Introduction

Text Data and Machine Learning for Social Science

Elliott Ash

Bocconi 2018

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

Lecture Times

1. Thursday Sept 6, 10:30am-12:00pm
2. Friday Sept 7, 10:30am-12:00pm
3. Monday Sept 10, 10:30am-12:00pm
4. Tuesday Sept 11, 10:30am-12:00pm
5. Thursday Sept 13, 10:30am-12:00pm
6. Friday Sept 14, 10:30am-12:00pm

Course Web Site

- http://elliottash.com/text_course

Readings

- The material in the slides is based on these materials, but a lot is skipped.
  - It would be reasonable to focus on the slides for study, and refer to the texts based on what is included.
- *Natural Language Processing in Python* (`http://www.nltk.org/book/`)
  - Chapters 1, 2, 3, 5, 7, 8
- *Hands-on Machine Learning with Scikit-learn & TensorFlow* (O'Reilly 2017)
  - Chapters 2, 3, 4, 7, 8 (code and text)
  - Chapters 10, 11, 13, 14, 15 (text, not code)
- *Mastering Metrics* or *Mostly Harmless Econometrics* (Angrist and Pischke)
  - for an applied micro refresher, if needed
- See syllabus for other recommended readings.

Python

- Python is the best programming language for text data and machine learning.
- I recommend Miniconda 3.6.
  - `continuum.io/downloads`
  - See the course web site for download instructions by platform.
  - I ask that the problem sets be submitted as jupyter notebooks.

Problem Sets and Exam

- Problem Sets (24%):
  - six problem sets, four points each
  - asks you to implement the major methods for text analysis on the New Zealand parliament speech corpus (Ash, Morelli, and Osnabruegge 2018).
  - questions might change depending on how much gets covered.
- Exam (26%):
  - based on the material covered in the slides
  - will provide some practice questions beforehand

Term Paper (50%)

- The main course product: empirical paper using text data
  - Can be done individually or in groups of two.
  - In consultation with instructor, form a research design using methods learned in the course.
- Deliverables:
  - 2+ page proposal (due September 25, 5%)
  - 12+ page paper (due early November, 45%)

Office Hours Etc

- I will be available for meetings outside of the lectures.
  - Set up a time by email: `ashe@ethz.ch`.
- We can talk about the course material, your research, anything you want.

The Era of Big Data

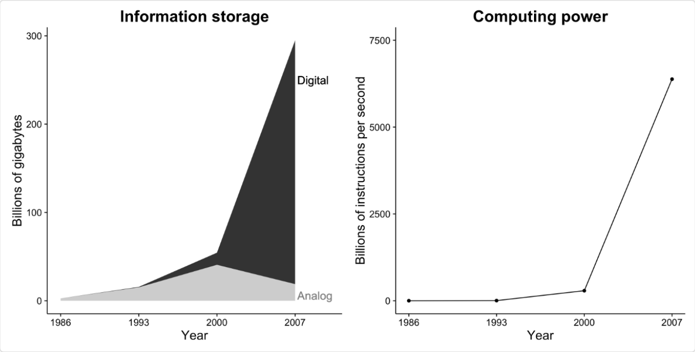
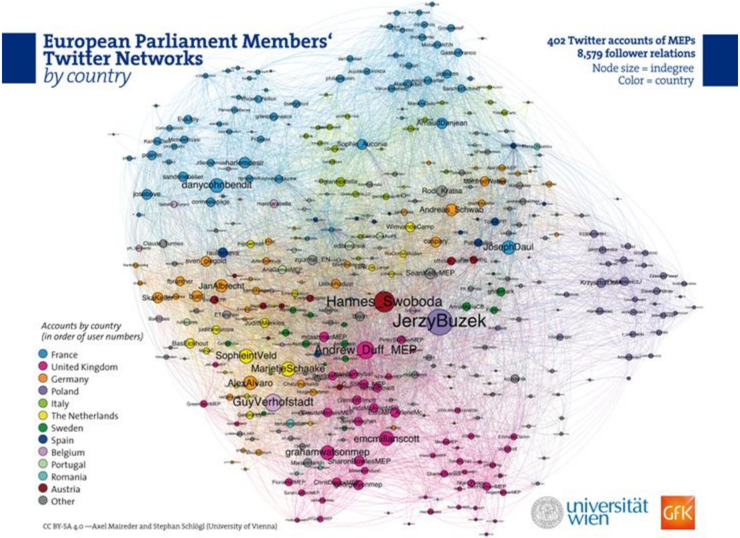


Figure 1.1: Information storage capacity and computing power are increasing dramatically. Further, information storage is now almost exclusively digital (Hilbert and López 2011). These changes create incredible opportunities for social researchers.

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.

The statistical problem

- We have a corpus, with a set of documents D, say the text of political speeches, whose features can be represented as a big matrix X.
- We have some outcome variables that depend on this corpus; for example: voter turnout Y is a function of the speeches X and other factors ϵ:
$$Y = f(X, \epsilon)$$
- What can we learn about $f(\cdot)$?

Constructing X

- First, we will work on transforming a corpus D into a matrix of features X:
  - 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”)

Understanding X

- The second question is how to understand 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

Predicting $f(X)$

- The third task is to predict 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
- In particular, we need to form approximations of $f(\cdot)$ that generalize to held-out data:
  - cross-validation

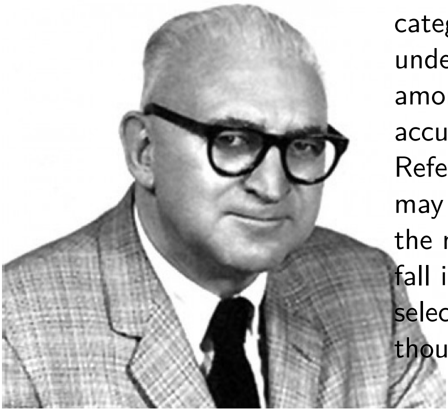
Causal estimates for $f(X)$:

- Consider the linear model

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

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.
  - exogenously changing speech X will not have the estimated effect β.
- On the last day we explore new methods for causal inference in high dimensions:
  - regularized instrumental variables
  - orthogonalized machine learning

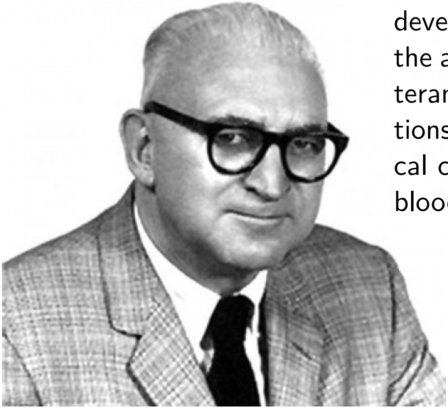
Propagandist Harold Lasswell



"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
2017	Journal of Economic Literature paper on "Text as Data"
2018	Text Analysis Course at Bocconi

Diversification of Text Methods

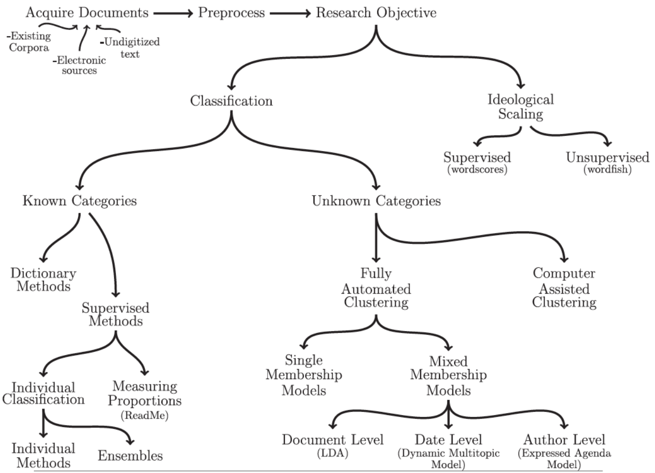


Fig. 1 An overview of text as data methods.

Source: Stewart and Grimmer (2013).

What's next

- Today – introductions to :
  - corpora
  - featurizing texts
  - machine learning
- Tomorrow – more on:
  - corpora
  - features
  - machine learning