The ABC of Computational Text Analysis

Supplements

Alex Flückiger
Faculty of Humanities and Social Sciences
University of Lucerne

14 April 2023

Purpose

Here I present some stuff that we did not cover in class.

Tasks

- find various ngrams with wildcards
- check gender specific language

  what follows `she/he` or `her/his`

Topics

Techniques

- text processing
- extracting and aggregating information
- creating simple visualizations
- optical character recognition (OCR)
- scraping files

Data

- using existing datasets
- creating new datasets

🤓 *inputs are more than welcome!*

Data vs. C apta

> «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.»

> «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.»

Forms of Data

- content data

  clean, plain text data

  preferable as `.txt`

- metadata ~ information about the actual data

  publishing date, authors, source, version

  preferable as `.csv`

show with default application (GUI)

```
open text.txt                 # macOS
wslview text.txt              # WSL Ubuntu (Windows)
```

Key Word in Context (KWIC)

```
ptx -f -w 50 */*.txt > ptx.txt
egrep -i "[a-z]  word" ptx.txt
```

Select Column in Dataset

```
cut -d\t -f1    # extract the 2nd column from a tab-separated file
```

Extract texts from tsv:

- http://www.theunixschool.com/2012/05/shell-read-text-or-csv-file-and-extract.html

Variables

```bash
echo "Starting program at $(date)"
```

Better Tokenization

- tokenization ~ splitting into words

```
# new, improved approach
cat text.txt | tr -sc "[a-zäöüA-ZÄÖÜ0-9-]" "\n"

# old approach
cat text.txt | tr ' ' '\n'
```

Batch Processing

```
for file in *.txt; do           # loop over all text files
 cat "$file" | pipe commands > "proc_$file"
done
```

Batch Renaming

```
rename  " " "_" *.txt   # replace spaces with underscores
# since there are different versions, if this doesn't work try:
# rename 's/ /_/' *.txt
```

```
i=1
for file in *.txt; do           # loop over all text files
 mv -- "$file" "text_$i.txt"    # rename each file with a sequential
 i=$((i+1))
done
```

Imperfect Data: A Tail of Bias

- social bias

  view from somewhere, stereotypes
- data/archive holes

  lost, uncollected
- corpus curation

  supposition that key-word indicates topic
- noise in data

  OCR errors, inconsistent spelling, non-content

👉 think about the data and mitigate issues

Outlook: NLP is on Fire 🔥

- supervised machine learning
- you can do basically anything with modern NLP

  train on human-annotated data

- effort, insights and quality may differ

  for better or worse

Mind your Data

- Who has a voice in your data?

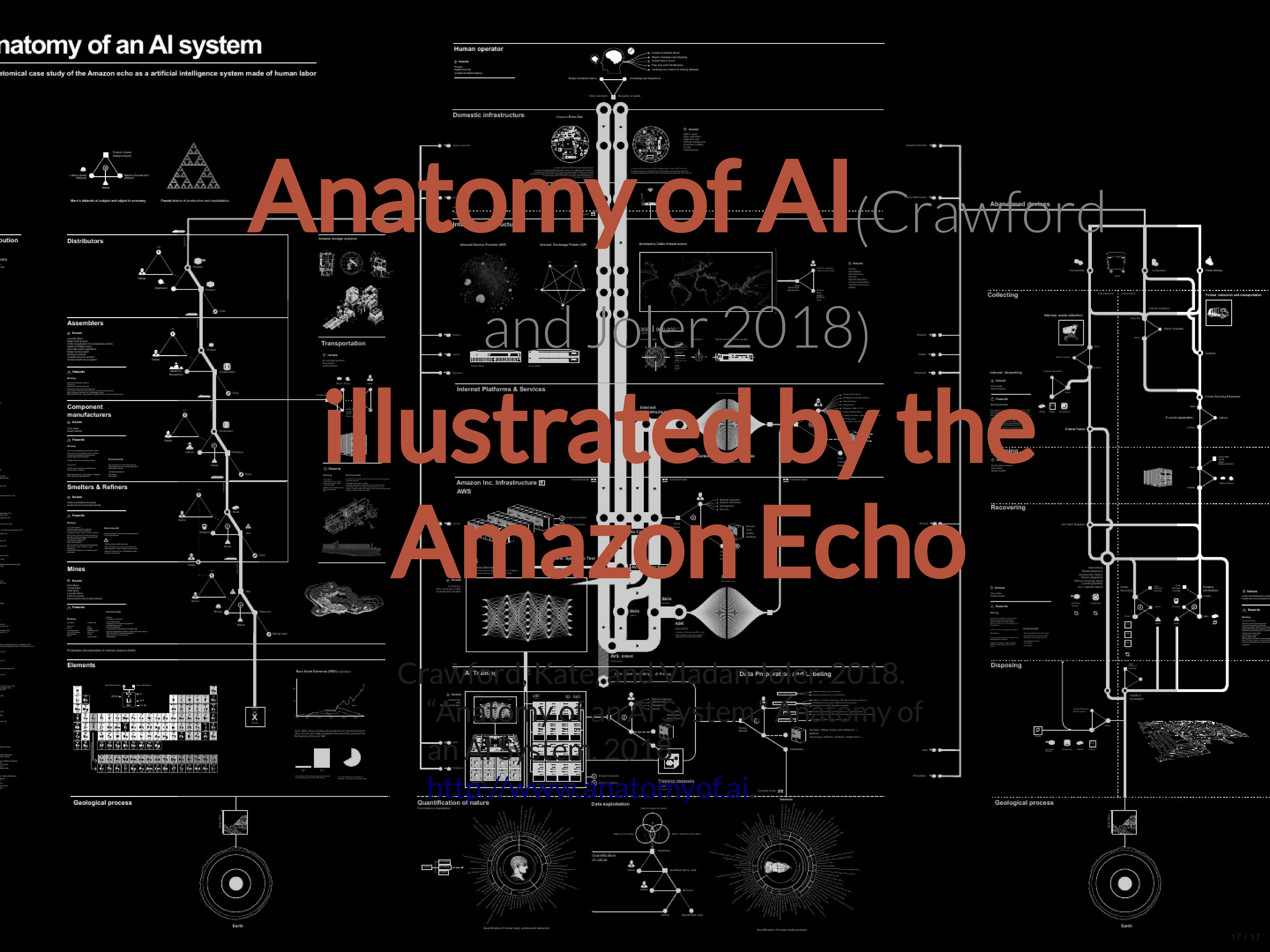
  social context

- bigger is not necessarily better

  more vs. more diverse data

- clean your data thoroughly

  noisy vs. clean data



Anatomy of AI (Crawford and Joler 2018) illustrated by the Amazon Echo

Crawford, Kate, and Vladan Joler. 2018. “Anatomy of an AI System.” Anatomy of an AI System. 2018. http://www.anatomyof.ai.