An Interdisciplinary Data Science Design for Undergraduate Students

dB-SERC Presentation 1

University of Pittsburgh
November 2017
Lucas Mentch, Dept. of Statistics

Brief Overview

STAT 1361: Statistical Learning and Data Science

- Overall Goal: The development of new data science related course in the statistics department aimed at sophomore/junior level undergraduate students
  - Part of larger (longer-term) effort to modernize curriculum
- Introductory level course also being developed (STAT 1261; Sungkyu Jung)
- Trial run of STAT 1361 offered Spring 2017 as topics course (STAT 1291)

Big Questions

1. What is Data Science?
2. Why should we care?
3. How can we create effective courses within existing major/department structures?

What is Data Science?

What is Data Science?

- *The field of data science is emerging at the intersection of the fields of social science and statistics, information and computer science, and design*

  -UC Berkeley School of Information

- *Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, machine learning, data mining, and predictive analytics*

  -Wikipedia

What is Data Science?

- *At its core, data science involves using automated methods to analyze massive amounts of data and to extract knowledge from them.*

  -NYU

- *Data science is a multidisciplinary blend of data inference, algorithmm development, and technology in order to solve analytically complex problems.*

  -datajobs.com

What is Data Science?

- *While there is not yet a consensus on what precisely constitutes data science, three professional communities,* **all within computer science and/or statistics**, *are emerging as foundational to data science: (i) Database Management enables transformation, conglomeration, and organization of data resources, (ii) Statistics and Machine Learning convert data into knowledge, and (iii) Distributed and Parallel Systems provide the computational infrastructure to carry out data analysis.*

-American Statistical Association (ASA)
ASA Statement on the Role of Statistics in Data Science

What is Data Science?

Working Definition: The general process of collecting, storing, and analyzing data in order to extract useful information.

$$\text{Data} \implies \text{knowledge}$$

- Utilizes skills and ideas from the fields of
  1. Statistics
  2. Computer Science
  3. Information Science
  4. Mathematics

Where does that leave us?

- So is "data scientist" just the new word for statistician?
- Are "data scientists" just statisticians with good computational skills?

- Bin Yu (UC Berkeley) 2014 IMS Presidential Address:
  - *"let us own data science"*
  - *"No existing discipline does more of the job of a data scientist "*
  - *"We do the job so let us call ourselves data scientists!"*
    - **But do we ... ?**

Breiman's Two Cultures

The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and ***has kept statisticians from working on a large range of interesting current problems.*** *Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics...* ***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.***

-Leo Breiman, 2001

"Statistical Modeling: The Two Cultures"

Where does that leave us?

- Question still remains: How did this thing (data science) become a thing?

Two primary reasons:

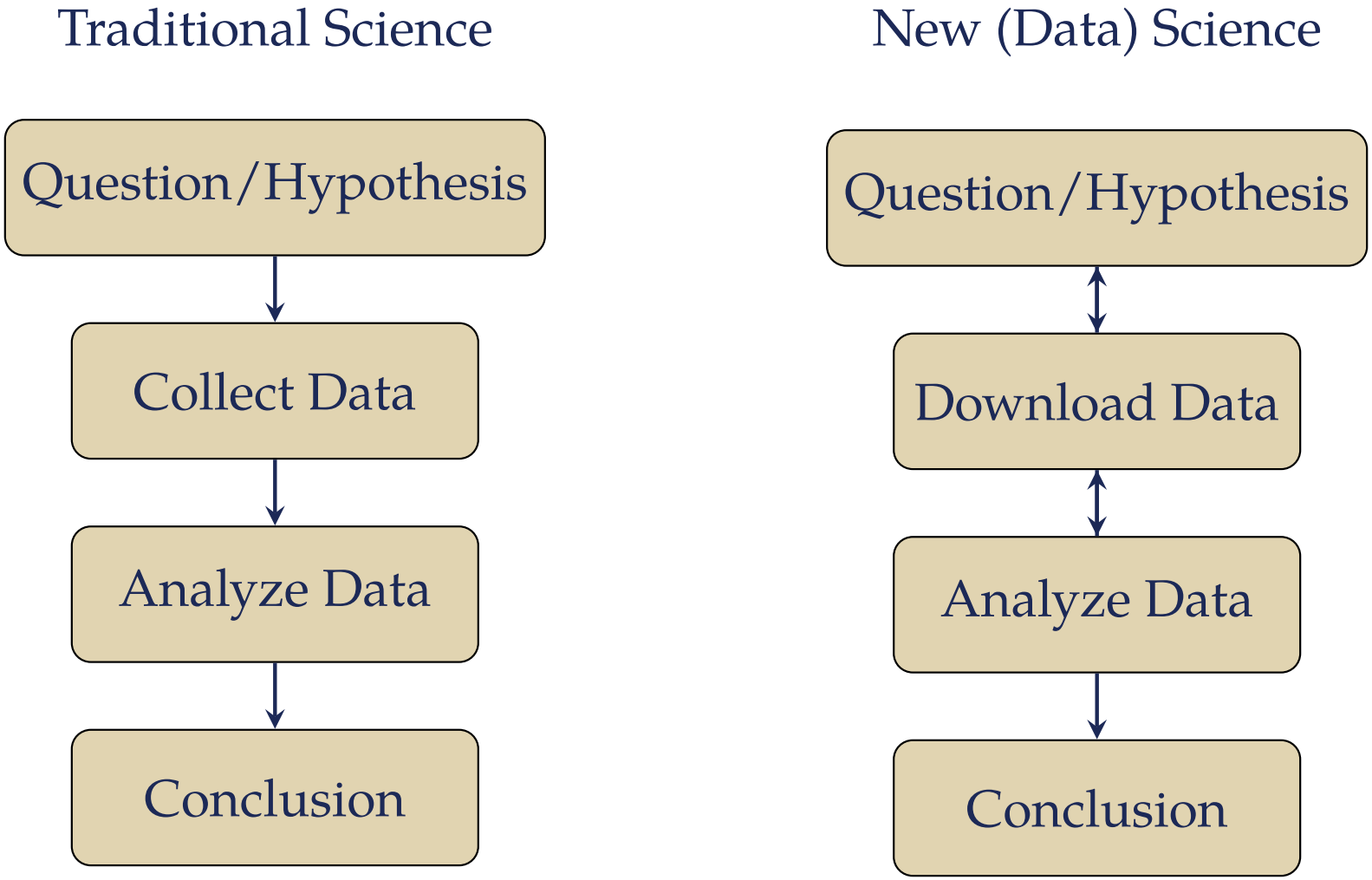
- Statisticians (and others in related fields) can be stubborn
  - *"Statistics has a bad PR department"*
- The rapid growth in storage and accessibility of data has lead (and continues to lead) to changes in scientific thinking and practice

The Science in Data Science

- 30+ years ago, scientists (mostly) followed the traditional scientific method
  - New ideas (hypotheses) were formed on the basis of prior experiments
  - To test these ideas, scientists collected and analyzed data to determine the strength of evidence
  - Much of the data collection took significant time, effort, and training (i.e. $)

- Now fast forward into the modern age:
  - **Before:** "Were going to need a $1M and 3 years to collect 30 samples on 3 variables"
  - **After:** "Were going to download 10,000 samples from the internet measured on 500 variables"

The Science in Data Science



The Good News

- For much of scientific history, little emphasis was placed on the analysis of the data
  - Reliance on simple (often inappropriate) methods and models
  - Scientists received little to no training in statistics / data analysis
  - Lead to "p-value crises" and bad science
- The rise of data analysis and big data has forced scientists to gradually move away from simple, unreliable models
  - Better more robust results
  - Development of state-of-the-art methodology
  - More work for statisticians :)
- Bigger and more data should never be a bad thing

Who Cares?

Who Cares?

Even though statistics fastest growing STEM degree ...

- Indeed Job Search, 12/11/2016:
  - "Statistician" in Pittsburgh, PA:

Who Cares?

Even though statistics fastest growing STEM degree ...

- Indeed Job Search, 12/11/2016:
  - "Statistician" in Pittsburgh, PA:
    - 13 Jobs Found

Who Cares?

Even though statistics fastest growing STEM degree ...

- Indeed Job Search, 12/11/2016:
  - "Statistician" in Pittsburgh, PA:
    - 13 Jobs Found
  - "Data Scientist" in Pittsburgh, PA

Who Cares?

Even though statistics fastest growing STEM degree ...

- Indeed Job Search, 12/11/2016:
  - "Statistician" in Pittsburgh, PA:
    - 13 Jobs Found
  - "Data Scientist" in Pittsburgh, PA
    - 937 Jobs Found

Who Cares?

Even though statistics fastest growing STEM degree ...

- Indeed Job Search, 12/11/2016:
  - "Statistician" in Pittsburgh, PA:
    - 13 Jobs Found
  - "Data Scientist" in Pittsburgh, PA
    - 937 Jobs Found

- Glassdoor 25 Best Jobs in America:
  1. Data Scientist
     - Median Base Salary: $ 116,840
     - Highest Career Opportunity Score

Related Buzz Words

- Statistics
- Machine Learning
- Statistical Learning
- Deep Learning
- Big Data
- Data Engineer
- Data Mining
- Data Analyst
- Data Architect
- Business Analytics
- Artificial Intelligence
- Analytics Associate
- Information Analyst
- Predictive Analytics

Takeaway

- Lots of statistics going on outside of statistics
  - But ... most (if not all) of these ideas are firmly grounded in statistical thinking
- Need to think about how we can integrate the necessary skills into a statistics department curriculum while stressing ideas related to assessing uncertainty and appealing to a wider audience

How do we do it?

Course Structure: Evaluations

- Many low-stakes assessments:
  - Quizzes (5-6) (20% total)
  - Homework (8-10) (30% total)
- One large semester-long group project:
  - Project Proposals (5%)
  - Project Updates (5%)
  - Project Summaries (5%)
  - Oral Presentation (15%)
  - Written Report (20%)
- Extra credit for attending course related seminars and writing report

Goals

- Appeal to a wide variety of undergraduate students
- Design course in a scalable fashion to accommodate larger classes and/or more sections
- Cover the range of data analysis and modeling techniques ranging from traditional statistics (e.g. linear models) to modern machine learning (e.g. random forests, SVMs)
- Stress the "thinking" aspect of data science and statistical modeling
- Integrate hands-on experience and real-world application; confront and assess common practical problems

Meeting the Goals:

Appealing to variety of students:

- Classes appealed to many stat majors, but also majors from:

Finance, Mathematics, Economics, Political Science, Bioinformatics, Chemistry, Philosophy, Biological Sciences, Actuarial Mathematics, Accounting, Mathematical Biology, Neuroscience, Computer Science, Psychology, Communications, Spanish, Supply Chain Management, and Chinese

Meeting the Goals:

Scalability:

- Most course materials can easily scale to larger and/or more sections:
  - Creating standard set of course material
- Major bottleneck: Oral presentation component of course project
  - Could be scaled back (more ideal) or omitted (less ideal)

Meeting the Goals:

Covering a range of modeling techniques:

- Unlike most math/stats courses, stress breadth over depth
  - Total of 15+ modeling techniques covered in one semester
  - Many methods too mathematically complex to derive formally, but providing a brief overview with motivation and pictures is often enough to get the main ideas across
    - e.g. Linear models → LASSO; Ridge Regression
- Emphasize settings where models might perform better/worse; encourage students to think about how contradictions might arise and be resolved/further investigated
  - Get's into the *thinking* aspect of the course

Meeting the Goals:

Stressing hands-on experience and thinking about fundamental issues:

- Weekly (roughly) homework assignments each contain at least one problem stressing each aspect
  - Real data walkthrough of the modeling method
  - Compare contrast output (sometimes hypothetical) with other approaches
- Discussion questions in class to get students perspectives on the different techniques
- Course project

Course Project:

Semester-long course project overview:

- Within first few weeks of the course, students find a dataset and propose possible projects
- Based on proposals, students get into "paired" groups of 4-5 and a single project is chosen for each group
- As we cover material/methods in the course, group members experiment on their dataset
- Project (and course) concludes with a write-up and oral presentation

Course Project:

- Students are given relatively few guidelines in selecting a dataset/project proposal (just need a reasonably big dataset and/or interesting question(s))
- At least two "check-in" days scheduled throughout the semester to get feedback/answer questions
  - The hope is that new discoveries may push the project in different directions
- By the final days of the course, groups should have constructed several models and be able to describe their results in a "real-world" fashion (without relying on too much statistical jargon)

Difficulties:

- Students stuck in the "student-only" mindset
  - Don't like the open-endedness – want to be told specifically what to do
  - Real pushback to the idea of "practical thinking" – want to conclude only exact model interpretations
    - Failure to think about real-world implications
  - The involvement of multiple modeling techniques encourages a "this model has lowest error so I trust it most" mindset
  - Often fearful of stating "negative" outcomes (e.g. should have used different data / theres nothing to see here)

Difficulties:

Potential Solutions:

- Stress the "big-picture-mindset" throughout the course
- Give examples where simple, high-level thinking could be better than hours of model-building and sifting
- Integrate small portions of the project into homework assignments throughout (and earlier in) the course
- Attempt to grade based on quality/innovation/defensibility of results as opposed to mere quantity