ORIGINAL ARTICLE LEADERSHIP

Big Data and Machine Learning—Strategies for Driving This Bus: A Summary of the 2016 Intersociety Summer Conference

SA-CME

Jonathan B. Kruskal, MD, PhD[a], Seth Berkowitz, MD[a], J. Raymond Geis, MD[b], Woojin Kim, MD[c], Paul Nagy, PhD[d], Keith Dreyer, DO, PhD[e]

Abstract

The 38th radiology Intersociety Committee reviewed the current state and future direction of clinical data science and its application to radiology practice. The assembled participants discussed the need to use current technology to better generate and demonstrate radiologists' value for our patients and referring providers. The attendants grappled with the potentially disruptive applications of machine learning to image analysis. Although the prospect of algorithms' interpreting images automatically initially shakes the core of the radiology profession, the group emerged with tremendous optimism about the future of radiology. Emerging technologies will provide enormous opportunities for radiologists to augment and improve the quality of care they provide to their patients. Radiologists must maintain an active role in guiding the development of these technologies. The conference ended with a call to action to develop educational strategies for future leaders, communicate optimism for our profession's future, and engage with industry to ensure the ethics and clinical relevance of developing technologies.

Key Words: Intersociety Committee, ACR, big data, deep learning, machine learning, data science, radiology, imaging informatics

J Am Coll Radiol 2017;■:■-■. Copyright © 2017 American College of Radiology

The ACR's Intersociety Conference meets annually to discuss timely topics of mutual interest to all 53 professional radiology societies, encompassing diagnostic and interventional radiology, radiation oncology, and radiologic physics. The goal of the meeting is to promote collegiality within the broad field of radiology, to foster communication among national societies, and to make recommendations on areas of current or emerging concern to our field [1].

[a]Department of Radiology, Beth Israel Deaconess Medical Center, Boston, Massachusetts.

[b]Advanced Medical Imaging Consultants, Fort Collins, Colorado.

[c]Nuance Communications, Inc. Los Angeles, California.

[d]Department of Radiology, Johns Hopkins Medical Institute, Baltimore, Maryland.

[e]Department of Radiology, Massachusetts General Hospital, Boston, Massachusetts.

Corresponding author and reprints: Jonathan B. Kruskal, MD, PhD, Department of Radiology, Beth Israel Deaconess Medical Center, One Deaconess Road, Boston, MA 02215; e-mail: jkruskal@bidmc.harvard.edu.

Dr Kruskal is an author for Up-To-Date. Dr Kim has received salary from and has shares in Nuance Communications; has received salary from and has shares, ownership, and options in Montage Healthcare Solutions; and has shares in Zebra Medical Vision. All other authors have no conflicts of interest related to the material discussed in this article.

The 38th Intersociety Summer Conference, held August 5 to 7, 2016, in Olympic Valley, California, discussed machine learning (ML) and big data. The assembled group reviewed the current and emerging states of this new "clinical data science" milieu. We speculated how it will affect health care and radiology practice and what are the associated opportunities and challenges for education and research.

Data science already informs radiologists' traditional roles and offers innumerable opportunities and innovations. With the exponential growth in imaging data, radiology knowledge bases, and data from other sources such as electronic health records (EHR) and the Internet of things, how we effectively manage, analyze, and use these data will define radiologists' practice going forward. Radiologists will likely transform from image interpreters to machine-augmented radiology data scientists.

These topics prompt debate and concern about future career opportunities for radiologists in an era of automation. Some even predict radiologists' obsolescence [2]. In contradistinction to this unease, a more likely outcome is that radiologists will assume even larger and

1546-1440/17/$36.00 ■ http://dx.doi.org/10.1016/j.jacr.2017.02.019

more central roles in health care if proper steps are taken today to develop radiology data science and define new practice and business models.

THE CURRENT STATE OF DATA SCIENCE

Data science is an umbrella field encompassing processes and techniques to extract meaning and insight from data. It includes data management, advanced statistics, software, hardware, and business intelligence, all combined with specific domain knowledge. "Big data" is a buzzword for data of such high volume, velocity, variability, and veracity that it requires unique hardware, software, and statistical methods to gain insights from the data [3]. Although radiology imaging data sets are large, at the moment they rarely need true big data techniques.

ML, deep learning, and artificial neural networks represent different flavors of computer algorithms that improve on their own through experience with data; in this article we use "ML" to describe the entire spectrum. These algorithms learn on training data sets. Virtually all current ML algorithms use supervised learning, whereby the algorithm trains on data labeled with the "ground truth" about the data, such as an image of a dog labeled as such. Once these algorithms produce models that fit training data, they can then be used to predict patterns in new data. The task of classifying a dog or cat in an image is the classic example. It is trivial for a human but, until recently, overwhelmingly difficult for a computer. Computers now perform some of these tasks with startling accuracy that outperforms human observers, such as the Large Scale Visual Recognition Challenge [4]. One of the challenge's organizers, Olga Russakovsky, noted in 2015 that the algorithms only had to find images as belonging to 1 of 1,000 categories. Humans can recognize a larger number of categories and, unlike programs, are able to judge the context of an image. Machines have difficulty both with multiple categories and with context, both of which are integral to interpreting radiology examinations [5].

Because radiology images provide large amounts of machine-readable digital data that also fascinate the public, radiology image interpretation appeals to ML specialists. Medical image analysis algorithms are in their infancy but improving, driven by advances in computing power and massive interest from largest corporations through small, high-powered startups. Many entrants to the field are new to health care IT and lack robust, if any, radiology domain knowledge.

DATA SCIENCE APPLICATIONS AND OPPORTUNITIES FOR RADIOLOGY

Radiologists' goal is to produce actions based on new insights from data. This process is a synergy of human and machines. Roentgen's discovery of x-rays first enabled this. Each new technological advancement, from fluoroscopy to CT to the latest complex imaging device, has augmented the relationship between radiologists and the use of technology.

Machine-generated insights will be the foundation for future data-driven radiology advancements. This new paradigm requires sophisticated data management, including the ability to integrate disparate data sources, robust understanding of ML algorithms and their underlying structure and statistics, and better computing power and technology. Radiologists have access to growing quantities of machine-consumable data from EHRs, PACS, and new sources such as the Internet of things, genomics, and social media.

Image Interpretation

The process of creating a useful ML tool can be broken into steps. First, define the problem to be tackled. Second, identify and prepare the data. In contrast to dog and cat images, radiology lacks similarly large data sets. Depending on the clinical question to be answered, the required training data set may range from hundreds of cases to multiple thousands.

Medical images have unique issues that make radiology image interpretation more complex than it may first appear to non–domain-expert data scientists. The cost to develop data sets is very high, the process is fraught with legal issues, and, even in best case scenarios, the data sets will constitute only a fraction of the number of animal photos on Facebook and Instagram. A data set of images must be classified into relevant categories, such as "disease" and "normal." This labor-intensive process depends on humans to classify the data. "Ground truth" is the term for classification accuracy. There are no standardized definitions of ground truth. Radiology reports are not always accurate, nor do they rigorously describe every variation or so-called incidentaloma. Pathologic results may be available for positive cases but not for those declared normal or disease free. Those cases may require clinical follow-up. For a wrist fracture, clinical follow-up time may be only 2 weeks, but how long should it be for lung cancer screening CT? How does one classify a case that, though originally labeled normal, on follow-up at a later date shows disease?