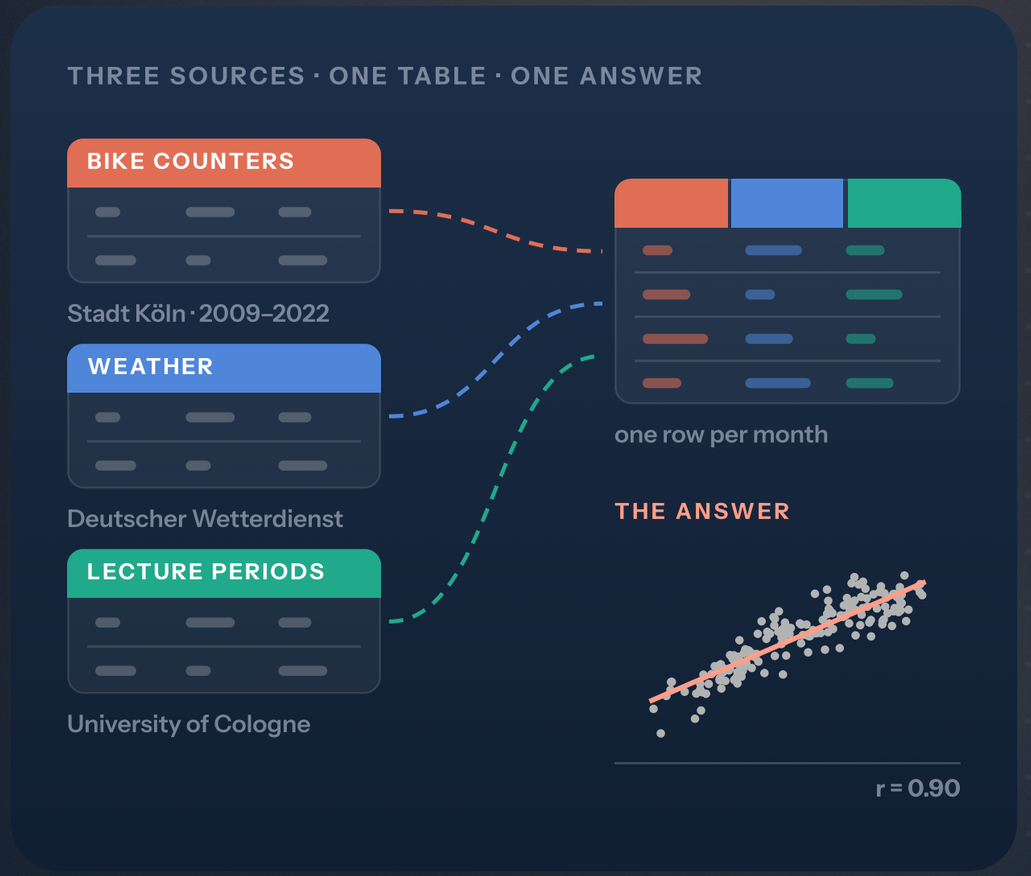
Many sources, one answer.

Kickoff: the whole course at a glance

Winter 2026/27 · Timo Breuer · TH Köln



Does Cologne cycle more when it is warm?

TWO OPEN DATASETS

SOURCE 1 · STADT KÖLN

dl-de/zero-2.0

Bike counters

Cyclists per month at every counting station — one CSV file per year, 2009 to 2022

\+ joined on year and month

SOURCE 2 · DWD

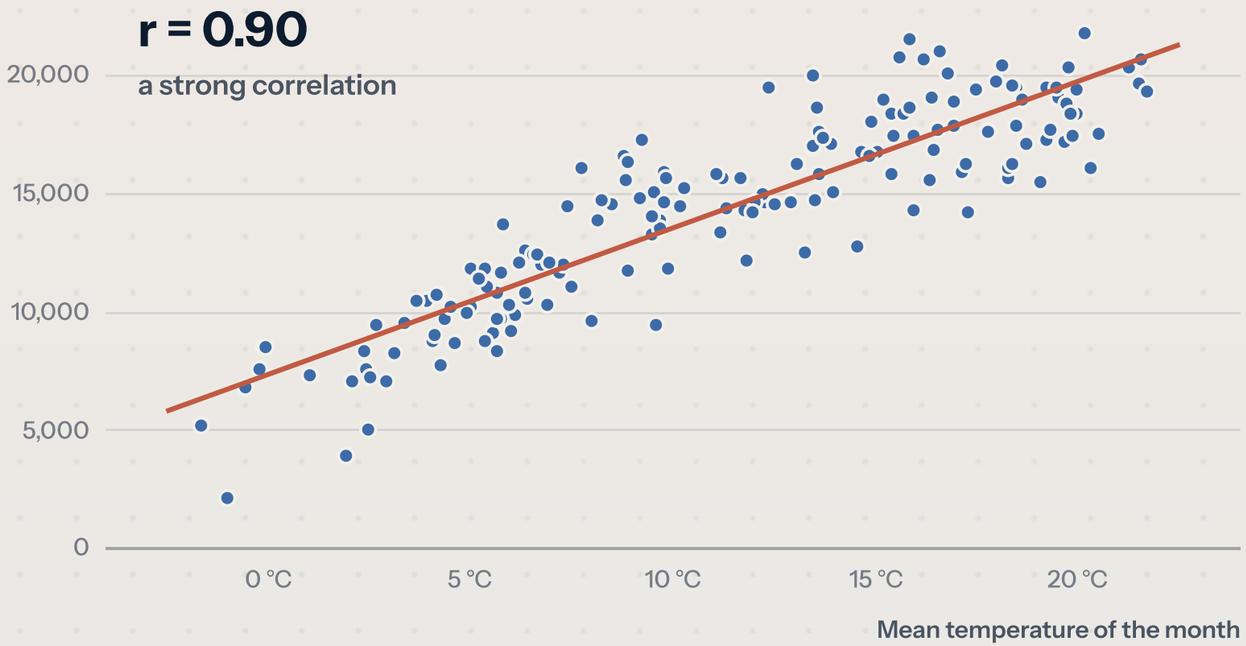
CC BY 4.0

Weather

Monthly temperature, rain and sunshine — Deutscher Wetterdienst, station Köln/Bonn

EACH DOT IS ONE MONTH · 168 MONTHS, 2009–2022

Cyclists per day · the four oldest counters



Only together do they answer it: about 620 more cyclists a day for every degree warmer.

Stadt Köln: Fahrrad-Verkehrsdaten 2009–2022, offenedaten-koeln.de (dl-de/zero-2.0) · DWD: monthly climate data, station Köln/Bonn (CC BY 4.0)

Before the answer: fourteen messy files

THE SAME SOURCE · THREE OF ITS FOURTEEN FILES

Radverkehr für Offene Daten Köln 2009.csv

```
;Deutzer Brücke;Hohenzollernbrücke;Neumarkt;Zülpicher Straße
Januar;46319;24610;59344;78727
…
Jahressumme;924253;582968;1222838;1615262
```

1 3

Fahrrad_Zaehlstellen_Koeln_2017.csv

```
Jahr 2016;Deutzer Brücke;Hohenzollernbrücke;Neumarkt;Zülpiche…
Jan 17;52.972;38.958;74.797;96.297;49.498;91.246;15.868;9.692…
```

4

Radverkehr für Offene Daten Köln 2022.csv

```
;Deutzer Brücke;Hohenzollernbrücke;Neumarkt;Zülpicher Straße;…
Januar;88.423;52.307;84.708;104.967;63.658;124.574;68.031;110…
```

5 2

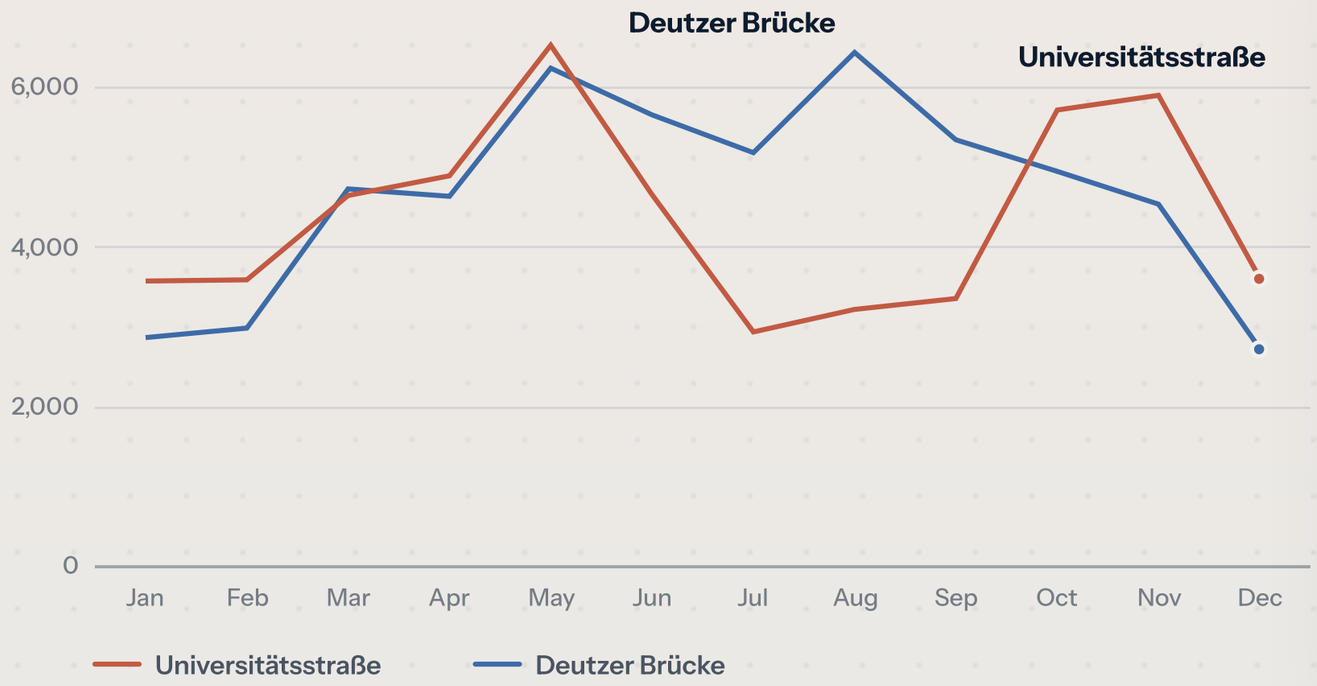
WHAT WE FOUND — AND FIXED

1. Two encodings: Latin-1 in 9 files, UTF-8 in 5. Read the wrong way, **Brücke** turns into **BrÃ¼cke**.
2. From 2016 on, dots separate thousands: **88.423** means 88,423 cyclists — not 88.423.
3. Until 2015, the last row is the annual total: add up a column and every cyclist counts twice.
4. Months are **Januar** in some years, **Jan 17** in others — and the file for 2017 says **Jahr 2016**.
5. 4 counters in 2009, 17 in 2022: new columns, a new order, empty cells.
6. The 2012 file repeats the 2011 numbers of Bonner Straße — month for month.

Files as published on offenedaten-koeln.de · pandas 3.0.6: `pd.read_csv(…, sep=";", encoding="latin-1")` reads 88.423 as a decimal number — `thousands="."` fixes it

Then the data surprises you

CYCLISTS PER DAY IN 2022 · TWO OF THE 17 COUNTERS



Stadt Köln: Radverkehr 2022 (dl-de/zero-2.0)

WHAT IS GOING ON HERE?

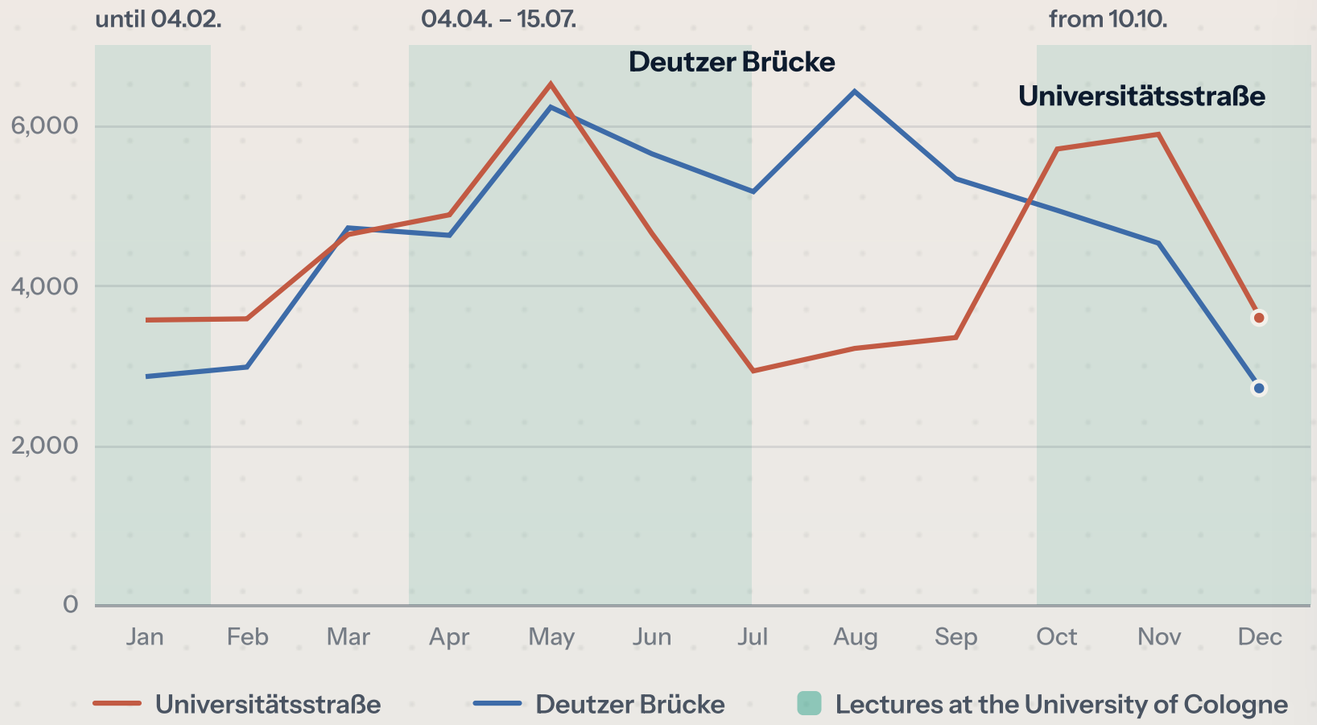
Universitätsstraße does not follow the weather.

- May → August: **−51%** here, **+3%** at Deutzer Brücke
- August → November: **+84%** here, **−30%** there
- The only one of 17 counters with a below-average August

What could explain it?

Then the data surprises you

CYCLISTS PER DAY IN 2022 · TWO OF THE 17 COUNTERS



A THIRD SOURCE EXPLAINS IT

Universitätsstraße follows the semester — not the weather.

- The street runs past the main campus: the university library is at no. 33.
- No lectures in August and September: half as many cyclists as in May.
- Lectures again from 10 October: busier than in any summer month.

Every answer raises a new question — and needs one more source.

Stadt Köln: Radverkehr 2022 (dl-de/zero-2.0) · lecture periods: University of Cologne, “Dates and deadlines”, winter term 21/22 to 22/23 (web.archive.org)

What we just did is this course

O
Obtain
Find and collect data

Open data · 23.10.

Web scraping · 27.11.

IN THE EXAMPLE

14 CSV files and a weather station

S
Scrub
Clean and structure it

Data cleaning and data structures · 04.12.

IN THE EXAMPLE

Two encodings, 88.423 cyclists

E
Explore
Combine and look for patterns

Data science packages · 11.12.

IN THE EXAMPLE

168 months in one scatter plot

M
Model
Test your hypotheses

Hypothesis testing · 08.01.

IN THE EXAMPLE

r = 0.90 — chance or not?

N
iNterpret
Explain and present

Your presentation · 25.01.

IN THE EXAMPLE

The semester, not the weather

TOOLS FOR EVERY STEP

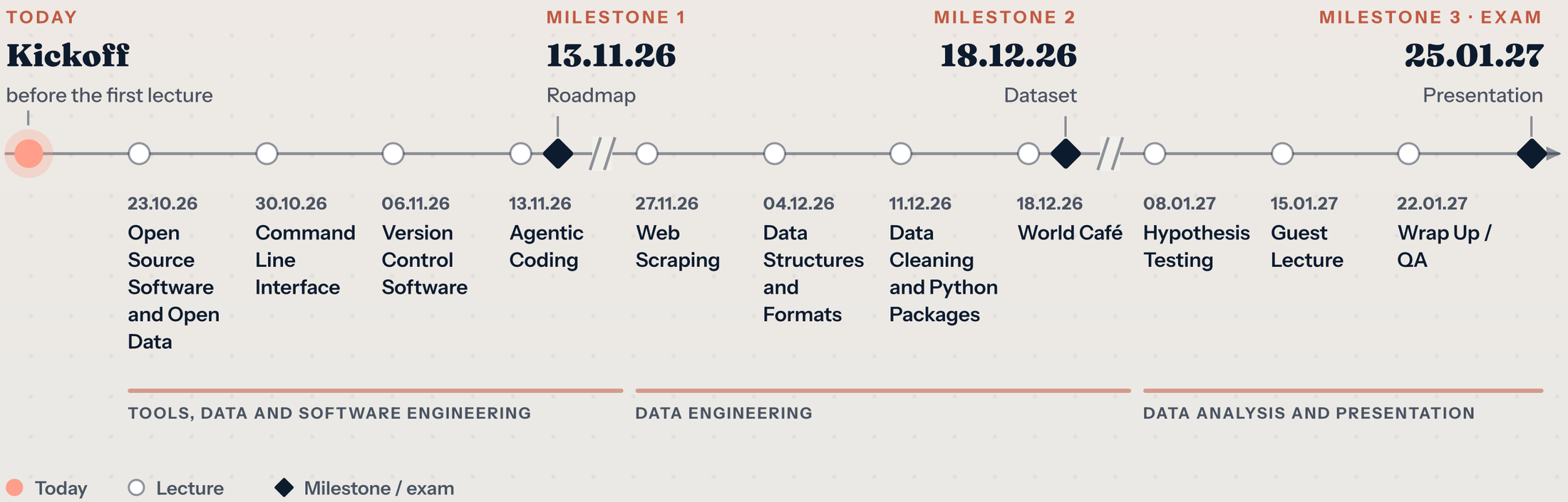
30.10. **Command line**

06.11. **Version control with Git**

13.11. **Agentic coding**

OSEMN after Hilary Mason and Chris Wiggins (2010), A Taxonomy of Data Science

Timeline


TODAY
Kickoff
before the first lecture
MILESTONE 1
13.11.26
Roadmap
MILESTONE 2
18.12.26
Dataset
MILESTONE 3 · EXAM
25.01.27
Presentation
23.10.26
Open Source Software and Open Data
30.10.26
Command Line Interface
06.11.26
Version Control Software
13.11.26
Agentic Coding
27.11.26
Web Scraping
04.12.26
Data Structures and Formats
11.12.26
Data Cleaning and Python Packages
18.12.26
World Café
08.01.27
Hypothesis Testing
15.01.27
Guest Lecture
22.01.27
Wrap Up / QA
TOOLS, DATA AND SOFTWARE ENGINEERING
DATA ENGINEERING
DATA ANALYSIS AND PRESENTATION
Today
Lecture
Milestone / exam

How the course works

1

Lectures

Nine lectures in three blocks, a guest lecture and a wrap-up with Q&A — on Fridays from 23.10.26 to 22.01.27.

2

Labs

One lab per lecture: practise right away what you just learned. In every lab you gain hands-on experience.

3

Group project

Your own data pipeline, in a team, on a question of your choice. Three milestones — the presentation is the exam.

NO PRIOR KNOWLEDGE NEEDED

We start from zero — with the command line (topic of the 2nd lecture on 30.10.26).

You learn every tool step by step and practise it in the labs.

LANGUAGE

Slides, labs and materials are in English. Your documentation and presentation may be in **English or German**.

Your question, your data

1
A question
you care about — sports, the environment, politics, health, culture

2
Open data
from portals such as GovData, Wikidata or Our World in Data

3
+ data you collect
from the web: scraped pages, APIs and downloads

4
Your answer
a pipeline, a published dataset and a presentation

TOPICS FROM PREVIOUS YEARS

Football
Market value vs. Champions League games · wins vs. weather · when do most goals fall?

Olympics
Medals vs. gross domestic product

Video games
Sales across platforms

Movies
Trends and patterns in IMDb data

Choose a topic you are curious about: you will spend a whole semester with it.

Three milestones, one per block

MILESTONE 1

13.11.26

Roadmap

Your topic, your datasets, your research questions — and a plan. Written in Markdown in your repository.

End of block 1 · tools

MILESTONE 2

18.12.26

Dataset

Your enriched dataset, hosted online, and a first version of your pipeline — presented in the World Café.

End of block 2 · data engineering

MILESTONE 3 · EXAM

25.01.27

Presentation

Your findings, a demo of your notebook, and your open repository with its documentation.

End of block 3 · analysis

Each milestone builds on the one before — start early, and commit often.

Grading in one slide

Your grade comes from your group project — code, documentation and presentation — and from the points it scores on a rubric of ten criteria.

below 50	50 points	55 points	60 points	65 points	70 points	75 points	80 points	85 points	90 points	95 points
n.b.	4.0	3.7	3.3	3.0	2.7	2.3	2.0	1.7	1.3	1.0

TEN CRITERIA — DETAILS IN THE FIRST LECTURE

- Dataset and motivation
- Documentation
- Repository organisation
- Version control and collaboration
- Web scraping
- Data cleaning
- Data analysis
- Hypothesis testing and modelling
- Visualisation and presentation
- Q&A

BONUS POINTS

+10 for passing the World Café on 18.12.26

The tools of the trade

30.10.

The command line

`bash` `Git Bash` `curl`

06.11.

Version control

`Git` `GitHub`

27.11.

Web scraping

`requests` `Beautiful Soup`

FROM 04.12.

Python for data

`Jupyter` `pandas` `NumPy` `matplotlib` `seaborn` `SciPy`

All free, almost all open source — yours to keep after the course.

AND AN AI PAIR PROGRAMMER

Agentic coding • 13.11.

Coding agents such as GitHub Copilot, Claude Code or Codex write and run code for you. You learn to steer them — and to check what they did.

IN THIS COURSE

Allowed — even encouraged. Document every use in `AI_USAGE.md`, and be able to explain every line.

Most need a paid plan — GitHub Copilot is free for verified students.

In January, you will be able to …

1. find open data — and collect more from the web
2. clean, structure and combine messy sources
3. explore data and test hypotheses in Python
4. work with the command line, Git and AI coding agents
5. present your findings — and defend them

AND YOU WILL HAVE

- **A public GitHub repository with your pipeline**
  Your code, its history and its README.
- **A dataset you built and published**
  Open data plus data you collected.
- **A presentation of your findings**
  Your answer to a question you chose.

Something to show when you apply for an internship, a student job or a thesis.

Before the first lecture on 23.10.26

1 **Create a GitHub account** at github.com — and verify your student status at education.github.com.

2 **Bring a laptop** to the labs: Windows, macOS or Linux. No laptop? Talk to me.

3 **Find teammates** — the project is teamwork from the start.

4 **Collect questions** you are curious about — and look for open data on them.

5 **Watch the course repository** on GitHub: **github.com/breuert/data-integration**

AND THAT IS ALL

Nothing to install yet — we set up the tools together, step by step.

Communication and support

1

The Moodle course is
DIS3.3 Data Integration [WS 2026/2027]
The passphrase is shared during the lecture.

`https://elearning.iws.th-koeln.de/moodle/course/view.php?id=3051`

2

We will use Moodle for
communication and announcements.

All resources are shared on the website
breuert.github.io/data-integration/

3

Maximilian L. Jäger
will assist you as a tutor during the lecture period.

The contact information is shared on Moodle.

You can also reach Timo Breuer via e-mail:

timo.breuer@th-koeln.de

Questions?

See you at the first lecture: **Open Source Software and Open Data** on 23.10.26.

timo.breuer@th-koeln.de

breuert.github.io/data-integration/

TODAY'S SOURCES

- offenedaten-koeln.de — Fahrrad-Verkehrsdaten Köln, one dataset per year, 2009–2022 (dl-de/zero-2.0)
- opendata.dwd.de — monthly climate data of the DWD, station Köln/Bonn (CC BY 4.0)
- Lecture periods of the University of Cologne, archived on web.archive.org: winter term 21/22, summer term 22, winter term 22/23
- Mason & Wiggins (2010): *A Taxonomy of Data Science* — the OSEMN steps