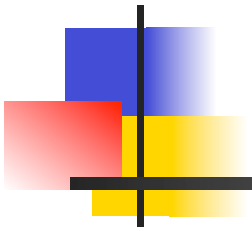
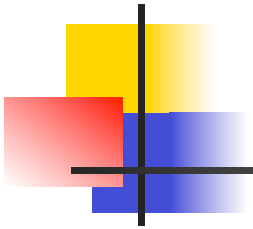
The Role of Statistics in Data Science, and Vice Versa



Jessica Utts
Professor of Statistics
University of California, Irvine
President, American Statistical Association

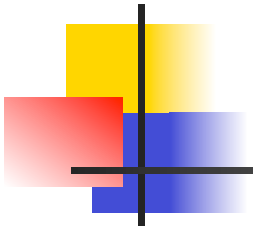
Nicholas Horton
Professor of Statistics
Amherst College

Some Issues for Discussion

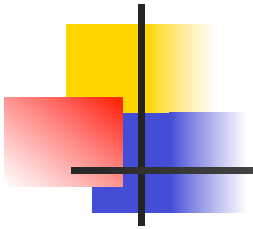


- How does statistics (as a discipline) view the emerging field of data science?
- What can statisticians contribute to data science?
- What elements of statistics are essential for data science education?

Overview and History



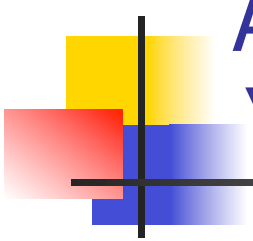
- Statistics has evolved along with technology and the growth of data
  - Statistics from the 1990s ≠ Statistics today!
- Foundational goal is the same
  - ASA's vision statement says it well: "A world that relies on data and statistical thinking to drive discovery and inform decisions"
- But methods for achieving that goal have changed and expanded



A Very Early Adopter: John Tukey
1962, *Annals of Mathematical Statistics*

Identified four driving forces in the "new science":

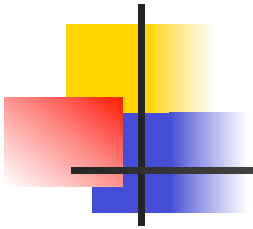
1. The formal theories of statistics
2. Accelerating developments in computers and display devices
3. The challenge, in many fields, of more and ever larger bodies of data
4. The emphasis on quantification in an ever wider variety of disciplines



A Less Early Adopter: Leo Breiman, 2001 "Statistical Modeling: The Two Cultures"

There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models… Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

(Statistical Science, 2001, with discussants)



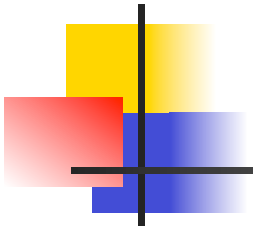
A Side Comment

David Donoho's "50 Years of Data Science" (2015) is worth reading. His version of Breiman's 2 cultures:

- The "Generative [stochastic data] Modeling" culture seeks to develop stochastic models which fit the data, and then make inferences about the data-generating mechanism based on the structure of those models.
- The "Predictive [algorithmic] Modeling" culture prioritizes prediction… is effectively silent about the underlying mechanism generating the data, and allows for many different predictive algorithms, preferring to discuss only accuracy of prediction made by different algorithms on various datasets.

Fast Forward 14 Years: ASA Statement on Role of Data Science in Statistics, 2015

- Identifies foundational data science fields:
  - Database management
  - Statistics and machine learning
  - Distributed and parallel systems
- Encourages greater, mutually beneficial collaboration across these three fields
- Intersects with numerous disciplines and related research areas



Many ongoing disciplinary collaborations

Some examples:

- Genomics (and personalized medicine)
- Health services research (electronic medical records)
- Business analytics (customer tracking)
- Smart cities (and sensor networks)
- Astronomy (data streams)

And others…

ASA Statement, Continued

- Notes that statistics education must evolve to meet needs
  - For example, address inclusion of data science in K-12, community college
  - More later on other aspects of education
- Elucidates role of statistics in data science

From the ASA Statement: The Role of Statistics

- Framing questions statistically allows researchers to leverage data resources to extract knowledge and obtain better answers.
- The central dogma of statistical inference, that there is a component of randomness in data, enables researchers to formulate questions in terms of underlying processes and to quantify uncertainty in their answers.
- A statistical framework allows researchers to distinguish between causation and correlation and thus to identify interventions that will cause changes in outcomes.

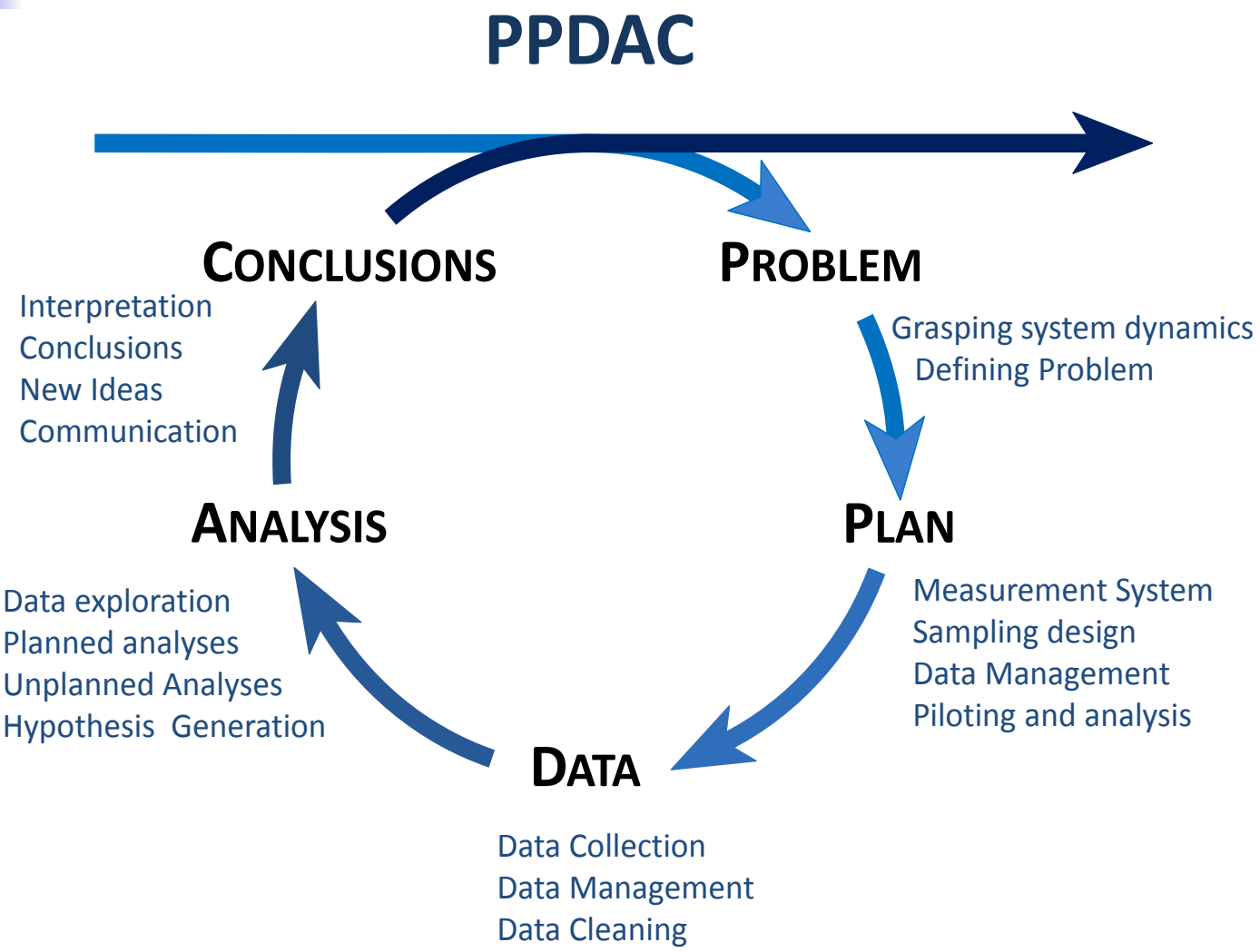
The ASA Statement, continued

- It also allows them to establish methods for prediction and estimation, to quantify their degree of certainty, and to do all of this using algorithms that exhibit predictable and reproducible behavior.
- In this way, statistical methods aim to focus attention on findings that can be reproduced by other researchers with different data resources.
- Simply put, statistical methods allow researchers to accumulate knowledge.

The Statistical Inquiry Cycle

Wild and Pfannkuch, 1999, *International Statistical Review*

Problem, Plan, Data, Analysis, Conclusions



How to carry out PPDAC?

"This scientific approach to statistical problem-solving is important for all data analysts. It needs to start in the first course and be a consistent theme in all subsequent courses." - American Statistical Association Guidelines for Undergraduate Programs in Statistics (2014), http://www.amstat.org/asa/education/Curriculum-Guidelines-for-Undergraduate-Programs-in-Statistical-Science.aspx

How to carry out PPDAC?

"Working with data requires extensive computing skills. To be prepared for statistics and data science careers, students need facility with professional statistical analysis software, the ability to wrangle data in various ways and algorithmic problem-solving. Students should be fluent in higher-level programming languages and facile with database systems." - American Statistical Association Guidelines for Undergraduate Programs in Statistics (2014), http://www.amstat.org/asa/education/Curriculum-Guidelines-for-Undergraduate-Programs-in-Statistical-Science.aspx

How to carry out PPDAC?

- Statistical Methods and Theory: Need to
  - understand issues of design, confounding, and bias,
  - have a foundation in theoretical statistics principles for sound analyses,
  - develop knowledge and gain experience applying a variety of statistical methods,
  - assess appropriateness of methods, and
  - communicate results

How to carry out PPDAC?

- Data Wrangling and Computation: Need to
  - be facile with professional statistical software
  - program in a higher-level language and think algorithmically,
  - use simulation-based statistical techniques and undertake simulation studies,
  - manage and wrangle data, and
  - undertake analyses in reproducible manner

How to carry out PPDAC?

- Statistical Practice and Communication: Need to
  - write clearly, speak fluently, and construct effective visual displays and compelling summaries,
  - demonstrate ability to collaborate in teams and to organize and manage projects,
  - incorporate ethical precepts into all aspects of their work, and
  - communicate complex statistical methods in basic terms to managers and other audiences

How to carry out PPDAC?

- Discipline-Specific Knowledge: Need to
  - apply statistical reasoning to domain-specific questions,
  - translate research questions into statistical questions, and
  - communicate results appropriate to different disciplinary audiences.

- Skills taken from undergraduate guidelines, but relevant at other levels as well

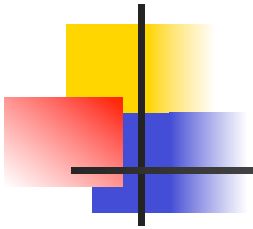
Park City Group Report (2016)

Curriculum Guidelines for Undergraduate Programs in Data Science (DeVeaux + 24 other authors)

- Data science as science
- Interdisciplinary nature
- Data at the core
- Analytical (computational and statistical) thinking and problem-solving
- (New pathways for) mathematical foundations
- Flexibility

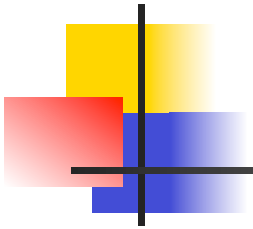
http://www.amstat.org/asa/files/pdfs/EDU-DataScienceGuidelines.pdf

What do statisticians bring to the table?



- Importance of context
- Accounting for variability
- Design, confounding, and analysis of found (observational) data
- Understanding of inference, multiplicity and reproducibility issues
- Statistical analysis (PPDAC) cycle
- Long history of making decisions with data
- Experience working on multidisciplinary teams

Some Issues for Discussion



- How does statistics (as a discipline) view the emerging field of data science?
- What can statisticians contribute to data science?
- What elements of statistics are essential for data science education?