

Veridical Data Science

Bin Yu

Departments of Statistics and EECS
University of California, Berkeley
statistics.berkeley.edu/~binyu

ABSTRACT

Veridical data science extracts reliable and reproducible information from data, with an enriched technical language to communicate and evaluate empirical evidence in the context of human decisions and domain knowledge. Building and expanding on principles of statistics, machine learning, and the sciences, we propose the predictability, computability, and stability (PCS) framework for veridical data science. Our framework is comprised of both a workflow and documentation and aims to provide responsible, reliable, reproducible, and transparent results across the entire data science life cycle. Moreover, we propose the PDR desiderata for interpretable machine learning as part of veridical data science (with PDR standing for predictive accuracy, predictive accuracy and relevancy to a human audience and a particular domain problem).

The PCS framework will be illustrated through the development of the DeepTune framework for characterizing V4 neurons. DeepTune builds predictive models using DNNs and ridge regression and applies the stability principle to obtain stable interpretations of 18 predictive models. Finally, a general DNN interpretation method based on contextual decomposition (CD) will be discussed with applications to sentiment analysis and cosmological parameter estimation.

Author Keywords

data science, predictability, computability, stability (PCS), interpretable machine learning (PDR), iterative random forests (iRF), contextual decomposition (CD) for deep neural nets (DNNs)

BIOGRAPHY

Bin Yu is Chancellor's Professor in 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 M.A. and Ph.D. degrees in Statistics from the University of California at Berkeley in 1987 and 1990, respectively. She held faculty positions at the University of Wisconsin-Madison and Yale University and was a Member of Technical Staff at Bell Labs, Lucent. She was 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 Chair of the Scientific Advisory Committee of the Statistical Science Center at Peking University.

She is 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 President of IMS (Institute of Mathematical Statistics) in 2013-2014, and will be the Rietz Lecturer of IMS in 2016. She is a Fellow of IMS, ASA, AAAS and IEEE.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

WSDM'20, February 3-7, 2020, Houston, TX, USA

© 2020 Copyright held by the owner/author.

ACM ISBN 978-1-4503-6822-3/20/02.

DOI: <https://doi.org/10.1145/3336191.3372191>

She served on the Board of Mathematics Sciences and Applications (BMSA) of NAS and as 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).

REFERENCES

- [1] Basu, K. Kumbier, J. B. Brown, and B. Yu (2018) iterative Random Forests to discover predictive and stable high-order interactions PNAS, 115 (8), 1943-1948.
- [2] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu (2019). definitions, methods, and applications in interpretable machine learning.. PNAS, 116 (44) 22071-22080.
- [3] J. Murdoch, P. Liu, and B. Yu (2018). Beyond word importance: contextual decomposition to extract interactions from LSTMs. Proc. ICLR. <https://arxiv.org/abs/1705.07356>
- [4] C. Singh, W. J. Murdoch, and B. Yu (2019) Hierarchical interpretations for neural network predictions. Proc. ICLR. <https://arxiv.org/abs/1806.05337>
- [5] K. Kumbier, S. Sumanta, J. B. Brown, S. Celniker, and B. Yu (2018) Refining interaction search through signed iterative Random Forests. <https://arxiv.org/abs/1810.07287>
- [6] B. Yu and K. Kumbier (2019) Three principles of data science: predictability, computability, and stability (PCS) <https://arxiv.org/abs/1901.08152> Dahshu Journal Club talk slides, Oct. 4, 2019