# **Democratically healthy news recommendation:** aligning NLP with society, theory, and evaluation

Myrthe Reuver
CLTL, VU Amsterdam

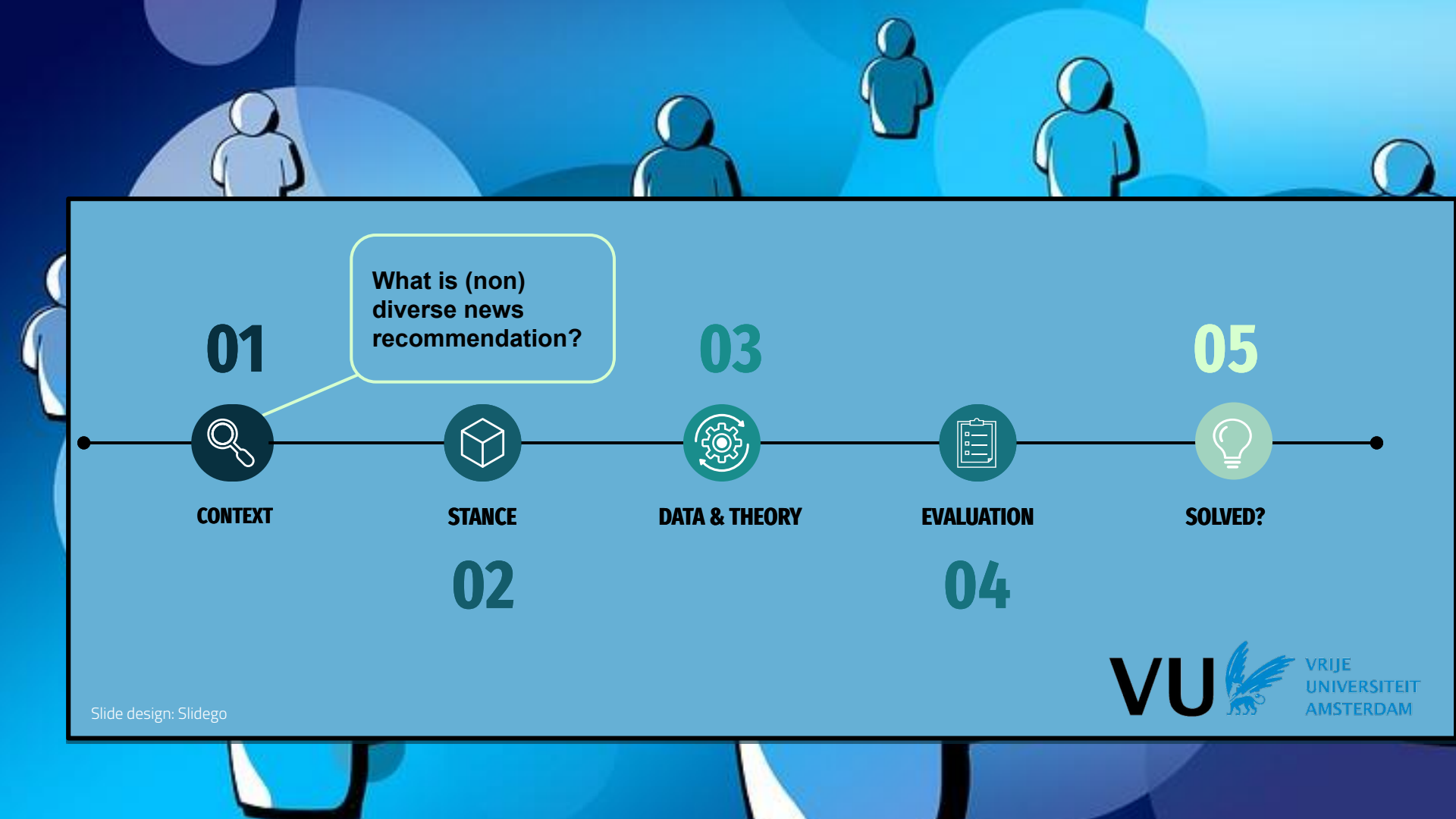
VU VRIJE UNIVERSITEIT AMSTERDAM

# Who am I?

Myrthe Reuver, 4th year (!) PhD candidate at CLTL, VU.

supervisors: prof.dr. Antske Fokkens (VU), prof.dr.Suzan Verberne (Leiden University)

Computational linguist in an **interdisciplinary project** on diversity in news recommendation.

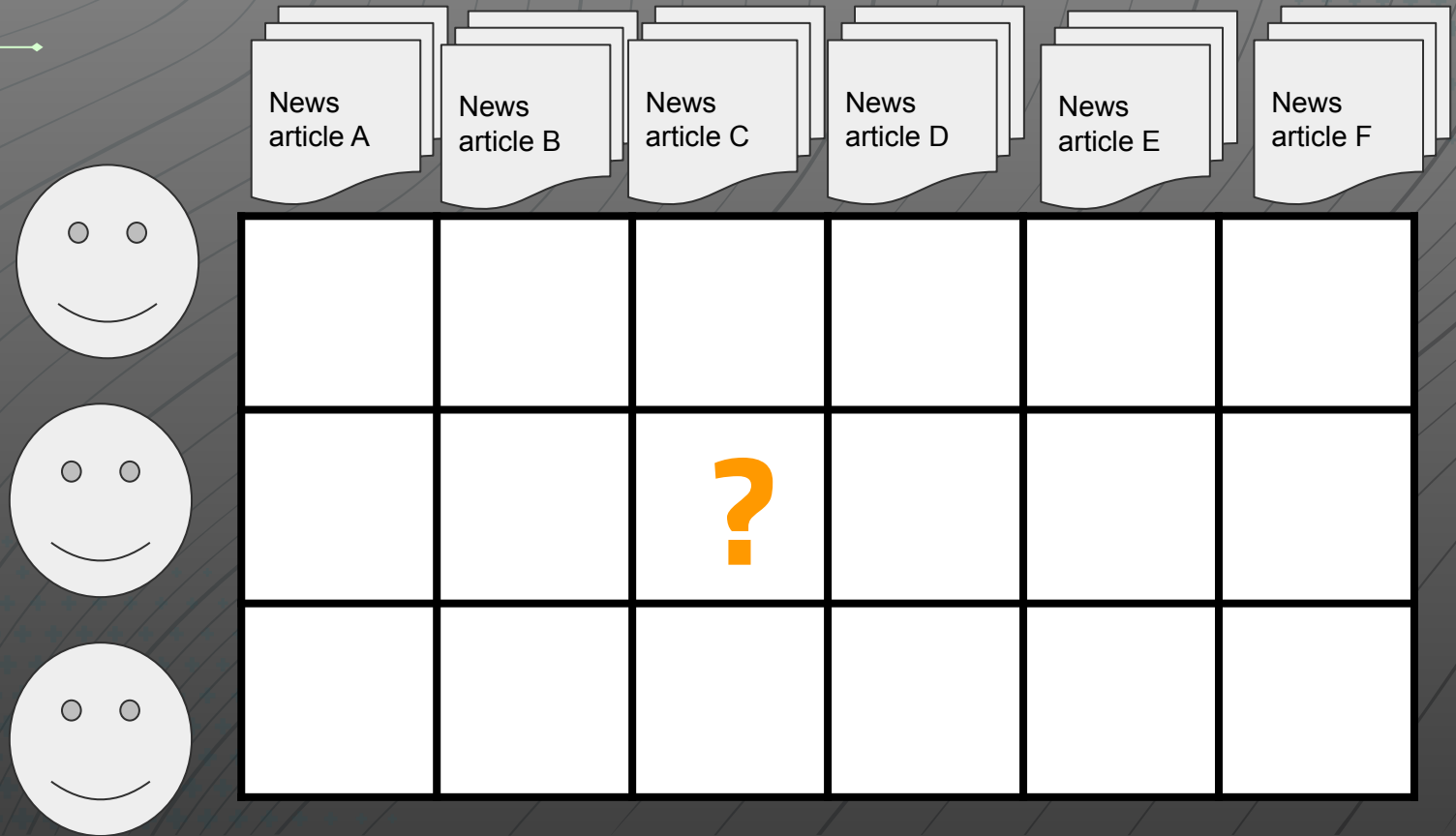



Slide design: Slidego

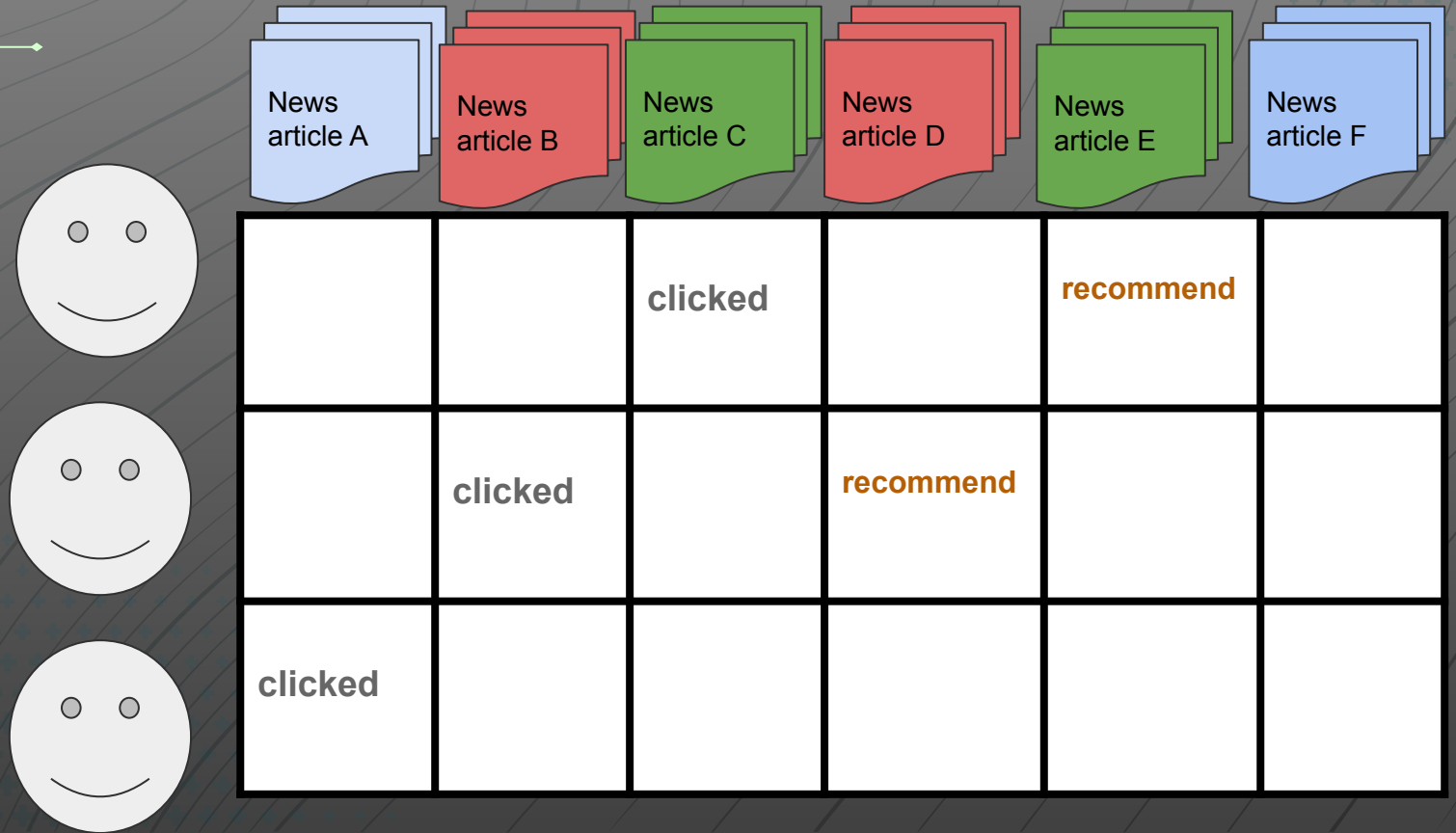
VU VRIJE UNIVERSITEIT AMSTERDAM

# What is a (news) recommender system?

# What is a (news) recommender system?

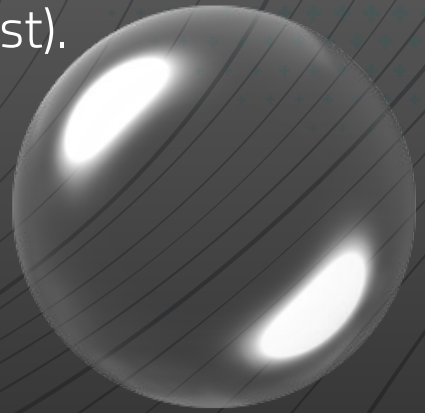


# What is a (news) recommender system?



# Optimization in News Recommendation

Usually in RecSys: **click-accuracy** (as proxy for user interest).
Consequence: Showing users more of the same.



- Could lead to ‘filter bubbles’
- Potentially problematic for democracy.

I started my PhD consulting experts (social scientists and theorists):

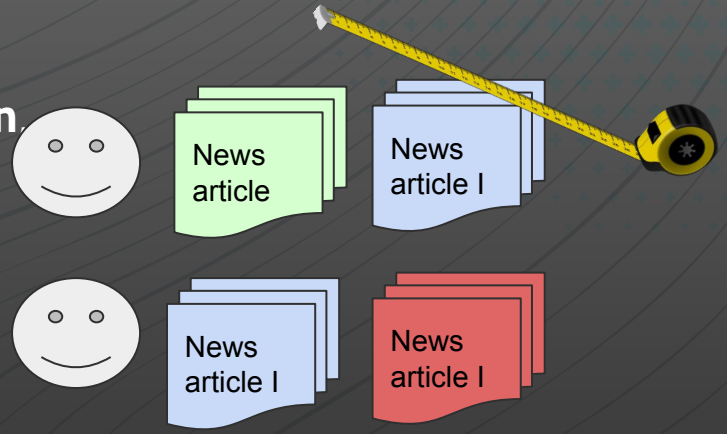
**Why** is this problematic for democracy and society?

# Models of Democracy


- Theoretical models that define **a functioning, ideal democracy**.
- **Deliberative model:** democracy requires **rational debate** , and actors encountering **a diverse set of viewpoints and ideas** on (societal) issues.
- Helberger (2019) connected this and other models to news recommendation.
- Supporting a deliberative model of democracy = recommenders promoting **rational rather than emotional content**, and **diverse viewpoints** on issues.

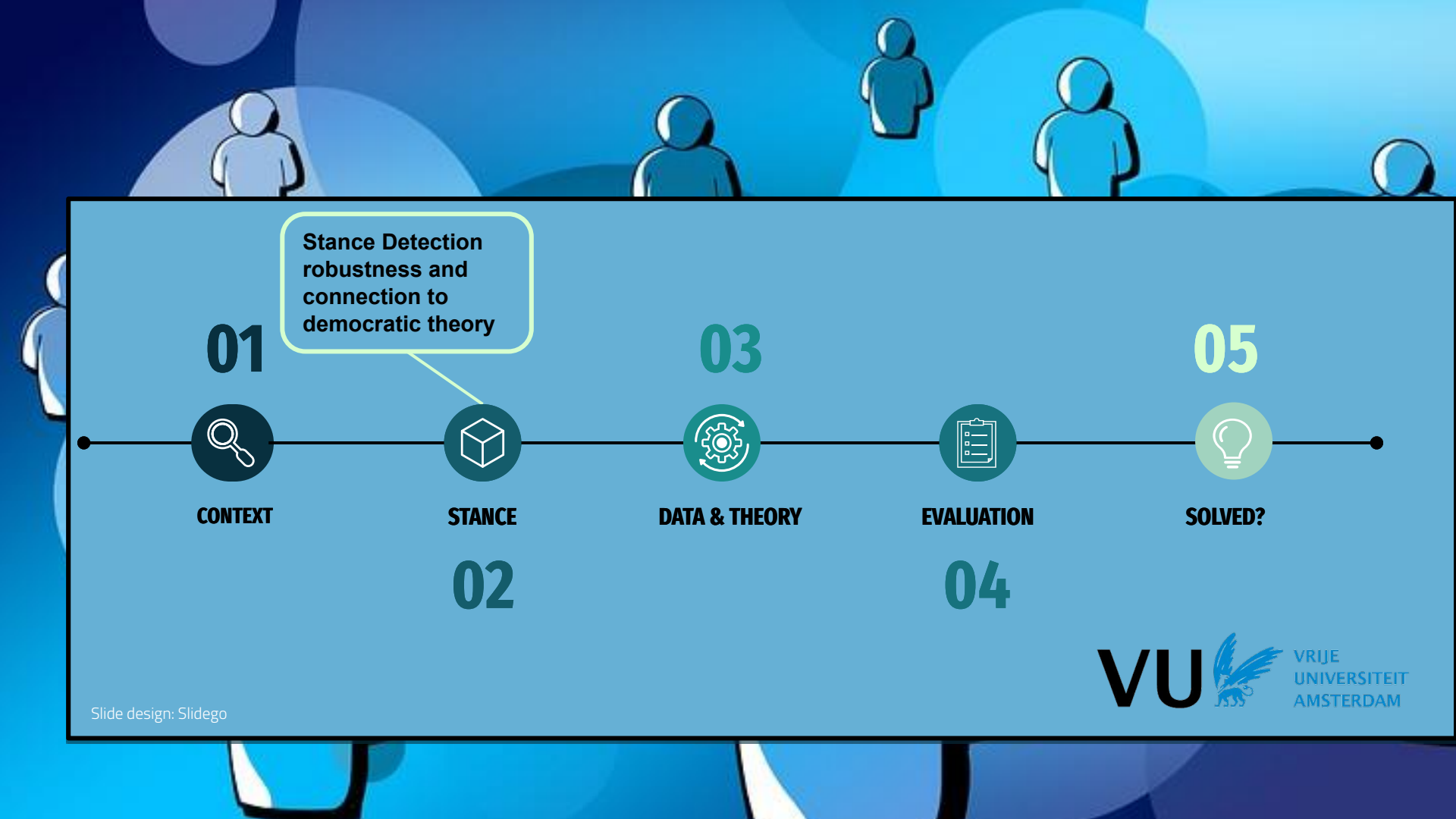
# Measuring health of recommendations

Optimization **beyond individual user satisfaction**, but on the **collective information environment,** with new metrics:




1) **Fragmentation:** shared public sphere
2) **Representation:** diverse actors and viewpoints
3) **Alternative Voices:** non-mainstream opinions
4) **Calibration:** personalization
5) **Affect:** emotional content

- **Deliberative debate** requires low affect, low fragmentation
- **Critical democracy model** (where viewpoints clashing is considered healthy) requires high affect.


Stance Detection robustness and connection to democratic theory
01
CONTEXT
02
STANCE
03
DATA & THEORY
04
EVALUATION
05
SOLVED?
Slide design: Slidego
VU
VRIJE UNIVERSITEIT AMSTERDAM

# Operationalizing ‘different viewpoints’:

## Viewpoint + difference

**Theoretical concept:**

Rather vague, e.g. “a wide range of perspectives on a given issue” (Griswold, 1998)

**Operationalization with NLP:**

a relationship between topics, texts, and tasks (stance, sentiment)

- tuple of entity, a topic related to this entity & sentiment towards topic (Ren et al, 2016)
- the standpoint of one or several authors on a set of topics (Thonet, 2016, 2017)
- stance + sentiment score of news articles towards refugee statements (Alam et. al., 2022)

Key: looking beyond **islands** of tasks, task definitions, and datasets

# Different tasks for "different viewpoint"

- Frames (Mulder et. al., 2021)
- Stances (Reuver et. al, 2021)
- Ideology (left-wing/right-wing, conservative/liberal, etc.)
- Perspective (content+stance, Fokkens et. al., 2017)
- Values of argument (e.g. "Economic Prosperity" vs "Mental Health", Liscio et. al., 2021)
- Morals from Moral Foundation Theory (Kobbe et. al, 2020)
- Types of argument ("Moral", "Civic", "Economic",Baden, 2017; Draws et. al. 2022)

.. or: combination of the above → e.g. multidimensional, see Draws et. al. 2022 combining **stance + argument type**, or **stance + moral foundation** (Kobbe et. al., 2020), or **stance + sentiment** score specifically in news articles (Alam et. al., 2022)

# Thinking beyond Task

Theory on democracy requires **nuanced aspects beyond task definition**:

- Our ultimate goal is not e.g. detecting stances, but **supporting democracy.** That might mean some stances, viewpoints, or ideas (attacking democracy, or inherently violent stances) should **not** be recommended, and **may require a separate class;**
- We also need to not only detect stances, but **find diverse, different, or opposing** viewpoints and ideas.

Reuver et. al. (2021) "No NLP Task Should be an Island: Multi-disciplinarity for Diversity in News Recommender Systems" Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation.

# What is (going on with) stance?

Stance detection, common definition: **classification task** (on texts, often tweets) with labels Pro, Con, Neutral towards an issue or topic

*"Abortion is a sin, and should never be practiced."*

Topic: Abortion, Stance: Con

# Why stances?

Built upon the linguistic phenomenon of **actors communicating their evaluation of targets**;

placing **themselves and their targets on "dimensions in the sociocultural field"** (Du Bois, 2007).

Directional (pro/con) →Immediate **connected to (deliberative) debate & democratic decision making**

→ agree/disagree with laws, proposals, etc.

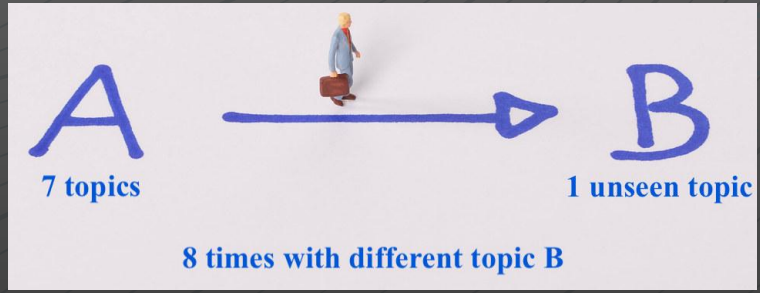
# Cross-topic stance classification

Train: 7 topics, test: 8th topic

Fine-tuning BERT (base & large)

Findings:

- avg. F1 (10 seeds) = 0.633
- +0.20 over reference model (LSTM)
- Results are *"very promising and stress the feasibility of the task"* (**Reimers et al. 2019**, p. 575)




Marco Verch @ Flickr, Creative Commons 2.0.
https://foto.wuestenigel.com/businessman-walking-from-a-to-b-point/

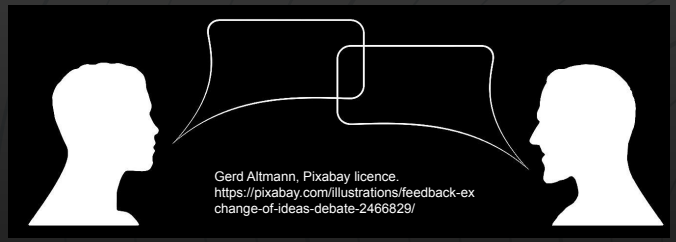
# **Dataset: UKP Dataset** (Stab et. al., 2018)

25,492 arguments on 8 topics, in 3 classes:

- For or against "the use, adoption, or idea" of the topic, or no argument

- 8 controversial debate topics from internet forums:

  *abortion, cloning, death penalty, gun control, marijuana legalization, minimum wage, nuclear energy* and *school uniforms*.



# Reproduction

- Reproduction is important for science, and NLP specifically! (Fokkens et al., 2013; Belz et al., 2021).

- **Systematic reproduction:** 3 dimensions of reproduction (Cohen et. al.,2018): numeric values, findings, conclusions.

- Non-deterministic results of BERT:
  - **Standard deviation (SD) over seeds**;
  - value is reproduced if it falls **within 2 SDs.**

Reuver et. al. (2021b). Is Stance Detection Topic-Independent and Cross-topic Generalizable?- A Reproduction Study. In P*roceedings of the 8th Workshop on Argument Mining.*

| Mean (stdv) over 10 seeds | F1 |
|---|---|
| **Reimers et. al. (2019)** | |
| LSTM (baseline) | **.424** |
| BERT-base | .613 (-) |
| BERT-large | .633 (-) |
| **Reuver et. al (2021)** | |
| **SVM+tf-idf (baseline)** | **.517** |
| Reproduction BERT-base | .617 (.006) |
| **Reproduction BERT-large (all)** | **.596 (.043)** |
| **BERT-large - 5 good seeds** | **.636 (.007)** |

## Results:

BERT-large under-performs in 50% of seeds

SVM+tf-idf model outperforms the LSTM reference model from the original study (F1 of .517 > .424)

# Cohen et. al. [2018]'s 3 dimensions of reproducibility

1. **(numeric) values:**

   Within 2 standard deviations

2. **findings** (relationship between variables, e.g. model & result):

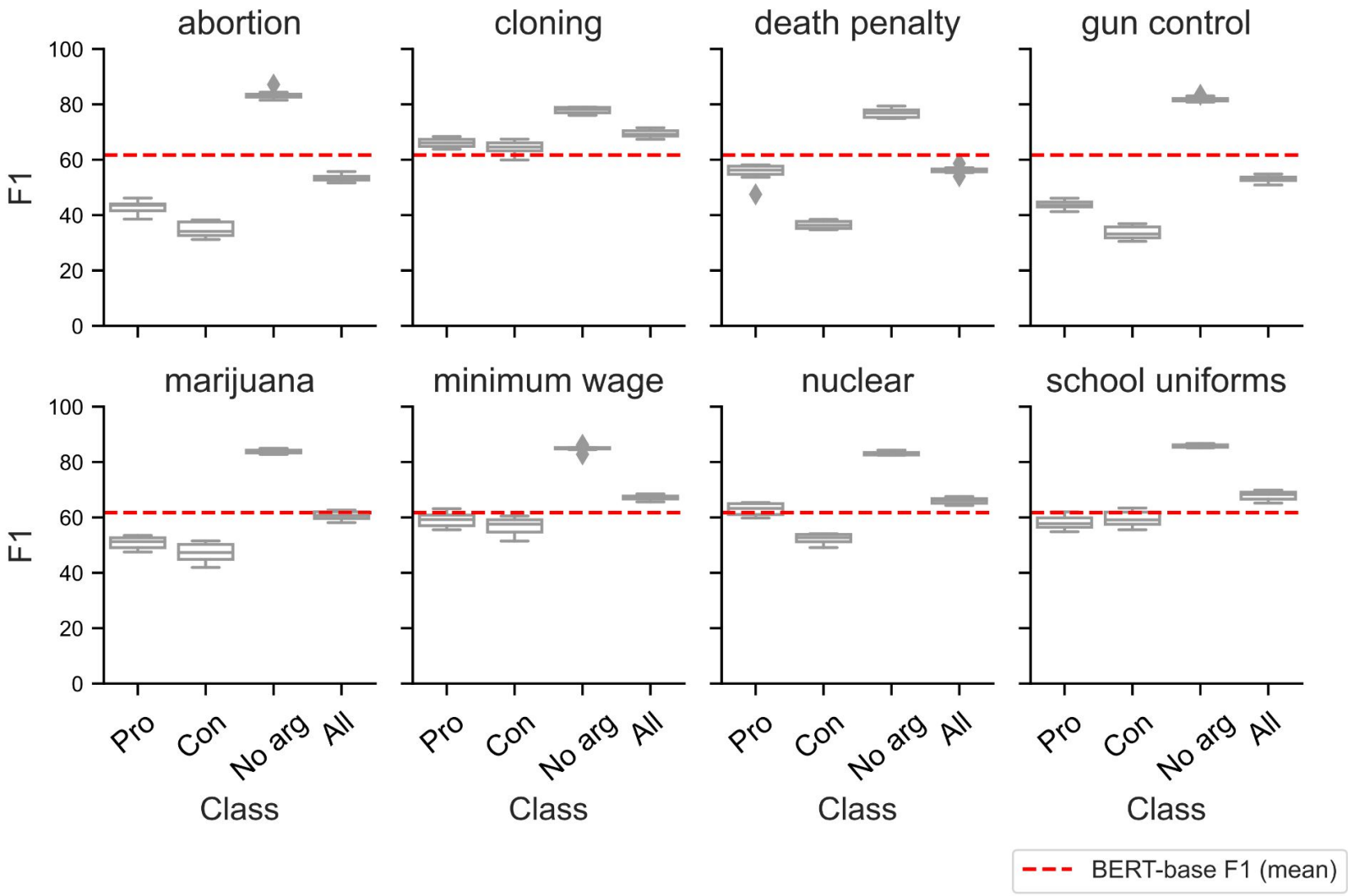
   baseline < BERT-base < BERT-large,

3. **conclusion(s):**

   How feasible is cross-topic stance detection?

# Cohen et. al. [2018]'s 3 dimensions of reproducibility

1. **(numeric) values:**
   - ✓ Within 2 standard deviations (BERT-large = large SD)
2. **findings** (relationship between variables, e.g. model & result):
   - ✓ baseline < BERT-base < BERT-large,
   - ✗ .20 improvement over baseline is (much) smaller with SVM
3. **conclusion(s):**
   - ? How feasible is cross-topic? Let's investigate some more, especially on topics.


abortion
cloning
death penalty
gun control
marijuana
minimum wage
nuclear
school uniforms
F1
100
80
60
40
20
0
Pro
Con
No arg
All
Class
BERT-base F1 (mean)

# What does this mean?

Successful reproduction of cross-topic stance classification (Reimers et. al., 2019) on most dimensions, **but**:

- **random seed** does matter for BERT-large;
- **baseline** matters;

**Topic matters!** Stance not as topic-independent as seems with one averaged F1 metric reported.

- See also: Thorn Jakobsen et. al. (2021)

A **class/topic interaction effect in stance**

OpenClipArt, Public domain

Building data from theory, and theory from data

01 CONTEXT

02 STANCE

03 DATA & THEORY

04 EVALUATION

05 SOLVED?


Slide design: Slidego

# Data and **topic-specificity** in stance

Stance often framed as **topic-agnostic**, but:

- **Aspects of stances** are **specific to the topics under discussion**
  example: abortion → rights of individuals, nuclear energy → harms vs benefits
- Earlier work has focussed on "when topics are similar enough" in the **modelling phase.** However, we can already capture this in **task definition & data.**
- **Topic-specific stance data and tasks** can increase the impact of stance detection on societally relevant research.

# **Topic-specificity** in social science theory

Social scientists often study **debate dynamics** of specific topics

**Theories and findings** from these studies can be used in stance definitions and datasets, to better reflect stance-taking activities on these topics.

In my pilot study: debates on **sustainability initiatives.**

# What drives (a lack of) support for sustainable initiatives expressed in online discussions?

Project with Ana Isabel Lopes (VU Communication Science) and Alessandra Polimeno (as VU student assistant) - funded by Network Institute grant.

# Value-belief-norm theory (Stern et al., 1999).

People who provide support believe:

- valued objects are **threatened**
  → (threat/no threat dimension)
- their actions can help **restore** those values
  → (power/ no power dimension)

image: Free SVG, CC licence

# Trust in Sustainable Initiatives

image: Free SVG, CC licence

**Valued object:** environment

**Initiative:** consumption of locally produced food

Distrusting people may not believe;

- non-local food production affects the environment (**no threat**)
- changing food habits has a collective effect (**no power to restore**)

# Argument Mining / Computational Argumentation

- Trust/distrust and threat/power → **different types of arguments for** or **against** an initiative
- Stances can be **implicit**, only mention the trust/distrust or threat/power aspect instead of saying "I am against X".

  → Example: "This is not going to help the environment because plastic bag payments are too small to have effect"
- For now: **an exploratory guided approach** in this manner of topic-specific data creation

# Data: Reddit discussions, **European** context

- **3** EU Reddit boards, **2.073** individual discussions on initiatives, **46.285 comments** → found with annotated word list, expanded with embeddings
- **5 years (2017-2022),** Language: **English** (but some multilingual comments), **preprocessing**: removing comments of bots
- **Currently:** 100 annotated posts, working on improving & expanding

# Examples

| Post title (topic) | Comment | Stance | Threat | Power |
|---|---|---|---|---|
| Spanish should eat less meat to limit climate crisis, says minister | He’s right. High levels of meat consumption and bio industry is a threat to all of humanity | Pro | Threat | Not mentioned |
| The road to sustainability: the superhighway built from paper waste instead of cement | it seems like a dumb idea. its a solution for something that is not a problem | Con | No threat | Not mentioned |
| Recycling rate of plastic packaging waste | Recycling plastic is mostly pointless. Far better to reduce the use of plastics in packaging as much as possible." | Con | Threat | No Power |

# Annotation study - work in progress


- 91 examples: 42 have a clear stance
- Power and threat receive fair agreement with 5 annotators (Fleiss K = 33), but improvement on the way

**In annotations:**

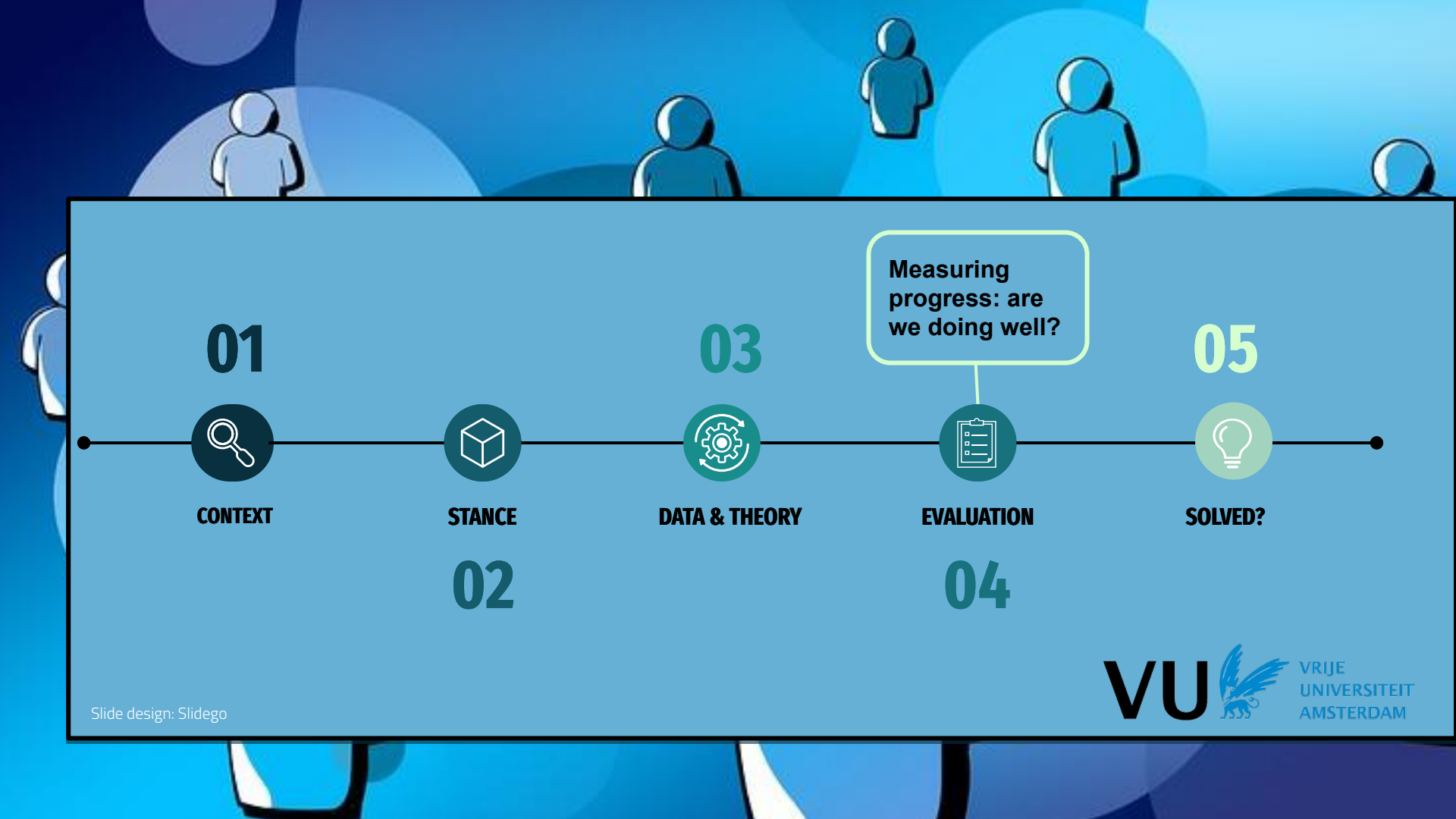
- more **threat** (28 comments) than no threat (10)

  → seeing a threat is related to positive stance for initiative
- More **lack of power** (24 comments) than power (8 comments)

  → **lack of power** is related to negative stance to initiative

# Future Questions

- Do theoretical aspects on stances help;
  - Annotation of stances?
  - Modelling stances?
  - Impact of these models?




Slide design: Slidego

VU VRIJE UNIVERSITEIT AMSTERDAM

# Normative metric: Fragmentation

## Fragmentation:

are citizens in a society **aware of the same news events** when receiving news recommendations?

If not, this can lead to **fragmentation of the public sphere.**


Paper: Polimeno et. al. (2023) Improving and Evaluating the Detection of Fragmentation in News Recommendations with the Clustering of News Story Chains. Proceedings of Normalize 2023, at RecSys 2023

# How can we best measure and evaluate the detection of Fragmentation?

**Requirements for operationalization:**

→ detecting **different articles mentioning the same** event or story, across news outlets

Related tasks: News *story chain* clustering (e.g. Van Hoof et. al., 2019)

What is needed to operationalize:

- a task ✓
- and a fitting **dataset for evaluating our approach**

# HeadLine Grouping Dataset (Laban et. al., 2021)

(American) English dataset of news titles

- 10 diverse events
- with human ground truth labels

| # | Topic | Size |
|---|---|---|
| 1 | Human Cloning | 108 |
| 2 | International Space Station | 215 |
| 3 | Ireland Abortion Vote | 170 |
| 4 | US Bird Flu Outbreak | 75 |
| 5 | Facebook Privacy Scandal | 172 |
| 6 | Wikileaks Trials | 153 |
| 7 | Tunisia Protests | 86 |
| 8 | Ivory Coast Army Mutiny | 104 |
| 9 | Equifax Breach | 156 |
| 10 | Brazil Dam Disaster | 247 |

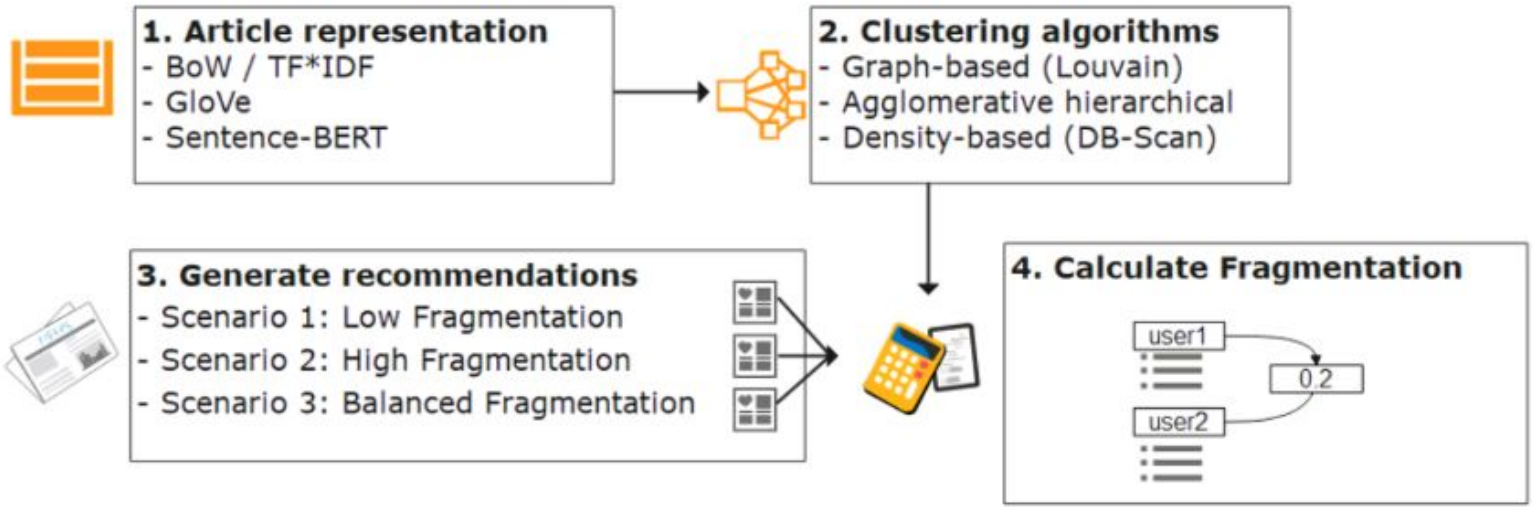
Procedure:

- scrape URLs;
- use entire news articles.

Final dataset: 1,394 articles in 10 events

→ 3 held-out events for testing

# Experiments: intrinsic (2) vs extrinsic (3) evaluation



# Intrinsic: Clustering News Story Chains

| Setup | H ↑ | C ↑ | V ↑ | S ↑ | DBI ↓ |
|---|---|---|---|---|---|
| Baseline | 0.166 | 0.156 | 0.161 | -0.060 | 12.441 |
| AHC*SBERT | **0.921** | **0.844** | **0.881** | 0.290 | 1.933 |
| AHC*GloVe | 0.762 | 0.708 | 0.734 | 0.183 | **1.913** |
| AHC*BoW | 0.813 | 0.658 | 0.727 | **0.413** | 1.965 |
| DB*SBERT | 0.694 | **0.872** | **0.773** | 0.231 | 1.509 |
| DB*GloVe | 0.002 | 0.236 | 0.004 | **0.390** | 0.387 |
| DB*BoW | **0.993** | 0.283 | 0.441 | 0.213 | **0.218** |

# Extrinsic: Do we capture Fragmentation?

| Scenario | Chains per user | Fragmentation |
|---|---|---|
| Scenario 1 | 7 | Low |
| Scenario 2 | 1 | High |
| Scenario 3, profile 1 (70%) | 5 | Balanced |
| Scenario 3, profile 2 (15%) | 2 | Balanced |
| Scenario 3, profile 3 (15%) | 7 | Balanced |

| Setup | Scen. 1 ↓ | Scen. 2 ↑ | Scen. 3 | Variation |
|---|---|---|---|---|
| Gold | 0.00 | 0.85 | 0.58 | 0.85 |
| Baseline | 0.67 | 0.73 | 0.70 | 0.06 |
| AHC*SBERT | 0.31 | 0.87 | 0.64 | 0.56 |
| AHC*GloVe | 0.38 | 0.84 | 0.63 | 0.46 |
| AHC*BoW | 0.62 | 0.85 | 0.63 | 0.23 |
| DB*SBERT | 0.16 | 0.74 | 0.48 | 0.58 |
| DB*GloVe | 0.01 | 0.01 | 0.00 | 0.01 |
| DB*BoW | 0.99 | 0.99 | 0.99 | 0.00 |

- **DB*GloVe:** only 3 clusters
- **Low Fragmentation** is hard to detect;
- **AHC-based approaches with embeddings** show most difference between different scenarios



41

# Take-aways

Cluster coherence evaluation without ground truth can give misleading results

→ *Human-labelled evaluation data is important!*

Intrinsic: SBERT*AHC vs TF-IDF baseline (V = .88 vs .16)

Extrinsic: SBERT*AHC detects difference in scenarios best

scores a **low** Fragmentation with .31;

All implementations scored a low scenario with => .16

Implementers focus on **contrasting** scores rather than absolute scores

"this set of users exhibits significantly lower Fragmentation compared to other users"

"this set of users shows low Fragmentation"

01

CONTEXT

02

STANCE

03

DATA & THEORY

04

EVALUATION

05

SOLVED?

Have we solved this problem?

VU VRIJE UNIVERSITEIT AMSTERDAM

Slide design: Slidego

# There is not one answer to what is the “best” NLP for democratically healthy news recommendation

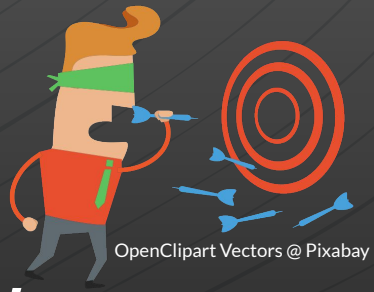
Rather, this is dependent on:

- **Which values and democratic model** stake-holders want to support;
- The **topic under discussion**, and its context-dependent aspects;
- important to align these with **evaluation** approach also.

… and: **user behaviour**!

# One important additional factor: **user behavior!**

Nicolas Mattis, VU PhD in social science in my project, is running experiments on **how users respond to different diverse and less diverse news.**


OpenClipart Vectors @ Pixabay

Heitz et. al. (2022): **Benefits of Diverse News Recommendations for Democracy: A User Study** indicated news recommender users **appreciate** different opinions (weak labelled stances) in their news recommendations.

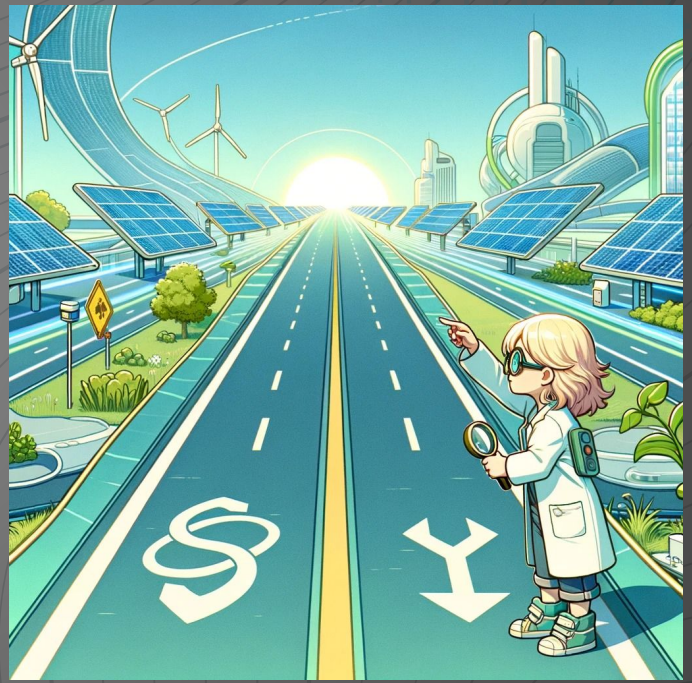
# Conclusion

NLP in news recommendation means juggling different key decisions: theoretical concept (of viewpoint and of democracy), task, data, and evaluation. Also, input from different experts!

There's **no single answer** when it comes to what is the "best" democratically healthy news recommendation, or an NLP model for it.

# The Future

Among others: next semester visiting Prof. Gabriella Lapesa at GESIS in Cologne to continue work on NLP with **social science theory, conceptualization, and careful evaluation**: this time with instruction-tuned models!



This is what GPT generates with prompts in the realm of "a female scientist researching a responsible future" looks like →