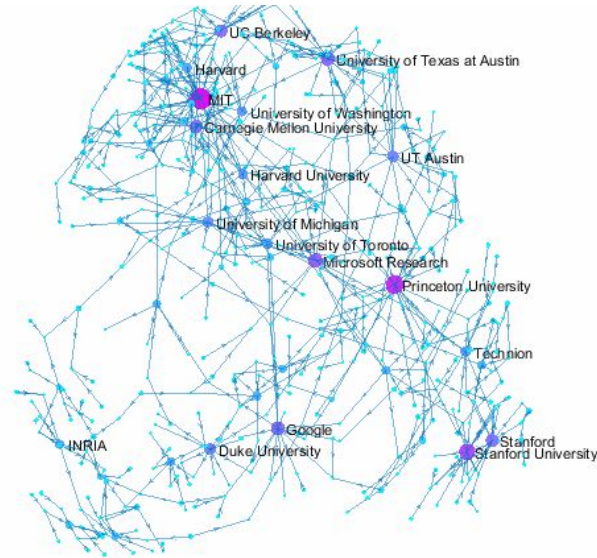


# Semantic Annotations for Frequent Patterns with Context Analysis

Difan Gu  
Wen Long  
Yanyue Wang

# Agenda

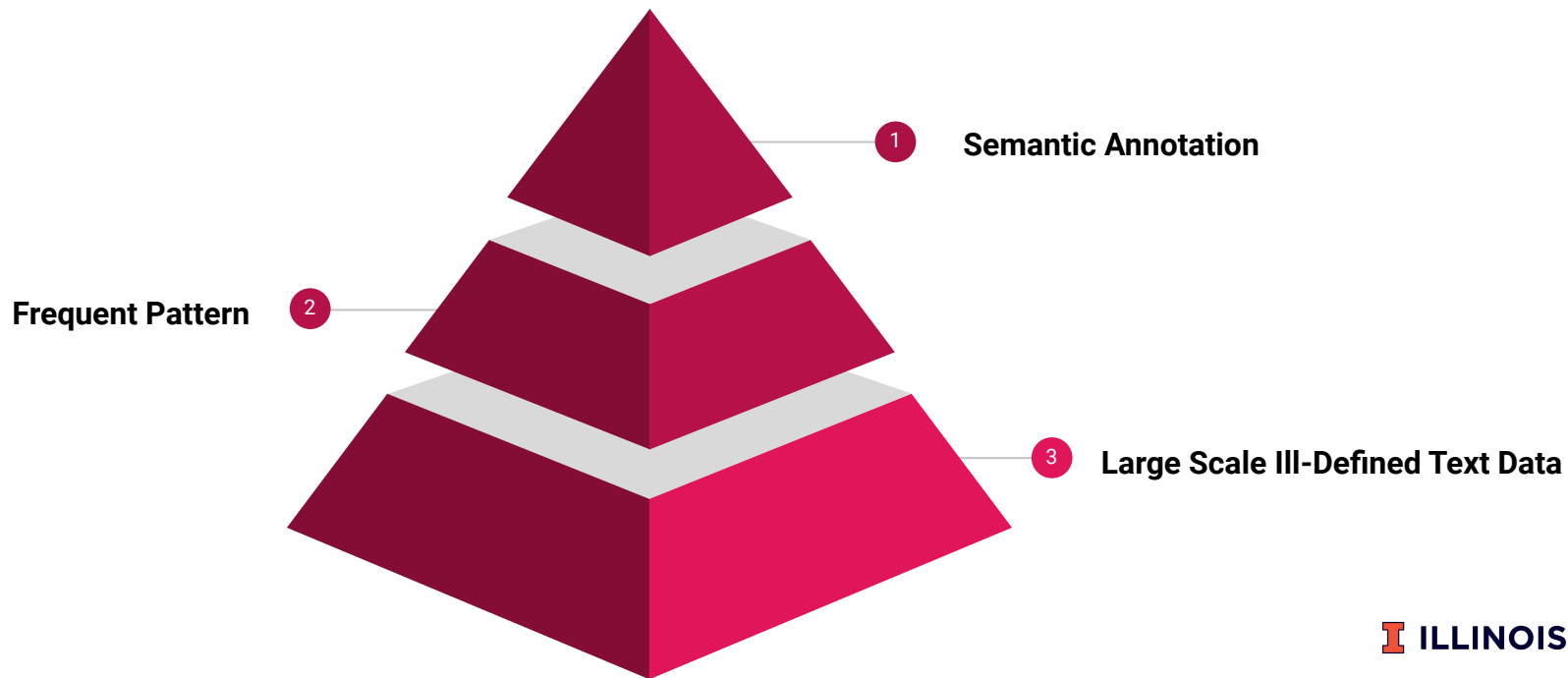
1. Project Objective
2. Methodology
3. Use Case: DBLP
4. Setup
5. Demo
6. Result



Loren on the Art of MATLAB

## Project Objective

**Goal:** to generate semantic annotations by providing contextual information to help readers summarize ill-defined or vague information.



## Methodology

### Data Preprocessing

- Extract Data from XML
- Tokenize Text Data
- Remove Stop Words
- Extract Frequent Terms

### Pattern Mining

- Pattern Preprocessing
- Extract Frequent Pattern
- Remove Pattern Redundancy

### Ranking

- Rank by Mutual Information
- Word Conversion
- Generate Result

## Use Case: DBLP

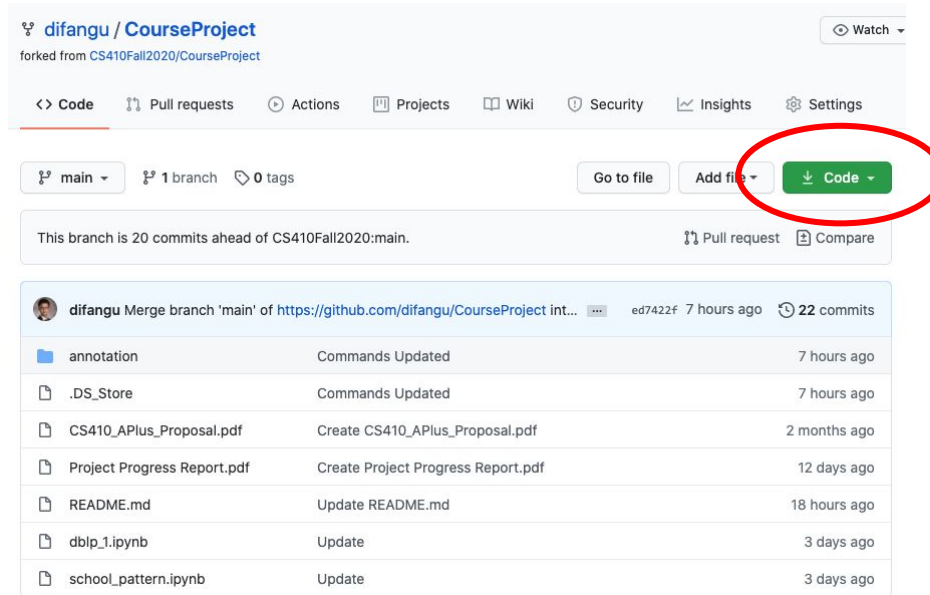
Given a large set of data, what are the most distinguished **academic focuses** of each college in major computer science conferences?

Desired Result:

- MIT: Artificial Intelligence, Robotics
- GT: Algorithm, Database
- UoM: Data Management, System Design

## Setup

1. Go to <https://github.com/difangu/CourseProject>
2. Hit green “Code” button and download to any directory
3. Memorize the directory the “annotation” folder you saved



The screenshot shows the GitHub repository page for **difangu / CourseProject**, which is forked from **CS410Fall2020/CourseProject**. The page includes navigation tabs for Code, Pull requests, Actions, Projects, Wiki, Security, Insights, and Settings. Below these, there are buttons for 'Go to file', 'Add file', and a green 'Code' button with a download icon, which is circled in red. The repository is on the 'main' branch, 1 branch ahead of the upstream, and has 0 tags. A status bar indicates 'This branch is 20 commits ahead of CS410Fall2020:main.' and provides links for 'Pull request' and 'Compare'. The commit history shows a merge of the 'main' branch from the upstream repository, with a list of files and their update times:

File	Commit Message	Time
annotation	Commands Updated	7 hours ago
.DS_Store	Commands Updated	7 hours ago
CS410_APlus_Proposal.pdf	Create CS410_APlus_Proposal.pdf	2 months ago
Project Progress Report.pdf	Create Project Progress Report.pdf	12 days ago
README.md	Update README.md	18 hours ago
dblp_1.ipynb	Update	3 days ago
school_pattern.ipynb	Update	3 days ago

**Demo**

## Demo

1. Open Terminal and cd to the location where holds the folder “annotation”
2. Run the following code in the oder:
  - a. `python school_pattern_creator.py`
  - b. `java -jar spmf.jar run CloSpan ./school_output/school9.txt ./pattern_output/output9.txt 1%`
  - c. `java -jar spmf.jar run CloSpan ./school_output/school2.txt ./pattern_output/output2.txt 0.5%`
  - d. `java -jar spmf.jar run CloSpan ./school_output/school8.txt ./pattern_output/output8.txt 3%`
  - e. `python pattern_decipher.py`
3. Check out the result either in terminal printout or in the “outcome” file under “annotation”



## Result

Top 5 semantic annotation for each school:



MIT	GT	UoM
System Control	Programming	Resource, Wireless
Power, Energy	Framework	Embedded, Systems
Control, Analysis	Management	Network, Resource
Power, Applications	Problems	Recognition, Object
Systems, Large	Approach	Social

# Thank You

## Reference

KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining August 2006 Pages 337–346 <https://doi.org/10.1145/1150402.1150441>

CloSpan: Mining Closed Sequential Patterns in Large Datasets, by X. Yan, J. Han, and R. Afshar. Proc. of 2003 SIAM Int. Conf. Data Mining (SDM'03), 2003