Surgery 000 (2018) 1–3

Contents lists available at ScienceDirect

Surgery

journal homepage: www.elsevier.com/locate/surg

# Big data: More than big data sets ☆,☆☆

Adrienne N. Cobb, MD [a,b,1], Andrew J. Benjamin, MD [c,1], Erich S. Huang, MD, PhD [d,e,f], Paul C. Kuo, MD [g,*]

[a] Department of Surgery, Loyola University Medical Center, Maywood, IL
[b] One: MAP Surgical Analytics, Department of Surgery, Loyola University Chicago, Maywood, IL
[c] Department of Surgery, University of Chicago Medical Center, Chicago, IL
[d] Institute for Genome Sciences & Policy, Duke University, Durham, NC
[e] Department of Surgery, Duke University School of Medicine, Durham, NC
[f] Sage Bionetworks, 1100 Fairview Avenue North, Seattle, WA
[g] Department of Surgery, University of South Florida, Tampa, FL

ARTICLE INFO

*Article history:*
Accepted 12 June 2018
Available online xxx

ABSTRACT

The term *big data* has been popularized over the past decade and is often used to refer to data sets that are too large or complex to be analyzed by traditional means. Although the term has been utilized for some time in business and engineering, the concept of big data is relatively new to medicine. The reception from the medical community has been mixed; however, the widespread utilization of electronic health records in the United States, the creation of large clinical data sets and national registries that capture information on numerous vectors affecting healthcare delivery and patient outcomes, and the sequencing of the human genome are all opportunities to leverage big data. This review was inspired by a lively panel discussion on big data that took place at the 75th Central Surgical Association Annual Meeting. The authors' aim was to describe big data, the methodologies used to analyze big data, and their practical clinical application.



## What are big data?

The term *big data*, popularized over the past decade, is often used to refer to data sets that are too large or complex to be analyzed by traditional means. It has often been taught that big data refer to any data set that cannot be readily stored or analyzed using spreadsheet programs such as Excel. Opinions on the usefulness and promise of big data vary widely. Some believe big data are the means by which novel insights into rare disease processes can be gained, whereas others believe big data simply add additional noise. Regardless of whether someone feels big data are the future or a nuisance, the data are here to stay. The widespread utilization of electronic health records (EHRs) in the United States, the creation of large clinical data sets and national registries that capture information on numerous vectors affecting healthcare delivery and patient outcomes, and the sequencing of the human genome are all opportunities to leverage these big data to our advantage and that of our patients. In fact, we now live in an era in which data repositories are being generated at an ever increasing pace. It has been estimated that more data have been created in the past 2 years than in the entire history of the human race.

Although collecting such large amounts of data can at times sacrifice data quality, comprehensive data collection often has the potential benefit of mitigating bias when compared with use of high-quality samples of data (the presence of fewer human assumptions in the algorithm also decreases bias). The sheer amount of data requires analytic techniques that can handle not only the volume of information, but also the potential interactions between data that would amplify bias when using traditional statistical methods.

Because large amounts of data are now collected on patients and their surgical outcomes, techniques such as regression and multivariate analysis often fall short in leveraging the advantages that such large data sets potentially offer. Statistical methods were developed in the context of detecting and summarizing relationships from small data sets that are purposefully constructed and structured. Statistical methods are theory driven, are inductive in

☆ Presented as a panel discussion during the Central Surgical Association Annual Meeting held on March 15–17, 2018 in Columbus, Ohio.

☆☆ Erich S. Huang discloses that he is the founder of kēlaHealth, Stratus Medicine, and MedBlue Data (all startups). The remaining authors have no disclosures.

* Corresponding author: 2 Tampa General Circle, Room 7015, Tampa, FL 33606.
*E-mail address:* paulkuo@health.usf.edu (P.C. Kuo).

[1] Adrienne N. Cobb and Andrew J. Benjamin contributed equally to this work.

https://doi.org/10.1016/j.surg.2018.06.022
0039-6060/

nature, and have a confirmatory approach. Alternatively, newer data science methodologies are utilized to discover new patterns and new knowledge in data sets that are realistic, opportunistic, and often messy. These newer methodologies are data driven, are deductive in nature, and have an exploratory approach. Inappropriate use of statistical approaches with "big data" may lead to finding a "signal" in any large enough data set, even if it is just noise. In addition, machine learning algorithms are often better able to make more accurate predictions when used with large data sets.[1]

One of the more promising data science tools available to researchers to make accurate predictions from data is machine learning. Machine learning is a subfield of artificial intelligence focused on constructing algorithms that can learn from and make predictions on data. Although the term *machine learning* was first coined in 1959 by Arthur Samuel, it was not until recently that advances in computing power and accessibility have allowed for widespread utilization of machine learning algorithms as "big data sets" have become more readily available. Machine learning is often thought of as an algorithm that learns to perform a task or make a decision automatically from data as opposed to being explicitly programmed. In reality, however, machine learning and statistics exist along a continuum from fully human-guided data analysis to fully machine-guided data analysis.[2] The fewer human assumptions placed in an algorithm, the higher up it moves on the spectrum of machine learning.

## Supervised and unsupervised learning

Machine learning algorithms can "learn" in two fundamentally different ways: supervised and unsupervised. Supervised machine learning algorithms are trained using examples of a known output or target. The goal is to create a model that is capable of predicting the desired target from a novel data set. Supervised machine learning is often done in the context of classification or regression. Example algorithms include logistic regression, support vector machines, artificial neural networks, and random forests. The goal is to create a model that will take input data and produce correct output data (which are determined from the training data). Alternatively, unsupervised machine learning is used with unlabeled data and is used to find naturally occurring patterns or groupings within the data. Interpreting the results of unsupervised machine learning algorithms is inherently more difficult, and often the utility of findings is determined by performance in subsequent supervised learning tasks.[3]

Another major advantage of machine learning algorithms is the ability of the models to "evolve" over time. As the model is used, it produces feedback data, which, in combination with collection of new data, allow the model to continue refining itself. As long as a sufficient stream of data is available, the predictive capacity of the model will continue to improve and can even adapt to changes in the underlying phenomenon being measured.

Machine learning also has fundamentally changed the types of raw data that can be analyzed. For example, consider a medical image such as that obtained with computed tomography (CT). Previously, we might have used as data points the interpretation of the scan by a radiologist or the size of a lesion; however, with advances in computational power, algorithms such as convolutional neural networks can analyze an image on a pixel-by-pixel basis. These pixels are analyzed and allow the algorithm to identify lung nodules and predict the presence or development of Alzheimer's disease.[4] Given that the International Business Machines researchers estimate that medical images now account for 90% of all imaging data, the promise of machine learning to analyze the raw data contained within medical images will surely lead to advances in the future.

## Examples of machine learning

Because of access to such large data sources and advances in computing power over the last decade, advanced machine learning algorithms have become more practical and useful as tools for analysis and prediction. This advance is key, because traditional statistical analyses are often overwhelmed not only by the sheer volume of data, but also by the inability to deal with nonlinear data. We discuss three machine learning algorithms commonly used to deal with big data: support vector machines, random forest models, and computational neural networks.

Support vector machines (SVMs) constitute a supervised learning method that can be used for both classification and regression. Existing data train the algorithm to then classify new or test data. SVMs perform classification through the development of a multidimensional hyperplane that partitions variables into groups. Both linear and nonlinear data can be used to train the algorithm. There are four main tuning parameters in SVM. The first, "kernel," defines whether we want a linear separation as opposed to a circular line, depending on the amount of transformation needed. The "regularization parameter" tells SVM optimization how much you want to avoid misclassifying each training sample. The "gamma parameter" determines how far the influence of a single training example reaches, with low values being "close" and high values being "far." Last but most important is the "margin," which describes how far each respective class is from the line of separation.[5] The goal of SVMs is to create a maximum-margin hyperplane that lies in a transformed input space and splits the example classes, while maximizing the distance to the nearest cleanly split examples.[6] SVMs are useful in real-life, practical classification problems, such as text categorization and facial recognition,[7] both of which have potential indications in healthcare. With the advent of EHRs, there are an abundance of unstructured data in the form of progress notes, discharge summaries, and other written communications that could potentially be useful in improving healthcare quality. SVMs have tremendous potential to help people better organize electronic resources. The same algorithms utilized for face recognition can be applied to evaluation of modalities such as magnetic resonance imaging.[7]

A decision tree is a model that splits data variables at discrete cut points, which are then often depicted graphically as "branches of a tree." Traditional decisions trees often have subpar predictive ability and are prone to overfitting; however, there are modified decision tree models, such as random forest (RF) models, which provide significantly improved predictive accuracy. RF models are a form of bagged tree model, where multiple trees are combined together, making the final model a collection of many trees. In addition, only random samples of predictor variables are considered at each split of the tree. These features allow RF models to automatically investigate interactions and nonlinear effects of predictors. This approach is starkly different from traditional models such as logistic regression in which such effects must be prespecified. One of the most common criticisms of machine learning is that the algorithms are "black boxes," which often leads to suspicion in the field of medicine. Advantages of RF models, however, are their ability to determine feature importance and their easy-to-visualize outputs with discrete branch points and cutoffs for several variables. In addition, when several models are tested, machine learning algorithms often outperform traditional methods such as logistic regression, with a better C statistic and clear delineation of important variables.

Convolutional neural networks (CNNs) are a deep learning algorithm used most commonly with image data. A CNN consists of a series of "nodes" inspired by the structure of the human visual cortex. In general, an advantage of CNNs is that they require minimal preprocessing, imparting an independence

Please cite this article as: A.N. Cobb et al., Big data: More than big data sets, Surgery (2018), https://doi.org/10.1016/j.surg.2018.06.022

from human input, which can be a substantial advantage. An interesting medical application of this algorithm is the management of pulmonary nodules in screening for lung cancer. Ciompi et al.[8] used CNNs to better manage the large amounts of CT data that are now being produced with the advent of screening for lung cancer in heavy smokers. Using multiscale, multidimensional, convolutional neural networks, they were able process raw CT data without any additional information, such as nodule size and segmentation. The training data then learned a 3-dimensional representation by analyzing an arbitrary number of 2-dimensional views of a given nodule. Ciompi et al. went on to illustrate that the use of CNNs achieved a performance at classifying the type of nodule that surpassed classic models of machine learning and was within the interobserver variability among 4 experienced human observers. Because the algorithm automatically classifies all of the nodules by type that are relevant for continued workup, this approach has the potential to increase the efficiency of diagnosis and treatment of pulmonary nodules.[8] Classification of lung nodules is only one of many promising uses of CNNs in clinical medicine. For example, models have been developed that diagnose breast cancer metastases from digital pathology slides,[9] diagnose diabetic retinopathy,[10] and identify malignant skin lesions.[11] As data continue to be collected, the capability of these algorithms will continue to evolve, further increasing their accuracy and usefulness to physicians.

## What is next?

We have discussed the meaning of big data and some of its applications in healthcare, but where do we go from here? What will contribute to advances in the use of machine learning? Dr. Erich S. Huang of Duke University's Forge Center for Health Data Science discussed with us some of the potential for learning health in an era of data science. He describes the Sage Bionetworks–DREAM Breast Cancer Prognosis Challenge (BCC), which provided a community of data analysts with "a common platform for data access and blinded evaluation of model accuracy in predicting breast cancer survival on the basis of data from gene expression, copy number, and clinical covariates."[12] Because they have shown promise for clinical decision making in breast cancer, these molecular markers can be utilized to distinguish biologically relevant groupings beyond clinical measures and have the potential to inform treatment strategies. The goal of the challenge was to see if any of the data analyst teams could produce a model to predict overall survival that outperformed the current models for breast cancer prognosis using predefined performance criteria, real-time feedback, transparent sharing of source code, and a blinded final validation set. The authors state that their study was not initially designed for direct clinical deployment of a suite of complex biomarkers, but rather to lay the groundwork for future challenges designed to tackle clinically actionable questions. Each team was given a training set on which they built their models, which were subsequently tested and validated on a separate, hold-out data set. This process was completed for several rounds as more than 1400 models were produced by the community of data analysts. They found that the best-performing model significantly outperformed available best-in-class methodologies. Even still, the improvement of the best-performing model (CI 0.76) was moderate with respect to the score achieved by aggregating standard clinical information (CI 0.72). This finding illustrates that the models themselves are not enough. Advances in healthcare will come in data processing and will allow the machine learning algorithms to make better predictions. Dr. Huang pointed out that the most difficult portion of learning health data science is not the execution of the models, but the feature of engineering called "data munging" and the preparation of the data prior to modeling. In addition, the strongest determinant of model performance was not simply the model itself, but the size of the data being utilized. There are several levels at which data can be analyzed, from single-hospital to national data or a single exome to the entire genome. The tools are there, but we must learn how to conduct this research properly to maximize its benefits in a clinical setting.

Big data and machine learning promise to fundamentally change how medicine is practiced. Machine learning algorithms utilizing big data have already proven to be highly effective clinical tools when implemented correctly, but widespread implementation of machine learning algorithms will require that physicians understand the key differentiating aspects between conventional approaches using "small data" and the different approaches needed to use "big data." As physicians become more comfortable with big data approaches, their willingness and desire to collect and process the large amounts of unbiased data will lead necessarily to advances that improve patient outcomes, streamline physician workflow, and uncover novel associations that may go unnoticed with smaller, more biased data sets.

## Acknowledgments

The authors thank the Central Surgical Society for the opportunity to present on this novel topic and for supporting innovation in the surgical community.

## References

1. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. *Critical Care Medicine*. 2016;44:368–374.
2. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319:1317–1318.
3. Deo RC. Machine learning in medicine. *Circulation*. 2015;132:1920–1930.
4. Morra JH, Tu ZW, Apostolova LG, Green AE, Toga AW, Thompson PM. Comparison of AdaBoost and support vector machines for detecting Alzheimer's Disease through automated hippocampal segmentation. *IEEE Trans Med Imaging*. 2010;29:30–43.
5. Patel S. Support vector machine—Theory. In: Machine Learning 101. Available at: https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72 Accessed: May 29, 2018.
6. Shmilovici A. *Support vector machines. In: Data mining and knowledge discovery handbook*. Boston: Springer; 2009:231–247.
7. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. *IEEE Intelligent Syst Their Appl*. 1998;13(4):18–28.
8. Ciompi F, Chung K, Van Riel SJ Setio AA, Gerke PK, Jacobs C, et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Sci Rep*. 2017;7:464–479.
9. Liu Y, Gadepalli K, Norouzi M, Dahl GE, Kohlberger T, Boyko A, et al. Detecting cancer metastases on gigapixel pathology images. In review. Available at: http://arxiv.org/abs/1703.02442v2 [cs.CV] Accessed: June 1, 2018.
10. Gulshan V, Peng L, Coram M, Stumpe MC, Derek Wu D, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402–2410.
11. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115–118.
12. Margolin AA, Bilal E, Huang E, Norman TC, Ottestad L, Mecham BH, et al. Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci Transl Med*. 2013;5(181) 181re1.