Introduction: Text Data for Social Science

Elliott Ash

Max Planck Summer School 2017

Welcome

- This course provides an introduction to social-science research with text data.

Goals of the Course

- Think about research questions that require text data to answer
- Prepare text corpora and transform them into matrices of text features
- Applications of machine learning methods for describing and analyzing high-dimensional data
- Applications of causal inference approaches to text data

Goals of the Course

- Think about research questions that require text data to answer
- Prepare text corpora and transform them into matrices of text features
- Applications of machine learning methods for describing and analyzing high-dimensional data
- Applications of causal inference approaches to text data

Goals of the Course

- Think about research questions that require text data to answer
- Prepare text corpora and transform them into matrices of text features
- Applications of machine learning methods for describing and analyzing high-dimensional data
- Applications of causal inference approaches to text data

Goals of the Course

- Think about research questions that require text data to answer
- Prepare text corpora and transform them into matrices of text features
- Applications of machine learning methods for describing and analyzing high-dimensional data
- Applications of causal inference approaches to text data

Outline

Lecture Times

- Monday through Friday, 11am-12:30pm
- Monday and Thursday, 4:30pm-5:15pm
- The exercise sessions will have some lecture material, and I will help with the homework assignment.

Course Web Site

- elliottash.com/text_course

Readings

- No required reading for the course outside the slides.
  - I highly recommend chapters 1 and 2 of *Natural Language Processing in Python*
  - I'd rather you spend more time on the assignment than reading.
- The syllabus has a long list of articles that the lecture slides are drawn from, so go there for more detail.

Python

- If you are going to learn a programming language, learn Python.
  - This is especially true for text data applications.
- I recommend Anaconda's Python 3.6 distribution:
  - `continuum.io/downloads`
- Refer to syllabus for required packages:
  - Natural Language Processing: nltk, spacy, gensim
  - Machine Learning: sklearn
  - Data Management: pandas
  - Statistics: statsmodels
  - Visualization: wordcloud, seaborn

Python

- If you are going to learn a programming language, learn Python.
  - This is especially true for text data applications.
- I recommend Anaconda's Python 3.6 distribution:
  - `continuum.io/downloads`
- Refer to syllabus for required packages:
  - Natural Language Processing: nltk, spacy, gensim
  - Machine Learning: sklearn
  - Data Management: pandas
  - Statistics: statsmodels
  - Visualization: wordcloud, seaborn

Python

- If you are going to learn a programming language, learn Python.
  - This is especially true for text data applications.
- I recommend Anaconda's Python 3.6 distribution:
  - `continuum.io/downloads`
- Refer to syllabus for required packages:
  - Natural Language Processing: nltk, spacy, gensim
  - Machine Learning: sklearn
  - Data Management: pandas
  - Statistics: statsmodels
  - Visualization: wordcloud, seaborn

Course Assignment

- The course assignment asks you to go through the major steps of a text analysis project.
- Tasks for later days might change depending on how much gets covered.

Exam

- The exam will be based on the slides.
  - I will write and grade it to avoid penalizing students with different language/coding/statistics backgrounds.
- I will provide some practice questions on Friday.

Office Hours Etc

- I will be in my office most of the day outside of the lectures.
  - Office: Room 208, this building
  - Set up a time by email: `etash@princeton.edu`.
- We can talk about the course material, your research, anything you want.

Outline

The Era of Big Data

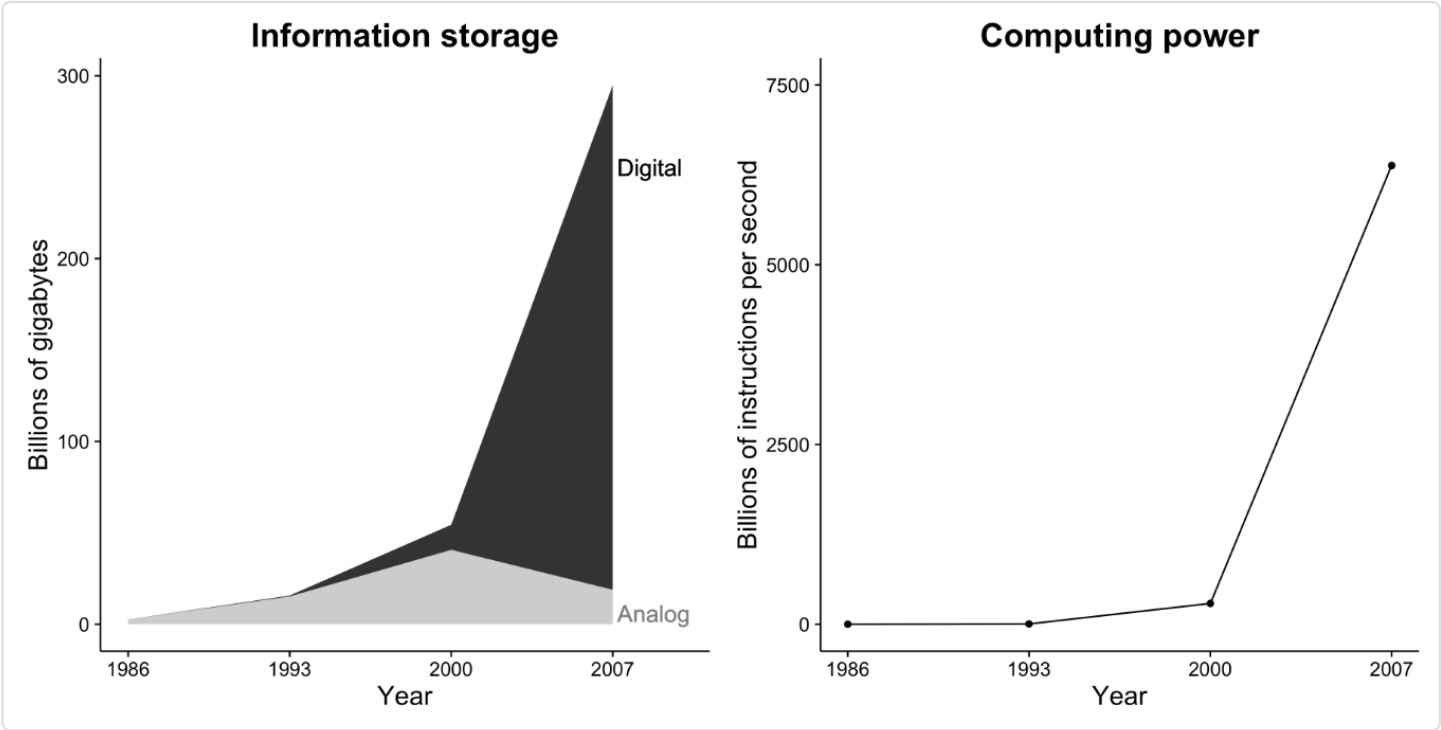
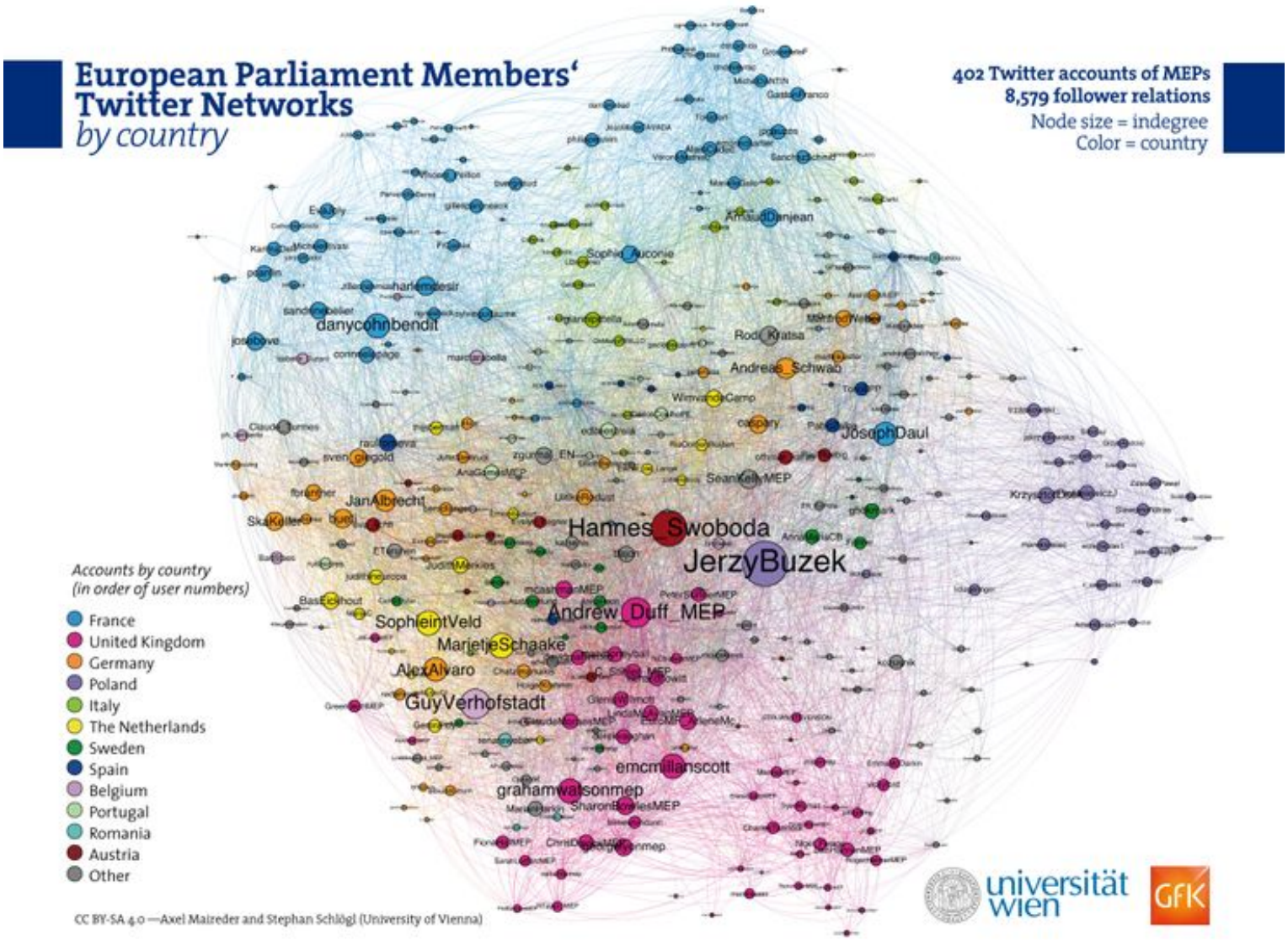


Figure 1.1: Information storage capacity and computing power are increasing dramatically. Further, information storage is now almost exclusively digital (Hilbert and López 2011). These changes create incredible opportunities for social researchers.

New Data, New Possibilities



New Data, New Challenges

- What do we do with millions (or even billions) of rows of data like this?

```
'<!DOCTYPE html>\n<html lang="en">\n<head>\n <meta charset="utf-8"/>\n
<meta http-equiv="Content-Language" content="en"/>\n <meta
name="language" content="en_us"/>\n <meta name="viewport"
content="width=device-width,initial-scale=1"/>\n\n \n <meta
name="description" content="Opinion for People v.  Germany, 674 P.2d
345"/>\n <link rel="author" href="/humans.txt" type="text/plain"/>\n\n
\n <link rel="search"\n type="application/opensearchdescription+xml"\n
title="CourtListener"\n href="/static/xml/opensearch.xml" />\n\n \n
<meta name="application-name" content="CourtListener"/>\n <meta
name="msapplication-tooltip" content="Create alerts, search for and
browse the latest court opinions."/>\n <meta
name="msapplication-starturl"
content="https://www.courtlistener.com"/>\n <meta
name="msapplication-navbutton-color" content="#6683B7"/>\n\n \n <meta
name="twitter:card" content="summary">\n <meta name="twitter:creator"
content="@freelawproject">'
```

"Text Data" is not a new field

- Text data is not a new field – but text data provide an avenue toward answering new questions, or providing new answers to old ones.

Outline

The statistical problem

- We have a corpus, with a set of documents D, say the criminal code, whose features can be represented as a big matrix X.
- We have some outcome variables that depend on this corpus; for example: crime levels Y are a function of the criminal code X and other factors ε:

$$Y = f(X, \varepsilon)$$

- What can we learn about $f(\cdot)$?

The statistical problem

- We have a corpus, with a set of documents D, say the criminal code, whose features can be represented as a big matrix X.
- We have some outcome variables that depend on this corpus; for example: crime levels Y are a function of the criminal code X and other factors ε:

$$Y = f(X, \varepsilon)$$

- What can we learn about $f(\cdot)$?

The statistical problem

- We have a corpus, with a set of documents D, say the criminal code, whose features can be represented as a big matrix X.
- We have some outcome variables that depend on this corpus; for example: crime levels Y are a function of the criminal code X and other factors ε:

$$Y = f(X, \varepsilon)$$

- What can we learn about $f(\cdot)$?

Part 1: Constructing X

- Today and tomorrow we will work on transforming a corpus D into a matrix of features X:
  - First, we need to find and prepare an interesting corpus.
- Featurization:
  - removal of uninformative content, such as capitalization and punctuation
  - frequency counts over words and phrases
  - extraction of syntactic relations (e.g. "defendant is 24 years old")

Part 1: Constructing X

- Today and tomorrow we will work on transforming a corpus D into a matrix of features X:
  - First, we need to find and prepare an interesting corpus.
- Featurization:
  - removal of uninformative content, such as capitalization and punctuation
  - frequency counts over words and phrases
  - extraction of syntactic relations (e.g. "defendant is 24 years old")

Part 2: Understanding X

- The second part reviews methods for understanding X, which is an unwieldy high-dimensional object.
  - Normal descriptive methods for low-dimensional data do not work.
- Unsupervised learning and dimension reduction:
  - topic models
  - word embeddings
  - clustering
  - document similarity

Part 2: Understanding X

- The second part reviews methods for understanding X, which is an unwieldy high-dimensional object.
  - Normal descriptive methods for low-dimensional data do not work.
- Unsupervised learning and dimension reduction:
  - topic models
  - word embeddings
  - clustering
  - document similarity

Part 3: Predicting $f(X)$

- The third part reviews methods for predicting an outcome Y given X, that is, constructing an approximation of $f(X)$.
  - With high-dimensionality and multi-collinearity, normal regression methods do not work.
- Supervised learning:
  - regularized regression
  - random forests
  - cross-validation

Part 3: Predicting $f(X)$

- The third part reviews methods for predicting an outcome Y given X, that is, constructing an approximation of $f(X)$.
  - With high-dimensionality and multi-collinearity, normal regression methods do not work.
- Supervised learning:
  - regularized regression
  - random forests
  - cross-validation

Part 4: Causal estimates for $f(X)$:

- Consider the linear model

$$Y_i = \alpha + X_i'\beta + A_i + \varepsilon_i$$

where X_i and A_i (unobserved) are correlated: $\mathbb{E}(X_i A_i) \neq 0$

  - we have omitted variable bias; least-squares estimates for β are biased.
  - changing a law X will not have the estimated effect β.

- In Part 4 we explore new methods for causal inference in high dimensions:
  - regularized instrumental variables
  - orthogonalized machine learning

Part 4: Causal estimates for $f(X)$:

- Consider the linear model

$$Y_i = \alpha + X_i'\beta + A_i + \varepsilon_i$$

where X_i and A_i (unobserved) are correlated: $\mathbb{E}(X_i A_i) \neq 0$

  - we have omitted variable bias; least-squares estimates for β are biased.
  - changing a law X will not have the estimated effect β.

- In Part 4 we explore new methods for causal inference in high dimensions:
  - regularized instrumental variables
  - orthogonalized machine learning

Part 4: Causal estimates for $f(X)$:

- Consider the linear model

$$Y_i = \alpha + X_i'\beta + A_i + \varepsilon_i$$

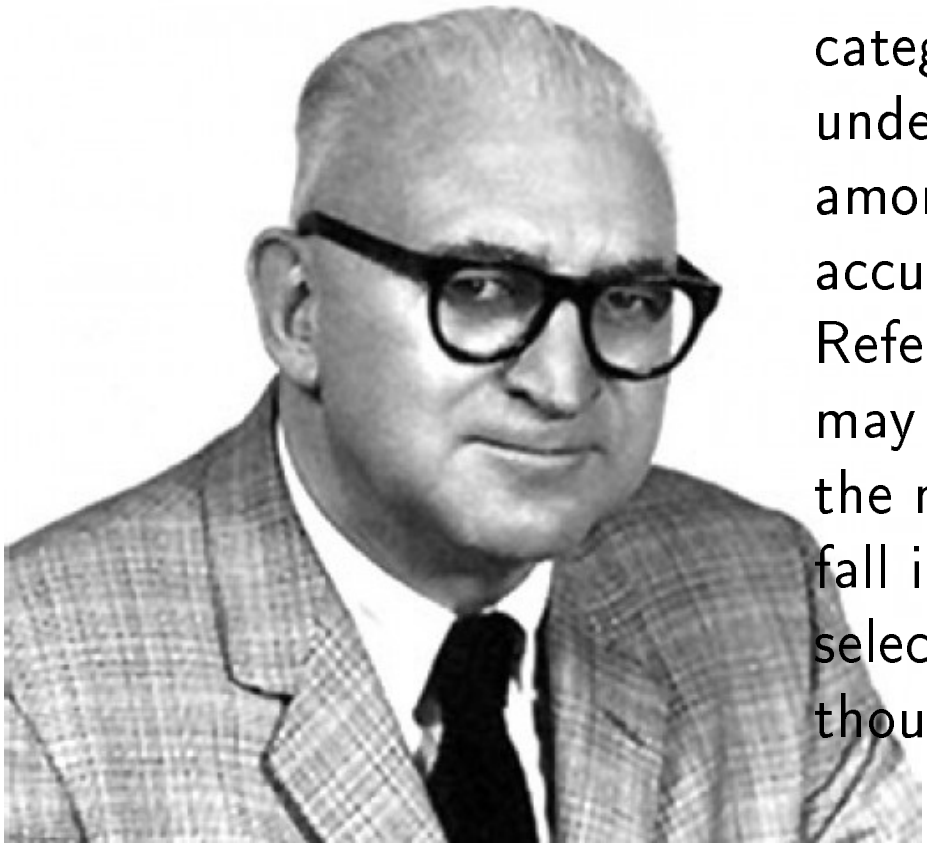
where X_i and A_i (unobserved) are correlated: $\mathbb{E}(X_i A_i) \neq 0$

  - we have omitted variable bias; least-squares estimates for β are biased.
  - changing a law X will not have the estimated effect β.

- In Part 4 we explore new methods for causal inference in high dimensions:
  - regularized instrumental variables
  - orthogonalized machine learning

Outline

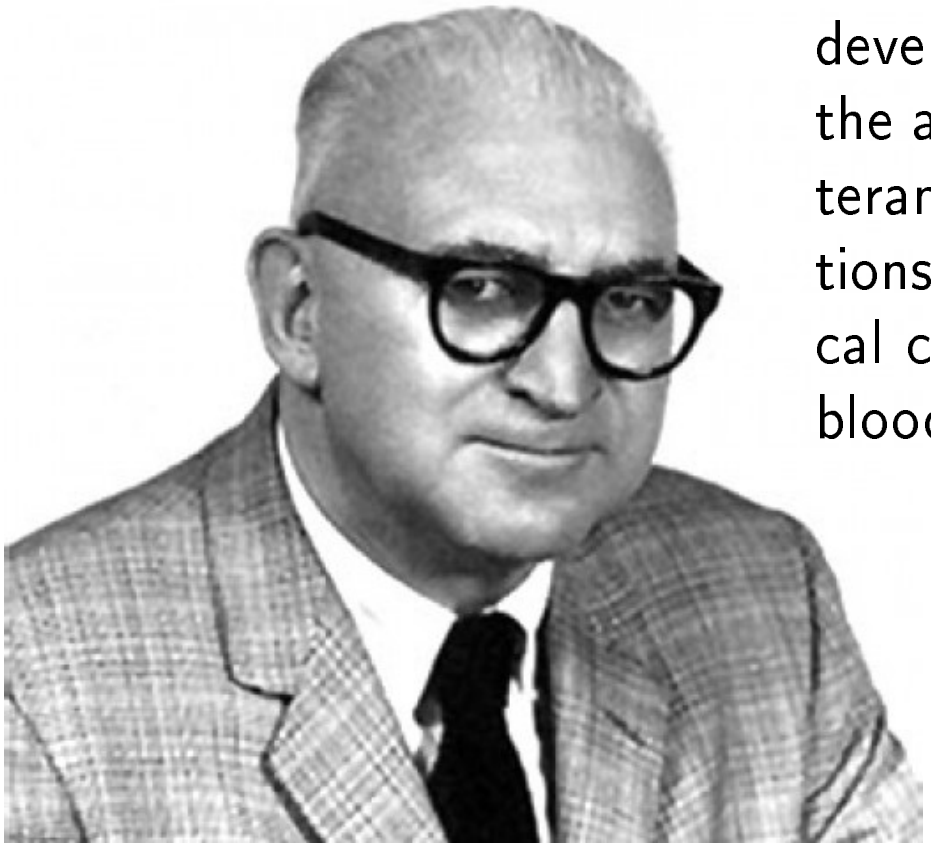
Propagandist Harold Lasswell



"We may classify references into categories according to the understanding which prevails among those who are accustomed to the symbols. References used in interviews may be quantified by counting the number of references which fall into each category during a selected period of time (or per thousand words uttered)."

-Lasswell (1938:198)

Ahead of his time?



In 1935 (age 21) Lasswell was developing methods that tracked the association between word utterances and physiological reactions (e.g. pulse rate, electrical conductivity of the skin, and blood pressure)

Timeline of Quantitative Text Analysis

Time	Activity
1934	Laswell Produces first Key-Word Count
1950	Gottschalk Uses Content Analysis to Track Freudian Themes
1950	Turing Applies AI to text
1952	Bereleson Publishes First Textbook on Content Analysis
1954	First Automatic Translation of Text (Georgetown Experiment)
1963	Msteller and Wallace analyze Federalist Papers

Timeline of Quantitative Text Analysis

Time	Activity
1966	Stone and Bales measure psychometric properties of text at RAND
1980	Machine Learning is Applied to NLP
1981	Weintraub counts parts of speech
1985	Schrodt Introduces Automated Event Coding
1986	Pennebaker develops LIWC
1989	Franzosi brings Quantitative Narrative Analysis to Social Science

Timeline of Quantitative Text Analysis

Time	Activity
1998	First Topic Models Developed
2001	Blei et al. develop LDA
2005	Quin et al use analyze political speeches using topic models
2010	Genztkow Shapiro *Econometrica* paper on media slant
2013	Mikolov et al develop Word2Vec
2016	Chernozhukov et al double-machine-learning paper
2017	Text Analysis Workshop at Max Planck Summer School

Diversification of Text Methods

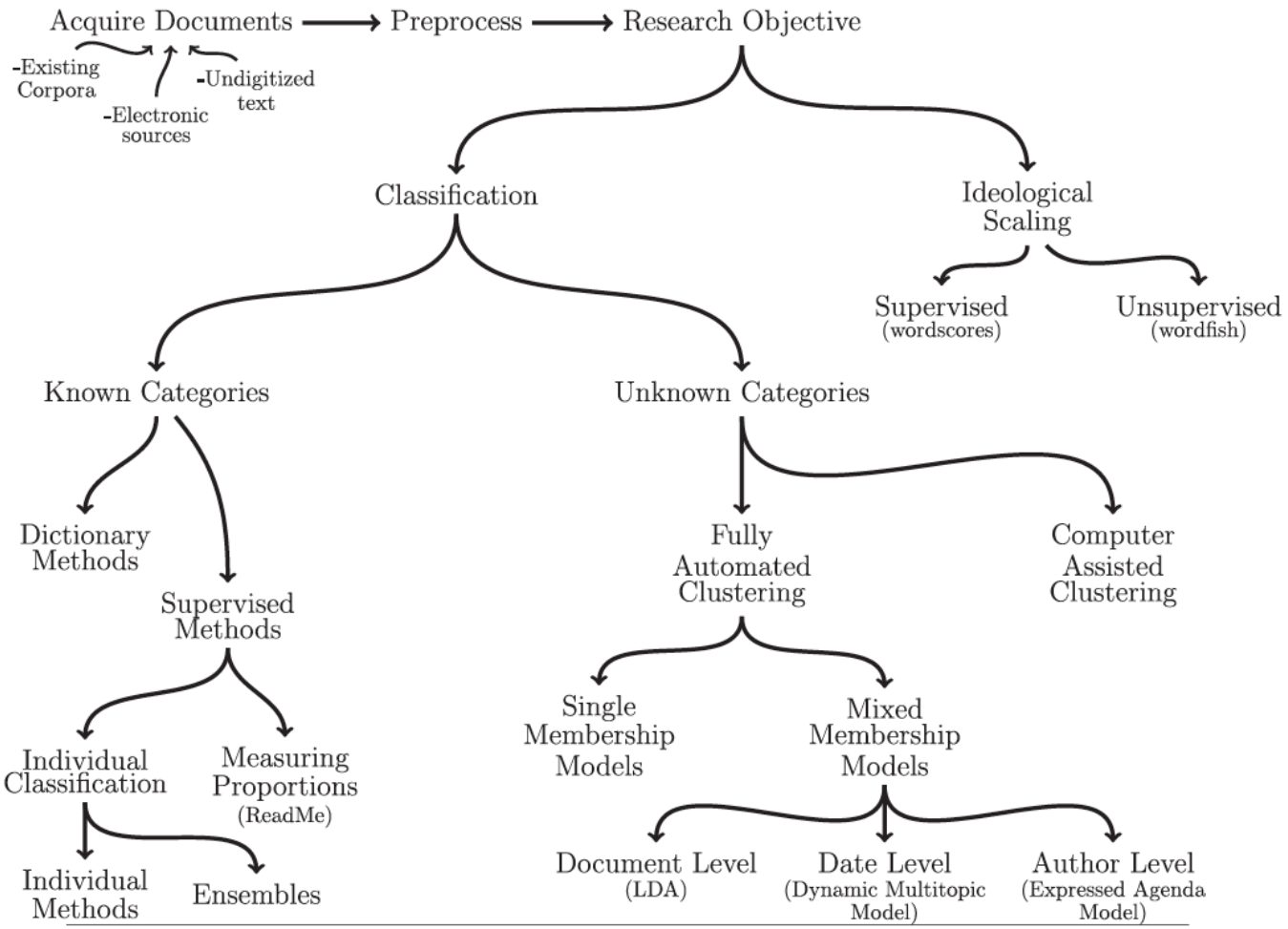


Fig. 1 An overview of text as data methods.

From Stewart and Grimmer (2013).