The Hiring Gambit: In Search of the Twofer Data Scientist

Radu V. Craiu

The setup

2013 was a tough year for this statistician. I recently dug out a piece I wrote back then, and it started like this:

> The Data Science revolution—that sweet promise of groundbreaking truths revealed from massive information—is an elusive one. As I keep looking for Big Data, people keep telling me that they *are everywhere around us*. This does not help my self-esteem. And when I finally start to get the big picture, I realize that I am already out of it. Statisticians out, data scientists in. I understand that my skills are good, but also that part of my training is holding me back. I know statistics, but somehow I have too much theory in me and not enough 'just do it.' All of a sudden, I am a Franken-data scientist.

Looking back, I no longer think that I was the only one at a crossroads. Entire departments changed names, merged or split, started new programs and dissolved old ones. In the effort to catch up to the data science train, statistics has gained a lot of traction. At the center of the changes implemented by most departments, my own included, was the difficult task of defining what data scientists are, and to recruit as many of them as possible. The latter motivated this discussion on some of the challenges involved in hiring data scientists into statistics departments.

The Challenges

The attempt to define a data scientist brings us to our first challenge. Even after so many of us have firmly embraced the field of data science, I am still not sure that we have a complete answer to what a data scientist does. This unsureness isn't necessarily the weakness many consider it. After all, who can *really say* what makes someone a poet or a scientist? Even so, some things can be said. In broad strokes, we are looking for someone with a data driven research agenda, who adheres to or aspires to using a principled implementation of statistical methods and uses efficient computation skills.

Facing these realities, departments have developed various strategies. Under the leadership of our former chair, Jamie Stafford, our department has taken a 'follow the data' approach and initiated a sustained campaign of joint hires, many of which could be said to target the two-for-one, or 'twofer,' data scientist. As we all know, hiring has to do with vision, usually for the future of one's department, but from time to turbulent time, the future of the discipline is at stake and must be considered, too.

In our department's collective vision, the future involves generating new methods motivated by cross-disciplinary collaborations between scientists and statisticians. These positions are usually initiated as collaborations between our department and one with a rich, data-driven research program. Finding

good people working at the intersection of statistics and a domain-specific field may sound trivial, but it is more complicated than it sounds. For instance, ask any committed statistical geneticist or astrostatistician how many hours they have spent in lab meetings encountering new problems while simultaneously absorbing a different scientific culture and learning new jargon. Given the enormous commitment, a data scientist taking this path can only do so knowing that the department has their back. This means that tenure-worthy work can be published in subject-matter journals, that papers filled with theorems are not a necessary condition for peer respect, and that students in the department are encouraged to take on research projects that are data-driven. Any data scientist that we pursue asks about the ways the department encourages and supports interdisciplinary research.

The committee will look for someone who gets their hands dirty with the kind of messy, unfiltered, and possibly unclean data - tainted by heteroskedasticity, complex dependence and missingness patterns - that until recently were avoided in polite conversations between more traditional statisticians, but will also not neglect theoreticians working on problems relevant to practice. Not surprisingly, it has become increasingly apparent that, for many of the subject matter departments, data scientists applying for these positions are using machine learning tools for data analysis. Their curricula vitae contain a staggering number of proceedings papers, and a traditional statistician may not recognize all the terms used in their titles. For such searches, we try to have on the hiring committee a mix of statisticians and computer scientists who can together decipher the research proposal. Acknowledging the cultural similarities between statistics and machine learning while not balking at their differences is a key element for a successful hire.

Hiring committee members will have to discuss and navigate important cultural differences between departments. For instance, on one such committee the statistician's favorite had four to five papers, each with fewer than twenty citations, and was contrasted with our partners' preferred candidate who boasted 40 publications and had a h-index greater than 20. It is clear that compromise and understanding must be part of the process, and this is achievable as long as both parties seek the best option for advancing the mission intended for the joint position. A certain level of institutional altruism is also required for success.

The trickiest part is to reach a comfortable level of confidence that the candidate has committed firmly and for the long-term to perform research in both areas of focus required for the search. This is difficult to establish for fresh graduates. Committees are often torn between taking a chance on a 'green' star-to-be versus chasing the mature candidate with great offers from all over the world.

The tension between maturity and elasticity must also be hedged in light of the future role played by the successful hire. It is clear to all parties involved that the 'twofer' data scientist is expected to catalyze collaborations between two departments. Co-supervision and coordination will be part of the

job, but as much as we hope to define the terms and conditions *a priori*, issues relating to practical matters such as funding, student supervision, and teaching are also factors.

The Benefits

Given our early start on this path, we are now positioned to see the benefits of our choices. The department recently organized a Faculty Research Day where a number of our 'twofers' presented their research projects. It was awesome! Gone are the days when one kind of knew what everyone else was working on. We now all relish our amazement at the sheer diversity of topics. In addition to the enriched research program, our students, in their new data science courses and research seminars, get to see messy data, learn how to practically tackle new problems, and maybe even get a better grasp of how statistical ideas can penetrate and ultimately change the scientific world. Who could wish for more?