The Data Science Revolution

``It's a revolution," says Harvard's Gary King recently in the New York Times [1] ``...the march of quantification ... will sweep through academia, business and government. There is no area that is going to be untouched." [1]

Nearly every field of human endeavor is changing -- or even being transformed -- by unprecedented computational resources available to obtain, share, and analyze data. As a consequence, statistics has never been so central and in such high demand. More specifically, both in the academy and in industry, more research and commerce demands the ability to compute with abundant real world data. This set of skills has been called applied statistics, computational statistics, or simply `data science', a term coined over a decade ago by the statistician William Cleveland [2] but increasingly popular over the past two to three years [3].

The skills needed include those in the traditional curriculum as well as some more practical matters usually learned through research or projects classes. As outlined in a recent post which I coauthored [4], I think of these skills as:

- Obtaining data: Data Science requires the computational fluency (often in scripting languages such as UNIX shell scripting or python) for obtaining (usually via download or via interaction with a market or a set of customers) large sets of usable data.
- Scrubbing data: Real data are often noisy, corrupted by error, or awkwardly formatted. Needed is the ability not only to translate data from one format to another, but to strip away the chaff from the wheat.
- Exploring data: As the Princeton statistician John Tukey emphasized even in the 1970s [4], a great deal can be learned by choosing the right simple exploration tools, including pairwise statistical comparisons and clever exploratory data analysis. Often `unsupervised learning' (e.g., clustering) provides the necessary insight to begin a mathematical analysis.
- Modeling data: Statisticians bring traditional mathematical skills learned in the classroom, including supervised learning, regression, model selection (that is, how *not* to over fit data), and feature selection (that is, identifying interpretable features or measured variables which are the most predictive).
- Interpreting data: The role of the statistician is not only to calculate but to bring insights. This often requires patient understanding of domain knowledge, and, increasingly, fluency with visualization tools which can render ones model graphically. Particularly for high-dimensional or rich, structured datasets, a simple scatterplot or histogram may not suffice.

Most importantly one should remember that these separate aspects of modern `data science' inform each other: for example, feature selection in the course of supervised learning often reveals how the data should have been cleaned, or which data should have been obtained. A second insight to keep in mind: particularly in modern industrial applications where data are obtained online through millions of online transactions (purchases, clicks, or simply `page views'), the statistician's skill of experimental design takes on whole new dimensions. The insight of a `data scientist' as to what data can be obtained from online transactions can mean new insights into analytics but also suggest entirely new products or business models.

How is a student to prepare for `data science'? The relevant skillset includes core ideas of statistics, particularly predictive modeling, but also includes literacy in sufficient software engineering for steps 'O'

and 'S' above. In addition, one needs experience with visualizing the results of one's analysis in a way interpretable (step 'E' and 'I') to others. This requires statistical depth to understand the core insights of one's modeling, yet the domain expertise to be able to communicate these insights, usually graphically, in a language a domain expert can interpret. As the Stanford statistician Trevor Hastie recently said [5] ``Statistics ... needs others; applied statistics demands sufficient understanding of ones model to communicate insight to the non-statistician in the domain of application.

To come back to the first sentence: the difference in the past several years has been the abundance of data being computationally analyzed. For that reason familiarity with scientific computation --- including both numerical linear algebra and computational methods for approximate solution --- are key. Both in academic research and in industry, the stumbling block to implementing a statistical method is increasingly the methods needed to approximate a model (step 'M') in a way that large-scale data analysis is feasible on human time scales.

[1] http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html
[2] http://cm.bell-labs.com/cm/ms/departments/sia/doc/datascience.pdf
[3] http://tech.fortune.cnn.com/2012/01/06/data-scientist-jobs/
[4] http://www.dataists.com/2010/09/a-taxonomy-of-data-science/
[5] http://en.wikipedia.org/wiki/Exploratory_Data_Analysis
[6] http://bits.blogs.nytimes.com/2012/01/26/what-are-the-odds-that-stats-would-get-this-popular/

Chris Wiggins
Department of Applied Physics and Applied Mathematics Center for Computational Biology and Bioinformatics Columbia University New York, NY