The Capstone in Everyone's Delivery Room: Placing 'Practice' at the Center of Data Science Education

Antonietta Mira^{1,2} Ernst Wit^{3,4}

¹Antonietta Mira: Institute of Computational Science, Università della Svizzera italiana, Lugano, Switzerland,

²Department of Science and High Technology, University of Insubria, Como, Italy,

³Statistics and Probability Unit, University of Groningen, Bernoulli Institute, Groningen, the Netherlands,

⁴Institute of Computational Science, Università della Svizzera italiana, Lugano, Switzerland

Published on: Feb 25, 2021

DOI: <https://doi.org/10.1162/99608f92.539432b5>

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

Statistics is one of the few scientific disciplines without its own subject field. Together with probably only philosophy, its subject matter can be anything and it shares with that discipline the fact that its purpose is fundamentally an enabling one. Just as Socrates claimed to be only the midwife of truth, so must statistics prepare its practitioners to be able to assist pregnant data to deliver new information.

Just as a midwife is not a doctor or a surgeon, so a statistician is not a mathematician or a computer scientist but needs skills from either. In fact, a statistician needs a whole array of skills, from mathematical insights and computational prowess all the way to adequate problem analysis and efficient communication. This is basically the argument that the authors of this article (Kolaczyk et al., this issue) make in describing how putting Statistics Practicum at the center of a one-year M.S. in Statistical Practice (MSSP) helps to instill this broad skillset.

With an extremely forward-looking attitude, Boston University granted, five years ago, sufficient resources to design this master's program starting *tabula rasa*: a rare privilege, often hard to replicate due to limited financial resources. With little surprise, the resulting MSSP has a practicum as its capstone. What are the successes and the challenges that this practice poses? To answer this question, one has to first define the KPIs (key performance indicators) that the authors identify in the following goal: train “students well-positioned to have immediate impact upon employment” or, in other words, “transform individuals from students focused largely on courses to already experienced data science professionals in as little as 9 months.”

As in architecture, the capstone is the final piece placed during construction, in the middle of the arch, locking all the stones into position, allowing the arch to bear weight, likewise the practicum stands, in the study program, at the center of two electives and four core courses (in computing, methods and modeling, and statistical theory) and its effectiveness relies on careful timing and extreme coordination of two conditions that are not easy to achieve. Although an arch or a vault cannot be self-supporting until the capstone is placed, the capstone experiences the least stress of any of the *voussoirs*, due to its position at the apex. Likewise, if the core courses and the electives are well integrated and properly interact with the practicum, the latter should shine as a natural achievement of what comes before, on one hand, and a natural premise, on the other, of what comes after it in the student learning experience: it should ultimately be self-supporting.

In 2009, Xiao-Li Meng wrote, “We no longer simply enjoy the privilege of playing in or cleaning up everyone's backyard (referring to the famous quote attributed to John Tukey, ‘The best thing about being a statistician is that you get to play in everyone's backyard.’). We are now being invited into *everyone's study or living room* and trusted with the task of being their offspring's first quantitative nanny.” In his first XL File, Professor Meng (2013, p. 7) promoted us to *bedroom guests* to reflect “how seriously we statisticians now are involved in substantive investigations.” As of today, the claim is that statisticians have entered *everyone's delivery room* for all intents and purposes. We have gained even greater intimacy: Scientists in pretty much all disciplines need to work side by side with data scientists that, as mentioned at the beginning, assist pregnant

data to deliver new information. This poses a great deal of responsibility on us, and we need to properly prepare our students to be able to effectively bear such responsibility.

Although we have a lot of sympathy with the idea of making a practical statistics course the core of an educational program in statistics, the main issue with such a statement of needing a broad skillset is that it is a mere platitude. Yes, by doing everything we might learn a bit about everything and perhaps 'within 9 months' students are transformed 'into experienced data science professionals.' The more interesting question is what the three crucial components in this transformation are.

- **Problem clarity:** As the authors rightly state, a clear problem formulation is a crucial component in any statistical enterprise. This is one place where the beginning data scientist often takes a shortcut by diving into the data without any clear question and therefore no clear aim or purpose. Posing the right question is a valuable skill every data scientist should master. It is a skill hard to teach and pass on even in a practicum because it is acquired with experience and requires a multidisciplinary and curious attitude. Moreover, these days the traditional paradigm of designing a study by advancing a hypothesis and subsequently collecting data able to verify or falsify it, is flanked by the alternative approach of identifying the right question that available data can help in answering. This new data-driven research paradigm is motivated by the fact that data lakes exist, simply because data is easily produced and fairly cheaply stored. Furthermore, the value of data is not reduced if queried with many questions—on the contrary, it may possibly be increased by merging different data sets that jointly can motivate and drive new research lines.
- **Experimental detective:** Just as data do not constitute a question (see the first point), they also do not constitute an experiment or study. Data are like fishes out of water. The point *really* is *how* they got there and *why* they got there. This requires careful detective work. It is important to experience how the same data within different study designs could lead to very different answers.
- **Adversarial reasoning:** A statistician is a professional skeptic, always imagining how she might be fooled by what is staring her in the face. This takes a number of guises.
 - **All data are wrong:** Data, whether they come from the medieval mortality tables or from high-throughput technology, often are not quite what they seem to be. Before taking data at face value, can we really be certain that we know what they mean and whether they contain no obvious errors? In a careful exploratory data analysis, often months of pointless complex data analyses can be avoided. By making one familiar with the nature of the data, one can move on to the inferential stage with more confidence and understanding.
 - **All models are wrong:** Modeling is a bridge between the question of interest and the data trickling out of a faraway place called Reality. By making sure that the bridge is constructed carefully, the data can reach the question of interest and begin to shed light on it.
 - **Don't be fooled by randomness:** Even the most seasoned statistician can get fooled by randomness. Even the best estimate will be off, and the best prediction will never get rid of randomness. Even if my

perfect prediction of a certain count is 10 cases, I should not be surprised to actually see 4 or 17 cases. No data science magic can change that.

- **Why all my conclusions might be wrong:** If a student can imagine the circumstances under which the conclusion from the statistical analyses are in fact wrong, then she can evaluate the confidence one might reasonably have in the conclusions. A conclusion that holds on a narrow set of assumptions might be considered less trustworthy than a conclusion that is based on a much broader set of assumptions.

Now all these aspects are clearly present in the Statistics Practicum that takes center stage in the MSSP program. One might only look for an example at Figure 2 describing the topics of the first semester of the practical course and see many similar themes. However, it is not merely that a data scientist needs to be broadly skilled or a great communicator. He or she needs specific skills, some of which—such as adversarial reasoning—do not always carry universal approval with clients or the wider public. Nevertheless, data science should preserve its integrity, otherwise its role as the midwife of truth comes under serious strain.

Disclosure Statement

Antonietta Mira and Ernst Wit have no financial or non-financial disclosures to share for this article.

References

Meng, X-L. (2009). Desired and feared—What do we do now and over the next 50 years? *The American Statistician*, 63(3), 202–210. <https://doi.org/10.1198/tast.2009.09045>

Meng, X-L. (2013). The XL files: Statisticians' impact, from backyard to bedroom? *IMS Bulletin*, 42(1), 7. https://imstat.org/wp-content/uploads/Bulletin42_1.pdf

©2021 Antonietta Mira and Ernst Wit. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.