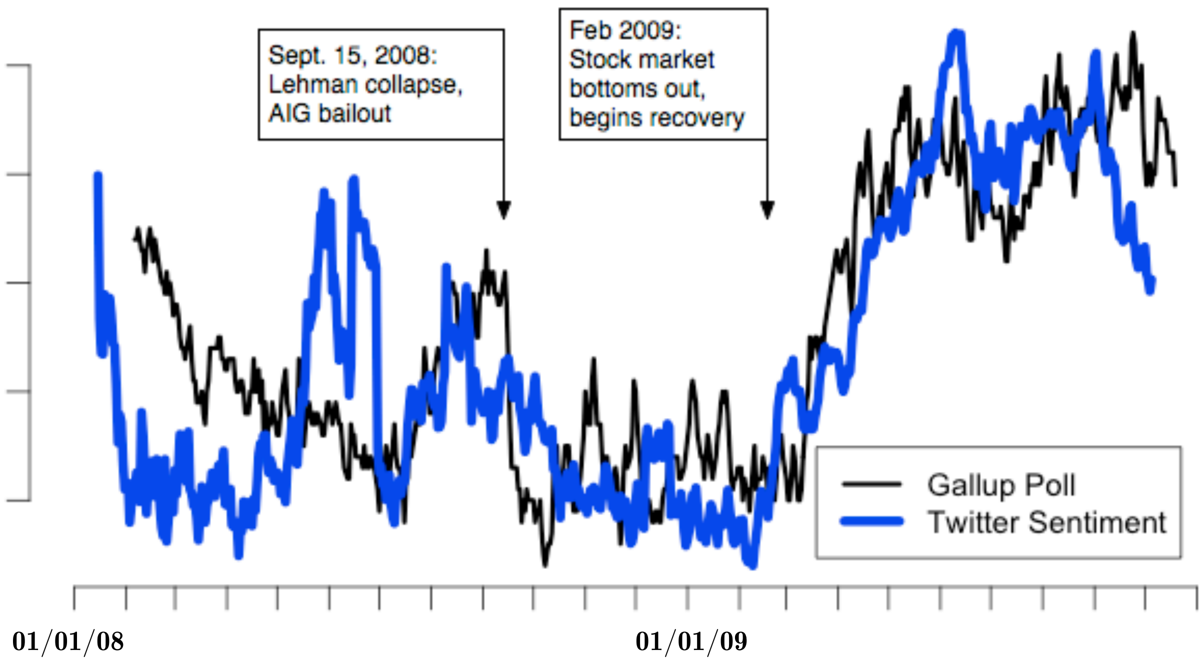
Using Text to Predict the Real World #textworld

Noah Smith*
School of Computer Science
Carnegie Mellon University
nasmith@cs.cmu.edu
@nlpnoah

Philip Resnik
Department of Linguistics, UMIACS
University of Maryland
resnik@umd.edu

***Joint work with Ramnath Balasubramanyan, Dipanjan Das, Jacob Eisenstein, Kevin Gimpel, Mahesh Joshi, Shimon Kogan, Dimitry Levin, Brendan O'Connor, Bryan Routledge, Jacob Sagi, Eric Xing.**

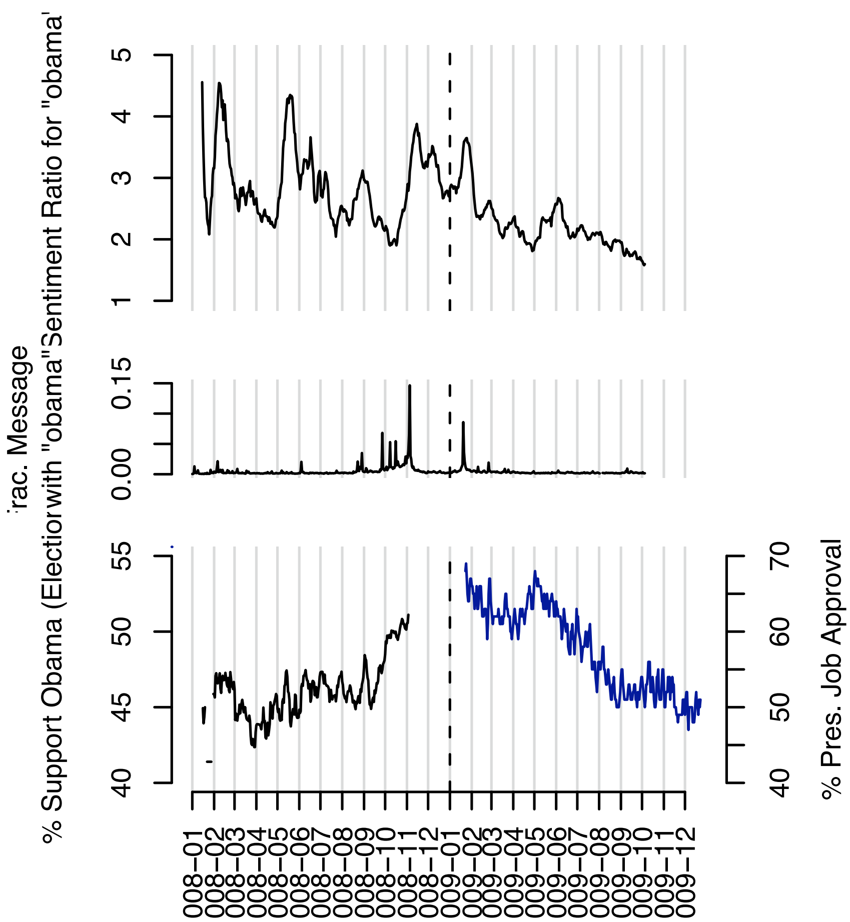
jobs on Twitter



$$r = 0.794$$

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; Smith, N. A. 2010. From tweets to polls: linking text sentiment to public opinion time series. *Proc. ICWSM* pp. 122-129.

obama on Twitter



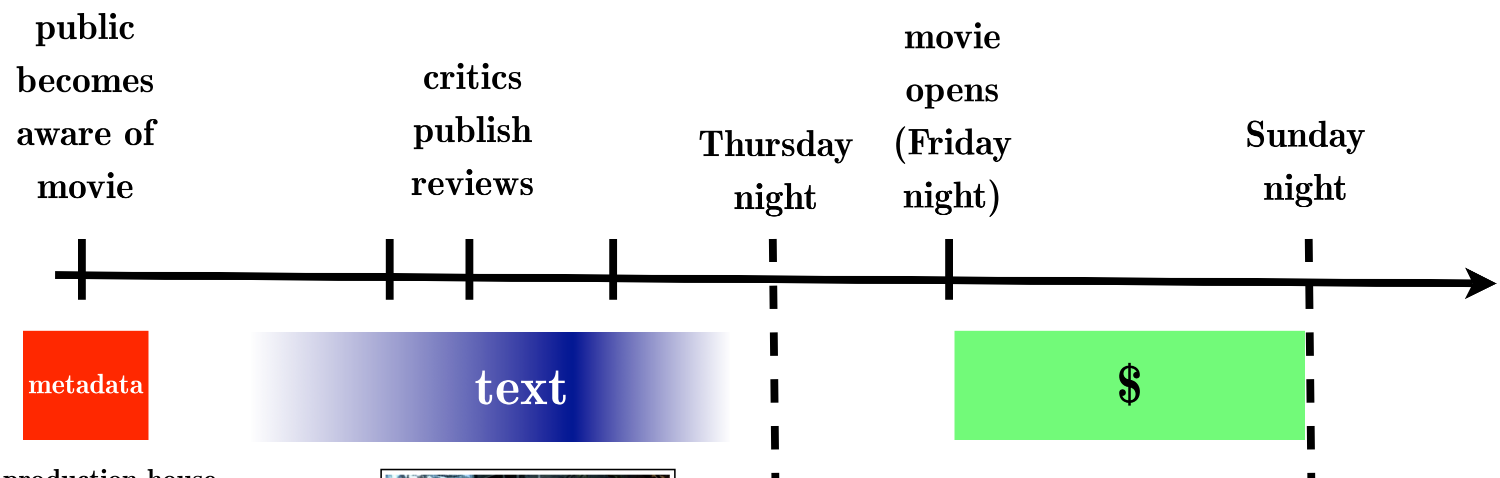
$r = 0.725$
(approval)

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; Smith, N. A. 2010. From tweets to polls: linking text sentiment to public opinion time series. *Proc. ICWSM* pp. 122-129.

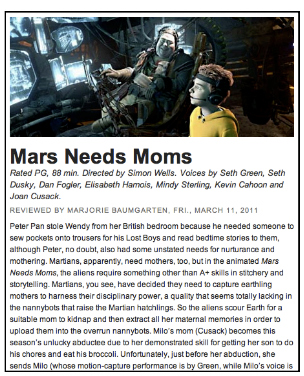
Conjecture

Text,
written by everyday people
in large volumes,
or by specialized experts,
can tell us about the social world.

An Example: Movie Reviews & Revenue



production house, genre(s), scriptwriter(s), director(s), country of origin, primary actors, release date, MPAA rating, running time, production budget (Simonoff & Sparrow, 2000; Sharda & Delen, 2006)



Mars Needs Moms

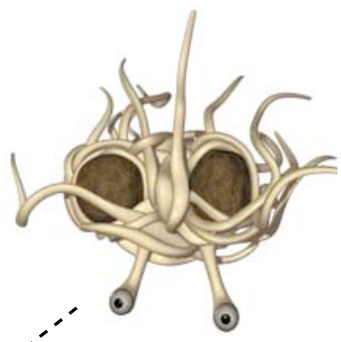
Rated PG, 88 min. Directed by Simon Wells. Voices by Seth Green, Seth Dusky, Dan Fogler, Elisabeth Harnois, Mindy Sterling, Kevin Cahoon and Joan Cusack.

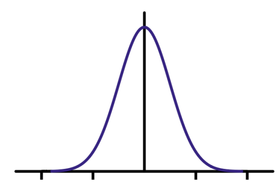
REVIEWED BY MARJORIE BAUMGARTEN, FRI., MARCH 11, 2011

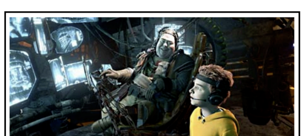
Peter Pan stole Wendy from her British bedroom because he needed someone to sew pockets onto trousers for his Lost Boys and read bedtime stories to them, although Peter, no doubt, also had some unstated needs for nurturance and mothering. Martians, apparently, need mothers, too, but in the animated *Mars Needs Moms*, the aliens require something other than A+ skills in stitchery and storytelling. Martians, you see, have decided they need to capture earthling mothers to harness their disciplinary power, a quality that seems totally lacking in the nannybots that raise the Martian hatchlings. So the aliens scour Earth for a suitable mom to kidnap and then extract all her maternal memories in order to upload them into the overrun nannybots. Milo's mom (Cusack) becomes this season's unlucky abductee due to her demonstrated skill for getting her son to do his chores and eat his broccoli. Unfortunately, just before her abduction, she sends Milo (whose motion-capture performance is by Green, while Milo's voice is

Joshi, M.; Das, D.; Gimpel, K.; Smith, N. A. 2010. Movie reviews and revenues: an experiment in text regression. *Proc. NAACL* pp. 293-296.

Model







Mars Needs Moms

Rated PG, 88 min. Directed by Simon Wells. Voices by Seth Green, Seth Dusky, Dan Fogler, Elisabeth Harnois, Mindy Sterling, Kevin Cahoon and Joan Cusack.

REVIEWED BY MARJORIE BAUMGARTEN, FRI., MARCH 11, 2011

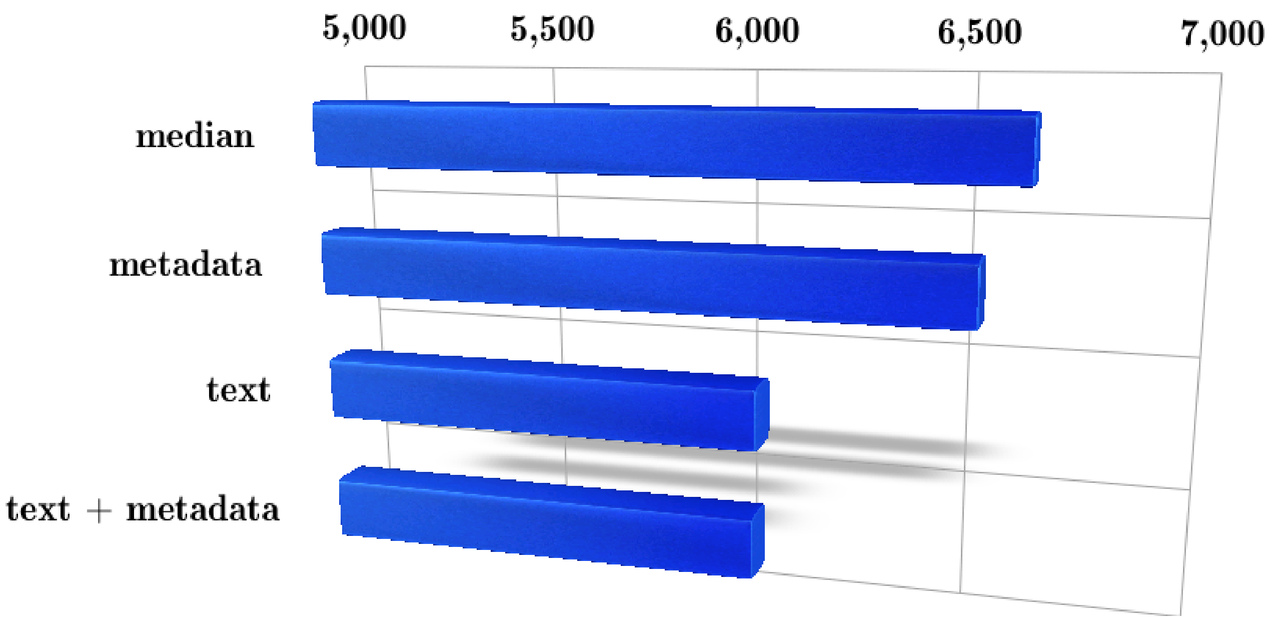
Peter Pan stole Wendy from her British bedroom because he needed someone to sew pockets onto trousers for his Lost Boys and read bedtime stories to them, although Peter, no doubt, also had some unstated needs for nurturance and mothering. Martians, apparently, need mothers, too, but in the animated *Mars Needs Moms*, the aliens require something other than A+ skills in stitchery and storytelling. Martians, you see, have decided they need to capture earthling mothers to harness their disciplinary power, a quality that seems totally lacking in the nannybots that raise the Martian hatchlings. So the aliens scour Earth for a suitable mom to kidnap and then extract all her maternal memories in order to upload them into the overrun nannybots. Milo's mom (Cusack) becomes this season's unlucky abductee due to her demonstrated skill for getting her son to do his chores and eat his broccoli. Unfortunately, just before her abduction, she sends Milo (whose motion-capture performance is by Green, while Milo's voice is

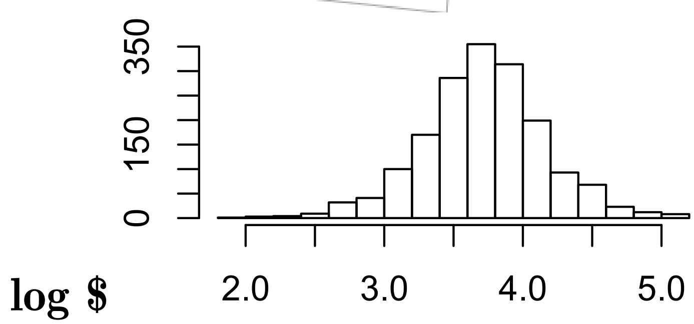
Experiment

- ✦ **1,718 films from 2005-9:**
  - **7,000 reviews (up to 7 reviews per movie)**
  - **Metadata from `metacritic.com` and `the-numbers.com`**
  - **Opening weekend gross and number of screens (`the-numbers.com`)**
- ✦ **Train the probabilistic model (elastic net linear regression) on movies from 2005-8.**
- ✦ **Evaluate on movies from 2009.**

- **Data available at `www.ark.cs.cmu.edu`**

Mean Absolute Error Per Screen ($)





Features ($M)

rating	pg	+0.085
	adult	-0.236
	rate r	-0.364
sequels	this series	+13.925
	the franchise	+5.112
	the sequel	+4.224
people	will smith	+2.560
	brittany	+1.128
	^ producer brian	+0.486

genre	testosterone	+1.945
	comedy for	+1.143
	a horror	+0.595
	documentary	-0.037
	independent	-0.127
sent.	best parts of	+1.462
	smart enough	+1.449
	a good thing	+1.117
	shame $	-0.098
	bogeyman	-0.689
plot	torso	+9.054
	vehicle in	+5.827
	superhero $	+2.020

***Also* ...** **of the art**, **and cgi**, **shrek movies**, **voldemort**, **blockbuster**, **anticipation**, **summer movie**; **cannes** **is bad.**

Discussion

✦ **Can we do it on Twitter?**

- **Yes! See Asur & Huberman (2010).**

✦ **Was that sentiment analysis?**

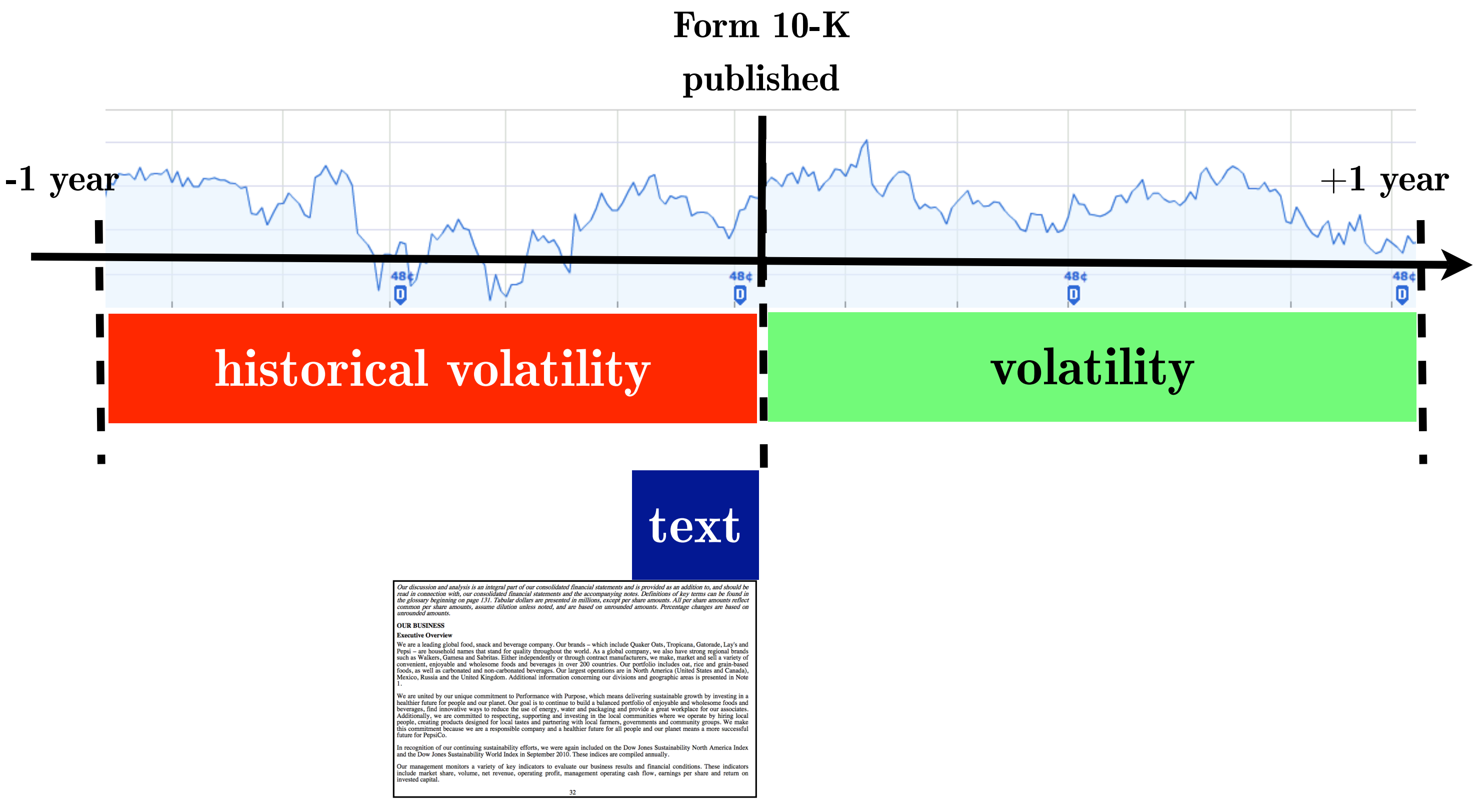
- **Sort of, but "sentiment" was measured in revenue.**
- **And standard linguistic preprocessing didn't really help us.**

Another Example: Financial Disclosures

- **The SEC mandates that publicly traded firms report to their shareholders.**
  - **Form 10-K, section 7: "Management's Discussion and Analysis," a disclosure about risk.**

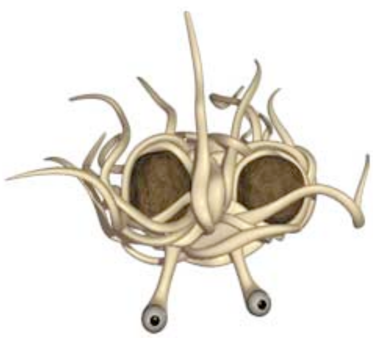
- **Does the text in an MD&A predict return volatility?**
  - **We're not predicting *returns*, which would require finding *new* information (hard).**

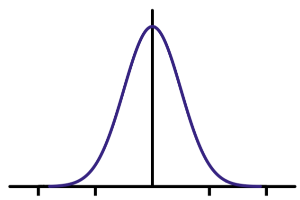
Disclosures and Volatility



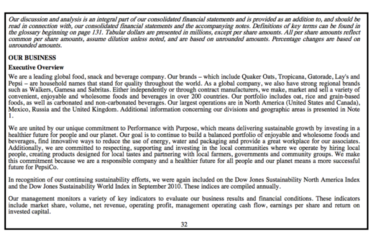
Kogan, S.; Levin, D.; Routledge, B. R.; Sagi, J. S.; Smith, N. A. 2009. Predicting risk from financial reports with regression. *Proc. NAACL* pp. 272-280.

Model





volatility



Our discussion and analysis is an integral part of our consolidated financial statements and is provided as an addition to, and should be read in connection with, our consolidated financial statements and the accompanying notes. Definitions of key terms can be found in the glossary beginning on page 131. Tabular dollars are presented in millions, except per share amounts. All per share amounts reflect common per share amounts, assume dilution unless noted, and are based on unrounded amounts. Percentage changes are based on unrounded amounts.

OUR BUSINESS

Executive Overview

We are a leading global food, snack and beverage company. Our brands – which include Quaker Oats, Tropicana, Gatorade, Lay's and Pepsi – are household names that stand for quality throughout the world. As a global company, we also have strong regional brands such as Walkers, Gamesa and Sabritas. Either independently or through contract manufacturers, we make, market and sell a variety of convenient, enjoyable and wholesome foods and beverages in over 200 countries. Our portfolio includes oat, rice and grain-based foods, as well as carbonated and non-carbonated beverages. Our largest operations are in North America (United States and Canada), Mexico, Russia and the United Kingdom. Additional information concerning our divisions and geographic areas is presented in Note 1.

We are united by our unique commitment to Performance with Purpose, which means delivering sustainable growth by investing in a healthier future for people and our planet. Our goal is to continue to build a balanced portfolio of enjoyable and wholesome foods and beverages, find innovative ways to reduce the use of energy, water and packaging and provide a great workplace for our associates. Additionally, we are committed to respecting, supporting and investing in the local communities where we operate by hiring local people, creating products designed for local tastes and partnering with local farmers, governments and community groups. We make this commitment because we are a responsible company and a healthier future for all people and our planet means a more successful future for PepsiCo.

In recognition of our continuing sustainability efforts, we were again included on the Dow Jones Sustainability North America Index and the Dow Jones Sustainability World Index in September 2010. These indices are compiled annually.

Our management monitors a variety of key indicators to evaluate our business results and financial conditions. These indicators include market share, volume, net revenue, operating profit, management operating cash flow, earnings per share and return on invested capital.

32

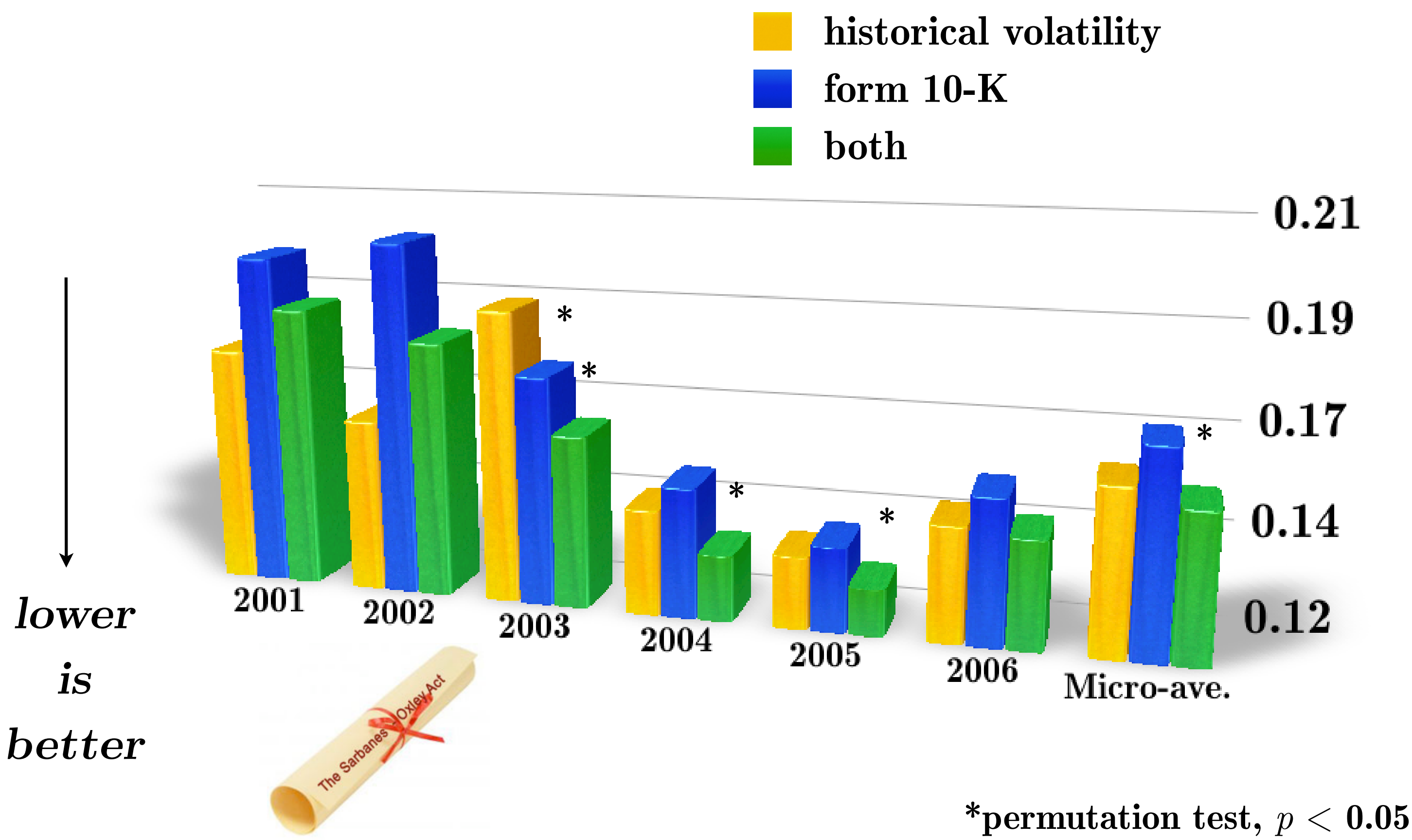
Data

✦ **26,806 10-K reports from 1996-2006 (`sec.gov`)**

- **Section 7 automatically extracted (noisy)**
- **Volatility in the previous year and the following year (Center for Research in Security Prices: U.S. Stocks Databases)**

✦ **Data available at `www.ark.cs.cmu.edu`**

MSE of Log-Volatility



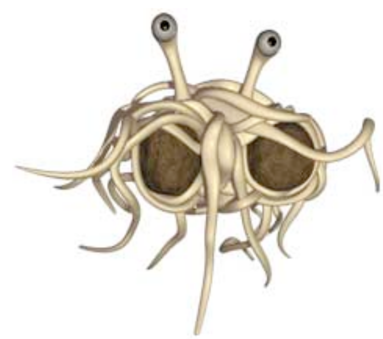
Dominant Weights (2000-4)

loss	0.025	net income	-0.021
net loss	0.017	rate	-0.017
year #	0.016	properties	-0.014
expenses	0.015	dividends	-0.013
going concern	0.014	lower interest	-0.012
a going	0.013	critical accounting	-0.012
administrative	0.013	insurance	-0.011
personnel	0.013	distributions	-0.011

high volatility terms **low volatility terms**

More Examples

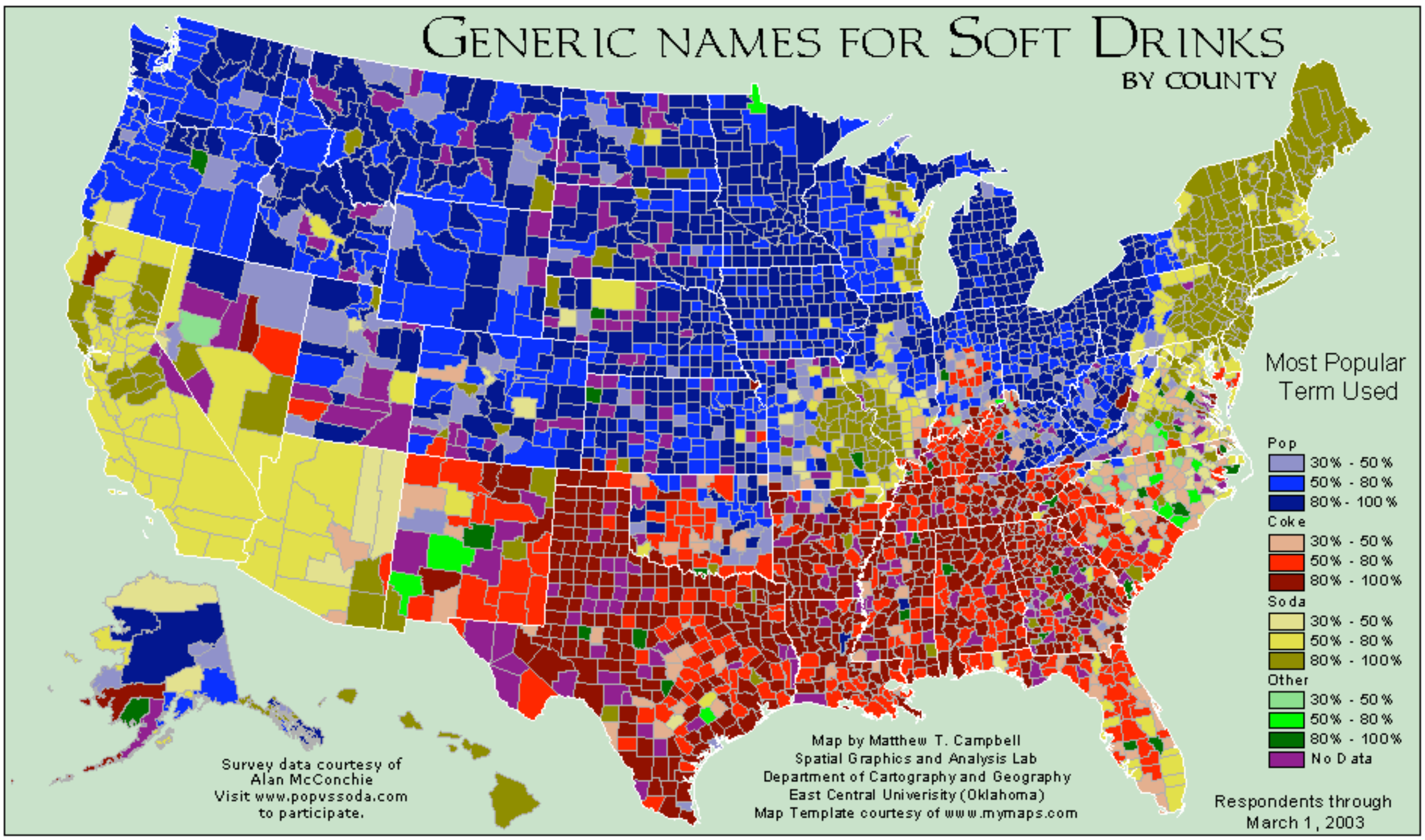
- ✦ **Will a political blog post attract a high volume of comments?**
- ✦ **Will a piece of legislation get a long debate, a partisan vote, success?**
- ✦ **Will a scientific article be heavily downloaded, cited?**



A Different Kind of Prediction

- ✦ So far, we've looked at what people have written, and made predictions about future measurements.
- ✦ Next, we'll consider how text reveals context.

Language Variation



Quantitative Study of Language Variation

✦ **Strong tradition:**

- **dialectology (Labov et al., 2006)**
- **sociolinguistics (Labov, 1966; Tagliamonte, 2006)**

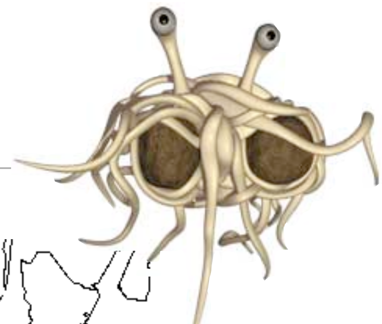
Data

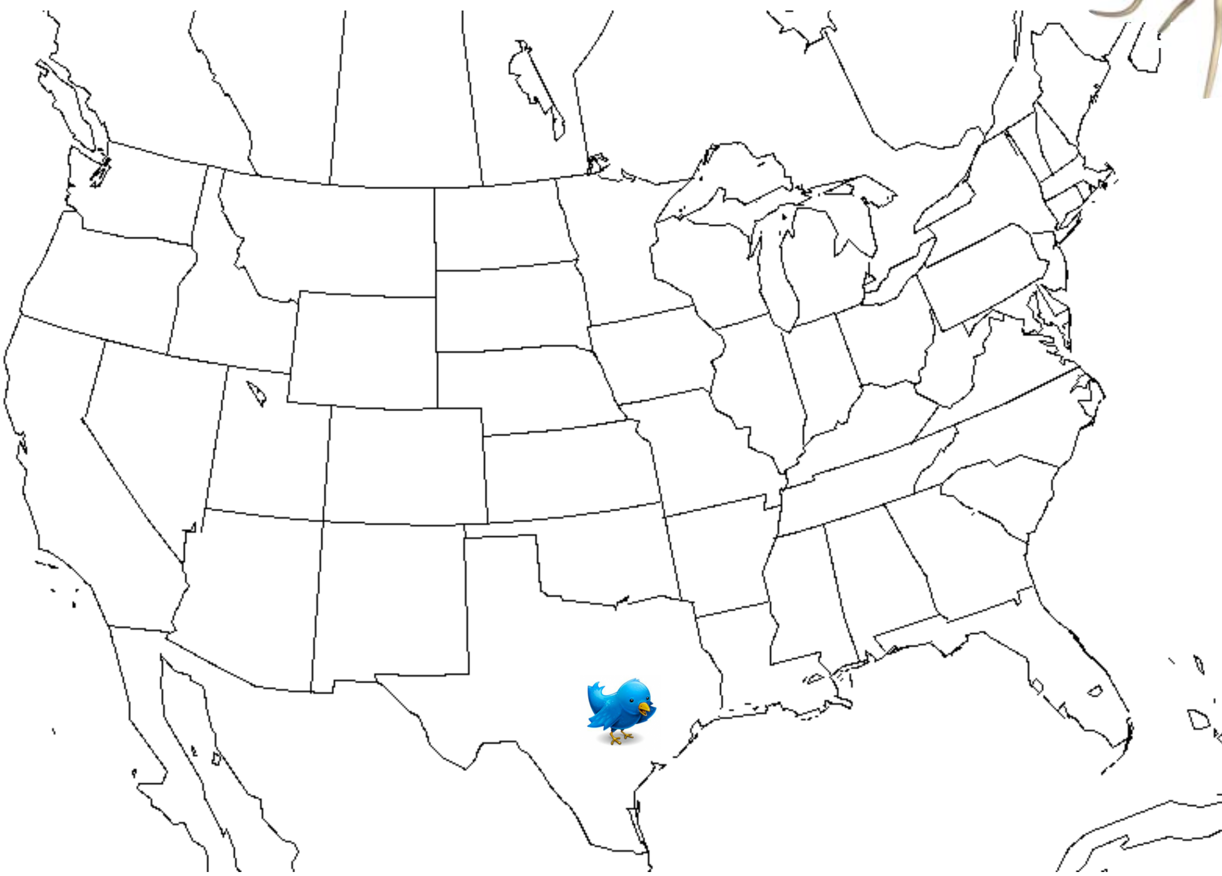
- **380,000 geo-tagged tweets from one week in March 2010**
  - **9,500 authors in (roughly) the United States**
  - **Informal: 25% of the most common words are not in standard dictionaries**
  - **Conversational: more than 50% of messages mention another user**

- **Data available at `www.ark.cs.cmu.edu`**

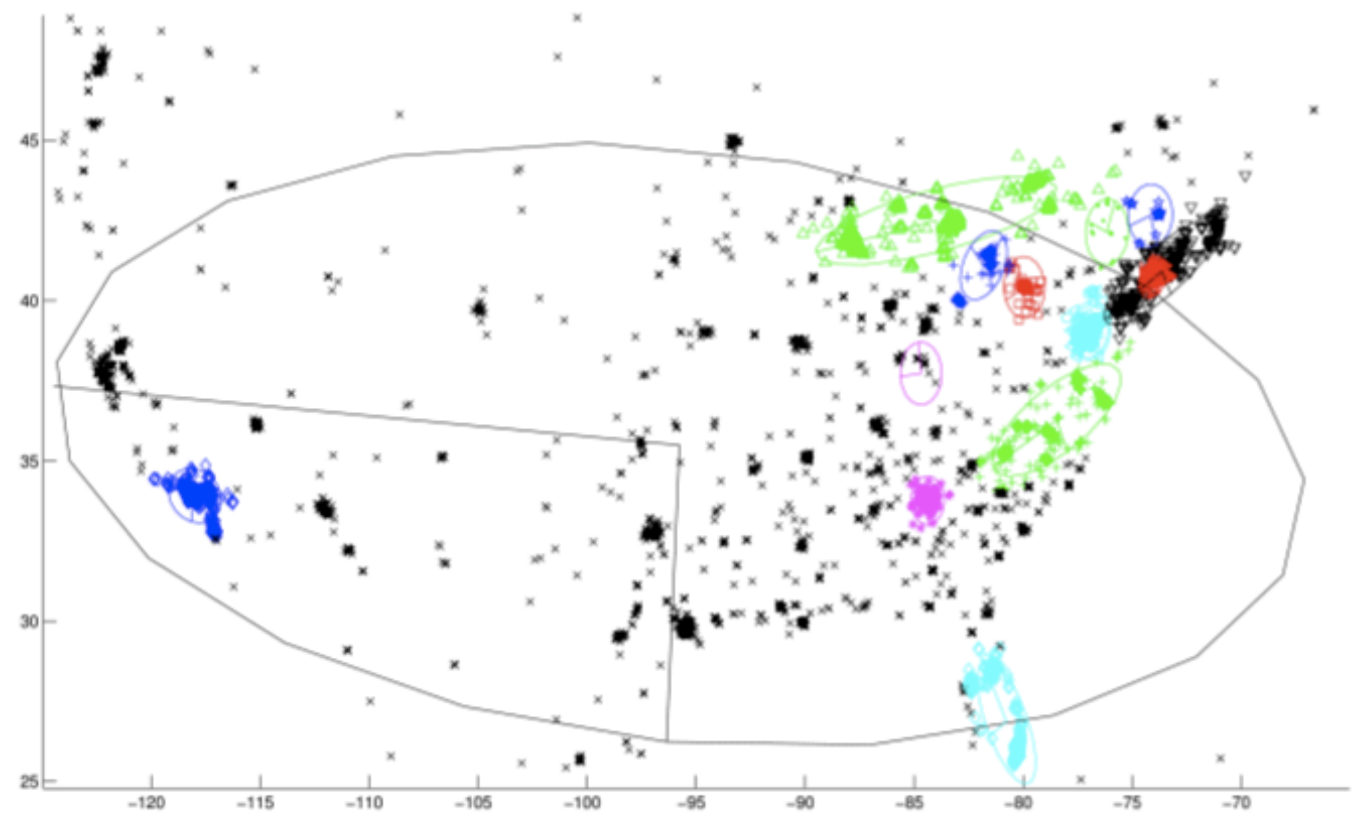
Eisenstein, J.; O'Connor, B.; Smith, N. A.; Xing, E. P. 2010. A latent variable model for geographic lexical variation. *Proc. EMNLP* pp. 1277-1287.

Model (Part 1)

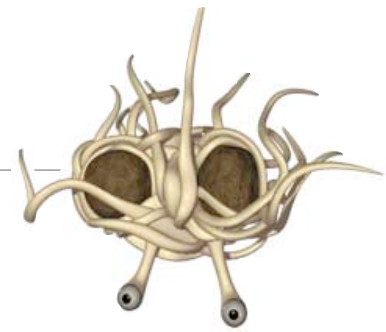




Gaussian Mixtures over Tweet Locations



Model (Part 2)



- ✦ What will you talk about (topics)?
- ✦ Pick words on those topic.
- ✦ Tweet.

Model

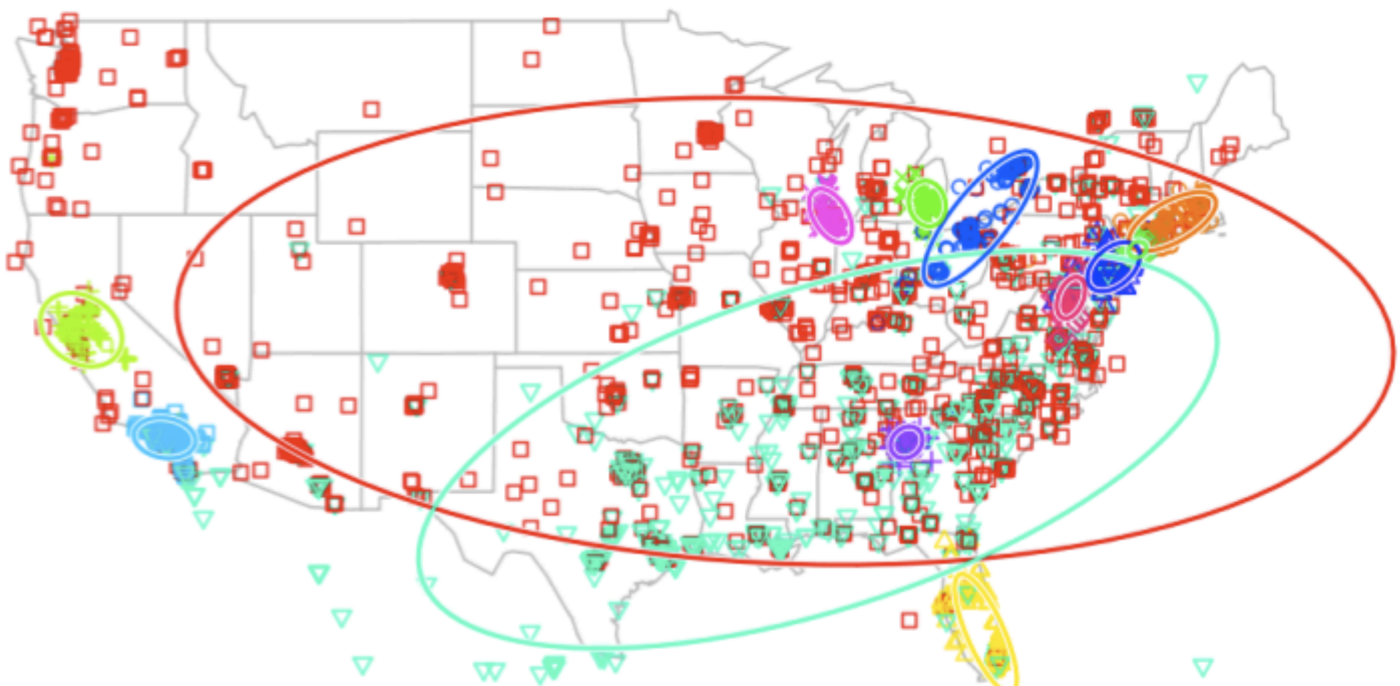
✦ **We can combine the two FSM myths:**

- **Generate location and text.**
- **Each topic gets *corrupted* in each region.**

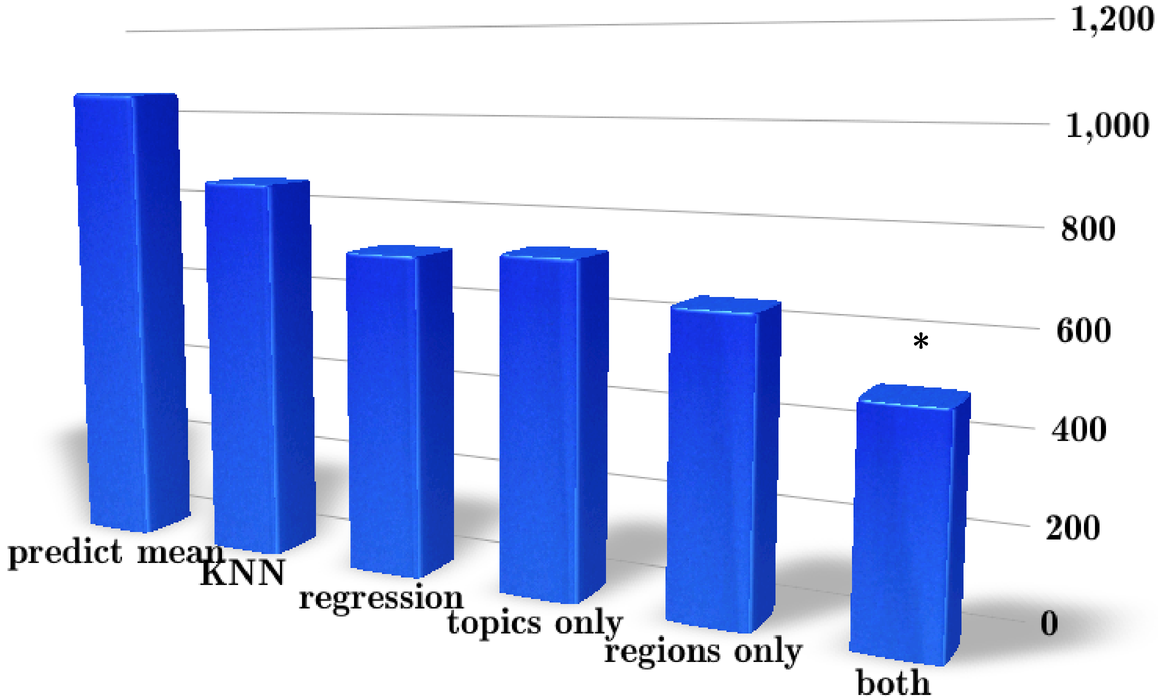
Topic: Food

dinner
delicious
snack
tasty

Regions from Text Content



Location Prediction (Error in km)

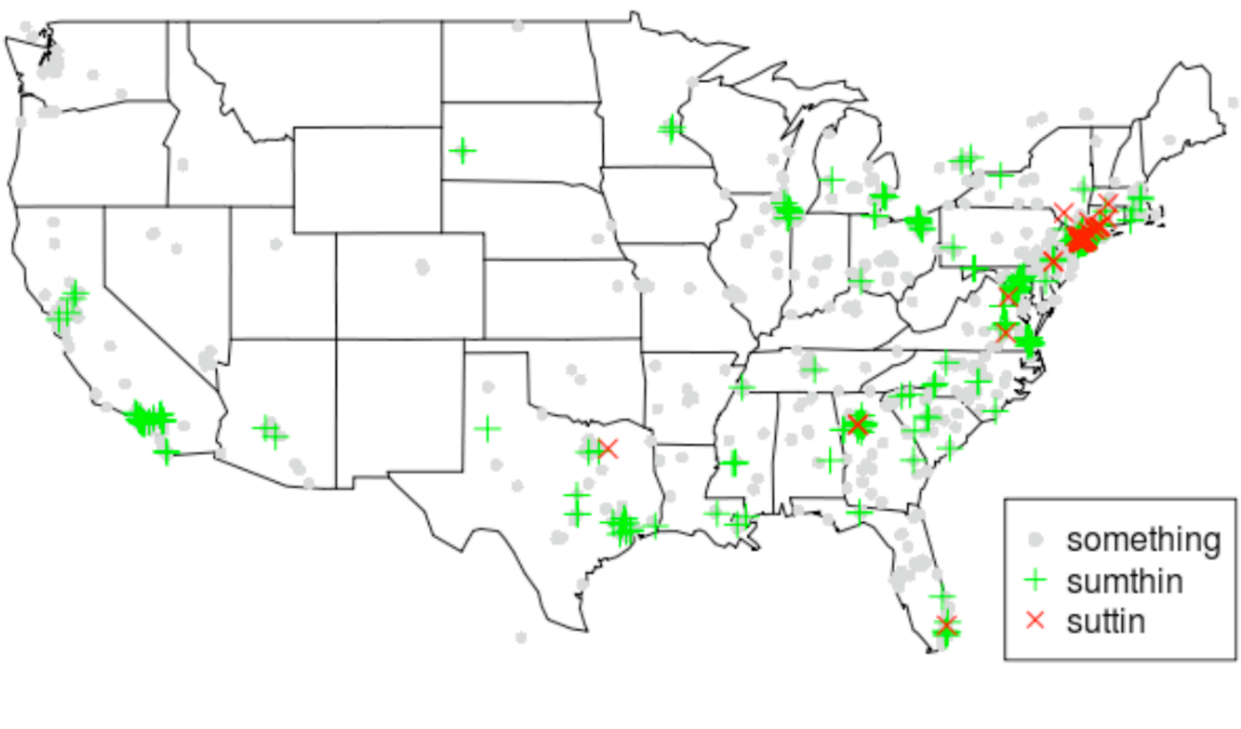


*Wilcoxon-Mann-Whitney, $p < 0.01$

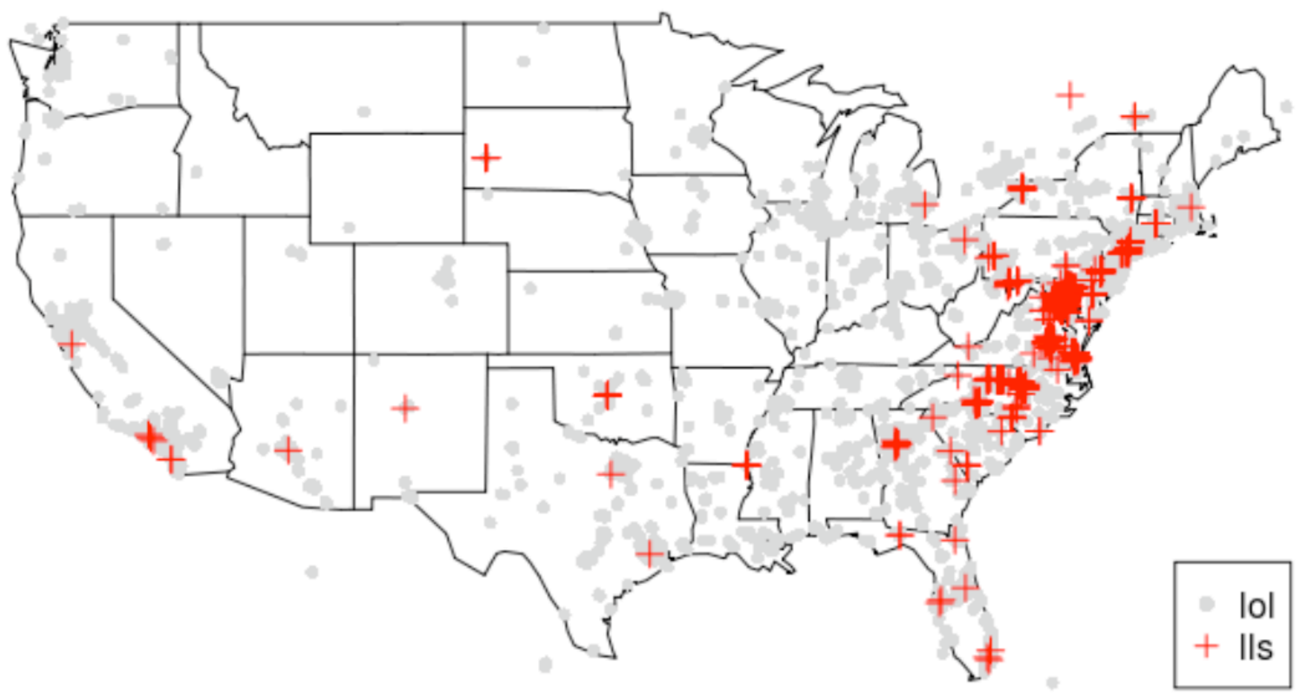
Qualitative Results

- Geographically-linked proper names are in the right places
  boston, knicks, bieber
- Some words reflect local prominence
  tacos, cab
- Geographically distinctive slang
  hella (Bucholtz et al., 2007), fasho, coo/koo, ;p
- Spanish words in regions with more Spanish speakers
  ese, pues, papi, nada

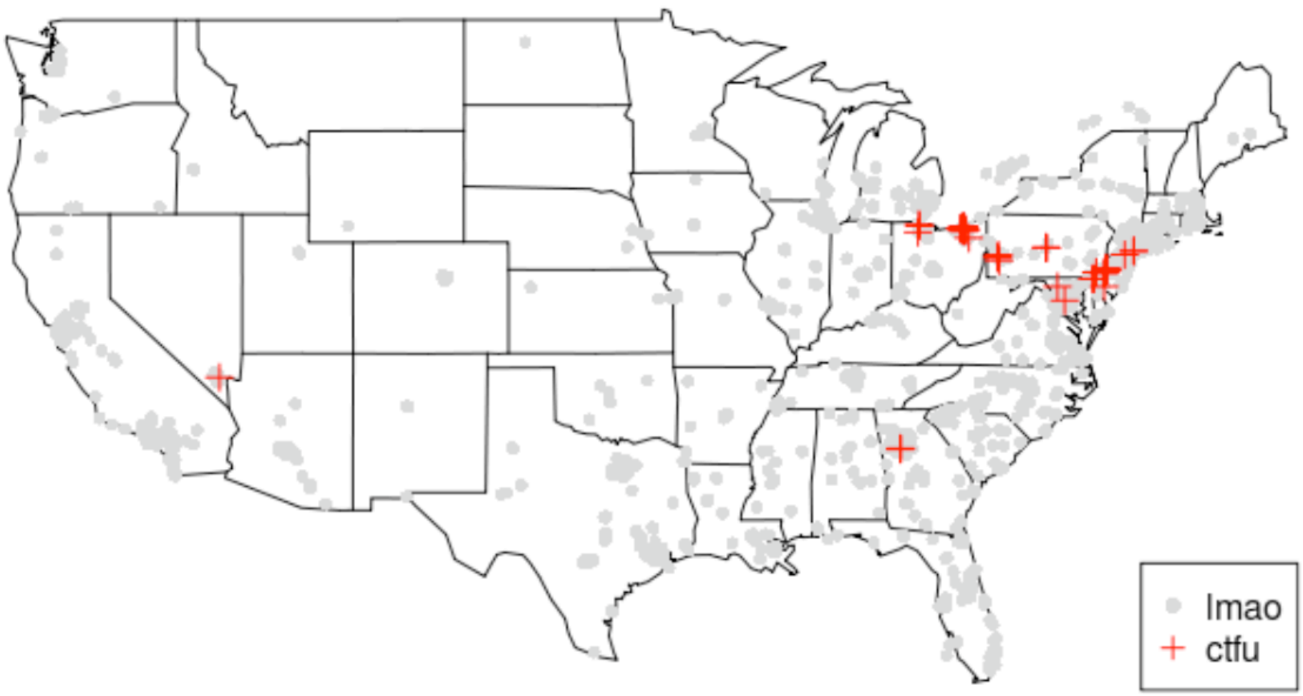
something/sumthin/suttin



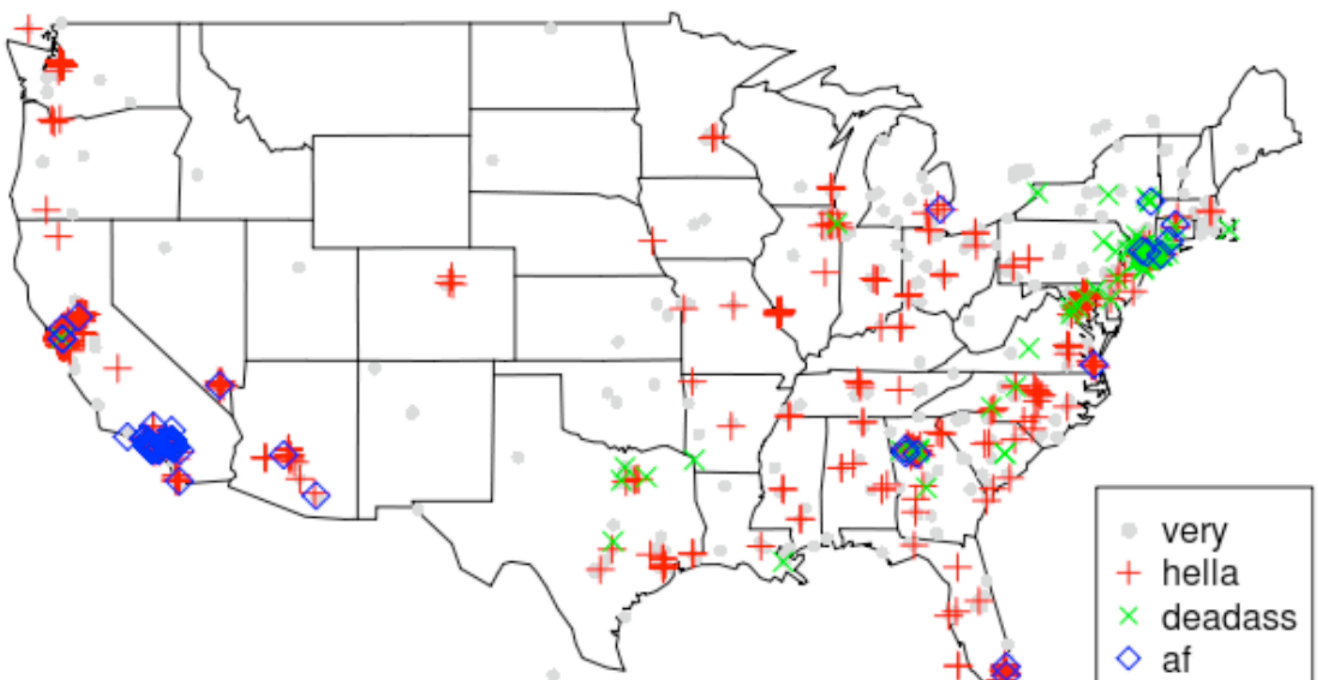
lol/lls



lmao/ctfu



Intensifiers



Ongoing Work

- From location to demographics*
- Languages other than American Twitter English
- Language change over time

*Eisenstein, J.; Smith, N. A.; Xing, E. P. 2011. Discovering sociolinguistic associations with structured sparsity. *Proc. ACL* (to appear).

Key Messages

✦ **Text is data.**

- **It carries useful information about the social world.**
- **Models based on text can "talk to us."**
- **We are just beginning to figure out how to extract quantitative, social information from text data.**

✦ **If you want to study/exploit language, look at the data.**

- **Statistical modeling is a powerful tool.**