Journal of Econometrics 237 (2023) 105519

Contents lists available at ScienceDirect

Journal of Econometrics

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

What is uncertainty in today's practice of data science?

Bin Yu

Departments of Statistics and EECS, and Center for Computational Biology at UC Berkeley, 367 Evans Hall # 3860, Berkeley, CA 94720, United States of America

ARTICLE INFO

Article history:
Received 7 August 2023
Received in revised form 7 August 2023
Accepted 13 August 2023
Available online 8 September 2023

AI has made impressive advances in recent years as showcased in generative AI technology such as chatGPT and BARD based on generative large language models (LLMs) that are transformer neural networks. Such LLMs are trained on a vast amount of data, making data science a key component of today's AI technology. Data science has three pillars: computer science, mathematics/statistics, and domain knowledge, with machine learning (including deep learning) belonging to both computer science and mathematics/statistics. To solve a domain problem (e.g. economic policy decision making and new economic knowledge generation), data science leverages the power of algorithms and data, resulting in useful data conclusions when both the algorithms capture good domain or economic knowledge and the data is of good quality and contains relevant information to the domain or economic problem under consideration.

Uncertainty quantification is central to statistics, and a corner-stone for building trust in data conclusions for an economic (or other domain) problem. Traditional statistics formally addresses uncertainty arising from sample-to-sample-variability under a generative stochastic model, which is unfortunately often not model-checked or not checked enough in today's practice of statistics. In a data science life cycle (DSLC) that each data analysis goes through in practice, there are many other important sources of uncertainty. We need realistic and trustworthy uncertainty quantification in the entire DSLC. However, reality-check requirements in a DSLC should go before considerations of uncertainty.

A data science life cycle (DSLC) includes 1. economic (domain) problem formulation in narratives and its translation into a statistical problem (with necessary considerations into why not another statistical problem translation), 2. data collection (or access to public data) and explanations on why this data is useful for solving the economic problem (with human judgment calls made on not using other available data), 3. data cleaning and pre-processing (with human judgment calls made on why certain choices are made and why not others), 4. data exploration analysis (EDA) and visualization (certain choices are made regarding which variables, summaries, plots and colors, for example), 5. model/algorithm development choices (with explanations on why not other ones), 6. post-hoc EDA and visualization (human judgment calls made on why not other choices regarding results to summarize and visualize, and plots and color choices), and communication and reporting in a format appropriate for the intended audience (with human judgment calls on what is "appropriate"). The reality-check considerations need to enter every step of a DSLC. For example, the whole DSLC is useless if the domain problem is ill formulated or does not have solid grounding in domain knowledge, or the data collection process is such poor quality that data does not reflect reality, or the data cleaning process removes the relevance of the data to the population of interest by cleaning away too many data points, or the models or algorithms developed for a balanced binary classification problem has accuracy close to that from a random toss or 50%.

Every step in DSLC is a source of uncertainty due to data collection process or data choice and human judgment calls; that is, different data choices and different human judgment calls are likely to lead to different final data conclusions.

E-mail address: binyu@berkeley.edu.

https://doi.org/10.1016/j.jeconom.2023.105519

Specifically, in our room of uncertainty, there are two big pink elephants or two major sources of uncertainty that we cannot afford to ignore anymore: one is that from the data cleaning/pre-processing step and the other in the modeling step due to human choices on the algorithms/models. In a graduate class project of Statistics 215 A on applied statistics and machine learning at UC Berkeley in Fall 2021, three teams of students were asked to clean a clinical data set independently and each team had a UCSF doctor as a consultant. They ended up with three different cleaned data sets with one team cleaning away 23% of the raw data and getting the best prediction results on their cleaned data set. Mimicking the data cleaning choices made by the three teams we created more cleaned data sets to show that the uncertainty due to the data cleaning choices is similar in magnitude as that from bootstrapping one cleaned data set or that in the traditional statistical uncertainty calculation[1]. To make the point that algorithm choices also give rise to a major uncertainty source, we quote from the abstract of a 2022 PNAS paper by Breznau et al. (2022) that "... Seventy-three independent research teams used identical cross-country survey data to test a prominent social science hypothesis: that more immigration will reduce public support for government provision of social policies. Instead of convergence, teams' results varied greatly, ranging from large negative to large positive effects of immigration on social policy support. The choices made by the research teams in designing their statistical tests explain very little of this variation; a hidden universe of uncertainty remains....".

Working towards trustworthy data conclusions through reality checking and addressing the myriad of uncertainty sources in a DSLC including the aforementioned two, a predictability-computability-stability (PCS) framework and documentation have been introduced in Yu and Kumbier (2022) for veridical data science. PCS unifies, streamlines, and expands on the ideas and best practices of statistics and machine learning. It uses predictability as a stand-in for reality check, uses stability assessing uncertainties in a DSLC from human judgment calls (expanding on traditional uncertainty assessment arising from randomness in a data collection process), and extends computability from computer science considerations (e.g. algorithm scalability and memory space) to include data-inspired simulations. A PCS documentation records as much as possible human reasoning and judgment calls in a DSLC in terms of narratives and codes. A template of PCS documentation can be found at https://yu-group.github.io/vdocs/TCGA-BRCA-Example.html. To ease the use of PCS, a Python software package called veridical-flow (v-flow) can be found at https://github.com/Yu-Group/veridical-flow and an R package simChef https://github.com/Yu-Group/simChef is available to support easy simulations of multiple reasonable models based on real data in order to help assess stability of new algorithms/methods in multiple scenarios that are grounded in real data.

PCS is a conceptual framework (and documentation) to cultivate critical thinking and transparent practice of data science in the entire DSLC in order to solve domain problems in responsible and trustworthy manner. It has been successfully employed to develop new methodologies such as iterative random forests (iRF) (Basu et al., 2018) and stability-driven nonnegative matrix factorization (staNMF) (Wu et al., 2016), to stress-test existing clinical decision rules for internal validity (Kornblith et al., 2022), and to discover calibrated subgroups based on randomized clinical trials (staDISC) (Dwivedi et al., 2020). It is worth noting that the iRF paper (Basu et al., 2018) contains also a biological case study of using PCS to reduce design space or recommend targets for external studies (wet-lab experiments or randomized experiments) to prove causality.

A new textbook based on PCS called "Veridical Data Science: The Practice of Responsible Data Analysis and Decision Making" by Bin Yu and Rebecca Barter has been accepted for publication by the MIT Press (in its series on "Adaptive computing and machine learning"). It covers every step of a DSLC under the guidance of PCS framework and documentation. It is complementary to existing statistical and machine learning books since it fills the missing gaps to connect symbols better with reality. It contains concrete steps to construct PCS prediction intervals with coverage calculated using a test set. This PCS prediction interval formally addresses the two major sources of uncertainty above from data cleaning and algorithm choice. On-going research is conducted to develop PCS perturbation intervals for parameter estimation in linear and other models while a PCS p-value has been proposed in a method called epiTree for epistasis detection in genomics (Behr et al., 2020).

PCS for veridical data science is gaining attention in the statistics and data science community. In particular, it has been extended to veridical spatial data science to include external validity checking and evidence accumulation (Kedron and Bardin, 2021), to veridical network analysis (Ward et al., 2020), and used to design reinforcement learning experiments (Trelia et al., 2022).

A free on-line copy of the textbook by Yu and Barter is expected to be available in fall 2023 while a hard copy is expected to appear in 2024. Please watch out for announcements at https://binyu.stat.berkeley.edu.

References

Basu, S., Kumbier, K., Brown, J.B., Yu, B., 2018. Iterative random forests to discover predictive and stable high-order interactions. Proc. Natl. Acad. Sci. USA 115 (8), 1943–1948. http://dx.doi.org/10.1073/pnas.1711236115.

Behr, M., Kumbier, K., Cordova-Palomera, A., Aguirre, M., Ashley, E., Butte, A.J., Arnaout, R., Brown, B., Priest, J., Yu, B., 2020. Learning epistatic polygenic phenotypes with Boolean interactions. https://www.biorxiv.org/content/10.1101/2020.11.24.396846v1.

Breznau, N., et al., 2022. Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty (Edited by Douglas Massey). Proc. Natl. Acad. Sci. USA 119 (44), e2203150119. http://dx.doi.org/10.1073/pnas.2203150119.

[1] Thanks to the TA Omer Ronen for his help on creating 8 cleaned data sets.

Dwivedi, R., Tan, Y.S., Park, B., Wei, M., Horgan, K., Madigan, D., Yu, B., 2020. Stable discovery of interpretable subgroups via calibration in causal studies. Internat. Statist. Rev. 88 (S1), S135–S178. http://dx.doi.org/10.1111/insr.12427.

Kedron, P., Bardin, S., 2021. A vision for veridical data science. https://escholarship.org/uc/item/1k1566jf.

Kornblith, A.E., Singh, C., Devlin, G., Addo, N., Streck, C.J., Holmes, J.F., Kuppermann, N., Grupp-Phelan, J., Fineman, J., Butte, A.J., Yu, B., 2022. Predictability and stability testing to assess clinical decision instrument performance for children after blunt torso trauma. PLOS Digit. Health http://dx.doi.org/10.1371/journal.pdig.0000076.

Trelia, A.L., Zhang, K.W., Nabum-Shani, I., Shetty, V., Doshi-Velez, F., Murphy, S.A., 2022. Designing reinforcement learning algorithms for digital interventions: Pre-implementation guidelines. https://arxiv.org/abs/2206.03944.

Ward, W., Huang, Z., Davison, A., Zheng, T., 2020. New waves in veridical network embedding. In: Statistical Data Analysis and Data Mining. http://dx.doi.org/10.1002/sam.11486.

Wu, S., Joseph, A., Hammonds, A.S., Celniker, S., Yu, B., Frise, E., 2016. Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. Proc. Natl. Acad. Sci. USA 113 (16), 4290–4295. http://dx.doi.org/10.1073/pnas.1521171113.

Yu, B., Kumbier, K., 2022. Veridical data science. Proc. Natl. Acad. Sci. USA 117 (8), 3920–3929. http://dx.doi.org/10.1073/pnas.1901326117.