NLP caught in the wild

As used in my research

Julius Koschnick

SDU

We’ve got new tools – now how to apply them?

→ How can we apply NLP methods to text data so that the outcome corresponds to economic concepts

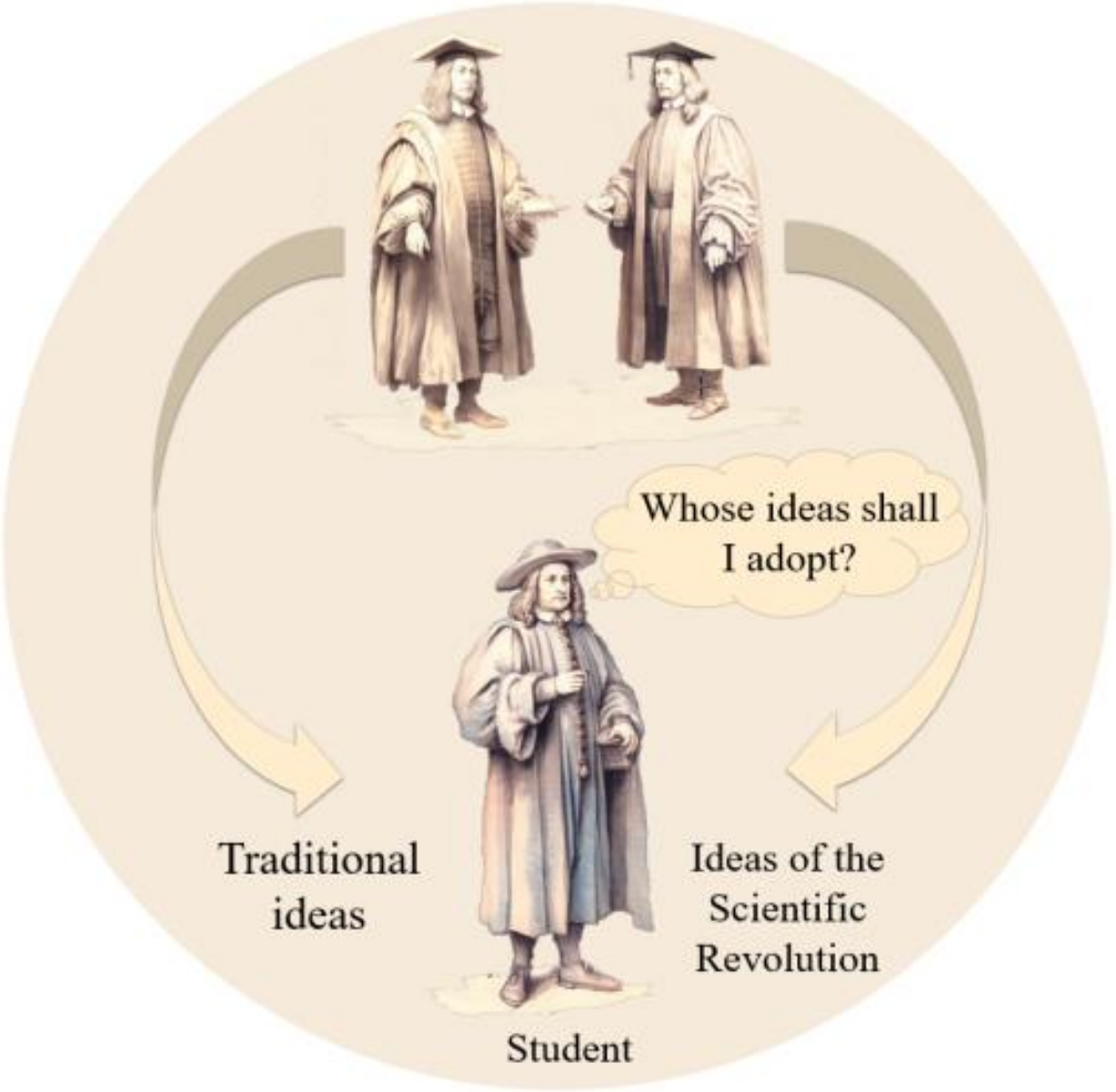
The data

→ Records of enrolled students at the University of Oxford and Cambridge, 1600-1800
→ Universe of everything published in England, 1600-1800

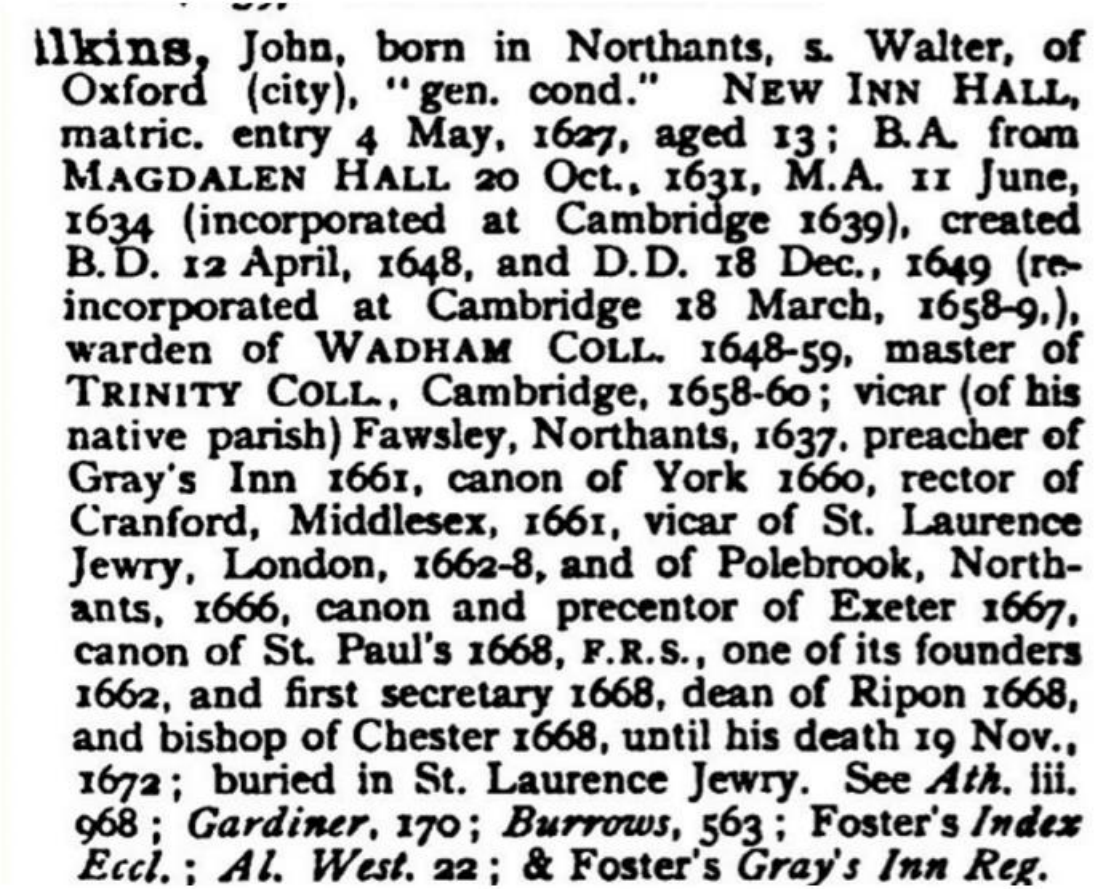
→ Lots of unstructured text material

→ OK, so what to do with this?

The question



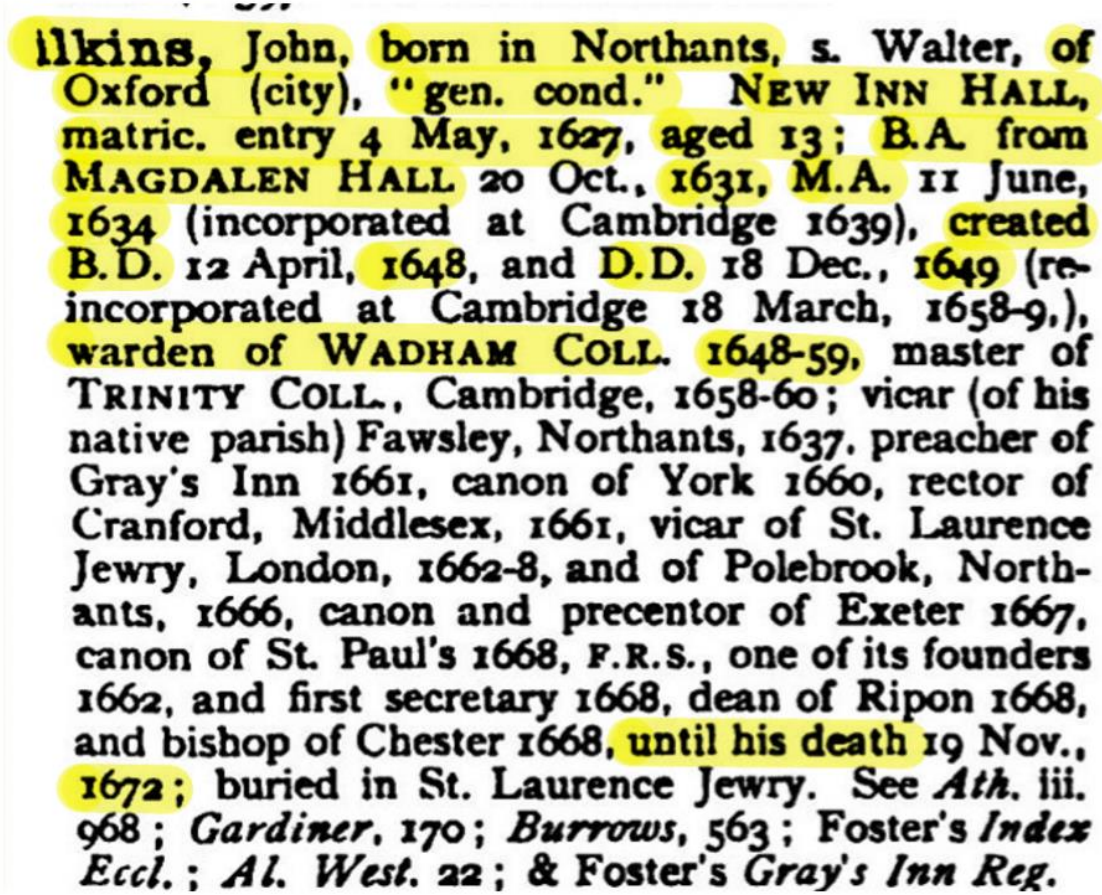
Deterministic structures



ilkins, John, born in Northants, s. Walter, of Oxford (city), "gen. cond." NEW INN HALL, matric. entry 4 May, 1627, aged 13; B.A. from MAGDALEN HALL 20 Oct., 1631, M.A. 11 June, 1634 (incorporated at Cambridge 1639), created B.D. 12 April, 1648, and D.D. 18 Dec., 1649 (re-incorporated at Cambridge 18 March, 1658-9,), warden of WADHAM COLL. 1648-59, master of TRINITY COLL., Cambridge, 1658-60; vicar (of his native parish) Fawsley, Northants, 1637, preacher of Gray's Inn 1661, canon of York 1660, rector of Cranford, Middlesex, 1661, vicar of St. Laurence Jewry, London, 1662-8, and of Polebrook, Northants, 1666, canon and precentor of Exeter 1667, canon of St. Paul's 1668, F.R.S., one of its founders 1662, and first secretary 1668, dean of Ripon 1668, and bishop of Chester 1668, until his death 19 Nov., 1672; buried in St. Laurence Jewry. See *Ath.* iii. 968; *Gardiner*, 170; *Burrows*, 563; Foster's *Index Eccl.*; *Al. West.* 22; & Foster's *Gray's Inn Reg.*

Figure: John Wilkins (1614–1672), entry from *Alumni Oxonienses 1500–1714*

Deterministic structures



ilkins, John, born in Northants, s. Walter, of Oxford (city), "gen. cond." NEW INN HALL, matric. entry 4 May, 1627, aged 13; B.A. from MAGDALEN HALL 20 Oct., 1631, M.A. 11 June, 1634 (incorporated at Cambridge 1639), created B.D. 12 April, 1648, and D.D. 18 Dec., 1649 (re-incorporated at Cambridge 18 March, 1658-9,), warden of WADHAM COLL. 1648-59, master of TRINITY COLL., Cambridge, 1658-60; vicar (of his native parish) Fawsley, Northants, 1637, preacher of Gray's Inn 1661, canon of York 1660, rector of Cranford, Middlesex, 1661, vicar of St. Laurence Jewry, London, 1662-8, and of Polebrook, Northants, 1666, canon and precentor of Exeter 1667, canon of St. Paul's 1668, F.R.S., one of its founders 1662, and first secretary 1668, dean of Ripon 1668, and bishop of Chester 1668, until his death 19 Nov., 1672; buried in St. Laurence Jewry. See *Ath.* iii. 968; *Gardiner*, 170; *Burrows*, 563; Foster's *Index Eccl.*; *Al. West.* 22; & Foster's *Gray's Inn Reg.*

Figure: John Wilkins (1614–1672), entry from *Alumni Oxonienses 1500–1714*

Deterministic structures

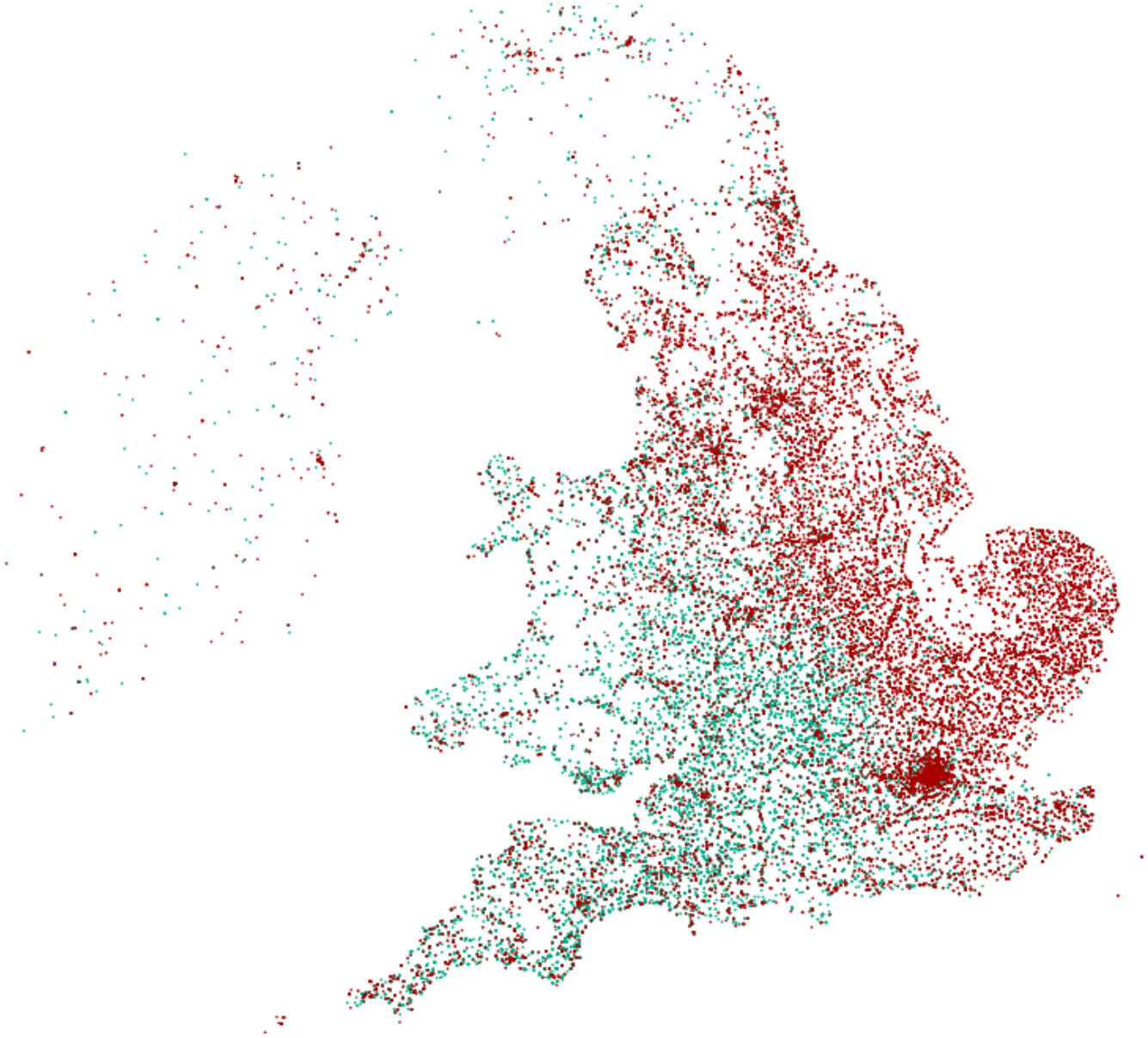


Figure: Places of origin of Oxford and Cambridge students

SDU

Solution: Regular expressions + add. info

```
# Extract and move content within square brackets to county
bracket_content <- regmatches(df[[place_var]], gregexpr("\\[(.*?)\\]", df[[place_var]]))
```

SDU

Solution: Regular expressions + add. info

```r
# Extract and move content within square brackets to county
bracket_content <- regmatches(df[[place_var]], gregexpr("\\[(.*?)\\]", df[[place_var]]))
```

(…)

```r
for (county_var in all_county_vars) {
  # Standardize county names and set non-matching entries to NA
  df[[county_var]] <- trimws(df[[county_var]])  # Trim whitespace
  df[[county_var]] <- county_lookup[df[[county_var]]]  # Map abbreviations to full names
  df[[county_var]][is.na(df[[county_var]])] <- NA  # Set non-matches to NA
}
```

```r
clean_county <- function(df) {
  # Define recognized counties and their common abbreviations
  county_mapping <- list(
    # English counties
    "Bedfordshire" = c("Beds", "Bedfordshire"),
    "Berkshire" = c("Berks", "Berkshire"),
    "Buckinghamshire" = c("Bucks", "Buckinghamshire"),
    "Cambridgeshire" = c("Cambs", "Cambridgeshire"),
    "Cheshire" = c("Cheshire"),
    "Cornwall" = c("Cornwall"),
    "Cumberland" = c("Cumberland"),
    "Derbyshire" = c("Derby", "Derbs", "Derbyshire"),
    "Devon" = c("Devon", "Devonshire"),
    "Dorset" = c("Dorset"),
    "Durham" = c("Durham"),
    "Essex" = c("Essex"),
    "Gloucestershire" = c("Gloucs", "Glouc.", "Gloucestershire"),
    "Hampshire" = c("Hants", "Hampshire"),
    "Herefordshire" = c("Hereford", "Herefordshire"),
    "Hertfordshire" = c("Herts", "Hertfordshire"),
    "Huntingdonshire" = c("Hunts", "Huntingdonshire"),
    "Kent" = c("Kent"),
    "Lancashire" = c("Lancs", "Lancashire"),
    "Leicestershire" = c("Leics", "Leicestershire"),
    "Lincolnshire" = c("Lincs", "Linc.", "Lincolnshire"),
    "Middlesex" = c("Middlesex"),
    "Norfolk" = c("Norfolk"),
    "Northamptonshire" = c("Northants", "Northamptonshire"),
    "Northumberland" = c("Northumberland"),
    "Nottinghamshire" = c("Notts", "Nottinghamshire"),
    "Oxfordshire" = c("Oxon", "Oxfordshire"),
    "Rutland" = c("Rutland"),
    "Shropshire" = c("Salop", "Shropshire"),
    "Somerset" = c("Somerset"),
    "Staffordshire" = c("Staffs", "Staffordshire")
```

SDU

Using identity recognition / LLMs to extract deterministic structures

→ Potentially time saving
→ But not necessarily expanding the frontier of possibilities

```python
# Define function to translate text and return translated text and original language
def translate_title(title):
    response = client.chat.completions.create(
        model="gpt-3.5-turbo-0125",
        response_format={"type": "json_object"},
        messages=[
            {"role": "system", "content": "You are a helpful assistant who outputs translations in JSON
            {"role": "user", "content": title}
        ],
        temperature=0.1,  # Strike balance between creativity and factualness - we choose a very low val
        seed=12345      #set seed for reproducibility
    )

    # Extract the relevant parts of the response
    translated_data = json.loads(response.choices[0].message.content)
    translated_data_list.append(translated_data)  # Save translated data to the list
    return translated_data['translated_text'], translated_data['original_language']
```

SDU

Non-deterministic structures

→ Titles on everything published in England between 1600-1800
→ Information that's
  → Dense in information
  → Unstructured
  → Context dependent

Dioptrica nova. A treatise of dioptricks, in two parts. Wherein the various effects and appearances of spherick glasses, both convex and concave, single and combined, in telescopes and microscopes, together with their usefulness in many concerns of humane life, are explained (William Molyneux, 1692)

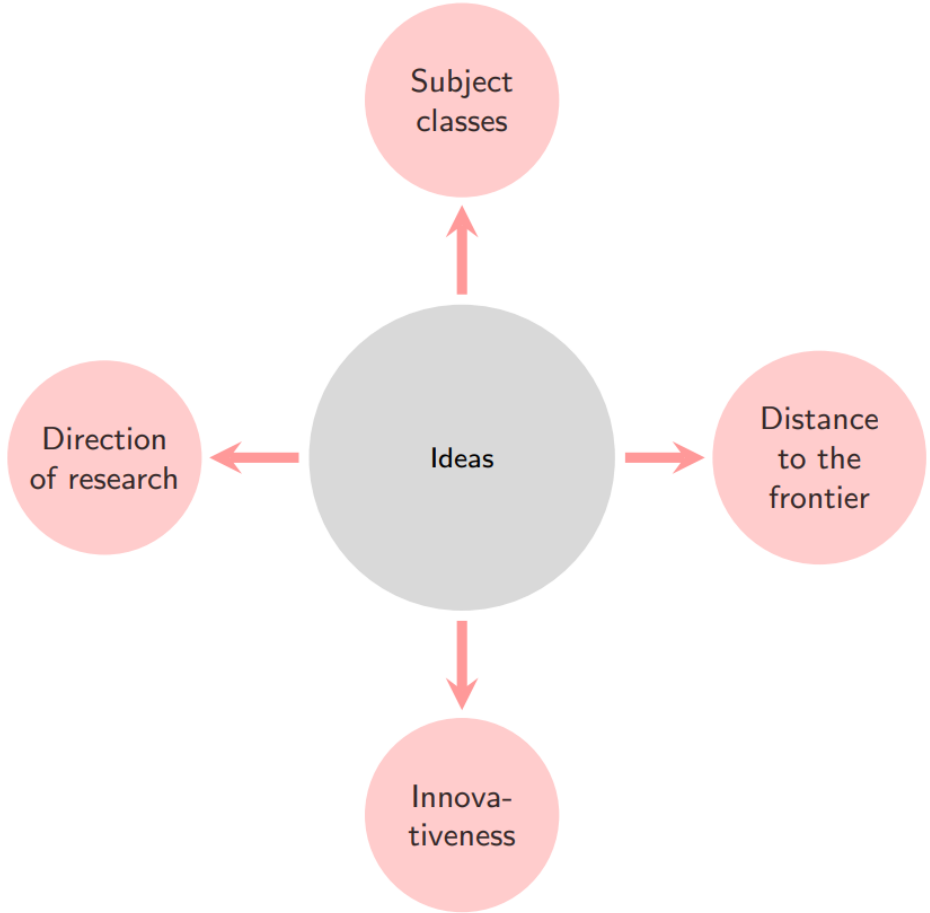
Strange and wonderful news from Ireland: of a whale of a prodigious size, being eighty two foot long, cast ashore on the third of this instant February, near Dublin, and there exposed to publick view. In a letter to a person of quality (Patrick Simmons, 1683)

Data at hand

→ English Short Title Catalogue:
  → Includes titles of all surviving printed items in Great Britain before 1800
  → For 1600-1800: 469,962 printed items
  → 4% non English titles, translated to English using the Google Translate API
→ Matched to students and teachers:
  → Overall, 94,378 title matches for 7,265 students
  → So, 5% of all Oxford and Cambridge graduates make up for at least 35% of all publications by real people in Britain for 1600-1800

Non-deterministic structures

→ Turning it into economic concepts

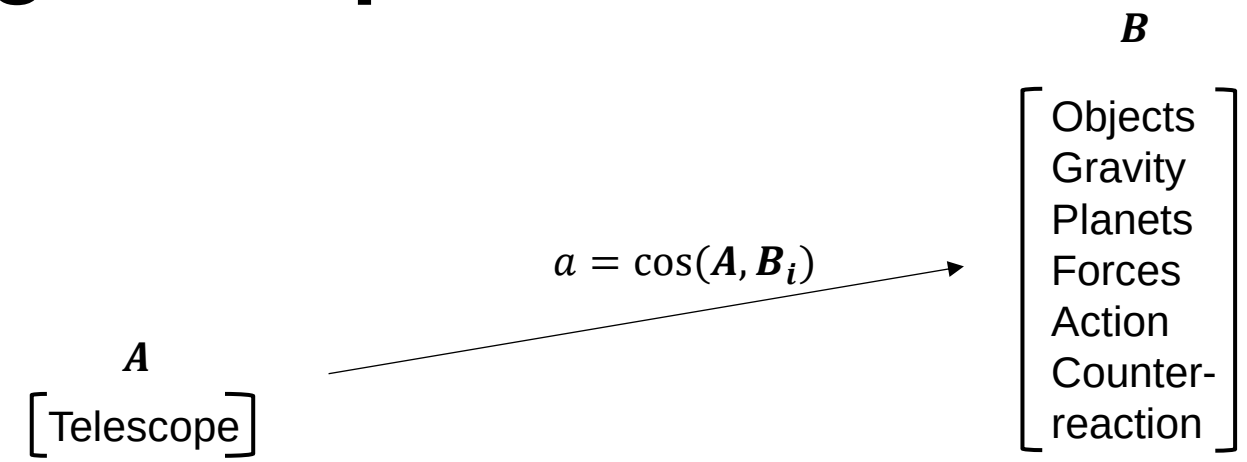


SDU

Conceptual example: How scientific is a title?

→ Not a classification exercise
  → What is it on?
→ Instead: Looking for quantitative measurement

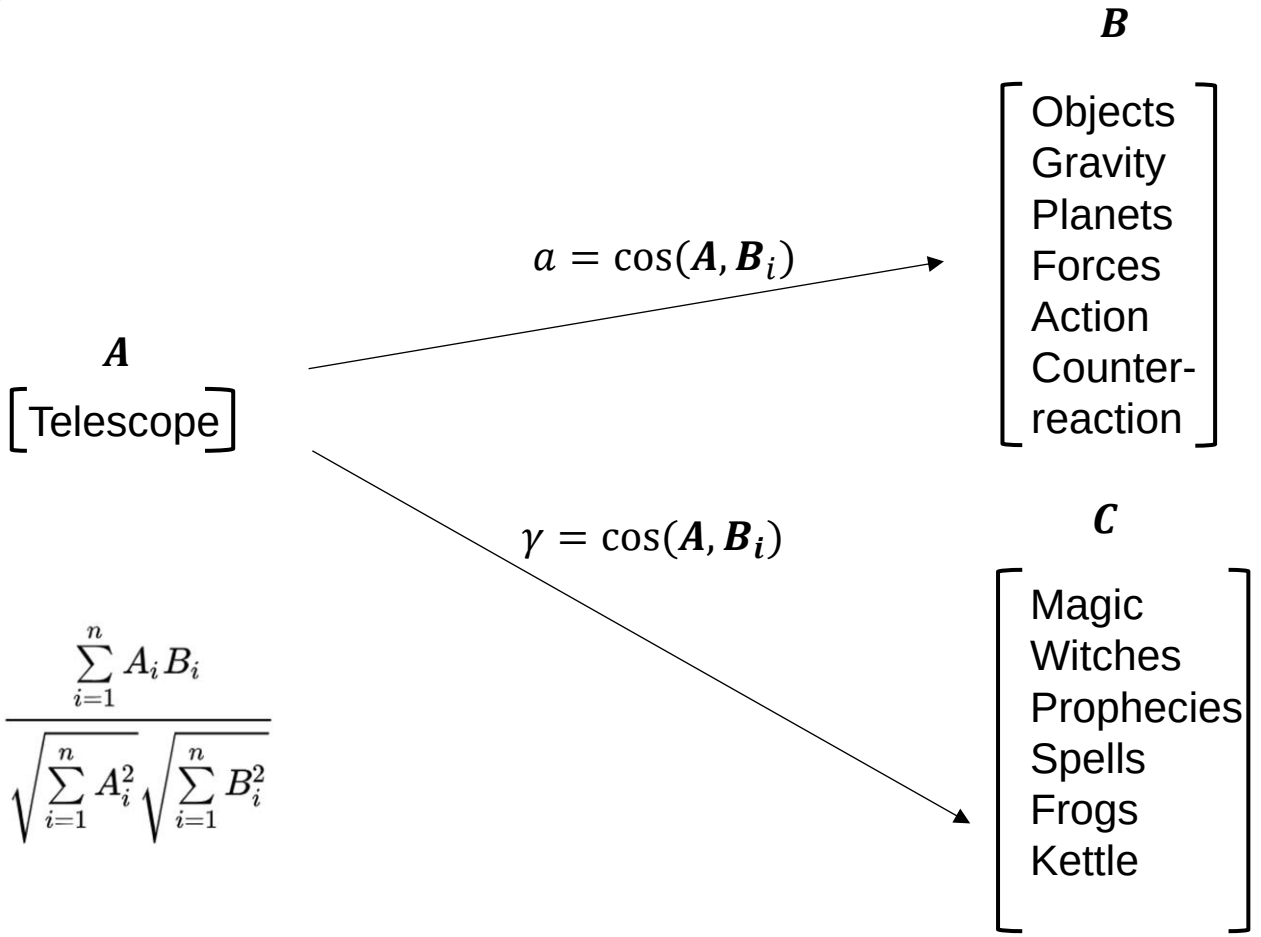
Measuring conceptual distance



$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

SDU

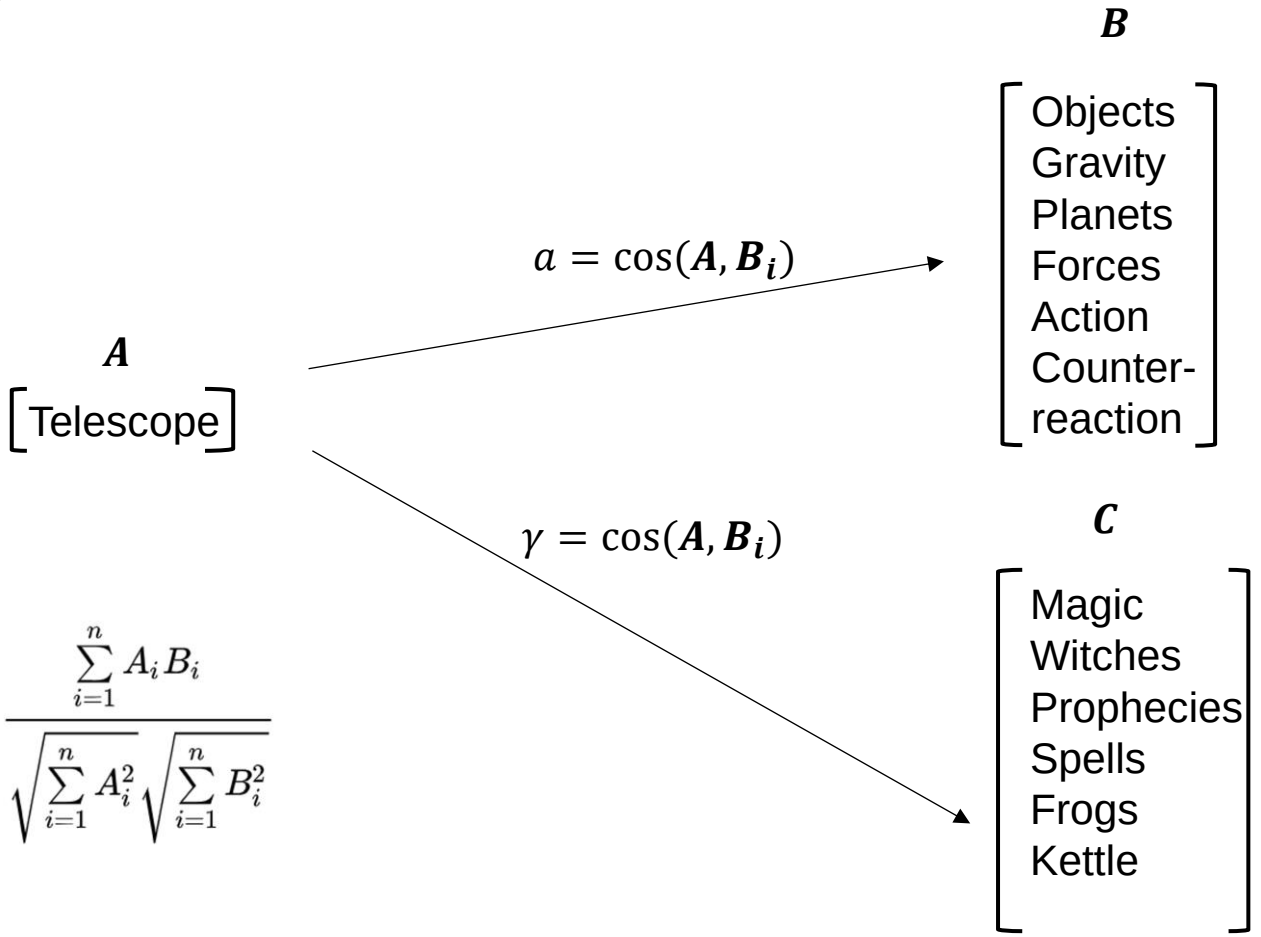
Measuring conceptual distance



$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

SDU

Measuring conceptual distance

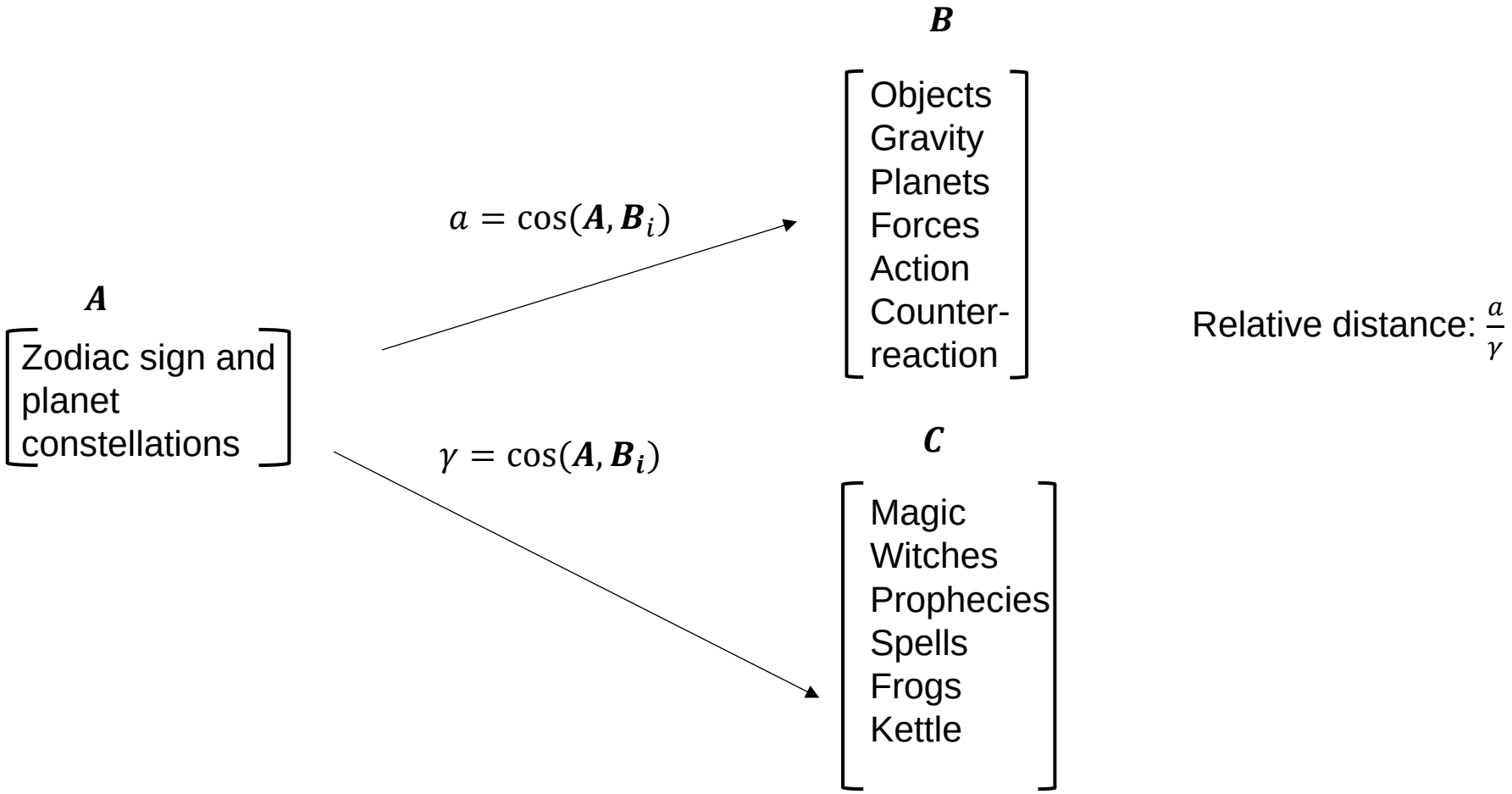


Relative distance: $\frac{a}{\gamma}$

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

SDU

Measuring conceptual distance

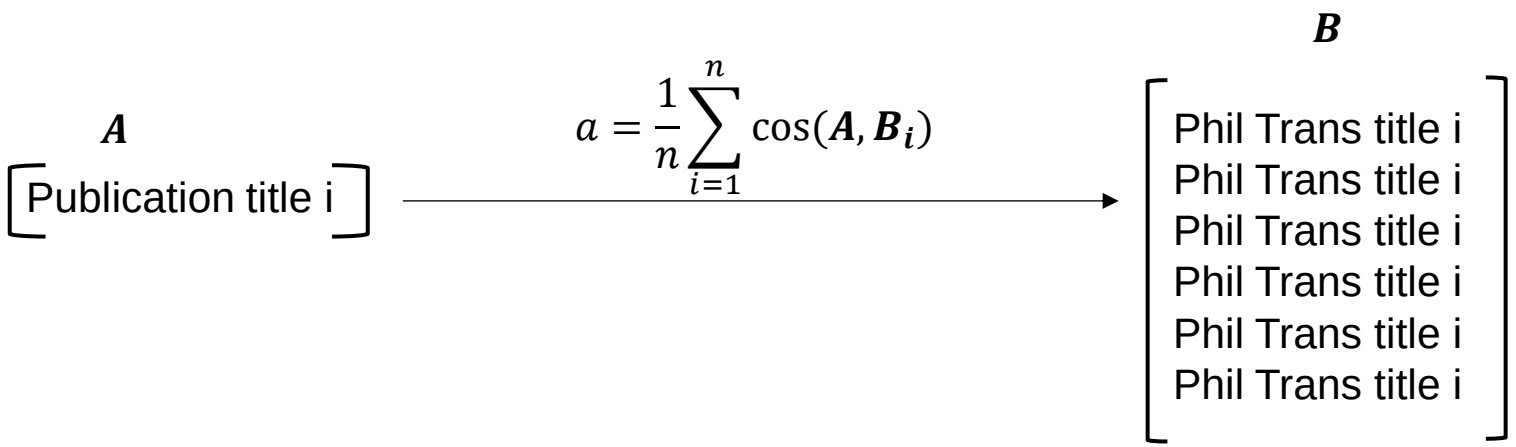


Relative distance: $\frac{a}{\gamma}$

Back to real history: Distance to the frontier

→ Now, moving from words to full titles
→ Sentence transformer model
  → transforms short text pieces into embedding space vectors

→ Economic concept: Distance to the frontier
→ What is the frontier
  → Publications in the journal of the Royal Society, the *Philosophical Transactions*



SDU

A measure of innovation

→ Measure of innovation based on publication titles
  → Reason: We don’t have citation counts back in the past

Intuition: What is innovation?

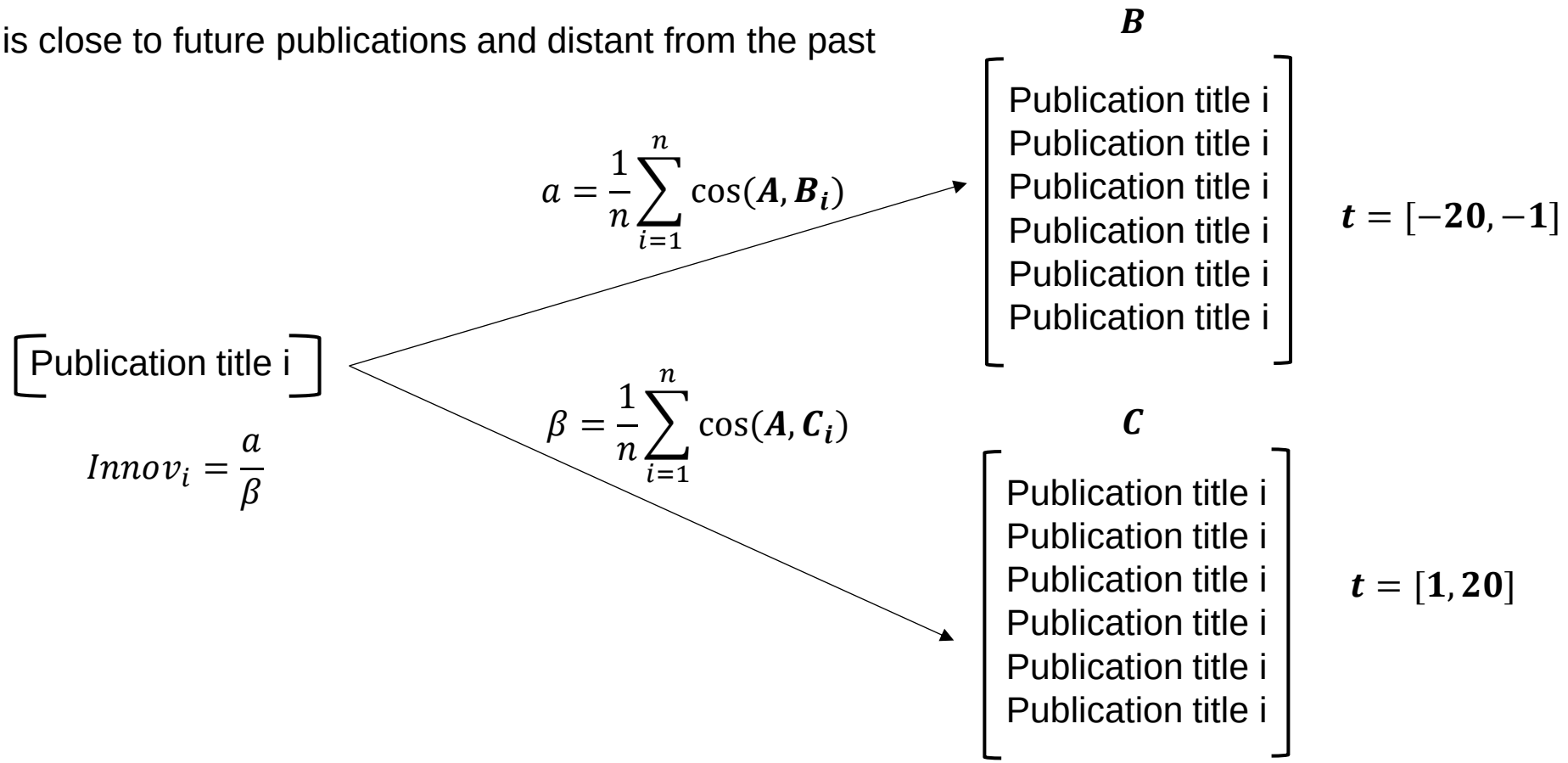
→ A title that is close to future publications?
→ No, has predictive value but something else could have been there before

→ A title is innovative when it is close the future & unrelated to the past?
→ No, could just be stupid

→ A title that is close to future publications and distant from the past
→ Sounds like a good definition

A measure of innovation

→ A title that is close to future publications and distant from the past



SDU

A measure of innovation

author	Titles	Publi-cation Year	Break-through in-dex
Boyle, Robert, 1627-1691.	new experiments physico mechanicall touching the spring of the air and its effects made for the most part in new pneumatical engine written by way of letter to the right honorable charles lord vicount of dungarvan eldest son to the earl of corke by the honorable robert boyle esq	1660	1.60
Gordon, Andreas, 1712-1750 or 1751.	attempt at an explanation of electricity	1745	1.54
Bacon, Francis, 1561-1626.	the great establishment of francis de verulam chancellor of england supreme	1620	1.47
Bacon, Francis, 1561-1626.	francis de verulamius chancellor of england great instauration	1620	1.37
Watson, William, Sir, 1715-1787.	experiments and observations tending to illus-trate the nature and properties of electricity in one letter to martin folkes esq president and two to the royal society by william watson	1745	1.36

SDU

Classification

→ Coming up in class soon

→ Intuition: Only ca. 40% of titles classified by British Library
→ Train a BERT model on the classification of the British Library
→ And predict classification of the rest of the titles

SDU

Classification

→ What we get from the classification approach:
→ Direction of research

$$\mathbf{v} = \begin{bmatrix} \frac{b_1}{n} \\ \frac{b_2}{n} \\ \frac{b_3}{n} \\ \frac{b_4}{n} \\ \frac{b_5}{n} \\ \frac{b_6}{n} \\ \frac{b_7}{n} \\ \frac{b_8}{n} \\ \frac{b_9}{n} \end{bmatrix} \begin{matrix} \text{(Astronomy)} \\ \text{(Almanacs)} \\ \text{(Applied Physics)} \\ \text{(Mathematics)} \\ \text{(Chemistry)} \\ \text{(Biology)} \\ \text{(Geography)} \\ \text{(Medicine)} \\ \text{(Scientific Instruments)} \end{matrix}$$

Estimate at the within-topic level

→ Oxford and Cambridge in the 17th century:
  → Teaching happens at the college level
  → Each college employs its own teachers
→ Treatment: Variation across colleges and variation across topics
  → Setup allows for student fixed effects

$$v_{jict} = \beta_1 p_{jict} + \boldsymbol{X'_{ct}} \beta_2 + \delta_i + \gamma_c + \zeta_j + \alpha_t + \varepsilon_{jict}$$

with:

- v_{jict} = Student i's publication share in field j at college c in cohort t
- p_{jict} = Teachers' publication share in field j at college c in cohort t
- $\boldsymbol{X'_{ct}}$: Vector of control variables for teacher characteristics
- $\delta_i + \gamma_c + \zeta_j + \alpha_t$ = Student-, i, field, j, college-, c, and time-, t fixed effects

Panel A: University of Oxford

	Share of each topic in student publications		
	(1)	(2)	(3)
	Mean topic	Mean topic	Mean topic
Log share of each topic in teacher publications	0.0662**	0.0285**	0.0297**
	(0.0201)	(0.00968)	(0.0112)
Teacher and college level controls	Yes	Yes	—
Student publication controls	Yes	Yes	—
Year fixed effects	Yes	Yes	—
College fixed effects	Yes	Yes	—
Topic fixed effects	No	Yes	Yes
Student fixed effects	No	No	Yes
Observations	11484	11484	11484
R-squared	0.02	0.04	0.17

Panel B: University of Cambridge

	Log share of each topic in student publications		
	(1)	(2)	(3)
	Mean topic	Mean topic	Mean topic
Log share of each topic in teacher publications	0.0482**	0.0124***	0.00996***
	(0.0148)	(0.00359)	(0.000598)
Teacher and college level controls	Yes	Yes	—
Student publication controls	Yes	Yes	—
Year fixed effects	Yes	Yes	—
College fixed effects	Yes	Yes	—
Topic fixed effects	No	Yes	Yes
Student fixed effects	No	No	Yes
Observations	12231	12231	12231
R-squared	0.02	0.04	0.17

Conclusion

→ We started with a question
  → Do teachers influence the research direction of their students

→ We started with lots of unstructured text data

→ We used NLP techniques to
  → Extract information from the data
  → Quantify economic concepts
    → Distance to the frontier
    → Innovation
    → Research direction

→ Then, we used this new information to estimate teacher-student effects

→ New insights from variation hidden in text data