STAT 1291: Data Science

Lecture 1 - What is Data Science?

Sungkyu Jung

STAT 1291

- Topics in Applied Statistics
- "Basics of Data Science"
- Meant to be "STAT 1261: Principles of Data Science"
- Bookmark your course webpage: http://www.stat.pitt.edu/sungkyu/course/pds/

Today

- What is Data Science?

Data Science

20th Century Innovation

Engineering and Computer Science played key role

- nuclear power
- airplanes & automobiles
- the digital computer
- radio
- internet
- imaging

(https://dataorigami.net/blogs/napkin-folding/17543555-datas-use-in-the-21st-century)

But how about these 20th Century questions?

- Does fertilizer increase crop yields?
- Does Streptomycin cure Tuberculosis?
- Does smoking cause lung-cancer?

What is the difference?

- Deterministic versus random
- Deductive versus empirical
- Solutions deduced mostly from theory versus solutions deduced from mostly from **data**

Data

- Does fertilizer increase crop yields? Answer: Collect and analyze agricultural experimental data
- Does Streptomycin cure Tuberculosis? Collect and analyze randomized trials data
- Does smoking cause lung-cancer? Collect and analyze observational studies data

But

That's what statisticians are already doing.

21st century

Figure 1:

21st century

- NYT (http://www.nytimes.com/2009/08/06/technology/06stats.html?mcubz=0)

"I keep saying that the sexy job in the next 10 years will be statisticians," said Hal Varian, chief economist at Google. "And I'm not kidding."

Hal Varian says

"The ability to take **data** - to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data."

- This is a good definition of data science

Data Science and Statistics

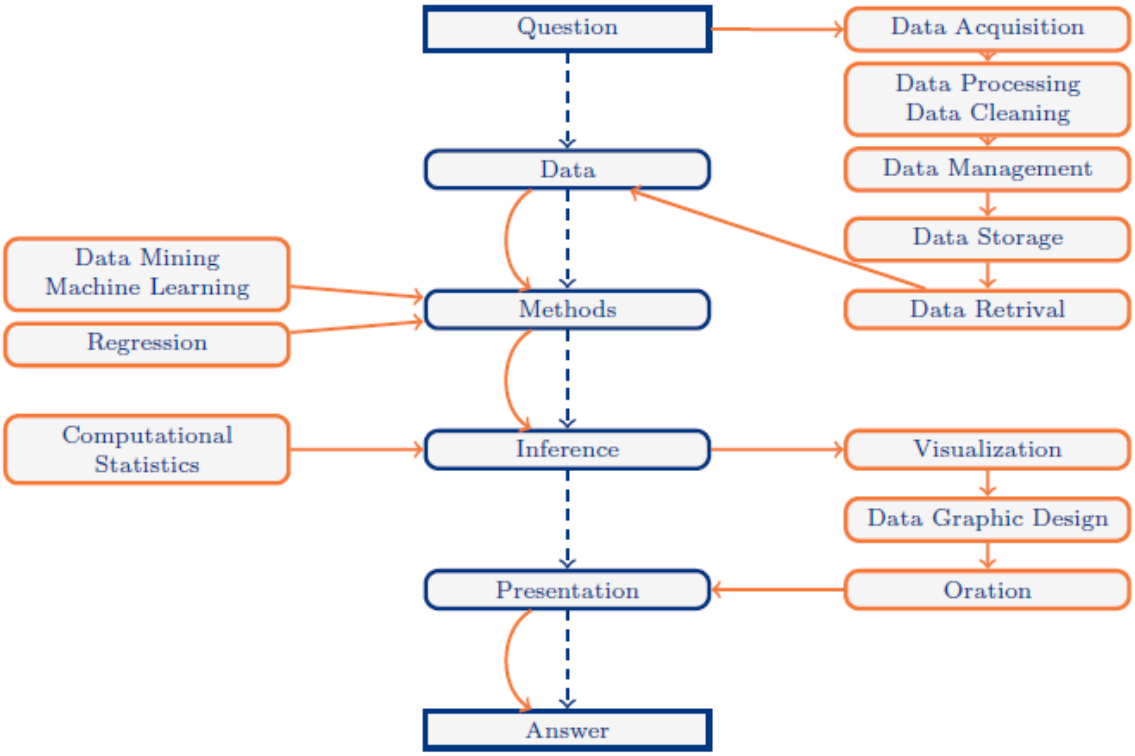

Figure 2:

- Schematic of the modern statistical analysis process
- We focus on the bubbles to the left and right

Typical data science project

- I chose a research project (in sociology)
- There are industrial/business projects; see e.g. https://www.datascienceweekly.org/articles/aspiring-data-scientist-here-are-some-at-work-project-ideas or https://www.analyticsvidhya.com/blog/2014/11/data-science-projects-learn/

More Tweets, More Votes

- Here is a recent example of data science.
- Rojas and his colleagues pose the question:

Is social media a valid indicator of political behavior?

- And answer this question using a random sample of 537,231,508 tweets from August 1 to November 1, 2010 and data from 406 competitive U.S. congressional elections provided by the Federal Election Commission.

Imagine that you were asked to answer

- The data being analyzed were scraped from the Internet
- The research question was addressed by combining domain knowledge
- A large amount of data
- Need experts in sociology and in data science

Exercise

- Read the draft of the paper
- Pair up
- Critically review the paper

(https://ssrn.com/abstract=2235423)

(http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0079449)

Final question:

How would you reproduce this study?