Data science as culture

The origins of data science

The fathers of data science are the statisticians John Tukey (see Tukey (1962), Donoho (2015)) and Leo Breimam (Breiman 2001). Both emphasized the need for flexibility and fluidity in data analysis. To quote Tukey (1962):

> Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.

Breiman's paper (2001) contrasts "data modeling culture" with "algorithmic modeling culture", where the algorithmic modeling community is "often called machine learning". He describes his approach to data analysis as (Breiman 2001):

> (a) Focus on finding a good solution - that's what consultants get paid for.
> (b) Live with the data before you plunge into modeling.
> (c) Search for a model that gives a good solution, either algorithmic or data.
> (d) Predictive accuracy on test sets is the criterion for how good the model is.
> (e) Computers are an indispensable partner.

Not a field but a culture

The term "data science" is another name for the culture of data analysis that both Tukey and Breiman are describing. The term has emerged now because this culture has shown itself to be successful in adapting to a wide range of difficult data analysis problems. As Breiman implies in his point (e) above, part of this success has been because there are a much larger number of scientists who can write valid and effective code. This in turn is due to the success of open source scientific computing, and with it, the increasing use of standard tools to improve code quality.

We are going to see rapid change in data analysis across many fields, as the benefits of data science methods become more obvious. If we are early in this movement, we will have a chance to reap these benefits, making us more effective at research, and more attractive as a teaching university, for future employment in industry and research.

Recruiting for data science

It would be most effective to concentrate on searching for individuals who are leaders in this new *culture*, across domains. This culture is not exactly "machine learning", but it does involve flexibility in data analysis, and a willingness to engage with computational methods. We want to spread this culture, so we want to hire people who have already shown that they too want to spread this culture, and have done so in their home institutions. Meanwhile, we want to foster this culture in our own students and post-docs, by teaching and by example. Undergraduate teaching should involve substantial open data analysis projects, where students are encouraged to explore the data and find new methods of exploration and inference. We should give students a foundation in simple programming that will allow them to make new algorithms, and be flexible in applying old ones.

References

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16 (3). Institute of Mathematical Statistics:199–231. https://projecteuclid.org/euclid.ss/1009213726.

Donoho, David. 2015. "50 Years of Data Science." In *Princeton NJ, Tukey Centennial Workshop*. http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf.

Tukey, John W. 1962. "The Future of Data Analysis." *The Annals of Mathematical Statistics* 33 (1). JSTOR:1–67. http://projecteuclid.org/euclid.aoms/1177704711.