Brian Kim
WRIT 340
Professor Warford
Illumin Article

Big Data: From Facts to Future

With technology continuing to revolutionize countless facets of society, its rapid expansion in the age of information has given rise to a generation that is scrambling to leverage technology and comprehend the language it is spoken in. One of the archetypal buzzwords of this movement is "Big Data" -- a phrase few modern companies and technical pundits hold back from touting, but what does it really mean? While the fanfare surrounding it is likened to new fad diets or fashion trends, the advent of big data has serious contemporary uses as well as major implications for the future of our society. It's predicting future economic trends before they happen, advancing artificial intelligence to uncharted heights, and telling everyone what song to listen to next. In an age where technical literacy is seemingly no longer an option, it is critical that people understand Big Data's role in shaping the future.

A Formal Introduction to Small Data

On first glance, the phrase big data itself conjures an image of streaming numbers and armies of computers at work, but while this perception is justified, the reality in fact is that data has a surprisingly human face. In fact, it may even be more human than humanity is willing to admit. To truly understand Big Data however, one must start small.

Data, according to the Free On-Line Dictionary of Computing (FOLDOC) is defined as "numbers, characters, images, or other method of recording, in a form which can be assessed by a human or input into a computer. Data only when interpreted by some kind of data processing system does it take on meaning and become information

[1].” As the last part states data requires referencing and analysis to become information that in turn becomes knowledge. This relationship with data and knowledge has existing long before the invention of computers. In the scientific method, data plays a key role in that scientists use data to infer facts and create reproducible conclusions [2].

In a simple science experiment meant to find the melting temperature of ice, data can be the state of ice after some interval of time in different temperatures. In recording the observation that ice becomes water at temperatures over 0 degrees Celsius and remains ice at temperatures under 0 degrees Celsius, a conclusion can be drawn that the melting point of water exists at 0 degrees Celsius.

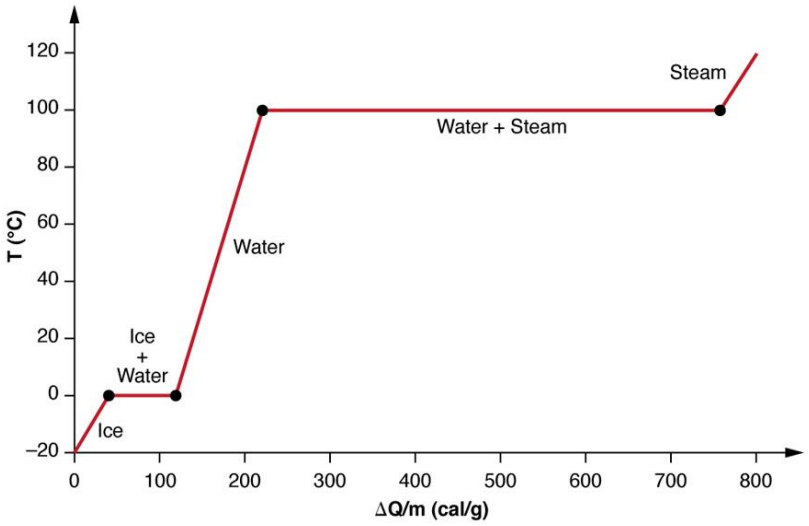

Figure 1: State of Water in varying temperatures. (Source: www.cnx.org)

Data however is no longer limited to the simple empirical observations and handwritten recordings of a few meant to gain explanations to individual phenomena. Now, modern tools allow people to leverage the scientific method paradigm to ask bigger questions and in turn gain more meaningful answers.

Enter Big Data

Big Data's emergence comes from the ability to store massive amounts of information in a variety of electronic devices encoded into binary bits. Thanks to computers data has become much more malleable, indexable, and searchable. As more and more information was stored onto disk drives, IBM was one of the very first to characterize Big Data in 2001 by the 3 V's: Volume, velocity, and variety [3].

Volume refers to the almost 10x growth in the amount of meta-data (peripheral information) that may be stored regarding an individual interaction a person might have such as a transaction on a website. Instead of simply storing the other party and the amount involved, other information such as the exact time stamp, the device used, the geolocation of the individual parties, the time it took to complete the transaction, may all be stored as well. In this regard, smartphones have contributed immensely to the volume of data accumulated in that whether the public is cognizant of it or not, data tracking along with the digital world have gone mobile.

Velocity is considered with regards to how many of these individual transactions may occur at a time, meaning all this data continues to stream into a company's databases.

Variety concerns the different shapes and types that data may come in and how to reconcile these differences to extract information from the conglomerate of records.

These factors combined result in data influx on the order of hundreds of terabytes (a thousand gigabytes) taken in everyday by single companies like Facebook and Boeing [4]. To put that in perspective, a character is represented in a computer by a

single byte in memory. If a word has on average 5 characters, a page has 200 words, and a book has 300 pages, a single gigabyte is sufficient enough to store over 3000 books [5]. 100 terabytes = 300,000,000 books a day.

The consensus is in, data has gotten big. So much so that it would seem impossible to infer anything from so much noise. Instead modern advancements have allowed them to infer more than they could have ever imagined.

Why now more than ever?

The Infrastructure

First point to address is physically storing all this information. Given that a traditional computer has about half a terabyte of space, when taking in data at the scale mentioned before, nontrivial amounts of space, both virtual and physical need to be allocated as shown in a photo of a Google Data Center in Figure 2.

Figure 2: Google Data Center in Council Bluffs, Iowa. (Source: www.google.com/about/datacenters/gallery/#/all/2)

Processing the information introduces another buzzword -- cloud computing. This requires armies of computers to manipulate the data at scales proportional to that of data centers. Through careful setup it is possible to coordinate them to perform either or individualized or collective tasks. By allotting each computer some division of labor, programs can read in the information at the same speed it comes in.

With volume and velocity accounted for, variety seems to pose the largest obstacle in devising programs that can make sense of these collections of raw data in a way that is semantically useful to humans. Thankfully, two of the hottest research fields in all of computer science are dedicated to this task in data mining and machine learning.

The Tools

Data Mining and Machine Learning both deal with “the extraction of hidden predictive information from large databases [6].” Before databases were used to answer single-dimensional questions such as a business’s total capital, which might entail a formula that takes in rows of numbers stored under assets and subtracts rows of numbers stored under liabilities. Such questions could be answered by humans given time by hand. Now, very much like the progress of the scientific method paradigm, data mining and machine learning are geared towards not just finding the information apparent in data like the melting point of water, but creating consistent theories, patterns, and relationships to form new bodies of knowledge entirely.

Figure 3 refers to a simplification of this learning process. Step by step, this approach will be applied to analyzing a person’s collection of music.

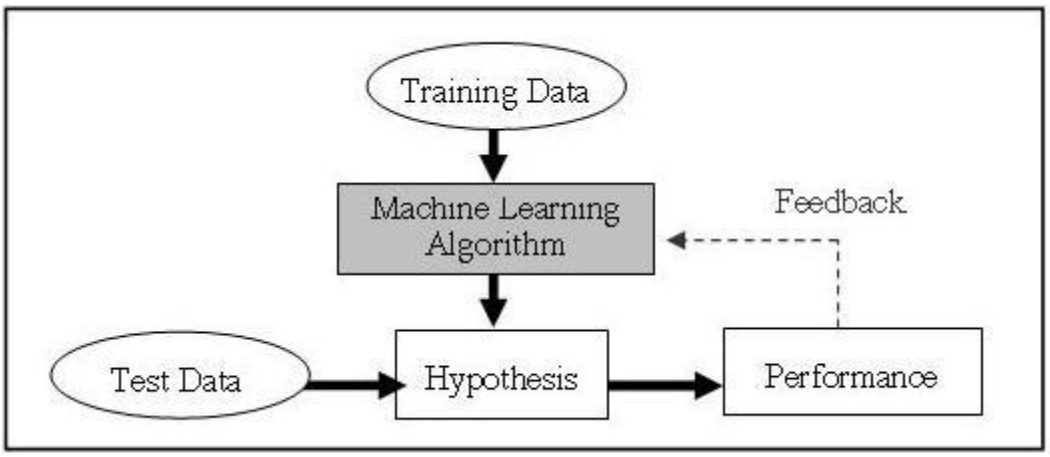

Figure 3: Model representing flow of information in a data mining / machine learning program. (Source: http://machinelearningsoftware.org/wp-content/uploads/2013/03/Machine_Learning_Technique.1.jpg)

1. Create a hypothesis, referred to as a model, which contains a set of features to measure. (Features can be things such as genre of music, number of total plays, or any other characteristic of the data at all -- length of a song, number of words, etc. These features should be scaled down to be represented as numbers from 0 to 1. To start, any random number in that range can be a feature's starting value.)
2. Input data and see how its features compare to the values present in the model.
3. Calculate the error, which is some special measure of the difference between the model and the data inputted.
4. Adjust the values of the model based on the error calculated.
5. Repeat until all the "training" data that adjusts or trains the model has been entered.

6. Extract patterns by viewing feature values of the model. This entails observing what high feature values correspond to other high or low feature values (e.g. Electronic Dance Music songs have few words, but high total plays).
7. Now given new data entries with some features missing their values, such as its genre and song length to see what the other features like its projected number of plays might be based on similar data entries. [6]

However, it's also very important to ask the right questions and use the right type of data, otherwise nonsensical, albeit hilarious relations can be inferred by data as shown in Figure 4.

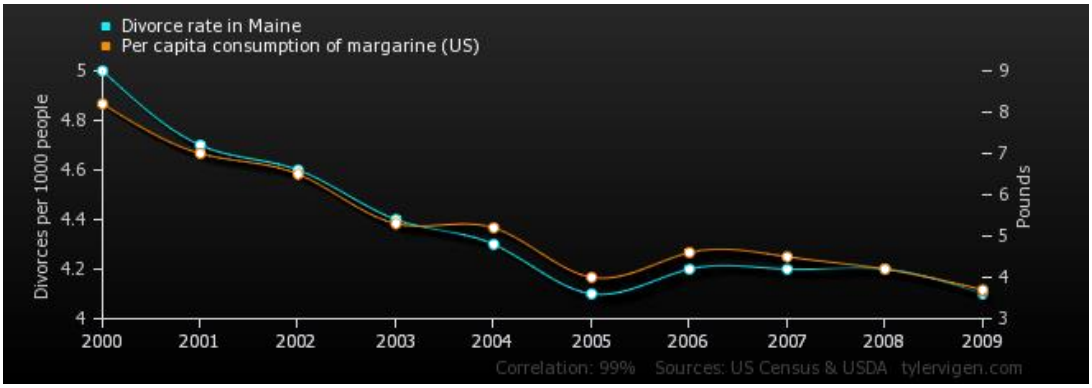

Figure 4: A comical example of how models can unearth information that may not be as useful. (Source: www.tylervigen.com)

At the same time, it is now clear how this information extraction approach can be applied beyond just technology. How much money can business X expect to make this quarter given the state of the economy? What type of clothes is this particular person most likely to buy?

Big Data Today

With the wealth of information that can be extracted, it's no wonder the world is talking. But what exactly are they doing with this technique? Obviously businesses are asking how to make more money, but how and to what extent? According to Federal Reserve Economists, patents, trademarks and copyrights regarding big data could be worth more than $8 trillion [5].

Companies like Google, Facebook, Amazon, and countless others are using this information to serve what type of search results one would most likely click on, compose news feeds that translate to more time spent on the page, and recommend products similar to ones somebody has previously ordered. These insights can very clearly tell us what makes money and what does not. It's exactly why billionaire LinkedIn Founder and investor, Reid Hoffman, urges startups to listen to data when making strategic decisions and why startups stress doing just that [8].

Outside of business, Big Data is playing a huge role in changing the landscape of the public as well. Projects like mapping many different human genomes are now working towards creating a branch of "precision medicine" which Obama is seeking to fund with $215 million [9]. By analyzing how drugs affect people with different genomes it can be made clear what aspects of a drug work for particular people and why.
At the same time, Big Data can be the crux of controversial topics. With data collection at the heart of gaining these insights, agencies like the NSA with regards to recent Edward Snowden leaks have also shown that the drastic measures they are willing to take to gain insight into who is a potential terrorist or protect general national security. These measures can cross ethical lines and cause concern for those who value their personal privacy. As data collection methods in the form of general surveillance and

application backdoors increase, with the help of analytics, tasks such as tracking the movements or full internet activity of an individual are most likely already in development if not already possible.

Big Data is a powerful concept and if there's data to be collected there is information to be extracted. As data for just about anything can be recorded and scales increase, there is no end near in sight for how far this technology can go.

Implications for the Future and Humanity

Pattern recognition seems to come easy to humans, but Big Data continues to not only see the patterns humans cannot, but also continues to get better at patterns humans are meant to recognize every day. Algorithms have been created like the one detailed in the paper titled "Surpassing Human-Level Face Verification Performance on LFW with GaussianFace", which details a program that can recognize a person's face in pictures with varied lighting, viewing angles, and facial expressions better than humans can [10]. Figure 5 shows a sample data set used in this experiment.

Figure 5: Data set used in "Surpassing Human-Level Face Verification Performance on LFW with GaussianFace". (Source: Department of Information Engineering, The Chinese University of Hong Kong)

This slightly counterintuitive and shocking implication that data can enable more than humanity is capable of has been hinted at for a while as well. For example, in the book Alan Turing: The Enigma which inspired the film “The Imitation Game”, Turing explains how World War II was won by not necessarily by the superior weapons or not even necessarily by the valor or superior fighting ability of the soldiers, but by game theory based decisions driven by data they had [11]. Given that data can predict people’s shopping decisions, greatly alter the course of history, or recognize humans better than humans can recognize each other, while there may be a limit to what Big Data can reveal, it seems very, very far away.

Even concepts outside of human control, such as pure randomness are not off limits to data. Something as classically associated with randomness itself like the flip of a coin can be determined if data points like the speed of the coin’s rotation, height above landing surface, and air pressure are collected and analyzed.

If even concepts of randomness, words applied to events outside human understanding, can theoretically be predicted, given a person’s entire life history, body composition down to an atomic level, or a comprehensive state of the world what will Big Data say? Perhaps more than humans are ready for. Whatever Big Data has in store, it is most certainly a buzzword worth buzzing about.

References

[1] Data, *Free On-Line Dictionary of Computing*, [Online]. 1/31/2015, Available: http://foldoc.org/data

[2] Jim Bogen. (2013). "Theory and Observation in Science." *The Stanford Encyclopedia of Philosophy.* [Online]. (Summer 2014 Edition), Available: http://plato.stanford.edu/archives/sum2014/entries/science-theory-observation/

[3] Doug Laney. (2001, Feb.). “3D Data Management: Controlling Data Volume, Velocity, and Variety.” *Application Delivery Strategies.* [Online]. Available: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[4] Big Data Explained, *MongoDB*, [Online]. Available: http://www.mongodb.com/big-data-explained

[5] Karun Chennuri, “How many books can you store in 1 GB hard drive?” *DailyRaaga*. [Online]. Available: https://dailyraaga.wordpress.com/2010/12/06/how-many-books-can-you-*store*-in-1-gb-hard-drive/

[6] An Introduction To Data Mining, *Thearling,* [Online]. Available: http://www.thearling.com/text/dmwhite/dmwhite.htm

[7] Vipal Monga. “The Big Mystery: What’s Big Data Really Worth?” *The Wall Street Journal*. [Online]. Available: http://www.wsj.com/articles/whats-all-that-data-worth-1413157156

[8] Eugene Kim, “The Billionaire Founder Of LinkedIn Explains How To Run A Great Startup.” *Business Insider*. [Online]. Available: http://www.businessinsider.com/reid-hoffman-how-to-be-a-great-founder-2014-11

[9] Obama calls on Congress to fund 'precision medicine'. *Al Jazeera.* Available:

http://america.aljazeera.com/articles/2015/1/30/obama-asks-for-215-million-for-precision-medicine.html

[10] Chaochao Lu, Xiaoou Tang. “Surpassing Human-Level Face Verification Performance on LFW with GaussianFace.” Cornell University Library. Available: http://arxiv.org/abs/1404.3840

[11] Andrew Hodges*, Alan Turing: The Enigma.* US, Simon & Schuster, 2014