Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Perspective:

Artificial intelligence and statistics*

Bin YU[†‡1,2], Karl KUMBIER[1]

[1]*Department of Statistics, University of California, Berkeley, CA 94720, USA*

[2]*Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720, USA*

[†]E-mail: binyu@stat.berkeley.edu

Received Dec. 7, 2017; Revision accepted Jan. 10, 2018; Crosschecked Jan. 28, 2018

Abstract: Artificial intelligence (AI) is intrinsically data-driven. It calls for the application of statistical concepts through human-machine collaboration during the generation of data, the development of algorithms, and the evaluation of results. This paper discusses how such human-machine collaboration can be approached through the statistical concepts of population, question of interest, representativeness of training data, and scrutiny of results (PQRS). The PQRS workflow provides a conceptual framework for integrating statistical ideas with human input into AI products and researches. These ideas include experimental design principles of randomization and local control as well as the principle of stability to gain reproducibility and interpretability of algorithms and data results. We discuss the use of these principles in the contexts of self-driving cars, automated medical diagnoses, and examples from the authors' collaborative research.

Key words: Artificial intelligence; Statistics; Human-machine collaboration
https://doi.org/10.1631/FITEE.1700813 **CLC number:** TP391; C8

Modern artificial intelligence (AI) can be traced back to work from at least 1943 that highlighted the connection between neural events and propositional logic (McCulloch and Pitts, 1943). Over the years, AI has grown into a transdisciplinary field, integrating and transforming ideas from computer science, statistics/machine learning, psychology, neuroscience, materials science, mechanical engineering, and computer hardware design. Excitement surrounding AI is now exploding. Ideas drawn from the field form the core of both start-ups and academic divisions, and new developments are being reported throughout the media with increased frequency. This excitement is driven in part by the empirical success of AI products that are now available to consumers worldwide. The 'magic' of AI can be captivating, with new AI products like the Amazon Echo responding almost effortlessly with in-depth answers to user queries. However, once one recognizes that these detailed responses barely extend beyond quoted Wikipedia articles, the substantial human input behind the 'magic' of AI is illuminated.

The Echo is a smart speaker that uses a wireless connection to search information over the Internet. This information is created by humans in the form of writing, speech, and music. In other words, the Echo's responses are derived from human-machine collaboration, analyzing manually generated data through algorithms designed and tested by Amazon's researchers (with the help of powerful computing and IT technologies). Similarly, AI products based on computer vision rely on powerful human-machine collaborations through deep

‡ Corresponding author

* Project supported by the Army Research Office (No. W911NF1710005), the National Science Foundation (Nos. DMS-1613002 and IIS 1741340), the Center for Science of Information, a US National Science Foundation Science and Technology Center (No. CCF-0939370), and the National Library of Medicine of the NIH (No. T32LM012417)

ORCID: Bin YU, http://orcid.org/0000-0002-6803-7964

learning algorithms engineered by researchers and manually generated data such as the ImageNet database, which contains roughly 14 million labeled images representing over 1000 categories. Current AI research shares this reliance on human-machine collaboration, in both the data-generation phase and algorithm design and testing. For instance, the authors' research group (Yu Group), in collaboration with the Gallant Neuroscience Lab at UC Berkeley, is combining convolutional neural networks (CNNs) (trained on ImageNet) and regression methods to characterize neurons in primate visual cortex area V4.

Poster children of today's AI applications include self-driving cars and automated medical diagnoses, such as those that identify the cause of a stroke using computed tomography (CT) scans (https://www.newyorker.com/magazine/2017/04/03/ai-versus-md). Both applications rely heavily on computer vision algorithms, which in turn rely on manually generated data. Mr. Tim Bradshaw declared in his 2017 Financial Times article (https://www.ft.com/content/36933cfc-620c-11e7-91a7-502f7ee26895): "Self-driving cars prove to be labor-intensive for humans." He went on to describe that most self-driving car companies hire hundreds or thousands of people to label video footage to teach algorithms to recognize obstacles such as pedestrians. He then quoted Matt Bencke, the founder and chief executive of Mighty AI, saying "AI practitioners, in my mind, have collectively had an arrogant blind spot, which is that computers will solve everything."

Properly framing data collection and analysis is critical for AI products, and can be achieved through human-machine collaboration using the statistical framework of population, question of interest, representativeness of training data, and scrutiny of results (PQRS). The PQRS workflow represents key steps in arriving at data-driven decisions, and is coined by the first author in the process of co-creating and co-teaching a new advanced undergraduate data science course at Berkeley (http://www.ds100.org/sp17/). A population (P) reflects the conditions under which observations are generated. Understanding P helps one recognize randomness in the data generating process, and hence the uncertainty (or errors) in a data result. The question of interest (Q) provides context for an analysis, allowing one to incorporate domain knowledge. Representativeness of training data (R) is closely related to P and assesses whether the available training data provides relevant information on a population (relative to the question asked). The thought process of asking whether the population has changed, or whether the training and test data are similar, addresses P and R simultaneously. Finally, scrutiny (S) describes the process of evaluating data results or algorithm outputs in the context of PQR.

The PQRS workflow provides four concrete steps to think through the cycle of data analysis and algorithm development for data-driven decisions, including those required for self-driving cars and automated medical diagnosis. For instance, answering how dynamic weather, traffic, and construction conditions affect pedestrian recognition can be viewed through the lens of PQRS. Similarly, the relationship between patient characteristics such as age, gender, and previous medical conditions and automated medical diagnoses can be approached using the steps of PQRS. These steps require human input from domain experts who understand a problem's context and from analysts who must obtain data results. It is always the case that such data results will be applied to new individuals or situations. Framing the data collection and analysis can prevent incorrect answers that result from improper context, which can be fatal in the case of self-driving cars and medical diagnoses. PQRS provides effective conceptual devices to integrate human input into these tasks, rescuing the 'magic' of AI from failure and meeting the challenges of dynamic environments head-on.

The final component of the PQRS workflow, S, builds on notions of interpretability to evaluate data results. Interpretability comes in a variety of forms that include, but are not limited to, algorithmic interpretability (i.e., how an algorithm maps features to responses) and domain interpretability (i.e., what a data result says in the context of a particular problem). Human input is critical here as well, because interpretability must be defined with respect to an individual (e.g., expert vs. non-expert). In the area of automated medical diagnosis, and more broadly, human interpretability of algorithms and data results is becoming a necessity. In fact, the EU General Data Protection Regulation (2016) has stipulated the 'right' of users to explanations of algorithms and data results. Thus, automated medical diagnosis algorithms have to be explainable to both doctors and

patients.

Currently, many supervised learning algorithms that are widely used in AI products cannot be well explained. For example, deep learning algorithms are notoriously difficult to interpret even for deep learning researchers, despite the fact that they deliver state-of-the-art prediction performance. To aid interpretability and increase reproducibility of algorithms and data results, Yu (2013) has advocated for the use of the stability principle. This principle is conceptually simple to understand and practically easy to use. It unifies a myriad of works in the literature, starting at least in the 1940s, and provides a platform for developing new stability-based methods. On one hand, it combines the philosophical principle of knowledge stability with the reproducibility principle of science. On the other hand, it connects to statistical inference or uncertainty assessment. Applying the stability principle requires human input to clearly define both appropriate perturbation(s) to data and/or models and stability measure(s). For instance, deep learning algorithms are stable for prediction-based metrics, but not for interpretability metrics that rely on the fitted weights. Appropriateness is a heavily loaded word and should be judged carefully by humans in terms of both the data generation process and domain knowledge.

For algorithm development related to automated medical diagnosis, at least two forms of data perturbations seem appropriate. One is to use a sub-sample of all CT scans from all patients in the training set and study the stability of the algorithm outputs relative to the different sub-samples. The other is to add a small amount of noise to the scans to see how the diagnosis changes. The tolerable level of instability is a domain matter that users must develop in context and in collaboration with subject matter experts such as doctors. It is one measure of uncertainty to take into account when conveying the diagnosis result to a patient.

The stability principle can be applied to interpret supervised learning algorithms whose means of prediction are otherwise impenetrable, making it easier for humans to scrutinize results. For example, the authors' research group incorporates stability into its current genomics work to identify candidate regulatory interactions. Specifically, the group stabilizes random forest decision paths through the iterative random forests (IRF) algorithm (Basu et al., 2018) to recover the high-order, non-linear interactions learned by the popular supervised learning method. The algorithm integrates domain knowledge regarding the thresholding phenomenon of biomolecular interactions (Wolpert, 1969) through the thresholding mechanism of decision trees. IRF empirically demonstrates the value of the stability principle, identifying a high-quality set of stable interactions, of which 80% of the pairwise interactions have been previously reported in fruit-fly genomics experiments. This holds great promise for effectively directing experimental efforts to discover third- or higher-order interactions at the frontiers of systems biology. Note that the scrutiny step in this project required both stable, interpretable interactions and human-generated wet-lab data to evaluate the quality of pairwise results.

Causal effects can also be viewed through the lens of stability as interpretable and stable mechanisms underlying a data generating process. To help doctors decide on drug treatment plans, randomized experiments (or A/B tests) are used to assign patients to treatment and control groups and evaluate the effect of a drug. This brings up the randomization principle of statistical experimental design for effective data collection in causal inference (Imbens and Rubin, 2015). For personal or precision medical diagnosis and treatment, it can be preferable to find a smaller subgroup of patients who are similar to the patient under consideration and carry out the stability analysis for this group. This type of analysis represents an instance of the 'local control' principle of statistical experimental design, which reduces uncertainty or variability induced by conditioning, or grouping, according to features of a patient that are related to the outcome. This is a challenging proposition because it is difficult to find the relevant dimensions by which patients are grouped, even with 'big data', and such groups can be very small with low estimation power. Once again, interpreting algorithmic outputs so they can be scrutinized by subject matter experts, relative to a question of interest, can aid in this decision process.

Data-driven decisions are at the core of AI. These decisions often rely on human input, particularly for cutting-edge AI products such as Amazon Echo, self-driving cars, and automated medical diagnosis. For these particular products, reliance on manual inputs will probably decrease. However, the

demand will be taken up by new AI applications. The PQRS workflow provides one approach for incorporating human input into AI products through simple statistical ideas including experimental design principles (Box et al., 2005) of randomization and local control as well as the stability principle. Integrating these concepts into analyses is useful for efficient and effective collection and use of data and for interpretability and reproducibility of AI algorithms and data results. The authors view it as AI's holy grail to reproduce the unconscious mind, which is yet to be clearly defined for human intelligence, and see an AI future in which humans and statistics continue to play indispensable roles.

Acknowledgements

The authors thank Bryan Liu and Rebecca Barter for their helpful comments.

References

Basu S, Kumbier K, Brown JB, et al., 2018. Iterative random forests to discover predictive and stable high-order interactions. *PNAS*, 115(8):1-6. https://doi.org/10.1073/pnas.1711236115

Box GE, Hunter JS, Hunter WG, 2005. Statistics for Experimenters: Design, Innovation, and Discovery (2nd Ed.). Wiley-Interscience, New York, USA.

Imbens GW, Rubin DB, 2015. Causal Inference for Statistics, Social, and Biomedical Sciences. Cambridge University Press, UK. https://doi.org/10.1017/cbo9781139025751

McCulloch WS, Pitts W, 1943. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys*, 5(4):115-133. https://doi.org/10.1007/BF02478259

Wolpert L, 1969. Positional information and the spatial pattern of cellular differentiation. *J Theor Biol*, 25(1):1-47. https://doi.org/10.1016/s0022-5193(69)80016-0

Yu B, 2013. Stability. *Bernoulli*, 19(4):1484-1500. https://doi.org/10.3150/13-bejsp14

Bin Yu is Chancellor's Professor at the departments of Statistics and of Electrical Engineering & Computer Sciences at the University of California at Berkeley. Her current research interests focus on statistics and machine learning theory, methodologies, and algorithms for solving high-dimensional data problems. Her group is engaged in interdisciplinary research with scientists from genomics, neuroscience, and remote sensing.

She obtained her B.S. degree in mathematics from Peking University in 1984, her MA and PhD degrees in statistics from the University of California at Berkeley in 1987 and 1990, respectively. She was the Chair of Department of Statistics at UC Berkeley from 2009 to 2012, and is a founding co-director of the Microsoft Lab on Statistics and Information Technology at Peking University, China, and the Chair of the Scientific Advisory Committee for the Statistical Science Center at Peking University.

She is a member of the U.S. National Academy of Sciences and Fellow of the American Academy of Arts and Sciences. She was a Guggenheim Fellow in 2006, an invited speaker at ICIAM in 2011, and the Tukey Memorial Lecturer of the Bernoulli Society in 2012. She was the President of the Institute of Mathematical Statistics (IMS) in 2013–2014, and the Rietz Lecturer of the IMS in 2016. She is a Fellow of IMS, ASA, AAAS, and IEEE.

She served on the Board of Mathematics Sciences and Applications of the NAS and as a co-chair of SAMSI Advisory Committee. She is serving on the Board of Trustees at ICERM and Scientific Advisory Board of IPAM. She has served or is serving on numerous editorial boards, including *Journal of Machine Learning Research* (*JMLR*), *Annals of Statistics*, and *Journal of American Statistical Association* (*JASA*).