An Overview of CS512 @Spring 2018

JIAWEI HAN
COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

JANUARY 16, 2018
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)
Coursera Data Mining Specialization

- Coursera Data Mining Specialization
  - **Data Visualization**: John Hart
  - **Pattern Discovery in Data Mining**: Jiawei Han
  - **Text Retrieval and Search Engines**: ChengXiang Zhai
  - **Cluster Analysis in Data Mining**: Jiawei Han
  - **Text Mining and Analytics**: ChengXiang Zhai
  - **Data Integration and Data Warehousing**: Kevin Chang
  - **Capstone Data Mining Capstone (6 weeks)**
  - Online MCS Data Science Master program ([https://online.illinois.edu/mcs-ds](https://online.illinois.edu/mcs-ds))
Topic Coverage: CS512 @ 2017

- Class introduction (0.5 wk)
- An overview on recent data mining research (0.5 wk)
- Text mining: phrase mining and entity typing (2 weeks)
- Textcube construction and exploration (2 weeks)
- 1st midterm exam (0.5 week) — 1st Lect. of 6th week
- Mining heterogeneous information networks (3 weeks)
- Truth finding (1 week)
- Mining social media and spatiotemporal data (1 week)
- Stream data mining (1 week) if time permits
- Selected class survey presentation (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:**
  - Ahmed El-Kishky (25%), Carl Yang (25%), Qi Zhu (50%, online TA), Honglei Zhuang (50%, lead TA)

- **Prerequisites (course preparation):**
  - CS412 (offered every semester) or consent of instructor
  - General background: Knowledge on statistics, machine learning, and data and information systems 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)

- **Textbook:**
  - 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
  - A set of recent published research papers (see course syllabus)
  - J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques, 3rd ed.*, Morgan Kaufmann, 2011
Textbook & Recommended Reference Books

Textbook
- Yizhou Sun and Jiawei Han, *Mining Heterogeneous Information Networks: Principles and Methodologies*, Morgan & Claypool, 2012
- Jiawei Han, Micheline Kamber, Jian Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2011

Recommended reference books

Reference papers
- A list of reference papers will be made available at course “resource” page
Course Work: Assignments, Exams and Course Project

- **Assignments:** (2 assignments, equal weight) **15%** total
- **Two midterm exams** (equal weight): **40%** in total
- **Research project proposal (one-page):** 0% (due at the end of 4th week)
- **Class attendance (3%):** Max misses w/o penalty: 3, then −0.3% for each miss
  - For online students, 3% will be folded into survey report
- **Class presentation and/or research survey (12% total)**
  - **Survey report** [expect to be comprehensive and in high quality, ≈ 20 pages]
    - Encourage to align with your research project topic domain
    - Report plus companion presentation slides [due at the end of 12th week]
  - **Class presentation:** May use 10 min. class survey presentation to replace the survey report (consent of instructor)—contents must closely aligned with the class content and in very high technical quality
- **Final course project: 30%** (due at the end of semester)
  - Evaluated by class (50%) and TA + instructor (50%) collectively!
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 list of papers at the end of each set of chapter slides
- Major conference proceedings that will be used
  - DM conferences: ACM SIGKDD (KDD), ICDM (IEEE, Int. Conf. Data Mining), SDM (SIAM Data Mining), ECMLPKDD (Principles KDD), PAKDD (Pacific-Asia)
  - DB conferences: ACM SIGMOD, VLDB, ICDE
  - ML conferences: NIPS, ICML
  - IR and Web conferences: SIGIR, CIKM, WWW, WSDM
  - Social network confs: ASONAM
- Other related conferences and journals
  - IEEE TKDE, ACM TKDD, DMKD, ML
- Use course Web page, DBLP, Google Scholar, Citeseer
From Data to Networks to Knowledge: An Evolutionary Path!