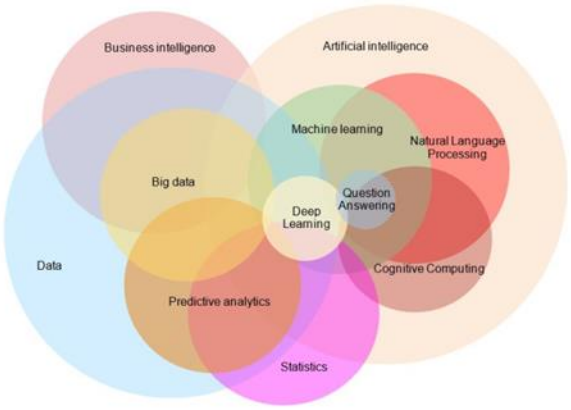
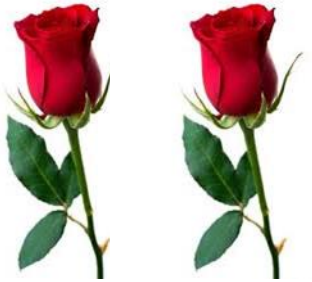
The Role of Statistics in Modern Data Analysis
or
A Rose by Any Other Name





Hal Stern
Department of Statistics
University of California, Irvine
sternh@uci.edu

UCI Department of Statistics
Donald Bren School of Information & Computer Sciences

A quote from a paper entitled “Data Analysis and Statistics: An Expository Overview”

Four major influences act on data analysis today:

1. The formal theories of statistics.
2. Accelerating developments in computers and display devices.
3. The challenge, in many fields, of more and ever larger bodies of data.
4. The emphasis on quantification in an ever wider variety of disciplines.

A quote from a paper entitled “Data Analysis and Statistics: An Expository Overview”

Four major influences act on data analysis today:

1. The formal theories of statistics.
2. Accelerating developments in computers and display devices.
3. The challenge, in many fields, of more and ever larger bodies of data.
4. The emphasis on quantification in an ever wider variety of disciplines.

DATA ANALYSIS AND STATISTICS: AN EXPOSITORY OVERVIEW *

1966, Proceedings Fall Joint Computer Conference

J. W. Tukey and M. B. Wilk

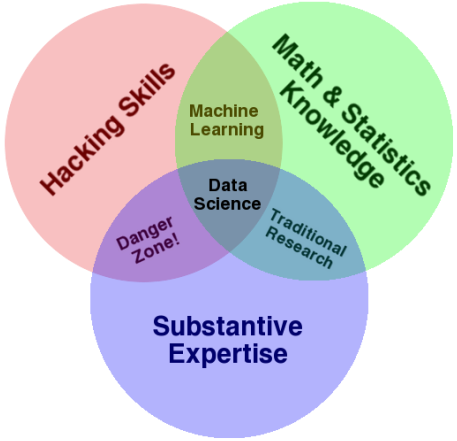
Princeton University and
Bell Telephone Laboratories, Inc.
Princeton and Murray Hill, New Jersey

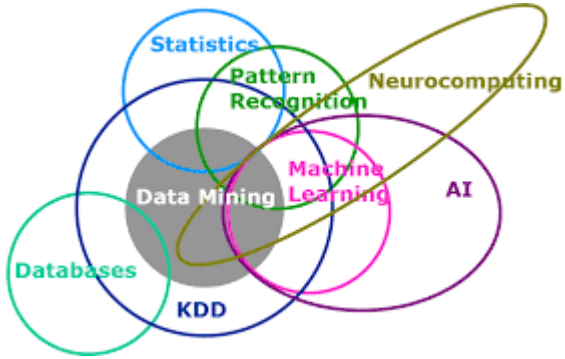
Terminology

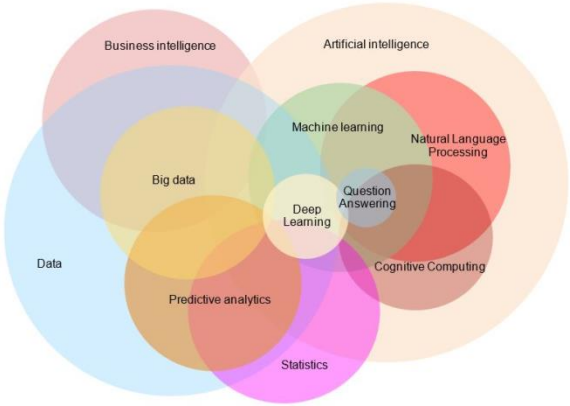
- There are many terms associated with data analysis these days. Examples include:
  - Statistics
  - Machine Learning (ML)
  - Data Science
  - Big Data
  - Artificial Intelligence (AI)
  - Deep Learning (Deep Neural Networks)
- This has proven confusing and led to many attempts to clarify …

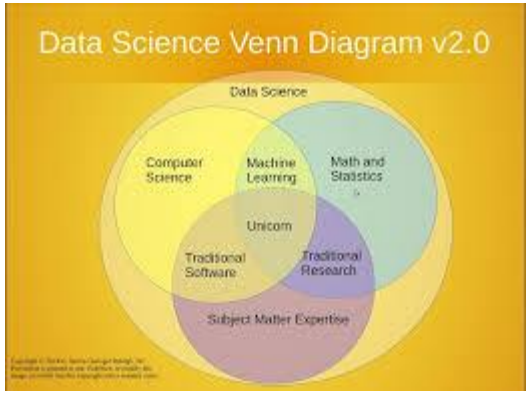
A Venn Diagram to Explain it All

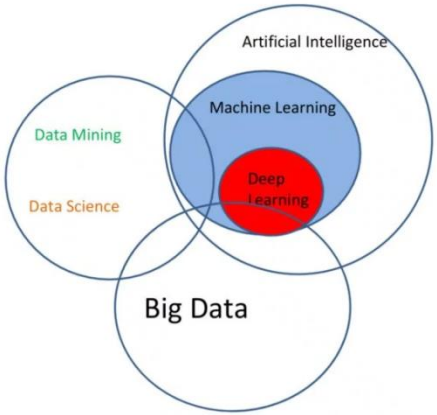
One of the first by Drew Conway

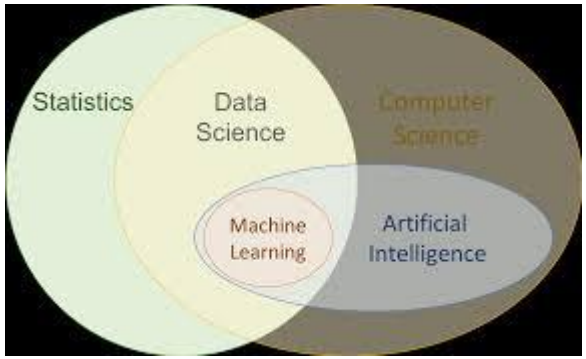






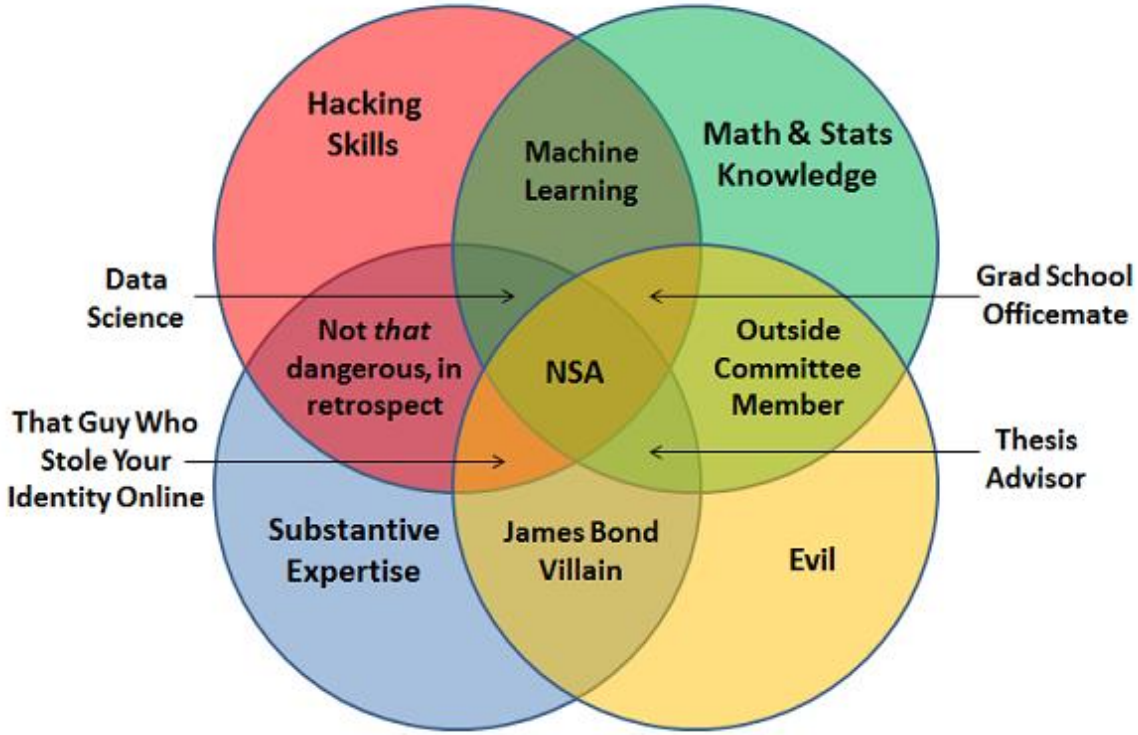






Terminology

A humorous take credited to Joel Grus:

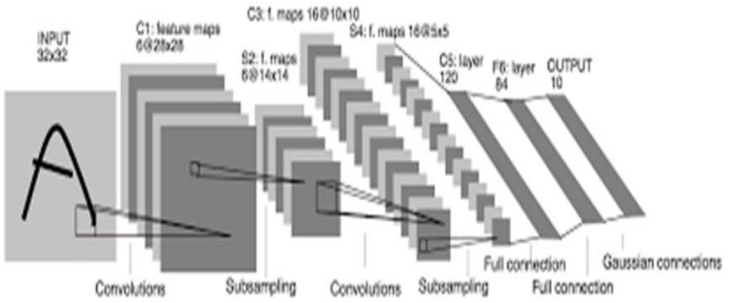


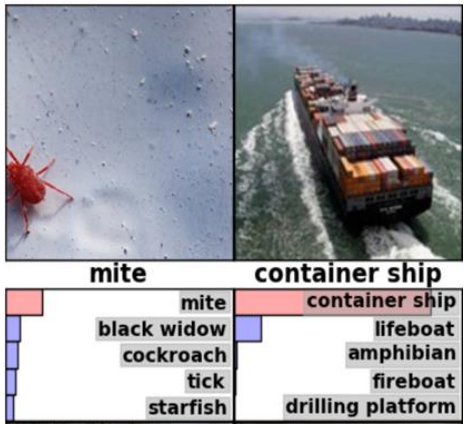
Terminology

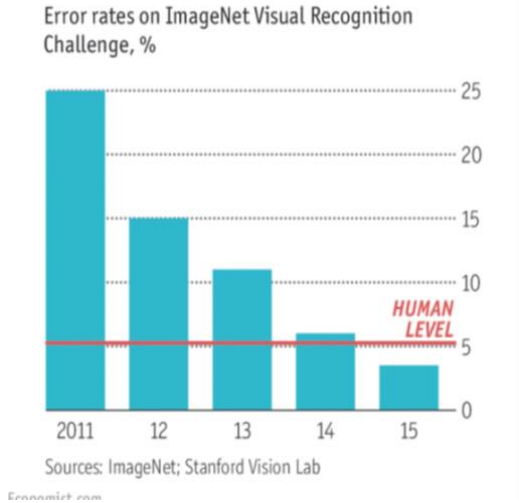
- Some parts of the story are reasonably clear:
  - AI contains ML contains Deep Learning
  - AI goes back to at least the early 1950s
    - Had several incarnations
    - Development of programmable computers
    - Neural networks
    - Expert systems
  - Machine Learning terminology was coined early (1959) but emerged as a force in the 1990s
    - Essentially recognition of some in CS that true AI was “too hard”
    - Led to a split of the AI community
    - Better to find algorithms to solve particular problems

Deep Learning

- Deep Learning refers to a particular class of machine learning algorithms (deep neural networks) that have proven **incredibly effective** at certain tasks (e.g., Siri: "What can I help you with?")
  - Neural networks have a very long history (computational models of the neuron date back to the 1940s)
  - Neural networks were extremely popular in the 1980s
  - Essentially died in the 1990s (they were not better than any other predictive technology)
  - Came back with a vengeance!!







Terminology

- The relationship of AI / ML to Statistics is much less clear
- Definitions of statistics tend to focus on collection, analysis and interpretation of numerical data
- Definitions of artificial intelligence usually refer to enabling computer systems to perform tasks normally assumed to require human intelligence
- Definitions of machine learning
  - Originally emphasized the capacity of a computer to modify itself based on the results of its processing
  - Now tends to refer to the study of algorithms / models for performing tasks (e.g., classification)

Terminology

- The (stereotyped) view of ML from Statistics
  - ML represents computer scientists discover the power of probability and statistical models to solve problems / analyze data
  - ML folks are not concerned about where the data come from
- The (stereotyped) view of Statistics from ML
  - Statistics is focused on mathematical theories for data analysis
  - Statisticians think more about interpretation / testing of models and less how algorithms work
  - Can't handle very large data sets

Contributions of CS

- Key areas of expertise
  - Databases
  - Algorithms
  - Programming innovation
- These skills allowed ML researchers to
  - More easily obtain data (e.g., web crawling)
  - Manipulate large, heterogeneous data sources
  - Scale up algorithms to handle the large data sets

Contributions of Statistics

- Key areas of expertise
  - Experimental design and statistical sampling for data collection
  - Underlying mathematical theory that justifies procedures
  - Measures of uncertainty (e.g., confidence intervals)
  - Emphasis on the distinction of correlation and causation

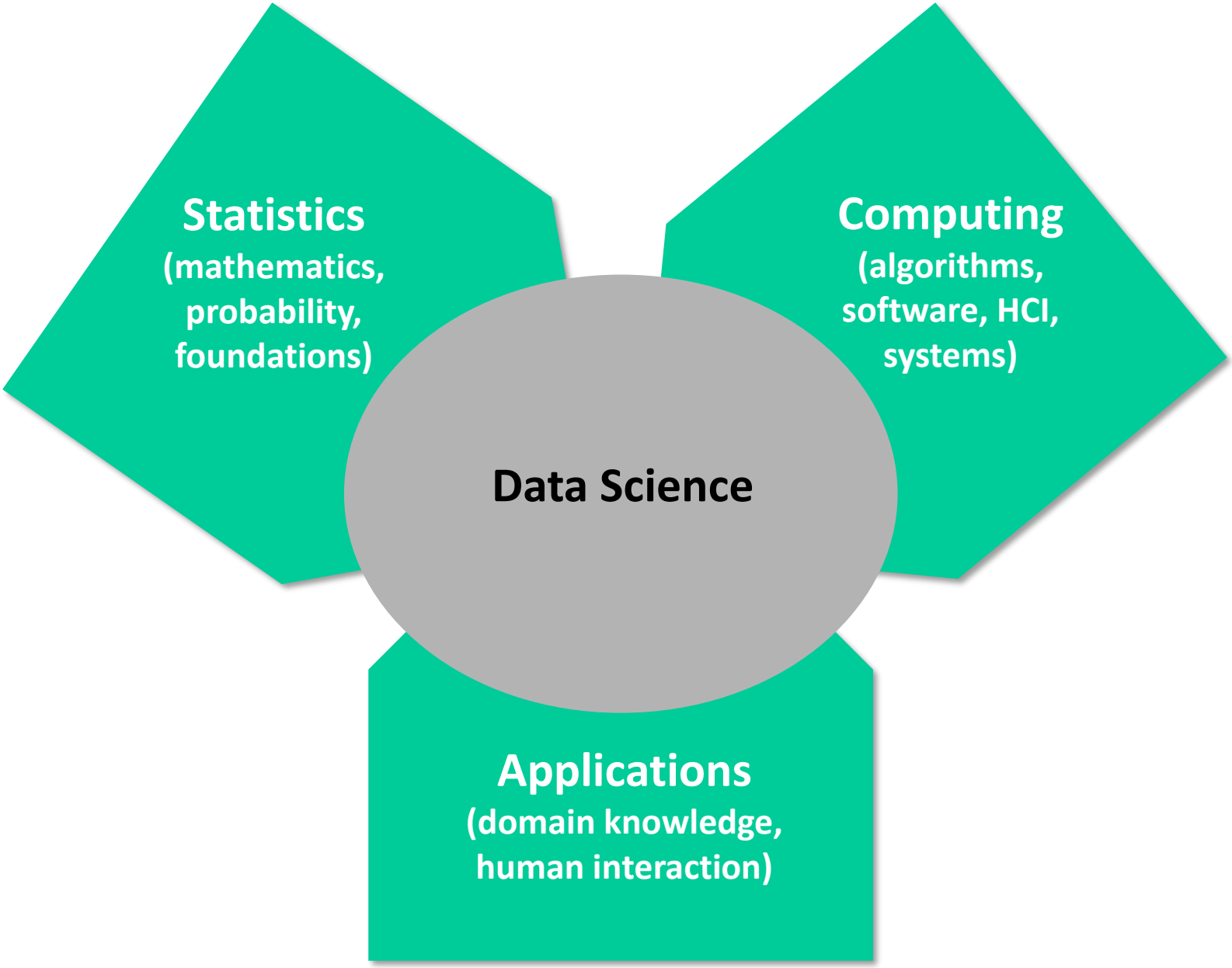
Modern Data Analysis

- Seems clear that a happy marriage is possible

"You got peanut butter on my chocolate!"

"You got chocolate in my peanut butter!"

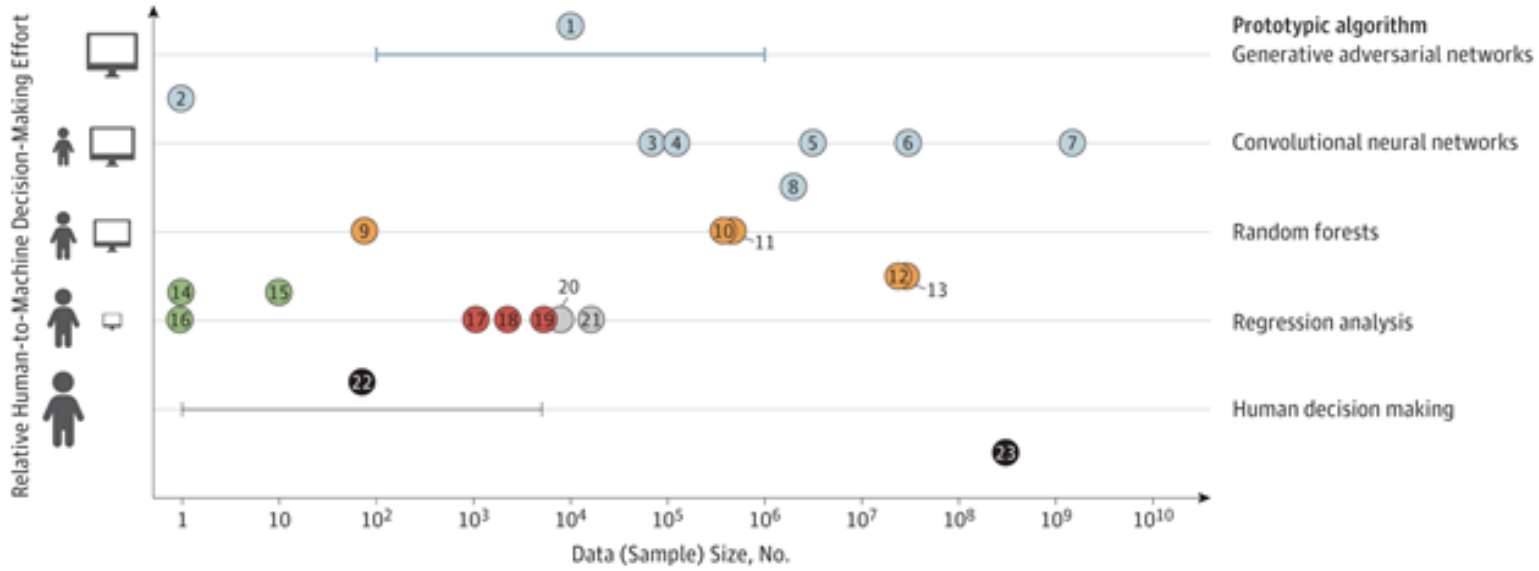
- Some call the marriage "Data Science" (but others really hate the idea!)


Statistics
(mathematics, probability, foundations)
Computing
(algorithms, software, HCI, systems)
Data Science
Applications
(domain knowledge, human interaction)

Modern Data Analysis

- There exist a vast array of data analysis methods/strategies
- Includes techniques from statistics, machine learning, artificial intelligence
- Methods / models vary on a number of dimensions
  - Human input required to build the model
  - Size / dimension of the model (number of parameters)
  - Amount of data required to fit the model
  - Strength of the assumptions required
  - Interpretability of the results of applying the model

Beam and Kohane, JAMA 2018



Deep learning

1. Generative adversarial networks (2014)
2. Google AlphaGo Zero (2017)
3. ATM check readers (1998)
4. Google diabetic retinopathy (2016)
5. ImageNet computer vision models (2012-2017)
6. Google AlphaGo (2015)
7. Facebook Photo Tagger (2015)
8. Prediction of 1-y all-cause mortality (2017)

Classic machine learning

9. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling (2002)
10. EHR-based CV risk prediction (2017)
11. Netflix Prize winner (2006)
12. Google Search (1998)
13. Amazon product recommendation (2003)

Expert AI systems

14. MYCIN (1975)
15. CASNET (1982)
16. DXplain (1986)

Risk calculators

17. CHA_2DS_2-VASc Score for atrial fibrillation stroke risk (2017)
18. MELD end-stage liver disease risk score (2001)
19. Framingham CV risk score (1998)

Randomized Clinical Trials

20. Celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis (2002)
21. Use of estrogen plus progestin in healthy postmenopausal women (2002)

Other

22. Clinical wisdom
23. Mortality rate estimates from US Census (2010)

Modern Data Analysis

- Applications also vary
  - Objectives
    - Prediction / interpretation
    - Effects of causes / data exploration
  - Amount of data
    - e.g., logistic regression on millions of cases
    - e.g., presence / absence of training data
  - Heterogeneity of data types (images, text)
  - Frequency of the analysis (one-time vs repeated model fitting)
  - Need for fairness / equity

What's a researcher to do?

- Use an appropriate method for the scientific context
  - The right method depends on the objectives
    - Purely predictive (computer vision, Siri)
      - Deep neural networks
    - Interpretation and prediction
      - Random forests, linear models
    - Inference / effects of causes
      (program evaluation, treatment effectiveness)
      - Regression analyses (linear, logistic), causal inference
    - Data description / exploration
      - Visualization tools, clustering algorithms

What's a researcher to do?

- Use an appropriate method for the scientific context
  - The right method depends on characteristics of the application
    - Amount of data available
      - e.g., deep learning in forensics?
    - Heterogeneity of data types
      - e.g., image analysis and the definition of relevant features
    - Need for fairness
      - e.g., automated sentencing guidelines
  - There is a need for transparency (i.e., being clear about what approach was used and why)

Conclusions

- Statistics continues to have a large role to play in modern data analysis
  - Expertise in experimental design and data collection
  - Ensure that the importance of uncertainty and variability is recognized
  - Wide-range of techniques that have proven useful in science and policy
- The impact of statistics and statisticians will grow if
  - The field embraces new methods/models and studies their strengths and weaknesses
  - The field embraces new developments in computer science (and statistics students are taught this material)
  - Statisticians work well in collaborative teams
- Thank you!

Comments/questions: sternh@uci.edu

UCI Department of Statistics
Donald Bren School of Information & Computer Sciences