Statistics and Big Data

"A significant constraint on realizing value from Big Data will be a shortage of talent, particularly of people with deep expertise in statistics and machine learning ... we project that demand for deep analytical positions in a big data world could exceed the supply being produced on current trends by 140,000 to 190,000 positions." -McKinsey *Big Data* report, 2011.

What is Statistics and what is Big Data?

- Statistics is the science of collecting, analyzing and understanding data, and accounting for the relevant uncertainties. As such, it permeates the physical, natural and social sciences; public health; medicine; business; and policy.
- Big Data is the collection and analysis of data sets that are complex in terms of the volume and variety, and in some cases the velocity at which they are collected. Big Data are especially challenging because some of them were not collected to address a specific scientific question.

How are Big Data problems being tackled?

- Big data problems usually require multidisciplinary teams by their very nature. At the very least, they typically require subject area (domain) experts, computational experts, machine learning experts, data miners, AND statisticians.

Why is it important for statistics to be one of the key disciplines for Big Data?

- Statistics is fundamental to ensuring meaningful, accurate information is extracted from Big Data. The following issues are crucial and are only exacerbated by Big Data:
  - Data quality and missing data
  - Observational nature of data
  - Quantification of the uncertainty of predictions, forecasts and models

> *In Big Data, statistical sciences and domain sciences are more intertwined than ever before, and statistical methodology is absolutely critical to making inferences.**

- Like in any data, one will find bias, false positives and uncertainty through the analysis of big data
- The scientific discipline of statistics brings sophisticated techniques and models to bear on these issues
- Statisticians help translate the scientific question into a statistical question, which includes carefully describing data structure; the underlying system that generated the data (the model); and what we are trying to assess (the parameter or parameters we wish to estimate) or predict

What does statistics bring to Big Data and where are the opportunities?

- Big Data will often not be served well by "off the shelf" methods or black box computational tools that work in low-dimensional and less complicated settings, and therefore require tailored statistical methods.
- Statisticians are skillful at assessing and correcting for bias; measuring uncertainty; designing studies and sampling strategies; assessing the quality of data; enumerating limitations of studies; dealing with issues such as missing data and other sources of non-sampling error; developing models for the analysis of complex data structures; creating methods for causal inference and comparative effectiveness; eliminating redundant and uninformative variables; combining information from multiple sources; and determining effective data visualization techniques.
- See the ASA Whitepaper: Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society.

The American Statistical Association (ASA) is a scientific and educational society of 19,000 members who serve in industry, government and academia in more than 90 countries, advancing research and promoting sound statistical practice to inform public policy and improve human welfare.

*Roger Peng, Johns Hopkins School of Public Health