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-21

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 is for everything else.

Introductions

Course

🌐 https://github.com/intro-to-data-science-21

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 is for everything else.

Me

📇 I'm Simon Munzert [si'mən munsɜrt], or just Simon [saɪmən].

✉ munzert@hertie-school.org

🎓 Assistant Professor (data science and public policy)

Introductions

Course

🌐 https://github.com/intro-to-data-science-21

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 is for everything else.

Me

📇 I'm Simon Munzert [si'mən munsɜrt], or just Simon [saɪmən].

✉ munzert@hertie-school.org

🎓 Assistant Professor (data science and public policy)

You

What's your name? Why are you here? And would you share a funny fact about yourself?

The labs

Who & how

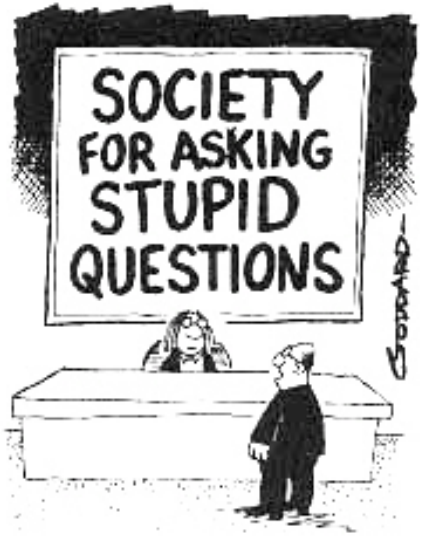
- This course is accompanied by labs administered by **Lisa Oswald** and **Tom Arend**.
- The labs are mandatory. Please attend them.
- As with the regular classes, please stick to the lab you are assigned to.

What for

- What these sessions are meant for:
  - Discussion of issues related to the content covered in the lecture
  - Discussion of issues related to the assignments
  - Boosting your R skills
- What these sessions are **not** meant for:
  - Solving the assignments for you

Class etiquette

- Learning programming with R can be challenging and might lead you out of your comfort zone. If you have problems with the pace of the course, let me 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. We are differentially affected by it. We are located in different time zones. **Let's support each other.**
- **Be respectful** to each other, all the time.
- **Ask questions** whenever you feel the need to do so!



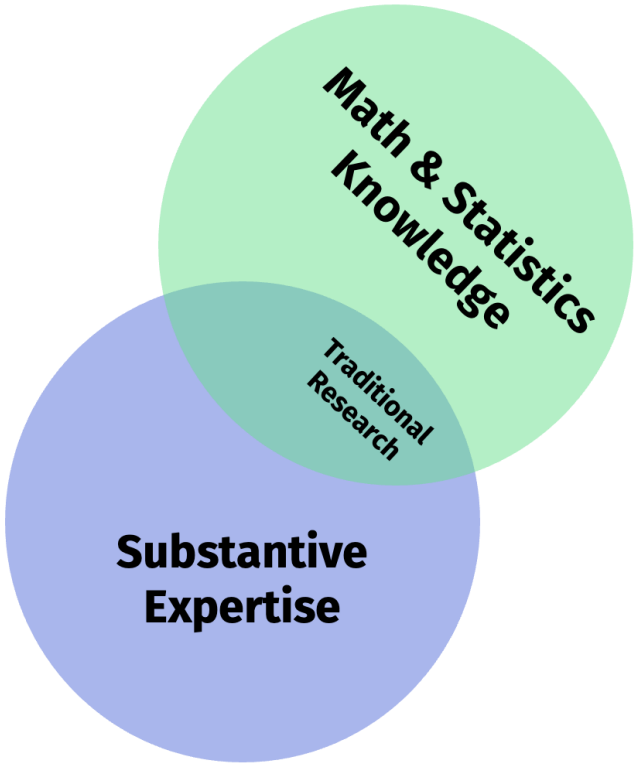
"Excuse me, but is this The Society for Asking Stupid Questions?"

Table of contents

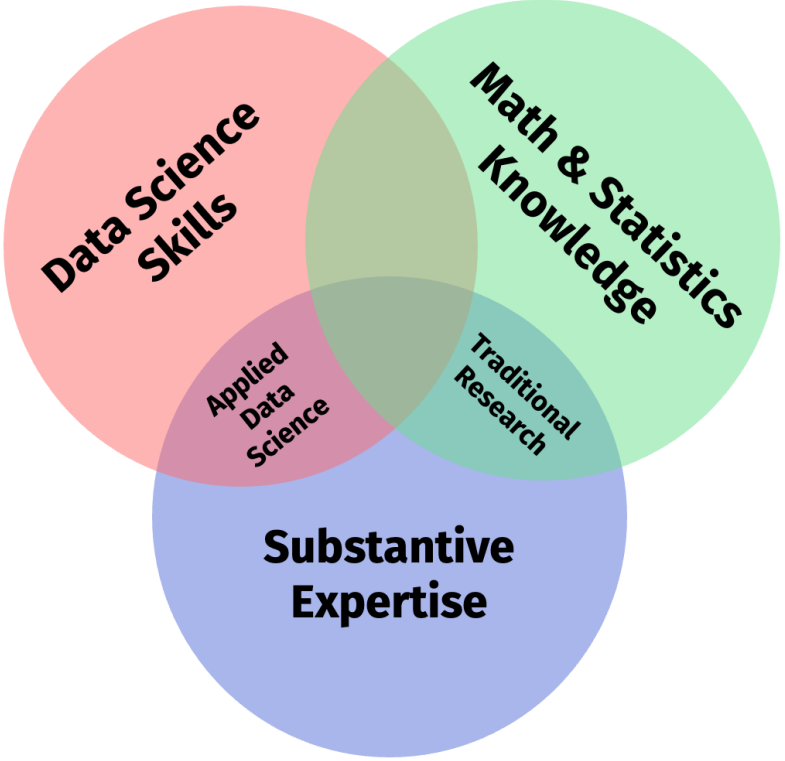
What is data science?

An old classic

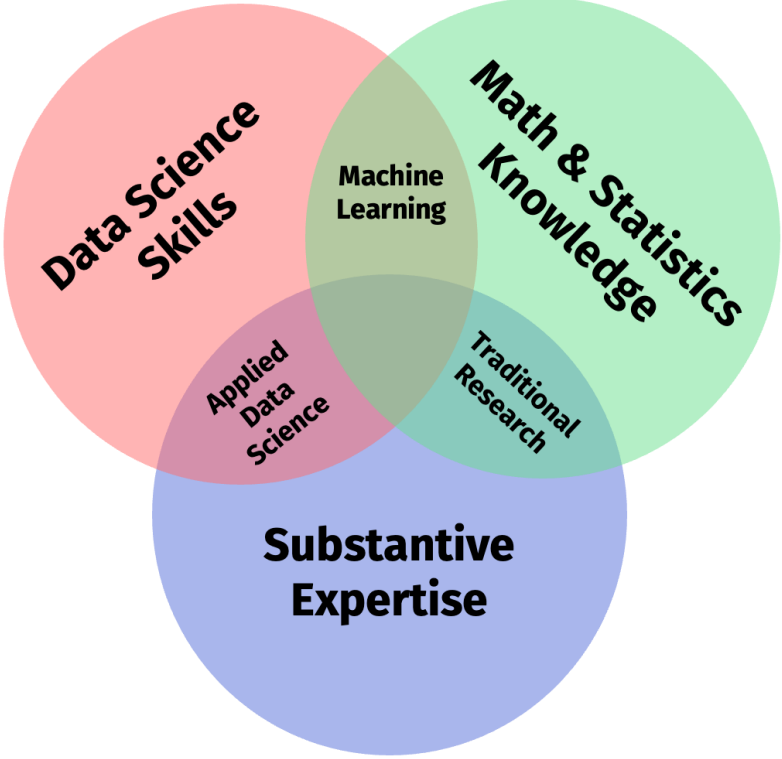
An old classic



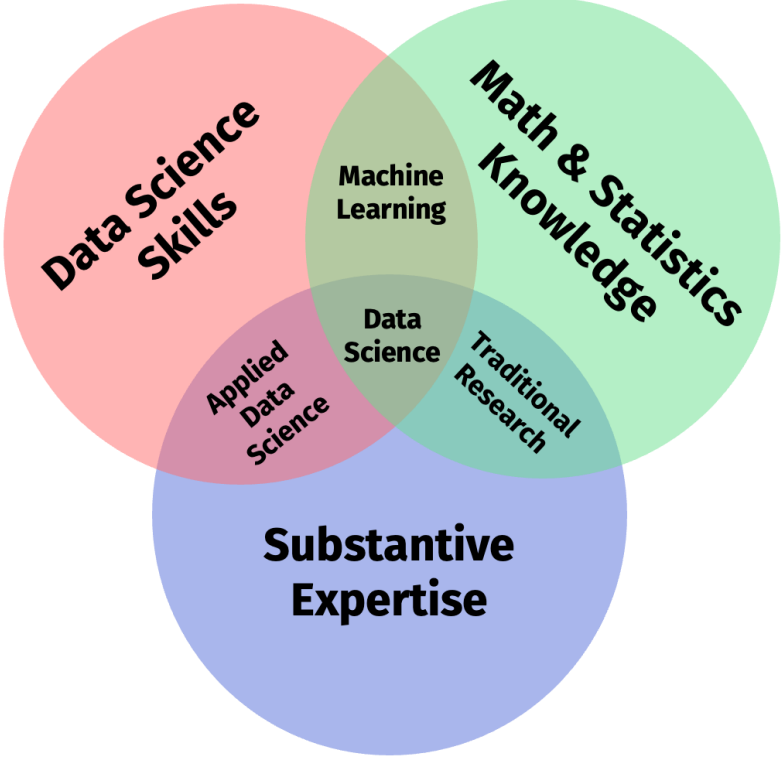
An old classic



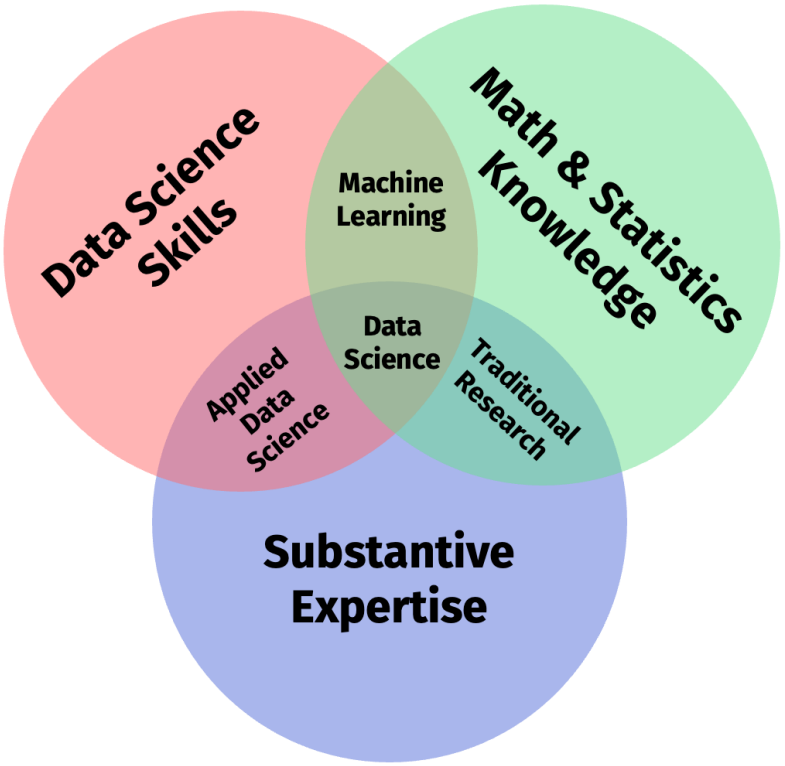
An old classic



An old classic



An old classic



© Drew Conway
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

The data science pipeline

The data science pipeline

Preparatory work

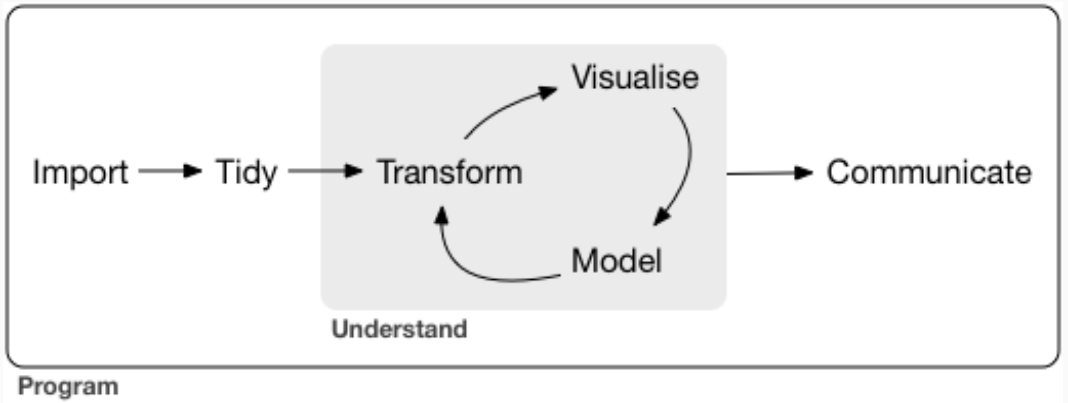
- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

The data science pipeline

Preparatory work

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

Data operation



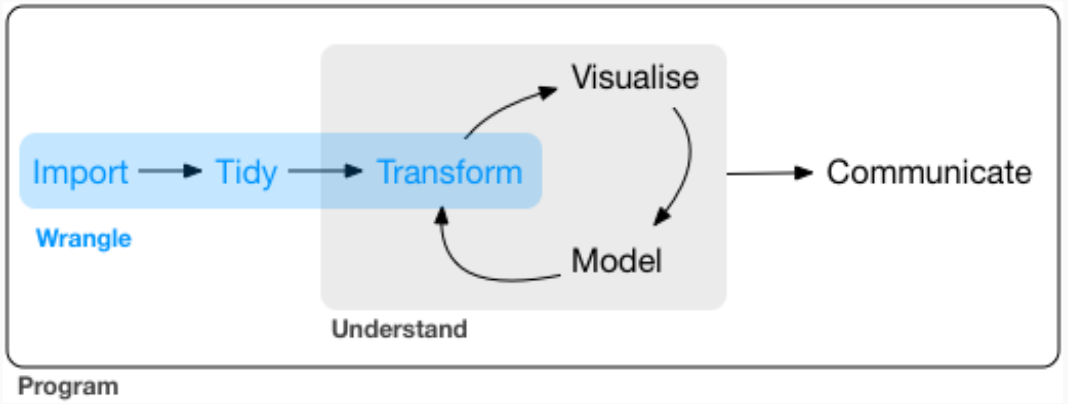
The data science pipeline

Preparatory work

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

Data operation

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



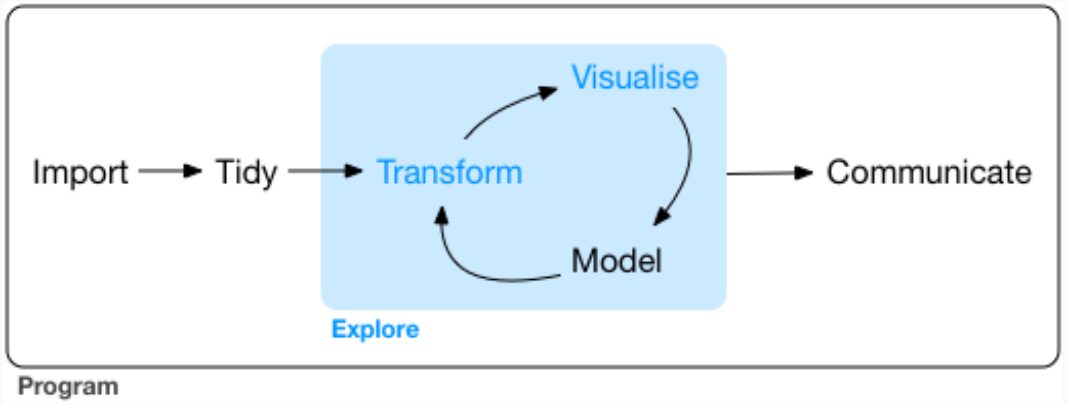
The data science pipeline

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



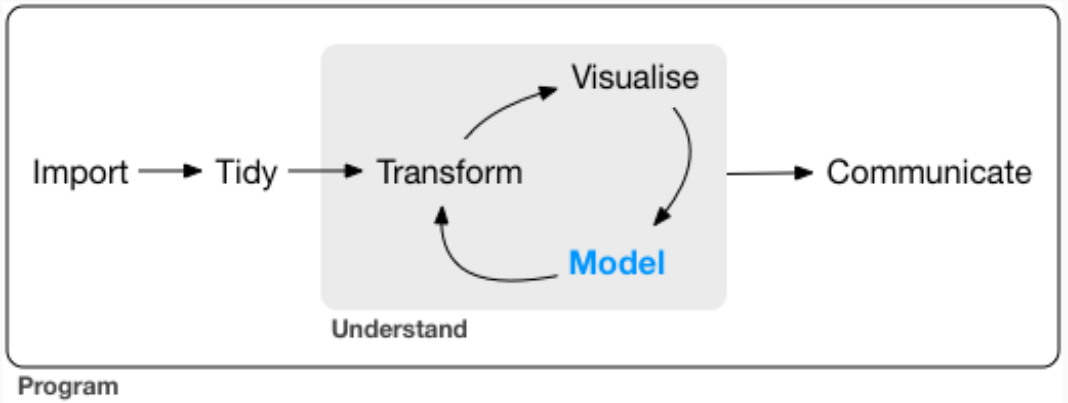
The data science pipeline

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

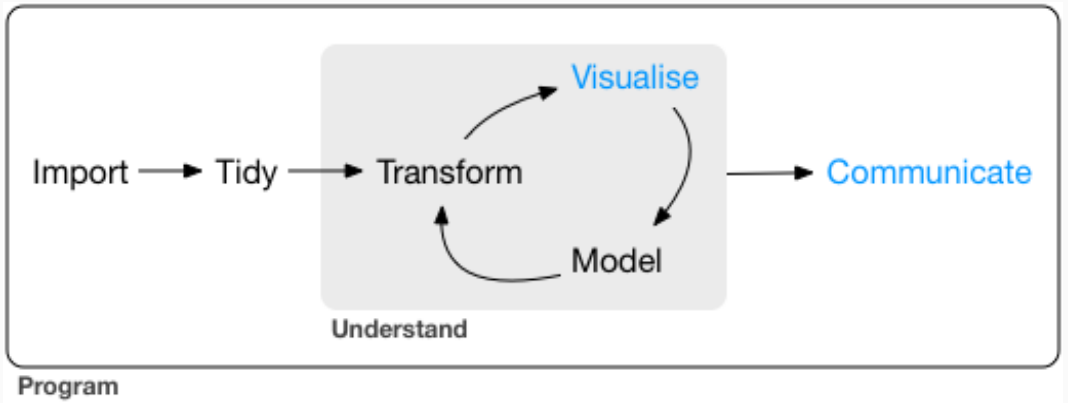
- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

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

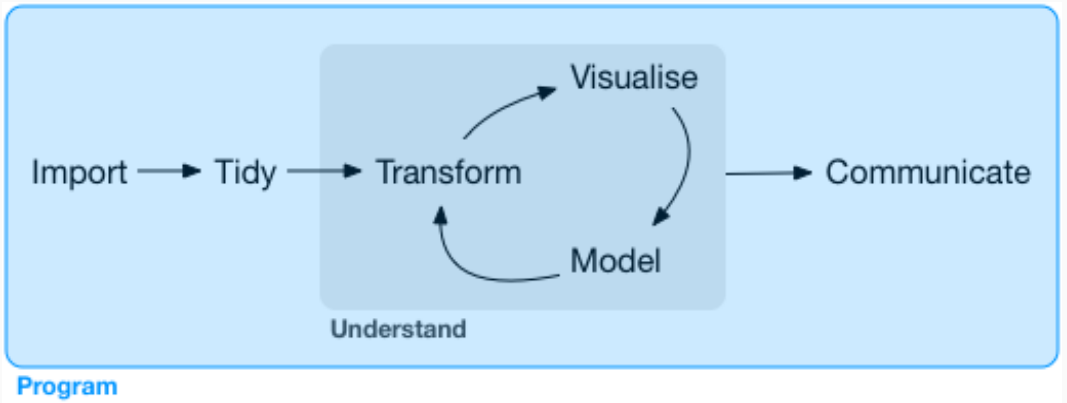
- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

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



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

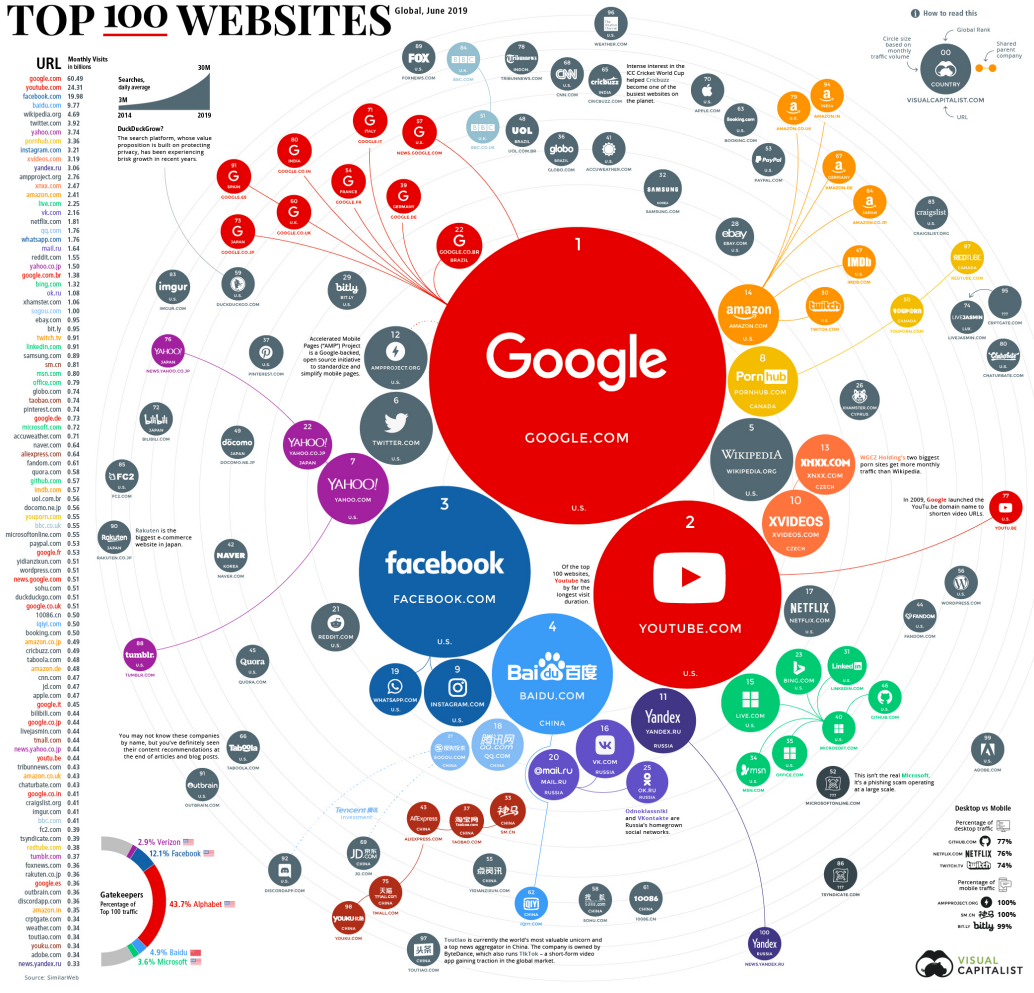
Sneak preview

Sneak preview
Learning to love a programming environment

The tidyverse

Sneak preview
Collecting web data at scale

Most popular websites 2019



Scraping the web for social research

How Censorship in China Allows Government Criticism but Silences Collective Expression

GARY KING *Harvard University*
JENNIFER PAN *Harvard University*
MARGARET E. ROBERTS *Harvard University*

We offer the first large scale, multiple source analysis of the outcome of what may be the most extensive effort to selectively censor human expression ever implemented. To do this, we have devised a system to locate, download, and analyze the content of millions of social media posts originating from nearly 1,400 different social media services all over China before the Chinese government is able to find, evaluate, and censor (i.e., remove from the Internet) the subset they deem objectionable. Using modern computer-assisted text analytic methods that we adapt to and validate in the Chinese language, we compare the substantive content of posts censored to those not censored over time in each of 85 topic areas. Contrary to previous understandings, posts with negative, even vitriolic, criticism of the state, its leaders, and its policies are not more likely to be censored. Instead, we show that the censorship program is aimed at curtailing collective action by silencing comments that represent, reinforce, or spur social mobilization, regardless of content. Censorship is oriented toward attempting to forestall collective activities that are occurring now or may occur in the future—and, as such, seem to clearly expose government intent.

British Journal of Political Science (2021), page 1 of 11
doi:10.1017/S0007123420000897

British Journal of Political Science

LETTER

The Comparative Legislators Database

Sascha Göbel[1]* and Simon Munzert[2]

[1]Faculty of Social Sciences, Goethe University Frankfurt am Main, Germany; and [2]Data Science Lab, Hertie School, Berlin, Germany
*Corresponding author. E-mail: sascha.goebel@soz.uni-frankfurt.de

(Received 7 June 2020; revised 12 November 2020; accepted 2 December 2020)

Abstract
Knowledge about political representatives' behavior is crucial for a deeper understanding of politics and policy-making processes. Yet resources on legislative elites are scattered, often specialized, limited in scope or not always accessible. This article introduces the Comparative Legislators Database (CLD), which joins micro-data collection efforts on open-collaboration platforms and other sources, and integrates with renowned political science datasets. The CLD includes political, sociodemographic, career, online presence, public attention, and visual information for over 45,000 contemporary and historical politicians from ten countries. The authors provide a straightforward and open-source interface to the database through an R package, offering targeted, fast and analysis-ready access in formats familiar to social scientists and standardized across time and space. The data is verified against human-coded datasets, and its use for investigating legislator prominence and turnover is illustrated. The CLD contributes to a central hub for versatile information about legislators and their behavior, supporting individual-level comparative research over long periods.

The Billion Prices Project: Using Online Prices for Measurement and Research
Alberto Cavallo and Roberto Rigobon
NBER Working Paper No. 22111
March 2016, Revised April 2016
JEL No. E31,F3,F4

ABSTRACT

New data-gathering techniques, often referred to as "Big Data" have the potential to improve statistics and empirical research in economics. In this paper we describe our work with online data at the Billion Prices Project at MIT and discuss key lessons 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. We emphasize how Big Data technologies are providing macro and international economists with opportunities to stop treating the data as "given" and to get directly involved with data collection.

SCIENCE ADVANCES | RESEARCH ARTICLE

SOCIAL NETWORKS

Leaking privacy and shadow profiles in online social networks

David Garcia

Social interaction and data integration in the digital society can affect the control that individuals have on their privacy. Social networking sites can access data from other services, including user contact lists where nonusers are listed too. Although most research on online privacy has focused on inference of personal information of users, this data integration poses the question of whether it is possible to predict personal information of nonusers. This article tests the shadow profile hypothesis, which postulates that the data given by the users of an online service predict personal information of nonusers. Using data from a disappeared social networking site, we perform a historical audit to evaluate whether personal data of nonusers could have been predicted with the personal data and contact lists shared by the users of the site. We analyze personal information of sexual orientation and relationship status, which follow regular mixing patterns in the social network. Going back in time over the growth of the network, we measure predictor performance as a function of network size and tendency of users to disclose their contact lists. This article presents robust evidence supporting the shadow profile hypothesis and reveals a multiplicative effect of network size and disclosure tendencies that accelerates the performance of predictors. These results call for new privacy paradigms that take into account the fact that individual privacy decisions do not happen in isolation and are mediated by the decisions of others.

Sneak preview
Applying data science to tackle social problems

Tracking the usage of a contact tracing app

nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-020-01044-x

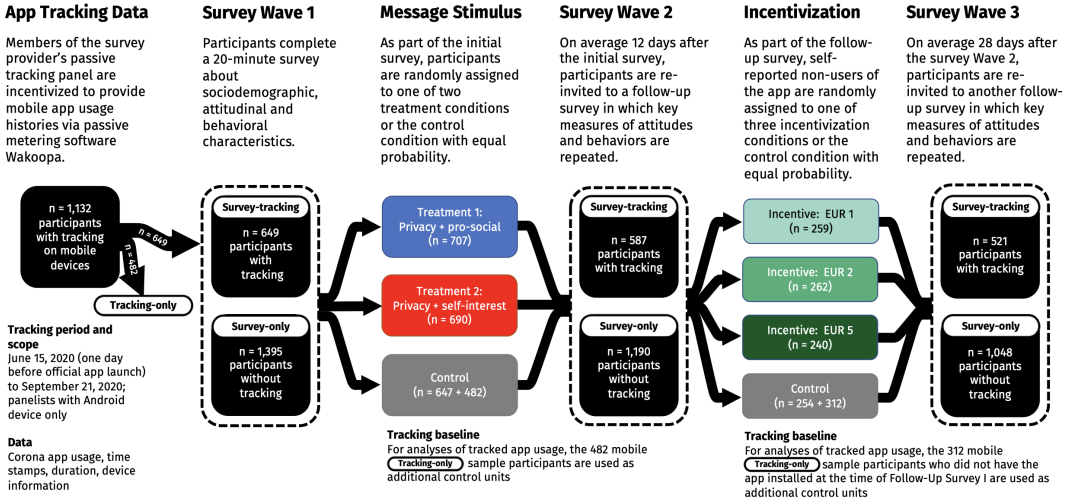
Check for updates

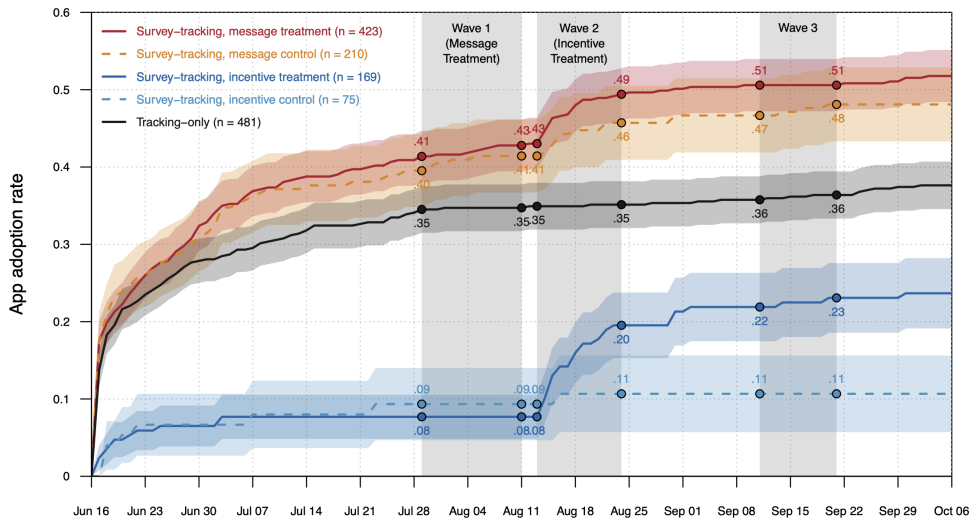
Tracking and promoting the usage of a COVID-19 contact tracing app

Simon Munzert[1]✉, Peter Selb[2], Anita Gohdes[1], Lukas F. Stoetzer[3] and Will Lowe[1]

Digital contact tracing apps have been introduced globally as an instrument to contain the COVID-19 pandemic. Yet, privacy by design impedes both the evaluation of these tools and the deployment of evidence-based interventions to stimulate uptake. We combine an online panel survey with mobile tracking data to measure the actual usage of Germany's official contact tracing app and reveal higher uptake rates among respondents with an increased risk of severe illness, but lower rates among those with a heightened risk of exposure to COVID-19. Using a randomized intervention, we show that informative and motivational video messages have very limited effect on uptake. However, findings from a second intervention suggest that even small monetary incentives can strongly increase uptake and help make digital contact tracing a more effective tool.

CORONA WARN-APP





Reducing hate speech on social media

Journal of Experimental Political Science (2021), **8**, 102–116
doi:10.1017/XPS.2020.14

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Don't @ Me: Experimentally Reducing Partisan Incivility on Twitter

Kevin Munger*

Pennsylvania State University, Pond Lab, State College, PA, USA
Corresponding author. Email: kmm7999@psu.edu

Abstract

I conduct an experiment which examines the impact of moral suasion on partisans engaged in uncivil arguments. Partisans often respond in vitriolic ways to politicians they disagree with, and this can engender hateful responses from partisans from the other side. This phenomenon was especially common during the contentious 2016 US Presidential Election. Using Twitter accounts that I controlled, I sanctioned people engaged partisan incivility in October 2016. I found that messages containing moral suasion were more effective at reducing incivility than were messages with no moral content in the first week post-treatment. There were no significant treatment effects in the first day post-treatment, emphasizing the need for research designs that measure effect duration. The type of moral suasion employed, however, did not have the expected differential effect on either Republicans or Democrats. These effects were significantly moderated by the anonymity of the subjects.

Keywords: affective polarization; Twitter; field experiment

(a) Example Bot–Clinton Condition

Change in Incivility, Full Sample (N=310)

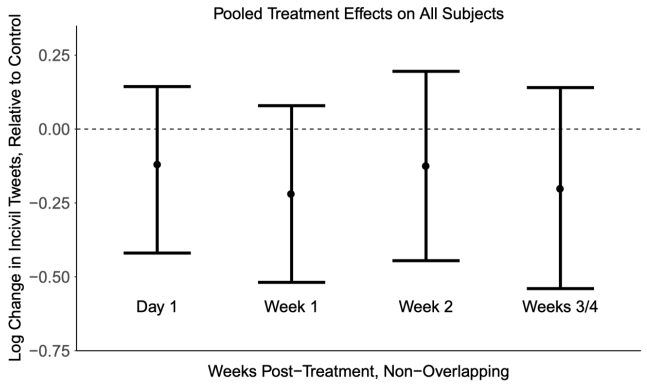


Figure 4
Pooled treatment effects on the entire sample, controlling for the log of the number of pre-treatment uncivil tweets sent by each subject. Lines represent 95% confidence intervals.

Monitoring the effects of climate change on health

The 2020 report of The *Lancet* Countdown on health and climate change: responding to converging crises

Nick Watts, Markus Amann, Nigel Arnell, Sonja Ayeb-Karlsson, Jessica Beagley, Kristine Belesova, Maxwell Boykoff, Peter Byass, Wenjia Cai, Diarmid Campbell-Lendrum, Stuart Capstick, Jonathan Chambers, Samantha Coleman, Carole Dalin, Meaghan Daly, Niheer Dasandi, Shouro Dasgupta, Michael Davies, Claudia Di Napoli, Paula Dominguez-Salas, Paul Drummond, Robert Dubrow, Kristie L Ebi, Matthew Eckelman, Paul Ekins, Luis E Escobar, Lucien Georgeson, Su Golder, Delia Grace, Hilary Graham, Paul Haggar, Ian Hamilton, Stella Hartinger, Jeremy Hess, Shih-Che Hsu, Nick Hughes, Slava Jankin Mikhaylov, Marcia P Jimenez, Ilan Kelman, Harry Kennard, Gregor Kiesewetter, Patrick L Kinney, Tord Kjellstrom, Dominic Kniveton, Pete Lampard, Bruno Lemke, Yang Liu, Zhao Liu, Melissa Lott, Rachel Lowe, Jaime Martinez-Urtaza, Mark Maslin, Lucy McAllister, Alice McGushin, Celia McMichael, James Milner, Maziar Moradi-Lakeh, Karyn Morrissey, Simon Munzert, Kris A Murray, Tara Neville, Maria Nilsson, Maquins Odhiambo Sewe, Tadj Oreszczyn, Matthias Otto, Fereidoon Owfi, Olivia Pearman, David Pencheon, Ruth Quinn, Mahnaz Rabbaniha, Elizabeth Robinson, Joacim Rocklöv, Marina Romanello, Jan C Semenza, Jodi Sherman, Liuhua Shi, Marco Springmann, Meisam Tabatabaei, Jonathon Taylor, Joaquin Triñanes, Joy Shumake-Guillemot, Bryan Vu, Paul Wilkinson, Matthew Winning, Peng Gong, Hugh Montgomery*, Anthony Costello**

Executive summary

The *Lancet* Countdown is an international collaboration established to provide an independent, global monitoring system dedicated to tracking the emerging health profile of the changing climate.

The 2020 report presents 43 indicators across five sections: climate change impacts, exposures, and vulnerabilities; adaptation, planning, and resilience for health; mitigation actions and health co-benefits; economics and finance; and public and political engagement. This report represents the findings and consensus of the 35 leading academic institutions and UN agencies that make up The *Lancet* Countdown, and draws on the expertise of climate scientists, geographers, engineers, experts in energy, food, and transport, economists, social, and political scientists, data scientists, public health professionals, and doctors.

trends within and between countries. An examination of the causes of climate change revealed similar issues, and many carbon-intensive practices and policies lead to poor air quality, poor food quality, and poor housing quality, which disproportionately harm the health of disadvantaged populations.

Vulnerable populations were exposed to an additional 475 million heatwave events globally in 2019, which was, in turn, reflected in excess morbidity and mortality (indicator 1.1.2). During the past 20 years, there has been a 53·7% increase in heat-related mortality in people older than 65 years, reaching a total of 296 000 deaths in 2018 (indicator 1.1.3). The high cost in terms of human lives and suffering is associated with effects on economic output, with 302 billion h of potential labour capacity lost in 2019 (indicator 1.1.4). India and Indonesia were among the worst affected countries, seeing losses of potential

*Co-chairs
Institute for Global Health (N Watts MA, J Beagley BA, S Coleman MSE, Prof I Kelman PhD, A McGushin MSc, M Romanello PhD), **Office of the Vice Provost for Research** (Prof A Costello FmedSci), **Energy Institute** (S-C Hsu MSc, I Hamilton PhD, H Kennard PhD, Prof T Oreszczyn PhD), **Institute for Sustainable Resources** (C Dalin PhD, P Drummond MSc, Prof P Ekins PhD, N Hughes PhD, M Winning PhD), **Institute for Environmental Design and Engineering** (Prof M Davies PhD), **Department of Geography** (Prof M Maslin PhD), **and**

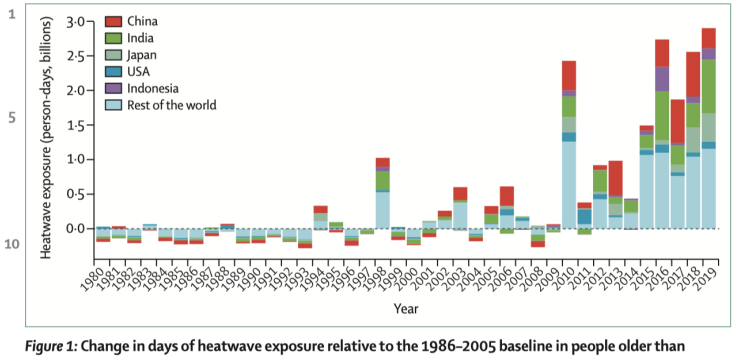
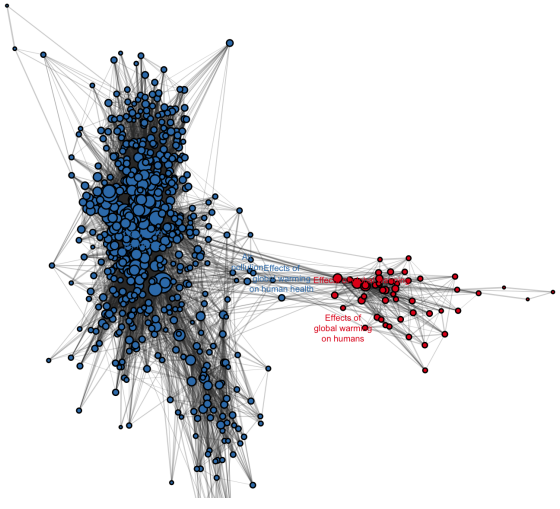
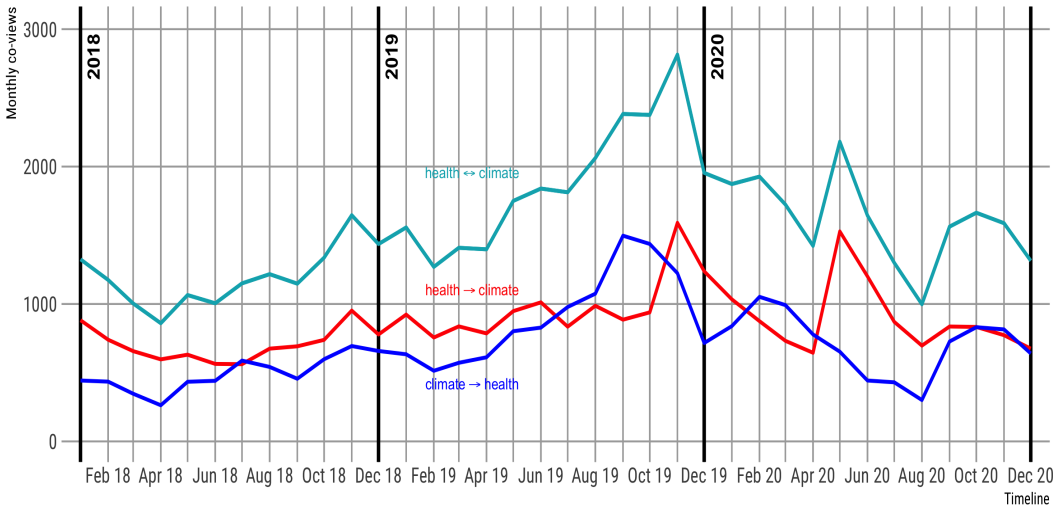


Figure 1: **Change in days of heatwave exposure relative to the 1986–2005 baseline in people older than 65 years**
The dotted line at 0 represents baseline.





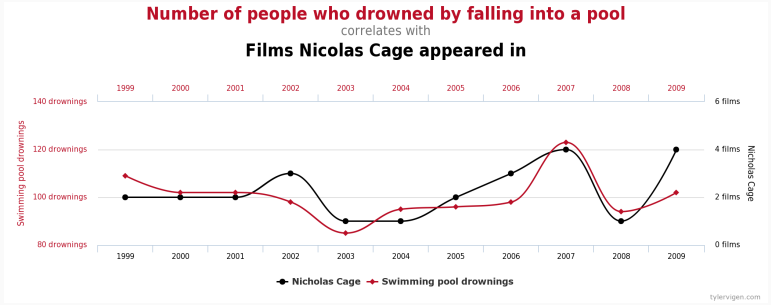
Sneak preview
Getting to know the limits of data

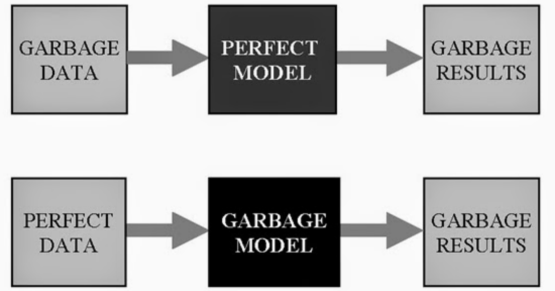
Calling bullsh*t when you see it

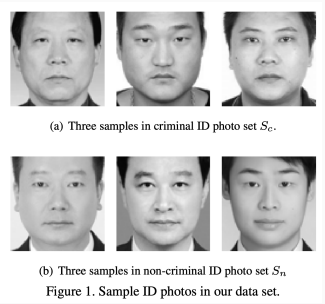
Learn not to be fooled by

- big data
- garbage data
- garbage models
- weird samples
- claims of generality
- statistical significance
- implausibly large effect sizes
- highly precise forecasts
- overfitted models

And much more...

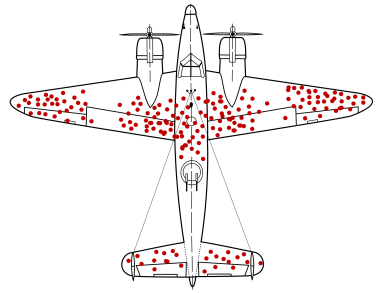


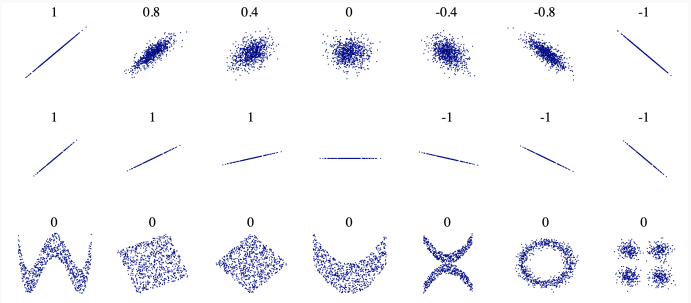



(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.





Sneak preview
Reflecting everyday ethics in data science

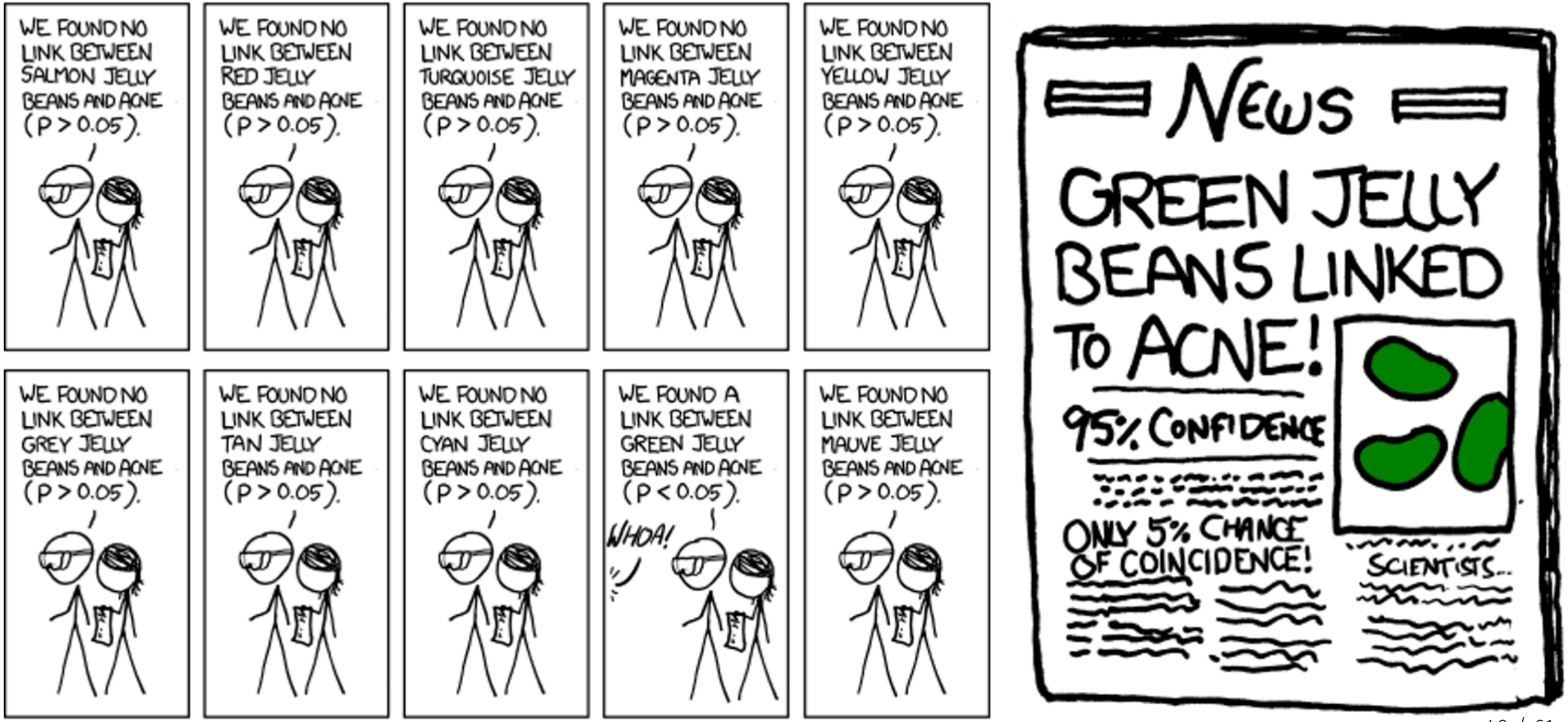
How do I pay clickworkers fairly?

How do I respect intellectual property?

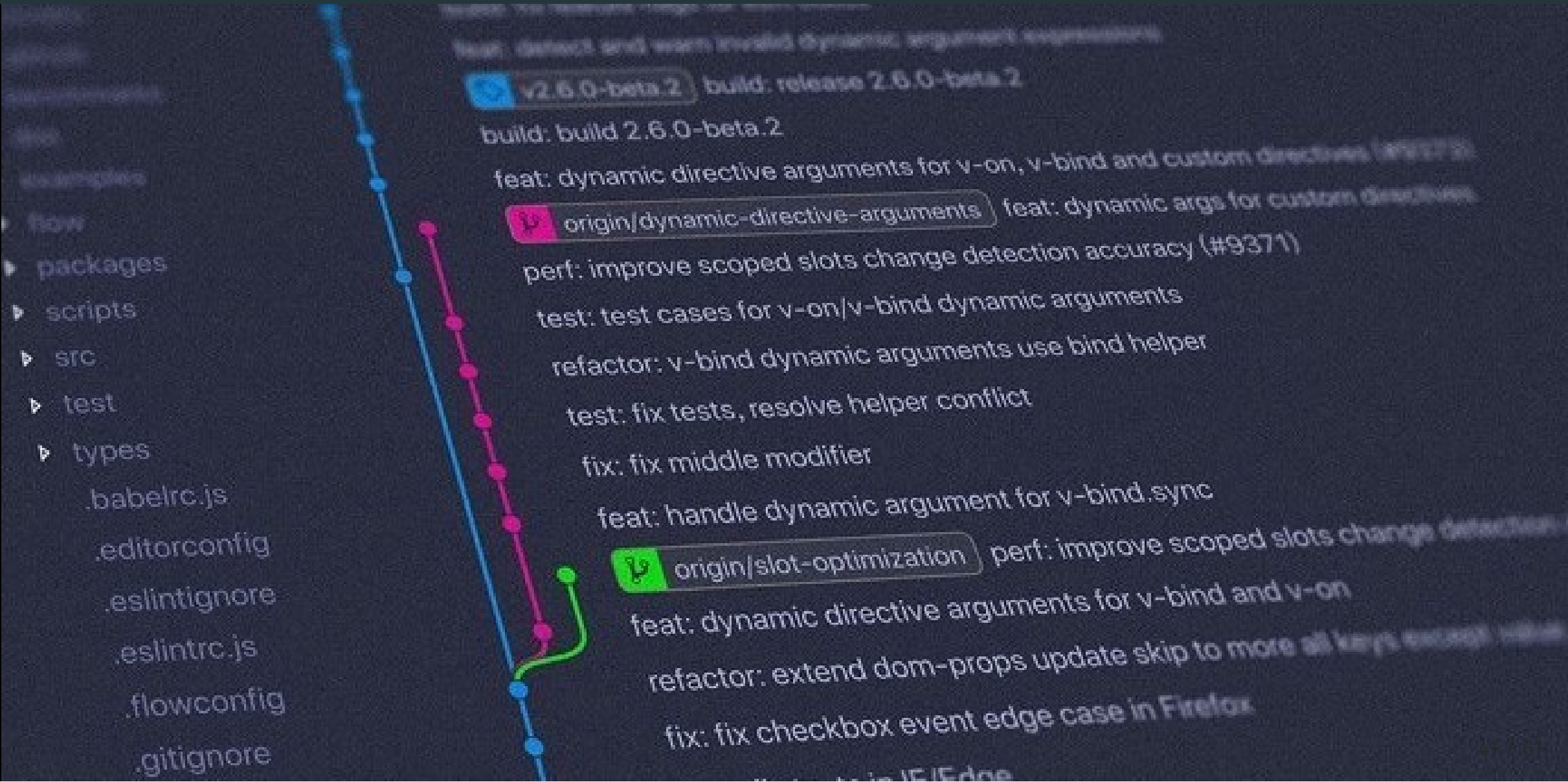
How do I protect the privacy of my research subjects?

How do I protect the safety of my research subjects?

How do I ensure statistical, measurement validity, etc.?



How do I ensure an open science workflow?



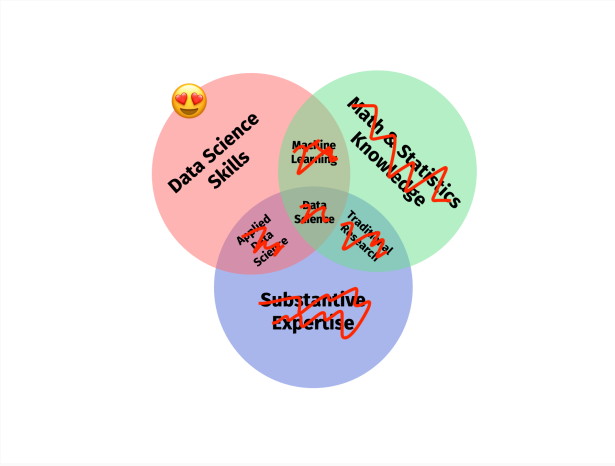
How do I communicate results honestly?

Class logistics

The plan

Goals of the course

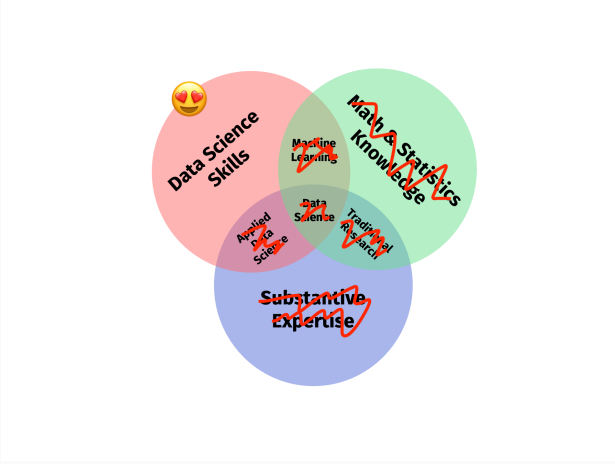
- This course equips you with conceptual knowledge about the data science pipeline and coding workflow, data structures, and data wrangling.
- It enables you to apply this knowledge with statistical software.
- It prepares you for our other core courses and electives as well as the master's thesis.



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
- Relational databases and SQL
- Web data and technologies
- Model fitting and simulation[1]
- Viusalization
- The command line
- Programming workflow: debugging, automation, packaging
- Monitoring and communication
- Data science ethics

[1] Not yet: mathematical and statistical foundations, ML.

You at the beginning of the course

You at the end of the course

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?"

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, seemless integration.

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?"

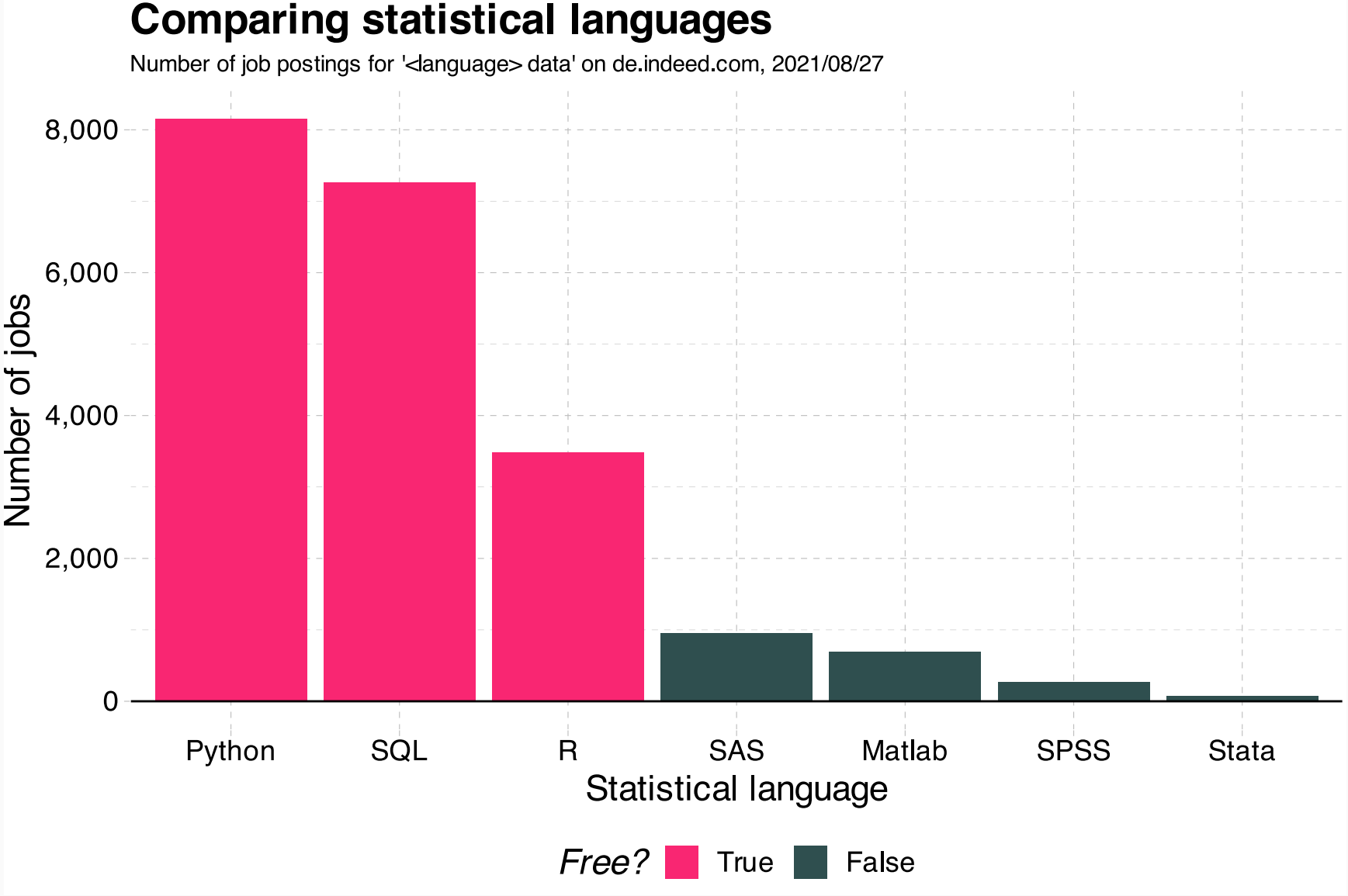
Bridge to multiple other programming environments, with statistics at heart

- 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, seemless integration.

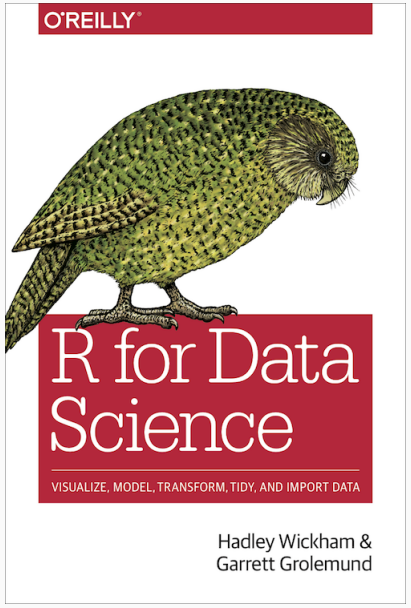
Path dependency

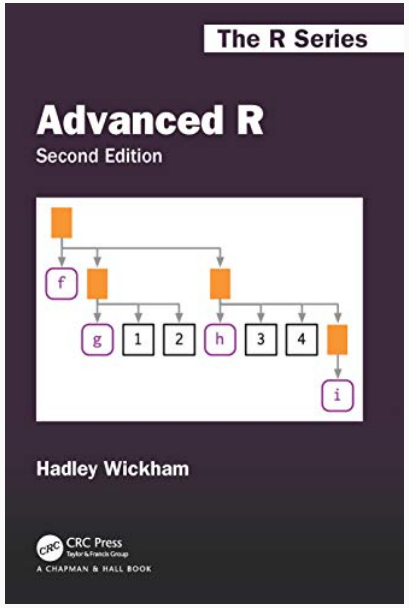
- It's also the language that I know best.
- (Learning multiple languages is a good idea, though.)

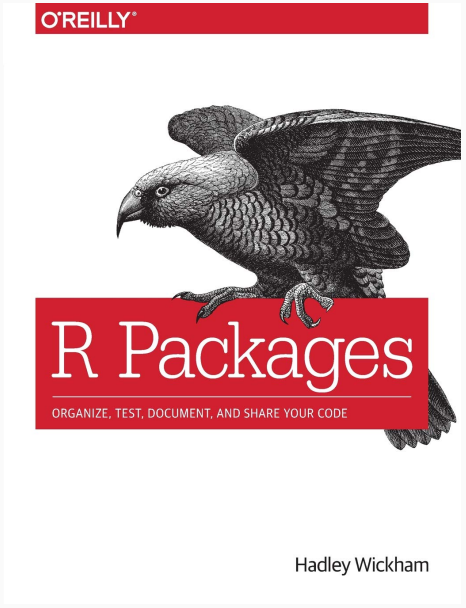
Why R and RStudio? (cont.)

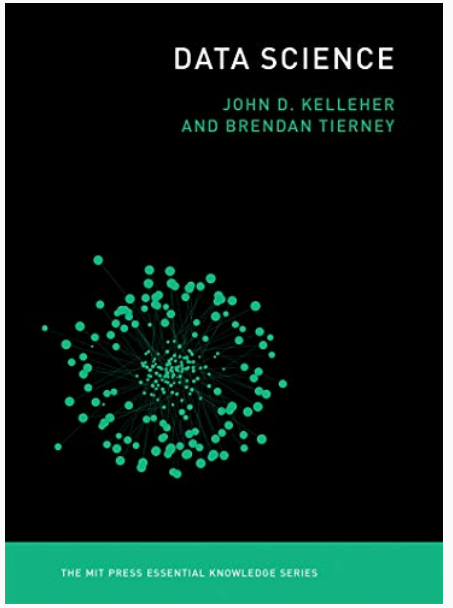


Core (and optional) readings









Attendance

General rules

- As a general rule, switching on-site groups will not be possible during the course of the semester. - However, if you happen to be unavailable on your slot but not on the other, please tune in (**only online!**). Don't email me about it - this will result in a crowded mailbox on my end.
- You cannot miss more than two sessions. If you have to miss a session for medical reasons or personal emergencies, please inform Examination Office.
- We will check attendance on-site + online. If you attend online, type your full name in the chat when you log in.
- For on-site attendance, the current Hertie hygiene rules apply!

Attendance

General rules

- As a general rule, switching on-site groups will not be possible during the course of the semester. - However, if you happen to be unavailable on your slot but not on the other, please tune in (**only online!**). Don't email me about it - this will result in a crowded mailbox on my end.
- You cannot miss more than two sessions. If you have to miss a session for medical reasons or personal emergencies, please inform Examination Office.
- We will check attendance on-site + online. If you attend online, type your full name in the chat when you log in.
- For on-site attendance, the current Hertie hygiene rules apply!

On-site rotation scheme

- There is more interest in attending the course on-site than we have space.
- I have set up a rotation scheme that tries to balance on-site attendance across participants, within groups. Check it out on Moodle.
- If you are not listed as on-site participant for a particular session, please do not sneak into the on-site session.
- **If you plan to attend online instead of on-site on short notice**, please drop Ayamba (kwoyila@hertie-school.org) a line until Wednesday noon so that she can inform successors.

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, get in touch with me (✉ munzert@hertie-school.org).
- If you want to discuss any other matters with me, just drop me a message and we'll arrange a meeting.
- My (virtual) office hours (by appointment) are Tuesdays, 3-4pm CET.
- For general technical advice, the Research Consulting Team at the Data Science Lab is there for you.

Assignments and grading

Component	Weight
4 × homework assignments (10% each)	40%
1 × workshop presentation	25%
1 × final data science project	35%

Assignments and grading

Component	Weight
4 × homework assignments (10% each)	40%
1 × workshop presentation	25%
1 × final data science project	35%

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) and the 4 best will contribute to the final grade.
- You'll have roughly 2 weeks to work on each assignment.
- You submit your solutions via GitHub Classroom.
- Grades will be based on (1) the accuracy of your solutions and (2) the adherence of a clean and efficient coding style that you will learn in the first sessions.

Assignments and grading

Component	Weight
4 × homework assignments (10% each)	40%
1 × workshop presentation	25%
1 × final data science project	35%

Workshop presentation

- About halfway through the semester, we will flip the roles and you will become the instructor for one session of a workshop on **Tools for Data Science**. Whether this will happen onsite (hybrid) or online will still be determined.
- You, in groups of 2 students, will present a preexisting tool or package that is useful for the data science workflow.
- Topics are listed on GitHub and will be 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 talk and the prepared exercise materials, which must also be hosted openly on GitHub, will be graded.

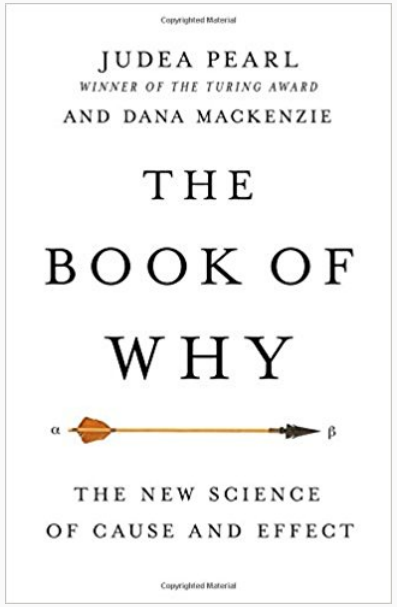
Assignments and grading

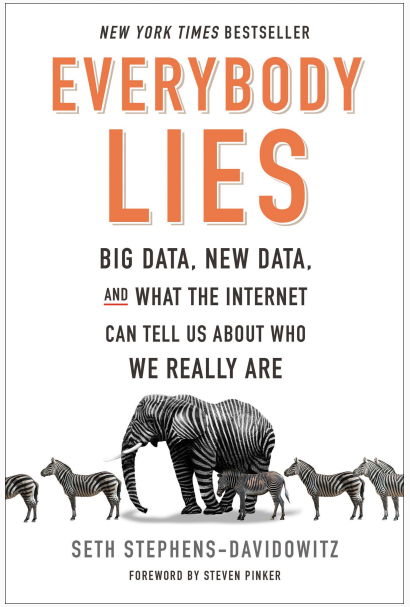
Component	Weight
4 × homework assignments (10% each)	40%
1 × workshop presentation	25%
1 × final data science project	35%

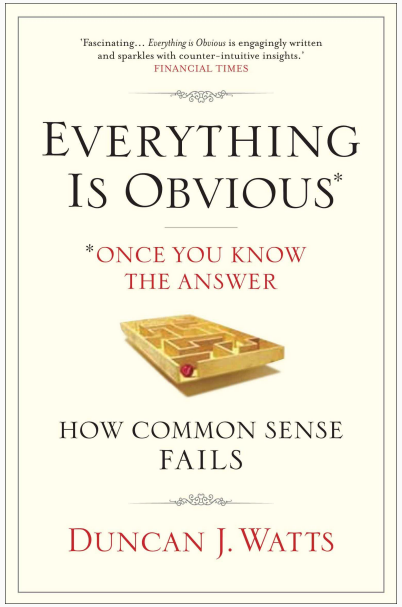
Final data science project

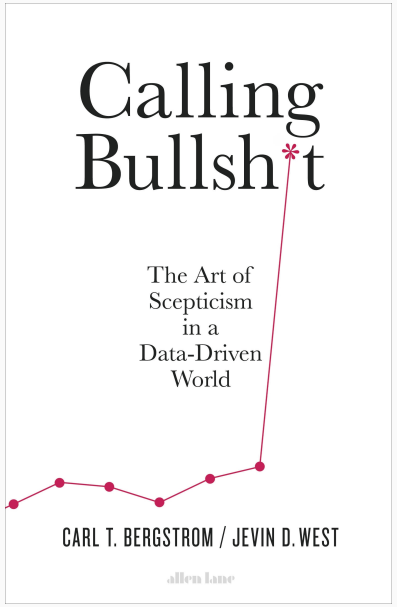
- In the component, to be submitted a couple of weeks after classes have finished, you will design and implement your own data science project.
- You are supposed to collaborate in groups of two or three students.
- Student groups choose their topic subject to approval by the instructor.
- A wide variety of project types are imaginable and include
  1. A report about a statistical data analysis
  2. A policy explainer project enriched with data (think: a piece of data journalism)
  3. A dashboard to make data / analyses accessible in an interactive fashion
  4. An R package with a dedicated use case (e.g., API binding, visualization tools, etc.)

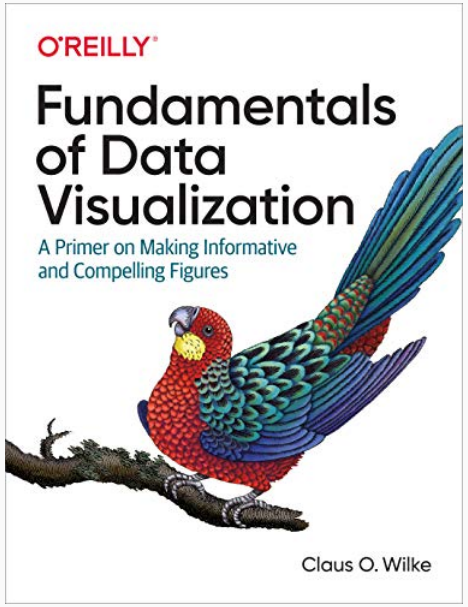
Further reading



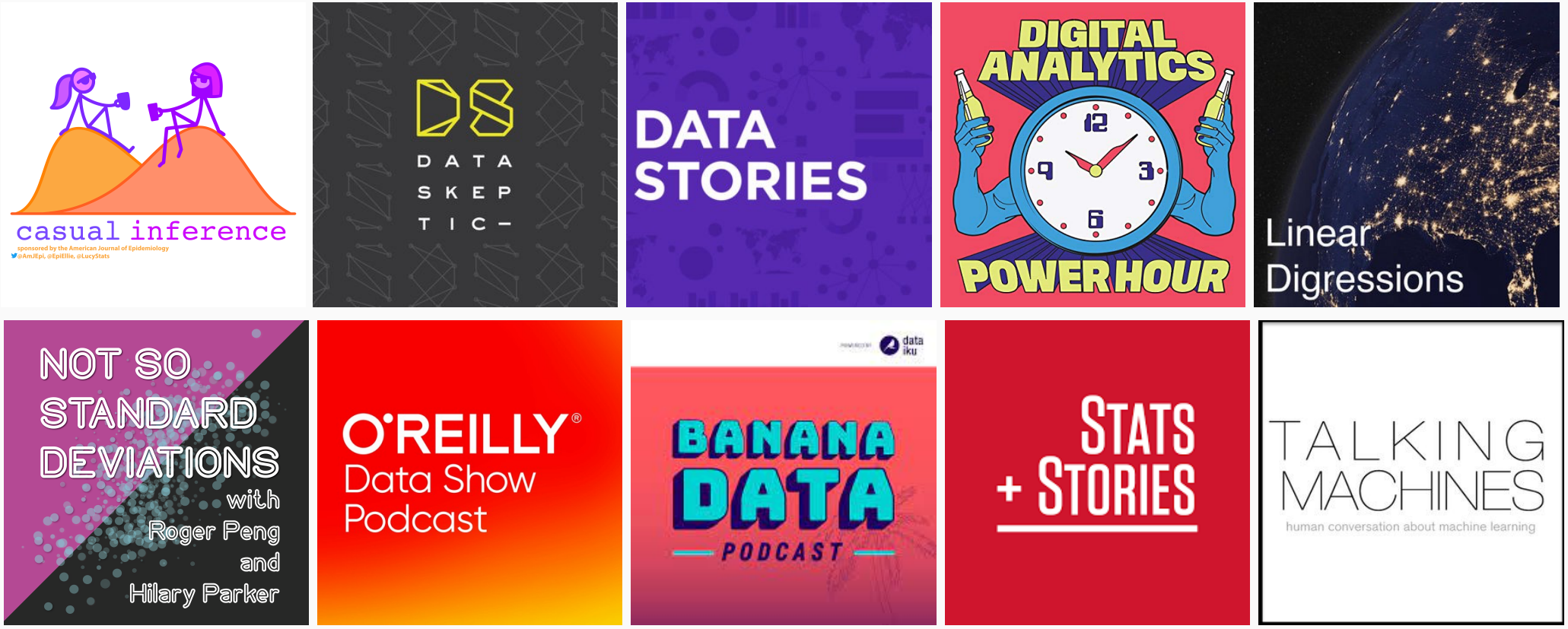




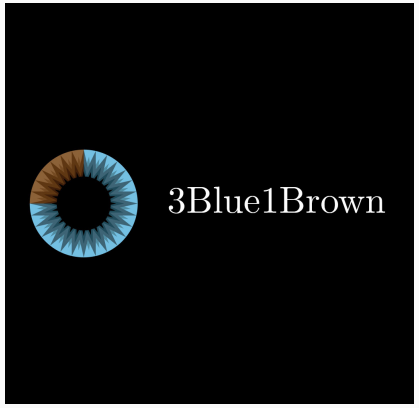


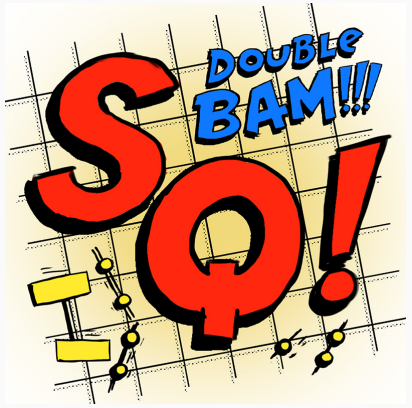


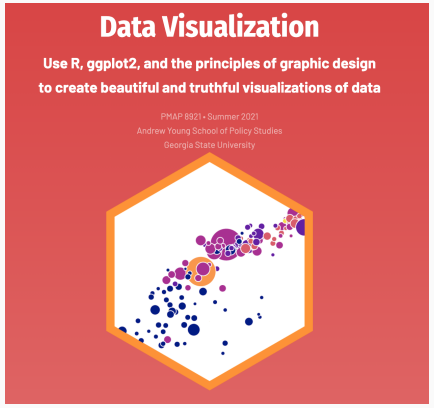
Further listening



Further watching







Getting started for the course

Software

1. Download R.
2. Download RStudio.
3. Download Git.
4. Create an account on GitHub and register for a student/educator discount. You will soon receive an invitation to the course organization on GitHub, as well as GitHub classroom, which is how we'll disseminate and submit assignments, receive feedback and grading, etc.

Getting started for the course

Software

1. Download R.
2. Download RStudio.
3. Download Git.
4. Create an account on GitHub and register for a student/educator discount. You will soon receive an invitation to the course organization on GitHub, as well as GitHub classroom, which is how we'll disseminate and submit assignments, receive feedback and grading, etc.

OS extras

- **Windows:** Install Rtools. You might also want to install Chocolately.
- **Mac:** Install Homebrew.
- **Linux:** None (you should be good to go).

Checklist

☑ Do you have the most recent version of R?

```
R> version$version.string
```

```
## [1] "R version 4.1.0 Patched (2021-07-20 r80657)"
```

☑ Do you have the most recent version of RStudio? (The preview version is fine.)

```
R> RStudio.Version()$version
R> ## Requires an interactive session but should return something like "[1] '1.4.1100'"
```

☑ Have you updated all of your R packages?

```
R> update.packages(ask = FALSE, checkBuilt = TRUE)
```

Checklist (cont.)

Open up the shell.

- Windows users, make sure that you installed a Bash-compatible version of the shell. If you installed Git for Windows, then you should be good to go.

☑ Which version of Git have you installed?

```
$ git --version
```

```
## git version 2.30.1 (Apple Git-130)
```

☑ Did you introduce yourself to Git? (Substitute in your details.)

```
$ git config --global user.name 'Simon Munzert'
$ git config --global user.email 'munzert@hertie-school.org'
$ git config --global --list
```

☑ Did you register an account in GitHub?

Coming up

The first lab session

Wednesday is lab day. Get to know Lisa, Tom, R, and RStudio, four of your best friends for the next months!

Next lecture

Version control and project management