Data Science in Modern Medicine:
How Data Science is Revolutionizing our Understanding of Behavior, Health, & Disease

Sean R. Zion
Readings in Data Science
Annotated Bibliography
June 6th 2018

Executive Summary

I study how peoples' thoughts, feelings, and behaviors impact their health and well-being. My research aims to integrate empirically supported psychological science into the practice of medicine to better understand illness, improve treatment outcomes, and optimize patient care. I entered this interdisciplinary field at an interesting time. In the psychological sciences, issues of reproducibility and replicability have recently surfaced, calling into question some foundational assumptions upon which the field was built. Similarly, initiatives like the human genome project and the recent focus on precision medicine have taken aim at improving our ability to diagnose and treat the rapidly increasing rates of chronic illnesses. The common thread running through both of these paradigm shifts is the need for skilled researchers with the ability to reliably and accurately collect, manage, and analyze large sets of complex data. As these two fields modernize and recognize the wealth of both behavioral and health data that is routinely being collected, the need for data science will become increasingly apparent.

To explore the current state of data science in the fields of behavioral health and medicine, I've organized this annotated bibliography into three sections. First, I outline the importance of integrating psychological science into the field of data science. My goal here is to highlight the need for social scientists in the growing field of data science, especially in industries for which the end user of the product is a human. Humans are complex, and merging data science with psychological theory can improve our ability to understand and predict human behavior. Second, I'll present a series of empirical papers, review articles, and opinion pieces on the emergence of data science in healthcare. The applications of data science in the medical context and the enormous potential for data scientists to work alongside medical researchers, physicians, and nurses is explored. Finally, I close the gap between these two seemingly disparate fields by honing in on the emerging interest in digital phenotyping. I'll present a series of recent, high impact articles that highlight how 'big data' has incredible potential to revolutionize how we diagnose and treat mental illness.

Synthesis

Given that my annotated bibliography is divided into three related sections, I'll briefly touch upon the interesting and insightful findings from each of the three sections. First, I'll describe the relationship between data science and psychology, and how the two industries can benefit from collaboration and integration. Second, I'll discuss the importance and incredible opportunities of data science in the field of medicine. Finally, I'll review the advancements in

digital phenotyping and how moment-to-moment tracking of human behavior can help us understand mental illness and save lives.

Humans behavior is often irrational and illogical, but it is nevertheless predictable. This notion is the foundation upon which the field of behavioral economics is built. However, as psychologists, we have long focused on the causal mechanisms underlying patterns of behavior. This does not lend itself well to making behavioral predictions. The field of data science has a lot to offer psychology in this regard, and integrating techniques and practices from data science offers promise for improving our ability to predict human behavior. This is especially true within academia, where psychology is undergoing a major methodological shift because of lingering issues of reproducibility. Psychologists can learn from the machine learning approaches that are commonly used by data scientists. While these newer approaches to working with data seem daunting to many, open source research tools have made the integration of big data approaches into research relatively accessible.

This relationship between behavioral science and data science, however, is a two way street: data scientists can also learn a lot from collaborating with psychologists. For any organization whose end user is a human being or has the goal of changing human behavior, data scientists and psychologists can benefit from a collaborative effort to more effectively understand and act upon data. In industries that rely heavily on data science and increasingly take data driven approaches to decision making and business strategies, psychologists can offer insight into how these decisions may impact consumer behavior. Finally, creating efficient teams of data scientists requires understanding skill sets, personality styles, and fundamentals of organizational behavior. As such, psychologists can help create better teams of data scientists and improve the quality and quantity of the insight they are able to generate.

Data science can certainly benefit from insight from the behavioral sciences. However, the field of medicine presents even greater opportunities: integrating data science into healthcare can help make medicine more precise, efficient, and accurate, and can generate new opportunities for data scientists to have a meaningful impact on patient care. The potential opportunities cannot be understated. Integrating data science could transform medicine in the same way as the microscope or antibiotic medications. Medicine cannot be a precise science if it is not driven by data, and data, by themselves, are useless. Data need to be analyzed, interpreted, and turned into actionable insights about how to diagnose, treat, and predict patient trajectories. Making medicine a precise science, therefore, requires data scientists.

30% of all stored data that exists comes from the healthcare industry and this data is valued at upwards of $300 billion. This data, which often consists of electronic health records, biomedical data, genetic information, and behavioral measures captured

by wearables and smartphones, is often given the catch-all descriptive of 'big data'. This big data can improve medicine in two key ways. First, it has the potential to increase clinician understanding of how and when to prescribe different treatments. Second, and critically, it can help predict disease outcomes and therefore could be an important prognostic tool. Machine learning is also emphasized, and recent studies have found that machine learning approaches to diagnosis and prognosis preform on par with trained physicians. Machine learning techniques have also been used to successfully predict suicidal behavior three years in advance using a large sample of patient health records.

These issues are especially relevant in psychiatry, a field for which there are essentially no objective or precise diagnostic methods. Depression, for instance, cannot be diagnosed with a blood test or with even the most advanced medical imaging. Psychiatry is the opposite of precision medicine, and as a result, millions of people have inadequate care and billions of dollars go to waste. Digital phenotyping has emerged in the past few years as a way of addressing the lack of precision. Digital phenotyping can be thought of as the moment-by-moment measurement of individual-level human behavior. Data is collected though active and passive data collection from personal digital devices. Because the average person interacts with their smartphone over 2500 times each day in a variety of unique ways, this technique has the potential to gather and analyze unprecedented amounts of behavioral data. This data can be used to understand connections between genetic factors, behavioral trends, and self-reported symptoms. Working with this health data will necessitate a new type of data scientist with behavioral, medical, and research expertise.

Digital phenotyping has been used by Mindstrong Health to detect depression by quantifying changes in patterns of interactions with smartphone and by Sharecare to monitor stress levels by analyzing data from phone conversations. It has also recently been employed by Facebook, which has started using artificial intelligence to detect language patterns suggestive of suicidal ideation. Working with experts in the field, the team at Facebook used a machine learning engine to train a classifier to recognize and flag posts that indicated the user was at high risk of attempting suicide.

Digital phenotyping may be a novel method of collecting useful data that can be used to predict mental health conditions. However these data are meaningless on their own – it is the analytic strategies that allow for the generation of actionable insight. Integrating data science into medicine and behavioral health will allow these fields to modernize, become more precise, and has the potential to improve and save lives.

Top 3 Papers

While there were many excellent empirical studies, review papers, blog posts, and thought pieces, several stood out particularly thought provoking and impactful. I selected the single most interesting paper from each of the three subsections of the annotated bibliography. The first describes the last mile problem and argues for an integration of behavioral sciences and data sciences. The second takes a look at a novel tool designed by Facebook to detect suicidal behaviors using artificial intelligence to detect linguistic patterns in social media posts. The third (which I strongly recommend) is an article from The New England Journal of Medicine Catalyst that describes the massive amount of healthcare data that has been generated and the comparative absence of data science from this field.

1. The last-mile problem: How data science and behavioral science can work together; Jim Guszcza (Jan 26 2015)
    - To act on data that captures human behaviors, data scientists need to understand how people think about the world and make decisions, not just the data analytics.
    - Behavioral scientists with strong data science skills can be an invaluable asset to any company for which the end user is a human being.
    - When the goal of an organization is behavior change, data science and psychologists can work together to more effectively understand and act on data.

2. Under the hood: Suicide prevention tools powered by AI (Facebook); Dan Muriello, Lizzy Donahue, Danny Ben-Davic, Umut Ozertem, & Reshef Shilon (Feb 21, 2018)
    - An excellent real world example of at the crossroads of data science, psychology, and medicine.
    - Facebook has started using artificial intelligence to detect language patterns that may suggest increased suicide risk.
    - Working with experts, the team at Facebook identified key words and phrases that may be associated with suicidality.
    - Using a machine learning engine, a classifier was trained to recognize and flag posts suggesting suicidality. The most effective classifier consisted of a combination of DeepText and linear regression models.

3. Using It or Losing It? The Case for Data Scientists Inside Health Care; March Huesch & Timmothy Mosher (May 4, 2017)
    - 30% of all data that exists worldwide comes from the healthcare industry. This enormous backlog of patient health data is estimated to be worth upwards of $300 billion.
    - However, according to LinkedIn, only 180 of the approximately 6,000 data scientists in the US work in the healthcare sector.

- Most of the key players in the healthcare industry are reluctant to adopt the tools used in the finance and tech sectors to harass the potential of this mass of data.
- The field needs to invest in data analytics and dramatically increase the number of data scientists to produce actionable insights from this data.

Annotated Bibliography

1. The Role of Psychology in The Field of Data Science

Behavioral Analytics : When Psychology collides with analytics; Tavish Srivastava (Apr 19, 2017)

- Consumer behavior is complex and is driven by underlying psychological tendencies.
- Rules of behavioral economic govern the often illogical and irrational decisions people make in their daily lives.
- Understanding, analyzing, and visualizing consumer data requires some degree of understanding of these psychological and behavioral principles.

Thinking machines: What has psychology got to do with data science?; Brad Love (Apr 5, 2017)

- Human brains and computers process information in different but complementary ways, and understanding how they complement one another can foster new ideas.

From Social Scientist to Data Scientist; Emily Robinson (Feb 7, 2017)

- Social scientists have a rigorous training in designing and carrying out experiments, which is a critically useful skill in the field of data science.
- Certain skills that are needed to be a data scientist, however, might not be included in the curricula of a social scientist, so data science boot camps have become a popular option to hone these skills before entering the workforce.

The last-mile problem: How data science and behavioral science can work together; Jim Guszcza (Jan 26 2015)

- To act on data that captures human behaviors, data scientists need to understand how people think about the world and make decisions, not just the data analytics.
- When the goal of an organization is behavior change, data science and psychologists can work together to more effective understand and act on data.

Why behavior science is the new data science in Consulting; Patrick Gormley (Apr 5, 2017)

- As companies focus on data driven approaches to decision making and business strategy, they will need to understand how these advances affect consumer behavior.
- Behavioral experts can serve as consultants to help quantify the effects these technological advances have on consumer and employee behavior.

The Psychology of Data Science: The who's who game in a data science team; Francesco Corea (Jan 26, 2017).

- Creating efficient teams of data scientists requires understanding skill sets, personality styles, and organizational behavior.
- Psychologists can help organizations create better teams of data scientists.

An algorithm can predict human behavior better than humans; Olivia Goldhill (Oct 18, 2015)

- Algorithms can successfully predict human behavior, and can do so more accurately than humans in certain instances.
- Using large scale behavioral data and prediction algorithms can increase our ability to quantify and predict human behaviors.

Big Data in Psychology: A Framework for Research Advancement; Idris Adjerid & Ken Kelley (Feb 22, 2018)

- The integration of 'big data' and techniques traditionally used in data science into the field of psychology presents valuable opportunities for the field.
- This opportunity often seems daunting for the average academic psychologist.
- Open source research tools make the integration of big data into research more accessible than most psychologists realize and diverse opportunities are available for integrating data science techniques into psychological research.

Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning; Tal Yarkoni & Jacob Westfall (Aug 25, 2017)

- Psychology has long been concerned with understanding the causal mechanisms that underlie human behavior.
- This overemphasis on causal mechanisms does not translate well to behavioral prediction.
- Integrating machine learning approaches to behavioral sciences has the potential to make psychology a more predictive science.
- Focusing on prediction over explanation will provide more useful information about the nature of human behavior.

2. The Emerging Role of Data Science in Modern Medicine

The Inevitable Application of Big Data to Health Care; Travis Murdoch, Allan Detsky (Apr 3, 2013)

- The amount of health data that is routinely collected has increased dramatically in recent years.
- This data presents opportunities for advancing clinical care.

Detecting Chemotherapeutic Skin Adverse Reactions in Social Health Networks Using Deep Learning; Julia Ransohoff, et al (Apr 1 2018)

- Traditional clinical trials are underpowered to detect serious but rare adverse effects.
- Once approved, most treatments do not undergo continuous monitoring to understand long term or rare side effects.
- Deep learning models using data from social networks can help capture important side effects than may have been overlooked in clinical trials.

Big Data and Machine Learning in Health Care; Andrew Beam & Isaac Kohane (Apr 3, 2018)

- Machine learning has allowed for significant advancements in the consumer tech industry, but has yet to be widely applied to health and medicine.
- Recent studies suggest that machine learning algorithms can perform on par with physicians.
- Integrating data science and basic understanding of machine learning into the medical education curricula may increase the use of these techniques, which initially appear out of the reach of many clinicians.

Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations; Jonathan Chen & Steven Asch (June 29, 2017)

- Machine learning techniques have the potential to improve the accuracy and efficiency of modern medicine.
- However, machine learning has taken on the status of a buzzword, with substantial hype surrounding the promise of revolutionizing healthcare.
- Future disillusionment with machine learning can be abated by educating the medical profession about its capabilities and limitations.

Implementing Machine Learning in Health Care — Addressing Ethical Challenges; Danton Char, Nigam Shah, & David Magnus (March 15, 2018)

- Machine learning aims to improve medicine and has the potential to revolutionize decision making and diagnosis in certain subspecialties like radiology and pathology.
- Ethical concerns surrounding the use of such techniques have surfaced and may cause strain between the data scientists who design these tools and the end users (physicians) who implement them.
- Medical ethics will likely need to adapt to this changing landscape.

Machine learning and electronic health records: A paradigm shift; Daniel Adkins (Feb 1, 2017)

- The amount of health and behavioral data in electronic health records (EHR) has exceeded the amount of information a physician can integrate into clinical decision making.
- Recent studies have employed machine learning models using longitudinal data from EHR.
- One such study of 1.7 million patient EHR accurately predicted suicidal behavior 3 years in advance of such behavior.

The Parable of Google Flu: Traps in Big Data Analysis; David Lazer, Ryan Kennedy, Gary King, & Alessandro Vespignani (March 14, 2014)

- Google Flu Trends, which has been hailed as one of the more prominent success stories of using big data in the domain of health, was found to be consistently over reporting cases of flu.
- This inaccuracy is blamed on two factors: Algorithm dynamics and big data hubris.

Learning from Big Health Care Data; Sebastian Schneeweiss (June 5, 2014)

- Big data can improve medicine in two key ways. First, it can help us understand the efficacy of treatments more accurately. Second, it can help predict health outcomes.
- To leverage big data in healthcare more effectively, the tools must be designed to appeal to busy clinicians in a way that can be integrated into their daily workflow.

Using It or Losing It? The Case for Data Scientists Inside Health Care; March Huesch & Timmothy Mosher (May 4, 2017)

- 30% of all data that exists worldwide comes from the healthcare industry, worth upwards of $300 billion.
- Most of the key players in the healthcare industry are reluctant to adopt the tools used in the finance and tech sectors to harass the potential of this mass of data.
- Medicine needs to invest in data analytics and dramatically increase the number of data scientists to produce actionable insights from this data.

Why Every Health Care Organization Needs a Data Science Strategy; Kathrin Cresswell, David Bates, Aziz Sheikh (March 22, 2017)

- Organization wide data science strategies are needed to harness insights from healthcare data.
- Effective utilization of data scientists within healthcare organizations have five key components: a secure repository, data integration across sources, governance frameworks to ensure security, use and reuse of data to improve care, and organizational capacity development.

Healthcare Big Data and the Promise of Value-Based Care; NEJM Catalyst (Jan 1, 2018)

- The use of big data has ushered in one of the most influential paradigm shifts in modern medicine.
- The challenges of this data comes from its large volume, the speed at which the data is generated, and the variability of the sources.
- Appropriate tools are needed to leverage this data and turn it into useful, actionable information.

Care Redesign Survey: What Data Can Really Do for Health Care; Amy Compton-Philips (March 9, 2017)

- Providers have expected the use of big data to solve many of healthcare's fundamental problems, and have become disillusioned by how difficult it has been to leverage this data in useful ways.
- Most healthcare organizations do not believe they are using their data effectively.
- A major barrier to effective data use is the lack of interoperability of the electronic health record systems. That is, it's difficult for data to follow patients across providers, networks, and care settings.

Predicting the Future — Big Data, Machine Learning, and Clinical Medicine; Zaid Obermeyer & Ezekiel Emanuel (Oct 10, 2016)

- Data itself is not transformative, but approaches to analyzing and yielding actionable insight from datasets can be transformative.
- In most other fields, machine learning has become a standard tool for solving complex problems. Within the healthcare industry, it can have a similar impact.
- Machine learning has the potential to improve diagnosis accuracy, reduce the workload of certain types of clinicians, and establish and improve prognosis.

Top 7 Data Science Use Cases in Healthcare; Igor Bobriakov (March 27, 2017)

- Data science has the potential to improve a number of domains within clinical medicine, but several show the most promise.
- Analysis of medical imaging, utilizing genetic data, creating novel treatments, predicting outcomes, and improving diagnosis are amongst the use cases that may have the highest impact on the field.

Data Science and Predictive Analytics in Healthcare; Thomas Blanchard (Sept 4, 2017)

- Integrating data science into the healthcare industry can save people's lives.
- The industry lags behind other industries in adopting methods from the field of data science and has yet to regularly employ data scientists.
- Healthcare presents unique challenges related to protected health information, data security, and the implications of poorly implemented data science.

3. **Digital Phenotyping: At the Crossroads of Psychology, Medicine, and Data Science**

Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health Jukka-Pekka Onnela & Scott Rauch (Jan 28, 2016)

- Digital phenotyping can be thought of as moment-by-moment measurement of individual-level human behavior by capturing data from personal digital devices.
- Digital phenotyping has the potential to gather and analyze behavioral data such that we can make connections between phenotypic expression, disease type, and genetic variations.
- It involves both active and passive data collection, largely from smartphones.

Digital Phenotyping: Technology for a New Science of Behavior; Thomas Insel (Oct 3, 2017)

- Digital phenotyping is ideal for capturing social and behavioral dimensions of psychiatric illness.
- Quantity of data and analytic strategies appear daunting.
- Need for skilled experts, like data scientists with an interest in human behavior, mental health, or psychiatry, to take on this challenge.

Digital phenotyping: an overarching framework to capture our extended mental states; Andrea Raballo (March 1, 2018)

- Digital phenotyping is needed because psychiatric classifications emerged through a complex historical process, not a systematic construction. Digital phenotyping will all for a bottom up revision of diagnostic criteria.
- A long-term roadmap and conceptual toolbox for digital phenotyping is needed.
- The substantial data that can be collected will require a new type of clinician and researcher with strong skills in data science

23andMe, Big Data, and the Genetics of Depression; Jennifer Abbasi (Jan 3, 2017)

- Big data approaches may be useful for identifying genetic associations for psychiatric conditions.
- The immense amount of genetic information collected by private firms like 23andMe presents new opportunities for integrating data science into medicine.
- Many companies collect and sell data, but the use of genetic data in this way has brought up ethical concerns that data scientists will need to grapple with in the future.

Bigger Data, Harder Questions—Opportunities Throughout Mental Health Care; Adam Chekroud (Dec 1, 2017)

- Big data isn't new in medicine, but we're at a point where we have machine learning tools that allow us to make sense of large complex data sets.

- A big data pipeline in which clinicians translate the patients' needs to data scientists who are able to design, build, test, and implement tools that will facilitate greater insight into patient care.

How Companies Scour Our Digital Lives for Clues to Our Health; Natasha Singer (Feb 25, 2018)

- The average person interacts with their smart phone over 2500 times each day. These interactions can hold clues to a person's health and well-being.
- Digitial phenotyping has been used by facebook to detect suicide risk, by MindStrong Health to detect depression, and by Sharecare to detect stress levels.
- This behavioral health data will necessitate a type of data scientist – one with clinical or research training – to provide relevant context to the health and behavior data.

Under the hood: Suicide prevention tools powered by AI (Facebook); Dan Muriello, Lizzy Donahue, Danny Ben-Davic, Umut Ozertem, & Reshef Shilon (Feb 21, 2018)

- Facebook has started using artificial intelligence to detect language patterns that may suggest increased suicide risk.
- Working with experts, the team at Facebook identified key works and phrases that may be associated with suicidality.
- Using a machine learning engine, a classifier was trained to recognize and flag posts suggesting suicidality. The most effective classifier consisted of a combination of DeepText and linear regression models.