Lecture 1: What is Data Science?

STAT 1361/2360: Statistical Learning and Data Science

University of Pittsburgh
Prof. Lucas Mentch

Instructors

Professor: Dr. Lucas Mentch

Dept of Statistics, University of Pittsburgh

Email: lkm31@pitt.edu

Office: 1807 WW Posvar Hall

Office Hours: Tuesday 2:00 PM - 3:00 PM*** (Virtual)

*** May need to change

Instructors

Head TA: Ryan Cecil

Dept of Statistics

Email: RMC144@pitt.edu

Office: 1824 WW Posvar Hall

Office Hours: Thur 11:00 AM - 1:00 PM & By Appt (In Person)

TA / Grader: Alex Dukart

Dept of Statistics

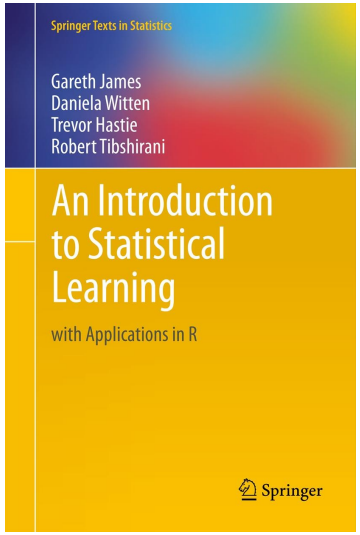
Email: ALD349@pitt.edu

Office: 1824 WW Posvar Hall

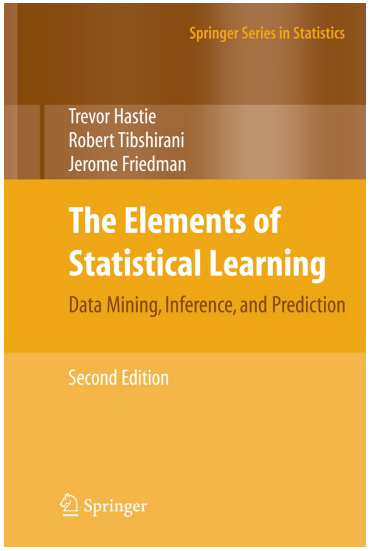
Office Hours: By Appt

Course Overview

- Full syllabus available on canvas
- Goal of the course is to provide a high-level overview of data science, with a particular focus on statistical aspects and tools
- Note: Textbook now has 2^{nd} edition; freely available online



Undergraduate Version



Graduate Version

Course Overview

Detailed (**tentative) schedule available on canvas**. General overview:

- Data science basics (today); homework overview (Next Wednesday)
- Stat vs ML; Bias/variance; underfitting/overfitting (ISLR Ch. 2)
- Regression basics (ISLR Ch. 3)
- Classification basics (ISLR Ch. 4)
- New ideas: resampling; validation; variable importance; permutation tests (ISLR Ch. 5 + Notes)
- Regularization; High dimensional problems; linear model extensions (ISLR Ch. 6-7)
- Statistical & machine learning methods; boosting; forests; neural networks (ISLR Ch. 8, 10)

Grades

- Quizes – 25%
- Homeworks – 50%
- Final Project – 25%

Quizzes

- Expect a total of 5 quizzes which will be given **at the start of class** – planned to be in person but will be given via canvas if we're still virtual when scheduled
- You will be given advance notice of all quizzes (i.e. not pop-quizzes)
- Designed to take 15 minutes; You will have 20-25 minutes to complete
- These are designed to be a review of fundamental concepts covered in past lectures **and from relevant textbook chapters**
- Quizzes are closed book and closed note. No aids of any kind are permitted.

Homework

- Homework assignments should be turned in via Canvas
- Expect a total of 7 assignments; **these will often require significant time and effort. Start them early!**
- **Homeworks are due at 5:00 PM Fridays – Barring extreme circumstances or prior written permission (>24 hours), late homework will not be accepted.**
- Homework assignments **must** be neatly done (this should not be your first draft of a problem solution) and show all relevant work, code, and results
- The TA reserves the right to refuse any homework that does not follow these guidelines

Project

- A primary component of this course is an end-of-semester data science project
- Analyze a dataset; provide both technical and “high-level" reports
- Goal is to simulate a real-world setting – you decide how to shape the data, what models to fit, and what you think can be learned and acted upon
- Project will be done during the last ≈ 2 weeks of class – more information and specific dates will be provided in class and on canvas

R Software

- We will make extensive use of the R programming language
  - You should have some experience with R as a prerequisite, or significant experience with another programming language
- Please take time early in the course to get acquainted with the environment
- In addition to the regular homeworks, there is a "Homework 0" that involves some basic programming tasks. This will not be graded – use this as a benchmark for your own skills.

R Software

- **Warning:** Again, we will not teach R in this class – you are expected to either be or quickly become comfortable with programming basics. Students with limited experience are *strongly* encouraged to take the course at a later date
- You are welcome to ask myself and the TA about specific R-related issues as they arise, but you should not rely on the TA to help write *or debug* your code
- If you do need additional R-related help, please utilize the stat lab and open office hours through the library (see syllabus for details)
- Please begin Homework 0 and judge for yourself whether you are prepared for the course. I am happy to discuss this aspect with you during office hours if you're feeling uncertain.

Where to go?

- There are 3 distinct roles in running this class: instructing (me), assisting (head TA - Ryan), grading (TA/grader - Alex)
- Think of the head TA (Ryan) as your primary go-to for the vast majority of issues
- The primary role of the TA/grader (Alex) is grading – you should really only need to contact him for questions related to how something was graded
- For anything more personal or "big picture", you can come to me (and you're certainly always welcome to come to office hours for whatever reason)

Some Disclaimers

- Course material will usually be presented via slides and recorded zoom lectures when university is virtual. The course slides are not meant to be a substitute for notes
- Slides will generally be made available before being discussed in class, but you will likely want to take additional notes
- The textbook is **exceptionally well written**. It is expected that you will read the relevant material in advance of class.

Copyright

Copyright Notice

Course materials may be protected by copyright. United States copyright law, 17 USC section 101, et seq., in addition to University policy and procedures, prohibit unauthorized duplication or retransmission of course materials. See Library of Congress Copyright Office and the University Copyright Policy.

Additional Course Materials

- We are using a very popular book and as such there are many additional resources at your disposal:
  - Book Homepage where you can find code, figures, errors, etc.
  - Slides and Videos by Rob Tibshirani and Trevor Hastie for their statistical learning MOOC

Final Notes

- All of this material is available in the syllabus on canvas
- Most course material (homeworks, due dates etc.) is already (or will soon be) posted on canvas – you are responsible for going through this material on your own in addition to what is mentioned in class

Final Notes

- This course should be extremely useful, especially for anyone not previously exposed to ideas in data science and statistical/machine learning
- We'll stress *breadth* over *depth* – cover a lot of different ideas and methods around a unified theme
- It **will involve a lot of work**. There are no separate lab/recitation sections – you will be responsible for picking up any material you need. Please get help early and often as needed.

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

- *A term introduced by industry to justify hiring non-statisticians to do statistics.*

  -Lucas Mentch, 2015

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.

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

Who Cares?

Even though statistics fastest growing STEM degree ...

- Indeed Job Search for Pittsburgh, PA:
  - "Statistician" – 10 jobs found (2023: 14, 2022: 14, 2021: 14, 2020: 14, 2019: 3)
  - "Data Scientist" – 84 jobs found (2023: 114, 2022: 271, 2021: 100, 2020: 140, 2019: 199) but ..
  - Jobs using "Statistics" – 248 jobs found (2023: 303, 2022: 602, 2021: 267, 2020: 644, 2019: 691)
- Glassdoor Best Jobs in America:
  - 2016 – 2019: # 1: Data Scientist
  - 2020: # 3: Data Scientist
  - 2021: # 2: Data Scientist
  - 2022-23: # 3: Data Scientist

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

Where does that leave us?

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

- 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?
  - Excellent Forbes article traces the history of the term

- Data Science arose for two primary reasons:
  1. Statisticians (and others in related fields) can be stubborn
     - *"Statistics has a bad PR department"*
  2. The rapid growth in storage and accessibility of data has lead (and continues to lead) to changes in scientific thinking and practice

The Tale of Leo Breiman

1. Statisticians (and others in related fields) can be stubborn

Leo Breiman, a founding father of Machine Learning:

- 1954: PhD in Mathematics from UC Berkeley
- 1961: Leaves tenured faculty position at UCLA to work as private consultant
- 1980: Joins Statistics Faculty at UC Berkeley
- 1993: "Retires" (much of his best work is done after retirement)

The Tale of Leo Breiman

- Upon rejoining academia in a statistics department, Breiman quickly realizes that the problems statisticians seem most interested in often do not correspond to real-world "cutting-edge" problems
  - Predicting next-day ozone levels.
  - Using mass spectra to identify halogen-containing compounds.
  - Predicting the class of a ship from high altitude radar returns.
  - Using sonar returns to predict the class of a submarine.
  - On-line prediction of the cause of a freeway traffic breakdown.
  - Speech recognition
  - The sources of delay in criminal trials in state court systems.

– Examples from "Statistical Modeling: The Two Cultures"

Similar example: predicting type of warship based on aerial photograph:

https://en.wikipedia.org/wiki/North_Carolina-class_battleship

https://en.wikipedia.org/wiki/French_battleship_Richelieu

- In grayscale, each image becomes 16×16 grid of pixel values from 0 to 255:

$$\text{Ship Type} = f(X_1, \ldots, X_{256}) + \epsilon$$

  - How do we begin to define an appropriate "model" for f?

Takeaway

- Reluctance on the part of statisticians to move towards this new way of thinking about using data to solve problems
  - Still quite prevalent today, though slowly dying off
- Funny thing: problems don't go away simply because a given field "doesn't do that"
- Stubbornness + natural aversion to new technology + necessity of significant computational skills makes room for a new field
  - Faced with "expand or die" most statisticians were quite content to stay put
  - Still an ongoing debate as to what should be considered statistics

The Other Half of the Story

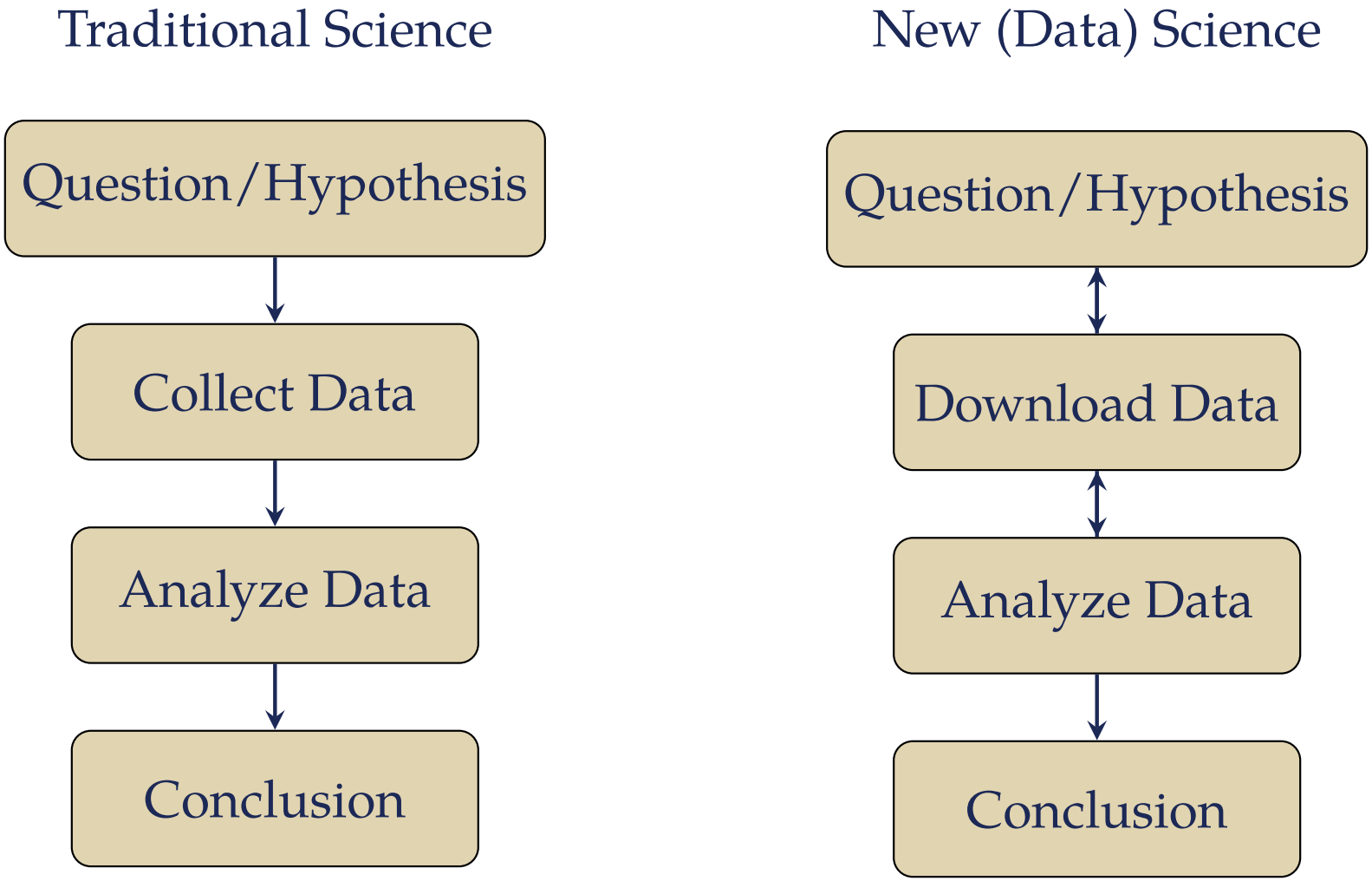
- This is still only one side of the coin
  - Gives a feel for how data science grew outside of statistics, but still doesn't explain its tremendous popularity

2. 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 of computers and internet:
  - **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 Science in Data Science

- Shift away from analyzing small datasets from controlled experiments to trying to make sense of big, messy datasets
  - "Big data" movement

The Good, the Bad, and the Ugly

The Good

- 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
  - Led 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 *

The Bad

- The move towards more sophisticated models has not necessarily lead to more advanced analytical training in science
  - "Blind faith" and misuse/misinterpretation of new methodologies and results
  - More statisticians getting involved, but still a strong sense of "we can do this on our own"
- Significant confusion in both academia and industry as to what exactly data science is and what data scientists should do
  - A statisticians view of data science may vary considerably from someone in information or computer science
  - Should you be a computer scientist with a significant background in math/statistics or a statistician with a strong background in programming and databases?

The Ugly

- Formal data science degrees, training, institutes etc. are beginning to evolve, but often very little quality control and no agreement as to what these programs should entail
  - If you were hiring a statistician (CS, Math, etc.), you would expect they have a significant background in statistics (CS, Math, etc.), but "anyone can be a data scientist"

- High demand (and growing) = Huge rush to cash in on the latest buzz word
  - "Data Scientist: The Sexiest Job of the 21^{st} century"

    – Harvard Business Review
  - "Become a data scientist in 12 weeks"

    – NYC Data Science Academy

The Ugly

reddit

Search

17

Become a data science wizard in just 1 month! You'll learn all the right skills for success in data science from the basics (stats & math) to the highly advanced and most in-demand (python & machine learning) in one super affordable course.