540: Statistics Case Studies
Use Data to Tell a Story

Fan Li

Department of Statistical Science

Duke University

Fall 2021

Data science and scientist

- Foundation of data science
  - Data engineering
  - Software engineering
  - Machine learning
  - Statistics
- A good data scientist is a complete package:
  - Statistical thinking and knowledge
  - Substantive knowledge
  - Computing
  - Communication (oral and writing)

Data science skill sets

- How to ask the right question given data
  - The most serious mistakes are not being made as a result of wrong answers. **The true dangerous thing is asking the wrong question.**" – *Peter Drucker*
  - Intellectual curiosity, experience, common sense, substantive knowledge
- How to think about data vs. problem
  - Math/Stat/Machine learning
- How to handle data
  - Modern computing and programing tool: Python, R, Java, Hadoop, Tableau…
- Teamwork and collaboration skills – how to work with others
- How to convince others about your results
  - Visualization and story telling
  - Communication: oral and written

Good and bad traits of a data scientist

- Good traits
  - Be open-minded: Bayesian or not? Nobody cares! Ultimate goal is to solve problems, whatever suits the best
  - Be creative, think out of the box
- Bad traits (how to fail)
  - Focusing only on the solution
  - Forgetting the basics
  - Ineffectively communicating
- More at https://www.techrepublic.com/article/how-to-failas-a-data-scientist-3-common-mistakes/

Use data to tell a story: a spectacular example

- Hans Rosling's famous lectures combine enormous quantities of public data with a sport's commentator's style to reveal the story of the world's past, present and future development.
- In this spectacular section of 'The Joy of Stats' he tells the story of the world in 200 countries over 200 years using 120,000 numbers - in just four minutes.
- Plotting life expectancy against income for every country since 1810, Hans shows how the world we live in is radically different from the world most of us imagine.

https://www.youtube.com/watch?v=jbkSRLYSojo

Use data to tell a story: another spectacular example

- **Watch the Growth of Walmart and Sam's Club**
- Walmart (blue) started slow in 1962 and then spread like wildfire in the southeast, starting in 1970, and then made its way towards the west coast. Sam's Club starts to sprout up in the 1980s with bursts up to present.

  http://projects.flowingdata.com/walmart/

Use data to tell a story: a disastrous example

- One should never fabricate data under any circumstance
- Data fabrication is hard to catch; even fabrication takes some thinking
- Evidence of Fraud in an Influential Field Experiment About Dishonesty

https://datacolada.org/98

  - Data on customer self-reported odometer reading of cars covered by their insurance policy.
  - Data distribution: how are such data distributed? Normal or uniform?
  - Rounding: Some people would round odometer reading, e.g. rounding 43537 to 43500

- Let's look at the data in the PNAS (2012) paper: just use common sense

How this course can help

- No formal instruction on statistics/machine learning topics (a few exceptions, e.g. causal inference)
- Project-based course. Learning by doing
  - Problem identification via teamwork and discussion
  - Problem solving by using existing skills or new skills, learn new things “on the job”, and learn from your peers
  - Present your codes, your results and your story (try to sell them)
  - There will be many things I cannot answer but let’s learn together
- Get hands dirty with data
- Gain experience and good habits

Course structure

- I will design open-ended challenges, each of which focuses on a slightly different area in data science
- In each challenge,
  - Start with information/knowledge we already have about the problem
  - Identify knowledge/skills we need to solve the problem
  - Articulate the above thinking process in a team and implement an inquiry as a team
  - Present the process in class and write a report
- I will provide case studies and some tutorials to provide guidance on aspects of the above processes

Group projects: working as a team

- You don’t have to be in the same room at the same time to work together
- Here are several ways you will work together in this course
  - Face-to-face brainstorm
  - Online discussion in group forum
  - Zoom video chat with screen share
  - **GitHub collaboration**

Project assignment

- Teacher assigns students into random groups.
- Each group should have a separate repository to work on (private to group members and teaching staff before due).
- Create new branches for temporary work and merge to master when you feel there are no errors.
- The Project name and membership can be managed later but the most important part is we get all the teams/groups set up automatically.
- Everyone from your team should install Git and know how to collaborate with version control through terminal, GitHub Desktop or Rstudio.
- **Reproducible data science:** use rmarkdown and “knitr” package to create reproducible HTML or PDF reports.

Some good resources

- Open case studies at Johns Hopkins Bloomberg School of Public Health
  - https://americanhealth.jhu.edu/open-case-studies
  - Several excellent data science examples on health studies; detailed information on data cleaning, wrangling, visualization, analysis, and related substantive knowledge; stat methods are relatively easy
- The “Applied data science” course at Columbia Statistics:

  http://tzstatsads.github.io/
  - Excellent case studies and tutorials on a variety of topics

Project 1: Our world with Covid-19

- We have been living in this Covid-19 pandemic for the last 1.5 years
- Covid-19 deeply influences every aspect of the human society and beyond
- Abundant Covid data online, e.g.
  - Johns Hopkins University Covid Github (aggregating many government data): https://github.com/govex/COVID-19
  - New York Times data and map https://www.nytimes.com/interactive/2021/world/covid-cases.html
  - Covid-19 R repository (by Mine Cetinkaya-Rundel) https://github.com/mine-cetinkaya-rundel/covid19-r
  - Most U.S. states health department has a dashboard on Covid cases, hospitalization, vaccination, etc.

Project 1: Assignment

- Assignment:
  - Each team comes up with a concrete problem related to any aspect of Covid-19
  - Collect your own data to investigate the problem.
  - Create an R shiny app to visualize the data and present it in class to discuss your main findings
  - Write a group report: the problem, the data acquiring and cleaning process, statistical analysis, main findings, limitations
- Some criteria:
  - Some statistical analysis, even simple, is necessary
  - A simply dashboard is not enough
  - Spatio-temporal trend is desirable, but not required

Project 1: Example ideas

1. How is the vaccination rate associated with 2020 U.S. presidential voting pattern, at both state and county level?
   - An overall correlation is useful, as well as a zoom-in analysis of the top and bottom counties/states.
   - Most public data repository only provides state level data, but for this topic, county level information is more important because of the huge county level heterogeneity
2. Spatio-temporal relationship between Covid cases and some important aspects of the society, e.g. crime rate; traffic accidents; number of new born; labor cost
   - This requires find additional data beside Covid-19

Project 1: Example ideas

3. Temporal trend of commodity prices during the Covid-19 pandemic, e.g. real estates, cars, labor
    - This can be limited to a specific area/city or nation wide
4. Temporal trend of the relationship between vaccination rate and number of cases and/or hospitalization numbers, at state or county or both level
5. Covid-19 and air pollution: how is Covid-19 associated with CO2 emissions by country?
6. Covid-19 and BMI: everyone gains weight? Is there any data supporting that?

Cautionary note: association does not imply causation. Be careful with interpretation.

Plan of the next class

- Each group present their main idea and preliminary results, discuss challenges
- Funda Gunes will give a 30 minute presentation with Q&A on a COVID project
- Some tutorial on presentation and writing

Acknowledgements

- I borrowed several slides from Professor Tian Zheng’s lecture notes on the Applied Data Science course at Columbia University.