Everybody Lies: Big Data, New Data, & what the Internet can tell us about who we really are, Seth Stephens (2017)

At the privacy of their keyboards, people confess things (as in dating sites or searches for professional advice) because they have real-life consequences, at other times precisely because they don't have consequences. **Google Trends**, a tool released in 2009, tells users how frequently any word or phrase has been searched in different locations at different times. At first Google search data didn't seem to be a proper source of information for "serious" academic research. Google was invented so that people could learn about the world, not so researchers could learn about people. But it turns out that people's search for information is, in itself, information, although my study was initially rejected by 5 academic journals. Google searches seemed like such a bizarre dataset. But the act of typing a word or phrase leaves a trace of truth that, when multiplied by millions, eventually reveals profound realities.

Surveys and conventional wisdom placed modern racism predominantly in the South and mostly among Republicans. But the places with the highest racist search rates included upstate NY, western PA, eastern OH, WV, southern LA, and MS. The true divide, was not South versus North; it was East versus West. Google searches, in other words, helped draw a new map of racism in the US. In parts of the country with a high number of racist searches in 2008, Obama did worse than Kerry had 4 years earlier. Racist searches did not predict poor performant for any *other* Democratic candidate. Only for Obama. Now that we have witnessed the inauguration of President Trump, my finding seems more plausible. Google searches revealed a darkness and hatred among a meaningful number of Americans that pundits missed. Search data revealed that we live in a very different society from the one academics and journalists, relying on polls, thought we lived in. It revealed a widespread rage that was waiting for a candidate to give voice to.

Google searches for "how to vote" or "where to vote" weeks before an election can accurately predict which parts of the country are going to have a big showing at the polls. A person is also significantly more likely to put the candidate they support first in a paired search. The order the candidates were searched was predictive of which way a particular state would go. This indicator could contain information that polls miss because voters are either lying to themselves or uncomfortable revealing their true preferences to pollsters. There were more searches for "Trump Clinton" than "Clinton Trump" in key states in the Midwest that Clinton was expected to win. Black Americans told polls they would turn out in large numbers to oppose Trump. But Google searches for information on voting in heavily black areas were way down. On election day, Clinton would be hurt by low black turnout.

The revelations have kept coming. Mental illness; human sexuality; child abuse; abortion; advertising; religion; health. Not exactly small topics and this dataset, which didn't exist a couple of decades ago, offered surprising new perspectives on all of them. Economists and other social scientists are always hunting for new sources of data. I am now convinced that Google searches are the most important dataset ever collected on the human psyche. Much of the new information flows from Google and social media. Some of it is a product of digitization of information that was previously hidden in cabinets and files. On an average day, human beings generate 2.5 million trilling bytes of data.

There are higher search rates for anxiety in rural upstate NY than NYC. I tested how much Google searches reflecting anxiety rose in a country in the days, weeks, and months following every major European or American terrorist attack since 2004. They didn't. At all. Searches for jokes are lowest on Monday, the day when people report they are most unhappy. They are lowest on cloudy and rainy days. People are actually more likely to seek out jokes when things are going well in life than when they aren't.

Too many businesses are drowning in data. They have lots of terabytes but few insights. The smartest Big Data companies are often cutting down their data. At Google, major decisions are based on only a tiny sampling of all their data. A major reason that Google searches are so valuable is not that there are so many of them; it is that people are so honest in them.

When I was 33 years old at a Thanksgiving gathering of family, my grandmother told me, "Seth, you need a nice girl. Not too pretty. Very smart. Good with people. Social so you will do things. Sense of humor, because you have a good sense of humor." She gave the best advice of the family's many comments on finding a wife for me because she has access to the largest number of data points. She is Big Data. Good data science is less complicated than many think. The best data science, in fact, is surprisingly intuitive. It is about spotting patterns and predicting how one variable will affect another. People do this all the time.

A team of researchers from Columbia University and Microsoft wanted to find what symptoms predict pancreatic cancer. This disease has a low 5-year survival rate—only about 3%--but early detection can double a patient's chances. What symptoms predict pancreatic cancer? Searching for back pain and then yellowing skin turned out to be a sign of pancreatic cancer; searching for just back pain alone made it unlikely someone had it. Similarly, searching for indigestion and then abdominal pain was evidence of pancreatic cancer, while searching for just indigestion without abdominal pain meant a person was unlikely to have it. The researchers could identify 5-15% of cases with almost no false positives.

Who else uses this methodology to figure out whether someone has a disease? Husbands and wives, mothers and fathers, and nurses and doctors. Based on experience and knowledge, they try to connect fevers, headaches, runny noses, and stomach pains to various diseases. The size of the dataset matters. Sometimes there is insufficient experience for our unaided gut to draw upon. While our gut may usually give us a good general sense of how the world works, it is frequently not precise. We need data to sharpen the picture. In winter months, warm climates, such as Honolulu, have 40% fewer depression searches than cold climates, such as Chicago. A move from Chicago-to-Honolulu would be twice as effective as that of the best antidepressant.

One cognitive trap is to exaggerate the relevance of our own experience and be thrown off by the basic human fascination with the dramatic. People consistently rank tornadoes as a more common cause of death then asthma. In fact, asthma causes 70X more deaths. While the methodology of good data science is often intuitive, results are sometime counterintuitive.

Do most NBA players grow up in poverty—like LeBron James? The best way to get the right answer is to combine all available data. Black men are about 40X more likely than white men to reach the NBA and have a substantially better chance of reaching the NBA if they were born in a wealthy county. The family backgrounds of the best black NBA players suggest that a comfortable background is a big advantage for achieving success. Among African-Americans, poor, uneducated, and single moms tend to give their kids different names from middle-class, educated, and married parents. California-born NBA players were half as likely to have unique names as the average black male, a statistically significant difference. The county of birth, the marital status of the mothers of the top scorers, and the first names of players all support the same story. Better socioeconomic status means a higher chance of making the NBA. Poor men tend to end up shorter. The average man in the developed world is now 4" taller than a century and a half ago. The average NBA player is 6'7"; the average American man is 5'9". Each additional inch roughly doubles your odds of making it to the NBA. Economists have found that middle-class, 2-parent families are on average substantially better at raising kids who are trusting, disciplined, persistent, focused, and organized. The data tell us that in worse-off families and communities there are NBA-level talents who are not in the NBA. These men had the genes and ambition, but never developed the temperament to become basketball superstars.

Offering up **new types of data** and **honest** data are the first 2 powers of Big Data. In the digital age, people still hide their thoughts from other people, but not from the Internet. Allowing us to **zoom in on small subsets** of people is the 3rd power of Bid Data, in rapid, controlled experiments to test for causality, not merely correlations. Allowing us to do **many causal experiments** is the 4th power of Big Data.

Searches such as "flu symptoms" and "muscle aches" have proven important indicators of how fast a flu is spreading. A service called **Google Correlate** gives researchers the means to experiment with the same type of analysis across a wide range of fields, not just health. We were able to show which searches most closely track housing prices. When prices are rising, Americans tend to search for such phrases as "80/20 mortgage," "new home builder," and "appreciation rate." Searches of porn sites and "Spider Solitaire" are a predicter of the unemployment rate--all are diversion-related.

Google didn't dominate search merely by collecting more data than everyone else. They did it by finding a *better* type of data. The Big Data revolution is less about collecting more data than collecting the right data.

Jeff Seder was an agent who provided information on which race horses to buy at auction. On his recommendation, an Egyptian owner bought one horse for $300,000 out of 152 for sale in the summer of 2013 by. 62 horses at the auction sold for more, with 2 fetching more than $1mn. Three months later the horse was renamed American Pharaoh and 18 months later became the first horse in more than 3 decades to win the Triple Crown. Seder had found over the previous 12 years 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. He also found that certain gaits correlated with racetrack success. Seder's genius was to look for data where others hadn't looked before, to consider nontraditional sources of data.

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. When trying to make predictions, you needn't worry too much about why your models work. Examples: 1) Strawberry Pop-Tarts sell 7x faster than normal in the days leading up to a hurricane. Who knows why? Does it matter? 2) The quality of wine can be broken down to a simple formula: Price = 12.145 + 0.00117 winter rainfall + 0.0614 average growing season temperature – 0.00386 harvest rainfall. The rest is hooey. For a data scientist, a fresh and original perspective can pay off.

Language is being used so much now that there is an entire field devoted to it--"text as data." We can predict whether a man and woman will to move on to a second date based on how they speak on the first date. A monotone voice is often seen by women as masculine, which implies that men, perhaps subconsciously, exaggerate their masculinity when they like a woman. A woman is likely to be interested when she talks about herself. Physical appearance trumps all else in predicting whether a man reports a connection. 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.

Scientist can now estimate how happy or sad a particular passage of text is by counting positive words, such as "happy, love and awesome," or negative words, such as "sad, death and depression." They can estimate a country's Gross National Happiness every day. Christmas is one of the happiest days of the year. Suicides drop around the holidays. The likelihood that stories become viral depends on how positive the content. Note this contrasts with the conventional wisdom that people are attracted to violent and catastrophic stories. "If it bleeds, it leads." It may suggest a new adage: "If it smiles, it's emailed." Democrats and Republicans may use different phrases to describe the same concept, depending on whether they view it favorably or not: estate tax or death tax, privatize social security or reform social security, and workers' rights or private property rights. Liberal *Washington Post* used the phrase "estate tax" 13.7X more frequently than "death tax." Conservative *Washington Times* used "death tax and "estate tax" about the same amount. A paper's owner has less effect on its slant than its market. Newspapers are inclined to give their readers what they want.

Light is data. We can measure GDP based on how much light there is in countries at night. In very poor parts of the world, people struggle to pay for electricity. As a result, when economic conditions are bad, households and villages dramatically reduce the amount of light they allow themselves at night. Combining both flawed government data with the imperfect night light data gives a better GDP estimate than either source alone.

Pictures are data. In developing countries, long lines at gas stations are a leading indicator of economic trouble. So are unavailable or unripe apples. *Premise* is a company which employs people in developing countries with smartphones to take pictures of interesting things that might have economic import. Their on-the-ground pictures of China helped discover food inflation there in 2011 and food deflation in 2012, long before official data came in. Premise sells this information to banks or hedge funds and collaborates with the World Bank. It now makes tens of millions of dollars in annual revenue from pictures--in the same league as *Playboy.*

Why do people misinform anonymous surveys? Roger Tourangeau, perhaps the world's foremost expert on social desirability bias explained, "About 1/3 of the time, people lie in real

life. The habits carry over to surveys. Then there's that odd habit we sometimes have of lying to ourselves." This may explain why so many say they are above average. If you are deluding yourself, you can't be honest in a survey. The more impersonal the conditions, the more honest people will be. Internet surveys are better than phone surveys, which are better than in-person surveys. People will admit more if they are alone than if others are in the room with them. Also, people have no incentive to tell surveys the truth. If you think you may be suffering from depression, you don't have an incentive to admit this to a survey. You do have an incentive to ask Google for symptoms and potential treatments. People are 7X more likely to ask Google whether they will regret not having children than whether they will regret having children. Adults with children are 3.6X more likely to tell Google they regret their decision than are adults without children.

I was able to find evidence of implicit prejudice against young girls harbored by their parents. Parents are 2.5x more likely to ask "Is my son gifted?" than "Is my daughter gifted?" They show a similar bias when using other phrases like, "Is my son a genius?" Their overriding concern regarding their daughters is related to appearance. Even though about 28% of girls are overweight, while 35% of boys are, parents see—or worry about—overweight girls much more frequently then overweight boys.

The data tell us the chances that 2 people visiting the same news site have different political views is about 45%. So, the internet is close to desegregated. Liberals and conservatives are "meeting" each other on the web all the time. You are more likely to come across someone with opposing views online than offline. The internet news industry is dominated by a few massive sites. Yahoo news remains the most popular news site among Americans. Many people with strong political opinions visit sites of the opposite view point, if only to get angry and argue. Someone who visits rushlimbaugh.com or glennbeck.com is more likely than the average internet user to visit nytimes.com. None of the top 10 most visited websites is pornographic.

In the Facebook world, family life seems perfect. We see other people's social media posts but not their searches. In the Google world, family life is messy. Netflix learned a similar lesson early on in its life cycle: don't trust what people tell you; trust what they do. Originally, Netflix users were filling their queues for future viewing with aspirational and highbrow movies, such as WW2 documentaries or foreign films. But days later, when they were reminded of the movies on the queue, they rarely clicked. Netflix stopped asking people to tell them what they wanted to see in the future and started building a model based on millions of clicks and viewers from similar customers. The result: customers visited Netflix more frequently and watched more movies. The algorithms know you better than you know yourself.

When we lecture angry people, the search data implies that their fury can grow. But subtly provoking their curiosity, giving new information, and offering new images of the group that is stoking their rage may turn their thoughts in different, more positive directions. We see this from how racist searches change after a black quarterback is drafted in a city or how sexist searches change after a woman is elected to office. We see how racism responds to community policing or how sexism responds to new sexual harassment laws.

A huge predictor of Mets fandom is whether the Mets won a World Series when they were around the age of 8. The most important year in a man's life, for the purposes of cementing his favorite baseball team as an adult, is when he is about 8 years old. For women, the patterns are much less sharp, but the peak age appears to be 22 years old. Many of our adult behavior and interests can be explained by the arbitrary facts of when we were born and what was going on in key years while we were young. Between the key ages of 14-24, many Americans will form their political views based on the popularity of the current president. And these views will, on average, last a lifetime. The Eisenhower experience resulted in a 10% boost for Republicans born in 1941. The Kennedy, Johnson, and Nixon experience gave Democrats a 7% advantage among Americans born in 1952.

Big Data is not just about doing the same things we would have done with surveys, except with more data. Data with hundreds of millions of people allows us to spot patterns among cities, towns, and neighborhoods, large and small. In some parts of the US the chance of a poor kid succeeding is as high as in any developed country in the world. In others, it is lower than in any developed country in the world. Scientists can zoom in on the small groups of people who moved from city to city to see how that might have affected their prospects. This allowed them to test for causation, not just correlation. And, yes, moving to the right city in one's formative years made a significant difference. Areas that spend more on education provide a better chance to poor kids. Places with more religious people and lower crime do better.

American women in the top 1% of income live 10 years longer than those in the bottom 1%. For men, the gap is 15 years. For the wealthiest Americans, life expectance is hardly affected by where they live. Rich people everywhere tend to develop healthier habits—on average they exercise more, eat better, smoke less, and are less likely to suffer from obesity. For the poor, life expectancy varies greatly depending on where they live. It can add 5 years to life expectancy.

What are the chances of becoming notable enough to warrant a Wikipedia entry? Early exposure to innovation and a college town increased the rate by 20x. A black child born in Tuskegee had the same probability of becoming a notable in a field outside of sports as a white child born on some of the highest-scoring, majority-white college towns. The second attribute was the presence in the county of a big city. NYC produces notable journalists; Boston produces notable scientists, and Los Angeles produces notable actors. Suburban counties, unless they contained a major college, performed far worse than their urban counterparts. Growing up near big ideas is better than growing up with a big backyard. The greater the percentage of foreign-born residents in an area, the higher the proportion of children who go on to notable success. Education spending did not correlate with notable writers, artist, or business leaders. Spending on education helps kids reach the upper middle class but does little to help them become a notable writer, artist, or business leader. Many of these huge successes hated school. Some dropped out.

The doppelganger search (as highlighted in *Moneyball)* is the best method ever used to predict baseball player performance. This zooms in on individuals and even on the traits of individuals. Apple uses something like a doppelganger search to suggest what books you might like. They see what people similar to you select and base their recommendations on that. This could be used to organize and collect all of our health information so that instead of using a one-size-fits-all approach, doctors can find patients just like you. Then they can employ more personalized, more focused diagnoses and treatments. A diagnosis really is a statement that you share properties with previously studied populations. A diagnosis is, in essence, a primitive kind of doppel-

ganger search. Most medical reports still exist on paper, buried in files, and for those that are computerized, they're often locked up in incompatible formats. We often have better data on baseball than health.

Randomized, controlled experiments are the most trusted evidence in any field. One potential reason students in India struggle so much is that teachers don't show up consistently. On a given day in some schools in rural India, more than 40% of teachers are absent. Easther Duflo, a French economist at MIT, ran a test. She and her colleagues randomly divided schools into 2 groups. In one, in addition to their base pay, teachers were paid a small amount—50 rupees, or about $1.15—for every day they showed up to work. By the end of the experiment, girls in schools where teachers were paid to come to class were 7% more likely to be able to write.

Experiments in the digital world have a huge advantage over those in the offline world. Offline experiments can cost thousands or hundreds of thousands of dollars and take months or years to conduct. In the digital world, randomized experiments can be cheap and fast. It makes randomized experiments, which can find truly causal effects, much easier to conduct. This insight quickly spread through Google and the rest of Silicon Valley where randomized controlled experiments have been renamed "A/B testing." This gold-standard testing is cheap, easy and frees us from reliance upon our intuition. We can test literally everything and seemingly small changes can have big effects. There are a thousand people on the other side of the screen whose job it is to break your self-control. Find enough winners of A/B tests and you have an addictive site, an increasing tool of the gaming industry. As expensive as the Super Bowl ads are, they are so effective in upping demand that companies are actually dramatically underpaying for them.

Economists use the arbitrariness of life to test for causal effects. They've found that successful assassinations dramatically alter world history, taking countries on radically different paths. A new leader causes previously peaceful countries to go to war and previously warring counties to achieve peace. Such natural experiments are powerful and will take on increasing importance in an era with more, better, and larger datasets.

Economists found that prisoners assigned to harsher conditions were more likely to commit additional crimes once they left. The tough prison conditions, rather than deterring them from crime, hardened them and made them more violent once they returned to the outside world.

On the "Elite Illusion," the effects of top schools on either side of admission cut-offs is nil. The factors that make you successful are your talent and your drive, not who gives your commencement speech or other advantages that the biggest name-brand school offer. While going to a good school is important, there is little gained from going to the greatest possible school. If future salary is the measure, similar students accepted to similarly prestigious schools who choose to attend different schools end up in about the same place.

People lie—to friends, to surveys, and to themselves—to make themselves look better. But the world also lies to us by presenting us with faulty, misleading data. Companies can learn how to get more customers. The government can learn how to use reimbursement to best motivate doctors. Students can learn what schools will prove most valuable. Experiments demonstrate the potential of Big Data to replace guess, conventional wisdom, and shoddy correlations with what actually works—*causally.* But Big Data does not eliminate the need for all the other ways human have developed over the millennia to understand the world. They complement each other.

Let's apply this to language and loan applications. The more assertive the borrower's promise, the more likely he will break it. Giving a detailed plan of how he can make payments and mentioning commitments he has kept in the past are evidence someone will pay back a loan. Making promises and appealing to your mercy or God is a clear sign someone will default.

People have long been judged by factors not directly related to job performance—the firmness of their handshakes, the neatness of their dress. But a danger of the data revolution is that, as more of our life is quantified, these proxy judgments can get more esoteric and intrusive—and more nefariously discriminating. Casinos gather all the information about a customer they can and find gamblers who are similar to her—her doppelgangers. Then push her to her "pain point."

On the other hand, Big Data also enables consumers to score blows against businesses that overcharge them or deliver shoddy product. One important weapon is sites such as Yelp, that publish reviews of restaurants and other services. Sales data in the state of Washington showed that one fewer star on Yelp will drop a restaurant's revenues 5-9%. Data can tell businesses which customers to avoid and which they can exploit. It can also tell customers the businesses they should avoid and who is trying to exploit them.

Google searches related to criminal activity correlate with criminal activity. Searches related to suicide correlate strongly with state-level suicide rates. But it is a large leap for data science to go from trying to predict the actions of a city to trying to predict the actions of an individual. Suicidal ideation is incredibly common. Suicide is not. In 2015, there were 12,000 searches in the US for "kill Muslims." There were only 12 murders of Muslims reported as hate crimes.

Social science is becoming a real science which is poised to improve our lives.

[People's search for information is, in itself, information. Google searches are the most important dataset ever collected on the human psyche. A major reason that Google searches are so valuable is that people are so honest in them. **new types of data** and **honest** data are the first 2 powers of Big Data. Allowing us to do **many causal experiments** is the 4th power of Big Data. Don't trust what people tell you; trust what they do. The algorithms know you better than you know yourself. Growing up near big ideas is better than growing up with a big backyard. It makes randomized experiments, which can find truly causal effects, much easier to conduct. While going to a good school is important, there is little gained from going to the greatest possible school. Social science is becoming a real science which is poised to improve our lives.
]