President's Column: Practice what we preach. Work for what we wish for.

Xiao-Li Meng writes his final President's Column, before handing on the gavel to the next IMS President, Susan Murphy, at JSM…

Photo: Martha Stewart

It has been so long since I was quarantined by the joy of learning as a student, a form of joy whose purity many of us only recognize decades after we lost our innocence. I was therefore in debt to the organizers of the 2019 IEEE Data Science Workshop (DSW: https://2019.ieeedatascience.org/). They provided me the opportunity to experience that joy again; on a breezy, refreshing Sunday I limbered up with "Large-scale Optimization for Machine Learning" in the morning and tangoed with "Tensors in Data Science" in the afternoon.

However, a real "aha" moment came during the welcome reception that evening. A dean and an ex-president of IEEE's Signal Processing Society (SPS) delivered welcoming remarks, and reminded the mixed audience of engineers, applied mathematicians, computer scientists, and statisticians that the mission of SPS has long been about "generation, transformation, extraction, and interpretation of information." Isn't that pretty much what Data Science (DS) is about? After all, who would care much about data if they don't ultimately lead to actionable or at least understandable information?

I share much of my fellow statisticians' and probabilists' frustration that our consistent and substantial contributions to DS have generally not been properly recognized. But this remark reminded me that we are still the luckier ones. What is the percentage of the Venn diagrams on data science you can find online that include "signal processing" either as a participating discipline or a skill set? So far that percentage from my search is smaller than the probability that my mother country would win the 2026 world cup. The OR (Operations Research) community is in a similar situation; its contributions to optimization methods, which are the bread and butter of machine learning, are essentially infinite compared to the attention the community has received in the media frenzy over DS or AI.

No matter how frustrated, or even outraged, any individual group or discipline in DS is, there is no DS deity we can blame for unfairly favoring some groups over others. If anything is to blame, it is our long and collective failure to communicate with and learn from each other. Period.

The good news is that this period is about to end. There is an increasing awareness that it is much more effective to engage in outreach than in outrage, so to speak. That computer scientists and statisticians were invited to IEEE DSW represents the SPS's effort. That the ACM and IMS reached out to each other last year is another such indication. As I wrote in my second President's column, this outreach resulted in the establishment of the IMS task force, co-chaired by Liza Levina (Michigan) and David Madigan (Columbia), on the partnership with ACM, the world's largest computing society with nearly 100,000 members. I am very happy to report that this effort is now expanding to a much larger-scale collaboration by multiple disciplines, as encouraged by NAS (National Academies of Sciences, Engineering, and Medicines), and with ACM and IMS as its co-leading organizations.

Specifically, the first ACM-IMS Interdisciplinary Summit on the Foundations of Data Science was held on June 15, 2019 in the grandiose Palace Hotel of San Francisco, just prior to the ACM award ceremony, which conferred the latest Turing Award to the "Fathers of the deep learning revolution." The Summit co-chair, Columbia computer scientist Jeannette Wing, concluded her opening remarks [*which you can watch on the Livestream at* https://www.acm.org/data-science-summit/livestream—*see screenshot below*] by emphasizing that,

"*While today's event focuses primarily on computer science and statistics, I want to acknowledge that the foundations of data science also draw on other fields—for example, signal processing from Electronic*

Continues on **page 10**

Continued from **page 9**

Engineering, optimization from Operation Research, analysis from Applied Mathematics, and more. David and I expect that the future events in the foundations of data science will reach out to these fields."

A panel on Robustness and Stability in Data Science at the ACM-IMS summit. L–R: moderator Ryan Tibshirani, panelists Xiao-Li Meng, Bin Yu, Richard J. Samworth, Aleksander Madry

The joint leadership of ACM and IMS in reaching out to many disciplines is a *Big Deal.* I am deeply grateful to the ACM leadership team, especially its Executive Director and CEO, Vicki Hanson, to the Summit co-chairs, Jeannette Wing and David Madigan, and to all the members of the Steering Committee, which include IMS representatives Chris Holmes (Oxford), Ryan Tibshirani (CMU), and Daniela Witten (UW), for having formally kicked off this joint effort in less than eight months. The first joint Summit was a great success by almost all measures. And it is about this "almost" qualification that I am writing to ask for your help, urgently.

As you will see from the program of the Summit [*on the next page*], it was an extremely well-crafted program in terms of coverage of the topics and representatives of the presenters. Indeed, the six-hour program packed with keynotes and panel debates was very inspiring and intense, so much so that one panel trigged the fire alarm—you can search for the recording to see how long we had to leave the auditorium. However, while the auditorium was packed with about 250 participants, the size of the IMS registered audience was smaller than the number of statisticians on the program.

I realize that the membership ratio of ACM to IMS is about 25:1, and hence the ratio at the Summit was not completely out of proportion. Nevertheless, if IMS truly wants to be a leading voice in DS, we have to move our collective feet to where our mouths say we want to be. We cannot keep complaining that we don't have a seat at the table but not show up in numbers when we are invited or, worse, when we're the co-hosts. The matter is very simple. If we don't take these seats reserved for us, many others will. And few would keep reserving seats for those who don't show up, no matter how important they are.

Of course, the IMS leadership needs to be more creative in finding ways to encourage members to attend such outreach events. For that, I am particularly grateful to David Madigan, together with Jeannette Wing, for leading the effort to secure an NSF (US National Science Foundation) grant which sponsored over 35 students and young researchers' attendance at the Summit. It is telling is that all of these funds were taken within 24 hours of the award announcement, almost surely by CS students and young researchers.

This last observation makes me particularly appreciate a new emphasis by another IMS task force, co-chaired by Joseph Blitzstein (Harvard) and Deborah Nolan (Berkeley), which was inspired by Jon Wellner's 2017 Presidential Address, *Teaching Statistics in the age of data science*. Its general task is as hard to accomplish as it is easy to state: **to determine what the PhD curriculum for statistics should be, in the age of data science**. The task force is charged with complimenting the work done at the NSF's 2018 "Statistics at a Crossroads" workshops, one of which focused on PhD education. The complementary roles IMS can play are in (at least) two dimensions: going beyond the United States, and going deeper into probability. Its membership therefore reflects these dimensions: David Aldous (Berkeley), Emmanuel Candès (Stanford), Antonietta Mira (Università della Svizzera Italiana), Guy Nason (Bristol), Richard Samworth (Cambridge), Nike Sun (MIT), Qi-Man Shao (Southern University of Science and Technology of China), and Harrison Zhou (Yale). I am extremely grateful to this most prominent task force, which has been working diligently via monthly conference calls: no small feat considering the wide range of the time zones! (I will leave this as a trivia question: what is the optimal call time the task force identified?)

The task force is working on a report that consists of four major parts:

International Training: Compare and contrast the programs in different countries, using various metrics, such as median length of program, number of required courses, topic breadth in required courses, and the depth of professional development.

Resources: Create, curate, and share course materials on emerging topics that are not easy to find a textbook-style reference, and work out how to incentivize such efforts.

Leadership: Develop more PhD students into outstanding communicators and ambassadors for the importance of statistics and statistical thinking, in an era where the general public often hears about AI and ML but may have little understanding the critical roles statistics plays, or even what it is.

Probability: Update the probability curriculum to better reflect the statistical and data scientific challenges students are

Continues on **page 11**

starting to encounter, addressing the old debate on how much measure theory to include in the core probability course, and recent questions about the roles of CS and DS in the probability curriculum.

I am particularly grateful for and pleased to see the task force's emphasis on building leadership while one is still a student. It is not a secret that for too long "leadership" has not been viewed as an essential skill, and in some faculty members' minds it was (and perhaps still is) even a distraction, subtracting from one's scholarship. The end result is that our profession simply does not have enough "outstanding communicators and ambassadors" out there to explain —and promote the importance of—what we do. Promotion is not a dirty word as long as we have substance to be promoted, and we absolutely do. The lack of general leadership training in statistics is hurting us in real terms, including in our pockets. At the latest NAS Committee on Applied and Theoretical Statistics (CATS) meeting I attended, representatives from NSF reminded the committee once again of a painful reality: the suggestions regarding what kinds of DS research the NSF should fund come almost exclusively from outside of the statistical community.

This was why I invited Juan Meza, the Director of the Division of Mathematical Sciences at NSF, to write to us directly last November [http://bulletin.imstat.org/2018/11/seeking-novelty-in-data-sciences/]. Meza told us about the **Harnessing the Data Revolution** initiative and asserted that, as DS evolves, "new strategies, methods, and theory will be needed to address all of the complex data issues arising." He concluded with a call to action for statisticians and probabilists: "And who better to do this than those who have already contributed so much to data sciences?" But apparently such messages need to be repeated periodically, as we are simply a shy profession, especially compared to CS which has a much faster-paced and action-oriented culture.

Regardless of whether or not we feel our fellow disciplines are moving too aggressively, no one can hear us if all we do is to complain to each other that others don't hear us. If we want to be a leading voice in the DS era, we must go out, communicate with other disciplines, speak to funding agencies, talk to the general public, etc. That is, we must work for what we wish for, just as we should always practice what we preach.

This is my departing wish as the IMS President. I look forward to thanking you in person for your trust in me when I see you at an ACM symposium or an AMS meeting or an IEEE workshop or an INFORMS conference.

Until then, please consider giving one presentation to your favorite high school. Thank you!

ACM–IMS Interdisciplinary Summit on the Foundations of Data Science June 15, 2019, San Francisco

Program

9:00-9:05 AM – Introduction, **Jeannette Wing**, Columbia University

9:05-9:40 AM – Keynote Talk: *"Making the Black Box Effective: What Statistics Can Offer,"* **Emmanuel Candès**, Stanford University, with introduction by **David Madigan**, Columbia University

9:40-10:20 AM – Panel: *Deep Learning, Reinforcement Learning, and Role of Methods in Data Science*. Moderator: **Joseph Gonzalez**, University of California Berkeley. Panelists: **Shirley Ho**, Flatiron Institute, **Sham Kakade**, University of Washington, **Suchi Saria**, Johns Hopkins University, **Manuela Veloso**, J.P. Morgan AI Research, Carnegie Mellon University

10:20-10:35 AM – Break

10:35-11:15 AM – Panel: *Robustness and Stability in Data Science.* Moderator: **Ryan Tibshirani**, Carnegie Mellon University. Panelists: **Aleksander Madry**, Massachusetts Institute of Technology, **Xiao-Li Meng**, Harvard University, **Richard J. Samworth**, University of Cambridge, The Alan Turing Institute, **Bin Yu**, University of California, Berkeley

11:15-11:55 AM – Panel: *Fairness and Ethics in Data Science.* Moderator: **Yannis Ioannidis**, National and Kapodistrian University of Athens. Panelists: **Joaquin Quiñonero Candela**, Facebook, **Alexandra Chouldechova**, Carnegie Mellon University, **Andrew Gelman**, Columbia University, **Kristian Lum**, Human Rights Data Analysis Group (HRDAG)

11:55 AM-1:00 PM – Lunch

1:00-1:35 PM – Keynote Talk: *"Deep Learning for Tackling Real-World Problems,"* **Jeffrey Dean**, Google, with introduction by **Suchi Saria**, Johns Hopkins University

1:35-2:10 PM – Keynote Talk: *"Machine Learning: A New Approach to Drug Discovery,"* **Daphne Koller**, insitro, with introduction by **Kristian Lum**, Human Rights Data Analysis Group

2:10-2:20 PM – Break

2:20-2:55 PM – Panel: *Future of Data Science.* Moderator: **David Madigan**, Columbia University. Panelists: **Michael I. Jordan**, University of California, Berkeley, **Jeannette Wing**, Columbia University

2:55-3:00 PM – *Closing Remarks:* **David Madigan** and **Jeannette Wing**, Columbia University