Testimony of Michael Kearns
Professor, University of Pennsylvania
Before the Subcommittees on Communications and Technology and Digital Commerce and Consumer Protection Of the Committee on Energy and Commerce
"Algorithms: How Companies' Decisions About Data and Content Impact Consumers"
November 29, 2017

Introduction

Chairmen Blackburn and Latta, Ranking Members Doyle and Schakowsky, and other distinguished Members of the Subcommittees, thank you for the opportunity to testify at this important hearing. My name is Michael Kearns, and I am a Computer and Information Science Professor at the University of Pennsylvania. I am appearing in a personal capacity today, and the views I express are my own. I have been an active and leading researcher and educator in the field of machine learning since the late 1980s, and, in addition to my academic work, I have consulted extensively on the use of machine learning in the technology and finance industries.

The fields of machine learning and artificial intelligence now play a central role in virtually every domain of science, technology, and business in which large data sets and challenging prediction problems are present. The number of instances in which the use of machine learning has provided tangible societal benefits, such as in medical diagnosis and, more recently, agriculture, is large and growing. Machine learning is also used in consumer-friendly activities such as detecting fraudulent banking or credit card activity.

Machine learning also increasingly plays a central role in the data collection and use practices of consumer-facing technology companies. Today I will discuss "data intimacy," the notion that machine learning enables companies to routinely draw predictions and inferences about consumers that go far deeper than the face-value of data collected as part of consumers' online activities.

It is not simply a question of whether consumer-facing technology companies are collecting large volumes of data; such companies are collecting information that provides, or allow inferences regarding, intimate details about our personal lives. Search engine queries permit inferences about our financial, physical, and psychological conditions. Social media users routinely reveal opinions, beliefs, or affiliations that might carry social stigma, and that they would be more reluctant to reveal in everyday life.

For example, a recent study showed that, using machine learning, anonymous social relationship data permits accurate identification of romantic partners for over 55% of users --- orders of magnitude higher than random guessing.[1] Another study concluded that Facebook's data, algorithms, and models are capable of identifying social relationships of which its users are themselves unaware.[2] It has also

[1] Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook, L. Backstrom and J. Kleinberg, Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing, 2014, available at: https://arxiv.org/pdf/1310.6753v1.pdf. *See also*, Facebook Inches Closer to Finding the Formula for Love, *Wired Magazine*, November 12, 2014, available at: https://www.wired.com/2013/11/can-facebook-really-predict-our-love-lives/.

[2] Facebook Figured Out My Family Secrets, and it Won't Tell Me How, Gizmodo, August 25, 2017, available at:

been established that religious and political beliefs can be accurately predicted from apparently unrelated social, search, and shopping activity undertaken by consumers.

The volume, diversity, intimacy, and modeling of consumers' information has substantial, and rapidly evolving, consequences for consumer privacy and implications for public policy. I will now provide an overview of how machine learning works, the value of information derived through machine learning, and how machine learning is utilized to predict consumer preferences and behaviors.

Machine Learning

Machine learning is the modern science underlying the construction of large-scale predictive models from massive data sets. It is a mixture of topics from areas as diverse as statistics, probability theory, pattern recognition, algorithms, artificial intelligence, and, most recently, distributed systems.

While its origins lie in the 1980s, in recent years, the data explosion enabled by the Internet has made machine learning one of the most important scientific fields, and one that has even entered the popular consciousness. The original efforts to catalog consumers' use of the Internet through hand-coded human expertise or knowledge were quickly overwhelmed by the exponential growth of consumers' online activities. As a result, machine learning has been employed to sift through vast volumes of data to improve the algorithms used for search results and other Internet-related queries.

The algorithms of machine learning and the models they produce are largely automated once in operation. But the development of these algorithms, their improvement and evolution, their implementation in a distributed, cloud-based computing environment, and their specialization to the idiosyncrasies of new and ever-changing data sets remains a highly technical, research-intensive, and human-centric activity. Machine learning has enabled technology companies to create highly predictive models for collective and individual consumer behavior, and to make subtle and accurate inferences about consumers' interests and preferences.

Machine Learning Process

The first step in the machine learning process is known as "feature extraction" or "feature engineering," which are the terms used to describe processes that transform the raw data streams into higher-level abstractions that have more structure, and encode more directly the underlying meaning and intent in the data. Examples include identifying objects and edges in images, or parsing an English sentence in a social media post. The development of algorithms for such feature extractions is actually extremely challenging, and has been the source of many decades of intense research.

Feature engineering turns the raw, unstructured data streams into structured objects with more meaningful and informative representations that are also much more amenable to machine understanding. For many machine learning tasks, the next step is to annotate such data with user feedback, which in the field's terminology is sometimes referred to as "labels" or "supervision." The basic idea is that if individual data items or events (such as sentences, photos, documents, or web pages) can be identified as relevant or irrelevant, good or bad, etc., then one can use sample data to train a predictive statistical model.

http://gizmodo.com/facebook-figured-out-my-family-secrets-and-it-wont-tel-1797696163.

The combination of feature extraction with user feedback or supervision sets up a classical statistical modeling problem: the raw user streams have now been transformed into <x,y> pairs, where x is some structured representation of complex data items like documents, sentences or images, and y is a signal indicating whether x is “good,” “bad,” or in between. The challenge is then to take a (very) large sample of such data pairs, and build a predictive model -- i.e. a model that, given a new, previously unseen x, can accurately predict the associated feedback y. This challenge is precisely the domain of modern machine learning. An example of the end-product of the machine learning process would be a model that takes as input all of a user’s activity on a social network, search engine, or shopping service, and outputs predictions of the ads in which the user would be most interested.

Value of Data Volume, Diversity and Intimacy

Machine learning and artificial intelligence would not be effective without large data sets to analyze. Consumer-facing technology companies in the United States have amassed an almost unimaginable set of data about consumers, both collectively and individually, which enable machine learning and artificial intelligence to draw conclusions about consumer behavior and preferences.

As mentioned above, these large and diverse data sets are the foundation for effective algorithms. Companies compete vigorously to acquire these diverse data sets to support their machine learning capabilities through the development of services as well as acquisitions. For example, search engines provide vast amounts of data about consumers’ interests and the manner in which they conduct searches. Search data can reveal consumers’ financial, medical, and mental conditions. Similarly, mobile operating system data provides a treasure trove of information regarding virtually everything a consumer does on a mobile device.

Services that consolidate location data also provide companies with vast information about consumers’ physical location, and enable such companies to develop inferences about consumers’ activities based upon those locations. Such location information goes far beyond what the GPS receiver in a consumer’s mobile device may divulge because they are amassed using WiFi, Bluetooth, and other technologies as well. Some services seek not only to determine where users are, but where they will be or plan to be in the future; examples include calendar apps, flight and travel shopping services, and navigation apps.

Social media platforms also provide substantial amounts of raw data. In addition to knowing with whom a consumer affiliates directly, social media platforms are able to accumulate information about who a consumer follows or what he or she likes. However, while the quantity of data is critical to develop accurate algorithms, the quality (and intimacy) of such data is important to discern consumer preferences and behaviors.

Use of Machine Learning to Predict Consumer Preferences and Behaviors

Increasingly, machine learning-based algorithms are utilized not only to determine consumer purchasing habits, but also to determine consumers’ emotions. While these algorithms are employed most commonly (and pervasively) to target advertising, as we’ve seen in media recently, such algorithms are also being utilized to generate (or incite) certain emotional responses.

For example, beginning with a substantial set of raw data, researchers recently used a sophisticated “dimensionality reduction” method know as singular value decomposition to automatically extract a much smaller set of informative "features" to represent each consumer.[3] These feature representations were in turn given to a standard statistical algorithm to produce predictive models for each of the targeted categories (sexual orientation, race, political party, etc.). Thus, the raw data on the collective population is transformed into a higher-level model that permits accurate (and intrusive) inferences about specific individuals that were not present in their raw data at all. This model, and the highly detailed information produced by the model, is developed almost entirely through machine learning methods. Other recent research has demonstrated the extent to which people use search engines to express their most private and intimate thoughts and concerns, as though they were entirely unobserved.[4] When combined with other data sources and the use of machine learning, the detailed insights and predictions that are possible are effectively unlimited.

Public Policy Implications of Machine Learning

From a privacy perspective, perhaps the most important overarching conclusion is that the "intimacy" of consumer data cannot be measured by the number of bits crossing a pipe, or similarly crude metrics that fail to account for the nature, diversity, and content of the data and its potential uses for modeling and inference. It is both possible and common that the highest volume data sources (such as the fragmented and possibly encrypted packets passing through a core router in the Internet) can reveal virtually nothing about the consumers who generate that traffic, whereas much lower-volume and more-specialized data sources can both directly and indirectly reveal the most private and personal details about consumers. In fact, the widespread application of machine learning to specialized consumer data sources is deliberately designed to extract personal and actionable insights about both individual users and collective behaviors.

Thus, it would be wrong to formulate privacy policy or metrics based only on the amount or apparent source of data --- one must evaluate the sensitivity of the data as well as anticipate how private or intimate the *inferences* that could be made from the data might be. And such anticipation, for policymakers or computer scientists, is extremely challenging.

This challenge argues for a privacy framework that comprehensively covers the diverse range of data being used commercially, and applies consistent privacy requirements. Policymakers should also take a forward-looking approach to privacy, and not overly focus on specific data types or practices (which are likely to become obsolete shortly due to the rapidly changing nature of technology). A technology-neutral approach can adapt quickly to new technical and market developments.

Conclusion

Thank you again for the opportunity testify before you today. Machine learning and artificial intelligence present significant challenges for policymakers because of the rapidly evolving nature of the

[3] Private Traits and Attributes are Predictable from Digital Records of Human Behavior, M. Kosinski, D. Stillwell, and T. Graepel, Proceedings of the National Academy of Sciences, 110(15), 2013, available at: http://www.pnas.org/content/110/15/5802.full.pdf.

[4] *See e.g.*, Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are, Harper Collins, 2017; Essays Using Google Data, Seth Stevens-Davidowitz, Doctoral Thesis, Harvard University 2013, available at: https://dash.harvard.edu/bitstream/handle/1/10984881/StephensDavidowitz_gsas.harvard_0084L_11016.pdf?sequence=1).%25C2%25A0); links to published research articles and items published in the *New York Times*, available at: http://sethsd.com/.

technology, as well as its pervasive use among consumer-facing technology companies to predict consumer preferences and draw inferences about intimate aspects of consumers lives. While policymakers should be mindful that machine learning and artificial intelligence also produce many of the sizeable benefits inherent in consumers' online experiences, such technology enables companies to shape commerce, and even belief and emotions. This hearing is therefore an important opportunity for the Members of these Subcommittees to understand and evaluate the risks inherent in such technology.