VIEWPOINT

Big Data and Machine Learning in Health Care

Andrew L. Beam, PhD
Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts.

Isaac S. Kohane, MD, PhD
Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts.

+ Supplemental content

Corresponding Author: Andrew Beam, PhD, Department of Biomedical Informatics, Harvard Medical School, 10 Shattuck St, Boston, MA 02115 (Andrew_Beam@hms.harvard.edu).

Nearly all aspects of modern life are in some way being changed by big data and machine learning. Netflix knows what movies people like to watch and Google knows what people want to know based on their search histories. Indeed, Google has recently begun to replace much of its existing non–machine learning technology with machine learning algorithms, and there is great optimism that these techniques can provide similar improvements across many sectors.

It is no surprise then that medicine is awash with claims of revolution from the application of machine learning to big health care data. Recent examples have demonstrated that big data and machine learning can create algorithms that perform on par with human physicians.[1] Though machine learning and big data may seem mysterious at first, they are in fact deeply related to traditional statistical models that are recognizable to most clinicians. It is our hope that elucidating these connections will demystify these techniques and provide a set of reasonable expectations for the role of machine learning and big data in health care.

Machine learning was originally described as a program that learns to perform a task or make a decision automatically from data, rather than having the behavior explicitly programmed. However, this definition is very broad and could cover nearly any form of data-driven approach. For instance, consider the Framingham cardiovascular risk score, which assigns points to various factors and produces a number that predicts 10-year cardiovascular risk. Should this be considered an example of machine learning? The answer might obviously seem to be no. Closer inspection of the Framingham risk score reveals that the answer might not be as obvious as it first seems. The score was originally created[2] by fitting a proportional hazards model to data from more than 5300 patients, and so the "rule" was in fact learned entirely from data. Designating a risk score as a machine learning algorithm might seem a strange notion, but this example reveals the uncertain nature of the original definition of machine learning.

It is perhaps more useful to imagine an algorithm as existing along a continuum between fully human-guided vs fully machine-guided data analysis. To understand the degree to which a predictive or diagnostic algorithm can said to be an instance of machine learning requires understanding how much of its structure or parameters were predetermined by humans. The trade-off between human specification of a predictive algorithm's properties vs learning those properties from data is what is known as the *machine learning spectrum*. Returning to the Framingham study, to create the original risk score statisticians and clinical experts worked together to make many important decisions, such as which variables to include in the model, the relationship between the dependent and independent variables, and variable transformations and interactions. Since considerable human effort was used to define these properties, it would place low on the machine learning spectrum (#19 in the **Figure** and Supplement). Many evidence-based clinical practices are based on a statistical model of this sort, and so many clinical decisions in fact exist on the machine learning spectrum (middle left of Figure). On the extreme low end of the machine learning spectrum would be heuristics and rules of thumb that do not directly involve the use of any rules or models explicitly derived from data (bottom left of Figure).

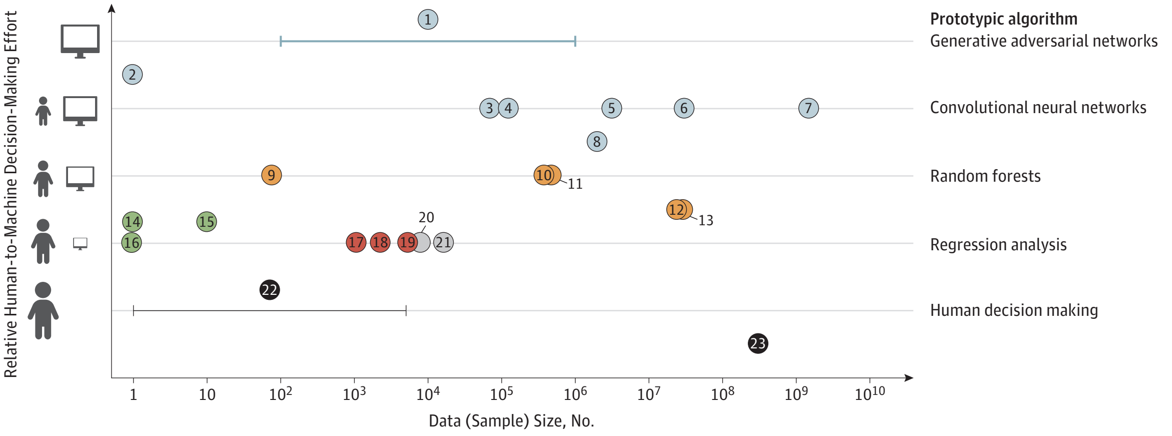
Suppose a new cardiovascular risk score is created that includes possible extensions to the original model. For example, it could be that risk factors should not be added but instead should be multiplied or divided, or perhaps a particularly important risk factor should square the entire score if it is present. Moreover, if it is not known in advance which variables will be important, but thousands of individual measurements have been collected, how should a good model be identified from among the infinite possibilities?

This is precisely what a machine learning algorithm attempts to do. As humans impose fewer assumptions on the algorithm, it moves further up the machine learning spectrum. However, there is never a specific threshold wherein a model suddenly becomes "machine learning"; rather, all of these approaches exist along a continuum, determined by how many human assumptions are placed onto the algorithm.

An example of an approach high on the machine learning spectrum has recently emerged in the form of so-called *deep learning models*. Deep learning models are stunningly complex networks of artificial neurons that were designed expressly to create accurate models directly from raw data. Researchers recently demonstrated a deep learning algorithm capable of detecting diabetic retinopathy (#4 in the Figure, top center) from retinal photographs at a sensitivity equal to or greater than that of ophthalmologists.[1] This model learned the diagnosis procedure directly from the raw pixels of the images with no human intervention outside of a team of ophthalmologists who annotated each image with the correct diagnosis. Because they are able to learn the task with little human instruction or prior assumptions, these deep learning algorithms rank very high on the machine learning spectrum (Figure, light blue circles).

Though they require less human guidance, deep learning algorithms for image recognition require enormous amounts of data to capture the full complexity, variety, and nuance inherent to real-world images. Consequently, these algorithms often require hundreds of thousands of examples to extract the salient image features that are correlated with the outcome of interest. Higher placement on the machine learning spectrum does not imply superiority, because different tasks require different levels of human involvement. While algorithms high on the spectrum are often very flexible and can learn many tasks, they are often uninterpretable

Figure. The Axes of Machine Learning and Big Data



Deep learning

1 Generative adversarial networks (2014)
2 Google AlphaGo Zero (2017)
3 ATM check readers (1998)
4 Google diabetic retinopathy (2016)
5 ImageNet computer vision models (2012-2017)
6 Google AlphaGo (2015)
7 Facebook Photo Tagger (2015)
8 Prediction of 1-y all-cause mortality (2017)

Classic machine learning

9 Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling (2002)
10 EHR-based CV risk prediction (2017)
11 Netflix Prize winner (2006)
12 Google Search (1998)
13 Amazon product recommendation (2003)

Expert AI systems

14 MYCIN (1975)
15 CASNET (1982)
16 DXplain (1986)

Risk calculators

17 CHA_2DS_2-VASc Score for atrial fibrillation stroke risk (2017)
18 MELD end-stage liver disease risk score (2001)
19 Framingham CV risk score (1998)

Randomized Clinical Trials

20 Celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis (2002)
21 Use of estrogen plus progestin in healthy postmenopausal women (2002)

Other

22 Clinical wisdom
23 Mortality rate estimates from US Census (2010)

Traditional clinical studies analyze data from hundreds or thousands of patients using a carefully designed statistical model and thus are low on the machine learning spectrum. Deep learning models are at the top of the spectrum. At the very top are generative adversarial networks, which can learn to generate new images by examining a large database of existing images. See the Supplement for details including supporting references and expansions of abbreviations.

and function mostly as "black boxes." In contrast, algorithms lower on the spectrum often produce outputs that are easier for humans to understand and interpret. Also, the flexibility offered by the high end of the spectrum requires vast amounts of computational resources must be used to develop and deploy these algorithms.

It is precisely because there is access to much larger sources of clinical data and faster computers in the last decade that algorithms on the high end of the machine learning spectrum have become practical and useful. Health care data can come from a diverse set of sources, including the electronic health care record (which includes laboratory results, imaging studies, and diagnosis codes), fitness trackers, genetic testing, among many others.[3] At its core, big data represents an opportunity, and this is especially true for applications in health care. Machine learning is one such tool to integrate and make sense of health care data at this scale.

Machine learning is not a magic device that can spin data into gold, though many news releases would imply that it can. Instead, it is a natural extension to traditional statistical approaches. Machine learning is a valuable and increasingly necessary tool for the modern health care system. Considering the vast amounts of information a physician may need to evaluate[3]—such as the patient's personal history, familial diseases, genomic sequences, medications, activity on social media, admissions to other hospitals—deriving insight to guide clinical decision may be an overwhelming task for any one person. As more control is ceded to algorithms, it is important to note that these new algorithmic decision-making tools come with no guarantees of fairness, equitability, or even veracity. Although we are reluctant to repeat the cliché, even with the best machine learning algorithms the maxim of "garbage in, garbage out" remains true. Whether an algorithm is high or low on the machine learning spectrum, best analytic practices must be used to ensure that the end result is robust and valid. This is especially true in health care because these algorithms have the potential to affect the lives of millions of patients.

ARTICLE INFORMATION

Published Online: March 12, 2018. doi:10.1001/jama.2017.18391

Conflict of Interest Disclosures: The authors completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402-2410.

2. Brand RJ, Rosenman RH, Sholtz RI, et al. Multivariate prediction of coronary heart disease in the Western Collaborative Group Study compared to the findings of the Framingham study. *Circulation*. 1976;53(2):348-355.

3. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. *JAMA*. 2014;311(24):2479-2480.