

Adapting Pretrained Representations for Text Mining

Yu Meng, Jiaxin Huang, Yu Zhang, Jiawei Han

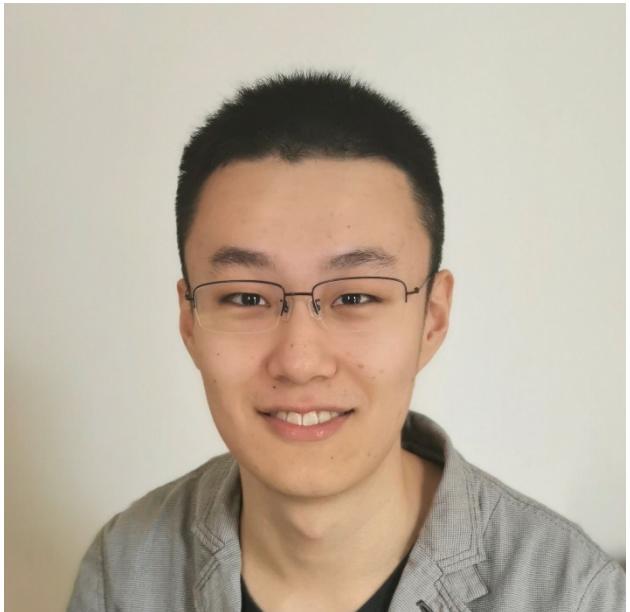
**Department of Computer Science
University of Illinois at Urbana-Champaign**

Aug 14, 2022

Estimated Timeline for This Tutorial

- ❑ Introduction: **15 mins (13:00-13:15 Han)**
- ❑ Part I: Pretrained Language Models: **35 mins (13:15-13:50 Meng)**
- ❑ Part II: Revisiting Text Mining Fundamentals with Pretrained Language Models: **35 mins (13:50-14:25 Huang)**
- ❑ Break: **10 mins (14:25-14:35)**
- ❑ Part III: Embedding-Driven Topic Discovery: **35 mins (14:35-15:10 Yu Zhang, Yu Meng, Jiaxin)**
- ❑ Part IV: Weakly-Supervised Text Classification: Embeddings with Less Human Effort: **35 mins (15:10-15:45 Zhang)**
- ❑ Part V: Advanced Text Mining Applications: **10 mins (15:45-15:55 Han)**
- ❑ Summary and Future Directions: **5 mins (15:55-16:00)**

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



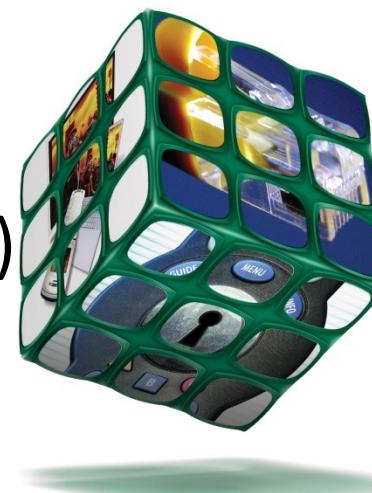
- Yu Zhang**
Ph.D. Candidate, UIUC
- Recipient of 2022 Yunni
and Maxine Pao
Memorial Fellowship



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

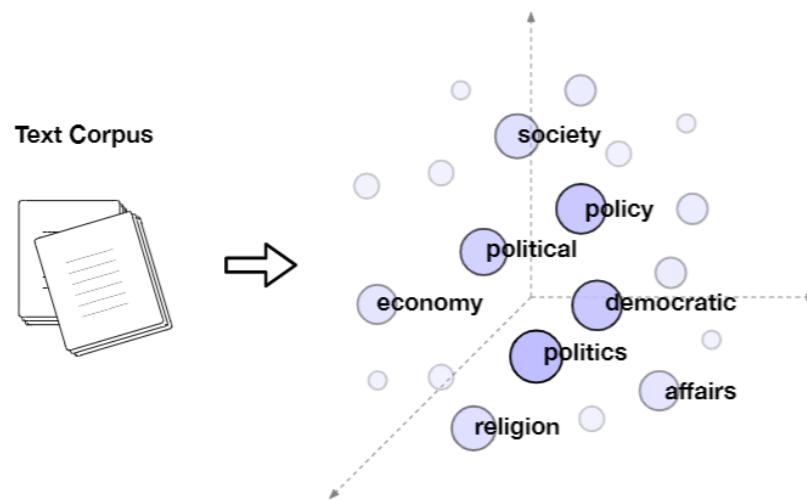
Over 80% of Big Data is Unstructured Text Data

- 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



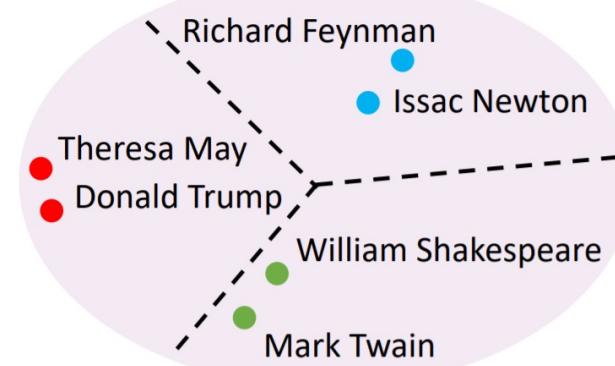
Text Representation: Embeddings & Language Models

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



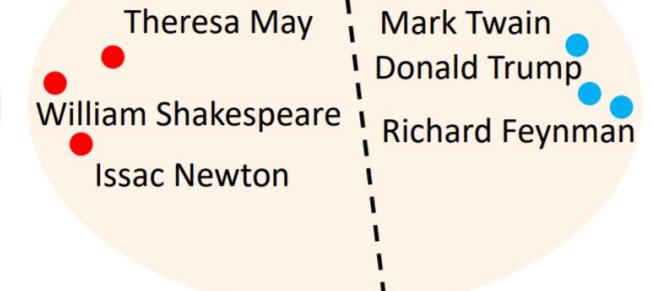
Unsupervised word embeddings:
learned from corpus statistics

Field Discriminative Embedding Space
(Category Name: Politics, Science, Literature)



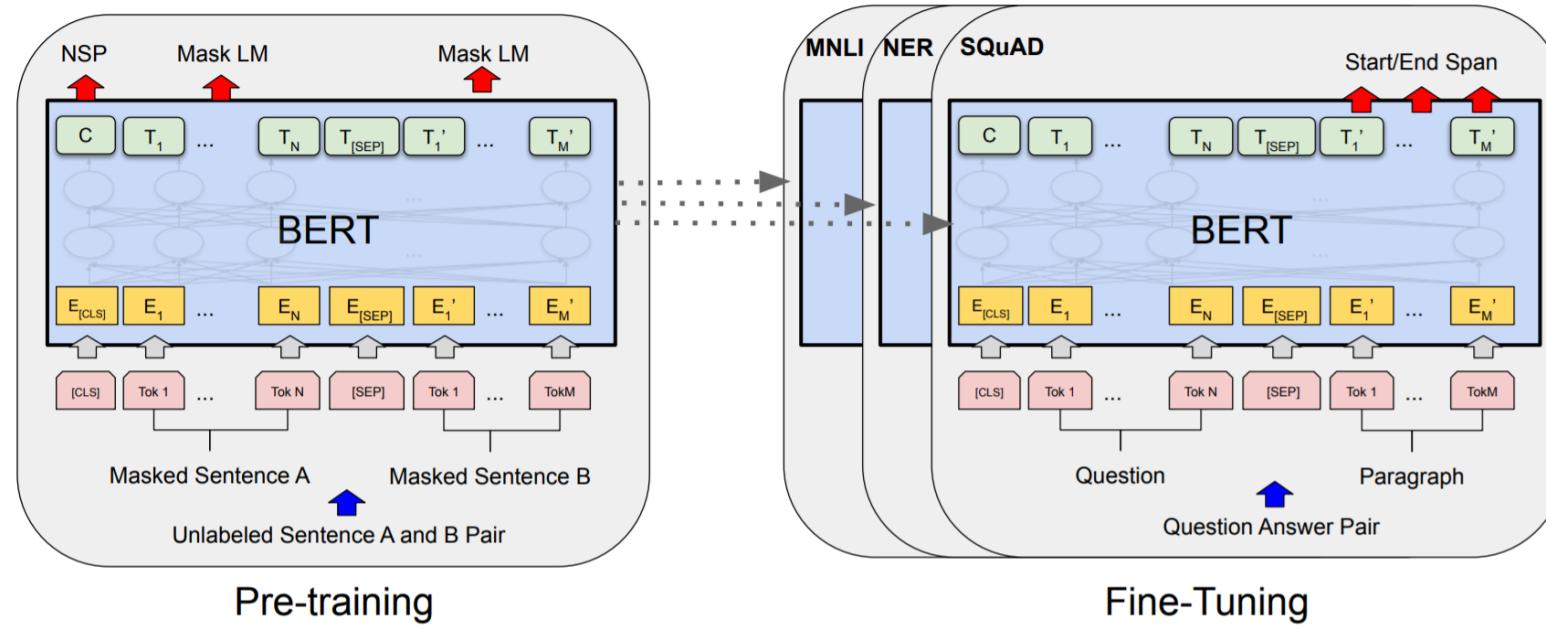
(Weakly-)supervised word embeddings:
learned from corpus statistics & user guidance

Location Discriminative Embedding Space
(Category Name: England, United States)



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”

The screenshot shows the TripAdvisor PriceFinder interface for New York City Hotels. At the top, there's a map of New York City with various landmarks labeled. Below the map are search filters: check-in date (09/08/2015), check-out date (09/08/2015), number of rooms (1 room), and number of guests (2 guests). A red box highlights the "Collections" section on the left sidebar, which contains a list of curated hotel collections: Walk to Penn Station (13), Times Square Views (9), Urban Oasis (12), Trendy Soho (11), Central Park Views (10), Art Deco Classic (12), Catch a Show (22), Design Hotels (12), and a "More" link. The main content area displays two hotel listings: "Hyatt Times Square New York" and "Hilton Times Square". Each listing includes a thumbnail image, the hotel name, a star rating, the number of reviews (2,576), the location (#46 of 469 hotels in New York City), and two recent reviews: "Great Location!" from 08/23/2015 and "Loved our stay here" from 08/23/2015.

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

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

Adv. Structuring: Multidimensional Nature of Texts

- ❑ The same document can naturally describe things across multiple dimensions

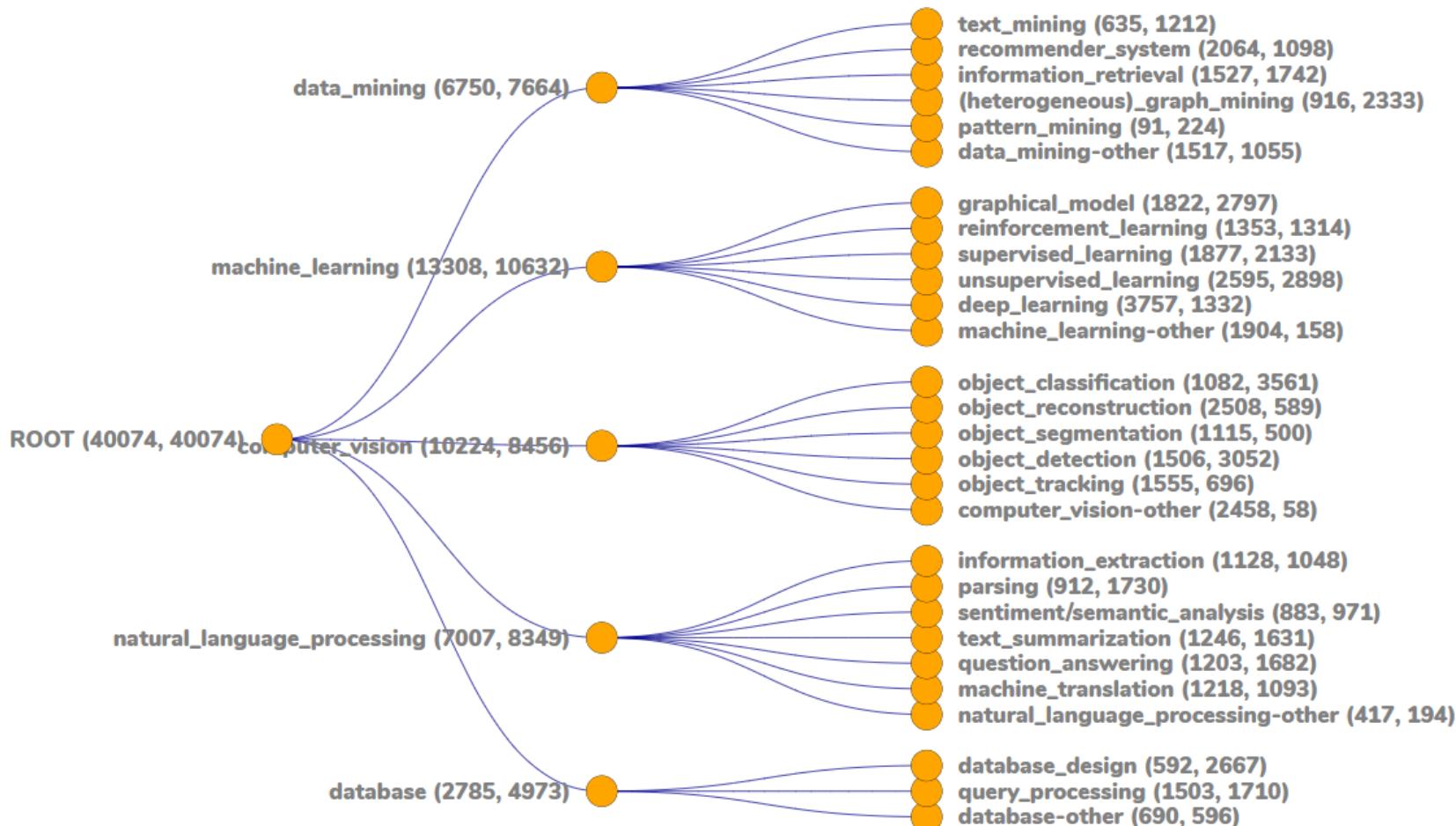
 - ❑ Example:
 - ❑ A technical review may cover
 - ❑ Brands
 - ❑ Products
 - ❑ Aspects
 - ❑ Years
 - ❑ ...
- Apple's 10th anniversary iPhone X sets a new gold standard for the next decade of iPhones. Coming hot on the heels of the [iPhone 8](#) and [iPhone 8 Plus](#), the [iPhone X](#) stole the show 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](#).

... ...

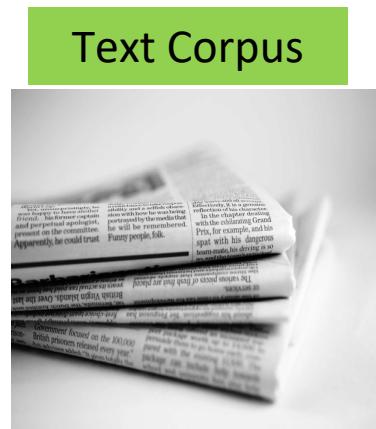
Advanced Structuring: Automatic Taxonomy Generation

Automatically Generated Taxonomy Visualization
Numbers in () from left to right represents the number of main papers and the number of secondary papers respectively.

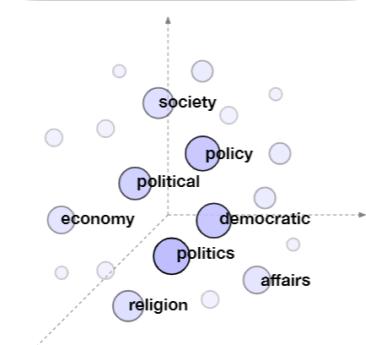
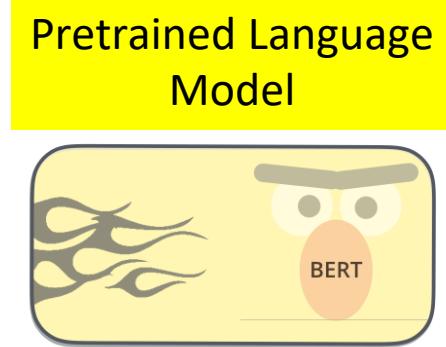


Adv. Structuring: Multi-Dimensional Text Cube Construction

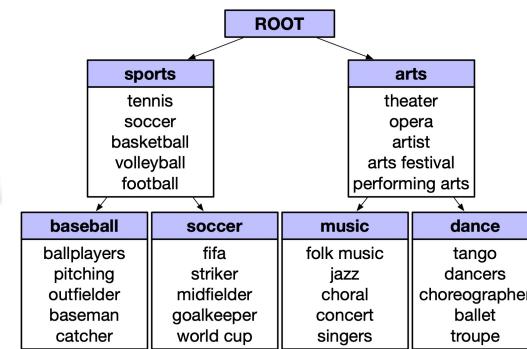
- Understand and Extract Information from Massive Text Corpora
- Organize and Analyze Information using **Multidimensional** Text Analysis



Existing KB

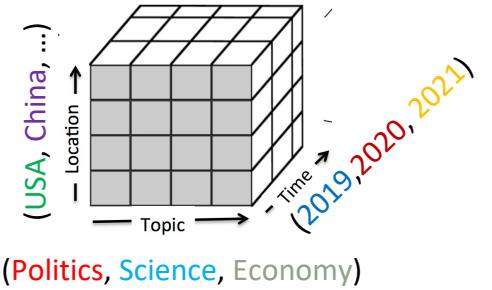


Text Mining Fundamentals
(Phrase/Entity Mining, Taxonomy)



Topic Discovery

Weakly-Supervised Text Classification



(Politics, Science, Economy)

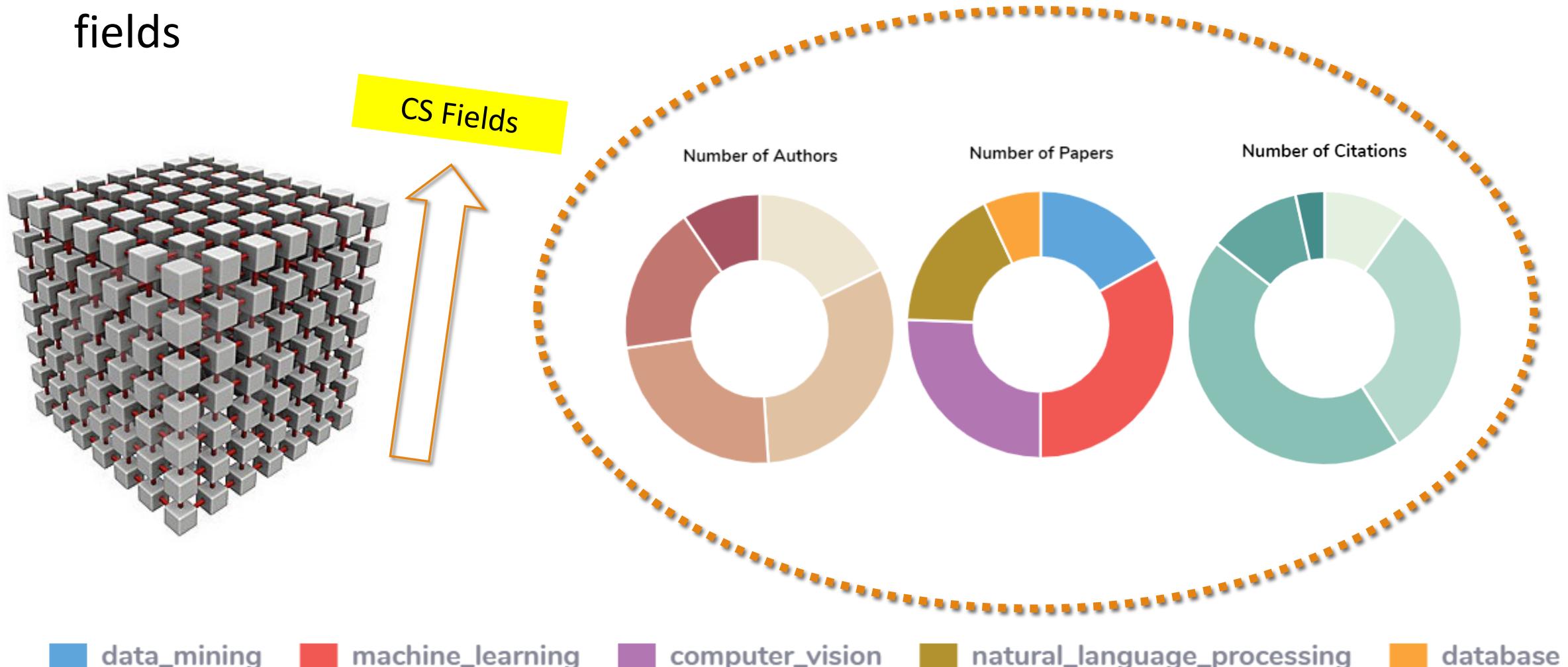


Sentiment analysis

Advanced Text Mining Applications
(Sentiment Analysis, Summarization)

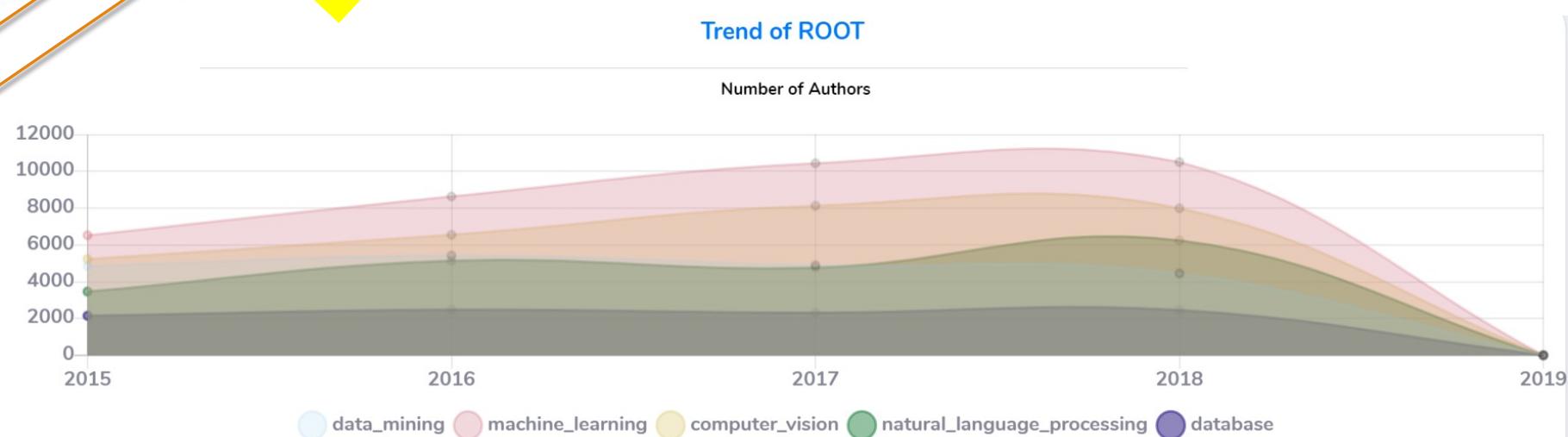
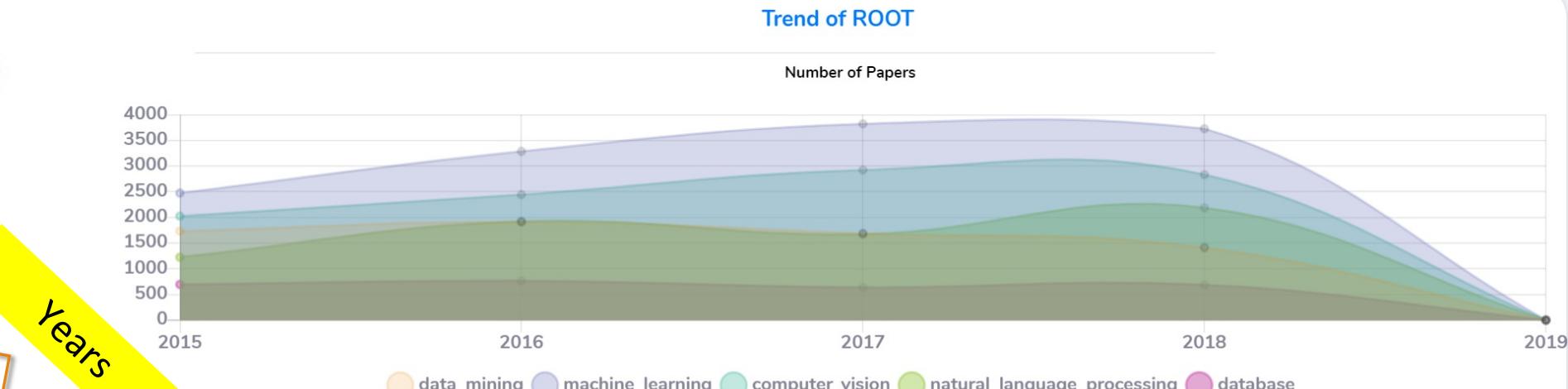
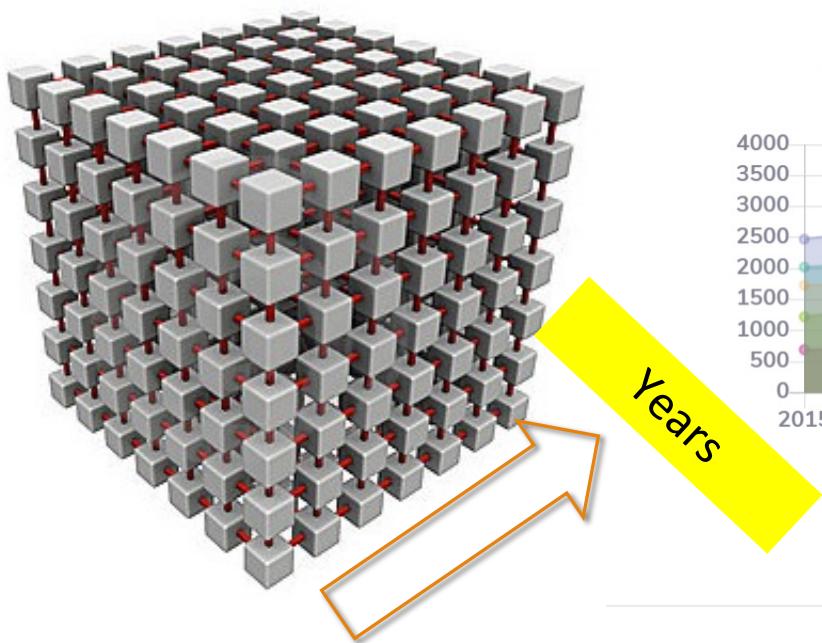
Application: DBLP—Automatic Paper Categorization

- Multidimensional text categorization and exploration across different CS fields

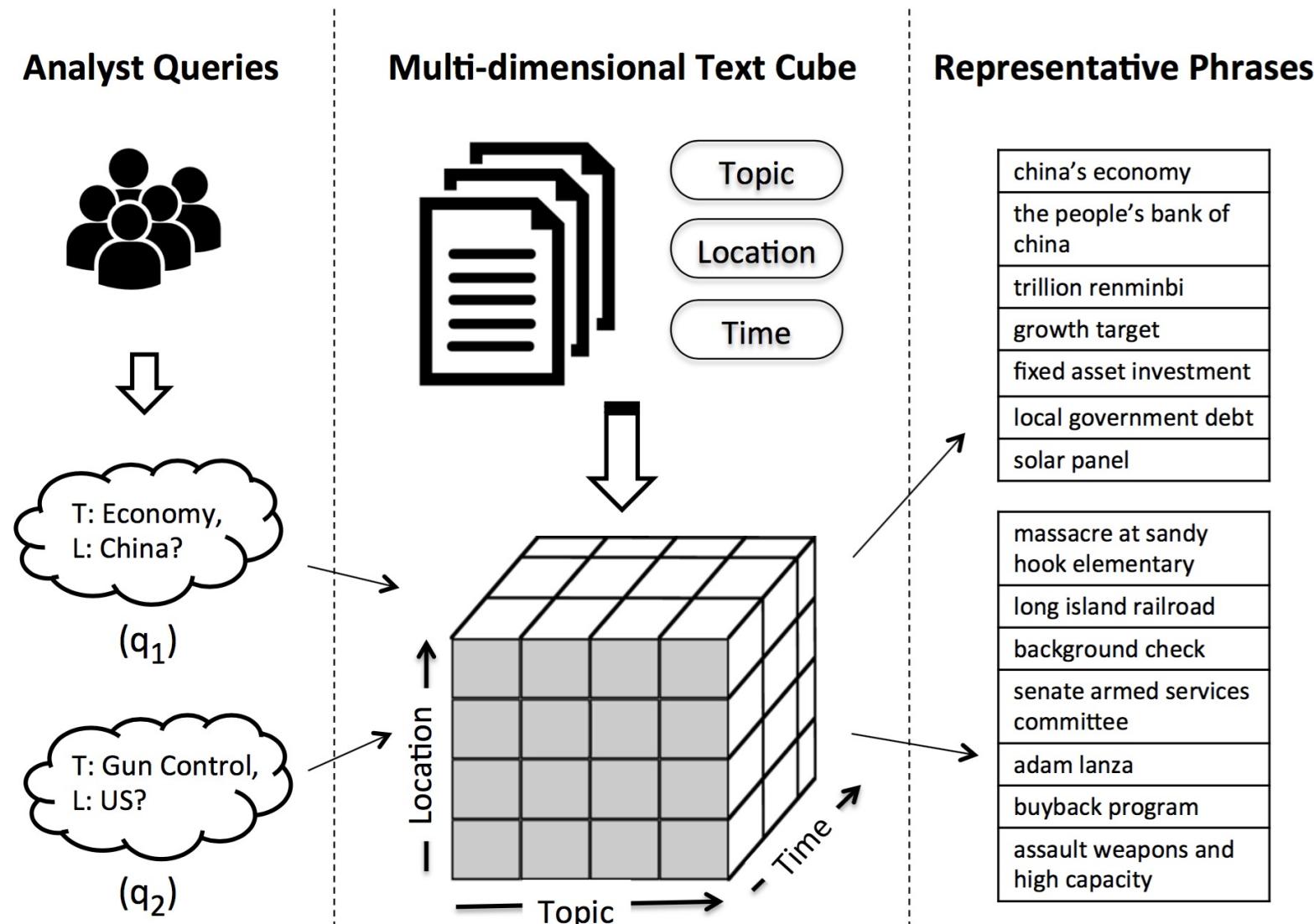


Application: DBLP—Trending Analysis

- Trending analysis on CS field development



Application: Comparative Summarization



Tutorial Outline

- Introduction
- Part I: Pretrained Language Models
- Part II: Revisiting Text Mining Fundamentals with Pretrained 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 Pretrained Embeddings
- Summary and Future Directions

Our Roadmap of This Tutorial

