Rebecca Goodwin

rj2339

Big Data

It seems that 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.) For better or worse, big data is changing the way we perceive the world around us. Big data not only tells us about the world as it currently is, but has the potential to shape human interactions in the future. This is a powerful ability and we must be careful to use it wisely.

The big data is, in general, about using huge sets of data to draw conclusions. The term is a very broad one, encompassing everything 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. Big data and data mining are not exactly interchangeable, though they have similar meanings. Data mining refers to specific pattern-finding algorithms used on large datasets, while big data more generally refers to the general analysis of large datasets. 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. As Martin Hilbert says in his article "Big Data for Development: From Information- to Knowledge Societies", "the key feature of the paradigmatic change is that analytic treatment of data is systematically placed at the forefront of intelligent decision-making." (Hilbert) We 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.

A key driver of big data is, as would be expected, money. 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) This kind of data analysis has obvious advantages from a business perspective. Maximizing profits requires finding the delicate balance between supply and demand, and big data is a key to give insight on 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." (Lohr) The implications of these types of applications is wonderful multiple perspectives. Less wasted product means better revenue for retailers, less wasted fuel means less pollution. And customers' needs are met better than ever. Big data isn't just about optimizing profits for business. The goal of big data analytics 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) 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. Hunn 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) This example again shows us that data can be used to benefit the customer, since inefficiencies can be detected and repaired to save money. However, not everyone agrees that monitoring and analyzing is always a good thing.

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 (pun intended) 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. While this kind of customer stalking seems a bit creepy, when it comes to tracking people's actions, there is a more terrifying entity than retailers: the government.

A common concern is that when data is collected about our every move and analyzed by the government, we have 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. While the program was officially discontinued, the core of the program remains. The biggest concern is, of course, that of mistakenly identifying a law-abiding citizen as a potential threat. 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?" 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. 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.) While hummus correlation is not an overt accusation, it still challenges a basic principle that citizens should be allowed a clean, unbiased slate. Suppose that a murder is committed and cell phone data reveals I was within walking distance of the crime a few minutes before it happened. This piece of data is evidence that no doubt should weigh into the decision of my guilt. It is not enough by itself to convict me, but combined with other evidence could lead to a strong case for my guilt. But suppose instead the piece of data is that I belong to a race that has been statistically shown more likely to commit this type of crime. Should this sort of evidence weigh into the judge's decision? Many reasonable people would say no, it is important for the judge to be unbiased. But in the growing paradigm of big data, this could be viewed as evidence just like any other piece of data. If I am indeed innocent, other evidence should weigh in my favor. But the question is whether my race should have weighed into the decision at at all. 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." (Lohr) He compares data to guns--we should control their abuse, but not take them (or data analysis) away.

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, the Founder and Executive Director of the Center for Advanced Studies in Science and Technology said at the Eighteenth Annual Aspen Institue Roundtable on Information Technology, "[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) But there are much deeper implications than just trying to game the system. Making decisions based on big data can influence society itself.

Big data has been used to identify trends and correlation in society. However, I make the case by merely identifying current trends can have an influence on the direction (or lack thereof) these trends can take. Let us examine the common stereotype that men are naturally more suited to math than women. Studies have shown that women consistently score lower on math tests than men. However, it is interesting that other studies have shown that when young girls are told females do worse on particular tasks, they in fact do worse. However, if they are ignorant of this "disadvantage," they perform the same (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 absolutely have an effect on the future. Suppose data shows people of minority races are more likely to commit crime. Let us suppose this data is without question. The issue is that this very knowledge can impact future events. Minorities may be treated with suspicion, and may on a subconscious level think, "If I'm already being treated like a criminal, why not act like one?" This of course will generate more data confirming the race/crime correlation which will reinforce people's beliefs in a positive feedback loop. The tragedy is that had this data not been known, perhaps the outcome could have been different. In an age inundated with even more data that has the potential to shape our beliefs, it is imperative that we recognize the human ability to shape future society based on beliefs about the present.

While I have argued that there is the potential for unintended consequences as a result of data mining, there is also the potential of very intentionally shaping society on a global scale. Global Pulse is a fascinating project in this vein. 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 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)

Like it or not, we are in the information age. 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 fruitless to argue that we should stop collecting or fail to utilize it. However, it is important to realize that data does not always paint an accurate picture, and 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.