Text as Data: Introduction

Guest Course – January 2026

Germain Gauthier, Philine Widmer[1]

[1]Bocconi Unversity, Paris School of Economics

USI Lugano

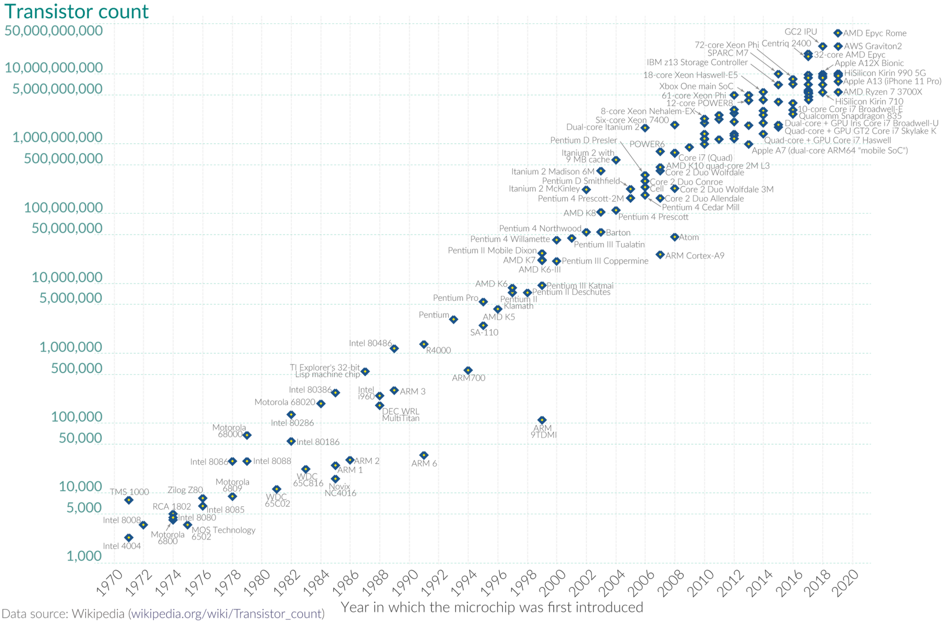
The Rise of Text Data

- The digital era generates considerable amounts of text.
    - Social media and Internet queries
    - Wikipedia, online newspapers, TV transcripts
    - Digitized books, political speeches and manifestos, laws
- It is matched with a similar increase in computational resources.
    - Moore's law = processing power of computers doubles every two years (since the 70s!)

Moore’s Law: The number of transistors on microchips doubles every two years

Our World in Data

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.



Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)

OurWorldinData.org – Research and data to make progress against the world’s largest problems.

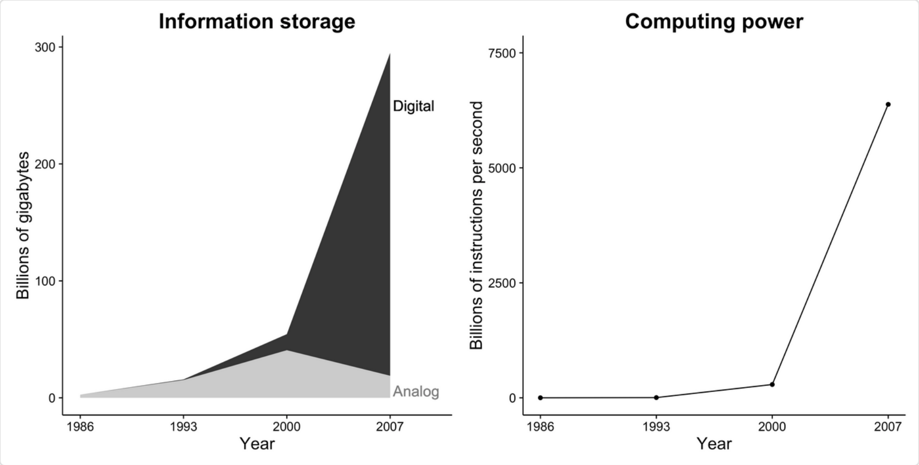


Figure 1.1: Information storage capacity and computing power are increasing dramatically. Further, information storage is now almost exclusively digital. These changes create incredible opportunities for social researchers. Adapted from Hilbert and López (2011), figures 2 and 5.

Source: Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.

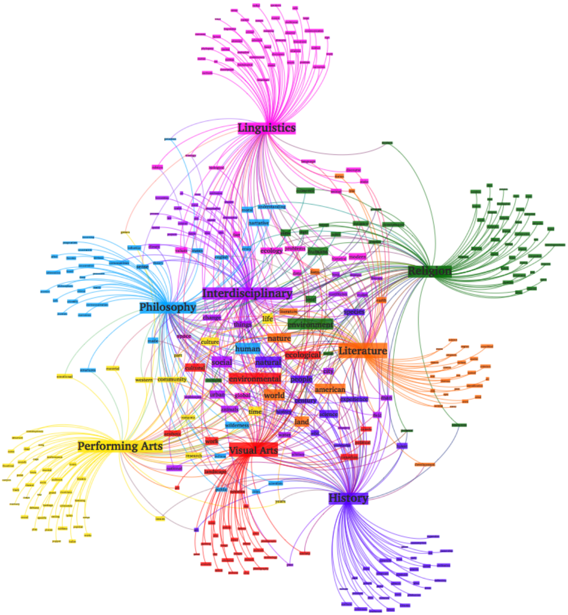
Natural Language Processing

- Natural language processing is a *data-driven* approach to the analysis of text documents.
- Applications in your everyday life:
  - Search engines, translation services, spam detection
  - ChatGPT, Claude, Gemini
- Applications in social science:
  - Measuring economic policy uncertainty, news sentiment, racial and misogynistic bias, political and economic narratives, speech polarization
  - Predicting protests, conflicts, GDP growth, financial market fluctuations

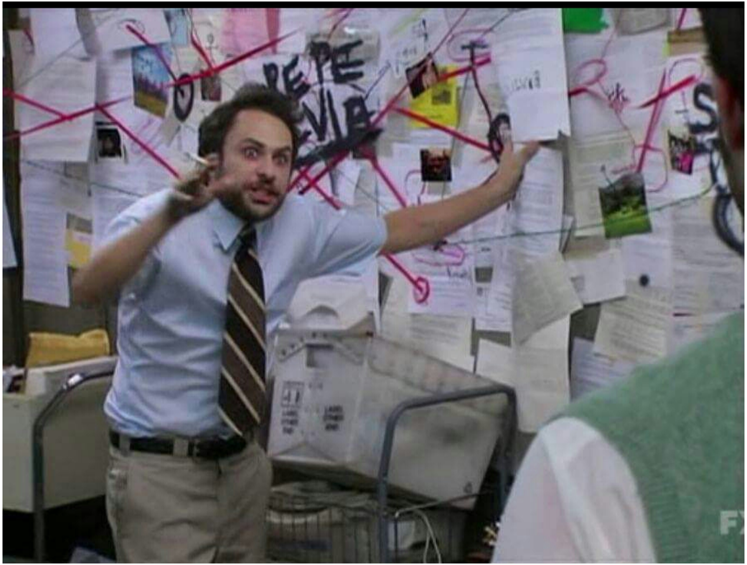
This course

- Focus on natural language processing in applied economic research
- Contents: dictionaries, text regressions, topic models, embeddings, semantic parsing, sequence models
- Why is this useful for economic research?
  - Measure economic/political/social concepts in texts and use them as regressors or outcomes (e.g., economic policy uncertainty, news sentiment, racial and misogynistic bias, political and economic narratives, speech polarization)
  - Predict some outcomes from texts (e.g., protests, conflicts, GDP growth, financial market fluctuations)

We want to do cool graphs like the one below...



But for this you have to bear with me...



Organization

- Course Days
  - January 12, 09:00-12:00: Introduction and basic NLP
  - January 13, 09:00-12:00: Topics + Embeddings Models
  - January 14, 09:00-12:00: Semantic Parsers + Deep Latent Factor Models
  - January 15, 09:00-12:00: Large Language Models (LLMs)
  - January 16, 09:00-12:00: Presentations and Feedback
- Communication: philine.widmer@gmail.com, germain.jean.gauthier@gmail.com
- Examination parts
  - Research proposal (due January 15 at 3 PM)
  - Accompanying slides (due January 16 at 9 AM)
  - Discussion of another person's proposal (due January 16 at 9 AM)

Organization

- Bibliography: see course fact sheet
- Programming language
    - Python 3
    - We will be using Jupyter notebooks; you are free to use any editor
    - Please make sure to have Python up and running (or access to Google Colab)
    - Python 3 is the “lingua franca” of NLP

Modeling Text

- Raw text is typically **unstructured**.
- The information we are after is mixed in with irrelevant information.
- To put some structure on the text, we need a **a statistical model**.
- All models simplify and throw some information away (i.e., make assumptions).
- But good models retain essential information.

The Curse of High-dimensionality

- Raw text is typically **high-dimensional**.
- Suppose each document is composed of w words drawn from a vocabulary of p possible words. Then the unique representation of these documents has dimension p^w.
- This quickly leads to absurdly large numbers.
- We will run into **high-dimensional statistics**.

  e.g., dimension reduction, feature selection, etc.

The Typical Research Pipeline[1]

- Let a corpus $\mathcal{C}$ be a collection of text documents.
- A typical research pipeline involves:

1. **A Featurization Approach**

   We wish to represent $\mathcal{C}$ as numerical array W.

2. **A Measurement Approach**

   A mapping f from features W to outcomes of interest Y.

3. **Some Insights (optional)**

   A causal or descriptive analysis of $\hat{Y}$.

1. Featurization

- First, we need a mathematical representation of the text.
- We wish to represent the corpus $\mathcal{C}$ as a numerical array W.
- The most common (and simple) representation of text is **word frequencies**.
- But we will also see richer (and more demanding) representations in the second part of the course.

2. Measurement

- We have a numerical representation W of our corpus $\mathcal{C}$.
- Next, we need a mapping f from features W to outcomes Y.
- Two common families in machine learning.

1. **Supervised learning**

   We learn f based on *labeled* data.

   e.g., Predict the GDP from newspapers text.

2. **Unsupervised learning**

   We learn f based on *unlabeled* data.

   e.g., Uncovering latent trending topics in news.

3. Insights

- We have a prediction of the outcome: $\hat{Y} = f(W)$.
- Next, we can use $\hat{Y}$ for standard statistical analysis.
- Two common approaches in the social sciences.

1. **Descriptive**

   We study $\hat{Y}$ with no attempt to make causal claims.

   e.g., What narratives do Democrats emphasize in the U.S. Congress?

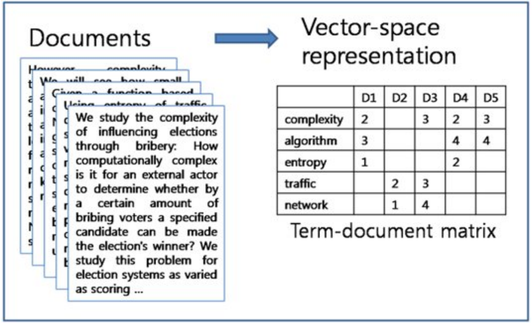
2. **Causal**

   For a set of observed variables X, we attempt to understand whether $\hat{Y}$ *causes* or *is the result of* X.

   e.g., After attending an economics training program, are judges more likely to use economic jargon in their rulings?

Counting Words

- The simplest way to represent text documents is word frequencies.
- This is referred to as the **bag-of-words** approach.
- The corpus is featurized as a term-document matrix **W**:



We study the complexity of influencing elections through bribery: How computationally complex is it for an external actor to determine whether by a certain amount of bribing voters a specified candidate can be made the election's winner? We study this problem for election systems as varied as scoring ...

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Some Refinements: Text "cleaning"

- The general unit for counts is called a **token**.
- The final set of tokens considered is the **vocabulary**.
- Tokens can also include symbols and digits.

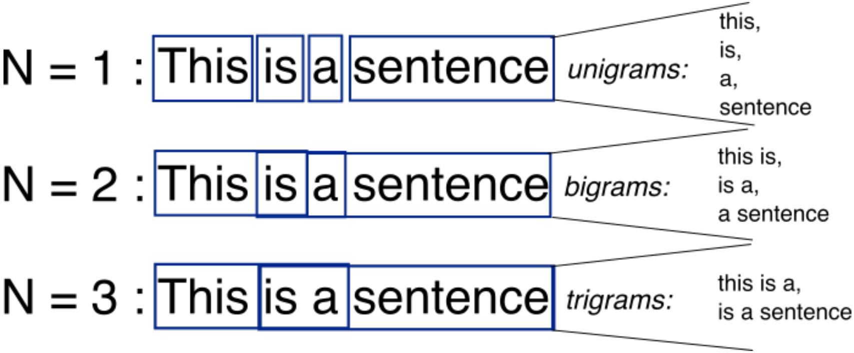
  e.g., #, !, ?, haha, 2008, etc.
- Depending on the application, some tokens are uninformative and can be removed (i.e., stopwords).

  e.g., she, he, the, a, etc.
- Some tokens mean the same thing and can be grouped together (via stemming or lemmatization).

  e.g., animal and animals, eating and eat, etc.

Some Refinements: N-grams

- The bag-of-words approach can be generalized to arbitrarily large token sequences called **n-grams**.



Some Refinements: TF-IDF

- Raw counts in the term-document matrix W tend to be dominated by very frequent words.
- TF–IDF rescales counts to emphasize words that are frequent in a document but rare in the corpus.
- For term t in document d:

$$\mathrm{tf}_{d,t} = \mathrm{count}(t \text{ in } d), \qquad \mathrm{df}_t = \#\{d : t \in d\}, \qquad \mathrm{idf}_t = \log\left(\frac{N+1}{\mathrm{df}_t + 1}\right) + 1$$

$$\mathrm{tfidf}_{d,t} = \mathrm{tf}_{d,t} \cdot \mathrm{idf}_t.$$

- Intuition: common words (high df_t) get downweighted; distinctive words get upweighted.

So what can we do with W?

References I

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019). “Text as data”. In: *Journal of Economic Literature* 57.3, pp. 535–574.