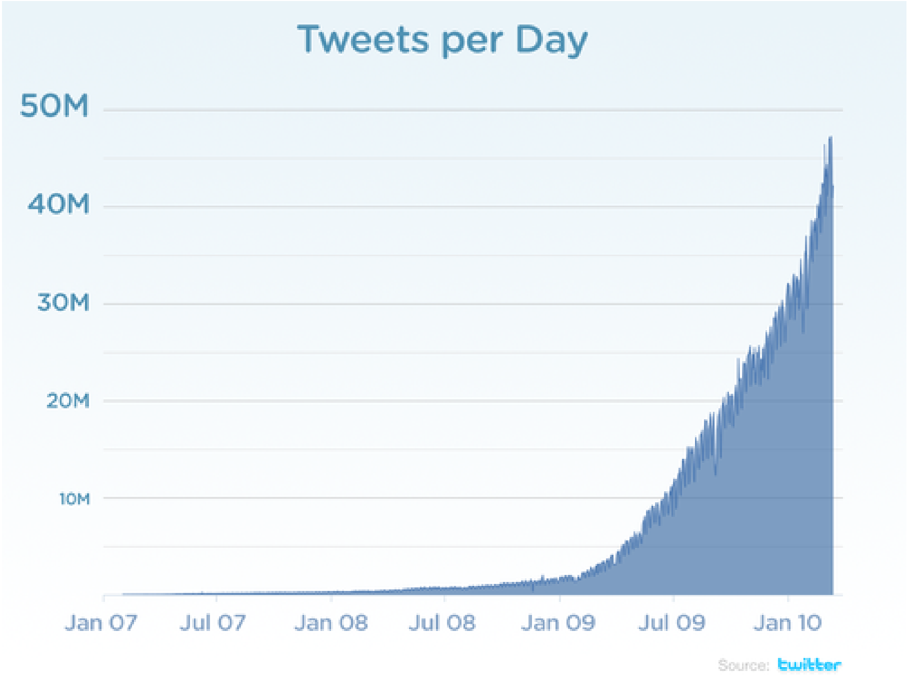
Statistical Exploration of Geographical Lexical Variation in Social Media

Jacob Eisenstein
Brendan O'Connor
Noah A. Smith
Eric P. Xing

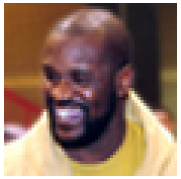
Carnegie Mellon University

Social media

- *Social media* links online text with social networks.
- Increasingly ubiquitous form of social interaction



- Social media text is often conversational and informal.



THE_REAL_SHAQ THE_REAL_SHAQ
@loveJBieber_90 I mite jump on stage and do baby baby baby again u r the best shawty main
28 Oct

Is there geographical variation in social media?

Searching for dialect in social media

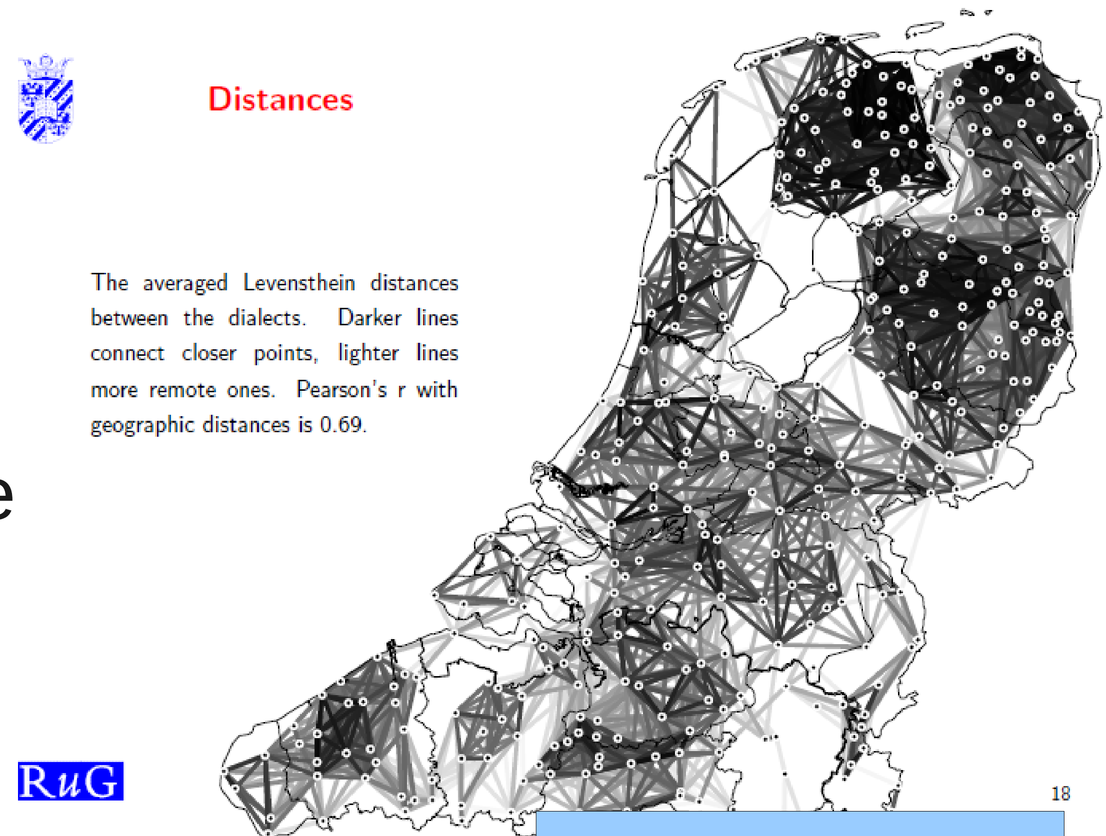
- One approach: search for known variable alternations, e.g. you / yinz / yall

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

Variables and dialect regions

- Given the dialect regions, we could use hypothesis testing to find variables.

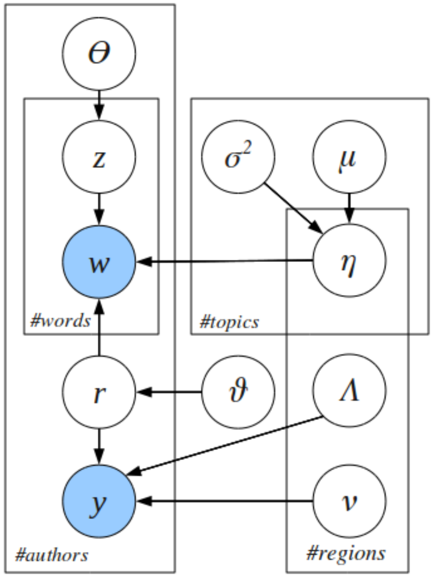
- Given the variables, we could use clustering to find the regions.

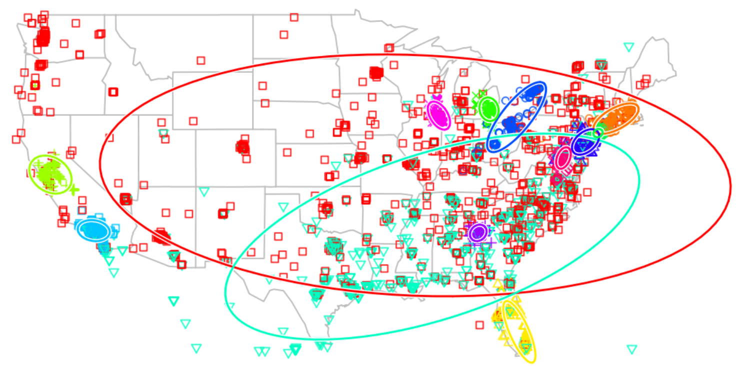


Nerbonne, 2005

- ***Can we infer both the regions and the variables from raw data?***

Outline



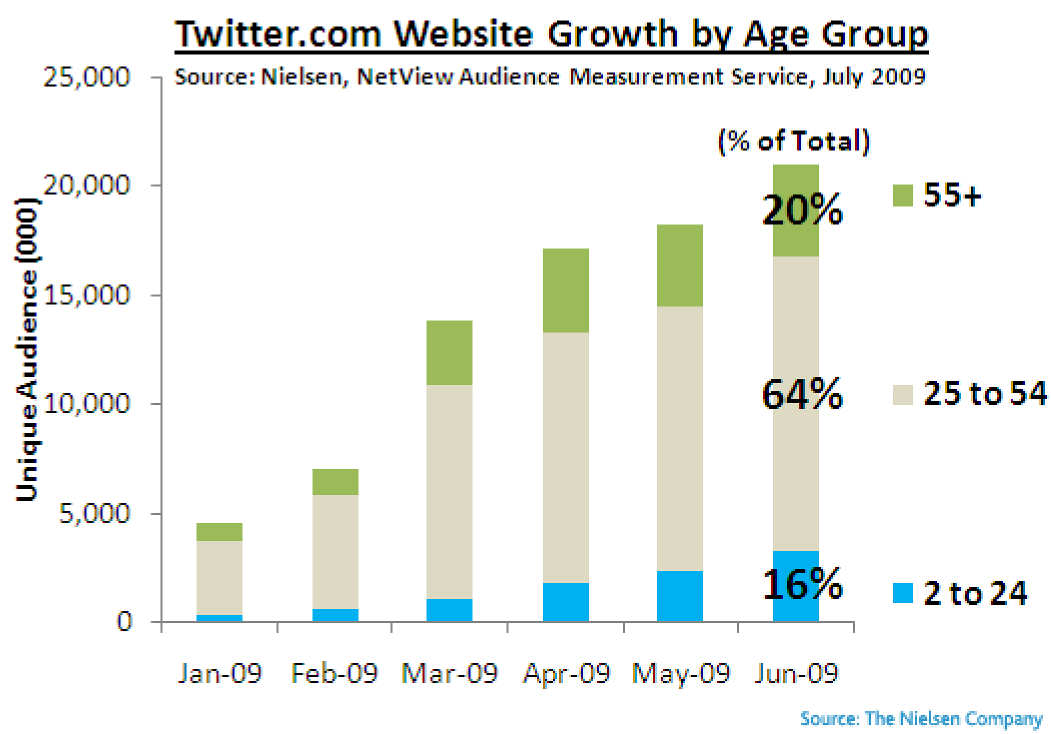


data **model** **results**

Data

Combines *microblogs* and social network.

- Messages limited to 140 characters.
- 65 million “tweets” per day, mostly public
- 190 million users
  - Diverse age, gender, and racial diversity

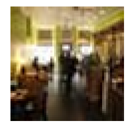


A partial taxonomy of Twitter messages

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

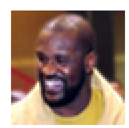


bigdogcoffee bigdogcoffee
Back to normal hours beginning tomorrow........Monday-Friday 6am-10pm Sat/Sun 7:30am-10pm
2 Jan

Links to blog and web content

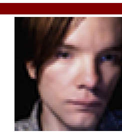
crampell Catherine Rampell
Casey B. Mulligan: Assessing the Housing Sector - http://nyti.ms/hcUKK9
10 hours ago

Celebrity self-promotion



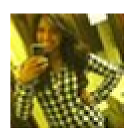
THE_REAL_SHAQ THE_REAL_SHAQ
fill in da blank, my new years shaqalution is ___________
4 Jan

Status messages



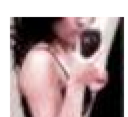
emax electronic max
1.1.11 - britons and americans can agree on the date for once. happy binary day!
1 Jan

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

Geotagged text

- Popular cellphone clients for Twitter encode GPS location.
- We screen our dataset to include only geotagged messages sent from iPhone or Blackberry clients.

Our corpus

- We receive a stream that included 15% of all public messages.
- During the first week of March 2010, we include all authors who:
  - ≥ 20 geotagged messages in our stream
  - From the continental USA
  - Social connections with fewer than 1000 users
- Quick and dirty!
  - Author location = GPS of first post

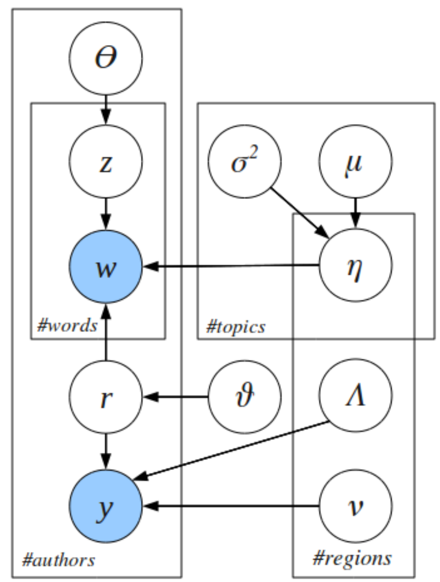
Corpus statistics

- 9500 authors
- 380,000 messages
- 4.7 million tokens
- Highly informal and conversational
  - 25% of the 5000 most common terms are not in the dictionary.
  - More than half of all messages mention another user.

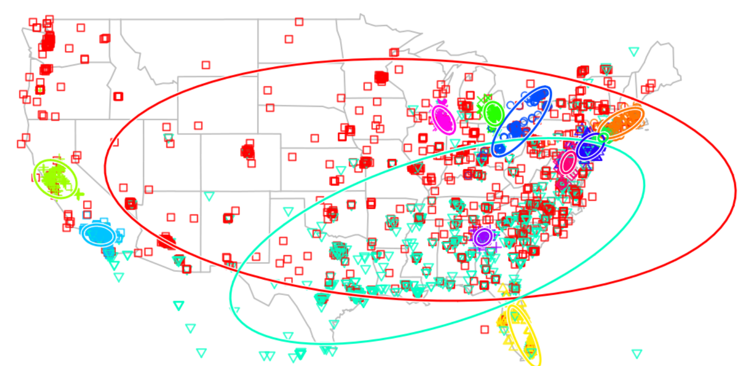
Online at: http://www.ark.cs.cmu.edu/GeoText

Outline

data



model



results

Generative models

- How to simultaneously discover dialect regions and the words that characterize them?
- Probabilistic generative models
  - a.k.a. graphical models
  - Examples:
    - Hidden markov model
    - Naïve Bayes
    - Topic Models a.k.a. Latent Dirichlet Allocation (Blei et al., 2003)

Generative models in 30 seconds

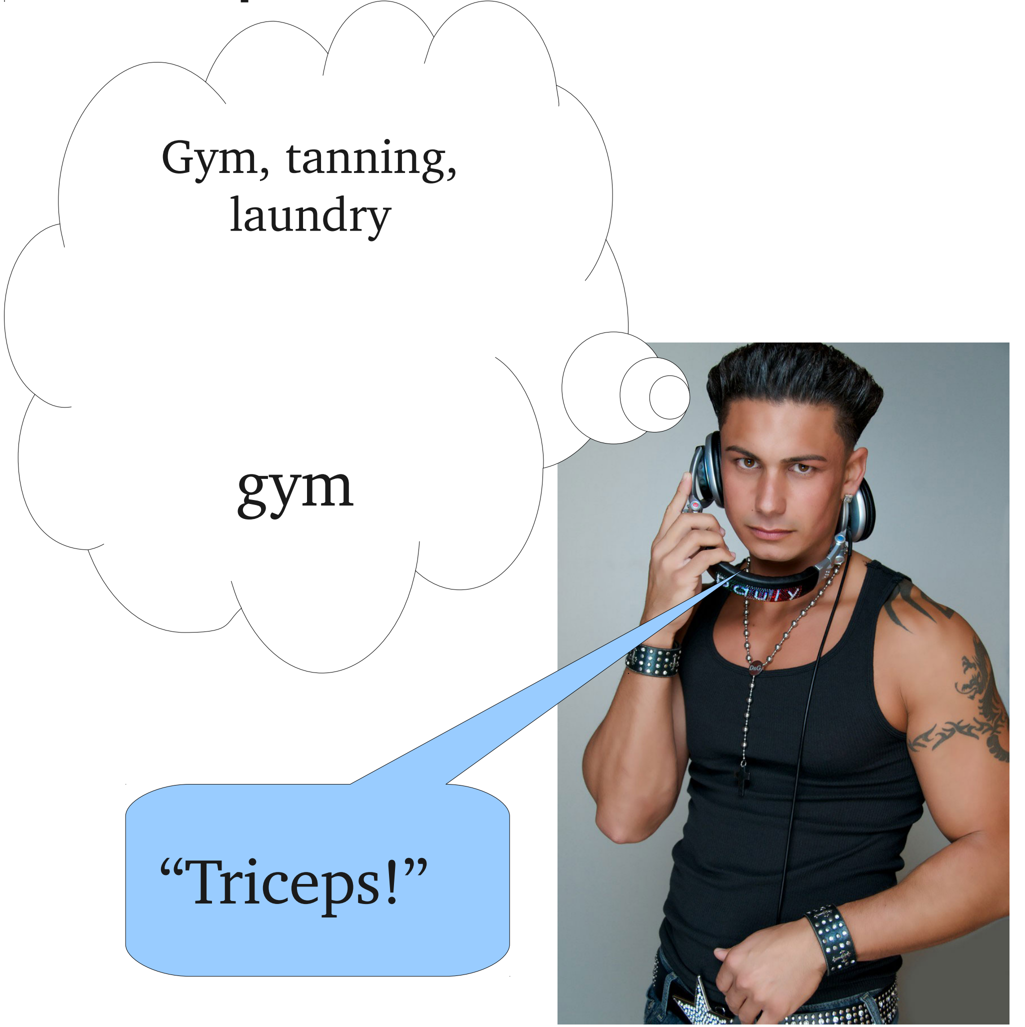
- We hypothesize that text is the output of a stochastic process. For example:

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

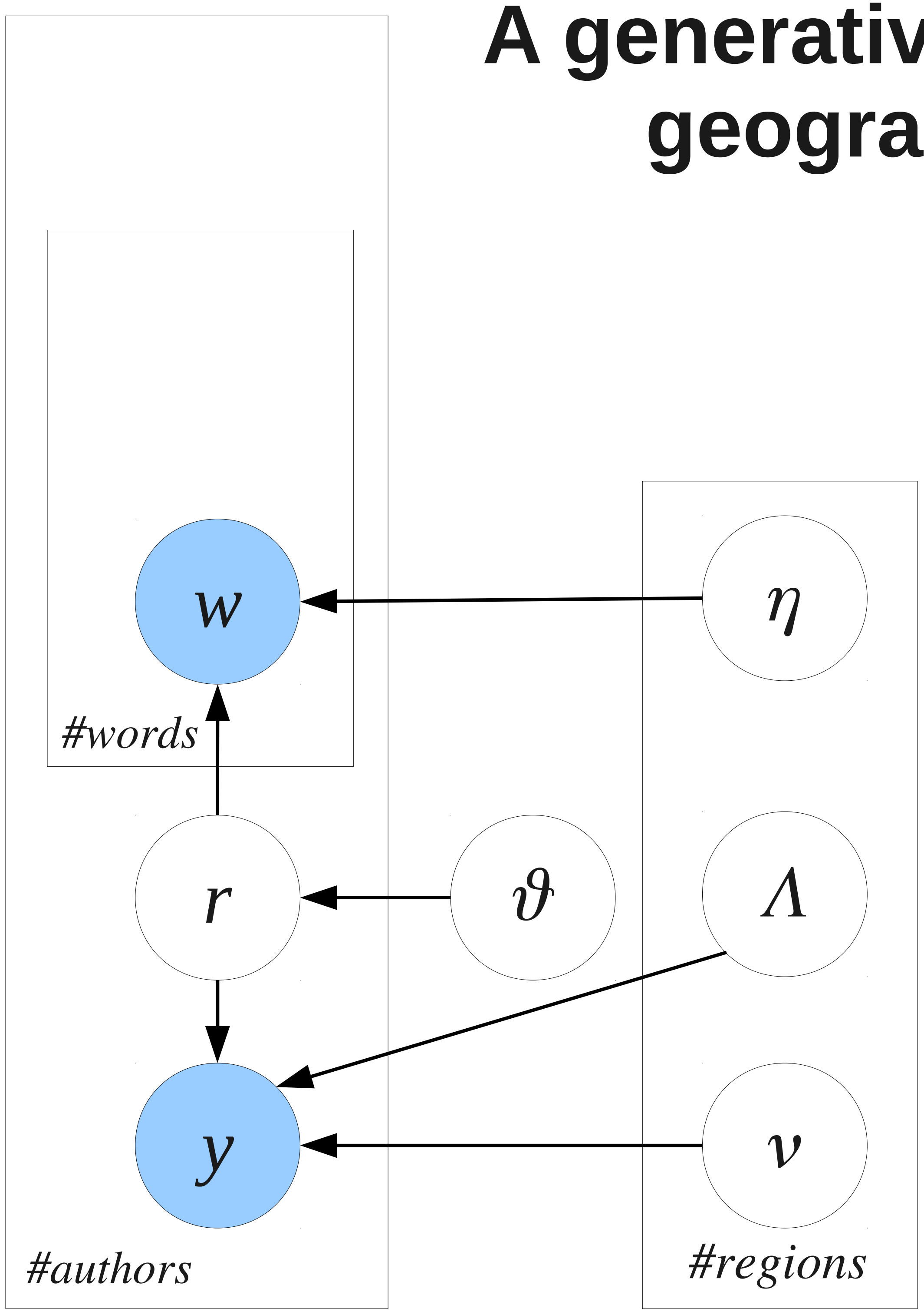
- We only see the output of the generative process.
- Through statistical inference over large amounts of data, we make educated guesses about the hidden variables.

Gym, tanning, laundry

gym

"Triceps!"

A generative model of lexical geographic variation



For each author

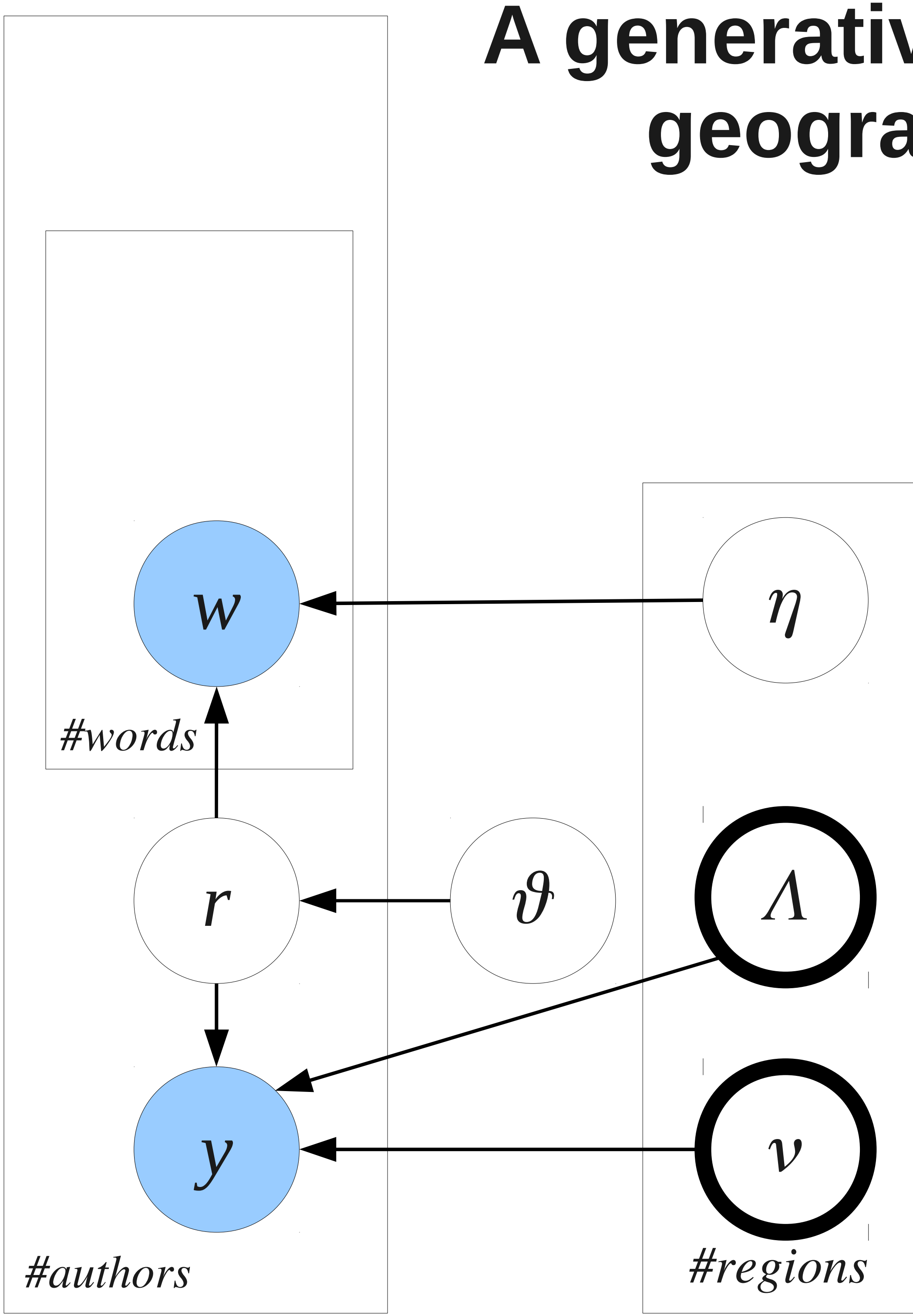
Pick a region from $P(r \mid \vartheta)$

Pick a location from $P(y \mid \Lambda_r, \nu_r)$

For each token

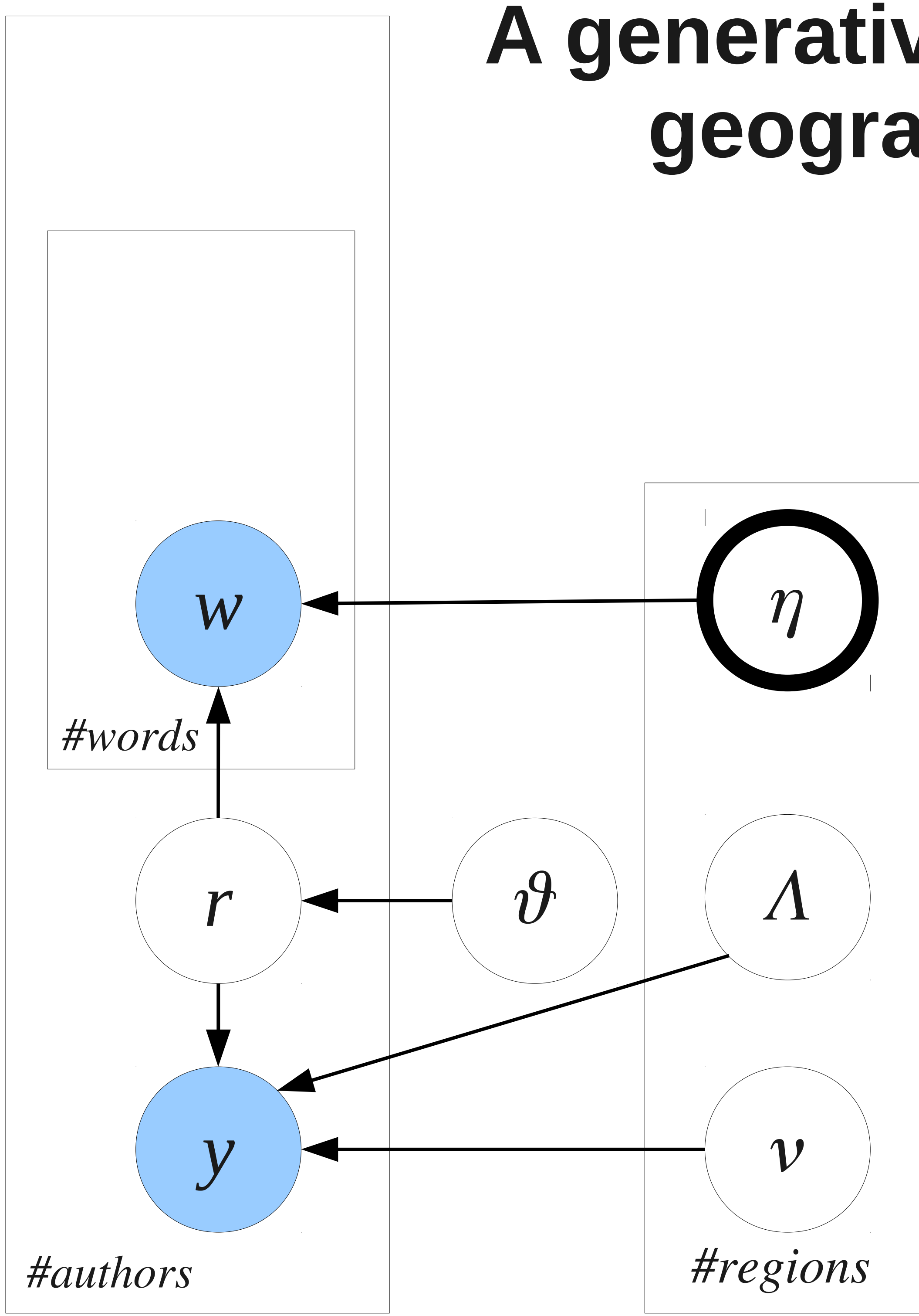
Pick a word from $P(w \mid \eta_r)$

A generative model of lexical geographic variation



ν and Λ define the location and extent of dialect regions

A generative model of lexical geographic variation



ν and Λ define the location and extent of dialect regions

η defines the words associated with each region

Topic models for lexical variation

- Discourse topic is a confound for lexical variation.
- **Solution**: model topical and regional variation jointly
  - Each author's text is shaped by both dialect region and topic
  - Each dialect region contains a unique version of each topic

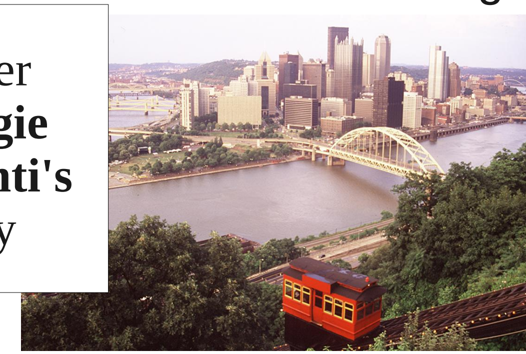
"Food"

Dinner
Delicious
Snack
Tasty

San Francisco

Delicious
Snack
Sprouts
Avocados

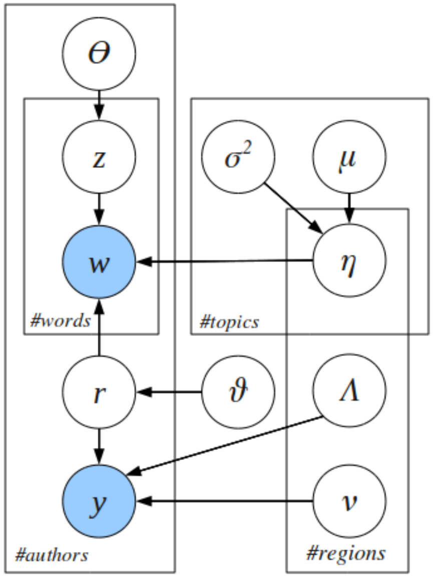
Pittsburgh

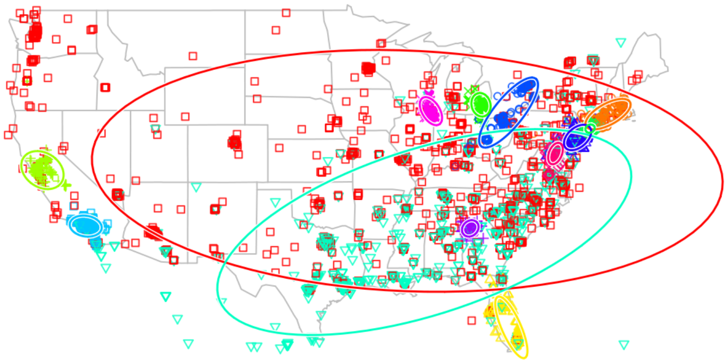


Dinner
Pierogie
Primanti's
Tasty

See our EMNLP 2010 paper for more details

Outline





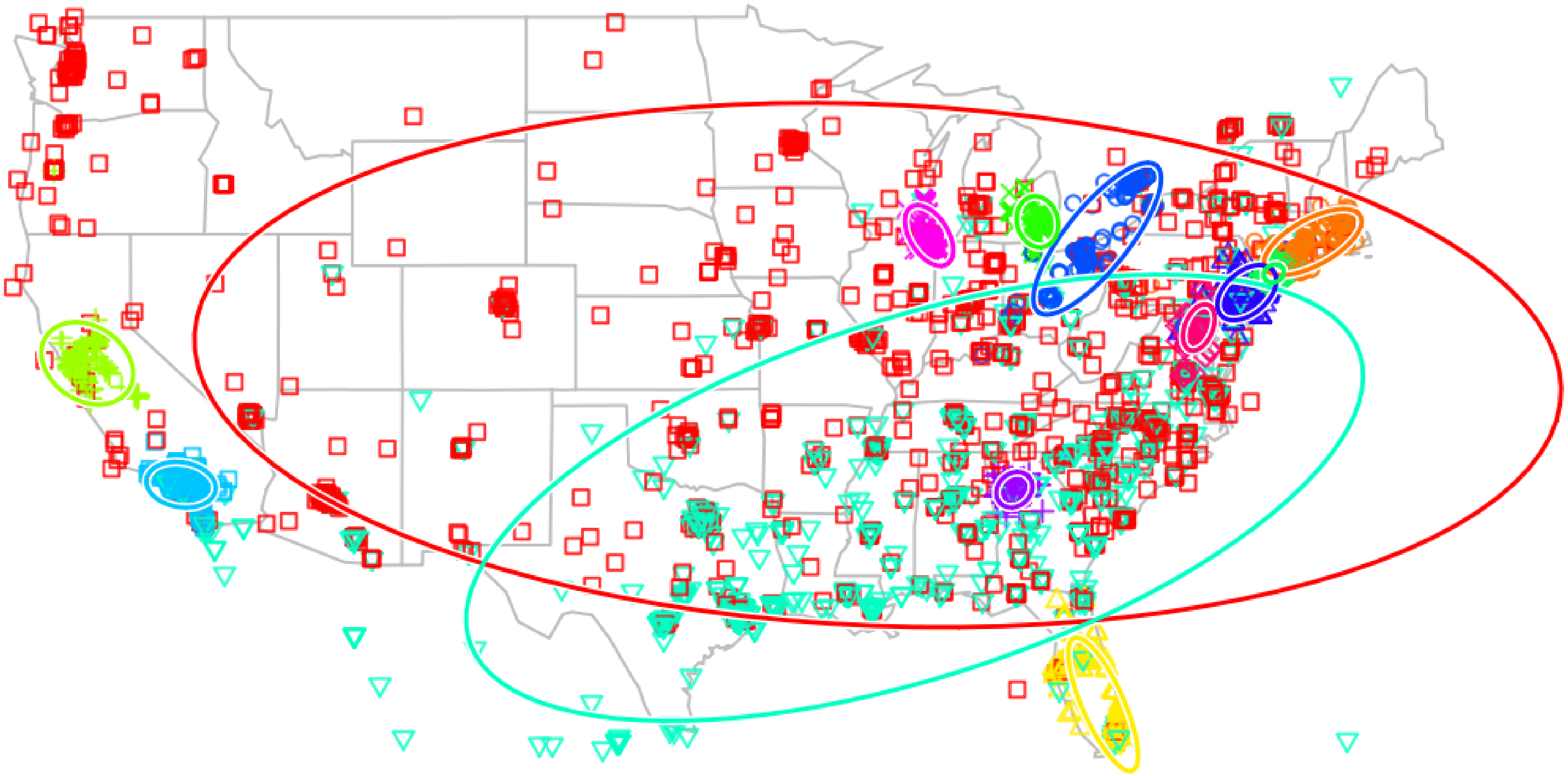
data **model** **results**

Does it work?

Task: predict author location from raw text

METHOD	MEAN ERROR (KM)	MEDIAN ERROR (KM)
Mean location	1148	1018
Text regression	948	712
Generative, no topics	947	644
Generative, topics	***900***	***494***

Induced dialect regions



- Each point is an individual in our dataset
- Symbols and colors indicate latent region membership

Observations

- Many sources of geographical variation
  - Geographically-specific proper names

    boston, knicks (NY), bieber (Lake Eerie)
  - Topics of local prominence:

    tacos (LA), cab (NY)
  - Foreign-language words

    pues (San Francisco), papi (LA)
  - Geographically distinctive “slang” terms

    hella (San Francisco; Bucholtz et al., 2007)

    fasho (LA), suttin (NY)

    coo (LA) / koo (San Francisco)

Discovering alternations

soda / pop / coke

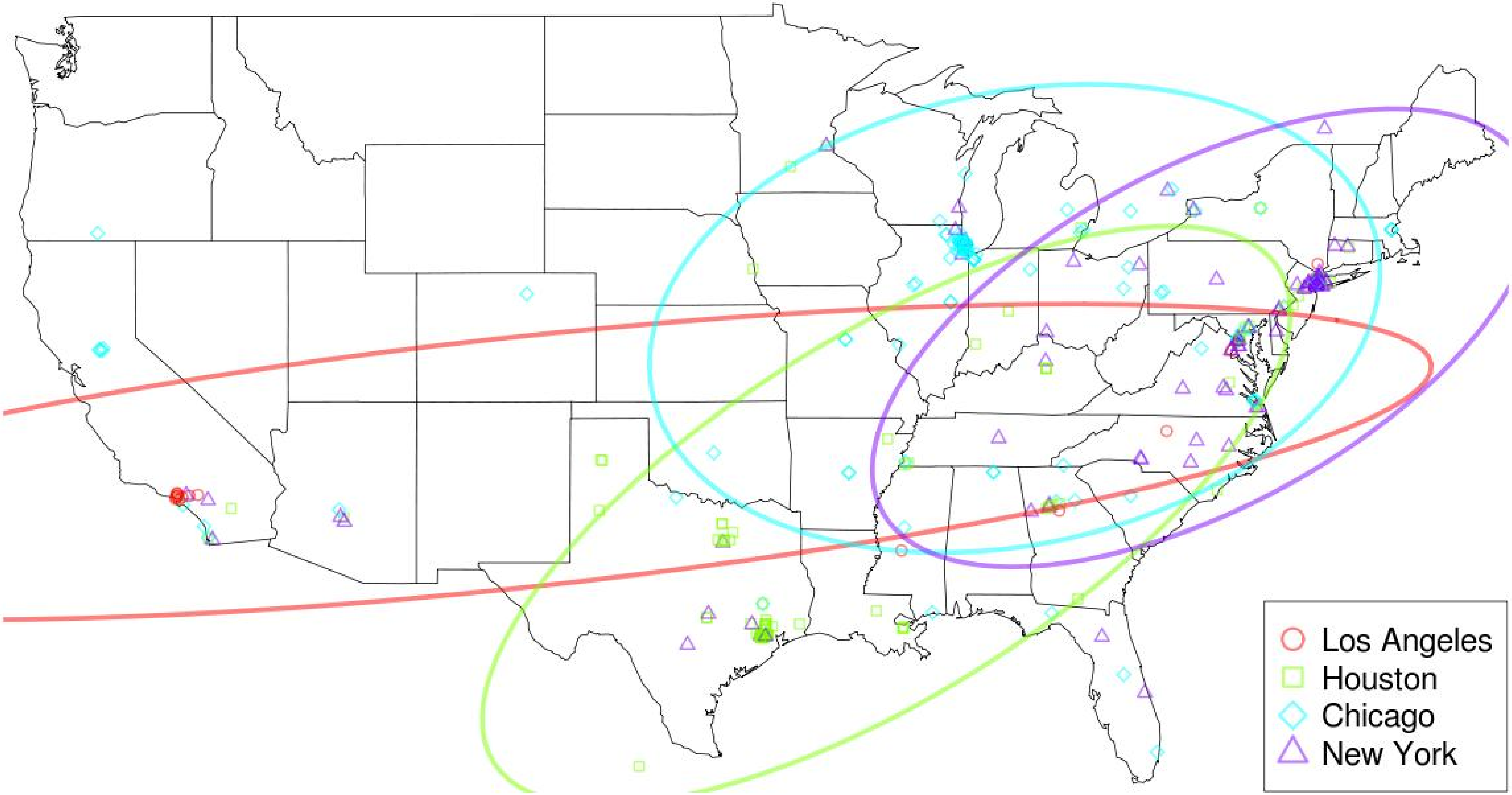
- Criteria:
  - **Geographically distinct**

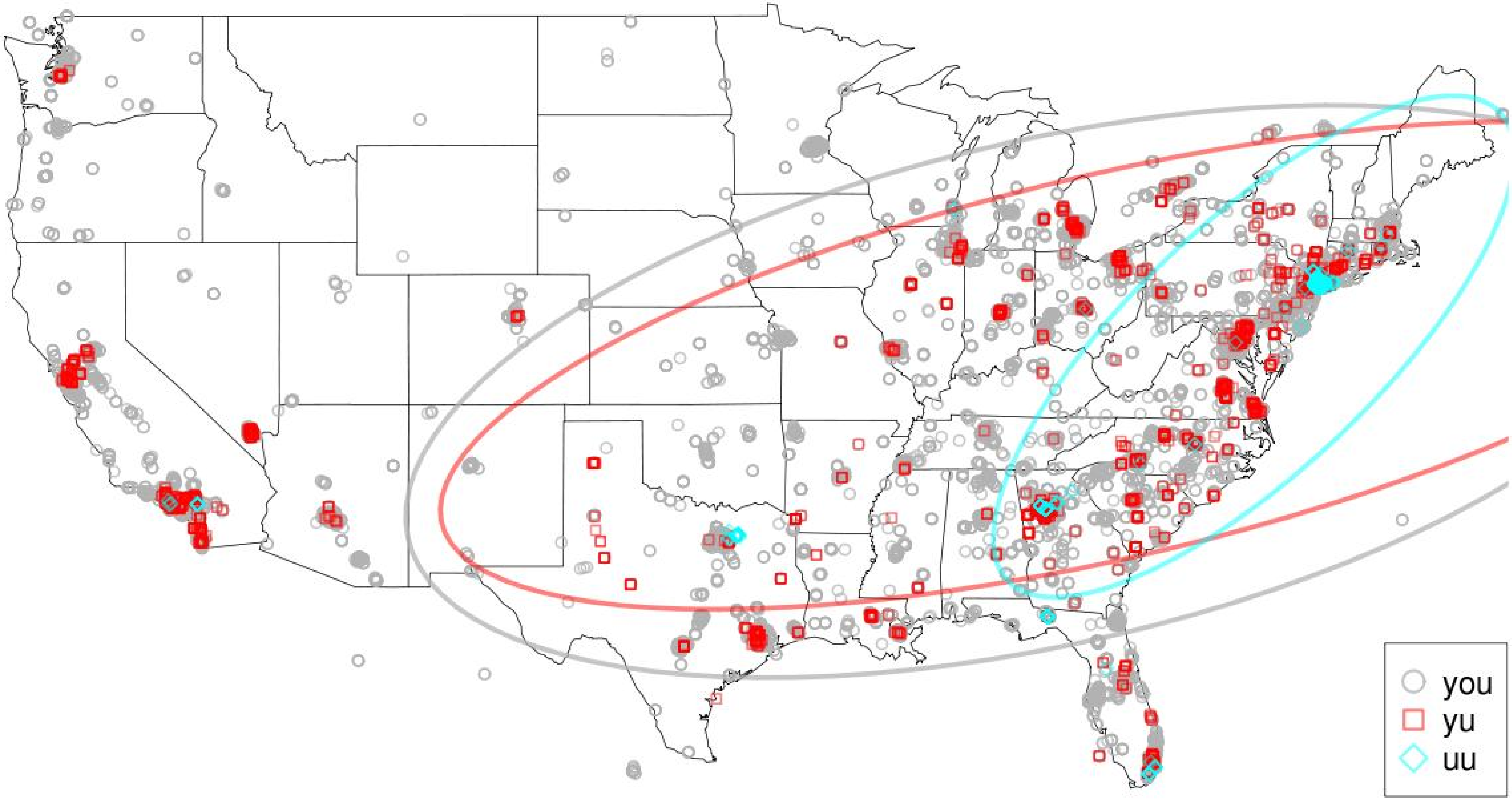
    Maximize divergence of
    P(Region | Word)

  - **Syntactically and** (hopefully) **semantically equivalent**

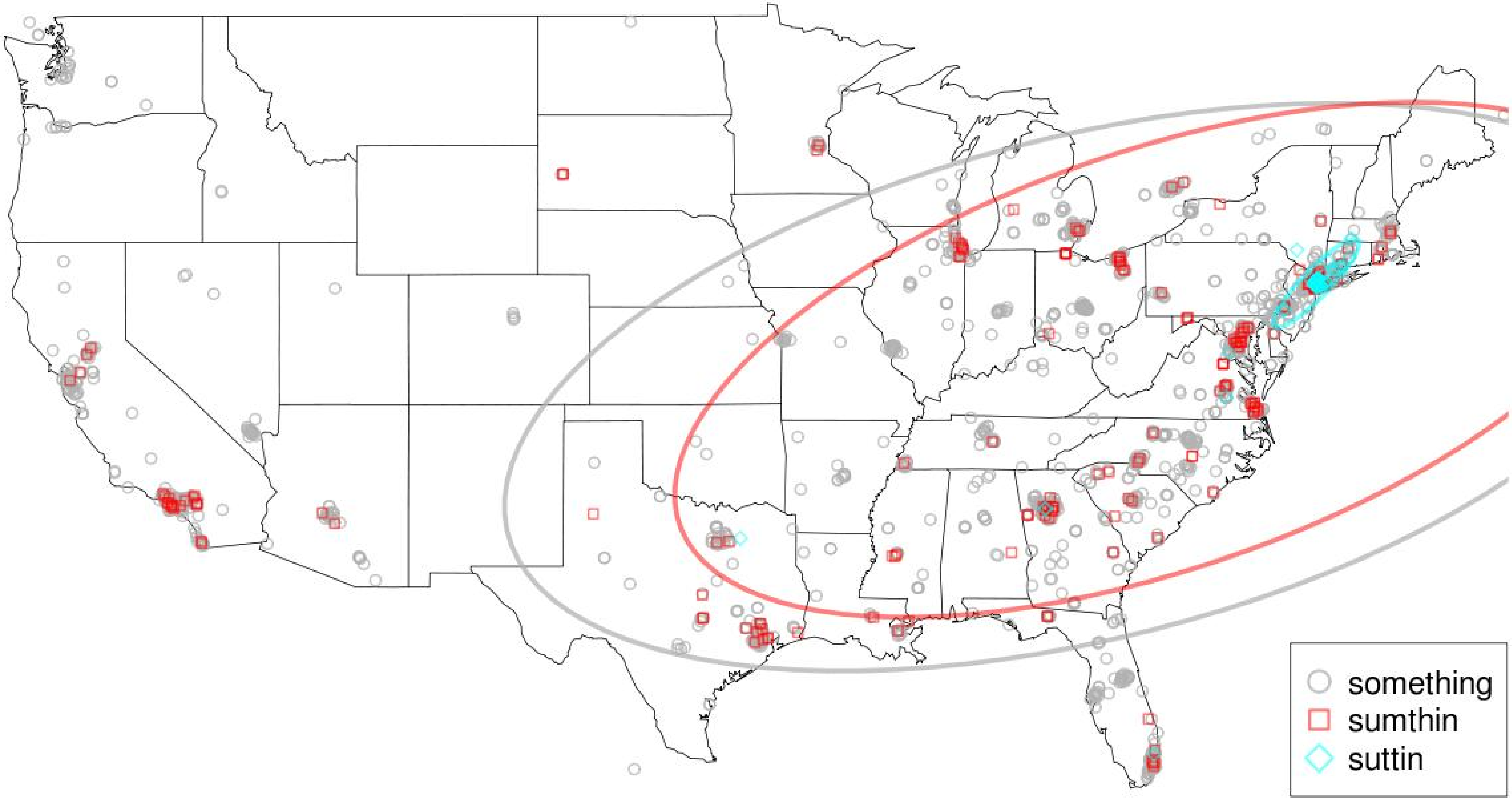
    Minimize divergence of
    P(Neighbors | Word)

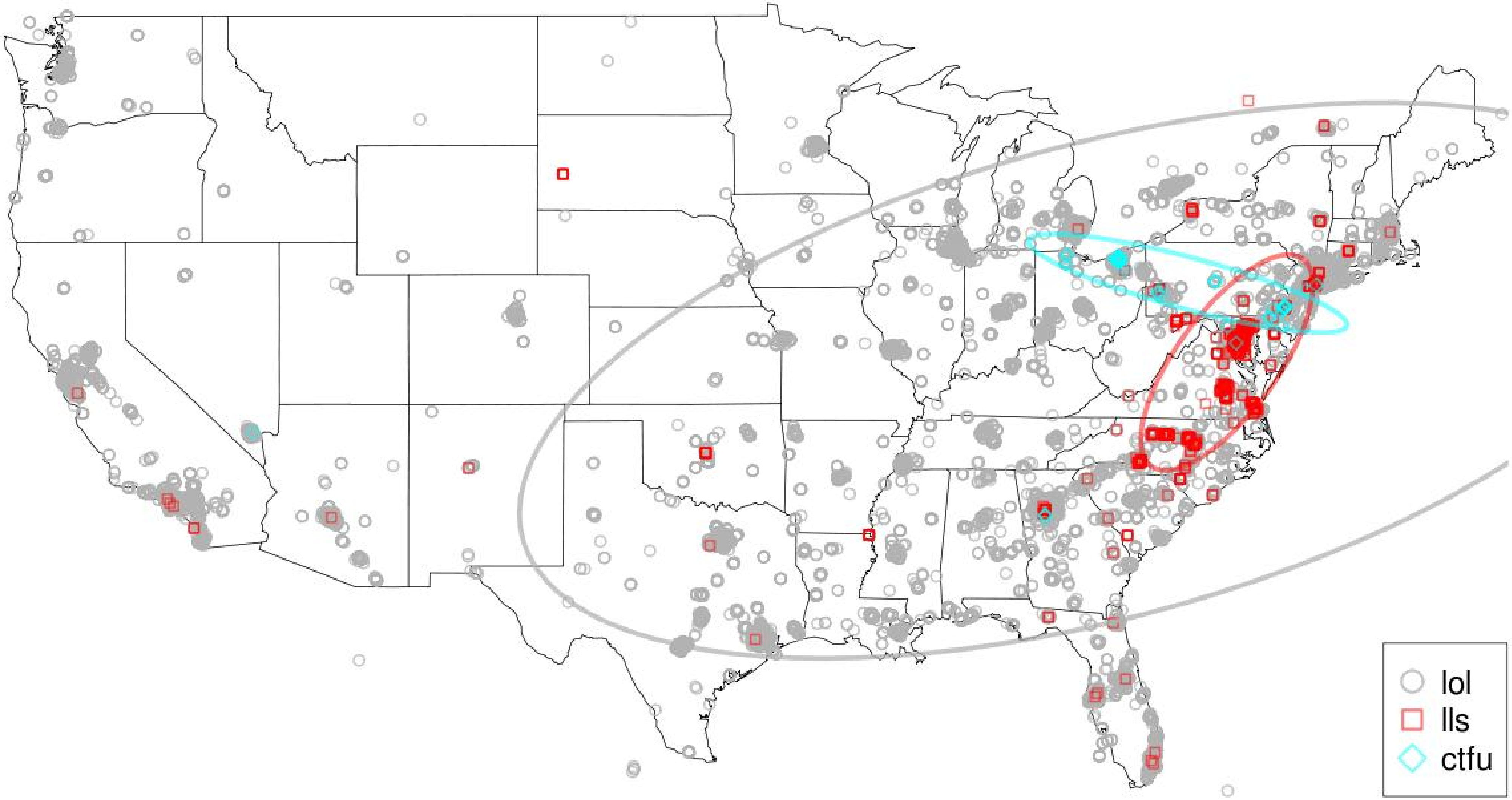
Examples


Los Angeles
Houston
Chicago
New York

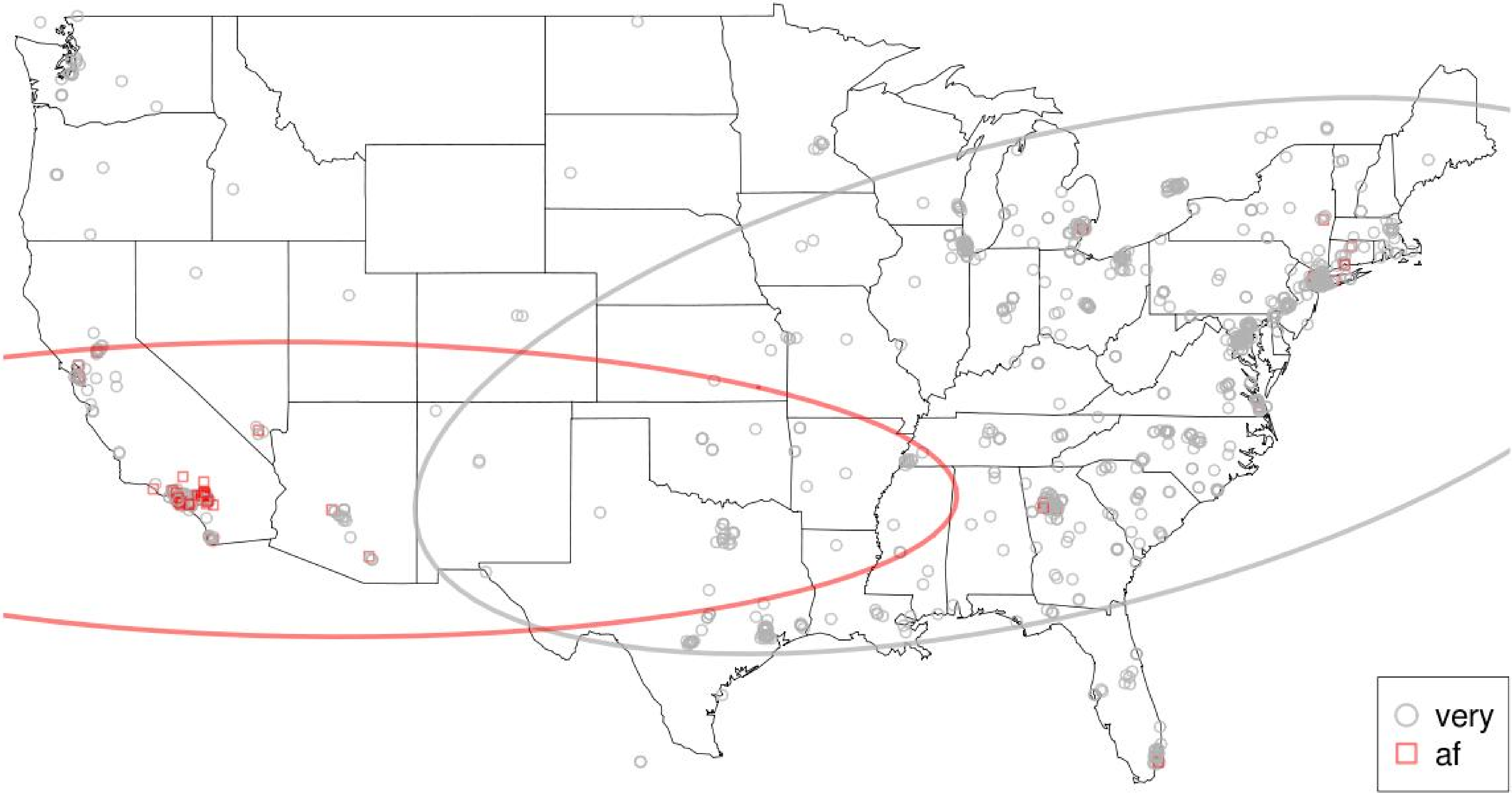


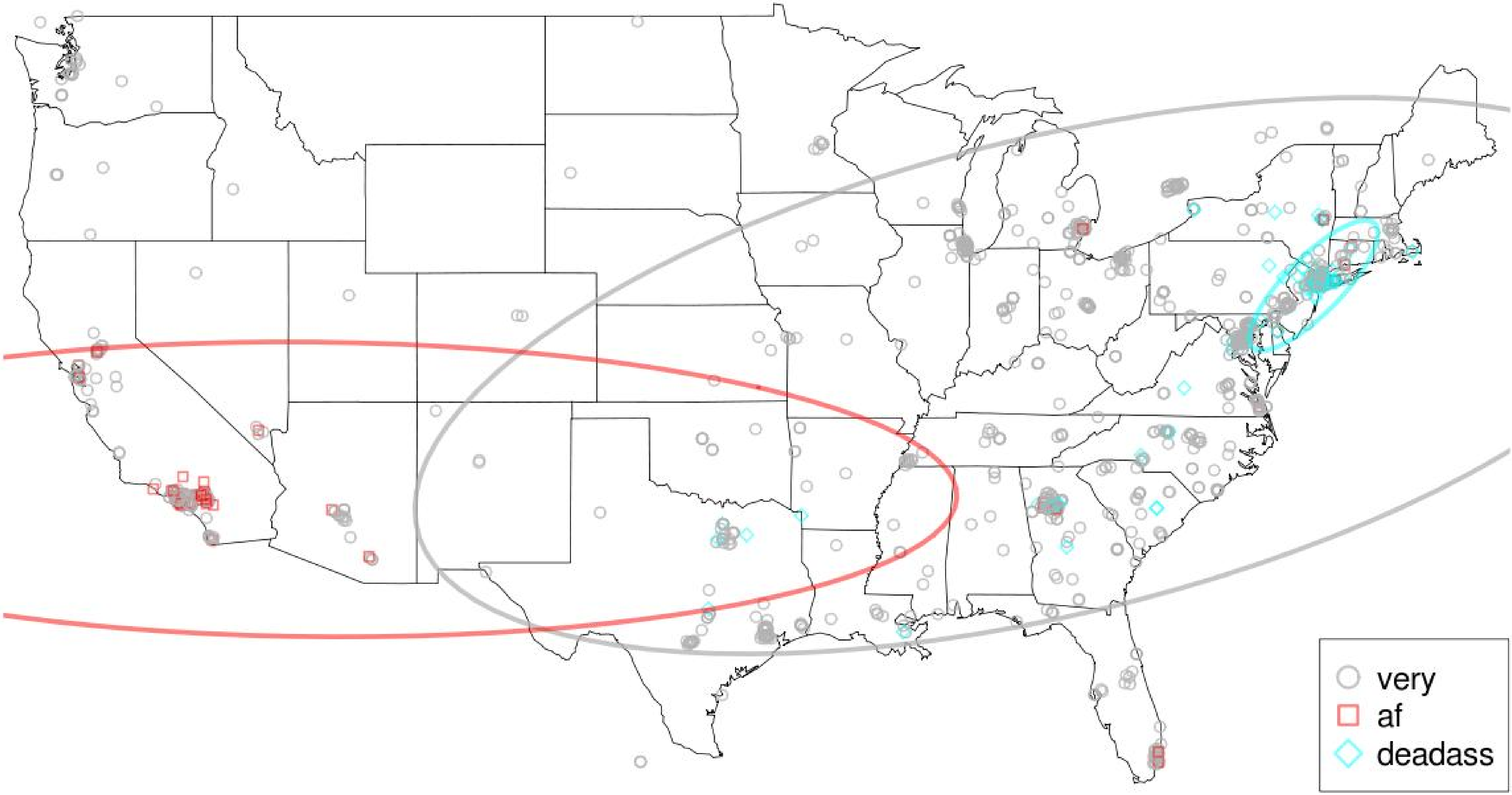
you
yu
uu

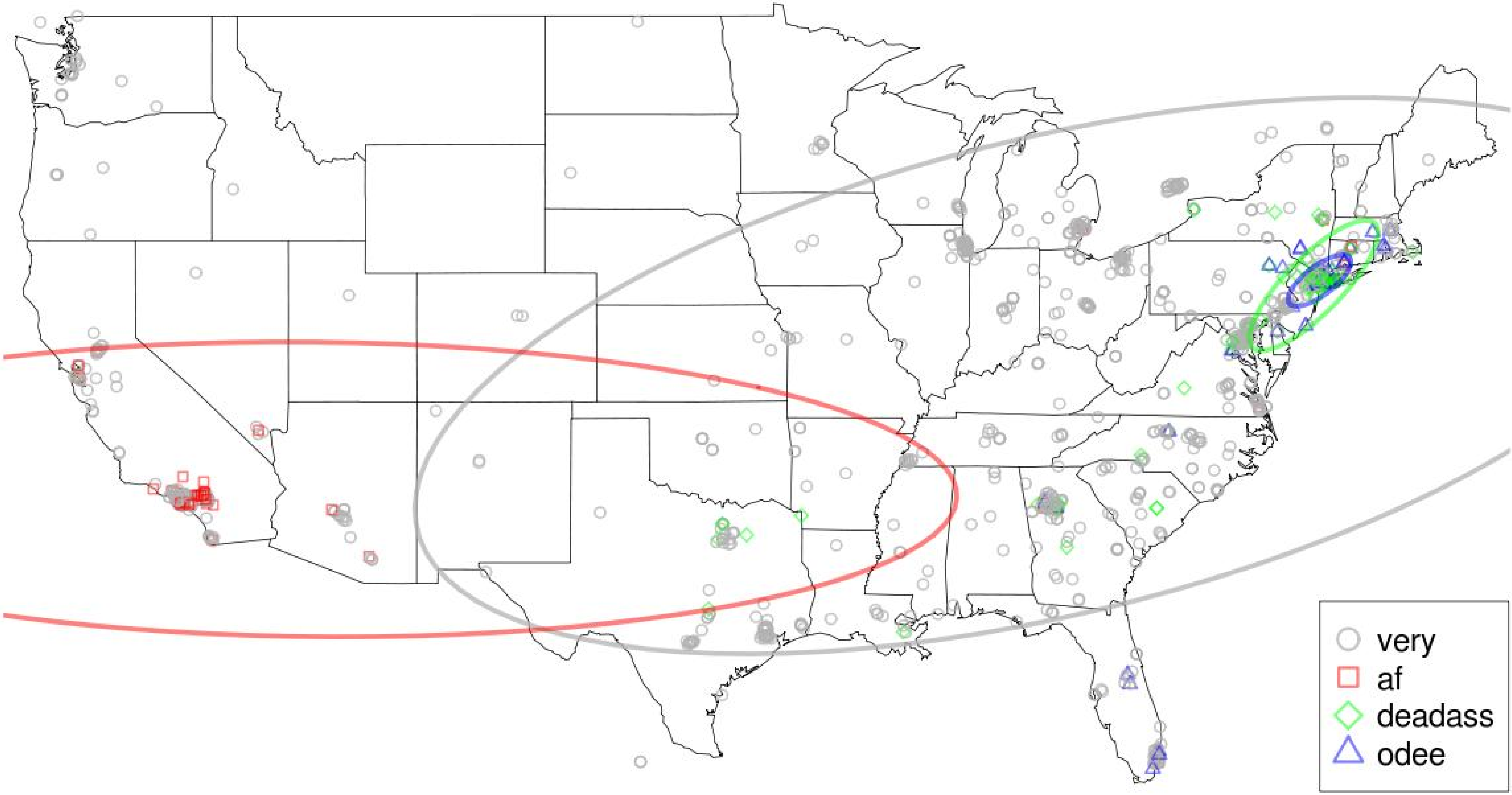

something
sumthin
suttin

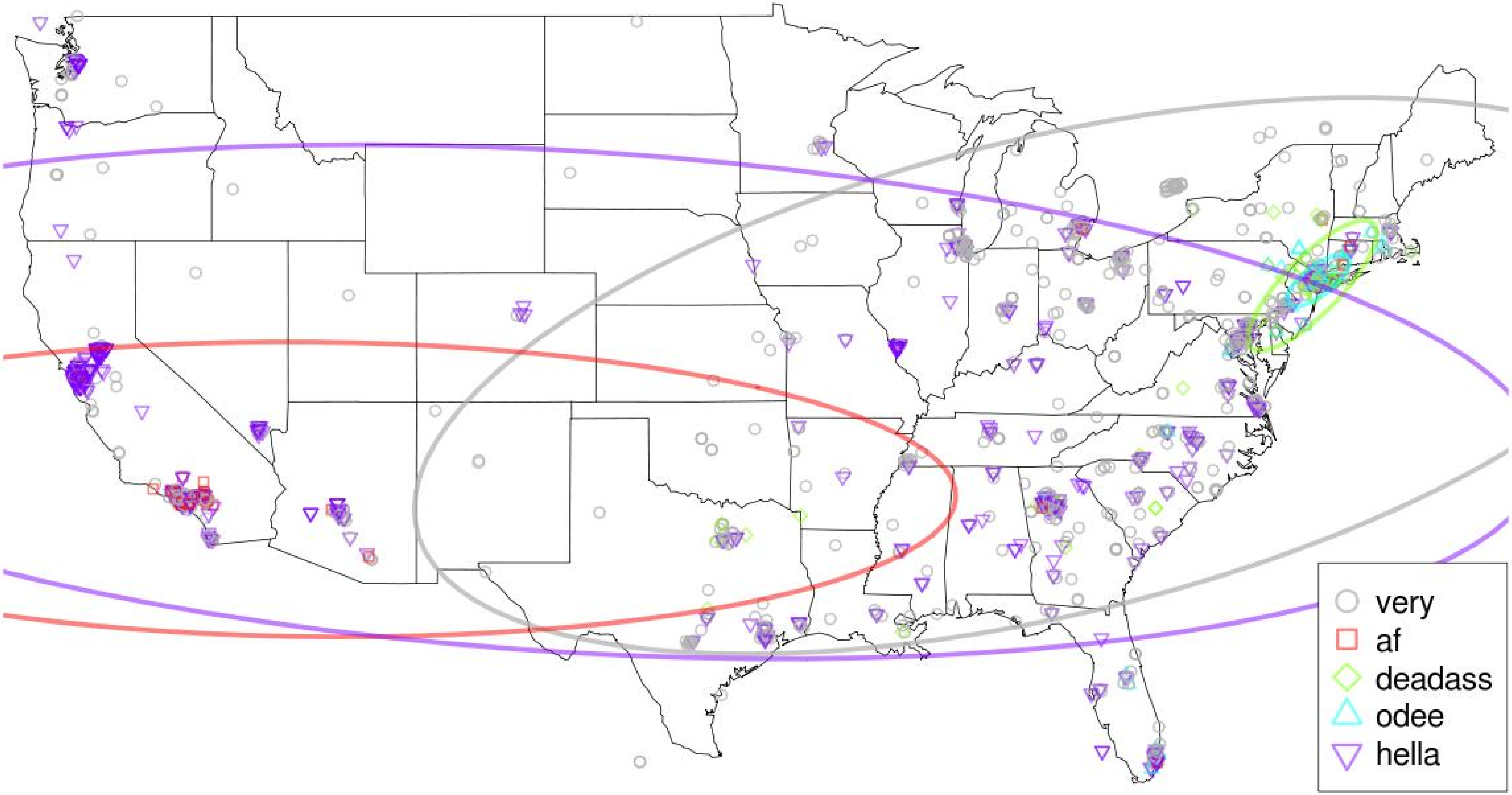

lol
lls
ctfu


very


very
af


very
af
deadass


very
af
deadass
odee


very
af
deadass
odee
hella

Summary (1)

- We can mine raw text to learn about lexical variation:
  - Discover geographic language communities and geographically-coherent sets of terms
  - Disentangle geographical and topical variation
  - Predict author location from text alone

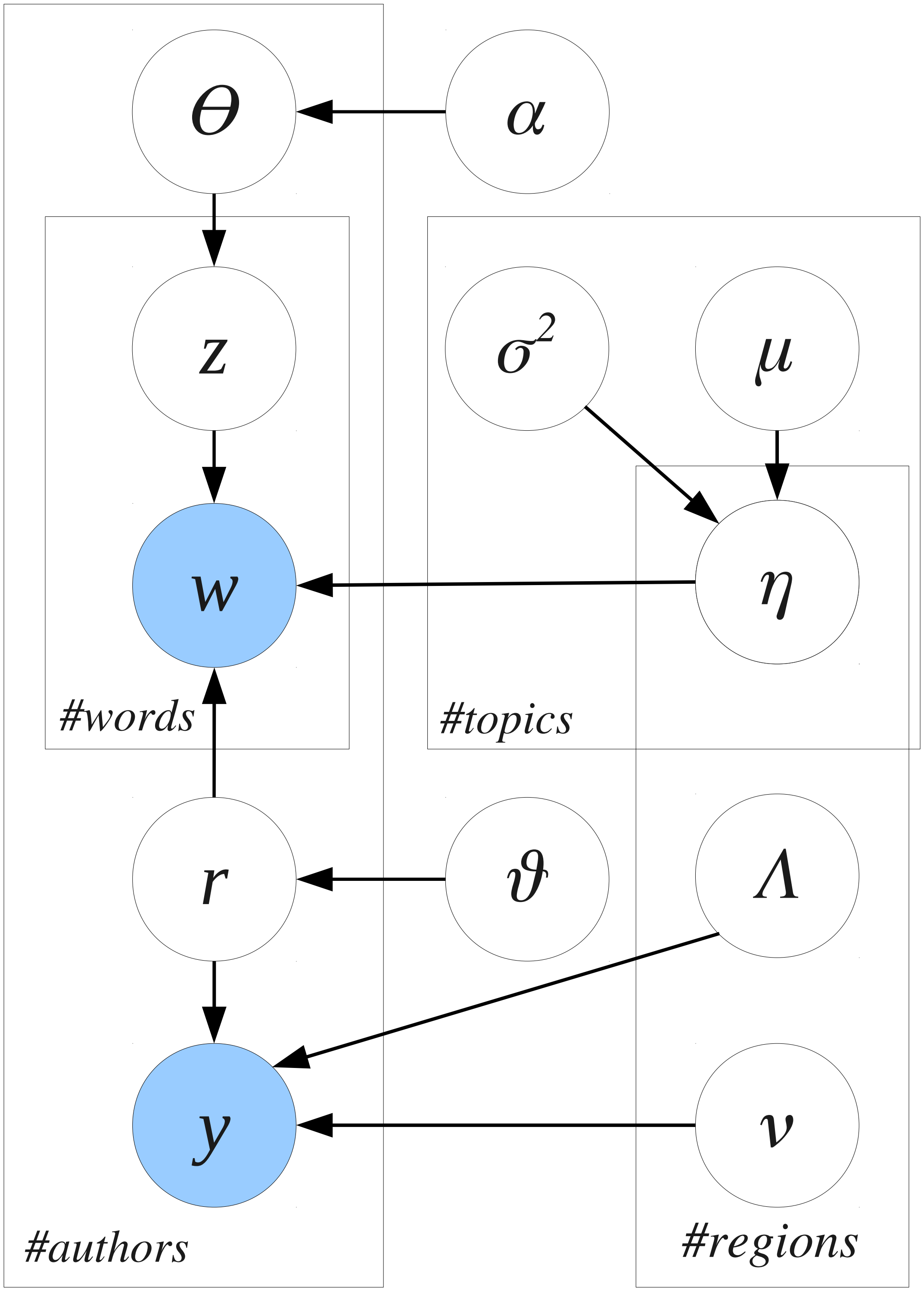
http://www.ark.cs.cmu.edu/GeoText

Summary (2)

- Social media text contains a variety of lexical dialect markers
  - Some are known to relate to speech: e.g., hella
  - Others appear to be unique to computer-mediated communication: coo/koo, lmao/ctfu, you/u/uu, …
  - **Future work: systematic analysis of the relationship between dialect in spoken language and social media text**

Thx!! R uu gna ask me suttin?

Adding topics



For each author

Pick a region from $P(r \mid \vartheta)$

Pick a location from $P(y \mid \Lambda_r, \nu_r)$

Pick a distribution over topics from $P(\Theta \mid \alpha)$

For each token

Pick a topic from $P(z \mid \Theta)$

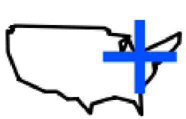
Pick a word from $\boldsymbol{P(w \mid \eta_{r,z})}$

Results

METHOD	MEAN ERROR (KM)	MEDIAN ERROR (KM)
Mean location	1148	1018
K-nearest neighbors	1077	853
Text regression	948	712
Supervised LDA	1055	728
Mixture of unigrams	947	644
Geographic Topic Model	***900***	***494***

Wilcoxon-Mann-Whitney: $p < .01$

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