Data Analytics for Social Science
Wordscores & wordfish

Johan A. Elkink

School of Politics & International Relations
University College Dublin

18 April 2017

Text analysis in political science

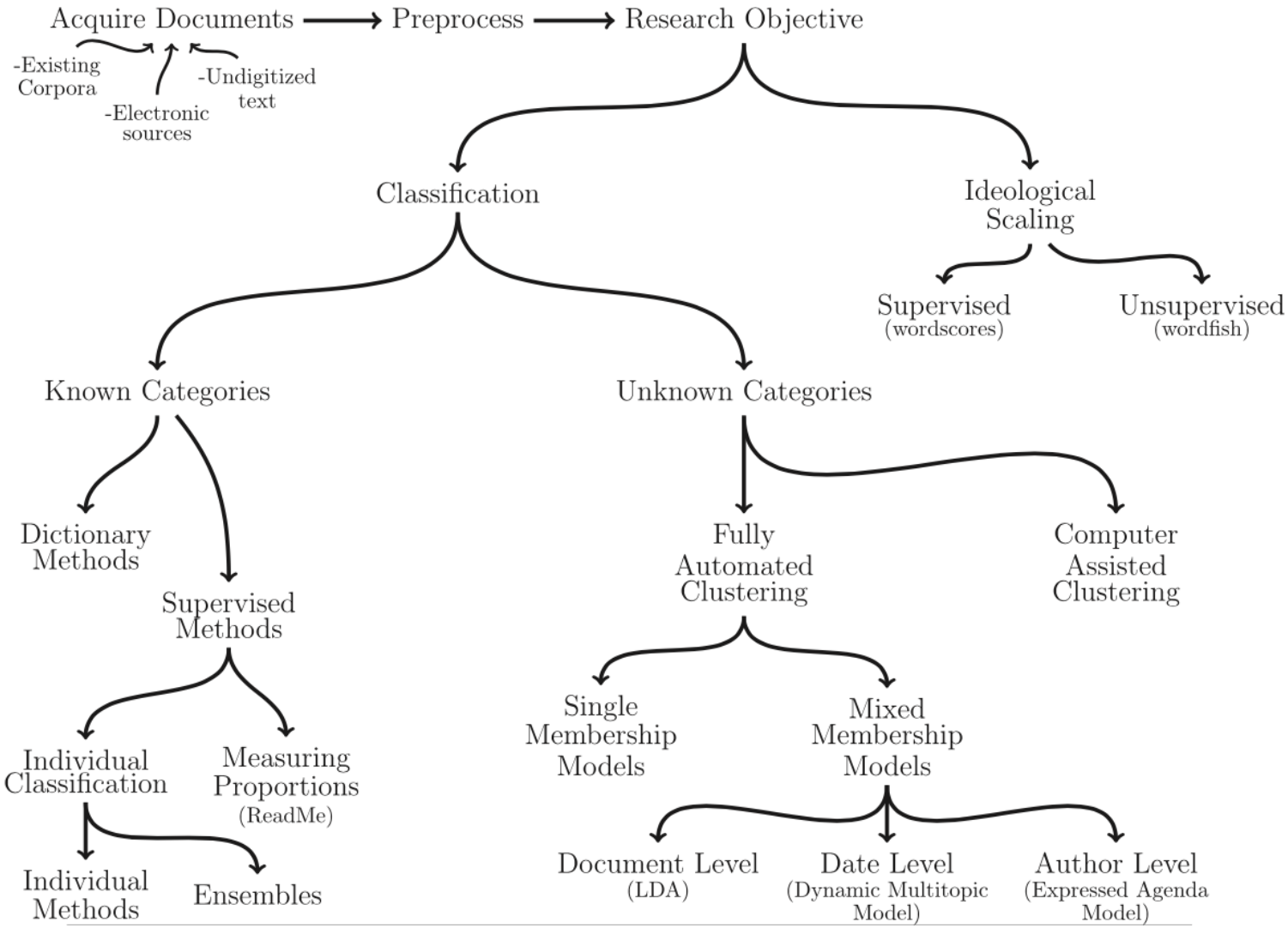
Within machine learning and data science, there are a number of applications for text analysis, such as:

- Identifying topics, e.g. what does Trump tweet about?
- Identifying sentiment, e.g. are tweets about Trump positive or negative?
- Identifying authorship, e.g. who is that ghostwriter?[1]
- Categorising documents, e.g. which emails are SPAM?

In political science, the main usage has been identifying the **ideological position** of political actors.

[1] An early example of this in political science concerns author identification of unsigned *Federalist Papers* (Mosteller and Wallace, 1964).

Classification of text methods



(Grimmer and Stewart, 2013, 2)

Ideological placement

Many attempts have been made to identify the ideological position of political actors:

- Manual coding of manifestos (Budge, Robertson and Hearl, 1987; Budge et al., 2001; Benoit and Laver, 2007)
- Elite voting behaviour (e.g. on laws in parliament or on cases in courts) (Poole and Rosenthal, 1985)
- Expert surveys (Castles and Mair, 1984; Laver and Hunt, 1992; Benoit and Laver, 2007; Bakker et al., 2015)
- Voter perceptions or voting behaviour (Cunningham and Elkink, forthcoming)
- Self-placement

The newest trend is to use statistical text analysis of manifestos or speeches to get at the ideological placement of actors.

Wordscores and wordfish

While the focus is now on techniques developed in machine learning and data science, the first statistical text analysis applications in political science were based on specifically designed procedures.

Wordscores: developed by Laver, Benoit and Garry (2003), estimates the ideological position of a number of documents based on the word similarity with two reference texts defining the extreme ends of the scale.

Wordfish: developed by Slapin and Proksch (2008), estimates an underlying latent variable model reflecting ideology, based on word frequencies.

Outline

Wordscores

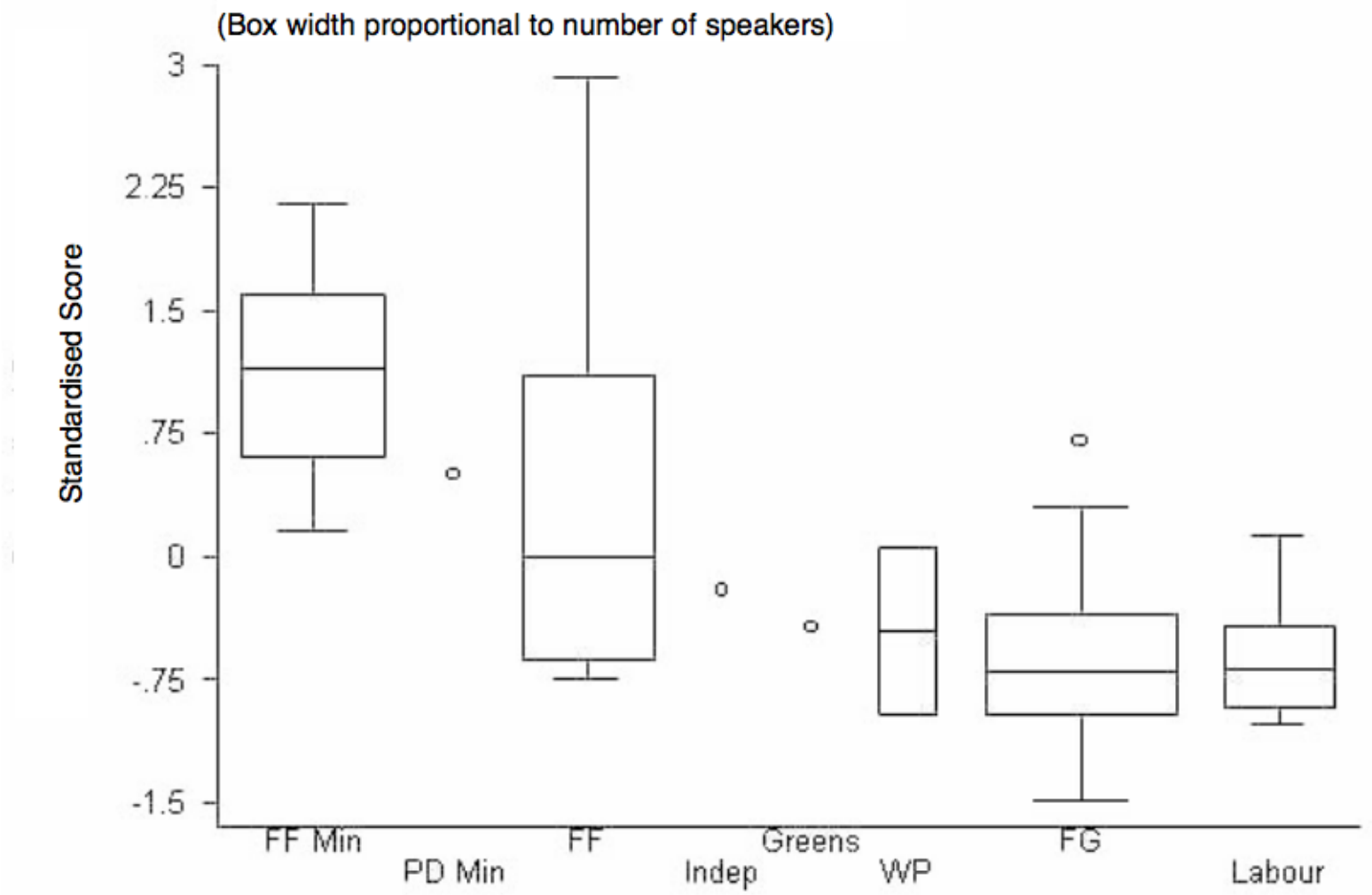
1. Select a set of **reference texts** (or training data) with known ideological scores, at extreme ends of the dimension.
2. Score words based on frequency in each reference text.
3. Take all remaining texts, **virgin texts** (or test data), and calculate document scores by taking the weighted average word score, weighted by the frequency.

Can be too heavily influenced by terms that are irrelevant for ideology and is very sensitive to the selection of the reference texts.

(Grimmer and Stewart, 2013, 26)

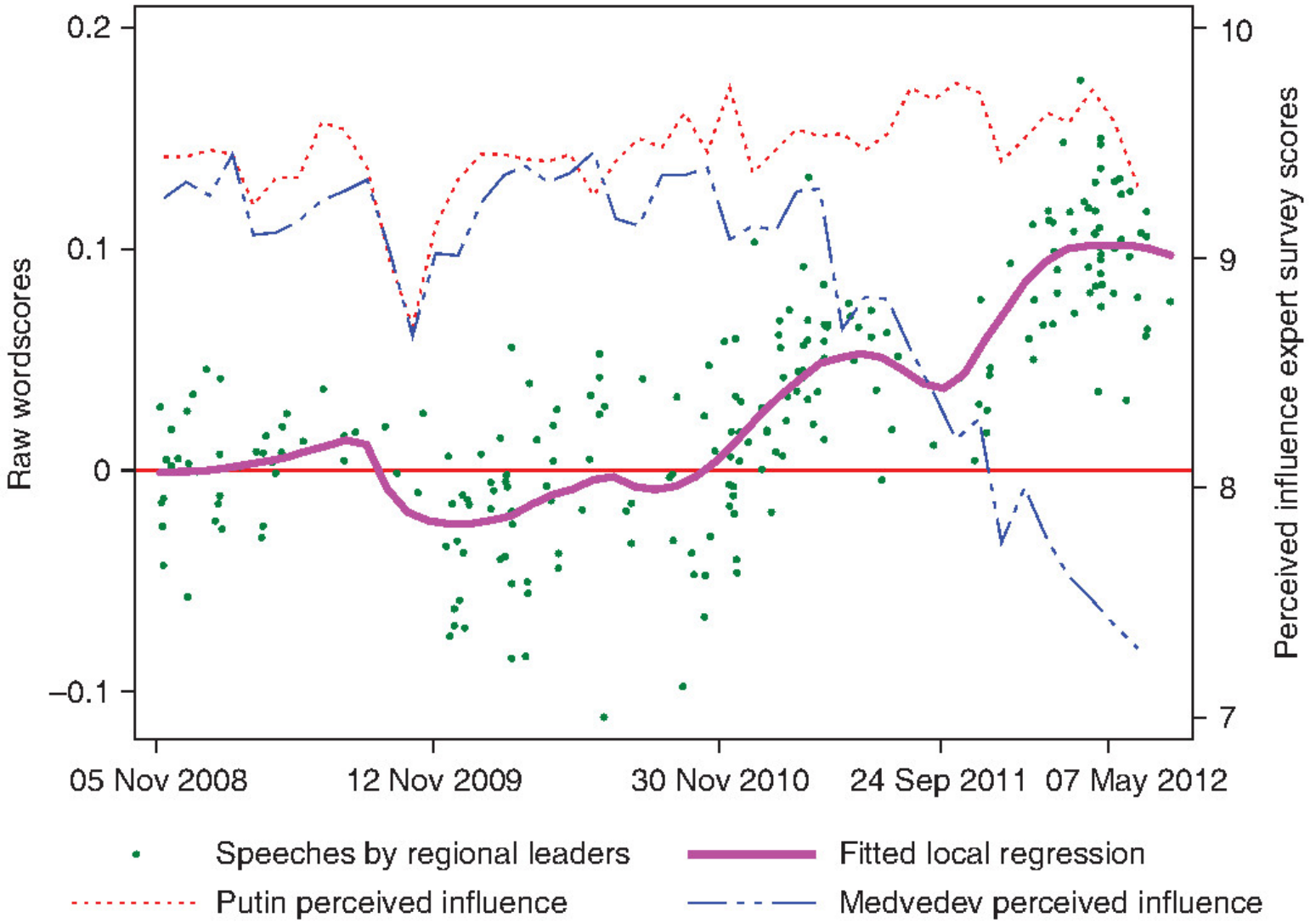
Laver and Benoit (2002)

BOX PLOT OF STANDARDISED SCORES OF SPEAKERS IN 1991 CONFIDENCE DEBATE ON 'PRO- VERSUS ANTI- GOVERNMENT 'DIMENSION, BY CATEGORY OF TD



Benoit and Laver (2003)

Baturo and Mikhaylov (2013)



Outline

Item Response Theory

- If all students answer the same set of exam questions and we measure the score on each answer, then
- the difficulty of the question varies by question, but is the same across students, and
- the capability of the student varies by students, but is the same across questions.
- So if you estimate, on average, how students perform on a question, you have a measure of difficulty; if you estimate, on average, how a student performs across questions, you have a measure of capability.

Item Response Theory

- If all students answer the same set of exam questions and we measure the score on each answer, then
- the difficulty of the question varies by question, but is the same across students, and
- the capability of the student varies by students, but is the same across questions.
- So if you estimate, on average, how students perform on a question, you have a measure of difficulty; if you estimate, on average, how a student performs across questions, you have a measure of capability.
- Instead of questions and students, we have here words and documents, and instead of difficulty and capability we have ideological dimensions of term usage and documents.

Wordfish

$$word_{ij} \sim Poisson(\lambda_{ij})$$
$$\lambda_{ij} = exp(\alpha_i + \phi_j + \beta_j \theta_i)$$

α = document fixed effect (captures document length)
ϕ = word fixed effect (captures average frequency)
β = word-specific weight
θ = estimated ideological position

An earlier attempt at using a Bayesian model to estimate underlying dimensions of speech data is Monroe and Maeda (2004).

Basically a task-specific model that is somewhat related to the factor analysis approach.

(Slapin and Proksch, 2008; Grimmer and Stewart, 2013, 27)

Outline

Preprocessing

Stemming is related to **lemmatization** and concerns the removal of all endings of words, so that words such as "walking", "walk", "walks" all become the same word, "walk".

Other transformations include the removal of **stop words**, changing all text to **lower case**, removing **numerals**.

The final transformation is the removal of **sparse terms** that rarely occur in the texts.

The lab will show all the relevant code in R, made easy by using the `tm` package.

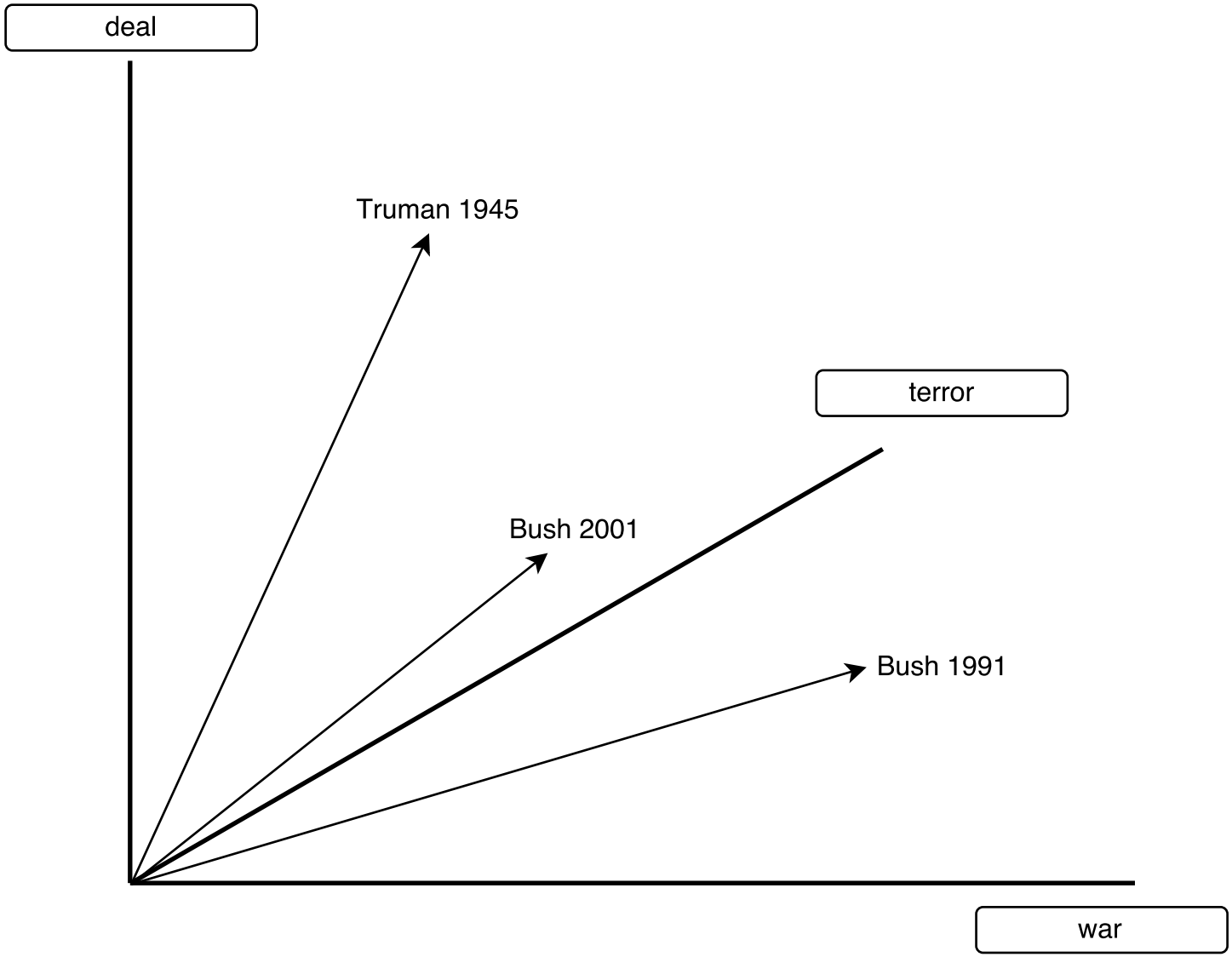
Project 3: State of the Union speeches

After cleaning the data (removing stop words, upper case, stemming, etc.) we can produce a data set where each observation is a document, each variable a word (term), and each observation the number of times a term appears in a document.

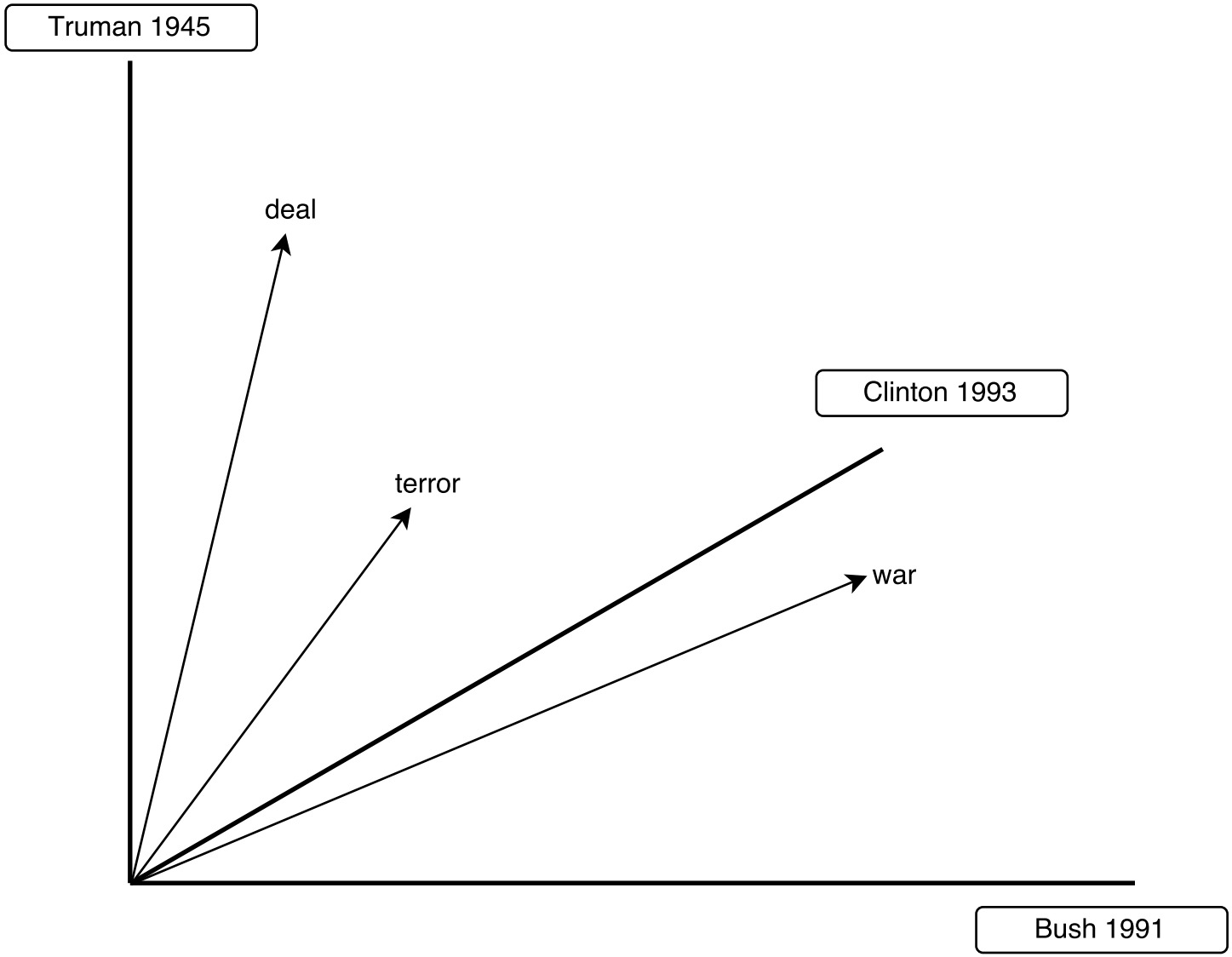
	achiev	act	action	ago	alreadi	also	america
Truman 1945	2	0	0	0	1	1	11
Truman 1946	15	38	21	3	25	43	2
Truman 1947	3	4	5	1	2	7	1
Truman 1948	11	6	4	2	1	9	0
Truman 1949	5	6	3	1	0	4	1
Truman 1950	18	2	4	4	3	9	0

This is the input matrix for earlier cluster analysis and principal components analysis, and it is also the input for Wordscores and Wordfish.

Text in high-dimensional space



Text in high-dimensional space



Bakker, Ryan, Catherine De Vries, Erica Edwards, Liesbet Hooghe, Seth Jolly, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen and Milada Anna Vachudova. 2015. "Measuring party positions in Europe: The Chapel Hill expert survey trend file, 1999–2010." *Party Politics* 21(1):143–152.

Baturo, Alexander and Slava Mikhaylov. 2013. "Life of Brian Revisited: Assessing Informational and Non-Informational Leadership Tools." *Political Science Research and Methods* 1(1):139–157.

Benoit, Kenneth and Michael Laver. 2003. "Estimating Irish party policy positions using computer wordscoring: The 2002 election–a research note." *Irish Political Studies* 18(1):97–107.

Benoit, Kenneth and Michael Laver. 2007. "Estimating party policy positions: Comparing expert surveys and hand-coded content analysis." *Electoral Studies* 26(1):90–107.

Budge, Ian, David Robertson and Derek Hearl. 1987. *Ideology, strategy and party change: spatial analyses of post-war election programmes in 19 democracies*. Cambridge University Press.

Budge, Ian, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara and Eric Tanenbaum. 2001. *Mapping policy preferences: estimates for parties, electors, and governments, 1945-1998*. Oxford University Press.

Castles, Francis G. and Peter Mair. 1984. "Left–right political scales: Some expertjudgments." *European Journal of Political Research* 12(1):73–88.

Cunningham, Kevin and Johan A. Elkink. forthcoming. Ideological dimensions in the 2016 elections. In *The post-crisis Irish voter: Voting behaviour in the Irish 2016 general election*, ed. David M. Farrell, Michael Marsh and Theresa Reidy.

Grimmer, Justin and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political Analysis* 21(3):267–297.

Laver, Michael and Kenneth Benoit. 2002. "Locating TDs in policy spaces: the computational text analysis of Dáil speeches." *Irish Political Studies* 17(1):59–73.

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Extracting policy positions from political texts using words as data." *American Political Science Review* 97(02):311–331.

Laver, Michael and W. Ben Hunt. 1992. *Policy and party competition*. Routledge.

Monroe, Burt L. and Ko Maeda. 2004. Talks cheap: Text-based estimation of rhetorical ideal-points. In *Annual Meeting of the Society for Political Methodology*. pp. 29–31.

Mosteller, Frederick and David Wallace. 1964. "Inference and disputed authorship: The Federalist.".

Poole, Keith T. and Howard Rosenthal. 1985. "A spatial model for legislative roll call analysis." *American Journal of Political Science* pp. 357–384.

Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. "A scaling model for estimating time-series party positions from texts." *American Journal of Political Science* 52(3):705–722.