Harvard Data Science Review • Issue 2.3, Summer 2020

Stability Expanded, in Reality

Bin Yu

Published on: Sep 30, 2020

DOI: 10.1162/99608f92.09889b4c

It is thought-provoking to read the pair of articles on 10 challenges in data science by Xuming He and Xihong Lin from a statistics perspective and Jeannette Wing from a computer science perspective. Unsurprisingly, there is a good overlap of important topics including multimodal and heterogenous data, data privacy, fairness and interpretability, and causal inference or reasoning. This overlap reflects and confirms the foundational and shared roles of statistics and computer science in data science, which is the merging of statistical and computing thinking in the context of solving domain problems. The challenges in both articles are presented as separate, not integrated, topics, and mostly decoupled from domain problems, possibly because of the mandate of "10 challenges."

In my mind, the most exciting 10 challenges in data science are to solve 10 pressing real-world data problems with positive impacts. For example, how is data science going to help control covid-19 spread while allowing a healthy economy? To mitigate climate change so that its negative impact on human and economics can be minimized and in time? To bring precision medicine to every patient safely and timely? To unlock the mysteries of the unconscious brain? To design genomic therapies for Alzheimer's? To design wearables that interact with multiple sclerosis patients to keep them safe? To help discover chip materials for the next generation of computers? To understand the origins of universe? To prevent cyberattacks on democracies all over the world? To self-regulate interactions of digital media with kids? To help people retool skills needed by the rapidly changing economy while allowing them to stay in familiar physical environments of friends, families, mountains, and rivers? Such real-world problems have to be the mission, the anchor, and the goal of data science, while methodologies/algorithms, approaches, and theories have to be at their service and appraised relative to how well they help solve them.

To solve any of these 10 real-world challenges and more, an integrated- and system-framing of data science needs to be embraced. Real-world data science problems are multidisciplinary, multidimensional, and multiphased. Each data science life cycle (DSLC) consists of domain problem formulation, data collection, data cleaning/preprocessing, visualization, analytical problem formulation/modeling, interpretation, evaluation/validation, data conclusions and decisions, and communication of decisions and conclusions. The steps are not at all linear but nonlinear and iterative. The challenges in He and Lin (this issue) fall mostly in the analytical problem-formulation or modeling stage and some on data preprocessing and one on issues in decision making. They do not touch other important steps such as data cleaning, problem formulation, and communication of decisions. Wing (this issue) covers emerging conceptual topics such as trustworthy AI and automating data preparation/preprocessing. Even though I believe some automation in the data cleaning step is necessary, I believe humans have to be in the loop to monitor, check, and make judgment calls in

ambiguous situations flagged by machines. That is, I see a human–machine collaboration future, not automation, for "front-end stages of the data life cycle" (Wing 2020).

The challenges in both articles are important, yet incomplete, components of a data science life cycle or system. Unless the entire system or all the components are integrated and connected together and owned as the traditional topics, there is no insurance that real-world problems such as the 10 challenges above will be solved with positive impacts. In particular, neither article recognizes the many human judgment calls in DSLC or discusses the stability or robustness or reproducibility issues in, say, the choices of data leaning and algorithm in solving a data problem. Data cleaning/preprocessing and coding irreproducibility has led to grave consequences in the past. An article called "Growth in a Time of Debt" was published by economists Carmen Reinhart and Kenneth Rogoff (2010). They concluded that public debt is not good for growth. Such a conclusion was widely used as evidence to argue for austerity policies in Europe and the United States after the 2008 financial crisis. Four years later, Thomas Herndon, Michael Ash, and Robert Pollin (2014) invalidated this conclusion when they included the few data points from New Zealand and corrected the coding errors. (It is not clear why these data points were omitted in the first place.)

When we embrace the data science life cycle as a system, it is clear that the elephant in the room is the human judgment calls made in every step. That is, stability (or robustness) relative to reasonable or appropriate perturbations to the system, including human judgment calls on data-cleaning choices, data perturbation, and model choices, has to be among the core considerations and a key metric for success. This is to makes sure that these perturbations and judgment calls are not driving the data conclusions and decisions, unless justified with well-explained documents. Equally important is to ensure a reality check through prediction into the future (or its good surrogate). Stability is a fundamental and common-sense principle in knowledge seeking and decision making. In fact, when I asked philosopher colleague Branden Fitelson at Northeastern whether considerations of stability of belief/judgment go back to the Greeks, his answer was an affirmative yes and he pointed me to Plato's quotes, here.

In the Meno, Plato writes:

> For true opinions, as long as they remain, are a fine thing and all they do is good, but they are not willing to remain long, and they escape from a man's mind, so that they are not worth much until one ties them down . . . That is why knowledge is prized higher than correct opinion, and knowledge differs from correct opinion in being tied down.

And, in Protagoras, Plato writes:

> [K]nowledge is something noble and able to govern man, and that whoever learns what is good and what is bad will never be swayed by anything to act otherwise than as knowledge bids, and that intelligence is a sufficient succor for mankind.

Fitelson also told me, "Hume was one of the first to emphasize that even (mere) belief needs to be stable (if it is to guide action in the right ways, etc.). Much of the contemporary work has shifted to arguing that even (mere) belief must also be stable in various ways, in order to perform its functions." (For more information on stability of belief, please see Leitgeb, 2017).

In order for a data science life cycle to "perform its functions" and "guide action in the right ways," say, to find a gene therapy for Alzheimer's, the DSLC process has to be stable and capture reality in the data and neuroscience. Predictability (reality check), stability, and computability were argued as the three pillars to support the PCS (predictability, computability, stability) framework for veridical data science (Yu, 2013; Yu & Kumbier, 2020). The PCS framework bridges Breiman's two cultures. It unifies and expands on ideas from machine learning (P and C) and statistics (P and S). Stability in PCS is a significant expansion on the concept of sample-to-sample variability in statistical uncertainty assessment and robust statistics to the entire DSLC including linguistic stability of the same word meaning the same thing for a multidisciplinary team. The PCS framework contains PCSF workflow and PCS documentation on GitHub in R Markdown or Jupyter Notebook to record human judgment calls and choices in the DSLC.

PCS was motivated and developed in the context of multidisciplinary projects in neuroscience and genomics. It has led to the developments of cutting-edge statistical machine learning algorithms ESCV (estimation stability with cross-validation) for Lasso model selection (Lim & Yu, 2015), staNMF for statbility-driven NMF (nonnegative matrix factorization) (Wu et al., 2016), iterative random forests (iRF) for predictive and stable discovery of high-order Boolean interactions (Basu et al., 2018), and DeepTune for visually characterizing V4 neurons (Abbasi-Asl et al., 2018) (corresponding codes can be found at https://www.stat.berkeley.edu/~yugroup/code.html). A recent article of ours (Dwivedi et al., 2020) articulated PCS in the context of causal inference to propose staDISC (stable discovery of subgroups via calibration). It is the first to propose a general model-checking device in causal studies, or calibration as reality checking, as an implementation of P from PCS. Simultaneous to the development of statDISC, we reanalyzed the 1999–2000 VIGOR study, which is an 8,076-patient randomized controlled trial that compared the risk of adverse GI and TC events from a then newly approved drug, rofecoxib (Vioxx), to that from an older drug, naproxen. StaDISC found a subgroup of patients with a prior history of GI events not only has a disproportionately reduced risk of GI events but also does not experience an increased risk of TC events. Building and employing the PCS framework, my group has had very fruitful outcomes in solving multidisciplinary data problems and

developing new general machine learning methodologies. I hope other teams join us in using it in their data science projects and developing if further together.

Finally, I believe a healthy and imperative criterion for designing a new data science algorithm, concept, or framework is to make a serious attempt at solving at least one new data problem as we did in developing algorithms such as iRF. It is a disturbing problem and wasteful of human and computing resources that in statistics, machine learning, or data science, we have way too many new algorithms (and way too many articles) relative to the new data problems that we solve. To solve real-world problems most efficiently from the point of view of society, the reward system in academia needs revamping so that research quality and positive impact are more valued and better incentivized. I believe that, if we willing to improve our reward system, and if we take on the real-world data challenges, embrace reality-check and stability considerations in the entire DSLC, we stand a much higher chance to meet the challenges outlined in the pair of articles by He and Lin, and Wing, respectively.

Disclosure Statement

Partial supports are gratefully acknowledged from ONR grant N00014-17-1-2176, NSF grants DMS-1613002, and IIS 1741340, and the Center for Science of Information (CSoI), a US NSF Science and Technology Center, under grant agreement CCF-0939370.

Acknowledgments

The author would like to thank the editor-in-chief Xiao-Li Meng for helpful comments on a draft of this discussion and Branden Fitelson for the permission to quote him.

References

Abbasi-Asl, R., Chen, Y., Bloniarz, A., Oliver, M., Willmore, B. D. B., Gallant, J. L., & Yu, B. (2018). The DeepTune framework for modeling and characterizing neurons in visual cortex area V4. *bioRxiv*. https://doi.org/10.1101/465534

Basu, S., Kumbier, K., Brown, J. B., & Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences, 115*(8), 1943–1948.

Dwivedi, R., Tan, Y., Park, B., Wei, M., Hogan, K., Madigan, D., & Yu, B. (2020). Stable discovery of interpretable subgroups via calibration in causal studies (staDISC). *arXiv*. https://arxiv.org/abs/2008.10109

Herndon, M., Ash, T., & Pollin, R. (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics, 38*(2), 257–279.

Leitgeb, H. (2017). *The stability of belief: How rational belief coheres with probability*. Oxford University Press.

Lim, C., & Yu, B. (2016). Estimation stability with cross-validation (ESCV). *Journal of Computational and Graphical Statistics, 25*(2), 464–492.

Plato. *Meno.* Translated by Benjamin Jowett. Retrieved from http://classics.mit.edu/Plato/meno.html

Plato. *Protagoras.* Translated by Benjamin Jowett. Retrieved from http://classics.mit.edu/Plato/protagoras.html

Reinhart, C., & Rogoff, K. (2010). Growth in a time of debt. *American Economic Review: Papers and Proceedings, 100*(2), 573–578.

Wu, S., Joseph, A., Hammonds, A., Celniker, S., Yu, B., & Fris, E. (2016). Stability driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. *Proceedings of the National Academy of Sciences, 113*(16), 4290–4295.

Yu, B. (2013). Stability. *Bernoulli, 19*(4), 1484–1500.

Yu, B., & Kumbier, K. (2020). Veridical data science. *Proceedings of the National Academy of Sciences, 117*(8), 3920–3929.
