Social Text Data Analysis

Brendan O'Connor
Machine Learning Department
Carnegie Mellon University

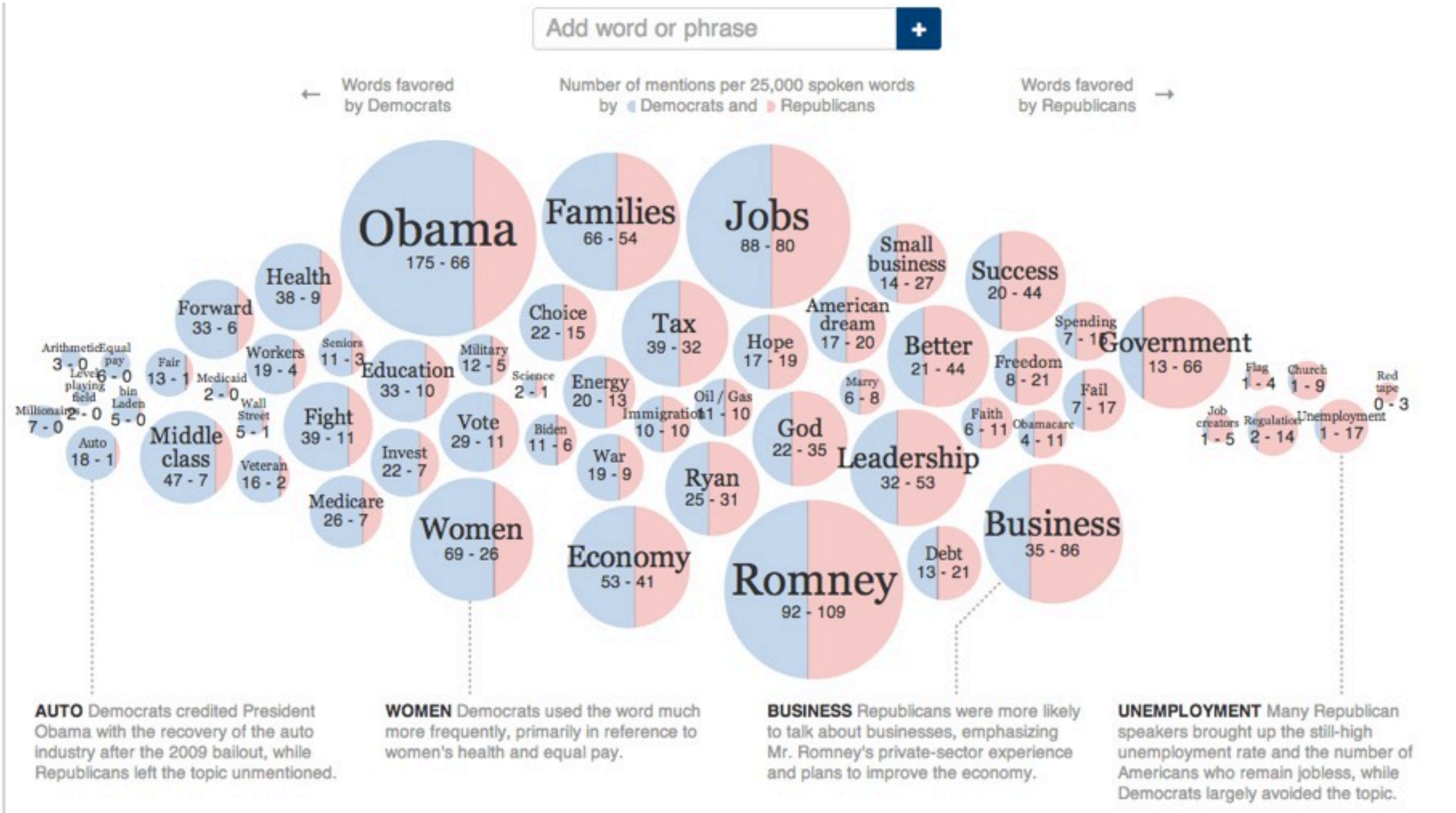
Talk at UChicago
Computational Social Science Workshop seminar

November 16, 2012

What can **statistical text analysis** tell us about **society**?

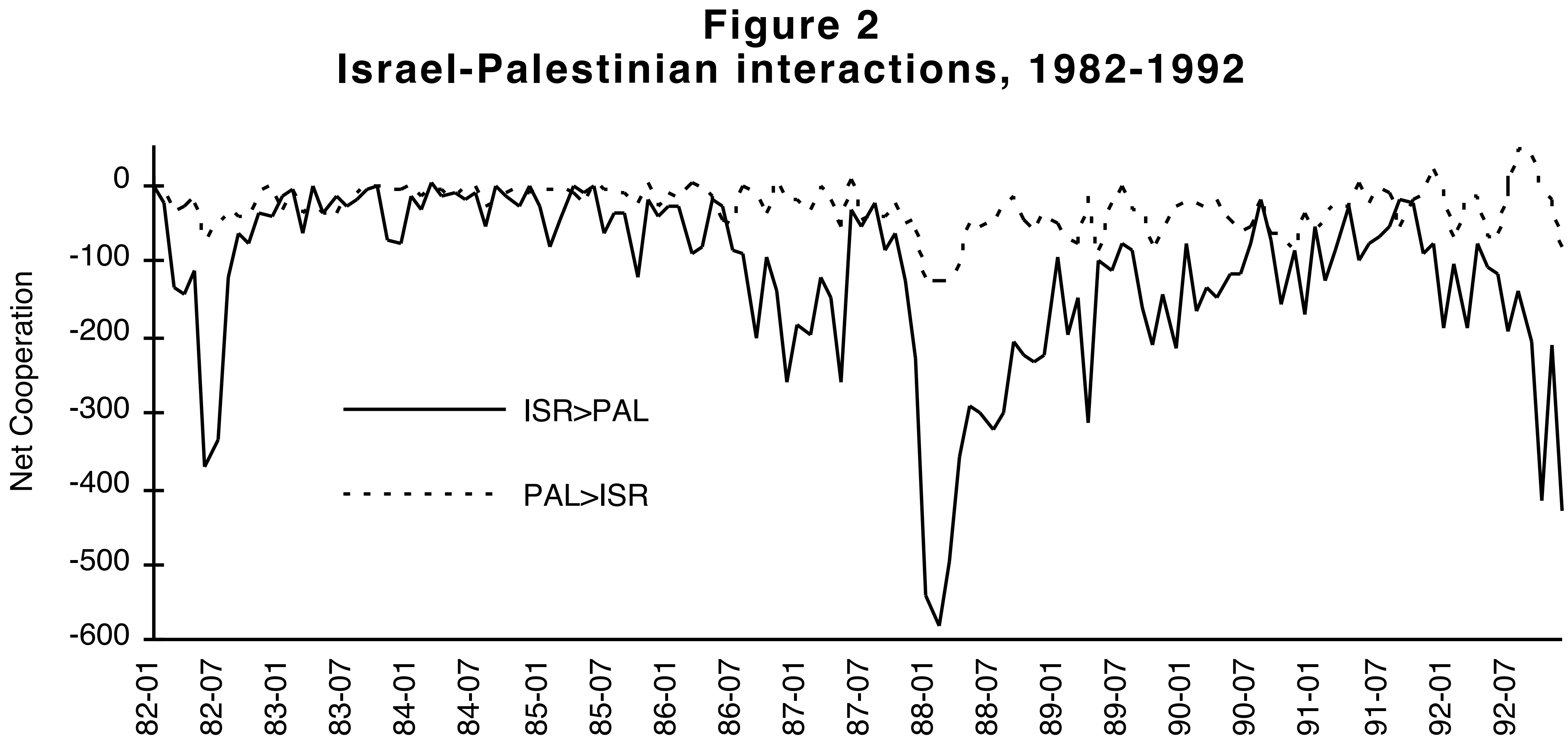
- Manual content analysis: analyze ideas, concepts, opinions etc. in text. Long and rich history *(Krippendorff 2012)* -- but very labor-intensive
- From manual to automated content analysis
  - Quantitative comparisons
  - Pattern recognition
  - Qualitative drilldowns
- Many emerging examples of automated text content analysis
  [Political science, media studies, economics, psychology, sociology of science, sociolinguistics, public health, history, literature...]
- Appropriating tools from natural language processing, information retrieval, data mining, machine learning as quantitative social science methodology

Text as measurement?: concepts



U.S. convention speeches' word frequencies, by party
http://www.nytimes.com/interactive/2012/09/06/us/politics/convention-word-counts.html

Text as measurement: events



Automatic event extraction from news reports (Schrodt 1993)

Taxonomy of text analysis methods

Vanilla topic models

... with (partial) labels

... with lexical constraints

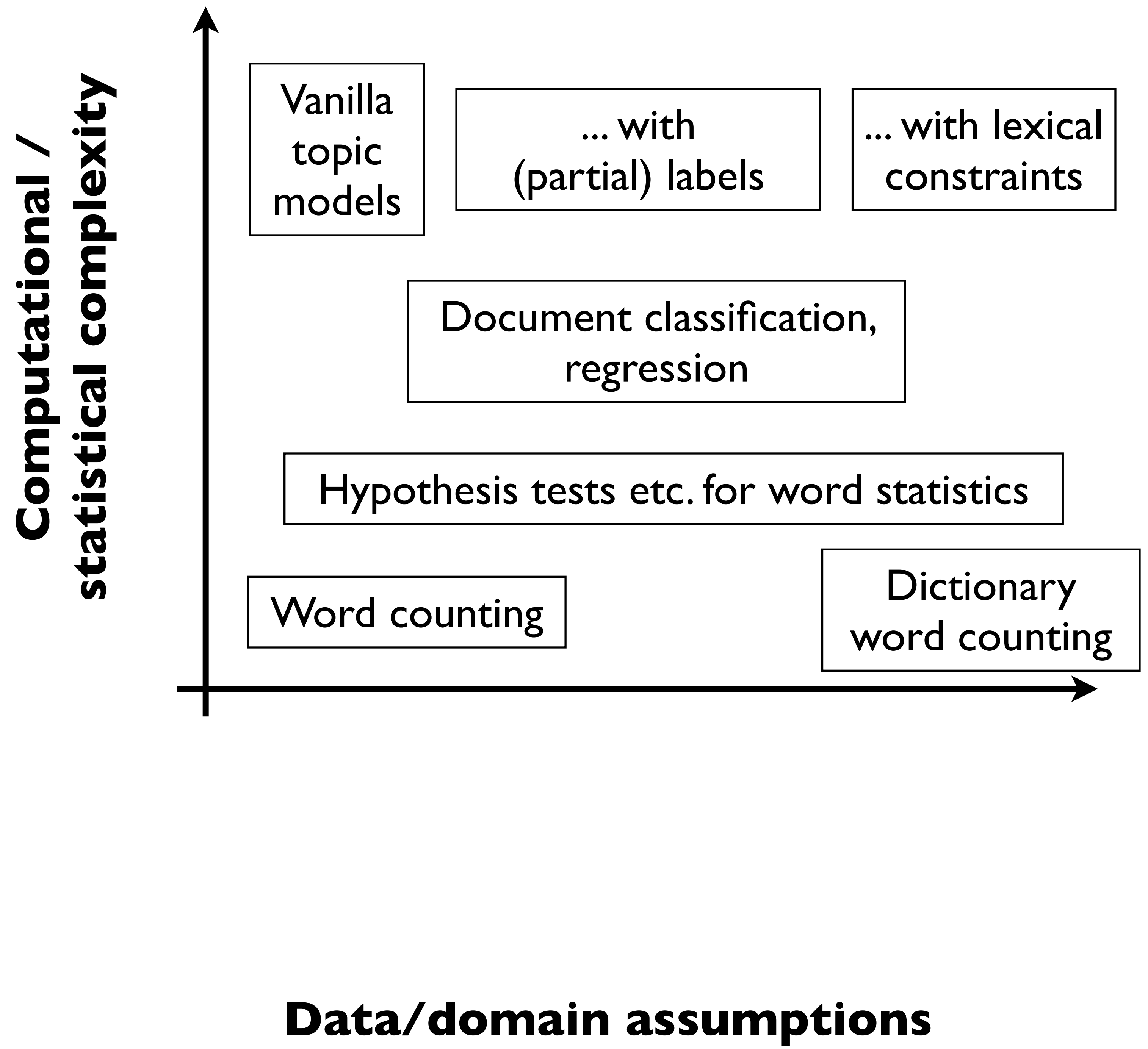
Document classification, regression

Hypothesis tests etc. for word statistics

Word counting

Dictionary word counting

Taxonomy of text analysis methods



Data and Prior Knowledge

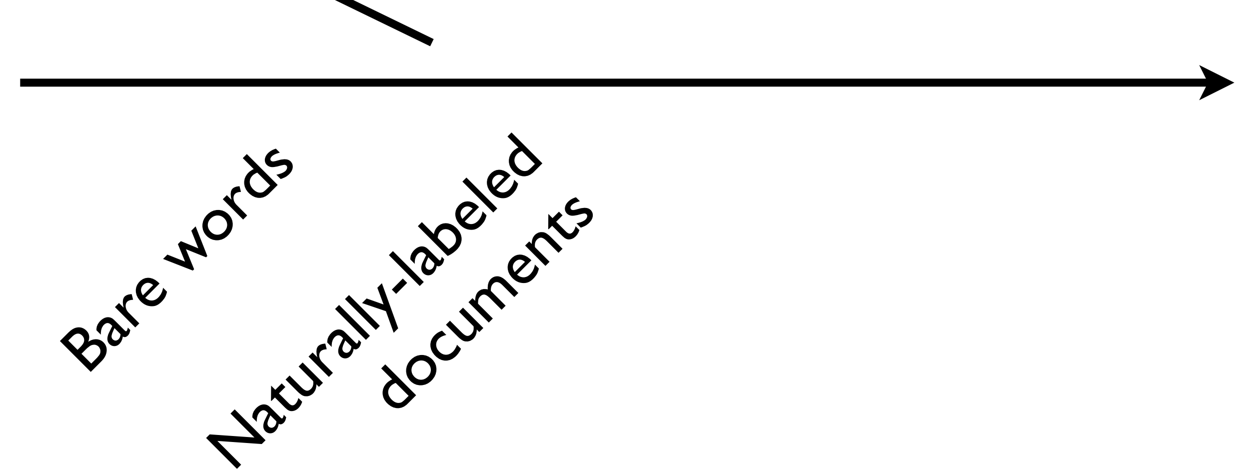
Bare words

Can interpret text
only through context

Data/domain assumptions

Data and Prior Knowledge

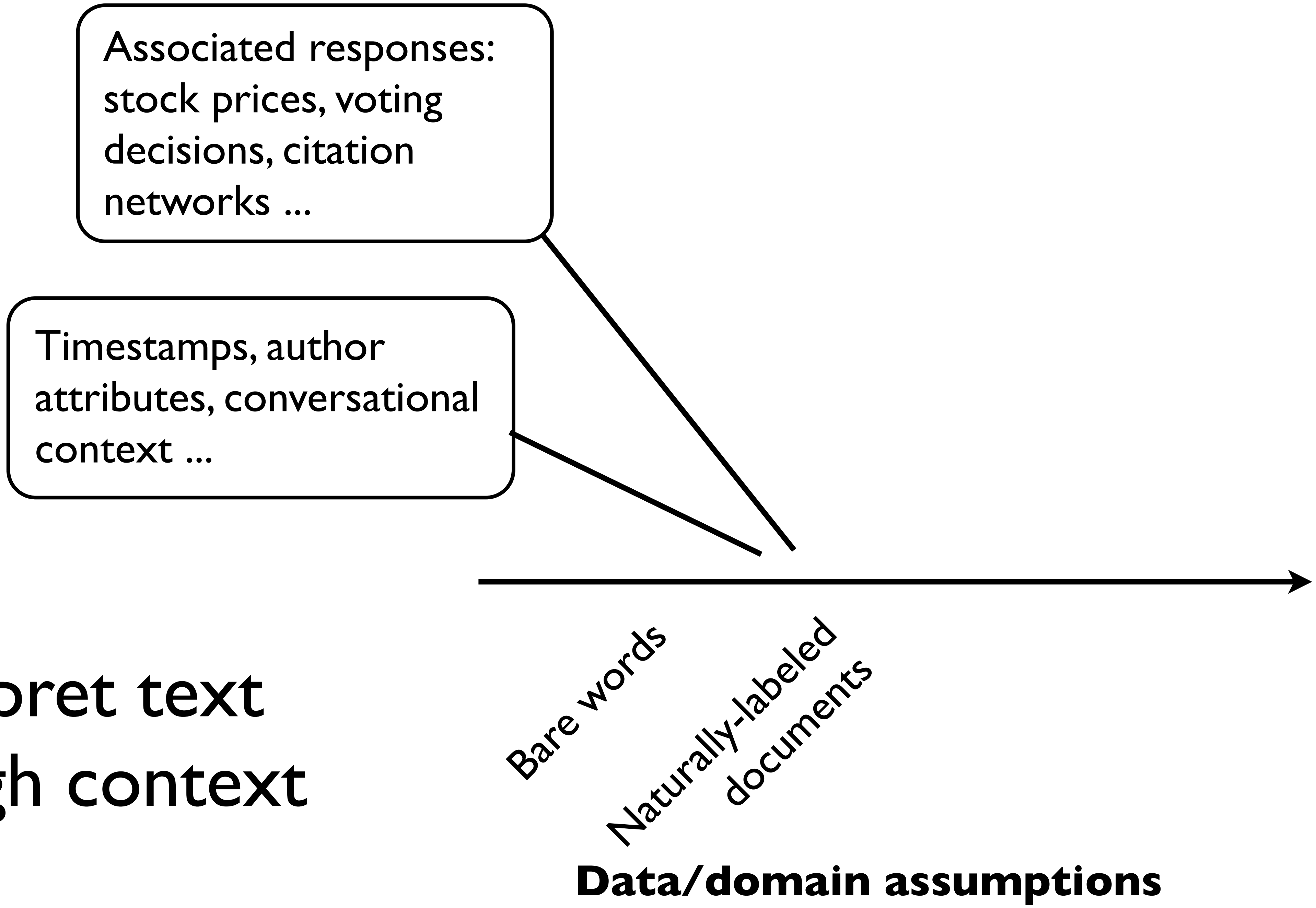
Timestamps, author attributes, conversational context ...



Data/domain assumptions

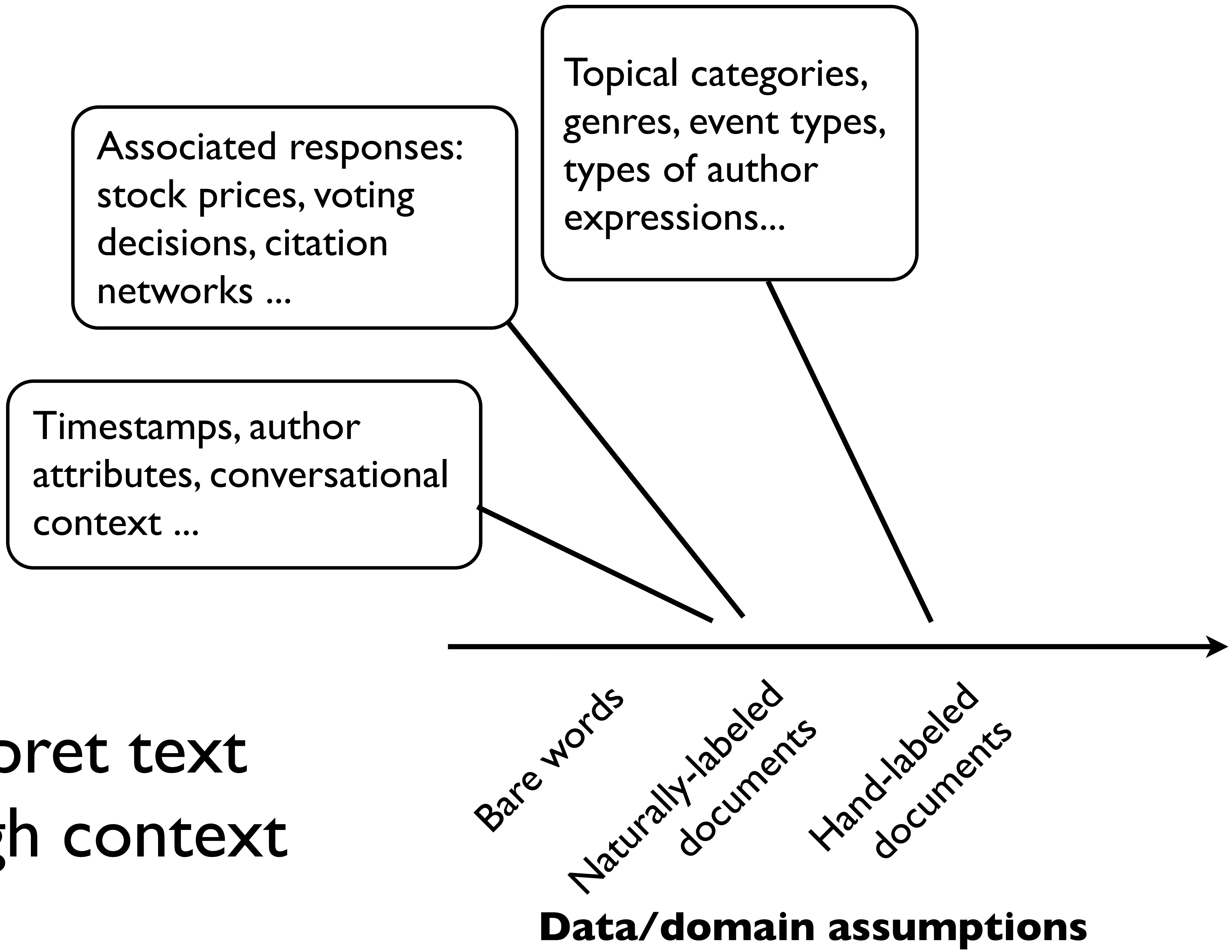
Can interpret text only through context

Data and Prior Knowledge



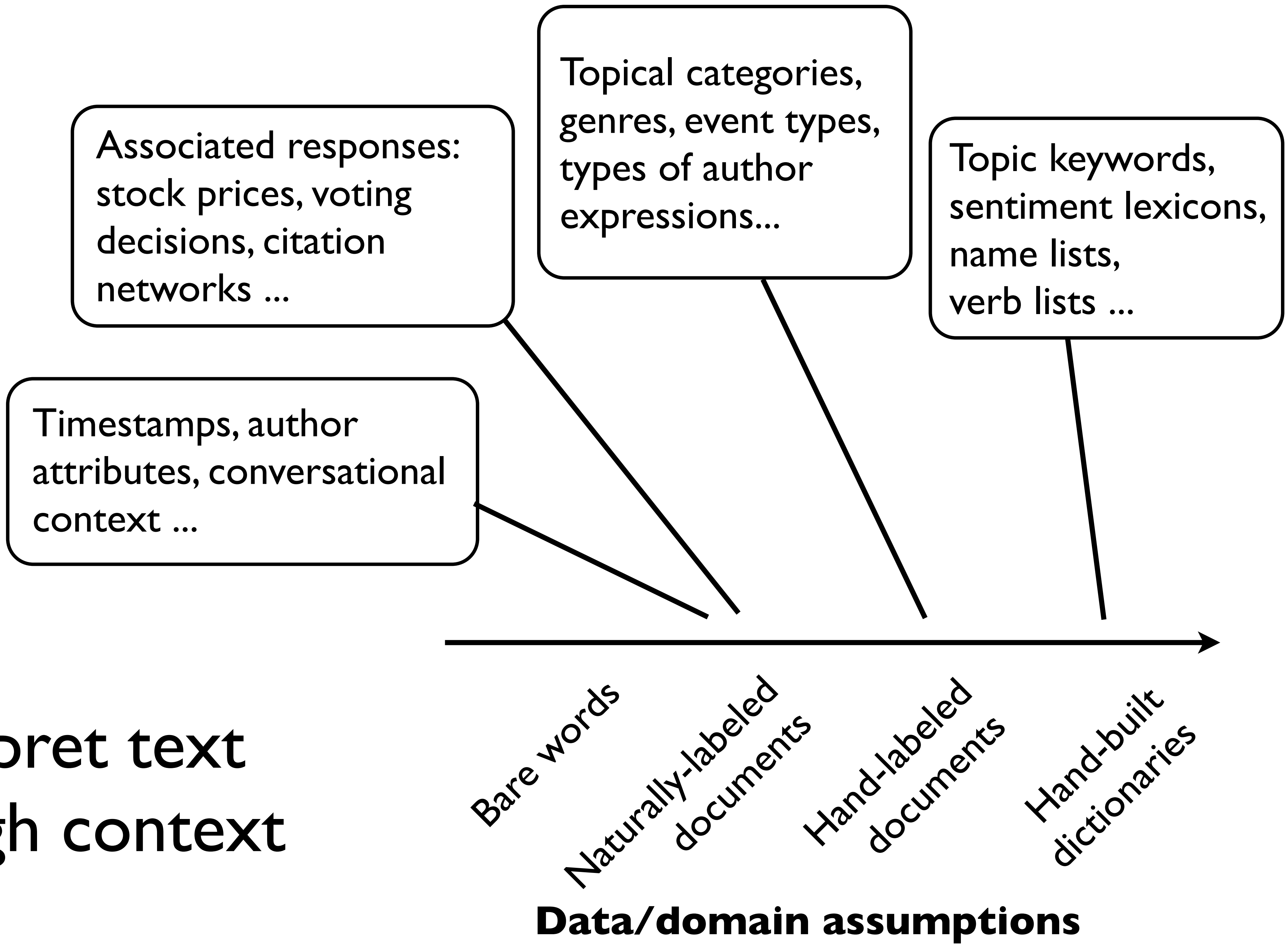
Can interpret text only through context

Data and Prior Knowledge



Can interpret text only through context

Data and Prior Knowledge

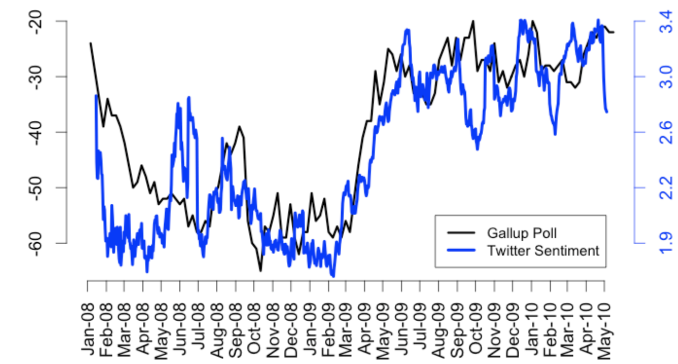


Can interpret text only through context

Detecting cultural phenomena in textual social media

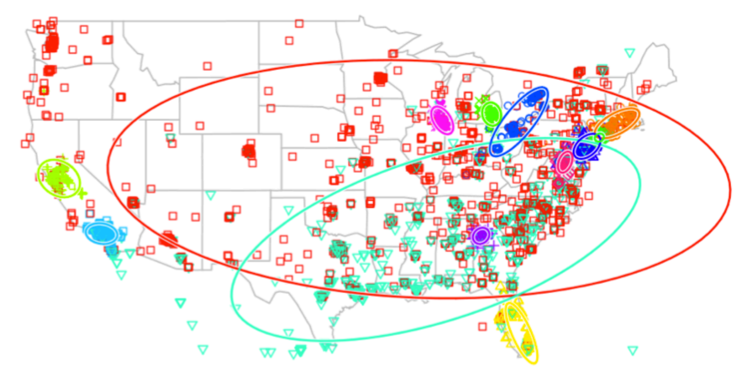
Opinions and Time

Brendan O'Connor, Ramnath Balasubramanyan, Bryan Routledge, Noah Smith



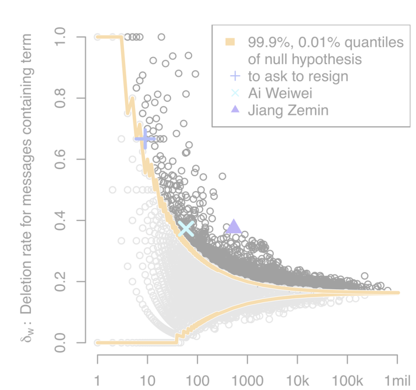
Language and Geography

Jacob Eisenstein, Brendan O'Connor, Noah Smith, Eric Xing



Internet Censorship

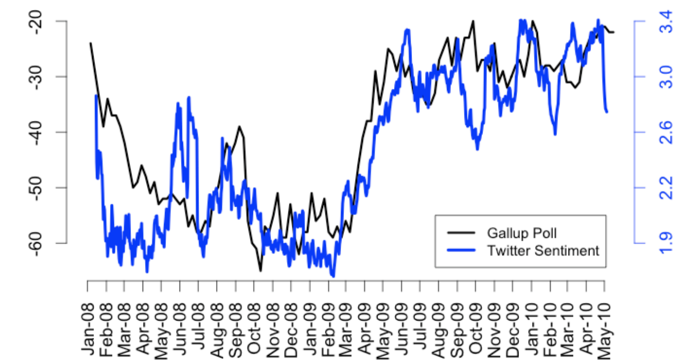
David Bamman, Brendan O'Connor, Noah Smith



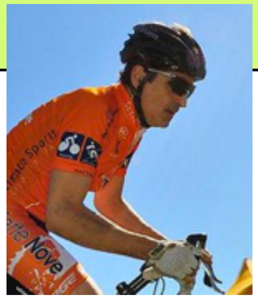
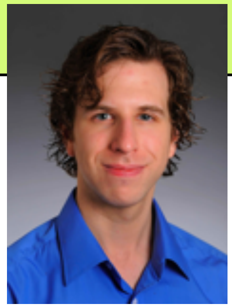
Detecting cultural phenomena in textual social media

Opinions and Time

Brendan O'Connor, Ramnath Balasubramanyan,
Bryan Routledge, Noah Smith



From Tweets to Polls:
Linking Text Sentiment
to Public Opinion Time Series.
ICWSM 2010

Measuring public opinion
through social media?

People in U.S.

STAN HONDA / AFP/Getty Images

Measuring public opinion
through social media?

People in U.S.

STAN HONDA / AFP/Getty Images

Query

Measuring public opinion
through social media?

People in U.S.

STAN HONDA / AFP/Getty Images

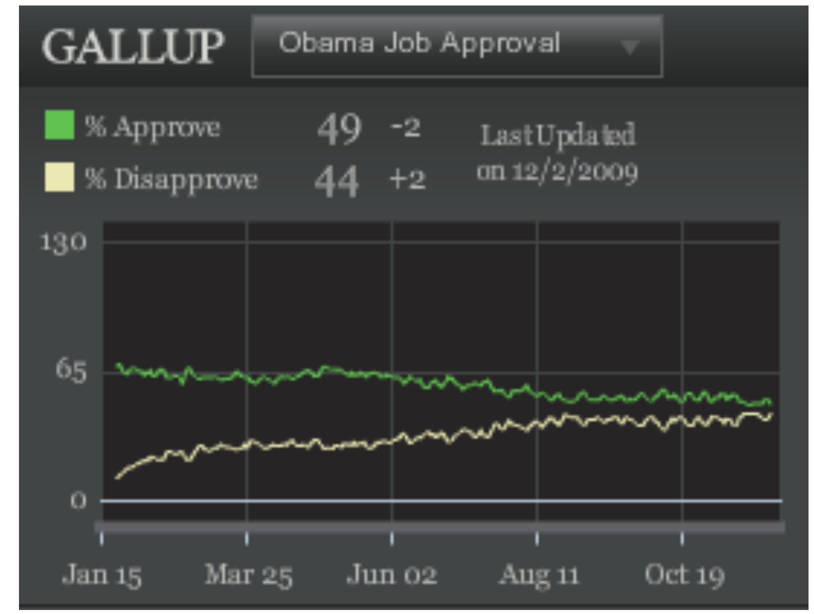
Query

Measuring public opinion
through social media?

People in U.S.

STAN HONDA / AFP/Getty Images

Query



Write

facebook

twitter

Query

Aggregate
Text Sentiment
Measure

Measuring public opinion
through social media?

People in U.S.

STAN HONDA / AFP/Getty Images

Query

Can we derive a similar measurement?

Write

facebook
twitter

Query

Aggregate Text Sentiment Measure

Method

Method

Poll Data

- Consumer confidence, 2008-2009
  - Index of Consumer Sentiment (Reuters/ Michigan)
  - Gallup Daily
- 2008 Presidential Elections
  - Aggregation, Pollster.com
- 2009 Presidential Job Approval
  - Gallup Daily

Which tweets correspond to which polls?

Method

Poll Data

- Consumer confidence, 2008-2009
  - Index of Consumer Sentiment (Reuters/ Michigan)
  - Gallup Daily
- 2008 Presidential Elections
  - Aggregation, Pollster.com
- 2009 Presidential Job Approval
  - Gallup Daily

Topic Keywords

Subset of Gardenhose public tweets over 2008-2009

→ "economy" "jobs" "job"

→ "obama" "mccain"

→ "obama"

Which tweets correspond to which polls?

Analyzed subsets of messages that contained manually selected topic keyword

Method

Poll Data

- Consumer confidence, 2008-2009
  - Index of Consumer Sentiment (Reuters/ Michigan)
  - Gallup Daily
- 2008 Presidential Elections
  - Aggregation, Pollster.com
- 2009 Presidential Job Approval
  - Gallup Daily

Which tweets correspond to which polls?

Topic Keywords

Subset of Gardenhose public tweets over 2008-2009

→ "economy" "jobs" "job"

→ "obama" "mccain"

→ "obama"

Analyzed subsets of messages that contained manually selected topic keyword

Sentiment Analysis

Subjectivity Clues lexicon from OpinionFinder (Wilson et al 2005)

$$\textit{sentiment}_t(\text{topic\_word}) = \frac{p(\text{pos. word} \mid \text{topic word}, t)}{p(\text{neg. word} \mid \text{topic word}, t)}$$

A note on the sentiment list

- Not well-suited for social media English
  - “sucks” “:)” “:(“

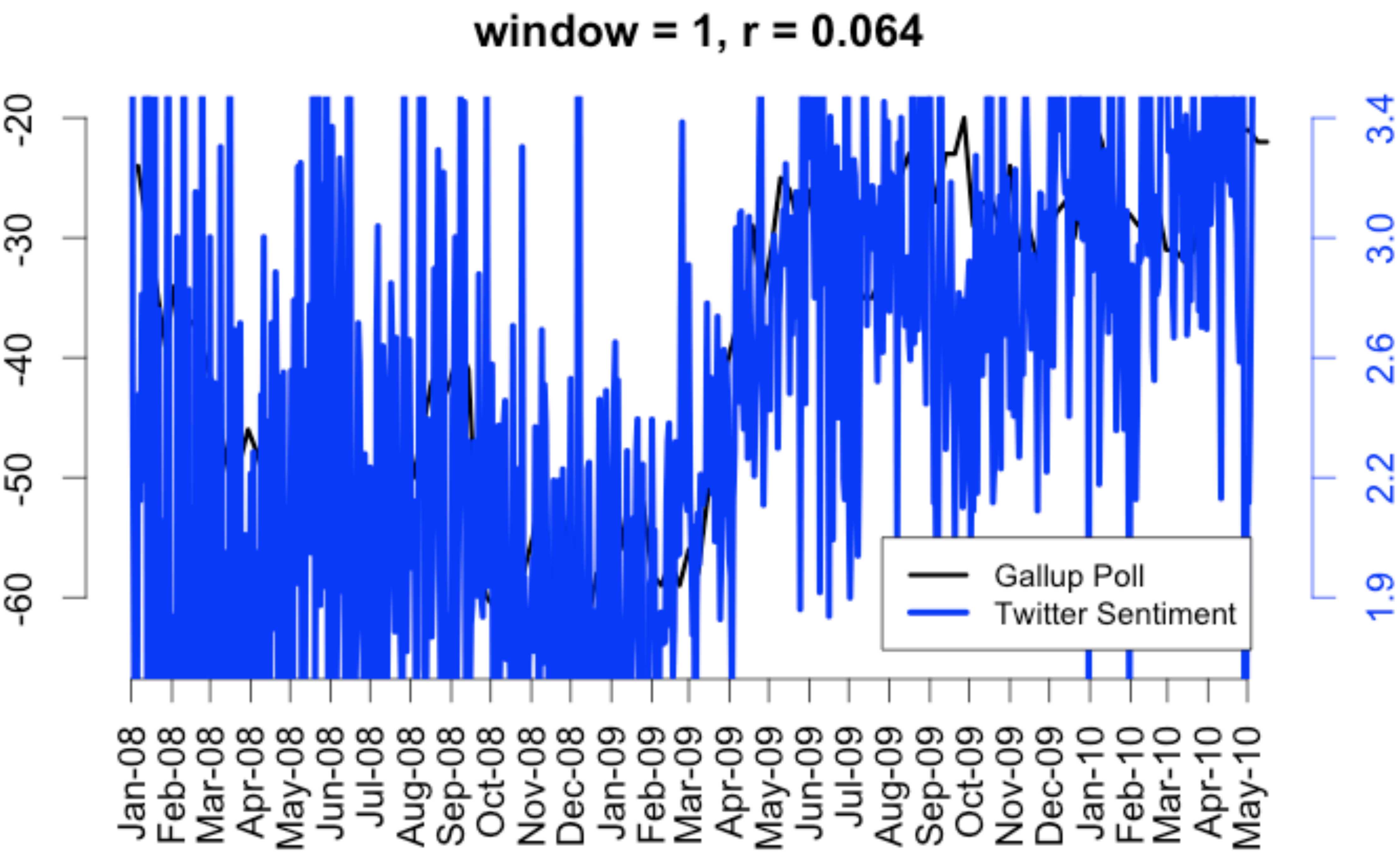
(Top examples)

word	valence	count
will	positive	3934
bad	negative	3402
good	positive	2655
help	positive	1971

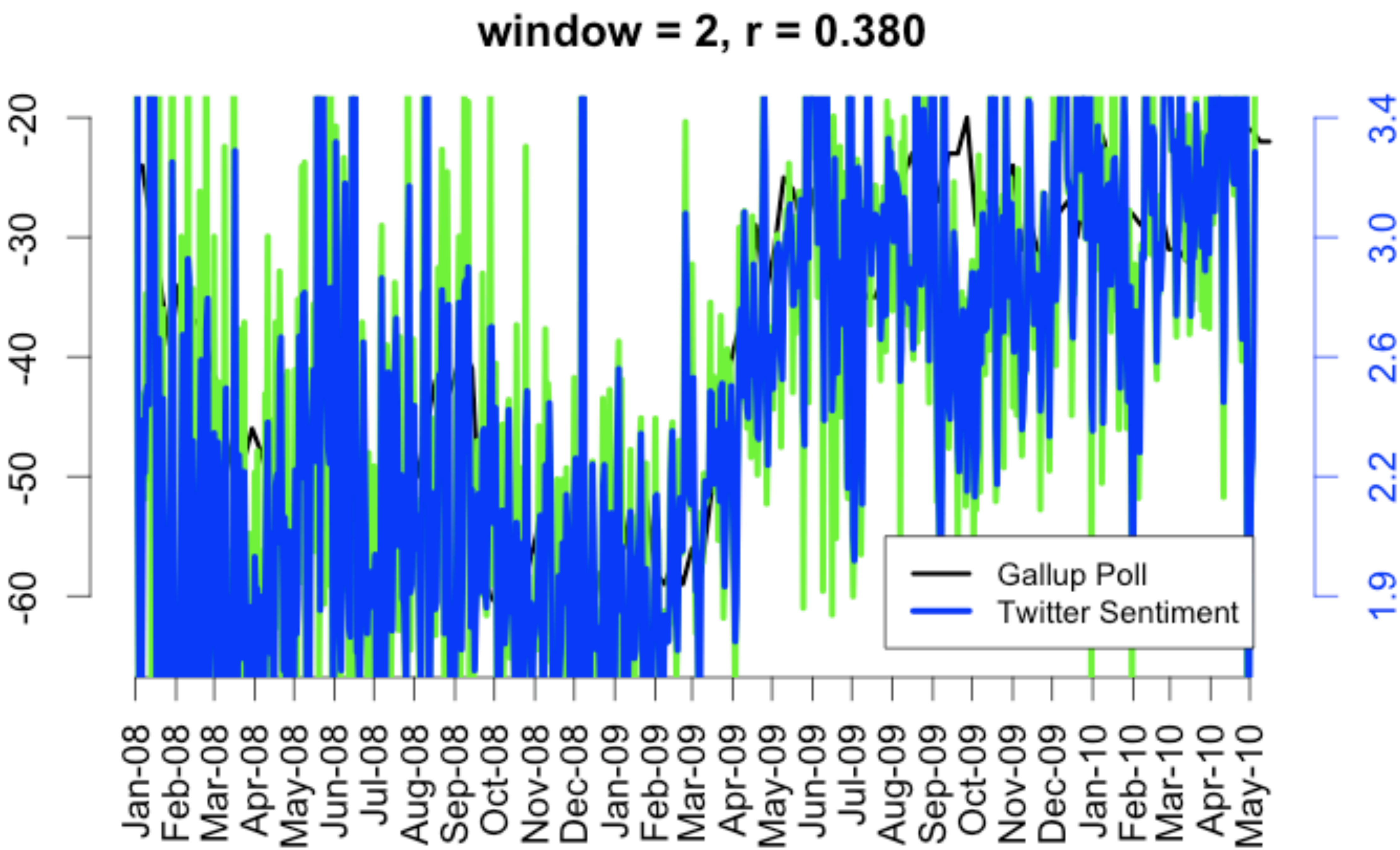
(Random examples)

word	valence	count
funny	positive	114
fantastic	positive	37
cornerstone	positive	2
slump	negative	85
bearish	negative	17
crackdown	negative	5

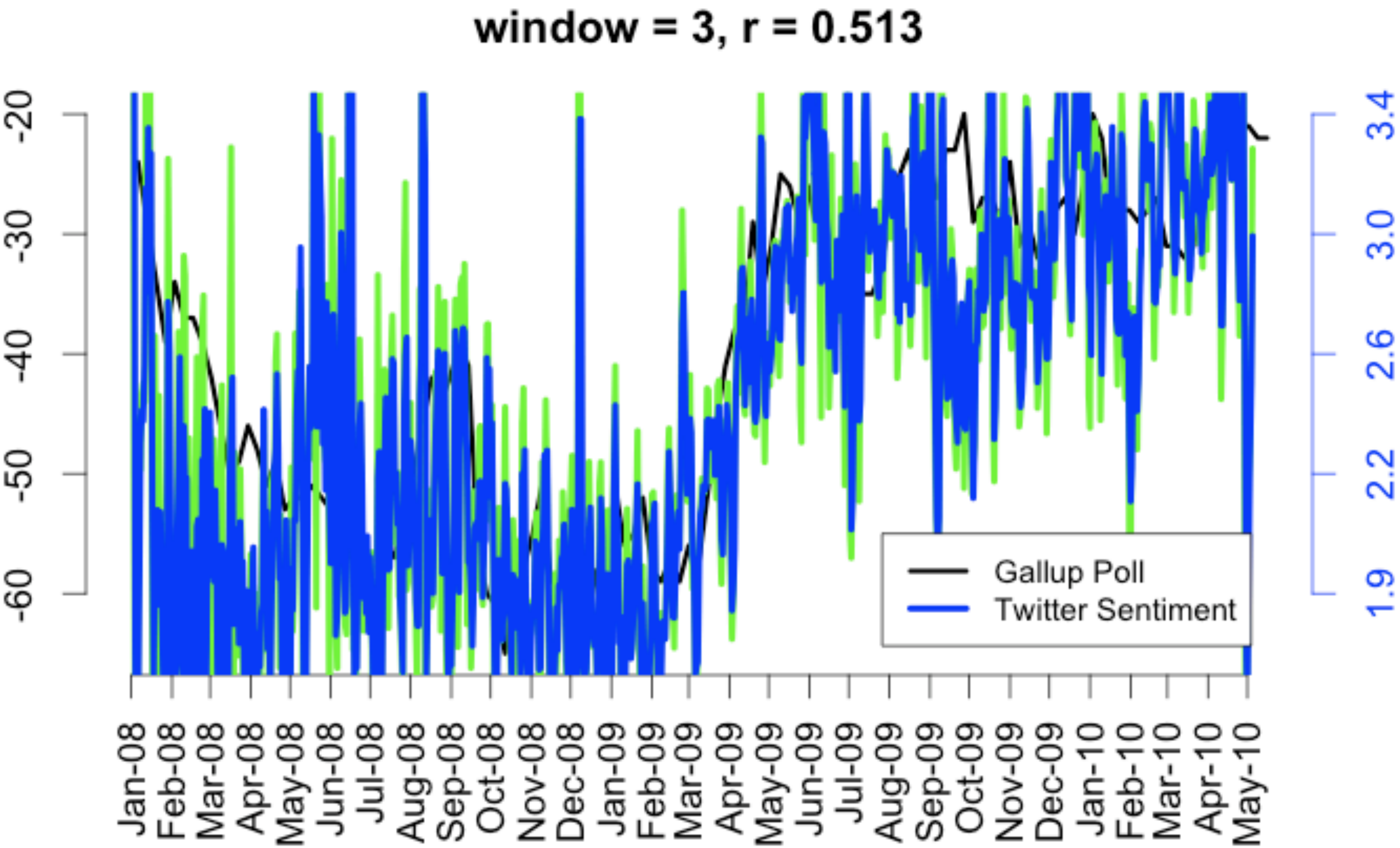
Smoothed comparisons:
“jobs” sentiment vs. consumer confidence



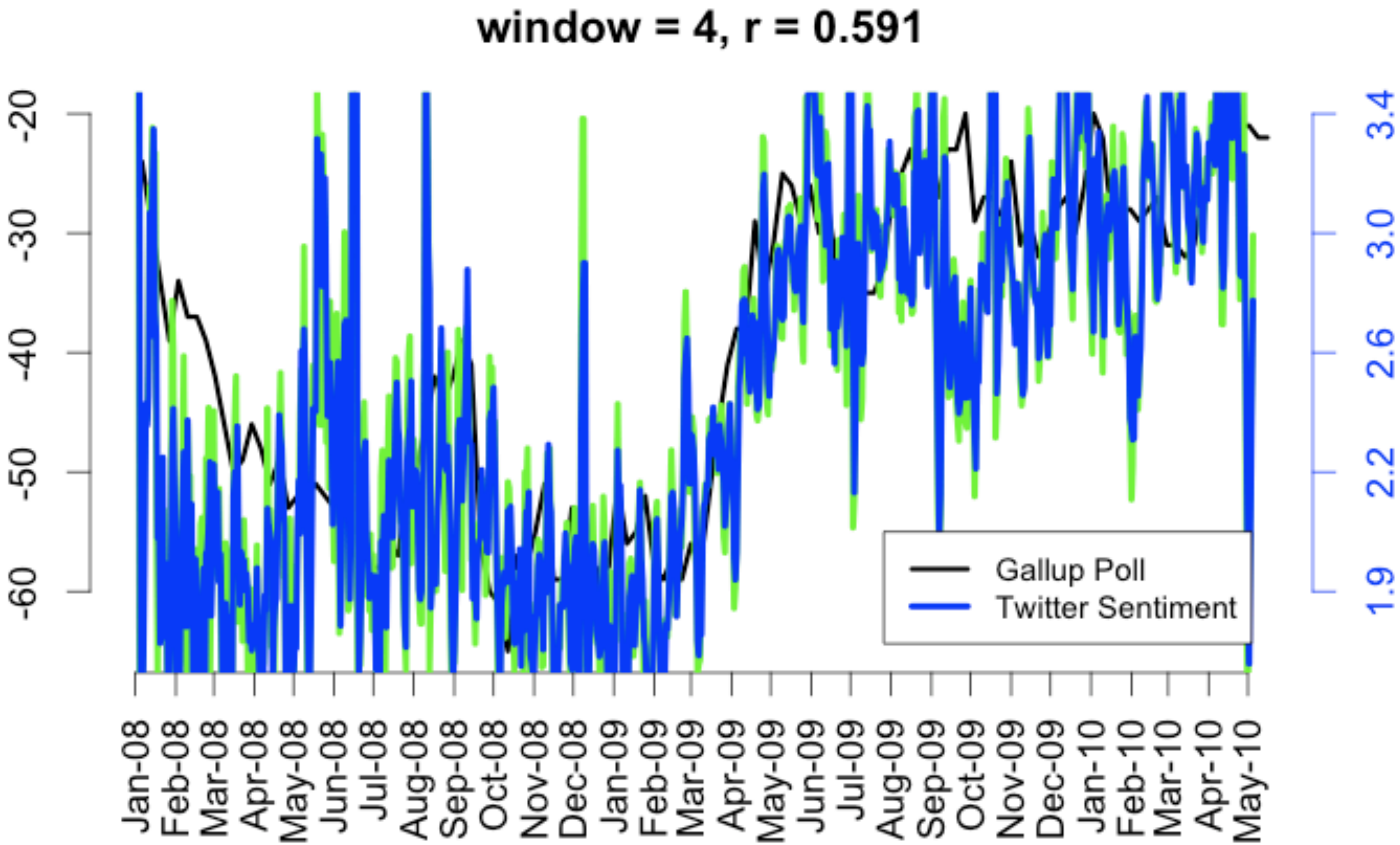
Smoothed comparisons:
"jobs" sentiment vs. consumer confidence



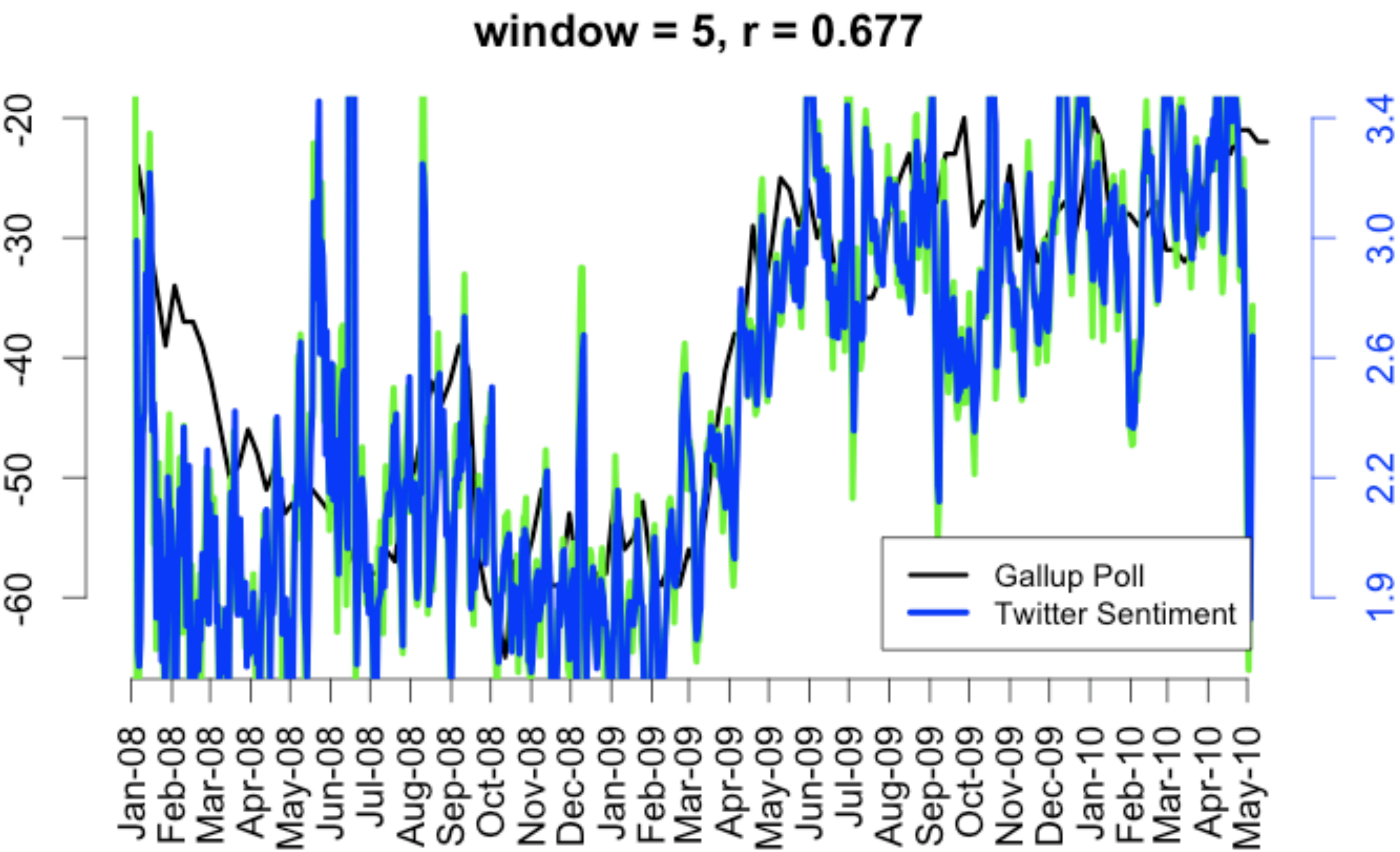
Smoothed comparisons:
"jobs" sentiment vs. consumer confidence



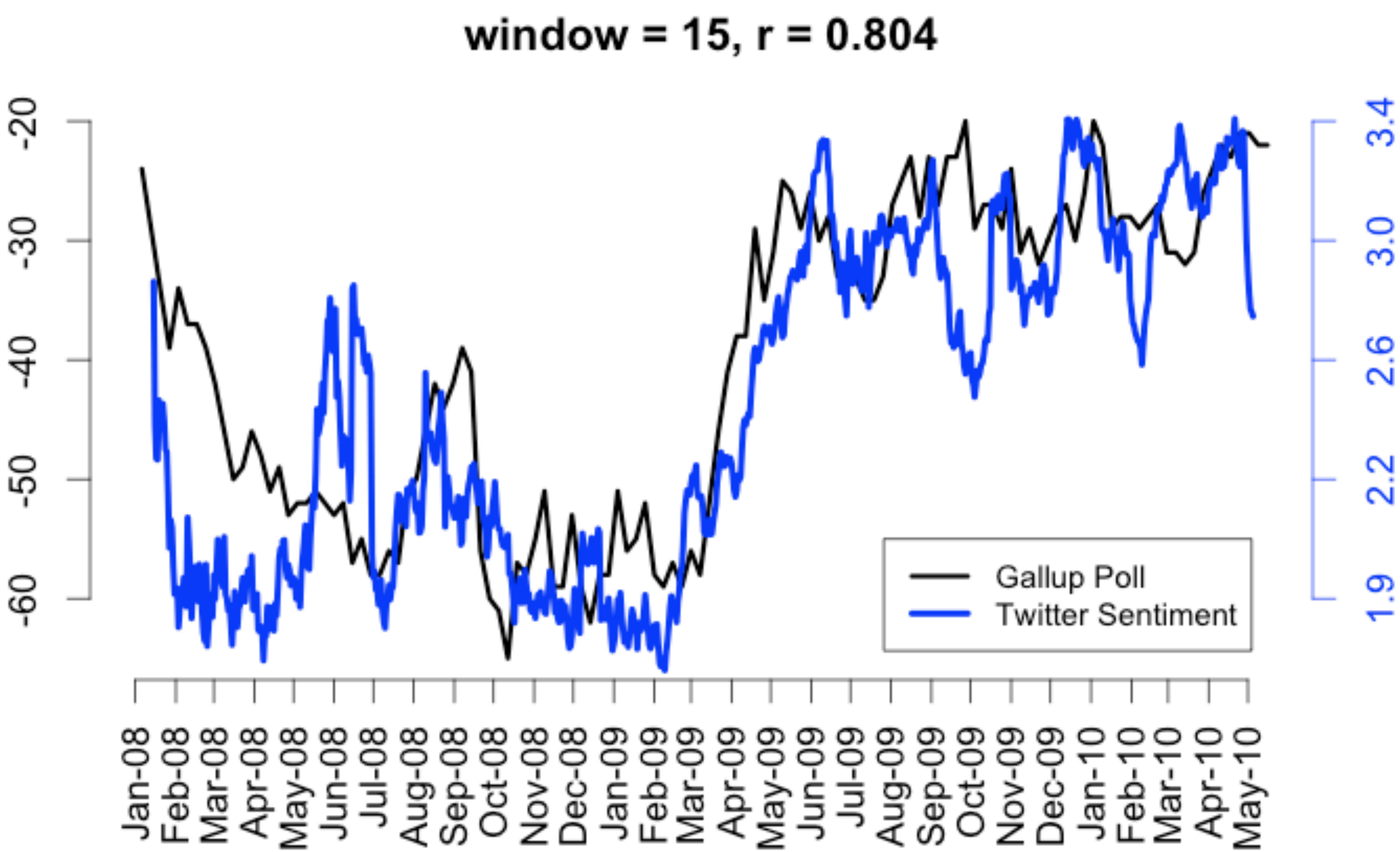
Smoothed comparisons:
"jobs" sentiment vs. consumer confidence



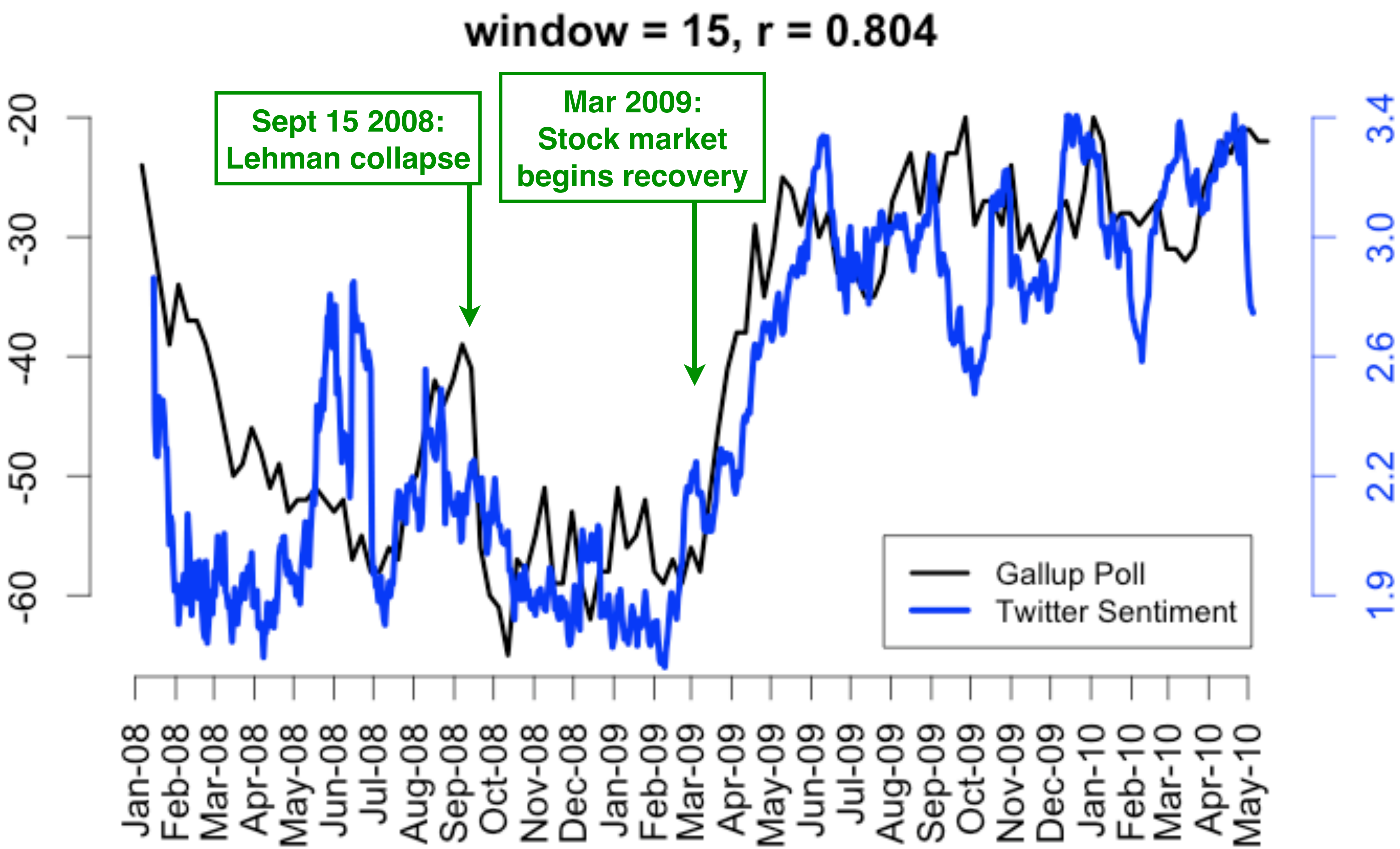
Smoothed comparisons:
"jobs" sentiment vs. consumer confidence



Smoothed comparisons:
"jobs" sentiment vs. consumer confidence

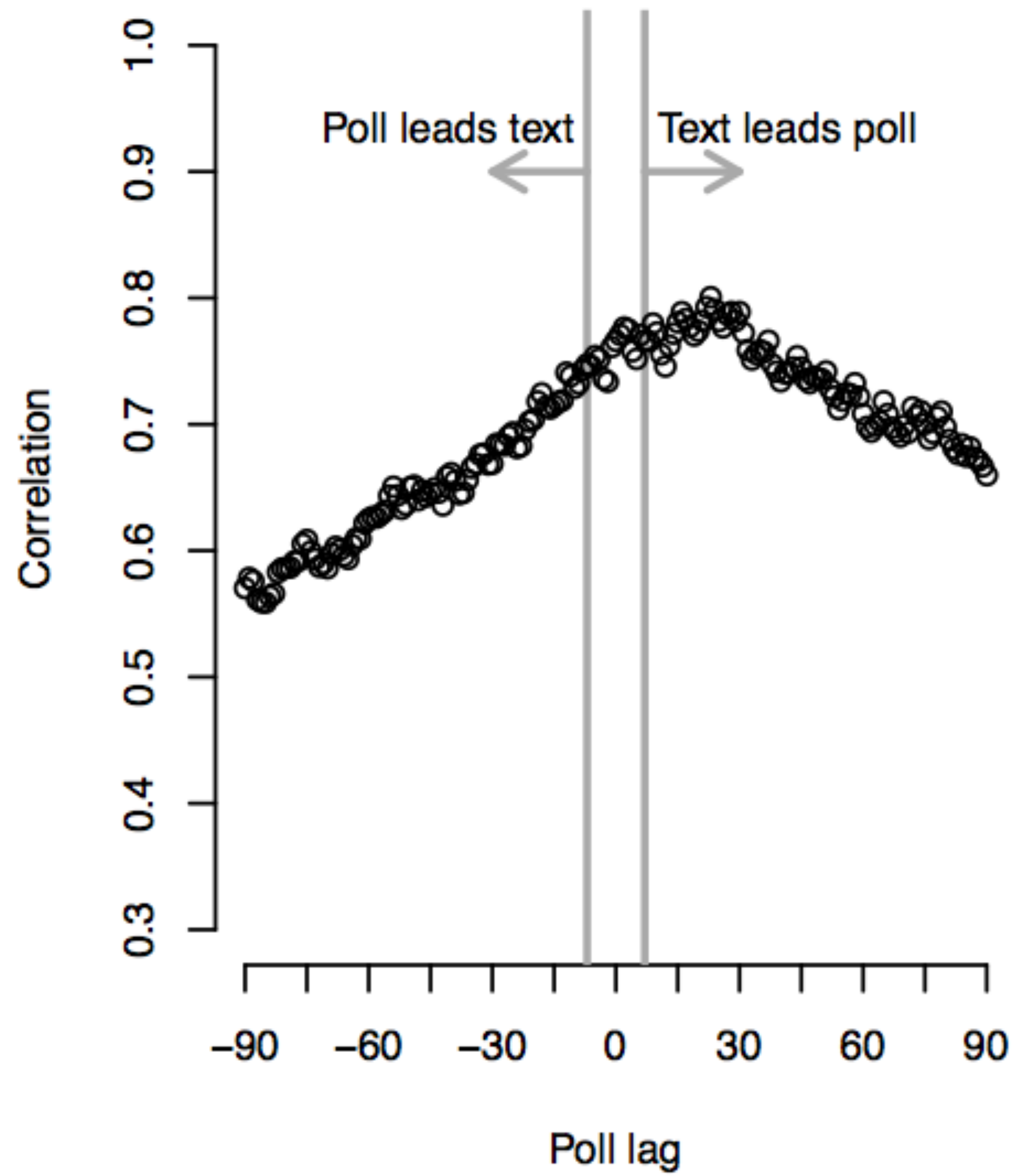


Smoothed comparisons:
“jobs” sentiment vs. consumer confidence



Which leads, poll or text?

- Cross-corr between
  - Sentiment score on day *t*
  - Poll day *t+L*
- *sentiment*(“jobs”) is leading indicator for poll



Keyword message selection

15-day windows, no lag

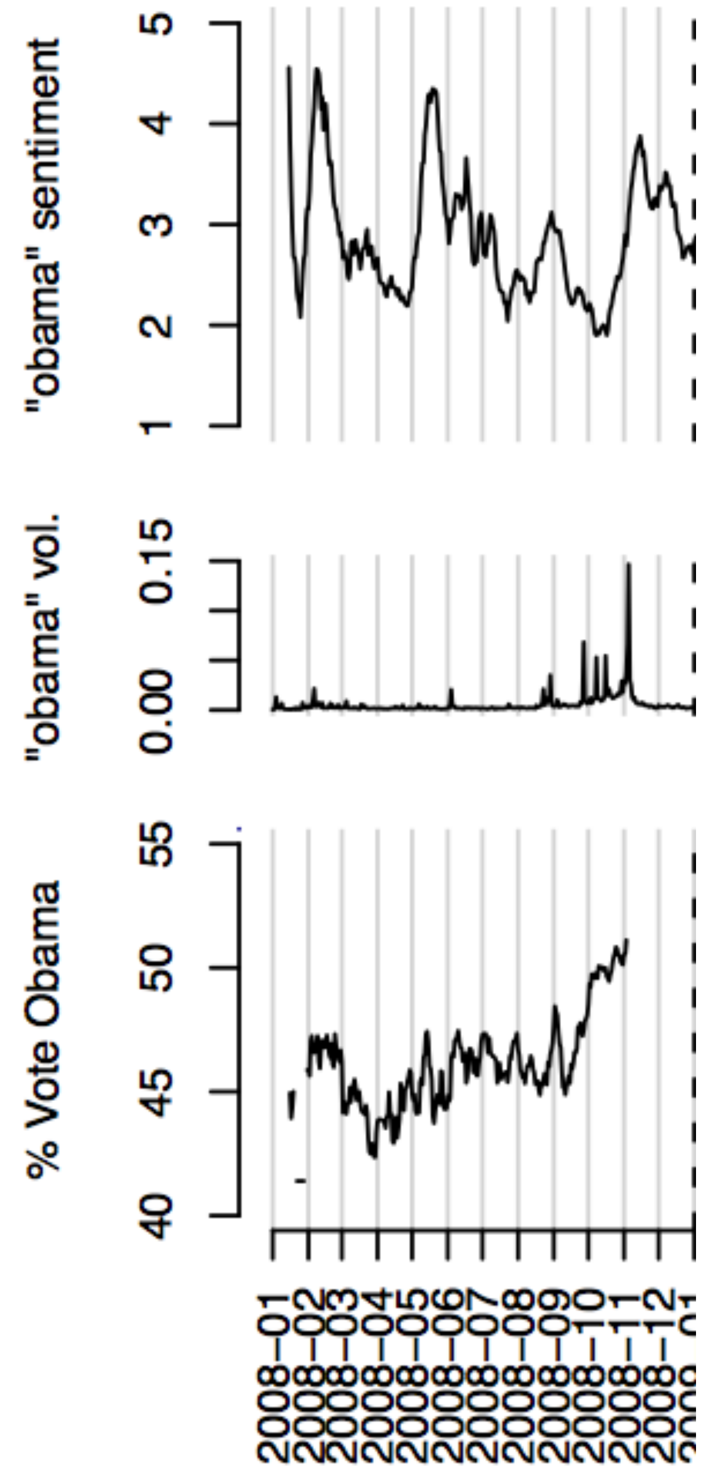
sentiment(“jobs”)	r = +0.80
sentiment(“job”)	r = +0.07
sentiment(“economy”)	r = –0.10

Look out for stemming

sentiment(“jobs” OR “job”)	r = +0.40

Presidential elections [doesn't work]

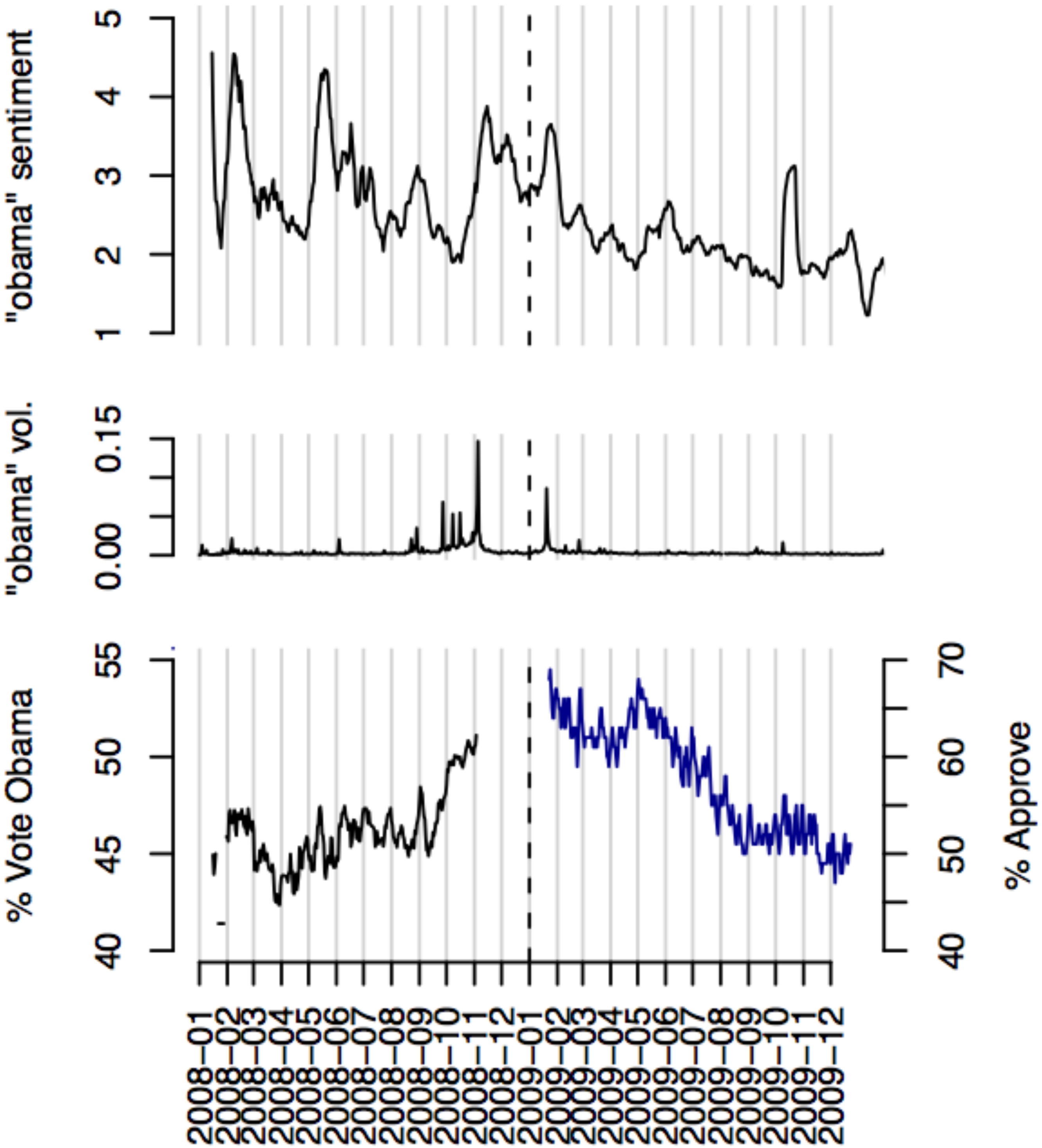
- 2008 elections *sen*("obama"), *sen*("mccain") do <u>not</u> correlate to polls



Presidential elections [doesn't work]
Presidential job approval [~works]

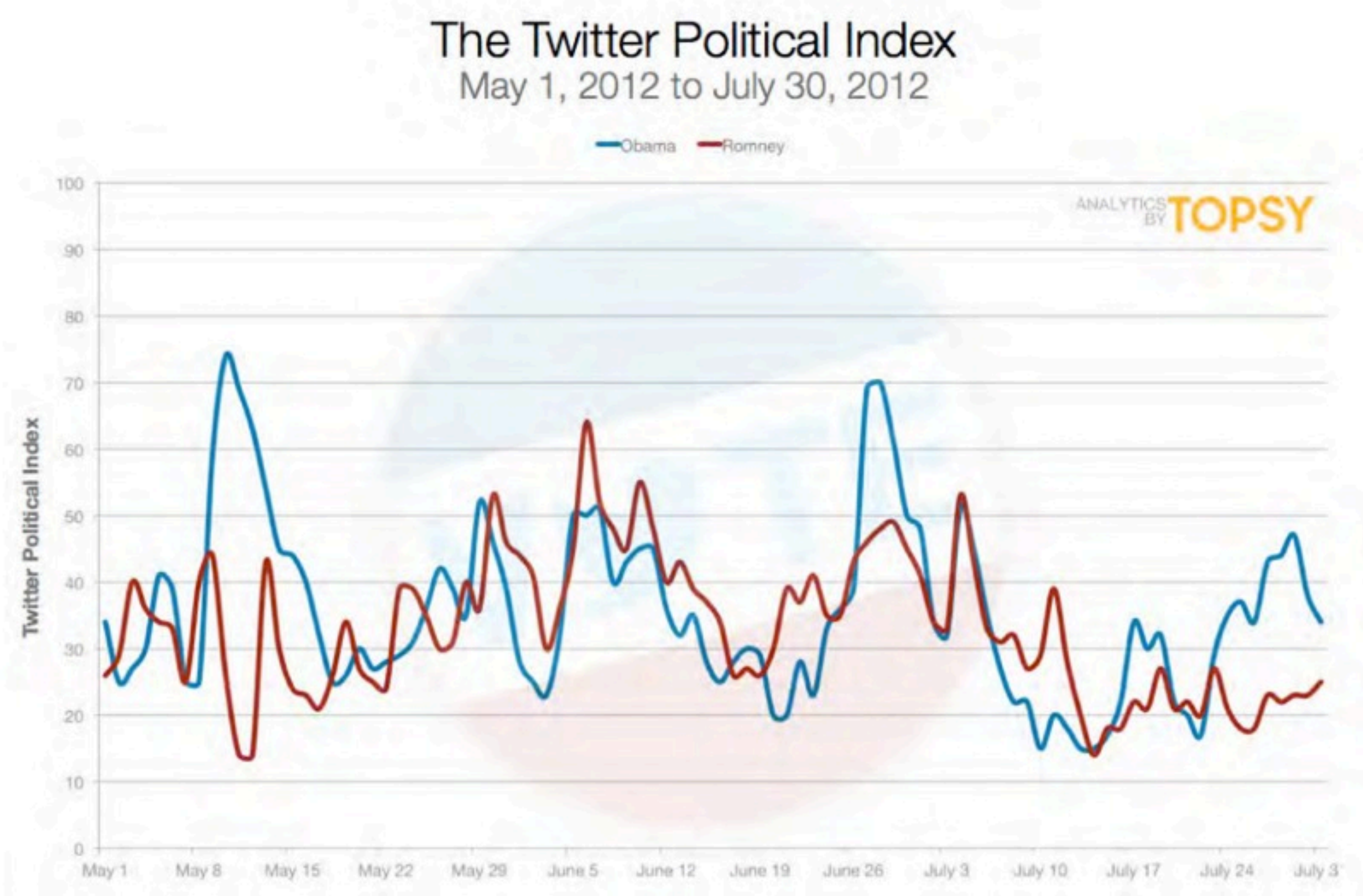
- 2008 elections
  sen("obama"),
  sen("mccain") do <u>not</u>
  correlate to polls

- 2009 job approval
  sen("obama") => r = 0.72
  Looks easy: simple decline



election.twitter.com

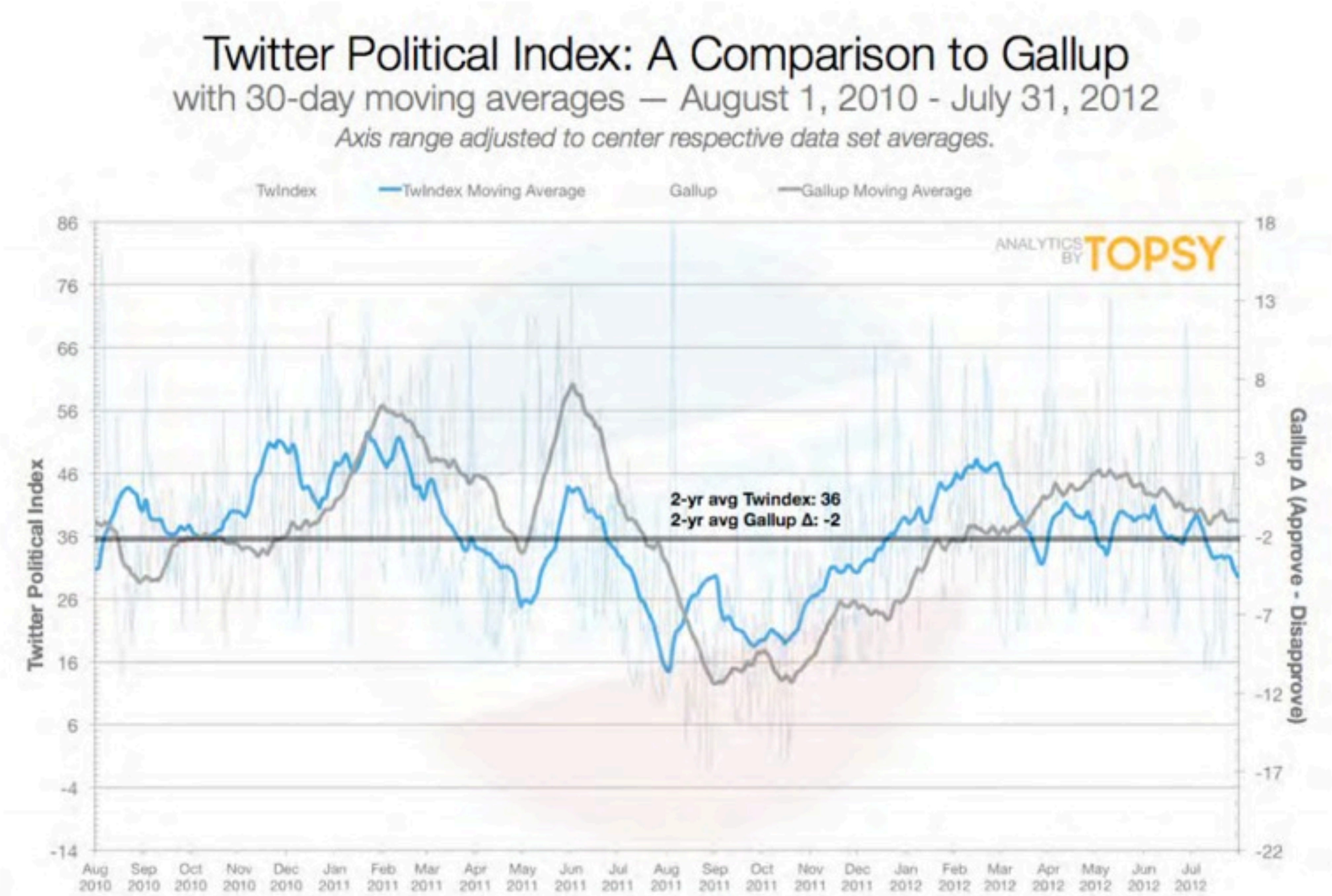
Proprietary sentiment analyzer over Obama vs Romney name-containing tweets
http://about.topsy.com/wp-content/uploads/2012/08/Twindex-report1.pdf



election.twitter.com

Proprietary sentiment analyzer over Obama vs Romney name-containing tweets

http://about.topsy.com/wp-content/uploads/2012/08/Twindex-report1.pdf



Twitter and Polls

- Preliminary results that sentiment analysis on Twitter data can give information similar to opinion polls
  - But, still not well-understood!
- Who is using Twitter?
  - Massive changes over time (2008 Twitter != 2012 Twitter)
  - News vs. opinion?
  - Other data sources might better distinguish?
- Better text analysis
  - Very wide linguistic variation on Twitter
  - Word sense ambiguity: "steve jobs"
  - Better data sources
- Suggestion for future work: analyze correlations to pre-existing surveys and other attitude measurements
- Not a replacement for polls, but seems potentially useful.
  Between ethnography and surveys?

- See also
  "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" -- A Balanced Survey on Election Prediction using Twitter Data
  Daniel Gayo-Avello, arXiv 2012

Taxonomy of text analysis methods

Computational / statistical complexity

Correlations, ratios, counts

Word counting

Dictionary word counting

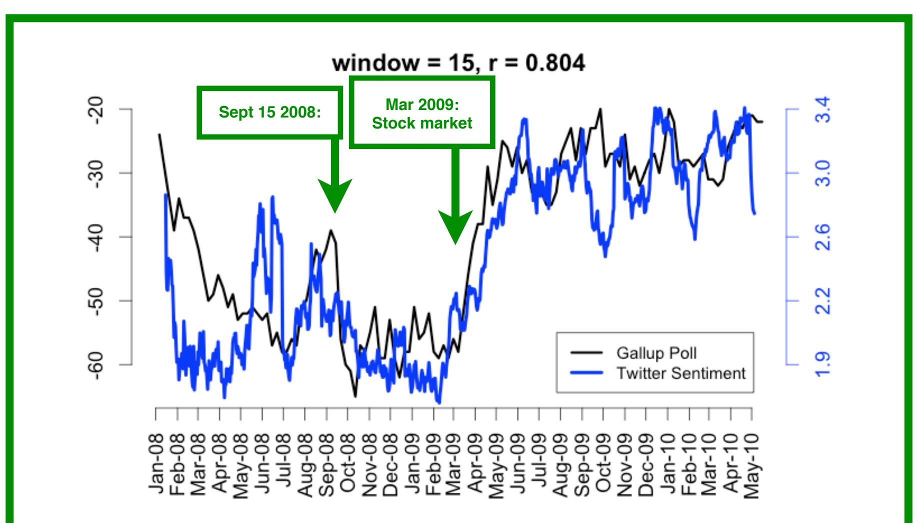
Bare words

Naturally-labeled documents

Hand-labeled documents

Hand-built dictionaries

Data/domain assumptions

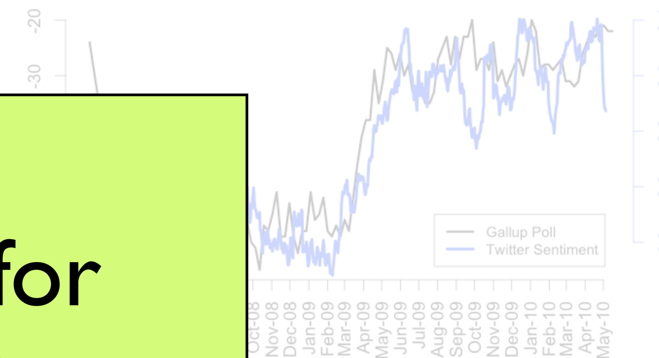


Detecting cultural phenomena in textual social media

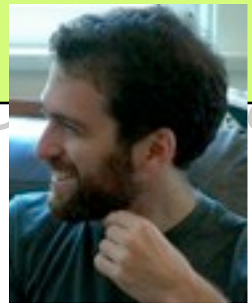
Opinions and Time

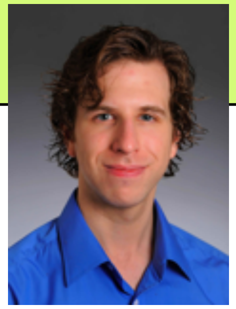
Brendan O'Conn

Bryan Routledge



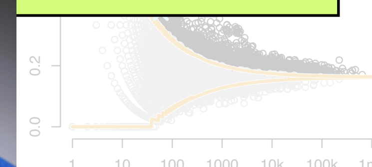
A Latent Variable Model for Geographic Lexical Variation. *EMNLP* 2010.





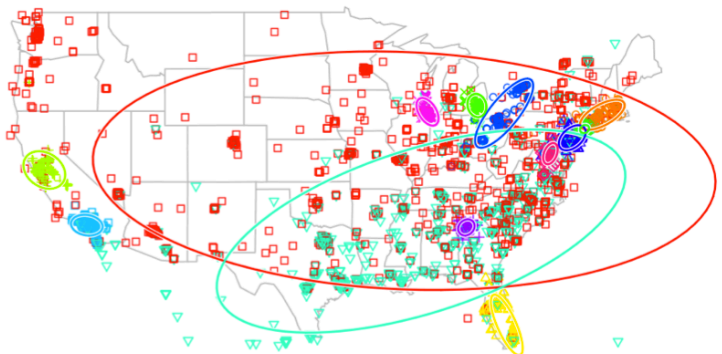
Internet C

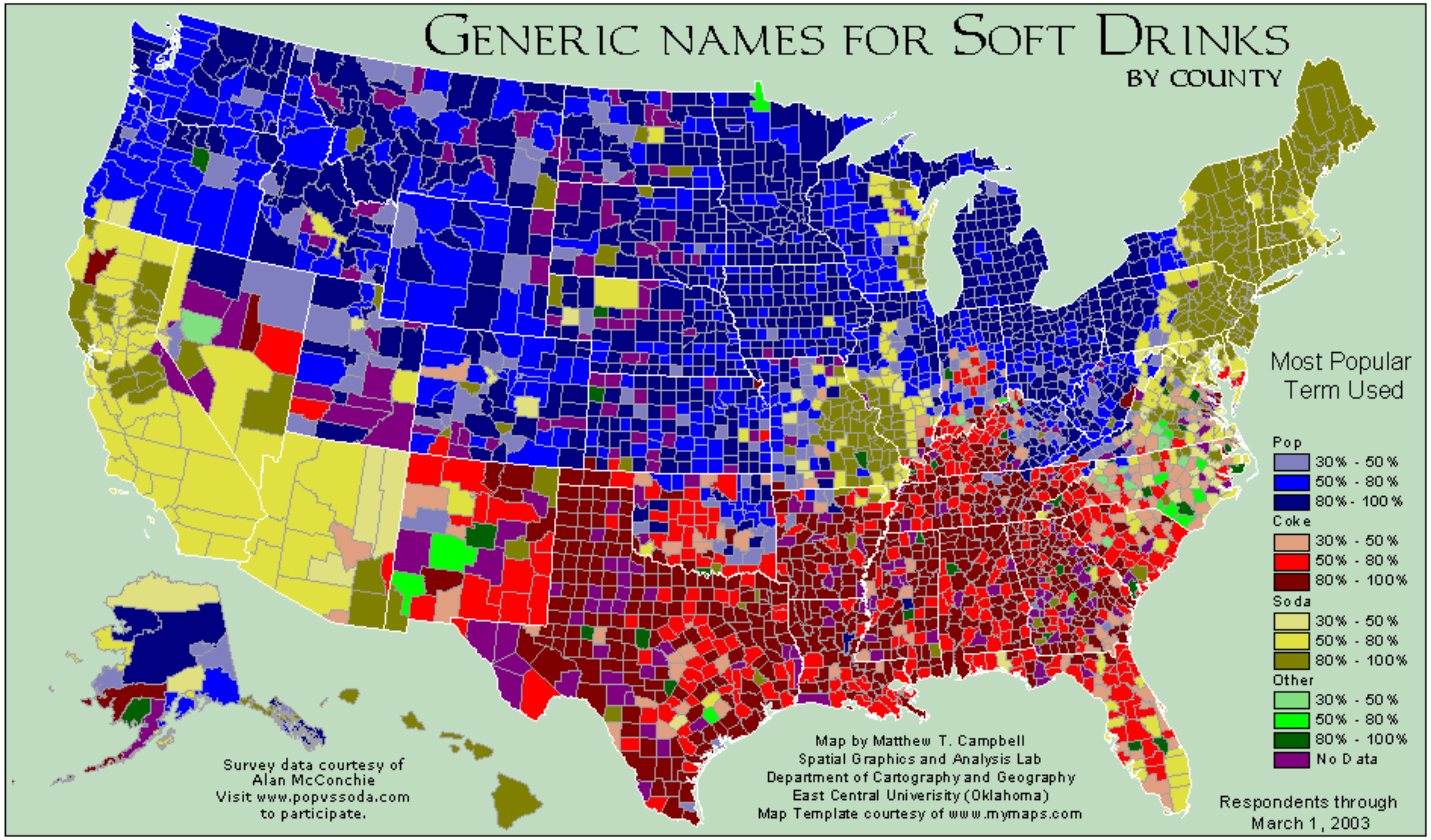
David Bamman, Brendan O'C



Language and Geography

Jacob Eisenstein, Brendan O'Connor,
Noah Smith, Eric Xing





- Languages exhibit variation, reflecting geography, status, race, gender, etc.

Searching for dialect in social media

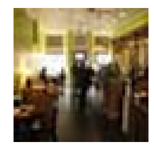
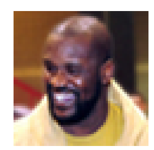
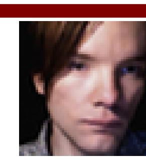
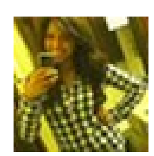
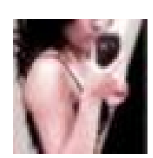
- One approach: search for known variable alternations, e.g. you / yinz / yall

  (Kurath 1949, …, Boberg 2005)
- Known variables like “yinz” don't appear much
- Are there new variables we don't know about?

Data

- Mobile clients for Twitter allow encoding of GPS location
- Our corpus: 380K messages from 9500 authors in the USA (March 2010)
  - Informal and conversational
  - 25% of the most common words not in the dictionary
- More than half of messages mention another user

A partial taxonomy of Twitter messages

Official announcements	**BritishMonarchy** TheBritishMonarchy On 6 Jan: Changing the Guard at Buckingham Palace - Starts at approx 11am http://www.royal.gov.uk/G 17 hours ago
Business advertising	 **bigdogcoffee** bigdogcoffee Back to normal hours beginning tomorrow........Monday-Friday 6am-10pm Sat/Sun 7:30am-10pm 2 Jan
Links to blog and web content	 **crampell** Catherine Rampell Casey B. Mulligan: Assessing the Housing Sector - http://nyti.ms/hcUKK9 10 hours ago
Celebrity self-promotion	 **THE_REAL_SHAQ** THE_REAL_SHAQ fill in da blank, my new years shaqalution is __________ 4 Jan
Status messages	 **emax** electronic max 1.1.11 - britons and americans can agree on the date for once. happy binary day! 1 Jan
Group conversation	 **_siddx3** Evelyn Santana RT @_LusciousVee: **#EveryoneShouldKnow** Ima Finally Be 18 This Year ^.^ 3 minutes ago
Personal conversation	 **xoxoJuicyCee** CeeCee'♥ @fxknnCelly aha kayy goodnightt (: 4 Jan

Generative Text Models

- How to simultaneously discover dialect regions and the words that characterize them?
- Probabilistic generative models
  - a.k.a. directed graphical models
  - Examples for text:
    - Hidden Markov Model
    - Naive Bayes
    - Topic Models, e.g. Latent Dirichlet Allocation

Generative models in 30 seconds

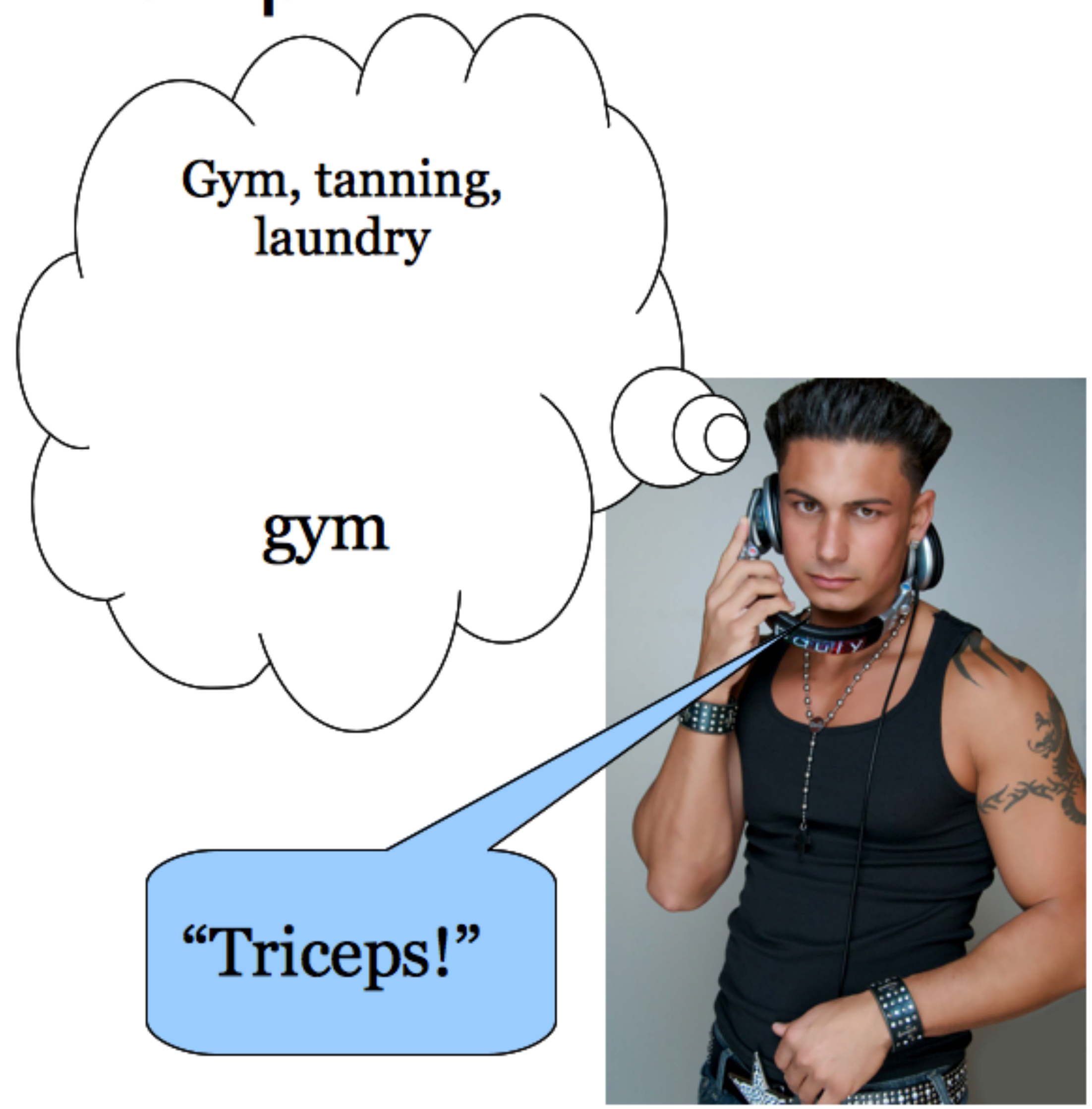
- We hypothesize that text is the output of a stochastic process. For example:

Pick some things to talk about

For each word,

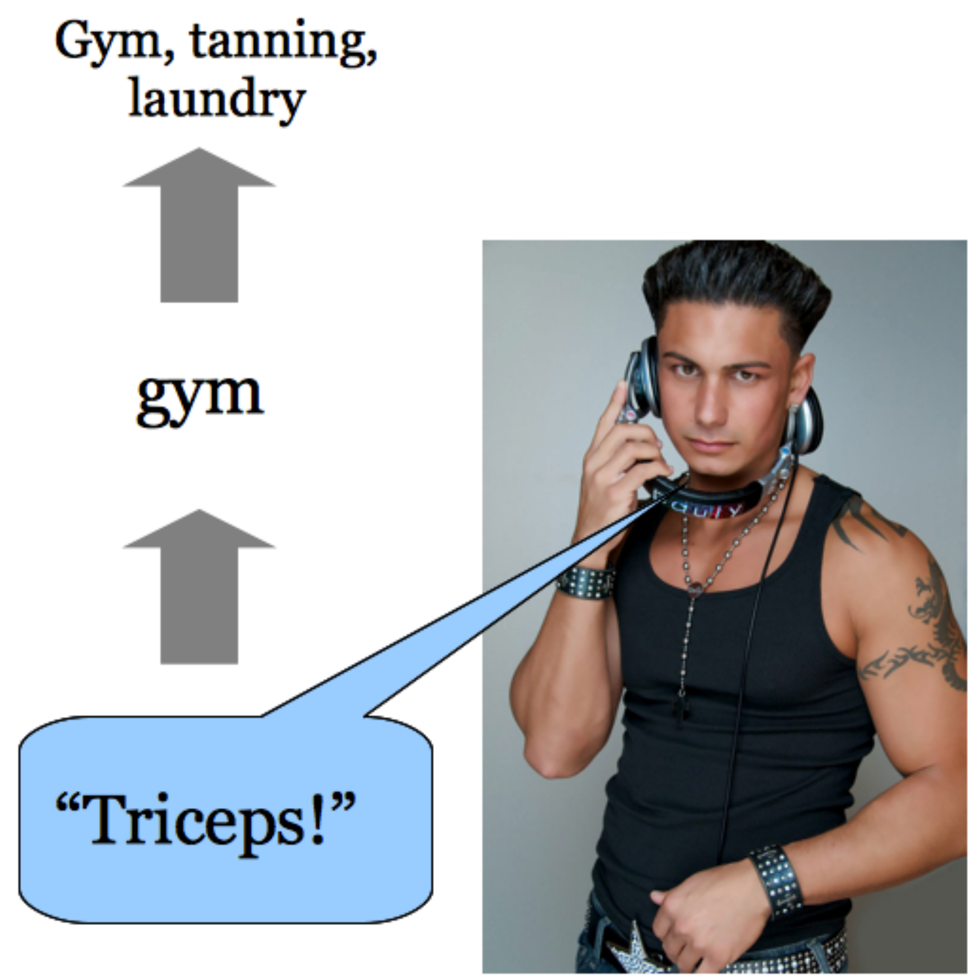
pick one thing to talk about

pick a word associated with that thing

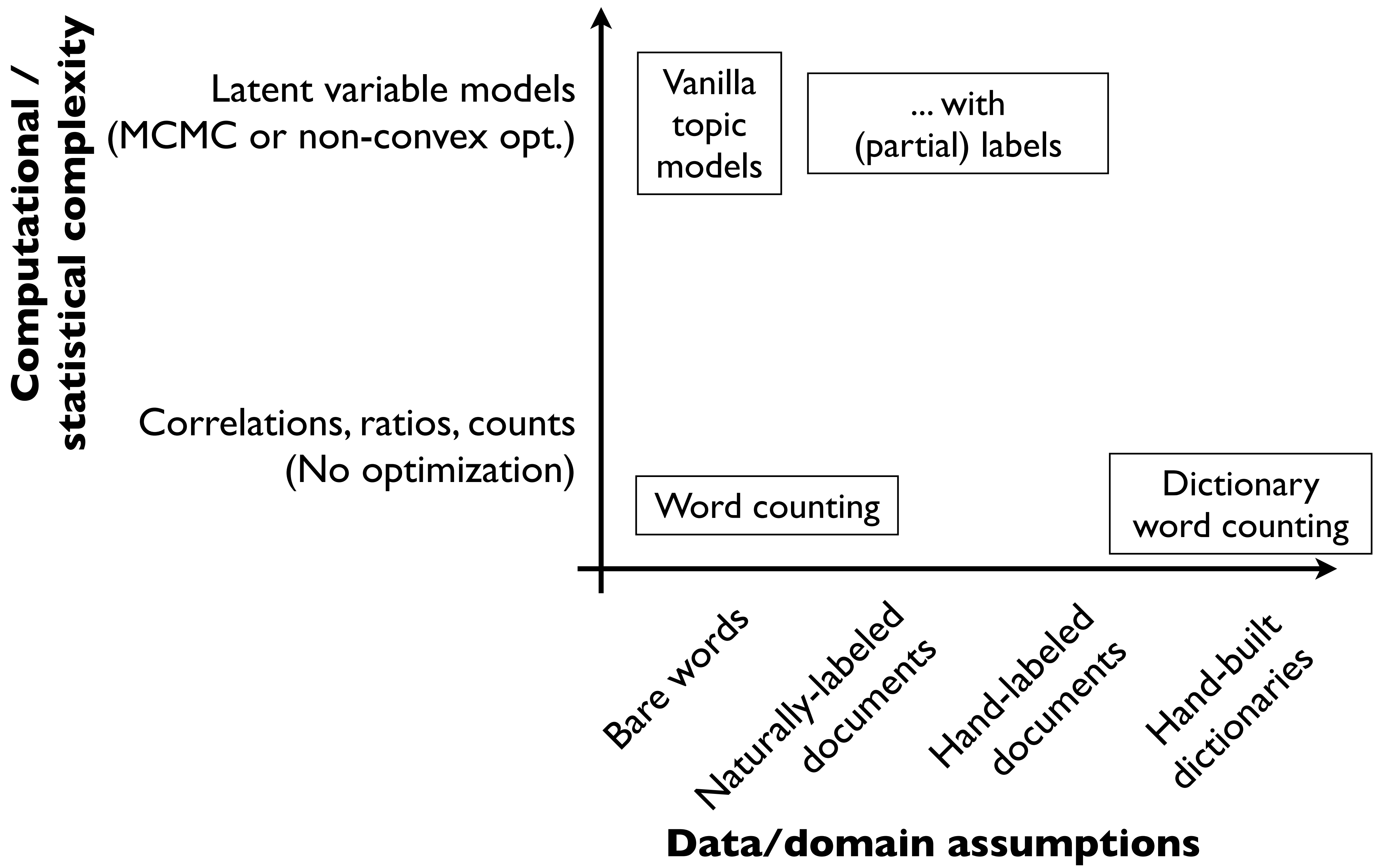


Generative models in 30 seconds

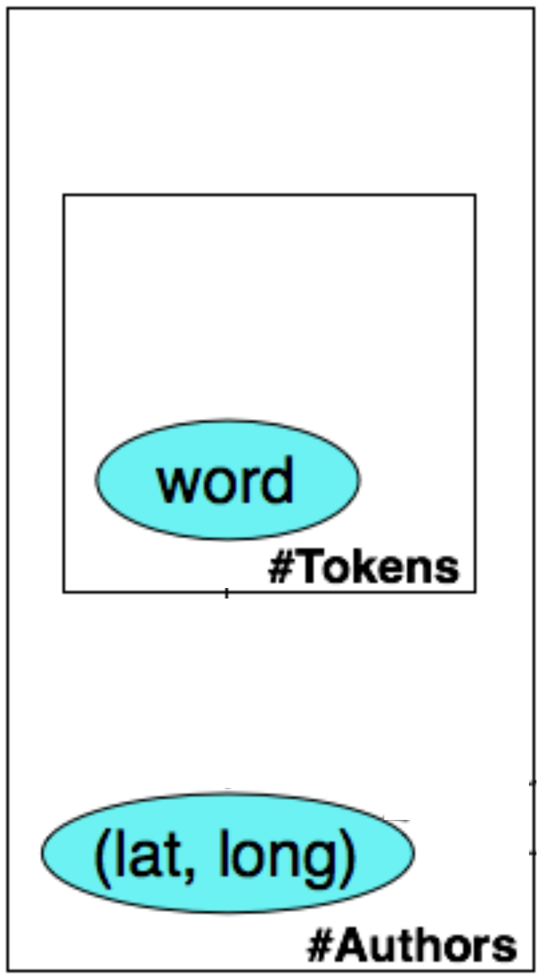
- We only see the output of the generative process.
- Through statistical inference over large amounts of data, we make educated guesses about the hidden variables.



Taxonomy of text analysis methods



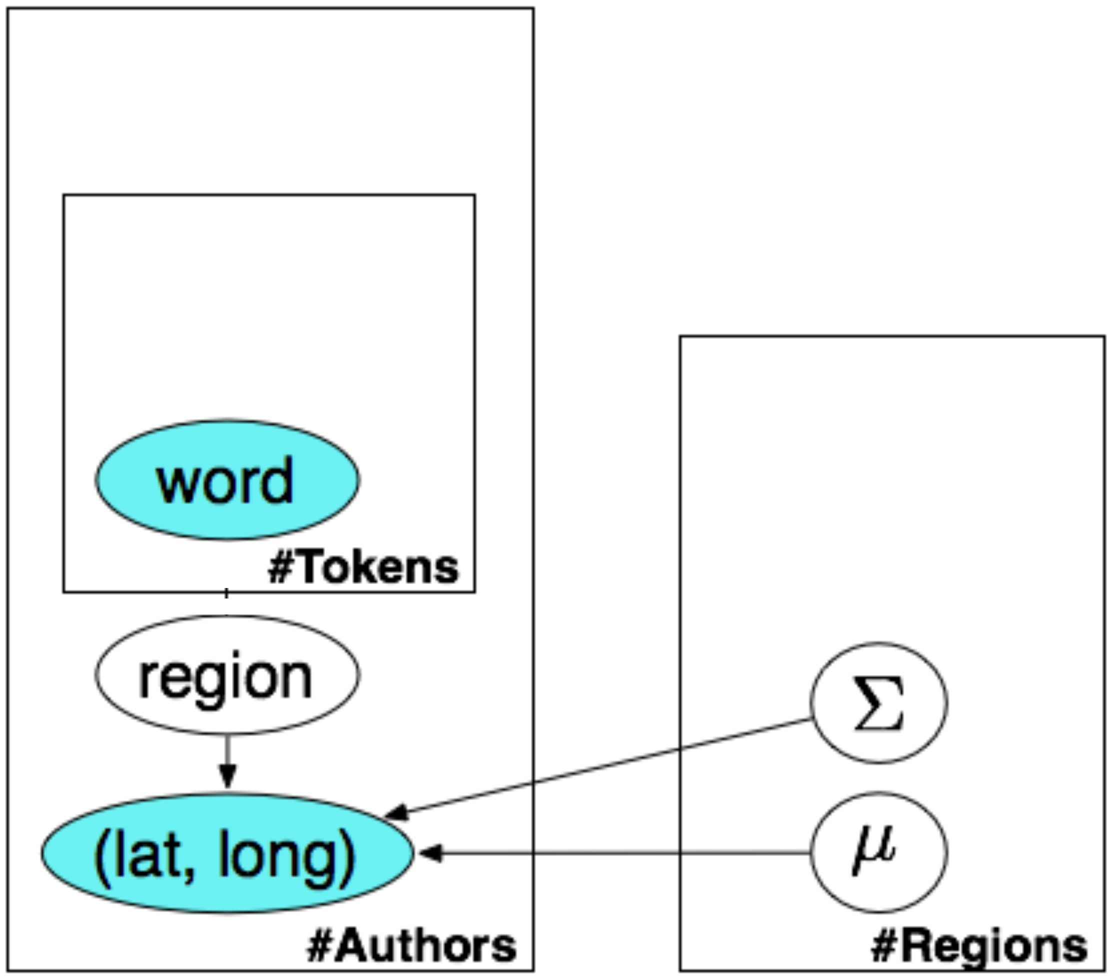
Model



All of an author's tweets collapsed into one "document"

Author is assigned one location

Model

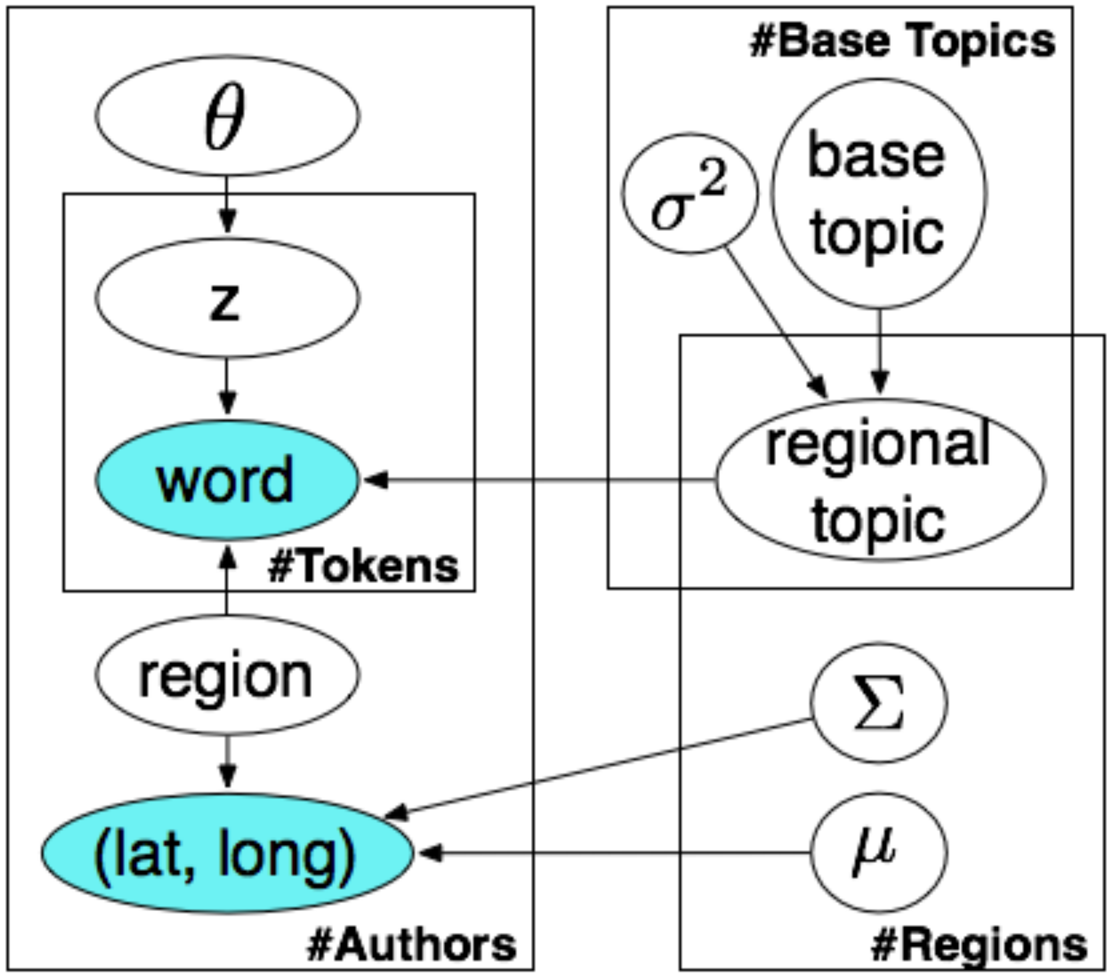


Locations are Gaussian mixture over space

$$r \sim \vec{\pi}$$

$$(lat,\ lon) \sim N(\vec{\mu}_r, \mathbf{\Sigma}_r)$$

Model



Author's words are admixture over regional topics

$$\theta \sim Dir(\vec{\alpha})$$

$$z \sim \vec{\theta}$$

$$w \sim \exp(\vec{\eta}_{zr})$$

Topics are logistic-normal over words

$$\vec{\phi}_k \sim N(\vec{a}, b^2\mathbf{I})$$

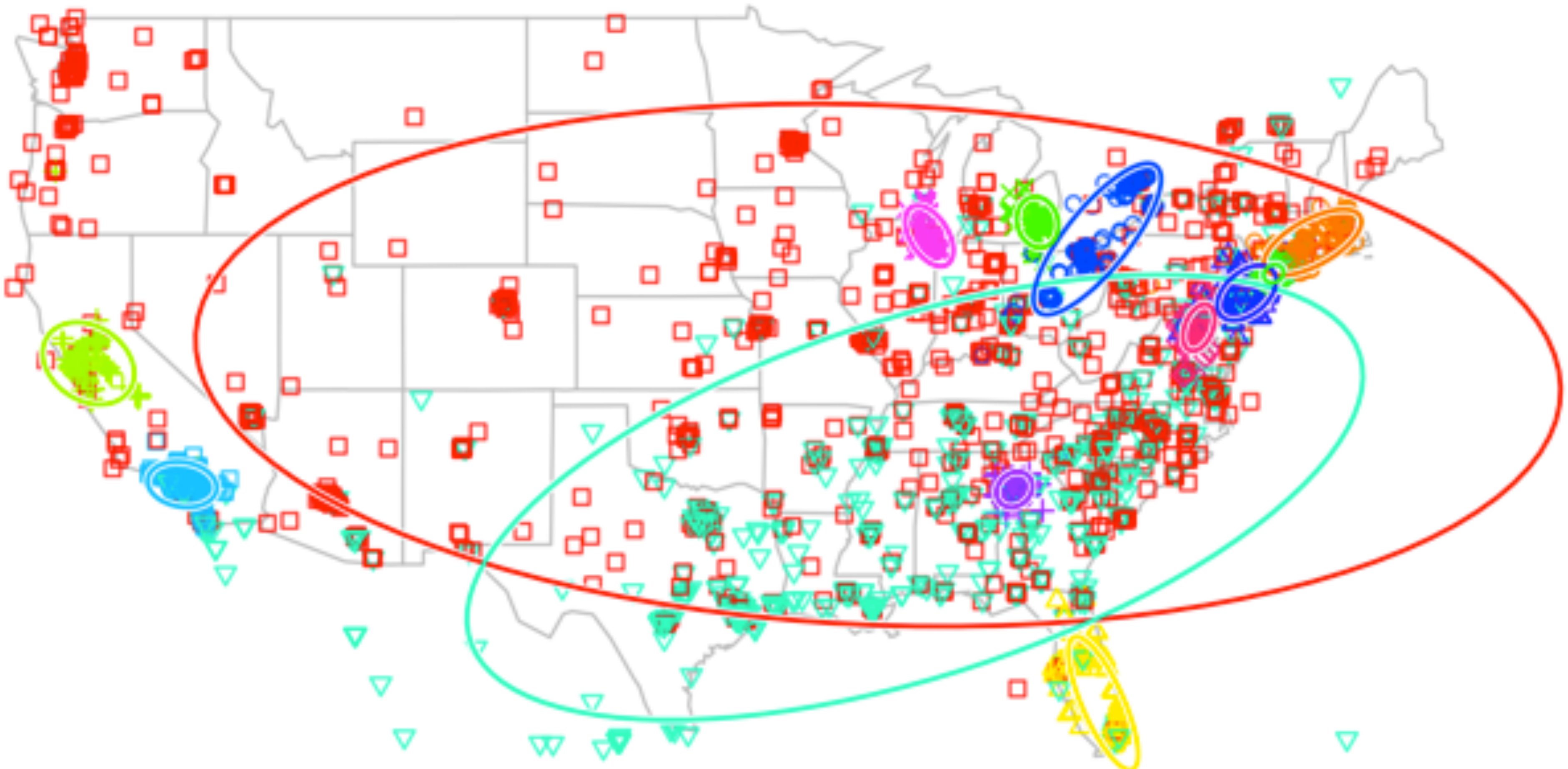
$$\vec{\eta}_{kj} \sim N(\vec{\phi}_k, s_k^2\mathbf{I})$$

Locations are Gaussian mixture over space

$$r \sim \vec{\pi}$$

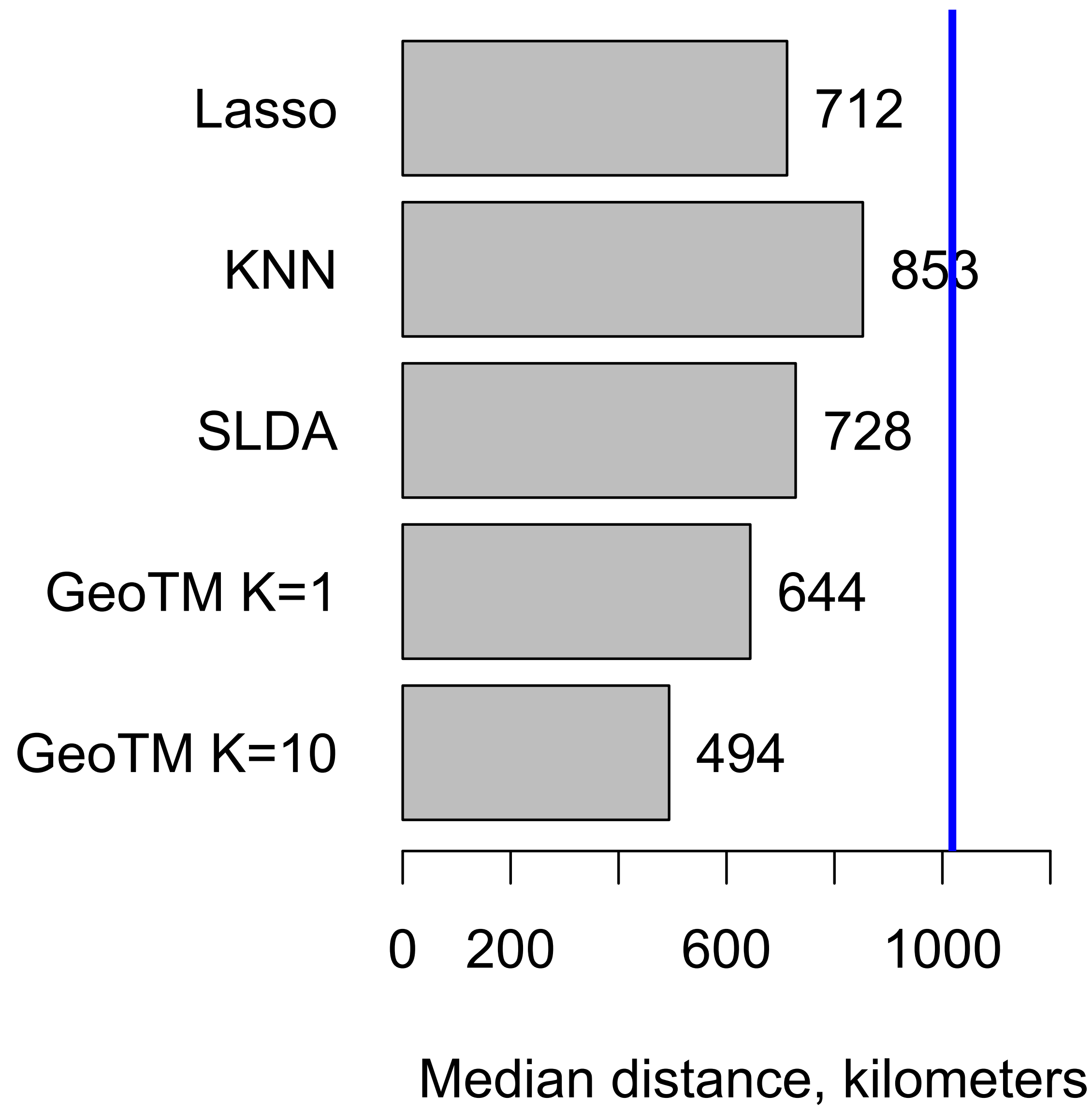
$$(lat,\ lon) \sim N(\vec{\mu}_r, \mathbf{\Sigma}_r)$$

Model fitting: variational mean field (Blei and Lafferty 06; Penny 01)



- Regions blend: text and geography jointly determine region membership

Validation: location prediction



- www.ark.cs.cmu.edu/GeoTwitter

Analysis

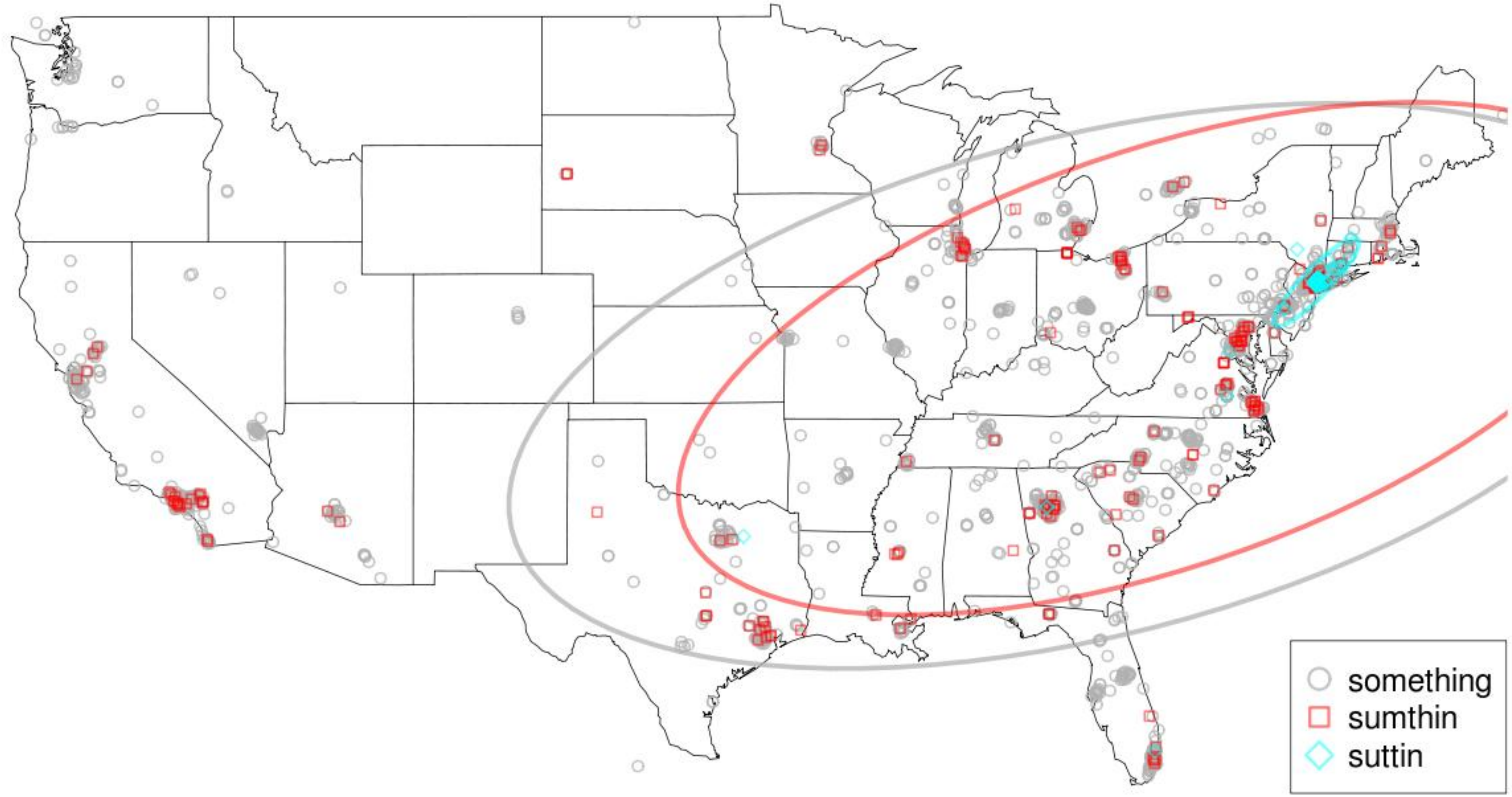
	“basketball” PISTONS KOBE LAKERS game DUKE NBA CAVS STUCKEY JETS KNICKS
Boston 	CELTICS victory BOSTON CHARLOTTE
N. California 	THUNDER KINGS GIANTS pimp trees clap
New York 	NETS KNICKS
Los Angeles 	#KOBE #LAKERS AUSTIN
Lake Erie 	CAVS CLEVELAND OHIO BUCKS od COLUMBUS

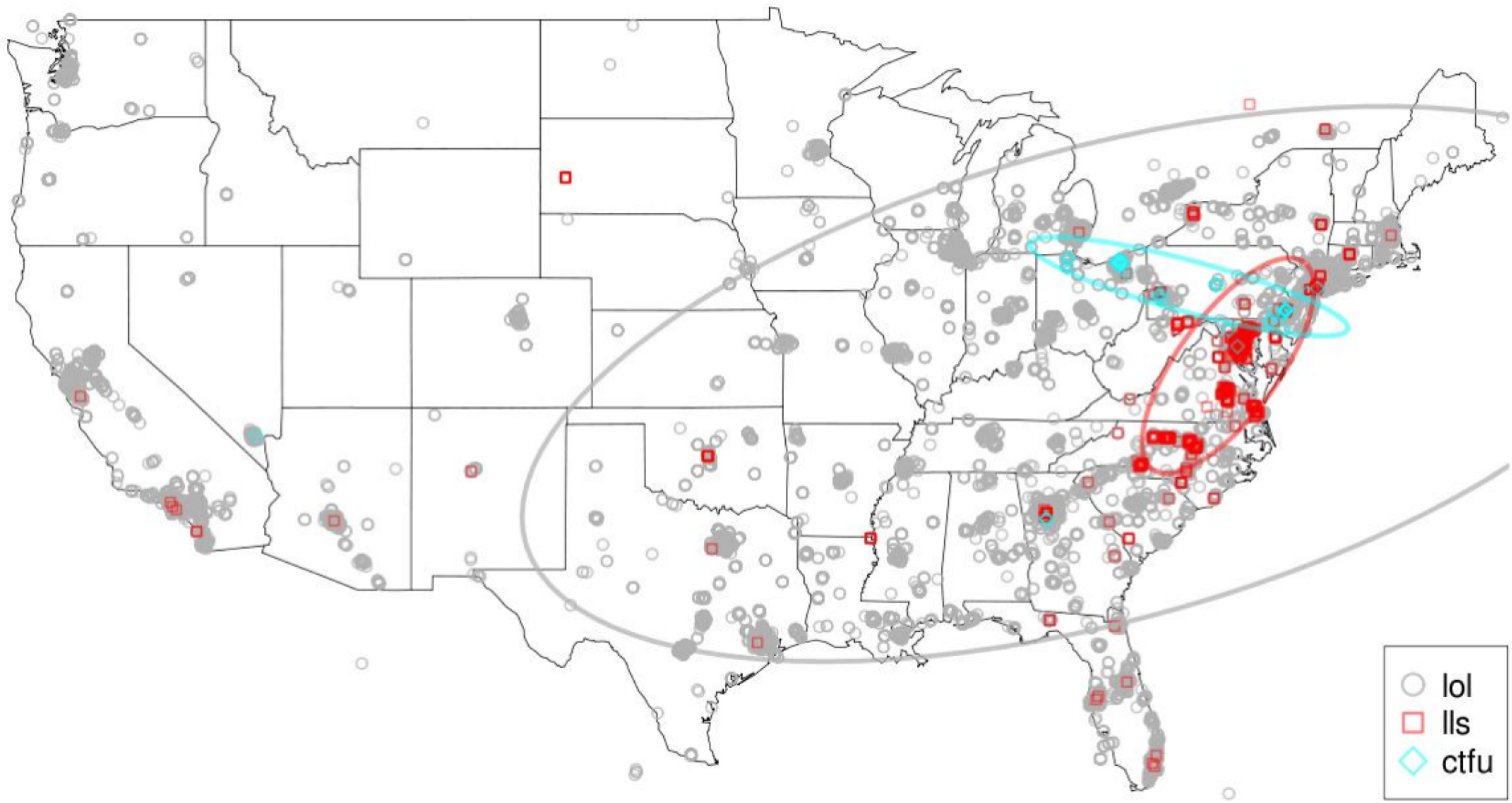
Analysis

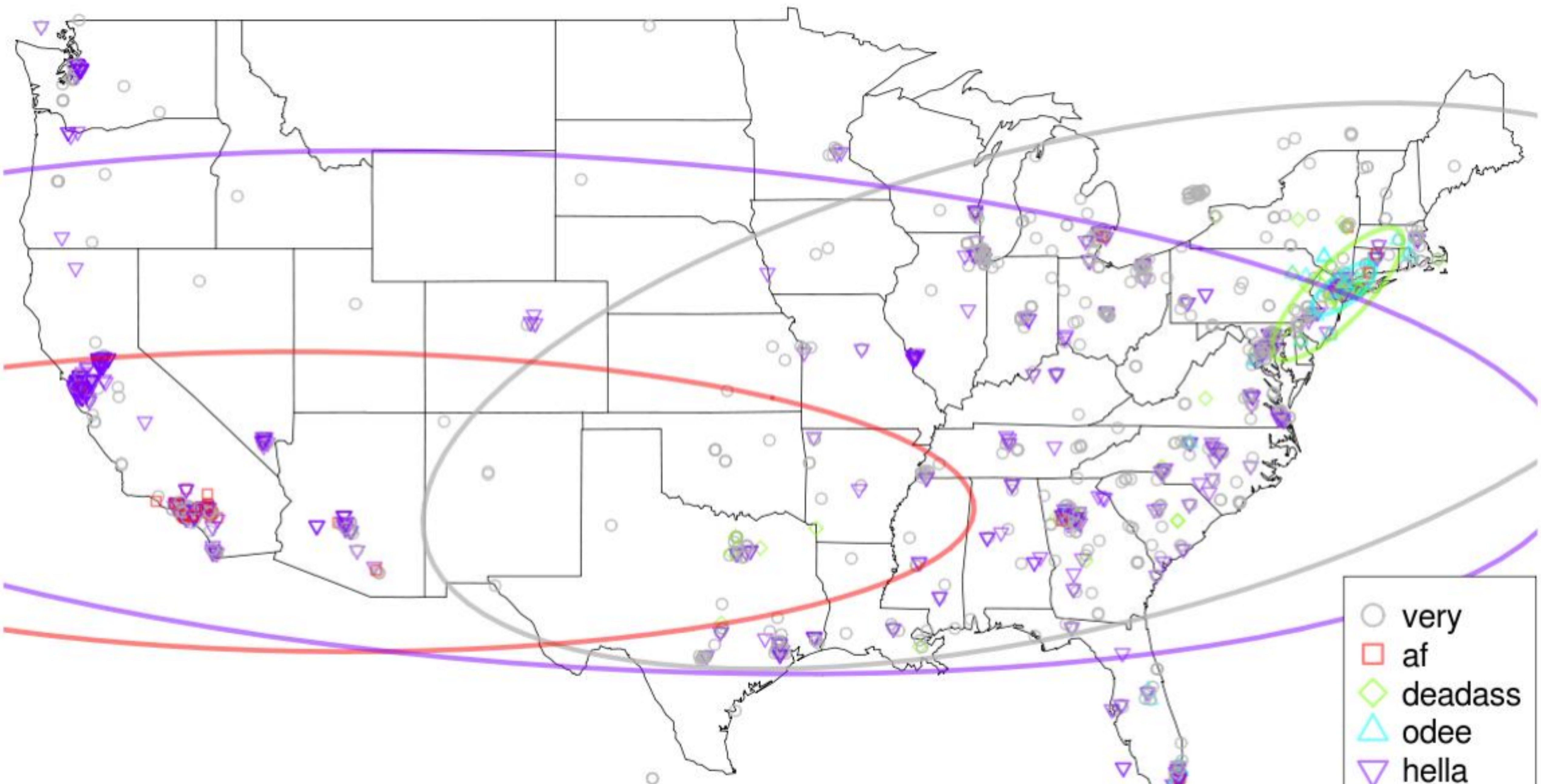
	"basketball"	"popular music"	"daily life"	"emoticons"	"chit chat"
	PISTONS KOBE LAKERS game DUKE NBA CAVS STUCKEY JETS KNICKS	album music beats artist video #LAKERS ITUNES tour produced vol	tonight shop weekend getting going chilling ready discount waiting iam	:) haha :d :(;) :p xd :/ hahaha hahah	lol smh jk yea wyd coo ima wassup somethin jp
Boston 	CELTICS victory BOSTON CHARLOTTE				
N. California 	THUNDER KINGS GIANTS pimp trees clap				
New York 	NETS KNICKS				
Los Angeles 	#KOBE #LAKERS AUSTIN				
Lake Erie 	CAVS CLEVELAND OHIO BUCKS od COLUMBUS				

Analysis

	"basketball"	"popular music"	"daily life"	"emoticons"	"chit chat"
	PISTONS KOBE LAKERS game DUKE NBA CAVS STUCKEY JETS KNICKS	album music beats artist video #LAKERS ITUNES tour produced vol	tonight shop weekend getting going chilling ready discount waiting iam	:) haha :d :(;) :p xd :/ hahaha hahah	lol smh jk yea wyd coo ima wassup somethin jp
Boston 	CELTICS victory BOSTON CHARLOTTE	playing daughter PEARL alive war comp	BOSTON	;p gna loveee	*ese* exam suttin sippin
N. California 	THUNDER KINGS GIANTS pimp trees clap	SIMON dl mountain seee	6am OAKLAND	*pues* hella koo SAN fckn	hella flirt hut iono OAKLAND
New York 	NETS KNICKS	BRONX	iam cab	oww	wasssup nm
Los Angeles 	#KOBE #LAKERS AUSTIN	#LAKERS load HOLLYWOOD imm MICKEY TUPAC	omw tacos hr HOLLYWOOD	af *papi* raining th bomb coo HOLLYWOOD	wyd coo af *nada* tacos messin fasho bomb
Lake Erie 	CAVS CLEVELAND OHIO BUCKS od COLUMBUS	premiere prod joint TORONTO onto designer CANADA village burr	stink CHIPOTLE tipsy	;d blvd BIEBER hve OHIO	foul WIZ salty excuses lames officer lastnight


something
sumthin
suttin


lol
lls
ctfu


very
af
deadass
odee
hella

Linguistic Diffusion

- Which groups or geographic regions influence others? Where do new linguistic trends start?
- How do geographic or demographic factors affect linguistic transmission?
- Work in progress:
  Mapping the geographical diffusion of new words
  http://arxiv.org/abs/1210.5268
  Eisenstein, O'Connor, Smith, Xing
  NIPS 2012 Workshop on Social Network and Social Media Analysis

How do linguistic innovations spread?

- Distance
  - *Wave* and *Gravity* models: linguistic innovations spread throughout a person's life; likelihood of contact based on distance (Bailey 1973, Trudgill 1974)
  - *Cascade* model: innovations travel from largest city to second-largest, etc. (Labov 2003)
- Cultural factors
  - e.g. African-American influences on standard American English: *cool*, *rip off*, *uptight*

How to study language change?

- Small number of carefully chosen linguistic variables -- esp. phonology
- Variation over *apparent time*: compare people of different ages
- Complementary approach from social media data
  - Time of posting (over a few years)
  - Lexical variables
- Fit simple regression-based model to large amount of real data
  - Contrast: theoretical quantitative models (dynamical systems, Nash equilibria, Bayesian learners, agent-based simulations...)

Data

- 86 weeks (Dec 2009 to May 2011)
  - 44 million geotagged messages (500k authors)
  - Aggregated by week
- Assigned to 200 Metropolitan Statistical Areas; demographics from U.S. Census
  - Simple interpretability
- 10,000 most frequent words => 1,818 (used at least 5 times in one week within one MSA)

Diffusion of novel words

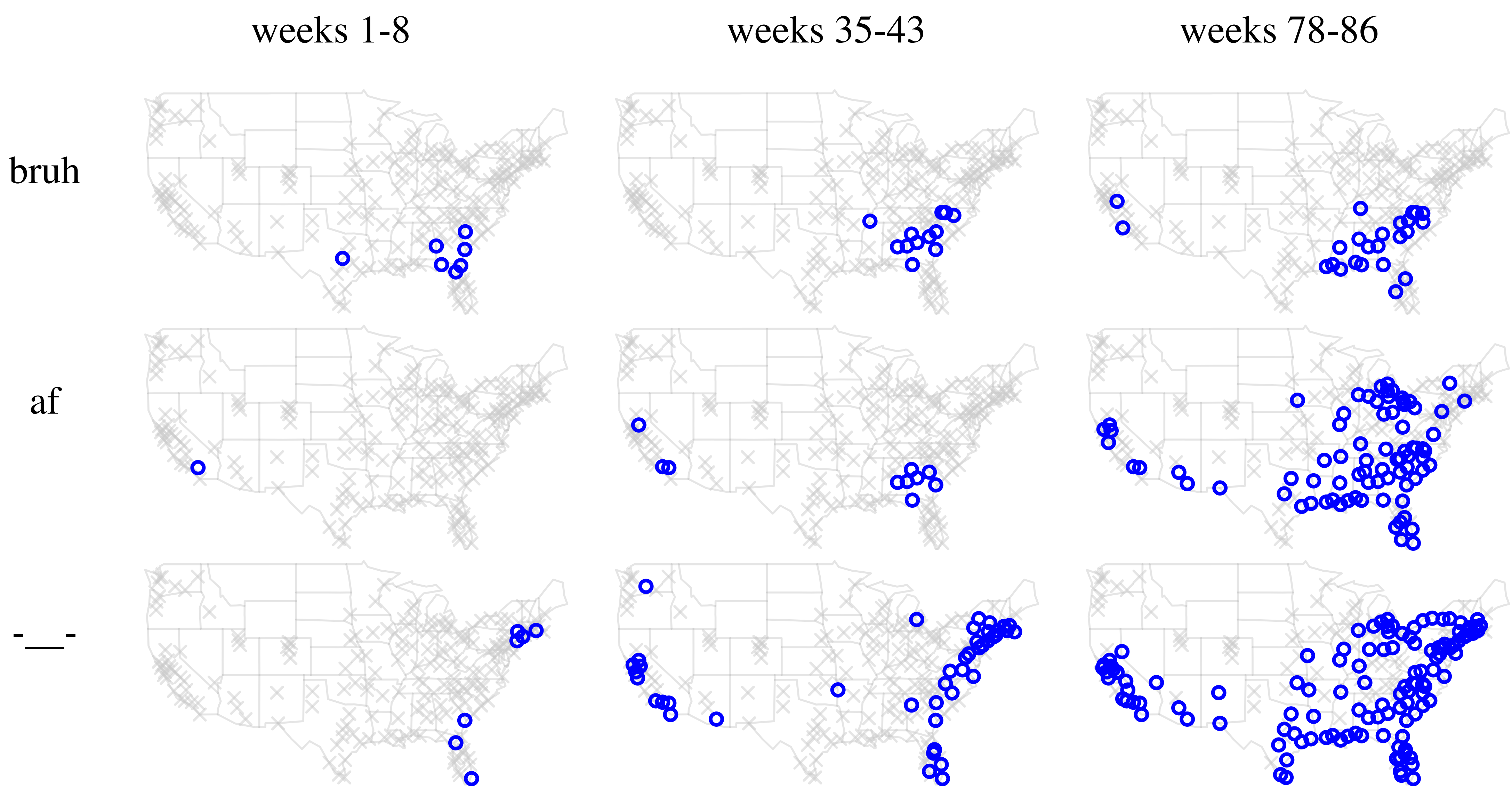


Figure 1: Change in popularity for three words: *bruh*, *af*, -__-. Blue circles indicate cities in which the probability of an author using the word in a week in the listed period is greater than 1% (2.5% for *bruh*).

Model

- MSAs as nodes in diffusion network
- "Influence"/"Transmission" from ***r*** -> ***s***: if words popular in region *r* later become popular in *s*. *[Lead-lag notion of influence/transmission]*
- Can't use simple cross-correlations: confounds for size of region, global events (e.g. TV shows, holidays), etc.
- Logistic binomial probability whether an author uses a word at a particular (region, time)
- "Influence" represented as linear dynamical system (latent vector autoregression)

Model: word frequencies by time and region

$$c_{i,r,t} \sim \text{Binomial}(s_{r,t}, \sigma(\nu_i + \tau_{r,t} + \eta_{i,*,t} + \eta_{i,r,t}))$$

- $c_{i,r,t}$: count of authors who use word *i* in region *r* at time *t*
- $s_{r,t}$: count of authors who post messages in region *r* at time *t*
- ν_i: overall (logodds) frequency of word *i*
- $\tau_{r,t}$: general activation of region *r* at time *t*
- $\eta_{i,*,t}$: global activation of word *i* at time *t*

$$\boldsymbol{\eta}_{i,t} \sim \text{Normal}(\mathbf{A}\boldsymbol{\eta}_{i,t-1}, \boldsymbol{\Gamma}) \qquad E[\eta_{i,t,r}] = \sum_s A_{r,s}\eta_{i,t-1,s}$$

$\mathbf{A}$	autoregressive coefficients (size $R \times R$)
$\boldsymbol{\Gamma}$	variance of the autoregressive process (size $R \times R$)

- Model fitting with EM
  - (30 million variables, takes a while)
  - Fitting eta: Gaussian approximation to logistic-binom
  - Kalman smoother to learn diag(A); non-diagonals from VAR fits to eta samples

Smoothing examples

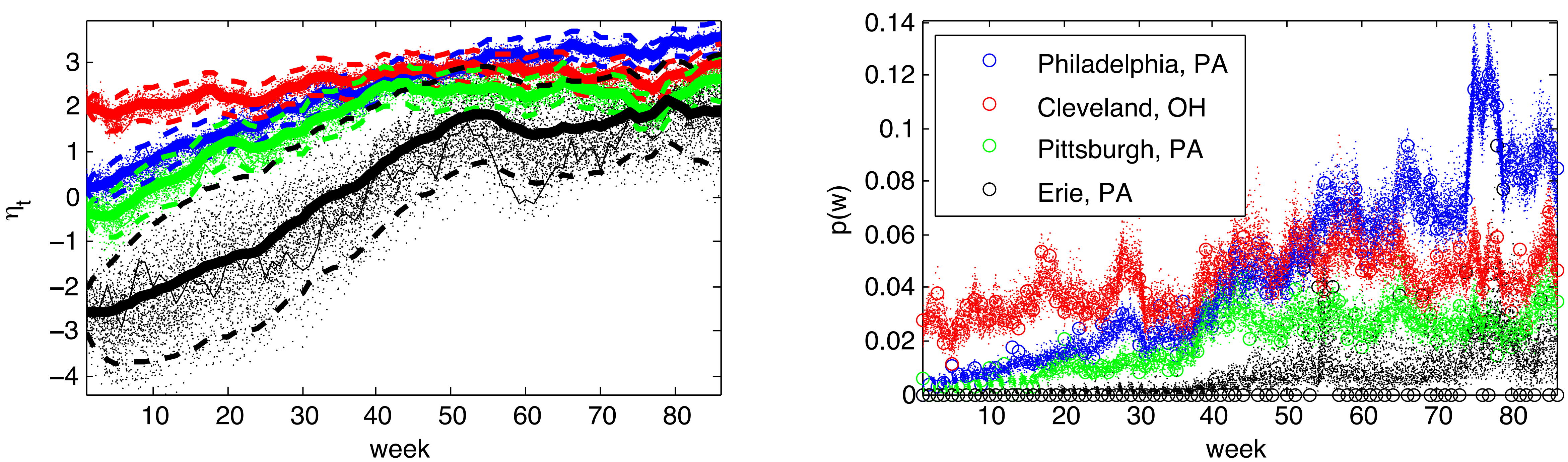
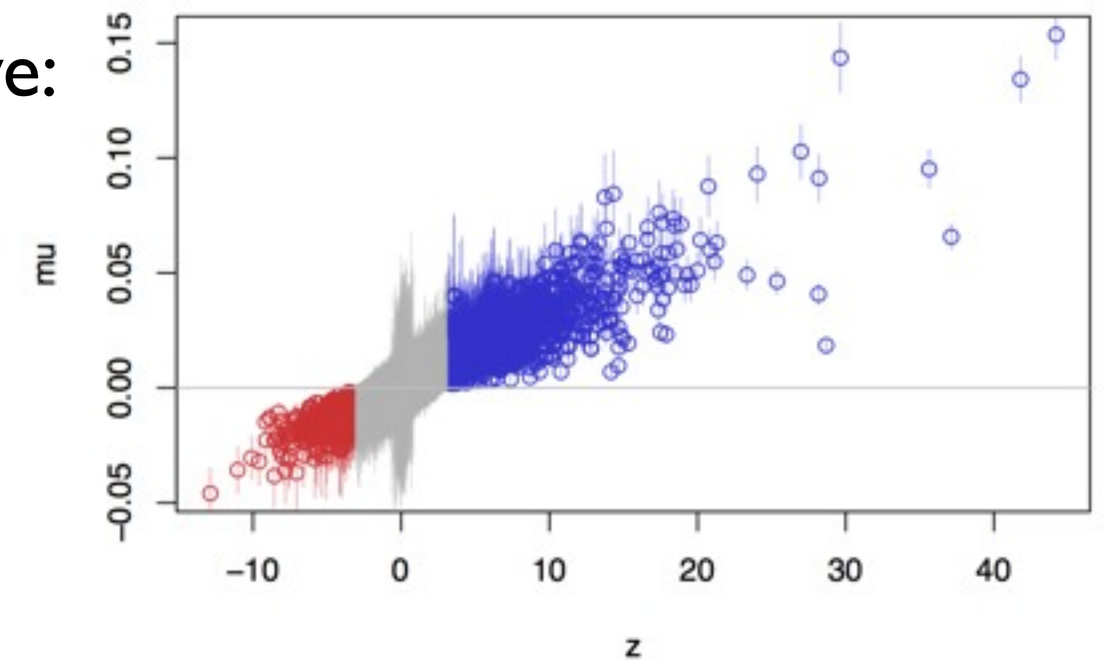


Figure 2: **Left**: Monte Carlo and Kalman smoothed estimates of η for the word *ctfu* in five cities. **Right**: estimates of term frequency in each city. Circles show empirical frequencies.

Transmission edge selection

- Simulated posterior samples of A matrix
- Which of the 39,600 non-diagonals (edges) are significantly non-zero? Multiple hypothesis testing
- Bonferroni correction too conservative: controlling familywise error rate
  P(>= 1 false positive)
- Benjamini-Hochberg: control the False Discovery Rate
  E[% false positive] <= 0.01



$$FDR = 0.01 = \frac{\mathbb{E}[\#\{\text{that pass under null hypothesis}\}]}{\#\{\text{that pass empirically}\}} = \frac{200(199)(1-\Phi(\bar{z}))}{\#\{z_{m,n} > \bar{z}\}}$$

yields 3544 edges

Transmission network

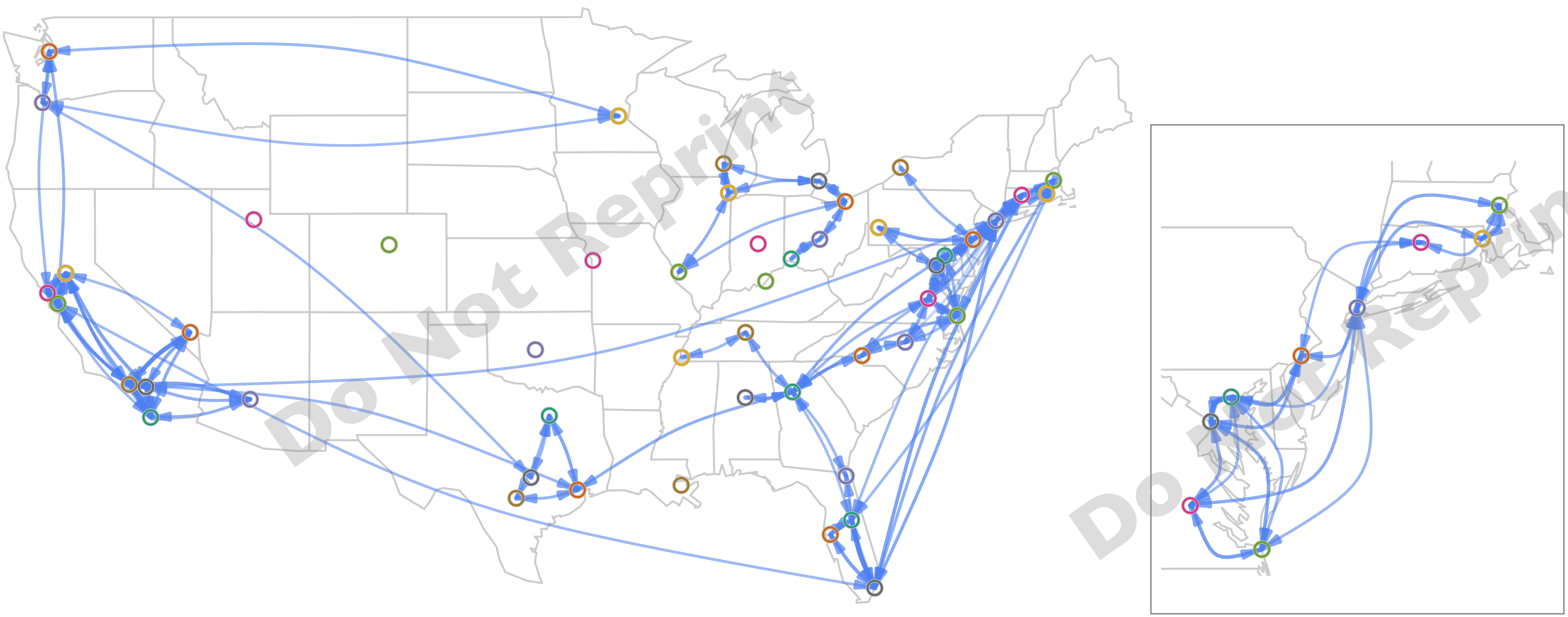
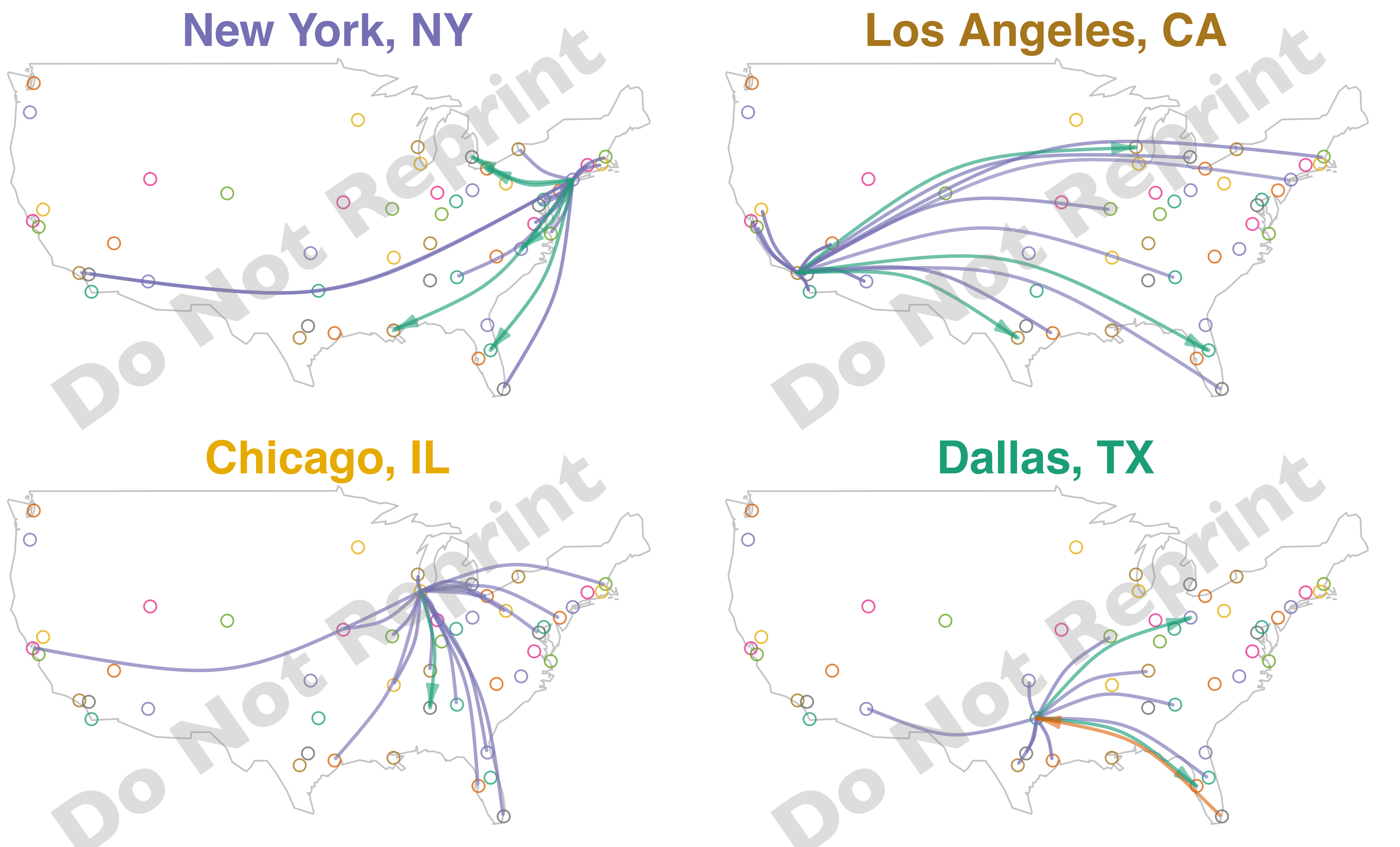
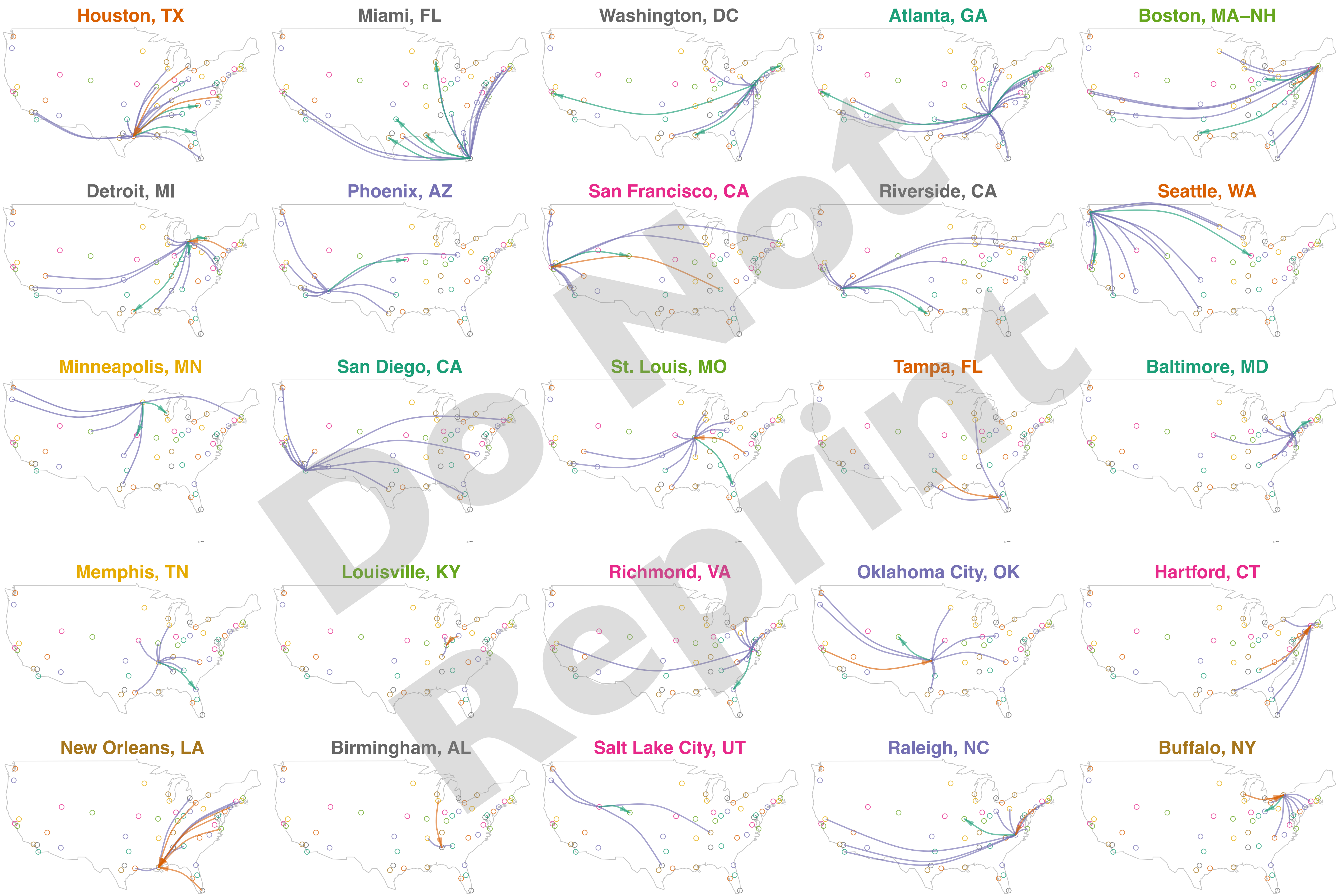


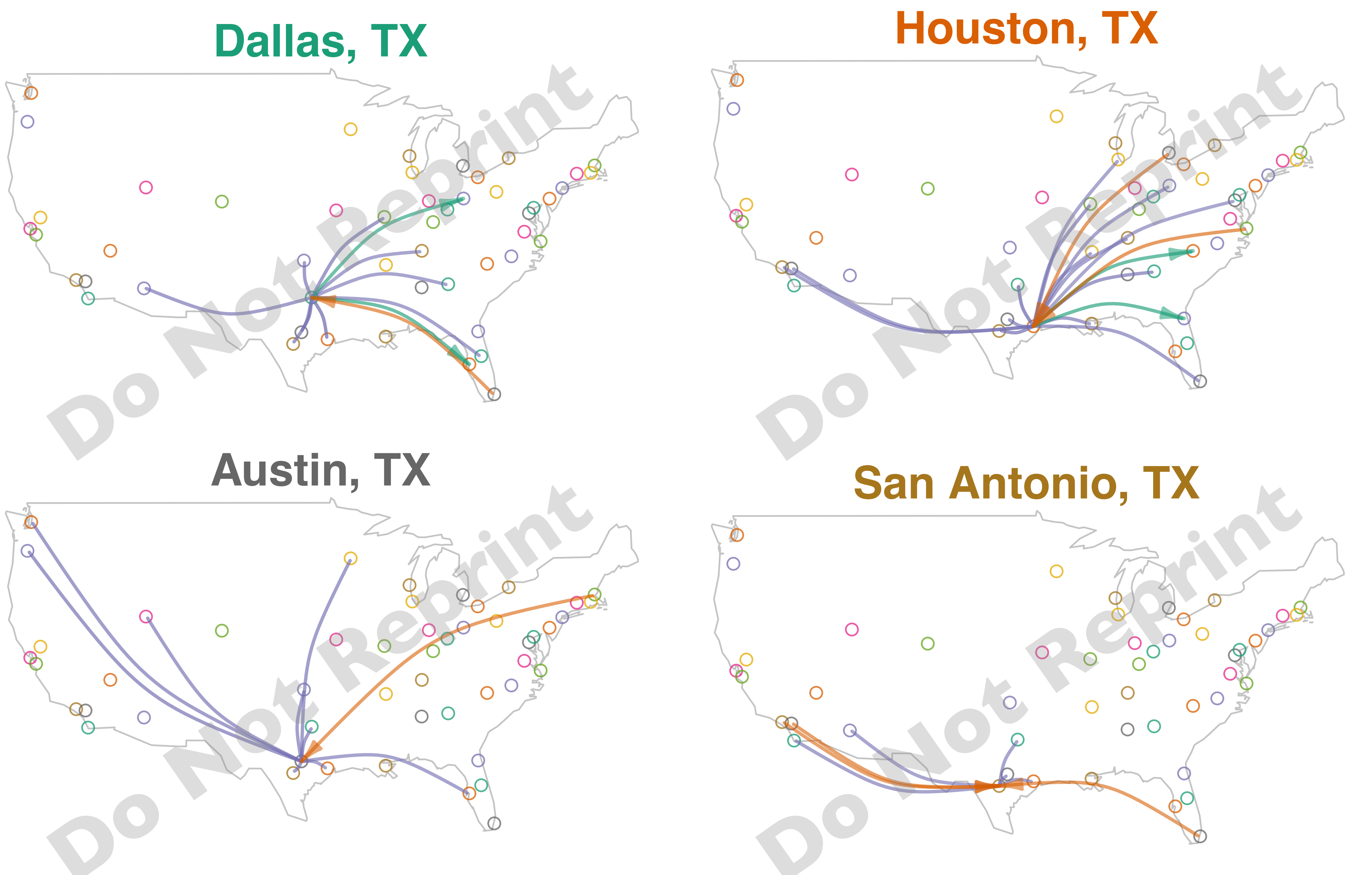
Figure 4: Lexical influence network: high-confidence, high-influence links ($z > 3.12, \mu > 0.025$). **Left:** among all 50 largest MSAs. **Right:** subnetwork for the Eastern seaboard area. See also Figure 5, which uses same colors for cities.

Transmission ego networks




Houston, TX
Miami, FL
Washington, DC
Atlanta, GA
Boston, MA–NH
Detroit, MI
Phoenix, AZ
San Francisco, CA
Riverside, CA
Seattle, WA
Minneapolis, MN
San Diego, CA
St. Louis, MO
Tampa, FL
Baltimore, MD
Memphis, TN
Louisville, KY
Richmond, VA
Oklahoma City, OK
Hartford, CT
New Orleans, LA
Birmingham, AL
Salt Lake City, UT
Raleigh, NC
Buffalo, NY

How does distance matter?



Symmetric properties of transmission

	geo distance	% White	% Af. Am.	% Hispanic	% urban	%renter	log income
linked	10.5 ± 0.5	10.2 ± 0.4	8.45 ± 0.37	9.88 ± 0.64	9.55 ± 0.34	6.30 ± 0.24	0.181 ± 0.006
unlinked	20.8 ± 0.6	16.2 ± 0.5	16.3 ± 0.5	15.2 ± 0.8	12.0 ± 0.4	6.78 ± 0.25	0.201 ± 0.007

Table 2: Geographical distances and absolute demographic differences for linked and non-linked pairs of MSAs. Confidence intervals are $p < .01$, two-tailed.

(a) Logistic regression coefficients predicting influence links between MSAs. Bold typeface indicates statistical significance at $p < .01$.

	estimate	s.e.	t-value
intercept	-0.0601	0.0287	-2.10
product of populations	0.393	0.048	8.22
distance	-0.870	0.033	-26.1
abs. diff. % White	-0.214	0.040	-5.39
abs. diff. % Af. Am.	-0.693	0.042	-16.7
abs. diff. % Hispanic	-0.140	0.030	-4.63
abs. diff. % urban	-0.170	0.030	-5.76
abs. diff. % renters	-0.0314	0.0304	-1.04
abs. diff. log income	0.0458	0.0301	1.52

(b) Accuracy of predicting influence links, with ablated feature sets.

feature set	accuracy	gap
all features	72.3	
-population	71.6	0.7
-geography	67.6	4.7
-demographics	66.9	5.4

Senders vs. Receivers

	Log pop.	**% White**	**% Af. Am**	% Hispanic	**% Urban**	**% Renters**	**Log income**
difference	0.968	-0.0858	0.0703	0.0094	0.0612	0.0231	0.0546
s.e.	0.0543	0.0065	0.0063	0.0098	0.0054	0.0041	0.0113
z-score	17.8	-13.2	11.1	0.950	11.3	5.67	4.82

Table 4: Differences in demographic attributes between senders and receivers of lexical influence. Bold typeface indicates statistical significance at $p < .01$.

	Log pop.	% White	**% Af. Am**	% Hispanic	% Urban	% Renters	Log income
weights	2.22	-0.246	1.08	0.0914	-0.129	-0.0133	0.225
s.e.	0.290	0.315	0.343	0.229	0.221	0.180	0.194
t-score	7.68	-0.78	3.15	0.40	-0.557	-0.0736	1.16

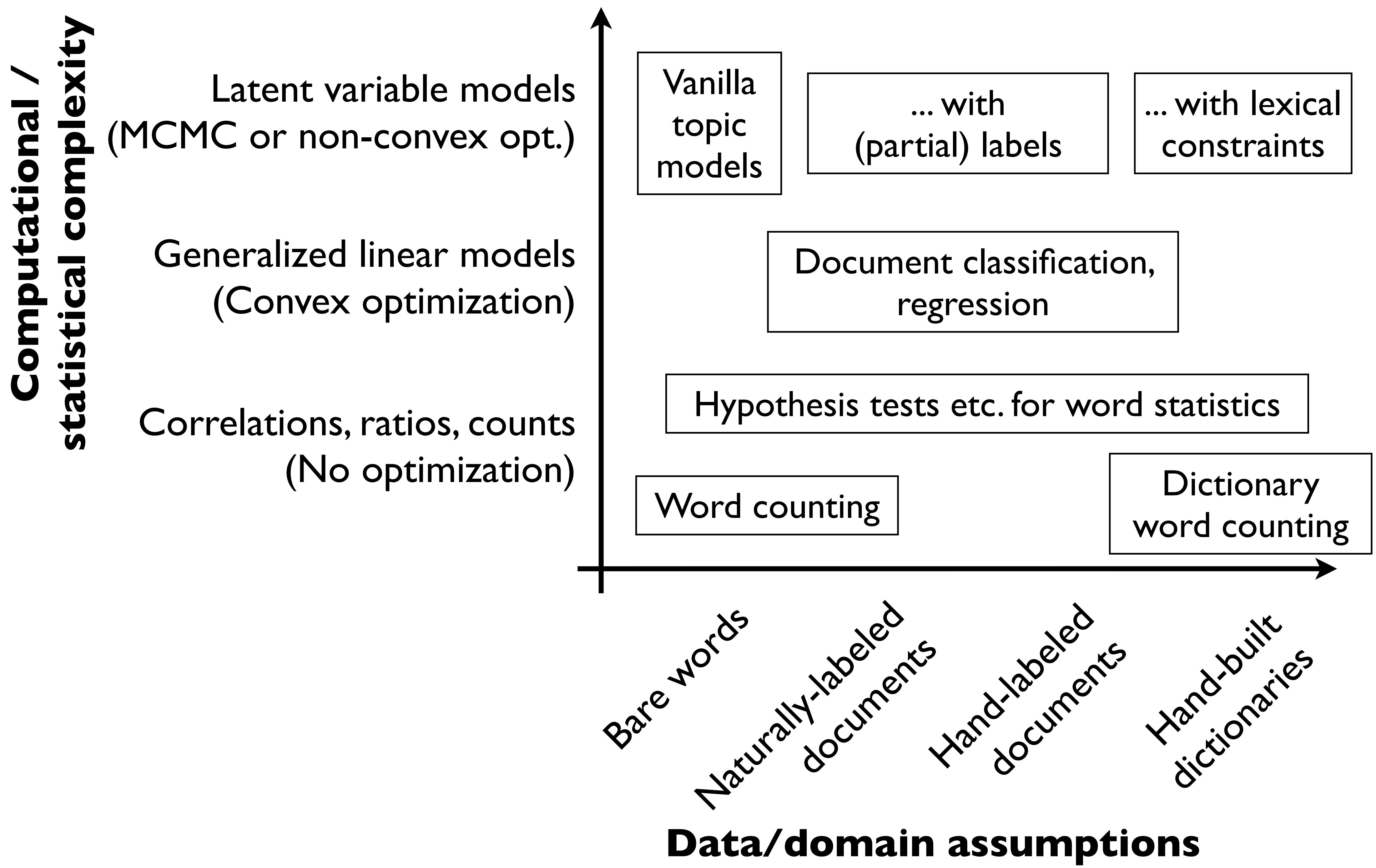
Table 5: Regression coefficients for predicting direction of influence. Bold typeface indicates statistical significance at $p < .01$.

- Remember biases in sample: Twitter skews to minorities, though exact demographic composition of our sample is not known

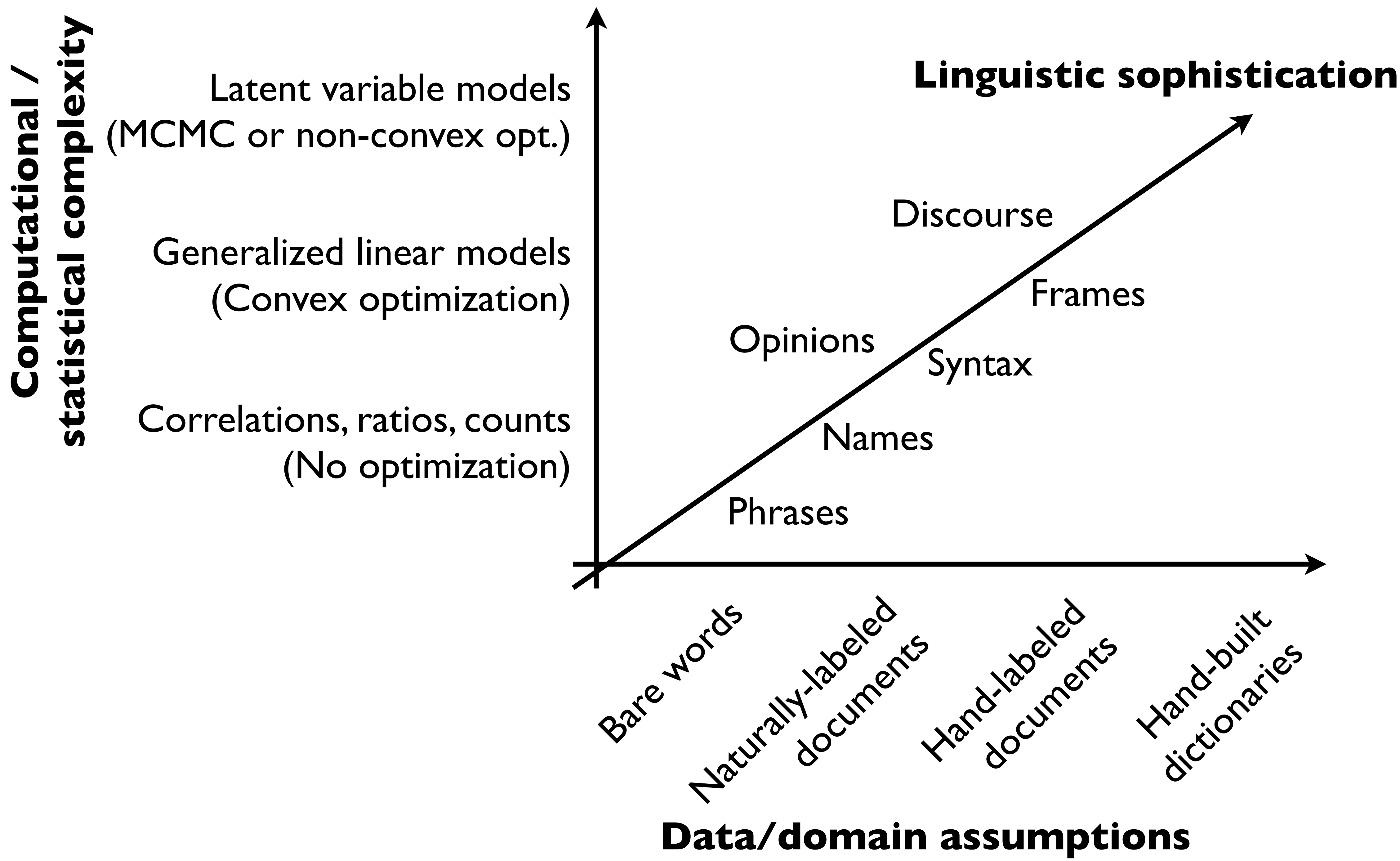
Tentative conclusions

- Strong roles for both geography and demographics
  - But demographics remarkable, given metropolitan area-level aggregation
  - Need neighborhood-level aggregation?
- Racial homophily most important?
  - Note socioeconomic class difficult to assess from this Census data
- Issues with word independence assumption?
  - e.g. *bruh* related to *bro*?

Taxonomy of text analysis methods



Taxonomy of text analysis methods

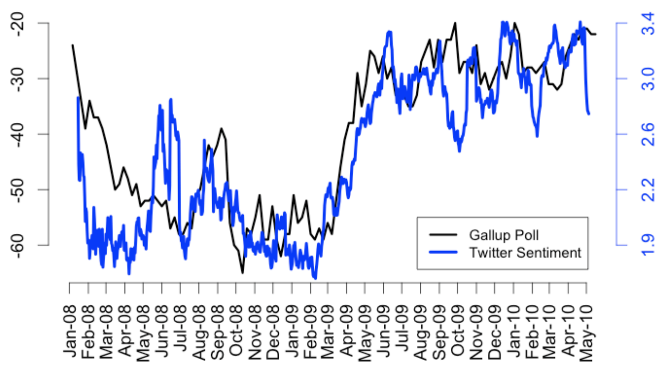


Thanks!

All papers, slides available at http://brenocon.com
Feedback welcome!

Opinions and Time

Brendan O'Connor, Ramnath Balasubramanyan, Bryan Routledge, Noah Smith



Language and Geography

Jacob Eisenstein, Brendan O'Connor, Noah Smith, Eric Xing