The ABC of computational Text Analysis

08: Create your own Data Sets + Ethics

Alex Flückiger
23 April 2020

Recap last Lecture

- cleaning with regular expression
- finding data sources

Outline

- feedback assignment #2
- use your texts as data ✅

  anything
  from anytime
  from anywhere

- care about ethics 🙈🙉🙊

Assignment #2

Feedback Assignment #2

example solution

- make patterns more general

  date: `DD* Month DDDD`

- keep it simple

  name of month ~ any word ~ `\w+`

- avoid false positives with positional information

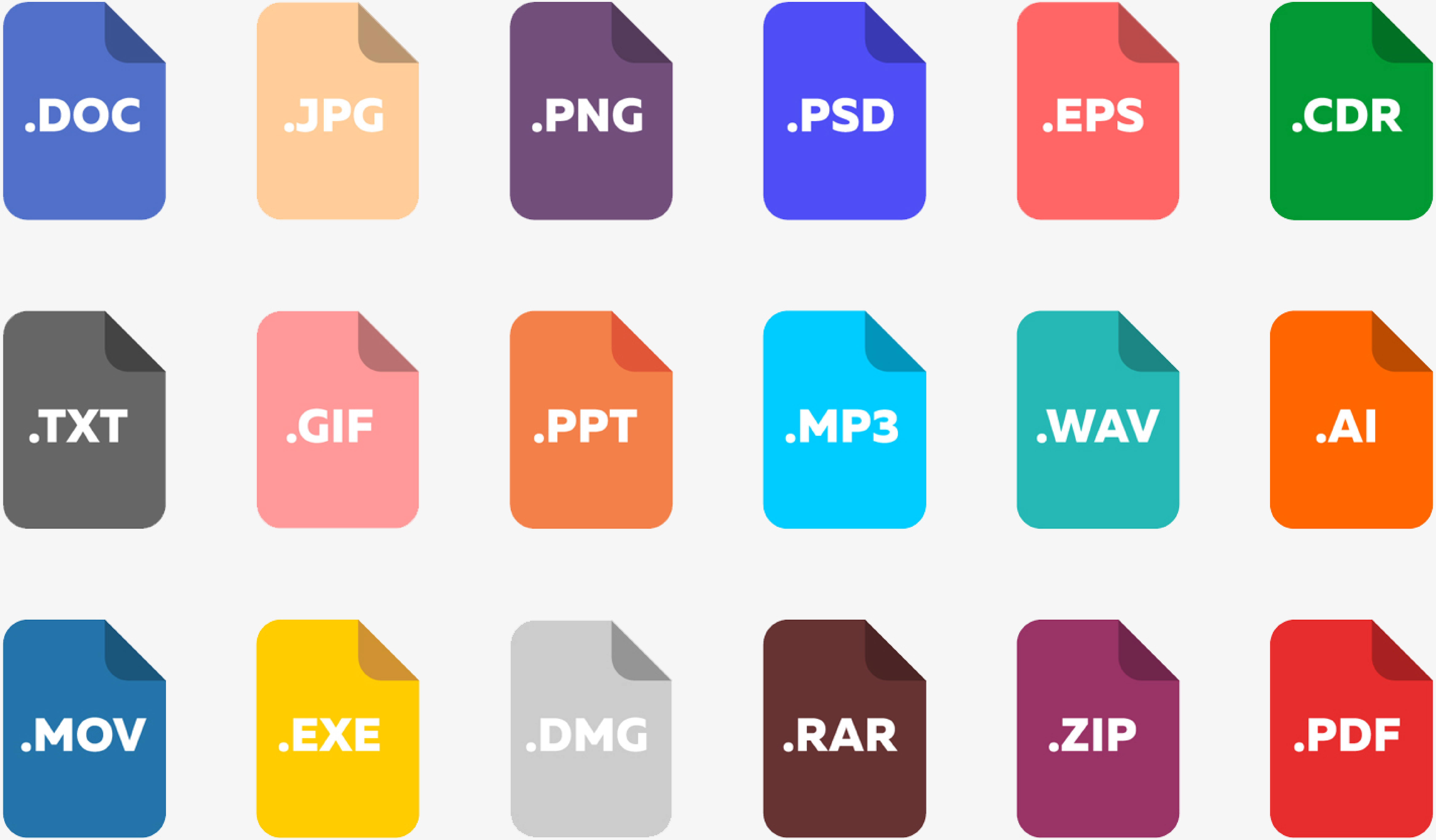
  start of line: `^`

- names are hard to extract

  variation + inconsistency

🤓 check the count of matches

Converting Documents


.DOC
.JPG
.PNG
.PSD
.EPS
.CDR
.TXT
.GIF
.PPT
.MP3
.WAV
.AI
.MOV
.EXE
.DMG
.RAR
.ZIP
.PDF

A world for humans …

news, press releases, reports from organizations

digital documents

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

⬇️

convert to `.txt`

scans of (old) documents

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

⬇️

Optical Character Recognition (OCR)

✅

machine-readable

Conversion of DOCX

use case: news articles from Nexis

- `pandoc` to convert file formats
- download as single articles in `.docx` on Nexis

```
# convert docx to txt
pandoc file_in.docx -t plain -o file_out.txt

### Install first with
brew install pandoc     # macOS
sudo apt install pandoc # Ubuntu
```

Conversion of digital PDF

use case: Swiss party programmes

```
# convert digital native pdf to txt
pdftotext -nopgbrk -eol unix file_in.pdf

### Install first with
brew install poppler                # macOS
sudo apt install poppler-utils  # Ubuntu
```

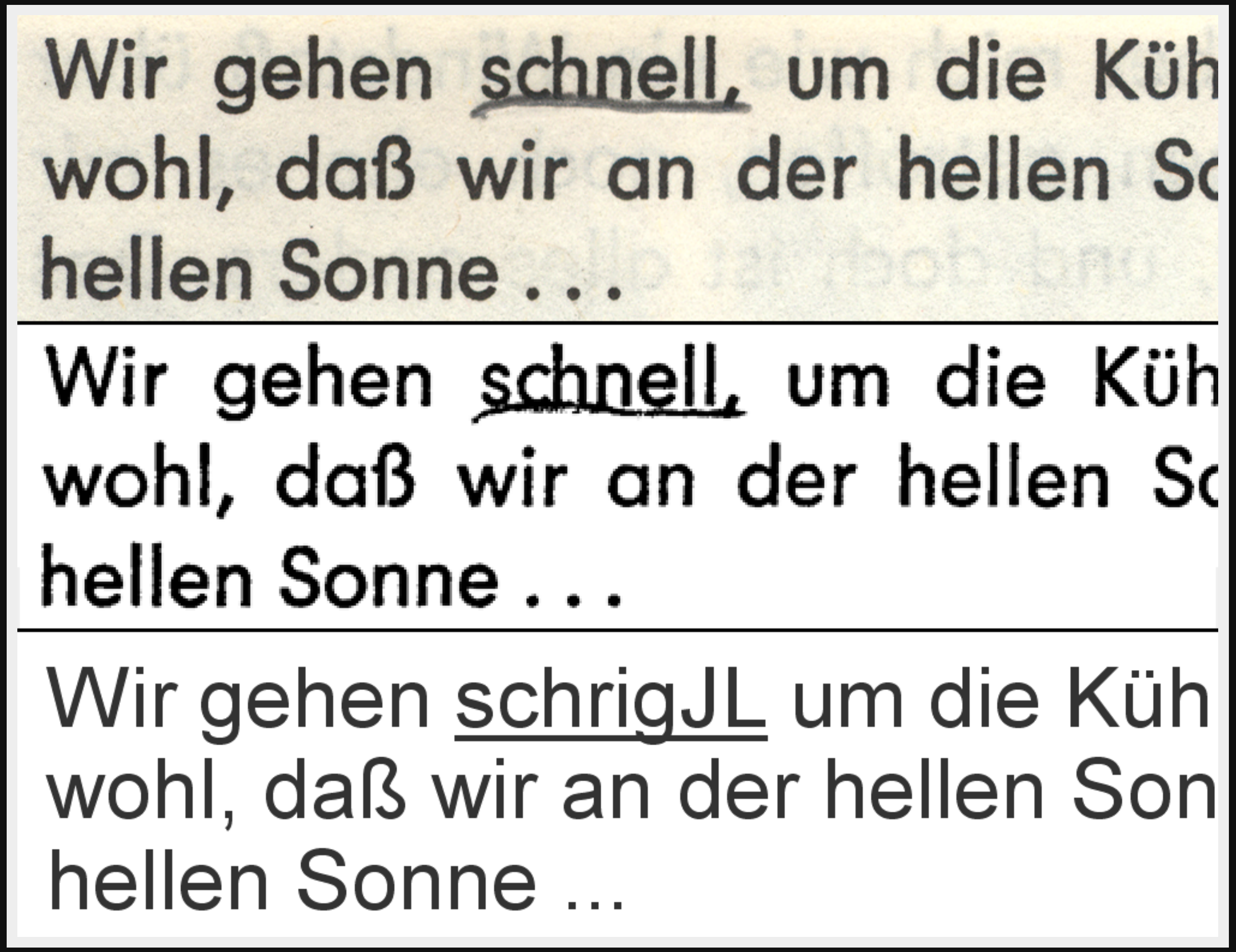
Optical Character Recognition (OCR)

- OCR ~ convert ../images into text

    text from scans/../images
    handwriting + Fraktur texts

- image quality is crucial
- open-source software: `tesseract`

    language-specific models



Wir gehen schnell, um die Küh
wohl, daß wir an der hellen Sc
hellen Sonne . . .

Wir gehen schnell, um die Küh
wohl, daß wir an der hellen Sc
hellen Sonne . . .

Wir gehen schrigJL um die Küh
wohl, daß wir an der hellen Son
hellen Sonne ...

example OCR (Wikipedia)

Conversion of digitalized PDF

use-case: historical party programmes

1. extract image from PDF + improve contrast
2. run optical character recognition (OCR) on the image

```
# convert scanned pdf to tiff, control quality with parameters
convert -density 300 -depth 8 -strip -background white -alpha off
file_in.pdf temp.tiff

echo test \
t

# run OCR for German ("eng" for English, "fra" for French)
tesseract -l deu temp.tiff file_out


### Install first with
brew install imagemagick              # macOS
sudo apt install imagemagick-6.q16    # Ubuntu
```

#LifeHack: Make a PDF searchable

use case: scanned book chapters

```
# output searchable pdf instead of txt
convert -density 300 -depth 8 -strip -background white -alpha off
file_in.pdf temp.tiff
tesseract -l deu temp.tiff file_out pdf
```

Scraping PDF from Websites

use case: Swiss voting booklet

- `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/abstimmungsbuech

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

Example Sources

- Party Programmes across Europe
- Swiss voting booklets
- 1 August speeches by Swiss Federal Councillors
- Nestlé Annual Reports
- … any organization of your interest 👍

Foundation of Batch Processing

perform the same operation on many files

```bash
# loop over all txt files
for file in *.txt; do

    # indent all commands in loop with a tab

    # rename each file
    # e.g. a.txt -> new_a.txt
    mv $file new_$file

done
```

Perform Batch OCR from PDF

```
for FILEPATH in *.pdf; do
    # convert pdf to image
    convert -density 300 $FILEPATH -depth 8 -strip \
    -background white -alpha off temp.tiff

    # define output name (remove .pdf from input)
    OUTFILE=${FILEPATH%.pdf}

    # perform OCR on the tiff image
    tesseract -l deu temp.tiff $OUTFILE

    # remove the intermediate tiff image
    rm temp.tiff

done
```

Preprocessing → RegEx

Bias & Ethics

Don’t be a fool …

… be wise, think twice.

Data = Digital Traces

- collecting, curating, preserving traces → uncover patterns
- data don’t disclose anything, you can speak with it though

Imperfect Data: A Tail of Bias

- data/archive holes

  lost, uncollected

- noise in data

  OCR errors, inconsistent spelling, non-content

- corpus curation

  supposition that key-word indicates topic

- social context

Raw data is an oxymoron.

(Gitelman 2013)

Data vs. Capta

Differences in the etymological roots of the terms data and capta make the distinction between constructivist and realist approaches clear. *Capta* is **"taken"** actively while *data* is assumed to be a **"given"** able to be recorded and observed. From this distinction, a world of differences arises. Humanistic inquiry acknowledges the situated, partial, and constitutive character of knowledge production, the recognition that knowledge is constructed, *taken*, **not simply given as a natural representation** of pre-existing fact.

(Drucker 2011)

Key Principles

- Who has a voice in your data?

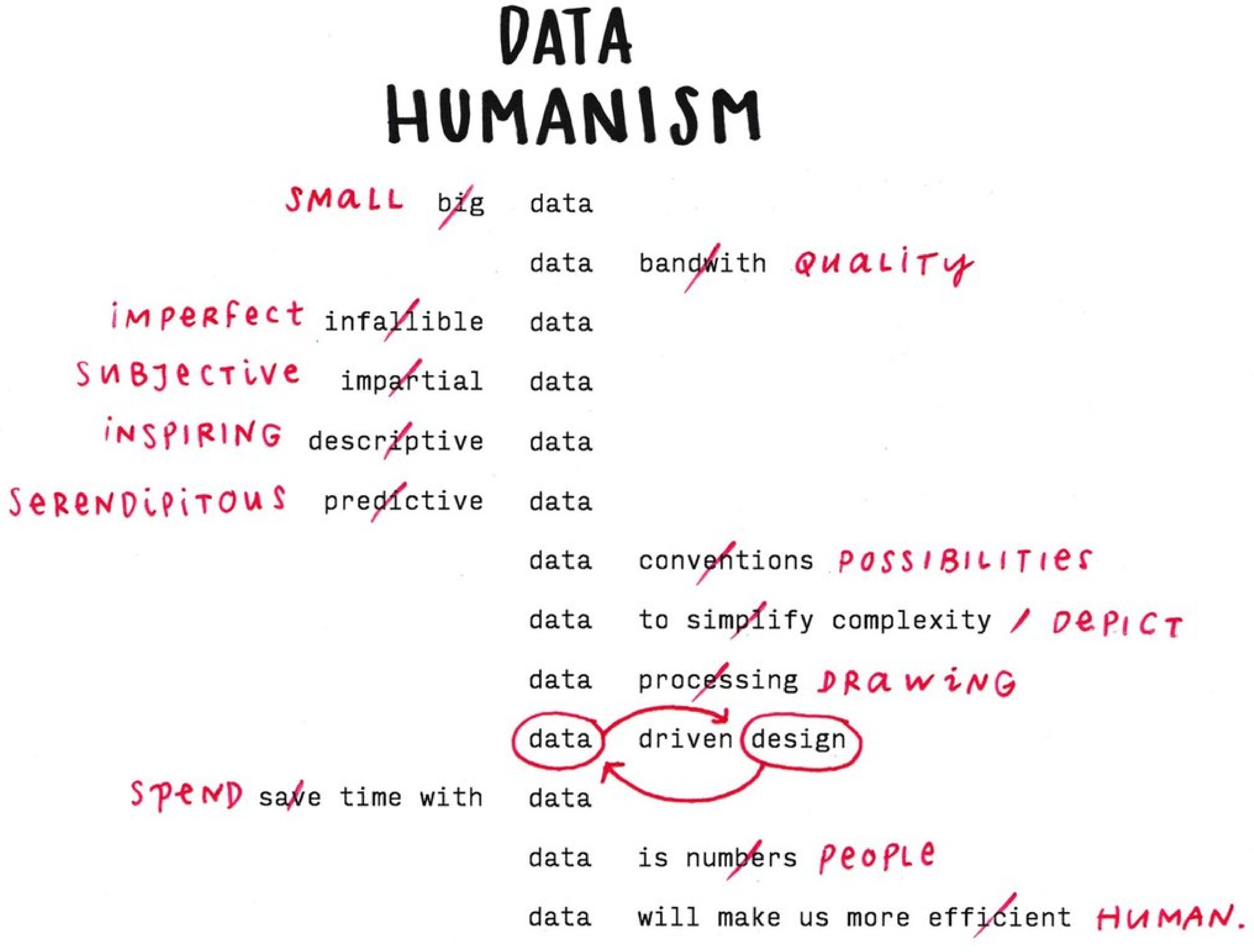
  social context

- bigger is not necessarily better

  more vs. more diverse data

- clean your data thoroughly

  noisy vs. clean data



DATA HUMANISM

SMALL ~~big~~ data

data ~~bandwith~~ QUALITY

imperfect ~~infallible~~ data

SUBJECTIVE ~~impartial~~ data

INSPIRING ~~descriptive~~ data

SERENDIPITOUS ~~predictive~~ data

data ~~conventions~~ POSSIBILITIES

data to ~~simplify~~ complexity / DEPICT

data ~~processing~~ DRAWING

data driven design

SPEND ~~save~~ time with data

data is ~~numbers~~ PEOPLE

data will make us more ~~efficient~~ HUMAN.

@giorgialupi

Data represents real life.

In-class: Exercises I

1. Make sure that your local copy of the Github repository KED2020 is up-to-date with `git pull`. Check out the data samples and scripts in `materials/`.
2. Install the missing tools with the commands given on the respective slides: `pandoc, imagemagick, poppler`
3. **Digest the commands. Test them. Check the resources. Ask questions. Think about your mini-project.**
4. Download one or all *cogito* issues (PDF files) from the UniLu website.
5. `wget` is a powerful tool. Have a look at its arguments and search for more examples in tutorials.

Resources

Make a more sophisticated script for PDF conversion

- Erick Peirson. 2015. Tutorial: Text Extraction and OCR with Tesseract and ImageMagick - Methods in Digital and Computational Humanities - DigInG Confluence. online

References

Drucker, Johanna. 2011. "Humanities Approaches to Graphical Display." *Digital Humanities Quarterly* 5 (1). http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

Gitelman, Lisa. 2013. *Raw Data Is an Oxymoron*. Cambridge: MIT.