Data Handling: Import, Cleaning and Visualisation

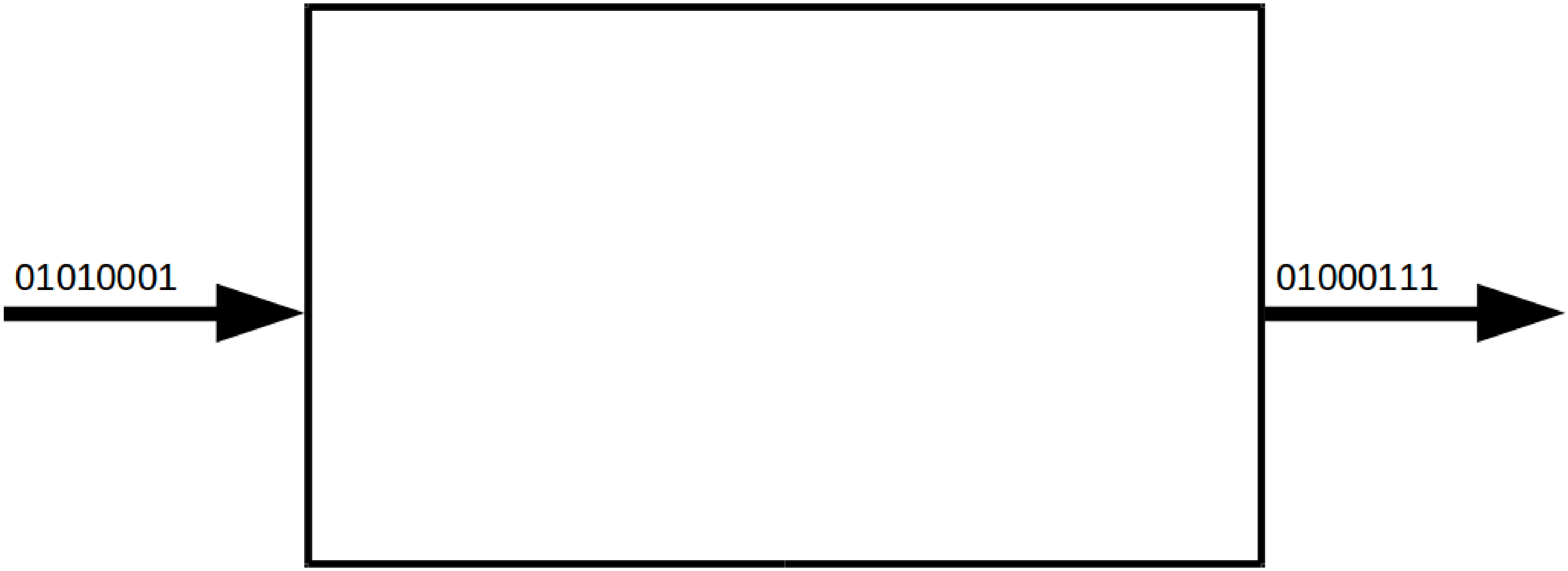
Lecture 1 :

Introduction

Prof. Dr. Ulrich Matter

Welcome to Data Handling 2022!

- Go to this page (or use the QR code): https://bit.ly/datahandling-2022
- Use one row to respond to the questions in the column headers (see the first two rows for examples).

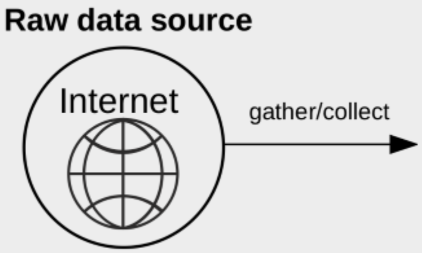

01010001
01000111

Data (science) pipeline

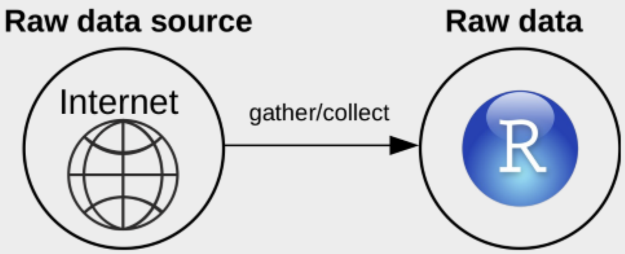
Raw data source

Internet

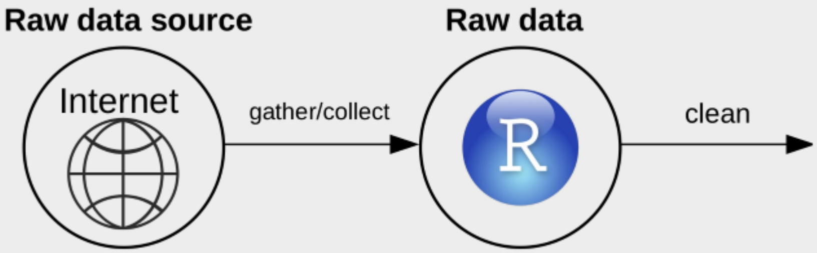
Data (science) pipeline



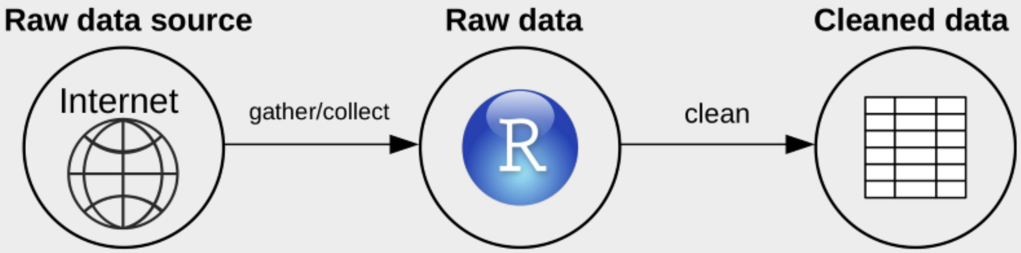
Data (science) pipeline



Data (science) pipeline



Data (science) pipeline



Raw data source

Internet

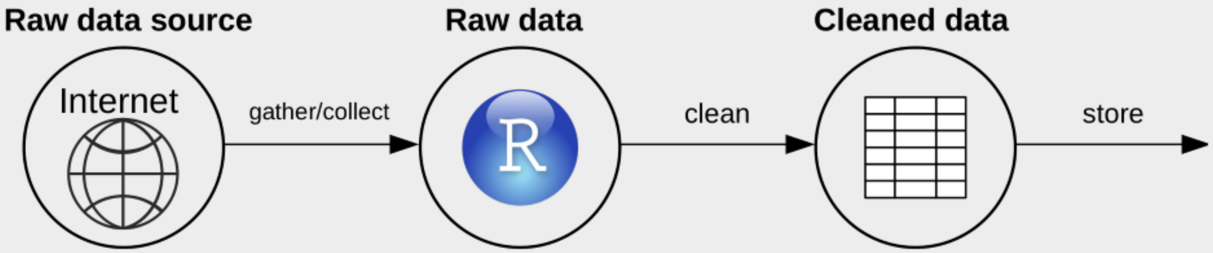
gather/collect

Raw data

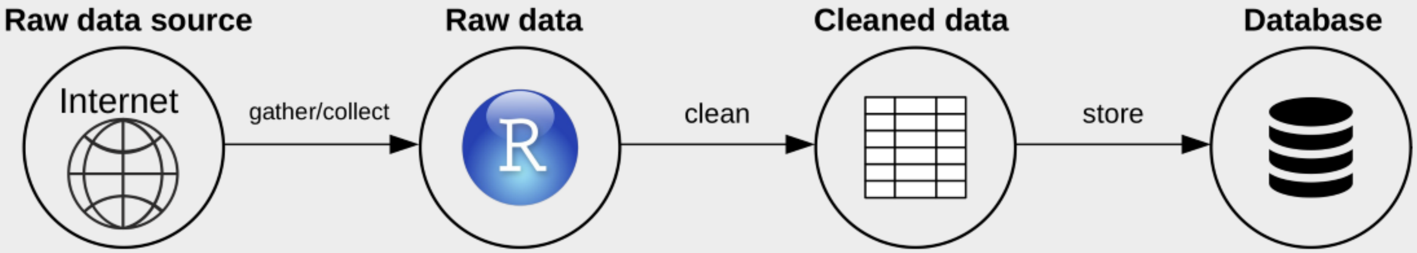
clean

Cleaned data

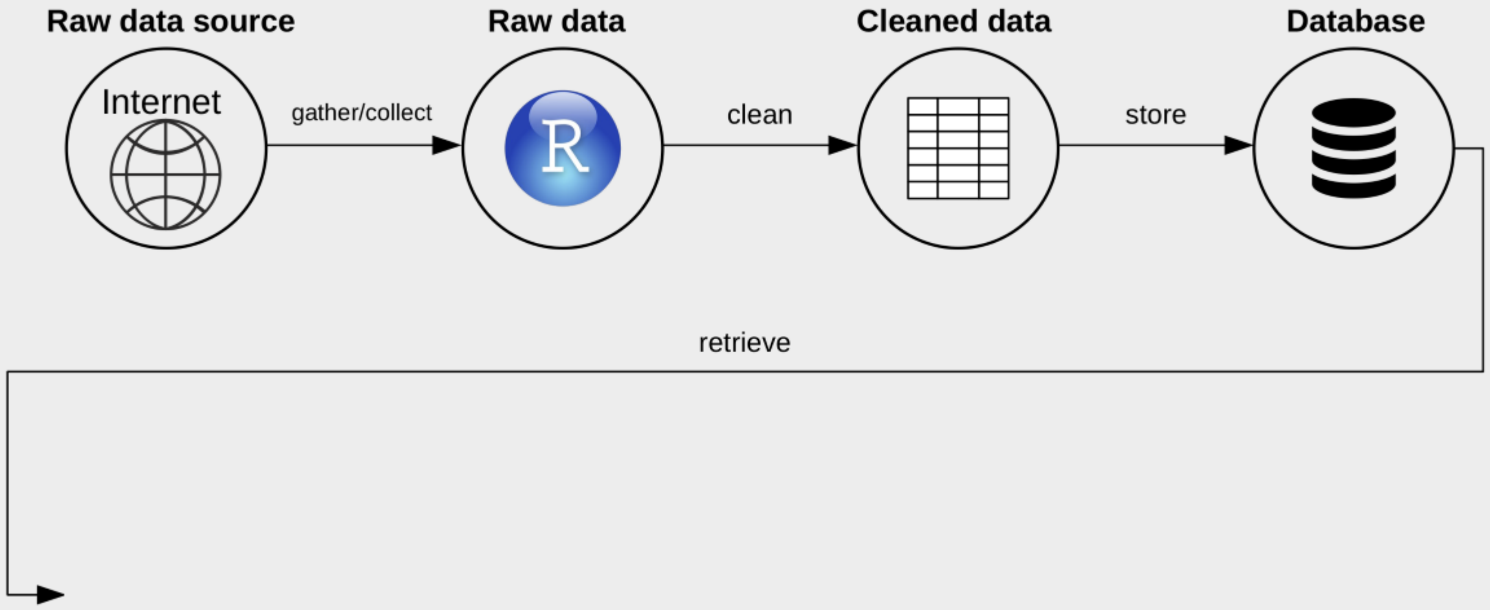
Data (science) pipeline



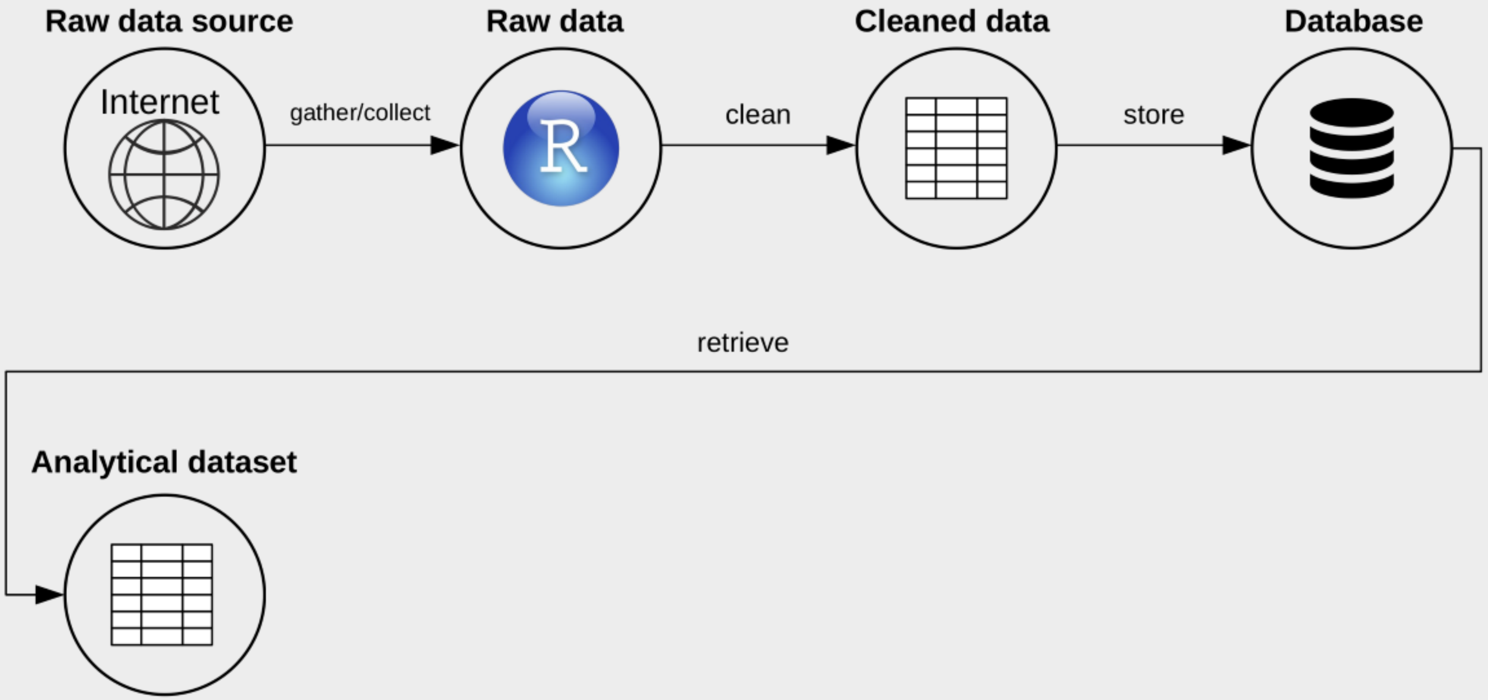
Data (science) pipeline



Data (science) pipeline



Data (science) pipeline



Raw data source

Internet

gather/collect

Raw data

clean

Cleaned data

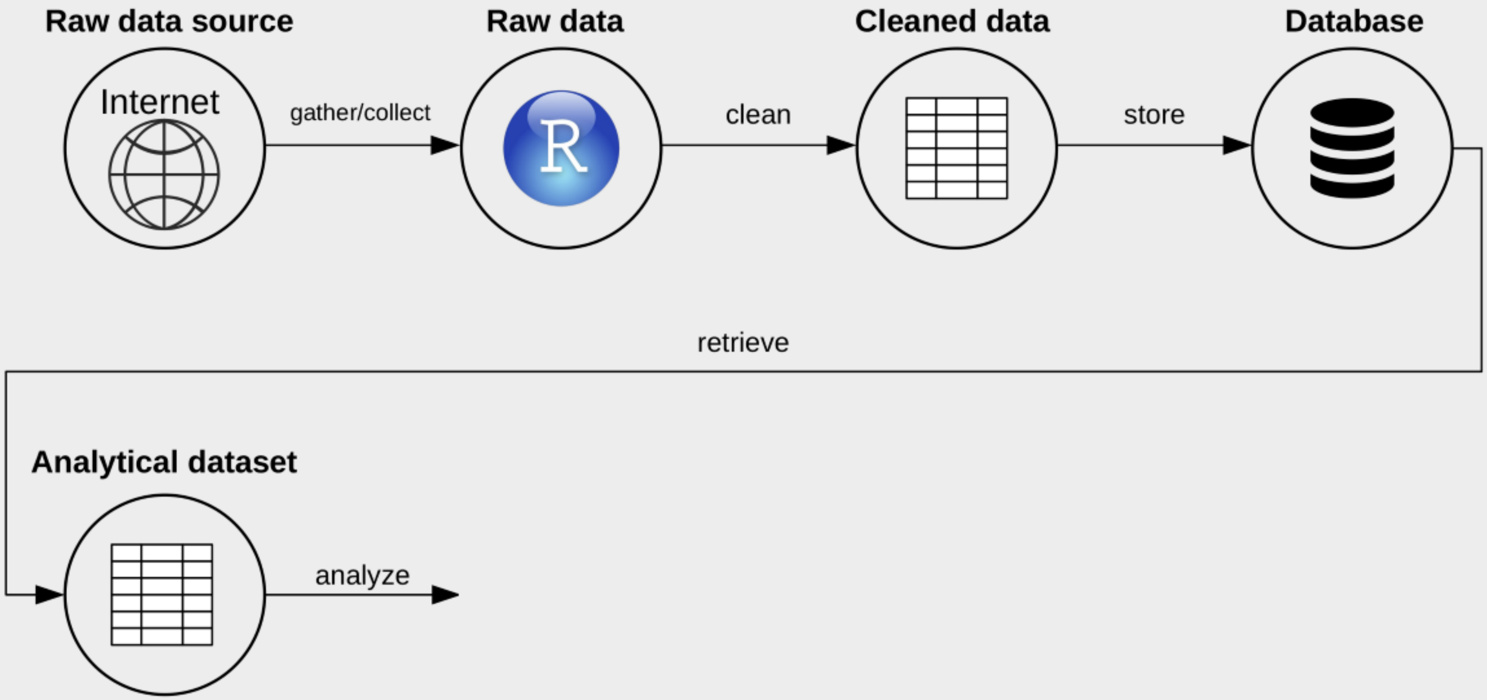
store

Database

retrieve

Analytical dataset

Data (science) pipeline



Raw data source

Internet

gather/collect

Raw data

clean

Cleaned data

store

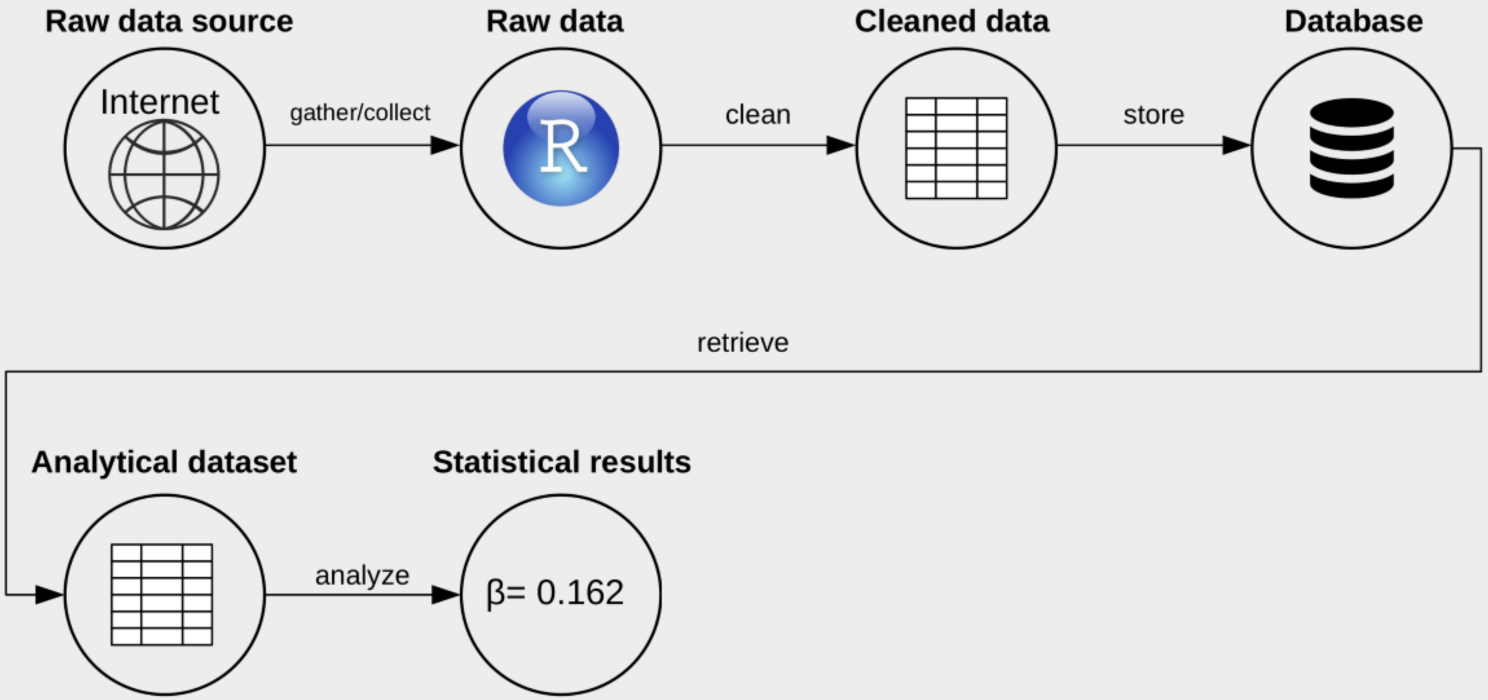
Database

retrieve

Analytical dataset

analyze

Data (science) pipeline



Raw data source

Internet

gather/collect

Raw data

clean

Cleaned data

store

Database

retrieve

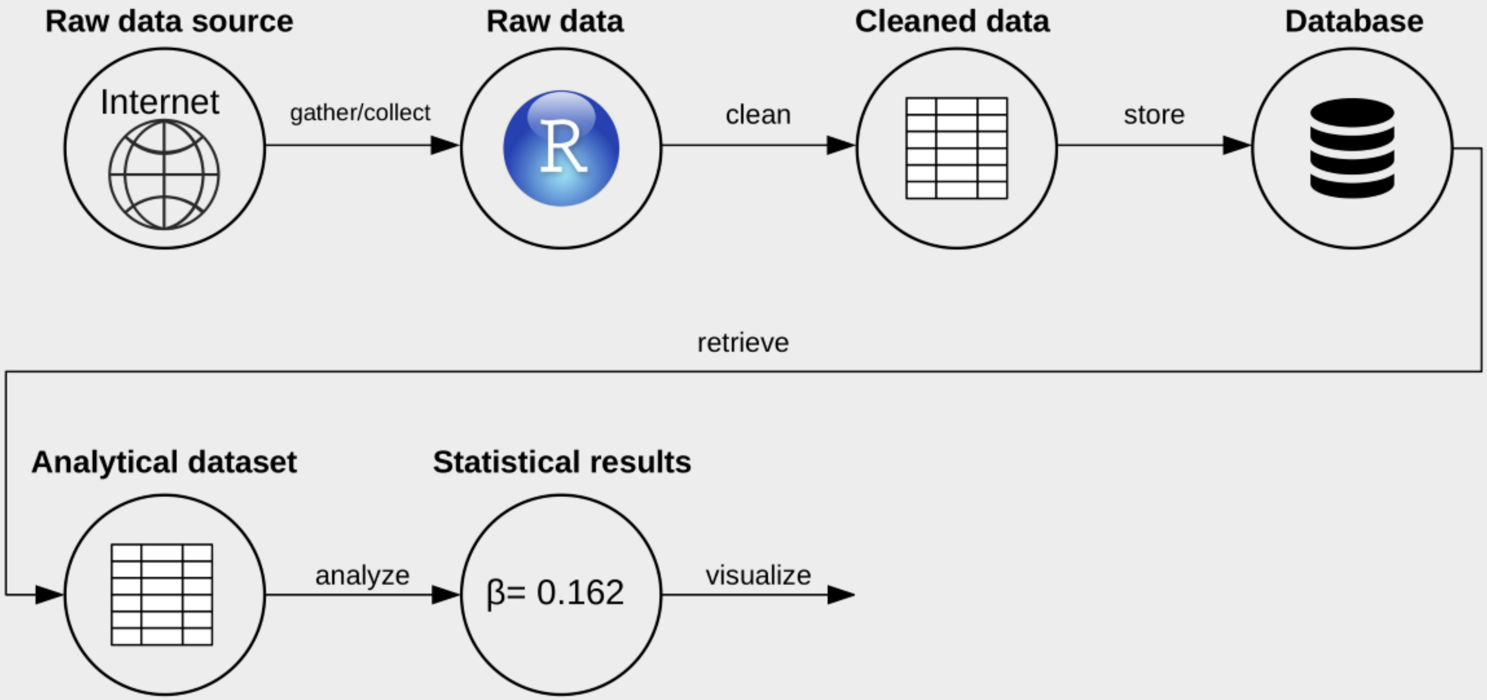
Analytical dataset

analyze

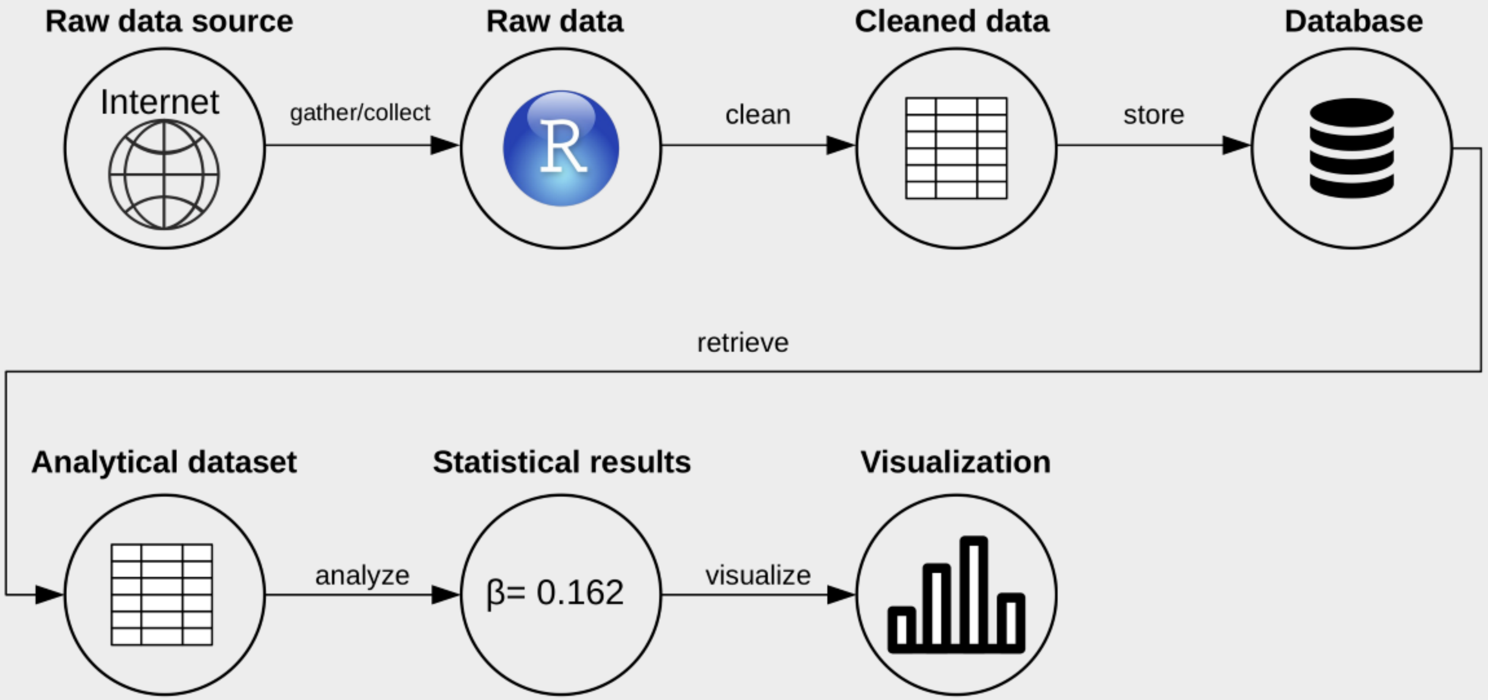
Statistical results

β= 0.162

Data (science) pipeline



Data (science) pipeline



Raw data source

Internet

gather/collect

Raw data

R

clean

Cleaned data

store

Database

retrieve

Analytical dataset

analyze

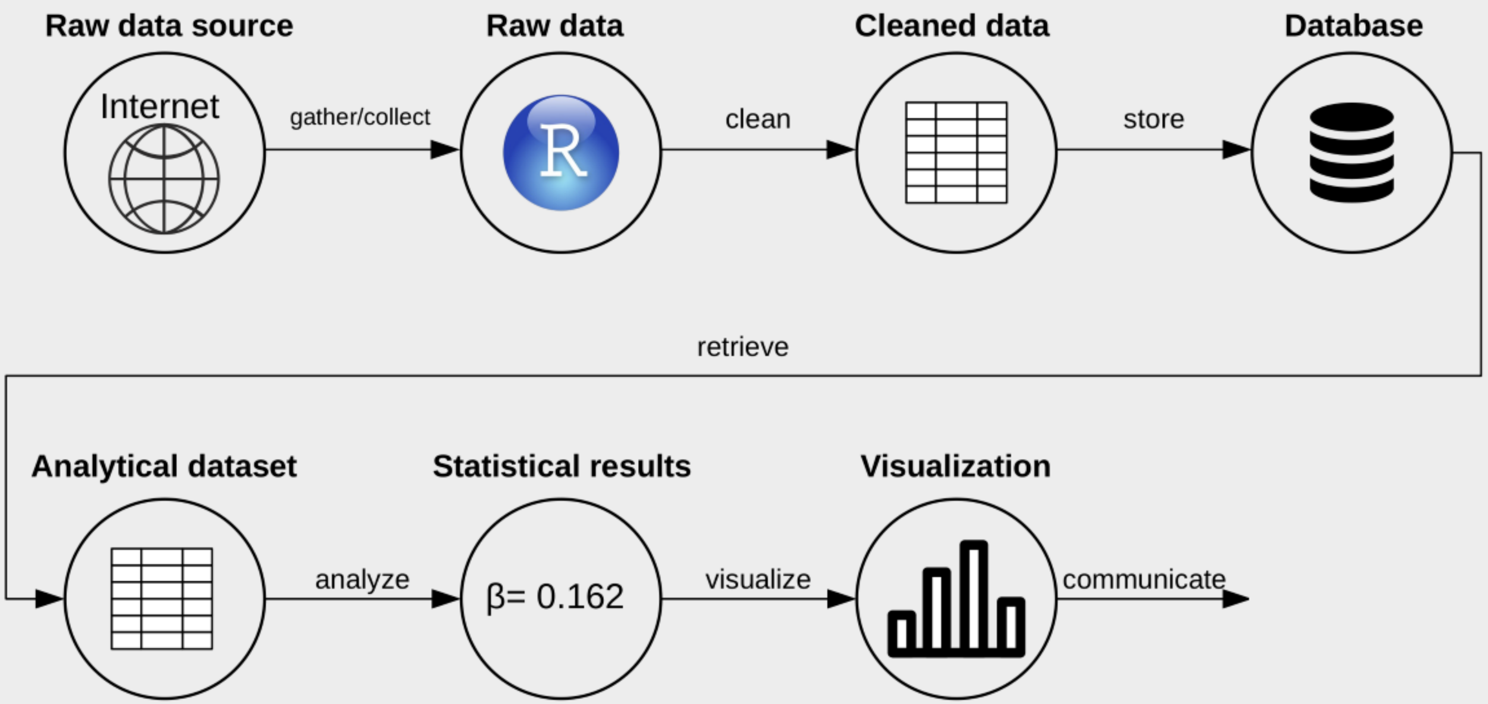
Statistical results

β= 0.162

visualize

Visualization

Data (science) pipeline



Raw data source

Internet

gather/collect

Raw data

clean

Cleaned data

store

Database

retrieve

Analytical dataset

analyze

Statistical results

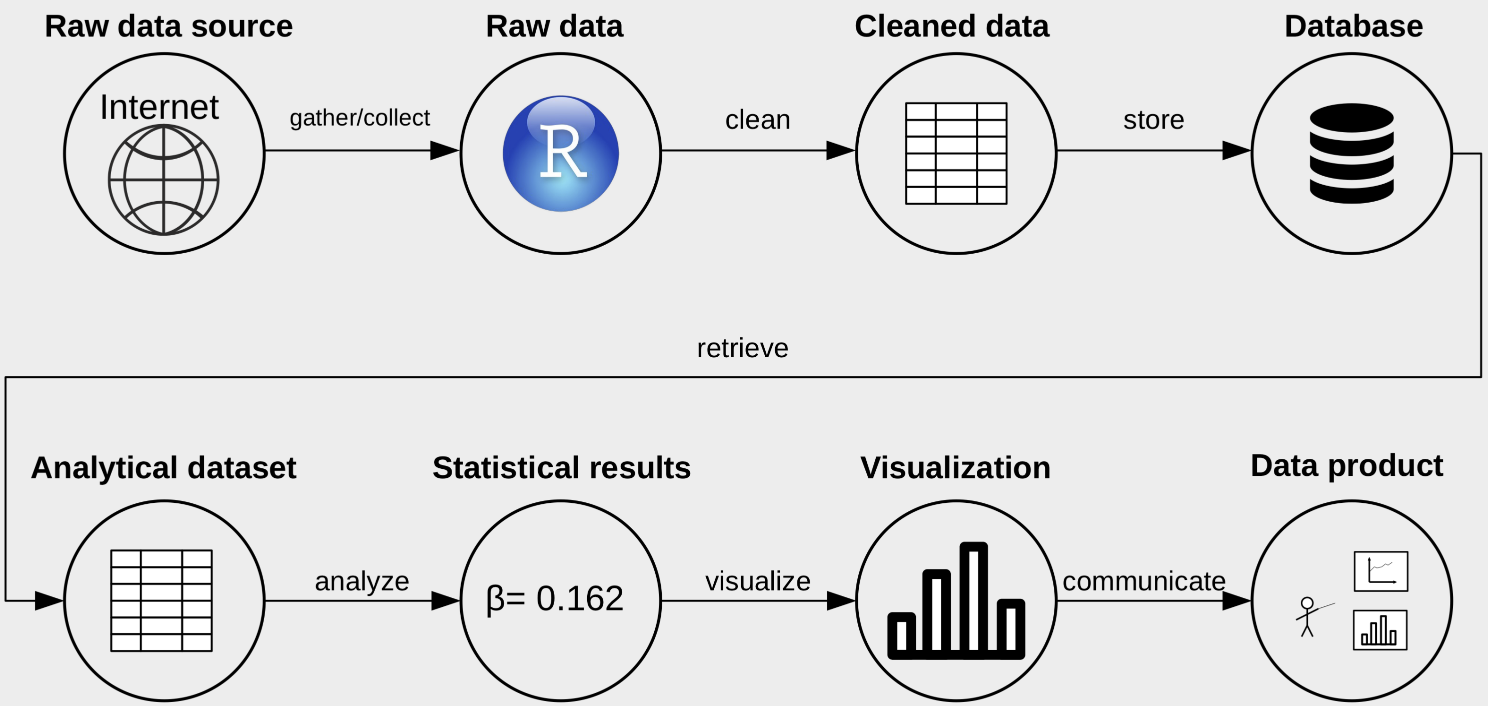
β= 0.162

visualize

Visualization

communicate

Data (science) pipeline



Raw data source

Internet

gather/collect

Raw data

clean

Cleaned data

store

Database

retrieve

Analytical dataset

analyze

Statistical results

β= 0.162

visualize

Visualization

communicate

Data product

Background

‘Data Science’?

“This coupling of scientific discovery and practice involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of scientific, translational, and inter-disciplinary applications.”

University of Michigan ‘Data Science Initiative’, 2015

But, what about statistics?!

“Seemingly, statistics is being marginalized here; the implicit message is that statistics is a part of what goes on in data science but not a very big part. At the same time, many of the concrete descriptions of what the DSI will actually do will seem to statisticians to be bread-and-butter statistics. Statistics is apparently the word that dare not speak its name in connection with such an initiative!”

David Donoho (2015). 50 years of Data Science

What’s new about all this?

“All in all, I have come to feel that my central interest is in data analysis, which I take to include, among other things: …”

What’s new about all this?

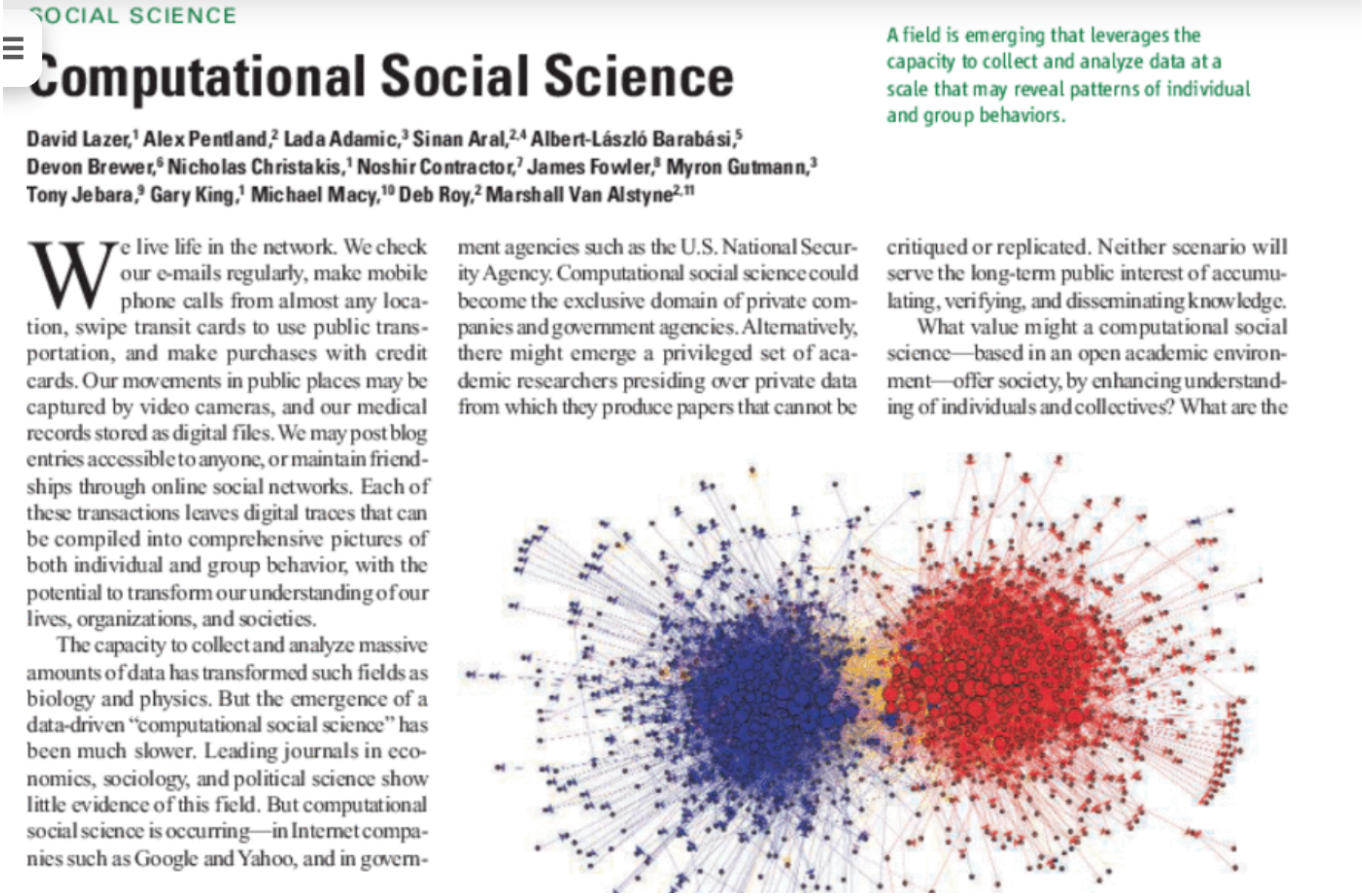
“All in all, I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.”

What's new about all this?

John Tukey (The Future of Data Analysis, 1962!)

Technological change

Relevance for modern economic research



SOCIAL SCIENCE

Computational Social Science

A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

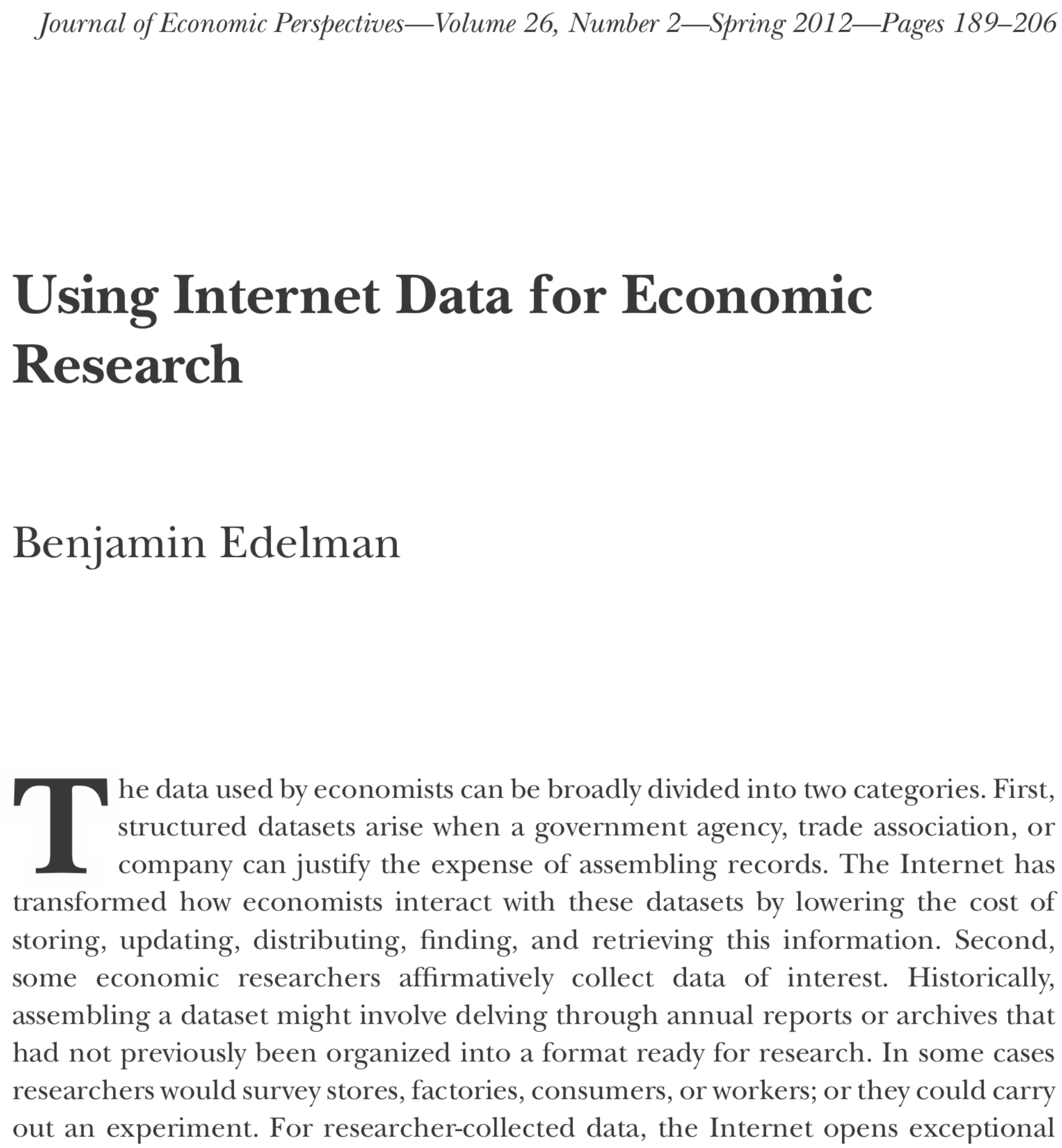
David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3] Tony Jebara,[9] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven "computational social science" has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring—in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

What value might a computational social science—based in an open academic environment—offer society, by enhancing understanding of individuals and collectives? What are the

Relevance for modern economic research



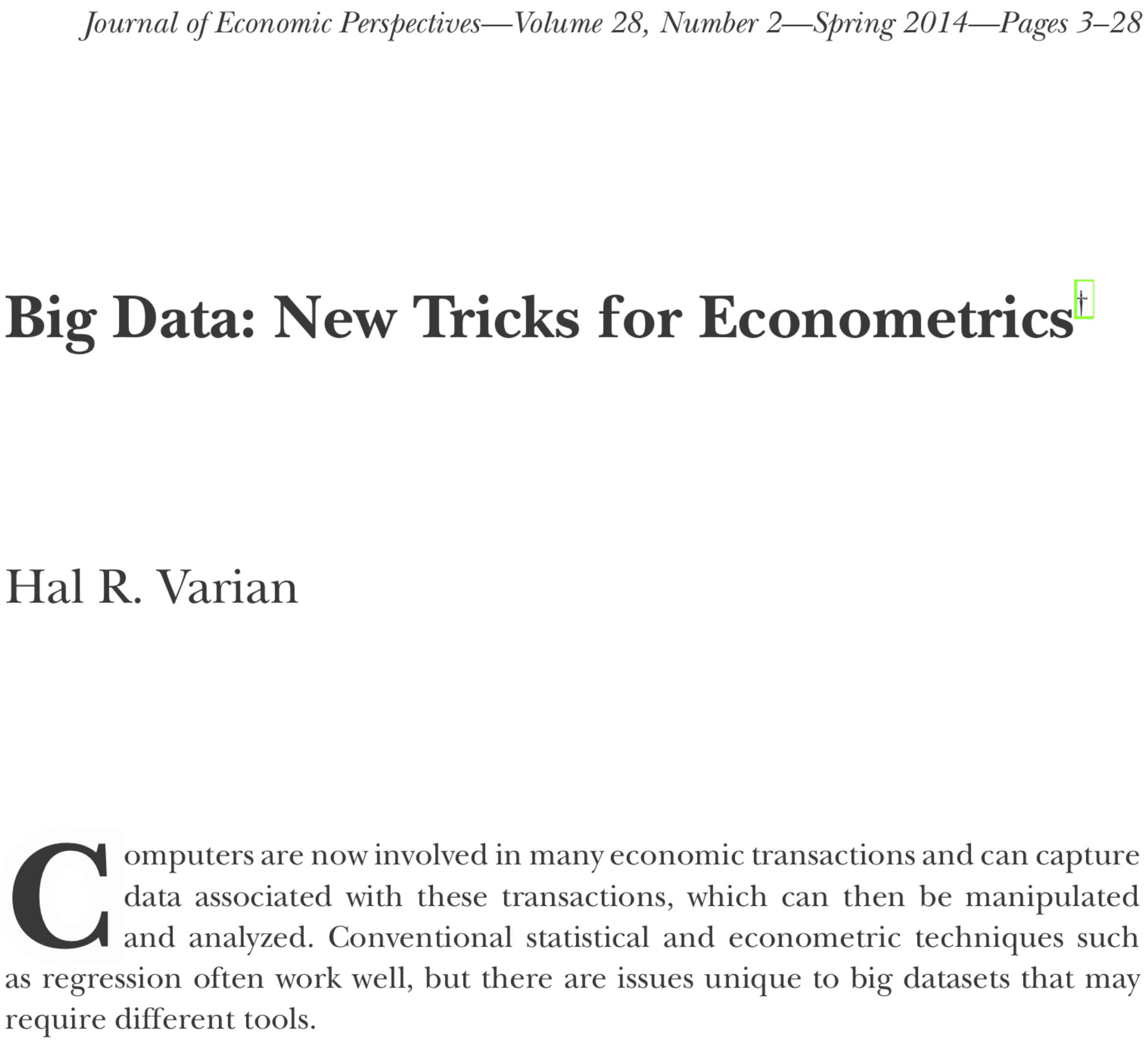
Journal of Economic Perspectives—Volume 26, Number 2—Spring 2012—Pages 189–206

Using Internet Data for Economic Research

Benjamin Edelman

The data used by economists can be broadly divided into two categories. First, structured datasets arise when a government agency, trade association, or company can justify the expense of assembling records. The Internet has transformed how economists interact with these datasets by lowering the cost of storing, updating, distributing, finding, and retrieving this information. Second, some economic researchers affirmatively collect data of interest. Historically, assembling a dataset might involve delving through annual reports or archives that had not previously been organized into a format ready for research. In some cases researchers would survey stores, factories, consumers, or workers; or they could carry out an experiment. For researcher-collected data, the Internet opens exceptional

Relevance for modern economic research



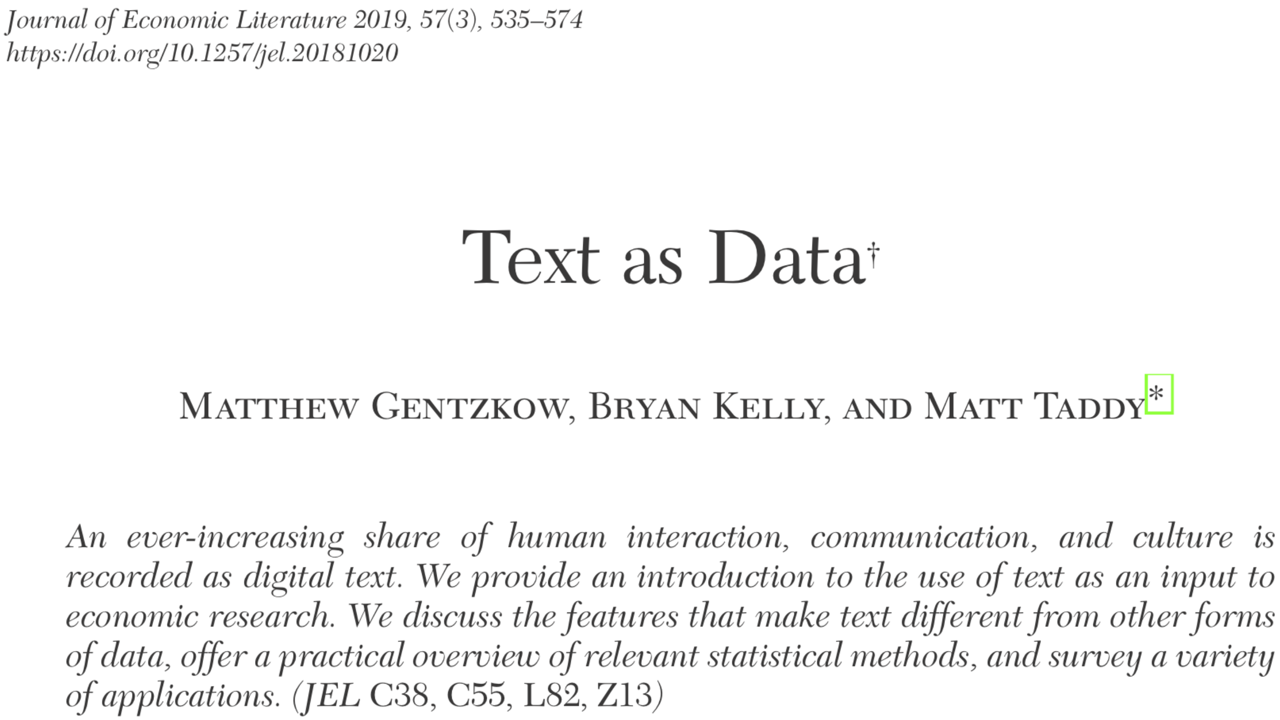
Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 3–28

Big Data: New Tricks for Econometrics[†]

Hal R. Varian

Computers are now involved in many economic transactions and can capture data associated with these transactions, which can then be manipulated and analyzed. Conventional statistical and econometric techniques such as regression often work well, but there are issues unique to big datasets that may require different tools.

Relevance for modern economic research



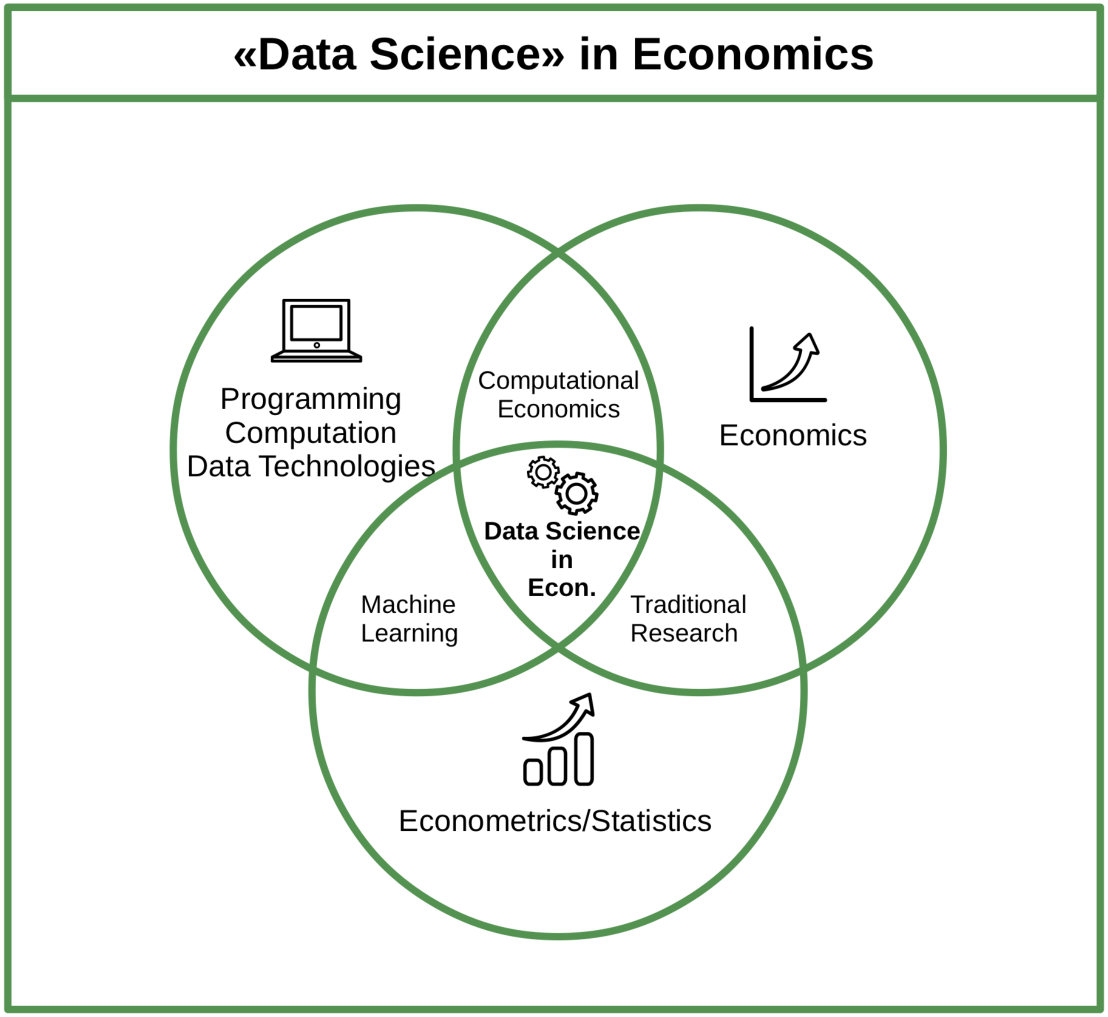
Journal of Economic Literature 2019, 57(3), 535–574
https://doi.org/10.1257/jel.20181020

Text as Data[†]

Matthew Gentzkow, Bryan Kelly, and Matt Taddy[*]

An ever-increasing share of human interaction, communication, and culture is recorded as digital text. We provide an introduction to the use of text as an input to economic research. We discuss the features that make text different from other forms of data, offer a practical overview of relevant statistical methods, and survey a variety of applications. (JEL C38, C55, L82, Z13*)*

Data science in economics skill set



«Data Science» in Economics

Programming
Computation
Data Technologies

Computational
Economics

Economics

Data Science
in
Econ.

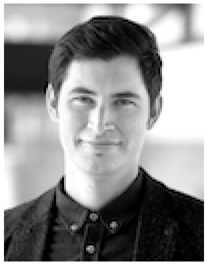
Machine
Learning

Traditional
Research

Econometrics/Statistics

Organisation of the Course

Our Team - At Your Service



Aurélien Sallin | Michael Tüting | Ulrich Matter

Introduction: Ulrich Matter

- 2017-today: Assistant Professor of Economics, University of St.Gallen
- 2021-today: Guest Lecturer (Applied Data Science), University of Lucerne

Previously:

Universität Basel

UNIVERSITY OF OXFORD

Saïd Business School

VE RI TAS

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

Introduction: Ulrich Matter

Research:

- Can the personalization of Google search results lead to political polarization?
- Does YouTube’s recommender algorithm lead to radicalization?
- Do politicians vote in the interest of money donors when voters are distracted?

Introduction: Ulrich Matter

Teaching:

- Well, this course…
- Big Data Analytics (Master)
- Introduction to Web Mining (Master)
- Economics in Practice (Master)

Course Structure

Course concept: lectures

- Lectures (Thursday morning)
    - Background/Concepts
    - Illustration concepts
    - Illustration of ‘hands-on’ approaches

Course concept: special lectures

- 27.10.2022: Research insights
  - Ulrich Matter: Web Data
  - Aurélien Sallin: Text as Data
  - Michael Tüting: Images as Data

24/11/2022: Guest lecture: Economic Data Science, SNB

Dr. Matthias Gubler

Head of Economic Data Science, SNB

Swiss National Bank

Dr. Helge Liebert

Economist, SNB

Course concept: exercises

- Exercise sheets (handed out every other week)
  - Some conceptual questions
  - Hands-on exercises/tutorials in R
  - Detailed solution videos
  - **First Exercises (set up R/RStudio) is available on StudyNet/Canvas today**

Course concept

- Learning mode in this course: Prepare with reading, visit the lecture, recap key concepts in lecture notes (self-study), work on exercises, watch solution video, come to exercise session, repeat…

- Strongly encouraged: (virtual) learning groups!
  - Biweekly exercises provide opportunity.
  - Tackle the tricky exercises together!

Course concept: exercise sessions

- In-class exercise sessions (bi-weekly evening sessions)
    - Discussion of exercises and additional input
    - Recap of concepts
    - Q&A, support
    - time for more coding!

Part I: Data (Science) fundamentals

Date	Topic
22.09.2022	Introduction: Big Data/Data Science, course overview
29.09.2022	Programming with R
29.09.2022	Exercises/Workshop 1: Tools, programming
06.10.2022	An introduction to data and data processing
13.10.2022	Data storage and data structures
13.10.2022	Exercises/Workshop 2: Data storage and data structures
20.10.2022	Web data, text, and images
27.10.2022	Research insights
27.10.2022	Exercises/Workshop 3: Web data, text, and images

Part II: Data gathering and preparation

Date	Topic
17.11.2022	Data sources, data gathering, data import
24.11.2022	Guest Lecture
24.11.2022	Exercises/Workshop 4: Data gathering, data import
01.12.2022	Data preparation and manipulation

Part III: Analysis, visualisation, output

Date	Topic
08.12.2022	Basic statistics and data analysis with R
08.12.2022	Exercises/Workshop 5: Data preparation and applied data analysis with R
15.12.2022	Visualisation, dynamic documents
21.12.2022	Exercises/Workshop 6: Visualization, dynamic documents
22.12.2021	Summary, Wrap-Up, Q&A, Feedback
22.12.2021	Exam for Exchange Students

Core course resources

- All information and materials (notes, slides, course sheet, syllabus, etc.) are available on StudyNet/Canvas.
- Core materials will also be made available on Nuvolos.

Main textbooks

Data Handling Pocket Reference

Murrell, Paul (2009). **Introduction to Data Technologies**, London: Chapman & Hall/CRC.

Wickham, Hadley and Garred Grolemund (2017). **R for Data Science**, 1st Edition. Sebastopol, CA: O'Reilly.

Further resources

- Stackoverflow
- Get inspired in the R blogsphere

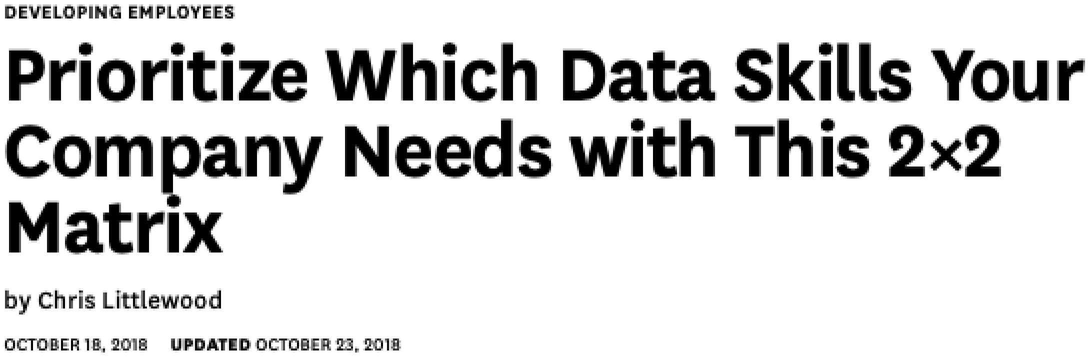
Exam information

- Central, written examination: **digital, BYOD!**, we will have an instructional session by the head of the digital examinations team (data TBD).
- Multiple choice questions.
- A few open questions.
- Theoretical concepts and practical applications in R (questions based on code examples).

Exam information II

- We will release samples of multiple choice questions via Quizzes on Canvas/Studynet (exact same format and style of exam questions).
- Exchange students who need to take the exam before the central exam block:
  - Date, time place, : 22.12.2022, 16:15-18:00, room 01-013.
  - Questions: **michael.tueting@unisg.ch**

And now this…



DEVELOPING EMPLOYEES

Prioritize Which Data Skills Your Company Needs with This 2×2 Matrix

by Chris Littlewood

OCTOBER 18, 2018 **UPDATED** OCTOBER 23, 2018

Q&A

References