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 2022-F Teaching Team

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

DSCI-553 2022-F Teaching Team

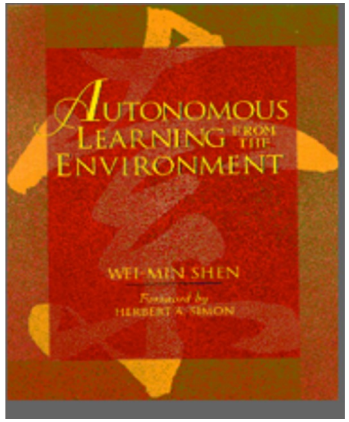
	Name	Office Hours	
	wmshen@usc.edu	When?	
		Where?	
	Wenxuan Li	When?	Wednesday 9 AM - 10 AM, Friday 9 AM - 10 AM
		Where?	https://usc.zoom.us/j/3212237726?pwd=L2pSMWtJaIF5V1hNdk9jL20yR1lHQT09
	Vishal Ajaybhai Kapadia	When?	Every Wednesday and Thursday: 11 AM PST
		Where?	https://usc.zoom.us/j/93441024543?pwd=T0Mzamxnd0dkVnRDa2dUdkl0anpnZz09
	Tirth Patel	When?	Every Wednesday and Thursday: 9 AM PST
		Where?	https://usc.zoom.us/j/99748457852? pwd=eEROR0ZlaE9ucTU3MmN2L2dVT1ZIZz09
	Weiwei Duan	When?	Monday 9 AM - 10 AM PST, Thursday 1 PM - 2PM PST
		Where?	https://usc.zoom.us/j/2332986718
	Jiahang Song	When?	Every Wednesday and Friday: 8 AM to 9 AM PST
		Where?	https://usc.zoom.us/j/96082301893
	Shaswat Anand	When?	Every Wednesday and Friday: 12 PM to 1 PM PST
		Where?	https://usc.zoom.us/j/96647039553? pwd=TzFmVDV1bXJCSnN3SDVJc3JuT0dmZz09
	Akash Ram Praveen Raj	When?	Monday: 10 AM to 11 AM PST, Friday: 2PM to 3PM PST
		Where?	https://usc.zoom.us/j/6431475007? pwd=UXA3Wk9rN0w1V1NOOCsyckR0ZTE2QT09
	Henil Shelat	When?	Every Monday and Friday: 11 AM to 12 PM PST
		Where?	
	Isha Manoj Chaudhari	When?	
		Where?	
	Sai Charan	When?	
		Where?	
	Harshita Bhorshetti	When?	Every Monday and Wednesday: 10AM - 11AM PST
		Where?	https://usc.zoom.us/j/98307180031?pwd=U1pPZzBzV1d4dXdRZzZleThjNDFZQT09
	Viraj Krishnakant Mehta	When?	Every Thursday and Friday: 9 AM to 10 AM
		Where?	https://usc.zoom.us/j/6513257794
	Jishnu Chander Ravichanderan	When?	
		Where?	

USC Polymorphic Robotics Lab

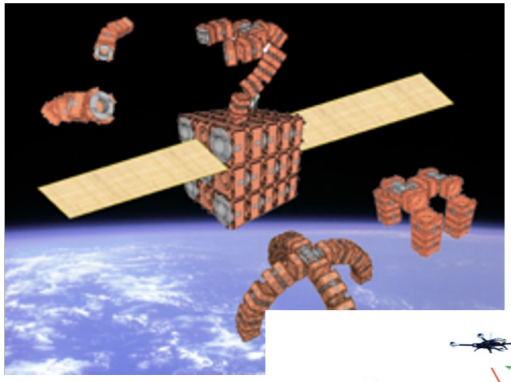
http://www.isi.edu/robots

Transformer Robots: Self-Reconfigurable & Modular

Surprise-Based Learning



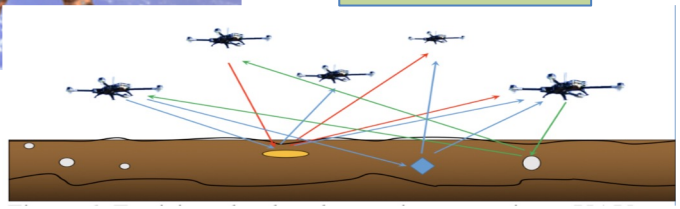
Welcome



Teaching
CS561 (AI)
CS360 (AI)
DSCI-552
DSCI-553

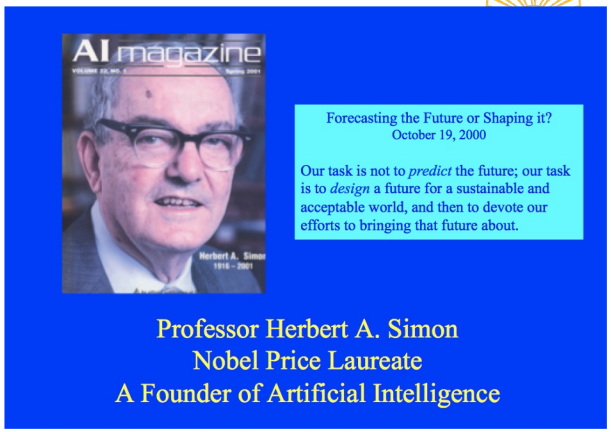
Projects

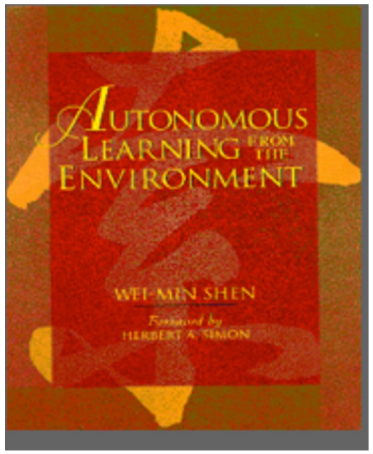
We conduct research in **adaptive, self-reconfigurable, autonomous robots and systems**, including **StarCell**, modular, multifunctional and self-reconfigurable **SuperBot**, Hormone-

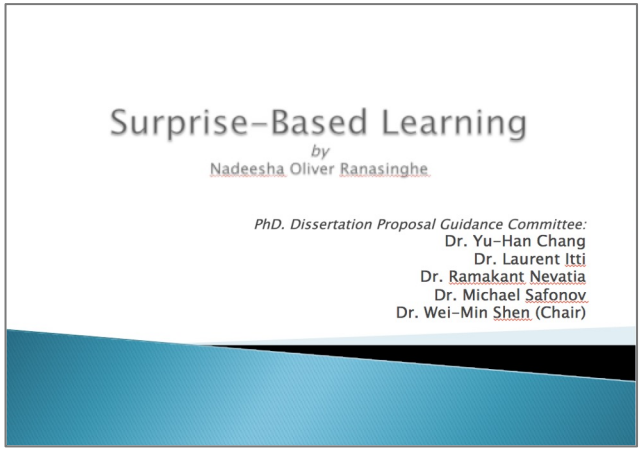


Surprise-Based Learning

- Herbert A. Simon (1916-2001)
  - Nobel Prize Winner (AI, Machine Discovery)
- Wei-Min Shen (1983 – present)
  - Autonomous Learning from the Environment
- Nadeesha Ranasinghe (2005 – present)
  - Learn and predict unexpected changes
- Thomas J. Collins (2013 – present)
  - Discovery of hidden/latent structures





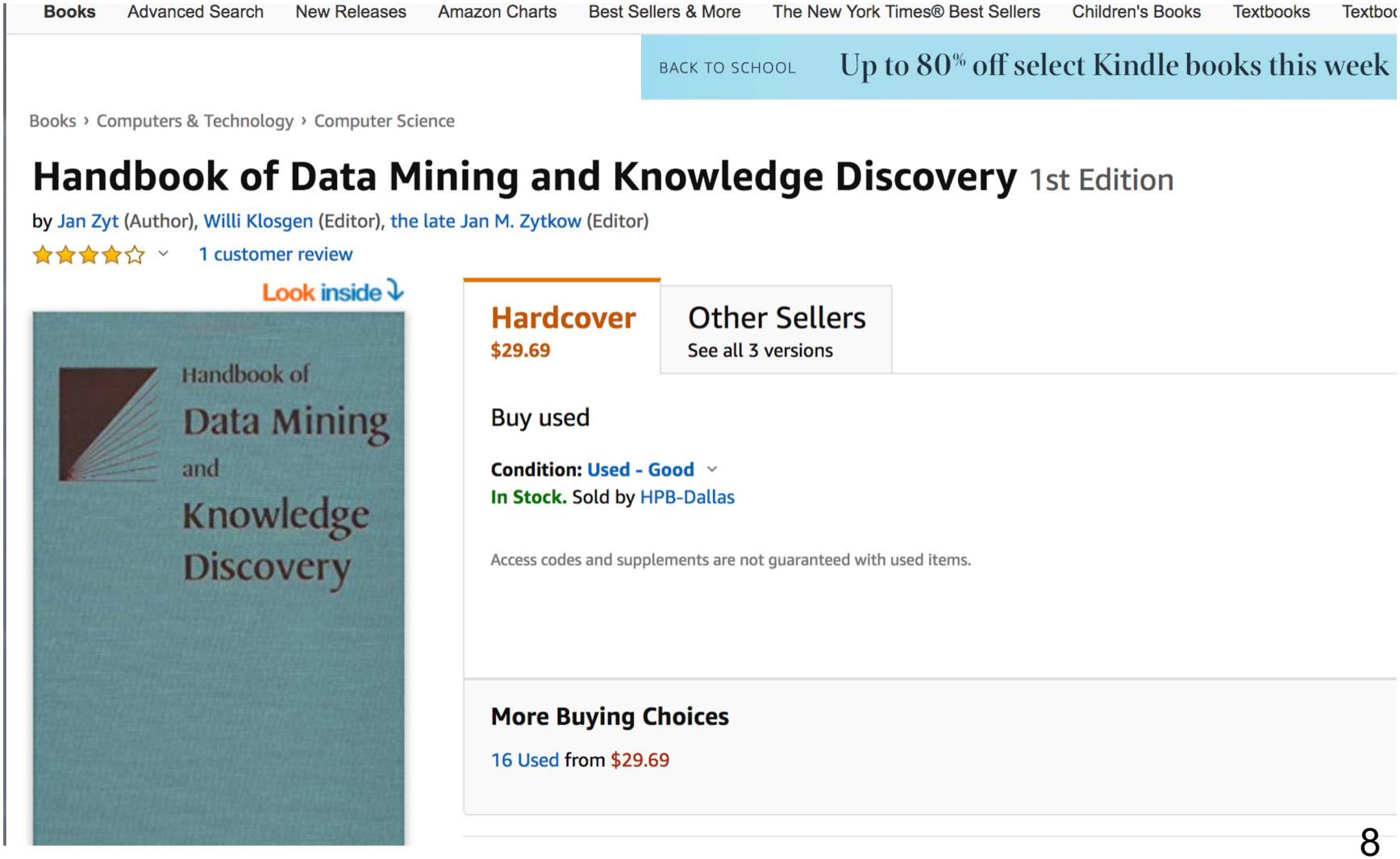


My Research on Data Mining

- *DataCrystal* for knowledge discovery from big data (NSF)
- Discovery hidden variables from complex systems (DARPA)
- Human Meridian Energy Sensing & Harmonization (USC+)
- Digital hormone for distributed device coordination (NASA)
- Biological Morphogenesis and digital hormones (Patents)
- Human health as a distributed, complex, intelligent system

- Editorial Board Member, Handbook of Data Mining and Knowledge Discovery, Oxford University Press, 2001
- NSF Panel Member on Data Mining
- One of the early starters of KDD at AAAI 1992

My Research on Data Mining



1997 RoboCup World Champion (CNN) (USC Dreamteam)

1997 RoboCup World Champion (CNN) (USC Dreamteam)

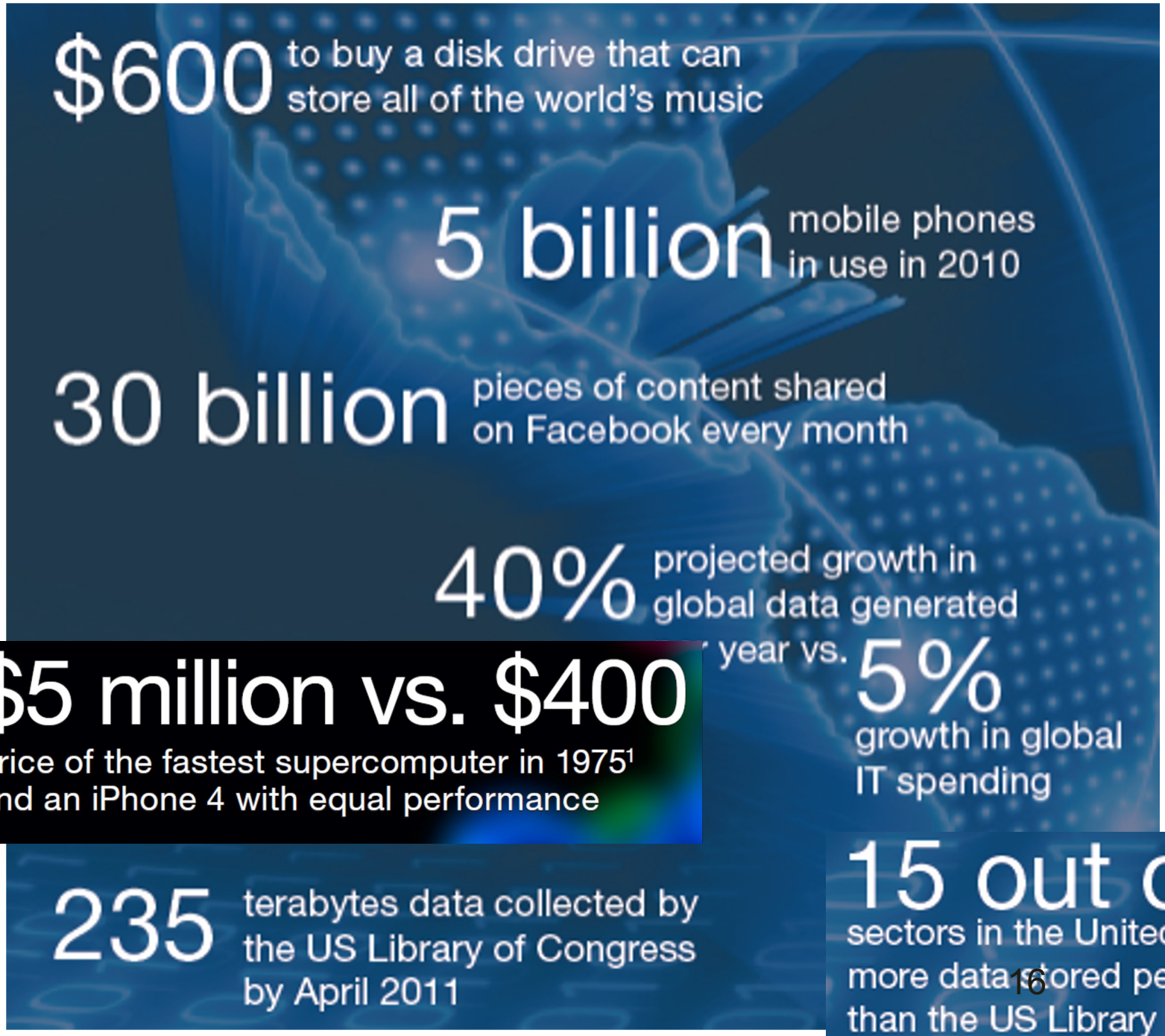
Introduction to SuperBot

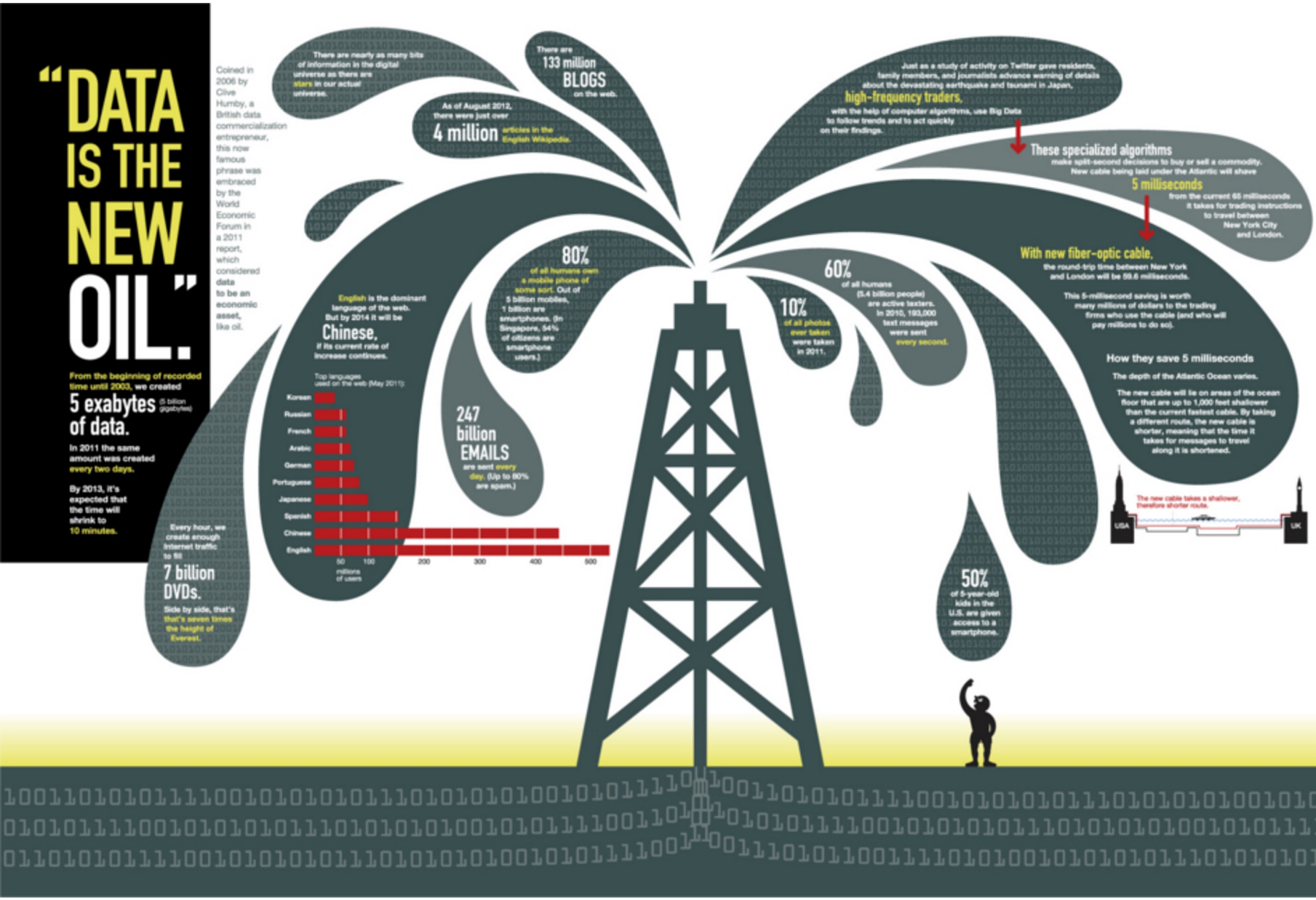

FOX 11 Ten O'Clock News
DVR
0:00
0:36
0:13
NEWS

Please introduce yourself !

What is Data Mining?
About THIS Course

What is Data Mining?
Knowledge Discovery from Data


$600 to buy a disk drive that can store all of the world's music
5 billion mobile phones in use in 2010
30 billion pieces of content shared on Facebook every month
40% projected growth in global data generated
year vs. 5% growth in global IT spending
$5 million vs. $400
Price of the fastest supercomputer in 1975[1] and an iPhone 4 with equal performance
235 terabytes data collected by the US Library of Congress by April 2011
15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress



Data contains value and knowledge

Data Mining

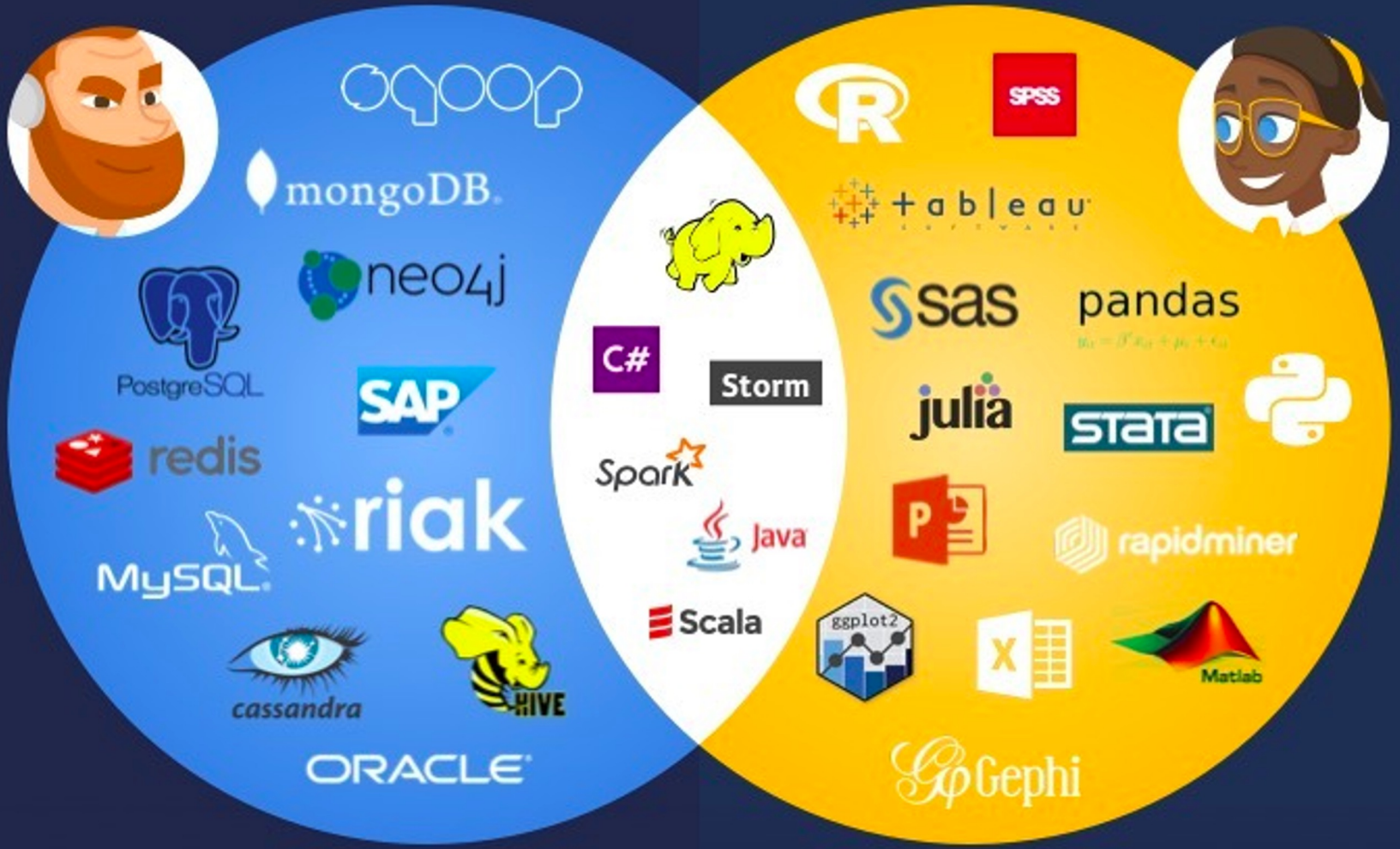
- But to extract the knowledge, data needs to be
  - Stored
  - Managed
  - And ANALYZED <= this class

Big Data Lifecycle

Data Mining ≈ Big Data ≈
Predictive Analytics ≈ Data Science


DATA
Engineer
DataCamp
Learn Data Science By Doing
DATA
Scientist
Develops, constructs, tests, and maintains architectures. Such as databases and large-scale processing systems.
Cleans, massages and organizes (big) data. Performs descriptive statistics and analysis to develop insights, build models and solve a business need.

Languages, Tools & Software



The Challenges of Big Data?

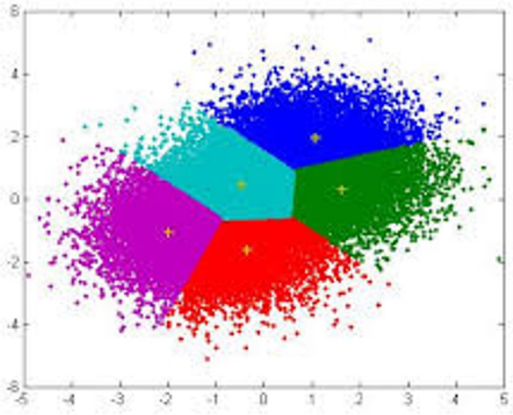
- Cannot store in one place
  - Must be distributed
- Failures for access may be unexpected
- Unpredictable diversity
- Messy, noisy, and errors are inevitable
- Dynamic
- ……

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

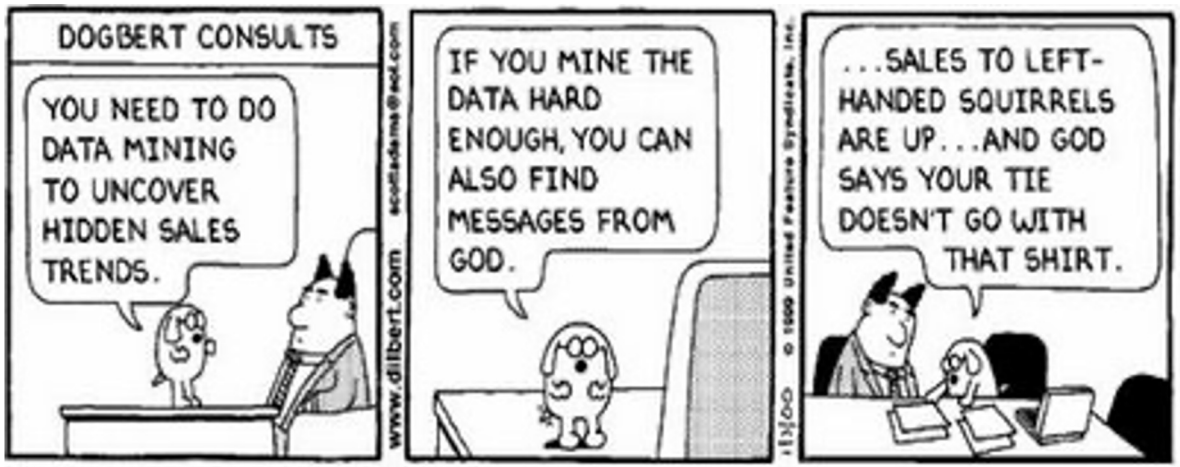
- **Descriptive methods**
  - Find human-interpretable patterns that describe the data
    - **Example:** Clustering



- **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 crap



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)

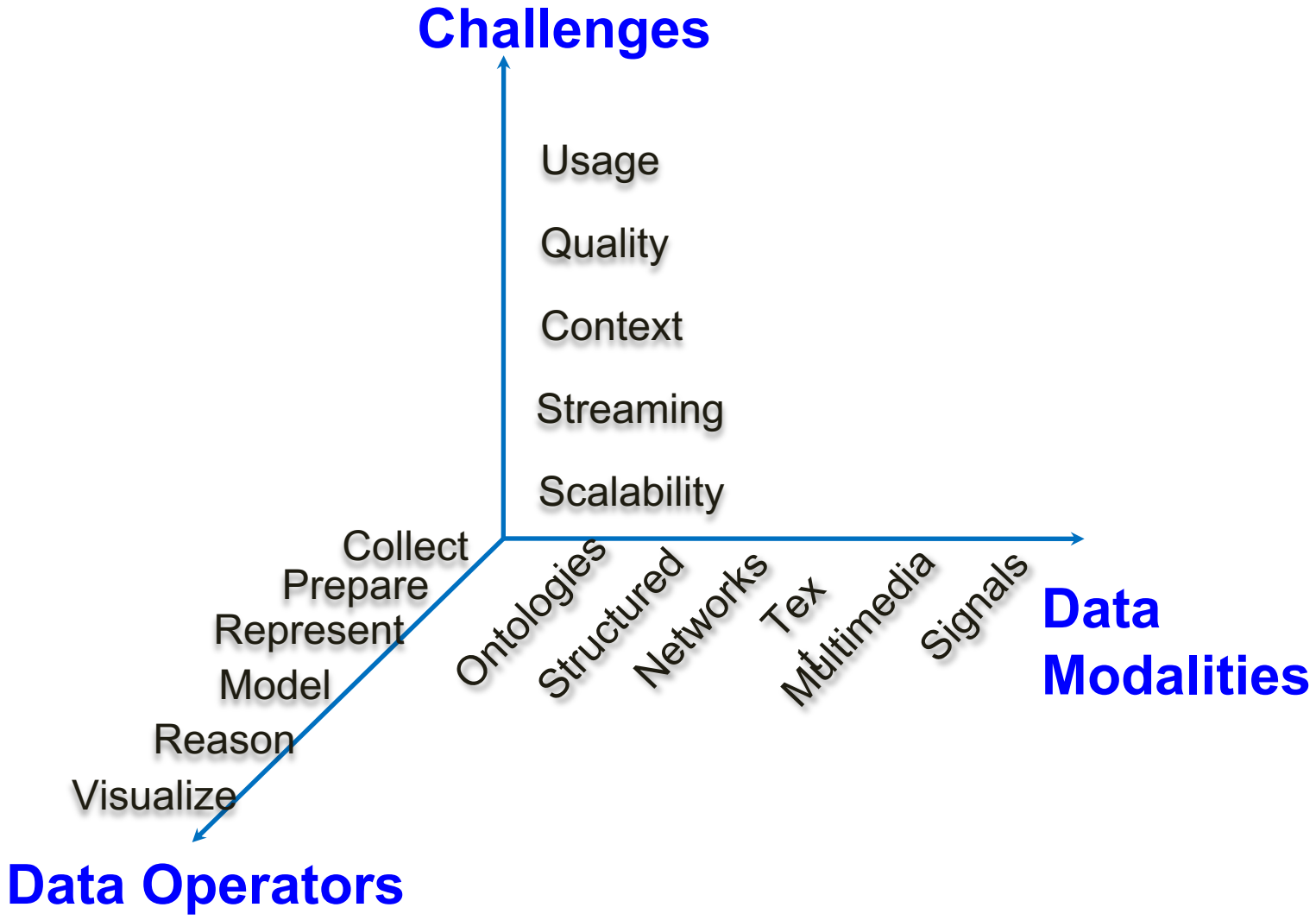
- 10^9 people, 1,000 days, 1% hotel stay, 10^5 hotels
  - The probability of any two people both deciding to visit a hotel on any given day is 0.0001 (i.e., 1%*1%)
  - The chance that they will visit the ***same*** hotel for one day is

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

    $$\binom{n}{k} = \frac{n!}{k!(n-k)!}; \quad \text{for large } n, \binom{n}{2} \text{ is about } n^2/2$$

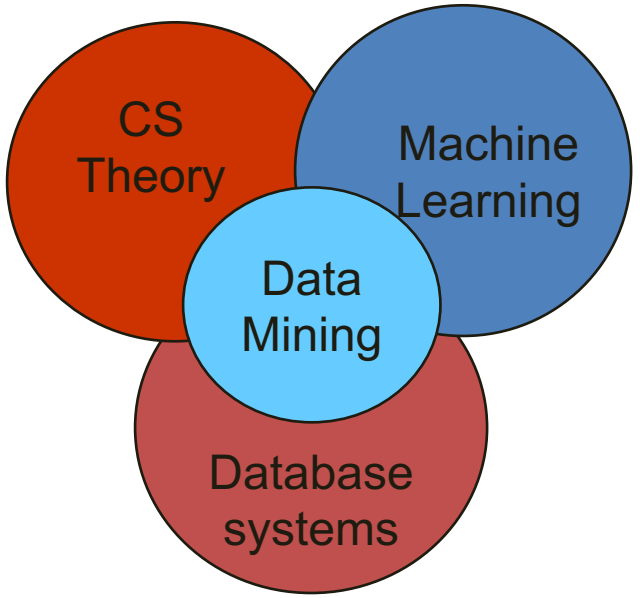
  - The number of pairs of people is $C(10^9, 2) = 5 \times 10^{17}$
  - The number of pairs of days is $C(10^3, 2) = 5 \times 10^5$
  - **Expected number of "suspicious" pairs of people:**
    - $5 \times 10^{17} \times 5 \times 10^5 \times 10^{-18} = 250{,}000$ (Wow!)
    - … too many combinations to 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

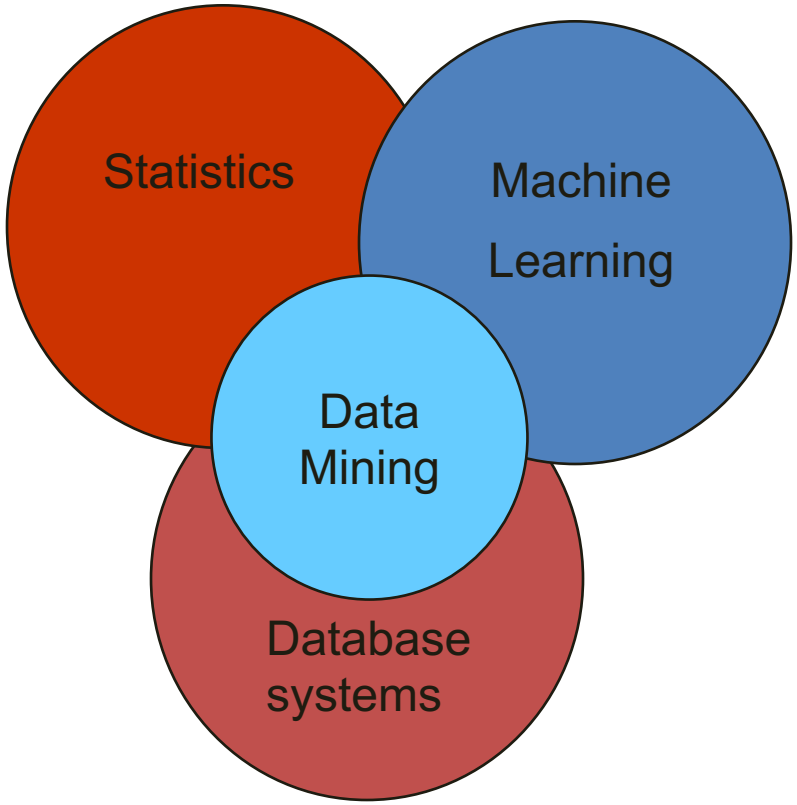
- **Data mining overlaps with:**
  - **Databases:** Large-scale data, simple(r) queries
  - **Machine learning:** Small data, Complex models
  - **CS Theory:** Algorithms
- **Different cultures:**
  - To a DB person, data mining is an extreme form of **analytic processing** – queries that examine **large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

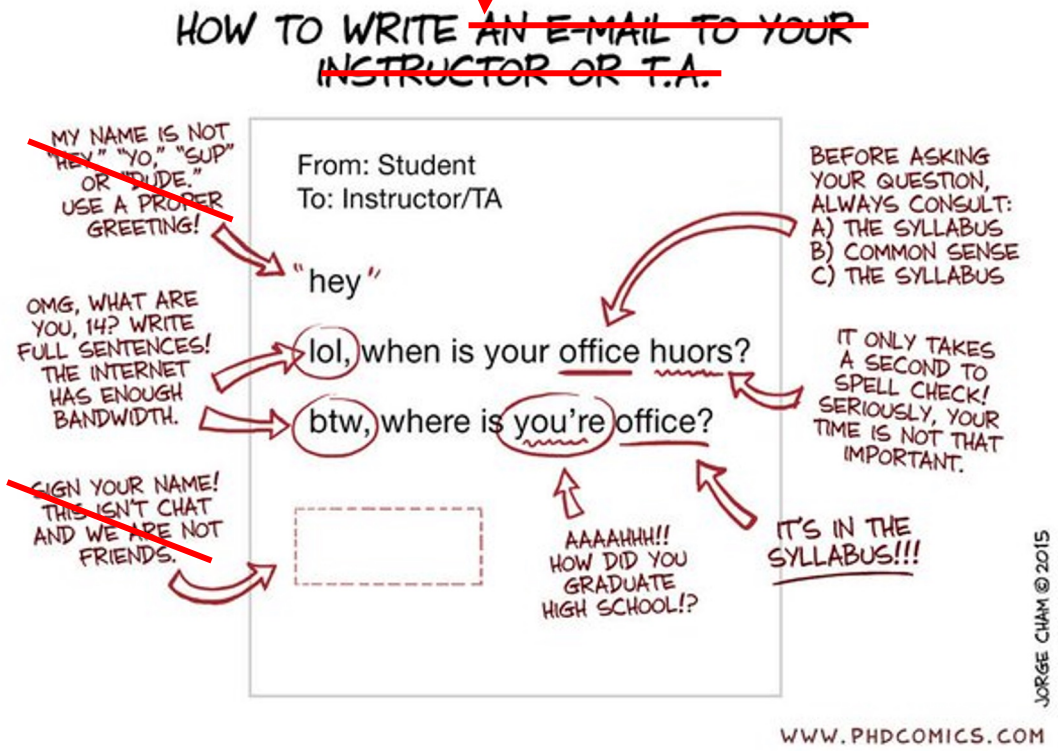
Course Logistics

- **Website:** courses.uscden.net/d2l/home/23448
  - 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

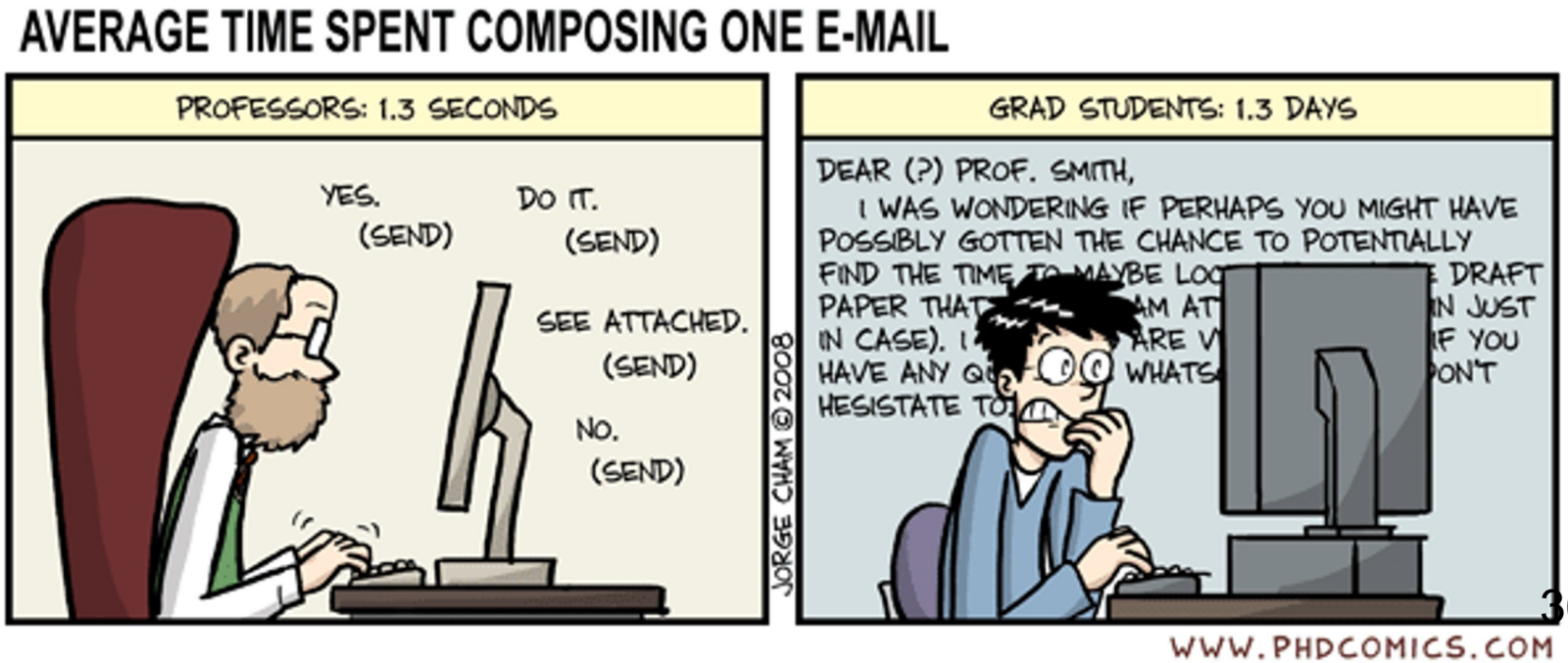
- **Discussion board on Piazza (Blackboard)**
  - Please enroll yourself, and you are all invited on Piazza
    - Questions and public communication with the course staff
    - Help your fellow students by contributing answers What a friend is for!

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!
  - No regrading
  - One week late penalty: –20%
  - 0 points after one week
  - Free five-day extensions
    - Use the extension days on homework however you want
    - No more extension days will be given for any reason

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:
    - Other students
    - TAs / Course Producers / Graders
- 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 Spring)

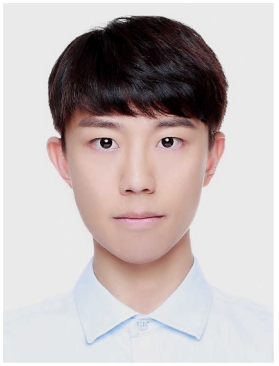
- USC Data Mining Competition (2022S)
  1. Feichi Yang *
  2. Shaswat Anand (your TA)
  3. Jheel Ketan Patel (your TA)
  4. Srinag Vinil Tummala
  5. Shayan Javid Yazdi
  6. Viraj Mehta (your TA)

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

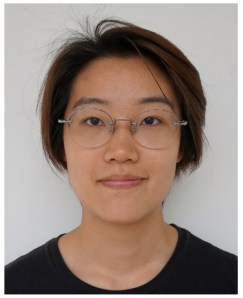
Previous Winners (2021 Spring)

- 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)

- 5:45 PM (quiz group zoom)
  - Join you group zoom, report to your proctor TA
  - See Piazza posts for your group assignment
- 6:00-6:20 PM (quiz group zoom)
  - Quiz, using Lockdown Browser and Password
- 6:25 – 9:20PM (lectures on DEN)
  - Join lecture on DEN (follow link on DEN)
- Notes
  - 2nd Week (practice quiz), real quizzes starts on 3rd

Prerequisites

- **Algorithms**
  - Dynamic programming, basic data structures, ….
- **Basic probability**
  - Moments, typical distributions, MLE, …
- **Programming**
  - Your choice, but Python will be very useful
  - 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+ 81 – 85 = B 78 – 80 = B-

75 – 80 =C+ 70 – 75 = C

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

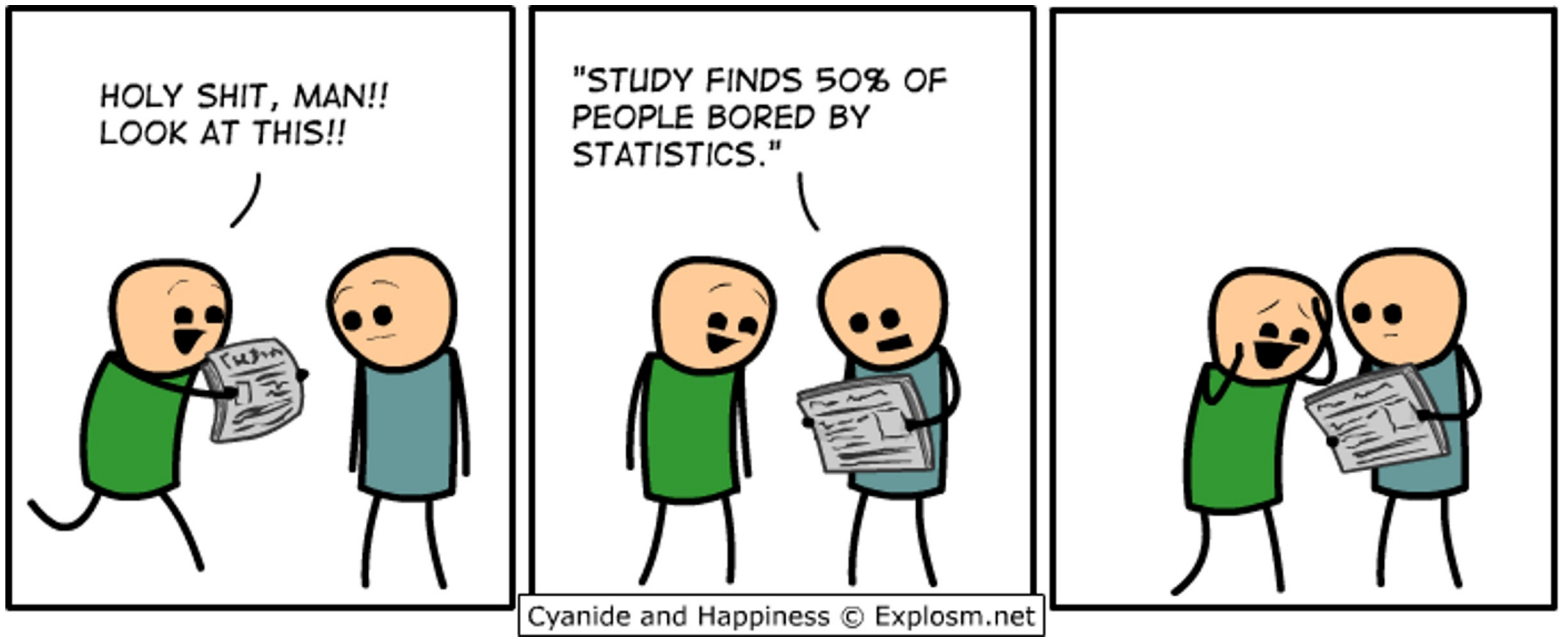
Below 60 is an F

What's After the Class?

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

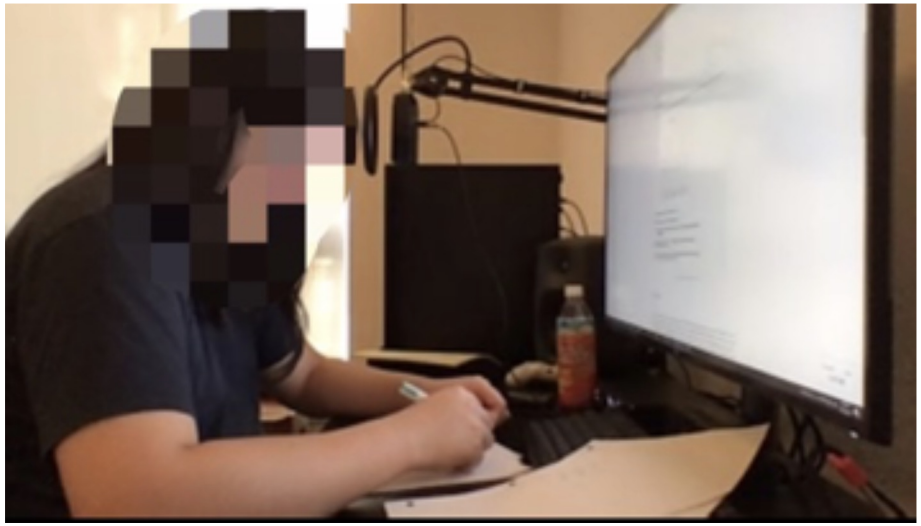
To-do items

- 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

- Alternate lecture days for a holiday
  - Please attend the lectures on the other lecture day
  - Study the slides and zoom videos yourself
- Quizzes will start on the 3rd week (2nd week practice)
  - Zoom 15min before, quiz on time for 20 min, lectures
  - 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

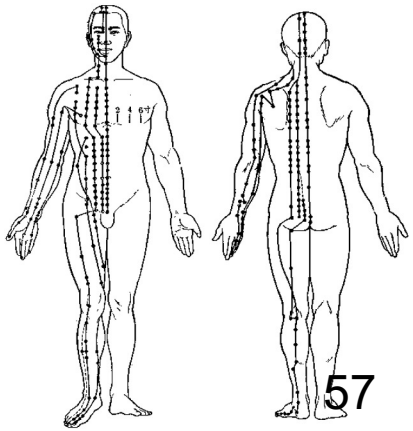
My Other Research Topics

- 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



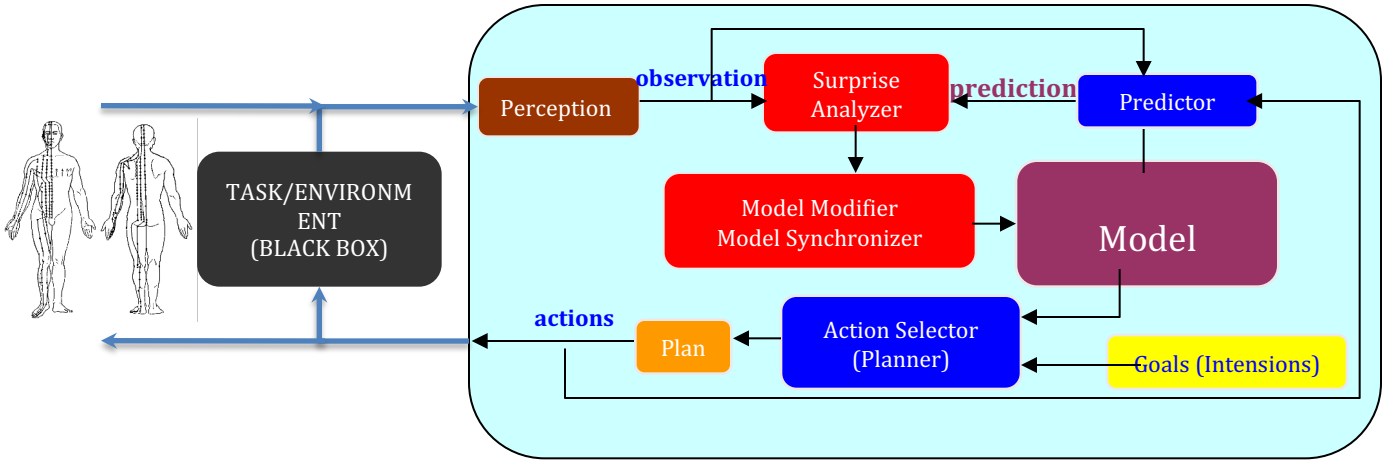
Machine Learning and Discovery

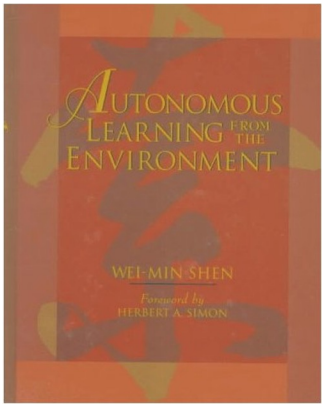
- Supervised Learning
- Unsupervised Learning
  - Type I: Clustering the data
    - Automatically group the data into clusters
  - Type II: Parameter Learning (states are known)
    - Learn transitions, sensor models, & current state
  - Type III: Structural & non-parametric Learning
    - States have 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