Rebecca Goodwin

rj2339

Someone, somewhere is recording everything we do on a daily basis. Data such as what we buy, our phone's location, what terms we search through Google, are all collected (often, but not always, anonymously). The massive amounts of data generated constantly has led to the rise of algorithms designed to make find correlations and draw conclusions. When this kind of analysis is done on personal data generated by humans, there are societal implications. With this new ability to learn about people, a slew of questions arises. Do applications of Big Data analysis help or hurt customers whose data is tracked? To what extent are we entitled privacy from companies? Does the government's access to massive amounts of personal data imply that innocent people's lives are being scrutinized? Can we use Big Data to solve major sociological issues?

Big Data is changing the way we perceive the world around us. Big Data is a driving force in our society today, as we will see in the case of businesses and the government. This paper will examine ways in which Big Data is being analyzed to gain insight. Furthermore, we will see that Big Data not only tells us about the world as it currently is, but has the potential to shape human interactions in the future. Decisions made based on Big Data analysis stand to have a real effect on society. Therefore, we must be careful to use this analysis wisely.

In recent years, there has been an explosion in how much data is generated and collected from sources such as cell phones and web interactions. The aggregation of this data has been coined "Big Data." Applications of Big Data analysis are quite varied, ranging from drawing scientific conclusions from datasets, to analyzing vast amounts of search terms to try to target ads for a particular user. Some of the data analysis techniques discussed in this paper are examples of data mining. Three general categories of data mining are anomaly detection, association learning, and cluster detection (Furnas). Speaking more generally about Big Data, the buzzword does not just refer to making use of ever-increasingly large data sets. "The key feature of the paradigmatic change is that analytic treatment of data is systematically placed at the forefront of intelligent decision-making" (Hilbert 4). Data scientists are not simply collecting data and coming to simple conclusions, but approaching the world from the viewpoint that with enough data, and enough analysis, we can gain an accurate picture of the way things are. With this information, informed decisions can be made and actions taken.

Using data to make decisions is not a new idea. Databases have been around for decades (Bollier 1). There are several factors that have come in to play in recent years. One is that technological advances have allowed more data to be stored and processed more quickly (Bollier 1). Another is that algorithms for finding patterns have grown in sophistication and prevalence (Bollier 1). These algorithms have enabled processing of unstructured data. In the past, data needed to be entered into a database to draw useful conclusions from it. However, algorithms now make it possible to find patterns in unstructured data such as texts, social network posts, and blog entries. (Simon 32-34) Unstructured data makes up the vast majority of what is called "Big Data" (Simon 35).

A huge application of Big Data is in business. Companies use Big Data to ensure that their product choices and prices match customers' desires. "Retailers, like Walmart and Kohl’s, analyze sales, pricing and economic, demographic and weather data to tailor product selections at particular stores and determine the timing of price markdowns" (Lohr, "The Age of Big Data" 3). This kind of data analysis has obvious advantages from a business perspective. Maximizing profits requires pinpointing the delicate balance between supply and demand, and Big Data is a key to finding this sweet spot. As another example of Big Data's use in business, "shipping companies, like U.P.S., mine data on truck delivery times and traffic patterns to fine-tune routing" ("The Age of Big Data" 3). Less wasted product means better revenue for retailers. It is clearly beneficial to retailers to make use of Big Data analysis.

Big Data isn't just about optimizing profits for business. The goal of Big Data analytics in business is to get an accurate picture of how things are. In some cases, these insights help customers, as well. Valocchi, from IBM, pointed out in an interview that "there could be billing disputes that come from customers. Using technology, you can show the lineage of data, to point to errors or proof [about whether the customer's bill correctly reflects] exact consumption" (Danigelis 4). This type of data could either help or hurt customers depending on particular situations, but it is hard to argue against fairness for all parties, which Big Data can help achieve. He gives an example of how Big Data can help customers detect inefficient appliances: "We've developed a range of clip-on sensors [...] that customers can self-install. We're starting to put in new sets of algorithms to look at how efficient your appliances are. [...] We can give maintenance information back to consumers" (Danigelis 4). This example again shows us that data can be used to benefit the customer, since inefficiencies can be detected and repaired to save money.

Retailers hope to learn everything they can about their customers so they can most effectively target their advertisements and tailor their products. However, the tables can also be turned when customers use Big Data to learn about a company rather than vice versa. Big Data allows customers to make better decisions as consumers. For example, customers can save a great of money by comparative shopping when buying a car. "Whereas the automotive industry traditionally had obscure pricing, it's now easy to find automobile prices online. [...] It's estimated that the public is saving $1 billion per year" (Lesk 88). Customers can compare other types of products as well. Amazon's price check app allows customers to compare prices between products in a store and on Amazon (Lesk 88). With this information in hand, the customer can make the choice that gives him the most value for his money. From this perspective, Big Data is having a positive influence on society, as customers are in charge and companies are encouraged to price their items competitively.

However, not everyone agrees that monitoring and analyzing is always a good thing. Bill Stensrud, Chairman and Chief Executive Officer of InstantEncore, also points out the imbalance between the technological ability of an online user and a company. "Vendors are using Big Data to try to acquire the consumer," [...] and they are doing that by using technologies that are beyond the reach of the consumer by orders of magnitude" (Bolier 24). Earlier we examined how Big Data could aid customers in comparison shopping. However, even if Big Data does sometimes help out the customer, large companies have much greater access to Big Data in terms of both data and analysis techniques. Furthermore, when customers make use of Big Data to target ads, they don't necessarily have the customers' best interests in mind. Harrah's Entertainment used data mining techniques to target its gambling customers very effectively. This was very beneficial for the company, but not for the targeted customers. "Many gambling addicts were trying to quit, only to be lured back to the tables by remarkably enticing offers in the mail" (Simon 69). In this case, a company profits while the customer suffers. This does not seem fair from the customer's perspective.

One of the biggest concerns about Big Data is privacy, especially since in many cases, people are not aware that data is even being collected and stored. Earlier, we looked at the advantages Big Data collection can have for customers. However, a somewhat more sinister side of this is that companies may not necessarily want customers to know exactly how much they know. Charles Duhigg of the New York Times did a story on how the retailer Target carefully chooses its ads for customers. Combinations of items can suggest, for instance, that a customer is pregnant. And these combinations are not necessarily obvious--Andrew Pole, Target's marketing statistician, found that purchases such as unscented lotion and cotton balls could signal a woman was approaching her due date. Target found that customers were put off when they received ads only for maternal items, so it began to mix these ads in with other ads so that targeting would be less apparent (Duhigg). While there is nothing overtly immoral in this method of sending out ads, it is a bit unsettling that this kind of information is collected and tracked without the customer's knowledge. It probably does not occur to the average customer that their purchases are being tracked by credit card number and that seemingly unrelated purchases are analyzed.

Should we be concerned about how much companies know about us? Eric Schmidt, previous CEO of Google, made the unsettling comment, "we know where you are. We know where you've been. We can more or less know what you're thinking about" (Simon 188). It is possible that some people are fine with effectively sharing their thoughts with Google. However, even if a particular user has no issue with his data used for whatever purpose a company or social website desires, that user is not the only one whose data is involved. "Social media providers are gathering and exploiting vast amounts of personal data, without quality controls and without the consent of the individuals to whom that data relates. Individuals who volunteer such data have moral responsibility for their actions, but little or no legal responsibility" (Wigan, Clark 49). Tagging a friend in a photograph gives Facebook information about someone who never willingly gave the company access to his data.

How do customers feel about their data being examined? Though there is an argument that targeted ads could help customers find products that suit their needs, online users don't necessarily like that their information is tracked. A survey done by two professors found that two thirds of online users take exception to online tracking by advertisers. They specifically objected to websites tailoring ads based on individual online behavior (Bollier 24). As Marc Rotenberg sums up, "People start getting very uneasy when buying suggestions are made based on how much we know about this particular person, a practice that takes us into the realm of behavioral targeting" (Bollier 23). Privacy is even more of a concern when we consider that online companies may not always be scrupulous. The Federal Trade Commission fined Google in August 2012 "to settle charges that it had bypassed privacy settings in Apple's Safari browser to be able to track users of the browser and show them advertisements" (Simon 186). The heart of the matter is that in many cases, users or customers do not realize they are being tracked.

While customer stalking is a bit creepy, when it comes to tracking people's actions, there is a more terrifying entity than retailers: the government. When data is collected about our every move and analyzed by the government, have we taken our first step towards a Big Brother society? One example of a disconcerting program is the Total Information Awareness program (TIA). TIA was proposed to monitor American citizens' activities, including emails, phone calls, and medical records (Markoff). The program was officially discontinued, but skeptics have argued that the US government is still trying to achieve the goal of total information. Michael McCanne of the "The New Inquiry" argues that the surveillance program PRISM is merely a continuation of TIA. "The PRISM program -- like its various predecessors -- is part of a continuous effort to achieve surveillance beyond the limits of laws and existing technology" (McCanne). Where is the government getting this data? We saw earlier that data can be used to target ads for particular consumers. However, it is not just companies using this data. "What’s new today is the amount of data that can be kept and searched, the sense of being constantly observed [...] Searches, emails, geographic position, and other observations combine to deliver ads for specific products in your particular location. So, when the government decided to look for terrorists, it went to the corporations that had the data" (Lesk 86). This puts companies' collection of customer data in an even more sinister light.

Edward Snowden, former NSA contractor, revealed that the PRISM database combines personal information from companies such as Google, Apple, and Verizon into a single database (McCanne). "[The government has] been collecting metadata about everybody’s telephone calls for more than a decade, while telephone companies have been using the same information commercially. Mobile telephone carriers exploit location data from your calls and sell it to others" (Lesk 86).

It is possible that the government can use phone data to not only determine an individual's current location, but also predict future behavior. An M.I.T study has found that "geo-location data patterns can successfully predict people's future locations and social interactions” (Eagle, Pentland 267). While this was a civilian study, it is disconcerting that this kind of information can be determined from one's phone activity. There is evidence that the government does in fact try to predict individual behavior. Christopher Ketcham interviewed "a former military operative regularly briefed by members of the intelligence community" who spoke of software that makes predictions about a "target's" behavior and keeps track of their associations using social network analysis and artificial intelligence (Ketcham 83). The source said, "the more data you have on a particular target, the better [the software] can predict what the target will do, where the target will go, who it will turn to for help" (Ketcham 83).

One would hope that to make use of such data, the government would at least need a warrant before sifting through innocent citizens' data. However, "in four new orders, which remain classified, the court defined massive data sets as “facilities” and agreed to certify periodically that the government had reasonable procedures in place to minimize collection of “U.S. persons” data without a warrant" (Gellman, Poitras 1).

A reasonable concern about this kind of analysis and targeting is that a law-abiding citizen may be unjustly identified as a potential threat. It seems unfair that someone guilty of no crime should be treated with suspicion. An open letter to the government by the Association for Computing Machinery club asks the question, "Is any level of false positive acceptable - and Constitutional - in such a system?" (Simons, Spafford). If we are simply analyzing data to, say, classify spam, this question is not nearly so important. We will want to minimize false positives, of course, but the mistake is not nearly so grave as if a human is falsely accused of a crime of which he is innocent. When we are talking about actual humans, a mistake becomes a much bigger deal. A founding premise of our country is that anyone is innocent until proven guilty. The uses of Big Data by the government examined in previous paragraphs indicate that the government has access to large amounts of data, which it may examine without citizens' knowledge and without concrete reasons for suspicion. It seems this use of Big Data could cause us to away from "innocent until proven guilty" and towards a Big Brother type world where citizens are innocent until they begin to look a little suspicious.

While probabilistic data is evidence, the people examining such evidence may be too quick to put too much stock in it. For instance, Stephen Baker of Business Week came across a story about an FBI agent who found a correlation between hummus consumption and a neighborhood's potential to be a refuge for terrorists (Bollier 34). He says, "The prejudices of a society are reflected in the algorithms that are searched" (Bollier 34). While hummus correlation is not an overt accusation, it still challenges a basic principle that citizens should be allowed a clean, unbiased slate.

An alternative opinion is that probabilistic data is a legitimate basis for probable cause. Kim Taiple, founder of executive director of the Center for Advanced Studies in Science and Technology, has the opinion that "'If you start from the premise that data is going to exist and the data may be relevant for making a judgment that is important to society', then the goal should not be to ban the use of correlations and data analysis" (Bollier, 34). He compares data to guns--we should control their abuse, but not take them (or data analysis) away. One way to control the abuse of probabilistic data is to make the reasons for a decision available. What makes probabilistic data so disconcerting is that complicated algorithms obfuscate the reasons why an individual is suspicious. Marc Rotenberg has proposed that when using probabilistic evidence, transparency is necessary--perhaps to the point of disclosing algorithms used to make an accusation (Bollier 34). On the one hand we have the concerning information that “Big Data in government hands is already used for activities like the “no fly” list and border searches" (Lesk 87). Rotenberg's counter to this is "if you're going to make decisions about people--such as preventing them from boarding a plan or detaining them as a security risk--then there has to be some fact that someone will stand behind that provides the basis of the decision" (Bollier 34).

In the Foundation Trilogy, Isaac Asimov explores the idea that the future can be predicted by a combination of history, psychology, and statistics. In the prequel, "Prelude to the Foundation," he brings up an interesting point--that the "predicting" the future itself has an effect on what will happen. The plot focuses on a world leader who is trying to get a "prediction" of his continued beneficial reign. Have we seen this kind of "predicting" in our own world?

As Kim Taipale puts it, "[it is] the classic Heisenberg principle problem. Once data is made available and visualized, people can actually act in ways to change the data as a result of having seen it. He was specifically referring to "Google Bombing," the practice of gaming the Google search engine to "raise the ranking of a given page in the search results" (Bollier 6). But there are much deeper implications than just trying to game the system. Making decisions based on Big Data can influence society itself.

Eli Pariser has come up with the idea of a "filter bubble," referring to the fact that online search services tend to show opinions with which the user already agrees (Lesk 87). This practice is intended to help the user, but it begs the question of whether we should be occasionally exposed to more diverse content. "Does that increase divisiveness and partisanship in society, or is it just saving you reading time?" (Lesk 87). Google customizes its search results based on individual search history (Sullivan). Other websites steer our behavior as well. Steve Lohr uses the example of friend suggestions from Facebook, which are based on the personal data you present on your page and on tracked clicks and searches (Lohr "Sure, Big Data is Great, but Don't Forget About Intution" 3). Again, the user's behavior is being directed by the feedback from his own choices. The user clicks on certain links, and then finds recommendations for more of the same--it is harder for the user to happen to come accross something different. Lohr refers to what data scientists call a behavioral loop: "A person feeds in data, which is collected by an algorithm that then presents the user with choices, thus steering behavior" (Lohr 3). What begins as a convenience could be slowly turning into a society in which each person is not forced to challenge preconceived beliefs. Surely, to achieve society's full potential, each individual must look at alternate opinions to form an informed view. Taken to the extreme, if we begin to rely entirely on websites like Google to feed us content particular to our tastes, "we’d never see anything new, just things we already believe" (Lesk 87).

Big Data has been used to identify trends and correlation in society. However, merely identifying current trends can have an influence on the direction (or lack thereof) these trends can take. For instance, take the stereotype that men are naturally more suited to math than women. A study by Andrei Cimpian has shown that when young girls were told that females were less suited for particular mathematical tasks, they did in fact perform worse (Cimpian). However, if they were ignorant of this "disadvantage," they performed at the same level as the male participants (Cimpian). This is not necessarily to say whether men or women are intrinsically better at math, but it is a clue into how societal expectations can influence actual trends. A different study by Robert Rosenthal has shown that when teachers are told they have some students who have scored highly on a test that predicts growth in IQ, those students in fact do better--despite the fact the fact that the "high-scoring" children were actually chosen at random (Rosenthal). Both of these studies tell us something: prior beliefs can have an effect on future outcomes. Wigan and Clark have speculated that premature conclusions from Big Data analysis could serve to reinforce negative stereotypes. "Some harmful inferences will arise from what careful analysis could reveal as false matches. In other cases, ambiguities will provide fertile ground for speculation, innuendo, and the exercise of preexisting biases for, and particularly against, racial, ethnic, religious, and socioeconomic stereotypes" (Wigan, Clark 49). In an age inundated with huge amounts of data that has the potential to influence our beliefs, it is imperative that we recognize the possibility that humans can shape future society based on their beliefs about the present.

While I have argued that there is the potential for unintended consequences as a result of analyzing Big Data, there is also the potential of very intentionally shaping society. Global Pulse is a fascinating project with this intent. Its goal is to reduce major issues on a global scale. Global Pulse "[...] will conduct so-called sentiment analysis of messages in social networks and text messages — using natural-language deciphering software — to help predict job losses, spending reductions or disease outbreaks in a given region. The goal is to use digital early-warning signals to guide assistance programs in advance to, for example, prevent a region from slipping back into poverty" (Lohr). Big Data could change our very future--previously unpredictable crises could now potentially be diverted.

As another example of Big Data being used to predict societal problems, "in economic forecasting, research has shown that trends in increasing or decreasing volumes of housing-related search queries in Google are a more accurate predictor of house sales in the next quarter than the forecasts of real estate economists" (Lohr, "The Age of Big Data" 4).

Kavel Leetaru of the University of Illinois is working on another exciting project. He is developing a supercomputer, Nautilus, which will predict global conflicts. He feeds Nautilus, the computer, blogs and articles to determine the social climate of a particular area. The super-computer found indications of a negative mood in Egypt right before the Gulf War and the Arab Spring in Egypt (Bauer, 1). So far, it has only worked to "predict" the past, but we can foresee a future in which data analysis becomes good enough to predict conflicts in time to react to them.

Like it or not, we are in the information age. Big Data is proving to be an important force in our society. Big Data could turn our nation into a police state, or it could be a key factor to worldwide peace. We must recognize that the implications of Big Data collection and analysis go beyond short-term profit or relevant Google results.

I have examined the societal influences of Big Data, both positive and negative. While this paper brings up several concerns about Big Data's implications, I am not at all arguing that Big Data processing should cease. Indeed, given the sheer amount of data generated constantly, it would be fruitless to argue that we should stop collecting or fail to utilize it. However, it is important to recognize that data analysis can have negative repercussions. Big Data has the potential to change the way society functions, and it is our responsibility to ensure this influence will be a positive one.