Social Data Science

Introduction

Dr. Thomas Davidson

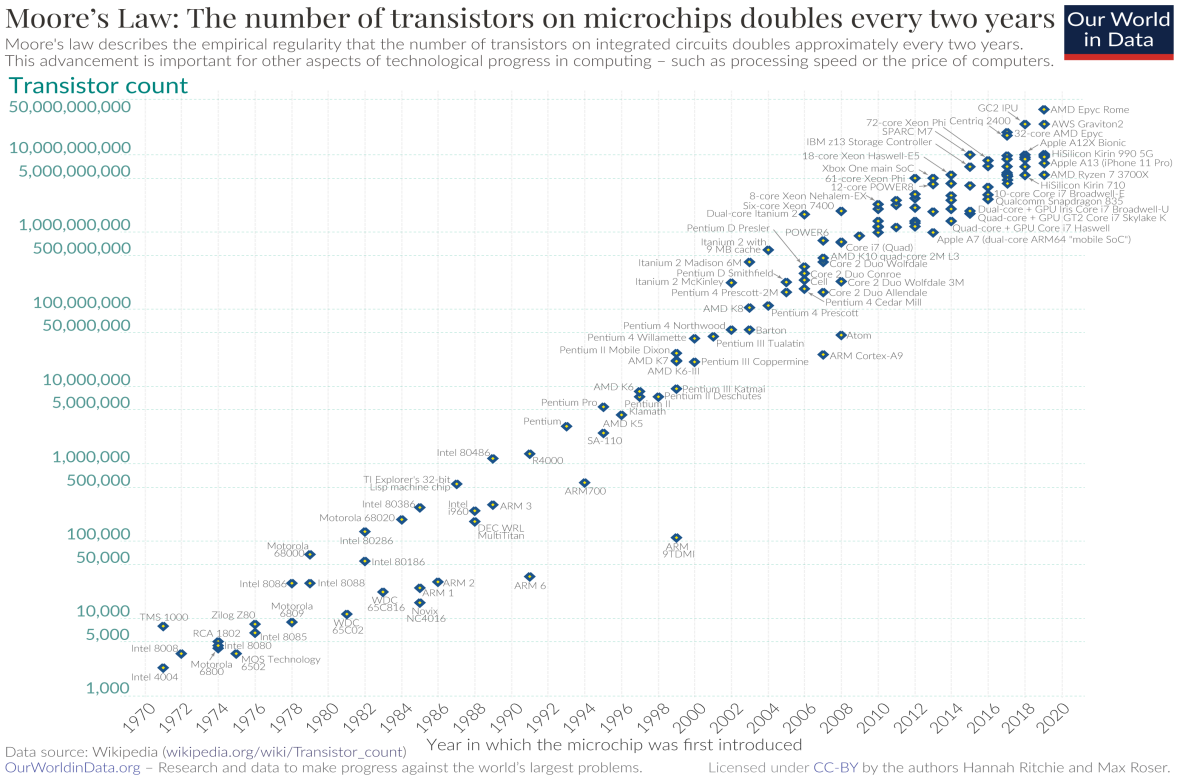
Rutgers University

September 1, 2021

Plan

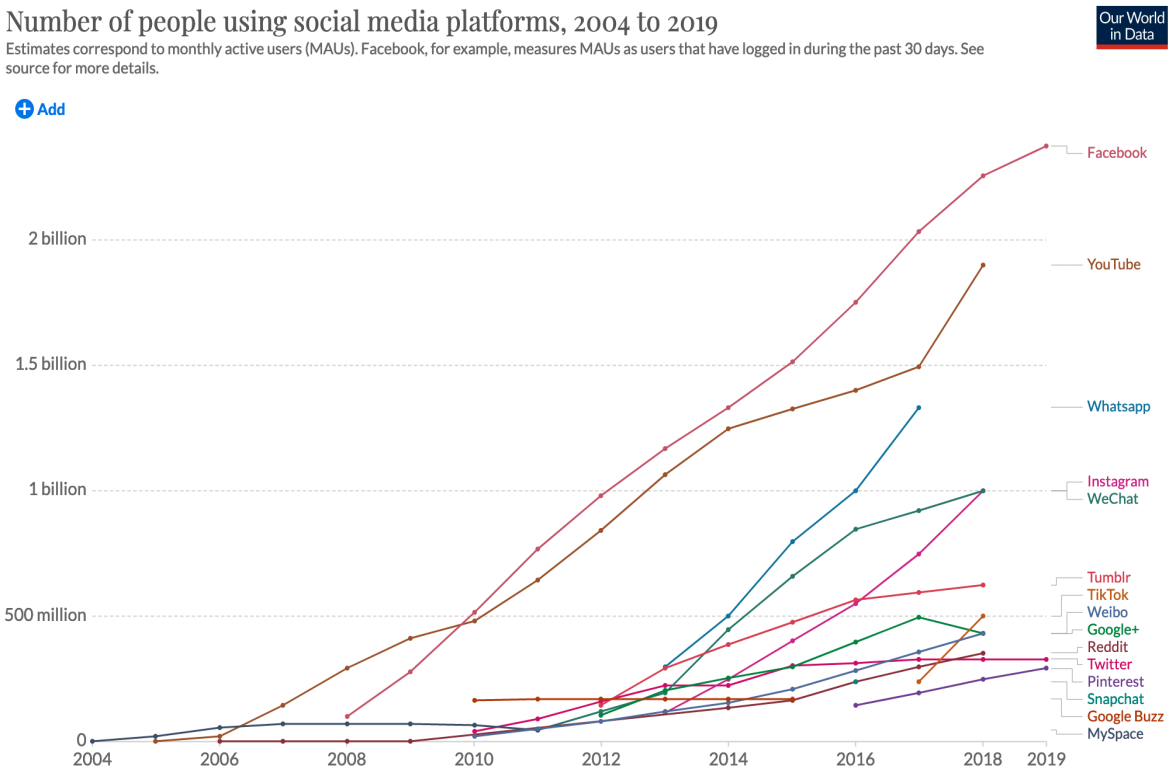
- Introductions
- A brief introduction to Social Data Science
- Course Outline
- R and RStudio
- Resources

Introduction to Social Data Science



https://en.wikipedia.org/wiki/Moore%27s_law#/media/File:Moore's_Law_Transistor_Count_1970-2020.png

Introduction to Social Data Science



https://ourworldindata.org/grapher/users-by-social-media-platform

Introduction to Social Data Science

Data Science and Social Science

- Contemporary society is characterized by a proliferation of data related to social life
- 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 Social Data Science

Data Science and Social Science

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

Introduction to Social Data Science

Data Science and Social Science

- Target audiences
  - Data scientists working with social data
    - Software engineers, project managers, research scientists, etc.
  - Social scientists using data science
    - Researchers, teachers, civil servants, etc.
- By the end of this course, the two groups should be hard to distinguish

Introduction to Social Data 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.

Lazer et al. 2009 make the case for computational social science (CSS)

Introduction to Social Data 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 Social Data Science

The emergence of a field

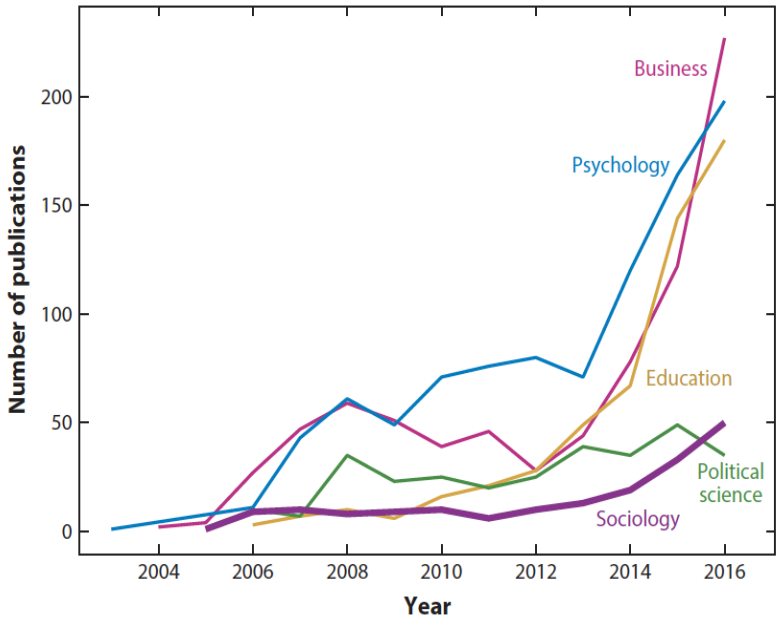


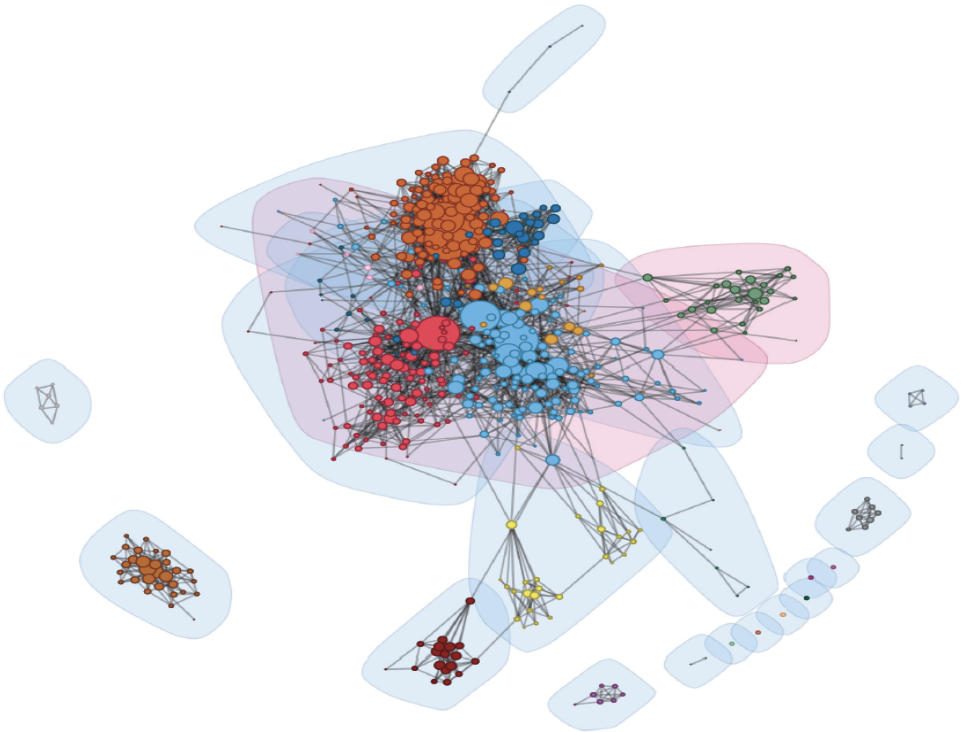
Figure 1

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

Edelmann et al. 2020

Introduction to Social Data Science

The emergence of a field



Edelmann et al. 2020

Introduction to Social Data 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 Social Data Science

Preview of cutting-edge research

- What do word embeddings reveal about our understanding of culture?
- How can we use machine-learning to study the dynamics of protest?
- How can we use Google Streetview to estimate demographic composition of neighborhoods?

Introduction to Social Data Science

Ethical considerations

- What is the legality of using data scraped from the web?
- Why do computer vision models discriminate against racial minorities?
- Should we trust predictive modeling systems with consequential decisions?

Course outline

Goals

- By the end of this course you should be able to
  - Understand the field of social data science / 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

Course outline

Prerequisites

- Data 101 or equivalent programming experience
- Some basic statistical knowledge

Course outline

Structure

1. Programming in R (Weeks 1-3)
2. Data collection (4-6)
3. Natural language processing (7-9)
4. Machine learning (10-13)
5. Agent-based modeling (14)

Course outline

Assessment

- Participation (10%)
- Homework assignments (60%)
  - Programming fundamentals
  - Data collection using APIs
  - Natural language processing

Course outline

Assessment

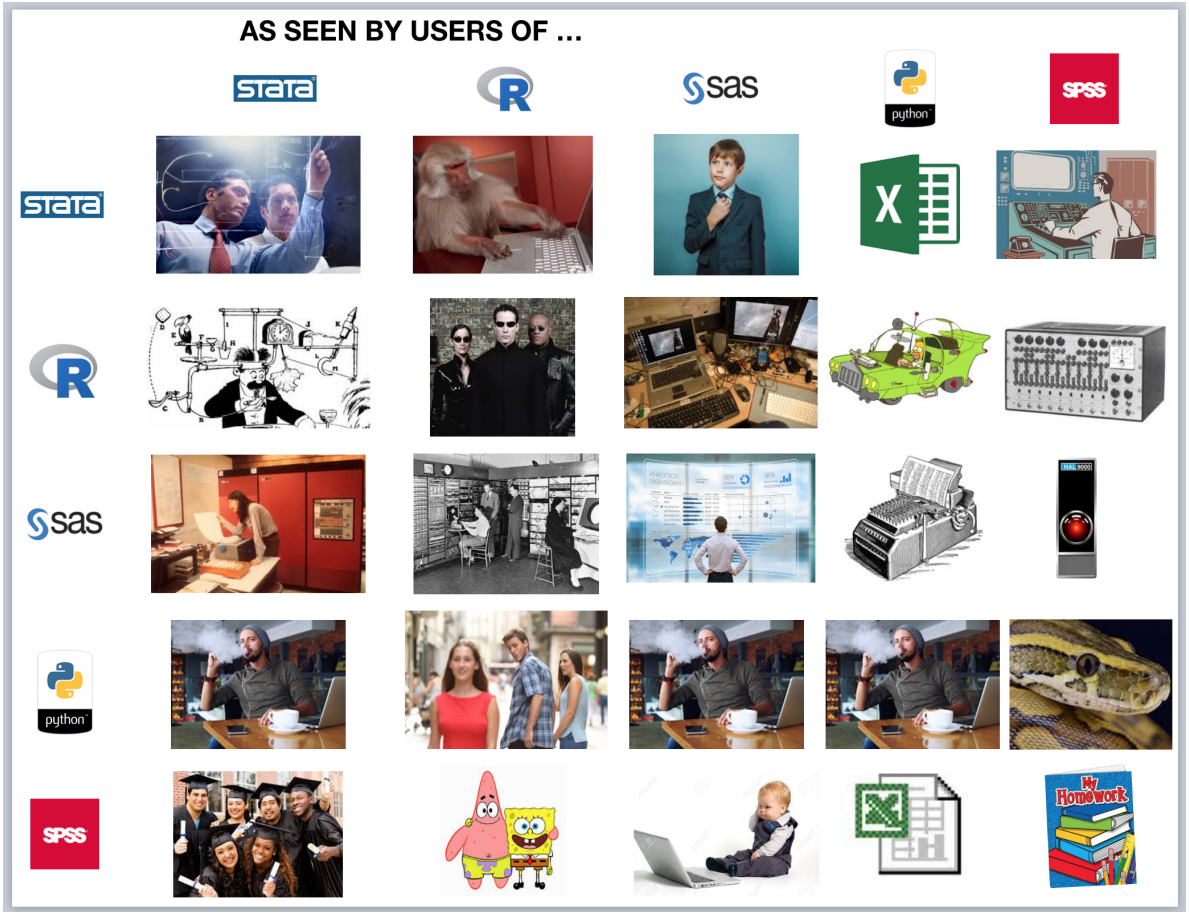
- Group project (30%)
  - Phase 0: Develop project ideas
  - Phase 1: Submit project proposal
  - Phase 2: Prototype
  - Phase 3: Present final project

Course outline

Policies

- Read the syllabus
  - Diversity and inclusion
  - Academic integrity
  - Accommodations
  - COVID-19

Why R?



https://kieranhealy.org/blog/archives/2019/02/07/statswars/

Why R?

- Alongside Python, it is one of the main programming languages used by data scientists
- Free and open-source
- A statistical programming language
- An active developer community
- RStudio

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.
  - Write in Markdown or Visual Editor
  - These slides are rendered using RMarkdown
  - You will be using RMarkdown for your homework assignments and hopefully your papers

Other resources

- R4DS Community
  - An online community associated with the R4DS book, including a Slack channel https://www.rfordatasci.com/
- R Reddit
  - https://www.reddit.com/r/rstats/
- R Twitter
  - Follow #rstats
- An Introduction to R
  - Free online R textbook https://intro2r.com/index.html

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

Questions?