# Computational Social Science

## Introduction

Dr. Thomas Davidson

Rutgers University

January 18, 2024

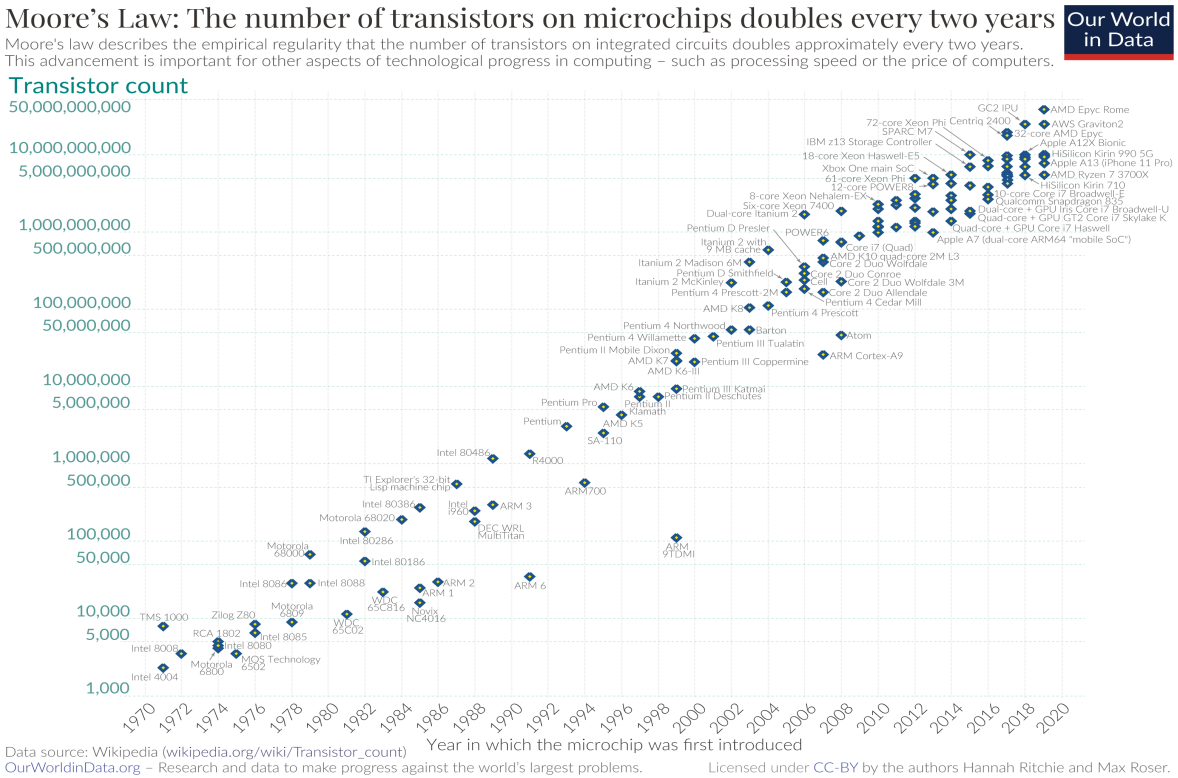
# Plan

- Introductions
- A brief introduction to Computational Social Science
- Course outline
- R and RStudio
- Resources
- Readings and set-up

# Introductions

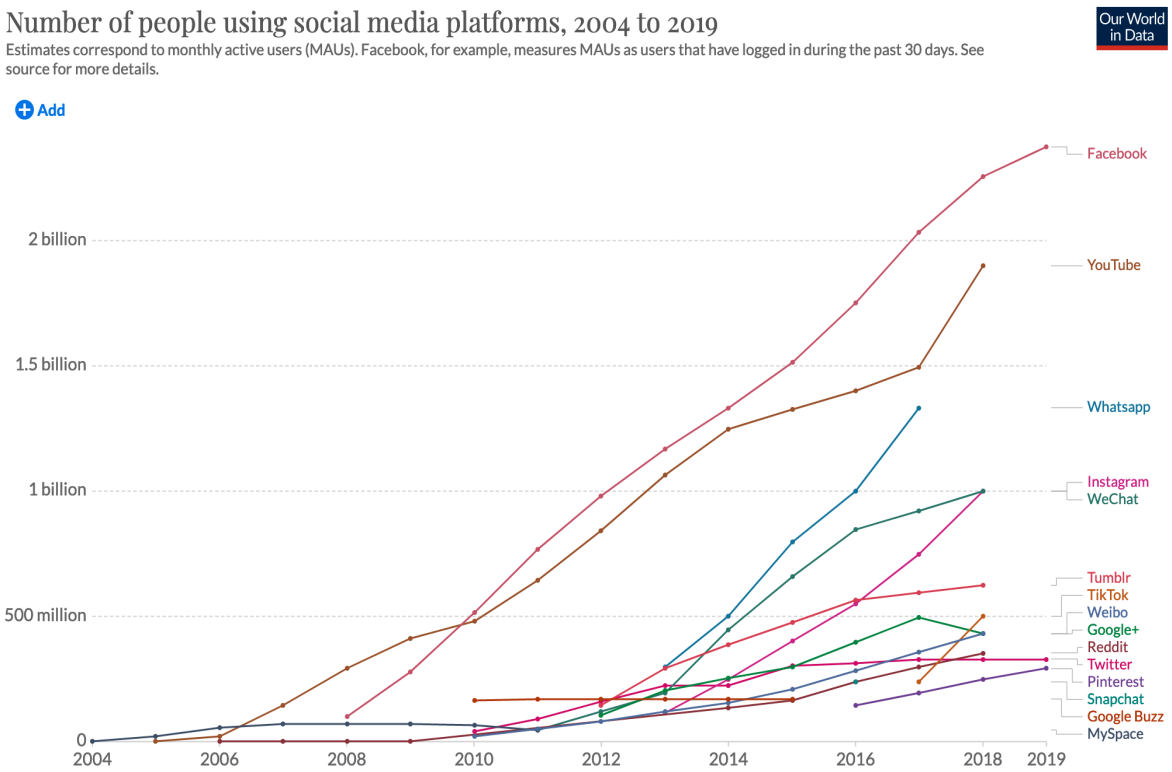
- Name
- Area of study (major, minors, etc.) and year
- Interests in data science / social science

# Introduction to Computational Social Science




Source: Wikipedia

# Introduction to Computational Social Science




Source: Our World in Data

# Introduction to Computational Social Science

## Data Science and Social Science

- Contemporary society is characterized by a proliferation of data related to social life and ubiquitous computation
- These data are being used by an array of different actors to make decisions and influence people's lives
  - Technology companies, government, business, finance, healthcare, insurance, non-profits, etc.

# Introduction to Computational Social Science

## Digital Traces and Big Data

> *[J]ust as the invention of the telescope revolutionized the study of the heavens, so too by rendering the unmeasurable measurable, the technological revolution in mobile, Web, and Internet communications has the potential to revolutionize our understanding of ourselves and how we interact . . . . [T]hree hundred years after Alexander Pope argued that the proper study of mankind should lie not in the heavens but in ourselves, we have finally found our telescope. Let the revolution begin.*
>
> —Duncan Watts (2011, p. 266)

Quoted in Golder and Macy 2014.

# Introduction to Computational Social Science

## Data Science and Social Science

- Traditional social science relies upon qualitative and quantitative methodologies that are insufficient alone to work with large, complex datasets
- Data scientists have access to more powerful methodological tools, but often have a limited understanding of social scientific principles or knowledge
- Computational Social Science combines the best of both approaches, bringing together social scientific and computational knowledge and methods

# Introduction to Computational Social Science

## Computational Social Science

> The capacity to collect and analyze massive amounts of data has unambiguously transformed such fields as biology and physics. The emergence of such a data-driven "computational social science" has been much slower, largely spearheaded by a few intrepid computer scientists, physicists, and social scientists. If one were to look at the leading disciplinary journals in economics, sociology, and political science, there would be minimal evidence of an emerging computational social science engaged in quantitative modeling of these new kinds of digital traces. However, computational social science is occurring, and on a large scale, in places like Google, Yahoo, and the National Security Agency. Computational social science could easily become the almost exclusive domain of private companies and government agencies. Alternatively, there might emerge a "Dead Sea Scrolls" model, with a privileged set of academic researchers sitting on private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest in the accumulation, verification, and dissemination of knowledge.

David Lazer and colleagues make the case for computational social science in their 2009 article in *Science*

# Introduction to Computational Social Science

## Data Science and Social Science

- Target audiences
  - Data scientists working with social data
    - Computer science and statistics majors, etc.
    - Software engineers, project managers, research scientists, etc.
  - Social scientists using data science
    - Sociology, political science, economics majors, etc.
    - Researcher analysts, teachers, civil servants, etc.

# Introduction to Computational Social Science

## The emergence of a field

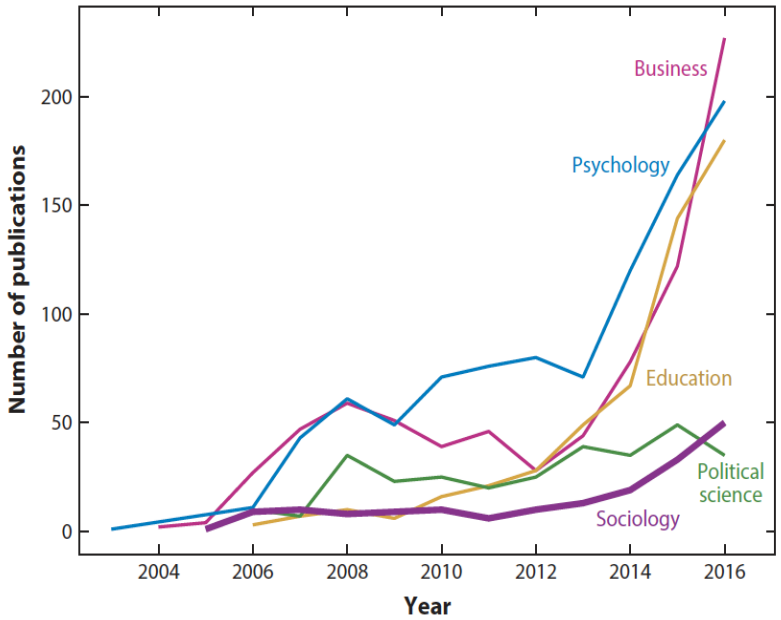



**Figure 1**

Number of computational social science publications by year—2003–2016—across four scholarly disciplines.

Source: Edelmann et al. 2020

# Introduction to Computational Social Science

## Towards a definition

1. Use of complex, multimodal data (digital records, image, text, video)
2. New digital modes of data collection (web-scraping, APIs, online experiments)
3. Application of methods developed by computer scientists to address social scientific questions (topic modeling, word embeddings, deep learning)
4. Combination of these methods with traditional social scientific approaches (hypothesis testing, theory-building)

# Introduction to Computational Social Science

## Preview of cutting-edge research

- What do word embeddings reveal about our understanding of culture?
- Can we use machine learning to predict important life events?
- How can we use Google StreetView to estimate demographic composition of neighborhoods?
- What can generative AI contribute to social science?

# Introduction to Computational Social Science

## Ethical considerations

- What is the legality of using data scraped from the web?
- Should we trust predictive modeling systems with consequential decisions?
- Why do hate speech detection algorithms discriminate against racial minorities?

# Course outline

## Goals

- By the end of this course you should be able to
  - Understand the field of Computational Social Science
  - Code using R at an intermediate level
  - Implement various computational methodologies for data collection and analysis
  - Think critically about the use of "big data" and computational methods to study social life
  - Build an app for data analysis and visualization

# Course outline

## Prerequisites

- Data 101 or equivalent programming experience
- Some basic statistical knowledge is helpful but not required

# Course outline

**Structure**

1. Programming in R (Weeks 1-3) and some background to CSS
2. Data collection (4-6)
3. Natural language processing (7-9)
4. Machine learning (10-13)
5. Generative AI (14)
6. Presentations (15)

# Readings

- This week's readings consist of the first chapter of *Bit by Bit* by Matthew Salganik and selected chapters from *R for Data Science (R4DS)* by Garrett Grolemund and Hadley Wickham.
- You will continue to read *R4DS* over the next few of weeks to build familiarity with R and the `tidyverse` packages.
  - *I recommend reviewing this material even if you are familiar with R.*

# Course outline

## Assessment

- Participation (10%)
- Homework assignments (60% total, 15% each)
  - Basics of data science in R
  - Online data collection
  - Natural language processing
  - Machine learning

# Course outline

## Assessment

- Final project (30%)
  - Build an app for data analysis and visualization using RShiny
- Four phases:
  1. Develop project ideas
  2. Submit project proposal
  3. Prototype
  4. Present final project and submit app
- *Projects can be completed individually or as part of a group*

# Course outline

## Policies

- Read the syllabus
  - Accommodations
  - Diversity and inclusion
  - Academic integrity
  - ChatGPT and AI (more shortly)

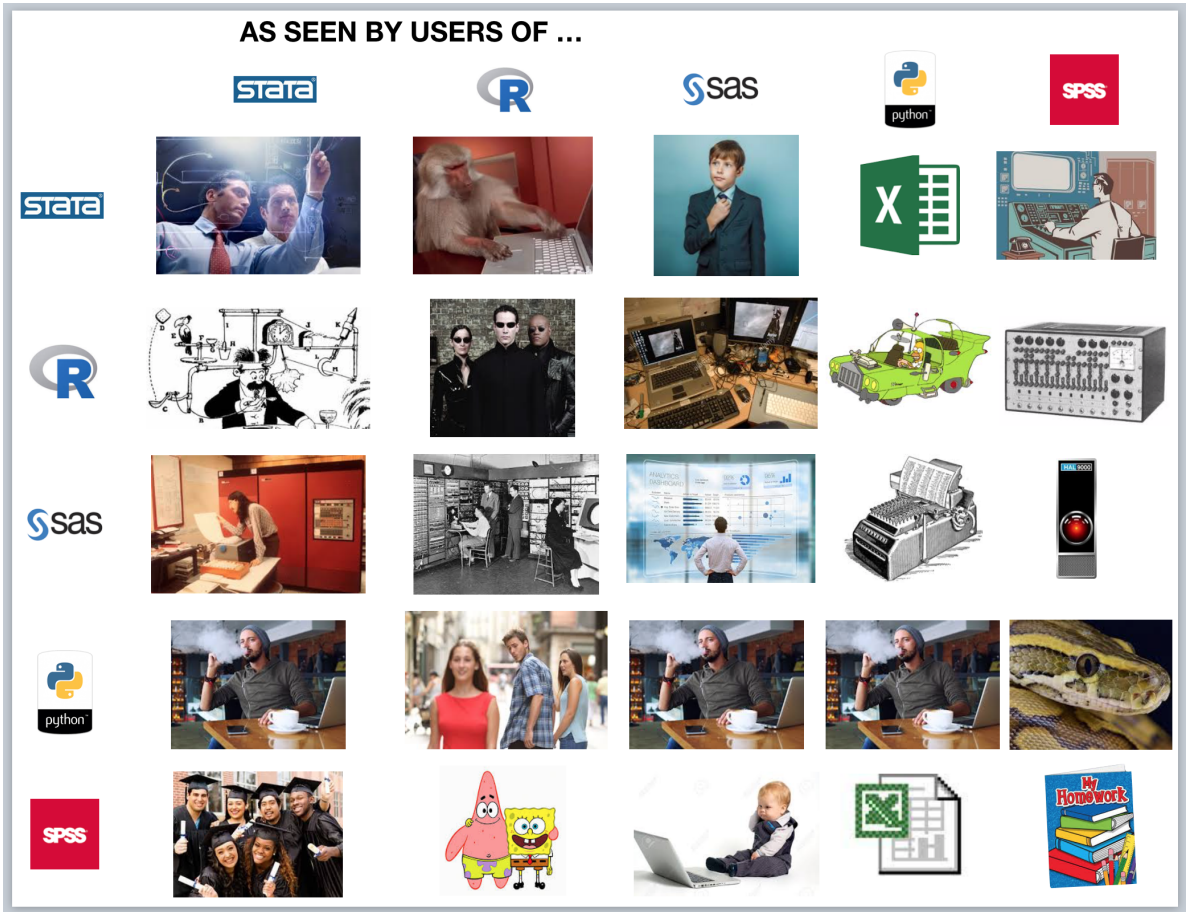
# Course outline

## Office hours and communication

- Office hours, Monday 5-6pm, 109 Davison Hall (Douglass) or Zoom.
- Email me to schedule an appointment if you cannot make office hours.

# Questions?

# Why R?




Source: Kieran Healey

# Why R?

- Alongside Python, it is one of the main programming languages used by data scientists
- A statistical computing language
- Free and open-source
- An active developer community
- RStudio
- A unified approach to data manipulation via the `tidyverse`

# RStudio

## Overview

- RStudio is an Integrated Development Environment for programming in R
  - Run code in the console or in scripts
  - Easy to view data, objects in memory, plots
  - Easy to create output such as papers or slides
  - Terminal interface
  - Integration with Github and Python

# RMarkdown

## Overview

- RMarkdown is an interactive coding environment
  - RMarkdown documents can combine text, LaTeX code, R code, and any output.
    - For example, these slides are rendered using RMarkdown
  - You will be using RMarkdown for your homework assignments and hopefully your papers

# Other resources

- StackOverflow
  - An online community for coding questions
    - Search for error messages or snippets. In most cases you should be able to find answers to your issues.
    - Sometimes it can take a while to figure out the appropriate query to use to find an answer.
    - If you can't find an answer, you can make your own question - but the formatting requirements are quite strict and users can be unforgiving.
  - A useful thread for posting an R question and example: https://stackoverflow.com/questions/5963269/how-to-make-a-great-r-reproducible-example

# Generative AI

- ChatGPT and other models are trained on large amounts of text data as well as code from Github
  - This means they can "understand" R and can write and comment on code
  - And these models have "knowledge" about most of the topics covered during the class
  - Some can even write and run code and produce results!

# Generative AI

## Two uses of generative AI

- Generative AI as a *research method*
  - Later in the semester, we will discuss these technologies in more detail and how they can advance computational social science
- Generative AI as a *pedagogical tool*
  - These tools can help you learn computational methods, but they also pose challenges to academic integrity

# Generative AI as a pedagogical tool

## Benefits and drawbacks

- *Benefits*
  - "Bespoke" problem solving reduces need to tailor queries for answers
  - Often correct for simple types of problems
  - Conversational modality can help you to learn
    - e.g. Ask for comments on your code or explanations
- *Drawbacks*
  - Overreliance may hamper ability to learn for yourself
  - Solutions can be incorrect and struggles for more complex problems
  - Less reliable for niche tasks and packages

# Generative AI and academic integrity

## Policies

- Attempt to solve problems yourself first to avoid overreliance
- *You are not permitted to use ChatGPT to solve homework problems or write code/text for your projects*
  - But you can use these to help you to learn
  - Contact me first if in doubt
- Document usage of these tools in any submissions
  - e.g. "I used ChatGPT to help me to understand how to use the `dplyr` package"
- *You are not permitted to use Github CoPilot in RStudio*
  - Autocompletion violates academic integrity
  - And the solutions are often distracting and not very helpful

# Exercise

## Problem solving in R

- Download and inspect this code in RStudio (paste it into a new R file)
  - http://tinyurl.com/badcode2024
- Task: Identify and fix the errors in the code

# Exercise

## Problem solving in R

- Three groups:
  1. Solve it without using the internet
  2. Solve it by searching StackOverflow
  3. Solve it by using ChatGPT

# Homework tasks

- Download and install RStudio
- Familiarize yourself with RMarkdown

# Questions?