Data Analytics for the Public Good

Spring 2019

Contents

Welcome

DATA ANALYTICS FOR THE PUBLIC GOOD:

Building Data-Driven Organizations in the Public and Nonprofit Sector

Watts College of Public Service and Community Solutions, Arizona State University

---

This collaborative text was created through efforts of students in the ASU course PAF 586: Data for the Public Good. The class covers topics about the sources and uses of data in modern organizations, with goals of understanding management approaches to:

- Harness large-scale data to inform policy design, increase stakeholder engagement, and improve service delivery.
- Intelligently consider the social, political, and ethical considerations of data in the public sector.

This text is meant to provide a broad overview of these topics, broken out into themes, with suggestions about how to best integrated new methods and best practices of data collection, management, and utlization. We draw heavily from the following texts:

1. Pentland, A. (2015). *Social Physics: How social networks can make us smarter.* Penguin.
2. Meier, P. (2015). *Digital humanitarians: how big data is changing the face of humanitarian response.* Routledge.
3. Eagle, N., & Greene, K. (2014). *Reality mining: Using big data to engineer a better world.* MIT Press.
4. O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy.* Broadway Books.
5. Duhigg, C. (2016). *Smarter faster better: The secrets of being productive.* Random House.
6. Sutherland, J., Sutherland, J. J. (2014). *Scrum: the art of doing twice the work in half the time.*

Authors

The following students in the course have contributed the chapter summaries and best practice overviews.

1. Lindsey Duncan
2. Rachael Goodwin
3. Erin Hart
4. Thomas Kolwicz
5. Carlos Lopez
6. Joseph Lynch
7. Julie Moore
8. Marcela Morales
9. Lorna Romero
10. William Seeley
11. Matthew Simon
12. Dennis Stockwell
13. Justin Stoker
14. Lauren Zajac

Chapter 1

The Big Promise of Big Data

William Seeley and Lauren Zajac (Team 1)

1.1 Topic Overview

The readings this week spent a great deal of time unveiling the potential, power and promise of big data including what McKinsey and Company referred to as:

"the new frontier for innovation, competition and productivity".

Big data has the promise to drive changes as profound as the industrial revolution according to GE and in 2017 the Economist stated that:

"Data is to this century what oil was to the last century: a driver of growth and change".

One of the key insights in the readings this week is that information overload, or too much information, can be as bad as a lack of information. Industries and companies who can utilize data specialists and new technology and tools to work with the data will have a competitive edge.

1.2 Chapter Summaries

1.2.1 Social Physics CH1 From Ideas to Action

Alex Pentland examined the tremendous potential of the new frontier of big data during chapter 1, entitled "From Ideas to Actions" when he presented his concept of social physics. He contends that while the internet makes our lives more connected, they also make things go faster, and that we are "drowning in information, so much that we don't know what items to pay attention to, and what to ignore" (p. 2). This overwhelming volume and speed of information forms virtual crowds across the world in minutes, leading to catastrophic events like stock market crashes and downfall of governments (p.2). He recognizes that people can no longer be seen as independent decision makers, and that the internet and social media-fueled interactions must be examined to fully explain and predict human behavior.
Social physics is defined as "a quantitative social science that describes reliable, mathematical connections between information and idea flow… and people's behavior" (p. 4). We are able to understand how ideas flow between people using social learning, and how this flow "shapes norms, productivity, creative output" (p. 4). Understanding the information flow and resulting changes in behavior is critical to understanding social physics. Furthermore, social physics is made possible through Big Data, which tracks our "digital bread crumbs" of all aspects of our lives and choices (p. 8). In addition to assisting social scientists with "reality mining" or analyzing patterns within these digital bread crumbs, big data also allows us the opportunity to view society and its complexity "many orders of magnitude over prior social science sets" (p. 11). Pentland

also issues some warnings on the age of Big Data, and he urges that scholars and researchers follow strict scientific policies to ensure the protection of privacy at all costs.

1.2.2 Digital Humanitarians CH1 Rise of Digital Humanitarianism

Patrick Meier dives further into the potential of Big Data, and then discusses his own personal entry into Digital Humanitarianism in Chapter 1, "The Rise of Digital Humanitarianism". He describes his own personal brush with tragedy when his wife was in Port-au-Prince, Haiti, conducting research in 2010 when the devastating earthquake hit. Feeling hopeless and powerless, and struggling with the lack of communication or information, Patrick and his growing network of friends and contacts and volunteers began to launch a "Crisis Response Map" and became digital humanitarianisms. They used the power of social networks and big data information to create, publish and maintain this digital map that pinpointed the worst areas of the disaster and earthquake impact and cries for help in a single map. This was made possible with thousands of volunteers world-wide who devoted hours for many weeks, translating and posting messages, and the map became the source of information from the press, emergency response teams, and friends and family members seeking information on their missing and affected loved ones. Big data and the internet allowed these thousands of volunteers to do something tangible to help during the crisis, and has forever changed the way we can respond to disasters. Meier cautions that Big Data and the overflow of information and data can also be "as paralyzing as the absence of data" (p. 18) and needs to be managed.

1.2.3 NYT The Age of Big Data

Steve Lohr continues this explanation of the power and promise of Big Data in his 2012 Article, *"The Age of Big Data"* published in the *New York Times*. He states that Big Data is a *"new economic asset, like currency or gold"*. Big Data has given rise to a fast growing career field of data consultants who are available and adept at analyzing the data to help industries "make decisions, trim costs and lift sales" using Big Data (p.2). And this Big Data is growing, as it "more than doubles every two years", and creates an "on-line fishbowl…into the real time behavior of huge numbers of people" (p. 6). Lohr goes on to provide a number of illustrations on how pervasive the use of Big Data is by presenting a number of examples in all industries, including business, sports, governments, and academics. Big Data has also fueled the growth of new computer technologies to harness and analyze the data, including artificial intelligence, natural language processing, pattern recognition and machine learning. Lohr also mentions that there is a feedback loop and that additional data actually helps make these tools better. Finally, Lohr also cautions that there are some drawbacks to the data, including false discoveries, all of the data makes as it is hard to focus on what is meaningful, biased fact finding, where people look for data that support their theory, and limits to statistical and mathematical modeling. But drawbacks aside, Big Data has caused a revolution and paradigm shift. People are no longer relying on intuition or feelings in their industries, but are rather focused on the data and analysis.

1.3 Key Take-Aways (for Yellowdig)

VIDEO

1.3.1 Discussion Questions

1. Meier cautions that Big Data and the "overflow of information and data" can also be "as paralyzing as the absence of data" (p. 18). Describe a time when you encountered "too much of a good thing (data)" and what were some strategies you used to overcome the problem?
2. In his article, "The age of Big Data", Lohr mentions that enthusiasts say that the Big Data has the potential to be "humanity's dashboard" with numerous helpful and positive uses, while critics argue that it is just "Big Brother" invading people's privacy. What is your feeling on Big Data?
3. Meier talks about the tangible results (Digital Humanitarianism!) That they found from the use of big data, is there a time that you used data and had immediate results?

1.4 References

1. Pentland, A. (2015). *Social Physics: How social networks can make us smarter.* Penguin. **CH1 From Ideas to Action**
2. Meier, P. (2015). *Digital humanitarians: how big data is changing the face of humanitarian response.* Routledge. **CH1 Rise of Digital Humanitarianism**
3. The Age of Big Data: New York Times LINK

Chapter 2

Information Blindness

Team 2 - Matthew Simon and Carlos Lopez

2.1 The Challenge of Big Data: Information Blindness

2.2 Topic Overview

In this module's reading the three authors pose challenges to the idea of "big data" and its actual usefulness to those that intend to use it. Patrick Meier defines big data as high volume, velocity and variety (Page 28). He uses the idea of social media to make this point. Overall, the theme between these three readings are how people are blind to the data they are collecting, its usefulness and how the user is analyzing the data they are receiving.

Let's take measuring impact for example. Non-profit and government agencies are always trying to demonstrate how they are measuring impact and what their contribution to society is. For business, this story is a little less complicated as they are primarily focused on profitability. However, in social organizations the overarching vision is a little more complex than measuring profitability. A common theme throughout the readings discuss clarity in purpose and focus on the outcome. Are we truly measuring what it is we (the organization) cares about? Are we actually collecting the data in which we need to in order to be able to measure what we care about? Or an even larger philosophical question posed by Gugerty and Karlan, do we (the organization) actually even know what we care about.

In our ever-increasing technological world, we are bombarded with immense amount of information. All of the texts outline human's ability to process large amounts of information. But they also demonstrate the short comings of humans being able to analyze data or how we react to it if it is not digestible in a way to be useful or meaningful. Duhig calls this the shoving everything in the drawer response.

For us, the biggest over-arching theme to these readings is data's connection to the work an organization is actually doing and how it can be utilized to amplify work. You can't just be ok with massive data collection and not using it. Then it is just a waste of resources. Or if you are attempting to use it and using it in a way where people are blind to it or overwhelmed by it, then it becomes an even further waste of resources. You have to be clear about what it is you are striving for and what it is you want to collect. You cannot collect data and analyze before you are clear about why you are collecting and what it is you want to do with it.

Duhig makes excellent points with regards to the education parallels he draws. Government policies have been pushing big data collection on students and student achievement for the past 20 years. The reality is that policymaker's hearts were in the right place, but local schools were ill-equipped to process this data. Teachers would become overwhelmed to this data and weren't using it in a way that could be useful to their students. Teachers first needed to understand what they were assessing and why they were assessing it. Then they needed to understand how far into the data they were going. If a teacher gives a comprehensive

assignment on many topics from an English unit and they only look at the overall grades on tests; they are not going to know what they need to do in order to better prepare different groups of students on their individual needs. The data sets of which they were already collected were too blunt. They needed to understand how students performed on various questions. The data needed to be disaggregated in a better format. Further, there needed to be an investment of time and training in order to better support teachers in utilizing this data. In addition to receiving information and data, teachers were forced to engage with it. They did their own analyses, tested hypothesis, tracked tests and measurements. By engaging directly with the data they were better able to use it to improve student performance.

One of the biggest best practices from these readings comes from Gugerty and Karlan. The reflection questions they pose about theory of change and how to proceed on measuring outcomes is extremely informative. For example, they write:

"Validating the initial steps in the theory of change is a critical step before moving on to measuring impact. Consider a program to deliver child development, health, and nutrition information to expectant mothers in order to improve prenatal care and early childhood outcomes. Starting an impact evaluation before knowing if expectant mothers will actually attend the training and adopt the practices makes little sense. First establish that there is a basic take-up of the program and that some immediate behaviors are being adopted. Before starting an impact evaluation of a program providing savings accounts, determine whether people will actually open a savings account when offered, and that they subsequently put money into the account. If not, the savings account design should be reconsidered."

2.3 Chapter Summaries

Digital Humanitarians by Patrick Meier (Pages 25-31) This section of reading basically looks at the impact of social media as big data and its applicable uses to disaster relief. They talk about the immense amount of social media postings and content and how to parse through it. Not all of the posts are going to be relevant or timely. However, they do discus an opportunity with using this type of massive data availability in response to humanitarian efforts. It all about identifying what you are looking for.

Smarter Faster Better by Charles Duhig (Chapter 8) This chapter gives a variety of practical real-life examples of how people absorb data. From examples in the school system, which were discussed in the topic overview, to examples about people being able to choose retirement accounts. He uses all of these examples to show that people need to be able to absorb and digest data in an effective way in order to process it and make a decision. He calls the human ability to make these choices and breakdown data as scaffolding and winnowing. When people are able to process data effectively it has huge implications for the impact that it is able to have on business operations and even the lives of students.

Ten Reasons Not to Measure Impact – and What to Do Instead by Mary Kay Gugerty and Dean Karlan This article focuses on organizations innate want to measure their impact and sometimes being blinded by what they are collecting. Governments and funders are increasingly calling on these organizations to demonstrate what it is they are doing and how those dollars are being used. They layout some of the missteps that current organizations fall into and what to do alternatively. For example, they discuss clarifying a theory of change, deciding on what programs to actually evaluate over others and how to effectively integrate data collection into current workstreams.

2.4 Key Take-Aways (for Yellowdig)

2.4.1 Discussion Questions

1. How are you blinded by data in your current organization? Do you feel overwhelmed by any data that you receive? What do you do when you receive this data?

2. Do you feel like you or your organization collect any data that is not used for anything? What is the data point? Do you know why it started being collected?

3. Do you feel that your current data procedures in your organization take away time from your work? Do you find data to be informative or not in your current practice? Why?
4. Disaster affected communities are increasingly becoming "digital communities" that turn to social media to communicate during disasters and to self-organize in response to crises. Do you have your own examples of "digital communities" related to your organization and how does your organization work with them?

2.5 References

- Duhigg, C. (2016). *Smarter faster better: The secrets of being productive. Random House.* **CH8 pp 238-247, 252-267**
- Meier, P. (2015). *Digital humanitarians: how big data is changing the face of humanitarian response. Routledge.* **CH2 the rise of big crisis data pp 25-31**

- Gugerty, M. K., & Karlan, D. (2018). Ten reasons not to measure impact—And what to do instead. Stanf. Soc. Innov. Rev.

Chapter 3

Challenges of Organizational Change

Joseph Lynch Marcela Morales

3.1 The Challenges of Big Data: Organizational Change

3.2 Topic Overview

With over 2.5 Quintillion bytes of data created every day, the greatest challenge to business is how to use this data to improve businesses while making it profitable.

3.3 Chapter Summaries

Desouza, K. C., & Smith, K. L. (2014). Big data for social innovation (Links to an external site.)Links to an external site.. Stanford Social Innovation Review, 2014, 39-43. **Big Data for social innovation**

The term "big data" is used to describe the growing proliferation of data and our increasing ability to make productive use of it. The business community has also been a heavy user of big data. Each month Netflix collects billions of hours of user data to analyze the titles, genres, time spent viewing, and video color schemes to gauge customer preferences to continually update their recommendation algorithms and programming to give the customer the best possible experience. There, a large chasm exists between the potential of data-driven information and its actual use in helping solve social problems. Social problems are often what are called "wicked" problems. Not only are they messier than their technical counterparts, they are also more dynamic and complex because of the number of stakeholders involved and the numerous feedback loops among inter-related components. Numerous government agencies and nonprofits are involved in tackling these problems, with limited cooperation and data sharing among them. Then there are policy and regulatory challenges that need to be faced, such as building data-sharing agreements, ensuring privacy and confidentiality of data, and creating collaboration protocols among various stakeholders tackling the same type of problem. There are multiple dimensions to big data, which are encapsulated in the handy set of seven "V"s that follow. Volume: considers the amount of data generated and collected. Velocity: refers to the speed at which data are analyzed. Variety: indicates the diversity of the types of data that are collected. Viscosity: measures the resistance to flow of data. Variability: measures the unpredictable rate of flow and types. Veracity: measures the biases, noise, abnormality, and reliability in datasets. Volatility: indicates how long data are valid and should be stored. Barriers creating and using big data include the storage of big data in proprietary systems, the regulation on data capture, storage, and curating for accountability, unreliability of data, and the unintended consequence of big data usage. Recommendations: Building global data banks on critical issues Engaging citizens and citizen science (Citizens can also be enlisted to help create and analyze these datasets) Build a cadre of data curators and analysts (We need to equip students and analysts with the necessary skills to curate data so as to create large datasets.)

Making advanced analytics work for you Barton, D., & Court, D. (2012). Making advanced analytics work for you. Harvard business review, 90(10), 78-83.

1. Choose the right date by mastering the environment you already have and exploring surprising sources of information. Be specific about the business problem that needs to be solved or opportunities they hope to exploit. Get the right technology and IT infrastructure to help integrate siloes information (huge issue in government). It will be a continuous flow of information so IT infrastructure that reports in "batches" will not be helpful.
2. Identify the business opportunity and determine how the model can improve performance. Use hypothesis-led modeling to generate faster outcomes and outcomes that are more broadly understood by managers.
3. Make it simple.

Despite big investments in data, many companies have not made it profitable

Despite big investments in data, many companies have not made it profitable: https://www.theregister.co.uk/2017/06/07/go_small_on_big_data/

Mountains of cash keep pouring into the titans of big data despite the world's inability to do much of value with their software. Companies like Cloudera and Hortonworks subsequently arose to help mainstream enterprises put this otherwise complex software to work. It's been a lucrative gig, with each company raising hundreds of millions of dollars and, in turn, generating hundreds of millions of dollars in revenue. What none of them has managed, however, is profit, and that's cause for concern. In other words, the money keeps pouring into the big data companies even as their customers generally struggle to figure out how to turn those investments into meaningful outcomes. These big data vendors then have to spend mountains of cash to convince would-be customers that this time it's different, that this time their investment will return "actionable insights" – that illusive dream of data scientists everywhere. Indeed, IDG Research nails it when it finds that "abundant data by itself solves nothing." Companies need to scale back their ambitions to invest in projects that are more evolutionary than revolutionary in nature, looking to tweak rather than overhaul existing operational practices.

Why Managers hate agile management

Why managers hate agile management:

In the traditional model, there is a top down model where a vision or product is created and this follows a "relay race" through the various managers, line staff, and sales teams. Each level is assigned a different aspect of the vision or product to achieve an end result. As noted in the article "" Why Do Managers Hate Agile?" (Forbes, 2015), the goal of the traditional model was to "have semi-skilled employees…perform repetitive activities competently and efficiently" and coordinating those efforts so that products could be produced in large quantities." In the Agile model, speed to service or product is the goal which conflicts with the traditional model by using the concurrent work of many (including private entities) to enhance the product. The traditional manager is used to having control of the outcome of the vision or product and this just does not work in the Agile model which is causing the "tension." To illustrate the differences between traditional and Agile, the Apple IPhone is a good model. If Apple had designed the IPhone using the traditional model, they would release the IPhone with 40 preset applications that they believed were best using consumer input. Once released, Apple would add applications based on consumer demand which would be vetted by management, created by Apple coders, prioritized for release, tested and placed on the platform. This process would be slow and the variation between Apple IPhone would be non-existent. All IPhone would have the same applications loaded. Consumers could seek out competitors with different variations of applications that met their needs. In the Agile method (which Apple uses), Apple created an IPhone with a number of preset applications, however, they have allowed outside entities to create applications based on the public demand. As of March 2018, there were 2.1 million apps available in the Apple App Store. In July of 2008, there were only 800. (https://www.lifewire.com/how-many-apps-in-app-store-2000252) This Agile approach allows the product to stay relevant to the demands of the consumer rather than the vision of the company. Instead of convincing the consumer to buy their product, Apple is giving the consumer what they want as fast as possible. The Agile method releases or lessens the control that the traditional Manager used to enjoy

for the speed and variation that a wider population can create. The speed to market on consumer demand is far beyond what a traditional model can keep up with. The loss of control and power that the traditional Manager has in their product or service is tough to swallow and that is why Mangers hate agile.

3.4 Key Take-Aways (for Yellowdig)

3.4.1 Discussion Questions

Can data be used to solve social issues deemed "wicked problems" since the infrastructure of non-profit agencies and government do not have the share data in the same way as business.

How can companies and agencies find a way to use big data? Is there a good roadmap to success?

If data is the such a key to success, why are the largest data companies have a problem making profit? Why does the Agile model of business conflict with traditional methods of management?

Why do you think that big data is so important in public sector yet the availability is so limited?

3.5 References

- Desouza, K. C., & Smith, K. L. (2014). Big data for social innovation. Stanford Social Innovation Review, 2014, 39-43.
- Barton, D., & Court, D. (2012). Making advanced analytics work for you. Harvard business review, 90(10), 78-83.
- Despite big investments in data, many companies have not made it profitable: LINK
- Why managers hate agile management: LINK

Chapter 4

Challenges of Big Data

Team 4 - Lindsey Duncan and Justin Stoker

4.1 Challenges of Big Data: Ethics and Privacy

4.2 Topic Overview

According to the Pew Research Center, 95% of all adults own some form of a cell phone and as many as 77% of those are smart phones (Mobile Fact Sheet, 2018). Verizon has recently announced their intent to shutdown their 2G and 3G data streams on December 31, 2019, effectively pushing people to 4G or the emerging 5G technology for mobile data (Morris, 2018). Smart phones are just one of the many tools that collect and report anonymous data based upon its user's location, social activities, financial transactions, browsing history, and information searches. This data is collected and passed through algorithms such as Apple's Siri, Google Maps and Google AdWords to help predict a user's schedule, interests, traffic patterns and delays, shopping habits, and more. Big Data is being collected all the time and often without the knowledge of the individual contributors of that data. This section discusses the Challenges of Big Data: Ethics and Privacy.

Whether it is the GPS on a cell phone, traffic cameras, license plate readers, macroscopic infrared imaging, or each other, we are becoming increasingly aware of the amount of information that is being collected about our individual lives. To quote the old English saying, "just because we can, doesn't mean we should." This is an example where technology is outpacing policy makers – where policy makers are often just as oblivious to the what's happening as everyone else.

It is important to discuss the ethics and privacy concerns that come about from the collection of all our individual data. While the vast majority of Americans have no concern and claim to have nothing to hide about their life's data, many are worried about the eventual public access and public use of that data. "There are 3 Big Data concerns that should keep people up at night: Data Privacy, Data Security and Data Discrimination" (Marr, 2018). Questions that people are likely to ask include:

- Is the information truly anonymous or can it be tracked back to me?
- Can my information be used against me?
- Is my data going to be used for corporate enrichment or political battles?
- Would my data contribute to racial or other discriminatory profiling by government or law enforcement?

While the data collection, in general, benefits everyone by helping with traffic or travel time prediction, allowing your phone to store hours of your favorite stores, sports scores or news from your favorite teams, people have begun to express concern about the use or the public exposure of their personal data for reasons

Figure 4.1: Image

not in the public good. The book Social Physics establishes the social nature of individuals and groups and makes the point how information is passes through those social networks. Currently, social media is used extensively for everything from sharing personal updates, to business marketing, to news and press releases. Even in a book that argues the virtues of the sharing of ideas through social networks, Social Physics acknowledges, "Maintaining protection of personal privacy and freedom is critical to the success of any society" (Pentland, 2015 p. 17).

4.3 Chapter Summaries

4.3.1 Data that turned the world upside down

This article is about an individual researcher named Michal Kosinski and a Big Data company called Cambridge Analytica. Kosinski's research in the field of psychometrics (measuring psychological traits) led to the development of algorithms associating a person's Facebook likes to the OCEAN (openness, conscientiousness, extroverted, agreeableness, neuroticism) personality instrument. Kosinski found a person's digital footprint to be extremely predictive of not only personality, but also other preferences. Though Kosinski was positive about the uses of his research, he worried about the potential ramifications. "What would happen, wondered Kosinski, if someone abused his people search engine to manipulate people? He began to add warnings to most of his scientific work. His approach, he warned, 'could pose a threat to an individual's well-being, freedom, or even life' " (Grassegger, 2017).

Kosinski was concerned when he discovered the work of Cambridge Analytica which has been associated with President Trump's election campaign and Great Britain's exit from the European Union (Brexit). Cambridge Analytica was claimed to have profiled all adults in the U.S., using the data for very targeted electronic marketing during the 2016 presidential election. Xx Nix, spokesperson for Cambridge Analytica's marketing strategy, "Cambridge Analytica buys personal data from a range of different sources, like land registries, automotive data, shopping data, bonus cards, club memberships, what magazines you read, what churches you attend… in the U.S. almost all personal data is for sale"(Grassegger, 2017). The company then matches this data and aligns with voter information and the personality profile to identify the target market.

4.3.2 Eye in the Sky Podcast

Theme is the availability of the data can do a lot of good things, like solve murders, property crimes, etc. but on the other hand there are those that call it a "grotesque violation of privacy" (Eye in the Sky). At what point can public data be taken by a person to track down a cheating spouse? When can it be abused, where do the lines exist?

The Eye in the Sky Podcast details the story of Ross McNutt, a former military officer that utilized surveillance equipment that continuously takes pictures every second over the Town of Fallujah in Afghanistan, to be able to track those that would plant roadside improvised explosive devices (IED). The surveillance equipment would be attached to the underside of an aircraft flying well above the town so that people were nothing more than pixels on a screen. When it was determined that an IED was planted, it was possible to go back and track the person that set the device forward to where they hid or met up with others. The method was effective in tracking down those that would plant the devices. McNutt later separated from the military and established a private company called Persistent Surveillance Systems that would do the same for more domestic towns and cities.

In one example, McNutt demonstrated the use of the technology to track crime in Juarez, Mexico and ultimately pitched the technology in his hometown of Dayton, Ohio. Despite reaching out to the American Civil Liberties Union (ACLU) and local residents, a vocal minority was able to shut down the proposal over concerns for individual privacy.

4.3.3 Weapons of Math Destruction – Intro pages 1-13

The introduction to the Weapons of Math Destruction text recognizes how the success of Big Data has actually been problematic. Big Data has been described as more objective than the application of human opinion in decision making. However, Big Data has also served to reinforce human bias when it is programmed into the systems used to collect and to analyze data. Further, O'Neil points out how difficult it is to challenge the verdict of Big Data because the algorithms and coding are a closely guarded proprietary secret or are so complex they are difficult to decipher. "Like gods, these mathematical models were opaque, their workings invisible to all but the highest priests in their domain,: mathematicians and computer scientists" (O'Neil, 2017, p 3).

The text highlights the problematic use of data, specifically in the Washington D.C. schools to evaluate teachers. The schools were using data to evaluate the success of teachers. Those who scored in the lowest percentiles were separated from employment. This shows how problems occur with data and algorithms when they are used rather as doctrine rather than suggestions or indicators. It highlights the story of Sarah Wysocki who scored well one year and then was fired the next. People couldn't explain the algorithm and failed to consider suggestive information that prior year test results on the students may have been altered by their teachers. Recall the disincentives that occur when what gets measured gets managed.

The underlying purpose of this text is that there are situations where Big Data is being misused and it is done by people that don't understand what they are doing. The author proudly proclaims at the end of the Introduction, "Big Data has plenty of evangelists, but I'm not one of them. This book will focus sharply in the other direction, on the damage inflicted by WMDs *weapons of math destruction* and the injustice they perpetuate. Welcome to the dark side of Big Data." (O'Neil, 2016, p.13)

4.4 Key Take-Aways (for Yellowdig)

4.4.0.1 Ethics

Users of Big Data should be thoughtful in their approach. As Cathy O'Neil suggest in Weapons of Math Destruction data can be used for harm even when intended for good. Programmers and administrators may inadvertently program personal biases into analytical algorithms. They should be conscientious in their application of the data, ensuring that it is not the only means for evaluating success. Success is measured as

the selection of a candidate for a job, the termination of an employee, the identification of a personal match for dating, the funding of a program, etc.

Data systems should also be subject to monitoring, evaluation, and adjustment. If the means for analyzing the data is flawed and hidden behind a proprietary veil, then the system should be opened up to scrutiny. If you cannot defend it, you probably shouldn't be doing it.

4.4.0.2 Privacy

The United States needs something similar to the European Union General Data Protection Regulation (EUGDPR) to establish policies with teeth to protect data from breeches and preserve privacy of its citizens. Better clarity is needed in terms of notification of how businesses are using data. Notifying customers that video recording is in process or a phone call is being recorded doesn't necessarily mean that people are consenting for their images to be used and linked to other forms of data collection to track personal habits or trends. As Hannes Grassegger and Michael Krogerus note in The Data That Turned The World Upside Down, "The company [Cambridge Analytica] is incorporated in the US, where laws regarding the release of personal data are more lax than in European Union countries. Whereas European privacy laws require a person to 'opt in' to a release of data, those in the US permit data to be released unless a user 'opts out.' "

The United States needs a data protection standard that encourages respect of personal data. Penalties for violating data security according to the EUGDPR can be as much as 4% of the annual revenue or €20 million whichever is greater (GDPR, 2019). This penalty is sizable enough to take data security seriously.

4.4.1 Discussion Questions

- All organizations collect and store data in some form or another, whether it is for billing, research, marketing, or a host of other reasons. As a manager, am I doing what's necessary to protect the data that I have from security and privacy breaches?
- Laws currently exist to provide basic security for data protection. Should I be doing more, beyond what is necessary, to protect the data that I have access to?
- Often data can be collected and then processed through algorithms to provide objective performance standards. However data processing is only as good as the programmers that prepared the algorithm. Am I considering the Human element when drawing conclusions from the data I have?
- Do I understand the algorithms or computational methods used to interpret the data? Are they accurate? Are they constantly being improved to consider additional factors/understandings?

4.5 References

- Eye in the sky: PODCAST
- Eye in the sky: Washington Post
- European Union: General Data Protection Regulation (2019). Retrieved January 17, 2019 from: LINK
- Feimberg, H. (2016, January 7). FTC Warns Against Use and Misuse of Big Data Analytics. Retrieved January 11, 2019 from LINK
- Grassegger, H., & Krogerus, M. (2017, January 28). The Data That Turned the World Upside Down. Retrieved January 15, 2019, from LINK
- Marr, B. (2017, June 15). 3 Massive Big Data Problems Everyone Should Know About. Retrieved January 17, 2019 from LINK
- "Mobile Fact Sheet." Pew Research Center, 5 Feb. 2018, LINK
- Morris, J. (2018, July 2). Verizon 2G and 3G Sunset Starts. Retrieved January 17, 2019 from LINK

- O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books. Introduction pp 1-13
- Pentland, A. (2015). Social Physics. Penguin Books. p 17
- Header image "cybersecurity" By Titima Ongkantong/Shutterstock.com

Chapter 5

Collecting Group Data

Collecting Group Data (Team 5)

5.1 Topic Overview

In Chapter 3 of Reality Mining and Chapter 5 of Social Physics we see examples of how we can learn from small groups. In Social Physics, Pentland worked with employers and their employees to conduct short-term experiments in order to improve the productivity of the workplace. He and his team were able to draw out key learnings from the data collected and discovered important links between behavior and social learning opportunities. **Their data showed that the pattern of ideas flowing is central to driving productivity. To take this a step further, their research showed that face to face engagement and exploration were the largest factors influencing productivity and creative output.** (Pentland, 2014) In other words, the more time people spending time interacting with each other, the more ideas could be generated, discussed, and their colleagues would be in the loop, thus resulting in higher efficiency and productivity.

In Reality Mining, Eagle and Greene discuss the need to expand collecting data from small groups outside the world of academia and opportunities to scale it to address wider community needs using sensors and technology. They acknowledge that small group data collection is often the most difficult type of information to gather because participants are often wary of sharing information that is trackable back to the individual. In small group settings, there is heightened concern about anonymity. **The authors' central belief is that incentives should be used more by those wanting to collect group data and that given the right incentive, data could be shared freely.** (Eagle and Greene, 2014)

DISCUSSION

There is no doubt about the benefit that can come from capturing small group data as these two texts illustrate. However, how realistic is it that nonprofits or government agencies can collect and use such data? Despite the relative ease of collecting the data now, compared to a decade ago, organizations must have the ability to capture the right data that is most meaningful to their work and the expertise to use it to affect the delivery of their programs or services. Further, organizations must have a strategy to use the data and processes to interact and learn from the data. (Barton and Court, 2012) (Duhigg, 2016) Using big data can feel overwhelming for these reasons. Additionally, agencies must commit the resources needed to extract and interpret the date and promote the validity of the investment. Far too often this type of research is considered nonessential and bypassed due to funding or staffing shortages, or simply because we are all too busy. If data and statistics are to be the foundation for decision making, then it must be treated as an essential tool, not a secondary accessory.

Perhaps a better and more manageable role for nonprofits and government agencies is collecting targeted data on a few key indicators and performing micro tests of focused hypotheses. This would allow them to

more easily track, manage and use the data to answer a key research question. For example, an advocacy organization could test what subject lines drive click-throughs to increase the number of advocates taking action via email. A government agency like the Motor Vehicle Division could test alternative strategies to find out what best reduces customer wait times. Some ideas might include creating an express lane, having staff specialize in a particular function and having them focus on those tasks, having staff keep track of frequently asked questions and communicating the answers, thus cutting down the amount of time needed when being helped. It seems like the idea of using big data is good in theory when you have the ability and technology to use it well. For both nonprofit and government settings, using the data in a more strategic, targeted approach could be more manageable and practical.

5.2 Chapter Summaries

Reality Mining: Using Big Data to Engineer a Better World Chapter 3: Gathering Data from Small Heterogeneous Groups

This chapter explores the potential of collecting data from small groups of people. It discusses the challenges, costs, and benefits of a number of different types of technologies that collect data from small groups. Detailed examples from multiple fields are given, including a look at smart conference badges that track movement and engagement of conference attendees, companies mining data from their employee activity to create a more productive work environment, neighborhoods collecting data to help solve social problems such as reducing pollution and determining the best bike paths, and audio surveillance systems that enable media companies to have a real-time look at how people consume multiple types of media. The authors cite concerns throughout the chapter about maintaining user privacy, legality of data collection, the expense of collecting data, and having the technology available to gather the data efficiently. Despite these concerns, the authors believe that the biggest challenge of group data collection is finding the right incentives. They argue that if the benefits of the data collection outweigh the costs, individuals would be more willing to share their data and companies who want that data should compensate people appropriately.

Social Physics: How Social Networks Can Make Us Smarter Chapter 5: Collective Intelligence-How Patterns of Interaction translate into Collective Intelligence

In Chapter 5 of *Social Physics,* researchers detail their findings regarding networks and exploration. Their findings indicate that the pattern of idea flow was the single most indicative element of a successful group. When groups featured characteristics such as equal participation, turn taking, and high engagement with other members, the group was likely to have high performance. Researchers summarized that when these group conditions exist, it will likely result in a large volume of ideas being generated and filtered in a group process with a majority consensus being gained. When a network functions in this capacity, research indicates they have a higher efficiency and yield a more creative output. These groups had a higher level of collective intelligence than group that did not function in this format.

In order to develop the volume of ideas, the members of the group must actively engage in 'exploration' whereby they venture outside of the network to gather unique data and return to the network to add it to the collective stockpile of information. Once this "idea dump" has been completed, the network reconvenes as a unit to filter and evaluate it. Once completed, any ideas or information that are deemed valuable or actionable by the majority are put into practice.

5.3 Key Take-Aways (for Yellowdig)

Collecting Group Data

5.3.1 Discussion Questions

In one case, a company subsidized the cell phone equipment and service for people participating in a study that tracked their media consumption. Their phones were enabled to listen to everything that user experienced and took short snippets of sound and matched those against a database to see what movies, songs,

TV shows, or other media a person may be listening or watching. For you, if a company offered to pay your cell phone bill, would you give them an all-access pass to listen to your life? Describe your answer.

Follow Up: This was published before the mass introduction of Amazon's popular "Alexa" gadget. Do the recent incidents around Alexa's unauthorized recordings and subsequent invasion of privacy influence your answer? In regards to the previously mentioned cell phone research, do assurance that conversations and other non-research related recordings are to be deleted and discarded appease you?

In another case, the authors talked about companies monitoring their employee's activities, including phone calls, activity on their computers, emails, recording of phone calls, movement around the office or text messages. Companies do this to better understand the social networks of their employees, connect staff with similar needs or interests, detect potential fraud, track productivity, and other things. Would you be comfortable working for an employer who is watching your every move? What benefits or incentives would your employer have to provide to make you comfortable with this?

In one case study, the author noted the productivity of a team increased when the members of the team were given a collective break, as opposed to staggered break times, allowing them to have more face-to-face interactions. The conclusion was that more engagement translated to better efficiency as a team. Do you find this to be true in your own work setting? Is there a point of too much "collectiveness"?

Does the research showing high functioning networks are a result of equality, turn taking, and engagement surprise you? Is this something we all learned in kindergarten but have to relearn in adulthood?

5.4 References

- Barton, D., & Court, D. (2012). Making advanced analytics work for you. Harvard Business Review, 90(10), 78-83.
- Duhigg, C. (2016) *Smarter faster better: The secrets of being productive in life and business.* Random House. New York, NY.
- Eagle, N., & Greene, K. (2014). *Reality mining: Using big data to engineer a better world.* MIT Press. CH3 gathering group data

- Pentland, A. (2015). *Social Physics: How social networks can make us smarter.* Penguin. CH5 observing people in organizations

Chapter 6

Processing Satellite Data

Team 6 - Tommy Kolwicz and Dennis S Stockwell

6.1 Topic Overview

The three references take us through the current best practices for collecting and analyzing different types of imagery and how the public can be used to help process large volumes of data. It is not humanly possible for imagery experts and scientists to analyze the massive amount of satellite, and soon UAV, imagery collected on a daily basis. Crowd sourcing the detection and flagging of whatever the item of interest happens to be is the most efficient method in place today. Incredibly, no matter what the topic of interest is, the public wants to help. Humanitarian, scientific, wildlife preservation, archeological, all have found a following.

The only things that have been shown to do a better job than humans in analyzing imagery are machines. The second reading takes us to the role AI and machine learning is playing in the processing of data and how we are teaching AI, through crowd sourced training sets, to process data so that interpretation can be done at a larger scale and faster. Time is the element that we are continually attempting to harness. The last reading transitions us to consider the implications of massive campaigns of data collection. Clearly this is a technological capability that we currently possess. The article poses the question, “Should we?”. The article also ponders whether people truly understand how the data will be used and what privacy we may be giving up for the proposed positive outcomes associated with the data collection.

6.2 Chapter Summaries

6.2.1 Crowd Computing Satellite and Aerial Imagery

Digital Humanitarians by Patrick Meier (Chapter 4)

Big Data Fusion, or using multiple different sensor systems and web tools in an integrated manner to produce a coherent operational picture, is the future of crowd sourced Digital Humanitarian work; this is referred to as microtasking. Recent history has shown that there are a number of sensor systems that can be used to collect needed imagery to do the necessary analysis required in a crisis.

On one end of the spectrum, you have extensive, and expensive, satellite imagery. In 2014, satellite imagery was provided to digital humanitarians online to help in the search and rescue operations of Malaysia Airlines flight 370. In just four days, 8 million volunteers had combed over 400,000 square miles of ocean and land. Crowd sourced imagery analysis has helped try to find the grave of Genghis Khan, search for Steve Fossett’s missing airplane, identify the number of shelters in Somalia, and tag galaxies for astrophysicists.

For much cheaper, digital humanitarians can buy Unmanned Aerial Vehicles (UAVs). These are becoming less expensive by the day and can be used to collect images of large areas or many images of focused areas.

Either way, UAVs are unaffected by clouds or other atmospherics that can cause trouble for satellite images. On the very inexpensive end of the spectrum are kites and balloons. In 2010, balloons and kites that could produce pseudosatellite images were used to document the BP oil spill and the magnitude of destruction.

No matter the method of collection, the crowd sourcing of analysis is what can make this methodology so powerful. Now, some organizations' are even experimenting with crowdsourcing the crowdsourcing. In this manner, the most accurate and efficient digital humanitarians are assigned to review the work of others. The most important images are then sent to professional analysts to review and eventually disseminate to the stakeholders on the ground.

Just as texts and images were the medium used to create crisis maps in aiding emergency response during natural disasters and search and rescue operations, aerial selfies, using small personal UAVs, are already beginning to join the world of crisis mapping. Now aerial views of points of interest to digital humanitarians can be added to a crisis map and combined with tweets and pictures from Image Clickers. The result…Big Data Fusion.

6.2.2 Artificial Intelligence in the Sky

Digital Humanitarians by Patrick Meier (Chapter 6)

When it comes to Satellite imagery the sheer quantity of images is now outpacing even the ability of microtasking by crowdsourcing. "Microtasking alone can't actually keep up with 1.5 million square miles of new satellite imagery produced every day – a figure that will increase substantially within just a few years." Enter the machines. That's right, just like something out of a 80s movie with Arnold Schwarzenegger, one solution being researched is machine learning and artificial intelligence. Machine learning has shown a lot of capability in the organizing and prioritizing of pictures from disaster response using many of the same techniques as with letters and words. Automated imagery analysis has shown to be extremely effective. In Haiti, the European Commission's Joint Research Center (JRC) showed a 92% accuracy rate in automatically identifying rubble-filled areas. Unfortunately, "image-based machine-learning classifiers do not 'port' well." In other words, one classification system for one area does not necessarily work for classification in another disaster response in another area.

Satellite imagery has also been shown to be useful in following refugee migration, estimating building size, and other aspects of disaster relief using such imagery characteristics as shadows. The hard part is getting satellite imagery inside the 24-hour threshold window for being able to have an impact.

Organizations like Galaxy Zoo have demonstrated that using the crowdsourced human-classified imagery to "train" the machines, they have been able to achieve classifications accuracy greater than 90%. Eventually, at least for Galaxy Zoo, the machines began to outperform the humans and put the volunteers on the sidelines. However, volunteer microtaskers are still needed to create new machine training sets.

UAV imagery is about to become as much of a "big data problem" as satellite imagery. But creation of training sets for machine learning to make sense of UAV images is already underway via crowd sourcing and microtasking. The University of Maryland's Institute for Advanced Computer Studies has already developed specialized software to automatically identify Rhino poachers and even the type of weapon they are carrying via pattern-recognition algorithms. These techniques could be easily applied to humanitarian disasters.

6.2.3 Eye in the Sky

Radiolab Podcast https://www.wnycstudios.org/story/eye-sky

Radiolab Podcast: Eye in the Sky was sponsored by Radio Lab a Radio Program on WYNC a public radio station in New York City. The podcast included Manoush Zomorodi and Alex Goldmark from the podcast "Note to Self" and Mr. Ross McNutt the CEO of Persistent Surveillance systems. Mr. Ross McNutt had been a 20 year Airforce Veteran that was part of an Air Force Team looking at how to reduce the deaths as a result of IED attacks in IRAQ. McNutts team developed a system (Project Angel Fire) that photographed a large area and took photos at 1 per second creating a snapshot in time that can be reviewed later. Once an incident has occurred or has been reported, those photos can be looked at in reverse time. Starting from the

incident in 1-second intervals in an effort to discover when the device was planted. If the vehicle responsible for the IED can be identified, the technology can also help answer multiple other questions. Where did the car come from? Did it pickup anyone up? Where did it start? The photographs can also be played forward post-event to see where the vehicle went after planting the device. As a result of the review of the photographs Security Forces can be sent to the most likely location of the terror cell.

The Podcast looks at the Project Angel Fire concept and examines the civilian impact of this technology, primarily law enforcement applications and the tradeoff of security and privacy; could this technology be used in the future by divorce lawyers, real estate people to monitor property, etc? The podcast recognized the potential for its significant ability to support Law Enforcement. A prime example is the city of Juarez Mexico after a killing of a Police Officer. The technology was used to identify the vehicles involved, who they came in contact with, and where they went. That data was used to map the locations these vehicles had in common so that law enforcement could make arrests at these mapped locations. This ultimately took down a drug cell that had been linked to multiple murders.

The podcast also discusses an incident involving the city of Dayton, Ohio. The city attempted to implement this program and had an open forum to allow residents to discuss their concerns. While many supported the program, the ones that did not were the loudest and prevented implementation. It did recognize that the program might be implemented in the future if the city can develop a better communication strategy, explaining the program and its limitations to reassure the community on what the program will be used for.

The Podcast participants clearly recognize the potential positive impacts this technology brings to the table, but raise questions about our government and data. Can we trust our government to use the data as they say they will? Just because we can, should we? My opinion is this technology will eventually be utilized in those cites that can afford it, it is my personal opinion programs like this are always better in the open under legal oversight rather than in the shadows as intelligence programs, where they may lack the oversight required to ensure they don't exceed their guidance.

6.3 Key Take-Aways (for Yellowdig)

6.3.1 Discussion Questions

1. Chapter 4- Have you heard the term micro tasking or crowd sourcing? Have you been an internet volunteer or a digital humanitarian? Maybe you were and you didn't even know it.
2. Chapter 6- What do you think about the development of AI (artificial Intelligence) to help humans interpret big data? Do you trust AI to make decisions for us?

3. Eye in the sky - Because we can, should we? Do we trust those with the data to use it as it was intended? What are the privacy concerns and can we protect it once it enters the public domain? Although the information is being used for one purpose now will there be future purposes for the information that we are collecting which we haven't even thought of yet.

6.4 References

- Meier, P. (2015). Digital humanitarians: how big data is changing the face of humanitarian response. Routledge. CH4 crowd computing satellite and aerial images
- Meier, P. (2015). Digital humanitarians: how big data is changing the face of humanitarian response. Routledge. CH6 artificial intelligence in the sky
- Eye in the Sky Radiolab Podcast

Chapter 7

Using Administrative Data

(Team 7)

7.1 Overview

The following summaries provide an overview of the multiple uses of administrative data and topics reagrding privacy and transparency.

7.2 Reality Mining: Using big data to engineer a better world. Chapter 7 - Taking the Pulse of a Nation: Census, Mobile Phones, and Internet Giants

National scale researchers & entrepreneurs get access to data sources from national censuses data, call data records or call detail records (CDRs), major internet companies (Google, Facebook, Twitter), and banks.

National Census Data – National census data is the easiest to obtain and it's publicly available. World Bank conducts international surveys and compiles census data. Google has integrated this data in a visualization tool in its search results. The U.S. government makes U.S. consensus data resources available through the American Fact Finder. (P.113) In 2009 the government began data.gov, where many interactive data sets have their own Application Programming Interface (APIs) to integrate maps and charts into web applications. (P.114) The international World Bank accumulates data from more than 200 countries with over 7,000 indicators (such as GDP and gas prices) to produce a data catalog on land, literacy rate, health, climate change, and much more. Although World Bank is one of the "richest collections" of worldly data, some inconsistencies exist due to a nation's timing and reporting practices. Data can also be accessed by indicators, operations, and financial data. Google uses World Bank data in Google Public Data Explorer, which allows one to "slice and dice data" from World Bank. (P. 115)

Call Data Records (CDRs) – CDRs were historically only used for billing purposes, but since 2005, it was determined how valuable this data can be used for modeling human mobility. Very few researchers and entrepreneurs have access to CDRs and must abide to legal stipulations, such as "no proprietary or personal information will ever be made public." (P.112) Some researchers must agree to show how they can assist with future predictions related to the mobile carrier to obtain access. Mobile operators must be careful about making the data available because, even with anonymized data, it's still possible to identify an individual by cross-referencing data. With just a DOB, gender, and zip code, 63-87% of the U.S. population can be identified. (P.117) Sprint researchers had a data set of 30 billion calls made by 25 million mobile users in the U.S. Researchers felt that by giving enough time to figure out the user's top locations (home and work) and combined it with census data, the researchers could identify almost all the users. Due to this conclusion,

researchers suggest that data should be "coarse in time and in space" (P.117) In other words, data should be collected in one day versus over a month period, and also collected from a larger geographical area, not just from one tower. AT&T operates a Work & Home Extracted Regions (WHERE) project in which a synthetic model of mobile users has been developed for a particular City. This concept has been fairly successful in maintaining individual privacy in NYC & LA, but is still in the research phase. WHERE can become an important tool if it can be applied outside of city limits. Currently, AT&T is working on projects such as AirGig, 5G, and Acumos to bring superfast internet to suburban and rural areas over power lines, combat biases in advertising, and help accelerate the adoptions of various technologies.

Internet Companies - Google services include Gmail, Google+ social network, YouTube, and the Chrome web browser. Google keeps a log of searches and URLs if the "Instant" feature is enabled on Chrome. Google also "logs YouTube videos watched, activity on Google+, and the text of emails sent through Gmail." (P.119) Google's Adsense program runs advertisements based on information gathered through key words and phrases.

Facebook offers a similar advertising network like Google's Adsense. Both programs operate on "clicks" and "impressions" based on key words. Key words are set by the advertiser and "pulled" from FB user's profiles. An advertiser then receives metrics of general statistics like the number of emails or profiles the ad appears in. This information is "mined" to get the overall feel of a product across demographics. FB also has an API which allows a programmer to have access to phone numbers, contact lists, etc., if the FB user allows this in their privacy settings. Zynga, an FB application, collects habits and behaviors of people who play their games. The most popular game, FarmVille, has been questioned for its ethnicity due to requirement of participants to trade information about their friends, likes, desires, and consumption habits to participate in the game.

Twitter is different than Google and FB because most of the user's data is publically available and can be used to "mine for the sentiment of a nation" based on the location of a tweet. (P. 121) However, it's a challenge to determine the signal from the "noise" from fragmented conversations, links, hashtags, and abbreviations.

Banking Transaction – Banking transactions are the most difficult to access. Banks use data analytics for purposes of fraud, to predict when someone might switch banks, or to adjust interest rates depending on spending habits or risk. (P. 121) Since banking transactions are tied to a location and a specific action, researchers can obtain a "fine-grained picture of a person's economic behavior." (P. 121) In 2008, Bank of America partnered with MIT using a using a sample of 10,000 customers, with various metrics, over a 3-year period. Recently, Tresata, an analytics software company in Charlottesville North Carolina, secured a $50-million-dollar growth capital investment from GCP Capital Partners. Mint, owned by Intuit, is a free, web-based personal financial management service for the US and Canada. Mint tracks on-line spending and uses the data to advertise financial products to customers based on their spending habits.

7.3 The NYPD Was Systematically Ticketing Legally Parked Cars for Millions of Dollars a Year- Open Data Just Put an End to It,

This blog shows how Open Data can be used by citizens to help themselves and government.

In late 2008, NYC passed legislation allowing drivers to park in front of a sidewalk pedestrian ramp as long as it's not connected to a sidewalk. A citizen continued to received parking tickets for parking in a "legal" spot. The tickets were always dismissed after a time consuming process.

This citizen used NYC's Open Data Portal to determine common parking spots in NYC where cars were ticketed for blocking pedestrian ramps. As a sample, 30 random spots were chosen that received more than 5 tickets in the last 2.5 years, and using google maps it was confirmed all spots were legal. Based on the data, there were 1,966 spots that were generating about $1.7 million annually. Note: It's possible some of these spots were illegal, but the majority are legal based on the sample results.

The citizen posted a map of 1,000 pedestrian ramp parking spots with the number of tickets each spot had received to date, for others to view on-line. The citizen determined Brooklyn's 70th precinct had the most cars wrongly ticketed generating over $107k fines per year, with the 77th precinct bringing in over $101k per year. Next, the citizen reached out to the NYDP via the Mayor's Office of Analytics and Manhattan Borough President Gale Brewer's Office and got feedback, which stated patrol officers were unfamiliar with the rule change. The officers have been trained and are digitally monitoring tickets to limit erroneous ticketing from happening in the future.

7.4 Open Data Reveals $791 Million Error in Newly Adopted NYC Budget

In 2016, New York City launched a searchable database of the municipal budget. Although it only consisted of the current year budget, it allowed a breakdown of specific budget unit per department. The transparency lead to the realization that NYC had a adopted a budget with a nearly $800 million error.

The new searchable database categorized expenses all the way down to the "Object Code" Name. These codes included elements including full time position costs, overtime and postage. All of these codes are assigned to an individual agency, making it easy to track expenses. Prior to the database, in order to analyze and assess the NYC budget, someone would have to sort through hundreds of pages of PDFs.

While sorting through the data, a question arose about the largest expenses within the NYPD. While analyzing the largest budget codes, it became clear that "Protection of Foreign Missions" was the largest expense. By way of comparison, more money was going to be spent on protecting foreign missions than school safety, transit, housing and narcotics combined. The analysis show that this budget category along amounted to about 1% of NYC's budget and 15% of the NYPD's entire budget.

By looking at prior year budget it was obvious that this $791 million line item was an error. But how on earth does a mistake like this happen? How does a typo like this make it through the entire budget process? Human error is understandable, and acceptable, but examples like this make the case for why open data and transparency is necessary to hold public entities accountable.

7.5 Payer or Prayer- A Look at NYC's $650 Million Property Tax Breaks Related to Religion

A team downloaded every individual tax bill via PDF to create a map that shows the locations of religious property tax exemptions as well as the type of exemption, amount and the name of the religious institution. The exemption types included: "House of Worship," "Religious Dormitory," "Clergy," "Parsonage," "Religious Mission," "Bible," and "Salvation Army."

A first look at the data revealed that the sum all exemptions totaled $12.9 billion of a total tax due of about $21.6 billion. But many of these tax abatements go to public agencies such parks, the department of education as well as the port authority. But listed as number six on the list is houses of worship, totaling $650 million a year approximately 1% of the entire city's budget. These exemptions also add up to $76 dollars per NYC resident.

Here is some of the highlights of the analysis: *The neighborhoods receiving the largest amount of exemptions per resident are some of our wealthiest* Communities with Large Jewish Populations Have the Most Religious Schools per Capita *Clergy May be Priced out of Manhattan, but there are Plenty Living in Ocean Parkway South* South Jamaica Has the Most Houses of Worship per Capita *There is a "Bible" exemption taken by just two properties in NYC

7.6 8 principles of open data

In 2007, number of open data advocates met and established principles of open government data. Below is a brief recap of the principles: 1. Complete: All data considered public is made available. Public data is defined as "data that is not subject to valid privacy, security or privilege limitations." 2. Primary: All data is produced from the original sources 3. Timely: Data is given as quickly as possible 4. Accessible: Data should be to a wide variety of users for the multiple of purposes. This includes: accessible via the internet, consideration for the disabled, and follow current industry standard protocols and formats. 5. Machine-processable: Data is provided in a format that allows for automated processing which require proper encoding. 6. Non-discriminatory: Data is available to all persons, including those who request it anonymously. 7. Non-proprietary: Data is available in a format that no one entity has exclusivity, therefore not in a proprietary format. 8. License-free: Often government data is a mixture personal data, copyrighted information or others forms of non-open data. Therefore, the experts determine that open data should not subject to any copyright, patent, trademark or trade secret regulation, but "reasonable privacy, security and privilege restrictions may be allowed."

There are seven additional principles that the group could have considered but did not. Those include: 1. Open data should be online and free 2. The open data should be permanent meaning the information should be provided in a stable location and accessible for as long as possible. 3. The data should be trusted which includes attestation or digital signatures or publication dates verifying authenticity; 4. A presumption of openness, the government will make public information available proactively with little to no barriers for use and access. 5. Documented, it is as important for users to know the data is current as for the data itself to be current, so users can assess accuracy. 6. Safe to Open Data content should be free of malware, viruses, worms, etc. 7. Designed with Public Input so appropriate information technology are utilized for public use and dissemination.

7.7 Key Take-aways:

7.7.1 Summary #1:

National scale researchers & entrepreneurs get access to data sources from: National censuses - World Bank American Fact Finder Data.gov Call data records or call detail records (CDRs) - Modeling human mobility Caution due to cross referencing Coarse in time & in space AT&T WHERE project – synthetic users Major internet companies - Google - Gmail, Google+ social network, YouTube, Chrome web browser/Adsense Facebook – advertising, Zinga, Farmville Twitter – "mine for the sentiment of a nation" based on the location of a tweet Banking Transactions – Most difficult to access Transactions are tied to a location & a specific action "Fine-grained picture of a person's economic behavior"

7.7.2 Summary #2:

NYC passed legislation allowing drivers to park in front of a sidewalk pedestrian ramp as long as it's not connected to a sidewalk

Patrol Officers continue to ticket vehicles parked in "legal" spots

Citizen was able to use NYC Open Portal and Google Maps to determine millions of dollars being fined to citizens parking in legal spots.

Citizen made appropriate notifications to NYC and NYPD provided training to officers & established a digital monitoring system to ensure citizens weren't being "legally" ticketed.

7.7.3 Summary #3:

Open Data of the NYC budget for one fiscal year revealed a nearly $800 million typo in the approved budget. The new searchable database categorized expenses all the way down to the "Object Code" Name. These

codes included elements including full time position costs, overtime and postage. All of these codes are assigned to an individual agency, making it easy to track expenses. It also made it easy to find errors.

7.7.4 Summary #4:

A team downloaded every individual tax bill via PDF to create a map that shows the locations of religious property tax exemptions as well as the type of exemption, amount and the name of the religious institution. The exemption types included: "House of Worship," "Religious Dormitory," "Clergy," "Parsonage," "Religious Mission," "Bible," and "Salvation Army." This allowed the group to create an analysis of every religiously affiliated property tax exemption in the city based on location and exemption type.

7.7.5 Summary #5:

According the leading experts, Open Data should be complete, primary, timely, accessible, machine processable, non-discriminatory, non-proprietary, and license-free.

7.8 Discussion Questions:

1. Do you think Zinga is being ethical (through Google) allowing FarmVille to gather information regardi
2. Are you comfortable with Tresata having access to your financial personal management, as well as an i
3. Do you like personalized advertisements regarding on-line spending based on your personal spending ti
4. Do you think a City allowing citizen's access to an Open Portal is good for the citizens and/or a Ci
5. NYC will lose a substantial dollar amount in parking fines, now that a citizen took the initiate, usi
6. Do you think City's which aren't transparent, are purposefully doing so to "hide" the facts?
7. How concerning is it that there was such a large error in the approved budget?
8. Have you encountered a situation where an error was exposed by a third party who analyzed your organi
9. Should all budget line items be available for public scrutiny? What potential challenges might this p
10. From a taxpayer perspective, how valuable is this specific property tax data? How should policy mal
11. Have you used similar mapping strategies to better understand where resources are allocated in your
12. Are there privacy concerns about visually mapping out specific taxpayer information including addres
13. In your current role/organization, how many of these 8 principles do you utilize when providing oper
14. For those in the public sector, what are some of the challenges government entities face when attemp

7.9 Video:

7.10 References:

- Eagle, N., & Greene, K. (2014). Reality mining: Using big data to engineer a better world. MIT Press. CH7 mobile and internet data
- Project AirGig Gets Closer to Initial Commercial Deployment, Dallas, Sept 10, 2018 https://about.att.com/story/project_airgig_trials_georgia.html
- A History of Firsts: AT&T Labs is Still Creating the Future 35 Years After the First Cellular Service, Andre Fuetsch, October 11, 2018 https://about.att.com/newsroom/2018/35th_mobile_anniversary_call.html
- Zynga's FarmVille, social games, and the ethics of big data mining, Michele Willson & Tama Leaver, June 10, 2015 https://www.tandfonline.com/doi/full/10.1080/22041451.2015.1048039
- Tech company Tresata just became Charlotte's third unicorn, Caroline Hudson - Staff Writer, Charlotte Business Journal, Oct 10, 2018, 10:55am EDT https://www.bizjournals.com/charlotte/news/2018/10/10/tech-company-tresata-just-became-charlottes-third.html
- The NYPD Was Systematically Ticketing Legally Parked Cars for Millions of Dollars a Year- Open Data Just Put an End to It, May 11, 2016, http://iquantny.tumblr.com/
- Open Data Reveals $791 Million Error in Newly Adopted NYC Budget July 15, 2016 http://iquantny.tumblr.com/

- A Look at NYC's $650 Million Property Tax Breaks Related to Religion http://iquantny.tumblr.com/
- 8 Principles of Open Government Data LINK

Chapter 8

Harnessing Social Media Data

William Seeley and Lauren Zajac Team 1

8.1 Topic Overview

The readings this week focused on new innovations to manage Big Data and the volume of information on social media. As discussed in earlier chapters, all of this information can be overwhelming, and can cause information blindness. Patrick Meier explains that sifting through this vast quantity of information is even harder than finding a "needle in a hay stack…you are actually trying to find a needle in a meadow" (p. 96). Chapter three and five are used to explore new tools, concepts and platforms in development and to help digital humanitarians accurately and quickly make sense of vast quantities of information. These tools can be used in a crisis to help to quickly classify information on social media into requests for help and asks for assistance, and offer built in digital "rules", using multiple people verify the same information, to ensure that there is quality control and assurance. The faster processing time enables disaster response teams to respond faster. While the readings focused almost exclusively on the use of artificial intelligence and other data platforms for digital humanitarian response teams, there are some important management lessons for all of us, including the hope that we can better understand many of these technical innovations, and begin to determine ways that they can help us tackle Big Data in our daily work.

8.2 Chapter Summaries

Chapter 3 describes some of the early attempts to find the needle in the meadow, including a "match" program used in 2010 in response to the fires in Russia. Patrick Meier explains that this was a "crisis map with purpose…to match people in need of help with those willing to provide help" (p. 50). The digital humanitarians also ran a "phone match" service for those who did not have access to the internet (p. 52). This match concept allowed the team the ability to better manage Big Data.

Meier and his team of digital humanitarians, now known as the Standby Volunteer Task Force (SBTF) liked the concept of a match program to quickly get resources to people in need, but they also knew that manual processes to make sense of Big Crisis data was ineffective and can be discouraging for volunteers. They began to experiment with "micro tasking" which is a processes of slicing and dicing information into smaller manageable sections to increase accuracy and ability to process a great deal of information (p. 62). This way, instead of 100 people trying to work through 10,000 tweets, they can each look at only 100 distinct tweets. Crowd crafting was an open source platform available at the time that used this concept, and with some alterations, it was used by the digital humanitarians following the typhoon in the Philippines. It also had an important quality control feature, and it required that pieces of information are verified by three to five other people to increase accuracy and reliability (p.65).

This concept led Meier and his team to develop and implement their "MicroMappers Platform" to allow their digital humanitarians to quickly tag tweets, images, and videos. The program also has a voting methodology, which requires 3-5 people review and agree on information, which helps to ensure quality control and agreement on the severity of the damage (p. 69). Micromappers also has an algorithm to sort and identify only unique tweets, which allowed a faster way to sort through a great deal of information (p. 71).

These innovations were critical, but still imperfect. In Chapter 5, Patrick Meier begins to explain some additional technological advances that are also helpful to sort through all of the vast quantities of social media information available following a disaster. One innovation is data mining, or an "automatic analysis of large data sets" (p. 97). First used to assist the World Health Organization monitor outbreaks, using a "Health Map" Harvard University helped the digital humanitarians use the same concept to document human rights abuse and violence in Syria (p. 98). The program uses text classifications for social media posts on Facebook and twitter. In order to be effective, these text classifications or "codes" need to be modified based upon the type of disaster, the location, language and culture (p. 104). Patrick Meier and his team first applied this technology in the response to the tornado in Oklahoma, and were able to discover that out of over two million tweets, only several hundred were actually individuals seeking or offering to help (p. 103).

Because the text codes and classifications need to be customized, Meier and his team went on to build the Artificial Intelligence for Disaster Platform or AIDR (p. 104). This platform is user friendly, and open source, and allows digital humanitarians the ability to…."quickly crowdsource the creation of hundreds of classifiers" in response to a disaster (p. 104). AIDR also been designed to require that multiple volunteers verify tweets and images to ensure quality control and accuracy, which also helps the AIDR program learn and improve their accuracy (p.105). In addition to social media uses, Meier and his team are also partnering with UNICEF to develop the AIDR platform to use the same concept for SMS text messaging (p. 108).

8.3 Key Take-Aways (for Yellowdig)

In conclusion, the readings this week explored a number of new concepts, platforms and strategies to analyze a great deal of information in a very short amount of time. This is critical for digital humanitarian efforts, but will also be increasingly critical and important in our daily lives as Big Data makes it harder to process information in any manual way. We are truly looking for a needle in a meadow

YELLOWDIG BOARD

8.3.1 Discussion Questions

Questions: 1. Meier described microtasking as a method where information is sliced and diced into smaller, more manageable sections to assist in processing a great deal of information in a short amount of time (p. 62). Have you used any similar techniques in your data analysis? 2. Artificial Intelligence (AI) seems like science fiction, but we continue to be exposed to more examples of it in our daily life- "Just ask Siri". What are some ways AI has or can impact your daily work? 3. Emergency Management has very rigid standards for data collection and responses, which makes additional data collection nearly impossible. Does your industry have the same issue? How can one overcome this barrier, especially in the age of social media?

8.4 References

- Meier, P. (2015). Digital humanitarians: how big data is changing the face of humanitarian response. Routledge. **CH3 crowd computing social media**

- Meier, P. (2015). Digital humanitarians: how big data is changing the face of humanitarian response. Routledge. **CH5 artificial intelligence for disaster response**

Chapter 9

Remote Sensors

Team 2: Matthew Simon and Carlos Lopez

9.1 Topic Overview

Technology is changing the way we do business, the way we travel, and the way we live. In this week's readings the authors include interesting perspectives regarding how a data driven city could operate to create a more productive, healthier, and engaging society. For example, both authors touched on transportation and the available opportunities with GPS data in our phones. Basically, when one sits in a traffic jam, Google can capture the minute by minute data on our phones and when combined with the data from other drivers it is able to accurately and anonymously demonstrate traffic conditions and delays (Eagle & Greene, 2014).

Google Maps is a great tool, one that is anonymous, and can be used daily for commutes. Alex Pentland in "Social Physics" offers the "smart city" and takes it to the next level where your commuting patterns along with the rest of the population in your city are part of a model that can provide you with the optimized time and route for your trip so that you have the most efficient travel experience on your way to work. Similarly, commercial vehicles can identify the optimized travel times reducing conflicts with passenger vehicles and improving their efficiency. Further, Pentland describes a "smart city" where people getting the flu and their recent whereabouts could be mapped and when that is overlaid with others that also have the flu the location of where the flu started could be identified and contained before it spreads further.

Smart cities may be the future or perhaps elements of this are already in place. We produce data every day such as how long we are on the road, what we ate for lunch, and how many steps per day we take. The authors basically propose a smart city that takes the individual life logging efforts per se that could be documented on our phones and when combined with the life logs of the rest of the population it can clearly demonstrate how the city operates, identify energy and transportation peaks and valleys, improve emergency services and security, and develop a city that works for the people living in it. The overarching concern and what seems to be the theme for this course is privacy. The authors explain the need for privacy and customer buy-in for long term success; however, if there are data breaches where data is comprised it would deter and setback the smart city effort back to more conventional ways.

9.2 Chapter Summaries

Eagle, N., & Greene, K. (2014). Reality mining: Using big data to engineer a better world. MIT Press. CH5 urban analytics: traffic data, crime stats, and closed-circuit cameras

This chapter begins with an intriguing problem statement: The cost of congestion is exceeding $100 billion per year and wasting 34 hours per commuter per year according to the 2011 Urban Mobility Report. The conventional ways to acquire a morning or evening traffic reports included the local radio station but now

there are applications such as Google Maps, cameras on the freeways that can be viewed online, and websites that can identify any accident or construction delays.

Eagle and Green provided examples of companies seeking to provide a better traffic report service such as Intirx – which aggregates a variety of traffic data to make sure navigational devises in people's cars are up to date. Intrix has partnered with Audi, Nissan, and Ford to provide traffic data for their cars' built in navigation systems as well as the University of Maryland and the Interstate 95 (I-95) Corridor Coalition. The 20,000 miles of highway traversing 10 states along I-95 from Florida to Maine has helped Departments of Transportation determine where and how to better allocate transportation resources.

This is a similar issue facing the Arizona Department of Transportation where the revenue forecasts allows for the preservation and maintenance of the existing highway system with little or no funding to expand outside of the metro regions of Phoenix and Tucson.

Eagle and Greene also discuss data for predicting crime. At the core of predicting crime is a solid database of the crimes that have occurred previously. Incident and arrest reports including the time, date, location, crime code, and persons involved provide data that help the department keep a history and trends in their region.

The next generation of data for predicting crime includes mapping and providing all of this information in a visual format. More recently, specialized algorithms and real time crime data is frequently updating crime maps that can position police in locations before crimes occur. One example is in Memphis, Tennessee where police have run a program since 2005 called "Blue crush" (Criminal Reduction Utilizing Statistical History) that examines current activity levels and shifts in crime levels due to previous changes in police coverage (Eagle & Greene, 2014).

Lastly, Eagle and Greene touch on video data to catch and possibly deter criminal activity. The Department of Homeland Security (DHS) has funded many police department's cameras in an effort to combat terrorist threats; in 2009 DHS spent $15 million in seven cities. In 2010, more than $830 million went to 64 metropolitan areas; in 2011 31 cities used $662 million of funding under their initiative. Research is mixed about the effectiveness of cameras. Studies in Los Angeles, London, Chicago, and Baltimore found some areas to have little or no effect in deterring criminals and in other areas such as Chicago crime dropped 12%.

Pentland, A. (2015). Social Physics: How social networks can make us smarter. Penguin. CH8 sensing citiesThis chapter focuses on two types of data thac can define the rhytm of a city: traffic metrics and crime statistics

This chapter sets the current conditions of how cities currently operate and offers a concept of what they could be if cities adopted data-driven and "smart city" initiatives.

Pentland sets the stage by going back to the 1800's when the industrial revolution spurred rapid urban growth and created huge social and environmental problems. The remedy then was to build centralized networks that delivered clean water and safe food, enabled commerce, removed waste, provided energy, facilitated transportation, and access to health care, police, and education.
However, the author points out that these solutions are outdated and becoming "increasingly obsolete" as cities struggle with transportation, health care, and education issues and more. Pentland offers a different framework of rather than having static systems that are separated by function – water, food, transport, to consider them as dynamic and holistic.
What would be the data source for these dynamic and holistic efforts to take shape? You guessed it via the mobile phone. Pentland explains that wireless devices and networks could become the eyes and ears of an all controlling "smart city."

But how does this happen?

First, this requires social physics, specifically the visualization tools that will allow citizens to use these new data streams to manage the city. The visualization will inform citizens about where people live, where industrial/commercial areas are located, and the demographics of its residents. The second step the author outlines is a new deal on data, an architecture and legal policy that guarantees privacy, stability and efficient government.

Pentland provides and emphasis on Behavior demographics. Pentland explains,

"For most people, the primary pattern is the workday, that is, going to work and coming home, usually along the same path day after day. The second most pronounced pattern is the weekend and days off, often with the characteristic behavior of sleeping in and spending that night out in a location besides home or work. Perhaps surprisingly, the places we go and things we do during our free time are almost as regular as our work patterns. The third patters is a wild card – days spent exploring, usually a shopping trip or an outing; together these three patterns typically account for 90 percent of our behavior. In summary, by combining these habits in time with the behavior demographics it can allow us to better plan city transportation, services, and growth; these data driven forecasts allow us to prepare for peaks in demand and manage them better. The ability to know where and when the people who are at risks of diabetes eat, or where the people who have trouble handling money shop also has great potential to improving public health and education."

Health issues could also be diagnosed based on individual behavior. Pentland describes that people feeling sick tend to behave differently; for example, those with a sore throat and cough symptoms were found to have their normal pattern of socialization disrupted, and they began to interact with more and different people. Those with a common cold, their overall number of interactions and nighttime interactions increased. People with fever limited their movement and people feeling stressed, sad, depressed became socially isolated on symptomatic days.

Behaviors can be signs of an emerging illness and the author points out an app such as Ginger.io that could identify that change in behavior and figure out if an illness is coming. To take it a step further, the author explains that by crowdsourcing this behavior across a population and then combining that info with data about where and when people went the infection risk area can be figured out. If this was known then action could be taken to avoid spreading the disease further.

9.3 Key Take-Aways (for Yellowdig)

9.3.1 Discussion Questions

1) Traffic, crime, and health technology are rapidly improving to better predict and identify incidents. What opportunities or concerns do you find with technology and the data driven approach?

2) What are examples of emerging technology in your industry? What are the strengths and/or weaknesses?

3) Would you be interested in living in a "smart city" as Pentland describes? What would you change or improve?

9.4 References

- Eagle, N., & Greene, K. (2014). Reality mining: Using big data to engineer a better world. MIT Press. **CH5 urban analytics: traffic data, crime stats, and closed-circuit cameras**

- Pentland, A. (2015). Social Physics: How social networks can make us smarter. Penguin. **CH8 sensing cities**

Chapter 10

Challenges of Data Quality

Marcela Morales and Joseph Lynch

10.1 Topic Overview

Big Data (social media) has been used during crisis situtations with differing levels of effectiveness. During the 3 chapters, the prevalance of big date, veratcity of big data, and the usability of big data in crisis situations is explored.

10.2 Chapter Summaries

CH2 the rise of big crisis data pp 31-47

This section discussed in great detail the rise in the use of big data in business and humanitarian efforts. Meier asserted that, " If government and humanitarian organizations do not actively or explicitly create demand for relevant, findable, and high quality social media content during disasters, then why should a supply of high quality data follow"(Meier, p. 31)? The use of hashtags by the Filipino government directly during Typhoon Pablo (#PabloPH) helped providing important updates during this natural disaster. Standardizing hashtags during crisis events can help in reporting accurate and high quality information to the public and from the public. What about False data? That is something that needs to be considered when getting data from social media feeds. False information on urgent humanitarian aid could be sent to the area which can result in wasted time, resources, or even cost a life. Access to information during disasters is equally as important as access to food. Nobody wants poisoned or rotten food which false is. The New York Times is considered by many to bed the gold standard of high quality journalism and yet they have to make 7,000 corrections to articles every year (Meier, 2015). Collecting information from traditional and nontraditional sources can help create a reasonably accurate picture of the situation. In the Twitter world the amount of tweets pinged from one location versus another can tell us a lot about the situation during the crisis. If the number of tweets are lower than we can potentially assume mass casualties, no electricity, or internet. If the number of tweets is the same or higher than average then we can assume that area was not hit as hard. Lastly, what is our responsibility with big data? What responsibility do we have to protect information that can create a safety risk? A suburban New York newspaper provided the names and addresses of all who held handgun permits including police officers, prison guards, and other position sensitive occupations.

CH7 verifying big crisis data via crowd computing

In this section identifying the authenticity of big data during a crisis. Humanitarian volunteers were being screened to ensure they were not trying to sabotage humanitarian efforts in Libya. In the survey they were asked to provide professional or academic email addresses. Twitter handles and Facebook pages were then used to get more information on past tweets and posts to verify the authenticity of this person. Several

large pieces of big data, where by themselves didn't mean much, however used contextually provided a clearer picture of the volunteer that was being vetted. According to the text, the quality of crowdsourced information simply mirrors the reliability of society. If we have low confidence in the reliability of the crowdsourced information then the belief is that this is a diagnosis of society and not the crowdsourcing tool itself. Additionally, an investigative strategy is imperative when going through the information and such strategies include asking sourced pointed questions and triangulating content to get a clear picture. Additionally, one must consider the platform being used and what type of information is being shared. For example, Redditt is probably best used for sharing pictures not necessarily sharing information about developing crisis. Redditt doesn't encourage the analysis of a situation in order to discredit rumors.

CH8 verifying big crisis data via artificial intelligence

"No technology can automatically verify a piece of User-generated Content with 100% certainty. The human eye or traditional investigations aren't enough either. It is the combination of the two" (Meier, 145). If everyone had perfect information during a crisis then it really wouldn't be a crisis. Also, while bad information can have far reaching effects we need to balance the effects of no information. While verifying data is very important we need to realize that we will need to deal with bad information as well and use the tools at our disposal to siphon out the accurate information. All information is good information—even when it is bad.

10.3 Key Take-Aways (for Yellowdig)

Overview

In Chapter 2, 7, and 8, the readings focused on the use of data in crisis or disaster situations, how to verify the information, what tools can be used to verify the data. Some of the larger questions to review are:

How to make big data usable in disasters?

"If governments and humanitarian organizations do not actively or explicitly create demand for relevant, findable, and high- quality social media content during disasters, then why should supply of high-quality data follow?"

The variance of reliable data on social media during disasters vary greatly. The Joplin tornado (2011) had 10% relevant data shared on social media, the Australian Bush Fires (2009) had 65% relevant data, Hurricane Sandy (2012) had .001% relevant data.

Can the Data be Trusted?

During the 2010 Chilean Earthquake, emergency services were responding to requests on social media that included fake information. The false data was responsible for sending emergency response teams on wild goose chases instead of being at true emergencies. False data was repeated again with the hurricane in Haiti, Hurricane Sandy, hacked AP social media sites saying the White House was attacked (briefly wiping out $130 Billion in the stock market),

How Much Data is Right?

When responding to disaster social media, is one tweet as valuable as 1,000 tweets. Does the quantity add to the validity? If Social media is being used to prepare disaster response, do areas without social media receive the same service. A large portion of the world's population does not use social media.

Can crowd sourcing be used to verify Data?

Libya screened volunteers to sourced data during humanitarian efforts. Identification and social media of the volunteers were vetted to ensure that information was not sabotaged. In Russia during the election, the crowd sourced election information function of social media was turned off due to massive reports of election violations. The Government did not want this shared.

Can Artificial Intelligence verify Big Data?

No technology can verify information alone. The blend of human "cognition" along with technical processes will deliver the most accurate result.

To best used data in a crisis situation, being "right is more important than being first"

10.4 Synopsis

10.5 Discussion Questions

1. Does the potential value of big data (social media/crowds sourcing) in crisis situations outweigh the potential false data and misuse of emergency response in crisis situations?

2. Does the variability of usable information (0.001% to 65%) cause you to limit your expectations for the viability of crowd sourced data?

10.6 References

- Meier, P. (2015). Digital humanitarians: how big data is changing the face of humanitarian response. **Routledge. CH2 the rise of big crisis data pp 31-47**

- Meier, P. (2015). Digital humanitarians: how big data is changing the face of humanitarian response. **Routledge. CH7 verifying big crisis data via crowd computing**

- Meier, P. (2015). Digital humanitarians: how big data is changing the face of humanitarian response. **Routledge. CH8 verifying big crisis data via artificial intelligence**

Chapter 11

Relationship Discovery

(Team 4)

11.1 Topic Overview

This chapter will focus on the exploration of ideas and how to harness big data for thought development.

The first focus is on the flow of ideas and how networking better develops ideas in any environment. Online forums or social groups can be used to collect and discuss topics, or provide a sounding board for ideas or potential solutions to a particular problem. Professionals from around the world can participate together to tackle general issues or specific problems alike.

Several years ago, an experiment was performed in the AT&T Bell Laboratories regarding employee productivity. In their discussions, people said, "In fields like computer programming, an eight-to-one difference between the productivity of stars and average workers has been reported" and "Ten to fifteen percent of our scientists and engineers are stars, while the vast majority are simply good, solid middle performers." (Kelley and Caplan, 1993). When managers were questioned about the difference between star performers and good performers, they received answers that it was related to IQ, a drive to succeed, better problem solvers, or a will to win (Kelley and Caplan, 1993). With regards to making good performers, great; there is little that can be done to change a person's personality or natural skill-set, so more was looked into what star performers do that makes them productive.

They found that individuals that were the star performers in the organization also engaged in a form of social exploration. Social exploration are those that take independent thoughts and then take them into the social networks, whether it be in-person or online and then discuss those ideas. In the course of this exploration, ideas can be greater developed by those that may have had similar experiences or have a diverse background that can contribute a perspective to the thought in a way that enhances it.

Although, the information age has brought about a number of information sharing platforms that have allowed for a greater ability to share ideas, Alex Pentland the author of Social Physics has said in his 2014 TedTalk, "Twitter and Facebook could make us smarter… but they are full of echo chambers. Face-to-face connections are still most important in changing minds, and making us smarter" (Pentland, 2014). He emphasizes that along with online social networks, we cannot pass on the personal connections that we have in social exploration.

The second focus is the actual use of online forums and other online social platforms can be used as a repository of data that can be accessed or analyzed to determine trends or anticipate needs. For those that aren't familiar with online forums, there are a number of tools that are available for data mining, idea development, and social exploration. With those networks though, is the warning to avoid the over-reliance

on them for thought, or echo chambers which will be discussed later, and the need to maintain personal associations, as previously mentioned.

11.2 Chapter Summaries

11.2.1 Social Physics, Chapter 2 - Exploration

Chapter 2 in Social Physics discusses the need for social learning. That by tapping into our social networks, we can rely upon information that has already been discovered and vetted. It allows us to identify those in our network with particular strengths that could be a source or a sounding board for new ideas. The key is to make sure that we have a sufficiently broad enough network in terms of numbers that we can harness the diversity of thought. Pentland says,

"The most consistently creative and insightful people are explorers. They spend an enormous amount of time seeking out new people and different ideas, without necessarily trying very hard to find the 'best' people or 'best' ideas. Instead, they seek out people with different views and different ideas." p.26

Pentland discusses the need to "meet many different sorts of people" to be able to harness as many different viewpoints as reasonably possible in an attempt to develop the thought (p.26). "The most productive people are constantly developing and testing a new story, adding newly discovered ideas to the story and then trying it out on everyone they meet (pgs.26-27).

The text identifies the need for social learning, diversity of thought in the exploration and flow of ideas, and to seek out contrarians to avoid feedback loops (p. 39-40). If there is an overreliance on other people's ideas and if your network is too small, then ideas tend to circulate around upon itself. Pentland refers to this as an echo chamber. Echo chambers begin to harm idea development when ideas are borrowed from others rather than receiving feedback on original ideas. Effectively, you get a recirculation of thought rather than idea development.

In the text, Pentland cites an experiment that he performed with MIT graduate assistants were they reviewed an online day trading platform called eToro. eToro allows traders to see what trades other traders are making and even grant a commission for trades that are copied. In his experiment he found that ideas that were shared and developed resulting in a higher rate of return for those investors. But he warned of isolationism and the echo chamber where either too little or too much feedback was sought. The following chart found on page 32 of the text, illustrates his point:

To further discuss potential pitfalls of social exploration, Pentland warns against the dangers of groupthink and echo chambers where individuals have started to follow the herd mentality and stopped relying upon their own thought to enhance the ideas in the network. Where social networks can enhance the development of ideas, groupthink has the potential to do more harm to the network than through individual thinkers. Groupthink or echo chambers can result in confidence bubbles that pop dramatically and potentially cause catastrophic damage.

You may be interested in the TedTalk that Alex Pentland did on this subject:

11.2.2 Reality Mining, Chapter 8 - Engineering and Policy: Addressing National Sentiment, Economic Deficits, and Disasters

The text focuses on using the "national repositories of data" and how to use them (p.125). It focuses on the five major points, demographic snapshots, mobility matters, roll your own polling app, tweets of crisis, and mining for financial futures.

11.2.2.1 Demographic Snapshots:

Data can be used to track trends or progress. While some data moves slowly like a national census, other data is very quick like phone data. "Journalists as well as policy makers are also interested in historic data

and trends over time that illustrate improvements, lack of improvements, and general changes in population, economics, political affiliation, and religion, among other metrics" (p.128).

11.2.2.2 Mobility Matters:

In government, it can take years of planning and appropriation to have major facilities constructed, such as schools or government centers. Understanding the mobility of people is key to being able to properly plan. "Ultimately, CDR (Call Detail Records) analysis can be applied to … help governments better understand the dynamics of of impermanent settlements such as slums in order to allocate resources better and to predict the repercussions of natural disasters such as earthquakes on those settlements (p.130).

11.2.2.3 Roll Your Own Polling App:

Google Adwords and Facebook Ads have become a great way to determine the sentiment of a particular group of people or demographic. Even items items as simple as "click if your a Republican" or "click if your a Democrat" are ways to get a quick feel for how people are thinking and feeling about certain issues or about politics in general.

11.2.2.4 Tweets of Crisis:

While Twitter isn't as great a place to poll the audience, the use of hashtags and keywords inside of Twitter are easy to mine with algorithms and are a way to determine trending topics of stress or crisis.

11.2.2.5 Mining for Financial Futures:

With more financial companies with personal information (and the book cites the Intuit corporation purchasing Mint), it is possible to use people's personal financial planning information to judge consumer confidence, the likeliness people are to borrow money, even to what types of investments people are interested in (p.138-139). This allows companies to better plan and shape their marketing packages to sentiment of the demographic they wish to target.

11.3 Key Take-Aways (for Yellowdig)

11.3.1 Discussion Questions

- How can we avoid an overreliance on our social networks and groupthink mentality?
- To what limit to we provide incentives for (or nudge) idea development into the sweet spot of development without influencing the outcomes?
- When using data to predict trends, do we over-rely on extrapolation models and correlations for definitive answers?

11.4 References

- Eagle, N., & Greene, K. (2014). Reality mining: Using big data to engineer a better world. MIT Press. CH8 economic deficits and disasters
- Gonzalez, Marta C, Hidalgo, Cesar A, and Barabasi, Albert-Laszlo. (2008). Understanding Individual Human Mobility Patterns, Nature 453 no. 7196. Pgs. 779-782.
- Kelley, R., & Caplan, J. (1993, July/August). How Bell Labs Creates Star Performers. Retrieved February 1, 2019, from LINK
- Pentland, A. (2014, May 20). How Social Networks Make Us Smarter. TedxBeaconStreet. Retrieved from LINK

- Pentland, A. (2015). Social Physics: How social networks can make us smarter. Penguin. CH2 exploration
- Pentland, A. (2017, March 20). Beyond the Echo Chamber, Harvard Business Review. Retrieved February 1, 2019, from LINK

Chapter 12

Playing Moneyball with Prediction

(Team 5)

12.1 Topic Overview

Summary

These chapters discuss the use of data for predictive purposes. Eagle and Greene discuss how big data has the ability to be a "real-time crystal ball" to track health threats, crime and traffic. Used in a mathematical model, big data also has the ability to predict behavior and simplify decision making in order to optimize resources. O'Neil discusses how these models are used every day to inform decisions in Major League Baseball, access to credit and financing (FICO scores), crime (CompStat), recidivism and more. But do the costs outweigh the benefits, and does the over simplification of big data lead to effectiveness in solving the problems they are created to address?

Eagle and Greene make the point that trust should not be put solely into big data and suggest that the data needs to be backed up to validate what is happening in real life. The concerns that O'Neil raises in Weapons of Math Destruction about the dangers of mathematical models and the assumptions they use, could be enhanced with this solutions-oriented approach of validating what is happening on the ground. Instead of creating their own "pernicious feedback loops," as she calls them, what if data models added a human touch by gathering real-time feedback and information from people affected by those models? Could it be possible that they could create their own feedback loop that was not grounded in assumptions?

This leads to another question: "Does data ever tell the full story?" Arizona students take an annual statewide exam called the AzMERIT. The test provides information at the end of the year to know whether students are proficient in math and English. AzMERIT produces both small and big data. For example, parents use their child's score to understand how their child is doing – is he on track or behind? Schools use the scores at the school and classroom levels to know how well their teachers are teaching and students are learning. Policymakers and administrators use the scores at the largest scale to determine if the school needs help or if they should receive a financial reward for their good performance. But does the data tell the full story of how a teacher taught during the school year or what a child fully understands? What if an A student woke up sick and had a bad day and did poorly on the test as a result? What if students did not have breakfast or had a traumatic event happen the day before? Could they perform as their best? Further, for teachers, does one good test score mean they are an incredible teacher, or a bad test score mean they are terrible? There are so many variables that go into test score results that are not reflected in the data. Test scores may not show that the majority of the class started the year three grade levels behind and the teacher worked overtime providing tutoring and Saturday classes so that the students could catch up one and a half grade levels in a year, however, they still fell a year behind their peers making their scores look terrible. It also may not show that the 5th grade class was using the 6th grade class curriculum (their school's norm),

thus blowing away all other 5th graders across the state. In this case it was not the teacher teaching that led to their success, but the school's accelerated curriculum and the relative wealth of the student population. It also doesn't show the obstacles that students have to overcome just to show up to class every day, let alone achieve. The point being that data does not tell the whole story.

Data models often prioritize efficiency of decision making over understanding the entirety of the situation that the data reflects. Data models could be made stronger by adding a human level that would increase the understanding of the data. This could be accomplished through surveys in person or via an electronic mechanism, or the creators of the data models spending time in the field building a greater understanding of the people's lives that they are trying to account for in the models. What if mathematicians calculating FICO scores spent a day in the life of individuals with low, medium and high levels of credit and debriefed their observations at the end of the day with their peers? Could they learn something that would better inform their models or change the way they count their data? IDEO does this famously with product design, using a design thinking process that puts the designers of products in the shoes of its ultimate users. Their job is to understand how the product would be used in real-life and to build empathy for the users to create a better product. Why can't mathematicians do the same? Could this personal touch add greater value to data models so that they are not creating the "pernicious feedback loops" that O'Neil warns of in her text?

Further, what happens when data models are flawed? Often these flaws are accepted and those impacted are marginalized as collateral damage for the sake of progress. There is often little recourse and little compassion for those who are stuck in this no-win situation. Those who do seek justice are often those that have the means (money) and resources (connections) to make waves. But it is often the under privileged who fall victim to the Weapons of Math Destruction and are the least likely to have the capacity to make said waves. What if the people affected by these models had the ability to understand what was in the model, how it was computed, and the ability to refute its contents? Often, when decisions are made by the WMD, the end user has no idea they have even been a victim. They are unaware of the decision process or how the conclusion was drawn. They are simply informed they are not qualified for the loan or that they have not been offered the job. If the process was more transparent, would that nudge this trend of mechanical decision making in the other direction? Could people make the case that they are worthy of credit, despite having a poor FICO score, by using other measures? FICO includes payment history (making payments on time), current levels of debt, types of credit held, credit history, and new accounts. (The Motley Fool) If someone has $2 million in cash in the bank, but has never used a credit card or had a loan, their FICO score would be poor and they would pay more in interest than someone with a stronger score. However, with $2 million cash-on-hand this person is likely to not pose a threat to the credit company and should have a lower interest rate. What if it was possible for individuals to use other measures, like the amount they have in savings, their personal net worth, amount invested, their ability to pay for their short-term expenses in cash, or others to make the case that their FICO score should increase? While this runs against the very nature of a model's purpose, it makes the case that models can be flawed, and their simplified nature does not tell the whole story of the people it is trying to represent.

It has been made clear that the introduction of data was intended to help the decision process become less bias and more quantitative. It was intended to add a finite element of measurement and eliminate arbitrary, qualitative 'feelings' that can interfere with human judgement. But it can be argued that this change has only led to a similar outcome. The use of data as the primary driver for decisions has resulted in unfair judgments, bias calculations and systematic stereotyping; all of which where the impetus for using it in the first place.

12.2 Chapter Summaries

Reality Mining: Using Big Data to Engineer a Better World

Chapter 6: Optimizing Resource Allocation

In Chapter 6, Eagle and Greene discuss how big data can be used to allocate resources in the most efficient manner. They discuss how crime and traffic data can be used to make predictions about future needs. Examples include proactively placing police officers in specific locations known for high levels of crime

incidents, predicting needs road infrastructure, and tracking the spread of disease. First, they look at how technology has enabled companies like Inrix and Google to track the speed of traffic and estimate travel times. They also discuss the idea of surprise modeling, which grew out of a project at Microsoft and enables a more accurate prediction of unexpected events on the road (like a traffic jam or when the jam clears) and how much time those might entail. Engle and Greene also discuss the Vehicle Probe Project, which is crowdsourcing data from GPS units in cars and other sources. The data provides more accuracy for drivers themselves, but also gives transportation planners the ability to solve for infrastructure needs and to evacuate people more efficiently in a disaster. They also discuss how traffic flow can be used to track how germs move within a city and can give public health professionals a window into how quickly and in what places a health issue may be spreading.

Chapter 10: Engineering a Safer and Healthier World

In Chapter 10, Eagle and Greene discuss how various forms of technology are being used to inform public health by tracking, predicting and stopping the spread of various communicable diseases. They discuss how Google searches, public posts on Facebook and Twitter, quick surveys, air travel and shipping routes, and digital footprints from cell phone use can help with these efforts. Air and travel shipping routes, for example, track the movement of people and goods from place to place. If people are ill and infect others, disease spreads. Likewise, if there is produce coming from one country to another that has infected mosquitoes living in it, those mosquitoes could move from one country to another and facilitate the spread of disease. Eagle and Green note that possibly a more direct predictor of the movement of people is to use data gathered from actual cell phone use and location. CDRs (Call Detail Records) can be used to see how people move at a much greater scale than air routes. The authors discuss pairing CDR data with syndromic surveillance to ascertain public health trends, which includes things like video of people coughing at train stations or orange juice sales. Likewise, Eagle and Greene discuss how Google uses search terms and locations of users to track possible incidences of the flu and how it may be spreading across the country. Google Flu Trends has tracked the spread of the flu in multiple countries using their technologies and expanded it to also track Dengue as well through Google Dengue Trends. Others have found similar success in tracking the flu via Facebook and Twitter.

Engle and Greene also discuss the potential of how big data and mining its contents has the potential to be a "real-time crystal ball" about health threats. However, they also argue that trust should not be put solely into big data. The data needs to be backed up to validate what is actually happening on the ground, which could be accomplished through surveys delivered to people in the affected areas.

Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy

Chapter 1: Bomb Parts

O'Neil introduces the idea of Weapons of Math Destruction (WMDs) in this chapter in greater depth. She outlines the critical elements of what makes a WMD: opacity, scale and damage. Opacity is the extent to which the model is transparent. Can people affected by the mathematical model be made aware of and understand the data included and how it works? The second element is scale: does the model have the potential to grow and be applied to other people or to an entire industry? The third criteria is damage: does the model have the potential to negatively impact a person or many people's lives?

She provides three examples of models that have been used and their alignment with these criteria. The first being the Moneyball example, where Major League Baseball teams use models to inform their offense and defensive strategies. For example, using this data model, teams are better able to shift field position to be in the location where the batter is most likely to hit in that situation, with say two outs and with a left handed pitcher. O'Neil believes this type of model is generally transparent, as data is available publicly and to the players who are in the models. It is also constantly updated with new information to adjust the models and uses relevant data, rather than proxies.

The second example the author provided was one related to her family meals. She shared that the users of the model (her family) could question the model and its results and she could explain it, they could weight the criteria differently, and that the model is useful for her family but is unlikely to scale.

The third model is focused on recidivism. She cites the LSI-R questionnaire that prisoners fill out. They are not made aware of its purpose, nor do they ever learn the results. The model is not transparent and she believes, has the ability to destroy lives, given its pernicious feedback loop. She discusses the scalability of the model and its use already in a number of states.

Chapter 5: Justice in the Era of Big Data

The focus of this chapter is on the overall fairness of the algorithms and formulas used in big data. O'Neil's point is that the models are used to increase efficiency abut often at the cost of equality. She looks at several examples relating to crime prevention programs as well as "Stop and Frisk" policies as being primary examples of how Weapons of Math Destruction (WMD's) are mistakenly believed to be both scientific and fair. She highlights how many crime prevention programs (PredPol, CompStat) are actually self-fulfilling in that they highlight areas where crime is expected to be more likely, such as a poorer neighborhood or community. As a result of this data output, police step up enforcement in that area as a means of deterring criminal activity. They make arrests on nominal charges, believing this will prevent larger, more heinous incidents from occurring. While that is one side of the arguments, these actions also inadvertently confirm the model and thus increase the data for future predictions. She offers the comparison of police cracking down in the Gold Coast, the very affluent neighborhood off of Lake Shore Drive in Chicago. Police might find violations like unpaid parking tickets, jaywalking, or other nominal infractions. Over time, these arrests or citations would create a data trail indicting crime was running rampant in this part of town! Data is used in the interest of efficiency, but at the cost of fairness. She goes on to offer a further comparison in recidivism and the idea that 'high risk' prisoners often get longer sentences, which translates into more difficulty in finding employment and thus making it difficult to secure a stable income, dependable housing, reliable transportation etc. These factors make it more likely that this newly released inmate will become a repeat offender. If and when that happens, the data is added and once again the self-fulfilling cycle is confirmed. The flaw in this type of Big Data application is summed up by O'Neil in this way: "All too often, they use data to justify the working of the system but not to question or improve the system."

Chapter 8: Landing Credit

In chapter 8, O'Neil takes us further down the scary path of Big Data and its impact on the finance landscape. Be it borrowing and buying or getting hired and promoted, data analytics and number crunching often have more to do with these decisions than we realize. Just as would-be criminals were identified by data algorithms as probable offenders, lending data algorithms offer the same type of 'probability bucket' linking potential borrowers with others that have similar data points. O'Neil references this as the "Birds of a Feather" approach, meaning that subjects who exhibit similar qualities are assumed to have other qualities in common. If everyone in your zip code that drives the same model car and makes the same salary range default on their mortgage payments, then its is inferred that you will too- its simple association by data points. And while many may point to the truism in this concept, the bigger picture is that computers can only account for so much. They cannot see the nuances. The data produces a general picture but cannot account for the finer details of context, situation, or common sense. The rub is that we as a society have essentially come full circle as data collection was inteneded as a means to incorporate fairness and equality, eliminating judgements on race and religion. But the overuse of data and the resulting typecasting has introduced unintentional consequences by being unforgivingly flawed and lacking human reasonability.

12.3 Key Take-Aways (for Yellowdig)

https://youtu.be/Q8SFAbqHYu4

12.3.1 Discussion Questions

1 - From your perspective, have data models added value to society or are they causing more harm than they should?

2- Have data models eliminated human bias or just disguised it using technology? Do you think some positive things can come from the use of data models?

3 - Has a data model ever affected your life? How did you find out about it? How would you rank it on O'Neil's three criteria of: opacity (or transparency of the model), ability to scale and damage caused?

4 – Should the industry of big data be regulated? Meaning, should companies that collect and sell people's data be given a set of rules and industry guidelines to be held accountable to, including to allow people transparency to see what is being collected about them and have the ability to change things that are incorrect?

5 - What role do you think data algorithms should play in our day to day life? Is there a level to which you feel it is appropriate? When is it too much?

6 - Have you experienced a situation where data was used to confirm or refute a decision that resulted in others being negatively impacted? (The proverbial collateral damage)

12.4 References

- Eagle, N., & Greene, K. (2014). Reality mining: Using big data to engineer a better world. MIT Press. **CH6 optimizing resource allocation**

- Eagle, N., & Greene, K. (2014). Reality mining: Using big data to engineer a better world. MIT Press. **CH10 engineering a safer and healthier world**

- O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books. **CH1 bomb parts**

- O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books. **CH5 justice in an era of big data**

- O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books. **CH8 landing credit**

- The Motley Fool. "What is a FICO Score." Retrieved from: https://www.fool.com/knowledge-center/what-is-a-fico-score.aspx

Chapter 13

Using Big Data for Evaluation

(Team 6)Tommy Kolwicz and Dennis S Stockwell

13.1 Topic Overview

This chapter focuses on how evaluators can use Big Data as another tool to provide data points that can be used for analysis. The overall theme is that there is a significant role for Big Data to play. The first reference lays out recommendations, that if Big Data is going to be collected and used that its use should be considered early, you need to understand the role that the selected big data will play, understand the limitations of the big data that you are using (demographics of users), how you will collect data from the demographics not represented by a specific social media platform, the role language may play in the collection of data, and there may be a role of shaping Big Data thru the use of hashtags or campaigns that encourage responses that can be measured. The second reference provides insights on how big data can be used to augment and fill in a lack of knowledge about communities from a program evaluation perspective. The authors argue for the immediate requirement for evaluators to learn and use big data sources, technologies, and methodologies. The bottom line is that there is a role for Big Data as long as users understand limitations.

13.2 Chapter Summaries

Can big data be used for evaluation? A UN Women feasibility study.

This week's reading is based on "Can Big Data be used for Evaluation" a UN Women feasibility study. The study was to "determine to what extent big data could help strengthen traditional UN women evaluations" The study focused on 2 countries Mexico and Pakistan. The study focused on Mexico due to its high internet penetration 65.3% and one of the top three twitter users representing approx. 20% of the population and Pakistan at the other side of the spectrum due to its lower internet penetration (22% of population) and over 31 million Facebook users. The Study focused on analysis of Twitter in Mexico and Facebook in Pakistan. Some of the pros and cons of Facebook were: Pros - conversations are more in depth, access to historical data, data can be collected via app (requires consent) and allows for analysis of social networks. Cons – Conversations are stovepiped, and data can only be obtained from public pages. For Twitter the Pros were analysis can be done in real time which allows for analysis of current issues and possibly identify key influencers within the network. Cons- Twitter limits conversation to so many characters, personal issues may not be discussed to the public nature of the media, and a small portion of the population are twitter users. Ultimately the study did conclude yes, "Big Data can be used for an evaluation", however there were several recommendations to keep in mind if you are going to use Big Data for an evaluation. Make sure that you understand the role that the selected big data will play, understand the limitations of the big data you are using (demographics of users), how you will collect data from the demographics not represented by a specific

social media platform, the role language may play in the collection of data, and there may be a role of shaping Big Data thru the use of hashtags or campaigns that encourage responses that can be measured.

Integrating big data into the monitoring and evaluation of development programs

This report is meant to be a "call to action"! Big data is here. We must use it to make the world a better place. The insights from big data can be used to augment and fill in a lack of knowledge about communities from a program evaluation perspective. The authors argue for the immediate requirement for evaluators to learn and use big data sources, technologies, and methodologies. Specifically the article is focusing on the real need to gain greater insight on the "impact of development programs on the poor and vulnerable."
This reading is an extremely in-depth report on two broad topics. Part 1 covers development evaluation in the age of big data; it's implications, possibilities, and challenges. This part of the paper essentially lays out all of the different types of data and all the varied methods of collection, while also putting types of data into bins essentially based on quantity. Like Digital Humanitarians, it hits a lot on the potential of real-time data. Then in chapter 2, the authors discuss the challenges of big data from a standpoint of how stakeholders can begin to integrate big data into an already existing framework of program monitoring, evaluation, and learning (MEL).
Part 2 takes us deep into a program evaluation discussion. In chapter 3 the theory of how, specifically, big data could be implemented into an existing MEL framework is discussed. The discussion hits a lot on where the holes are in existing evaluation systems and how big data can fill those holes. Chapter 4 discusses building big data into program monitoring. And chapter 5 discusses how to build big data into the MEL frameworks. Chapter 6 finishes off with a discussion on managing big data inclusive evaluations, the manager's role, and the special challenges associated with big data management. "This chapter stresses the critical role of the evaluation manager in ensuring that all evaluations address the key questions of concern to stakeholders and that the kinds of information generated can be used by a wide range of stakeholders and for different purposes."
The UN's Sustainable Development Goals (SDGs) are used in this paper to explain why it is important to rethink current MEL processes. The SDGs define the vision and goals for creating the "world we want" by focusing on the 5 Ps: People, Planet, Prosperity, Peace, and Partnership. There are 17 goals and 169 targets. One example is SDG–4, "Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all." Certainly we can all probably think of ways that big data could be used to facilitate the evaluation of that SDG.
There are three main ways that big data can be integrated into MEL frameworks: you can simply fit it into a conventional evaluation, you can use the data to strengthen an existing design, or you can create a new integrated design from the bottom up. Chapter 5 has some interesting case studies that I will share here.
• Using high frequency metering data for high–quality information about energy consumption and demand in rural solar micro–grids in India.
• Tablet–based financial education in Colombia. Using savings and transaction data combined with survey and telemetric tablet data.
• Assessing the effects of a government tax increase on smoking using changes in search query volume to assess the effects of a major increase in cigarette smoking in the USA. • Evaluating causal interactions between labor market shocks and internal mobility.

This paper serves as an excellent and informative primary source for how to integrate big data into program monitoring and evaluation, and what can be learned from that integration.

13.2.1 Discussion Questions

1) How can you use Big Data to evaluate one of your programs?
2) What are some blindspots in the evaluation that would result from relying only on Big Data as an evaluation tool?
3) Can you see a vision in your community for creation of new evaluation frameworks built from the bottom up integrating Big Data?

13.3 References

- Claudia Abreu Lopes, Savita Bailur, Giles Barton-Owe (2018). "Can big data be used for evaluation? A UN Women feasibility study." United Nations Entity for Gender Equality and the Empowerment of Women.

- Michael Bamberger (2016). Integrating big data into the monitoring and evaluation of development programs. UN Global Pulse

Chapter 14

Motivating People

Team 7 Lorna & Julie

14.1 Topic Overview

14.2 Chapter Summaries

14.2.1 Daniel Pink (2010) What Motivates Us? Harvard Business Review

Pink is interviewed by Katherine Bell on "What motivates us?" based on his book Drive: The Surprising Truth About What Motivates Us.

Pink states that based on 40-50 years of behavioral science research, people are not motivated by the common belief of the carrot/stick concept. Meaning, if you reward someone that person will display more of the desired behavior, and if you punish someone that person will display less of the undesired behavior. This concept is known as the punishment/reward drive. A second drive is the biological drive, meaning we eat when we are hungry. Pink states the biological and punishment/reward drives do not work well in the business world. He states people are motivated based on a third drive, which is contribution.

The carrot/stick motivator can actually be a de-motivator in business. For example, studies show that when people are given a high-stake reward for achieving a short-term goal, people will actually finagle the system or flat out cheat. A high-stake reward, used as a motivator, can narrow one's vision due to focusing on the reward. Pink states that these rewards are less effective for creative workers.

Managers should motivate creative individuals by giving their people a "sense of making progress." This can be done by helping people see their progress, by recognizing it, celebrating it, and keeping people on a path towards making more progress.

However, Pink states, that although people are motivated in business when they feel like they are contributing and making progress, people must be paid enough money for them to focus on work and not worry about money.

Pink discusses an Australian company who established "FedEx" days where employees could do whatever they wanted with whomever they chose, and then produced their work at the end of 24 hours. FedEx days produced amazing results, including software fixes and new product ideas. He believes people were motivated due to "a greater sense of autonomy, allowing people to make progress, and animate what the people are doing with a greater sense of purpose."

It's important for people to understand the source of their de-motivation. Popular reasons people are demotivated is the perception or reality of being underpaid, not making progress, or the feeling of being

micromanaged. Once the source is discovered, it's an opportunity to "re-sculpt" your job or possibly find a new workplace.

Pink said the writing and research he did for his new book has changed his work habits by establishing helpful measures to make progress. He developed a "do it yourself performance review." At the beginning of each month, he writes down what he wants to learn, performance goals, and learning goals. He then reviews himself at the end of each month. He also asks himself every night if he was better today than yesterday. He said the answer is usually "no." However, it's rare that he answers "no" two days in a row. Pink said this book has changed his personal life too. He now believes allowances for children should not be given based on the completion of chores. He believes chores are a moral obligation and shouldn't be compensated through economic transactions.

14.2.2 Chapter 4 Social Physics- Engagement

Alex Pentland discusses how best to get individuals to work together. According to the text, "working together requires more than shared habits; it requires habits that result in cooperation." But how do you actually get people to work in a group effectively? Many of these patterns can be witnessed in animals and primates. Examples include signaling mechanism such as body postures and vocalization are aspects of the decision-making process within groups.

In Bob Kelly's Bell Stars study, Kelly analyzed the difference between average and star performers. He found that star performers encouraged and pushed everyone in their group toward "joint ownership" of the goal, strategy, activities, etc. The average performers just did their part of the group work.

But how do you translate this type of engagement from face-to-face interaction to the digital space? One example Pentland points out is a Facebook experiment in the 2010 election cycle. A group of scientists launched a "Get Out the Vote" message to 61 million Facebook users and analyzed the impact of the different types of messages.

One group of Facebook users received a 'Get out and vote" message, while others received a "Vote" message and saw the faces of friends who has already voted. The experiment revealed that those who received the "Vote" message with faces of their friends had a higher rate of mobilization. It turned out, close friends "exerted about four times more influence on the total number of voters" than the message itself.

Why did this impact behavior? Pentland argues that it is Social Pressure. Seeing our peers adopt certain habits and behavior is motivation to join in. This is because these are trusted sources. So how do you use social network incentives (an incentive to alter interaction between people) to change behavior and increase cooperation?

Pentland states that by creating social pressure and "increasing the amount of interaction around specific, targeted ideas" creates an environment where people are more likely to adopt these habits. One example used to show social network incentives is "FunFit" which encouraged friends and family members to remain active. Everyone was assigned to two buddies, some were people they interacted with a lot, others were acquaintances. They created clusters that centered around one target person and they were given a small cash reward based on the behavior change of the main person, thus creating social pressure to be more active. For those who had more direct interaction with their target, the social network incentive worked nearly 8 times better than the standard market approach. Also, those who received social network incentives maintained high levels of activity after the incentives went away.

The amount of direct interaction between people predicts shared trust, effectiveness of peer pressure and influence to create behavior change.

In an energy conservation experiment mentioned in the book, homeowners received feedback on how much energy they consumed in comparison to the average person. The result, no changed behavior. When the comparison was between a homeowner and people in their neighborhood, behavior changed. Pentland explains that this is the "social network effect" because identifying with a group of people increases trust and the social pressure and influence of the group. This strategy can be adopted to digital networks as long as there is trust among peers.

Pentland cautions that just as the prior experiments showed cooperative interactions, if the majority of interaction within a social network is combative and negative it will destroy trust. He gives the example that some people dislike politicians and lawyers as a group but may enjoy specific individuals. This type of behavior creates a "don't trust them" mentality which can create discrimination between groups.

14.2.3 Sutherland, J.J. (2014) Scrum: the art of doing twice the work in half the time. CH1

The FBI had one goal and one goal only when trying to develop a new data system, prevent another 9/11. The inability of the department to track and share information failed to show Al Qaeda terrorists entering the country because there was no effective mechanism for capturing or sharing its institutional knowledge.

Senator Patrick Leahy of Vermont told the media: "We had information that could have stopped 9/11… was not acted upon… I haven't seen them correct the problems." In 2005 the FBI announced a new program, Sentinel, costing $451 million, to be operational by 2009. By 2010 the program, being built by Lockheed Martin, needed an additional 350 million. That's when to project was terminated and the group decided to go in a different direction.

The problem wasn't resources or the intelligence of the people working on the project, it was the "way" people were working together. When a plan doesn't work, people need to be flexible and modify, not pretend as though the plan is working.

They were using Henry Gantt charts which looked good with colors and graphs. Gantt invented his charts in 1910 which were used in World War I by General William Crozier. The trouble was, managers hired people "to make it look as if the plan is working. Essentially, they're paying people to lie to them. Johnson felt that they had smart people in place, but the systems were failing due to the way people were working. He wanted to remove all impediments. Ohno's Toyota Production System formed many of the ideas behind SCRUM. Ohno states "production should flow swiftly and calmly throughout the process, and one of management's key tasks is to identify and remove impediments to that flow." Johnston implemented SCRUM to complete the development of the Sentinel system.

The framework of the approach is simple: looking at "how" people work, rather than how they "say" they work. The term scrum comes from the game of Rugby and refers to the way a team works together to move the ball down the field: alignment, unity and clarity of goal. Traditional management operates in a world of control and predictability. But the best scenario rarely unfolds, and this approach restricts people to the "plan".

Scrum asks questions, and embraces creativity and uncertainty. It focuses on the learning process and gives people the tools to self-organize and rapidly improve speed and quality of work.

How SCRUM Works: It begins with an "Inspect and Adapt" cycle, a "Honey do List", setting sequential goals that must be completed in a fixed length of time usually 2 weeks.
This step involves: * Setting a date to show a finished increment of product to stakeholders, users of the product * Providing real time feedback

At the beginning of each cycle there is a meeting to plan the "Sprint". Each member decides how much work they can get done in two weeks. The items on their priority lists are all tallied and documented with sticky notes. Members begin to have a baseline of how much they can complete in two weeks – their velocity. The team gets faster and faster by eliminating impediments.

With this strategy, the FBI was able to create Sentinel in 20 months total. They joked that they were able to create something with 5% of the budget that Lockheed had and couldn't develop in over 10 years. The system is used for everything, sorting evidence, paying informant, case files, meetings, etc. By using the SCRUM approach they measured progress, demonstrated the product throughout the process and were transparent about their progress.

SCRUM has made a dramatic difference for the FBI and the ability to communicate and share information has changed how the Bureau operates. Due to SCRUM, a million-dollar transfer to another country, bank

approved, was stopped.

14.2.4 Duhigg, C. (2016) "What we learned from Google's Efforts to Build a Perfect Team." The New York Times Magazine, Feb 25, 2016

Before Julie Rozovsky was hired by Google to study people's habits and tendencies, she attended Harvard where she was associated with two groups of people; a study group and a case team. Although the study group was comprised of intelligent, productive, and competitive members, she found the daily group meetings to be very draining and stressful. Her case team members had a variety of professional careers, but they "clicked" well and talked with each other in a non-formal, casual way on a personal level. Although studies have proven working in teams or groups results in faster innovation, better results, and reports of higher job satisfaction, Rozovsky was bewildered why the study group and case team experiences were so different.

Top executives used to believe that teams should be made up of like-minded individuals, i.e. introverts working together, members should be friends outside of work, etc. Many firms have come to realize that analyzing and improving individual workers, a practice known as "employee performance optimization," is just the beginning of increasing productivity. Researchers found the most productive teams possess: * Equality in distribution of conversational turn-taking. If one person did all the talking, collective intelligence declines. * Average social sensitivity. Members are skilled at intuiting how others feel based on tone of voice, expressions, and nonverbal cues.

Members of a team are most comfortable being themselves, which fosters creativity, when members show mutual respect and trust with one another. People who scored high on a social sensitivity exam were most successful working on a team. The social sensitivity exam comprised of how people interpreted pictures of sets of eyes.

Julie was assigned to Google's Project Aristotle to review how teams work. "Project Aristotle researchers concluded that understanding and influencing group norms were the keys to improving Google's teams". Teams operate best when the following are present: * Psychological safety * Clear goals * Creating a culture of dependability

Another Google employee, Matt Sakaguchi formerly from Silicon Valley, linked psychological safety to emotional conversations. He states a bond between team members is needed to establish psychological safety, conversation turn-taking, and empathy. The key is to "figure out how to create psychological safety faster, better, and in more productive ways." Project Aristotle is now encouraging emotional conversations and discussions of norms to create effective teams.

14.2.5 The Book Smarter Faster Better-The secretes of being Productive in Life and Business by Charles Duhigg, Chapter 5 Managing Others

This is the story of Frank Janssen who was kidnapped because his daughter prosecuted a high ranking gang member, Kevin Melton. This case demonstrates how people being empowered and people working together, combined with the FBI's Sentinel system, and led to success.

Key Takeaways: * "Employees work smarter and better when they believe they have more decision-making authority and when they believe their colleagues are committed to their success." * "A sense of control can fuel motivation, but that drive to produce insights and innovations, people need to know their suggestions won't be ignored."

Building a culture of commitment and trust is key when a great idea comes along. "The biggest mistake is when there is never an opportunity for an employee to make a mistake."

14.3 Key Take-Aways (for Yellowdig)

14.3.0.1 What Motivates Us:

This article is an interview with Daniel Pink regarding his book titled Drive: The Surprising Truth About What Motivates Us. He says the carrot & stick motivation concept doesn't work for business. He states managers should motivate their members by giving them a "sense of making progress." This can be done by helping members see their progress, by recognizing it, celebrating it, and keeping people on a path towards making more progress. He also says it's a combination of intrinsic and extrinsic motivation factors that keeps people motivated. He firmly believes members should be paid enough so that one can entirely focus on work and not worry about money. Pink said the writing and research he did for his new book has changed his work habits by establishing helpful measures to making progress for himself. He developed daily work habits such as a monthly "do it yourself performance review" and he also asks himself every night if he was better today than yesterday. His answer is usually "no". He says he doesn't answer "no" two nights in a row because he subconsciously wants a better day after the next day, so this daily questioning of himself improves his life.

14.3.0.2 Social Physics Engagement

According to the text, "working together requires more than shared habits; it requires habits that result in cooperation." Pentland argues that "social pressure" is a key tool to change behavior and help people work together. Seeing our peers adopt certain habits and behavior is motivation to join in. This is because these are trusted sources.

Rules of Engagement: Three key things to remember * Engagement requires interaction * Engagement requires cooperation * Engagement requires trust

14.3.0.3 SCRUM:

This is the story of Jeff Johnson, Assistant Director of the IT Engineering Division for the Federal Bureau of Investigation (FBI). He initially worked on a system, which was very outdated, antiquated and porous, called the Automated Case Support System. When the FBI tried to upgrade to a Virtual Case File (VCF) system, it fell apart at the expense of $170 million in taxpayer money. Johnson felt that they had smart people in place but the systems were failing due to the way people were working. He wanted to remove all impediments. Ohno's Toyota Production System formed many of the ideas behind SCRUM. Ohno states "Production should flow swiftly and calmly throughout the process, and one of management's key tasks is to identify and remove impediments to that flow." Johnston implemented SCRUM to complete the development of the Sentinel system.

Key Takeaways: * Planning is useful. Blindly following Plans is Stupid.
* Inspect and Adapt. Like a "Honey do List." In this step a team sets sequential goals, a prioritized list of tasks each member will complete in a 2-week cycle, called "Sprints." The team will start to build "velocity" – knowing how much they can achieve in a 2-week period, and their work continually progresses faster and faster because they are more productive. The plan is constantly reviewed to determine "what brings value to the project." At the end of 2- weeks, the team presented a "Demo" for all stakeholders. * Change or Die. Don't cling to the old ways of doing things. Be open-minded and creative. * Fail Fast so you can fix early. Working in product short cycles "sprints", 2 weeks, allows early user feedback and can eliminate wasteful effort.

In regards to the Sentinel project - what was projected to take up 90% of the budget and ten years only took 5% of the budget and 20 months. SCRUM has made a dramatic difference for the FBI, the ability to communicate and share information has changed how the Bureau operates.

14.3.0.4 "What we learned from Google's Efforts to Build a Perfect Team":

This article is about Julie Rozovsky she was hired by Google to study people's habits and tendencies. Julie was assigned to Google's Project Aristotle to review how teams work. Her team concluded that understanding

and influencing group norms were the keys to improving Google's teams (and motivating team members to participate). Teams operate best when the following are present: * Psychological safety-meaning members must feel comfortable being themselves. There should be an equality in distribution of conversation turn-taking. Members should be skilled at intuiting how others feel based on tone of voice, expressions, and nonverbal cues. * Clear goals * Creating a culture of dependability (trust)

Based on these findings, Project Aristotle encourages emotional conversations and discussions of norms between team members. Getting to know members on a personal basis increase team compatibility.

14.3.1 Discussion Questions

1. Have you ever found yourself buried in a project, knowing it wouldn't come to fruition, but because of the pressure from other team members you continued the course?
2. Do you feel like "chitchat" before a meeting is a waste of time, or do you think it's important to understand your team members on a personal level?
3. Have you had a situation where social pressure changed your habits or behavior?
4. What are the potential downsides to social pressure, especially in the digital space?

14.4 References

- Daniel Pink (2010). What Motivates Us? Harvard Business Review. LINK VIDEO
- Pentland, A. (2015). Social Physics: How social networks can make us smarter. Penguin. **CH4 engagement**
- Sutherland, J., & Sutherland, J. J. (2014). Scrum: the art of doing twice the work in half the time. Currency. **CH1 fixing management with flow**
- Duhigg, C. (2016). "What we learned from Google's Efforts to Build a Perfect Team." The New York Times Magazine, Feb. 25, 2016. LINK

Chapter 15

Building Effective Teams

(Team 1)

15.1 Topic Overview

The readings this week spent a great deal of time discussing social learning, idea flow, and happiness and the impact it has on productivity in the workplace, and the role that leaders can play to foster this culture.

15.2 Chapter Summaries

Team 1 William Seeley Lauren Zajac

- Social Physics: CH3 idea flow
- SCRUM: CH7 happiness
- Social Physics: CH5 collective intelligence
- Smarter faster better: CH5 managing others
- Social Physics: CH6 shaping organizations

The readings this week spent a great deal of time discussing social learning, idea flow, and happiness and the impact it has on productivity in the workplace, and the role that leaders can play to foster this culture. As managers, it is imperative that we learn how to create, sustain, and encourage these connections and interactions, as they are key to creating high performing, productive teams. Managers also need to ensure that they are creating happiness in their workforce, as happy people are more productive at work at home and in their personal life. Finally, the readings ended with an evaluation of workplace culture, and the research behind the commitment culture. This is a culture where employees know that the organization is committed to them, and they are committed to the success of the organization. We end the discussions this week with an examination of the strategies leaders can use to foster connections and develop teams who excel at collaboration, idea flow and are be happy.

Pentland kicked off the concept of building effective teams through idea flow in Chapter 3. He stresses that managers can increase productivity by the "elimination of barriers to idea flow, and the rate of idea flow" (p. 44). Team communications, engagement and idea flow help to develop a collective intelligence, which far superior to individual intelligence (p. 44). Our behavior as individuals also can be predicted based upon our exposure to the behaviors of other people, known as social learning (p. 45). This social conformity ("everyone else is doing it") can be positive or negative, and managers need to harness the idea flow to ensure a productive outcome for the company (p. 55). The adoption of habits and preferences is a slow processes that requires repeated exposure (p. 58). Pentland describes a scenario where there is a "rush of new ideas through social exploration, followed by a slower quieter process to determine which are converted into personal habits" (p. 58).

Charles Duhigg further explores the role of the manager to help develop happiness (p. 146). Duhigg argues that happiness is critical for your business, as happy people are more productive and successful (p. 147). Happiness is fostered when employees have the "ability to control their own destiny, the feeling that they are getting better at a task, and knowing that they are serving something bigger than themselves" (p. 153). Management needs to encourage these qualities, and structure the work in such a way to promote happiness. Duhigg also stressed the importance of transparency, so the team is aware of everyone, and can help address individual and group performance issues. One example of this transparency can be achieved through the use of a SCRUM board so all members of the team are aware of what needs to be done and who is doing what, and all information is shared (p. 155). Managers also need to ensure that the team does not "rest on their laurels" and get trapped in a "happiness bubble" and Duhigg recommends that happiness is measured and monitored and that managers are prepared to intervene when necessary (p. 168).

Pentland continued the exploration of building effective and productive teams by exploring collective intelligence in chapter five. He stated that one of the greatest predictors of group intelligence is when the members are equal when taking turns talking, and that no one person dominates (p. 88). High performing groups have a large number of ideas exchanged, have very quick and "dense" interactions, where ideas are "quickly validated or invalidated" and when there is a wide diversity of ideas (p. 89). Group problem solving is also enhanced when groups are more connected (p. 93). When all team members are "in the loop" there is a greater ability for people to learn from each other and to also develop more creativity (p. 96). Creating this environment of engagement and connectedness is critical to increase opportunities for "social learning, sharing of vital resources operational knowledge and developing successful work habits" (p. 104).

Duhigg presents concrete examples of these successful work cultures in Chapter 5. He discusses the concepts of "lean manufacturing" where companies actively "push decisions to the lowest possible level so that those who saw the problems first had the greatest authority to find the solutions" (p. 144). The idea behind lean manufacturing are based upon the premise that if you put people in a situation to succeed, they will. Baron and Hannan were two researchers who set out to determine which workplace culture led to the best longevity and performance in their landmark study of Silicon Valley startup companies. They overwhelmingly discovered that the commitment model culture, where workers are committed to the culture, and the company is equally committed to the employees, far outpaced any other culture (p. 149). This culture fosters a deep trust between the management and employees that leads to success and performance (p. 150). Duhigg explores variations of the lean manufacturing and commitment culture then by examining the diverse industries NUMMI Toyota automotive factory in California and the FBI. The leaders in these diverse industries were all able to achieve success by following the lean manufacturing principles of allowing those closest to a problem to make decisions, and to encourage a culture of commitment built upon trust, collaboration and idea exchange (p. 160). These examples demonstrate that employees work smarter and are more motivated when they have more decision making authority and believe that their managers and colleagues are committed to their success (p. 165).

Finally, Pentland discusses the ways leaders can improve idea flow and thus improve performance in Chapter 6, "Shaping Organizations". He argues that managers need to move away from managing organizational charts, and become managers who monitor, foster and encourage idea flow by increasing interactions (p. 106). Making group members aware of the patterns of communication between and within groups is very important. Frequent interactions and engagement creates a shared understanding of what needs to be changed, and there is social pressure for the team to adopt these "agreed upon patterns" (p. 106). Innovation is another driver of performance, and managers can help encourage innovation by helping to forge diverse connections between diverse team members to establish a number of connections (p. 107). Cooperation is key with engagement, and everyone needs to share and contribute equally (p. 111). Some managers and teams have even begun to use new tools to visualize group patterns and interactions to ensure team cohesion and cooperation, using sociometric feedback tools (p. 116). These tools allow teams to visualize their cooperation and ensure a good idea flow and improve the social intelligence of the group (p. 113). Finally, Pentland explores the personal influence of leadership, and the practical charisma of effective leaders who are both "energetic and systematically engage with others, to grow the interaction patterns in the right direction" (p. 117). He describes these leaders as "charismatic connectors" who show a genuine interest and curiosity in everyone and everything, and make people feel good (p. 118). This behavior and leadership trait can be learned as

people pay attention to new ideas, bounce those ideas off others to see their reaction, and try to expand social networks to gather as many diverse ideas as possible (p. 118).

In conclusion, managers and leaders can play an important role in shaping the cultures of their workplaces to encourage social learning, idea flow, and happiness. As managers, it is imperative that we learn how to create, sustain, and encourage these connections and interactions, as they are key to creating high performing, productive teams.

15.3 Key Take-Aways

Voicethread link:

https://voicethread.com/share/12098991/

15.3.1 Discussion Questions

1. Duhigg states that happiness is fostered when employees have the "ability to control their own destiny, the feeling that they are getting better at a task, and knowing that they are serving something bigger than themselves" (p. 153). What are some ways that you can structure your team and assignments to encourage happiness?
2. Pentland explores the personal influence of leadership, and the practical charisma of effective leaders who are both "energetic and systematically engage with others, to grow the interaction patterns in the right direction" (p. 117). He describes these leaders as "charismatic connectors". Have you ever worked for someone with these characteristics?

3. Duhigg explores the concept of "lean manufacturing", where companies actively "push decisions to the lowest possible level so that those who saw the problems first had the greatest authority to find the solutions" (p. 144). Have you encountered these tools in your workplace?

15.4 References

- Pentland, A. (2015). Social Physics: How social networks can make us smarter. Penguin. **CH3 idea flow**

- Sutherland, J., & Sutherland, J. J. (2014). Scrum: the art of doing twice the work in half the time. Currency. **CH7 happiness**

- Pentland, A. (2015). Social Physics: How social networks can make us smarter. Penguin. **CH5 collective intelligence**

- Duhigg, C. (2016). Smarter faster better: The secrets of being productive. Random House. **CH5 managing others**

- Pentland, A. (2015). Social Physics: How social networks can make us smarter. Penguin. **CH6 shaping organizations**

Chapter 16

Managerial Experiments

(Team 2) Matthew Simon and Carlos Lopez

16.1 Topic Overview

As we have continued to dive into the different aspects of data and technology in the public sector, we have been challenged to think about its applications in real life experiences. This collection of readings starts to challenge our thinking of how data, social networks and technology can impact the workplace and societal change.

Two of the readings (Social Physics and Reality Mining), really leverage each other and focus on the power of collective groups to increase societal change. For example, they discuss and analyze the power of social networks. In Social Physics, the focus is how social networks can create instant organizations that actually work effectively without having to create complex bureaucracies. In Reality Mining, they make more closely aligned applications to the workplace and social change with the use of data collection. One of the key connections for me is that they both include a sense of opting in. I mean this in a way that there was a sense of autonomy in contributing to this collective approach. One of the potential problems from a public sector government perspective is when these approaches no longer contain an opt-in function and they become compulsory.

The other two examples from Iceland and the schools are little bit different and show how data can be used for effective programs and outcomes. For me, these had a more practical approach of program analysis and implementation that could be used. I think that they show a clearer approach of how to use data effectively in effecting desired outcomes.

One thing that I think should be challenged was the notion of human interest and participation in Social Physics. They almost start to make the connection of a socialist economy and society and social networks show the altruism of wanting to do good or contribute. I would challenge human nature a bit more on this perspective. They even reference a political scientist to make this connection, but I think that they do so incorrectly. Humans want connection and purpose, but I don't know how sustainable the power of the "red balloon" exercise is past achieving that one simple outcome for a short period of time.

16.2 Chapter Summaries

1) Duhigg, C. (2016). Smarter faster better: The secrets of being productive. Random House. CH8: This chapter gives a variety of practical real-life examples of how people absorb data. From examples in the school system, which were discussed in the topic overview, to examples about people being able to choose retirement accounts. He uses all of these examples to show that people need to be able to absorb and digest data in an effective way in order to process it and make a decision. He calls the human ability to make these

choices and breakdown data as scaffolding and winnowing. When people are able to process data effectively it has huge implications for the impact that it is able to have on business operations and even the lives of students.

2) Social Physics: How social networks can make us smarter. CH7: This section of reading focuses on how social network incentives can be used to create organizations and be used for other purposes. They use an example of Red Balloon Challenge and how they used social media to create an immediate network to solve an almost impossible task. The break down this idea of social connections and the power behind them into three key components: instant organizations, organizations in stress and trust. The first component basically describes how utilizing social networks allows you to easily and almost instantly create a functioning organization of people all moving to the same purpose without really having to do much organizing. Consider Wikipedia, an immense online encyclopedia built by volunteers that is often cited as a great example of crowdsourcing; while many people contribute content there was a core group of dedicated editors who worked for years to organize the content as it was added who were driven by social network incentives rather than monetary (p.125). Pentland further takes this idea and talks about how the strength of the social networks are resilient when stress is introduced and how that helps the organization continue to function when something is injected, like layoffs. Immediately after layoffs, the level of engagement between employees skyrocketed as they began to adapt to the new situation by generation new patterns of interaction (p. 128). Interestingly, it was the employees with the highest level of engagement before the layoffs that had the easiest time adapting to the new patterns of interaction (p.128). The final component discusses trust and the critical social ties that exist because these people already have a bond and know each other from some previous connection and the power behind that social capital with one another. Cooperation is most effective when it leverages preexisting personal social ties, and the more active the social tie, the greater the level of cooperation (p.129). Building strong ties with people is good for idea flow, but strong ties also can be used to exert social pressure (p.129). This connection between engagement, trust, and people's ability to act cooperative highlights the relationship between civic engagement and the health of society (p.130). We are traders in ideas, goods, favors, and information and not simply the traditional market thinking would make us (p.130). The section of reading ends with contemplating how this can be used in practical larger scale decision making and organizational structures like cities and larger societies.

3) Reality mining: Using big data to engineer a better world. CH4: This section of reading discusses the implications of using user generated data to create social change for the benefit of the provider of the data. The author uses two lenses to show how this can translate in the office space or in the city context. In the office context, they discuss how people can basically opt-in to a system where if they want to information share in order to support work efforts they can do so. It would be mutually beneficial because both parties have to opt-in in order to protect the privacy of both individuals. This is basically Facebook for an internal office. The authors provide a workplace example about Tacit a system that analyze employees' work and communication behavior, helping them to make new connections with each other, sharing of knowledge, interest, and networks all while ensuring everyone's privacy (p.70). In large organizations with thousands of employees this system could be helpful to connect people working on similar projects but are located in different office locations allowing them to create new opportunities that otherwise would have come about. The other component this translated to was in the city context. They discussed a variety of user applications that allow city residents to disclose or collect certain data points that can be used for the cities overall benefit. For example, this includes developing better transportation systems, cleaner streets and better services for residents. There is FixMyStreet app that allows people to send geotagged, time stamped photos to their city government (p.76). In some cases these photos connect to the workflow system and create work orders! Other applications allow users to share the best biking/jogging routes to other users. These efforts can improve citizen's quality of life utilizing tools and mining projects. The main drawback from all of these benefits that was discussed was about the need for privacy and how the user needs to proactively opt-in to sharing this information. A parallel was drawn with dating applications and how people choose to share certain aspects of their personal life for their profiles and data about their locations. However, there are potential drawbacks to the overall benefits and these privacy concerns need to be discussed before a large-scale use and approach is used by government.

4) BBC World Service: "How Iceland Saved Its Teenagers" November 11, 2017: This section

of content was a podcast from BBC in which the authors explore policies used in Iceland are potentially linked to the decrease in teen drinking and smoking. Primarily, Iceland used anonymous survey data of high school students to better understand their habits and identify high-yield protective factors into their everyday lives. To determine the protective factors that would be most effective they needed timely and accurate data. All Icelandic high school students would fill out an annual survey, which outlines their risk factors and how often they were engaging in the undesirable behavior. Honesty in the survey was key, which had a high rate because it is anonymous. This information and data were then fed back into local communities to implement programs and policies, which would increase the amount of activities which students were engaged in that were considered protective factors. This approach covered about 85% of young people in Iceland. Some of the protective factors included: spending time with parents, physical activity and organized activities, implementing a curfew for kids under 17 determined by parents and cultivating relationships between parents.

16.3 Key Take-Aways (for Yellowdig)

16.3.1 Discussion Questions

1. Eagle and Greene provided a workplace example about Tacit a system that analyzes employees' work and communication behavior (including emails), helping them to make new connections with each other and improve sharing of knowledge to improve communication and efficiency. Does your workplace have a similar system like this or would you be interested in participating in one?
2. Pentland explains cooperation is most effective when it leverages preexisting personal social ties, and the more active the social tie, the greater the level of cooperation (p.129). What is an example where you have experienced this?
3. Do you have any social networks that are available at work that make your work life better and more productive? Are they mandatory or optional? Why did you choose to join one of them in the first place? How does it contribute to your work?
4. In Iceland, they used local data to make decisions about how to reduce teen drinking and smoking. They talked about data being quote "fresh" and being used quickly and timely. Do you feel that the data you use in your work environment is fresh and timely? Or is it old and stale by the time it gets through the bureaucracy? What are the implications of this if it is the latter?

16.4 References

- Duhigg, C. (2016). Smarter faster better: The secrets of being productive. Random House. **CH8 Charlotte Fludd pp 247-252, the data room 252-256**
- Eagle, N., & Greene, K. (2014). Reality mining: Using big data to engineer a better world. MIT Press. **CH4 engineering public policy**
- Pentland, A. (2015). Social Physics: How social networks can make us smarter. Penguin. **CH7 organizational change**
- BBC World Service: "How Iceland Saved Its Teenagers" November 11, 2017. LINK

Chapter 17

Amazon vs Zappos

Team 3: Marcela Morales and Joseph Lynch

A Tale of Two Data-Driven Management Systems: Amazon and Zappos

17.1 Topic Overview

The articles below examine the use of data and their effects on management. From the tightly controlled business at Amazon to the more freely run Zappos, these articles attempt to show the use of data on measuring employee effectiveness, employee retention, and morale. Both Amazon and Zappos are highly successful, but each has a very different philosophy. These different environments are even more strange since Amazon owns Zappos!

17.2 Chapter Summaries

Sutherland, J., & Sutherland, J. J. (2014). Scrum: the art of doing twice the work in half the time. Currency. CH7 happiness

In the chapter "Happiness" the author explores the impact of attitude and happiness on business. Sutherland and Sutherland assert that happiness happens when pushing yourself. Happiness of the team is crucial to business. Further examined, the authors states that "happiness leads to success in nearly every domain of your life." Happiness is not the outcome of being successful, it "precedes" it. In the SCRUM process, the goal of the team is not to assign blame when things don't work as intended but rather to review the process and find the solutions, as a team. The authors instituted a concept called the happiness metric, plainly put, it identifies how the employee feels and identifies what one thing would make them happier. As noted, autonomy, mastery, and purpose are a core component of employee happiness. Sutherland and Sutherland then examined Zappos and the core concepts of happiness that made the company successful. To achieve Wow! Moments for the customers and make them happy, Zappos felt the employees had to be happy as well. Connection to others is a major component of the strategy. Zappos has a 12% turnover rate, and this is due to letting employees go who lack happiness as measure by customer interactions and complacency. To ensure new employees understand the culture, they are put through a 4 week "boot camp" to ensure they understand the company's philosophies and culture. Sutherland and Sutherland used Zappos to show that happiness is a truly effective approach to business. The chapter is summarized in 8 key points. It's the Journey and not the Destination. Happiness is the New Black. Quantify Happiness. Get better everyday and measure it. Secrecy is poison. Make work visible. Happiness is autonomy, mastery, and purpose. Pop the Happy Bubble. Zappos made these ideas work and grew from a $1.6 million dollar company in 2000 to a billion dollar company in 2008.

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books. CH7 sweating bullets

The article in Weapons of math destruction is broken into 3 distinct subchapters. The first section deals with how data is used to impact the workforce. In the article, workers in the food service and retail industry were examined. The article asserts that 2/3 of food service and ½ of retail workers receive their schedule less than one week before it starts. This is done with the help of scheduling technology using Operations Research (OR). The technology matches employees to the busiest times, hours available, and other business needs. As a result, employees have a challenge maintaining a routine which inhibits activities like attending college, spending time with children, or managing child care. The effect on their children is quite similar, a lack of routines. The varied schedule was linked to sleep deprivation and heightened anxiety. Some employees have to close a business and open it the next morning, termed "clopening." This technology also works to limit employees less than 30 hours, so they often do not receive benefits. The inability for these workers to advance themselves enhances the inequalities in the system while maximizing profits for the business. The Second section is focused on employee metrics. A company called Cataphora designed algorithms that could scour emails and other data to identify the best idea generators in the workforce. These algorithms would rank employees for management, so they could determine, through technology, the best workers. During the recession of 2008, this data was used to identify who to lay off. This technology used assumptions and often without other types of manager input. The technology had a low correlation to employee performance. The potential outcome of this technology could have been the culling of very effective employees from the businesses. Cataphora went out of business. In the third section, a review of the testing of students and monitoring of teachers in response to the 1983 report "A Nation at Risk" by the Regan administration. The report warned of a "rising tide of mediocrity" in our nation's schools. Tests were used to determine teacher effectiveness with one major flaw. The tests were not measured on an absolute scale. Scores were inaccurately tied to school type, social inequality, and a failed try at fairness. Teachers scores varied wildly from year to year although the learning plans and lessons taught remained the same. One teacher examined in the article scored a 6 out of 100 one year and 96 out of 100 the next year. Nothing about his style had changed. In addition, the reporting of lower SAT scores was examined. The total overall score had changed, however the assumptions were incorrect. Over time, the number of students taking the test expanded greatly. When each subgroup of test taker was examined, each area had shown improvement which is called Simpson's Paradox. The total was lower, but each subgroup improved.

"Zappos: A Workplace Where No One And Everyone Is The Boss." NPR, July 21, 2015. PODCAST

This article discusses the idea of "holacracy" and how it was successfully implemented at the Zappos company. The idea behind it was to create more autonomy in the employees while maintaining the small-firm culture in a big-firm environment. Holacratic governance is compared to cell biology. Each of your cells works on its own without asking other cells for permissions to do anything. He seeks adaptability and nimbleness in his companies that focuses on the groups interests and not individual interests. This workplace is more employee centered with the idea that happy employees have a "trickle down" effect when it comes with handling customer requests.

Richard Feloni (Feb 19, 2016). "A former Zappos manager explains how her job changed after the company got rid of bosses." Business Insider. LINK

This worked well for Zappos because without the structure of a hierarchy employees are empowered and understand the meaning of their work and have purpose instead of menial goals. Holocracy is a style or work that relies on autonomy and self-governance. When Zappos decided to make the change to a holocratic environment, bosses were no longer needed. Zappos gave each employee the ability to opt out of this culture shift by offering a buyout of those who didn't want to "participate." 14% of the workforce took this offer. Each member joins teams to solve issues or create programs. Every employee has an equal say and the decisions are made where the information is rather than through a hierarchal structure.

17.3 Key Take-Aways (for Yellowdig)

Two companies, Amazon and Zappos have entirely different cultures, but both have success.

At Amazon, the corporate culture uses statistical and employee-initiated data to improve performance. Employees can a report other employees anonymously when they think they are not working hard enough. Turnover is high, the stress is high, and the ability for employees to sabotage one another is high. The culture brings in highly ambitious employees who work as rivals to achieve results. The articles paint Amazon as a very tough culture in which to succeed and quite unforgiving. Even with this tough work environment, Amazon has become the largest retailer in the world.

Zappos is a company that believes that happiness of the employees leads to happiness of the customer. Zappos does not use a hierarchal approach, rather, they use "holacracy" which gives employees autonomy to make decisions that create Wow moments for their customers. Zappos has no managers, they use teams to make decisions and most closely resemble the SCRUM approach. Zappos has grown significantly and is also quite successful. Zappos invests in employee culture and dedicates 4 weeks of "boot-camp" for every new employee. Zappos, in terms of culture, is the exact opposite of Amazon. Surprisingly, Amazon owns Zappos.

17.3.1 Discussion Questions

Discussion Question 1: Do the risks of bad data or assumptions outweigh the benefits of using data to identify workplace effeciencies or employee performance?

Discussion Question 2: Would holacracy, where there is no management, and a design that emphasizes autonomy and self-governance work in your agency? How would this benefit your Agency and what risks would it bring?

17.4 References

- Sutherland, J., & Sutherland, J. J. (2014). Scrum: the art of doing twice the work in half the time. Currency. **CH7 happiness**
- O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books. **CH7 sweating bullets**
- Kantor, J. & Streitfeld, D. "Inside Amazon: Wrestling Big Ideas in a Bruising Workplace." The New York Times, August 15, 2015.
- "Zappos: A Workplace Where No One And Everyone Is The Boss." NPR, July 21, 2015. PODCAST
- Richard Feloni (Feb 19, 2016). "A former Zappos manager explains how her job changed after the company got rid of bosses." Business Insider. LINK

Chapter 18

Manipulating Crowds

(Team 4)

18.1 Topic Overview

History is filled with stories of the old salesman that would enter town and then sell snake oil or other mysterious elixirs that promise to cure any and all ailments. Even further back are those that proclaimed to be purveyors of religious relics. In 1957 the musical "The Music Man" was released that detailed the story of a predatory salesman that entered a small town, found out about the people, that they had a new pool table, and then exploited that data to sell musical instruments for a boy's band.

While the methods really haven't changed, the tactics have become more sophisticated. Predatory salespeople are using big data to identify, manipulate, and take advantage of people - usually those that are poor, ignorant, or in desperate circumstances. Joanna Redden authored an article identifying six ways that big data is harming society (Redden, 2017). She lists those as:

1. Targeting based on vulnerability
2. Misuse of personal information
3. Discrimination
4. Data Breaches
5. Political manipulation and social harm
6. Data and system errors

Vulnerability:

Predatory marketing sought to target those that were vulnerable long before big data. Sales or marketing people, past and present, customize their message and apply pressure to those that may be the most desperate circumstances. This could be an auto mechanic in a small town inflating prices because a person needs their car fixed right away, weight loss pills, or debt consolidation. Big data has allowed businesses to profile and find those that are desperate for help and then prey upon them.

Misuse: In 2009, Intuit the parent company of TurboTax and Quicken financial software packages, acquired Mint, a web-based financial planning software. With that Intuit acquired the budgets, the spending habits, and financial goals for Mint's "1.5 million users tracking nearly $50 billion in assets and $200 billion in transactions" (Wortham, 2009). With that information combined with information that it has already collected through TurboTax and Quicken, Intuit has the potential for an incredible depth of information regarding people's finances. They are even asking users to share their data for services such as credit card offers and student loan refinancing. Even without permission there is very little that can stop Intuit from using the data to profile people or sell their data to other businesses.

Discrimination: "A 2017 investigation by ProPublica and Consumer Reports showed that minority neighbourhoods pay more for car insurance than white neighbourhoods with the same risk levels" (Redden, 2017). Data from social media and census records can provide a glimpse into a predominant demographic of particular neighborhoods. That information is used to profile and discriminate regarding the availability and cost of goods and services. The use of digital fingerprints and facial recognition further adds data that categorizes people.

Data Breach: In 2017, Equifax, one of the three national credit reporting agencies reported to the Federal Trade Commission of a data breach that "exposed the sensitive personal information of 143 million Americans" (FTC, 2017). What's even more concerning is that the guilty party has yet to be found (Fazzini, 2019). Credit reporting agencies hold a tremendous amount of personal information. There is a tremendous risk if holders of sensitive data is subject to data breaches.

Political Manipulation: The readings consisted of Cambridge Analytica and how it profiled people using social media and played a role in the 2016 presidential elections. Additionally, and whether or not it helped, Clinton used data and an algorithm named Ada that was relied upon to identify states and counties that would play a critical role in campaigning (Wagner, 2016).

Data and System Errors: The role of data is only as good as the analysis and interpretation. Both Trump and Clinton used data and computer algorithms to improve the efficiency of their campaign time. However, it is only as good as the programmers make it to be. "Like much of the political establishment {Clinton's} Ada appeared to underestimate the power of rural voters in Rust Belt states" and ultimately downplayed the importance of Wisconsin and Michigan which Clinton didn't make an effort and lost (Wagner, 2016).

The use of Big Data has only sharpened the tools that markets use to be able to target individuals. "This [Big Data] establishes a powerful basis for legitimate ad campaigns, but it also fuels their predatory cousins: ads that pinpoint people in great need and sell them false or overpriced promises" (O'Neil, 2016, p.70).

Our readings this week come primarily from Weapons of Math Destruction and the harm that big data is causing to society. Chapters 4 and 10 of Weapons of Math Destruction go into detail on many of Ms Reddin's six aspects. The article The Data that Turned the World Upside Down details how Cambridge Analytica and big data was used to influence the presidential election in 2016.

18.2 Chapter Summaries

18.2.1 Weapons of Math Destruction, Chapter 4: Propaganda Machine, Online Advertising

Highlights the practices of for-profit colleges/universities that spend millions of dollars in recruitment, promise a student free money (federal student loans) and lock them into a student loan that the person can't repay with a diploma that is next to worthless. Research done by CALDERA/American Institutes for Research created 9,000 fictitious resumes. Some with high school diplomas, some with community college degrees, and others with degrees from for-profit colleges/universities. They sent these fictitious resumes out to job postings and measured the response. They found that diplomas from the for-profit colleges were worth less than community colleges and about the same as a high school diploma. (O'Neil, 2016, p.80)

O'Neil discusses how these colleges utilize Big Data to either generate or purchase leads for targeted marketing. Increasingly, such entities are using the internet to identify prospective student's personal habits and unmet needs to produce marketing aimed at addressing his/her apparent vulnerabilities. Unfortunately, this leads to predatory tactics which exploit vulnerable populations. The wealth gap is related to this practice of predatory sales where "the bottom 40 percent … has a net debt of $14,800, much of it in extortionate credit card accounts" (O'Neil, 2016, p.81). "They sell them the promise of an education and a tantalizing glimpse of upward mobility - while plunging them deeper into debt" (O'Neil, 2016, p.81).

Even worse, the for-profit colleges are teaching employees to leverage this information to secure paying students. "A 2012 Senate committee report on for-profit colleges described Vatterott's [a career-training institute] recruiting manual… It directs recruiters to target 'Welfare Mom w/Kids. Pregnant Ladies. Recent

Divorce. Low Self-Esteem. Low Income Jobs…" (O'Neil, 2016, p. 72). The list continues to name other populations. Other colleges acquire leads from third-parties and immediately begin phone-based outreach campaigns to secure new students, especially those likely to be eligible for federally backed student loans.

The Big Data ad campaigns grow in sophistication as they continually test alternative ads for success. Successful campaigns are measured by clicks and other lead generation tools. They are further refined and retried against alternate campaigns until the optimal messaging is achieved. This testing and refining is often conducted by intelligent machines, "The algorithm finds patterns on its own, and then, through time, connects them with outcomes. In a sense, it learns" (O'Neil, 2016, p. 75).

18.2.2 Weapons of Math Destruction, Chapter 10: The Targeted Citizen, Civic Life

O'Neil shares the story of how Facebook researchers have utilized the social media giant's algorithms to refine individual's feeds and test their subsequent posts. They altered filters to reveal either more positive or or more negative posts and observed subsequent activity. They found this alteration lead to corresponding increases in positive activity for those receiving positive feeds and the opposite for those receiving more negative feeds. O'Neil notes people on Facebook don't understand that their news feeds are being manipulated. Many think that they are seeing everything that their friends are posting (O'Neil, 2016, p.183).

Beyond Facebook manipulation, data gives political campaigns the data they need to be even more meticulous in their direct messages and in a way where they can be multiplicitous in speech and be more likely to get away with it (O'Neil, 2016, pg.187-188). Rayid Ghani performed an experiment with Accenture Laboratories in Chicago where he studied the shopping habits of people in a grocery store. By doing this, he was able to identify groups of shoppers that catered to discounts, were prone to advertisements, were impulse shoppers, or that shopped organic produce. Using this data, he could predict how many people could be influenced in their shopping habits, and then customized a marketing message. He then took it a step further and adapted his experiment politically for the 2012 Presidential campaign. He published outreach on LinkedIn, recruiting other data scientists to "help guide election strategy" (O'Neil, 2016, p.188). He and his team refined data mapping techniques originating in the grocery store marketing aimed at converting shoppers sitting on the proverbial fence. In this way, he ended up identifying and marketing up to 15 million swing voters (O'Neil, 2016, pg.189-190).

The chapter also describes how similar micro-targeting is used in campaign financing. Data scientists are able to utilize information mined from massive datasets to profile potential donors. They refine the messaging for these targets, hoping to entice donations and avoid offending their particular sensibilities. In this way, politicians may be more open and candid with certain donors without offending other populations who will never know the candid tale. Data is not only used for campaigns but to "deliver ideological bombs that politicians will only hint at on the record" (O'Neil, 2016, p.194).

Deep dossiers are kept on everyone and are used to influence shopping, political campaign, ideologies, a frankly shape the decisions that we make in our lives.

18.2.3 The Data that Turned the World Upside Down

Cambridge Analytica is a Big Data marketing company which has been associated with President Trump's election campaign and Great Britain's exit from the European Union (Brexit). While the Big Data analysts in President Obama's re-election campaign were said to have developed micro-targeted marketing based on supermarket research, Cambridge Analytica was claimed to have developed personality and preference based profiles of all adults in the U.S., using the data for very targeted electronic marketing during the 2016 presidential election.

Alexander James Ashburner Nix, spokesperson for Cambridge Analytica, says of their marketing strategy, "Cambridge Analytica buys personal data from a range of different sources, like land registries, automotive data, shopping data, bonus cards, club memberships, what magazines you read, what churches you attend… in the U.S. almost all personal data is for sale" (Grassegger, 2017). The company then matches this data and aligns with voter information and the personality profile to identify the target market. Specific campaign ads

are then developed and launched to play on the profile of the individual as expressed through their revealed preferences. Voters are typically unaware they are receiving this narrow and very targeted marketing. They simply believe the candidate to be well aligned with their beliefs and needs.

18.3 Key Take-Aways (for Yellowdig)

Increasingly, Big Data is being used to profile individuals based upon their demographics, purchasing habits, social-media presence, known associates, and other readily available information. In many ways, Big Data is being used to enhance the world in which we live. Unfortunately, it is also being used in ways unknown to most of the population to manipulate everything from elections to college attendance through micro-targeted (personalized) marketing. Individuals often believe they are receiving the same message as everyone else and do not realize their specific vulnerabilities are being targeted by the modern version of snake-oil salesmen.

18.3.1 Discussion Questions

1. What is the best mechanism for oversight of the micro-targeted marketing practices in the for-profit college business? No oversight? Further restrictions on federal funding? Public outcry? Other methods?
2. There is a national student loan debt of $1.5 trillion across 44 million people with 10.7% in default (Friedman, 2018). It appears that much of the US student loan debt is from federal funding and linked to degrees with little market value. With the predatory nature of "for-profit" colleges and their lack of value according to O'Neil in Chapter 4 of Weapons of Math Destruction, should student loan availability to "for-profit" colleges and universities be tied to performance value of the degree? To what standard?
3. How can individuals be trained to recognize the signs that they are the subject of micro-targeted marketing? Further, how can you avoid being biased by such messaging?
4. Should politicians in the American political system be subject to regulations regarding the disclosure of micro-targeted marketing tools used in their campaigns.

18.4 References

- Fazzini, Kate. (2019, Feb. 13). The great Equifax mystery: 17 months later, the stolen data has never been found, and experts are starting to suspect a spy scheme. CNBC. Retrieved 2/14/2019 from: LINK
- Federal Trade Commission (FTC). (2017-2018). The Equifax Data Breach. Retrieved 2/14/2019 from: LINK
- Friedman, Zach (2018, June 13). Student Loan Debt Statistics In 2018: A $1.5 Trillion Crisis. Retrieved 2/14/2019 from: LINK
- Grassegger & Krogerus (2017). The data that turned the world upside down. Motherboard Jan 28 2017. LINK
- O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books. **CH4 propaganda machine**
- O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books. **CH10 targeting citizens**
- Redden, Joanna (2017, Dec. 7). Six ways (and counting) that big data systems are harming society. Retrieved 2/14/2019 from: LINK
- Wagner, John (2016, Nov. 9). Clinton's data-driven campaign relied heavily on an algorithm named Ada. What didn't she see? - Washington Post. Retrieved 2/14/2019 from: LINK
- Wortham, Jenna (2009, Sep. 14). Intuit Buys Mint, a Web-Based Finance Competitor. New York Times. Retrieved 2/14/2019 from: LINK

Chapter 19

Best Practices for Privacy

(Team 5)

19.1 Topic Overview

The concept of managing data privacy by putting it in the hands of users themselves served as a central theme in this week's readings. In Social Physics, Pentland discusses the OpenPDS system that he has created to put data in the hands of users directly, giving them control of the content of their data and its use. Brandom discusses the General Data Protection Regulation (GDPR) in the European Union which is requiring user opt-in when sharing their data and requiring greater transparency for companies using the data.

Generally this idea of making a user's data more accessible and transparent in how it is being used is progress in the right direction, however, there remains concern over helping users understand the implications of this shift in ownership. Once you put the power in the hands of the user, you have to equip them to be successful. For example, Arizona is home to one of the strongest school-choice landscapes in the nation. Parents have the option to send their child to just about any public school they desire. What has resulted is a need to help parents understand the options available to them and how to pick the best school for their child. Parents need to be critical consumers of information about schools. Instead of sending their child to the neighborhood school, they need to assess a school's philosophy of education, curriculum, assessments, principal leadership style, parent group involvement, access to athletics and electives, and more to find the best fit for their child's education. The parents who understand this information and how to use it win.

This same concept applies with regard to shifting the power of data ownership to its users. If we are going to put the control of data into their hands, then we must educate them on what the data means, why it is important, how it will be used, and how to change their data when needed. Educating the public would require a massive, multi-year public outreach, paid media, and public relations campaign. It could also spur the creation of an industry of organizations to help consumers manage their data, nonprofit and for-profit alike.

Also important is that those businesses that hold the data have practices that truly enable users to manage their data with ease. When users access the data it should be in a format that they can understand and manipulate easily, which will require some type of graphic interface, the ability for a user to create an account to have ongoing access, and request help when needed. Additionally, if users need to make changes or want to revoke access to their data, those businesses who have the data cannot make the process overwhelmingly difficult. For example related to today's environment, when users call to correct something on their credit histories, story after story illustrate that it takes multiple phone calls and months of work to get those items resolved. The process is so difficult and complicated that it is discouraging to the users and does not inspire confidence. As users have access to their data, these processes should be streamlined and made more efficient in order to support user requests.

Conversely, there are a host of other questions that arise regarding data privacy, in addition to public understanding and education, including developing technology to integrate billions of pieces of individual data, hardware and software that support such technology, guidelines or laws about usage and storage, and how to collect data on non-technology users, just to name a few. The power that personal data holds has potential to bring advancement in almost every aspect of daily life, but it holds the equivalent potential to wreak havoc. There is a give and take with data that must be considered and the forthcoming decisions regarding personal data is putting the decision burden onto the individual. Instead of an institutionalized process, each person will have to take personal ownership and accountability for their data. Much like HIPPA laws are in place to protect medical health data, the next wave of data collection looks to institute a similar concept on all personal information.

Proponents of this type of data advancement present concepts about how the data would be self-managed and controlled, allowing users to share only the elements they are comfortable with or feel is of value in trading: Do I want my mapping software to know where I am so it in return can help me navigate? As the data models become more sophisticated and the data collection becomes broader, there is concern that assumptions and conclusions will be drawn, thereby nullifying personal choice, individuality and independence. What inferences are being made about the individual? Will they be accurate? Will a user be able to refute the interpretation? Will decisions be made on data alone with no regard to context or background? These are practical and ethically based question that are worthy of proactive debate.

Often, data users are painted as the big bad wolf, but individual responsibility may be cause for suspect as well. Consider if someone is looking for new healthcare and the carriers require personal data about weight, diet, and exercise to be shared in order to consider coverage. If an individual could control that data, would they manipulate it to their advantage? If the carrier required the individual to average 10,000 steps per day would FitBits get strapped to bustling 4-years-olds worldwide, just to make the data work? Clearly this is just but one example, but the idea is that gaming the data system is a very likely possibility. With this in mind, how much regulation are users comfortable with in order to prevent the 'gaming'? Would fingerprints or retina scans be required to prove authenticity? Some may see this as a farfetched question, but the possibility is not outrageous.

Data collection and evaluation is a clearly a compelling and motivating arena. Numbers do not lie, but that cannot be said of the collectors or the users of the data. We have discussed throughout this class that data alone is not sufficient for decision making, but is there reason to be concerned that this is exactly what may happen with the rise of the personal data store concept? Advocates for this type of data instrument show how these tools can unlock a better, brighter future, not only on a personal level but for society as whole. Concerns like climate change, food shortages, and pollution are all potential areas that can be addressed and improved with the use of better data collection and more importantly, better data sharing. But there seems to be convincing reasons to be cautions as this level of data collection and distribution has equal potential for converse effects and unintended consequences. The production and circulation of data does not come without risk, and there is no such thing as having your cake and eating it too.

19.2 Chapter Summaries

Social Physics: How Social Networks Can Make us Smarter Chapter 10: Data Driven Societies Appendix 2: OpenPDS

Pentland charges into the idea of "a new deal on data" advocating that privacy concerns can be dramatically minimized if the ownership of personal data belonged to the individual. He proposes that in order to develop more effective and efficient use of big data, the information that is collected should not be stored in individual "silos" by each collecting entity but should be stored and shared by each person on an individual level. Currently, the model of data collection is that each app a person uses, each social media subscribed to, and each technology gadget used collects data singularly. This data is not shared nor aggregated. It is horded for internal use only therefore limiting it potential to be used multilaterally for bigger, more universal purposes. Pentland goes on to point out that extensive data collection is already happening online, though often unaware by the contributor. This unregulated format has been called into question and is already on

the road to making changes, or at a minimum, placing the share process in the hands of the user. Web giants such as Google have begun the discovery process and are leading the charge for change through more transparency about what information is or is not stored within their products.

The result of this idea is the concept of Personal Data System (PDS). This system would essentially function as a portfolio of your personal data. No two portfolios would be a like because they are person specific, just as fingerprints or DNA are unique to each individual. This concept already exists within distinct technology platforms- these personal portfolios are how songs, or movies, or books are suggested to users. But Pentland's concept takes all of those preferences and data points and compiles them by person, resulting in a unique data set for every individual. Then the PDS would be used individually, sharing what information, when and with whom they see fit. The information would be owned and governed by the individual.

Pentland recognizes his concepts would be a radical change from current methods and would require a society-wide paradigm shift in order to succeed. Today, the influx of big data is just too great to be managed in traditional formats of fixed confines and closed systems. Pentland argues that the stream has developed into a rushing river and the pond is just not sufficient to contain it. He advocates for a 'living laboratory' where real time data is used to test connections and new hypothesis. In essence, Pentland argues for a 'guinea pig' society that is willing to try out of the box options in order to find revolutionary improvements. In so doing, our human understanding of data and its potential will require a much deeper connection and language if society is to integrate the full potential of big data. But while societal change is quite the obstacle, Pentland and other advocates of big data offer that this change must happen if the world is to improve and conquer the problems of the future.

Reality Mining: Using Big Data to Engineer a Better World Chapter 2: Using Personal Data in a Privacy Sensitive Way to Make a Person's Life Easier and Happier

Eagle and Greene use this chapter to highlight a number of big data-based solutions that have the potential to have benefit to users; however, warn the public that these tools must be implemented with care with regard to data privacy. They share examples of data solutions that can break bad habits, like smoking or other behavioral issues that identify the triggers that lead to the habit and provide a warning to the user. They also discuss tracking personal behavior in order to send alerts when behavior is out of the norm to monitor the elderly or add to antitheft systems in cars. They also discuss employer-based health incentive programs that reduce the cost for healthcare (and sometimes provide a financial incentive to the insured), auto insurance programs that reduce premiums based on behavior, and efforts to protect victims of domestic violence.

These data-based solutions hold a lot of promise in changing behaviors, keeping people safe, and saving people money. However, Eagle and Greene caution that data ownership, legality, and other impacts to the users need to be carefully thought through and meaningfully addressed. They argue for "various approaches to data collection and analysis that offer a range of privacy options." Some options might include giving users the ability to opt out, having clear and easily understood information about how data will be shared or used, and more.

Everything You Need to Know about the EU's New Privacy Law: GDPR

General Data Protection Regulation (GDPR) is a new law in the European Union (EU) that "sets new rules for how companies manage and share user data." Because of the global nature of the Internet, the rules extend far beyond just the EU and are affecting businesses across the globe, which is why there is a sudden proliferation of "click to proceed" requests online. GDPR requires that companies get permission from users any time they are collecting personal data from those who are EU citizens. It also puts users more in the driver's seat by giving them the ability to revoke that privilege and request all of the data on themselves from the company that collects the information.

The author notes that perhaps the most important thing that GDPR does is place restrictions on sharing behind the scenes between companies. The impact is that "companies have to rethink how they approach analytics, logins, and, above all, advertising." Sites who share data with other companies now will have to bring those relationships into the open and the companies receiving the data will have to justify its use.

The law took effect in May 2018 and since then companies have been scrambling to rewrite their terms of service and contracts. Companies that violate GDPR face a stiff penalty of up to 4 percent of their global turnover, or $20 million, whichever is higher.

Brandom calls this a "sea change for how data is handled across the world" and cites that GDPR "could fundamentally flip the relationship between massive tech companies that gather data, and the users they gather it from.

19.3 Key Take-Aways (for Yellowdig)

Video: https://voicethread.com/share/12143546/

19.3.1 Discussion Questions

GDPR shifts the power of online data to the user for EU citizens. Do you believe something like this would be beneficial to consumers in the United States? Would it give you greater comfort in knowing that you had the ability to request your data and revoke the websites' ability to collect your data? Do you think it matters to the general public?

Would you sign up to have your car insurance company track your driving behavior so that you could get a discount if you were a "safe" driver? Likewise, would you sign up for your employer's health incentive program where they would help you track your personal health data in exchange for a less expensive insurance premium? Is the tradeoff of a cheaper insurance bill worth having these elements of your life tracked?

Do you have concerns that if data was self-monitored, it would be accurate and/or unbiased? If you knew you could edit your data to show a more positive or 'cleaner' version of yourself, would you? Do you think others would? If others could see your data (much like they can see your age, birthday and location on social media) would that influence what you edited and what you didn't? And if so, would the value of the data degrade, making it less impactful in decision making?

If a Personal Data System was developed, would you adopt it? Would it be as tightly guarded as a social security number? How would non-technology users navigate without a PDS?

19.4 References

- Pentland, A. (2015). Social Physics: How social networks can make us smarter. Penguin. CH10 data-driven societies, **Appendix 2 openPDS**

- Eagle, N., & Greene, K. (2014). Reality mining: Using big data to engineer a better world. MIT Press. **CH2 using personal data in a privacy-sensitive way**

- Everything you need to know about the EU's new privacy law GDPR: LINK

Chapter 20

Best Practices for Open Data

(Team 6) Tommy and Dennis

20.1 Topic Overview

These chapters discuss why the sharing and licensing of data is important and the current US government implementation of its Project Open Data.

20.2 Chapter Summaries

20.2.1 Sunlight Foundation: Open Data Policy Guidelines link

In this reading, the Sunlight Foundation has laid out some best practices for how we, as a society, can determine what data should be public, how we can go about making data public, and how to implement policy for making data public. There are 31 recommendations in this paper, all of which have merit and should be considered. I just want to highlight three that need to be considered but will be difficult, to say the least, to implement.

First, specifying methods of determining the prioritization of data release. This is one that could get very political. Due to the necessity to ensure that data will not harm any entity through release to the world, all data should have some kind of delay for release. This delay could be as short as hours, or as long as decades, depending on the type and potential harm of the data. Additionally, this is definitely an issue that would be politicized and hampered by special interests.

Second, mandate data be explicitly license-free. Where do we start? The biggest issue here is information that may be held by the government, but in some way be proprietary. The government has an interest in protecting information from which it gains an advantage against other nations. The easiest example of this is military, but there are probably other agencies that deal in data created for government use, are proprietary, and although not classified, should not be immediately released for public consumption.

Lastly, ensure sufficient funding for implementation. In a fiscally constrained environment, this is a tough nut to crack. It will need to be made explicitly clear how and why this open data vision is a requirement over and above other items in the budget. Not to mention the question of who is actually paying for it. Gaining agreement within our own country is hard enough. Gaining agreement among all the nations of the world as to how data will be used, maintained, and paid for may be considered the next giant leap for mankind.

20.2.2 Open Data Licenses link

This reading focuses on the importance of licensing your data, how to license your data, the different kinds of rights (intellectual property or other) and how other countries interpret those rights. The material has a pretty serious disclaimer that should be considered "This information is collected by altruistic individuals most of whom are not lawyers; those who are lawyers are not your lawyers nor experts in your situation. You use this information at your own risk. Nothing in this page should be considered as legal advice". If that isn't reassuring I don't know what is. The material wants to educate the reader to the importance of licensing data not just to prevent others from using your data but if you want your data to be available to others, where they may be legal limitations to others use.

The reading does emphasis that if you are planning to make your data available you should put a license on it. The material is fairly technical but helps the reader think through the process and importance of licensing their data. The reading also seeks to help the reader understand the language of data collection and the importance of ensuring that the language is understood example: "the structural elements of a database will generally be covered by copyright. However, we need to be a bit careful because the word isn't particularly precise: "data" can mean a few or even single items (for example a single bibliographic record, a lat/long etc) or "data" can mean a large collection (e.g. all the material in the database). To avoid confusion we shall reserve the term "contents" to mean the individual items, and data to denote the collection". As my old boss use to say words have meaning.

Forms of protection for your data will fall most likely into two cases, copyright for compilations and a sui generis (constituting a class alone, unique, peculiar) right for collections of data. It's critical that you understand how countries will interpret the protection that will afforded based on copyrights and licenses. The material does a good job of providing the reader an over view of how the European Union, Canada and the United States will interpret data protection. The overall reading is fairly technical but required if the owner of data is wanting to understand how their data will be protected and what they have to do it the want others to have access to use the data.

20.2.3 Project Open Data

This reading is designed to educate the reader in regards to the US Government project called "Project Open Data". The reading is a located on the GitHub website that allows readers to participate in the process and refinement of the website information. The US government recognized that "data is a valuable national resource and a strategic asset to the U.S. Government, its partners, and the public." The website seeks to assist data owners with "wherever possible, release it to the public in a way that makes it open, discoverable, and usable." The website seeks to assist with the implementation of the US Government Open Data program by providing tools to assist in the process.

The reading provides definitions of the principles of open data(public, accessible, described, reusable, complete, timely and managed post release), the standards, specifications and formats developmental process, open data glossary of common terms and a project open data metadata Schema (guidance to support the use of the Project Open Data metadata to list agency datasets and application programming interfaces (APIs). The definitions provide a common language that all organizations can reference to make sure that miscommunication is minimized and all participants can operate on a shared baseline of information that will be recognized by others. There is also references on how the open data will be implemented by the US Government and tools that are available. The references are the Executive Orders and President Memorandums that provide the framework and authority for implementation of the Open Data program. The references even provide how APIs will be documented. APIs are a program that allows one computer software to talk to another computer software. The tools provide technical data to assist with the data transmission.

There is information on resources that are available to organizations that are required to implement the Open data Program. Resources like Metadata Resources, Business Case for Open Data, and Examples of Policy Documents just to name a few, there is also a potion that covers case studies of organizations that are leading the way ion open data so that others can use as information to see how their own organization is doing and may be find a best case to assist them with implementation.

The final portion is dedicated to engagement with others and your own organization to seek feedback and act on that feedback. Provides format on how to hold engagements and who to achieve the best results by setting objectives and how to seeking engagement with external partners. This reference helps educate not just the government employee, but those organizations that seek access to government data for their own studies. By understanding the government's requirement to implement open data, users may have a better understanding of what data and when will be available for external use.

20.2.4 Discussion Questions

1. Do you see contradictions in your field between the goods and bads of open data?

2. Do you know if you your data will be protected? Why should I license my data and if I do need to license my data how do I do it.

3. What is the US Government's policy in regards to the data that they create each and every day, is there data that I will be able to use in the future.

YouTube Summary Video LINK

20.3 References

- Sunlight Foundation: Open Data Policy Guidelines LINK

- Open Data Licenses: LINK

- Project Open Data: LINK

Chapter 21

Ethics of Algorithms

(Team 7) Julie Moore and Lorna Romero

21.1 Topic Overview

This section of Weapons of Math Destruction we explore how big data can have a negative effect for an individual getting insurance, specifically auto, and a comments on health care related to organizations. The major themes of the book are discussed during the conclusion as well as an aspirational goal for the proper use of data and the non-proliferation of WMDs in this context.

21.2 Chapter Summaries

21.2.1 WMD Chapter 9 NO SAFE ZONE: Getting Insurance

Segregating people into different risk classes began back in the late nineteenth century with Frederick Hoffman at Prudential Life Insurance. He used data to determine "race was a powerful predictor of life expectancy." It was discovered he enveloped all African Americans as uninsurable, he lumped laboring sharecroppers and professionals into one group. Hoffman "published a 330-page report that set back the cause of racial equality in the U.S. and reinforced the status of millions as second-class citizens" "not meaning to harm".

John Graunt and other mathematicians began calculating big data to determine the most probable arc of a person's life span. Insurers began drawing their scores from credit reports. Meaning "how you manage money can matter more than how you drive a car." In other words, as an example in Florida, if you have a clean driving record and poor credit scores one would pay an average of $1,552 per year more than a convicted drunk driver with a good credit score.

Allstate charges individuals based on if they think a consumer will "shop around" for better pricing, if not Allstate will charge an individual more money for a policy. Stated by the former Texas insurance commissioner and CFA director of insurance "Allstate's insurance pricing has become untethered from the rules of risk-based premiums ad from the rule of law." The industry charges citizens based on a host of proxies versus true driving skills. This results in a feedback loop that poorer drivers pay higher premiums, so if a driver continues to experience bad debt (because they are paying higher premiums) they may default on other expenses such as auto loans, credit cards, etc., which ultimately spirals the person "into an even more forlorn microsegment." It's a never-ending vicious circle.

Consumer Reports got involved and launched a campaign directed at the National Association of Insurance Commissioners (NAIC) titled "Price me how I drive not by who you think I am!" In other words, premiums should be based on driving records, not other proxies. The trucking industry is used as an example. Seven hundred truckers die per year in America. The average cost of a fatal crash is $3.5 million according to the

Federal Motor Carrier Safety Administration. So, leading insurers offer discounted rates to truckers who agree to share their driving data, such as a telemetric unit, which logs speed, braking, and accelerating, and a GPS monitor to track a vehicle's movement. Ideally, this coincides with the Consumer Reports campaign. It's pointed out that if a driver drives in a risky neighborhood, where car thefts and drunken driving are prevalent, insurers can take this into consideration, which sounds like an improvement. Interestingly, an example of a barista was given where she drives home late night through a bar/stripper street to save time and toll charges and now falls into a higher premium rate. In the early years, we had to "opt in" to be tracked, now everyone can be tracked all the time. The use of big data, without human intervention, is categorizing everyone. Insurance company's original purpose was to help society balance its risk, however with big data, citizens who can least afford high insurance rates are getting hit the hardest.

The next example in Chapter 9 regards big data used concerning health care. Employer's new trend is to establish "wellness programs" (being incentivized by the Affordable Car Act), which springs from "good intentions" to offset coverage costs. However, will employers "sift through job candidates?" Companies such as Michelin tire company and CVS have the employees not reaching targeted requirements pay more in health premiums up to $1,000. "Humiliation and fat-shaming" were used by a columnist from the "Bitch Media" regarding CVS. Using Body Mass Index (BMI) was determined to be flawed because of so many different body types and is said to "mold bodies to the corporation's ideal and infringes on freedom." The book states there is "scant evidence that mandatory wellness programs actually make workers healthier." It's stated that even when people lose weight participating in a wellness program, they tend to gain the weight back in the near future. Although breaking a smoking cigarette habit has shown to be successful.

The last point made in Chapter 9 states "while it is true that people are more likely to suffer from health problems, these tend to come later in life, when they're off the corporate health plan and on Medicare. In fact, the greatest savings from wellness programs come from the penalties assessed on the workers. In other words, like scheduling algorithms, they provide corporations with yet another tool to raid their employee's paychecks." It is believed employers are "trying to map our thoughts and friendships, and predict our productivity," which could lead to a "full-fledged WMD."

21.2.2 WMD Conclusion

In the conclusion of Weapons of Math Destruction, Cathy O'Neil provides a recap of the major themes of the book and aspirational goal for the proper use of data and the non-proliferation of WMDs in this context.

According to O'Neil, "being poor in the world of WMDs is getting more and more dangerous and expensive." For example, poor people with bad credit who live in bad neighborhoods are likely to be served predatory ads for pay-day or loans for-profit universities. Companies and organizations use data for targeting purposes, creating marketing or advertising silos based on a number of factors such as social economic class. O'Neil contends that this is not something the free market can correct on its own, it requires action and regulation.

One example given was while former Democratic President Bill Clinton signed to Defense of Marriage Act, IBM made the decision to give medical benefits to same-sex partners. O'Neil speculates that IBM made that decision to remain competitive in the growing internet and etch marketplace; fairness was a byproduct of the business decision.

Eliminating some WMDs does not always create a financial benefit. Some companies are solely modeled off of this type of manipulation.

WMDs also target the middle class, by preventing job opportunities for qualified candidates or reducing pay for someone who doesn't meet model health requirements.

But although human decision making is flawed, it can evolve, but automated systems cannot. They process the information of the past not invent the future.

O'Neil suggest that in order to make changes moving forward, data scientists must pledge to a Hippocratic Oath: * I did not make the world and it does not satisfy my equations * I will use models boldly to estimate values, but I will not be overly impressed by mathematics * I will not sacrifice reality for elegance without

explaining why I have done so * Nor will I give the people who use my model false comfort about its accuracy, instead I will make explicit its assumptions and oversights.

O'Neil also notes that a regulatory framework for WMDs must include hidden or non-numerical costs. We must also acknowledge that technology alone cannot do everything. To disarm WMDs, we must measure impact and conduct audits. This will allow us to track biases in automated systems. For example, Netflix uses algorithms to optimize the content feature for their subscribers. But audits often face resistance from "web giants" such as Facebook and Google who wish to keep their information internal. O'Neil argues that these web giants "must be accountable to all of us—which means opening its platform to more data auditors.

We need to demand transparency, disclosing the input data as well as the results of their targeting, data collected must be approved by the user, an opt-in approach, and models should be available to the public.

The main problem for data analysts is not just to identify the problem but present possible solutions. Some mathematical models can used for good, if they are not abused. Examples include tracking goods to see if they were produced by forced labor or predictive models that can determine if a child is likely to be in an abusive home to provide proper services.

Math and data can solve problems, but we need to demand fairness and accountability in order to do good and not cause harm.

21.3 Key Take-Aways (for Yellowdig)

21.3.1 Summary #1:

Chapter 9 of WMD reports on how big data can have a negative effect for an individual getting insurance, specifically auto, and a comments on health care related to organizations.

Auto insurance companies use credit reports to determine insurance rates, in other words "how you manage your money can matter more than how you drive a car." Allstate charges individuals based on if they think a consumer will "shop around" for better pricing, if not, Allstate will charge an individual more money for a policy. This results in a feedback loop that poorer drivers paying higher premiums.

The industry as a whole charges citizens based on a host of proxies versus true driving skills. Consumer Reports got involved and launched a campaign directed at the National Association of Insurance Commissioners (NAIC) titled "Price me how I drive not by who you think I am!" In response, leading insurers began offering discounts to drivers who share their data… although this coincides with the Consumer Reports campaign, drivers were still "dinged" even if they had valid reasons for driving through "bad" areas like bars and strip clubs.

Regarding Health care insurance, the chapter mentions the new employer trend of establishing wellness programs, which brings up concerns of employers "sifting through job candidates. There's also an interesting discussion of Body Mass Index (BMI) and how it can be misinterpreted.

The last point made in Chapter 9 states "while it is true that people are more likely to suffer from health problems, these tend to come later in life, when they're off the corporate health plan and on Medicare. In fact, the greatest savings from wellness programs come from the penalties assessed on the workers. In other words, like scheduling algorithms, they provide corporations with yet another tool to raid their employee's paychecks." It is believed employers are "trying to map our thoughts and friendships, and predict our productivity," which could lead to a "full-fledged WMD."

21.3.2 Summary #2:

In the conclusion of Weapons of Math Destruction, Cathy O'Neil provides a recap of the major themes of the book and aspirational goal for the proper use of data and the non-proliferation of WMDs in this context.

According to O'Neil, "being poor in the world of WMDs is getting more and more dangerous and expensive."

But although human decision making is flawed, it can evolve, but automated systems cannot. They process the information of the past not invent the future.

O'Neil suggest that in order to make changes moving forward, data scientists must pledge to a Hippocratic Oath.

Math and data can solve problems, but we need to demand fairness and accountability in order to do good and not cause harm.

21.3.3 Discussion Questions

1. How do you feel about sharing your driving data with auto insurers? The Chapter mentions a barista driving a road through strip clubs, on her way home from work late at night, to avoid tolls and cutting time off her trip…
2. Do you feel like you are forced to participate in a work wellness program? Chapter 9 suggests these do nothing but "raid your paycheck."
3. Does this area need increased regulation, or can the free market correct this abuse?
4. How much should personal responsibility be a factor when considering future regulation of big data?

21.4 References

- O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books. **CH9 getting insurance**

- O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books. **conclusion**

Bonus Paper (presents a counter-argument to O'Neil, ok to skim to get basic outline of argument):

- Kleinberg, Jon and Ludwig, Jens and Mullainathan, Sendhil and Sunstein, Cass R., Discrimination in the Age of Algorithms (February 5, 2019). Available at SSRN: [LINK]