SciencesPo

Research Methods in Public Policy

Isabelle Guinaudeau, chargée de recherches CNRS
Benjamin Guinaudeau, Universität Konstanz

Master Science Politique
Sciences Po

Class 4- Text-as-Data Methods for Policy Analysis

4.1 Using text as (quantitative) data : Intro

SciencesPo

- The choice of methods is driven by the research question:

Research question	Most appropriate study design
Does this policy work?	Meta-analysis, experimental designs
What caused outcome Y?	Regression analysis, experimental designs, process-tracing
What did people perceive or think?	Ethnographic work, interviews, focus groups

- Today, we will deal with research questions and designs involving texts

4.1 Using text as (quantitative) data : Intro

SciencesPo

- Text has always been an important data source in political science (in particular policy research)

- Formal and written documents are a key feature of bureaucracies, both public and private (Weber 2015)

- Long tradition of document analysis in the social and political sciences more generally (think of Tocqueville's, Marx's or Moore's analyses of official reports, censuses, newspapers, laws and statues…)

4.1 Using text as (quantitative) data : Intro

SciencesPo

Official documents

- Policies or policy directives
- Official statement and declarations, parliamentary debates or questions
- Official position papers, party manifesto

Legal documents

- Laws
- Regulations, decrees
- Cooperation agreements, international treaties
- Committee reports

Implementation documents

- Midterm or final reports, evaluation reports
- Financial analyses
- Operational plans
- Funding requests

Other documents

- Emails
- Mission reports
- Drafts

Scholarly work

- Scientific or peer-reviewed publications
- Master or doctoral dissertations
- Textbooks or other course materials
- Project reports

Public and media documents

- Twitter / social media content
- Newspaper articles
- Podcasts, video, ratio or TV segments
- Advertisements, posters
- Wikipedia articles

4.1 Using text as (quantitative) data : Intro
(based on Wilkerson & Casas 2017)

SciencesPo

- The internet provides a wealth of data related to politics (public records, newspaper online archives, Gutenberg Project or Google Books, Wikipedia, Twitter or Facebook posts…)

- New methods have emerged to (1) collect and (2) analyze text data, which are labelled with the term “text-as-data”.

- ”Text-as-data methods are a broad set of techniques and approaches relying on the automated or semi-automated analysis of text" (Gilardi & Wüest 2020)

- Text-as-data approaches are becoming mainstream in political science… and this should be increasingly the case also in policy analysis

4.1 Using text as (quantitative) data : Intro

SciencesPo

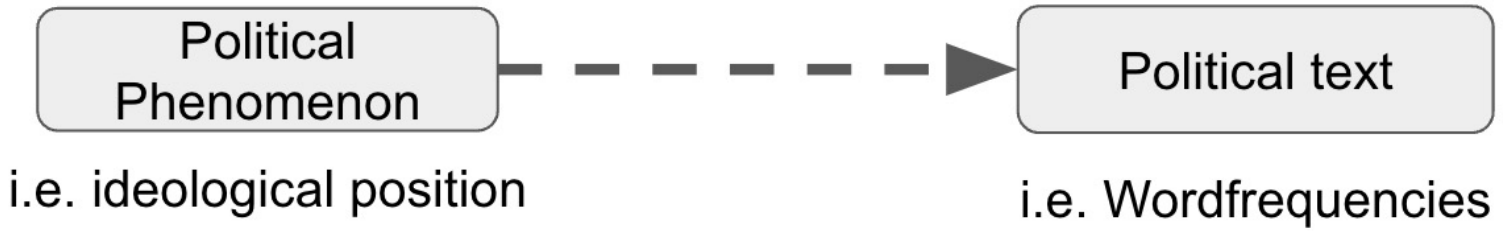
Objectives

- Provide an overview of text-as-data applications
- … and related opportunities in public policy research
- Learn the main steps of text-as-data collection and analysis
- Examine and assess examples of applications

4.1 Using text as (quantitative) data : Intro

SciencesPo

- Daga Generating Process



- Measurement

4.2 How to make text machine-readable?

SciencesPo

- Computer can only manipulate numbers

128	Ç	144	É	160	á	176	▒	193	┴	209	╤	225	ß	241	±
129	ü	145	æ	161	í	177	▓	194	┬	210	╥	226	Γ	242	≥
130	é	146	Æ	162	ó	178	▓	195	├	211	╙	227	π	243	≤
131	â	147	ô	163	ú	179	|	196	─	212	╘	228	Σ	244	⌠
132	ä	148	ö	164	ñ	180	┤	197	┼	213	╒	229	σ	245	⌡
133	à	149	ò	165	Ñ	181	╡	198	╞	214	╓	230	µ	246	÷
134	å	150	û	166	ª	182	╢	199	╟	215	╫	231	τ	247	≈
135	ç	151	ù	167	º	183	╖	200	╚	216	╪	232	Φ	248	°
136	ê	152	ÿ	168	¿	184	╕	201	╔	217	┘	233	Θ	249	·
137	ë	153	Ö	169	⌐	185	╣	202	╩	218	┌	234	Ω	250	·
138	è	154	Ü	170	¬	186	║	203	╦	219	█	235	δ	251	√
139	ï	156	£	171	½	187	╗	204	╠	220	▄	236	∞	252	_
140	î	157	¥	172	¼	188	╝	205	═	221	▌	237	φ	253	²
141	ì	158	₧	173	¡	189	╜	206	╬	222	▐	238	ε	254	■
142	Ä	159	ƒ	174	«	190	╛	207	╧	223	▀	239	∩	255	
143	Å	192	└	175	»	191	┐	208	╨	224	α	240	≡		

4.2 How to make text machine-readable?

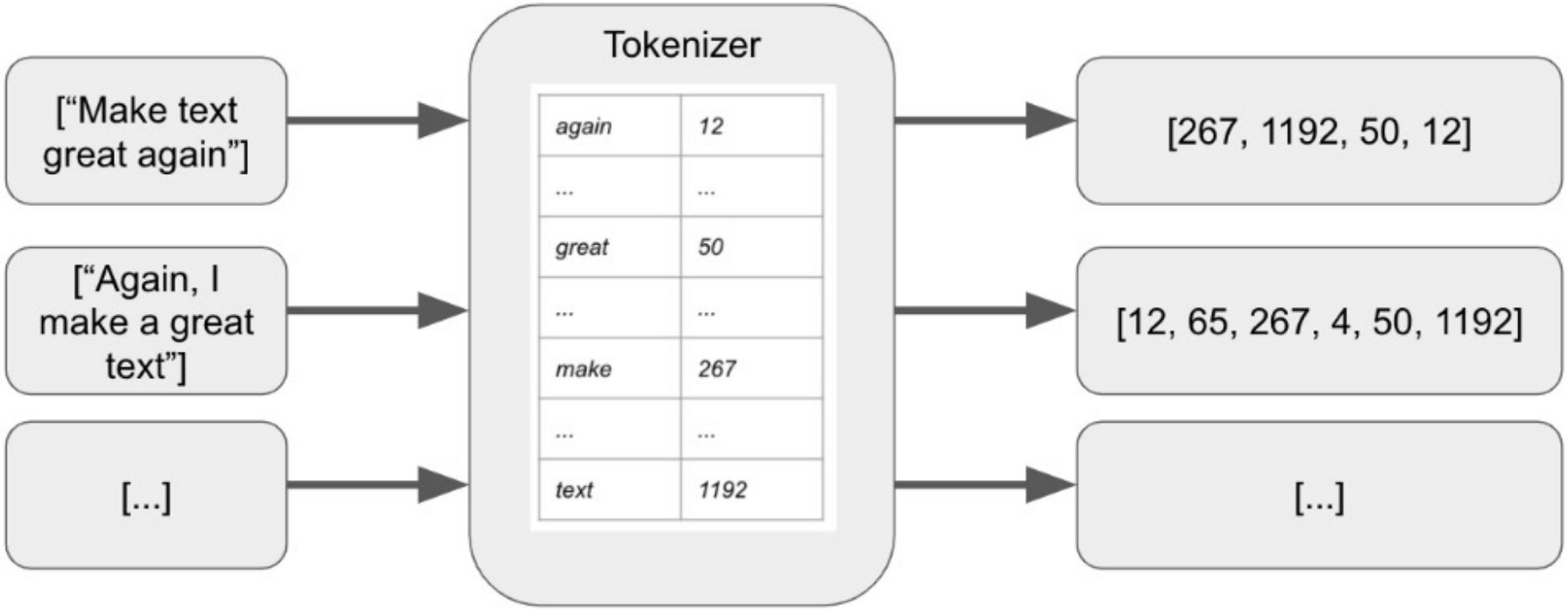
SciencesPo

- Potential numeric representation of texts:
  - Sequence of characters
  - Bag-of-Words (aka DTM or Word frequencies)
  - Sequence of words
  - Semantic vectors (embedding)

→ No representation is right. It has to match the task and method of analysis.

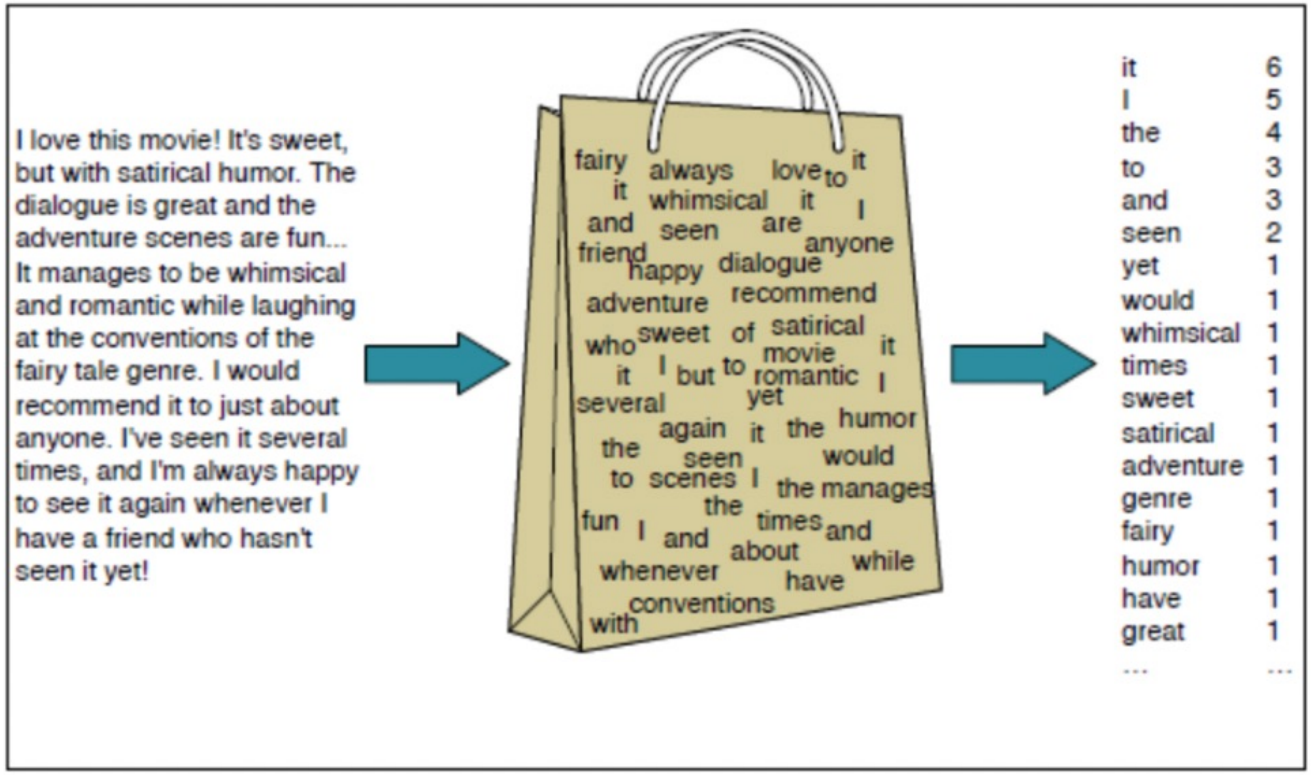
4.2 Text representation – Sequence of Words

SciencesPo



4.2 Text representation – Bag of Words

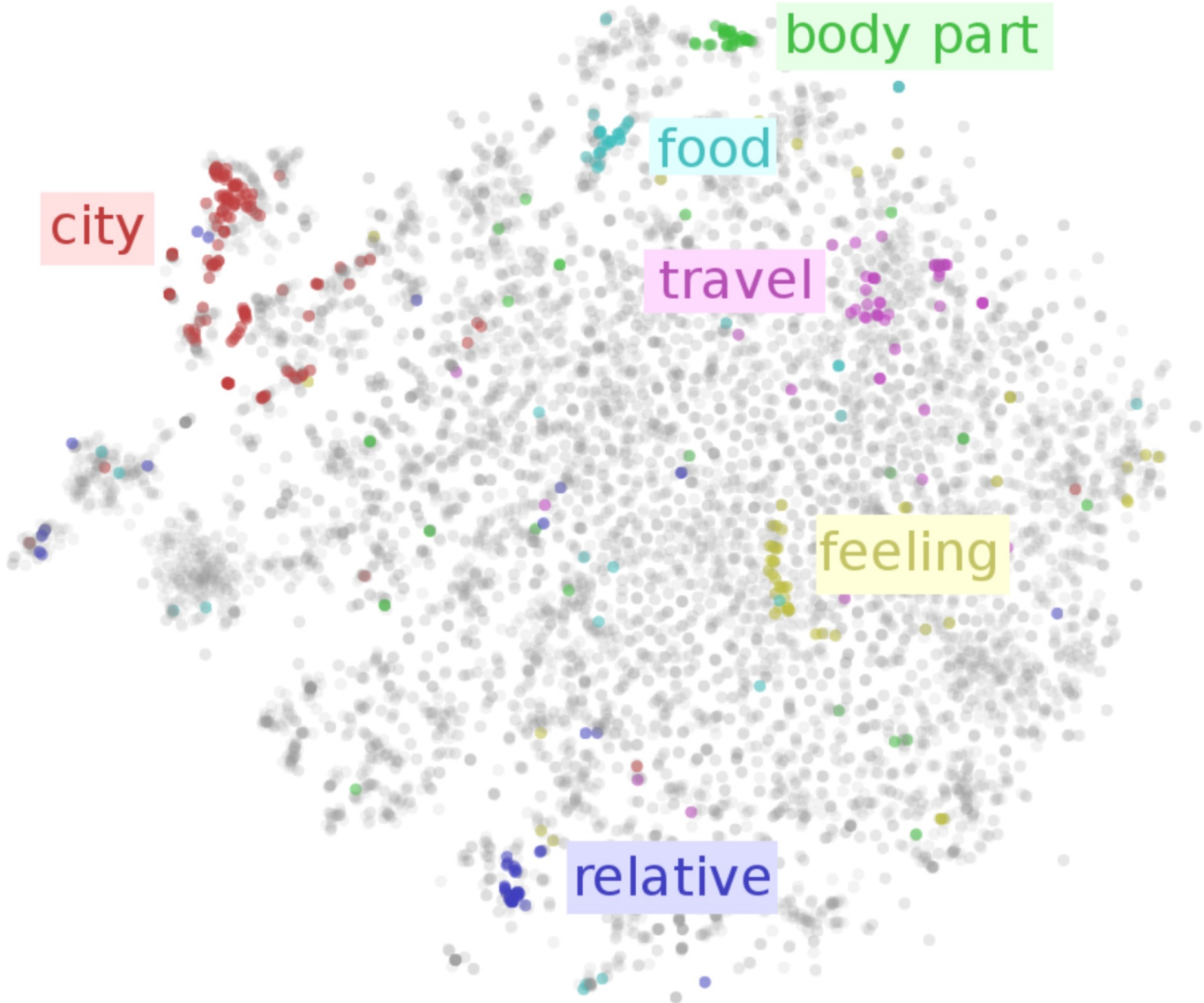
SciencesPo



source: Jurafsky et al., 2018

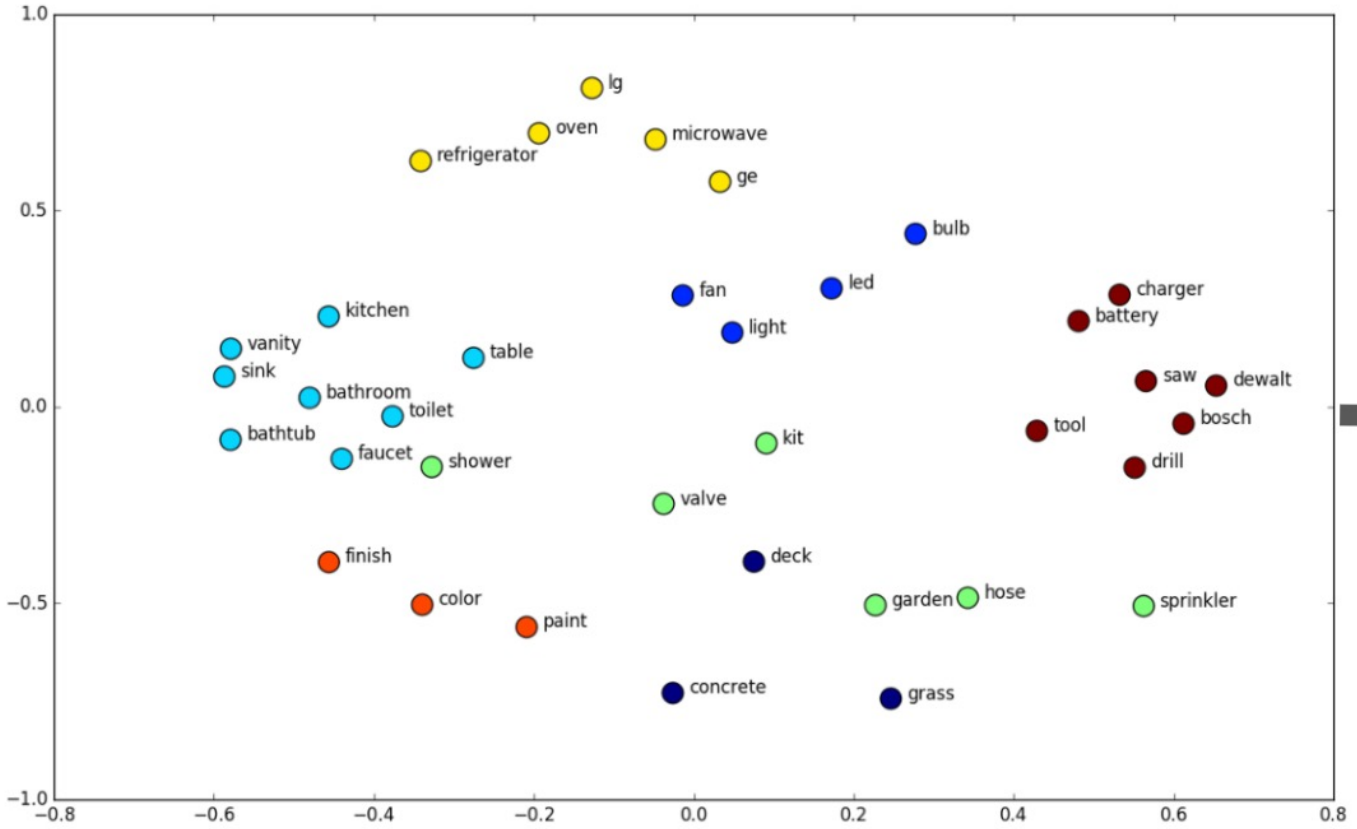
4.2 Text representation – Embeddings

SciencesPo



4.2 Text representation – Embeddings

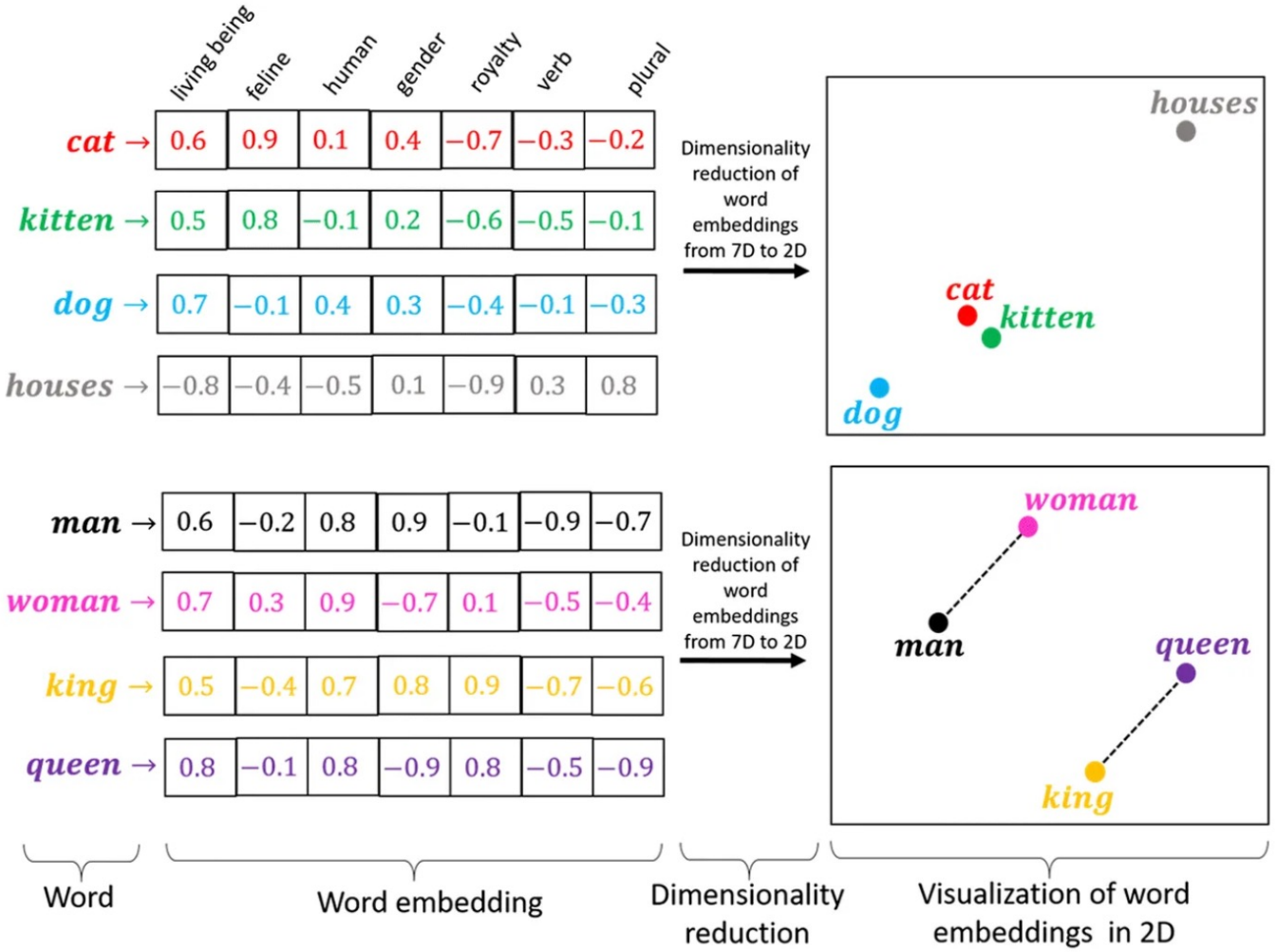
SciencesPo



Wort <chr>	Dim 1 <dbl>	Dim 2 <dbl>
refrigerator	−0.37	0.65
oven	−0.21	0.70
bulb	0.23	0.44
bathroom	−0.45	−0.20
kitchen	−0.44	0.23
drill	0.59	−0.10
...	0.00	0.00

4.2 Text representation – Embeddings

SciencesPo



4.3 Measurement models

SciencesPo

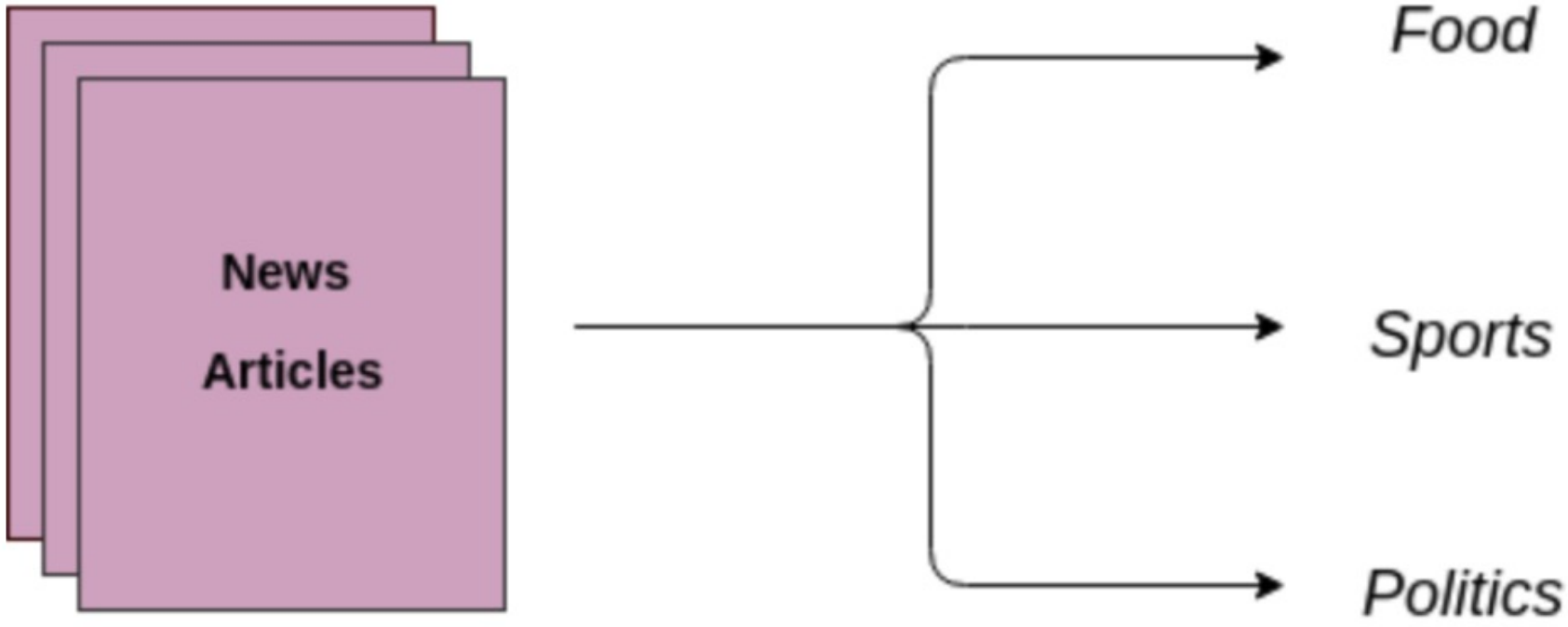
There are three main strategies to use text in social science research:

1. Supervised classification
2. Unsupervised classification – Topic Model
3. Text-Scaling – Latent variable model

The list is non-exhaustive: Text-Reuse ; NLP

4.3 Supervised classification

SciencesPo



4.3 Supervised classification

SciencesPo

Objective: train a model able to label unseen documents

Research question: What proportion of the German oral parliamentary questions regards international affairs?

```
Rows: 3,119
Columns: 6
$ lp       <chr> "18", "18", "18", "18", "18", "18", "18", "18", "18", "18", "18", "18", "18", …
$ mp_party <chr> "BÜNDNIS 90/DIE GRÜNEN", "DIE LINKE", "DIE LINKE", "BÜNDNIS 90/DIE GRÜNEN", "B…
$ ministry <chr> "Bundeskanzleramt", "Bundesministerium für Wirtschaft und Technologie", "Bunde…
$ titel    <chr> "Einfluss des ehemaligen Staatsministers Eckart von Klaeden auf Entscheidungen…
$ text     <chr> "Inwieweit hat Staatsminister a. D. Eckart von Klaeden in seiner Amtszeit Einf…
$ date     <chr> "25.11.2013", "25.11.2013", "25.11.2013", "25.11.2013", "25.11.2013", "25.11.2…
```

datum	titel	MdB	inhalt	international
25.11.201	Perspektive für die Östlichen Partnerschaften mit der EU	Cornelia Pieper	Wie sieht die Bundesregierung vor dem Vilnius-Gipfel die Perspektive für die Östliche Partnerschaft angesichts der Tatsache, dass die Ukraine die Vorbereitung zur Unterzeichnung des Assoziierungsabkommens mit der EU	?
16.05.201	EU-Mission in der Ukraine im Rahmen der "GSVP" (Gemeinsame Sicherheits- und Verteidigungspolitik)	Michael Roth	Welchen personellen und finanziellen Beitrag (inklusive Ausstattungshilfe) beabsichtigt die Bundesregierung zu der EU-GSVP-Mission in der Ukraine zu leisten, die nach dem Ratsbeschluss vom 12. Mai 2014 derzeit vom	?
16.09.201	Meldefrist für die Kennzeichnung von Kälbern mit Ohrmarken	Peter Bleser	Sieht die Bundesregierung eine Möglichkeit, die Sieben-Tage-Frist für die Kennzeichnung von Kälbern mit Ohrmarken, die sich aus der Umsetzung der Verordnung (EG) Nr. 1760/2000 (vorher EU-Verordnung 820/97) des	?
10.02.201	Ausländische Straftäter in deutschen Strafvollzugseinrichtungen in den Jahren 2015 und 2016	Christian Lange	Welche Kenntnis hat die Bundesregierung zu Zahlen, genauer zum Anteil ausländischer Straftäter in deutschen Strafvollzugseinrichtungen jeweils in den Jahren 2015 und 2016?	?

4.3 Supervised classification

SciencesPo

Objective: train a model able to label unseen documents

Research question: What proportion of the German oral parliamentary questions regards international affairs?

Steps:

1. Obtain a pre-labeled/Manually label a dataset
2. Train a model, able to associate text patterns with labels
3. Evaluate the model
4. Use the model to label unseen documents

4.3 Supervised classification: Pre-labeled data

SciencesPo

Example of pre-labeled dataset:

- Binary classification: international vs. domestic affairs
- 24k German parliamentary questions

```
Rows: 24,105
Columns: 7
$ lp            <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ drucksache    <chr> "01/3493", "01/3858", "01/3859", "01/3864", "01/4205", "01/4305", "01/44…
$ anfragedatum  <date> 1952-06-24, 1952-11-17, 1952-11-18, 1952-11-18, 1953-03-20, 1953-04-30,…
$ titel         <chr> "Einfuhr- und Vorratsstellen", "Bekanntgabe der Note der Bundesregierung…
$ anfragesteller <chr> "Fraktion der FDP", "Fraktion der SPD", "Fraktion der SPD", "Fraktion de…
$ inhalt        <chr> "Einfuhr- und Vorratsstellen. 1. Was gedenkt die Bundesregierung zu tun,…
$ international <dbl> 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, …
```

datum	titel	fraktion	inhalt	international
14.02.52	Schutz deutscher Interessen im Ausland	FU	Schutz deutscher Interessen im Ausland. Wir fragen die Bundesregierung: 1. Ist der Bundesregierung bekannt, daß bei den Ereignissen, die sich am 26. Januar 1952 in Ägypten, insbesondere in Kairo, abspielten, auch das Eigentum	1
13.09.12	Pflege-Transparenzvereinbarung (so genannter Pflege-TueV)	Grünen	Pruefung stationaerer Altenpflegeeinrichtungen sowie ambulanter Pflegedienste gemaess den Pflege-Transparenzvereinbarungen stationaer (PTVS) bzw. ambulant (PTVA), Ergebnisse, Pruefpersonal der Medizinischen	0
27.01.16	Änderungen der bisherigen Rüstungsexportpolitik und ihrer gesetzlichen Grundlagen	Linke	Abgegebene Verpflichtungserklärungen zur Anwendung der Kleinwaffengrundsätze "Neu für Alt" bzw. "Neu, Vernichtung bei Aussonderung" bei Genehmigungsentscheidungen für Brasilien, Hongkong,	1
27.01.16	Neue Erkenntnisse und Pläne der Bundesregierung zum Einsatz der Fracking-Technik in Deutschland	Grünen	Hinterfragung der weiteren Unterstützung des Gesetzespakets zur Fracking-Regulierung (BT-Drs 18/4713 und 18/4714) vor dem Hintergrund der Beschlüsse der VN-Klimakonferenz in Paris, Bericht der Bundesregierung zum	0

4.3 Supervised classification: Train a model

SciencesPo

Train-test split to avoid overfitting:

- Fit a model (i.e. logistic regression) on 80% of the data (train)
- Use the remaining 20% (test) to evaluate the out-of-sample performance

$$Pr(y^{train} = 1) = f(text^{train})$$

4.3 Supervised classification: Evaluate the model

SciencesPo

Evalute the model: confusion matrix based on test dataset

- Train differenct models (LR, Random Forrest, SVM, etc…) with different parameters (L1/L2 Regularization, etc…)
- Choose the model with the best out-of-sample performance

	Predicted: 0	Predicted: 1
Real: 0	Correct (True negative)	Error (False Positive)
Real: 1	Error (False Negative)	Correct (True positive)

4.3 Supervised classification: Evaluate the model

SciencesPo

Evalute the model: confusion matrix based on test dataset

	Predicted: 0	Predicted: 1
Real: 0	100	25
Real: 1	75	100

4.3 Supervised classification: Evaluate the model

SciencesPo

Evalute the model: confusion matrix based on test dataset

	Predicted: 0	Predicted: 1
Real: 0	100	25
Real: 1	75	100

Precision:
100/125 = .8

4.3 Supervised classification: Evaluate the model

SciencesPo

Evalute the model: confusion matrix based on test dataset

	Predicted: 0	Predicted: 1
Real: 0	100	25
Real: 1	75	100

Recall: 100/175 = .57

4.3 Supervised classification: Evaluate the model

SciencesPo

Evalute the model: confusion matrix based on test dataset

	Predicted: 0	Predicted: 1
Real: 0	100	25
Real: 1	75	100

Recall: 100/175 = .57

4.3 Supervised classification: Label new data

SciencesPo

What proportion of the oral parliamentary questions concerns international affairs?

```
Rows: 3,119
Columns: 6
$ lp       <chr> "18", "18", "18", "18", "18", "18", "18", "18", "18", "18", "18", "18", "18", …
$ mp_party <chr> "BÜNDNIS 90/DIE GRÜNEN", "DIE LINKE", "DIE LINKE", "BÜNDNIS 90/DIE GRÜNEN", "B…
$ ministry <chr> "Bundeskanzleramt", "Bundesministerium für Wirtschaft und Technologie", "Bunde…
$ titel    <chr> "Einfluss des ehemaligen Staatsministers Eckart von Klaeden auf Entscheidungen…
$ text     <chr> "Inwieweit hat Staatsminister a. D. Eckart von Klaeden in seiner Amtszeit Einf…
$ date     <chr> "25.11.2013", "25.11.2013", "25.11.2013", "25.11.2013", "25.11.2013", "25.11.2…
```

datum	titel	MdB	inhalt	international
25.11.201	Perspektive für die Östlichen Partnerschaften mit der EU	Cornelia Pieper	Wie sieht die Bundesregierung vor dem Vilnius-Gipfel die Perspektive für die Östliche Partnerschaft angesichts der Tatsache, dass die Ukraine die Vorbereitung zur Unterzeichnung des Assoziierungsabkommens mit der EU	?
16.05.201	EU-Mission in der Ukraine im Rahmen der "GSVP" (Gemeinsame Sicherheits- und Verteidigungspolitik)	Michael Roth	Welchen personellen und finanziellen Beitrag (inklusive Ausstattungshilfe) beabsichtigt die Bundesregierung zu der EU-GSVP-Mission in der Ukraine zu leisten, die nach dem Ratsbeschluss vom 12. Mai 2014 derzeit vom	?
16.09.201	Meldefrist für die Kennzeichnung von Kälbern mit Ohrmarken	Peter Bleser	Sieht die Bundesregierung eine Möglichkeit, die Sieben-Tage-Frist für die Kennzeichnung von Kälbern mit Ohrmarken, die sich aus der Umsetzung der Verordnung (EG) Nr. 1760/2000 (vorher EU-Verordnung 820/97) des	?
10.02.201	Ausländische Straftäter in deutschen Strafvollzugseinrichtungen in den Jahren 2015 und 2016	Christian Lange	Welche Kenntnis hat die Bundesregierung zu Zahlen, genauer zum Anteil ausländischer Straftäter in deutschen Strafvollzugseinrichtungen jeweils in den Jahren 2015 und 2016?	?

4.3 Unsupervised classification/Topic Model

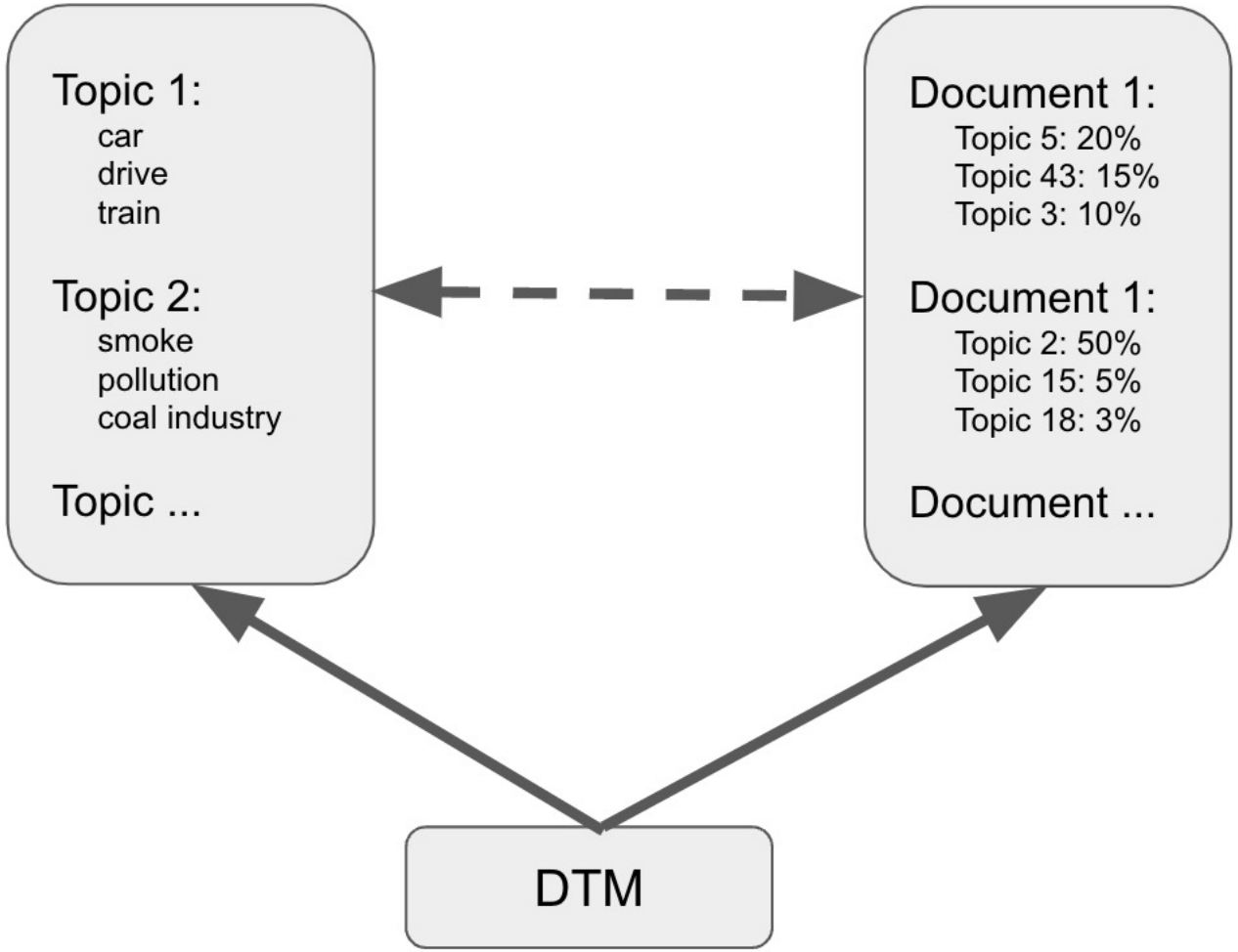
SciencesPo

Objectives:

1. Identify groups of words, which usually appear together (topics)
2. Estimate the proportion of each document relating to a given topic

4.3 Unsupervised classification/Topic Model

SciencesPo



4.3 Unsupervised classification/Topic Model

SciencesPo

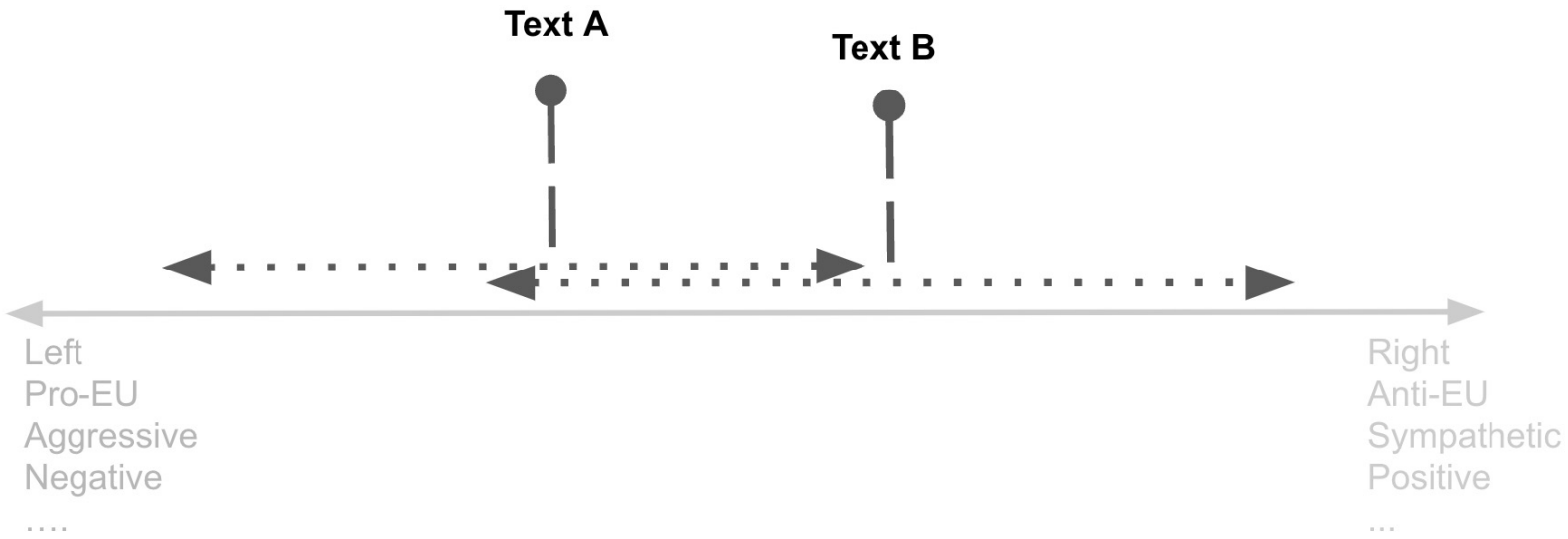
2 steps:

1. Fit the model on an unlabeled corpus
2. Interpret/validate the topics
   - Interpretation is qualitative
   - No clear validation framework

4.3 Text scaling

SciencesPo

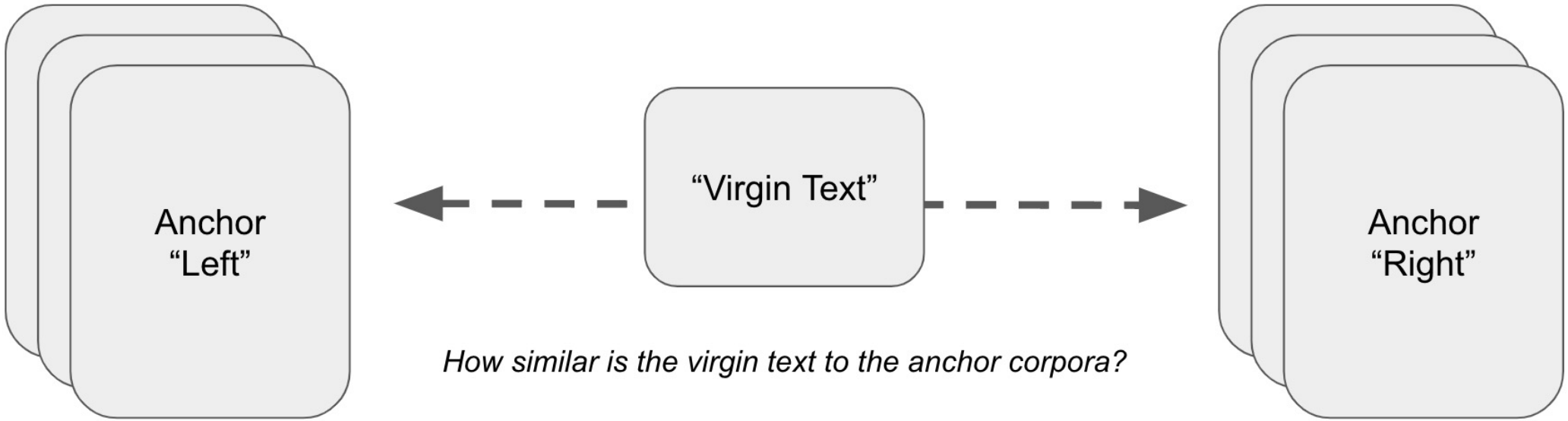
Objective: Estimate the spatial position of texts/actors on a latent dimension



4.3 Text scaling – Wordscore (Benoit and Laver 2003)

SciencesPo

Supervised scaling: Similarity with anchors

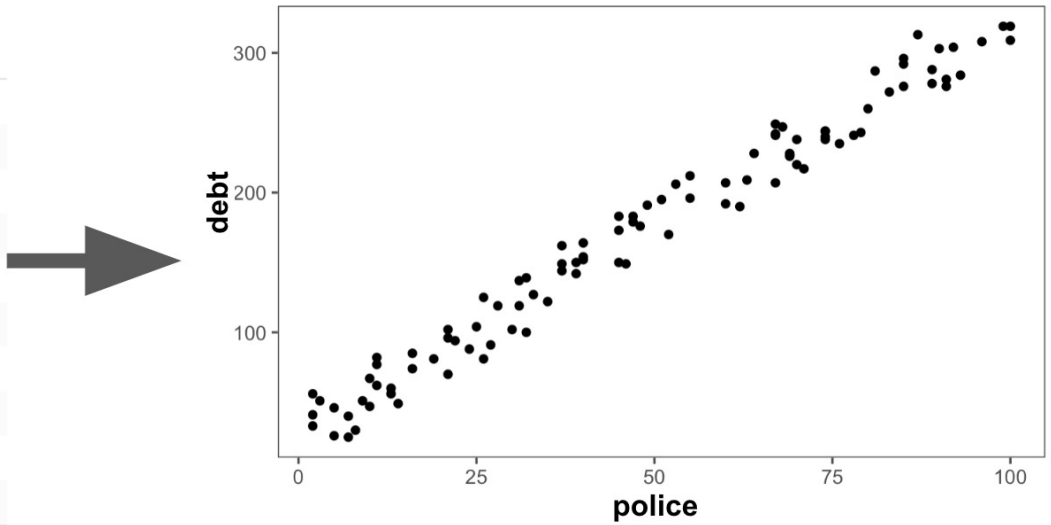


4.3 Text scaling – Wordfish (Proksch and Slapin 2008)

SciencesPo

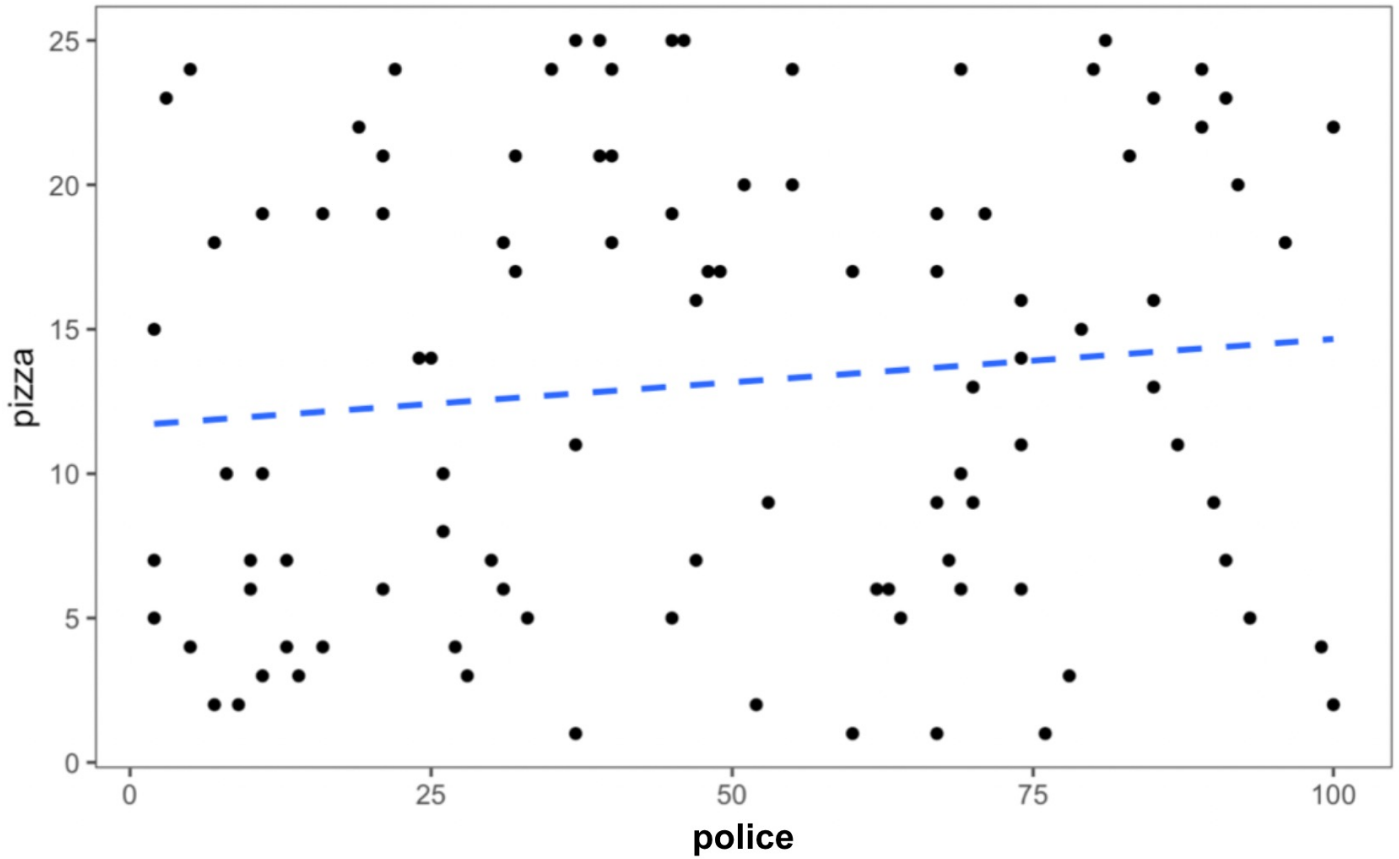
Unsupervised scaling: Captures the most-salient dimension in the text

doc <int>	police <int>	debt <dbl>
1	46	149
2	13	60
3	7	25
4	32	139
5	35	122
6	45	150
7	74	244
8	2	56
9	51	195
10	22	94



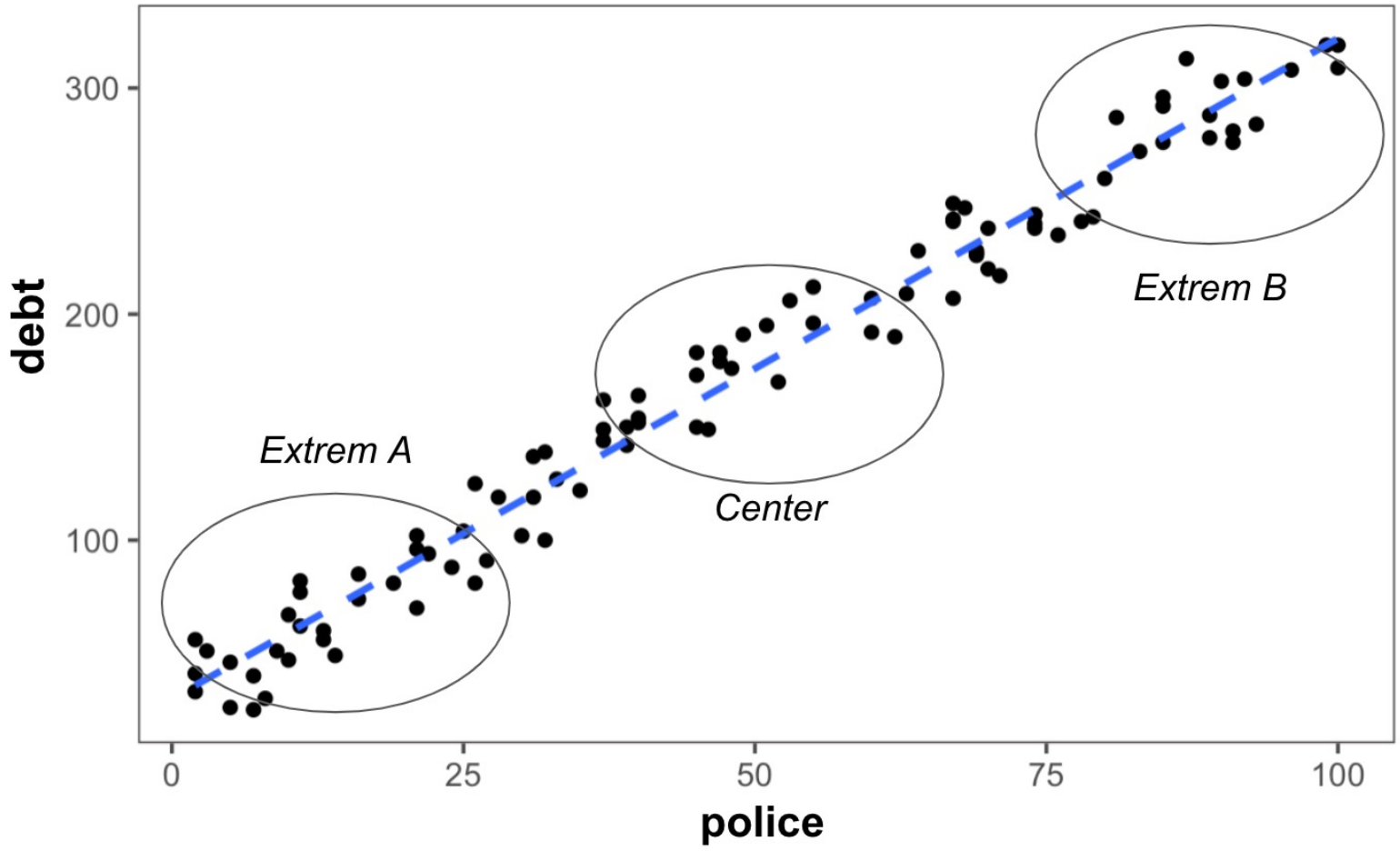
4.3 Text scaling – Wordfish (Proksch and Slapin 2008)

SciencesPo



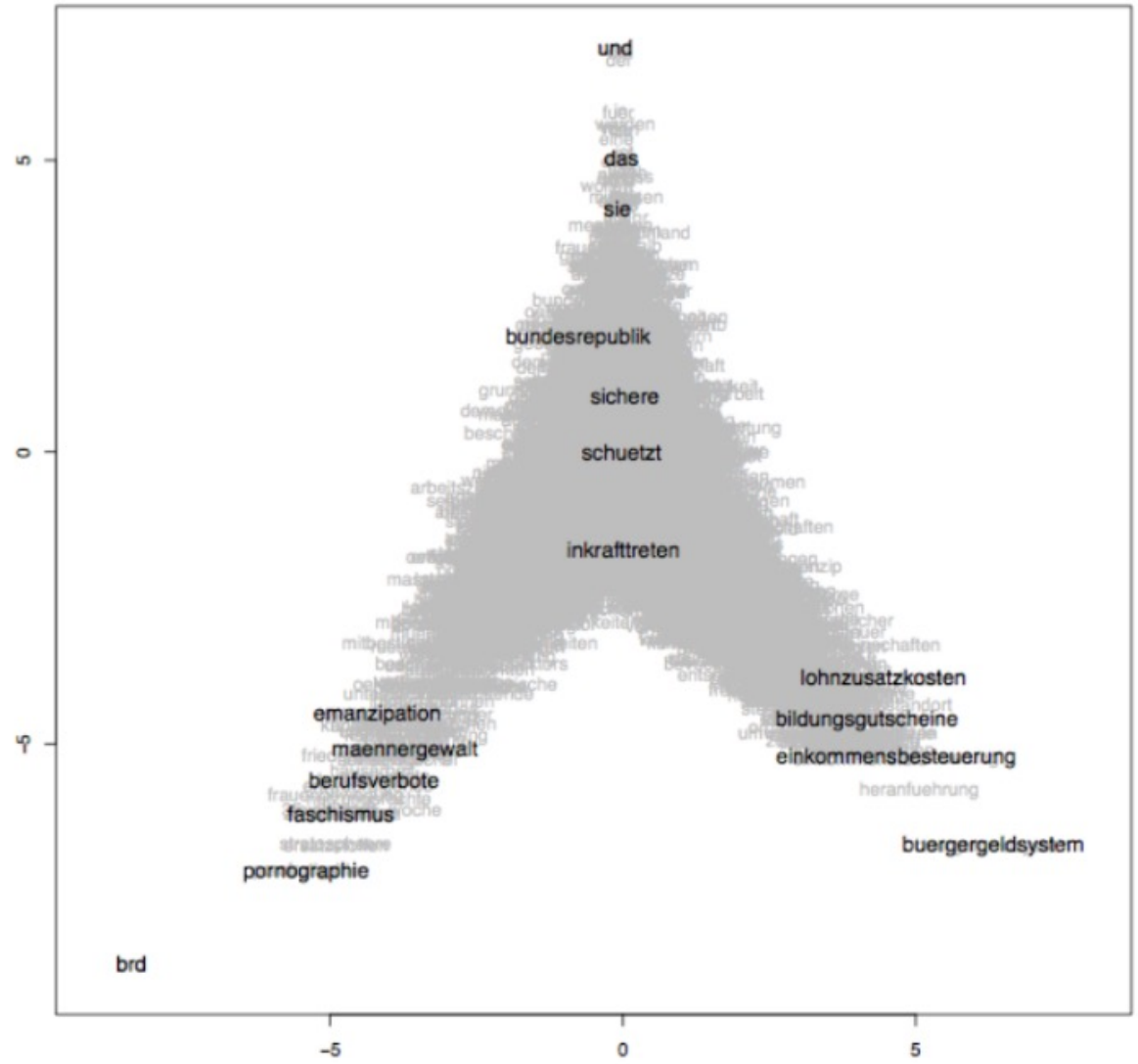
4.3 Text scaling – Wordfish (Proksch and Slapin 2008)

SciencesPo



4.3 Text scaling – Wordfish (Proksch and Slapin 2008)

SciencesPo



4.3 Text scaling – Key questions

SciencesPo

Wordscores (supervised):

- How to choose well-suited anchors?

Wordfish (unsupervised):

- How to validate the discovered dimension?

4.3 Beyond the three main strategies

SciencesPo

Classification (supervised and unsupervised) and scaling amounts to the vast majority of text-based measures.

Nonetheless, other strategies exist:

1. Text Reuse
2. Entity recognition
3. POS tagging (gramatical natur and function)
4. ….

4.4 Presentations

SciencesPo

4.4 Wilkerson et al. 2015

SciencesPo

- Theoretical argument: laws originate to large extent in failed bills (some of which put sponsored by opposition MPs): the progress of policy ideas is more interesting than the progress of single bills.

- New approach: “text reuse” approach based on computer science to compare the substance of law sections to sections of bills

4.4 Wilkerson et al. 2015

SciencesPo

Alternative approaches?

	Advantages	Drawbacks
Expert (manual) coding	Based on fine-grained contextual expertise and knowledge	Resource- and cost-intensive Unreliable (CRS not reliable) Level of laws instead of law sections
Computer-assisted (text-reuse and machine-learning) approach	Ability to process big data Reliable and replicable	Potential validity and causality concerns

4.4 Wilkerson et al. 2015

SciencesPo

Approach

- Identify and format the corpus of data (text of all bills introduced in the 111th Congress)
- Delimitate unit of analysis and standardize information (bill sections without summaries, titles, etc.)
- Design (semi-)automated procedure depending on the goal: concept identification, classification, scaling or discovery of categories… (text-reuse for identification, machine learning approach accounting for word embedding for classifying cases)
- Assess validity
  - Recall tests: fraction of relevant instances that were retrieved (how many false negative?)
  - Precision tests: fraction of relevant instances among the retrieved instances (how many false positive?)

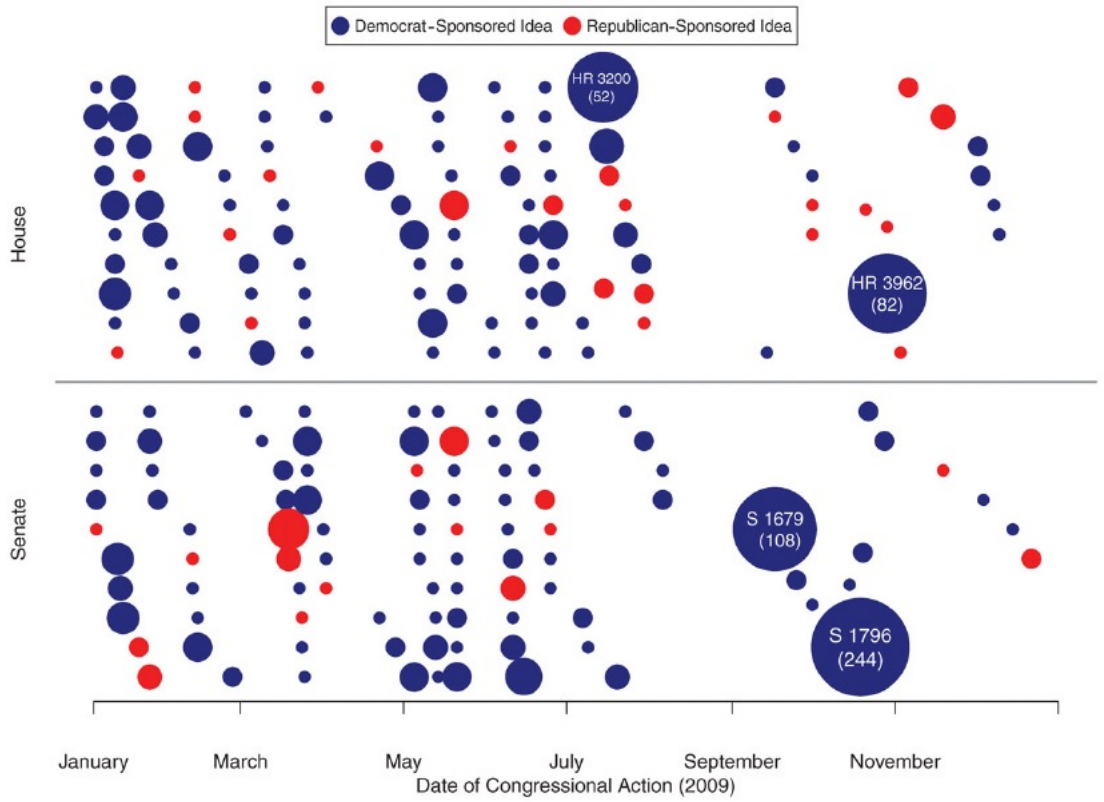
4.4 Wilkerson et al. 2015

SciencesPo

Data visualization

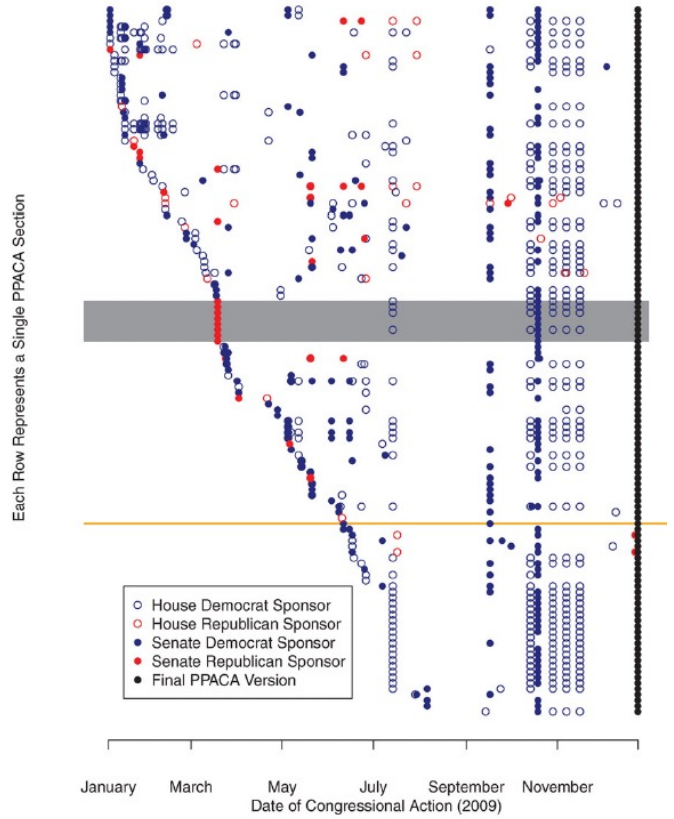
(1) First graph shows rather which bill had most influence on the Obamacare law

FIGURE 3 Bills Sharing Policy Ideas with the PPACA (by Date of Introduction)



(2) 2nd graph focuses on *ideas* and shows which ones can be traced furthest back in time

FIGURE 4 Sections of Other Bills Sharing Policy Ideas with PPACA Sections



4.4 Wilkerson et al. 2015

SciencesPo

- Causality concerns: the authors claim that earlier bills influence later laws. Are there alternative chains of causality?

4.4 Wilkerson et al. 2015

SciencesPo

- Causality concerns: the authors claim that earlier bills influence later laws. Are there alternative chains of causality?
  - For instance, it is possible that a same interest group pressures different MPs to put forward a same bill.
  - Or that MPs respond similarly to a “hot” issue.

4.4 Wilkerson et al. 2015

SciencesPo

- Which other questions could be tackled with the text-reuse or classification approaches used in the article?

4.4 Wilkerson et al. 2015

SciencesPo

- Which other questions could be tackled with the text-reuse or classification approaches used in the article?
  - Text-reuse: impact of interest group positions papers? Are there sectors where legislators take up more ideas from the opposition?
  - Classification: what drives attention to topic x in parliamentary questions?

4.5 Bräuninger et al. 2019

SciencesPo

- Theoretical argument: parties seek to form coalitions with ideologically close parties

- New approach: scaling applied to party manifesto to compare their positions along several relevant dimensions

4.5 Bräuninger et al. 2019

SciencesPo

Alternative approaches?

	Advantages	Drawbacks
Expert (manual) coding	Based on fine-grained contextual expertise and knowledge	Resource- and cost-intensive Not necessarily reliable
Computer-assisted (text as data) approach based on scaling departing from reference documents	Ability to process big data Reliable and replicable	Potential validity concerns

4.5 Bräuninger et al. 2019

SciencesPo

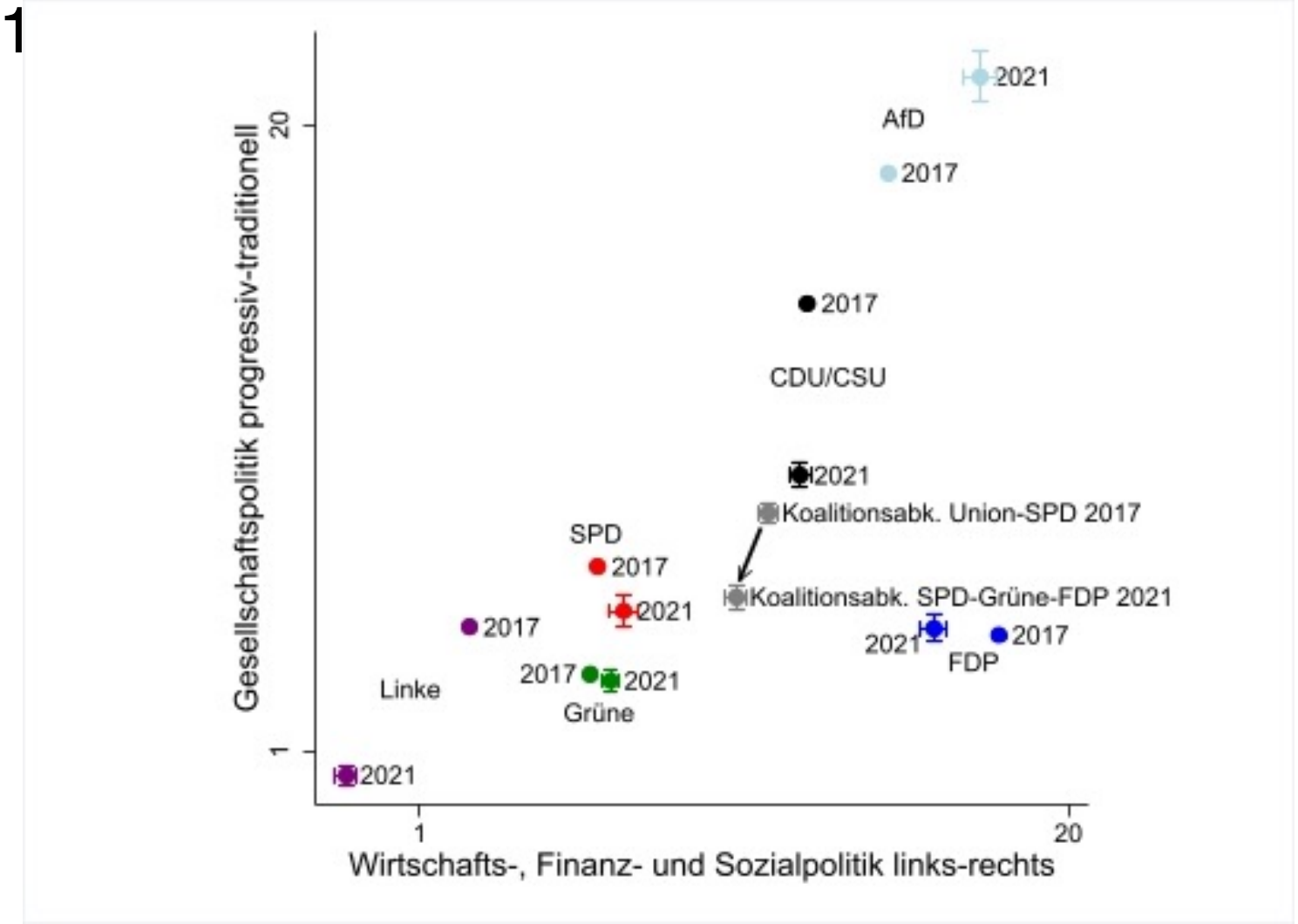
Approach

- Identify and format the corpus of data (party manifestos, coalition negotiation papers, coalition agreements)
- Delimitate unit of analysis and standardize information (full document)
- Design (semi-)automated procedure for measuring ideological distance (scaling)
- Assess validity
  - Face validity?

4.5 Bräuninger et al. 2019

SciencesPo

Update for 2021



Source: https://twitter.com/DebusMarc/status/1463520319201812484

4.5 Bräuninger et al. 2019

SciencesPo

- Which other questions could be tackled with the text-reuse or classification approaches used in the article?

4.5 Bräuninger et al. 2019

SciencesPo

- Which other questions could be tackled with the text-reuse or classification approaches used in the article?
  - Predict outcome of legislative bargaining (between institutions, coalition partners or within parties)
  - Predict policy outcomes depending on the position of the coalition agreement (or of single governments parties)

4.6 Current challenges

SciencesPo

1. Accessing data
   - Public API/Scraping against TOS
   - Twitter's policy shift
2. Reproducibility
   - Validation framework for topic model
   - Transparency about modelling decisions
3. Larger models require more computational power
   - LLM and transformers
4. Model biases
   - Model reproduce human biases