A First Course in Data Science

Donghui Yan

University of Massachusetts Dartmouth

January 6, 2021

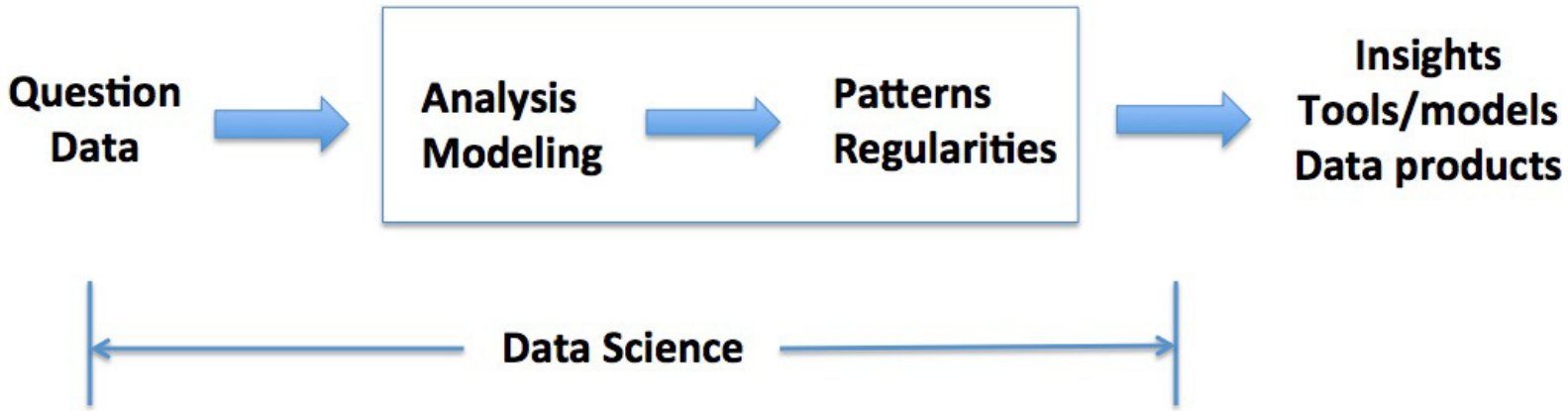
Outline

- Introduction
- Course design based on data science life cycle
- Teaching in practice

What is data science?

- Data Science is the science of data for analysis

  A discipline that provides principles, methodology, or guidelines for the analysis of data for tools, values, or insights.



An example - item recommendation

- Large number of user accesses to a typical e-commerce site
  - e.g., tens of millions of user access at walmart.com each day
  - Every mouse click captured by e-commerce server
    - e.g., view, cart, purchase of an item
  - ♠ Huge data of sales records, how to leverage those?

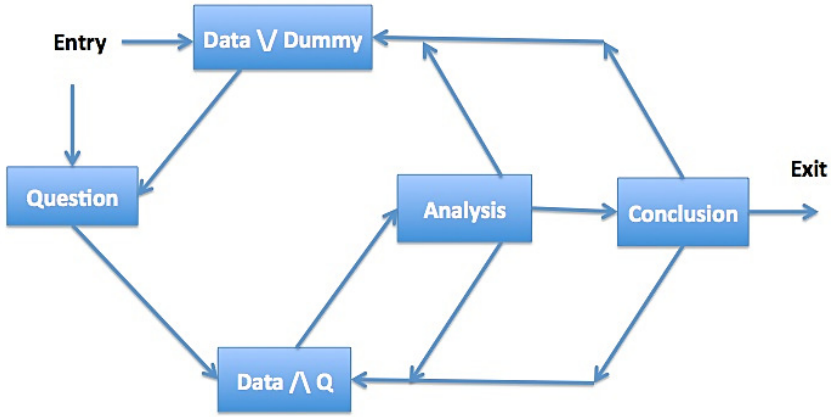
- *Observation*: users typically buy items together
  - e.g., items $B_1, ..., B_t$ bought in same transaction
  - Items $B_1, ..., B_t$ called co-bought items
  - If a user buys an item, he tends to buy co-bought items as well.

Item recommendation (continued)

- Co-bought stats can be used to build a recommendation model
  - Users buying/viewing product A may also like B
  - An effective way to promote sales
    - e.g., *Walmart* boosts its sales by 5-10%

A model for data science practice

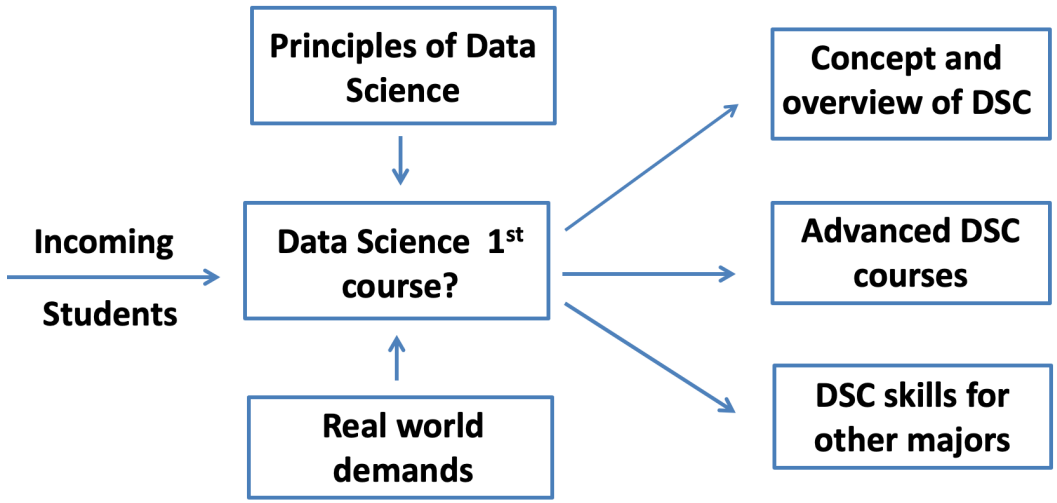
- The data science life cycle
  - State or analytical tasks driven



- Alternative models
  - Data flow and operations on data: Schutt and O'Neal (2013), Wickham and Grolemund (2016)
  - Implementation tasks dependency: Guo (2012).

Design methodology - a structured approach

- What are the inputs?
  - Incoming students, mostly *freshmen*
  - Principles of data science
  - Demands/requirements from advanced courses, industry etc
- Our education goal (outputs)
  - Concepts and overview of data science
  - Prepare for advanced courses
  - Positively enable students and inspire their interests.



Topics in our course

- Center around the data science life cycle
  - Coherent body of knowledge
- Concept of data science life cycle with examples
- Ask or derive interesting questions from data
- Data collection
  - Various potential bias in data collection
  - Random sampling
- Exploratory data analysis
  - Summary statistics
  - Data visualization
  - Data cleaning, transformation and feature engineering
  - Clustering (hierarchical and k-means)

Topics in our course (continued)

- Linear regression
  - Data visualization aspect and modeling technique.
- Hypothesis testing
  - Framework for data-driven confirmatory analysis.
- R programming
  - Concept of programming
  - Basic data structures in R programming
  - Structured programming constructs
  - Functions
  - Dealing with data input/ouput.

Philosophy in developing course materials

- Focus on concepts, ideas, and culture
- Use of *real* examples, stories and applications
- Emphasis on exploratory data analysis
  - Freshman course no calculus required
  - Data visualization tools, ideas, and examples
  - Visualization aspect of tools or methods (e.g., hierarchical clustering, regression)
- Real world data for *authentic* data experience.

Other course components

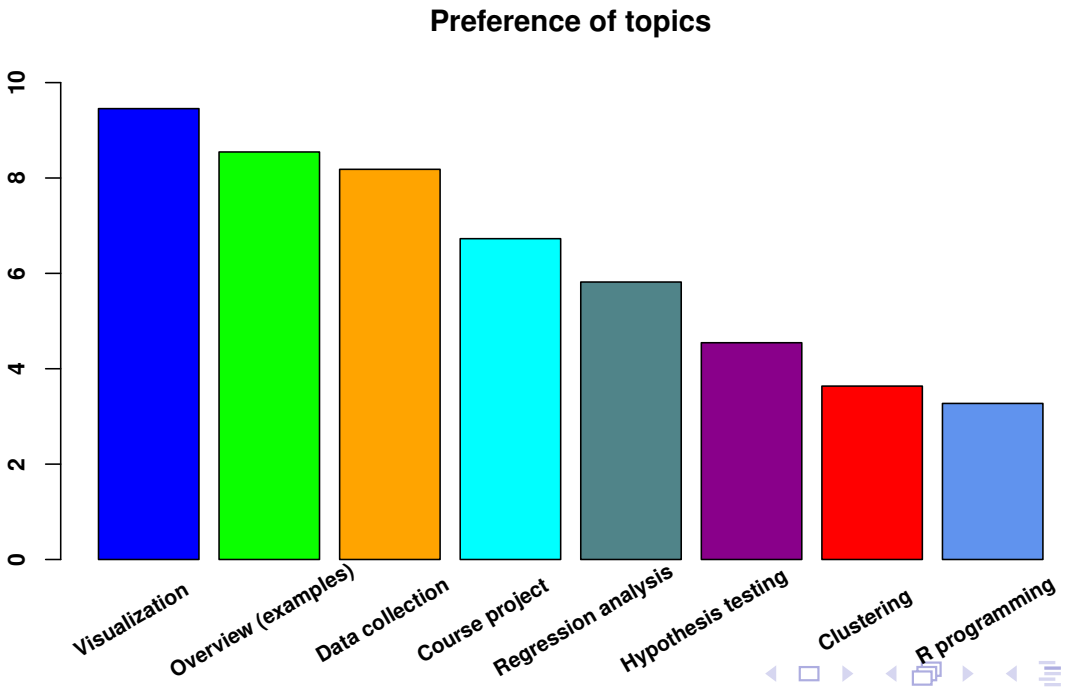
- Homework or labs
  - Readings on data science articles, examples of biased data collections in news, media etc
  - Data visualization
  - Clustering
  - Linear regression
  - Hypothesis testing
- Course project
  - On either one of data visualization, regression or clustering
  - Project presentation.

Sample project topics

- Global terrorism trend analysis
- Analysis of used car sales prices
- Daily counts of Covid-19 cases and the Benford's law
- Analysis on effects of location on earthquake size and depth
- An exploratory analysis of US serial killers
- On the economics of US electric semi-trucks.

Summary

- Course launched in Fall 2015 at UMass Dartmouth
- Positive experience from many majors on campus
  - Actively attracted students to data science major or minor
  - Inspire some students to work on their own DSC projects
  - Motivate further study in subject to understand analysis results.



Challenges in teaching

- Highly diverse student body
  - Non- to highly quantitative majors
    - Data Science, Mathematics, Computer Science, Engineering, Biology, English, Psychology, Political Science, Accounting, Economics, Business, Crime Justice, Visualization and Performance Arts etc
  - Different levels of preparation
    - Some already done projects in high school while few others not even computer literate
- Topics challenging for some students
  - R programming (skill)
  - Clustering (technique)
  - Hypothesis testing (concept).

Summary

- Data science life cycle based course design
- Real examples, real stories, and real data are *very* important
- Teaching experience at UMass Dartmouth.

The end

Thank you!

References

1. P. Guo (2012). Software Tools to Facilitate Research Programming. *Ph.D. Dissertation, Stanford University.*
2. C. O'Neil and R. Schutt (2013). Doing Data Science: Straight Talk From the Frontline, *Sebastopol, CA: O'Reilly Media.*
3. W. Wickham and G. Grolemund (2016). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data, *Sebastopol, CA: O'Reilly Media.*
4. D. Yan and G. Davis (2019). A first course in data science. *Journal of Statistics Education*, 27(2), 99-109.