting this challenge would demonstrate such a great leap toward humanlike capabilities that it invokes the "I" word: intelligence. A computer pulling it off would appear as magical and mysterious as the human mind. Despite my own 20-odd years studying, teaching, and researching all things artificial intelligence (AI), I was a firm skeptic. But this task required a leap so great that seeing it succeed might leave me, for the first time, agreeing that the term *AI* is justified.

AI is a loaded term. It blithely presumes a machine could ever possibly qualify for this title. Only with great audacity does the machine-builder bestow the honor of "intelligence" upon her own creations. Invoking the term comes across as a bit self-aggrandizing, since the inventor would have to be pretty clever herself to pull this off.

The *A* isn't the problem—it's the *I*. Intelligence is an entirely subjective construct, so AI is not a well-defined field. Most of its various definitions boil down to "making computers intelligent," whatever that means! AI ordains no one particular capability as the objective to be pursued. In practice, AI is the pursuit of philosophical ideals and research grants.

What do God, Groucho Marx, and AI have in common? They'd never be a member of a club that would have them as a member. AI destroys itself with a logical paradox in much the same way God does in Douglas Adams's *Hitchhiker's Guide to the Galaxy*:[4]

> *"I refuse to prove that I exist," says God, "for proof denies faith, and without faith I am nothing."*
>
> *"But," says Man, "The Babel fish [which translates between the languages of interplanetary species] is a dead giveaway isn't it? It could not have evolved by chance. It proves that you exist, and so therefore, by your own arguments, you don't. QED."*
>
> *"Oh dear," says God, "I hadn't thought of that," and promptly disappears in a puff of logic.*

AI faces analogous self-destruction because, once you get a computer to do something, you've necessarily trivialized it. We conceive of as yet unmet "intelligent" objectives that appear big, impressive, and unwieldy, such as transcribing the spoken word (*speech recognition*) or defeating the world chess champion. They aren't easy to achieve, but once we do pass such benchmarks, they suddenly lose their charm. After all, computers can manage only mechanical tasks that are well understood and well specified. You might be impressed by its lightning-fast speed, but its electronic

[4] Watson's avatar, its visual depiction shown on *Jeopardy!*, consists of 42 glowing, crisscrossing threads as an inside joke and homage that references the significance this number holds in Adams's infamous *Hitchhiker's Guide*.

execution couldn't hold any transcendental or truly humanlike qualities. If it's possible, it's not intelligent. Conversely, as famed computer scientist Larry Tesler succinctly put it, "Intelligence is whatever machines haven't done yet."

Suffering from an intrinsic, overly grandiose objective, AI inadvertently equates to "getting computers to do things too difficult for computers to do"—artificial impossibility.

LEARNING TO ANSWER QUESTIONS

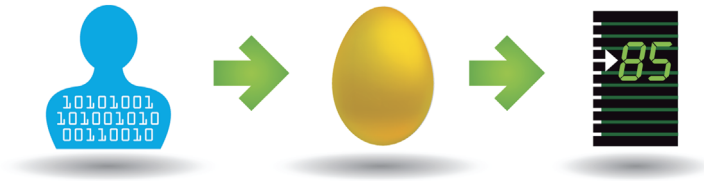
But in fact, IBM did face a specific, well-defined task: answering *Jeopardy!* questions. And if the researchers succeeded and Watson happened to appear intelligent to some, IBM would earn extra credit on this homework assignment.

As a rule, anticipating all possible variations in language is not possible. NLP researchers derive elegant, sophisticated systems to deconstruct phrases in English and other natural languages, based on deep linguistic concepts and specially designed dictionaries. But, implemented as computer programs, the methods just don't scale. It's always possible to find phrases that seem simple and common to us as humans, but trip up an NLP system. The researcher, in turn, broadens the theory and knowledge base, tweaking the system to accommodate more phrases. After years of tweaking, these hand-engineered methods still have light-years to go before we'll be chatting with our laptops just the same as with people.

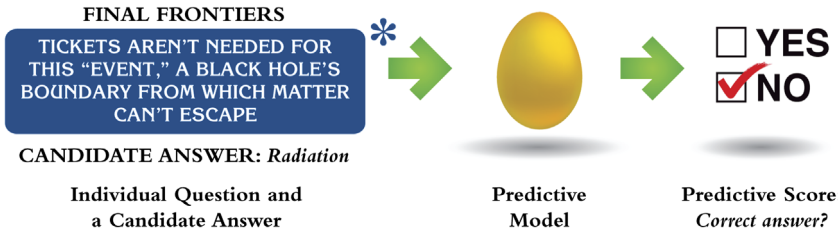
There's one remaining hope: Automate the researchers' iterative tweaking so it explodes with scale as a *learning* process. After all, that is the very topic of this book:

> ***Predictive analytics (PA)***—*Technology that learns from experience (data) to predict the future behavior of individuals in order to drive better decisions.*

Applying PA to question answering is a bit different from most of the examples we've discussed in this book. In those cases, the predictive model foretells whether a human will take a certain action, such as *click*, *buy*, *lie*, or *die*, based on things known about that individual:



IBM's Watson computer includes models that predict whether human experts would consider a *Jeopardy!* question/answer pair correct:



If the model is working well, it should give a low score, since *event horizon*, not *radiation*, is the correct answer (*Star Trek* fans will appreciate this question's category, "Final Frontiers"). Watson did prudently put a 97 percent score on *event horizon* and scored its second and third candidates, *mass* and *radiation*, at 11 percent and 10 percent, respectively. This approach frames question answering as a PA application:

PA APPLICATION: OPEN QUESTION ANSWERING

1. **What's predicted:** Given a question and one candidate answer, whether the answer is correct.
2. **What's done about it:** The candidate answer with the highest predictive score is provided by the system as its final answer.

Answering questions is not *prediction* in the conventional sense—Watson does not predict the future. Rather, its models "predict" the correctness of an answer. The same core modeling methods apply—but unlike other applications of predictive modeling, the unknown thing being "predicted" is

already known by some, rather than becoming known only when witnessed in the future. Through the remainder of this chapter, I employ this alternative use of the word *predict*, meaning, "to imperfectly infer an unknown." You could even think of Watson's predictive models as answering the predictive question: "Would human experts agree with this candidate answer to the question?" This semantic issue also arises for predicting clinical diagnosis (Central Table 4), fraud (Central Table 5), human thought (Central Table 8) and other areas—all marked with $\mathscr{D}$ (for "detect") in the Central Tables.

WALK LIKE A MAN, TALK LIKE A MAN

IBM needed data—specifically, example *Jeopardy!* questions—from which to learn. Ask and ye shall receive: Decades of televised *Jeopardy!* provide hundreds of thousands of questions, each alongside its correct answer (IBM downloaded these from fan websites, which post all the questions). This wealth of learning data delivers a huge, unprecedented boon for pushing the envelope in human language understanding. While most PA projects enjoy as data a good number of example individuals who either did or did not take the action being predicted (such as all those behaviors listed in the left columns of this book's Central Tables of PA applications), most NLP projects simply do not have many previously solved examples from which to learn.

With this abundance of *Jeopardy!* history, the computer could learn to become humanlike. The questions, along with their answer key, contribute examples of human behavior: how people answer these types of questions. Therefore, this form of data fuels machine learning to produce a model that mimics how a human would answer, "Is this the right answer to this question?"—the learning machine models the human expert's response. We may be too darn complex to program computers to mimic ourselves, but the model need not derive answers in the same manner as a person; with predictive modeling, perhaps the computer can find some innovative way to program itself for this human task, even if it's done differently than by humans.

As Alan Turing famously asked, would a computer program that exhibits humanlike behavior qualify as AI? It's anthropocentric to think so, although I've been called worse.

But having extensive *Jeopardy!* learning data did not in itself guarantee successful predictive models, for two reasons:

1. Open question answering presents tremendous unconquered challenges in the realms of language analysis and human reasoning.
2. Unlike many applications of PA, success on *Jeopardy!* requires high predictive *accuracy*; The Prediction Effect from Chapter 1—*a little prediction goes a long way*—does not apply here.

When IBM embarked upon the *Jeopardy!* challenge in 2006, the state of the art fell severely short. The most notable source of open question answering data was a government-run competition called TREC QA (Text REtrieval Conference—Question Answering). To serve as training data, the contest provided questions that were much more straightforward and simply phrased than those on *Jeopardy!*, such as, "When did James Dean die?" Competing systems would pore over news articles to find each answer. IBM had a top-five competitor that answered 33 percent of those questions correctly, and no competing system broke the 50 percent mark. Even worse, after IBM worked for about one month to extend the system to the more challenging arena of *Jeopardy!* questions, it could answer only 13 percent correctly, substantially less than the 30 percent achieved by just using Internet search.

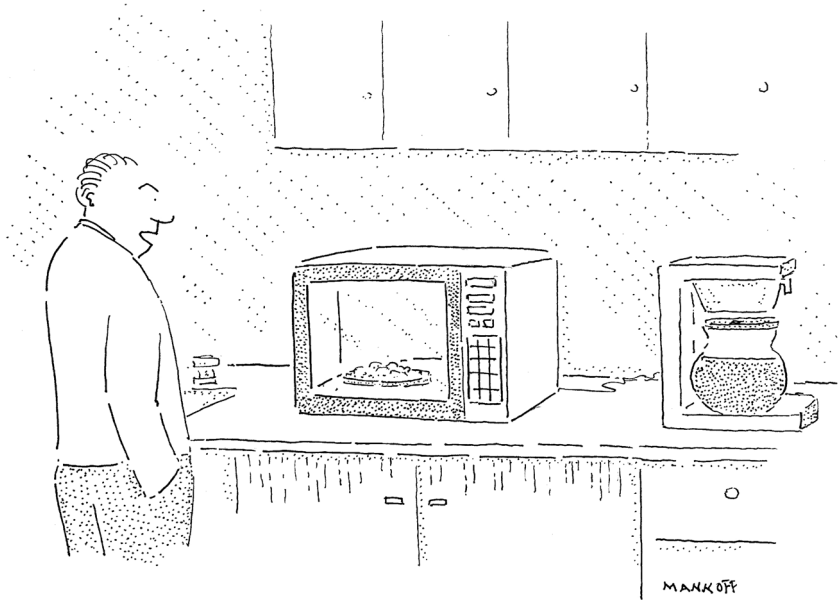
PUTTING ON THE PRESSURE

> Scientists often set their own research goals. A grand challenge takes this control out of the hands of the scientist to force them to work on a problem that is harder than one they would pick to work on themselves.
>
> —Edward Nazarko, Client Technical Advisor, IBM

Jumping on the *Jeopardy!* challenge, IBM put its name on the line. Following the 1997 chess match in which IBM's Deep Blue computer defeated then

world champion Garry Kasparov, the 2011 *Jeopardy!* broadcast pitted man against machine just as publicly, and with a renewed, healthy dose of bravado. A national audience of *Jeopardy!* viewers waited on the horizon.

As with all grand challenges, success was not a certainty. No precedent or principle had ensured it would be possible to fly across the Atlantic (Charles Lindbergh did so to win $25,000 in 1927); walk on the moon (NASA's Apollo 11 brought people there in 1969, achieving the goal John F. Kennedy set for that decade); beat a chess grandmaster with a computer (IBM's Deep Blue in 1997); or even improve Netflix's movie recommendation system by 10 percent (2009, as detailed in the previous chapter).



"No, I don't want to play chess. I just want you to reheat the lasagna."

Reproduced with permission.

In great need of a breakthrough, IBM tackled the technical challenge with the force only a megamultinational enterprise can muster. With over $92 billion in annual revenue and more than 412,000 employees worldwide, IBM is the third-largest U.S. company by number of employees. All told, its investment to develop Watson is estimated in the tens of millions of dollars,

including the dedication of a team that grew to 25 PhD's over four years at its T. J. Watson Research Center in New York (which, like the *Jeopardy!*-playing computer, was named after IBM's first president, Thomas J. Watson).

The power to push really hard does not necessarily mean you're pushing in the right direction. From where will scientific epiphany emerge? Recall the key innovation that the crowdsourcing approach to grand challenges helped bring to light, *ensemble models*, introduced in the prior chapter. It's just what the doctor ordered for IBM's *Jeopardy!* challenge.

THE ANSWERING MACHINE

David Gondek and his colleagues at IBM Research could overcome the daunting *Jeopardy!* challenge only with *synthesis*. When it came to processing human language, the state of the art was fragmented and partial—a potpourri of techniques, each innovative in conception but severely limited in application. None of them alone made the grade.

How does IBM's Watson work? It's built with ensemble models. Watson merges a massive amalgam of methodologies. It succeeds by fusing technologies. There's no secret ingredient; it's the overall recipe that does the trick. Inside Watson, ensemble models select the final answer to each question.

Before we more closely examine how Watson works, let's look at the discoveries made by a PA expert who analyzed *Jeopardy!* data in order to "program himself" to become a celebrated (human) champion of the game show.

MONEYBALLING *JEOPARDY!*

On September 21, 2010, a few months before Watson faced off on *Jeopardy!*, televisions across the land displayed host Alex Trebek speaking a clue tailored to the science fiction fan.

ZACHARY QUINTO SHOWED US THE LOGIC AS THIS CHARACTER IN 2009'S "STAR TREK"

Contestant Roger Craig avidly buzzed in. Like any technology PhD, he knew the answer was Spock.

As Spock would, Roger had taken studying to its logical extreme. *Jeopardy!* requires inordinate cultural literacy, the almost unattainable status of a Renaissance man, one who holds at least basic knowledge about pretty much every topic. To prepare for his appearance on the show, which he'd craved since age 12, Roger did for *Jeopardy!* what had never been done before. He *Moneyballed* it.

Roger optimized his study time with prediction. As a mere mortal, he faced a limited number of hours per day to study. He rigged his computer with *Jeopardy!* data. An expert in predictive modeling, he developed a system to learn from his performance practicing on *Jeopardy!* questions so that it could serve up questions he was likely to miss in order to efficiently focus his practice time on the topics where he needed it most. *He used PA to predict himself.*

PA APPLICATION: EDUCATION—GUIDED STUDYING FOR TARGETED LEARNING

1. **What's predicted:** Which questions a student will get right or wrong.
2. **What's done about it:** Spend more study time on the questions the student will get wrong.

This bolstered the brainiac for a breakout. On *Jeopardy!*, Roger set the all-time record for a single-game win of $77,000 and continued on, winning more than $230,000 during a seven-day run that placed him as the third-highest winning contestant (regular season) to date. He was invited back a year later for a "Tournament of Champions" and took its $250,000 first place award. He estimates his own ability to correctly answer 90 percent of *Jeopardy!* questions, placing him among a small handful of all-time best players.

Analyzing roughly 211,000 *Jeopardy!* questions (downloaded as IBM did from online archives maintained by fans of the game show), Roger gained perspective on its knowledge domain. If you learn about 10,000 to 12,000 answers, he told me, you've got most of it covered. This includes countries,

states, presidents, and planets. But among many categories, you only need to go so far. Designed to entertain its audience, *Jeopardy!* doesn't get too arcane. So you only need to learn about the top cities, elements, movies, and flowers. In classical music, knowing a couple of dozen composers and the top few works of each will do the trick.

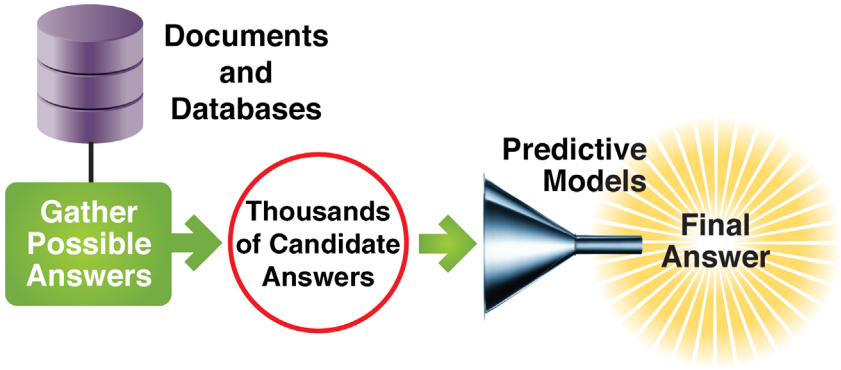
These bounds are no great relief to those pursuing the holy grail of open question answering. Predictive models often choose between only two options: *Will the person click, buy, lie, or die—yes or no?* As if that's not hard enough, for each question, Watson must choose between more than 10,000 possible answers.

The analytical improvement of human competitors was more bad news for Watson. Allowed by Roger to access his system, Watson's soon-to-be opponent Ken Jennings borrowed the study-guiding software while preparing for the big match, crediting it as "a huge help getting me back in game mode."

AMASSING EVIDENCE FOR AN ANSWER

Here's how Watson works. Given a question, it takes three main steps:

1. Collect thousands of *candidate answers*.
2. For each answer, amass *evidence*.
3. Apply predictive models to *funnel down*.



Predictive modeling has the final say. After gathering thousands of candidate answers to a question, Watson funnels them down to spit out the single answer scored most highly by a predictive model.

Watson gathers the answers and their evidence from sources that IBM selectively downloaded, a snapshot of a smart part of the Internet that forms Watson's base of knowledge. This includes 8.6 million written documents, consisting of 3.5 million Wikipedia articles (i.e., a 2010 copy of the entire English portion thereof), the Bible, other miscellaneous popular books, a history's worth of newswire articles, entire encyclopedias, and more. This is complemented by more structured knowledge sources such as dictionaries, thesauri, and databases such as the Internet Movie Database.

Watson isn't picky when collecting the candidate answers. The system follows the strategy of casting a wide, ad hoc net in order to ensure that the correct answer is in there somewhere. It rummages through its knowledge sources in various ways, including performing search in much the same way as Internet search engines like Google do (although Watson searches only within its own internal store). When it finds a relevant document, for some document types such as Wikipedia articles, it will grab the document's title as a candidate answer. In other cases, it will nab "answer-sized snippets" of text, as Watson developers call them. It also performs certain lookups and reverse lookups into databases and dictionaries to collect more candidate answers.

Like its fictional human namesake, the partner of Sherlock Holmes, Watson now faces a classic whodunit: Which of the many suspected answers is "guilty" of being the right one?[5] The mystery can only be solved with diligent detective work in order to gather as much evidence as possible for or against each candidate. Watson pounds the pavement by once again surveying its sources.

With so many possible answers, uncertainty looms. It's a serious challenge for the machine to even be confident what *kind* of thing is being asked for. An actor? A movie? State capital, entertainer, fruit, planet, company, novel, president, philosophical concept? IBM determined that *Jeopardy!* calls for

[5] Watson was not named after this fictional detective—it was named after IBM founder Thomas J. Watson.

2,500 different types of answers. The researchers considered tackling a more manageable task by covering only the most popular of these answer types, but it turned out that even if they specialized Watson for the top 200, it could then answer only half the questions. The range of possibilities is too wide and evenly spread for a shortcut to work.

ELEMENTARY, MY DEAR WATSON

Evidence counterattacks the enemy: *uncertainty*. To this end, Watson employs a diverse range of language technologies. This is where the state of the art in NLP comes into play, incorporating the research results from leading researchers at Carnegie Mellon University, the University of Massachusetts, the University of Southern California, the University of Texas, Massachusetts Institute of Technology, other universities, and, of course, IBM Research itself.

Sometimes, deep linguistics matters. Consider this question:

When David Gondek addressed Predictive Analytics World with a keynote, he provided an example phrase that could threaten to confuse Watson:

> *In May, Gary arrived in India after he celebrated his anniversary in Portugal.*

So many words match, the system is likely to include *Gary* as a candidate answer. Search methods would love a document that includes this phrase. Likewise, Watson's evidence-seeking methods built on the comparison of words would give this phrase a high score—most of its words appear in the question at hand.

Watson needs linguistic methods that more adeptly recognize how words relate to one another so that it pays heed to, for example:

On the 27th of May 1498, Vasco da Gama landed in Kappad Beach.

Other than *in, of*, and *the*, only the word *May* overlaps with the question. However, Watson recognizes meaningful correspondences. Kappad Beach is in India. *Landed in* is a way to paraphrase *arrived in*. A 400th anniversary in 1898 must correspond to a prior event in 1498.

These matches establish support for the correct answer, Vasco da Gama. Like all candidate answers, it is evaluated for compatibility with the answer type—in this case, *explorer*, as determined from *this explorer* in the question. Vasco da Gama is indeed famed as an explorer, so support would likely be strong.

These relationships pertain to the very meaning of words, their *semantics*. Watson works with databases of established semantic relationships and seeks evidence to establish new ones. Consider this *Jeopardy!* question:

IN CELL DIVISION, MITOSIS SPLITS THE NUCLEUS AND CYTOKINESIS SPLITS THIS LIQUID CUSHIONING THE NUCLEUS

Watson's candidate answers include organelle, vacuole, cytoplasm, plasma, and mitochondria. The type of answer sought being a liquid, Watson finds evidence that the correct answer, cytoplasm, makes the cut. It looks up a record listing cytoplasm as a fluid, and has sufficient evidence that fluids are often liquids to boost cytoplasm's score on that basis.

Here, Watson performs a daredevil stunt of logic. Reasoning as humans do in the wide-open domain of *Jeopardy!* questions is an extreme sport. Fuzziness pervades—for example, most reputable sources Watson may access would state all liquids are fluids, but some are ambiguous as to whether glass is definitely solid or liquid. Similarly, all people are mortal, yet infamous people have attained immortality. Therefore, a strict hierarchy of concepts just can't

apply. Because of this, as well as the vagueness of our languages' words and the difference context makes, databases of abstract semantic relationships disagree madly with one another. Like political parties, they often fail to see eye to eye, and a universal authority—an absolute, singular truth—to reconcile their differences simply does not exist.

Rather than making a vain attempt to resolve these disagreements, Watson keeps all pieces of evidence in play, even as they disagree. The resolution comes only at the end, when weighing the complete set of evidence to select its final answer to a question. In this way, Watson's solution is analogous to yours. Rather than absolutes, it adjusts according to context. Some songs are both a little bit country and a little bit rock and roll. With a James Taylor song, you could go either way.

On the other hand, keeping an "open mind" by way of this sort of flexible thinking can lead to embarrassment. Avoiding absolutes means playing fast and loose with semantics, leaving an ever-present risk of gaffes—that is, mistaken answers that seem all too obvious to us humans. For example, in Watson's televised *Jeopardy!* match, it faced a question under the category "U.S. Cities":

ITS LARGEST AIRPORT IS NAMED FOR A WORLD WAR II HERO; ITS SECOND LARGEST, FOR A WORLD WAR II BATTLE *

Struggling, Watson managed to accumulate only scant evidence for its candidate answers, so it would never have buzzed in to attempt the question. However, this was the show's "Final *Jeopardy!*" round, so a response from each player was mandatory. Instead of the correct answer, Chicago, Watson answered with a city that's not in the United States at all, Toronto. Canadian game show host Alex Trebek poked a bit of fun, saying that he had learned something new.

English grammar matters. To answer some questions, phrases must be properly deconstructed. Consider this question:

HE WAS PRESIDENTIALLY PARDONED ON SEPT. 8, 1974

In seeking evidence, Watson pulls up this phrase, which appeared in a *Los Angeles Times* article:

Ford pardoned Nixon on Sept. 8, 1974.

Unlike you, a computer won't easily see the answer must be Nixon rather than Ford. Based on word matching alone, this phrase provides equal support for Ford as it does for Nixon. Only by detecting that the question takes the passive voice, which means the answer sought is the receiver rather than the issuer of a pardon, and by detecting that the evidence phrase is in the active voice, is this phrase correctly interpreted as stronger support for Nixon than Ford.[6]

NLP's attempts to grammatically deconstruct don't always work. Complementary sources of evidence must be accumulated, since computers won't always grok the grammar. Language is tricky. Consider this question:

MILORAD CAVIC ALMOST UPSET THIS MAN'S PERFECT 2008 OLYMPICS, LOSING TO HIM BY ONE HUNDREDTH OF A SECOND *

A phrase like this could be stumbled upon as evidence:

Sam was upset before witnessing the near win by Milorad Cavic.

If *upset* is misinterpreted as a passively voiced verb rather than an adjective, the phrase could be interpreted as evidence for Sam as the question's answer.

[6] Watson employs as its main method for grammatical parsing the *English Slot Grammar*, by IBM's own researcher Michael McCord (I had the pleasure to use this tool for my doctoral research in the mid-1990s).

However, it was swimmer Michael Phelps who held on to his perfect 2008 Olympics performance. Even detecting the simplest grammatical structure of a sentence depends on the deep, often intangible meaning of words.

MOUNTING EVIDENCE

There's no silver bullet. Whether interpreting semantic relationships between words or grammatically deconstructing phrases, language processing is brittle. Even the best methods are going to get it wrong a lot of the time. This predicament is exacerbated by the clever, intricate manner in which questions are phrased on *Jeopardy!* The show's question writers have adopted a playful, informative style in order to entertain the TV viewers at home.

The only hope is to accumulate as much evidence as possible, searching far and wide for support of, or evidence against, each candidate answer. Every little bit of evidence helps. In this quest, diversity is the name of the game. An aggregate mass of varied evidence stands the best chance, since neither the cleverest nor the simplest method may be trusted if used solo. Fortunately, diversity comes with the territory: As with scientific research in general, the NLP researchers who developed the methods at hand each worked to distinguish their own unique contribution, intentionally differentiating the methods they designed from those of others.

Watson employs an assorted number of evidence routines that assess a candidate answer, including:

- **Passage search.** After inserting the candidate answer into the question to try it on for size (e.g., "*Nixon* was presidentially pardoned on Sept. 8, 1974") and searching, do many matches come up? How many match word for word, semantically, and after grammatical deconstruction? What's the longest similar sequence of words that each found phrase has in common with the question?
- **Popularity.** How common is the candidate answer?
- **Type match.** Does the candidate match the answer type called for by the question (e.g., entertainer, fruit, planet, company, or novel)? If it's a person, does the gender match?

- **Temporal.** Was the candidate in existence within the question's time frame?
- **Source reliability.** Does the evidence come from a source considered reliable?

For each question, you never know which of these factors (and the hundreds of variations thereof that Watson measures) may be critical to arriving at the right answer. Consider this question:

Although the correct answer is Argentina, measures of evidence based on simple search show overwhelming support for Bolivia due to a certain border dispute well covered in news articles. Fortunately, enough other supporting evidence such as from logically matched phrases and geographical knowledge sources compensates and wins out, and Watson answers correctly.

Some may view this ad hoc smorgasbord of techniques as a hack, but I do not see it that way. It is true that the most semantically and linguistically intricate approaches are brittle and often just don't work. It can also be said that the remaining methods are harebrained in their oversimplicity. But a collective capacity *emerges* from this mix of components, which blends hundreds of evidence measurements, even if each alone is crude.[7]

[7] Watson and PA in general are not designed to simulate how people think, predict, learn language, or answer questions. But it may be worth considering that, although as a human you experience a feeling of confidence and certainty in your answer to some questions, some components of the cognition that lead you there may be just as harebrained in isolation as Watson's components. Sometimes you have a specific recollection of the answer, as Watson does in certain cases of strong singular evidence. At other times, your confidence may only feel like a strong hunch, possibly based on a large number of weak factors.

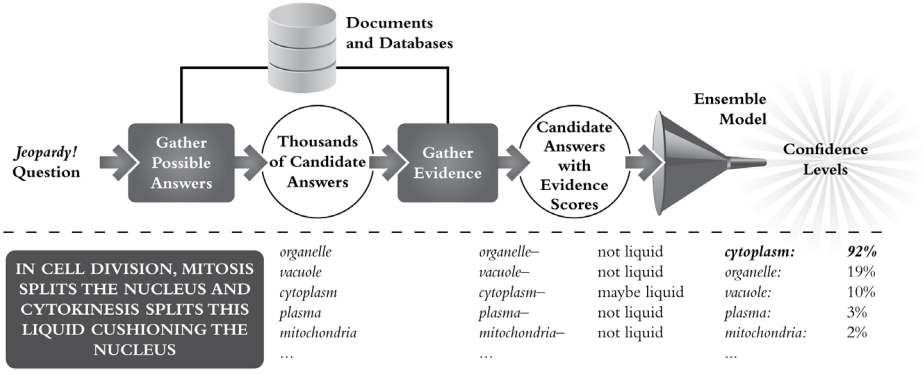
The Ensemble Effect comes into full play: The sheer count and diversity of approaches make up for their individual weaknesses. As a whole, the system achieves operational proficiency on a previously unachievable, far-off goal: open human language question answering.

WEIGHING EVIDENCE WITH ENSEMBLE MODELS

> *There are two ways of building intelligence. You either know how to write down the recipe, or you let it grow itself. And it's pretty clear that we don't know how to write down the recipe. Machine learning is all about giving it the capability to grow itself.*
>
> —Tom Mitchell, founding chair of the world's first Machine Learning Department (at Carnegie Mellon University)

The key to optimally joining disparate evidence is machine learning. Guided by the answer key for roughly 25,000 *Jeopardy!* questions, the learning process discovers how to weigh the various sources of evidence for each candidate answer. To this end, David Gondek led the application of machine learning in developing Watson. He had his hands on the very process that brings it all together.

Synthesizing sources of evidence to select a single, final answer propels Watson past the limits of Internet search and into the formerly unconquered domain of question answering. Here's a more detailed overview:



An overview of key steps Watson takes for each question, with an example question and its candidate answers along the bottom. An ensemble model selects the final answer from thousands of candidates.

As shown, Watson gathers candidate answers and then evidence for each candidate. Its ensemble model then scores each candidate answer with a confidence level so that it may be ranked relative to the other candidates. Watson then goes with the answer for which it holds the highest confidence, speaking it out loud when prompted to do so on *Jeopardy!*

PA APPLICATION: OPEN QUESTION ANSWERING

1. **What's predicted:** Given a question and one candidate answer, whether the answer is correct.
2. **What's done about it:** The candidate answer with the highest predictive score is provided by the system as its final answer.

AN ENSEMBLE OF ENSEMBLES

David led the design of Watson's innovative, intricate machine learning components, of which the ensembling of models is part and parcel. Moving from document search to open question answering demands a great leap, so the design is a bit involved. Watson incorporates ensembling in three ways:

1. **Combining evidence.** Hundreds of methods provide evidence scores for each candidate answer. Instead of tallying a simple vote across contributing evidence scores, as in some work with ensembles described in the prior chapter, the method takes it a step further by training a model to decide how best to fuse them together.[8]
2. **Specialized models by question type.** Watson has separate specialized ensemble models for specific question types, such as puzzle, multiple choice, date, number, translation, and etymology (about the

[8] *Ensemble model* commonly refers to the combination of trained predictive models. However, many of Watson's evidence-scoring methods themselves were hand-designed by experts rather than developed by learning over data, so I am using the term a bit more broadly. But The Ensemble Effect is at play; the strengths of cooperating methods make up for one another's weaknesses.

history and origin of words) questions. In this way, Watson consists of *an ensemble of ensembles*.

3. **Iterative phases of predictive models.** For each question, Watson iteratively applies several phases of predictive models, each of which can compensate for mistakes made by prior phases. Each phase filters candidates and refines the evidence. The first phase filters down the number of candidate answers from thousands to about one hundred, and subsequent phases filter out more. After each phase's filtering, the evidence scores are reassessed and refined relative to the now-smaller list of candidate answers. A separate predictive model is developed for each phase so that the ranking of the shrinking list of candidates is further honed and refined. With these phases, Watson consists of *an ensemble of ensembles of ensembles*.

MACHINE LEARNING ACHIEVES THE POTENTIAL OF NATURAL LANGUAGE PROCESSING

Despite this complexity, Watson's individual component models are fairly straightforward: they perform a weighted vote of the evidence measures. In this way, some forms of evidence count more, and others count less. Although David tested various modeling methods, such as decision trees (covered in Chapter 4), he discovered that the best results for Watson came from another modeling technique called *logistic regression*, which weighs each input variable (i.e., measure of evidence), adds them up, and then formulaically shifts the resulting sum a bit for good measure.[9]

Since the model is made up of weights, the modeling process learns to literally *weigh the evidence* for each candidate answer. The predictive model filters out weak candidate answers by assigning them a lower score. It doesn't

[9] After the weighted sum, logistic regression transforms the result with a function called an *S-curve* (aka *sigmoid squashing function*). The S-curve is designed to help predictive models with binary (twofold) target outputs, such as answering a yes/no question: *Is this answer correct, given the cumulative evidence?*

help Watson derive better candidate answers—rather, it cleans up the bulky mass of candidates, narrowing down to one final answer.

To this end, the predictive models are trained over 5.7 million examples of a *Jeopardy!* question paired with a candidate answer. Each example includes 550 predictor variables that summarize the various measures of evidence aggregated for that answer (therefore, the model is made of 550 weights, one per variable). This large amount of training data was formed out of 25,000 *Jeopardy!* questions. Each question contributes to many training examples, since there are many incorrect candidate answers. Both the correct and incorrect answers provide experience from which the system learns how to best weigh the evidence.

Watson leverages The Ensemble Effect, propelling the state of the art in language processing to achieve its full potential and conquer open question answering. Only by learning from the guidance provided by the archive of *Jeopardy!* questions was it possible to successfully merge Watson's hundreds of language-processing methods. Predictive modeling has the effect of measuring the methods' relative strengths and weaknesses. In this way, the system quantifies how much more important evidence from linguistically and semantically deep methods can be, and just how moderately simpler word-matching methods should be weighed so that they, too, may contribute to question answering.

With this framework, the IBM team empowered itself to incrementally refine and bolster Watson in anticipation of the televised *Jeopardy!* match—and moved the field of question answering forward. The system allows researchers to experiment with a continually growing range of language-processing methods: Just throw in a new language-processing technique that retrieves and reports on evidence for candidate answers, retrain the system's ensemble models, and check for its improved performance.

As David and his team expanded and refined the hundreds of evidence-gathering methods, returns diminished relative to efforts. Performance kept improving, but at a slower and slower pace. However, they kept at it, squeezing every drop of potential out of their brainshare and data, right up until the final weeks before the big match.

CONFIDENCE WITHOUT OVERCONFIDENCE

Both experts and laypeople mistake more confident predictions for more accurate ones. But overconfidence is often the reason for failure. If our appreciation of uncertainty improves, our predictions can get better too.

—Nate Silver, *The Signal and the Noise: Why So Many Predictions Fail—but Some Don't*

You got to know when to hold 'em, know when to fold 'em.

—Don Schlitz, "The Gambler" (sung by Kenny Rogers)

Jeopardy! *wasn't built for players with no self-doubt.*

—Chris Jones, *Esquire* magazine

Besides answering questions, there's a second skill each *Jeopardy!* player must hone: assessing self-confidence. Why? Because you get penalized by answering incorrectly. When a question is presented, you must decide whether to attempt to buzz in and provide an answer. If you do, you'll either gain the dollar amount assigned to the question or lose as much.

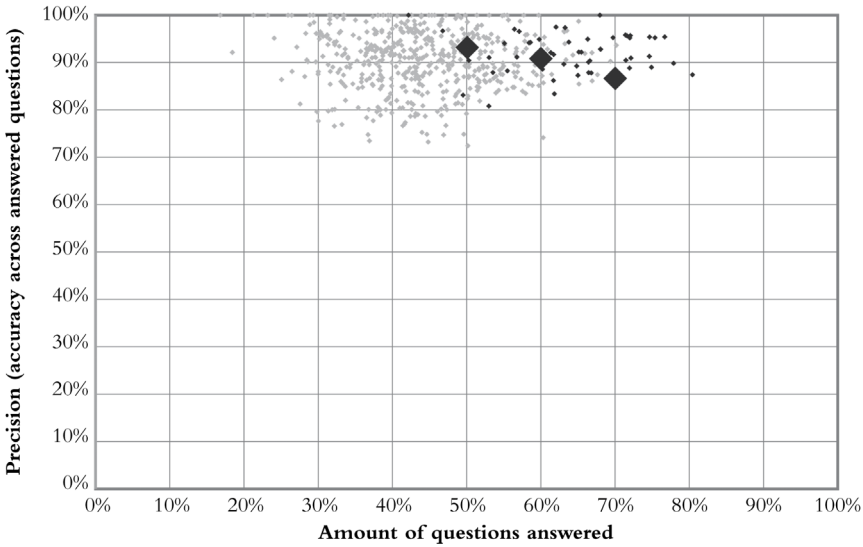
In this way, *Jeopardy!* reflects a general principle of life and business: *You need not do everything well; select the tasks at which you excel.* It's the very practice of putting your best foot forward. In fact, many commercial uses of PA optimize on this very notion. Just as Watson must predict which questions it can successfully answer, businesses predict which customers will be successfully sold to—and therefore are worth the expenditure of marketing and sales resources.

Calculating a measure of self-confidence in each answer could be a whole new can of worms for the system. Is it a tall order to require the machine to "know thyself" in this respect?

David Gondek showed that this problem could be solved "for free." The very same predictive score output by the models that serves to select the best answer also serves to estimate confidence in that answer. The scores are probabilities. For example, if a candidate answer with a score of 0.85 has a higher score than every other candidate, it will be Watson's final answer, and Watson will consider its chance of being correct at 85 percent. As the IBM team put it, "Watson knows what it knows, and it knows what it doesn't know."

Watching Watson's televised *Jeopardy!* matches, you can see these self-confidence scores in action. For each question, Watson's top three candidate answers are displayed at the bottom of your TV screen along with their confidence scores (for example, see the second figure in this chapter). Watson bases its decision to buzz in on its top candidate's score, plus its position in the game relative to its opponents. If it is behind, it will play more aggressively, buzzing in even if the confidence is lower. If ahead in the game, it will be more conservative, buzzing in to answer only when highly confident.

A player's success depends not only on how many answers are known, but on his, her, or its ability to assess self-confidence. With that in mind, here's a view that compares *Jeopardy!* players:



***Jeopardy!* player performances. Each dot signifies a winner's game (the dark dots represent Ken Jennings's games). The three large diamonds represent the per-game performance Watson can achieve.[10]**

[10] Graph adapted from D. Ferrucci et al., "Building Watson: An Overview of the DeepQA Project," *AI Magazine* 31, no. 3 (2010), 59–79.

Players strive for the top right of this graph. Most points on the graph depict the performance of an individual human player. The horizontal axis indicates what proportion of questions they successfully buzzed in for, and the vertical axis tells us, for those questions they answered, how often they were correct. Buzzing in more would put you further to the right, but would challenge you with the need to know more answers.

Human *Jeopardy!* winners tend toward the top, since they usually answer correctly, and some also reach pretty far to the right. Each light gray dot represents the performance of the winner of one game. The impressively positioned dark gray dots that stretch further to the right represent the outstanding performance of champion player Ken Jennings, whose breath-taking streak of 74 consecutive wins in 2004 demonstrated his prowess. He is one of the two champions against whom Watson was preparing to compete.

Watson performs at the level of human experts. Three example points (large diamonds) are shown to illustrate Watson's potential performance. When needed, Watson sets itself to buzz in more often, assuming an aggressive willingness to answer even when confidence is lower. This moves its performance to the right and, as a result, also a bit down. Alternatively, when playing more conservatively, fewer questions are attempted, but Watson's answer is more often correct—precision is higher (unlike politics, on this graph left is more conservative).

Human sweat empowered Watson's human level of performance. The machine's proficiency is the product of four painstaking years of perseverance by the team of researchers.[11]

THE NEED FOR SPEED

There was one more requirement. Watson had to be fast.

[11] The industry is taken with Watson. A Predictive Analytics World keynote address by Watson's machine learning leader, David Gondek, dazzled a ballroom of industry insiders, who on average rated the speech's content at an unmatched 4.7 out of 5 in a subsequent poll.

A *Jeopardy!* player has only a few seconds to answer a question, but on a single computer (e.g., 2.6 gigahertz), determining an answer can take a couple of hours. It's a lengthy process because Watson employs hundreds of methods to search a huge number of sources, both to accrue candidate answers and to collect evidence measurements for each one. It then predictively scores and ranks the candidates by applying the series of predictive models (I refer here only to the deployed use of Watson to play *Jeopardy!*, after the machine learning process is completed and the models are being employed without further learning).

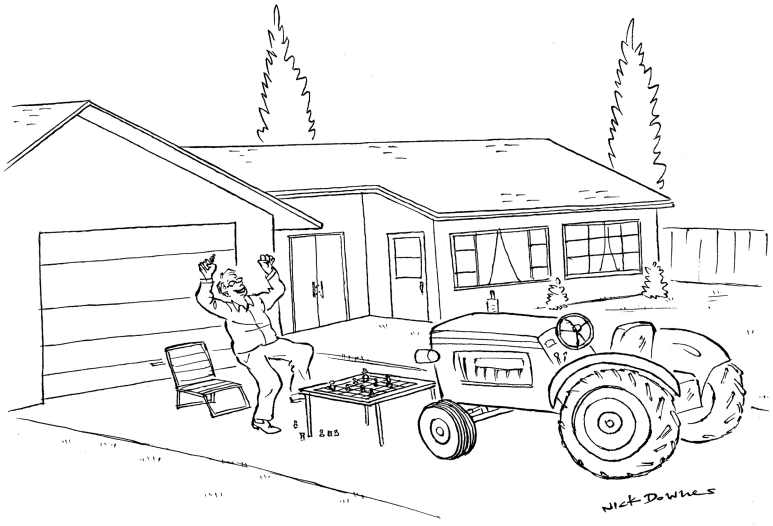
To make it thousands of times faster, Watson employs thousands of CPUs. This supercomputer clobbers bottlenecks and zips along, thanks to a cluster of 90 servers consisting of 2,800 core processors. It handles 80 trillion operations per second. It favors 15 terabytes of RAM over slower hard-drive storage. The cost of this hardware brawn is estimated to come to $3 million, a small fraction of the cost to develop its analytical software brains.

Having thousands of CPUs means that thousands of tasks can be done simultaneously, in parallel. Watson's process lends itself so amenably to taking advantage of this hardware by way of distribution into contemporaneous subtasks that the research team considers it *embarrassingly parallel*. For example, each evidence-seeking, language-processing routine can be assigned to its own processor.

Better is bigger. To assemble Watson, IBM crated in a mammoth configuration of hardware, about 10 refrigerators' worth. Watson didn't go to *Jeopardy!*; *Jeopardy!* came to Watson, setting up a temporary game show studio within IBM's T. J. Watson Research Center.

DOUBLE *JEOPARDY!*—WOULD WATSON WIN?

Watson was not sure to win. During sparring games against human champions, Watson had tweaked its way up to a 71 percent win record. It didn't always win, and these trial runs didn't pit it against the lethal competition it was preparing to face on the televised match: all-time leading *Jeopardy!* champions Ken Jennings and Brad Rutter.



"Once again, man beats machine!"

Reproduced with permission.

The *Jeopardy!* match was to gain full-scale media publicity, exposing IBM's analytical prowess or failure. The top-rated quiz show in syndication, *Jeopardy!* attracts nearly 9 million viewers every day and would draw an audience of 34.5 million for this special man-versus-machine match. If the massive popularity of *Jeopardy!* put on the pressure, so too was it the only reason this grand challenge might be doable. As the United States' greatest pop culture institution of human knowledge, *Jeopardy!*'s legacy provided the treasure trove of historical question/answer pairs from which Watson learns.

Beyond impressing or disappointing your average home viewer, Watson's impending performance held enormous professional ramifications. Within both the practical realm of information technology and the research world of artificial intelligence, IBM had loudly proclaimed that it was prepared to run a three-and-a-half-minute mile. After the immense investment, one can only imagine the seething pressure the research team must have felt from the powers at IBM to defend the corporate image and ensure against public humiliation. At this juncture, the researchers saw clear implications for their scientific careers as well as for science itself.

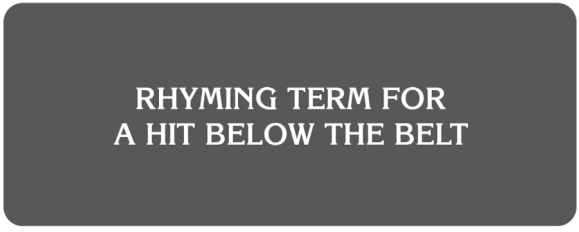
During its formative stages, Watson's most humorous mistakes entertained, but threatened to embarrass IBM on national TV. Under the category "The Queen's English":

Watson said: *urinate* (correct answer: call on the phone).

Under the category "*New York Times* Headlines":

Watson said: *a sentence* (correct answer: World War I).

Under the category "Boxing Terms":



Watson said: *wang bang* (correct answer: low blow).

The team rallied for the home stretch. Watson principal investigator David Ferrucci, who managed the entire initiative, moved everyone from their offices into a common area he considered akin to a war room, cultivating a productive but crisislike level of eustress. Their lives were flipped on their

heads. David Gondek moved temporarily into a nearby apartment to eliminate his commuting time. The team lived and breathed open question answering. "I think I dream about *Jeopardy!* questions now," Gondek said. "I have nightmares about *Jeopardy!* questions. I talk to people in the form of a question."

JEOPARDY! JITTERS: DEPLOYING A PROTOTYPE

There's no such thing as human error. Only system error.

—Alexander Day Chaffee, software architect

Core Watson development team member Jennifer Chu-Carroll tried to stay calm. "We knew we probably were gonna win, but . . . what if we did the math wrong for some reason and lost by a dollar instead of won by a dollar?" There were provisions in their agreement with the *Jeopardy!* producers for do-overs in the case of a hardware crash (the show was taped, not broadcast live, and like any computer, sometimes you need to turn off Watson and then start it back up again). However, if Watson spat out an embarrassing answer due to a software bug without crashing, nothing could be done to take it back. This was going to national television.

Groundbreaking deployments of new technology—whether destined to be in orbit or intelligent—risk life and limb, not only because they boldly go where no one has gone before, but because they launch a prototype. Moon-bound *Apollo 11* didn't roll off the assembly line. It was the first of its kind. The Watson system deployed on *Jeopardy!* was beta. Rather than conducting the established, sound process of "productizing" a piece of software for mass distribution, this high-speed, real-time behemoth was constructed not by software engineers who build things, but by the same scientific researchers who designed and developed its analytical capabilities. On the software side, the deployed system and the experimental system were largely one and the same. There was no clear delineation between some of the code they used for iterative, experimental improvement with machine learning and code within the deployed system. Of course, these were world-class researchers, many

with software design training, but the pressure mounted as these scientists applied virtual hammer to nail to fashion a vessel that would propel their laboratory success into an environment of high-paced, unforeseen questions.

Shedding their lab coats for engineering caps, the team members dug in as best they could. As David Gondek told me, changes in Watson's code continued even until and including the very day before the big match, which many would consider a wildly unorthodox practice in preparing for a mission-critical launch of software. Nobody on the team wanted to be the programmer who confused metric and English imperial units in their code, thus crashing NASA's Mars Climate Orbiter, as took place in 1998 after a $327.6 million, nine-month trip to Mars. Recall the story of the Netflix Prize (see Chapter 5), which was won in part by two nonanalysts who found that their expertise as professional software engineers was key to their success.

The brave team nervously saw Watson off to meet its destiny. The training wheels were off. Watson operates on its own, self-contained and disconnected from the Internet or any other knowledge source. Unlike a human *Jeopardy!* player, the one connection it does need is an electrical outlet. It's scary to watch your child fly from the nest. Life has no safety net.

As a machine, Watson was artificial. The world would now witness whether it was also intelligent.

FOR THE WIN

You are about to witness what may prove to be an historic competition.

—Alex Trebek

If functional discourse in human language qualifies, then the world was publicly introduced to the greatest singular leap in artificial intelligence on February 14, 2011.

As the entertainment industry would often have it, this unparalleled moment in scientific achievement was heralded first with Hollywood cheese, and only secondarily with pomp and circumstance. After all, this

was a populist play. It was, in a sense, the very first conversant machine ever, and thus potentially easier for everyone to relate to than any other computer. Whether perceived as *Star Trek*-ian electronic buddy or HAL-esque force to be reckoned with, 34.5 million turned on the TV to watch it do its thing.

The *Jeopardy!* theme song begins to play,[12] and a slick, professional voice manically declares, "From the T. J. Watson Research Center in Yorktown Heights, New York, this is *Jeopardy!*, the IBM Challenge!"

When colleagues and I watch the footage, there's a bit of culture shock: We're looking for signs of AI, and instead see glitzy show business. But this came as no surprise to the members of Team Watson seated in the studio audience, who had been preparing for *Jeopardy!* for years.

Once the formalities and introductions to Watson pass, the show moves along jauntily as if it's just any other episode, as if there is nothing extraordinary about the fact that one of the players spitting out answer after answer is not an articulate scholar with his shirt buttoned up to the top, but instead a robot with a synthetic voice straight out of a science fiction movie.

But for David Gondek and his colleagues it was anything but ordinary. The team endured a nail-biting day during the show's recording, one month before its broadcast. Watching the two-game match, which was televised over a three-day period, you see dozens of questions fly by. When the camera turns for audience reactions, it centers on the scientists, David Ferrucci, David Gondek, Jennifer Chu-Carroll, and others, who enjoy moments of elation and endure the occasional heartache.

On this day, Machine triumphed over Man. Watson answered 66 questions correctly and eight incorrectly. Of those eight, only the answer that categorized Toronto as a U.S. city was considered a gaffe by human standards. The example questions covered in this chapter marked with an asterisk ("*") were fielded by Watson during the match (all correctly except

[12] This well-known tune is a simple exercise in major fifths composed by Merv Griffin, *Jeopardy!*'s creator. In contradiction with what some consider a mind-numbing quality, the song's title is the same as the IBM motto coined by the company's founder, Thomas Watson: "Think."

the one answered with Toronto). The final scores, measured in *Jeopardy!* as dollars, were Watson: $77,147, Jennings: $24,000, and Rutter: $21,600.[13]

Prompted to write down his answer to the match's final question, Ken Jennings, invoking a *Simpsons* meme originating from an H. G. Wells movie, appended an editorial: "I, for one, welcome our new computer overlords." He later ruminated, "Watson has lots in common with a top-ranked human *Jeopardy!* player: It's very smart, very fast, speaks in an uneven monotone, and has never known the touch of a woman."

AFTER MATCH: HONOR, ACCOLADES, AND AWE

> *I would have thought that technology like this was years away, but it's here now. I have the bruised ego to prove it.*
>
> —Brad Rutter

> *This was to be an away game for humanity, I realized.*
>
> —Ken Jennings

> *Maybe we should have toned it down a notch.*
>
> —Sam Palmisano, then CEO, IBM

One million-dollar first place award for the *Jeopardy!* match? Check (donated to charities). American Technology Awards' "Breakthrough Technology of the Year Award"? Check. *R&D* magazine "Innovator of the Year" award? Check.

Webby "Person of the Year" award? Unexpected, but check.

Riding a wave of accolades, IBM is working to reposition components of Watson and its underlying question-answering architecture, which the company calls *DeepQA*, to serve industries such as healthcare and finance. Consider medical diagnosis. The wealth of written knowledge is so great, no doctor could read it all; providing a ranked list of candidate diagnoses for each

[13] This strong lead was due at least in part to the speed with which Watson could buzz in to answer questions, although that issue is involved and debated; it is complicated to truly level the playing field when human and machine compete.

patient could mean doctors miss the right one less often. Guiding the analysis of knowledge sources by learning from training data—answers in the case of *Jeopardy!* and diagnoses in the case of healthcare—is a means to "capture and institutionalize decision-making knowledge," as Robert Jewell of IBM Watson Solutions put it to me.

IAMBIC IBM AI

Is Watson intelligent? The question presupposes that such a concept is scientific in the first place. The mistake has been made, as proselytizers have often "over-souled" AI (credit for this poignant pun goes to Eric King, president of the consultancy he dubbed with the double entendre The Modeling Agency). It's easy to read a lot into the thing. Case in point: I once designed a palindrome-generation system (a palindrome reads the same forward and backward) when teaching the AI course at Columbia University that spontaneously derived "Iambic IBM AI." This one is particularly self-referential in that its meter is iambic.

Some credit Watson with far too much smarts. A guard working at IBM's research facility got David Gondek's attention as he was leaving for the day. Since this was a machine that could answer questions about any topic, he suggested, why not ask it who shot JFK?

Strangely, even technology experts tend to answer this philosophical question with a strong opinion in one direction or the other. It's not about right and wrong. Waxing philosophical is a dance, a wonderful, playful pastime. I like to join in the fun as much as the next guy. Here are my thoughts:

> *Watching Watson rattle off one answer after another to diverse questions laced with abstractions, metaphors, and extraneous puns, I am dumbfounded. It is the first time I've felt compelled to anthropomorphize a machine in a meaningful way, well beyond the experience of suspending disbelief in order to feel fooled by a magic trick. To me, Watson looks and feels adept, not just with information but with knowledge. My perceptions endow it with a certain capacity to cogitate. It's a sensation I never thought I'd have cause to experience in my lifetime. To me, Watson is the first artificial intelligence.*

If you haven't done so, I encourage you to watch the *Jeopardy!* match (see the Notes at www.PredictiveNotes.com for a YouTube link).

PREDICT THE RIGHT THING

Predictive models are improving and achieving their potential, but sometimes predicting what's going to happen misses the point entirely. Often, an organization needs to decide what next action to take. One doesn't just want to predict what individuals will do—one wants to know what to do about it. To this end, *we've got to predict something other than what's going to happen*—something else entirely. Turn to the next chapter to find out what.

Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die, Revised and Updated Edition. Eric Siegel.

CHAPTER 7

Persuasion by the Numbers

How Telenor, U.S. Bank, and the Obama Campaign Engineered Influence

What is the scientific key to persuasion? Why does some marketing fiercely backfire? Why is human behavior the wrong thing to predict? What should all businesses learn about persuasion from presidential campaigns? What voter predictions helped Obama win in 2012 more than the detection of swing voters? How could doctors kill fewer patients inadvertently? How is a person like a quantum particle? Riddle: What often happens to you that cannot be perceived and that you can't even be sure has happened afterward—but that can *be predicted in advance?*

In her job in Norway, Eva Helle stood guard to protect one of the world's largest cell phone carriers from its most dire threat. Her company, Telenor, had charged her with a tough assignment because, as it happens, the mobile business was about to suddenly turn perilous.

A new consumer right exerted new corporate strain: Mobile phone numbers became portable. Between 2001 and 2004, most European countries passed legislation to mandate that, if you switch to another wireless service provider, you may happily bring your phone number along with you—you need not change it (the United States did this as well; Canada, a few years later).

As customers leaped at the chance to leave, Eva faced an old truth. You just never know how fickle people are until they're untied. The consumer gains power, and the corporation pays a price.

But, as Eva and her colleagues would soon learn, the game had changed even more than they realized. Their method to woo customers and convince

them to stay had stopped working. A fundamental shift in how customers respond to marketing forced Eva to reconsider how things were done.

CHURN BABY CHURN

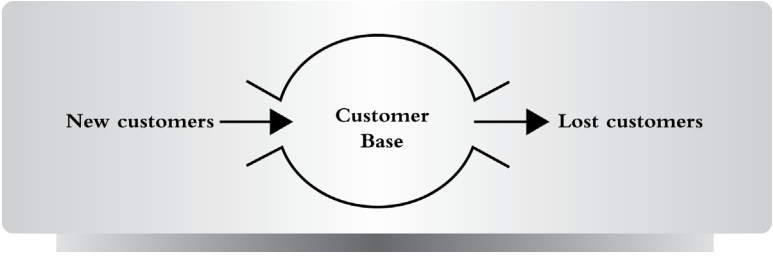
Before this change, Telenor had been successfully applying the industry's leading technique to hold on to its cell phone subscribers—a technique that applies predictive analytics (PA):

PA APPLICATION: CUSTOMER RETENTION WITH *CHURN MODELING*

1. **What's predicted:** Which customers will leave.
2. **What's done about it:** Retention efforts target at-risk customers.

Churn modeling may be the hottest marketing application of PA, and for good reason. Any seasoned executive will tell you retention is all-important because it's usually cheaper to convince a customer to stay than to acquire a new one.

Picture customer turnover as air flowing into and out of a balloon:



Retaining more customers is akin to clamping down on the nozzle on the right. Lessening the rate of loss just a bit, the balloon blossoms, magnifying its rate of expansion—that is, the growth rate of the company's customer base. This growth is the *raison d'être* of business.

Prediction and proaction are musts. Persuading someone to stay often sets a mobile carrier back a free phone or a hefty discount. A company must target this generosity where it's needed: those customers predicted to leave. Like most major cell phone carriers, Telenor had been enjoying a clear win with churn modeling.[1]

What could possibly go wrong?

[1] This book's Central Table 2 lists several more examples of applied churn modeling, and Chapter 4 reveals how Chase applied the prediction of customer departure in a unique way.

SLEEPING DOGS

If I leave here tomorrow
Would you still remember me?
For I must be traveling on, now
'Cause there's too many places I've got to see.

—From "Free Bird" by Lynyrd Skynyrd

Imagine you received an alluring brochure from your cell phone company that says:

Tantalized? Imagining a higher-tech toy in your pocket?

Now imagine you are newly emancipated, recently granted the liberty to take your phone number with you to another carrier. You've been aching to change to another carrier to join your friends who say they love it over there. In fact, your provider may have sent you this offer only because it predicted your likely departure.

Big mistake. *The company just reminded you that your contracted commitment is ending and you're free to defect.*

Contacting you backfired, increasing instead of decreasing the chance you'll leave. If you are a sleeping dog, they just failed to let you lie.

Bad news piled on. While already struggling against rising rates of defection, Eva and her colleagues at Telenor detected this backfiring of their efforts to retain, a detrimental occurrence that was now happening more often. More customers were being inadvertently turned away, triggered to leave when they otherwise, if not contacted, might have stayed. It was no longer business as usual.

A NEW THING TO PREDICT

You didn't have to be so nice; I would have liked you anyway.

—The Lovin' Spoonful, 1965

D'oh!

—Homer Simpson

This newly dominant phenomenon brought up for Telenor the question of what PA should be used to predict in the first place. Beyond predicting departure, must a company secondarily predict how customers will respond when contacted? Must we predict the more complicated, two-part question, "Who is leaving but would stay if we contacted them?" This sounds pretty

convoluted. To do so, it seems like we'd need data tracking when people *change their minds*!

This question of integrating a secondary prediction also pertains to another killer app of PA, the utterly fundamental targeting of marketing:

PA APPLICATION: TARGETED MARKETING WITH *RESPONSE MODELING*

1. **What's predicted:** Which customers will purchase if contacted.
2. **What's done about it:** Contact those customers who are more likely to do so.

Despite *response modeling*'s esteemed status as the most established business application of PA (see the 12 examples listed in this book's Central Table 2), it falls severely short because it predicts the outcome for those we *do* contact, but not for those left uncontacted. Assume we have contacted these individuals:

If the dark gray individuals made a purchase, we may proceed with patting ourselves on the back. We must have done a great job of targeting by way of astute predictions about who would buy if contacted, since so many actually did so—relative to how direct marketing often goes, achieving response rates of a few percent, 1 percent, or even less.

One simple question jolts the most senior PA expert out of a stupor: *Which of the dark gray individuals would have purchased anyway, even if we hadn't contacted them?* In some cases, up to half of them—or even more—are so prone to purchasing, they would have either way.

Even an analytics practitioner with decades of experience tweaking predictive models can be floored and flabbergasted by this. She wonders to herself, "Have I been predicting the wrong thing the whole time?" Another bonks himself on the head, groaning, "Why didn't I ever think of that?" Analytics labs echo with the inevitable Homer Simpson exclamation, "D'oh!"

Let's step back and look logically at an organization's intentions:

- The company wants customers to stay and to buy.
- The company does not intend to *force* customers (they have free will).
- Therefore, the company needs to *convince* customers—to influence, to persuade.

If persuasion is what matters, shouldn't that be what's predicted? Let's try that on for size.

Prediction goal: *Will the marketing brochure* persuade *the customer?*

Mission accomplished. This meets the company's goals with just one predictive question, integrating within it both whether the customer will do what's desired and whether it's a good idea to contact the customer.

Predicting impact impacts prediction. PA shifts substantially, from predicting a behavior to predicting *influence on behavior.*

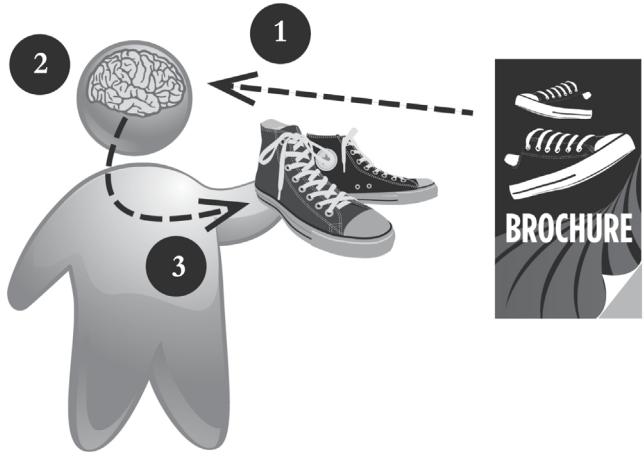
Predicting influence promises to boost PA's value, since an organization doesn't just want to know what individuals will do—it wants to know *what it can do about it.* This makes predictive scores actionable.

I know I asked this earlier but, what could possibly go wrong?

EYE CAN'T SEE IT

Houston, we have another problem.

How can you know something happened if you didn't see it? Take a look at this possible instance of influence:



1. The individual perceives the sales brochure.
2. Something happens inside the brain.
3. The individual buys the product.

Is it safe to assume influence took place? How do we know the brochure made a difference? *Perhaps the individual would have purchased anyway.*

The brain's a black box into which we cannot peek. Even if we were conducting neuroscience, it's not clear if and when that field of science will progress far enough to detect when one changes one's mind (and even if it could, we'd need brain readings from each consumer to employ it!).

Introspection doesn't work, either. You cannot always report on how your own decision making took place. You just can't be certain what made a difference, whether your friend, client, sister, or even you yourself would have made a purchase if circumstances had been different.

To observe influence, we'd need to detect *causality*: Did the brochure *cause* the individual to purchase? As explored in Chapter 3, our knowledge about causality is limited. To truly know causality would be to fully understand how things in the world affect one another, with all the detail involved, the chain reactions that lead one event to result in another. This is the domain of physics, chemistry, and other sciences. It's How the World Works. Ultimately, science tells us only a limited amount.

Therefore, *influence cannot be observed.* We can never witness an individual case of persuasion with complete certainty.

How, then, could we ever predict it?

PERCEIVING PERSUASION

No man ever steps in the same river twice.

—Heraclitus

Good grief. The most valuable thing to predict can't even be detected in the first place.

The desire to influence drives every move we make. As organizations or individuals, out of self-interest or altruistically, almost everything we do is meant to produce a desired effect, including:

- Send a brochure to a customer (or voter).
- Prescribe a medication to a patient.
- Provide social benefits intended to foster self-sufficiency.

Each action risks backfiring: The customer cancels, the patient suffers an adverse reaction, or the beneficiary becomes dependent on assistance. So we make choices not only to pursue what will work, but also to avoid what would do more harm than good.

In one arena in particular, do we feel the pangs of misstep and failure: dating. In courtship, you are both the director of marketing and the product. You're not in the restaurant for food—rather, it is a sales call. Here are some tips and pointers to persuade. Don't be overly assertive, too frequently contacting your prospect. Yet don't remain overly passive, risking that a competitor will swoop in and steal your thunder. Try to predict what you think is the right message, and avoid communicating the wrong thing.

In the movie *Groundhog Day*, our hero Bill Murray acquires a kind of superpower: the coveted ability to perceive influence. Stuck in a magical loop, reliving the same dull day over and over, he faces a humbling sort of purgatory, apparently designed to address the character's flamboyant narcissism. He cannot escape, and he becomes despondent.

Things turn around for Bill when he recognizes that his plight in fact endows him with the ability to *test different marketing treatments on the same subject under exactly the same circumstances*—and then observe the outcome. Desperate to win over the apple of his eye (Andie MacDowell) and immune to the fallout and crush of failure, he endeavors in endless trial and error to eventually learn just the right way to woo her.

Only in this wonderful fantasy can we see with certainty the difference each choice makes. That's life. You never know for sure whether you made the optimal choice about anything. Should I have admitted I love the Bee

Gees? Should we have sent that brochure? Would the other surgical treatment have gone better? Woulda, coulda, shoulda.

In real life, there are no do-overs, so our only recourse is to predict beforehand as well as possible what will work. But, in real life, what's real? If we can't observe influence, how do we know it ever really happens at all?

PERSUASIVE CHOICES

Think before you speak.

Even in dating, there's science to persuasion. Dating website OkCupid showed that messages initiating first contact that include the word *awesome* are more than twice as likely to elicit a response as those with *sexy*. *Howdy* is better than *hey*. *Band* does better than *literature* and *video games* (go figure).

Psychology professor Robert Cialdini persuaded people to commit less crime, and proved it worked. Visitors regularly steal a precious resource from Arizona's Petrified Forest National Park: chunks of petrified wood. Cialdini measured the result of posting the following sign:

With that sign in place, the rate of theft was 1.67 percent. Next he tested another message that more strongly emphasizes the negative effect of theft:

You might expect that would further reduce theft, but it backfired. This message has the effect of destigmatizing theft, since it implies the act is common—"Everybody does it." Possibly for that reason, it resulted in more than four times as much theft as the first sign, 7.92 percent. Regardless of the psychological interpretation and whether the result is a surprise, persuasion has been proven. We can safely conclude that *relaying the first message rather than the second influences people to steal less*. Similar effects have been shown in the persuasion of hotel room towel recycling and decreasing home energy usage, as explored in Cialdini's coauthored book, *Yes! 50 Scientifically Proven Ways to Be Persuasive*.[2]

These studies prove influence takes place across a group but ascertain nothing about any one individual, so the choice of message still cannot be individually selected according to what's most likely to influence each person.

[2] Although psychological interpretations such as this destigmatizing effect are not conclusively supported by the data analysis, it is also true that persuasion "by the numbers"—the focus of this chapter—depends on the creative design of messages (more generally, treatments) to test in the first place. As always, human creativity, such as that in the field of psychology, and number crunching—the soft and the hard sciences—complement one another and are mutually interdependent.

In the field of medicine, most clinical studies do this same thing—compare two treatments and see which tends to work better overall. For knee surgery after a ski accident, I had to select a graft source from which to reconstruct my busted anterior cruciate ligament (ACL, the knee's central ligament—previously known to me as the Association for Computational Linguists). I based my decision on a study that showed subsequent knee walking was rated "difficult or impossible" by twice as many patients who donated their own patellar tissue rather than hamstring tissue.[3]

It's good, but it's not personalized. I can never know if my choice for knee surgery was the best for my particular case (although my knee does seem great now). The same holds true for any treatment decision based on such studies, which provide only a one-size-fits-all result. We're left with uncertainty for each individual patient. If you take a pill and your headache goes away, you can't know for sure that the medicine worked; maybe your headache would have stopped anyway.

More generally, if you prevent something bad, how can you be sure it was ever going to happen in the first place?

BUSINESS STIMULUS AND BUSINESS RESPONSE

Many of your everyday clicks contribute to the Web's constant testing of how to improve overall persuasiveness. Google has compared 41 shades of blue to see which elicits more clicks. Websites serve the ads that get clicked the most and run random AB tests to compare which Web page design and content lead to the most buying. Facebook conducts controlled experiments to see how changes to the rules driving which friends' posts get displayed influence your engagement and usage of their website (see Central Table 1).

[3] The decision was mine alone, with no personalized guidance from a physician. I found each knee surgeon to be almost entirely devoted to one graft source or another and therefore unable to provide balanced guidance for my choice. My only option was to first select a surgical procedure and then choose a doctor who focused on that procedure.

I tested titles for this book, following in the footsteps of *SuperCrunchers* and *The 4-Hour Workweek*. Placed as ads on Google Adwords, *Predictive Analytics*, when displayed on tens of thousands of screens of unsuspecting experimental subjects across the country, was clicked almost twice as often as *Geek Prophecies* and also beat out *I Knew You Were Going to Do That* and *Clairvoyant Computers*, plus six other book titles that I also entered into this contest. It was convenient that the field's very name came out as the top contender, an unquestionably fitting title for this book.

In both medicine and marketing, this scheme to test *treatments* reveals the impact of selecting one outward action over another—but only as a trend across the group of subjects as a whole. After this sort of experiment, the best an organization can do is run with the one most effective treatment, applying it uniformly for all individuals.

In this practice, the organization is employing a blunt instrument. Looking back, we still don't know for whom the treatment was truly effective. Looking forward, we still don't know how to make personalized choices for each individual.

THE QUANTUM HUMAN

> *Here's the thing about the future. Every time you look at it, it changes. Because you looked at it.*
>
> —Nicolas Cage's clairvoyant in *Next*

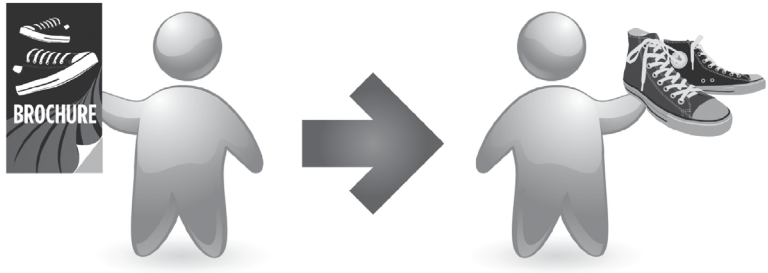
> *Heisenberg might have slept here.*
>
> —Anonymous

As in quantum physics, some things are unknowable. Although you may protest being reduced to a quantum particle, there's a powerful analogy to be drawn between the uncertainty about influence on an individual and *Heisenberg's uncertainty principle*. This principle states that we can't know everything about a particle—for example, both its position and speed. It's a trade-off. The more precisely you measure one, the less precisely you can measure the other.

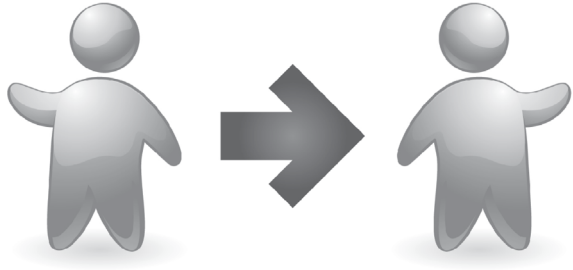
Likewise, we can't know everything about a human. In particular, we can't know both things that we'd need to know in order to conclude that a person could be influenced. For example:

1. Will Bill purchase if we send him a brochure?
2. Will Bill purchase if we *don't* send him a brochure?

If we did know the answer to both, we'd readily know this most desired fact about Bill—whether he's *influenceable*. In some cases, the answers to the two questions disagree, such as:



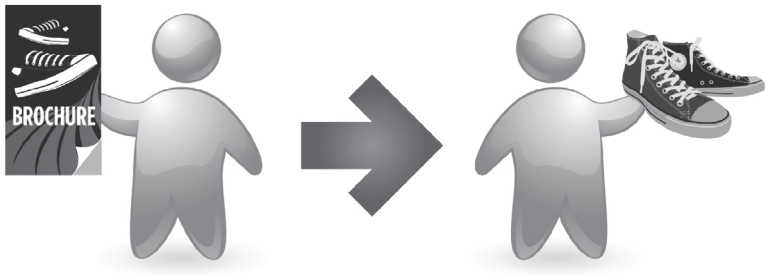
The answer to (1) is "Yes"—Bill receives a brochure and then purchases.



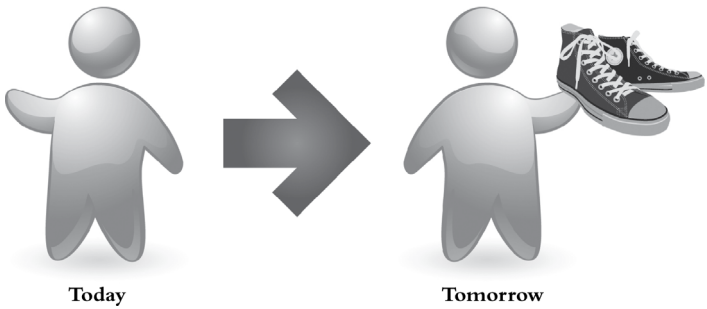
The answer to (2) is "No"—Bill does not receive a brochure and does not purchase.

In this case, we would conclude that the choice of treatment does have an influential effect on Bill; he is persuadable.

In other cases, the answers to the questions agree, such as:



The answer to (1) is "Yes"—Bill receives a brochure and then purchases.



The answer to (2) is also "Yes"—Bill does not receive a brochure but then purchases anyway.

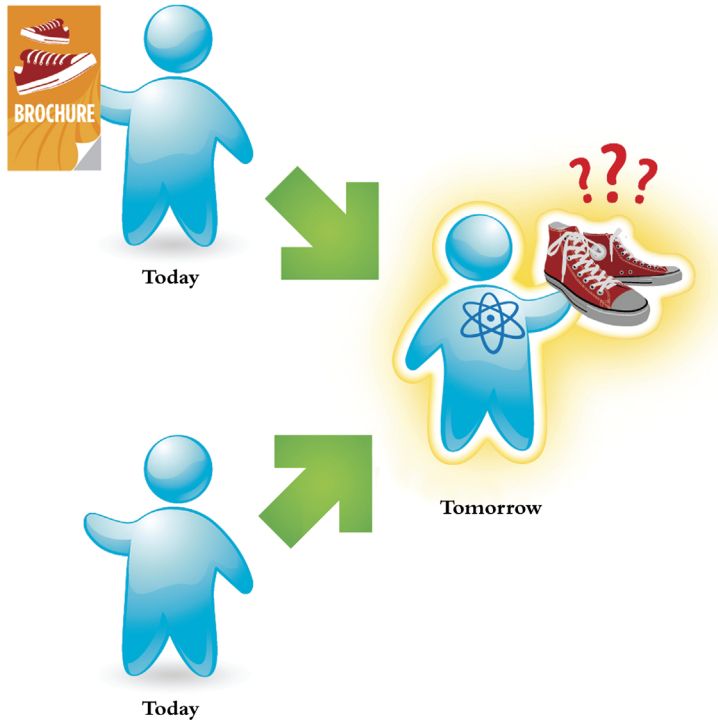
In this case, we conclude the choice of treatment has no influence; he would buy either way. This type of customer is called a *sure thing*.

Other scenarios exist. Sometimes a brochure backfires and adversely influences a customer who would otherwise buy not to.

But this is a fantasy—we *can't* know the answer to both questions. We can find out (1) by sending Bill a brochure. We can find out (2) by not sending him a brochure. But we can't both contact and not contact Bill. We can't administer medicine and not administer medicine. We can't try two different forms of surgery at once. In general, you can't test an individual with both treatments.

This uncertainty leaves us with philosophical struggles akin to those of quantum physics. Given that we could never know both, does a particle ever

really have both a true position and a true speed? Similarly, do answers to both of the previous questions about a person truly exist? Answering one renders the other purely hypothetical. It's like the tree falling in the forest with no one to perceive the sound, which becomes only theoretical. This most fundamental status of a human as influenceable or not influenceable holds only as an ethereal concept. It's only observable in aggregate across a group, never established for any one person. Does the quality of influenceability exist only in the context of a group, emergently, defying true definition for any single individual? If influenceable people do walk among us, you can never be certain who they are.



The quantum human—is he or she influenceable?

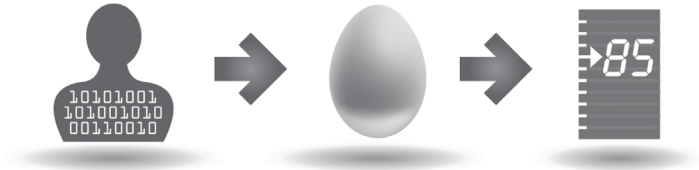
This unknowability equates the past and the future. We don't know whether a person *was* influenced, and we don't know whether the person *could be* influenced—whether he or she is *influenceable*. It's kind of a refreshing change that prediction is no more difficult than retrospection, that tomorrow

presents no greater a challenge than yesterday. Both previous and forthcoming influence can only at best be estimated. Clearly, the future is the more valuable one to estimate. If we can know *how likely* each person is to be influenced, we can drive decisions, treating each individual accordingly.

But how can you predictively model influence? That is, how could you train a predictive model when there are no learning examples—no individual known cases—of the thing we want to predict?

PREDICTING INFLUENCE WITH UPLIFT MODELING

A model that predicts influence will be a predictive model like any other:



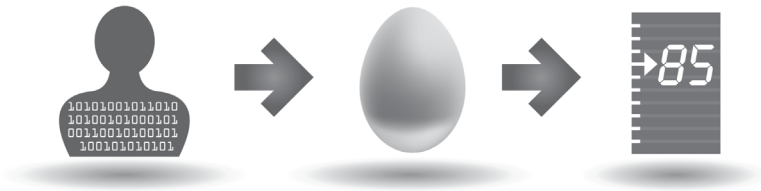
Like all the models we've covered in this book, it takes characteristics of the individual as input and provides a predictive score as output.

But it will be a special case of predictive models. Instead of predicting an outright behavior, we need *a model that scores according to the likelihood an individual's behavior will be influenced*. We need an *uplift model*:

> ***Uplift model***—*A predictive model that predicts the influence on an individual's behavior that results from applying one treatment over another.*[4]

The uplift score answers the question, *"How much more likely is this treatment to generate the desired outcome than the alternative treatment?"* It guides an organization's

[4] Not to be confused with the *lift* of a predictive model covered in Chapter 4, uplift modeling is also known as *differential response, impact, incremental impact, incremental lift, incremental response, net lift, net response, persuasion, true lift,* or *true response modeling*.



choice of treatment or action, what to do or say to each individual.[5] The secondary treatment can be the passive action of a *control set*—for example, make no marketing contact or administer a placebo instead of the trial drug—in which case an uplift model effectively decides whether or not to treat.

How do you learn about something you can't see? We never have at our disposal learning examples of the very thing we want to predict: *influenceable individuals*. We don't have the usual training data from which to directly learn.

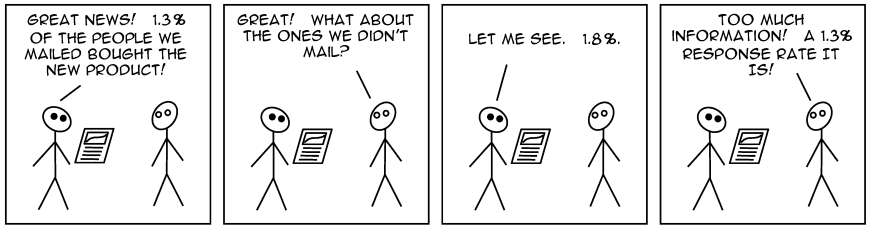
To do the seemingly impossible, *uplift modeling* needs a clever work-around. To see how it works, let's explore a detailed example from U.S. Bank.

BANKING ON INFLUENCE

U.S. Bank Assistant Vice President Michael Grundhoefer isn't satisfied with good. In the mid-1990s, the bank's direct marketing efforts to sell financial products such as lines of credit fared well. Most mail campaigns turned a satisfactory profit. Michael, who headed up the analytics behind many of these campaigns, kept a keen eye on the underlying response models and how they could be improved.

Companies often misinterpret marketing campaign results. Here's where they go terribly wrong: They look at the list of customers contacted and ask, "How many responded?" That's the *response rate*. One of the original inventors of uplift modeling, Nicholas Radcliffe (now an independent consultant and sometimes visiting professor in Edinburgh), drew a cartoon about that measure's drawbacks:

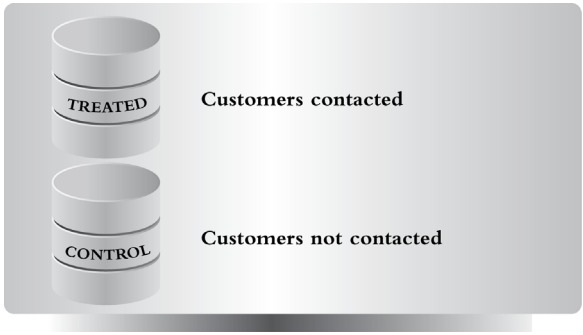
[5] While *prescriptive analytics* might be a suitable synonym for uplift modeling, it is not usually used this way. Ill-defined, it is a problematic term; see the Notes for more.



Cartoon reproduced with permission.

The response rate completely overlooks how many would buy anyway, even if not contacted. Some products just fly off the shelves and sell themselves. For business, that's a good thing—but if so, it's important not to credit the marketing. You could be wasting dollars and chopping down trees to send mail that isn't actually helping.

Just as with medicine, marketing's success—or lack thereof—is revealed by comparing to a control set, a group of individuals suppressed from the treatment (or administered a placebo, in the case of medicine). Therefore, we need to collect two sets of data:



If the treated customers buy more than the control customers, we know the campaign successfully persuades. This proves some individuals were influenced, but, as usual, we don't know which.

PREDICTING THE WRONG THING

If you come to a fork in the road, take it.

—Yogi Berra

To target the marketing campaigns, Michael and his team at U.S. Bank were employing the industry standard: response models, which predict who will

buy if contacted. That's not the same thing as predicting who will buy *because* they were contacted; it does not predict influence. Compared to a control set, Michael showed the campaigns were successful, turning a profit. But he knew the targeting would be more effective if only there were a way to predict which customers would be *persuaded* by the marketing collateral.

Standard response models predict the wrong thing and are in fact falsely named. Response models don't predict response *caused by* contact; they predict buying *in light of* contact. But predicting for whom contact will be the cause of buying is more pertinent than predicting buying in general. Knowing who your "good" customers are—the ones who will buy more—may be nice to know, but it takes second place.[6]

For some projects, conventional response models have it backward. By aiming to increase response rate, they complacently focus on the metric that's easiest to measure. As former U.S. Secretary of Defense Robert McNamara said, "We have to find a way of making the important measurable, instead of making the measurable important." A standard response model will gladly target customers who would buy anyway, doing little to address how much junk mail we as consumers receive. Instead, it's only a small sliver of persuadable customers who are actually worth mailing to, if we can identify them.

Standard response modeling predicts:

1. Will the customer buy if contacted?

Uplift modeling changes everything by adding just one word:

2. Will the customer buy **only** if contacted?

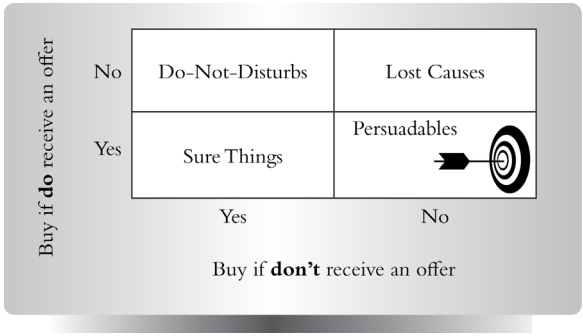
[6] Driving decisions by only predicting the outcome of one treatment without predicting the result of the other is a form of *satisficing*. It's a compromise. Instead of compromising, marketing needs all the help it can get to better target. As a data miner, I actually receive e-mail inquiries from drilling supply vendors. I'm not that kind of miner. Eric King of The Modeling Agency receives job inquiries from (human) models seeking opportunities in the fashion industry.

Although the second question may appear simple, it answers the composite of two questions: "Will the customer buy if contacted and not buy otherwise?" This two-in-one query homes in on the difference that will result from one treatment over another. It's the same as asking, "Would contacting the customer *influence* him or her to buy?"

RESPONSE UPLIFT MODELING

Weigh your options.

By addressing a composite of two questions, each individual belongs in one of four conceptual segments that distinguish along two dimensions:



Conceptual response segments. The lower-right segment is targeted with uplift modeling.[7]

This quad first distinguishes from top to bottom which customers will buy in light of marketing contact, which is the job of conventional response modeling. But then it further distinguishes along a second dimension: Which customers will make a purchase even if not contacted?

[7] Table derived from Nicholas Radcliffe, "Generating Incremental Sales: Maximizing the Incremental Impact of Cross-Selling, Up-Selling and Deep-Selling through Uplift Modeling," Stochastic Solutions Limited, February 16, 2008, and Suresh Vittal, "Optimal Targeting through Uplift Modeling: Generating Higher Demand and Increasing Customer Retention While Reducing Marketing Costs," Forrester Research white paper, 2008.

Michael at U.S. Bank wanted to target the lower-right quadrant, those worthy of investing the cost to contact. These persuadables won't buy if not contacted, but will buy if they are. These are the individuals an uplift model aims to flag with the affirmative prediction.

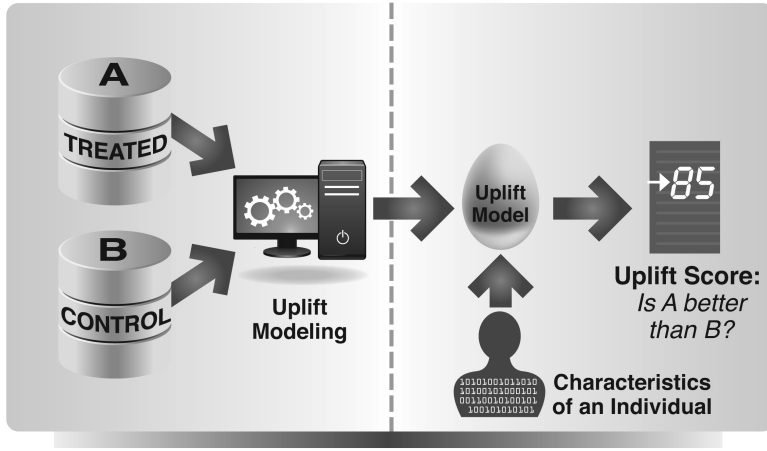
PA APPLICATION: TARGETED MARKETING WITH *RESPONSE UPLIFT MODELING*

1. **What's predicted:** Which customers will be persuaded to buy.
2. **What's done about it:** Target persuadable customers.

An uplift model provides the opportunity to reduce costs and unnecessary mail in comparison to a traditional response model. This is achieved by suppressing from the contact list those customers in the lower-left quadrant, the so-called sure things who will buy either way.

THE MECHANICS OF UPLIFT MODELING

Uplift modeling operates simultaneously on two data sets—both the treated set and the control set—learning from them both:

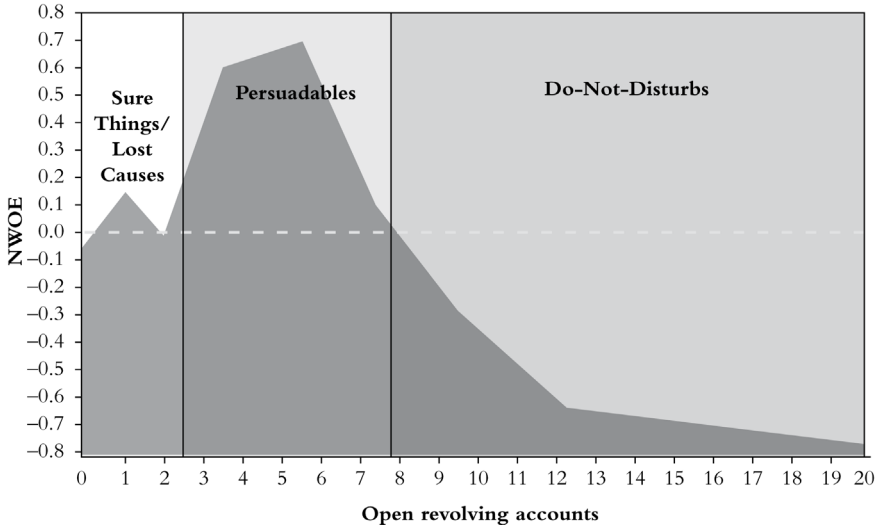


Two training sets are used together to develop an uplift model.

To learn to distinguish influenceables—those for whom the choice of treatment makes a difference—uplift modeling learns from both customers

who were contacted and others who weren't. Processing two data sets represents a significant paradigm shift after decades of predictive modeling and machine learning research almost entirely focused on tweaking a modeling process that operates across a single data set.

Starting first with a single-variable example, we can see that it is possible to predict uplift by comparing behavior across the two data sets:



Net weight of evidence (NWOE, a measure of uplift) varies by a customer's number of open revolving accounts. Graph courtesy of Kim Larsen.

This fictional but typical example of a financial institution's direct-marketing results illustrates that mildly engaged customers are hot, readily persuadable by direct mail. The vertical axis represents *net weight of evidence* (NWOE), a measure of uplift, and the horizontal axis represents the number of open revolving accounts the customer already holds. In this case, it turns out that customers in the middle region, who don't already hold too many or too few open revolving accounts, will be more likely to be persuaded by direct mail.

Less engaged customers on the left are unmoved—whether they were already destined to open more accounts or not, their plans don't change if contacted. This includes both sure things and lost causes—either way, it isn't worth contacting them.

Avoid at all costs contacting customers on the right—they are "do-not-disturbs." Contacting these individuals, who already hold a good number of accounts, actually decreases the chance they'll buy. The curve dips down into negative numbers—a veritable *downlift*. The explanation may be that customers with many accounts are already so engaged that they are more sensitive to, aware of, and annoyed by what they consider to be unnecessary marketing contact. An alternative explanation is that customers who have already accumulated so many credit accounts are susceptible to impulse buys (e.g., when they come into a bank branch), but when contacted at home will be prone to respond by considering the decision more intently and researching competing products online.

This shows the power of one variable. How can we leverage PA's true potential by considering multiple variables, as with the predictive models of Chapter 4? Let's turn back to Michael's story for a detailed example.

HOW UPLIFT MODELING WORKS

Despite their marketing successes, Michael at U.S. Bank had a nagging feeling things could be better. Unlike many marketers, he was aware of the difference between a campaign's response rate and the sales generated by it. Inspecting reports, he could see the response models were less than ideal. He tried out some good ideas of his own to attempt to model persuasion, which provided preliminary yet inconsistent and unstable success.

One time, Michael noted failure for a certain group within a direct mail campaign selling a home-equity line of credit to existing customers. For that group, the campaign not only failed to cover its own printing and mailing costs, it in fact had the detrimental effect of decreasing sales, a slight *downlift* overall.

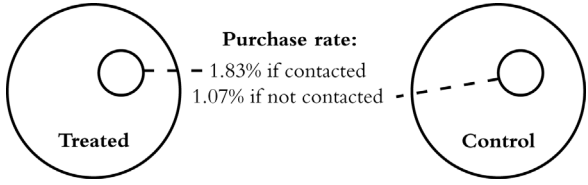
Michael was beginning to collaborate with a small company called Quadstone (now Pitney Bowes Software) that provided a new commercial approach to uplift modeling. The system could derive marketing segments that reveal persuadable customers, such as:[8]

[8] Thanks to Patrick Surry at Pitney Bowes Software for this example segment derived across U.S. Bank data. The segment is simplified for this illustration.

Has paid back more than 17.3% of current loan
-AND-
Is using more than 9.0% of revolving credit limit
-AND-
Is designated within a certain set of lifestyle segments

A segment of persuadable individuals.

This is not your usual marketing segment. It doesn't designate customers more likely to buy. It doesn't designate customers less likely to buy. It is customers *more likely to be influenced by marketing contact.* The difference marketing makes for this segment can be calculated only by seeing how its purchase rate differs between the treated and control sets:[9]



Purchase rates of the persuadable segment described above differ, depending on whether marketing contact is received.

Success! Within this segment, the direct mail elicits more responses from customers who were contacted (the treated set) than those not contacted (the control set). By automatically deriving its defining characteristics, uplift modeling has discovered a segment of customers for which this direct mail campaign succeeds after all.

[9] A simpler alternative to analyzing both sets at once is to make a separate predictive model for each treatment, as was the approach behind the online ad-selection case study described in Chapter 1. Michael at U.S. Bank evaluated this simpler method and concluded that the tree-based approach to uplift modeling provided stronger and more consistent results.

The uplift modeling method that discovers such segments is an expansion of decision trees (see Chapter 4) called *uplift trees*. Normal decision trees strive to identify segments extreme in their response rates—many responses or few responses. Uplift trees use variables to mechanically "segment down" in the same way, but seek to find segments extreme in the difference treatment makes—segments that are particularly influenceable. A single uplift tree is composed of a number of segments such as the one shown above.[10]

For U.S. Bank, response uplift modeling delivered an unprecedented boost, increasing the marketing campaign's return on investment (ROI) by a factor of five in comparison with standard response model targeting. This win resulted from reducing both the amount of direct mail that commanded no impact (sent to lost causes or sure things) and the amount that instigated an adverse response (sent to sleeping dogs, aka do-not-disturbs).

CASE STUDY: U.S. BANK

Business case: Direct mail for a home-equity line of credit to existing customers.
Approach: Target campaign with an uplift model.
Resulting improvements over prior conventional analytical approach:

- Return on investment (ROI) increased five times over previous campaigns (from 75 percent to 400 percent).
- Campaign costs cut by 40 percent.
- Revenue gain increased by over 300 percent.

Uplift practitioners at Fidelity Investments also see the light: *Spend less, earn more*. By avoiding sure things and do-not-disturbs, "Uplift modeling empowers your organization to capture more than 100 percent of responses

[10] Ensemble models (see Chapter 5) of decision trees are recommended when employing this analytical approach to uplift modeling to help ensure stable results. Although predicting influence rather than outright behavior, The Ensemble Effect still applies (as do The Prediction, Data, and Induction Effects).

by contacting less than 100 percent of the target population," says Kathleen Kane, Fidelity's principal decision scientist.

THE PERSUASION EFFECT

Uplift modeling conquers the imperceivable—influence—by newly combining two well-trodden, previously separate paradigms:

1. comparing treated and control results; and
2. predictive modeling (machine learning, statistical regression, etc.).

Only by cleverly combining these two practices does the newfound ability to predict persuasion for each individual become possible. I call this *The Persuasion Effect*:

> **The Persuasion Effect:** *Although imperceptible, the persuasion of an individual can be predicted by* uplift modeling*, predictively modeling across two distinct training data sets that record, respectively, the outcomes of two competing treatments.*

If you haven't already figured it out, this answers the riddle posed at the beginning of this chapter. *Being influenced* is the thing that often happens to you that cannot be witnessed and that you can't even be sure has happened afterward—but that can be predicted in advance. In this way, PA transcends human perception.

INFLUENCE ACROSS INDUSTRIES

Uplift modeling applies everywhere: marketing, credit risk, electoral politics, sociology, and healthcare. The intent to influence is common to almost all organizations, so The Persuasion Effect is put into play across industry sectors.

Application of Uplift Modeling	Treatment Decision	Objective
Targeted marketing with response uplift modeling	*Should we contact the customer or not (active or passive treatment)?*	Positive impact of **direct marketing** campaigns
Customer retention with churn uplift modeling	*Should we provide the customer a retention offer or not (active or passive treatment)?*	Positive impact of **retention campaigns**
Content selection	*With which ad, illustration, choice of words, or product should we solicit the customer?*	Response rate of **direct marketing**, **cross-sell**, and **online and offline ads**
Channel selection	*Through which channel should we contact the customer (e.g., mail, e-mail, or telephone)?*	Positive impact of **direct marketing** campaigns
Dynamic pricing and discounting	*Should we offer the customer a higher price or a lower price?*	Revenue of **sales**
Collections	*Should we offer the debtor a deeper write-off?*	Revenue of **accounts receivable**
Credit risk	*Should we offer the customer a higher or lower credit limit? A higher or lower APR?*	Revenue from **interest payments** and savings from **fewer defaults**
Electoral politics	*Should we market to the constituent/in the state (swing voter/swing state)?*	Positive votes resulting from **political election campaigns** *(see this chapter's sidebar for how Obama's 2012 campaign employed uplift modeling)*
Sociology	*Should we provide benefits to this individual?*	Improved **social program** outcome: long-term self-sufficiency

(*continued*)

(*continued*)

Application of Uplift Modeling	Treatment Decision	Objective
Personalized medicine	*Which medical treatment should the patient receive?*	Favorable patient outcome in **clinical healthcare**

This chapter covers in detail the first two areas on marketing in the table above, as well as a case study in electoral politics (in the sidebar about Obama's 2012 presidential campaign at the end of this chapter). Here's a bit more about the rest of them (note that for some of these application areas, no public case studies or proofs of concept yet exist—uplift modeling is an emerging technology).

Content and channel selection. Uplift modeling selects for each user the ad, offer, content, product, or channel of contact (phone, e-mail, etc.) most likely to elicit a response. In these cases, there is no passive option and therefore no control set—both data sets test an active treatment.

Dynamic pricing and collections. As for any decision, a certain risk is faced for each treatment option when pricing: The higher price may turn a customer away, but the lower price (or deeper discount or write-off, for collections) unnecessarily sacrifices revenue if the customer would have been willing to pay more.

Credit risk. The balance between risk and upside profitability for each debtor is influenced by both the credit limit and the APR offered. Raising one or both may result in higher revenue in the form of interest payments, but may also increase the chance of the debtor defaulting on payments and an ensuing write-off.

Electoral politics. As a resident of California, I see few if any ads for presidential campaigns—the state is a lock; depending on your political affiliation, it could be viewed as either a sure thing or a lost cause. Just as so-called swing clients (influenceables) are potentially persuaded by marketing contact, the same benefit is gained where this term originates: political campaigns that target swing voters. The constituents with the most potential

to be influenced by campaign contact are worth the cost of contact. Analogously, only the swing states that could conceivably be persuaded as a whole are worth expending great campaign resources. For more on elections and uplift modeling, see this chapter's sidebar, "Beyond Swing Voters: How Persuasion Modeling Revolutionized Political Campaigns for Obama and Beyond."

Sociology: targeting social programs. Speaking of politics, here is a concept that could change everything. Social programs such as educational and occupational support endure scrutiny as possibly more frequently awarded to go-getters who would have succeeded anyway. For certain other beneficiaries, skeptics ask, does support backfire, leaving them more dependent rather than more self-sufficient? Only by predicting how a program will influence the outcome for each individual prospect can programs be targeted in a way that addresses these questions. In so doing, *might such scientifically based, individualized economic policies help resolve the crippling government deadlock that results from the opposing fiscal ideologies currently held by conservative and liberal policymakers?*

Personalized medicine. While one medical treatment may deliver better results on average than another, this one-size-fits-all approach commonly implemented by clinical studies means treatment decisions that help many may in fact hurt some. In this way, healthcare decisions backfire on occasion, exerting influence opposite to that intended: They hurt or kill—although they kill fewer than following no clinical studies at all. *Personalized medicine* aims to predict which treatment is best suited for each patient, employing analytical methods to predict treatment impact (i.e., medical influence) similar to the uplift modeling techniques used for marketing treatment decisions. For example, to drive beta-blocker treatment decisions for heart failure, Harvard University researchers "use two independent data sets to construct a systematic, subject-specific treatment selection procedure." A certain HIV treatment is shown to be more effective for younger children. Treatments for various cancers are targeted by genetic markers—a trend so significant the Food and Drug Administration is increasingly requiring for new pharmaceuticals, as *The New York Times* puts it, "a companion test that could reliably detect the [genetic] mutation so that

the drug could be given to the patients it is intended to help," those "most likely to benefit."

IMMOBILIZING MOBILE CUSTOMERS

It wasn't long after phone number portability came, raising a hailstorm in the telecommunications industry, that Quadstone spoke with Eva at Telenor about the new uplift modeling technique. It was a revelation. Eva had already confirmed that Telenor's retention efforts triggered some customers to leave rather than persuading them to stay, but she wasn't aware of any established technique to address the issue. The timing was fortuitous, as Quadstone was just starting out, seeking its first few clients to prove uplift modeling's value.

PA APPLICATION: CUSTOMER RETENTION WITH *CHURN UPLIFT MODELING*

1. **What's predicted:** Which customers can be persuaded to stay.
2. **What's done about it:** Retention efforts target persuadable customers.

Customers can be as easily scared away as a skittish bunny. Traditional churn models often inadvertently frighten these rabbits, since customers most likely to leave are often those most easy to trigger—sleeping dogs easy to wake up. This includes, for example, the health club member who never gets to the gym and the Netflix subscriber who rarely trades in each DVD movie rental—both just need a reminder before they get around to canceling (it would be more ideal to reengage them). Someone once told me that, when he received an offer to extend his camera's warranty, it reminded him that coverage was soon ending. He promptly put his camera into the microwave to break it so he could return it. It would inevitably be more cost-effective to avoid triggering such criminal activity than to prosecute for it after the fact.

Prompting a cell phone customer to leave can be especially costly because it may trigger a social domino effect: People tend to stick with the same wireless carrier as their friends. One major North American carrier showed that a customer is seven times more likely to cancel if someone in the person's calling network cancels.

For Telenor, churn uplift modeling delivered an astonishing boost to the effectiveness of its retention initiatives: The ROI increased by a factor of 11, in comparison with its prior, established practice of targeting with standard churn models. This came from decreasing the number of sleeping dog customers the company had been inadvertently waking, and secondarily from reducing the total number of mail pieces sent—like U.S. Bank, Telenor got more for less.

CASE STUDY: TELENOR, THE WORLD'S SEVENTH-LARGEST MOBILE OPERATOR

Business case: Retention campaign for cell phone subscribers.
Approach: Target campaign with an uplift model.
Resulting improvements over the conventional approach to analytical retention:

- Campaign ROI increased by a factor of 11.
- Churn reduced a further 36 percent.
- Campaign costs reduced by 40 percent.

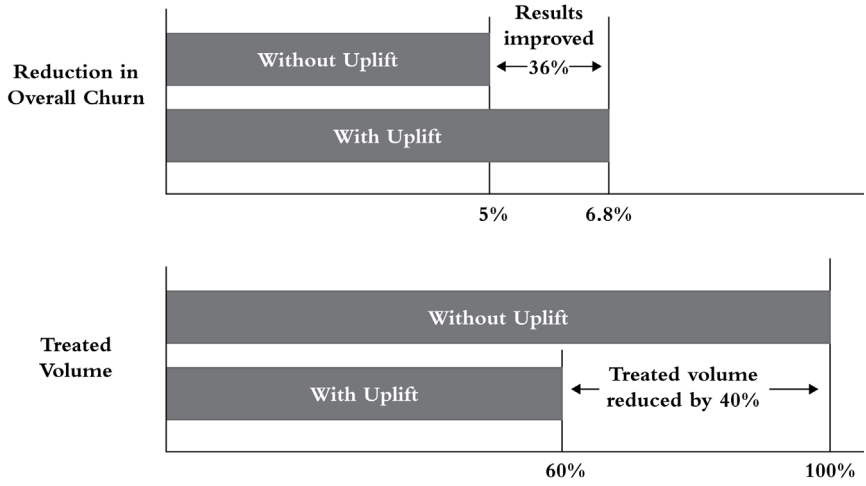


Figure permission of Pitney Bowes Software.

For the international mobile carrier, which serves tens of millions of cell phone subscribers across 11 markets, this was a huge win. Beyond addressing the new business challenges that came of phone number portability, it alleviated the systematic "sleeping dog" problem inherent to churn modeling, one Telenor likely had suffered from all along. Even when there's a net benefit from marketing, offers could be triggering some customers to leave who would have otherwise stayed.

For Eva, who has since been promoted to head of customer analytics, and for the rest of the world, this only marks the beginning of the emerging practice of inducing influence and predicting persuasion.

BEYOND SWING VOTERS

No other presidential campaign [besides Obama's] has relied so heavily on the science of analytics, using information to predict voting patterns. Election day may have changed the game.

—Christi Parsons and Kathleen Hennessey, *Los Angeles Times*, November 13, 2012

Elections hang by a thinner thread than you think.

You may know that President Barack Obama's 2012 campaign for a second term *Moneyballed* the election, employing a team of over 50 analytics experts.

You may also know that the tremendous volume of any presidential campaign's activities, frenetically executed into the eleventh hour in pursuit of landing the world's most powerful job, ultimately serves only to sway a thin slice of the electorate: swing voters within swing states.

But what most people do not realize is that presidential campaigns must focus *even more narrowly than that*, taking microtargeting to a whole new level. The Obama campaign got this one right, breaking ground for election cycles to come by applying uplift modeling to

BEYOND SWING VOTERS (CONTINUED)

drive millions of per-voter campaign decisions, thereby boosting persuasive impact.

However, the buzz in 2012 was about something else. Rather than learning about campaign targeting, when it came to the math behind the election, we heard a great deal about Nate Silver. Silver emerged as the media darling of poll analyzers, soaring past the ranks of guru quant or sexy scientist to become the very face of prediction itself. If mathematical "tomorrowvision" had a name, it was Nate. Even before his forecasts were vindicated by the election results, it was hard to find a talk show host—at least among the left—who hadn't enjoyed a visit from Silver, probing him with evident, slack-jawed fascination.

An election poll does not constitute prognostic technology—it does not endeavor to calculate insights that foresee human behavior. Rather, a poll is plainly the act of voters explicitly telling you what they're going to do. It's a minielection dry run. There's a craft to aggregating polls, as Silver has mastered so adeptly, but even he admits it's no miracle of clairvoyance. "It's not really that complicated," he told late night talk show host Stephen Colbert the day before the election. "There are many things that are *much* more complicated than looking at the polls and taking an average . . . right?"

You want power? *True power comes in influencing the future rather than only speculating on it.* Nate Silver publicly competed to win election forecasting, while Obama's analytics team discreetly competed to win the election itself.

This reflects the very difference between forecasting and predictive analytics (PA). Forecasting calculates an aggregate view for each U.S. state, while PA delivers more detailed insights that drive action: predictions for each individual voter.

THE RARE BIRD: PERSUADABLE VOTERS

Swing voters are a myth. The concept is ill-defined and subjective. In one approach, the Democratic National Committee (DNC) labels as

(*continued*)

BEYOND SWING VOTERS (CONTINUED)

"not very partisan" those voters who self-reported as independent, or for whom their party is (for any reason) unknown. Despite being labeled "swing," many such voters have indeed made up their minds and are unswingable.

Instead of mythical swing voters—or unicorns, for that matter—what matters to a campaign is concrete and yet quite narrow: *Who will be influenced to vote for our candidate by a call, door knock, flier, or TV ad?*

Presidential campaigns must hold themselves to a higher standard than most marketing campaigns. In this unparalleled, ruthless competition of optimal tweaking, the notion of expending resources—such as a paid mailing or a campaign volunteer's precious time—to contact a constituent who was already going to vote your way is abhorrent. Even worse, it is known that, for some cases, contact will inadvertently change the voter's mind in the wrong direction—it can backfire and cause the person to vote for the other candidate.

In the business world, marketing campaigns often withstand such cases without wincing. They inadvertently hit some "sure thing" and "do-not- disturb" customers, yet carry on happily with high profits. As long as the overall campaign is doing more good than harm, taking on the more sophisticated methods needed to smooth these imperfect edges is often seen as too high an investment relative to the expected payoff (although this determination is often just inertia speaking; *uplift modeling* is new and not yet widely practiced).

But a presidential campaign comes along only once every four years. Its extraordinarily high stakes demand that all stops be pulled out. It was only a matter of time before campaigns began predicting their potential to influence each voter in order to optimize that influence.

ANOTHER RARE BIRD: PERSUASION MODELING EXPERTS

Enter Rayid Ghani, chief data scientist of the presidential campaign Obama for America 2012. He was the man for the job. With a master's

BEYOND SWING VOTERS (CONTINUED)

degree in machine learning from Carnegie Mellon (the first university to have a machine learning department), plus 10 years of research work at the labs of consulting behemoth Accenture, Rayid had rare, sought-after experience in uplift modeling—which the campaign called *persuasion modeling*. His background included research determining which medical treatment will provide the best outcome for each patient, and which price will provide the best profit from each retail customer. At Obama for America, he helped determine whether campaign contact would provide the right vote from each constituent.

It's a deep analytical challenge. A predictive model that foresees the ability to persuade is not your average predictive model. Beyond identifying voters who would come out for Obama if contacted, the persuasion models developed by Rayid's staff needed to distinguish those voters who would come out for Obama in any case (the sure things), as well as those who in fact were at risk of being turned off by campaign contact and switching over to vote for the other guy, Mitt Romney (the do-not-disturbs). If you think it through, you'll see the single idea of "can be positively persuaded" actually involves all these distinctions.

PA APPLICATION: POLITICAL CAMPAIGNING WITH VOTER PERSUASION MODELING

1. **What's predicted:** Which voter will be positively persuaded by political campaign contact such as a call, door knock, flier, online ad, or TV ad.
2. **What's done about it:** Persuadable voters are contacted, and voters predicted to be adversely influenced by contact are avoided.

For this project, the campaign needed to collect not donations but data. No matter how smart, the brains on Obama's staff could only tackle the persuasion problem with just the right data sets. To this end, they tested across thousands of voters the very actions they would later

(*continued*)

BEYOND SWING VOTERS (CONTINUED)

decide on for millions. Batches of voters received campaign contact—door knocks, fliers, and phone calls—and, critically, other batches received no contact at all (the control groups). All the batches were then later polled to see whether they would support Obama in the voting booth.

ACTIVELY CAMPAIGNING ON PERSUASION

> *[The Obama campaign job listing for "predictive modeling"] read like politics as done by Martians.*
>
> —Peggy Noonan, *The Wall Street Journal*, July 30, 2011

> *The Martians have landed.*
>
> —Christi Parsons and Kathleen Hennessey, *Los Angeles Times*, November 13, 2012

The data proved that campaigning generally helps, which was good news for the team—but then, analysis had only just begun. Rayid's team faced the ultimate campaign imperative: Learn to discriminate, voter by voter, whether contact would persuade. This is where persuasion modeling (the technique described in the rest of this chapter as *uplift modeling*) came in and took over by storm.

"Our modeling team built persuasion models for each swing state," Rayid said. "The models were then employed to predict the potential to persuade for each of millions of individuals in swing states. It told us which were most likely to be won over to Obama's side, and which we should avoid contacting entirely."* A small group of only three quants led the hands-on execution of persuasion modeling for the campaign.

* Beyond persuasion modeling, the team also employed predictive modeling to gauge the propensity to vote for Obama regardless of campaign contact, the probability of voting at all (turnout), and the probability of donating in order to target fund-raising efforts.

BEYOND SWING VOTERS (CONTINUED)

Persuasion modeling identified nonpartisan voters, according to Director of Statistical Modeling Daniel Porter, one of the three-member unit. Daniel and his colleagues tweaked the models, experimenting extensively across avant-garde techniques designed to identify which factors predict whether a voter is persuadable. The process delivered, pinpointing certain behaviors that seemed to reveal a voter is not strictly partisan, such as supporting Bush in 2004 but Obama in 2008, or being registered as a Democrat in combination with having voted Republican or living in a highly Republican location.

The available data sources were rich. Although campaign staff have not disclosed many other details about the data elements available to discern persuadability, their related effort predicting a constituent's propensity to vote for Obama (regardless of campaign contact) employed more than 80 fields, including demographics, voting history, and magazine subscriptions. The campaign's most cherished data source was the DNC's database, which includes notes regarding each voter's observed response to door knocks—welcoming or door-slamming—during prior presidential election cycles.

The potential persuadability of each voter predicted by these models guided the massive army of campaign volunteers as they pounded the pavement and dialed phone numbers. When knocking on a door, the volunteer wasn't simply canvassing the local neighborhood—this very voter had been predictively targeted as persuadable. As Daniel Wagner, the campaign's chief analytics officer, told the *Los Angeles Times*, "White suburban women? They're not all the same. The Latino community is very diverse with very different interests." This form of microtargeting delved deeper, even bringing volunteers to specific houses within the thick of strongly Republican neighborhoods, and in so doing, moved beyond protocols that had become standard during prior election cycles.

Fliers also targeted the persuadables. As with door knocks, a voter received the flier only if predicted to be influenced, if that voter's mind

(*continued*)

BEYOND SWING VOTERS (CONTINUED)

was likely to be changed. Traditional marketing sends direct mail to those expected to buy *in light of* contact, rather than *because of* it. It's a subtle difference, but all the difference in the world. Putting it another way, rather than determining whether contact is *a good idea*, persuasion modeling determines whether contact is *a better idea* than not contacting.

Persuasion modeling worked. This method was shown to convince more voters to choose Obama than traditional campaign targeting. "These models showed significant lift over just targeting voters who were undecided or had registered as nonpartisan," Rayid said.

This relative boost came in part by avoiding those voters for whom contact was predicted to backfire (the "do-not-disturbs"). As one might expect, for certain voters, campaign contact hurt more than helped.[†] So, during the full-scale efforts ultimately guided by the persuasion models, many such voters were predictively identified and shrewdly left uncontacted.

Persuasion modeling also targeted the campaign's TV ad buying, which delivered a dramatic improvement. A TV spot—such as Fox News in Tampa during evening hours—sells its ad slots by providing a demographic breakdown of its viewers. Team Obama viewed these breakdowns through the filter of their persuasion models in order to decide which spots to hit. Their postcampaign analysis showed this made the TV ad buy 18 percent more effective—they could persuade 18 percent more voters with the same level of investment, which is a meaningful effect given the TV budget magnitude with which they were working: $400 million.

[†] Even during the analysis of results collected from campaign testing, this is not self-evident from the data, since no individual voter could be both contacted and not contacted to determine which would lead to a better outcome. Detecting the influence of campaign contact, be it positive influence or negative influence, requires modeling, even retrospectively.

BEYOND SWING VOTERS (CONTINUED)

Unsurprisingly, 2016 presidential campaigns are gearing up to apply persuasion modeling. The specifics are well-guarded secrets, but the trend is undeniable. Even as early as July 2015, Hillary Clinton's "analytics team is looking for data nerds," said her campaign website. Shown as one of 11 campaign job categories, analytics included five types of open roles. Analytics job postings for the campaign on relevant industry portals enlisted staff for "helping the campaign determine which voters to target for persuasion." Bernie Sanders' campaign website included "Director of Data and Analytics" as one of only five posted job listings.

Years after the 2012 election, Daniel Porter's perspective hasn't changed. "It remains clear that persuasion modeling is extraordinarily valuable for political campaigns. In fact, after the experience accrued last time around, it's sure to be done by 2016 campaigns even more effectively than in 2012." There's also going to be better data for this work, at least on the Democratic side. "The DNC is building out further its data infrastructure about voters in battleground states."

It's advanced, it's analytical, but it's not arcane. Persuasion modeling is the final chapter of this book, and has begun a whole new chapter for politics.