SAMUEL SHERMAN

You can find me at:

github.com/scsherm

linkedin.com/in/samuelcsherman

Congressional Bill Modeling

Samuel Sherman scsherm@gmail.com github.com/scsherm linkedin.com/in/samuelcsherman

Data Collection

- **Congress.gov**
  - Bill text

- **Sunlight Foundation API**
  - Bill data (JSON)
  - Votes data (JSON)

- **MongoDB**

- **Tools**
  - Pandas, pymongo, sklearn, nltk, numpy, scipy, AWS ec2

Samuel Sherman scsherm@gmail.com github.com/scsherm linkedin.com/in/samuelcsherman

Motivations

- **Percent of yes votes for a given party**
  - **Polarization in government**
- **Whether a bill will reach a vote**
  - **Important features in bills**
- **Latent topics in bill text**
  - **Prevalence over time**
  - **Distinguishable characteristics between presidencies**

Natural Language Processing

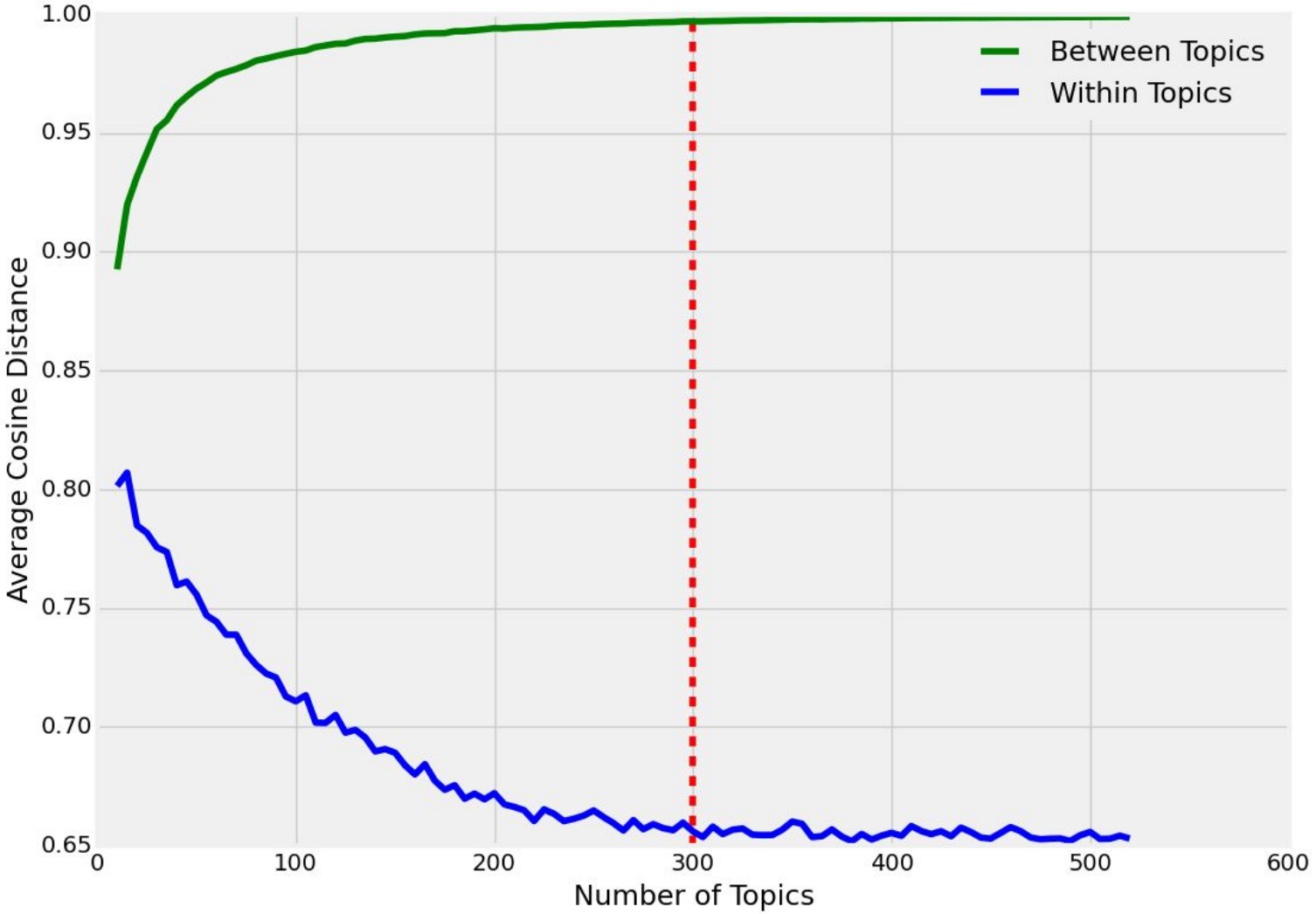
- **Stopwords**
  - Regular words: “and”, “the”, “be”, “it”, “there”
  - Congressional: “amendment”, “bill”, “quorum”, “act”
- **Maximum document frequency**
- **TFIDF**
  - TF: Frequency of words across a single document
  - IDF: Frequency of words across all documents

Non-Negative Matrix Factorization

- **Similarity within topics → high**

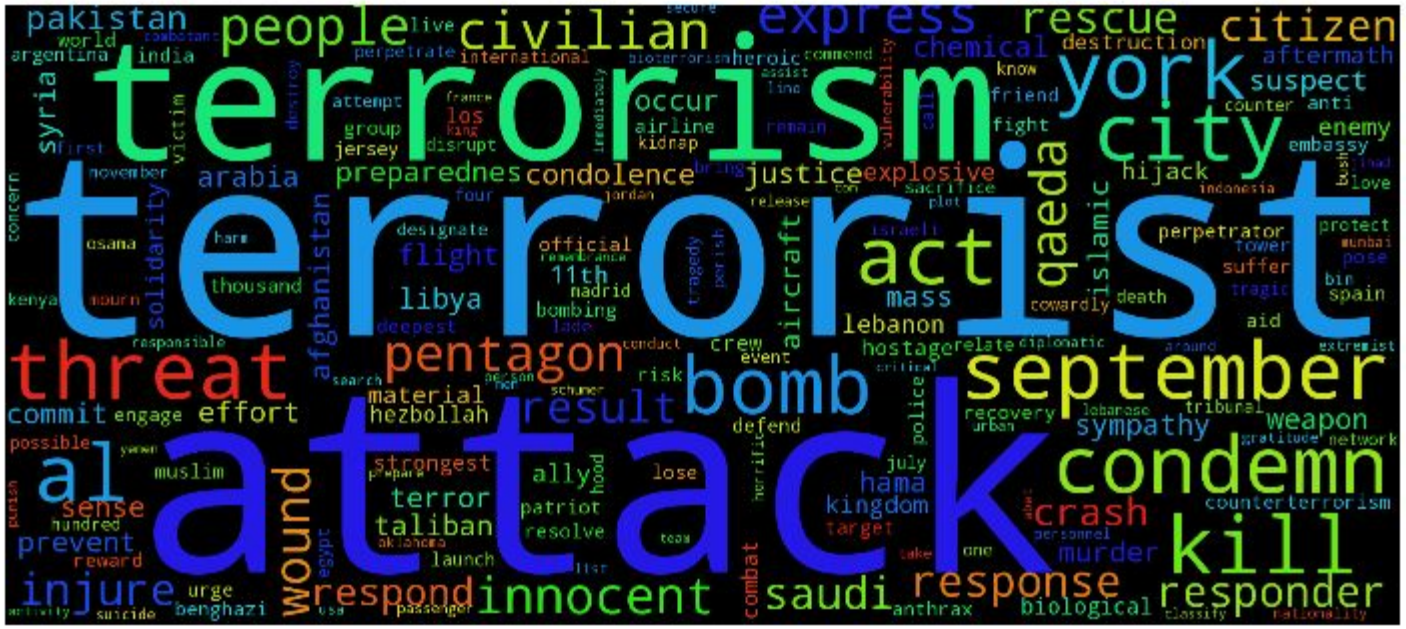
- **Similarity between topics → low**

Choosing K Topics



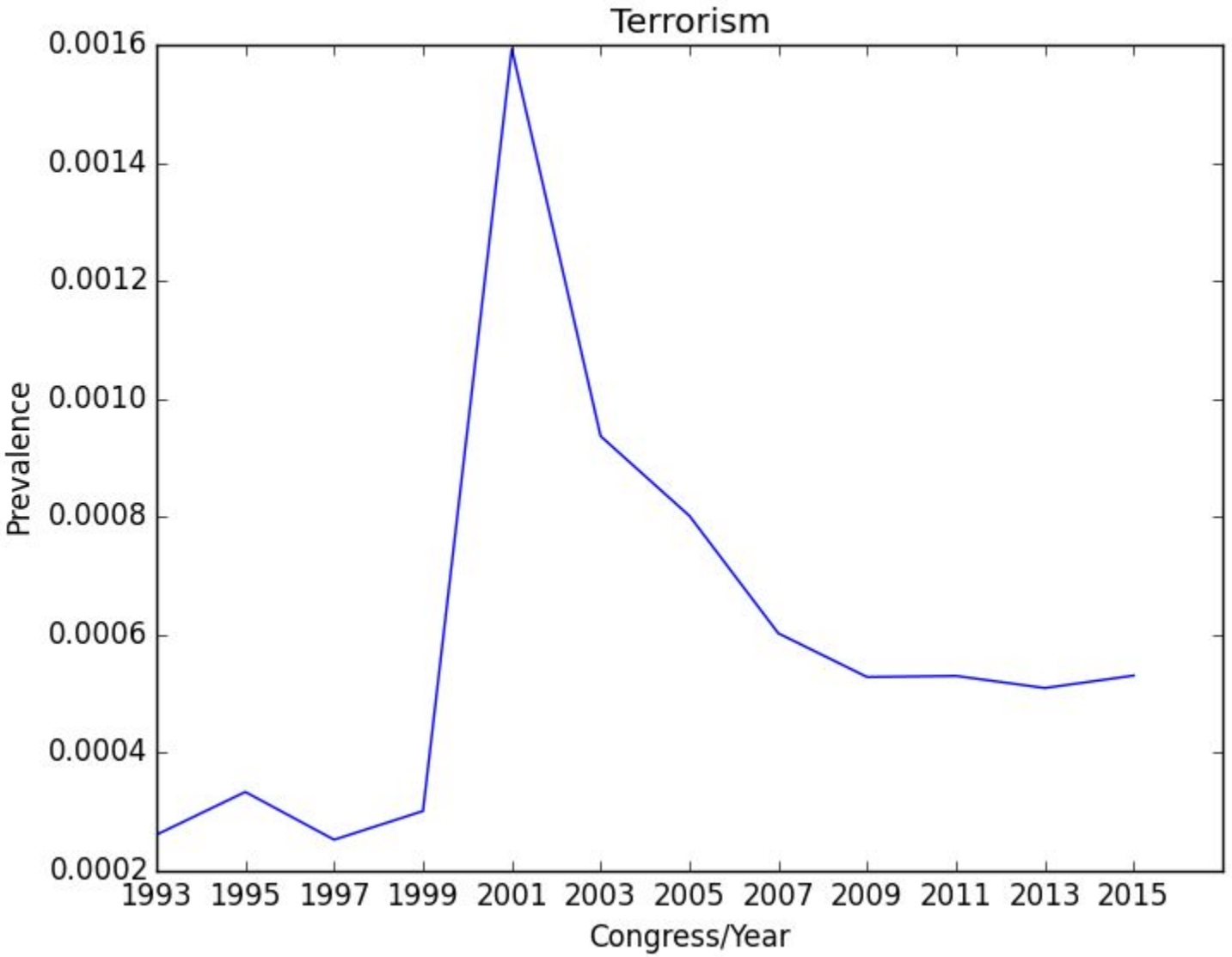
Measuring Topic Distance

Topic 110



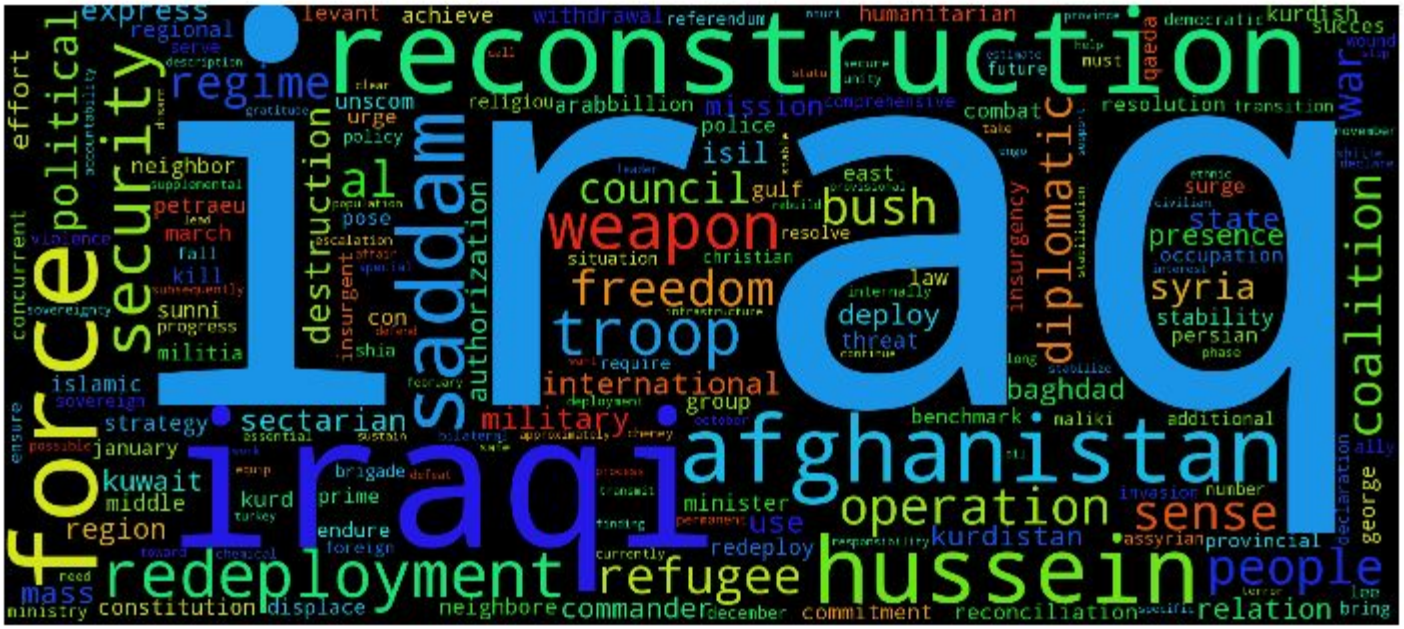
Terrorism

Samuel Sherman scsherm@gmail.com github.com/scsherm linkedin.com/in/samuelcsherman



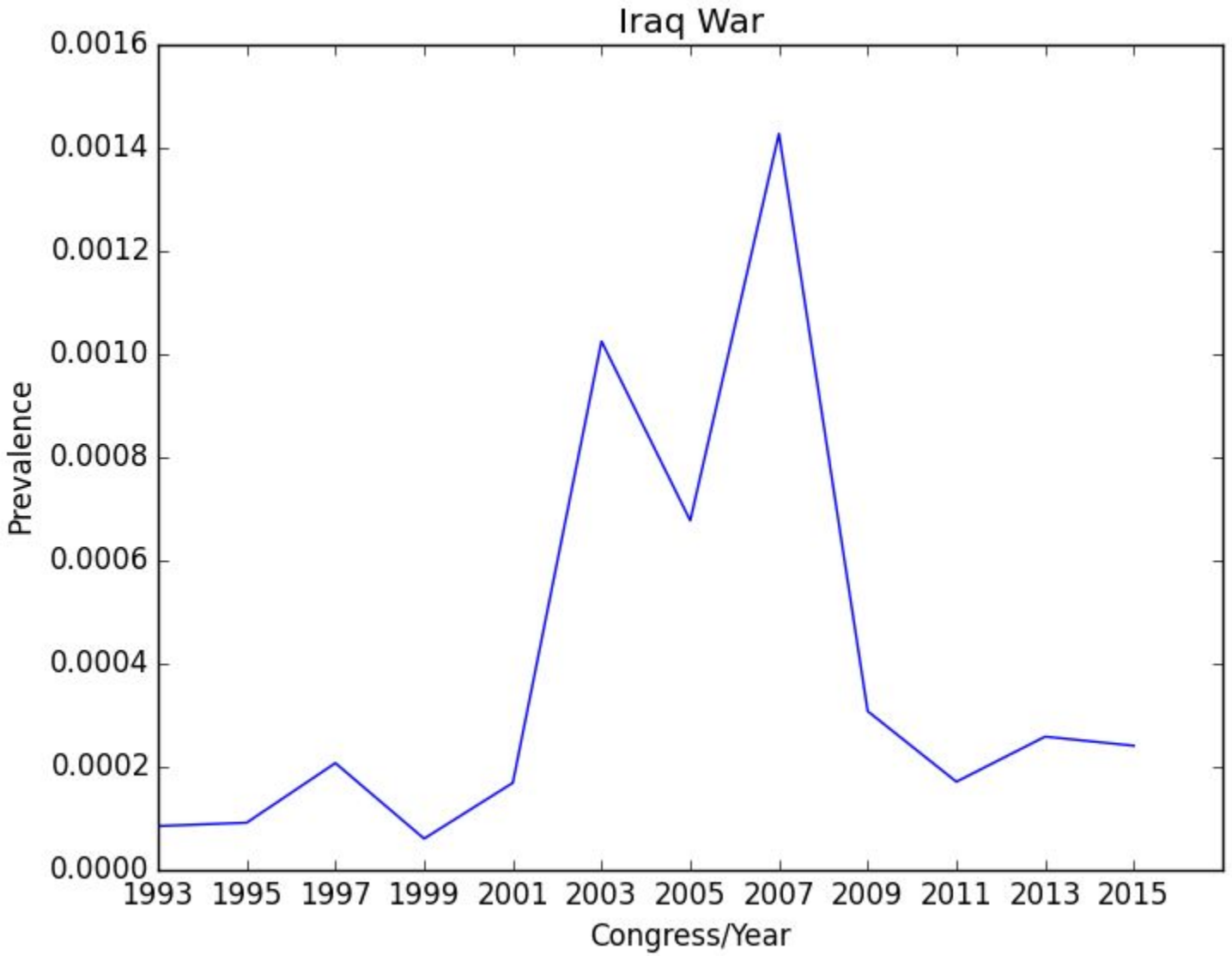
Terrorism in Bills Over Time

Topic 76



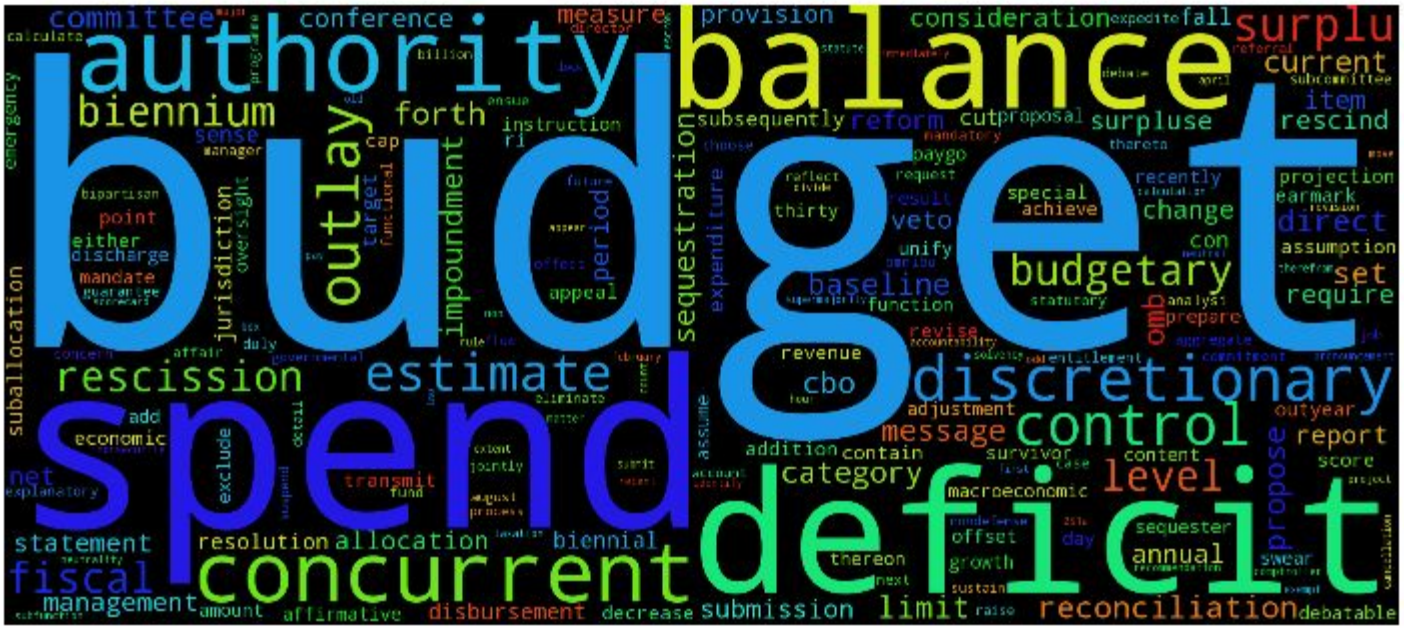
Iraq War

Samuel Sherman scsherm@gmail.com github.com/scsherm linkedin.com/in/samuelcsherman

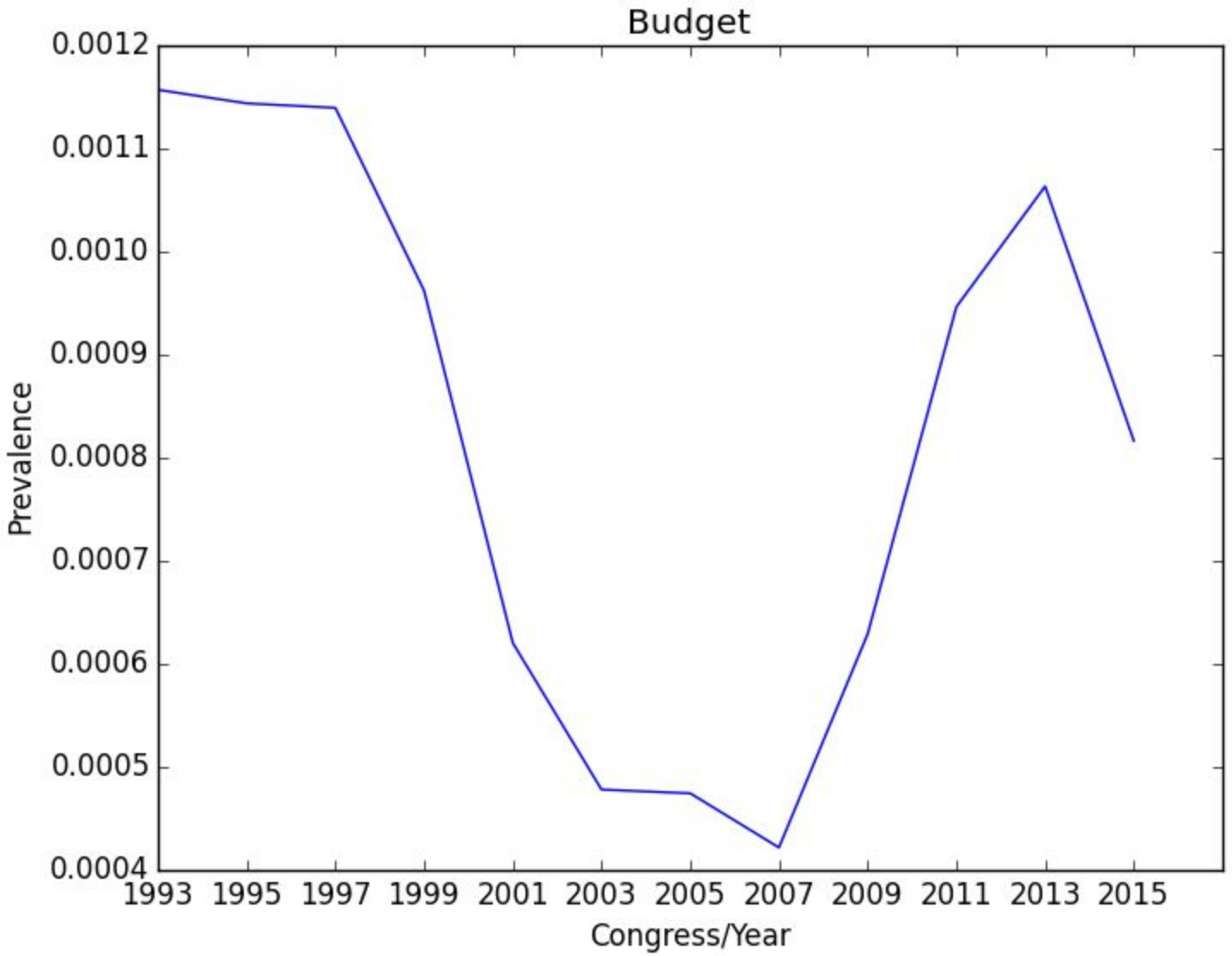


Iraq War in Bills Over Time

Topic 188

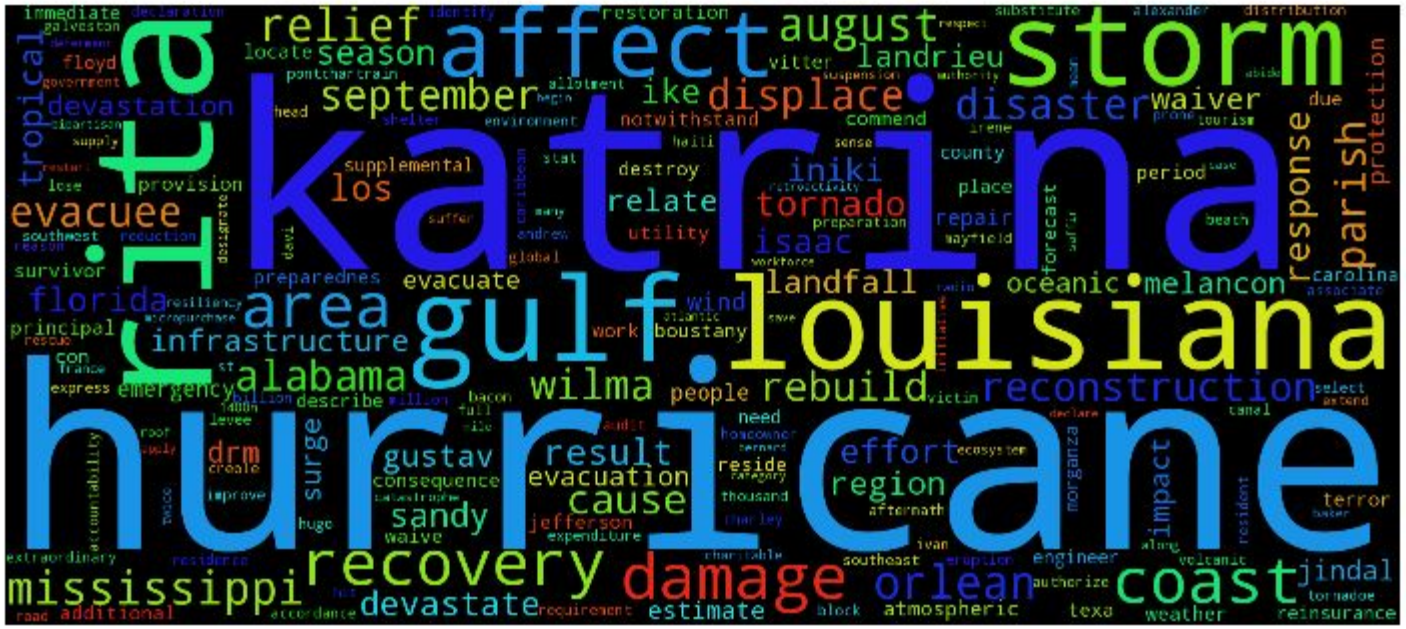


Budget



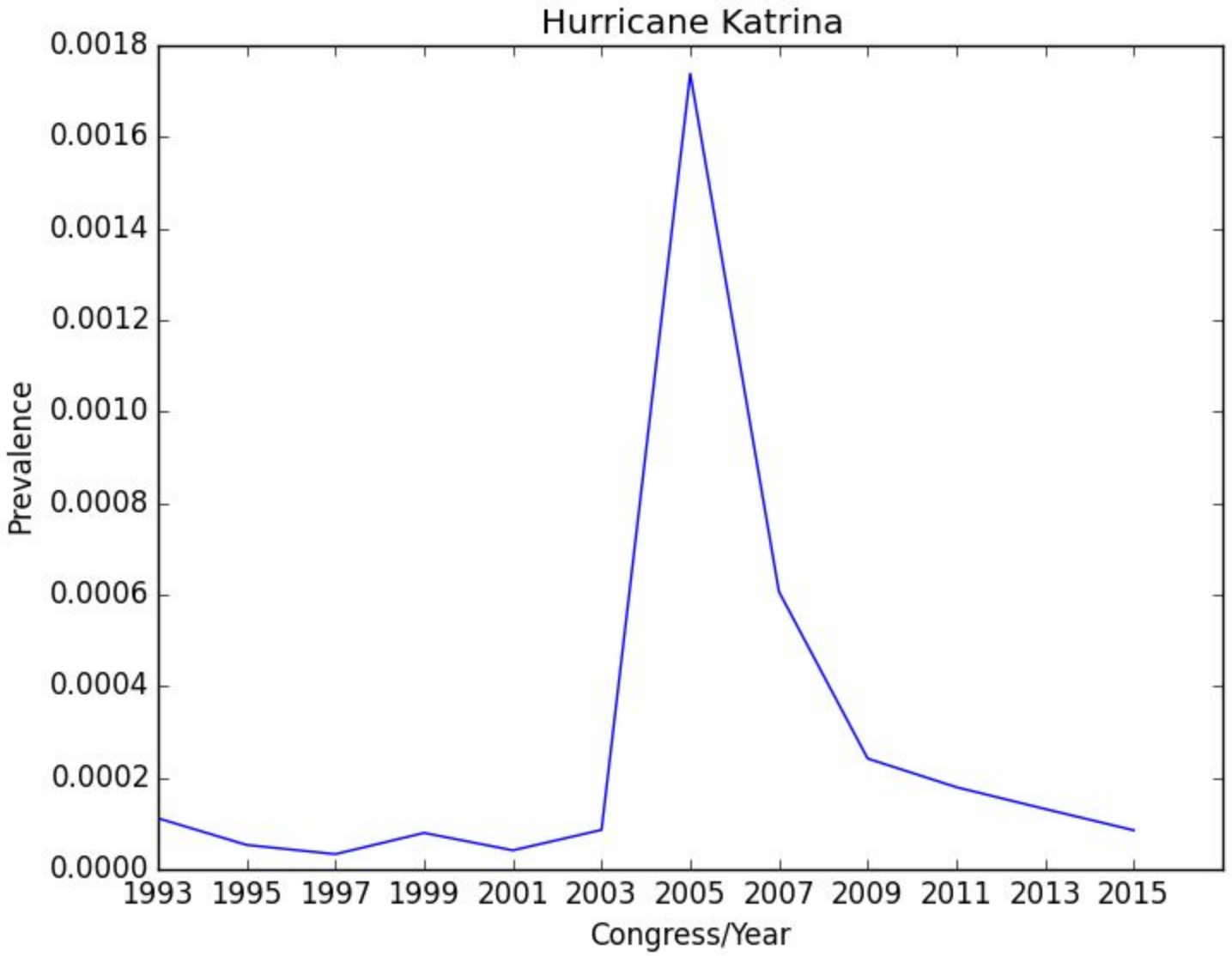
Budget in Bills Over Time

Topic 205



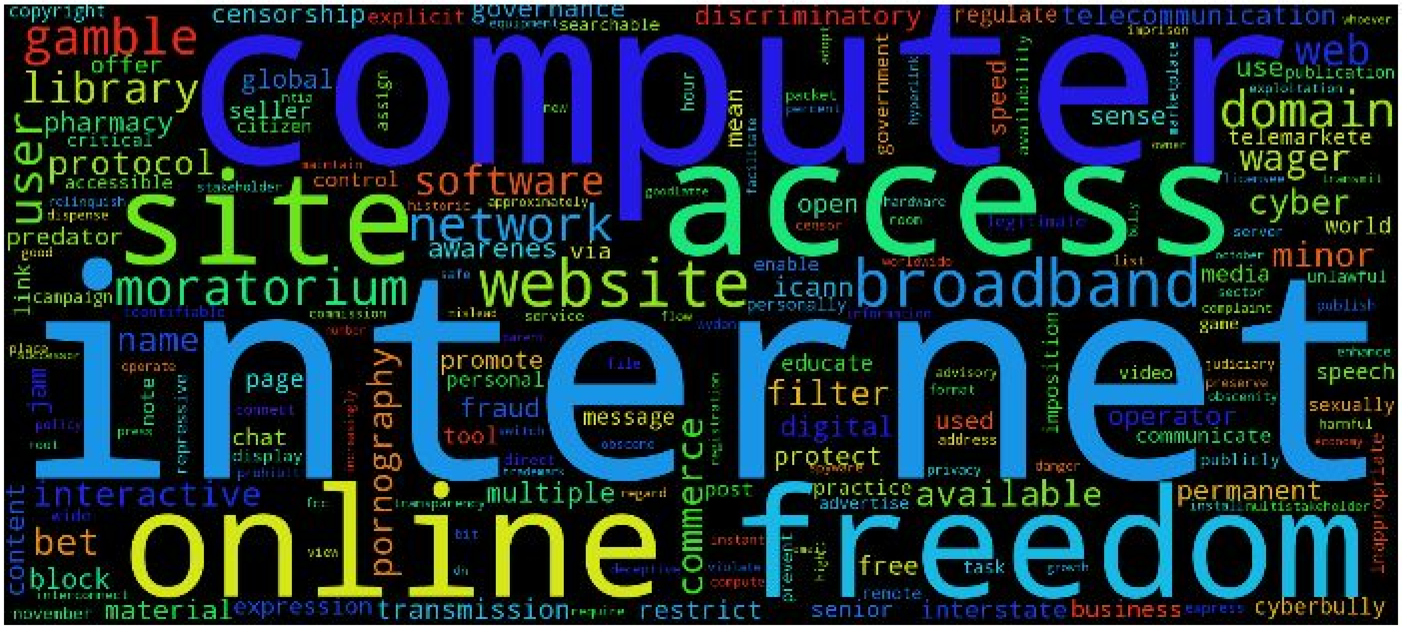
Hurricane Katrina

Samuel Sherman scsherm@gmail.com github.com/scsherm linkedin.com/in/samuelcsherman



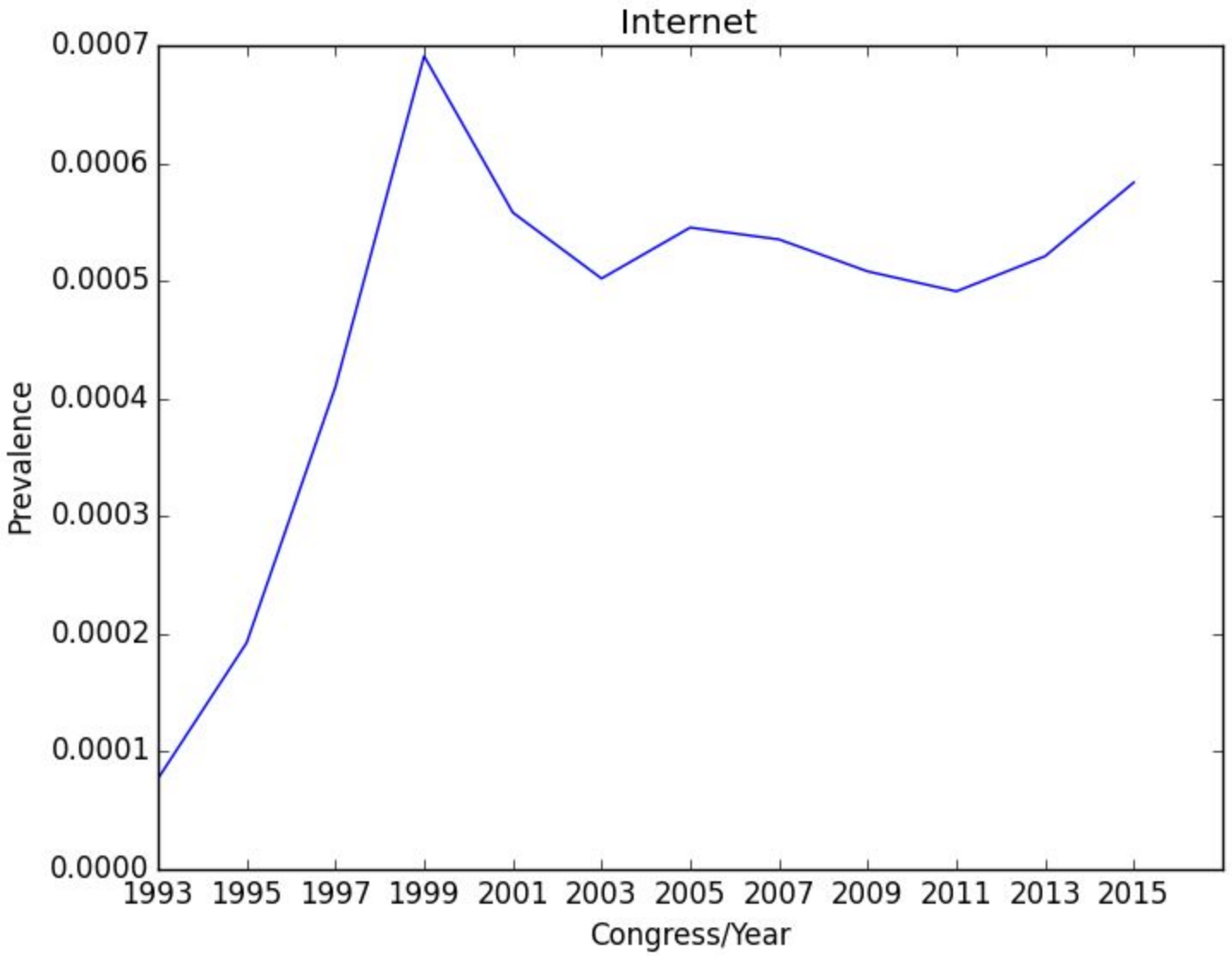
Hurricane Katrina in Bills Over Time

Topic 220



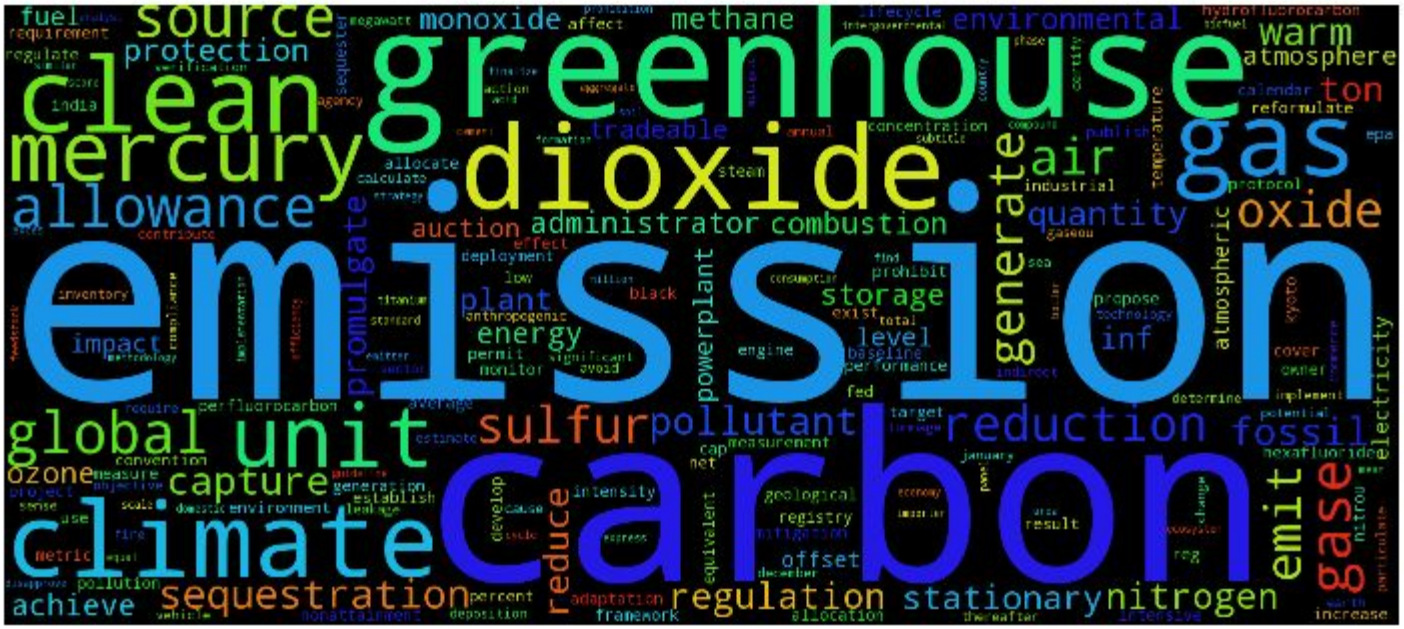
Internet

Samuel Sherman scsherm@gmail.com github.com/scsherm linkedin.com/in/samuelcsherman



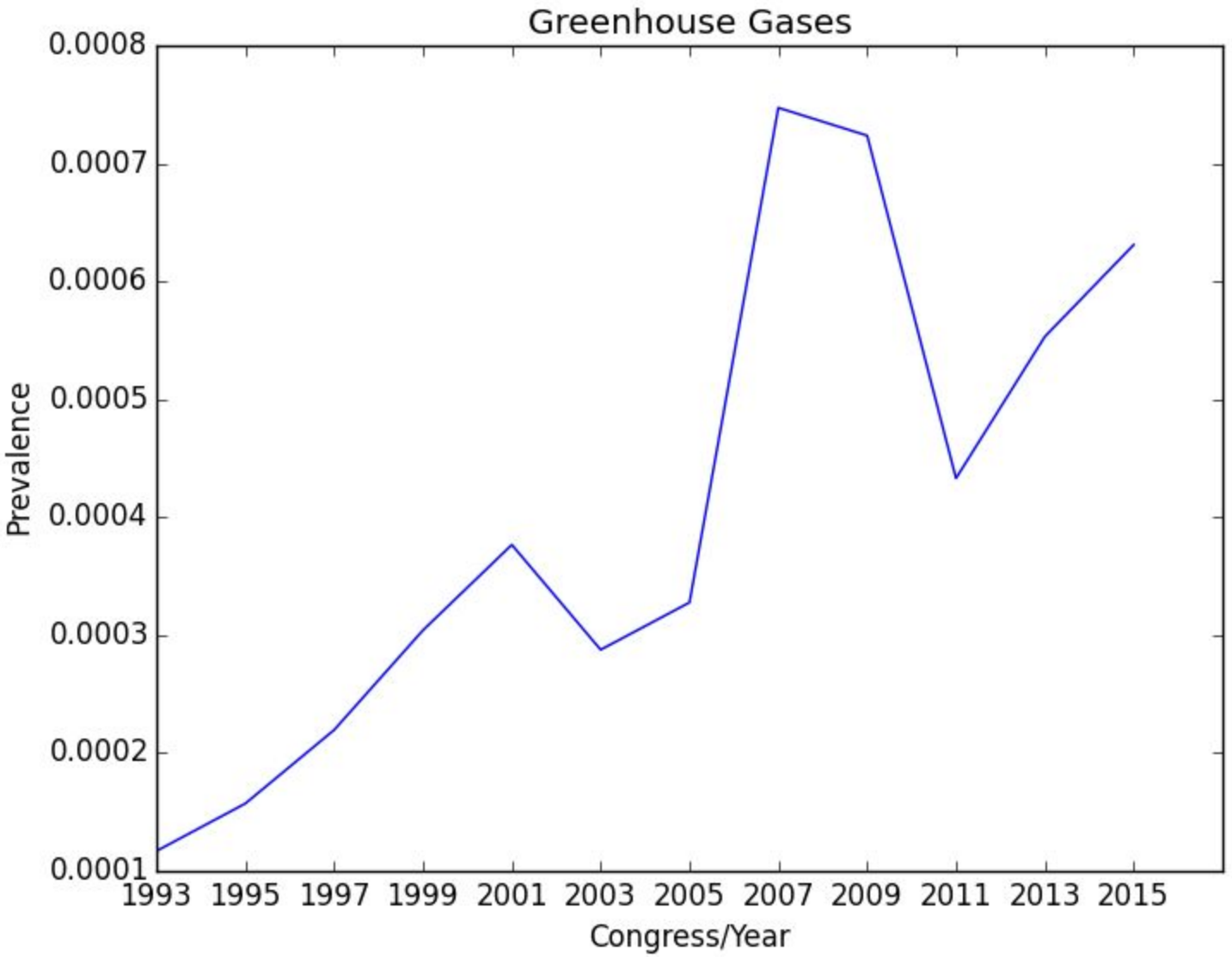
Internet in Bills Over Time

Topic 277



Greenhouse Gases

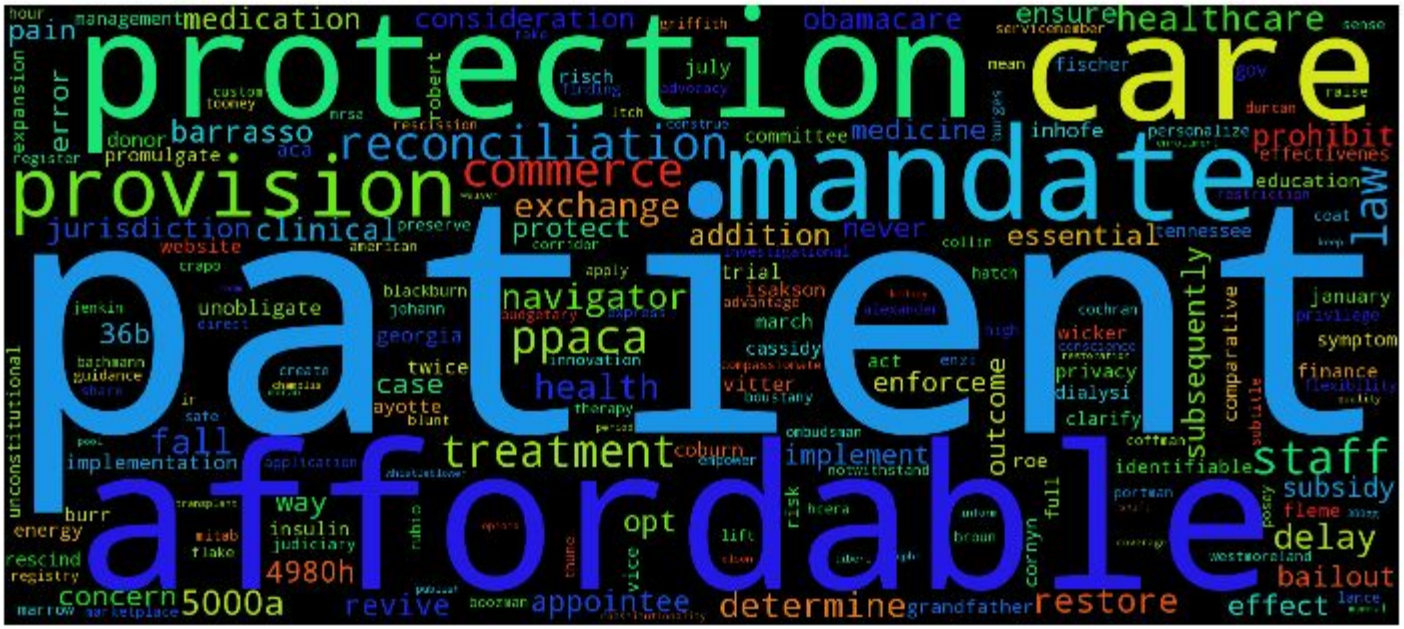
Samuel Sherman scsherm@gmail.com github.com/scsherm linkedin.com/in/samuelcsherman



Greenhouse Gases in Bills Over Time

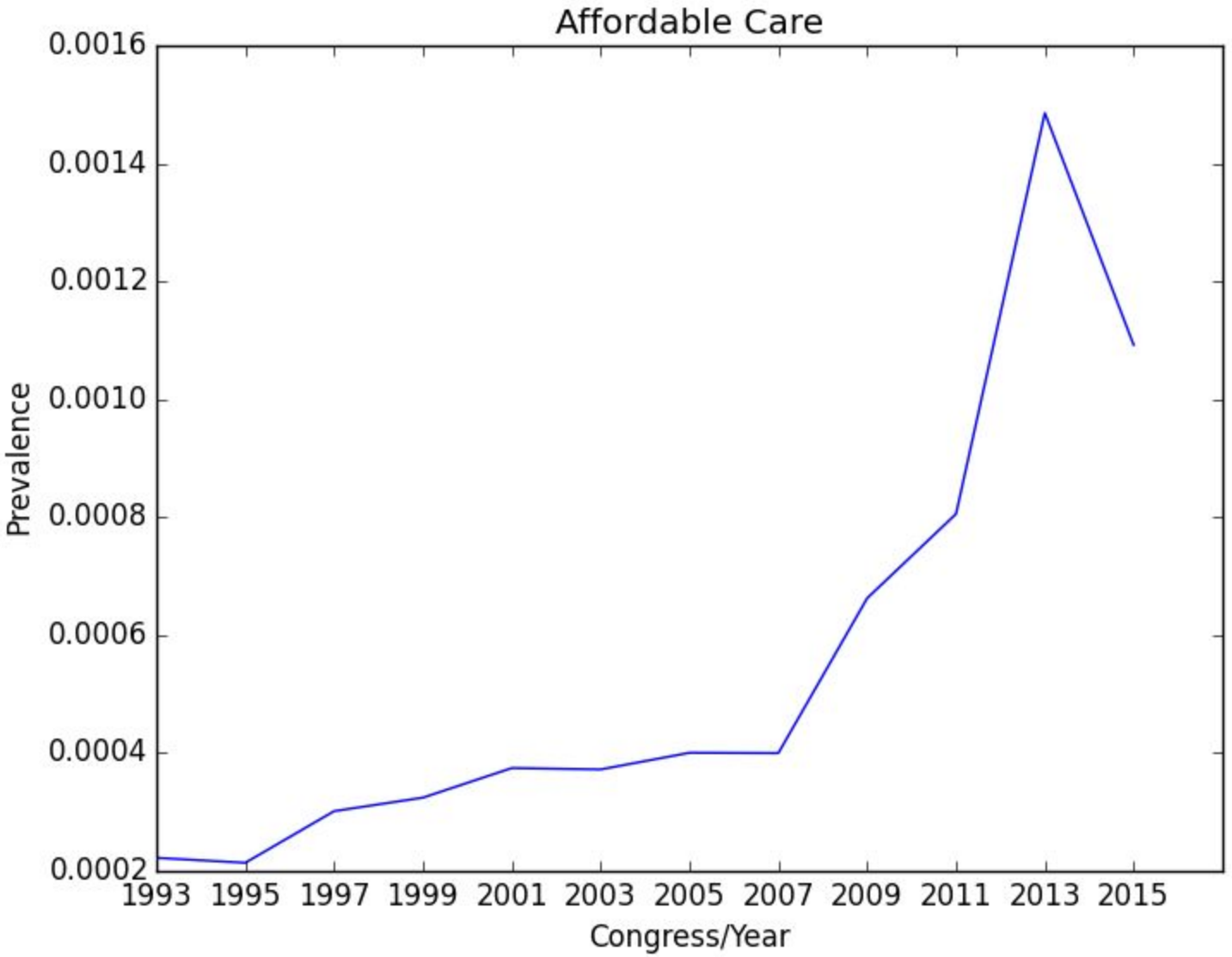
Samuel Sherman scsherm@gmail.com github.com/scsherm linkedin.com/in/samuelcsherman

Topic 286



Affordable Care Act

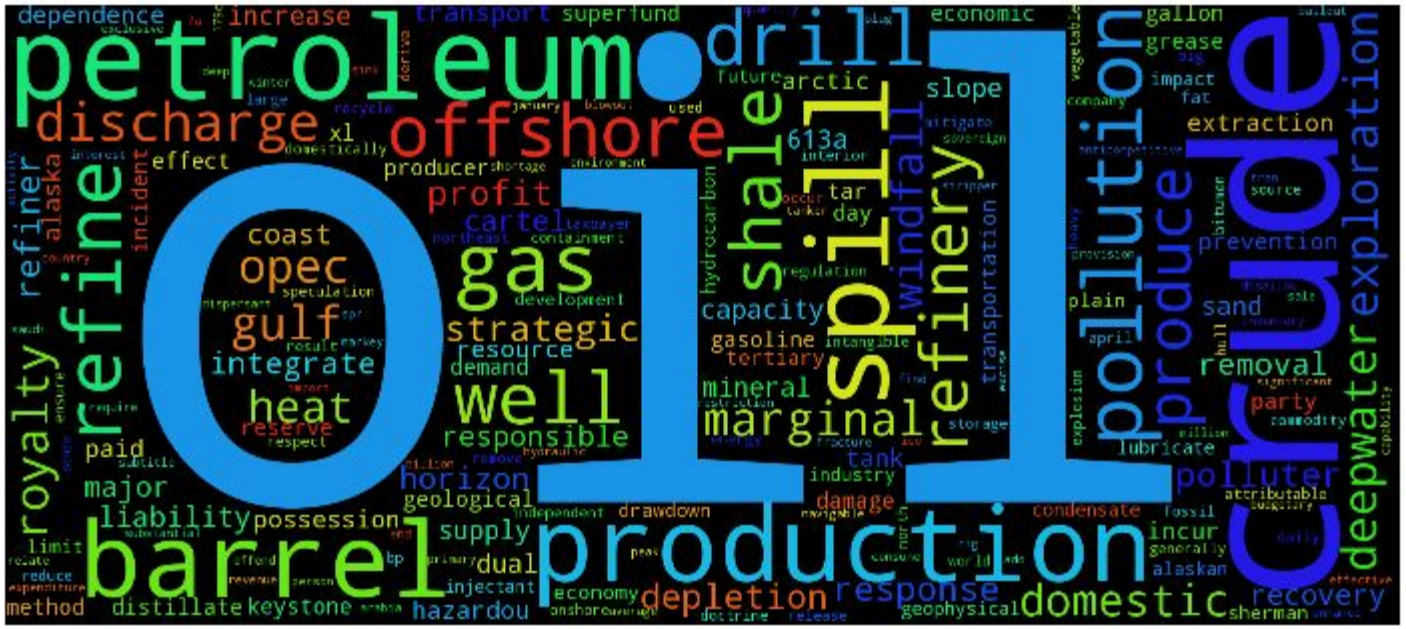
Samuel Sherman scsherm@gmail.com github.com/scsherm linkedin.com/in/samuelcsherman



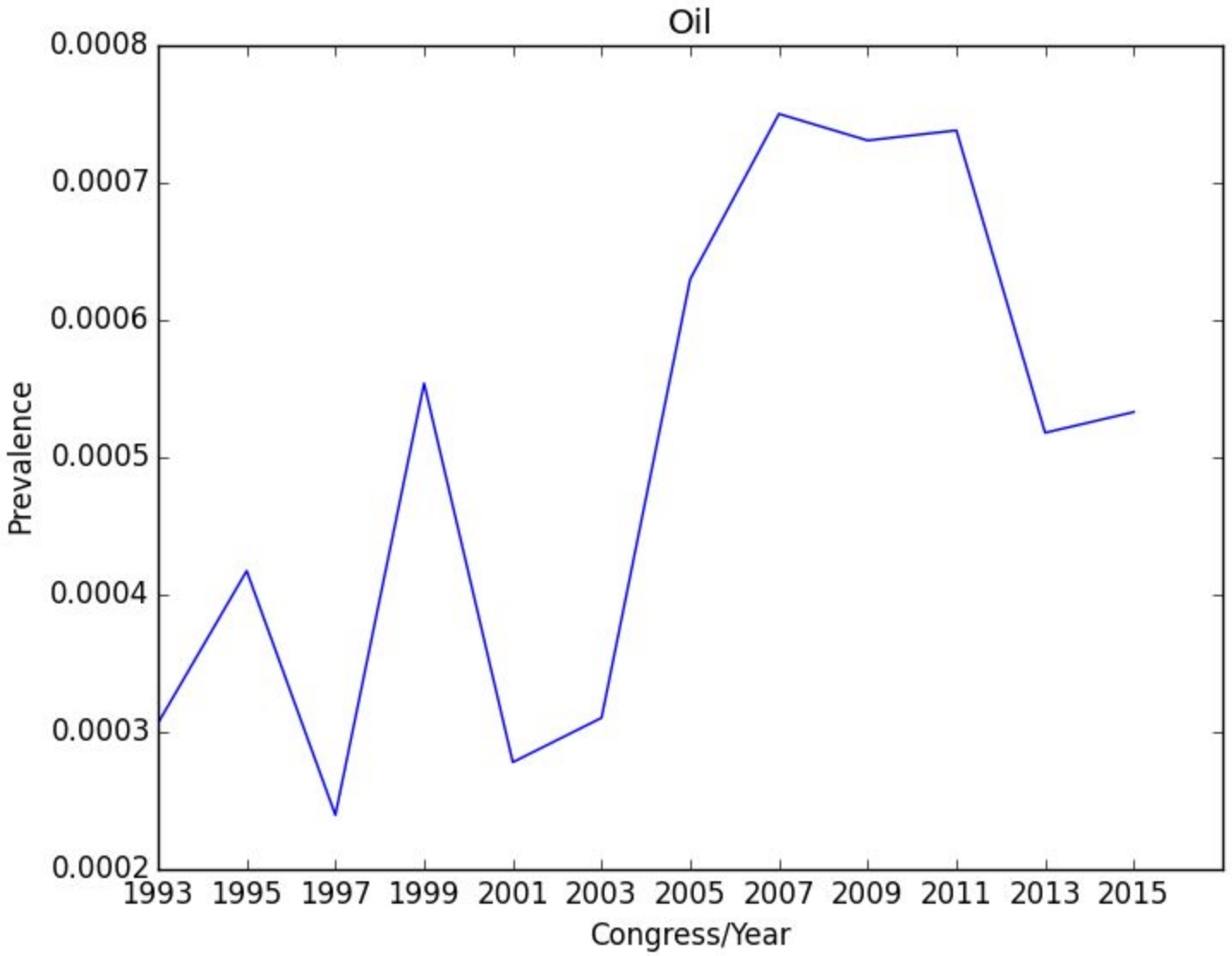
Affordable Care in Bills Over Time

Samuel Sherman scsherm@gmail.com github.com/scsherm linkedin.com/in/samuelcsherman

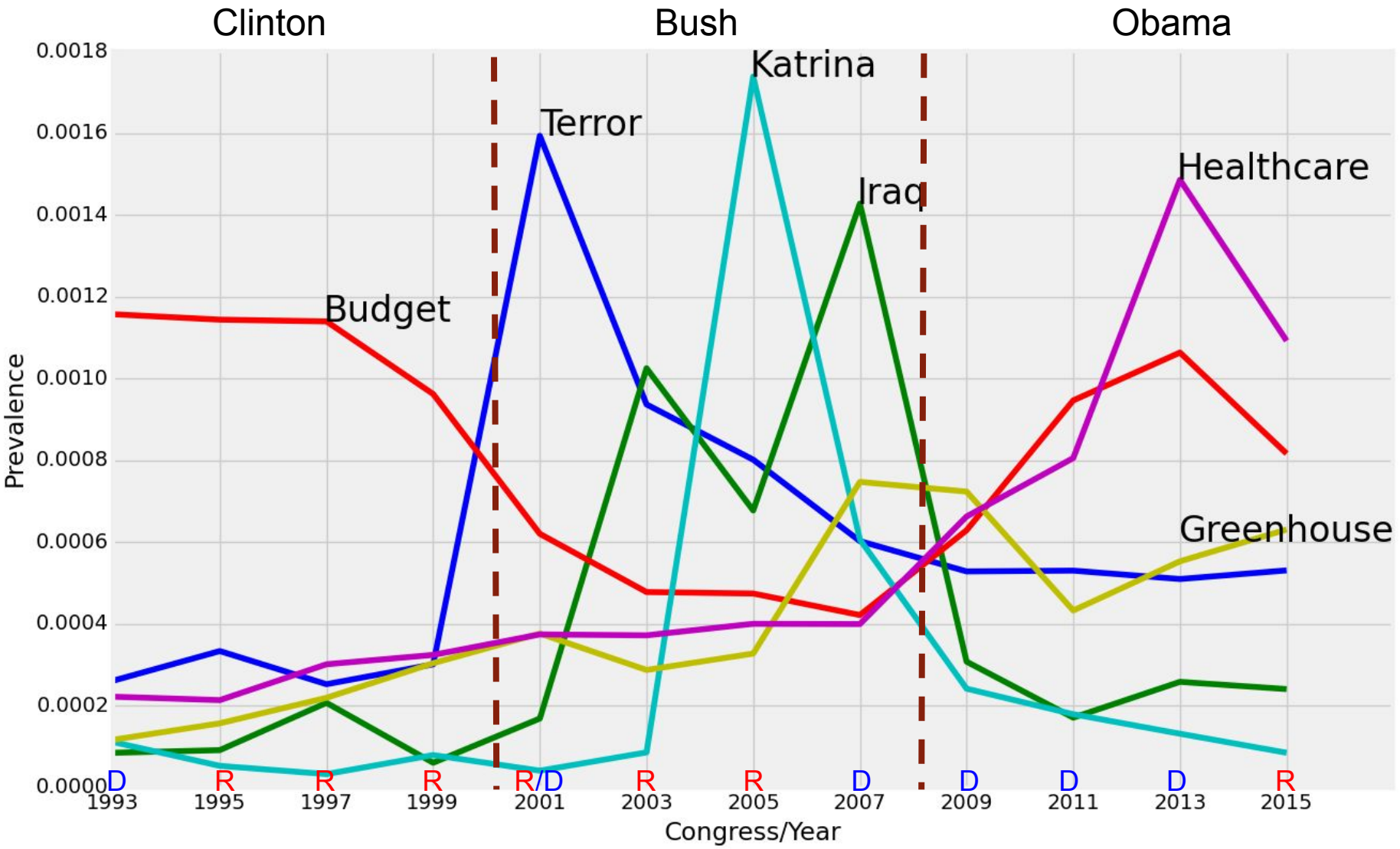
Topic 235



Oil



Oil in Bills Over Time



Samuel Sherman scsherm@gmail.com github.com/scsherm linkedin.com/in/samuelcsherman

Thanks!

ANY QUESTIONS?

You can find me at:

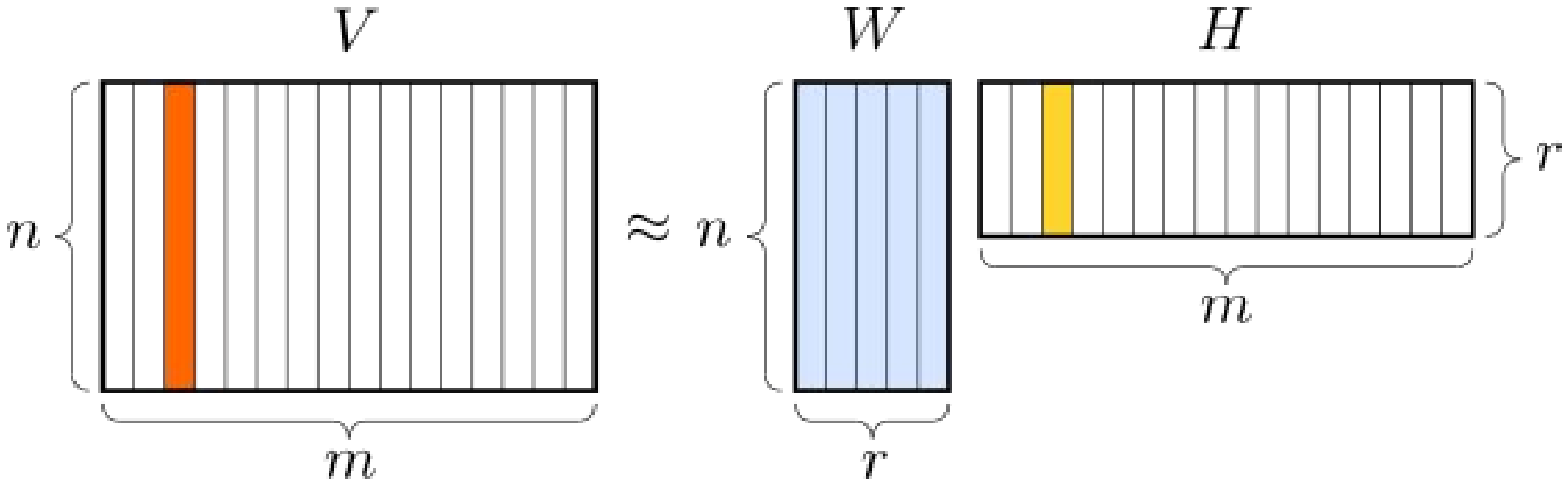
linkedin.com/in/samuelcsherman

github.com/scsherm

scsherm@gmail.com

Appendix

Non-Negative Matrix Factorization

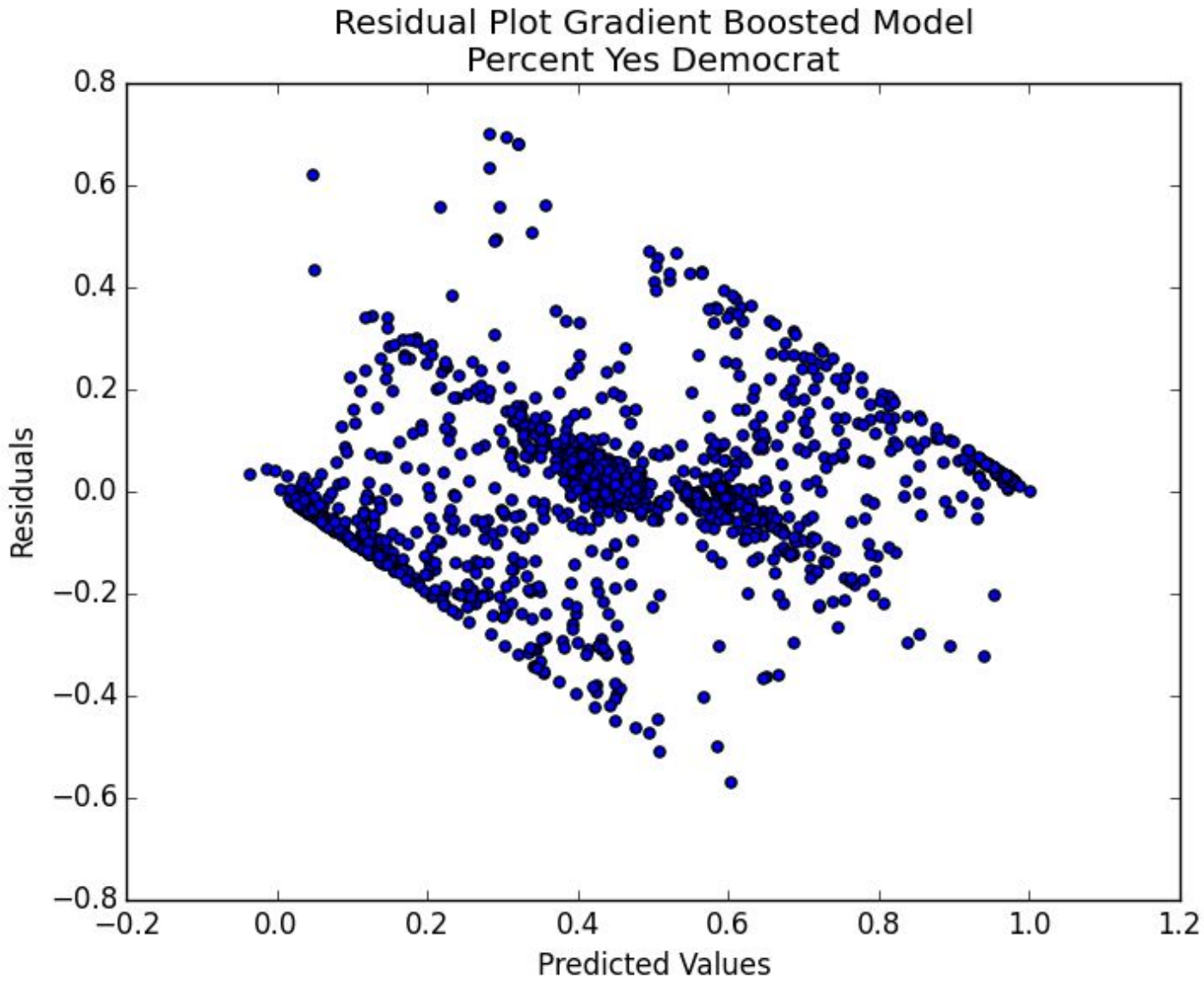


- Recommendation Systems
  - Null space
- Trend modeling/finance/stocks
  - Social Media
- Even Image Processing
  - Pattern/facial recognition

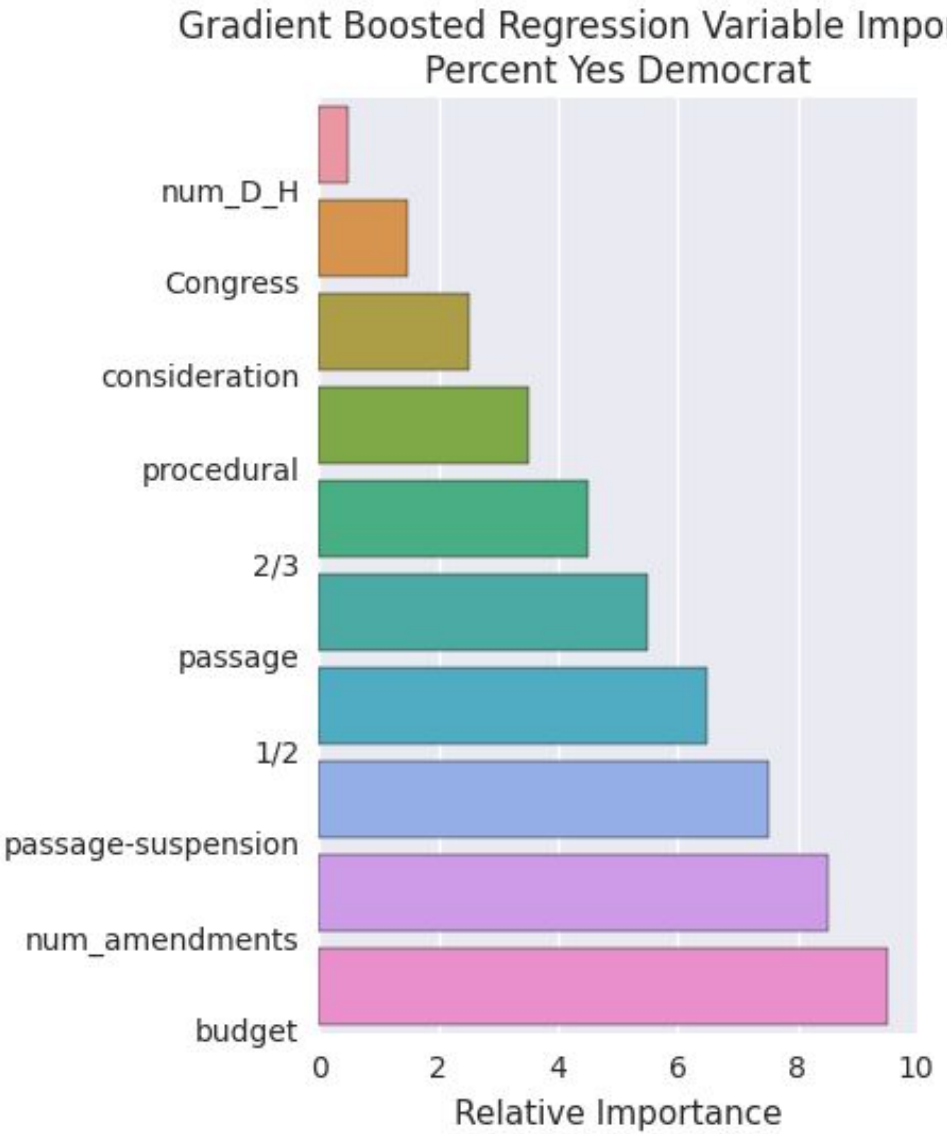
Regression

Percent Yes Democrat

	Mean Squared Error	Root Mean Squared Error	R^2 score
Random Forest	**0.0205**	**0.143**	**0.740**
Bagging	**0.0207**	**0.144**	**0.737**
Linear	**0.0417**	**0.204**	**0.417**
Gradient Boosted	**0.0203**	**0.143**	**0.743**


Residual Plot Gradient Boosted Model
Percent Yes Democrat
Residuals
Predicted Values
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2

Residuals



Feature Importances

Classifier

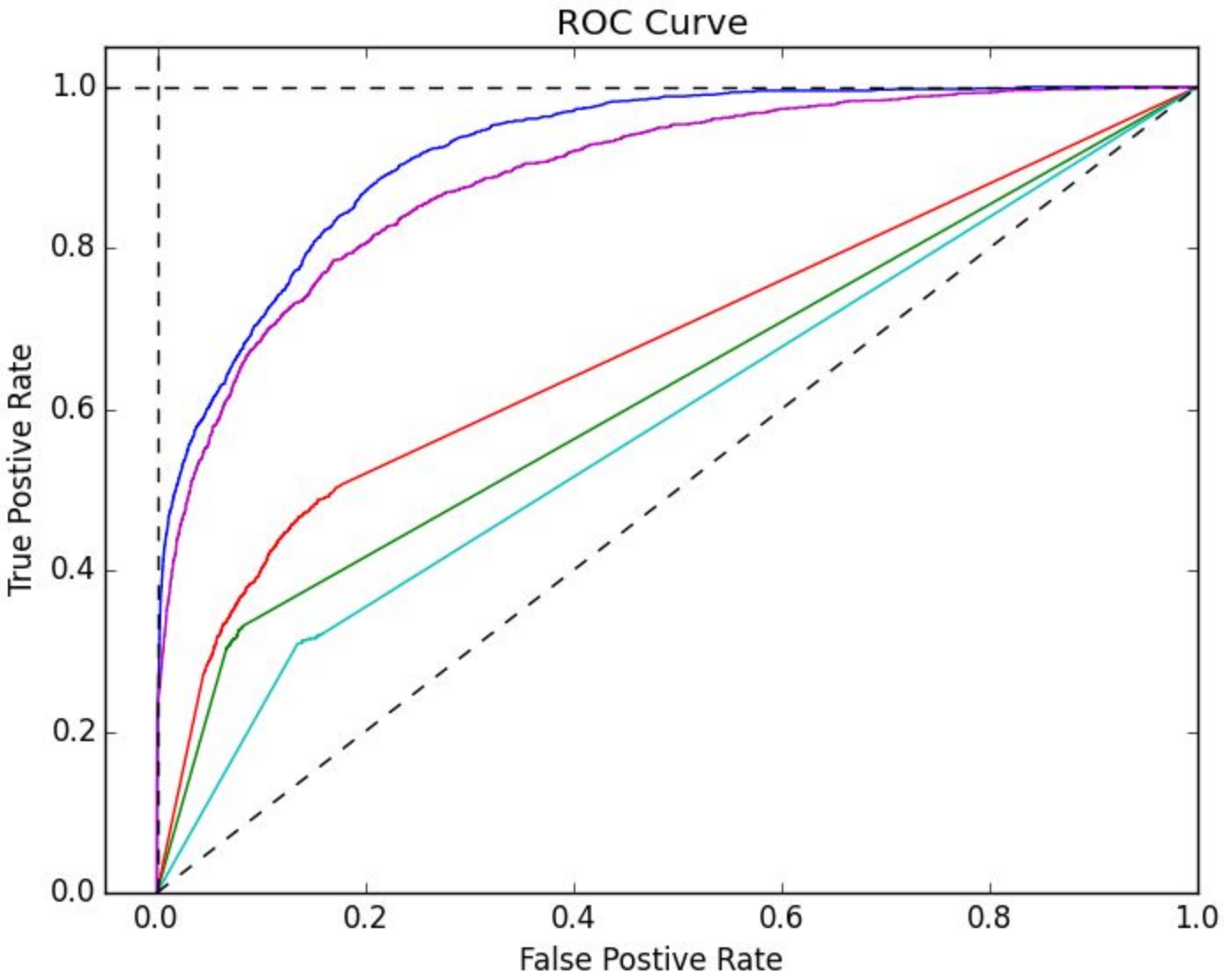
130,000 Bills

Since 1993

6% voted on

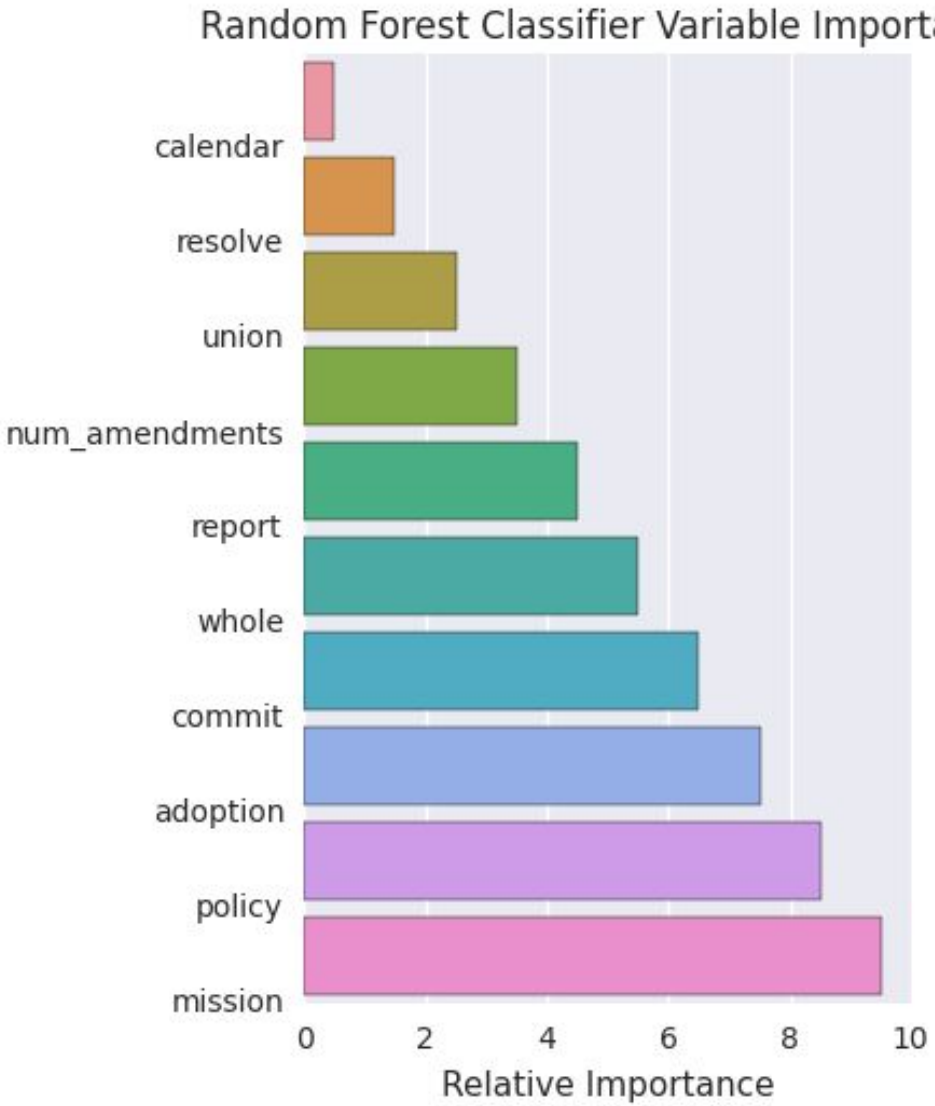
7,600

Samuel Sherman scsherm@gmail.com github.com/scsherm linkedin.com/in/samuelcsherman



RFC AUC = 0.921
GNB AUC = 0.626
MNB AUC = 0.681
BNB AUC = 0.584
logr AUC = 0.89

- **Oversampling**
  - **SMOTE**
- **Undersampling**
- **Recall/Precision**



Feature Importances