Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Data Come First: Discussion of "Co-citation and Co-authorship Networks of Statisticians"

David Donoho

To cite this article: David Donoho (2022) Data Come First: Discussion of "Co-citation and Co-authorship Networks of Statisticians", Journal of Business & Economic Statistics, 40:2, 491-491, DOI: 10.1080/07350015.2022.2055356

To link to this article: https://doi.org/10.1080/07350015.2022.2055356

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2022, VOL. 40, NO. 2, 491
https://doi.org/10.1080/07350015.2022.2055356

Data Come First: Discussion of "Co-citation and Co-authorship Networks of Statisticians"

David Donoho

Stanford University, Stanford, CA

I salute the authors for their gift to the world of this new dataset! They have clearly invested plenty of time, effort, and IQ points in the study of the statistics literature as a bibliometric laboratory, and our field will grow and develop because of this dataset, as well as methodology the authors developed and/or fine-tuned with those data.

Strikingly, the article also conveys a great deal of enthusiasm for the data! This seems such a departure from the pattern of many articles in statistics today.

The enthusiastic spirit reminds me of some classic work by great figures in the history of statistics, who often were fascinated by new kinds of data which were just becoming available in their day, and who were inspired by the new data to invent fundamental new statistical tools and mathematical machinery. Francis Galton was interested in the relationships between father's height and son's height, himself compiling an extensive bivariate dataset of such heights, leading to the invention of the bivariate normal distribution and the correlation coefficient.

Time and time again, new types of data came first, new types of models and methodology later. Indeed, this seems almost inevitable. As new technologies come onstream, new kinds of measurements become available, and new settings for data analysis and statistical inference emerge. This is plain to see in recent decades, where computational biology produced gene expression data, DNA sequence data, SNP data, and RNA-Seq data, each new data type leading to interesting methodological challenges and scientific progress.

For me, each effort by a statistics researcher to understand a newly available type of data enlarges our field; it should be a primary part of the career of statisticians to cultivate an interest in cultivating new types of datasets, so that new methodology can be discovered and developed.

However, many Ph.D. statisticians, particularly those who are early in their careers, might not agree; they might even have difficulty "getting" the point I'm trying to make. The data-first approach I cited earlier, giving the example of Galton, has not been academically dominant in recent decades. This approach might be considered something like footprints left in the forest by our ancestors: and in this case, those footprints became overgrown and hidden over the last century. It's a wonder they were not lost forever.

Since the 1930s really, the literature of statistics—the very topic of this article (!)—has focused on models and methodology first; in some articles, occasionally data are provided as a kind of afterthought, simply to illustrate the article's methodology concretely—the same way we might "tack on" a bibliography at the end of a article, we "tack on" a data example.

Indeed, this article uncovers a statistics triangle, revealing that our field's literature is clustered around specific types of models and methodology. If the field were instead data-first, its literature might instead be organized around datasets; in which case the authors might have uncovered a very different type of clustering.

Remarkably, the authors discovered this model-first structure of the literature because they were willing to depart from the models-first tradition and work instead within the data-first tradition, developing a fresh high-quality dataset which could support scrutiny and discovery.

The authors could, amazingly, transcend their own modern, theory-driven upbringing and reinvent or relearn the largely forgotten data-first attitudes of an era prior to today's. But doing so has paid off! Kudos to the authors for showing us the way, I hope that others can be inspired and that the literature of statistics will grow because of this shining example of the fruits of a data-first orientation.

CONTACT David Donoho *donoho@stanford.edu* Stanford University, Stanford, CA.