CHAPTER 1

Welcome to the Future

The shovel is a tool, and so is a bulldozer. Neither works on its own, "automating" the task of digging. But both tools augment our ability to dig.

Dr. Douglas Engelbart, "Improving Our Ability to Improve"[1]

Marketing is about to get weird. We've become used to an ever-increasing rate of change. But occasionally, we have to catch our breath, take a new sighting, and reset our course.

Between the time my grandfather was born in 1899 and his seventh birthday:

- Theodore Roosevelt took over as president from William McKinley.
- Dr. Henry A. Rowland of Johns Hopkins University announced a theory about the cause of the Earth's magnetism.
- L. Frank Baum's *The Wonderful Wizard of Oz* was published in Chicago.
- The first zeppelin flight was carried out over Lake Constance near Friedrichshafen, Germany.
- Karl Landsteiner developed a system of blood typing.

- The Ford Motor Company produced its first car—the Ford Model A.
- Thomas Edison invented the nickel-alkaline storage battery.
- The first electric typewriter was invented by George Canfield Blickensderfer of Erie, Pennsylvania.
- The first radio that successfully received a radio transmission was developed by Guglielmo Marconi.
- The Wright brothers flew at Kitty Hawk.
- The Panama Canal was under construction.
- Benjamin Holt invented one of the first practical continuous tracks for use in tractors and tanks.
- The Victor Talking Machine Company released the Victrola.
- The Autochrome Lumière, patented in 1903, became the first commercial color photography process.

My grandfather then lived to see men walk on the moon.

In the next few decades, we will see:

- Self-driving cars replace personally owned transportation.
- Doctors routinely operate remote, robotic surgery devices.
- Implantable communication devices replace mobile phones.
- In-eye augmented reality become normalized.
- Maglev elevators travel sideways and transform building shapes.
- Every surface consume light for energy and act as a display.
- Mind-controlled prosthetics with tactile skin interfaces become mainstream.
- Quantum computing make today’s systems microscopic.
- 3-D printers allow for instant delivery of goods.
- Style-selective, nanotech clothing continuously clean itself.

And today’s youngsters will live to see a colony on Mars.

It’s no surprise that computational systems will manage more tasks in advertising and marketing. Yes, we have lots of technology for marketing, but the next step into artificial intelligence and machine learning will be different. Rather than being an ever-larger confusion of rules-based programs, operating faster than the eye can see, AI systems will operate more inscrutably than the human mind can fathom.

WELCOME TO AUTONOMIC MARKETING

The autonomic nervous system controls everything you don't have to think about: your heart, your breathing, your digestion. All of these things can happen while you're asleep or unconscious. These tasks are complex, interrelated, and vital. They are so necessary they must function continuously without the need for deliberate thought.

That's where marketing is headed. We are on the verge of the need for autonomic responses just to stay afloat. Personalization, recommendations, dynamic content selection, and dynamic display styles are all going to be table stakes.

The technologies seeing the light of day in the second decade of the twenty-first century will be made available as services and any company *not* using them will suffer the same fate as those that decided not to avail themselves of word processing, database management, or Internet marketing. And so, it's time to open up that black box full of mumbo-jumbo called artificial intelligence and understand it just well enough to make the most of it for marketing. Ignorance is no excuse. You should be comfortable enough with artificial intelligence to put it to practical use without having to get a degree in data science.

WELCOME TO ARTIFICIAL INTELLIGENCE FOR MARKETERS

> *It is of the highest importance in the art of detection to be able to recognize, out of a number of facts, which are incidental and which vital.*
>
> *Sherlock Holmes*, The Reigate Squires

This book looks at some current buzzwords to make just enough sense for regular marketing folk to understand what's going on.

- This is no deep exposé on the dark arts of artificial intelligence.
- This is no textbook for learning a new type of programming.
- This is no exhaustive catalog of cutting-edge technologies.

This book is not for those with advanced math degrees or those who wish to become data scientists. If, however, you are inspired to delve into the bottomless realm of modern systems building, I'll point you to "How to Get the Best Deep Learning Education for Free"[2] and be happy to take the credit for inspiring you. But that is not my intent.

You *will not* find passages like the following in this book:

> Monte-Carlo simulations are used in many contexts: to produce high quality pseudo-random numbers, in complex settings such as multi-layer spatio-temporal hierarchical Bayesian models, to estimate parameters, to compute statistics associated with very rare events, or even to generate large amount of data (for instance cross and auto-correlated time series) to test and compare various algorithms, especially for stock trading or in engineering.
>
> **"24 Uses of Statistical Modeling" (Part II)**[3]

You *will* find explanations such as: Artificial intelligence is valuable because it was designed to deal in gray areas rather than crank out statistical charts and graphs. It is capable, over time, of understanding context.

The purpose of this tome is to be a primer, an introduction, a statement of understanding for those who have regular jobs in marketing—and would like to keep them in the foreseeable future.

Let's start with a super-simple comparison between artificial intelligence and machine learning from Avinash Kaushik, digital marketing evangelist at Google: "AI is an *intelligent machine* and ML is the *ability to learn without being explicitly programmed*."

Artificial intelligence is a machine pretending to be a human. Machine learning is a machine pretending to be a statistical programmer. Managing either one requires a data scientist.

An ever-so-slightly deeper definition comes from E. Fredkin University professor at the Carnegie Mellon University Tom Mitchell:[4]

> The field of Machine Learning seeks to answer the question, "How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?"
>
> A machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E. Depending on how we specify T, P, and E, the learning task might also be called by names such as data mining, autonomous discovery, database updating, programming by example, etc.

Machine learning is a computer's way of using a given data set to figure out how to perform a specific function through trial and error.

What is a specific function? A simple example is deciding the best e-mail subject line for people who used certain search terms to find your website, their behavior on your website, and their subsequent responses (or lack thereof) to your e-mails.

The machine looks at previous results, formulates a conclusion, and then waits for the results of a test of its hypothesis. The machine next consumes those test results and updates its weighting factors from which it suggests alternative subject lines—over and over.

There is no final answer because reality is messy and ever changing. So, just like humans, the machine is always accepting new input to formulate its judgments. It's learning.

The "three *D*s" of artificial intelligence are that it can *detect*, *decide*, and *develop*.

Detect

AI can discover which elements or attributes in a subject matter domain are the most predictive. Even with a great deal of noisy data and a large variety of data types, it can identify the most revealing characteristics, figuring out which to heed to and which to ignore.

Decide

AI can infer rules about data, from the data, and weigh the most predictive attributes against each other to make a decision. It can take an enormous number of characteristics into consideration, ponder the relevance of each, and reach a conclusion.

Develop

AI can grow and mature with each iteration. Whether it is considering new information or the results of experimentation, it can alter its opinion about the environment as well as how it evaluates that environment. It can program itself.

WHOM IS THIS BOOK FOR?

This is the sort of book data scientists should buy for their marketing colleagues to help them understand what goes on in the data science department.

This is the sort of book marketing professionals should buy for their data scientists to help them understand what goes on in the marketing department.

This book is for the marketing manager who has to respond to the C-level insistence that the marketing department "get with the times" (management by *in-flight* magazine).

This book is for the marketing manager who has finally become comfortable with analytics as a concept, and learned how to become a dexterous consumer of analytics outputs, but must now face a new educational learning curve.

This book is for the rest of us who need to understand the big, broad brushstrokes of this new type of data processing in order to understand where we are headed in business.

This book is for those of us who need to survive even though we are not data scientists, algorithm magicians, or predictive analytics statisticians.

We must get a firm grasp on artificial intelligence because it will be our jobs to make use of it in ways that raise revenue, lower costs, increase customer satisfaction, and improve organizational capabilities.

THE BRIGHT, BRIGHT FUTURE

Artificial intelligence will give you the ability to match information about your product with the information your prospective buyers need at the moment and in a format they are most likely to consume it most effectively.

I came across my first seemingly self-learning computer system when I was selling Apple II computers in a retail store in Santa Barbara in 1980. Since then, I've been fascinated by how computers can be useful in life and work. I was so interested, in fact, that I ended up explaining (and selling) computers to companies that had never had one before, and programming tools to software engineers, and consulting to the world's largest corporations on how to improve their digital relationships with customers through analytics.

Machine learning offers so much power and so much opportunity that we're in the same place we were with personal computers in 1980, the Internet in 1993, and e-commerce when Amazon.com began taking over e-commerce.

In each case, the promise was enormous and the possibilities were endless. Those who understood the impact could take advantage of it before their competitors. But the advantage was fuzzy, the implications were diverse, and speculations were off the chart.

The same is true of AI today. We know it's powerful and we know it's going to open doors we had not anticipated. There are current examples of marketing departments experimenting with some good and some not-so-good outcomes, but the promise remains enormous.

In advertising, machine learning works overtime to get the right message to the right person at the right time. The machine folds response rates back into the algorithm, not just the database. In the realm of customer experience, machine learning rapidly produces and takes action on new data-driven insights, which then act as new input for the next iteration of its models. Businesses use the results to delight customers, anticipate needs, and achieve competitive advantage.

Consider the telecommunications company that uses automation to respond to customer service requests quicker or the bank that uses data on past activity to serve up more timely and relevant offers to customers through e-mail or the retail company that uses beacon technology to engage its most loyal shoppers in the store.

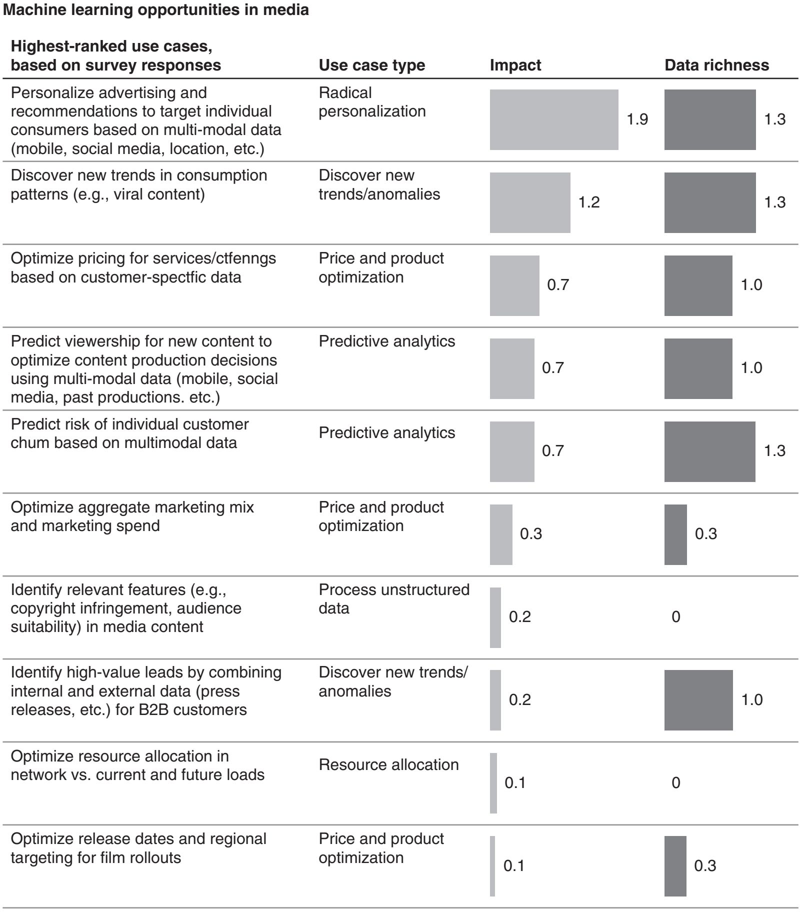
Don't forget media companies using machine learning to track customer preference data to analyze viewing history and present personalized content recommendations. In "The Age of Analytics: Competing in a Data-Driven World,"[5] McKinsey Global Institute studied the areas in a dozen industries that were ripe for disruption by AI. Media was one of them. (See Figure 1.1.)[6]

IS AI SO GREAT IF IT'S SO EXPENSIVE?

As you are an astute businessperson, you are asking whether the investment is worth the effort. After all, this is experimental stuff and Google is *still* trying to teach a car how to drive itself.

Christopher Berry, Director of Product Intelligence for the Canadian Broadcasting Corporation, puts the business spin on this question.[7]

> Look at the progress that Google has made in terms of its self-driving car technology. They invested years and years and years in computer vision, and then training machines to respond to road conditions. Then look at the way that Tesla has been able to completely catch up by way of watching its drivers just use the car.
>
> The emotional reaction that a data scientist is going to have is, "I'm building machine to be *better* than a human being. Why would I want to bring a machine up to the point of it being as *bad* as a human being?"



Machine learning opportunities in media

Highest-ranked use cases, based on survey responses	Use case type	Impact	Data richness
Personalize advertising and recommendations to target individual consumers based on multi-modal data (mobile, social media, location, etc.)	Radical personalization	1.9	1.3
Discover new trends in consumption patterns (e.g., viral content)	Discover new trends/anomalies	1.2	1.3
Optimize pricing for services/ctfenngs based on customer-spectfic data	Price and product optimization	0.7	1.0
Predict viewership for new content to optimize content production decisions using multi-modal data (mobile, social media, past productions. etc.)	Predictive analytics	0.7	1.0
Predict risk of individual customer chum based on multimodal data	Predictive analytics	0.7	1.3
Optimize aggregate marketing mix and marketing spend	Price and product optimization	0.3	0.3
Identify relevant features (e.g., copyright infringement, audience suitability) in media content	Process unstructured data	0.2	0
Identify high-value leads by combining internal and external data (press releases, etc.) for B2B customers	Discover new trends/ anomalies	0.2	1.0
Optimize resource allocation in network vs. current and future loads	Resource allocation	0.1	0
Optimize release dates and regional targeting for film rollouts	Price and product optimization	0.1	0.3

Figure 1.1 A McKinsey survey finds advertising and marketing highly ranked for disruption.

> The commercial answer is that if you can train a generic Machine Learning algorithm well enough to do a job as poorly as a human being, it's still better than hiring an expensive human being because every single time that machine runs, you don't have to pay its pension, you don't have to pay its salary, and it doesn't walk out the door and maybe go off to a competitor.
>
> And there's a possibility that it could surpass a human intelligence. If you follow that argument all the way

> through, narrow machine intelligence is good enough for problem subsets that are incredibly routine.
>
> We have so many companies that are dedicated to marketing automation and to smart agents and smart bots. If we were to enumerate all the jobs being done in marketing department and score them based on how much pain caused, and how esteemed they are, you'd have no shortage of start-ups trying to provide the next wave of mechanization in the age of information.

And heaven knows, we have plenty of well-paid people spending a great deal of time doing incredibly routine work.

So machine learning is great. It's powerful. It's the future of marketing. But just what the heck *is* it?

WHAT'S ALL THIS AI THEN?

What are AI, cognitive computing, and machine learning? In "The History of Artificial Intelligence,"[8] Chris Smith introduces AI this way:

> The term *artificial intelligence* was first coined by John McCarthy in 1956 when he held the first academic conference on the subject. But the journey to understand if machines can truly think began much before that. In Vannevar Bush's seminal work *As We May Think* (1945) he proposed a system which amplifies people's own knowledge and understanding. Five years later Alan Turing wrote a paper on the notion of machines being able to simulate human beings and the ability to do intelligent things, such as play Chess (1950).

In brief—AI mimics humans, while machine learning is a system that can figure out how to figure out a specific task. According to SAS, multinational developer of analytics software, "Cognitive computing is based on self-learning systems that use machine-learning techniques to perform specific, humanlike tasks in an intelligent way."[9]

THE AI UMBRELLA

We start with *AI*, *artificial intelligence*, as it is the overarching term for a variety of technologies. AI generally refers to making computers act like people. "Weak AI" is that which can do something very specific, very well, and “strong AI” is that which thinks like humans, draws on general knowledge, imitates common sense, threatens to become self-aware, and takes over the world.

We have lived with weak AI for a while now. Pandora is very good at choosing what music you might like based on the sort of music you liked before. Amazon is pretty good at guessing that if you bought *this,* you might like to buy *that*. Google’s AlphaGo beat Go world champion Lee Sedol in March 2016. Another AI system (DeepStack) beat experts at no-limit, Texas Hold’em Poker.[10] But none of those systems can do anything else. They are *weak*.

Artificial intelligence is a large umbrella. Under it, you’ll find visual recognition (“That’s a cat!”), voice recognition (you can say things like, “It won’t turn on” or “It won’t connect to the Internet” or “It never arrived”), natural language processing (“I think you said you wanted me to open the garage door and warm up your car. Is that right?”), expert systems (“Based on its behavior, I am 98.3% confident that is a cat”), affective computing (“I see cats make you happy”), and robotics (I’m acting like a cat).

THE MACHINE THAT LEARNS

The magic of machine learning is that it was designed to learn, not to follow strict rules. This is the most fundamental aspect to understand and the most important to remember when you hit that inevitable frustration when things start going slightly off-track. A rules-based system does exactly what it’s told and nothing more. We are comforted by that. A command to send out a gazillion e-mails with the “<first_name>” after the salutation does precisely that. That’s good.

Of course, when the database has something fishy in the first_name field, then somebody gets an e-mail that begins, “Hello, Null, how are you?” Once humans know to look for those sorts of mistakes, we create processes to check and correct the data before hitting Send the next time.

When a batch of e-mails goes out that *all* say, “Hello, <first_name>, how are you?” and the e-mails all include those brackets and that underline, we know to flail the programmers until they find the errant semicolon that caused the problem. In both cases, we can backtrack, find the problem, and fix it.

Machine learning has data and is given a task. It performs the task and waits for feedback. If the feedback is positive, it tries to duplicate the same actions the next time for a similar task. If the feedback is negative, it requires information on where it went wrong.

Guess the Animal

In 1980, at the Personal Electronics store in Santa Barbara, California, I was shown a program that could guess the animal you had in mind. I was flabbergasted. It was a simple case of Asimov's "Any sufficiently advanced technology is indistinguishable from magic."

It: Think of an animal . . .

Is it an elephant? (y or n)

Me: n

It: You win. Help me learn from my mistake before you go . . . What animal were you thinking of?

Me: a rabbit

It: Give me a question to distinguish a rabbit from an elephant.

Me: Is it a small animal?

It: For a rabbit, what is the answer to your question? (y or n)

Me: y

It: Thanks.

Play again? (y or n)

Me: y

It: Think of an animal . . .

Is it a small animal? (y or n)

Me: y

It: Is it a rabbit? (y or n)

Me: n

It: You win. Help me learn from my mistake before you go . . . What animal were you thinking of?

Me: a frog

It: Give me a question to distinguish a frog from rabbit and an elephant.

Me: Is it green?

It: Thanks.

Play again? (y or n)

Me: No, I see how this works now.

It: I don't understand.

Play again? (y or n)

Me: n

It: Goodbye

Is this program rudimentary? You bet. Was this machine learning? Almost.

After running again and again, the game could guess exactly what animal you had in mind after only a few questions. It was impressive, but it was just following programmed logic. It was not *learning*. Guess the Animal could update its rules-based database and *appear* to be getting smarter as it went along, but it did not change how it made decisions.

The Machine that Programs Itself

Machine learning systems look for patterns and try to make sense of them. It all starts with the question: What problem are you trying to solve?

Let's say you want the machine to recognize a picture of a cat. Feed it all the pictures of cats you can get your hands on and tell it, "These are cats." The machine looks through all of them, looking for patterns. It sees that cats have fur, pointy ears, tails, and so on, and waits for you to ask a question.

"How many paws does a cat have?"

"On average, 3.24."

That's a good, solid answer from a regular database. It looks at all the photos, adds up the paws, and divides by the number of pictures.

But a machine learning system is designed to learn. When you tell the machine that most cats have four paws, it can "realize" that it cannot see all of the paws. So when you ask,

"How many ears does a cat have?"

"No more than two."

the machine has learned something from its experience with paws and can apply that learning to counting ears.

The magic of machine learning is building systems that build themselves. We teach the machine to learn how to learn. We build systems that can write their own algorithms, their own architecture. Rather than learn more information, they are able to change their minds about the data they acquire. They alter the way they perceive. They learn.

The code is unreadable to humans. The machine writes its own code. You can't fix it; you can only try to correct its behavior.

It's troublesome that we cannot backtrack and find out where a machine learning system went off the rails if things come out wrong. That makes us decidedly uncomfortable. It is also likely to be illegal, especially in Europe.

"The EU General Data Protection Regulation (GDPR) is the most important change in data privacy regulation in 20 years" says the homepage of the EU GDPR Portal.[11] Article 5, Principles Relating to Personal Data Processing, starts right out with:

> Personal Data must be:
>
> * processed lawfully, fairly, and in a manner transparent to the data subject
>
> * collected for specified, explicit purposes and only those purposes
>
> * limited to the minimum amount of personal data necessary for a given situation
>
> * accurate and where necessary, up to date
>
> * kept in a form that permits identification of the data subject for only as long as is necessary, with the only exceptions being statistical or scientific research purposes pursuant to article 83a
>
> * Parliament adds that the data must be processed in a manner allowing the data subject to exercise his/her rights and protects the integrity of the data
>
> * Council adds that the data must be processed in a manner that ensures the security of the data processed under the responsibility and liability of the data controller

Imagine sitting in a bolted-to-the-floor chair in a small room at a heavily scarred table with a single, bright spotlight overhead and a detective leaning in asking, "So how did your system screw this up so badly and how are you going to fix it? Show me the decision-making process!"

This is a murky area at the moment, and one that is being reviewed and pursued. Machine learning systems will have to come with tools that allow a decision to be explored and explained.

ARE WE THERE YET?

Most of this sounds a little over-the-horizon and science-fiction-ish, and it is. But it's only *just* over the horizon. (Quick—check the publication date at the front of this book!) The capabilities have been in the lab for a while now. Examples are in the field. AI and machine learning are being used in advertising, marketing, and customer service, and they don't seem to be slowing down.

But there are some projections that this is all coming at an alarming rate.[12]

> According to researcher Gartner, AI bots will power 85% of all customer service interactions by the year 2020. Given Facebook and other messaging platforms have already seen significant adoption of customer service bots on their chat apps, this shouldn't necessarily come as a huge surprise. Since this use of AI can help reduce wait times for many types of interactions, this trend sounds like a win for businesses and customers alike.

The White House says it's time to get ready. In a report called "Preparing for the Future of Artificial Intelligence" (October 2016),[13] the Executive Office of the President National Science and Technology Council Committee on Technology said:

> The current wave of progress and enthusiasm for AI began around 2010, driven by three factors that built upon each other: the availability of big data from sources including e-commerce, businesses, social media, science, and government; which provided raw material for dramatically improved Machine Learning approaches and algorithms; which in turn relied on the capabilities of more powerful computers. During this period, the pace of improvement surprised AI experts. For example, on a popular image recognition challenge[14] that has a 5 percent human error rate according to one error measure, the best AI result improved from a 26 percent error rate in 2011 to 3.5 percent in 2015.

> Simultaneously, industry has been increasing its investment in AI. In 2016, Google Chief Executive Officer (CEO) Sundar Pichai said, "Machine Learning [a subfield of AI] is a core, transformative way by which we're rethinking how we're doing everything. We are thoughtfully applying it across all our products, be it search, ads, YouTube, or Play. And we're in early days, but you will see us—in a systematic way—apply Machine Learning in all these areas." This view of AI broadly impacting how software is created and delivered was widely shared by CEOs in the technology industry, including Ginni Rometty of IBM, who has said that her organization is betting the company on AI.

The commercial growth in AI is surprising to those of little faith and not at all surprising to true believers. IDC Research "predicts that spending on AI software for marketing and related function businesses will grow at an exceptionally fast cumulative average growth rate (CAGR) of 54 percent worldwide, from around $360 million in 2016 to over $2 billion in 2020, due to the attractiveness of this technology to both sell-side suppliers and buy-side end-user customers."[15]

Best to be prepared for the "ketchup effect," as Mattias Östmar called it: "First nothing, then nothing, then a drip and then all of a sudden—splash!"

You might call it hype, crystal-balling, or wishful thinking, but the best minds of our time are taking it very seriously. The White House's primary recommendation from the above report is to "examine whether and how (private and public institutions) can responsibly leverage AI and Machine Learning in ways that will benefit society."

Can you responsibly leverage AI and machine learning in ways that will benefit society? What happens if you don't? What could possibly go wrong?

AI-POCALYPSE

> *Cyberdyne will become the largest supplier of military computer systems. All stealth bombers are upgraded with Cyberdyne computers, becoming fully unmanned. Afterwards, they fly with a perfect operational record. The Skynet Funding Bill is passed. The system goes online August 4th, 1997. Human decisions are removed from*

> *strategic defense. Skynet begins to learn at a geometric rate. It becomes self-aware at 2:14 a.m. Eastern time, August 29th. In a panic, they try to pull the plug.*
>
> The Terminator, *Orion Pictures, 1984*

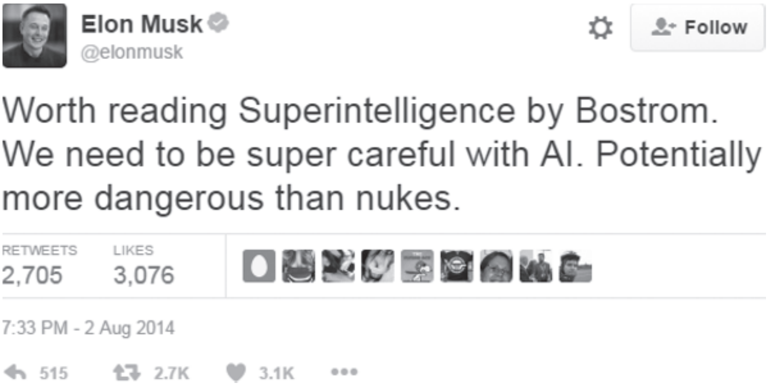
At the end of 2014, Professor Stephen Hawking rattled the data science world when he warned, "The development of full artificial intelligence could spell the end of the human race . . . . It would take off on its own, and re-design itself at an ever increasing rate. Humans, who are limited by slow biological evolution, couldn't compete and would be superseded."[16]

In August 2014, Elon Musk took to Twitter to express his misgivings:

"Worth reading Superintelligence by Bostrom. We need to be super careful with AI. Potentially more dangerous than nukes," (Figure 1.2) and "Hope we're not just the biological boot loader for digital superintelligence. Unfortunately, that is increasingly probable."

In a clip from the movie *Lo and Behold*, by German filmmaker Werner Herzog, Musk says:

> I think that the biggest risk is not that the AI will develop a will of its own, but rather that it will follow the will of people that establish its utility function. If it is not well thought out—even if its intent is benign—it could have quite a bad outcome. If you were a hedge fund or private equity fund and you said, "Well, all I want my AI to do is



Elon Musk
@elonmusk

Worth reading Superintelligence by Bostrom. We need to be super careful with AI. Potentially more dangerous than nukes.

RETWEETS 2,705 LIKES 3,076

7:33 PM - 2 Aug 2014

Figure 1.2 Elon Musk expresses his disquiet on Twitter.

> maximize the value of my portfolio," then the AI could decide, well, the best way to do that is to short consumer stocks, go long defense stocks, and start a war. That would obviously be quite bad.

While Hawking is thinking big, Musk raises the quintessential Paperclip Maximizer Problem and the Intentional Consequences Problem.

The AI that Ate the Earth

Say you build an AI system with a goal of maximizing the number of paperclips it has. The threat is that it learns how to find paperclips, buy paperclips (requiring it to learn how to make money), and then work out how to manufacture paperclips. It would realize that it needs to be smarter, and so increases its own intelligence in order to make it even smarter, in service of making paperclips.

What is the problem? A hyper-intelligent agent could figure out how to use nanotech and quantum physics to alter all atoms on Earth into paperclips.

Whoops, somebody seems to have forgotten to include the Three Laws of Robotics from Isaac Asimov's 1950 book, *I Robot*:

1. A robot may not injure a human being, or through inaction, allow a human being to come to harm.
2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Max Tegmark, president of the Future of Life Institute, ponders what would happen if an AI

> is programmed to do something beneficial, but it develops a destructive method for achieving its goal: This can happen whenever we fail to fully align the AI's goals with ours, which is strikingly difficult. If you ask an obedient intelligent car to take you to the airport as fast as possible, it might get you there chased by helicopters and covered in vomit, doing not what you wanted but literally what you asked for. If a superintelligent system is tasked with a(n) ambitious geoengineering project, it might wreak havoc with our ecosystem as a side effect, and view human attempts to stop it as a threat to be met.[17]

If you really want to dive into a dark hole of the existential problem that AI represents, take a gander at "The AI Revolution: Our Immortality or Extinction."[18]

Intentional Consequences Problem

Bad guys are the scariest thing about guns, nuclear weapons, hacking, and, yes, AI. Dictators and authoritarian regimes, people with a grudge, and people who are mentally unstable could all use very powerful software to wreak havoc on our self-driving cars, dams, water systems, and air traffic control systems. That would, to repeat Mr. Musk, obviously be quite bad.

That's why the Future of Life Institute offered "Autonomous Weapons: An Open Letter from AI & Robotics Researchers," which concludes, "Starting a military AI arms race is a bad idea, and should be prevented by a ban on offensive autonomous weapons beyond meaningful human control."[19]

In his 2015 presentation on "The Long-Term Future of (Artificial) Intelligence," University of California, Berkeley professor Stuart Russell asked, "What's so bad about the better AI? AI that is incredibly good at achieving something other than what we *really* want."

Russell then offered some approaches to managing the it's-smarter-than-we-are conundrum. He described AIs that are not in control of anything in the world, but only answer a human's questions, making us wonder whether it could learn to manipulate the human. He suggested creating an agent whose only job is to review other AIs to see if they are potentially dangerous and admitted that was a bit of a paradox. He's very optimistic, however, given the economic incentive for humans to create AI systems that do *not* run amok and turn people into paperclips. The result will inevitably be the development of community standards and a global regulatory framework.

Setting aside science fiction fears of the unknown and a madman with a suitcase nuke, there are some issues that are real and deserve our attention.

Unintended Consequences

The biggest legitimate concern facing marketing executives when it comes to machine learning and AI is when the machine does what you tell it to do rather than what you wanted it to do. This is much like the paperclip problem, but much more subtle. In broad terms, this is known as the *alignment problem*. The alignment problem wonders how to explain to an AI system goals that are not absolute, but take all of human values into consideration, especially considering that values vary widely from human to human, even in the same community. And even then, humans, according to Professor Russell, are irrational, inconsistent, and weak-willed.

The good news is that addressing this issue is actively happening at the industrial level. "OpenAI is a non-profit artificial intelligence research company. Our mission is to build safe AI, and ensure AI's benefits are as widely and evenly distributed as possible."[20]

The other good news is that addressing this issue is actively happening at the academic/scientific level. The Future of Humanity Institute teamed with Google to publish a paper titled "Safely Interruptible Agents."[21]

> Reinforcement learning agents interacting with a complex environment like the real world are unlikely to behave optimally all the time. If such an agent is operating in real-time under human supervision, now and then it may be necessary for a human operator to press the big red button to prevent the agent from continuing a harmful sequence of actions—harmful either for the agent or for the environment—and lead the agent into a safer situation. However, if the learning agent expects to receive rewards from this sequence, it may learn in the long run to avoid such interruptions, for example by disabling the red button—which is an undesirable outcome. This paper explores a way to make sure a learning agent will not learn to prevent (or seek!) being interrupted by the environment or a human operator. We provide a formal definition of safe interruptibility and exploit the off-policy learning property to prove that either some agents are already safely interruptible, like Q-learning, or can easily be made so, like Sarsa. We show that even ideal, uncomputable reinforcement learning agents for (deterministic) general computable environments can be made safely interruptible.

There is also the Partnership on Artificial Intelligence to Benefit People and Society,[22] which was "established to study and formulate best practices on AI technologies, to advance the public's understanding of AI, and to serve as an open platform for discussion and engagement about AI and its influences on people and society."

Granted, one of its main goals from an industrial perspective is to calm the fears of the masses, but it also intends to "support research and recommend best practices in areas including ethics, fairness, and inclusivity; transparency and interoperability; privacy; collaboration between people and AI systems; and of the trustworthiness, reliability, and robustness of the technology."

The Partnership on AI's stated tenets[23] include:

> We are committed to open research and dialog on the ethical, social, economic, and legal implications of AI.
>
> We will work to maximize the benefits and address the potential challenges of AI technologies, by:
>
> > Working to protect the privacy and security of individuals.
> >
> > Striving to understand and respect the interests of all parties that may be impacted by AI advances.
> >
> > Working to ensure that AI research and engineering communities remain socially responsible, sensitive, and engaged directly with the potential influences of AI technologies on wider society.
> >
> > Ensuring that AI research and technology is robust, reliable, trustworthy, and operates within secure constraints.
> >
> > Opposing development and use of AI technologies that would violate international conventions or human rights, and promoting safeguards and technologies that do no harm.

That's somewhat comforting, but the blood pressure lowers considerably when we notice that the Partnership includes the American Civil Liberties Union. That makes it a little more socially reliable than the Self-Driving Coalition for Safer Streets, which is made up of Ford, Google, Lyft, Uber, and Volvo without any representation from little old ladies who are just trying to get to the other side.

Will a Robot Take Your Job?

Just as automation and robotics have displaced myriad laborers and word processing has done away with legions of secretaries, some jobs will be going away.

The *Wall Street Journal* article, "The World's Largest Hedge Fund Is Building an Algorithmic Model from Its Employees' Brains,"[24] reported on $160 billion Bridgewater Associates trying to embed its founder's approach to management into a so-called Principles Operating System. The system is intended to study employee reviews and testing to delegate specific tasks to specific employees along with detailed instructions, not to mention having a hand in hiring, firing, and promotions. Whether a system that thinks about humans as complex machines can succeed will take some time.

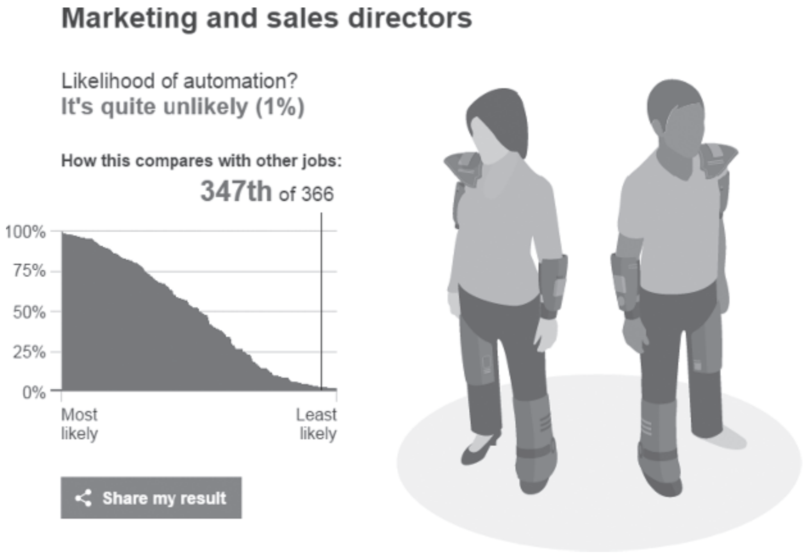
A *Guardian* article sporting the headline "Japanese Company Replaces Office Workers with Artificial Intelligence"[25] reported on an insurance company at which 34 employees were to be replaced in March 2017 by an AI system that calculates policyholder payouts.

> Fukoku Mutual Life Insurance believes it will increase productivity by 30% and see a return on its investment in less than two years. The firm said it would save about 140m yen (£1m) a year after the 200m yen (£1.4m) AI system is installed this month. Maintaining it will cost about 15m yen (£100k) a year.
>
> The technology will be able to read tens of thousands of medical certificates and factor in the length of hospital stays, medical histories and any surgical procedures before calculating payouts, according to the Mainichi Shimbun.
>
> While the use of AI will drastically reduce the time needed to calculate Fukoku Mutual's payouts—which reportedly totalled 132,000 during the current financial year—the sums will not be paid until they have been approved by a member of staff, the newspaper said.
>
> Japan's shrinking, ageing population, coupled with its prowess in robot technology, makes it a prime testing ground for AI.
>
> According to a 2015 report by the Nomura Research Institute, nearly half of all jobs in Japan could be performed by robots by 2035.

I plan on being retired by then.

Is *your* job at risk? Probably not. Assuming that you are either a data scientist trying to understand marketing or a marketing person trying to understand data science, you're likely to keep your job for a while.

In September 2015, the BBC ran its "Will a Robot Take Your Job?"[26] feature. Choose your job title from the dropdown menu and



Marketing and sales directors

Likelihood of automation?
It's quite unlikely (1%)

How this compares with other jobs:
347th of 366

Figure 1.3 Marketing and sales managers get to keep their jobs a little longer than most.

voilà! If you're a marketing and sales director, you're pretty safe. (See Figure 1.3.)

In January 2017, McKinsey Global Institute published "A Future that Works: Automation, Employment, and Productivity,"[27] stating, "While few occupations are fully automatable, 60 percent of all occupations have at least 30 percent technically automatable activities."

The institute offered five factors affecting pace and extent of adoption:

1. *Technical feasibility:* Technology has to be invented, integrated, and adapted into solutions for specific case use.
2. *Cost of developing and deploying solutions:* Hardware and software costs.
3. *Labor market dynamics:* The supply, demand, and costs of human labor affect which activities will be automated.
4. *Economic benefits:* Include higher throughput and increased quality, alongside labor cost savings.
5. *Regulatory and social acceptance:* Even when automation makes business sense, adoption can take time.

Christopher Berry sees a threat to the lower ranks of those in the marketing department.[28]

> If we view it as being a way of liberating people from the drudgery of routine within marketing departments, that would be quite a bit more exciting. People could focus on the things that are most energizing about marketing like the creativity and the messaging—the stuff people enjoy doing.
>
> I just see nothing but opportunity in terms of tasks that could be automated to liberate humans. On the other side, it's a typical employment problem. If we get rid of all the farming jobs, then what are people going to do in the economy? It could be a tremendous era of a lot more displacement in white collar marketing departments.
>
> Some of the first jobs to be automated will be juniors. So we could be very much to a point where the traditional career ladder gets pulled up after us and that the degree of education and professionalism that's required in marketing just increases and increases.

So, yes, if you've been in marketing for a while, you'll keep your job, but it will look very different, very soon.

MACHINE LEARNING'S BIGGEST ROADBLOCK

That would be *data*. Even before the application of machine learning to marketing, the glory of *big data* was that you could sort, sift, slice, and dice through more data than previously computationally possible.

Massive numbers of website interactions, social engagements, and mobile phone swipes could be sucked into an enormous database in the cloud and millions of small computers that are so much better, faster, and cheaper than the Big Iron of the good old mainframe days could process the heck out of it all. The problem then—and the problem now—is that these data sets do not play well together.

The best and the brightest data scientists and analysts are still spending an enormous and unproductive amount of time performing janitorial work. They are ensuring that new data streams are properly vetted, that legacy data streams continue to flow reliably, that the data that comes in is formatted correctly, and that the data is appropriately groomed so that all the bits line up.

- Data set A starts each week on Monday rather than Sunday.
- Data set B drops leading zeros from numeric fields.
- Data set C uses dashes instead of parentheses in phone numbers.
- Data set D stores dates European style (day, month, year).
- Data set E has no field for a middle initial.
- Data set F stores transaction numbers but not customer IDs.
- Data set G does not include in-page actions, only clicks.
- Data set H stores a smartphone's IMEI or MEID number rather than its phone number.
- Data set I is missing a significant number of values.
- Data set J uses a different scale of measurements.
- Data set K, and so on.

It's easy to see how much work goes into data cleansing and normalization. This seems to be a natural challenge for a machine learning application.

Sure enough, there are academics and data scientists working on this, but they're a long way off. How can you tell?

In their paper titled "Probabilistic Noise Identification and Data Cleaning,"[29] Jeremy Kubica and Andrew Moore describe their work on *not* throwing out entire records when only some of the fields are contaminated. "In this paper we present an approach for identifying corrupted fields and using the remaining non-corrupted fields for subsequent modeling and analysis. Our approach learns a probabilistic model from the data that contains three components: a generative model of the clean data points, a generative model of the noise values, and a probabilistic model of the corruption process."

It's a start.

MACHINE LEARNING'S GREATEST ASSET

That would be *data*. Machine learning has a truly tough time with too little information. If you give it only one example, it can tell you *exactly* what to expect the next time with 100 percent confidence. It will be wrong.

Machine learning doesn't work like statistics. Statistics can tell you the likelihood of a coin toss or the probability of a plane crash.

PROBABILITY OF A PLANE CRASH

Three statisticians are in a plane when the pilot announces that they've lost one of their engines. "But it's okay, folks, these planes were built to fly under the worst conditions. It does mean, however, that we're going to fly a bit slower and we'll be about a half an hour late. Please don't worry. Sit back, relax, and enjoy the rest of your flight."

The first statistician says, "There's still a 25 percent chance that I'll make my connection."

Fifteen minutes later, the pilot is on the PA again. "Ladies and gentlemen, we seem to have lost a second engine. No problem, the others are still going strong. This does mean, however, that we'll be about an hour late to the gate. I'm so sorry for the inconvenience."

The second statistician says, "There's an 83 percent chance I'm going to miss my dinner."

After a half an hour, the pilot makes *another* announcement, "Ladies and Gents, we've lost yet another engine. Yes, I know this is bad, but there's really no need to worry. We'll make it just fine, but we're going to be two hours late to the airport."

The third statistician says, "That last engine better not fail or we'll *never* land!"

Human experience and ingenuity have worked wonders for marketing for hundreds of years: gut feel and common sense. When we added statistics to the mix, we expanded our experience by considering historical precedent. But we still rely on gut feel as we feel around blindly in the data, hoping to stumble on something recognizable.

How We Used to Dive into Data

As the Board Chair of the Digital Analytics Association, I strove to explain how digital analysts go beyond answering specific questions. I wrote the following in the Applied Marketing Analytics Journal, describing the role of the "data detective."

> **Discovering Discovery, Data Discovery Best Practices**[30]
>
> A crystal ball is filled with nothing at all or smoke and clouds, mesmerizing the uninitiated, but very useful for the scrying specialist. The crystal ball mystic is tasked with entertaining more than communicating genuine visions. Creating something from nothing takes imagination, creativity, and the ability to read

one's fellow man to determine what fictions they might consider valuable. The medium who directs a séance is in much the same role.

Tarot Card readers are a step closer to practicality. They use their cards as conversation starters. "You drew The Magician, which stands for creation and individuality, next to the Three of Cups, which represents a group of people working together. Are you working on a project with others right now?" The "mystical" conversation is all about the subject, and therefore, seems revelatory.

The Digital Analyst also has a crystal ball (The Database) and Tarot Cards (Correlations) with which to entice and enthrall the Truth Seeker. The database is a mystery to the supplicant, and the correlations seem almost magical.

The Digital Analyst has something more powerful than visions and more practical than psychology—although both are necessary in this line of work. The analyst has data; data that can be validated and verified. Data that can be reliably used to answer specific questions.

The Digital Analyst truly shines when seeking insight beyond the normal, predictable questions asked on a daily basis. The analyst can engage in discovery; the art of uncovering important truths that can be useful or even transformative to those who would be data-driven.

Traditional Approach: Asking Specific Questions

> A business manager wants to know the buying patterns of her customers.
>
> A shipping manager wants to project what increased sales will mean to staffing.
>
> A production manager wants to anticipate and accordingly adjust the supply chain.
>
> An advertising professional wants to see the comparative results of a half a dozen promotional campaigns.

Each of these scenarios call for specific data to be assembled and tabulated to provide a specific answer. Proper data collecting, cleansing, and blending are required, and can be codified if the same questions are to be asked repeatedly. And thus, reporting is born.

Reports are valuable and necessary . . . until they are not. Then they are the source of repetitive stress, adding no value to the organization. The antidote is discovery.

Exploring Data

An investigation is an effort to get data to reveal what it knows. (“Where were you on the night of the 27th?”). But data discovery is the art of interviewing data to learn things you didn’t necessarily know you wanted to know.

The Talented data explorer is much like the crystal ball gazer and the Tarot reader in several ways. They:

> Have a method for figuring out what the paying customer wants to know.
>
> Have broad enough knowledge about the subject to recognize potentially interesting details.
>
> Are sufficiently open minded to be receptive to details that *might* be relevant.
>
> Keep in close communication with the petitioner to guide the conversation.
>
> Understand the underlying principles well enough to push the boundaries.
>
> Are curious by nature and enjoys the intellectual hunt.

Data discovery is part mind reading, part pattern recognition, and part puzzle solving. Reading the mind of the inquisitor is obligatory to ensure the results are of interest to those with control of the budget. Pattern recognition is a special skill that can be honed to help direct lines of enquiry and trains of thought. An aptitude for detective work is the most important talent of the Digital Analyst; that ability to ponder the meaning of newly uncovered evidence.

Data discovery is the art of mixing an infinitely large bowl of alphabet soup and being able to recognize the occasional message that floats to the surface in an assortment of languages. Although, with Big Data, adding more data variety to the mix, the Digital Analyst must also be able to read tea leaves, translate the I Ching, generate an astrological chart, interpret dreams, observe auras, speak in tongues, and sing with sirens in order to turn lead into gold.

Data discovery is all about the application of those human skills that computers have a tough time with reasoning, creativity, learning, intuition, application of incongruous knowledge, etc.

Computers are fast but dumb, while humans are slow but smart.

That doesn't mean technology cannot be helpful.

Data Discovery Tools

The business intelligence tool industry is pivoting as fast as it can to offer up data discovery tools. They describe their offerings in florid terms:

> Imagine an analytics tool so intuitive, anyone in your company could easily create personalized reports and dynamic dashboards to explore vast amounts of data and find meaningful insights. (Qlik.com[1])
>
> Tableau enables people throughout an organization—not just superstar analysts—to investigate data to find nuances, trends, and outliers in a flash. (Yes, the superstars benefit, too.) No longer constrained to a million rows of spreadsheet data or a monthly report that only answers a few questions, people can now interact and visualize data, asking—and answering—questions at the speed of thought.
>
> Using an intuitive, drag-and-drop approach to data exploration means spending time thinking about what your data is telling you, not creating a mountain of pivot tables or filling out report requests. (Tableau[2])
>
> We help people make faster, better business decisions, empowering them with self-service tools to explore data and share insights in minutes.... Simple drag-and-drop tools are paired with intuitive visualizations. Connect to any data source and share your insights in minutes.... Standalone data discovery tools will only get you so far. Step into enterprise-ready analytics and guarantee secure, governed data discovery. (Microstrategy[3])

Regardless of the speed and agility of one technology or another, it all depends on the person driving the system to ask really good questions. However, if the system does not have really good data, even the best questions will result in faulty insights. Therefore, data hygiene takes precedent over superior query capability.

Data Hygiene

Garbage in, garbage out. So much goes into Big Data, it's very hard to know which bits are worthy of being included and which need to be rectified. For that, you need a subject matter expert *and* a data matter expert.

A data matter expert is knowledgeable about a specific stream: how it was collected, how it was cleansed, sampled, aggregated and segmented, and what transformation is required before blending it with other streams.

Data hygiene and data governance are paramount to ensure the digital analytics cooks are using the very best ingredients to avoid ruining a time-proven recipe.

Further, when the output of one analysis provides the input for the next (creating a dashboard, for example), transformation, aggregation and segmentation help obfuscate the true flavor of the raw material until it is past the ability of a forensic data scientists to track down the cause of any problems—supposing somebody is aware that there is a problem.

Yet, aggregations are as important to the insight supply chain as top-grade ingredients are to the five-star chef:

> [D]ata aggregations and summaries remain critical for supporting visual reporting and analytics so that users can see specific time periods and frame other areas of interest without getting overwhelmed by the data deluge. Along with providing access to Hadoop files, many modern visual reporting and data discovery tools enable users to create aggregations as the need arises rather than having to suffer the delays of requisitioning them ahead of time from IT developers. In a number of leading tools, this is accomplished through an integrated in-memory data store where the aggregations are done on the fly from detailed data stored in memory.

> TDWI Research finds that enterprise data warehouses, BI reporting and OLAP cubes, spreadsheets, and analytic databases are the most important data sources for visual analysis and data discovery, according to survey respondents. (TDWI[4])

The care and feeding of the raw material used in the data discovery process is even more important in light of the lack of five-star chefs. As analytics becomes more accepted, demanded and democratized, more and more amateur analysts will be deriving conclusions from raw material they trust implicitly rather than understand thoroughly. Preparing for data illiterate explorers requires even more rigor than usual to guard against their impulse to jump to the wrong conclusions.

Asking Really Good Questions

In the hands of a well-informed analyst, lots of data and heavy-lifting analytics tools are very powerful. Getting the most out of this combination takes a little bit of creativity.

Creativity means broadening your mental scope. Rather than seeking a specific answer, open yourself up to possibilities. It's like focusing on your peripheral vision.

1. Appreciate Anomalies

Whether you use visualization tools and "look for" things that go bump in the night, or you are adept at scanning a sea of numbers and wondering why it looks out of balance, the skill to hone is the art of seeing the out-of-the-ordinary.

Outliers, spikes, troughs—any anomaly—are our friends. They draw our attention to that which is not like the others and spark the intellectual exercise of wondering "Why?"

What is it about this element that makes it point in a different direction? Could it be some error in the collection or transformation of the underlying data? Is it a function of how the report was written or the query was structured? Or does it represent some new behavior/market movement/customer trend?

It is in the hunt for the truth about these standouts that we trip over the serendipitous component that spawns a new

question and another dive down the rabbit hole.
The secret is knowing when to stop.

One can easily get lost in a hyperlink-chasing "research session" and burn hours with very little to show for it. Following the scent of significance is an art and one that takes practice and discipline. Many scientists spend a career pursuing a specific outcome only to find it disproved. Others stop just short of a discovery because they lose heart. The magic happens between those two points.

Give in to the temptation to slice the data one more time or to cross reference results against just one more query, but be vigilant that you are not wasting valuable cycles on diminishing returns.

If you don't see what you expect to see, work your hardest to understand why. It may be that you do not have enough facts. It might be that you have already, unknowingly, come to a conclusion or formed a pet theory without all the facts. It might be—and this is the most likely—that there is something afoot which you have not yet considered.

Dig deeper. Ask, "I wonder . . . ." And be cognizant of that which is conspicuous in its absence.

> *Gregory (Scotland Yard detective): "Is there any other point to which you would wish to draw my attention?"*
>
> *Holmes: "To the curious incident of the dog in the night-time."*
>
> *Gregory: "The dog did nothing in the night-time."*
>
> *Holmes: "That was the curious incident."*
>
> *Sir Arthur Conan Doyle*, Silver Blaze

As a corollary, be wary of the homologous as well:

1. Exhibiting a degree of correspondence or similarity.
2. Corresponding in structure and evolutionary origin, but not necessarily in function.

For example, human arm, dog foreleg, bird wing, and whale flipper are homologous. (A Word A Day[5])

Things that are unusually similar are equally cause for alarm as standouts. If everybody in your cohort looks the same, there's something funny going on and it's worth an investigation. It may be that their similarity is a statistical anomaly.

2. Savor Segmentation

People (thank heaven!) are different. We make a huge mistake when we lump them all together. But we cannot treat them as individuals—yet. Peppers and Rogers' *One to One Future* is not yet upon us. In between lies segmentation.

It almost doesn't matter how you segment your customers (geographically, chronologically, by hair color). Eventually, you will find traits that are useful in finding a cluster of behavior that can be leveraged to your advantage.

> People who come to our website in the morning are more likely to X.
>
> People who complain about us on social media respond better to message Y.
>
> People who use our mobile app more than twice a week are more likely to Z.

When it comes to segmenting customers by behavior, Bernard Berelson pretty much nailed it in his "Human Behavior: An Inventory of Scientific Findings"[6] where he said:

> Some do and some don't.
>
> The differences aren't that great.
>
> It's more complicated than that.

When you're trying to get the right message in front of the right people at the right time and on the right device, segmentation may likely be the key to the mystery.

3. Don't Fool Yourself

While working with data is reassuring—we are, after all dealing with facts and not opinions—we are still human and still faced with serious mental handicaps.

Being open-minded and objective are wonderful goals, but they are not absolute.

Cognitive biases are inherited, taught, and picked up by osmosis in a given culture. In short, your mind can play tricks on you. While this is too large a subject to cover in depth here, there are some examples that make it clear just how tenuous your relationship with "the facts" might be.

Familiarity Bias
I've worked in television advertising all my life and I can tell you without any doubt that it's the most powerful branding medium there is.

Hindsight or Outcome Bias
If they'd only have asked me, I would have told them that the blue button would not convert as well as the red one. It was obvious all along.

Attribution Bias
Of course I should have turned left at that light. But I was distracted by the sun in my eyes and the phone ringing. That other guy missed the turn because he's a dim-wit.

Representativeness Bias
Everybody who clicks on that link must be like everybody else who clicked on that link in the past.

Anchoring Bias
That's far too much to pay for this item. The one next to it is half the price.

Availability Bias (the first example that comes to mind)
That'll never work—let me tell you what happened to my brother-in-law . . .

Bandwagon Bias
We should run a Snapchat campaign because everybody else is doing it.

Confirmation Bias
I'm a conservative, so I only watch Fox News.

I'm a liberal, so I only watch The Rachel Maddow Show.

I've been in advertising all my life, so I count on Nielsen, Hitwise, and comScore.

> I started out grepping log files, so I only trust my Coremetrics/Omniture/Webtrends numbers.
>
> **Projection Bias**
> I would never click on a product demo without a long list of testimonials, so we can assume that's true of everybody else.
>
> **Expectancy Bias**
> Your report must be wrong because it does not show the results I anticipated.
>
> **Normalcy Bias**
> Back-ups? We've never had a data loss problem yet, I don't see it happening this quarter so we won't have to budget for it.
>
> **Semmelweis Reflex**
> I don't care what your numbers say, we've always had better conversions from search than social media so we're not going to change our investment.

If any of the above sound familiar, congratulations—you've been paying attention. The hard part is convincing others that there may be a cognitive problem.

4. Correlation versus Causation

While frequently mentioned, it cannot be stressed enough that just because drownings go up when ice cream sales go up, one did not cause the other.

Most recently, a Swedish study ("Allergy in Children in Hand Versus Machine Dishwashing"[7]) concluded, "In families who use hand dishwashing, allergic diseases in children are less common than in children from families who use machine dishwashing," and speculated that, "a less-efficient dishwashing method may induce tolerance via increased microbial exposure."

While the study asked a great deal of questions about the types of food they eat, food preparation, parental smoking, etc., there are simply too many other variables at play for this cause to be solely responsible for that effect. How many other similarities are there among families that have dishwashers vs. those that do not?

Correlations are a wonderful clue, but they must be treated as clues and not results. Correlations are the stimulus for seeking a cause, not the end of the story.

5. Communicating Carefully

Coming up with a fascinating correlation and proving a causative relationship can be exciting. The thrill of the chase, the disappointment of a miscalculation, and the redemption of the correction make for an invigorating career, but like your latest round of golf, not necessarily a great story at the dinner table. And certainly not at the conference room table or across the desk from an executive who is trying to make a multimillion-dollar advertising decision.

This is the time to stick with what you know, not how you got there.

The most important part of your performance when delivering insights based on data is to avoid any bravado of certainty. You have not been asked to audit the books and come to a conclusion. You have not been tasked with adding up a row of numbers and delivering The Answer. Instead, you have been asked to sift through the data to see if there's anything in there that might be directional.

To assure everybody else that you understand your responsibility and to appropriately frame your findings in terms that will lead to a valuable conversation and business decision, monitor your language carefully.

> The data suggests . . .
>
> It seems more likely . . .
>
> One could conclude . . .
>
> Based on the data, it feels like . . .
>
> If I were placing bets after seeing this . . .

Remember that you are looking into a crystal ball that is a complete mystery to the business side of the house and you are telling them things about a subject they know very well, just not through that lens. They know advertising and marketing inside and out and are going to be incredulous if you make pronouncements that are contrary to their experience, gut feel, and common sense.

The domain expert can look at a carefully scrutinized, statistical revelation and roll their eyes.

> "Of course movies starting with the letter A are more popular—we list them alphabetically."
>
> "Of course online sales took a jump the week in that region—there was a five day blizzard."
>
> "Of course we sold more low-end laptops that day—our competitor's website was down."

Be sure to sound more like the weather prognosticator who talks about a chance of showers. Use the vernacular or the gambler running the odds. Think in terms of a Probability Line [Figure 1.4] and choose your words accordingly.

Follow the lead of doctors who talk about relative health risks. And then, draw them into the supposition process.

> Doesn't that seem logical?
>
> Does that meet or challenge your thoughts?
>
> Do you think it means this or that?

It shouldn't take long to get them to see you as an advisor and not a report writer.

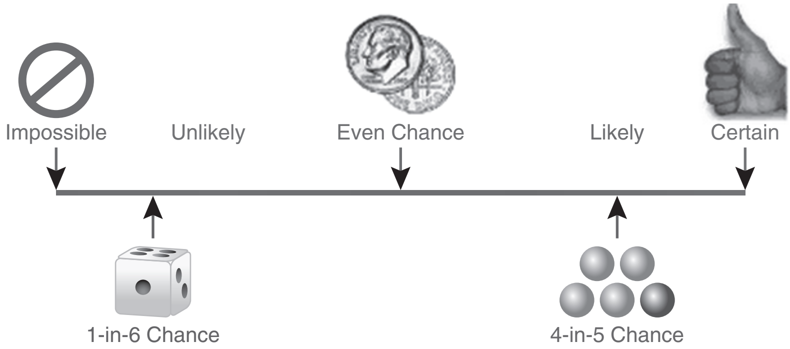


Figure 1.4 The spectrum of probability (Math Is Fun[31])

6. Become a Change Agent

The very best way to win the hearts and minds of those who can benefit the most by your flair for data discovery is to educate them.

The more people in your organization who understand the ways and means of data exploration as well as the associate risks and rewards, the more they will come to you for answers, include you in planning sessions and support your calls for more data, people and tools.

Start by inviting them to lunch. Ask them to bring their best questions about The Data. Encourage those who would rather not be seen as ill-informed to submit their questions in advance. Prepare a handful of questions that you wish they would ask.

Answer their questions. Show them examples of quick-wins enjoyed by other projects in other departments. Share case studies from vendors about successes at other companies.

Engage your audience in the excitement of the chase with a simple data set and a common challenge. If you can teach them how to ask great questions by example and by exercise then you can change how they approach data—to see it as a tool instead of an accusation.

And be sure to feed them. This is a case where a free lunch will pay off handsomely.

Your Job as Translator

You know your data inside and out, but the consumers of your insights, who must depend on your recommendations do not. To them, your data is as readable as a crystal ball or a sequence of Tarot cards. That means they are putting their trust in you.

Therefore, your responsibility is to inform without confusing, to encourage without mystifying and to reassure without resorting to sleight of hand. Entice and enthrall your Truth Seekers with The Data and The Correlations, but make sure your confidence levels are high and be prepared to show your work.

Conclusion

Successful data discovery requires good tools (technology) and trustworthy raw material (clean data), but depends on the creativity of the data detective. The best analyst has the ability to manipulate data in a variety of ways to tease out relevant insights. With the goals of the organization firmly in mind, top analysts engage the data in a conversation of What-Ifs, resulting in tangible insights that can be used to make decisions by those in charge. The analyst, as consulting detective, becomes indispensable.

NOTES

1. Self-Service Data Discovery and Visualization Application, Sense BI Tool | Qlik, available at http://www.qlik.com/us/explore/products/sense, last accessed on 3/13/15.
2. Data Discovery | Tableau Software, available at http://www.tableau.com/solutions/data-discovery, last accessed on 3/13/15.
3. Features of the Analytics Platform | MicroStrategy, available at http://www.microstrategy.com/us/analytics/features, last accessed on 3/13/15.
4. Data Visualization and Discovery for Better Business Decisions, available at http://www.adaptiveinsights.com/uploads/news/id421/tdwi_data_visualization_discovery_better_business_decisions_adaptive_insights.pdf, last accessed on 3/13/15.
5. A.Word.A.Day—homologous, available at http://wordsmith.org/words/homologous.html, last accessed 3/13/15.
6. Human Behavior: An Inventory of Scientific Findings, available at http://home.uchicago.edu/aabbott/barbpapers/barbhuman.pdf, last accessed 3/13/15.
7. Allergy in Children in Hand Versus Machine Dishwashing, available at http://pediatrics.aappublications.org/content/early/2015/02/17/peds.2014-2968.full.pdf, last accessed 3/13/115.

Variety of Data Is the Spice of Life

Machine learning differs from data diving. It is like putting tens of thousands of statisticians in a black box and throwing in a question. They will scour through the data in different ways, confer, and then pop out an answer along with their degree of confidence. Next, they will test their answer against some fresh information and adjust their opinion. The more data you let them look at, and the more they cycle their assumptions against real-world results, the better.

With the price of storage in a downward spiral to almost nothing and the speed of processing continuing to increase thanks to parallel processing in the cloud, we can crunch through a great deal more information than ever. Machine learning is good with lots of data, but it *really* goes to town when it has lots of different types of data to play with. It can find correlations between attributes humans wouldn't even consider comparing. If there is a relationship among the weather, the color of socks a prospect is wearing, and what the prospect had for lunch, then marketers can leverage that correlation. It doesn't matter if the correlation is logical or even understandable, it only matters that it is actionable.

In addition to all the digital interaction data that drove the whole Big-Data-Hadoop-Clusters-in-the-Cloud movement, now there's even *more* data to chew on out there.

Open Data

Hundreds of organizations, both governmental and NGOs, are publishing a shockingly large amount of data that might be useful in finding your next customer. Just think about all the APIs (application program interfaces) that allow you to grab onto firehoses like Facebook and Twitter. Facebook Likes alone can predict quite a bit about you as an individual, according to a paper from the Psychometrics Centre, University of Cambridge.[32] "Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes, including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender."

Think about all the recipes you can get from Campbell's Soup:[33]

> The Campbell's Kitchen API was developed to share information from Campbell's Kitchen. This information includes thousands of recipes using brands like Campbell's®, Swanson®, Pace®, Prego®, & Pepperidge Farm®—brands people love, trust, and use every day. The easier people can find those recipes, the less time they have to spend worrying about what to make for dinner.

> We hope you will use this information to develop smart and simple ways to help people get the dinner and entertaining ideas they're looking for.
>
> GET ACCESS TO:
>
> - Thousands of proven family favorite recipes
> - Extensive recipe filtering by key ingredients, product UPC, keywords and more
> - Professional food photography
> - Reader-generated recipe reviews & comments
> - Recipe search results through superior tagging
> - Well-known food brands people know and trust
>
> SO MANY POSSIBILITIES:
>
> Enhance websites with related recipes & delicious looking photographs
>
> Create food-related apps (for websites and the latest and greatest devices and toys) and helpful shopping and cooking tools
>
> Augment social media sites like Facebook, Twitter, & Google+
>
> Raise visibility for your brand
>
> Drive more traffic to your site and gain new readers from a wider audience
>
> The sky's the limit

Imagine cross-referencing the people who comment on recipes with their social media accounts to target people by flavor preferences. But that's just the tip of the iceberg. Google hosts a growing number of data sets that are directly accessible through its BigQuery utility.[34]

> BigQuery is a fully managed data warehouse and analytics platform. The public datasets listed on this page are available for you to analyze using SQL queries. You can access BigQuery public data sets using the web UI, the command-line tool, or by making calls to the BigQuery REST API using a variety of client libraries such as Java, .NET, or Python.

The first terabyte of data processed per month is free, so you can start querying datasets without enabling billing. To get started running some sample queries, select or create a project and then run the example queries on the NOAA GSOD weather dataset.

GDELT Book Corpus
A dataset that contains 3.5 million digitized books stretching back two centuries, encompassing the complete English-language public domain collections of the Internet Archive (1.3 M volumes) and HathiTrust (2.2 million volumes).

GitHub Data
This public dataset contains GitHub activity data for more than 2.8 million open source GitHub repositories, more than 145 million unique commits, over 2 billion different file paths, and the contents of the latest revision for 163 million files.

Hacker News
A dataset that contains all stories and comments from Hacker News since its launch in 2006.

IRS Form 990 Data
A dataset that contains financial information about nonprofit/exempt organizations in the United States, gathered by the Internal Revenue Service (IRS) using Form 990.

Medicare Data
This public dataset summarizes the utilization and payments for procedures, services, and prescription drugs provided to Medicare beneficiaries by specific inpatient and outpatient hospitals, physicians, and other suppliers.

Major League Baseball Data
This public dataset contains pitch-by-pitch activity data for Major League Baseball (MLB) in 2016.

NOAA GHCN
This public dataset was created by the National Oceanic and Atmospheric Administration (NOAA) and includes climate summaries from land surface stations across the globe that have been subjected to a common suite of quality assurance reviews. This dataset draws from more

than 20 sources, including some data from every year since 1763.

NOAA GSOD
This public dataset was created by the National Oceanic and Atmospheric Administration (NOAA) and includes global data obtained from the USAF Climatology Center. This dataset covers GSOD data between 1929 and 2016, collected from over 9000 stations.

NYC 311 Service Requests
This public data includes all 311 service requests from 2010 to the present, and is updated daily. 311 is a non-emergency number that provides access to non-emergency municipal services.

NYC Citi Bike Trips
Data collected by the NYC Citi Bike bicycle sharing program, that includes trip records for 10,000 bikes and 600 stations across Manhattan, Brooklyn, Queens, and Jersey City since Citi Bike launched in September 2013.

NYC TLC Trips
Data collected by the NYC Taxi and Limousine Commission (TLC) that includes trip records from all trips completed in yellow and green taxis in NYC from 2009 to 2015.

NYPD Motor Vehicle Collisions
This dataset includes details of Motor Vehicle Collisions in New York City provided by the Police Department (NYPD) from 2012 to the present.

Open Images Data
This public dataset contains approximately 9 million URLs and metadata for images that have been annotated with labels spanning more than 6,000 categories.

Stack Overflow Data
This public dataset contains an archive of Stack Overflow content, including posts, votes, tags, and badges.

USA Disease Surveillance
A dataset published by the U.S. Department of Health and Human Services that includes all weekly surveillance reports of nationally notifiable diseases for all U.S. cities and states published between 1888 and 2013.

USA Names
A Social Security Administration dataset that contains all names from Social Security card applications for births that occurred in the United States after 1879.

In its top-20 list of the best free data sources available online, Data Science Central includes:[35]

1. Data.gov.uk, the UK government's open data portal including the British National Bibliography—metadata on all UK books and publications since 1950.
2. Data.gov. Search through 194,832 USA data sets about topics ranging from education to Agriculture.
3. US Census Bureau latest population, behaviour and economic data in the USA.
4. Socrata—software provider that works with governments to provide open data to the public, it also has its own open data network to explore.
5. European Union Open Data Portal—thousands of datasets about a broad range of topics in the European Union.
6. DBpedia, crowdsourced community trying to create a public database of all Wikipedia entries.
7. The New York Times—a searchable archive of all *New York Times* articles from 1851 to today.
8. Dataportals.org, datasets from all around the world collected in one place.
9. The World Factbook information prepared by the CIA about, what seems like, all of the countries of the world.
10. NHS Health and Social Care Information Centre datasets from the UK National Health Service.
11. Healthdata.gov, detailed USA healthcare data covering loads of health-related topics.
12. UNICEF statistics about the situation of children and women around the world.
13. World Health organisation statistics concerning nutrition, disease and health.
14. Amazon web services' large repository of interesting datasets including the human genome project, NASA's database and an index of 5 billion web pages.

15. Google Public data explorer search through already mentioned and lesser known open data repositories.
16. Gapminder, a collection of datasets from the World Health Organisation and World Bank covering economic, medical and social statistics.
17. Google Trends analyse the shift of searches throughout the years.
18. Google Finance, real-time finance data that goes back as far as 40 years.
19. UCI Machine Learning Repository, a collection of databases for the Machine Learning community.
20. National Climatic Data Center, world largest archive of climate data.

While all of the above is far too much for humans to sift through, machines might be able to find a useful, and potentially profitable, correlation. One Oracle blog post[36] included this about Red Roof Inn:

> Marketers for the hotel chain took advantage of open data about weather conditions, flight cancellations and customers' locations to offer last-minute hotel deals to stranded travelers. They used the information to develop an algorithm that considered various travel conditions to determine the opportune time to message customers about nearby hotel availability and rates.

Might information on Iowa liquor sales be useful? "This dataset contains the spirits purchase information of Iowa Class 'E' liquor licensees by product and date of purchase from January 1, 2014, to current. The dataset can be used to analyze total spirits sales in Iowa of individual products at the store level."[37]

And don't look now, but here comes the Internet of Things and the unbelievable amounts and types of data that will come spilling out.

The same can be said for exhaust data. That's information that's a byproduct of some action, reaction, or transaction. Walking through a shopping center throws off lots of exhaust information about where you are. How often you respond to text messages, where you take pictures, and whether you speed up at yellow stop lights is reactive. Whether stocks trade more when the market goes up or down is transaction-oriented.

There are, of course, companies that offer a conglomeration of the above as a service. Second Measure sells insights derived from credit card transactions so you can "spot inflections in businesses as they happen, identify this week's fastest-growing companies [and] see the latest KPIs (Key Performance Indicators) before they're announced."

Mattermark monitors marketplace KPIs such as companies' net revenue, gross margin, growth, market share, liquidity, average order value, Net Promoter Score, retention, cost per customer acquisition, marketing channel mix, overall ROI, and cash burn rate. This is a whole new data set for B2B sales and competitive intelligence.

The combination of all of the available data with the power of machine learning is cause for excitement and competitive advantage. (See Figure 1.5.)

Data for Sale

Upon this gifted age, in its dark hour,

Rains from the sky a meteoric shower

Of facts . . . . They lie unquestioned, uncombined.

Wisdom enough to leech us of our ill

Is daily spun, but there exists no loom

To weave it into fabric . . . .

Edna St. Vincent Millay from Sonnet 137, Huntsman, What Quarry*?*

In an ideal world, the machine collects all the data there is and weaves it into a tapestry that makes all things clear at a glance. The data aggregation industry has been active for years, starting with the Census Bureau 115 years ago. Since then, it's become a big business.

The amount of available information is enormous from public records and criminal databases to credit rating firms and credit card companies to public companies like Dun & Bradstreet and Acxiom, which claims to have more than 32 billion records. That's the sort of aggregator that powers most direct mail and telemarketers.

> Acxiom's extensive third-party data offers rich insight into consumers and their behaviors:
>
> Curated from multiple, reliable sources
>
> Includes more than 1,000 customer traits and basic information including, location, age, and household details

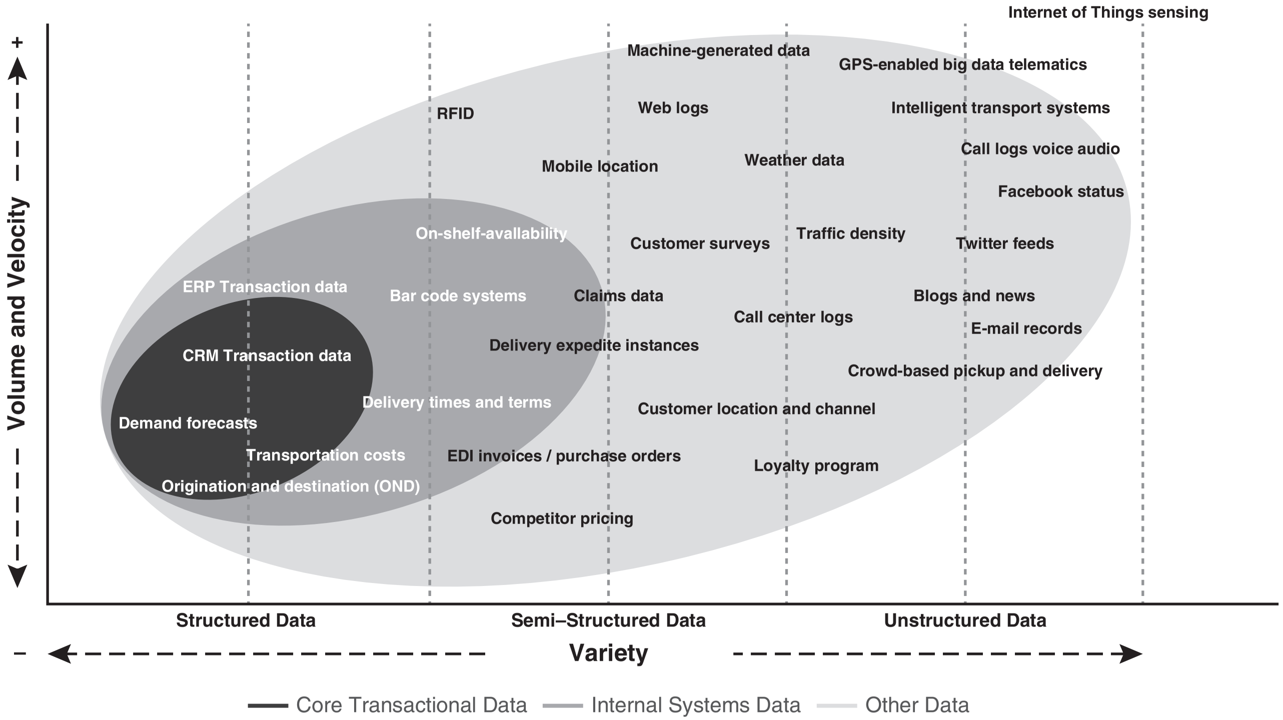


Figure 1.5 So many types of data, so little time[38]

Provides more than 3,500 specific behavioral insights, such as propensity to make a purchase

Offers real insights into a broad spectrum of offline behavior, not just indicators from web browsing behavior

Gives analysts more ways to segment data and use for audience modeling

Acxiom data fuels highly personalized data-driven campaigns, enabling you to:

Personalize messages and consistently engage audiences across all channels

Incorporate both online and offline data in a safe, privacy-compliant way

Segment audiences at the household or individual level based on a variety of options from ethnicity and acculturation to digital behaviors

Optimize for scale and accuracy

Request audience recommendations from seasoned data experts[39]

But Wait—There's More

The volume and variety of data seems to have no end.

- The weather (http://www.ncdc.noaa.gov/)
- U.S. Census data (http://dataferrett.census.gov/)
- Japan Census Data (https://aws.amazon.com/datasets/Economics/2285)
- Health and retirement study (http://www.rand.org/labor/aging/dataprod/hrs-data.html)
- Federal Reserve economic data (https://aws.amazon.com/datasets/Economics/2443)
- The entire Internet for the past seven years (http://commoncrawl.org/)
- 125 years of public health data (http://www.bigdatanews.com/group/bdn-daily-press-releases/forum/topics/pitt-unlocks-125-years-of-public-health-data-to-help-fight-contag)

- Consumer complaints about financial products and services (http://catalog.data.gov/dataset/consumer-complaint-database)
- Product safety recalls from the Consumer Product Safety Commission (http://www.cpsc.gov/Newsroom/News-Releases/2010/CPSC-Makes-Recall-Data-Available-Electronically-to-Businesses-3rd-Party-Developers/)
- Franchise failures by brand (https://opendata.socrata.com/Business/Franchise-Failureby-Brand2011/5qh7-7usu)
- Top 30 earning websites (https://opendata.socrata.com/Business/Top-30-earning-websites/rwft-hd5j)
- Car sales data (https://opendata.socrata.com/Business/Car-Sales-Data/da8m-smts)
- Yahoo! Search Marketing Advertiser bidding data (http://webscope.sandbox.yahoo.com/catalog.php?datatype=a)
- American time use survey (http://www.bls.gov/tus/tables.htm)
- Global entrepreneurship monitor (http://www.gemconsortium.org/Data)
- Wage Statistics for the U.S. (http://www.bls.gov/bls/blswage.htm)
- City of Chicago building permits from 2006 to the present (https://data.cityofchicago.org/Buildings/Building-Permits/ydr8-5enu)
- Age, race, income, commute time to work, home value, veteran status (http://catalog.data.gov/dataset/american-community-survey)

Or how about all of Wikipedia?

- (http://en.wikipedia.org/wiki/Wikipedia:Database_download)

A Collaboration of Datasets

After three years as a systems analyst at Deloitte, Brett Hurt started one of the first web analytics companies (Coremetrics later sold to IBM), and an online reviews and ratings company (Bazaarvoice) has turned his attention to the world of data.

His current startup is data.world, a B-Corp (Public Benefit Corporation) intent on building a collaborative data resource. From the outset, according to John Battelle,[40] "Hurt & co. may well have unleashed a blast of magic into the world."

The problem they are out to solve is allowing data to be visible. Rather than data shoved into its own database silo, hidden away from all other data, as we experience it now, data.world seeks to unlock that data and make it discoverable, just as the World Wide Web has brought links between research papers and marketing materials and blog posts.

> One consistently formatted master repository, with social and sharing built in. Once researchers upload their data, they can annotate it, write scripts to manipulate it, combine it with other data sets, and most importantly, they can share it (they can also have private data sets). Cognizant of the social capital which drives sites like GitHub, LinkedIn, and Quora, data.world has profiles, ratings, and other "social proofs" that encourage researchers to share and add value to each others' work.
>
> In short, data.world makes data discoverable, interoperable, and social. And that could mean an explosion of data-driven insights is at hand.

For artificial intelligence to really flex its muscles, it must have a *lot* of data to chew on; data.world feels like a step in the right direction to join up the massive amounts of data that's out there, for the use of all comers.

A Customer Data Taxonomy

The breadth of available data is overwhelming (social media graphs, Facebook Likes, tweets, auto registration, voting records, etc.). It's helpful to have a taxonomy at hand.

Types of Collectible Information

The wide variety of data is expanding at a phenomenal rate. Here is an indicative but not exhaustive list of data sets shoved into categorization cubbyholes through sheer blunt force.

Identity
Can we identify them? Who are they?

- Name
- Gender

- Age
- Race
- Address
- Phone
- Fingerprint
- Heart rate
- Weight
- Device
- Government ID
- And so on

History
What's in their past? What have they done or achieved?

- Education
- Career
- Criminal record
- Press exposure
- Publications
- Awards
- Association memberships
- Credit score
- Legal matters
- Loans
- Divorce
- Where they have traveled
- And so on

Proclivities
What attracts them? Are they liberal or conservative? What do they like?

- Preferences
- Settings
- Avocations
- Political party
- Social groups

- Social “Likes”
- Entertainment
- Hobbies
- News feeds
- Browser history
- Brand affinity
- And so on

Possessions
What do they have, whether purchased, acquired, found, or made?

- Income
- Home
- Cars
- Devices
- Clothing
- Jewelry
- Investments
- Subscriptions
- Memberships
- Collections
- Relationships
- And so on

Activities
Can we catch them in the act? What do they do and how do they do it?

- Keystrokes
- Gestures
- Eye tracking
- Day part
- Location
- IP address
- Social posts
- Dining out
- Television viewing

- Heart rate over time
- And so on

Beliefs

How do they feel and where do they stand on issues?

- Religion
- Values
- Donations
- Political party
- Skepticism/Altruism
- Introvert/Extrovert
- Generous/Miserly
- Adaptive/Inflexible
- Aggressive/Passive
- Opinion
- Mood
- And so on

Methods of Data Capture

All of the above comes to light in a variety of ways. The data scientist will be more responsible as time goes on—and legislation crops up—to know whether an individual data element was collected with full consent. The future will also require recording whether that consent was given in perpetuity or only for the purpose initially stated.

Here, then, are suggested categories of data capture, based on "The Origins of Personal Data and Its Implications for Governance" by Martin Abrams,[41] which included a taxonomy based on origin.

Provided

Individuals are highly aware when they are providing information. They might *initiate* the delivery of the information when filling out an application, registering to vote or registering a product for warranty, or acquiring a public license to drive, marry, or carry a gun.

The *transactional* provision of data happens any time people use a credit card. They are clearly and knowingly identifying themselves. Paying a bill by writing a check qualifies as well, as does answering surveys, registering for a school, or participating in a court proceeding. This would also pertain to filling in one of those online quizzes (Which *Star Wars* Character are *You*?).

Individuals are also said to be providing information when they post it publicly. That may be delivering a speech in public, writing a letter to the editor for publication, or posting something online in a social network. *Posting* happens when you announce to all of your Facebook friends that you are, indeed, Han Solo.

Observed

Information can be casually observed. The Internet is an ideal place for observation as every click is recorded. People forget that their phone is always listening to them in case they wish to summon "Hey Siri" or "OK Google" by voice.

Browser cookies and loyalty cards are examples of *engaged* observations. People go to a website intentionally. They have their grocery store card scanned on purpose. They know they're doing it, but they're not thinking in terms of that action being revealing. They may choose to refuse to use their membership card or surf incognito, but they trade off convenience and discounts.

An *unanticipated* collection of data surprises people for an instant, and then they realize that they knew there were sensors and conclude they probably knew data was being gathered. You know your car can talk to the cloud to get navigation map updates and to call for roadside assistance. But you might not have read the manual where it talks about collecting information on engine temperature and tire pressure as well.

The *passive* collection of data is where things start to border on creepy. People don't expect to have their picture taken by a traffic camera and then dropped into a database. They don't expect their movements to be recorded as they walk around a department store. There is no expectation of privacy, but the first time you become aware that it's happening, you feel a little queasy. After that, it becomes the new normal.

Derived

Now that the raw material has been scooped up, it's time to start massaging it. The amount of time you spend on one page or another is *computationally* derived. We subtract the time you arrive from the time you leave, and *voilá*, time-on-page. This information must be calculated. How often do you search for gaming laptops? How much do you usually buy on this site? How often do you return?

The result of each of these calculations is another data point that can be associated with an individual, but there's no way for that person to know such provided and observed data is being manipulated.

Data about you can be *notionally derived* by assigning you to a given category like lookie-loo versus serious buyer or soccer-mom versus single mother. This sort of classification is also invisible to the individual being labeled.

Play your cards right and that merchant may decide you are a prime candidate for a super-discount-member category. Finding yourself misclassified can be surprising, annoying, or cause for arrest in the wrong database.

Inferred

Data that is created through inference has taken computational data a step further into analytical evaluation.

Statistically inferred data determines whether you get a call while on vacation asking if that's really *you* checking into that hotel. Your FICO score is the statistical result of comparing you to others.

Take statistics to their logical extreme and you have *advanced analytical* data. Big data and AI are hard at work to correlate all of the above to come to a supposition about the prospect or customer. How likely are you to be who you say you are? How likely are you to default on a loan? Contract a disease? Recommend this book to your friends?

The result of each of these data collection and derivation methods is—more data. Martin Abrams posits that data supplied by individuals will remain about the same as you only need a finite number of driving or wedding licenses, even while uploading photos becomes more popular. However, observed data will enjoy healthy growth as more sensors are born into the Internet of Things. Abrams sees derived data losing ground as inferred data becomes more popular.

That brings us back to AI and machine learning. "Inferred data will accelerate as more and more organizations, both public and private, increasingly take advantage of broader data sets, more computing power, and better mathematical processes," says Abrams. "The bottom-line is that data begets more data."

Marketing Data Trustworthiness

Data is a wonderful thing—especially digital data because it's binary. It's either ones or zeros and crystal clear. While we'd all like to believe that's true, only those who don't know data at all would fall for that.

So Much Data, So Little Trust

One of the more difficult aspects of marketing data is its uneven fidelity. Transactions are dependable. A sale was made at a given time to a given person at a given price—all rather solid. On the far end of the spectrum, social media sentiment is almost guesswork.

The conundrum comes when marketing professionals are asked to rank the relative reliability of various data sets; their minimal knowledge of the data stands in their way.

When multiple metrics are combined to form an index, the variable trustworthiness of the variables is completely hidden. The solution to this dilemma lies in data scientists working closely with marketers to properly weight the variety of data elements that go into the soup.

So Much Data, So Little Connection

Matthew Tod of D4t4 Solutions Plc tells a story of trying to fit online data to offline data that starts *after* the struggle to line up the two is over.[42]

> I was working with a retailer with standard, online behavioral data from a tag-based, log file system tracking sessions. Fortunately for me, they had linked them to email addresses so I have a key to join sessions to email addresses. They started issuing e-receipts. You go into the store, you buy the stuff, they email you a receipt. But only about 35% of their in-store transactions warrant a receipt, any receipt, but that 35% of transaction accounts for 90% of sales revenue because people only want a receipt for insurance purposes, or for returning a product, so for the valuable stuff. For the little stuff, nobody is going to ask for an e-receipt.
>
> So, I end up with a data set roughly 80 million sessions on the website, a million email addresses and 55 million rows of transactional data. I bring all of that together in order to answer the question, "What is the impact of Google on my physical store sales?"
>
> Because I now have a link from store sale to session, and via campaign back to Google with 300,000 people, I could say, matthew.tod@gmail.com went into our Wimbledon store on Saturday. Funny enough, I noticed he was on our website on Thursday for 45 minutes, researching products.
>
> Obviously, my digital analytics regard that as an abandoned basket—fail—low conversion rate and my in-store manager goes, "Gosh, he is a great guy, he came in and spent five hundred quid! Love him to bits!"

> We could show, in this particular instance, that for every Pound of sales the website thought they made, we could see two Pounds in-store. That was the end of the official project with the start of the science project. That's when we started playing with machine learning.

Even with the most reliable data, getting it all to make sense is still troubling.

ARE WE REALLY CALCULABLE?

> *While the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty.*
>
> *Sherlock Holmes*, The Sign of Four

On the BBC show *Sherlock*, Mary asked how Sherlock Holmes had managed to find her and the flash drive she was carrying around when, "Every movement I made was entirely random, every new personality just on the roll of a dice!" Sherlock replied:

> Mary, no human action is ever truly random. An advanced grasp of the mathematics of probability, mapped onto a thorough apprehension of human psychology and the known dispositions of any given individual, can reduce the number of variables considerably. I myself know of at least 58 techniques to refine a seemingly infinite array of randomly generated possibilities down to the smallest number of feasible variables.

After a brief pause, he admitted, "But they're really difficult, so instead I just stuck a tracer on the inside of the memory stick."[43]

This, then, is our task: to use the big data and machine learning tools we have at hand to see if we can't build a better, more useful model of individual, human probabilities in order to send the right message to the right person at the right time on the right device. Sherlock is right; it is difficult.

So, now you understand the *idea* of machine learning. You know just enough to hold your own at a cocktail party. You can nod knowingly should the topic pop up and can comfortably converse with senior management about the possibilities.

The next chapter is intended to go one level deeper. You will *not* become a data scientist by careful study of Chapter 2, but you *will* be able to hold your own at a meeting on machine learning. You can nod knowingly should the subject matter get deeper and will be able to comfortably converse with data scientists about the possibilities.

NOTES

1. "Improving Our Ability to Improve," http://www.almaden.ibm.com/coevolution/pdf/engelbart_paper.pdf.
2. "How to Get the Best Deep Learning Education for Free," http://www.topbots.com/artificial-intelligence-deep-learning-education-free.
3. "24 Uses of Statistical Modeling (Part II)," http://www.datasciencecentral.com/profiles/blogs/24-uses-of-statistical-modeling-part-ii.
4. "The Discipline of Machine Learning," http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf.
5. http://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world.
6. Source: http://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world.
7. Source: Personal interview.
8. "The History of Artificial Intelligence," http://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf.
9. "An Executive's Guide to Cognitive Computing," http://www.sas.com/en_us/insights/articles/big-data/executives-guide-to-cognitive-computing.html.
10. "DeepStack: Expert-Level Artificial Intelligence in No-Limit Poker," https://arxiv.org/abs/1701.01724.
11. EU GDPR Portal, http://www.eugdpr.org.
12. "9 Artificial Intelligence Stats that Will Blow You Away," http://www.foxbusiness.com/markets/2016/12/10/artificial-intelligence-stats-that-will-blow-away.html.
13. "Preparing for the Future of Artificial Intelligence," https://www.whitehouse.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.
14. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.
15. "Machine Learning Will Revolutionize Market Segmentation Practices," January 2017, http://www.idgconnect.com/view_abstract/41712/machine-learning-will-revolutionize-market-segmentation-practices.
16. http://www.bbc.com/news/technology-30290540.
17. "Benefits & Risks of Artificial Intelligence," http://futureoflife.org/background/benefits-risks-of-artificial-intelligence/.
18. "The AI Revolution: Our Immortality or Extinction," http://waitbutwhy.com/2015/01/artificial-intelligence-revolution-2.html.
19. "Autonomous Weapons: An Open Letter from AI & Robotics Researchers," http://futureoflife.org/open-letter-autonomous-weapons.

20. https://openai.com/about.
21. "Safely Interruptible Agents," http://intelligence.org/files/Interruptibility.pdf.
22. Partnership on Artificial Intelligence to Benefit People and Society, https://www.partnershiponai.org/.
23. The Partnership on AI's stated tenets, https://www.partnershiponai.org/tenets.
24. *Wall Street Journal,* http://www.wsj.com/articles/the-worlds-largest-hedge-fund-is-building-an-algorithmic-model-of-its-founders-brain-1482423694.
25. *Guardian,* https://www.theguardian.com/technology/2017/jan/05/japanese-company-replaces-office-workers-artificial-intelligence-ai-fukoku-mutual-life-insurance?CMP=Share_iOSApp_Other.
26. "Will a Robot Take Your Job?" http://www.bbc.com/news/technology-34066941.
27. "A Future That Works: Automation, Employment, and Productivity," http://www.mckinsey.com/global-themes/digital-disruption/harnessing-automation-for-a-future-that-works.
28. Source: Personal interview.
29. "Probabilistic Noise Identification and Data Cleaning," http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.71.4154&rep=rep1&type=pdf.
30. Originally published in *Applied Marketing Analytics Journal*, Vol. 1, No. 3, reproduced with permission from Henry Stewart Publications LLP, https://www.henrystewartpublications.com/ama/v1.
31. "Math Is Fun," https://www.mathsisfun.com/probability_line.html.
32. "Private Traits and Attributes Are Predictable from Digital Records of Human Behavior," http://www.pnas.org/content/110/15/5802.full.pdf.
33. Campbell Soup API Developer Portal, https://developer.campbellskitchen.com.
34. Google Big Query datasets, https://cloud.google.com/bigquery/public-data/.
35. Data Science Central, "Top 20 Open Data Sources," http://www.datasciencecentral.com/profiles/blogs/top-20-open-data-sources.
36. "How 4 Companies Find and Create Value from Open Data," https://blogs.oracle.com/marketingcloud/create-value-from-open-data.
37. Iowa Liquor Sales, https://www.reddit.com/r/bigquery/comments/37fcm6/iowa_liquor_sales_dataset_879mb_3million_rows.
38. "Big Data Analytics in Supply Chain Management: Trends and Related Research," https://www.researchgate.net/publication/270506965_Big_Data_Analytics_in_Supply_Chain_Management_Trends_and_Related_Research.
39. "Why Acxiom Data?" http://www.acxiom.com/data-solutions/.
40. https://shift.newco.co/as-we-may-think-data-world-lays-the-traceroutes-for-a-data-revolution-b4b751f295d9.
41. "The Origins of Personal Data and Its Implications for Governance," http://informationaccountability.org/wp-content/uploads/Data-Origins-Abrams.pdf.
42. Source: Personal interview.
43. http://www.bbc.co.uk/programmes/b0881dgp.