DSCI 553
Foundations and Applications of Data Mining

Professor Wei-Min Shen

University of Southern California

USC Viterbi School of Engineering

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

Outline

- Introduction of Teaching Team
- About This Course
- Introduction to Data Mining
- Map-Reduce (Part I)

DSCI-553 2023-S Teaching Team

- **Professor Wei-Min Shen**
  - Office Hours
    - After Lecture (by email appointment)
- **Teaching Assistants and Course Producers**
  - Office Hours
    - See Blackboard and Piazza (resources/staff)

DSCI-553 2023-S Teaching Team

	Name	Office Hours	
	wmshen@usc.edu	When?	After Lecture by appointment
		Where?	After Lecture by appointment
	Shaswat Anand	When?	
		Where?	
	Sahana Kalakonda	When?	Monday: 8 AM - 9 AM, Monday: 12 PM - 1PM
		Where?	https://usc.zoom.us/j/3682583875?pwd=UUI1eHhjOFVubXNNRm40TXltWGJSUT09
	Lokesh Sipani	When?	
		Where?	
	Alex Bisberg	When?	Tue/Thu 10am-11am
		Where?	https://usc.zoom.us/j/8018679775?pwd=YWRuTEw2MWRmSTV0ZVp1bWRjekNjdz09
	Yilei Zeng	When?	
		Where?	
	Viraj Krishnakant Mehta	When?	
		Where?	
	Hetvi Shah	When?	
		Where?	
	Jheel Ketan Patel	When?	
		Where?	
	Akash Ram Praveen Raj	When?	Tuesday: 9 AM-10 AM , Friday:3 PM-4 PM
		Where?	https://usc.zoom.us/j/6431475007?pwd=UXA3Wk9rN0w1V1NOOCsyckR0ZTE2QT09
	Henil Shelat	When?	
		Where?	
	Cole Howard	When?	
		Where?	
	Gautam Pranjal	When?	Monday (6 PM - 7 PM) + Friday (5 PM to 6 PM)
		Where?	https://usc.zoom.us/j/3042958534?pwd=enE5UWR3OS9XTlZPUHpPaW8zWTlsQT09
	JINGPING YU	When?	
		Where?	
	Yunhe Wang	When?	
		Where?	

Please introduce yourself !

What is Data Mining?
About THIS Course

What is Data Mining?
Knowledge Discovery from Data

Data Mining

- But to extract the knowledge, data needs to be
  - Stored
  - Managed
  - And ANALYZED <= this class

Big Data Lifecycle

Data Mining ≈ Big Data ≈
Predictive Analytics ≈ Data Science

The Challenges of Big Data?

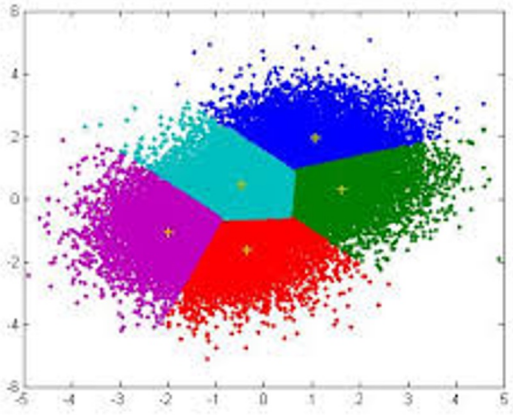
- Cannot store in one place
  - Must be distributed
- Failures for access may be unexpected
- Unpredictable diversity
- Messy, noisy, and errors are inevitable
- Dynamic
- ……

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

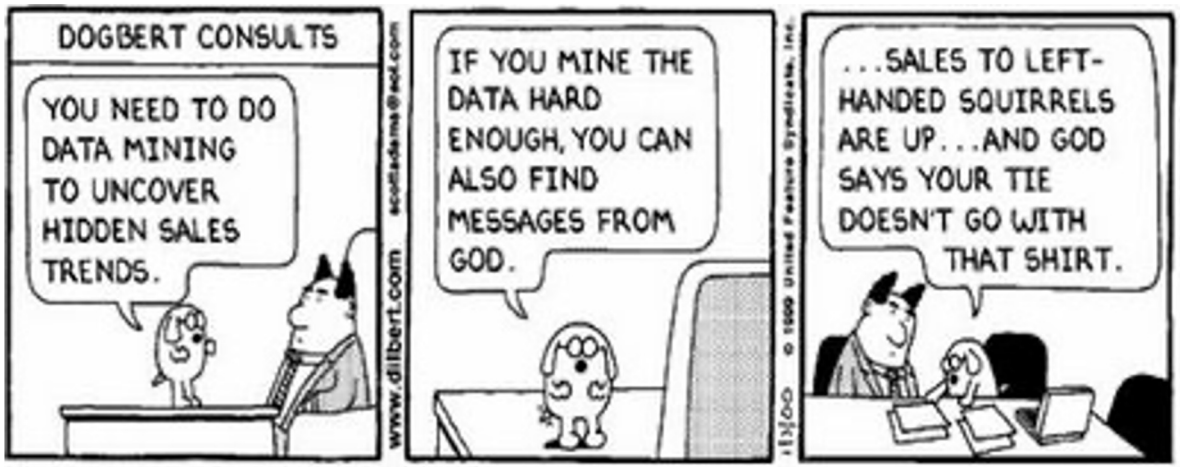
- **Descriptive methods**
  - Find human-interpretable patterns that describe the data
    - **Example:** Clustering



- **Predictive methods**
  - Use some variables to predict unknown or future values of other variables
    - **Example:** Recommender systems

Meaningfulness of Analytic Answers

- **A risk with "Data mining" is that an analyst can "discover" patterns that are meaningless**
- **Bonferroni's principle**:
  - If you look in more places for interesting patterns than your amount of data will support, you are bound to find some "craps"!



Meaningfulness of Analytic Answers

Example:

- We want to find (unrelated) people who **at least twice have stayed at the same hotel on the same day**
  - 10^9 people being tracked – 1 billion
  - 1,000 days ~ 3 years
  - Each person stays in a hotel 1% of time (1 day out of 100)
  - Hotels hold 100 people (so 10^5 hotels)
    - enough to hold the 1% of a billion people who visit a hotel on any given day
  - **If everyone behaves randomly will the data mining detect anything suspicious?**

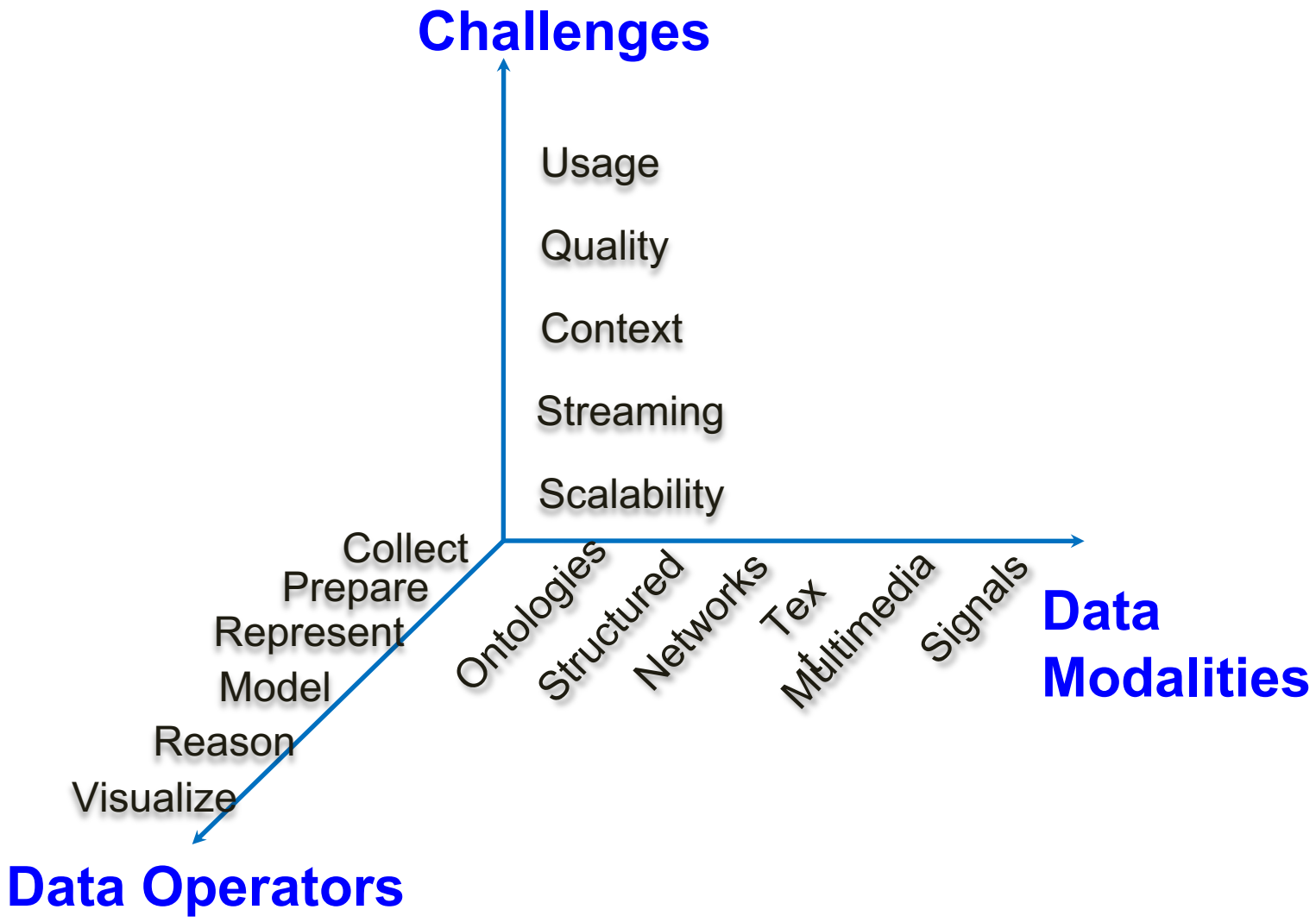
Meaningfulness of Analytic Answers (Cont'd)

- 10^9 people, 1,000 days, 1% hotel stay, 10^5 hotels
  - The probability of any two people both deciding to visit a hotel on any given day is 0.0001 (i.e., 1%*1%)
  - The chance that they will visit the ***same*** hotel for one day is

    $0.0001 / 10^5 = 10^{-9}$; for two given days = 10^{-18}

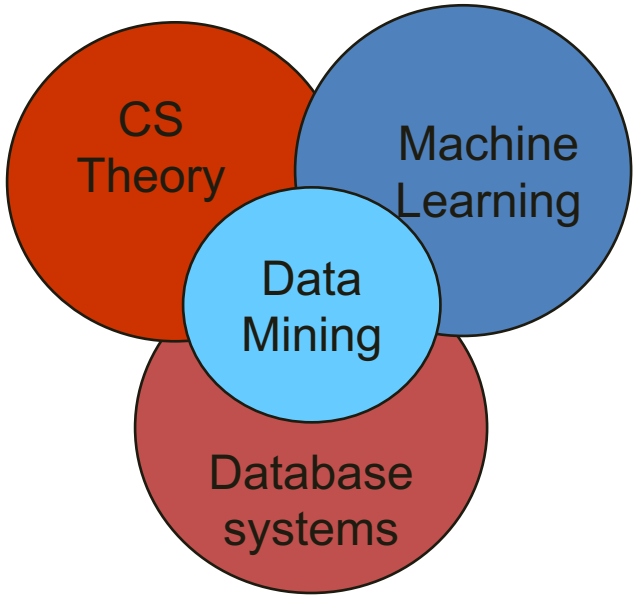
    $$\binom{n}{k} = \frac{n!}{k!(n-k)!}; \quad \text{for large } n, \binom{n}{2} \text{ is about } n^2/2$$
  - The number of pairs of people is $C(10^9, 2) = 5 \times 10^{17}$
  - The number of pairs of days is $C(10^3, 2) = 5 \times 10^5$
  - **Expected number of "suspicious" pairs of people:**
    - $5 \times 10^{17} \times 5 \times 10^5 \times 10^{-18} = 250{,}000$ (Wow!)
    - … too many combinations to check – we need to have some additional evidence to find "suspicious" pairs of people in some more efficient way

What matters when dealing with data?



Data Mining: Cultures

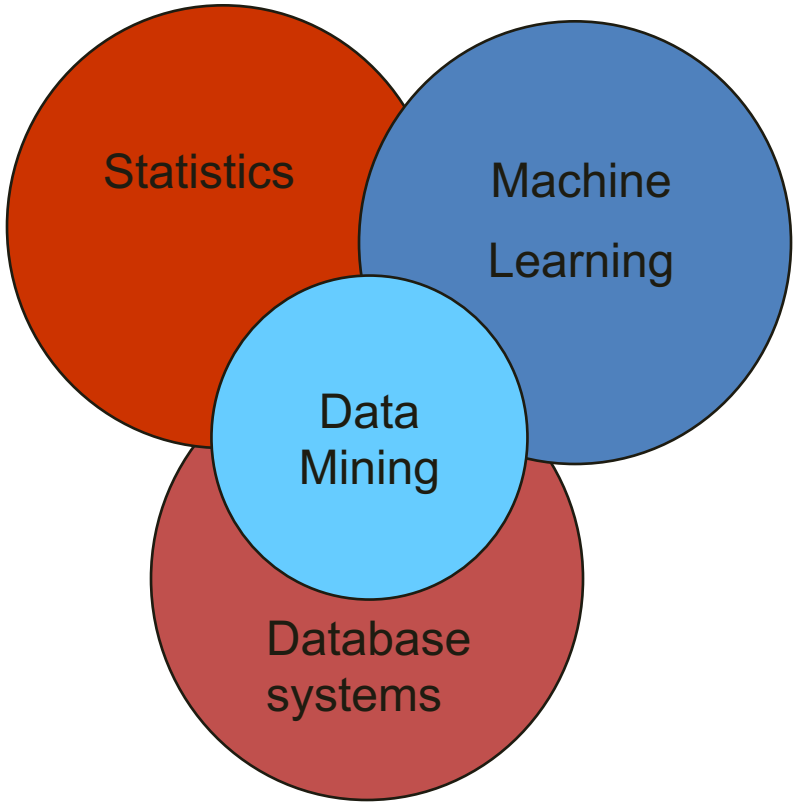
- **Data mining overlaps with:**
  - **Databases:** Large-scale data, simple(r) queries
  - **Machine learning:** Small data, Complex models
  - **CS Theory:** Algorithms
- **Different cultures:**
  - To a DB person, data mining is an extreme form of **analytic processing** – queries that examine **large 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
- **In this class we will do both!**



About THIS Course

This Course

- This course overlaps with machine learning, statistics, artificial intelligence, databases but more stress on
  - **Scalability** (big data)
  - **Algorithms**
  - **Computing Architectures**
  - **Automatic Handling** of large data



What will we learn?

- **We will learn to mine different types of data:**
  - Data are high dimensional
  - Data are graphs
  - Data are infinite/never-ending
  - Data are 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 (Recomm. Sys., Communities)
  - Optimization (stochastic gradient descent)
  - Dynamic programming (frequent itemsets)
  - Hashing (LSH, Bloom filters)

How It All Fits Together

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

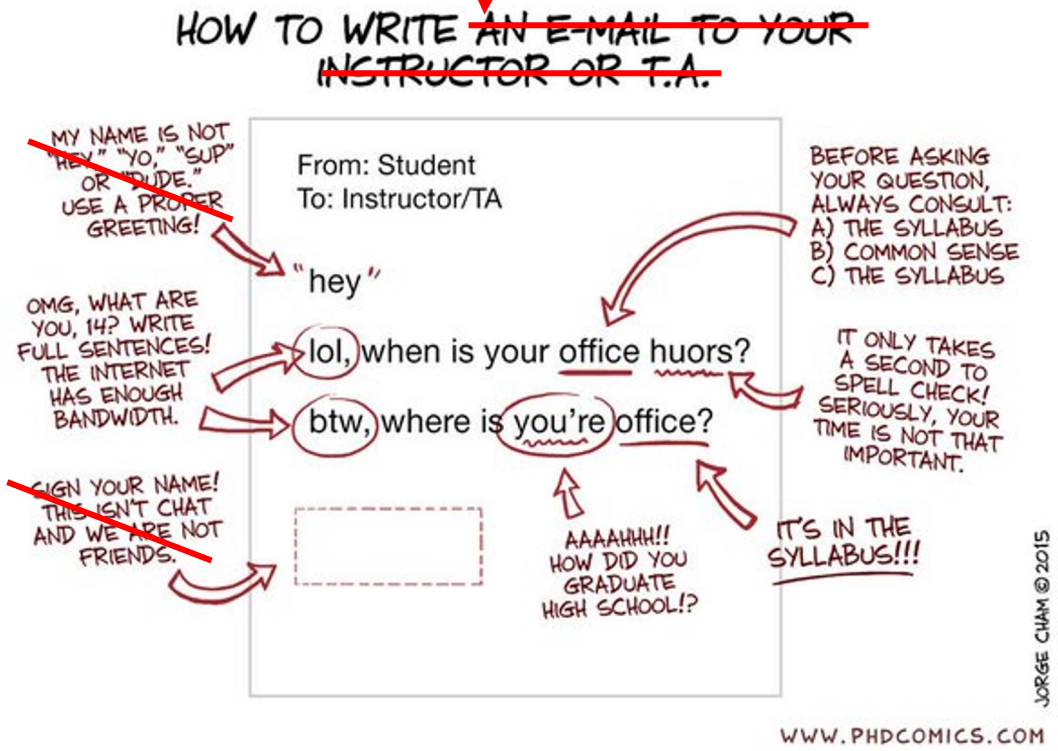
Course Logistics

- **Website:** courses.uscden.net/d2l/
  - Lecture Notes and Video Records on DEN:
    - Post on the same day of the lecture
  - Quizzes, Homework, Exam, Competition
  - Readings
    - **Mining of Massive Datasets**
      - J. Leskovec, A. Rajaraman and J. Ullman
      - Free online: http://www.mmds.org
    - Other relevant papers

Logistics: Communication

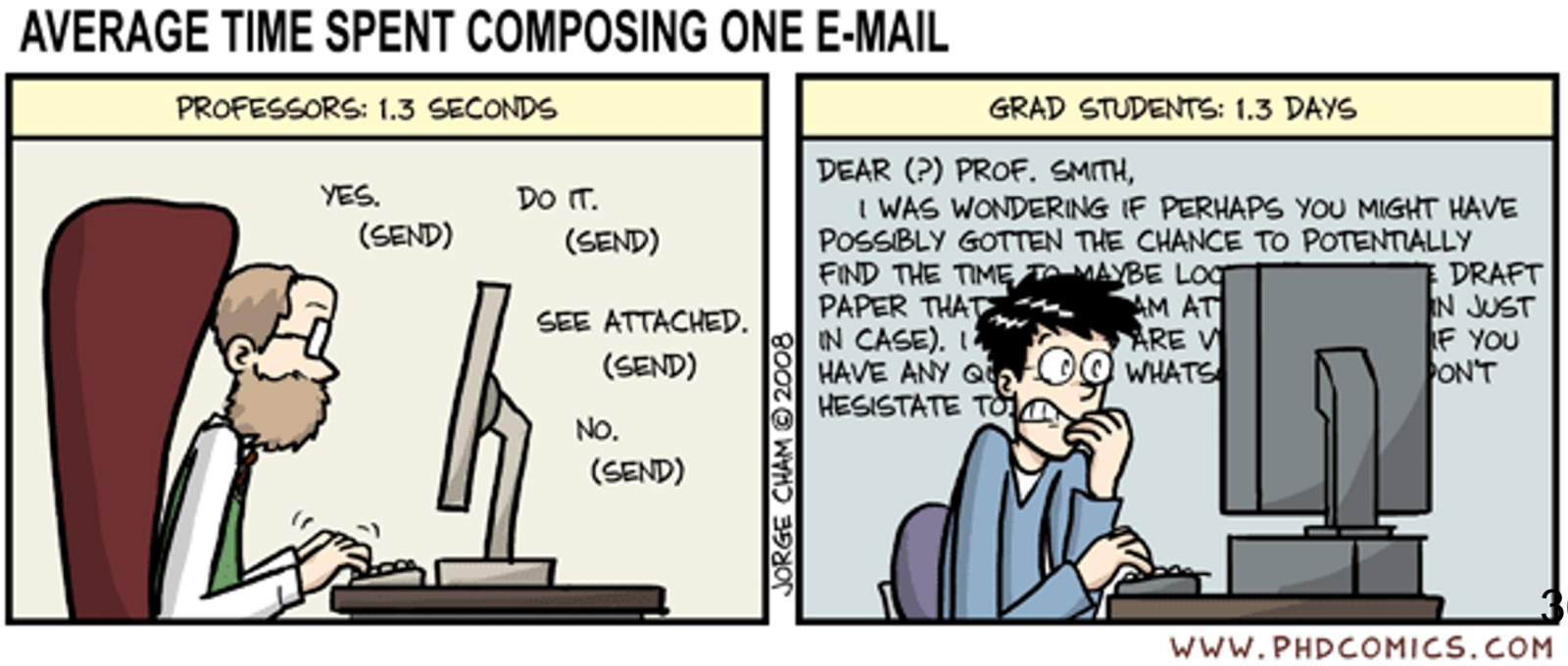
- **Discussion board on Piazza (Blackboard)**
  - Please enroll yourself, and you are all invited on Piazza
    - Questions and public communication with the course staff
    - Help your fellow students by contributing answers What a friend is for!

a post on the discussion board



Logistics: Communication

- We will post course announcements to
  - Piazza and DEN/Blackboard (please check regularly)
- **Emails:**
  - Do not use emails unless it's personal!



Work for the Course

- **Six homework: 42%**
  - Programming assignments on Vocareum.com
    - MapReduce (2 weeks)
    - Frequent Itemsets (2 weeks)
    - Recommendation Systems (3 weeks***)
    - Detecting Communities in a Social Network (2 weeks)
    - Streaming data analysis (1 week)
    - Clustering big datasets (1 week)
  - **Assignments take lots of time. Start early!!**
- **One Grand Competition Project: 8% + bonus**
  - **Recommendation Systems**
- **Please work on your own code! (very smart detective agents we have)**

Work for the Course

- **Homework policy:**
  - On Vocareum.com
    - You can test and submit your code
    - Your work will be auto-graded by scripts
      - Be careful to use the ***exact*** formats!
  - Late Penalty: One week late penalty: –20%,
    - 0 points after one week
  - Your submitted code will be auto-checked by an advanced detective agent for similarities against all other submissions in the class
    - 50% deduction if your submitted code has >75% similarity matches with some other submissions.
  - Free five-day extensions
    - Use the extension days on homework however you want
    - Submit your request for using free-extensions before deadline
    - No more extension days will be given for any other reasons

Work for the Course

- Final Project and Grand Competition
  - We will be building a recommendation system based on homework-3
  - You will continue improving your recommendation system (homework-3) throughout the course
  - Your recommendation system will compete with systems from all other students in the class
- Why are we doing the competition?

Work for the Course

- **Why are we doing the competition?**
  - Learn how to handle real (big) datasets, e.g.,
    - http://grouplens.org/datasets/movielens/
  - You will have the opportunity to put something like this on your resume:
    - First Place, USC Data Mining Competition (202x)
  - I will have something to say if you want me to be your reference ☺

Previous Winners (2022 Fall)

- USC Data Mining Competition (2022F)

1. Zhang Zhiyu
2. Li Chaoyu, Wang Yunhe
3. Tamazian Alain, Minoofar Amir, Siddhart Ganesh
4. Tian Jiayu
5. Ankit Nitin Kumar Shah
6. Lin Chih-Hsien

Previous Winners (2022 Spring)

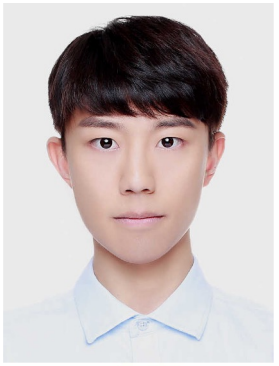
- USC Data Mining Competition (2022S)
  1. Feichi Yang *
  2. Shaswat Anand (your TA)
  3. Jheel Ketan Patel (your TA)
  4. Srinag Vinil Tummala
  5. Shayan Javid Yazdi
  6. Viraj Mehta (your TA)

Previous Winners (2021 Fall)

- USC Data Mining Competition
  1. Li WenXuan (your TA)
  2. Daniel Hao
  3. Tirth Patel (your TA)
  4. Dhruvil Anil Trivedi
  5. Vishal Ajaybhai Kapadia (your TA)
  6. Deep Prakash Amin

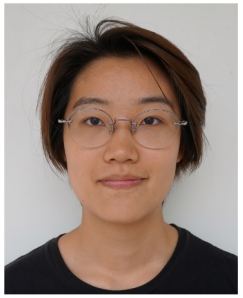
Previous Winners (2021 Spring)

- USC Data Mining Competition
  1. **Zeyang Gong** (0.9721) (your TA)
  2. Jiemin Tang (0.9744)
  3. Nitin Chandra Perumandl (0.9745)
  4. Shiyang Chen (0.9749)
  5. Matheus Schmitz (0.9750)
  6. Yuxin Jiang (0.9752)



Previous Winners (2019)

- USC Data Mining Competition
- First Place, **Yang Zheng**



- Second Place, Peilun Yan

- Third Place, Nivedetha Kumaram

Previous Winners (2018)

- USC Data Mining Competition
- First Place, **Hongtao Yang**

- Second Place, **Chen Lou**

- Third Place, **Bufan Zeng**

Previous Winners (2017)

- USC Data Mining Competition
- First Place, **Priyambada Jain**

- Second Place, **Zihao Zhai**

- Third Place, **Vijayakumar Gedigeri**

Work for the Course

- **Not-So-Short weekly quizzes: 30%**
  - Not-So-Short in-class quizzes every week
  - Covers the material learned in the previous week
  - You may drop two your lowest quizzes
- **Comprehensive exam: 20%**
  - **Tentative Date: Week16,** 7:00 – 9:00 pm
  - The exam will cover everything taught in class
- No final exam but the final project will due online at the scheduled final date / time

- **It's going to be fun and hard work.**

Quiz and Lecture (where and when)

- 5:15 PM (in Class/DEN)
  - In class, please bring your laptop computers
  - DEN: please join the DEN site and turn on your camera as instructed
- 5:30-6:00 PM (in Class/DEN)
  - Must use Lockdown Browser and Password
- 6:00 – 8:50PM (lectures in Class/DEN)
- Notes
  - 2nd Week: Practice Quiz-0 on DEN (pswd=”welcome”)
  - 3nd Week: start the real quizzes

Prerequisites (Hard)

- **Algorithms**
  - Dynamic programming, basic data structures, ….
- **Basic probability**
  - Moments, typical distributions, MLE, …
- **Programming**
  - Python will be required
  - Spark and/or Scala (for homework)
- **We provide some background, but the class will be fast paced**

Course Grade

92 – 100 = A 88 – 92 = A-

85 – 88 = B+ 80 – 85 = B 78 – 80 = B-

75 – 78 =C+ 70 – 75 = C

67 –70 = C- 65 – 67 = D+ 63 – 65 = D 60 – 63 = D-

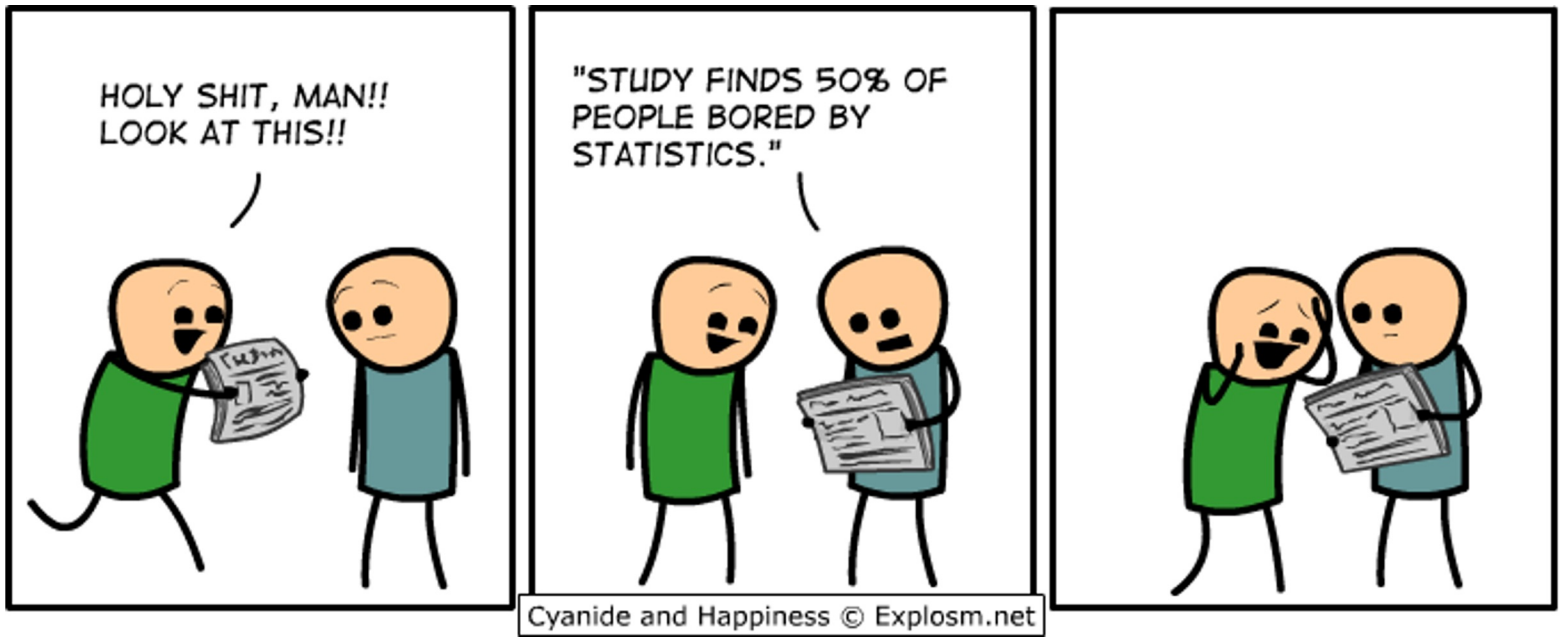
Below 60 is an F

What's After the Class?

- Directed Research
- Course producer or grader positions
- Paid TA positions
- Paid RA positions (maybe)

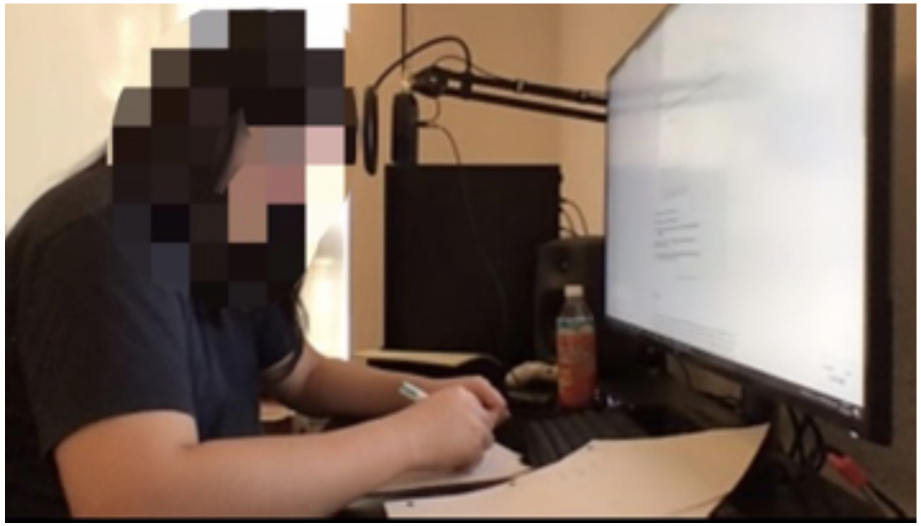
To-do items

- Download the textbook
- Install Spark on your machine (http://spark.apache.org/)
- Get familiar with **Vocareum** and how to use the terminal window
- Play with datasets http://grouplens.org/datasets/movielens/ and http://jmcauley.ucsd.edu/data/amazon/links.html



Schedule around Holidays

- If a lecture day falls on a holiday
  - Please watch the lectures video on DEN
  - Please study the slides and zoom videos yourself
- Quizzes will start on the 3rd week (2nd week practice)
  - Join 15min before, quiz on time for 20 min. lectures after
  - Please arrange your camera as this



About Auditing …

- Auditing the class is fine if we have space
  - Please email me if you want to audit the class
- Check with your advisor for the last day to drop a class without a mark of “W”

Some Backup Slides

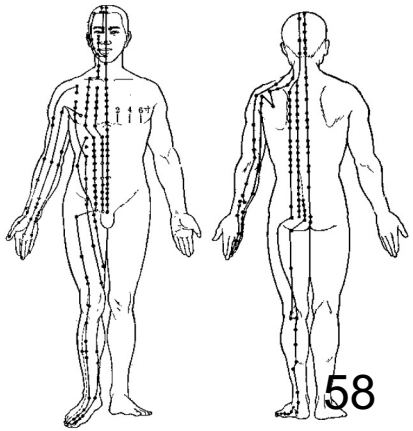
My Other Research Topics

- Biological morphogenesis
- Self-organizing swarm robots and UAVs
- Self-healing (scalable) systems
- Underwater robotic swarm
- Self-assembly in space
- Oil and energy robotics
- Self-configureble networks
- Distributed Power/Resource Sharing
- Evolution of brains
- Surprise-Based Learning (SBL) discovering hidden variables

Why Surprise-Based Learning?

- New Environment/Task → Learning → Knowledge/Skill
- Key Idea: Detect and learn from "surprises"
  - Adaptive to new environments (no priori knowledge, "swim")
  - Learn to accomplish new tasks (goals may change dynamically)
  - Self-heal unexpected failures or dynamics (e.g., inverted visions)
    - Know-how → Surprises → Learn → Recovery
- Human health as a complex system to be discovered



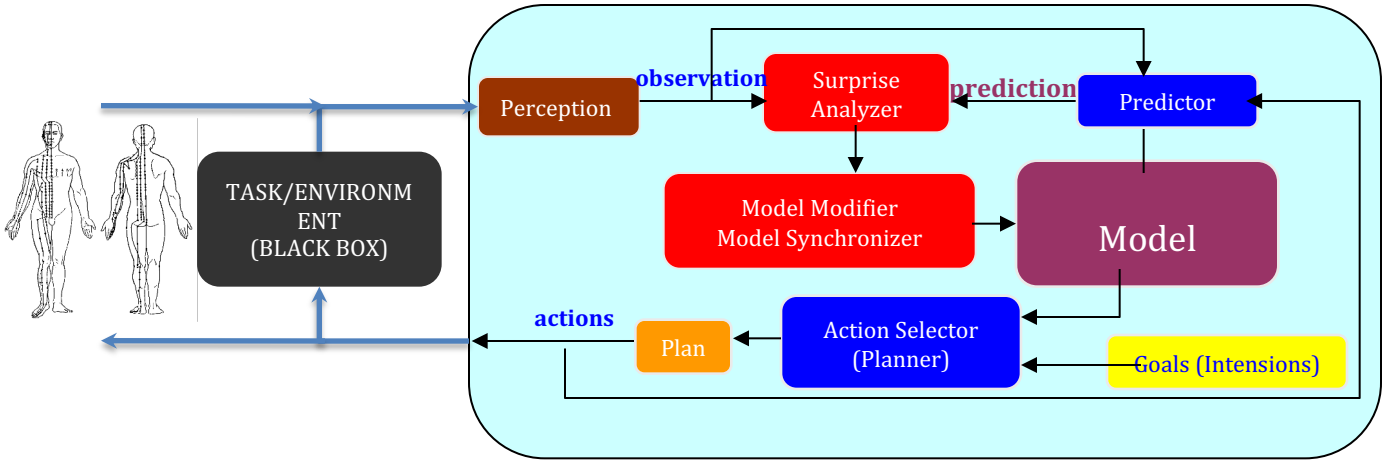
Machine Learning and Discovery

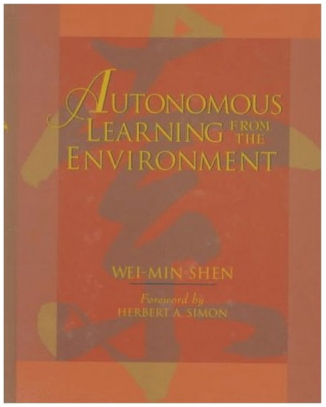
- Supervised Learning
- Unsupervised Learning
  - Type I: Clustering the data
    - Automatically group the data into clusters
  - Type II: Parameter Learning (states are known)
    - Learn transitions, sensor models, & current state
  - Type III: Structural & non-parametric Learning
    - States have internal structures (not just “symbols”)
    - States are not known and must be learned
- Machine Discovery

SBL Objectives

- Autonomously extracts the states and corresponding state machines from the interactive experience with the environment
  - Number of states, observed features of states & transitions between states are not necessarily known or predetermined
- Top-down approach to meet low-level sensor measurements and evidences
  - Learn the structure of states for description while detecting, explaining and learning from surprises in the experience
  - “Surprise” connects to “anomaly” and “unexpected interference”
- End-to-end Learning
- Life-long Learning
- Incremental and online learning

Surprise-Based Learning (Structure)





- The Learner continuously makes predictions, detects surprise, analyzes surprises, extracts critical information from surprises, and improves and use its action models to achieve goals

Surprise ==> Model ==> Prediction

Surprise Types and Key Problems

- Types of surprise
  - Unexpected failures
  - Unexpected successes
  - Null prediction surprise
    - When there is no a priori model
- Differentiate “new information” from noise
  - Focus attention on the relevant features
  - Seek differences that are statistically significant