Welcome

- This course provides an introduction to social-science research with text data.
- Goals of the course:
  - Learn about techniques to analyze text as data
  - Learn how to apply these techniques in a practical way using the programming language R
  - Allow graduate students to work on a research project that they will, hopefully, be able to use for their dissertations

Readings

- The slides are 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.
  - See syllabus for other recommended readings.

The Era of Big Data

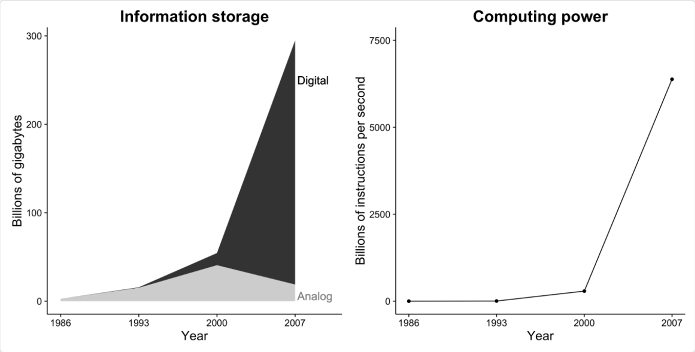


Figure 1.1: Information storage capacity and computing power are increasing dramatically. Further, information storage is now almost exclusively digital (Hilbert and López 2011). These changes create incredible opportunities for social researchers.

Opens up new avenues for research

- In finance, text from financial news, social media, and company filings is used to predict asset price movements and study the causal impact of new information.
- In macroeconomics, text is used to forecast variation in inflation and unemployment, and estimate the effects of policy uncertainty.
- In media economics, text from news and social media is used to study the drivers and effects of political slant.
- In political economy, text from politicians' speeches is used to study the dynamics of political agendas and debate.
- In economic history, used to match census records over time or identify religious identity of regions after Reformation using Universal Short Title Catalogue.

Traditional Econometrics Methods Often Can't Handle These Questions

- Imagine a situation where you need to predict what email messages go to spam or not.
- For simplicity, each message is 30 words long and only uses the most common 1,000 words in the English language.
- The unique representation of a message has dimension 1000^{30}. example
- If you have a sample of emails, the dimension of this sample quickly approaches the number of atoms in the universe.

The Usual Workflow

1. Represent raw text D as a numerical array $\mathbf{C}$; details
2. Map $\mathbf{C}$ to predicted values $\hat{\mathbf{V}}$ of unknown (or "latent") outcomes $\mathbf{V}$; details
3. Use $\hat{\mathbf{V}}$ in subsequent descriptive or causal analysis.

$\rightarrow$ In the spam example, $\mathbf{V}$ is an indicator for whether or not a message is spam or not. In a supervised learning exercise, we may want to train a model on a subset of $\mathbf{C}$ and then test it (or cross-validate it) using the held back data.

$\rightarrow$ Crucially, we're usually not going to get worthwhile causal estimates about the estimated parameters of the model. What we care about is predictive power.

$\rightarrow$ This does not mean we can't use these tools for causal analysis though.

In general, we will...

Get Texts → **Document Term Matrix** → **Operate** → **Inference**

> An expert hospital consultant has written to my hon. Friend...

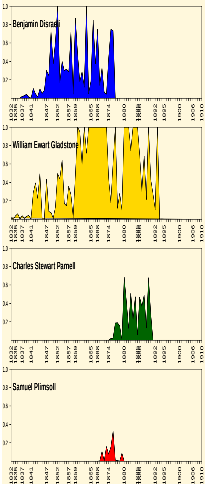
> Order. The Minister must be allowed to reply without interruption.

> I am grateful to my hon. Friend for her question. I pay tribute to her work with the International Myeloma Foundation...

> My constituent, Brian Jago, was fortunate enough to receive a course of Velcade, as a result of which he does not have to...

$$\begin{array}{cc} & \begin{array}{cccc} a & an & \dots & ze \end{array} \\ \begin{array}{c} MP_{001} \\ MP_{002} \\ \\ MP_i \\ \\ MP_{654} \\ MP_{655} \end{array} & \begin{pmatrix} 2 & 0 & \dots & 1 \\ 0 & 3 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 2 \end{pmatrix} \end{array}$$

- (dis)similarity
- diversity
- readability
- scale
- classify
- topic model
- sentiment
- ...



2 words long, 3 possible words (cat, hat, bat)

$3^2 = 9$

(1) (cat, cat)
(2) (cat, hat)
(3) (cat, bat)
(4) (hat, cat)
(5) (hat, hat)
(6) (hat, bat)
(7) (bat, cat)
(8) (bat, hat)
(9) (bat, bat)

Constructing **C**

- First, we will work on transforming a corpus D into a matrix of features $\mathbf{C}$:
  - 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. "nigerian prince", "bank account", "account hacked")

C will often look like a frequency count of words or group of words (tokens) by document (e.g emails)

$$
\begin{array}{cc}
 & \begin{array}{cccc} token_1 & token_2 & \dots & token_n \end{array} \\
\begin{array}{c} email_1 \\ email_2 \\ \\ email_i \\ \\ email_{29} \\ email_{30} \end{array} &
\begin{pmatrix}
2 & 0 & \dots & 1 \\
0 & 3 & \dots & 0 \\
\vdots & \vdots & \dots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \dots & 2
\end{pmatrix}
\end{array}
$$

Understanding **C**

- The second question is how to understand **C**, 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

return

Predicting $\mathbf{V}$

- The third task is to predict an outcome $\hat{\mathbf{V}}$ given $\mathbf{C}$, that is, constructing an approximation of $f(\mathbf{C})$.
  - 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 generalizes to held-out data:
  - cross-validation

return