Working with Metadata

Tatjana Scheffler
Universität Potsdam
tatjana.scheffler@uni-potsdam.de
@tschfflr

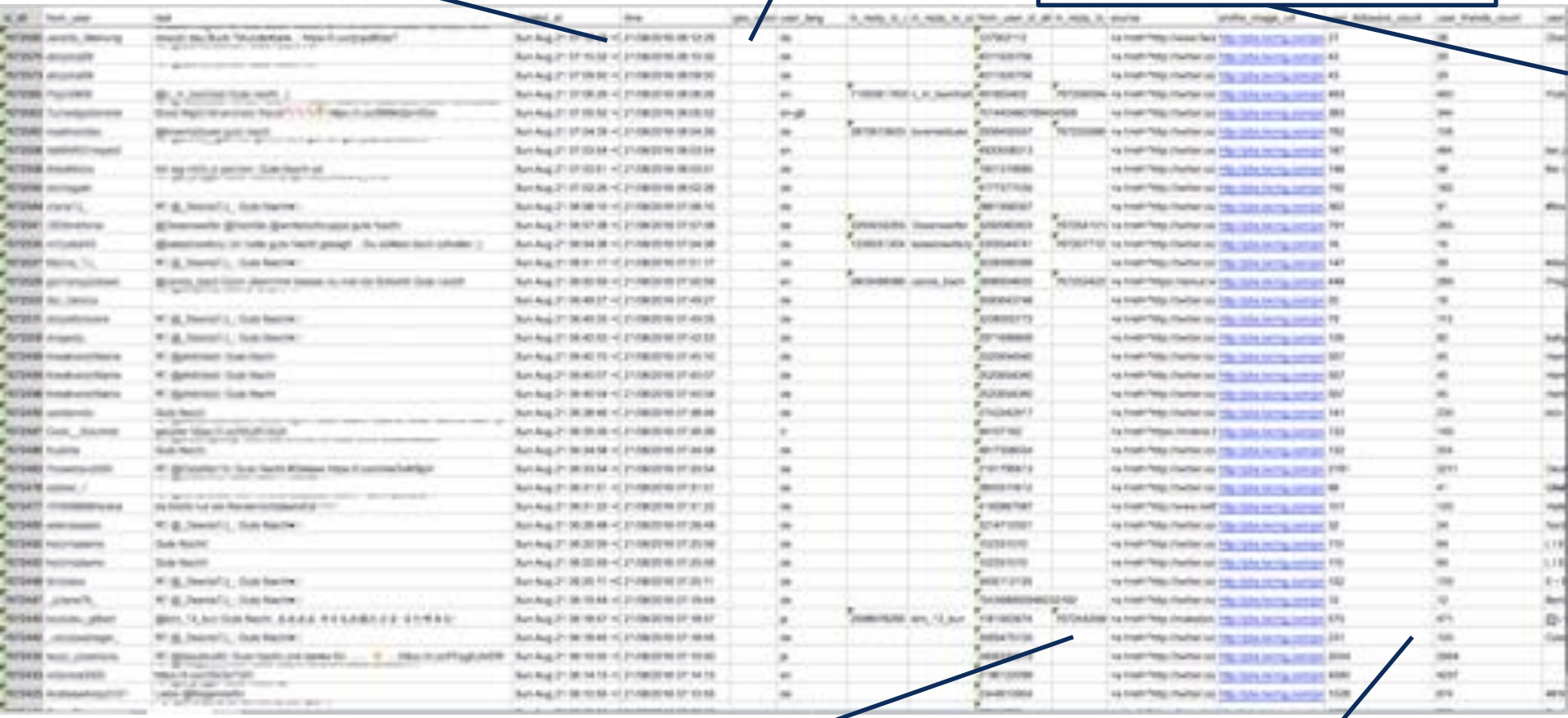
Computational Sociolingustics

Discovering properties of the author in tweets

time

geo_coordinates

user profile info



in_reply_to

user network

Metadata

1. geographic coordinates
2. time stamps
3. reply info
4. user network
   - followers
   - friends
   - retweets
   - mentions
5. user profile information

Geolocation of Tweets

Joint work with Johannes Gontrum

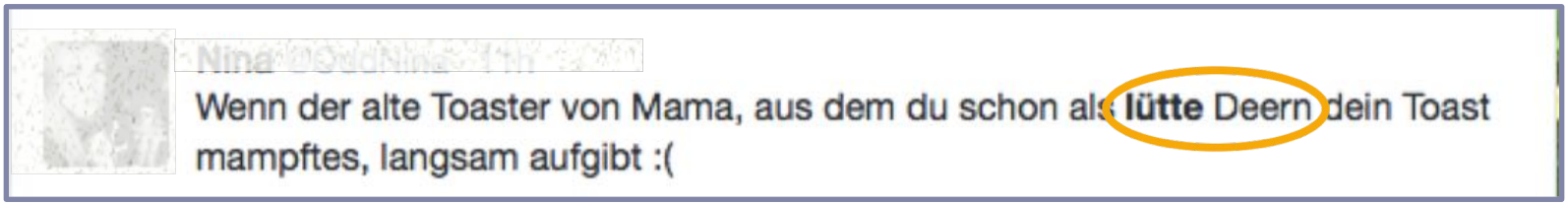
Recovering location

Johannes Gontrum and Tatjana Scheffler. Text-based Geolocation of German Tweets. 2015.

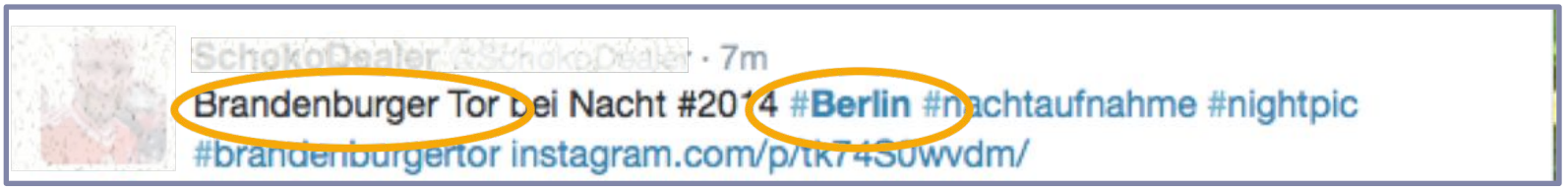
Map: „Niederdeutsche Dialekte seit 1945 BRD“ von Et Mikkel/Dudy001 - Transferred from de.wikipedia; transferred to Commons by User:Saibo using CommonsHelper.

Regional influences on tweets

1. Dialect origin

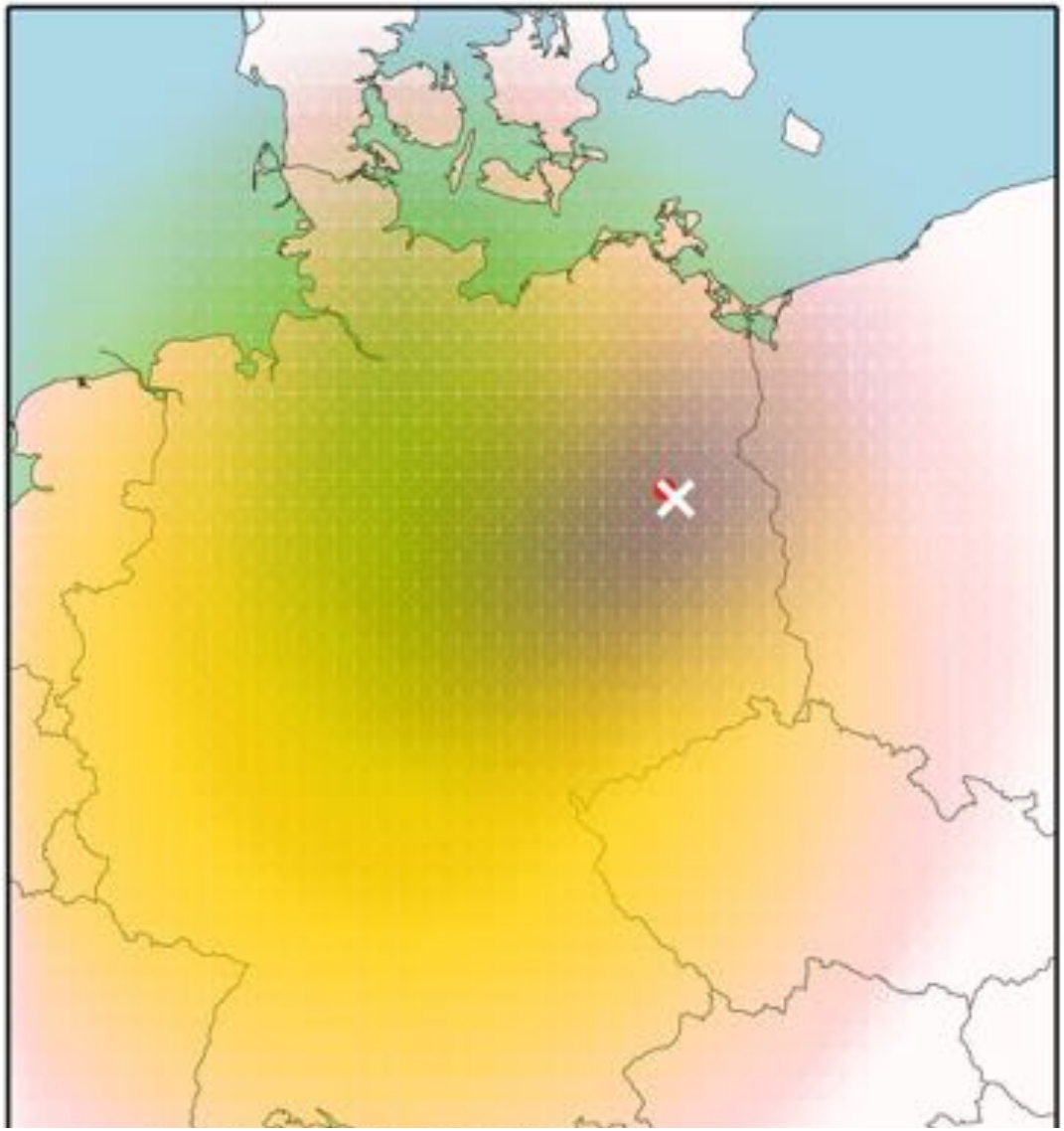


2. Current location



Idea

- Some words are used at certain locations more often than others
- Derive a probability distribution for each word
- High variance vs. low variance
- Common words vs. highly informative words



green: hhwahl
blue: berlin
red: nordbahnhof
yellow: rest

x = true location

“balken gucken und so **hhwahl** pa **nordbahnhof** in **berlin**”

Location of a tweet

$$Loc(t) = \frac{\sum_{i=0}^{n} \sigma_i^{-1} * m_i}{\sum_{i=0}^{n} \sigma_i^{-1}}$$

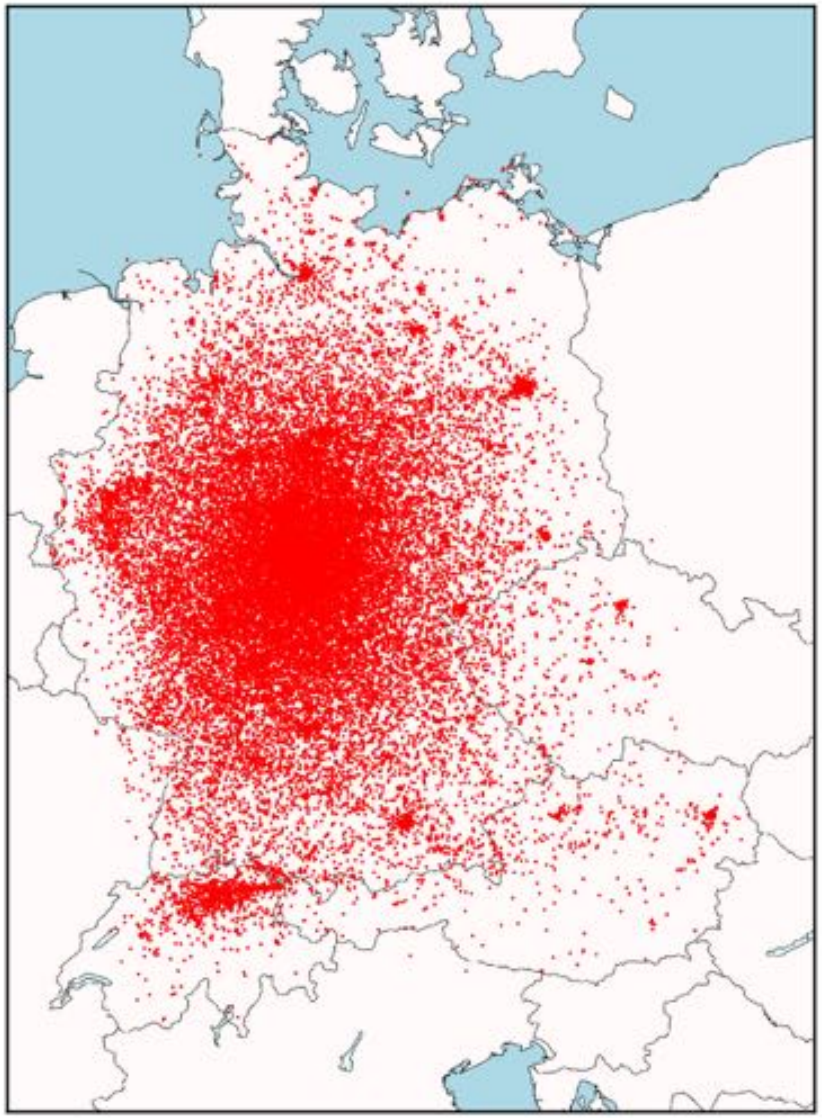
t = tweet with tokens $t_0 \ldots t_n$

variance $\sigma_0 \ldots \sigma_n$

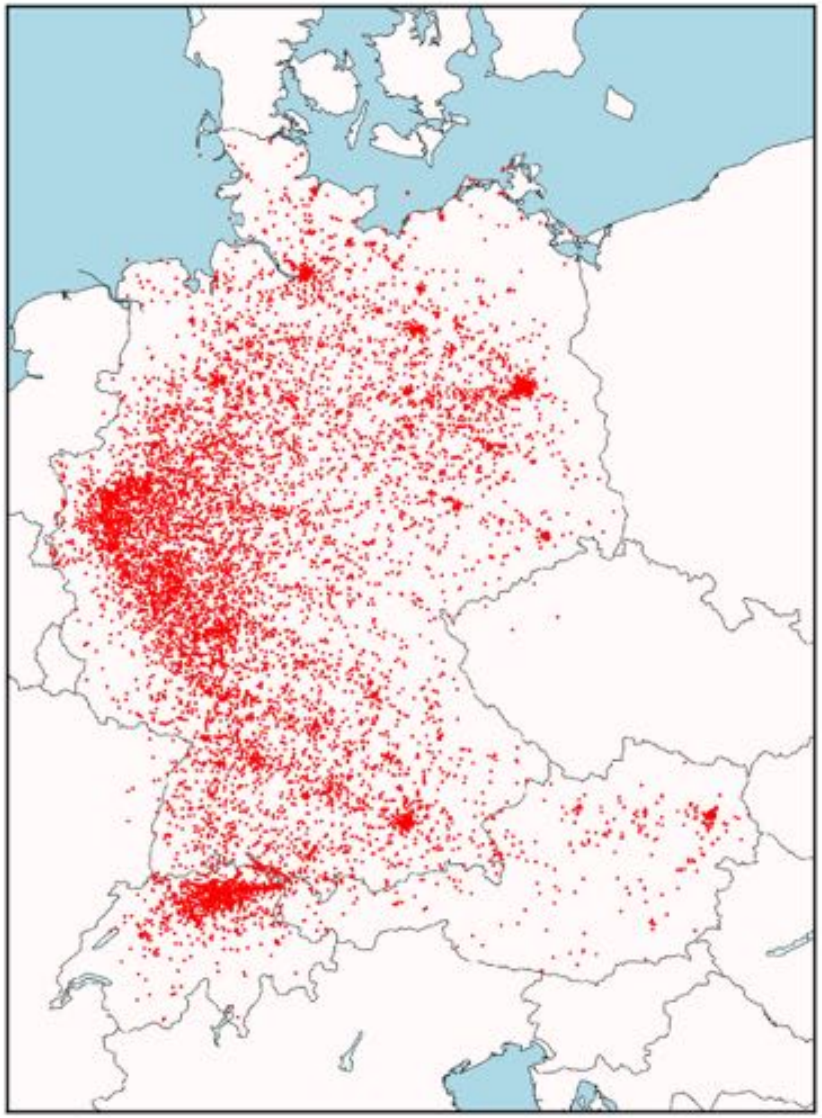
median location $m_0 \ldots m_n$

"balken gucken und so **hhwahl** pa **nordbahnhof** in **berlin**"

Midpoints of tokens (training corpus)

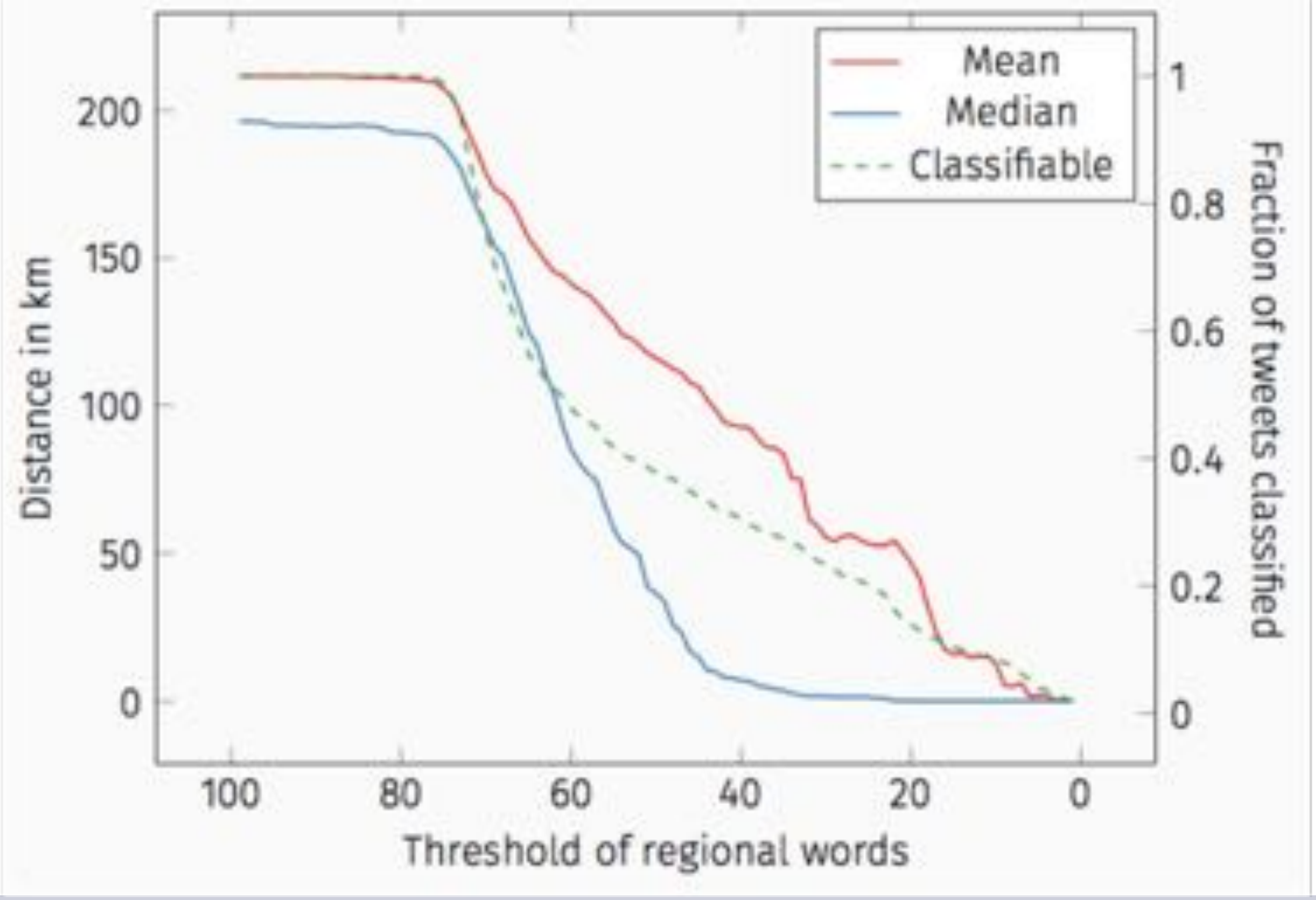


all tokens (100%)



25% of tokens

Regionality threshold



Results (test corpus)

Threshold	Mean	Median	#Tweets
100	212km	196km	1000
75	207km	188km	988
50	116km	36km	377
40	93km	7km	306
30	55km	1.56km	233
20	47km	0.06km	139
10	12km	0.00km	84

Regionally salient tokens

Berlin	Zurich	Essen
kadewe	tagi	rheinische
kudamm	uf	hattingen
alexanderplatz	het	herne
friedrichshain	isch	westfalen
brandenburg	scho	ddorf
fernsehturm	au	ruhr
dit	zuerichsee	thyssenkrupp
morjen	gseh	duisburg

Conclusion: geocoding of tweets

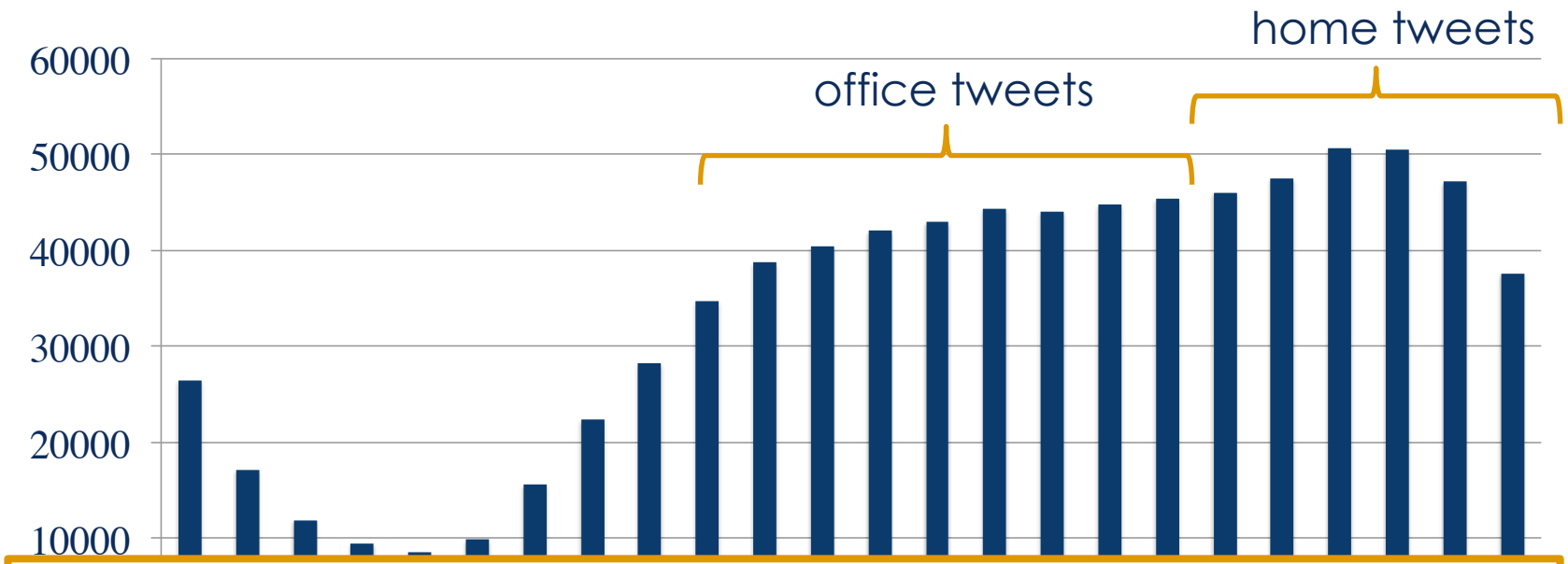
- Classify tweets only based on their text
- Geographic probability distribution for tokens
- Filtering and weighting tokens by variance
- Removal of wide-spread words
- Accurate reconstruction of tweets' location

But: How do geocoded and non-geocoded tweets differ?

Twitter + Circadian Rhythm

Joint work with Christopher Kyba, GFZ Potsdam

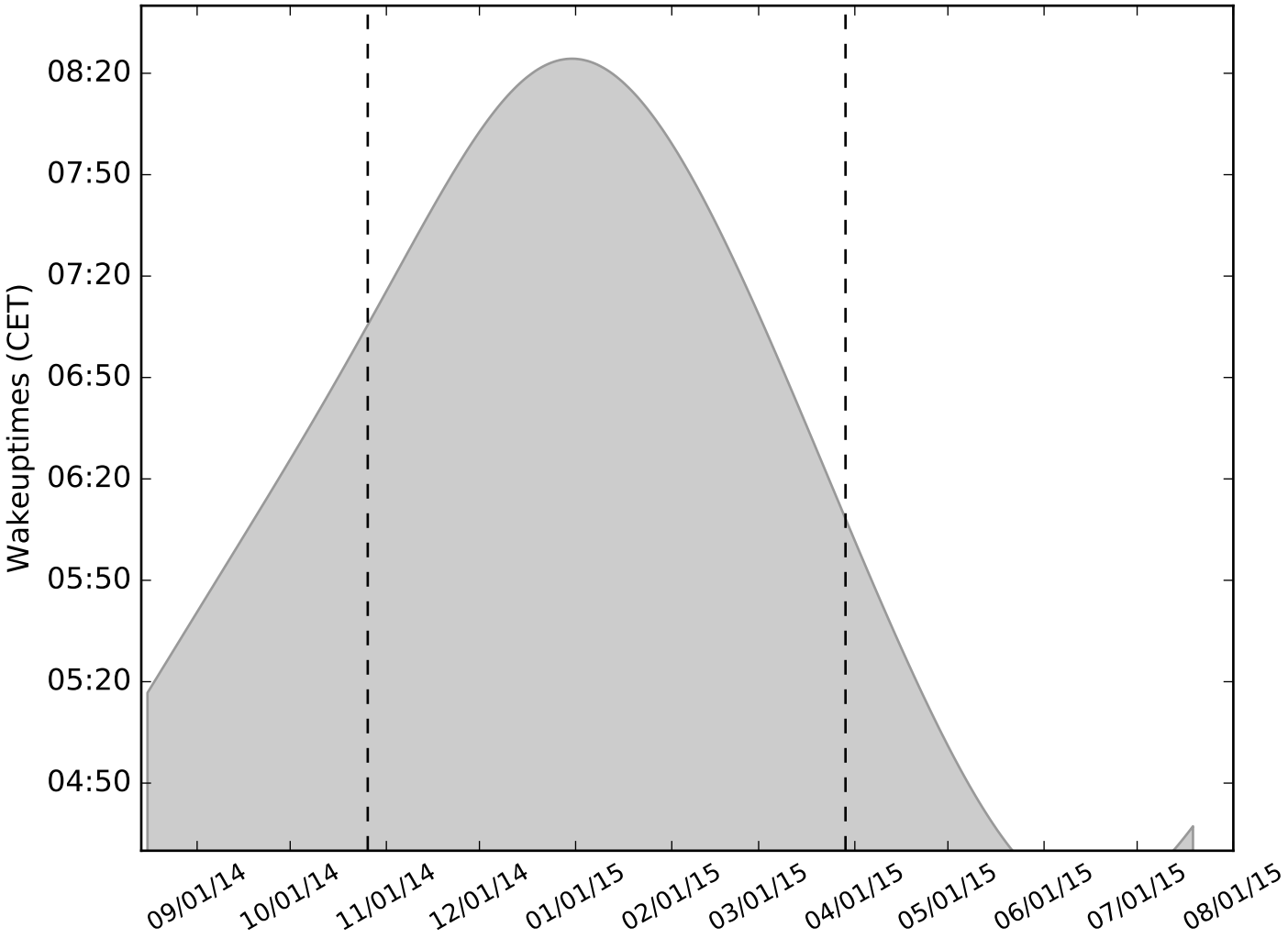
When do users tweet?

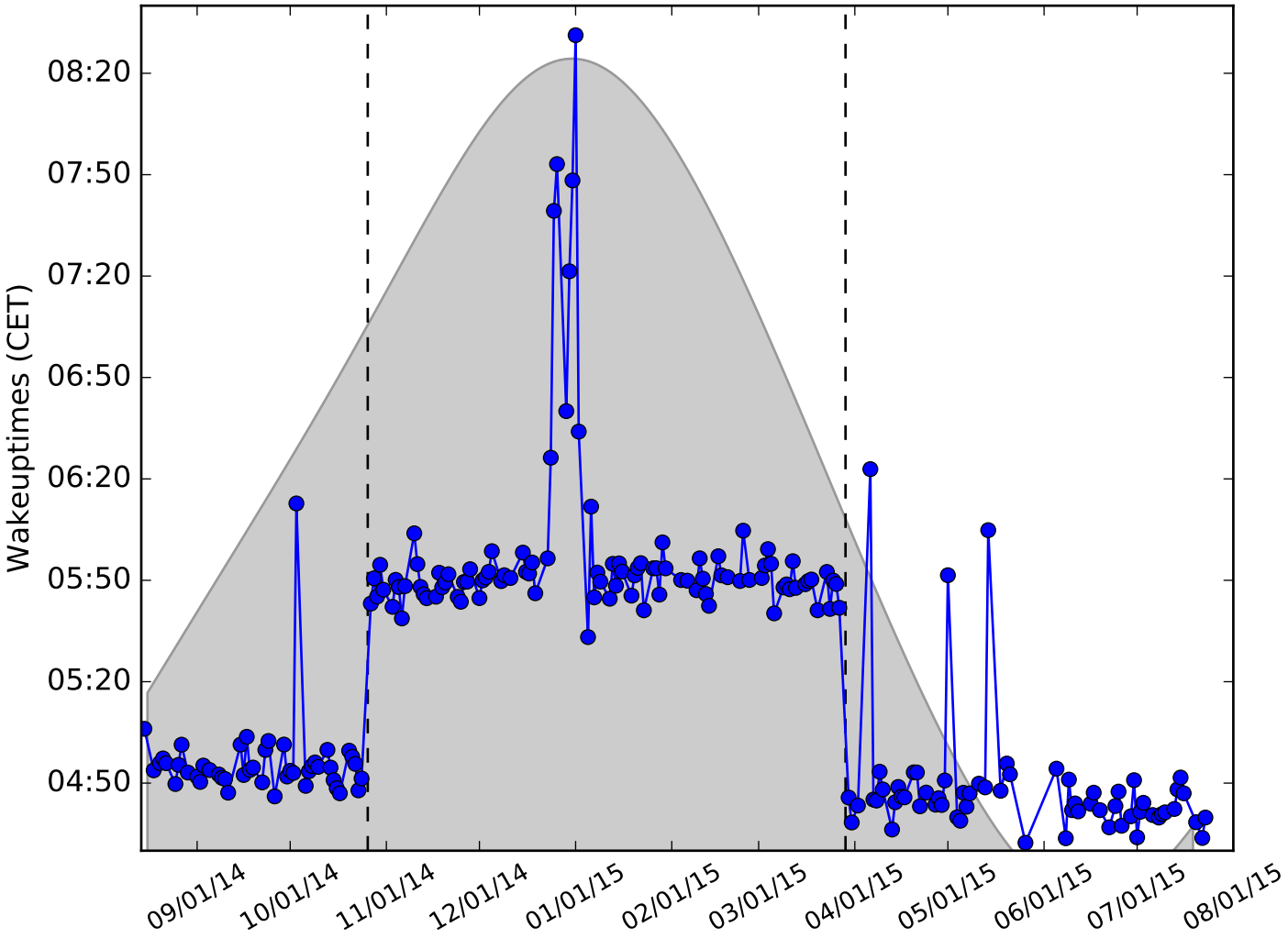


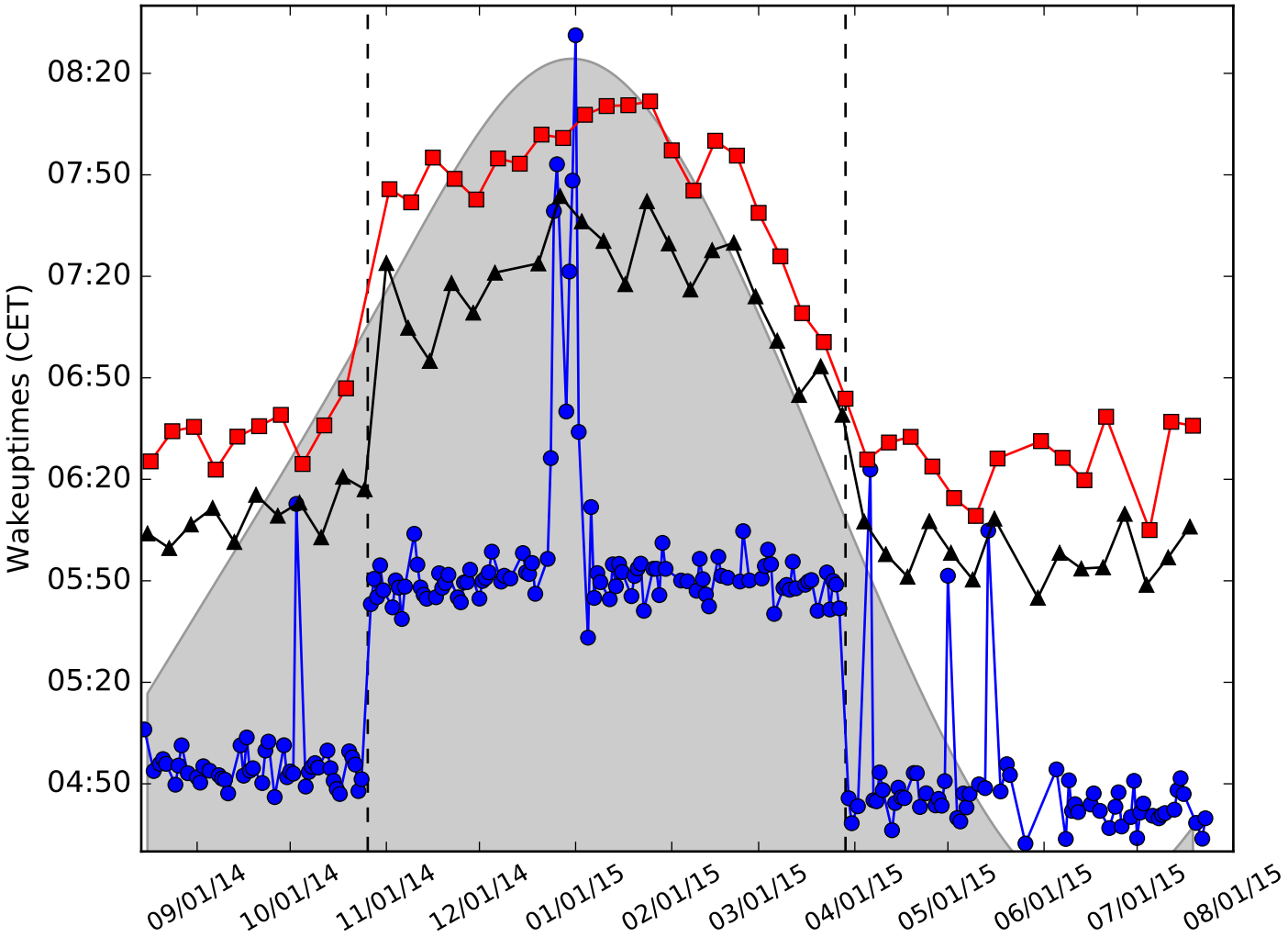
Use a Twitter time series to measure the sleep-wake pattern in humans and its interaction with DST.

Analysis

- "Onset of Twitter activity" = time at which the rate of 'good morning'-tweets reached half of the maximum
- The relation of this time to the sunrise time and social time was studied

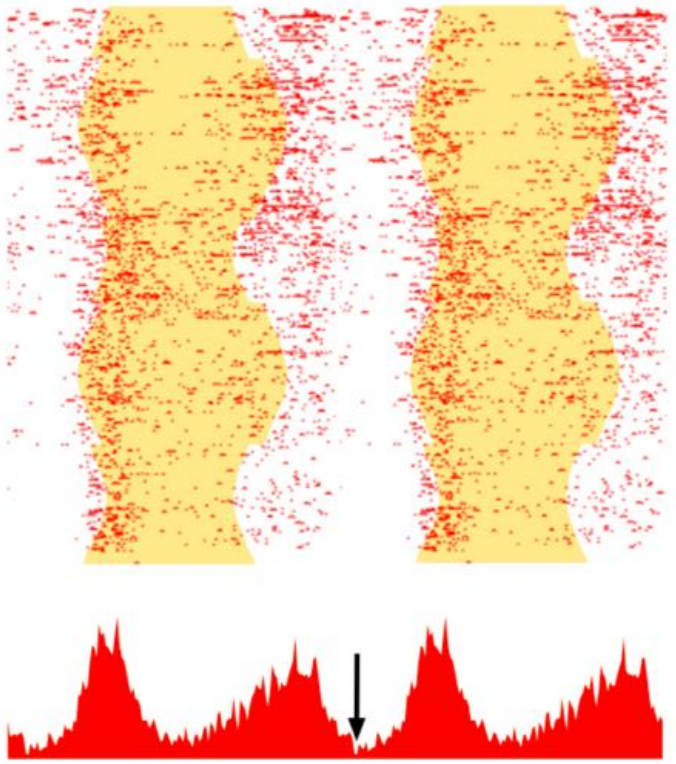

Wakeuptimes (CET)
08:20
07:50
07:20
06:50
06:20
05:50
05:20
04:50
09/01/14
10/01/14
11/01/14
12/01/14
01/01/15
02/01/15
03/01/15
04/01/15
05/01/15
06/01/15
07/01/15
08/01/15


Wakeuptimes (CET)
08:20
07:50
07:20
06:50
06:20
05:50
05:20
04:50
09/01/14
10/01/14
11/01/14
12/01/14
01/01/15
02/01/15
03/01/15
04/01/15
05/01/15
06/01/15
07/01/15
08/01/15


Wakeuptimes (CET)
08:20
07:50
07:20
06:50
06:20
05:50
05:20
04:50
09/01/14
10/01/14
11/01/14
12/01/14
01/01/15
02/01/15
03/01/15
04/01/15
05/01/15
06/01/15
07/01/15
08/01/15

Individual circadian rhythm

- @realdonaldtrump
- Dec 2014-Mar 2017
- Least active (~ mid-sleep time): 1:30am
- Sleep duration: max. 6.5 hours on 70% of days



Roenneberg, T. (2017). Twitter as a means to study temporal behaviour. *Current Biology*, *27*(17), R830-R832.

Summary

- Language on social media is variable in structured ways
- Using linguistic information to recover metadata
- Social media data as a sensor for human behavior / real-world events

Thank you.

tatjana.scheffler@uni-potsdam.de