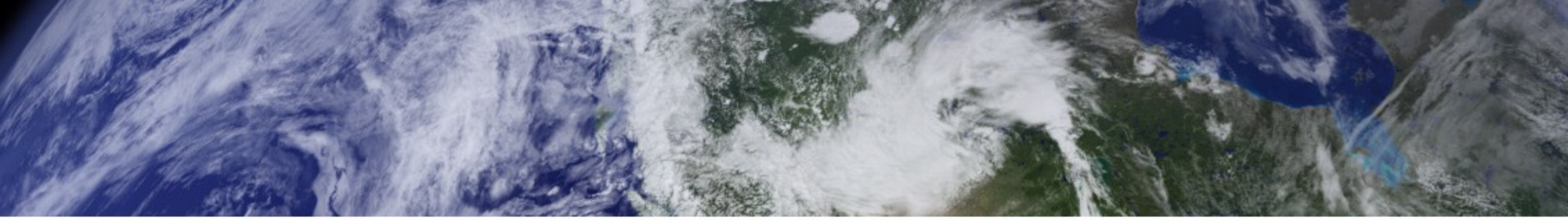
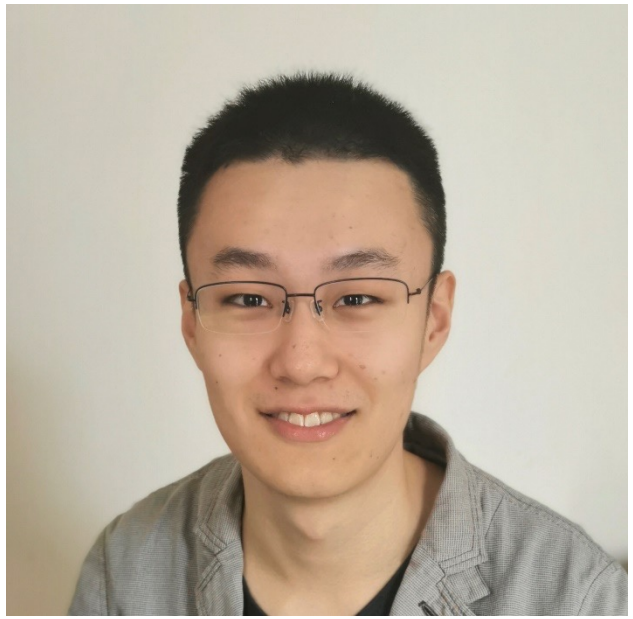


Pre-Trained Language Representations for Text Mining

Yu Meng, Jiaxin Huang, Yu Zhang, Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign
February 23, 2022

About Instructors



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

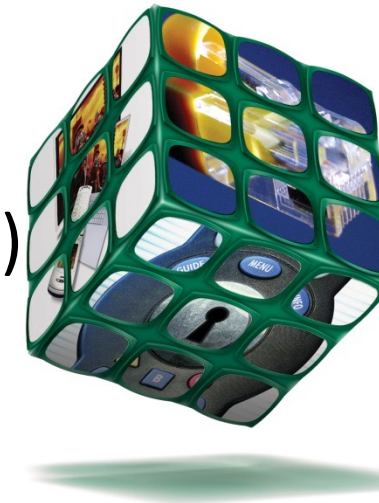
- Jiaxin Huang
  Ph.D. Candidate, UIUC
- Recipient of 2021 Microsoft PhD Fellowship

- Yu Zhang
  Ph.D. Candidate, UIUC
- Recipient of WWW'18 Best Poster Award Honorable Mentioning

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

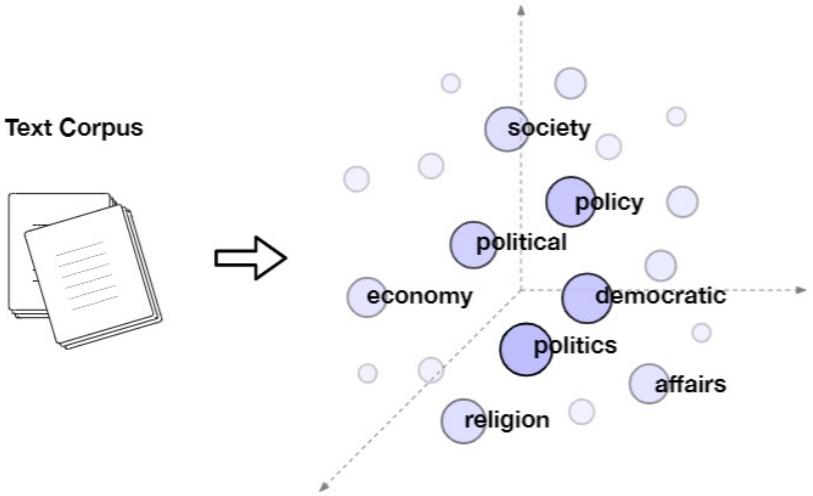
Over 80% of Big Data is Unstructured Text Data

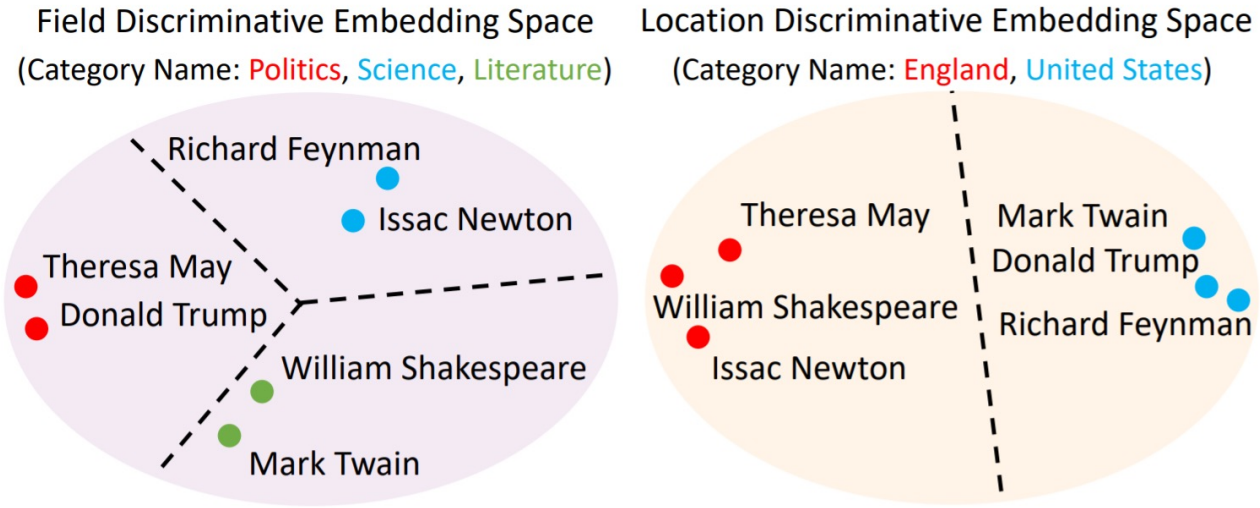
- Ubiquity of big unstructured, text data
  - Big Data: Over 80% of our data is from text (e.g., news, papers, social media): unstructured/semi-structured, noisy, dynamic, inter-related, high-dimensional, …
- How to mine/analyze such big data systematically?
  - **Text Representation** (i.e., computing 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



Text Representation: Embeddings & Language Models

- Word embeddings map words into a vector space which reflects semantic similarity



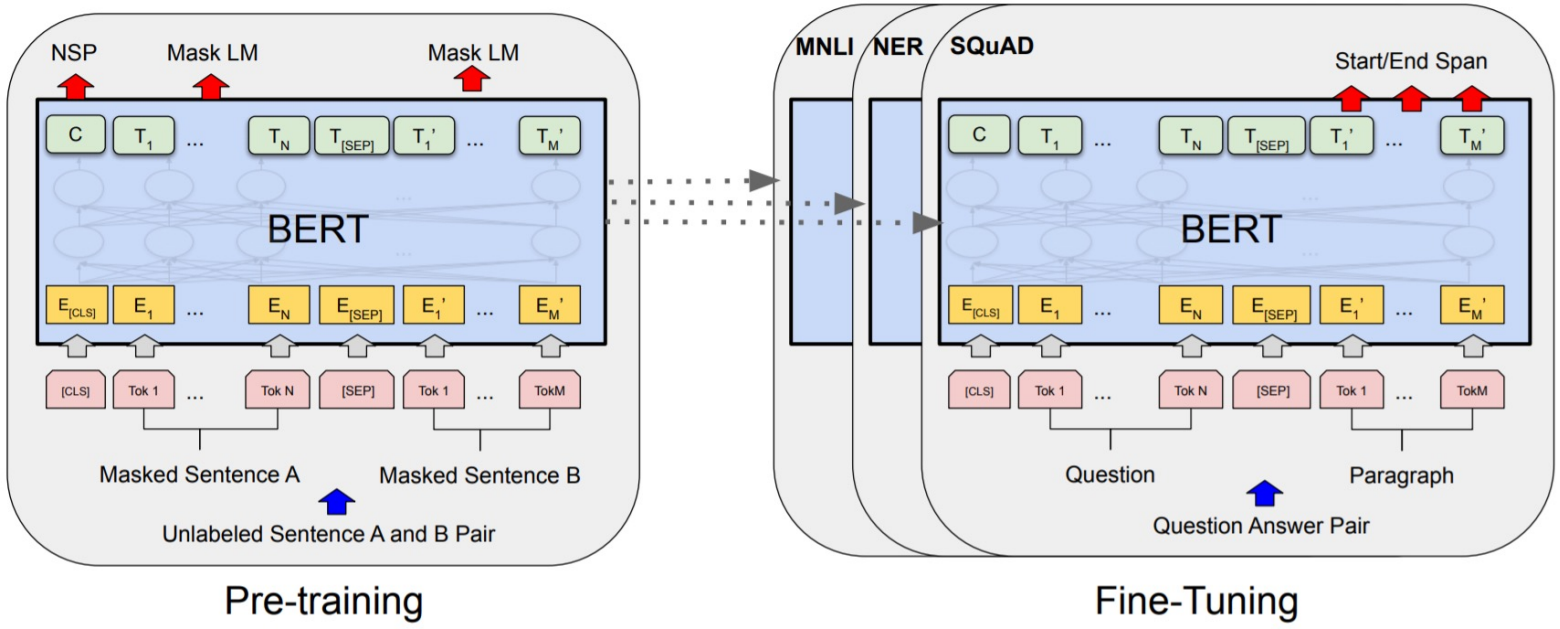


Unsupervised word embeddings:
learned from corpus statistics

(Weakly-)supervised word embeddings:
learned from corpus statistics & user guidance

Text Representation: Embeddings & 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

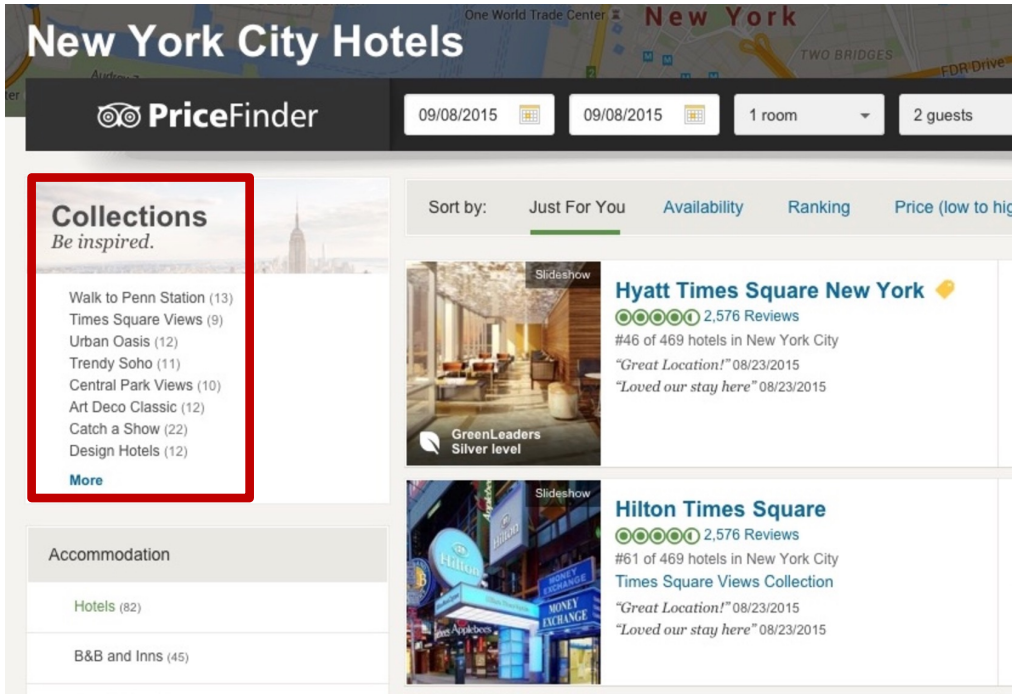


Unsupervised/Self-supervised;
On large-scale general domain corpus

Task-specific supervision;
On target corpus

Basic Structuring: Phrase Mining and Information Extraction

Example: Finding "Interesting Hotel Collections"



Grouping hotels based on structured facts extracted from the review text

Different Dimensions of Information

Features for "**Catch a Show**" collection

1	broadway shows
2	beacon theater
3	broadway dance center
4	broadway plays
5	david letterman show
6	radio city music hall
7	theatre shows

Features for "**Near The High Line**" collection

1	high line park
2	chelsea market
3	highline walkway
4	elevated park
5	meatpacking district
6	west side
7	old railway

http://engineering.tripadvisor.com/using-nlp-to-find-interesting-collections-of-hotels/

Basic Structuring: Automated Named Entity Recognition & Typing

Angiotensin-converting enzyme 2 ***GENE_OR_GENOME*** *(* *ACE2* ***GENE_OR_GENOME*** *) as a SARS-CoV-2* ***CORONAVIRUS*** *receptor* ***CHEMICAL****: molecular mechanisms and potential therapeutic target.*

SARS-CoV-2 **CORONAVIRUS** has been sequenced [3] . A phylogenetic **EVOLUTION** analysis [3 , 4] found a bat **WILDLIFE** origin for the SARS-CoV-2 **CORONAVIRUS** . There is a diversity of possible intermediate hosts **NORP** for SARS-CoV-2 **CORONAVIRUS** , including pangolins **WILDLIFE** , but not mice **EUKARYOTE** and rats **EUKARYOTE** [5] . There are many similarities of SARS-CoV-2 **CORONAVIRUS** with the original SARS-CoV **CORONAVIRUS** . Using computer modeling , Xu et al **PERSON**. [6] found that the spike proteins **GENE_OR_GENOME** of SARS-CoV-2 **CORONAVIRUS** and SARS-CoV **CORONAVIRUS** have almost identical 3-D structures in the receptor binding domain that maintains Van der Waals forces **PHYSICAL_SCIENCE** . SARS-CoV spike proteins **GENE_OR_GENOME** has a strong binding affinity **DISEASE_OR_SYNDROME** to human ACE2 **GENE_OR_GENOME** , based on biochemical interaction studies and crystal structure analysis [7] . SARS-CoV-2 **CORONAVIRUS** and SARS-CoV spike proteins **GENE_OR_GENOME** share identity in amino acid sequences and , importantly, the SARS-CoV-2 **CORONAVIRUS** and SARS-CoV spike proteins **GENE_OR_GENOME** have a high degree of homology [6, 7] . Wan et al **PERSON**. [4] reported that residue 394 **CARDINAL** (glutamine **CHEMICAL**) in the SARS-CoV-2 **CORONAVIRUS** receptor-binding domain ….

Automatic Recognition of 75 entity types in COVID-19 Biomedical Literature (ISMB 2020 Demo)

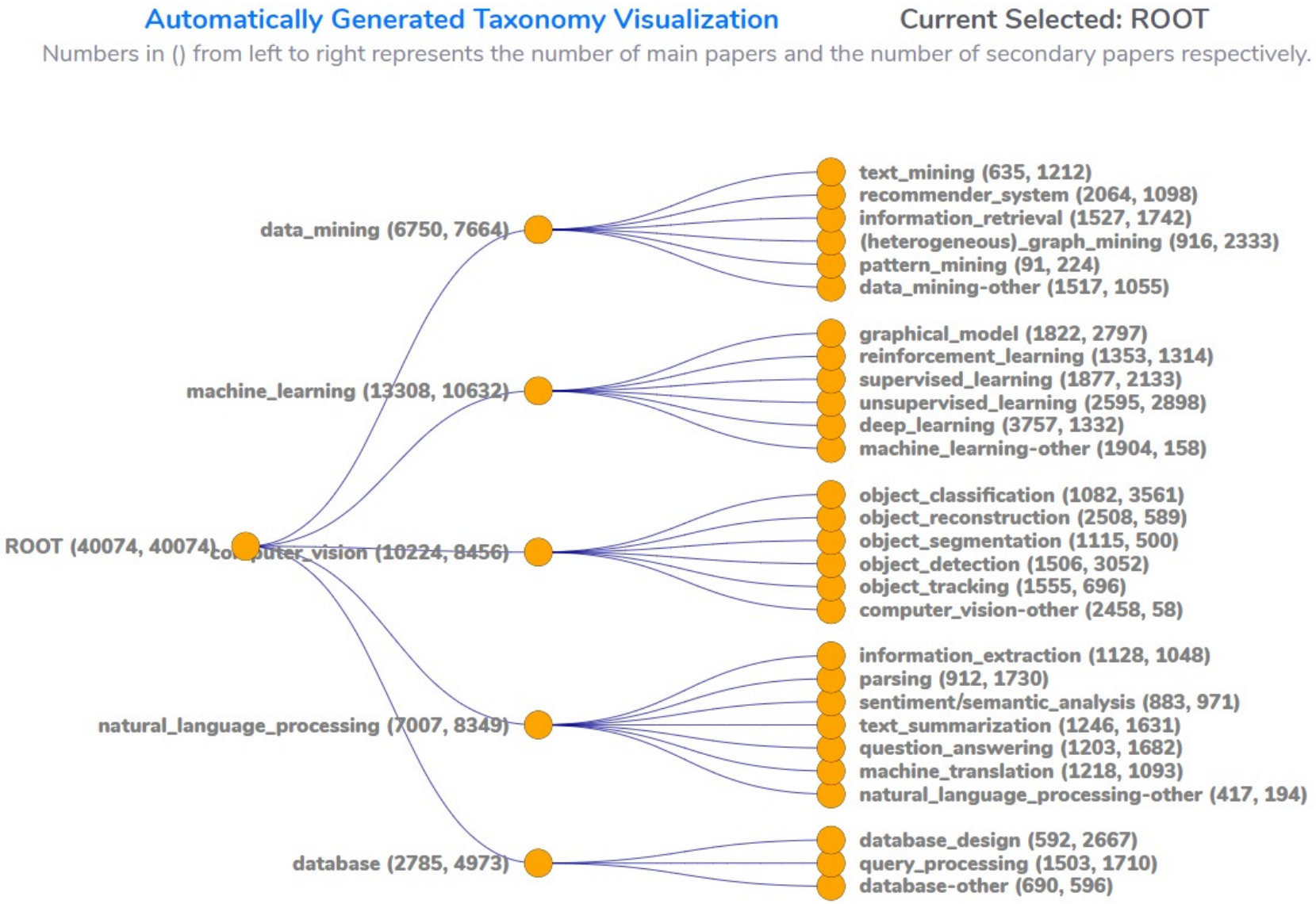
Adv. Structuring: Multidimensional Nature of Texts

- The same document can naturally describe things across multiple dimensions

- Example:
  - A technical review may cover
    - Brands
    - Products
    - Aspects
    - Years
    - …

Apple's 10th anniversary iPhone X sets a new gold standard for the next decade of iPhones. Coming hot on the heels of the iPhone 8and iPhone 8 Plus, the iPhone X stole the show despite sharing nearly identical internal hardware. The X (pronounced "ten," like the Roman numeral) is a beautiful, modern sculpture, and iPhone owners finally have a reason to show off their phones again.
As we're now about four months from Apple's next iPhone launch, we're revisiting the iPhone X to see if it's still worth the high price tag.
… …

Advanced Structuring: Automatic Taxonomy Generation



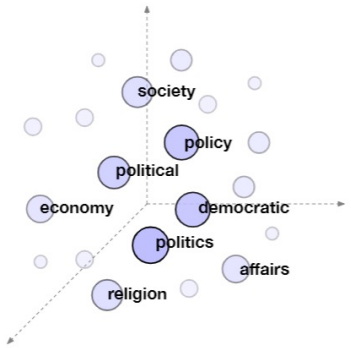
Adv. Structuring: Multi-Dimensional Text Cube Construction

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

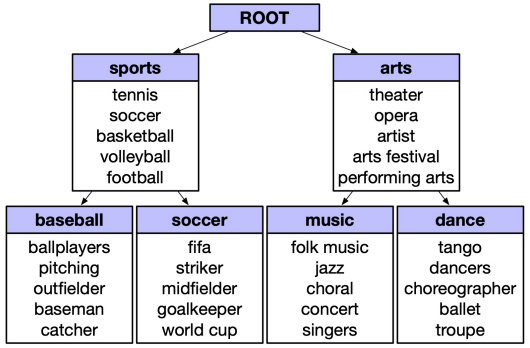
Text Corpus

Existing KB

Pre-Trained Language Model

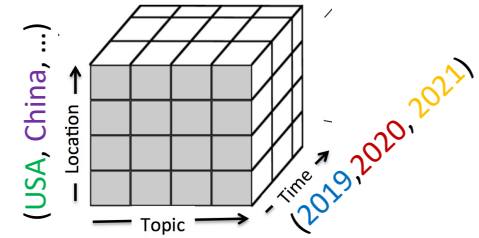


Text Mining Fundamentals (Phrase/Entity Mining, Taxonomy)

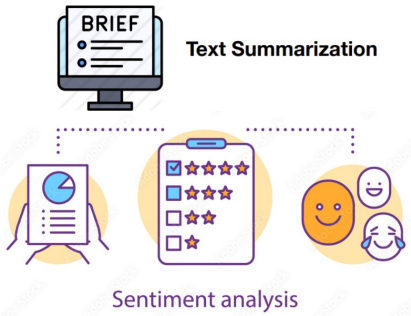


Topic Discovery

Weakly-Supervised Text Classification



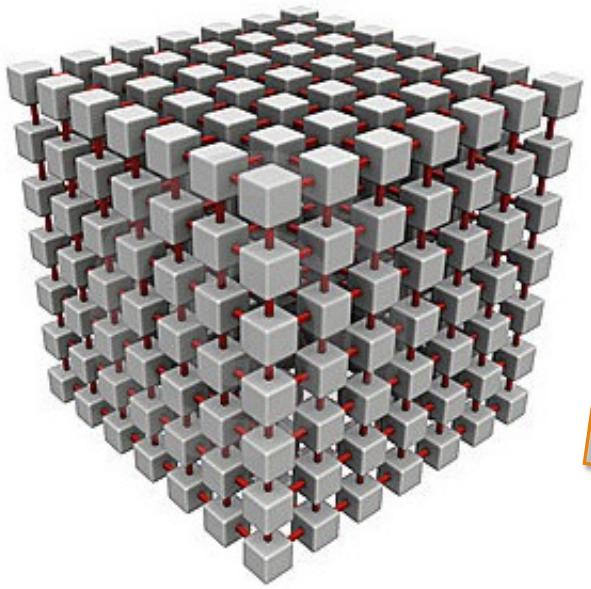
(Politics, Science, Economy)



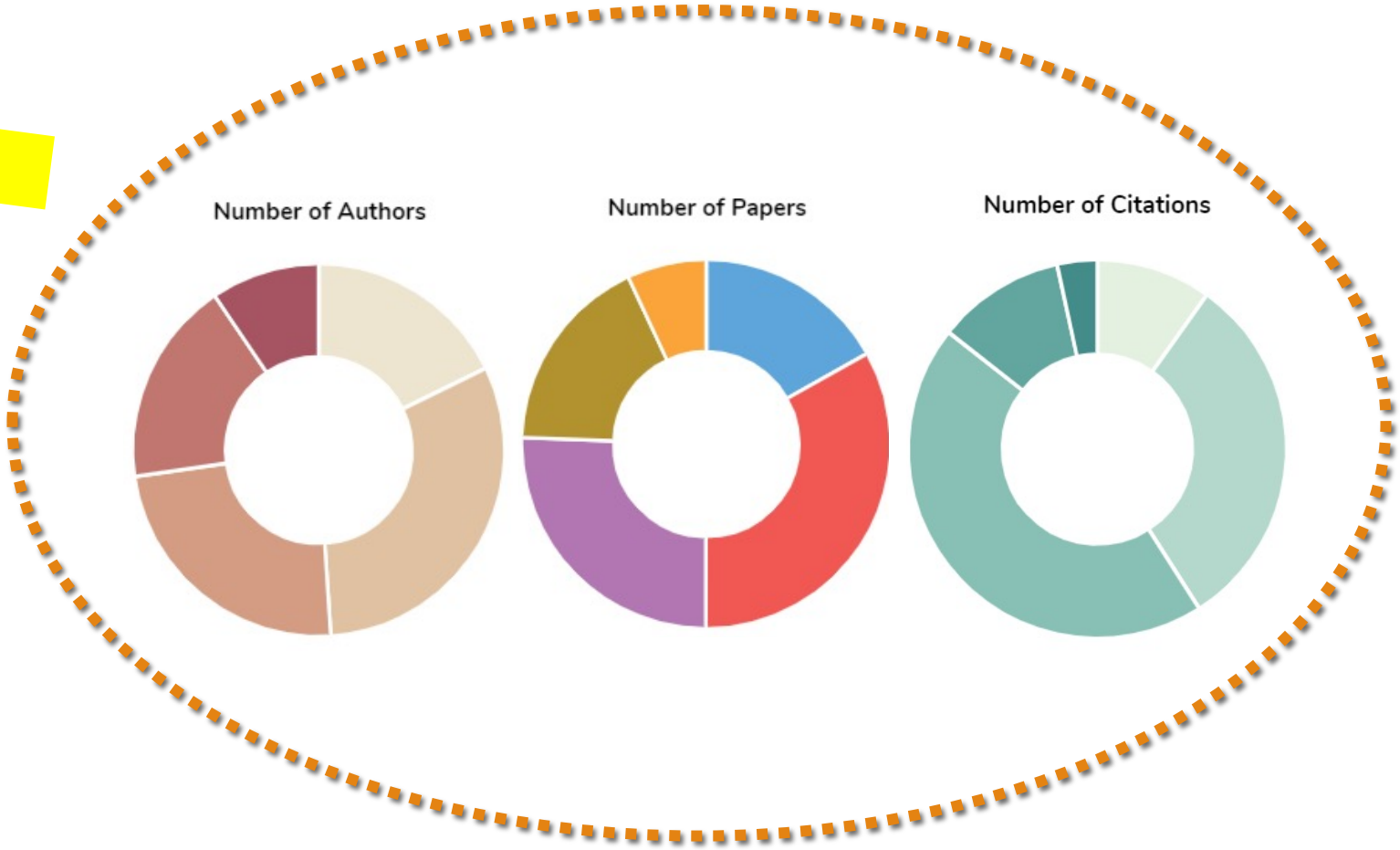
Advanced Text Mining Applications (Sentiment Analysis, Summarization)

Application: DBLP—Automatic Paper Categorization

- Multidimensional text categorization and exploration across different CS fields

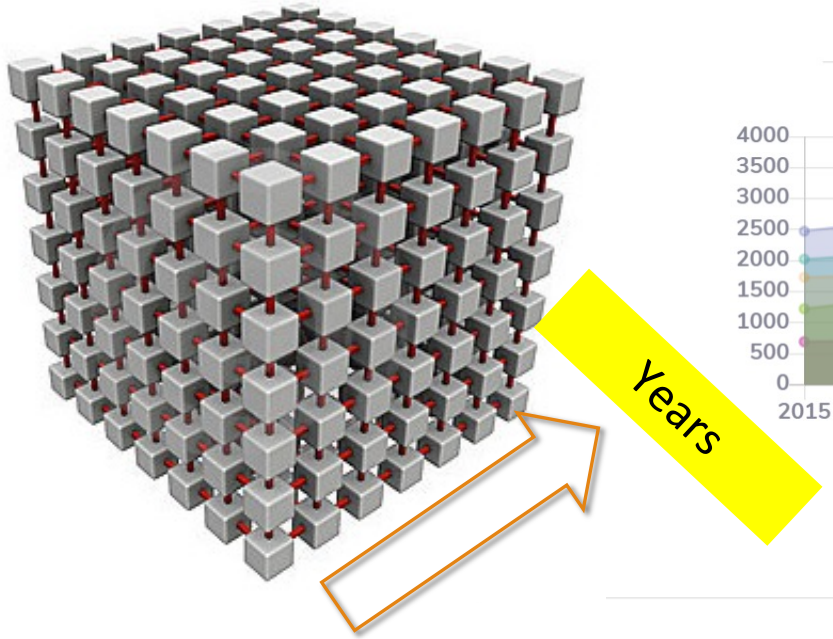


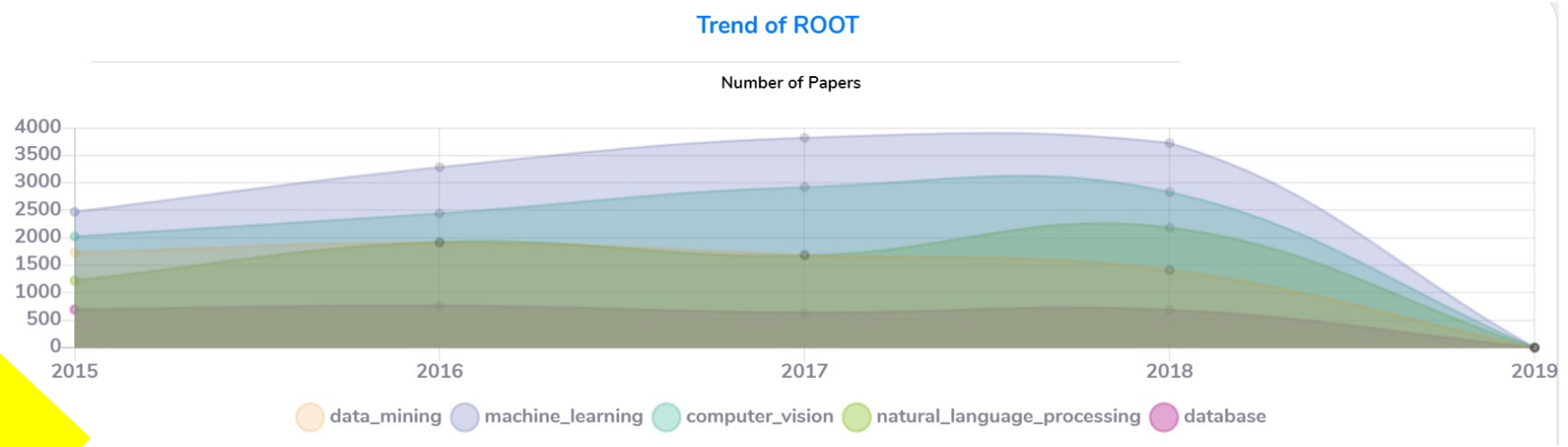
CS Fields

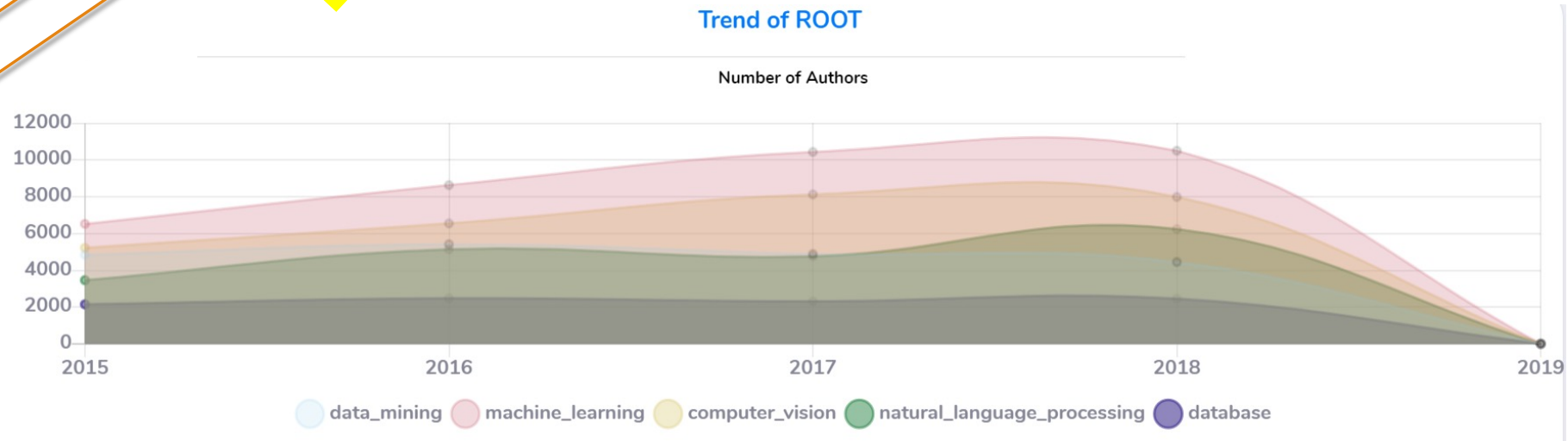


Application: DBLP—Trending Analysis

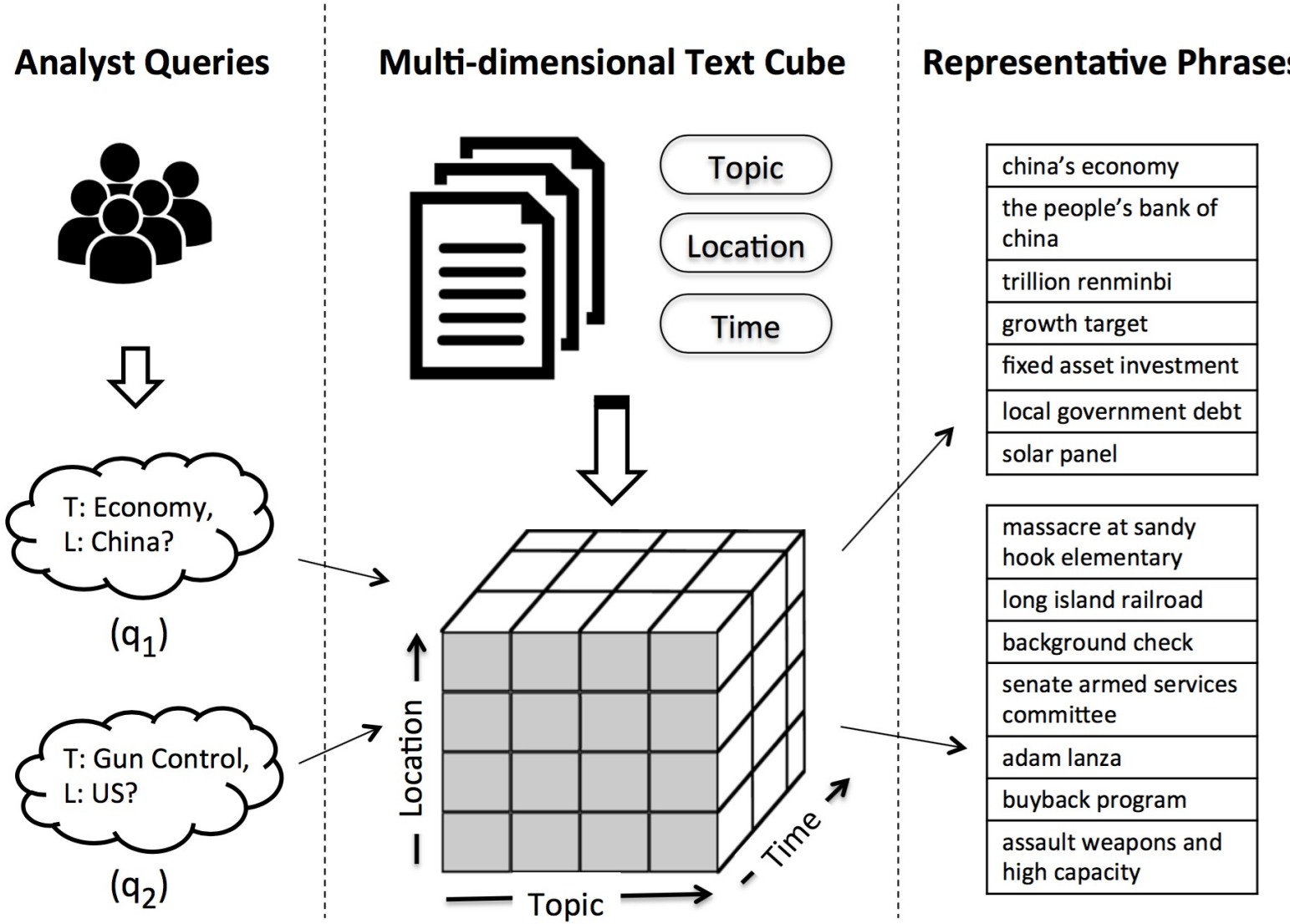
- Trending analysis on CS field development







Application: Comparative Summarization



Tutorial Outline

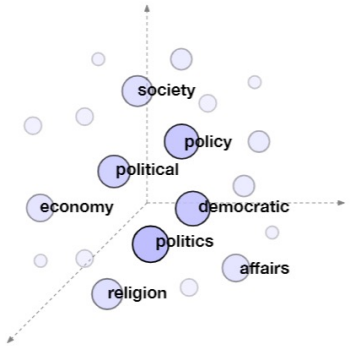
- Introduction
- Part I: Pre-Trained Language Models
- Part II: Revisiting Text Mining Fundamentals with Pre-Trained Language Models
- Part III: Embedding-Driven Topic Discovery
- Part IV: Weakly-Supervised Text Classification: Embeddings with Less Human Effort
- Part V: Advanced Text Mining Applications Empowered by Pre-Trained Embeddings
- Summary and Future Directions

Our Roadmap of This Tutorial

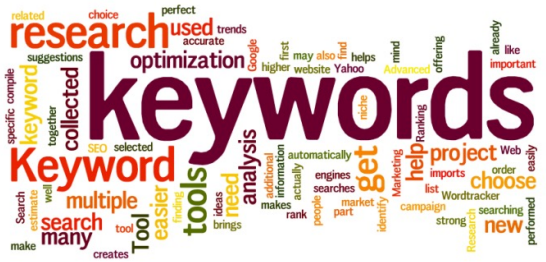
Text Corpus

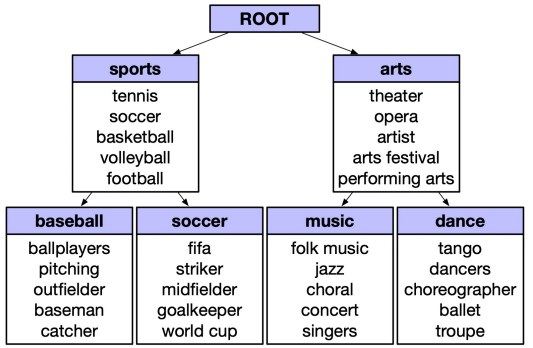
Existing KB

Part I: Pre-Trained Language Model



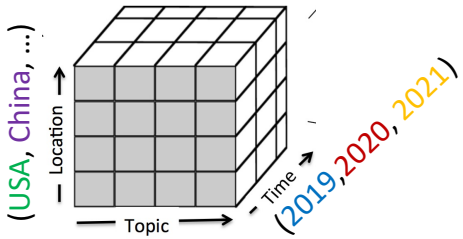
Part II: Text Mining Fundamentals (Phrase/Entity Mining, Taxonomy)



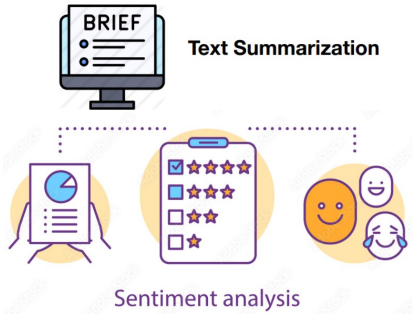


Part III: Topic Discovery

Part IV: Weakly-Supervised Text Classification



(Politics, Science, Economy)



Part V: Advanced Text Mining Applications (Sentiment Analysis, Summarization)