open space

Feel free to approach us in case of questions…
(microphone or chat)

«Critical Social Media Analysis using Mixed Methods»

Clustering and Visualization

Michael Tebbe, Dr. Simon David Hirsbrunner

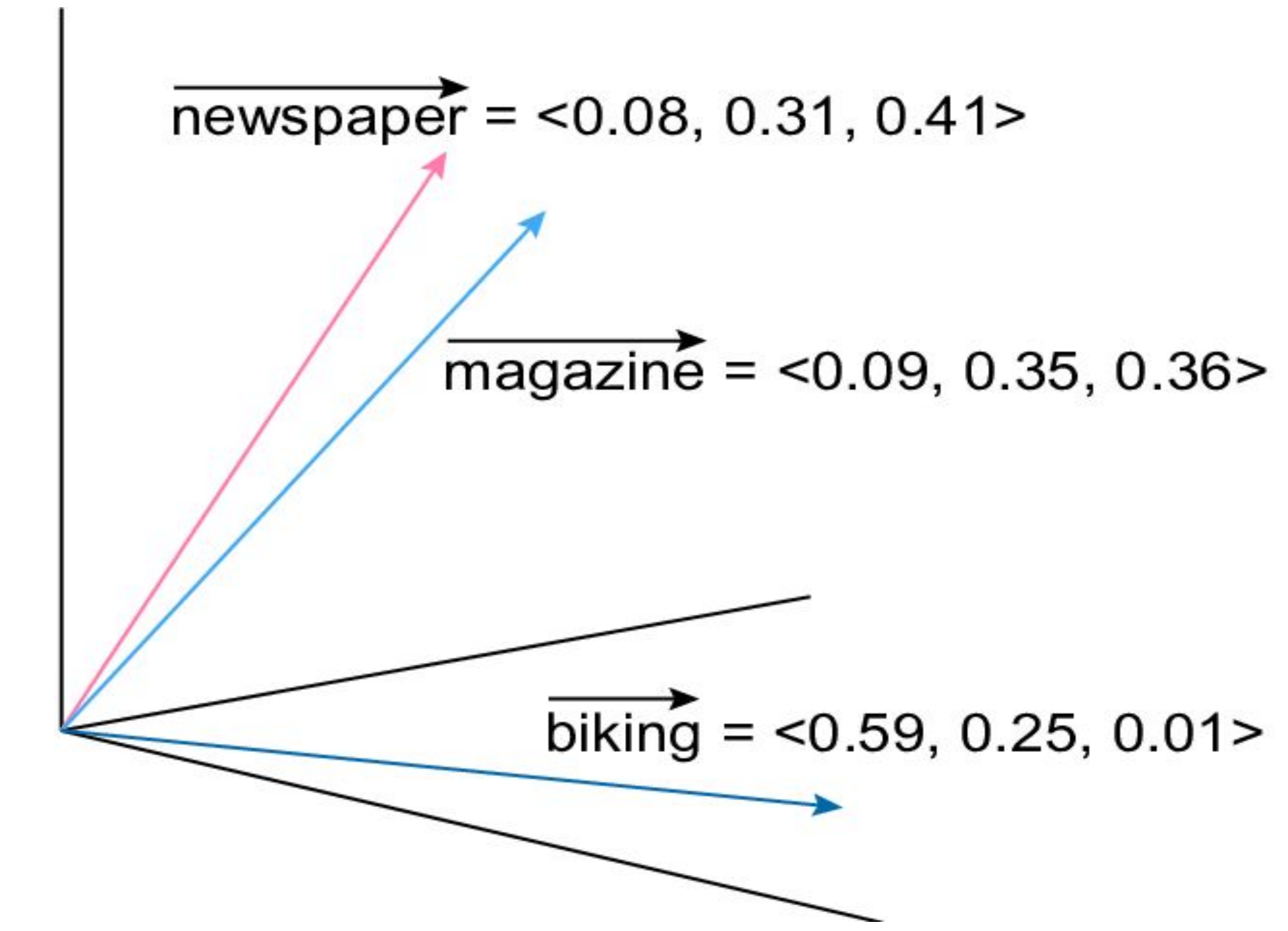
Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

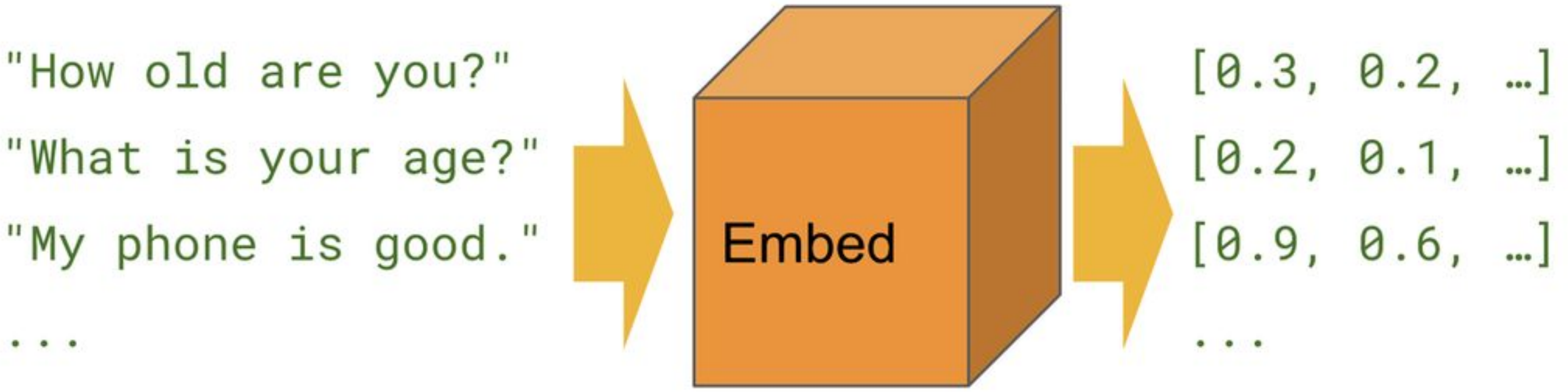
Session III, 19 Nov 2020

Recap last session

- Language Models
- Sentence Embeddings (Vector Space Model)
- Pipeline Part 1 of 2


$\overrightarrow{newspaper}$ = <0.08, 0.31, 0.41>
$\overrightarrow{magazine}$ = <0.09, 0.35, 0.36>
$\overrightarrow{biking}$ = <0.59, 0.25, 0.01>

http://mbenhaddou.com/2019/12/14/word2vec-concepts-from-scratch/



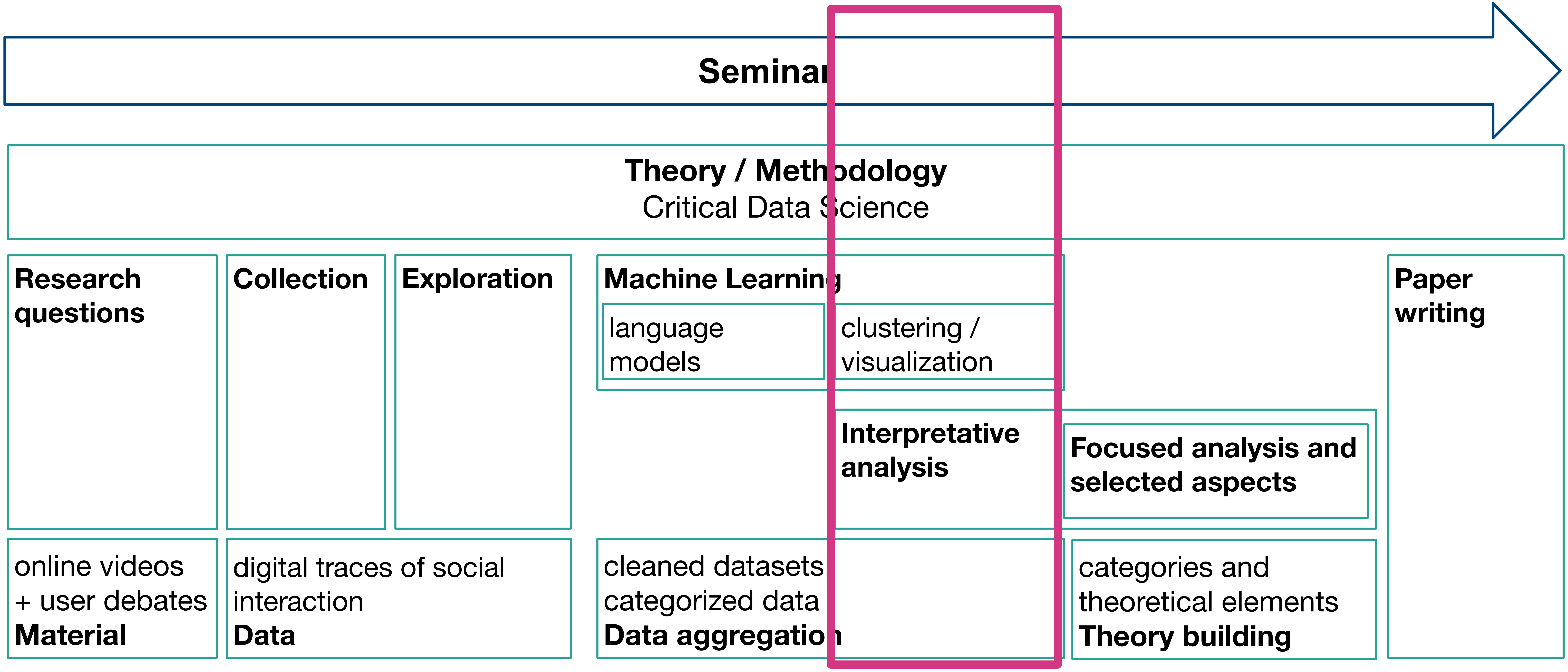
"How old are you?"
"What is your age?"
"My phone is good."
...

[0.3, 0.2, …]
[0.2, 0.1, …]
[0.9, 0.6, …]
...

Plan for today

- Discussion previous Assignment
- Cluster Analysis (Pipeline part 2)
- Guest Talk Dr. Kinkeldey: Visualizing High-dimensional Data
- (Short break: 5 minutes)
- Assignment for next week
- Collaborative collection of ideas and meeting peers

Seminar progress / today



Seminar

Theory / Methodology
Critical Data Science

Research questions

Collection

Exploration

Machine Learning

language models

clustering / visualization

Interpretative analysis

Focused analysis and selected aspects

Paper writing

online videos
+ user debates
Material

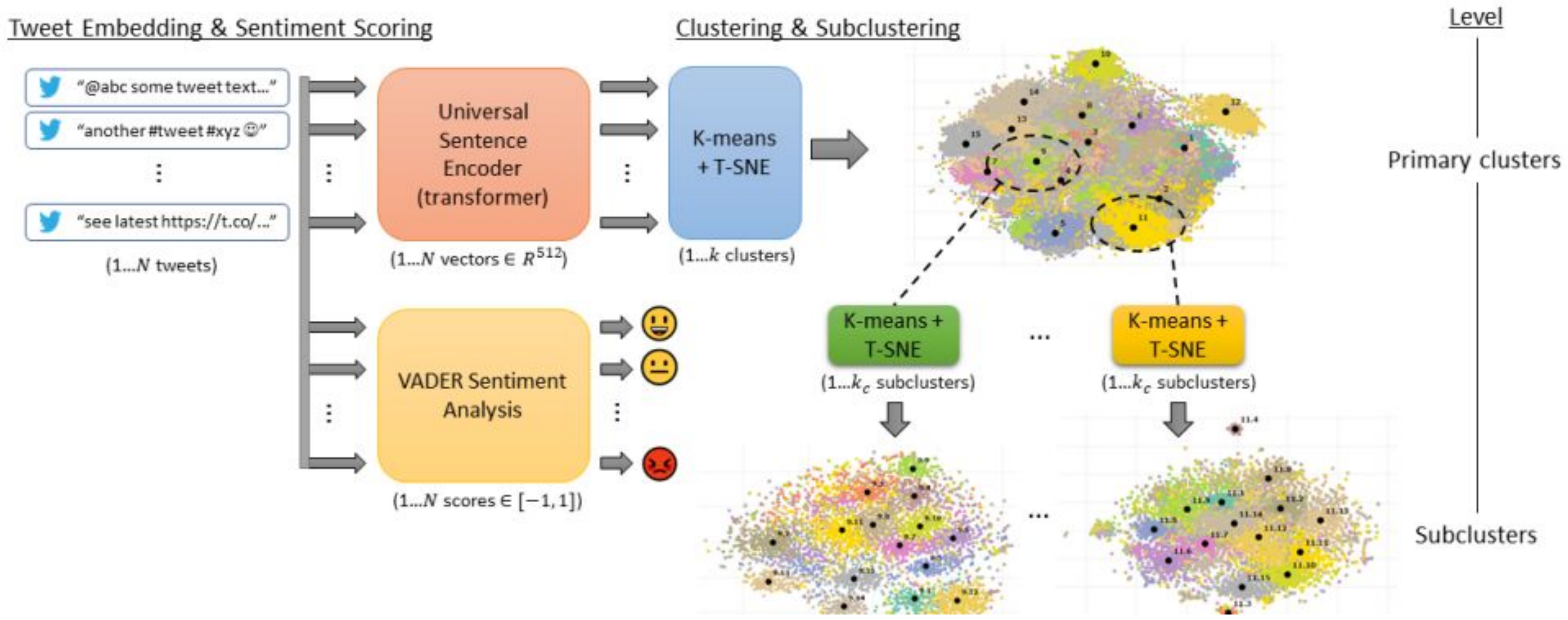
digital traces of social interaction
Data

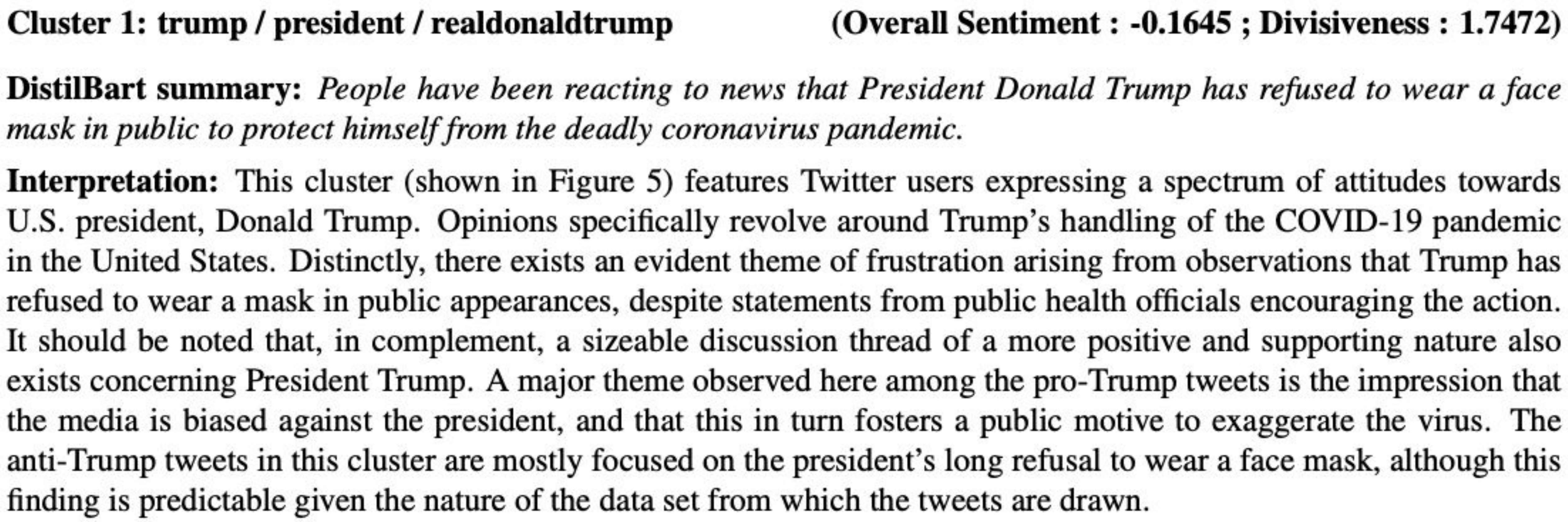
cleaned datasets
categorized data
Data aggregation

categories and theoretical elements
Theory building

Clustering - Example

Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse





Cluster 1: trump / president / realdonaldtrump **(Overall Sentiment : -0.1645 ; Divisiveness : 1.7472)**

DistilBart summary: *People have been reacting to news that President Donald Trump has refused to wear a face mask in public to protect himself from the deadly coronavirus pandemic.*

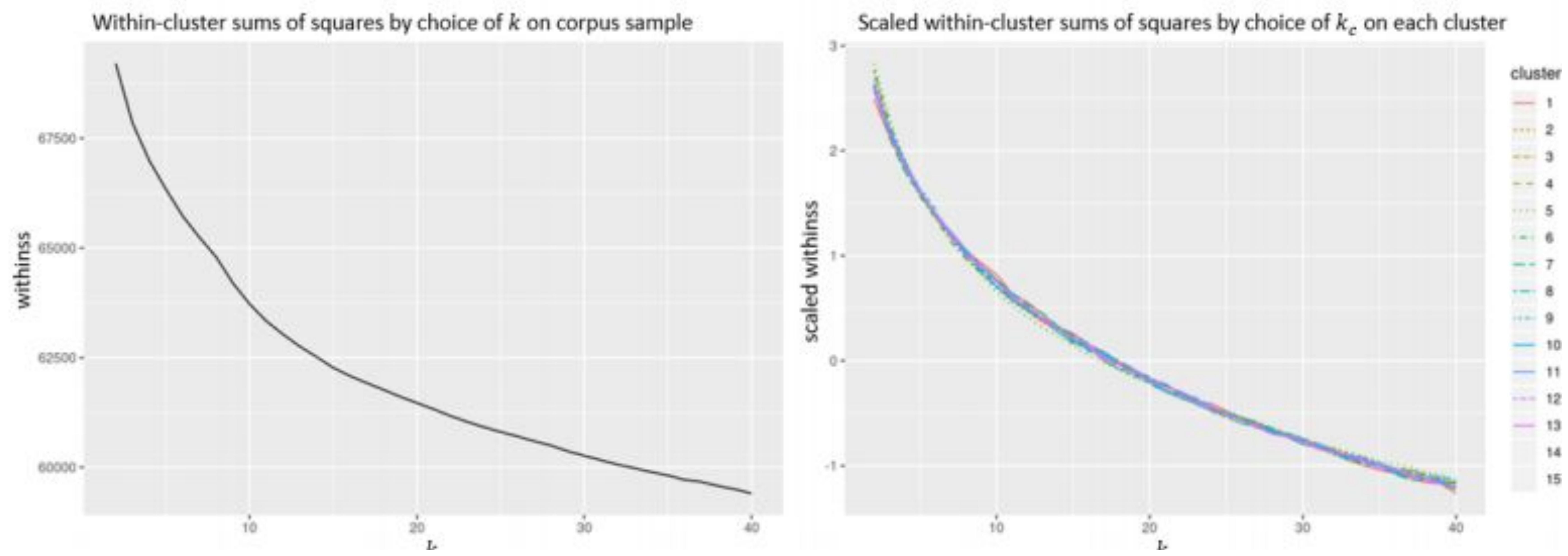
Interpretation: This cluster (shown in Figure 5) features Twitter users expressing a spectrum of attitudes towards U.S. president, Donald Trump. Opinions specifically revolve around Trump's handling of the COVID-19 pandemic in the United States. Distinctly, there exists an evident theme of frustration arising from observations that Trump has refused to wear a mask in public appearances, despite statements from public health officials encouraging the action. It should be noted that, in complement, a sizeable discussion thread of a more positive and supporting nature also exists concerning President Trump. A major theme observed here among the pro-Trump tweets is the impression that the media is biased against the president, and that this in turn fosters a public motive to exaggerate the virus. The anti-Trump tweets in this cluster are mostly focused on the president's long refusal to wear a face mask, although this finding is predictable given the nature of the data set from which the tweets are drawn.

Pipeline

Results

Sanders, Abraham, Rachael White, Lauren Severson, Rufeng Ma, Richard McQueen, Haniel Campos Alcanatara Paulo, Yucheng Zhang, John S Erickson, und Kristin P Bennett. „Unmasking the Conversation on Masks: Natural Language Processing for Topical Sentiment Analysis of COVID-19 Twitter Discourse". Preprint. Health Informatics, 1. September 2020. https://doi.org/10.1101/2020.08.28.20183863.

Clustering - Example

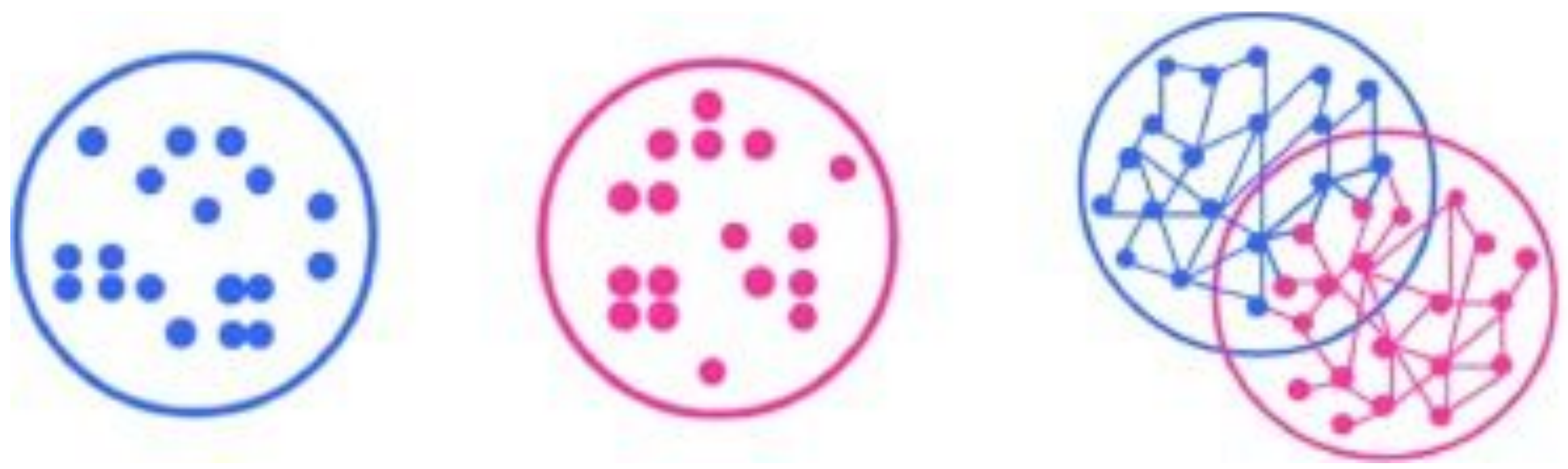


“To find suitable choices for k and k_c we use the **elbow method**, where the within-cluster sums of squares objective function is measured over a range of choices for k and k_c in an attempt to find the point which strikes a balance between minimization of the objective and avoiding over-clustering.”

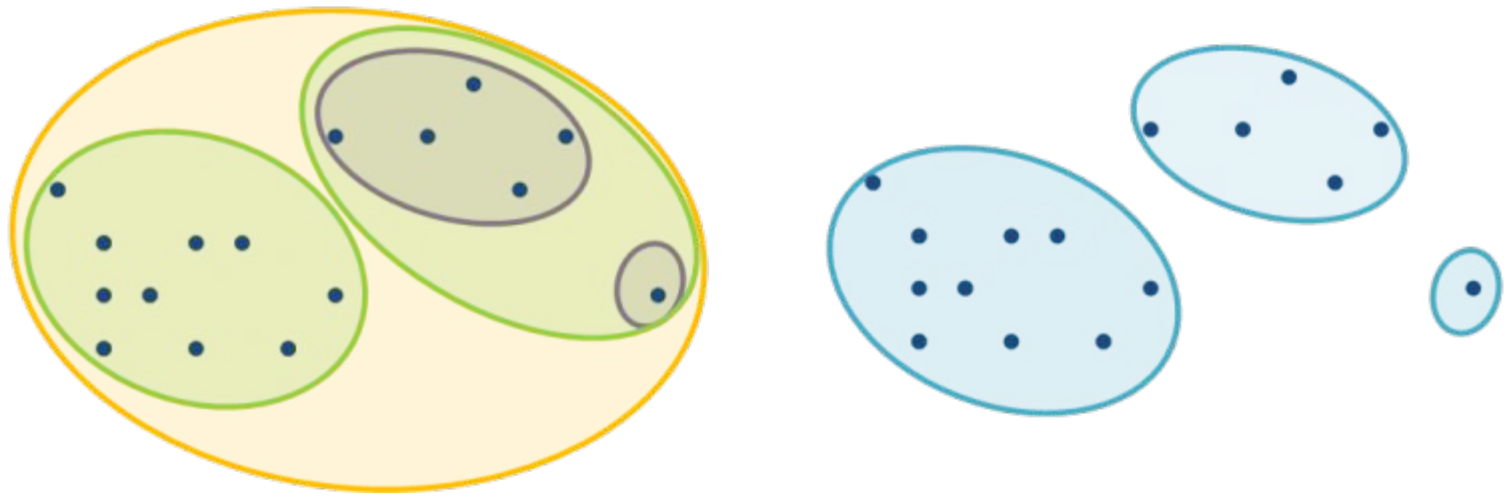
https://therensselaeridea.github.io/COVID-masks-nlp/paper_supplement.pdf

Sanders, Abraham, Rachael White, Lauren Severson, Rufeng Ma, Richard McQueen, Haniel Campos Alcanatara Paulo, Yucheng Zhang, John S Erickson, und Kristin P Bennett. „Unmasking the Conversation on Masks: Natural Language Processing for Topical Sentiment Analysis of COVID-19 Twitter Discourse“. Preprint. Health Informatics, 1. September 2020. https://doi.org/10.1101/2020.08.28.20183863.

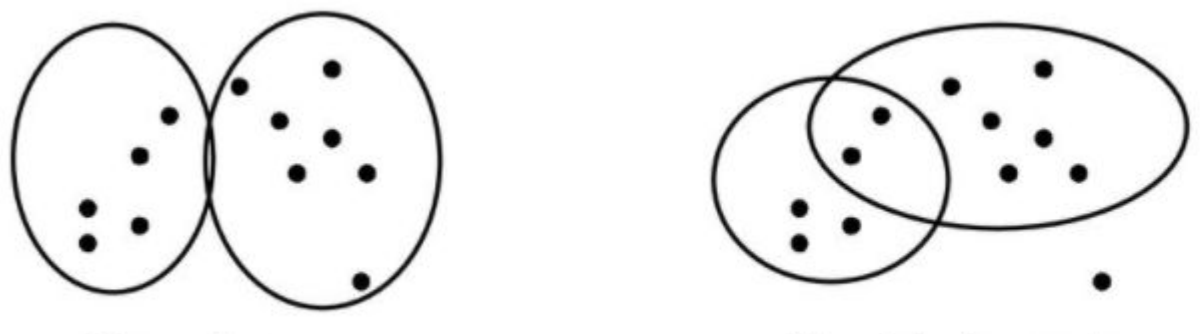
Clustering Overview of Methods



exclusive vs. overlapping



hierarchical vs. partitional



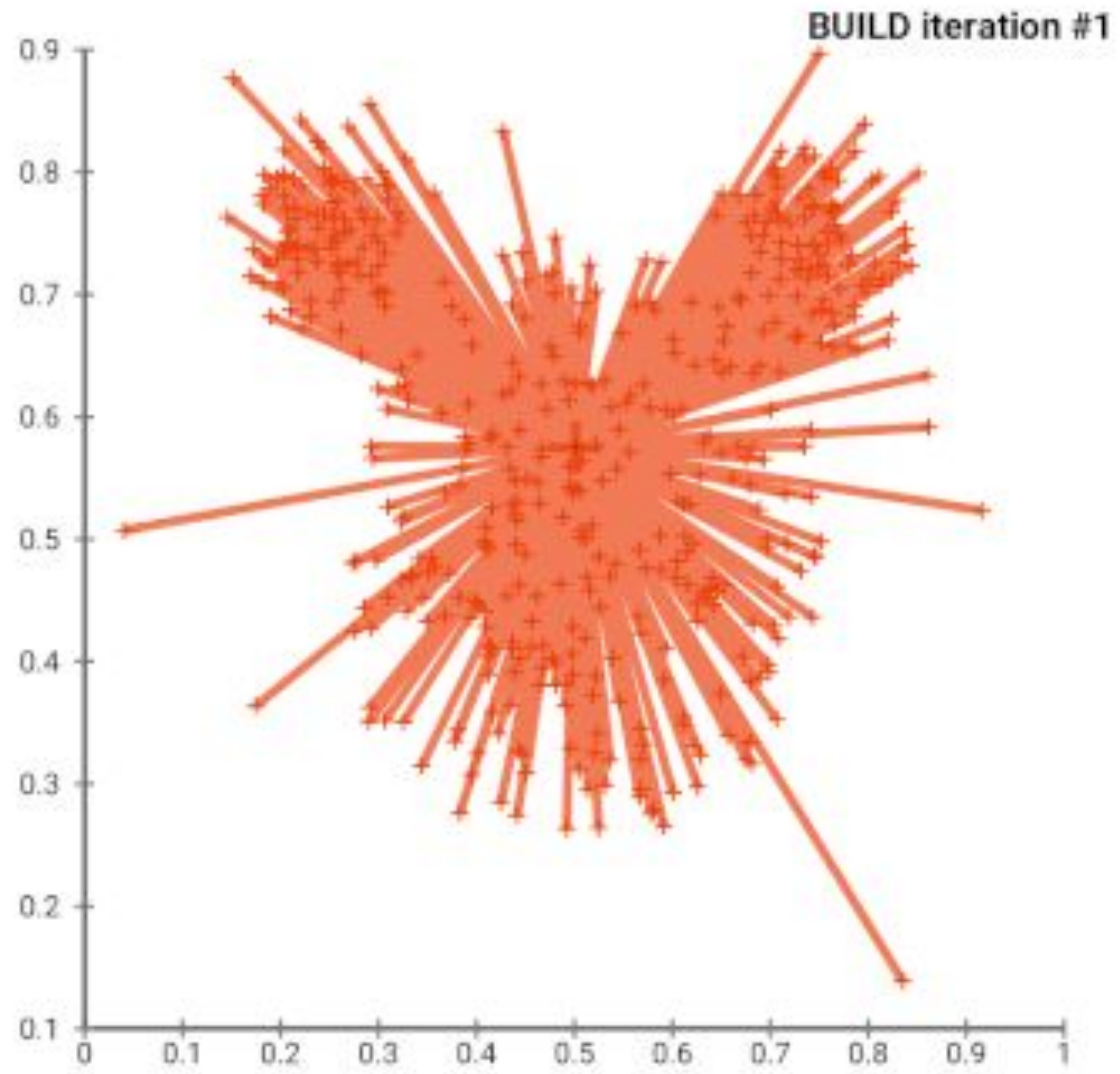
complete vs. partial

Algorithms:

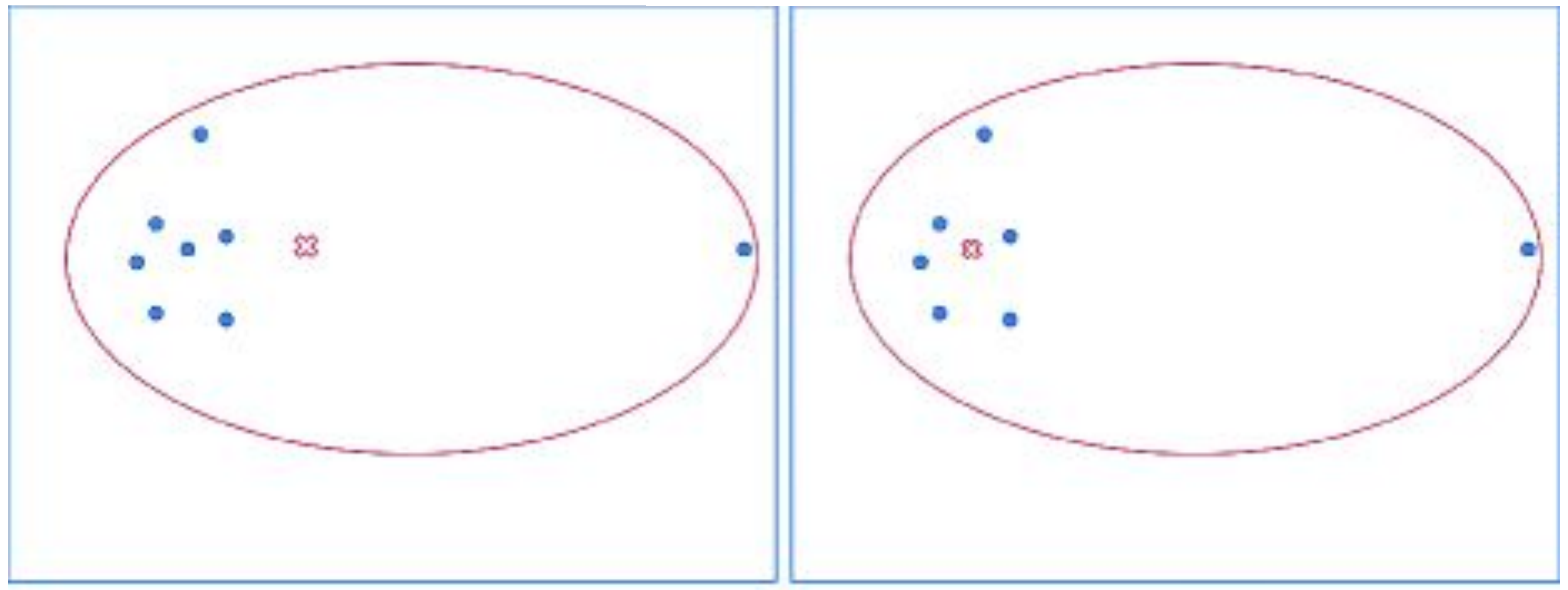
- graph-based (e.g. Affinity Propagation)
- density-based (e.g. DBSCAN)
- prototype-based (e.g. K-Means)
- ...

Tan, P., Steinbach, M., Karpatne, A., & Kumar, V. (2018). Introduction to Data Mining (2nd Edition).Chapter 7 (available online: https://www-users.cs.umn.edu/~kumar001/dmbook/ch7_clustering.pdf)

Clustering - K-Medoids



1. Initialize: Select *k* of the *n* data points as the medoids to minimize the cost
2. Associate each data point to the closest medoid.
3. While the cost of the configuration decreases:
    1. For each medoid *m*, and for each non-medoid data point *o*:
        1. Consider the swap of *m* and *o*, and compute the cost change
        2. If the cost change is the current best, remember this *m* and *o* combination
    2. Perform the best swap of m_best and o_best, if it decreases the cost function. Otherwise, the algorithm terminates.



(a) Mean

(b) Medoid

Pro: + Less sensitive to outliers than k-means;

+ can use cosine distance as metric

Con: - Number of clusters has to be defined;

- Assumes convex clusters (i.e. ‘round’)

Park, H.S.and Jun, C.H., 2009. A simple and fast algorithm for K-medoids

Demo: Pipeline part 2

Park, H.S.and Jun, C.H., 2009. A simple and fast algorithm for K-medoids

Visualizing High-dimensional data

Guest talk by Dr. Christoph Kinkeldey

Postdoctoral Researcher

at Human-centered Computing

Short break: 5 Minutes

Assignments for next week

1 Reading assignment

- Read Paper:
  - Baumer, Eric & Mimno, David & Guha, Shion & Quan, Emily & Gay, Geri. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?. Journal of the Association for Information Science and Technology. 68. 10.1002/asi.23786.
- Share one personal insight in a commentary of 150 words as a reply to this issue (e.g. an aspect you found interesting, a point you disagree, a perspective you want to explore further)

2 Cluster Analysis

- Create a backup of your Output of the last assignment.
- Download and setup the Jupyter notebook **Assignment_5** as described in our GitHub repository (https://github.com/FUB-HCC/seminar_critical-social-media-analysis)
- Load your preprocessed data and embeddings from the previous assignment.
- Optimize the number of clusters for k-medoids by maximizing the average silhouette score while minimizing the inertia for your data.
- Sample 2 clusters you deem interesting, print them and interpret them.
- Answer the following questions in a summary of ~150 words:
  - What is the content of the clusters? What is the quality of the clusters?
  - Would you suggest a purely quantitative approach to optimizing the clustering pipeline? Why or why not?
- Commit your Notebook with outputs to GitHub: create a new folder named [name]_assignment_session5 within the folder /Pipeline/Assignment_5
- Share your notebook URL in your assignment submission

Submit on Github (reply to issue) until 9 Dec 12h00 (noon)

Github issue for assignment: https://github.com/FUB-HCC/seminar_critical-social-media-analysis/issues/23

Collaborative brainstorming and meeting peers

Discussion on Discord voice channels

- Based on your preliminary research (assignments) and material (videos, post-video discussion data), discuss first project ideas to be implemented using the ML pipeline.
- Instructors will drop by and can give you feedback
  - You can also ask instructors to drop by on WebEx or Discord.

Seminar project groups

- You can enter group participant names and/or first ideas, elements here: (see GitHub)
  - Not a must at this stage, but it will help instructors to taylor future inputs to your project ideas and give further assistance.
  - Help finding peers for your project

Indications for research projects

- Indications about the research project and seminar paper can be found here: https://github.com/FUB-HCC/seminar_critical-social-media-analysis/issues/15
- If needed, you can also go back to the flinga.fi board from last session:

What’s up next session?

Combining perspectives of ML
and interpretative analysis

Recommended readings

Andy Coenen, Adam Pearce. Understanding UMAP | Google PAIR. https://pair-code.github.io/understanding-umap/

Tan, P., Steinbach, M., Karpatne, A., & Kumar, V. (2018). Introduction to Data Mining (2nd Edition). Chapter 7 (available online: https://www-users.cs.umn.edu/~kumar001/dmbook/ch7_clustering.pdf)

open space

Feel free to approach us in case of questions…
(microphone or chat)