



# **Turning Web-Scale Texts to Knowledge: Transferring Pretrained Representations to Text Mining Applications**

**Yu Meng, Jiaxin Huang, Yu Zhang, Jiawei Han**

Department of Computer Science  
University of Illinois at Urbana-Champaign  
April 30, 2023

Tutorial Website:



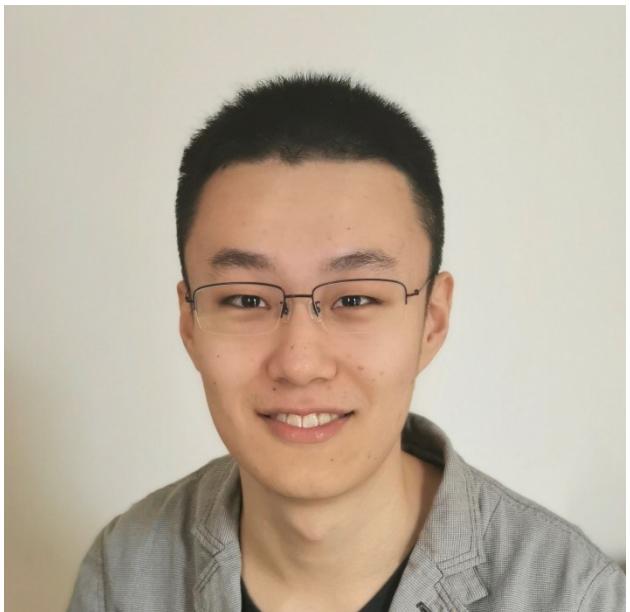
# Estimated Timeline for This Tutorial

---

- ❑ Introduction: **10 mins (11:00-11:10 Han)**
- ❑ Part I: Pretrained Language Models: **15 mins (11:10-11:25 Meng)**
- ❑ Part II: Embedding-Driven Topic Discovery: **35 mins (11:25-12:00 Meng & Huang)**
- ❑ Part III: Weakly-Supervised Text Classification: **25 mins (12:00-12:25 Zhang)**
- ❑ Summary and Future Directions: **5 mins (12:25-12:30 Han)**

# About Instructors

---



- Yu Meng**  
Ph.D. Candidate, UIUC
- Recipient of 2021  
Google PhD Fellowship  
in Structured Data and  
Database Management

- Jiaxin Huang**  
Ph.D. Candidate, UIUC
- Recipient of 2021  
Microsoft PhD  
Fellowship

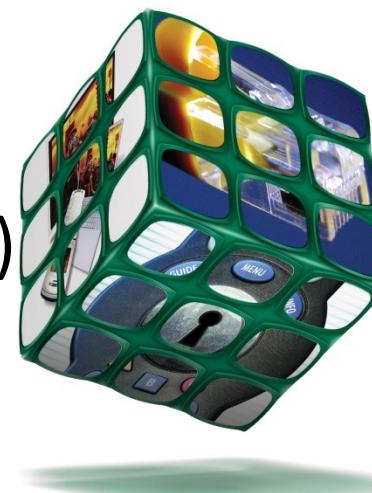
- Yu Zhang**  
Ph.D. Candidate, UIUC
- Recipient of 2022 Yunni  
and Maxine Pao  
Memorial Fellowship

- Jiawei Han**  
Michael Aiken Chair  
Professor at UIUC
- ACM SIGKDD  
Innovation Award  
Winner (2004)

# Over 80% of Big (Web) Data is Unstructured Text Data

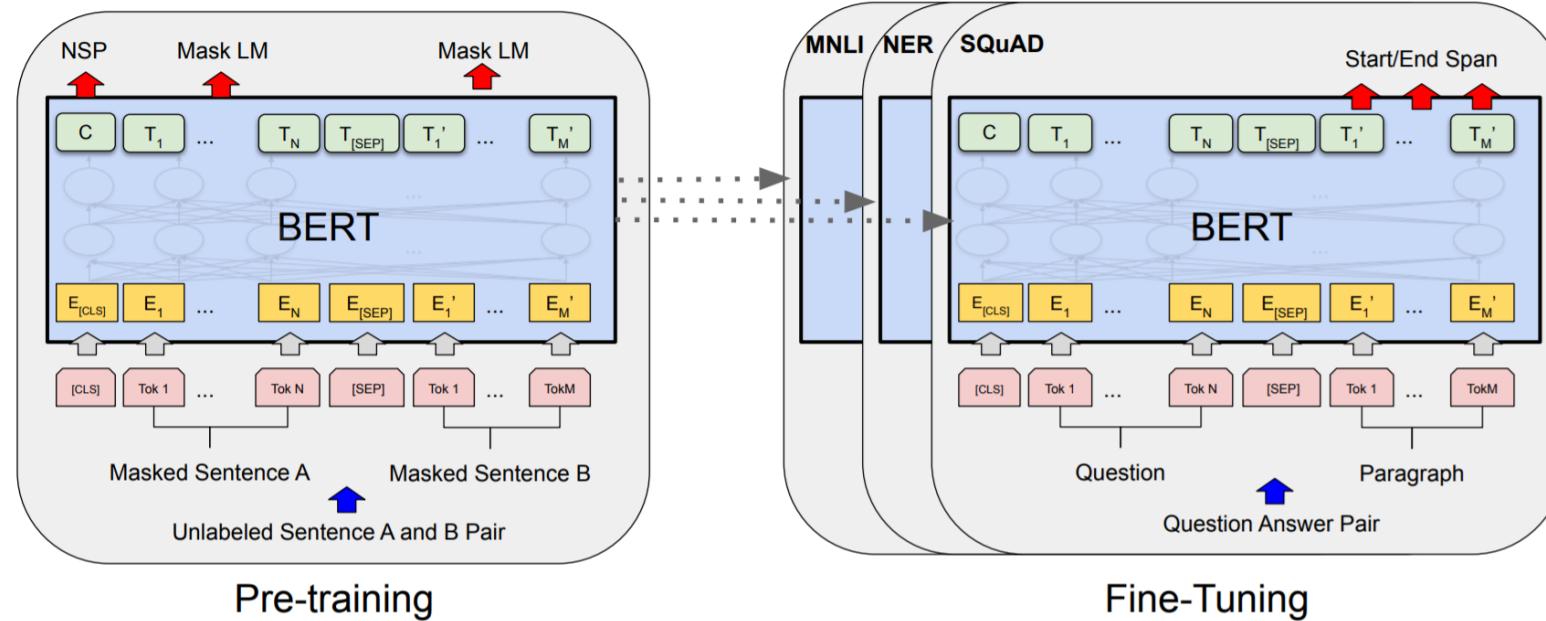
---

- Ubiquity of big unstructured, text data
  - **Big Data:** Over 80% of our data is from text (e.g., news, papers, social media): unstructured/semi-structured, noisy, dynamic, inter-related, high-dimensional, ...
- How to mine/analyze such big data systematically?
  - **Text Representation** (i.e., computing vector representations of words/phrases/sentences)
  - **Basic Structuring** (i.e., phase mining & transforming unstructured text into structured, typed entities/relationships)
  - **Advanced Structuring:** Discovering Hierarchies/taxonomies, exploring in multi-dimensional space



# Contextualized Text Representation: Language Models

- Language models are pre-trained on large-scale general-domain corpora to learn universal/generic language representations that can be transferred to downstream tasks via fine-tuning

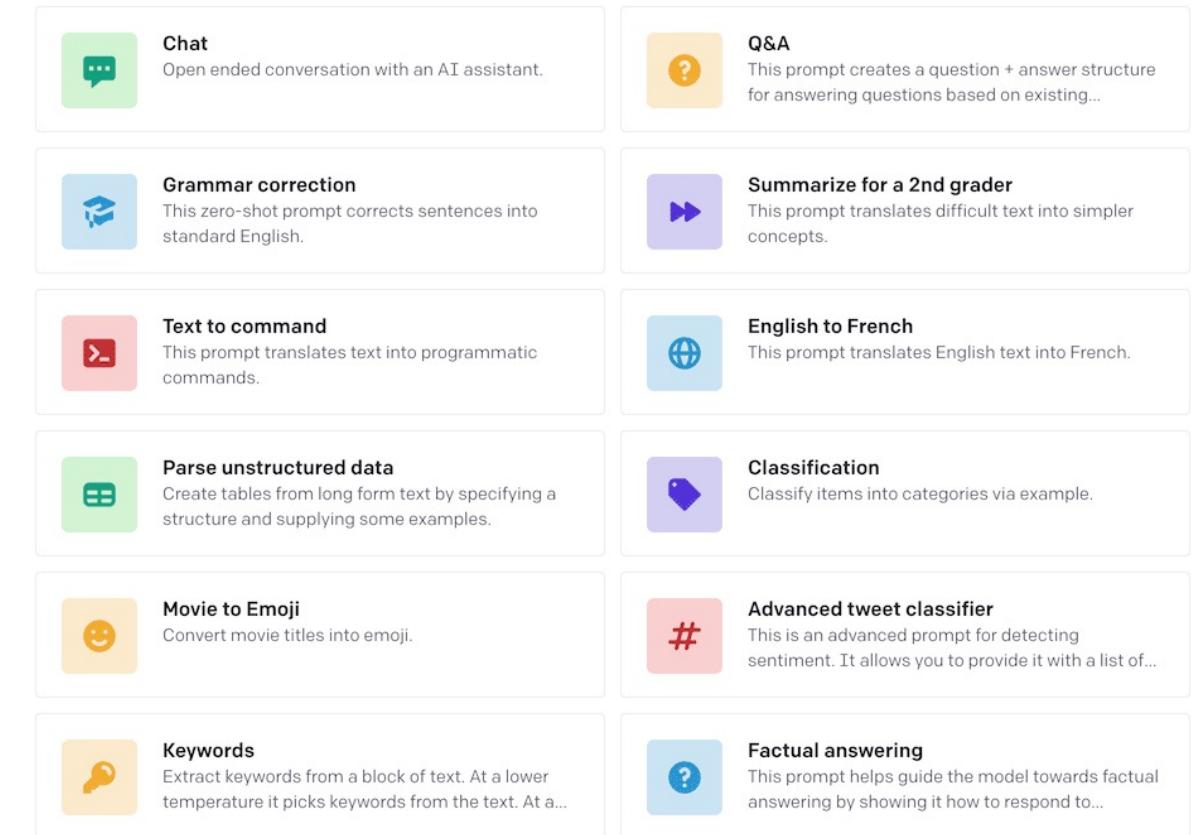
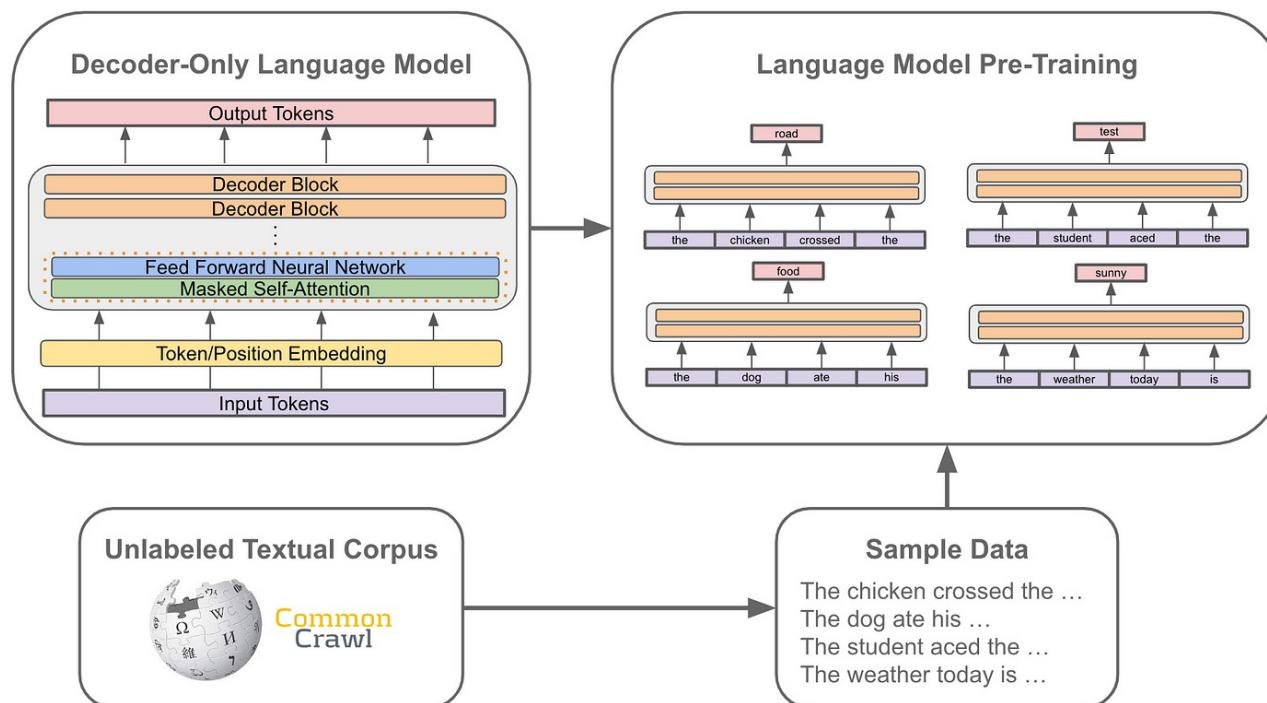


Unsupervised/Self-supervised;  
On large-scale general domain corpus

Task-specific supervision;  
On target corpus

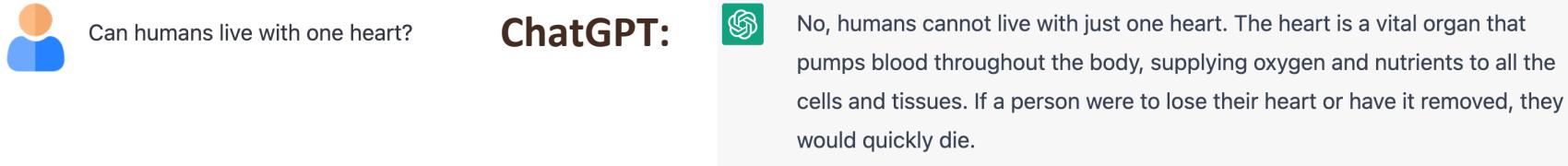
# Generative Large Language Models: The GPT Series

- GPT models: Large language models (LLMs) trained for text generation
- Applicable to a wide range of tasks

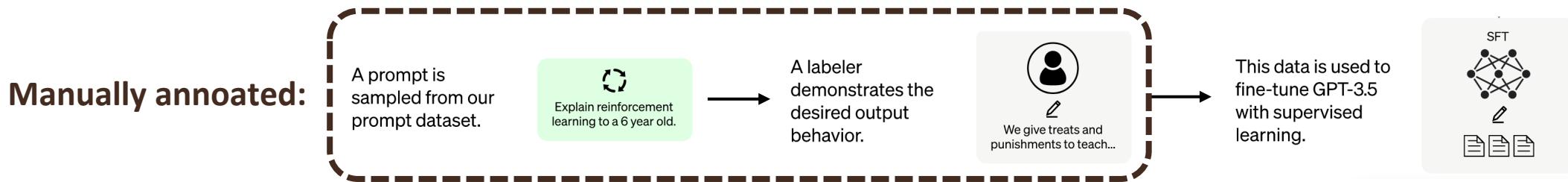


# Challenges of Large Language Models

- ❑ Not factually guaranteed: May generate wrong information

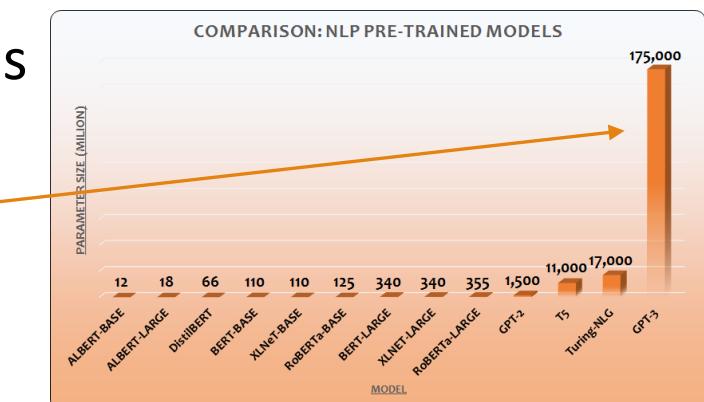


- ❑ Heavy supervision required: Trained on massive annotated data



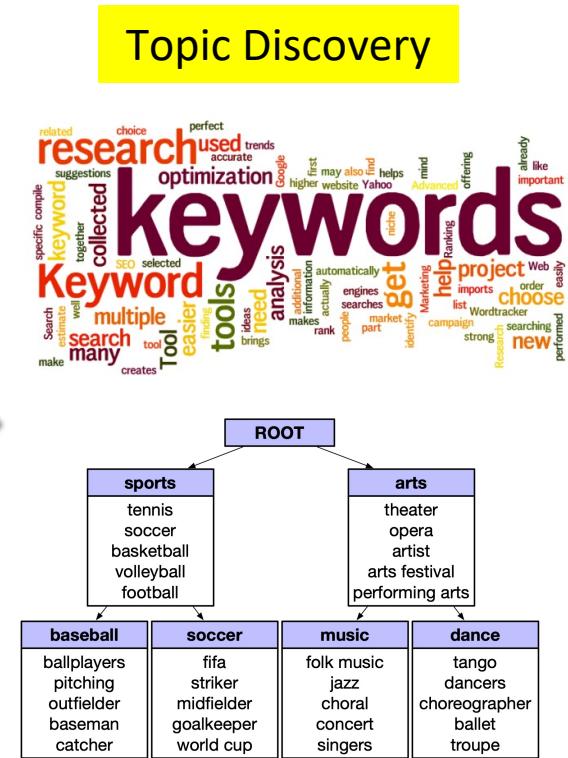
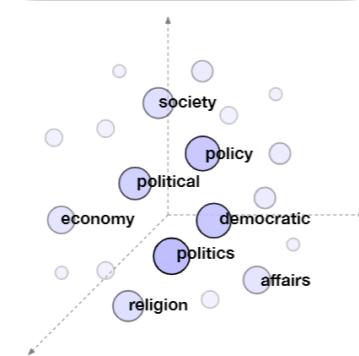
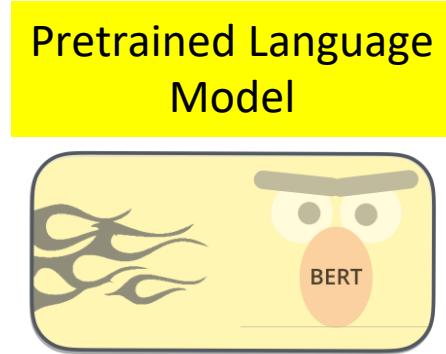
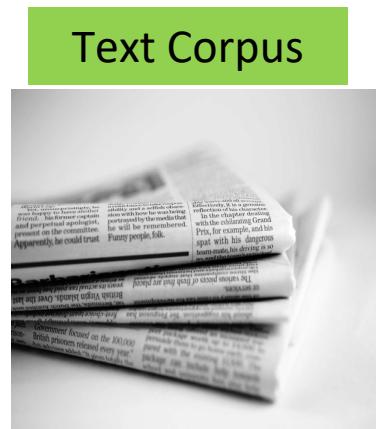
- ❑ Costly & Inefficient: Too large to be used in many applications

GPT3 has 175B parameters (ChatGPT/GPT-4 may have more!)

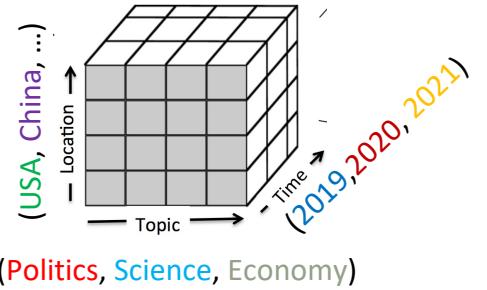


# Towards Factual, Automatic, and Efficient Text Mining

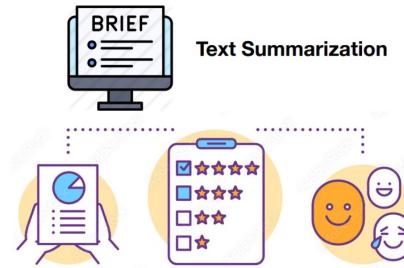
- Understand and Extract Information from Massive Text Corpora
- Organize and Analyze Information using **Multidimensional** Text Analysis



Weakly-Supervised Text Classification



(Politics, Science, Economy)

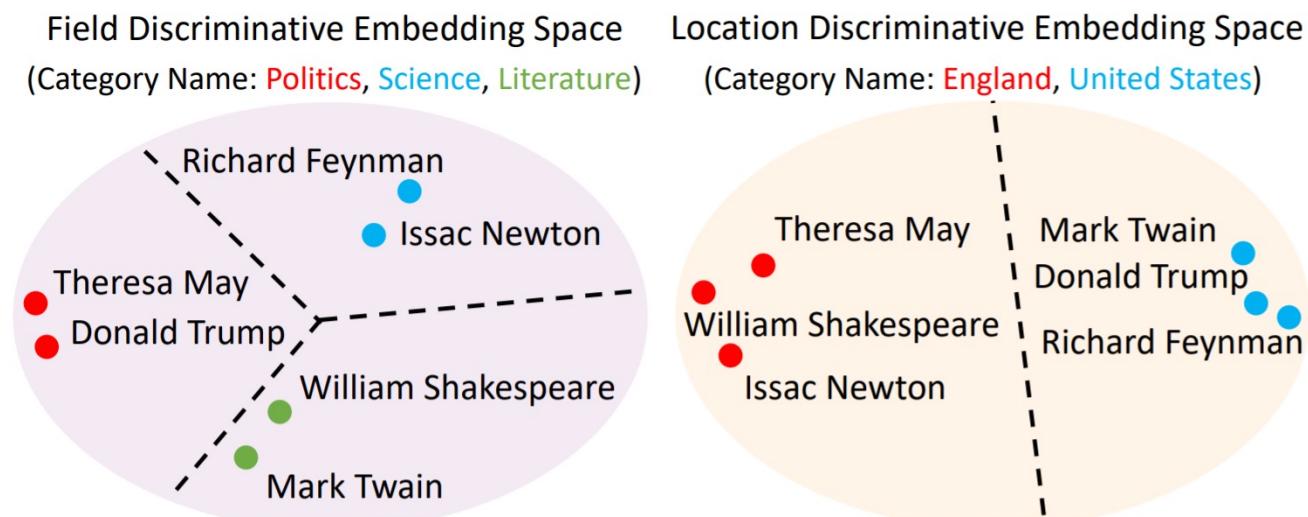


Sentiment analysis

Advanced Text Mining Applications  
(Sentiment Analysis, Summarization)

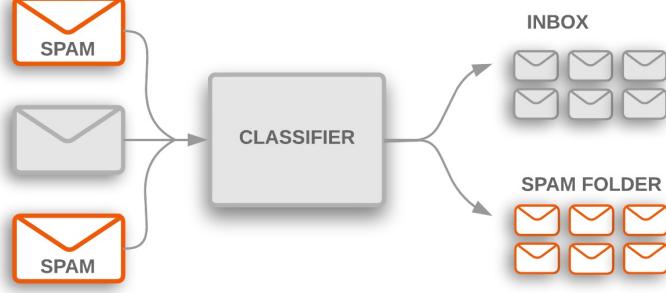
# Overview of Seed-Guided Topic Discovery

- ❑ Mining topic structures from massive corpora is crucial for text understanding
- ❑ The same set of concepts/topics/entities may be organized via different aspects
- ❑ How to incorporate user interests/preferences?
  - ❑ Manually labeling documents requires non-trivial human efforts and is hard to scale
  - ❑ Use seed words instead to guide topic discovery!

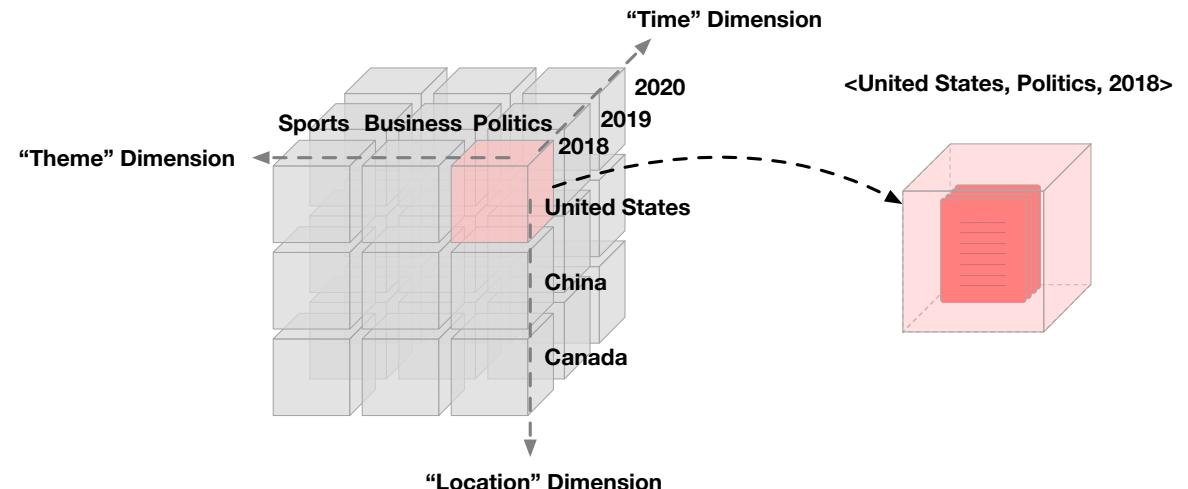


# Overview of Weakly-Supervised Text Classification

- Text classification is a core task for document organization and understanding
- Text classifiers are typically trained on massive manually-labeled data
- How to build text classifiers with fewer human annotations?
- Weakly-supervised text classification: Use label names & keywords as weak supervision



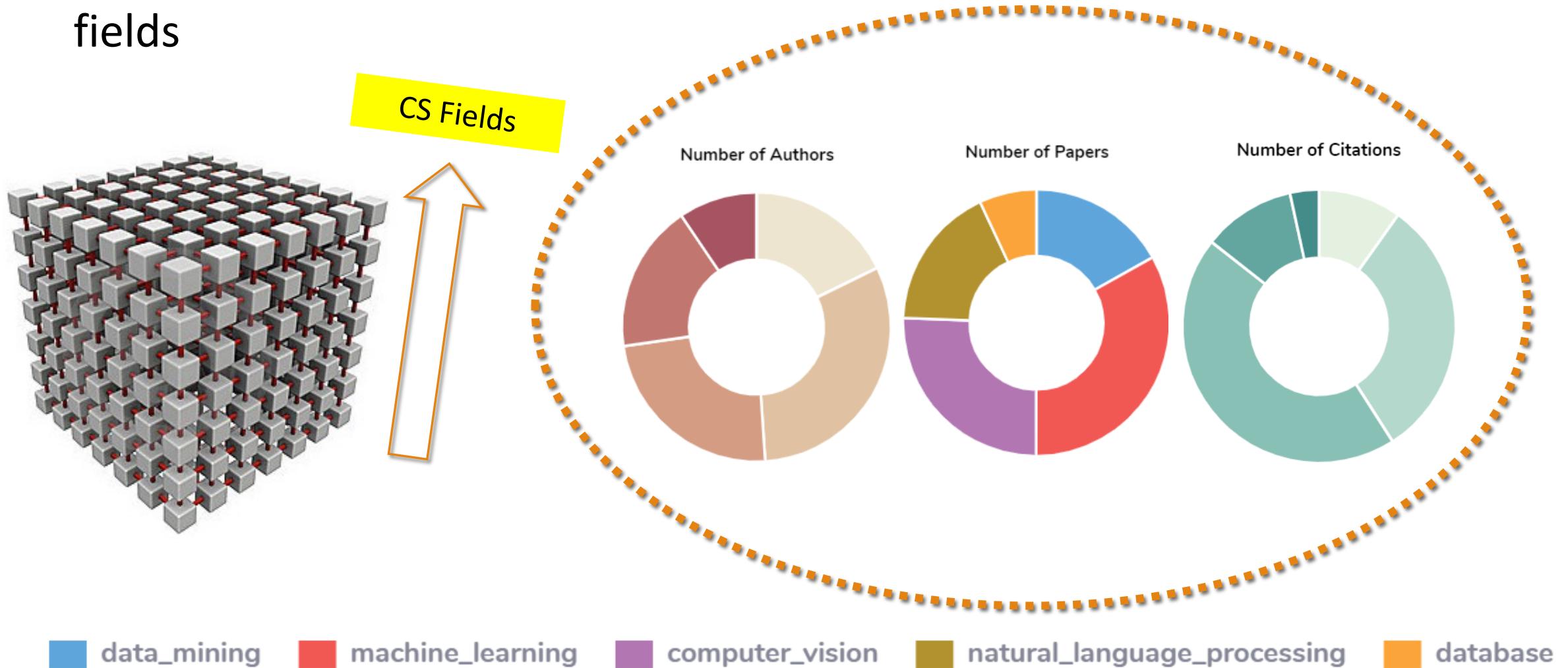
Text classifiers



Weakly-supervised text classification only leverages label names as supervision

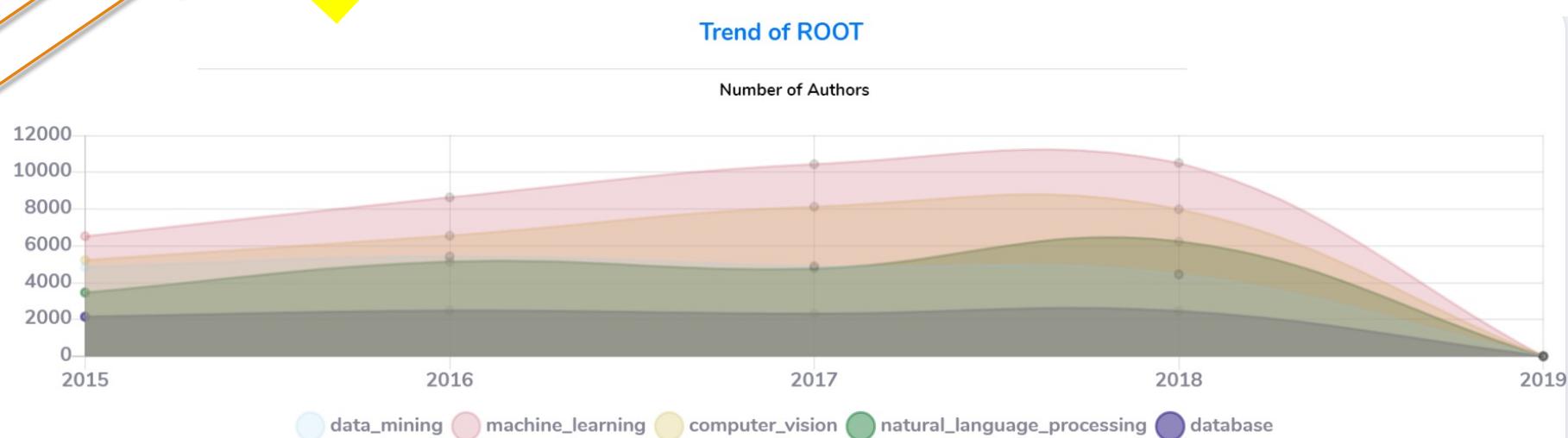
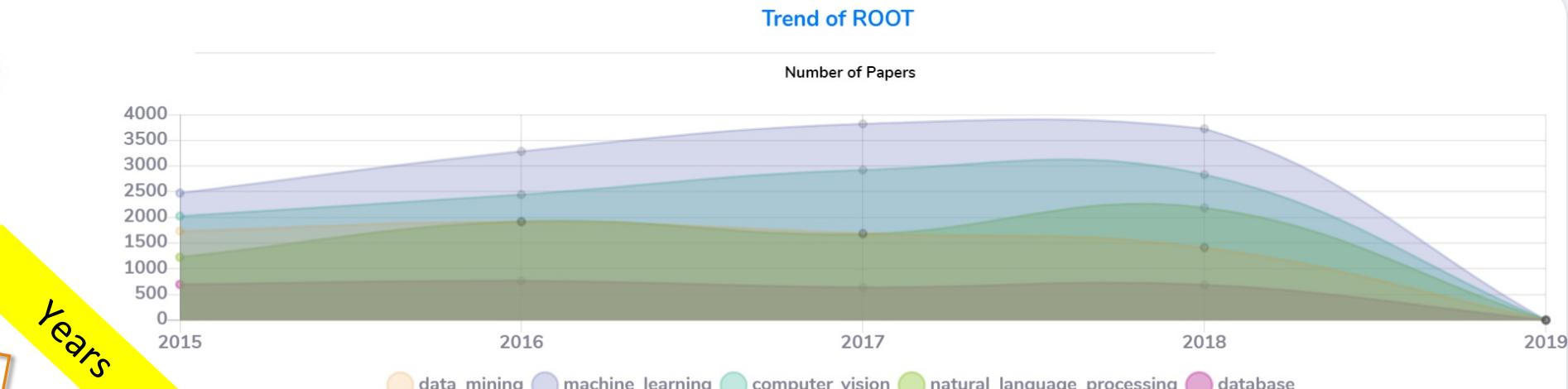
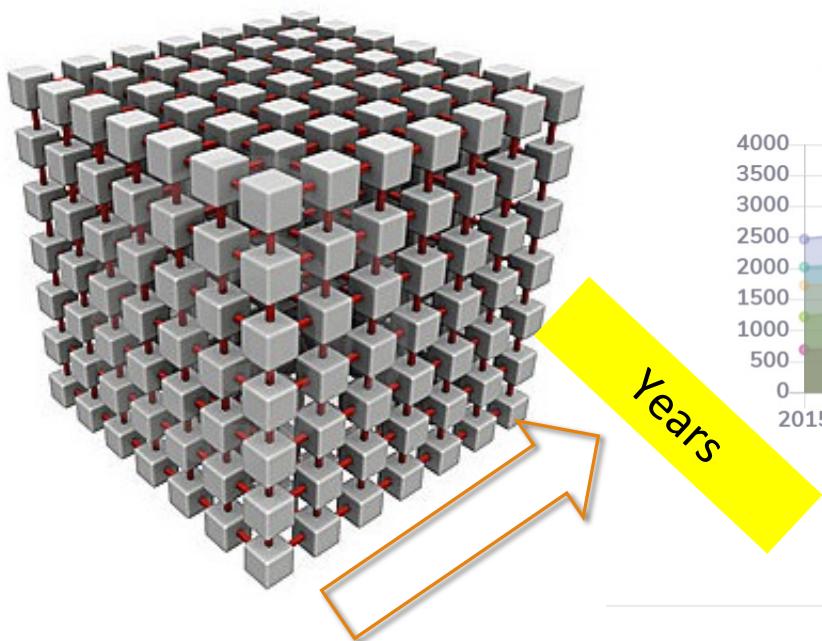
# Application: DBLP—Automatic Paper Categorization

- Multidimensional text categorization and exploration across different CS fields

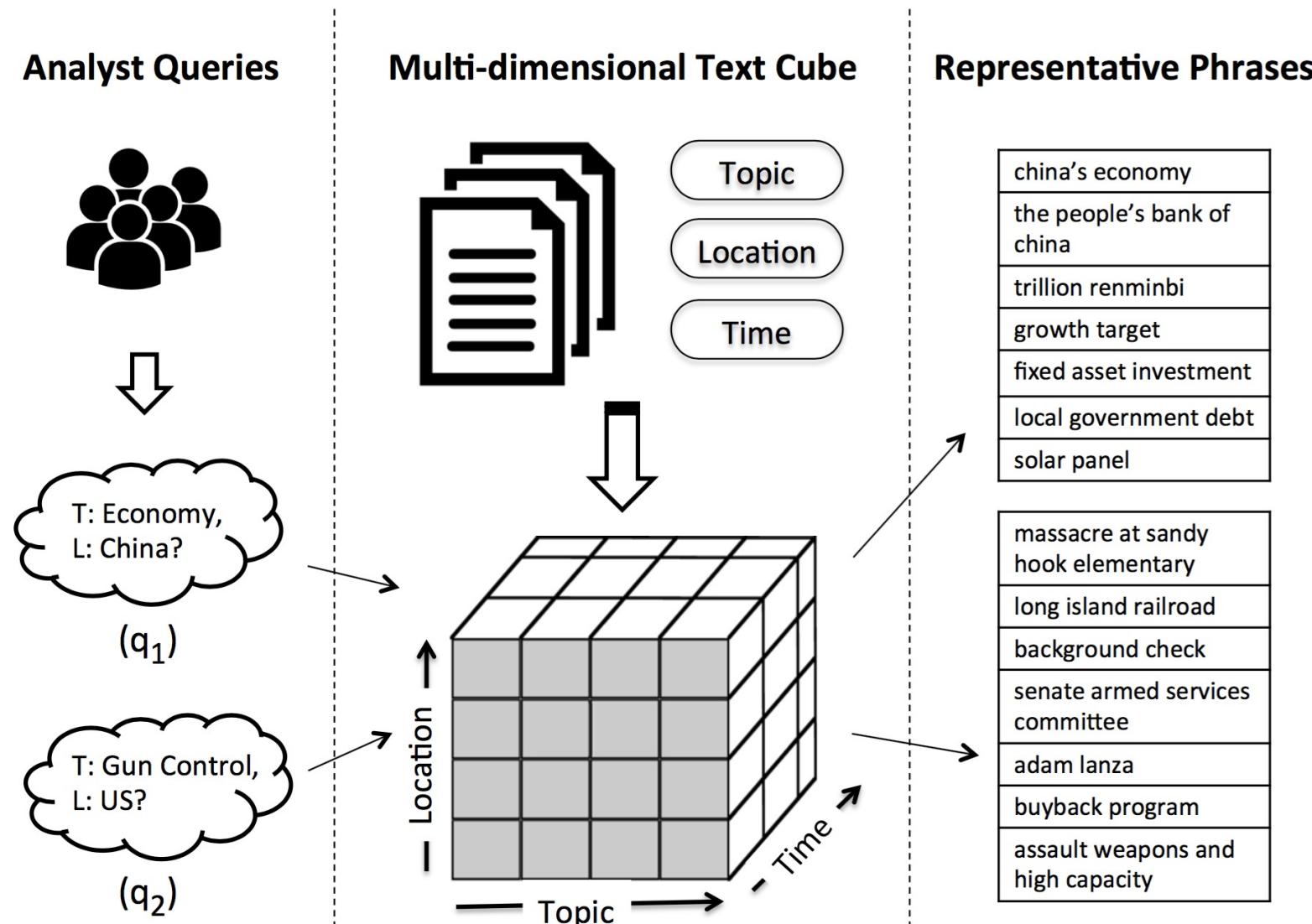


# Application: DBLP—Trending Analysis

- Trending analysis on CS field development



# Application: Comparative Summarization



# Tutorial Outline

---

- Introduction
- Part I: A Brief Introduction to Pretrained Language Models
- Part II: Embedding-Driven Topic Discovery
- Part III: Weakly-Supervised Text Classification
- Summary and Future Directions

# Our Roadmap of This Tutorial

Text Corpus

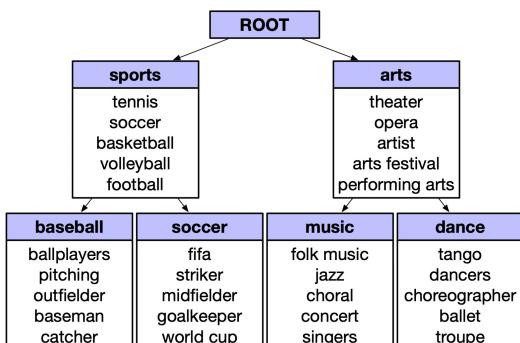
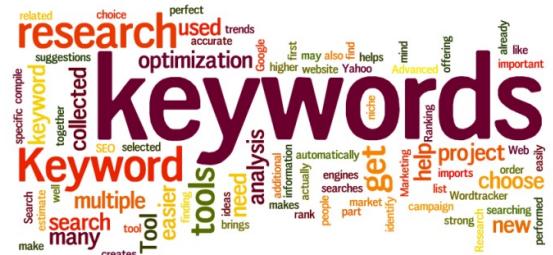


Part I: Pretrained Language Model



Existing KB

Part II: Topic Discovery



Part III: Weakly-Supervised Text Classification

