An Overview of CS512 @Spring 2017

JIAWEI HAN
COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

JANUARY 12, 2017
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

- Course: Data Visualization
  Instructor: John Hart

- Course: Pattern Discovery in Data Mining
  Instructor: Jiawei Han

- Course: Text Retrieval and Search Engines
  Instructor: ChengXiang Zhai

- Course: Cluster Analysis in Data Mining
  Instructor: Jiawei Han

- Course: Text Mining and Analytics
  Instructor: ChengXiang Zhai

- Course: Data Integration and Data Warehousing
  Instructor: Kevin Chang

- Course: Capstone Data Mining Capstone (6 weeks)

- Online MCS Data Science Master program ([https://online.illinois.edu/mcs-ds](https://online.illinois.edu/mcs-ds))
Another Related Course: CS 598 tar: Social Sensing

- CS 598 tar: Social Sensing (Tarek Abdelzaher) Wed/Fri: 12:30-1:45pm

- An important emerging category of cyber-physical systems are those that function in social spaces. These systems are ushered in by the proliferation of "sensing" devices (e.g., phones, wearables, and "smart home" technologies) in the possession of the average individual, as well as our ubiquitous mobile connectivity and the rise of social media (e.g., Twitter and Instagram).

- New data fusion systems are envisioned that will use social networks as sensor networks, and will integrate inputs from IoT devices, physical sensors, mobile phones, and social media to develop myriads of new services and applications for future smart cities and connected communities.

- This paper-reading course covers the unfolding research challenges and directions in distributed social sensing, and overviews the broader landscape of cyber-physical systems, discusses common misconceptions, presents the underlying theoretical foundations, and sheds light on related recent technologies and publications.

- The course includes a group project (in groups of 2 or 3) that will utilize one or more of the social sensing technologies (e.g., sensors, smart-phones, or APIs to social media such as Twitter and Instagram) to build novel applications or services. In the past, many projects resulted in publications. Budget will be available for individual projects to buy devices needed, if not already on hand.
Topic Coverage: CS512 @ 2017

- Class introduction (0.5 wk)
- An overview on recent data mining research (0.5 wk)
- Introduction to networks (1.5 weeks)
- Mining heterogeneous information networks (3 weeks)
  - 1\textsuperscript{st} midterm exam (0.5 week) — 1\textsuperscript{st} Lect. of 7\textsuperscript{th} week
- Text mining (3 weeks)
- Truth finding (1 week)
  - Spring break (1 week)
- Mining social media and spatiotemporal data (1 week)
- Stream data mining (1 week) if time permits
- Selected class survey presentation (1 week)
  - 2\textsuperscript{nd} midterm exams (0.5 week) — 2\textsuperscript{nd} Lect. of 15\textsuperscript{th} 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**:
  - Shi Zhi (shizhi2), Xiang Ren (xren7), Wenqi He, Dongming Lei

- **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**:
  - Yizhou Sun and Jiawei Han, Mining Heterogeneous Information Networks: Principles and Methodologies, Morgan & Claypool, 2012
  - J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, 2011
  - A set of recent published research papers (see course syllabus)
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) **10%** total
- Two midterm exams: (20% each) **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
- **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:** **35%** (due at the end of semester)
  - Evaluated by class (50%) and TA + instructor (50%) collectively!
Research Projects Evaluation

- **Final course project:** 35% (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!

- Han, Kamber and Pei, Data Mining, 3rd ed. 2011
- Sun and Han, Mining Heterogeneous Information Networks, 2012
  Y. Sun: SIGKDD’13 Dissertation Award
- Wang and Han, Mining Latent Entity Structures, 2015
  C. Wang: SIGKDD’15 Dissertation Award