Textual Predictors of Bill Survival in Congressional Committees

Tae Yano, LTI, CMU

Noah Smith, LTI, CMU

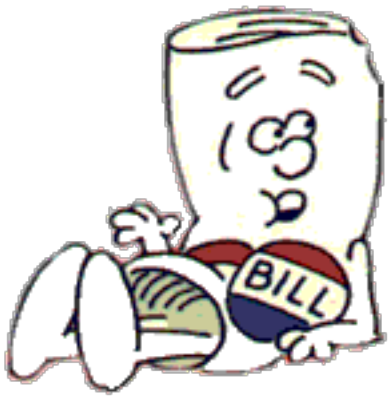
John Wilkerson, Political Science, UW

Thanks: David Bamman, Justin Grimmer, Michael Heilman, Brendan O'Connor, and Dani Yogatama. This research was supported by DARPA grant N10AP20042.

Outline

1. A little background on U.S. government
2. A task: predicting bill survival
3. Baseline model
4. Three ways to do better with text
5. Data release

The Early Life of a Bill

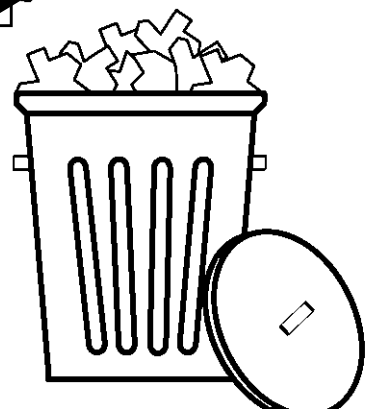


Formally proposed by one member of Congress (sponsor), routed to 1+ committees.

Congressional committee

~20 committees in the House of Representatives, each with a chairman, subcommittees, and more structure.
No consistent transcript availability, and no transcripts for bills that don't survive.

~13% of bills survive



"The fight on the floor of Congress between Matthew Lyon and Roger Griswold."
Unknown artist, 1798.

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.
- 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?

The Only Formula Slide

- 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.

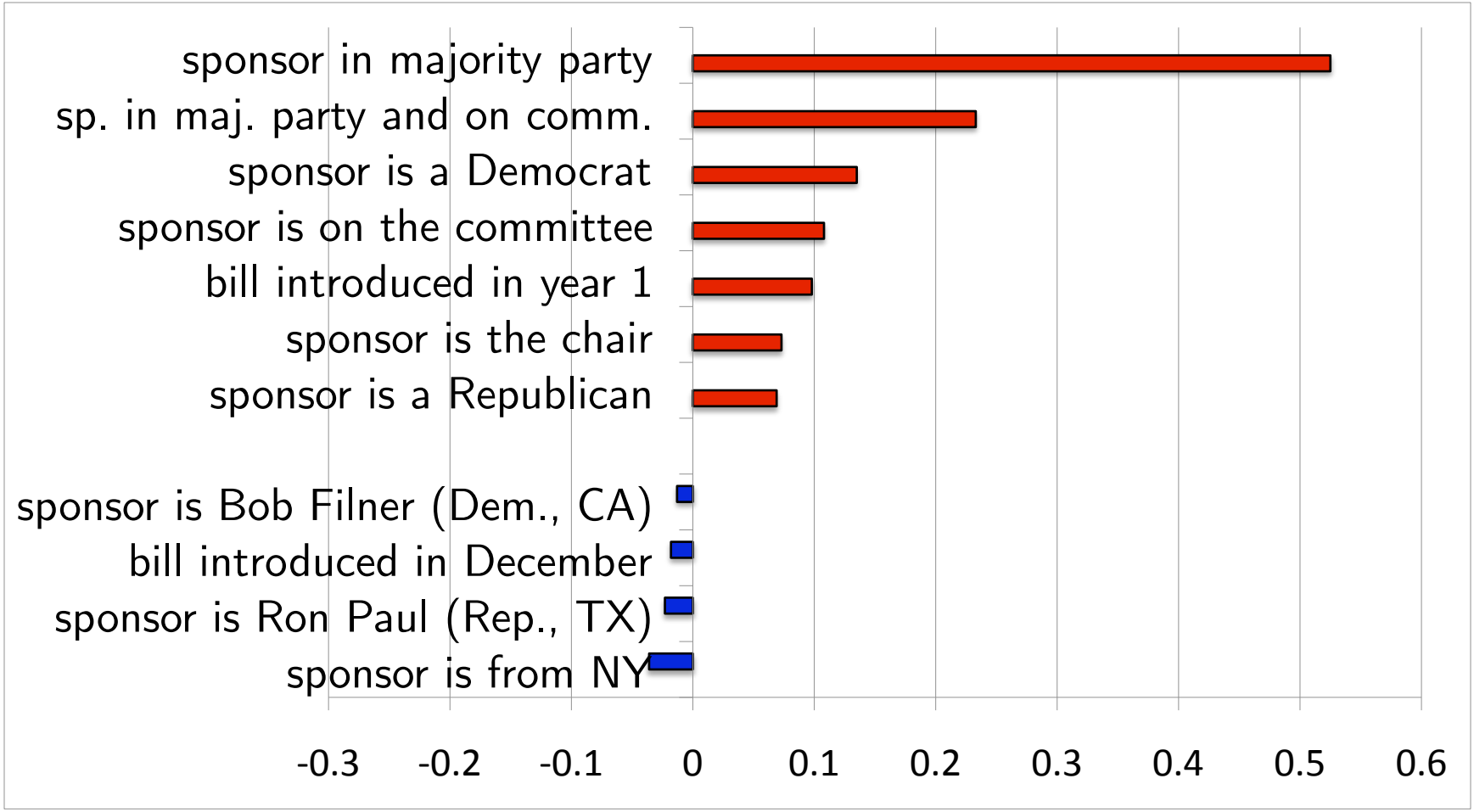
Baseline Error

	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?

Impact of Features on Test-Set Predictions



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 $\times$ 2 differently regularized models $\times$ 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 $\exp \text{cosine-similarity}(x, x_{past})$, 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

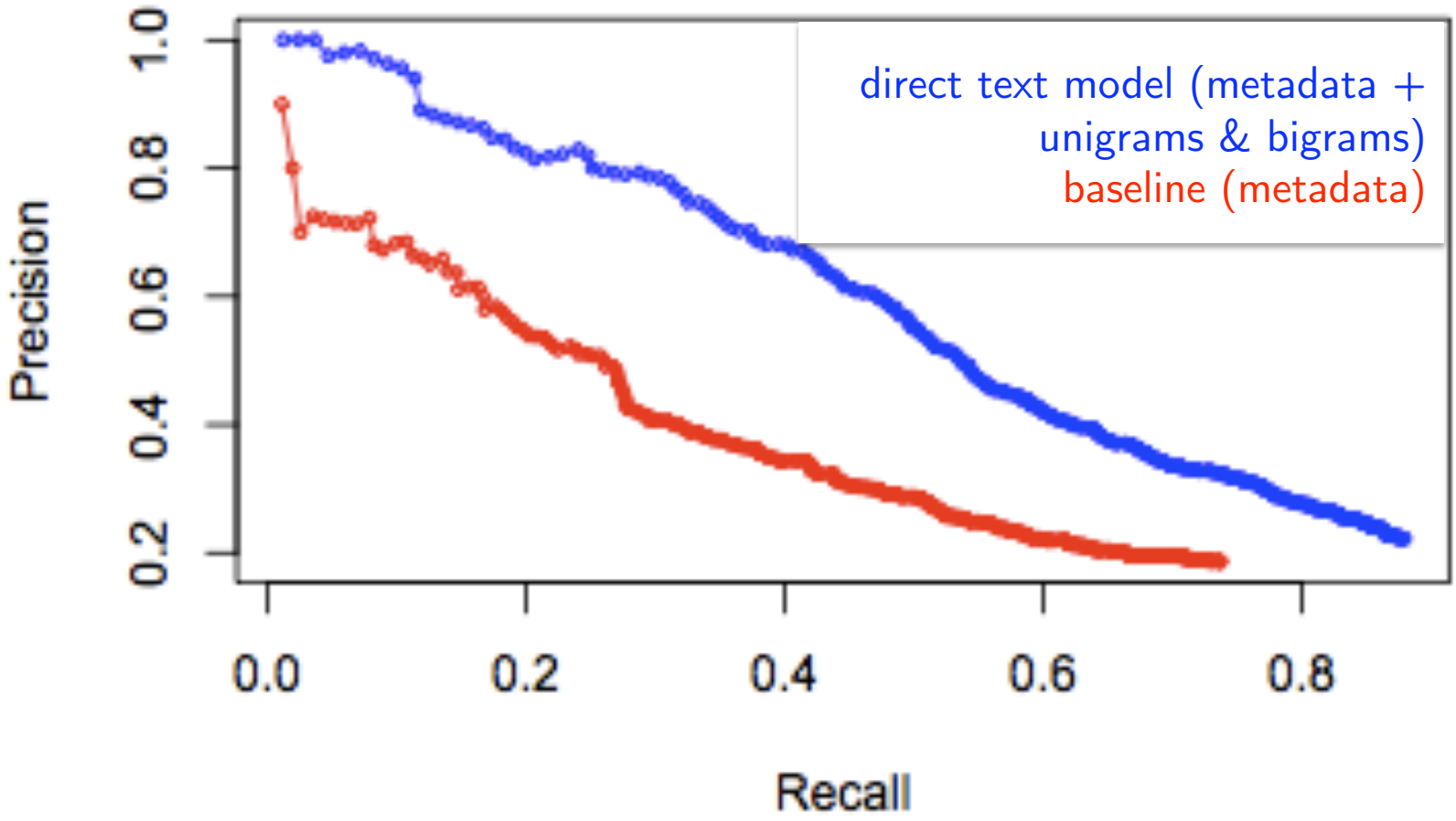
- Unigram indicators from bill body
- Unigram and bigram indicators from bill title (separate)
- Punctuation removed, numerals collapsed, filter to terms with document frequency between 0.5% and 30%.
- 24,515 lexical features considered
  - Baseline was 3,731

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

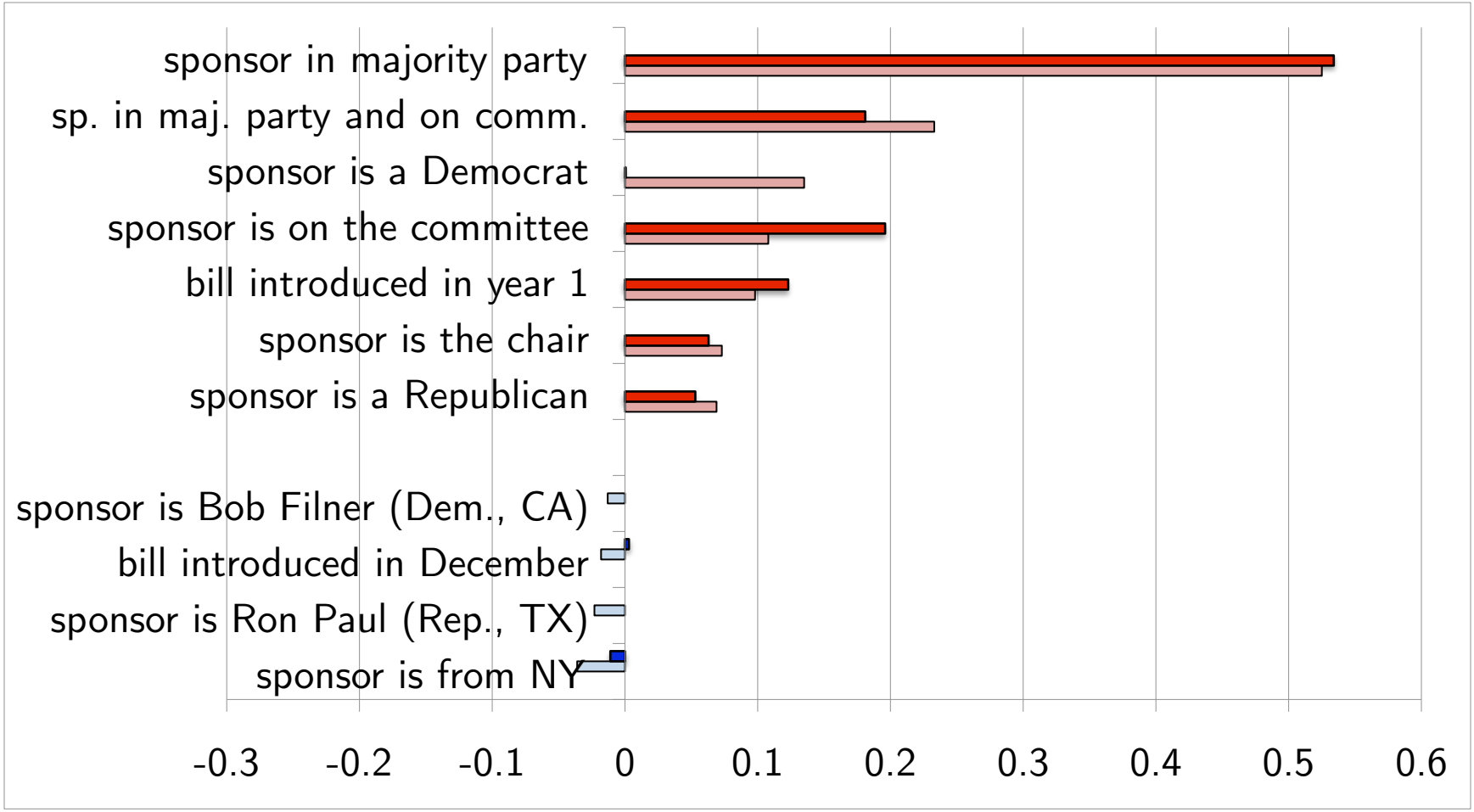
	% nonzero-weighted features with impact	Test on 111th (2009-2011)
Majority class from training set		12.6
Baseline (metadata)	36	11.8
Metadata + functional bill categories (from textcat model)	55	11.7
Metadata + text-based proxy vote	58	10.9
Metadata + unigrams & bigrams	98	9.8


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

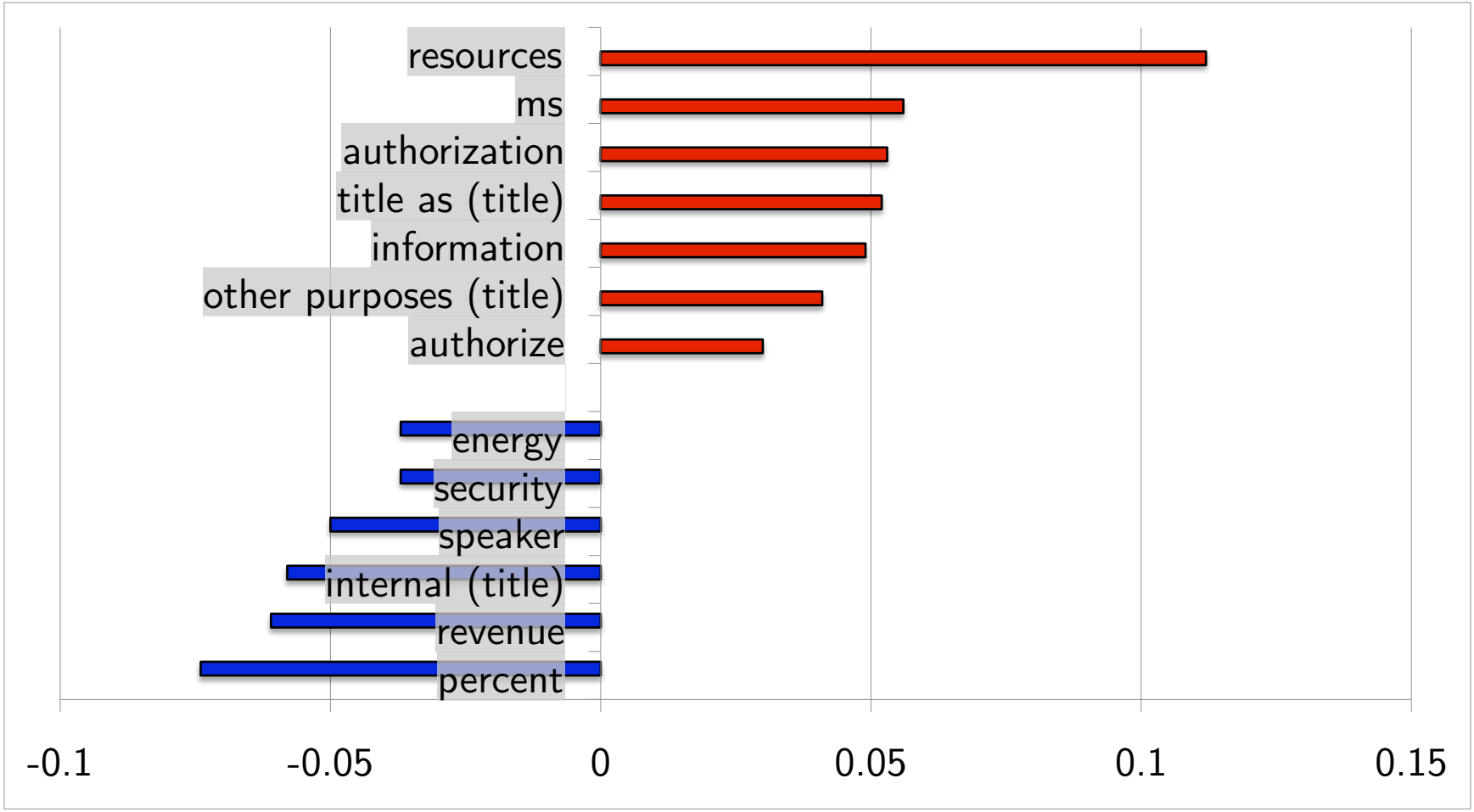
Full Model Error

	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:
  - Some evidence for "position-taking" (*energy*, *security*, *human*) – sponsoring bills on principle, not because they can survive.
  - Tax and social security bills tend to die; their contents are often packaged into larger bills.
  - Bill numbers that are "reserved for the speaker" are often not introduced. In our data, these "die."

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).

Thanks!