How is Big Data Different? A Paradigm Shift

Jennifer Clarke, Ph.D.

Associate Professor
Department of Statistics
Department of Food Science and Technology
University of Nebraska Lincoln

ASA Snake River Chapter, Meridian, ID, May 29 2015

Outline

Shift 1. Trickle to Firehose

Shift 2. Experiment to Observation?

Shift 3. Low Dimension ($n > p$) to High Dimension ($n < p$)

Shift 4. Modeling to Prediction

Trickle to Firehose

- As statisticians we enjoy working with data - preprocessing, visualizing, modeling, inference, analysis, sharing, etc. These activities become difficult or impossible when the amount of data is large.
- Each piece of analysis requires considerable time and effort, and consideration of computational expense
- What do we do when data is so large it can't fit on one computer? What about 'garbage in, garbage out'?
- We rethink data as dynamic and emphasize random sampling

Trickle to Firehose

- Think about distributed processing. Can analysis be done in parallel or online? Hadoop, MapReduce
- Data gets bigger, not smaller, during preprocessing so plan ahead
- Computer scientists and system administrators are your friends
- Learn new computational skills (Python, SQL) and learn from others (social media)

Experiment to Observation?

- Much of Big Data is collected without thought. Period. Collecting is easy, analysis is hard.
- Experimental design has played a limited role, but is critical to inference and prediction.
- Sample size may not correlate with data size, e.g., genomics, social networks.
- Need for modern experimental design with detailed data provenance

Experiment to Observation?

- Do we need statistics? YES.
- 'To consult the statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.' Fisher
- Still relevant: sampling populations, confounders, multiple testing, bias, and overfitting
- Usually not randomization
- Unstructured data: information that either does not have a pre-defined data model and/or is not organized in a predefined manner. (www.mapr.com)

Low Dimension ($n > p$) to High Dimension ($n < p$)

- Figure out ways to make p smaller: variable selection, variable summarization, variable sampling.
- Variable selection: sure independence screening (Fan and Lv 2008), LASSO (Tibshirani 1996), regularization
- In the context of linear regression $E(y|X = x) = \alpha + \beta x$, estimates are chosen to minimize $\arg\min_{\beta} \sum (y_i - \alpha - \beta x_i)^2$ subject to $\sum |\beta_j| < t$. This is equivalent to minimizing $\frac{1}{2n} \sum (y_i - \alpha - \beta x_i)^2 + \lambda \sum |\beta_j|$.
- one can summarize variables by PCA, cluster medioids, etc.

Low Dimension ($n > p$) to High Dimension ($n < p$)

- Variable sampling (huh?). Can model subsets of data where subsets are random samples of observations AND variables.
- Overfitting is a BIG problem.
- Correct for multiple testing ... FDR, Westfall-Young
- Problem drives solution ... don't 'hit all the nails'

Shift 4. Modeling to Prediction

- Idea: When data is unstructured or otherwise complex, model uncertainty is high
- If the goal is prediction, average or ensemble. Think bagging, boosting. This will reduce variability without increasing bias (while accounting for model uncertainty).
- Focus on accurate prediction and sequential/prequential analysis (interactive)
- Approach inference by assessing and dissecting predictor
- Uncertainty and variability based on random sampling and/or permutation
- Error rate depends on correlation between models and strength of individual models

Shift 4. Modeling to Prediction

- Linear models are like donkeys. Treat them right and they'll carry you, grudgingly, as far as they can. They will bray when they are unhappy but you will know how to feed and water them. They don't travel far.
- Neural Networks are like a big pile of snakes. Some are poisonous though most aren't. There are different sizes and colors. You can grab one that looks right but since you can't see the whole thing it will coil around in unexpected ways.
- Trees are like foxes. They're clever and run around all over the place. Catch the ones that will solve your problem.
- Ensemble methods are packrats. They always have something that works well but it may be a mess. This can save time catching snakes, chasing foxes, or beating donkeys. But, you don't know what you've really got.