Introduction

The Prediction Effect

I'm just like you. I succeed at times, and at others I fail. Some days good things happen to me, some days bad. We always wonder how things could have gone differently. I begin with seven brief tales of woe:

1. In 2009 I just about destroyed my right knee downhill skiing in Utah. The jump was no problem; it was landing that presented an issue. For knee surgery, I had to pick a graft source from which to reconstruct my busted ACL (the knee's central ligament). The choice is a tough one and can make the difference between living with a good knee or a bad knee. I went with my hamstring. *Could the hospital have selected a medically better option for my case?*
2. Despite all my suffering, it was really my health insurance company that paid dearly—knee surgery is expensive. *Could the company have better anticipated the risk of accepting a ski jumping fool as a customer and priced my insurance premium accordingly?*
3. Back in 1995 another incident caused me suffering, although it hurt less. I fell victim to identity theft, costing me dozens of hours of bureaucratic baloney and tedious paperwork to clear up my damaged credit rating. *Could the creditors have prevented the fiasco by detecting*

that the accounts were bogus when they were filed under my name in the first place?

4. With my name cleared, I recently took out a mortgage to buy an apartment. Was it a good move, or *should my financial adviser have warned me the property could soon be outvalued by my mortgage?*
5. While embarking on vacation, I asked the neighboring airplane passenger what price she'd paid for her ticket, and it was much less than I'd paid. *Before I booked the flight, could I have determined the airfare was going to drop?*
6. My professional life is susceptible, too. My business is faring well, but a company always faces the risk of changing economic conditions and growing competition. *Could we protect the bottom line by foreseeing which marketing activities and other investments will pay off, and which will amount to burnt capital?*
7. Small ups and downs determine your fate and mine, every day. A precise spam filter has a meaningful impact on almost every working hour. We depend heavily on effective Internet search for work, health (e.g., exploring knee surgery options), home improvement, and most everything else. We put our faith in personalized music and movie recommendations from Spotify and Netflix. After all these years, my mailbox wonders why companies don't know me well enough to send less junk mail (and sacrifice fewer trees needlessly).

These predicaments matter. They can make or break your day, year, or life. But what do they all have in common?

These challenges—and many others like them—are best addressed with *prediction*. Will the patient's outcome from surgery be positive? Will the credit applicant turn out to be a fraudster? Will the homeowner face a bad mortgage? Will the airfare go down? Will the customer respond if mailed a brochure? By predicting these things, it is possible to fortify healthcare, combat risk, conquer spam, toughen crime fighting, boost sales, and cut costs.

PREDICTION IN BIG BUSINESS—THE DESTINY OF ASSETS

There's another angle. Beyond benefiting you and me as consumers, prediction serves the organization, empowering it with an entirely new form of competitive armament. Corporations positively pounce on prediction.

In the mid-1990s, an entrepreneurial scientist named Dan Steinberg delivered predictive capabilities unto the nation's largest bank, Chase, to assist with their management of millions of mortgages. This mammoth enterprise put its faith in Dan's predictive technology, deploying it to drive transactional decisions across a tremendous mortgage portfolio. What did this guy have on his résumé?

Prediction is power. Big business secures a killer competitive stronghold by predicting the future destiny and value of individual assets. In this case, by driving mortgage decisions with predictions about the future payment behavior of homeowners, Chase curtailed risk, boosted profit, and witnessed a windfall.

INTRODUCING . . . THE CLAIRVOYANT COMPUTER

Compelled to grow and propelled to the mainstream, predictive technology is commonplace and affects everyone, every day. It impacts your experiences in undetectable ways as you drive, shop, study, vote, see the doctor, communicate, watch TV, earn, borrow, or even steal.

This book is about the most influential and valuable achievements of computerized prediction, and the two things that make it possible: the people behind it, and the fascinating science that powers it.

Making such predictions poses a tough challenge. Each prediction depends on multiple factors: The various characteristics known about each patient, each homeowner, each consumer, and each e-mail that may be spam. How shall we attack the intricate problem of putting all these pieces together for each prediction?

The idea is simple, although that doesn't make it easy. The challenge is tackled by a systematic, scientific means to develop and continually improve prediction—to literally *learn* to predict.

The solution is *machine learning*—computers automatically developing new knowledge and capabilities by furiously feeding on modern society's greatest and most potent *unnatural* resource: data.

"FEED ME!"—FOOD FOR THOUGHT FOR THE MACHINE

Data is the new oil.

—European Consumer Commissioner Meglena Kuneva

The only source of knowledge is experience.

—Albert Einstein

In God we trust. All others must bring data.

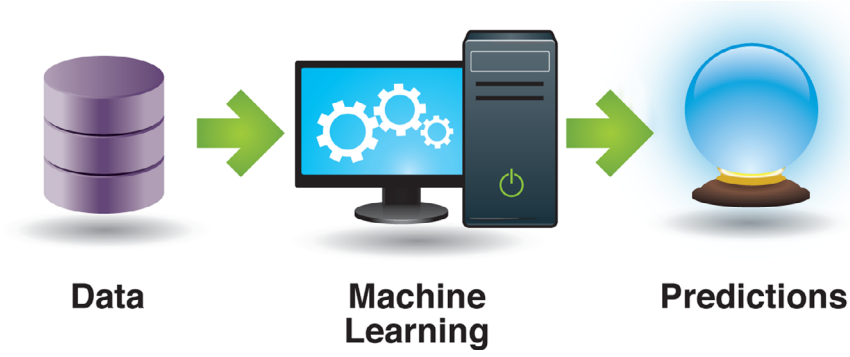
—William Edwards Deming (a business professor famous for work in manufacturing)

Most people couldn't be less interested in data. It can seem like such dry, boring stuff. It's a vast, endless regimen of recorded facts and figures, each alone as mundane as the most banal tweet, "I just bought some new sneakers!" It's the unsalted, flavorless residue deposited en masse as businesses churn away.

Don't be fooled! The truth is that data embodies a priceless collection of experience from which to learn. Every medical procedure, credit application, Facebook post, movie recommendation, fraudulent act, spammy e-mail, and purchase of any kind—each positive or negative outcome, each successful or failed sales call, each incident, event, and transaction—is encoded as data and warehoused. This glut grows by an estimated 2.5 quintillion bytes per day (that's a 1 with 18 zeros after it). And so a veritable Big Bang has set off, delivering an epic sea of raw materials, a plethora of examples so great in number, only a computer could manage to learn from them. Used correctly, computers avidly soak up this ocean like a sponge.

As data piles up, we have ourselves a genuine gold rush. But data isn't the gold. I repeat, data in its raw form is boring crud. The gold is what's discovered therein.

The process of machines learning from data unleashes the power of this exploding resource. It uncovers what drives people and the actions they take—what makes us tick and how the world works. With the new knowledge gained, prediction is possible.



This learning process discovers insightful gems such as:[1]

- Early retirement decreases your life expectancy.
- Online daters more consistently rated as attractive receive *less* interest.
- Rihanna fans are mostly political Democrats.
- Vegetarians miss fewer flights.
- Local crime increases after public sporting events.

Machine learning builds upon insights such as these in order to develop predictive capabilities, following a number-crunching, trial-and-error process that has its roots in statistics and computer science.

[1] See Chapter 3 for more details on these examples.

I KNEW YOU WERE GOING TO DO THAT

With this power at hand, what do we want to predict? Every important thing a person does is valuable to predict, namely: *consume*, *think*, *work*, *quit*, *vote, love*, *procreate*, *divorce*, *mess up*, *lie*, *cheat*, *steal*, *kill*, and *die*. Let's explore some examples.[2]

PEOPLE CONSUME

- Hollywood studios predict the success of a screenplay if produced.
- Netflix awarded $1 million to a team of scientists who best improved their recommendation system's ability to predict which movies you will like.
- The Hopper app helps you get the best deal on a flight by recommending whether you should buy or wait, based on its prediction as to whether the airfare will change.
- Australian energy company Energex predicts electricity demand in order to decide where to build out its power grid, and Con Edison predicts system failure in the face of high levels of consumption.
- Wall Street firms trade algorithmically, buying and selling based on the prediction of stock prices.
- Companies predict which customer will buy their products in order to target their marketing, from U.S. Bank down to small companies like Harbor Sweets (candy) and Vermont Country Store ("top quality and hard-to-find classic products"). These predictions dictate the allocations of precious marketing budgets. Some companies literally predict how to best influence you to buy more (the topic of Chapter 7).
- Prediction drives the coupons you get at the grocery cash register. U.K. grocery giant Tesco, the world's third-largest retailer, predicts which discounts will be redeemed in order to target more than

[2] For more examples and further detail, see this book's Central Tables.

100 million personalized coupons annually at cash registers across 13 countries. Similarly, Kmart, Kroger, Ralph's, Safeway, Stop & Shop, Target, and Winn-Dixie follow in kind.

- Predicting mouse clicks pays off massively. Since websites are often paid per click for the advertisements they display, they predict which ad you're mostly likely to click in order to instantly choose which one to show you. This, in effect, selects more relevant ads and drives millions in newly found revenue.
- Facebook predicts which of the thousands of posts by your friends will interest you most every time you view the news feed (unless you change the default setting). The social network also predicts the suggested "people you may know," not to mention which ads you're likely to click.

People Love, Work, Procreate, and Divorce

- The leading career-focused social network, LinkedIn, predicts your job skills.
- Online dating leaders Match.com, OkCupid, and eHarmony predict which hottie on your screen would be the best bet at your side.
- Target predicts customer pregnancy in order to market relevant products accordingly. Nothing foretells consumer need like predicting the birth of a new consumer.
- Clinical researchers predict infidelity and divorce. There's even a self-help website tool to put odds on your marriage's long-term success (www.divorceprobability.com).

People Think and Decide

- Obama was reelected in 2012 with the help of voter prediction. The Obama for America campaign predicted which voters would be positively persuaded by campaign contact (a call, door knock, flier, or TV ad), and which would actually be inadvertently influenced to

(*continued*)

(*continued*)
vote adversely by contact. Employed to drive campaign decisions for millions of swing state voters, this method was shown to successfully convince more voters to choose Obama than traditional campaign targeting. Hillary for America 2016 is positioning to apply the same technique.

- "What did you mean by that?" Systems have learned to ascertain the intent behind the written word. Citibank and PayPal detect the customer sentiment about their products, and one researcher's machine can tell which Amazon.com book reviews are sarcastic.
- Student essay grade prediction has been developed for possible use to automatically grade. The system grades as accurately as human graders.
- There's a machine that can participate in the same capacity as humans in the United States' most popular broadcast celebration of human knowledge and cultural literacy. On the TV quiz show *Jeopardy!*, IBM's Watson computer triumphed. This machine learned to work proficiently enough with English to predict the answers to free-form inquiries across an open range of topics and defeat the two all-time human champs.
- Computers can literally read your mind. Researchers trained systems to decode a scan of your brain and determine which type of object you're thinking about—such as certain tools, buildings, and food—with over 80 percent accuracy for some human subjects.

People Quit

- Hewlett-Packard (HP) earmarks each and every one of its more than 300,000 worldwide employees according to "Flight Risk," the expected chance he or she will quit their job, so that managers may intervene in advance where possible and plan accordingly otherwise.
- Ever experience frustration with your cell phone service? Your service provider endeavors to know. All major wireless carriers

predict how likely it is you will cancel and switch to a competitor—possibly before you have even conceived a plan to do so—based on factors such as dropped calls, your phone usage, billing information, and whether your contacts have already defected.

- FedEx stays ahead of the game by predicting—with 65 to 90 percent accuracy—which customers are at risk of defecting to a competitor.
- The American Public University System predicted student dropouts and used these predictions to intervene successfully; the University of Alabama, Arizona State University, Iowa State University, Oklahoma State University, and the Netherlands' Eindhoven University of Technology predict dropouts as well.
- Wikipedia predicts which of its editors, who work for free as a labor of love to keep this priceless online asset alive, are going to discontinue their valuable service.
- Researchers at Harvard Medical School predict that if your friends stop smoking, you're more likely to do so yourself as well. Quitting smoking is contagious.

PEOPLE MESS UP

- Insurance companies predict who is going to crash a car or hurt themselves another way (such as a ski accident). Allstate predicts bodily injury liability from car crashes based on the characteristics of the insured vehicle, demonstrating improvements to prediction that could be worth an estimated $40 million annually. Another top insurance provider reported savings of almost $50 million per year by expanding its actuarial practices with advanced predictive techniques.
- Ford is learning from data so its cars can detect when the driver is not alert due to distraction, fatigue, or intoxication and take action such as sounding an alarm.
- Researchers have identified aviation incidents that are five times more likely than average to be fatal, using data from the National Transportation Safety Board.

(*continued*)

(*continued*)

- All large banks and credit card companies predict which debtors are most likely to turn delinquent, failing to pay back their loans or credit card balances. Collection agencies prioritize their efforts with predictions of which tactic has the best chance to recoup the most from each defaulting debtor.

People Get Sick and Die

I'm not afraid of death; I just don't want to be there when it happens.

—Woody Allen

- In 2013, the Heritage Provider Network handed over $500,000 to a team of scientists who won an analytics competition to best predict individual hospital admissions. By following these predictions, proactive preventive measures can take a healthier bite out of the tens of billions of dollars spent annually on unnecessary hospitalizations. Similarly, the University of Pittsburgh Medical Center predicts short-term hospital readmissions, so doctors can be prompted to think twice before a hasty discharge.
- At Stanford University, a machine learned to diagnose breast cancer better than human doctors by discovering an innovative method that considers a greater number of factors in a tissue sample.
- Researchers at Brigham Young University and the University of Utah correctly predict about 80 percent of premature births (and about 80 percent of full-term births), based on peptide biomarkers, as found in a blood exam as early as week 24 of pregnancy.
- University researchers derived a method to detect patient schizophrenia from transcripts of their spoken words alone.
- A growing number of life insurance companies go beyond conventional actuarial tables and employ predictive technology to establish mortality risk. It's not called *death insurance*, but they calculate when you are going to die.

- Beyond life insurance, one top-five *health* insurance company predicts the probability that elderly insurance policyholders will pass away within 18 months, based on clinical markers in the insured's recent medical claims. Fear not—it's actually done for benevolent purposes.
- Researchers predict your risk of death in surgery based on aspects of you and your condition to help inform medical decisions.
- By following one common practice, doctors regularly—yet unintentionally—sacrifice some patients in order to save others, and this is done completely without controversy. But this would be lessened by predicting something besides diagnosis or outcome: healthcare *impact* (impact prediction is the topic of Chapter 7).

PEOPLE LIE, CHEAT, STEAL, AND KILL

- Most medium-size and large banks employ predictive technology to counter the ever-blooming assault of fraudulent checks, credit card charges, and other transactions. Citizens Bank developed the capacity to decrease losses resulting from check fraud by 20 percent. Hewlett-Packard saved $66 million by detecting fraudulent warranty claims.
- Predictive computers help decide who belongs in prison. To assist with parole and sentencing decisions, officials in states such as Oregon and Pennsylvania consult prognostic machines that assess the risk a convict will offend again.
- Murder is widely considered impossible to predict with meaningful accuracy in general, but within at-risk populations predictive methods can be effective. Maryland analytically generates predictions as to which inmates will kill or be killed. University and law enforcement researchers have developed predictive systems that foretell murder among those previously convicted for homicide.
- One fraud expert at a large bank in the United Kingdom extended his work to discover a small pool of terror suspects based on their

(*continued*)

(*continued*)
banking activities. While few details have been disclosed publicly, it's clear that the National Security Agency also considers this type of analysis a strategic priority in order to automatically discover previously unknown potential suspects.

- Police patrol the areas predicted to spring up as crime hot spots in cities such as Chicago, Memphis, and Richmond, Va.
- Inspired by the TV crime drama *Lie to Me* about a microexpression reader, researchers at the University at Buffalo trained a system to detect lies with 82 percent accuracy by observing eye movements alone.
- As a professor at Columbia University in the late 1990s, I had a team of teaching assistants who employed cheating-detection software to patrol hundreds of computer programming homework submissions for plagiarism.
- The IRS predicts if you are cheating on your taxes.

THE LIMITS AND POTENTIAL OF PREDICTION

An economist is an expert who will know tomorrow why the things he predicted yesterday didn't happen.

—Earl Wilson

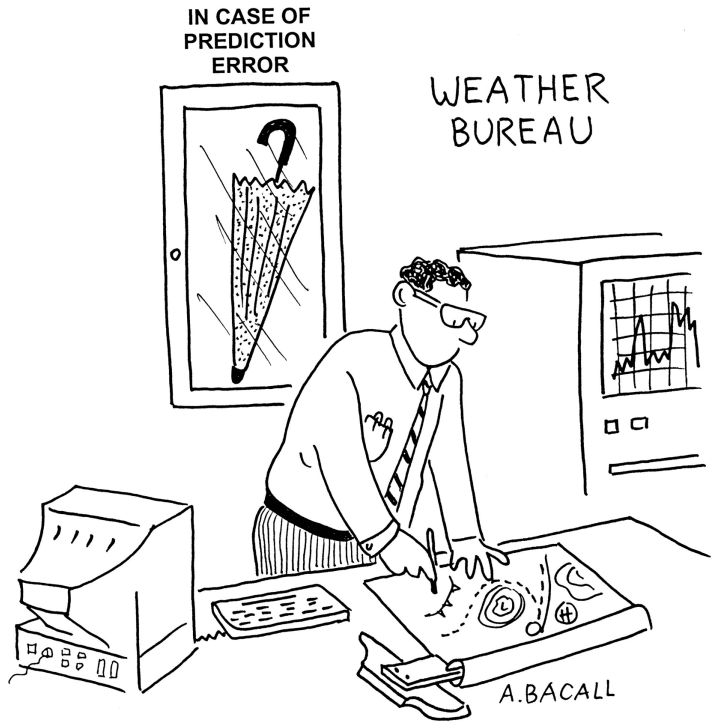
How come you never see a headline like "Psychic Wins Lottery"?

—Jay Leno

Each of the preceding accomplishments is powered by prediction, which is in turn a product of machine learning. A striking difference exists between these varied capabilities and science fiction: They aren't fiction. At this point, I predict that you won't be surprised to hear that those examples represent

only a small sample. You can safely predict that the power of prediction is here to stay.

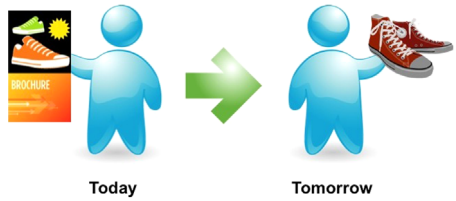
But are these claims too bold? As the Danish physicist Niels Bohr put it, "Prediction is very difficult, especially if it's about the future." After all, isn't prediction basically impossible? The future is unknown, and uncertainty is the only thing about which we're certain.



Let me be perfectly clear. It's fuzzy. Accurate prediction is generally not possible. The weather is predicted with only about 50 percent accuracy, and it doesn't get easier predicting the behavior of humans, be they patients, customers, or criminals.

Good news! Predictions need not be accurate to score big value. For instance, one of the most straightforward commercial applications of

predictive technology is deciding whom to target when a company sends direct mail. If the learning process identifies a carefully defined group of customers who are predicted to be, say, three times more likely than average to respond positively to the mail, the company profits big-time by preemptively removing likely *nonresponders* from the mailing list. And those nonresponders in turn benefit, contending with less junk mail.



Prediction—A person who sees a sales brochure today buys a product tomorrow.

In this way the business, already playing a sort of numbers game by conducting mass marketing in the first place, tips the balance delicately yet significantly in its favor—and does so without highly accurate predictions. In fact, its utility withstands quite poor accuracy. If the overall marketing response is at 1 percent, the so-called hot pocket with three times as many would-be responders is at 3 percent. So, in this case, we can't confidently predict the response of any one particular customer. Rather, the value is derived from identifying a group of people who—in aggregate—will tend to behave in a certain way.

This demonstrates in a nutshell what I call *The Prediction Effect*. Predicting better than pure guesswork, even if not accurately, delivers real value. A hazy view of what's to come outperforms complete darkness by a landslide.

The Prediction Effect: *A little prediction goes a long way.*

This is the first of five Effects introduced in this book. You may have heard of the butterfly, Doppler, and placebo effects. Stay tuned here for the *Data*, *Induction*, *Ensemble*, and *Persuasion Effects*. Each of these Effects encompasses the fun part of science and technology: an intuitive hook that reveals how it works and why it succeeds.

THE FIELD OF DREAMS

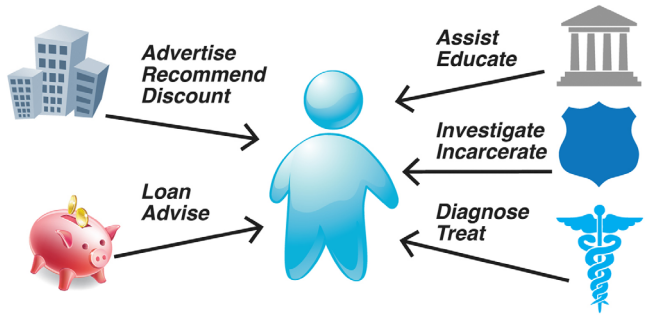
> *People . . . operate with beliefs and biases. To the extent you can eliminate both and replace them with data, you gain a clear advantage.*
>
> —Michael Lewis, *Moneyball: The Art of Winning an Unfair Game*

What field of study or branch of science are we talking about here? Learning how to predict from data is sometimes called *machine learning*—but it turns out this is mostly an academic term you find used within research labs, conference papers, and university courses (full disclosure: I taught the Machine Learning graduate course at Columbia University a couple of times in the late 1990s). These arenas are a priceless wellspring, but they aren't where the rubber hits the road. In commercial, industrial, and government applications—in the real-world usage of machine learning to predict—it's called something else, something that in fact is the very topic of this book:

> ***Predictive analytics (PA)***—*Technology that learns from experience (data) to predict the future behavior of individuals in order to drive better decisions.*[3]

[3] In this definition, *individuals* is a broad term that can refer to people as well as other organizational elements. Most examples in this book involve predicting people, such as customers, debtors, applicants, employees, students, patients, donors, voters, taxpayers, potential suspects, and convicts. However, PA also applies to individual companies (e.g., for business-to-business), products, locations, restaurants, vehicles, ships, flights, deliveries, buildings, manholes, transactions, Facebook posts, movies, satellites, stocks, *Jeopardy!* questions, and much more. Whatever the domain, PA renders predictions over scalable numbers of individuals.

Built upon computer science and statistics and bolstered by devoted conferences and university degree programs, PA has emerged as its own discipline. But beyond a field of science, PA is a movement that exerts a forceful impact. Millions of decisions a day determine whom to call, mail, approve, test, diagnose, warn, investigate, incarcerate, set up on a date, and medicate. PA is the means to drive *per-person* decisions empirically, as guided by data. By answering this mountain of smaller questions, PA may in fact answer the biggest question of all: *How can we improve the effectiveness of all these massive functions across government, healthcare, business, nonprofit, and law enforcement work?*



Predictions drive how organizations treat and serve an individual, across the frontline operations that define a functional society.

In this way, PA is a completely different animal from *forecasting*. Forecasting makes aggregate predictions on a macroscopic level. How will the economy fare? Which presidential candidate will win more votes in Ohio? Whereas forecasting estimates the total number of ice cream cones to be purchased next month in Nebraska, PA tells you which *individual* Nebraskans are most likely to be seen with cone in hand.

PA leads within the growing trend to make decisions more "data driven," relying less on one's "gut" and more on hard, empirical evidence. Enter this fact-based domain and you'll be attacked by buzzwords, including *analytics*, *big data*, *data science,* and *business intelligence*. While PA fits

underneath each of these umbrellas, these evocative terms refer more to the culture and general skill sets of technologists who do an assortment of creative, innovative things with data, rather than alluding to any specific technology or method. These areas are broad; in some cases, they refer simply to standard Excel reports—that is, to things that are important and require a great deal of craft, but may not rely on science or sophisticated math. And so they are more subjectively defined. As Mike Loukides, a vice president at the innovation publisher O'Reilly, once put it, "Data science is like porn—you know it when you see it." Another term, *data mining*, is often used as a synonym for PA, but as an evocative metaphor depicting "digging around" through data in one fashion or another, it is often used more broadly as well.

ORGANIZATIONAL LEARNING

> *The powerhouse organizations of the Internet era, which include Google and Amazon . . . have business models that hinge on predictive models based on machine learning.*
>
> —Professor Vasant Dhar, Stern School of Business, New York University

> *A breakthrough in machine learning would be worth 10 Microsofts.*
>
> —Bill Gates

An organization is sort of a "megaperson," so shouldn't it "megalearn"? A group comes together for the collective benefit of its members and those it serves, be it a company, government, hospital, university, or charity. Once formed, it gains from division of labor, mutually complementary skills, and the efficiency of mass production. The result is more powerful than the sum of its parts. Collective learning is the organization's next logical step to further leverage this power. Just as a salesperson learns over time from her positive and negative interactions with sales leads, her successes, and failures, PA is the process by which an organization learns from the experience it has

collectively gained across its team members and computer systems. In fact, an organization that doesn't leverage its data in this way is like a person with a photographic memory who never bothers to think.

With only a few striking exceptions, we find that organizations, rather than individuals, benefit by employing PA. Organizations make the many, many operational decisions for which there's ample room for improvement; organizations are intrinsically inefficient and wasteful on a grand scale. Marketing casts a wide net—junk mail is marketing money wasted and trees felled to print unread brochures. An estimated 80 percent of all e-mail is spam. Risky debtors are given too much credit. Applications for government benefits are backlogged and delayed. And it's organizations that have the data to power the predictions that drive improvements in these operations.

In the commercial sector, profit is a driving force. You can well imagine the booming incentives intrinsic to rendering everyday routines more efficient, marketing more precisely, catching more fraud, avoiding bad debtors, and luring more online customers. Upgrading how business is done, PA rocks the enterprise's economies of scale, optimizing operations right where it makes the biggest difference.

THE NEW SUPER GEEK: DATA SCIENTISTS

> *The alternative [to thinking ahead] would be to think backwards . . . and that's just remembering.*
>
> —Sheldon, the theoretical physicist on *The Big Bang Theory*

Opportunities abound, but the profit incentive is not the only driving force. The source, the energy that makes it work, is Geek Power! I speak of the enthusiasm of technical practitioners. Truth be told, my passion for PA didn't originate from its value to organizations. I am in it for the fun. The idea of a machine that can actually learn seems so cool to me that I care more about what happens inside the magic box than its outer usefulness.

Indeed, perhaps that's the defining motivator that qualifies one as a geek. We love the technology; we're in awe of it. Case in point: The leading free, open-source software tool for PA, called R (a one-letter, geeky name), has a rapidly expanding base of users as well as enthusiastic volunteer developers who add to and support its functionalities. Great numbers of professionals and amateurs alike flock to public PA competitions with a tremendous spirit of "coopetition." We operate within organizations, or consult across them. We're in demand, so we fly a lot. But we fly coach, at best Economy Plus.

THE ART OF LEARNING

Whatcha gonna do with your CPU to reach its potentiality?
Use your noggin when you log in to crank it exponentially.
The endeavor that will render my obtuse computer clever:
Self-improve impeccably by way of trial and error.

Once upon a time, humanity created The Ultimate General Purpose Machine and, in an inexplicable fit of understatement, decided to call it "a computer" (a word that until this time had simply meant a person who did computations by hand). This automaton could crank through any demanding, detailed set of endless instructions without fail or error and with nary a complaint; within just a few decades, its speed became so blazingly brisk that humanity could only exclaim, "Gosh, we really cranked that!" An obviously much better name for this device would have been the appropriately grand *La Machine*, but a few decades later this name was hyperbolically bestowed upon a food processor (I am not joking). *Quel dommage*. "What should we do with the computer? What's its true potential, and how do we achieve it?" humanity asked of itself in wonderment.

A computer and your brain have something in common that renders them both mysterious, yet at the same time easy to take for granted. If while

pondering what this might be you heard a pin drop, you have your answer. They are both silent. Their mechanics make no sound. Sure, a computer may have a disk drive or cooling fan that stirs—just as one's noggin may emit wheezes, sneezes, and snores—but the mammoth grunt work that takes place therein involves no "moving parts," so these noiseless efforts go along completely unwitnessed. The smooth delivery of content on your screen—and ideas in your mind—can seem miraculous.[4]

They're both powerful as heck, your brain and your computer. So could computers be successfully programmed to think, feel, or become truly intelligent? Who knows? At best these are stimulating philosophical questions that are difficult to answer, and at worst they are subjective benchmarks for which success could never be conclusively established. But thankfully we do have some clarity: There is one truly impressive, profound human endeavor computers *can* undertake. They can learn.

But how? It turns out that learning—generalizing from a list of examples, be it a long list or a short one—is more than just challenging. It's a philosophically deep dilemma. Machine learning's task is to find patterns that appear not only in the data at hand, but in general, so that what is learned will hold true in new situations never yet encountered. At the core, this ability to generalize is the magic bullet of PA. There is a true art in the design of these computer methods. We'll explore more later, but for now I'll give you a hint. The machine actually learns more about your next likely action by studying *others* than by studying *you*.

While I'm dispensing teasers that leave you hanging, here's one more. This book's final chapter answers the riddle: *What often happens to you that*

---

[4] Silence is characteristic to solid state electronics, but computers didn't have to be built that way. The idea of a general-purpose, instruction-following machine is abstract, not affixed to the notion of electricity. You could construct a computer of cogs and wheels and levers, powered by steam or gasoline. I mean, I wouldn't recommend it, but you could. It would be slow, big, and loud, and nobody would buy it.

cannot be witnessed, and that you can't even be sure has happened afterward—but that can *be predicted in advance?*

Learning from data to predict is only the first step. To take the next step and *act on predictions* is to fearlessly gamble. Let's kick off Chapter 1 with a suspenseful story that shows why launching PA feels like blasting off in a rocket.

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

CHAPTER 1

Liftoff! Prediction Takes Action

How much guts does it take to deploy a predictive model into field operation, and what do you stand to gain? What happens when a man invests his entire life savings into his own predictive stock market trading system? Launching predictive analytics means to act on its predictions, applying what's been learned, what's been discovered within data. It's a leap many take—you can't win if you don't play.

In the mid-1990s, an ambitious postdoc researcher couldn't stand to wait any longer. After consulting with his wife, he loaded their entire life savings into a stock market prediction system of his own design—a contraption he had developed moonlighting on the side. Like Dr. Henry Jekyll imbibing his own untested potion in the moonlight, the young Dr. John Elder unflinchingly pressed "go."

There is a scary moment every time new technology is launched. A spaceship lifting off may be the quintessential portrait of technological greatness and national prestige, but the image leaves out a small group of spouses terrified to the very point of psychological trauma. Astronauts are in essence stunt pilots, voluntarily strapping themselves in to serve as guinea pigs for a giant experiment, willing to sacrifice themselves in order to be part of history.

From grand challenges are born great achievements. We've taken strolls on our moon, and in more recent years a $10 million Grand Challenge prize was awarded to the first nongovernmental organization to develop a reusable manned spacecraft. Driverless cars have been unleashed—"Look, Ma, no hands!" Fueled as well by millions of dollars in prize money, they navigate autonomously around the campuses of Google and BMW.

Replace the roar of rockets with the crunch of data, and the ambitions are no less far-reaching, "boldly going" not to space but to a new final

frontier: predicting the future. This frontier is just as exciting to explore, yet less dangerous and uncomfortable (outer space is a vacuum, and vacuums totally suck). Millions in grand challenge prize money go toward averting the unnecessary hospitalization of each patient and predicting the idiosyncratic preferences of each individual consumer. The TV quiz show *Jeopardy!* awarded $1.5 million in prize money for a face-off between man and machine that demonstrated dramatic progress in predicting the answers to questions (IBM invested a lot more than that to achieve this win, as detailed in Chapter 6). Organizations are literally keeping kids in school, keeping the lights on, and keeping crime down with predictive analytics (PA). And success is its own reward when analytics wins a political election, a baseball championship, or . . . did I mention managing a financial portfolio?

Black-box trading—driving financial trading decisions automatically with a machine—is the holy grail of data-driven decision making. It's a black box into which current financial environmental conditions are fed, with buy/hold/sell decisions spit out the other end. It's black (i.e., opaque) because you don't care what's on the inside, as long as it makes good decisions. When working, it trumps any other conceivable business proposal in the world: Your computer is now a box that turns electricity into money.

And so with the launch of his stock trading system, John Elder took on his own personal grand challenge. Even if stock market prediction would represent a giant leap for mankind, this was no small step for John himself. It's an occasion worthy of mixing metaphors. By putting all his eggs into one analytical basket, John was taking a healthy dose of his own medicine.

Before continuing with the story of John's blast-off, let's establish how launching a predictive system works, not only for black-box trading but across a multitude of applications.

GOING LIVE

Learning from data is virtually universally useful. Master it and you'll be welcomed nearly everywhere!

—John Elder

New groundbreaking stories of PA in action are pouring in. A few key ingredients have opened these floodgates:

- wildly increasing loads of data;
- cultural shifts as organizations learn to appreciate, embrace, and integrate predictive technology;
- improved software solutions to deliver PA to organizations.

But this flood built up its potential in the first place simply because predictive technology boasts an inherent generality—there are just so many conceivable ways to make use of it. Want to come up with your own new innovative use for PA? You need only two ingredients.

EACH APPLICATION OF PA IS DEFINED BY:

1. **What's predicted:** the kind of behavior (i.e., action, event, or happening) to predict for each individual, stock, or other kind of element.
2. **What's done about it:** the decisions driven by prediction; the action taken by the organization in response to or informed by each prediction.

Given its open-ended nature, the list of application areas is so broad and the list of example stories is so long that it presents a minor data-management challenge in and of itself! So I placed this big list (182 examples total) into nine tables in the center of this book. Take a flip through to get a feel for just how much is going on. That's the sexy part—it's the "centerfold" of this book. The Central Tables divulge cases of predicting: stock prices, risk, delinquencies, accidents, sales, donations, clicks, cancellations, health problems, hospital admissions, fraud, tax evasion, crime, malfunctions, oil flow, electricity outages, approvals for government benefits, thoughts, intention, answers, opinions, lies, grades, dropouts, friendship, romance, pregnancy, divorce, jobs, quitting, wins, votes, and more. The application areas are growing at a breakneck pace.

Within this long list, the quintessential application for business is the one covered in the Introduction for mass marketing:

PA APPLICATION: TARGETING DIRECT MARKETING

1. **What's predicted:** Which customers will respond to marketing contact.
2. **What's done about it:** Contact customers more likely to respond.

As we saw, this use of PA illustrates *The Prediction Effect*.

The Prediction Effect: *A little prediction goes a long way.*

Let's take a moment to see how straightforward it is to calculate the sheer value resulting from The Prediction Effect. Imagine you have a company with a mailing list of a million prospects. It costs $2 to mail to each one, and you have observed that one out of 100 of them will buy your product (i.e., 10,000 responses). You take your chances and mail to the entire list.

If you profit $220 for each rare positive response, then you pocket:

$$\begin{aligned}\text{Overall profit} &= \text{Revenue} - \text{Cost} \\ &= (\$220 \times 10,000 \text{ responses}) - (\$2 \times 1 \text{ million})\end{aligned}$$

Whip out your calculator—that's $200,000 profit. Are you happy yet? I didn't think so.

If you are new to the arena of direct marketing (welcome!), you'll notice we're playing a kind of wild numbers game, amassing great waste, like one million monkeys chucking darts across a chasm in the general direction of a dartboard. As turn-of-the-century marketing pioneer John Wanamaker famously put it, "Half the money I spend on advertising is wasted; the trouble is I don't know which half." The bad news is that it's actually more than half; the good news is that PA can learn to do better.

A FAULTY ORACLE EVERYONE LOVES

The first step toward predicting the future is admitting you can't.

—Stephen Dubner, Freakonomics Radio, March 30, 2011

The "prediction paradox": The more humility we have about our ability to make predictions, the more successful we can be in planning for the future.

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

Your resident "oracle," PA, tells you which customers are most likely to respond. It earmarks a quarter of the entire list and says, "These folks are three times more likely to respond than average!" So now you have a short list of 250,000 customers of whom 3 percent will respond—7,500 responses.

Oracle, shmoracle! These predictions are seriously inaccurate—we still don't have strong confidence when contacting any one customer, given this measly 3 percent response rate. However, the overall IQ of your dart-throwing monkeys has taken a real boost. If you send mail to only this short list then you profit:

$$\begin{aligned}\text{Overall profit} &= \text{Revenue} - \text{Cost} \\ &= (\$220 \times 7,500 \text{ responses}) - (\$2 \times 250,000)\end{aligned}$$

That's $1,150,000 profit. You just improved your profit 5.75 times over by mailing to *fewer* people (and, in so doing, expending fewer trees). In particular, you predicted who wasn't worth contacting and simply left them alone. Thus you cut your costs by three-quarters in exchange for losing only one-quarter of sales. That's a deal I'd take any day.

It's not hard to put a value on prediction. As you can see, even if predictions themselves are generated from sophisticated mathematics, it takes only simple arithmetic to roll up the plethora of predictions—some accurate, and others not so much—and reveal the aggregate bottom-line effect. This isn't just some abstract notion; The Prediction Effect means business.

PREDICTIVE PROTECTION

Thus, value has emerged from just a little predictive insight, a small prognostic nudge in the right direction. It's easy to draw an analogy to science fiction, where just a bit of supernatural foresight can go a long way. Nicolas Cage kicks some serious bad-guy butt in the movie *Next,* based on a story by Philip K. Dick. His weapon? Pure prognostication. He can see the future, but only two minutes ahead. It's enough prescience to do some damage. An unarmed civilian with a soft heart and the best of intentions, he winds up marching through something of a war zone, surrounded by a posse of heavily armed FBI agents who obey his every gesture. He sees the damage of every booby trap, sniper, and mean-faced grunt before it happens and so can command just the right moves for this Superhuman Risk-Aversion Team, avoiding one calamity after another.

In a way, deploying PA makes a Superhuman Risk-Aversion Team of the organization just the same. Every decision an organization makes, each step it takes, incurs risk. Imagine the protective benefit of foreseeing each pitfall so that it may be avoided—each criminal act, stock value decline, hospitalization, bad debt, traffic jam, high school dropout . . . and each ignored marketing brochure that was a waste to mail. *Organizational risk management*, traditionally the act of defending against singular, macrolevel incidents like the crash of an aircraft or an economy, now broadens to fight a myriad of microlevel risks.

Hey, it's not all bad news. We win by foreseeing good behavior as well, since it often signals an opportunity to gain. The name of the game is "Predict 'n' Pounce" when it pops up on the radar that a customer is likely to buy, a stock value is likely to increase, a voter is likely to swing, or the apple of one's online dating eye is likely to reciprocate.

A little glimpse into the future gives you power because it gives you options. In some cases the obvious decision is to act in order to avert what may not be inevitable, be it crime, loss, or sickness. On the positive side, in the case of foreseeing demand, you act to exploit it. Either way, prediction serves to drive decisions.

Let's turn to a real case, a $1 million example.

A SILENT REVOLUTION WORTH A MILLION

When an organization goes live with PA, it unleashes a massive army, but it's an army of ants. These ants march out to the front lines of an organization's operations, the places where there's contact with the likes of customers, students, or patients—the people served by the organization. Within these interactions, the ant army, guided by predictions, improves millions of small decisions. The process goes largely unnoticed, under the radar . . . until someone bothers to look at how it's adding up. The improved decisions may each be ant-sized, relatively speaking, but there are so many that they come to a powerful net effect.

In 2005, I was digging in the trenches, neck deep in data for a client who wanted more clicks on their website. To be precise, they wanted more clicks on their sponsors' ads. This was about the money—more clicks, more money. The site had gained tens of millions of users over the years, and within just several months' worth of tracking data that they handed me, there were 50 million rows of learning data—no small treasure trove from which to learn to predict . . . *clicks*.

Advertising is an inevitable part of media, be it print, television, or your online experience. Benjamin Franklin forgot to include it when he proclaimed, "In this world nothing can be said to be certain, except death and taxes." The flagship Internet behemoth Google credits ads as its greatest source of revenue. It's the same with Facebook.

But on this website, ads told a slightly different story than usual, which further amplified the potential win of predicting user clicks. The client was a leading student grant and scholarship search service, with one in three college-bound high school seniors using it: an arcane niche, but just the one over which certain universities and military recruiters were drooling. One ad for a university included a strong pitch, naming itself "America's leader in creative education" and culminating with a button that begged to be clicked: "Yes, please have someone from the Art Institute's Admissions Office contact me!" And you won't be surprised to hear that creditors were also placing ads, at the ready to provide these students another source of funds: loans. The sponsors would pay up to $25 per lead—for each

would-be recruit. That's good compensation for one little click of the mouse. What's more, since the ads were largely relevant to the users, closely related to their purpose on the website, the response rates climbed up to an unusually high 5 percent. So this little business, owned by a well-known online job-hunting firm, was earning well. Any small improvement meant real revenue.

But improving ad selection is a serious challenge. At certain intervals, users were exposed to a full-page ad, selected from a pool of 291 options. The trick is selecting the best one for each user. The website currently selected which ad to show based simply on the revenue it generated on average, with no regard to the particular user. The universally strongest ad was always shown first. Although this tactic forsakes the possibility of matching ads to individual users, it's a formidable champion to unseat. Some sponsor ads, such as certain universities, paid such a high bounty per click, and were clicked so often, that showing any user a less powerful ad seemed like a crazy thing to consider, since doing so would risk losing currently established value.

THE PERILS OF PERSONALIZATION

By trusting predictions in order to customize for the individual, you take on risk. A predictive system boldly proclaims, "Even though ad A is so strong overall, for this particular user it is worth the risk of going with ad B." For this reason, most online ads are not personalized for the individual user—even Google's AdWords, which allows you to place textual ads alongside search results as well as on other Web pages, determines which ad to display by Web page context, the ad's click rate, and the advertiser's bid (what it is willing to pay for a click). It is not determined by anything known or predicted about the particular viewer who is going to actually see the ad.

But weathering this risk carries us to a new frontier of customization. For business, it promises to "personalize!," "increase relevance!," and "engage one-to-one marketing!" The benefits reach beyond personalizing marketing treatment to customizing the individual treatment of patients and suspected criminals as well. During a speech about satisfying our widely varying preferences in choice of spaghetti sauce—chunky? sweet? spicy?—Malcolm

Gladwell said, "People . . . were looking for . . . universals, they were looking for one way to treat all of us[;] . . . all of science through the nineteenth century and much of the twentieth was obsessed with universals. Psychologists, medical scientists, economists were all interested in finding out the rules that govern the way all of us behave. But that changed, right? What is the great revolution of science in the last 10, 15 years? It is the movement from the search for universals to the understanding of variability. Now in medical science we don't want to know . . . just how cancer works; we want to know how your cancer is different from my cancer."

From medical issues to consumer preferences, individualization trumps universals. And so it goes with ads:

PA APPLICATION: PREDICTIVE ADVERTISEMENT TARGETING

1. **What's predicted:** Which ad each customer is most likely to click.
2. **What's done about it:** Display the best ad (based on the likelihood of a click as well as the bounty paid by its sponsor).

I set up PA to perform ad targeting for my client, and the company launched it in a head-to-head, champion/challenger competition to the death against their existing system. The loser would surely be relegated to the bin of second-class ideas that just don't make as much cash. To prepare for this battle, we armed PA with powerful weaponry. The predictions were generated from machine learning across 50 million learning cases, each depicting a microlesson from history of the form, "User Mary was shown ad A and she did click it" (a positive case) or "User John was shown ad B and he did not click it" (a negative case).

The learning technology employed to pick the best ad for each user was a Naïve Bayes model. Rev. Thomas Bayes was an eighteenth-century mathematician, and the "Naïve" part means that we take a very smart man's ideas and compromise them in a way that simplifies yet makes their application feasible, resulting in a practical method that's often considered good enough at prediction and scales to the task at hand. I went with this method for its relative simplicity, since in fact I needed to generate 291 such models, one for each ad. Together, these models predict which ad a user is most likely to click on.

DEPLOYMENT'S DETOURS AND DELAYS

As with a rocket ship, launching PA looks great on paper. You design and construct the technology, place it on the launchpad, and wait for the green light. But just when you're about to hit "go," the launch is scrubbed. Then delayed. Then scrubbed again. The Wright brothers and others, galvanized by the awesome promise of a newly discovered wing design that generates lift, endured an uncharted, rocky road, faltering, floundering, and risking life and limb until all the kinks were out.

For ad targeting and other real-time PA deployments, predictions have got to zoom in at warp speed in order to provide value. Our online world tolerates no delay when it's time to choose which ad to display, determine whether to buy a stock, decide whether to authorize a credit card charge, recommend a movie, filter an e-mail for viruses, or answer a question on *Jeopardy!* A real-time PA solution must be directly integrated into operational systems, such as websites or credit card processing facilities. If you are newly integrating PA within an organization, this can be a significant project for the software engineers, who often have their hands full with maintenance tasks just to keep the business operating normally. Thus, the *deployment* phase of a PA project takes much more than simply receiving a nod from senior management to go live: It demands major construction. By the time the programmers deployed my predictive ad selection system, the data over which I had tuned it was already about 11 months old. Were the facets of what had been learned still relevant almost one year later, or would prediction's power peter out?

IN FLIGHT

This is Major Tom to Ground Control
I'm stepping through the door
And I'm floating in a most peculiar way . . .

—"Space Oddity" by David Bowie

Once launched, PA enters an eerie, silent waiting period, like you're floating in orbit and nothing is moving. But the fact is, in a low orbit around Earth you're actually screaming along at over 14,000 miles per hour. Unlike the drama of launching a rocket or erecting a skyscraper, the launch of PA is a

relatively stealthy maneuver. It goes live, but daily activities exhibit no immediately apparent change. After the ad-targeting project's launch, if you checked out the website, it would show you an ad as usual, and you could wonder whether the system made any difference in this one choice. This is what computers do best. They hold the power to silently enact massive procedural changes that often go uncredited, since most aren't directly witnessed by any one person.

But, under the surface, a sea change is in play, as if the entire ocean has been reconfigured. You actually notice the impact only when you examine an aggregated report.

In my client's deployment, predictive ad selection triumphed. The client conducted a head-to-head comparison, selecting ads for half the users with the existing champion system and the other half with the new predictive system, and reported that the new system generated at least 3.6 percent more revenue, which amounts to $1 million every 19 months, given the rate at which revenue was already coming in. This was for the website's full-page ads only; many more (smaller) ads are embedded within functional Web pages, which could potentially also be boosted with a similar PA project.

No new customers, no new sponsors, no changes to business contracts, no materials or computer hardware needed, no new full-time employees or ongoing effort—solely an improvement to decision making was needed to generate cold, hard cash. In a well-oiled, established system like the one my client had, even a small improvement of 3.6 percent amounts to something substantial. The gains of an incremental tweak can be even more dramatic: In the insurance business, one company reports that PA saves almost $50 million annually by decreasing its loss ratio by *half a percentage point*.

So how did these models predict each click?

ELEMENTARY, MY DEAR: THE POWER OF OBSERVATION

Just like Sherlock Holmes drawing conclusions by sizing up a suspect, prediction comes of astute observation: What's known about each individual provides a set of clues about what he or she may do next. The chance a user will click on a certain ad depends on all sorts of elements, including the individual's current school year, gender, and e-mail domain

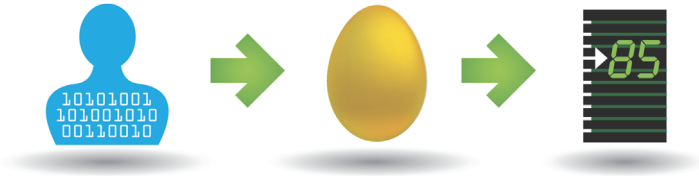
(Hotmail, Yahoo, Gmail, etc.); the ratio of the individual's SAT written-to-math scores (is the user more a verbal person or more a math person?), and on and on.

In fact, this website collected a wealth of information about its users. To find out which grants and scholarships they're eligible for, users answer dozens of questions about their school performance, academic interests, extracurricular activities, prospective college majors, parents' degrees, and more. So the table of learning data was long (at 50 million examples) and was also wide, with each row holding all the information known about the user at the moment the person viewed an ad.

It can sound like a tall order: *harnessing millions of examples in order to learn how to incorporate the various factoids known about each individual so that prediction is possible*. But we can break this down into a couple of parts, and suddenly it gets much simpler. Let's start with the contraption that makes the predictions, the electronic Sherlock Holmes that knows how to consider all these factors and roll them up into a single prediction for the individual.

> ***Predictive model***—*a mechanism that predicts a behavior of an individual, such as click, buy, lie, or die. It takes characteristics of the individual as input and provides a* predictive score *as output. The higher the score, the more likely it is that the individual will exhibit the predicted behavior.*

A predictive model (depicted throughout this book as a "golden" egg, albeit in black and white) scores an individual:



A predictive model is the means by which the attributes of an individual are factored together for prediction. There are many ways to do this. One is to weigh each characteristic and then add them up—perhaps females boost their score by 33.4, Hotmail users decrease their score by 15.7, and so on.

Each element counts toward or against the final score for that individual. This is called a *linear model*, generally considered quite simple and limited, although usually much better than nothing.

Other models are composed of *rules*, like this real example:

> IF the individual
> is still in high school
> AND
> expects to graduate college within three years
> AND
> indicates certain military interest
> AND
> has not been shown this ad yet
> THEN the probability of clicking on the ad for the Art Institute is 13.5 percent.

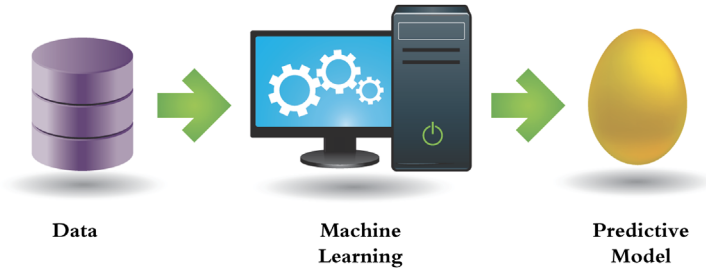
This rule is a valuable find, since the overall probability of responding to the Art Institute's ad is only 2.7 percent, so we've identified a pocket of avid clickers, relatively speaking.

It is interesting that those who have indicated a military interest are more likely to show interest in the Art Institute. We can speculate, but it's important not to assume there is a *causal* relationship. For example, it may be that people who complete more of their profile are just more likely to click in general, across all kinds of ads.

Various types of models compete to make the most accurate predictions. Models that combine a bunch of rules like the one just shown are—relatively speaking—on the simpler side. Alternatively, we can go more "supermath" on the prediction problem, employing complex formulas that predict more effectively but are almost impossible to understand by human eyes.

But all predictive models share the same objective: They consider the various factors of an individual in order to derive a single predictive score for that individual. This score is then used to drive an organizational decision, guiding which action to take.

Before using a model, we've got to build it. Machine learning builds the predictive model:



Machine learning crunches data to build the model, a brand-new prediction machine. The model is the product of this learning technology—it is itself the very thing that has been learned. For this reason, machine learning is also called *predictive modeling*, which is a more common term in the commercial world. If deferring to the older metaphorical term *data mining*, the predictive model is the unearthed gem.

Predictive modeling generates the entire model from scratch. All the model's math, weights, or rules are created automatically by the computer. The machine learning process is designed to accomplish this task, to mechanically develop new capabilities from data. This automation is the means by which PA builds its predictive power.

The hunter returns back to the tribe, proudly displaying his kill. So, too, a data scientist posts her model on the bulletin board near the company ping-pong table. The hunter hands over the kill to the cook, and the data scientist cooks up her model, translates it to a standard computer language, and e-mails it to an engineer for integration. A well-fed tribe shows the love; a psyched executive issues a bonus.

TO ACT IS TO DECIDE

Knowing is not enough; we must act.

—Johann Wolfgang von Goethe

Once you develop a model, don't pat yourself on the back just yet. Predictions don't help unless you do something about them. They're just thoughts, just

ideas. They may be astute, brilliant gems that glimmer like the most polished of crystal balls, but displaying them on a shelf gains you nothing—they just sit there and look smart.

Unlike a report sitting dormant on the desk, PA leaps out of the lab and takes action. In this way, it stands above other forms of analysis, data science, and data mining. It desires deployment and loves to be launched—because, in what it foretells, it mandates movement.

The predictive score for each individual directly informs the decision of what action to take with that individual. Doctors take a second look at patients predicted to be readmitted, and service agents contact customers predicted to cancel. Predictive scores issue imperatives to *mail, call, offer a discount, recommend a product, show an ad, expend sales resources, audit, investigate, inspect for flaws, approve a loan,* or *buy a stock*. By acting on the predictions produced by machine learning, the organization is now applying what's been learned, modifying its everyday operations for the better.

To make this point, we have mangled the English language. Proponents like to say that PA is *actionable*. Its output directly informs actions, commanding the organization about what to do next. But with this use of vocabulary, industry insiders have stolen the word *actionable*, which originally meant *worthy of legal action* (i.e., "sue-able"), and morphed it. They did so because they're tired of seeing sharp-looking reports that provide only a vague, unsure sense of direction.

With this word's new meaning established, "your fly is unzipped" is *actionable* (it is clear what to do—you can and should take action to remedy), but "you're going bald" is not (there's no cure; nothing to be done). Better yet, "I predict you will buy these button-fly jeans and this snazzy hat" is actionable to a salesperson.

Launching PA into action delivers a critical new edge in the competitive world of business. One sees massive commoditization taking place today as the faces of corporations appear to blend together. They all seem to sell pretty much the same thing and act in pretty much the same ways. To stand above the crowd, where can a company turn?

As Thomas Davenport and Jeanne Harris put it in *Competing on Analytics: The New Science of Winning*, "At a time when companies in many industries offer similar products and use comparable technology, high-performance business

processes are among the last remaining points of differentiation." Enter PA. Survey results have in fact shown that "a tougher competitive environment" is by far the strongest reason why organizations adopt this technology.

But while the launch of PA brings real change, it can also wreak havoc by introducing new risk. With this in mind, we now return to John's story.

A PERILOUS LAUNCH

Dr. John Elder bet it all on a predictive model. He concocted it in the lab, packed it into a black box, and unleashed it on the stock market. Some people make their own bed in which they must then passively lie. But John had climbed way up high to take a leap of faith. Diving off a mountaintop with newly constructed, experimental wings, he wondered how long it might take before he could be sure he was flying rather than crashing.

The risks stared John in the face. His and his wife's full retirement savings were in the hands of an experimental device, launched into oblivion and destined for one of the same two outcomes achieved by every rocket: glory or mission failure. Discovering profitable market patterns that sustain is the mission of thousands of traders operating in what John points out is a brutally competitive environment; doing so automatically with machine learning is the most challenging of ambitions, considered impossible by many. It doesn't help that a stock market scientist is completely on his own, since work in this area is shrouded in secrecy, leaving virtually no potential to learn from the successes and failures of others. Academics publish, marketers discuss, but quants hide away in their Batcaves. What can look great on paper might be stricken with a weakness that destroys or an error that bankrupts. John puts it plainly: "Wall Street is the hardest data mining problem."

The evidence of danger was palpable, as John had recently uncovered a crippling flaw in an existing predictive trading system and personally escorted it to its grave. Opportunity had come knocking on the door of a small firm called Delta Financial in the form of a black-box trading system purported to predict movements of the Standard & Poor's (S&P) 500 with 70 percent accuracy. Built by a proud scientist, the system promised to make millions, so stakeholders were flying around all dressed up in suits, actively lining up

investors prepared to place a huge bet. Among potential early investors, Delta was leading the way for others, taking a central, influential role. The firm was known for investigating and championing cutting-edge approaches, weathering the risk inherent to innovation. As a necessary precaution, Delta sought to empirically validate this system. The firm turned to John, who was consulting for them on the side while pursuing his doctorate at the University of Virginia in Charlottesville. John's work for Delta often involved inspecting, and sometimes debunking, black-box trading systems.

How do you prove a machine is broken if you're not allowed to look inside it? Healthy skepticism bolstered John's resolve, since the claimed 70 percent accuracy raised red flags as quite possibly too darn good to be true. But he was not granted access to the predictive model. With secrecy reigning supreme, the protocol for this type of audit dictated that John receive only the numerical results, along with a few adjectives that described its design: *new, unique, powerful!* With meager evidence, John sought to prove a crime he couldn't even be sure had been committed.

Before each launch, organizations establish confidence in PA by "predicting the past" (aka backtesting). The predictive model must prove itself on historical data before its deployment. Conducting a kind of simulated prediction, the model evaluates across data from last week, last month, or last year. Feeding on input that could only have been known at a given time, the model spits out its prediction, which then matches against what we now already know took place thereafter. Would the S&P 500 go down or up on March 21, 1991? If the model gets this retrospective question right, based only on data available by March 20, 1991 (the day just before), we have evidence the model works. These retrospective predictions—without the manner in which they had been derived—were all John had to work with.

HOUSTON, WE HAVE A PROBLEM

Even the most elite of engineers commit the most mundane and costly of errors. In late 1998, NASA launched the Mars Climate Orbiter on a daunting nine-month trip to Mars, a mission that fewer than half the world's launched probes headed for that destination have completed successfully. This $327.6 million

calamity crashed and burned, due not to the flip of fate's coin, but rather a simple snafu. The spacecraft came too close to Mars and disintegrated in its atmosphere. The source of the navigational bungle? One system expected to receive information in metric units (newton-seconds), but a computer programmer for another system had it speak in English imperial units (pound-seconds). Oops.

John stared at a screen of numbers, wondering if anything was wrong and, if so, whether he could find it. From the long list of impressive—yet retrospective—predictions, he plainly saw the promise of huge profits that had everyone involved so excited. If he proved there was a flaw, vindication; if not, lingering uncertainty. The task at hand was to reverse engineer: Given the predictions the system generated, could he infer how it worked under the hood, essentially eking out the method in its madness? This was ironic, since all predictive modeling is a kind of reverse engineering to begin with. Machine learning starts with the data, an encoding of things that have happened, and attempts to uncover patterns that generated or explained the data in the first place. John was attempting to deduce what the other team had deduced. His guide? Informal hunches and ill-informed inferences, each of which could be pursued only by way of trial and error, testing each hypothetical mess-up he could dream up by programming it by hand and comparing it to the retrospective predictions he had been given.

His perseverance finally paid off: John uncovered a true flaw, thereby flinging back the curtain to expose a flustered Wizard of Oz. It turned out that the prediction engine committed the most sacrilegious of cheats by looking at the one thing it must not be permitted to see. It had looked at the future. The battery of impressive retrospective predictions weren't true predictions at all. Rather, they were based in part on a three-day average calculated across yesterday, today . . . and tomorrow. The scientists had probably intended to incorporate a three-day average leading up to today, but had inadvertently shifted the window by a day. Oops. This crippling bug delivered the dead-certain prognosis that this predictive model would not perform well if deployed into the field. Any prediction it would generate today could not incorporate the very thing it was designed to foresee—tomorrow's stock price—since, well, it isn't known yet. So, if foolishly deployed, its accuracy could never match the exaggerated performance falsely demonstrated across

the historical data. John revealed this bug by reverse engineering it. On a hunch, he handcrafted a method with the same type of bug and showed that its predictions closely matched those of the trading system.

A predictive model will sink faster than the *Titanic* if you don't seal all its "time leaks" before launch. But this kind of "leak from the future" is common, if mundane. Although core to the very integrity of prediction, it's an easy mistake to make, given that each model is backtested over historical data for which prediction is not, strictly speaking, possible. The relative future is always readily available in the testing data, easy to inadvertently incorporate into the very model trying to predict it. Such temporal leaks achieve status as a commonly known gotcha among PA practitioners. If this were an episode of *Star Trek*, our beloved, hypomanic engineer Scotty would be screaming, "Captain, we're losing our temporal integrity!"

It was with no pleasure that John delivered the disappointing news to his client, Delta Financial: He had debunked the system, essentially exposing it as inadvertent fraud. High hopes were dashed as another fairy tale bit the dust, but gratitude quickly ensued as would-be investors realized they'd just dodged a bullet. The wannabe inventor of the system suffered dismay but was better off knowing now; it would have hit the fan much harder postlaunch, possibly including prosecution for fraud, even if inadvertently committed. The project was aborted.

THE LITTLE MODEL THAT COULD

Even the young practitioner that he was, John was a go-to data man for entrepreneurs in black-box trading. One such investor moved to Charlottesville, but only after John Elder, PhD, new doctorate degree in hand, had just relocated to Houston in order to continue his academic rite of passage with a postdoc research position at Rice University. He'd left quite an impression back in Charlottesville, though; people in both the academic and commercial sectors alike referred the investor to John. Despite John's distance, the investor hired him to prepare, launch, and monitor a new black-box mission remotely from Houston. It seemed as good a place as any for the project's Mission Control.

And so it was time for John to move beyond the low-risk role of evaluating other people's predictive systems and dare to build one of his own. Over several months, he and a small team of colleagues built upon core insights from the investor and produced a new, promising black-box trading model. John was champing at the bit to launch it and put it to the test. All the stars were aligned for liftoff except one: The money people didn't trust it yet.

There was good reason to believe in John. Having recently completed his doctorate degree, he was armed with a fresh, talented mind, yet had already gained an impressively wide range of data-crunching problem-solving experience. On the academic side, his PhD thesis had broken records among researchers as the most efficient way to optimize for a certain broad class of system engineering problems (machine learning is itself a kind of optimization problem). He had also taken on predicting the species of a bat from its echolocation signals (the chirps bats make for their radar). And in the commercial world, John's pregrad positions had dropped him right into the thick of machine learning systems that steer for aerospace flight and that detect cooling pipe cracks in nuclear reactors, not to mention projects for Delta Financial looking over the shoulders of other black-box quants.

And now John's latest creation absolutely itched to be deployed. Backtesting against historical data, all indications whispered confident promises for what this thing could do once set in motion. As John puts it, "A slight pattern emerged from the overwhelming noise; we had stumbled across a persistent pricing inefficiency in a corner of the market, a small edge over the average investor, which appeared repeatable." Inefficiencies are what traders live for. A perfectly efficient market can't be played, but if you can identify the right imperfection, it's payday.

PA APPLICATION: BLACK-BOX TRADING

1. **What's predicted:** Whether a stock will go up or down.
2. **What's done about it:** Buy stocks that will go up; sell those that will go down.

John could not get the green light. As he strove to convince the investor, cold feet prevailed. It appeared they were stuck in a stalemate. After all, this

guy might not get past his jitters until he could see the system succeed, yet it couldn't succeed while stuck on the launchpad. The time was now, as each day marked lost opportunity.

After a disconcerting meeting that seemed to go nowhere, John went home and had a sit-down with his wife, Elizabeth. What supportive spouse could possibly resist the seduction of her beloved's ardent excitement and strong belief in his own abilities? She gave him the go-ahead to risk it all, a move that could threaten their very home. But he still needed buy-in from one more party.

Delivering his appeal to the client investor raised questions, concerns, and eyebrows. John wanted to launch with his own personal funds, which meant no risk whatsoever to the client and would resolve any doubts by field-testing John's model. But this unorthodox step would be akin to the dubious choice to act as one's own defense attorney. When an individual is without great personal means, this kind of thing is often frowned upon. It conveys overconfident, foolish brashness. Even if the client wanted to truly believe, it would be another thing to expect the same from coinvestors who hadn't gotten to know and trust John. But with every launch, proponents gamble something fierce. John had set the rules for the game he'd chosen to play.

He received his answer from the investor: "Go for it!" This meant there was nothing to prevent moving forward. It could have also meant the investor was prepared to write off the project entirely, feeling there was nothing left to lose.

HOUSTON, WE HAVE LIFTOFF

Practitioners of PA often put their own professional lives a bit on the line to push forward, but this case was extreme. Like baseball's Billy Beane of the Oakland A's, who literally risked his entire career to deploy and field-test an analytical approach to team management, John risked everything he had. It was early 1994, and John's individual retirement account (IRA) amounted to little more than $40,000. He put it all in.

"Going live with black-box trading is really exciting and really scary," says John. "It's a roller coaster that never stops. The coaster takes on all these thrilling ups and downs, but with a very real chance it could go off the rails."

As with baseball, he points out, slumps aren't slumps at all—they're inevitable statistical certainties. Each one leaves you wondering, "Is this falling feeling part of a safe ride, or is something broken?" A key component to his system was a cleverly designed means to detect real quality, a measure of system integrity that revealed whether recent success had been truly deserved or had come about just due to dumb luck.

From the get-go, the predictive engine rocked. It increased John's assets at a rate of 40 percent per year, which meant that after two years his money had doubled.

The client investor was quickly impressed and soon put in a couple of million dollars himself. A year later, the predictive model was managing a $20 million fund across a group of investors, and eventually the investment pool increased to a few hundred million dollars. With this much on tap, every win of the system was multiplicatively magnified.

No question about it: All involved relished this fiesta, and the party raged on and on, continuing almost nine years, consistently outperforming the overall market all along. The system chugged, autonomously trading among a dozen market sectors such as technology, transportation, and healthcare. John says the system "beat the market each year and exhibited only two-thirds its standard deviation—a home run as measured by risk-adjusted return."

But all good things must come to an end, and just as John had talked his client up, he later had to talk him down. After nearly a decade, the key measure of system integrity began to decline. John was adamant that they were running on fumes, so with little ceremony the entire fund was wound down. The system was halted in time, before catastrophe could strike. In the end, all the investors came out ahead.

A PASSIONATE SCIENTIST

The early success of this streak had quickly altered John's life. Once the project was cruising, he had begun supporting his rapidly growing family

with ease. The project was taking only a couple of John's hours each day to monitor, tweak, and refresh what was a fundamentally stable, unchanging method within the black box. What's a man to do? Do you put your feet up and sip wine indefinitely, with the possible interruption of family trips to Disney World? After all, John had thus far always burned the candle at both ends out of financial necessity, with summer jobs during college, part-time work during graduate school, and this black-box project, which itself had begun as a moonlighting gig during his postdoc. Or do you follow the logical business imperative: Pounce on your successes, using all your free bandwidth to find ways to do more of the same?

John's passion for the craft transcended these self-serving responses to his good fortune. That is to say, he contains the spirit of the geek. He jokes about the endless insatiability of his own appetite for the stimulation of fresh scientific challenges. He's addicted to tackling something new. There is but one antidote: a growing list of diverse projects. So, two years into the stock market project, he wrapped up his postdoc, packed up his family, and moved back to Charlottesville to start his own data mining company.

And so John launched Elder Research, now the largest predictive analytics services firm (pure play) in North America. A narrow focus is key to the success of many businesses, but Elder Research's advantage is quite the opposite: its diversity. The company's portfolio reaches far beyond finance to include all major commercial sectors and many branches of government. John has also earned a top-echelon position in the industry. He coauthors massive textbooks, frequently chairs or keynotes at Predictive Analytics World conferences, takes cameos as a university professor, and served five years as a presidential appointee on a national security technology panel.

LAUNCHING PREDICTION INTO INNER SPACE

With stories like John's coming to light, organizations are jumping on the PA bandwagon. One such firm, a mammoth international organization, focuses the power of prediction introspectively, casting PA's keen gaze on its own employees. Read on to witness the windfall and the fallout when scientists dare to ask: Do people like being predicted?

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

CHAPTER 2

With Power Comes Responsibility

Hewlett-Packard, Target, the Cops, and the NSA Deduce Your Secrets

How do we safely harness a predictive machine that can foresee job resignation, pregnancy, and crime? Are civil liberties at risk? Why does one leading health insurance company predict policyholder death? Two extended sidebars explore: (1) Does the government undertake fraud detection more for its citizens or for self-preservation, and (2) for what compelling purpose does the National Security Agency (NSA) need your data even if you have no connection to crime whatsoever, and can the agency use machine learning supercomputers to fight terrorism without endangering human rights?

> *Predictive analytics . . . is right at the fulcrum point of utopian and dystopian visions of the future.*
>
> –Andrew Frank, Research Vice President, Gartner

What would happen if your boss were notified that you're allegedly going to quit—even though you had said this to no one? If you are one of the more than 300,000 who work at Hewlett-Packard (HP), your employer has tagged you—and all your colleagues—with a "Flight Risk" score. This simple number foretells whether you're likely to leave your job. As an HP employee, there's a good chance you didn't already know that. Postpone freaking out until you finish reading the full explanation in this chapter.

This story about HP arrived in the wake of media outcry against Target in 2012 after learning the big-box retailer had taken to predicting customer pregnancy. The media firestorm invoked misleading accusations, fear of corporate power, postulations by television personalities, and, of course, predictive analytics (PA). To my surprise, I ended up in the thick of it.

TV news programs strike like a blunt instrument, but often in the right general direction. The media assault was reactionary and chose to misinform, yet legitimate quandaries lurk below the surface. Target's and HP's predictive power brings to focus an exceptionally challenging and pressing ethical question. Within the minefield that is the privacy debate, the stakes just rose even higher.

Why? Because prediction snoops into your private future. These cases involve the corporate deduction of previously unknown, sensitive facts: Are you considering quitting your job? Are you pregnant? This isn't a case of mishandling, leaking, or stealing data. Rather, it is *the generation of new data*, the indirect discovery of unvolunteered truths about people. Organizations predict these powerful insights from existing innocuous data, as if creating them out of thin air. Are they equipped to manage their own creation?

While we come to terms with the sheer magnitude of prediction's power, we've only begun to fathom the privacy concerns it introduces. A chain reaction triggers and surprises even the experts: Organizations exert newfound capabilities, consumers rise up, the media stir the pot, and scientists dodge bullets and then reexamine scruples.

The journey eventually takes us to a particularly uncomfortable dilemma. Beyond expectant moms and departing employees, PA also flags potential criminals and actively helps law enforcement decide who stays in prison and who goes free.

This tale follows my journey from carefree technologist to unwitting talking head and the journey of organizations from headstrong to humbled. The asocial domain of data and analytics is not so irrelevant after all.

THE PREDICTION OF TARGET AND THE TARGET OF PREDICTION

In 2010, I invited an expert at Target, Andrew Pole, to keynote at Predictive Analytics World, the conference series I founded. Pole manages dozens of analytics professionals who run various PA projects at Target. In October of that year, Pole delivered a stellar speech on a wide range of PA deployments at Target. He took the stage and dynamically engaged the audience,

revealing detailed examples, interesting stories, and meaningful business results that left the audience clearly enthused. Free to view, here it is: www.pawcon.com/Target.

Toward the end, Pole described a project to predict customer pregnancy. Given that there's a tremendous sales opportunity when a family prepares for a newborn, you can see the marketing potential.

But this was something pointedly new, and I turned my head to scan the audience for any reactions. Nothing. Nada. Zilch. Normally, for marketing projects, PA predicts buying behavior. Here, the thing being predicted was not something marketers care about directly, but rather, something that could itself be a strong predictor of a wide range of shopping needs. After all, the marketer's job is to discover demand and pounce on it. You can think of this predictive goal as a "surrogate" (sorry) for the pertinent shopping activities a retail marketer is paid to care about.

PA APPLICATION: PREGNANCY PREDICTION

1. **What's predicted:** Which female customers will have a baby in coming months.
2. **What's done about it:** Market relevant offers for soon-to-be parents of newborns.

From what data did Target learn to predict pregnancy, given that predictive modeling requires a number of known cases from which to learn? Remember, the predictive modeling process is a form of automated data crunching that learns from *training examples*, which must include both positive and negative examples. An organization needs to have positively identified in the past some cases of what it would like to predict in the future. To predict something like "will buy a stereo," you can bet a retailer has plenty of positive cases. But how can you locate Target customers known to be pregnant?

You may be surprised how simple it is to answer this puzzle. Can you guess? Let's assume no medical information or pharmaceutical data is employed for this project. Why does a customer inform Target she is pregnant? The answer: the Target baby registry. Registrants not only

disclose they're pregnant, but they also reveal their due date. In addition, Target has indicated there are other marketing programs through which more moms-to-be identify themselves, thus also serving as positive learning examples.

Target pulled together training data by merging the baby registry data with other retail customer data and generated a "fairly accurate" predictive model. The store can now apply the model to customers who have *not* registered as pregnant. This identifies many more pregnant customers, since we can assume most such customers in fact do not register.

The model predictively evaluates a customer based on what she has purchased, which can include baby-related products, but may include combinations of other products not necessarily directly related to babies. Deriving the model is an automated act of trend spotting that explores a broad range of factors. I doubt Target's system confirmed that buying pickles and ice cream turns out to be a good indicator of pregnancy, but any and all product categories were analyzed and considered. The model identified 30 percent more customers for Target to contact with pregnancy-oriented marketing material—a significant marketing success story.

A PREGNANT PAUSE

Strutting charismatically across the conference stage, Pole boldly lauded this unorthodox endeavor, which he led at Target. The business value was clear, the story entertaining. It's likely he was delivering what had gone over well for internal Target presentations, but now to an open forum. It made for great material and engaged the audience.

I wondered for a moment if there had been any concerns but assumed, as one engrossed in the core technology itself may tend to do, that this project had been vetted, that concerns had been allayed and put to rest by folks at Target. Emerging from inside the PA practitioner's dark data cave, squinting at the world outside, it can be hard to imagine how unsuspecting folks walking down the street might respond to such a project. In fact, Pole reassured the audience that Target carefully adheres to all privacy and data-use laws. "Target wants to make sure that we don't end up in the newspaper or on TV because

we went out and we used something that we're not supposed to be using." Little did we know where this was headed.

MY 15 MINUTES

Because the ensuing media storm around Target's pregnancy prediction pulled me into its wake, I witnessed from a front-row seat how, if one reporter sets off just the right spark, the pundits will obediently burn and the news cycle will fan the flames.

Who spilled the beans in the first place? A few months after Pole's presentation, *New York Times* reporter Charles Duhigg interviewed me. Exploring, he asked for interesting new ways PA was being used. I rattled off a few and included pregnancy prediction, pointing him to the online video of Pole's talk, which had thus far received no media attention, and connecting him to Pole via e-mail. I must admit that by now the privacy question had left my mind almost entirely.

One year later, in February 2012, Duhigg published a front-page *New York Times Magazine* article, sparking a viral outbreak that turned the Target pregnancy prediction story into a debacle. The article, "How Companies Learn Your Secrets," conveys a tone that implies wrongdoing is a foregone conclusion. It punctuates this by alleging an anonymous story of a man discovering his teenage daughter is pregnant only by seeing Target's marketing offers to her, with the unsubstantiated but tacit implication that this resulted specifically from Target's PA project. The *Times* even produced a short video to go with the article, which features dramatic, slow-motion, color-muted images of Target shoppers checking out, while creepy, suspenseful music plays and Duhigg himself states, "If they know when [your life is changing], then they can . . . manipulate you . . . so that your habits put dollars in their pockets." He refers to the practice of data-driven marketing as "spying on" customers.

This well-engineered splash triggered rote repetition by press, radio, and television, all of whom blindly took as gospel what had only been implied—that the teen's story stemmed from Target's pregnancy prediction—and ran with it. Not incidentally, the article was excerpted from and helped launch

Duhigg's book, *The Power of Habit: Why We Do What We Do in Life and Business* (Random House, 2012), which hit the *New York Times* bestseller list.

The tornado sucked me in because the article quoted me in addition to Pole who, along with Target as a whole, had now unsurprisingly clammed up. As an independent consultant, I enjoyed unfettered freedom to make public appearances. I had no prudent employer that might hold me back.

THRUST INTO THE LIMELIGHT

This techie transmogrified into a pundit, literally overnight, as I raced to New York City on a red-eye to appear on Fox News. But placing my talking head on millions of TVs does not magically prepare me for such a role. Thriving in an abstract pool of data, the PA professional occasionally surfaces for air, usually only by accident. For the most part, this work is an exercise in math and algorithms to discover patterns that promise to hold true tomorrow—a strange, magical game to almost defy whatever laws of physics prohibit time travel. Inside this petri dish, you're insulated, knowing nothing of the visceral angst of broken hearts or broken privacy. In asking me to shed my lab coat for a suit and tie, the powers that be declared that our previously esoteric activities buried beneath these murky depths of data are truly important after all.

The morning news program *Fox & Friends* positioned me behind a desk, and I struggled to sit still in what was clearly the hot seat. Celebrity host Gretchen Carlson looked over and raised her voice to greet me from across the studio just before we started: "Hi, Eric!" I greeted her back as if it were just another day in the studio: "Hi, Gretchen!"

Then we were live to an estimated two million viewers. Falling in line behind the *Times*, Carlson read Target the riot act for exposing a girl's pregnancy, neglecting to mention the story was only an unsubstantiated allegation and implying this kind of collateral damage is innate to PA's application. A third talking head, a professor of medical ethics, reinforced the theme that all applications of PA ought best be shut down, at least pending further investigation. The millions of TVs tuned to Fox at that moment

displayed a Target store, overlaid with the question, "Are stores spying on you?" Later the screen proclaimed, "Target has got you in its aim."

It quickly became clear I was to serve as a foil as the news show demonized my profession. For the moment, I was the face of PA, and I had to fight back. If there is a certain carelessness in how organizations wield the increasing power to predict, so too is there carelessness in misleading media coverage. I took a deep breath and asserted that the *New York Times* article was misleading because it implied Target has a "supernatural" ability to accurately predict who is pregnant, and because it established an unsubstantiated connection to the pregnant teen's alleged story. Target's predictions are not medical diagnosis and are not based on medical information. Finally, I managed to squeeze into my allotted seconds the main point: It is really important that PA not be universally stigmatized. You can watch the televised clip at www.pawcon.com/target-on-fox.

In another interview, I was confronted with a quote from privacy advocate Katherine Albrecht, who said, "The whole goal [of retailers] is to figure out everything you can learn about your customer. We're creating a retail zoo, where customers are the exhibits." My reply? Unlike the social sciences, PA's objective is to improve operational efficiency rather than figure people out for its own sake—and, either way, just because you're observing a person does not mean that person is being treated like an animal.

The media coverage was broad and, within a few weeks, it seemed like everyone I spoke with both inside and outside my work life had at least caught wind of the Target pregnancy story. Even comedian Stephen Colbert covered it, suggesting Target's next move will be to predict from your spouse's shopping habits that she is having an affair, and therefore send you a coupon for a hot plate that will go perfectly with your new studio apartment (more than just a joke, divorce prediction is included in this book's Central Table 1).

As the dust settles, we're left with a significant challenge: How can true privacy concerns be clearly defined, even as media overblows and confuses?

YOU CAN'T IMPRISON SOMETHING THAT CAN TELEPORT

Information about transactions, at some point in time, will become more important than the transactions themselves.

—Walter Wriston, former chairman and CEO of Citicorp

Information wants to be free.

—Stewart Brand to Steve Wozniak at the first Hackers Conference, 1984

Data matters. It's the very essence of what we care about.

Personal data is not equivalent to a real person—it's much better. It takes no space, costs almost nothing to maintain, lasts forever, and is far easier to replicate and transport. Data is worth more than its weight in gold—certainly so, since data weighs nothing; it has no mass.

Data about a person is not as valuable as the person, but since the data is so much cheaper to manage, it's a far better investment. Alexis Madrigal, senior editor at *The Atlantic*, points out that a user's data can be purchased for about half a cent, but the average user's value to the Internet advertising ecosystem is estimated at $1,200 per year.

Data's value—its power, its meaning—is the very thing that also makes it sensitive. The more data, the more power. The more powerful the data, the more sensitive. So the tension we're feeling is unavoidable. If nobody cared about some piece of data, nobody would try to protect it, and nobody would want to access it or even bother to retain it in the first place. John Elder reflected, "The fact that it's perceived as dangerous speaks to its power; if it were weak, it wouldn't be a threat."

Ever since the advent of paper and pen, this has been the story. A doctor scribbled a note, and the battle to establish and enforce access policies began.

But now, digital data travels so far, so fast, between people, organizations, and nations. Combine this ability of data to go anywhere at almost no cost with the intrinsic value of the stuff that's traveling, and you have the makings of a very fickle beast, a swarm of gremlins impressively tough to control. It's like trying to incarcerate the *X-Men*'s superhero Nightcrawler, who has the ability to teleport. It's not confined to our normal three dimensions of movement, so you just can't lock it up.

Data is such a unique thing to ship, we have a special word for its telekinetic mode of transport. We call it telecommunication.

Data wants to spread like wildfire. As privacy advocate David Sobel put it, "Once information exists, it's virtually impossible to limit its use. You have all this great data lying around, and sooner or later, somebody will say, 'What else can I do with it?'"

This new, powerful currency proves tough to police. A shady deal to share consumer records is completed with the press of a button—no covert physical shipment of goods required.

LAW AND ORDER: POLICIES AND POLICING OF DATA

> *[Privacy is] the most comprehensive of all rights and the one most cherished by a free people.*
>
> —Supreme Court Justice Louis Brandeis, 1928

And yet, we must try our darnedest to tame this wild creature. An open free-for-all is surely not an option. The world will continue struggling to impose order on the distribution of medical facts, financial secrets, and embarrassing photos. Consternation runs deep, with an estimated one in four Facebook users posting false data due to privacy concerns.

Each organization must decide data's who, what, where, when, how long, and why:

Retain—What is stored and for how long.

Access—Which employees, types of personnel, or group members may retrieve and look at which data elements.

Share—What data may be disseminated to which parties within the organization, and to what external organizations.

Merge—What data may be joined together, aggregated, or connected.

React—How may each data element be acted upon, determining an organization's response, outreach, or other behavior.

To make it even more complicated, add to each of these items ". . . under which circumstances and for what type of intention or purpose."

Pressing conundrums ensue. Which data policies can and should be established via legislation, and which by industry best practices and rules of etiquette? For which data practices may the organization default the consumer in, in which case she must take explicit action to opt out if so desired? How are policies enforced: What security standards—encryption, password integrity, firewalls, and the like—promise to earn Fort Knox's reputation in the electronic realm?

We have our work cut out for us.

THE BATTLE OVER DATA

> *The Internet of free platforms, free services, and free content is wholly subsidized by targeted advertising, the efficacy (and thus profitability) of which relies on collecting and mining user data.*
>
> —Alexander Furnas, writer for *The Atlantic*

The stakes increase and the opponents' resolve hardens like cooling lava.

In one corner we have privacy advocates, often loath to trust organizations, racing to squeeze shut data's ebb and flow: Contain it, delete it, or prevent it from being recorded in the first place.

In the other corner we have the data hustlers, salivating: the hoarders and opportunists. This colorful group ranges from entrepreneurs to managers, techies, and board members.

Data prospectors see value, and value is exciting—from more than just a selfish or economic standpoint. We love building the brave new world: increasing productivity and efficiency, decreasing junk mail and its environmental impact, improving healthcare, and suggesting movies and music that will better entertain you. And we love taking on the scientific challenges that get us there.

And yet, even the data hustlers themselves can feel the pain. I was at Walgreens a few years ago, and upon checkout an attractive, colorful coupon spit out of the machine. The product it hawked, pictured for all my fellow

shoppers to see, had the potential to mortify. It was a coupon for Beano, a medication for flatulence. I'd developed mild lactose intolerance but, before figuring that out, had been trying anything to address my symptom. Acting blindly on data, Walgreens' recommendation system seemed to suggest that others not stand so close.

Other clinical data holds a more serious and sensitive status than digestive woes. Once, when teaching a summer program for talented teenagers, I received data I felt would have been better kept away from me. The administrator took me aside to inform me that one of my students had a diagnosis of bipolar disorder. I wasn't trained in psychology. I didn't want to prejudge the student, but there is no "delete" button in the brain's memory banks. In the end, the student was one of my best, and his supposed disorder never seemed to manifest in any perceivable way.

Now we are witnessing the increasing use of location data from cell phones and cars. Some people are getting into serious trouble with their bosses, spouses, and other law enforcement agencies. Tom Mitchell, a professor at Carnegie Mellon University and a world leader in the research and development of machine learning capabilities, wrote in a *Science* article: "The potential benefits of mining such data [from cell phones that track location via GPS] are various; examples include reducing traffic congestion and pollution, limiting the spread of disease, and better using public resources such as parks, buses, and ambulance services. But risks to privacy from aggregating these data are on a scale that humans have never before faced."

These camps will battle over data for decades to come. Data hustlers must hone their radar for land mines, improving their sensitivity to sensitivity. Privacy advocates must see that data-driven technology is a tool that can serve both good and evil—like a knife. Outlawing it completely is not an option. There's no objectively correct resolution; this is a subjective, dynamic arena in which new aspects of our culture are being defined. Dialogue is critical, and a "check here to agree to our lengthy privacy policy that you are too busy to read" does not count as dialogue. Organizations and consumers are not speaking the same language. Striking a balance, together, is society's big new challenge. We have a long way to go.

DATA MINING DOES NOT DRILL DOWN

Exonerate the data scientists and their darling invention. PA in and of itself does not invade privacy—its core process is the *opposite* of privacy invasion. Although it's sometimes called *data mining*, PA doesn't "drill down" to peer at any individual's data. Instead, PA actually "rolls up," learning patterns that hold true in general by way of rote number crunching across the masses of customer records. Data mining often appears to be a culprit when people misunderstand and completely reverse its meaning.

But PA palpably intensifies the battle over data. Why? It ignites fire under data hustlers across the world with a greater and more urgent hunger for more data. Having more data elements per customer means better odds in number crunching's exploration for what will prove most predictive. And the more rows of customer records, the better the predictive model resulting from PA's learning process.

Don't blame the sun when a thirsty criminal steals lemonade. If data rules are fair and right, PA activities that abide by them cannot contribute to abuse or privacy invasion. In this case, PA will be deemed copacetic and be greeted with open arms, and all will be well in our happy futuristic world of prediction. Right?

Fade to black and flash forward to a dystopia. You work in a chic cubicle, sucking chicken-flavored sustenance from a tube. You're furiously maneuvering with a joystick, remotely operating a vehicle on a meteor digging for precious metals. Your boss stops by and gives you a look. "We need to talk about your loyalty to this company."

The organization you work for has deduced that you might be planning to quit. It predicts your plans and intentions, possibly before you have even conceived them.

HP LEARNS ABOUT ITSELF

In 2011, two crackerjack scientists at HP broke ground by mathematically scrutinizing the loyalty of each and every one of their more than 300,000

colleagues. Gitali Halder and Anindya Dey developed predictive models to identify all "Flight Risk" employees, those with a higher expected chance of quitting their jobs.

Retaining employees is core to protecting any organization. After all, an organization's defining characteristic is that it's a collection of members. One of five ideological tenets set forth by a founder of HP is: "We achieve our common objectives through teamwork." Employees contribute complementary skills and take on complementary roles. They learn how to work together. It's bad news when a good one goes. The management of employee turnover is a significant challenge for all companies. For example, another multinational corporation looked to decrease turnover among customer service agents at a call center in Barcelona. Folks would come just to spend the summer in that beautiful city and then suddenly give notice and split. It would help to identify such job applicants in advance.

In this endeavor, the organization is aiming PA inwardly to predict its own staff's behavior, in contrast to the more common activity of predicting its patrons' behavior. As with predicting which customers are most likely to leave in order to target retention efforts, HP predicts which of its staff are likely to leave in order to do the same. In both cases, it's like identifying leaks in a boat's hull in order to patch them up and keep the ship afloat.[1]

PA Application: Employee Retention

1. **What's predicted:** Which employees will quit.
2. **What's done about it:** Managers take the predictions for those they supervise into consideration, at their discretion. This is an example of decision *support* rather than feeding predictions into an automatic decision process.

[1] This and related workforce applications of PA are emerging rapidly enough that the field warranted the 2015 launch of its own annual conference: Predictive Analytics World for Workforce.

Reproduced with permission.

INSIGHT OR INTRUSION?

HP is the iconic success story. It literally started in the proverbial garage and now leads the worldwide manufacturing of personal computers. The company came in as the twenty-seventh largest employer of 2011, amassing $127 billion in revenue, which makes it one of the highest-earning technology companies in the world.

HP is an empire of sorts, but by no means a locked-up citadel. Some working groups report turnover rates as high as 20 percent. On a ship this big, there are bound to be some leaks, especially given the apparent short attention span of today's technology worker.

HP is a progressive analytics leader. Its analytics department houses 1,700 workers in Bangalore alone. They boast cutting-edge analytical capabilities across sales, marketing, supply chain, finance, and human resources (HR)

domains. Their PA projects include customer loss prediction, sales lead scoring, and supplier fraud detection.

Gitali Halder leads HP's analytics team in Bangalore focused on HR applications. With a master's in economics from the Delhi School of Economics and several years of hands-on experience, Halder is your true PA powerhouse. Confident, well spoken, and gregarious, she compels and impresses. Having teamed with HP consultant Anindya Dey, also in Bangalore, the two shine as a well-presented dynamic duo, as evidenced by their polished presentation on this project at the Predictive Analytics World conference in November 2011 in London.

Halder and Dey compiled a massive set of training data to serve as learning material for PA. They pulled together two years of employee data such as salaries, raises, job ratings, and job rotations. Then they tacked on, for each of these employee records, whether the person had quit. Thus, HP was positioned to learn from past experience to predict a priceless gem: which combinations of factors define the type(s) of employees most likely to quit their jobs.

If this project helps HP slow its employee turnover rate, Halder and Dey may stand above the crowd as two of its most valuable employees—or become two of the most resented, at least by select colleagues. Some devoted HP workers are bound to be uncomfortable that their Flight Risk score exists. What if your score is wrong, unfairly labeling you as disloyal and blemishing your reputation?

A whole new breed of powerful HR data emerges: speculative data. Beyond personal, financial, or otherwise private data about a person, this is an estimation of the future and thus speaks to the heart, mind, and intentions of the employee. Insight or intrusion?

It depends on what HP does with it.

FLIGHT RISK: I QUIT!

On the other side of the world, Alex Beaux helps Halder and Dey bring the fruits of their labor to bear upon a select niche of HP employees. It's 10.5 hours earlier in Houston, where Beaux sits as a manager for HP's

internal Global Business Services (GBS). With thousands of staff members, GBS provides all kinds of services across HP to departments that have something they'd like to outsource (even though "outsourcing" to GBS technically still keeps the work within HP).

Beaux, Halder, and Dey set their sights on GBS's Sales Compensation team, since its roughly 300 employees—spread across a few countries—have been exhibiting a high attrition rate of up to 20 percent. A nicely contained petri dish for a pilot field test of Flight Risk prediction, this team provides support for calculating and managing the compensation of salespeople internationally.

The message is clear: Global enterprises are complex! This is not a team of salespeople. It isn't even a regular HR team that supports salespeople. Rather, it is a global team, mostly in Mexico, China, and Poland, that helps various HR teams that support salespeople. And so this project is multilevel: It's the analytical HR management of a team that helps HR (that supports salespeople).

Just read that paragraph five more times and you'll be fine. I once worked on an HP project that predicted the potential demand of its corporate clients—how many computers will the company need to buy, and how much of that need is currently covered by HP's competitors? Working on that project for several months, I was on conference calls with folks from so many working groups named with so many acronyms and across so many time zones that it required a glossary just to keep up.

This organizational complexity means there's great value in retaining sales compensation staff. A lot of overhead must be expended to get each new hire ramped up. Sales compensation team members boast a very specific skill set, since they manage an intricate, large-scale operation. They work with systems that determine the nitty-gritty as to how salespeople are compensated. A global enterprise does not follow an orderly grid designed by a city planner—it takes on a patchwork quality since so much organizational growth comes of buying smaller companies, thus absorbing new sales teams with their own compensation rules. The GBS Sales Compensation team handles an estimated 50 percent of the work to manage sales compensation across the entire global organization.

INSIGHTS: THE FACTORS BEHIND QUITTING

The data showed that Flight Risk depends on some of the things you would expect. For example, employees with higher salaries, more raises, and increased performance ratings quit less. These factors pan out as drivers that decrease Flight Risk. Having more job rotations also keeps employees on board; Beaux conjectures that for the rote, transactional nature of this work, daily activities are kept more interesting with periodic change.

One surprise is that getting a promotion is not always a good thing. Across all of HP, promotions do decrease Flight Risk, but within this Sales Compensation team, where a number of promotions had been associated with relatively low raises, the effect was reversed: Those employees who had been promoted more times were more likely to quit, unless a more significant pay hike had gone along with the promotion.

The analysis is only as good as the data (garbage in, garbage out). In a similar but unrelated project for another company, I predictively modeled how long new prospective hires for a Fortune 1000 business-to-business (B2B) provider of credit information would stay on if hired for call center staffing. Candidates with previous outbound sales experience proved 69 percent more likely to remain on the job at least nine months. Other factors included the number of jobs in the past decade, the referring source of the applicant, and the highest degree attained. This project dodged a land mine, as preliminary results falsely showed new hires without a high school degree were 2.6 times as likely to stay on the job longer. We were only days away from presenting this result to the client—and recommending that the company hire more high school dropouts—when we discovered an unusual combination of errors in the data the client had delivered.[2] Error-prone data—noise—usually just means fewer conclusions will be drawn, rather than strong false ones, but this case was an exceptional perfect storm—a close call!

[2] Encodings for the highest degree attained were inconsistent and the inconsistency corresponded with non-random portions of the dataset. Discovering this was largely serendipitous; with less luck it could easily have continued to go unnoticed.

As for any domain of PA, the predictive model zips up these various factors into a single score—in this case, a Flight Risk score—for each individual. Even if many of these phenomena seem obvious or intuitive, the model is where the subtle stuff comes in: how these elements weigh in relative to one another, how they combine or interact, and which other intuitive hunches that don't pan out should be eliminated. A machine learning process automates these discoveries by crunching the historical data, literally learning from it.

Halder and Dey's Flight Risk model identified $300 million in estimated potential savings with respect to staff replacement and productivity loss across all HP employees throughout all global regions. The 40 percent of HP employees with highest Flight Risk scores included 75 percent of the quitters (a predictive *lift* of 1.9).

I asked the two, who themselves are HP employees, what their own Flight Risk scores were. Had they predicted themselves likely to quit? Halder and Dey are quick to point out that they like their jobs at HP very much, but admit they are in fact members of a high-risk group. This sounds likely, since analytics skills are in high demand.

DELIVERING DYNAMITE

When chemists synthesize a new, unstable element, they must *handle with care*.

HP's Flight Risk scores deploy with extreme caution, under lock and key. Beaux, Halder, and Dey devised a report delivery system whereby only a select few high-level managers who have been trained in interpreting Flight Risk scores and understanding their limitations, ramifications, and confidentiality may view individual employee scores—and only scores for employees under them. In fact, if unauthorized parties got their hands on the report itself, they would find there are no names or identifying elements for the employees listed there—only cryptic identifiers, which the authorized managers have the key to unscramble and match to real names. All security systems have vulnerabilities, but this one is fairly bulletproof.

For the GBS Sales Compensation team of 300 employees, only three managers see these reports. A tool displays the Flight Risk scores in

a user-friendly, nontechnical view that delivers supporting contextual information about each score in order to help explain why it is high or low. The consumers of this analytical product are trained in advance to understand the Flight Risk scores in terms of their accompanying explanations—the factors about the employee that contributed to the score—so that these numbers aren't deferred to as a forceful authority or overly trusted in lieu of other considerations.

A score produced by any predictive model must be taken with a very particular grain of salt. Scores speak to trends and probabilities across a large group; one individual probability by its nature oversimplifies the real-world thing it describes. If I were to miss a single credit card payment, the probability that I'd miss another in the same year may quadruple, based on that factor alone. But if you also take into account that my roof caved in that month (this is a fictional example), your view will change. In general, the complete story for an individual is in fact more than we can ever know. You can see a parallel to another scrutinized practice: diagnosing someone with a psychological disorder and thus labeling them and influencing how they're to be treated.

Over time, the Flight Risk reports sway management decisions in a productive direction. They serve as early warning signals that guide management in planning around loss of staff when it can't be avoided, and working to keep key employees where possible. The system informs what factors drive employee attrition, empowering managers to develop more robust strategies to retain their staffs in order to reduce costs and maintain business continuity.

THE VALUE GAINED FROM FLIGHT RISK

And the results are in. GBS's Sales Compensation staff attrition rates that were above 20 percent in some regions have decreased to 15 percent and continue to trend downward. This success is credited in large part to the impact of Flight Risk reports and their well-crafted delivery.

The project gained significant visibility within HP. Even HP's worldwide vice president of sales compensation heartily applauded the project. Flight

Risk reports continue to make an impact today, and their underlying predictive models are updated quarterly over more recent data in order to remain current.

These pioneers may not realize just how big a shift this practice is from a cultural standpoint. The computer is doing more than obeying the usual mechanical orders to retain facts and figures. It's producing new information that's so powerful, it must be handled with a new kind of care. We're in a new world in which systems not only divine new, potent information but must carefully manage it as well.

Managed well and delivered prudently, Flight Risk scores can perhaps benefit an organization without ruffling too many feathers. Given your established relationship with your boss, perhaps you'd be comfortable if he or she received a Flight Risk score for you, assuming it was considered within the right context. And perhaps it's reasonable and acceptable for an employer to crunch numbers on employee patterns and trends, even without the employees necessarily knowing about it. There's no universally approved ethical framework yet established—the jury is still out on this new case.

But, moving from employment record to criminal record, what if law enforcement officers appeared at your door to investigate you, Future Crime Risk report in hand?

PREDICTING CRIME TO STOP IT BEFORE IT HAPPENS

> *What if you could shift the intelligence paradigm from "sense, guess, and respond" to "predict, plan, and act"?*
>
> —Sgt. Christopher Fulcher, Chief Technology Officer of the Vineland, New Jersey, Police Department

Cops have their work cut out for them. Crime rates may ebb and flow, but law enforcement by its nature will always face the impossible challenge of optimizing the deployment of limited resources such as patrolling officers and perusing auditors.

Police deploy PA to predict the location of crime and to direct cops to patrol those areas accordingly. One system, backtested on two years of

data from Santa Cruz, California, correctly predicted the locations of 25 percent of burglaries. This system directs patrols today, delivering 10 hot spots each day within this small city to send police vehicles to. The initiative was honored by *Time* magazine as one of the 50 best inventions of 2011.

PA APPLICATION: CRIME PREDICTION (AKA *PREDICTIVE POLICING*)

1. **What's predicted:** The location of a future crime.
2. **What's done about it:** Police patrol the area.

Another crime prediction system, revealed at a 2011 conference by Chief Information Officer Stephen Hollifield of the Richmond, Virginia, police department, serves up a crime-fighting display that marks up maps by the risk of imminent crime and lists precincts, neighborhoods, and crime types by risk level. Since this system's deployment, Richmond crime rates have decreased. Similar systems are in development in Chicago; Los Angeles; Vineland, New Jersey; and Memphis, where prediction is credited with reducing crime by 31 percent. In 2009, the U.S. National Institute of Justice awarded planning grants to seven police departments to create crime prediction capabilities.

Lightning strikes twice. The predictive models leverage discoveries such as the trend that crimes are more—not less—likely to soon reoccur in nearby locations, as detected in Santa Cruz. In Richmond, the predictive model flags for future crime based on clues such as today's city events, whether it's a payday or a holiday, the day of the week, and the weather.

What's not to like? Law enforcement gains a new tool, and crime is defrayed. Any controversy over these deployments appears relatively tame. Even the American Civil Liberties Union gave this one a nod of the head. No harm, no foul.

In fact, there's one type of crime that elicits loud complaints when predictive models *fail* to detect it: fraud. To learn more, see the sidebar on fraud detection. After the sidebar, we continue on to explore how crime-predicting computers inform how much time convicts spend in prison.

SPECIAL SIDEBAR ON FRAUD DETECTION

Criminals can be such nice guys. I became friends with one in 1995. I was pursuing my doctorate in New York City and he was the new boyfriend of my girlfriend's sister. Extremely charismatic and supposedly a former professional athlete, the crook wooed, wowed, and otherwise ingratiated himself into our hearts and home. I'll never forget the really huge, fun dinner he treated us to at the famous Italian restaurant Carmine's. I didn't think twice about letting him use my apartment when I went on a vacation.

A year or two later I discovered he had acquired my Social Security number, stolen my identity, and soiled my sparkly clean credit rating. He had started a small water bottling business in the Los Angeles area, posing as me. Despite being a decade older than I, on the wrong coast, and not even attempting to emulate my signature, he had attained numerous credit accounts, including credit cards and leases on water bottling equipment. After building considerable debt, he abandoned the business and defaulted on the payments. It took a couple of years of tedious paperwork to clear my name and clean up my credit rating.

Where's a good predictive model when you need one? Why couldn't these credit applications have been flagged or quarantined, checking with me by way of the contact information established in my credit files? After all, once all the evidence was gathered and submitted, most auditors immediately perceived the case as obvious fraud.

While some deployments of PA give rise to concern, the absence thereof does as well. Enter *fraud detection*.

A WOLF IN SHEEP'S CLOTHING

Fraud, defined as "intentional deception made for personal gain," is the very act of a wolf dressing up in sheep's clothing. It's when someone pretends to be someone else or to be authorized to do something the fraudster is not authorized to do. A student copies another's homework, a sumo wrestler throws a match, an online

SPECIAL SIDEBAR ON FRAUD DETECTION (CONTINUED)

gambler cheats with illegal information as part of an inside job, inauthentic Twitter accounts spread misinformation about a political candidate, or a death is faked in order to make a claim against a life insurance policy. All such crimes have been detected analytically.

It's a good time to be a fraudster since they enjoy a massive, expanding stomping ground: the Internet, a transaction infrastructure for global commerce. But by connecting to everybody, we've connected to folks with malicious intent. The easier it is to conduct consumer and business transactions, the easier it is to fake them as well. And with the buyer, seller, goods, and payment spread across four different physical locations, there is an abundance of vulnerabilities that may be exploited.

As transactions become increasingly numerous and automated, criminal opportunities abound. Fraudulent transactions such as credit card purchases, tax returns, insurance claims, warranty claims, consumer banking checks, and even intentionally excessive clicks on paid ads incur great cost. The National Insurance Crime Bureau says that insurance criminals steal over $30 billion annually, making such fraud the second most costly white-collar crime in the United States—behind tax evasion—resulting in $200 to $300 of additional insurance premiums per U.S. household; we are paying these criminals out of our pockets.

"It is estimated that the nation's banks experience over $10 billion per year in attempted check fraud," says former Citizens Bank Vice President Jay Zhou, now a data mining consultant. Credit card fraud losses approach $5 billion annually in the United States, and Medicaid fraud is estimated to be the same amount for New York State alone. According to the most recent report published by the Federal Trade Commission, 2011 brought over 1.8 million complaints of fraud, identity theft, or other intentional deceit in business, about 40 percent more than in 2010.

(*continued*)

SPECIAL SIDEBAR ON FRAUD DETECTION (CONTINUED)

Aggregate fraud loss in the United States sees estimates from $100 billion to $1 trillion.

Prediction helps. Predictively scoring and ranking transactions dramatically boosts fraud detection. A team of enforcement workers can inspect only a fixed number of suspected transactions each week. For example, Progressive Insurance employs about 200 "special investigations professionals" on this task. Delivering a more precisely identified pool of candidate transactions—fewer false alarms (false positives)—renders their time more effectively spent; more fraud is detected, and more losses are prevented or recouped.

PA APPLICATION: FRAUD DETECTION[3]

1. **What's predicted:** Which transactions or applications for credit, benefits, reimbursements, refunds, and so on are fraudulent.
2. **What's done about it:** Human auditors screen the transactions and applications that are predicted most likely to be fraudulent.

Math is fighting back. Most large—and many medium-sized—financial institutions employ fraud detection. For example, Citizens Bank developed a fraud prediction model that scores each check, predicting well enough to decrease fraud loss by 20 percent. One automobile insurance carrier showed that PA delivers 6.5 times the fraud detection capacity of that attained with no means to rank or score insurance claims. Online transaction giant PayPal suffered an almost 20 percent fraud rate soon after it was launched, a primary threat to its

[3] Rather than performing prediction in the conventional sense of the word, this application of PA performs detection. As with predicting the future, such an application imperfectly infers an unknown.

SPECIAL SIDEBAR ON FRAUD DETECTION (CONTINUED)

success. Fraud detection methods brought the rate down to a reported less than 1 percent. The people behind each of these stories have spoken at the Predictive Analytics World conference, as have those telling similar stories from 1–800-FLOWERS, Activision, the Belgian government, the U.S. Postal Service, the Internal Revenue Service (IRS), administrators of Medicare and Medicaid, and a leading high-tech company that catches warranty claims from repair shops that didn't actually do the service at all.

GOVERNMENT, PROTECT THYSELF

The government is working hard on fraud management—but unlike its efforts enforcing against crimes like theft and assault, most of this effort isn't focused on protecting you, or even any business. When it comes to fraud, the U.S. government is fighting to protect its own funds. In fact, fraud detection is the most evident government application of PA, providing a means to decrease loss in the face of tightening budgets.

Elder Research (John Elder's company) headed a fraud modeling project for the IRS that increased the capacity to detect fraudulent returns by a factor of 25 for a certain targeted segment. A similar effort has been reported by the Mexican Tax Administration, which has its own Risk Models Office.

The U.S. Defense Finance and Accounting Service, responsible for disbursing nearly all Department of Defense funds, executes millions of payments on vendor invoices. Dean Abbott, a top PA consultant who has also consulted for the IRS, led the development of a predictive model capable of detecting 97 percent of known cases of fraudulent invoices. The model scores invoices based on factors such as the time since the last invoice, the existence of other payees at the same postal address, whether the address is a P.O. box, and whether the vendor submitted invoices out of order.

(*continued*)

SPECIAL SIDEBAR ON FRAUD DETECTION (CONTINUED)

Beyond these possible signs of fraud, other innovative clues turbocharge the predictive model, helping determine which cases are flagged. 1-800-FLOWERS improved its ability to detect fraud by considering the social connections between prospective perpetrators. In fact, one fraud scheme can't be detected without this kind of social data. (Oxymoron, anyone?) A group of criminals open financial accounts that improve their respective credit ratings by transferring funds among themselves. Since the money transfers take place only between these accounts, the fraudsters need not spend any real money in conducting these transactions; they play their own little zero-sum game. Once each account has built up its own supposedly legitimate record, they strike, taking out loans, grabbing the money, and running. These schemes can be detected only by way of social data to reveal that the network of transactors is a closed group.

Naturally, criminals respond by growing more creative.

THE FRAUD DETECTION ARMS RACE

> *The fraudsters were also good, and nimble, too, devising new scams as soon as old ones were compromised.*
>
> —Steven Levitt and Stephen Dubner, *SuperFreakonomics*

Just as competing businesses in the free market push one another to better themselves, fraud detection capabilities drive criminals toward self-improvement by the design of smarter techniques. The act of fraud strives to be stealthy, sneaking under the predictive model's radar. As with the possibility of superbacteria emerging from the overuse of antibiotics, we are inadvertently creating a stronger enemy.

But there's good news. The white hats sustain a great advantage. In addition to exerting human creativity like our opponents, we have the data with which to hone fraud detection models. A broad set of data containing historical examples of both fraudulent and legitimate

SPECIAL SIDEBAR ON FRAUD DETECTION (CONTINUED)

transactions intrinsically encodes the inherent difference between the two. PA is the very means by which to discover this difference from data. And so, beyond storing and indexing a table of "signatures" that betray the perpetration of known fraud schemes, the modeling process generates detection schemes that cast a wider net. It predicts forthcoming forms of fraud by generalizing from previously observed examples. This is the defining characteristic of a learning system.

THIS MEANS WAR

It's a war like any other. In fact, cyberwarfare itself follows the same rules. PA bolsters information security by detecting hackers and viruses that exploit online weaknesses, such as system bugs or other vulnerabilities. After all, the Internet's underlying networking technology, TCP/IP, is a platform originally designed only for interactions between mutually entrusted parties. As the broad, commercial system it evolved to be, the Internet is, underneath the hood, something of a slapped-together hack with regard to security. Like an unplanned city, it functions, but like a Social Security number awaiting discovery in an unlocked drawer, it holds intrinsic weaknesses.

PA APPLICATION: NETWORK INTRUSION DETECTION

1. **What's predicted:** Which low-level Internet communications originate from imposters.
2. **What's done about it:** Block such interactions.

PA boosts detection by taking a qualitatively new step in the escalating arms race between white and black hats. A predictive detection system's field of vision encompasses a broad scope of potential attacks that cannot be known by perpetrators, simply because they don't

(*continued*)

SPECIAL SIDEBAR ON FRAUD DETECTION (CONTINUED)

have access to the same data used to develop the predictive model. Hackers can't know if their techniques will be detected. PA's deployment brings a qualitative change in the way we compete against malicious intent.

But beware! Another type of fraud attacks you and every one of us, many times a day. Are you protected?

LIPSTICK ON A PIG

> *An Internet service cannot be considered truly successful until it has attracted spammers.*
>
> —Rafe Colburn, Internet development thought leader

Alan Turing (1912–1954), the father of computer science, proposed a thought experiment to explore the definition of what would constitute an "intelligent" computer. This so-called *Turing test* allows people to communicate via written language with someone or something hidden behind a closed door in order to formulate an answer to the question: Is it human or machine? The thought experiment poses this tough question: If, across experiments that randomly switch between a real person and a computer, subjects can't correctly tell human from machine more often than the 50 percent correctness one could get from guessing, would you then conclude that the computer, having thereby passed the test by proving it can trick people, is intelligent? I'll give you a hint: There's no right answer to this philosophical conundrum.

In practice, computers attempt to fool people for money every day via e-mail. It's called spam. As with androids in science fiction movies like *Aliens* and *Blade Runner*, successful spam makes you believe. Spam's cousin, *phishing*, persuades you to divulge financial secrets. *Spambots* take the form of humans in social networks and dating sites in

SPECIAL SIDEBAR ON FRAUD DETECTION (CONTINUED)

order to grab your attention. And spammy Web pages trick search engines into pointing you their way.

Spam filters, powered by PA, are attempting their own kind of Turing test every day at an e-mail in-box near you.

PA APPLICATION: SPAM FILTERING

1. **What's predicted:** Which e-mail is spam.
2. **What's done about it:** Divert suspected e-mails to your spam e-mail folder.

Unfortunately, in the spam domain, white hats don't exclusively own the arms race advantage. The perpetrators can also access data from which to learn, by testing out a spam filter and reverse engineering it with a model of their own that predicts which messages will make it through the filter. University of California, Berkeley researchers showed how to do this to render one spam filter useless.

ARTIFICIAL ARTIFICIAL INTELLIGENCE

In contrast to these precocious computers, we sometimes witness a complete role reversal: a person pretends to be a machine. The Mechanical Turk, a hoax in the eighteenth century, created the illusion of a machine playing chess. The Turk was a desk-sized box that revealed mechanical gears within and sported a chessboard on top. Seated behind the desk was a mannequin whose arm would reach across the board and move the pieces. A small human chess expert who did not suffer from claustrophobia (chess is a long game) hid inside the desk, viewing the board from underneath and manipulating the mannequin's arm. Napoleon Bonaparte and Benjamin Franklin had the pleasure of losing to this wonder of innovation—I mean, this crouching, uncomfortable imposter.

(*continued*)

SPECIAL SIDEBAR ON FRAUD DETECTION (CONTINUED)

In the modern-day equivalent, human workers perform low-level tasks for the Amazon Mechanical Turk, a crowdsourcing website by Amazon.com that coordinates hundreds of thousands of workers to do "things that human beings can [still] do much more effectively than computers, such as identifying objects in a photo . . . [or] transcribing audio recordings." Its slogan is "Artificial Artificial Intelligence." (This reminds me of the vegetarian restaurant with "mock mock duck" on the menu—I swear, it tastes exactly like mock duck.) As NASA put it in 1965 when defending the idea of sending humans into space, "Man is the lowest-cost, 150-pound, nonlinear, all-purpose computer system which can be mass-produced by unskilled labor."

But for some tasks, we don't have to pretend anymore. Everything changed in 1997 when IBM's Deep Blue computer defeated then world chess champion Garry Kasparov. Predictive modeling was key. No matter how fast the computer, perfection at chess is impossible, since there are too many possible scenarios to explore. Various estimates agree there are more chess games than atoms in the universe, a result of the nature of exponential growth. So the computer can look ahead only a limited number of moves, after which it needs to stop enumerating scenarios and evaluate game states (boards with pieces in set positions), predicting whether each state will end up being more or less advantageous.

PA APPLICATION: PLAYING A BOARD GAME

1. **What's predicted:** Which game board state will lead to a win.
2. **What's done about it:** Make a game move that will lead to a state predicted to lead to a win.

Upon losing this match and effectively demoting humankind in its standoff against machines, Kasparov was so impressed with the strategies Deep Blue exhibited that he momentarily accused IBM of

SPECIAL SIDEBAR ON FRAUD DETECTION (CONTINUED)

cheating, as if IBM had secretly hidden another human grandmaster chess champion, squeezed in there somewhere between a circuit board and a disk drive like a really exorbitant modern-day Mechanical Turk. And so IBM had passed a "mini Turing test" (not really, but the company did inadvertently fool a pretty smart guy).

From this upset emerges a new form of chess fraud: humans who employ the assistance of chess-playing computers when competing in online chess tournaments. And yet another arms race begins, as tournament administrators look to detect such cheating players. This brings us full circle, back to computers that pose as people, as is the case with spam.

So computer "intelligence" has flipped the meaning of fraud on its head, reversing it. Rather than a chess-playing person pretending to be a machine (the Mechanical Turk), we have a machine masking as a person (cheating in human chess tournaments). It's rather like *Star Trek*'s Commander Data, an emotionally stunted android afflicted with the Pinocchio Syndrome of wanting to be more human.

THE DATA OF CRIME AND THE CRIME OF DATA

PA has taken on an enormous crime wave. It is central to tackling fraud and promises to bolster street-level policing as well.

In these efforts, PA's power optimizes the assignment of resources. Its predictions dictate how enforcers spend their time—which transactions auditors search for fraud and which street corners cops search for crime.

But how about giving PA the power to help decide who belongs in prison?

To help make these tough decisions, judges and parole boards consult predictive models. To build these models, Philadelphia's Adult Probation and Parole Department enlisted a professor of statistics and criminology from the University of Pennsylvania. The parole department's research director,

Ellen Kurtz, told *The Atlantic*, "Our vision was that every single person, when they walked through the door [of a parole hearing], would be scored by a computer" as to his or her risk of recidivism—committing crime again.

Oregon launched a crime prediction tool to be consulted by judges when sentencing convicted felons. The tool is on display for anyone to try out. If you know the convict's state ID and the crime for which he or she is being sentenced, you can enter the information on the Oregon Criminal Justice Commission's public website and see the predictive model's output: the probability the offender will be convicted again for a felony within three years of being released.

PA Application: Recidivism Prediction for Law Enforcement

1. **What's predicted:** Whether a prosecuted criminal will offend again.
2. **What's done about it:** Judges and parole boards consult model predictions when making decisions about an individual's incarceration.

The predictive model behind Oregon's tool performs admirably. Machine learning generated the model by processing the records of 55,000 Oregon offenders across five years of data. The model then validated across 350,000 offender records across 30 years of history. Among the least risky tenth of criminals—those for whom the model outputs the lowest predictive scores—recidivism is just 20 percent. Yet among the top fifth receiving the highest scores, recidivism will probably occur; over half of these offenders will commit a felony again.

Law enforcement's deployment of PA to predict for individual convicts is building steam. In these deployments, PA builds upon and expands beyond a longstanding tradition of crime statistics and standard actuarial models. Virginia's and Missouri's sentencing guidelines also prescribe the consideration of quantitative risk assessment, and Maryland has models that predict murder. The machine is a respected adviser that has the attention of judges and parole boards.

Humans could use some help with these decisions, so why not introduce an objective, data-driven voice into the process? After all, studies have shown that arbitrary extraneous factors greatly affect judicial decisions. A joint study

by Columbia University and Ben Gurion University (Israel) showed that hungry judges rule negatively. Judicial parole decisions immediately after a food break are about 65 percent favorable, but then drop gradually to almost zero percent before the next break. If your parole board judges are hungry, you're much more likely to stay in prison.

With this reasoning accepted, the convict's future now rests in nonhuman hands. Given new power, the computer can commit more than just prediction errors—it can commit injustice, previously a form of misjudgment that only people were in a position to make. It's a whole new playing field for the machine, with much higher stakes. Miscalculations in this arena are more costly than for other applications of PA. After all, the price is not as high when an e-mail message is wrongly incarcerated in the spam folder or a fraud auditor's time is wasted on a transaction that turns out to be legitimate.

MACHINE RISK WITHOUT MEASURE

In the movie *Minority Report*, Tom Cruise's science fiction cop tackles and handcuffs individuals who have committed no crime (yet), proclaiming stuff like: "By mandate of the District of Columbia Precrime Division, I'm placing you under arrest for the future murder of Sarah Marks and Donald Dubin." Rather than the punishment fitting the crime, the punishment fits the precrime.

Cruise's bravado does not go unchecked. Colin Farrell's Department of Justice agent confronts Cruise, and the two brutes stand off, mano a mano. "You ever get any false positives?" accuses Farrell.

A *false positive*, aka *false alarm*, is when a model incorrectly predicts yes when the correct answer is no. It says you're guilty, convicting you of a crime you didn't (or in this case, won't) commit.

As self-driving cars emerge from Google and BMW and begin to hit the streets, a new cultural acceptance of machine risk will emerge as well. The world will see automobile collision casualty rates decrease overall and eventually, among waves of ire and protest, will learn to accept that on some occasions the computer is to blame for an accidental death.

But when a criminal who would not reoffend is kept in prison because of an incorrect prediction, we will never have the luxury of knowing. You can

prove innocent a legitimate transaction wrongly flagged as fraudulent, but an incarcerated person has no recourse to disprove unjust assumptions about what his or her future behavior outside prison would have been. If you prevent something, how can you be certain it was ever going to happen?

We're entrusting machines to contribute to life-changing decisions for which there can be no accountability: We can't measure the quality of these decisions, so there's no way to determine blame. We've grown comfortable with entrusting humans, despite their cherished fallibility, to make these judgment calls. A culture shift is nigh as we broaden this sacred circle of trust. PA sometimes makes wrong predictions but often proves to be less wrong than people. Bringing PA in to support decision making means introducing a new type of bias, a new fallibility, to balance against that of a person.

The development of computerized law enforcement presents extraordinarily tough ethical quandaries:

- Does the application of PA for law enforcement fly in the face of the very notion of judging a person as an individual? Is it unfair to predict a person's risk of bad behavior based on what other people—who share certain characteristics with that person—have done? Or, isn't the prediction by a human (e.g., a judge) of one's future crimes also intrinsically based only on prior observations of others, since humans learn from experience as well?
- A crime risk model dehumanizes the prior offender by paring him or her down to the extremely limited view captured by a small number of characteristics (variables input to a predictive model). But, if the integration of PA promises to lower the overall crime rate—as well as the expense of unnecessary incarceration—is this within the acceptable realm of compromises to civil liberties (on top of incarceration) that convicts endure?
- With these efforts under way, should not at least as much effort go into leveraging PA to improve offender rehabilitation; for example, by targeting those with the highest risk of recidivism? (In one groundbreaking case, the Florida Department of Juvenile Justice does just this—see Central Table 5.)

PA threatens to attain too much authority. Like an enchanted child with a Magic 8 Ball toy (originated in 1950), which is designed to pop up a *random* answer to a yes/no question, insightful human decision makers could place a great deal of confidence in the recommendations of a system they do not deeply understand. What may render judges better informed could also sway them toward less active observation and thought, tempting them to defer to the technology as a kind of crutch and grant it undue credence. It's important for users of PA—the judges and parole board members—to keep well in mind that it bases predictions on a much more limited range of factors than are available to a person.

THE CYCLICITY OF PREJUDICE

Yet another quandary lurks. Although science promises to improve the effectiveness and efficiency of law enforcement, when you formalize and quantify decision making, you inadvertently instill existing prejudices against minorities. Why? Because prejudice is cyclic, a self-fulfilling prophecy, and this cycling could be intensified by PA's deployment.

Across the United States, crime prediction systems calculate a criminal's probability of recidivism based on things like the individual's age, gender, and neighborhood, as well as prior crimes, arrests, and incarcerations. No government-sponsored predictive models explicitly incorporate ethnic class or other minority status.

However, ethnicity creeps into the model indirectly. Philadelphia's recidivism prediction model incorporates the offender's ZIP code, known to highly correlate with race. For this reason, redlining, the denying of services by banks, insurance companies, and other businesses by geographical region, has been largely outlawed in the United States.

Similarly, terrorist prediction models factor in religion. Levitt and Dubner's book *SuperFreakonomics* (HarperCollins, 2009) details a search for suspects among data held by a large UK bank. Informed in part by attributes of the September 11 perpetrators, as well as other known terrorists, a fraud detection analyst at the bank pinpointed a very specific group of customers to forward to the authorities. This *microsegment* was defined by factors such as

the types of bank accounts opened, existence of wire transfers and other transactions, record of a mobile phone, status as a student who rents, and a lack of life insurance (since suicide nullifies the policy). But to get the list of suspects down to a manageable size, the analyst filtered out people with non-Muslim names, as well as those who made ATM withdrawals on Friday afternoons—admittedly a proxy for practicing Muslims. Conceptually, this may not be a huge leap from the internment of suspected enemies of the state, although it should be noted that this was not a government-sponsored analysis. While this work has been criticized as an "egregious piece of armchair antiterrorism," the bank analyst who delivered the suspect list to the authorities may exert power by way of his perceived credibility as a bank representative.

But even if such factors are disallowed for prediction, it's still a challenge to avoid involving minority status.

Bernard Harcourt, a professor of both political science and law at the University of Chicago and author of *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*, told *The Atlantic* that minority group members discriminated against by law enforcement, such as by way of profiling, are proportionately more likely to show a prior criminal record (e.g., since they may be screened more often), which artificially inflates the minority group's incidence of criminal records. Rather than race being a predictor of prior offenses, prior offenses are indicative of race. By factoring in prior offenses to predict future crimes, "you just inscribe the racial discrimination you have today into the future." It's a cyclic magnification of prejudice's already self-fulfilling prophecy.

Even Ellen Kurtz, who champions the adoption of the crime model in Philadelphia, admits, "If you wanted to remove everything correlated with race, you couldn't use anything. That's the reality of life in America."

But don't make data a scapegoat. It isn't solely a petri dish in which racial discrimination grows—it's also a tool that serves the fight against discrimination. Government departments outside law enforcement, such as the Federal Housing Finance Agency, the Education Department, and the Department of Housing and Urban Development, collect data for the very purpose of detecting discriminatory practices in banking loans, public education, affordable housing, and employment opportunities.

Within law enforcement, the math getting us in trouble could also remedy the problem by quantifying prejudice. However, that could be done only by introducing the very data element that—so far—remains outside the analysis, albeit inside the eye of every profiling police officer: race. Technically, there could be an analytical means to take this on if race were input into the system. This would require addressing new questions and debates analogous to those that arise with the implementation of equal-opportunity practices.

GOOD PREDICTION, BAD PREDICTION

Privacy is a compromise between the interests of the government and the citizen.
—Eric Schmidt, former Executive Chairman and CEO, Google

Information technology has changed just about everything in our lives. . . . But while we have new ethical problems, we don't have new ethics.
—Michael Lotti

When we think in terms of power, it is clear we are getting a raw deal: We grant private entities—with no interest in the public good and no public accountability—greater powers of persuasion than anyone has ever had before and in exchange we get free e-mail.
—Alexander Furnas, writer for *The Atlantic*

With great power comes great responsibility.
—Spider-Man's wise uncle (paraphrasing the Bible, Voltaire, and others)

Pregnancy prediction faces the opposite dilemma of that faced by crime prediction. Crime prediction causes damage when it predicts *wrong*, but predicting sensitive facts like pregnancy can cause damage when it's *right*. Like X-ray glasses, PA unveils new hot-button data elements for which all the fundamental data privacy questions must be examined anew. Sherlock Holmes, as well as his modern-day doppelganger Dr. Gregory House, size you up and embarrass you: A few scuff marks on your shoe and the detective knows you're having an affair. Likewise, no one wants her pregnancy unwittingly divulged; it's safe to assume organizations generally don't wish to divulge it, either.

It's tempting to write off these matters as benign in comparison to the qualms of crime prediction. KDnuggets, a leading analytics portal, took a poll: "Was Target wrong in using analytics to identify pregnant women from changes in their buying behavior?" The results were 17 percent "Yes," 74 percent "No," and 9 percent "Not sure" among the analytics community. One written comment pointed out that intent is relevant, asking, "When I yield a seat on a train to elderly people or a pregnant woman, am I 'trying to infer sensitive personal data such as pregnancy or elderliness'? Or just trying to provide the person with her needs?"

But knowledge of a pregnancy is extremely potent, and leaking it to the wrong ears can be life-changing indeed. As one online pundit proclaimed, imagine the pregnant woman's "job is shaky, and your state disability isn't set up right yet, and, although she's working on that, to have disclosure could risk the retail cost of a birth ($20,000), disability payments during time off ($10,000 to $50,000), and even her job."

As with pregnancy, predictive models can also ascertain minority status—from behavior online, where divulging demographics would otherwise come only at the user's discretion. A study from the University of Cambridge shows that race, age, sexual orientation, and political orientation can be determined with high levels of accuracy based on one's Facebook likes. This capability could grant marketers and other researchers access to unvolunteered demographic information.

Google itself appears to have sacrificed a significant boon from predictive modeling in the name of privacy by halting its work on the automatic recognition of faces within photographs. When he was Google's CEO, Eric Schmidt stated his concern that facial recognition could be misused by organizations that identify people in a crowd. This could, among other things, ascertain people's locations without their consent. He acknowledges that other organizations will continue to develop such technology, but Google chose not to be behind it.

Other organizations agree: Sometimes it's better not to know. John Elder tells of the adverse reaction from one company's HR department when the idea of predicting employee death was put on the table. Since death is one

way to lose an employee, it's in the data mix. In a meeting with a large organization about predicting employee attrition, one of John's staff witnessed a shutdown when someone mentioned the idea. The project stakeholder balked immediately: "Don't show us!" Unlike healthcare organizations, this HR group was not meant to handle and safeguard such prognostications.

Predicting death is so sensitive that it's done secretly, keeping it on the down low even when done for benevolent purposes. One top-five health insurance company predicts the likelihood an elderly insurance policyholder will pass away within 18 months, based on clinical markers in the insured's recent medical claims. On the surface, this sounds potentially dubious. With the ulterior motives of health insurance often under scrutiny, one starts to imagine the terrible implications. Might the insurance company deny or delay the coverage of treatment based in part on how likely you are to die soon anyway? Not in this case. The company's purposes are altruistic. The predictions serve to trigger end-of-life counseling (e.g., regarding living wills and palliative care). An employee of the company told me the predictive performance is strong, and the project is providing clear value for the patients. Despite this, those at the company quake in their boots that the project could go public, agreeing only to speak with me under the condition of anonymity. "It's a very sensitive issue, easily misconstrued," the employee said.

The media goes too far when it sounds alarms that imply PA ought to be sweepingly indicted. To incriminate deduction would be akin to outlawing thought. It's no more than the act of figuring something out. If I glance into my friend's shopping cart and, based on certain items, draw the conclusion that she may be pregnant, have I just committed a *thoughtcrime*—the very act enforced against by Big Brother in George Orwell's *Nineteen Eighty-Four*? And so the plot twists, since perhaps critics of Target who would compare this kind of analysis to that of Big Brother are themselves calling the kettle black by judging Target for thoughtcrime. Pregnancy prediction need not be viewed as entirely self-serving—as with any marketing, this targeting does have potential to serve the customer. In the end, with all his eccentricities,

Sherlock Holmes is still our hero, and his revealing deductions serve the greater good.

"Privacy and analytics are often publicly positioned as mortal enemies, but are they really?" asks Ari Schwartz of the U.S. Department of Commerce's National Institute of Standards and Technology. Indeed, some data hustlers want a free-for-all, while others want to throw the baby out with the bathwater. But Schwartz suggests, "The two worlds may have some real differences, but can probably live a peaceful coexistence if they simply understand where the other is coming from."

It's not what an organization comes to know; it's what it *does* about it. Inferring new, powerful data is not itself a crime, but it does evoke the burden of responsibility. Target does know how to benefit from pregnancy predictions without actually divulging them to anyone (the alleged story of the pregnant teen is at worst an individual albeit significant gaffe). But any marketing department must realize that if it generates quasimedical data from thin air, it must take on, with credibility, the privacy and security practices of a facility or department commonly entrusted with such data. *You made it, you manage it.*

PA is an important, blossoming science. Foretelling your future behavior and revealing your intentions, it's an extremely powerful tool—and one with significant potential for misuse. It's got to be managed with extreme care. The agreement we collectively come to for PA's position in the world is central to the massive cultural shifts we face as we fully enter and embrace the information age.

THE SOURCE OF POWER

New questions arise as we move from predicting the repeat offenses of convicts to the discovery of new potential suspects within the general populace of civilians. The following sidebar on *automatic suspect discovery* brings these questions to the surface, after which the next chapter turns to the source of predictive power—data—and explores the most bizarre insights it reveals, and how easy it is to be fooled by it.

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA

Synopsis: *It's a foregone conclusion that the world's largest spy organization running the country's largest surveillance data center and employing the world's largest number of PhD mathematicians considers predictive analytics (PA) a strategic priority. Can the NSA use machine learning supercomputers to fight terrorism—and can other agencies do so to fight crime in general—without endangering civil liberties?*

Today's data privacy debate falters, because both sides are underinformed.

The NSA has endured intense scrutiny and suffered heavy backlash over its mass data collection that was unveiled in detail by whistleblower Edward Snowden in 2013. But don't give too much credence to the news or even the books—public discourse leaves out the greatest power law enforcement stands to gain from this data.

SUMMARY OF THE MAINSTREAM DEBATE REGARDING NSA DATA COLLECTION:

Privacy advocates: The NSA is violating civil liberties by collecting data on a massive scale about private citizens, including the majority who are not even suspected of any wrongdoing. Access to this data, whether in-house or by proxy via telecom companies, facilitates arbitrary snooping.

The NSA (and supportive legislators): We require comprehensive data in-house so we can rapidly investigate specific individuals when they become of interest. We do not inspect the activities of ordinary civilians in general.

This contentious dialogue only touches on half the story. Both sides fail to address what's really at stake for law enforcement: *Data empowers*

(*continued*)

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

not only the investigation of established suspects, but also the discovery of new suspects. I would like to propose the following term for this emerging form of data-driven law enforcement:

> **Automatic Suspect Discovery (ASD)**—The identification of previously unknown potential suspects by applying PA to flag and rank individuals according to their likelihood to be worthy of investigation, either because of their direct involvement in, or relationship to, criminal activities.
>
> ***A note on automation:*** *ASD flags new persons of interest who may then be elevated to suspect by an ensuing investigation. By the formal law enforcement definition of the word, an individual would not be classified as a suspect by a computer, only by a law enforcement officer.*

ASD provides a novel means to unearth new suspects. Using it, law enforcement can hunt scientifically, more effectively targeting its search by applying PA, the same state-of-the-art, data-driven technology behind fraud detection, financial credit scoring, spam filtering, and targeted marketing.

THE SPY WHO LOVED MY DATA

To harness this potential, law enforcement needs the whole haystack. The government doesn't desire data about you just to spy at will—on the off chance you turn out to be a suspect. Rather, they actually require this data as a baseline in order to pursue their greater objective with ASD. This approach relies on wide-scale data access, even including data about both you and me—a full regimen of data about normal, innocent civilian activity unrelated to crime of any sort. Mathematically speaking, the broader a swath of noncriminal cases fed into the analysis, the better it works.

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

Given this, ASD only amplifies the stakes of the contentious security-versus-privacy debate; both sides are bound to dig in and redouble their conviction. The promise of a novel technique for suspect discovery emboldens law enforcement's rationale for collecting as much data as possible. On the other side, privacy advocates perceive law enforcement's now stronger incentive as an even greater cause for alarm. Viewing the bulk collection of personal data itself as a violation of civil liberties, they argue the price is too high—especially given that any quantitative approach such as ASD cannot guarantee results a priori.

HOW IT WORKS: SHRINK THE HAYSTACK

Law enforcement (antiterrorism or otherwise) is a numbers game, a quest to find needles in the haystack that is the general population.

(*continued*)

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

The working hours spent by agents, officers, and analysts constitute a precious, finite resource that must be allocated as effectively as possible. As staff collects evidence, follows leads, and studies forensics, there is no magic oracle to focus these efforts and ensure the quest is efficient. But PA can better target a portion of the work.

PA APPLICATION: AUTOMATIC SUSPECT DISCOVERY (ASD)

1. **What's predicted:** Whether an individual is a "person of interest."
2. **What's done about it:** Individuals with a sufficiently high predictive score are considered or investigated.

As with fraud detection, prediction shrinks the haystack to be searched. This multiplies the effectiveness of available human resources. By focusing time on the top echelon, those with the highest predictive scores, an investigator is more likely to come across worthy suspects. While it is reasonable to assume ASD pays off over time, investigators must understand the odds have only shifted; it's not a magic crystal ball. Most targets of investigation still turn out to be innocent—that is to say, the false positive rate will be lowered but by no means eliminated; the haystack is smaller but still large.

For best results, ASD may be applied repeatedly over a batch of predictive objectives. Its success depends on creatively defining *person of interest*, that is, the class of potential suspect being sought. For example, to predict known perpetrators of a rare crime such as terrorism, there may be too few known positive examples—"needles"—with which to train the predictive model. Therefore, ASD may be more likely to succeed when targeting instead a broader category of "interesting" persons, which could be defined as, for example, members

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

of an active surveillance group or persons with certain links to key criminal networks.[4]

As with all application areas, PA learns from data that encodes both positive and negative cases—in the case of ASD, both the known needles and the vast haystack, respectively. The analytical number-crunching process builds models (e.g., patterns or other formulations) to distinguish needles from hay. Models are then used to score each individual according to the probability of being a person of interest. This is the very purpose and function of core PA methods, such as *decision trees* and *ensemble models* (covered in Chapters 4 and 5, respectively).

EXAMPLE PATTERNS: WHAT IT COULD DISCOVER

Data brims with predictive potential. Even when the data about each individual is limited—such as with *metadata*, which characterizes e-mail and telephone communications by their time, date, destination, and the like—there's a lot to work with. These are the nuts and bolts of behavior that are often at least as revealing as, not to mention much easier to process than, communication content, that is, the typed message of an e-mail or spoken words during a phone call.

The experts see the predictive potential. Dean Abbott, a senior hands-on consultant who's applied PA for fraud detection for both the private and public sectors, agrees that ASD is a worthy application of

(*continued*)

[4] However, as a counterexample, note that the pattern (aka microsegment) designed by a UK bank analyst covered earlier in this chapter (from *Super-Freakonomics*) was formed vis-à-vis a target set of known terrorism perpetrators.

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

PA. "Yes, I absolutely think it would be worth the effort to build a predictive model based on metadata that identifies new leads for a given hotlist—especially one that incorporates link data of who's called whom," he said.

A predictive model acts as a choosy, discriminating fishing net. It may include patterns that capture a wide yet precisely defined spectrum of possibilities, arbitrarily abstract and multidimensional. Investigation activities target the individuals who match such patterns; those matches define the now smaller haystack to be searched.

For example, Defense Department-funded university research identified certain circumstances—characterized by the following pattern—that present an 88 percent probability of an attack by the South Asian terrorist organization Lashkar-e-Taiba:

- **PATTERN:** *Between five and 24 of the organization's operatives have been arrested and operatives are on trial in India or Pakistan.*

In a similar vein, such patterns could serve to identify attackers rather than impending attacks. Here is the controversial pattern designed by a UK bank analyst to discover terrorism suspects covered earlier in this chapter (from the book *SuperFreakonomics*—due to its intentional religious discrimination, I consider this example ethically prohibitive; despite that, I include this rare, public, data-driven example to illustrate the mechanics of patterns):

- **PATTERN:** *The individual has opened a certain type of bank account, has placed certain types of wire transfers or other transactions,*

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

has a mobile phone (this example is from the early 2000s), is listed as a student who rents, shows no life insurance policy (suicide would nullify the policy), and holds certain attributes that indicate a likelihood to be Muslim.

To further illustrate the concept, here are three simple example patterns to identify possible suspects that could be generated by PA (fictional, for illustrative purposes only):[5]

- **PATTERN:** *The caller has placed calls from at least two countries per week for eight months, calls from an average of four countries per week, has placed two calls to numbers two degrees of separation from a hotlist of numbers, and received a call from a hotlist number within the last four hours* (such a rule could trigger a real-time alert to analysts).
- **PATTERN:** *The caller shows typical calling patterns (regarding frequency, variance of call durations, and the number of both frequent and infrequent correspondents), but with the addition of calls to more than four never-before-called government phone numbers per week on most weeks, across more than seven countries, for three months.*
- **PATTERN:** *The e-mail address, logged into at a flagged Internet café, is likely a proxy for another e-mail address that has second-degree*

(*continued*)

[5] Patterns like these could be derived by decision trees in combination with specialized data preparation (predictor variables designed for call pattern detection). The adeptness of such patterns improves by combining a larger number of such pattern-matching rules—hundreds rather than only several—as achieved by *ensemble models*.

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

> *ties to a hotlist of e-mail addresses. The proxy pairing is based on the frequency of forwards between the two that are not replied to, the overlap in the sets of correspondents, and similar geolocation login patterns.*

Although a particular pattern may "catch no fish" and come up empty, when a number of even the most arcane patterns are applied across a large population of civilians, there's an opportunity to evenutally find matches. Applied with tactical panache, I believe that iteratively running ASD projects that incorporate human creativity and law enforcement expertise is bound to deliver.

Law enforcement has an unfair advantage. Criminals lack one key resource required to compete against this form of intelligence: the data. Criminal organizations generally cannot recreate law enforcement's surveillance of persons of interest, let alone the much larger dataset of negative examples, the civilians. So they have no means to ascertain the predictive patterns that crime fighters derive from this data, which leaves them with no insight to evade being detected by such patterns. As with *network intrusion detection*, ASD achieves a qualitatively unparalleled advancement in this escalating arms race, the ongoing competition between detection and evasion.

PRESUMPTION: THE NSA USES PREDICTIVE ANALYTICS

It's a foregone conclusion the NSA considers PA a strategic priority. Any use of PA by the NSA is necessarily a secret; the lack of public examples is the nature of the beast. However, wondering whether they use it is like speculating whether a chef who bought flat pasta, meat sauce, mozzarella, and ricotta is making lasagna. Beyond a reasonable doubt, the

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

world's largest spy organization running the country's largest surveillance data center and employing the world's largest number of PhD mathematicians strives to analytically learn from data.

There's much supporting the assumption that the NSA has worked with PA and will continue to do so (see the corresponding section in this book's Notes at www.PredictiveNotes.com for details pertaining to the following summary list):

- NSA documents and official documents about the NSA explicitly indicate established capabilities in machine learning and pattern discovery.
- The NSA has purchased intelligence software solutions that include PA capabilities from two companies, Palantir and Cognito.
- NSA job postings for "data scientists" seek candidates experienced with machine learning and other related technologies.
- The NSA's domestic counterpart conducts ASD: The U.S. Department of Justice released a report describing how the FBI applies PA to assign terrorism "risk scores" to possible suspects.
- PA stands clear as an increasingly common practice for law enforcement of all kinds, including U.S. Armed Forces-funded terrorism prediction, predictive policing, recidivism prediction, and fraud detection, arguably the leading government application of PA.
- Data-driven suspect discovery is a publicly established concept. The popular book *SuperFreakonomics* even covers a specific example of iteratively redefining a pattern to discern terrorism suspects (summarized earlier in this chapter).

(*continued*)

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

THE ARGUMENT FOR COLLECTING THE WHOLE HAYSTACK

Law enforcement is intrinsically destined to apply PA, which serves to discover potential suspects who would otherwise continue undetected. Just as Santa Claus defies scale and visits every single household overnight with lightning speed, this virtual cop sizes up every civilian by matching against scientifically established patterns. And just as companies screen each transaction for fraud and each employee for propensity to quit their job, so too does a government strive to screen each civilian for connection to crime.

Without an understanding of ASD, privacy advocates trip up on fallacies. Wisconsin Rep. James Sensenbrenner, who himself introduced the Patriot Act in the House, argued, "The bigger haystack makes it harder to find the needle." It's a common misconception. Even with regard to the private sector, journalists warn of "drowning" in too much data. But PA practitioners recognize that the data glut is not a problem—it's an opportunity.

Public figures overlook an irony intrinsic to ASD: Wide-scale data collection can serve to identify the few who should be actively surveilled, rather than spy on the many. But some pundits presume the opposite necessarily holds true, in part because ASD is not widely known. Robert Scheer, author of *They Know Everything about You: How Data-Collecting Corporations and Snooping Government Agencies Are Destroying Democracy*, inadvertently invoked ASD when he wrote, "Intelligence should be about learning what you need to know and don't already, not just about sucking up unmanageable gigabytes of minutiae everywhere in the world, which has been the NSA's enormously costly and ineffectual game of choice."

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

A comparable controversy plays out in the field of medicine, where the potential for lifesaving insights also compels open data. Healthcare data-sharing proponent John Wilbanks argues that privacy protections on clinical research data slow down research. "These are tools that we created to protect us from harm, but what they're doing is protecting us from innovation now," he said in a TED talk. "When I tell cancer survivors that this tool we created to protect them is actually preventing their data from being used, . . . their reaction is not, 'Thank you, God, for protecting my privacy.' It's outrage that we have this information and we can't use it."

And so law enforcement by its nature lusts for ever-growing surveillance, just as users of PA for all purposes across all sectors perpetually crave bigger data.

THE COUNTERARGUMENT: CURTAIL MONITORING TO PROTECT CIVIL LIBERTIES

For all its promise, mass government surveillance risks civil liberties and therefore cannot go unrestrained. Those civilians whose data is considered up close by law enforcement personnel, although constituting a minority of cases, are vulnerable to high degrees of potentially unfounded scrutiny and other enforcement activities. With data collection capabilities growing in scope, an agent of the law is armed with more information about the person of interest than he or she may reasonably have known was being tracked. This data then becomes the subject of the particular prejudices of the agent, who considers the demographic profile in combination with perceived aspects of the suspect's private online and telecommunication activities. Over large

(*continued*)

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

numbers of cases, for some this inevitably leads to grave inconveniences, further invasion into their personal life, or even harassment and unjust prosecution.

The presence of this potential infliction upon the few curtails liberties for the many. Glenn Greenwald, author of *No Place to Hide: Edward Snowden, the NSA, and the U.S. Surveillance State* and lead journalist on the 2013 disclosures, wrote that "it is in the realm of privacy where creativity, dissent, and challenges to orthodoxy germinate. A society in which everyone knows they can be watched by the state—where the private realm is effectively eliminated—is one in which those attributes are lost, at both the societal and the individual level. . . . Mass surveillance by the state is therefore inherently repressive."

Brazilian President Dilma Rousseff brought this reasoning to its natural conclusion, following revelations that the NSA had monitored Brazilian citizens and allegations that the intelligence organization had even intercepted official e-mail and telephone communications of the president herself. She declared, "In the absence of the right to privacy, there can be no true freedom of expression and opinion, and therefore no effective democracy."

Besides requiring wide-scale data collection to feed as input, ASD's *outputs*—the predictions it generates—also incur risk to liberty. Potential suspects flagged by ASD face the risk of invasive treatment. Innocent civilians, the inevitable false positives among ASD's targets, could fall subject to unjust scrutiny, drilling down into the data collected about them. When ASD flags an individual, this does not necessarily mean reasonable suspicion has been established by way of specific evidence. However, the personal data previously collected about the individual will not continue to lay dormant—a law

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

enforcement officer will access and leverage it, which may in turn lead to unwarranted acts of search, seizure, or detention.

The ACLU calls this profiling. In a discussion of ASD with Allen Gilbert, the executive director at the American Civil Liberties Union of Vermont, he told me: "Predictive analytics is in essence a form of profiling. It provides an excuse rather than evidence to target someone as a criminal suspect. It short-circuits the Fourth Amendment's protections against search and seizure without reasonable suspicion of crime. A civil libertarian gasps that such pre-judging—prejudice—is considered justified in modern-day crime fighting."

CONCLUSION: A SMARTER DEBATE

Want a productive debate? Then learn more—whichever side you're on. Any simple, sweeping resolution put forth overlooks a great depth of multilayered gray area. Sound bites don't cut it.

We face two extremely challenging tasks:

- To balance the great value aggregated data bears against the danger it holds. The agreed-upon extent of active government surveillance can range across a continuum. At one extreme, at least some minimal level of tracking is broadly accepted without controversy, such as each time we drive through a tollbooth or tunnel, every flight we take, and each time we cross an international border. At the other extreme, there's also general agreement that particularly high levels of monitoring would be too much, for example if the government required a video feed for every room in every building.

(*continued*)

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

- To determine whether and how ASD may safely target law enforcement activities. By design, data-driven investigation more effectively targets and could *decrease* the prevalence of inaccurate human discrimination that accompanies investigations driven by "gut" or "hunch." Is it possible for law enforcement to investigate analytically-derived leads in a prudent manner, or does ASD entail an unacceptably high intrinsic risk for law enforcement abuse?

The position of agreed compromise on these two questions—destined to continuously evolve, by the way—must be set by *more deeply informed debate and negotiation*. Both opposing sides must learn more about the other's concerns:

A. **Data hustlers** who support increased data collection by law enforcement must become deeply familiar with the philosophy, practicalities, and political history that illustrate how compromised privacy brings a loss of liberty and incurs the risk of abuse by law enforcement officers. Furthermore, to inform the risks at hand, law enforcement must render data collection practices and internal data access regulations publicly transparent.
B. **Privacy advocates** who support decreased data collection by law enforcement must come to understand why ASD presents a much stronger incentive for broad-scale data collection than if data were only to serve for investigating individuals: It provides a means to unearth new suspects who might otherwise go undetected, a key capability for the war on terror as well as other law enforcement efforts.

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

Whatever the extent of data collection, as ASD continues to develop it must be carefully managed. I contacted an expert on the ramifications PA holds for *reasonable suspicion*, a legal standard for everyday law enforcement activity. His name is Andrew Ferguson, a law professor of the University of the District of Columbia. He put it this way: "Predictive analytics is clearly the future of law enforcement. The problem is that the forecast for transparency and accountability is less than clear."

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

CHAPTER 3

The Data Effect

A Glut at the End of the Rainbow

We are up to our ears in data, but how much can this raw material really tell us? What actually makes it predictive? What are the most bizarre discoveries from data? When we find an interesting insight, why are we often better off not asking why? In what way is bigger data more dangerous? How do we avoid being fooled by random noise and ensure scientific discoveries are trustworthy?

Spotting the big data tsunami, analytics enthusiasts exclaim, "Surf's up!"

We've entered the golden age of predictive discoveries. A frenzy of number crunching churns out a bonanza of colorful, valuable, and sometimes surprising insights:[1]

- People who "like" curly fries on Facebook are more intelligent.
- Typing with proper capitalization indicates creditworthiness.
- Users of the Chrome and Firefox browsers make better employees.
- Men who skip breakfast are at greater risk for coronary heart disease.
- The demand for Pop-Tarts spikes before a hurricane.
- Female-named hurricanes are more deadly.
- High-crime neighborhoods demand more Uber rides.

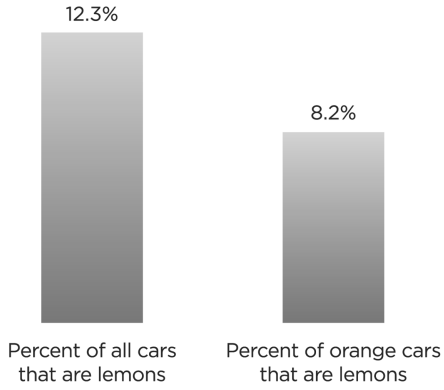
[1] For more details on these findings, see the "Bizarre and Surprising Insights" tables later in this chapter; for the specific citations, see the corresponding Notes (at www.PredictiveNotes.com).

A CAUTIONARY TALE: ORANGE LEMONS

Look like fun? Before you dive in, be warned: This spree of data exploration must be tamed with strict quality control. It's easy to get it wrong and end up with egg on your face.

In 2012, a *Seattle Times* article led with an eye-catching predictive discovery: "An orange used car is least likely to be a lemon."[2] This insight came from a predictive analytics (PA) competition to detect which used cars are bad buys (*lemons*). While insights also emerged pertaining to other car attributes—such as make, model, year, trim level, and size—the apparent advantage of being orange caught the most attention. Responding to quizzical expressions, data wonks offered creative explanations, such as the idea that owners who select an unusual car color tend to have more of a "connection" to and take better care of their vehicle.

Examined alone, the "orange lemon" discovery appeared sound from a mathematical perspective. Here's the specific result:



This shows orange cars turn out to be lemons one third less often than average. Put another way, if you buy a car that's *not* orange, you increase your risk by 50 percent.

Well-established statistics appeared to back up this "colorful" discovery. A formal assessment indicated it was *statistically significant*, meaning that the

[2] This discovery was also featured by *The Huffington Post*, *The New York Times*, *National Public Radio*, *The Wall Street Journal*, and the *New York Times* Bestseller *Big Data: A Revolution That Will Transform How We Live, Work, and Think*.

chances were slim this pattern would have appeared only by random chance. It seemed safe to assume the finding was sound. To be more specific, a standard mathematical test indicated there was less than a 1 percent chance this trend would show up in the data if orange cars weren't actually more reliable.

But something had gone terribly wrong. The "orange car" insight later proved inconclusive. The statistical test had been applied in a flawed manner; the press had ran with the finding prematurely. As data gets bigger, so does a potential pitfall in the application of common, established statistical methods. We'll dive into this dilemma later—but for now here's the issue in a nutshell: *Testing many predictors means taking many small risks of being fooled by randomness, adding up to one big risk.*

This chapter first establishes just how important an opportunity data represents, and then shows how to securely tap it—here's the flow of topics:

The source: where data comes from.

- Why logs of transactions aren't boring
- Why *social data* isn't always an oxymoron
- Estimating the mass mood of the public
- The massive recycling effort that supplies data for PA

The enormousness: how much there is and what the *big* in big data actually means.

The excitement: why data is so predictive—The Data Effect.

The gold rush: what data tells us—46 fascinating discoveries.

Caveat #1: why causality is generally an unknown.

Caveat #2: what went wrong with the "orange lemons" case and how to tap data's potential without drawing false conclusions.

THE SOURCE: OTHERWISE BORING LOGS FUEL PREDICTION

Today's predictive gold mine occurred by happy accident. Most data accumulates not to serve analytics, but as the by-product of routine tasks.

Consider all the phone calls you make. Your wireless provider logs your communications for billing and other transactional purposes. Boring! And

yet these logs also reveal a wellspring of behavioral trends that characterize you and your contacts (and serve law enforcement activities, as discussed in the previous chapter). Companies leverage the predictive power of such consumer behavior to, for example, keep consumers around. By predicting who's going to leave, companies target offers—such as a free phone—in order to retain would-be defectors.

"Social data" may sound like an oxymoron to many, but data about social behavior predicts like nobody's business. Optus, a leading cell phone carrier in Australia, doubled the precision of predicting whether a customer will cancel by incorporating the behavior of each customer's social contacts: If the people you regularly call defect to another wireless provider, there's an up to sevenfold greater risk you will also do so, as more than one telecom has discovered.[3]

Beyond the telecom industry, another immense sector of modern society stockpiles records of person-to-person interactions: social media sites like Facebook, Twitter, and an endless assortment of blogs. Seeing the potential, the financial industry taps these sites to help assess the creditworthiness of would-be debtors, and the Internal Revenue Service taps them to check out taxpayers. City health departments predict restaurant health code violations via Yelp reviews.

In short, what you've posted online may help determine whether your application for a credit card is approved, whether your tax return is audited, and whether a restaurant is inspected.

SOCIAL MEDIA AND MASS PUBLIC MOOD

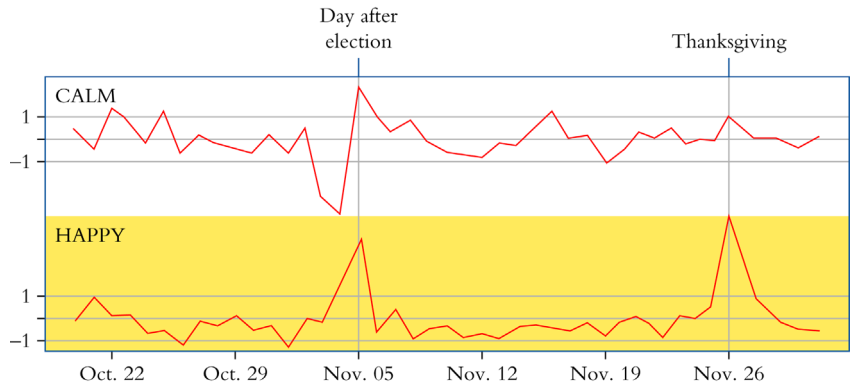
Can a population's overall average mood predict mass behavior? Many bet yes. A trending area of research taps social media posts to gauge the aggregate mood of the public. Researchers evaluate these readings of mass mood for their ability to predict all kinds of population-level behaviors, including the stock market, product sales, top music hits, movie box-office revenue, Academy Award and Grammy winners, elections, and unemployment statistics.

[3] As with the law enforcement examples in Chapter 2's sidebar on automatic suspect discovery, this represents another application area where cellphone *metadata* alone proves to be predictively valuable.

Emotions don't usually fall within the domain of PA. Feelings are not concrete things easily tabulated in a spreadsheet as facts and figures. They're ephemeral and subjective. Sure, they may be the most important element of our human condition, but their subtleties place them outside the reach of most hard science. While a good number of neuroscientists are wiring up the noggins of undergraduate students in exchange for free pizza, many data scientists view this work as irrelevant, far removed from common applications of PA.

But social media blares our emotions. Bloggers, tweeters, and posters broadcast their thoughts, thereby transforming from private, introverted "Dear Diary" writers into vocal extroverts. A mass chorus expresses freely, unfettered by any preordained purpose or restriction. Bloggers alone render an estimated 864,000 posts per day, and in so doing act as an army of volunteers who express sentiment on the public's behalf.

Take a look at how our collective mood moves. Here's sample output of a word-based measure of mood by researchers at Indiana University. Based on a feed from Twitter, it produces daily readings of mass mood for the dimensions *calm* versus *anxious*, and *happy* versus *unhappy* (shown from October 2008 to December 2008):[4]



[4] Johan Bollen, Huina Mao, and Xiao-Jun Zeng, "Twitter Mood Predicts the Stock Market," *Journal of Computational Science*, 2, no. 1 (March 2011). Figure reproduced with permission.

As we oscillate between elation and despair, this jittery movement reveals that we are a moody bunch. The time range shown includes a U.S. presidential election as well as Thanksgiving. Calmness rebounds once the voting of Election Day is complete. Happiness spikes on Thanksgiving.

A tantalizing prospect lingers for black-box trading if the mass mood approach bears fruit predicting the stock market. While there's not yet publicly known proof that it could predict the market well enough to make a killing, optimistic pioneers believe mass mood will become a fundamental component of trading analysis, alongside standard economic gauges. Entrepreneurial quant Randy Saaf said, "We see 'sentiment' as a diversified asset class like foreign markets, bonds, [and] gold."

RECYCLING THE DATA DUMP

One man's trash is another man's treasure.

By leveraging social media in a new way, researchers discover newfound value in oversharing. People tweet whatever the heck suits their fancy. If someone tweets, "I feel awesome today! Just wanted to share," you might assume it interests only the tweeter's friends and family, and there's no value for the rest of the world. As with most applications of PA, though, the data at hand is readily repurposed.

This repurposing signifies a mammoth recycling initiative: the discovery of new value in the data avalanche. Like the millions of chicken feet the United States has realized it can sell to China rather than throw away, our phenomenal accumulation of 1's and 0's surprises us over and over with newfound applications. Calamari was originally considered junk, as was the basis for white chocolate. My mom, Lisa Schamberg, makes photographic art of compost, documenting the beauty inherent in organic waste. Mad scientists want to make use of nuclear waste. I can assure you, data scientists are just as mad.

Growing up watching *Sesame Street*, I got a kick out of the creature Oscar the Grouch, who lives in a garbage can and sings a song about how much he loves trash. It turns out Oscar isn't so crazy after all.

If social media amounts to large-scale, unregulated graffiti, there's a similar phenomenon with the millions of encyclopedias' worth of organizational data scrawled onto magnetic media for miscellaneous operational functions. It's a zillion tons of human refuse that does not smell. What do ScarJo, Iceland, and borscht have in common with data? They're all beautiful things with unwelcoming names.

Most data is not accumulated for the purpose of prediction, but PA can learn from this massive recording of events in the same fashion that you can learn from your accumulation of life experience. As a simple example, take a company's record of your e-mail address and membership status—utilitarian, yet also predictive. During one project, I found that users who signed up with an Earthlink.com (an Internet provider) e-mail address were almost five times more likely to convert from a free trial user level to the premium paid level than those with a Hotmail.com e-mail address. This could be because those who divulged only a temporary e-mail account—which is the intent for some users of free e-mail services like Hotmail—were, on average, less committed to their trial membership. Whatever the reason, this kind of discovery helps a company predict who will be acquired as a paying customer.

THE INSTRUMENTATION OF EVERYTHING WE DO

> *Count what is countable, measure what is measurable, and what is not measurable, make measurable.*
>
> —Galileo

> *Intangibles that appear to be completely intractable can be measured.*
>
> —Douglas Hubbard, *How to Measure Anything*

Some historians assert that we are now experiencing the information revolution, following the agricultural and industrial revolutions. I buy it.

Colin Shearer, a PA leader at IBM, eloquently states that the key to the information revolution is "the instrumentation of everything." More and more, each move you make, online and offline, is recorded, including transactions conducted, websites visited, movies watched, links clicked, friends called, opinions posted, dental procedures endured, sports games won (if you're a professional athlete), traffic cameras passed, flights taken, Wikipedia articles edited, and earthquakes experienced. Countless sensors deploy daily. Mobile devices, robots, and shipping containers record movement, interactions, inventory counts, and radiation levels. Personal health monitors watch your vital signs and exercise routine. The mass migration of online applications from your desktop to the cloud (aka *software as a service*) makes even more of your computer use recordable by organizations.

Free public data is also busting out, so a wealth of knowledge sits at your fingertips. Following the *open data* movement, often embracing a not-for-profit philosophy, many data sets are available online from fields like biodiversity, business, cartography, chemistry, genomics, and medicine. Look at one central index, www.kdnuggets.com/datasets/, and you'll see what amounts to lists of lists of data resources. The Federal Chief Information Officer of the United States launched Data.gov "to increase public access to high value, machine readable datasets generated by . . . the Government." Data.gov sports over 390,000 data sets, including data about marine casualties, pollution, active mines, earthquakes, and commercial flights. Its growth is prescribed: A directive in 2009 obliged all U.S. federal agencies to post at least three "high-value" data sets.

Far afield of government activities, a completely different accumulation of data answers the more forbidden question, "Are you having *fun* yet?" For a dating website, I predicted occurrences of *online flirtation*. After all, as data shows, you're much more likely to be retained as a customer if you get some positive attention. When it comes to recording and predicting human behavior, what's more fundamental than our mating rituals? For this project, actions such as a virtual "wink," a message, or a request to connect as "friends" counted as "flirtatious." Working up a sort of digital tabloid magazine, I produced reports such as the average waiting times before a

flirt is reciprocated, depending on the characteristics of the customer. For example:

Sexual orientation:	Average hours before reciprocal flirt (if any):
Man seeking man	40
Woman seeking man	33
Man seeking woman	43
Woman seeking woman	55

For your entertainment, here's an actual piece of code from a short 175-line computer program called "Flirtback" that I wrote (in the computer language AWK, an oldie but goodie):

```
sex = sexuality[flirt_to]; # sexual orientation
sumbysex[sex] += (delta/(60*60));
nPairsSex[sex]++
```

Come on, you have to admit that's some exciting stuff—enough to keep any computer programmer awake.

Data expresses the bare essence of human behavior. What it doesn't capture is the full dimension and innuendo of human experience—and that's just fine for PA. Because organizations record the aspects of our actions important to their function, one extraordinarily elusive, daunting task has already been completed in the production of raw materials for PA: abstracting the infinite complexity of everyday life and thereby defining which of its endless details are salient.

A new window on the world has opened. Professor Erik Brynjolfsson, an economist at the Massachusetts Institute of Technology, compares this mass instrumentation of human behavior to another historic breakthrough in scientific observation. "The microscope, invented four centuries ago, allowed people to see and measure things as never before—at the cellular level," said *The New York Times*, explaining Brynjolfsson's perspective. "It was a revolution in measurement. Data measurement is the modern

equivalent of the microscope." But rather than viewing things previously too small to see, now we view things previously too big.

BATTEN DOWN THE HATCHES: TMI

There are over 358 million trillion gallons of water on Earth.
—A TV advertisement for Ice Mountain Spring Water

The world now contains more photographs than bricks.
—John Szarkowski, Director of Photography,
Museum of Modern Art (back in 1976)

All this tracking dumps upon us a data glut. Six hundred blog posts are published per minute; by 2011, there were over 100 million blogs across WordPress and Tumblr alone. As for Twitter, "Every day, the world writes the equivalent of a 10-million-page book in Tweets or 8,163 copies of Leo Tolstoy's *War and Peace*," says the official Twitter blog. Stacking that many copies of the book "would reach the height of about 1,470 feet, nearly the ground-to-roof height of Taiwan's Taipei 101, the second tallest building in the world."

YouTube gains an hour of video each second. Estimates put the World Wide Web at over 8.32 billion Web pages. Millions of online retail transactions take place every hour. More photos are taken daily than in the first 100 years of photography, more in two minutes than in all of the 1800s, with 200 million uploaded to Facebook every day. Femto-photography takes a trillion frames per second to capture light in motion and "see around corners." Over 7 billion mobile devices capture usage statistics. More than 100 things per second connect to the Internet, and this rate is increasing; by 2020 the "Internet of Everything" will connect 50 billion things, Cisco projects.

Making all this growth affordable, the cost of data storage is sinking like a rock. The cost per gigabyte on a hard drive has been exponentially decaying since the 1980s, when it approached $1 million. By 2014, it reached 3 cents. We can afford to never delete.[5]

[5] When first released, Google's free e-mail service, Gmail, had no option to delete a message, only to archive it.

Government intelligence aims to archive vast portions of all communication. The U.S. National Security Agency's $2 billion Utah Data Center, a facility five times the size of the U.S. Capitol, is designed to store mammoth archives of human interactions, including complete phone conversations and e-mail messages.

Scientific researchers are uncovering and capturing more and more data, and in so doing revolutionizing their own paradigms. Astronomers are building a new array of radio telescopes that will generate an exabyte of data per day (an exabyte is a quintillion bytes; a byte is a single value, an integer between 0 and 255, often representing a single letter, digit, or punctuation mark). Using satellites, wildlife conservationists track manta rays, considered vulnerable to extinction, as the creatures travel as far as 680 miles in search of food. In biology, as famed futurist Ray Kurzweil portends, given that the price to map a human genome has dropped from $1 billion to a few thousand dollars, information technology will prove to be the domain from which this field's greatest advances emerge.

Overall, data is growing at an incomprehensible speed, an estimated 2.5 quintillion bytes (exabytes) of data per day. A quintillion is a 1 with 18 zeros. In 1986, the data stored by computers, printed on double-sided paper, could have covered the Earth's landmasses; by 2011, it could have done so with two layers of books.

The growth is exponential. Data more than doubles every three years. This brought us to an estimated 8 zettabytes in 2015—that's 8,000,000,000,000,000,000,000 (21 zeros) bytes. Welcome to Big Bang 2.0.

The next logical question is: What's the most valuable thing to do with all this stuff? This book's answer: *Learn from it how to predict.*

WHO'S YOUR DATA?

Good, better, best, bested. How do you like that for a declension, young man?
—Edward Albee, *Who's Afraid of Virginia Woolf?*

Bow your head: The hot buzzword *big data* has ascended to royalty. It's in every news clip, every data science presentation, and every advertisement for analytics solutions. It's a crisis! It's an opportunity! It's a crisis of opportunity!

Big data does not exist. The elephant in the room is that there is no elephant in the room. What's exciting about data isn't how much of it there is, but how quickly it is growing. We're in a persistent state of awe at data's sheer quantity because of one thing that does not change: There's always so much more today than yesterday. Size is relative, not absolute. If we use the word *big* today, we'll quickly run out of adjectives: "big data," "bigger data," "even bigger data," and "biggest data." The International Conference on Very Large Databases has been running since 1975. We have a dearth of vocabulary with which to describe a wealth of data.[6]

"Big data" is also grammatically incorrect. It's like saying "big water." Rather, it should be "a lot of data" or "plenty of data."

What's big about data is the excitement—about its rate of growth and about its predictive value.

THE DATA EFFECT: IT'S PREDICTIVE

The leg bone connected to the knee bone,
and the knee bone connected to the thigh bone,
and the thigh bone connected to the hip bone.

—From the song "Dry Bones"

There's a ton of it—so what? What guarantees that all this residual rubbish, this by-product of organizational functions, holds value? It's no more than an extremely long list of observed events, an obsessive-compulsive enumeration of things that have happened.

The answer is simple. Everything in the world is affected by connections to other things—things touch and cause one another in all sorts of ways—and this is reflected in data. For example:

- Your purchases relate to your shopping history, online behavior, and preferred payment method, and to the actions of your social

[6] Other buzzwords also have their issues. Calling this work *data science* is like calling a librarian a "book librarian." Calling it *data mining* is like calling gold mining "dirt mining."

contacts. Data reveals how to predict consumer behavior from these elements.
- Your health relates to your life choices and environment, and therefore data captures connections predictive of health based on type of neighborhood and household characteristics.
- Your job satisfaction relates to your salary, evaluations, and promotions, and data mirrors this reality.

Data always speaks. It always has a story to tell, and there's always something to learn from it. Data scientists see this over and over again across PA projects. Pull some data together and, although you can never be certain what you'll find, you can be sure you'll discover valuable connections by decoding the language it speaks and listening. That's *The Data Effect* in a nutshell.

The Data Effect: *Data is always predictive.*

This is the assumption behind the leap of faith an organization takes when undertaking PA. Budgeting the staff and tools for a PA project requires this leap, knowing not what specifically will be discovered and yet trusting that something will be. Sitting on an expert panel at Predictive Analytics World, leading UK consultant Tom Khabaza put it this way: "Projects never fail due to lack of patterns." With The Data Effect in mind, the scientist rests easy, secure the analysis will be fruitful.

Data is the new oil. It's this century's greatest possession and often considered an organization's most important strategic asset. Several thought leaders have dubbed it as such—"the new oil"—including European Consumer Commissioner Meglena Kuneva, who also calls it "the new currency of the digital world." It's not hyperbole. In 2012, Apple, Inc. overtook Exxon Mobil Corp., the world's largest oil company, as the most valuable publicly traded company in the world. Unlike oil, data is extremely easy to transport and cheap to store. It's a bigger geyser, and this one is never going to run out.

THE BUILDING BLOCKS: PREDICTORS

Prediction starts small. PA's building block is the *predictor variable*, a single value measured for each individual (known informally as a *factor, attribute, feature,* or *predictor*, and more formally as an *independent variable*). For example, *recency*, the number of weeks since the last time an individual made a purchase, committed a crime, or exhibited a medical symptom, often reveals the chances that individual will do it again in the near term. In many arenas, it makes sense to begin with the most *recently* active people first, whether for marketing contact, criminal investigation, or clinical assessment.

Similarly, *frequency*—the number of times the individual has exhibited the behavior—is also a common, fruitful measure. People who have done something a lot are more likely to do it again.

In fact, it is usually what individuals *have done* that predicts what they *will do*. And so PA feeds on data that extends past dry yet essential demographics like location and gender to include *behavioral predictors* such as recency, frequency, purchases, financial activity, and product usage such as calls and Web surfing. These behaviors are often the most valuable—it's always a *behavior* that we seek to predict, and indeed behavior predicts behavior. As Jean-Paul Sartre put it, "[A man's] true self is dictated by his actions."

PA builds its power by combining dozens—or even hundreds—of predictors. You give the machine everything you know about each individual, and let 'er rip. The core learning technology to combine these elements is where the real scientific magic takes place. That learning process is the topic of the next chapter; for now, let's look at some interesting individual predictors.

FAR OUT, BIZARRE, AND SURPRISING INSIGHTS

Some predictors are more fun to talk about than others.

Are customers more profitable if they don't think? Does crime increase after a sporting event? Does hunger dramatically influence a judge's life-altering decisions? Do online daters more consistently rated as attractive receive *less* interest? Can promotions *increase* the chance you'll quit your job? Do vegetarians miss fewer flights? Does your e-mail address reveal your intentions?

Yes, yes, yes, yes, yes, yes, and yes!

Welcome to the *Ripley's Believe It or Not!* of data science. Poring over a potpourri of prospective predictors, PA's aim isn't only to assess human hunches by testing relationships that seem to make sense, but also to explore a boundless playing field of possible truths beyond the realms of intuition. And so, with The Data Effect in play, PA drops onto your desk connections that seem to defy logic. As strange, mystifying, or unexpected as they may seem, these discoveries help predict.

Here are some colorful discoveries, each pertaining to a single predictor variable (for each example's citation, see the Notes at www.PredictiveNotes.com).

Bizarre and Surprising Insights—Consumer Behavior

Insight	Organization	Suggested Explanation[7]
Guys literally drool over sports cars. Male college student subjects produce measurably more saliva when presented with images of sports cars or money.	Northwestern University Kellogg School of Management	*Consumer impulses are physiological cousins of hunger.*
If you buy diapers, you are more likely to also buy beer. A pharmacy chain found this across 90 days of evening shopping across dozens of outlets (urban myth to some, but based on reported results).	Osco Drug	*Daddy needs a beer.*

(*continued*)

[7] Warning: Do not give much credence to the "Suggested Explanation" column's attempt to answer "why" for each insight. For each one, there are also other plausible explanations, and, in most cases, only intuition rather than scientific evidence behind the particular answer provided. This issue is explored in the next section immediately after these tables of "Bizarre and Surprising Insights."

Bizarre and Surprising Insights—Consumer Behavior (*continued*)

Insight	Organization	Suggested Explanation
Dolls and candy bars. Sixty percent of customers who buy a Barbie doll buy one of three types of candy bars.	Walmart	*Kids come along for errands.*
Pop-Tarts before a hurricane. Prehurricane, Strawberry Pop-Tart sales increased about sevenfold.	Walmart	*In preparation before an act of nature, people stock up on comfort or nonperishable foods.*
Staplers reveal hires. The purchase of a stapler often accompanies the purchase of paper, waste baskets, scissors, paper clips, folders, and so on.	A large retailer	*Stapler purchases are often a part of a complete office kit for a new employee.*
Higher crime, more Uber rides. In San Francisco, the areas with the most prostitution, alcohol, theft, and burglary are most positively correlated with Uber trips.	Uber	*"We hypothesized that crime should be a proxy for nonresidential population. . . . Uber riders are not causing more crime.* Right, guys?*"*
Mac users book more expensive hotels. Orbitz users on an Apple Mac spend up to 30 percent more than Windows users when booking a hotel reservation. Orbitz applies this insight, altering displayed options according to your operating system.	Orbitz	*Macs are often more expensive than Windows computers, so Mac users may on average have greater financial resources.*
Your inclination to buy varies by time of day. For retail websites, the peak is 8:00 PM; for dating, late at night; for finance, around 1:00 PM; for	Survey of websites	*The impetus to complete certain kinds of transactions is higher during certain times of day.*

Bizarre and Surprising Insights—Consumer Behavior (*continued*)

Insight	Organization	Suggested Explanation
travel, just after 10:00 AM. This is not the amount of website traffic, but the propensity to buy of those who are already on the website.		
Your e-mail address reveals your level of commitment. Customers who register for a free account with an Earthlink.com e-mail address are almost five times more likely to convert to a paid, premium-level membership than those with a Hotmail.com e-mail address.	An online dating website	*Disclosing permanent or primary e-mail accounts reveals a longer-term intention.*
Banner ads affect you more than you think. Although you may feel you've learned to ignore them, people who see a merchant's banner ad are 61 percent more likely to subsequently perform a related search, and this drives a 249 percent increase in clicks on the merchant's paid textual ads in the search results.	Yahoo!	*Advertising exerts a subconscious effect.*
Companies win by not prompting customers to think. Contacting actively engaged customers can backfire—direct mailing financial service customers who have already opened several	U.S. Bank	*Customers who have already accumulated many credit accounts are susceptible to impulse buys (e.g., when they walk into a bank branch) but, when contacted at home, will respond by*

(*continued*)

Bizarre and Surprising Insights—Consumer Behavior (*continued*)

Insight	Organization	Suggested Explanation
accounts *decreases* the chances they will open more accounts *(more details in Chapter 7).*		*considering the decision and possibly researching competing products online. They would have been more likely to make the purchase if left to their own devices.*
Your Web browsing reveals your intentions. Wireless customers who check online when their contract period ends are more likely to defect to a competing cell phone company.	A major North American wireless carrier	*Adverse to early termination fees, those intending to switch carriers remind themselves when they'll be free to change over.*
Friends stick to the same cell phone company (a social effect). If you switch wireless carriers, your contacts are in turn up to seven times more likely to follow suit.	A major North American wireless carrier; Optus (Australian telecom) saw a similar effect.	*People experience social influence and/or heed financial incentives for in-network calling.*

Bizarre and Surprising Insights—Finance and Insurance

Insight	Organization	Suggested Explanation
Low credit rating, more car accidents. If your credit score is higher, car insurance companies will lower your premium, since you are a lower driving risk. People with poor credit ratings are charged more for car	Automobile insurers	*"Research indicates that people who manage their personal finances responsibly tend to manage other important aspects of their life with that same level of responsibility, and that would include being responsible behind the wheel of their car," Donald*

Bizarre and Surprising Insights—Finance and Insurance (*continued*)

Insight	Organization	Suggested Explanation
insurance. In fact, a low credit score can increase your premium more than an at-fault car accident; missing two payments can as much as double your premium.		*Hanson of the National Association of Independent Insurers theorizes.*
Your shopping habits foretell your reliability as a debtor. If you use your credit card at a drinking establishment, you're a greater risk to miss credit card payments; at the dentist, lower risk; buy cheap, generic rather than name-brand automotive oil, greater risk; buy felt pads that affix to chair legs to protect the floor, lower risk.	Canadian Tire (a major retail and financial services company)	*More cautionary activity such as seeing the dentist reflects a more conservative or well-planned lifestyle.*
Typing with proper capitalization indicates creditworthiness. Online loan applicants who complete the application form with the correct case are more dependable debtors. Those who complete the form with all lower-case	A financial services startup company	*Adherence to grammatical rules reflects a general propensity to correctly comply.*

(*continued*)

Bizarre and Surprising Insights—Finance and Insurance (*continued*)

Insight	Organization	Suggested Explanation
letters are slightly less reliable payers; all capitals reveals even less reliability.		
Small businesses' credit risk depends on the owner's behavior as a *consumer*. Unlike business loans in general, when it comes to a small business, consumer-level data about the owner is more predictive of credit risk performance than business- level data (and combining both data sources is best of all).	Creditors to the leasing industry	*A small business's behavior largely reflects the choices and habits of one individual: the owner.*

Bizarre and Surprising Insights—Healthcare

Insight	Organization	Suggested Explanation
Genetics foretell cheating wives. Within a certain genetic cluster, having more genes shared by a heterosexual couple means more infidelity by the female.	University of New Mexico	*We're programmed to avoid inbreeding, since there are benefits to genetic diversity.*
Early retirement means earlier death. For a certain working category of males in Austria, each additional year of early	University of Zurich	*Unhealthy habits such as smoking and drinking follow retirement. Voltaire said, "Work spares us from three evils: boredom, vice, and*

Bizarre and Surprising Insights—Healthcare (*continued*)

Insight	Organization	Suggested Explanation
retirement decreases life expectancy by 1.8 months.		*need." Malcolm Forbes said, "Retirement kills more people than hard work ever did."*
Men who skip breakfast get more coronary heart disease. American men 45 to 82 who skip breakfast showed a 27 percent higher risk of coronary heart disease over a 16-year period.	Harvard University medical researchers	*Besides direct health effects—if any—eating breakfast may be a proxy for lifestyle: People who skip breakfast may lead more stressful lives and "were more likely to be smokers, to work full time, to be unmarried, to be less physically active, and to drink more alcohol."*
Google search trends predict disease outbreaks. Certain searches for flu-related information provide insight into current trends in the spread of the influenza virus.	Google Flu Trends	*People with symptoms or in the vicinity of others with symptoms seek further information.*
Smokers suffer less from repetitive motion disorder. In certain work environments, people who smoke cigarettes are less likely to develop carpal tunnel syndrome.	A major metropolitan newspaper, conducting research on its own staff's health	*Smokers take more breaks.*
Positive health habits are contagious (a social effect). If you quit smoking, your close contacts become 36 percent less likely to smoke. Your chance of	Research institutions	*People are strongly influenced by their social environment.*

(*continued*)

Bizarre and Surprising Insights—Healthcare (*continued*)

Insight	Organization	Suggested Explanation
becoming obese increases by 57 percent if you have a friend who becomes obese.		
Happiness is contagious (a social effect). Each additional Facebook friend who is happy increases your chances of being happy by roughly 9 percent.	Harvard University	*"Waves of happiness . . . spread throughout the network."*
Knee surgery choices make a big difference. After ACL-reconstruction knee surgery, walking on knees was rated "difficult or impossible" by twice as many patients who donated their own patellar tissue as a graft source rather than hamstring tissue.	Medical research institutions in Sweden	*The patellar ligament runs across your kneecap, so grafting from it causes injury in that location.*
Music expedites poststroke recovery and improves mood. Stroke patients who listen to music for a couple of hours a day more greatly improve their verbal memory and attention span and improve their mood, as measured by a psychological test.	Cognitive Brain Research Unit, Department of Psychology, University of Helsinki, and Helsinki Brain Research Centre, Finland	*"Music listening activates a widespread bilateral network of brain regions related to attention, semantic processing, memory, motor functions, and emotional processing."*
Yoga improves your mood. Long-term yoga practitioners showed benefits in a psychological test for mood in comparison to nonyoga practitioners, including a higher "vigor" score.	Research institutions in Japan	*Yoga is designed for, and practiced with the intent for, the attainment of tranquility.*

Bizarre and Surprising Insights—Crime and Law Enforcement

Insight	Organization	Suggested Explanation
Suicide bombers do not buy life insurance. An analysis of bank data of suspected terrorists revealed a propensity to not hold a life insurance policy.	A large UK bank	*Suicide nullifies a life insurance policy.*
Unlike lightning, crime strikes twice. Crime is more likely to repeat nearby, spreading like earthquake aftershocks.	Departments of math, computer science, statistics, criminology, and law in California universities	*Perpetrators "repeatedly attack clusters of nearby targets because local vulnerabilities are well-known to the offenders."*
Crime rises with public sporting events. College football upset losses correspond to a 112 percent increase in assaults.	University of Colorado	*Psychological theories of fan aggression are offered.*
Crime rises after elections. In India, crime is lower during an election year and rises soon after elections.	Researchers in India	*Incumbent politicians crack down on crime more forcefully when running for reelection.*
Phone card sales predict danger in the Congo. Impending massacres in the Congo are presaged by spikes in the sale of prepaid phone cards.	CellTel (African telecom)	*Prepaid cards denominated in U.S. dollars serve as in-pocket security against inflation for people "sensing impending chaos."*
Hungry judges rule negatively. Judicial parole decisions	Columbia University and Ben Gurion University (Israel)	*Hunger and/or fatigue leave decision makers feeling less forgiving.*

(*continued*)

Bizarre and Surprising Insights—Crime and Law Enforcement (*continued*)

Insight	Organization	Suggested Explanation
immediately after a food break are about 65 percent favorable, which then drops gradually to almost zero percent before the next break. If the judges are hungry, you are more likely to stay in prison.		

Bizarre and Surprising Insights—Miscellaneous

Insight	Organization	Suggested Explanation
Music taste predicts political affiliation. Kenny Chesney and George Strait fans are most likely conservative, Rihanna and Jay-Z fans liberal. Republicans can be more accurately predicted by music preferences than Democrats because they display slightly less diversity in music taste. Metal fans can go either way, spanning the political spectrum.	The Echo Nest (a music data company)	*Personality types entail certain predilections in both musical and political preferences (this is the author's hypothesis; the researchers do not offer a hypothesis).*
Online dating: Be cool and unreligious to succeed. Online dating messages that initiate first contact and include the word *awesome* are more than twice as likely to elicit a	OkCupid (online dating website)	*There is value in avoiding the overused or trite; video games are not a strong aphrodisiac.*

Bizarre and Surprising Insights—Miscellaneous (*continued*)

Insight	Organization	Suggested Explanation
response as those with *sexy*. Messages with "your pretty" get fewer responses than those with "you're pretty." "Howdy" is better than "Hey." "Band" does better than "literature" and "video games." "Atheist" far surpasses most major religions, but "Zeus" is even better.		
Hot or not? People consistently considered attractive get *less* attention. Online daters rated with a higher variance of attractiveness ratings receive more messages than others with the same average rating but less variance. A greater range of opinions—more disagreement on looks—results in receiving more contact.	OkCupid	*People often feel they don't have a chance with someone who appears universally attractive. When less competition is expected, there is more incentive to initiate contact.*
Users of the Chrome and Firefox browsers make better employees. Among hourly employees engaged in front-line service and sales-based positions, those who use these two custom Web browsers perform better on employment assessment metrics and stay on longer.	A human resources professional services firm, over employee data from Xerox and other firms	*"The fact that you took the time to install [another browser] shows . . . that you are an informed consumer . . . that you care about your productively and made an active choice."*

(*continued*)

Bizarre and Surprising Insights—Miscellaneous (*continued*)

Insight	Organization	Suggested Explanation
A job promotion can lead to quitting. In one division of HP, promotions increase the risk an employee will leave unless accompanied by sufficient increases in compensation; promotions without raises hurt more than help.	Hewlett-Packard	*Increased responsibilities are perceived as burdensome if not financially rewarded.*
More engaged employees have fewer accidents. Among oil refinery workers, a one percentage-point increase in team employee engagement is associated with a 4 percent decrease in the number of safety incidents per employee.	Shell	*More engaged workers are more attentive and focused.*
Higher status, less polite. Editors on Wikipedia who exhibit politeness are more likely to be elected to "administrative" status that grants greater operational authority. However, once elected, an editor's politeness decreases.	Researchers examining Wikipedia behavior	*"Politeness theory predicts a negative correlation between politeness and the power of the requester."*
Vegetarians miss fewer flights. Airline customers who preorder a vegetarian meal are more likely to make their flight.	An airline	*The knowledge of a personalized or specific meal awaiting the customer provides an incentive or establishes a sense of commitment.*
Smart people like curly fries. Liking "Curly Fries" on	Researchers at the University of	*An intelligent person was the first to like this Facebook*

Bizarre and Surprising Insights—Miscellaneous (*continued*)

Insight	Organization	Suggested Explanation
Facebook is predictive of high intelligence.	Cambridge and Microsoft Research	*page, "and his friends saw it, and by homophily, we know that he probably had smart friends, and so it spread to them . . . ," and so on.*
A photo's quality is predictable from its caption. Even without looking at the picture itself, key words from its caption foretell whether a human would subjectively rate the photo as "good." The words *Peru, tombs, trails*, and *boats* corresponded with better photos, whereas the words *graduation* and *CEO* tend to appear with lower-quality photos.	(Not available)	*Certain events and locations are conducive to or provide incentive for capturing more picturesque photos.*
Female-named hurricanes are more deadly. Based on a study of the most damaging hurricanes in the United States during six recent decades, the ones with "relatively feminine" names killed an average of 42 people, almost three times the 15 killed by hurricanes with "relatively male" names.	University researchers	*This may result from "a hazardous form of implicit sexism." Psychological experiments in a related study "suggested that this is because feminine- versus masculine-named hurricanes are perceived as less risky and thus motivate less preparedness. . . . Individuals systematically underestimate their vulnerability to hurricanes with more feminine names."*

(*continued*)

Bizarre and Surprising Insights—Miscellaneous (*continued*)

Insight	Organization	Suggested Explanation
Men on the *Titanic* faced much greater risk than women. A woman on the *Titanic* was almost four times as likely to survive as a man. Most men died and most women lived.	Miscellaneous researchers	*Priority for access to lifeboats was given to women.*
Solo rockers die younger than those in bands. Although all rock stars face higher risk, solo rock stars suffer twice the risk of early death as rock band members.	Public health offices in the UK	*Band members benefit from peer support, and solo artists exhibit even riskier behavior.*

CAVEAT #1: CORRELATION DOES NOT IMPLY CAUSATION

Satisfaction came in the chain reaction.

—From the song "Disco Inferno," by The Trammps

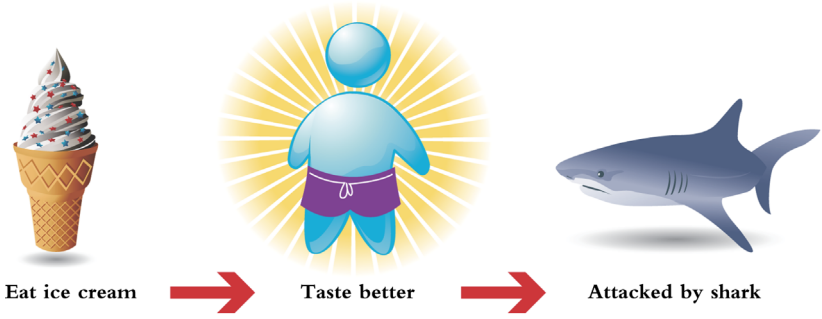
The preceding tables, packed with fun-filled facts, do not explain a single thing.

Take note, the third column is headed "*Suggested* Explanation." While the left column's discoveries are validated by data, the reasons behind them are unknown. Every explanation put forth, each entry in the rightmost column, is pure conjecture with absolutely no hard facts to back it up.

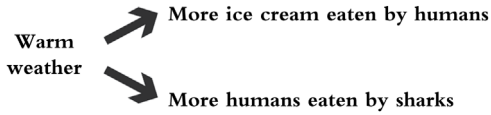
The dilemma is, as it is often said, *correlation does not imply causation*.[8] The discovery of a predictive relationship between A and B does not mean one causes the other, not even indirectly. No way, no how.

[8] The Latin phrase *Post hoc, ergo propter hoc* ("After this, therefore because of this") is another common expression that references the issue at hand; it refers to the unwarranted act of concluding a causal relationship.

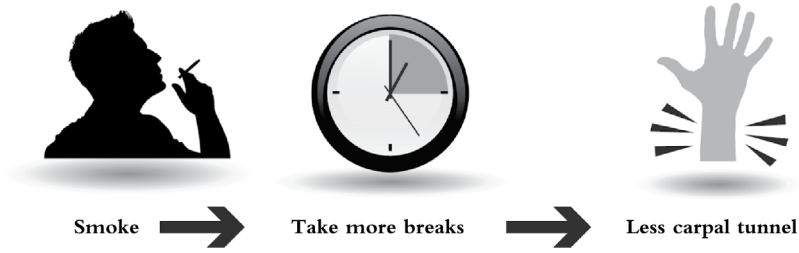
Consider this: Increased ice cream sales correspond with increased shark attacks. Why do you think that is? A causal explanation could be that eating ice cream makes us taste better to sharks:



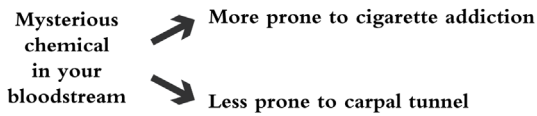
But another explanation is that, rather than one being caused by the other, they are both caused by the same thing. On cold days, people eat less ice cream and also swim less; on warm days, they do the opposite:



Take the example of smokers getting less carpal tunnel syndrome, from the table of healthcare examples. One explanation is that smokers take more breaks:



But another could be that there's some mysterious chemical in your bloodstream that influences both things:



I totally made that up. But the truth is that finding the connection between smoking and carpal tunnel syndrome in and of itself provides no evidence that one explanation is more likely than the other. With this in mind, take another look through the tables. The same rule applies to each example. We know the *what*, but we don't know the *why*.

When applying PA, we generally don't have firm knowledge about causation, and we often don't necessarily care. For many PA projects, the value comes from prediction, with only an avocational interest in understanding the world and figuring out what makes it tick.

Causality is elusive, tough to nail down. We naturally assume things do influence one another in some way, and we conceive of these effects in physical, chemical, medical, financial, or psychological terms. The noble scientists in these fields have their work cut out for them as they work to establish and characterize causal links.

In this way, data scientists have it easier with PA. It just needs to work; prediction trumps explanation. PA operates with extreme solution-oriented intent. The whole point, the "ka-ching" of value, comes in driving decisions from many individual predictions, one per patient, customer, or person of any kind. And while PA often delivers meaningful insights akin to those of various social sciences, this is usually a side effect, not the primary objective.

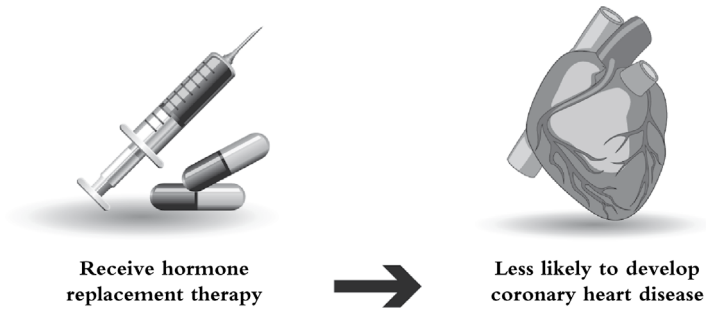
This makes PA a kind of "metascience" that transcends the taxonomy of natural and social sciences, abstracting across them by learning from any and all data sources that would typically serve biology, criminology, economics, education, epidemiology, medicine, political science, psychology, or sociology. PA's mission is to engineer solutions. As for the data employed and the insights gained, the tactic in play is: "Whatever works."

And yet even hard-nosed scientists fight the urge to overexplain. It's human nature, but it's dangerous. It's the difference between good science and bad science.

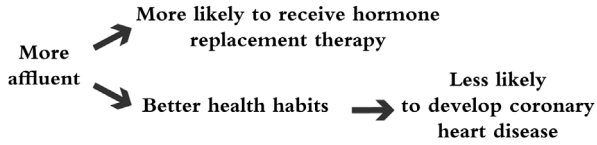
Stein Kretsinger, founding executive of Advertising.com and a director at Elder Research, tells a classic story of our overly interpretive minds. In the early 1990s, as a graduate student, Stein was leading a medical research meeting, assessing the factors that determine how long it takes to wean off a

respirator. As this was before the advent of PowerPoint projection, Stein displayed the factors, one at a time, via graphs on overhead transparencies. The team of healthcare experts nodded their heads, offering one explanation after another for the relationships shown in the data. After going through a few, though, Stein realized he'd been placing the transparencies with the wrong side up, thus projecting mirror images that depicted the *opposite* of the true relationships. After he flipped them to the correct side, the experts seemed just as comfortable as before, offering new explanations for what was now the very opposite effect of each factor. Our thinking is malleable—people readily find underlying theories to explain just about anything.

In another case, a published medical study discovered that women who happened to be receiving hormone replacement therapy showed a lower incidence of coronary heart disease. Could it be that a new treatment for this disease had been discovered?



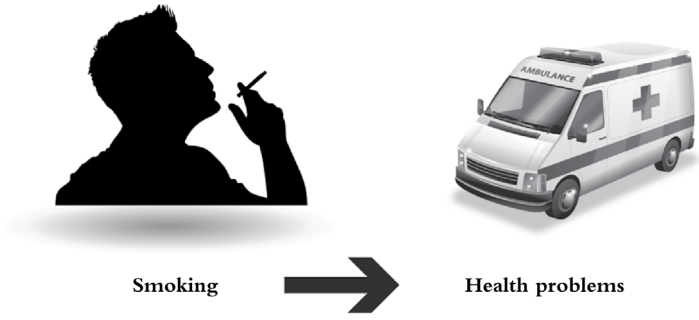
Later, a proper control experiment disproved this false conclusion. Instead, the currently held explanation is that more affluent women had access to the hormone replacement therapy, and these same women had better health habits overall:



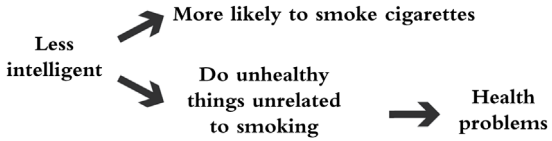
Prematurely jumping to conclusions about causality is bad science that leads to bad medical treatment. This kind of research snafu is not an isolated case.

According to *The Wall Street Journal*, the number of retracted journal publications has surged in recent years.

But, in this arena, the line between apt and inept sometimes blurs. Twenty years ago, while in graduate school, I befriended a colleague, a chain smoker who was nevertheless brilliant with the mathematics behind probability theory. He would hang you out to dry if you attempted to criticize his bad smoking habit on the basis of clinical studies. "Smoking studies have no control group," he'd snap.[9] He was questioning the common causal conclusion:



One day in front of the computer science building, as I kept my distance from his cloud of smoke, he drove this point home. New to the study of probability, I suddenly realized what he was saying and, looking at him incredulously, asked, "You mean to say that it's possible smoking studies actually reflect that stupid people smoke, and that these people also do other stupid things, and only those other things poorly affect their health?" By this logic, I had been stupid for not considering him quite possibly both stupid and healthy.



[9] This is because it's not reasonable to instruct one clinical group to smoke, nor to expect another to uniformly resist smoking. To statistically prove in this way that something kills, you would need to kill some people.

He exhaled a lungful of smoke triumphantly as if he'd won the argument and said with no irony, "Yes!" The same position had also been espoused in the 1950s by an early founder of modern statistics, Ronald Fisher. He was a pipe-smoking curmudgeon who attacked the government-supported publicity about tobacco risks, calling it egregious fearmongering.

In addressing the effects of tobacco, renowned healthcare statistician David Salsburg wrote that the very meaning of cause and effect is "a deep philosophical problem . . . that gnaws at the heart of scientific thought." Due in part to our understanding of how inhaled agents actively lead to genetic mutations that create cancerous cells, the scientific community has concluded that cigarettes are causal in their connection to cancer. While I implore scientists not to overinterpret results, I also implore you not to smoke.

CAVEAT #2: SECURING SOUND DISCOVERIES

> *The trouble with the world is that the stupid are cocksure and the intelligent are full of doubt.*
>
> —Bertrand Russell

Even before suggesting any causal explanation for a correlation observed in data, you had better verify it's actually a real trend rather than misleading noise.

At the beginning of this chapter, we saw that data can lead us astray, tempting us—and several mass media outlets—to believe orange cars last longer. In that data, used cars sporting this flashy color turned out to be lemons 33 percent less often. However, subsequent analysis has severely weakened the confidence in this discovery, relegating it to inconclusive. What went wrong?

Warning! Big data brings big potential—but also big danger. With more data, a unique pitfall often dupes even the brightest of data scientists. This hidden hazard can undermine the process that evaluates for statistical significance, the gold standard of scientific soundness. And what a hazard it is! A bogus discovery can spell disaster. You may buy an orange car—or

undergo an ineffective medical procedure—for no good reason. As the aphorisms tell us, bad information is worse than no information at all; misplaced confidence is seldom found again.

This peril seems paradoxical. If data's so valuable, why should we suffer from obtaining more and more of it? Statistics has long advised that having more examples is better. A longer list of cases provides the means to more scrupulously assess a trend. Can you imagine what the downside of more data might be? As you'll see in a moment, it's a thought-provoking, dramatic plot twist.

The fate of science—and sleeping well at night—depends on deterring the danger. The very notion of empirical discovery is at stake. To leverage the extraordinary opportunity of today's data explosion, we need a surefire way to determine whether an observed trend is real, rather than a random artifact of the data.

Statistics approaches this challenge in a very particular way. It tells us the chances the observed trend could randomly appear even if the effect were not real. That is, it answers this question:[10]

> **Question that statistics can answer:** *If orange cars were actually no more reliable than used cars in general, what would be the probability that this strong a trend—depicting orange cars as more reliable—would show in data anyway, just by random chance?*

With any discovery in data, there's always some possibility we've been *Fooled by Randomness*, as Nassim Taleb titled his compelling book. The

[10] Mini statistics lesson: The notion of the trend being untrue—the notion that orange cars have no advantage—is called the *null hypothesis*. And the probability the observed effect would occur in data if the null hypothesis were true (i.e., the answer to the question above) is called the *p-value*, a foundational concept brought to popularity in the 1920s by the same Ronald Fisher who criticized anti-tobacco propaganda in the 1950s. If the p-value is low enough—e.g., below 1 percent or 5 percent—then a researcher will typically reject the null hypothesis as too unlikely, and view this as support for the discovery, which is thereby considered *statistically significant*. This evaluation process is standard practice for executing on the scientific method itself.

book reveals the dangerous tendency people have to subscribe to unfounded explanations for their own successes and failures, rather than correctly attributing many happenings to sheer randomness. The scientific antidote to this failing is probability, which Taleb affectionately dubs "a branch of applied skepticism."

Statistics is the resource we rely on to gauge probability. It answers the orange car question above by calculating the probability that what's been observed in data would occur randomly if orange cars actually held no advantage. The calculation takes data size into account—in this case, there were 72,983 used cars varying across 15 colors, of which 415 were orange.[11]

Calculated answer to the question: 0.68 percent

Looks like a safe bet. Common practice considers this risk acceptably remote, low enough to at least tentatively believe the data. But don't buy an orange car just yet—or write about the finding in a newspaper for that matter.

WHAT WENT WRONG: ACCUMULATING RISK

In China when you're one in a million, there are 1,300 people just like you.

—Bill Gates

So if there had only been a 1 percent long shot that we'd be misled by randomness, what went wrong?

The experimenters' mistake was to not account for running many small risks, which had added up to one big one. In addition to checking whether being orange is predictive of car reliability, they also checked each of the other 14 colors, as well as the make, model, year, trim level, type of transmission, size, and more. For each of these factors, they repeatedly ran the risk of being fooled by randomness.

[11] The applicable statistical method is a *1-sided equality of proportions hypothesis test*, which produced a p-value under 0.0068. The p-value is the estimated chance we would have ended up with this data if in fact the observed effect were not real; that is, if being colored orange had no correlation with whether the car is a good or bad buy.

Probability is relative, affected entirely by context. With additional background information, a seemingly unlikely event turns out to be not so special after all. Imagine your friend calls to tell you, "I won the jackpot at hundred-to-one odds!" You might get a little excited. "Wow!"

Now imagine your friend adds, "By the way, I'm only talking about one of 70 times that I spun the jackpot wheel." The occurrence that had at first seemed special suddenly has a new context, positioned alongside a number of less remarkable episodes. Instead of exclaiming wow, you might instead do some arithmetic. The probability of losing a spin is 99 percent. If you spin twice, the chances of losing both is 99 percent × 99 percent, which is about 98 percent. Although you'll probably lose both spins, why stop at two? The more times you spin, the lower the chances of never winning once. To figure out the probability of losing 70 times in a row, multiple 99 percent times itself 70 times, aka 0.99 raised to the power of 70. That comes to just under 0.5. Let your friend know that nothing special happened—the odds of winning at least once were about 50/50.

Special cases aren't so special after all. By the same sort of reasoning, we might be skeptical about the merits of the famed and fortuned. Do the most successful elite hold talents as elevated as their singular status? As Taleb put it in *Fooled by Randomness*, "I am not saying that Warren Buffett is not skilled; only that a large population of random investors will almost necessarily produce someone with his track records just by luck."

Play enough and you'll eventually win. Likewise, press your luck repeatedly and you'll eventually lose. Imagine your same well-intentioned friend calls to tell you, "I discovered that orange cars are more reliable, and the stats say there's only a 1 percent chance this phenomenon would appear in the data if it weren't true." You might get a little impressed. "Interesting discovery!"

Now imagine your friend adds, "By the way, I'm only talking about one among dozens of car factors—my computer program systematically went through and checked each one." Both of your friend's stories enthusiastically led with a "remarkable" event—a jackpot win or a predictive discovery. But the numerous other less remarkable attempts—that often go unmentioned—are just as pertinent to each story's conclusion.

Wake up and smell the probability. Imagine we test 70 characteristics of cars that in reality are not predictive of lemons. But each test suffers a, say, 1 percent risk the data will falsely show a predictive effect just by random chance. The accumulated risk piles up. As with the jackpot wheel, there's a 50/50 chance the unlikely event will eventually take place—that you will stumble upon a random perturbation that, considered in isolation, is compelling enough to mislead.

THE POTENTIAL AND DANGER OF AUTOMATING SCIENCE: VAST SEARCH

The most exciting phrase to hear in science, the one that heralds new discoveries, is not "Eureka!" but rather "Hmm . . . that's funny . . ."

—Isaac Asimov

A tremendous potential inspires us to face this peril: Predictive modeling automates scientific discovery. Although it may seem like an obvious thing to do in this computer age, trying out each predictor variable is a dramatic departure from the classic scientific method of developing a single hypothesis and then testing it. Your computer essentially acts as hundreds or even thousands of scientists by conducting a broad, exploratory analysis, automatically evaluating an entire batch of predictors. This aggressive hunt for any novel source of predictive information leaves no stone unturned. The process is key to uncovering valuable, unforeseen insights.

Automating this search for valuable predictors empowers science, lessening its dependence on ever-elusive serendipity. Instead of waiting to inadvertently stumble upon revelations or racking our brains for hypotheses, we rely less on luck and hunches by systematically testing many factors. While necessity is the mother of invention, historically speaking, serendipity has long been its daddy. It was only by happy accident that Alexander Fleming happened upon the potent effects of penicillin, by noticing that an old bacteria culture he was about to clean up happened to be contaminated with some mold—which was successfully killing it. Likewise, Minoxidil was

inadvertently discovered as a baldness remedy in an unexpected, quizzical moment: "Look, more hair!"

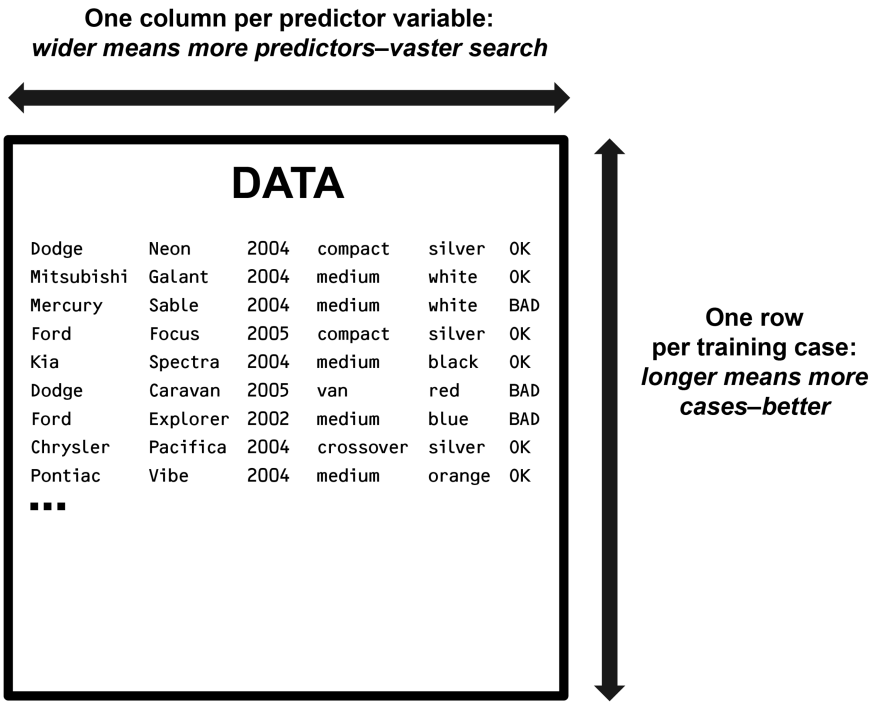
But as exciting a proposition as it is, this automation of data exploration builds up an overall risk of eventually being fooled—at one time or another—by randomness. This inflation of risk comes as a consequence of assessing many characteristics of used cars, for example. The power of automatically testing a batch of predictors may serve us well, but it also exposes us to the very real risk of bogus discoveries.

Let's call this issue *vast search*—the term that industry leader (and Chapter 1's predictive investor) John Elder coined for this form of automated exploration and its associated peril. Repeatedly identified anew across industries and fields of science, this issue is also called the *multiple comparisons problem* or *multiple comparisons trap*. John warns, "The problem is so widespread that it is the chief reason for a crisis in experimental science, where most journal results have been discovered to resist replication; that is, to be wrong!"

Statistics darling Nate Silver jumped straight to the issue of vast search when asked generally about the topic of big data on *Freakonomics Radio*. With a lot of data, he said, "you're going to find lots and lots of correlations through brute force . . . but the problem is that a high percentage of those, maybe the vast majority, are false correlations, are false positives. . . . They [appear] statistically significant, but you have so many lottery tickets when you can run an analysis on a [large data set] that you're going to have some one-in-a-million coincidences just by chance alone."

The casual "mining" of data—analysis of one sort or another to find interesting tidbits and insights—often involves vast search, making it all too easy to dig up a false claim. With this misstep so commonplace, there's a real possibility that some of the predictive discoveries listed in the tables earlier in this chapter could face debunking, depending on whether the researchers have taken proper care. As we'll see in the next chapter, one mischievous professor illustrated the problem of searching and "re-searching" too far and wide when he unearthed a cockamamie relationship between dairy products in Bangladesh and the U.S. stock market.

Bigger data isn't the problem—more specifically, it's *wider* data. When prepared for PA, data grows in two dimensions—it's a table:



A small sample of data for predicting bad buys among used cars. The complete data is both wider and longer.

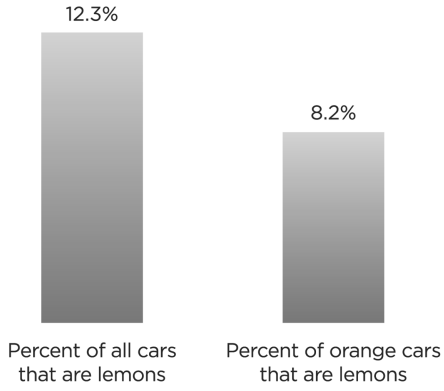
As you accrue more examples of cars, people, or whatever you're predicting, the table grows longer (more rows, aka *training cases*). That's always a good thing. The more training cases to analyze, the more statistically sound.[12] Expanding in the other dimension, each row widens (more columns) as more factors—aka predictor variables—are accrued. A certain factor such as car color may only amount to a single column in the data, but since we look at each possible color individually, it has the virtual effect of adding 15 columns to the width, one per color. Overall, the sample data in the figure above is not nearly as wide as data often gets, but even in this case the vast

[12] This only holds true under the assumption you have a representative sample, e.g., an unbiased, random selection of cases.

search effect is at play. With wider and wider data, we can only tap the potential if we can avoid the booby trap set by vast search.

A FAILSAFE FOR SOUND RESULTS

To understand what sort of failsafe mechanism we need, let's revisit the misleading "orange lemons" discovery.



This 12.3-versus-8.2 result is calculated from four numbers:

There were 72,983 cars, of which 8,976 were lemons.

There were 415 orange cars, of which 34 were lemons.

The standard method—the one that misled researchers as well as the press—evaluates for statistical significance based *only* on those four numbers. When fed these as input, the test provides a positive result, calculating there was only a 0.68 percent chance we would witness that extreme a difference in orange cars if they were in actuality no more prone to be lemons than cars of other colors.

But these four numbers alone do not tell the whole story—the *context* of the discovery also matters. How vast was the search for such discoveries? How many other factors were also checked for a correlation with whether a car is a lemon?

In other words, if a data scientist hands you these four numbers as "proof" of a discovery, you should ask what it took to find it. Inquire, "How many other things did you also try that came up dry?"

With the breadth of search taken into account, the "orange lemon" discovery collapses. Confidence diminishes and it shows as inconclusive. Even if we assume the other 14 colors were the only other factors examined, statistical methods estimate a much less impressive 7.2 percent probability of stumbling by chance alone upon a bogus finding that appears this compelling.[13] Although 7.2 percent is lower odds than a coin toss, it's no long shot; by common standards, this is not a publishable result. Moreover, 7.2 is an optimistic estimate. We can assume the risk was even higher than that (i.e., worse) since other factors such as car make, model, and year were also available, rendering the search even wider and the opportunities to be duped even more plentiful.

Inconclusive results are no results at all. It may still be true that orange cars are less likely to be lemons, but the likelihood this appeared in the data by chance alone is too high to put a lot of faith in it. In other words, there's not enough evidence to rigorously support the hypothesis. It is, at least for now, relegated to "a fascinating possibility," only provisionally distinct from any untested theories one might think up.

Want conclusive results? Then get *longer* data, i.e., more rows of examples. Adequately rigorous failsafe methods that account for the breadth of search set a higher bar. They serve as a more scrupulous filter to eliminate inconclusive findings before they get applied or published. To compensate for this strictness and increase the opportunity to nonetheless attain conclusive results, the best recourse is elongating the list of cases. If the search is vast—that is, if the data is wide—then findings will need to be more compelling in order to pass through the filter. To that end, if there are ample examples with which to confirm findings—in other words, if the data makes up for its width by also being longer—then legitimate findings will have the empirical support they need to be validated.

[13] This probability was estimated with a method called *target shuffling*, which does take the vastness of search into account. For details, see "Are Orange Cars Really not Lemons?" by John Elder and Ben Bullard of Elder Research, Inc. (elderresearch.com/orange-car)

The Data Effect will prevail so long as there are enough training examples to correctly discern which predictive discoveries are authentic.

A PREVALENT MISTAKE

Despite the seriousness of this mistake, the vast search pitfall regularly trips up even the most well-intentioned data scientists, statisticians, and other researchers. A perfect storm of influences leads to its prevalence:

- **It's elusive.** You have to think outside a certain box. The classic application of statistical methods has traditionally focused on evaluating for significance based entirely on the result itself. There's a conceptual leap in moving beyond that to also account for the breadth of search, the full suite of other predictors also considered.
- **It's new.** Since the advent of big data—to be specific, wide data—has more recently intensified this problem, awareness across the data science community still needs to catch up.
- **Simplicity can deceive.** Ironically, although bite-sized anecdotes are more likely to make compelling headlines and draw public attention, they're less likely to be properly screened against failure. It's widely understood that a predictive model, whose job is to combine variables in order to fit the data, can go too far and *overfit*—a primary topic of the next chapter. Since single-variable insights—such as the "orange lemons" claim and the many examples listed earlier in this chapter's tables—are so much simpler than multivariate models, their potential to hold a spurious aberration is underestimated and so they're often subjected to less rigorous scrutiny.
- **Falsehoods don't look wrong.** Without realizing that a pattern may only in actuality be random noise, people creatively formulate compelling causal explanations. This is human nature, but on many occasions it only increases one's attachment to a false discovery.
- **It's a buzzkill.** Given the strong incentives to make predictive discoveries, the temptation is there to be less than scrupulous, either intentionally—

or, more commonly, with a certain convenient forgetfulness—neglecting to account for the full scope of the search that led to the discovery.

In the big data tsunami, you've got to either sharpen your skills or get out of the water.

PUTTING ALL THE PREDICTORS TOGETHER

There ought to be a rock band named after this chapter's explosive topic, "The Predictors."[14]

The number of predictors at our disposal grows along with an unbridled trend: Exploding quantities of increasingly diverse data are springing forth, and organizations are innovating to turn all this unprocessed sap into maple syrup.

The next step is a doozy. To fully leverage predictor variables, we must deftly and intricately combine them with a predictive model. To this end, you can't just stir the bowl with a big spoon. You need an apparatus that learns from the data itself how best to mix and combine it.

Holy combinatorial explosion, Batman! This will make the vast search problem worse—much worse. By combining two predictors, as in, "Are cars with the color black and the make Audi liable to be lemons?" for example—or even more than two—we will build up a much larger batch of relationships to evaluate. This also means a much greater number of opportunities to be fooled by randomness.

Concerned? Overwhelmed? What if I told you there's an intuitive, elegant method for building a predictive model, as well as a simple way to confirm a model's soundness, *without the need to mathematically account for the vastness of search*? The next chapter shows you how it's done—20 pages on how machine learning works, plus another 12 covering the most practical yet philosophically intriguing question of data science: How can we ensure that what the machine has learned is real, that the predictive model is sound?

[14] I spoke too soon; there is one! They're in Australia. See www.thepredictors.com.au. I told you data rocks.

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