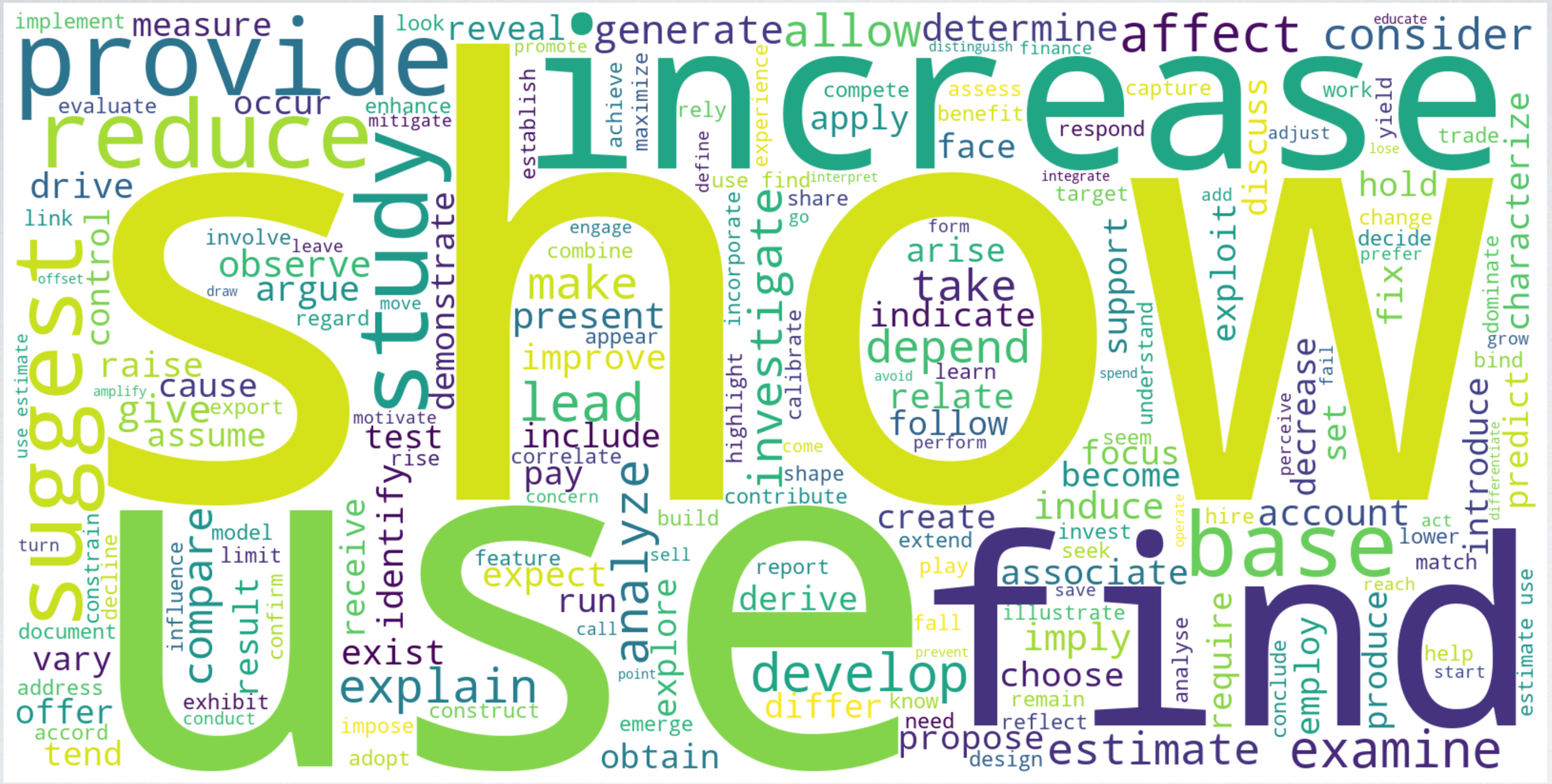
TEXT ANALYSIS FOR ECONOMIC LITERATURE

Presenter: Henry Liu
Advanced Data Analytic Seminar
Feb. 12, 2024

1. **Motivation**

- Economic Papers ≈ 5,000,000
- If one reads 10 papers per day, it takes around 1369 years.
- Previous studies have provided some general insights.
- The report tries to approach this question from Text Analysis perspective:
- **What trends do the economic literature present?**

I. **Motivation**



① Show, find
② Use

Figure 1: Word cloud of verb stems

Data source: European Economic Review

2. **Findings**

- Three Trends:

  ① Trends of Methodology

  ② Trends of Writing Features

  ③ Trends of Academic Equality (authors from different countries)

2.1 Findings - Trends of Methodology

- Model
- Theory
- Effect
- Data

Figure 2a: Share of papers whose abstract has methodology containing the keywords

Data source: EconLit Database

2.1 Findings - Trends of Methodology

- Identification
- Causal Revolution (2000 - today)

Figure 2b: Share of papers whose abstract has methodology containing "identification"

Data source: EconLit Database

2.2 Findings - Trends of Writing Style

- Word count
- Number count
- Capital letter count
- Punctuation count

Figure 3a: Word-level writing feature trends

Data source: EconLit Database

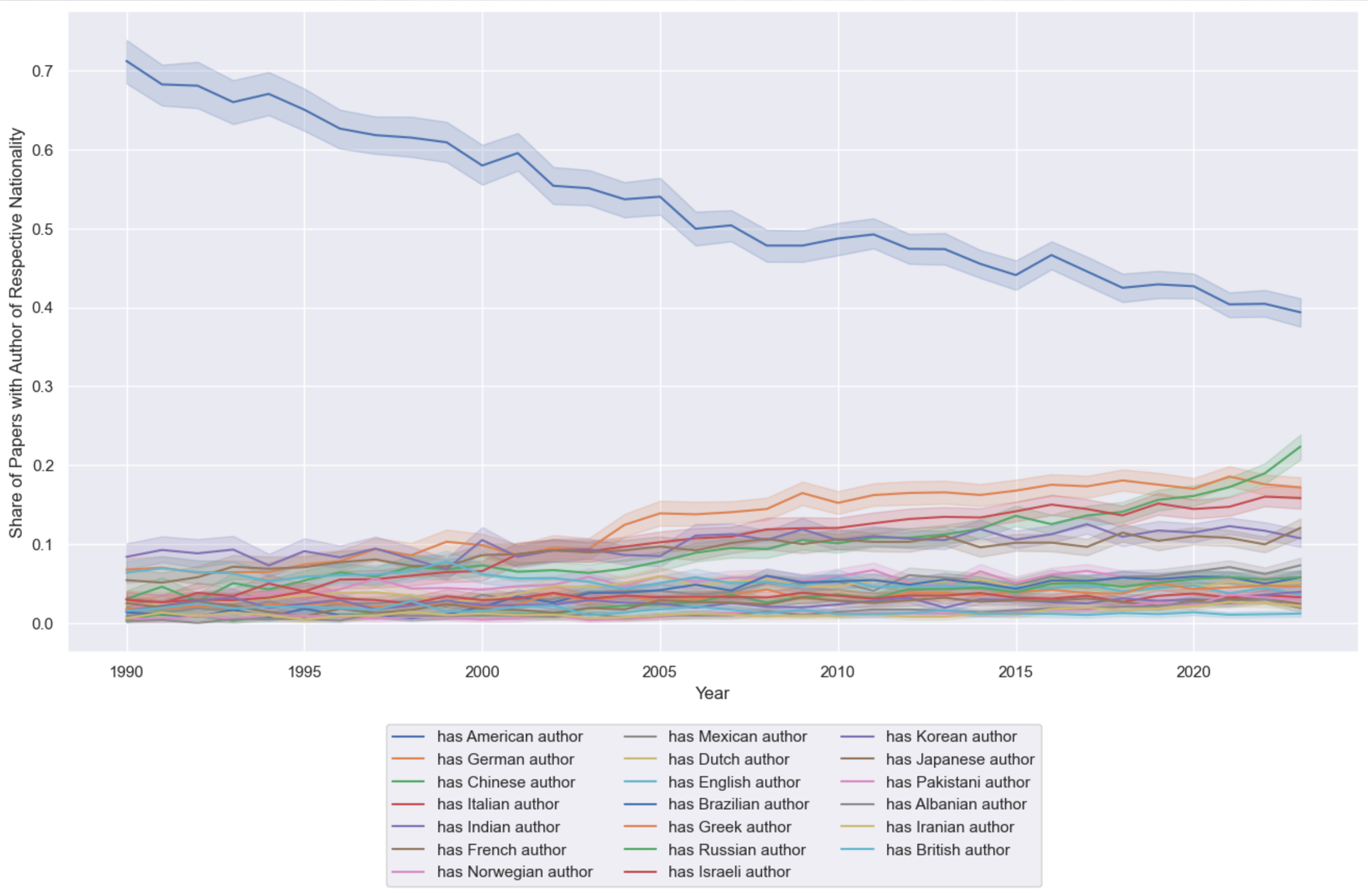
2.2 Findings - Trends of Writing Style

- Sentence count
- Short sentence count
- Long sentence count
- Capital letter per sentence

Figure 3b: Sentence-level writing feature trends

Data source: EconLit Database

2.3 Findings - Trends of Academic Equality



- American:
  70% → 40%
- German
- Chinese
- Italian
- Indian
- French

Figure 4: Share of author nationalities by year

Data source: EconLit Database

- **Take-aways**:

  ① Economics has becoming more data driven, with greater focus on causal inference.

  ② The abstract of an economic paper is 50% longer than it's in 1990s (word count 100 → 150).

  ③ US economists used to dominate most top journals (70%), but researchers with other backgrounds are now the majority.

3. **Methods**

- Models I trained:

  ① Word2Vec Model (for word embedding)

  ② Sentence Classification Model

  ③ Name to Nationality Prediction Model

3.1 Methods - Word2Vec

Figure 2a: Share of papers whose abstract has methodology containing the keywords

Data source: EconLit Database

- **Problem 1:**
- Problem: Single word count is not robust.
- Reasoning: If we can get a bag of similar words, they can do majority voting.
- Question: How to get similar words like below?

Target Word	Similar word 1	Similar word 2	Similar word 3	
model	framework	setup	theory	…
theory	argument	logic	theoretical	…
effect	impact	relationship	consequence	…
data	datum	dataset	microdata	…

3.1 Methods - Word2Vec

Figure 2a: Share of papers whose abstract has methodology containing the keywords

Data source: EconLit Database

- Idea: We can first convert words into **vectors**, and then use cosine similarity to measure.
- Solution: Use **Word2Vec** algorithm, a concurrence-based method, generate similar vectors for those words often appear together.

Target Word	Similar word 1	Similar word 2	Similar word 3	
model	framework	setup	theory	…
theory	argument	logic	theoretical	…
effect	impact	relationship	consequence	…
data	datum	dataset	microdata	…

3.2 Methods - Sentence Classifier

- **Another Problem**:
- In the methodology trend plot, I only want the information about methods.
- Reasoning: I need a model to classify which content is about methods.
- Question: How can I get the data to train such a language model?
- Idea: ChatGPT can do prompt-based annotations.

Figure 2a: Share of papers whose abstract has methodology containing the keywords

Data source: EconLit Database

3.2 Methods - Sentence Classifier

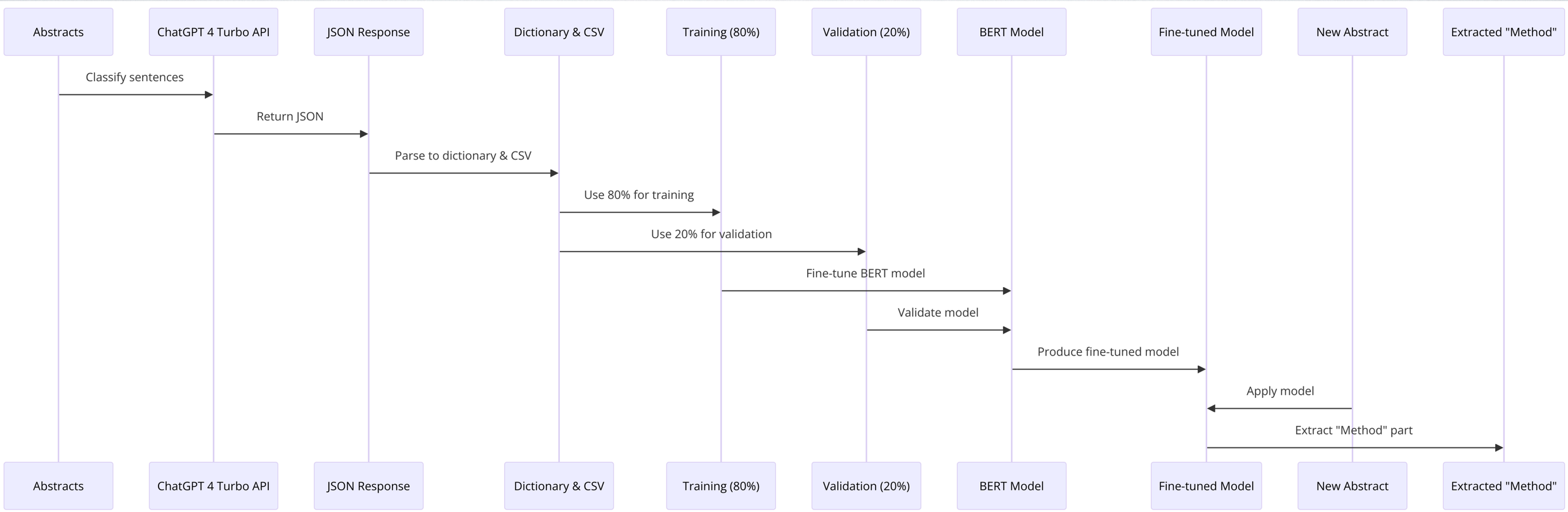


Figure 5: Sentence classifier training and application procedure diagram

3.3 Methods - Name to Nationality Predictor

- The name to nationality model utilizes a similar idea.
- It's also a fine-tuned version of BERT.
- The dataset is prepared by Kyubyong who scraped wikipedia and extract name and nationality from the content.

Source: https://github.com/Kyubyong/name2nat

4. **Summary**

- This report presents some **descriptive insights**, about the trends in **economic literatures**, from the perspective of **text analysis**.
- Drawbacks:
  - The people in Wikipedia are not balanced, this may cause bias in nationality prediction.
  - Some sentences are multi-categorical, could use fuzzy classifier.
  - No causal inference.
- Future directions:
  - Semantic correlation analysis based on word2vec embedding similarity.

5. **References**

Ash, E., & Hansen, S. (2023). Text algorithms in economics. *Annual Review of Economics*. https://doi.org/10.1146/annurev-economics-082222-074352

Card, D., & DellaVigna, S. (2013). Nine Facts about Top Journals in Economics. *Journal of Economic Literature, 51*(1), 144-161. https://doi.org/10.1257/jel.51.1.144

Cardoso, A. R., Guimarães, P., & Zimmermann, K. F. (2010). Trends in Economic Research: An International Perspective. *Kyklos, 63*(4), 479-494. https://doi.org/10.1111/j.1467-6435.2010.00484.x

Hudson, J. (1996). Trends in multi-authored papers in economics. *Journal of Economic Perspectives, 10*(3), 153-158. https://doi.org/10.1257/jep.10.3.153

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review, 84*(5), 905-949. https://doi.org/10.1177/0003122419877135

Paldam, M. (2021). Methods used in economic research: An empirical study of trends and levels. *Economics, 15*(1), 28–42. https://doi.org/10.1515/econ-2021-0003

Park, K. (2020). name2nat: a Python package for nationality prediction from a name. *GitHub repository*. https://github.com/Kyubyong/name2nat