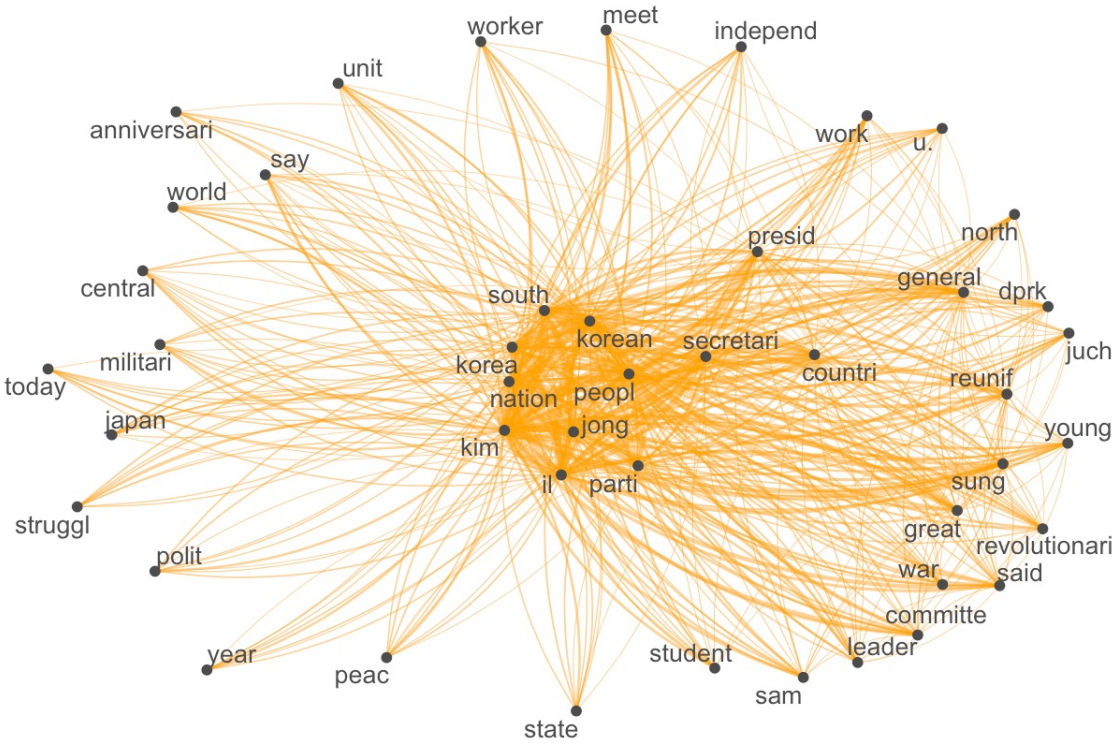
Text Analysis with R

Dr. Dani Madrid-Morales | dmmorales2@uh.edu | @DMadrid_M

Lee Kuan Yew School of Public Policy, National University of Singapore, 24 March 2022

UNIVERSITY of HOUSTON

What is this text?



Korean	*567*	*Il*	*254*
Kim	*492*	*struggle*	*238*
people	*361*	*law(s)*	*238*
south	*268*	*South*	*223*
worker(s)	*267*	*Jong*	*222*
Korea	*266*		

Outline

- Why Computational Text Analysis
- Bag of Words Approach
- QTA Approaches
  - Dictionary methods
  - Supervised ML
  - Unsupervised ML
- QTA in R: `quanteda`

Outline

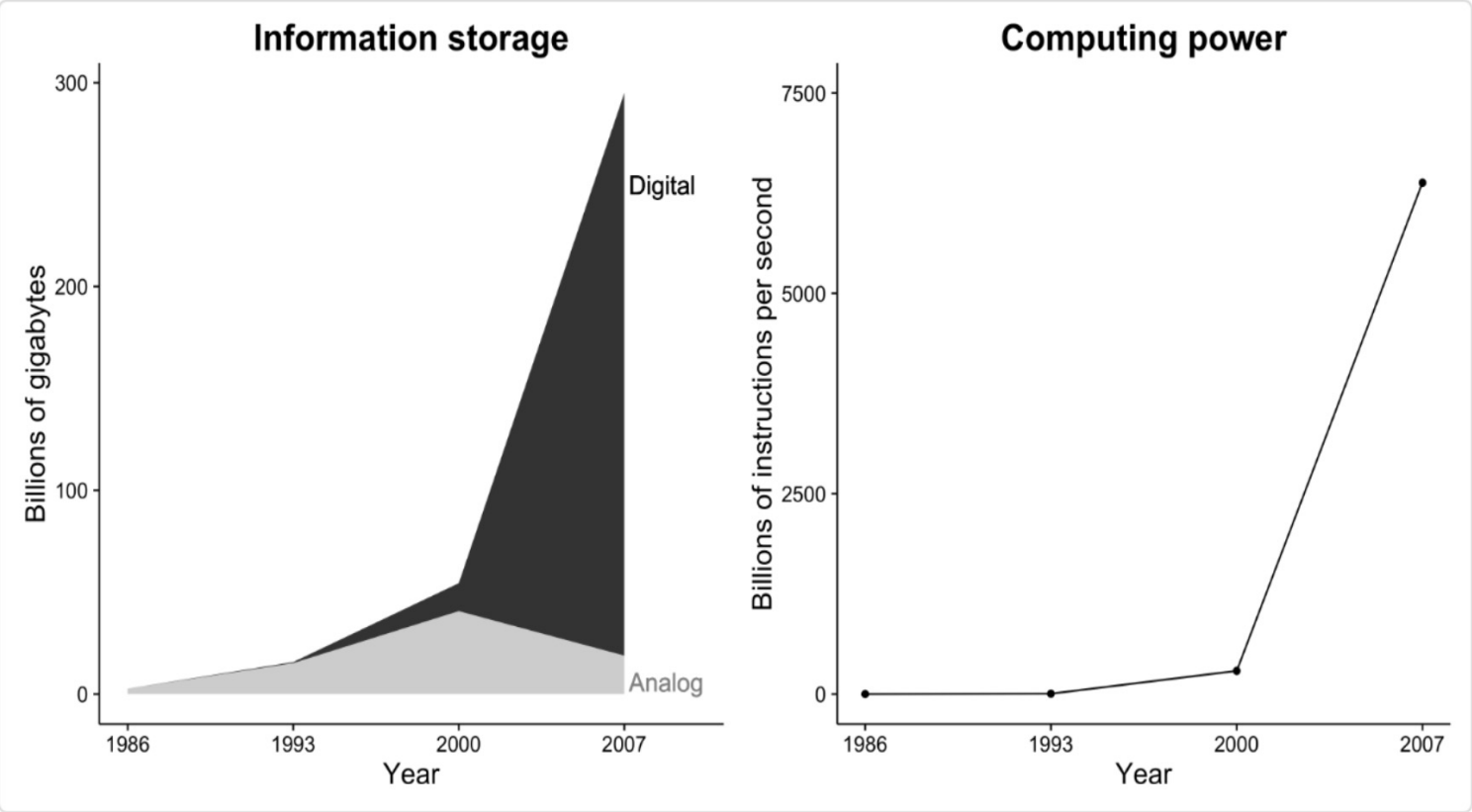
- **Why Computational Text Analysis**
- Bag of Words Approach
- QTA Approaches
  - Dictionary methods
  - Supervised ML
  - Unsupervised ML
- QTA in R: `quanteda`

Why Computational Text Analysis?

- Social Scientists have always used **texts as data**.
- There are costs (**human labor**) to large-scale text analysis.
- Computers can **lower these costs**.
  - Growth in computational power at relatively low costs.
  - Facilitated by widespread digitization of information.

Adapted from Terman (2018)

Why Computational Text Analysis?



Bail (2018)

Computational Social Sciences

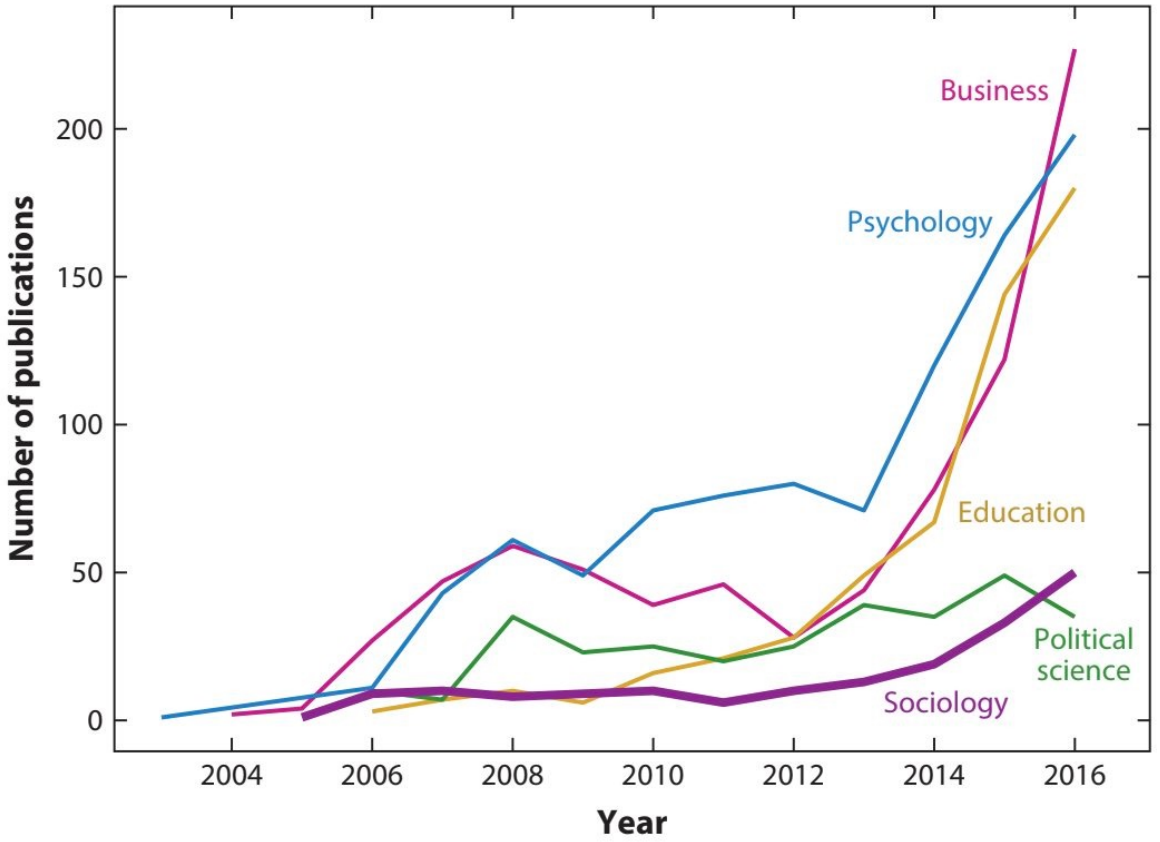
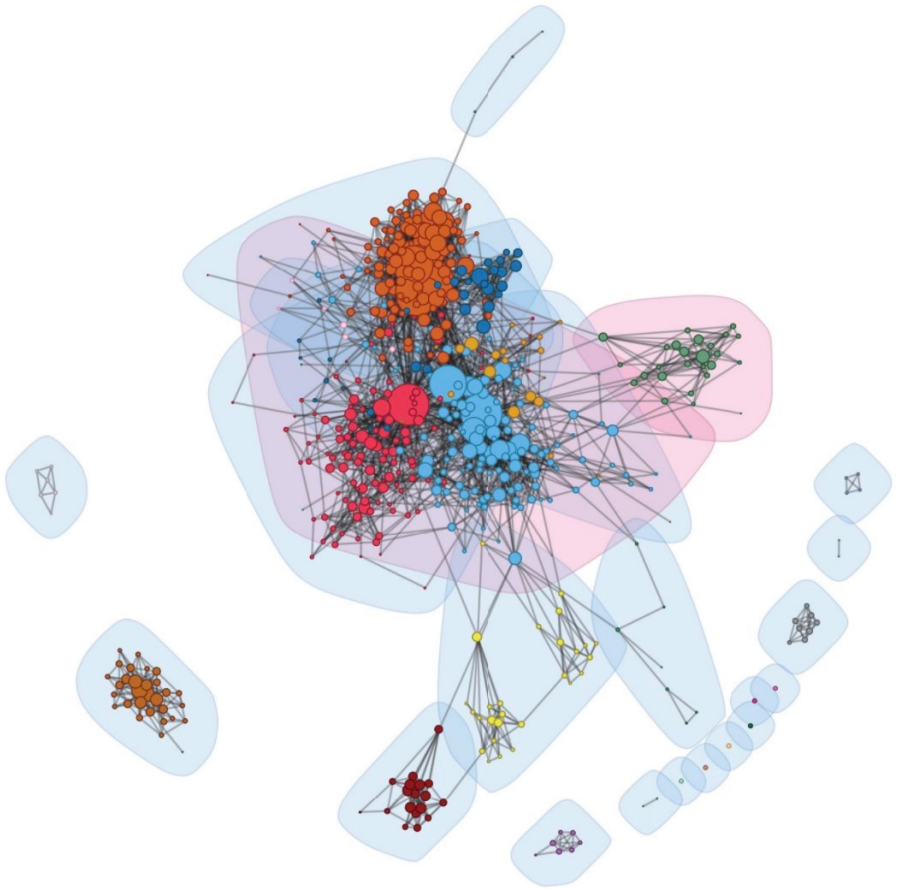


Figure 1

Number of computational social science publications by year—2003–2016—across four scholarly disciplines.

Eddelmann et al. (2020)

Computational Social Sciences



- Communication (25)
- Sociology (24)
- Political science (22)

- Management (74)
- Business (43)
- Information science and library science (20)

- Geography (28)
- Communication (23)
- Social sciences, interdisciplinary (22)

- Business and finance (20)
- Law (5)
- Business (4)

- Clinical psychology (17)
- Social sciences, interdisciplinary (14)
- Social sciences, biomedical (8)

- Experimental psychology (15)
- Behavioral sciences (10)
- Neurosciences (9)

- Sociology (19)
- Anthropology (15)
- Management (6)

- Geography (8)
- Multidisciplinary sciences (4)
- Sociology (3)

- Experimental psychology (10)
- Developmental psychology (5)
- Educational psychology (3)

- Social psychology (5)
- Multidisciplinary psychology (2)
- Educational psychology (1)

- Applied psychology (7)
- Business (3)
- Management (3)

- Management (3)
- Sociology (2)
- Business (1)

- Business (7)
- Management (1)

- Experimental psychology (4)
- Multidisciplinary psychology (2)
- Psychology (2)

- Business (3)
- Management (3)
- Sociology (1)

- Social psychology (2)
- Experimental psychology (1)
- Mathematics, interdisciplinary applications (1)

- Political science (2)
- International relations (1)

- Educational psychology (1)
- Social psychology (1)

- Multidisciplinary psychology (1)

- Business (1)
- Management (1)

- Sociology (1)

- Hospitality, leisure, sport and tourism (1)

- Multidisciplinary psychology (1)
- Psychology (1)

- Area studies (1)
- Political science (1)

(Caption appears on following page)

www.annualreviews.org • Computational Social Science and Sociology 24.5

Eddelmann et al. (2020)

Text and Political Science Pre 2000s

- Social interaction often **occurs in texts**
- (Some) Social Scientists **avoided** studying texts/speech
- Why?
  - Hard to find
  - Time Consuming
  - Not generalizable (each new data set implies a new coding scheme)
  - Difficult to store/search
  - Idiosyncratic to coders/researcher
  - Statistical methods/algorithms, computationally intensive

Grimmer (2018)

Text and Political Science Post 2000s

- Massive collections of texts are increasingly used as a data source in social science:

  - Congressional **speeches**, press releases, newsletters, ...
  - Facebook posts, **tweets**, emails, cell phone records, ...
  - Newspapers, magazines, **news broadcasts**, ...
  - Foreign **news sources**, treaties, sermons, fatwas, ...
  - Declarations, bilateral agreements, **UN resolutions**, ...

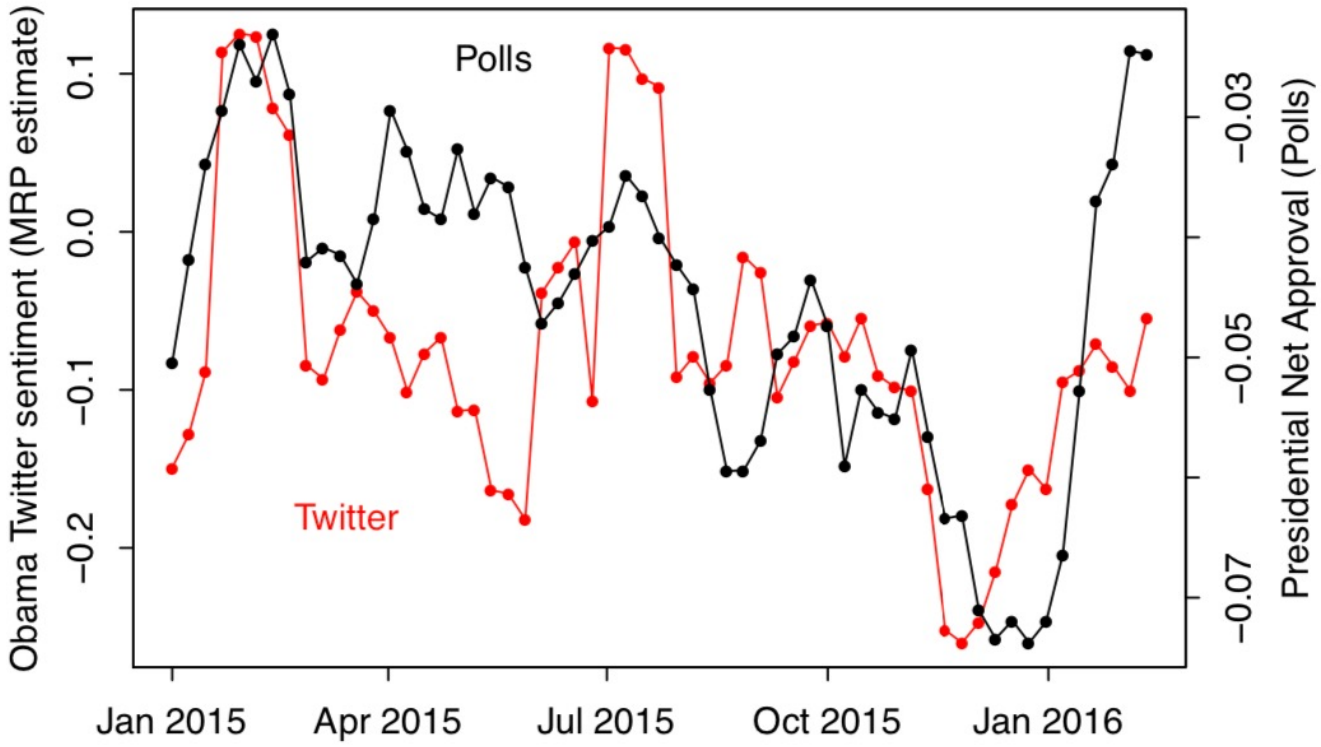
Grimmer (2018)

Example #1

- Question: How is the **tone** of social media discussions about US presidents related to their **approval ratings**?
- Data: Tweets about President Obama
- Method: Sentiment Analysis

#1: Presidential approval, Twitter & polls

Figure 3: Comparing Twitter- and survey-based measures of presidential job approval



Barberá (2015)

Example #2

- Question: What explains **coverage of news about taxation** around the world?
- Data: News articles (500.000+)
- Method: Document classification (supervised)

#2 Taxpaying, political system & the media

Table 3. Binary Logistic Regression Predicting the Framing of a Taxpayer in Public Spending Terms, With the Democracy Level Measured by the Reverse-Coded Freedom House Score.

	(Model 1)		(Model 2)		(Model 3)	
	B (*SE*)	Exp (*B*)	*B* (*SE*)	Exp (*B*)	*B* (*SE*)	Exp (*B*)
Country-level variables						
Democracy level	0.130*** (0.020)	1.139	0.004 (0.023)	1.004	0.024 (0.023)	1.025
Tax reliance	0.056*** (0.005)	1.057	0.044*** (0.005)	1.045	0.050*** (0.005)	1.052
Newspaper-level variables						
State ownership			−1.039*** (0.102)	0.354	1.194*** (0.365)	3.301
News agency			−0.834*** (0.116)	0.434	−0.920*** (0.119)	0.399
Tabloid			−0.056 (0.063)	0.945	−0.066*** (0.064)	0.936
Document-level variables						
Domestic context					−0.567*** (0.063)	0.567
State ownership × Domestic context					−2.259*** (0.360)	0.104
Constant	−0.533*** (0.112)	0.587	0.686*** (0.139)	1.986	0.892*** (0.146)	2.439
N	23,343[a]		23,343[a]		23,343[a]	
Nagelkerke R^2	.029		.055		.069	
Classification accuracy	84.3		84.5		84.9	

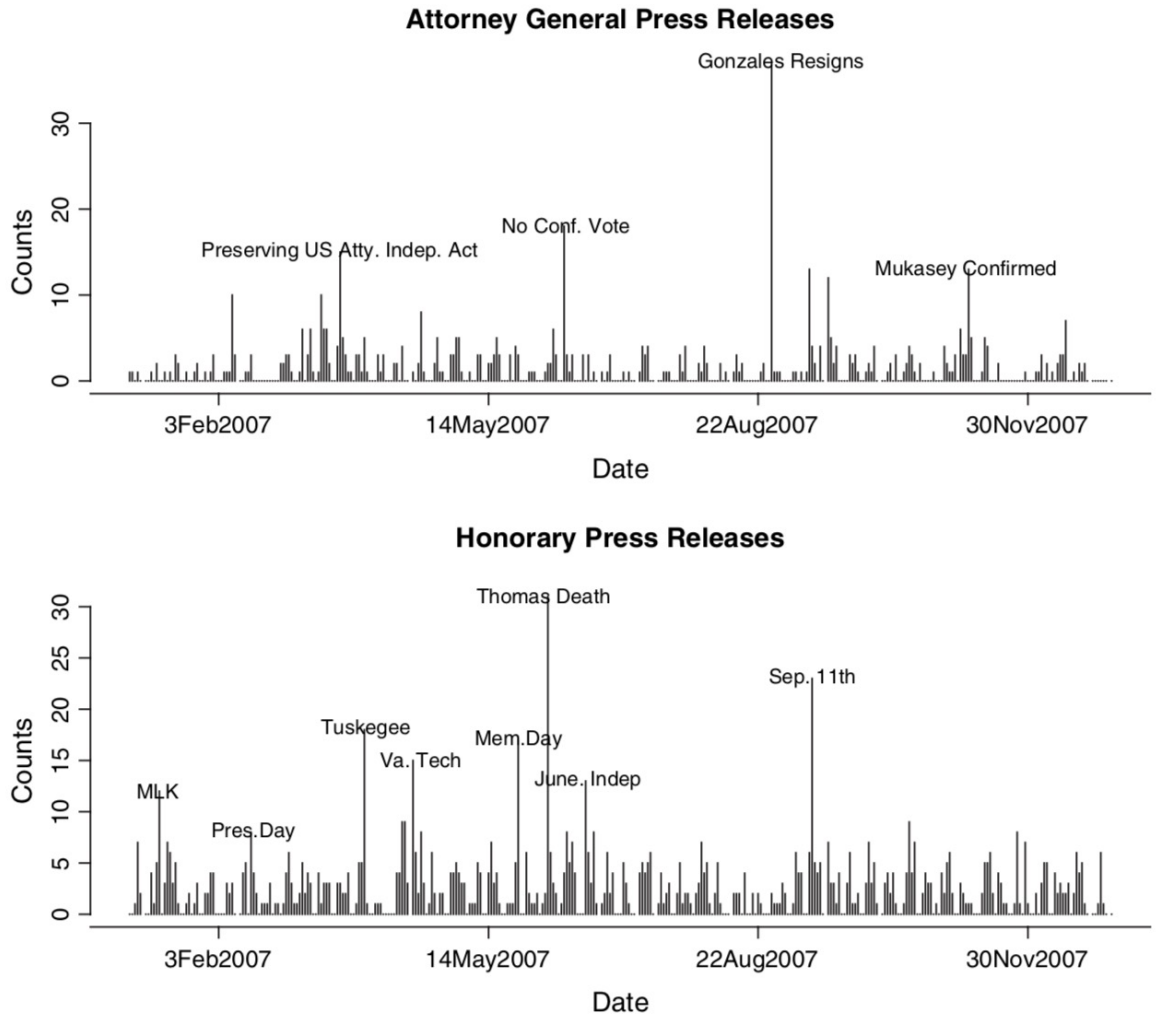
a. Lower than the original *N* = 25,191 due to missing data on the tax reliance measure for year 2015.
***$p \leq .001$, two-tailed.

Kananovich (2018)

Example #3

- Question: What **factors** shape US politicians' **issuance of press releases**?
- Data: Press releases of US Senators (24,000+)
- Method: Document classification (unsupervised)

#3: Press releases, Senate activity & news events



Grimmer (2010)

Weaknesses of New Text Data

- Incomplete
  - Social media posts are regularly **deleted, removed or hidden**;
- Inaccessible
  - There are **walls** around some type of data (e.g. Facebook);
- Non-representative
  - Not all **social groups** use the same type of (social) media;
- Algorithmically confounded
  - Predicted changes in human behavior sometimes simply mean changes in **human-computer interaction**

Adapted from Bail (2016)

Weaknesses of New Text Data

- Drift
  - Changes in usage and consumption of digital media
- Unstructured
  - Datasets can be **very messy** and difficult to reformat
- Sensitive
  - There are concerns with **privacy**, and how personal data is used
- Positivity (& Self-Presentation) Bias
  - Content published/posted by users does not reflect their **lived experience**.

Adapted from Bail (2016)

What Can Computational Text Methods Do?

Haystack metaphor ~ **Improve Reading**

✗ Interpreting meaning of a phrase [**Analyzing a straw of hay**]

- Humans: amazing! (Straussian political theory, analysis of English poetry…)
- Computers: struggle ☹

Comparing, Organizing, & Classifying Texts [**Organizing hay stack**]

- Humans: terrible. Tiny active memories 😐
- Computers: amazing!

Grimmer (2018a)

What Automated Text Methods Don't Do?

- Develop a **comprehensive statistical model** of language;
- Replace the **need to read**;
- Develop a **single tool** & evaluation for all possible tasks.

Grimmer (2018a)

Principles of Computational Text Analysis

1. All Quantitative Models of Language Are Wrong – But Some Are Useful.
2. Quantitative methods for text amplify resources and augment humans.
3. There is no globally best method for automated text analysis.
4. Validate! Validate! Validate!

Adapted from Grimmer and Stewart (2013)

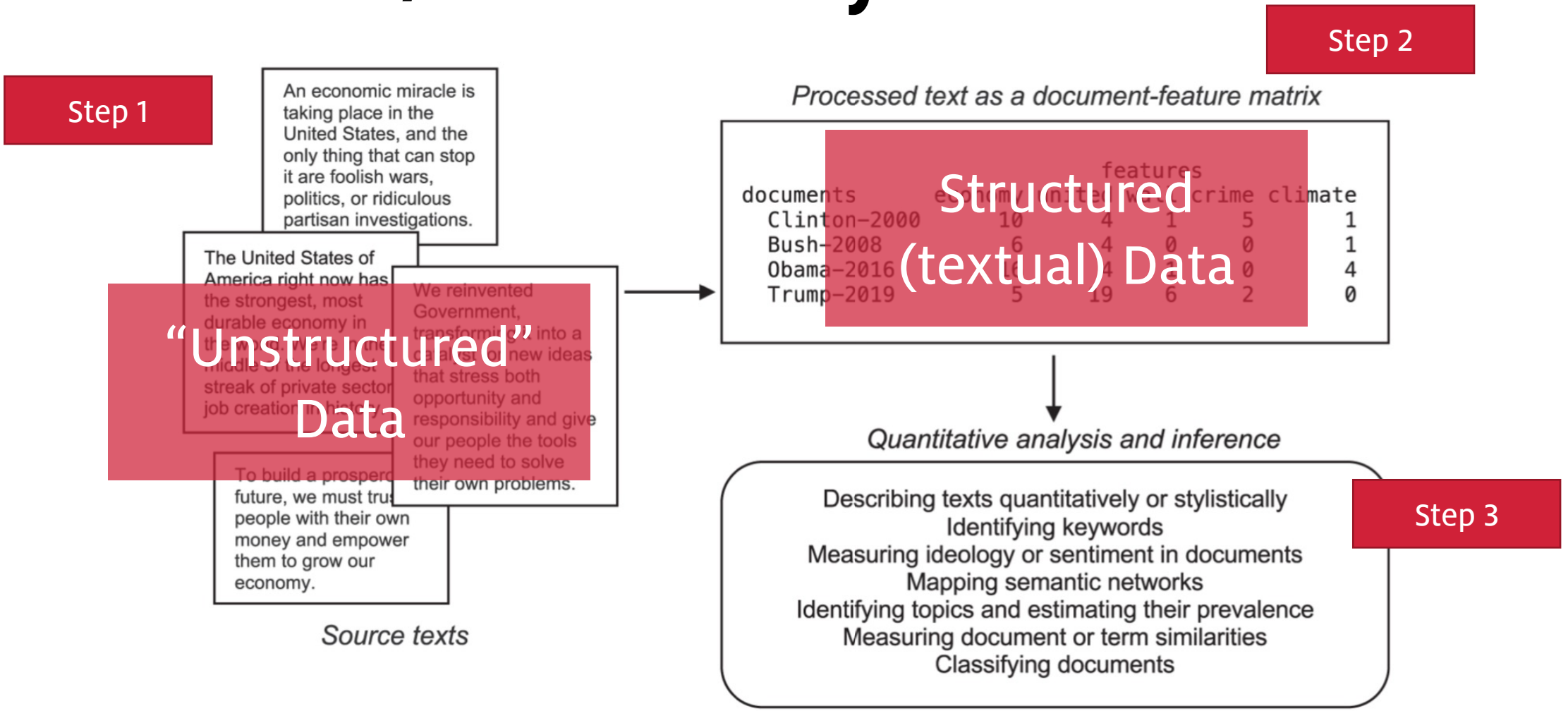
Outline

- ✓Why Computational Text Analysis
- Bag of Words Approach
- QTA Approaches
  - Dictionary methods
  - Supervised ML
  - Unsupervised ML
- QTA in R: `quanteda`

Outline

- ✓Why Computational Text Analysis
- **Bag of Words Approach**
- QTA Approaches
  - Dictionary methods
  - Supervised ML
  - Unsupervised ML
- QTA in R: `quanteda`

Text → DTM/DFM → Analysis



Benoit (2020)

Assumptions of QTA

- Texts can be represented through extracting their features
  - most common is the "**bag of words**" assumption
  - many other possible definitions of "features" (e.g., word embeddings)
- A **document-feature matrix** can be analyzed using quantitative methods to produce meaningful and valid estimates of the underlying characteristic of interest

Barberá & Benoit (2018)

Key features of quantitative text analysis

1. Selecting texts: Defining the **corpus**
2. Conversion of texts into a **common electronic format**
3. Defining documents: deciding what will be the **documentary unit** of analysis (segmentation or aggregation)

Barberá (2016)

Key features of quantitative text analysis

4. Defining and refining features. These can take a variety of forms, including **tokens**, equivalence classes of tokens (**dictionaries**), selected phrases, human-coded segments (of possibly variable length), linguistic features, and more.
5. Conversion of textual features into a **quantitative matrix**
6. A quantitative or statistical procedure to **extract information** from the quantitative matrix.
7. Summary and interpretation of the quantitative results

Barberá (2016)

Key concepts

(text) corpus	a large and structured set of texts for analysis
document	each of the units of the corpus
types	for our purposes, a unique word
tokens	any word – so token count is total words

Barberá & Benoit (2018)

Key concepts

Consider these two sentences:

> "A corpus is a set of documents."
> "This is the second document in the corpus."

This is a corpus with 2 **documents**, where each document is 1 **sentence**. The first document has 6 **types** and 7 **tokens**. The second document has 7 **types** and 8 **tokens**.

[Punctuation is considered a token too]

Barberá & Benoit (2018)

Preprocessing for QTA

One (of many) recipes for preprocessing. End goal is to **retain useful information only**.

1) Remove capitalization, punctuation
2) Discard Word Order (Bag of Words Assumption)
3) Discard stop words
4) Create Equivalence Class: Stem, Lemmatize, or synonym
5) Discard less useful features~ depends on application
6) Other reduction, specialization

Output: Count vector, each element counts **occurrence of stems**

Grimmer (2018b)

From words to numbers

1. Bag-of-words assumption
2. Pre-processing text
   - Capitalization, cleaning digits/URLs, removing stop words and sparse words…
   - Stemming
   - [Part-of-speech tagging]
3. Document-term matrix
   - **W**: matrix of N documents by M unique words
   - Wim = number of times m-th words appears in i-th document.
   - Usually large matrix, but sparse (so it fits well in memory)

Barberá (2016)

Document-term matrix (or DTM)

	Word 1	*Word 2*	*Word 3*	*Word 4*	*Word 5*	...	*M Words*
Document 1	1	3	2	0	0	...	
Document 2	0	0	1	1	0	...	
Document 3	1	1	0	2	3	...	
Document 4	3	1	0	0	0	...	
Document 5	0	1	0	3	1	...	
...							
Document n	0	1	1	0	1	...	

$$\times = \begin{pmatrix} 2 & 1 & 0 & \cdots & 2 \\ 1 & 0 & 1 & \cdots & 3 \\ 3 & 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & \cdots & 3 \end{pmatrix}$$

From words to numbers

1. Preprocess text (raw data)

Tweet 1 ***"@MEPcandidate thank you and congratulations, you're the best #EP2014"***

Tweet 2 ***"@MEPcandidate You're an inept, I would never vote for you"***

From words to numbers

1. Preprocess text: lowercase

Tweet 1 ***"@MEPcandidate thank you and congratulations, you're the best #EP2014"***

"@mepcandidate thank you and congratulations, you're the best #ep2014"

Tweet 2 ***"@MEPcandidate You're an inept, I would never vote for you"***

"@mepcandidate you're an inept, i would never vote for you"

From words to numbers

1. Preprocess text: lowercase, remove stop words, remove punctuation

Tweet 1 ***"@MEPcandidate thank you and congratulations, you're the best #EP2014"***

"@mepcandidate thank congratulations you're best #ep2014"

Tweet 2 ***"@MEPcandidate You're an inept, I would never vote for you"***

"@mepcandidate you're inept never vote"

From words to numbers

1. Preprocess text: lowercase, remove stop words, remove punctuation, stem, tokenize

Tweet 1 ***"@MEPcandidate thank you and congratulations, you're the best #EP2014"***

"@ thank congratul you'r best #ep2014"

Tweet 2 ***"@MEPcandidate You're an inept, I would never vote for you"***

"@ you'r inept never vote"

From words to numbers

	@	thank	congratul	you'r	#ep2014	inept	best	vote
Document 1	1	1	1	1	1	0	1	0
Document 2	1	0	0	1	0	1	0	1

Outline

- ✓Why Computational Text Analysis
- ✓Bag of Words Approach
- QTA Approaches
  - Dictionary methods
  - Supervised ML
  - Unsupervised ML
- QTA in R: `quanteda`

Outline

- ✓Why Computational Text Analysis
- ✓Bag of Words Approach
- **QTA Approaches**
  - Dictionary methods
  - Supervised ML
  - Unsupervised ML
- QTA in R: `quanteda`

An Overview of Methods

- Three broad approaches to computational text analysis:
  - Dictionary methods: We apply lists of words (or lexicons) to documents in order to **quantify the presence/absence** of certain latent characteristics in the texts
  - Supervised methods: We identify what we're interested in first, and then use computers to **extend our insights** to a larger population of unseen documents.
  - Unsupervised methods: We do not specify the conceptual structure of the texts beforehand. Instead, we use the model to **discover a structure** that best explains the documents.

Adapted from Terman (2018)

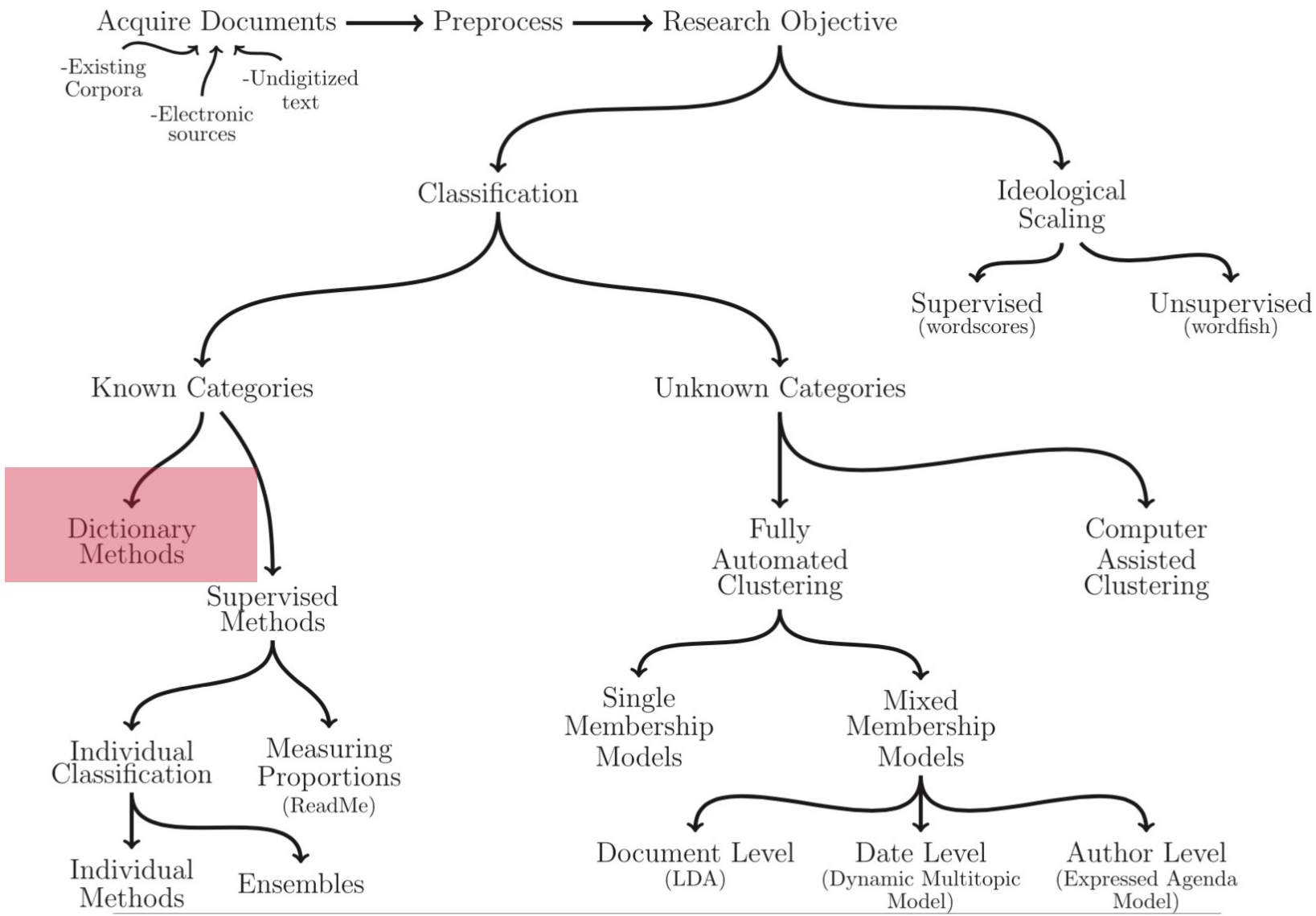


Fig. 1 An overview of text as data methods.

Grimmer and Stewart (2013)

Rationale for dictionaries

- Rather than count ALL words that occur in a text we count **pre-defined words** associated with specific meanings
- Dictionaries have two components:
  - key ~ the label for the **equivalence class** for the concept or canonical term
  - values ~ (multiple) terms or patterns that are declared equivalent occurrences of the key class

Barberá & Benoit (2018)

Sentiment Lexicons

```
> library(tidytext)

> get_sentiments("bing")

# A tibble: 6,788 x 2
          word sentiment
         <chr>     <chr>
 1     2-faced  negative
 2     2-faces  negative
 3          a+  positive
 4    abnormal  negative
 5     abolish  negative
 6  abominable  negative
 7  abominably  negative
 8   abominate  negative
 9 abomination  negative
10       abort  negative
# ... with 6,778 more rows
```

```
> get_sentiments("afinn")

# A tibble: 2,476 x 2
          word score
         <chr> <int>
 1     abandon    -2
 2   abandoned    -2
 3    abandons    -2
 4    abducted    -2
 5   abduction    -2
 6  abductions    -2
 7       abhor    -3
 8    abhorred    -3
 9   abhorrent    -3
10       abhors    -3
# ... with 2,466 more rows
```

```
> get_sentiments("nrc")

# A tibble: 13,901 x 2
          word sentiment
         <chr>     <chr>
 1      abacus     trust
 2     abandon      fear
 3     abandon  negative
 4     abandon   sadness
 5   abandoned     anger
 6   abandoned      fear
 7   abandoned  negative
 8   abandoned   sadness
 9 abandonment     anger
10 abandonment      fear
# ... with 13,891 more rows
```

Silge (2016)

Advantages of dictionary methods

- Very high (perfect) **reliability** because there is no human decision making as part of the text analysis procedure
- The **validity** of the results needs to be proven by the researcher

Adapted from Barberá & Benoit (2018)

Challenges in using dictionaries

- Building a reliable dictionary can be time-consuming and costly
- Most dictionaries are so **domain-specific** that can't be reused widely across projects.
- There is **no single 'best in breed'** approach to sentiment analysis.
  - A researcher's choice can severely affect the outcome – and thus validation is crucial.
- Language is **not static**, and therefore dictionaries can be obsolete quickly.

Adapted from Curini & Fahey (2020)

Dictionaries are highly specific to context

- Loughran and McDonald used the Harvard-IV-4 TagNeg (H4N) dictionary to classify sentiment for a **corpus of 50,115 firm-year 10-K filings** from 1994–2008
- They found that almost **three-fourths** (75%!!!) of the "negative" words of H4N were typically not negative in a financial context
  - e.g. mine or cancer, or tax, cost, capital, board, liability, foreign, and vice

Barberá & Benoit (2018)

Dictionaries are highly specific to context

- In this study, the researchers faced **two problems**
  - polysemes – words that have multiple meanings in different contexts
  - The Harvard IV dictionary lacked important negative financial words, such as felony, litigation, restated, misstatement, and unanticipated

Barberá & Benoit (2018)

Lessons learnt?

- Validate
- Validate
- Validate!

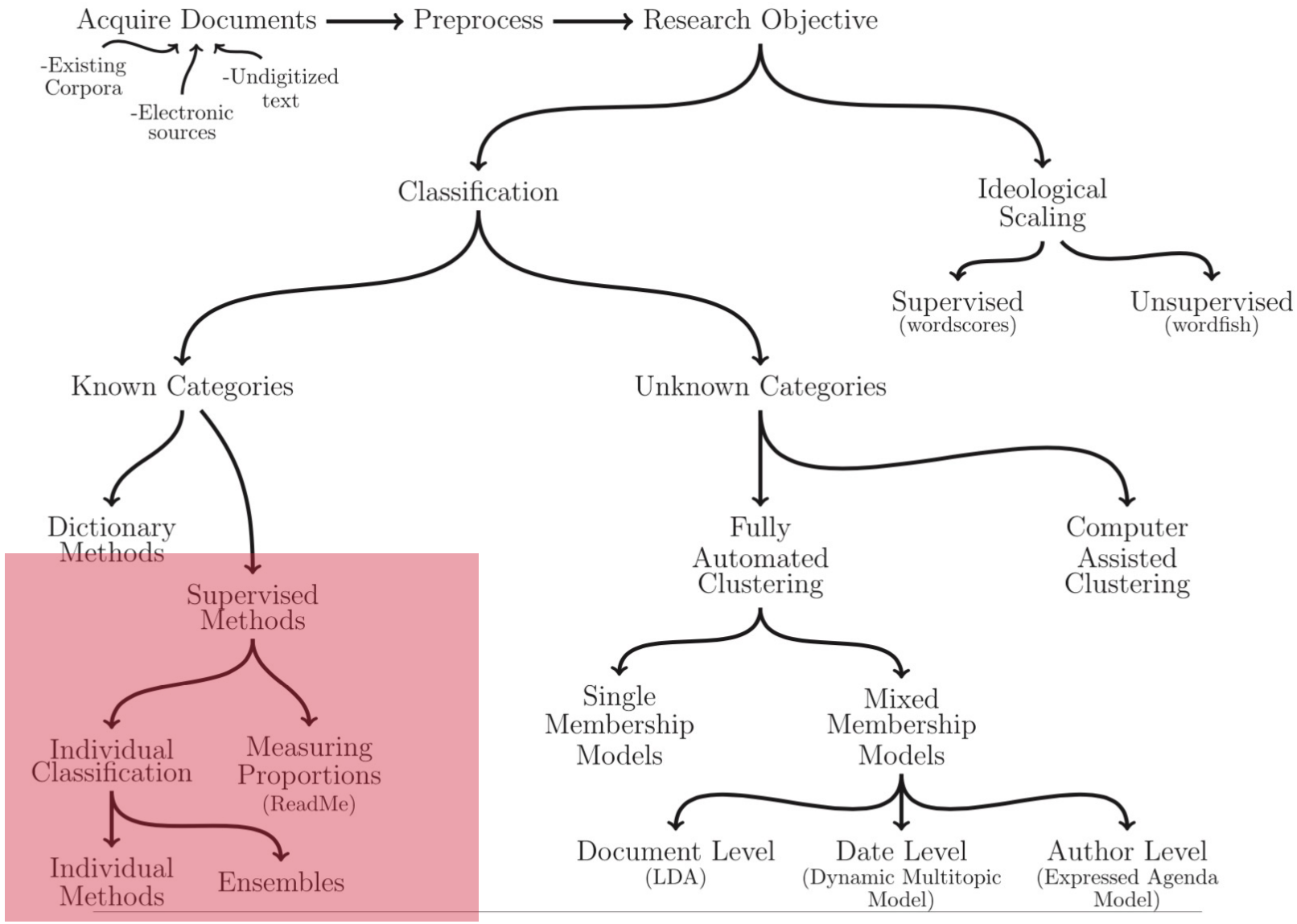


Fig. 1 An overview of text as data methods.

Grimmer and Stewart (2013)

Supervised Learning vs. Dictionary Methods

- Supervised learning can be conceptualized as a **generalization of dictionary methods**, where features associated with each categories (and their relative weight) are learned from the data
- Dictionary methods:
  - Advantage: not corpus-specific, cost to apply to a new corpus is trivial
  - Disadvantage: not corpus-specific, so performance on a new corpus is unknown (domain shift)

Barberá (2019)

Components to Supervised Learning Method

Hand coding is used to **train**, or supervise, statistical models to classify texts in pre-determined categories.

1. Set of **known categories**
   - Positive Tone, Negative Tone
   - Pro-war, Ambiguous, Anti-war
2. Set of **hand-coded documents**
   - Coding done by human coders
     - Training Set: documents we'll use to learn how to code
     - Validation Set: documents we'll use to learn how well we code

Terman (2018)

Components to Supervised Learning Methods

3. Set of **unlabeled documents** that we want to classify.
4. Method to extrapolate from hand coding to unlabeled documents (dictionary methods, logistic regression, Naïve Bayes...).
5. Validate by comparing **predicted label to actual** (hand-coded) labels.

Terman (2018)

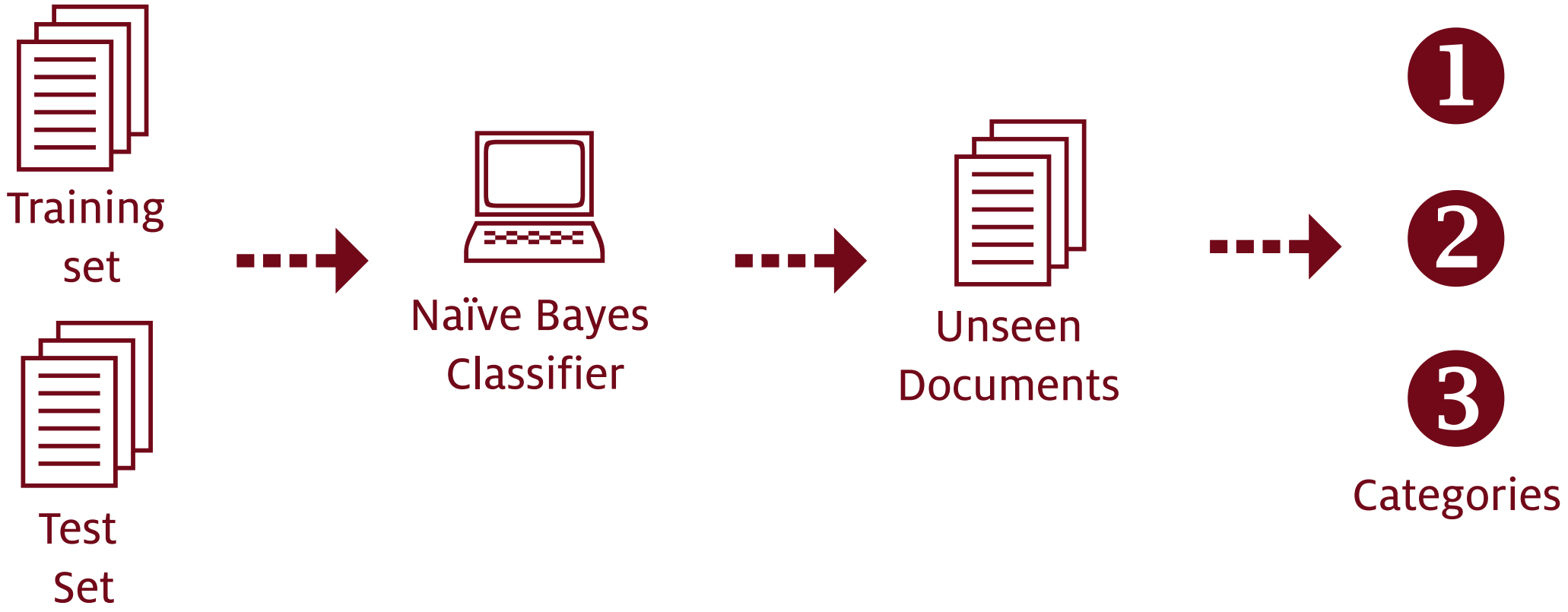
Example: Federalist Papers

- Mosteller & Wallace (1964) wanted to **determine the authorship** of 12 of the Federalist papers
  - Of the 85 documents, the writer of 12 papers was disputed (Hamilton or Madison)

- This study used **Bayesian methods** & **unparalleled computational power**, and a set of documents of known authorship

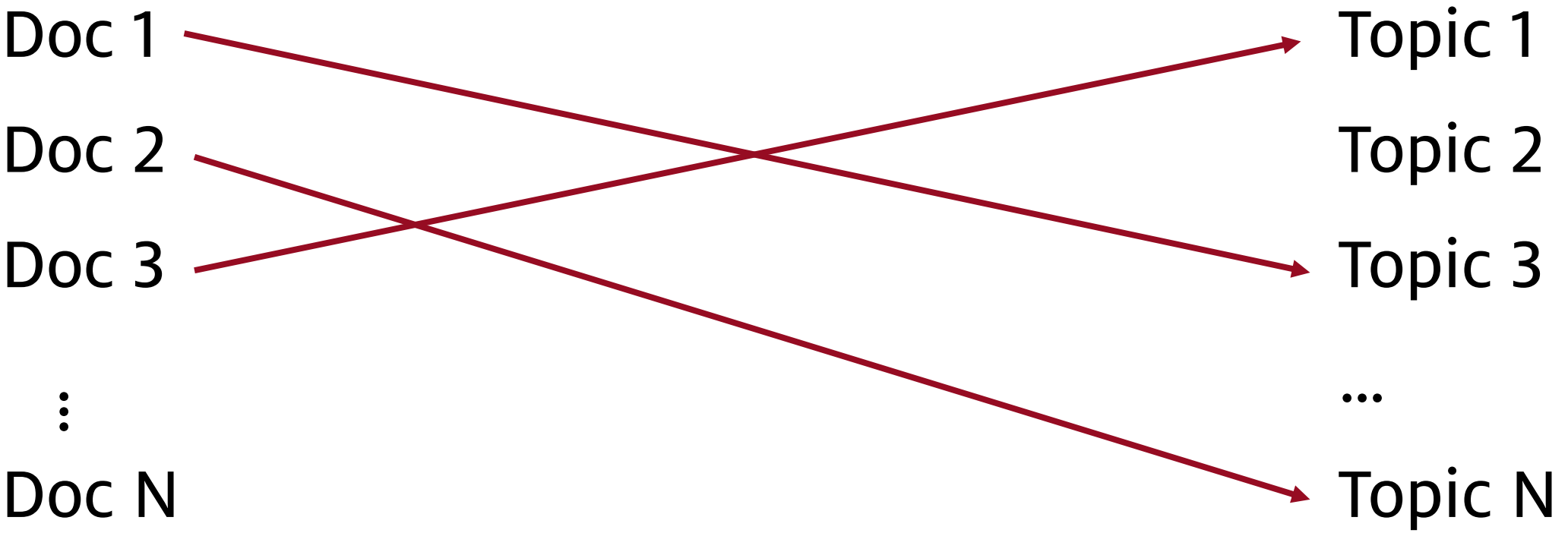
- Using **word frequencies and computing joint probabilities**, they eventually attributed the papers to Madison.

Supervised ML (Naïve Bayes Classifier)



Single Membership Models

Supervised ML (e.g. Naïve Bayes Classifier)



Adapted from Terman (2018)

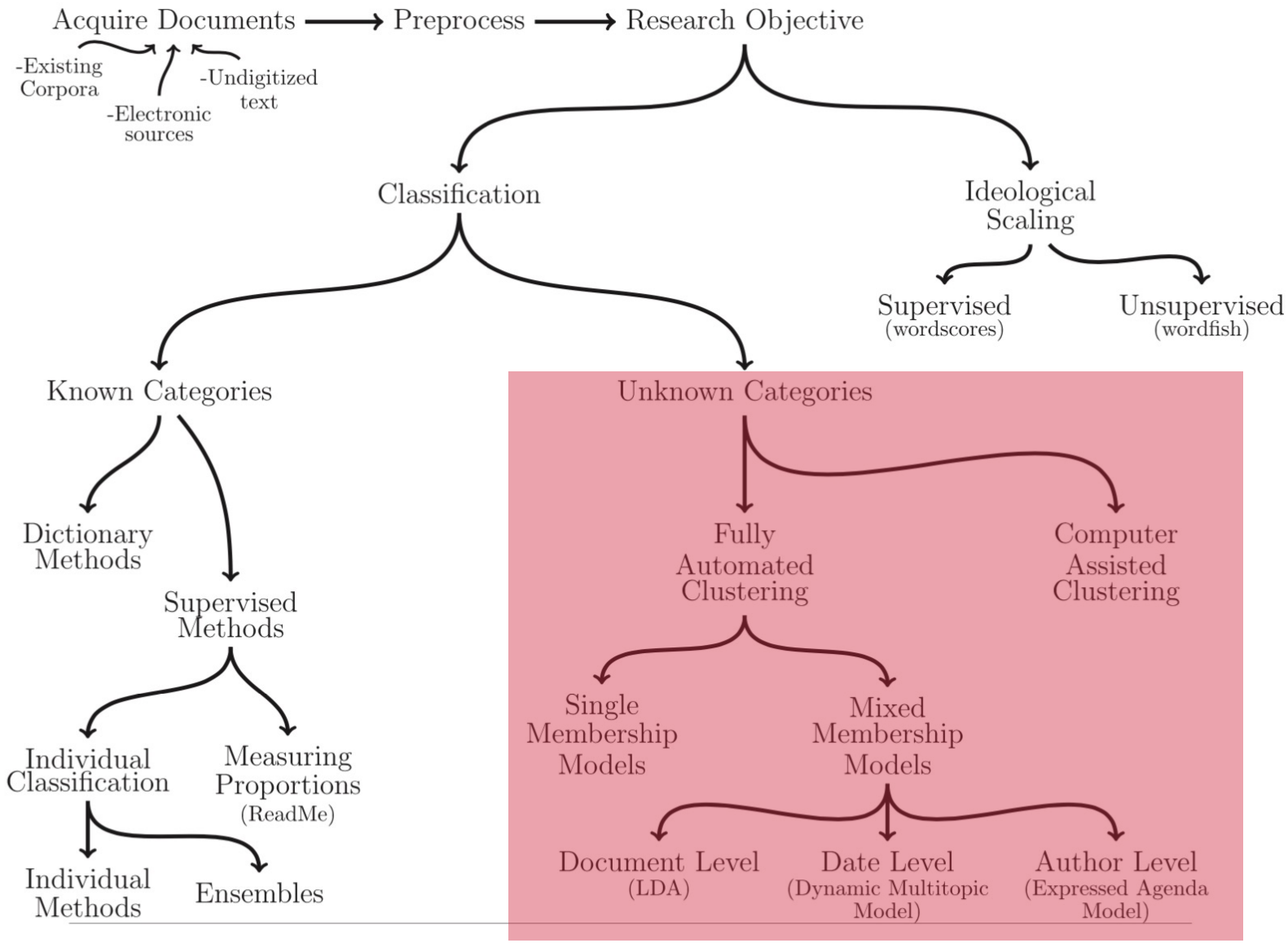


Fig. 1 An overview of text as data methods.

Grimmer and Stewart (2013)

Components of Unsupervised Learning Methods

Discover **new ways of organizing** texts that are theoretically useful, but perhaps understudied or previously unknown.

1. Set of **unlabeled documents** that we want to classify.
2. Method to **discover categories** and then classify documents into those categories (k-means clustering, topic models).
3. Interpretation skills to **assign labels to categories** and understand what they mean.

Terman (2018)

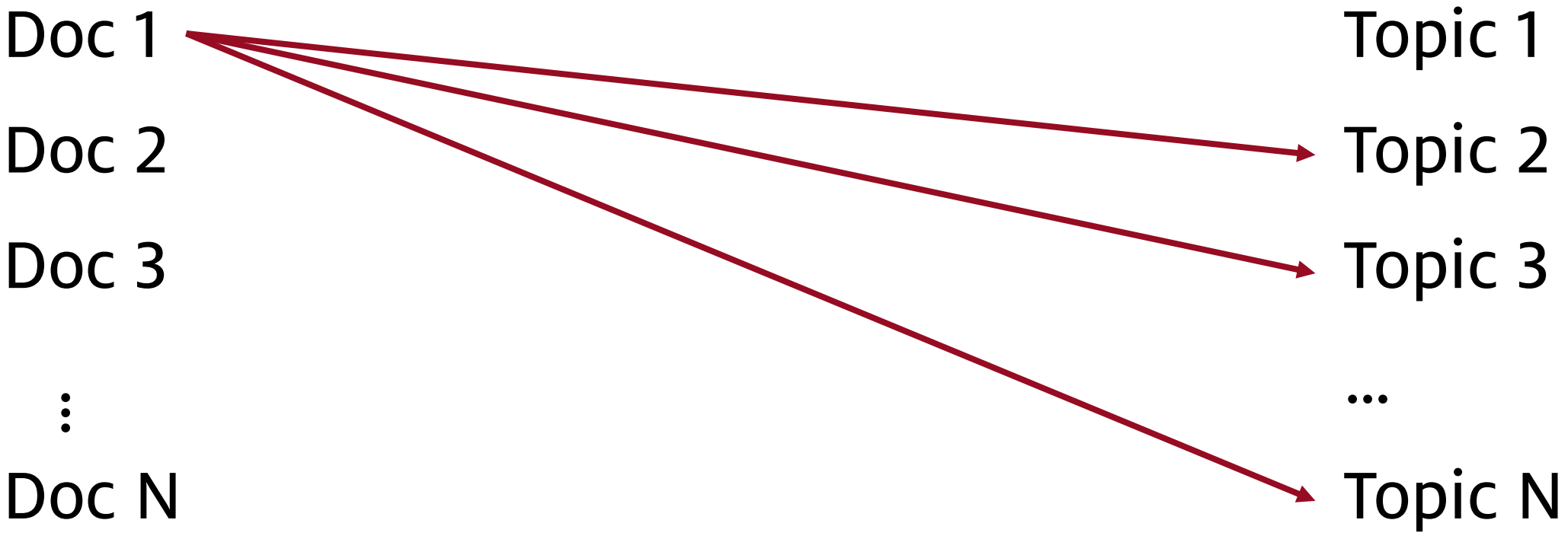
What is Topic Modeling?

- The algorithm "discovers" **abstract topics** that can be thought of as a constellation of words that tend to **show up together**.
- It is **mixed membership** method because it considers each document to be a mixture of different topics.

Terman (2018)

Mixed Membership models

Topic modeling (e.g. LDA)



Adapted from Terman (2018)

How does topic modeling work?

- Imagine our goal is to **topic model** the following documents in a corpus:
  1. Russia threatens Georgia and Ukraine with sanctions.
  2. New threat against Ukraine and Georgia over election interfering.
  3. The rising price of oil and gold.
  4. UAE reduces production of oil significantly.
  5. With new sanctions coming, an increase in oil prices looms over Georgia.

Adapted from Terman (2018)

How does topic modeling work?

- We suspect that our corpus contains **2 topics**.
- We want to get those topics from the **co-occurrence of words** in each document.

How does topic modeling work?

- Goal: Topic model the following documents:
  1. Russia threatens Georgia with sanctions again.
  2. New threat against Ukraine and Georgia over election interfering.
  3. The rising price of oil and gold.
  4. UAE reduces production of oil significantly.
  5. With new sanctions coming, an increase in oil prices looms over Georgia.

Adapted from Terman (2018)

How does topic modeling work?

- Goal: Topic model the following documents:
  1. Russia **threatens Georgia** with **sanctions** again.
  2. New **threat** against **Ukraine** and **Georgia** over **election** interfering.
  3. The rising **price** of **oil** and **gold**.
  4. UAE reduces **production** of **oil** significantly.
  5. With new **sanctions** coming, an increase in **oil prices** looms over **Georgia**.

Adapted from Terman (2018)

How does topic modeling work?

- Goal: Topic model the following documents:
  1. Russia threatens Georgia with sanctions again.
  2. New threat against Ukraine and Georgia over election interfering.
  3. The rising price of oil and gold.
  4. UAE reduces production of oil significantly.
  5. With new sanctions coming, an increase in oil prices looms over Georgia.

Topic A (interpreted as 'politics')

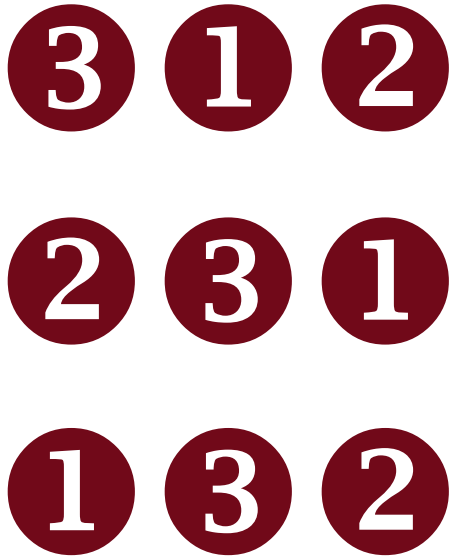
Topic B (interpreted as 'economy')

Adapted from Terman (2018)

Unsupervised ML (Topic Modeling)

Unseen Documents

Topic Modeling



Multiple Categories

Example: Putin vs. Medvedev

- Goal:
  - Did Putin and Medvedev **talk about different issues** when in office?
- Data:
  - All **speeches of the President of Russia** were scraped from the Kremlin website
  - All **speeches of the Prime Minister of Russia** for 2008–12 were scraped from the Premier website
- Method:
  - Unsupervised **Topic Modeling** (LDA)

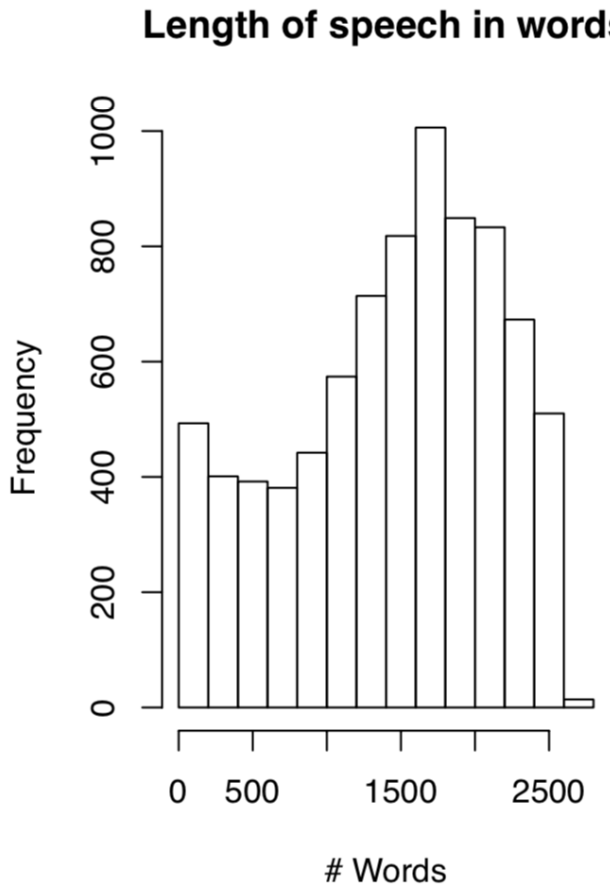
Adapted from Elkink (2018)

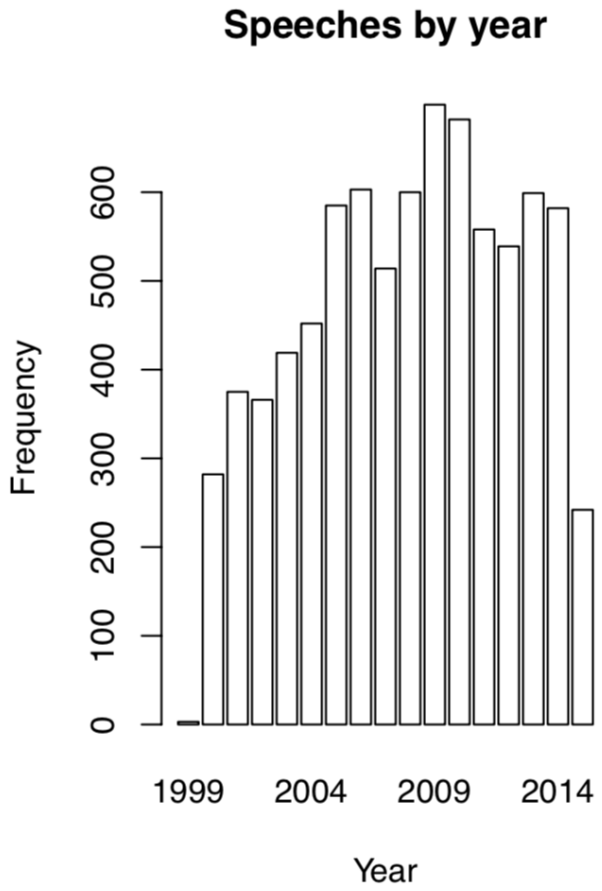
Example: Putin vs. Medvedev | Data cleaning

- Scraped data was **pre-processed**:
  - removing short words;
  - stemming words;
  - removing very frequent and very rare words;
  - removing stop words;
  - removing tiny paragraphs.
- Paragraphs were used as **unit of analysis**:
  - 34,499 paragraphs
  - 11,595 documents
  - 3,282 types

Adapted from Elkink (2018)

Example: Putin vs. Medvedev | Texts





Elkink (2018)

Example: Putin vs. Medvedev | LDA Output

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
defense	states	economies	company	investments
military	people	develop	invest	investor
forces	w	tasks	productions	zones
armed	its	social	market	investment
technician	solved	growth	project	environmental
Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
farms	negotiations	waspi	developed	vladimirovich
earth	worker	vladimir	roads	vladimir
land	problems	vladimir	project	colleagues
state council	meetings	finish	east	rescue
rural	affordable	rescue	transport	groups

Elkink (2018)

Overview of Text as Data Methods

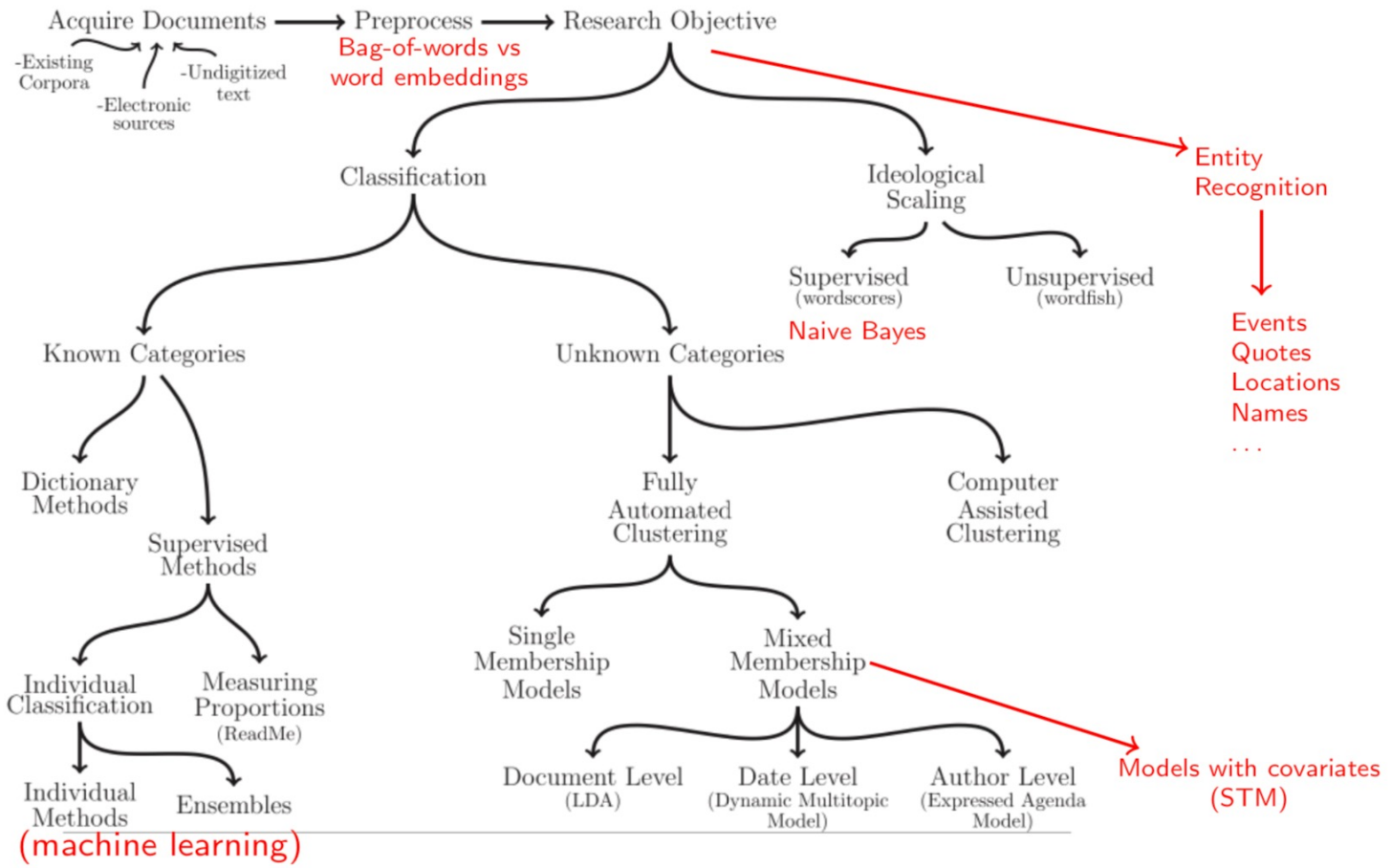


Fig. 1 in Grimmer and Stewart (2013)

Adapted from Barberá & Benoit (2018)

Outline

- ✓Why Computational Text Analysis
- ✓Bag of Words Approach
- ✓QTA Approaches
  - ✓Dictionary methods
  - ✓Supervised ML
  - ✓Unsupervised ML
- QTA in R: `quanteda`

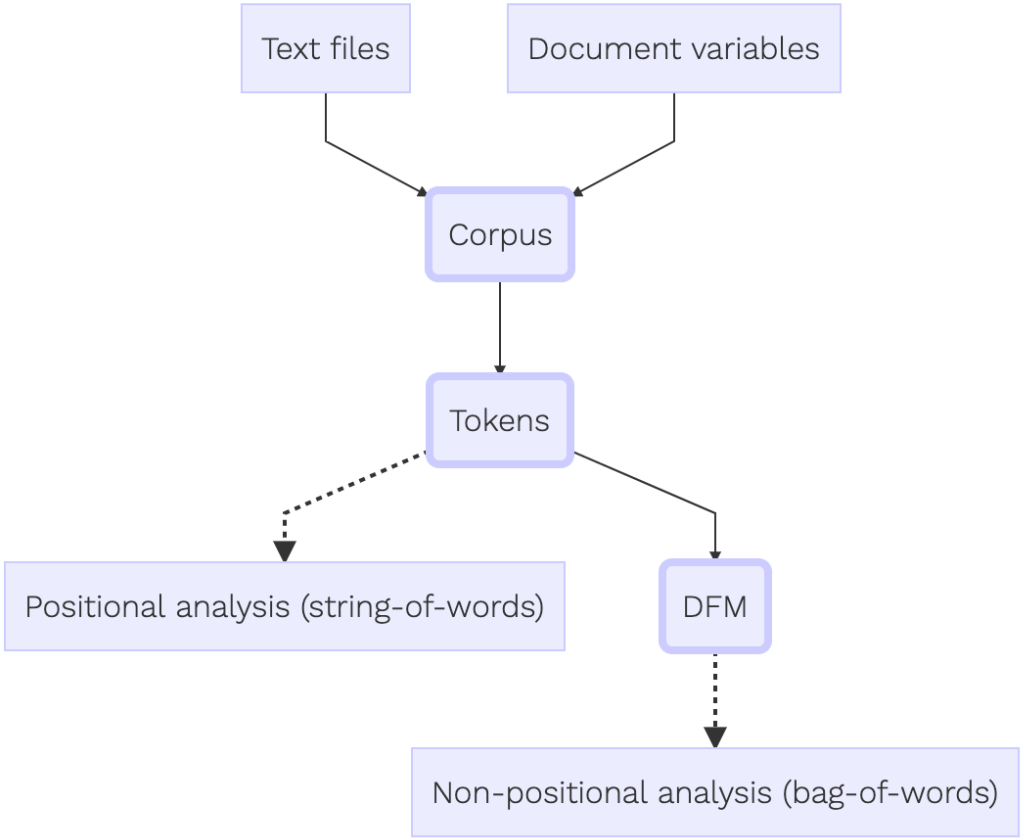
Outline

- ✓Why Computational Text Analysis
- ✓Bag of Words Approach
- ✓QTA Approaches
  - ✓Dictionary methods
  - ✓Supervised ML
  - ✓Unsupervised ML
- QTA in R: `quanteda`

R's `quanteda` package

- A powerful tool (think of a Swiss knife) to implement QTA.
  - There are other approaches such as `tidytext` and `tm`
  - It is powerful because
    - It has a simple and consistent grammar;
    - It includes tools to do most (if not all) of the processes described above;
    - It has been shown to be more resource efficient than other similar packages;
    - It was created by political scientists, and has used widely;
    - It is regularly updated (VERY important!).

Standard QTA procedure in quanteda



Outline

- ✓Why Computational Text Analysis
- ✓Bag of Words Approach
- ✓QTA Approaches
  - ✓Dictionary methods
  - ✓Supervised ML
  - ✓Unsupervised ML
- ✓QTA in R: `quanteda`

Text Analysis with R

Dr. Dani Madrid-Morales | dmmorales2@uh.edu | @DMadrid_M

Lee Kuan Yew School of Public Policy, National University of Singapore, 24 March 2022

UNIVERSITY of HOUSTON