Dialect variation in social media

Jacob Eisenstein
@jacobeisenstein

Georgia Institute of Technology

August 15, 2014

What would you do with a billion words?

Social media offers exciting new opportunities for dialectology...

- 10^9 words, 10^6 authors

What would you do with a ~~billion~~ trillion words?

Social media offers exciting new opportunities for dialectology...

- 10^9 words, 10^6 authors

What would you do with a ~~billion~~ trillion words?

Social media offers exciting new opportunities for dialectology...

- 10^9 words, 10^6 authors
- Metadata: geolocation, timestamps, and more.

What would you do with a ~~billion~~ trillion words?

Social media offers exciting new opportunities for dialectology...

- 10^9 words, 10^6 authors
- Metadata: geolocation, timestamps, and more.
- Natural conversations, no experimenter intervention

... but also new challenges.

What does social media mean for dialect?

What does social media mean for dialect?

Then

Now

- To an unprecedented degree, informal communication is now written.
- What does this mean for dialect? For writing?

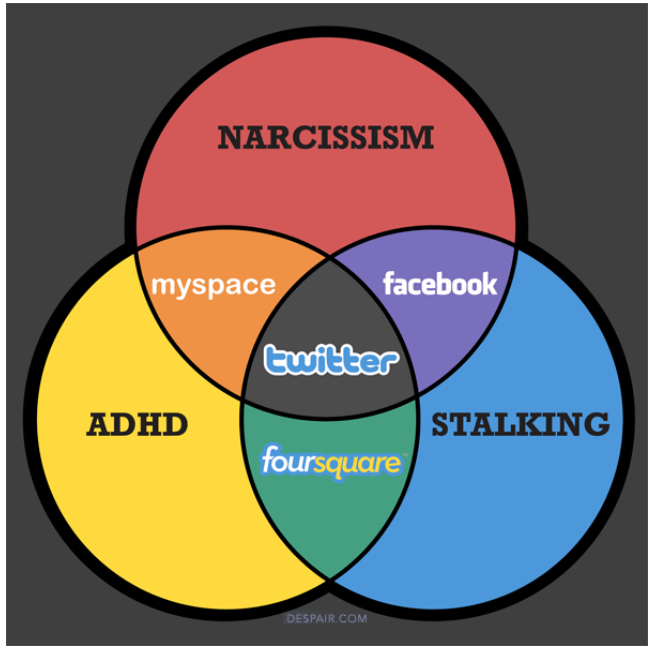
Two questions

1. What would you do with a billion words?
2. What does ubiquitous informal written communication mean for dialect?

Defining digital communication

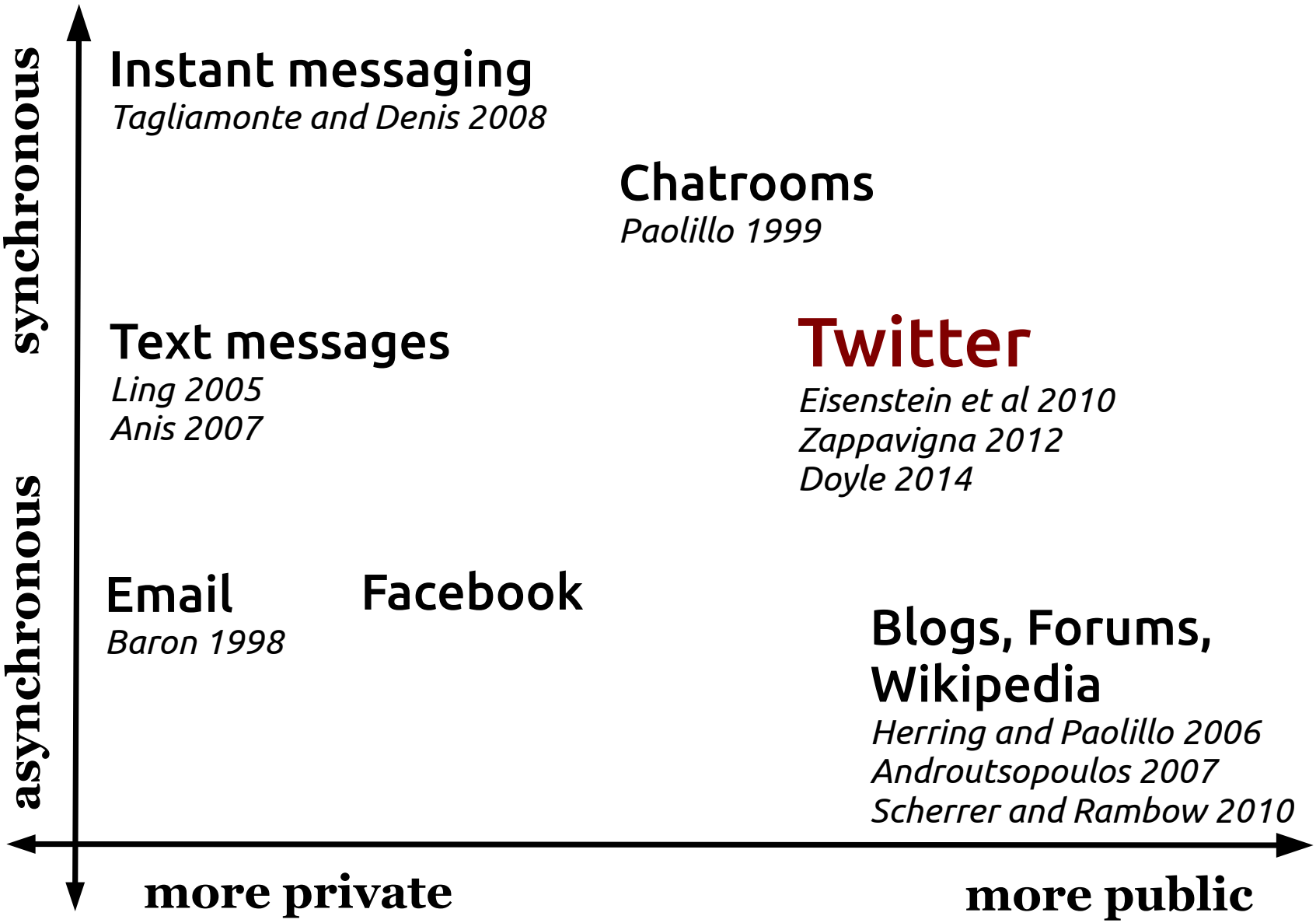
Androutsopoulus (2011):

Networked writing... carried out on digital technologies that enable private or public, asynchronous or near-synchronous exchange among individuals and groups...



In Coupland & Kristiansen (Eds.), Language Standardisation in Europe.

A landscape of digital communication

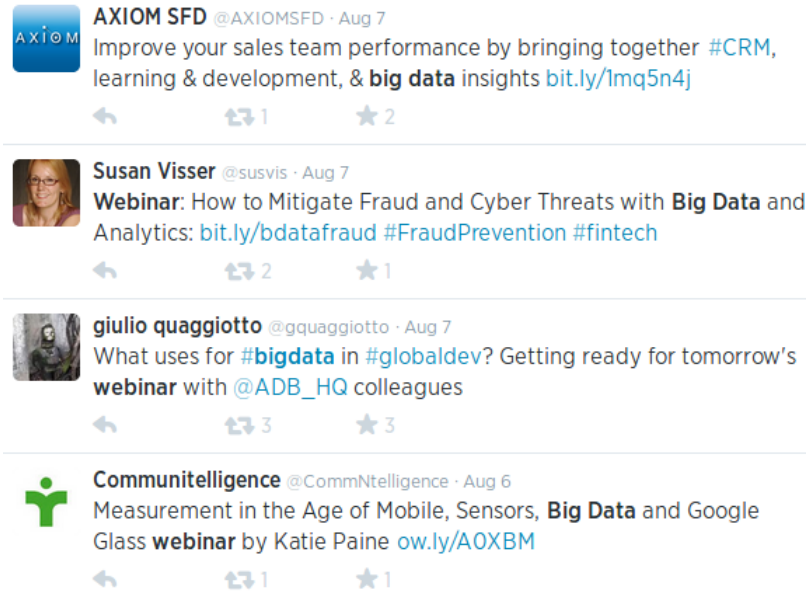


What is Twitter?

- 140-character messages
- Each user has a custom *timeline* of people they've chosen to *follow*.

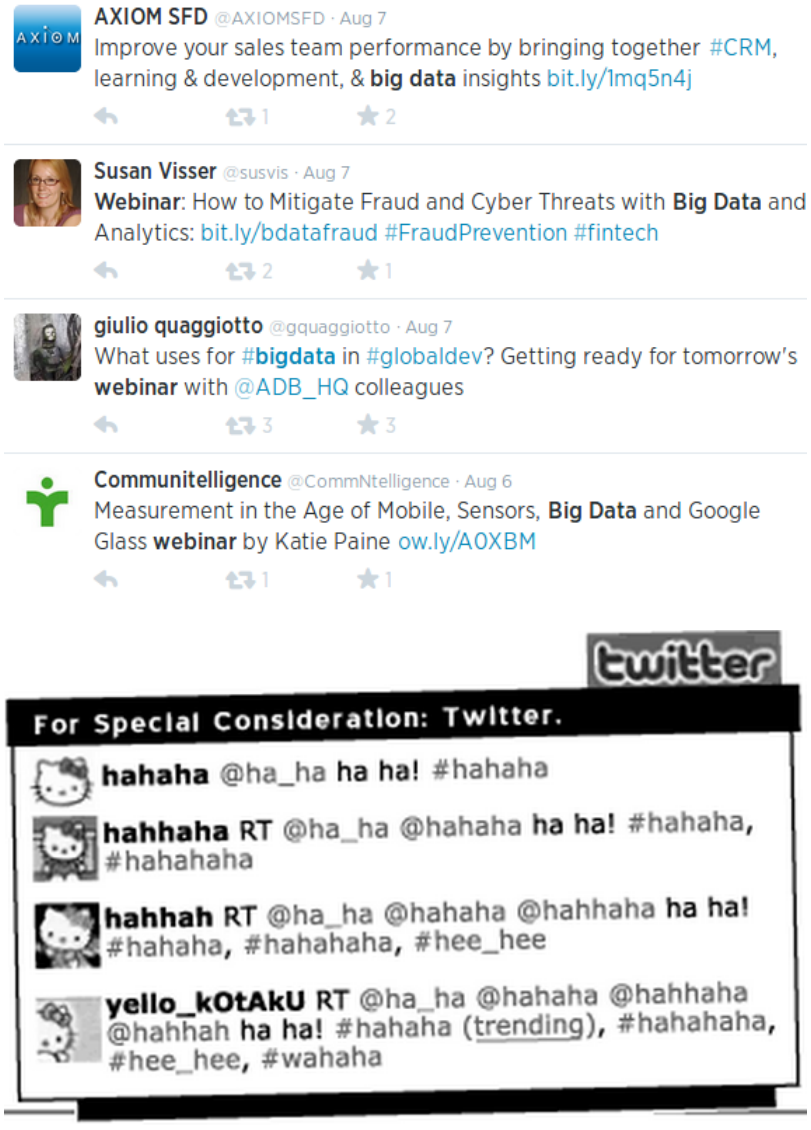
What is Twitter?

- 140-character messages
- Each user has a custom *timeline* of people they've chosen to *follow*.



What is Twitter?

- 140-character messages
- Each user has a custom *timeline* of people they've chosen to *follow*.



Who are these people?

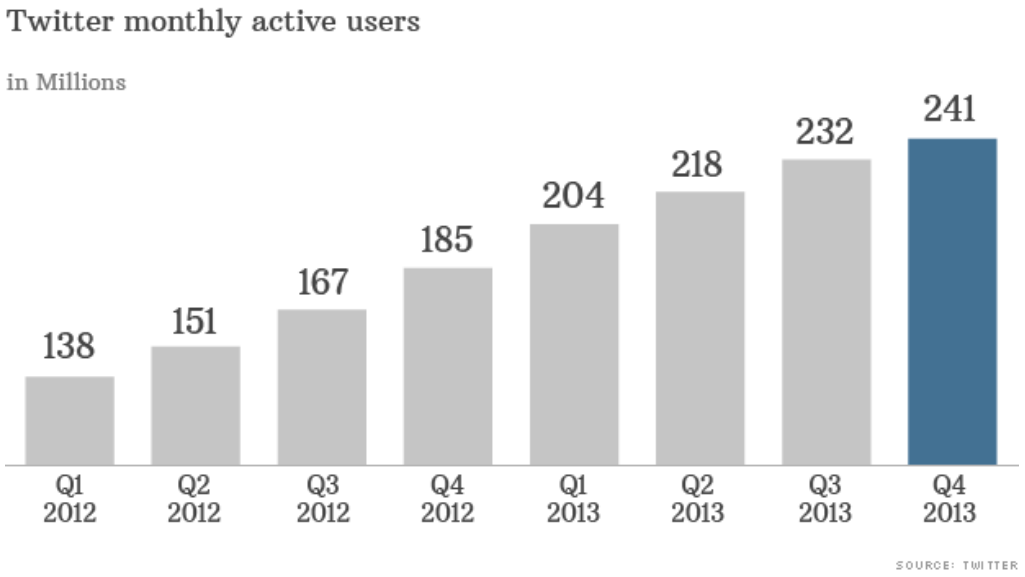
Twitter users

Among online adults, the % who use Twitter

		Use Twitter
	All internet users (n= 1,445)	*18%*
a	Men (n= 734)	17
b	Women (n= 711)	18
a	White, Non-Hispanic (n= 1,025)	16
b	Black, Non-Hispanic (n= 138)	29[ac]
c	Hispanic (n= 169)	16
a	18-29 (n= 267)	31[bcd]
b	30-49 (n= 473)	19[cd]
c	50-64 (n= 401)	9
d	65+ (n= 278)	5
a	High school grad or less (n= 385)	17
b	Some college (n= 433)	18
c	College+ (n= 619)	18
a	Less than $30,000/yr (n= 328)	17
b	$30,000-$49,999 (n= 259)	18
c	$50,000-$74,999 (n= 187)	15
d	$75,000+ (n= 486)	19
a	Urban (n= 479)	18[c]
b	Suburban (n= 700)	19[c]
c	Rural (n= 266)	11

- Per-message statistics will differ.
- Representativeness concerns are real...
- ... but there are potential solutions.
- Social media has important representativeness advantages too.

Getting Twitter data



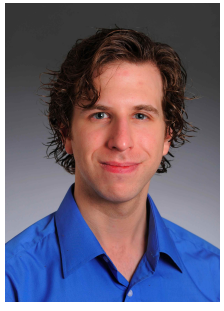
- Stream 1% of public tweets for free. (or...)
- Currently, it's 2GB per day, compressed.
- Roughly 1-2% of messages have GPS coordinates.
- *Data cannot be redistributed.*

The CMU Twitter Corpus

- August 2009 to September 2012
- 2.77 million user accounts
- 114 million messages, all geolocated to USA
- Filters:
  - No retweets, no links
  - No "celebrities"

Brendan O'Connor

Noah Smith

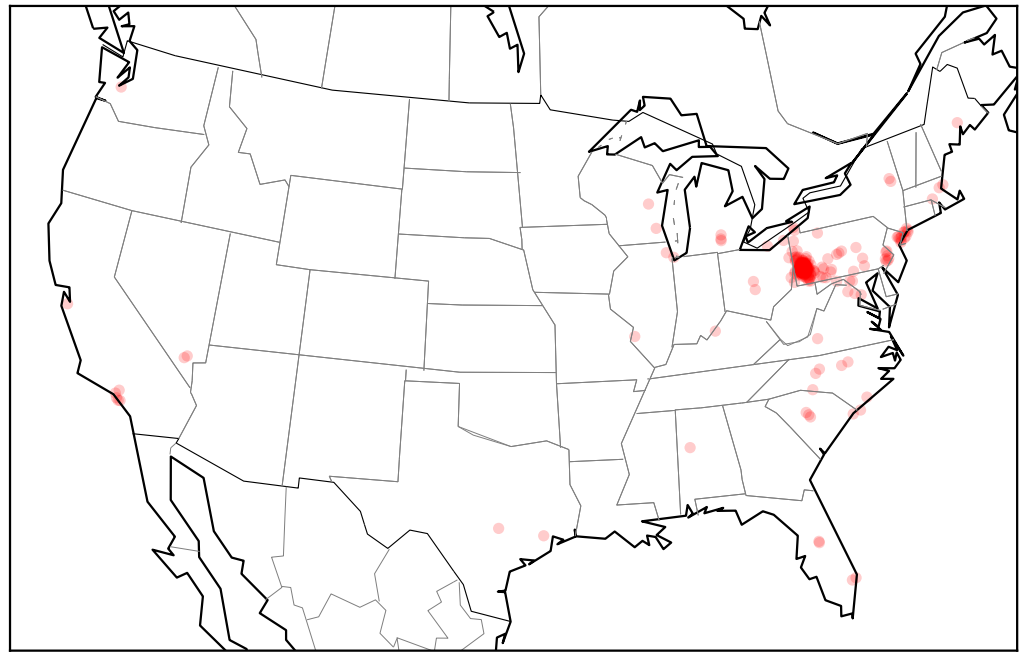


What should we do with a billion words?

1. Map known dialect terms.

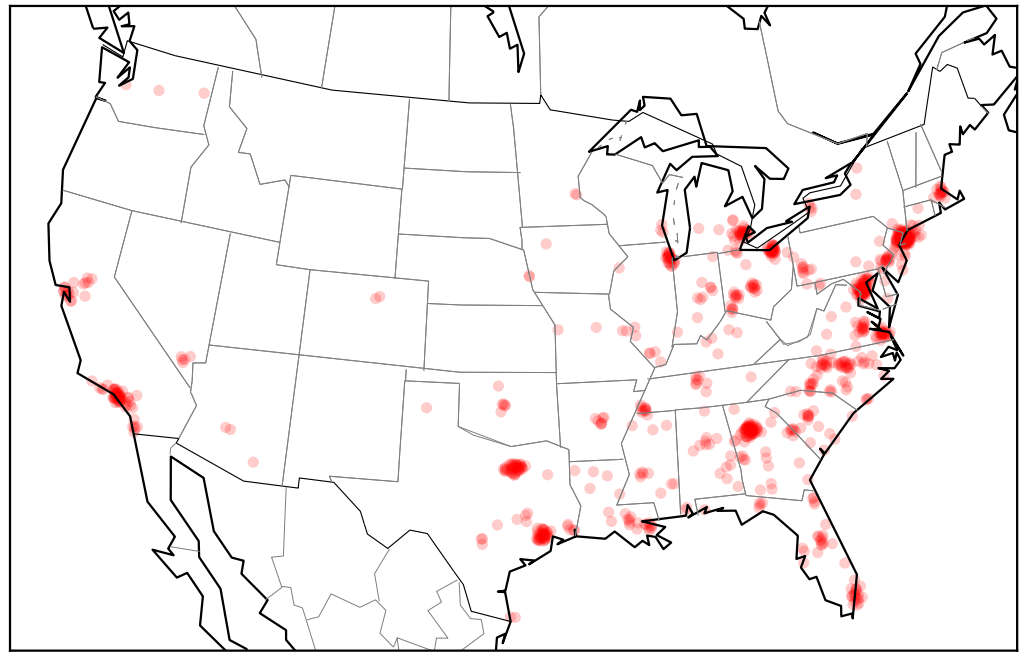
Yinz

- 2nd-person pronoun
- Western Pennsylvania
- Very rare: appears in 535 of 10^8 messages



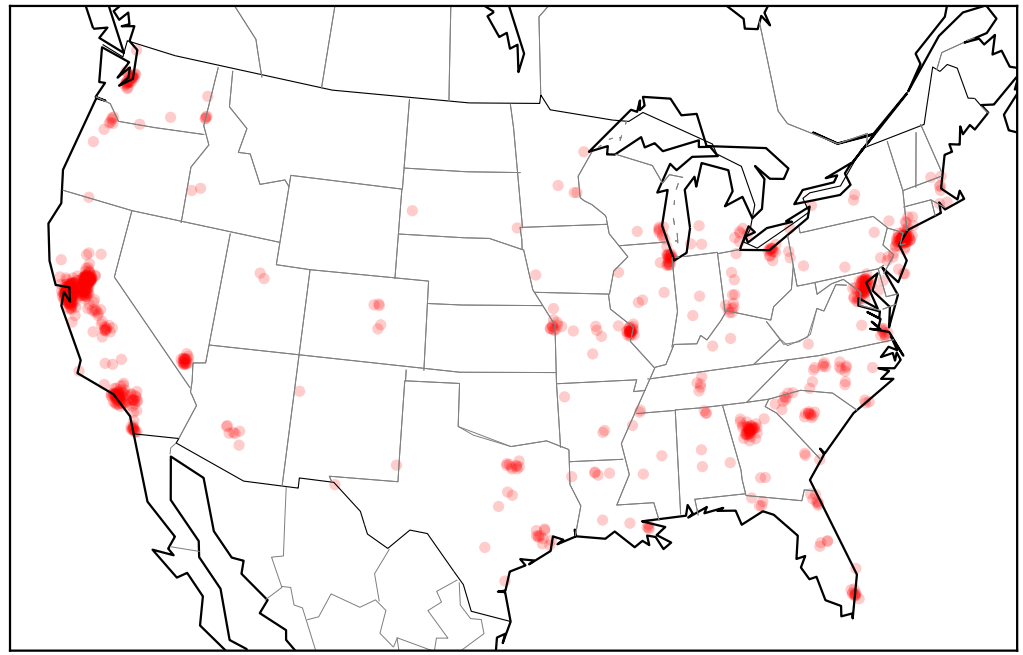
Yall

- 2nd-person pronoun
- Southeast, African-American English
- Once per 250 messages



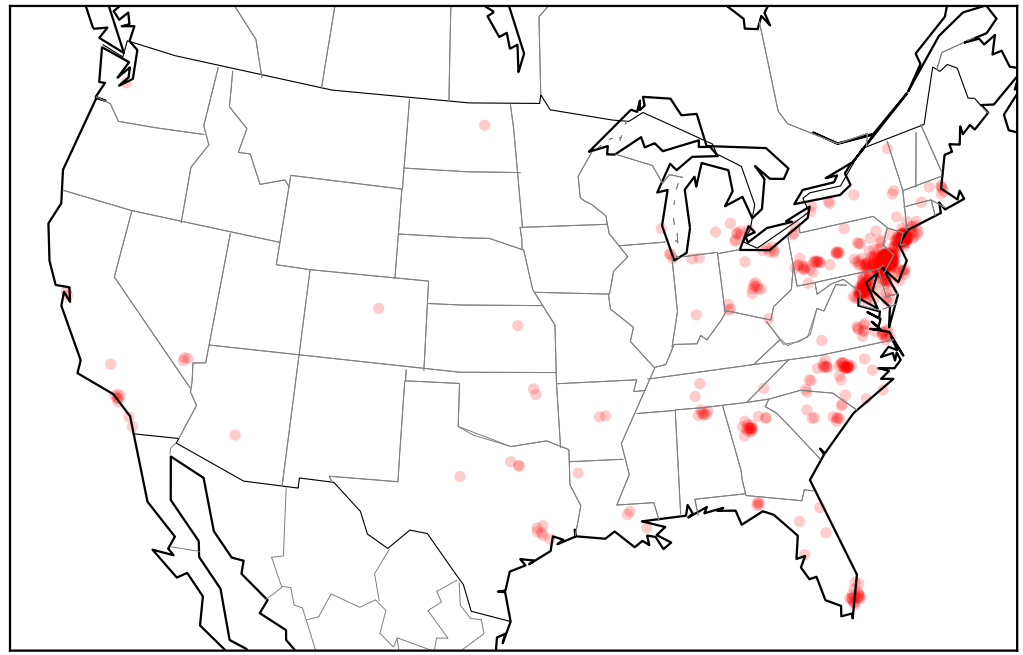
Hella

- Intensifier, e.g.
  - i got hella nervous
- Northern California (Bucholtz 2007)
- Once per 1000 messages



Jawn

- Noun, diffuse semantics
- Philadelphia, hiphop (Alim 2009)
- Once per 1000 messages



- @user ok u have heard this jawn right
- i did wear that jawn but it was kinda warm this week

Summary of spoken dialect terms

	rate	*region*
yinz	200,000	mainly used in Western PA
yall	250	ubiquitous
hella	1000	ubiquitous, but more frequent in Northern California
jawn	1000	mainly used in Philadelphia

- Overall: mixed evidence for spoken language dialect variation in Twitter.
- But are these the right words?

What should we do with a billion words?

1. Map known dialect terms.

What should we do with a billion words?

1. Map known dialect terms.
2. Search for unknown dialect terms.

Measuring regional specificity

Per region r,

- *Difference* in frequencies, $f_{i,r} - f_i$

over-emphasizes frequent words

Measuring regional specificity

Per region r,

- *Difference* in frequencies, $f_{i,r} - f_i$

  over-emphasizes frequent words

- *Log-ratio* in frequencies, $\log f_{i,r} - \log f_i$

  over-emphasizes rare words

Measuring regional specificity

Per region r,

- *Difference* in frequencies, $f_{i,r} - f_i$

  over-emphasizes frequent words

- *Log-ratio* in frequencies, $\log f_{i,r} - \log f_i$

  over-emphasizes rare words

- *Regularized* log-frequency ratio,
  $\eta_{i,r} \approx \log f_{i,r} - \log f_i$, where $|\eta_{i,r}|$ is penalized.

$$\hat{\eta}_r = \arg\max_{\eta} \quad \log P(w|\eta; f) - \lambda|\eta|$$

λ controls the tradeoff between rare and frequent words

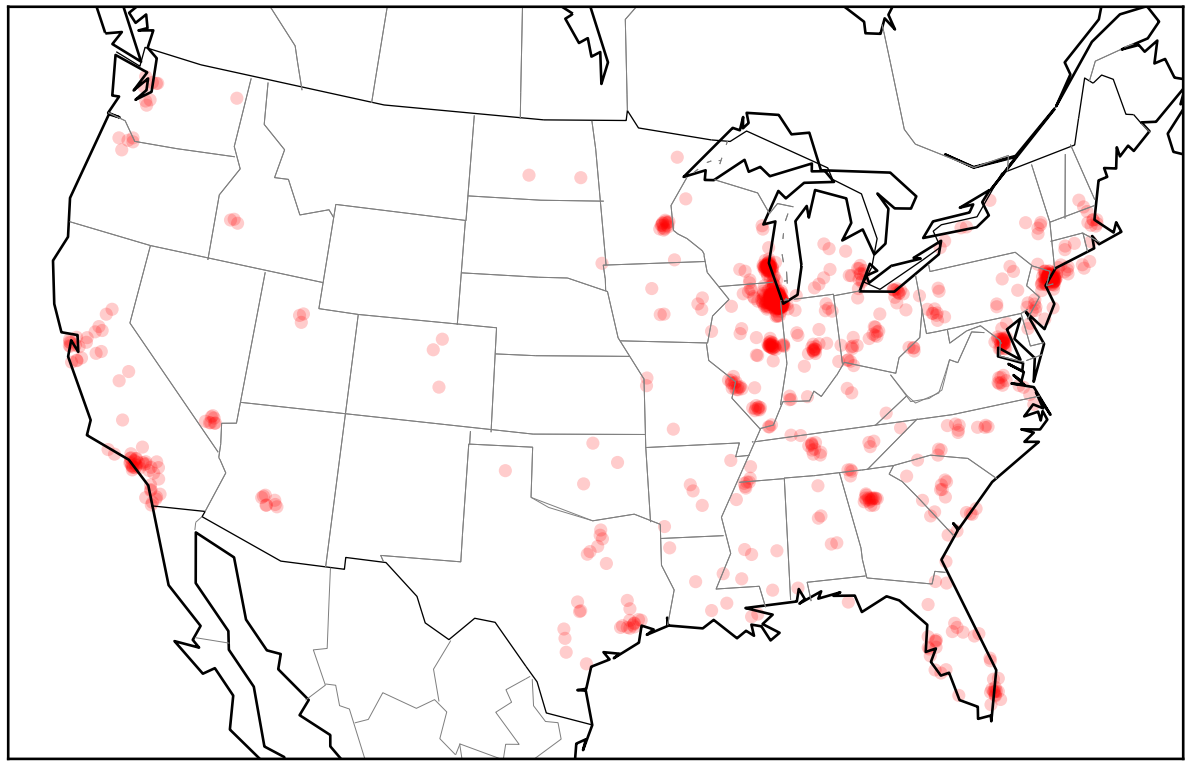
Discovered words

- **New York**: flatbush, baii, brib, bx, staten, *mta*, odee, soho, deadass, werd
- **Los Angeles**: pasadena, venice, anaheim, *dodger*, disneyland, angeles, compton, *ucla*, *dodgers*, melrose
- **Chicago**: #chicago, lbvs, chicago, *blackhawks*, *#bears*, *#bulls*, mfs, *cubs*, burbs, bogus
- **Philadelphia**: jawn, ard, *#phillies*, *sixers*, *phils*, *wawa*, *philadelphia*, delaware, *philly*, *phillies*

place names *entities* words

bogus

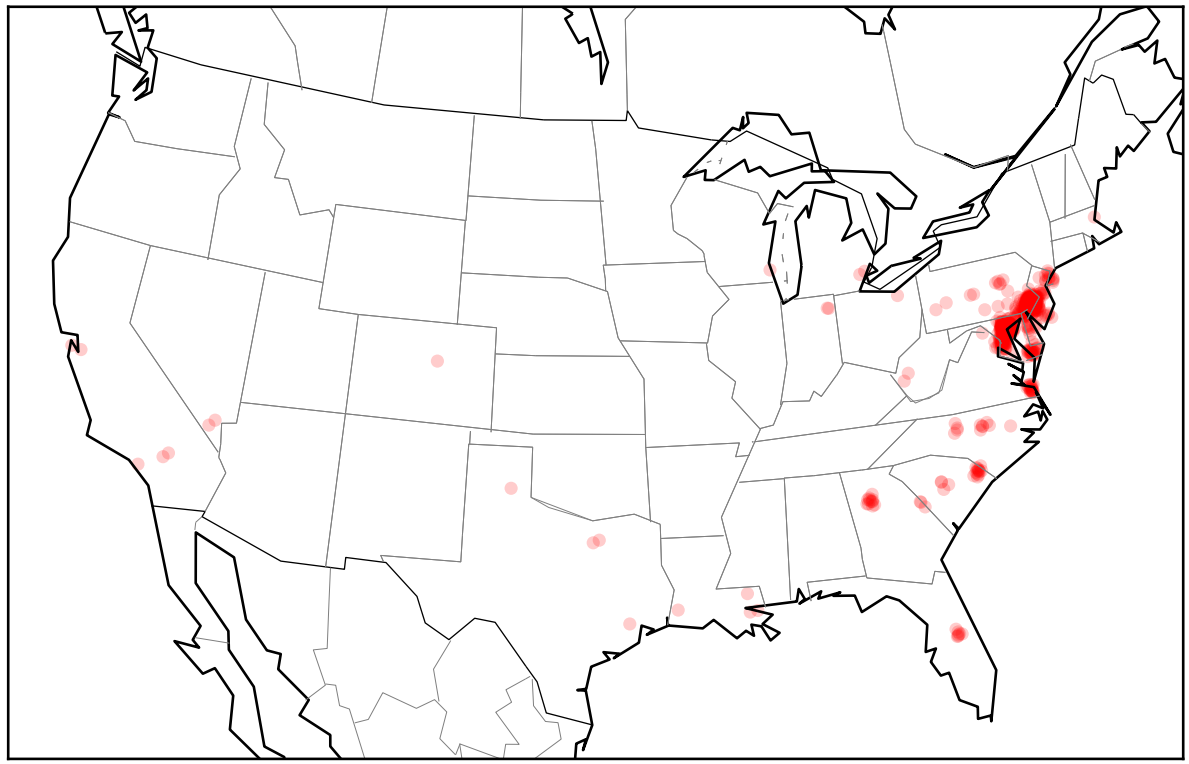
adjective, meaning fake



- That muffin was bogus

ard

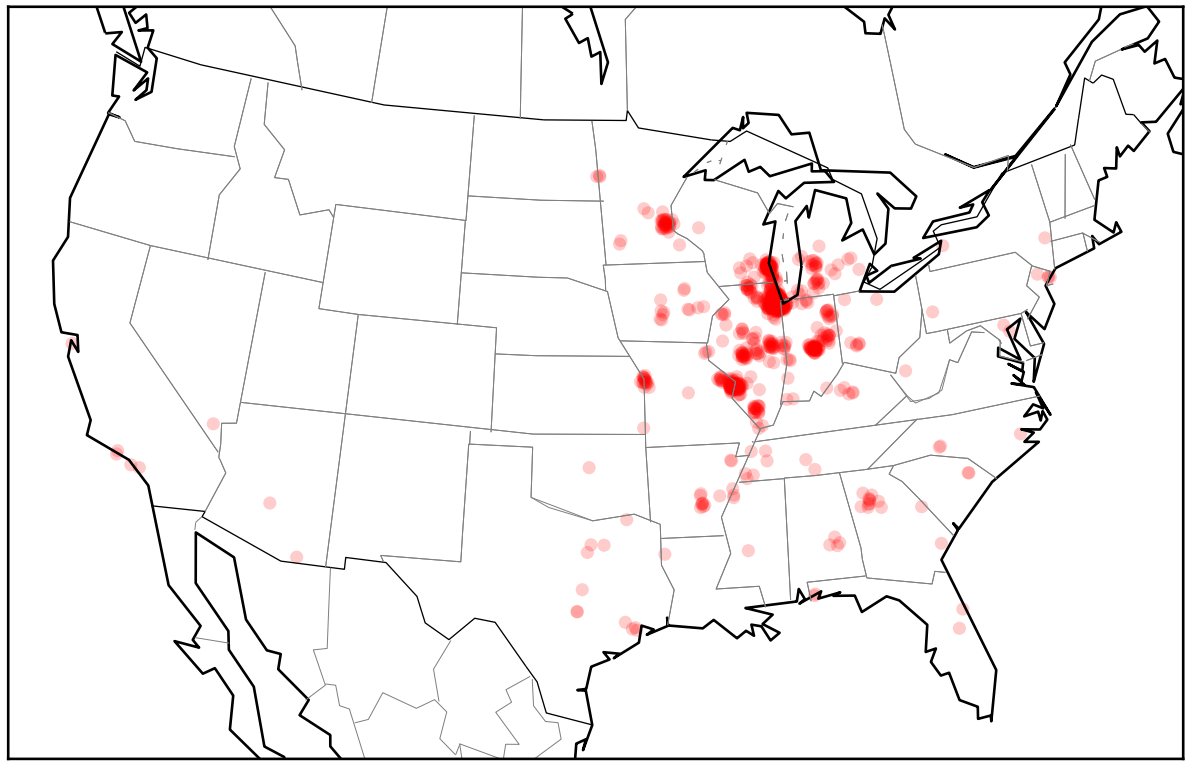
alternative spelling for alright



- @name ard let me kno
- lol u'll be ard

lbvs

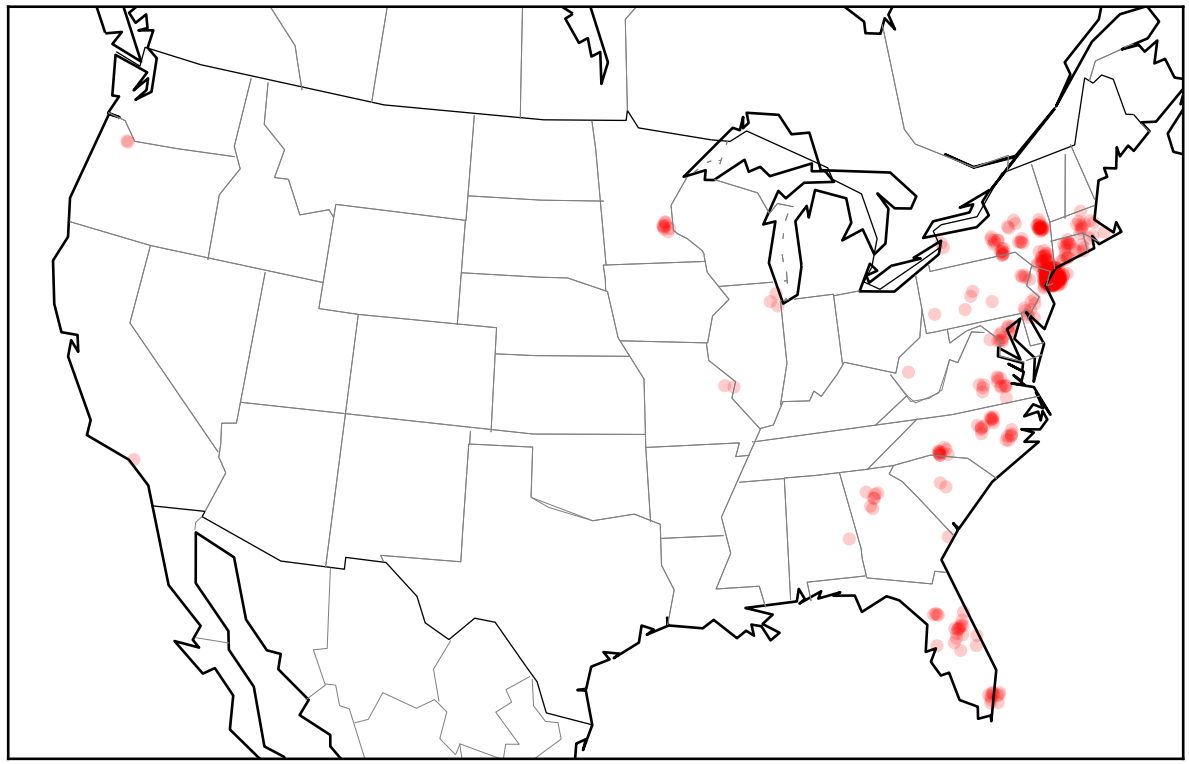
laughing but very serious



- i wanna rent a hotel room just to swim lbvs
- tell ur momma 2 buy me a car lbvs

odee

intensifier, related to overdose or overdone



- i'm odee sleepy
- she said she odee miss me
- its rainin odee :(

Summary of discovered dialect terms

1. Words from speech: bogus, jawn
2. Abbreviations: lbvs
3. Phonetic spellings: ard
4. Combinations of 2 and 3: odee
5. Emoticons: -_-

Summary of discovered dialect terms

1. Words from speech: bogus, jawn
2. Abbreviations: lbvs
3. Phonetic spellings: ard
4. Combinations of 2 and 3: odee
5. Emoticons: -_-

- "Netspeak" phenomena can also be regional.
- Are these regional differences ephemeral?
- Could they be as persistent as spoken language dialect differences?

What should we do with a billion words?

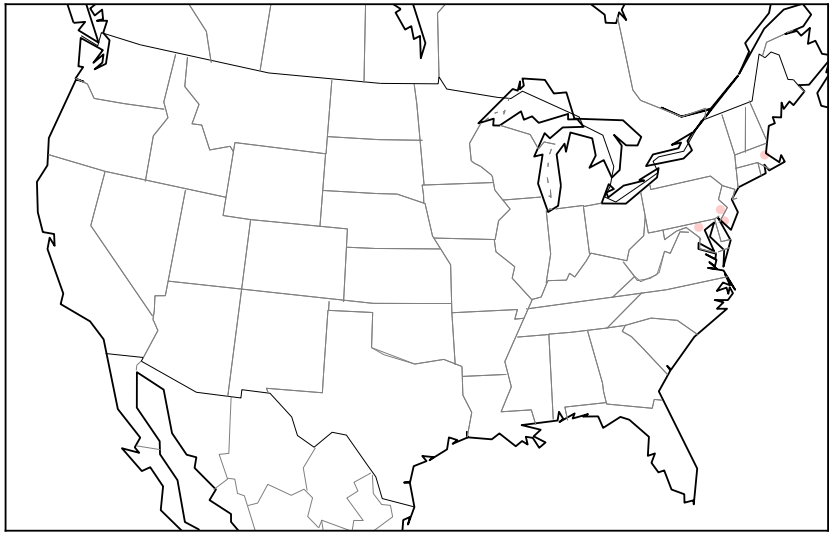
1. Map known dialect terms.
2. Search for unknown dialect terms.

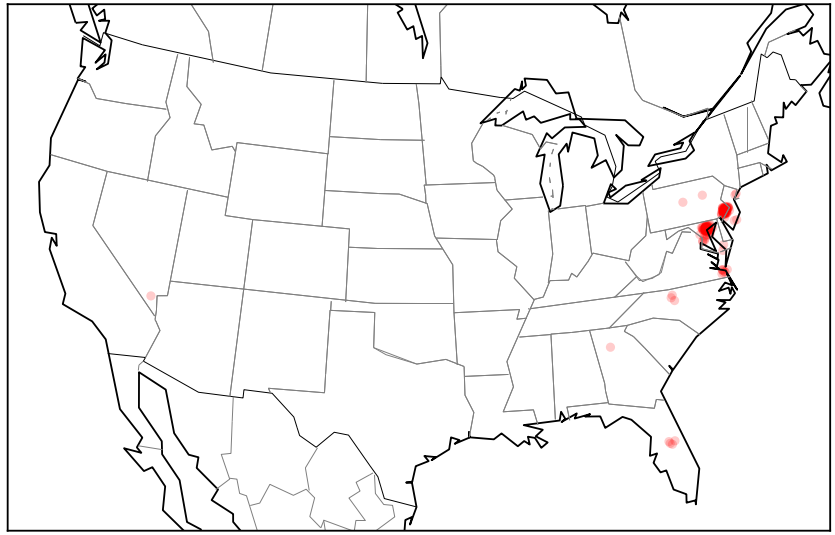
What should we do with a billion words?

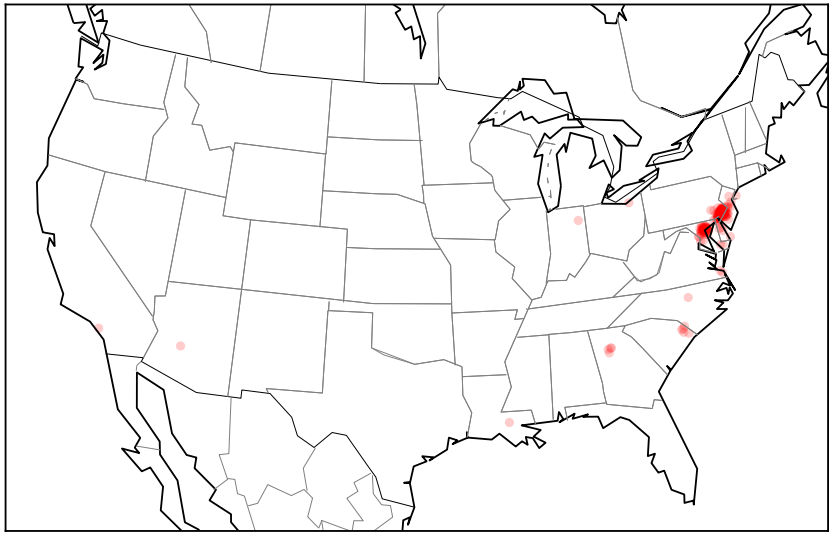
1. Map known dialect terms.
2. Search for unknown dialect terms.
3. Model change over time.

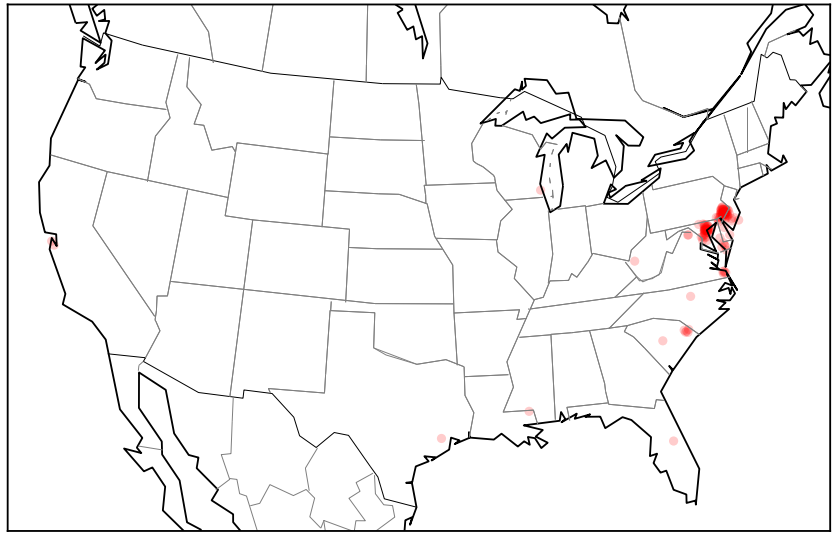
Change from 2009-2012: ard

lol u'll be ard



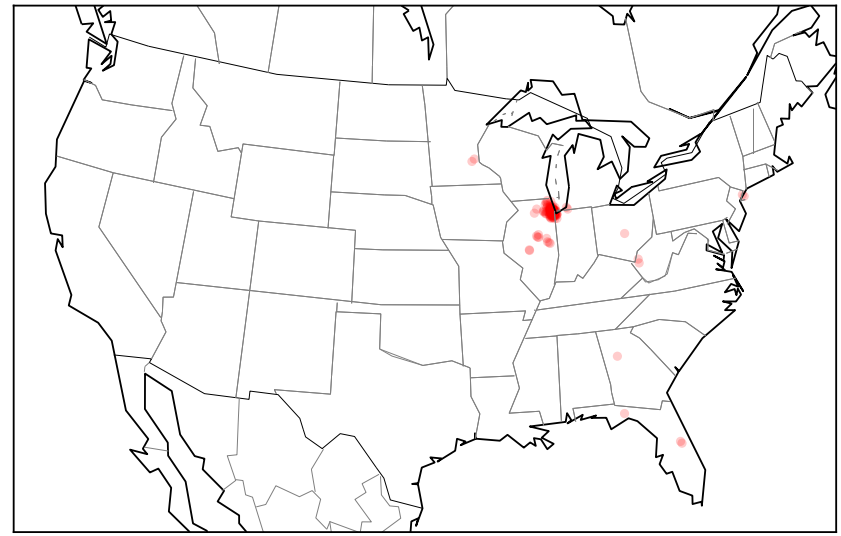


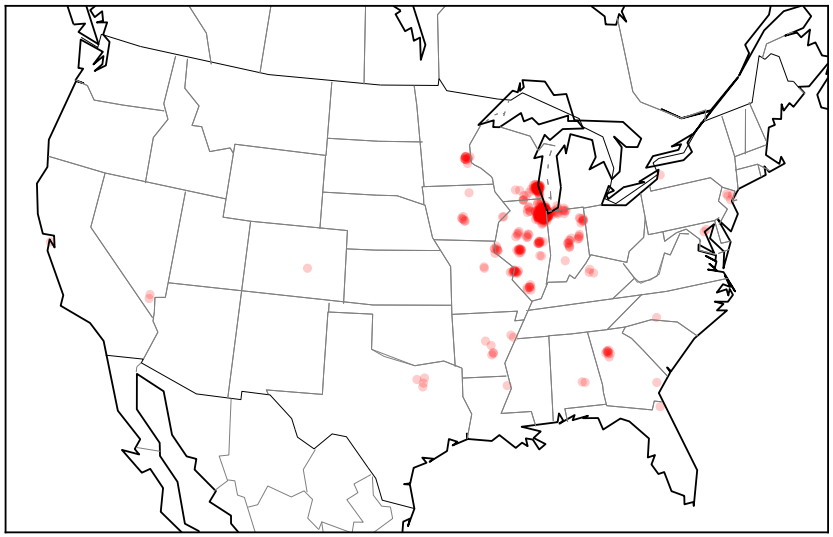


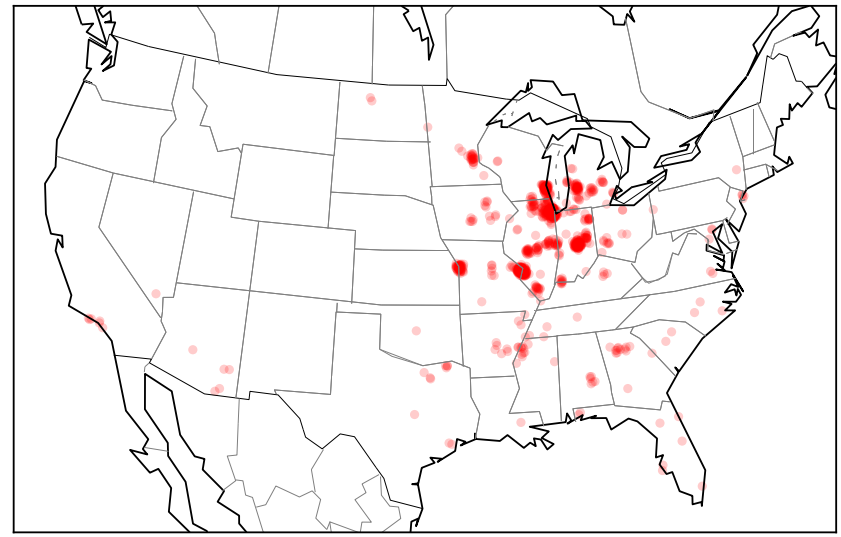


Change from 2009-2012: lbvs

tell ur momma 2 buy me a car lbvs

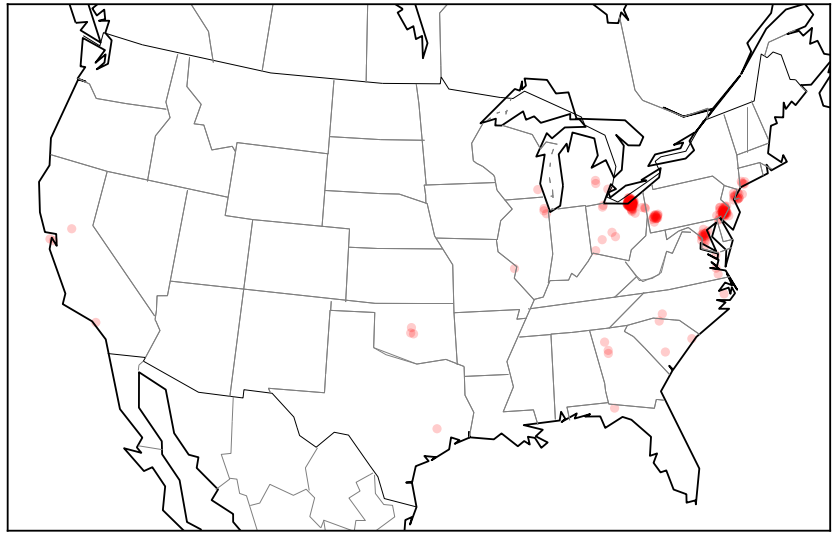


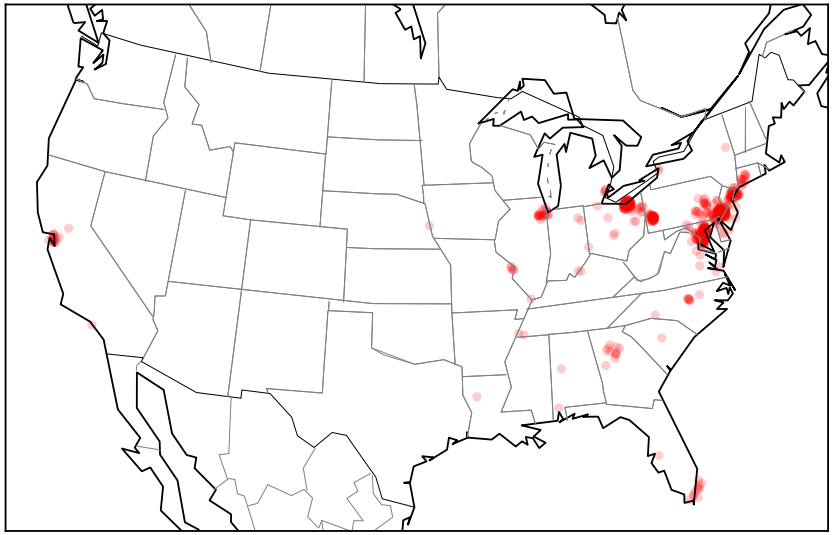


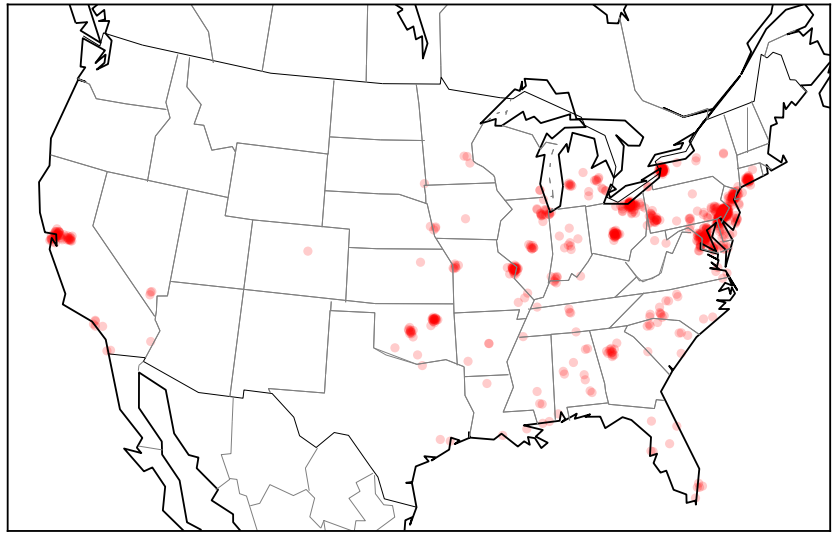


Change from 2009-2012: ctfu

@name lmao! haahhaa ctfu!







The voyage of ctfu

2009	Cleveland
2010	Pittsburgh, Philadelphia
2011	Washington DC, Chicago, NY
2012	San Francisco, Columbus

- Hard to explain with wave, gravity models.
- There are thousands of other words.
  How to integrate them?

Aggregating across words

- Let a represent region-to-region linguistic "influence"

  $a_{i,j}$ is the influence between region i and region j.

- Let w represent the word counts per region

  $w_{n,r,t}$ is the count of word n in region r at time t.

- We want the *maximum likelihood estimate*,

$$\hat{a} = \arg\max_{a} P(w; a)$$

But first we have to define this probability.

A probabilistic model of regional change

$$P(\text{words}; \text{influence}) \triangleq P(w; a)$$

$$= \sum_z P(w, z; a) = \sum_z \overbrace{P(w|z)}^{\text{emission}} \overbrace{P(z; a)}^{\text{transition}}$$

(z represents "activation")

A probabilistic model of regional change

$$P(\text{words}; \text{influence}) \triangleq P(w; a)$$

$$= \sum_z P(w, z; a) = \sum_z \overbrace{P(w|z)}^{\text{emission}} \overbrace{P(z; a)}^{\text{transition}}$$

(z represents "activation")

$$= \int P(w|z)P(z; a)dz \qquad \text{(uh oh...)}$$

A probabilistic model of regional change

$$P(\text{words}; \text{influence}) \triangleq P(w; a)$$

$$= \sum_z P(w, z; a) = \sum_z \overbrace{P(w|z)}^{\text{emission}} \overbrace{P(z; a)}^{\text{transition}}$$

(z represents "activation")

$$= \int P(w|z)P(z; a)dz \qquad \text{(uh oh...)}$$

$$\rightarrow z^{(k)}, k \in \{1, 2, \ldots, K\}$$

$$\approx \sum_k P(w|z^{(k)})P(z^{(k)}; a)$$

(Monte Carlo approximation to the rescue!)

A probabilistic model of regional change

$$P(\text{words}; \text{influence}) \triangleq P(w; a)$$

$$= \sum_z P(w, z; a) = \sum_z \overbrace{P(w|z)}^{\text{emission}} \overbrace{P(z; a)}^{\text{transition}}$$

(z represents "activation")

$$= \int P(w|z)P(z; a)dz \qquad \text{(uh oh...)}$$

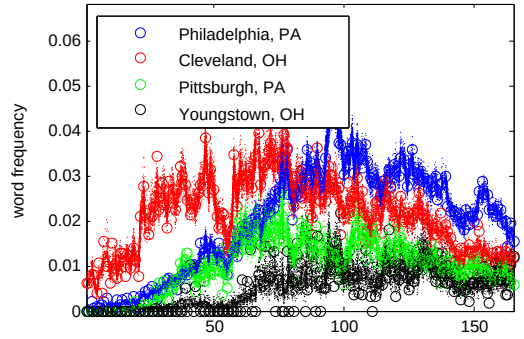
$$\rightarrow z^{(k)}, k \in \{1, 2, \ldots, K\}$$

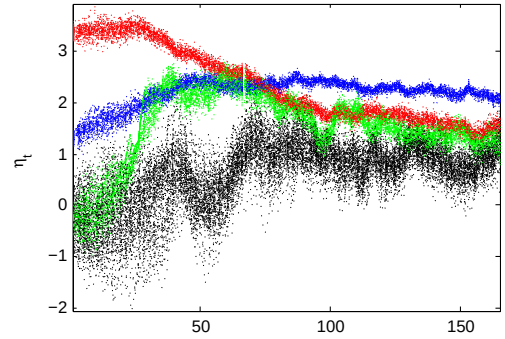
$$\approx \sum_k P(w|z^{(k)})P(z^{(k)}; a)$$

(Monte Carlo approximation to the rescue!)

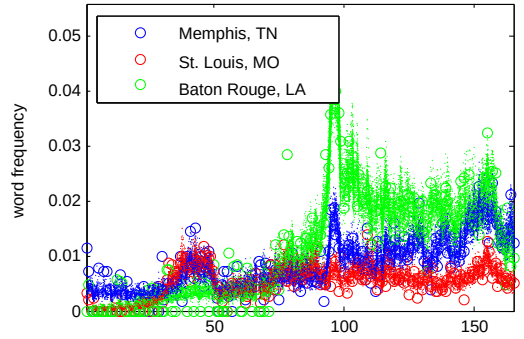
$$\hat{a} = \arg\max_a \sum_k P(w|z^{(k)})P(z^{(k)}; a)$$

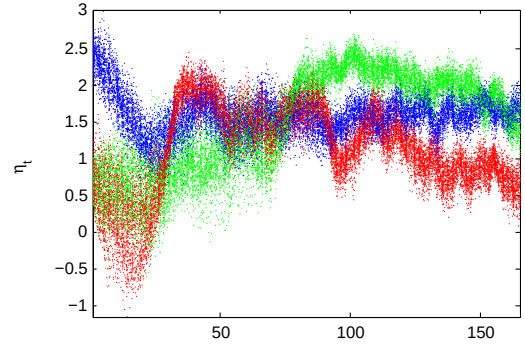
ctfu


Philadelphia, PA
Cleveland, OH
Pittsburgh, PA
Youngstown, OH
word frequency

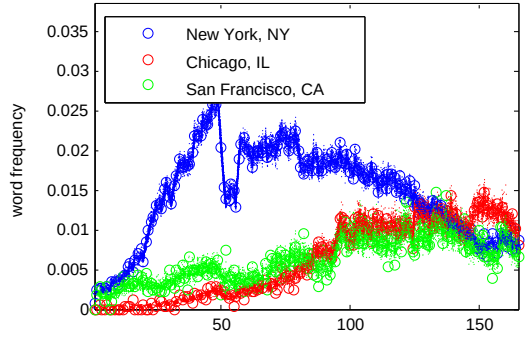

η_t

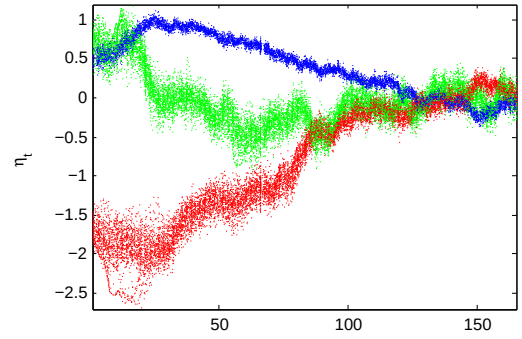
ion


Memphis, TN
St. Louis, MO
Baton Rouge, LA
word frequency

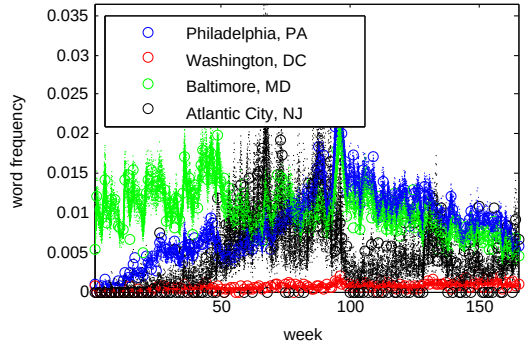

η_t

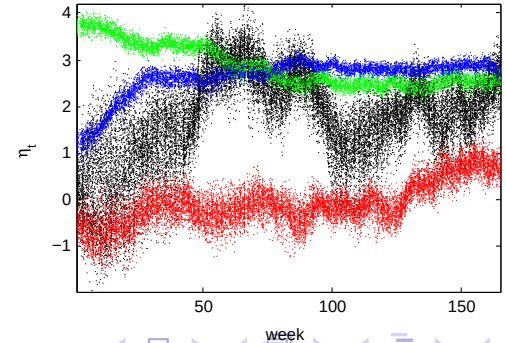
-__-


New York, NY
Chicago, IL
San Francisco, CA
word frequency


η_t

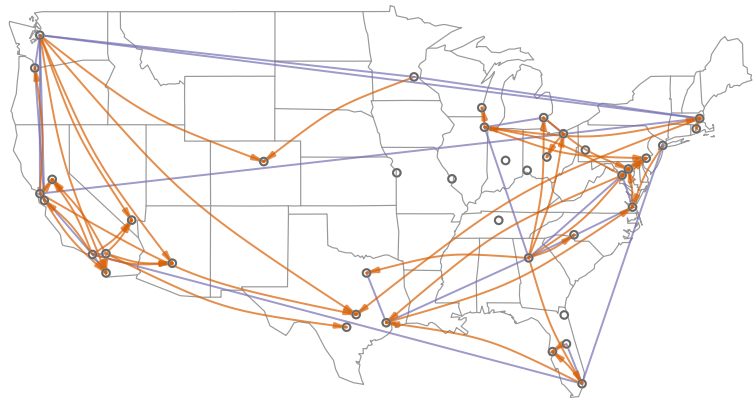
ard


Philadelphia, PA
Washington, DC
Baltimore, MD
Atlantic City, NJ
word frequency
week


η_t
week

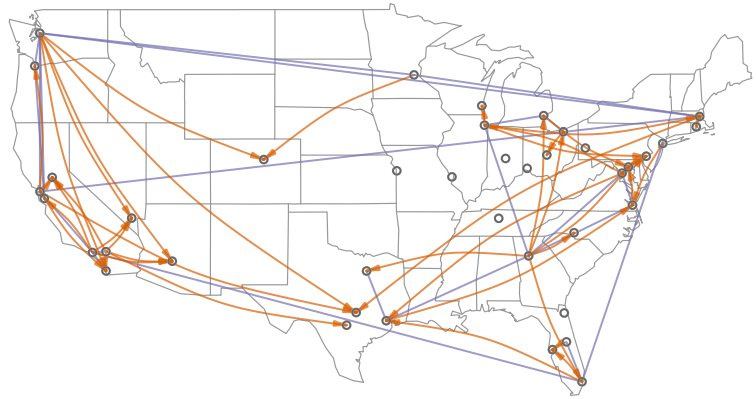
Aggregating region-to-region influence

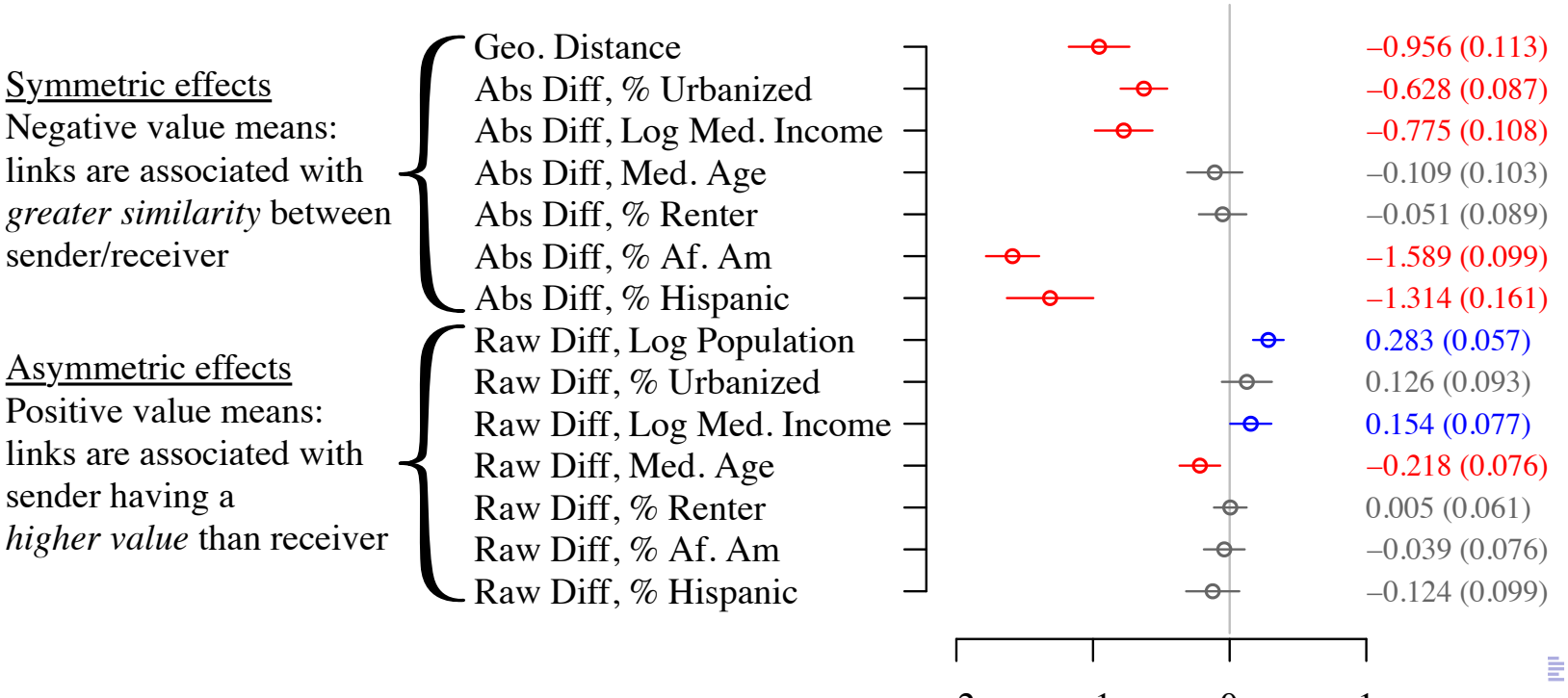
- Next, discretize a to form a network.
- (This is a subset of highly-confident edges)



Aggregating region-to-region influence

- Next, discretize *a* to form a network.
- (This is a subset of highly-confident edges)



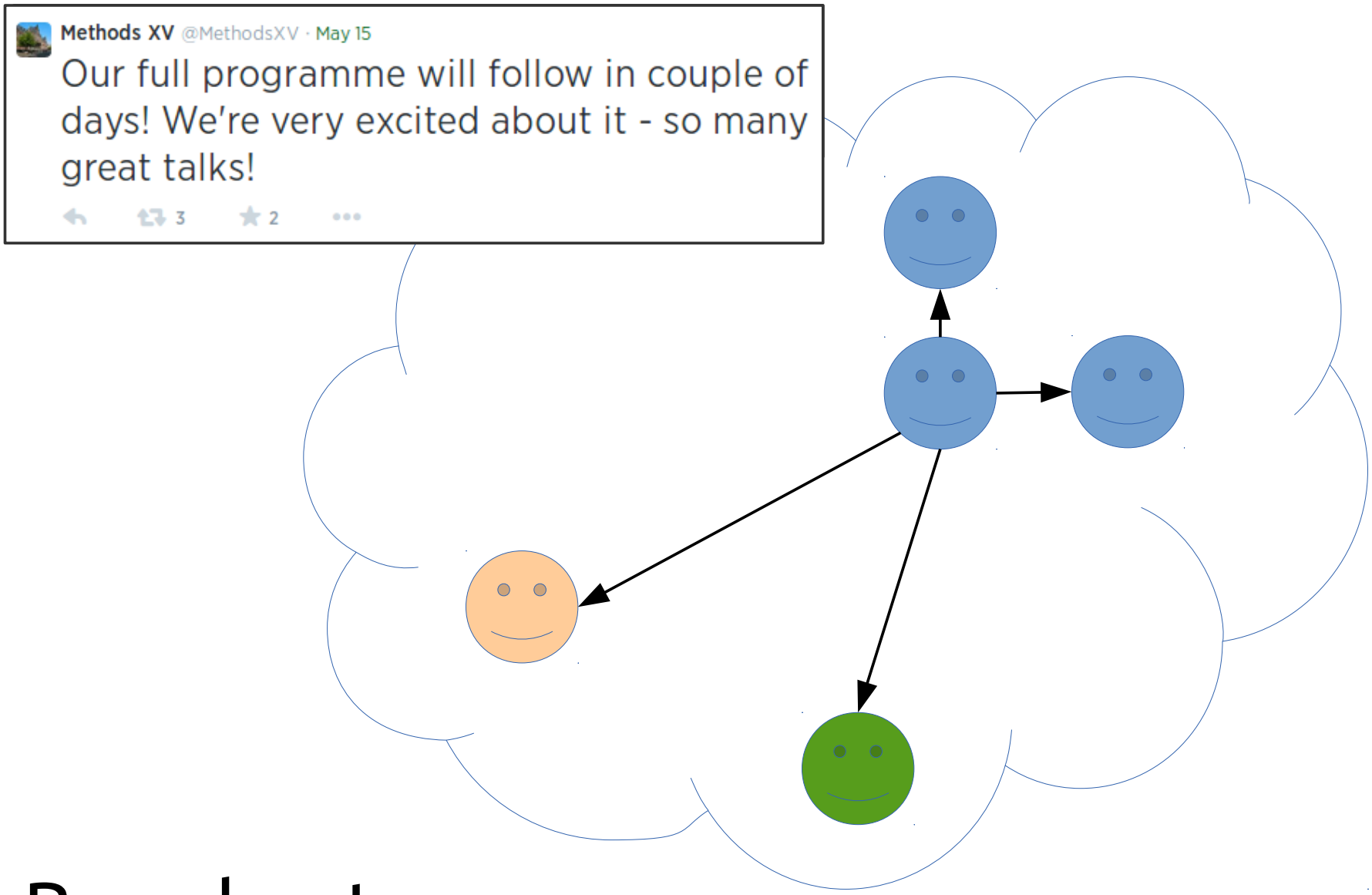


From macro to micro

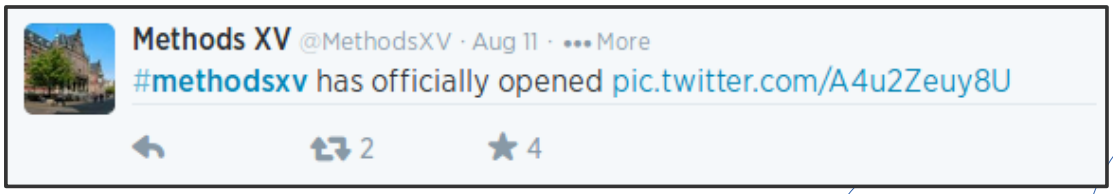
Macro-level variation and change must ground out in individual linguistic decisions.

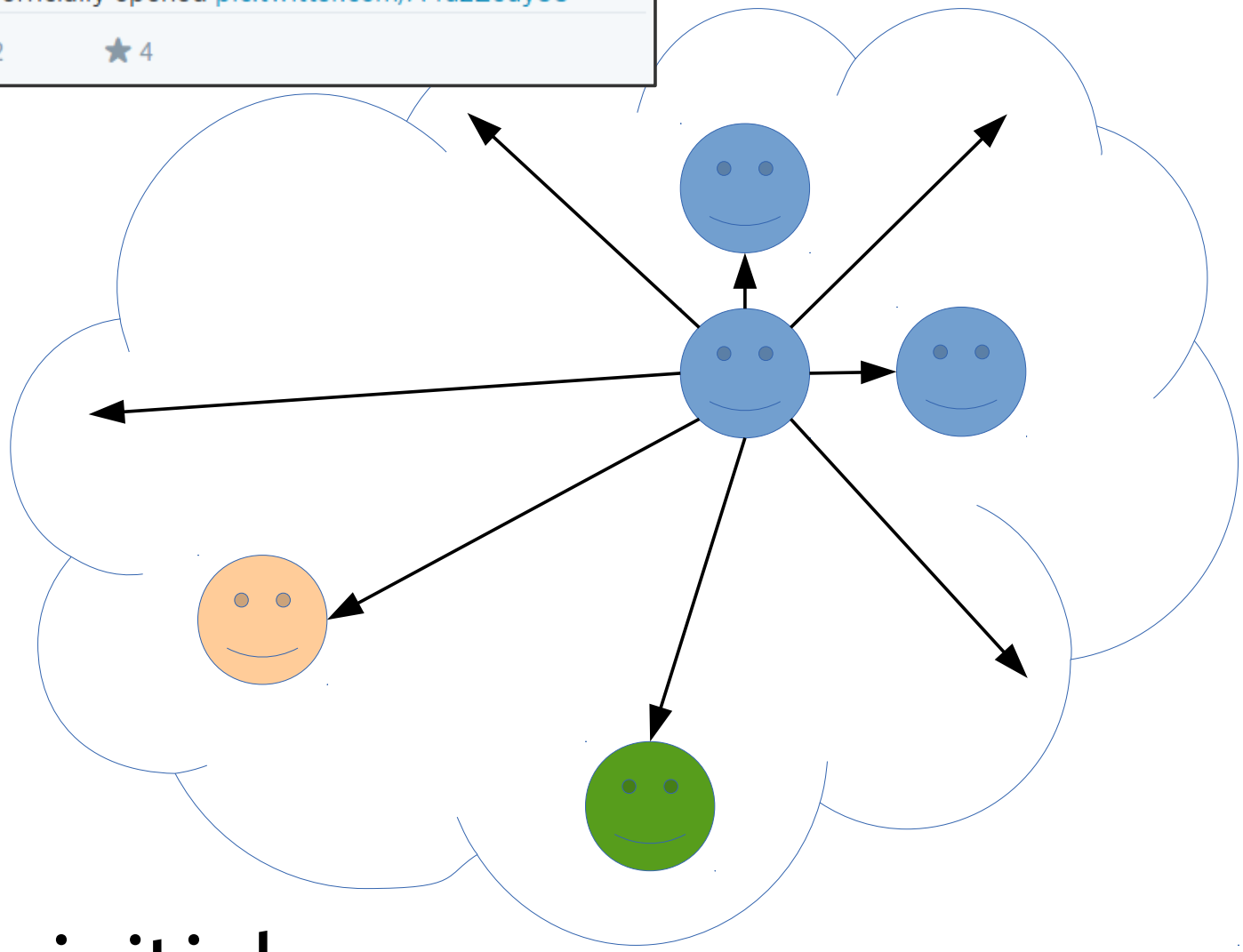
- With social media data, we can distinguish the *contexts* in which feature counts appear.
- One way to define context is by the *intended audience*.
- If a variable's frequency depends on context, this suggests social marking.

(with Umashanthi Pavalanathan)


Methods XV @MethodsXV · May 15
Our full programme will follow in couple of days! We're very excited about it - so many great talks!
3
2

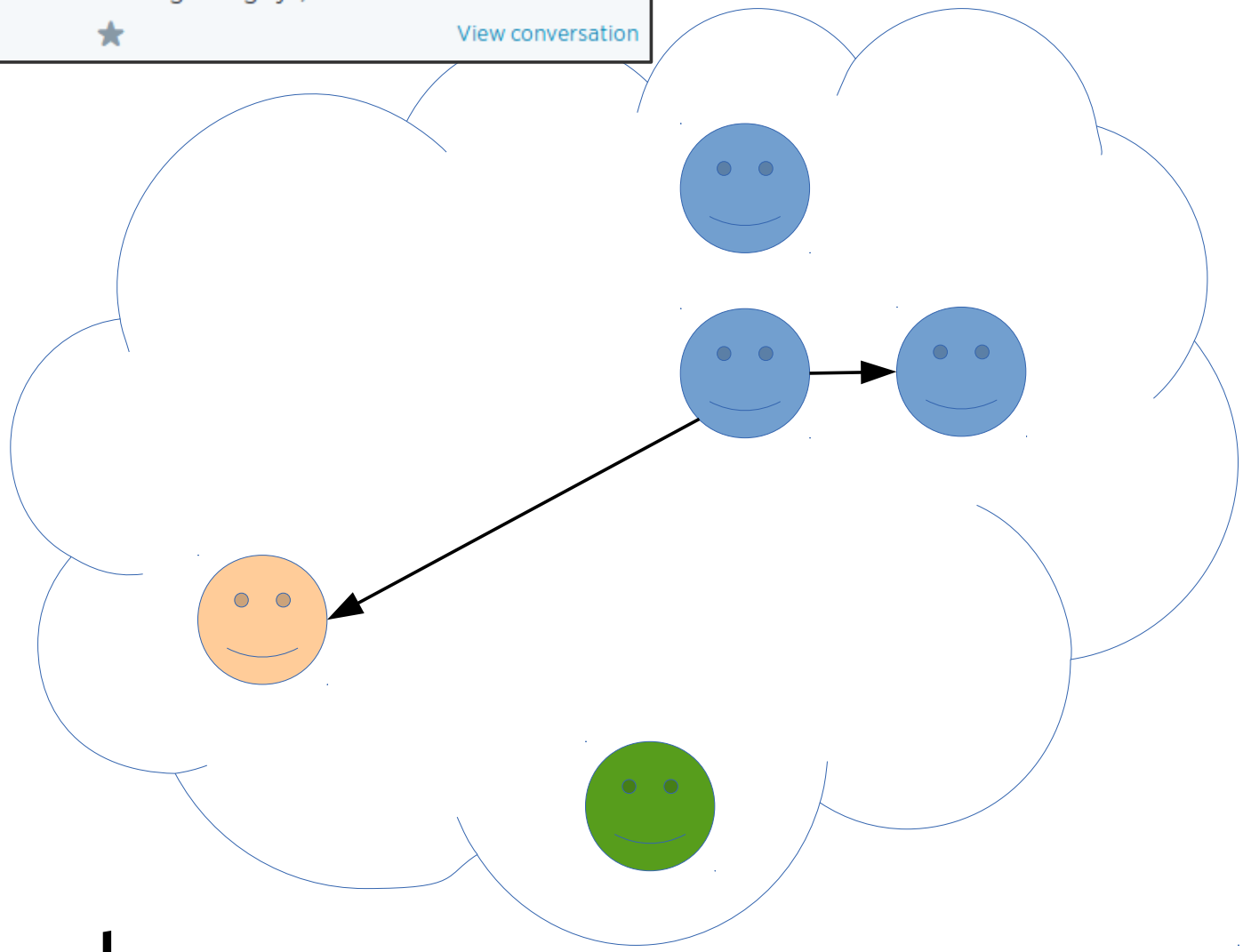
Broadcast





Hashtag-initial


Methods XV @MethodsXV · Aug 10 · ••• More
@ajvYUL @wgi_pr31ea sounds great guys, #MethodsXV it is!
View conversation

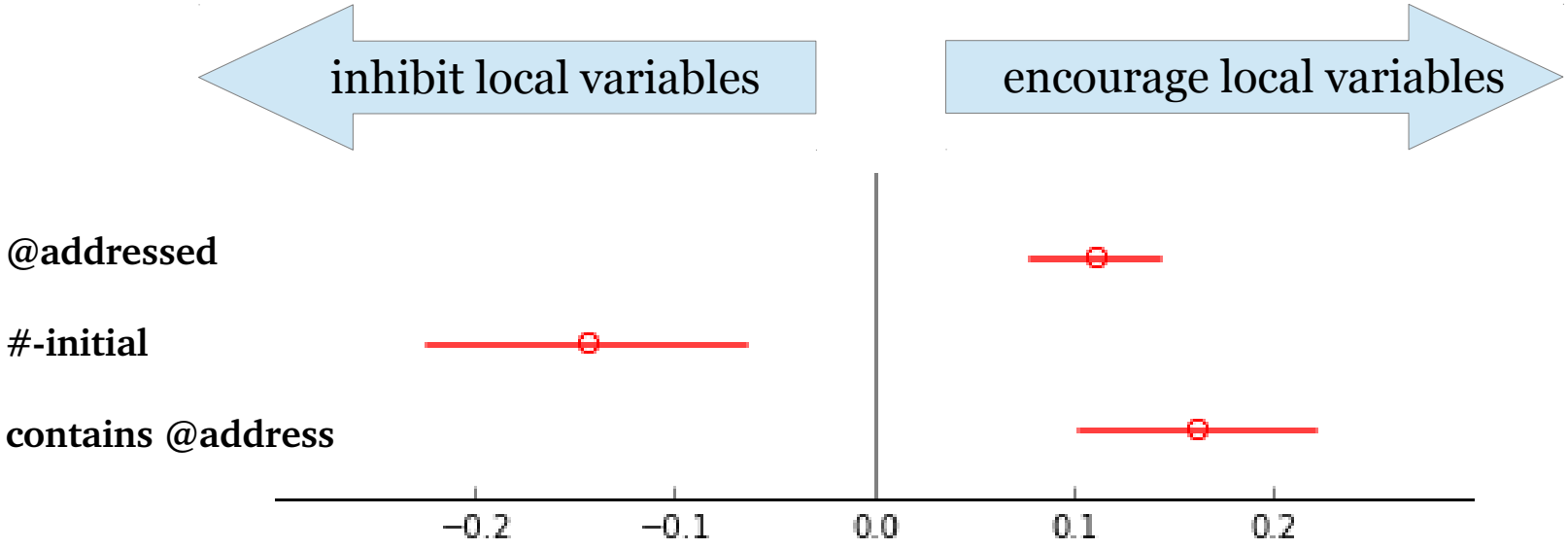


Addressed

Logistic regression

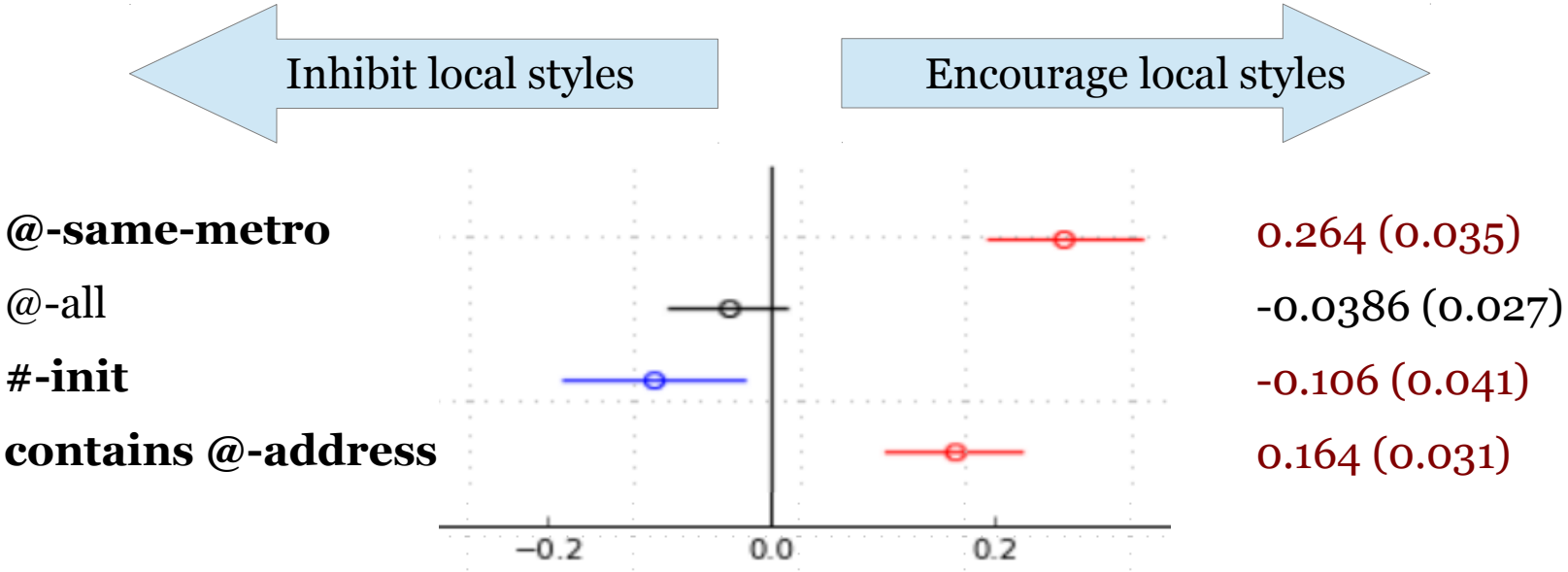
- *Dependent variable*: does the tweet contain a local word (e.g., lbvs, hella, jawn)
- *Predictors*
  - **Message type**: broadcast, addressed, #-initial
  - **Controls**: message length, author statistics

Results



- The more specific the audience, the more likely are local variables.
- Does it matter *who* is addressed?

Results



- Local variables are specifically encouraged when replying to users from the same metro area.
- These results are consistent with these variables being socially marked.

What should we do with a billion words?

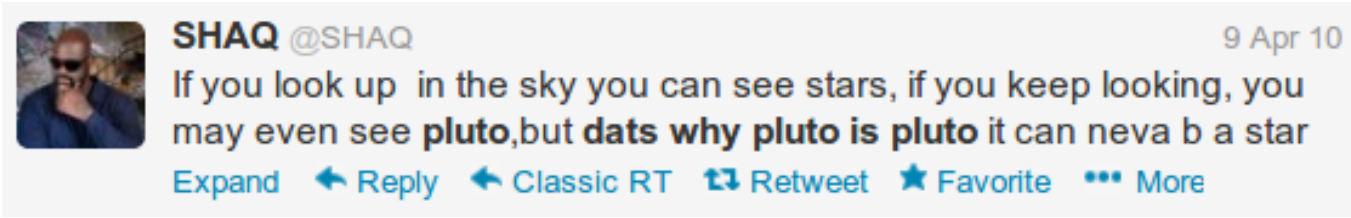
1. Map known dialect terms.
2. Search for unknown dialect terms.
3. Model change over time.

What should we do with a billion words?

1. Map known dialect terms.
2. Search for unknown dialect terms.
3. Model change over time.
4. Investigate how speech affects spelling.

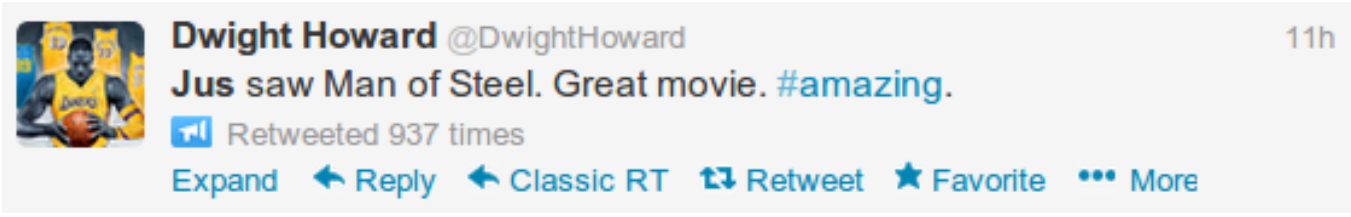
Traces of speech in social media writing

- Th-stopping (th $\rightarrow$ d)



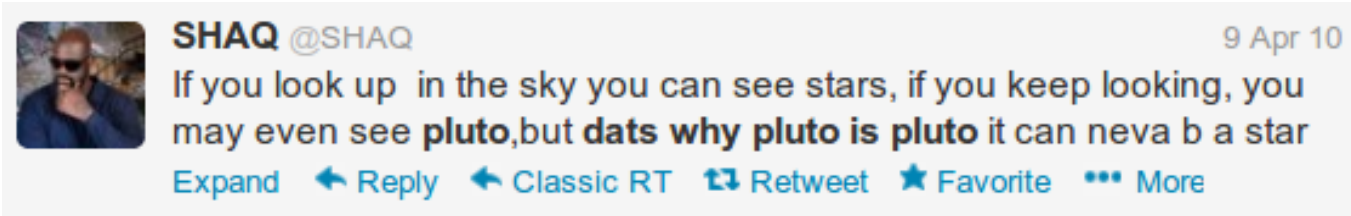
- G-dropping

- (T,D)-deletion



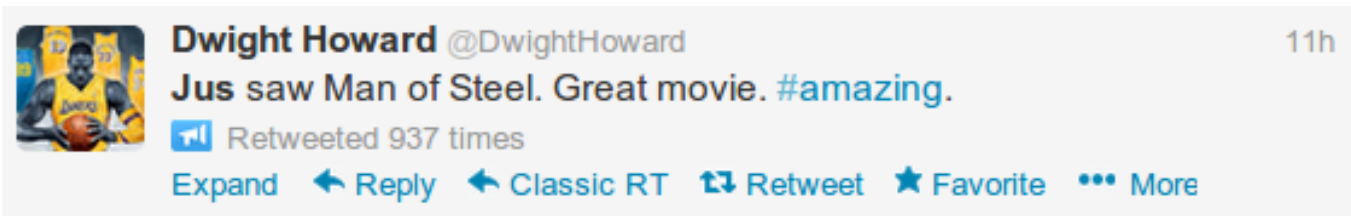
Traces of speech in social media writing

- **Th-stopping** (th $\rightarrow$ d)

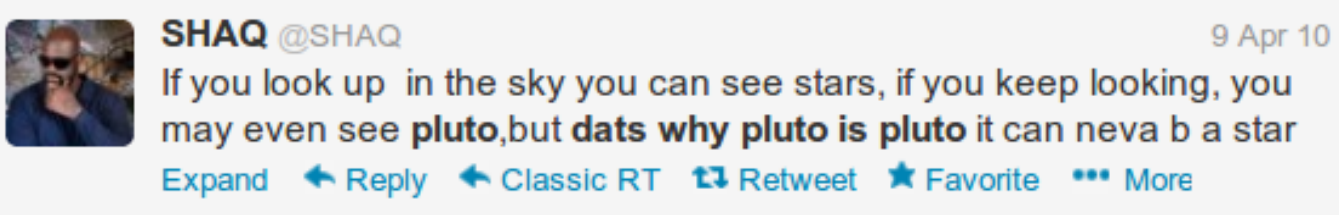


- G-dropping

- (T,D)-deletion

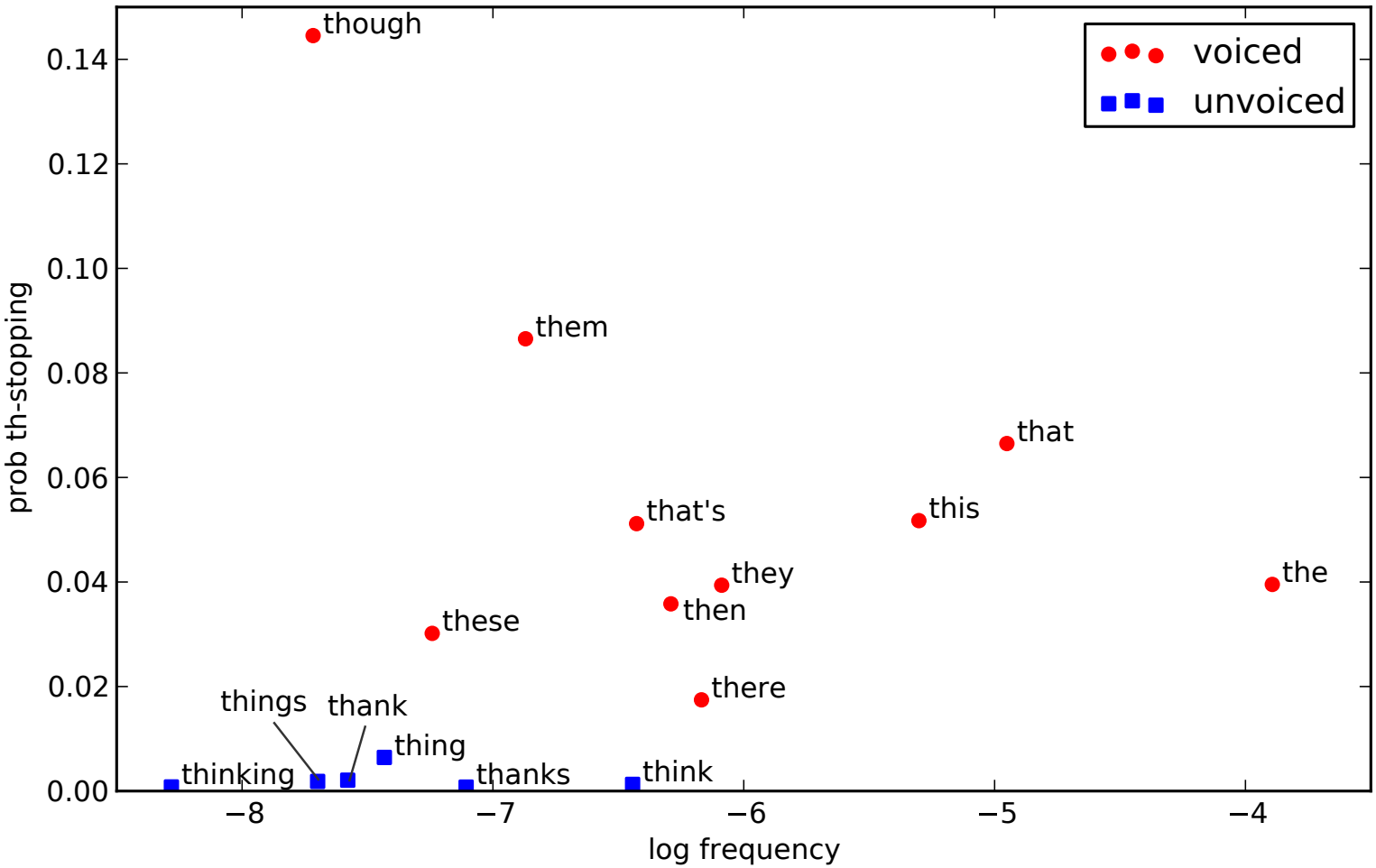


Th-stopping



- Informally, replacement of th with t or d
- But not all th are created equal!
  - In thing, it's unvoiced, θ
  - In this, it's voiced, ð
- In spoken English, voiced ð is much more likely to be stopped. *What about in writing?*

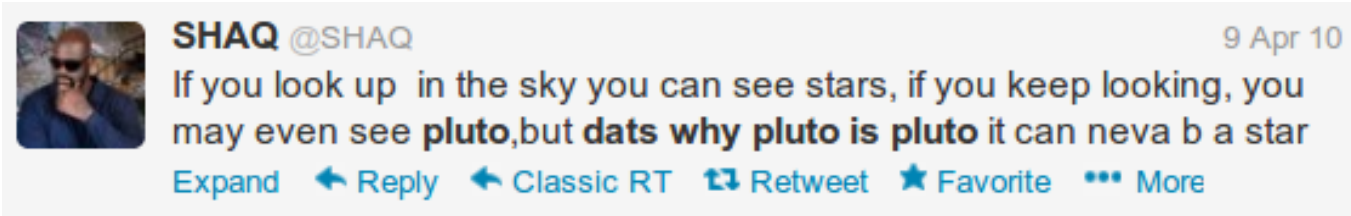
Th-stopping: type-level analysis



Twitter writers are consistently more likely to stop the voiced th, just like in speech.

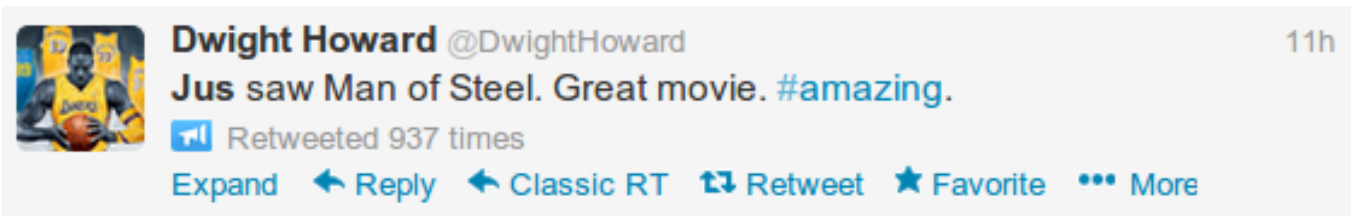
Traces of speech in social media writing

- **Th-stopping** (th $\rightarrow$ d)



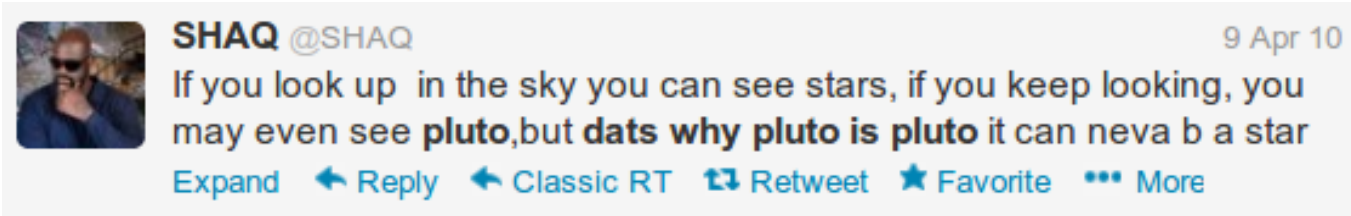
- G-dropping

- (T,D)-deletion



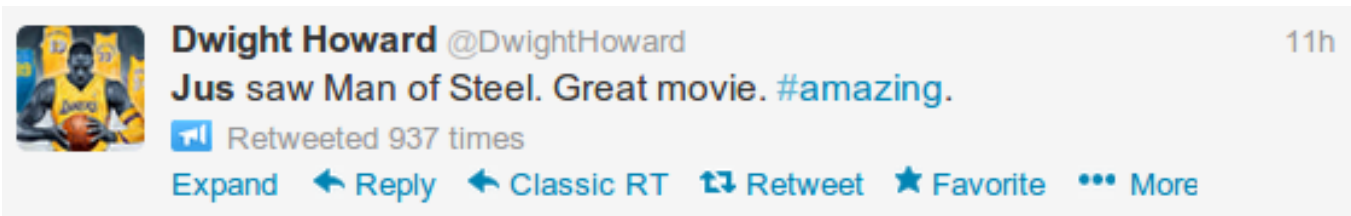
Traces of speech in social media writing

- **Th-stopping** (th $\rightarrow$ d): phonemes matter.



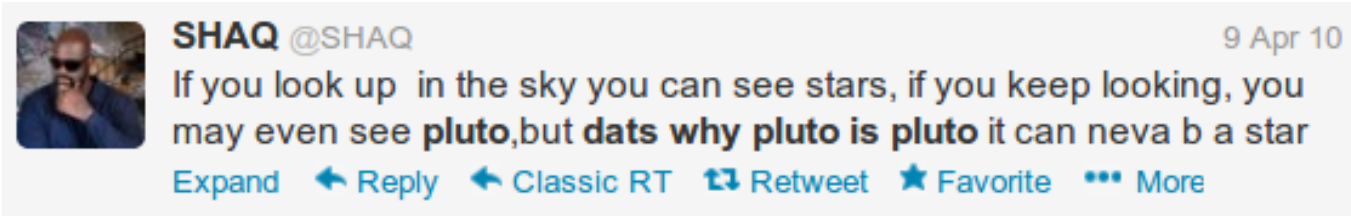
- G-dropping

- (T,D)-deletion



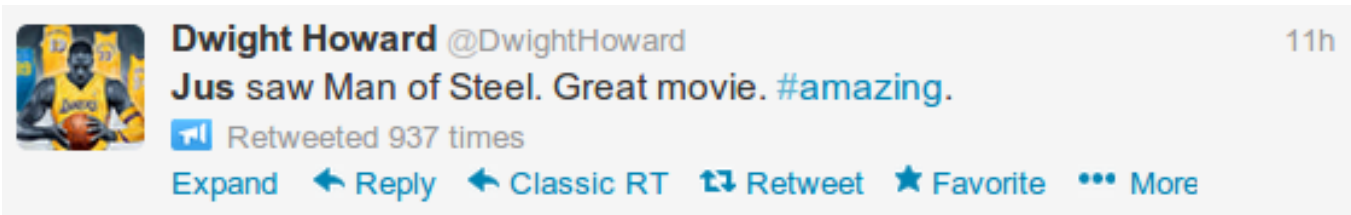
Traces of speech in social media writing

- Th-stopping (th $\rightarrow$ d): phonemes matter.



- **G-dropping**

- (T,D)-deletion



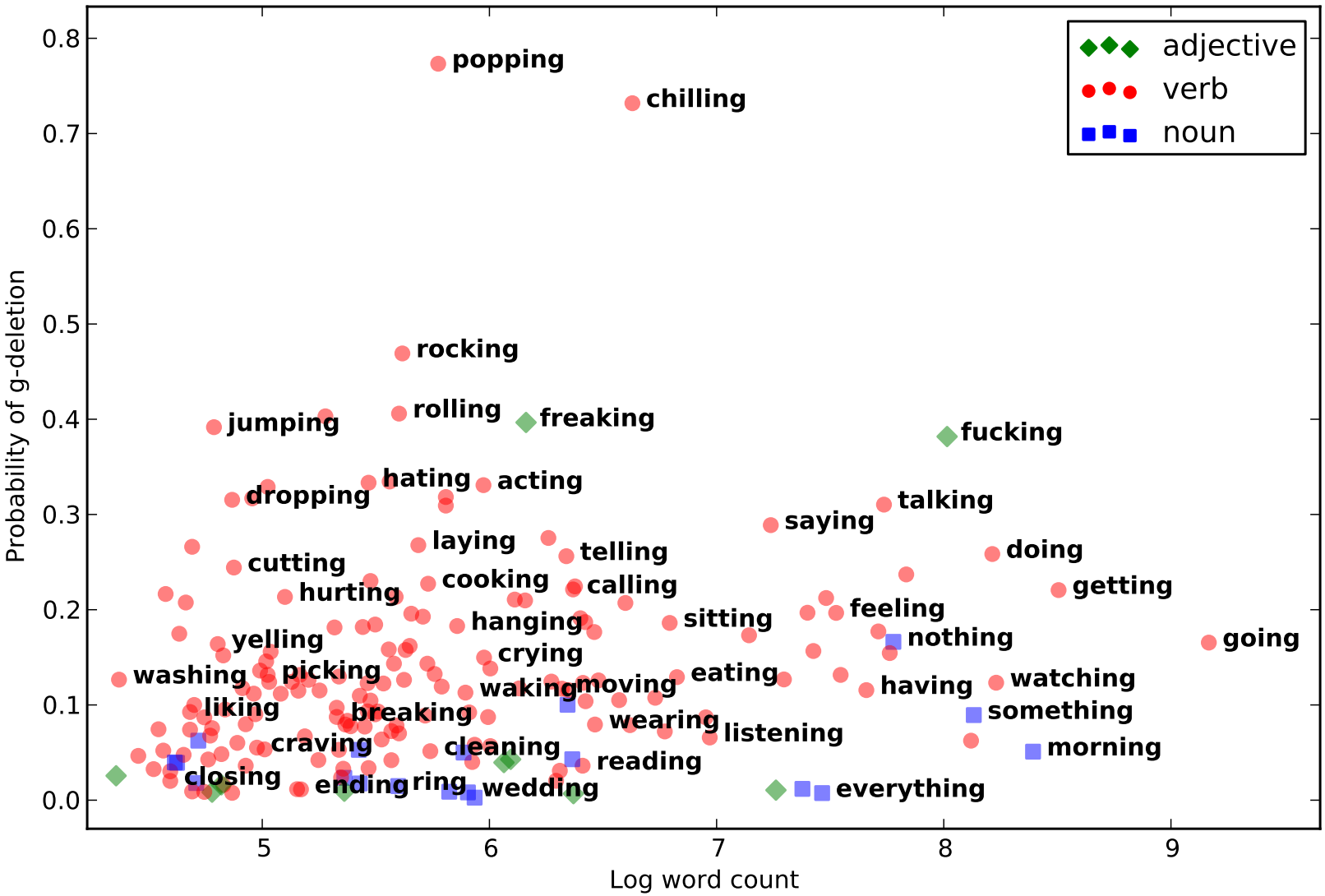
G-dropping

- In speech, "g" is deleted more often from verbs. *What about in writing?*

G-dropping

- In speech, "g" is deleted more often from verbs. *What about in writing?*
- Corpus: 120K tokens of top 200 unambiguous -ing words (ex. king, thing, sing)
- Part-of-speech tags from CMU Twitter tagger (Gimpel et al 2011).

G-dropping: type-level analysis



(Colored by most common POS tag)

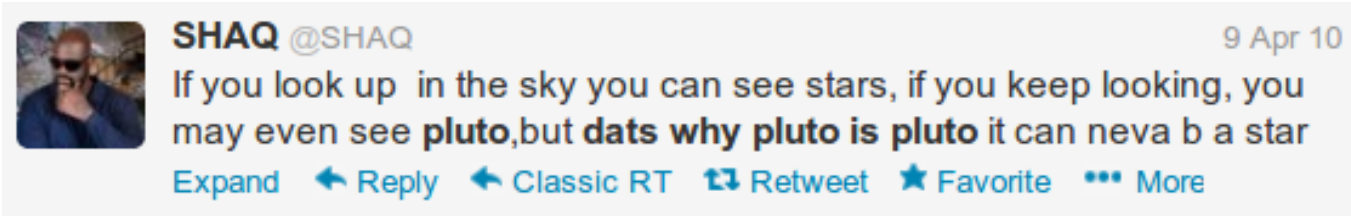
G-dropping: token-level analysis

	Weight	Log odds	%	N
Message type				
Conversational	.567	.27	.205	36,974
Broadcast	.433	-.27	.165	75,919
Number of syllables				
> 1	.959	3.161	.102	4089
1	.041	-3.161	.181	108,804
Part of speech				
Verb	.544	.176	.200	89,173
Noun	.508	.032	.083	18,756
Adjective	.448	-.208	.149	4,964
Total			.178	112,893

Table : Variable rule analysis of factors for g-deletion

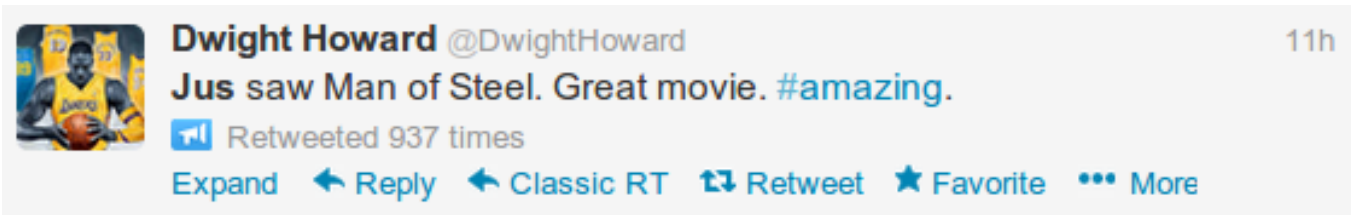
Traces of speech in social media writing

- Th-stopping (th $\rightarrow$ d): phonemes matter.



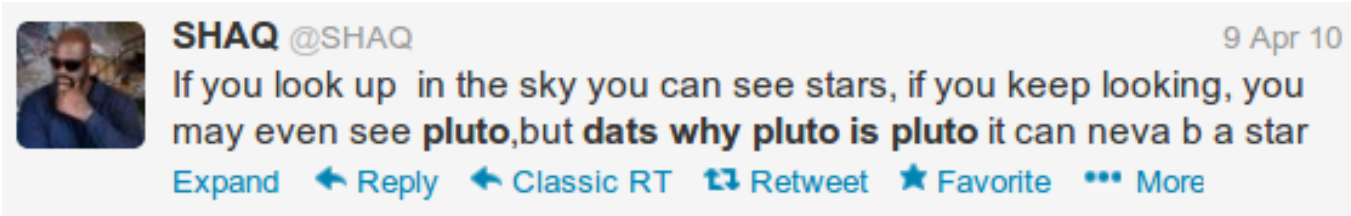
- **G-dropping**

- (T,D)-deletion



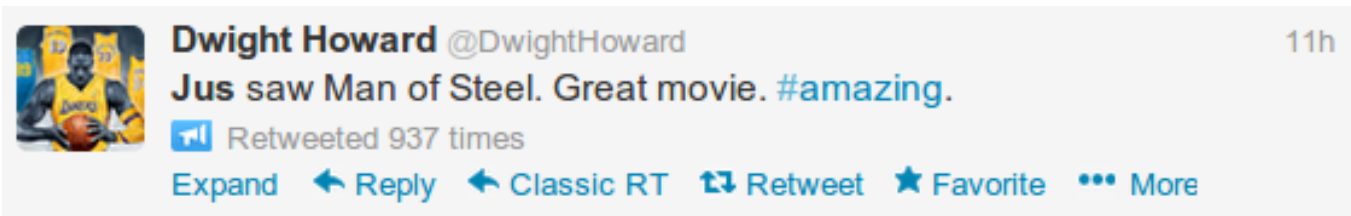
Traces of speech in social media writing

- Th-stopping (th $\rightarrow$ d): phonemes matter.



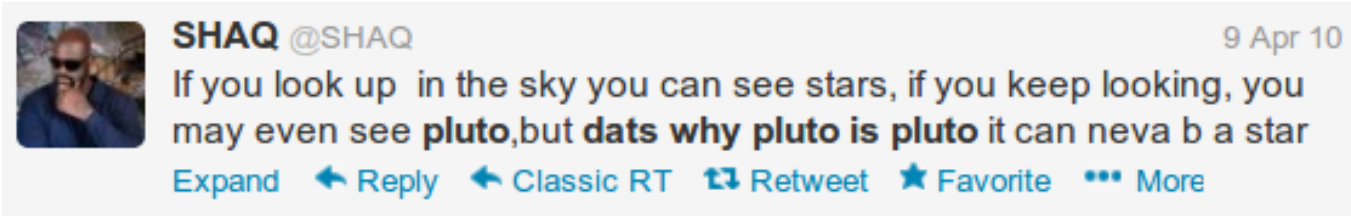
- **G-dropping**: syntax matters.

- (T,D)-deletion



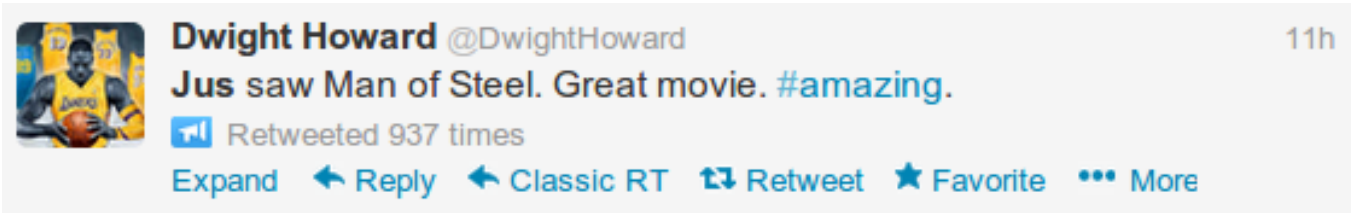
Traces of speech in social media writing

- Th-stopping (th → d): phonemes matter.

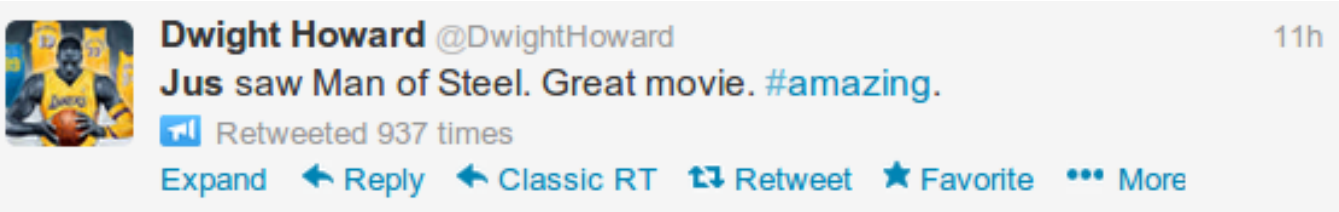


- G-dropping: syntax matters.

- **(T,D)-deletion**



(t,d)-deletion in context



In speech, (t,d)-deletion frequency depends on the phonological context:

$$P(\text{best peaches} \rightarrow \text{bes peaches})$$
$$> P(\text{best apples} \rightarrow \text{bes apples})$$

What about in writing?

(t,d)-deletion: Variable rules analysis

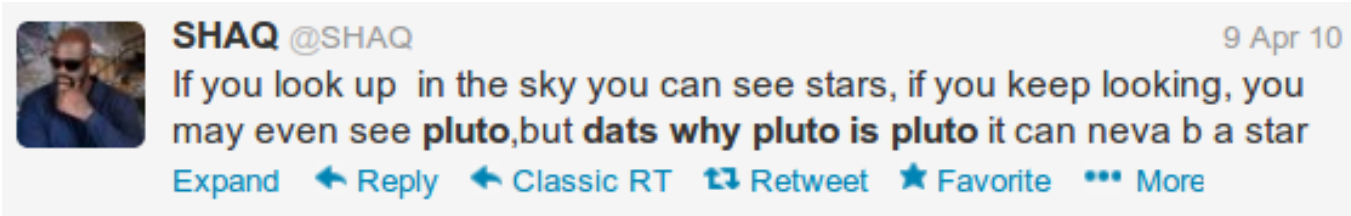
- *Dependent variable*: deletion or not?
- *Fixed effects*
  - Succeeding phoneme (vowel or consonant)
  - Message type
- *Random effects*
  - Frequent bigrams
  - Author
  - Year

(t,d)-deletion: Variable rules analysis

Group		Weight	Log odds	%	N
(t,d): just, old, aint, told, next					
Context	vowel	.467	-.131	.385	9,002
	consonant	.533	.131	.428	80,170
(ing): going, getting, fucking, talking, doing					
Context	vowel	.460	-.159	.477	27,005
	consonant	.540	.159	.506	75,257
non-phonetic: know, love, have, true, maybe					
Context	vowel	[.506]	[.023]	.517	28,856
	consonant	[.494]	[-.023]	.486	80,170

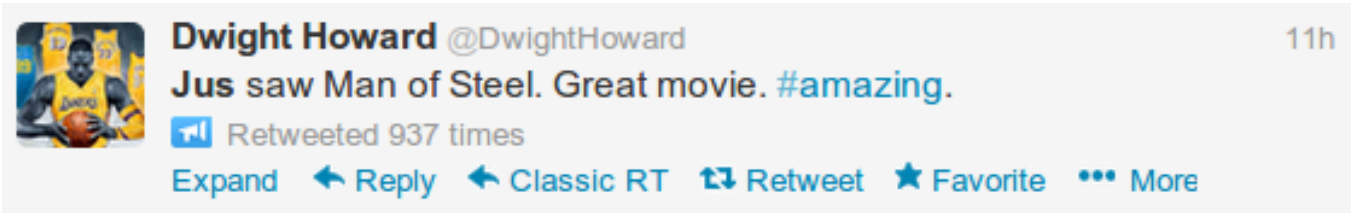
Traces of speech in social media writing

- Th-stopping (th → d): phonemes matter.



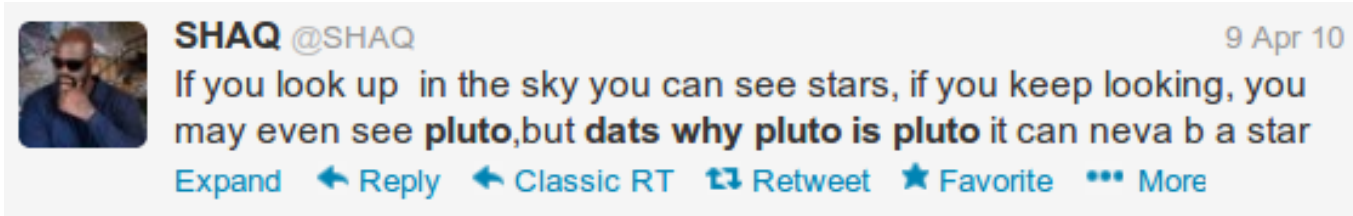
- G-dropping: syntax matters.

- **(T,D)-deletion**



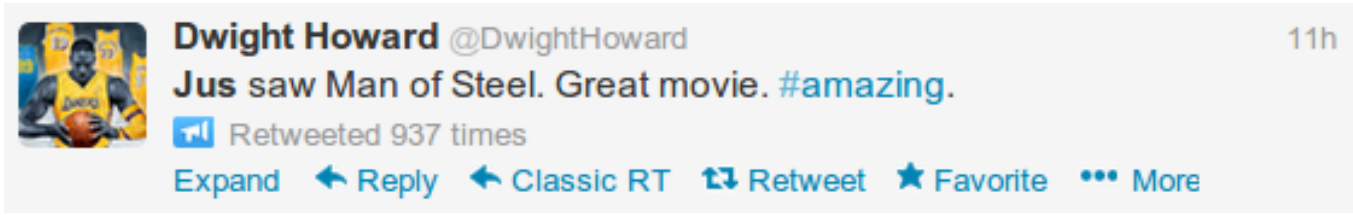
Traces of speech in social media writing

- Th-stopping (th → d): phonemes matter.



- G-dropping: syntax matters.

- **(T,D)-deletion**: phonology matters.



What should we do with a billion words?

1. Map known dialect terms.
2. Search for unknown dialect terms.
3. Model change over time.
4. Investigate how speech affects spelling.

What should we do with a billion words?

1. Map known dialect terms.
2. Search for unknown dialect terms.

   "Netspeak" dialect often has regional characteristics.
3. Model change over time.
4. Investigate how speech affects spelling.

What should we do with a billion words?

1. Map known dialect terms.
2. Search for unknown dialect terms.

   "Netspeak" dialect often has regional characteristics.
3. Model change over time.

   Real-time language change is sensitive to racial demographics, and variation is socially marked.
4. Investigate how speech affects spelling.

What should we do with a billion words?

1. Map known dialect terms.
2. Search for unknown dialect terms.

   "Netspeak" dialect often has regional characteristics.
3. Model change over time.

   Real-time language change is sensitive to racial demographics, and variation is socially marked.
4. Investigate how speech affects spelling.

   Social media writing transcribes phonological variation with a surprising degree of fidelity.

What should you do with a trillion words?

- Move beyond the North American context.

  (see Swanenberg; Perea & Tinoco; Ernestina & Pereira)

- More systematically investigate the relationship between writing and spoken dialect.
- Morphosyntactic variation (see Ernestina and Pereira)
- Better statistics for "found" data
  - Causal inference
  - Stratified sampling
- Relate Twitter to other social media, and to mass media.
- Integrate with other methodologies, such as ethnography.

What should you do with a trillion words?

- Move beyond the North American context.

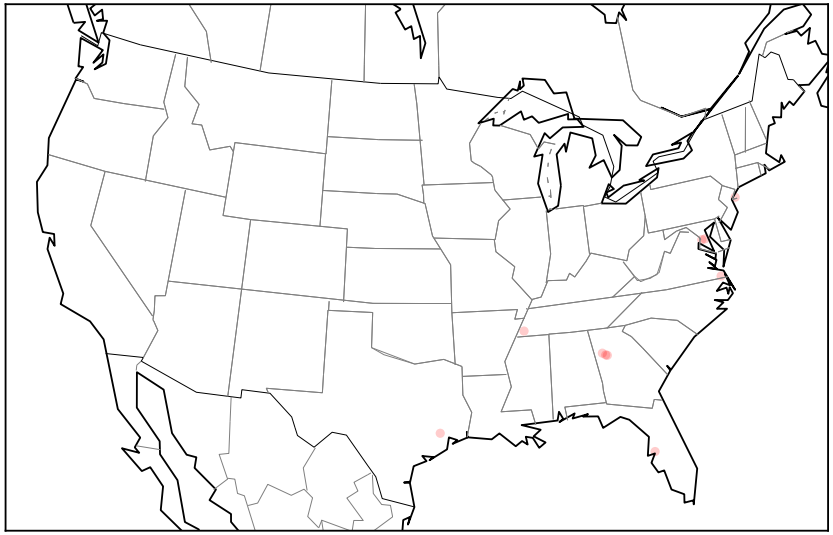
  (see Swanenberg; Perea & Tinoco; Ernestina & Pereira)

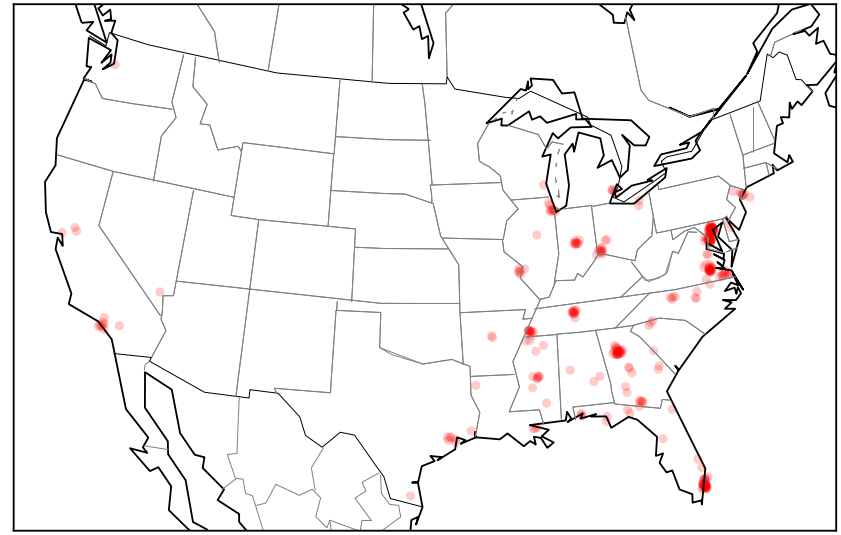
- More systematically investigate the relationship between writing and spoken dialect.
- Morphosyntactic variation (see Ernestina and Pereira)
- Better statistics for "found" data
  - Causal inference
  - Stratified sampling
- Relate Twitter to other social media, and to mass media.
- Integrate with other methodologies, such as ethnography.

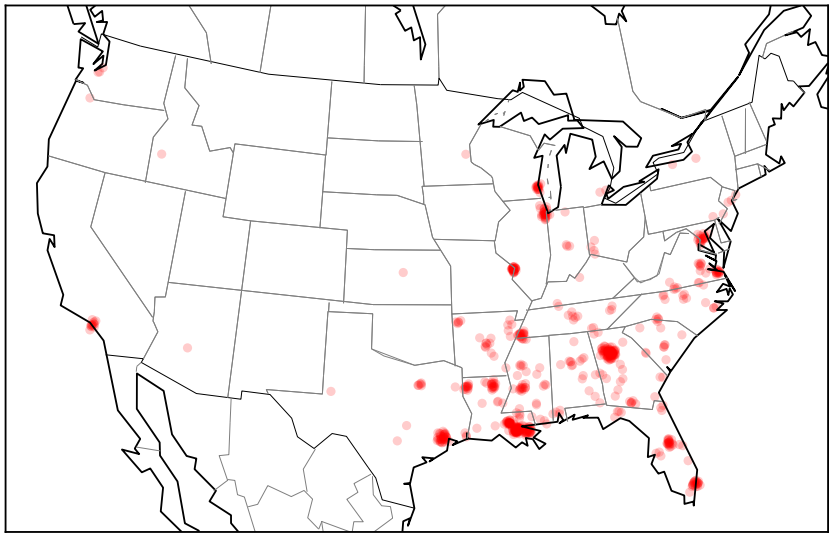
Thank you!

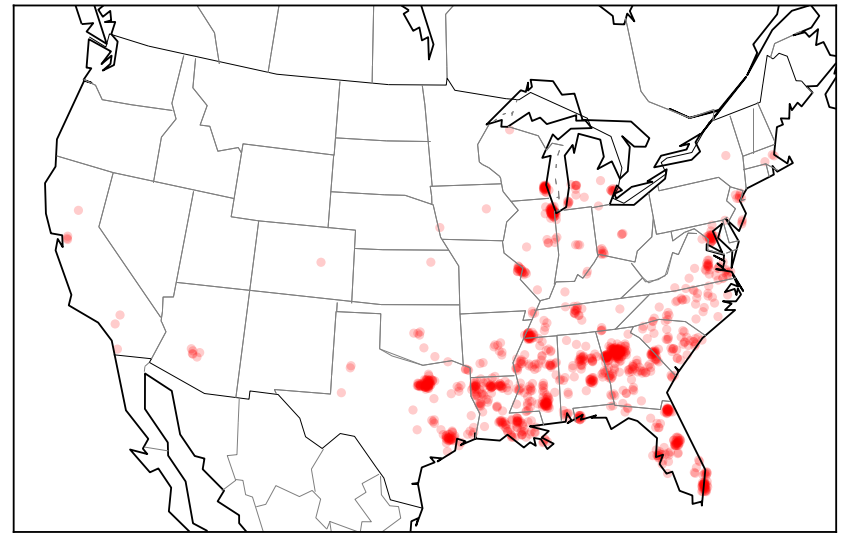
Change from 2009-2012: ion

i must be blind cuz ion see it









Change from 2009-2012: -_ _-

flight delayed -_ _- just what i need