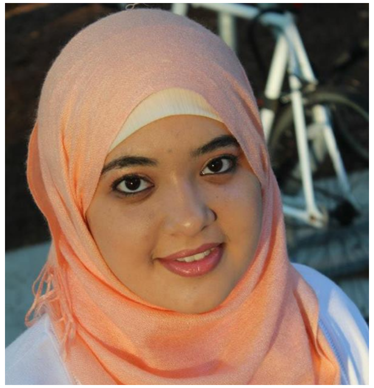
CS510 (Fall 2018)
Advanced Topics in Information Retrieval

Instructor: **ChengXiang ("Cheng") Zhai**

Department of Computer Science

University of Illinois, Urbana-Champaign



Teaching Assistants

Assma Boughoula

Xueqing Liu

DAIS The Data and Information Systems Laboratories at The University of Illinois at Urbana-Champaign *Large Scale Information Management and Mining*

TIMAN

Text data cover all kinds of topics

Topics:
People
Events
Products
Services, …

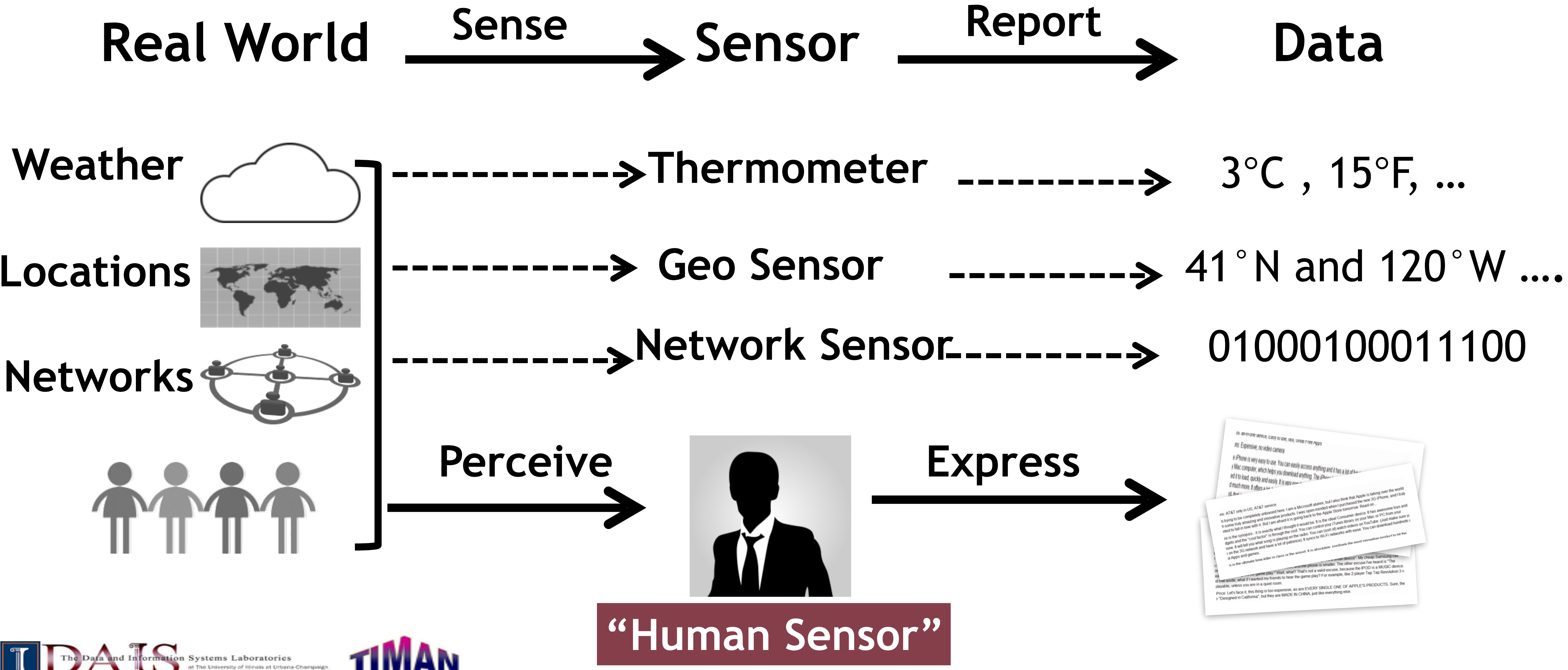
Sources:
Blogs
Microblogs
Forums
Reviews ,…

45M reviews ↑ 53M blogs ↑ 65M msgs/day ↑ 115M users ↑

1307M posts 10M groups

tripadvisor blogScope PREVIEW twitter YAHOO! GROUPS …

Humans as Subjective & Intelligent “Sensors”



Real World —Sense→ Sensor —Report→ Data

Weather ------------> Thermometer ----------> 3°C , 15°F, …

Locations ------------> Geo Sensor ----------> 41 ° N and 120 ° W ….

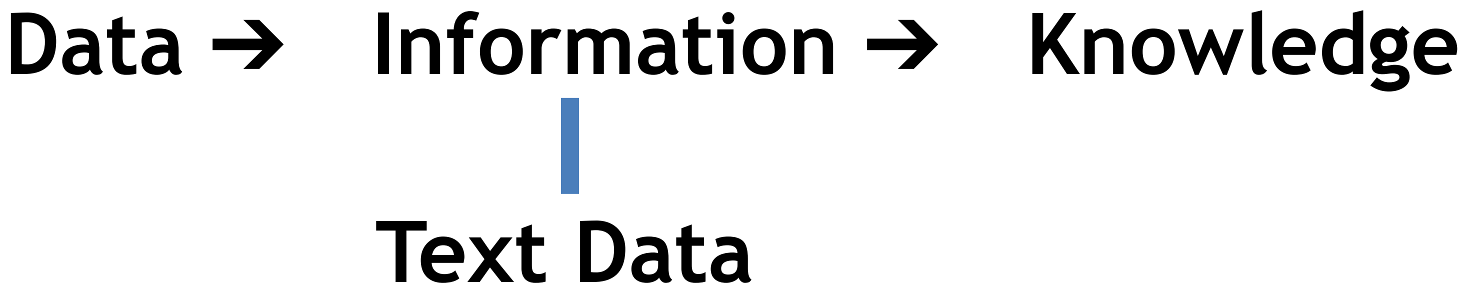
Networks ------------> Network Sensor----------> 01000100011100

Perceive → Express →

“Human Sensor”

Unique Value of Text Data

- Useful to **all** big data applications
- Especially useful for mining knowledge about **people's behavior, attitude**, and **opinions**
- **Directly** express knowledge about our world: **Small text data** are also useful!



Data → Information → Knowledge

Text Data

However, NLP is difficult!

“A man saw a boy *with a telescope*.” (who had the telescope?)

“He has quit smoking” ➔ he smoked before.

How can we leverage **imperfect** NLP to build a **perfect** general application?

Answer: Having humans in the loop!

TextScope to enhance human perception

Microscope

Telescope

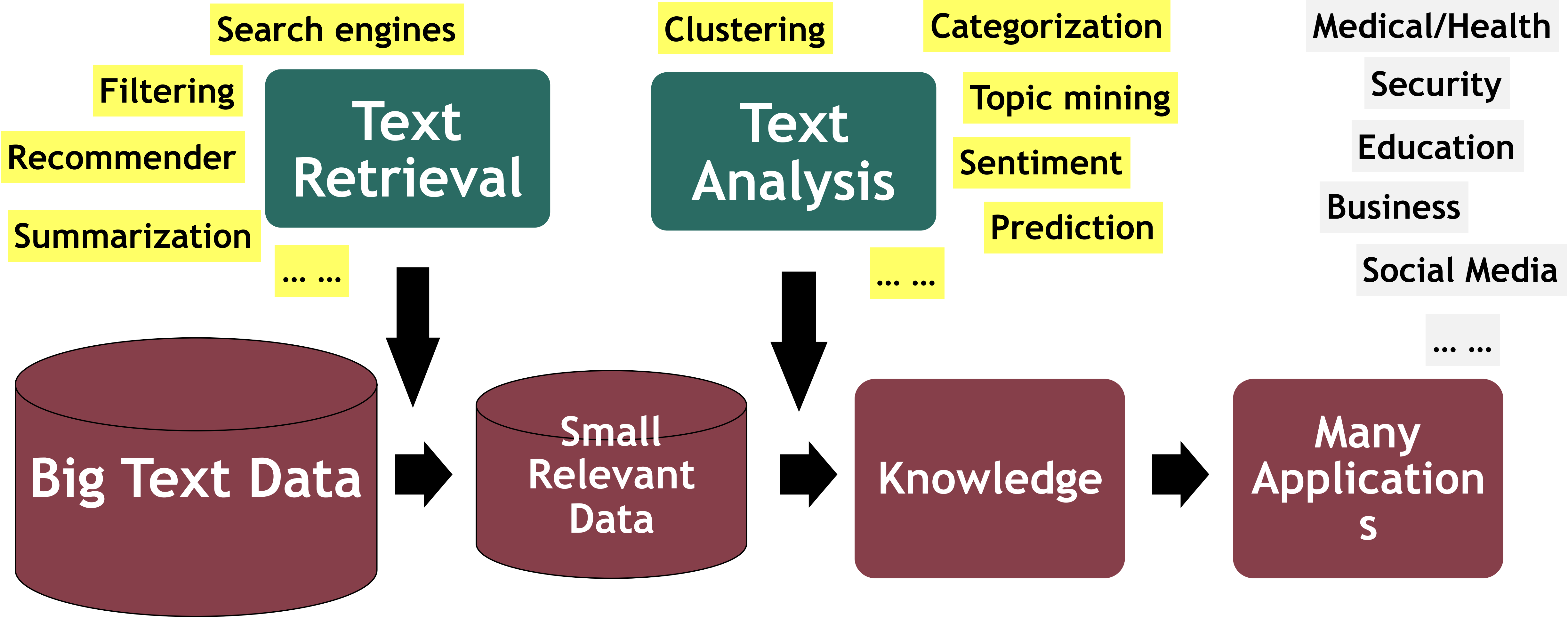
TextScope

Intelligent Interactive Retrieval & Text Analysis
for Task Support and Decision Making

Examples of TextScope Applications

- **Search**
  - Web search, enterprise search, desktop search, PubMed, …
- **Filtering/Recommender Systems**
  - spam email filter, news/literature/movie recommender
- **Categorization**
  - news categorization, help desk email routing, sentiment tagging, …
- **Topic mining**
  - discovery of topical trends in scientific research
  - discovery of major complaints from customers
  - business intelligence, bioinformatics, …
- **Text-based Prediction**
  - prediction of stock prices, voting results, …

Main Techniques for Building a TextScope: Text Retrieval + Text Analysis



Search engines

Filtering

Recommender

Summarization

… …

Text Retrieval

Clustering

Categorization

Topic mining

Sentiment

Prediction

… …

Text Analysis

Medical/Health

Security

Education

Business

Social Media

… …

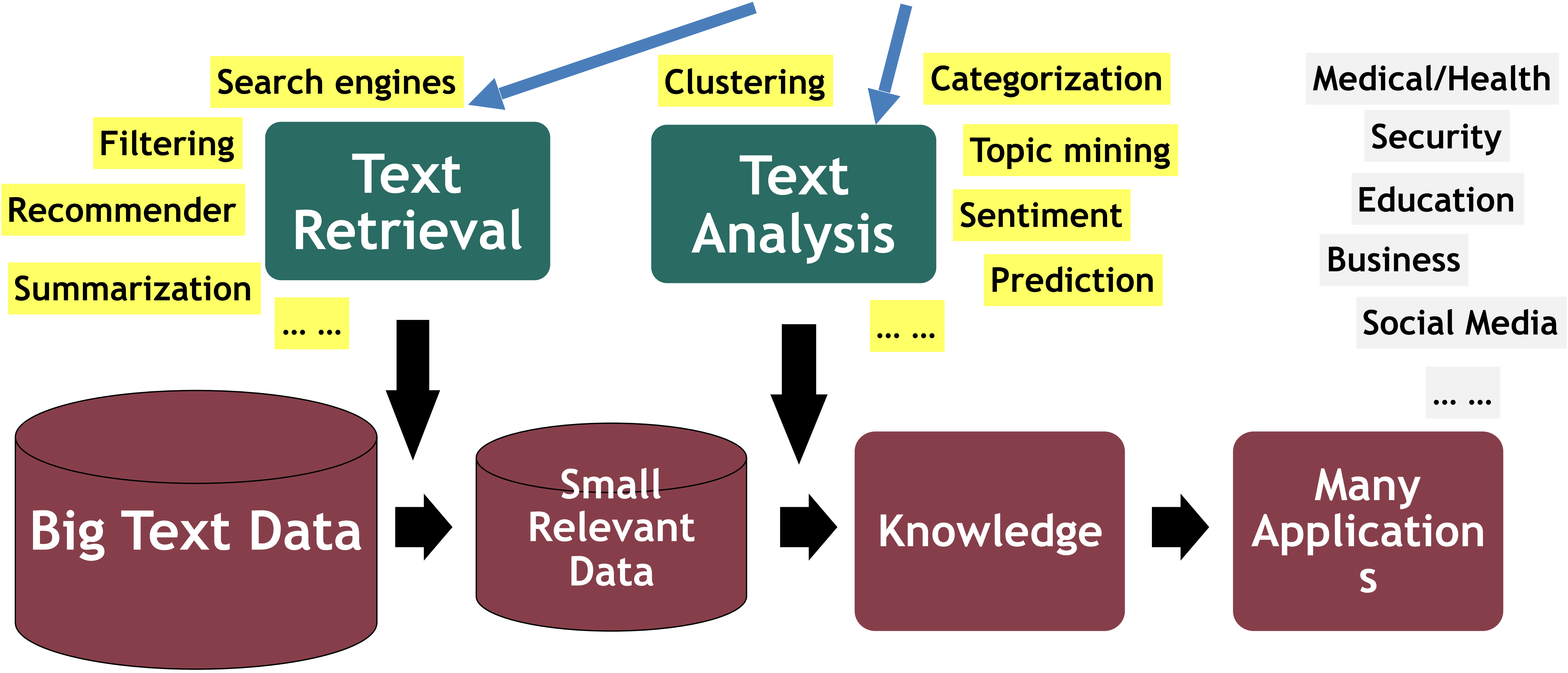
Big Text Data

Small Relevant Data

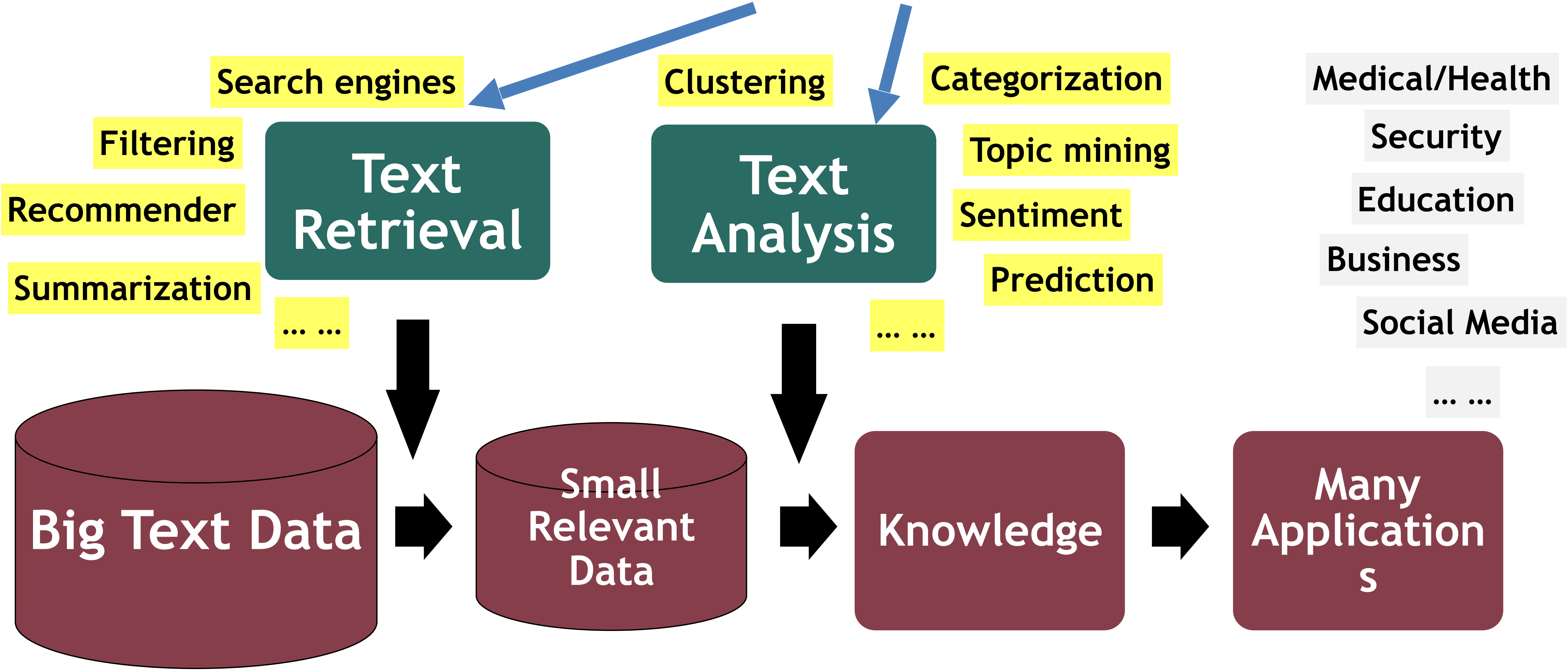
Knowledge

Many Application s

This Course: Statistical Language Models



Assignment: MeTA Toolkit



Search engines

Filtering

Recommender

Summarization

… …

Text Retrieval

Clustering

Categorization

Topic mining

Sentiment

Prediction

… …

Text Analysis

Medical/Health

Security

Education

Business

Social Media

… …

Big Text Data

Small Relevant Data

Knowledge

Many Application s

Course Goal

- Advanced (graduate-level) introduction to the field of information retrieval (IR), broadly including Text mining
- Goal
  - Provide a systematic introduction to statistical language models and their applications in text retrieval and text analysis
  - Provide an opportunity for students to explore frontier topics via course projects (customized toward the interests of students)
  - Give students enough training for doing research in IR or applying advanced IR techniques to applications
  - Tangible outcome: research paper, open source code, and application system

Prerequisites

- Basic concepts in CS410 Text Info Systems
- Programming skills: CS225 or equivalent level
- A good knowledge of basic probability and statistics
- Knowledge of one or more of the following areas is a plus, but not required: Information Retrieval, Machine Learning, Data Mining, Natural Language Processing
- Contact the instructor if you aren’t sure

Format

- Lectures (mostly by instructor)
- Short frequent written assignments (problem sets): ensure solid mastery of concepts, models, and algorithms
- Programming assignments: ensure solid mastery of skills of implementation and experimentation
- 2 Midterms (75 min each, in class): mostly to verify your mastery of concepts, models, and algorithms as covered in the assignments
- Course project: multiple options
  - In-depth study of a topic ➔ publication/submission
  - Implementation of a major algorithm ➔ open source
  - Development of a novel application ➔ useful application

Grading

- Assignments: 30%
- Midterm 1: 20%
- Midterm 2: 20%
- Project: 30%

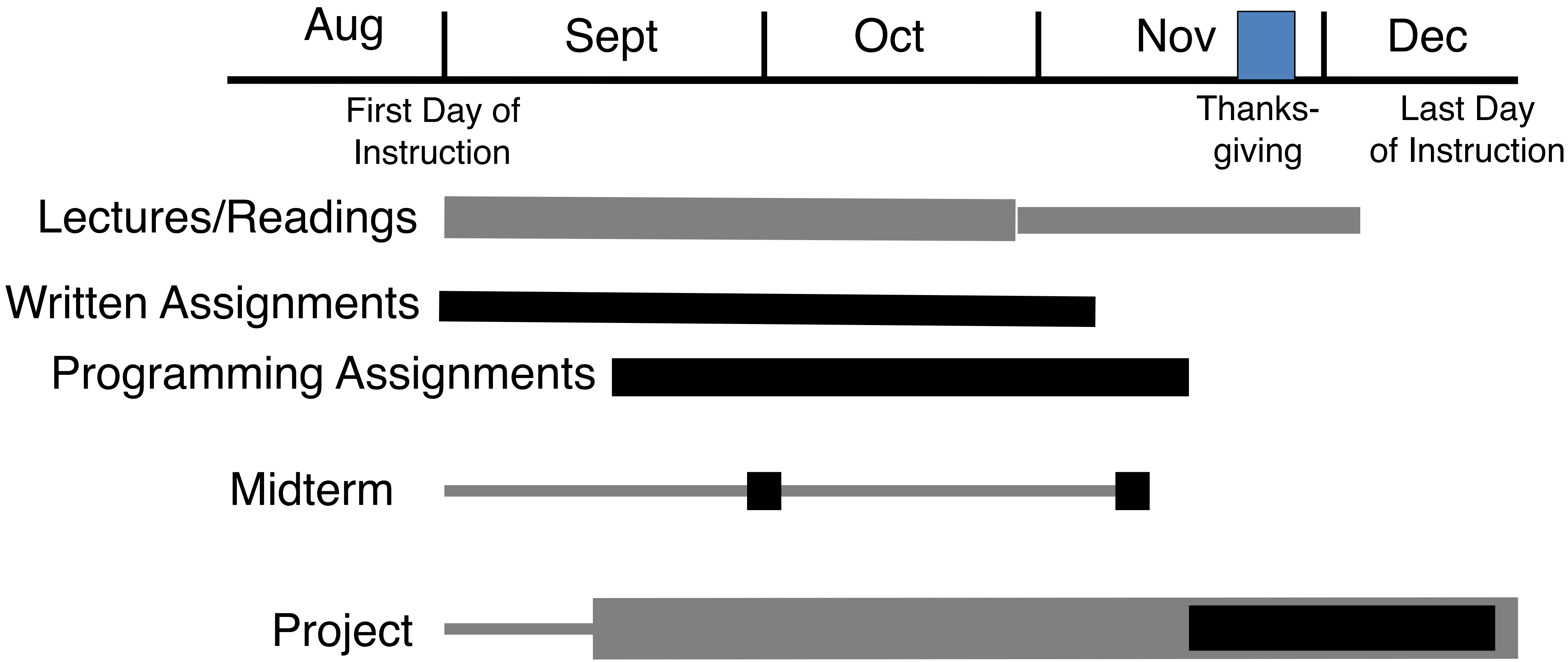
(Tentative) Office Hours

- Instructor:
  - 1:30pm-2:30pm Tuesdays & 3pm-4pm Thursdays
  - 2116SC
- TA (0207SC)
  - Assma Boughoula: 11am-12pm Mondays & 12pm-1pm Wednesdays
  - Xueqing Liu: 11am-12noon, Wednesdays & Fridays
- Post your question on Piazza as soon as you have it.

Schedule

- Background, overview of text retrieval & analysis; relevant math
- Overview of statistical language models (LMs)
- N-gram LMs (applications: text retrieval, text categorization)
- N-gram class LMs (applications: lexical relation discovery, text retrieval)
- Mixture LMs (PLSA, LDA, topic discovery and analysis)
- State-space LMs/Hidden Markov Models (applications: passage retrieval, sequential topic modeling)
- ==
- Contextualized LMs (applications: text mining, text-based prediction)
- Learning to rank
- Neural language models (word embedding, deep learning for IR)

Your Work Load



Reference Book

ChengXiang Zhai, Chase Geigle, *Statistical Language Models for Text Data Retrieval and Analysis*, forthcoming.

Draft will be available online

Other readings: mostly research papers, survey articles, and book chapters

- Synthesis Lectures Digital Library: http://www.morganclaypool.com/
- Foundations & Trends in IR: http://www.nowpublishers.com/ir/
- Recent papers from SIGIR, CIKM, WWW, WSDM, KDD, ACL, ICML,...

Questions?

Course website:
http://times.cs.uiuc.edu/course/510f18

Piazza:
https://piazza.com/illinois/fall2018/cs510