QnAs with Bin Yu

Farooq Ahmed, *Science Writer*

The explosion of available data in the past decades has birthed a myriad of statistical and machine-learning tools. These tools have allowed scientists from fields as disparate as genomics and cosmology to model and interpret data, draw conclusions, and move science forward. Building on computational advances and increased data availability, data science has emerged as a platform that integrates statistics, computer science, and other disciplines. It has now found commonplace usage, often by those untrained in the underlying statistics, methods, and algorithms. University of California, Berkeley professor Bin Yu trained in statistics but was driven to leverage new computational developments, including machine learning, to solve important scientific problems. The Chancellor's Distinguished Professor and Class of 1936 Second Chair in the departments of statistics and electrical engineering and computer sciences at Berkeley, Yu seeks to formalize the principles of data science while making it more accessible to researchers from other fields. In her Inaugural Article (1), Yu lays out a framework called PCS, which stands for the three principles of data science—predictability, computability, and stability—to guide those who solve domain data problems with data science tools. To do so, Yu leveraged her experience solving data problems across fields such as neuroscience, genomics, remote sensing, and precision medicine. PNAS recently spoke to Yu, who was elected to the National Academy of Sciences in 2014, about her current research.

Bin Yu. Image courtesy of Nan Zhao (photographer).

PNAS: Your Inaugural Article outlines a workflow for using data science to address scientific problems in diverse fields (1). Why did you develop the PCS framework?

Yu: Researchers from all fields have been using data science since it emerged. More and more people are affected by data science, and interpretation of data and the conclusions drawn from these interpretations have become very important.

Building on my work with students, postdocs, and collaborators, the PCS framework brings together some of the principles and best practices of the sciences, whether physical or biological, into data science. At the same time, it embraces the machine-learning platform that is part of modern statistics; predictability and computability are at the center of machine learning. The stability of choices and the decisions made by data scientists at all stages of the data life cycle is a minimum requirement for the interpretability of results as well as a key to their validity.

Moreover, one cannot avoid using natural language in a data science enterprise due to its transdisciplinary nature. Through the PCS framework, we are pushing stability upstream to say that we need linguistic stability even in basic problem formulation, so everyone can understand a particular formulation in the same way.

PNAS: How does the PCS framework work?

Yu: PCS helps researchers think through the entire data science investigation process to extract information

This is a QnAs with a member of the National Academy of Sciences to accompany the member's Inaugural Article, 10.1073/pnas.1901326117.

from data reliably and reproducibly, not just deliver them a number as an answer. The first part, predictability, is conceptual and important since it checks against reality. In the sciences it is, in fact, one important way to falsify a hypothesis.

The second part, computability, outlines how we must use reproducible and scalable computational platforms to collect, store, transmit, clean, manipulate, and compute with data. Moreover, it includes using data-inspired simulations to design algorithms. Finally, stability assesses the choices made in all of the steps of the data science life cycle, including problem formulation, preprocessing, modeling, and even post hoc analysis. Ad-hoc choices made by data scientists should not alter the data conclusions qualitatively.

That's a simplified version of PCS, but taken together it provides a workflow of displaying evidence in a transparent manner to help researchers or users of data results make a rigorous and data-driven decision. The PCS workflow backs up the data evidence with narratives and reproducible codes with documentation so that researchers can trust the process.

I am currently writing a practical data science textbook with my doctoral student, Rebecca Barter, guided by PCS. We hope to have an online version so that anyone can work through data case studies while reading the book.

PNAS: How long have you been developing this framework?

Yu: Since the early 2000s, when I became interested in machine learning with the encouragement of colleagues like the late Leo Breiman. It's a no-brainer now that we need to embrace machine learning in statistics, but things were different then.

After a decade in machine learning, I was invited to give the 2012 Tukey Lecture for the Bernoulli Society. I was trying to make a connection with my data perturbation work for lasso (L1 penalized least squares) and Tukey's robust statistics. I began developing the stability principle relative to data and model/algorithm perturbations.

Over the last several years, I pushed the framework upstream to linguistic stability and data cleaning and created a formal PCS workflow of data-driven decision making. Using PCS in collaborative projects with my group members, as well as with collaborators, has provided me [with] strong evidence for and confidence in the PCS framework.

PNAS: Why is this framework relevant to nonstatistical disciplines?

Yu: Because many, if not all, other disciplines, including genomics, astronomy, precision medicine, political science, and economics—to name just a few—use data science to extract meaningful information from the vast data available.

PCS helps the domain sciences by encouraging researchers to follow the scientific principles of prediction and replication, which manifest themselves as predictability and stability in PCS, so that they can extract reliable knowledge from data. In other words, PCS aims at veridical data science to make the whole data science pipeline as integrated and rigorous as possible by connecting models and algorithms to reality in a transparent manner.

Last but not least, my former doctoral student Karl Kumbier, now at [the University of California, San Francisco], joined in spring 2018 as a coauthor and made indispensable contributions to shaping up the paper.

[1] B. Yu, K. Kumbier, Veridical data science. *Proc. Natl. Acad. Sci. U.S.A.* 10.1073/pnas.1901326117 (2020).