UNIVERSITY OF LUCERNE

The ABC of Computational Text Analysis

#7 Working with (your own) Data

Alex Flückiger
Faculty of Humanities and Social Sciences
University of Lucerne

11 April 2024

Recap last lecture

- introducing Python 🐍
- learning programming concepts & syntax

  data types, loops, indexing, functions…
- working with VS Code Editor

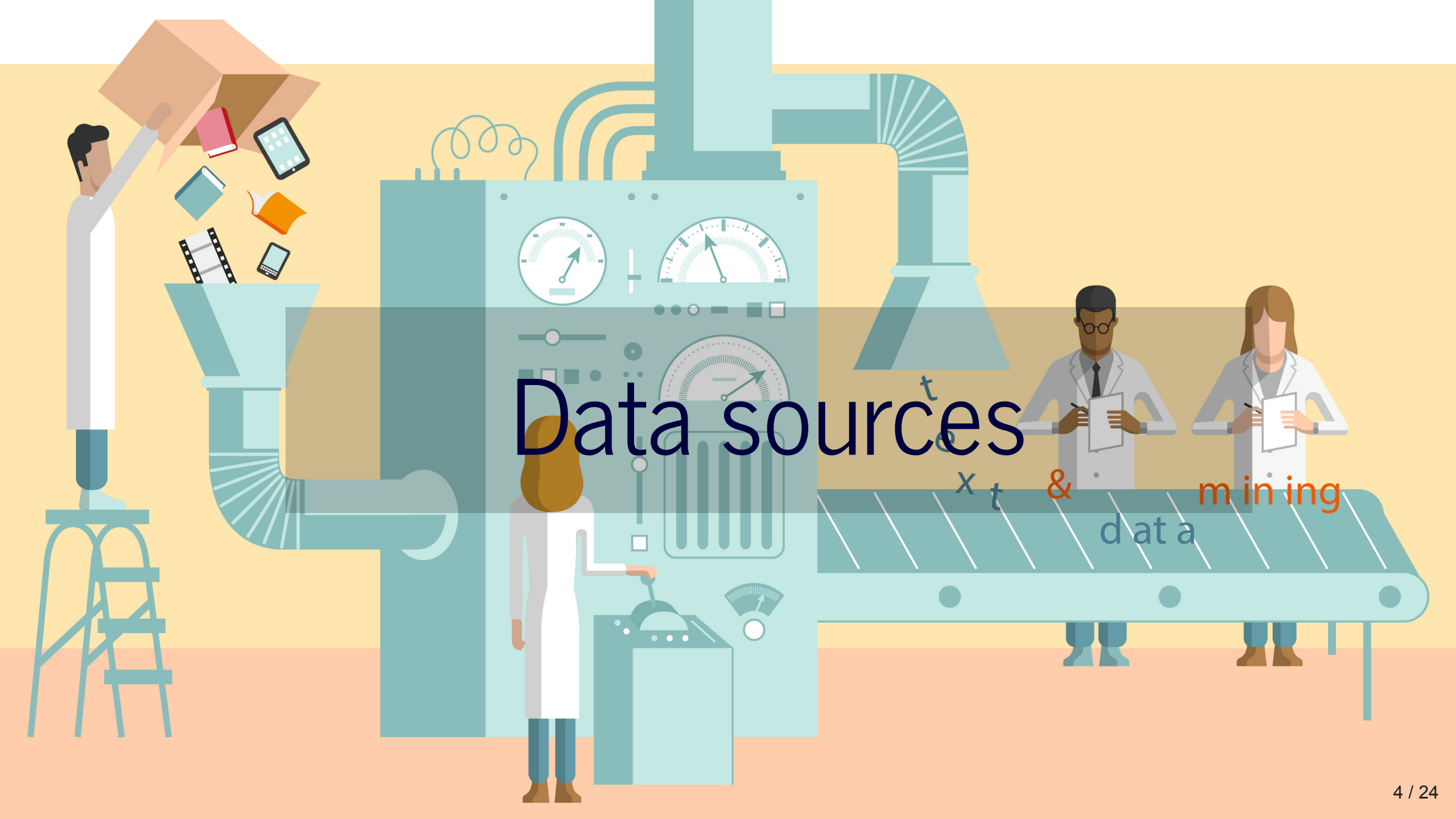
Outline

- learn about available data sources
- analyze and justify your own data

  any text ✅

  from anywhere ✅

  `.txt` or `.pdf` format ✅


Data sources
t
e
x t
&
m in ing
d at a

What data sources are there?

- broadly social
  - newspapers + magazines
  - websites + social media
  - reports by NGOs/GOs
- scientific articles
- economic
  - business plans/reports
  - contracts
  - patents

👉 basically, any textual documents…

How does the data look like?

Any text is data, yet some formats are more suitable.

1. datasets like `.csv` 🥰
2. plain text like `.txt` 🙂
3. text in other formats like `.pdf` 😬

Some great (historical) datasets

`.csv` ready off-the-shelf

- 1 August speeches by Swiss Federal Councilors

  provided via course repo
- Human Rights Reports by various NGOs
- United Nations General Debate Corpus
- Inaugural Speeches by US Presidents

😓 There are still not many.

Dedicated search engines for datasets

Use case: Search for existing datasets

- Harvard Dataverse

  open scientific data repository

- Google Dataset Search

  Google for datasets basically

👉 search for a topic followed by `corpus`, `text collection` or `text as data`

Search techniques 🔍

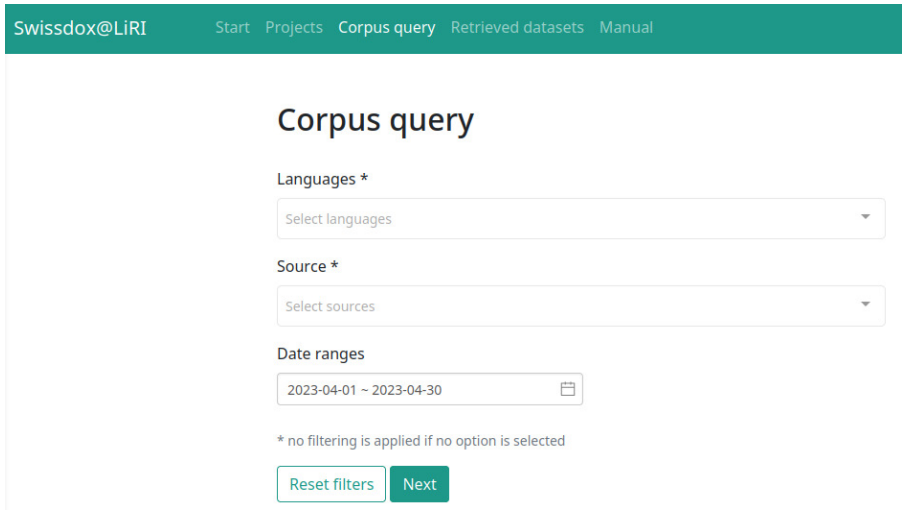
Make your (Google) web search more efficient by using dedicated tags.
Examples:

- `"computational social science"`
- `site:nytimes.com`
- `nature OR environment`

Swissdox: A game changer

Assemble a news dataset and export as `.csv`

- over 250 Swiss newspapers
- historical and updated daily
- needs registration (free)



More publishers

- Project Gutenberg and HathiTrust
  - massive collection of books
  - open, HathiTrust requires agreement
- Nexis Uni
  - international newspaper, business + legal reports
  - licensed by the university
- Constellate
  - scientific articles of JSTOR across disciplines
  - provides an easy dataset builder

👉 check out other resources licensed by ZHB

Interesting sources as PDFs

Any organization of your interest 👍

- Party Programmes across Europe

  covers over 1000 parties from 1920 until today in over 50 countries

- Swiss voting booklets

  from 1978 until today

- Swissvotes

  collection of resources on Swiss public votings

- 1 August speeches by Swiss Federal Councillors
- Nestlé Annual Reports
- University of Zurich Annual Reports

Scraping PDFs from websites

Use case: Swiss voting booklets

- `wget` to download any files from the internet

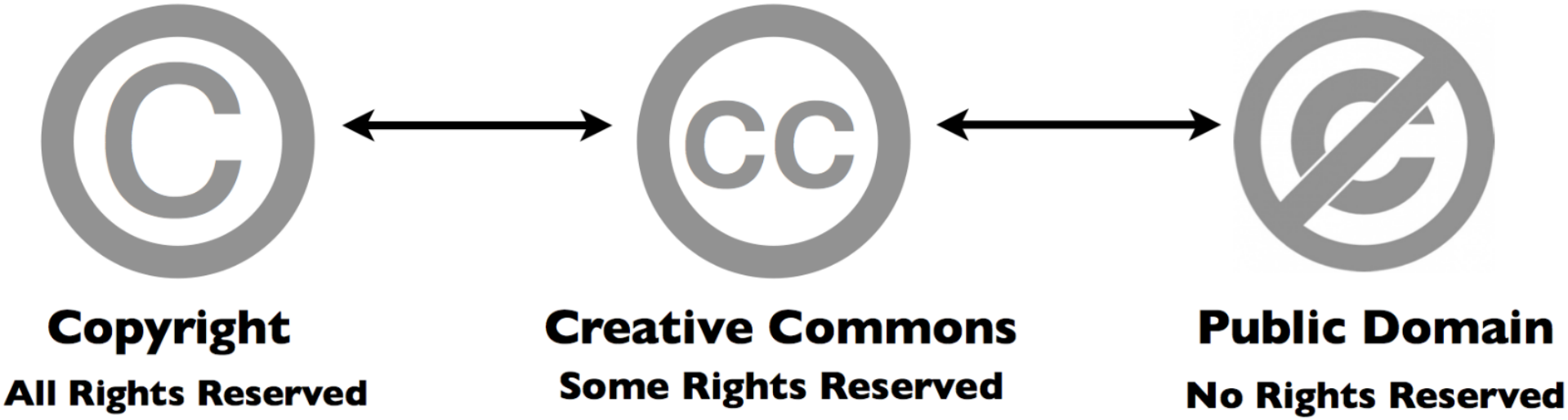
```
# get a single file
wget EXACT_URL

# get all linked pdf from a single webpage
wget --recursive --accept pdf -nH --cut-dirs=5 \
--ignore-case --wait 1 --level 1 --directory-prefix=data \
https://www.bk.admin.ch/bk/de/home/dokumentation/abstimmungsbuechlein.ht

# --accept FORMAT_OF_YOUR_INTEREST
# --directory-prefix YOUR_OUTPUT_DIRECTORY
```

Data is no free lunch 😢

Data is property 🚫

... and has rights too

- copyrights may further limit access to high quality data
- check the rights before processing data



Imperfect data: A tail of bias

- noise in text

  non-content (e.g. table of content), inconsistent spelling

- archive holes

  lost or uncollected data

- selective corpus curation

  supposition that key-word(s) captures topic

- social bias

  view from somewhere, stereotypes

👉 think about the data and mitigate issues

A world for humans ... ... and a jungle of file formats.

Common conversions

Your text is of a particular type

⬇️

digital native documents

`.pdf`, `.docx`, `.html`

⬇️

convert to `.txt`

⬇️

scans of (old) documents

`.pdf`, `.jpg`, `.png`

⬇️

Optical Character Recognition (OCR)

machine-readable ✅

What data are you interested in?

Think about your mini-project

Illustration of text analysis generated by Image Creator from Microsoft Copilot

Assignment #2 ✍️

- get/submit via OLAT
  - starting tonight
  - deadline: 19 April 2024, 23:59
- discuss issues on OLAT forum

Let’s start coding ✨


Questions?

In-class: Exercises I

1. Make sure that your local copy of the Github repository KED2024 is up-to-date with `git pull`. Check out the data samples in `ked2024/materials/data` and the scripts to extract their text in `ked2024/materials/code`.
2. Decide on one use-case that interests you most.
3. **Apply the commands to reproduce on the given data. Test them on your own data. Check the resources. Ask questions. Think about your mini-project.**

In-class: Exercises II

1. Use `wget` to download *cogito* and its predecessor *uniluAKTUELL* issues (PDF files) from the UniLu website. Start with downloading one issue first and then try to automatize the process to download all the listed issued using arguments for the `wget` command.
2. Convert the *cogito* and *uniluAKTUELL* PDF files into TXT files. You can use the code given for native, digital PDFs (no OCR needed). Try with a single issue first and then write a loop to batch process all of them.
3. What is the University of Lucerne talking about in its issues? Use the commands of the previous lectures to count the vocabulary.
4. Do the same as in 3.), yet analyze the vocabulary of *cogito* and *uniluAKTUELL* issues separately. Does the language and topics differ between the two magazines?