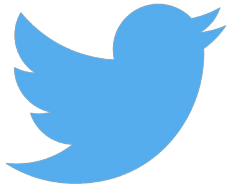
Alexander Bartel
Institute for Veterinary Epidemiology and Biostatistics

Freie Universität Berlin



Using Twitter for Research

Alexander Bartel

Outline

1. Social Media Research and Concepts
2. Why Twitter?
    1. How to access Twitter?
    2. What information can be obtained?
3. What to do with Twitter Data?
    1. Geocoding
    2. Sentiment Analysis
    3. Network Analysis

Social Media Research

Data Void

Criteria for successful topics

1. newsworthiness
2. excitement
3. provoking and confrontational

=> Engagement

Topics which are common knowledge or boring are not discussed.

Problems

- Huge age and gender bias
  - Age depending on platform (Twitter mostly millennials)
  - Gender bias mostly due to topics discussed
- Small proportion of highly active users (Source: twopcharts)
  - 44% percent of Twitter users never tweeted
  - 13% of registered accounts have tweeted more than 100 times

Why Twitter?

- Generous data access rules
  - Only social network which provides full access to all tweets
  - Facebook was only accessible for commercial partners
    - Scientific access was limited to randomly sampled content
    - Since Cambridge Analytica Scandal, access is even more limited

Why Twitter?

- Character limit of 280 (before 140)
  - Leads to simplified wording
  - Less nuance
  - Higher polarization

=> Easier to classify for algorithms

How to access twitter?

Search endpoint

- Limited history to 7 days
- Language and locale filtering

Stream endpoint

- Live filtering according to search words

Timeline endpoint

- Get all tweets of people you follow

Lookup endpoint

- Get tweets by id

Endpoint	Rate per 15 min
GET search/tweets	180
POST statuses/filter	1% of all Tweets ~60 or more
GET statuses/user_timeline	900
GET statuses/lookup	900

Apply for access here:
https://developer.twitter.com

How to access twitter?

Python:

- Tweepy: Good stream management
  (Pausing when rate limit is reached)
- Python-twitter: Similar functionality, integrates with django

R:

- twitteR: not updated since 2016
  but simpler usage and integrated database management

The Tweet Object

Tweets are supplied as JSON


```
{
"created_at" : "Thu May 10 15:24:15 +0000 2018" ,
 "id_str" : "850006245121695744" ,
 "text" : "Here is the Tweet message." ,
 "user" : {
 } ,
 "place" : {
 } ,
 "entities" : {
 } ,
 "extended_entities" : {
 }
}
```

The Tweet Object

Tweets are supplied as JSON

The Tweet Object

Tweets are supplied as JSON

Full User profile:
- Name
- Location (user specified)
- Profile Text
- Image Link

GPS Location
- Rarely enabled


```
{
"created_at" : "Thu May 10 15:24:15 +0000 2018" ,
 "id_str" : "850006245121695744" ,
 "text" : "Here is the Tweet message." ,
"user" : {
 } ,
 "place" : {
 } ,
 "entities" : {
 } ,
 "extended_entities" : {
 }
}
```

The Tweet Object

Tweets are supplied as JSON

Hashtags
Including start and end position

Also:

- If retweet or comment the whole cited tweet is included

Respect Privacy Protections

Severe Restrictions apply to:

- Sensitive information
- Off-Twitter matching
- Redistribution of Twitter content
- Surveillance, privacy, and user protection

https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases.html

Sensitive Information

You should be careful about using Twitter data to derive or infer potentially sensitive characteristics about Twitter users. Never derive or infer, or store derived or inferred, information about a Twitter user's:

- Health (including pregnancy)
- Negative financial status or condition
- Political affiliation or beliefs
- Racial or ethnic origin
- Religious or philosophical affiliation or beliefs
- Sex life or sexual orientation
- Trade union membership
- Alleged or actual commission of a crime

Sensitive Information

Aggregate analysis of Twitter content that does not store any personal data (for example, user IDs, usernames, and other identifiers) is permitted, provided that the analysis also complies with applicable laws and all parts of the Developer Agreement and Policy.

ANALYSIS

Geocoding

- GPS data rarely available
- ~60% of users fill out the location field in the user profile

- “Berlin” => lat 52.519444° long 13.406667°

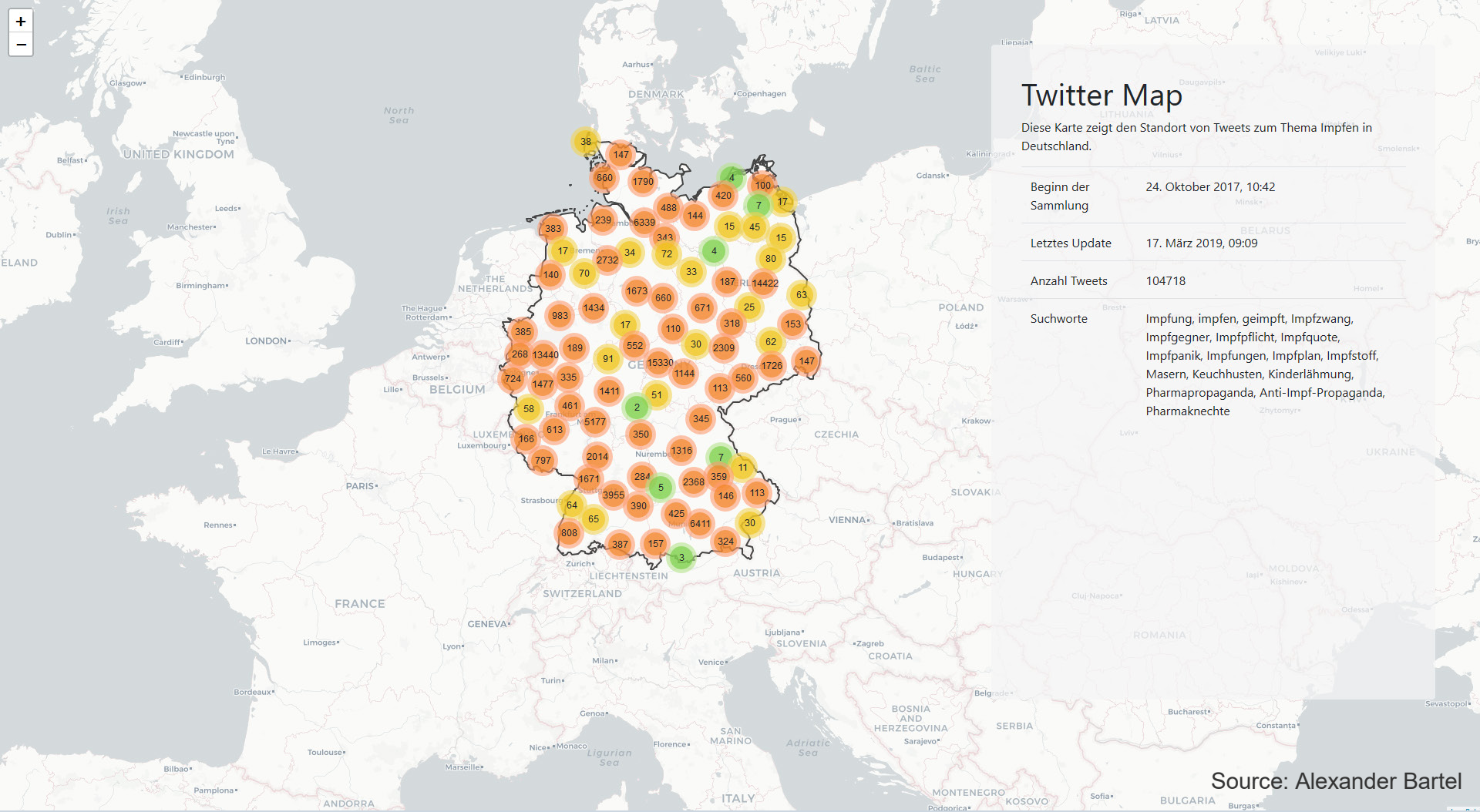
Available Services:

Google Maps API: free for up 200$ per month

ArcGIS API: free for 0.5-1 request per second (automatic slowdown)

OpenStreetMaps: self-hosted **DataScienceToolkit** for full privacy

or Nominatim API (multiple providers)


Twitter Map
Diese Karte zeigt den Standort von Tweets zum Thema Impfen in Deutschland.
Beginn der Sammlung
24. Oktober 2017, 10:42
Letztes Update
17. März 2019, 09:09
Anzahl Tweets
104718
Suchworte
Impfung, impfen, geimpft, Impfzwang, Impfgegner, Impfpflicht, Impfquote, Impfpanik, Impfungen, Impfplan, Impfstoff, Masern, Keuchhusten, Kinderlähmung, Pharmapropaganda, Anti-Impf-Propaganda, Pharmaknechte
Source: Alexander Bartel

Sentiment Analysis

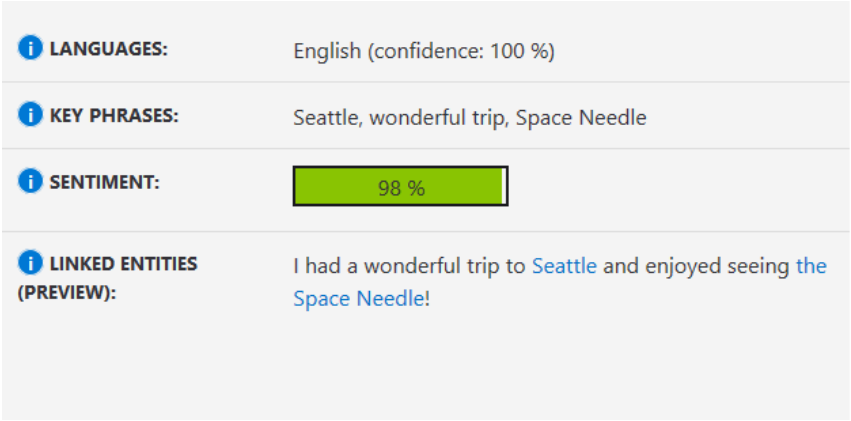
Standard approach:

- Calculate an overall score for how positive or negative the text is
- Implementation mostly language specific

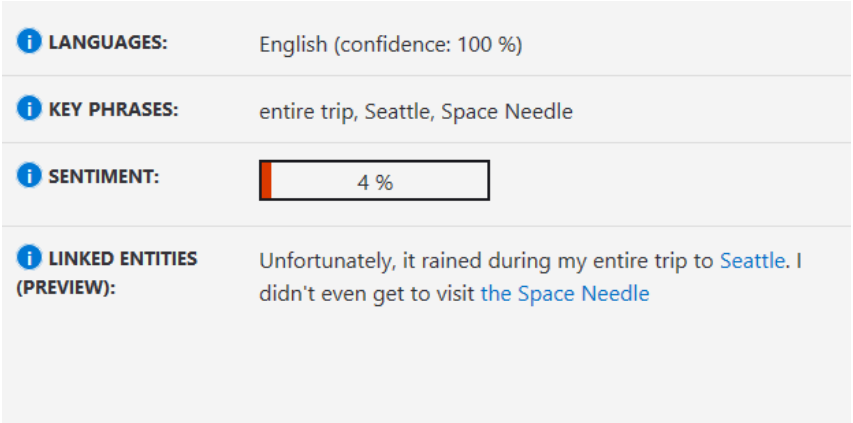
Services:

- Polyglot (40+ Languages), local, open-source, privacy friendly
  - http://polyglot-nlp.com
- Microsoft Azure Text Analytics API
  - https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/

I had a wonderful trip to Seattle and enjoyed seeing the Space Needle!



Unfortunately, it rained during my entire trip to Seattle. I didn't even get to visit the Space Needle



https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/

Neural Net Tweet tagging

- Tag a subsample of tweets by hand

- (Optional) Preprocessing of Tweets with NLP Librarys
  - Standford CoreNLP
  - UDPipe for R

- Train Neural Net to classify Tweets

Neural Net Tweet tagging

- Tag a subsample of tweets by hand

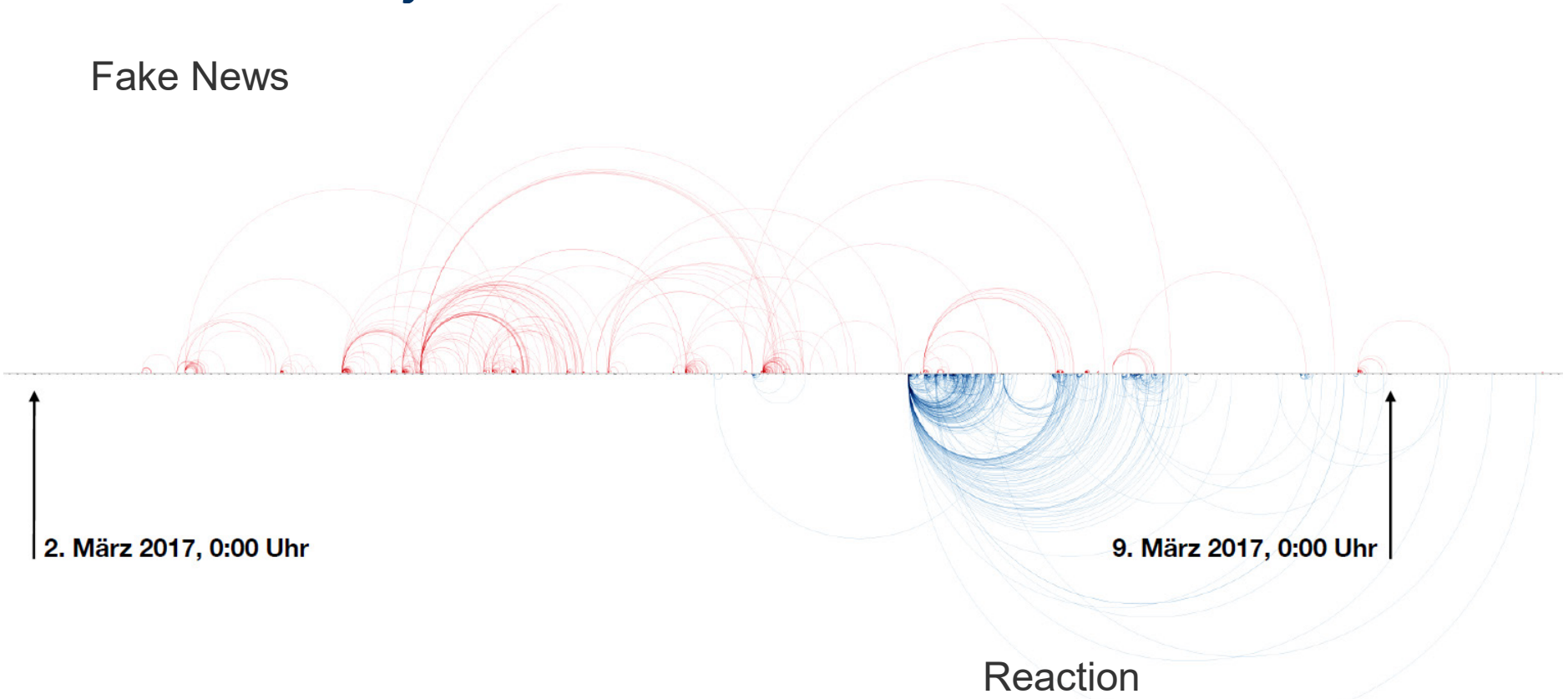
Tag all Tweets by hand and be done with it

- Student classifier:

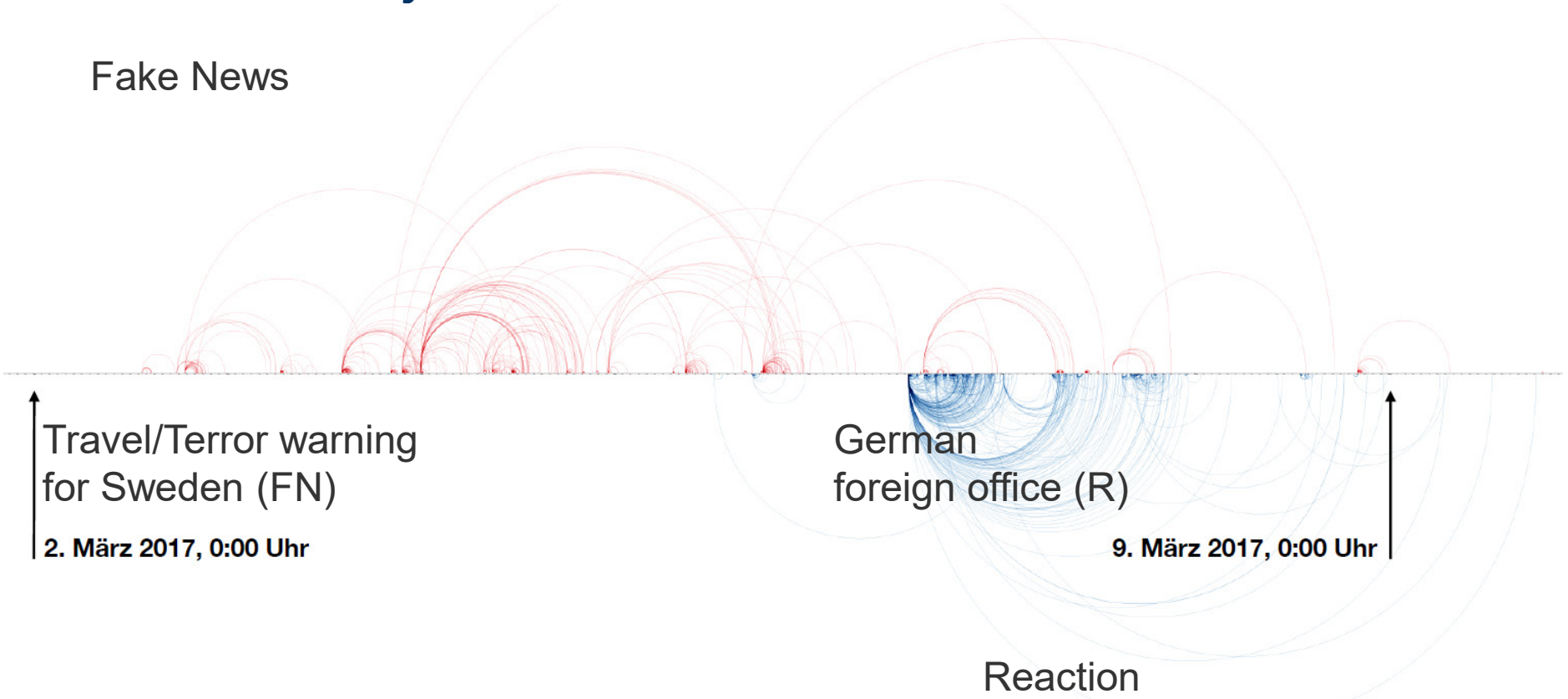
  Every student will have a better classification quality than any of the presented automatic methods.

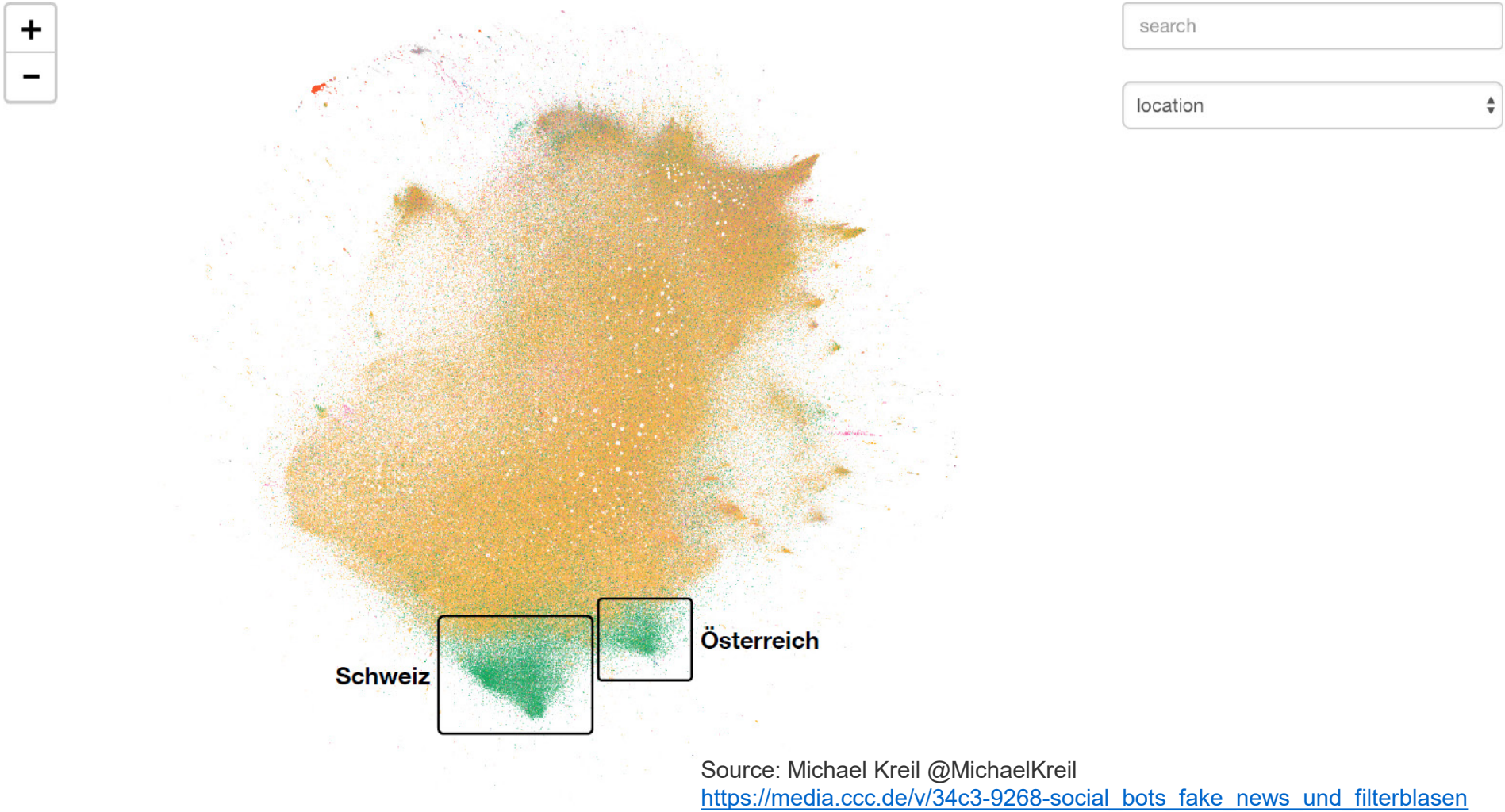
- ~10% double classification to calculate Inter-rater reliability

Network Analysis (Source: Michael Kreil @MichaelKreil)

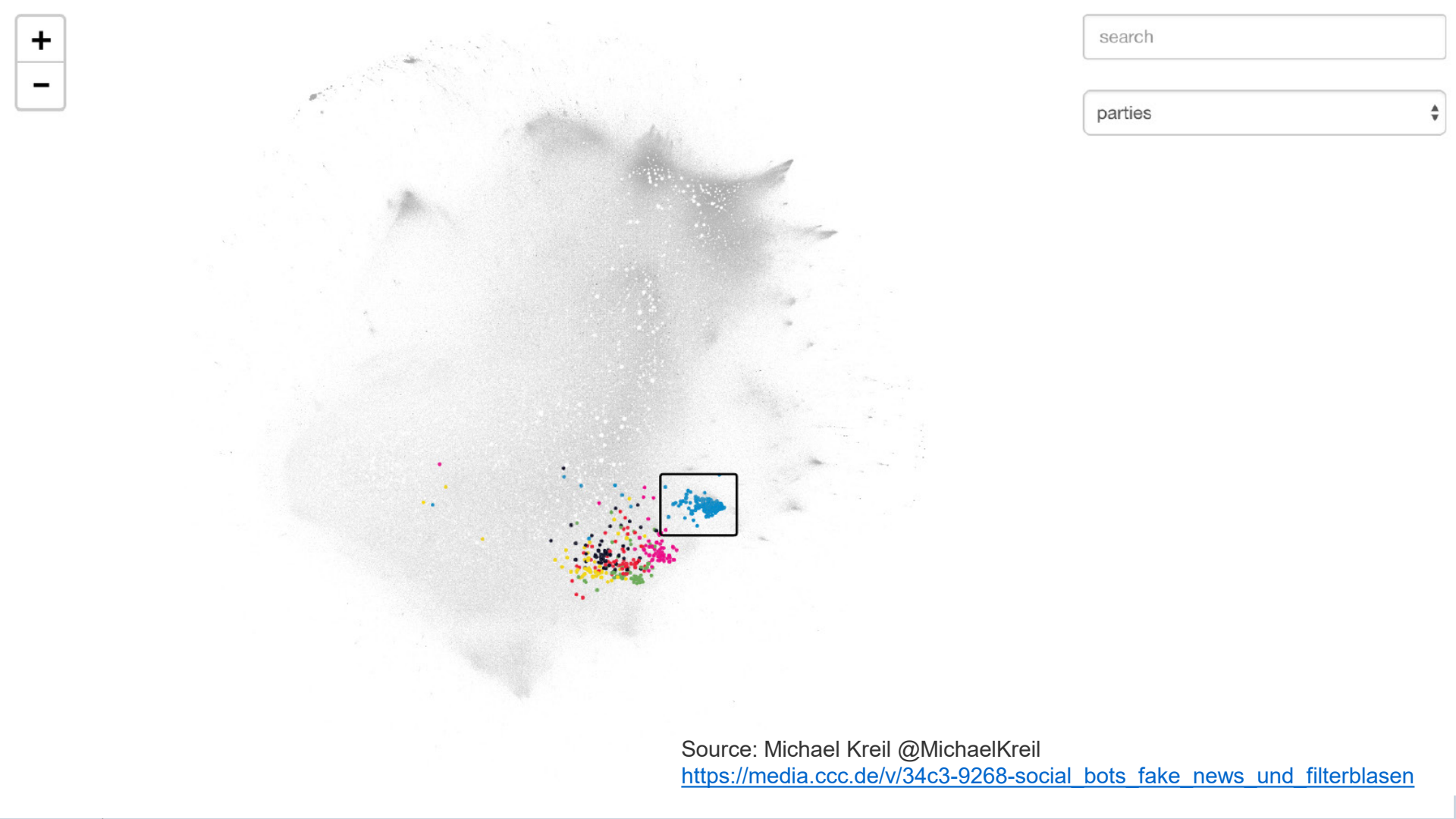


Network Analysis (Source: Michael Kreil @MichaelKreil)





Source: Michael Kreil @MichaelKreil
https://media.ccc.de/v/34c3-9268-social_bots_fake_news_und_filterblasen



Source: Michael Kreil @MichaelKreil
https://media.ccc.de/v/34c3-9268-social_bots_fake_news_und_filterblasen

Network Analysis

- Useful for explorative research
- Identify closed subgroups (filter bubble)
- Only low resolution evaluation possible to protect individual users

https://alexander-bartel.de/twitter

THANK YOU