Scalable Detection of Emerging Topics and Geo-spatial Events in Large Textual Streams

Erich Schubert[1,2], Michael Weiler[1], Hans-Peter Kriegel[1]

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

[1] Lehr- und Forschungseinheit Datenbanksysteme,
Ludwig-Maximilians-Universität München

[2] Lehrstuhl für Datenbanksysteme,
Ruprecht-Karls-Universität Heidelberg

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Lernen. Wissen. Daten. Analysen.
September 12–14, 2016, Potsdam, Deutschland

Scalable Detection of Emerging Topics

This presentation will summarize the following two publications:

E. Schubert, M. Weiler, and H.-P. Kriegel. "SigniTrend: Scalable Detection of Emerging Topics in Textual Streams by Hashed Significance Thresholds". In: Proceedings of the 20th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), New York, NY. 2014, pp. 871–880

E. Schubert, M. Weiler, and H.-P. Kriegel. "SPOTHOT: Scalable Detection of Geo-spatial Events in Large Textual Streams". In: Proceedings of the 28th International Conference on Scientific and Statistical Database Management (SSDBM), Budapest, Hungary. 2016, 8:1–8:12

SSDBM 2016 International Conference on Scientific and Statistical Database Management

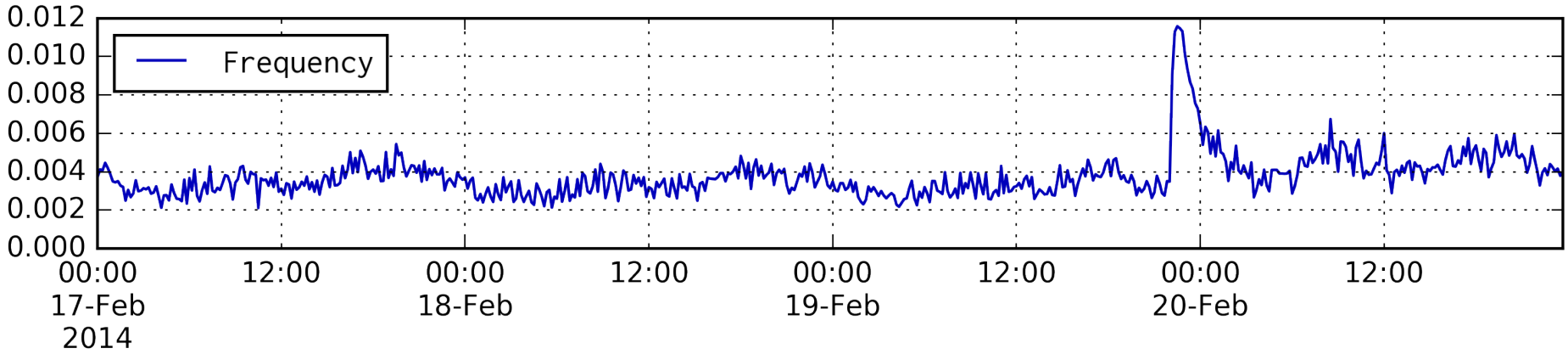
For details, please refer to these publications, and please ask!

Our Objective

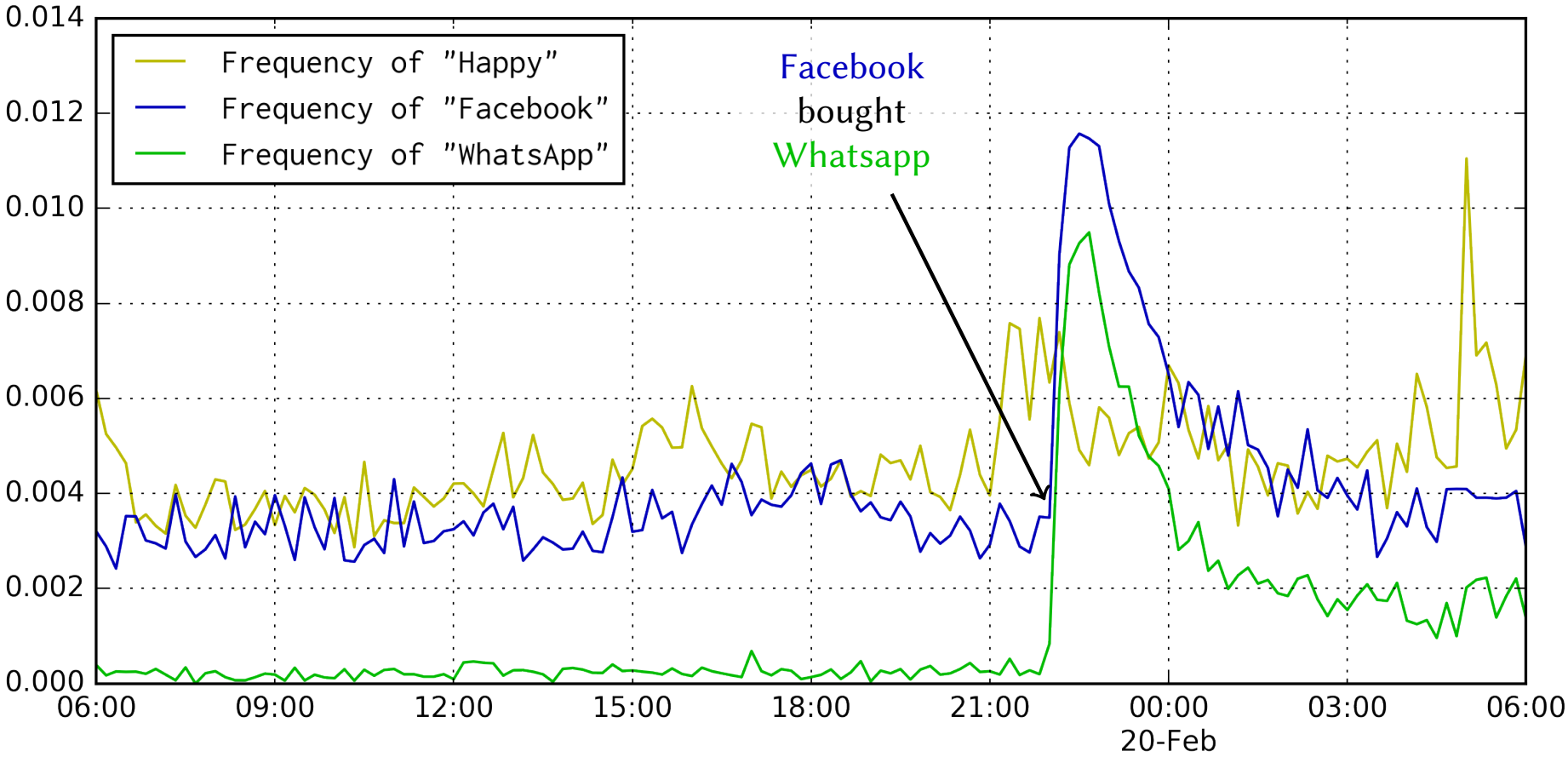
Scalable Detection of Emerging Topics and Geo-spatial Events

- Scalable: able to process years of news and Twitter data
- Detection: topics and keywords should not be defined beforehand
- Emerging: significant increase (c.f. “Trending Topics”)
- Topics: not every single message, but groups of related messages
- Geo-spatial Events: observe locality and detect geographic change

How do we find (and score) events such as this – at huge scale?



Motivation: Event Detection



Data: 1% Twitter sample, February 2014.

Objective: Detect such events without knowing the terms beforehand.

Limitations of Existing Approaches

- Often require terms to be specified beforehand
  (e.g. “Earthquake shakes Twitter users” [SOM10])
- Often only work on #hashtags (e.g. enBlogue [Alv+12])
- Often need to keep history in memory (e.g. EvenTweet [ASG13])
- Based on absolute increase in frequency (and thus can only detect events in very popular terms, e.g. TwitterMonitor [MK10])
- Cannot use geography, or observe only the top-k most popular places (e.g. GeoScope [Bud+13])
- Require multiple passes over the data
  (Most topic models – not applicable to large data streams)
- Will not scale to a billion tweets.

Key Ideas of our Solution

- From statistics: use exponentially weighted average + variance for detecting only significant change (contribution).
- From databases: Hashing and Count-Min sketches for scalability (contribution: "heavy hitters" for mean and variance).
- From computational linguistics: Word cooccurrences instead of single words for more meaningful results.
- From visualization: Word-cloud like visualization, but incorporating the co-trendiness of words (contribution).
- From data mining: Clustering of word pairs into simple "topics".
- Adjustment for rare words to reduce spurious events (contribution).
- Integration of geographic information:
  By mapping coordinates to tokens similar to text (contribution).

The big challenge is scalability to millions of words, word-pairs, and thousands of Tweets per second!

▸ Details on hashing for scalability

Significance via Moving Averages

For any word (and word pair), we monitor:

1. Moving average frequency (EWMA)
2. Moving variance (EWMVar)

▸ EWMA equations

We use exponentially weighted moving averages:

- Minimal memory requirement (two floats)
- Can be updated incrementally (based on [Fin09])
- Intuitive half-life time parameter

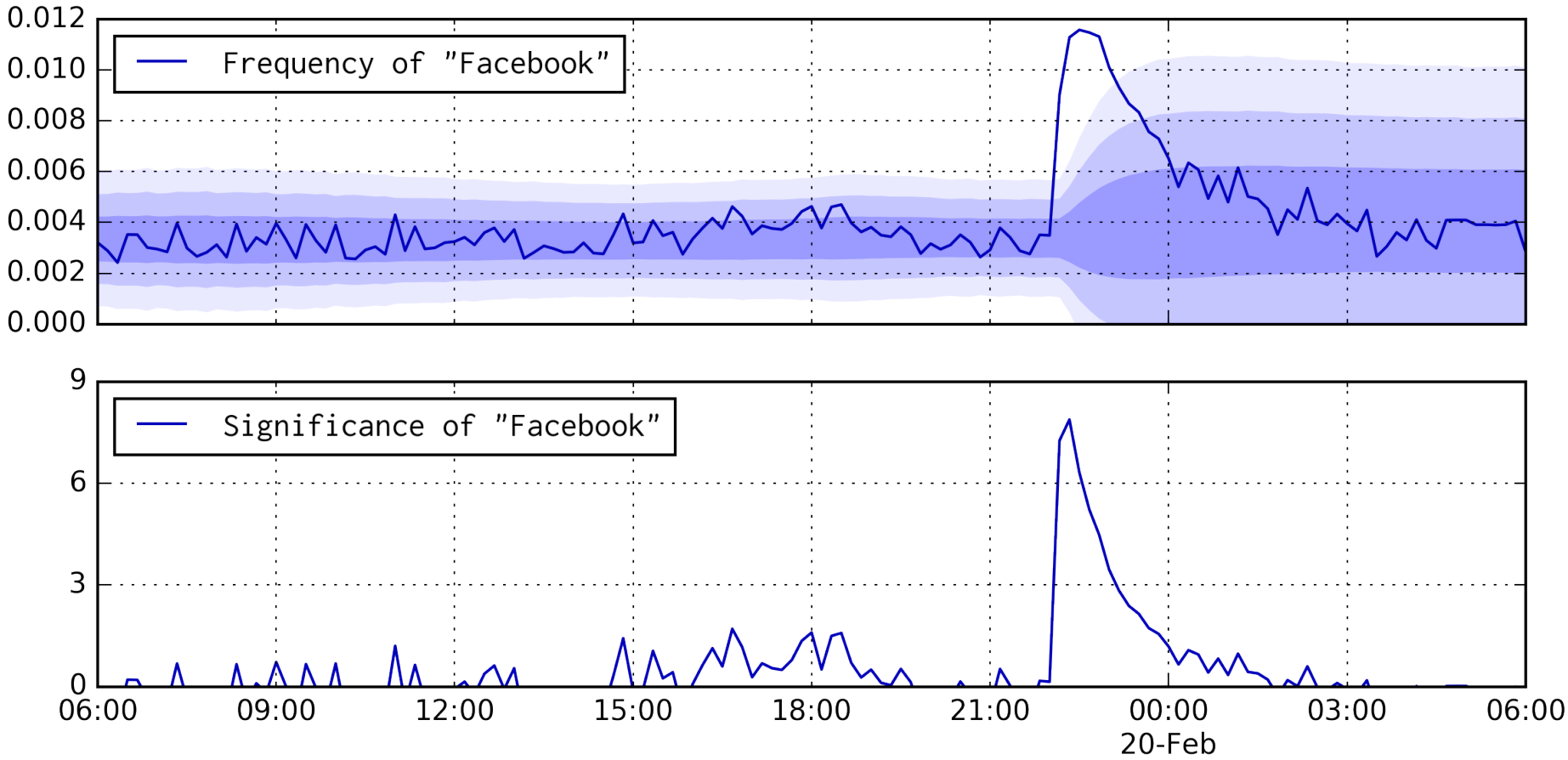
We get a z-score like significance score:

$$\mathrm{sig}_\beta(x) := \frac{x - \max\{\mathrm{EWMA}, \beta\}}{\sqrt{\mathrm{EWMVar}} + \beta}$$

Where β is a Laplace-like adjustment for unobserved occurrences.

"Only" need to scale this to all words and word pairs!

Example: Significance via Moving Averages



Modeling: Moving average and standard deviation.

Exponential aging (including exponential weighted standard deviation)

Hashing for Scalability

News and Twitter have millions of unique words (also typos, spam, ...).
Word-pairs further increase the number of time series that we need to track.

Related fixed-memory hashing based approaches are:

- Bloom filters [Blo70]
- Count-min sketches [CM05]

▸ Count-min example

Instead of bits (presence, Bloom filter), or integers (Count-min sketch), we store two floats for mean (EWMA) and variance (EWMVar).

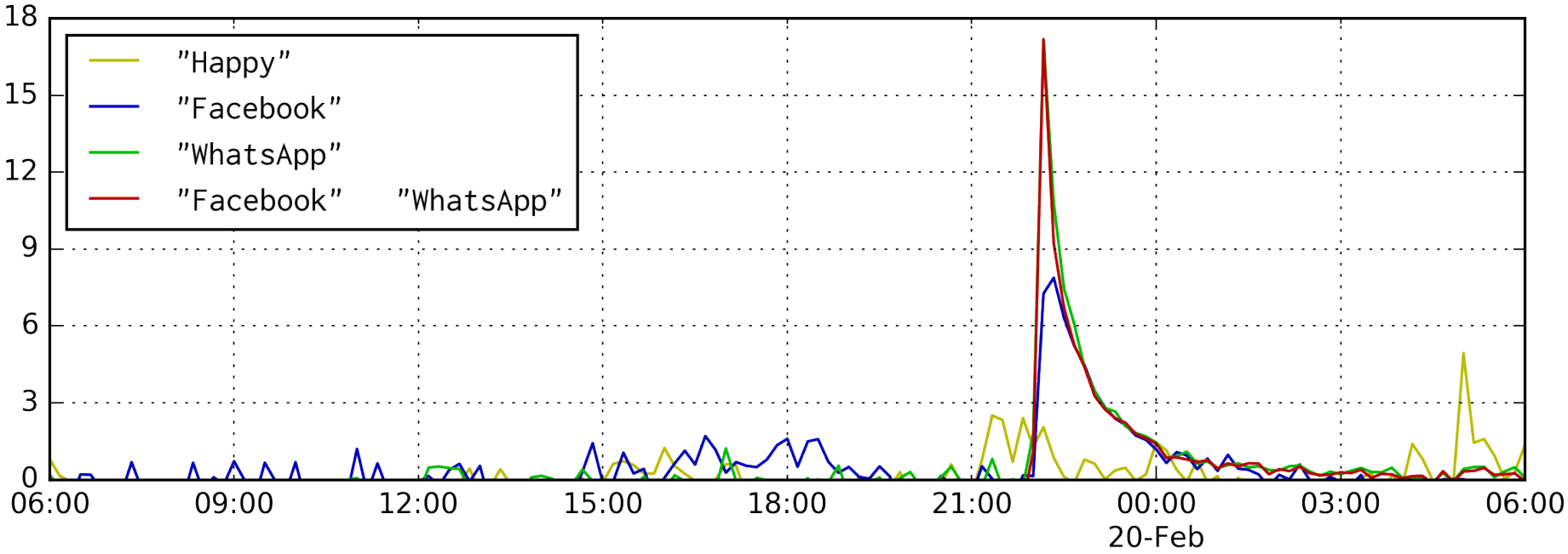
By using $h = 3$ hash functions and $2^{20} - 2^{22}$ buckets, we get very accurate estimates for frequent terms.

We overestimate rare terms, but if the frequency is less than β this does not effect event detection at all.

▸ Collision probabilities

Significance of Cooccurrences

Cooccurrences can be more significant than the individual words:



- The combination "Whatsapp" $\wedge$ "Facebook" is interesting!
- Facebook itself is less interesting (more background noise).
- "Happy Birthday" at midnight east coast – less significant.

Tracking all Word Cooccurrences

Why word cooccurrences and not just words?

Word combinations are interesting:

- "Facebook" bought "WhatsApp"
- Edward "Snowden" traveled to "Moscow"
- "Putin", "Obama" and "Merkel"
  — their interactions are more interesting than their frequency

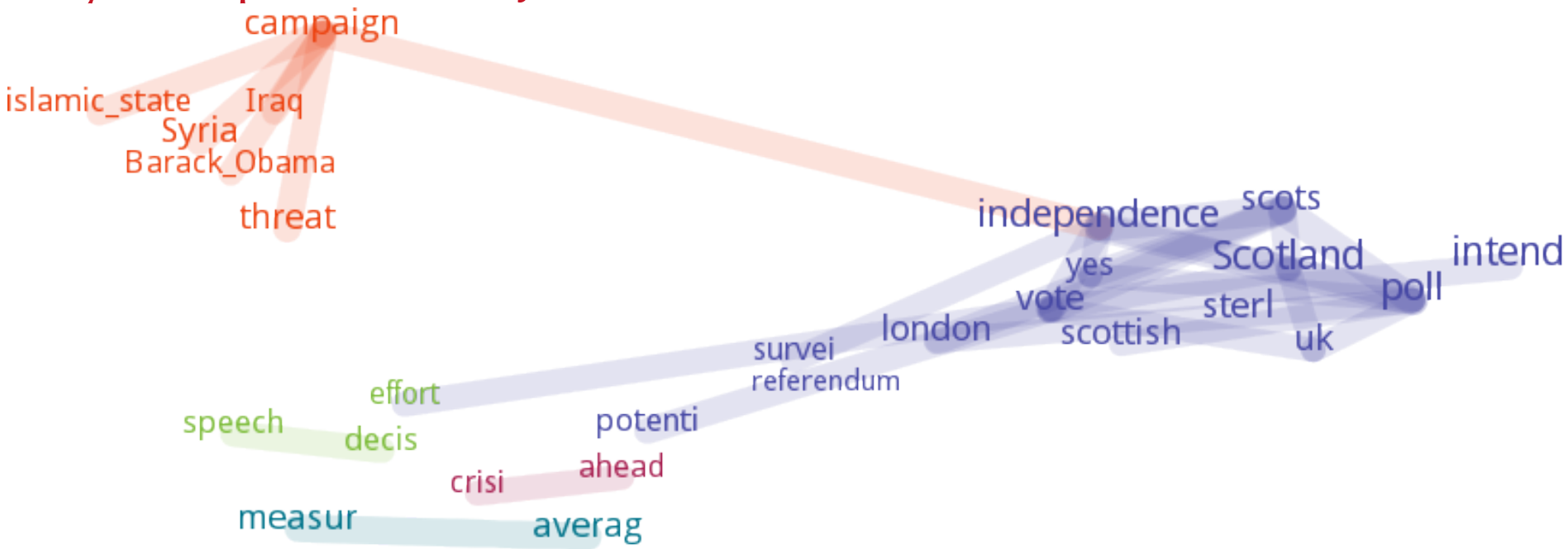
Why not the most popular terms?

Twitter is very biased:

- "@justinbieber" is always popular on Twitter
- Domain specific stopwords (e.g. "follow", "RT", "ILYSM")
- Cultural-, language- and geographic differences in usage

Tracking all Word Cooccurrences

Why word pairs and not just words?



Word relationships yields interesting structure

Uppercase or underscore: named entities,
Colors: clusters via hierarchical clustering,
Links: trending word pairs, Layout: MDS + spring graph

SigniTrend Examples

Explore online (best with a large screen):
`http://signi-trend.appspot.com/`

Top 10 events for news 2014 (chronological):

2014-03-08	Malaysia Airlines MH-370 missing in South China Sea
2014-04-17	Russia-Ukraine crisis escalates
2014-04-28	Soccer World Cup coverage: team lineups
2014-07-17	Malaysian Airlines MH-17 shot down over Ukraine
2014-07-18	Russian blamed for 298 dead in airline downing
2014-07-20	Israel shelling Gaza causes 40+ casualties in a day
2014-08-30	EU increases sanctions against Russia
2014-10-22	Ottawa parliament shooting
2014-11-05	U.S. mid-term elections
2014-12-17	U.S. and Cuba relations improve unexpectedly

Geo-spatial Event Detection

Our SigniTrend [SWK14] approach can answer

- What is the event (token combinations)
- When is the event (first significant occurrence)

In SPOTHOT [SWK16], we added the ability to answer Where,
and to detect a change in geography.

For example there is always a "concert" or "earthquake" somewhere,
so this word is not significant in the full data set.

Within a limited geographical context (e.g. city or state),
we may see a locally significant "concert".

This can also normalize to geographic differences in Twitter usage.

Integrating Geographic Information as Text

SigniTrend is designed for text, but can process arbitrary tokens.

- Named entities (e.g. Barack Obama)
- #hashtags and @usermentions
- Emoticons and Emojis
- URLs
- Location?

Integrating Geographic Information as Text

SigniTrend is designed for text, but can process arbitrary tokens.

- Named entities (e.g. Barack Obama)
- #hashtags and @usermentions
- Emoticons and Emojis
- URLs
- Location?

For this, we need a function

$$(\text{longitude}, \text{latitude}) \quad \rightarrow \quad \text{Symbol}$$

such that nearby locations produce the same symbol.

Integrating Geographic Information as Text

SigniTrend is designed for text, but can process arbitrary tokens.

- Named entities (e.g. Barack Obama)
- #hashtags and @usermentions
- Emoticons and Emojis
- URLs
- Location?

For this, we need a function

$$(\text{longitude}, \text{latitude}) \quad \rightarrow \quad \{\text{Symbol}, \ldots\}$$

such that nearby locations produce the same symbol.
Better results with multiple symbols at different resolution!

Tokenization with Geographic Information

Token generation example:

Presenting a novel event detection method at #SSDBM2016 in Budapest :-)

present novel event_detection method #ssdbm2016 Q1781:Budapest :)

(stem) (stop) (entity) (stop) (normalized) (stop) (entity) (norm.)

47.5323 19.0530

!geo0!46!18 !geo1!48!18 !geo2!48!20

(Overlapping grid cells)

!geo!Budapest !geo!Budapesti_kistérség !geo!Közép-Magyarország !geo!Hungary

(Hierarchical semantic location information)

We can now use a SigniTrend approach to detect frequent pairs:

$$(\texttt{!geo!Budapest}, \texttt{\#ssdbm2016})$$

Grid: three overlapping grids for worst-case guarantees [Cha98]. ▸ Details

Administrative boundaries from OpenStreetMap. ▸ Details

(Source code: `https://github.com/kno10/reversegeocode`)

Data Set for Geography Experiments

- 5–6 million geo-tagged tweets per day (no retweets!)
- Estimated 1/3rd of all geo-tagged tweets
- September 10, 2014 to February 19, 2015
- Over 1.1 billion tweets

Selected top geographies:

Region	Mil.	Share
United States	287.7	25.4%
Brazil	165.6	14.6%
Argentina	73.6	6.5%
Indonesia	72.0	6.4%
Turkey	59.3	5.2%
Japan	52.4	4.6%
United Kingdom	49.3	4.4%
⋮		

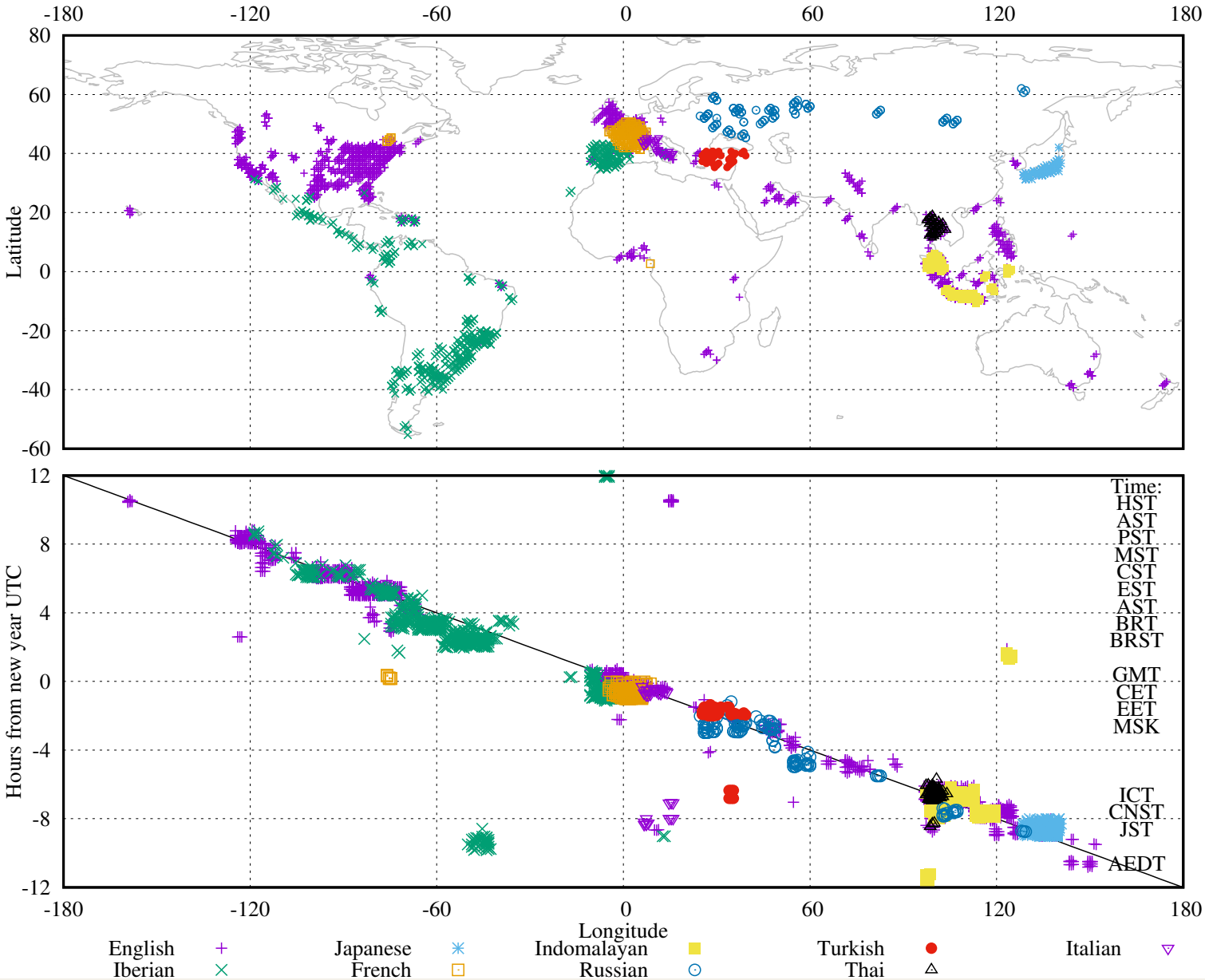
Region	Mil.	Share
London	7.6	0.67%
New York City	7.5	0.66%
Tokyo	7.4	0.66%
⋮		
Germany	3.5	0.31%
⋮		
Berlin	0.5	0.05%
⋮		

Results – Most Significant Geo-located Events

The most significant words in the most significant locations only:

σ	Time	Word	Location	Explanation
2001.8	2014-10-29 00:59	#voteluantvz	Brazil	Brazilian Music Award 2014
727.8	2014-09-23 02:21	allahımsenbüyüksün	Denizli (Turkey)	Portmanteau used in spam wave
550.1	2015-02-02 01:32	Missy_Elliott	United States of America	Super Bowl Halftime Show
413.5	2014-09-18 21:29	#gala1gh15	Spain	Spanish Big Brother Launch
412.2	2014-11-11 19:29	#murrayftw	Italy	Teen idol triggered follow spree
293.8	2014-10-21 12:05	#tarıkgüneşttyapıyor	Marmara Region	Hashtag used in spam wave
271.2	2015-02-02 02:28	#masterchefgranfinal	Chile	MasterChef Chile final
268.1	2015-01-30 19:28	#سباركيز	Saudi Arabia	Amusement park “Sparky’s”
257.7	2014-11-16 21:44	gemma	United Kingdom	Gemma Collins at jungle camp opening
249.1	2014-10-08 02:56	rosmeri	Argentina	Rosmery González joined Bailando 2014
223.1	2015-01-21 18:51	otortfv	Central Anatolia Region	Keyword used in spam wave
212.7	2014-09-11 18:58	#catalansvote9n	Catalonia	Catalan referendum requests
208.4	2014-12-02 20:00	#cengizhangençtürk	Northern Borders Region	Hashtag used in spam wave
205.3	2015-01-04 15:56	hairul	Malaysia	Hairul Azreen, Fear Factor Malaysia
198.7	2014-12-31 15:49	あけましておめでとうございます	Japan	New Year in Japan
198.5	2015-01-10 20:19	ВК	Russian Federation	“Russian Facebook” VK unavailable
179.7	2014-10-04 16:28	#hormonestheseries2	Thailand	Hormones: The Series Season 2
174.7	2014-11-28 21:29	chespirito	Mexico	Comedian “Chespirito” died
160.9	2014-09-21 21:27	#ss5	Portugal	Secret Story 5 Portugal launch
157.3	2014-09-24 01:57	maluma	Colombia	Maluma on The Voice Kids Colombia

Results – New Year Around the World



Top events we could match to WikiTimes [TA14]

Date σ	Event Term Cluster (!geo! omitted) Event description from Wikipedia, The Free Encyclopedia
09-18 25.6	Scotland, United_Kingdom, uk, England, Greater_London, London, David_Cameron Prime Minister David Cameron announces plans to devolve further powers to Scotland, as well as the UK's other constituent countries.
09-18 15.0	England, referendum, Greater_London, United_Kingdom, Alex_Salmond, Scotland, resign, London, salmond, Glasgow_City Alex Salmond announces his resignation as First Minister of Scotland and leader of the Scottish National Party following the referendum.
09-22 40.1	Isis, U_S_A, Syria, airstrikes, bomb, target, islamic_state, u_s, strike, air The United States and its allies commence air strikes against Islamic State in Syria with reports of at least 120 deaths.
09-23 17.7	Syria, strike, air, Isis The al-Nusra Front claims its leader Abu Yousef al-Turki was killed in air strikes.
10-08 60.5	di, patient, thoma, duncan, eric, dallas, hospital, diagnos, texas The first person who was diagnosed with Ebola in the United States, Thomas Eric Duncan, a Liberian man, dies in Dallas, Texas.
10-10 44.7	kailash, satyarthi, India, Nobel_Peace_Prize, malala, Malala_Yousafzai, congratul, #nobelpeaceprize, indian, pakistani, peace Pakistani child education activist Malala Yousafzai and Indian children's rights advocate Kailash Satyarthi share the 2014 Nobel Peace Prize.
10-14 34.4	Republic_of_Ireland, ireland, United_Kingdom, Germany, England, John_O'Shea, Leinster, County_Dublin, Scotland Ireland stuns world champion Germany in Gelsenkirchen, with Ireland drawing the match at 1–1 when John O'Shea scores in stoppage time.
10-14 30.6	Albania, Serbia, United_Kingdom, England, London, match, drone, flag The game between Albania and Serbia is abandoned after a drone carrying a flag promoting the concept of Greater Albania descends onto the pitch in Belgrade, sparking riots, mass brawling and an explosion.
10-15 17.8	posit, worker, tests, Ebola_virus_disease, texas, health A second health worker tests positive for the Ebola virus in Dallas, Texas.
10-22 26.4	soldier, Canada, Ottawa, shoot, Ontario, insid, canada, parliament A gunman shoots a Canadian Forces soldier outside the Canadian National War Memorial.

Thank you!

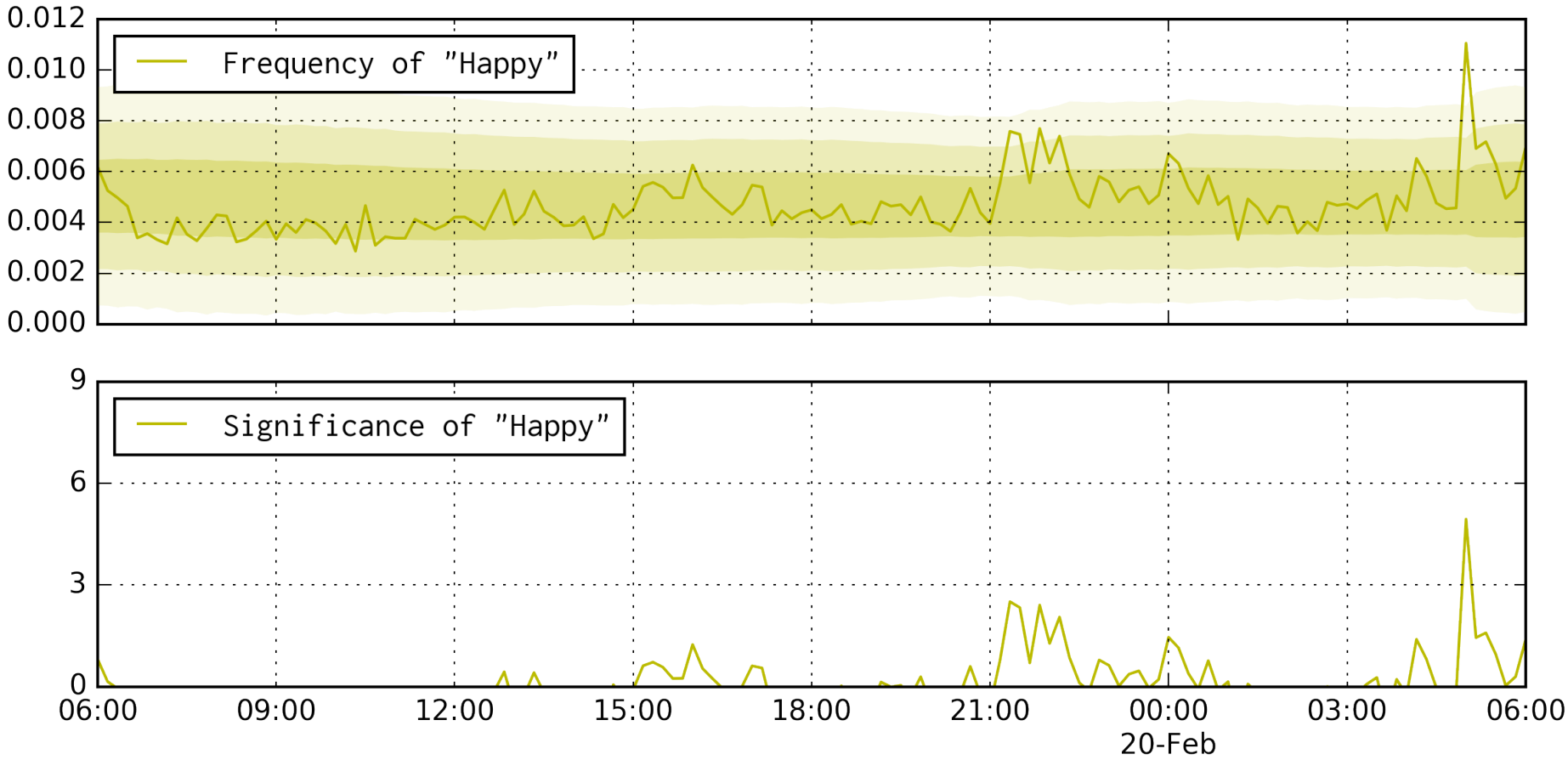
Questions & Discussion

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Outline

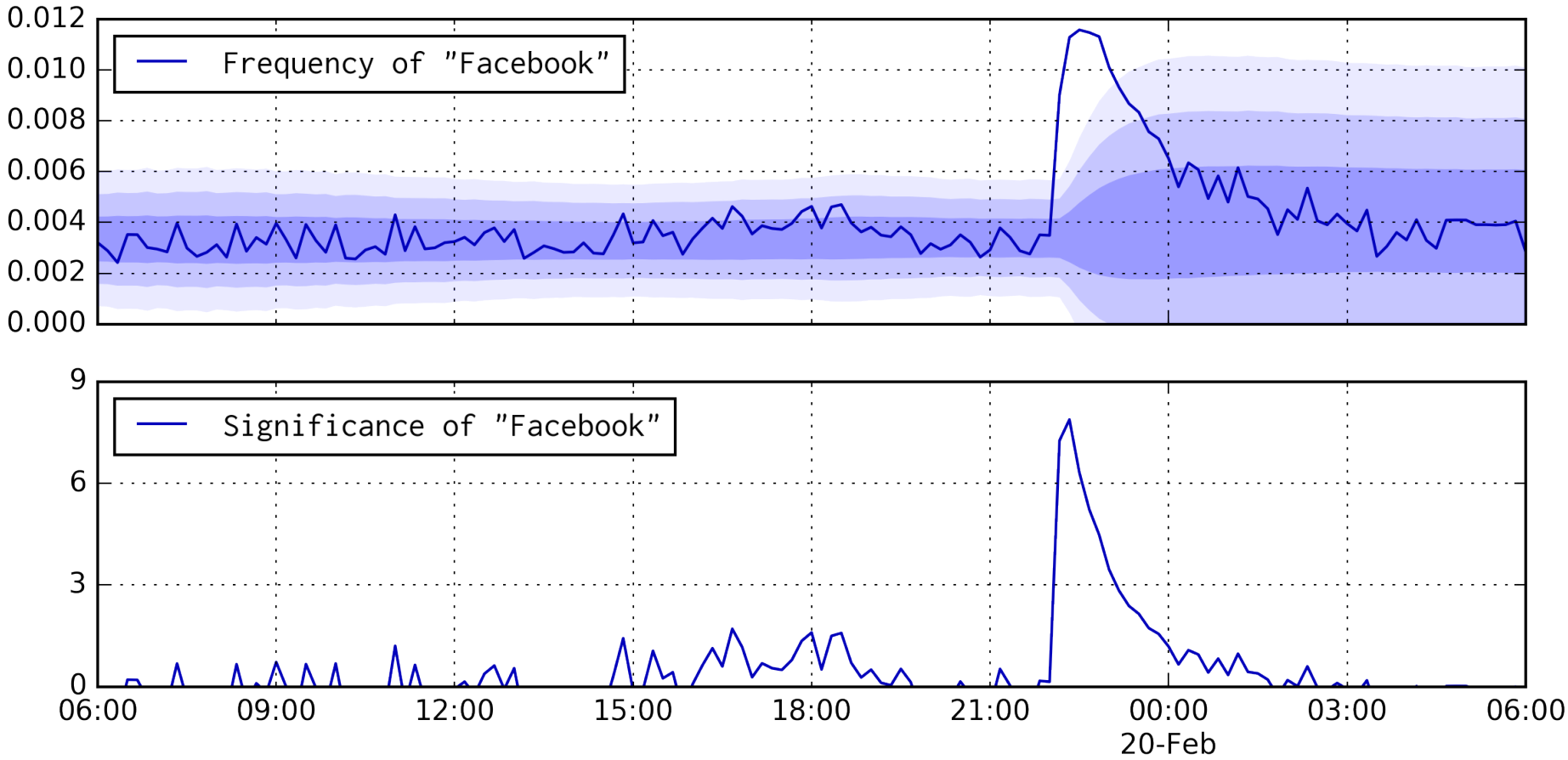
Motivation: Event Detection



Moving standard deviation normalizes for higher variance.

Return

Motivation: Event Detection



Moving standard deviation normalizes for higher variance.

Return

Motivation: Scale

Counting is still possible with 32 GB RAM – but is it interesting?

Data set	News 2013	Twitter	StackOverflow
Documents	424,704	94,127,149	5,932,320
Paragraphs	5,867,457	94,127,149	30,423,831
Unique Words	300,141	25,581,022	2,040,932
Total Words	56,661,782	245,140,695	138,205,636
Unique Pairs	71,289,359	179,105,233	91,460,397
Total Pairs	660,430,059	473,871,456	545,570,530

* These statistics include year 2013 of two news agencies; 114 days of 1% of Twitter.

1 year of 1% of Twitter uncompressed JSON is "just" around 15 TB.

Do we need to count everything, or can we accept errors?

We will try to make errors in a way that quality does not degrade!

Scalability via Hashing

Similar to Bloom filters [Blo70] and Count-min sketches [CM05] we use multiple hash functions and accept collisions.

1. Count all occurrences in a (small) time window.
2. Hash counts into one table, keeping the maximum only. (Using multiple hash functions, as in Bloom filters)
3. Normalize by the number of documents.
4. Update mean and variance estimates (in each bucket).
5. Predict new frequency (mean and standard deviation).

Event detection:

Report events if the observed frequency is more than τ standard deviations more than the expected value (from the previous iteration).

Estimate expected frequency using the minimum of all buckets (cf. Count-min sketch) and the associated variance.

Return

Incremental EWMA and EWMVar

Incremental updating (based on weighted variance [Fin09]):

$$\Delta \leftarrow x - \mathrm{EWMA}$$
$$\mathrm{EWMA} \leftarrow \mathrm{EWMA} + \alpha \cdot \Delta$$
$$\mathrm{EWMVar} \leftarrow (1-\alpha) \cdot (\mathrm{EWMVar} + \alpha \cdot \Delta^2)$$

Learning rate α can easily be set using the half-life time $t_{1/2}$:

$$\alpha_{\text{half-life}} = 1 - \exp\left(\log\left(\tfrac{1}{2}\right)/t_{1/2}\right)$$

Significance is then measured as z-score:

$$\mathrm{sig}_\beta(x) := \frac{x - \max\{\mathrm{EWMA}, \beta\}}{\sqrt{\mathrm{EWMVar}} + \beta}$$

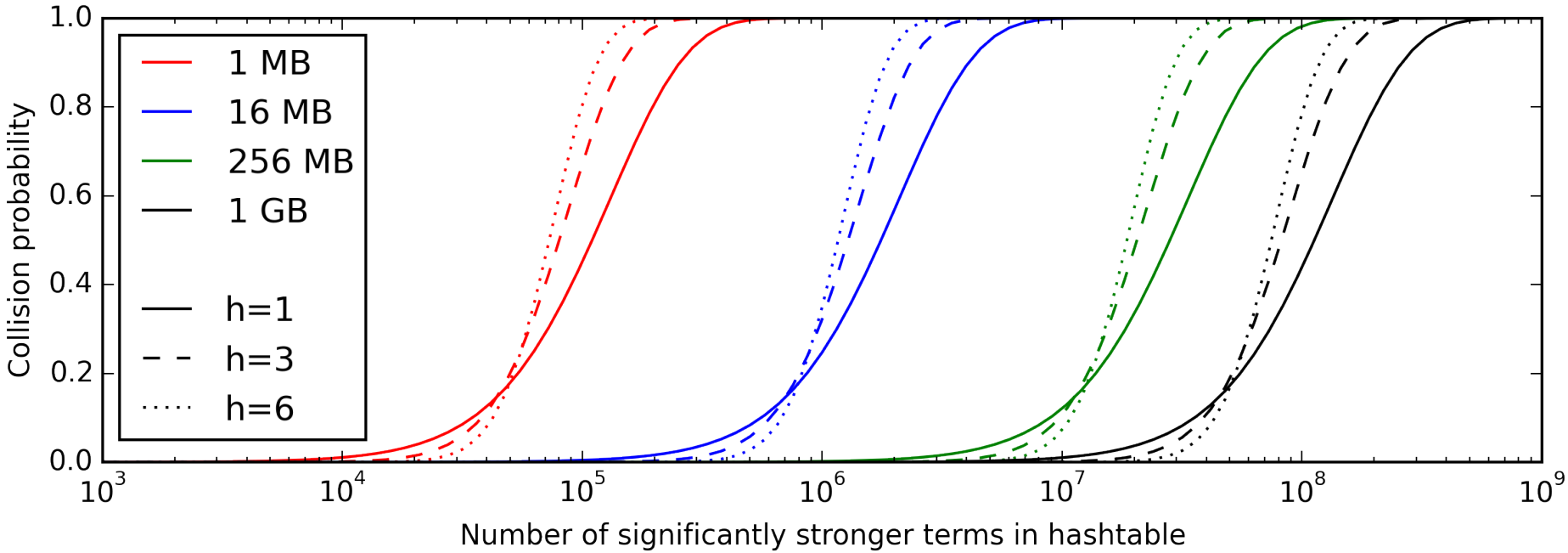
with a small correction term β (similar to Laplacian correction).

Return

Do we make errors?

Yes. This is intentional: to save memory – no free lunch!

Errors happen if we have h collisions with much more frequent terms.



We can track the top 1,000,000 w.h.p. with 256 MB of RAM.
(Also verified experimentally: we observed saturatation of recall with 22–24 bits when simulating artificial trends in text streams)

Return

Count-min Sketch [CM05]

Counting Bloom Filter:

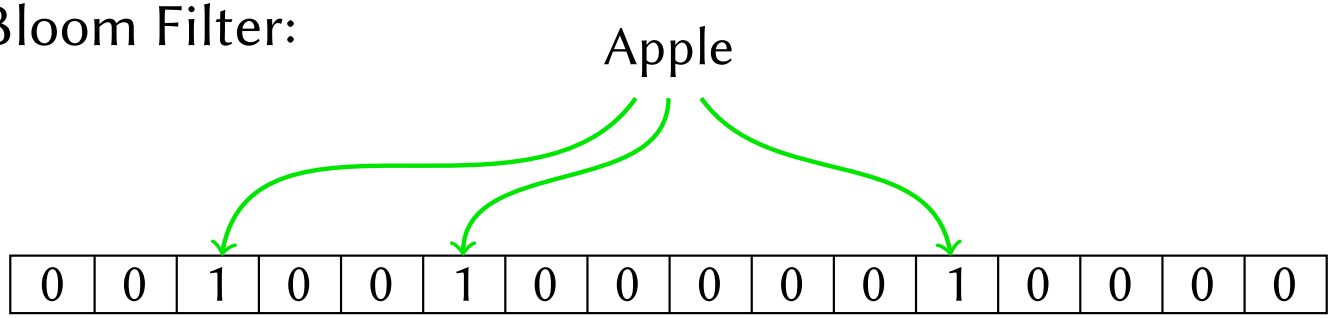
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Note: [CM05] uses a seprate table for each hash function.

◂ Back to method

Count-min Sketch [CM05]

Counting Bloom Filter:

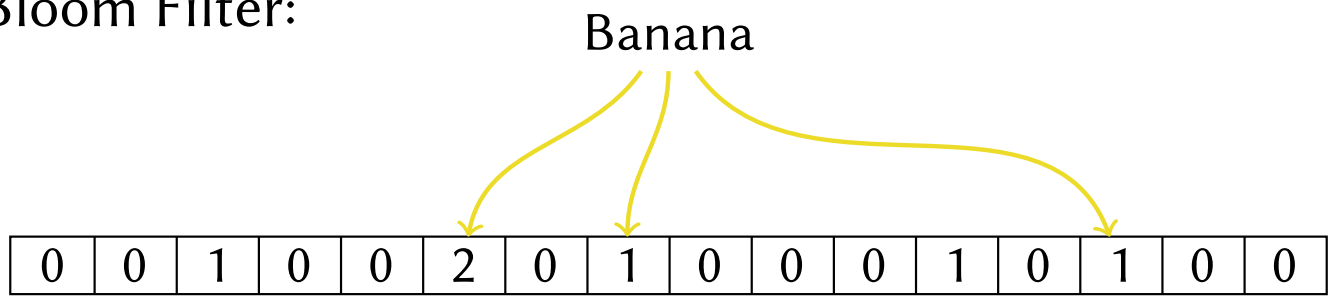


Note: [CM05] uses a seprate table for each hash function.

◂ Back to method

Count-min Sketch [CM05]

Counting Bloom Filter:

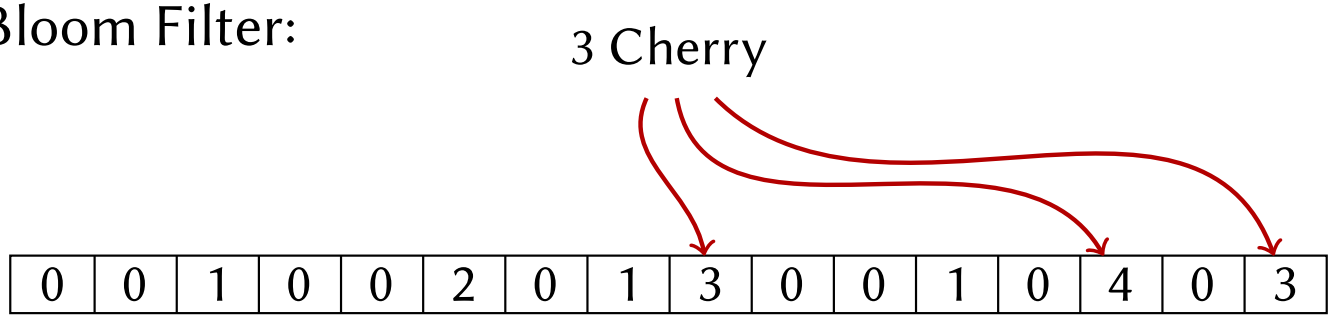


Note: [CM05] uses a seprate table for each hash function.

◂ Back to method

Count-min Sketch [CM05]

Counting Bloom Filter:

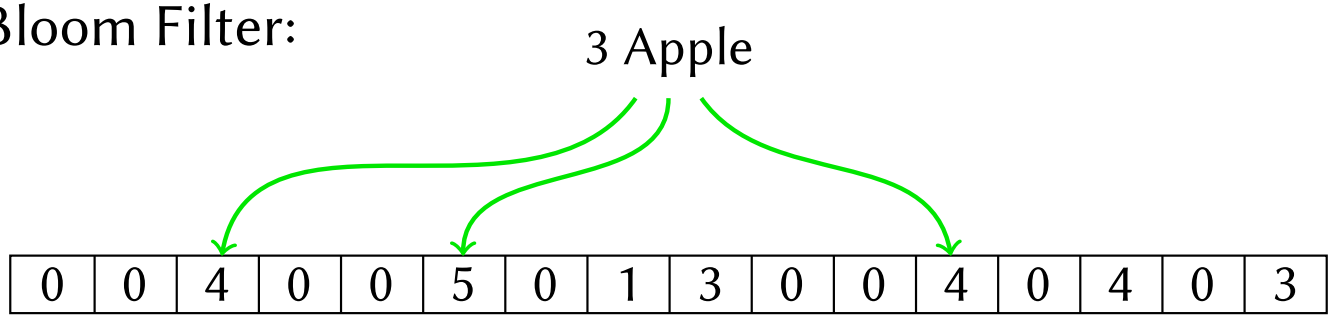


Note: [CM05] uses a seprate table for each hash function.

◂ Back to method

Count-min Sketch [CM05]

Counting Bloom Filter:

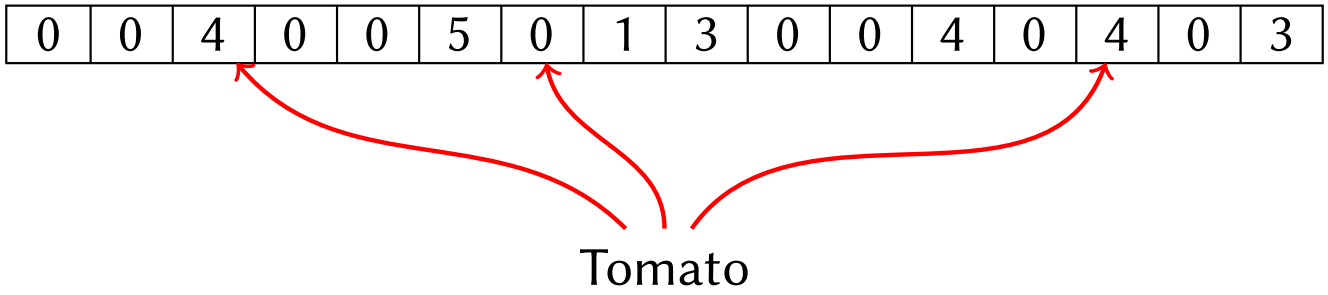


Note: [CM05] uses a seprate table for each hash function.

◂ Back to method

Count-min Sketch [CM05]

Counting Bloom Filter:



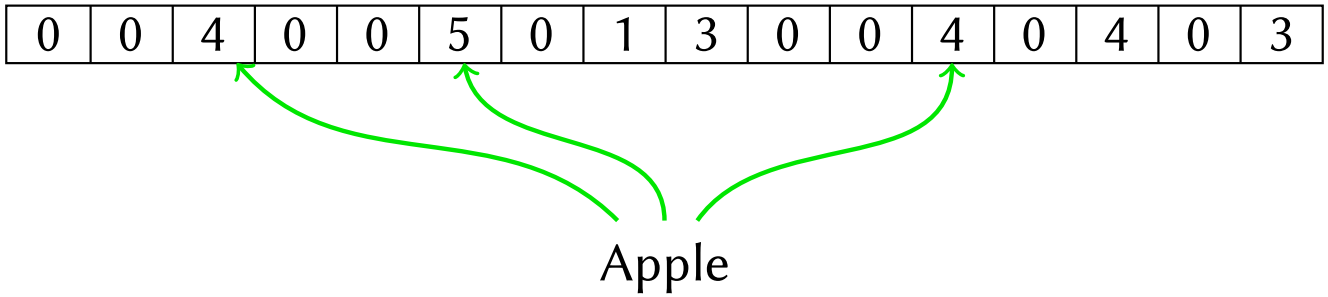
$$\min\{4, 0, 4\} = 0$$

Note: [CM05] uses a seprate table for each hash function.

◂ Back to method

Count-min Sketch [CM05]

Counting Bloom Filter:



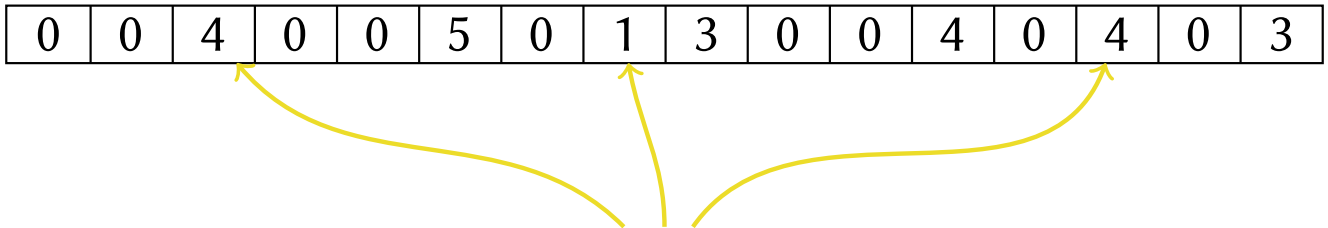
$$\min\{4, 5, 4\} = 4$$

Note: [CM05] uses a seprate table for each hash function.

◂ Back to method

Count-min Sketch [CM05]

Counting Bloom Filter:



$$\min\{4, 1, 4\} = 1$$

Count-min can overestimate.
But the most frequent terms are “with high probability” correct.

Note: [CM05] uses a seprate table for each hash function.

◂ Back to method

Mapping Location to Text

Lookup (reverse geocode) coordinates to a hierarchy of names:

Local Name	International	Wikipedia	Wikidata	OpenStreetMap ID
München	Munich	de:München	Q1726	r62428
Oberbayern	Upper Bavaria	de:Oberbayern	Q10562	r2145274
Bayern	Free State of Bavaria	de:Bayern	Q980	r2145268
Deutschland	Germany	de:Deutschland	Q183	r51477

Generate text-like tokens: ‘‘!geo!München’’, ‘‘!geo!Oberbayern’’, ‘‘!geo!Bayern’’, ‘‘!geo!Deutschland’’

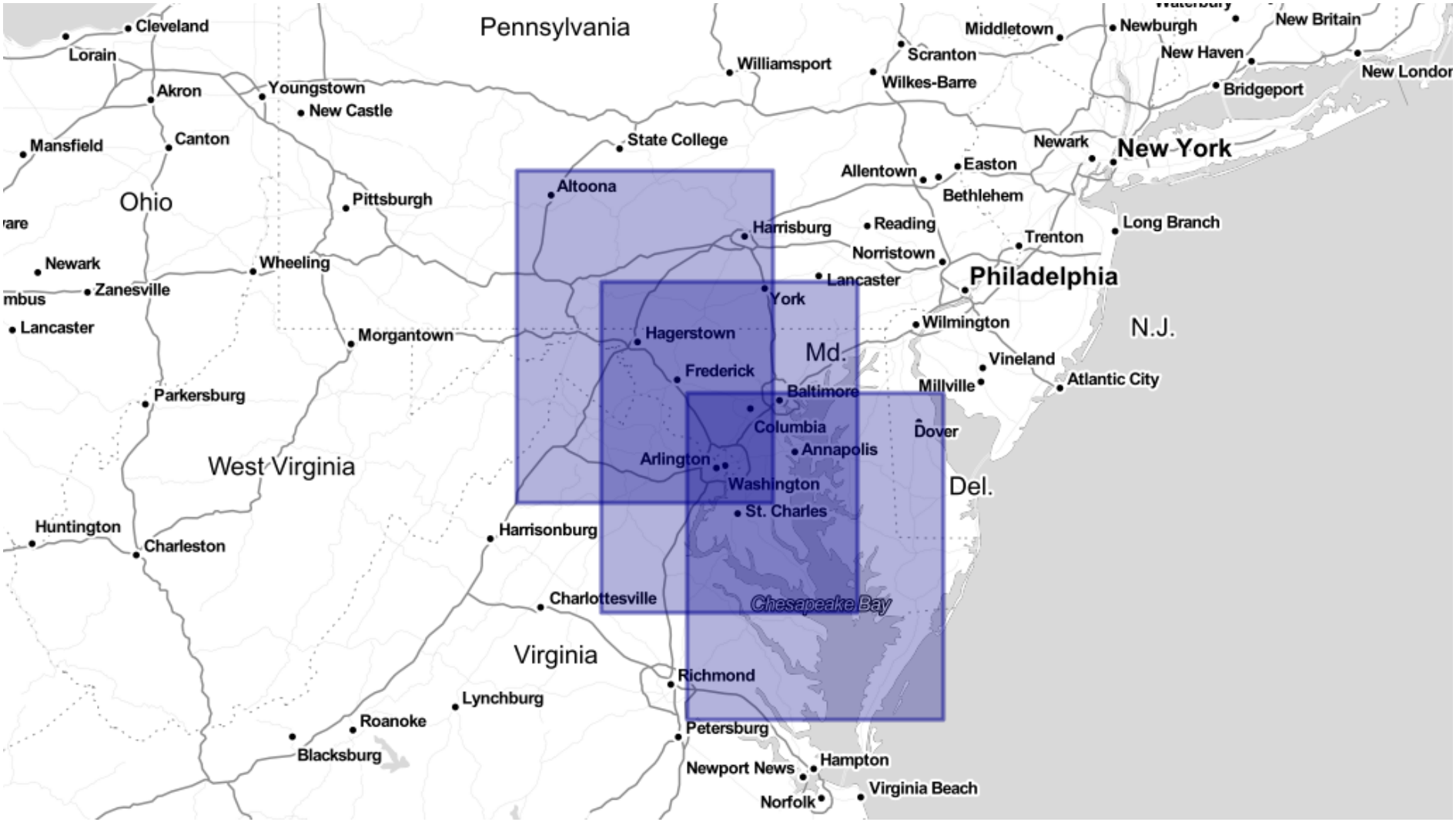
We can treat these as if these were regular words.

Then we can detect the pair: (‘‘!geo!Bayern’’, ‘‘Bundesliga’’)

Additional challenge: do this at Twitter speed.
And even faster, for archived data.
(Source code: https://github.com/kno10/reversegeocode)

Return

Grid-based Symbolic Representation



Use more than two overlapping grids for worst-case guarantees [Cha98].

Return

References I

[Alv+12] F. Alvanaki, S. Michel, K. Ramamritham, and G. Weikum. "See what's enBlogue: real-time emergent topic identification in social media". In: Proceedings of the 15th International Conference on Extending Database Technology (EDBT), Berlin, Germany. 2012, pp. 336–347.

[ASG13] H. Abdelhaq, C. Sengstock, and M. Gertz. "EvenTweet: Online localized event detection from Twitter". In: Proceedings of the VLDB Endowment 6.12 (2013), pp. 1326–1329.

[Blo70] B. H. Bloom. "Space/time trade-offs in hash coding with allowable errors". In: Communications of the ACM 13.7 (1970), pp. 422–426.

[Bud+13] C. Budak, T. Georgiou, D. Agrawal, and A. El Abbadi. "GeoScope: Online detection of geo-correlated information trends in social networks". In: Proceedings of the VLDB Endowment 7.4 (2013), pp. 229–240.

[Cha98] T. M. Chan. "Approximate Nearest Neighbor Queries Revisited". In: Discrete & Computational Geometry 20.3 (1998), pp. 359–373.

[CM05] G. Cormode and S. Muthukrishnan. "An improved data stream summary: the count-min sketch and its applications". In: J. Algorithms 55.1 (2005), pp. 58–75.

[Fin09] T. Finch. Incremental calculation of weighted mean and variance. Tech. rep. University of Cambridge, 2009.

References II

[MK10] M. Mathioudakis and N. Koudas. "Twittermonitor: trend detection over the Twitter stream". In: Proceedings of the ACM International Conference on Management of Data (SIGMOD), Indianapolis, IN. 2010, pp. 1155–1158.

[SOM10] T. Sakaki, M. Okazaki, and Y. Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors". In: Proceedings of the 19th International Conference on World Wide Web (WWW), Raleigh, NC. 2010, pp. 851–860.

[SWK14] E. Schubert, M. Weiler, and H.-P. Kriegel. "SigniTrend: Scalable Detection of Emerging Topics in Textual Streams by Hashed Significance Thresholds". In: Proceedings of the 20th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), New York, NY. 2014, pp. 871–880.

[SWK16] E. Schubert, M. Weiler, and H.-P. Kriegel. "SPOTHOT: Scalable Detection of Geo-spatial Events in Large Textual Streams". In: Proceedings of the 28th International Conference on Scientific and Statistical Database Management (SSDBM), Budapest, Hungary. 2016, 8:1–8:12.

[TA14] G. B. Tran and M. Alrifai. "Indexing and analyzing Wikipedia's current events portal, the daily news summaries by the crowd". In: Proceedings of the 23rd International Conference on World Wide Web (WWW), Seoul, Korea. 2014, pp. 511–516.