HPI Hasso Plattner Institut
Digital Engineering • Universität Potsdam

Internationales Zentrum für Ethik in den Wissenschaften (IZEW)

Eberhard Karls Universität Tübingen

# Responsible Data Science

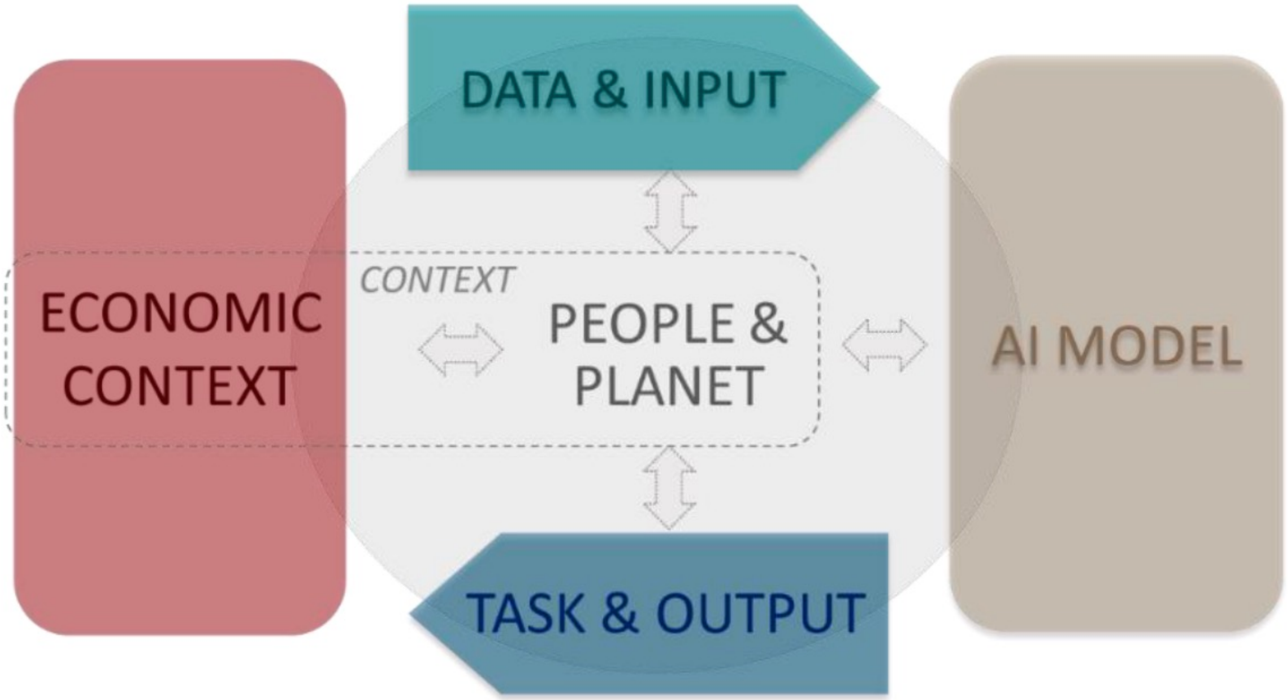
## Session 6: 25.05.2023, 15.15 – 19.30 h
## MA Seminar, SoSe 2023, Hasso-Plattner Institut

Dr. Simon David Hirsbrunner

# Today

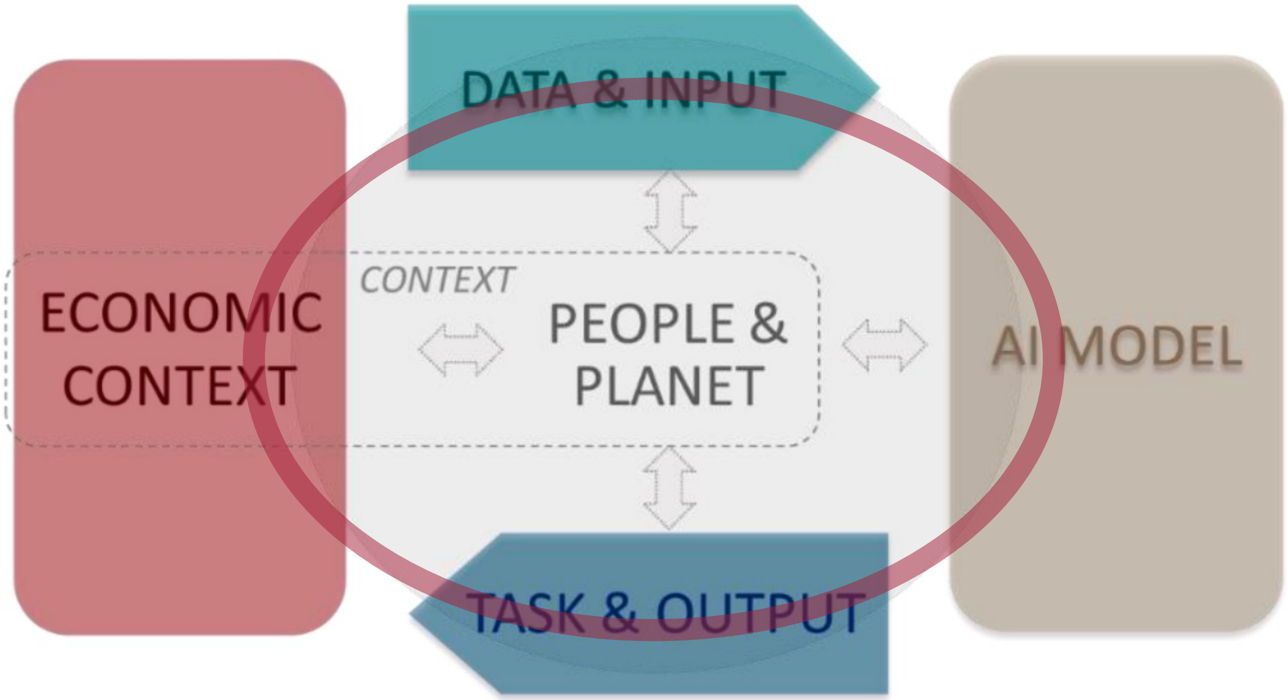
| topic | time |
|---|---|
| Introduction | 15h15 |
| Student presentation and discussion: Anatomy of AI (Pia Rissom) | 15h30 |
| Discussion: Sustainable AI | 16h30 |
| —— Break —— | 17h00 |
| Input: risk assessments / reflexivity | 17h30 |
| Exercise: simulation | 17h45 |
| —— Break —— | 18h30 |
| Seminar papers: brainstorming session in small groups | 18h45 |
| End | 19h30 |

# People and planet



Source: key high-level dimensions of the OECD Framework for the classification of AI Systems 2022

# People and planet



Source: key high-level dimensions of the OECD Framework for the classification of AI Systems 2022

# Presentation and discussion

Crawford, K., & Joler, V. (2019). Anatomy of an AI System. *Virtual Creativity*, *9*(1), 117–120.

(presentation by Pia Rissom)

# AI as wicked problem?

As described by Rittel and Webber (1974), wicked problems have 10 important characteristics:

1) They do not have a definitive formulation.

2)  They do not have a “stopping rule.” In other words, these problems lack an inherent logic that signals when they are solved.

3)  Their solutions are not true or false, only good or bad.

4)  There is no way to test the solution to a wicked problem.

5)  They cannot be studied through trial and error. Their solutions are irreversible so, as Rittel and Webber put it, “every trial counts.”

6) There is no end to the number of solutions or approaches to a wicked problem.

7) All wicked problems are essentially unique.

8) Wicked problems can always be described as the symptom of other problems.

9) The way a wicked problem is described determines its possible solutions.

10) Planners, that is those who present solutions to these problems, have no right to be wrong. Unlike mathematicians, “planners are liable for the consequences of the solutions they generate; the effects can matter a great deal to the people who are touched by those actions.”

# Presentation and discussion

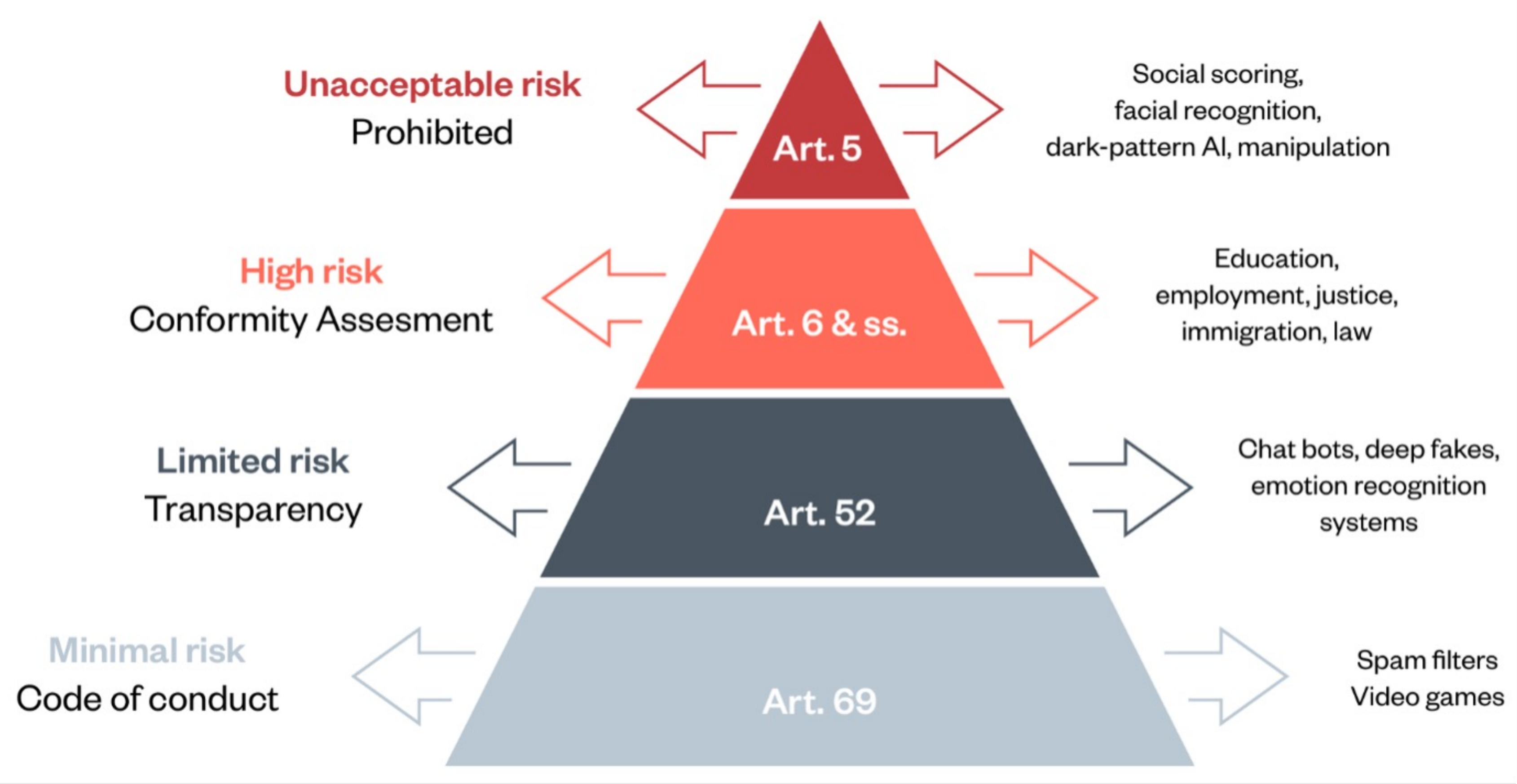
van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, *1*(3), 213–218. https://doi.org/10.1007/s43681-021-00043-6

# AI risks for stakeholders

**Definition of risk**
„general probability of negative consequences to actions“
(see Cambridge Dictionary, 2022 in Lütge et al. 2022).

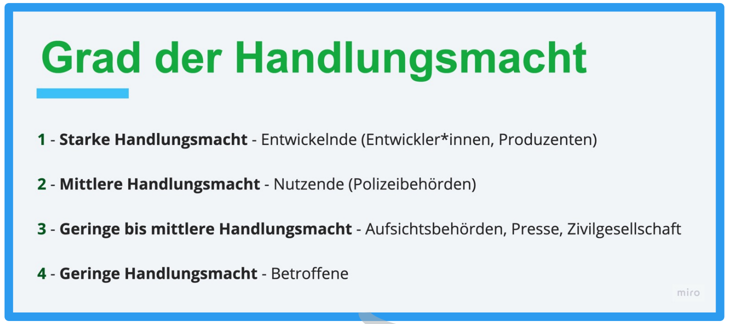
**Algorithm-related risks for stakeholders**
❑ Access to goods, benefits, or services ❑ Financial ❑ Property/material resources ❑ Reputation ❑ Emotional ❑ Life/security ❑ Privacy ❑ Liberty ❑ Rights/intellectual property (see City and County of San Francisco's Ethics and Algorithms Toolkit).

# Risk-based approach of the (proposed) EU AI regulation



Source: https://www.adalovelaceinstitute.org/resource/eu-ai-act-explainer/

# Tool
## Stakeholder and risk mapping



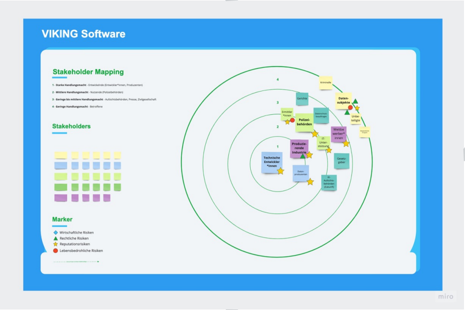
### Grad der Handlungsmacht

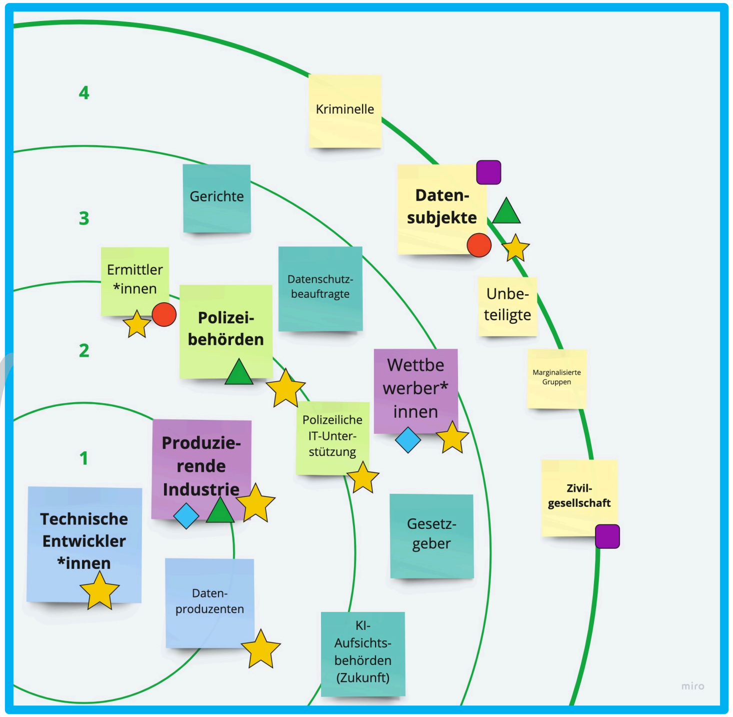
1 - **Starke Handlungsmacht** - Entwickelnde (Entwickler*innen, Produzenten)

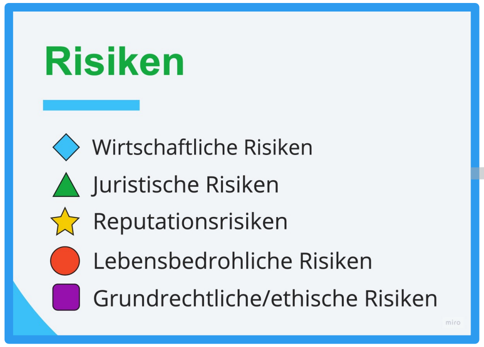
2 - **Mittlere Handlungsmacht** - Nutzende (Polizeibehörden)

3 - **Geringe bis mittlere Handlungsmacht** - Aufsichtsbehörden, Presse, Zivilgesellschaft

4 - **Geringe Handlungsmacht** - Betroffene

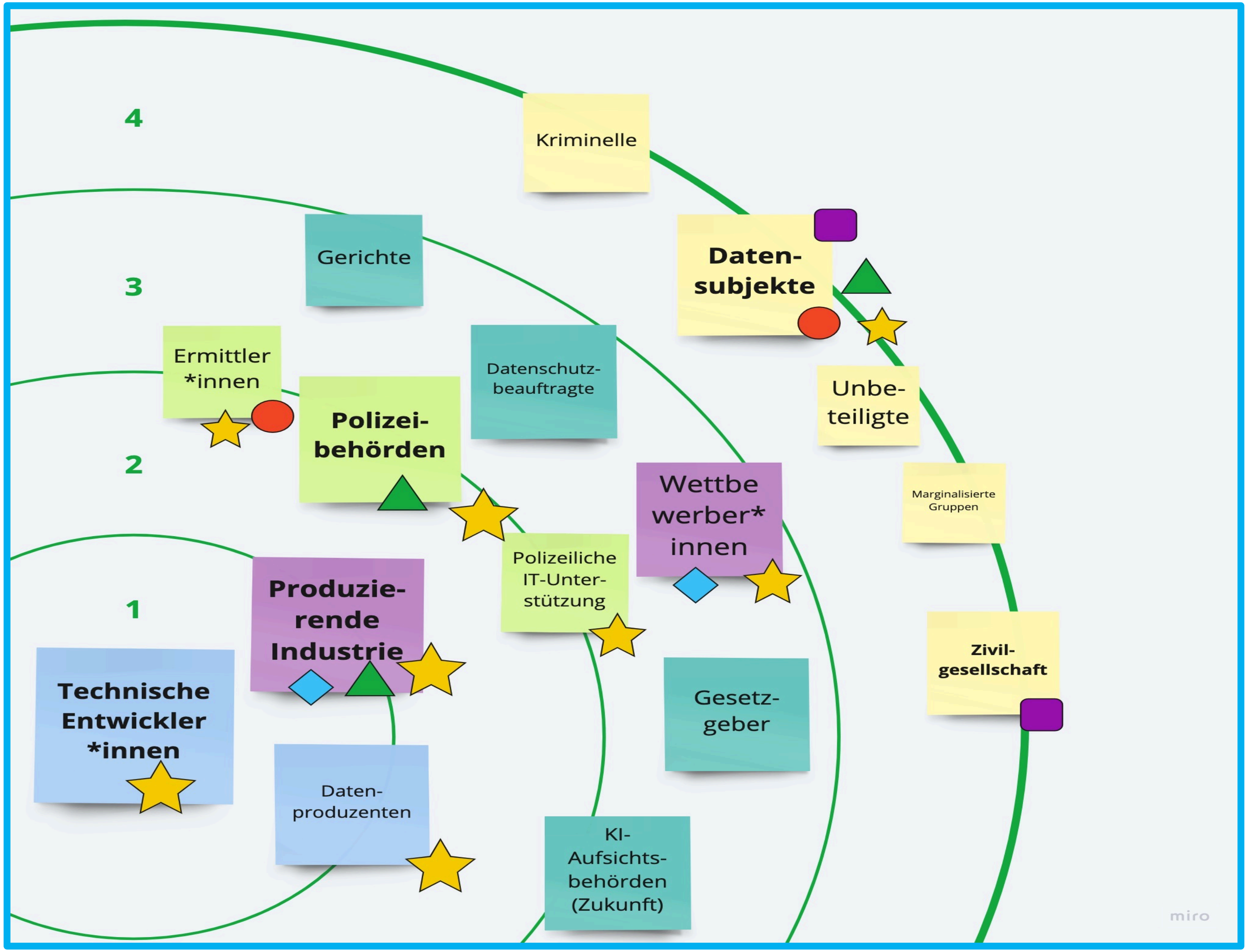






### Risiken

- ◆ Wirtschaftliche Risiken
- ▲ Juristische Risiken
- ☆ Reputationsrisiken
- ● Lebensbedrohliche Risiken
- ■ Grundrechtliche/ethische Risiken


4
Kriminelle
Gerichte
Daten-subjekte
3
Ermittler *innen
Polizei-behörden
Datenschutz-beauftragte
Unbe-teiligte
2
Wettbe werber* innen
Marginalisierte Gruppen
Polizeiliche IT-Unter-stützung
1
Produzie-rende Industrie
Zivil-gesellschaft
Technische Entwickler *innen
Gesetz-geber
Daten-produzenten
KI-Aufsichts-behörden (Zukunft)
miro

# Risks for different stakeholder groups

### Technical developers

- can lose their scientific / professional reputation;

### **Companies** offering or selling the software or service,

- may be banned from a market;
- be sued and lose money and reputation

### Users

- make mistakes that result in lost time, personal or material damage
- get sued or fired for using the software (incorrectly)

### Data subjects

- are wrongly suspected of having committed a crime or intending to commit a crime in the future
- can be discriminated
- be subjected to threats and violence
- be unjustly persecuted, arrested, injured or killed

### Data subjects

- are wrongly suspected of having committed a crime or intending to commit a crime in the future
- can be discriminated
- be subjected to threats and violence
- be unjustly persecuted, arrested, injured or killed

### Institutional actors

- lose confidence in AI technology and stop supporting it (e.g., governments, trade associations, corporations, private and institutional funders, mass media)

### (Civil) **society** actors

- are threatened in their exercise of fundamental rights.

# 1 Exercice

Come back to your stakeholder mapping addressing Palantir’s AIP (Artificial Intelligence Platform) for defense.

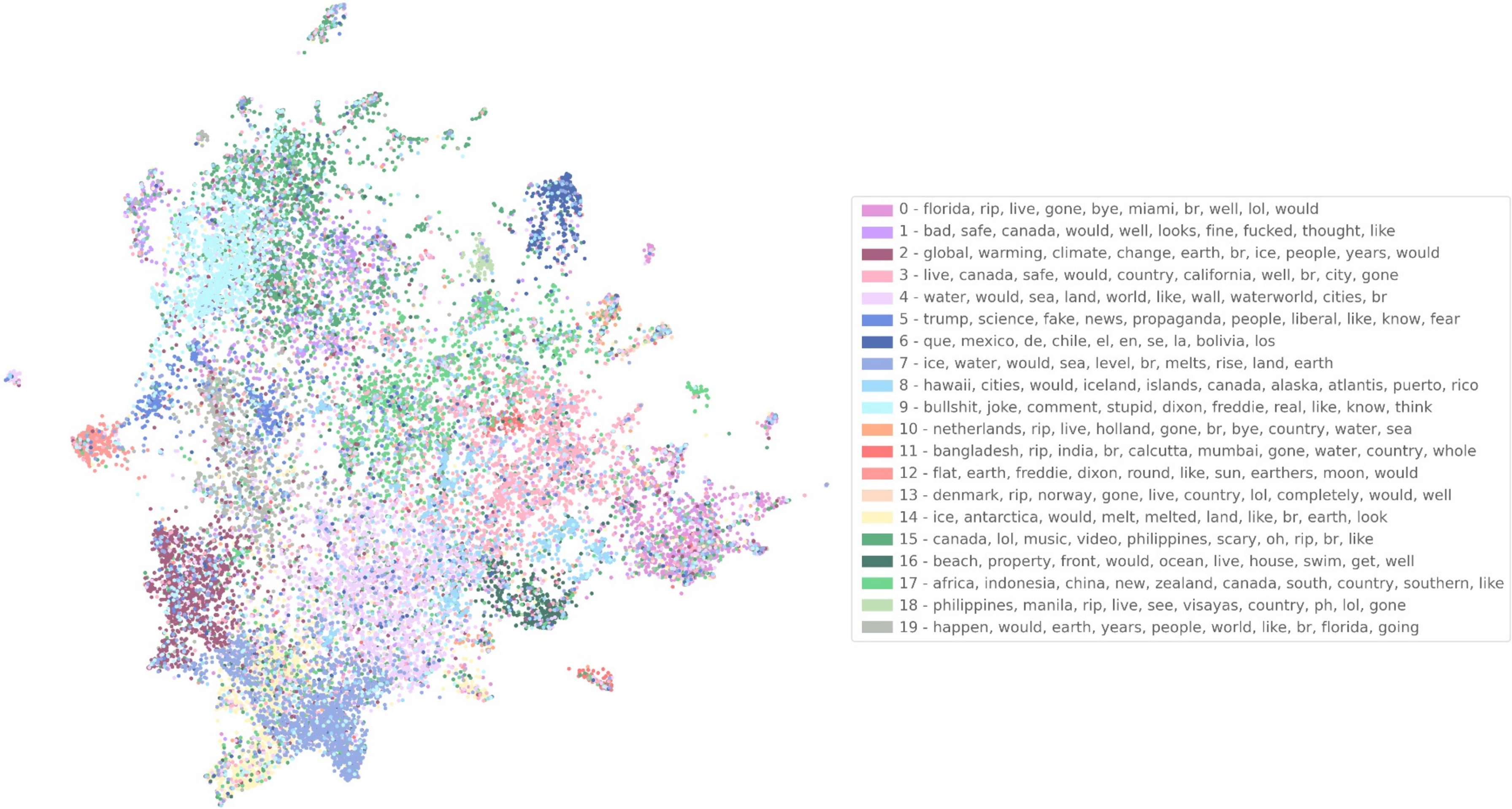
Identify concrete risks and attach them to different stakeholders on the map (post-its).

# Reflexivity in data science

Reflexivity understood as critical self-awareness should be a cornerstone of a *responsible data science*.

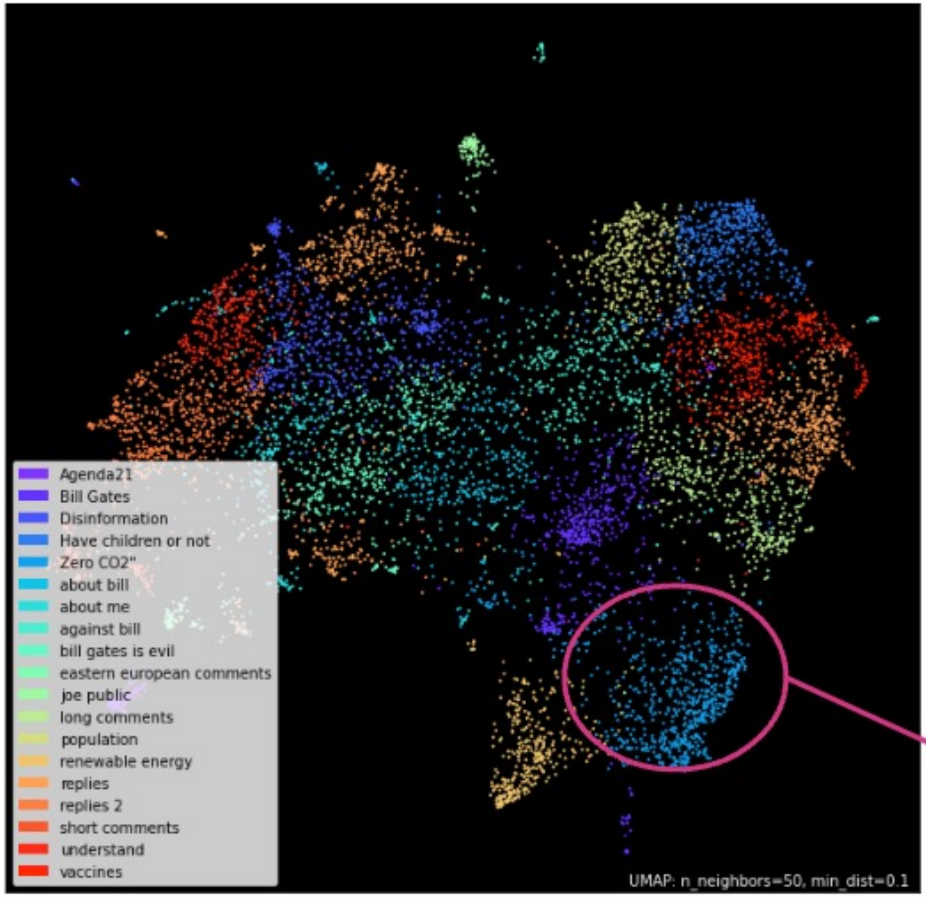
“Reflexive research cultivates a critical self-awareness, including itself among its objects of study and developing useful concepts for reflecting on the research as it is happening.” (Agre 1997a: 27)

# An experiment in reflexivity (I)



(Hirsbrunner et al. 2022)

# An experiment in reflexivity (II)



Example: "Innovating to Zero | Bill Gates", Video on YouTube posted on 21 Feb 2010

**14'426 comments in 10 years**

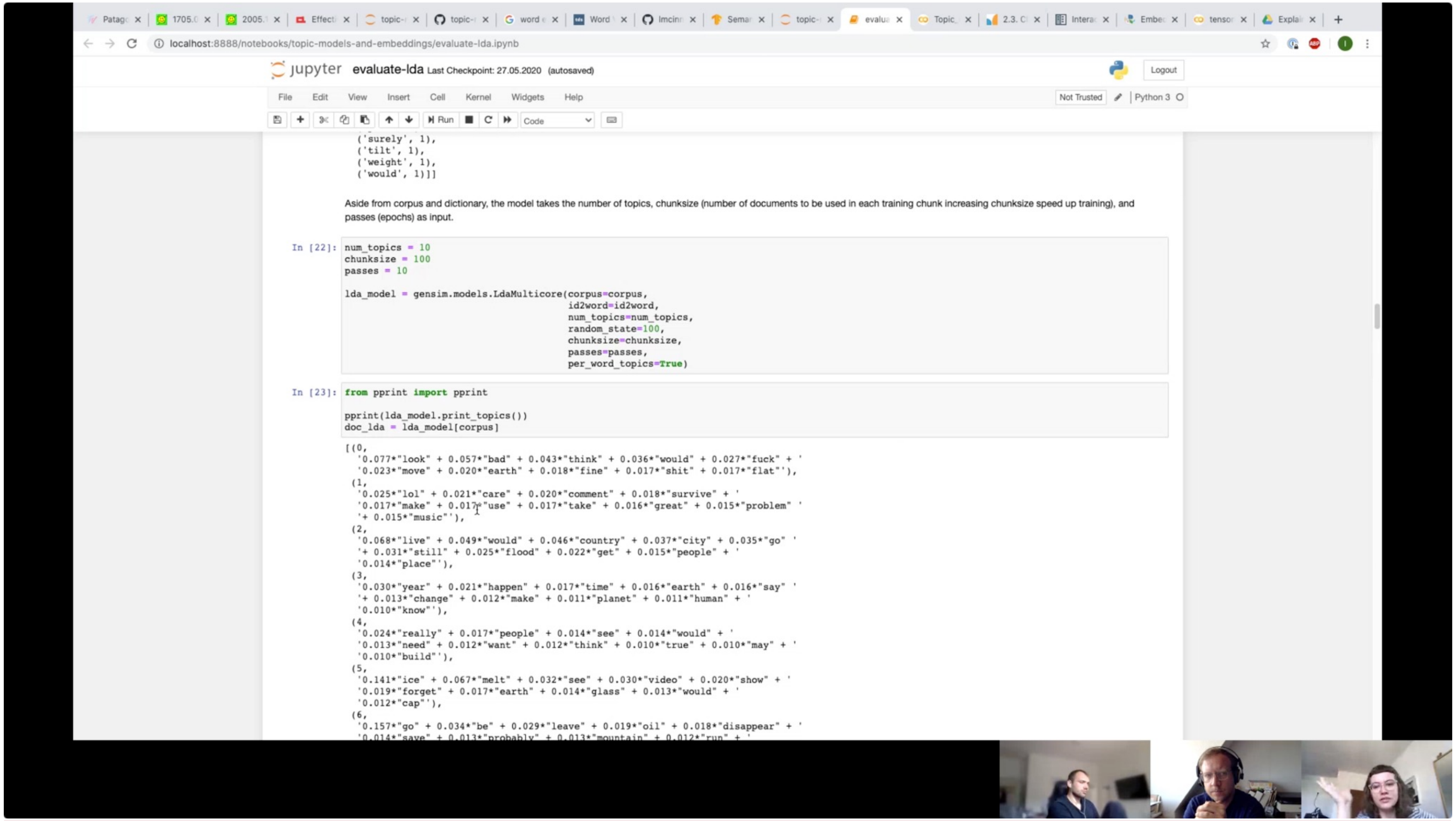
Source: https://www.youtube.com/watch?v=JaF-fq2Zn7I

Why do you hate humanity so much Bill? You do realize if you go 0 CO2 you will have to MURDER all humans off as well, since we breathe CO2 out

12

(Hirsbrunner et al. 2022)

# An experiment in reflexivity (III)


localhost:8888/notebooks/topic-models-and-embeddings/evaluate-lda.ipynb
jupyter evaluate-lda Last Checkpoint: 27.05.2020 (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code

```
 ('surely', 1),
 ('tilt', 1),
 ('weight', 1),
 ('would', 1)]]
```

Aside from corpus and dictionary, the model takes the number of topics, chunksize (number of documents to be used in each training chunk increasing chunksize speed up training), and passes (epochs) as input.

```
In [22]: num_topics = 10
         chunksize = 100
         passes = 10

         lda_model = gensim.models.LdaMulticore(corpus=corpus,
                                                id2word=id2word,
                                                num_topics=num_topics,
                                                random_state=100,
                                                chunksize=chunksize,
                                                passes=passes,
                                                per_word_topics=True)
```

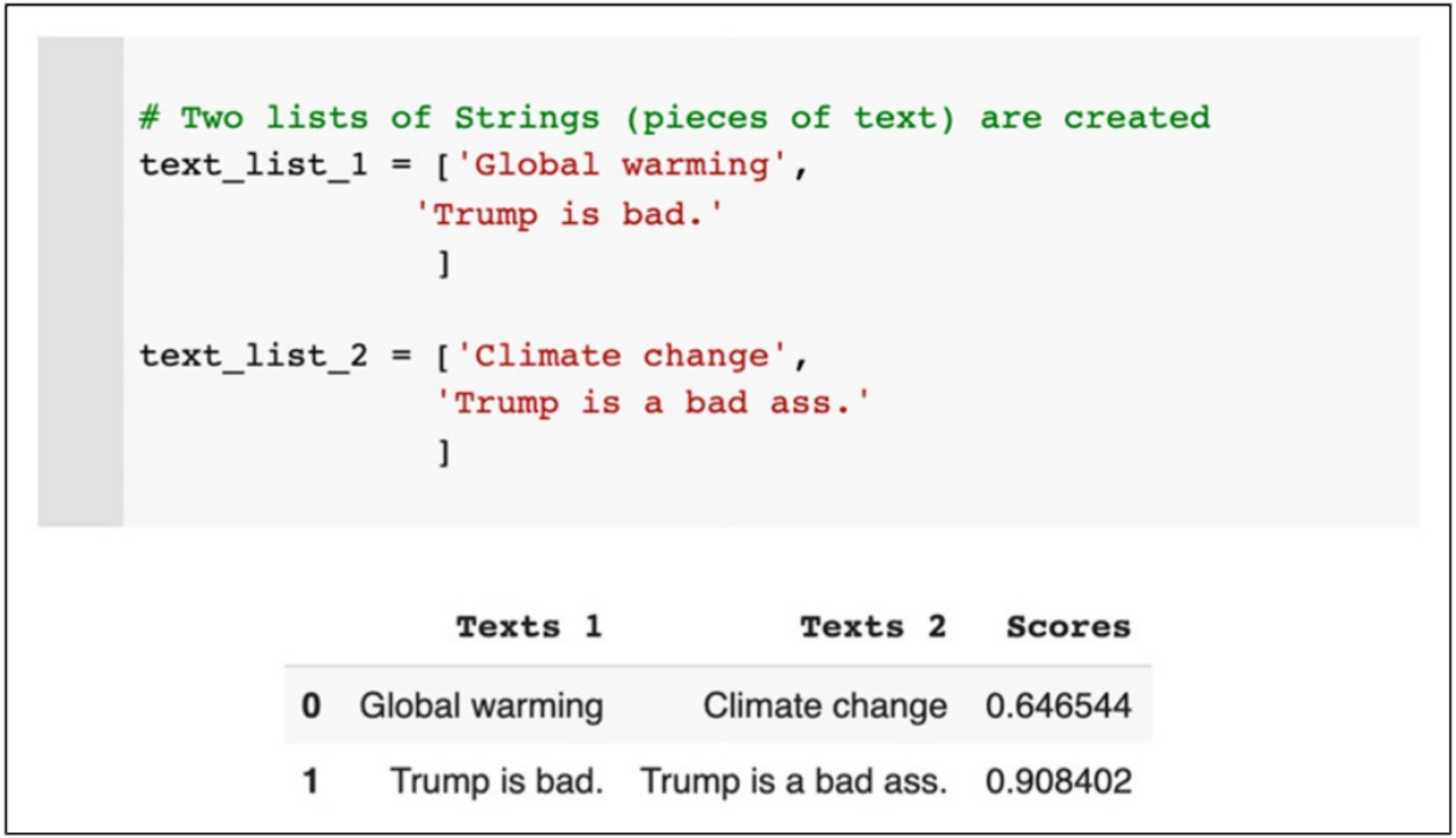
```
In [23]: from pprint import pprint

         pprint(lda_model.print_topics())
         doc_lda = lda_model[corpus]
```

```
[(0,
  '0.077*"look" + 0.057*"bad" + 0.043*"think" + 0.036*"would" + 0.027*"fuck" + '
  '0.023*"move" + 0.020*"earth" + 0.018*"fine" + 0.017*"shit" + 0.017*"flat"'),
 (1,
  '0.025*"lol" + 0.021*"care" + 0.020*"comment" + 0.018*"survive" + '
  '0.017*"make" + 0.017*"use" + 0.017*"take" + 0.016*"great" + 0.015*"problem" '
  '+ 0.015*"music"'),
 (2,
  '0.068*"live" + 0.049*"would" + 0.046*"country" + 0.037*"city" + 0.035*"go" '
  '+ 0.031*"still" + 0.025*"flood" + 0.022*"get" + 0.015*"people" + '
  '0.014*"place"'),
 (3,
  '0.030*"year" + 0.021*"happen" + 0.017*"time" + 0.016*"earth" + 0.016*"say" '
  '+ 0.013*"change" + 0.012*"make" + 0.011*"planet" + 0.011*"human" + '
  '0.010*"know"'),
 (4,
  '0.024*"really" + 0.017*"people" + 0.014*"see" + 0.014*"would" + '
  '0.013*"need" + 0.012*"want" + 0.012*"think" + 0.010*"true" + 0.010*"may" + '
  '0.010*"build"'),
 (5,
  '0.141*"ice" + 0.067*"melt" + 0.032*"see" + 0.030*"video" + 0.020*"show" + '
  '0.019*"forget" + 0.017*"earth" + 0.014*"glass" + 0.013*"would" + '
  '0.012*"cap"'),
 (6,
  '0.157*"go" + 0.034*"be" + 0.029*"leave" + 0.019*"oil" + 0.018*"disappear" + '
  '0.014*"save" + 0.013*"probably" + 0.013*"mountain" + 0.012*"run" + '
```

(Hirsbrunner et al. 2022)

# An experiment in reflexivity (IV)



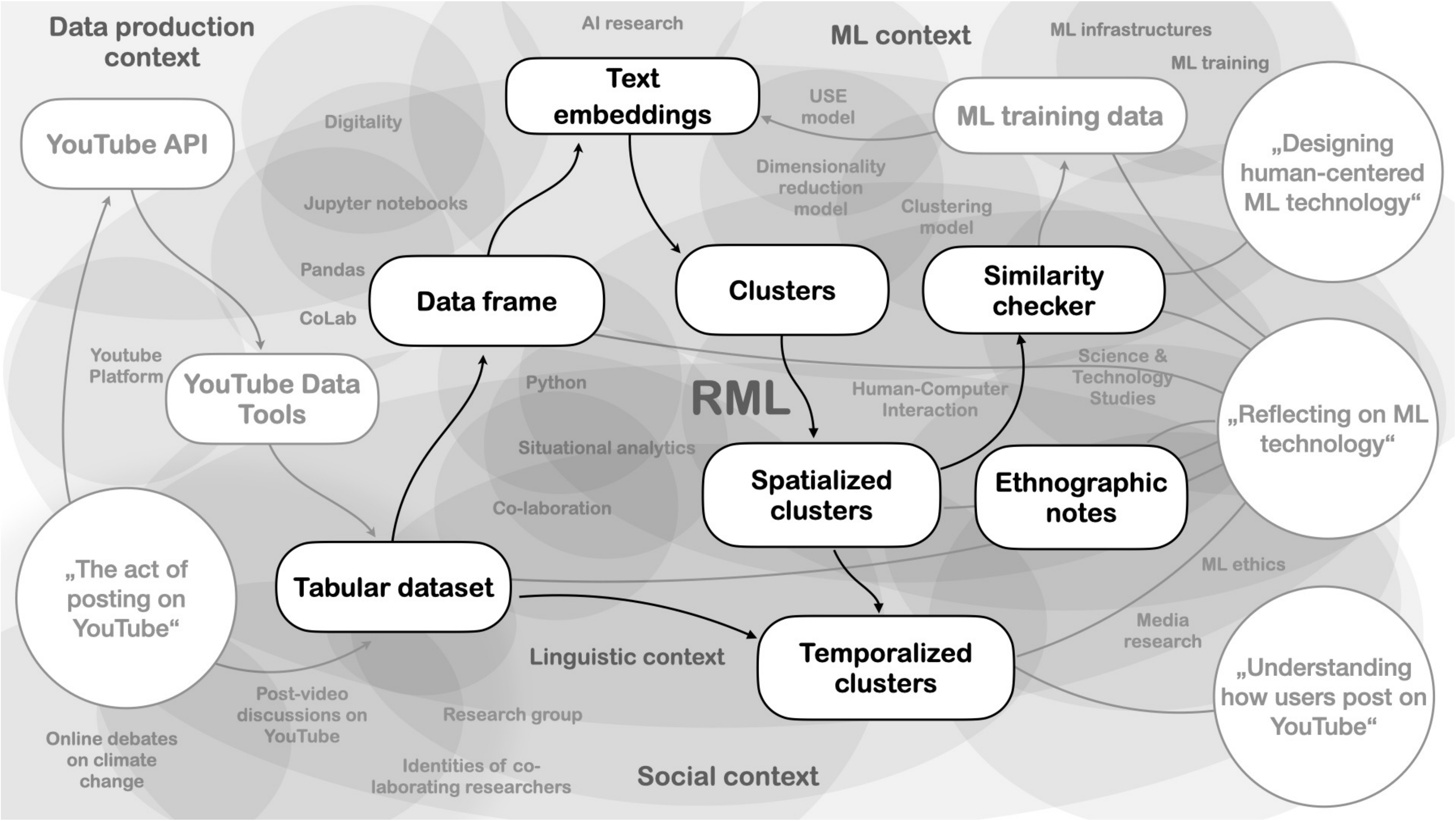
```
# Two lists of Strings (pieces of text) are created
text_list_1 = ['Global warming',
             'Trump is bad.'
              ]

text_list_2 = ['Climate change',
              'Trump is a bad ass.'
              ]
```

|   | Texts 1 | Texts 2 | Scores |
|---|---|---|---|
| 0 | Global warming | Climate change | 0.646544 |
| 1 | Trump is bad. | Trump is a bad ass. | 0.908402 |

(Hirsbrunner et al. 2022)

# An experiment in reflexivity (III)



(Hirsbrunner et al. 2022)

# 2 Exercice

Take up the role of one of the stakeholders. Make some notes about the interests of this stakeholder.

Defend the interests of the stakeholder within the staged design process.

# Sources

See entire list of course references on Github: https://github.com/simonsimson/responsible-data-science/tree/main/slides