Data Handling: Import, Cleaning and Visualisation

Lecture 1: Introduction

Dr. Aurélien Sallin
01/10/2023

Welcome to Data Handling 2024!

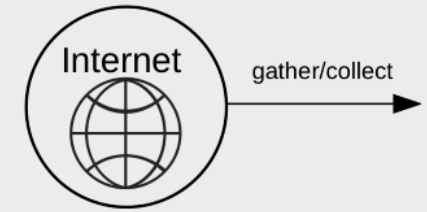
- Go to this app (use the QR code): https://datahandling.shinyapps.io/DataHandlingIntro/
- Use one row to respond to the questions in the column headers (see the first two rows for examples).

Data (science) pipeline

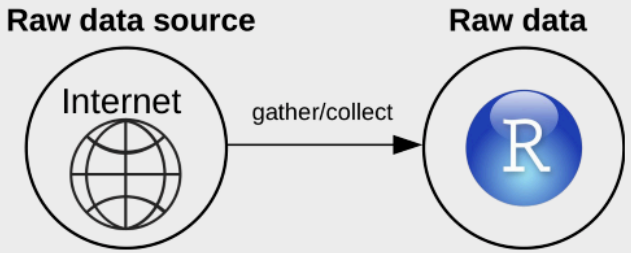
Raw data source

Data (science) pipeline

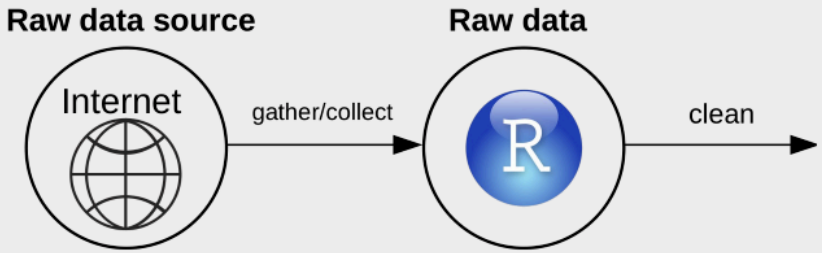
Raw data source



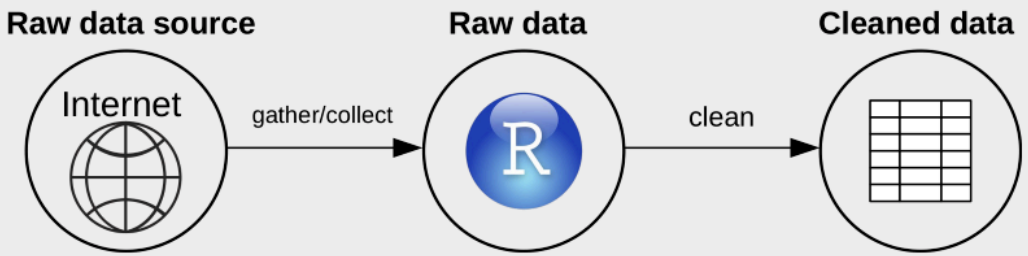
Data (science) pipeline



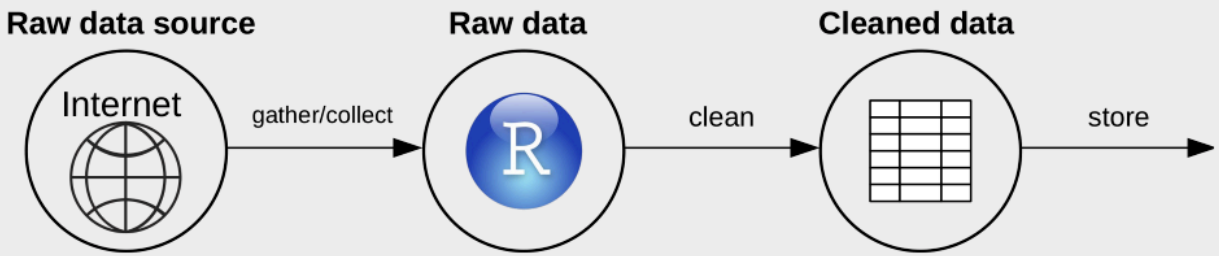
Data (science) pipeline



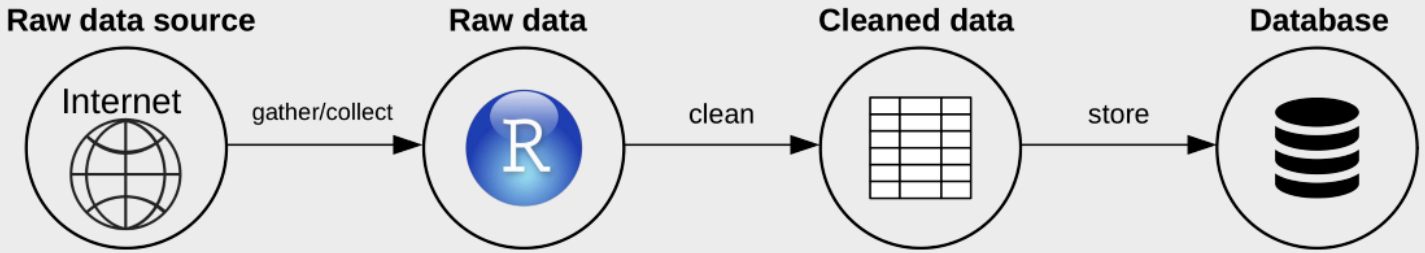
Data (science) pipeline



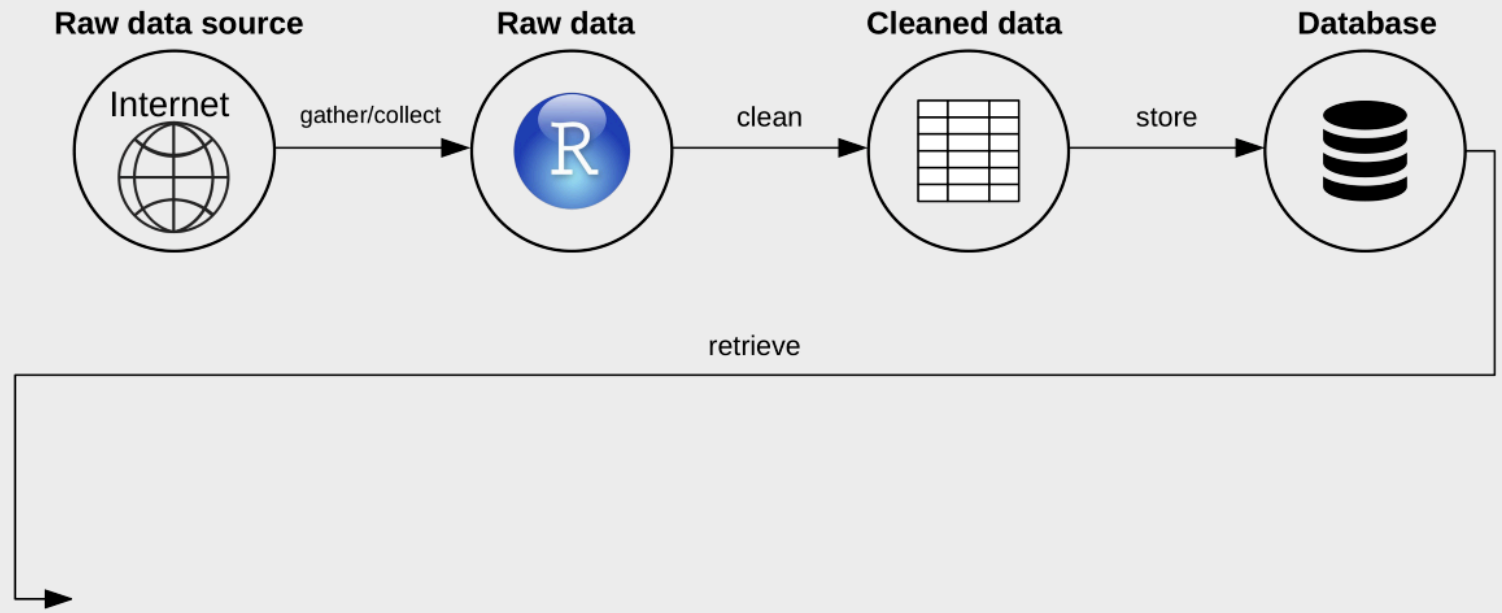
Data (science) pipeline



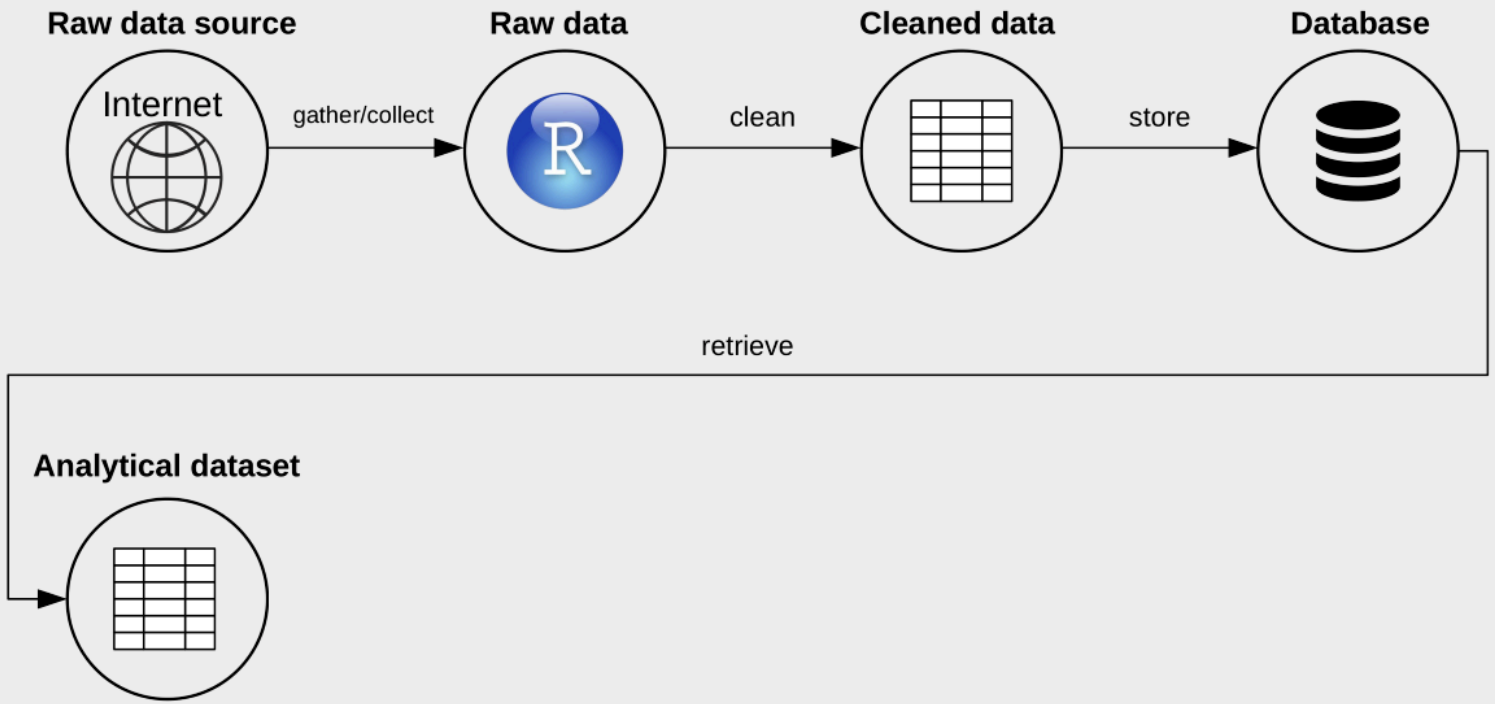
Data (science) pipeline



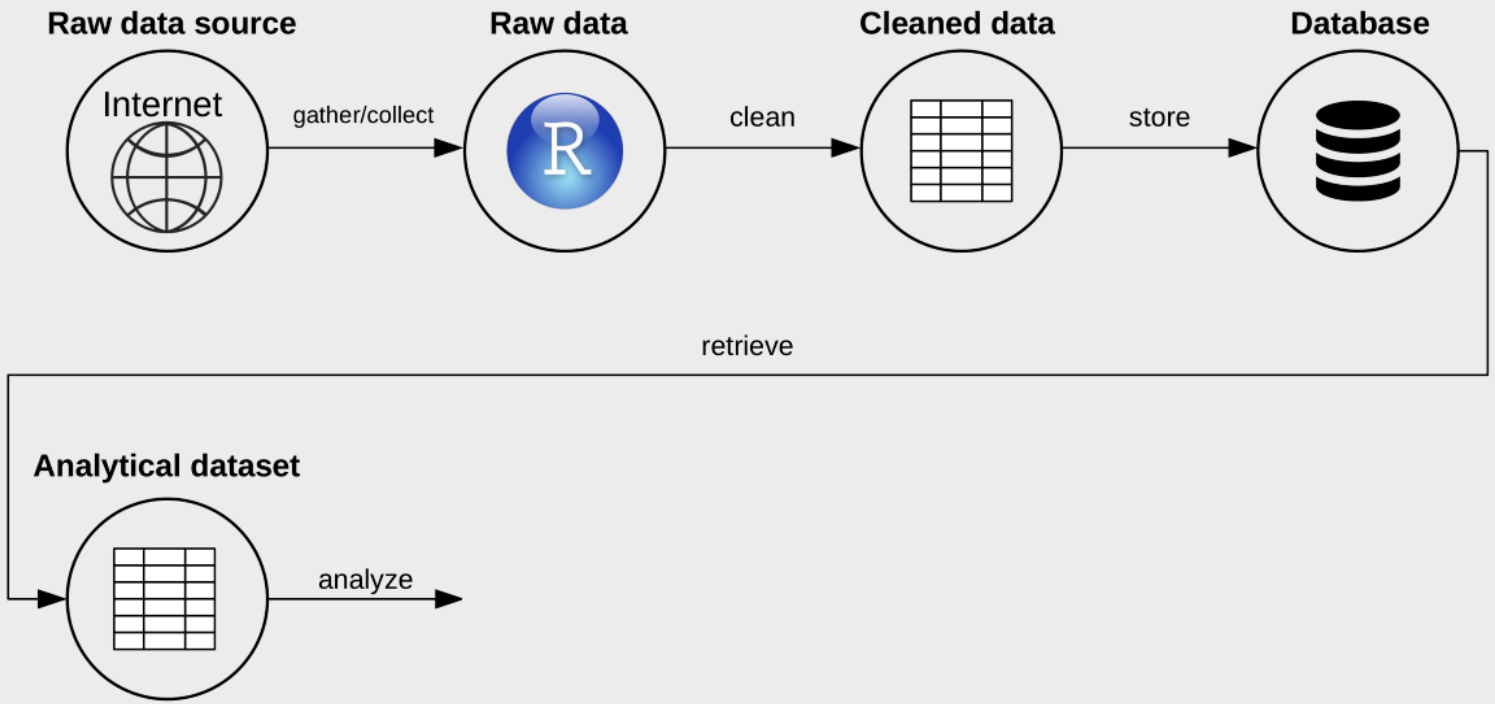
Data (science) pipeline



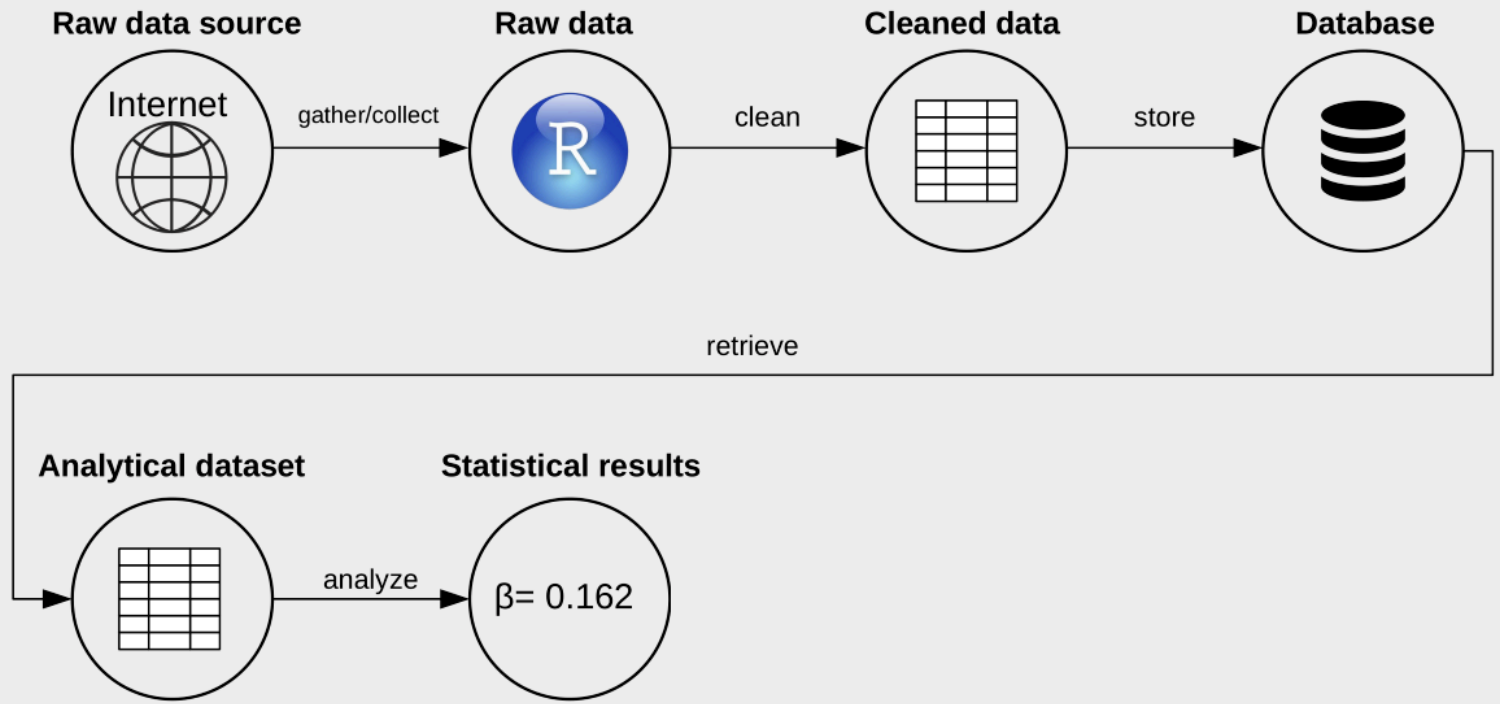
Data (science) pipeline



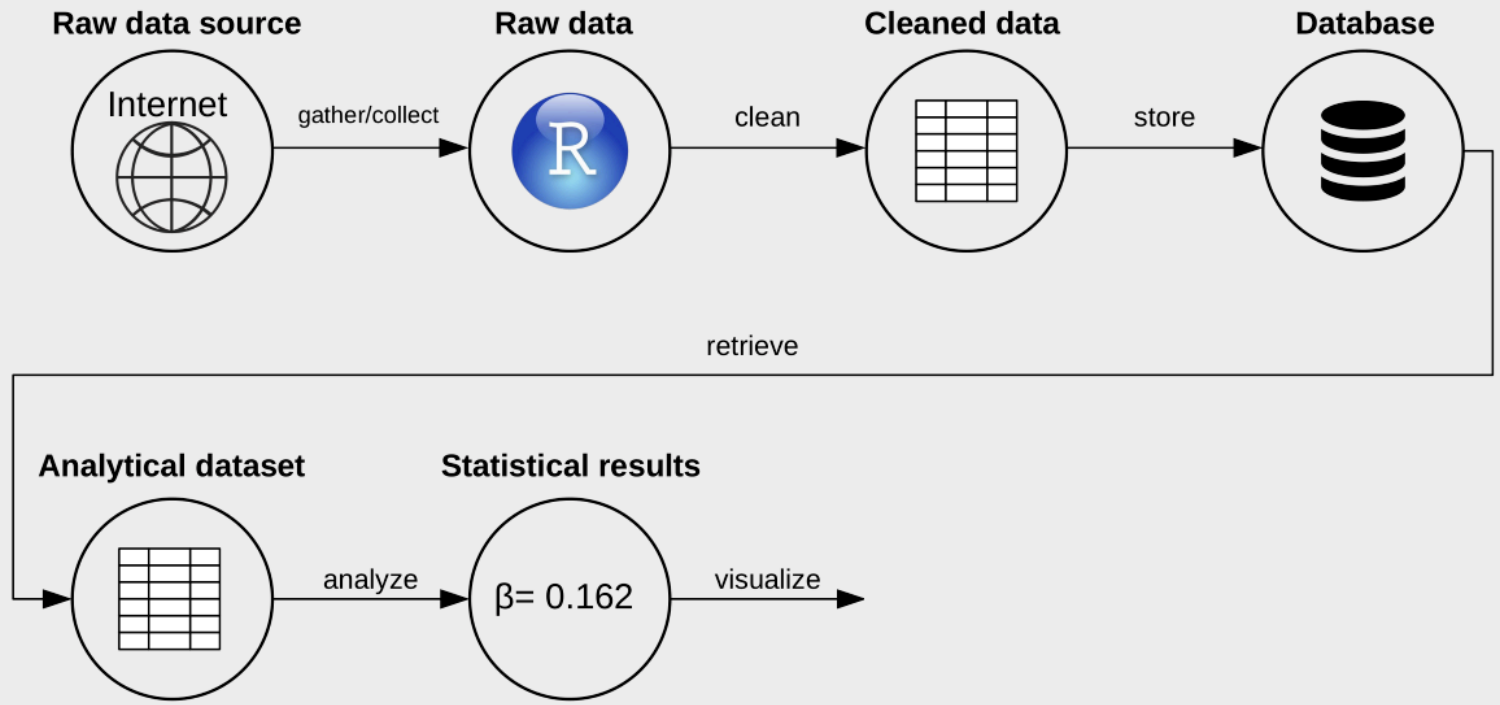
Data (science) pipeline



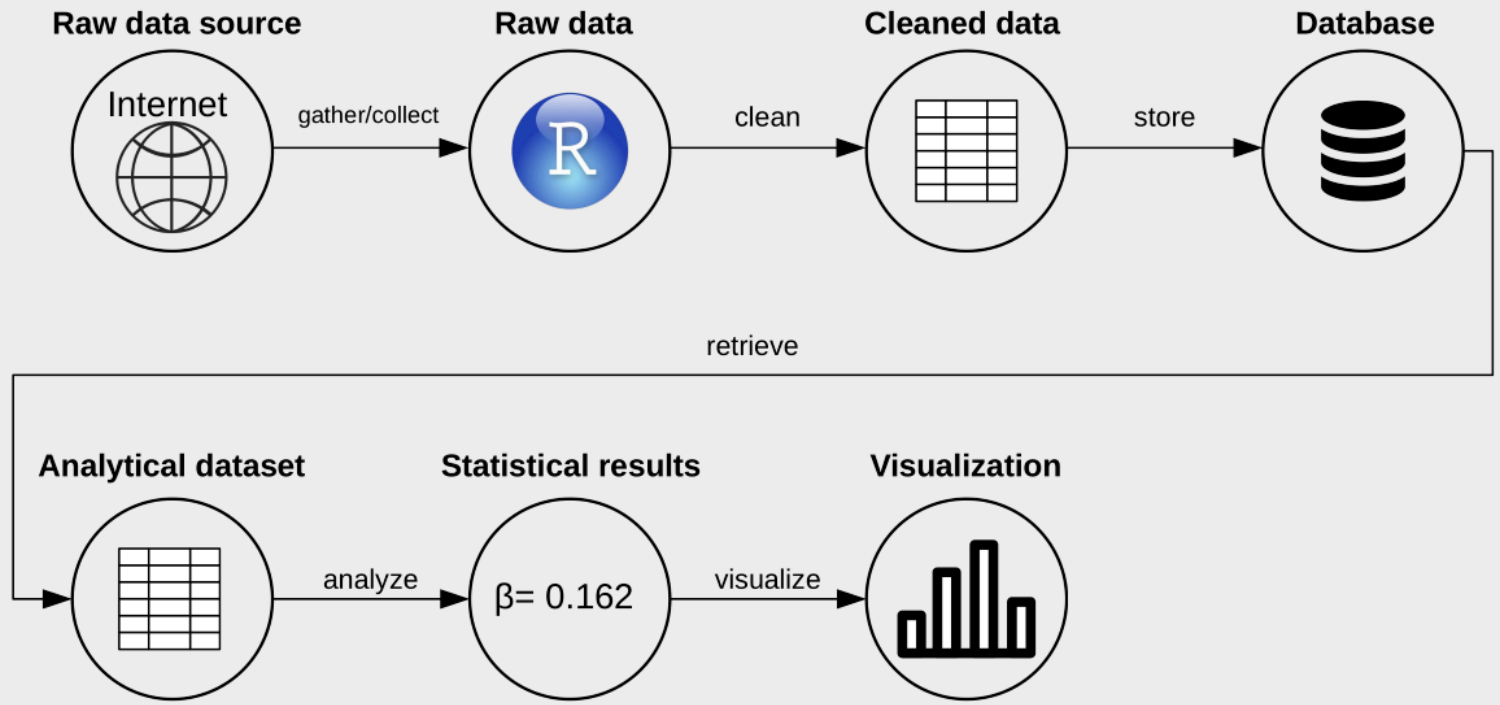
Data (science) pipeline



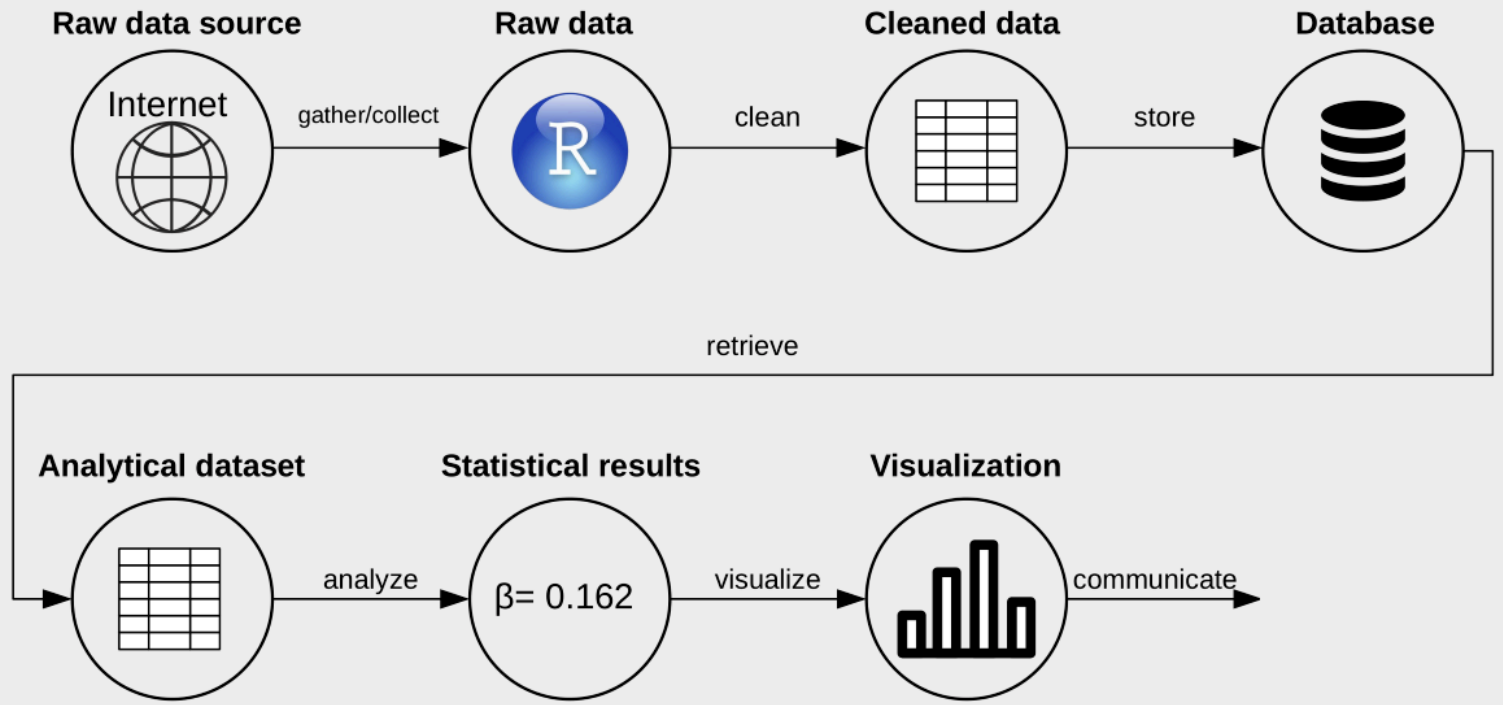
Data (science) pipeline



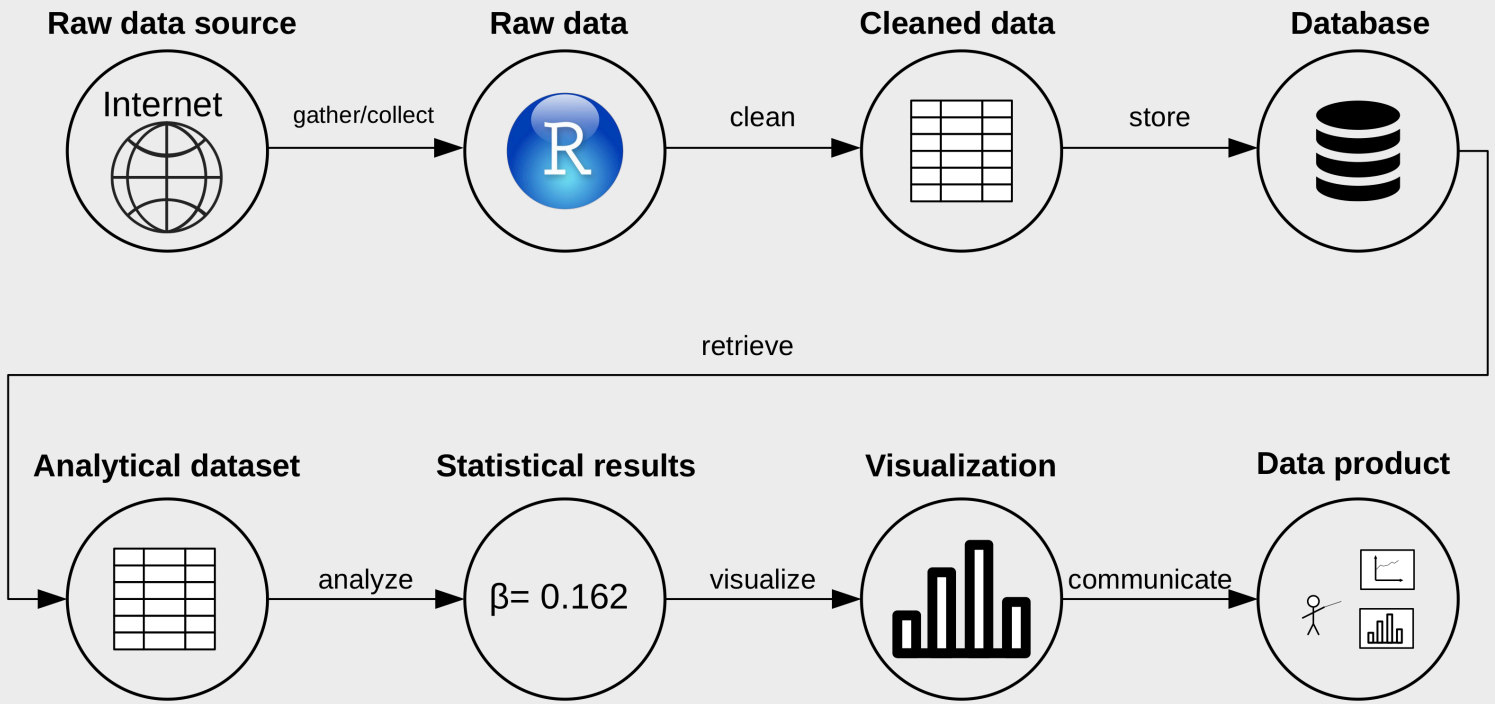
Data (science) pipeline



Data (science) pipeline



Data (science) pipeline



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

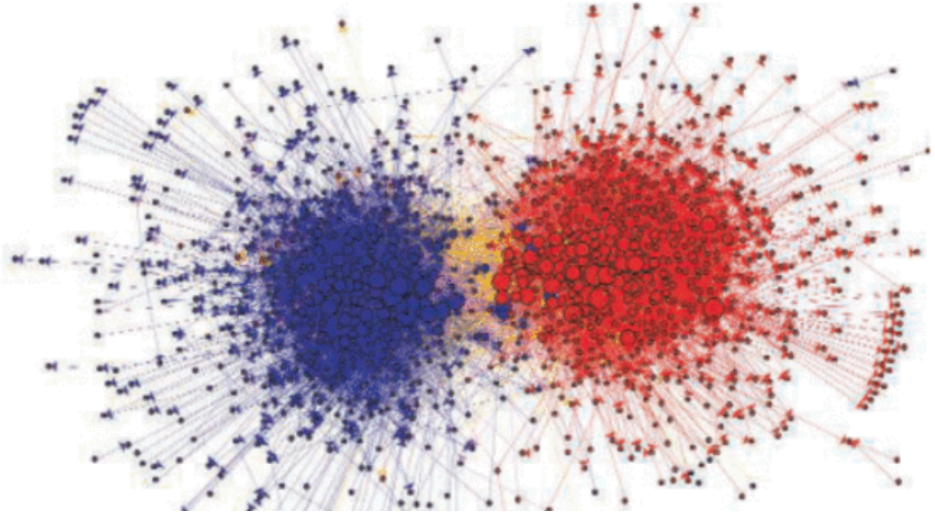
A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

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

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

Relevance for modern economic research

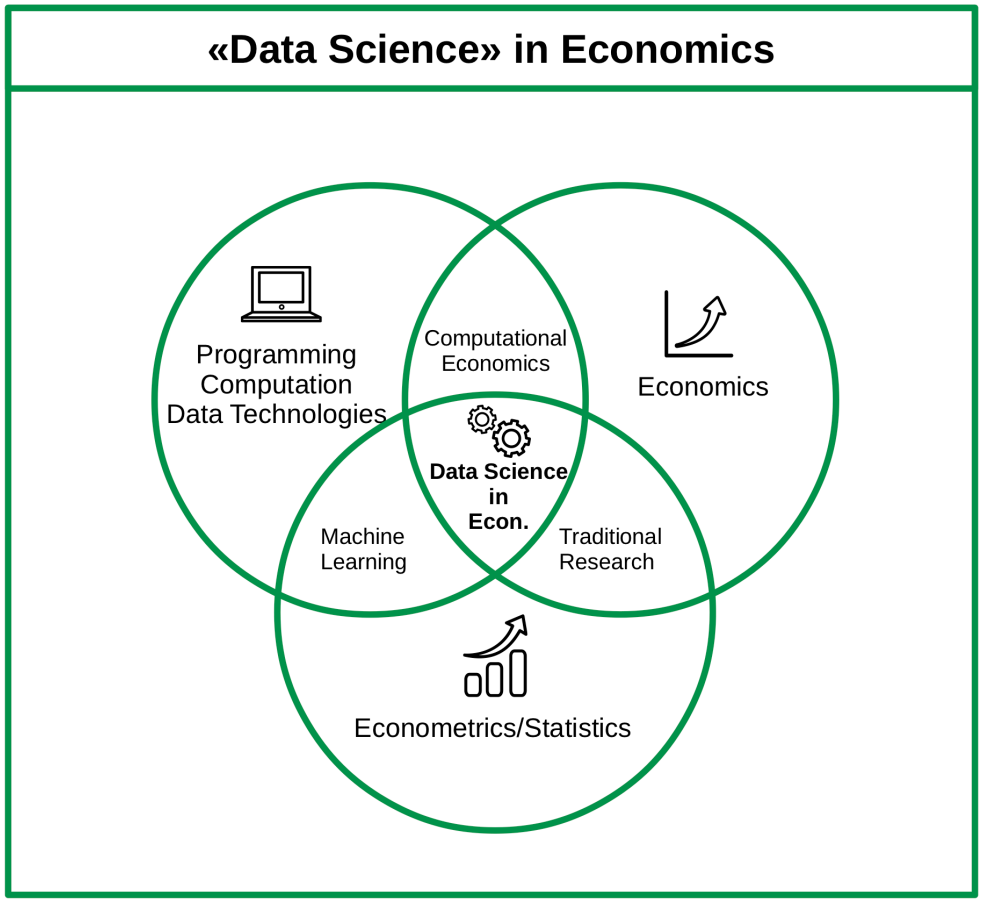
Journal of Economic Literature 2019, 57(3), 535–574
https://doi.org/10.1257/jel.20181020

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 as a life skill

Harvard Business Review

Latest Magazine Ascend Topics Podcasts Video Store The Big Idea Data & Visuals Case Selections

Analytics And Data Science

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and DJ Patil

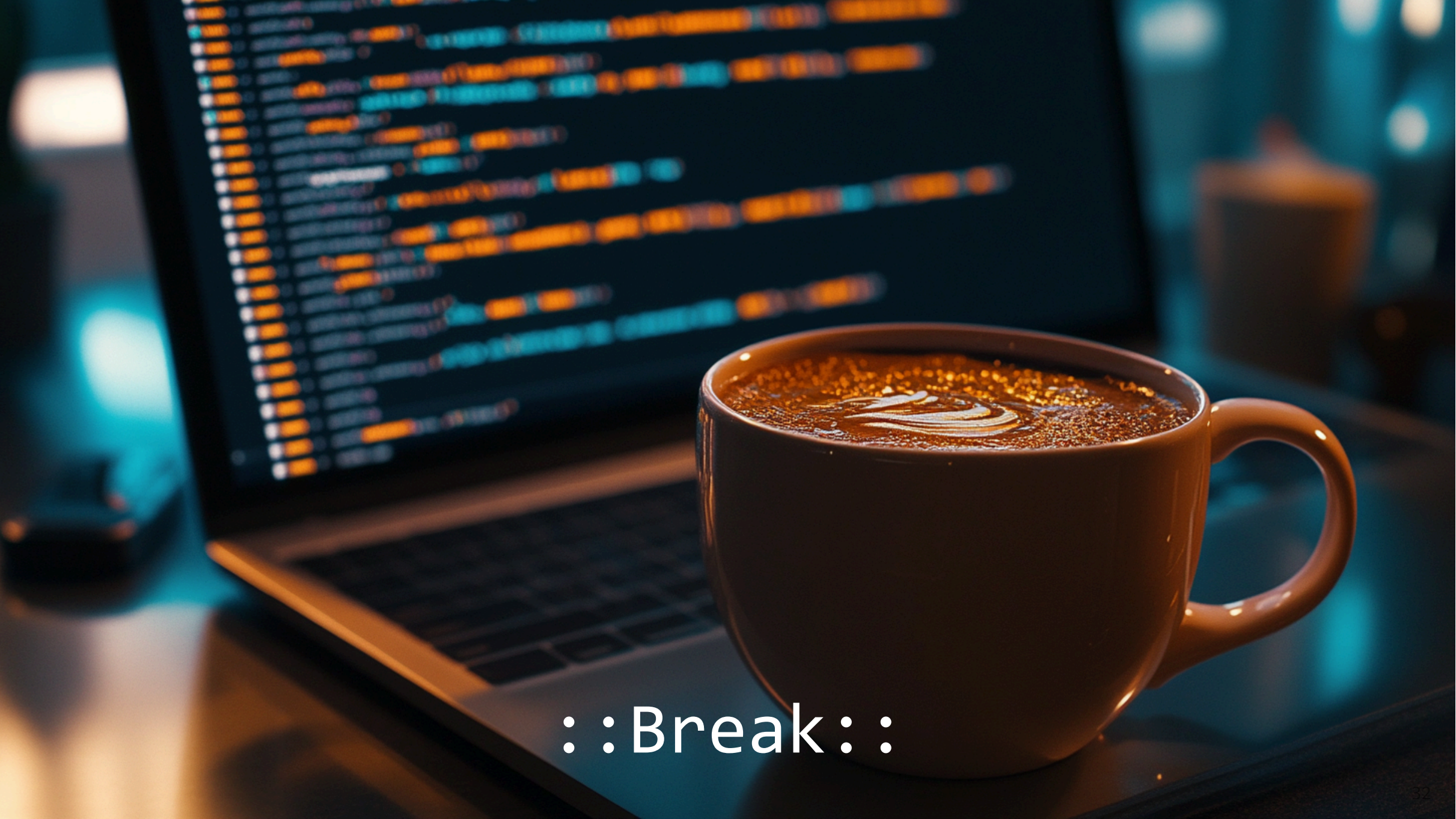
From the Magazine (October 2012)

Data science as a life skill

"More than anything, what data scientists do is **make discoveries while swimming in data.** ... As they make discoveries, they communicate what they've learned and suggest its implications for new business directions. Often they are *creative in displaying information visually and making the patterns they find clear and compelling*...

They advise executives and product managers on the implications of the data for *products, processes, and decisions*.

What kind of person does all this? *Think of him or her as a hybrid of data hacker, analyst, communicator, and trusted adviser. The combination is extremely powerful — and rare.*"


::Break::

Philosophy of this course

A shoemaker and his apprentice c.1914, Emile Adan

At the end of the course, you will be able to...

- **Understand the tools you need when working with data**
  We will use the programming language R, but principles are similar for any other programming language (👞 ⚙️)
- **Work independently with data**
  We will learn how to collect, clean, and analyze data so that you can conduct a data project in Economics (research/consulting/...) from start to finish
- **Ask the right questions to a dataset**
  We will learn how to ask the right questions to a dataset
- **Learn to communicate about data**
  We will learn to present our results in a clear and compelling way

My commitment to these goals and to your learning process

- Transferrable skills
- Hands-on approach
- Emphasis on real-world relevance
  (caveat: this course is mandatory for Econ students, I have limited freedom in the syllabus)
- As much fun as possible (as coding can be fun… 😎)

Your commitment to the course

- Prepare with reading, visit the lecture, recap key concepts in lecture notes (self-study)
- Work on exercises, come to exercise session, tackle the tricky exercises together!
- Code, code, and code. repeat...

```
try <- 0
while(try < 999) {
  try <- try + 1
}
cat("success!")
```

```
## success!
```

Our Team - At Your Service

Federica Mascolo

Andrea Burro

Aurélien Sallin

Introduction: Aurélien Sallin

- 2022-today: Expert in Health Care Research and Member of Management, SWICA Health Organization, Winterthur
- 2022-today: Lecturer, HSG
- 2018-2022: PhD Economic and Finance, HSG

Päpstliche Schweizergarde
Garde Suisse Pontificale
Guardia Svizzera Pontificia
Guardia Svizra Papala

UNI FR

SWICA

Introduction: Aurélien Sallin

Research at SWICA

- Use Real-World Data from claims to assess effectiveness of health technological tools
- Use (Causal) Machine Learning to evaluate the effect of health policies on doctors' prescription behaviors
- Develop financing models for mandatory health care in Switzerland

Other Research in Economics of Education (during my PhD Economic and Finance)

- Missclassification rates for gifted students
- Evaluation of Special Education programs

Organisation of the Course

Course concept: lectures

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

Course concept: exercises

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

Course concept: exercise sessions

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


The road ahead

Two special lectures

- 24.10.2024: R from a student's perspective
  - Minna Heim, BA from St. Gallen, student in Data Science at ETH Zurich

- 05.12.2024: Industry and Consulting Insights
  - Rachel Lund, PhD: Data Science Lead at Deloitte

Part I: Data (Science) fundamentals

Date	Topic
19.09.2024	Introduction: Big Data/Data Science, course overview
26.09.2024	Programming with R
26.09.2024	Exercises 1: Tools, programming
03.10.2024	An introduction to data and data processing
10.10.2024	Data storage and data structures
10.10.2024	Exercises/Workshop 2: Data storage and data structures
17.10.2024	Rectangular data
24.10.2024	Non-rectangular data. Guest spot: Minna Heim
24.10.2024	Exercises/Workshop 3: Web data, text, and images

Part II: Data gathering and preparation

Date	Topic
14.11.2024	Data preparation and manipulation
21.11.2024	Basic statistics and data analysis with R
21.11.2024	Exercises/Workshop 4: Data gathering, data import
28.11.2024	Visualisation
05.12.2024	Guest Lecture: Data Handling @Deloitte (Rachel Lund, Senior Economist)
05.12.2024	Exercises/Workshop 5: Data preparation and applied data analysis with R

Part III: Analysis, visualisation, output

Date	Topic
12.12.2024	Analytics, more visualisation, and data products
19.12.2024	Summary, Wrap-up, Final workshop
19.12.2024	Exercises/Workshop 6: Visualization, dynamic documents
19.12.2024	Exam for Exchange Students

Exam information

- Central, written examination: *digital, BYOD!*.
- 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, : *19.12.2024, 16:15-18:00, room tbd*.
  - Questions: *andrea.burro@unisg.ch*

The tools

Core course resources

- All information and materials (notes, slides, course sheet, syllabus, etc.) are available on StudyNet/Canvas.
- Use github to be always updated about the course material
  - Install git on your computer as explained here
  - Clone the course repository using

```
git clone git@github.com:ASallin/datahandling-lecture.git  # to clone

git pull origin main  # to update
```

Why R?

The data language

- Widely used in Data Science jobs.
- Originally designed as a tool for statistical analysis.
- Particularly useful to program with data.

High-level language

- Relatively easy to learn.
- A lot of free tutorials and support online.

Free, open-source, large community

- Used in various fields.
- Thousands of 'R-packages' covering diverse aspects of data analysis.
- Learn from open sources.

R

Install R from here!

RStudio

Install RStudio from here!

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.

Baumer, Kaplan and Norton (2023). *Modern Data Science with R*, 2nd Edition.

Further resources

- Stackoverflow
- Get inspired in the R blogsphere
- ChatGPT

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