NLP of Microblogs, Day 2: Working with Metadata

ESSLLI 2016

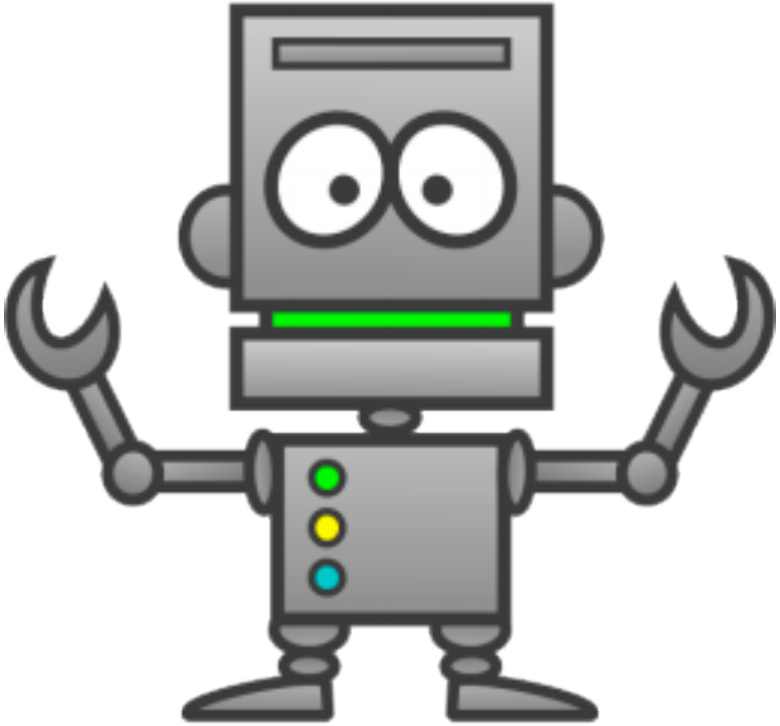
Bolzano/Bozen, Italy

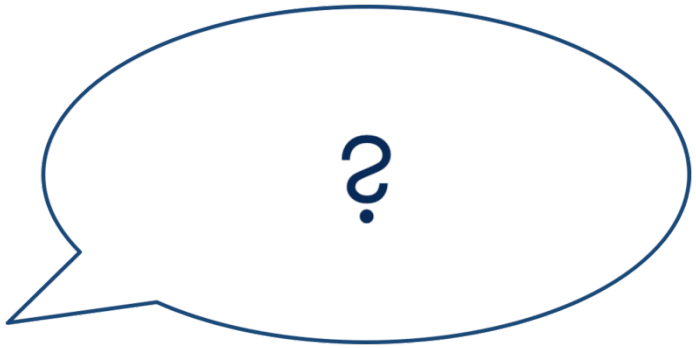
Tatjana Scheffler & Manfred Stede

tatjana.scheffler@uni-potsdam.de / stede@uni-potsdam.de

August 23, 2016

Build Twitter bots?





Twitter Metadata

time

geo_coordinates

user profile info

id_str	from_user	text	created_at	time	geo_coor	user_lang	in_reply_to_u	in_reply_to_sc	from_user_id_str	in_reply_to	source	profile_image_url	user_followers_count	user_friends_count	user
7672580	Janetts_Meinung	obwohl das Buch "Wunderbare... https://t.co/rjnpd8tdzT	Sun Aug 21 07:12:29 +0(	21/08/2016 08:12:29		de			127902112		<a href="http://www.face	http://pbs.twimg.com/pr	27	28	
7672575			Sun Aug 21 07:10:32 +0(	21/08/2016 08:10:32		de			4011535756		<a href="http://twitter.co	http://pbs.twimg.com/pr	43	29	
7672573			Sun Aug 21 07:09:50 +0(	21/08/2016 08:09:50		de			4011535756		<a href="http://twitter.co	http://pbs.twimg.com/pr	43	29	
7672565		@c_m_barnhart Gute nacht :)	Sun Aug 21 07:06:26 +0(	21/08/2016 08:06:26		en	71653617635	c_m_barnhart	491850402	767256094	<a href="http://twitter.co	http://pbs.twimg.com/pr	493	460	
7672563		Good Night All animals friend https://t.co/B89kQzn3Dw	Sun Aug 21 07:05:52 +0(	21/08/2016 08:05:52		en-gb			701440480769404928		<a href="http://twitter.co	http://pbs.twimg.com/pr	383	344	
7672560		@lovemebluee gute nacht	Sun Aug 21 07:04:39 +0(	21/08/2016 08:04:39		de	2872615833	lovemebluee	2939455557	767230588	<a href="http://twitter.co	http://pbs.twimg.com/pr	762	106	
7672558			Sun Aug 21 07:03:54 +0(	21/08/2016 08:03:54		en			4920008313		<a href="http://twitter.co	http://pbs.twimg.com/pr	187	484	
7672558		Ich leg mich jz pennen. Gute Nacht xd	Sun Aug 21 07:03:51 +0(	21/08/2016 08:03:51		de			1901319685		<a href="http://twitter.co	http://pbs.twimg.com/pr	748	58	
7672555			Sun Aug 21 07:02:26 +0(	21/08/2016 08:02:26		de			4777377435		<a href="http://twitter.co	http://pbs.twimg.com/pr	192	182	
7672544		RT @_DeenisTJ_: Gute Nacht♥	Sun Aug 21 06:58:10 +0(	21/08/2016 07:58:10		de			2861956357		<a href="http://twitter.co	http://pbs.twimg.com/pr	362	91	
7672541		@_Dosenwerfer @Nerdiiis @winterschnuppe gute Nacht	Sun Aug 21 06:57:08 +0(	21/08/2016 07:57:08		de	2293432255	Dosenwerfer	3292060903	767254101	<a href="http://twitter.co	http://pbs.twimg.com/pr	791	265	
7672535		@katastrawfairy ich hatte gute Nacht gesagt! .. Du solltest doch schlafen ;)	Sun Aug 21 06:54:38 +0(	21/08/2016 07:54:38		de	1235051204	katastrawfairy	3355544741	767207712	<a href="http://twitter.co	http://pbs.twimg.com/pr	16	16	
7672527		RT @_DeenisTJ_: Gute Nacht♥	Sun Aug 21 06:51:17 +0(	21/08/2016 07:51:17		de			3028395099		<a href="http://twitter.co	http://pbs.twimg.com/pr	147	59	
7672526		@carola_bach Dann übernimm besser du mal die Schicht! Gute nacht!	Sun Aug 21 06:50:59 +0(	21/08/2016 07:50:59		en	2803486586	carola_bach	2696904632	767252422	<a href="https://about.tw	http://pbs.twimg.com/pr	448	289	
7672522			Sun Aug 21 06:49:27 +0(	21/08/2016 07:49:27		de			3590643748		<a href="http://twitter.co	http://pbs.twimg.com/pr	20	16	
7672511		RT @_DeenisTJ_: Gute Nacht♥	Sun Aug 21 06:45:05 +0(	21/08/2016 07:45:05		de			3238305772		<a href="http://twitter.co	http://pbs.twimg.com/pr	79	113	
7672506		RT @_DeenisTJ_: Gute Nacht♥	Sun Aug 21 06:42:53 +0(	21/08/2016 07:42:53		de			2971699859		<a href="http://twitter.co	http://pbs.twimg.com/pr	126	90	
7672499		RT @philinked: Gute Nacht	Sun Aug 21 06:40:10 +0(	21/08/2016 07:40:10		de			2520654340		<a href="http://twitter.co	http://pbs.twimg.com/pr	557	45	
7672499		RT @philinked: Gute Nacht	Sun Aug 21 06:40:07 +0(	21/08/2016 07:40:07		de			2520654340		<a href="http://twitter.co	http://pbs.twimg.com/pr	557	45	
7672498		RT @philinked: Gute Nacht	Sun Aug 21 06:40:04 +0(	21/08/2016 07:40:04		de			2520654340		<a href="http://twitter.co	http://pbs.twimg.com/pr	557	45	
7672495		Gute Nacht	Sun Aug 21 06:38:46 +0(	21/08/2016 07:38:46		de			2742242917		<a href="http://twitter.co	http://pbs.twimg.com/pr	141	230	
7672487		gunner https://t.co/hhj3Fn3vjh	Sun Aug 21 06:35:39 +0(	21/08/2016 07:35:39		tr			94107162		<a href="https://mobile.t	http://pbs.twimg.com/pr	132	143	
7672486		Gute Nacht	Sun Aug 21 06:34:58 +0(	21/08/2016 07:34:58		de			4817308534		<a href="http://twitter.co	http://pbs.twimg.com/pr	132	204	
7672483		RT @Clearfan15: Gute Nacht #Ostsee https://t.co/cVeOvMSpV	Sun Aug 21 06:33:54 +0(	21/08/2016 07:33:54		de			2191795613		<a href="http://twitter.co	http://pbs.twimg.com/pr	2781	2211	
7672478			Sun Aug 21 06:31:51 +0(	21/08/2016 07:31:51		de			2800511612		<a href="http://twitter.co	http://pbs.twimg.com/pr	86	41	
7672477		da bleibt nur ein Kerzenlichtabendrot ♡♡♡	Sun Aug 21 06:31:22 +0(	21/08/2016 07:31:22		de			4192867587		<a href="http://www.twitt	http://pbs.twimg.com/pr	101	120	
7672465		RT @_DeenisTJ_: Gute Nacht♥	Sun Aug 21 06:26:48 +0(	21/08/2016 07:26:48		de			3214710001		<a href="http://twitter.co	http://pbs.twimg.com/pr	32	24	
7672450		Gute Nacht!	Sun Aug 21 06:20:59 +0(	21/08/2016 07:20:59		de			102331010		<a href="http://twitter.co	http://pbs.twimg.com/pr	115	64	
7672450		Gute Nacht!	Sun Aug 21 06:20:59 +0(	21/08/2016 07:20:59		de			102331010		<a href="http://twitter.co	http://pbs.twimg.com/pr	115	64	
7672448		RT @_DeenisTJ_: Gute Nacht♥	Sun Aug 21 06:20:11 +0(	21/08/2016 07:20:11		de			3435112155		<a href="http://twitter.co	http://pbs.twimg.com/pr	152	133	
7672447		RT @_DeenisTJ_: Gute Nacht♥	Sun Aug 21 06:19:44 +0(	21/08/2016 07:19:44		de			754398999248232192		<a href="http://twitter.co	http://pbs.twimg.com/pr	10	12	
7672445		@kim_13_kun Gute Nacht. あああああ 今日もお疲れさまでした明日な!	Sun Aug 21 06:18:57 +0(	21/08/2016 07:18:57		ja	2598978290	kim_13_kun	1181550674	767244256	<a href="http://makebot	http://pbs.twimg.com/pr	575	471	
7672445		RT @_DeenisTJ_: Gute Nacht♥	Sun Aug 21 06:18:45 +0(	21/08/2016 07:18:45		de			3569475135		<a href="http://twitter.co	http://pbs.twimg.com/pr	231	100	
7672435		RT @blaublu42: Gute Nacht und danke für. ..... https://t.co/FFpgIUArDR	Sun Aug 21 06:15:00 +0(	21/08/2016 07:15:00		ja			2456		<a href="http://twitter.co	http://pbs.twimg.com/pr	2044	2954	
7672433		https://t.co/m5c3xT12C	Sun Aug 21 06:14:13 +0(	21/08/2016 07:14:13		en			2196122099		<a href="http://twitter.co	http://pbs.twimg.com/pr	4260	4237	
7672425		Liebe @Regendelfin	Sun Aug 21 06:10:55 +0(	21/08/2016 07:10:55		de			2344913904		<a href="http://twitter.co	http://pbs.twimg.com/pr	1028	674	

in_reply_to

user network

Metadata

1. geographic coordinates
2. time stamps
3. reply info (Thursday!)
4. user network (not covered in this course)
   - followers
   - friends
   - retweets
   - mentions
5. user profile information

1. geographic coordinates

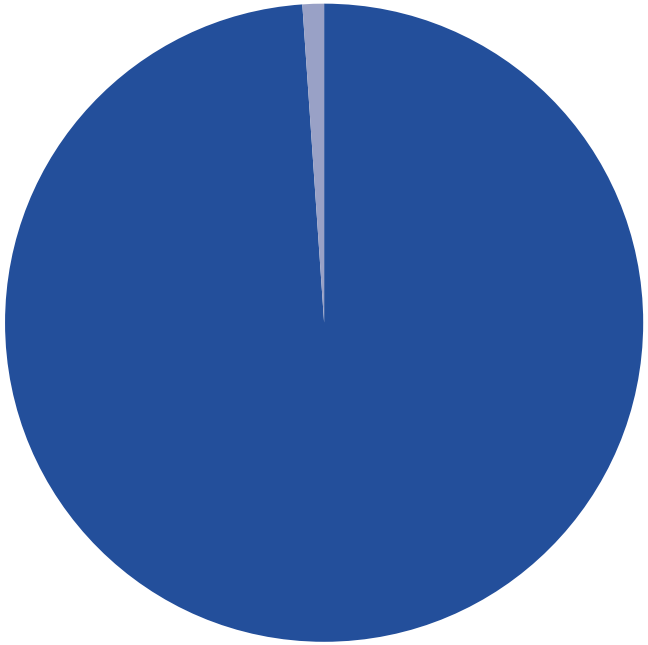
Geolocation information in Twitter

```
place (
| country =  "Germany"
| place_type =  "city"
| country_code =  "DE"
| name =  "Stuttgart"
| full_name =  "Stuttgart, Stuttgart"
| url =  "http://api.twitter.com/1/geo/id/
e385d4d639c6a423.json"
| id =  "e385d4d639c6a423"
| bounding_box (
| | coordinates => Array (1) (
| | |[...] )
| | type =  "Polygon" )
| attributes ( ))
```

Where do Germans tweet from?

images by Norman Rosner

Geo-tagged tweets



German tweets
April 2013

- non geo-tagged: 23,915,825
- geo-tagged: 263,364

Tatjana Scheffler. A German Twitter Snapshot. In: *Proceedings of LREC*, Reykjavik, Iceland. 2014.

Simone Biles @Simone_Biles · 19h

bye Rio De Janeiro

good things must end, Thank you RIO for unforgettable memories ❤️ Being the flag bearer was a cherry on top! Congrats to all the other USA athletes as well. We killed it out there! GO TEAM USA 🇺🇸✨

1.3K 8.4K

Mario D'Aloisio

@Mario_DAloisio

Follow

Il giardino ad Alexanderplatz, in pieno centro, parla chiaro. Ovunque è così...

holly stump

@hollycs_

Follow

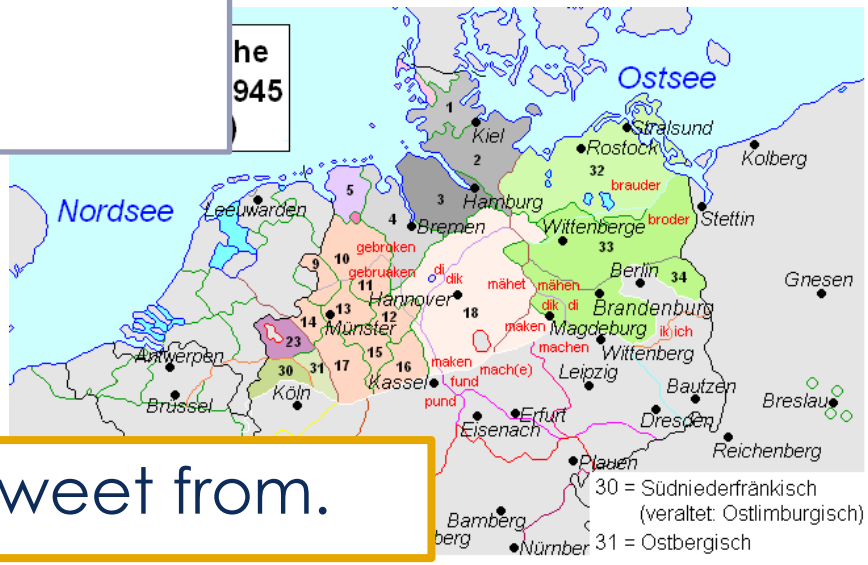
Last night uberd form bar to wawa, bought the driver a hoagie, uberd back to the bar an[illegible] my wawa at the bar

Simone Biles @Simone_Biles · Aug 20

beach day

Alexandra Raisman and Madison Kocian

Goal

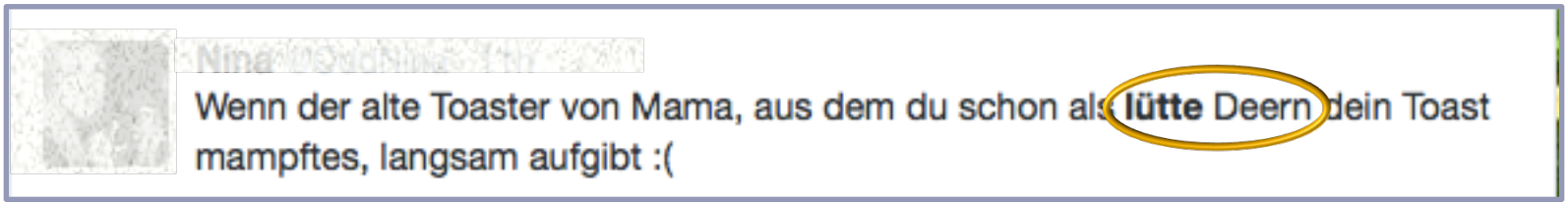


Find out where users tweet from.

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


Simone Biles @Simone_Biles · Aug 20
beach day
Alexandra Raisman and Madison Kocian

Assumptions for identifying tweet location

- dialectal region and current location are independent
- But:
  - regionally diverging tweets should be relatively rare
  - probabilistic model of regional salience allows for assignment of tweets to several regions equally
- use only a tweet's text as features
  - ignore user profile and other metadata (for now)
- majority of tweets are "non-regional"

Related work

- Sociolinguistic Twitter studies on dialects (Eisenstein et al., 2012; Grieve, 2014)
- Localizing Twitter users by aggregating all of their tweets (Cheng et al., 2010; Hecht et al., 2011)
- Deriving location specific words from geotagged tweets (Leetaru et al., 2013)
- Thesaurus based geolocation, using known dialectal words (Scheffler et al., 2014)
- Classifying tweets into dialect regions (Scheffler et al., 2014)

Geographic Mapping of Tweets

Joint work with Johannes Gontrum

Approach

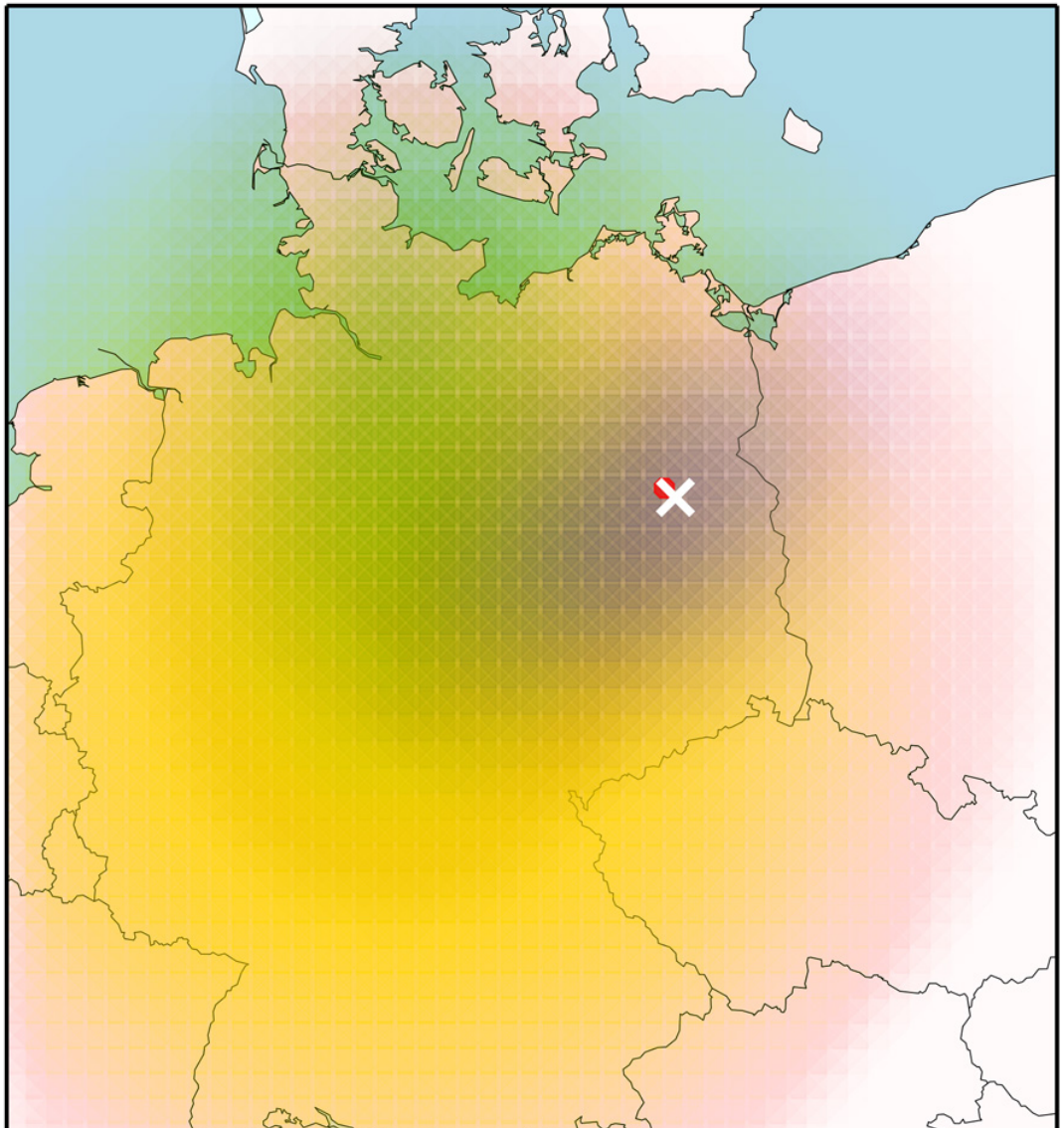
- Reconstructing the point of origin for German tweets.
  - Text based
  - Language independent
  - Using simple statistic properties
  - Exact coordinates
- Using the best 40% of all tokens, tweets can be reliably classified.
  - Median: 7km
  - Mean: 93km

Data

- 65 mio. German tweets
- Filtered by common German words
- Time span: February - May 2015
- Filtering corpus:
  - Removing tweets without geo tag
  - Removing tweets not from Germany, Austria, Switzerland
  - Removing generated tweets from bots
  - Removing numbers, URLs, mentions etc.
- Remaining: 360k tweets (0.55%)

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**"

Reconstructing the location

- The original position of a tweet is dependent on the variance and median location of each token.

1. Calculate variance and median location for each token.
2. Remove common/non-regional words by variance threshold
3. Compute the weighted midpoint for all other words in the tweet.
4. Weight: Inverse variance.

Location of a tweet

t = tweet with tokens $t_0 \ldots t_n$

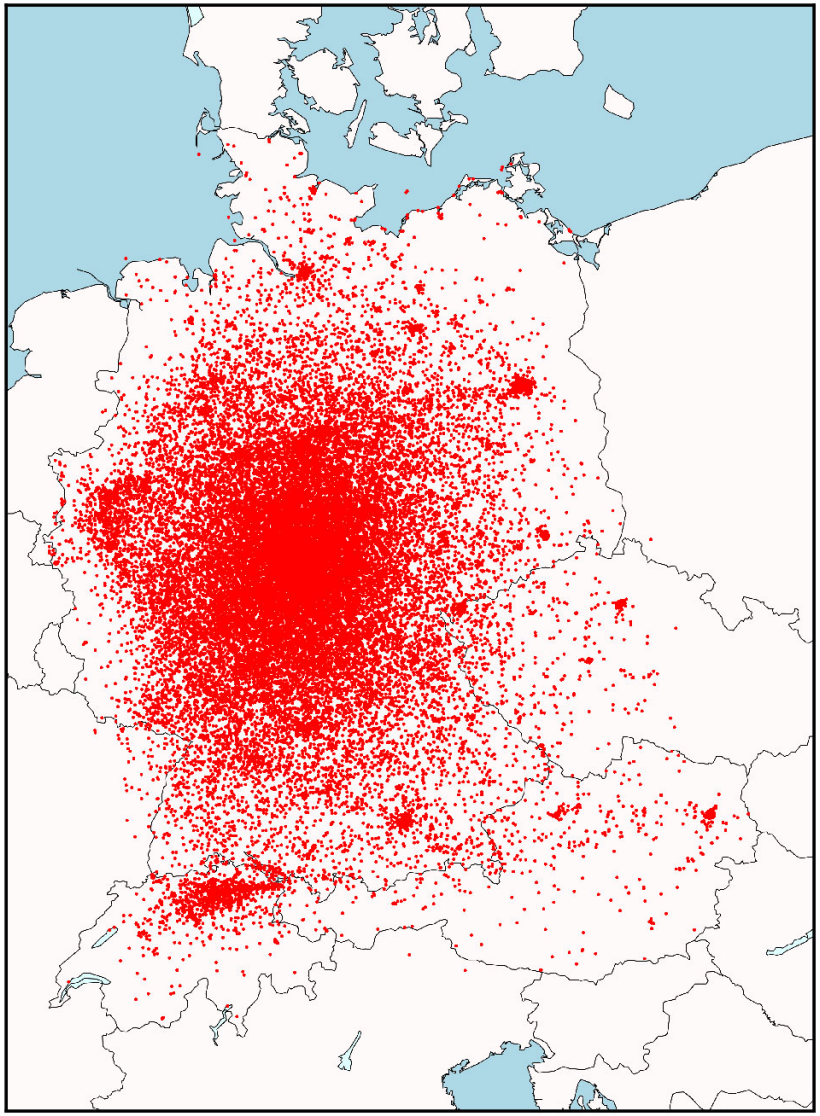
variance $\sigma_0 \ldots \sigma_n$

median location $m_0 \ldots m_n$

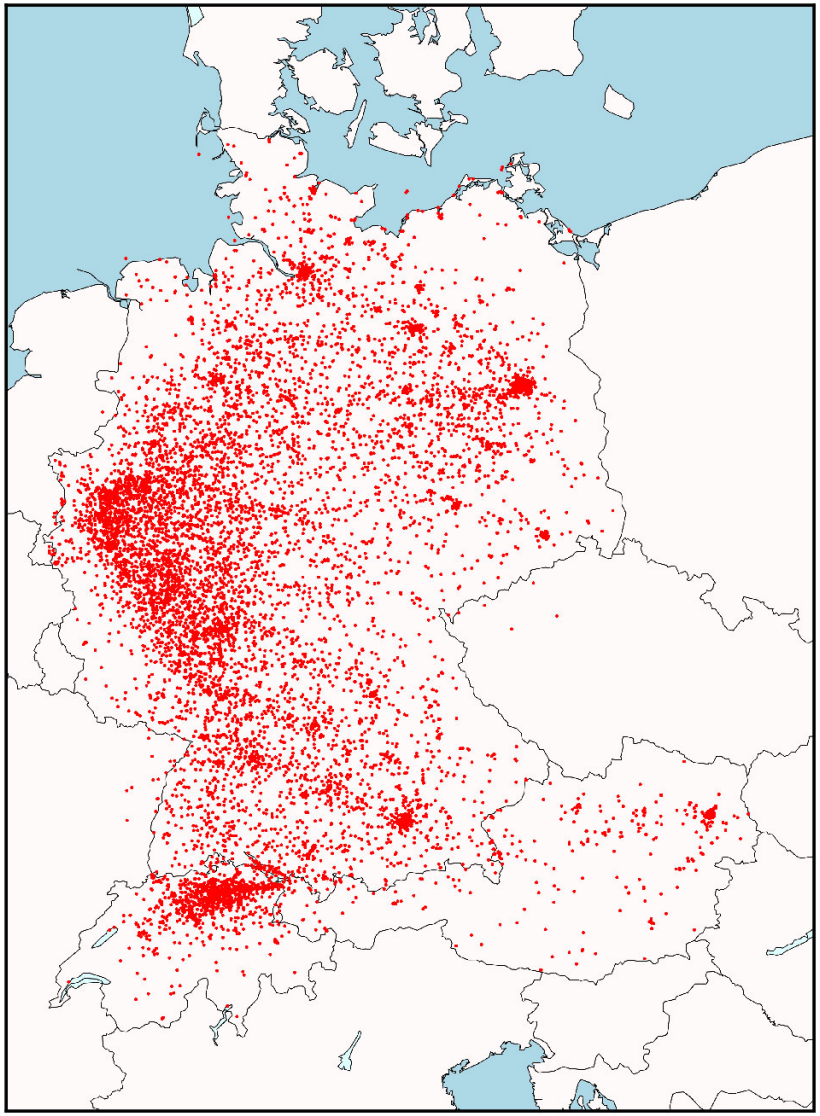
$$Loc(t) = \frac{\sum_{i=0}^{n} \vec{\sigma_i}^{-1} * \vec{m}_i}{\sum_{i=0}^{n} \vec{\sigma_i}^{-1}}$$

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

Midpoints of tokens (training corpus)



all tokens (100%)

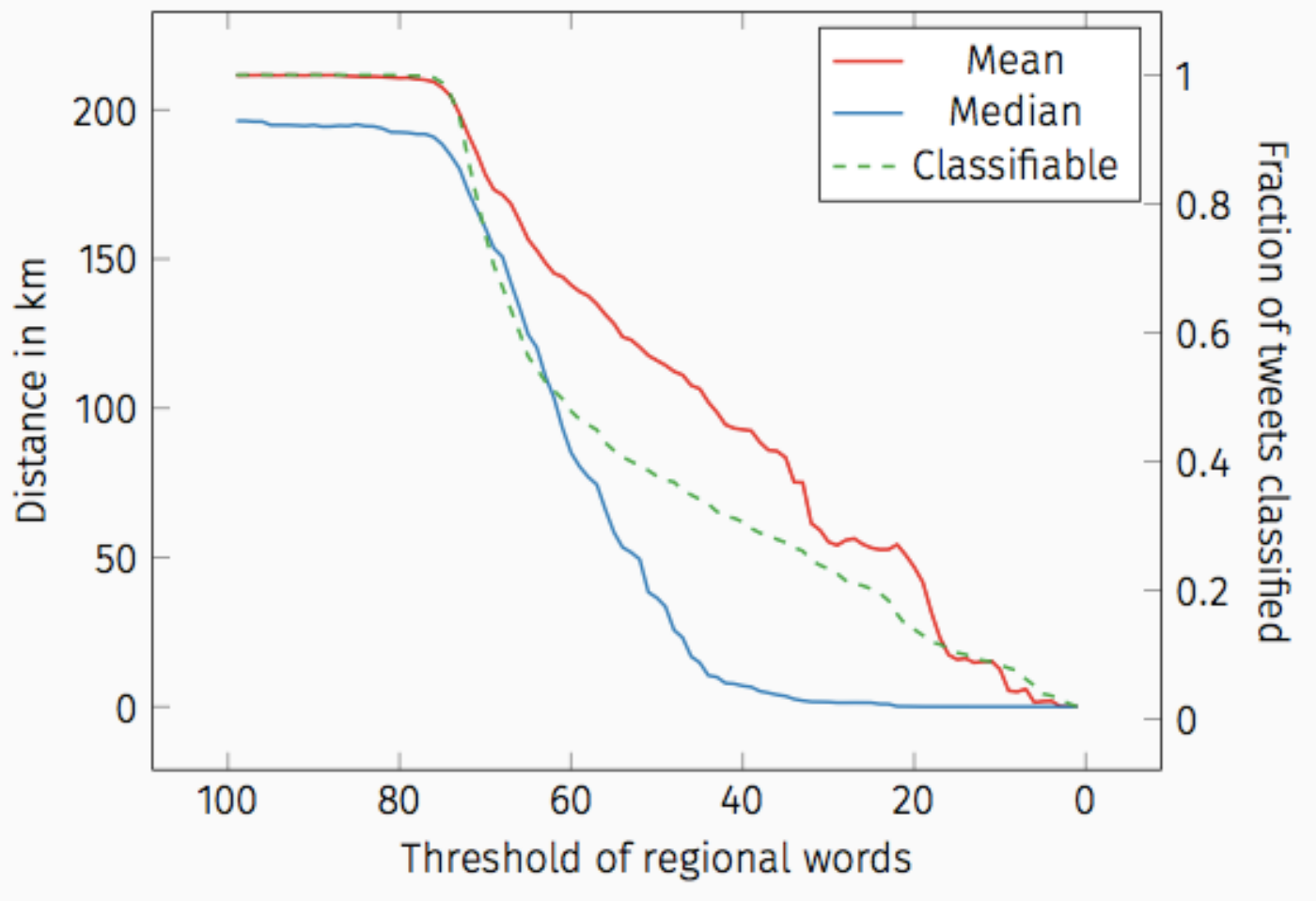


25% of tokens

Importance of filtering

- Not every token contains information about locality.
- Classifying a tweet with only common words will fail.
- False confidence in the result.
- Sort all tokens by variance, keep only lowest X%.
- Assumption: X% of words are regionally salient.

Regional threshold



Results

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

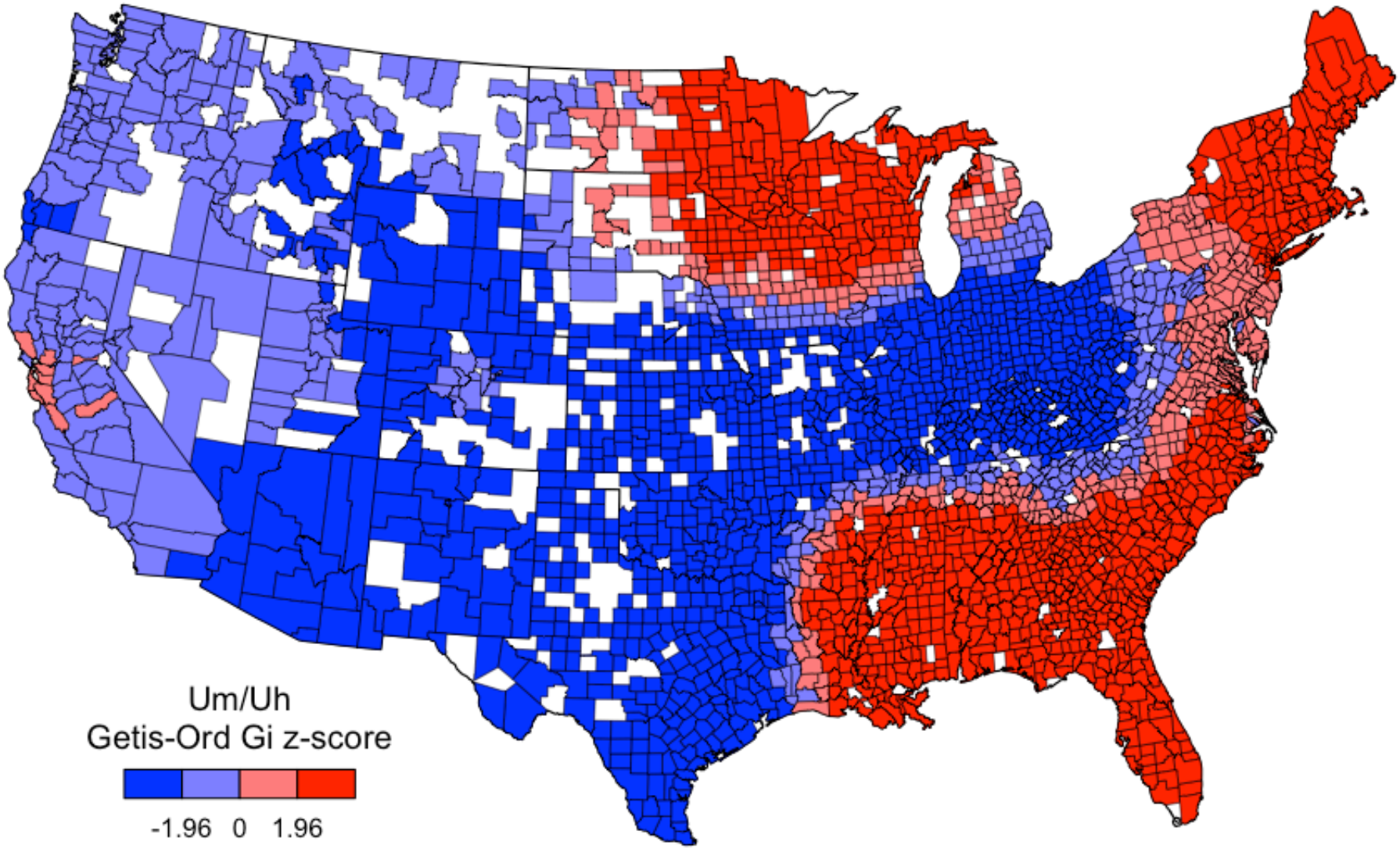
Regionally salient tokens

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

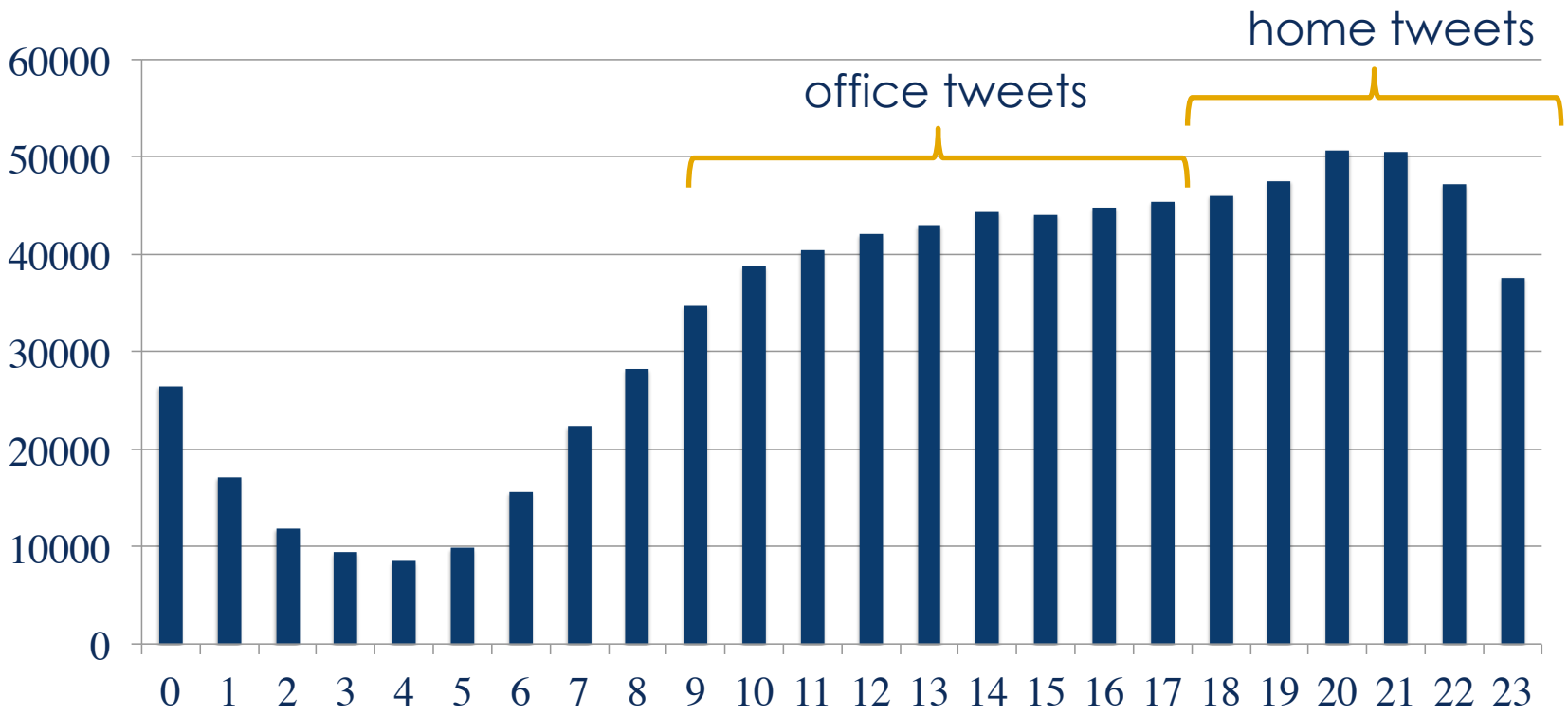
But: How do geocoded and non-geocoded tweets differ?

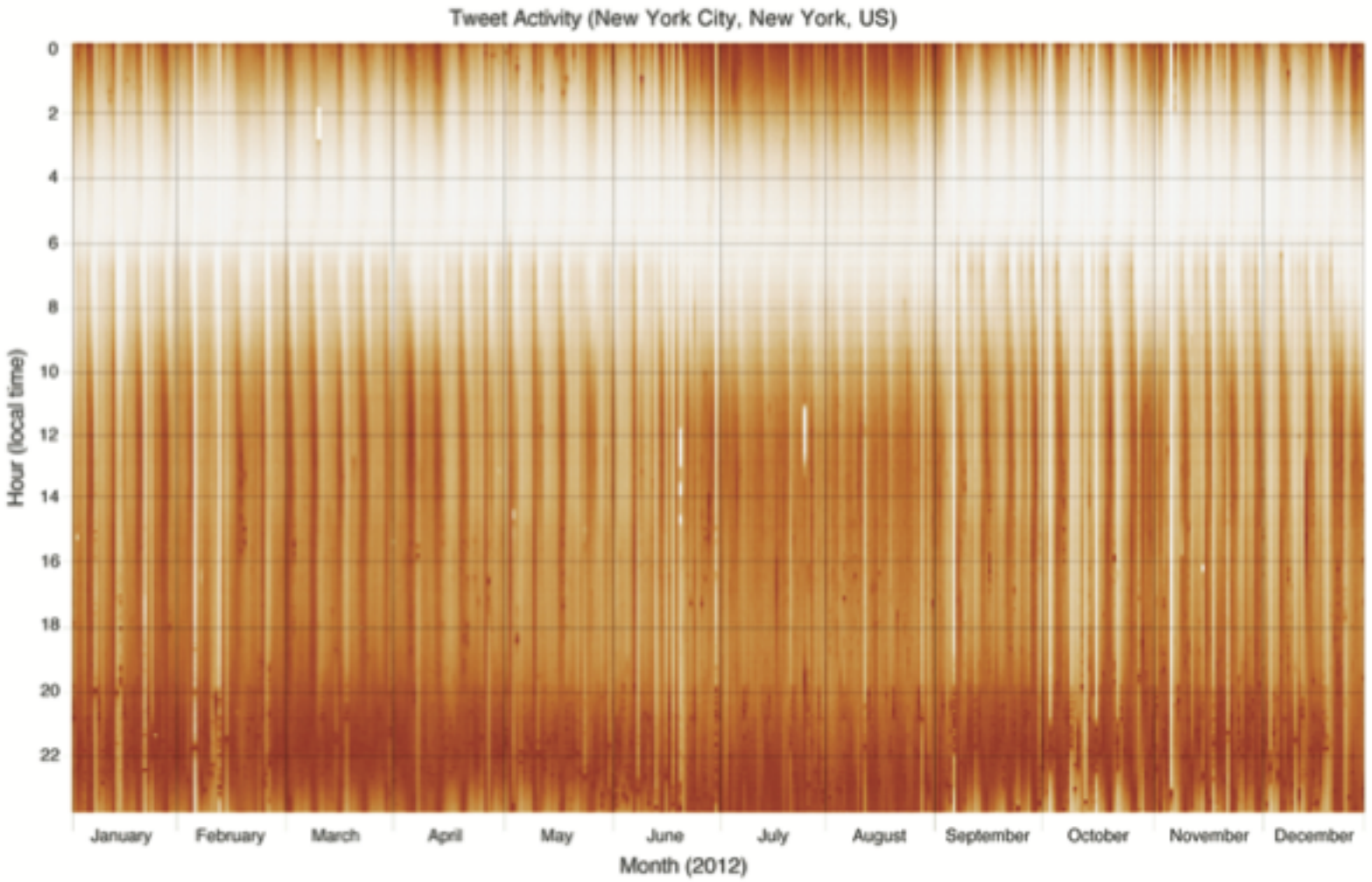


um / uh – Jack Grieve, https://sites.google.com/site/jackgrieveaston/treesandtweets , August 18, 2014

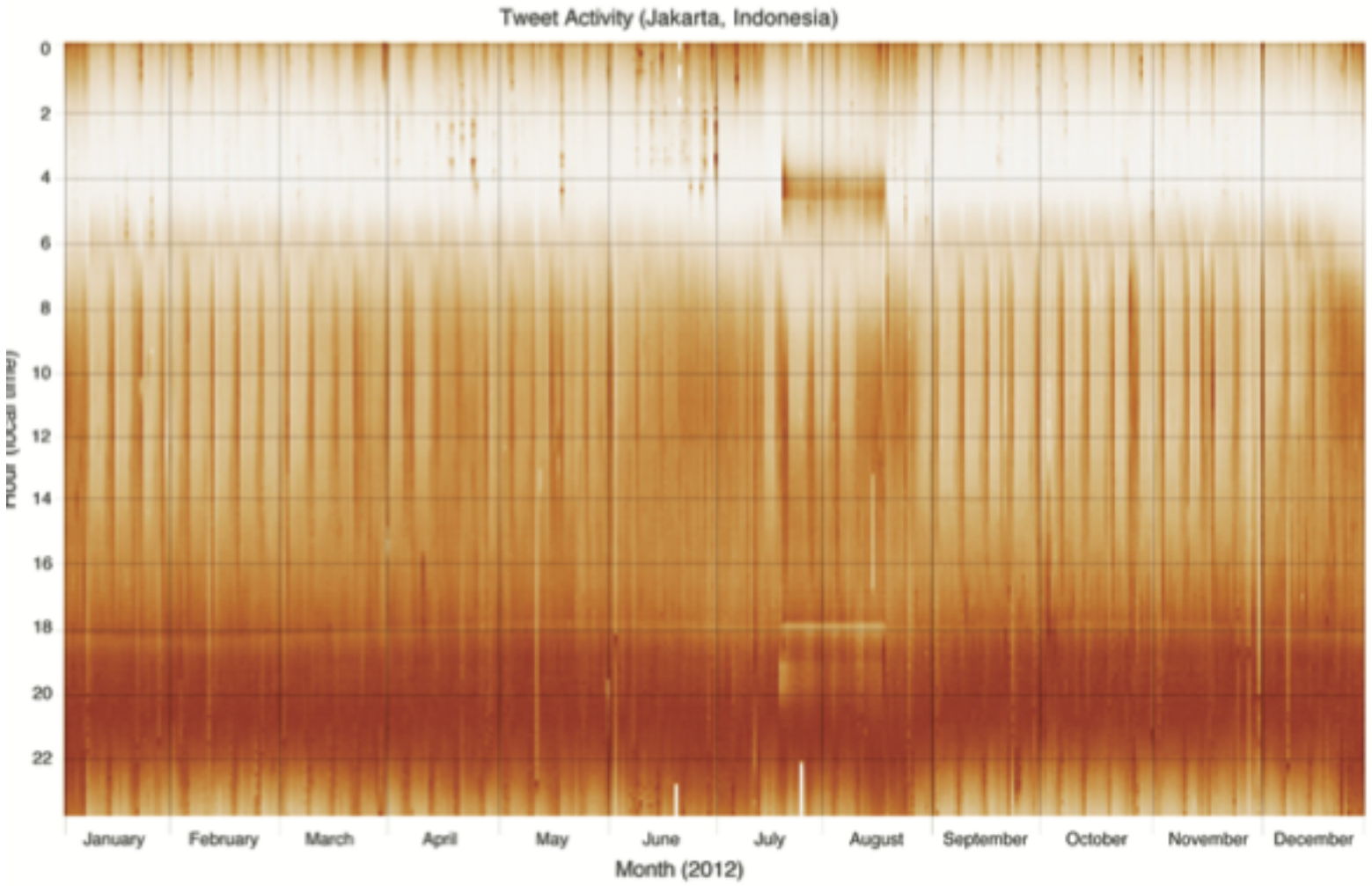
2. time stamps

When do users tweet?





Rios/Lin, 2013: Visualizing the "Pulse" of World Cities on Twitter. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.

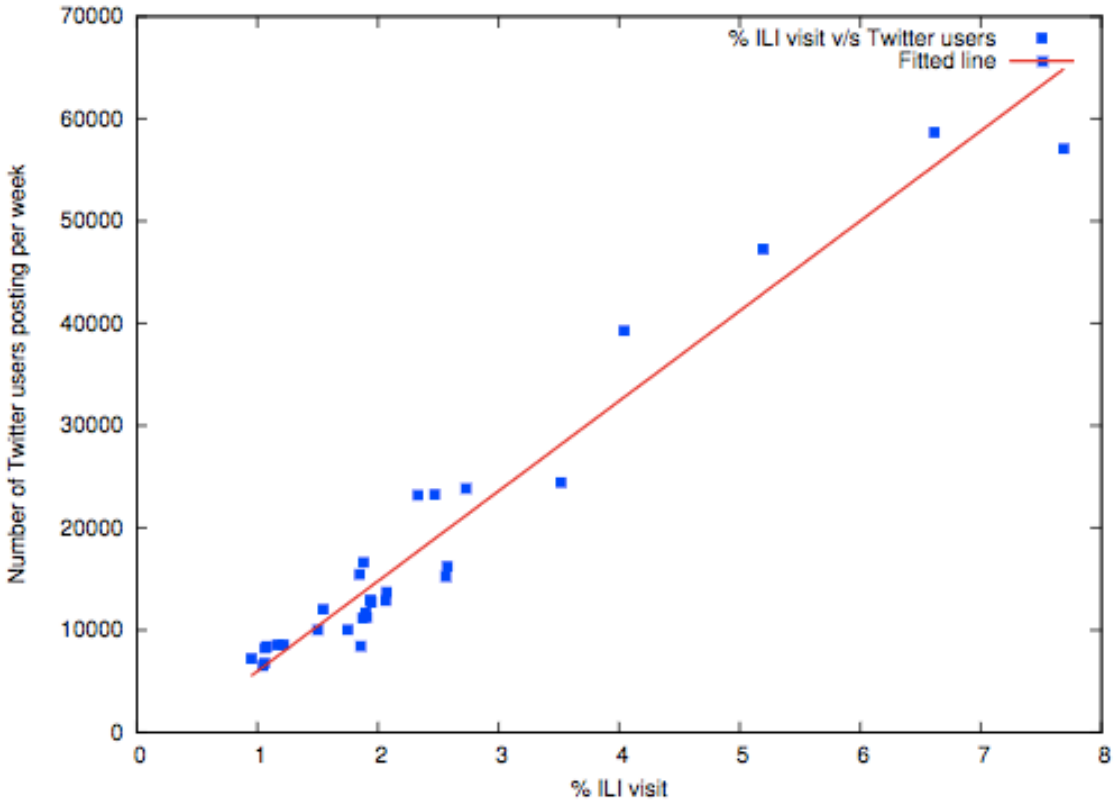


Rios/Lin, 2013: Visualizing the "Pulse" of World Cities on Twitter. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.

Twitter data as a social sensor

- earthquake observations (Sakaki et al., 2010)
- flu trend prediction (Achrekar et al., 2011)
- effect of weather on mood (Hannak et al., 2012)
- event detection (Ritter et al., 2012)

Flu trends (Achrekar et al., 2011)



Twitter + Circadian Rhythm

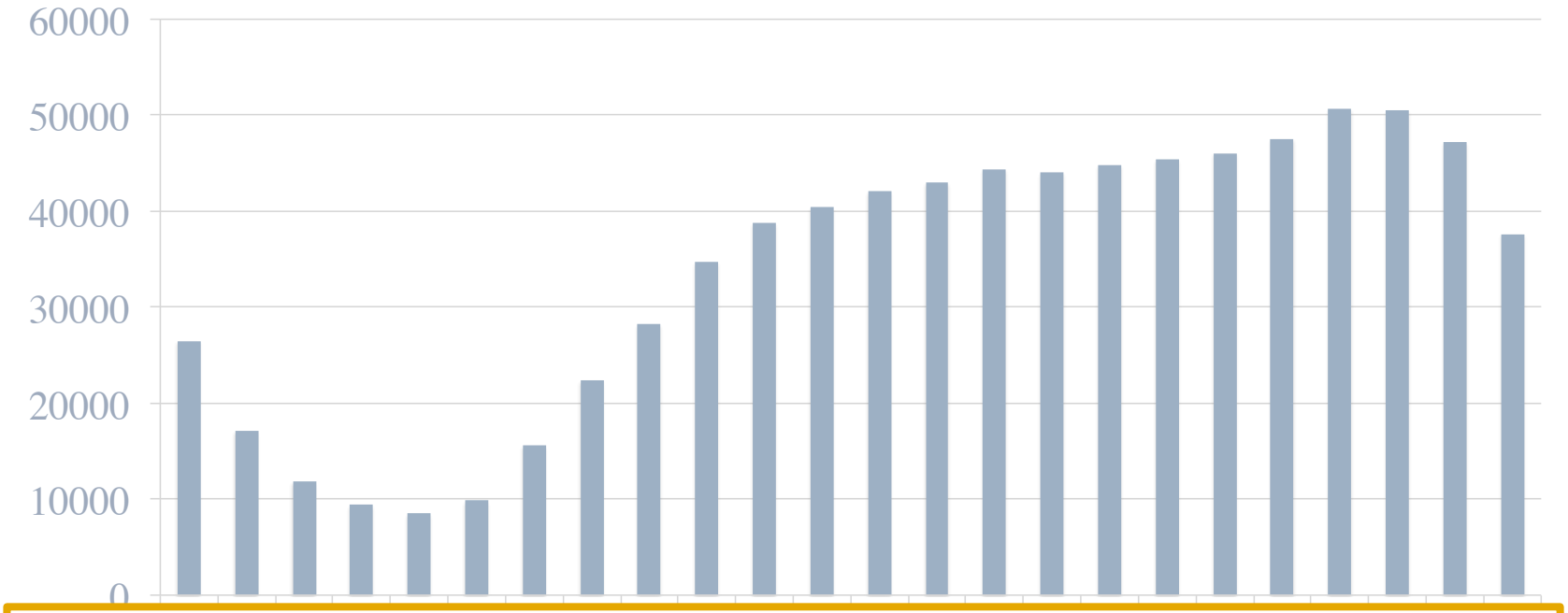
Joint work with Christopher Kyba, GFZ Potsdam

Image: Fotolia

Circadian rhythm

- Human schedules are determined by biological and social constraints
- Kantermann et al. (2007) showed that wake timing on free days tracks sunrise during standard time, but not during daylight savings time
- Sleep research is done in lab-based sleep studies or through surveys
- Roenneberg (2013) argues for much larger sample sizes of real-world data to improve our understanding of human sleep-wake patterns

Goal

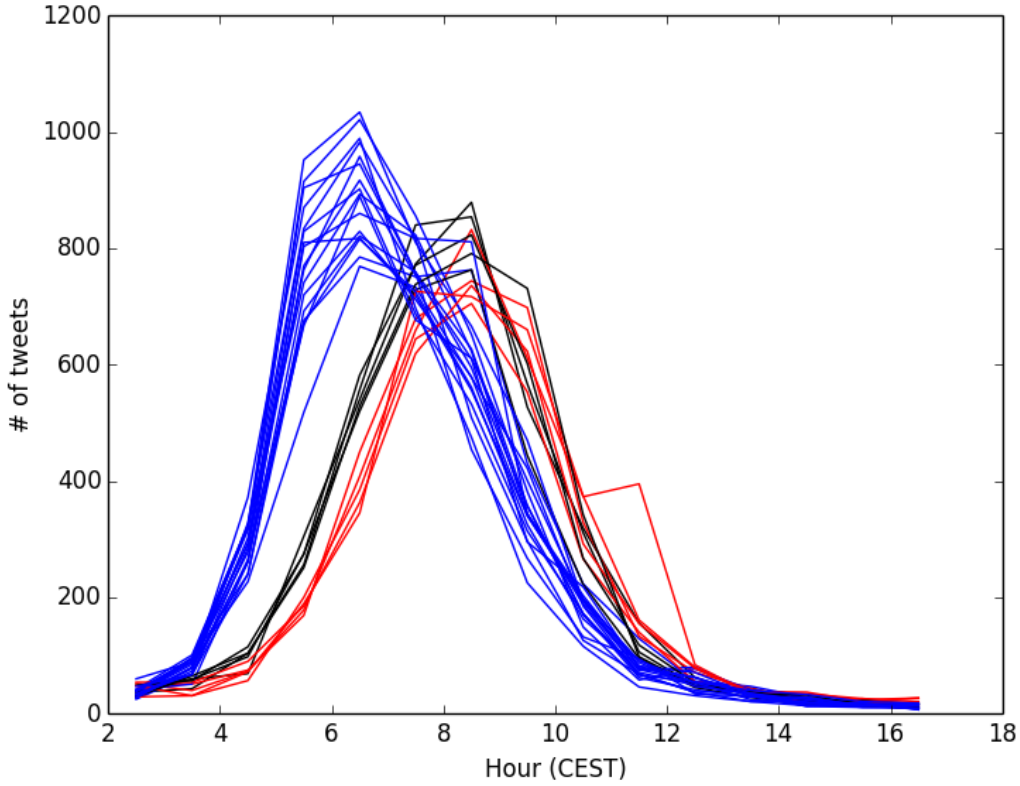


Use a Twitter time series to measure the sleep-wake pattern in humans and its interaction with DST.

Dataset

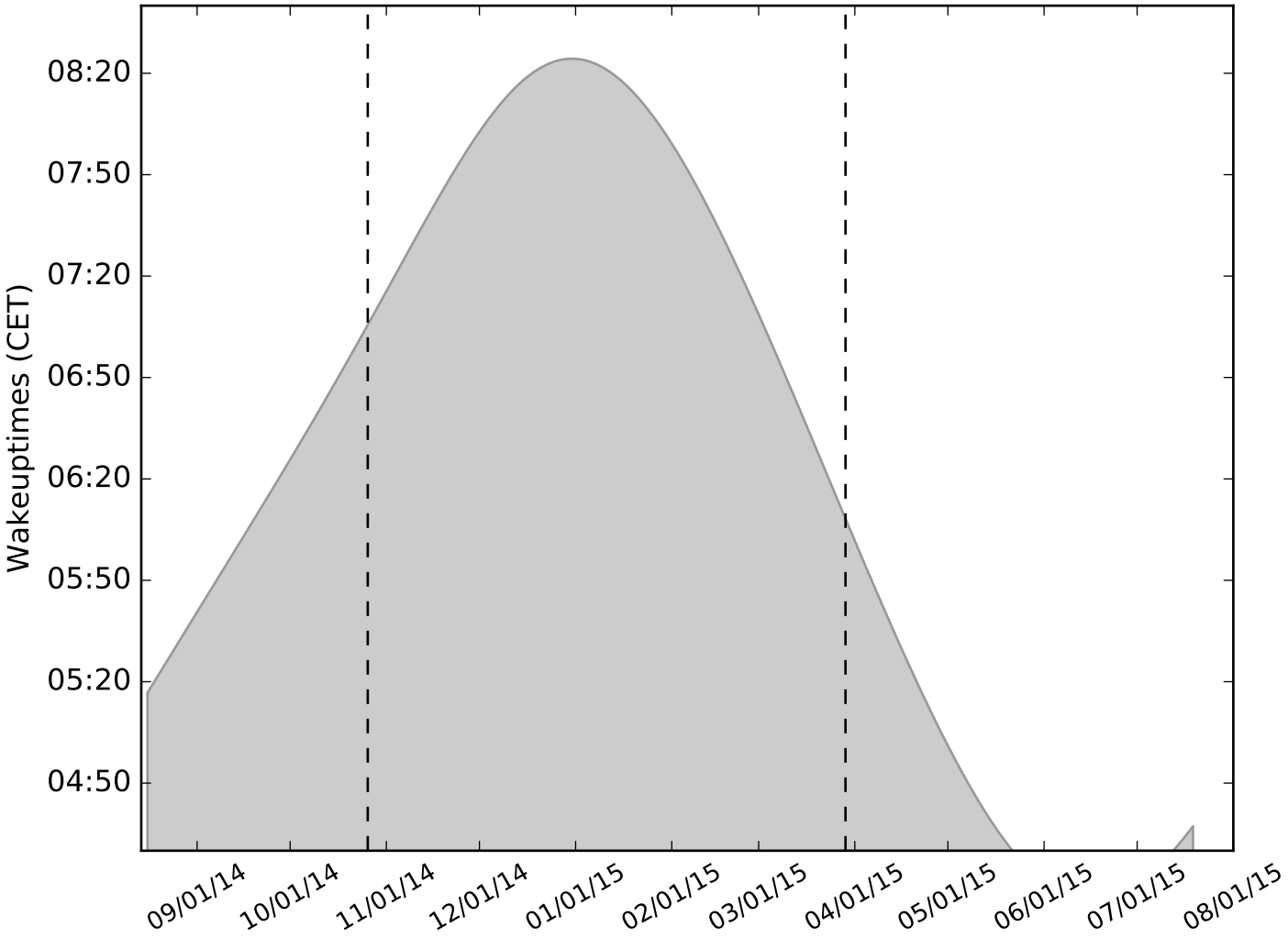
- one year of German tweets containing the phrase "guten morgen"
- August 15, 2014 to August 14, 2015
- dataset: 1,443,004 unique tweets from 206,633 individual users (retweets exluded)
- Tweets were binned by 15-minute windows

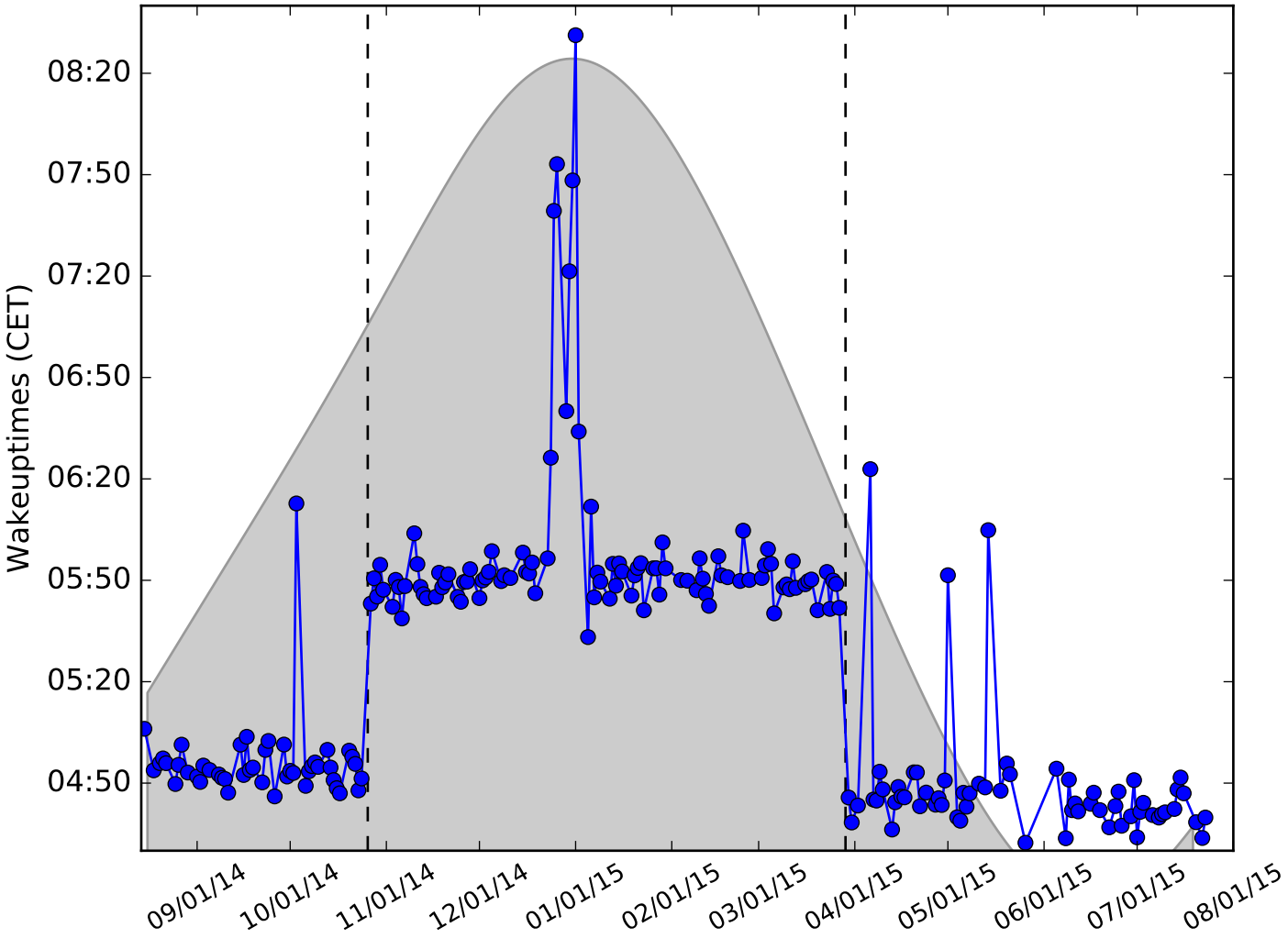
‘Good morning’-tweets

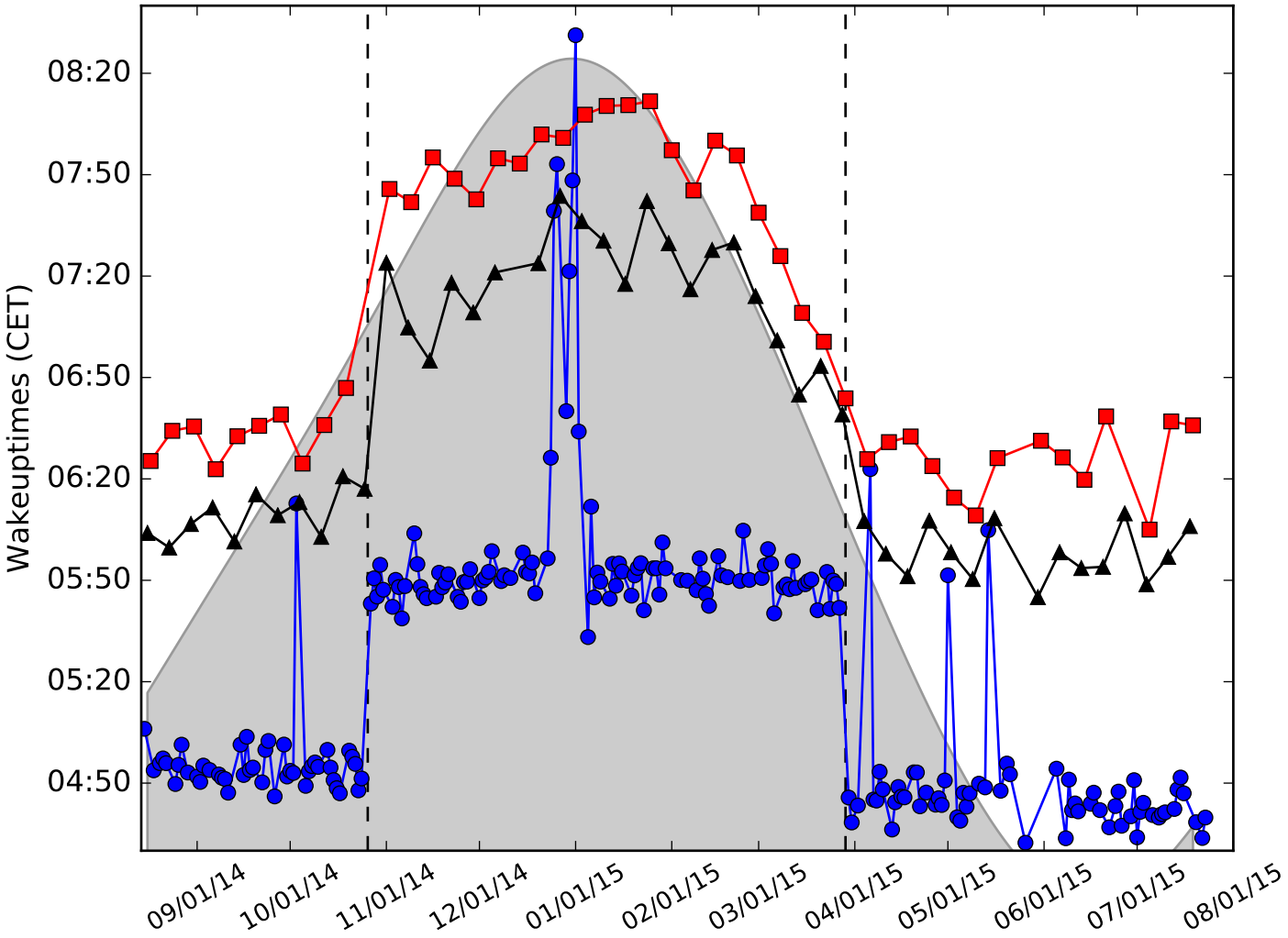


Analysis

- "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

General patterns

- morning greetings realistically reflect the onset of activity times (OTAT), between 4:30 and 8:30 local time on weekdays
- OTAT (determined geometrically) is related to the local sunrise and social time
- OTAT is much earlier on weekdays than on free days (public holidays resemble Saturdays)
- during the winter and standard time, OTAT on free days tracks dawn
- This relationship ends with the start of DST, and as a result the difference between wake times on free and work days grows considerably.

Interaction with social norms

We computed the difference in OTAT between work and free days for 42 weeks during the study period:

- Saturday: 79±14min
- Sunday: 103±25min
- weekday/weekend difference is largest in January (Saturday: ~99mins, Sunday: ~140mins)
- is smallest in Spring just up to the introduction of DST

Discussion

- Close tracking of dawn by OTAT on free days in Winter is disrupted by the introduction of DST.
- Large difference between Saturdays and Sundays
- Results are consistent with sleep survey data by Kantermann et al. (2007)
- Use of social media data in studying sleep patterns and other social norms?
- Implications for future policy change wrt. DST – could use social media to track effects?

3. Recovering user profile info

Demographic user data

- user profile:
  - language
  - location
  - name, description, profile image, …
- demographic information:
  - gender
  - age
  - location
  - income/profession
  - personality, languages, interests, …

What your language says about your occupation

- Hu et al., 2016
- Match data from different social media: Twitter + LinkedIn
- Twitter data used to represent styles, interests and personalities
- correlate linguistic info from Twitter and occupation info from LinkedIn

Data

Linguistic data:

- 3000 most recent tweets
- Twokenizer; remove too rare and too frequent n-grams

Occupation data:

- LDA topic model based on LinkedIn skill endorsements
- matrix of skill-job cluster relations
- matrix of person-job cluster relations

Jobs and skills

- 8 largest job clusters

Job Name	First Skill	Second Skill	Third Skill	Fourth Skill	Fifth Skill
1. Marketing	Digital MKTG	Social Media MKTG	Online MKTG	Digital Strategy	Advertising
2. Administrator	Public Speaking	Leadership	Fundraising	Event Planning	Coaching
3. Start-up	Start-ups	Entrepreneurship	Strategy	Business DEV	Management
4. Editer&Writer	Blogging	Editing	Journalism	Copy Editing	Storytelling
5. Software Engr	MySQL	CSS	JavaScript	PHP	jQuery
6. Office Clerk	Microsoft Office	Microsoft Excel	PowerPoint	Microsoft Word	Customer Service
7. Public Relation	Public Relations	Media Relations	Press Releases	Strategic COMM	Corporate COMM
8. Designer	Graphic Design	Web Design	Photography	Illustrator	Photoshop

HU, T.; XIAO, H.; LUO, J.; NGUYEN, T.. What the Language You Tweet Says About Your Occupation. International AAAI Conference on Web and Social Media, North America, mar. 2016.

Jobs and words

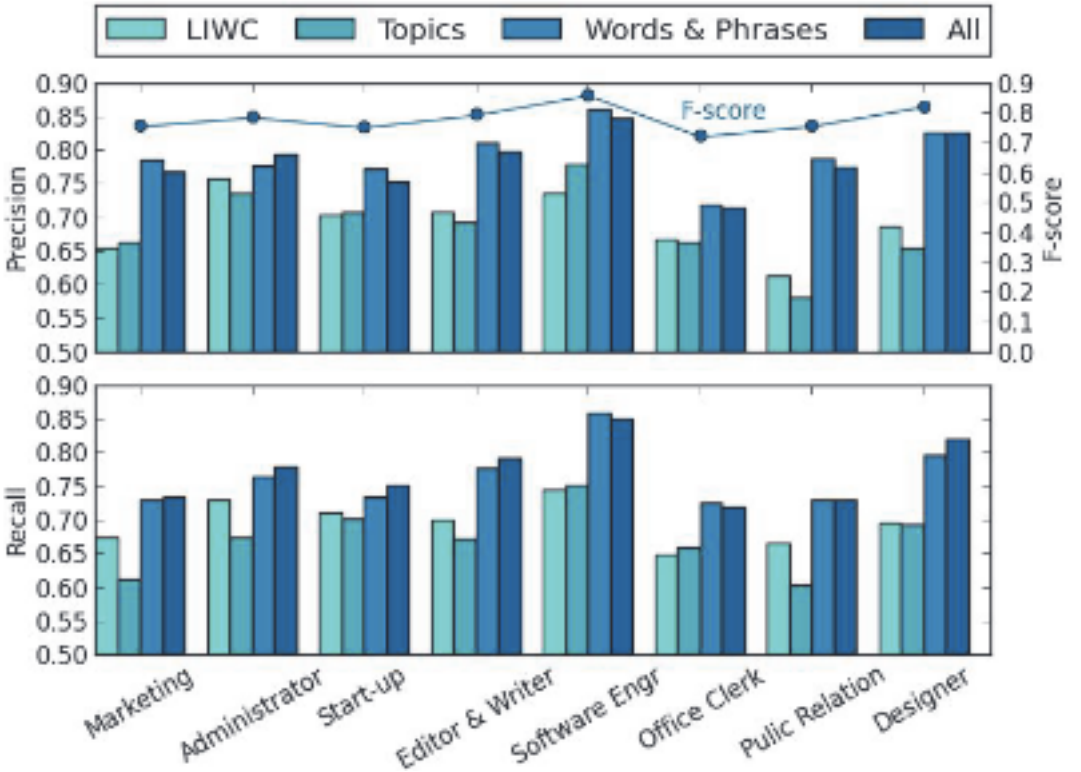
- most significant words by Pearson correlation of a word's TF-IDF weight for a certain job vs. others

(a) Word cloud for Editors&Writers

(b) Word cloud for Marketing

Predicting a job

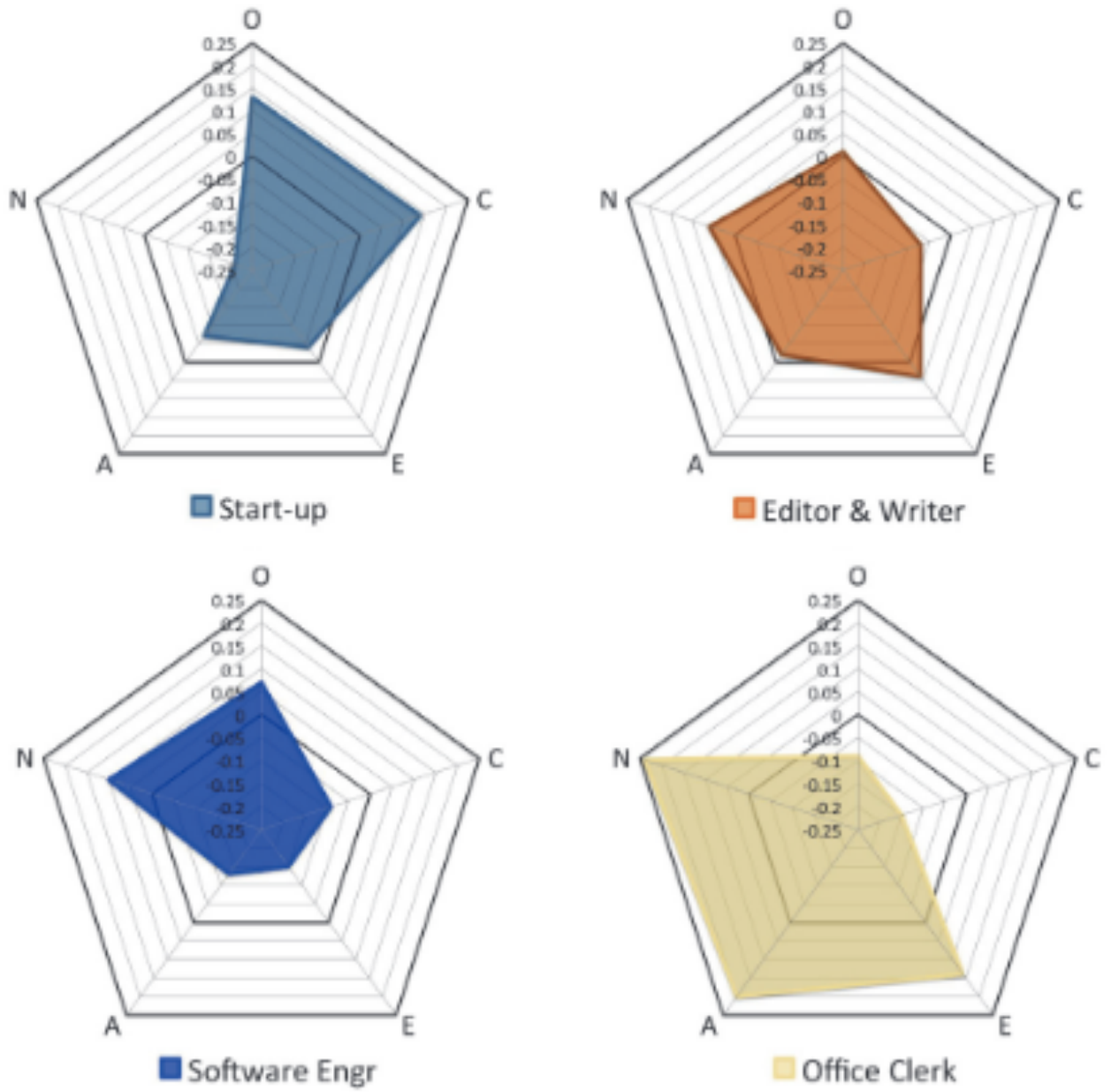
- 5-fold cross validation
- assign each user to the job whose weight is > 0.8
- overall F-score of all 8 jobs is 0.78
- tweet n-grams are very predictive for these occupations



HU, T.; XIAO, H.; LUO, J.; NGUYEN, T.. What the Language You Tweet Says About Your Occupation. International AAAI Conference on Web and Social Media, North America, mar. 2016.

Personality

- BM Watson Personality Insights service API: www.ibm.com/smarterplanet/us/en/ibmwatson/
- Personality traits based on a user's tweets
- Big Five:
  - Openness (to new things)
  - Conscientiousness (self-discipline)
  - Extraversion (social interactions)
  - Agreeableness (cooperativity, willing to compromise)
  - Neuroticism (instability of emotions)



HU, T.; XIAO, H.; LUO, J.; NGUYEN, T.. What the Language You Tweet Says About Your Occupation. International AAAI Conference on Web and Social Media, North America, mar. 2016.

Conclusions: Hu et al., 2016

- Combine data from different social media platforms
- Soft-clustering occupations to avoid uncertainty in self-reporting information
- Extract linguistic style information and Big Five personality traits from tweet text
- Pearson Correlation Coefficients used to explore differences in these properties across jobs

Summary Day 2

- Using metadata together with linguistic information
- Using linguistic information to recover metadata
- Social media data as a sensor for human behavior / real-world events

Coming up…

Wednesday: Sentiment (classification and clustering)

Thursday: Conversations and Discourse

Friday:

…. stay tuned!

Thank you.

tatjana.scheffler@uni-potsdam.de

Twitter Bots: to prepare…

1. Get a Twitter account!
2. Create a new Twitter application and receive consumer key and secret.
3. After the step above, you will be redirected to your app's page. Create an access token in the "Your access token" section.

If you want to get started:

Quick and easy Twitter bots: Make your own @HydrateBot

Python corpus based Twitter bots: Creative Twitter bots

Eisenstein J, B. O'Connor, N. Smith, and E.P. Xing. 2010. A latent variable model for geographic lexical variation. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1277–1287.

Gontrum J and Scheffler T. Text-based Geolocation of German Tweets. In: Proceedings of the NLP4CMC 2015 Workshop at GSCL, Duisburg-Essen, Germany. 2015.

HU, T.; XIAO, H.; LUO, J.; NGUYEN, T.. What the Language You Tweet Says About Your Occupation. International AAAI Conference on Web and Social Media, North America, mar. 2016. Available at: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13020/12738>.

Kantermann T, Juda M, Merrow M, Roenneberg T (2007) The human circadian clock's seasonal adjustment is disrupted by daylight saving time. Current Biology 17: 1996-2000.

Pavalanathan U. and J. Eisenstein. 2015. Confounds and Consequences in Geotagged Twitter Data.

Rios, M and Lin, J (2013) Visualizing the "pulse" of world cities on Twitter. In Proceedings of ICWSM, Boston, USA: AAAI.

Roenneberg, T (2013) Chronobiology: the human sleep project. Nature 498(7455):427–428.

Scheffler T (2014) A German Twitter snapshot, In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland: European Language Resources Association (ELRA).

Scheffler T, Kyba CCM (2016) Measuring Social Jetlag in Twitter Data, In Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016), Cologne, Germany: AAAI. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13080

… and references therein