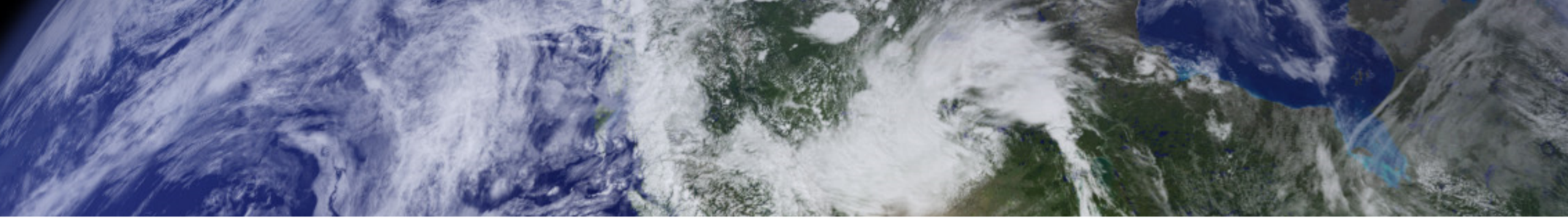


Embedding-Driven Multi-Dimensional Topic Mining and Text Analysis

Yu Meng, Jiaxin Huang, Jiawei Han

Department of Computer Science

University of Illinois at Urbana-Champaign

August 23, 2020

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?
  - **Basic Structuring** (i.e., phase mining & transforming unstructured text into structured, typed entities/relationships (IE))
  - **Embedding** (i.e., computing similarities among entities and relations)
  - **Advanced Structuring**: Discovering Hierarchies/taxonomies, exploring in multi-dimensional space

Multidimensional Nature of Texts

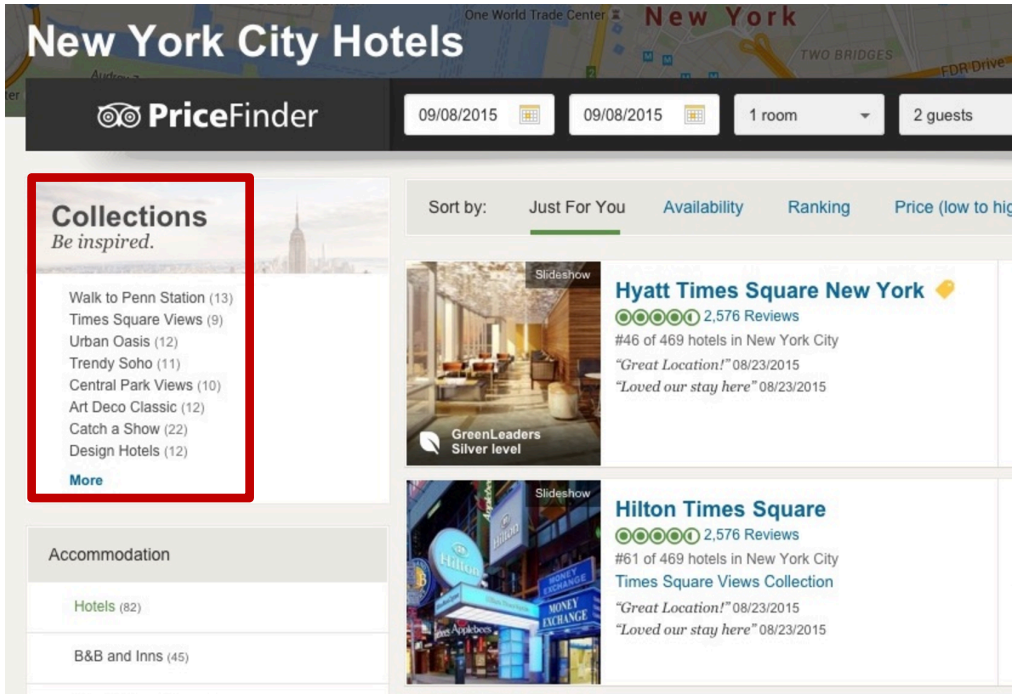
- The same document can naturally describe things across multiple dimensions

- 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.
… …

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

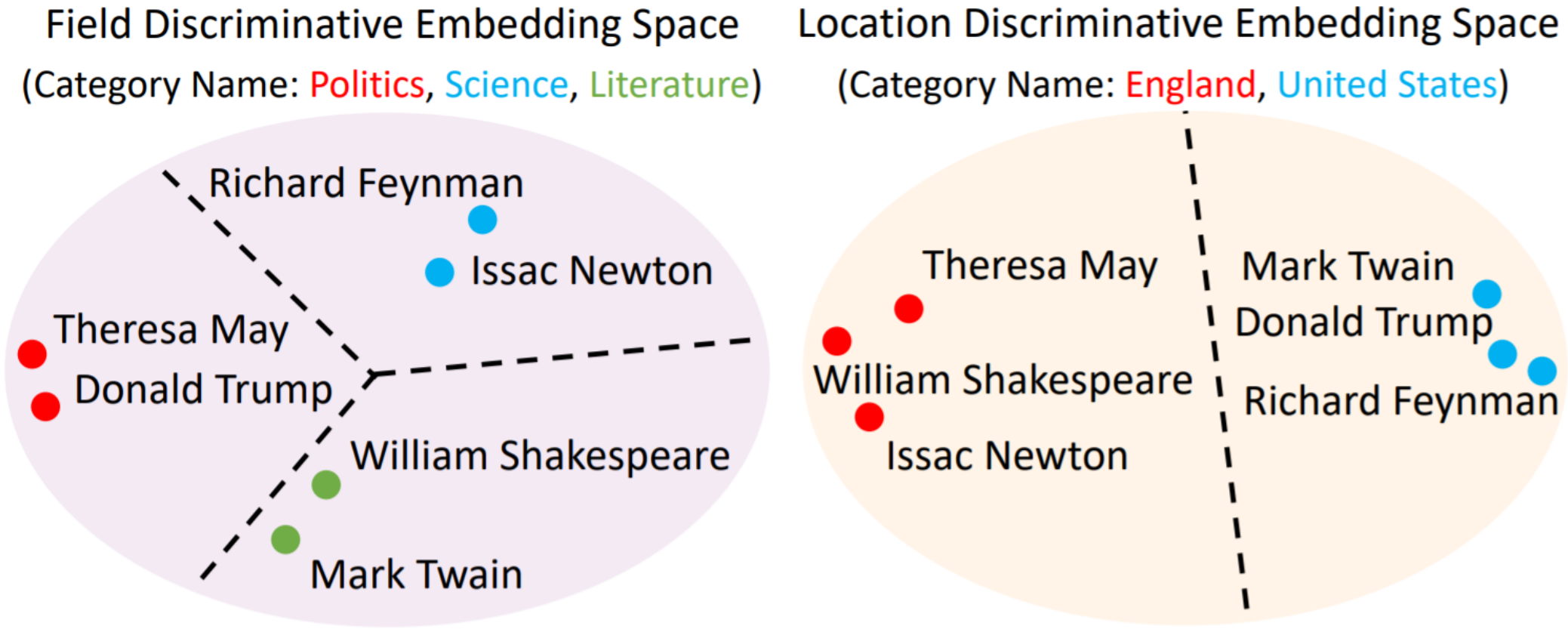
Angiotensin-converting enzyme 2 ***GENE_OR_GENOME*** *(* *ACE2* ***GENE_OR_GENOME*** *) as a SARS-CoV-2* ***CORONAVIRUS*** *receptor* ***CHEMICAL****: molecular mechanisms and potential therapeutic target.*

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 ….

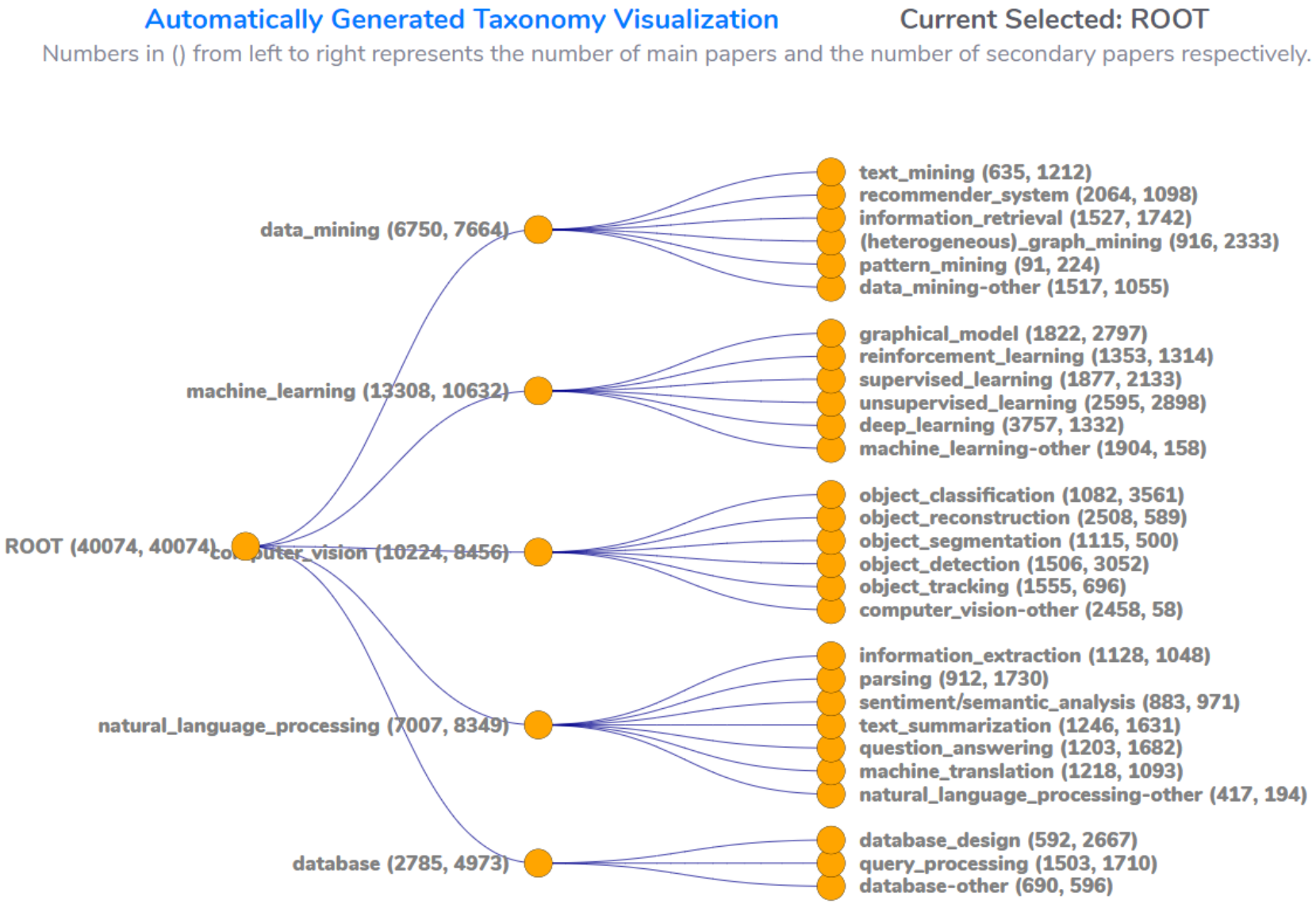
Automatic Recognition of 75 entity types in COVID-19 Biomedical Literature (ISMB 2020 Demo)

Text Embedding: Multi-faceted Topic Mining

- Mining a set of coherent and representative terms based on a set of user-given categories.



Advanced Structuring: Automatic Taxonomy Generation



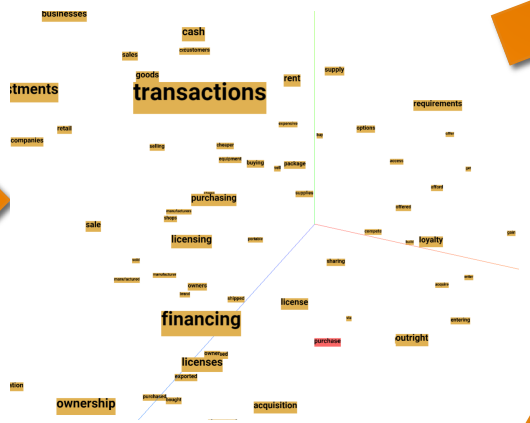
Adv. Structuring: Multi-Dimensional Text Cube Construction

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

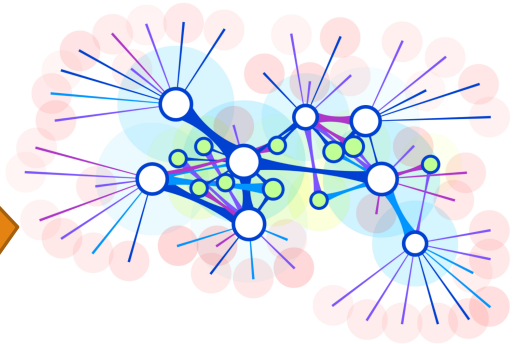
Text Corpus

Existing KB

Text Embedding

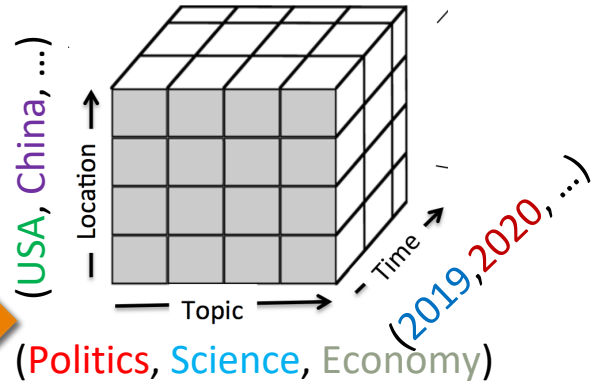


Taxonomy Construction



Topic Mining

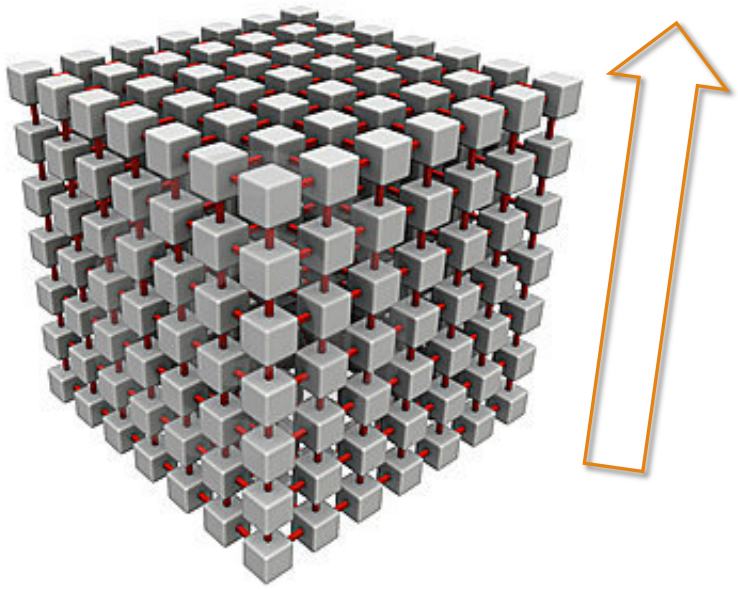
Multi-Dimensional Text Analysis



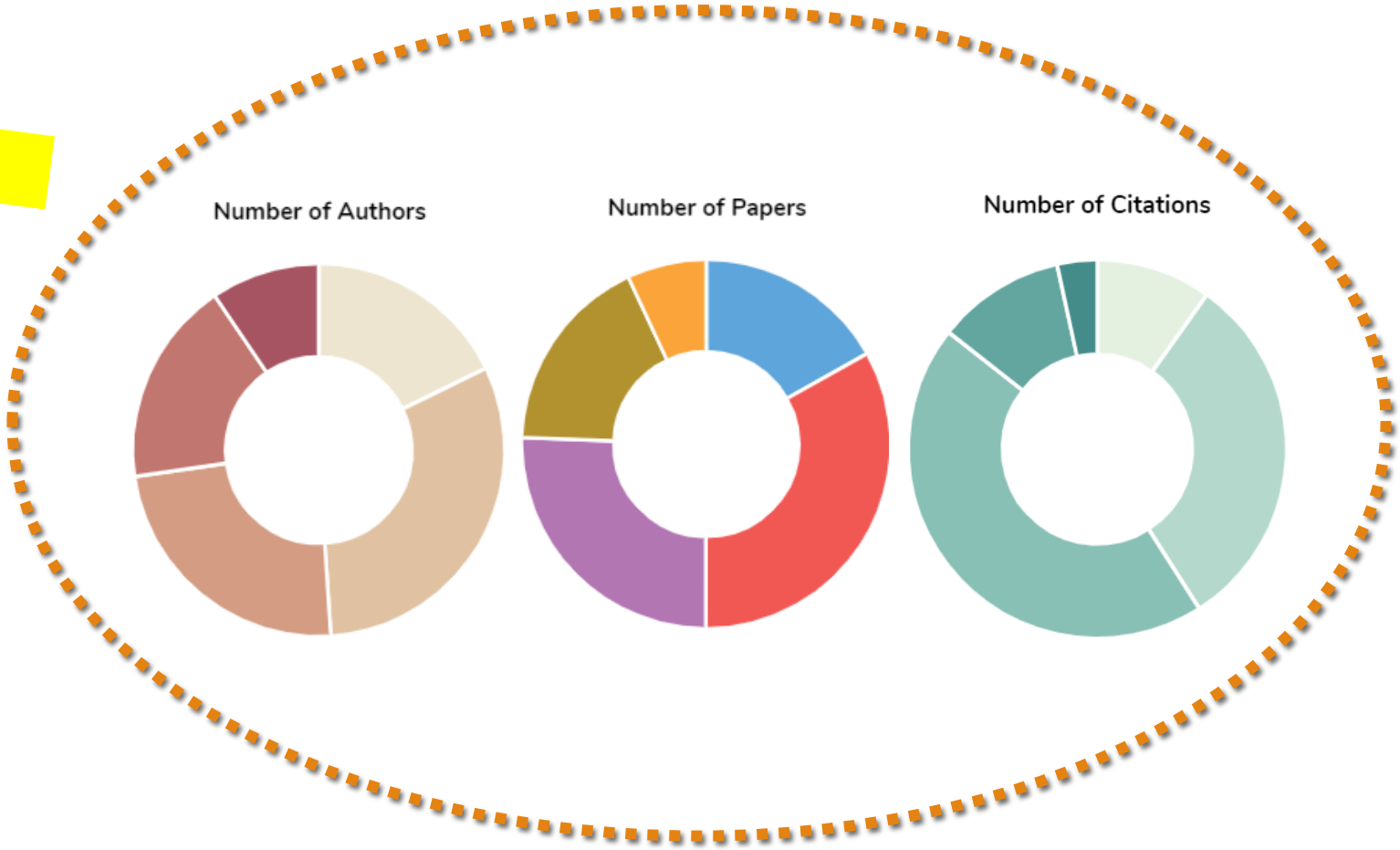
Knowledge

Application: DBLP—Automatic Paper Categorization

- Multidimensional text categorization and exploration across different CS fields

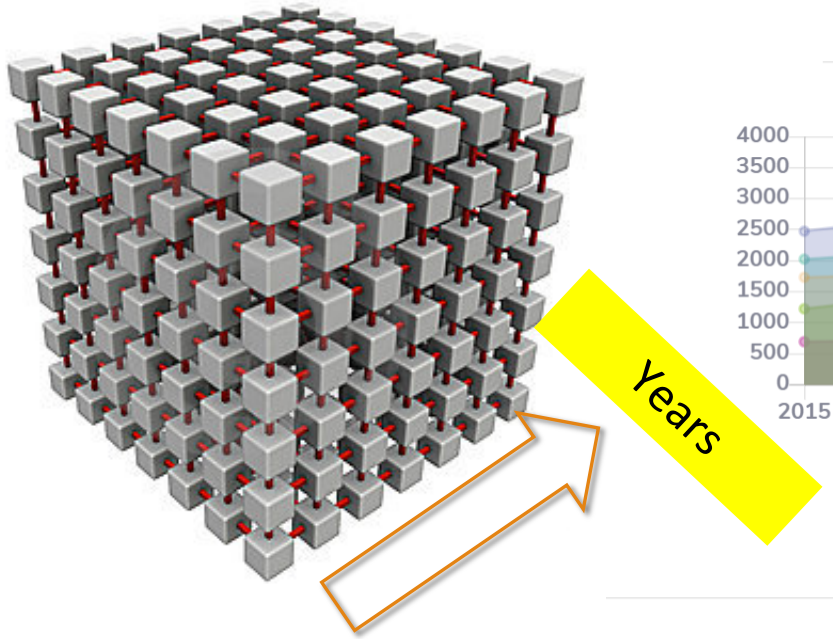


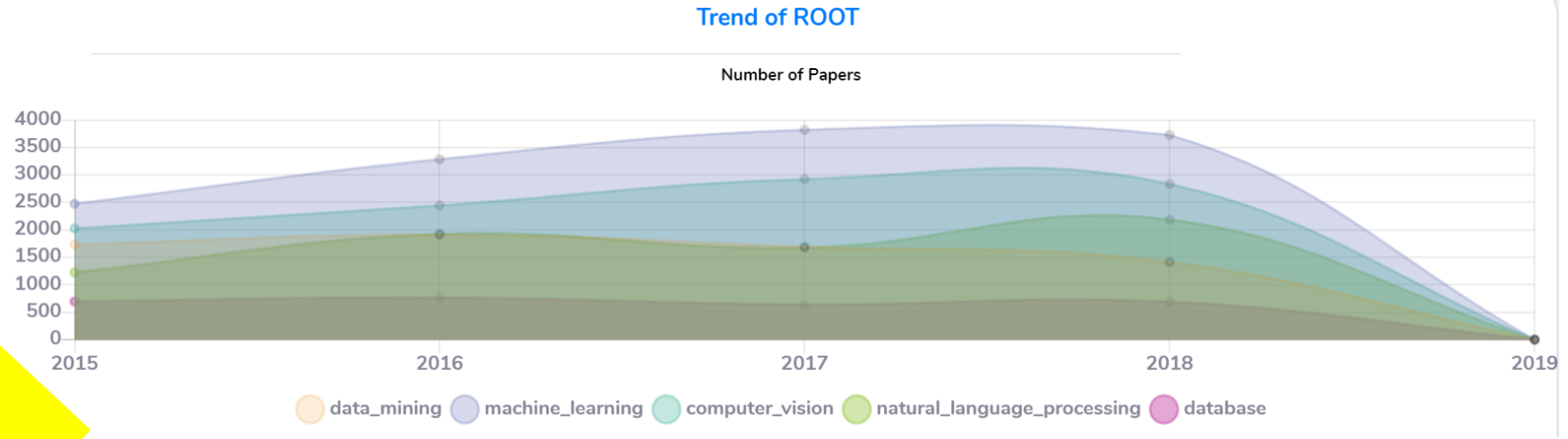
CS Fields

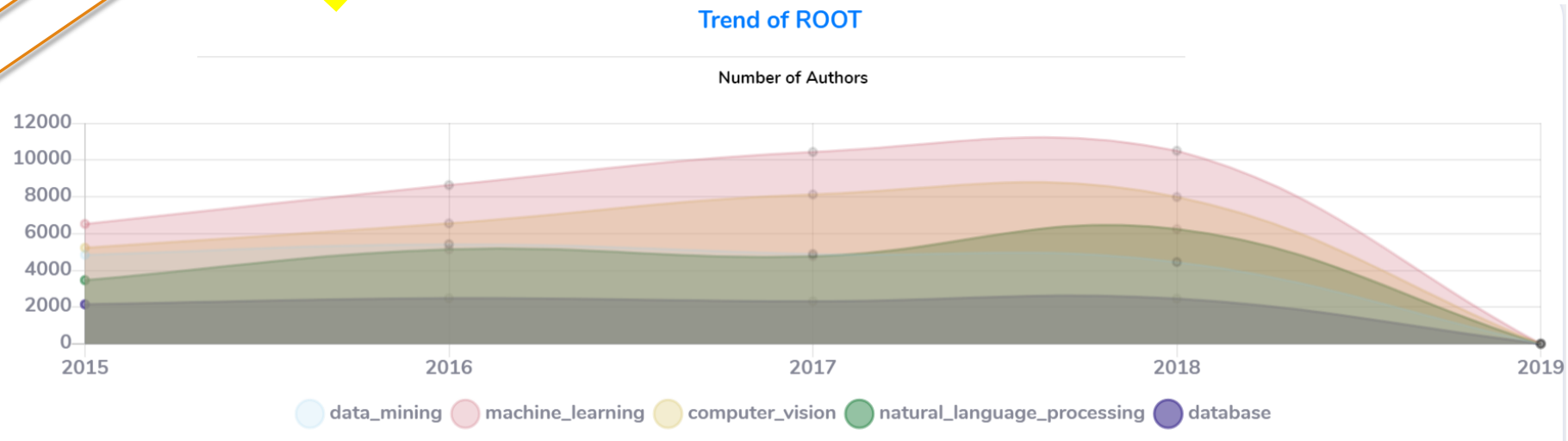


Application: DBLP—Trending Analysis

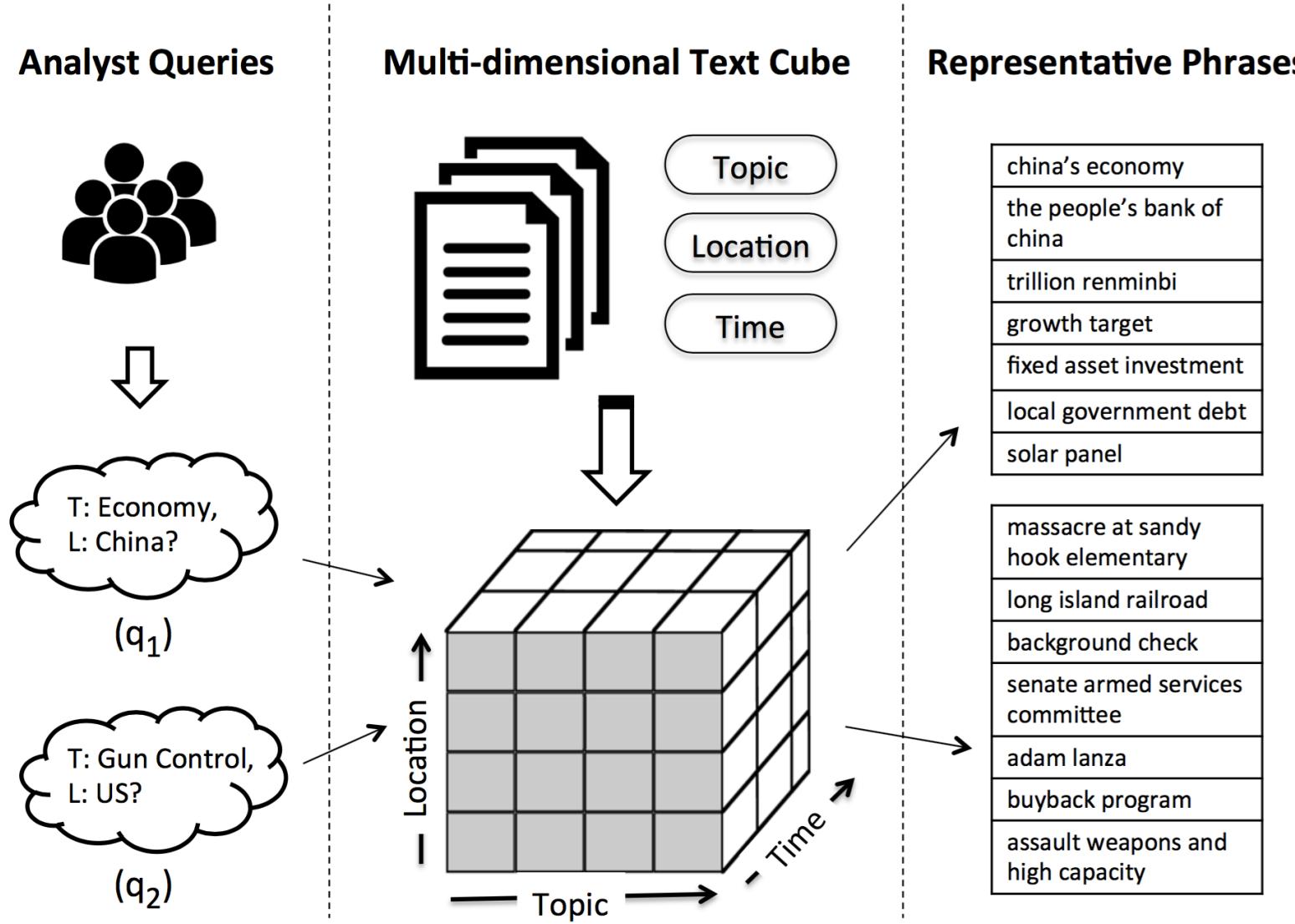
- Trending analysis on CS field development







Application: Comparative Summarization



Tutorial Outline

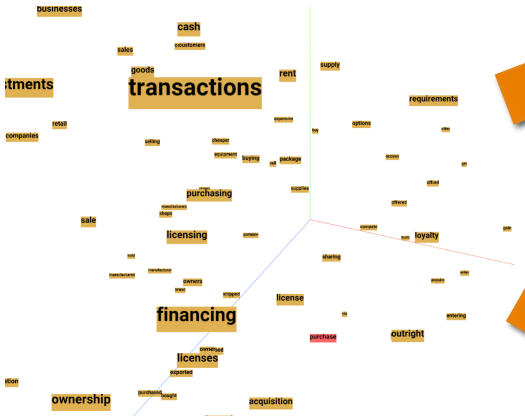
- Introduction
- Part I: Text Embedding
- Part II: Taxonomy Construction
- Part III: User-Guided Topic Mining
- Part IV: Multi-Dimensional Text Analysis
- Summary and Future Directions

Our Roadmap of This Tutorial

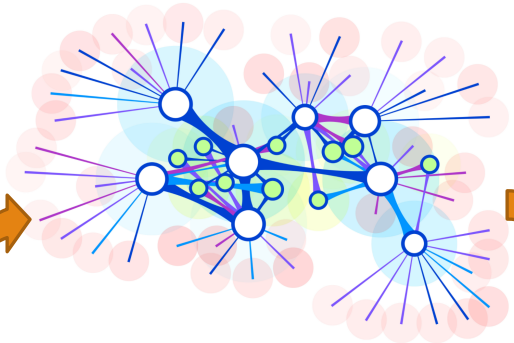
Text Corpus

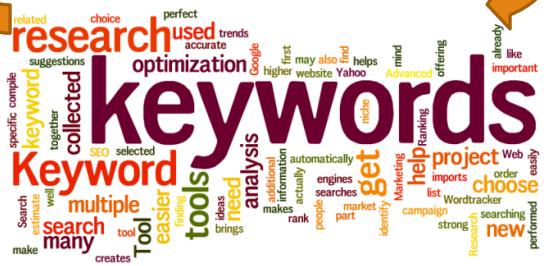
Existing KB

Part I: Text Embedding



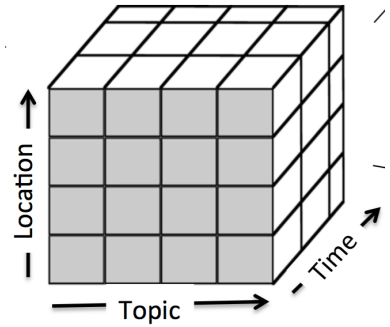
Part II: Taxonomy Construction





Part III: User-Guided Topic Mining

Part IV: Multi-Dimensional Text Analysis



Knowledge