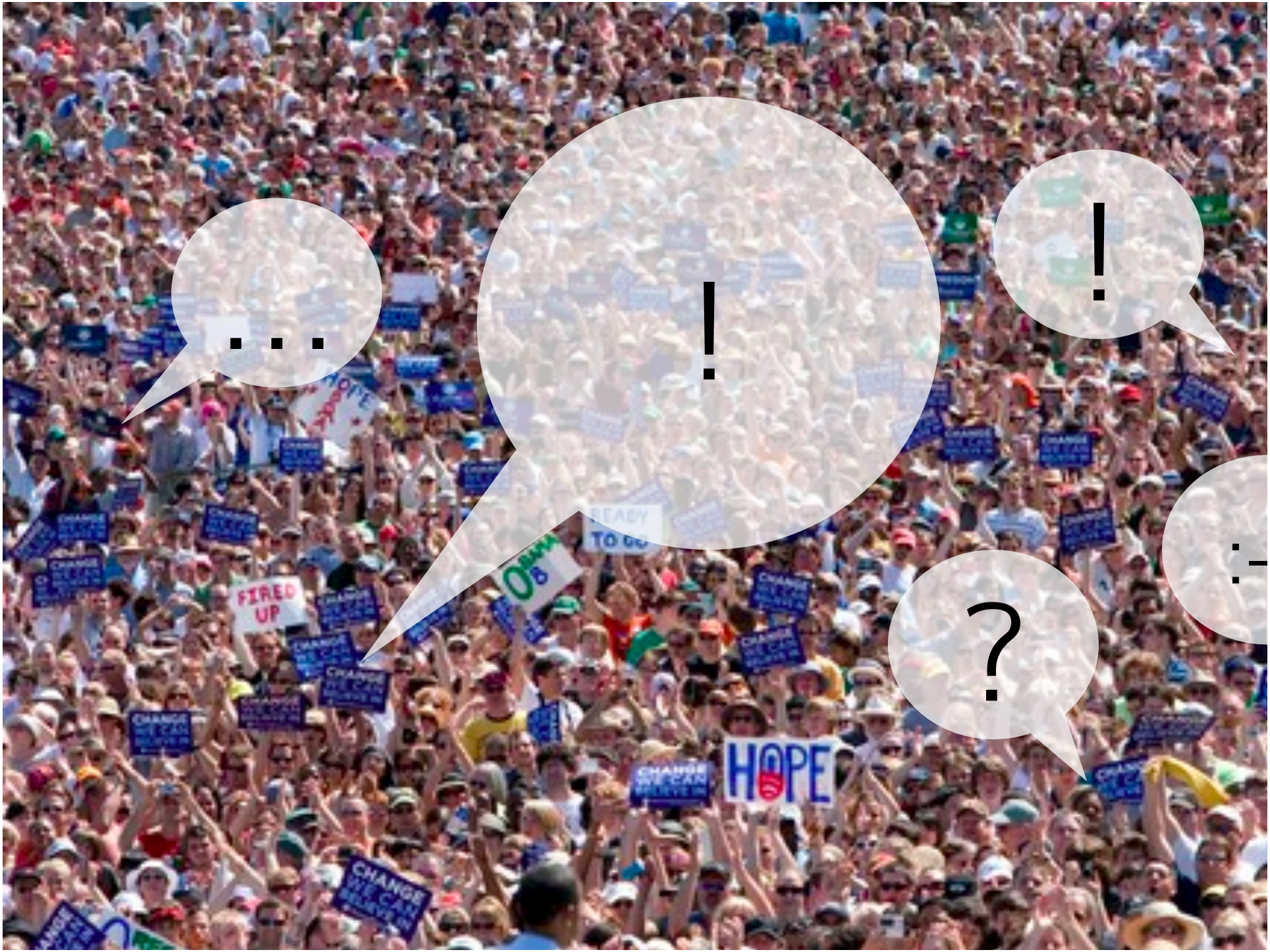
Text and Social Context: Analysis and Prediction

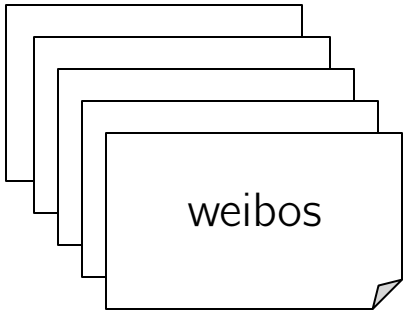
Noah Smith
Finmeccanica Associate Professor
School of Computer Science
Carnegie Mellon University
nasmith@cs.cmu.edu

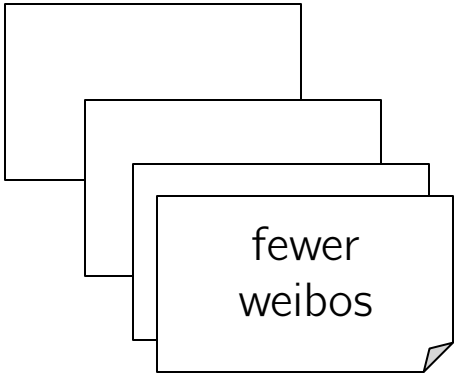
Joint work with: David Bamman, Chris Dyer, Michael Heilman (ETS), Brendan O'Connor, Bryan Routledge, John Wilkerson (UW), Tae Yano, and Dani Yogatama


. . .
!
!
?
:-
FIRED UP
HOPE
CHANGE WE CAN BELIEVE

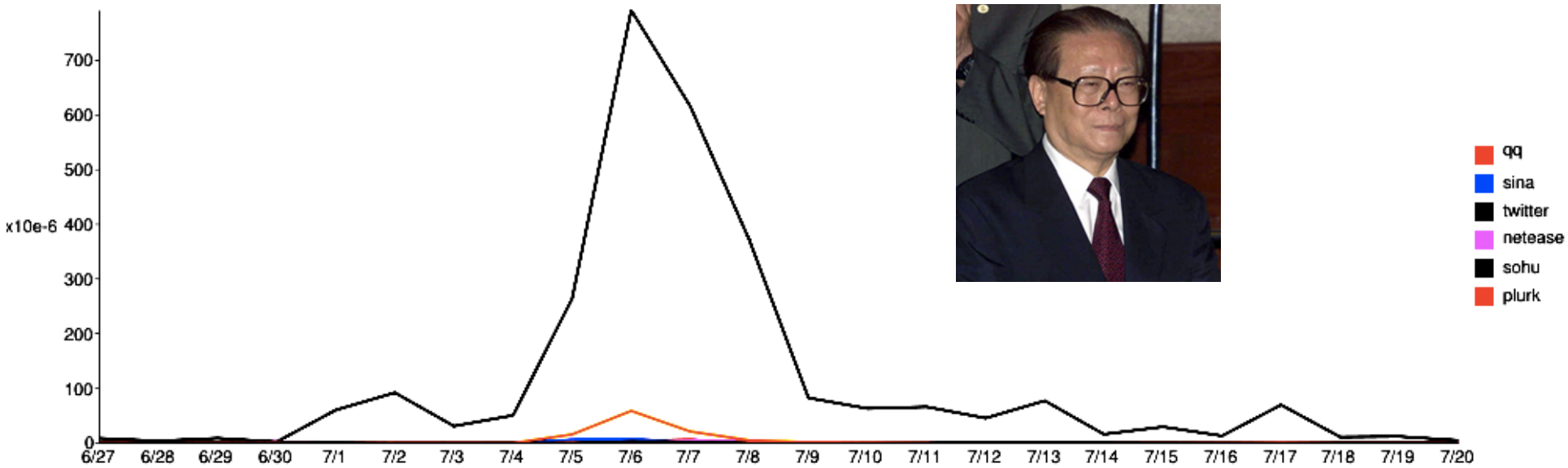
Outline

1. Analyzing social media content
   - Message deletion in China
2. Prediction of social outcomes using text
   - Will a scientific article get cited?
   - Will a congressional bill survive committee?


weibos

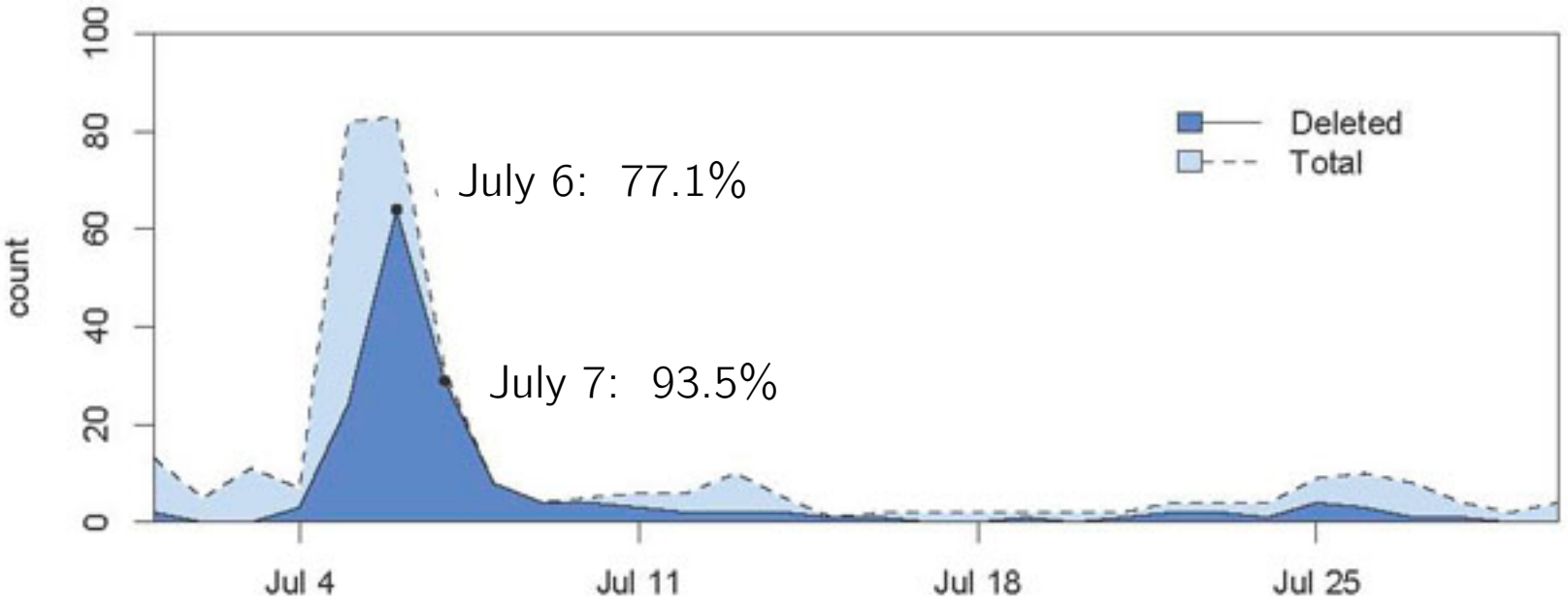

fewer
weibos

江泽民 (Jiang Zemin)



"We reported that Jiang was in critical condition or died based on several sources, but a figure in Beijing called us at noon to inform us that his condition has improved."

江泽民 (Jiang Zemin)



Searching for Jiang on `weibo.com`

"According to relevant laws, regulations and policies, the search results were not shown."

Background

- "Great Firewall of China": network filtering (Crandall et al., 2007, *inter alia*)
- Search filtering by various engines (Villeneuve, 2008a)
- Evidence for keyword-based censorship in chat (Villeneuve, 2008b) and blogs (MacKinnon, 2009)

- Domestic Chinese companies appear to police their own content.
- Overviews: MacKinnon (2011) and OpenNet Initiative (2009)

Sina Weibo Data

- 57 million messages collected June 27, 2011 – September 30, 2011.
- Sampled subset of 1.3 million from June 30 – July 25, subject to:
  - Remove duplicates based on Chinese-character content
  - At least five followers for author
  - Remove messages with hyperlinks if fewer than 100 followers

Terms

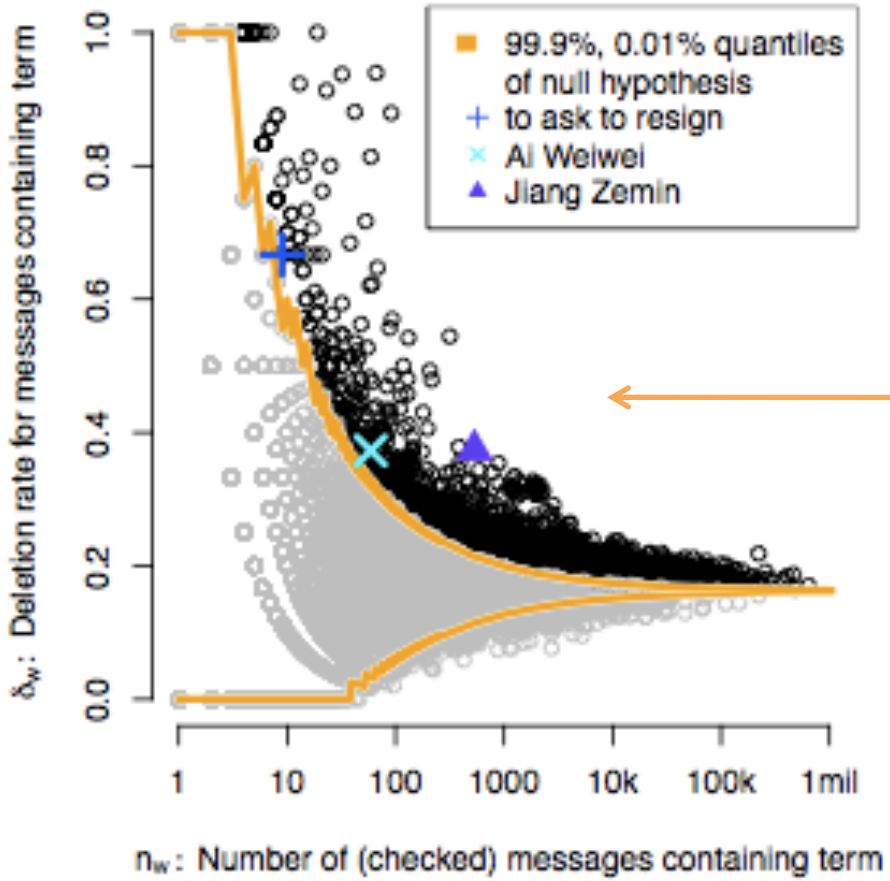
- 255,126-entry lexicon
  - Start with CC-CEDICT (open-source)
  - Add all Wikipedia page titles in Chinese with an English sibling page
- Consider all matches (2-5 characters) in a weibo; we count *document* frequency.

- Which terms have *abnormally* high deletion rates, accounting for randomness in sampling?

Message Deletions

- For the 1.3M: check three months after publications whether each message existed.
  - Overall deletion rate of 16.25%.

Deletion Rates and Document Frequencies



4.5%, not 0.1%, of points are above the orange line

False Discovery Rate

- For one term *w*, *p*-value is easy:

  $p_w = P_{null}(D > d_w \mid N = n_w)$

  $= 1 - \text{BinomialCDF}(d_w\ ;\ n_w,\ 0.1625)$

- Tens of thousands of hypothesis tests; require a correction.
- For (say) $p_w < 10^{-3}$, Benjamini and Hochberg (1995) give a calculation for the upper bound on FDR:

$$\text{FDR}_{p_w<0.001} < \frac{P_{null}(p_w < 0.001)}{\hat{P}(p_w < 0.001)} = \frac{0.001}{0.040} = 2.5\%$$

- At $p_w < 0.001$, we'd select 3,046 terms (out of 76K).

Finding Sensitive Terms

1. Comparison with Twitter
   - 11M tweets from 10K high-volume Twitter users (June 1-24): calculate log-likelihood ratio of Twitter vs. Sina.
2. Blocked searches
   - Of the top 2K terms from above, 6.8% (136 terms) were blocked on the Sina search interface.
3. Merge 136 high LLR terms with the list identified by Crandall et al. (2007) and the list on Wikipedia: **295** total.
4. Original 56M messages reduced to 33,363 containing such terms, of which 17.4% (5,811) were deleted
5. We considered FDR bound of 2.5%, giving 17 terms.

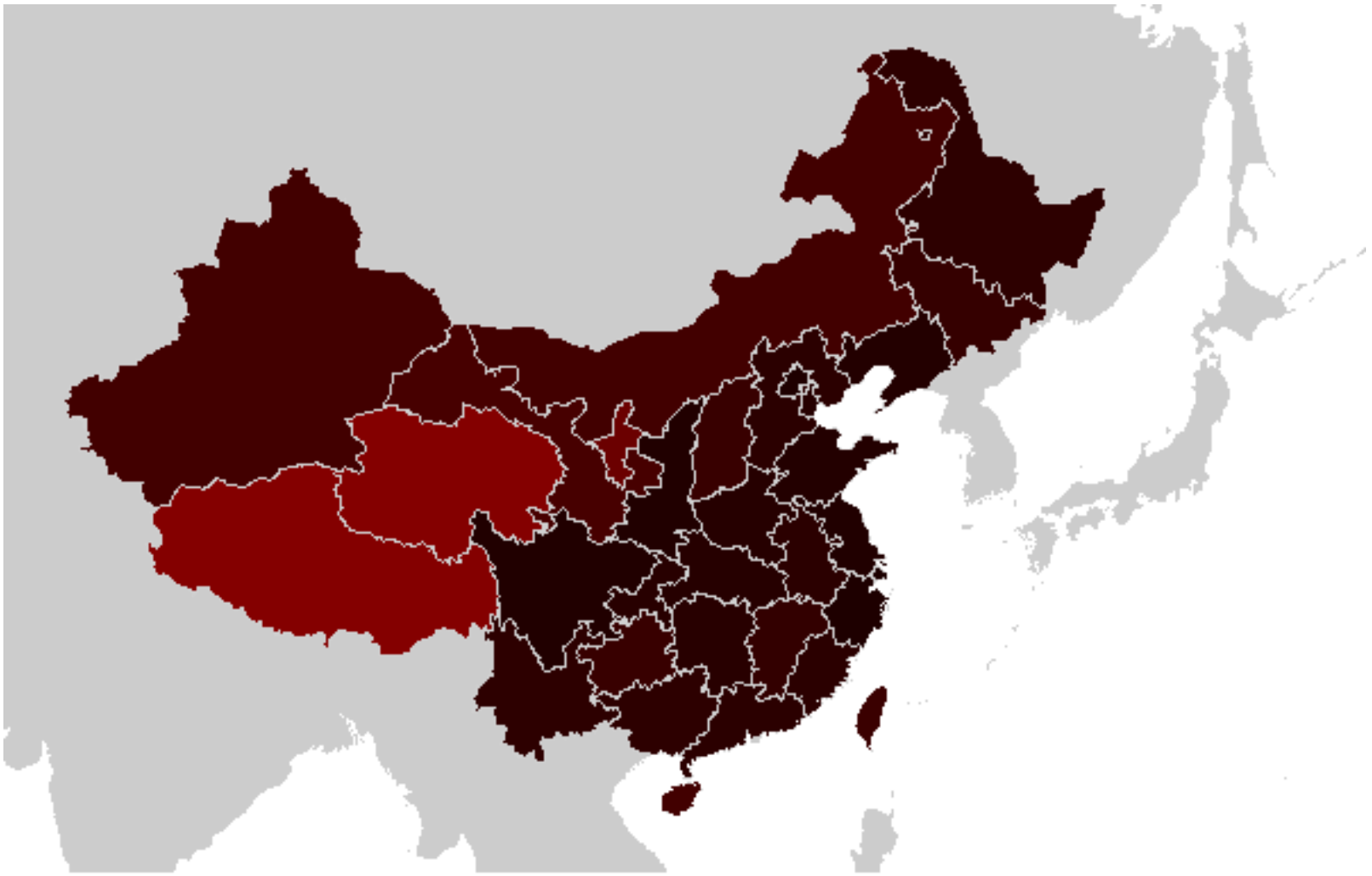
Sensitive Terms

方滨兴	Fang Binxing (architect of China's "Great Fire Wall")
真理部	Ministry of Truth (Orwell reference)
法轮功	Falun Gong (Chinese spiritual practice)
共匪	communist bandit
盛雪	Sheng Xue (human rights activist)
法轮	Falun
新语丝	*New Threads* (news website)
反社会	antisociety
江泽民	Jiang Zemin (1926-), President of PRC 1993-2003
艾未未	Ai Weiwei
不为人知的故事	"The Unknown Story"
流亡	to be exiled
驾崩	death of a king or emperor
浏览	to browse
花花公子	*Playboy*
封锁	to blockade
大法	Falun Dafa

Further Points

- Most terms are not deleted 100% of the time.
- Many terms from Crandall et al. (2007) were not being deleted at significant rates.
- Deletion seems to have no significant relationship with a message's rebroadcast count or author's follower count.
- *Ai Weiwei* positive sentiment messages: 11/16 not deleted.

- See also: new paper by King, Pan, and Roberts: "How Censorship in China Allows Government Criticism but Silences Collective Expression"



Related Work

- How does casual social text vary with geographic and demographic variables?
  - See Jacob Eisenstein's talk tomorrow evening!
- How do words describing food relate to its price?
  - "Word salad: relating food prices and descriptions." Victor Chahuneau, Kevin Gimpel, Bryan R. Routledge, Lily Scherlis, and Noah A. Smith. EMNLP 2012.

Outline

1. Analyzing social media content
   - ✓Message deletion in China
2. Prediction of social outcomes using text
   - Will a scientific article get cited?
   - Will a congressional bill survive committee?

Text-Driven Forecasting

- Build a model that predicts a real-world measurement in the future, using text.
- Train up to time t, test on data after t.
- Examples:
  - prediction markets from news (Lerman et al., 2008)
  - volatility of a firm's returns from its financial reports (Kogan et al., 2009)
  - volume of comments on blog posts (Yano and Smith, 2010)
  - a film's first-weekend box office revenues from its pre-release reviews (Joshi et al., 2010)

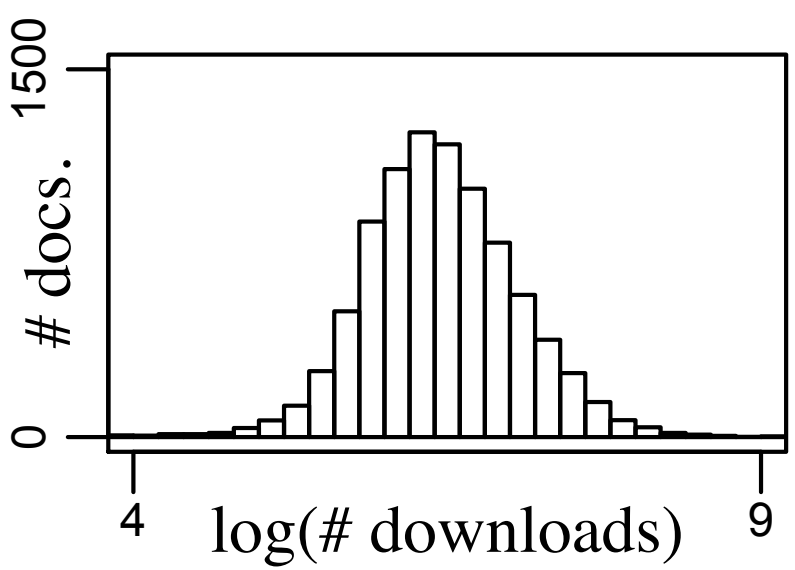
Example

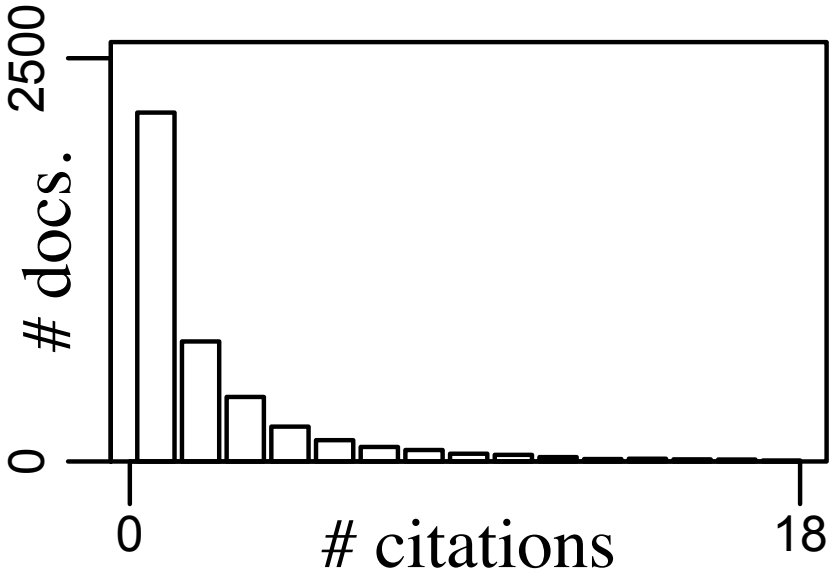
Title: Predicting a scientific community's response to an article
Authors: Yogatama, Heilman, O'Connor, Dyer, Routledge, Smith
Abstract: We consider the problem of predicting measurable responses to scientific articles based primarily on their text content. Specifically, we consider papers in two fields (economics and computational linguistics) and make predictions about downloads and within-community citations. Our approach is based on generalized linear models, allowing interpretability; a novel extension that captures first-order temporal effects is also presented. We demonstrate that text features significantly improve accuracy of predictions over metadata features like authors, topical categories, and publication venues.

Will this paper be cited? Will anyone read it? Should I?

Datasets

- Papers from the National Bureau of Economics Research (NBER), 1999-2010
  - Response: downloads within one year
  - Total documents: 8,814
- Papers from the Association for Computational Linguistics (ACL), 1980-2008 (Radev et al., 2009)
  - Response: cited or not within 3 years
  - Total documents: 4,026
- We do not use any structured citation data.


1500
docs.
0
4
log(# downloads)
9


2500
docs.
0
0
citations
18

Generalized Linear Models

- Represent input as a *feature vector* $\boldsymbol{x}$ and response as a boolean or real, y.

$$\hat{y} = \begin{cases} \boldsymbol{\beta}^\top \boldsymbol{x} & \text{continuous} \\ sign(\boldsymbol{\beta}^\top \boldsymbol{x}) & \text{binary} \end{cases}$$

- Parameterization is a linear model, so we need to learn coefficients for the features, denoted $\boldsymbol{\beta}$.
- Learning:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{X}, Y) + R(\boldsymbol{\beta})$$

  Continuous case uses MSE (linear regression); discrete uses logit link function (logistic regression).

Features

- Features in $\boldsymbol{x}$:
  - Author names (binary indicators)
  - NBER: Programs (e.g., "health economics")
  - ACL: publication venue
  - Words in title (binary and log-count)
  - NBER: abstract n-grams (binary and log-count), average length is 155 words
  - ACL: content n-grams (binary and log-count), average length is 3,966 words

Experimental Results

	ACL: 2006 test set (% accuracy)	NBER: 2009 test set (mean absolute error)
Majority class from training set (baseline)	50	
Median from training set (baseline)		397
Metadata only, one past year	62	375
Metadata only, all past data	60	378
All features, one past year	67	351
All features, all past data	**70**	**339**

Question

- Which data to train on?
- How to learn about *temporal trends*?

Revisiting Learning

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{X}, Y) + R(\boldsymbol{\beta})$$

- The regularizer R is usually set to be a quadratic penalty that discourages very large coefficients.
  - "Ridge" regression (Hoerl and Kennard, 1970)
  - Very, very important for models with high-dimensional input ($\boldsymbol{x}$)
- Many alternatives exist.

Time Series Model

- Intuition:
  - Importance of any feature (author, venue, term-expressed concept) will vary over time.
  - Learn different weights at different times.
  - Use the regularizer to encourage *smooth* change.
- New model: coefficients follow something like an AR(1) **time series**.

Example

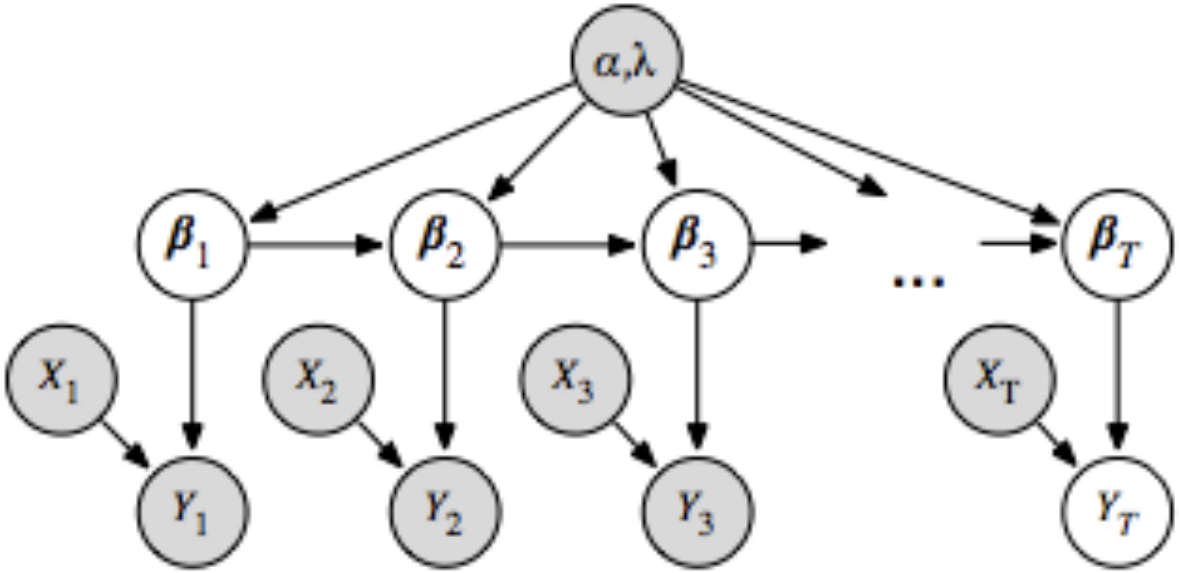
- Instead of a single feature for *unemployment rate* (bigram), we have a copy of the feature in each year:
  - $\textit{unemployment rate}_{1999}$
  - $\textit{unemployment rate}_{2000}$
  - $\textit{unemployment rate}_{2001}$
  - $\textit{unemployment rate}_{2002}$
- We encourage the feature coefficients to be similar to their adjacent copies.
  - See also Chelba and Acero (2006).

Time Series Regularizer

$$R(\boldsymbol{\beta}) = \lambda \sum_{t=1}^{T} \sum_{j=1}^{d} \beta_{t,j}^2 + \lambda\alpha \sum_{t=2}^{T} \sum_{j=0}^{d} (\beta_{t,j} - \beta_{t-1,j})^2$$

- Infinite α recovers our model trained on all data.
- $\alpha = 0$ recovers our model trained on just the most recent year (independent ridge regressions at each epoch).

Probabilistic Graphical Models View



Precision Matrix

$$\Lambda = \lambda \begin{bmatrix} 1+\alpha & -\alpha & 0 & 0 & \dots \\ -\alpha & 1+2\alpha & -\alpha & 0 & \dots \\ 0 & -\alpha & 1+2\alpha & -\alpha & \dots \\ 0 & 0 & -\alpha & 1+2\alpha & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Experimental Results

	ACL: 2006 test set (% accuracy)	NBER: 2009 test set (mean absolute error)
Metadata only, one past year ($\alpha = 0$)	62	375
Metadata only, all past data ($\alpha = \infty$)	60	378
Metadata only, time series (tuned α)	56	375
All features, one past year ($\alpha = 0$)	67	351
All features, all past data ($\alpha = \infty$)	70	339
All features, time series (tuned α)	**72**	**332**

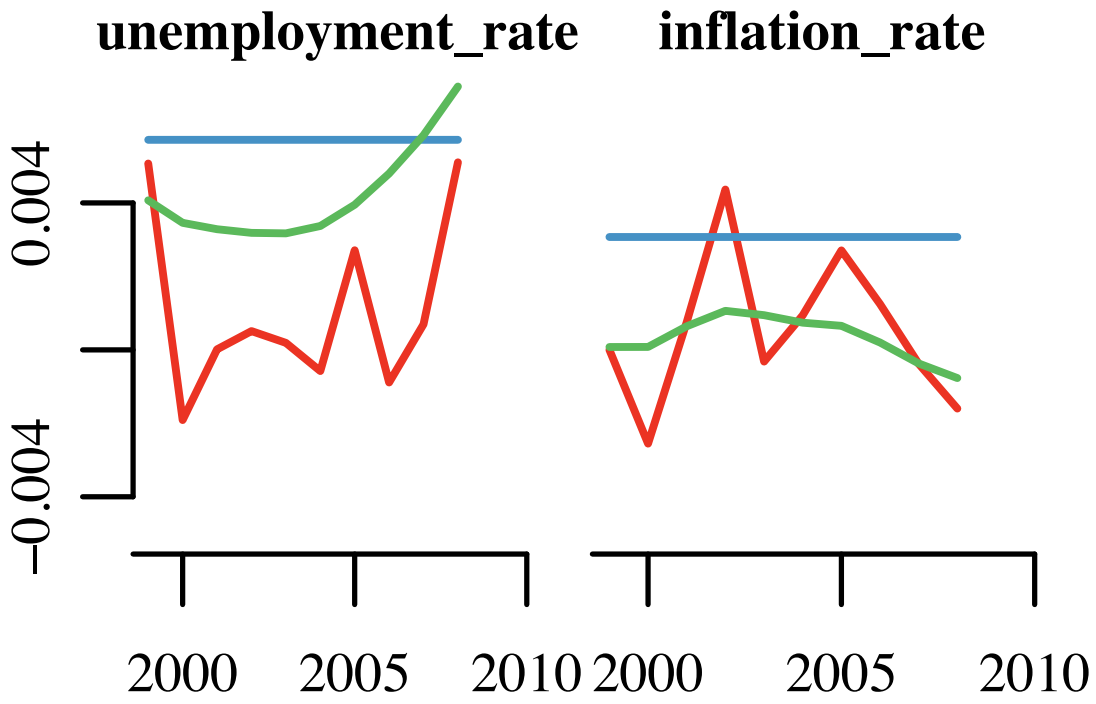
Time series model is advantageous (results are significant).

Experimental Results

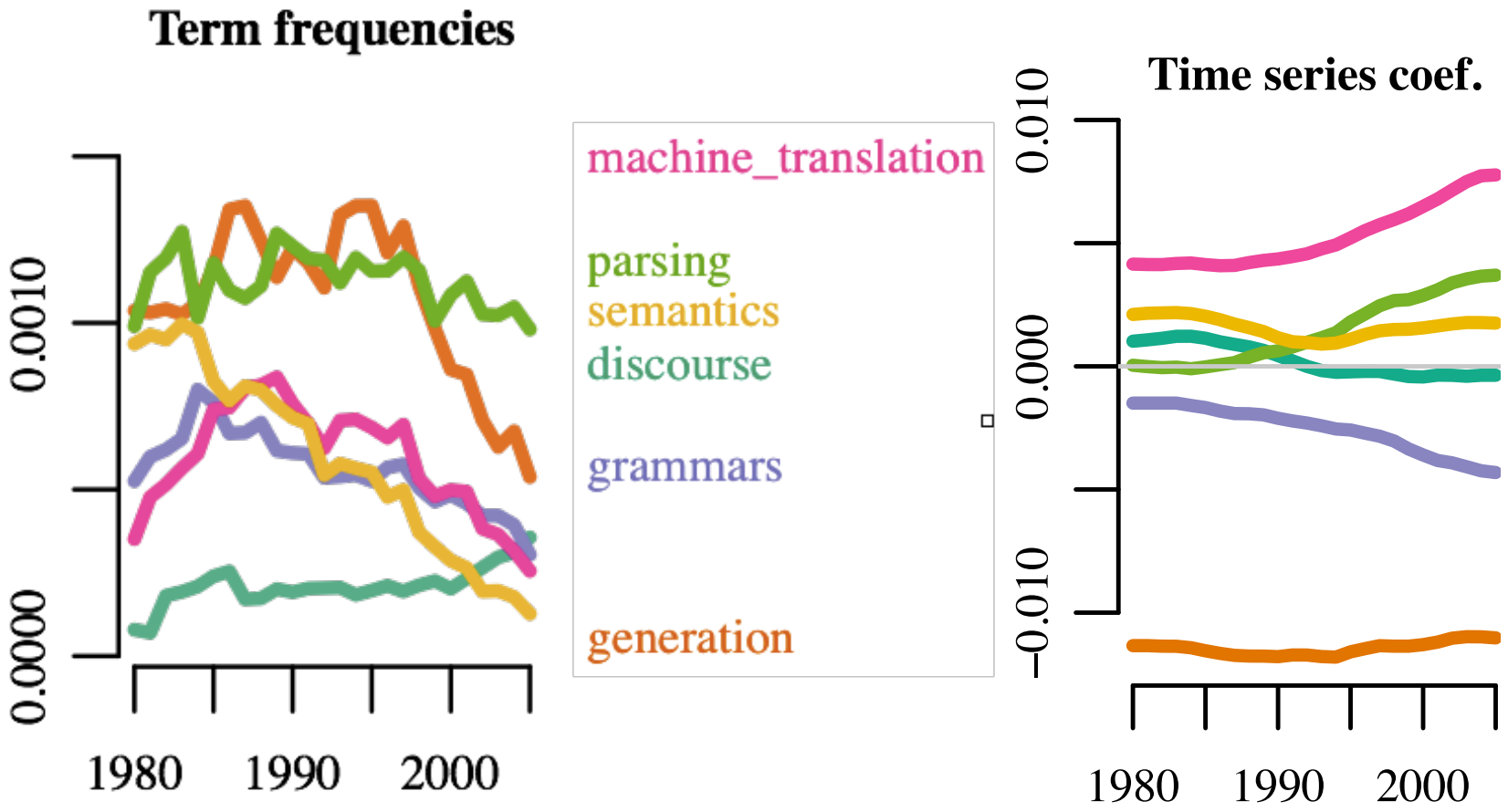
	ACL: 2006 test set (Kendall's τ)	NBER: 2009 test set (Kendall's τ)
Metadata only, one past year ($\alpha = 0$)	0.16	0.22
Metadata only, all past data ($\alpha = \infty$)	0.22	0.21
Metadata only, time series (tuned α)	0.22	0.17
All features, one past year ($\alpha = 0$)	0.33	0.31
All features, all past data ($\alpha = \infty$)	0.37	0.40
All features, time series (tuned α)	**0.38**	**0.43**

These corroborate the accuracy and MAE results.

Inside the Model

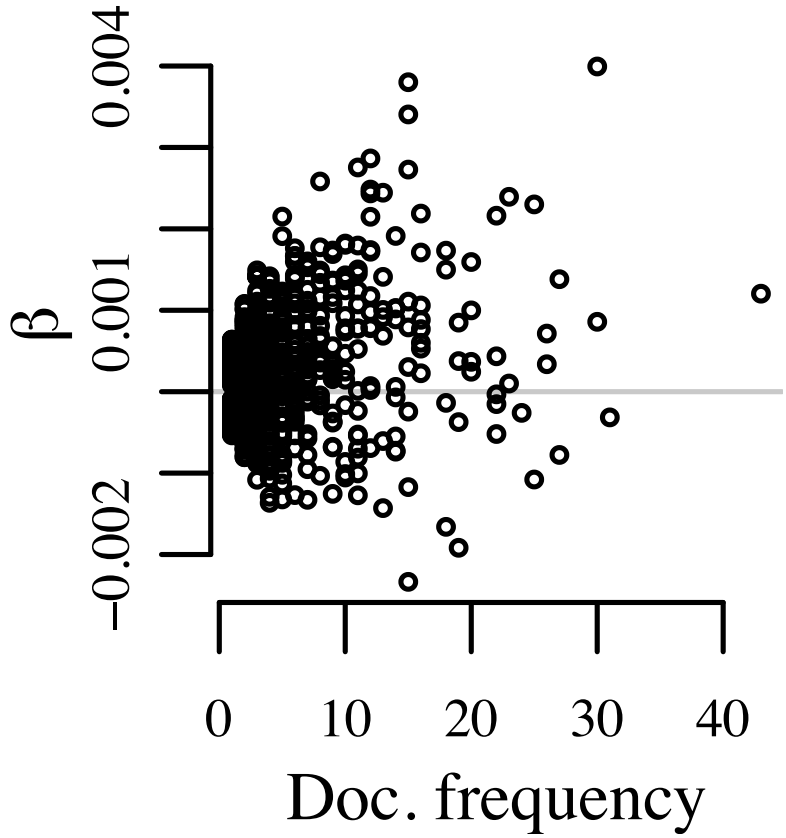


Trends

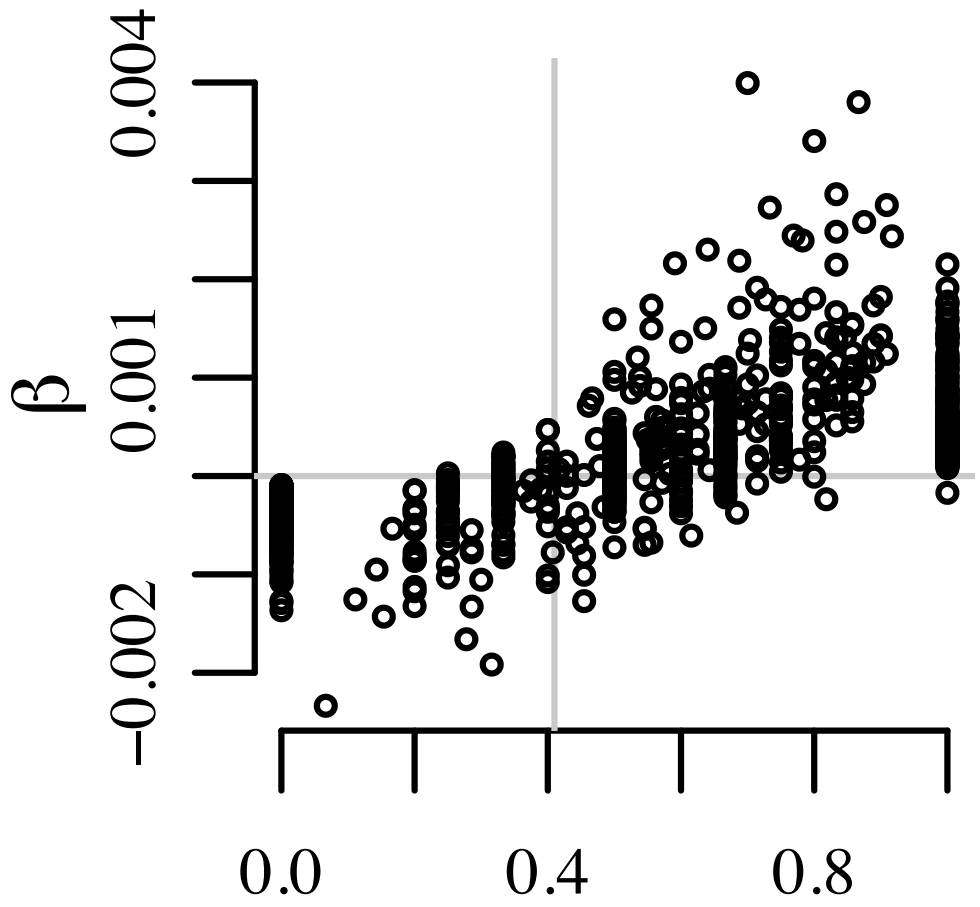


"Google N-Gram Books Viewer"
(Michel et al., 2011) approach.

ACL authors: Writing more papers doesn't necessarily increase citation odds.



ACL authors: coefficient relates closely to rate of citation, but is less brittle.



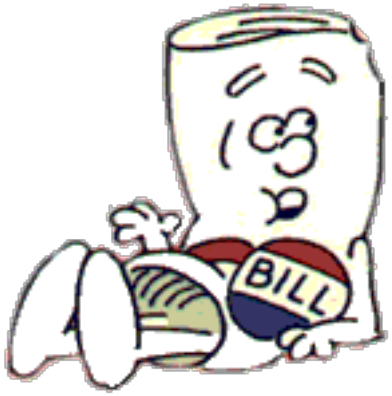
See Also

- Bayesian data analysis to discover factions that explain who cites whom, and with what words.
  - "Discovering factions in the computational linguistics community." Yanchuan Sim, Noah A. Smith, and David A. Smith. ACL "Rediscovering 50 Years of Discoveries" Workshop, 2012.

Outline

1. Analyzing social media content
   - ✓Message deletion in China
2. Prediction of social outcomes using text
   - ✓Will a scientific article get cited?
   - – Will a congressional bill survive committee?

Bill Survival



- ~85% of bills introduced in Congress do not survive committee.
  - Of those that *do*, 90% pass a floor vote.
- Committees are not transparent.
- Agenda-setting: what factors lead to success?

Our Dataset

- Nine Congresses (each 2 years, 1993-2011).
- We consider only the House of Representatives.
- Total 51,762 bills, downloaded from THOMAS, the Library of Congress website.
  - Additional data from Charles Stewart's resources (MIT) and the Congressional Bills Project (UW) – gratefully acknowledged!
  - Mean 1,972 words, s.d. 3,080.
  - 13.2% survival rate.
- We know a bill survives if it is *reported*.

An Example

- Identifier: C103-HR748
- Response: false
- Sponsor: Ken Calvert (Rep., CA)
  - (Sponsor is not in the majority party.)
- Introduced: February, year 1 of 2
- Committee: Judiciary
  - (Sponsor is not on the committee of referral.)
- Title: For the relief of John M. Ragsdale

103rd Congress, H.R. 748

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. COMPENSATION FOR WORK-RELATED INJURY.

(a) AUTHORIZATION OF PAYMENT- The Secretary of the Treasury shall pay, out of money in the Treasury not otherwise appropriated, the sum of $46,726.30 to John M. Ragsdale as compensation for injuries sustained by John M. Ragsdale in June and July of 1952 while John M. Ragsdale was employed by the National Bureau of Standards.

(b) SETTLEMENT OF CLAIMS- The payment made under subsection (a) shall be a full settlement of all claims by John M. Ragsdale against the United States for the injuries referred to in subsection (a).

SEC. 2. LIMITATION ON AGENTS AND ATTORNEYS' FEES.

It shall be unlawful for an amount that exceeds 10 percent of the amount authorized by section 1 to be paid to or received by any agent or attorney in consideration of services rendered in connection with this Act. Any person who violates this section shall be guilty of an infraction and shall be subject to a fine in the amount provided in title 18, United States Code.

Task Definition

- Given the sponsor (identity, party, state), committee makeup, date, and, optionally, title and text contents, predict whether a bill will survive.
  - Cf. Thomas, Pang, and Lee (2006), who modeled support/opposition for a bill from floor debate transcripts.
  - Cf. Gerrish and Blei (2011), who predicted survival on the *floor*, not in committee.

A Basic Model (No Text): 3,731 Features

1. Is the bill's sponsor affiliated with party p?
2. Is the sponsor in the majority party?
3. Is the sponsor on the committee?
4. $f_2 \wedge f_3$
5. Is the sponsor the chairman of the committee?
6. Was j the sponsor of the bill?
7. $f_5 \wedge f_6$
8. $f_2 \wedge f_6$
9. Is the sponsor from state s?
10. Was the bill introduced during month m?
11. Was the bill introduced during year y of 2?

Model

- We use L_1-regularized logistic regression:

$$\hat{y}(x) = \begin{cases} 1 & \text{if } \hat{\mathbf{w}}^\top \mathbf{f}(x) > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \sum_i \log \underbrace{p_{\mathbf{w}}(y_i \mid x_i)}_{\frac{\exp y_i \mathbf{w}^\top \mathbf{f}(x_i)}{1+\exp \mathbf{w}^\top \mathbf{f}(x_i)}} - \lambda \|\mathbf{w}\|_1$$

- λ tuned on development data.

Baseline Error

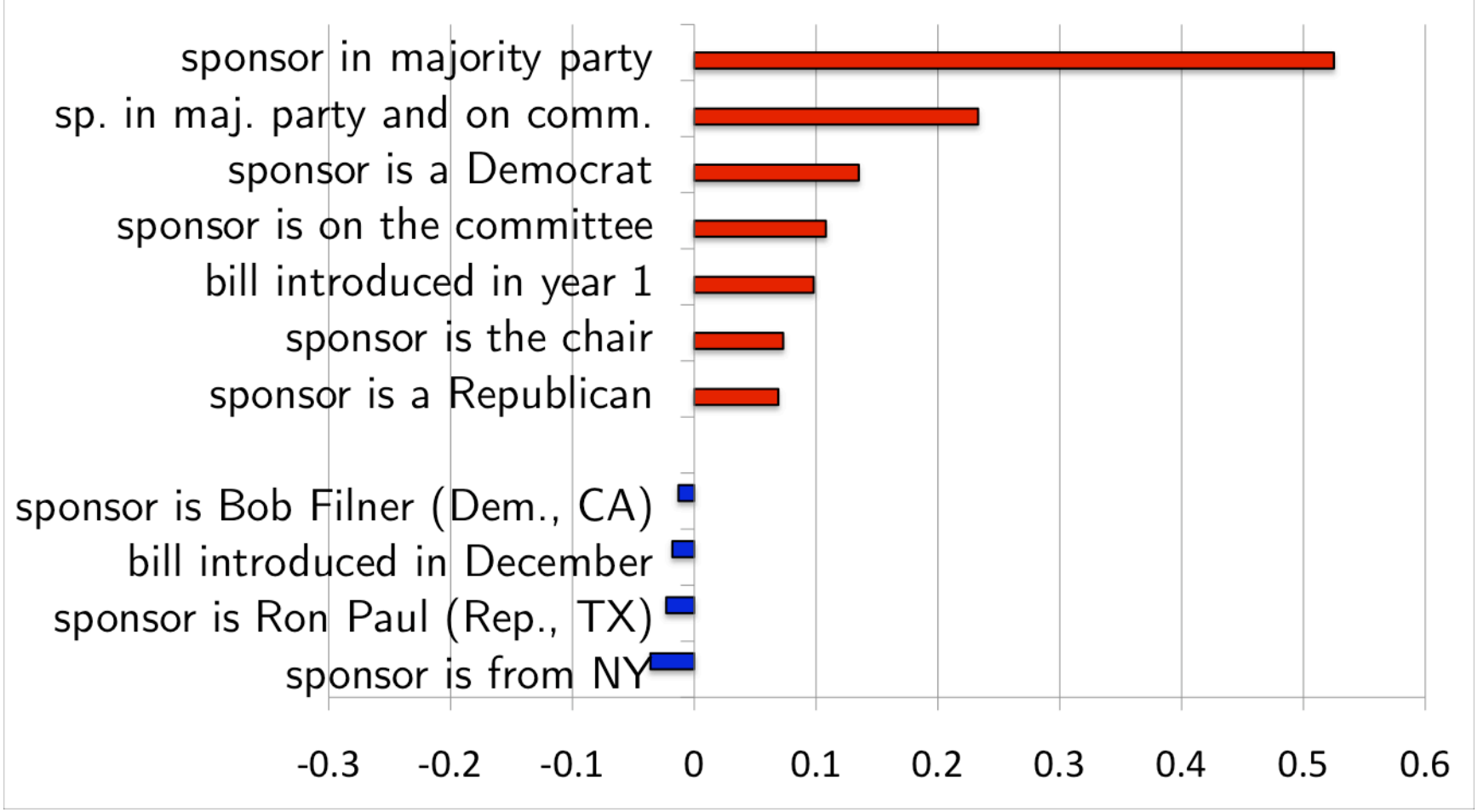
	Test on 109th (2005-2007)	Test on 110th (2007-2009)	Test on 111th (2009-2011)
Majority class from training set	11.8	14.5	12.6
Baseline (metadata)	11.1	13.9	11.8

Inspecting the Model

- Look at the weights: if you change the feature, how much do the log-odds change?
  - But rare features sometimes get large weights.
- Instead, we consider **impact** (credit: Brendan O'Connor).
  - How much effect does this feature actually have on the model's beliefs, summed over the test data?

$$w_j \cdot \frac{\sum_{i=1}^{N} f_j(x_i)}{N}$$

Impact of Features on Test-Set Predictions



Three Ways to Use Text

1. Functional categories (Adler and Wilkerson, 2005): trivial, recurring, important
    1. ~83% accuracy against expert coders
2. "Proxy vote" based on text similarity to past *floor* votes
3. Unigrams and bigrams

Text Model #1: Functional Categories

- Adler and Wilkerson (2005): functional category of a bill is an important factor in its success.
- Annotated data from 101-105th Congresses (103-105th in our data): bills can be **trivial** (11%), technical (1%), **recurring** (7%), **important** (10%).
  - Categories can overlap.
- In a cross-validation experiment, logistic regression on word features gets 83%.
  - Add 24 binary features based on posterior bins (3 labels x 2 differently regularized models x 4 bins).

Functional Category Error

	Test on 109th (2005-2007)	Test on 110th (2007-2009)	Test on 111th (2009-2011)
Majority class from training set	11.8	14.5	12.6
Baseline (metadata)	11.1	13.9	11.8
Metadata + functional bill categories (from textcat model)	10.9	13.6	11.7

Number of features with impact (111^{th}): 460 vs. 152

Text Model #2: Similarity to Past Bills

- Most committee members have voted on bills *on the floor* in the past.
- Perhaps voting behavior on similar bills is an estimate for the new bill?
- Features that tally **proxy votes** (estimates of *yea*, *nay*, and their ratio), quantized into bins.

Proxy Vote

- Simple way to estimate the **proxy vote**:
  - Assume each voter chooses a bill from the past x_{past} is chosen randomly, proportional to exp cosine-similarity(x, x_{past}), from the set of bills this individual voted on (out of 2,014).
  - Assume the vote on x = the vote on x_{past}.
  - Calculate the expected value of the vote, summing over past bills.

- Who "votes"? Chair only, majority party, or all.

Proxy Vote Error

	Test on 109th (2005-2007)	Test on 110th (2007-2009)	Test on 111th (2009-2011)
Majority class from training set	11.8	14.5	12.6
Baseline (metadata)	11.1	13.9	11.8
Metadata + functional bill categories (from textcat model)	10.9	13.6	11.7
Metadata + text-based proxy vote	9.9	12.7	10.9

The *chair* proxy vote features accounts for most performance gain.

Text Model #3: Direct

- Unigram indicators from bill body
- Unigram and bigram indicators from bill title (separate)
- Punctuation removed, numerals collapsed, filter to terms with document frequency between 0.5% and 30%.
- 24,515 lexical features considered
  - Baseline was 3,731

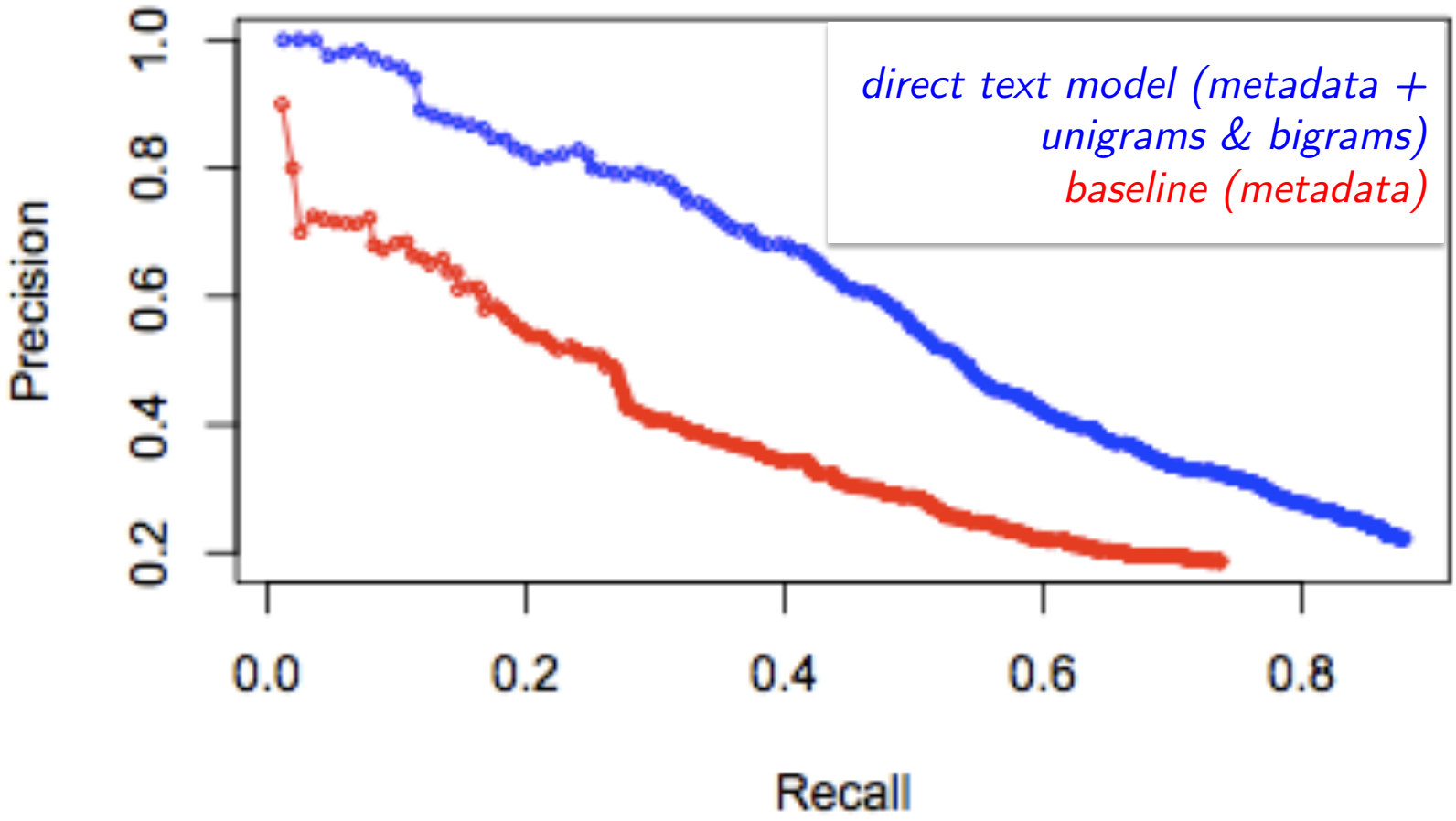
Direct Words Error

	Test on 109th (2005-2007)	Test on 110th (2007-2009)	Test on 111th (2009-2011)
Majority class from training set	11.8	14.5	12.6
Baseline (metadata)	11.1	13.9	11.8
Metadata + functional bill categories (from textcat model)	10.9	13.6	11.7
Metadata + text-based proxy vote	9.9	12.7	10.9
Metadata + unigrams & bigrams	8.9	10.6	9.8

Direct Words

	% nonzero-weighted features with impact	Test on 111th (2009-2011)
Majority class from training set		12.6
Baseline (metadata)	36	11.8
Metadata + functional bill categories (from textcat model)	55	11.7
Metadata + text-based proxy vote	58	10.9
Metadata + unigrams & bigrams	98	9.8

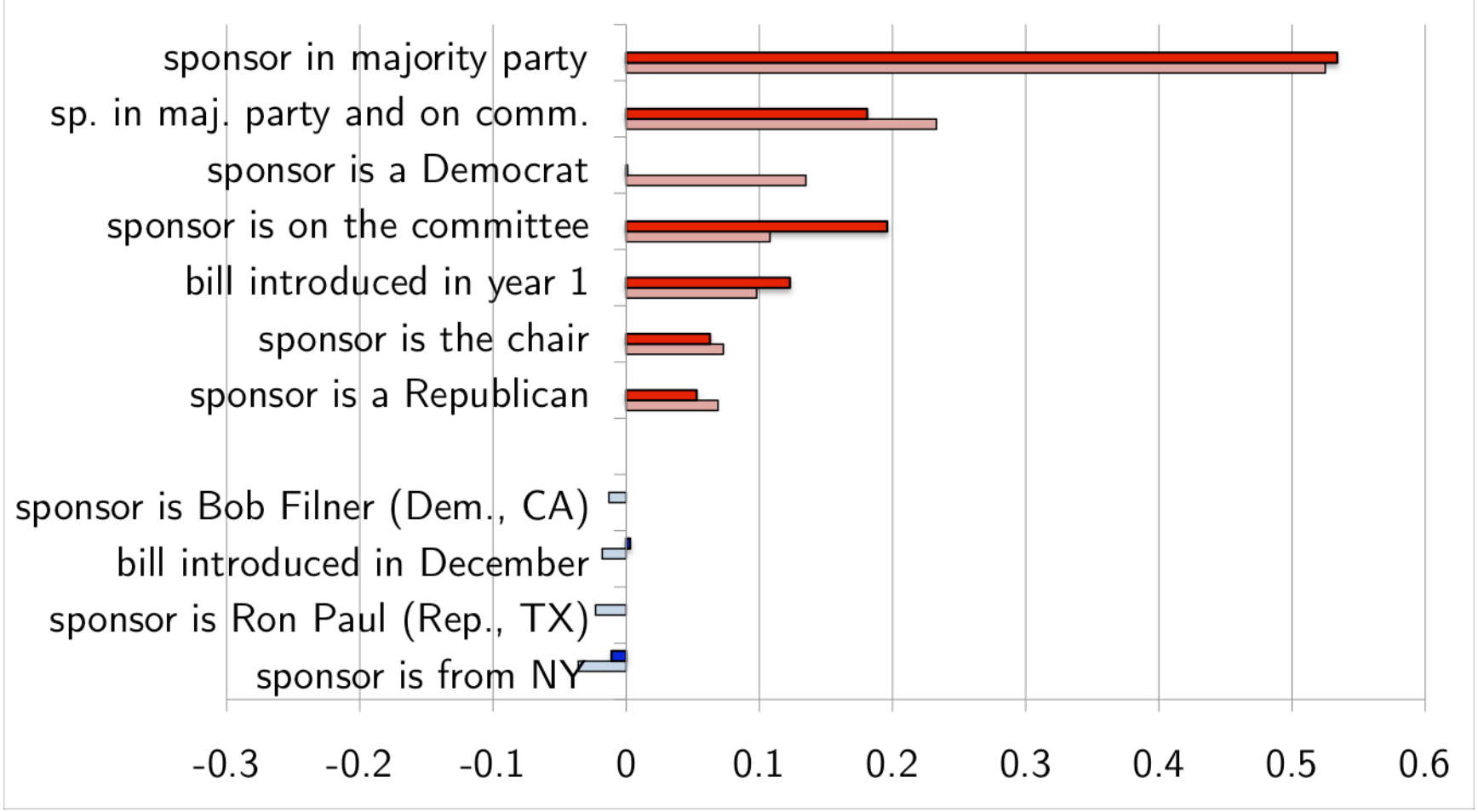
Number of features with impact (111th): 194


direct text model (metadata + unigrams & bigrams)
baseline (metadata)
Precision
1.0
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
Recall

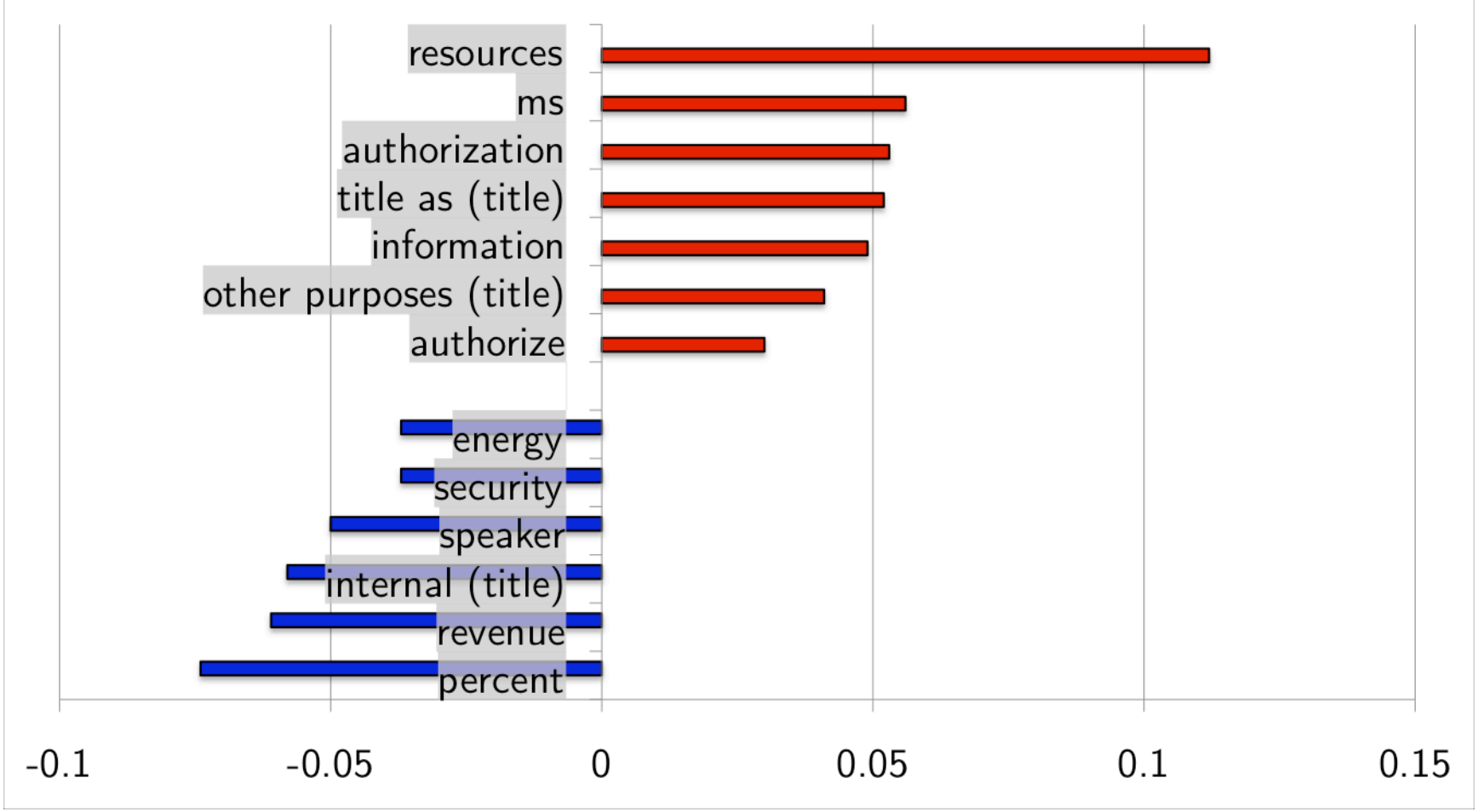
Full Model Error

	Test on 109th (2005-2007)	Test on 110th (2007-2009)	Test on 111th (2009-2011)
Majority class from training set	11.8	14.5	12.6
Baseline (metadata)	11.1	13.9	11.8
Metadata + functional bill categories (from textcat model)	10.9	13.6	11.7
Metadata + text-based proxy vote	9.9	12.7	10.9
Metadata + unigrams & bigrams	**8.9**	**10.6**	9.8
All	**8.9**	10.9	**9.6**

Impact of Features on Test-Set Predictions



Impact of Features on Test-Set Predictions



What Is Discovered?

- Survival features appear to focus on non-controversial issues (local land transfer, naming federal buildings).
- Death features:
  - Position-taking
  - "Omnibus" effect
  - Special bill numbers

The Data

- Congressional Bill Corpus v. 1.00 is available at http://www.ark.cs.cmu.edu/bills
  - Includes text, metadata, outcomes

Who Cares?

- How does the substance of a policy proposal relate to its progress?
- How are different types of issues managed in legislatures?
- Laws result from a complex social process; language is at the heart of it.
  - NLP as a tool for understanding the social world?
  - "Text as [quantitative] data" movement in political science (Grimmer and Stewart, 2012).

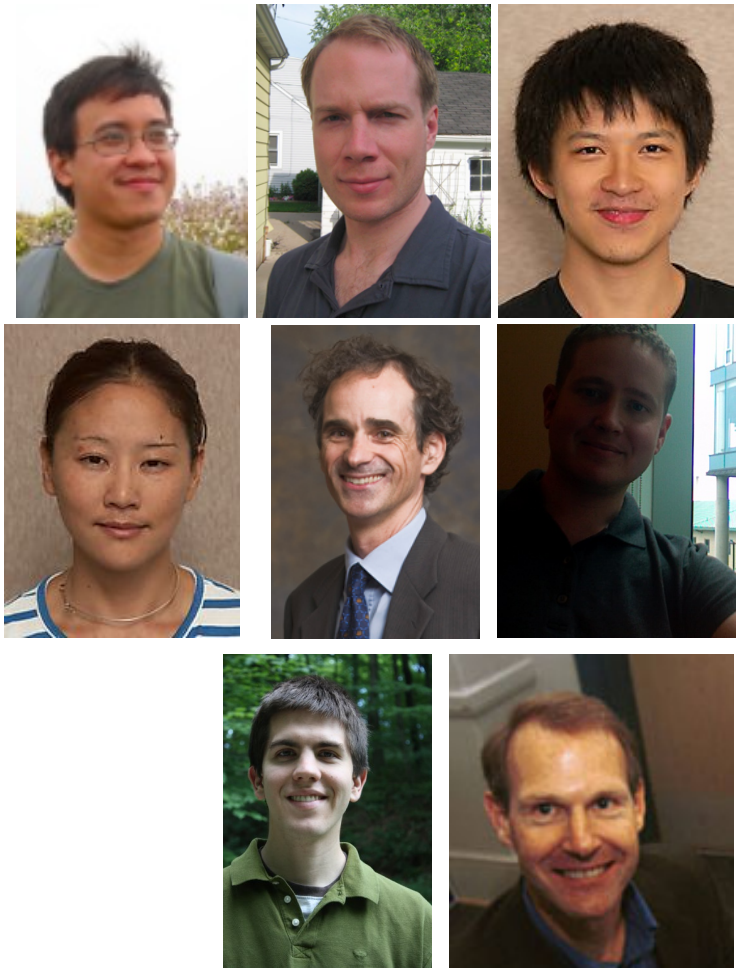
Outline

1. Analyzing social media content
   - ✓Message deletion in China
2. Prediction of social outcomes using text
   - ✓Will a scientific article get cited?
   - ✓Will a congressional bill survive committee?

Conclusions

- Social media reveal everyday language in context and let us use language to understand that context.
- Text data can predict and shed light on social phenomena.
- Familiar discriminative models provide an extensible framework for integrating text and non-text data to obtain:
  - Accurate predictions
  - Interpretable models

Where To Find More



Censorship and Content Deletion in Chinese Social Media. David Bamman, Brendan O'Connor, and Noah A. Smith. *First Monday* 17(3), March 2012.

Predicting a Scientific Community's Response to an Article. Dani Yogatama, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R. Routledge, and Noah A. Smith. EMNLP 2011.

Textual Predictors of Bill Survival in Congressional Committees. Tae Yano, Noah A. Smith, and John D. Wilkerson. NAACL 2012.

Acknowledgments:

- NSF (Smith)
- IARPA (Smith, Routledge)
- Google