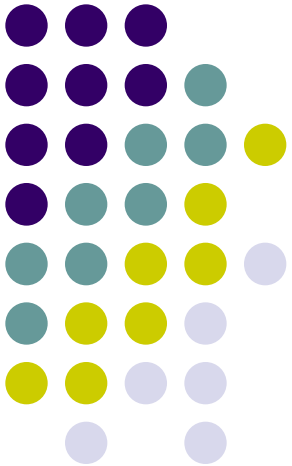
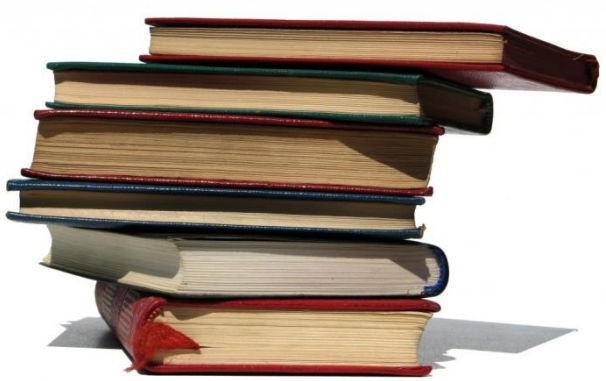
Applied Scaling & Classification Techniques in Political Science

Lecture 2

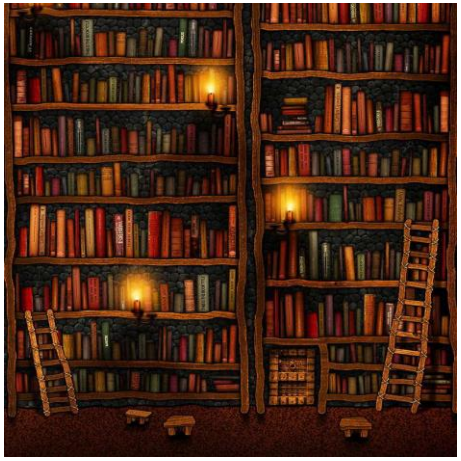
Unsupervised scaling algorithms: Wordfish



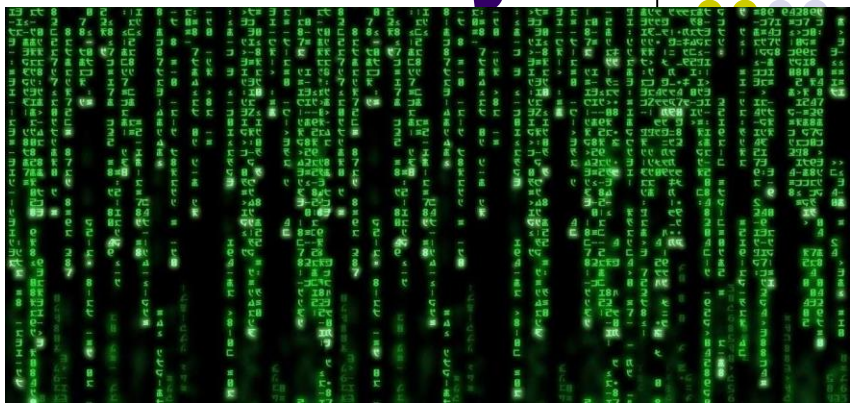
But before that…a summary



You pass to R the texts you want to analyze via `readtext`



You tell to R that those bunch of texts belong to the *same collection of texts* you want to analyze via `corpus`



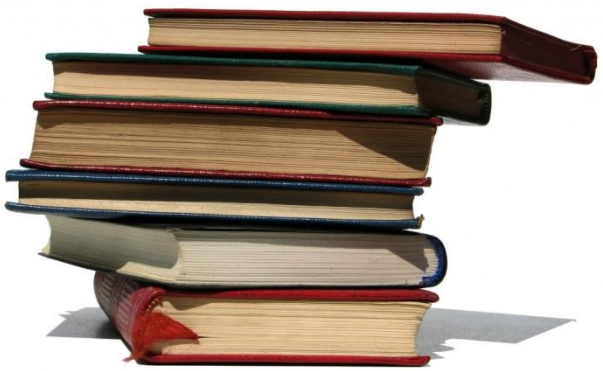
You extract from the corpus the relative document term (or feature) matrix via `dfm`. By doing that you apply the bag of words approach to that corpus of texts

But before that…a summary

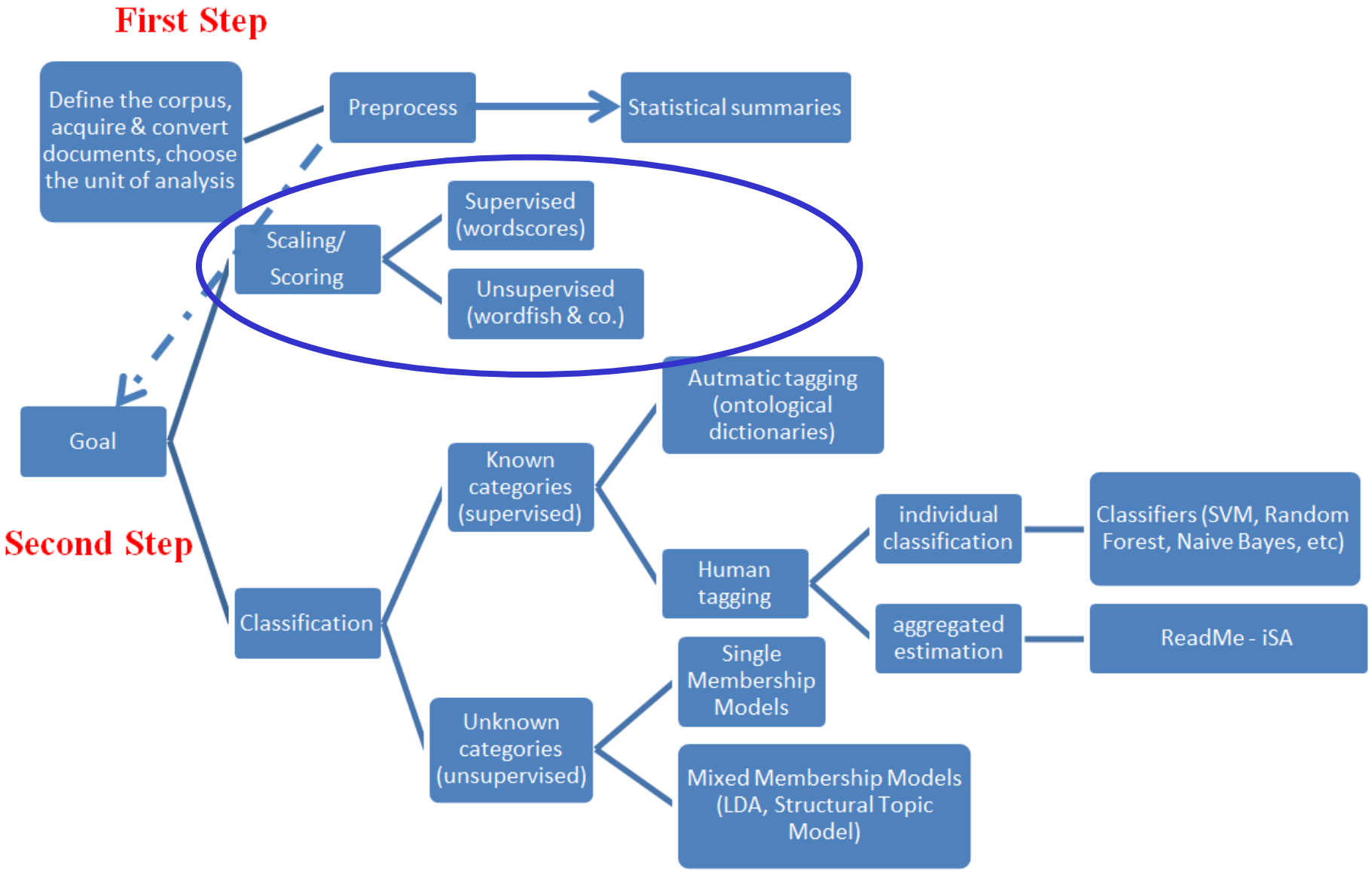
All the statistical models that we will see, **work on this**

docs	voto	programma	priorità	punti	piano	sostegno	famigli	natalità
FDI	1	5	4	5	7	11	4	1
FI	0	2	0	1	11	7	3	1
LEGA	1	9	5	6	18	32	7	3
LEU	0	2	1	0	15	7	3	0
M5S	6	18	13	12	45	20	22	0
PD	1	11	12	8	38	23	25	3

NOT on this



Our Course Map



References

- ✓ Proksch, Sven-Oliber, and Slapin, Jonathan B. 2008. “A Scaling Model for Estimating Time-Series Party Positions from Texts”. *American Journal of Political Science*, 52(3): 705-722.
- ✓ Proksch, Sven-Oliber, and Slapin, Jonathan B. 2009. “How to Avoid Pitfalls in Statistical Analysis of Political Texts: The Case of Germany”. *German Politics*, 18(3): 323-344
- ✓ Egerod, Benjamin C.K., and Robert Klemmensen (2020). Scaling Political Positions from text. Assumptions, Methods and Pitfalls. In Luigi Curini and Robert Franzese (eds.), *SAGE Handbook of Research Methods is Political Science & International Relations*, London, Sage, chapter 27
- ✓ Curini, Luigi, Airo Hino, and Atsushi Osaki. 2018. “Intensity of government–opposition divide as measured through legislative speeches and what we can learn from it. Analyses of Japanese parliamentary debates, 1953–2013”. *Government and Opposition*, DOI: 10.1017/gov.2018.15

Latent models

Textual data might focus on **manifest characteristics** whose significance lies primarily in **how they were communicated** in the text

To take an example, if we were interested in whether a political speaker used **racist language**, this language **would be manifest directly in the text itself** in the form of racist terms or references, and what would matter is whether they were used, not so much what they might represent

Latent models

However, sometimes the target of concern is not so much **what the text contains**, but what **its contents reveal as data about the latent characteristics** for which the text provides ***observable implications***

Is this important? YES!

Latent models

In the study of politics, some of our important theories about political and social actors concern qualities that are **unobservable through direct means**

Ideology, in particular, is fundamental to the study of political competition and political preferences, but we have **no direct measurement instrument** for recording an individual or party's relative preference for (for example) socially and morally liberal policies versus conservative ones

That is, ideology is not something that the researcher can **directly observe**…rather it must be indirectly estimated based upon **observable actions** taken by actors

Latent models

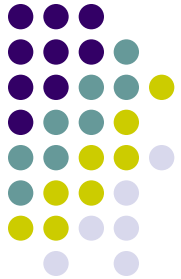
Observable actions…such as?

Roll-calls, for example! Still, voting in a legislature is subject to **party discipline** and may be **highly strategic**….and so?

Let's rely on something else then…like…what?!?

Texts, of course!

Scaling methods



The goal of methods **for scaling positions** is to use **some observed set of outcomes** to draw inferences about an actor's (in the widest sense of the word) unobservable position on a **latent dimension** relative to other actors

Position is here to be understood as the **preference on some dimension**. To get at such a position, the **observed outcomes** must reveal some kind of preference on the part of the actor

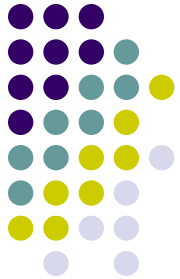
Scaling methods

When scaling the political positions of a corpus of texts, we can view the **choice of words as the observed outcome**

Whenever **certain statements are associated with particular political positions**, we can use them to discriminate between positions in a certain political space

In other words, the use of a particular (set of) word(s) provides us with a revealed preference for a specific (kind of) policy

Scaling methods



Whenever we can think of the **data generating process** in these terms, **scaling** intended as a way "to isolate a specific dimension on which texts are to be compared and provide a point estimate of this quantity, on some continuous scale", might be feasible

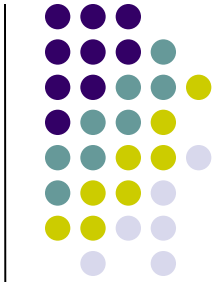
Moreover…estimating locations using existing data is often difficult and sometimes impossible…but nearly all political actors speak (or write)!

Scaling methods

(text) Scaling methods are designed to isolate a specific dimension on which texts are to be compared and provide a point estimate of this quantity, on some continuous scale

Such dimension could be related to ideology, or to some other policy (or non-policy) space

Scaling methods



Scaling methods can be differentiated between **Supervised & Unsupervised Methods**

Supervised models use human input, typically in the form of a set of reference texts that have been already validated (i.e., already classified as left, right, extreme right texts for example)

These estimates can then be used to predict (i.e., scale) the positions of texts the model has not encountered previously (i.e., virgin texts)

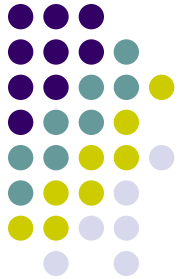
The reference texts also serves to **define the policy space** that the researcher seeks to estimate (if you use a set of reference-texts validated over a left-right economic scale, you will scale the virgin texts along such scale. More on this later on)

Scaling methods

Unsupervised Methods simultaneously learn about the latent space and estimate document positions in it, without input from the researcher, i.e., they "*discover*" words that distinguish locations on some dimensional spectrum (**not defined a-priori** as it happens in the case of the supervised scaling methods)

How is possible? Give me a moment…

Scaling methods



Which Assumptions are needed to Scale a Text?

First, even though it is called **text scaling**, what we most commonly want to draw inferences about is not the **political position** articulated in the text, but the **preference** held by the author

But if the cost to articulating a position is low, authors' might engage in cheap talk. Conversely, if costs are high, they might choose not to articulate the position for strategic reasons

All the scaling techniques we focus on, assume that authors do **not censor their statements for political reasons**. This assumption, in some given circumstances, could however cause significant measurement error

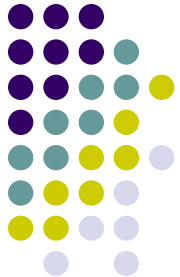
Scaling methods

Second, we need to make assumptions about how any given author translates her position into text, and how that relates to the other authors in the corpus

Specifically, the language used in the **texts must discriminate between the intended messages of different authors**. In other words, the authors should receive varying levels of utility from their choice of words, and this variation should be related to the political space, we want to measure

If authors of different preferences receive the same utility from similar choices of words, we cannot use the texts to discriminate between their positions.

Scaling methods



Accordingly, the documents **should be informative about the political differences** we seek to estimate

Particularly in contexts where there are **strong common norms about how to phrase a documen**t (as with highly technical legislative or legal documents) or the texts do not communicate any preference at all, it can be difficult to scale documents

An interesting special case of incomparability is when authors simply use **different languages**

Scaling methods

Third, and regarding the relation between documents, a set of texts is only be scalable, if they can be placed in the **same Euclidian space**

A possible violation in this respect would be if the language used in the documents is **incomparable** in the way meaning is ascribed to words

Analyzing text that is produced under very different conditions or in varying contexts; that are from different time periods or actors; or have very different audiences in mind would make it difficult to place them relative to each other, let alone in the same space

A similar problem can apply if texts refer to **different dimensionality** of the space

Wordfish

Unsupervised methods for scaling texts produce estimates using **only the information available** in the textual data itself

How to do that?

Let's introduce **Wordfish**!

Wordfish

Wordfish assumes that the **language** used by political actors expresses political ideology, that is…

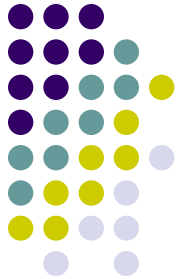
…**Ideology** manifests itself in the **word choice** of politicians when writing party documents or saying something for example

Wordfish

More specifically, Wordfish assumes that parties' **relative word** usage within party documents conveys information about their positions in a policy space

To give an example, the technique assumes that if one party uses the word '**freedom**' more frequently than the word '**equality**' in a document on economic policy while another party uses 'equality' more often than 'freedom' in a similar document, these **two words** – 'equality' and 'freedom' – **provide information** about party ideology with regard to an underlying policy dimension, and **discriminate** between the parties

Wordfish

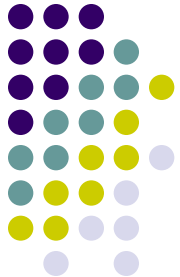


The interpretation of the **estimated dimension** in Wordfish **is then completely left to the researcher**

In the previous example, Wordfish **does not tell the researcher** whether ‘equality’ is a ‘left-wing word’ while ‘freedom’ is a ‘right-wing word’

The algorithm will simply use the relative frequencies of these words as data to locate the documents on a scale, and it is up to the researcher to make an assessment about what constitutes ‘left’ and ‘right’ based upon her **knowledge of politics** (*a-posteriori* method!)

Wordfish Estimation Process



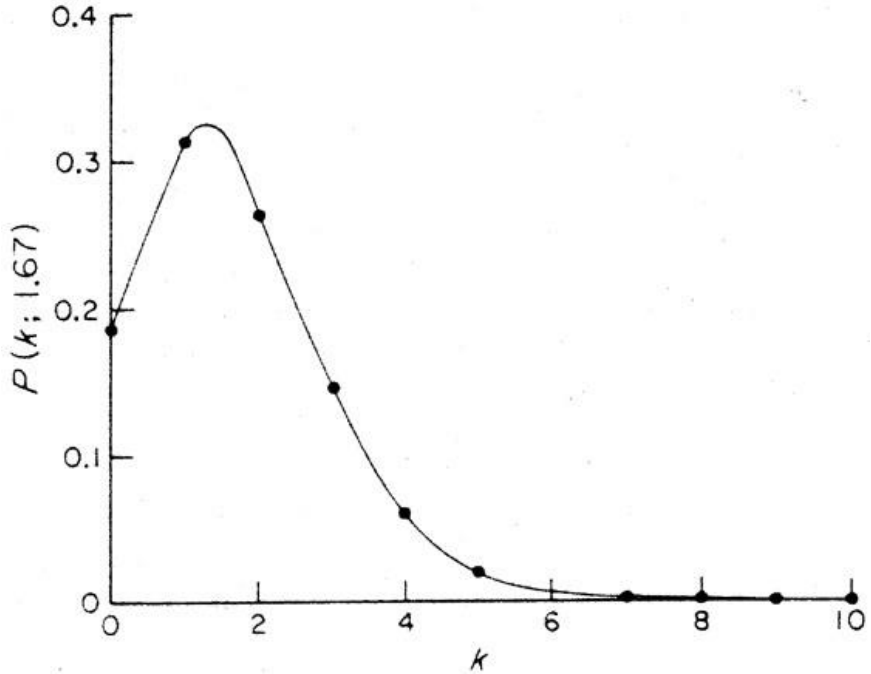
We already told that Wordfish, as all unsupervised scaling methods, "*discovers*" words that distinguish locations on a political spectrum

This is made possible by adopting some statistical assumptions on the **distribution of words** employed in texts

Wordfish Estimation Process

But which is the **statistical distribution** which most **accurately approximate word usage**?

Wordfish assumes that word frequencies (the number of times an actor *i* mentions word *j)* are generated by/drawn from a **Poisson process**, a distribution that is **heavily skewed**, as is the case of **word usage**



Wordfish Estimation Process

The **systematic component** of this process contains 4 parameters: 1) *word fixed effects*; 2) *document fixed effects*; 3) *document positions*; 4) *word weights* (discriminating parameters)

Word fixed effects are included to capture the fact that some words need to be used **much more often** in a language

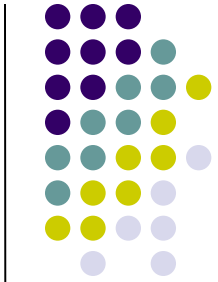
Such words may serve a grammatical purpose but they have no substantive or ideological meaning, such as conjunctions or definite and indefinite articles

Wordfish Estimation Process

The **document fixed effect** parameters control for the possibility that some documents in the analysis may be **significantly longer** than others

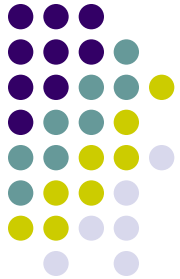
When using manifestos to estimate party positions, for example, this can happen when some parties in some years write much longer manifestos

Wordfish Estimation Process



The **document positions parameters** tells us the positions of each document relative to the other documents in the recovered latent space

Wordfish Estimation Process



The **word discrimination parameters** allow the researcher to analyze **which words differentiate documents (party) positions**

In previous example, '*equality*' would have a high absolute value for its discrimination value and its usage would most likely be associated with left-wing documents (and parties). The word '*freedom*' would also have a high absolute value **but with the opposite sign** because its usage would be associated with right-wing parties

This allows the researcher to estimate party positions and uncover the variations in political language that are responsible for placing parties on this dimension

More formally

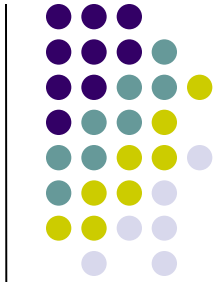
Formally the functional form of the model is as follows: $y_{ijt} \approx POISSON(\lambda_{ijt})$ where y_{ijt} is the count of word *j* in document *i*'s (i.e., party manifesto; speech; etc.) at time *t*

The lambda parameter has the following systematic component:

$$\lambda_{ijt} = exp(\alpha_{it} + \psi_j + \beta_j * \theta_{it})$$

with α as a set of **document fixed effects at time t**, Ψ (psi) as a set of **word fixed effects**, β as estimates of **word specific weights** capturing the importance of word *j* in discriminating between documents (manifestoes or speeches), and θ (theta) as the estimate of document (i.e., party if we are talking about parties' documents) *i*'s **position** at time *t* (therefore *it* is indexing one specific document)

More formally



WORDFISH uses an **expectation maximization (EM) algorithm** to retrieve maximum likelihood estimates for all parameters

The implementation of this algorithm entails an **iterative process**:

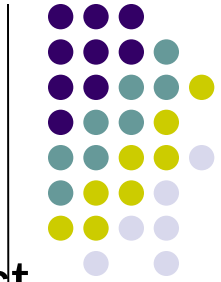
first document parameters are held fixed at a certain value while word parameters are estimated, **then** word parameters are held fixed at their new values while the document parameters are estimated

This process is **repeated until the parameter estimates** reach an acceptable level of convergence

Some challanges

1. Document processing
2. Interpretation
3. Dynamic pattern
4. Problems with confidence intervals

Document Processing



Document processing is essential and possibly the most tricky task in the estimation process in Wordfish (and not only for this method…)

Researchers should predefine the sets of texts to be analyzed

The model specification used by Wordfish works best as **more data is available**, meaning as **more documents** are used in the analysis and as those documents contain **more unique words**

If the documents do not contain a sufficient number of unique words, there will not be adequate information to estimate document parameters

Document Processing

Moreover, Wordfish will recognize differences in word use between two texts as indicative of their different political positions, but in reality these differences could be also due to the topics addressed by the authors

A special case of this, is in situations where texts use **completely incomparable language** or do not address **similar topics at all**

In these situations they cannot be scaled together, and if they are, it will often result in the main policy dimension being miss-specified

Document Processing

The selection of texts will depend on what kind of **policy dimension** should be analyzed

Wordfish estimates a **single policy dimension**, and the information contained in this **dimension depends upon the texts** that the researcher chooses to analyze

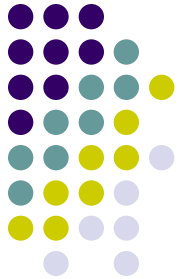
Therefore, the **selection of texts should depend** on the particular policy dimension the researcher wishes to examine

Document Processing

For instance, if a researcher is interested in comparing **foreign policy statements** of parties in country X, then only such texts should be included in the analysis

On the other hand, if the research question is to determine a **general ideological position** using all aspects of policy (e.g. left-right), then the analysis should potentially be conducted using all parts of an election manifesto, for example, assuming that such documents are encyclopedic statements of policy positions

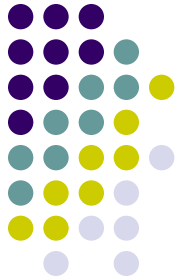
Document Processing



The estimated single dimension **will thus be a function** of the selection of the text corpus

This also implies that when the generative model specifies a unidimensional policy space, when it really is multidimensional, we risk miss-specifying the policy dimension

Document Processing



WORDFISH does not estimate **multiple dimensions**, only a single dimension, but it does allow the estimation of **different dimensions** if you use different text sources

For instance, if your interest is in estimating positions of **presidential candidates on foreign policy and economic policy**, then you could estimate separate positions using foreign policy speeches only on the one hand and economic policy speeches on the other hand and from this creating a **2-dimensional space**

Interpretation

Position estimates derived using Wordfish are based **only on the information in the texts**

This lack of an ex ante defined dimensionality is a **double-edged sword**: while Wordfish scales texts independently of prior information, it renders **uncertain** the exact nature of the dimension being estimated (as it happens in all unsupervised approaches!)

One important drawback of unsupervised algorithms is thus that the nature of the dimensions produced requires **intensive validation** before they can be applied across different sets of texts and contexts

Interpretation

Improving the validation of scales will help improve current models, which quite often rely on the strong assumption of **ideological dominance in speech** (i.e., that actors' ideological leanings determine what is discussed in texts)…sometimes this makes sense, other times no!

This is **not a shortcoming** of Wordfish!

In fact, next week we will see how also (non)ideological locations that Wordfish eventually identifies can be quite useful!

This simply suggests that one **should not blindly assume** that Wordfish output measures an ideological location of political actors without careful validation

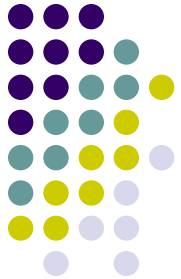
Dynamic Estimation

Using text to estimate party positions **over time** creates an additional challenge. On the one hand, we would like to use as much information in the texts as possible. On the other hand, we would like to estimate position change over time. This is a trade-off

For example, if the **political debate changes and new vocabulary** enters the political lexicon in election *t*, then this will differentiate the texts at point *t* from those at point *t-1*

In fact, in this instance, we are likely to pick up a **policy agenda** shift in texts, whereas we are interested in party position change

Dynamic Estimation



Potential route to addressing this issue: carefully select the **words** that enter the analysis!!!

Thus, if there is movement of parties, it can only be due to **different word usage**

This requires that the **word data over time** must be comparable at a minimum level

Dynamic Estimation

Take as an example the set of parties' manifestos in Germany since 1970 to 2005. Assume that you want to analyze such documents with Wordfish

Now assume that the political lexicon in the manifestos at election time *t* contains an issue that is no longer relevant at time *t+1*, e.g. official relations with the GDR (East Germany)

If all parties make a statement with regard to the GDR at point *t* but not at *t+1*, then the words **will not only distinguish** parties at point *t*, but also **distinguish the elections**

As a result, if all words are counted, even the rare ones, the parties are more likely to be **clustered by election**

Dynamic Estimation

The same is true if we have some changes in the **actual meaning** of some political words

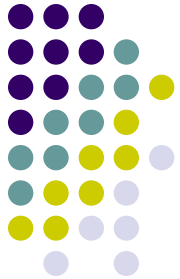
Which word inclusion criteria then?

Two (main) options

Dynamic Estimation

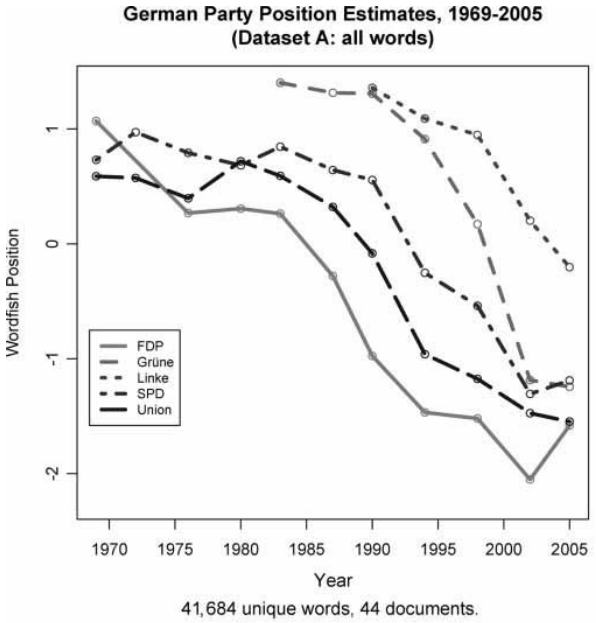
First alternative (non-informative priors): in the term-document matrix includes words that are **mentioned in a minimum number of documents** (say, in at least 20%), thus essentially keeping words that are deemed important enough to be mentioned either over time by one party or by several parties

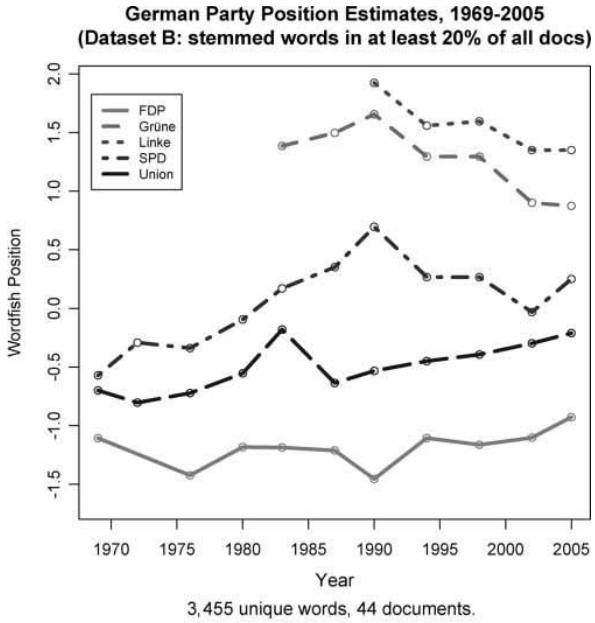
Dynamic Estimation

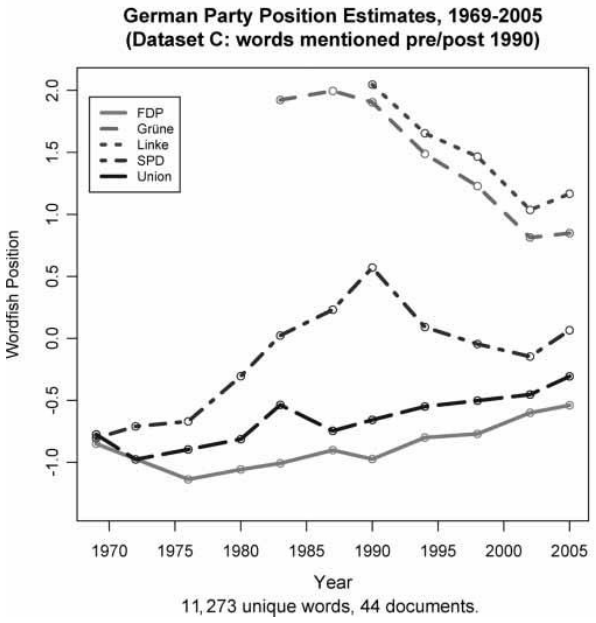


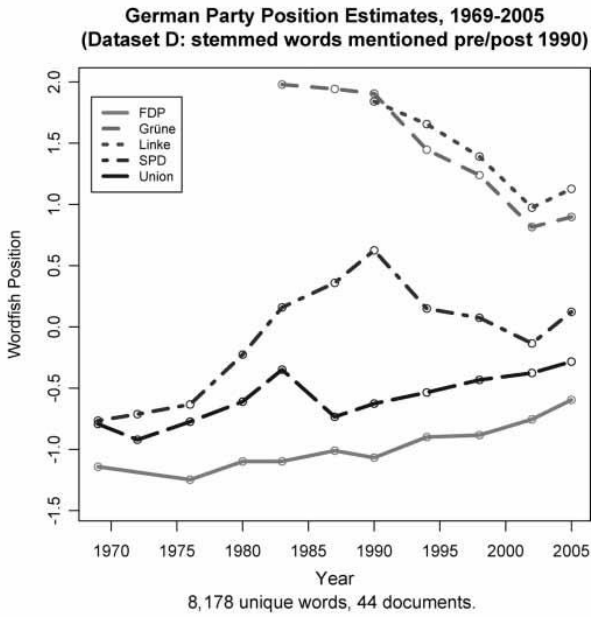
Second alternative (informative priors): in the term-document matrix includes **only those words that appear both pre- and post-1990**, i.e., reunification added words to the German political lexicon that were not in it previously. Likewise, some words that were previously important likely fell out of use.

If we do not control for this fact, we would see a **large jump** in all parties around 1990 as they all shift their word usage to account for new political realities


German Party Position Estimates, 1969-2005
(Dataset A: all words)
Wordfish Position
Year
FDP
Grüne
Linke
SPD
Union
41,684 unique words, 44 documents.


German Party Position Estimates, 1969-2005
(Dataset B: stemmed words in at least 20% of all docs)
Wordfish Position
Year
FDP
Grüne
Linke
SPD
Union
3,455 unique words, 44 documents.


German Party Position Estimates, 1969-2005
(Dataset C: words mentioned pre/post 1990)
Wordfish Position
Year
FDP
Grüne
Linke
SPD
Union
11,273 unique words, 44 documents.

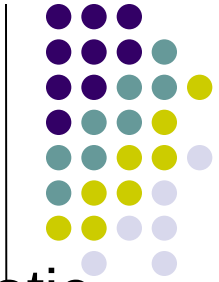

German Party Position Estimates, 1969-2005
(Dataset D: stemmed words mentioned pre/post 1990)
Wordfish Position
Year
FDP
Grüne
Linke
SPD
Union
8,178 unique words, 44 documents.

Dynamic Estimation

As suspected, **agenda effects over time** dominate the results when all words are used

Excluding **words that are specific to a given time-period** induces stability and the results are corroborated by their good face validity

Confidence intervals

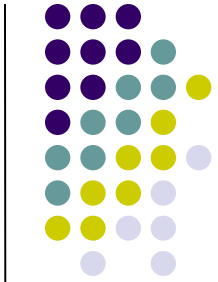


Wordfish in the Quanteda package implements asymptotic standard errors. These SEs rely however heavily on the model being correctly specified. As a result such SEs should really be treated with care cause quite often they will be too small

As a way of obtaining uncertainty estimates with weaker assumptions, Lowe and Benoit (2013) also introduced a **bootstrap procedure**, that basically iterates across different (bootstrapped) samples of the original DfM and then average the results

The Quanteda package supplies functionality for random sampling of Words [`dfm_sample`], which can be used to implement the above bootstrap procedure with relative ease

Confidence intervals



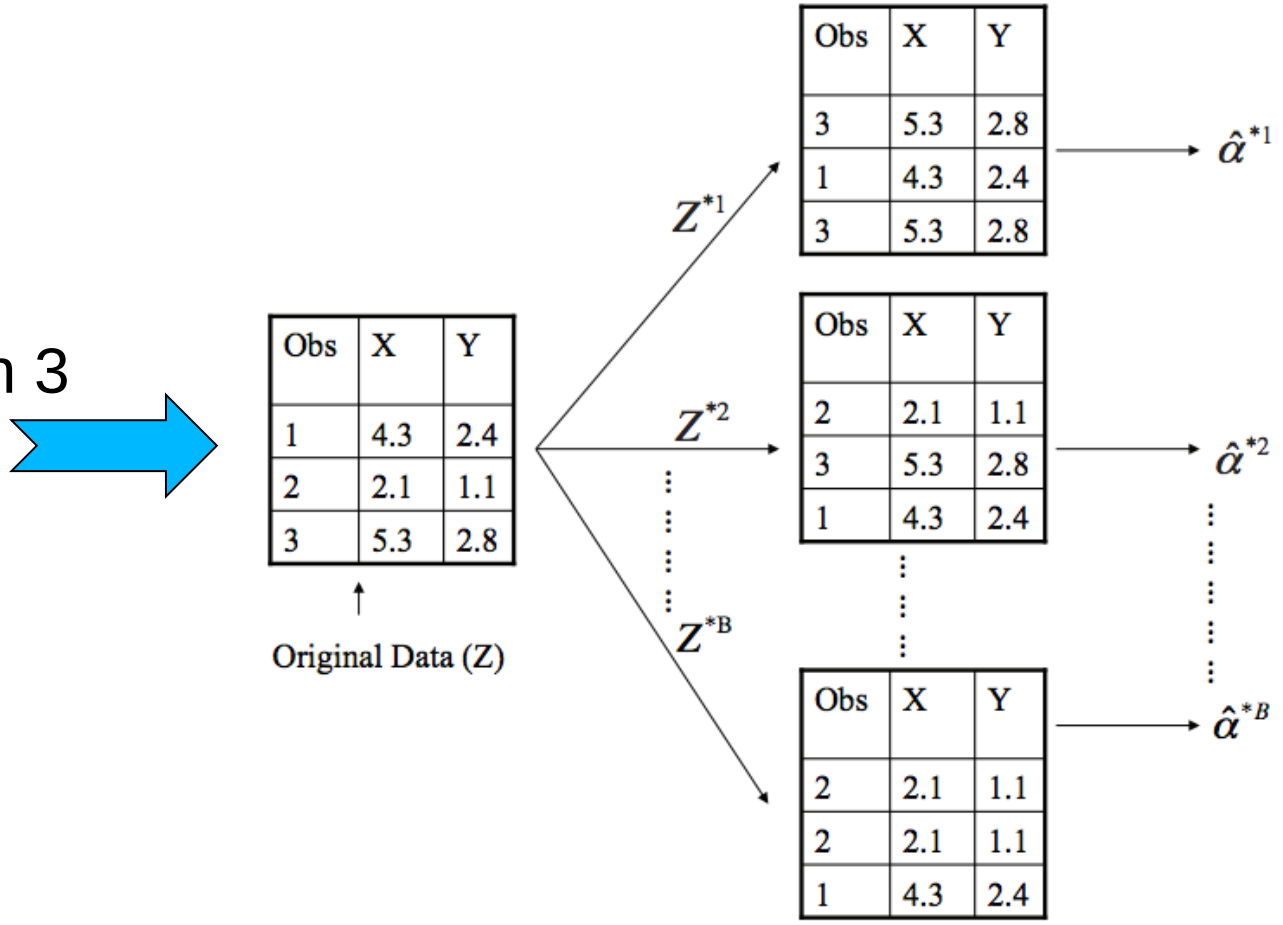
What do we mean by **bootstrapping**?

In essence bootstrapping rep**eatedly draws independent samples** from our data set to create bootstrap data sets. This sample is performed with *replacement*, which means that the same observation can be sampled more than once

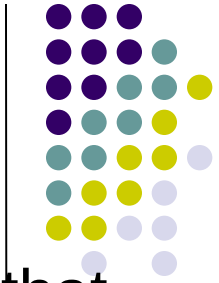
Each bootstrap data set contains *n* observations, sampled with replacement from the original data set. Each bootstrap is the used to compute the estimated statistic we are interested in (i.e., a mean or anything else – as the thetas of a Wordfish model!)

Confidence intervals

An example with 3 resamples

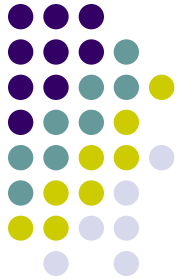


Confidence intervals



Bootstrapping is an extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method

We can in fact use all the bootstrapped data sets to compute the standard error of the desired statistics, or their 95% confidence intervals, etc.



An application of Wordfish to Japanese parliamentary debates, 1953–2013

The theoretical framework

Measuring **how confrontational parties** are within a legislature and in particular the ‘distance’ between cabinet and opposition parties (i.e. the extent to which a government and its opposition oppose each other) is a relevant **political metric** that explains several important facts (the ability of a cabinet to change the status quo, its survival, etc.)

Usually such distance is measured in terms of **ideological/policy distance**

But is that enough?

Beyond ideology?

After all, the **line of conflict** between government and opposition can underline **not only** the mere ideological distance between parties, but also **several other factors**, among them:

- ✓ mutual (dis)trust
- ✓ evolving parliamentary dynamics
- ✓ past behaviours
- ✓ forward expectations

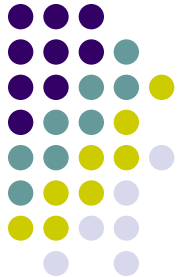
Beyond ideology?

As a result, the cabinet-opposition divide in some given circumstances could be **much (less) larger** than what would appear based on ideological considerations

How we estimate **the level of confrontational among parties within a parliament** (i.e. the actual content of parties' relative positions) is in few words very important!

This point has **substantial theoretical (and empirical) consequences**

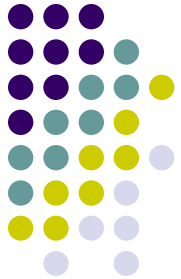
Beyond ideology?



In this sense, **words matter**!!!

…by focusing on the **type of words** that different political actors employ to express their positions with respect to the cabinet during a parliamentary debate, **we could be in a better position** to assess their relative degree of distance (the by-product of the several factors mentioned above) in that precise moment

Beyond ideology?



But be aware…

…as the language spoken on the floor is primarily directed at other delegates, cabinets, or opposition parties **rather than to voters**, it could be expected that the dimension of conflict (and cooperation) would be **possibly different from the ideological one** often found in different political texts primarily prepared for election campaigning!

Beyond ideology?

We demonstrate this by analysing the speeches made by prime ministers and party representatives in the parliamentary sessions of the Japanese Diet from 1953 to 2013 using the Wordfish algorithm

Why Japan?

1. The Japanese Diet is known for its adversarial nature
2. Japan shows a relatively high number of changes of cabinet
3. A quasi-experimental setting (pre- and post- 1993)

The Japanese case

We have selected all the speeches in which Prime Minister makes a general policy speech (*shoshin hyoumei enzetsu*) in the following situations:

i) after being nominated in the Special session

ii) after having succeeded a predecessor during a parliamentary session

iii) and in the beginning of the Extraordinary session

The Japanese case

Overall 439 speeches over 82 sessions, and almost 20,000 words/kanji

URL to get access to Japanese legislative speeches:

http://kokkai.ndl.go.jp/

Of course, we **tokenized** all the texts!!!

The temporal challenge

Using texts to estimate positions **over time** is quite tricky (remember!)

We chose to include in the analysis only words that fulfill a **minimum threshold** criterion based on informative priors**,** i.e., we kept in the analysis only those words that appear both pre- and post-1990

Choosing different temporal breaking points (such as 1993 or the early 80s when a change in the meaning of ideology seems to have happened in the Japanese case: Jou and Endo 2016) does not affect any of our conclusions reported below

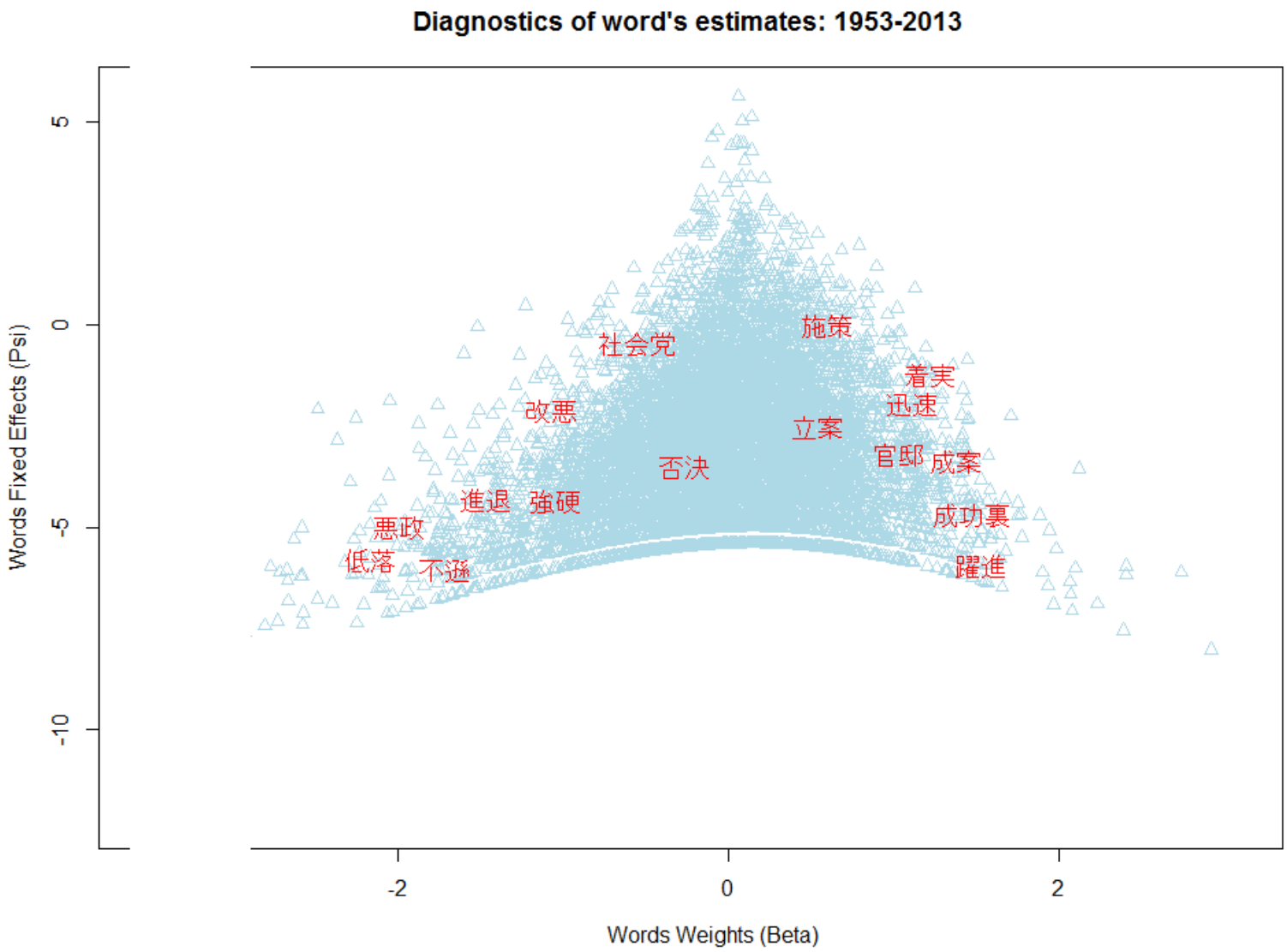
The same results are found if we use non-informative priors

The Japanese case

After this normalization, the average number of words for a typical legislative speech is 4119.7 (standard deviation: 1408.5)

The relative **large number of words** is reassuring, given that it has been shown that WORDISFH tends to estimate positions (more) accurately as the number of words increases

The discriminating words



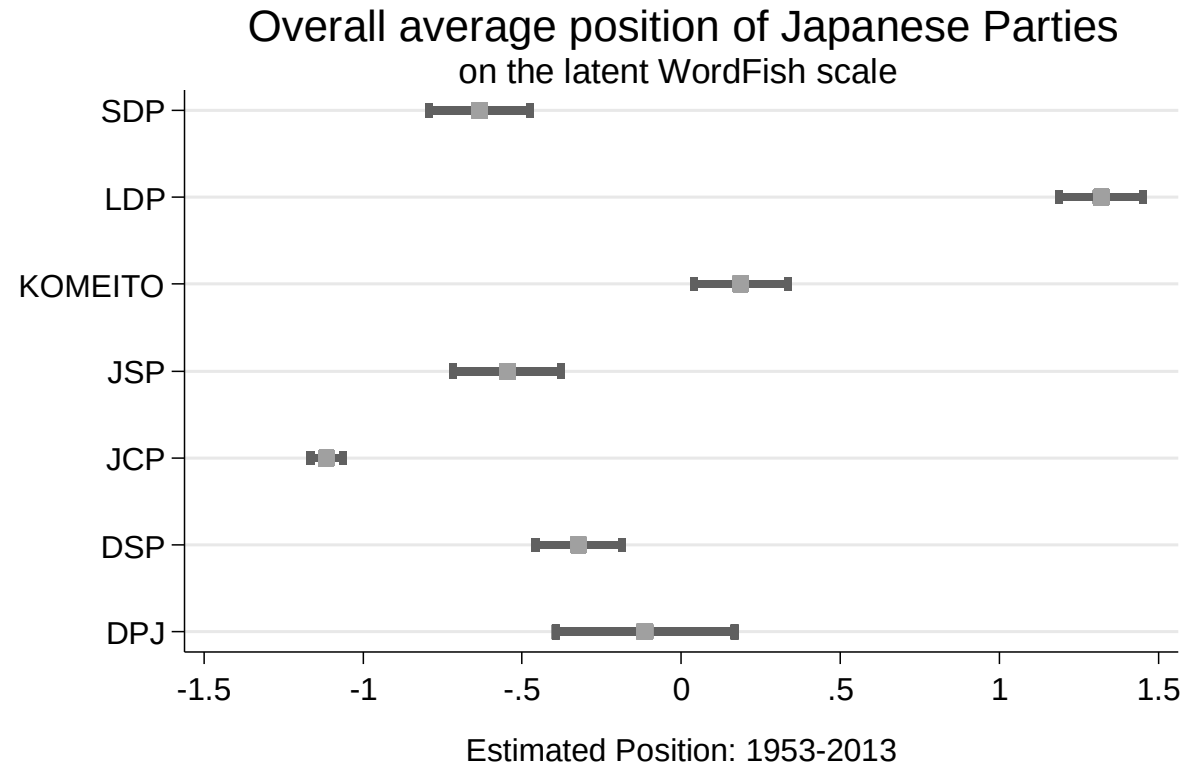
The discriminating words

Positive betas: *breakthrough, successfully, bills passed, steady, prompt, policy measure, policy making*

Negative betas: *decline, misgovernment, arrogance, decision to leave from a position, deterioration, by force, rejecting bills*

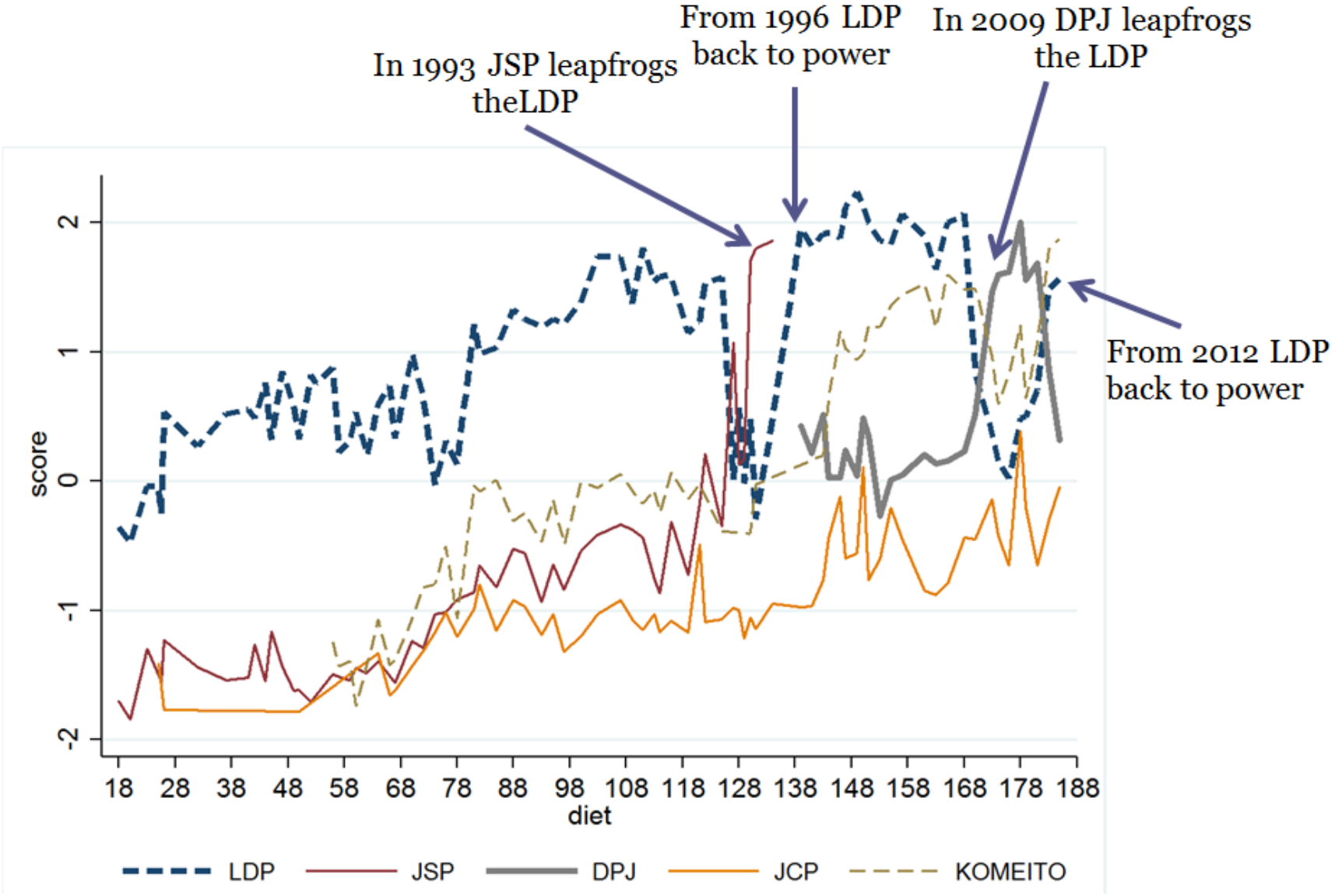
The two opposite sides of the words spectrum seem to define **different attitudes toward government very well**: a positive one (words with a positive beta) and a negative one (words with a negative beta)

And the speech positions?



Source: Word Fish analysis on Legislative Speeches 1953-2013

However, if we break down the estimated positions from legislative speeches over time…

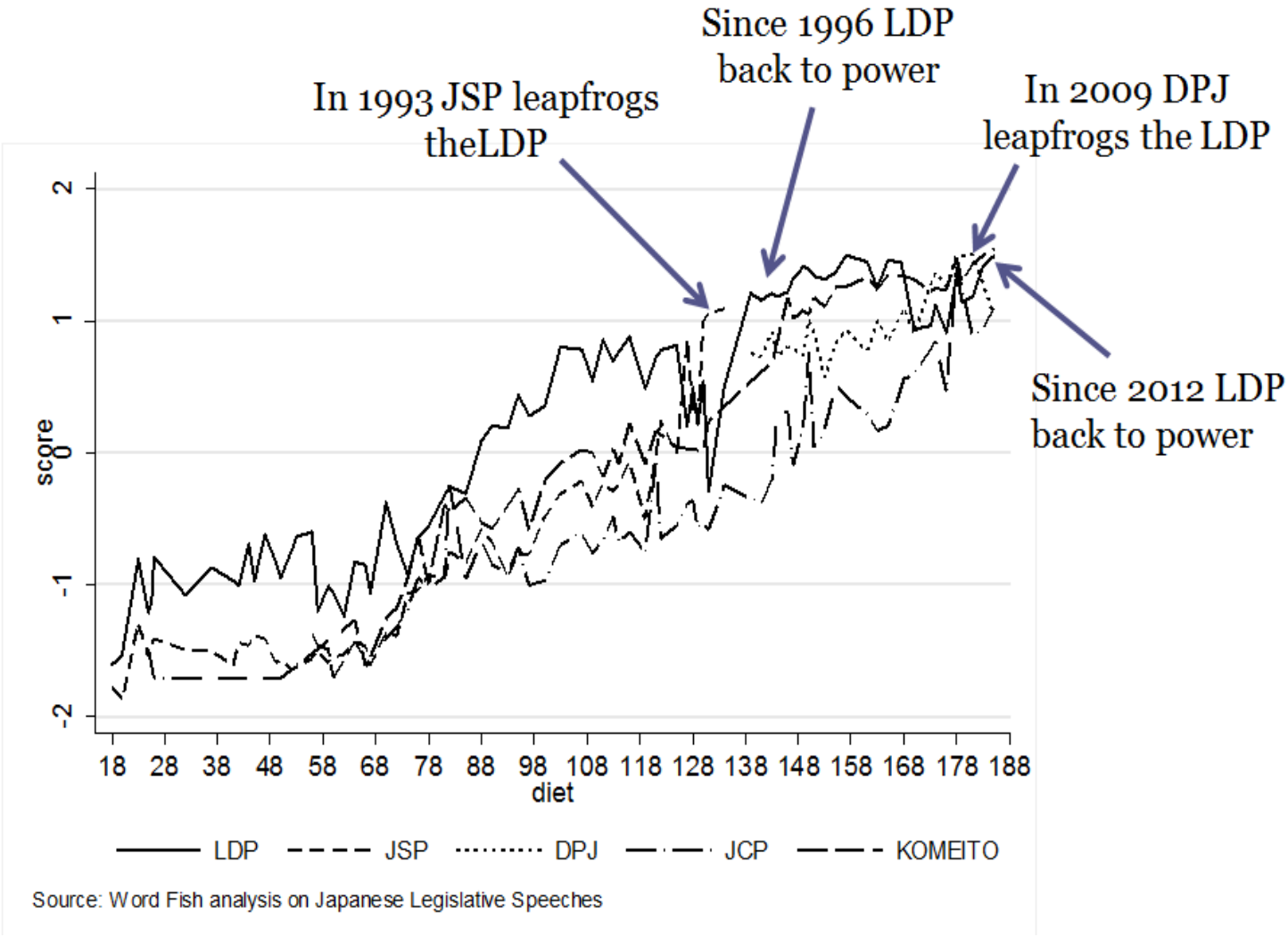

In 1993 JSP leapfrogs theLDP
From 1996 LDP back to power
In 2009 DPJ leapfrogs the LDP
From 2012 LDP back to power
score
2
1
0
-1
-2
18 28 38 48 58 68 78 88 98 108 118 128 138 148 158 168 178 188
diet
LDP
JSP
DPJ
JCP
KOMEITO

Findings

The extracted scores appear clearly related to the **confrontational nature of Parliamentary institutions**, and therefore to a government-opposition divide

This is not an artifact of WORDFISH **plus** Japanese language (*Hone* + *Tatemae*)! In Proksch et al. (ES 2011), Japanese parties were clearly located according to their ideological position when analyzing party programs with WORDFISH

Note that we have HUGE problems if we consider **all the words** over the entire period (as in the German case)!



Intensity of Government and Opposition (IGO)

We use the WORDISFH scores to estimate a measure of the **intensity** of the government-opposition divide over each session

To this aim, we adapted the **Dalton's index** (2004) of party system polarization, that, except for a constant, is mathematically the weighted population standard deviation of party positions in a given country

Of course, in our case, such index of polarization is based on positions that **go beyond ideology** to include several other factors related to the government-opposition divide

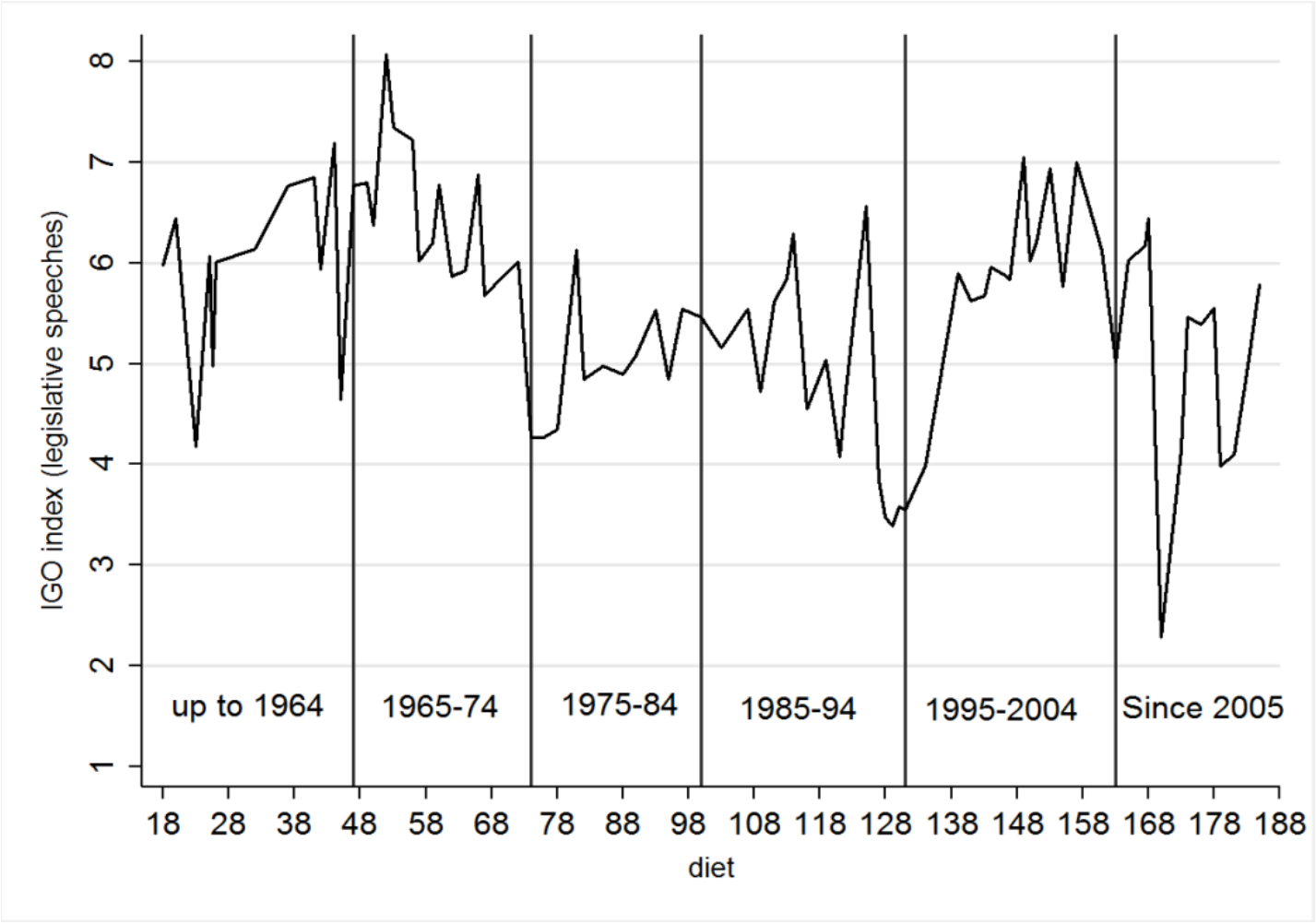
Intensity of Government and Opposition (IGO)

$$IGO_k = \sqrt{\sum_{j=1} VS_{jk} * \left(\left[P_{jk} - \bar{P}_k\right]/5\right)^2}$$

where IGO_k is the value of IGO during the parliamentary session k, VS_{jk} is the seat share of party j during session k, P_{jk} is the position of party j during session k over the latent government-opposition scale, and $\bar{P}_k$ is the average position of parties along the same scale during session k

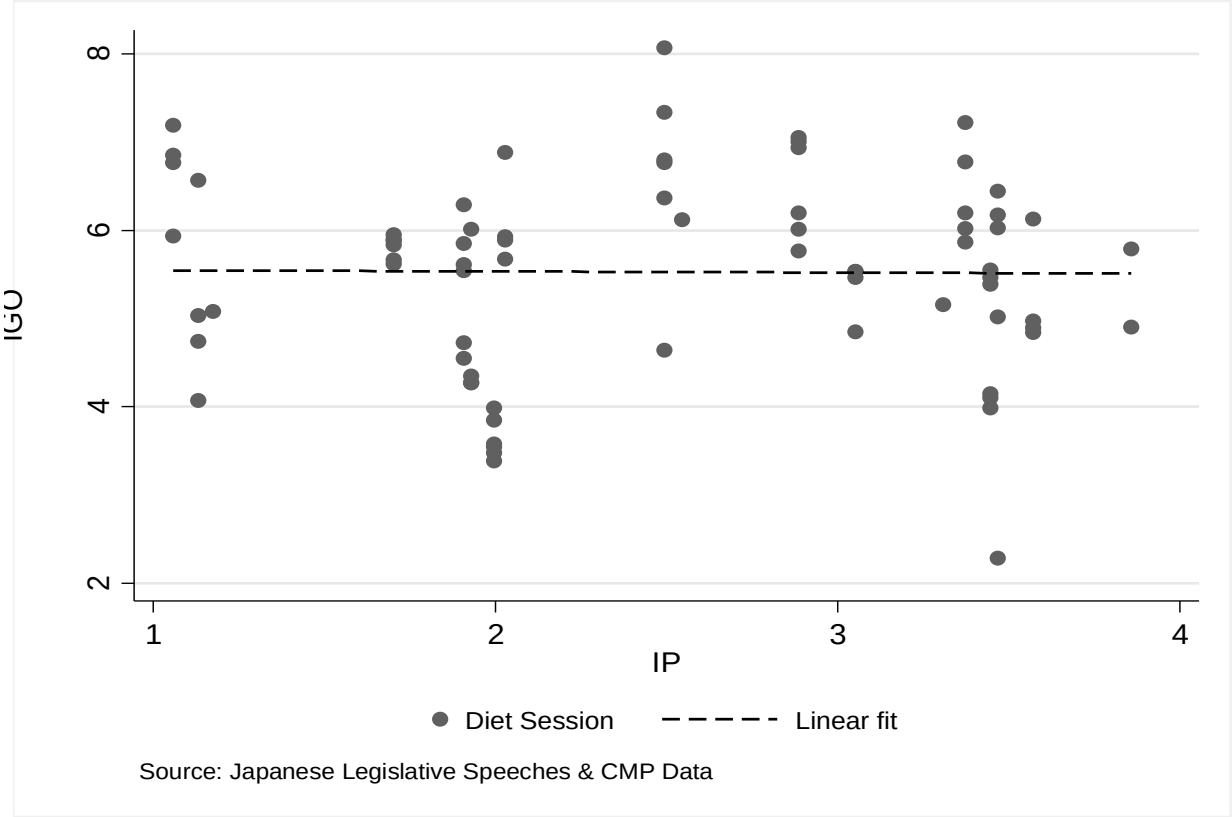
In estimating IGO, we have rescaled P_{jk} on a 0 to 10 scale. In our sample the average value of IGO is 5.6 (standard deviation: 1.07).

IGO index over time



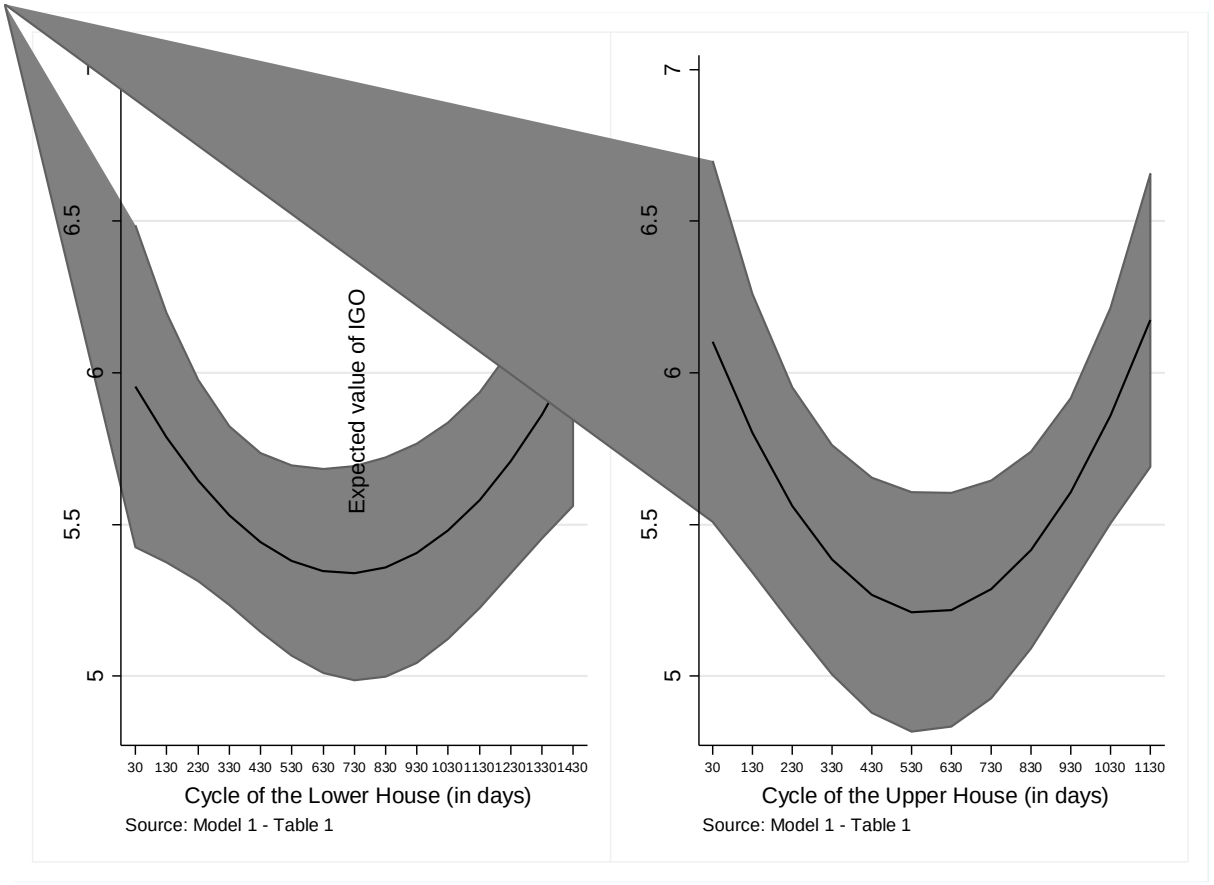
The determinants of the trend in the IGO index

Ideology? Not that much



The determinants of the trend in the IGO index

What seems to matter most: the **change in cabinet format** and…the **electoral cycle**!



And so? Much ado about nothing?

Does this new measure of distance help us solve the empirical puzzles that important legislative phenomena cannot be explained well by party competition **merely based** on ideological confrontations?

The answer is…YES!

Two applications

1. The **survival rate** of Japanese governments (1953-2013)
2. The **duration of cabinet bills** (i.e., time needed for governments to pass their proposed bills in the 1953-1996 period; source: Fukumoto 2000)

In both cases we contrast the results obtained by employing the **IGO** index with a different measure of the **level of complexity** in the bargaining environment in which a cabinet must operate based on a **pure ideological** polarization index using CMP data (1960-2005)

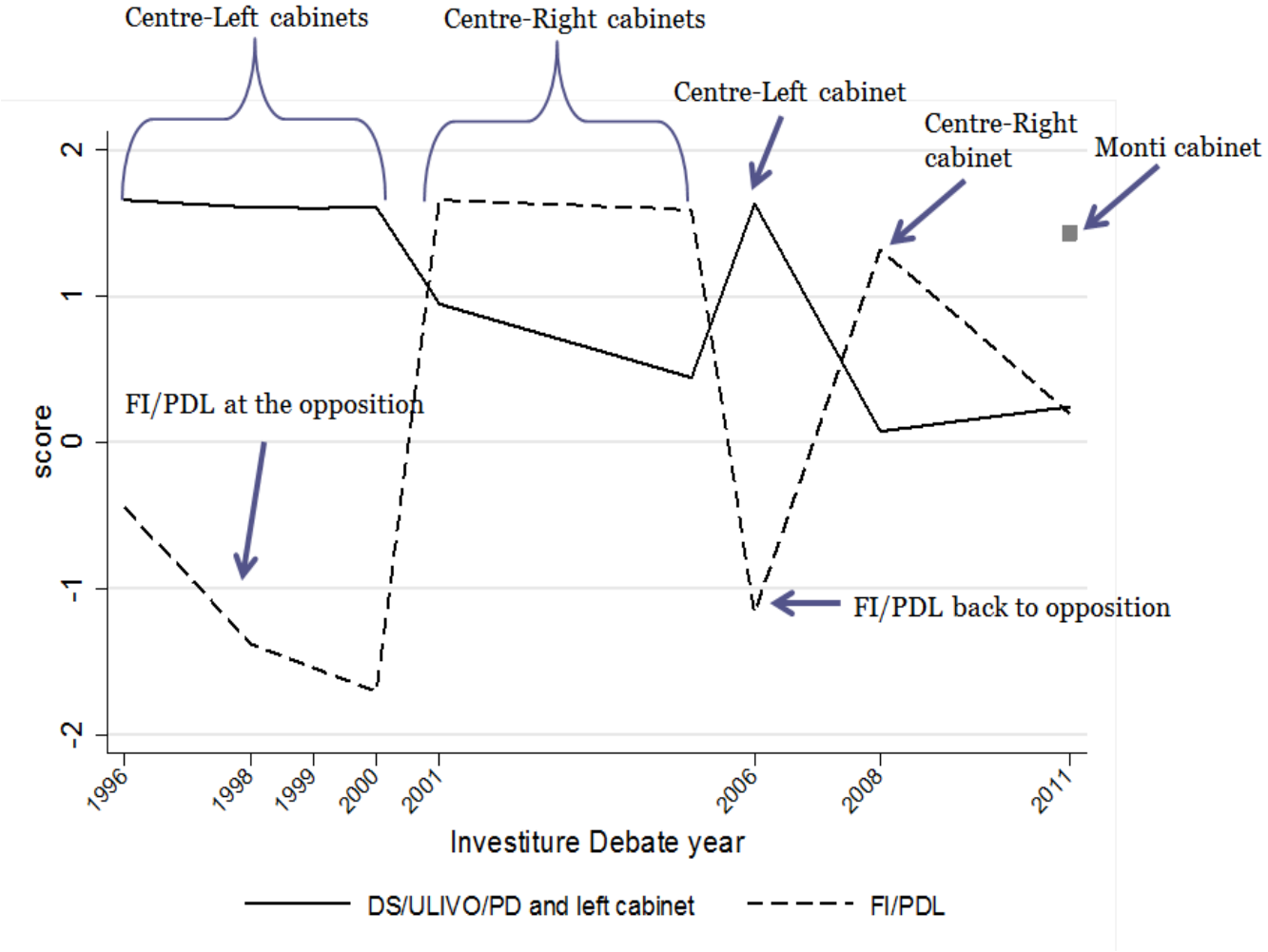
And in both case IGO turns out to be very significant in explaining our dependent variables!

Only in Japan?

We replicated the analysis in the case of the Italian Second Republic (1996-2012)

Different language, different political setting…but same results!

Only in Japan?

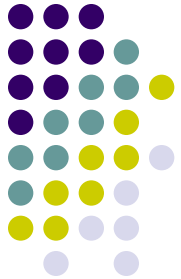


Conclusion: what did we learn?

Studying legislative speeches is **very relevant** and in some instances an irreplaceable opportunity, given that by analyzing them we can capture the position of parties and political actors and their change over time

Still, a researcher should devote an **extra care** about the **substantial content of the positions of political actors** that she gets by analyzing such speeches, especially when she decides to employ any automated scale algorithm to texts

Conclusion: what did we learn?



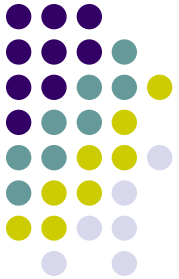
The recovered positions **may contain not only** ideological/policy considerations but also several other aspects that are however important to better define the intensity of the cabinet-opposition divide

In a nutshell, do not apply Wordfish (or any algorithm…) blindly!!! Always **validate** your results!!!

Remember the **fourth Principles** that we have studied in our first lecture!!!

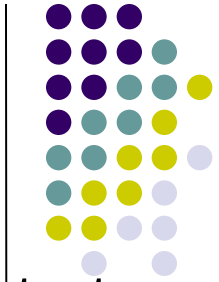
Before today's Lab

1. install.packages("cowplot", repos='http://cran.us.r-project.org')
2. install.packages("magicfor", repos='http://cran.us.r-project.org')
3. devtools::install_github("quanteda/quanteda.corpora")



IMPORTANT!!! (part 1)

Before using rtweet



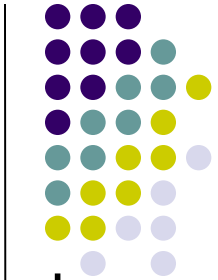
We will use in the next classes the rtweet package: so start to install it!

install.packages("rtweet", repos='http://cran.us.r-project.org')

install.packages("httpuv", repos='http://cran.us.r-project.org')

install.packages("ggmap", repos='http://cran.us.r-project.org')

Before using rtweet



Then open an R session and type the following commands. Plz let me know if you are able (or not) to download the 10 tweets:

```
library(rtweet)

library(httpuv)

rt <- search_tweets( "#rstats", n = 10, include_rts = FALSE)

print(rt$text[1:10])
```

Optional

Before we can start geocoding data, we need to obtain an API key from Google. Go to the registration page, and follow the instructions (select all mapping options)

The **geocoding API** is a free service, but you nevertheless need to associate a credit card with the account.

Please note that the Google Maps API is not a free service. There is a free allowance of 40,000 calls to the geocoding API per month, and beyond that calls are $0.005 each

This implies that basically you have a monthly free limit of $200 (more than enough…)

To register you need to have: a) a gmail account; b) a credit card

Optional

After you finish the registration (if everything hopefully works fine!) Google gives you back an API number. Save it!

Then type:

library(ggmap)

register_google(key = “NUMBER OF YOUR GOOGLE API!")

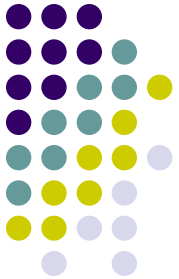
geocode(c("White House", "Uluru"))

You should get this result back:

```
# A tibble: 2 x 2
    lon   lat
  <dbl> <dbl>
1 -77.0  38.9
2 131.  -25.3
```

Optional

If you are able to get the Google API, but GGMAP does not get any results back, enable the “geocoding app” in your console developer. Check how to enable GOOGLE API here



IMPORTANT!!! (part 2)

CMP and R

We will use in the next classes also The ManifestoR package (https://manifesto-project.wzb.eu/information/documents/manifestoR): so start to install it!

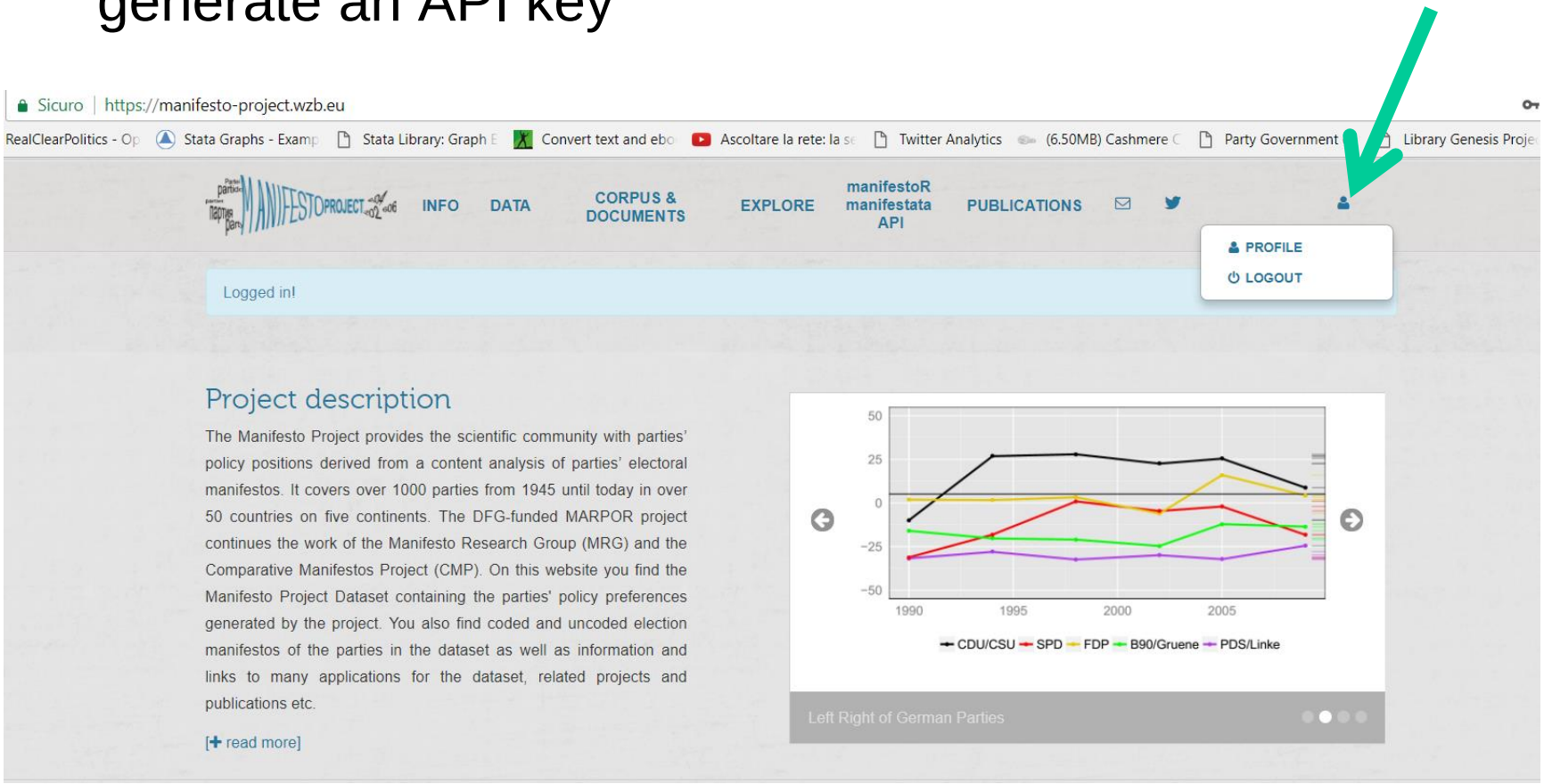
For using such package, however, you also need to have a personal API KEY to get access to the CMP database

How to get it?

Sign up on the Manifesto Project Database webpage to get an account (https://manifesto-project.wzb.eu/signup)

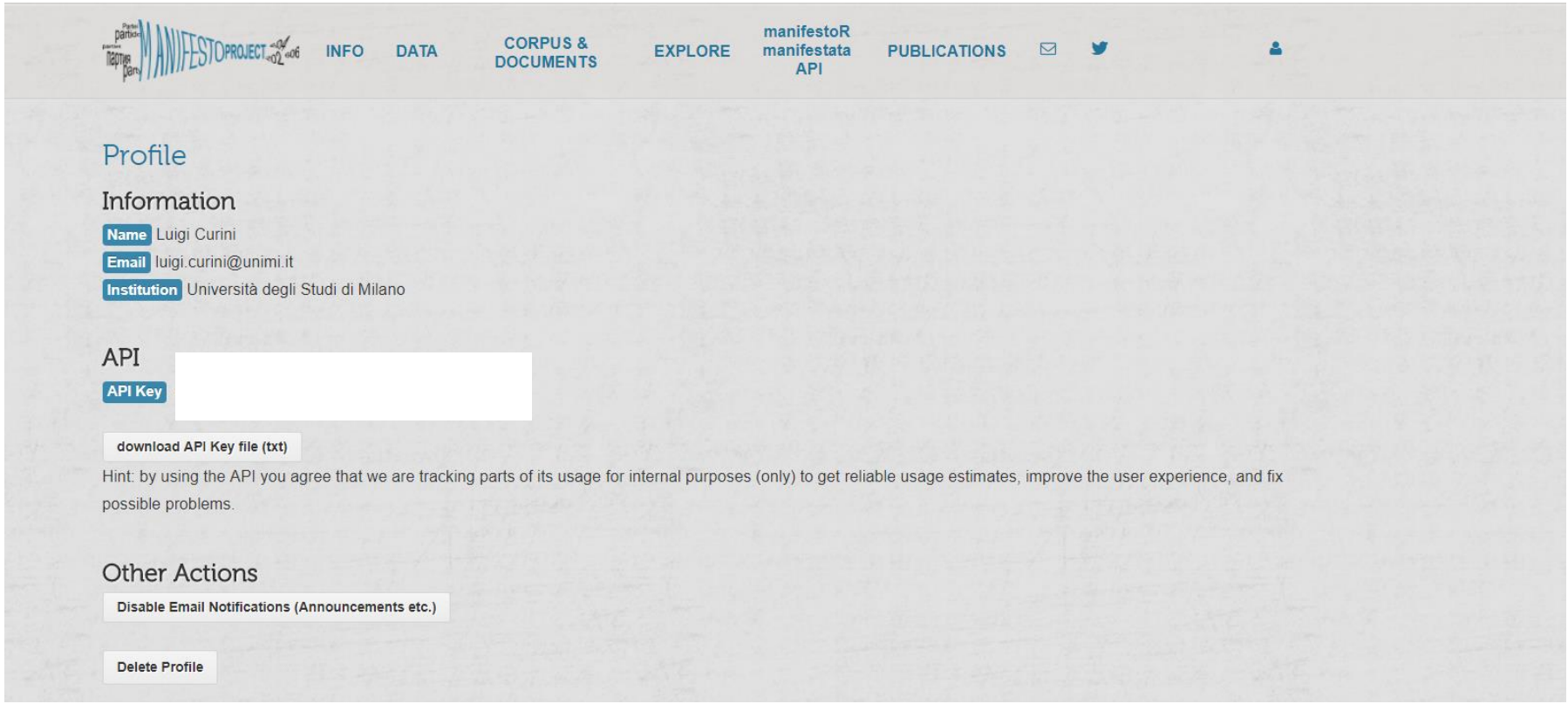
CMP and R

Then login to your account, go to your profile page and generate an API key



CMP and R

Then save such API key by writing down somewhere



CMP and R

Now install the R package «manifestoR»

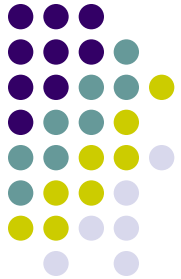
install.packages("manifestoR", repos='http://cran.us.r-project.org')

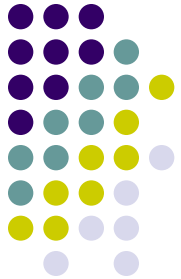
Then type:

library(manifestoR)

mp_setapikey(key.file = NULL, key = “THE API KEY YOU GOT”)

If you get an error, please let me know!





Please check that everything is ok with both rtweet and ManifestoR before 23 of October!!!