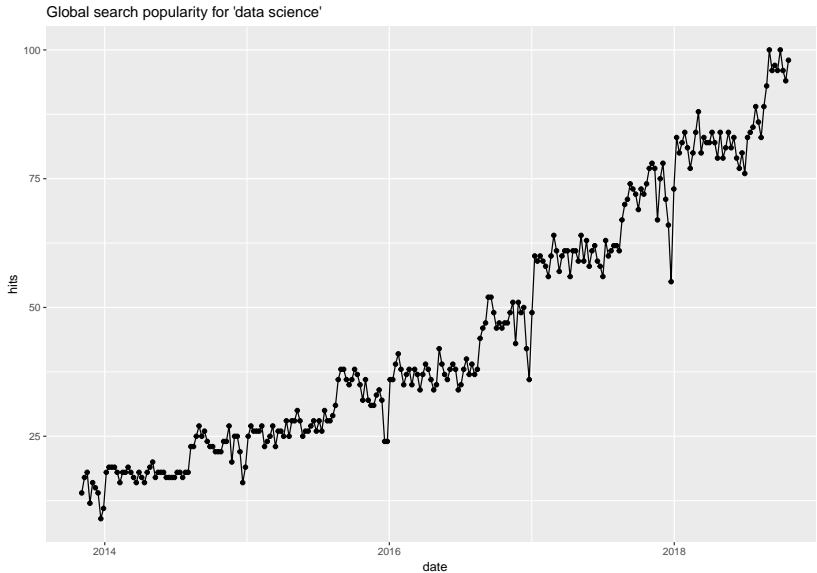


# Introduction to Computational Tools and Techniques in Social Science

Jae Yeon Kim

27 October, 2018

# Motivation



- ▶ Why we should we care?
- ▶ Big data is only a piece.
- ▶ New tools for
  - ▶ Data collection (e.g., webscaping)
  - ▶ Analysis (e.g., machine learning)
  - ▶ Visualization (e.g., maps, social networks)
- ▶ Diverse tools help us answer more interesting questions.

- ▶ You need to learn how to code *a little bit*.
- ▶ Learning on your own is inefficient.
- ▶ More important, you can get bad habits.

# Objectives

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



**Hadley Wickham** ✓

@hadleywickham

Follow



The only way to write good code is to write tons of shitty code first. Feeling shame about bad code stops you from getting to good code

7:11 AM - 17 Apr 2015

909 Retweets 1,005 Likes



41



909



1.0K

Figure 1: Hadley Wickham quote

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

- ▶ Bad habits are **bad**.
  - ▶ No commenting is a sin.
  - ▶ Not reusing your functions is a waste.
  - ▶ Organize, organize, and organize (R project, version control).



- ▶ **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
- ▶ Find answers to things we won't cover in this class

- ▶ **Don't expect** you become in the end of the semester:
  - ▶ A software programmer (we cover only a tip of the iceberg)
  - ▶ Get all the answers you need
  - ▶ We focus on learning **how to learn**

# Practical examples

## Private Lawyers Give Services to Agency

**Publication info:** Oakland Post (1968-1981) ; Oakland, Calif. [Oakland, Calif]11 Sep 1968: 22.

[ProQuest document link](#)

### Abstract:

Joseph E. Smith, president of the Alameda County Bar Ass'n, and former mayor of Oakland, was the first to serve. He was in the West Oakland Service Center, 1330 Chestnut St., Aug. 27th. The other centers are: North Oakland, 905 - 55th St.; East Oakland, 8924 Holly St., Fruitvale, 1470 Fruitvale Ave., Hayward, 22531 Watkins St. and Livermore, 2222 Second St.

**Links:** [UC-eLinks](#)

### Full text:

Alameda County lawyers are volunteering their services to the Legal Aid Society, the Society announced today. More than 75 local lawyers have responded to the Society's appeal to spend one evening per month at a neighborhood service center or contribute an equivalent amount of time in their offices.

Joseph E. Smith, president of the Alameda County Bar Ass'n, and former mayor of Oakland, was the first to serve. He was in the West Oakland Service Center, 1330 Chestnut St., Aug. 27th. The other centers are: North Oakland, 905 - 55th St.; East Oakland, 8924 Holly St., Fruitvale, 1470 Fruitvale Ave., Hayward, 22531 Watkins St. and Livermore, 2222 Second St.

Purpose of the new program is to improve the service now rendered by Legal Aid lawyers, as well as to give lawyers in private practice an opportunity to meet members of the poverty community in a professional relationship.

In addition, the time contributed helps fulfill federal requirements that 20 percent of the Society's budget be contributed from local sources. This year the federal contribution amounts to nearly \$300,000 and nearly \$70,000 must be raised locally.

The Legal Aid Society provides free legal service in non-criminal cases to persons who cannot afford to pay a lawyer. The Society has been instrumental in improving the administration of welfare laws, in protecting consumers from abusive sales practices and collection practices, in assisting tenants of private and public housing facilities, and served more than 8,000 clients in Oakland. It has a full-time staff of 12 lawyers in Oakland.

**Ethnicity:** African American/Caribbean/African

**Publication title:** Oakland Post (1968-1981); Oakland, Calif.

**Volume:** 5

**Issue:** 19

**Pages:** 22

**Number of pages:** 0

**Publication year:** 1968

**Publication date:** Sep 11, 1968

**Publisher:** Alameda Publishing Corp.

**Place of publication:** Oakland, Calif.

**Country of publication:** United States, Oakland, Calif.

**Publication subject:** General Interest Periodicals--United States, African American/Caribbean/African

**Source type:** Newspapers

**Language of publication:** English

Figure 2: Turning html documents into a dataset

- ▶ Performing the task with Excel
  - ▶ 3 mins for copying, pasting, and reorganizing one article
  - ▶ 80,000 newspaper articles
  - ▶ Taking **4,000** hours or **166 days**

```

    for i in range(len(doc.author)):
        sum_author.append(doc.author[i].find_previous().find_previous().find_previous().find_previous().text) # some articles do not have identified authors; this is a way to get around that problem

    # check

    print(len(sum_text), len(sum_date), len(sum_source), len(sum_author))
    print(sum_date[i], sum_source[i])

    # combine the results as a list
    newspaper_list = {'text': sum_text, 'date': sum_date, 'source': sum_source, 'author':sum_author}

    # return

    import pandas as pd

    return(pd.DataFrame(newspaper_list))

```

```

In [ ]: # get working directory

import os

os.chdir('/home/jae/Documents/Text_analysis/Data/Asian')

```

```

In [ ]: # for loop over entire page results

n = 0

temp_dataset = []
for filename in os.listdir(os.getcwd()):
    if filename.endswith(".html"):
        n = n + 1
        print("file",n, filename)
        temp_dataset.append(parsing_proquest(filename))

```

```

file 1 w1t1page80.html
100 100 100 100
Publication date: Jun 27, 1986 Publisher: Asian Week
file 2 w1t1page26.html
100 100 100 100
Publication date: Aug 25, 1983 Publisher: Asian Week
file 3 w1t2page15.html
100 100 100 100
Publication date: Oct 2, 1987 Publisher: Asian Week
file 4 w1t1page9.html
100 100 100 100

```

# Final projects

# Obsctacles to learning

# Class

- ▶ Participation (25%)
  - ▶ Be kind and nice to each other. We're all learning (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)