Master of Data Science for Public Policy

Taster Lecture: Introduction to Data Science

Simon Munzert
Hertie School | GRAD-C11

Welcome!

Roadmap

1. Welcome!
2. What is data science?
3. Sneak preview
4. Q&A

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 for everything else.

Me

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

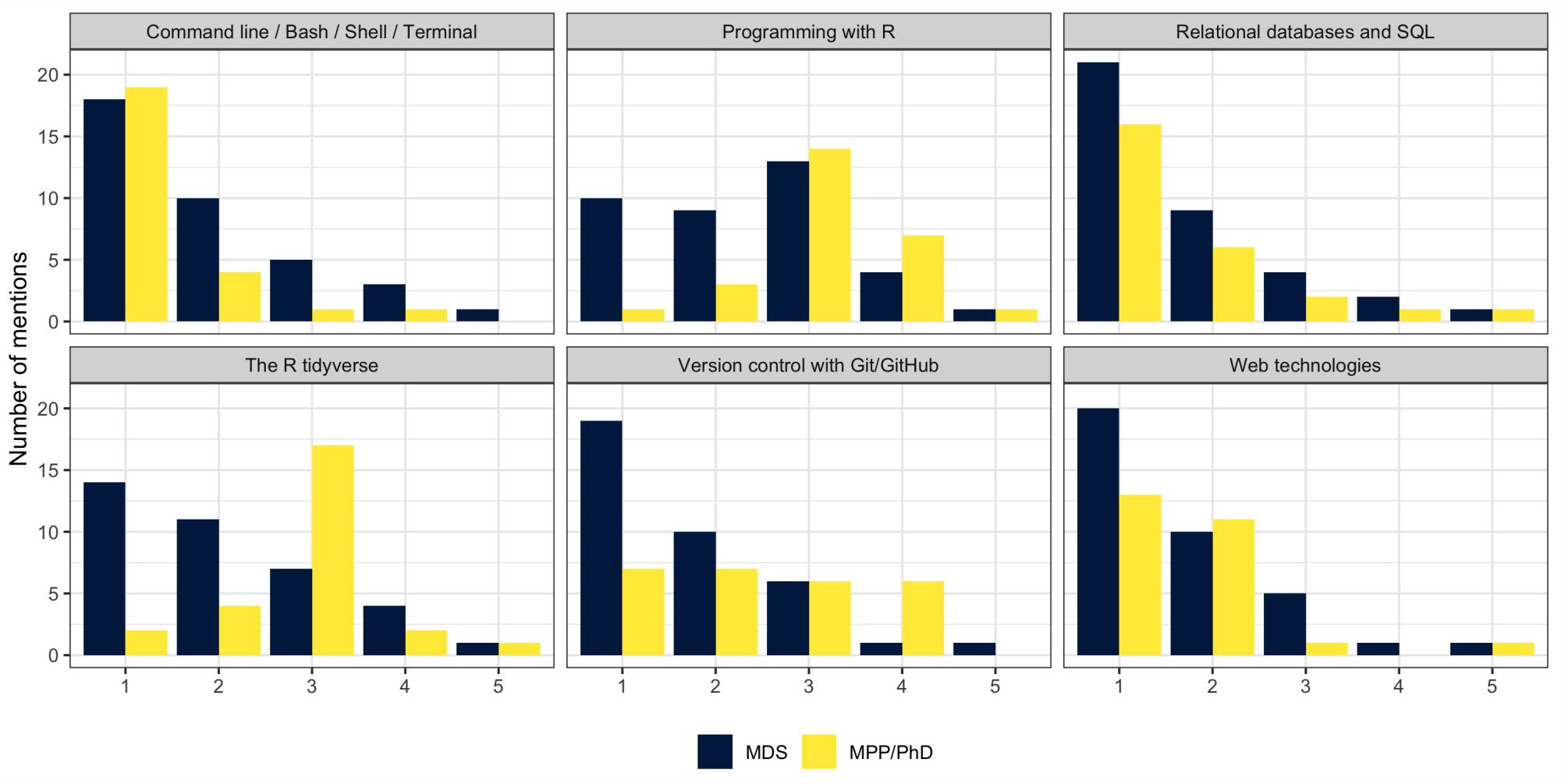
✉ munzert@hertie-school.org

🎓 Assistant Professor of Data Science and Public Policy

You

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

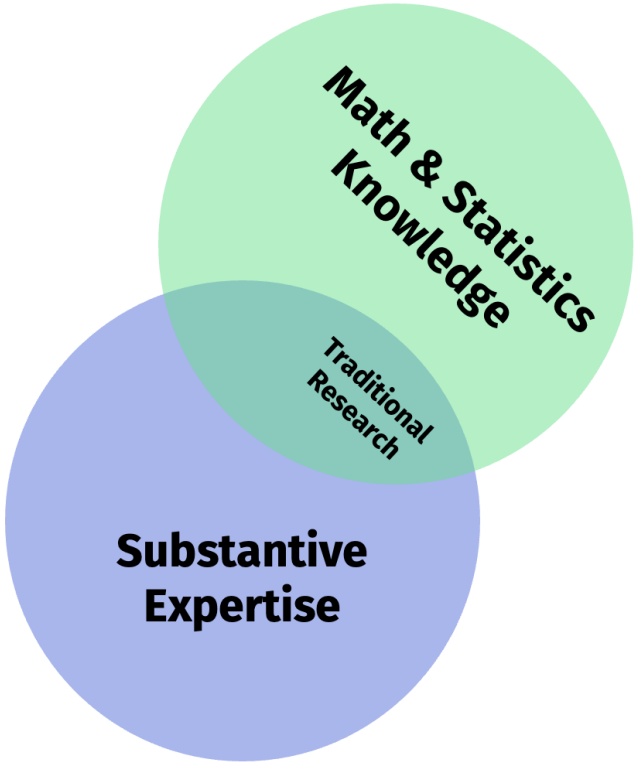
More about you (one year ago)



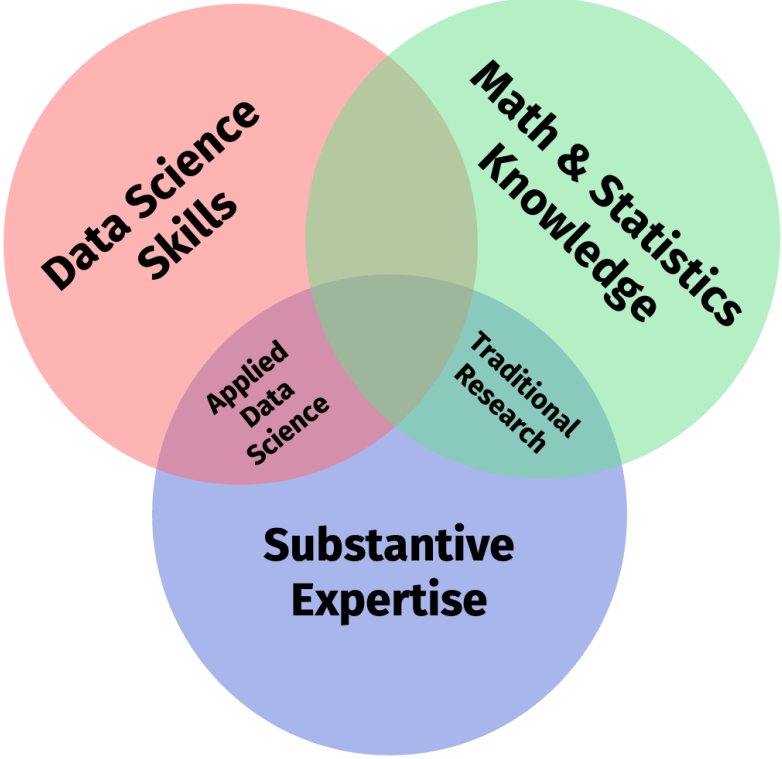
What is data science?

An old classic

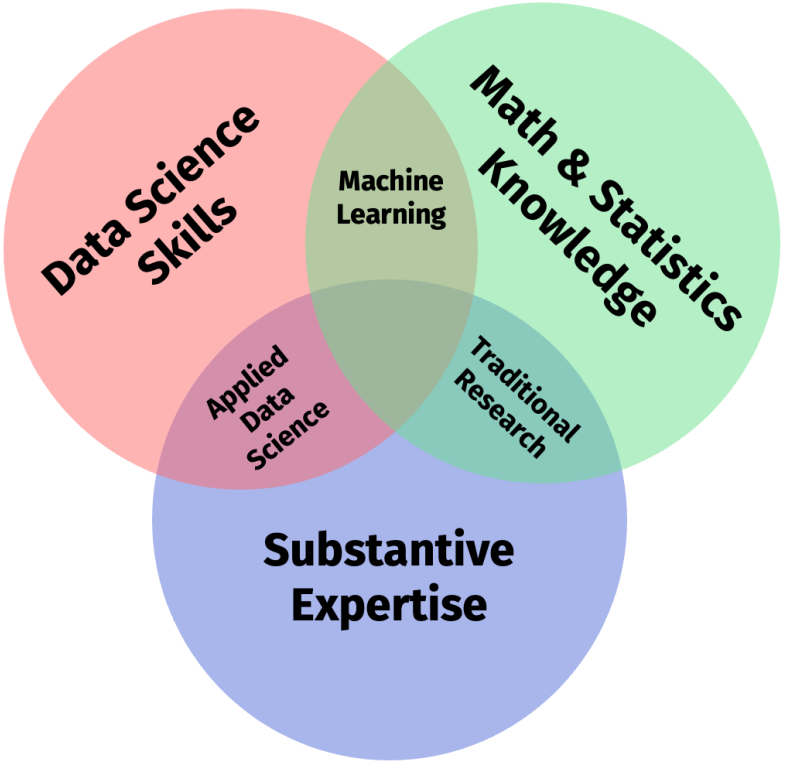
An old classic



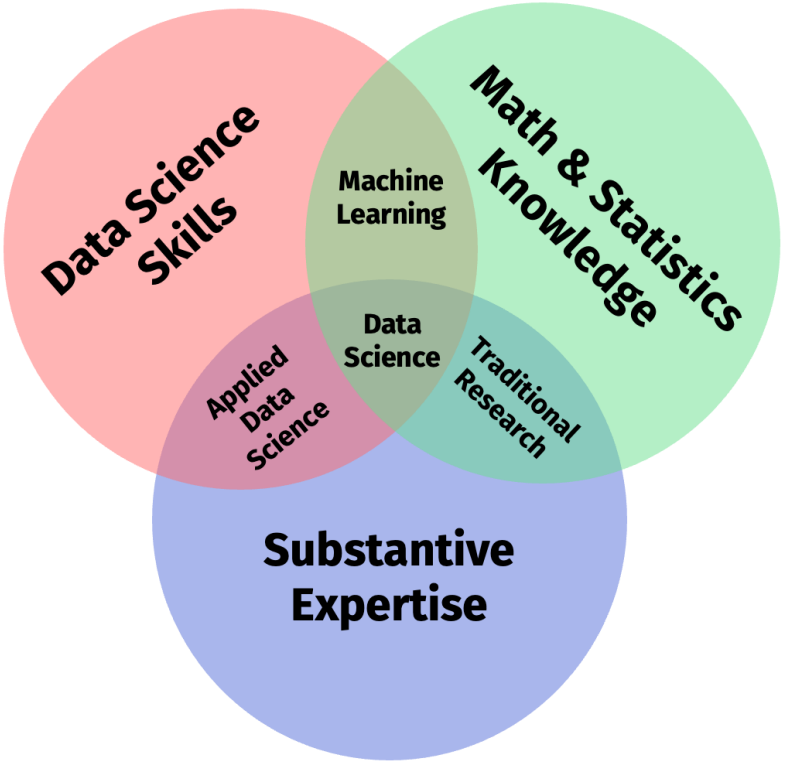
An old classic



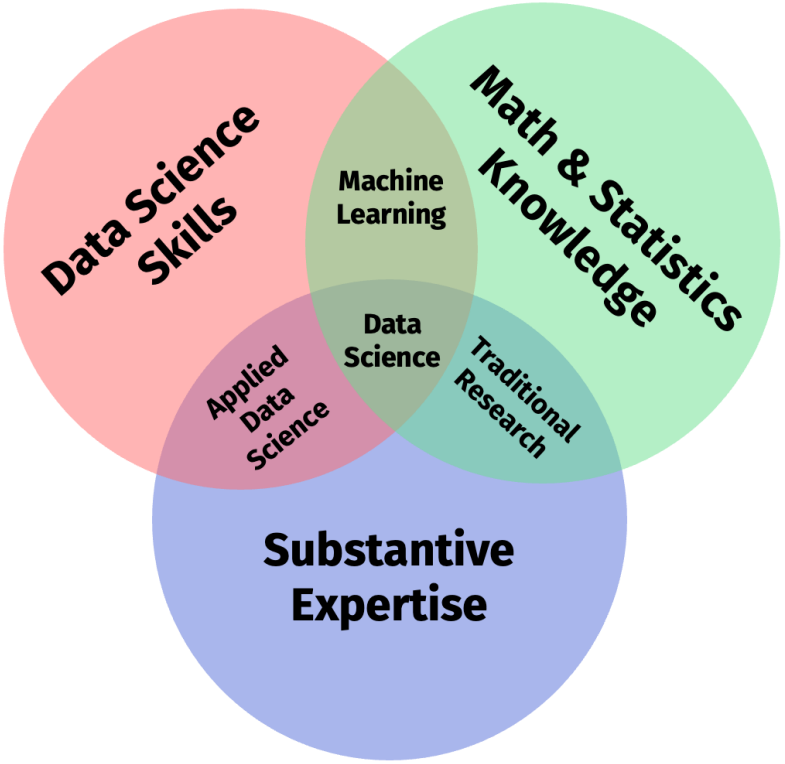
An old classic



An old classic



An old classic



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

The data science pipeline

Preparatory work

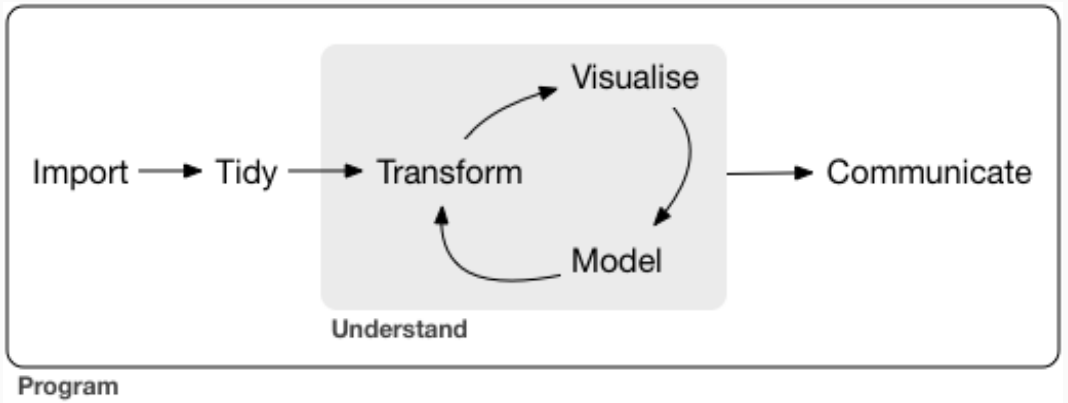
- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

The data science pipeline

Preparatory work

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

Data operation



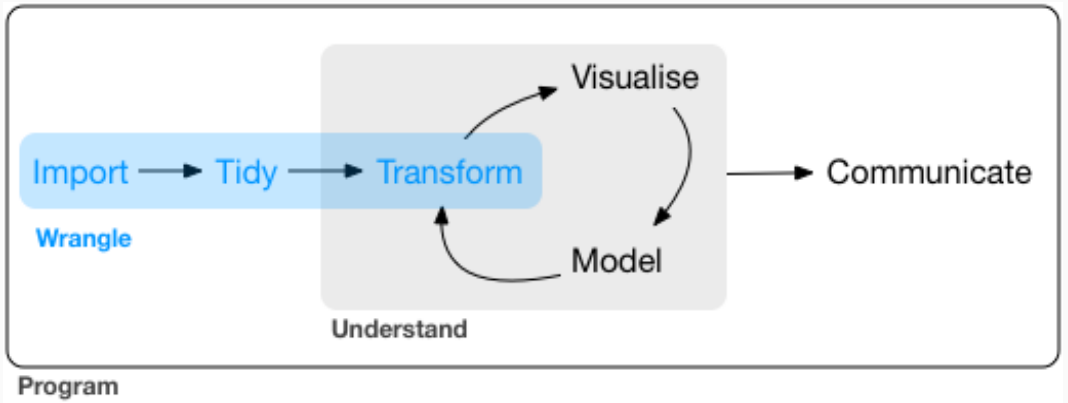
The data science pipeline

Preparatory work

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

Data operation

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



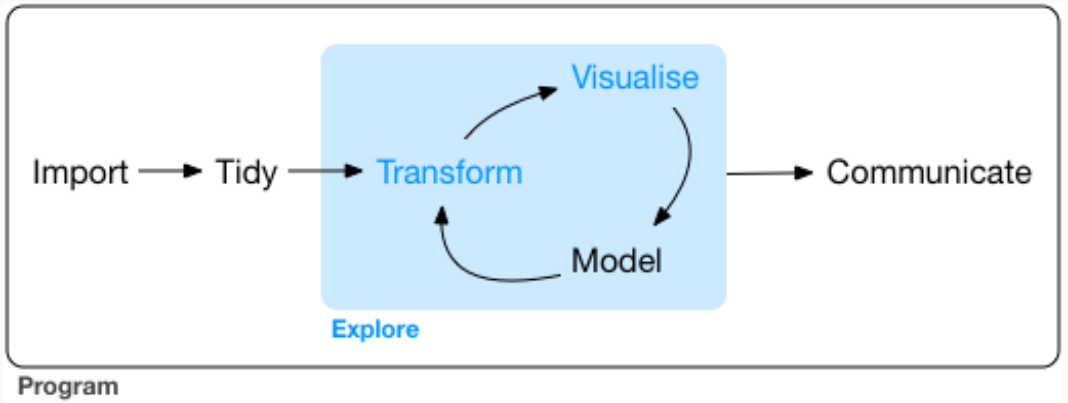
The data science pipeline

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



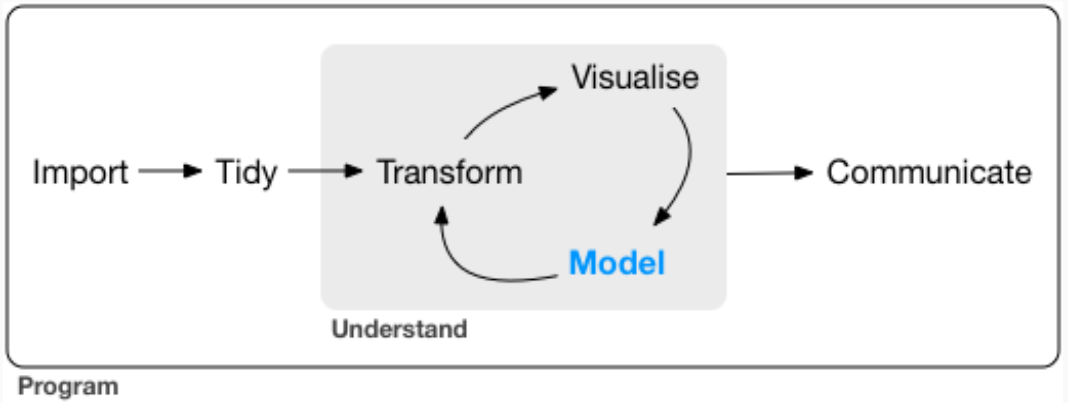
The data science pipeline

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

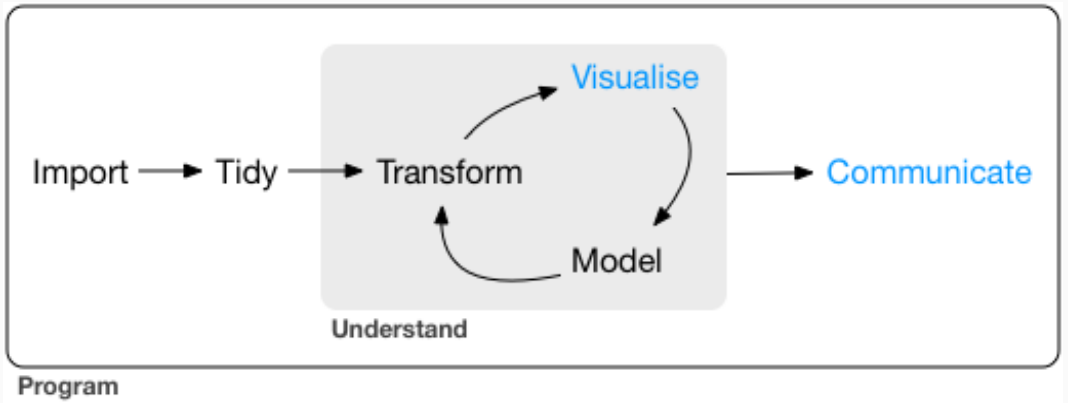
- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

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

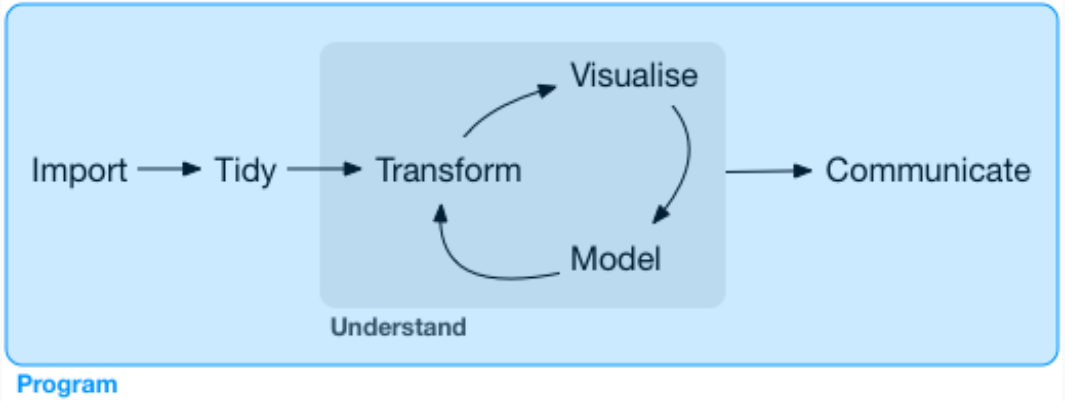
- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

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

Introduction to Data Science in a nutshell

Session	Session Date	Session Title
		Setting things up
1	09.09.2021	What is data science?
2	16.09.2021	Version control and project management
		Collecting and wrangling data
3	23.09.2021	R and the tidyverse
4	30.09.2021	Relational databases and SQL
5	07.10.2021	Web data and technologies
		Analyzing data
6	14.10.2021	Model fitting and simulation
		Mid-term Exam Week: 18.10 - 22.10.2021 – no class
7	28.10.2021	Visualization
8	04.11.2021	Workshop: Tools for Data Science
		Fine-tuning the workflow
9	11.11.2021	Working at the command line
10	18.11.2021	Debugging, automation, packaging
11	25.11.2021	Monitoring and communication
12	02.12.2021	Data science ethics
		Final Exam Week: 14.12 - 18.12.2021 – no class

Sneak preview
Learning to love a programming environment

The tidyverse

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?"

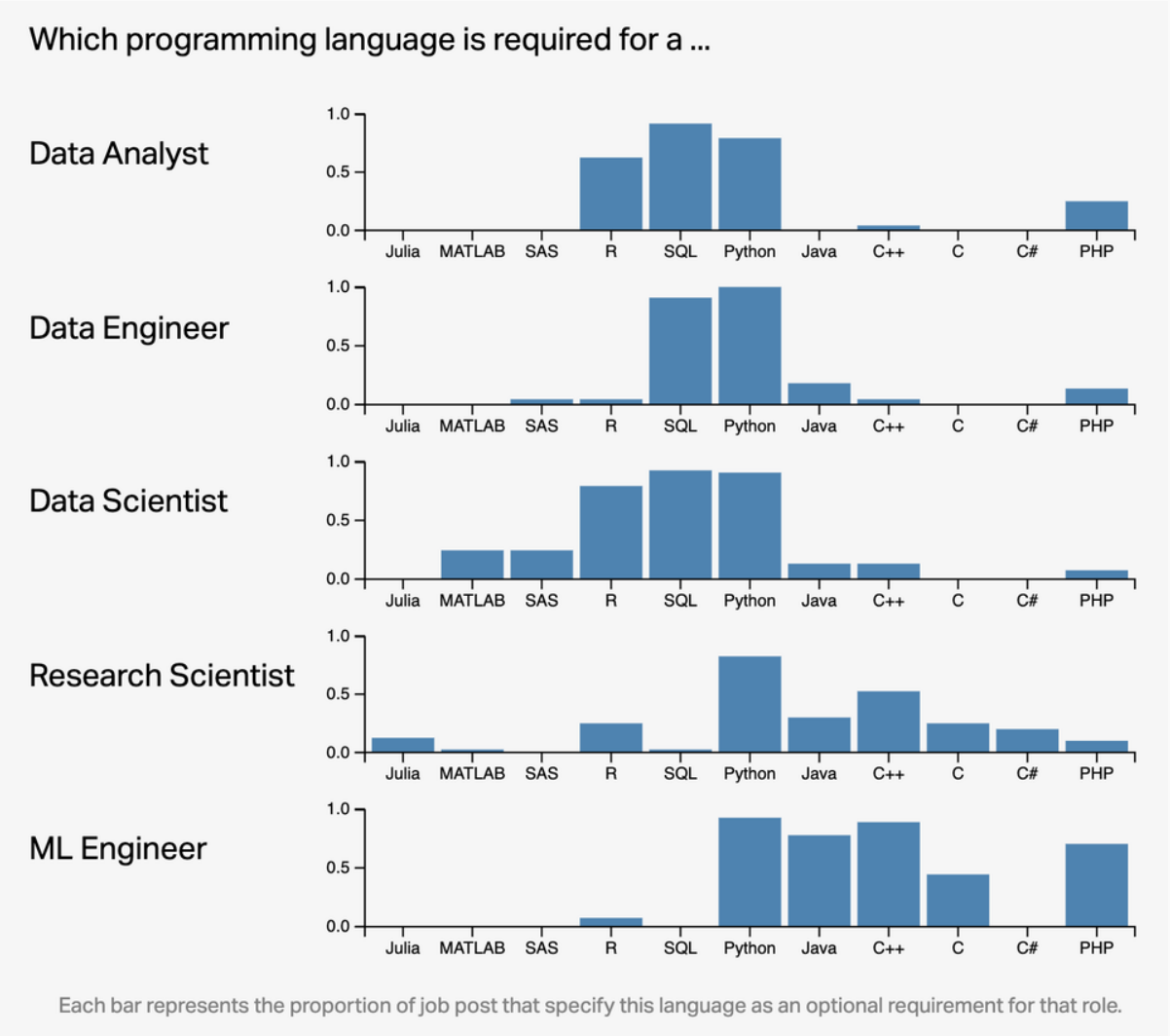
Bridge to multiple other programming environments, with statistics at heart

- 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 - and many data scientists with a foot in public policy - know best.
- (Learning multiple languages is a good idea, though.)

Why R and RStudio? (cont.)



`Credit` Left_Ad8361/Reddit

Sneak preview
Collecting web data at scale

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.

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.

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.

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.

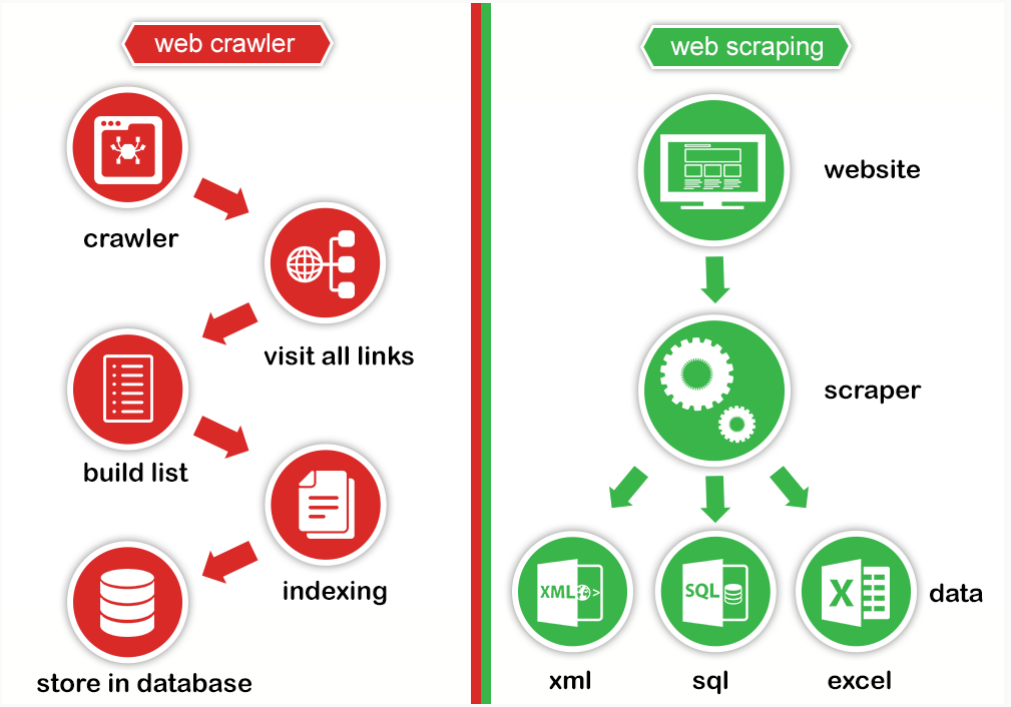
Web scraping

What is web scraping?

1. Pulling (unstructured) data from websites (HTMLs)
2. Bringing it into shape (into an analysis-ready format)

The philosophy of scraping with R

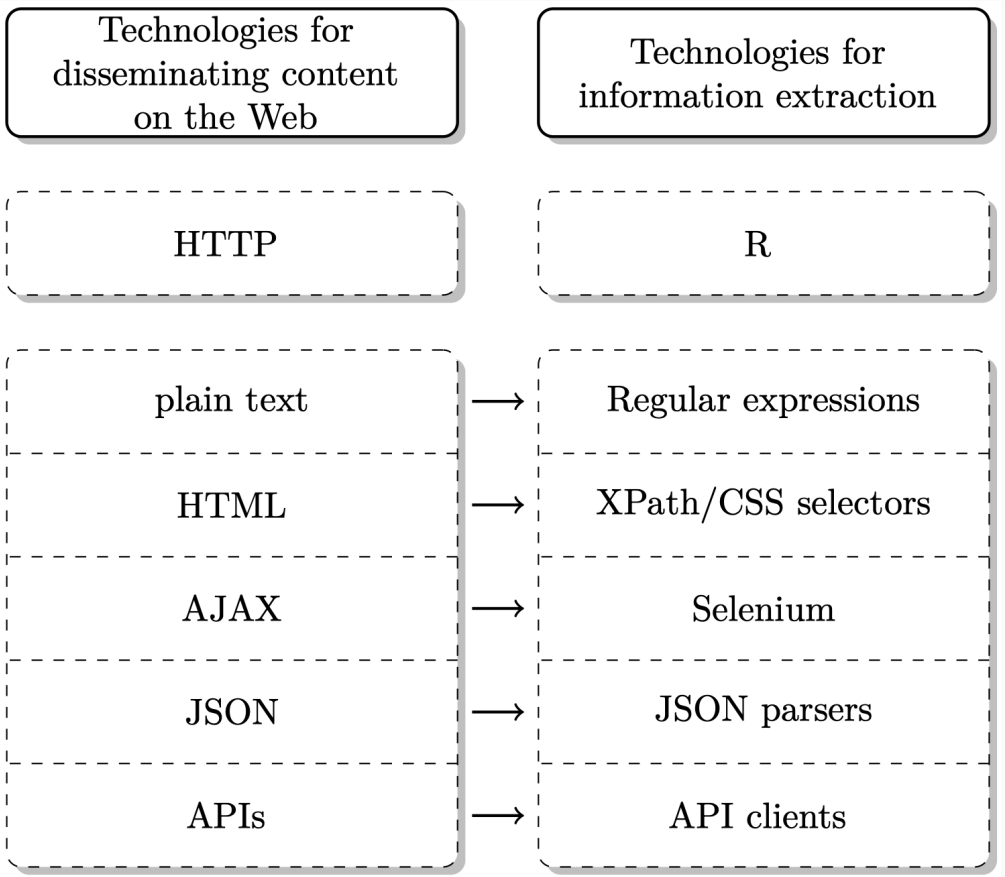
- No point-and-click procedure
- Script the entire process from start to finish
- **Automate**
  - The downloading of files
  - The scraping of information from web sites
  - Tapping APIs
  - Parsing of web content
  - Data tidying, text data processing
- Easily scale up scraping procedures
- Scheduling of scraping tasks



`Credit` prowebscraping.com

Technologies of the world wide web

- To fully unlock the potential of web data for data science, we draw on certain web technologies.
- Importantly, often a basic understanding of these technologies is sufficient as the focus is on web data collection, not web development.
- Specifically, we have to understand
  - How our machine/browser/R communicates with web servers (→ **HTTP/S**)
  - How websites are built (→ **HTML**, **CSS**, basics of **JavaScript**)
  - How content in webpages can be effectively located (→ **XPath**, **CSS selectors**)
  - How dynamic web applications are executed and tapped (→ **AJAX**, **Selenium**)
  - How data by web services is distributed and processed (→ **APIs**, **JSON**, **XML**)



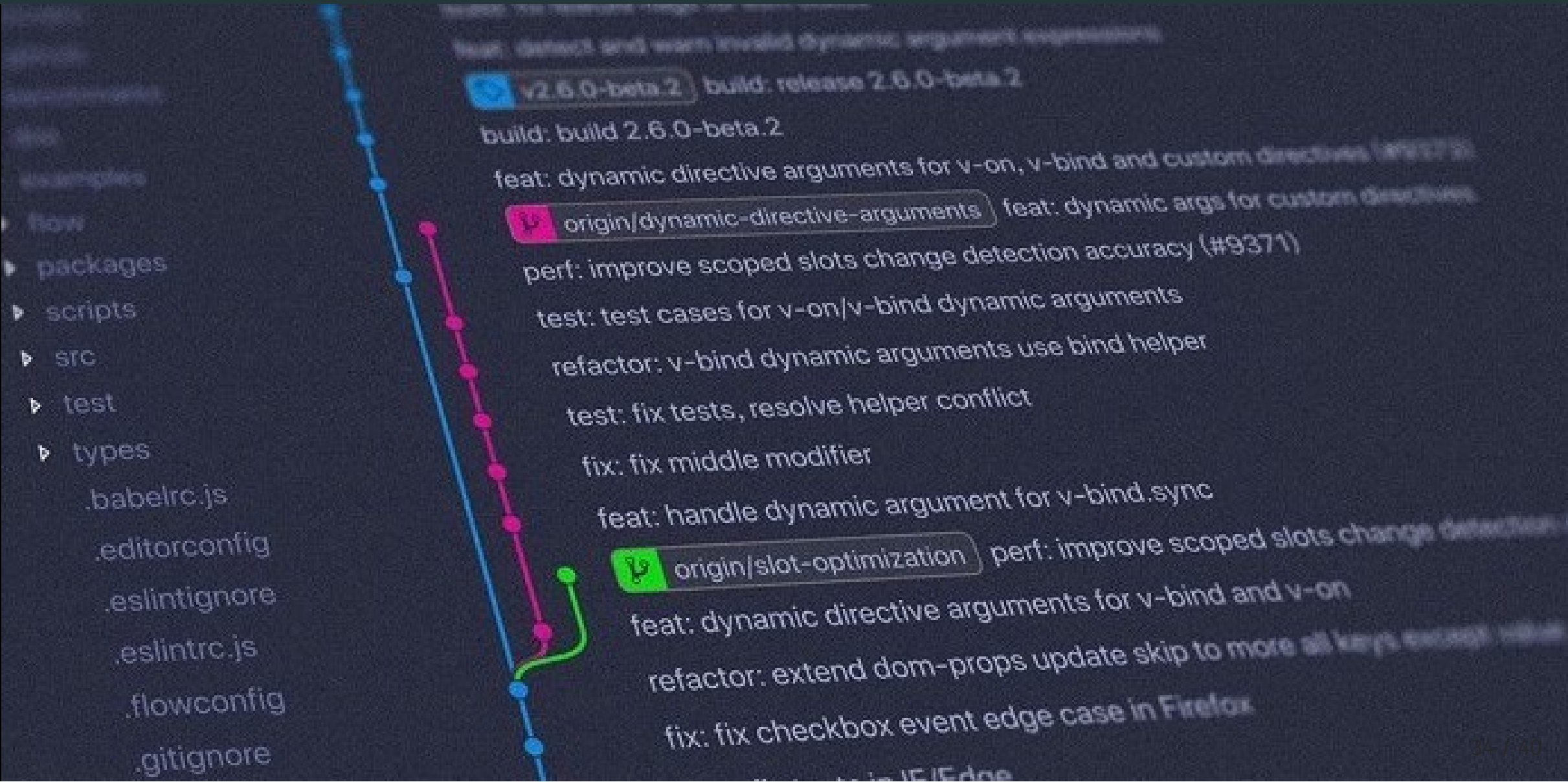
`Credit` ADCR

Sneak preview
Reflecting everyday ethics in data science

How do I respect intellectual property?

How do I protect the safety of my research subjects?

How do I ensure an open science workflow?



How do I communicate results honestly?

Sneak preview
Tools for Data Science Workshop

Tools for Data Science Workshop

https://intro-to-data-science-21-workshop.github.io/

Tools for Data Science Workshop

Wrangling data at scale with data.table

Julia Ellingwood, Renato Franco

Video Materials

Categorical variables with forcats

Janine De Vera, Victor Möslein

Video Materials

Working with JSON and jsonlite

Gabriel da Silva Zech, Francesco Danovi

Video Materials

Geocoding with sf

Andrew Wells, Fernanda Candido Gomes

Video Materials

Text analysis with quanteda

Kathryn Malchow, Federico Mammana

Video Materials

Regular expressions with stringr

Anna Clara Deniz, Angela Duarte Pardo

Video Materials

Creating web APIs with plumber

Hyebin Hong, Guilherme Lacerda

Video Materials

Dynamic web scraping with RSelenium

Reed Garvin, Francesca Giacco

Video Materials

Q & A

Question time!

Please AMA!

- About the course
- About the MDS curriculum
- About the Hertie School
- About the Data Science Lab
- About anything else

SEMESTER 1	Data structures and algorithms
	Introduction to data science
	Public policy
	Economics
SEMESTER 2	Mathematics for data science
	Causal inference
	Machine learning
	Law and governance
-	Internship
SEMESTER 3 + 4	2 Data Science concentration electives
	2 Governance and Management for Data Science electives
	2 Portfolio electives: Expand your policy and data science portfolio
	Master's thesis