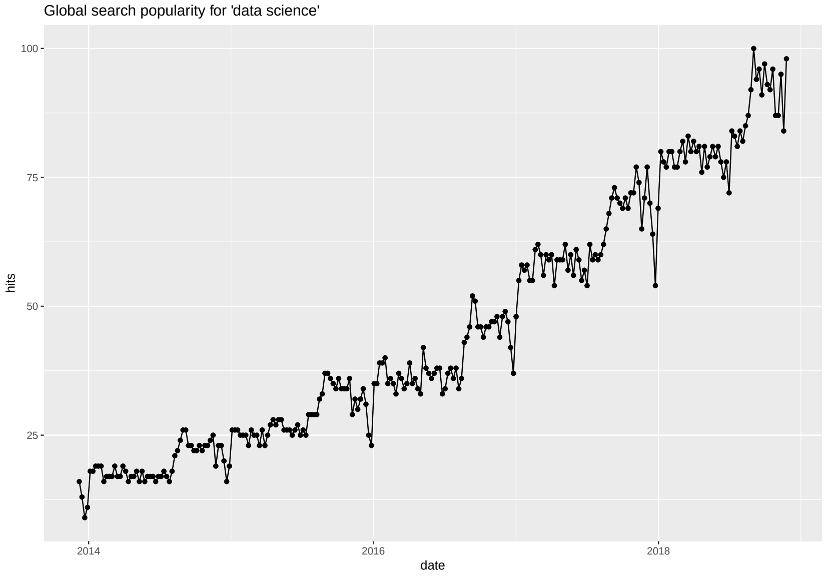
Introduction to Computational Tools and Techniques in Social Science

Jae Yeon Kim

02 December, 2018

Motivation


Global search popularity for 'data science'
hits
100
75
50
25
2014
2016
2018
date

- Why should we care?
- Yes, big data is a trend.
- But being good at computational tools and techniques has more immediate benefits.
- It can make your life **easy** and **organized**.
  - Don't repeat yourself: AUTOMATE!

- In addiiton, there are new tools for
    - Data collection (e.g., APIs, webscraping)
    - Analysis (e.g., machine learning)
    - Visualization (e.g., maps, social networks)
- In sum, you can do cool stuff.

- But it takes some **efforts** to take advantages of these new tools.
    - You need to learn how to code **a little bit**.
    - However, leanring on your own is inefficient.
    - More important, you can get **bad** habits.

- The following examples are adapted from https://style.tidyverse.org

```
# Good
fit_models.R

if (y < 0 && debug) {
  message("y is negative")
}

# Bad
fit models.R

if (y < 0 && debug)
message("Y is negative")
```

```
# Good
do_something_very_complicated(
  something = "that",
  requires = many,
  arguments = "some of which may be long"
)

# Bad
do_something_very_complicated("that", requires, many, argu
                              "some of which may be long"
                              )
```

Objectives

- Tasting a wide range of computational tools
- Getting programming fundamentals right
    - Concepts
    - Techniques
- Learning by doing
    - Learning from your own trials and erros
    - Learning from others

- Coding is similar to **cooking**.
    - So many different cuisines (programming languages).
    - But there are fudamentals.
        - Ingredients (data)
        - Techniques (logic)
        - Recipes (workflow)

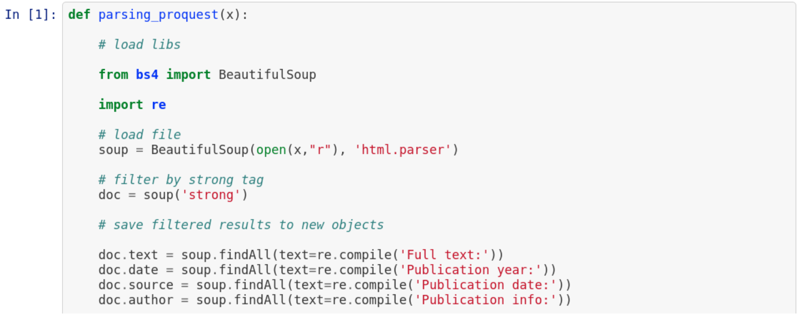
- Bad habits are **bad**.
    - Rule 1. Thou shall comment.
    - Rule 2. Thou shall reuse functions (no copy and paste).
    - Rule 3. Thou shall practice version control (no final_final_final.Rmd)

- **Learn to learn**
- Specifically, we are going to learn:
  - Navigate and operate effectively in a UNIX environment
  - Master basic Git and Github workflows
  - Write, execute, and debug R code for data cleaning, statistical analysis, data visualization and machine learning
  - Parse HTML, CSS, and Javascript for the purposes of using tools like APIs, webscraping, and Qualtrics
  - Write, execute, and debug R/Python code for text analysis, as well as other computing tasks

- **Don’t expect** you become:
    - A software programmer (we covere only a tip of the iceberg)
    - Get all the answers you need
- We focus on learning **how to learn**.
    - Programming is one endless Google Search (aka “Rochelle’s Law”)

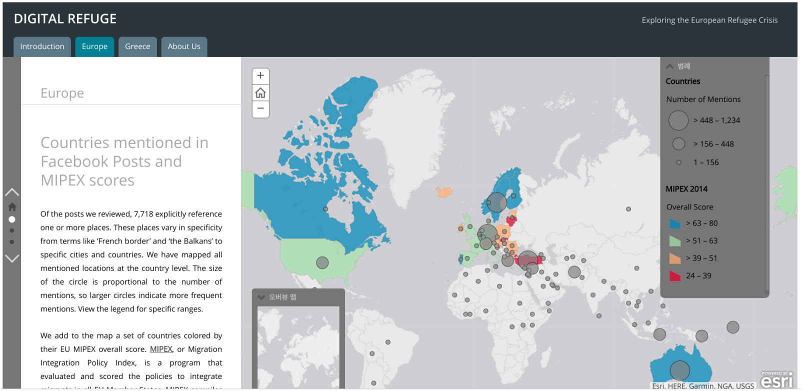
- Using Excel:
  - 3 mins for copying, pasting, and reorganizing one article
  - 80,000 newspaper articles
  - Taking **4,000** hours or **166 days**

- Using python:
  - A few hours for coding
  - Less than 5 mins for creating the dataset
  - Also, the code is reusable.



```
In [1]: def parsing_proquest(x):

            # load libs

            from bs4 import BeautifulSoup

            import re

            # load file
            soup = BeautifulSoup(open(x,"r"), 'html.parser')

            # filter by strong tag
            doc = soup('strong')

            # save filtered results to new objects

            doc.text = soup.findAll(text=re.compile('Full text:'))
            doc.date = soup.findAll(text=re.compile('Publication year:'))
            doc.source = soup.findAll(text=re.compile('Publication date:'))
            doc.author = soup.findAll(text=re.compile('Publication info:'))
```

Previous final projects by students



Europe

Countries mentioned in Facebook Posts and MIPEX scores

Of the posts we reviewed, 7,718 explicitly reference one or more places. These places vary in specificity from terms like 'French border' and 'the Balkans' to specific cities and countries. We have mapped all mentioned locations at the country level. The size of the circle is proportional to the number of mentions, so larger circles indicate more frequent mentions. View the legend for specific ranges.

We add to the map a set of countries colored by their EU MIPEX overall score. MIPEX, or Migration Integration Policy Index, is a program that evaluated and scored the policies to integrate

Focus on best practice

- Good habits are **good**.
  - Commenting serves you and many other people.
  - Reusing functions provides opportunities to learn and clean up your mess.
  - Practicing version control is how we become a mature researcher and a coder.

Class

- Participation (25%)
  - Be kind and nice to each other. We're all leanring (especially me).
- Homework (50%)
  - Every week.
  - Learning how to code is like learning how to drive.
- Final project (25%)
  - Feasibility is your friend. Late Feb proposal, April presentations.

Logistics

- Learning by doing
- Pair-programming on in-class challenges
- Section is required.
- Julia Christensen is a technical assistant to the course.

Special thanks

- Rochelle Terman (University of Chicago)
- Rachel Bernhard (University of Oxford, UC Davis)