Data Science Applications to Politics Research

Week 1

Daniel de Kadt & Zach Dickson

London School of Economics

GV330

Welcome to GV330

Dr Zach Dickson

Postdoctoral Fellow in Quantitative Methodology in the Department of Methodology and DSI Affiliate.

Dr Daniel de Kadt

Assistant Professor in Quantitative Research Methods in the Department of Methodology and DSI Affiliate.

Overview

Overview

Course learning outcomes

- Critically assess the application of data science methodologies in political science research.
- Practically apply quantitative methodologies using novel complex, or “big” data sources to address political science topics.
- Reflect critically on the practice and challenges of doing (data) science, including understanding reproducibility and the replication process and its importance.
- Replicate and reappraise the data analysis in important research papers and extend those analyses, using appropriate methodologies.
- Consider and respond to current debates about the uses of data in terms of ethics, policy, and privacy.

Course outline

Topic 1: Causality, credibility, and “big data” (weeks 2 & 3)

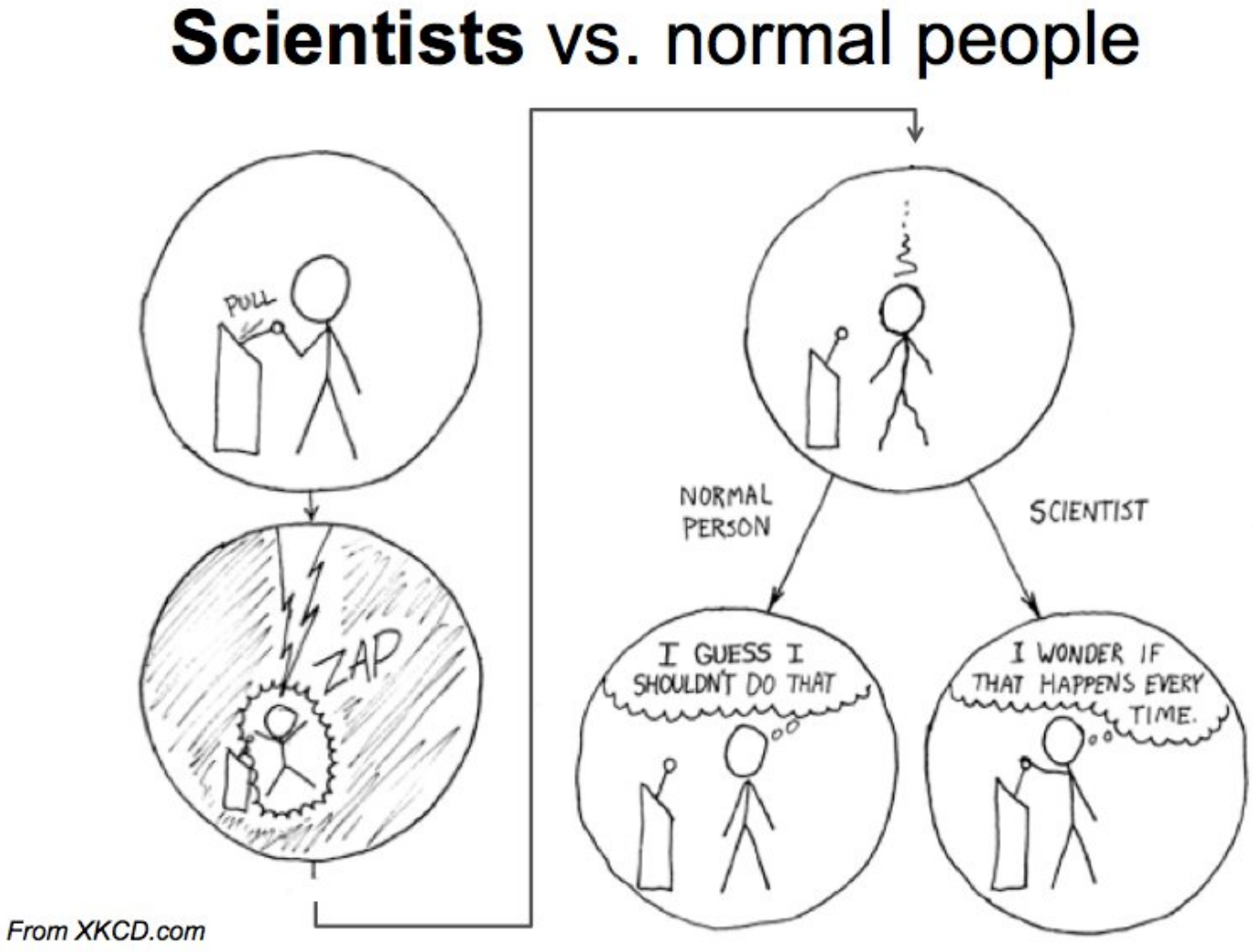
Topic 2: Administrative, urban, and open data (weeks 4 & 5)

Topic 3: Media, social media, and search data (weeks 7 & 8)

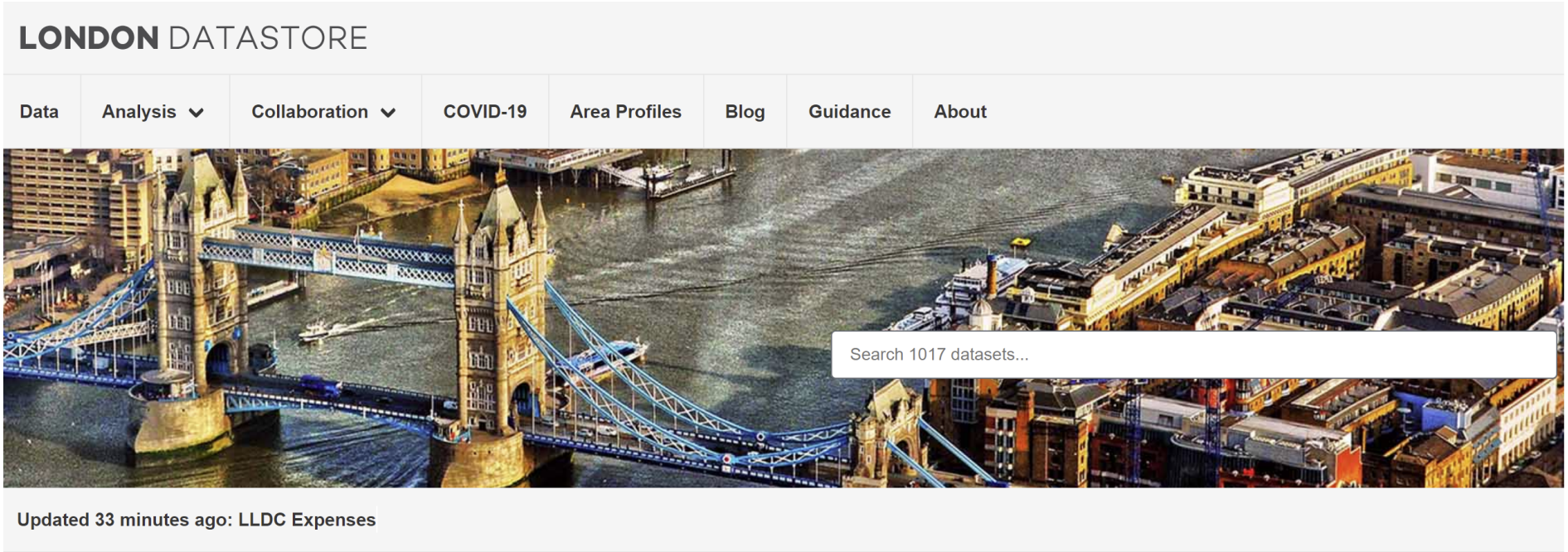
Topic 4: Text, image, video, and audio data (weeks 9 & 10)

Topic 5: Generative artificial intelligence (week 11)

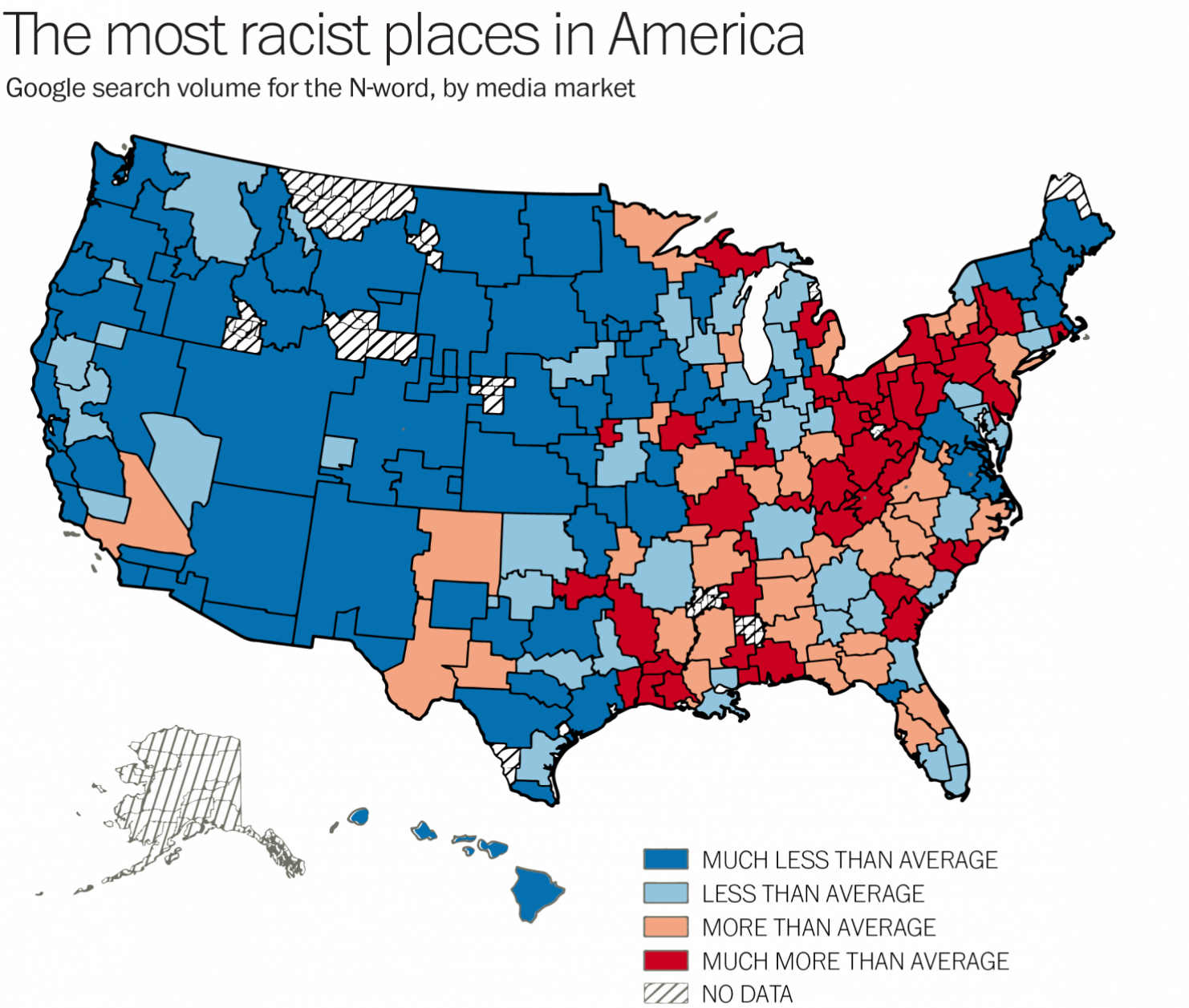
Topic 1: Causality, credibility, and “big data”



Topic 2: Administrative, urban, and open data

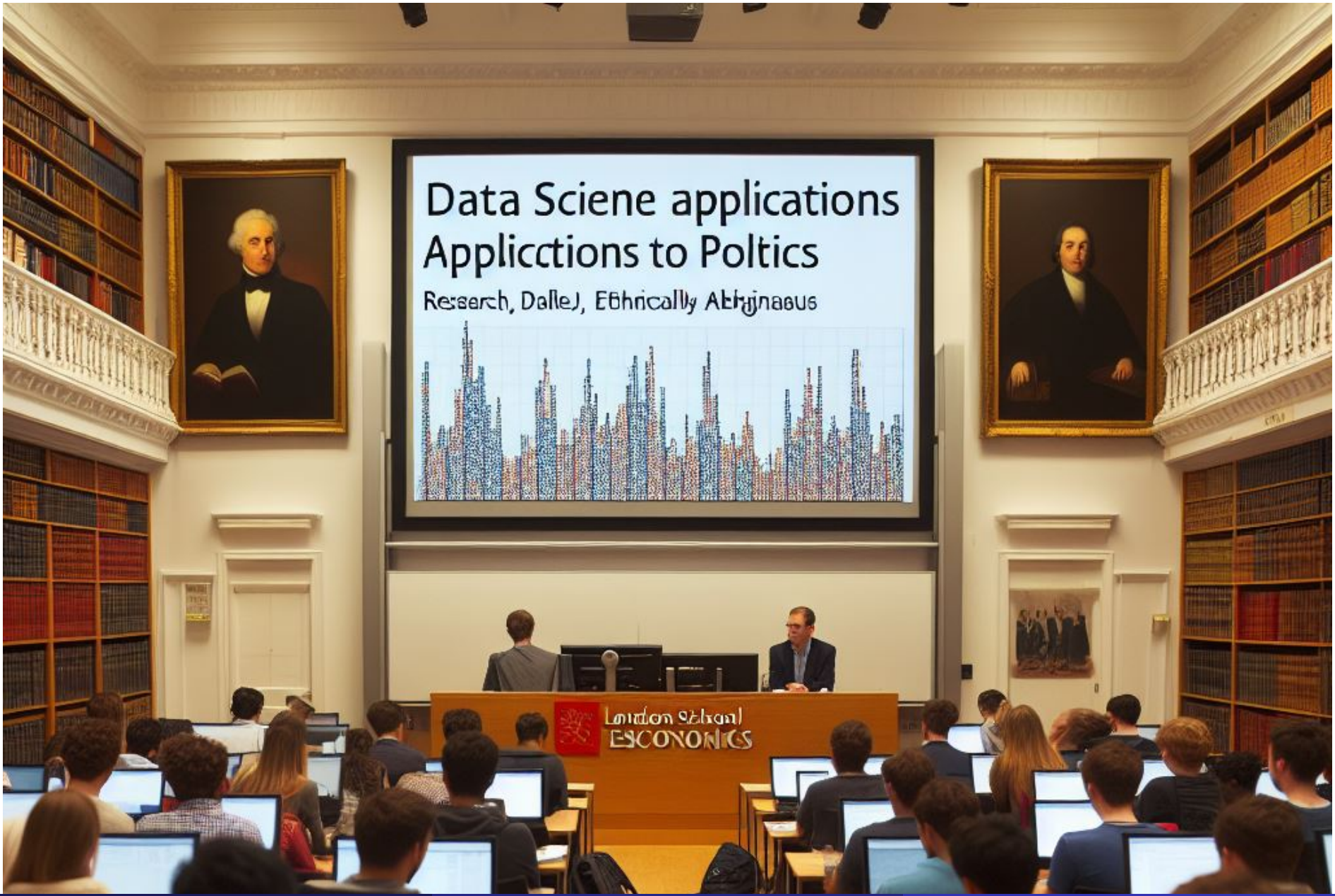


Topic 3: Media, social media, and search data



Topic 4: Text, image, video, and audio data

Topic 5: Special topics (e.g., Generative AI & LLMs)



Overview

Course delivery

- **Lectures**: Fridays, 13:30-15:00 in CKK.1.15
- **Seminars**:
  - Group 1: Fridays 15:00-16:30 in CKK.2.17
  - Group 2: Fridays 16:30-18:00 in CKK.2.17
- Dr Dickson (weeks 2, 4, 7, 9, 11) and Dr de Kadt (weeks 1, 3, 5, 8, 10) will trade off throughout the term.

Readings & assessments

- Readings
  - Available on Moodle/Leganto (linked on Github)
  - Required reading: 1-2 articles (or podcast or video) per week

- Formative
  1. Sign up to present one article in class from list of options (see Github site for link)
     - Presentations begin in the Seminar in Week 3
     - See presentation guidelines on Github (we will walk through expectations in class at a later date)
  2. Formative problem set - due 5pm 27th March (WT Week 10)

- Summative
  1. 100%: Replication exercise - due 5pm 25th April (ST)

Additional Support

- Talking to us (we don't bite!):
  - Book office hours via studenthub.
  - Additional appointments via Zoom may be available by request, but no guarantees.
  - In general, we won't answer substantive questions about course material over email. You can post these questions on the Moodle forum and we will get back to you as quickly as possible.
- LSE Life
- LSE Wellness

Overview

Computing: R

- Prior familiarity with `R` – or willingness to learn – is expected.
  - R is a programming language for statistical computing (based on the earlier language S)
  - It is primarily a functional programming language
- Resources to help you brush up on your `R` skills:
  - LSE Digital Skills Lab: R Fundamentals workshop series, openings starting next week
  - DataCamp, CodeAcademy

Computing: R

R and R Studio

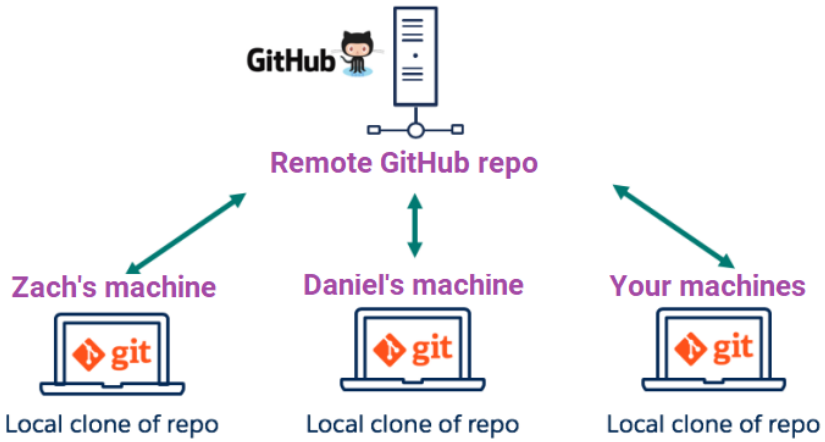
- Why R?
  - Many great languages for data science: R, Python, Julia, etc.
  - The wisest among us are not dogmatic about what tools we use.
  - In political science, R is probably dominant (Python is an increasingly close second).
  - Main advantages of R: relatively easy to learn and use, very powerful and mature packages/libraries for econometrics and causal inference, reasonably good for small-scale machine learning.
  - Main disadvantages of R: quite fragile, quite inefficient, highly package/library-dependent development process
- For this class, we recommend combining R + R Studio:
  - While R is a programming lanaguage, R Studio is an integrated development environment (IDE) – a useful/attractive shell for interacting with R

R Markdown and Quarto

- For seminars, we will sometimes use R Markdown or Quarto files (the course website is built in Quarto):
  - These are files – .Rmd and .Qmd – that you can edit and run in R Studio.
  - R Markdown (and Quarto) are “unified authoring framework(s) for data science, combining your code, its results, and your prose commentary" into a single document.
  - This facilitates what is called literate programming and is extremely useful when reproducibility and transparency are important.
- Your final assignment will be written in R Markdown or Quarto (don’t fret: it’s pretty easy to learn, and we will help you!)

Git and GitHub

- Git is software used for version control – you will need to install Git on your local computer
  - Organized around “repositories” or “repos”: Essentially a folder of sub-folders and files that Git keeps track of.
- GitHub is a suite of services built around Git – you will need an account with GitHub
  - You can create, manage, delete (be careful), and most of all remotely store repos on GitHub.
  - Your local version of Git will keep track of the remote files (on GitHub) and the local files (your computer).



(Adapted from sfdctechie.wordpress.com)

Git and GitHub

- The utility of Git and GitHub for software development and data science is not debatable.
- If you end up in any professional development environment you will use Git in some form.
- But how we will use Git and GitHub in this class?
  - You can follow the course materials by cloning or forking the class materials from the course_materials repo.
  - Your final assignment will be a replication and reappraisal of a published paper.
  - When replicating others work, it's critical that your work be transparent, traceable, and easy to review.
  - So, as part of the assignment submission, you will submit a GitHub repo that fully reproduces your final paper (which will be written as a R Markdown or Quarto document, knitted to a final .pdf file that you submit on Moodle).
  - Don't fret: We'll talk about this more as the term goes on!

Overview

What is Data Science?

Not universally agreed-upon, but most definitions emphasize:

- **Interdisciplinarity**: uses tools from statistics, computer science, and domain-specific expertise
- **Extracting insights** from data, often to inform decision-making (e.g., by making predictions)
- Various **activities**, including data cleaning and preprocessing, analyzing and modeling, and data visualization

What is Data Science?

Activities	Examples
Data gathering, preparation, and exploration	Survey data, experimental data, genomic data, textual data, administrative data, image data, web data, and sensor data Data cleaning and exploratory data analysis methods for checking on outliers and data quality
Data representation and transformation	Relational and nonrelational databases Networks and graphs Other mathematical structures for data
Computing with data	R and Python Programming packages, text manipulation languages Cluster and cloud computing Reproducible workflows
Data modeling	Determining or hypothesizing data generating probability functions, structural and predictive modeling
Data visualization and presentation	Types of visualizations and graphs Rules for labeling and presenting data Psychological impacts of various displays
Data archiving, indexing, and search and data governance	Standards for open data and reproducibility Determining rules for access and privacy protection where necessary
Science about data science	How people do data science Impacts of data science and big data on society

From Brady 2019 (adapted from Donoho 2017)

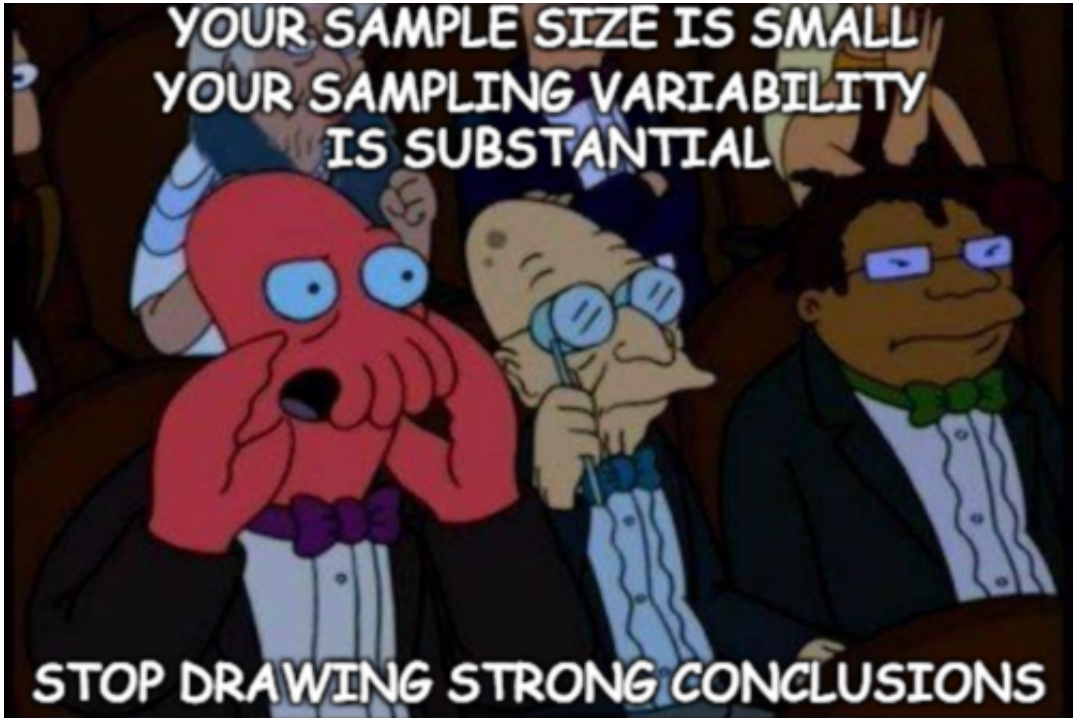
What is “data”?

- Data: Abstractions, representations, symbols, variables, related to underlying information
- Information: Signals that contain something other than random noise
- So, data is a representation of underlying information.
- As social scientists, our goal is to combine data with analyses (justified by explicit assumptions we make) to extract meaning.

What is “big data"?

- A highly multidimensional and complex body of information that is not inherently ordered; must go through a process of data reduction before it can be analyzed (Patty & Penn 2015)
- Rectangular array of information with n rows and p columns (Titiunik 2015)
  - Big data “as large n”: data sources in which the number of observations (n) is extremely large relative to the number of variables (p) available in the dataset
  - Big data “as large p”: data sources in which the number of variables (p) is very large
- “Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software...” (Wikipedia)

Why “big data"?

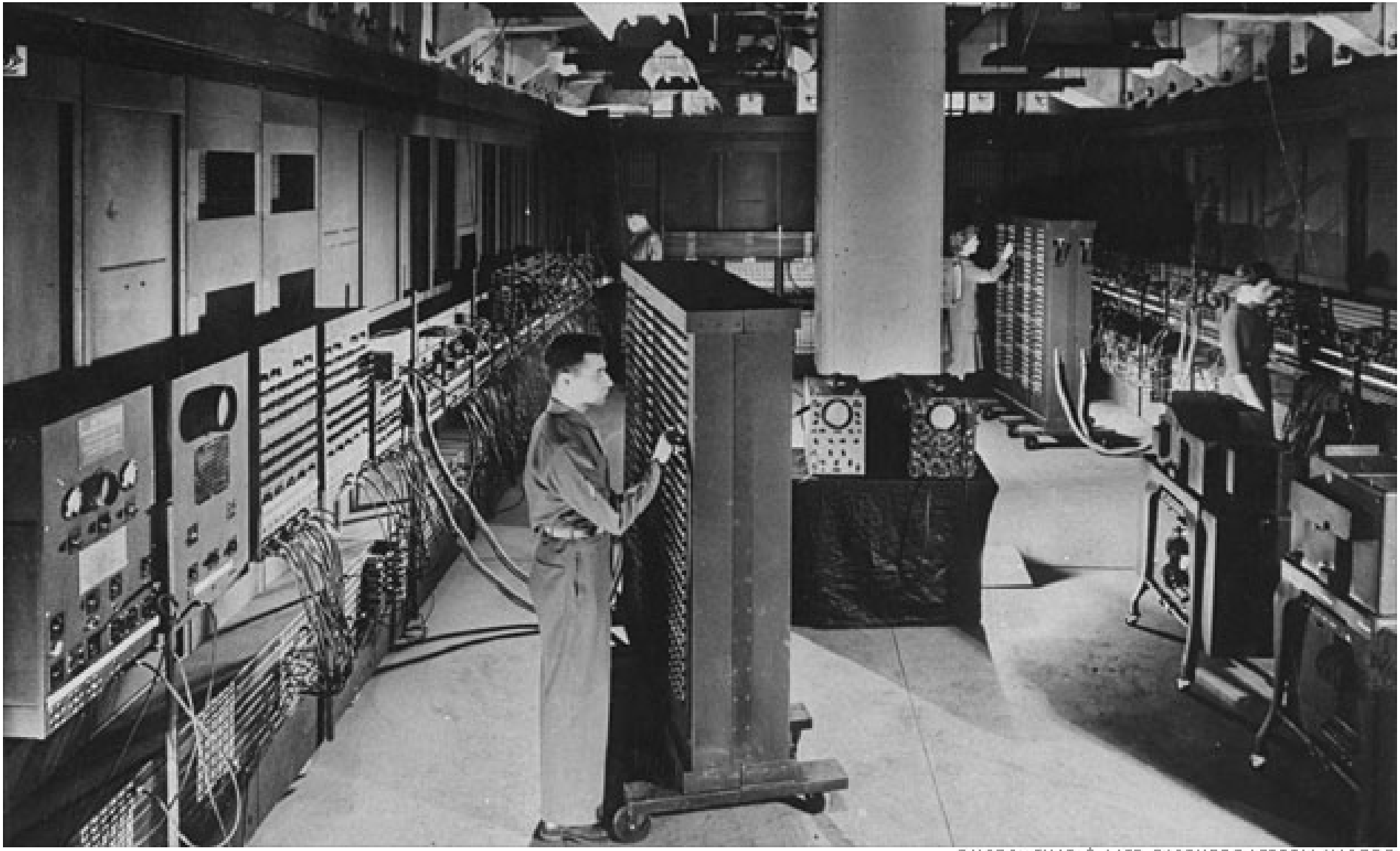

YOUR SAMPLE SIZE IS SMALL
YOUR SAMPLING VARIABILITY
IS SUBSTANTIAL
STOP DRAWING STRONG CONCLUSIONS


Keith Rabois
@rabois
Follow
with very large numbers of n’s you don’t need randomization.
2:41 am · 25 Mar 2020 from San Francisco, CA
48
150
36
53

Why “big data"?

- Intuitively, we often assume that more data is better. But how much better, and at what rate do improvements accrue?
- One school of thought is that latent dimensionality is more important than sample size
  - Intuition: At some point, more data isn’t very helpful for anything other than precision
  - Instead, your data needs to represent “more” (dimension of) underlying information
  - Is more p always useful?Depends on the p!
- The real gains to be made are in terms of increasing the underlying dimensionality of the data, and then applying the right tools to that data (Spirling & Morucci, 2025)
- This is unintuitive to social scientists, as we have spent the last 100 years specifically engaged in dimension reduction!

A moving target?



What can I do with training in Politics & DS?

- **Employers**: civil service, governments, international organizations, digital media firms, regulators, think tanks, political consultancies, political parties and campaigns, political risk analysis firms, and others.
- **Careers**: data science, political data analytics, data journalism, political risk analysis and forecasting, targeted campaigning, cyber security and cyber threat analysis , policy analysis, and more.
- **Academia**: one of us, one of us, one of us (it's not a cult, we promise)

Data Scientist

Harvard Business Review

Analytics And Data Science

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and DJ Patil

From the Magazine (October 2012)

Data Scientist

r/artificial • 8 mo. ago
Thick-Resident8775

Are data science/tech jobs going to get replaced by AI?

Discussion

I'm in the last year of my data science degree and I'm scared if it's even worth it as there are reports and AI godfather Geoffrey Hinton said itself that it can take away your jobs, especially tech jobs. I just used GPT-4 and it was really impressive, if they keep updating at such a fast rate then will it be smarter than data scientists/AI engineers?

8 25 Share

Express_Category3067 • 8mo ago

I said this before, prepare for a future where your intelligence is valuable in the market.

If that future does not exist, because AI can fully replace your intelligence as a data scientist, it can do the same for 90% of the jobs.

Worry not about that future, I say this in a nice way, you can't prepare for it.

perplex1 • 8mo ago

As I see it. Data scientists will be able to focus on the business goals better. Being creative in how to use data to drive and inform efforts in new ways. But that takes deep knowledge and familiarity with the processes, data availability, and business goals you are trying to solve for.

Overview

Pre-term Survey

If you have not already, please complete this brief survey:
`https://forms.gle/FLDPQG82vtbQnbte8`