POLICY FORUM

Data, privacy, and the greater good

Eric Horvitz[1]* and Deirdre Mulligan[2]*

Large-scale aggregate analyses of anonymized data can yield valuable results and insights that address public health challenges and provide new avenues for scientific discovery. These methods can extend our knowledge and provide new tools for enhancing health and wellbeing. However, they raise questions about how to best address potential threats to privacy while reaping benefits for individuals and to society as a whole. The use of machine learning to make leaps across informational and social contexts to infer health conditions and risks from nonmedical data provides representative scenarios for reflections on directions with balancing innovation and regulation.

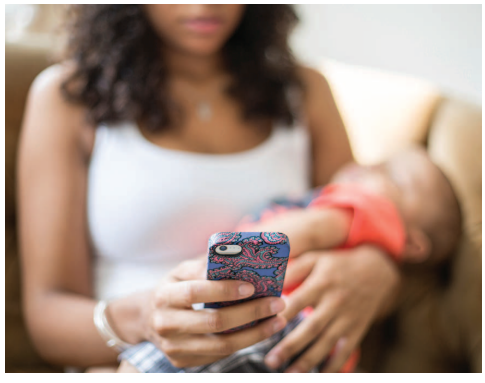
What if analyzing Twitter tweets or Facebook posts could identify new mothers at risk for postpartum depression (PPD)? Despite PPD's serious consequences, early identification and prevention remain difficult. Absent a history of depression, detection is largely dependent on new mothers' self-reports. But researchers found that shifts in sets of activities and language usage on Facebook are predictors of PPD (*1*) (see the photo). This is but one example of promising research that uses machine learning to derive and leverage health-related inferences from the massive flows of data about individuals and populations generated through social media and other digital data streams. At the same time, machine learning presents new challenges for protecting individual privacy and ensuring fair use of data. We need to strike a new balance between controls on collecting information and controls on how it is used, as well as pursue auditable and accountable technologies and systems that facilitate greater use-based privacy protections.

Researchers have coined terms, such as digital disease detection (*2*) and infodemiology (*3*), to define the new science of harnessing diverse streams of digital information to inform public health and policy, e.g., earlier identification of epidemics, (*4*) modeling communicability and flow of illness (*5*), and stratifying individuals at risk for illness (*6*). This new form of health research can also inform and extend understandings drawn from traditional health records and human subjects research. For example, the detection of adverse drug reactions could be improved by jointly leveraging data from the U.S. Food and Drug Administration's Adverse Event Reporting System and anonymized search logs (*7*). Search logs can serve as a large-scale sensing system that can be used for drug safety surveillance—pharmacovigilance.

Infodemiology studies are typically large-scale aggregate analyses of anonymized data—publicly disclosed or privately held—that yield results and insights on public health questions across populations. However, some methods and models can be aimed at making inferences about unique individuals that could drive actions, such as alerting or providing digital nudges, to improve individual or public health outcomes.

[1]Microsoft Research, Redmond, WA 98052, USA. [2]University of California, Berkeley, Berkeley, CA 94720, USA.
*Corresponding author. E-mail: horvitz@microsoft.com (E.H.); dmulligan@berkeley.edu (D.M.)

Although digital nudging shows promise, a recent flare-up in the United Kingdom highlights the privacy concerns it can ignite. A Twitter suicide-prevention application called Good Samaritan monitored individuals' tweets for words and phrases indicating a potential mental health crisis. The app notified the person's followers so they could intervene to avert a potential suicide. But the app was shuttered after public outcry drew regulator concern (*8*). Critics worried the app would encourage online stalkers and bullies to target vulnerable individuals and collected 1200 signatures on a petition arguing that the app breached users' privacy by collecting, processing, and sharing sensitive information. Despite the developers' laudable goal of preventing suicide, the nonprofit was chastised for playing fast and loose with the privacy and mental health of those it was seeking to save (*9*).



Machine learning can make "category-jumping" inferences about health. New mother's activities and language usage on social media are predictors of postpartum depression.

Machine learning can facilitate leaps across informational and social contexts, making "category-jumping" inferences about health conditions or propensities from nonmedical data generated far outside the medical context. The implications for privacy are profound. Category-jumping inferences may reveal attributes or conditions an individual has specifically withheld from others. To protect against such violations, the United States heavily regulates health care privacy. But, although information about health conditions garnered from health care treatment and payment must be handled in a manner that respects patient privacy, machine learning and inference can sidestep many of the existing protections.

Even when not category-jumping, machine learning can be used to draw powerful and compromising inferences from self-disclosed, seemingly benign data or readily observed behavior. These inferences can undermine a basic goal of many privacy laws—to allow individuals to control who knows what about them. Machine learning and inference makes it increasingly difficult for individuals to understand what others can know about them based on what they have explicitly or implicitly shared. And these computer-generated channels of information about health conditions join other technically created fissures in existing legal protections for health privacy. In particular, it is difficult to reliably deidentify publicly shared data sets, given the enormous amount and variety of ancillary data that can be used to reidentify individuals.

The capacities of machine learning expose the fundamental limitations of existing U.S. privacy rules that tie the privacy protection of an individual's health status to specific contexts or specific types of information a priori identified as health information. Health privacy regulations and privacy laws in the United States generally are based on the assumption that the semantics of data are relatively fixed and knowable and reside in isolated contexts. Machine learning techniques can instead be used to infer new meaning within and across contexts and is generally unencumbered by privacy rules in the United States. Using publicly available Twitter posts to infer risk of PPD, for example, does not run afoul of existing privacy law. This might be unsurprising, and seem unproblematic, given that the posts were publicly shared, but there are troubling consequences.

Current privacy laws often do double duty. At a basic level, they limit who has access to information about a person. This implicitly limits the extent to which that information influences decision-making and thus doubles as a limit on the opportunities for information to fuel discrimination. Because of the heightened privacy sensitivities and concerns with health-related discrimination, we have additional laws that regulate the use of health information outside the health care context. U.S. laws specifically limit the use of some health information in ways considered unfair. For example, credit-reporting agencies are generally prohibited from providing medical information to make decisions about employment, credit, or housing. The Americans with Disabilities Act (ADA) prohibits discrimination on the basis of substantial physical or mental disabilities—or even a mistaken belief that an individual suffers from such a disability. If machine learning is used to infer that an individual suffers from a physical or mental impairment, an employer who bases a hiring decision on it, even if the inference is wrong, would violate the law.

But the ADA does not prohibit discrimination based on predispositions for such disabilities (*10*). Machine learning might discover those, too. In theory, the Genetic Information Non-Discrimination Act (GINA) should fill this gap by protecting people genetically predisposed to a disease. But again, machine learning exposes cracks in this protection. Although GINA prohibits discrimination based on information derived from genetic tests or a family history of a disease (*11*), it does not limit the use of information

about such a disposition—even if it is grounded in genetics—inferred through machine learning techniques that mine other sorts of data. In other words, machine learning that predicts future health status from nongenetic information—including health status changes due to genetic predisposition—would circumvent existing legal protections (*12*).

Just as machine learning can expose secrets, it facilitates social sorting—placing individuals into categories for differential treatment—with good or bad intent and positive or negative outcomes. The methods used to classify individuals as part of beneficial public health programs and nudges can just as easily be used for more nefarious purposes, such as discrimination to protect organizational profits.

Policy-makers in the United States and elsewhere are just beginning to address the challenges that machine learning and inference pose to commitments to privacy and equal treatment. Although not specifically focused on health information, reports issued by the White House—discuss the potential for large-scale data analyses to result in discrimination (*13*)—and the Federal Trade Commission (FTC) have suggested new efforts to protect privacy, regulate harmful uses of information, and increase transparency.

The FTC is the key agency policing unfair and deceptive practices in the commercial marketplace, including those that touch on the privacy and security of personal information. Its proposed privacy framework encourages companies to combine technical and policy mechanisms to protect against reidentification. The FTC's proposed rules would work to ensure that data are both "not reasonably identifiable" and accompanied by public company commitments not to reidentify it. The same privacy rules should apply to downstream users of the data (*14*). This approach is promising for machine learning and other areas of artificial intelligence that rely on data-centric analyses. It allows learning from large data sets—and sharing them—by encouraging companies to reduce the risks that data pools and data sharing pose for individual privacy.

The FTC proposal grows, in part, from recent agency actions focused on inferences that we have deemed "context-jumping." In one high-profile case, Netflix publicly released data sets to support a competition to improve their recommendation algorithm. When outside researchers used ancillary data to reidentify and infer sensitive attributes about individuals from the Netflix data sets, the FTC worked with the company to limit future public disclosures—setting out the limits discussed above. In a similar vein, the FTC objected to a change in Facebook's defaults that exposed individuals' group affiliations from which sensitive information, such as political views and sexual orientation, could be inferred (*15*).

Additionally, the FTC has made efforts to ensure that individuals can control tracking in the online and mobile environments. These are in part due to the nonobvious inferences that can be drawn from vast collections of data (*16*–*18*) and the subsequent risks to consumers, who may be placed in classifications that single them out for specific treatment in the marketplace (*19*, *20*). In a related context, the FTC recommended that Congress require data brokers—companies that collect consumers' personal information and resell or share that information with others—to clearly disclose to consumers information about the data they collect, as well as the fact that they derive inferences from it (*21*). Here, too, the FTC appears concerned with not just the raw data, but inferences from its analysis.

The Obama Administration's Big Data Initiative has also considered the risks to privacy posed by machine learning and the potential downsides of using machine inferences in the commercial marketplace (*22*, *23*), concluding that we need to update our privacy rules, increase technical expertise in consumer protection and civil rights agencies to address novel discrimination issues arising from big data, provide individuals with privacy preserving tools that allow the to control the collection and manage the use of personal information, as well as increase transparency into how companies use and trade data. The Administration is also concerned with the use of machine learning in policing and national security. The White House report called for increased technical expertise to help civil rights and consumer protection agencies identify, investigate, and resolve uses of big data analytics that have a discriminatory impact on protected classes (*24*).

Note that reports and proposals from the Administration distinctly emphasize policies and regulations focused on data use rather than collection. While acknowledging the need for tools that allow consumers to control when and how their data is collected, the Administration recommendations focus on empowering individuals to participate in decisions about future uses and disclosures of collected data (*25*). A separate report by the President's Council of Advisors on Science and Technology (PCAST) concluded that this was a more fruitful direction for technical protections. Both reports suggest that use-based protections better address the latent meaning of data—inferences drawn from data using machine learning—and can adapt to the scale of the data-rich and connected environment of the future (*26*). The Administration called for collaborative efforts to ensure that regulations in the health context will allow society to reap the benefits and mitigate the risks posed by machine learning and inferences. Use-based approaches are often favored by industry, as well, which tends to view data as akin to a natural resource to be mined for commercial and public benefit, and industry is resistant to efforts to constrain data collection.

Although incomplete and unlikely to be acted upon by the current gridlocked Congress, adoption of these recommendations would increase transparency about data's collection, use, and consequences. Along with efforts to identify and constrain discriminatory or unfair uses of data and inferences, they are promising steps. They also align with aspects of existing European privacy laws concerned with the transparency and fairness of data processing, particularly the risks to individuals of purely automated decision-making.

Current European Union (EU) law requires entities to provide individuals with access to the data on which decisions are rendered, as well as information about decision criteria [see Articles 12 and 15 of (*21*)]. Although currently governed by a Europe-wide directive, both provisions are a matter of national law. What exactly individuals receive when they request access to their data and to processing logic varies by country, as does the implementation of the limitation on "purely automated" processing. The EU is expected to adopt a data privacy regulation that will supplant local law, with a single national standard. Although the current draft includes parallel provisions, their final form is not yet known nor is how they will ultimately be interpreted (*27*). In theory, a new EU requirement to disclose the logic of processing could apply quite broadly, with implications for public access to data analytics and algorithms. In the interim, a decision expected this summer in a case before the European Court of Justice may provide some detail as to what level of access to both data and the logic of processing is currently required under the EU Directive (*28*).

Improving the transparency of data processing to data subjects is both important and challenging. Although the goal may be to promote actual understanding of the workings or likely outputs of machine learning and reasoning methods, the workflows and dynamism of algorithms and decision criteria may be difficult to characterize and explain. For example, popular convolutional neural-network learning procedures (commonly referred to as "deep learning") automatically induce rich, multilayered representations that their developers themselves may not understand with clarity. Although high-level descriptions of procedures and representations might be provided, even an accomplished programmer with access to the source code would be unable to describe the precise operation of such a system or predict the output of a given set of inputs.

Data's meaning has become a moving target. Data sets can be easily combined to reidentify data sets thought deidentified, and sensitive knowledge can be inferred from benign data that are routinely and promiscuously shared. These pose difficulties for current U.S. legal approaches to privacy protection that regulate data on the basis of its identifiability and express meaning.

Use-based approaches are driven, in part, by the realization that focusing solely on limiting data collection is inadequate. In a way, this presupposes that data are an unalloyed good that should be collected on principle, whenever and wherever possible. Whereas we are not ready to abandon limits on data collection, we agree that use-based regulations, although challenging to implement, are an important part of the future legal landscape—and will help to advance privacy, equality, and the public good. To advance transparency and to balance the constraints they impose, use-based approaches would need to emphasize access, accuracy, and correction rights for individuals.

The evolution of regulations for health information, although incomplete, provides a useful map for thinking about the challenges and opportunities we face today and frames potential solutions. In health care, privacy rules were joined by nondiscrimination rules and always were accompanied by special provisions to support research. Today, they are being joined by collective governance models designed to encourage pooling of data in biobanks that support research on health conditions while protecting collective interests in privacy.

Despite practical challenges, we are hopeful that informed discussions among policy-makers and the public about data and the capabilities of machine learning, will lead to insightful designs of programs and policies that can balance the goals of protecting privacy and ensuring fairness with those of reaping the benefits to scientific research and to individual and public health. Our commitments to privacy and fairness are evergreen, but our policy choices must adapt to advance them, and support new techniques for deepening our knowledge.

REFERENCES AND NOTES

1. M. De Choudhury, S. Counts, E. Horvitz, A. Hoff, in *Proceedings of International Conference on Weblogs and Social Media* [Association for the Advancement of Artificial Intelligence (AAAI), Palo Alto, CA, 2014].
2. J. S. Brownstein, C. C. Freifeld, L. C. Madoff, *N. Engl. J. Med.* **360**, 2153–2155 (2009).
3. G. Eysenbach, *J. Med. Internet Res.* **11**, e11 (2009).
4. D. A. Broniatowski, M. J. Paul, M. Dredze, *PLOS ONE* **8**, e83672 (2013).
5. A. Sadilek, H. Kautz, V. Silenzio, in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* (AAAI, Palo Alto, CA, 2012).
6. M. De Choudhury, S. Counts, E. Horvitz, in *Proceedings of the* SIGCHI *Conference on Human Factors in Computing Systems* (Association for Computing Machinery, New York, 2013), pp. 3267–3276.
7. R. W. White, R. Harpaz, N. H. Shah, W. DuMouchel, E. Horvitz, *Clin. Pharmacol. Ther.* **96**, 239–246 (2014).
8. Samaritans Radar; www.samaritans.org/how-we-can-help-you/supporting-someone-online/samaritans-radar.
9. Shut down Samaritans Radar; http://bit.ly/Samaritans-after.
10. U.S. Equal Employment Opportunity Commission (EEOC), 29 Code of Federal Regulations (C.F.R.), 1630.2 (g) (2013).
11. EEOC, 29 CFR 1635.3 (c) (2013).
12. M. A. Rothstein, *J. Law Med. Ethics* **36**, 837–840 (2008).
13. Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values* (White House, Washington, DC, 2014); http://1.usa.gov/1TSOhiG.
14. Letter from Maneesha Mithal, FTC, to Reed Freeman, Morrison, & Foerster LLP, Counsel for Netflix, 2 [closing letter] (2010); http://1.usa.gov/1GCFyXR.
15. In re Facebook, Complaint, FTC File No. 092 3184 (2012).
16. FTC Staff Report, *Mobile Privacy Disclosures: Building Trust Through Transparency* (FTC, Washington, DC, 2013); http://1.usa.gov/1eNz8zr.
17. FTC, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers* (FTC, Washington, DC, 2012).
18. Directive 95/46/ec of the European Parliament and of The Council of Europe, 24 October 1995.
19. L. Sweeney, Online ads roll the dice [blog]; http://1.usa.gov/1KgEcYg.
20. FTC, "Big data: A tool for inclusion or exclusion?" (workshop, FTC, Washington, DC, 2014); http://1.usa.gov/1SR65cv
21. FTC, *Data Brokers: A Call for Transparency and Accountability* (FTC, Washington, DC, 2014); http://1.usa.gov/1GCFoj5.
22. J. Podesta, "Big data and privacy: 1 year out" [blog]; http://bit.ly/WHsePrivacy.
23. White House Council of Economic Advisers, *Big Data and Differential Pricing* (White House, Washington, DC, 2015).
24. Executive Office of the President, *Big Data and Differential Processing* (White House, Washington, DC, 2015); http://1.usa.gov/1eNy7qR.
25. Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values* (White House, Washington, DC, 2014); http://1.usa.gov/1TSOhiG.
26. President's Council of Advisors on Science and Technology (PCAST), *Big Data and Privacy: A Technological Perspective* (White House, Washington, DC, 2014); http://1.usa.gov/1C5ewNv.
27. European Commission, Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), COM(2012) 11 final (2012); http://bit.ly/1Lu5POv.
28. *M. Schrems* v. *Facebook Ireland Limited*, §J. Unlawful data transmission to the U.S.A. ("PRISM"), ¶166 and 167 (2013); www.europe-v-facebook.org/sk/sk_en.pdf.

10.1126/science.aac4520

REVIEW

Machine learning: Trends, perspectives, and prospects

M. I. Jordan[1]* and T. M. Mitchell[2]*

Machine learning addresses the question of how to build computers that improve automatically through experience. It is one of today's most rapidly growing technical fields, lying at the intersection of computer science and statistics, and at the core of artificial intelligence and data science. Recent progress in machine learning has been driven both by the development of new learning algorithms and theory and by the ongoing explosion in the availability of online data and low-cost computation. The adoption of data-intensive machine-learning methods can be found throughout science, technology and commerce, leading to more evidence-based decision-making across many walks of life, including health care, manufacturing, education, financial modeling, policing, and marketing.

Machine learning is a discipline focused on two interrelated questions: How can one construct computer systems that automatically improve through experience? and What are the fundamental statistical-computational-information-theoretic laws that govern all learning systems, including computers, humans, and organizations? The study of machine learning is important both for addressing these fundamental scientific and engineering questions and for the highly practical computer software it has produced and fielded across many applications.

Machine learning has progressed dramatically over the past two decades, from laboratory curiosity to a practical technology in widespread commercial use. Within artificial intelligence (AI), machine learning has emerged as the method of choice for developing practical software for computer vision, speech recognition, natural language processing, robot control, and other applications. Many developers of AI systems now recognize that, for many applications, it can be far easier to train a system by showing it examples of desired input-output behavior than to program it manually by anticipating the desired response for all possible inputs. The effect of machine learning has also been felt broadly across computer science and across a range of industries concerned with data-intensive issues, such as consumer services, the diagnosis of faults in complex systems, and the control of logistics chains. There has been a similarly broad range of effects across empirical sciences, from biology to cosmology to social science, as machine-learning methods have been developed to analyze high-throughput experimental data in novel ways. See Fig. 1 for a depiction of some recent areas of application of machine learning.

A learning problem can be defined as the problem of improving some measure of performance when executing some task, through some type of training experience. For example, in learning to detect credit-card fraud, the task is to assign a label of "fraud" or "not fraud" to any given credit-card transaction. The performance metric to be improved might be the accuracy of this fraud classifier, and the training experience might consist of a collection of historical credit-card transactions, each labeled in retrospect as fraudulent or not. Alternatively, one might define a different performance metric that assigns a higher penalty when "fraud" is labeled "not fraud" than when "not fraud" is incorrectly labeled "fraud." One might also define a different type of training experience—for example, by including unlabeled credit-card transactions along with labeled examples.

A diverse array of machine-learning algorithms has been developed to cover the wide variety of data and problem types exhibited across different machine-learning problems (*1*, *2*). Conceptually, machine-learning algorithms can be viewed as searching through a large space of candidate programs, guided by training experience, to find a program that optimizes the performance metric. Machine-learning algorithms vary greatly, in part by the way in which they represent candidate programs (e.g., decision trees, mathematical functions, and general programming languages) and in part by the way in which they search through this space of programs (e.g., optimization algorithms with well-understood convergence guarantees and evolutionary search methods that evaluate successive generations of randomly mutated programs). Here, we focus on approaches that have been particularly successful to date.

Many algorithms focus on function approximation problems, where the task is embodied in a function (e.g., given an input transaction, output a "fraud" or "not fraud" label), and the learning problem is to improve the accuracy of that function, with experience consisting of a sample of known input-output pairs of the function. In some cases, the function is represented explicitly as a parameterized functional form; in other cases, the function is implicit and obtained via a search process, a factorization, an optimization

[1]Department of Electrical Engineering and Computer Sciences, Department of Statistics, University of California, Berkeley, CA, USA. [2]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA.
***Corresponding author. E-mail: jordan@cs.berkeley.edu (M.I.J.); tom.mitchell@cs.cmu.edu (T.M.M.)**