

Introduction to Data Science

Session 1: What is data science?

Simon Munzert

Hertie School | GRAD-C11/E1339

Welcome!

Introductions

Course

 <https://github.com/intro-to-data-science-24>

Much of this course lives on GitHub. You will find lecture materials, code, assignments, and other people's presentations there. We also have Moodle, which is for everything else.

Me

 I'm **Simon Munzert** [si'mən munsərt], or just Simon [saɪmən].

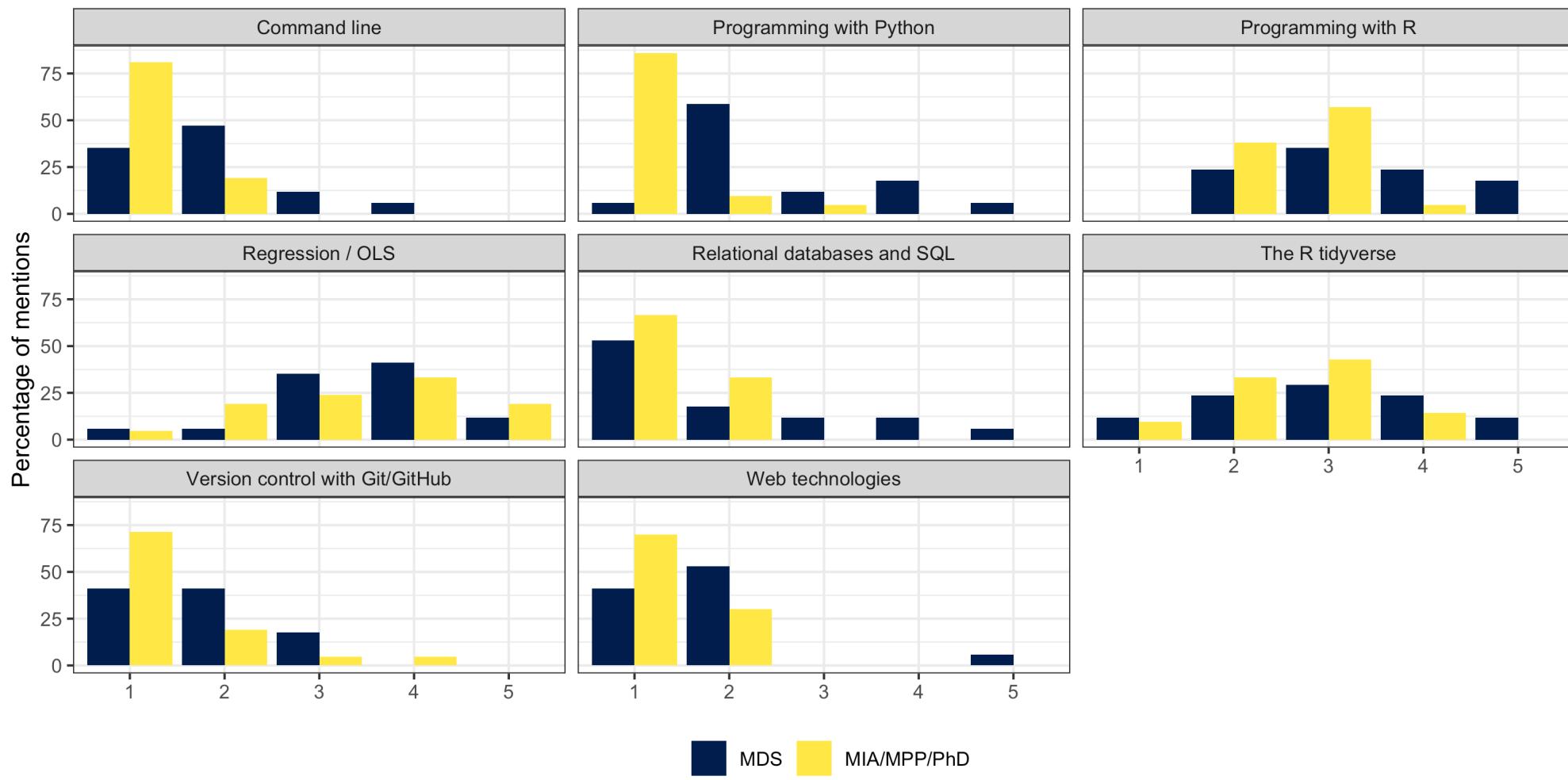
 munzert@hertie-school.org

 Professor of Data Science and Public Policy | Director of the Data Science Lab

You

What's your name? And would you share a fun fact about yourself?

More about you



More about you

MPP/MIA/PhD

A word cloud visualization showing the most frequent terms related to the field of statistics. The words are colored in shades of purple, teal, and green. The largest words are 'statistics' (purple), 'data' (purple), and 'analysis' (purple). Other prominent words include 'coding' (purple), 'research' (teal), 'machine_learning' (green), 'learning' (green), 'big_data' (teal), 'r' (teal), 'trash' (green), 'data_cleaning' (green), 'wrangling' (green), 'continuous_learning' (green), 'maths' (green), 'cleaning' (green), 'machine' (green), and 'learning' (teal).

learning
machine_learning machine
r trash
big data
analysis
coding
research
data_analysis
big_data
coding_coding

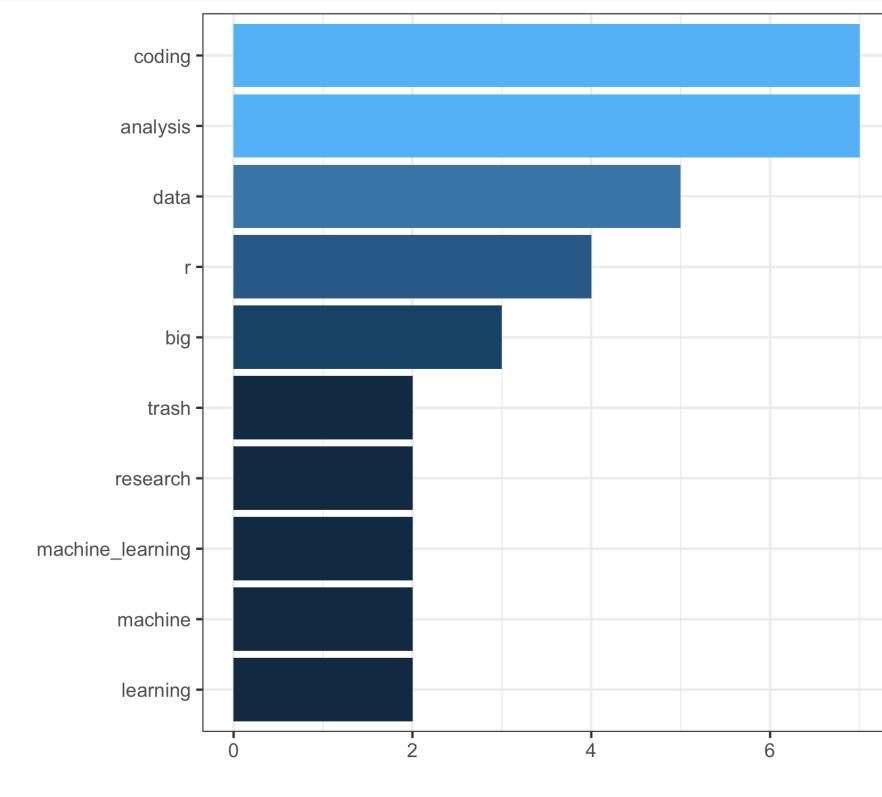
MDS

A word cloud visualization showing the most frequent terms related to the field of statistics. The words are colored in shades of purple, teal, and green. The largest words are 'statistics' (purple), 'data' (purple), and 'learning' (teal). Other prominent words include 'python' (green), 'data_cleaning' (green), 'wrangling' (green), 'continuous_learning' (green), 'maths' (green), 'cleaning' (green), 'machine' (green), 'r' (teal), 'learning' (teal), 'wrangling' (green), 'maths' (green), 'data_wrangling' (green), and 'continuous_learning' (green).

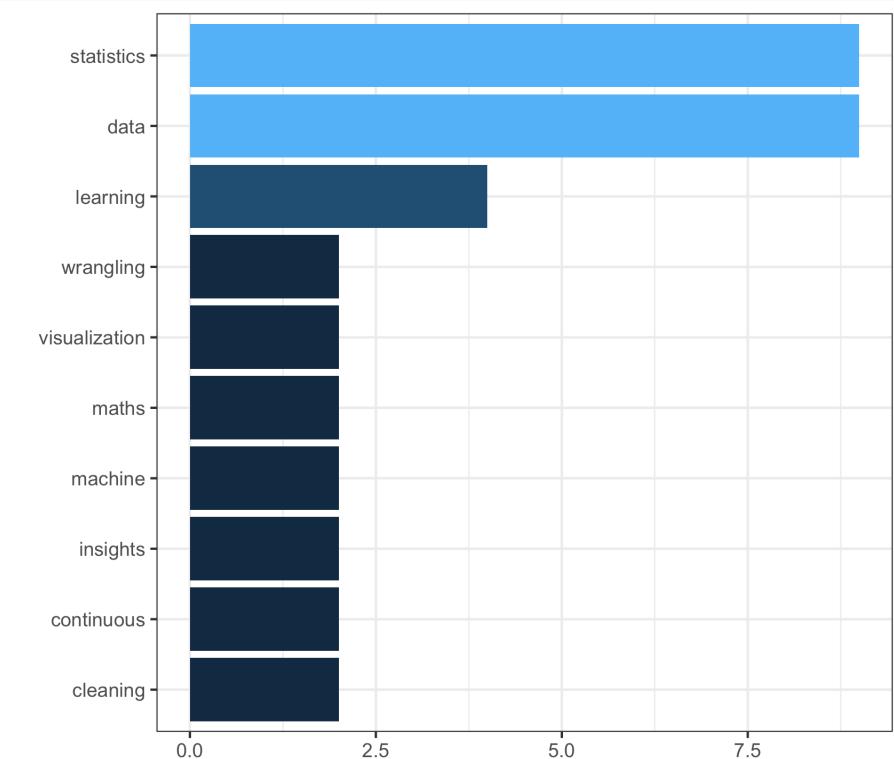
machine_learning
insights continuous
visualization
statistics
data
learning
python
data_cleaning
wrangling
continuous_learning
maths
cleaning
machine
r
learning
wrangling
maths
data_wrangling
continuous_learning

More about you

MPP/MIA/PhD



MDS



The labs

Who & how

- This course is accompanied by labs administered by **Carmen Garro** and **Sebastian Ramirez Ruiz**.
- The labs are mandatory (MDS) / optional (the rest). Please attend them in any case.
- As with the regular classes, please stick to the lab you are assigned to.



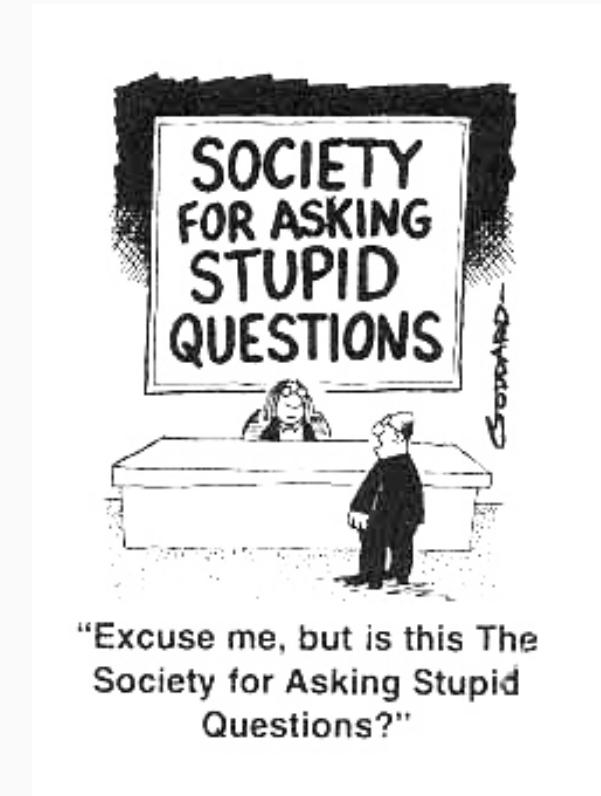
What for

- What these sessions are meant for:
 - Applying tools in practice
 - Discussion of issues related to the assignments
 - Boosting your R skills
- What these sessions are **not** meant for:
 - Solving the assignments for you
 - Taking care of developing your coding skills



Class etiquette

- Learning how to code can be challenging and might lead you out of your comfort zone. If you have problems with the pace of the course, let me and the TAs know. I expect your commitment to the class, but **I do not want anyone to fail.**
- You are all genuinely interested in data science. But there is also considerable variation in your backgrounds. This is how we like it! Some sessions will be more informative for you than others. If you feel bored, **look out for and help others**, or explore other corners of R you don't know yet.
- The pandemic is still around, and other crises have emerged. We are affected by them in different ways. **Let's support each other.**
- **Be respectful** to each other, all the time. This includes the TAs and me.
- **Ask questions** whenever you feel the need to do so!



"Excuse me, but is this The Society for Asking Stupid Questions?"

Table of contents

1. Welcome!
2. What is data science?
3. (Data) science for public policy
4. Class logistics

What is data science?

What is data science?

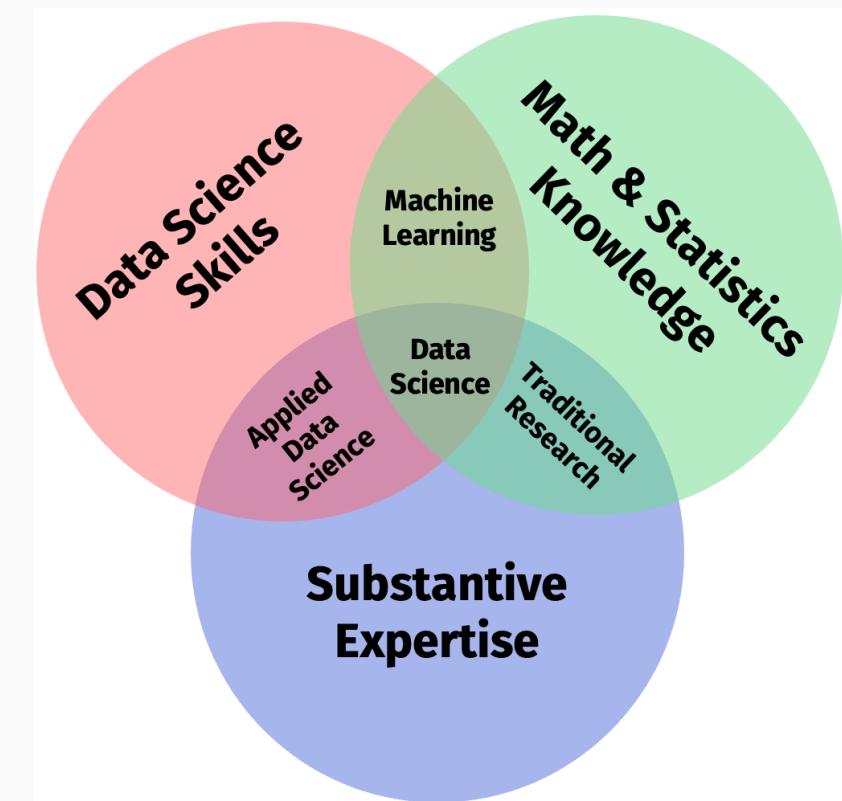
What is data science?

"Data science is an interdisciplinary academic field that uses statistics, scientific computing, scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from potentially noisy, structured, or unstructured data." - [Wikipedia](#)

"Data science is a concept to unify statistics, data analysis, informatics, and their related methods to understand and analyze actual phenomena with data." - [Chikio Hayashi](#)

Overall, there's **no consensus** - it is a buzzword after all.
We're going to carry on with Conway's working definition.

A working definition



Source [Drew Conway, 2010](#) (adapted)

Types of data-driven research and their role for policy

1. Description

- What is the state of the world?
- What are the trends over time?
- What are the differences between groups?

2. Explanation

- What is the effect of a policy?
- Does the effect vary across groups?
- What are the mechanisms behind the effect?

3. Prediction

- What is the path of an indicator?
- (When) will future events happen?
- What class does this observation most likely belong to?

The value for policy-making

- At the center of **monitoring**
- "How many people consume misinformation online?"
- "How many people are unemployed in a certain district?"
- "How does the distribution of income vary across educational segments of the population?"

The value for policy-making

- At the center of **evaluation**
- "Did the wage increase lead to a decrease in employment?"
- "Did the campaign affect the exposure to misinformation differently across groups?"
- "Why did the intervention not lead to the expected results?"

The value for policy-making

- At the center of **forecasting** but also **targeting** and **measurement**
- "Will there be conflict?"
- "How many people will be unemployed in a certain district next year?"
- "Which individuals are most likely to be affected by a policy?"

The data science pipeline

Preparatory work

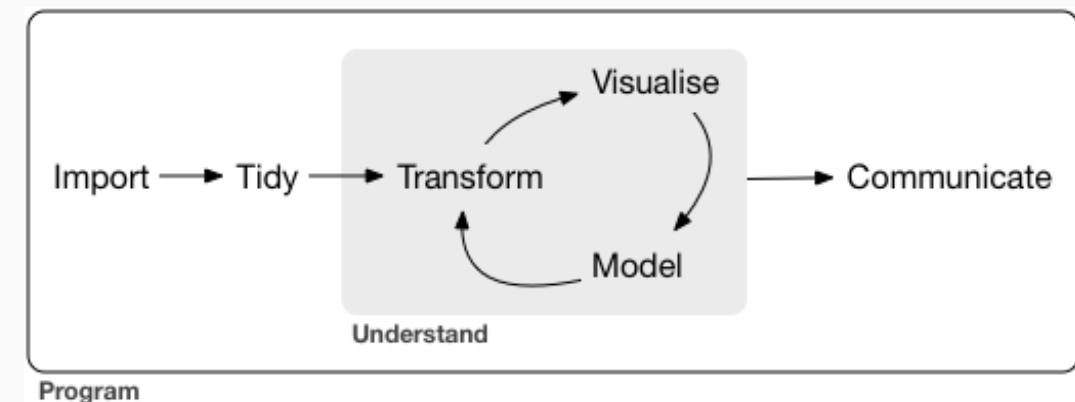
- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

The data science pipeline

Preparatory work

- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

Data operation



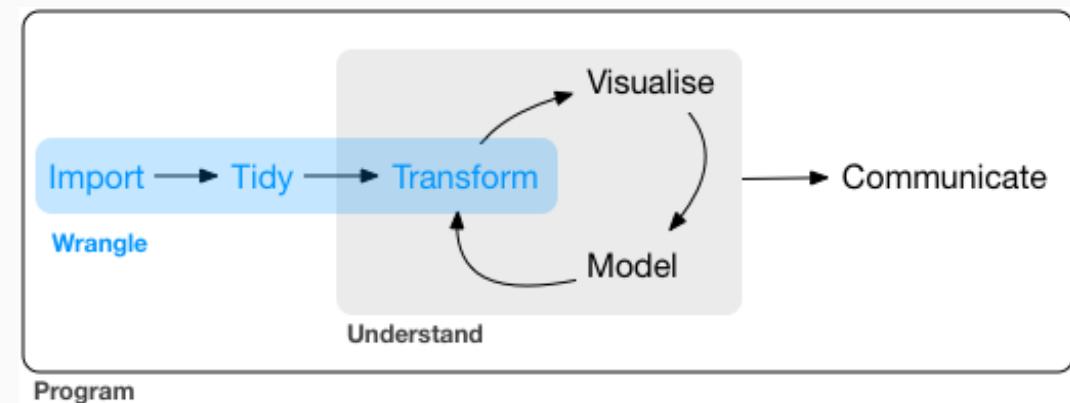
The data science pipeline

Preparatory work

- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

Data operation

- **Wrangle:** import, tidy, manipulate



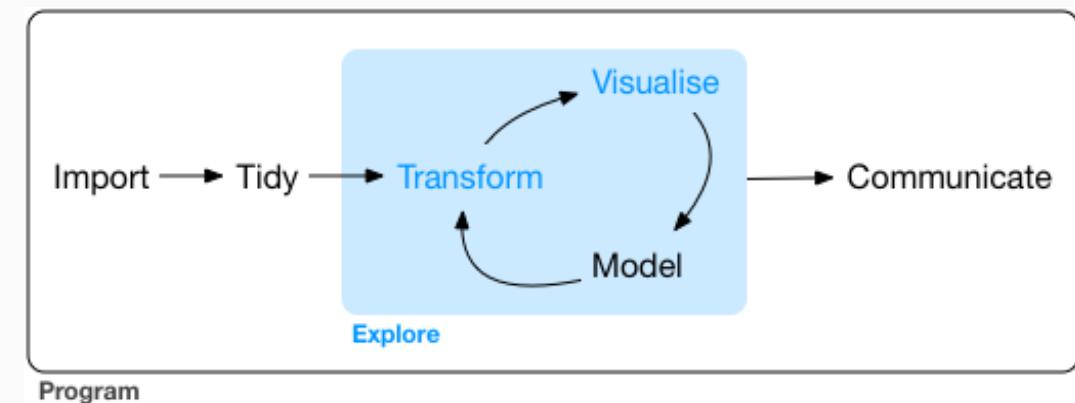
The data science pipeline

Preparatory work

- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

Data operation

- **Wrangle**: import, tidy, manipulate
- **Explore**: visualize, describe, discover



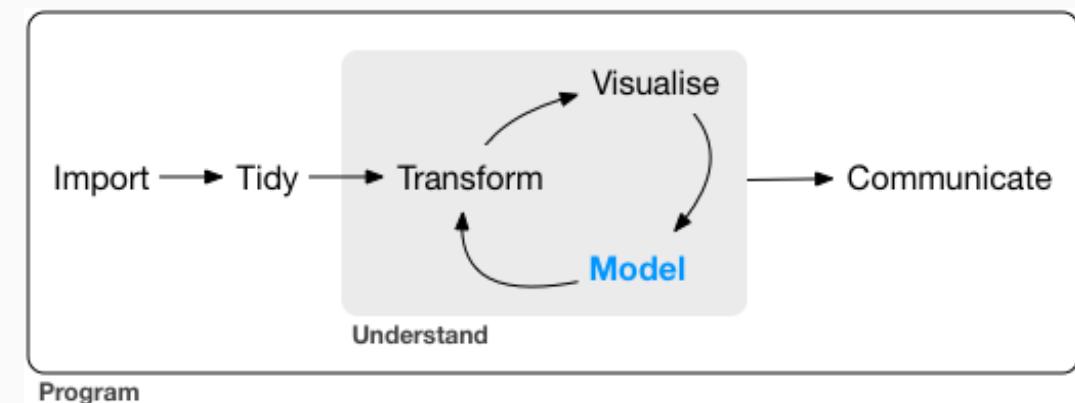
The data science pipeline

Preparatory work

- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

Data operation

- **Wrangle**: import, tidy, manipulate
- **Explore**: visualize, describe, discover
- **Model**: build, test, infer, predict



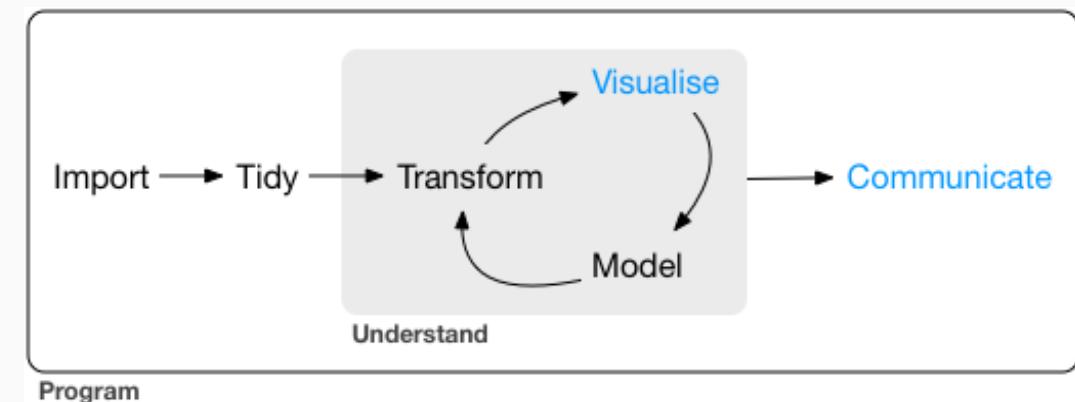
The data science pipeline

Preparatory work

- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

Data operation

- **Wrangle**: import, tidy, manipulate
- **Explore**: visualize, describe, discover
- **Model**: build, test, infer, predict



Dissemination

- **Communicate**: to the public, media, policymakers
- **Publish**: journals/proceedings, blogs, software
- **Productize**: make usable, robust, scalable

The data science pipeline

Preparatory work

- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

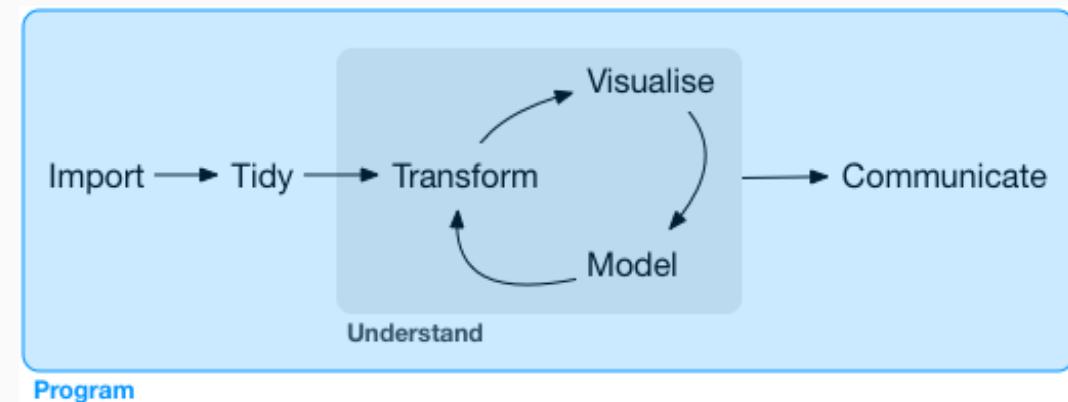
Data operation

- **Wrangle**: import, tidy, manipulate
- **Explore**: visualize, describe, discover
- **Model**: build, test, infer, predict

Dissemination

- **Communicate**: to the public, media, policymakers
- **Publish**: journals/proceedings, blogs, software
- **Productize**: make usable, robust, scalable

Meta skill: programming



The data science pipeline

Preparatory work

- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

Data operation

- **Wrangle**: import, tidy, manipulate
- **Explore**: visualize, describe, discover
- **Model**: build, test, infer, predict

Dissemination

- **Communicate**: to the public, media, policymakers
- **Publish**: journals/proceedings, blogs, software
- **Productize**: make usable, robust, scalable

Meta skill: programming with R



(Data) science for public policy

The MIT Billion Prices Project

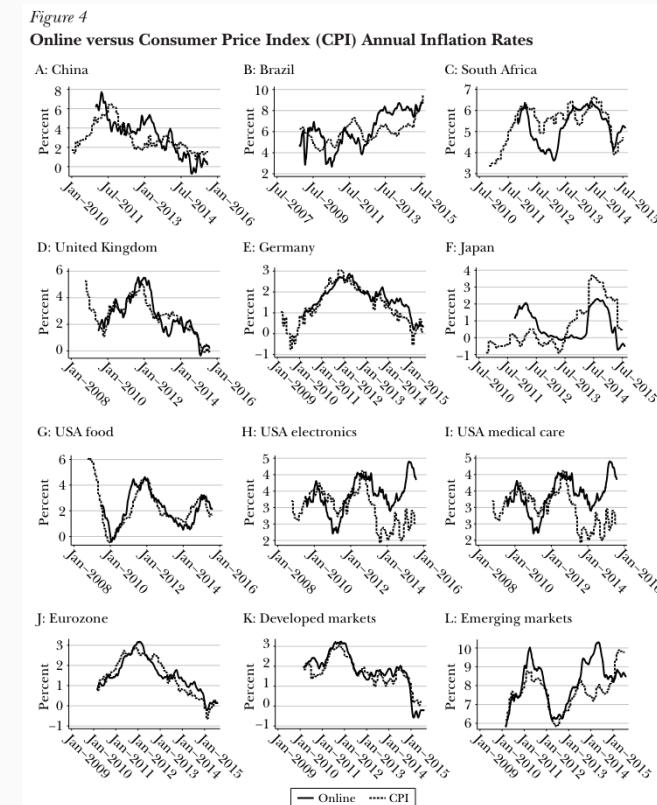
Journal of Economic Perspectives—Volume 30, Number 2—Spring 2016—Pages 151–178

The Billion Prices Project: Using Online Prices for Measurement and Research

Alberto Cavallo and Roberto Rigobon

New data-gathering techniques, often referred to as “Big Data,” have the potential to improve statistics and empirical research in economics. This paper presents one example of how this can be achieved by using the vast number of online prices displayed on the web. We describe our work with the Billion Prices Project at MIT, and emphasize key lessons that can be used for both inflation measurement and some fundamental research questions in macro and international economics. In particular, we show how online prices can be used to construct daily price indexes in multiple countries and to avoid measurement biases that distort evidence of price stickiness and international relative prices.

The basic procedure used in most countries to collect inflation data has remained roughly the same for decades. A large number of people working for national statistical offices visit hundreds of stores on a monthly or bimonthly basis to collect prices for a preselected basket of goods and services. The micro data are then processed and used to construct consumer price indexes and other related indicators. This process is expensive, complex, and often too slow for some users of the data. Infrequent sampling and slow updates to the baskets can complicate



Source: Authors using online price indexes computed by PriceStats and consumer price indexes sourced from the national statistical office in each country.

Notes: Figure 4 compares inflation as measured by online prices and by the offline prices in the official consumer price index for a selection of countries, sectors, and regions. Annual inflation rates for daily online price indexes are computed as the percentage change in the average of the previous 30 days compared to the same average 365 days before. The series are nonseasonally adjusted. Indexes are “all-items” with the exception of China, where an online supermarket index is shown next to the official food index. Global aggregates in the last row are computed using 2010 consumption weights in each country and CPIs from official sources.

See also: <https://thebillionpricesproject.com/> and <https://www.pricestats.com/>

The MIT Billion Prices Project

Journal of Economic Perspectives—Volume 30, Number 2—Spring 2016—Pages 151–178

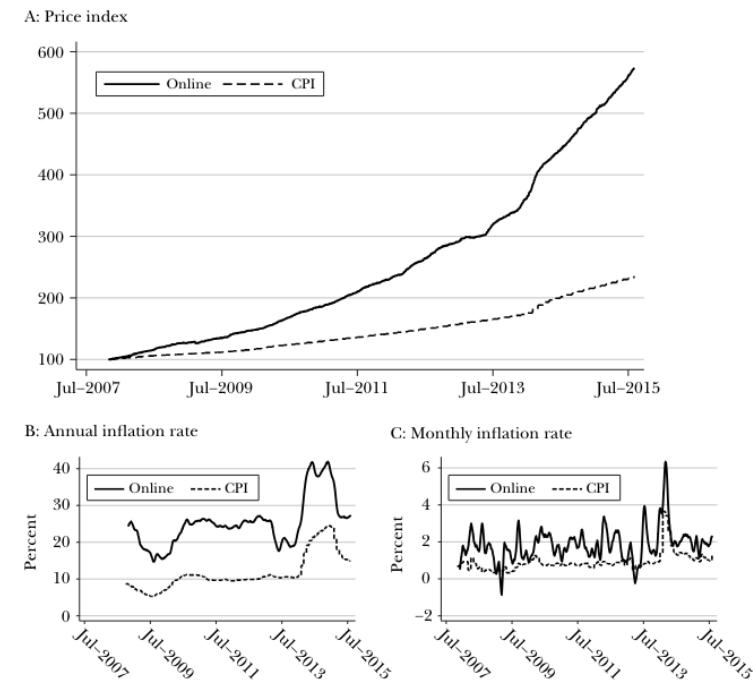
The Billion Prices Project: Using Online Prices for Measurement and Research

Alberto Cavallo and Roberto Rigobon

New data-gathering techniques, often referred to as “Big Data,” have the potential to improve statistics and empirical research in economics. This paper presents one example of how this can be achieved by using the vast number of online prices displayed on the web. We describe our work with the Billion Prices Project at MIT, and emphasize key lessons that can be used for both inflation measurement and some fundamental research questions in macro and international economics. In particular, we show how online prices can be used to construct daily price indexes in multiple countries and to avoid measurement biases that distort evidence of price stickiness and international relative prices.

The basic procedure used in most countries to collect inflation data has remained roughly the same for decades. A large number of people working for national statistical offices visit hundreds of stores on a monthly or bimonthly basis to collect prices for a preselected basket of goods and services. The micro data are then processed and used to construct consumer price indexes and other related indicators. This process is expensive, complex, and often too slow for some users of the data. Infrequent sampling and slow updates to the baskets can complicate

Figure 1
Argentina



Source: Authors using online price index computed by PriceStats and the consumer price index from the national statistical office in Argentina (INDEC).

Notes: The figure compares a price index produced with online data to a comparable official consumer price index (CPI) for the case of Argentina from 2007 to 2015. It also looks at annual and monthly inflation rates using each source of data. Monthly inflation rates for the online index are computed as the percentage change in the average of the previous 30 days compared to the same average a month before. Annual inflation rates for the online index are computed as the percentage change in the average of the previous 30 days compared to the same average 365 days before. All price indexes are nonseasonally adjusted.

The COMPAS algorithm to predict criminals' recidivism

Background

- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a decision support tool developed by Northpointe (now Equivant) used by U.S. courts to **assess the likelihood of recidivism**
- Produced several scales (Pretrial release risk, General recidivism, Violent recidivism) based on factors such as age, criminal history, and substance abuse
- The algorithm is proprietary and its inner workings are not public

Source [Practitioner's Guide to COMPAS Core](#)

Practitioner's Guide to COMPAS Core

The Practitioner's Guide provides an overview of the COMPAS Core Module in the Northpointe Suite. The Northpointe Suite is an integrated web-based assessment and case management system for criminal justice practitioners. The Northpointe Suite has modules designed for pretrial, jail, probation, prison, parole and community corrections applications. COMPAS Core is designed for both male and female offenders recently removed from the community or currently in the community. The Practitioner's Guide to COMPAS Core covers case interpretation, validity and reliability, and treatment implications. Most of the information provided is specific to COMPAS Core. Throughout this text we use the term COMPAS Core to distinguish an element (scale, typology, decile type) specific to COMPAS Core from general elements in the Northpointe Suite, such as scales found in both COMPAS Core and COMPAS Reentry.

COMPAS is a fourth generation risk and needs assessment instrument. Criminal justice agencies across the nation use COMPAS to inform decisions regarding the placement, supervision and case management of offenders. COMPAS was developed empirically with a focus on predictors known to affect recidivism. It includes dynamic risk factors, and it provides information on a variety of well validated risk and needs factors designed to aid in correctional intervention to decrease the likelihood that offenders will reoffend.

COMPAS was first developed in 1998 and has been revised over the years as the knowledge base of criminology has grown and correctional practice has evolved. In many ways changes in the field have followed new developments in risk assessment. We continue to make improvements to COMPAS based on results from norm studies and recidivism studies conducted in jails, probation agencies, and prisons. COMPAS is periodically updated to keep pace with emerging best practices and technological advances.

In overloaded and crowded criminal justice systems, brevity, efficiency, ease of administration and clear organization of key risk/needs data are critical. COMPAS was designed to optimize these practical factors. We acknowledge the trade-off between comprehensive coverage of key risk and criminogenic factors on the one hand, and brevity and practicality on the other. COMPAS deals with this trade-off in several ways; it provides a comprehensive set of key risk factors that have emerged from the recent criminological literature, and it allows for customization inside the software. Therefore, ease of use, efficient and effective time management, and case management considerations that are critical to best practice in the criminal justice field can be achieved through COMPAS.

The COMPAS algorithm to predict criminals' recidivism

The ProPublica and other investigations

- In 2016, ProPublica published an investigation showing that COMPAS was **biased against African Americans**
- **Bias:** The algorithm was more likely for African Americans to wrongly predict that defendants would re-offend.
- **Accuracy:** only 20% of people predicted to commit violent crimes actually went on to do so (in a later study estimated with 65%, still worse than a group of humans with little expertise)

Source ProPublica 2016

Machine Bias*

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner

There's software used across the country to predict future criminals. And it's biased against blacks.

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances—which belonged to a 6-year-old boy—a woman came running after them saying, “That’s my kid’s stuff.” Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late—a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of \$80.

Compare their crime with a similar one: The previous summer, 41-year-old Vernon Prater was picked up for shoplifting \$86.35 worth of tools from a nearby Home Depot store (Figure 6.1.1).

Prater was the more seasoned criminal. He had already been convicted of armed robbery and attempted armed robbery, for which he served five years in prison, in addition to another armed robbery charge. Borden had a record, too, but it was for misdemeanors committed when she was a juvenile.

Yet something odd happened when Borden and Prater were booked into jail: A computer program spat out a score predicting the likelihood of each committing a future crime. Borden—who is black—was rated a high risk. Prater—who is white—was rated a low risk.

Two years later, we know the computer algorithm got it exactly backward. Borden has not been charged with any new crimes. Prater is serving an eight-year prison term for subsequently breaking into a warehouse and stealing thousands of dollars’ worth of electronics.

Scores like this—known as risk assessments—are increasingly common in courtrooms across the nation. They are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts—as is the case in Fort Lauderdale—to

* Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner, “Machine Bias,” *ProPublica* (May 23, 2016). Reprinted with permission.

The COMPAS algorithm to predict criminals' recidivism

The ProPublica and other investigations

- In 2016, ProPublica published an investigation showing that COMPAS was **biased against African Americans**
- **Bias:** The algorithm was more likely for African Americans to wrongly predict that defendants would re-offend.
- **Accuracy:** only 20% of people predicted to commit violent crimes actually went on to do so (in a later study estimated with 65%, still worse than a group of humans with little expertise)

Source Dressel and Fair, 2018, Science Advances

SCIENCE ADVANCES | RESEARCH ARTICLE

RESEARCH METHODS

The accuracy, fairness, and limits of predicting recidivism

Julia Dressel and Hany Farid*

Algorithms for predicting recidivism are commonly used to assess a criminal defendant's likelihood of committing a crime. These predictions are used in pretrial, parole, and sentencing decisions. Proponents of these systems argue that big data and advanced machine learning make these analyses more accurate and less biased than humans. We show, however, that the widely used commercial risk assessment software COMPAS is no more accurate or fair than predictions made by people with little or no criminal justice expertise. In addition, despite COMPAS's collection of 137 features, the same accuracy can be achieved with a simple linear predictor with only two features.

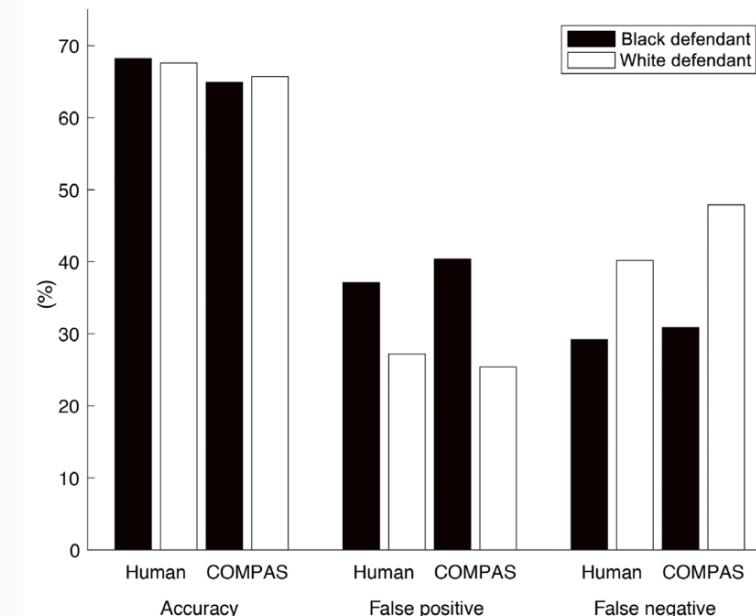


Fig. 1. Human (no-race condition) versus COMPAS algorithmic predictions (see also Table 1).

The Meta US 2020 Election study

How do social media feed algorithms affect attitudes and behavior in an election campaign?

Andrew M. Guess^{1*}, Neil Malhotra², Jennifer Pan³, Pablo Barberá⁴, Hunt Allcott⁵, Taylor Brown⁴, Adriana Crespo-Tenorio⁴, Drew Dimmery^{4,6}, Deen Freelon⁷, Matthew Gentzkow⁸, Sandra González-Bailón⁹, Edward Kennedy¹⁰, Young Mie Kim¹¹, David Lazer¹², Devra Moehler⁴, Brendan Nyhan¹³, Carlos Velasco Rivera⁴, Jaime Settle¹⁴, Daniel Robert Thomas⁴, Emily Thorson¹⁵, Rebekah Tromble¹⁶, Arjun Wilkins⁴, Magdalena Wojcieszak^{17,18}, Beixian Xiong⁴, Chad Kiewiet de Jonge⁴, Annie Franco⁴, Winter Mason⁴, Natalie Jomini Stroud¹⁹, Joshua A. Tucker²⁰

We investigated the effects of Facebook's and Instagram's feed algorithms during the 2020 US election. We assigned a sample of consenting users to reverse-chronologically-ordered feeds instead of the default algorithms. Moving users out of algorithmic feeds substantially decreased the time they spent on the platforms and their activity. The chronological feed also affected exposure to content: The amount of political and untrustworthy content they saw increased on both platforms, the amount of content classified as uncivil or containing slur words they saw decreased on Facebook, and the amount of content from moderate friends and sources with ideologically mixed audiences they saw increased on Facebook. Despite these substantial changes in users' on-platform experience, the chronological feed did not significantly alter levels of issue polarization, affective polarization, political knowledge, or other key attitudes during the 3-month study period.

Reshares on social media amplify political news but do not detectably affect beliefs or opinions

Andrew M. Guess^{1*}, Neil Malhotra², Jennifer Pan³, Pablo Barberá⁴, Hunt Allcott⁵, Taylor Brown⁴, Adriana Crespo-Tenorio⁴, Drew Dimmery^{4,6}, Deen Freelon⁷, Matthew Gentzkow⁸, Sandra González-Bailón⁹, Edward Kennedy¹⁰, Young Mie Kim¹¹, David Lazer¹², Devra Moehler⁴, Brendan Nyhan¹³, Carlos Velasco Rivera⁴, Jaime Settle¹⁴, Daniel Robert Thomas⁴, Emily Thorson¹⁵, Rebekah Tromble¹⁶, Arjun Wilkins⁴, Magdalena Wojcieszak^{17,18}, Beixian Xiong⁴, Chad Kiewiet de Jonge⁴, Annie Franco⁴, Winter Mason⁴, Natalie Jomini Stroud¹⁹, Joshua A. Tucker²⁰

We studied the effects of exposure to reshared content on Facebook during the 2020 US election by assigning a random set of consenting, US-based users to feeds that did not contain any reshares over a 3-month period. We find that removing reshared content substantially decreases the amount of political news, including content from untrustworthy sources, to which users are exposed; decreases overall clicks and reactions; and reduces partisan news clicks. Further, we observe that removing reshared content produces clear decreases in news knowledge within the sample, although there is some uncertainty about how this would generalize to all users. Contrary to expectations, the treatment does not significantly affect political polarization or any measure of individual-level political attitudes.

The Meta US 2020 Election study

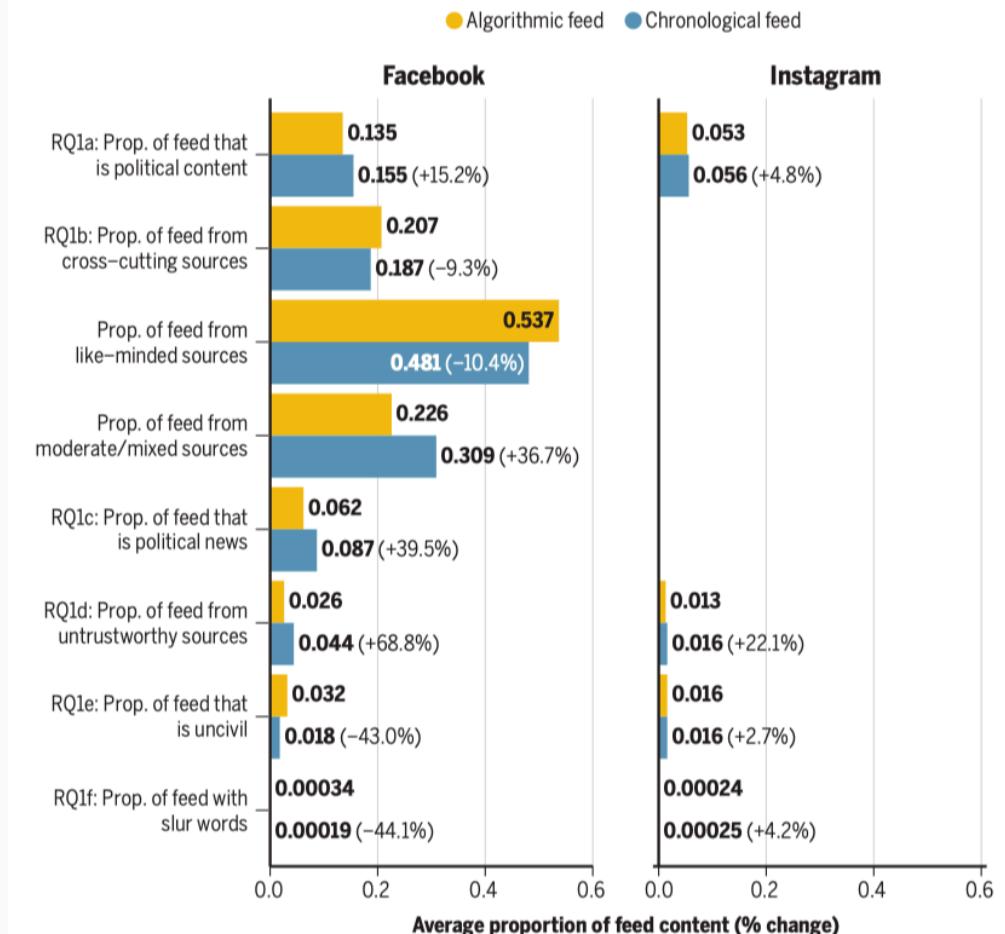


Fig. 2. Estimated changes in prevalence of feed content on both Facebook and Instagram. (Left)
Facebook. (Right) Instagram. Values are average unweighted proportions within each group, with percent changes relative to the Algorithmic Feed control group in parentheses. All differences are significant at the $p < 0.005$ level, except RQ1f for Instagram ($p < 0.05$); confidence intervals are thus not shown. RQ1b and RQ1c were not tested for Instagram because political and ideology classifications are not available on that platform. Fully specified regression models with survey weights are reported in the SM, section S2.2.

Source Guess et al. 2023, Science

The Meta US 2020 Election study

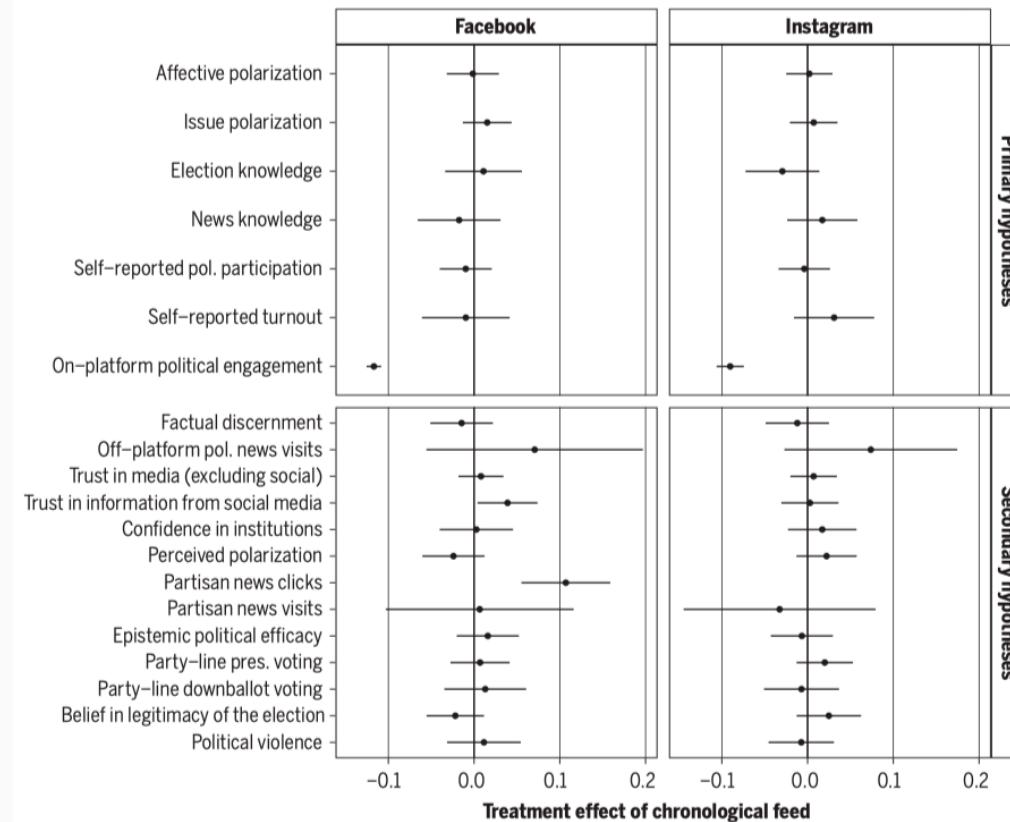


Fig. 3. Population average treatment effects of the Chronological Feed, relative to the Algorithmic Feed control group, on both Facebook and Instagram. (Left) Facebook. (Right) Instagram. Estimates are presented in standard deviations with 95% confidence intervals (not adjusted for multiple comparisons). Partisan news clicks are estimated only for Facebook because source-level estimates of political ideology are not available for Instagram. pol., political; pres., presidential.

Class logistics

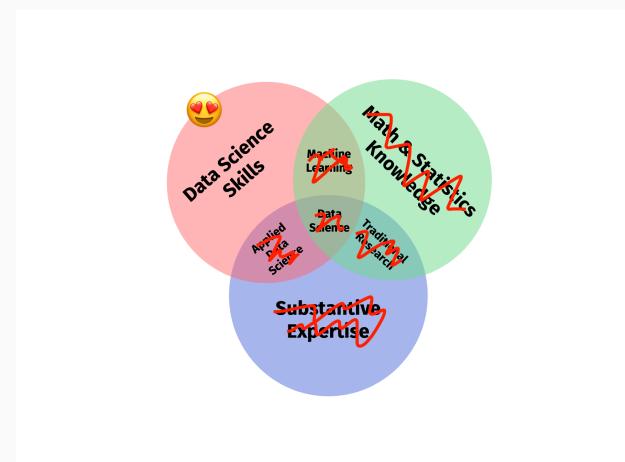
Introduction to Data Science in a nutshell

Session	Session Title	A	Date
<u>Fundamentals</u>			
0	R and the tidyverse	-	-
1	What is data science?	H	September 02
2	Version control and project management	Q	September 09
3	Programming: Functions and debugging	H	September 16
4	Relational databases and SQL	Q	September 23
<u>Collecting and wrangling data</u>			
5	Web data and technologies	Q	September 30
6	Web scraping and APIs	H	October 07
7	Modeling	Q	October 14
Mid-term Exam Week: no class			
<u>Analyzing data</u>			
8	Workshop: Tools for Data Science	-	October 28
9	Visualization	H	November 04
10	Automation, scheduling, and packages	Q	November 11
<u>Fine-tuning the workflow</u>			
11	Monitoring and communication	H	November 18
12	Data science ethics	-	November 25
Final Exam Week: no class			

The plan

Goals of the course

- This course equips you with conceptual knowledge about the data science pipeline and coding workflow, data structures, and data wrangling.
- It enables you to apply this knowledge with statistical software.
- It prepares you for our other core courses and electives as well as the master's thesis.



What we will cover

- Version control and project management
- R and the tidyverse
- Programming workflow: debugging, automation, packaging
- Relational databases and SQL
- Web data and technologies
- Model fitting and evaluation
- Visualization
- Monitoring and communication
- Data science ethics
- (The command line)

You at the beginning of the course



You at the end of the course



Why R?

Why R and RStudio?

Data science positivism

- Alongside Python, R has become the *de facto* language for data science.
 - See: *The Impressive Growth of R*, *The Popularity of Data Science Software*
- Open-source (free!) with a global user-base spanning academia and industry.
 - "Do you want to be a profit source or a cost center?"

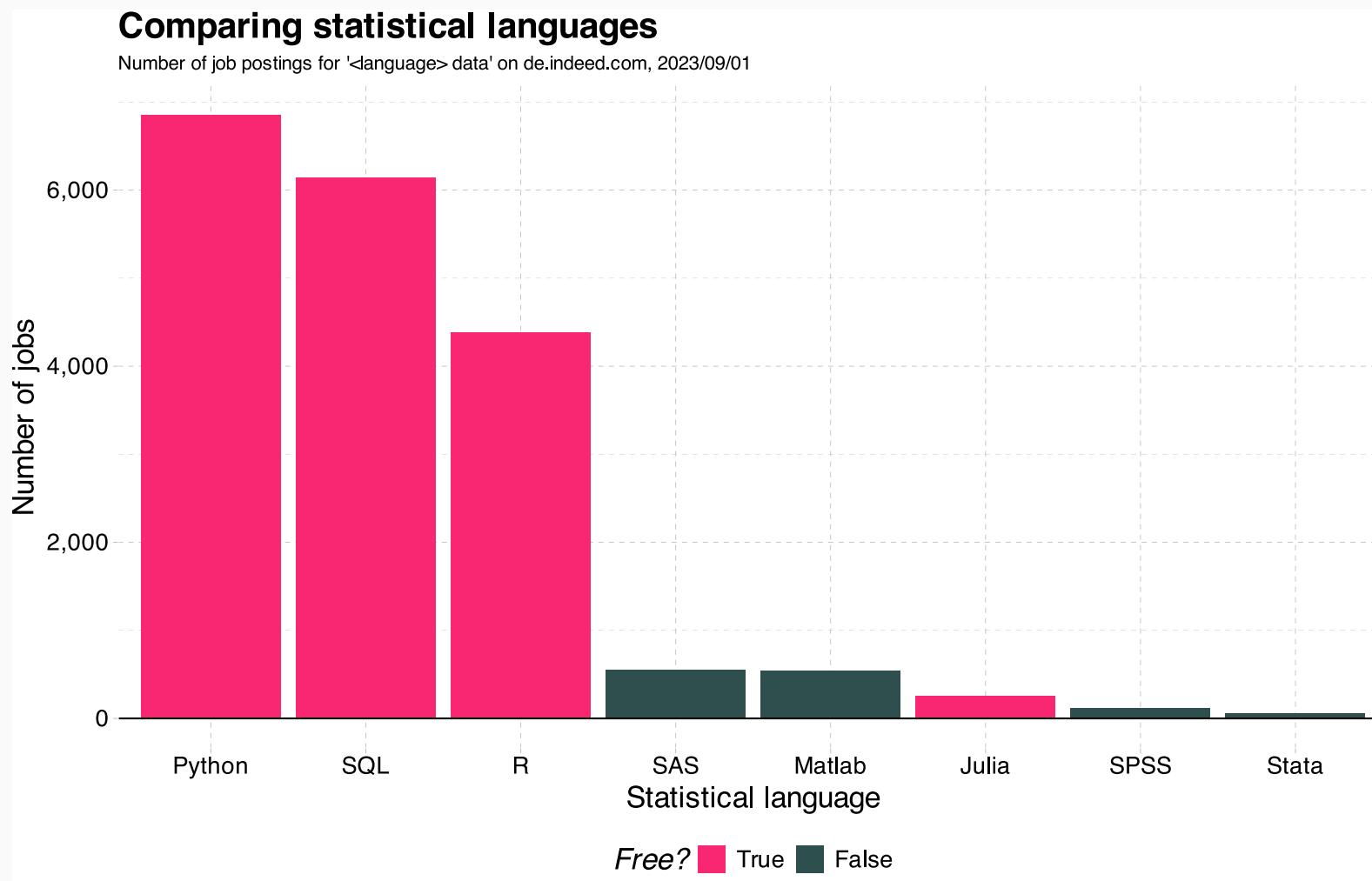
Bridge to multiple other programming environments, with statistics at heart

- Already has all of the statistics support, and is amazingly adaptable as a “glue” language to other programming languages and APIs.
- The RStudio IDE and ecosystem allow for further, seamless integration.

Path dependency

- It's also the language that I know best.
- (Learning multiple languages is a good idea, though.)

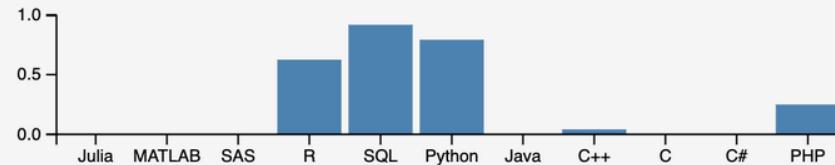
Why R and RStudio? (cont.)



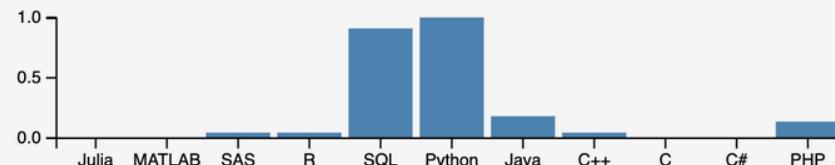
Why R and RStudio? (cont.)

Which programming language is required for a ...

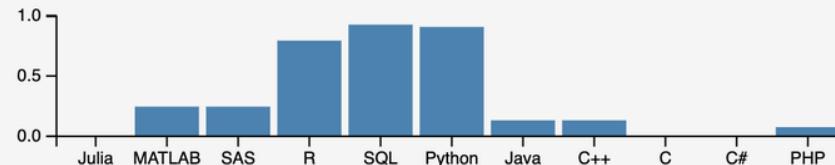
Data Analyst



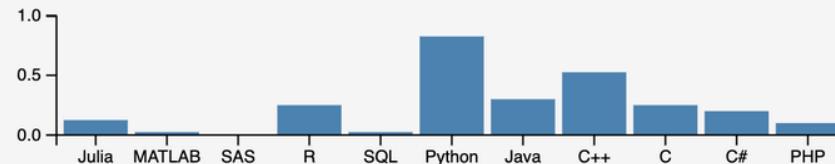
Data Engineer



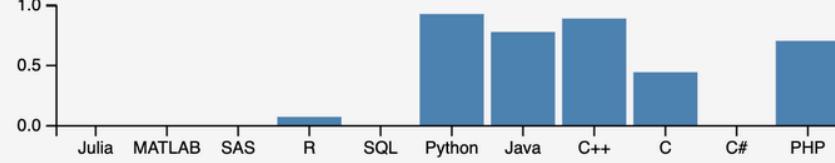
Data Scientist



Research Scientist

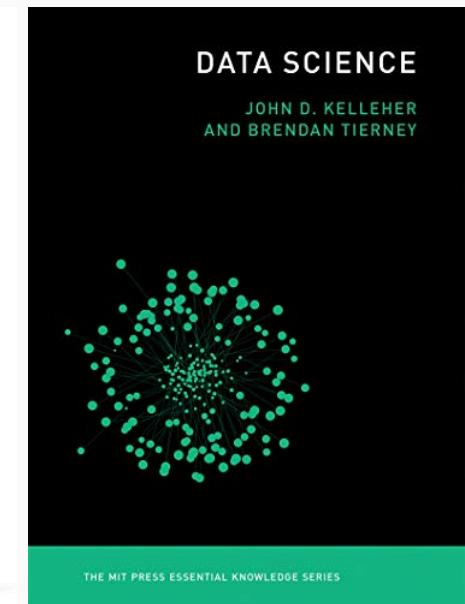
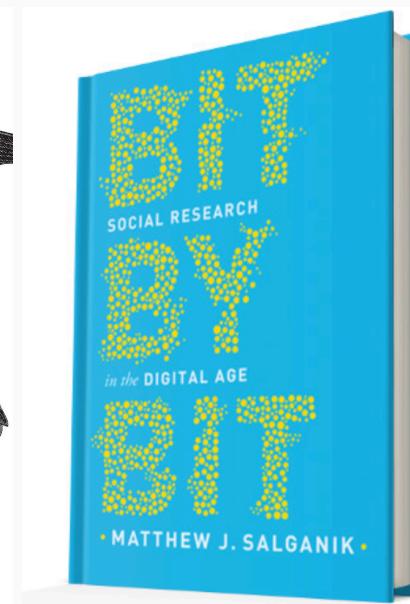
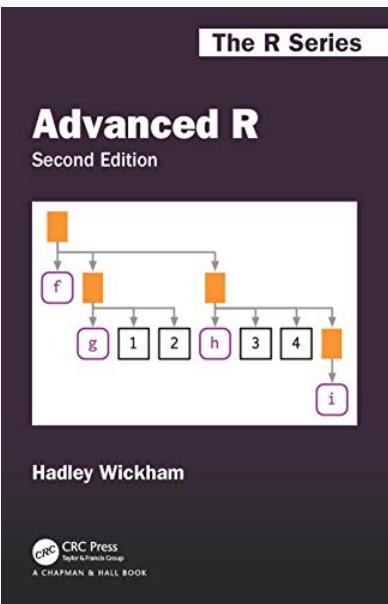
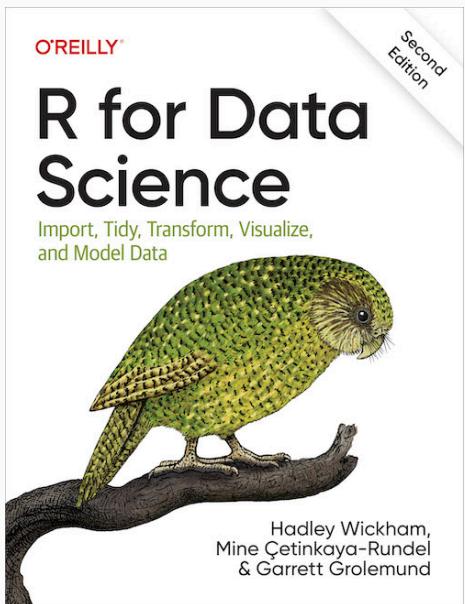


ML Engineer



Each bar represents the proportion of job post that specify this language as an optional requirement for that role.

Core (and optional) readings



This semester

 Hadley Wickham 
@hadleywickham

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

928 Retweets 1,113 Likes 

 40  928  1.1K

OUR MINDSET FOR THE COURSE THIS SEMESTER!

Attendance

General rules

- You cannot miss more than two sessions. If you have to miss a session for medical reasons or personal emergencies, please **inform Examination Office** and they will inform me about your absence. There is no need to notify me in advance or ex post.
- We will check attendance on-site.
- The current **Hertie hygiene rules** apply!

Office hours and advice

- If you want to discuss content from class, please first do so in the lab sessions.
- If you still need more feedback on course topics, use the Moodle forum.
- If you want to discuss any other matters with me, drop Alex Karras, my assistant, a message ( karras@hertie-school.org) and he will arrange a meeting.
- For general technical advice, the [Research Consulting Team at the Data Science Lab](#) is there for you.

Assignments and grading

Component	Weight
4(5) × homework assignments (10% each)	40%
4(5) × online quizzes (5% each)	20%
1 × workshop presentation/attendance	10%
1 × hackathon project	30%

Homework assignments

- The assignments are distributed via our own [GitHub Classroom](#).
- Each assignment is a mix of practical problems that are to be solved with R.
- You are encouraged to collaborate, but everyone will hand in a separate solution.
- There will be 5 assignments (one every ~2 weeks; see [overview on GitHub](#)) and the 4 best will contribute to the final grade.
- You'll have one week to work on each assignment (deadline: Tuesdays at 9:30am).
- You submit your solutions via GitHub.

Assignments and grading

Component	Weight
4(5) × homework assignments (10% each)	40%
4(5) × online quizzes (5% each)	20%
1 × workshop presentation/attendance	10%
1 × hackathon project	30%

Homework assignments

- Grades will be based on (1) the accuracy of your solutions and (2) the adherence of a clean and efficient coding style.
- Feedback will be verbal:
 - Excellent (95+)
 - Very good (90-94)
 - Good (85-89)
 - OK (80-84)
 - Acceptable (75-79)
 - Definitely needs improvement (below 75)

Assignments and grading

Component	Weight
4(5) × homework assignments (10% each)	40%
4(5) × online quizzes (5% each)	20%
1 × workshop presentation/attendance	10%
1 × hackathon project	30%

Online quizzes

- The short online quizzes will test your knowledge about the topics covered in class.
- There will be 5 quizzes and the 4 best will contribute to the final grade.
- You'll have one week to work on each assignment (deadline: Tuesdays at 9:30am).

Assignments and grading

Component	Weight
4(5) × homework assignments (10% each)	40%
4(5) × online quizzes (5% each)	20%
1 × workshop presentation/attendance	10%
1 × hackathon project	30%

Workshop presentation (MDS students)

- On October 28, 14-20h, we will flip roles and you will become instructor of a data science workshop session.
- You, in groups of 2 students, will present a data science workflow tool (randomly **allocated**).
- Your contribution will include:
 1. A lightning talk (recorded) where you briefly introduce and motivate the tool
 2. A hands-on session where you showcase the tool and provide practice material
- Both the recorded talk and the materials will be graded.
- Check out the materials from previous workshops online >2021< >2022< >2023<!
- **MPP/MIA students:** You will not give a talk, but have to actively participate in the workshop.

Assignments and grading

Component	Weight
4(5) × homework assignments (10% each)	40%
4(5) × online quizzes (5% each)	20%
1 × workshop presentation/attendance	10%
1 × hackathon project	30%

Hackathon project

- On December 2, 13-16h, there will be a hackathon hosted at Hertie.
- At the hackathon itself, we introduce the data and provide an environment that should facilitate you getting started with the project and form groups of 3-4 students.
- Two weeks later, on December 16, the project instructions will be made available. You will then have 48 hours to submit your solutions.
- The task is similar to the homework assignments but puts more emphasis on creative problem-solving using the tools and techniques you have learned in class.

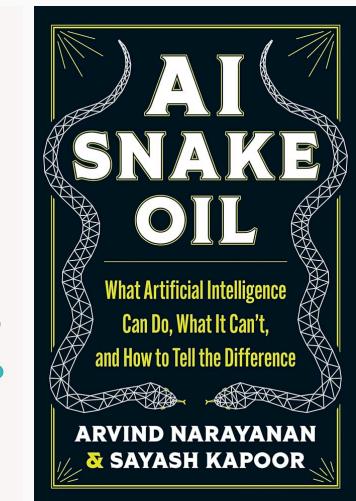
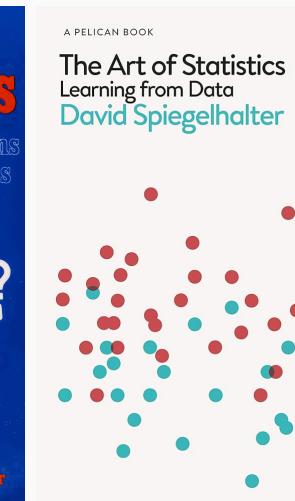
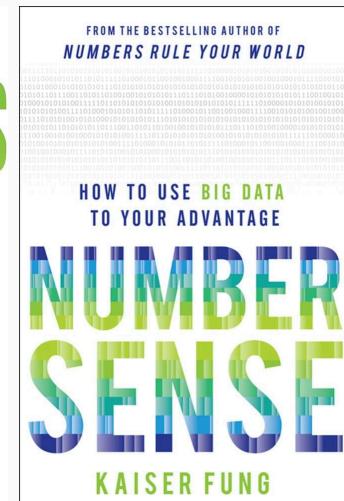
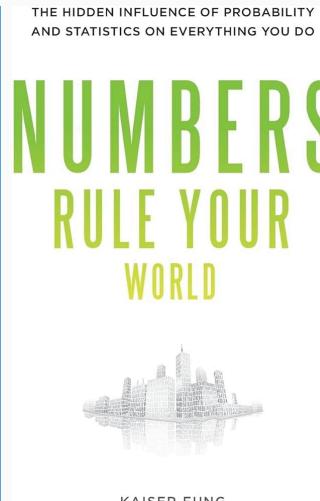
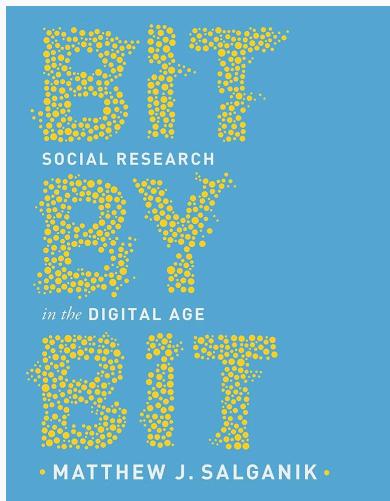
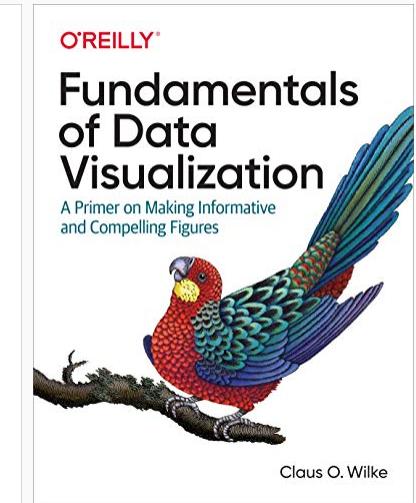
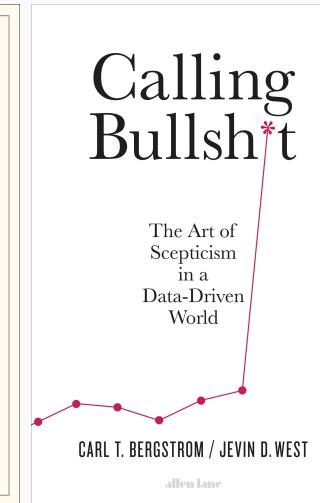
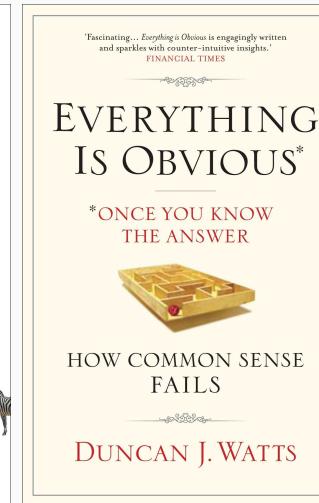
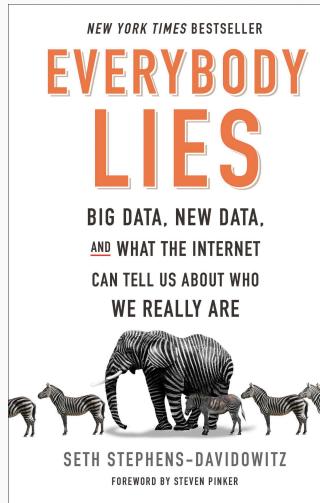
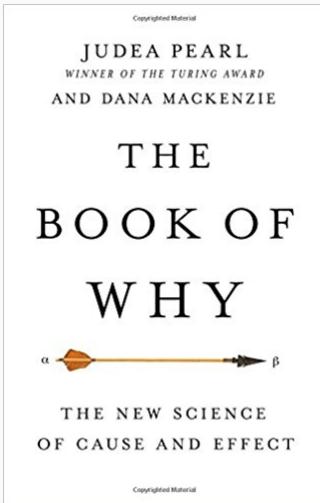
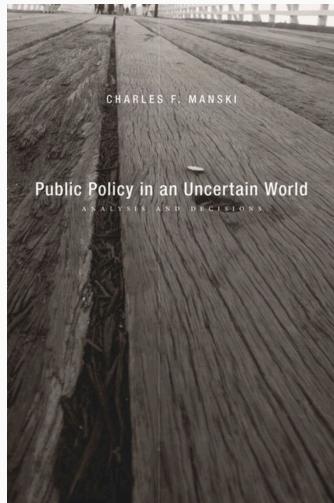
AI use in and for the course

Can AI tools (LLM interfaces, AI pair programming) be used for assignments?

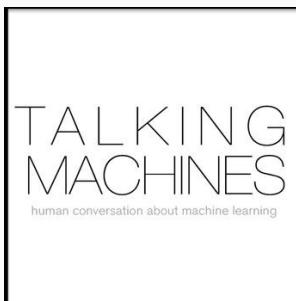
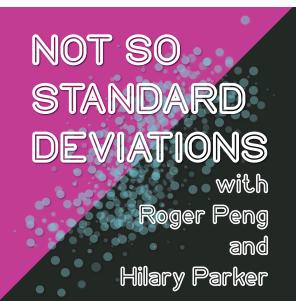
- Yes, but use them with care. You will not become an efficient programmer if you heavily rely on those tools without learning the basics.
- The Hertie School has installed [teaching guidelines on the use of AI Tools](#) in Spring 2023. We will stick to those guidelines.
- Some key elements from the guidelines:
 - "Familiarity with AI tools is helpful for the learning experience and the professional development of students afterwards, ..."
 - "... but needs to be done with clear guidelines on ethical use, biases, and limits of the tools that are currently available."
 - "[T]he use of AI tools for the preparation of assignments (...) needs to be clearly referenced in the text."



Further reading



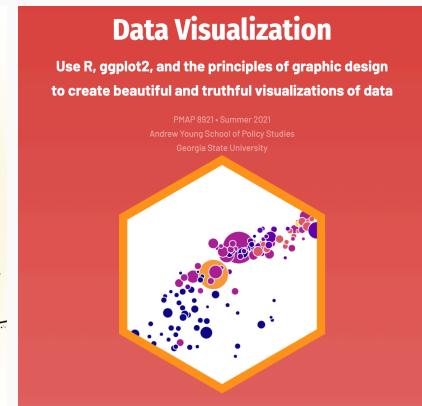
Further listening



Further watching



3Blue1Brown



Online
Causal
Inference
Seminar



Coming up

The first lab session

Get to know Carmen, Sebastian, R, and RStudio, four of your best friends for the next months!

Next lecture

Version control and project management