3

DATA REIMAGINED

At 6 A.M. on a particular Friday of every month, the streets of most of Manhattan will be largely desolate. The stores lining these streets will be closed, their façades covered by steel security gates, the apartments above dark and silent.

The floors of Goldman Sachs, the global investment banking institution in lower Manhattan, on the other hand, will be brightly lit, its elevators taking thousands of workers to their desks. By 7 A.M. most of these desks will be occupied.

It would not be unfair on any other day to describe this hour in this part of town as sleepy. On this Friday morning, however, there will be a buzz of energy and excitement. On this day, information that will massively impact the stock market is set to arrive.

Minutes after its release, this information will be reported by news sites. Seconds after its release, this information will be discussed, debated, and dissected, loudly, at Goldman and hundreds of other financial firms. But much of the real action in finance these days happens in milliseconds. Goldman and other financial firms paid tens of millions of dollars to get access to fiber-optic cables that reduced the time information travels from Chicago to New Jersey by just four milliseconds (from 17 to 13). Financial firms have algorithms in place to read the information and trade based on it—all in a matter of milliseconds. After this crucial information is released, the market will move in less time than it takes you to blink your eye.

So what is this crucial data that is so valuable to Goldman and numerous other financial institutions?

The monthly unemployment rate.

The rate, however—which has such a profound impact on the stock market that financial institutions have done whatever it takes to maximize the speed with which they receive, analyze, and act upon it—is from a phone survey that the Bureau of Labor Statistics conducts and the information is some three weeks—or 2 billion milliseconds—old by the time it is released.

When firms are spending millions of dollars to chip a millisecond off the flow of information, it might strike you as more than a bit strange that the government takes so long to calculate the unemployment rate.

Indeed, getting these critical numbers out sooner was one of Alan Krueger's primary agendas when he took over as President Obama's chairman of the Council of Economic Advisors in 2011. He was unsuccessful. "Either the BLS doesn't have the resources," he concluded. "Or they are stuck in twentieth-century thinking."

With the government clearly not picking up the pace anytime soon, is there a way to get at least a rough measure of the unemployment statistics at a faster rate? In this high-tech era—when nearly every click any human makes on the internet is recorded somewhere—do we really have to wait weeks to find out how many people are out of work?

One potential solution was inspired by the work of a former Google engineer, Jeremy Ginsberg. Ginsberg noticed that health data, like unemployment data, was released with a delay by the government. The Centers for Disease Control and Prevention takes one week to release influenza data, even though doctors and hospitals would benefit from having the data much sooner.

Ginsberg suspected that people sick with the flu are likely to make flu-related searches. In essence, they would report their symptoms to Google. These searches, he thought, could give a reasonably accurate measure of the current influenza rate. Indeed, searches such as "flu symptoms" and "muscle aches" have proven important indicators of how fast the flu is spreading.[*]

Meanwhile, Google engineers created a service, Google Correlate, that gives outside researchers the means to experiment with the same type of analyses across a wide range of fields, not just health. Researchers can take any data series that they are tracking over time and see what Google searches correlate most with that dataset.

For example, using Google Correlate, Hal Varian, chief economist at Google, and I were able to show which searches most closely track housing prices. When housing prices are rising, Americans tend to search for such phrases as "80/20 mortgage," "new home builder," and "appreciation rate." When housing prices are falling, Americans tend to search for such phrases as "short sale process," "underwater mortgage," and "mortgage forgiveness debt relief."

So can Google searches be used as a litmus test for unemployment in the same way they can for housing prices or influenza? Can we tell, simply by what people are Googling, how many people are unemployed, and can we do so well before the government collates its survey results?

One day, I put the United States unemployment rate from 2004 through 2011 into Google Correlate.

Of the trillions of Google searches during that time, what do you think turned out to be most tightly connected to unemployment? You might imagine "unemployment office"—or something similar. That was high but not at the very top. "New jobs"? Also high but also not at the very top.

The highest during the period I searched—and these terms do shift—was "Slutload." That's right, the most frequent search was for a pornographic site. This may seem strange at first blush, but unemployed people presumably have a lot of time on their hands. Many are stuck at home, alone and bored. Another of the highly correlated searches—this one in the PG realm—is "Spider Solitaire." Again, not surprising for a group of people who presumably have a lot of time on their hands.

Now, I am not arguing, based on this one analysis, that tracking "Slutload" or "Spider Solitaire" is the best way to predict the unemployment rate. The specific diversions that unemployed people use can change over time (at one point, "Rawtube," a different porn site, was among the strongest correlations) and none of these particular terms by itself attracts anything approaching a plurality of the unemployed. But I have generally found that a mix of diversion-related searches can track the unemployment rate—and would be a part of the best model predicting it.

This example illustrates the first power of Big Data, the reimagining of what qualifies as data. Frequently, the value of Big Data is not its size; it's that it can offer you new kinds of information to study—information that had never previously been collected.

Before Google there was information available on certain leisure activities—movie ticket sales, for example—that could yield some clues as to how much time people have on their hands. But the opportunity to know how much solitaire is being played or porn is being watched is new—and powerful. In this instance this data might help us more quickly measure how the economy is doing—at least until the government learns to conduct and collate a survey more quickly.

Life on Google's campus in Mountain View, California, is very different from that in Goldman Sachs's Manhattan headquarters. At 9 A.M. Google's offices

are nearly empty. If any workers are around, it is probably to eat breakfast for free—banana-blueberry pancakes, scrambled egg whites, filtered cucumber water. Some employees might be out of town: at an off-site meeting in Boulder or Las Vegas or perhaps on a free ski trip to Lake Tahoe. Around lunchtime, the sand volleyball courts and grass soccer fields will be filled. The best burrito I've ever eaten was at Google's Mexican restaurant.

How can one of the biggest and most competitive tech companies in the world seemingly be so relaxed and generous? Google harnessed Big Data in a way that no other company ever has to build an automated money stream. The company plays a crucial role in this book since Google searches are by far the dominant source of Big Data. But it is important to remember that Google's success is itself built on the collection of a new kind of data.

If you are old enough to have used the internet in the twentieth century, you might remember the various search engines that existed back then—MetaCrawler, Lycos, AltaVista, to name a few. And you might remember that these search engines were, at best, mildly reliable. Sometimes, if you were lucky, they managed to find what you wanted. Often, they would not. If you typed "Bill Clinton" into the most popular search engines in the late 1990s, the top results included a random site that just proclaimed "Bill Clinton Sucks" or a site that featured a bad Clinton joke. Hardly the most relevant information about the then president of the United States.

In 1998, Google showed up. And its search results were undeniably better those that of every one of its competitors. If you typed "Bill Clinton" into Google in 1998, you were given his website, the White House email address, and the best biographies of the man that existed on the internet. Google seemed to be magic.

What had Google's founders, Sergey Brin and Larry Page, done differently?

Other search engines located for their users the websites that most frequently included the phrase for which they searched. If you were looking for information on "Bill Clinton," those search engines would find, across the entire internet, the websites that had the most references to Bill Clinton. There were many reasons this ranking system was imperfect and one of them was that it was easy to game the system. A joke site with the text "Bill Clinton Bill Clinton Bill Clinton Bill Clinton Bill Clinton" hidden somewhere on its page would score higher than the White House's official website.[*]

What Brin and Page did was find a way to record a new type of information that was far more valuable than a simple count of words. Websites often would, when discussing a subject, link to the sites they thought were most helpful in understanding that subject. For example, the *New York Times*, if it mentioned Bill Clinton, might allow readers who clicked on his name to be sent to the White House's official website.

Every website creating one of these links was, in a sense, giving its opinion of the best information on Bill Clinton. Brin and Page could aggregate all these opinions on every topic. It could crowdsource the opinions of the *New York Times*, millions of Listservs, hundreds of bloggers, and everyone else on the internet. If a whole slew of people thought that the most important link for "Bill Clinton" was his official website, this was probably the website that most people searching for "Bill Clinton" would want to see.

These kinds of links were data that other search engines didn't even consider, and they were incredibly predictive of the most useful information on a given topic. The point here is that Google didn't dominate search merely by collecting more data than everyone else. They did it by finding a *better* type of data. Fewer than two years after its launch, Google, powered by its link analysis, grew to be the internet's most popular search engine. Today, Brin and Page are together worth more than $60 billion.

As with Google, so with everyone else trying to use data to understand the world. The Big Data revolution is less about collecting more and more data. It is about collecting the right data.

But the internet isn't the only place where you can collect new data and where getting the right data can have profoundly disruptive results. This book is largely about how the data on the web can help us better understand people. The next section, however, doesn't have anything to do with web data. In fact, it doesn't have anything to do with people. But it does help illustrate the main point of this chapter: the outsize value of new, unconventional data. And the principles it teaches us are helpful in understanding the digital-based data revolution.

BODIES AS DATA

In the summer of 2013, a reddish-brown horse, of above-average size, with a black mane, sat in a small barn in upstate New York. He was one of 152 one-year-old horses at August's Fasig-Tipton Select Yearling Sale in Saratoga Springs, and one of ten thousand one-year-old horses being auctioned off that year.

Wealthy men and women, when they shell out a lot of money on a racehorse, want the honor of choosing the horse's name. Thus the reddish-brown horse did not yet have a name and, like most horses at the auction, was instead referred to by his barn number, 85.

There was little that made No. 85 stand out at this auction. His pedigree was good but not great. His sire (father), Pioneerof [*sic*] the Nile, was a top racehorse, but other kids of Pioneerof the Nile had not had much racing success. There were also doubts based on how No. 85 looked. He had a scratch on his ankle, for example, which some buyers worried might be evidence of an injury.

The current owner of No. 85 was an Egyptian beer magnate, Ahmed Zayat, who had come to upstate New York looking to sell the horse and buy a few others.

Like almost all owners, Zayat hired a team of experts to help him choose which horses to buy. But his experts were a bit different than those used by nearly every other owner. The typical horse experts you'd see at an event like this were middle-aged men, many from Kentucky or rural Florida with little education but with a family background in the horse business. Zayat's experts, however, came from a small firm called EQB. The head of EQB was not an old-school horse man. The head of EQB, instead, was Jeff Seder, an eccentric, Philadelphia-born man with a pile of degrees from Harvard.

Zayat had worked with EQB before, so the process was familiar. After a few days of evaluating horses, Seder's team would come back to Zayat with five or so horses they recommended buying to replace No. 85.

This time, though, was different. Seder's team came back to Zayat and told him they were unable to fulfill his request. They simply could not recommend that he buy any of the 151 other horses offered up for sale that day. Instead, they offered an unexpected and near-desperate plea. Zayat absolutely, positively could not sell horse No. 85. This horse, EQB declared, was not just the best horse in the auction; he was the best horse of the year and, quite possibly, the decade. "Sell your house," the team implored him. "Do not sell this horse."

The next day, with little fanfare, horse No. 85 was bought for $300,000 by a man calling himself Incardo Bloodstock. Bloodstock, it was later revealed, was a pseudonym used by Ahmed Zayat. In response to the pleas of Seder, Zayat had bought back his own horse, an almost unprecedented action. (The rules of the auction prevented Zayat from simply removing the horse from the auction, thus necessitating the pseudonymous transaction.) Sixty-two horses at the auction sold for a higher price than horse No. 85, with two fetching more than $1 million each.

Three months later, Zayat finally chose a name for No. 85: American Pharoah. And eighteen months later, on a 75-degree Saturday evening in the suburbs of New York City, American Pharoah became the first horse in more than three decades to win the Triple Crown.

What did Jeff Seder know about horse No. 85 that apparently nobody else knew? How did this Harvard man get so good at evaluating horses?

I first met up with Seder, who was then sixty-four, on a scorching June afternoon in Ocala, Florida, more than a year after American Pharoah's Triple Crown. The event was a weeklong showcase for two-year-old horses, culminating in an auction, not dissimilar to the 2013 event where Zayat bought his own horse back.

Seder has a booming, Mel Brooks–like voice, a full head of hair, and a discernable bounce in his step. He was wearing suspenders, khakis, a black shirt with his company's logo on it, and a hearing aid.

Over the next three days, he told me his life story—and how he became so good at predicting horses. It was hardly a direct route. After graduating magna cum laude and Phi Beta Kappa from Harvard, Seder went on to get, also from Harvard, a law degree and a business degree. At age twenty-six, he was working as an analyst for Citigroup in New York City but felt unhappy and burnt-out. One day, sitting in the atrium at the firm's new offices on Lexington Avenue, he found himself studying a large mural of an open field. The painting reminded him of his love of the countryside and his love of horses. He went home and looked at himself in the mirror with his three-piece suit on. He knew then that he was not meant to be a banker and he was not meant to live in New York City. The next morning, he quit his job.

Seder moved to rural Pennsylvania and ambled through a variety of jobs in textiles and sports medicine before devoting his life full-time to his passion: predicting the success of racehorses. The numbers in horse racing are rough. Of the one thousand two-year-old horses showcased at Ocala's auction, one of the nation's most prestigious, perhaps five will end up winning a race with a significant purse. What will happen to the other 995 horses? Roughly one-third will prove too slow. Another one-third will get injured—most because their limbs can't withstand the enormous pressure of galloping at full speed. (Every year, hundreds of horses die on American racetracks, mostly due to broken legs.) And the remaining one-third will have what you might call Bartleby syndrome. Bartleby, the scrivener in Herman Melville's extraordinary short story, stops working and answers every request his employer makes with "I would prefer not to." Many horses, early in their racing careers, apparently come to realize that they don't need to run if they don't feel like it. They may start a race running fast, but, at some point, they'll simply slow down or stop running altogether. Why run around an oval as fast as you can, especially when your hooves and hocks ache? "I would prefer not to," they decide. (I have a soft spot for Bartlebys, horse or human.)

With the odds stacked against them, how can owners pick a profitable horse? Historically, people have believed that the best way to predict whether a horse will succeed has been to analyze his or her pedigree. Being a horse expert means being able to rattle off everything anybody could possibly want to know about a horse's father, mother, grandfathers, grandmothers, brothers, and sisters. Agents announce, for instance, that a big horse "came to her size legitimately" if her mother's line has lots of big horses.

There is one problem, however. While pedigree does matter, it can still only explain a small part of a racing horse's success. Consider the track record of full siblings of all the horses named Horse of the Year, racing's most prestigious annual award. These horses have the best possible pedigrees—the identical family history as world-historical horses. Still, more than three-fourths do not win a major race. The traditional way of predicting horse success, the data tells us, leaves plenty of room for improvement.

It's actually not that surprising that pedigree is not that predictive. Think of humans. Imagine an NBA owner who bought his future team, as ten-year-olds, based on their pedigrees. He would have hired an agent to examine Earvin Johnson III, son of "Magic" Johnson. "He's got nice size, thus far," an agent might say. "It's legitimate size, from the Johnson line. He should have great vision, selflessness, size, and speed. He seems to be outgoing, great personality. Confident walk. Personable. This is a great bet." Unfortunately, fourteen years later, this owner would have a 6'2" (short for a pro ball player) fashion blogger for *E!* Earvin Johnson III might be of great assistance in designing the uniforms, but he would probably offer little help on the court.

Along with the fashion blogger, an NBA owner who chose a team as many owners choose horses would likely snap up Jeffrey and Marcus Jordan, both sons of Michael Jordan, and both of whom proved mediocre college players. Good luck against the Cleveland Cavaliers. They are led by LeBron James, whose mom is 5'5". Or imagine a country that elected its leaders based on their pedigrees. We'd be led by people like George W. Bush. (Sorry, couldn't resist.)

Horse agents do use other information besides pedigree. For example, they analyze the gaits of two-year-olds and examine horses visually. In Ocala, I spent hours chatting with various agents, which was long enough to determine that there was little agreement on what in fact they were looking for.

Add to these rampant contradictions and uncertainties the fact that some horse buyers have what seems like infinite funds, and you get a market with rather large inefficiencies. Ten years ago, Horse No. 153 was a two-year-old who ran faster than every other horse, looked beautiful to most agents, and had a wonderful pedigree—a descendant of Northern Dancer and Secretariat, two of the greatest racehorses of all time. An Irish billionaire and a Dubai sheik both wanted to purchase him. They got into a bidding war that quickly turned into a contest of pride. As hundreds of stunned horse men and women looked on, the bids kept getting higher and higher, until the two-year-old horse finally sold for $16 million, by far the highest price ever paid for a horse. Horse No. 153, who was given the name The Green Monkey, ran three races, earned just $10,000, and was retired.

Seder never had any interest in the traditional methods of evaluating horses. He was interested only in data. He planned to measure various attributes of racehorses and see which of them correlated with their performance. It's important to note that Seder worked out his plan half a decade before the World Wide Web was invented. But his strategy was very much based on data science. And the lessons from his story are applicable to anybody using Big Data.

For years, Seder's pursuit produced nothing but frustration. He measured the size of horses' nostrils, creating the world's first and largest dataset on horse nostril size and eventual earnings. Nostril size, he found, did not predict horse success. He gave horses EKGs to examine their hearts and cut the limbs off dead horses to measure the volume of their fast-twitch muscles. He once grabbed a shovel outside a barn to determine the size of horses' excrement, on the theory that shedding too much weight before an event can slow a horse down. None of this correlated with racing success.

Then, twelve years ago, he got his first big break. Seder decided to measure the size of the horses' internal organs. Since this was impossible with existing technology, he constructed his own portable ultrasound. The results were remarkable. He found that the size of the heart, and particularly the size of the left ventricle, was a massive predictor of a horse's success, the single most important variable. Another organ that mattered was the spleen: horses with small spleens earned virtually nothing.

Seder had a couple more hits. He digitized thousands of videos of horses galloping and found that certain gaits did correlate with racetrack success. He also discovered that some two-year-old horses wheeze after running one-eighth of a mile. Such horses sometimes sell for as much as a million dollars, but Seder's data told him that the wheezers virtually never pan out. He thus assigns an assistant to sit near the finish line and weed out the wheezers.

Of about a thousand horses at the Ocala auction, roughly ten will pass all of Seder's tests. He ignores pedigree entirely, except as it will influence the price a horse will sell for. "Pedigree tells us a horse might have a very small chance of being great," he says. "But if I can see he's great, what do I care how he got there?"

One night, Seder invited me to his room at the Hilton hotel in Ocala. In the room, he told me about his childhood, his family, and his career. He showed me pictures of his wife, daughter, and son. He told me he was one of three Jewish students in his Philadelphia high school, and that when he entered he was 4'10". (He grew in college to 5'9".) He told me about his favorite horse: Pinky Pizwaanski. Seder bought and named this horse after a gay rider. He felt that Pinky, the horse, always gave a great effort even if he wasn't the most successful.

Finally, he showed me the file that included all the data he had recorded on No. 85, the file that drove the biggest prediction of his career. Was he giving away his secret? Perhaps, but he said he didn't care. More important to him than protecting his secrets was being proven right, showing to the world that these twenty years of cracking limbs, shoveling poop, and jerry-rigging ultrasounds had been worth it.

Here's some of the data on horse No. 85:

NO. 85 (LATER AMERICAN PHAROAH) PERCENTILES AS A ONE-YEAR-OLD

	PERCENTILE

Height	56
Weight	61
Pedigree	70
Left Ventricle	*99.61*

There it was, stark and clear, the reason that Seder and his team had become so obsessed with No. 85. His left ventricle was in the 99.61st percentile!

Not only that, but all his other important organs, including the rest of his heart and spleen, were exceptionally large as well. Generally speaking, when it comes to racing, Seder had found, the bigger the left ventricle, the better. But a left ventricle as big as this can be a sign of illness if the other organs are tiny. In American Pharoah, all the key organs were bigger than average, and the left ventricle was enormous. The data screamed that No. 85 was a 1-in-100,000 or even a one-in-a-million horse.

What can data scientists learn from Seder's project?

First, and perhaps most important, if you are going to try to use new data to revolutionize a field, it is best to go into a field where old methods are lousy. The pedigree-obsessed horse agents whom Seder beat left plenty of room for improvement. So did the word-count-obsessed search engines that Google beat.

One weakness of Google's attempt to predict influenza using search data is that you can already predict influenza very well just using last week's data and a simple seasonal adjustment. There is still debate about how much search data adds to that simple, powerful model. In my opinion, Google searches have more promise measuring health conditions for which existing data is weaker and therefore something like Google STD may prove more valuable in the long haul than Google Flu.

The second lesson is that, when trying to make predictions, you needn't worry too much about why your models work. Seder could not fully explain to me why the left ventricle is so important in predicting a horse's success. Nor could he precisely account for the value of the spleen. Perhaps one day horse cardiologists and hematologists will solve these mysteries. But for now it doesn't matter. Seder is in the prediction business, not the explanation business. And, in the prediction business, you just need to know that something works, not why.

For example, Walmart uses data from sales in all their stores to know what products to shelve. Before Hurricane Frances, a destructive storm that hit the Southeast in 2004, Walmart suspected—correctly—that people's shopping habits may change when a city is about to be pummeled by a storm. They pored through sales data from previous hurricanes to see what people might want to buy. A major answer? Strawberry Pop-Tarts. This product sells seven times faster than normal in the days leading up to a hurricane.

Based on their analysis, Walmart had trucks loaded with strawberry Pop-Tarts heading down Interstate 95 toward stores in the path of the hurricane. And indeed, these Pop-Tarts sold well.

Why Pop-Tarts? Probably because they don't require refrigeration or cooking. Why strawberry? No clue. But when hurricanes hit, people turn to strawberry Pop-Tarts apparently. So in the days before a hurricane, Walmart now regularly stocks its shelves with boxes upon boxes of strawberry Pop-Tarts. The reason for the relationship doesn't matter. But the relationship itself does. Maybe one day food scientists will figure out the association between hurricanes and toaster pastries filled with strawberry jam. But, while waiting for some such explanation, Walmart still needs to stock its shelves with strawberry Pop-Tarts when hurricanes are approaching and save the Rice Krispies treats for sunnier days.

This lesson is also clear in the story of Orley Ashenfelter. What Seder is to horses, Ashenfelter, an economist at Princeton, may be to wine.

A little over a decade ago, Ashenfelter was frustrated. He had been buying a lot of red wine from the Bordeaux region of France. Sometimes this wine was delicious, worthy of its high price. Many times, though, it was a letdown.

Why, Ashenfelter wondered, was he paying the same price for wine that turned out so differently?

One day, Ashenfelter received a tip from a journalist friend and wine connoisseur. There was indeed a way to figure out whether a wine would be good. The key, Ashenfelter's friend told him, was the weather during the growing season.

Ashenfelter's interest was piqued. He went on a quest to figure out if this was true and he could consistently purchase better wine. He downloaded thirty years of weather data on the Bordeaux region. He also collected auction prices of wines. The auctions, which occur many years after the wine was originally sold, would tell you how the wine turned out.

The result was amazing. A huge percentage of the quality of a wine could be explained simply by the weather during the growing season.

In fact, a wine's quality could be broken down to one simple formula, which we might call the First Law of Viticulture:

Price = 12.145 + 0.00117 winter rainfall + 0.0614 average growing season temperature – 0.00386 harvest rainfall.

So why does wine quality in the Bordeaux region work like this? What explains the First Law of Viticulture? There is some explanation for Ashenfelter's wine formula—heat and early irrigation are necessary for grapes to properly ripen.

But the precise details of his predictive formula go well beyond any theory and will likely never be fully understood even by experts in the field.

Why does a centimeter of winter rain add, on average, exactly 0.1 cents to the price of a fully matured bottle of red wine? Why not 0.2 cents? Why not 0.05? Nobody can answer these questions. But if there are 1,000 centimeters of additional rain in a winter, you should be willing to pay an additional $1 for a bottle of wine.

Indeed, Ashenfelter, despite not knowing exactly why his regression worked exactly as it did, used it to purchase wines. According to him, "It worked out great." The quality of the wines he drank noticeably improved.

If your goal is to predict the future—what wine will taste good, what products will sell, which horses will run fast—you do not need to worry too much about why your model works exactly as it does. Just get the numbers right. That is the second lesson of Jeff Seder's horse story.

The final lesson to be learned from Seder's successful attempt to predict a potential Triple Crown winner is that you have to be open and flexible in determining what counts as data. It is not as if the old-time horse agents were oblivious to data before Seder came along. They scrutinized race times and pedigree charts. Seder's genius was to look for data where others hadn't looked before, to consider nontraditional sources of data. For a data scientist, a fresh and original perspective can pay off.

WORDS AS DATA

One day in 2004, two young economists with an expertise in media, then Ph.D. students at Harvard, were reading about a recent court decision in Massachusetts legalizing gay marriage.

The economists, Matt Gentzkow and Jesse Shapiro, noticed something interesting: two newspapers employed strikingly different language to report the

same story. The *Washington Times,* which has a reputation for being conservative, headlined the story: “Homosexuals ‘Marry’ in Massachusetts.” The *Washington Post,* which has a reputation for being liberal, reported that there had been a victory for “same-sex couples.”

It’s no surprise that different news organizations can tilt in different directions, that newspapers can cover the same story with a different focus. For years, in fact, Gentzkow and Shapiro had been pondering if they might use their economics training to help understand media bias. Why do some news organizations seem to take a more liberal view and others a more conservative one?

But Gentzkow and Shapiro didn’t really have any ideas on how they might tackle this question; they couldn’t figure out how they could systematically and objectively measure media subjectivity.

What Gentzkow and Shapiro found interesting, then, about the gay marriage story was not that news organizations differed in their coverage; it was *how* the newspapers’ coverage differed—it came down to a distinct shift in word choice. In 2004, “homosexuals,” as used by the *Washington Times,* was an old-fashioned and disparaging way to describe gay people, whereas “same-sex couples,” as used by the *Washington Post,* emphasized that gay relationships were just another form of romance.

The scholars wondered whether language might be the key to understanding bias. Did liberals and conservatives consistently use different phrases? Could the words that newspapers use in stories be turned into data? What might this reveal about the American press? Could we figure out whether the press was liberal or conservative? And could we figure out why? In 2004, these weren’t idle questions. The billions of words in American newspapers were no longer trapped on newsprint or microfilm. Certain websites now recorded every word included in every story for nearly every newspaper in the United States. Gentzkow and Shapiro could scrape these sites and quickly test the extent to which language could measure newspaper bias. And, by doing this, they could sharpen our understanding of how the news media works.

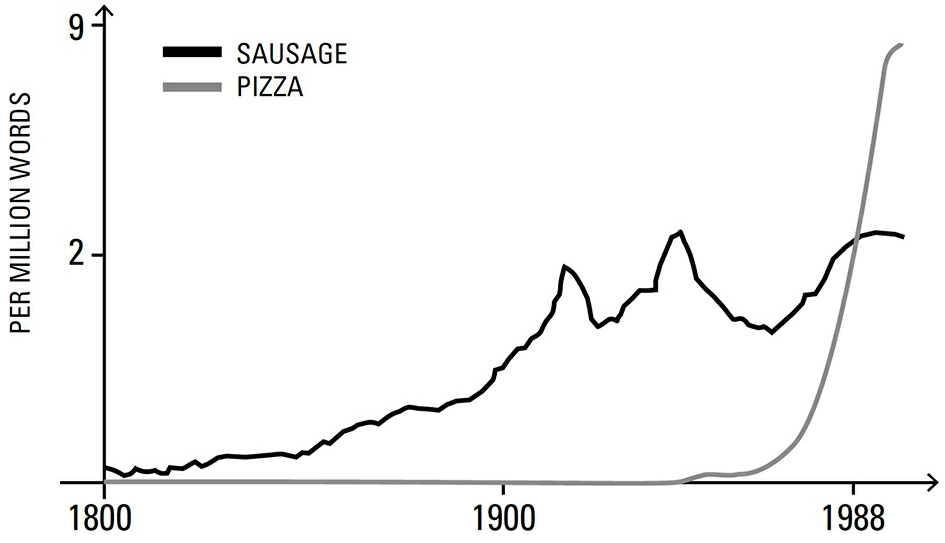
But, before describing what they found, let’s leave for a moment the story of Gentzkow and Shapiro and their attempt to quantify the language in newspapers, and discuss how scholars, across a wide range of fields, have utilized this new type of data—words—to better understand human nature.

Language has, of course, always been a topic of interest to social scientists. However, studying language generally required the close reading of texts, and turning huge swaths of text into data wasn’t feasible. Now, with computers and digitization, tabulating words across massive sets of documents is easy. Language has thus become subject to Big Data analysis. The links that Google utilized were composed of words. So are the Google searches that I study. Words feature frequently in this book. But language is so important to the Big Data revolution, it deserves its own section. In fact, it is being used so much now that there is an entire field devoted to it: “text as data.”

A major development in this field is Google Ngrams. A few years ago, two young biologists, Erez Aiden and Jean-Baptiste Michel, had their research assistants counting words one by one in old, dusty texts to try to find new insights on how certain usages of words spread. One day, Aiden and Michel heard about a new project by Google to digitize a large portion of the world’s books. Almost immediately, the biologists grasped that this would be a much easier way to understand the history of language.

“We realized our methods were so hopelessly obsolete,” Aiden told *Discover* magazine. “It was clear that you couldn’t compete with this juggernaut of digitization.” So they decided to collaborate with the search company. With the help of Google engineers, they created a service that searches through the millions of digitized books for a particular word or phrase. It then will tell researchers how frequently that word or phrase appeared in every year, from 1800 to 2010.

So what can we learn from the frequency with which words or phrases appear in books in different years? For one thing, we learn about the slow growth in popularity of sausage and the relatively recent and rapid growth in popularity of pizza.



But there are lessons far more profound than that. For instance, Google Ngrams can teach us how national identity formed. One fascinating example is presented in Aiden and Michel’s book, *Uncharted.*

First, a quick question. Do you think the United States is currently a united or a divided country? If you are like most people, you would say the United States is divided these days due to the high level of political polarization. You might even say the country is about as divided as it has ever been. America, after all, is now color-coded: red states are Republican; blue states are Democratic. But, in *Uncharted,* Aiden and Michel note one fascinating data point that reveals just how much more divided the United States once was. The data point is the language people use to talk about the country.

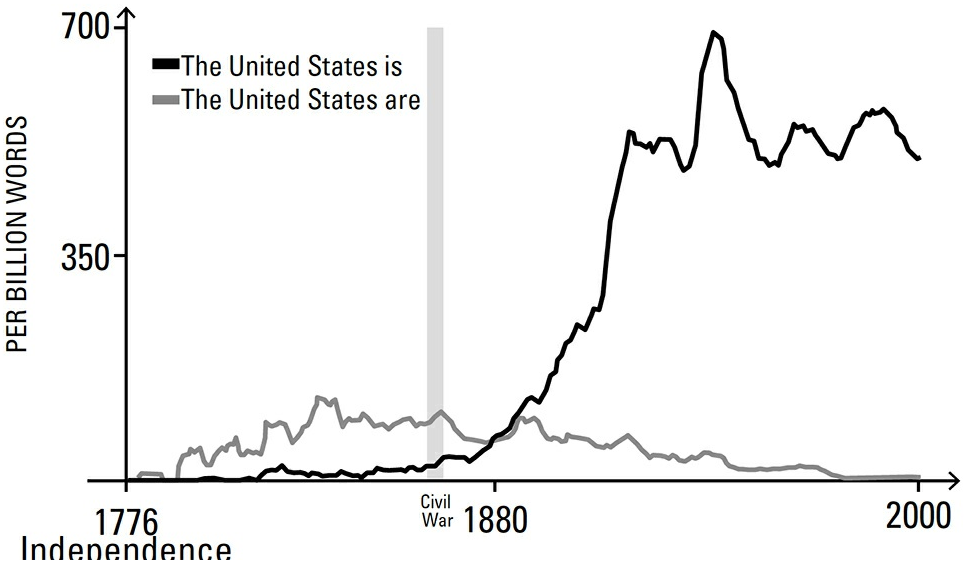
Note the words I used in the previous paragraph when I discussed how divided the country is. I wrote, “The United States *is* divided.” I referred to the United States as a singular noun. This is natural; it is proper grammar and standard usage. I am sure you didn’t even notice.

However, Americans didn’t always speak this way. In the early days of the country, Americans referred to the United States using the plural form. For example, John Adams, in his 1799 State of the Union address, referred to “the United States in *their* treaties with his Britanic Majesty.” If my book were

written in 1800, I would have said, “The United States *are* divided.” This little usage difference has long been a fascination for historians, since it suggests there was a point when America stopped thinking of itself as a collection of states and started thinking of itself as one nation.

So when did this happen? Historians, *Uncharted* informs us, have never been sure, as there has been no systematic way to test it. But many have long suspected the cause was the Civil War. In fact, James McPherson, former president of the American Historical Association and a Pulitzer Prize winner, noted bluntly: “The war marked a transition of the United States to a singular noun.”

But it turns out McPherson was wrong. Google Ngrams gave Aiden and Michel a systematic way to check this. They could see how frequently American books used the phrase “The United States are . . .” versus “The United States is . . .” for every year in the country’s history. The transformation was more gradual and didn’t accelerate until well after the Civil War ended.



Fifteen years after the Civil War, there were still more uses of “The United States are . . .” than “The United States is . . . ,” showing the country was still divided linguistically. Military victories happen quicker than changes in mindsets.

So much for how a country unites. How do a man and woman unite? Words can help here, too.

For example, we can predict whether a man and woman will go on a second date based on how they speak on the first date.

This was shown by an interdisciplinary team of Stanford and Northwestern scientists: Daniel McFarland, Dan Jurafsky, and Craig Rawlings. They studied hundreds of heterosexual speed daters and tried to determine what predicts whether they will feel a connection and want a second date.

They first used traditional data. They asked daters for their height, weight, and hobbies and tested how these factors correlated with someone reporting a spark of romantic interest. Women, on average, prefer men who are taller and share their hobbies; men, on average, prefer women who are skinnier and share their hobbies. Nothing new there.

But the scientists also collected a new type of data. They instructed the daters to take tape recorders with them. The recordings of the dates were then digitized. The scientists were thus able to code the words used, the presence of laughter, and the tone of voice. They could test both how men and women signaled they were interested and how partners earned that interest.

So what did the linguistic data tell us? First, how a man or woman conveys that he or she is interested. One of the ways a man signals that he is attracted is obvious: he laughs at a woman’s jokes. Another is less obvious: when speaking, he limits the range of his pitch. There is research that suggests a monotone voice is often seen by women as masculine, which implies that men, perhaps subconsciously, exaggerate their masculinity when they like a woman.

The scientists found that a woman signals her interest by varying her pitch, speaking more softly, and taking shorter turns talking. There are also major clues about a woman’s interest based on the particular words she uses. A woman is unlikely to be interested when she uses hedge words and phrases such as “probably” or “I guess.”

Fellas, if a woman is hedging her statements on any topic—if she “sorta” likes her drink or “kinda” feels chilly or “probably” will have another hors d’oeuvre—you can bet that she is “sorta” “kinda” “probably” not into you.

A woman *is* likely to be interested when she talks about herself. It turns out that, for a man looking to connect, the most beautiful word you can hear from a woman’s mouth may be “I”: it’s a sign she is feeling comfortable. A woman also is likely to be interested if she uses self-marking phrases such as “Ya know?” and “I mean.” Why? The scientists noted that these phrases invite the listener’s attention. They are friendly and warm and suggest a person is looking to connect, ya know what I mean?

Now, how can men and women communicate in order to get a date interested in them? The data tells us that there are plenty of ways a man can talk to raise the chances a woman likes him. Women like men who follow their lead. Perhaps not surprisingly, a woman is more likely to report a connection if a man laughs at her jokes and keeps the conversation on topics she introduces rather than constantly changing the subject to those he wants to talk about.* Women also like men who express support and sympathy. If a man says, “That’s awesome!” or “That’s really cool,” a woman is significantly more likely to report a connection. Likewise if he uses phrases such as “That’s tough” or “You must be sad.”

For women, there is some bad news here, as the data seems to confirm a distasteful truth about men. Conversation plays only a small role in how they respond to women. Physical appearance trumps all else in predicting whether a man reports a connection. That said, there is one word that a woman can use to at least slightly improve the odds a man likes her and it’s one we’ve already discussed: “I.” Men are more likely to report clicking with a woman who talks about herself. And as previously noted, a woman is also more likely to report a connection after a date where she talks about herself. Thus it is a great sign, on a first date, if there is substantial discussion about the woman. The woman signals her comfort and probably appreciates that the man is not hogging the conversation. And the man likes that the woman is opening up. A second date is likely.

Finally, there is one clear indicator of trouble in a date transcript: a question mark. If there are lots of questions asked on a date, it is less likely that both the man and the woman will report a connection. This seems counterintuitive; you might think that questions are a sign of interest. But not so on a first date. On a first date, most questions are signs of boredom. “What are your hobbies?” “How many brothers and sisters do you have?” These are the kinds of things people say when the conversation stalls. A great first date may include a single question at the end: “Will you go out with me again?” If this is the only question on the date, the answer is likely to be “Yes.”

And men and women don’t just talk differently when they’re trying to woo each other. They talk differently in general.

A team of psychologists analyzed the words used in hundreds of thousands of Facebook posts. They measured how frequently every word is used by men and women. They could then declare which are the most masculine and most feminine words in the English language.

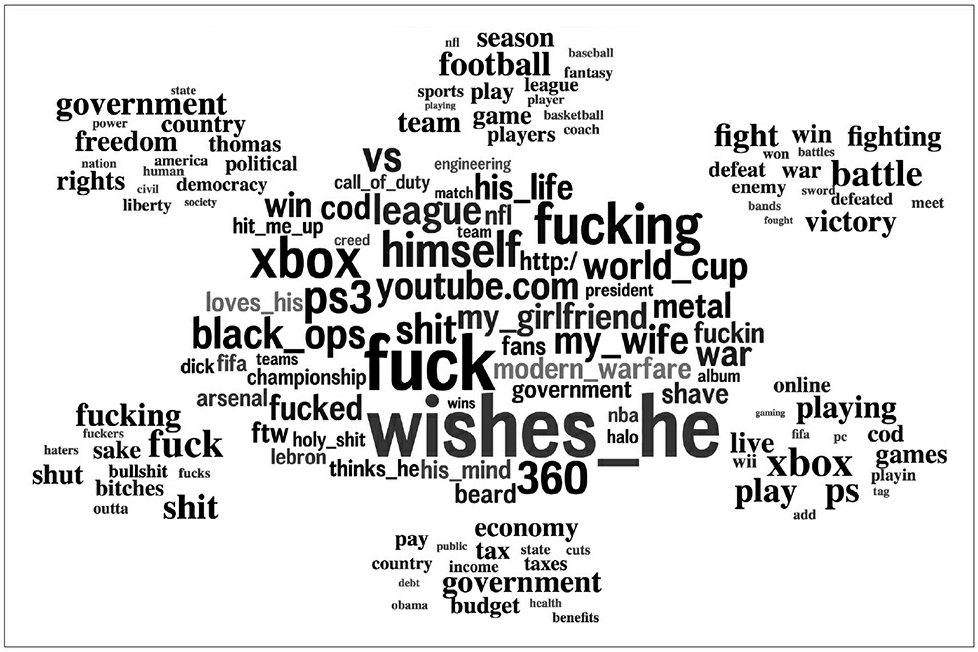
Many of these word preferences, alas, were obvious. For example, women talk about “shopping” and “my hair” much more frequently than men do. Men talk about “football” and “Xbox” much more frequently than women do. You probably didn’t need a team of psychologists analyzing Big Data to tell you that.

Some of the findings, however, were more interesting. Women use the word “tomorrow” far more often than men do, perhaps because men aren’t so great at thinking ahead. Adding the letter “o” to the word “so” is one of the most feminine linguistic traits. Among the words most disproportionately used by women are “soo,” “sooo,” “soooo,” “sooooo,” and “soooooo.”

Maybe it was my childhood exposure to women who weren’t afraid to throw the occasional f-bomb. But I always thought cursing was an equal-opportunity trait. Not so. Among the words used much more frequently by men than women are “fuck,” “shit,” “fucks,” “bullshit,” “fucking,” and “fuckers.”

Here are word clouds showing words used mostly by men and those used mostly by women. The larger a word appears, the more that word’s use tilts toward that gender.

Males



Females

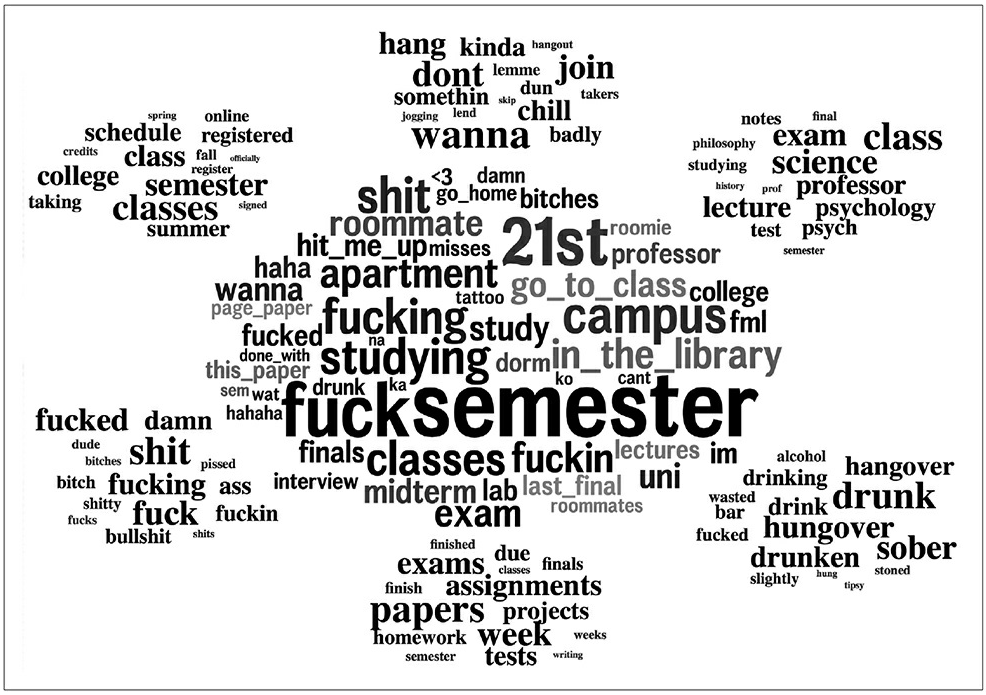
What I like about this study is the new data informs us of patterns that have long existed but we hadn't necessarily been aware of. Men and women have always spoken in different ways. But, for tens of thousands of years, this data disappeared as soon as the sound waves faded in space. Now this data is preserved on computers and can be analyzed by computers.

Or perhaps what I should have said, given my gender: "The words used to fucking disappear. Now we can take a break from watching football and playing Xbox and learn this shit. That is, if anyone gives a fuck."

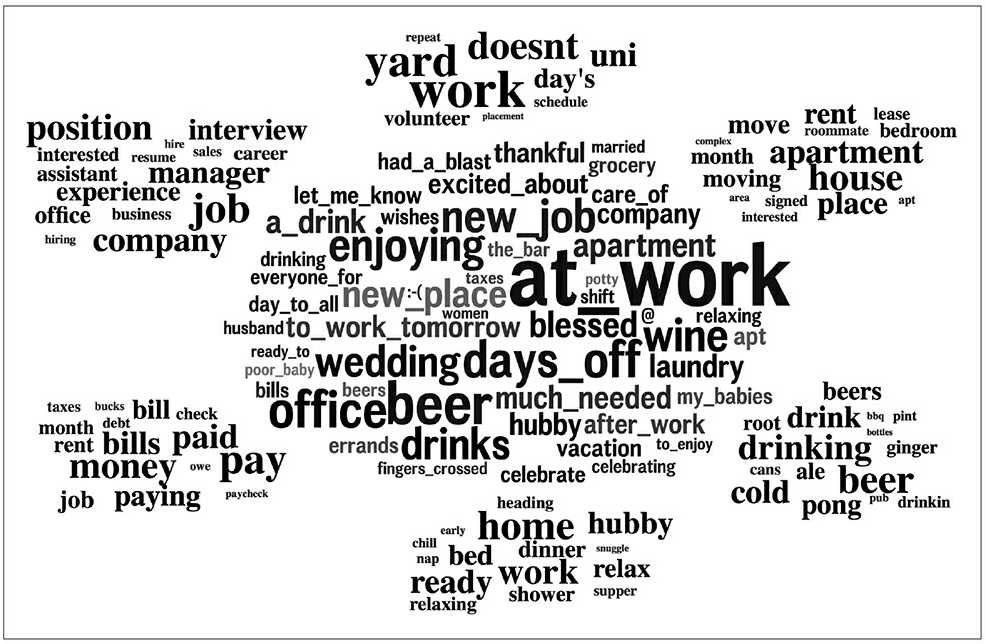
It isn't just men and women who speak differently. People use different words as they age. This might even give us some clues as to how the aging process plays out. Here, from the same study, are the words most disproportionately used by people of different ages on Facebook. I call this graphic "Drink. Work. Pray." In people's teens, they're drinking. In their twenties, they are working. In their thirties and onward, they are praying.

DRINK.WORK.PRAY

19- to 22-year-olds



23- to 29-year-olds



30- to 65-year-olds

A powerful new tool for analyzing text is something called sentiment analysis. Scientists can now estimate how happy or sad a particular passage of text is.

How? Teams of scientists have asked large numbers of people to code tens of thousands of words in the English language as positive or negative. The most positive words, according to this methodology, include “happy,” “love,” and “awesome.” The most negative words include “sad,” “death,” and “depression.” They thus have built an index of the mood of a huge set of words.

Using this index, they can measure the average mood of words in a passage of text. If someone writes “I am happy and in love and feeling awesome,” sentiment analysis would code that as extremely happy text. If someone writes “I am sad thinking about all the world’s death and depression,” sentiment analysis would code that as extremely sad text. Other pieces of text would be somewhere in between.

So what can you learn when you code the mood of text? Facebook data scientists have shown one exciting possibility. They can estimate a country’s Gross National Happiness every day. If people’s status messages tend to be positive, the country is assumed happy for the day. If they tend to be negative, the country is assumed sad for the day.

Among the Facebook data scientists’ findings: Christmas is one of the happiest days of the year. Now, I was skeptical of this analysis—and am a bit skeptical of this whole project. Generally, I think many people are secretly sad on Christmas because they are lonely or fighting with their family. More

generally, I tend not to trust Facebook status updates, for reasons that I will discuss in the next chapter—namely, our propensity to lie about our lives on social media.

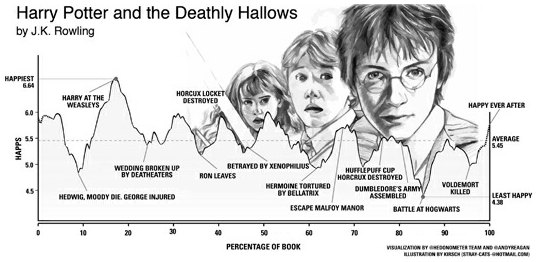
If you are alone and miserable on Christmas, do you really want to bother all of your friends by posting about how unhappy you are? I suspect there are many people spending a joyless Christmas who still post on Facebook about how grateful they are for their "wonderful, awesome, amazing, happy life." They then get coded as substantially raising America's Gross National Happiness. If we are going to really code Gross National Happiness, we should use more sources than just Facebook status updates.

That said, the finding that Christmas is, on balance, a joyous occasion does seem legitimately to be true. Google searches for depression and Gallup surveys also tell us that Christmas is among the happiest days of the year. And, contrary to an urban myth, suicides drop around the holidays. Even if there are some sad and lonely people on Christmas, there are many more merry ones.

These days, when people sit down to read, most of the time it is to peruse status updates on Facebook. But, once upon a time, not so long ago, human beings read stories, sometimes in books. Sentiment analysis can teach us a lot here, too.

A team of scientists, led by Andy Reagan, now at the University of California at Berkeley School of Information, downloaded the text of thousands of books and movie scripts. They could then code how happy or sad each point of the story was.

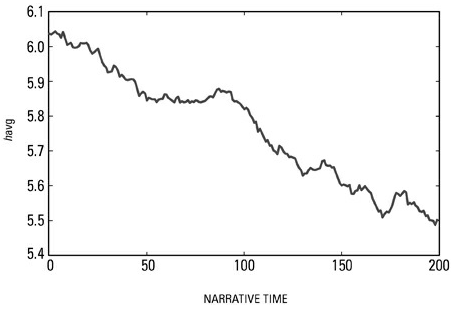
Consider, for example, the book *Harry Potter and the Deathly Hallows*. Here, from that team of scientists, is how the mood of the story changes, along with a description of key plot points.



Note that the many rises and falls in mood that the sentiment analysis detects correspond to key events.

Most stories have simpler structures. Take, for example, Shakespeare's tragedy *King John*. In this play, nothing goes right. King John of England is asked to renounce his throne. He is excommunicated for disobeying the pope. War breaks out. His nephew dies, perhaps by suicide. Other people die. Finally, John is poisoned by a disgruntled monk.

And here is the sentiment analysis as the play progresses.

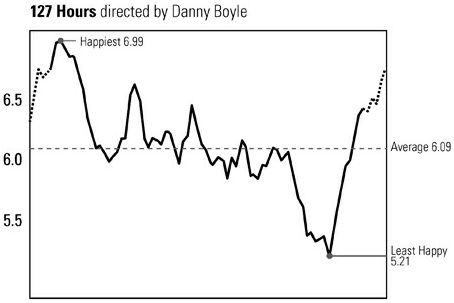


In other words, just from the words, the computer was able to detect that things go from bad to worse to worst.

Or consider the movie *127 Hours*. A basic plot summary of this movie is as follows:

> A mountaineer goes to Utah's Canyonlands National Park to hike. He befriends other hikers but then parts ways with them. Suddenly, he slips and knocks loose a boulder, which traps his hand and wrist. He attempts various escapes, but each one fails. He becomes depressed. Finally, he amputates his arm and escapes. He gets married, starts a family, and continues climbing, although now he makes sure to leave a note whenever he goes off.

And here is the sentiment analysis as the movie progresses, again by Reagan's team of scientists.



So what do we learn from the mood of thousands of these stories?

The computer scientists found that a huge percentage of stories fit into one of six relatively simple structures. They are, borrowing a chart from Reagan's team:

Rags to Riches (rise)
Riches to Rags (fall)
Man in a Hole (fall, then rise)
Icarus (rise, then fall)
Cinderella (rise, then fall, then rise)
Oedipus (fall, then rise, then fall)

There might be small twists and turns not captured by this simple scheme. For example, *127 Hours* ranks as a Man in a Hole story, even though there are moments along the way down when sentiments temporarily improve. The large, overarching structure of most stories fits into one of the six categories. *Harry Potter and the Deathly Hallows* is an exception.

There are a lot of additional questions we might answer. For example, how has the structure of stories changed through time? Have stories gotten more complicated through the years? Do cultures differ in the types of stories they tell? What types of stories do people like most? Do different story structures appeal to men and women? What about people in different countries?

Ultimately, text as data may give us unprecedented insights into what audiences actually want, which may be different from what authors or executives think they want. Already there are some clues that point in this direction.

Consider a study by two Wharton School professors, Jonah Berger and Katherine L. Milkman, on what types of stories get shared. They tested whether positive stories or negative stories were more likely to make the *New York Times*' most-emailed list. They downloaded every *Times* article over a three-month period. Using sentiment analysis, the professors coded the mood of articles. Examples of positive stories included "Wide-Eyed New Arrivals Falling in Love with the City" and "Tony Award for Philanthropy." Stories such as "Web Rumors Tied to Korean Actress' Suicide" and "Germany: Baby Polar Bear's Feeder Dies" proved, not surprisingly, to be negative.

The professors also had information about where the story was placed. Was it on the home page? On the top right? The top left? And they had information about when the story came out. Late Tuesday night? Monday morning?

They could compare two articles—one of them positive, one of them negative—that appeared in a similar place on the *Times* site and came out at a similar time and see which one was more likely to be emailed.

So what gets shared, positive or negative articles?

Positive articles. As the authors conclude, "Content is more likely to become viral the more positive it is."

Note this would seem to contrast with the conventional journalistic wisdom that people are attracted to violent and catastrophic stories. It may be true that news media give people plenty of dark stories. There is something to the newsroom adage, "If it bleeds, it leads." The Wharton professors' study, however, suggests that people may actually want more cheery stories. It may suggest a new adage: "If it smiles, it's emailed," though that doesn't really rhyme.

So much for sad and happy text. How do you figure out what words are liberal or conservative? And what does that tell us about the modern news media? This is a bit more complicated, which brings us back to Gentzkow and Shapiro. Remember, they were the economists who saw gay marriage described different ways in two different newspapers and wondered if they could use language to uncover political bias.

The first thing these two ambitious young scholars did was examine transcripts of the *Congressional Record.* Since this record was already digitized, they could download every word used by every Democratic congressperson in 2005 and every word used by every Republican congressperson in 2005. They could then see if certain phrases were significantly more likely to be used by Democrats or Republicans.

Some were indeed. Here are a few examples in each category.

PHRASES USED FAR MORE BY DEMOCRATS	PHRASES USED FAR MORE BY REPUBLICANS
Estate tax	Death tax
Privatize social security	Reform social security
Rosa Parks	Saddam Hussein
Workers rights	Private property rights
Poor people	Government spending

What explains these differences in language?

Sometimes Democrats and Republicans use different phrasing to describe the same concept. In 2005, Republicans tried to cut the federal inheritance tax. They tended to describe it as a "death tax" (which sounds like an imposition upon the newly deceased). Democrats described it as an "estate tax" (which sounds like a tax on the wealthy). Similarly, Republicans tried to move Social Security into individual retirement accounts. To Republicans, this was a "reform." To Democrats, this was a more dangerous-sounding "privatization."

Sometimes differences in language are a question of emphasis. Republicans and Democrats presumably both have great respect for Rosa Parks, the civil rights hero. But Democrats talked about her more frequently. Likewise, Democrats and Republicans presumably both think that Saddam Hussein, the former leader of Iraq, was an evil dictator. But Republicans repeatedly mentioned him in their attempt to justify the Iraq War. Similarly, "workers' rights" and concern for "poor people" are core principles of the Democratic Party. "Private property rights" and cutting "government spending" are core principles of Republicans.

And these differences in language use are substantial. For example, in 2005, congressional Republicans used the phrase "death tax" 365 times and "estate tax" only 46 times. For congressional Democrats, the pattern was reversed. They used the phrase "death tax" only 35 times and "estate tax" 195 times.

And if these words can tell us whether a congressperson is a Democrat or a Republican, the scholars realized, they could also tell us whether a newspaper tilts left or right. Just as Republican congresspeople might be more likely to use the phrase "death tax" to persuade people to oppose it, conservative newspapers might do the same. The relatively liberal *Washington Post* used the phrase "estate tax" 13.7 times more frequently than they used the phrase "death tax." The conservative *Washington Times* used "death tax" and "estate tax" about the same amount.

Thanks to the wonders of the internet, Gentzkow and Shapiro could analyze the language used in a large number of the nation's newspapers. The scholars utilized two websites, newslibrary.com and proquest.com, which together had digitized 433 newspapers. They then counted how frequently one

thousand such politically charged phrases were used in newspapers in order to measure the papers' political slant. The most liberal newspaper, by this measure, proved to be the *Philadelphia Daily News;* the most conservative: the *Billings* (Montana) *Gazette*.

When you have the first comprehensive measure of media bias for such a wide swath of outlets, you can answer perhaps the most important question about the press: why do some publications lean left and others right?

The economists quickly homed in on one key factor: the politics of a given area. If an area is generally liberal, as Philadelphia and Detroit are, the dominant newspaper there tends to be liberal. If an area is more conservative, as are Billings and Amarillo, Texas, the dominant paper there tends to be conservative. In other words, the evidence strongly suggests that newspapers are inclined to give their readers what they want.

You might think a paper's owner would have some influence on the slant of its coverage, but as a rule, who owns a paper has less effect than we might think upon its political bias. Note what happens when the same person or company owns papers in different markets. Consider the New York Times Company. It owns what Gentzkow and Shapiro find to be the liberal-leaning *New York Times*, based in New York City, where roughly 70 percent of the population is Democratic. It also owned, at the time of the study, the conservative-leaning, by their measure, *Spartanburg Herald-Journal*, in Spartanburg, South Carolina, where roughly 70 percent of the population is Republican. There are exceptions, of course: Rupert Murdoch's News Corporation owns what just about anyone would find to be the conservative *New York Post*. But, overall, the findings suggest that the market determines newspapers' slants far more than owners do.

The study has a profound impact on how we think about the news media. Many people, particularly Marxists, have viewed American journalism as controlled by rich people or corporations with the goal of influencing the masses, perhaps to push people toward their political views. Gentzkow and Shapiro's paper suggests, however, that this is not the predominant motivation of owners. The owners of the American press, instead, are primarily giving the masses what they want so that the owners can become even richer.

Oh, and one more question—a big, controversial, and perhaps even more provocative question. Do the American news media, on average, slant left or right? Are the media on average liberal or conservative?

Gentzkow and Shapiro found that newspapers slant left. The average newspaper is more similar, in the words it uses, to a Democratic congressperson than it is to a Republican congressperson.

"Aha!" conservative readers may be ready to scream, "I told you so!" Many conservatives have long suspected newspapers have been biased to try to manipulate the masses to support left-wing viewpoints.

Not so, say the authors. In fact, the liberal bias is well calibrated to what newspaper readers want. Newspaper readership, on average, tilts a bit left. (They have data on that.) And newspapers, on average, tilt a bit left to give their readers the viewpoints they demand.

There is no grand conspiracy. There is just capitalism.

The news media, Gentzkow and Shapiro's results imply, often operate like every other industry on the planet. Just as supermarkets figure out what ice cream people want and fill their shelves with it, newspapers figure out what viewpoints people want and fill their pages with it. "It's just a business," Shapiro told me. That is what you can learn when you break down and quantify matters as convoluted as news, analysis, and opinion into their component parts: words.

PICTURES AS DATA

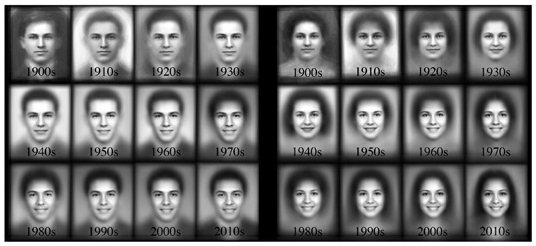
Traditionally, when academics or businesspeople wanted data, they conducted surveys. The data came neatly formed, drawn from numbers or checked boxes on questionnaires. This is no longer the case. The days of structured, clean, simple, survey-based data are over. In this new age, the messy traces we leave as we go through life are becoming the primary source of data.

As we've already seen, words are data. Clicks are data. Links are data. Typos are data. Bananas in dreams are data. Tone of voice is data. Wheezing is data. Heartbeats are data. Spleen size is data. Searches are, I argue, the most revelatory data.

Pictures, it turns out, are data, too.

Just as words, which were once confined to books and periodicals on dusty shelves, have now been digitized, pictures have been liberated from albums and cardboard boxes. They too have been transformed into bits and released into the cloud. And as text can give us history lessons—showing us, for example, the changing ways people have spoken—pictures can give us history lessons—showing us, for example, the changing ways people have posed.

Consider an ingenious study by a team of four computer scientists at Brown and Berkeley. They took advantage of a neat digital-era development: many high schools have scanned their historical yearbooks and made them available online. Across the internet, the researchers found 949 scanned yearbooks from American high schools spanning the years 1905–2013. This included tens of thousands of senior portraits. Using computer software, they were able to create an "average" face out of the pictures from every decade. In other words, they could figure out the average location and configuration of people's noses, eyes, lips, and hair. Here are the average faces from across the last century plus, broken down by gender:



Notice anything? Americans—and particularly women—started smiling. They went from nearly stone-faced at the start of the twentieth century to beaming by the end.

So why the change? Did Americans get happier?

Nope. Other scholars have helped answer this question. The reason is, at least to me, fascinating. When photographs were first invented, people thought of them like paintings. There was nothing else to compare them to. Thus, subjects in photos copied subjects in paintings. And since people sitting for portraits couldn't hold a smile for the many hours the painting took, they adopted a serious look. Subjects in photos adopted the same look.

What finally got them to change? Business, profit, and marketing, of course. In the mid-twentieth century, Kodak, the film and camera company, was frustrated by the limited number of pictures people were taking and devised a strategy to get them to take more. Kodak's advertising began associating photos with happiness. The goal was to get people in the habit of taking a picture whenever they wanted to show others what a good time they were having. All those smiling yearbook photos are a result of that successful campaign (as are most of the photos you see on Facebook and Instagram today).

But photos as data can tell us much more than when high school seniors began to say "cheese." Surprisingly, images may be able to tell us how the economy is doing.

Consider one provocatively titled academic paper: “Measuring Economic Growth from Outer Space.” When a paper has a title like that, you can bet I’m going to read it. The authors of this paper—J. Vernon Henderson, Adam Storeygard, and David N. Weil—begin by noting that in many developing countries, existing measures of gross domestic product (GDP) are inefficient. This is because large portions of economic activity happen off the books, and the government agencies meant to measure economic output have limited resources.

The authors’ rather unconventional idea? They could help measure GDP based on how much light there is in these countries at night. They got that information from photographs taken by a U.S. Air Force satellite that circles the earth fourteen times per day.

Why might light at night be a good measure of GDP? Well, in very poor parts of the world, people struggle to pay for electricity. And as a result, when economic conditions are bad, households and villages will dramatically reduce the amount of light they allow themselves at night.

Night light dropped sharply in Indonesia during the 1998 Asian financial crisis. In South Korea, night light increased 72 percent from 1992 to 2008, corresponding to a remarkably strong economic performance over this period. In North Korea, over the same time, night light actually fell, corresponding to a dismal economic performance during this time.

In 1998, in southern Madagascar, a large accumulation of rubies and sapphires was discovered. The town of Ilakaka went from little more than a truck stop to a major trading center. There was virtually no night light in Ilakaka prior to 1998. In the next five years, there was an explosion of light at night.

The authors admit their night light data is far from a perfect measure of economic output. You most definitely cannot know exactly how an economy is doing just from how much light satellites can pick up at night. The authors do not recommend using this measure at all for developed countries, such as the United States, where the existing economic data is more accurate. And to be fair, even in developing countries, they find that night light is only about as useful as the official measures. But combining both the flawed government data with the imperfect night light data gives a better estimate than either source alone could provide. You can, in other words, improve your understanding of developing economies using pictures taken from outer space.

Joseph Reisinger, a computer science Ph.D. with a soft voice, shares the night light authors’ frustration with the existing datasets on the economies in developing countries. In April 2014, Reisinger notes, Nigeria updated its GDP estimate, taking into account new sectors they may have missed in previous estimates. Their estimated GDP was now 90 percent higher.

“They’re the largest economy in Africa,” Reisinger said, his voice slowly rising. “We don’t even know the most basic thing we would want to know about that country.”

He wanted to find a way to get a sharper look at economic performance. His solution is quite an example of how to reimagine what constitutes data and the value of doing so.

Reisinger founded a company, Premise, which employs a group of workers in developing countries, armed with smartphones. The employees’ job? To take pictures of interesting goings-on that might have economic import.

The employees might get snapshots outside gas stations or of fruit bins in supermarkets. They take pictures of the same locations over and over again. The pictures are sent back to Premise, whose second group of employees—computer scientists—turn the photos into data. The company’s analysts can code everything from the length of lines in gas stations to how many apples are available in a supermarket to the ripeness of these apples to the price listed on the apples’ bin. Based on photographs of all sorts of activity, Premise can begin to put together estimates of economic output and inflation. In developing countries, long lines in gas stations are a leading indicator of economic trouble. So are unavailable or unripe apples. Premise’s on-the-ground pictures of China helped them discover food inflation there in 2011 and food deflation in 2012, long before the official data came in.

Premise sells this information to banks or hedge funds and also collaborates with the World Bank.

Like many good ideas, Premise’s is a gift that keeps on giving. The World Bank was recently interested in the size of the underground cigarette economy in the Philippines. In particular, they wanted to know the effects of the government’s recent efforts, which included random raids, to crack down on manufacturers that produced cigarettes without paying a tax. Premise’s clever idea? Take photos of cigarette boxes seen on the street. See how many of them have tax stamps, which all legitimate cigarettes do. They have found that this part of the underground economy, while large in 2015, got significantly smaller in 2016. The government’s efforts worked, although seeing something usually so hidden—illegal cigarettes—required new data.

As we’ve seen, what constitutes data has been wildly reimagined in the digital age and a lot of insights have been found in this new information. Learning what drives media bias, what makes a good first date, and how developing economies are really doing is just the beginning.

Not incidentally, a lot of money has also been made from such new data, starting with Messrs. Brin’s and Page’s tens of billions. Joseph Reisinger hasn’t done badly himself. Observers estimate that Premise is now making tens of millions of dollars in annual revenue. Investors recently poured $50 million into the company. This means some investors consider Premise among the most valuable enterprises in the world primarily in the business of taking and selling photos, in the same league as *Playboy*.

There is, in other words, outsize value, for scholars and entrepreneurs alike, in utilizing all the new types of data now available, in thinking broadly about what counts as data. These days, a data scientist must not limit herself to a narrow or traditional view of data. These days, photographs of supermarket lines are valuable data. The fullness of supermarket bins is data. The ripeness of apples is data. Photos from outer space are data. The curvature of lips is data. Everything is data!

And with all this new data, we can finally see through people’s lies.