The Summer Institutes in Computational Social Science

Bringing together graduate students, postdoctoral researchers, and junior faculty for 2 weeks of intensive study and interdisciplinary research

PROJECT FAIR

The Summer Institutes in Computational Social Science

SICSS 2024-Amsterdam

BEFORE WE START

We will present a few project ideas

You can work on something completely different

No pre-requisites on how your project should look like

Some guidance is possible (beginning of Week 2)

Present your findings on Friday Week 2

OVERVIEW OF PROJECTS

Roberto Cerina: ***Twitter US elections***

Rens Wilderom: ***IMDb data on movies***

Diliara Valeeva: ***Youtube data on conpiracies***

Johannes Aengenheyster: ***Art and music data***

Marije Peute: ***Instagram data***

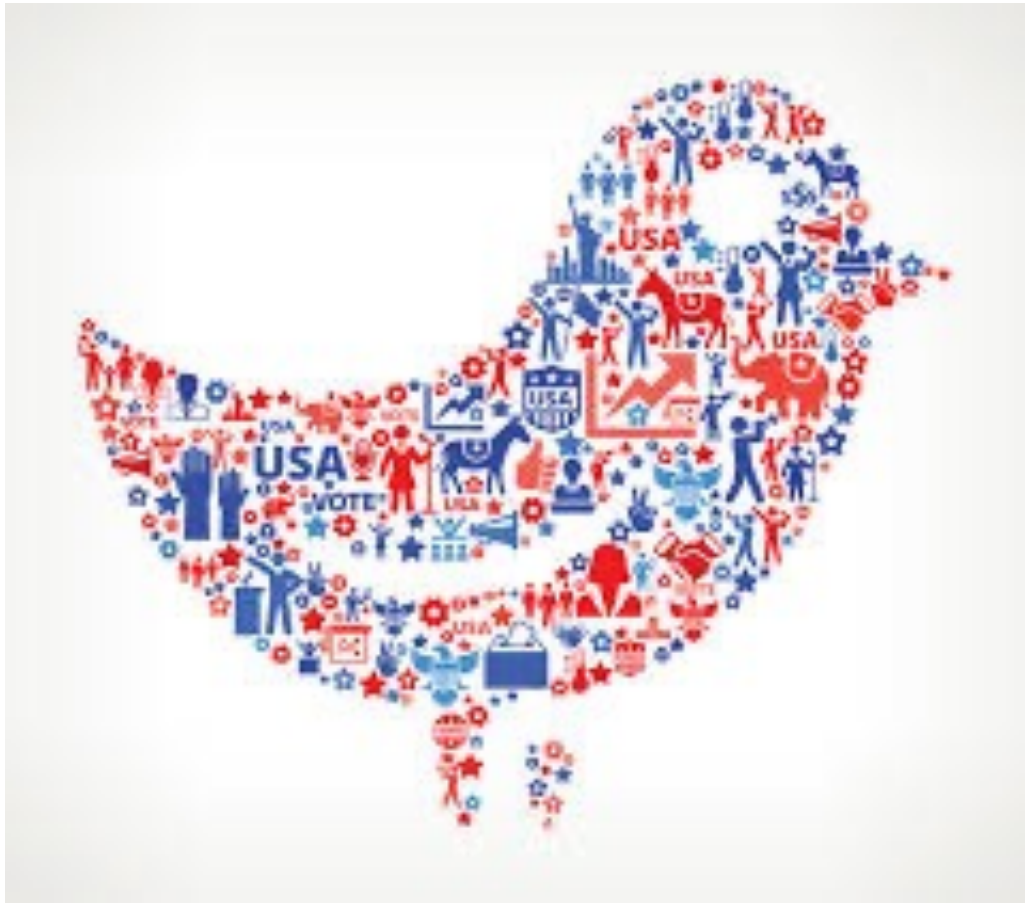
Malte Lukken & Gijs Schumacher: ***Federal debates data***

Daniel Mayerhoffer: ***UvA datasets***

+ Your projects?

TWITTER DATA ON US 2020 ELECTIONS

by Roberto Cerina



TWITTER DATA: US ELECTIONS + JAN 6

by Roberto Cerina

Data: Twitter API data around US 2020 Election + January 6th

Description: Tweets responding to query `Trump OR Biden' during the months around and including the 2020 US Election and Jan. 6th. Twitter API version 1.0

Link:
https://www.dropbox.com/scl/fo/mijwb0d5bz5q9ug55ac2a/AG2ylN53Nl5h3rjvM5XGTyE?rlkey=o eycvz1m5azurae3mnm67gzts&st=oyp6mbum&dl=0

Codebook: Details on the output can be found by navigating around this web-page https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview

Format: separate csv files, one for each day of the collection

TWITTER DATA: US ELECTIONS + JAN 6

Example Prediction Tasks

- Use a topic model to assign Tweets to a topic, and then predict the frequency of topics during a specific time-period (e.g. the 2020 Election campaign or the January 6th Event)

- Extract measures of candidate popularity, and use these in a model to `back-cast' the 2020 election, or predict the frequency of the measures themselves

- Train a classifier to detect instances of Fake news / Conspiracy Theories / Harmful content in general, and demonstrate its efficacy on the dataset.

TWITTER DATA: US ELECTIONS + JAN 6

Example Potential Inequalities:

- Ethically impermissible features of the data are used for prediction of harmful content

- The political leaning of the platform is such that any election prediction is inherently biased in favour of one side

- The demographic composition of Twitter is such that the topics extracted with topic modeling only reflect a selected subset of the population of interest, and certain topics' importance is enlarged relative to their prevalence in the `real world'.

IMDB DATA ON MOVIES

by Rens Wilderom

YOUTUBE DATA ON CONSPIRACY THEORIES

by Diliara Valeeva

YOUTUBE DATA ON CONSPIRACY THEORIES

- Popular videos (> 0.5 mln views) discussing conspiracy theories on elites/politicians/those who hold power. Keyword search e.g. 'freemasons', 'illuminati'
- Around 55 videos, only English-language videos
- Used Youtube API in 2021 to scrape the comments

Data: 1) video properties; 2) comments and their properties.

Example comments:

Juhotuho10: "Illuminati no longer exists", that's what someone connected to the Illuminati would say!!
Haslan: - "Are the Illuminati real?" - "No, we aren't"
Chandra Satrio That's sounds like someone from Illuminati would say. Ted: nervous laugh*
NPC 33331: This video brought to you by the Illuminati

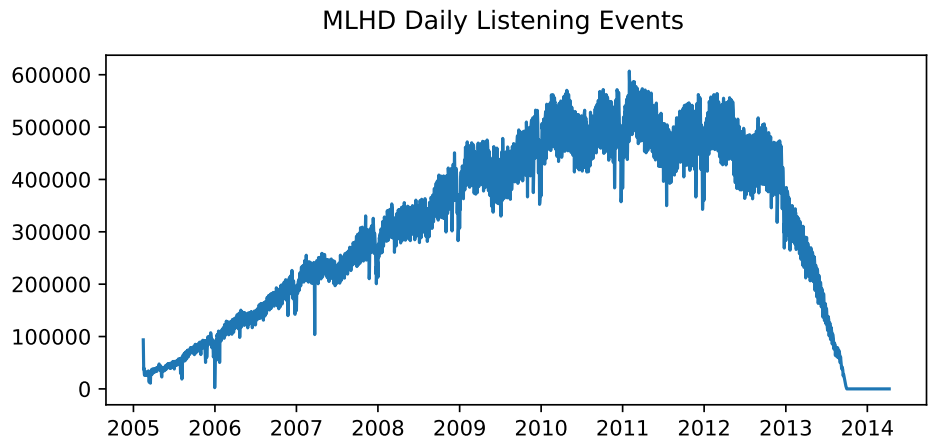
ART AND MUSIC DATA

by Johannes Aengenheyster

Music

combination of

- last.fm user-level listening events (MLHD)
- song acoustic features from AcousticBrainz
- song tags from last.fm API



dataschema

listening event logs: 1b downloaded (US; full 27b)

- time stamp
- user id (27k unique)
- song id (4.1m unique)

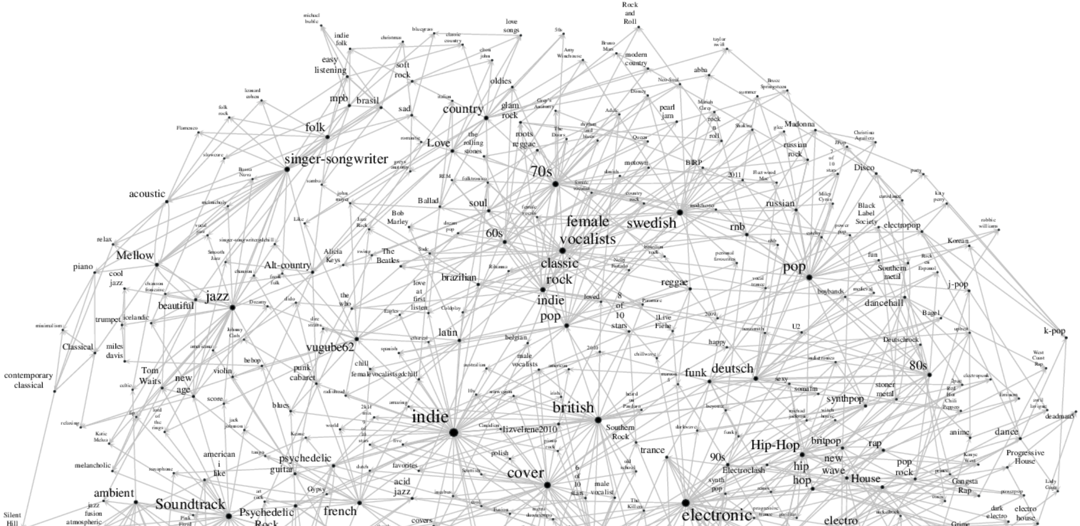
acoustic features: (1.3m)

- song id
- song name
- artist name
- acoustic features (timbre, tonality, danceability, mood (happy, sad, relaxed, party))

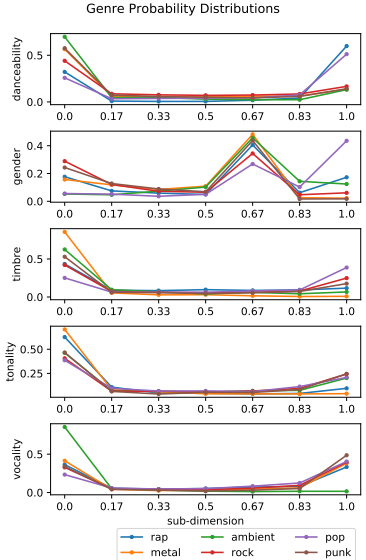
tags: 28m

- song id (1.8m unique)
- tag (885k unique, user-provided)
- weight

tags


Rock and Roll
indie folk
love songs
easy listening
soft rock
oldies
modern country
abba
mpb
brasil
sad
country
glam rock
pearl jam
folk
Madonna
Love
roots reggae
russian rock
singer-songwriter
70s
Disco
acoustic
Ballad
soul
russian
Bob Marley
female vocalists
swedish
rnb
Black Label Society
electropop
relax
60s
Mellow
classic rock
pop
piano
Alicia Keys
Alt-country
The Beatles
brazilian
cool jazz
love at first listen
reggae
indie pop
j-pop
jazz
beautiful
dancehall
trumpet
icelandic
latin
k-pop
Classical
miles davis
vugube62
80s
contemporary classical
Tom Waits
new age
violin
punk cabaret
funk
deutsch
synthpop
score
blues
indie
british
Southern Rock
trance
anime
dance
Hip-Hop
britpop
rap
Progressive House
american i like
psychedelic
guitar
90s
new wave
hip hop
cover
House
pop rock
Gangsta Rap
melancholic
ambient
Soundtrack
Psychedelic Rock
french
acid jazz
electronic
electro
Silent Hill

acoustic information per tag



Art

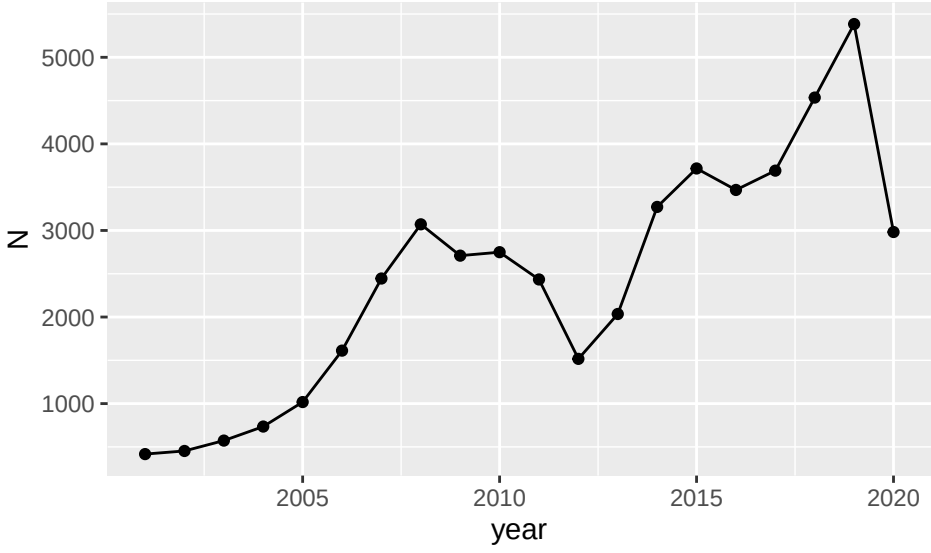
exhibition of Chinese artists, scraped from ArtLinkArt
info:

- artist: id, name (50k)
- exhibition: id, name, date (50k)
- space: id, name, location, category (gallery, museum, independent) (10k)

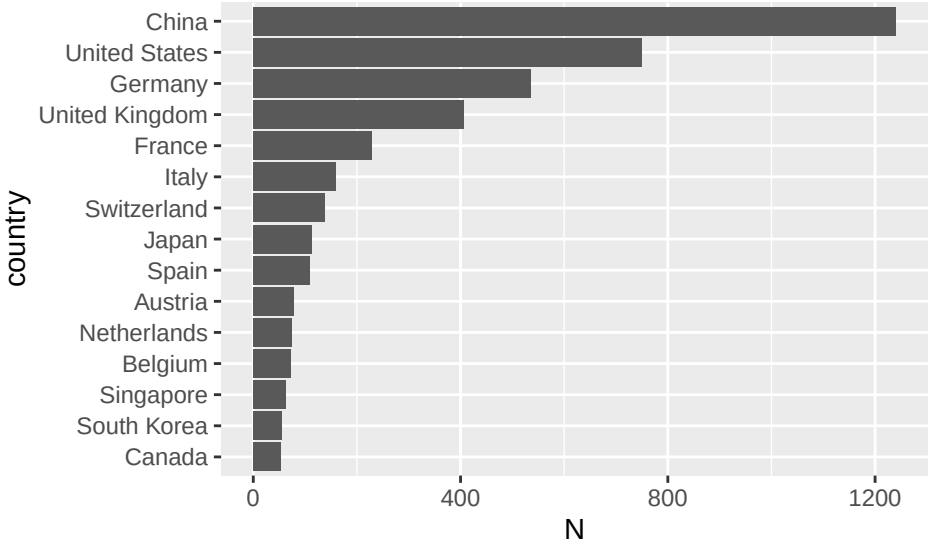
links:

- artist-exhibition: 156k
- exhibition-space: 58k

exhibitions



locations


China
United States
Germany
United Kingdom
France
Italy
Switzerland
Japan
Spain
Austria
Netherlands
Belgium
Singapore
South Korea
Canada
country
0
400
800
1200
N

INSTAGRAM DATA

by Marije Peute

GERMAN FEDERAL DEBATES DATA

by Malte Lukken & Gijs Schumacher

GERMAN FEDERAL DEBATES DATASET

- Available on Hugging Face Hub: https://tinyurl.com/german-debates
- Output from MEXCA pipeline for 12 federal election debates from the years 2002 to 2021
- Stored as 2 parquet files
- Has "filename" column with debate identifier
- Variables are documented in documentation of mexca package
- Notebook for efficient loading and processing in Python: https://tinyurl.com/data-loader-python
- Script for efficient loading and processing in R: https://tinyurl.com/data-loader-r

GERMAN FEDERAL DEBATES DATASET

Things to keep in mind:

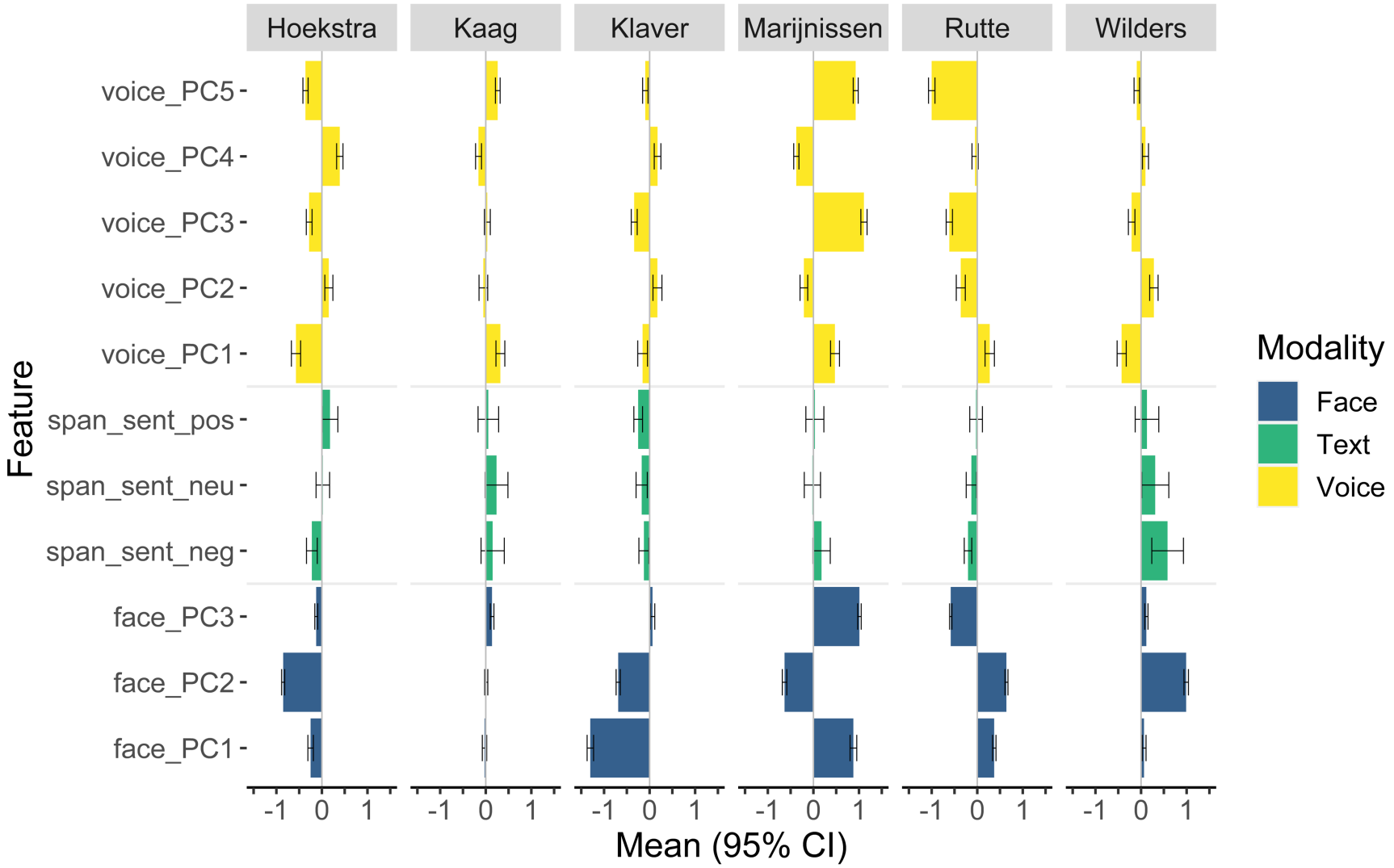
- Columns “face_label” and “segment_speaker_label” refer to different persons in different debates
- Some facial action units are predicted with low accuracy
- Many features are not normally distributed → transformations
- Lots of missing data!

GERMAN FEDERAL DEBATES DATASET

Suggestions for analyses:

- Look at structure of emotion expressions → factor analysis, dimensionality reduction
- Look at differences between speakers
- Look at interactions between speakers
- Look at temporal dynamics

EXAMPLE ANALYSIS



UVA DATASETS

by Daniel Mayerhoffer

Computational
Social Science

Databases UvA

SICSS Amsterdam

DM Mayerhoffer

June 13, 2024

University of Amsterdam

Retrieving Data from Existing Databases

Advantages

- Easy to get - no messy work to collect data
- Often tidy and high-quality
  - Only minimal prepossessing necessary
  - Easy to work with and merge
  - Documentation makes them easy to understand
- Examples of how to work with the data out there

Be careful!

- Don't forget to solve a research puzzle:
  - Don't let the data (availability) guide your research.
  - Check whether your research project is indeed novel.
- Note the **Often** above...

Database Access through the UvA Library

Check out the databases licensed by UvA **here**!

- Text corpora
- Corporate and economic data
- Population statistics
- Bibliometric data
- ...

How I can help

- Provide directions and advice.
- Do quick search queries for you.

NOT Manually retrieve large chunks of data.

Work with the Scopus Database

Potential Research Puzzles

- Evolution of (interdisciplinary) research fields or paradigms
- Importance of scholars, institutions or journals
- Collaboration structures
- (Co-) Citation structures
- Topics of titles, keywords and abstracts
- Differences in region or gender

YOUR PROJECT IDEAS

NEXT STEPS

CHOOSE YOUR PROJECT/GROUPMATES

NEXT STEPS

From tomorrow onwards, we will have 'Group work' timeslots in the schedule. We have two rooms booked for you. Lunch and snacks will be provided.

Except Monday: we have one workshop and one keynote

+ ***Tuesday***: office hours with the keynote Elisa Omodei to discuss your group work or indvidual projects

+ ***Thursday***: Lecture on data revolution by the AISSR (with drinks afterwards)

Also: Tomorrow during the lunch: Campus Tour by Aya and Leo (join if you want)