INF 553: Foundations and Applications of Data Mining

Wensheng Wu, Computer Science
Yao-Yi Chiang, Spatial Sciences Institute

Thanks for source slides and material to: Ann Chervenak, J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Basic Course Information

- **Lectures**
  - Tuesday/Thursday
  - **Sec. 32423D:** 9:30-10:50; Sec. **32444D:** 5-6:20pm
  - Both meet at KAP 163
- **Instructor**
  - Wensheng Wu (wenshenw@usc.edu)
  - Yao-Yi Chiang (yaoyic@usc.edu)
- **Office Hours**
  - Wensheng, 11am-12pm (Tu and Th), GER 204
  - Yao-Yi, Tuesday after class, AFH B55C
- **Grader/Course Supervisor**
  - Pooja Anand, poojaana@usc.edu
  - Siddharth Mahendra Dasani, sdasani@usc.edu

Outline of Course

- Map-Reduce and Hadoop
  - Infrastructure for analyzing/mining “big-data”
- Association rules (e.g., market basket analysis)
  - Apriori algorithm
- Finding similar items (e.g., web pages)
  - Minhashing: reduce large sets into small signatures
  - Locality-sensitive hashing: use signatures to estimate similarity
- Recommendation systems (e.g., suggest news, products)
  - Collaborative filtering (e.g., NetFlix challenge)
- Link analysis
  - PageRank (combating term spams)
  - TrustRank (topic-sensitive PageRank, combating link spams)

Outline (cont.)

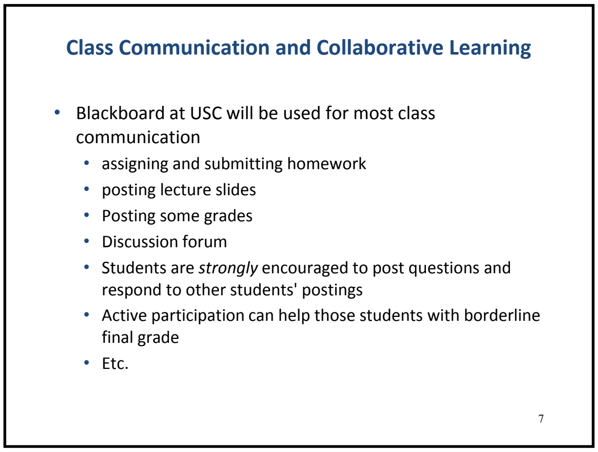
- Clustering data
  - Large amount of data
  - High-dimensional
  - Non-Euclidean
- Social network analysis (e.g., finding communities)
  - Edge betweenness
  - Spectral clustering
- Managing Web advertisements
  - Direct placement, e.g., ebay, craigslist
  - Display ads
  - Search ads

Outline (cont.)

- Mining data streams (e.g., sensors, satellite images, clickstream, etc.)
  - Sampling
  - Filtering
  - Counting/estimating distinct values (when # of distinct values is too large to fit in main memory)

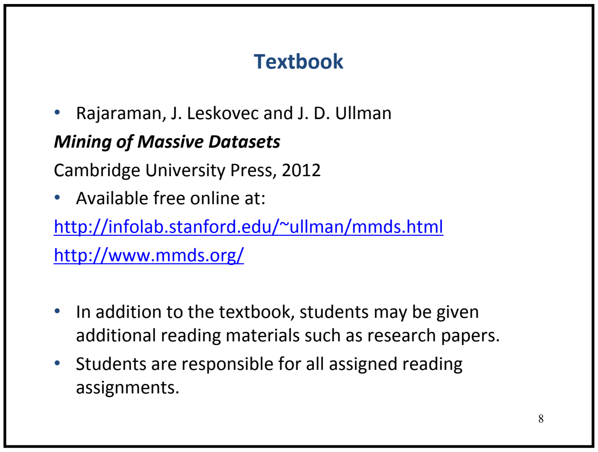
Prerequisites

- A basic understanding of engineering principles
- Programming skills
  - Familiarity with the Python language is desirable
  - Code Academy Python tutorials: http://www.codecademy.com/tracks/python
  - Google Python Class: https://developers.google.com/edu/python/
- Most assignments are designed for the Unix environment
  - Basic Unix skills will make programming assignments easier
- Mathematical background: probability, statistics, and linear algebra
- Some knowledge of machine learning is helpful, but not required



Class Communication and Collaborative Learning

- Blackboard at USC will be used for most class communication
  - assigning and submitting homework
  - posting lecture slides
  - Posting some grades
  - Discussion forum
  - Students are *strongly* encouraged to post questions and respond to other students' postings
  - Active participation can help those students with borderline final grade
  - Etc.



Textbook

- Rajaraman, J. Leskovec and J. D. Ullman

Mining of Massive Datasets

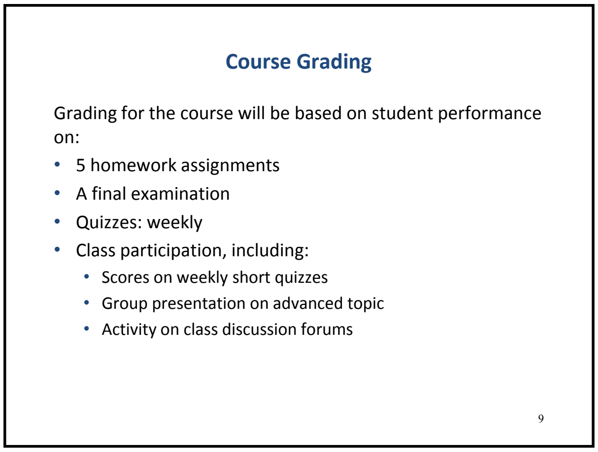
Cambridge University Press, 2012

- Available free online at:

http://infolab.stanford.edu/~ullman/mmds.html

http://www.mmds.org/

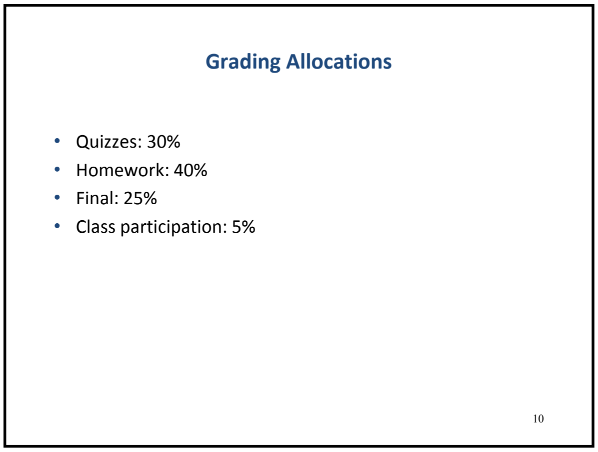
- In addition to the textbook, students may be given additional reading materials such as research papers.
- Students are responsible for all assigned reading assignments.



Course Grading

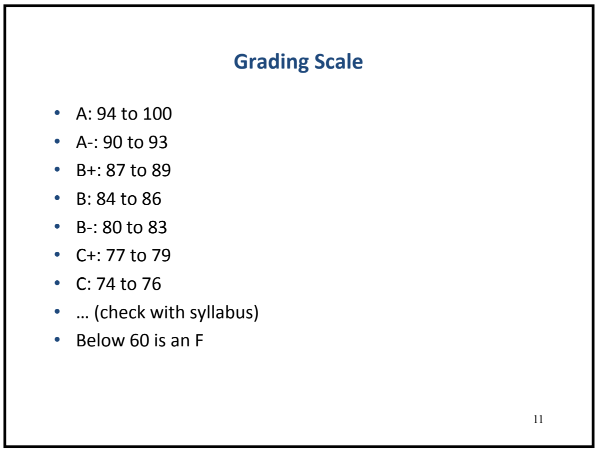
Grading for the course will be based on student performance on:

- 5 homework assignments
- A final examination
- Quizzes: weekly
- Class participation, including:
  - Scores on weekly short quizzes
  - Group presentation on advanced topic
  - Activity on class discussion forums



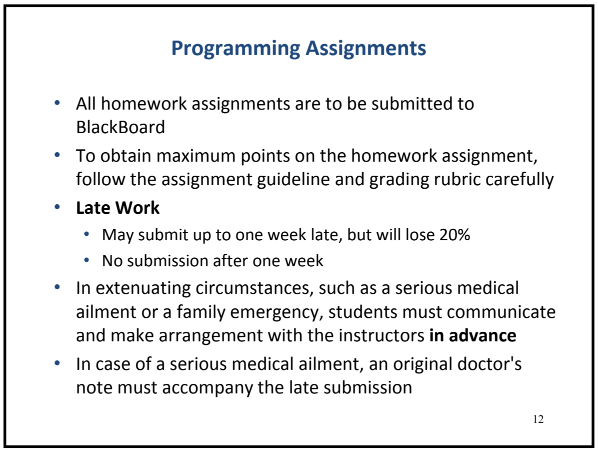
Grading Allocations

- Quizzes: 30%
- Homework: 40%
- Final: 25%
- Class participation: 5%



Grading Scale

- A: 94 to 100
- A-: 90 to 93
- B+: 87 to 89
- B: 84 to 86
- B-: 80 to 83
- C+: 77 to 79
- C: 74 to 76
- … (check with syllabus)
- Below 60 is an F



Programming Assignments

- All homework assignments are to be submitted to BlackBoard
- To obtain maximum points on the homework assignment, follow the assignment guideline and grading rubric carefully
- **Late Work**
  - May submit up to one week late, but will lose 20%
  - No submission after one week
- In extenuating circumstances, such as a serious medical ailment or a family emergency, students must communicate and make arrangement with the instructors **in advance**
- In case of a serious medical ailment, an original doctor's note must accompany the late submission

Class Participation

- There are 4 components for the class participation grade
- Paying attention, asking and answering questions in class
- Weekly quizzes on the previous lessons
  - The quizzes are designed to enforce class attendance, participation, and attention
  - Should be straightforward if you have attended class, kept up with the reading and done the programming assignments yourself
- Group presentation on an advanced topic
- Students are expected to actively participate in the class forum
  - For example: asking questions and posting answers to other students' questions

Grading Corrections

- **Grades are not negotiable**
- Any student who wastes the instructor's time with non-legitimate requests for additional points on an assignment or exams risks losing additional points as well as having their behavior affect their class participation grades

- Any legitimate request for re-grading **must be submitted in writing, with carefully worked out explanation** of why it is believed that an assignment has not been properly graded.

Academic Integrity

- **Cheating will not be tolerated**
- **All parties involved will receive a grade of F for the course and be reported to SJACS** (WITHOUT EXCEPTION)
- It is fine to answer questions from other students on the class discussion board, but DO NOT post your solution to an assignment.
- **We will be using the Moss system to detect software plagiarism**

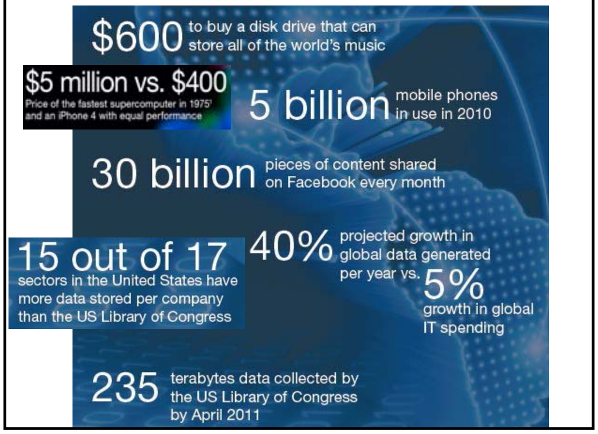
  http://theory.stanford.edu/~aiken/moss/
- If you have questions or concerns regarding what is permitted in terms of collaboration or teamwork, please ask the instructor/grader for clarifications

Example Moss Output

What is Data Mining?

What is Data Mining?

Knowledge Discovery from Data



$600 to buy a disk drive that can store all of the world's music

$5 million vs. $400
Price of the fastest supercomputer in 1975' and an iPhone 4 with equal performance

5 billion mobile phones in use in 2010

30 billion pieces of content shared on Facebook every month

15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress

40% projected growth in global data generated per year vs. 5% growth in global IT spending

235 terabytes data collected by the US Library of Congress by April 2011

Some Additional Information

- 97 million songs in the world (according to MusicHype CEO Kevin King)
  - http://www.marsbands.com/2011/10/97-million-and-counting/
- Keynote video by Kevin Kelly, senior Maverick, Wired
  - http://www.web2expo.com/webexsf2011/public/schedule/proceedings.2
  - 6TB enough to hold all the songs
- Big data: The next frontier for innovation, competition, and productivity, Mckinsey report, 2011.
  - http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

20

POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
PLUS
Juan Enriquez Reprograms Life
James Gleick Unsplits the Bit
AND Lawrence Weschler Questions the Cloud
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE

Data contains value and knowledge

Data Mining

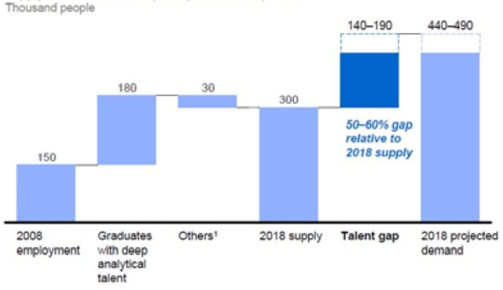
- But to extract the knowledge data needs to be
  - Stored
  - Managed
  - And ANALYZED ← this class

Data Mining ≈ Big Data ≈
Predictive Analytics ≈ Data Science

Good news: Demand for Data Mining

Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018
Thousand people



[1] Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

What is Data Mining?

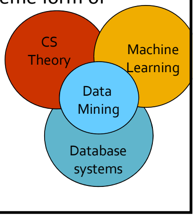
- Given lots of data
- Discover patterns and models that are:
  - Valid: hold on new data with some certainty
  - Useful: should be possible to act on the item
  - Unexpected: non-obvious to the system
  - Understandable: humans should be able to interpret the pattern

Data Mining Tasks

- **Descriptive methods**
  - Find human-interpretable patterns that describe the data
    - **Example:** Clustering
- **Predictive methods**
  - Use some variables to predict unknown or future values of other variables
    - **Example:** Recommender systems

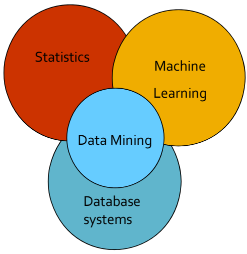
Data Mining: Cultures

- **Data mining overlaps with:**
  - **Databases:** Large-scale data, complex queries
  - **Machine learning:** Small data, complex models
  - **CS Theory:** (Randomized) Algorithms
- **Different cultures:**
  - To a DB person, data mining is an extreme form of **analytic processing** – queries that examine large amounts of data
    - Result is the query answer
  - To a ML person, data-mining is the **inference of models**
    - Result is the parameters of the model



This Class

- **This class overlaps with machine learning, statistics, artificial intelligence, databases but more stress on**
  - **Scalability** (big data)
  - **Algorithms**
  - **Computing architectures**
  - Automation for handling **large data**



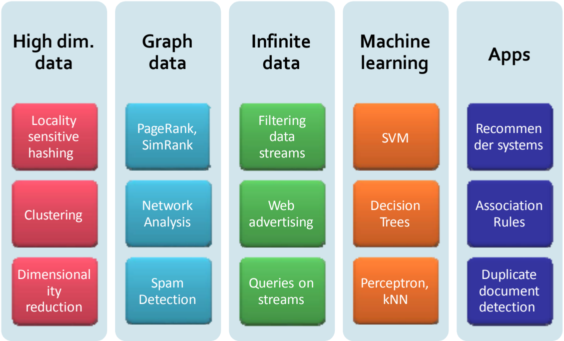
What will we learn?

- **We will learn to mine different types of data:**
  - Data is high dimensional
  - Data is a graph
  - Data is infinite/never-ending
  - Data is labeled
- **We will learn to use different models of computation:**
  - MapReduce
  - Streams and online algorithms
  - Single machine in-memory

What will we learn?

- **We will learn to solve real-world problems:**
  - Recommender systems
  - Market Basket Analysis
  - Spam detection
  - Duplicate document detection
- **We will learn various "tools":**
  - Linear algebra (SVD, Rec. Sys., Communities)
  - Dynamic programming (frequent itemsets)
  - Hashing (LSH, Bloom filters)
  - Optimization (stochastic gradient descent)

How the Class Fits Together



High dim. data	Graph data	Infinite data	Machine learning	Apps
Locality sensitive hashing	PageRank, SimRank	Filtering data streams	SVM	Recommender systems
Clustering	Network Analysis	Web advertising	Decision Trees	Association Rules
Dimensionality reduction	Spam Detection	Queries on streams	Perceptron, kNN	Duplicate document detection

Modeling Data: Summarization

- Summarize the data
- PageRank (Chapter 5)
  - Structure of the web is summarized by a single number for each web page (its PageRank)
  - Probability that a random walker on the graph would be on the page at any given time
  - Property: the PageRank reflects the *importance* of the page
  - How much a typical searcher wants that page returned as an answer to their search query

Modeling Data: Summarization

- Clustering (Chapter 7)
  - Data viewed as points in multidimensional space
  - Points “close” in space assigned to same cluster
  - Clusters are summarized: e.g., by centroid of cluster and average distance from centroid of points in cluster
  - Cluster summaries become summary of data set
  - Example: identify clusters of cholera cases around London intersections due to contaminated wells

Modeling Data: Feature Extraction

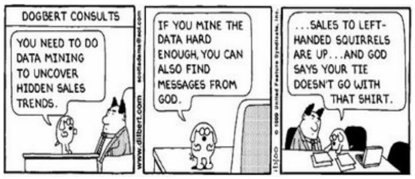
- A complex relationship between objects is represented by finding the strongest statistical dependencies among objects and using those to represent statistical connections
- Frequent itemsets (Chapter 6)
  - Model for data that consists of “baskets” of small sets of items (e.g., for brick and mortar stores or shopping sites)
  - Look for small sets of items that appear together in many baskets
  - These “frequent itemsets” characterize the data
  - Identify sets of items that people tend to buy together, can use to set prices
    - E.g., diapers and beer

Modeling Data: Feature Extraction

- Similar items (Chapter 3)
  - Data looks like a collection of sets
  - Objective: find pairs of sets that have fairly large number of items in common
  - E.g. represent customers as set of items they have bought, look for similar customers to *recommend* additional items those customers have bought
  - Or, for each customer, identify small number of customers with similar tastes
  - Recommendation Systems, Chapter 9

Meaningfulness of Analytic Answers

- A risk with “Data mining” is that an analyst can “discover” patterns that are meaningless
- Statisticians call it **Bonferroni’s principle**:
  - Roughly, if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap – avoid treating random occurrences as if they were real


DOGBERT CONSULTS
YOU NEED TO DO DATA MINING TO UNCOVER HIDDEN SALES TRENDS.
scottadams@aol.com
www.dilbert.com
IF YOU MINE THE DATA HARD ENOUGH, YOU CAN ALSO FIND MESSAGES FROM GOD.
© 1999 United Feature Syndicate, Inc.
...SALES TO LEFT-HANDED SQUIRRELS ARE UP...AND GOD SAYS YOUR TIE DOESN'T GO WITH THAT SHIRT.

Example of Bonferroni’ s Principle

A big objection to Total Information Awareness (searching for suspicious activity) was that it was looking for so many vague connections that it was sure to find things that were bogus and thus violate privacy of innocent people

Scenario

- Suppose we believe that certain groups of evil-doers are meeting occasionally in hotels to plot doing evil.
- We want to find (unrelated) people who at least twice have stayed at the same hotel on the same day.

The Details

- 10^9 (1 billion) people being tracked
- 1000 days
- Each person stays in a hotel 1% of the time (10 days out of 1000).
- Hotels hold 100 people
- 10^5 hotels to hold 1% of 10^9 people
  - $10^5 \times 100 = 10^7$
- If everyone behaves randomly (i.e., no evil-doers) will the data mining detect anything suspicious?

Calculations

- Probability that given persons p and q will be at the same hotel on given day d:
  - $\boxed{\overbrace{1/100}^{p \text{ at some hotel}}} \times \boxed{\overbrace{1/100}^{q \text{ at some hotel}}} \times \boxed{\overbrace{10^{-5}}^{\text{Same hotel}}} = 10^{-9}$
- Probability that p and q will be at the same hotel on given days d_1 and d_2:
  - $10^{-9} \times 10^{-9} = 10^{-18}$
- Recall: for large n, $\binom{n}{2}$ is about $n^2/2$
- Pairs of days: $\binom{1000}{2}$ is about 5×10^5

Calculations

- Probability that p and q will be at the same hotel on some two days:
  - (number of pairs of days) x (prob p and q at same hotel for 2 days)
  - $5{\times}10^5 \times 10^{-18} = 5{\times}10^{-13}$
- Pairs of people:
  - 10^9 people
  - About $n^2/2$ pairs of people: $5{\times}10^{17}$
- Expected number of "suspicious" pairs of people:
  - $5{\times}10^{17} \times 5{\times}10^{-13}$ = 250,000

Conclusion

- Suppose there are (say) 10 pairs of evil-doers who definitely stayed at the same hotel twice
- Analysts have to sift through 250,010 candidates to find the 10 real cases.
  - Not realistic
  - But how can we improve the scheme? (see also Ex. 1.2.1)

Moral

- When looking for a property (e.g., "two people stayed at the same hotel twice"), make sure that the property does not allow so many possibilities that random data will surely produce facts "of interest."