Predicting Risk from Financial Reports with Regression

Shimon Kogan, University of Texas at Austin
Dimitry Levin, Carnegie Mellon University
Bryan R. Routledge, Carnegie Mellon University
Jacob S. Sagi, Vanderbilt University
Noah A. Smith, Carnegie Mellon University

Talk In A Nutshell

financial risk = f(financial report)

volatility of returns

SV regression

Form 10-K, Item 7

What This Talk Isn't

New statistical models for NLP ...

Exciting text domains like **political blogs** ...

Advances in applications like **translation** and **summarization** ...

What This Talk Isn't

Shay Cohen,
10:40 am
yesterday

New statistical models for NLP ...

Tae Yano,
10:40 am
tomorrow

Exciting text domains like **political blogs** ...

Ashish
Venugopal,
right now

Advances in applications like **translation** and **summarization** ...

André Martins,
11 am Thursday

What This Talk Isn't

New statistical models for NLP ...

Exciting text domains like **political blogs** ...

Advances in applications like **translation** and **summarization** ...

What This Talk

Isn't

New statistical models for NLP ...

Exciting text domains like **political blogs** ...

Advances in applications like **translation** and **summarization** ...

Is

Bag of terms representation and SVR model.

Boring (to read) text domain of financial reports.

Under-explored application: **forecasting**.

See Also ...

- Lavrenko et al. (2000), Koppel and Shtrimberg (2004), and others: prices
- Blei and McAuliffe (2007): popularity
- Lerman et al. (2008): prediction markets

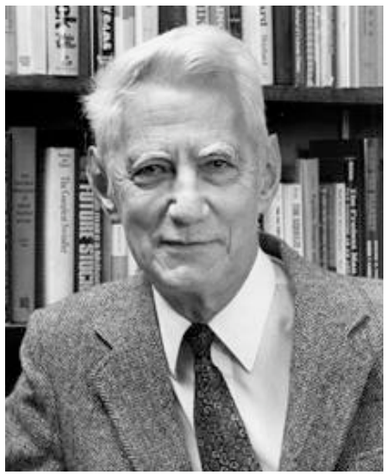
Outline

- Mini-lesson in finance
- A new text-driven forecasting task
- Regression models trained on text
- Experimental results and analysis
- Outlook

Finance

Allocation of wealth (e.g., money) across time and risk (states of nature).

Finance



From an NLP perspective: crucial information about your investments that's buried in documents you'd rather not read.

$$\text{financial risk} = f(\text{financial report})$$

financial risk = f(financial report)

volatility
of returns

What is Risk?

- Return on day t:

$$r_t = \frac{\text{closingprice}_t + \text{dividends}_t}{\text{closingprice}_{t-1}} - 1$$

- Sample standard deviation from day t - τ to day t:

$$v_{[t-\tau,t]} = \sqrt{\sum_{i=0}^{\tau} (r_{t-i} - \bar{r})^2 \Big/ \tau}$$

- This is called measured volatility.

Why Not Predict Returns, Get Rich, Retire Early?

- Hard: predicting a stock's *performance*.
- To predict *returns*, we would need to find new information.
- Our reports probably don't contain new information (10-Ks do not precede big price changes).

Will This Talk Make Anyone Rich?

- Some people think you can exploit accurate volatility predictions.
- I'm not really qualified to give financial advice.
- Consulting to portfolio/wealth managers is a huge industry.

So Then Why Do Finance Researchers Care?

- Models of economics and finance treat information simplistically.
  - No notion of *extracting* information from large amounts of raw *data*.
- These reports are produced at huge expense. Are they worth it?

Important Property of Volatility

- Autoregressive conditional heteroscedacity: volatility tends to be stable (over horizons like ours).
- $V_{[t - \tau, t]}$ is a strong predictor of $V_{[t, t + \tau]}$
- This is our strong baseline.

financial risk = f(financial report)

volatility of returns

Form 10-K, Item 7

Form 10-K, Item 7

General Motors Corp.

March 5, 2009

Item 7. *Management's Discussion and Analysis of Financial Condition and Results of Operations*

Overview

We are primarily engaged in the worldwide production and marketing of cars and trucks. We operate in two businesses, consisting of our automotive operations, which we also refer to as Automotive, GM Automotive or GMA, that includes our four automotive segments consisting of GMNA, GME, GMLAAM and GMAP, and our financing and insurance operations (FIO). Our finance and insurance operations are primarily conducted through GMAC, a wholly-owned subsidiary through November 2006. On November 30, 2006, we sold a 51% controlling ownership interest in GMAC to a consortium of investors. After the sale, we have accounted for our 49% ownership interest in GMAC under the equity method. GMAC provides a broad range of financial services, including consumer vehicle financing, automotive dealership and other commercial financing, residential mortgage services, automobile service contracts, personal automobile insurance coverage and selected commercial insurance coverage.

Automotive Industry

In 2008, the global automotive industry has been severely affected by the deepening global credit crisis, volatile oil prices and the recession in North America and Western Europe, decreases in the employment rate and lack of consumer confidence. The industry continued to show growth in Eastern Europe, the LAAM region and in Asia Pacific, although the growth in these areas moderated from previous levels and is beginning to show the effects of the credit market crisis which began in the United States and has since spread to Western Europe and the rest of the world. Global industry vehicle sales to retail and fleet customers were 67.1 million units in 2008, representing a 5.1% decrease compared to 2007. We expect industry sales to be approximately 57.5 million units in 2009.

Our Corpus

- Edgar database at http://www.sec.gov
- 26,806 examples of Item 7, 1996-2006
- 247.7 million words in total
- http://www.ark.cs.cmu.edu/10K

"Annotation"

- For each report at time t, we gathered
  - "Historical" volatility: $v_{[t - 1y, t]}$
  - "Future" volatility: $v_{[t, t + 1y]}$
- Source: Center for Research in Security Prices U.S. Stocks Databases

Methodology

- *Input:* Item 7 and/or historical volatility
- *Output:* predicted future volatility
- Test on (input, output) pairs from year Y
- Train on (input, output) from years < Y
- *Evaluation:* MSE of (log) volatility

financial risk = f(financial report)

volatility of returns

SV regression

Form 10-K, Item 7

Support-Vector Regression

(Drucker et al., 1997)

- Predicted future volatility is a function of a document (Item 7), **d**, and a weight vector **w**:

$$\hat{v} = f(\mathbf{d}; \mathbf{w})$$

- The training criterion:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{i=1}^{N} \max\left(0, \left|v_i - f(\mathbf{d}_i; \mathbf{w})\right| - \epsilon\right)$$

regularize

prediction within **ε** of correct

Representation

$$f(\mathbf{d}; \mathbf{w}) = h(\mathbf{d})^\top \mathbf{w}$$

- Vector-space model (tf, tfidf, etc.)
- So far, unigrams and bigrams
- Linear kernel (for interpretability)

Representation

$$f(\mathbf{d};\mathbf{w}) = h(\mathbf{d})^\top \mathbf{w} = \sum_{i=1}^{N} \alpha_i K(\mathbf{d}, \mathbf{d}_i) = \sum_{i=1}^{N} \alpha_i h(\mathbf{d})^\top h(\mathbf{d}_i)$$

- Vector-space model (tf, tfidf, etc.)
- So far, unigrams and bigrams
- Linear kernel (for interpretability)

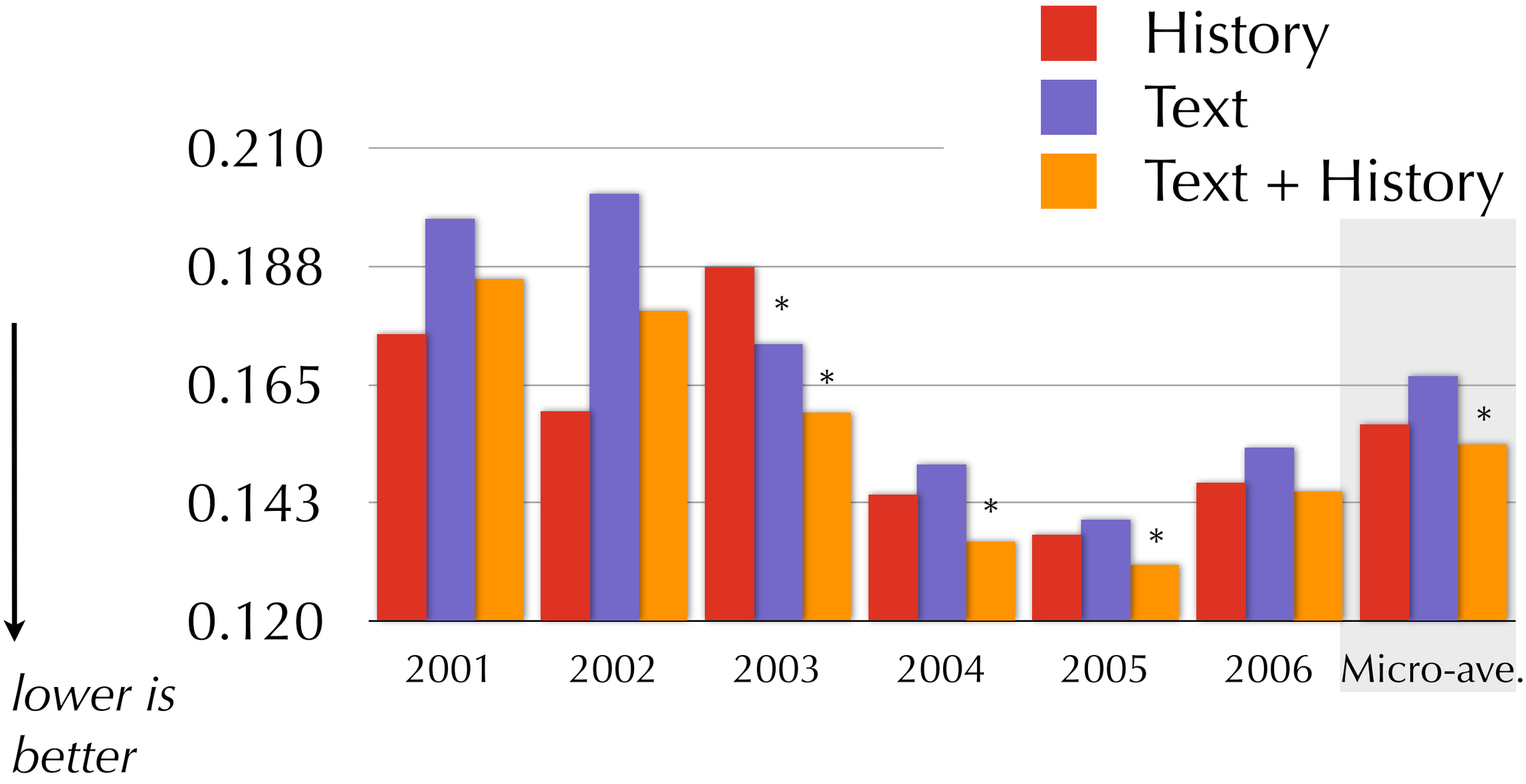
dual

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i h(\mathbf{d}_i)$$

Experiment

- Test on year Y.
- Train on (Y - 5, Y - 4, Y - 3, Y - 2, Y - 1).
- Six such splits.
- Compare history-only baseline, text-only SVR, combined SVR.

MSE of Log-Volatility



Using "log(1+freq.)" representation on *all* unigrams and bigrams. See paper.

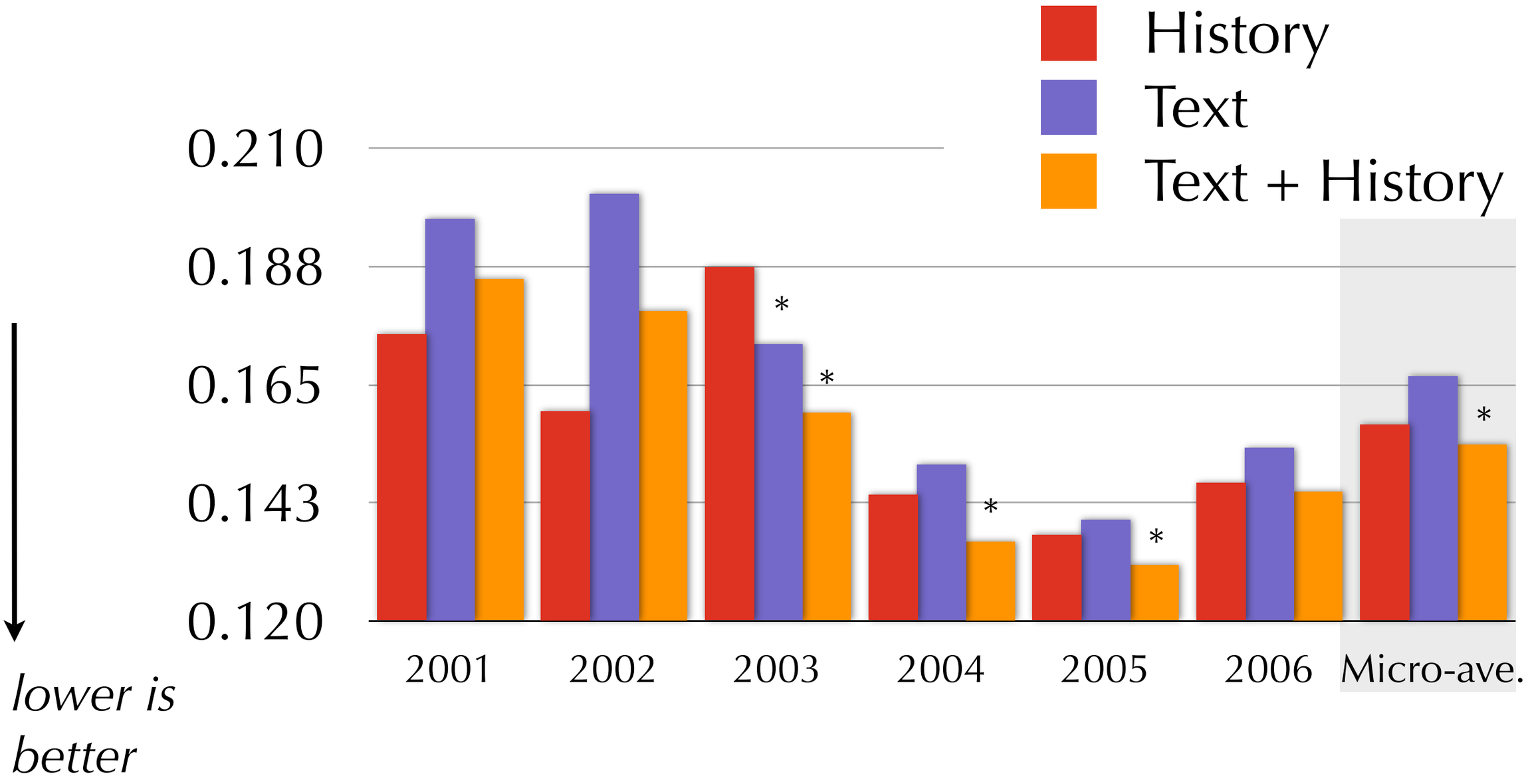
Dominant Weights (2000-4)

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 words

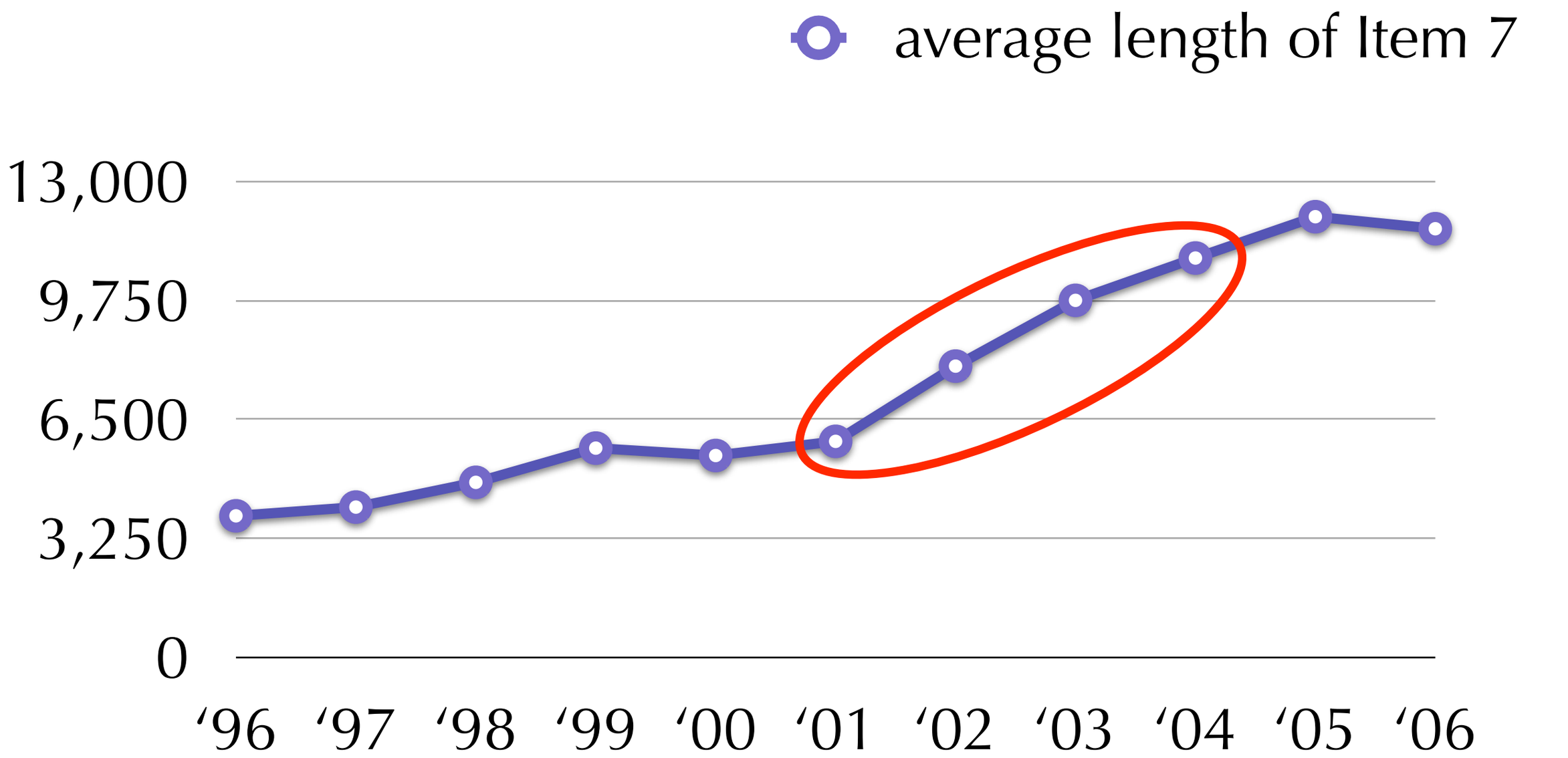
low volatility words

MSE of Log-Volatility



Using "log(1+freq.)" representation on *all* unigrams and bigrams. See paper.

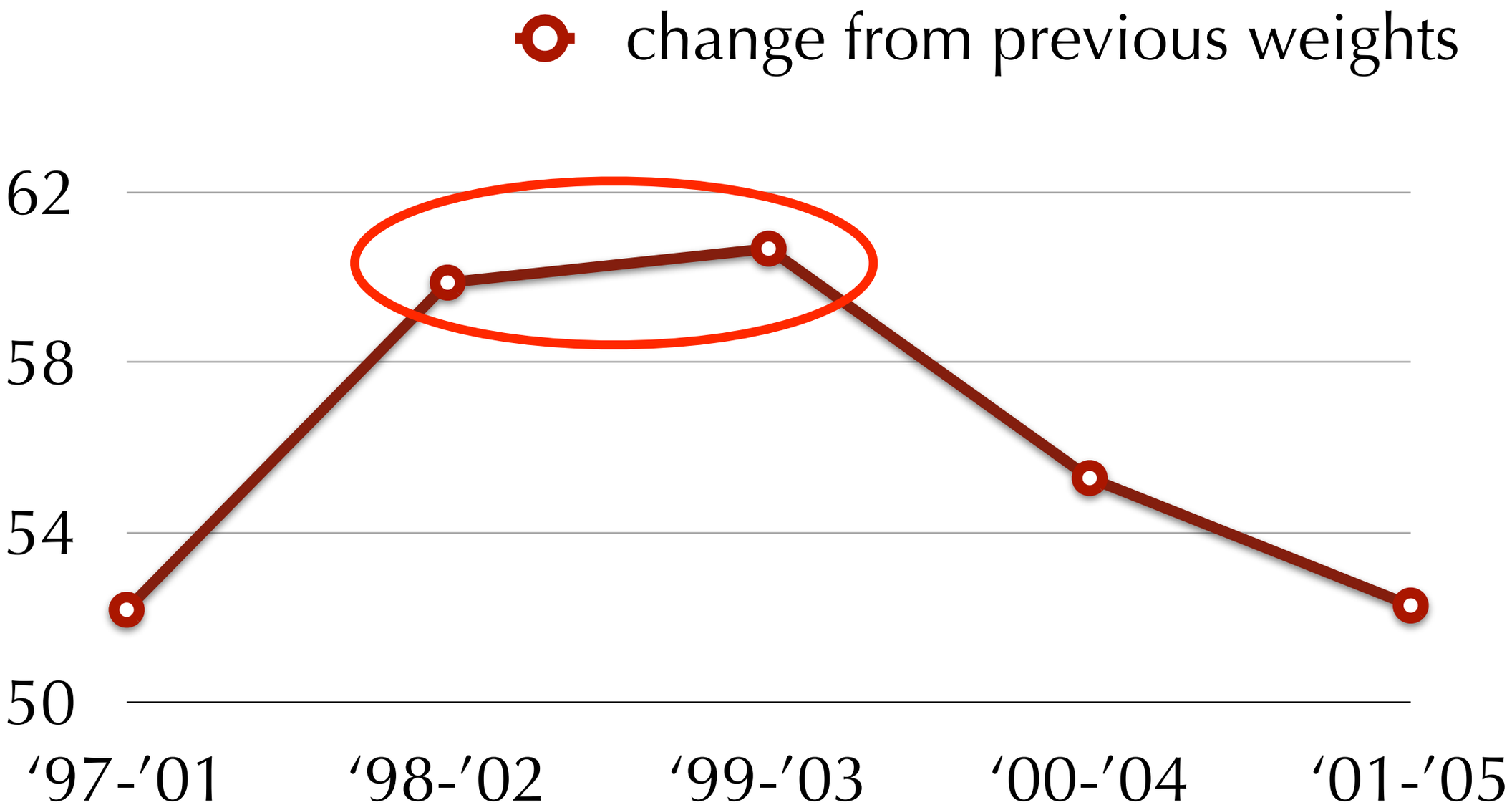
Changes Over Time



2002

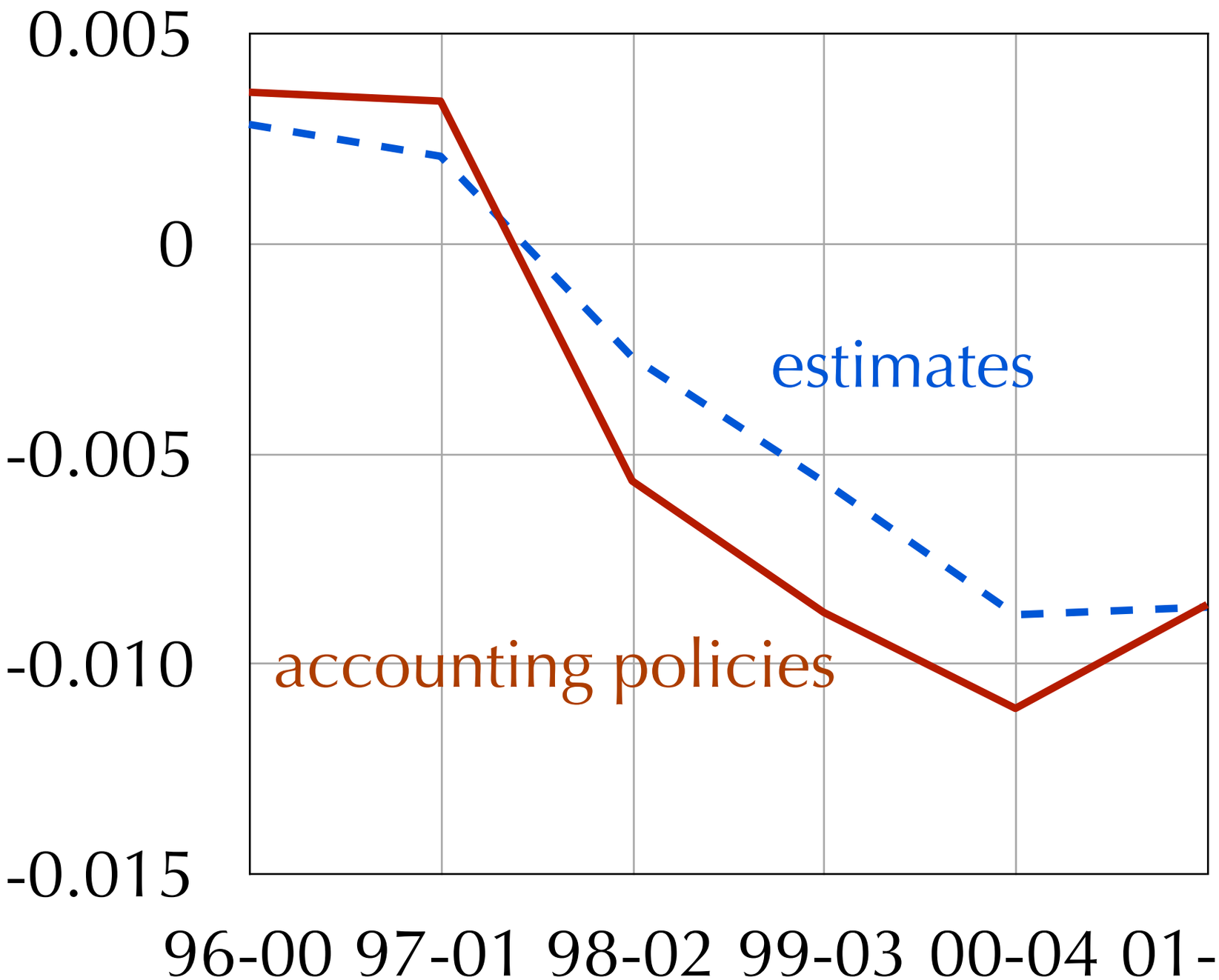
- Enron and other accounting scandals
- Sarbanes-Oxley Act of 2002
- Longer reports
- Are the reports more informative after 2002? Because of Sarbanes-Oxley?

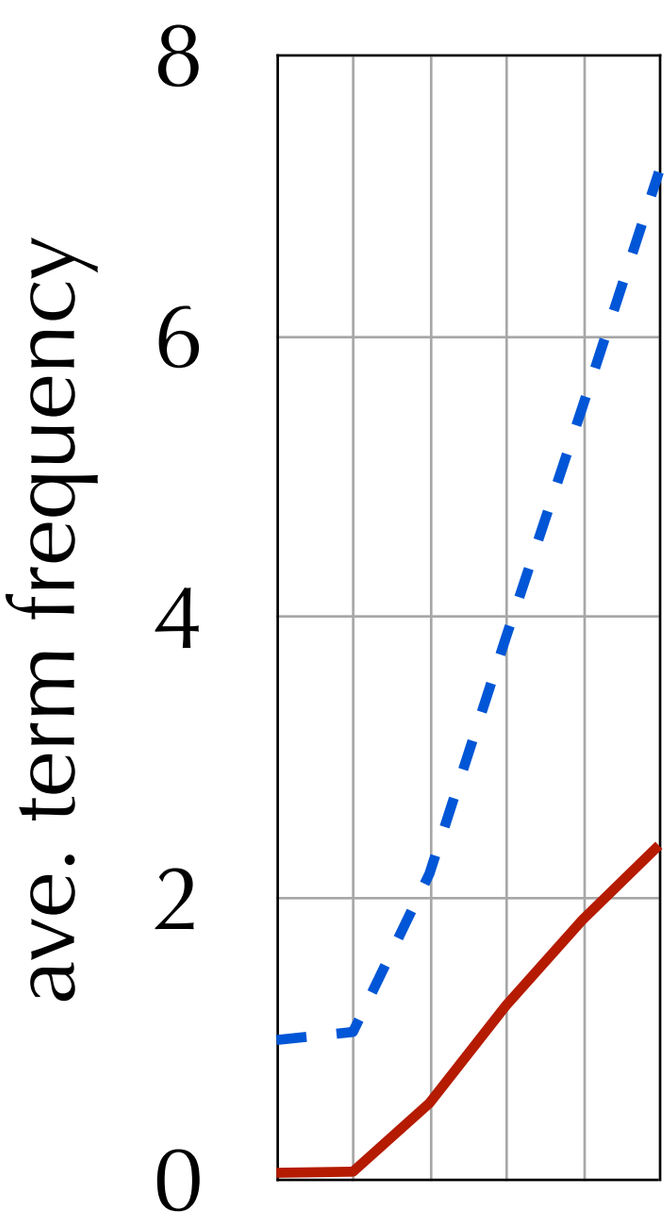
Changes In **w**



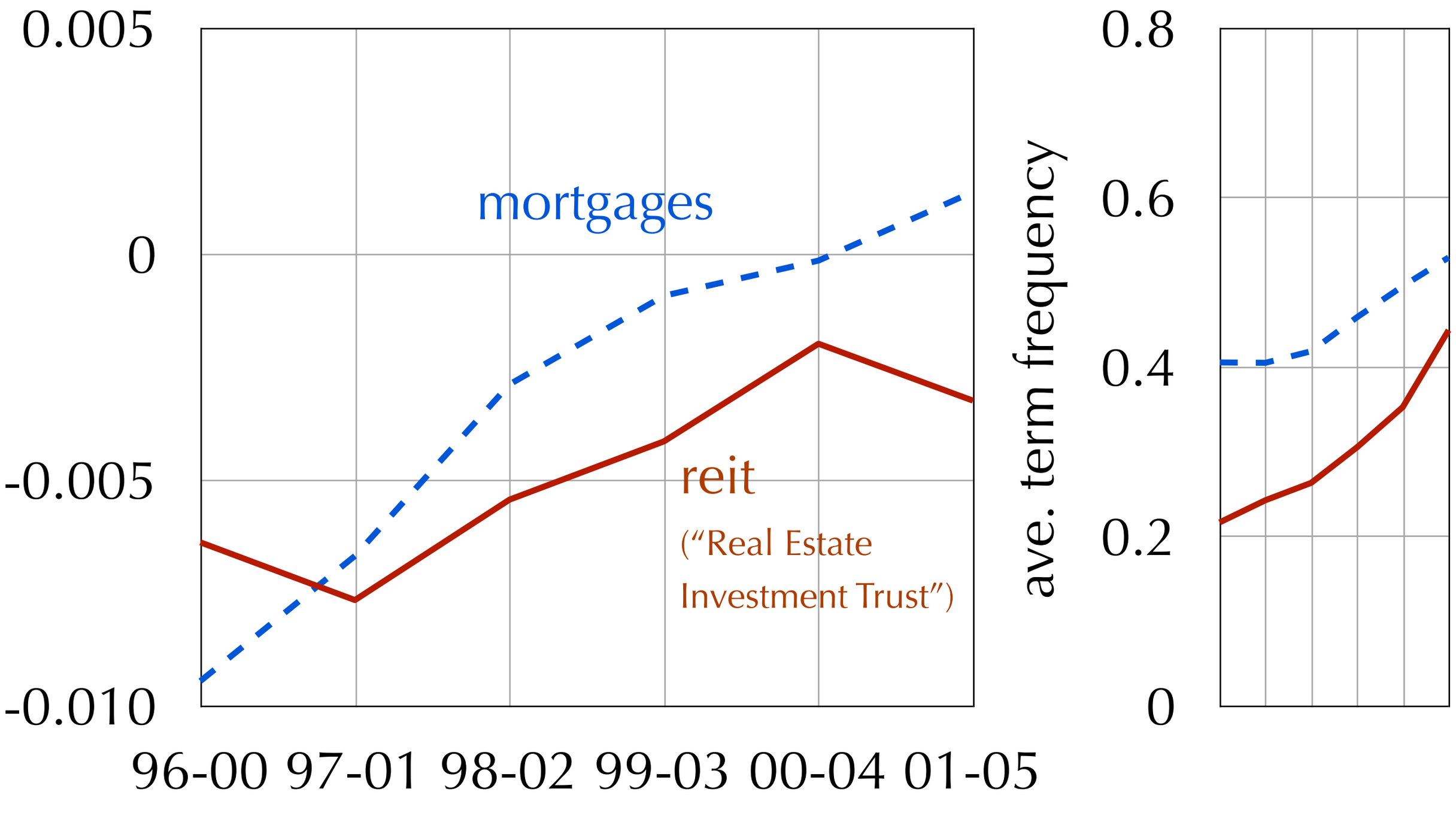
Measured in L_1 distance; based on unigram model with “log(1 + freq.)” representation.

Language Over Time

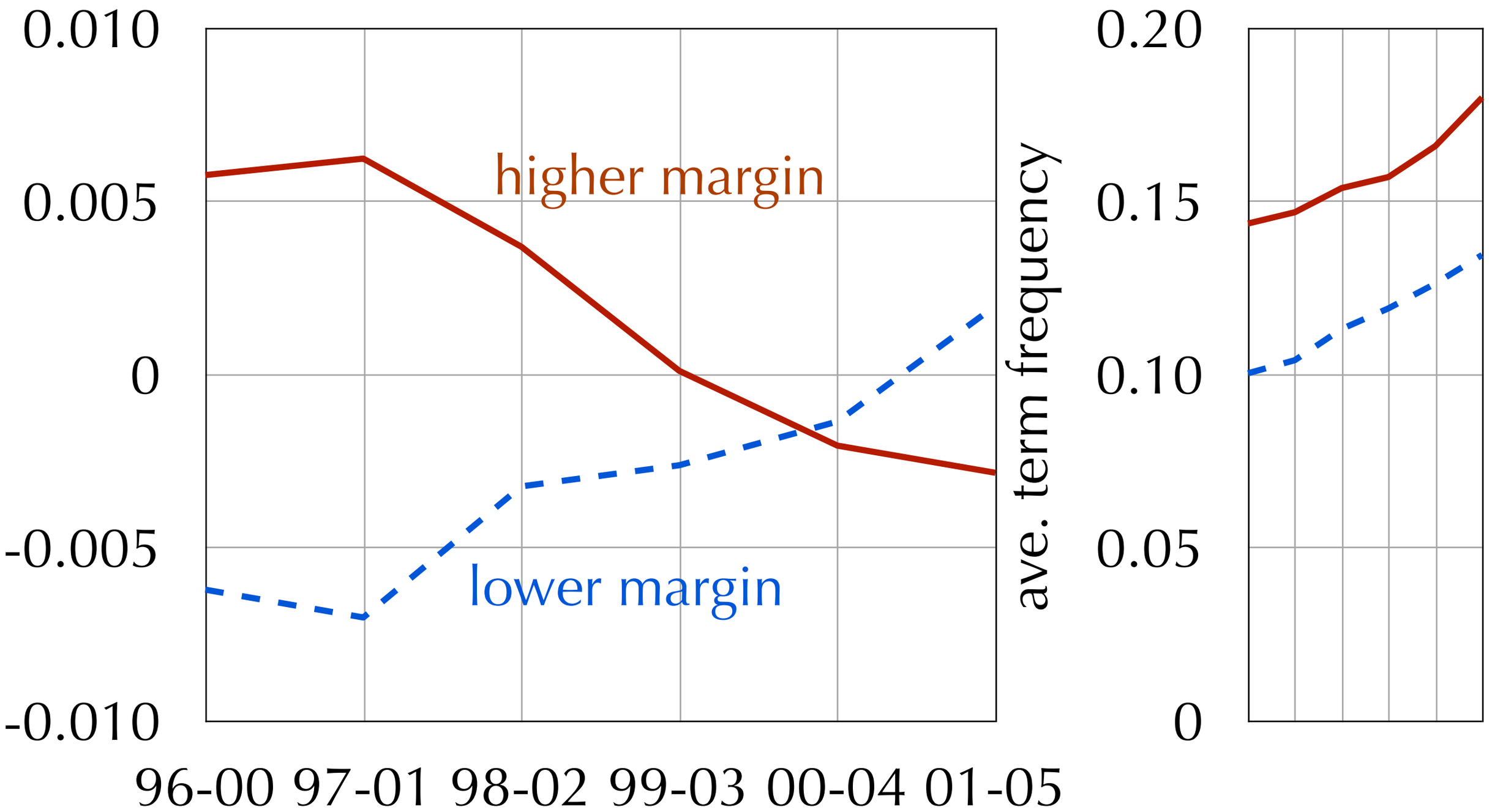




Language Over Time

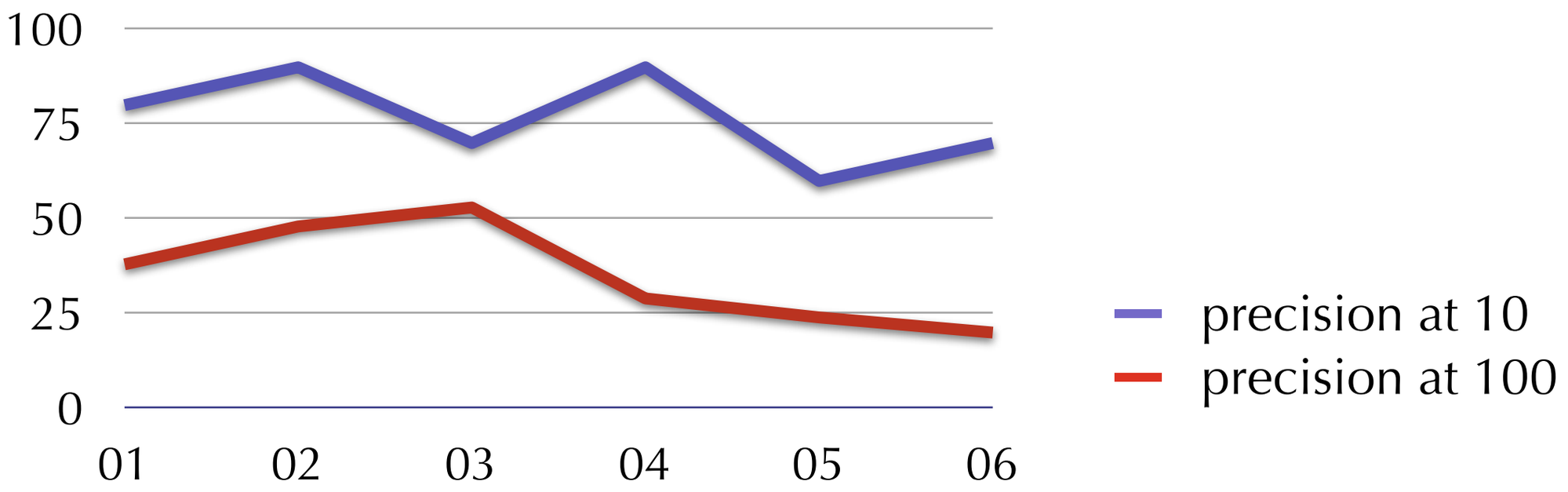


Language Over Time



Delisting

- Rare (4%) event: **delisting** due to dissolution after bankruptcy, merger, violation of rules.
- bulletin, creditors, dip, otc, court



Conclusions

- Text-driven forecasting of volatility, by regression.
  - Works nearly as well as strong history predictor.
  - Often works better in combination.
- Suggestion of effects of legislation on a real-world text-generating process.

Future Work

- Measuring the effect of Sarbanes-Oxley
- Other predictions
- Other text representations
- Other datasets

Future Work (Text-Driven Forecasting)

- Application for NLP: techniques that use text to make real-world predictions.
- Many potential domains (finance, politics, government, sales, ...)
- There's lots of room for improvement!