Text-Based Methods for Studying Innovation, Science and Technology

Universidad Autónoma de Barcelona

Joseph Emmens

An introduction to text-as-data tools in economics and how they can be used to study the organisation of innovation and science

Text is Everywhere

“By 2025, IDC estimates there will be 175 zettabytes of data globally (that's 175 with 21 zeros), with 80% of that data being unstructured. Ninety percent of unstructured data is never analyzed.”

[Forbes]

The New York Times

NEW YORK, TUESDAY, JANUARY 23, 2018

SHUTDOWN EN
SETTING UP CL
OVER 'DREAM

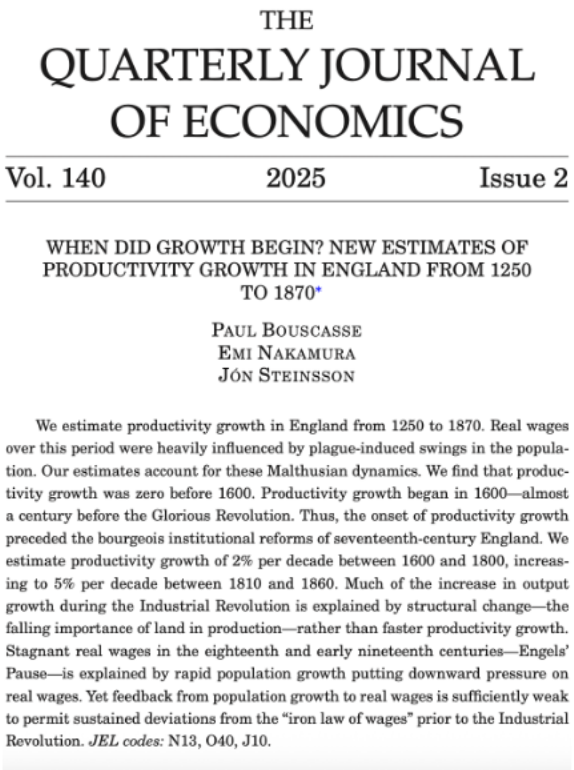
NEWS ANALYSIS

Hope for Bargain in a Swift Surrender

House members leaving the Capitol on Monday, top. Senator Mitch McConnell, right, promised Senator Chuck Schumer, left, and fellow Democrats that he would allow an immigration vote.

Tax-Law Bonuses Veil a Bonanza for Companies

Newspaper Articles



THE
QUARTERLY JOURNAL
OF ECONOMICS

Vol. 140 2025 Issue 2

WHEN DID GROWTH BEGIN? NEW ESTIMATES OF PRODUCTIVITY GROWTH IN ENGLAND FROM 1250 TO 1870*

PAUL BOUSCASSE
EMI NAKAMURA
JÓN STEINSSON

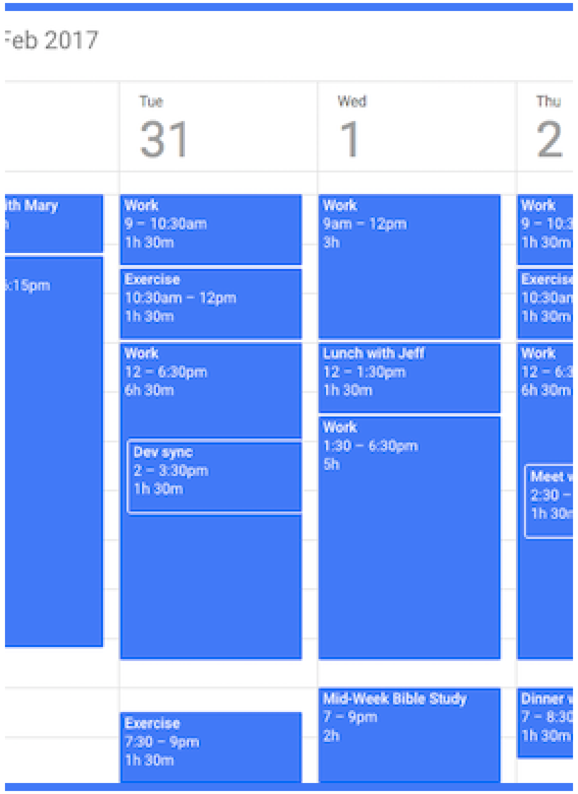
We estimate productivity growth in England from 1250 to 1870. Real wages over this period were heavily influenced by plague-induced swings in the population. Our estimates account for these Malthusian dynamics. We find that productivity growth was zero before 1600. Productivity growth began in 1600—almost a century before the Glorious Revolution. Thus, the onset of productivity growth preceded the bourgeois institutional reforms of seventeenth-century England. We estimate productivity growth of 2% per decade between 1600 and 1800, increasing to 5% per decade between 1810 and 1860. Much of the increase in output growth during the Industrial Revolution is explained by structural change—the falling importance of land in production—rather than faster productivity growth. Stagnant real wages in the eighteenth and early nineteenth centuries—Engels' Pause—is explained by rapid population growth putting downward pressure on real wages. Yet feedback from population growth to real wages is sufficiently weak to permit sustained deviations from the “iron law of wages” prior to the Industrial Revolution. *JEL codes:* N13, O40, J10.

Academic Papers

Text is Everywhere

“By 2025, IDC estimates there will be 175 zettabytes of data globally (that's 175 with 21 zeros), with 80% of that data being unstructured. Ninety percent of unstructured data is never analyzed.”

[Forbes]



Calander Events

Job Adverts

What Can Text Reveal?

01. Beliefs

Text from inspection reports can reveal bureaucrats' beliefs about compliance risks.

02. Tone

Managerial emails with urgent and directive language point to a hierarchical decision-making culture.

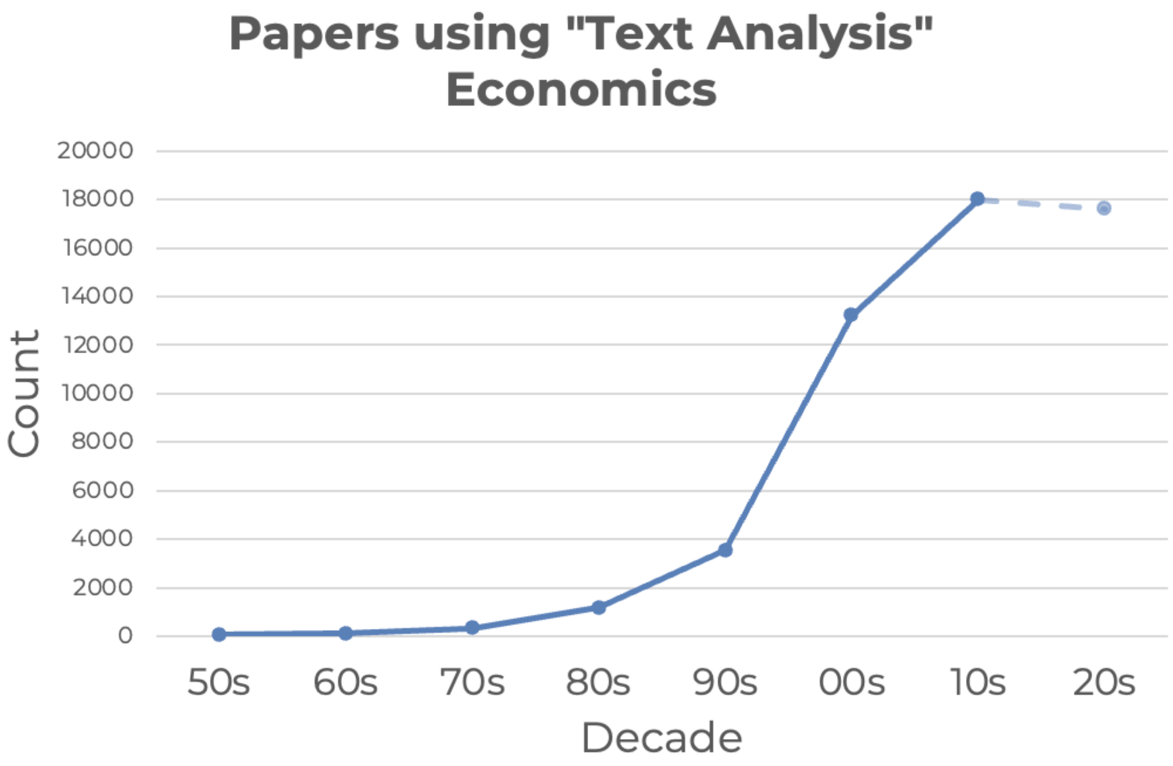
03. Strategy

Memos stressing "trust in teams" over "individual incentives" reveal a firm's organisational philosophy.

04. Knowledge

A patent, scientific paper or product description that is the first to combine two words indicates a novel creation.

Text Analysis is Growing in Economics



Models are getting bigger and faster

Huge resources are being put into developing computer science text-based models–which spill over into social sciences!

Models are open source and well supported

Huge communities of online programmers support the development of *python*, *C*, *R*, *Julia and even Stata* packages for text analysis!

Table of Contents

4 Essential Steps to Using Text as Data[1] (with example[2])

A. Clean raw text with pre-processing.

B. Represent text numerically.

C. Extract or map numerical form to values of interest.

D. Use these extracted values in empirical models.

1: Gentzkow, Kelly & Taddy, JEL Review (2019)

A) Clean Raw Text with Preprocessing.

- Pre-processing is crucial to extract structure from language.

What to remove?

- Standard to remove:
  - Punctionation
  - Numbers
  - Stop words
- But beware–one person's garbage is another person's treasure.

Stemming & lemmatization

E.g: managing, managed, manager

Stemming:	**Lemmatization:**
All become → "manag"	→ "manage","manage", "manager"
Crude chopping. Fast but loses meaning.	Slower, but preserves meaning and context.

B) Four Methods to Represent Text Numerically

01. Dictionary Methods

Assign meaning to text by counting the occurrence of predefined words associated with specific categories.

02. Bag-of-Words (TF-IDF)

A vector tracking which words appear in each document (TF-IDF: weighted by their relative frequency across all documents).

03. Topic Modelling

Identify hidden themes in text by modeling each document as a mixture of topics and each topic as a distribution over words.

04. Embedding Models

Represent words or documents as dense vectors in a high-dimensional space, where semantic similarity is reflected in similarity measures on these vectors.

Dictionary Methods

Assign meaning to text by counting the occurrence of predefined words associated with specific categories.

Azoulay, Graff Zivin, & Wang (2010)

- Extract MeSH terms — a curated dictionary of biomedical concepts from the National Library of Medicine
- For each pair of scientists, collect the MeSH terms used in their publications.
- Compute their proximity as roughly the proportion of overlapping MeSH terms.

Methodology

- Pre-defined word lists: Use curated dictionaries of words associated with categories (e.g. sentiment, technology) to count term frequency.

- Ignores grammar, word order, and context — focuses solely on whether and how often a term appears.

- Loop over each text and count the occurances of each target word– sometimes normalise by group / time counts.

Dictionary Methods

Assign meaning to text by counting occurrences of predefined words associated with specific categories.

Vol. 18, 1997, Special Issue: Computers Can Read as Well as Count: Computer-Aided Text Analysis in Organizational Research

Journal of Organizational Behavior

Published by: Wiley

Strength in simplicity

- They are fast and easy to run on local computers.
- Results are transparent to non-technical audiences.

No estimation needed

- Count the frequency of pre-defined words or phrases.

Intuitive outcomes

- Great at measuring exposure.
- Intuitively measure concepts like sentiment, or risk.

Long-standing methodology

- The JOB had a special issue in 1997 using many dictionary based methods!
- Simple but not outdated.

Bag-of-Words (TF-IDF)

A vector tracking which words appear in each document (TF-IDF: weighted by their relative frequency across all documents).

Kelly, Papanikolaou, Seru, Taddy (2021)

- Represent each patent in TF-IDF form.
- Use pairwise forward and backward patent similarity to measure quality (identify **breakthrough patents**)
- 9 million patents and 1,685,416 terms. Took 4 weeks to estimate on 60 servers; each server had 128 GB RAM & 64 cores.

Methodology

- Build a database (corpus) specific vocabulary – a list of all unique words.
- Vectorise the documents – Count how often each word appears in each document and store the result as a vector (e.g. in a document-term matrix).
- TF-IDF up-weights each word by how often it appears in a document (TF) and down-weights it based on how common it is across all documents (IDF).

Bag-of-Words (TF-IDF)

A vector tracking which words appear in each document (TF-IDF: weighted by their relative importance across all documents).

Simple and Straightforward

- It is simple and fast to implement.
- Common baseline method in natural language processing tasks.

Computationally Low-cost

- Word count vectors can be stored as sparse vectors.
- Doesn't require as much RAM.

Ignores Word/ Sentence Context

- BoW ignores context, semantics, and word order.
- This leads to worse performance on similarity.

Effective for Simple Tasks

- Performs well on straightforward tasks.
- For example, identifying topics of speeches.

Topic Modelling (LDA)

Identify hidden themes in text by modeling each document as a mixture of topics and each topic as a distribution over words.

Bandiera, Hansen, Prat & Sadun (2020):

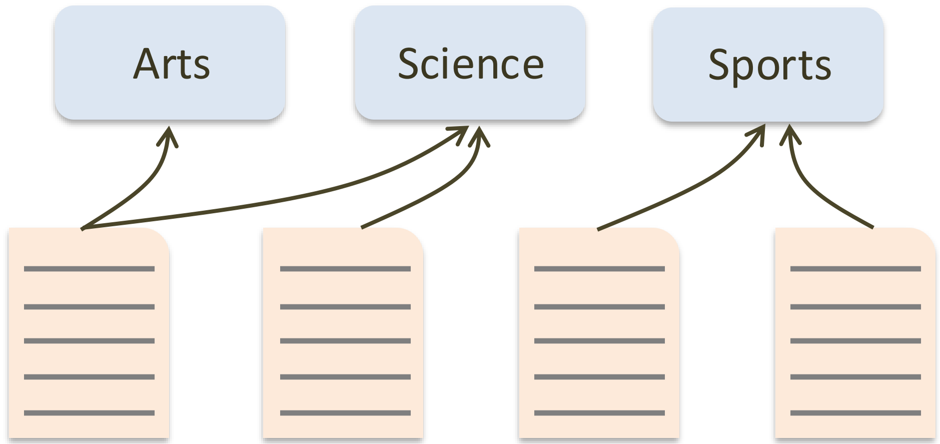
- Map manager calendar activities into two types – ***Manager*** and ***Leader***.
- Build a convex-combination of each types to order CEOs along a 1-D index.
- Use this measure to study the correlation between CEO behaviour and firm performance.

Methodology

- Each document is modelled as a probability distribution over topics, and each topic as a distribution over words.
- Estimate two things:
  - For each document, how much it talks about each topic.
  - For each topic, which words are most likely.
- Use an inference algorithm to repeatedly guess topic assignments for words until the guesses stabilise.

Topic Modelling (LDA)

Identify hidden thematic structures in documents by modeling each document as a mixture of topics and each topic as a distribution over words.



Intuitive Dimension Reduction

- Each vector representation of a text are human interpretable.
- This provides easier ways of validating your output!

Computationally Intermediate

- Require more than dictionary methods, but significantly less than embeddings!

Theoretically Flexible

- Thanks to their simplicity, you can link the generative process to many models of social behaviour.

Large Number of Variations

There are many variations:

- Correlated
- Dynamic
- Author
- Structural
- BERTopic

Embedding Models

Represent words or documents as dense vectors in a high-dimensional space, where semantic similarity is reflected in geometric proximity.

Chaturvedi, Mahajan, Siddique (2023)

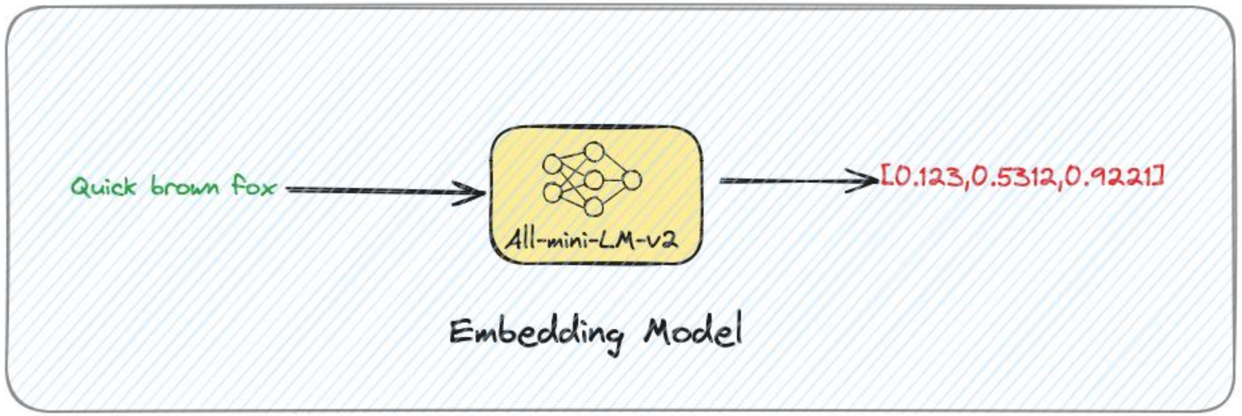
- Use data on 332,000 job ads to back out and explain **skill demands**.
- Represent job ads by embedding vector and cluster into groups.
- Estimate mincer regressions to correlate log wages and skill demands.

Methodology

- Create word representations that change depending on sentence context.
- Use a *transformer* model to:
  - Predict missing words in a sentence.
  - Predict if one sentence follows another.
- Optimise model weights by minimising prediction errors across huge text datasets.

Embedding Models

Represent words or documents as dense vectors in a high-dimensional space, where semantic similarity is reflected in geometric proximity.



Very precise similarities

- State of the art.
- Move beyond Bag-of-Words and allow for word and sentence context.
- *Consider 'bank'* in different sentences.

Rich but Opaque

- Precision increases, however, the dimensions are not human interpretable.
- BERTopic model is one solution!

Computationally intensive

- Training your own model gives you greater control over domain specific issues.
- But they require a lot of data/power!

Access pre-trained models

- Hugging Face has thousands of pre-trained and easy to implement models!

When to Use What?

Dimension	Bag-of-Words (LDA)	Embeddings
Interpretability	**High** — each word/topic is visible and labelled	**Low** — vector dimensions lack clear meaning
Complexity	**Low** — simple models, fast to compute	**High** — pretrained models, larger computation required
Context awareness	**None** — treats each word independently	**High** — captures meaning from surrounding words
Semantic similarity	**Poor** — based on co-occurrence	**Strong** — captures nuanced meanings and relationships

C) Mapping to Values of Interest

- Once you have represented the data numerically, normally you want to add more structure and interpretation.

1

Find the **distance between vectors**

- BoW, topics or embeddings
- Cosine similarity or Euclidean
- Measure novelty or polarisation

2

Topic proportions per document

- Taken from the topic distribution
- % of time talking about war, exports versus imports, uncertainty etc.

3

Measure tone or sentiment

- Use pre-defined lists or sentiment classification models.
- Give each manager a score on their anger, directness, empathy etc.

D) Using Values in Empirical Models

- Once you have extracted some value of interest we want to either **explain it, or explain something with it**!

1

Use as **X or Y in a reduced form empirical model.**

- Regression models
- Map topic share → outcome.
- E.g. the % of local news on the "conflict" topic to explain yield rates.

2

Classification or prediction

- Growing use in economics across fields.
- E.g. Conflict Forecast–Mueller & Rauh (2022)

3

Structural Modelling

- Give structural interpretation to model parameters!
- Gentzkow, Shapiro & Taddy (2019)

Application

Innovation, Science and Technology

Teams & Text

- How do collaboration and contribution dynamics change over an inventor's lifecycle?
- I demonstrate a quality-quantity trade off as inventor's become senior.

Research Fields

- A study into the rise and fall of research fields
- What is the role of public versus private financing in sparking new fields.

Topic Model

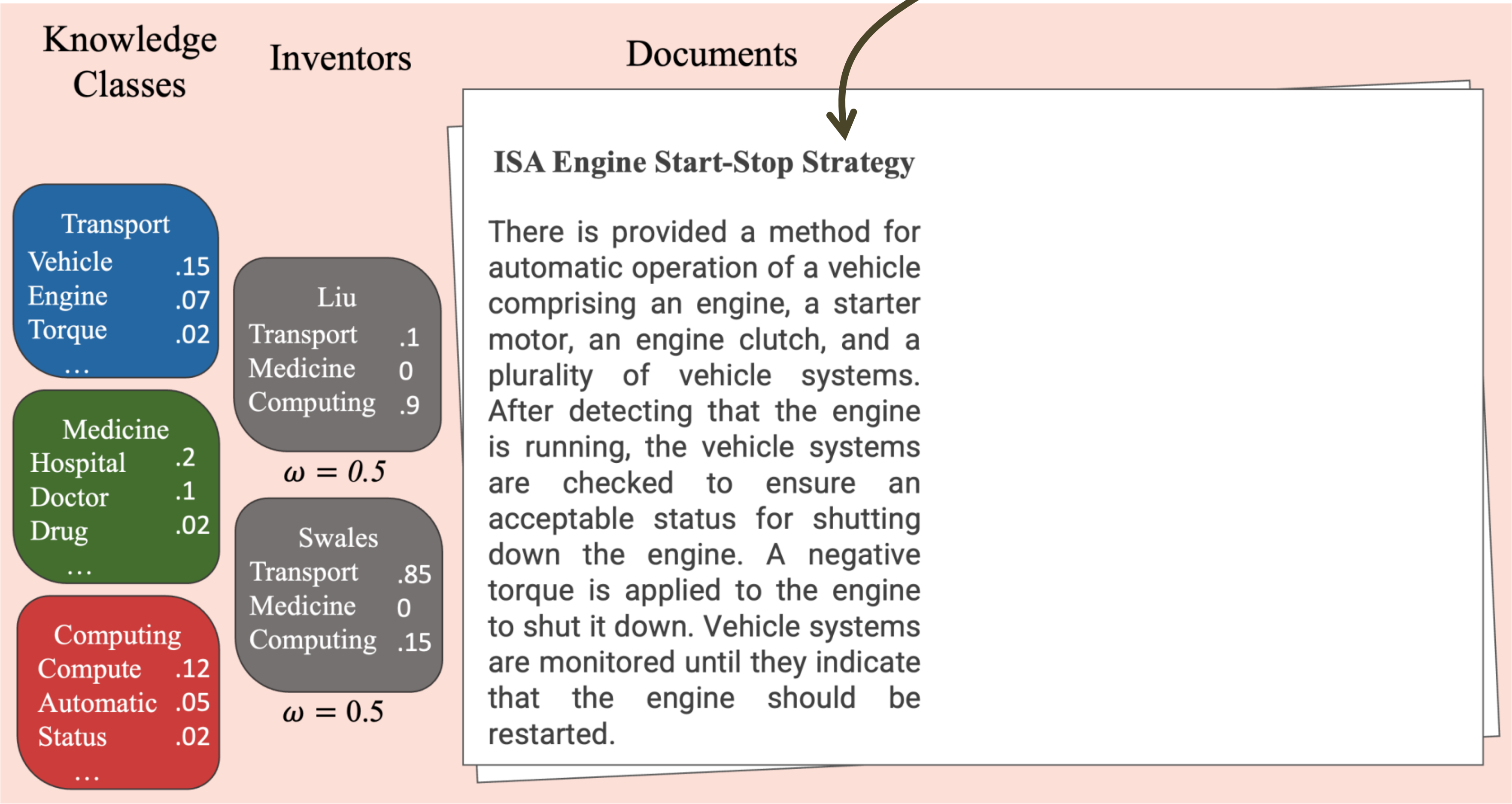
- Uses an Author-Topic model to disentangle contributions.

Embeddings

- Uses a pre-trained embedding model for scientific texts.

Teams and Text: Back out a contribution share

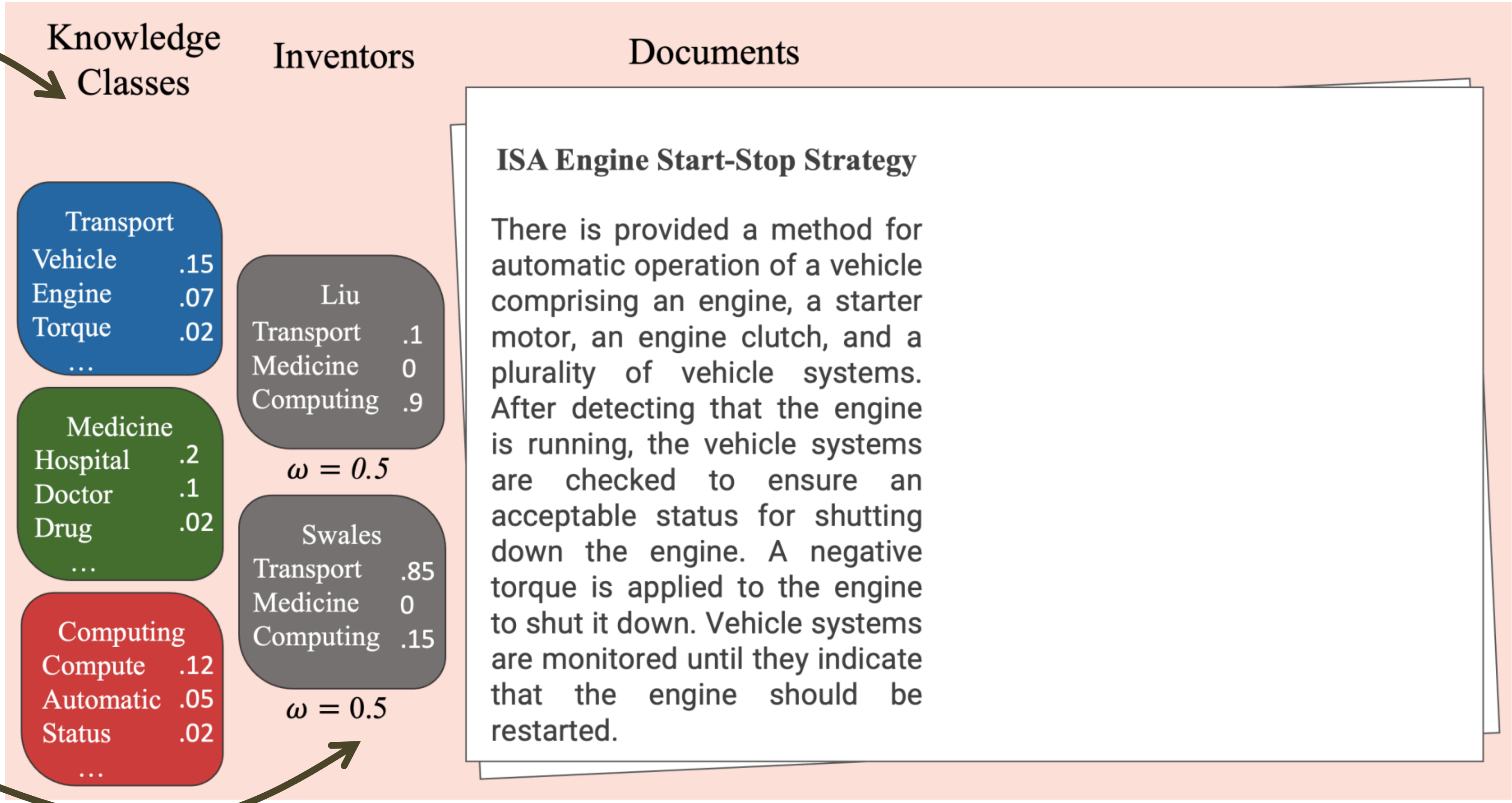
Start with a set of patent documents.



Teams and Text:
Back out a contribution share

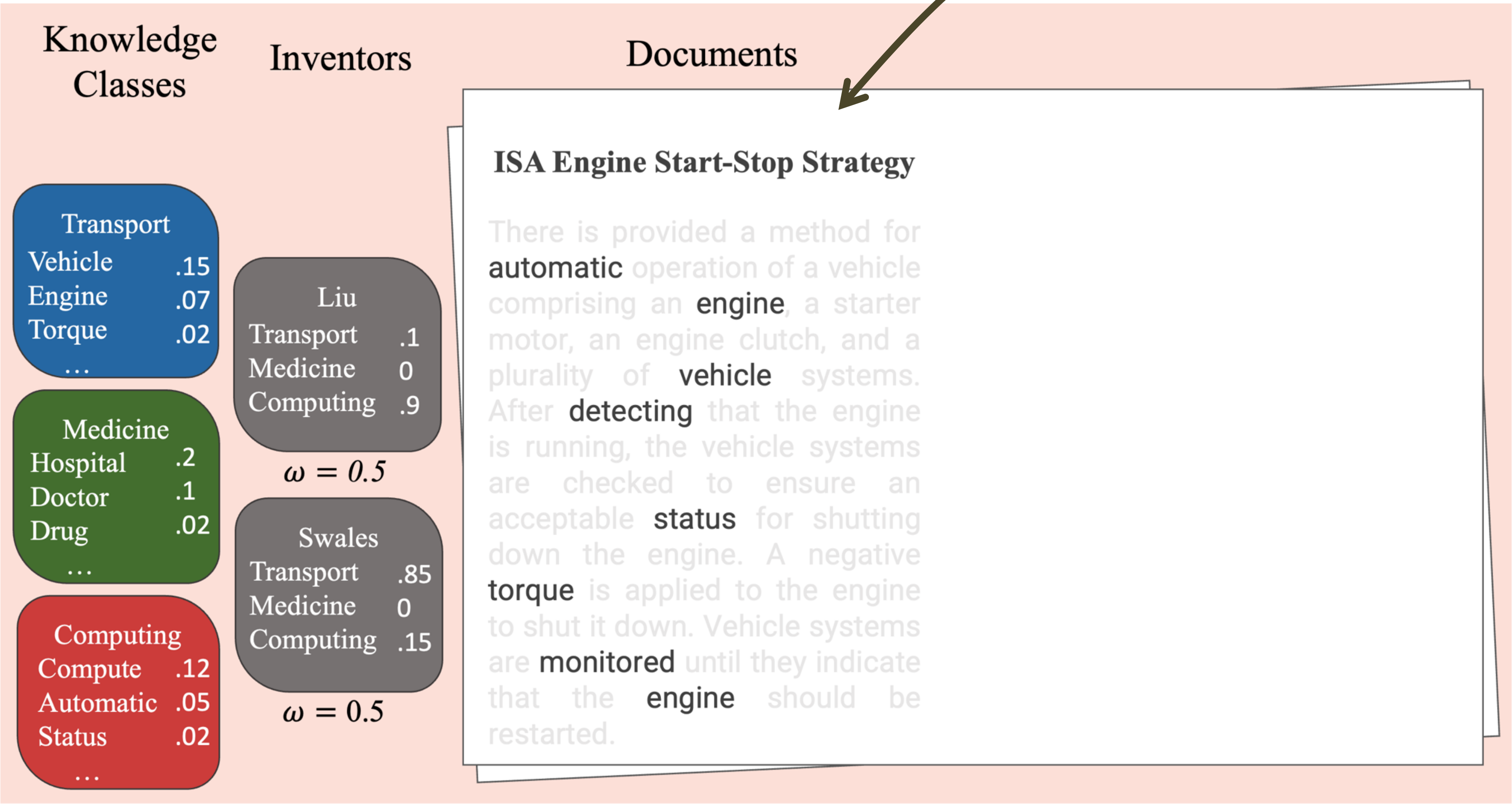
Assume all patents can be represented as a combination of *K* knowledge classes.

Assume that all inventors have a distribution across these knowledge classes.



Teams and Text: Back out a contribution share

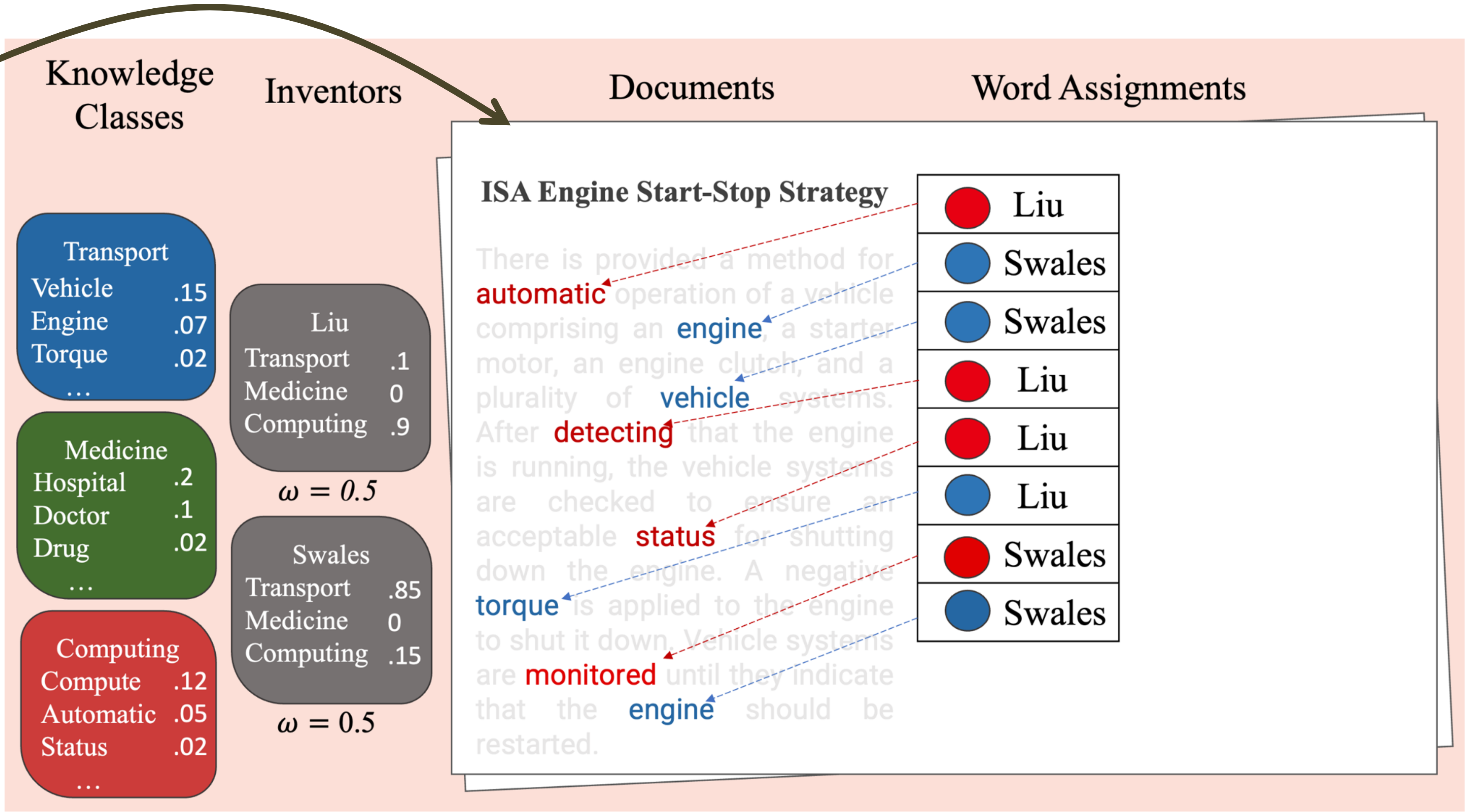
First pre-process the text to keep informative words



Teams and Text:
Back out a contribution share

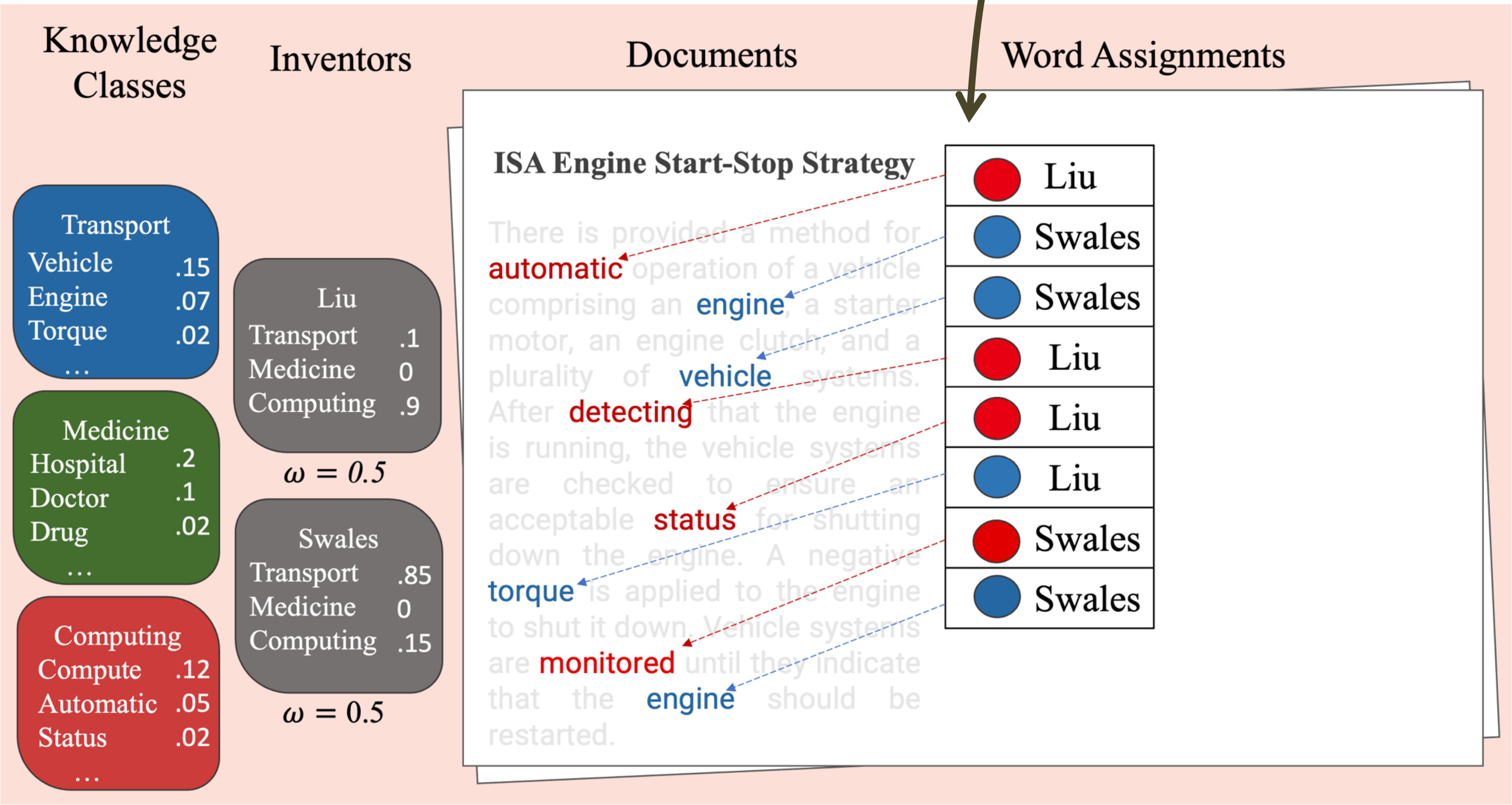
Iterate over every word in every document:

- Choose an (inventor x knowledge class) pair to **maximise the likelihood of the data**



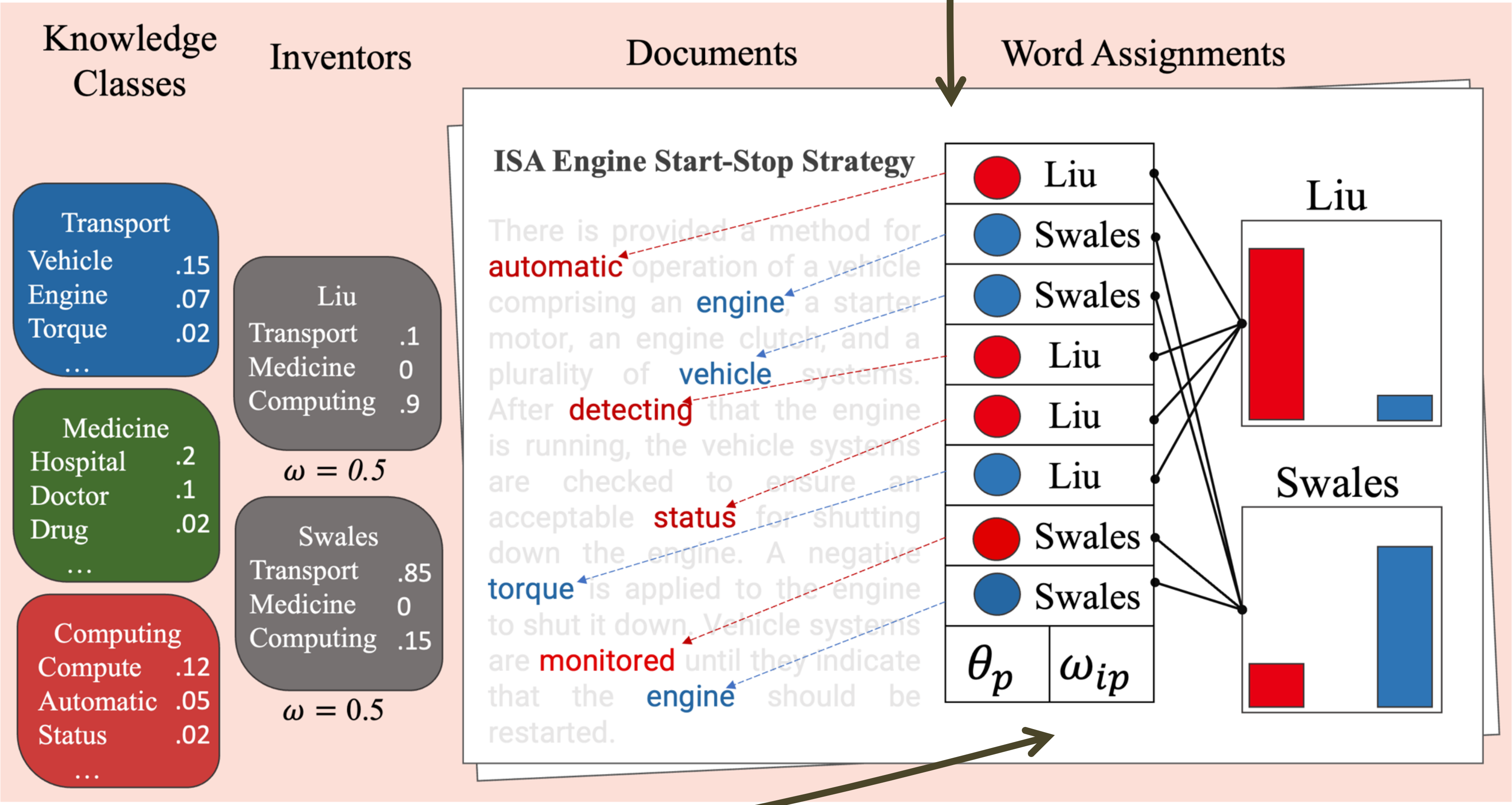
Teams and Text: Back out a contribution share

This choice is informed by an inventor's full history of patenting.



Teams and Text: Back out a contribution share

This **matching of inventors and classes to words** jointly backs out a set of latent parameters



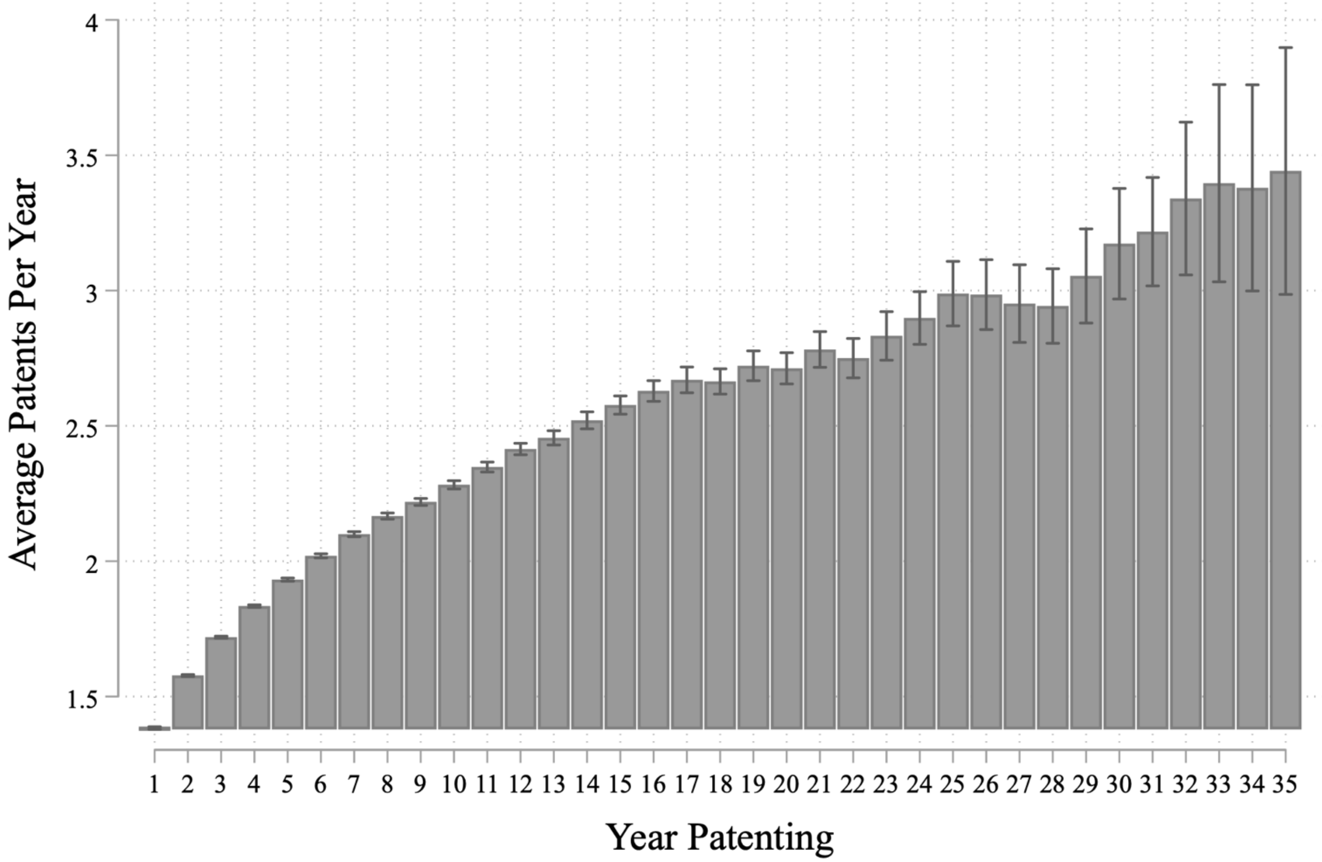
The main contribution is that that I learn a **contribution share** $\boldsymbol{\omega_{ip}}$ for inventor *i* to patent *p*

Summary of Results

- I use the learnt contribution shares to study collaboration patterns over the lifecycle of an inventors career.

- I demonstrate a **quantity-quality** trade off for inventors as they become senior inventors.

Fact 1: Seniors collaborate on more patents each year of experience.

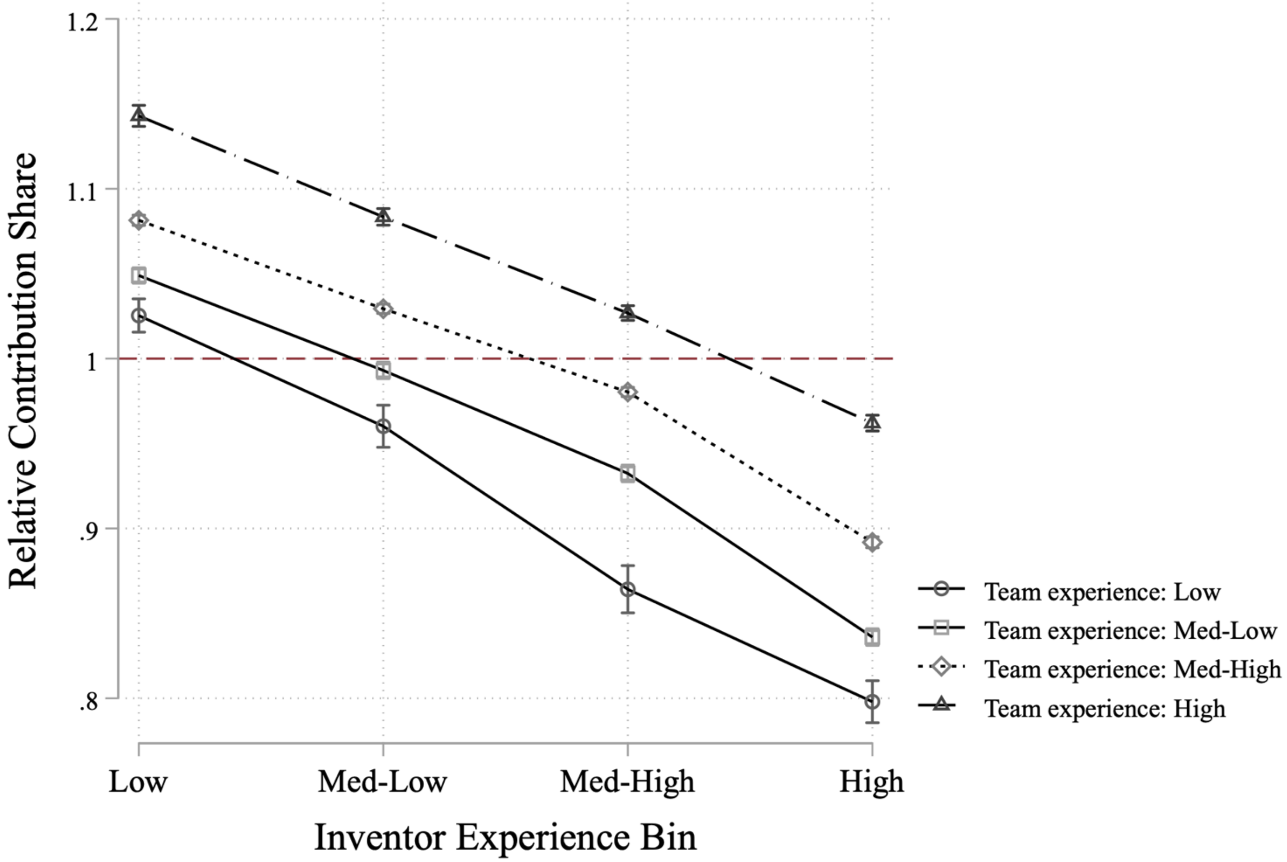


Summary of Results

- I measure the relative contribution share as the inventors share divided by the inverse of the team size

- If the relative contribution is above 1 then they are contributing more within the team.

Fact 2: Seniors contribute less when collaborating with juniors.



Summary of Results

- I measure the concentration of a team's contribution share by finding the distance between equal shares, and the team shares.
- A higher concentration value means a less equal distribution of the workload.

Fact 3: Concentrated contribution shares correlate with lower patent values.

TABLE 1.3

CONCENTRATION ON PATENT OUTCOMES

	(1) ln(Citations+1)	(2) ln(Market)	(3) ln(Novelty+1)	(4) ln(Impact+1)	(5) Pr(Break)
Concentration	-1.9334***	-4.6084***	-0.6339***	-1.0018***	-0.2366***
	(0.0844)	(0.2161)	(0.0629)	(0.0950)	(0.0229)
N	27103	14917	26738	26738	20111
R^2	0.323	0.227	0.189	0.219	0.194

Notes: This table presents regression estimates examining the relationship between team concentration and five innovation outcomes: citations, market value, novelty, impact and the likelihood of producing a breakthrough patent. Concentration is measured as the Euclidean distance between the vector of contribution shares and a uniform distribution. All models include year fixed effects, and robust standard errors are used.

The Rise and Fall of Research Fields

With Christian Fons-Rosen

01. Combine data on patent & paper abstracts (Arts et al., 2025)

They provide access to an already cleaned OpenAlex & USPTO database!

02. Use the Logic-Mill model to represent abstracts as vectors.

A pre-trained embedding model from the [Max-Planck Institute for Innovation and Competition].

03. Implement a BERTopic model to cluster vectors.

This combines the benefits of topic models with embeddings–it takes embedding vectors and clusters them into topics.

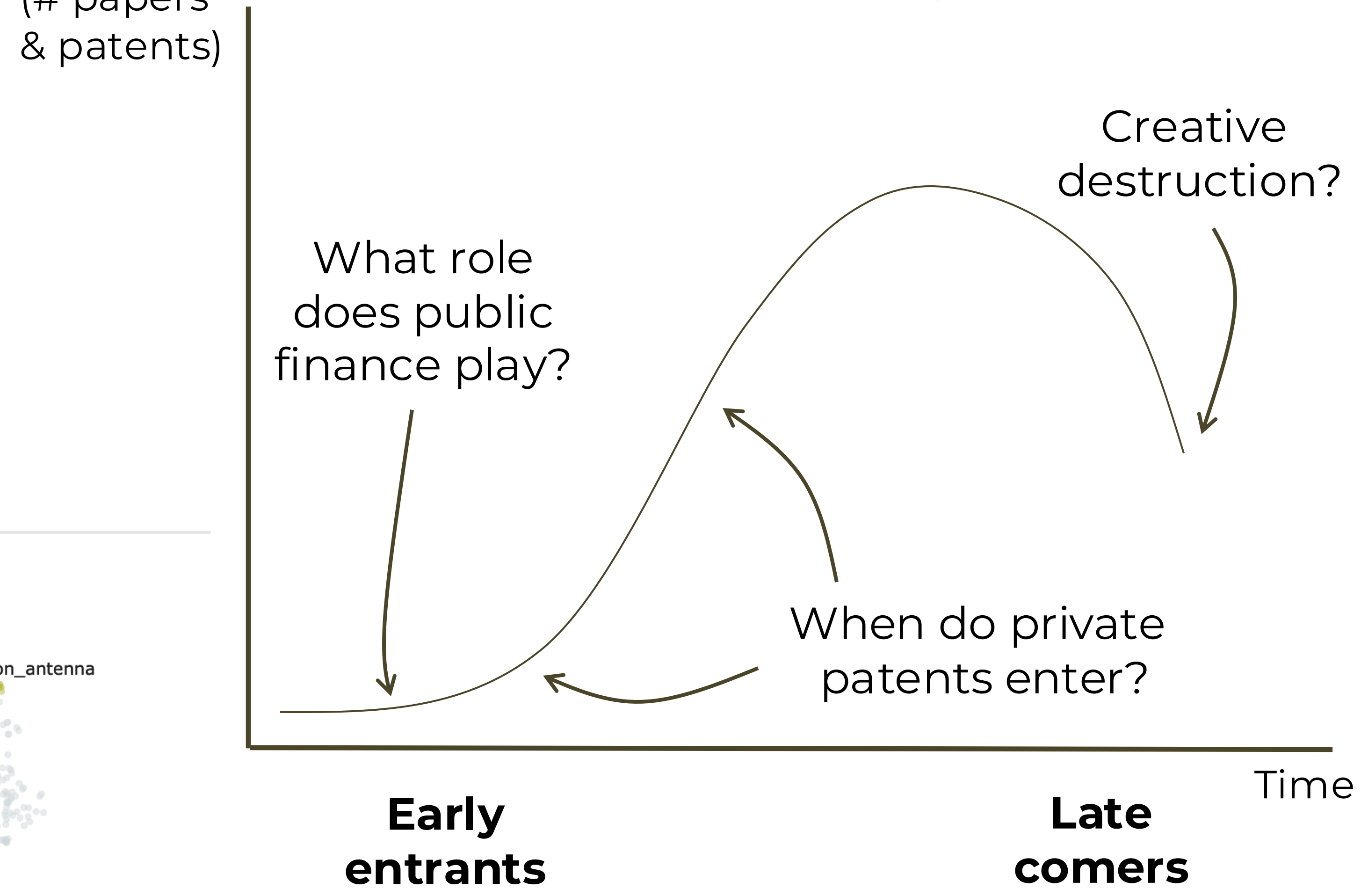
The Rise and Fall of Research Fields

1. Capture the creation and death of research fields

The Rise and Fall of Research Fields

1. Capture the creation and death of research fields

2. Study the life-cycle within one research field (hypothesis)



Two Papers with Two Different Methods

Teams and Text: Collaboration Patterns

Objective to disentangle individual contributions. → BOW structure very useful as simpler to disentangle individual words. → So, the LDA model was ideal.

The Rise and Fall of Research Fields

Finding similar texts is most important to correctly cluster into fields. → *LogicMill* allows us to combine both patent and paper domains! → So, I needed embeddings and could make use of the fantastic pre-trained model ready to go!

Open Technical Questions

01. Large Language Models

We haven't discussed LLMs! But they are becoming increasingly popular to analysing text (but they have their problems!

02. Generated Regressors

There are several concerns around biases *(Battaglia et al., (2024) critique)* from introducing noisy estimates to regressions and how to correct standard errors.

03. Causal Analysis

Use text to define instruments. Not only to access new data but also to capture quasi-random variation.

04. Validation Tests

Advances needed in cross-validation for unsupervised models! There are a lot of parameters and pre-processing steps to check!

Getting Started

- The beauty is that a lot of data and programs are open source!

Available Data

- USPTO–PatentsView
- OpenAlex (Academic Papers)
- Web-scraping–E.g. glassdoor reviews (Check T&Cs)
- Factiva–News stories (Paid!)

Code and Programs

- Hugging Face–Huge repository of pre-trained models!
- Gensim/spaCy–Leading packages for text analysis in Python.
- BERTopic – the best of both worlds?
- Learn the code: Stephen Hansen's Notebooks

Final Thoughts

- **Four steps** to text analysis in Economics
- Four top models of **transforming text to numbers**.
- Organisation economics applications are expanding:
  - Culture and Norms
  - Organisational Change and Adaptation
  - Leadership and Management
  - Really any area works…
- I am more than happy to go into more detail on any topics over the week!

Citations

Ash, Elliott, and Stephen Hansen. *'Text Algorithms in Economics'*. Annual Review of Economics 15, no. 1 (2023): 659–88. 01

Azoulay, Pierre, Joshua S. Graff Zivin, and Jialan Wang. "Superstar Extinction." The Quarterly Journal of Economics 125, no. 2 (2010): 549–589. 02

Bandiera, Oriana, Andrea Prat, Stephen Hansen, and Raffaella Sadun. *'CEO Behavior and Firm Performance'*. Journal of Political Economy 128, no. 4 (2020): 1325–69. 03

Chaturvedi, Sugat, Kanika Mahajan, and Zahra Siddique. '*Using Domain-Specific Word Embeddings to Examine the Demand for Skills*'. Research in Labor Economics (Working Paper 2023): 171–223. 04

Erhardt, Sebastian, Mainak Ghosh, Erik Buunk, Michael E. Rose, and Dietmar Harhoff. 'Logic Mill: A Knowledge Navigation System'. (Working Paper 2023) 05

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. *'Text as Data'*. Journal of Economic Literature 57, no. 3 (2019): 535–74. 06

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. *'Measuring Technological Innovation over the Long Run'*. American Economic Review: Insights 3, no. 3 (2021): 303–20. 07