Data and Information Systems (DAIS)
Course Structures at CS/UIUC

- Three main streams: Database, data mining and text information systems
  - Seminar: Yahoo!-DAIS Seminar: (Not CS591 seminar, no credit given)
- Database Systems:
  - Database management systems (CS411: Fall + Spring)
  - Advanced database systems (CS511: Fall)
  - Human-in-the-loop Data Management (CS 598: Aditya Parameswaran)
- Data mining
  - Intro. to data mining (CS412: Fall + Spring)
  - Data mining: Principles and algorithms (CS512: Spring (Han))
- Text information systems
  - Introduction to Text Information Systems (CS410: Spring (Zhai))
  - Advance Topics on Information Retrieval (CS 598 or CS510: Fall (Zhai))
  - Social & Economic Networks (CS 598: Hari Sundaram)
CS512 Coverage@2019: Mining Massive Text Corpora and Information Networks

- Class introduction + course technical overview (1 week)
- Text mining 1: Phrase mining (1 week)
- Text mining 2: Named entity/relation extraction and typing (1 week)
- Text mining 3: Text embedding (1 week)
- Text mining 4: Mining patterns, relations and claims (1 week)
- Text mining 5: Mining sets, synonyms and taxonomy (1 week)
- 1st midterm exam (0.5 week) — 1st Lect. of 7th week
- Text mining 6: Text classification and Text cubes (1 week)
- Text mining 7: Document retrieval and SetSearch (1 week)
- Network mining 1: Heterogeneous information networks and network clustering (1 week)
- Network mining 2: Classification and link prediction in hetero. info. networks (1 week)
- Network mining 3: Other issues at mining heterogeneous information networks (1 week)
- Truth finding (1 week)
- 2nd midterm exams (0.5 week) — 2nd Lect. of 15th week
- Class research project presentation (final week + exam week)
Class Information

- **Instructor:** Jiawei Han ([www.cs.uiuc.edu/~hanj](http://www.cs.uiuc.edu/~hanj))
- **Lectures:** Tues/Thurs 9:30-10:45am (0216 SC)
- **Office hours:** Tues/Thurs 10:45-11:30am (2132 SC)
- **Teach Assistants** (using Piazza to seek for help when needed)
  - Carl Yang (50% lead TA), Qi Zhu (50%, online TA), Xiaotao Gu (50%)
  - TA office hours: 9:30-10:15am Monday and 4-4:45pm Thursday
- **Prerequisites** (course preparation)
  - **CS412** (offered every semester) or consent of instructor
  - General background: Knowledge on **statistics, machine learning (CS446), and text information systems (CS410)** will help understand the course materials
- **Course website** (bookmark it since it will be used frequently!)
  - [https://wiki.cites.illinois.edu/wiki/display/cs512/Lectures](https://wiki.cites.illinois.edu/wiki/display/cs512/Lectures)
- **Major textbook:** Recent research papers plus
  - Charu C. Aggarwal, Machine Learning for Text, Springer 2017
Textbooks & Recommended References

- **Textbooks**
  - Xiang Ren and Jiawei Han, *Mining Structures of Factual Knowledge from Text: An Effort-Light Approach*, Morgan & Claypool Publishers, 2018
  - Jialu Liu, Jingbo Shang and Jiawei Han, *Phrase Mining from Massive Text and Its Applications*, Morgan & Claypool, 2017
  - Yizhou Sun and Jiawei Han, *Mining Heterogeneous Information Networks: Principles and Methodologies*, Morgan & Claypool, 2012
  - Recent published research papers (see course syllabus)

- **Other reference books**
  - Jiawei Han, Micheline Kamber, *Jian Pei, Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2011
Course Work: Assignments, Exams and Course Project

- **Assignments**: (Two assignments, equal weight) **22%** total
  - One programming assignment (10%)
  - One mini-research assignment (12%)

- **Two midterm exams** (equal weight): **40%** in total

- **Research project proposal (3-5 pages)**: 5% (due at the end of 5th week)

- **Class attendance** (3%): Max misses w/o penalty: 3, then −0.3% for each miss
  - For online students, 3% will be folded into research/survey report

- **Final course project**: **30%** (due at the end of semester)
  - Evaluated by class (50%) and TA + instructor (50%) collectively!

- **Class presentation on new papers and surveys** (Option: Extract small credit)
  - Topics and time slot (10-15 minutes): Consent with instructor
Research Projects Evaluation

- **Final course project:** 30% (due at the end of semester)
  - The final project will be evaluated based on (1) technical innovation, (2) thoroughness of the work, and (3) clarity of presentation
  - The final project will need to hand in: (1) project report (length will be similar to a typical 8-12 page double-column conference paper), and (2) project presentation slides (which is required for both online and on-campus students)
  - Each course project for every on-campus student will be evaluated collectively by instructor (plus TA) and other on-campus students in the same class
  - The course project for online students will be evaluated by instructors and TA only
  - Group projects (both survey and research): Single-person project is OK, also encouraged to have two as a group, and team up with other senior graduate students, and will be judged by them
Where to Find Reference Papers?

- Course research papers: Check reading list and references at the each chapter
- Major conference proceedings
  - DM conferences: ACM SIGKDD (KDD), ICDM (IEEE, Int. Conf. Data Mining), SDM (SIAM Data Mining), ECMLPKDD (Principles KDD), PAKDD (Pacific-Asia)
  - NLP conferences: ACL, EMNLP, NAACL
  - IR and Web conferences: SIGIR, CIKM, WWW, WSDM
  - ML conferences: NIPS, ICML
  - DB conferences: ACM SIGMOD, VLDB, ICDE
  - Social network conferences: ASONAM
- Other related conferences and journals
  - IEEE TKDE, ACM TKDD, DMKD, ML, ...
- Use course Web page, DBLP, Google Scholar, Citeseer
Questions for Short Discussion

- Two disciplines: Data mining vs. machine learning
  - What are the links and differences?
- Two courses: CS412 (Introduction to Data Mining) vs. CS512 (Advance Data Mining)
  - What are the links and differences?
- Two research projects: Mini-research assignment vs. your selected research projects
  - What are the links and differences?
- Discussion on course grading policy
From Data to Networks to Knowledge: An Evolutionary Path!

Han, Kamber and Pei, Data Mining, 3rd ed. 2011
Yu, Han and Faloutsos (eds.), Link Mining, 2010
Wang and Han, Mining Latent Entity Structures, 2015
Sun and Han, Mining Heterogeneous Information Networks, 2012

Yizhou Sun: SIGKDD’13 Dissertation Award
Chi Wang: SIGKDD’15 Dissertation Award
Xiang Ren: SIGKDD’18 Dissertation Award