A Better Wordscores: Scaling Text with the Class Affinity Model

Kenneth Benoit
London School of Economics
(and Australian National University)

Joint work with Patrick Perry (Oscar Health)

General Problem

We need to measure latent traits about authors from the contents of their texts:

- ideology from political documents (e.g. everybody)
- Government v opposition from legislative speeches (e.g. Herzog and Benoit 2015)
- Pro- v anti-petitioner from amicus briefs (e.g. Evans et al. 2007)
- Leave v Remain in Brexit (e.g. Amador-Lopez et al 2017)
- sentiment (e.g. half of everybody - more on this later)

and

We would like to **know what we are measuring** through using a supervised approach

Not to for prediction or classification, but for **measurement**

Specific Problem

"The Wordscores algorithm, by Laver, Benoit, and Garry (2003), represents a true breakthrough in the use of text as data in political science…

But facets of wordscores constrain the method and make it difficult to recommend for general use (see Lowe (2008) for an extended critique)."

- Grimmer, Roberts and Stewart (2022)

(By the way)
Running example

The Irish no-confidence debate from Laver, Benoit and Garry (2003) in which 56 speeches expressed support for or opposition to the government

"Trained" using party leader speeches

Classification v. Scaling

Classification Models	Reality
Every text has a "true" label (possibly unobserved)	Every text expresses a mixture of viewpoints
The world is black-and-white	The world is gray
Estimate P(black)/P(white)	Estimate degree of grayness

Votes: Uninformative

Party	Vote to Oppose	Vote to Support
Fianna Fáil	0	24
Progressive Dems.	0	1
Democratic Left	3	0
Fine Gael	22	0
Green	1	0
Labour	7	0

A Better Predictor

$$\hat{y} = \begin{cases} \text{Govt} & \text{if Party} = \text{FF or PD} \\ \text{Opp} & \text{otherwise} \end{cases}$$

Misclassification rate: 0%

(Classification is not an interesting problem)

Conceptual Model

Treat each text as a mixture of positions

1. Over the course of a speech, a speaker's orientation switches back and forth between Govt mode and Opp mode

2. When she is in Govt mode, she picks random words according to the Government probabilities

3. When she is in Opp mode, she picks random words according to the Opposition probabilities

Class Affinity Model: Over the course of a speech, the underling orientations switches back and forth between Govt and Opp.

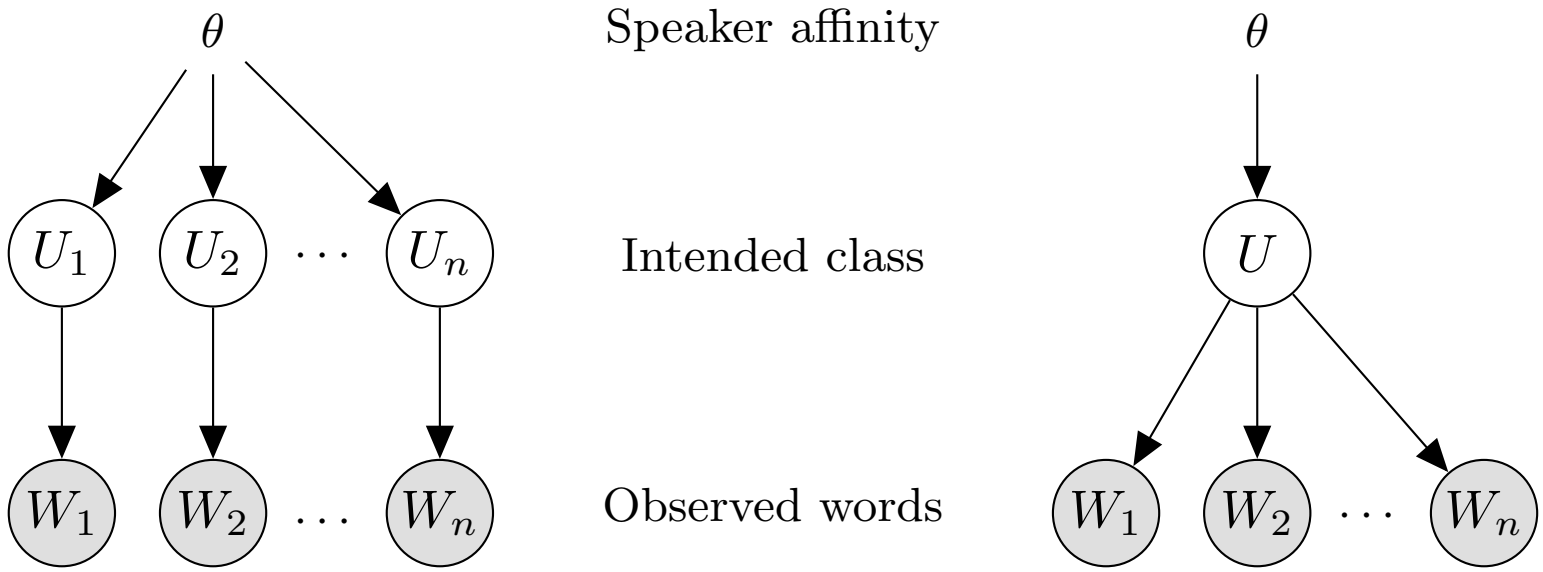
Ex. (Blaney, FF): *confidence in this Government that appears to be related… to the recent sandals, scams or allegations of such… These have not yet been proven. Nor has there been any proof… to connect the Government or their Ministers with any wrongdoing… it looks fairly likely that some of them will be proven. On the other hand there has been no evidence of Governmental or ministerial involvement*

Interpretable Parameters

- Assume that the speaker randomly picks a different orientation at each position in her speech
- Assume that orientations at different positions are independent
- Parameters ("Affinities"):

$$\theta_1 = \Pr(\text{mode} = \text{Govt})$$
$$\theta_2 = \Pr(\text{mode} = \text{Opp})$$

Generative Model



(a) Class affinity model

(b) Classification Model

Fig 2: Generative model for the underlying orientation U and the token sequence W, contrasting the class affinity model to the classification model.

General Model (K Classes)

Tokenized Text: $(W_1, W_2, \ldots, W_n)$

Underlying Orientation: $(U_1, U_2, \ldots, U_n)$

$$\Pr(W_i = v) = \sum_{k=1}^{K} \Pr(U_i = k) \Pr(W_i = v \mid U_i = k)$$

$$= \sum_{k=1}^{K} \theta_k \, p_{kv}.$$

(cf. Blei, Ng, and Jordan 2003)

Debate Speech (K = 2)

$$\theta_1 = \Pr(U_i = \text{Govt})$$

$$\theta_2 = \Pr(U_i = \text{Opp})$$

Likelihood-based Estimation

Given reference distributions $p_1, p_2, \ldots, p_K$ *:*

Pr(word v appears) $\mu_v(\theta) = \sum_{k=1}^{K} \theta_k p_{kv}$

Assuming independence:

Log likelihood $\ell(\theta) = \sum_{v \in \mathcal{V}} x_v \log\{\mu_v(\theta)\}$

(fit separately for each speech)

Special Case: K = 2

Reparametrization: $\beta = \theta_2 - \theta_1$

Score Function:

$$u(\beta) = \ell'(\beta)$$
$$= \frac{1}{2} \sum_{v \in \mathcal{V}} \frac{p_{2v} - p_{1v}}{\mu_v} x_v$$

Information Function:

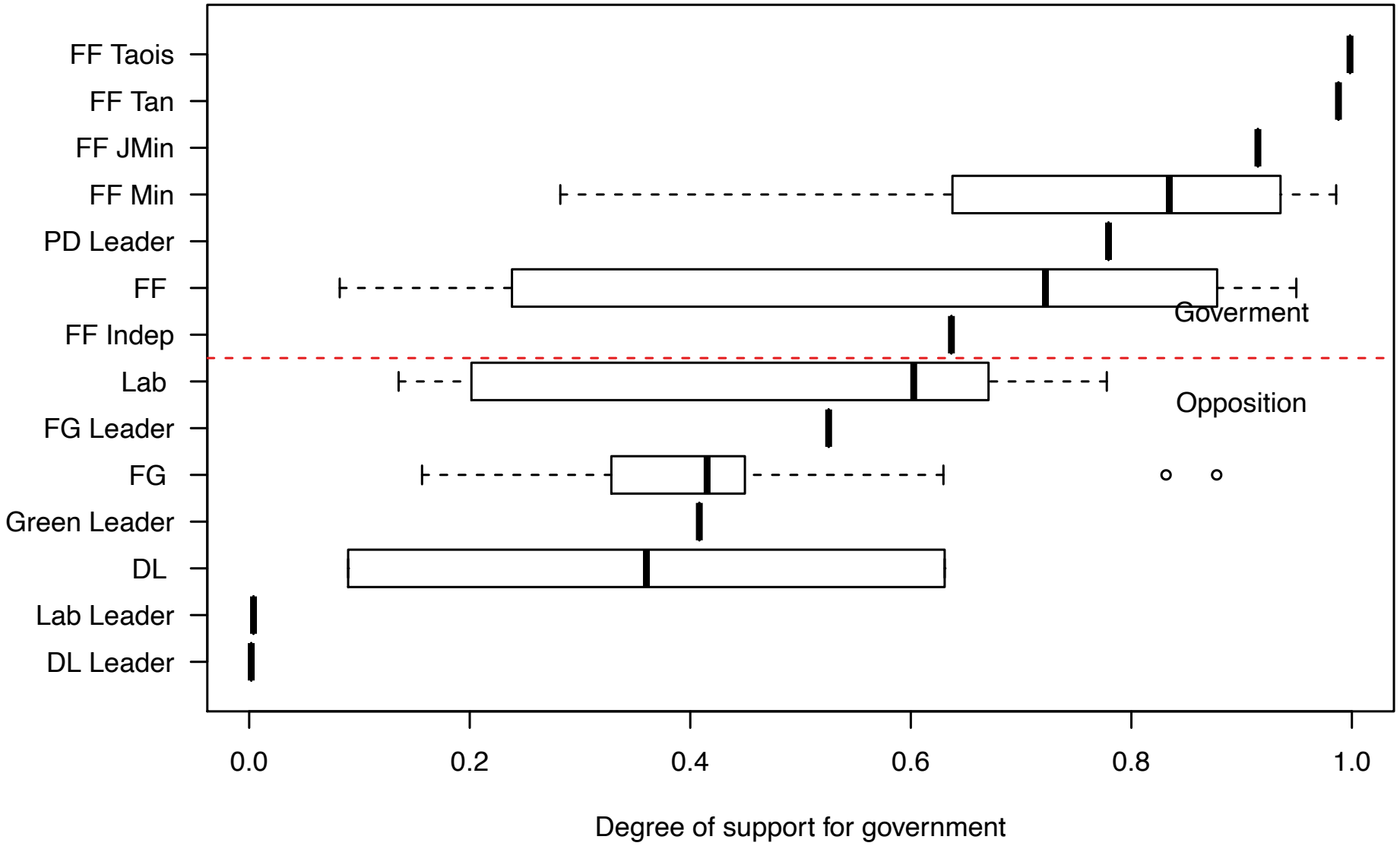
$$\mathcal{I}(\beta) = -\ell''(\beta)$$
$$= \frac{1}{4} \sum_{v \in \mathcal{V}} \frac{(p_{2v} - p_{1v})^2}{\mu_v^2} x_v$$

Maximization

$$\hat{\beta}^{(0)} = 0$$

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + [\mathcal{I}(\hat{\beta}^{(t)})]^{-1} u(\hat{\beta}^{(t)})$$

- The likelihood is log-concave, and so can be maximized efficiently (Newton-Raphson)
- Computing the Hessian takes time $O(V K^2)$
- In practice about 5 iterations suffice

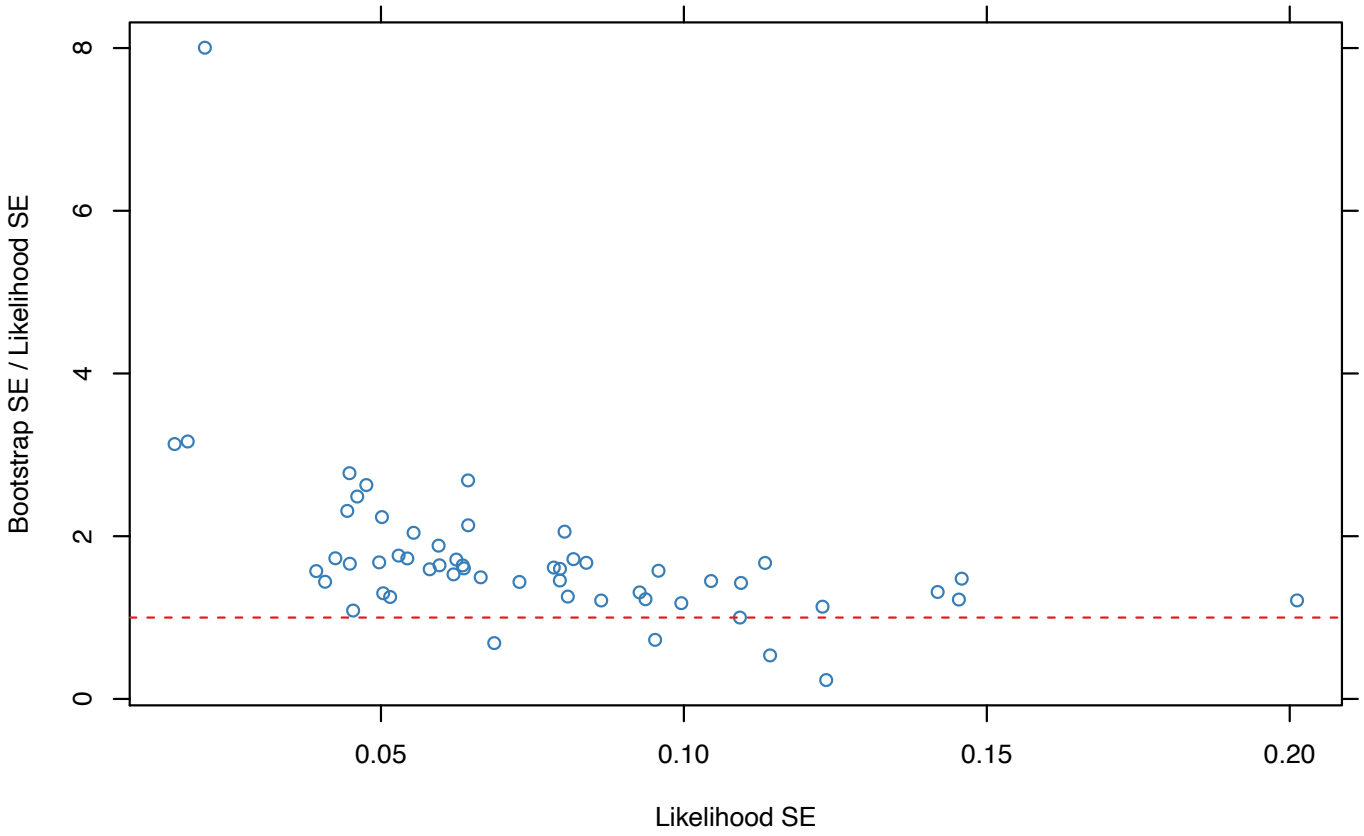


Application: It works

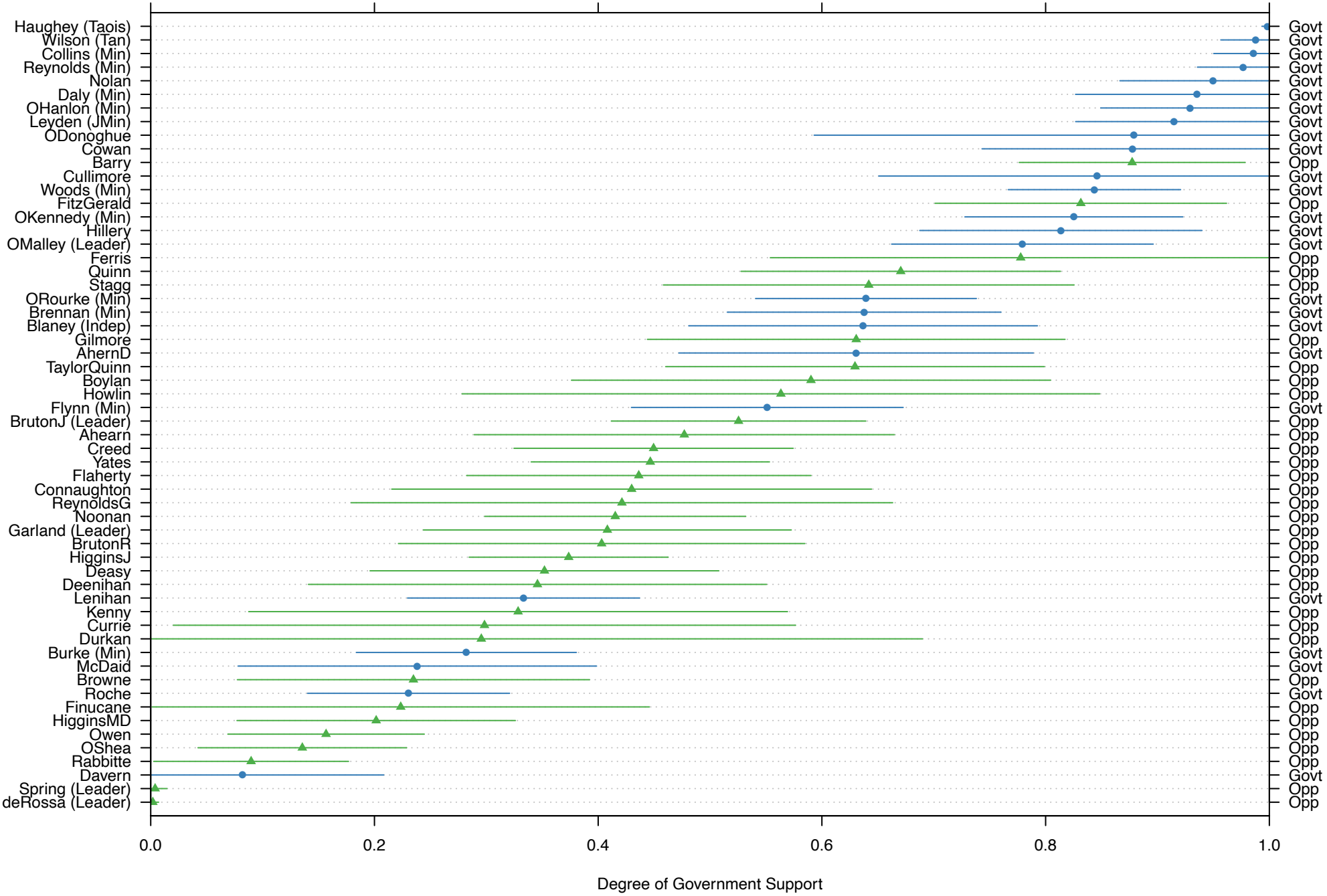
Uncertainty Quantification

- The standard errors gotten from the Fisher information rely on the "bag of words" assumption and are likely too narrow
- For better uncertainty estimates, we use a block bootstrap, resampling sentences

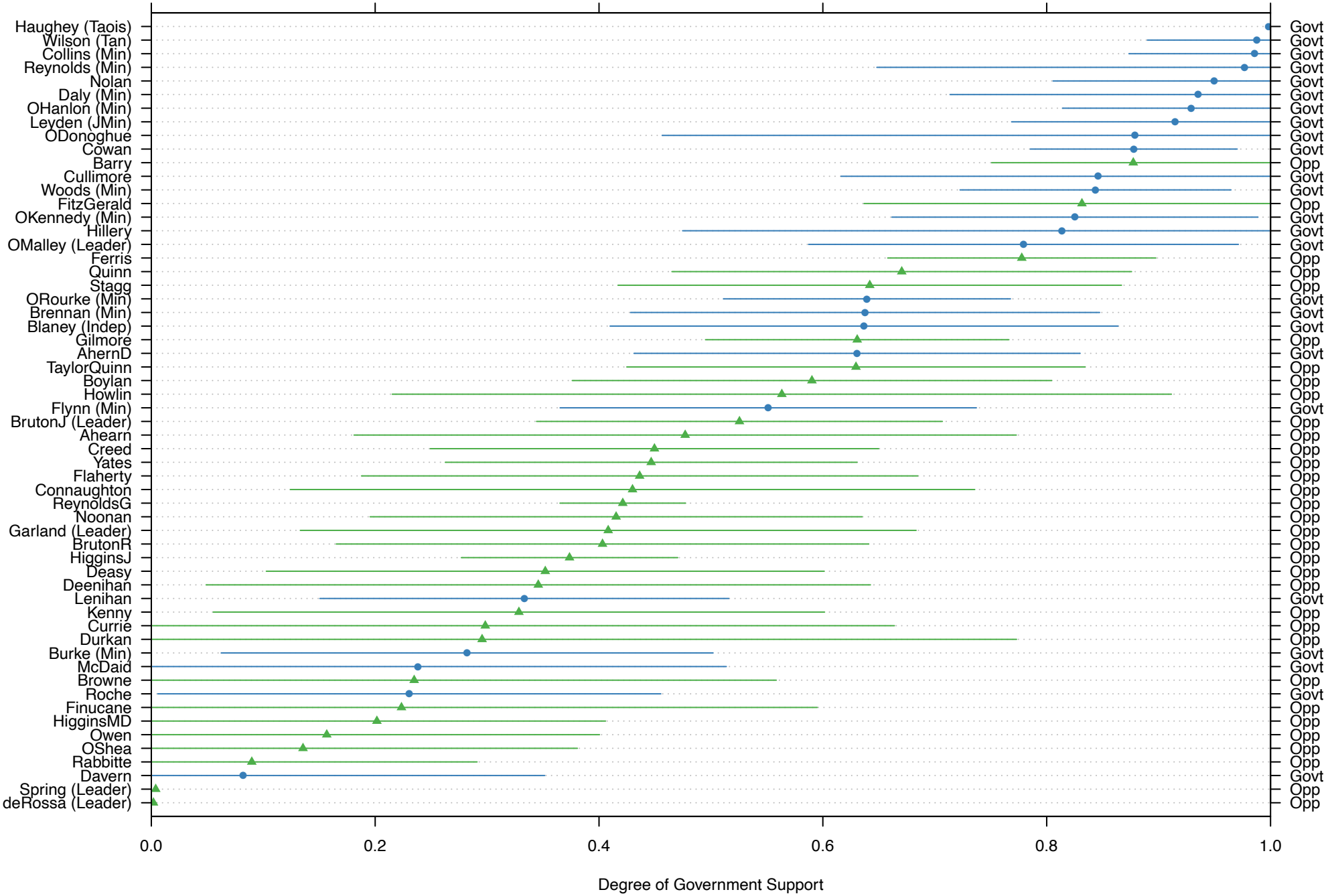
Bootstrap SEs



(Resample *sentences*, not tokens)



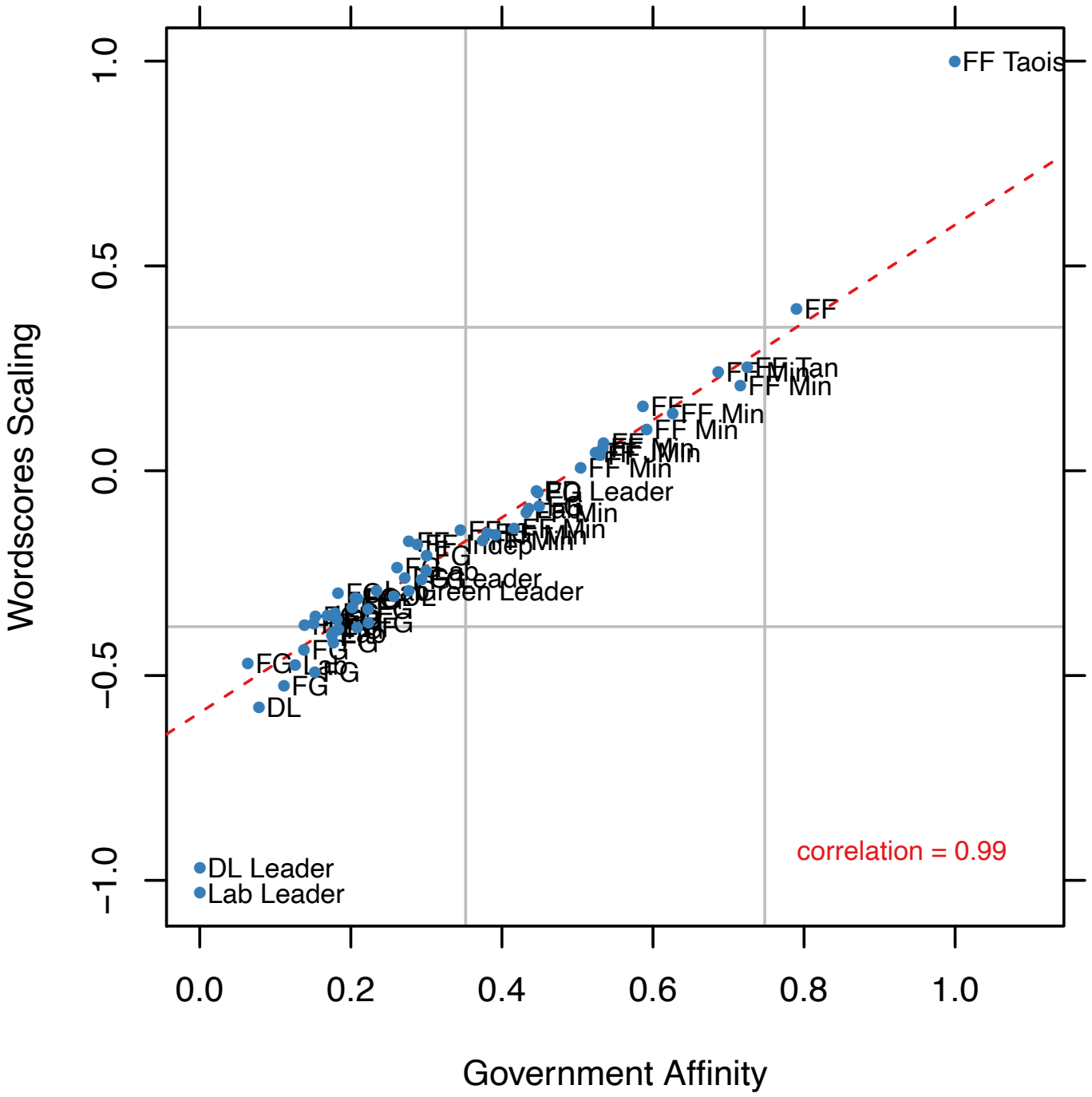
(Likelihood CIs)

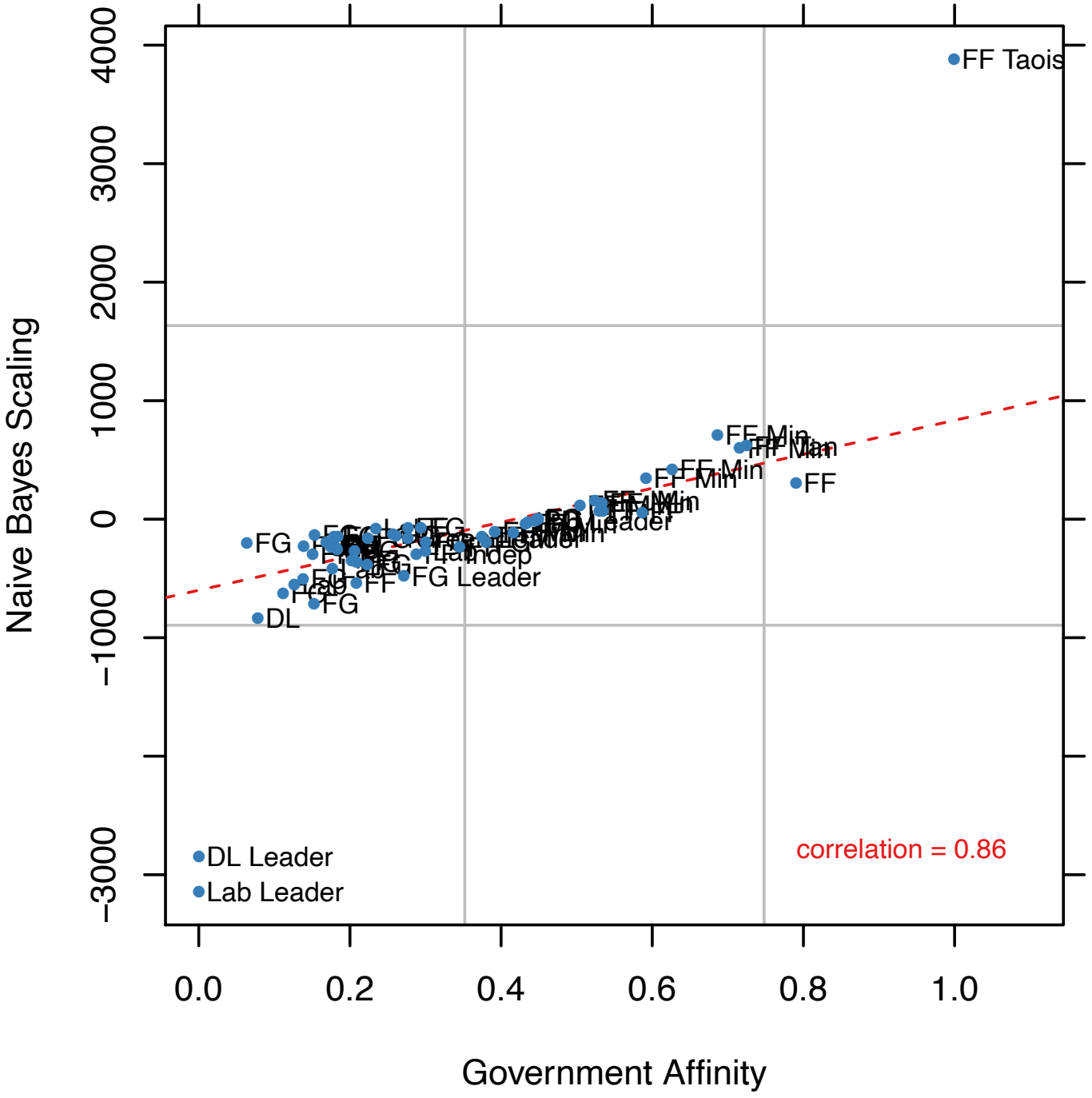


(Bootstrap CIs)

Connections to Other Methods

- If the the Government and Opposition reference distributions have disjoint supports, the method is exactly equivalent to dictionary scaling
- For moderate texts, this scaling highly correlated with Wordscores (Laver, Benoit, and Garry 2003)
- When the training set is small, this scaling highly correlated with scores from sLDA and other supervised topics models (Blei and McAuliffe 2008; Ramage et al. 2009; Ramage, Manning and Dumais 2011)


Wordscores Scaling
Government Affinity
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0
FF Taois
FF
FF Tan
FF Min
FF Min
FF Min
FF Min
FF Min
Leader
Min
Min
Leader
reen Leader
FG
FG
FG
FG
DL
DL Leader
Lab Leader
correlation = 0.99


Naive Bayes Scaling
4000
3000
2000
1000
0
−1000
−3000
0.0
0.2
0.4
0.6
0.8
1.0
Government Affinity
FF Taois
FF
FG
DL
FG Leader
DL Leader
Lab Leader
correlation = 0.86

Extensions

In the paper, we also develop and apply methods for:

1. Diagnostics
2. Feature selection
3. Uncertainty quantification

Style v. Substance

A common criticism of Wordscores, which excluded nothing. But this was not an essential component of the algorithm, but rather the way it was applied.

Solution: We define an influence measure similar to Cook's residual influence measure, but more computationally efficient

But in short: We recommend excluding common words (stop words) and happy legemona

Conclusions

- Generative language model whose parameters can be estimated, not an algorithmic approach
- Likelihood estimation with adjustable but robust parameters
- Avoids distortions of extreme texts
- Computationally efficient
- We provide additional methods for feature selection and uncertainty quantification

RIP Wordscores: 2003-2022

- Class Affinity Scaling is **always better than Wordscores**
  (and always better than dictionaries)
- Use it with two contrasting classes
- Remove stopwords and hapax legemona
  (and look at diagnostics)

Next steps

1. Political science delivery (suggestions welcome!)
2. Improving software

Implementation

R package **quanteda.textmodels**

```
textmodel_affinity(x, y, exclude = NULL,
                   smooth = 0.5, ref_smooth = 0.5)
```

Thank you

(in advance for the citations)

Outtakes

General Case

Reparametrization: $\beta \in \mathbb{R}^{K-1}$

$$\theta = \theta_0 + C\beta$$

$\theta_0 = (1/K, 1/K, \ldots, 1/K)$

(Center of parameter space)

$C \in \mathbb{R}^{K \times (K-1)}$

$C^T 1 = 0$

(Contrast matrix)

Text Classification

x = text

y = label (Govt/Opp)

Works for Authorship, Spam, and Sentiment…
why not Ideology?

(cf. Mosteller and Wallace 1963; Pang, Lee, and Vaithyanathan 2002; Yu, Kaufmann, and Diermeier 2008; Taddy 2013, …)

Naive Bayes Classification

- Model: tokens are chosen from a multinomial on V categories (V = vocabulary size)
- p_1 = Government usage probabilities
- p_2 = Opposition usage probabilities

$$\log \frac{\Pr(y = \text{Govt} \mid x)}{\Pr(y = \text{Opp} \mid x)} = \sum_{v \in \mathcal{V}} x_v \log(p_{1v}/p_{2v})$$

Naive Bayes Predictor: $$\hat{y} = \begin{cases} \text{Govt} & \text{if } \eta(x) > 0 \\ \text{Opp} & \text{otherwise} \end{cases}$$

Naive Bayes Scaling: $$\eta(x) = \sum_{v \in \mathcal{V}} x_v \log(p_{1v}/p_{2v})$$

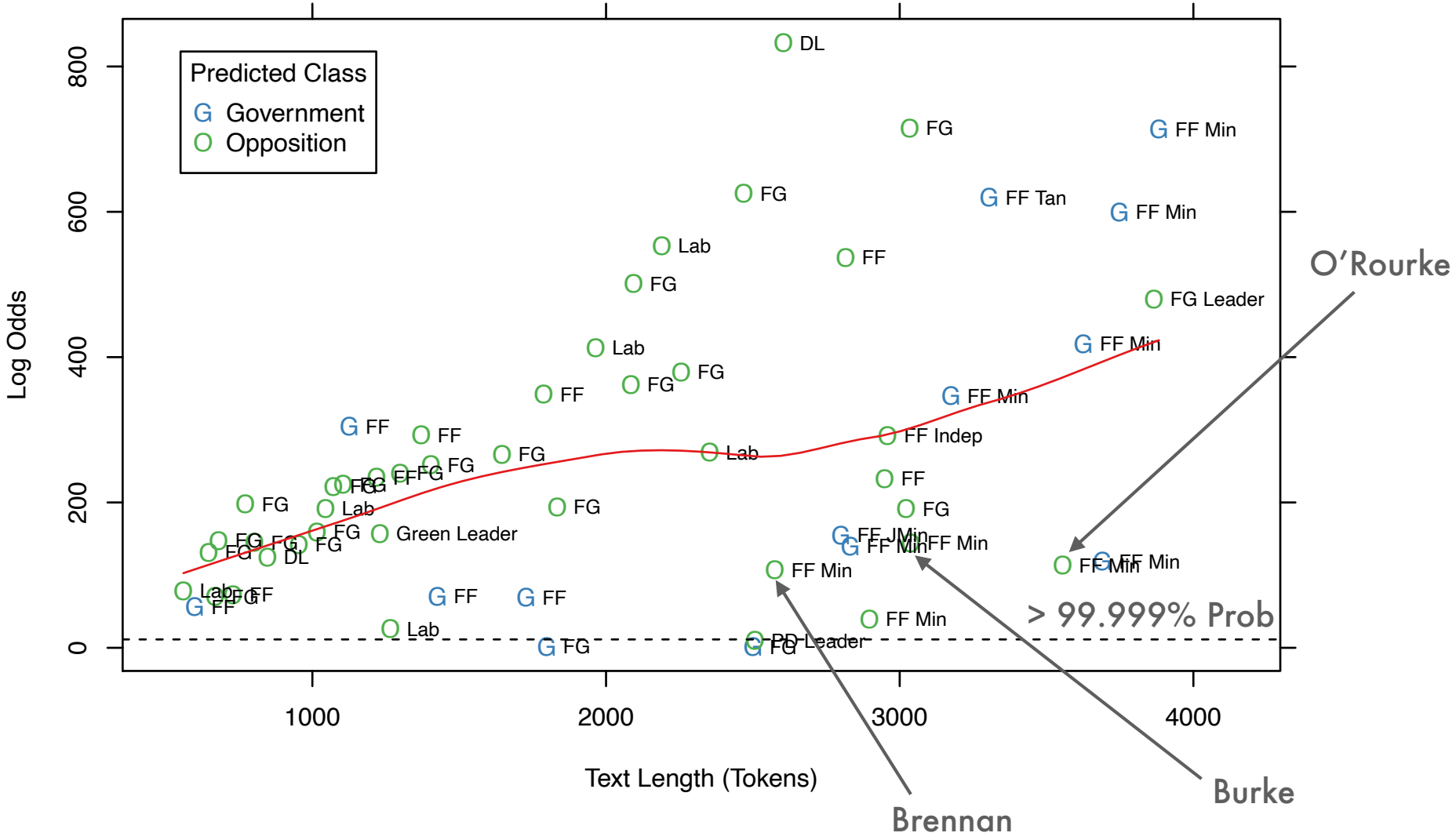
Brennan (FF Minister): *This is a strong, competent and determined Government*

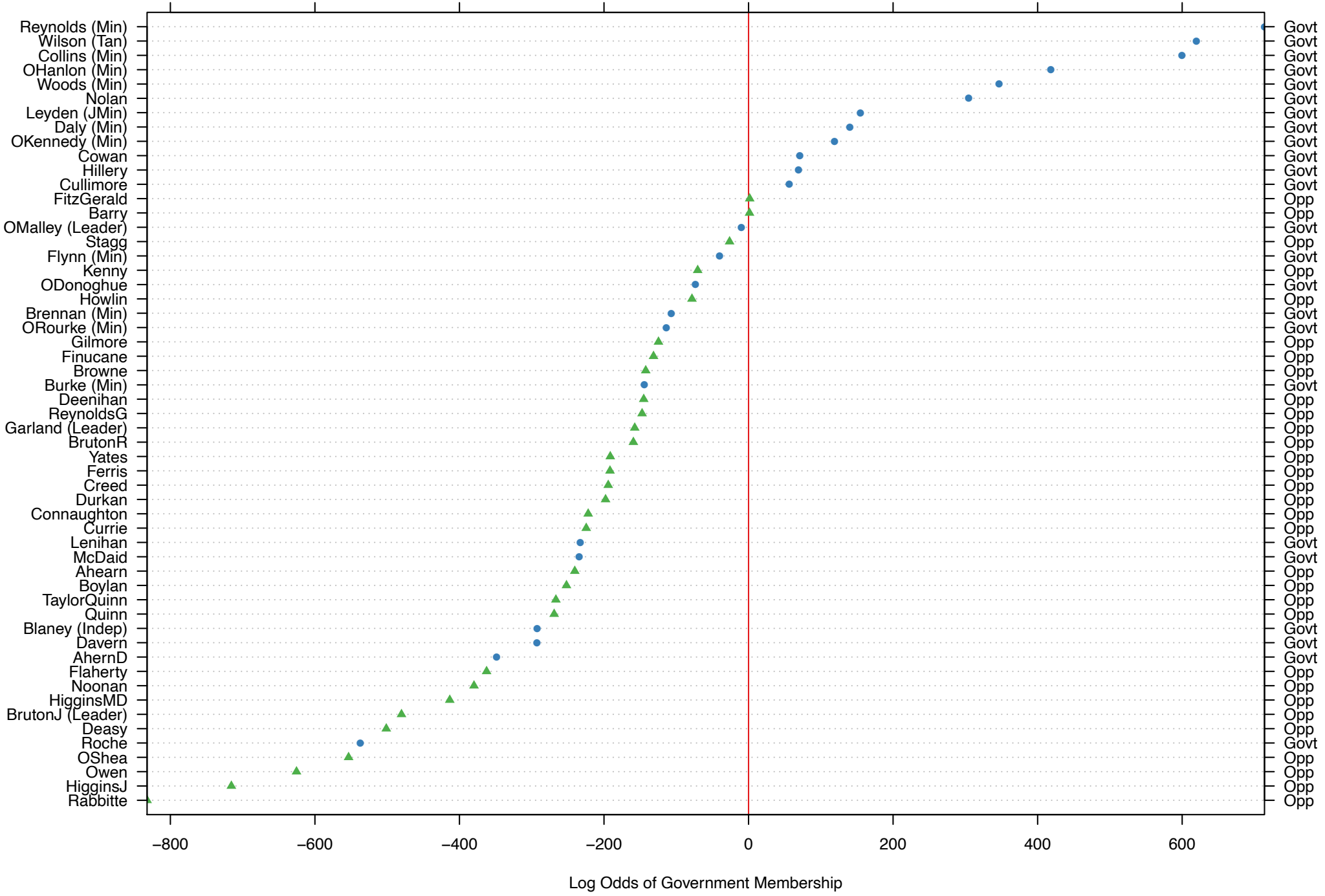
Burke (FF Minister): *I have no hesitation in urging this House to declare its confidence in this Government*

O'Rourke (FF Minister): *I would like to place on record the outstanding achievements of this Government*

Naive Bayes says P(Opp) > 99.999999999999%

Naive Bayes Behavior




Reynolds (Min) Govt
Wilson (Tan) Govt
Collins (Min) Govt
OHanlon (Min) Govt
Woods (Min) Govt
Nolan Govt
Leyden (JMin) Govt
Daly (Min) Govt
OKennedy (Min) Govt
Cowan Govt
Hillery Govt
Cullimore Govt
FitzGerald Opp
Barry Opp
OMalley (Leader) Govt
Stagg Opp
Flynn (Min) Govt
Kenny Opp
ODonoghue Govt
Howlin Opp
Brennan (Min) Govt
ORourke (Min) Govt
Gilmore Opp
Finucane Opp
Browne Opp
Burke (Min) Govt
Deenihan Opp
ReynoldsG Opp
Garland (Leader) Opp
BrutonR Opp
Yates Opp
Ferris Opp
Creed Opp
Durkan Opp
Connaughton Opp
Currie Opp
Lenihan Govt
McDaid Govt
Ahearn Opp
Boylan Opp
TaylorQuinn Opp
Quinn Opp
Blaney (Indep) Govt
Davern Govt
AhernD Govt
Flaherty Opp
Noonan Opp
HigginsMD Opp
BrutonJ (Leader) Opp
Deasy Opp
Roche Govt
OShea Opp
Owen Opp
HigginsJ Opp
Rabbitte Opp
−800
−600
−400
−200
0
200
400
600
Log Odds of Government Membership

Big Picture: Scaling Text

- Political Speech
- Legal Briefs
- Product Reviews
- News Articles
- Financial Reports
- Party Manifestos
- Meeting Transcripts
- Social Media Posts
- Medical Records
- etc.

Shades of Gray

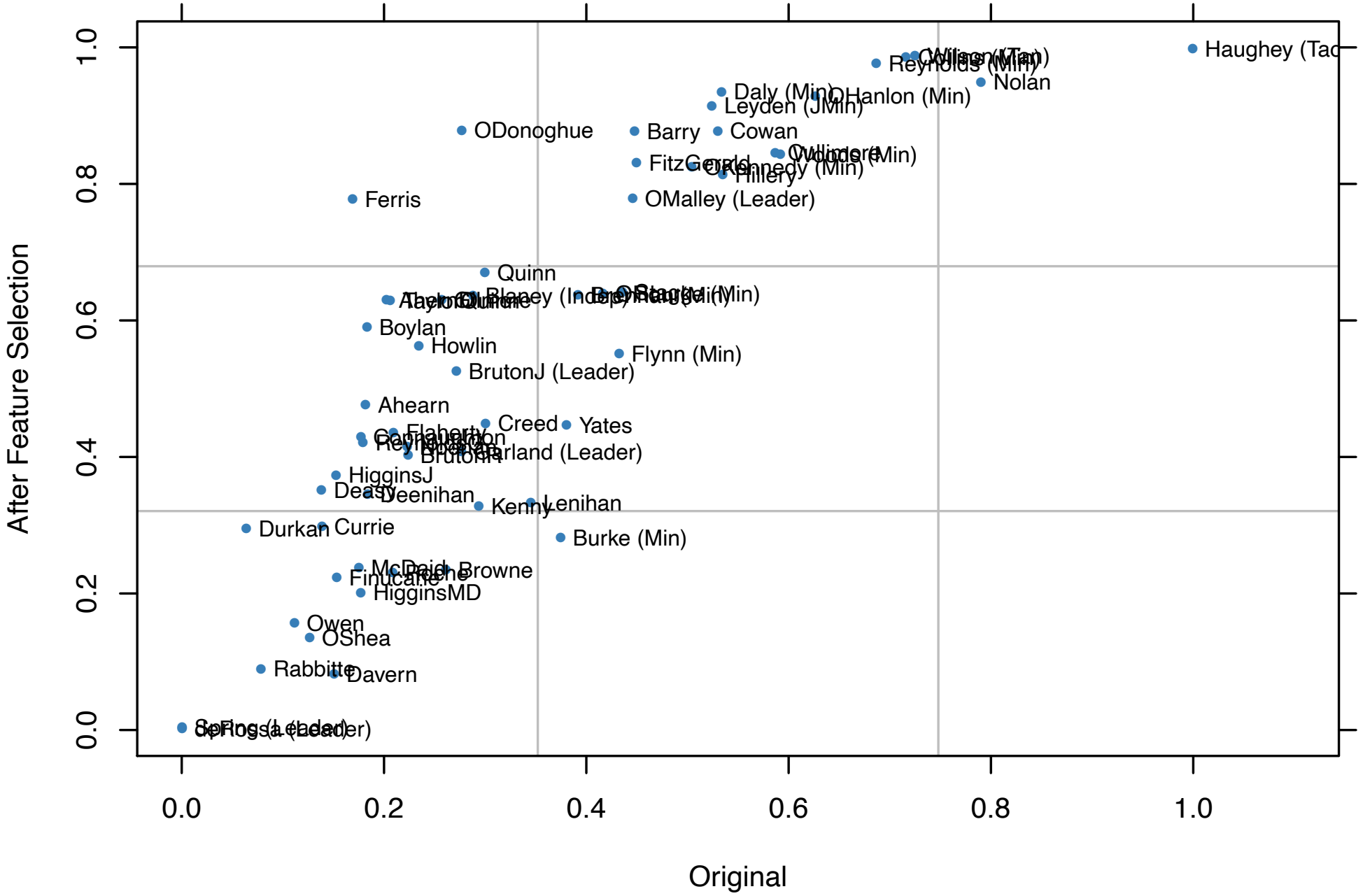
1. a statistical model with interpretable parameters
2. an efficient fitting procedure
3. diagnostics and uncertainty quantification

Implemented in **quanteda** *as* textmodel_affinity()

Thank You!

Votes: Uninformative

Party	Vote to Oppose	Vote to Support
Fianna Fáil	0	24
Progressive Dems.	0	1
Democratic Left	3	0
Fine Gael	22	0
Green	1	0
Labour	7	0


After Feature Selection
Original
0.0
0.2
0.4
0.6
0.8
1.0
Haughey (Tac
Reynolds (Min)
Nolan
Daly (Min)
OHanlon (Min)
Leyden (JMin)
ODonoghue
Barry
Cowan
FitzGerald
Hillery
Ferris
OMalley (Leader)
Quinn
Boylan
Howlin
Flynn (Min)
BrutonJ (Leader)
Ahearn
Creed
Yates
HigginsJ
Deenihan
Kenny
Lenihan
Durkan
Currie
Burke (Min)
Browne
Finucane
HigginsMD
Owen
OShea
Rabbitte
Davern

A Better Predictor

$$\hat{y} = \begin{cases} \text{Govt} & \text{if Party} = \text{FF or PD} \\ \text{Opp} & \text{otherwise} \end{cases}$$

Misclassification rate: 0%

(Classification is not an interesting problem)

Regularization

$$\tilde{\ell}(\theta) = \ell(\theta) + \frac{1}{2}\sum_{k=1}^{K} \log \theta_k$$

- Ensures parameter estimates are in the interior of the parameter space
- Introduces O(1/n) bias
- Reduces estimator variance

(cf. Firth 1993)

100+ Years of Text Analysis

- Markov's analysis of a Pushkin poem (1913)
- Zipf's Law for word usage rates (1949)
- *The Federalist* authorship inference (Mosteller and Wallace 1964)
- "How many words did Shakespeare know?" (Efron and Thisted, 1976)
- The Penn Treebank of annotated linguistic data (Marcus et al. 1993)
- Bayesian spam filterning (Sahami et al. 1998)
- Opinion Mining and Sentiment Analysis (e.g. Pang and Lee 2008)
- Vector space word embeddings (Bengio et al. 2003; Mikolov et al. 2013)
- (a billion other references)

Why Not Use sLDA?

- Lots of tuning parameters (#topics, prior hyperparameters, regularization parameter), overly complicated model
- Unreasonable results for the extremes
- Interpretation is different: Odds of membership (sLDA) vs. Degree of membership (affinity model)