Is “Official” Statistics Ready to Embrace the (Still) “Unofficial” World of Data Science?

Francesca Dominici^{1,2}

¹Harvard Data Science Initiative, Harvard University, Cambridge, Massachusetts, United States of America,

²Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, Cambridge, Massachusetts, United States of America

Published on: Oct 28, 2021

DOI: <https://doi.org/10.1162/99608f92.52e5fd67>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Editor’s Column

Hi! This is Francesca Dominici, one of the two Interim Editors-in-Chief of *HDSR* (together with David Parkes). As you might know, Xiao-Li is taking a well-deserved sabbatical. So, I am left with the unsurmountable task of introducing this new issue. I have big shoes to fill!

Is ‘Official’ Statistics Ready to Embrace The (Still) ‘Unofficial’ World of Data Science?

This issue includes a [Special Theme: Official Statistics From the Changing World of Data Science](#). Many of the papers in this special theme attempt to address the following question: *Is ‘official’ statistics ready to embrace the (still) ‘unofficial’ world of data science?* [This series of articles is summarized](#) and discussed by the two co-editors, John Bailer and Wendy Martinez.

What do we mean by official statistics? Official statistics result from the collection and processing of data into statistical information by a government institution or international organization. Official statistics make information on economic and social development accessible to the public, allowing the impact of government policies to be assessed, thus improving accountability. The papers of the special theme will attempt to address questions such as: What significant changes has the data science revolution brought, if any, to official statistics? What are the key challenges and opportunities that you see for official statistics in the digital age, especially with respect to data quality, including data collections and data privacy? Enjoy!

How Do We Decide What Should We Measure and What Not? Is Paper Count a Good Measure to Promote a Faculty Member?

In addition to the special theme, we have a [Panorama article](#) where the author Fisher provides a comprehensive overview of the issues, approaches, and opportunities in performance measurement. The author wrote, “Performance measures permeate our lives, whether or not we are aware of them. They can support or frustrate what we are trying to do, help or hinder enterprises going about their business, encourage or distort behaviors, clarify or confuse purposes.” One of the examples comes from the academic world. Ah! I am sure that I now have your attention. In the academic world, the most used metrics to evaluate the quality of the research and compare individuals for promotion cases is based on (a) paper count, b) citation count, and c) journal impact factors. Well, these metrics purport to provide some evidence of ‘quality’—of the individual or of the journal. The author raises critically important questions: 1) Who is the customer here? 2) Who are the people or groups with a vested interest in how well this work is done, and what does ‘quality’ mean to them? I hope you will find this research as enticing and provocative as I have!

Should Computer Algorithms Replace Radiologists? No...No Yet.

The rapid development of deep learning, a family of machine learning techniques, has spurred much interest in its application to medical imaging problems. In the **Cornucopia** section, [we present a paper](#) that relies on DeepMiner to discover interpretable representations for mammogram classification and explanations. The authors present a method that can generate mammogram classifications that are consistent with ground truth radiology reports. They show that their methodology not only enables better understanding of the nuances of classification decisions but also possibly discovers new visual knowledge relevant to medical diagnosis.

Data Science Education: Crowdsourcing in a Collaborative Setting

In **Stepping Stones**, the authors of [“Collaboratory at Columbia: An Aspen Grove of Data Science Education”](#) present a crowd-sourcing approach to creating new data science pedagogy. The authors of this paper advocate for offering seed funding to foster proactive efforts to embed data science ‘in context’ into more traditional domains. Importantly, it embeds ethics by situating data science in context. The authors wrote “Like an aspen grove, the Collaboratory has fostered a diverse and vibrant set of context specific pedagogy, a community of instructors that nurtures transdisciplinary research, and a groundswell of student interest able to benefit from the ‘canopy’ of Collaboratory offerings. It has also enriched the ‘soil’ of the university, building appreciation among university leadership for the contributions data science can make across the curriculum.” I am grateful to the authors for sharing these important insights as many academic institutions are trying to innovate in data science education. Also [see the video!](#)

Dynamic Data Citation

Last but not least, the acquisition and analysis of data related to the COVID-19 pandemic has taught us—again—that real-world data evolves all the time. Data changes over time and measurements are often corrected and linked to additional data sources. Yet, although this is happening often with data-driven science, the scientific community still lacks concrete guidance for reproducibility and for referencing dynamically evolving data sets. In the [Milestones and Millstones article](#), the authors review, challenge, and discuss several recommendations that are centered around time-stamping and versioning evolving data sources. This is an important piece that I will be sharing with my lab, as I believe any responsible data-centered study should follow these principles.

Finally, do not miss our columns!

Disclosure Statement

Francesca Dominici has no financial or non-financial disclosures to share for this editorial.

©2021 Francesca Dominici. This editorial is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the editorial.