CSE692: Adv Topics in CS
Systems Design for Massive Data

Fall 2007
Outline

✦ Administrative Information
✦ Course Schedule
✦ Grading
✦ About the instructor
✦ Discussion
Administrative Information

- CSE692: Systems Design for Massive Data
- 3-credit, Tu Th 12:50-2:10pm, CS 1441
- course mailing list cse692@cs.
- Instructor: Qin (Christine) Lv
  - CS 1411, 2-8426, qlv@cs.
- Office hours: Tu 2:10-3pm
Course Summary

✦ Efficient systems for managing and exploring massive amounts of digital data

✦ Systems
  ✦ search systems, storage systems, P2P

✦ Algorithms
  ✦ bloom filters, sketching, indexing

✦ Applications
  ✦ multimedia, bioinformatics, scientific data
Course Schedule

- CSE692 course website
- Tentative schedule, subject to change
- Suggestions for other topics, readings?
- Sign up for presentations
Outline

✦ Administrative Information
✦ Course Schedule
✦ Grading
✦ About the instructor
✦ Discussion
Grading

✦ Paper review (20%)
✦ Class participation (20%)
✦ Paper presentation (20%)
✦ Course project (40%)
Paper Review (20%)

✦ Each class

✦ 2-3 papers in the reading list

✦ pick one paper to review

✦ due at 5pm the day before

✦ first review due at 5pm 9/5 (tomorrow)
Paper Review (20%)

✦ A paper review may contain the following

✦ What is this paper about?
✦ What are the strengths and weaknesses?
✦ Are the evaluations convincing?
✦ Can you improve their technique?
✦ Can you apply their technique to other domains/problems?
✦ Any other observations/questions
Class Participation (20%)

✦ Class attendance
✦ Read all papers in reading list
  ✦ be prepared to answer questions in class
✦ Participate in discussions
  ✦ listen to others, ask questions, tell us your thoughts
  ✦ presentation style, suggestions
Paper Presentation (20%)

- One or two presentations per student
- Also counts as a paper review
- 45-minute presentation + discussion
  - motivation, background, related work
  - main techniques & evaluations
  - possible applications, improvements
- Meet w/ instructor one day before class
Course Project (40%)

✦ No midterm or final exam

✦ Important Dates

✦ proposal due: Oct 18
✦ checkpoint: Nov 20
✦ presentation: Dec 11
✦ final report due: Dec 13

✦ Start early!
Course Project (40%)

✧ A self-contained project related to this course’s topics
✧ Work alone
✧ Get instructor’s permission if work in pairs
✧ Possible project ideas will be posted at course website
✧ Students may also pick their own topics
✧ Talk to instructor
Project Proposal (Oct 18)

- A 15-minute presentation
  - motivation
  - literature survey
  - your technique
  - how to evaluate
  - milestones
- Submit a 3-page project proposal
Project Checkpoint (Nov 20)

✦ A 15-minute presentation
✦ proposal review: motivation, your technique, evaluation, milestones
✦ what you’ve achieved so far
✦ what remains to be done
✦ Submit a progress report
✦ updated version of your initial proposal
✦ highlight your progresses
Project Presentation (Dec 11)

✦ A 25-minute presentation
  ✦ motivation, literature survey, your technique, evaluation, conclusions, future work
  ✦ Presentations will be peer-reviewed
  ✦ technical depth, evaluations
  ✦ presentation: style, clarity
Follow the format of a regular research paper

- title, abstract
- introduction, related work
- main technique, evaluation
- conclusion, future work
- references
Academic Honesty

✦ Be personally accountable for all your submitted work.

✦ No academic dishonesty will be tolerated.

✦ Any suspected instance will be reported to the Computer Science Department and may be forwarded to the Graduate School.
Outline

- Administrative Information
- Grading
- Course Schedule
- About the instructor
- Discussion
Qin (Christine) Lv

- 2000: B.E. in Computer Science & Technology, Tsinghua University, China
- 2006: Ph.D. in Computer Science, Princeton University, USA
- 2006-2007: Postdoc, Princeton University
- Sep. 2007:
  - Assistant Professor, Stony Brook Univ.
Research Interests

- Develop efficient systems for managing and exploring massive amounts of digital data
- Search systems, data management, distributed systems, storage systems, networking
- Systems, algorithms, applications
Research Projects

✦ Networking
  ✦ self-organized (ad-hoc) networks
  ✦ performance monitoring & optimization

✦ Peer-to-peer networks
  ✦ search, replication, heterogeneity

✦ Storage systems
  ✦ content-addressable, distributed B-tree
Research Projects

✦ CASS: Content-Aware Search Systems

✦ feature-rich data: audio, video, digital photos, genomic data, scientific sensor data, ...

✦ content-based similarity search

✦ $L_1$ sketching, multi-probe LSH indexing, Ferret toolkit
Research Projects

✦ Massive Data Systems Lab

✦ distributed search systems
✦ similarity-aware storage systems
✦ data stream processing systems
✦ healthcare data management
✦ scientific data management
✦ and many more
Discussion

✦ Class meeting time
  ✦ conflicts?
  ✦ twice a week or once a week?
✦ Other topics to cover?
✦ Questions? Suggestions?
✦ Sign up for presentations
Next Lecture (9/6)

- Data, data, data
- MyLifeBits
- Memex
- How much information? 2003
- Paper review due at 5pm, 9/5
- Class starts at 1:10pm