2016 Stu Hunter Research Conference
Waterloo, Canada

Analyzing Behavioral Big Data: Methodological, Practical, Ethical, and Moral Issues

Galit Shmueli[1]*

Abstract

The term "Big Data" evokes emotions ranging from excitement to exasperation in the statistics community. Looking beyond these emotions reveals several important changes that affect us as statisticians and as humans. I focus on *Behavioral Big Data* (BBD), or very large and rich multidimensional datasets on human behaviors, actions and interactions, which have become available to companies, governments, and researchers. The paper describes the BBD landscape and examines opportunities and critical issues that arise when applying statistical and data mining approaches to Behavioral Big Data, including the move from macro- to micro-decisioning and its implications.

Keywords

big data, human behavior, data mining, statistical methods

[1] *Institute of Service Science, National Tsing Hua University, Hsinchu, Taiwan*

***Corresponding author**: galit.shmueli@iss.nthu.edu.tw

What Is Behavioral Big Data

Big Data has become available in many fields due to technological advancement in measurement and storage. The term Big Data can now be found in almost any field, including manufacturing, engineering, the physical and life sciences, business, and more recently in the social sciences. As [Hoerl et al., 2014] comment, "Big Data is here to stay, and is revolutionizing business, industry, government, science, education, medicine, and virtually all aspects of society."

The focus of this paper is the types of data, studies, and applications that capture human actions and interactions at a new level of detail. Such studies and applications are quickly growing in industry as well as in academic research. I call this new data type *Behavioral Big Data* (BBD), to highlight its focus on human behavior, and the novelty of its scale. The combination of "behavioral" and "big" create challenges and opportunities for statisticians and data miners, especially since the great majority of researchers in these communities are not trained in the behavioral sciences.

BBD Is a Special Kind of Big Data

BBD is different from other types of Big Data in a fundamental way: it captures people's actions and interactions, their self-reported opinions, thoughts, and feelings that pertain to their day-to-day life. This is different from Big Data collected on items and products (such as in manufacturing or engineering), as the human aspect involves issues of intention, deception, emotion, reciprocation, herding, and other human and social aspects. In many cases, people are aware that BBD is collected about them. They therefore might modify their behaviors accordingly, such as to avoid legal risks, embarrassment, or unwanted solicitation.

BBD offers a valuable addition (and sometimes alternative) to surveys, case studies, interviews and other traditional social science data collection methods that aim to capture human intentions, feelings, and thoughts. BBD therefore holds the promise to transform our understanding of individuals, communities, and societies.

BBD Is Different from Medical Big Data

What is common to medical big data and behavioral big data is that they involve human subjects. However, there are two fundamental differences between the two:

1. Medical big data is focused on physical measurements. Advances in technology allow cheap and accurate measurements of various physical phenomena using wearable devices such as watches that measure activity and heartbeat and bathrobes that measure a person's temperature (Hunter, 2013,

https://youtu.be/YwNqAd1N0Ho). The human genome project has big data at the DNA-level of humans. In contrast, BBD includes data on people's daily actions and interactions, as well as self-reported feelings, opinions, and thoughts.

2. Medical data that arises from large clinical trials is fundamentally different from large-scale behavioral experiments: the subjects of the experiments in clinical trials are aware that they are part of an experiment and have a vested interest in being part of the experiment. In contrast, subjects of behavioral experiments in the age of BBD are often unaware that they are part of an experiment. And in many cases, the experiment is aimed at helping the company, sometimes at the expense of the individual. In other words, unlike clinical trials, in large-scale behavioral experiments there usually is no direct value for the subjects.

BBD on Citizens and Customers

Governments and companies have long realized that data on human behavior can help them make better decisions. The technological ability to collect, store, and analyze detailed behavioral data on many individuals, and its usage by governments and companies has been widely discussed in the context of privacy and confidentiality ([Fienberg, 2006]). BBD has been long used by government agencies, who collect and analyze BBD for security, law enforcement, and other reasons. Today, many major cities have cameras and sensors that record traffic and human activity. Police and security departments use such data for crime detection; transportation departments use the data for managing traffic lights, issuing speeding tickets, and charging for road usage.

Another organization that has been collecting BBD for a long time are telecommunication companies. Telecoms have enormous databases that contain metadata about each call made: its origin and destination, call duration, time stamp, and more. Banks have information on individuals' complete financial activity, including ATM withdrawals, loan payments, salary deposits, bill payments, and more. Similarly credit card companies have information on every purchase the card holder performs as well as bill payment activity.

Hypermarkets and large chain stores have also long joined the BBD wagon, collecting detailed information about individual purchases and individual customers. They have been one of the earliest users of BBD for marketing purposes.

As tracking technology advances, more BBD types become available. For example, some insurance companies in the United States offer Usage-Based auto insurance (UBI) schemes, which rely on telematics information collected by a device in the car. The device collects information about driving habits, thereby giving the insurance companies access to driving BBD on all their insured drivers.

A major technology leading to the availability of BBD to many more (and smaller) companies, is the Internet. Companies and other organizations that offer online platforms, such as e-commerce, gaming, search, and social networking sites, obtain large amounts of user generated data. This data is typically unsolicited and includes micro-level details on the user's activities. The data consists of information actively and often voluntarily entered by the user, such as personal details, photos, comments, messages, search terms, bids in auctions, payment information, and connections with "friends", as well as passive footprints such as the duration they spent on the website, what pages were browsed, in what sequence, the website that referred them to the website, the Internet browser and operating system used, location, and IP address. Websites that require the user to login, or those that use the technology of cookies, can collect longitudinal data on their users.

As more human activities go online, new types of BBD become available. With the growth of online education, there is now an abundance of BBD on students who participate in online courses. Companies such as Coursera, EdX, FutureLearn and other platforms that offer online courses, have data on every participant who has taken a course: which pages were viewed, what material was downloaded, how long a video was played, their quiz results, posts on a discussion board, and more. Personal relationships have also gone online: dating websites (e.g., match.com) and marriage matrimonial websites (e.g., www.shadi.com) collect rich information about their users and their activities on the website. As drivers now use mobile phones in cars, apps such as Waze use the information from drivers' phones to map traffic and provide drivers with the traffic situation and suggested routes.

We note that most of the companies that have BBD were not created for the purpose of generating BBD. Instead, their business model was based on providing a service. They later discovered that the BBD in their possession is one of their most valuable assets. This is true even for mega-BBD-owners such as Google, which started out as a provider of web search capabilities.

BBD on Employees

Industrial statisticians have contributed greatly to advances in analyzing data from manufacturing environments. In his paper "Industrial statistics: The challenges and the research", [Steinberg, 2016] emphasized the importance of real-world industrial problems for stimulating valuable statistical research, and the decline in such work in recent years. He mentioned Big Data, data science and industry as a critical frontier where statisticians are mostly absent.

Steinberg mentioned the marked shift of Western economies from manufacturing to services in the last 30 years. Such a shift is also engulfing many Asian economies, such as India, which has become a giant provider of back-end services. The move from products to services has led to the availability of BBD on employees: the service and service provider are now the subject of measurement, quality control, and quality improvement. Service provider companies now have large datasets on their employees and the service each employee provided. For example, call centers have information on each call taken by an operator ("this call may be monitored"), used for purposes of quality control and productivity improvement. Many companies, governments and other organizations use Electronic Performance Monitoring (EPM) systems that record and even monitor employee behaviors: web surfing, e-mails sent and received, telephone use, video surveillance for security purposes, storage and review of computer files, video recording of employee job performance, recording and review of telephone conversations, and storage and review of voice-mail messages [Moussa, 2015].

Mobile devices now enable easy recording of "offline" service encounters that are not conducted on the web or by phone. For example, the Taiwan-based company iChef (www.ichef.com.tw) provides restaurants with a mobile Point-of-Sale (POS) solution for easier handling of the restaurant's operations. The restaurant staff use iPads to manage everything from the waitlist of arriving customers, to seating, taking orders, coordinating with the kitchen, and finally handling payment and accounting. iChef therefore has BBD on the restaurant staff and on consumers[1].

In an effort to improve service, some taxi companies (e.g., Meru Cabs in India, www.merucabs.com) monitor their drivers' locations and share the location with the customer, for purposes of safer rides and reduced uncertainty about pickup time. Such information results in BBD about the drivers: the routes they have taken, the speed traveled, etc.

BBD on Self

With the now trendy wearable technology ("fashion electronics") such as activity trackers, heath and fitness devices, and the ubiquity of mobile phones, a new source of BBD is the Quantified Self (QS). QS involves ordinary people recording and analyzing numerous aspects of their lives to understand and improve themselves ([Fawcett, 2016]). QS is more than simply medical data, because on top of data from physical measurements (e.g., heartbeat and sleeping cycles) it also includes cognitive measurements (response time), self-reported food consumption, mood, stress levels, and more. Such measurements make QS part of BBD. Users often upload their data to apps that help them make sense of the data. Thus, QS BBD is also available to providers of such apps.

The examples given in this section aim to give a flavor of the types of information now available to companies, and the ease with which BBD can now be collected by large and small enterprises and organizations and even individuals.

Research Using Behavioral Big Data

The availability of BBD in many new areas of daily life has enabled researchers in the social and behavioral sciences to examine new phenomena as well as to re-examine old phenomena with better data. However, because handling large amounts of data requires some technical capabilities and expertise, research using BBD has progressed fastest by those with a technical background. Two such communities are the information systems and marketing areas, typically housed in a business school, where researchers have mixed backgrounds, from computer science and engineering to economics and behavioral sciences. A third community is computational scientists in computer science departments or in corporate research labs such the "computational social science" group at Microsoft Research, which comprises of members with computer science and social sciences backgrounds. What makes the BBD research of these communities different from other research using Big Data, is that the research questions themselves are about

[1] iChef and NTHU's Institute of Service Science have recently partnered on a large-scale research project called "small restaurants, big data".

human behavior. Because of the close ties of such communities to industry, research is often geared towards advancing company goals.

[Watts, 2013], principal researcher at the computational social science group at Microsoft Research, defined the field of computational social science and its challenges:

> There's been surprisingly little progress on the "big" questions that motivated the field of computational social science... Of the many reasons for this state of affairs, I concentrate here on three. First, social science problems are almost always more difficult than they seem. Second, the data required to address many problems of interest to social scientists remain difficult to assemble. And third, thorough exploration of complex social problems often requires the complementary application of multiple research traditions — statistical modeling and simulation, social and economic theory, lab experiments, surveys, ethnographic fieldwork, historical or archival research, and practical experience — many of which will be unfamiliar to any one researcher.

Examples of BBD Studies

To give the reader an idea about types of research topics conducted by researchers using BBD, we briefly describe a few studies recently published in top journals in management, information systems, economics, and in *Science*.

Example 1: Consumption in Virtual Worlds

[Hinz et al., 2015] studied whether conspicuous consumption represents an investment in social capital, by analyzing the digital footprints of purchases and social interactions in different virtual worlds. Specifically, they used BBD from two virtual world websites. The first is Habbo, which offers a place to meet new and existing friends and play simple games (Habbo receives more than five million unique visitors per month on average, with an average visit duration of 41 minutes). The second website is a newer Massively Multiplayer Online Roleplaying Game (MMORPG), similar to Habbo. The authors conducted an experiment on this newer website in collaboration with the company[2]. This study relied mostly on the social networks information between friends. The authors relate to research in sociology and state that they aimed to answer an age-old sociology question with new BBD data:

> Conspicuous consumption affects anyone who cares about social status; it has intrigued sociologists and economists for more than 100 years. The idea that conspicuous consumption can increase social status, as a form of social capital, has been broadly accepted, yet researchers have not been able to test this effect empirically.

Example 2: Anonymous Browsing in Online Dating Sites

[Bapna et al., 2016] studied the effect of anonymous browsing on user behavior and matching outcomes in online dating websites. The authors partnered with one of the largest dating websites in North America, and ran an experiment to test the effect of anonymous browsing. They analyzed BBD on 100,000 users that included information from the users' profiles as well as their behavior on the website (browsing, messaging, etc.). The authors aimed to answer new questions about human behavior, which arise due to new technologies. They note:

> The growing popularity of online dating websites is altering one of the most fundamental human activities: finding a date or a marriage partner. Online dating platforms offer new capabilities, such as extensive search, big data–based mate recommendations, and varying levels of anonymity, whose parallels do not exist in the physical world... Our work fits under the broader umbrella of emerging research examining the societal impact of the new generation of big data–enabled online social platforms that connect people who either know each other or would like to know each other.

We note that here too, the authors related their work to social science theories, as can be seen by the citations to papers in journals in psychology, sociology, economics, and more.

Example 3: Social influence in Social News Websites

[Muchnik et al., 2014] used BBD from a social news aggregation website where users contribute news articles,

[2] In the acknowledgments, the authors thank Sulake Corporation for the provision of data.

discuss them, and rate comments. They studied the effects of social influence on users' ratings and discourse on the website. They addressed an important research question using data from the new online context:

> The recent availability of population-scale data sets on rating behavior and social communication enable novel investigations of social influence... The data therefore provide a unique opportunity to comprehensively study social influence bias in rating behavior.

Example 4: Impact of Teachers on Student Outcomes using Education and Tax BBD

[Chetty et al., 2014] combined BBD from administrative school district records and federal income tax records to study whether high value-added (VA) teachers improve students' long-term outcomes. The question of the long-term impact of teachers on student outcomes has been of interest in economic policy. The novelty of this study is its use of BBD, which includes many life events such as test scores, teachers, demographics, college attendance and quality, teenage pregnancy, childbirth, earnings, and more[3].

Example 5: Online Consumer Ratings of Physicians

In the context of health providers, [Gao et al., 2014] studied a question related to the new phenomenon of online ratings of service providers: how consumer-generated ratings of physician quality reflect the opinions of the population at large. They compared a large BBD of online reviews against an offline BBD, and stated the novelty of this approach:

> A distinctive feature that differentiates this study from prior work is a unique dataset that includes direct measures of both the offline population's perception of physician quality, and consumer generated online reviews. These data allow us to examine how closely the online ratings reflect patients' opinion about physician quality at large.

Example 6: Emotional Contagion in Social Networks

In collaboration with Facebook, researchers from Cornell conducted an experiment that manipulated the extent to which users were exposed to emotional expressions in their Facebook News Feed ([Kramer et al., 2014]). The authors showed that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness.

These examples are a small (and biased) sample in the now growing BBD-based literature. The aim is to show the diversity in the sources of BBD, the types of questions asked, and the venues in which the work is published[4].

Questions Asked Using BBD

BBD Studies in Academia

The studies presented in the previous section illustrate different types of questions asked by researchers using BBD. Five of the examples examined a causal research question. The study by [Gao et al., 2014] asked a descriptive question ("How do Online Ratings Reflect Population Perceptions of Quality?").

In the social sciences literature, causal questions are the most common. It is therefore not surprising that many of the BBD studies are causal in nature. Among causal BBD-based studies, some examine age-old questions with the newly-available BBD – e.g., whether conspicuous consumption represents an investment in social capital ([Hinz et al., 2015]), the effect of social influence ([Muchnik et al., 2014]), and the impact of teachers on student outcomes ([Chetty et al., 2014]) – while others identify and ask new questions, often related to new technological capabilities and their effect on behavior – e.g., the effect of anonymous browsing in online dating sites ([Bapna et al., 2016]) and the impact of broadband at school on student performance ([Belo et al., 2013])[5].

One challenge faced by non-statistician researchers conducting explanatory and descriptive modeling with statistical models is scalability. Technically, running regression models on very large high-dimensional samples is resource and time consuming, and is often solved by brute-force computation with more powerful computing power. Methodologically, researchers trained in classical statistical inference continue to rely on p-values for drawing conclusions, a practice that is misleading at best ([Hoerl et al., 2014, Lin et al., 2013]). A related methodological issue is multiple testing, that can lead to high chances of false discoveries, and which becomes much more acute in the presence of rich, high-dimensional

[3] Despite the authors claim "We find that teacher VA has substantial impacts on a broad range of outcomes", the reported magnitudes are practically insignificant and in my opinion far from substantial.

[4] The breadth of applications is clearly limited by my own work environment, and therefore might exclude other areas where BBD are used for answering social science type questions.

[5] This study was not described earlier, for brevity.

BBD.

While predictive modeling is rare in the social sciences ([Shmueli, 2010, Shmueli and Koppius, 2011]), there do exist BBD studies that ask predictive questions. These studies are often authored by researchers with a machine learning background. One example is [Hill et al., 2006], who used BBD from a large Telecom to evaluate the predictive value of calls network data to target marketing. As reported in AdWeek[6] (www.adweek.com, June 28, 2009):

> [the researchers] plowed through reams of the data AT&T collects on phone use. Calling patterns revealed to them that there was a direct correlation between the connectedness of consumers and their purchasing habits. More specifically, consumers shop quite a bit like their friends and are more likely to respond to marketing messages from a brand a friend uses.

In another study, [Dhar et al., 2014] used BBD from Amazon.com to create a network between books (based on user co-purchases) with two predictive purposes in mind:

> (a) does current and past information regarding neighboring entities contain predictive information? and (b) do the network's structural properties... contain additional predictive information?

In a third paper, [Junque de Fortuny et al., 2014] studied to what extent larger data on "low-level human behavior" leads to better predictive models. Examining nine BBD in a variety of contexts, from book reviews to banking transaction, they showed that the marginal increase in predictive performance of predictive models built from sparse, fine-grained behavioral data continues to increase even to very large scale. The authors concluded:

> Social scientists have long argued that one way to circumvent the poor predictive validity of attitudes and traits is to aggregate data across occasions, situations, and forms of actions. This provides an early suggestion that more (and more varied) data might indeed be useful when modeling human behavior data. The implication for predictive analytics based on data drawn from human behaviors is that by gathering more data over more behaviors or individuals (aggregated by the modeling), one could indeed hope for better predictions.

BBD Studies in Industry

Companies that own BBD do not always have the capability to analyze their data. BBD must be stored in a way that makes it accessible and amenable for analysis. This has been a major challenge for many companies. Companies that have made the leap into analysis-enabled BBD, have used a variety of methods to try and derive actionable insights. The first step into analytics is typically creating visual dashboards that give a picture of "what is happening" using summaries and charts - called *business intelligence*. The next step is the use of more advanced models and methods, most commonly predictive analytics and data mining algorithms, termed *business analytics*.

Predictive analytics generate predicted values at the individual observation level and have been used by companies for personalization, for various purposes such as increasing the duration a user spends on a website and increasing the number of purchased items. Personal recommendation engines are common on almost every type of website, from e-commerce to news sites, to social network sites, and beyond.

In the previous section we described several studies conducted by academic researchers in collaboration with companies, where the question asked was a scientific one and the study was published in academic journals. We now focus on studies conducted by companies for evaluating or improving their products, service, operations, etc. for the eventual purpose of increasing profit.

A famous recommendation system study by a company is the Netflix Prize contest, which took place in 2006-2008. At the time, Netflix was the largest movie rental company in North America, which provided DVDs by mail to members. The company's goal behind the contest was to improve their movie recommendation system. They decided to "outsource" the task by conducting a contest that was open to the public and had a $1 million prize. For the contest, the company released a large sample from their BBD on movie ratings by individual users. The contest concluded with a winning team that included computer scientists and a statistician ([Bell et al., 2010]). Interestingly, Netflix wanted to run another round of the contest, but was forced to cancel following a privacy

[6] www.adweek.com/news/technology/connect-thoughts-99712

lawsuit accusing them of indirectly exposing the movie preferences of Netflix users by publishing anonymized customer data[7]. While the company's initial goal for the contest was to improve their own recommendation system, this open contest ended up contributing to research on recommender systems but not providing an improved solution, because by the time the contest was over, Netflix was shifting from primarily DVD by mail to movie streaming, rendering the developed algorithm incompatible with their new type of BBD: "it turns out that the recommendation for streaming videos is different than for rental viewing a few days later"[8].

Another type of personalization studies performed by companies is choosing which content to display to a user. [Agarwal and Chen, 2016], who worked at Yahoo! Research and at LinkedIn, described two such applications, where the end goal was to serve the "best" content to users in an automated fashion to optimize metrics such as user engagement or ad click-through-ratio. They describe an explore-exploit approach that led to significant improvement in click-through rates on the Yahoo! Front page Today Module and the LinkedIn Today Module, as well as on LinkedIn self-serve Ads.

One of the earliest and still most common types of company-led studies is aimed at target marketing. This involves building predictive models, using BBD, for identifying which customers, employees, transactions, or other observations to act upon. E-commerce and mobile commerce companies use targeted marketing to serve customized offers, discounts, and coupons. A somewhat extreme example of using change in customers' behavioral data is the giant retailer Target, who made headlines in 2012 for its ability to identify when shoppers were pregnant. They did this by using predictive analytics to identify changes in shopping habits at early stages of pregnancy. The controversy apparently started following a case where the company's algorithm was able to predict a high school girl's pregnancy before her father did[9].

Some companies with BBD have run experiments to evaluate new features or to compare strategies. Amazon.com was one of the first online companies to massively use what is known as *A/B testing* - a two-level single factorial experiment. In one of their experiments, they manipulated the price of top-selling DVDs. However, users discovered the differential pricing scheme, sounded their anger, and the company discontinued the experiment and compensated customers[10]. Amazon.com continues to carry out A/B tests for studying various factors, such as new home page design, moving features around the page, different algorithms for recommendations, and changing search relevance rankings[11].

A/B testing is now also used in political campaigns to predict who should receive which message treatment. Campaigns now maintain extensive data on voters to help guide decisions about outreach to individual voters. Prior to the 2008 Obama campaign, the practice was to make rule-based decisions in accord with expert political judgment. Since 2008, it has increasingly been recognized that, rather than relying on judgment or supposition to determine whether an individual should be called, it is best to use the data to develop a model that can predict whether a voter will respond positively to outreach. This is typically done by first conducting a survey of voters to determine their inclination to vote for a certain party or candidate. Given the survey results, an A/B test is conducted, randomly promoting the party/candidate to half of the sample. The experiment is followed by another survey to evaluate whether voters' opinions have shifted. This combination of A/B testing with predictive models is known as *uplift modeling* ([Shmueli et al., 2016]).

Obtaining BBD for Research

Open Data and Publicly Available Data

In recent years there has been a growing number of governments and agencies making BBD publicly available. Websites such as data.gov, data.gov.uk, and data.taipei provide datasets collected by government: traffic accidents, consumer complaints, crimes, health surveys, and more. data.worldbank.org provides data on economic growth, education, etc. While this trend and the number of available datasets has been growing, data are often not easily available, due to limited APIs, inconvenient data formats (such as PDF files), and limiting sharing

[7] "Netflix Settles Privacy Lawsuit, Cancels Prize Sequel", Forbes.com, March 12, 2010.

[8] "Streaming has not only changed the way our members interact with the service, but also the type of data available to use in our algorithms." Netflix blog, Apr 6, 2012, http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html

[9] www.dailymail.co.uk/news/article-2102859/, from [Hardoon and Shmueli, 2013]

[10] "Amazon backs away from test prices", www.cnet.com, Jan 2, 2002

[11] "Amazon's business strategy and revenue model: A history and 2014 update", www.smartinsights.com, June 30, 2014.

rules ([Adar, 2015]). Also, many datasets are given at aggregation levels that do not support BBD research.

A growing number of companies have been making their data publicly available through simple download or through APIs. Twitter is probably one of the most heavily used BBD sources for researchers: one can download all tweets from the last 30 days. Amazon and eBay share some of their data via APIs. In contrast, Facebook does not provide data download.

Websites such as UCI Machine Learning Repository make available many datasets, some BBD. The datasets are heavily used by researchers in machine learning to test new algorithms.

A more recent publicly available source for BBD is data mining contest platforms such as kaggle.com and crowdanalytix.com. These platforms host contests for various companies who share a large dataset. Many of these contests include BBD. Examples include consumer ratings from the restaurant rating website yelp.com, Hillary Clinton's emails, customer bookings on airbnb.com, crimes in San Francisco, purchase and browsing behavior on ponpare.jp, restaurant bookings by customers on eztable.com.tw, and more.

Aside from download and the use of APIs, another tool commonly used by academic researchers is "web scraping" - automated programs that collect data from a website in a methodical way. Some websites disallow web scraping from some or all pages, by setting technological barriers and legal notices. Yet, many websites do tolerate web scraping by researchers, if they do not overload their servers or scrape massively. [Allen et al., 2006] discuss legal and ethical issues pertaining to web data collection and offer guidelines for academic researchers.

Institutional Review Board (IRB)

Collecting BBD through experiments or surveys typically requires researchers to obtain approval by an ethics committee. Academic researchers who conduct studies on human subjects are well familiar with Institutional Review Boards (IRB) and the process of obtaining IRB approval before carrying out a study that involves human subjects. The IRB is a university-level committee designated to approve, monitor, and review biomedical and behavioral research involving humans. In the United States, any university or body that receives Federal funds is required to have an IRB. Such "ethics committees" also exist in other countries with different names. The IRB performs a benefit-risk analysis for proposed studies, aimed at assuring that the study will potentially have a sufficient contribution to justify the risks for the human subjects involved. Guidelines focus on beneficence, justice, and respect for persons:

1. Risks to subjects are minimized (beneficence)
2. Risks are reasonable in relation to benefits (beneficence)
3. Selection of subjects is equitable (justice)
4. Provisions are adequate to monitor the data and ensure its confidentiality and the safety of subjects (beneficence)
5. Informed consent obtained and documented (respect for persons), including assuring information comprehension and voluntary agreement
6. Safeguards for vulnerable populations (respect for persons)

We note that *minimal risk* means that the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests[12].

Considerations Beyond IRB Approval'

Academic studies that collect BBD through experiments or surveys are usually required to obtain IRB approval, and top journals require the authors to confirm that their study has IRB approval.

We note that the university IRB sometimes does not require researchers to obtain IRB approval, even when human subjects are involved. The recent Facebook experiment ([Kramer et al., 2014]) which created controversy was followed by a note by the editor-in-chief of *Proceedings of the National Academy of Sciences (PNAS)*, the journal in which the paper was published. The editorial Expression of Concern stated:

> This paper represents an important and emerging area of social science research that needs to be approached with sensitivity and with vigilance regarding personal privacy issues. Questions have been raised about the principles of informed consent and opportunity

[12] "Code of Federal Regulations", www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html

> to opt out in connection with the research in this paper. The authors noted in their paper, “[The work] was consistent with Facebook’s Data Use Policy, to which all users agree prior to creating an account on Facebook, constituting informed consent for this research.” When the authors prepared their paper for publication in PNAS, they stated that: “Because this experiment was conducted by Facebook, Inc. for internal purposes, the Cornell University IRB [Institutional Review Board] determined that the project did not fall under Cornell’s Human Research Protection Program.” This statement has since been confirmed by Cornell University. Obtaining informed consent and allowing participants to opt out are best practices in most instances under the US Department of Health and Human Services Policy for the Protection of Human Research Subjects (the ”Common Rule”). Adherence to the Common Rule is PNAS policy, but as a private company Facebook was under no obligation to conform to the provisions of the Common Rule when it collected the data used by the authors, and the Common Rule does not preclude their use of the data. Based on the information provided by the authors, PNAS editors deemed it appropriate to publish the paper. It is nevertheless a matter of concern that the collection of the data by Facebook may have involved practices that were not fully consistent with the principles of obtaining informed consent and allowing participants to opt out.

[Adar, 2015] claims that the controversy around the Facebook emotional contagion study, as reflected by a broad variation in responses from the public, academics (where computational scientists tended to be more in favor compared to social scientists opposing it), the press, ethicists, and corporates, “demonstrates that we have not yet converged [to a] solution that can balance the demands of scientists, the public, and corporate interests.”

Crowdsourcing: Amazon Mechanical Turk (AMT)

The use of large online labor markets, such as Amazon Mechanical Turk (AMT), has been replacing the student population as laboratory subjects in social science experiments. The advantage of conducting experiments or surveys on AMT include easy access to a large, stable, and diverse subject pool, the low cost of doing experiments, and faster iteration between developing theory and executing experiments ([Mason and Suri, 2012]). Moreover, the large pool of subjects make it easier to conduct synchronous experiments, where multiple subjects must be present at the same time.

Among the labor platforms, AMT currently seems to be the most popular service used by academic researchers, as indicated by published papers. AMT is used for a variety of tasks, from using workers as experiment subjects, to survey respondents, as well as for cleaning data, tagging data, and performing other operations that humans are better at than computers (e.g., tagging the gender of people in photos). [Mason and Suri, 2012] illustrate the mechanics of putting a task on AMT, including recruiting subjects, executing the task, and reviewing the work that was submitted.

Partnering With a Company

Most of the studies described earlier were the result of a partnership between individual academic researchers and a company. The partnership can take different forms, from one where both parties are interested in the same question, to one where the data are purchased from the company, or obtained in exchange for another resource.

There are also partnerships between schools and companies or other organizations. The Living Analytics Research Lab is a joint research initiative between Singapore Management University (SMU) and Carnegie Mellon University (CMU) to conduct research on behavioral and social network analytics and behavioral experiments. LARC partners with companies and government organizations in Singapore such as Citi Asia Pacific, Resorts World Sentosa, and Starhub who make BBD available for research by the LARC researchers.

Large Behavioral Experiments: Issues to Consider

Four of the BBD-based studies mentioned earlier conducted randomized experiments: [Muchnik et al., 2014] manipulated the ratings of news article comments; [Hinz et al., 2015] provided the treatment group with a prestige good, while the control group did not receive it; [Bapna et al., 2016] gifted an anonymous browsing feature to a treatment group, while the control group continued to browse non-anonymously. [Kramer et al., 2014] manipulated the emotional content in users’ Facebook News Feed. In all these studies, the authors partnered with a company in order to perform the experiment. Such partnerships can take different forms, and can require NDAs, payment for data, and strict conditions regarding data sharing. In some studies, the authors work at the company or organization that is carrying out the experiment. This is more often the case in organizations that have a research culture or division.

Conducting large-scale behavioral randomized experiments poses challenges that differ from the industrial environment. We discuss practical, methodological, ethical, and moral issues that arise in this context.

Fast-Changing Environment

One major challenge is the fast-changing environment. In the book *Amazonia. Five years at the epicentre of the dot-com juggernaut*, [Marcus, 2014], an ex-Amazon.com employee, describes some of the challenges that Amazon encounters[13]:

> Amazon has a culture of experiments of which A/B tests are key components... These involve testing a new treatment against a previous control for a limited time of a few days or a week. The system will randomly show one or more treatments to visitors and measure a range of parameters such as units sold and revenue by category (and total), session time, session length, etc. The new features will usually be launched if the desired metrics are statistically significantly better. Statistical tests are a challenge though as distributions are not normal (they have a large mass at zero for example of no purchase) There are other challenges since multiple A/B tests are running every day and A/B tests may overlap and so conflict. There are also longer-term effects where some features are 'cool' for the first two weeks and the opposite effect where changing navigation may degrade performance temporarily. Amazon also finds that as its users evolve in their online experience the way they act online has changed. This means that Amazon has to constantly test and evolve its features.

Multiplicity and Scaling

[Agarwal and Chen, 2016] summarized several of the challenges in designing algorithms for computational advertising and content recommendation. One is the multivariate nature of outcomes, in multiple different contexts, with multiple objectives. He calls this "3Ms": Multi-response (Clicks, share, comments, likes), Multi-context (Mobile, Desktop, Email) modeling to optimize Multiple Objectives (Tradeoff in engagement, revenue, viral activities).

[Agarwal and Chen, 2016] also pointed out the challenge of scaling statistical model computations at runtime to avoid latency issues.

Spill-Over Effects

A methodological challenge with randomized experiments in BBD environments is that the treatment can sometimes affect the control group. This is especially challenging in social network environments, where control group members might become "contaminated" by the treatment through connections with members in the treatment group. [Bapna and Umyarov, 2015] who conducted an experiment on a social network emphasize:

> Researchers working on network experiments have to be careful in dealing with possible biases that can arise because of the presence of network structure among the peers of manipulated users... The failure to appropriately account for this intersection problem could introduce various kinds of biases threatening either internal or external validity of the experiment. How one deals with this issue depends on the particular methodology and design choices deployed by the researchers.

[Fienberg, 2015] pointed out the challenge of conducting experiments in network data: "How to design [a] network-based experiment with randomization and statistically valid results?". He raises the question of the role of randomization in this setting, and points out that even if the treatment and control members are chosen to be sufficiently far away as to avoid spill-over effects, analysis still must account for dependence among units.

Knowledge of Allocation and Gift Effect

Similar to clinical trials, where experiments are conducted on human subjects, a concern arises regarding the effect of subjects' knowledge of their allocation to

[13] from "Amazon's business strategy and revenue model: A history and 2014 update", www.smartinsights.com, June 30, 2014.

the treatment or control group on the outcome. Knowledge of allocation can also affect compliance levels of the subjects. In clinical trials solutions include blinding (single, double and even triple, where the person analyzing the data also does not know the allocation), and placebo. Blinding and placebo approaches can be difficult to employ in BBD experiments, especially when they are carried out online.

In an online environment, users can sometimes identify a manipulation and whether they belong to the manipulated group or not through various online communication channels. An example is the experiment by Amazon that manipulated prices of top-selling DVDs. Consumers quickly detected the price variations shown to different users.

[Hinz et al., 2015] discussed the issue of possible knowledge of the manipulation by subjects, and even conducted a survey to make sure users in both the treatment and control groups perceived their chances of receiving the premium gift as equal.

A related issue is a potential “gift effect”. In BBD experiments where the treatment group receives a gift or preferential treatment, it is possible that the treated members react to the act of receiving a gift rather than (or in addition) to the treatment itself. This is similar to the clinical trials scenario, where placebos are used to eliminate the effect. In the online dating experiment by [Bapna et al., 2016], the authors ruled out a gift effect by comparing the outcomes in the last week of the treatment month to the first week of the post-treatment month. Because the effect persisted in the last week of the treatment month and immediately disappeared in the first week of post-treatment month, they were able to rule out a gift effect.

Ethical and Moral Issues

The emotional contagion experiment by Facebook has highlighted ethical and moral issues that large-scale experiments on human subjects raise. The ease of running a large scale experiment quickly and at low cost holds the danger of harming many people at a quick rate. One suggestion to reduce such a risk is performing a small-scale pilot study to evaluate risk.

The growing use of crowdsourcing labor markets, such as Amazon Mechanical Turk, has raised issues related to fair treatment and payment to workers. A Wiki page[14] created by several AMT workers, and signed by over 60 researchers, provides guidelines for researchers (see also http://wiki.wearedynamo.org).

Observational BBD: Issues to Consider

In BBD environments, it is often impossible, unethical, or complicated to conduct randomized experiments. BBD studies therefore often rely on observational data.

Selection Bias

Inferring causality from observational data is tricky, due to possible selection bias that arises from individuals' self selecting the treatment group, where the choice of treatment group can also be driving the outcome. Inferring causality from observational data requires the use of specialized analysis methods as well as making assumptions about the selection mechanism. The two most common approaches are Propensity Score Matching (PSM) ([Rosenbaum and Rubin, 1983]) and the Heckman approach ([Heckman, 1979]). Both methods attempt to match the self-selected treatment group with a control group that has the same propensity to select the intervention. The methods differ mainly in terms of assuming that the selection process can be modeled using observable data (PSM), or it is unobservable (Heckman approach). With very rich BBD, it becomes more plausible to model the selection process, and therefore PSM is common in BBD studies.

[Lambert and Pregibon, 2007] described a self-selection challenge in the context of online advertising, where Google wanted to test a new feature, but could not randomize the advertisers that would receive the new feature. They authors mentioned the additional challenge in assessing “whether a new feature makes advertisers happier” due to “the irregularities in advertiser behavior that largely depend on business conditions... whether or not a new feature is introduced”. While propensity score matching can handle these issues, the authors note:

> Our main reservation about all variants of matching is the degree of care required in building the propensity score model and the degree to which the matched sets must balance the advertiser characteristics. If analysis is automated, then the care needed may not be taken. In our idealized view, we want our cake and we want to eat it too; specifically, we require an estimator that has good performance and that can be applied routinely by non-statisticians.

[14] http://wiki.wearedynamo.org/index.php/Guidelines_for_Academic_Requesters

Given the growing number of observational studies to infer causality by non-statisticians, both in industry and in academia, we fully agree that there is a need for more robust, automated, and user-understandable techniques. Moreover, matching techniques do not scale well to Big Data. Recently, [Yahav et al., 2016] developed a tree-based approach which offers an automated, data-driven, non-parametric, computationally scalable, and easy-to-understand alternative to PSM. They illustrated the usefulness of the tree-based approach in a variety of scenarios, such as heterogeneous treatment effects and a continuous treatment variable; the method also highlights pre-treatment variables that are unbalanced across the treatment and control groups, which helps the analyst draw insights about what might be driving the self-selection.

Simpson's Paradox

[Glymour et al., 1997] described another challenge that arises when using observational data for inferring causality: Simpson's paradox. Simpsons paradox describes the case where the direction of a causal effect is reversed in the aggregated data compared to the disaggregated data. Detecting whether Simpson's paradox occurs in a dataset used for decision making is therefore important. The issue of causality and Simpson's paradox was revisited in a recent discussion in the *The American Statistician* (Feb. 2014 issue). In the opening paper, [Armistead, 2014] suggested: "[w]hether causal or not, third variables can convey critical information about a first-order relationship, study design, and previously unobserved variables."

Given a large and rich dataset such as BBD, and a causal question, it is useful to be able to determine whether a Simpson's paradox is possible. [Shmueli and Yahav, 2014] introduced a method that uses Classification and Regression Trees for automated detection of potential Simpson's paradoxes in data with few or many potential confounding variables, and scales to large samples. Their approach relies on the tree structure and the location of the cause vs. the confounders in the tree.

Observational BBD Contaminated by Experiments

Because companies and other organizations constantly experiment with new features, observational data are typically contaminated by the effects of such experiments. Most often, there is no documentation of such experiments. If an experiment result in extreme behavior and outliers, then it might be possible to detect it. However, in most cases it is difficult to identify and clean the data appropriately.

Predictive Analytics and Sample Size and Dimension

When the study goal is predictive, the sample size and dimension needed for achieving the predictive level of interest will depend on the nature of the data. As mentioned earlier, [Junque de Fortuny et al., 2014] showed that the marginal increase in predictive power continues to grow significantly as more behavioral data are added (more measurements and more observations). They conclude that this property gives an advantage to large companies who have larger BBD than smaller companies. This also means that researchers should partner with companies that have sufficiently large and rich BBD in order to answer predictive questions.

Ethical and Moral Issues

As Netflix discovered by the privacy lawsuit, observational BBD, as anonymized and minimal as they might be, can reveal individuals and risk their well-being. Note that the contest BBD included nothing more than a movie ID, an anonymized user ID, and the rating that the user gave the movie. Even with these three variables per record an individual was revealed.

The implication for researchers using BBD is that they must protect the data very well. More distressing is the effect on reproducible research: to protect privacy, authors of BBD studies should not share the BBD with readers; yet, not sharing data underlying a study harms the important principle of reproducible research.

Moral issues arise when the study conclusions lead to operational actions that trade-off the company's interest with user well-being. A study by [Xiao and Benbasat, 2015] showed that product recommendations systems, which are now ubiquitous on websites, can be designed to produce their recommendations on the basis of benefiting e-commerce merchants rather than benefiting consumers.

Large Scale Surveys

We focused on two sources of BBD: large scale experiments and observational data. Both of these are unsolicited and are currently the major sources of BBD. We briefly mention another source for BBD: large scale online surveys. Large scale surveys, made possible by cheap and user-friendly online tools, now supplement observational and/or experimental BBD in many applications. Academic researchers and companies use such tools routinely to solicit opinions, preferences, and insights. In the studies we described earlier, several used surveys to supplement their analysis and robustness check (e.g., [Hinz et al., 2015] used a survey to ensure the treatment and control subjects perceived the chance of receiving a premium gift as equal).

Generalization

Platforms such as Amazon Mechanical Turk have further enhanced large survey deployment, with their large pool of available workers. While it is easier, cheaper, and faster to collect many responses using an online platform with a large worker population, non-sampling errors, and especially representativeness of the population of interest, are still a challenge. [Keiding and Louis, 2016] describe such challenges in the context of epidemiological surveys, and the same issues are relevant for survey BBD. They note:

> The central issue is whether conditional effects in the sample (the study population) may be transported to desired target populations. Success depends on compatibility of causal structures in study and target populations, and will require subject matter considerations in each concrete case.

Data Quality

Other challenges such as data quality, duplicate responses, and insincere responses persist and require different approaches at such large scale.

Paradata

Web-based surveys can collect not only the responses of individuals, but also detailed paradata – data on how the survey was accessed/answered: time stamps of when the invitation email was opened, when the survey was accessed, the duration for answering each question, etc. Some paradata is available even for non-responders and for those who started responding but did not submit the survey.

An example of the use of paradata directly related to the survey topic is the Survey of Adult Skills by the OECD[15], an international survey conducted in 33 countries as part of the Programme for the International Assessment of Adult Competencies (PIAAC). The survey measures the key cognitive and workplace skills needed for individuals to participate in society and for economies to prosper. By conducting the survey on a computer, data is captured directly on how the respondents manage the various tasks. These data are then used to evaluate computer skills.

Methodical Analysis Cycle of BBD

Large scale experiments are typically performed by companies, even when the study is in partnership with academic researchers. In a partnership scenario, the academic researchers typically create a methodical scientific process:

1. understand the context and BBD that the company collects
2. set up the research question and hypotheses (based on the literature and theoretization)
3. determine the needed experimental design
4. obtain IRB approval (if needed)
5. possibly perform a pilot experiment
6. communicate the experimental design with the company and assure feasibility
7. the company deploys the experiment and collects the data
8. the company shares the data with the researchers
9. the researchers analyze the data and arrive at conclusions
10. the researchers share the insights and conclusions with the company and with the research community
11. the company operationalizes the insights into actions to improve their business
12. the company deploys (ideally in collaboration with the researchers) an impact study to evaluate the change

This ideal process corresponds with the “life cycle view” by [Kenett, 2015], and [Hoerl et al., 2014]’s building blocks of statistical thinking. In practice, such a complete methodical process typically does not take place

[15] www.oecd.org/site/piaac/surveyofadultskills.htm

for various reasons. One reason is the lack of a clear roadmap such as the one spelled out above.

Another issue is the need for a framework to evaluate the potential of the behavioral experiment for answering the question of interest. The effort and risks involved in conducting a large scale behavioral experiment can be large. How can the researchers and the company know whether the BBD to be generated by the experiment can properly address the question of interest? (The same issue arises in company-driven experiments).

The Information Quality (InfoQ) framework by [Kenett and Shmueli, 2014] aims to address this problem. It provides guidelines for companies and researchers conducting large scale behavioral experiments as well as studies using observational data, where researchers can ask "what is the potential of this BBD to answer my research question?". The InfoQ framework examines each of four components: the study goal, the BBD, the analysis methods, and the utility. Because all four components are examined in a holistic way, the InfoQ framework can also help highlight and resolve potential conflicts in terms of company vs. academic researcher goals and utility.

Summary and Further Thoughts

The availability of large amounts of rich, granular behavioral data has been changing the way human-subject studies are conducted in industry and in academia. BBD arises from passive collection, experiments, surveys, and their combination. While the World Wide Web is an immense source of BBD, another recent source is The Internet of Things (IoT) - the network of physical objects embedded with electronics, software, sensors, and network connectivity - which enables these objects to collect and exchange data. BBD therefore now arises in fields and applications that earlier did not capture human behavior. Examples include manufacturing environments that now use employee monitoring systems, and domestic "smart heating" systems that learn the inhabitants' habits.

In contrast to classical statistical design of experiments, power analyses, and sampling designs that focus on efficient and minimal data collection, in the BBD environment data quantity is typically not a major constraint. The challenges that plague BBD studies include technical issues (e.g., data access, analysis scalability, a quick changing environment), methodological issues (e.g., sampling and selection bias, data contamination by undocumented experiments, lack of methodical life cycle), legal and ethical (e.g., privacy violation, risks to human subjects) and moral issues (misaligned goals of science and company, gains for companies and organizations at the expense of individuals, communities, societies, and science).

BBD and Privacy

BBD data and studies raise serious privacy-related concerns. First, publicly available BBD, even if very limited in terms of the number of variables, can disclose individuals in unanticipated ways, as Netflix discovered in the lawsuit against them. This can occur in sparse BBD, such as the ratings data that Netflix made public.

Social network data make privacy issues even more complicated. [Fienberg, 2011] describes the challenge of privacy protection for statistical network data and points the need for developing new ways to think about privacy protection for network data. [Cascavilla et al., 2015, Schwartz et al., 2016] show that censoring data on a social network, such as using initials instead of full names, does not necessarily maintain anonymity. They show examples of "unintentional information leakage" on Facebook, where posts that included only initials of individuals were de-anonymized when considering the comments on the post. The commentator's identity and their visible list of "friends" can (unintentionally) disclose the anonymized name.

Companies such as Google and Facebook have BBD on a broad set of user behaviors, because they offer a wide variety of services and because they constantly acquire companies that have BBD on more aspects of these users. For example, Facebook now owns WhatsApp, which means that its BBD includes not only information about a Facebook user's activity on facebook.com, but also the data from chats on WhatsApp. Integration of BBD across platforms poses even bigger threats to privacy and society. A "futuristic" scenario of such integration is described in the novel *The Circle* by [Eggers, 2013]:

> The Circle, run out of a sprawling California campus, links users' personal emails, social media, banking, and purchasing with their universal operating system, resulting in one online identity and a new age of civility and transparency.

The book raises not only privacy concerns, but also threats to society, human thought, and human interactions.

Generalizing from BBD Studies

In company studies that rely on their own collected BBD it can be reasonable to assume that results generalize to the company-specific population of interest. However, scientific studies based on a company's BBD do not necessarily lead to results that generalize to a larger population of interest. Sampling bias is therefore an important challenge that if overlooked can lead to wrong conclusions even with Big Data.

Generalization is also a concern when using BBD from a crowdsourced market such as AMT. One concern is whether the online community represents the offline community. In countries with very high Internet penetration rates this might not be critical, but in many countries online access is unavailable to entire populations, which differ markedly from the online "elite". Crowdsourcing platforms such as AMT have a population of workers that is driven by legal and technological constraints. Currently, workers are only from the USA and from India, and the legal minimal age for working at AMT is 18. According to [Mason and Suri, 2012], numerous studies show correspondence between the behavior of AMT workers and behavior offline or in other online contexts. They conclude:

> While there are clearly differences between Mechanical Turk and offline contexts, evidence that Mechanical Turk is a valid means of collecting data is consistent and continues to accumulate.

Company vs. Researcher Objectives

Many BBD-based studies published in top academic journals are based on a partnership between academic researchers and a company, where the data are obtained from the company's BBD. [Adar, 2015] notes that "there is rarely a perfect alignment between commercial and academic research interests. Clearly, the agenda of companies will focus [on] the kinds of questions they ask and consequently the kinds of data they capture: can I understand my customers and predict what they will do next?". It is therefore the responsibility of the academic researcher to carefully consider the potential impact of a study that is based on company BBD.

One way to ensure that the study is driven by the scientific goal is to make sure that answering the scientific question also benefits the company. [Adar, 2015] gives the example of the Facebook study, which offered a potential benefit to the company (e.g., understanding how your posting behavior varies from your friend's may be used to design interfaces or algorithms that encourage posting behavior), as well as to scientific inquiry (e.g., how emotional contagion may work). Another possible solution is purchasing the data from the company. A third option is to work with companies that do understand and see value in scientific research, such as companies that reward their staff for academic publications. Large corporations that have research divisions (such as Microsoft, Google, and IBM) have long traditions of publication and scientific-oriented research and even allow their research staff to perform independent research.

Another issue that distinguishes commercial from academic use of analytics is the near-complete focus of companies on predictive analytics, while academic research is focused more on causal modeling. Predictive modeling and assessment are necessary and beneficial for scientific development ([Shmueli, 2010, Shmueli and Koppius, 2011]). However, their use for theory building, evaluation, and development is different from the deployment of predictive analytics for commercial purposes. Companies are using predictive analytics for immediate actions such as direct marketing (which customer to send an offer to), customer churn (which customer to reach out to for retention), fraud detection (whether to accept a payment), Human Resources analytics (employee retention and employee training), personalized recommendations, and more. The latter provide immediate solutions and actions that are based on correlations, which are aimed at improving the company's operations in the short term. However, they do not provide an understanding of the underlying causes for problems such as employee churn or customer dissatisfaction. While we have advocated the use of predictive analytics in scientific work, it is equally important to advocate the use of explanatory modeling in the BBD industry.

Personal Thoughts

Lastly, I'd like to comment as a statistician who has collaborated with non-statistician colleagues and with companies in studies that use BBD. I've been blessed to work in this new environment with smart and creative colleagues, where challenges abound. Methodological challenges have led us to exciting new techniques and approaches. At the same time, the moral and ethical issues are troubling. Whereas in other domains it might be easier to say "I am just the statistician", when it comes to BBD, I constantly find myself debating ethical and moral issues that pertain to individuals, communities, societies, cultures, human nature and human interactions. Treating individuals as vectors of numbers can easily turn personalization efforts into de-personalization. And there is the "Law of unintended consequences" (also known as "The way to hell is paved with good intentions"): Even studies that aim to "do good", such as using education BBD to identify potential dropouts, can lead to more harm than intended by labeling students as failures before an event actually occurred (similar to crime analytics). It is therefore difficult to distinguish right from wrong. I try to be aware and give careful attention to the questions that we aim to answer, but the pace of research today and the inertia of collaborations does not leave much room for deep contemplation. The new reliance on companies for BBD research is both a challenge as well as an opportunity to "do good".

The Way Forward

As statisticians, we are not trained to consider ethical and moral dilemmas, yet we provide methods and knowledge that can have significant influence in this age of BBD. It is therefore imperative for statistics programs to include training on legal, ethical, and moral issues that arise in BBD collection and analysis. Recent changes have led social science programs to include programming and technical courses for their students. It is as important for technical programs, including statistics programs, to include courses on ethics and human subjects, which are the expertise of social scientists.

A better understanding of human behavior and interaction, along with their ethical and moral considerations, will also become invaluable to industrial statisticians, as the objects for which they have been designing experiments and monitoring methods become "smart objects" that measure human behavior, directly or indirectly.

In order to integrate statistical principles into BBD studies, and given the special challenges and pitfalls posed by BBD, a general and understandable framework is needed for BBD studies. Researchers and practitioners would benefit from guidelines that tie a BBD study goal with adequate data analysis methods and performance evaluation. Such a framework would also help bridge possibly-conflicting goals; it would create a common language; and provide a "checklist" of the main statistical principles that must be taken into account.

Acknowledgments

I thank the organizers of the 2016 Stu Hunter Conference for inviting me to gather my experiences and thoughts about BBD studies and to write this paper. Special thanks to Steve Fienberg, Ron Kenett, Boaz Shmueli, and Alon Wolf for their comments and suggestions on an earlier draft of the paper. This work was supported in part by grant 104-2410-H-007-001-MY2 from the Ministry of Science and Technology in Taiwan.

References

[Adar, 2015] Adar, E. (2015). The two cultures and big data research. *A Journal of Law and Policy for the Information Society*, 10 (3).

[Agarwal and Chen, 2016] Agarwal, D. K. and Chen, B.-C. (2016). *Statistical Methods for Recommender Systems*. Cambridge University Press.

[Allen et al., 2006] Allen, G. N., L., B. D., and Davis, G. B. (2006). Academic data collection in electronic environments: Defining acceptable use of internet resources. *MIS Quarterly*, 30 (3):599–610.

[Armistead, 2014] Armistead, T. (2014). Resurrecting the third variable: Critique of pearl's causal analysis of simpson's paradox. *The American Statistician*, 68 (2):1–7.

[Bapna et al., 2016] Bapna, R., Ramaprasad, J., Shmueli, G., and Umyarov, A. (2016). One-way mirrors in online dating: A randomized field experiment. *Management Science*.

[Bapna and Umyarov, 2015] Bapna, R. and Umyarov, A. (2015). Do your online friends make you pay? a randomized field experiment on peer influence in online social networks. *Management Science*, 61 (8):1902–1920.

[Bell et al., 2010] Bell, R. M., Koren, Y., and Volinsky, C. (2010). All together now: A perspective on the netflix prize. *Chance*, 23 (1):24–29.

[Belo et al., 2013] Belo, R., Ferreira, P., and Telang, R. (2013). Broadband in school: Impact on student performance. *Management Science*, 60 (2):265–282.

[Cascavilla et al., 2015] Cascavilla, G., Conti, M., Schwartz, D. G., and Yahav, I. (2015). Revealing censored information through comments and commenters in online social networks. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, pages 675–680, New York, NY, USA. ACM.

[Chetty et al., 2014] Chetty, R., Friedman, J, N., and Rockoff, J. E. (2014). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104 (9).

[Dhar et al., 2014] Dhar, V., Geva, T., Oestreicher-Singer, G., and Sundararajan, A. (2014). Prediction in economic networks. *Information Systems Research*, 25 (2):264–284.

[Eggers, 2013] Eggers, D. (2013). *The Circle*. Random House LLC.

[Fawcett, 2016] Fawcett, T. (2016). Mining the quantified self: Personal knowledge discovery as a challenge for data science. *Big Data*, 3 (4):249–266.

[Fienberg, 2006] Fienberg, S. (2006). Privacy and confidentiality in an e-commerce world: Data mining, data warehousing, matching and disclosure limitation. *Statistical Science*, 21 (2):143–154.

[Fienberg, 2011] Fienberg, S. (2011). The challenge of privacy protection for statistical network data. In *Proceedings of the 58th World Statistics Congress 2011*.

[Fienberg, 2015] Fienberg, S. (2015). The promise and perils of big data for statistical inference.

[Gao et al., 2014] Gao, G., Greenwood, B. N., McCullough, J., and Agarwal, R. (2014). Vocal minority and silent majority: How do online ratings reflect population perceptions of quality?. *MIS Quarterly*, 39 (3):565–589.

[Glymour et al., 1997] Glymour, C., Madigan, D., Pregibon, D., and Smyth, P. (1997). Statistical themes and lessons for data mining. *Data Mining and Knowledge Discovery*, 1:11–28.

[Hardoon and Shmueli, 2013] Hardoon, D. R. and Shmueli, G. (2013). *Getting Started With Business Analytics: Insightful Decision Making*. Taylor & Francis.

[Heckman, 1979] Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47 (1):153–161.

[Hill et al., 2006] Hill, S., Provost, F., and Volinsky, C. (2006). Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, 21 (2):256–276.

[Hinz et al., 2015] Hinz, O., Spann, M., and Hahn, I.-H. (2015). Can't buy me love...or can i? social capital attainment through conspicuous consumption in virtual environments. *Information Systems Research*, 26 (4):849–870.

[Hoerl et al., 2014] Hoerl, R., Snee, R., and De Veaux, R. (2014). Applying statistical thinking to 'big data' problems. *WIREs Comput Stat*, 6:222–232.

[Junque de Fortuny et al., 2014] Junque de Fortuny, E., Martens, D., and Provost, F. (2014). Predictive modeling with big data: Is bigger really better? *Big Data*, 1 (4):215–226.

[Keiding and Louis, 2016] Keiding, N. and Louis, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society, Series A*, 179 (2):1–28.

[Kenett, 2015] Kenett, R. (2015). Statistics: A life cycle view. *Quality Engineering*, 27:111–121.

[Kenett and Shmueli, 2014] Kenett, R. S. and Shmueli, G. (2014). On information quality. *Journal of the Royal Statistical Society, Series A*, 177 (1):3–38.

[Kramer et al., 2014] Kramer, A., Guillory, J., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academies of Sciences*, 111 (24):8788–8790.

[Lambert and Pregibon, 2007] Lambert, D. and Pregibon, D. (2007). More bang for their bucks: Assessing new features for online advertisers. In *Proceedings of the 1st International Workshop on Data Mining and Audience Intelligence for Advertising*, ADKDD '07, pages 7–15, New York, NY, USA. ACM.

[Lin et al., 2013] Lin, M., Lucas Jr., H., and Shmueli, G. (2013). Too big to fail: Large samples and the p-value problem. *Information Systems Research*, 24 (4):906–917.

[Marcus, 2014] Marcus, J. (2014). *Amazonia. Five years at the epicentre of the dot-com juggernaut*. The New Press, NY.

[Mason and Suri, 2012] Mason, W. and Suri, S. (2012). Conducting behavioral research on amazon's mechanical turk. *Behavior Research*, 44 (1):647–651.

[Moussa, 2015] Moussa, M. (2015). Monitoring employee behavior through the use of technology and issues of employee privacy in america. *SAGE Open*, 5 (2).

[Muchnik et al., 2014] Muchnik, L., Aral, S., and Taylor, S. (2014). Social influence bias: A randomized experiment. *Science*, 341 (6146):647–651.

[Rosenbaum and Rubin, 1983] Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70 (1):41–55.

[Schwartz et al., 2016] Schwartz, D., Yahav, I., and Silverman, G. (2016). News censorship in online social networks: A study of circumvention in the commentsphere. *Journal of the Association for Information Science and Technology*, forthcoming.

[Shmueli, 2010] Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25 (3):289–310.

[Shmueli et al., 2016] Shmueli, G., Bruce, P. C., and Patel, N. R. (2016). *Data Mining for Business Analytics: Concepts, Techniques, and Applications with XLMiner*. Wiley & Sons, 3rd edition.

[Shmueli and Koppius, 2011] Shmueli, G. and Koppius, O. (2011). To explain or to predict? *MIS Quarterly*, 35 (3):553–572.

[Shmueli and Yahav, 2014] Shmueli, G. and Yahav, I. (2014). The forest or the trees? tackling simpson's paradox with classification and regression trees.

[Steinberg, 2016] Steinberg, D. M. (2016). Industrial statistics: The challenges and the research. *Quality Engineering*, 28 (1):45–59.

[Watts, 2013] Watts, D. J. (2013). Computational social science: Exciting progress and future directions. *The Bridge on Frontiers of Engineering*, 43 (4):5–10.

[Xiao and Benbasat, 2015] Xiao, B. and Benbasat, I. (2015). Designing warning messages for detecting biased online product recommendations: An empirical investigation. *MIS Quarterly*, 26 (4).

[Yahav et al., 2016] Yahav, I., Shmueli, G., and Mani, D. (2016). A tree-based approach for addressing self-selection in impact studies with big data. *MIS Quarterly*, forthcoming.