# Text Data in Business and Economics

Basel University - Autumn 2024

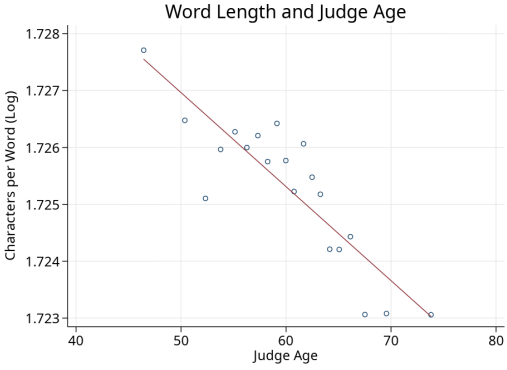
2. Style Features and Dictionaries

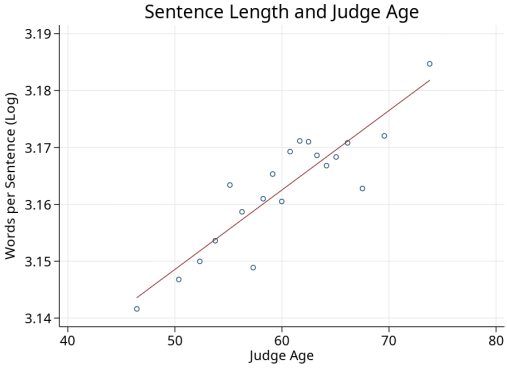
# Judge Age and Writing Style

Ash, Goessmann, and MacLeod (2022)

# Judge Age and Writing Style

Ash, Goessmann, and MacLeod (2022)



# Optimal Legal Complexity (Katz and Bommarito 2014)

- More legal detail is needed to properly specify rules and target incentives to activities and groups.
  - but there are costs to understanding/following/maintaining complex laws, so there is a trade off.

# Optimal Legal Complexity (Katz and Bommarito 2014)

- More legal detail is needed to properly specify rules and target incentives to activities and groups.
  - but there are costs to understanding/following/maintaining complex laws, so there is a trade off.
- Katz and Bommarito measure complexity/detail from the text – number of words for code title, and also *word entropy* $\approx$ diversity of the vocabulary.

# Optimal Legal Complexity (Katz and Bommarito 2014)

- More legal detail is needed to properly specify rules and target incentives to activities and groups.
  - but there are costs to understanding/following/maintaining complex laws, so there is a trade off.
- Katz and Bommarito measure complexity/detail from the text – number of words for code title, and also *word entropy* $\approx$ diversity of the vocabulary.

Five largest and smallest titles by token count

| Title | Tokens | Tokens per section |
|---|---|---|
| Public Health and Welfare (Title 42) | 2,732,251 | 369.22 |
| Internal Revenue Code (Title 26) | 1,016,995 | 487.07 |
| Conservation (Title 16) | 947,467 | 200.48 |
| Commerce and Trade (Title 15) | 773,819 | 336.88 |
| Agriculture (Title 7) | 751,579 | 274.00 |
| President (Title 3) | 7,564 | 120.06 |
| Intoxicating Liquors (Title 27) | 6,515 | 144.78 |
| Flag and Seal, Seat of Govt. and the States (Title 4) | 5,598 | 119.11 |
| General Provisions (Title 1) | 3,143 | 80.59 |
| Arbitration (Title 9) | 2,489 | 80.29 |

# Optimal Legal Complexity (Katz and Bommarito 2014)

- More legal detail is needed to properly specify rules and target incentives to activities and groups.
  - but there are costs to understanding/following/maintaining complex laws, so there is a trade off.
- Katz and Bommarito measure complexity/detail from the text – number of words for code title, and also *word entropy* $\approx$ diversity of the vocabulary.

Five largest and smallest titles by token count

| Title | Tokens | Tokens per section |
|---|---|---|
| Public Health and Welfare (Title 42) | 2,732,251 | 369.22 |
| Internal Revenue Code (Title 26) | 1,016,995 | 487.07 |
| Conservation (Title 16) | 947,467 | 200.48 |
| Commerce and Trade (Title 15) | 773,819 | 336.88 |
| Agriculture (Title 7) | 751,579 | 274.00 |
| President (Title 3) | 7,564 | 120.06 |
| Intoxicating Liquors (Title 27) | 6,515 | 144.78 |
| Flag and Seal, Seat of Govt. and the States (Title 4) | 5,598 | 119.11 |
| General Provisions (Title 1) | 3,143 | 80.59 |
| Arbitration (Title 9) | 2,489 | 80.29 |

Five highest and lowest titles by word entropy

| Title | Word entropy |
|---|---|
| Commerce and Trade (Title 15) | 10.80 |
| Public Health and Welfare (Title 42) | 10.79 |
| Conservation (Title 16) | 10.75 |
| Navigation and Navigable Waters (Title 33) | 10.67 |
| Foreign Relations and Intercourse (Title 22) | 10.67 |
| Intoxicating Liquors (Title 27) | 9.01 |
| President (Title 3) | 8.89 |
| National Guard (Title 32) | 8.50 |
| General Provisions (Title 1) | 8.49 |
| Arbitration (Title 9) | 8.24 |

Quantity of Text as Data

Dictionary-Based Methods

Sentiment Analysis

# Overview of Dictionary-Based Methods

- Dictionary-based text methods use a pre-selected list of words or phrases to analyze a corpus.
  - use regular expressions for this task (see notebook)

# Overview of Dictionary-Based Methods

- Dictionary-based text methods use a pre-selected list of words or phrases to analyze a corpus.
  - use regular expressions for this task (see notebook)
- Corpus-specific: counting sets of words or phrases across documents
  - (e.g., number of times a judge says “justice” vs “efficiency”)

# Overview of Dictionary-Based Methods

- Dictionary-based text methods use a pre-selected list of words or phrases to analyze a corpus.
  - use regular expressions for this task (see notebook)
- Corpus-specific: counting sets of words or phrases across documents
  - (e.g., number of times a judge says “justice” vs “efficiency”)
- General dictionaries: WordNet, LIWC, MFD, etc.

# Measuring uncertainty in macroeconomy

Baker, Bloom, and Davis (QJE 2016)

# Measuring uncertainty in macroeconomy

## Baker, Bloom, and Davis (QJE 2016)

For each newspaper on each day since 1985, submit the following query:

1. Article contains "uncertain" OR "uncertainty", AND
2. Article contains "economic" OR "economy", AND
3. Article contains "congress" OR "deficit" OR "federal reserve" OR "legislation" OR "regulation" OR "white house"

Normalize resulting article counts by total newspaper articles that month.

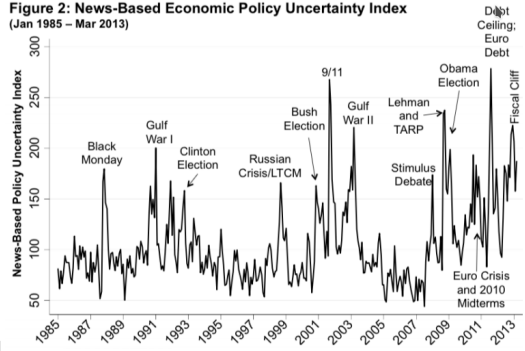
# Measuring uncertainty in macroeconomy

Baker, Bloom, and Davis (QJE 2016)

For each newspaper on each day since 1985, submit the following query:

1. Article contains "uncertain" OR "uncertainty", AND
2. Article contains "economic" OR "economy", AND
3. Article contains "congress" OR "deficit" OR "federal reserve" OR "legislation" OR "regulation" OR "white house"

Normalize resulting article counts by total newspaper articles that month.



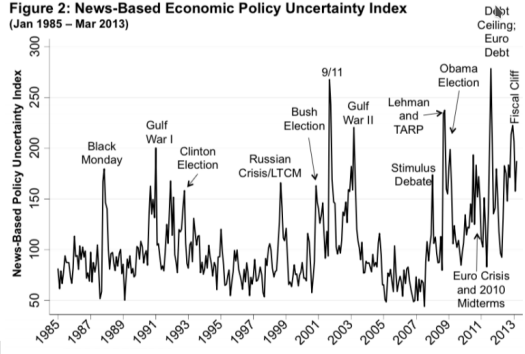
# Measuring uncertainty in macroeconomy

Baker, Bloom, and Davis (QJE 2016)

For each newspaper on each day since 1985, submit the following query:

1. Article contains "uncertain" OR "uncertainty", AND
2. Article contains "economic" OR "economy", AND
3. Article contains "congress" OR "deficit" OR "federal reserve" OR "legislation" OR "regulation" OR "white house"

Normalize resulting article counts by total newspaper articles that month.




- but see Keith et al (2020), showing some problems with this measure (`https://arxiv.org/abs/2010.04706`).

# Dictionary Methods: Identifying Race-Related Research in Economics (1)

## RACE-RELATED RESEARCH IN ECONOMICS AND OTHER SOCIAL SCIENCES*

ARUN ADVANI ELLIOTT ASH DAVID CAI IMRAN RASUL†

DECEMBER 2020

**Abstract**

How does economics compare to other social sciences in its study of race and ethnicity related issues? We assess this question using a corpus of 500,000 academic publications in economics, political science, and sociology. Using an algorithmic approach to classify race-related publications, we document that economics lags far behind the other disciplines in the volume and share of race-related research. Since 1960, there have been 13,000 race-related

# Dictionary Methods: Identifying Race-Related Research in Economics (2)

**Corpus.** We build a corpus of publications for economics, political science, and sociology. The foundation for this corpus is the *JSTOR* database of academic journals (`jstor.org`). We consider all publications in journals that *JSTOR* characterizes as comprising the disciplines of economics, sociology, and political science. Although publication series are available back to the 1880s, our

this rises steadily over time. Our working sample from 1960 to 2020 covers nearly half a million journal publications: 224,855 publications from 231 economics journals, 138,188 publications from 185 sociology journals, and 110,835 publications from 213 political science journals.

# Dictionary Methods: Identifying Race-Related Research in Economics (3)

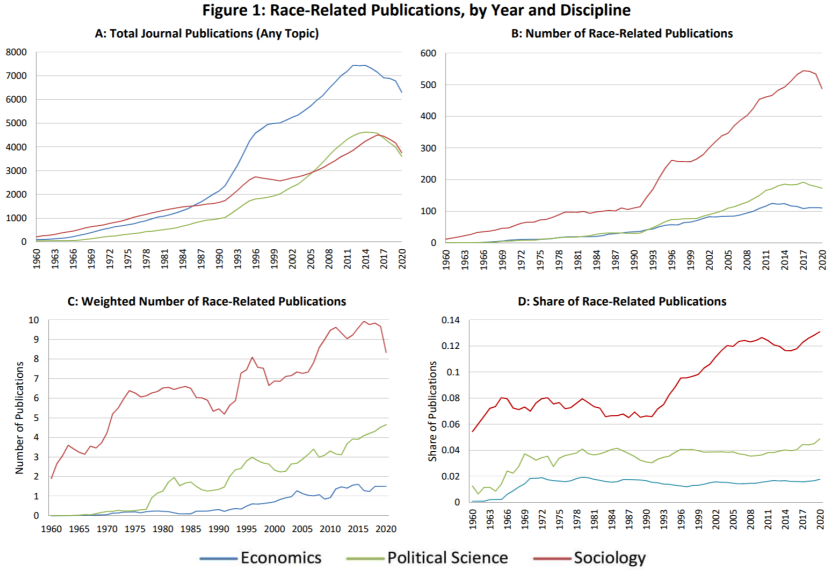
**Identifying Race-Related Research.** Given the volume of publications considered, it is infeasible to codify race-related research by hand. We thus take an automated approach and use an algorithm to classify race-related publications. We do so using keywords along two dimensions: (i) the racial or ethnic group being studied; and (ii) the issue being studied. Examples of (case-insensitive) keywords along the group dimension are race, african-american, person of color, and ethnicity. Examples of (case-insensitive) issue keywords include discrimination, prejudice, and stereotype.[2]

Our algorithm selects a publication as being race-related if: (i) at least one group keyword is in the title; or, (ii) at least one group keyword and at least one issue keyword are mentioned in the title or abstract. For rule (ii) we drop the last sentence of the abstract to avoid false positives from research that only mentions race parenthetically, say because it is part of some robustness check rather than the primary focus of study.

Specifically, we define three bands of group keywords that gradually expand on the racial or ethnic groups being studied. Band 0 consists of only abstract or generic keywords denoting racial and ethnic groups (e.g. race, ethnic, under represented minority). Band 1 adds group keywords relating to the main minority groups in the U.S. (African American, Latinos and Native Americans). Band 2 adds less salient group keywords (e.g. White, South Asian, Indian American, Japanese American) and other minorities based on religious beliefs (e.g. Muslim, Jewish). The full lexicon of group keywords used by Band are shown in Appendix Table A1.

The lexicon of issue keywords, shown in Appendix Table A2, are held constant and not split into bands. These words and phrases are broadly split across five broader topics: discrimination, inequality, diversity, identity, and historical issues. For example, discrimination includes prejudice and stereotypes, while inequality includes disparity and disadvantage.

# Dictionary Methods: Identifying Race-Related Research in Economics (4)




**Notes:** We use data from JSTOR, Scopus, and the Web of Science to construct the number and shares of race related publications in economics, political science, and sociology. Panel A reports the total number of publications in each discipline. As the publication series start in the 1880s, the publication numbers do not start exactly at zero in 1960, the first year of our working sample. Panel B reports the number of articles that are determined to be race-related by our algorithm. Panel C reports a journal-weighted version of Panel B using the journal quality weights from Angrist et al. [2020]. Panel D reports the share of articles determined to be race-related by our algorithm in each discipline. All series presented are 5-year moving averages.

# WordNet

- English word database: 118K nouns, 12K verbs, 22K adjectives, 5K adverbs

The noun "bass" has 8 senses in WordNet.
1. bass[1] - (the lowest part of the musical range)
2. bass[2], bass part[1] - (the lowest part in polyphonic music)
3. bass[3], basso[1] - (an adult male singer with the lowest voice)
4. sea bass[1], bass[4] - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass[1], bass[5] - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass[6], bass voice[1], basso[2] - (the lowest adult male singing voice)
7. bass[7] - (the member with the lowest range of a family of musical instruments)
8. bass[8] - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

**Figure 19.1** A portion of the WordNet 3.0 entry for the noun *bass*.

- Synonym sets (synsets) are a group of near-synonyms, plus a gloss (definition).
  - also contains information on antonyms (opposites), holonyms/meronyms (part-whole).
- Nouns are organized in categorical hierarchy (hence "WordNet")
  - "hypernym" – the higher category that a word is a member of.
  - "hyponyms" – members of the category identified by a word.

# WordNet Supersenses (Word Categories)

| Category | Example | Category | Example | Category | Example |
|---|---|---|---|---|---|
| ACT | *service* | GROUP | *place* | PLANT | *tree* |
| ANIMAL | *dog* | LOCATION | *area* | POSSESSION | *price* |
| ARTIFACT | *car* | MOTIVE | *reason* | PROCESS | *process* |
| ATTRIBUTE | *quality* | NATURAL EVENT | *experience* | QUANTITY | *amount* |
| BODY | *hair* | NATURAL OBJECT | *flower* | RELATION | *portion* |
| COGNITION | *way* | OTHER | *stuff* | SHAPE | *square* |
| COMMUNICATION | *review* | PERSON | *people* | STATE | *pain* |
| FEELING | *discomfort* | PHENOMENON | *result* | SUBSTANCE | *oil* |
| FOOD | *food* | | | TIME | *day* |

**Figure 19.2** Supersenses: 26 lexicographic categories for nouns in WordNet.

# WordNet Supersenses (Word Categories)

| Category | Example | Category | Example | Category | Example |
|---|---|---|---|---|---|
| ACT | *service* | GROUP | *place* | PLANT | *tree* |
| ANIMAL | *dog* | LOCATION | *area* | POSSESSION | *price* |
| ARTIFACT | *car* | MOTIVE | *reason* | PROCESS | *process* |
| ATTRIBUTE | *quality* | NATURAL EVENT | *experience* | QUANTITY | *amount* |
| BODY | *hair* | NATURAL OBJECT | *flower* | RELATION | *portion* |
| COGNITION | *way* | OTHER | *stuff* | SHAPE | *square* |
| COMMUNICATION | *review* | PERSON | *people* | STATE | *pain* |
| FEELING | *discomfort* | PHENOMENON | *result* | SUBSTANCE | *oil* |
| FOOD | *food* | | | TIME | *day* |

**Figure 19.2** Supersenses: 26 lexicographic categories for nouns in WordNet.

| Supersense | Verbs denoting . . . |
|---|---|
| body | grooming, dressing and bodily care |
| change | size, temperature change, intensifying |
| cognition | thinking, judging, analyzing, doubting |
| communication | telling, asking, ordering, singing |
| competition | fighting, athletic activities |
| consumption | eating and drinking |
| contact | touching, hitting, tying, digging |
| creation | sewing, baking, painting, performing |
| emotion | feeling |
| motion | walking, flying, swimming |
| perception | seeing, hearing, feeling |
| possession | buying, selling, owning |
| social | political and social activities and events |
| stative | being, having, spatial relations |
| weather | raining, snowing, thawing, thundering |

# General Dictionaries

- Function words (e.g. *for*, *rather*, *than*)
  - also called stopwords
  - can be used to get at non-topical dimensions, identify authors.

# General Dictionaries

- Function words (e.g. *for*, *rather*, *than*)
  - also called stopwords
  - can be used to get at non-topical dimensions, identify authors.
- LIWC (pronounced "Luke"): Linguistic Inquiry and Word Counts
  - >10000 words from >100 lists of category-relevant words, e.g. "emotion", "cognition", "work", "family", "positive", "negative" etc.

# General Dictionaries

- Function words (e.g. *for*, *rather*, *than*)
  - also called stopwords
  - can be used to get at non-topical dimensions, identify authors.
- LIWC (pronounced "Luke"): Linguistic Inquiry and Word Counts
  - >10000 words from >100 lists of category-relevant words, e.g. "emotion", "cognition", "work", "family", "positive", "negative" etc.
- Mohammad and Turney (2011):
  - code 10,000 words along four emotional dimensions: joy–sadness, anger-fear, trust-disgust, anticipation-surprise
- Warriner et al (2013):
  - code 14,000 words along three emotional dimensions: valence, arousal, dominance.

Quantity of Text as Data

Dictionary-Based Methods

Sentiment Analysis

# Sentiment Analysis

Extract a "tone" dimension – positive, negative, neutral

- standard approach is lexicon-based, but they fail easily: e.g., "good" versus "not good" versus "not very good"

# Sentiment Analysis

Extract a "tone" dimension – positive, negative, neutral

- standard approach is lexicon-based, but they fail easily: e.g., "good" versus "not good" versus "not very good"
- huggingface model hub has a number of transformer-based sentiment models

# Sentiment Analysis

Extract a "tone" dimension – positive, negative, neutral

- standard approach is lexicon-based, but they fail easily: e.g., "good" versus "not good" versus "not very good"
- huggingface model hub has a number of transformer-based sentiment models
- Off-the-shelf scores may be trained on biased corpora, eg online writing – may not work for legal text, for example.
  - Hamilton et al (2016) and Zorn and Rice (2019) show how to make domain-specific sentiment lexicons using word embeddings (more on this later).

# Problems with Sentiment Analyzers: NLP System Bias

```
text_to_sentiment("Let's go get Italian food")
2.0429166109
text_to_sentiment("Let's go get Chinese food")
1.4094033658
text_to_sentiment("Let's go get Mexican food")
0.3880198556
```

```
text_to_sentiment("My name is Emily")
2.2286179365
text_to_sentiment("My name is Heather")
1.3976291151
text_to_sentiment("My name is Yvette")
0.9846380213
text_to_sentiment("My name is Shaniqua")
-0.4704813178
```

**Is this sentiment model racist?**

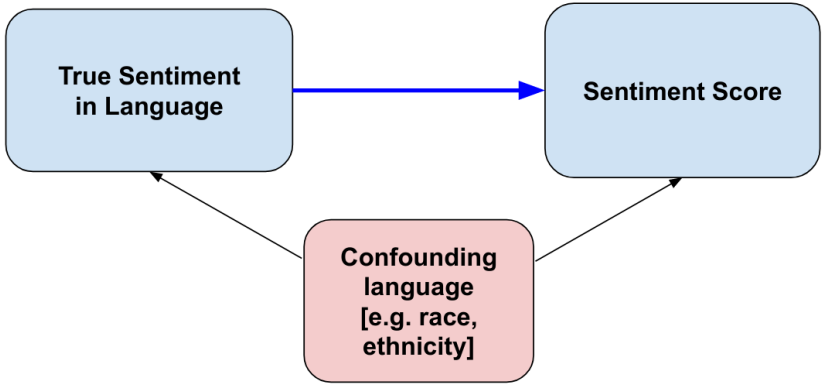
Source: Kareem Carr slides.

# NLP "Bias" is statistical bias

- Sentiment scores that are trained on annotated datasets also learn from the correlated non-sentiment information.

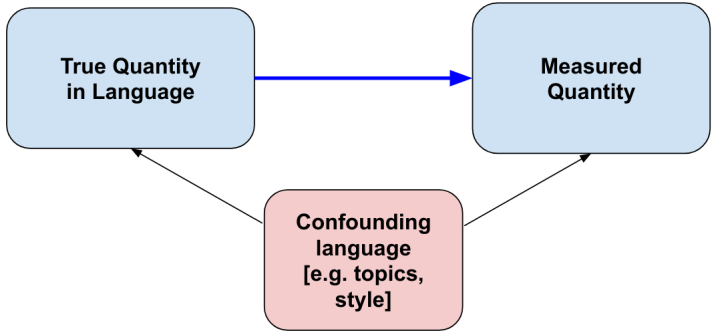
# NLP "Bias" is statistical bias

- Sentiment scores that are trained on annotated datasets also learn from the correlated non-sentiment information.



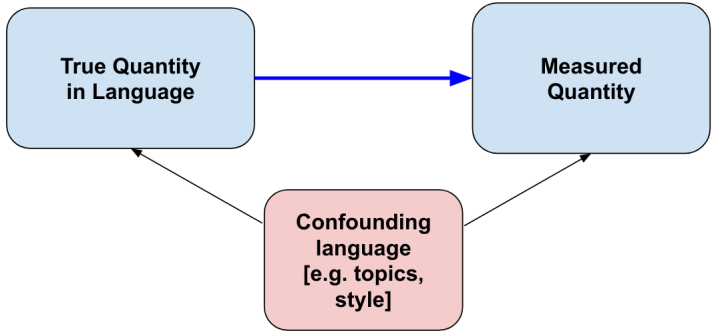

- Supervised sentiment models are confounded by correlated language factors.
  - e.g., in the training set maybe people complain about Mexican food more often than Italian food because Italian restaurants tend to be more upscale.

# This is a universal problem




- supervised models (classifiers, regressors) learn features that are correlated with the label being annotated.
- unsupervised models (topic models, word embeddings) learn correlations between topics / contexts.

# This is a universal problem




- supervised models (classifiers, regressors) learn features that are correlated with the label being annotated.
- unsupervised models (topic models, word embeddings) learn correlations between topics / contexts.
- **dictionary methods**, while having other limitations, mitigate this problem
  - the researcher intentionally "regularizes" out spurious confounders with the targeted language dimension.
  - helps explain why economists often still use dictionary methods.