



INF 553: Foundations and Applications of Data Mining

Professor Wei-Min Shen

University of Southern California



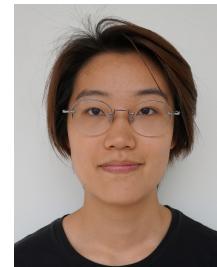
Outline

- Personal Introduction
- About this course
- Introduction to Data Mining
- Map-Reduce (Part I)



INF553 Course Staff

- **Instructor: Professor Wei-Min Shen**
- **TAs/CPs/Graders**
 - Chang Liu, liu599@usc.edu
 - Anshul Gupta, anshulg@usc.edu
 - Yang Zheng, zhen528@usc.edu
 - Jing Ouyang, jingo@usc.edu
 - Xiyuan Zhou, xzhou911@usc.edu
- **Office Hours**
 - Instructor
 - Monday/Tuesday after classes (by appointment)
 - TAs
 - See Blackboard for time and locations



USC Polymorphic Robotics Lab



<http://www.isi.edu/robots>

Transformer Robots: Self-Reconfigurable & Modular

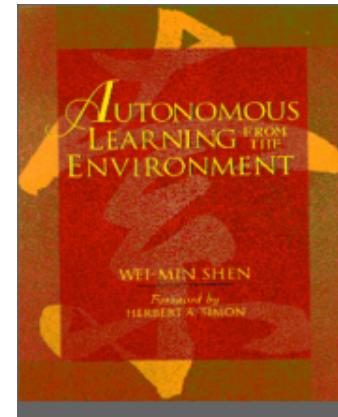


Welcome

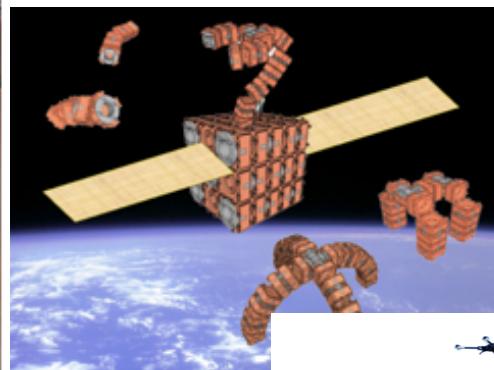


We conduct research in **adaptive, self-reconfigurable, autonomous robots and systems**, including **StarCell**, modular, multifunctional and self-reconfigurable **SuperBot**, Hormone-inspired Control (**HIC**), **Swarm-Based Learning (**SBL**)**, and

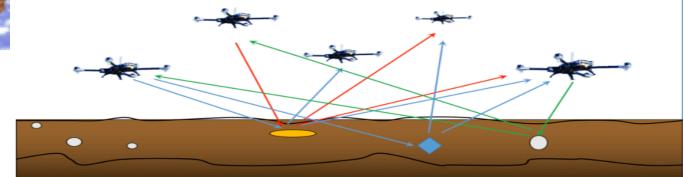
Surprise-Based Learning



Projects

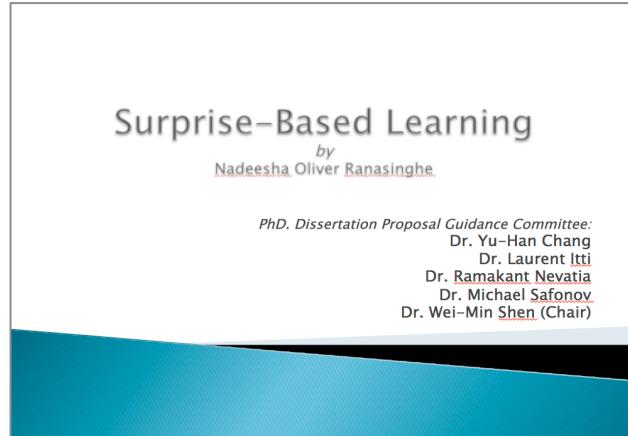
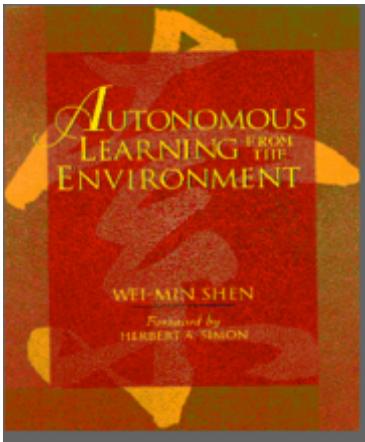


Teaching
CS561 (AI)
CS360 (AI)
INF552
INF553



Surprise-Based Learning

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



AI magazine
VOLUME 22, NO. 1, January 2001

Herbert A. Simon
1916 - 2001

Forecasting the Future or Shaping it?
October 19, 2000

Our task is not to *predict* the future; our task is to *design* a future for a sustainable and acceptable world, and then to devote our efforts to bringing that future about.

Professor Herbert A. Simon
Nobel Price Laureate
A Founder of Artificial Intelligence

PolyMorphic
Robotics
Laboratory

Welcome

Self-Reconfigurable Modular Robots
(System of Systems)

Surprise-Based Learning

Projects

We conduct research in adaptive, self-reconfigurable, autonomous robots and systems, including SuperCell, modular, multifunctional and self-reconfigurable SuperBot, Harmonic...

Active State Learning from Surprises in Stochastic and Partially-observable Environments

Thomas Joseph Collins

Defense Committee:
Prof. Wei-Min Shen (Chair)
Prof. Paul Rosenbloom
Prof. John Carlsson (Outside member)

USC Viterbi
School of Engineering
Information Sciences



My Research on Data Mining

Books Advanced Search New Releases Amazon Charts Best Sellers & More The New York Times® Best Sellers Children's Books Textbooks Textbook

BACK TO SCHOOL Up to 80% off select Kindle books this week

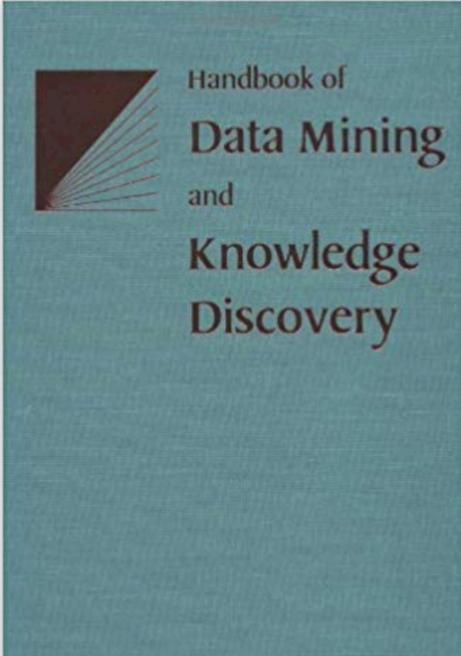
Books > Computers & Technology > Computer Science

Handbook of Data Mining and Knowledge Discovery 1st Edition

by [Jan Zyt](#) (Author), [Willi Klosgen](#) (Editor), [the late Jan M. Zytkow](#) (Editor)

1 customer review

[Look inside](#) ↴



Hardcover
\$29.69

Other Sellers
[See all 3 versions](#)

[Buy used](#)

Condition: [Used - Good](#) ▾
[In Stock](#). Sold by [HPB-Dallas](#)

Access codes and supplements are not guaranteed with used items.

More Buying Choices

[16 Used from \\$29.69](#)



First RoboCup World Champion (1997, USC Dreamteam)





Introduction to SuperBot





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

\$5 million vs. \$400

Price of the fastest supercomputer in 1975¹ and an iPhone 4 with equal performance

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

year vs. **5%** growth in global IT spending

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

"DATA IS THE NEW OIL."

Coined in 2006 by Clive Humby, a British data commercialization entrepreneur, this now famous phrase was embraced by the World Economic Forum in a 2011 report, which considered data to be an economic asset, like oil.

From the beginning of recorded time until 2003, we created
5 exabytes (5 billion terabytes) of data.

In 2011 the same amount was created every two days.

By 2013, it's expected that the time will shrink to 10 minutes.

Every hour, we create enough Internet traffic to fill
7 billion DVDs.

Side by side, that's that's seven times the height of Everest.

There are nearly as many bits of information in the digital universe as there are stars in our actual universe.

As of August 2012, there were just over
4 million articles in the English Wikipedia.

There are **133 million BLOGS** on the web.

English is the dominant language of the web. But by 2014 it will be
Chinese.

If its current rate of increase continues,

Top languages used on the web (May 2012):



247 billion EMAILS are sent every day. (Up to 80% are spam.)

80% of all humans own a mobile phone of some sort. Out of 5 billion mobiles, 1 billion are smartphones. In Singapore, 94% of citizens are smartphone users.

Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan, **high-frequency traders**,

with the help of computer algorithms, use Big Data to follow trends and to act quickly on their findings.

These specialized algorithms make split-second decisions to buy or sell a commodity.

New cable being laid under the Atlantic will shave **5 milliseconds** from the current 65 milliseconds it takes for trading instructions to travel between New York City and London.

With new fiber-optic cable, the round-trip time between New York and London will be 59.6 milliseconds.

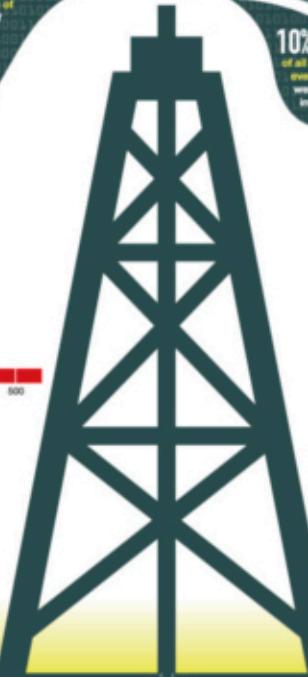
This 5-millisecond saving is worth many millions of dollars to the trading firms who use the cable (and who will pay millions to do so).

How they save 5 milliseconds

The depth of the Atlantic Ocean varies. The new cable will lie on areas of the ocean floor that are up to 1,000 feet shallower than the current fastest cable. By taking a different route, the new cable is shorter, meaning that the time it takes for messages to travel along it is shortened.



50% of 8-year-old kids in the U.S. are given access to a smartphone.



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

Develops, constructs, tests, and maintains architectures. Such as databases and large-scale processing systems.



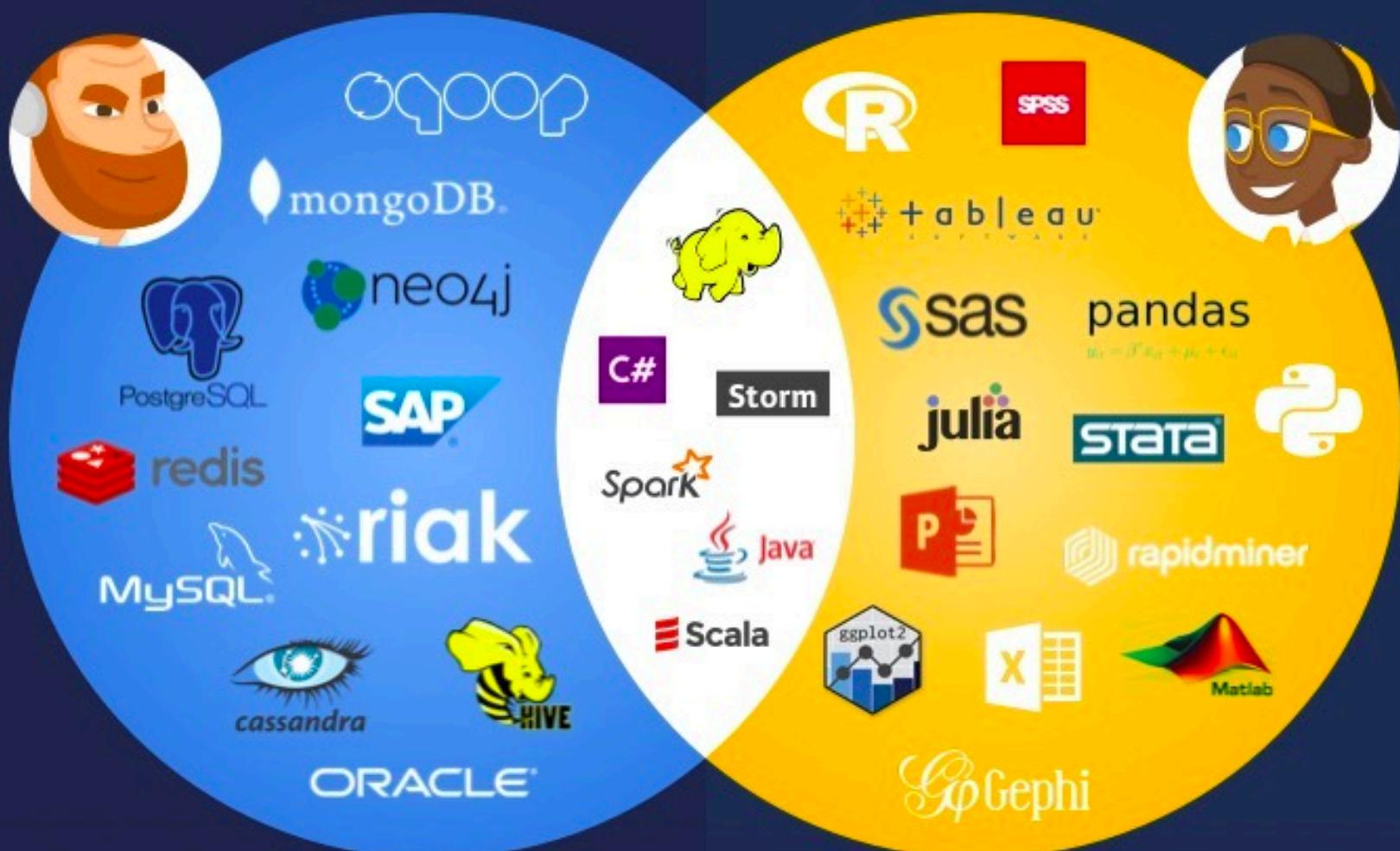
DataCamp
Learn Data Science By Doing

DATA Scientist

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
-



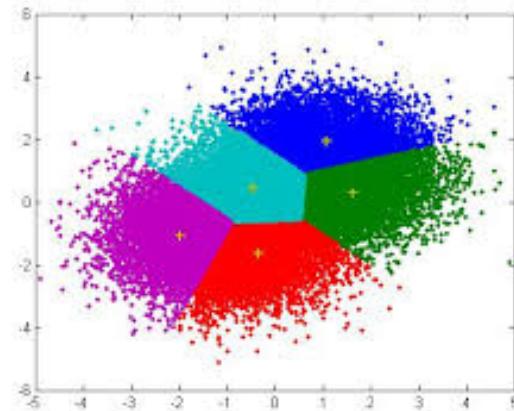
What is Data Mining?

- 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

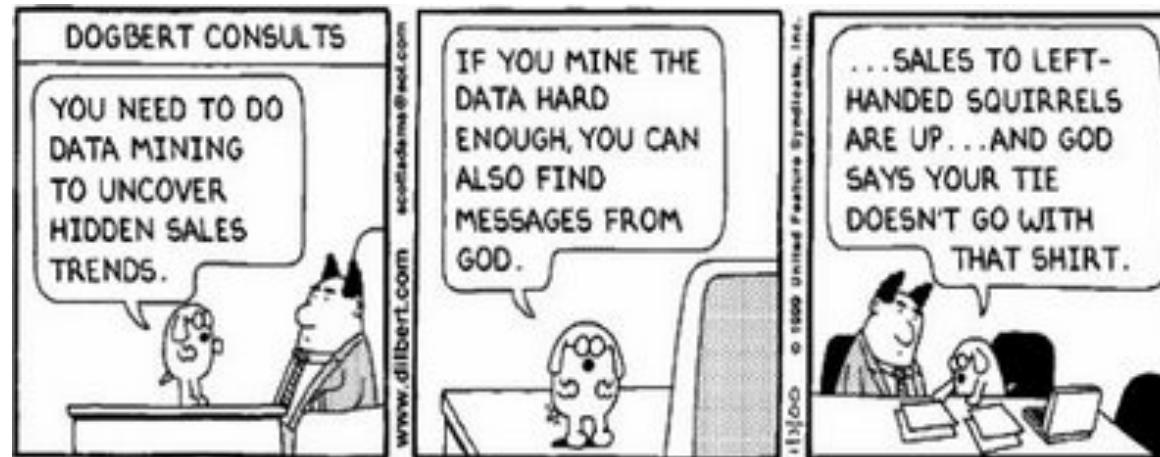
- **Descriptive methods**
 - Find human-interpretable patterns that describe the data
 - **Example:** Clustering
- **Predictive methods**
 - Use some variables to predict unknown or future values of other variables
 - **Example:** Recommender systems





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 \sim 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?**

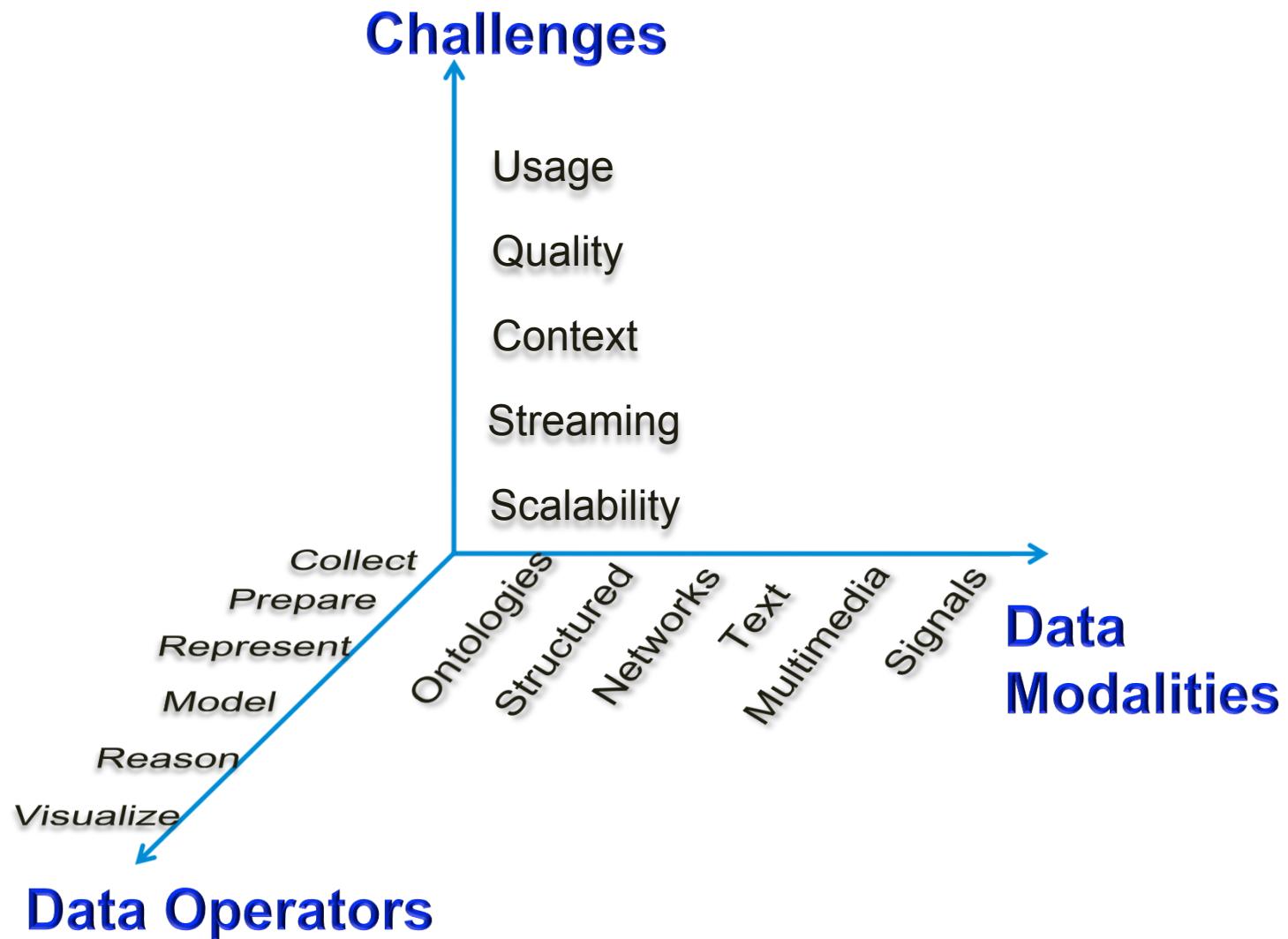


Meaningfulness of Analytic Answers (Cont'd)

- 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}
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \text{ for large } n, \binom{n}{2} \text{ is about } n^2/2$$
 - 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:**
 - $5 \times 10^{17} \times 5 \times 10^5 \times 10^{-18} = 250,000!$
 - ... too many combinations to check – we need to have some additional evidence to find “suspicious” pairs of people in some more efficient way



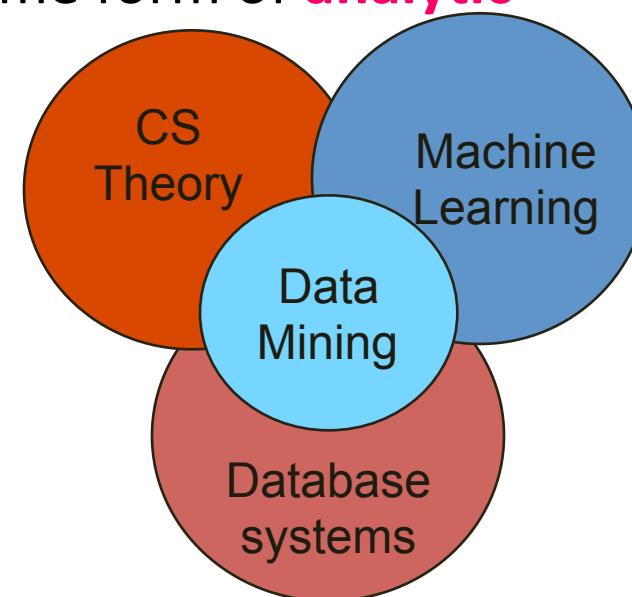
What matters when dealing with data?





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!**



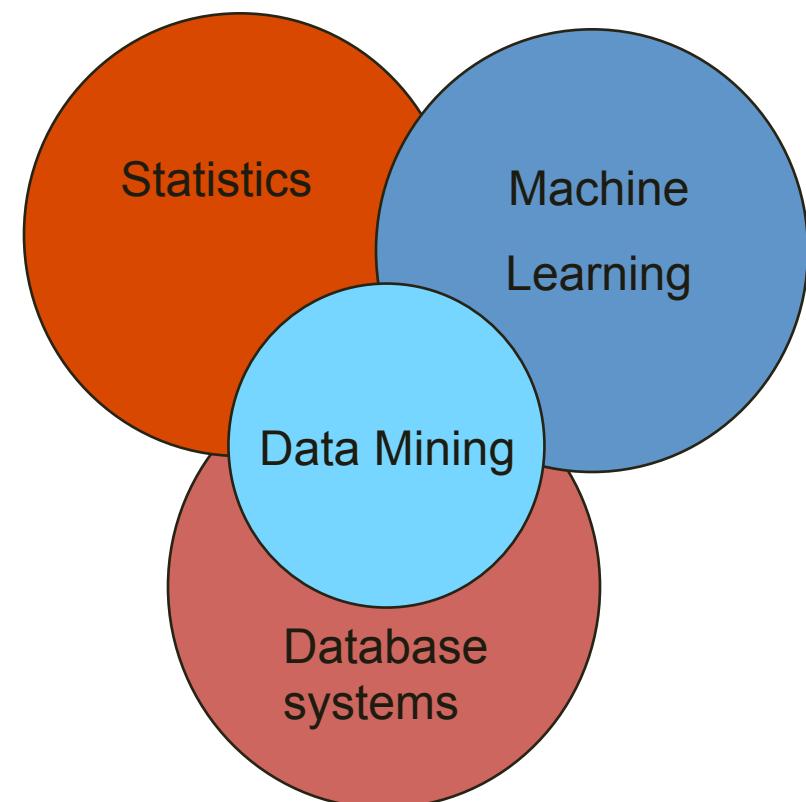


About THIS Course



This Course

- This course overlaps with machine learning, statistics, artificial intelligence, databases but more stress on
 - **Scalability** (big data)
 - **Algorithms**
 - **Computing architectures**
 - 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 (Rec. Sys., Communities)
 - Optimization (stochastic gradient descent)
 - Dynamic programming (frequent itemsets)
 - Hashing (LSH, Bloom filters)



How It All Fits Together

High dim. data

Locality
sensitive
hashing

Clustering

Dimensional
ity
reduction

Graph data

PageRank,
SimRank

Community
Detection

Spam
Detection

Infinite data

Filtering
data
streams

Web
advertising

Queries on
streams

Machine learning

SVM

Decision
Trees

Perceptron,
kNN

Apps

Recommen
der systems

Association
Rules

Duplicate
document
detection



Course Logistics

- **Course website:**

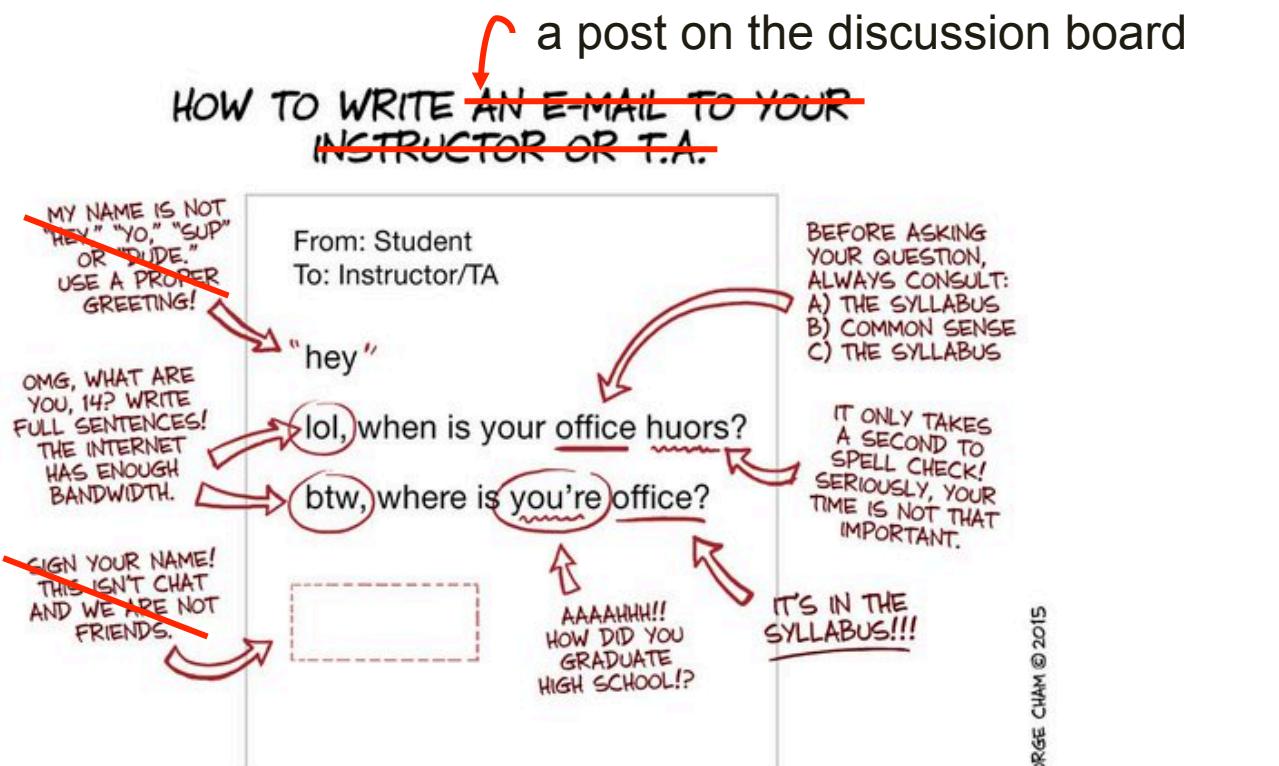
<https://blackboard.usc.edu/>

- Lecture slides (same day after the lecture)
- Homeworks, Quizzes, Exam, Competition
- Readings
 - **Mining of Massive Datasets**
 - J. Leskovec, A. Rajaraman and J. Ullman
 - Free online: <http://www.mmds.org> or on Blackboard
 - 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 your answers (with bonus)

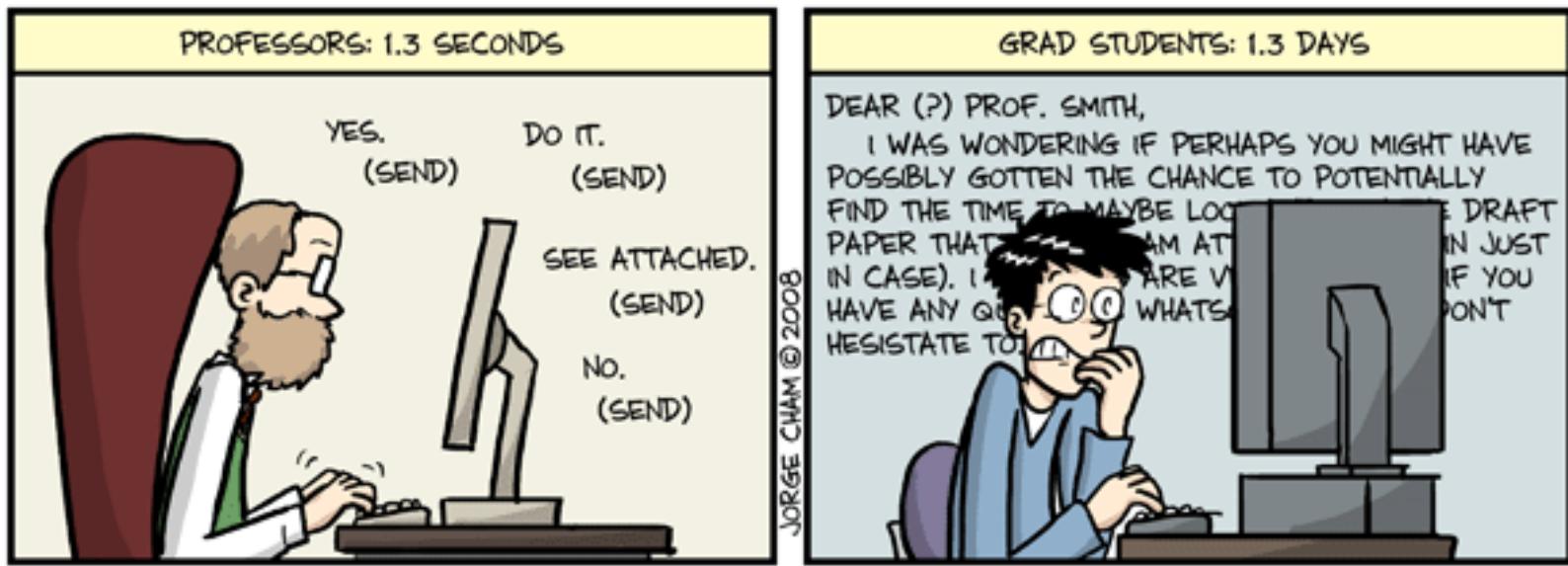




Logistics: Communication

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

AVERAGE TIME SPENT COMPOSING ONE E-MAIL





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 Project & Grand Competition: 8% + bonus**
 - **Recommendation Systems**
- **Please work on your own code!**



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
 - TA / grader
- 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 (Spring 2020)
 - I will have something to say if you want me to be your reference ☺



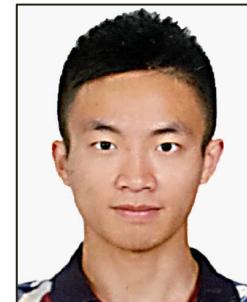


Previous Winners (2017)

- USC Data Mining Competition



- First Place, **Priyambada Jain**



- Second Place, **Zihao Zhai**



- Third Place, **Vijayakumar Gedigeri**



Previous Winners (2018)

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



- Second Place, **Chen Lou**



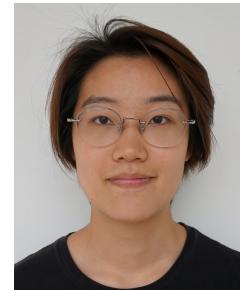
- Third Place, **Bufan Zeng**





Previous Winners (2019 Fall)

- USC Data Mining Competition



- First Place, **Yang Zheng**



- Second Place, Peilun Yan



- Third Place, Nivedetha Kumaram



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
 - We will drop your one lowest quiz
- **Comprehensive exam: 20%**
 - **Tentative Date: 4/27 and 4/28 3:30 – 5:30 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.**



Prerequisites

- **Algorithms**
 - Dynamic programming, basic data structures
- **Basic probability**
 - Moments, typical distributions, MLE, ...
- **Programming**
 - Your choice, but Python will be very useful
 - Some of the homework will require you to use Scala only
- **We provide some background, but the class will be fast paced**



Course Grade

Grades will range from A through F. The following is the breakdown for grading:

94 – 100 = A 73 – 77 = C

90 – 94 = A- 70 – 73 = C-

87 – 90 = B+ 67 – 70 = D+

84 – 87 = B 64 – 67 = D

80 – 84 = B- 60 – 64 = D-

77 – 80 = C+ Below 60 is an F



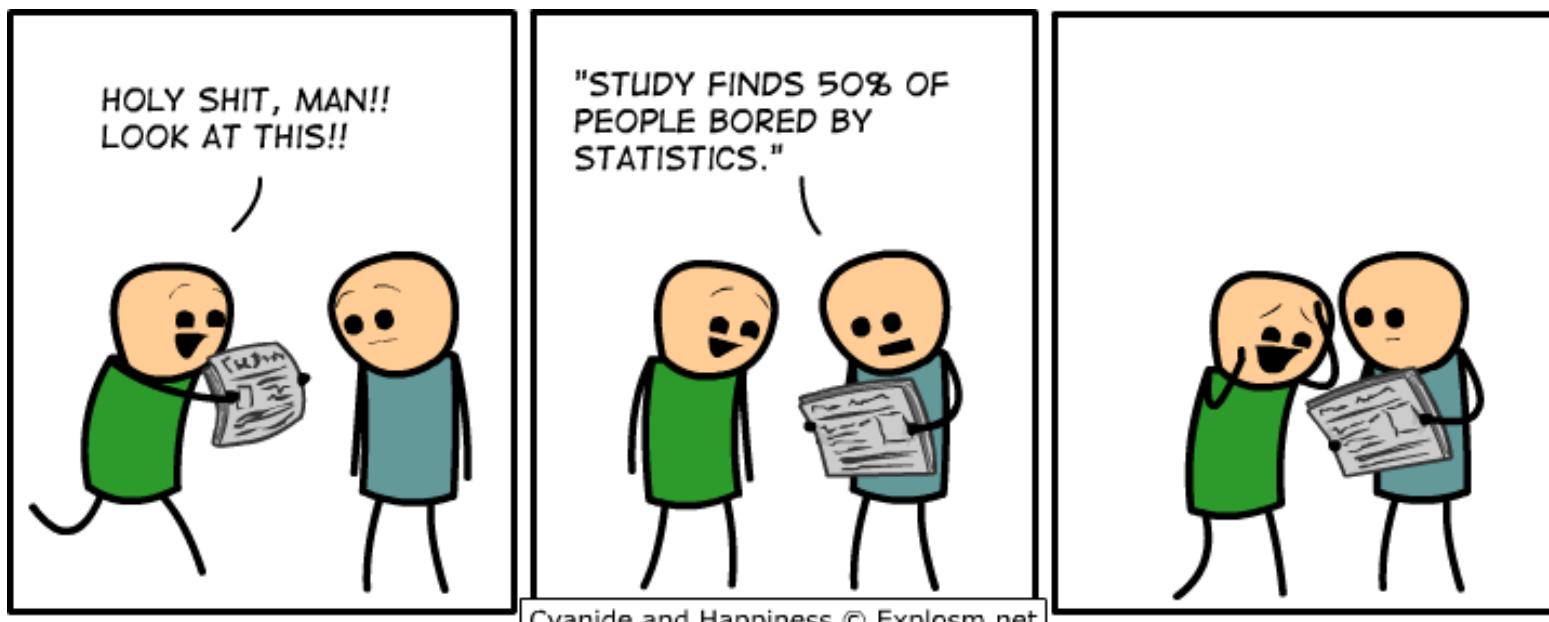
What's after the class

- Directed Research
- Course producer or grader positions
- Paid RA positions



To-do items

- Download the textbook
- Install Spark on your machine (<http://spark.apache.org/>)
- Get familiar with Vocareum.com and how to use the terminal there
- Play with datasets <http://grouplens.org/datasets/movielens/> and <http://jmcauley.ucsd.edu/data/amazon/links.html>
- Email me a photo of you if your GRS photo is outdated...





Schedule around Holidays

- Next Monday 1/20 is a holiday
 - Please come to my Tuesday 1/21 lectures
 - or study the slides yourself
- Quizzes will start on 1/27 onwards
- Monday 2/17 is also a holiday
 - Please come on Tuesday for quizzes/lectures



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”





My Research on Data Mining

- Human health as a distributed, complex, intelligent system
- 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)

- Served on Editor Board, Handbook of Data Mining and Knowledge Discovery, Oxford University Press, 2001
- NSF Panels on Data Mining



My Research on Data Mining

Books Advanced Search New Releases Amazon Charts Best Sellers & More The New York Times® Best Sellers Children's Books Textbooks Textbook

BACK TO SCHOOL Up to 80% off select Kindle books this week

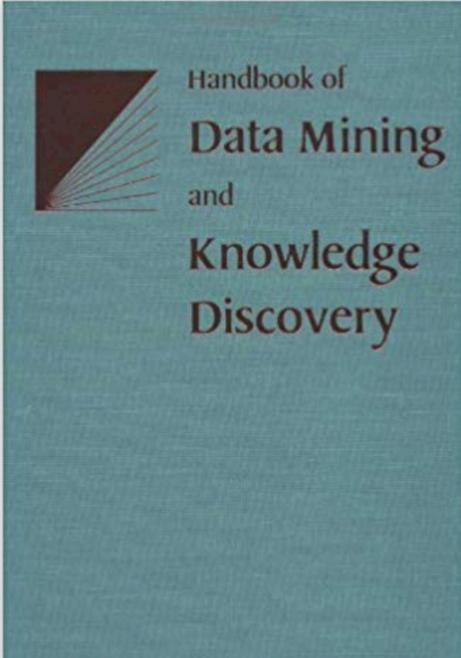
Books > Computers & Technology > Computer Science

Handbook of Data Mining and Knowledge Discovery 1st Edition

by [Jan Zyt](#) (Author), [Willi Klosgen](#) (Editor), [the late Jan M. Zytkow](#) (Editor)

1 customer review

[Look inside](#) ↴



Hardcover
\$29.69

Other Sellers
[See all 3 versions](#)

[Buy used](#)

Condition: [Used - Good](#) ▾
[In Stock](#). Sold by [HPB-Dallas](#)

Access codes and supplements are not guaranteed with used items.

More Buying Choices

16 Used from \$29.69



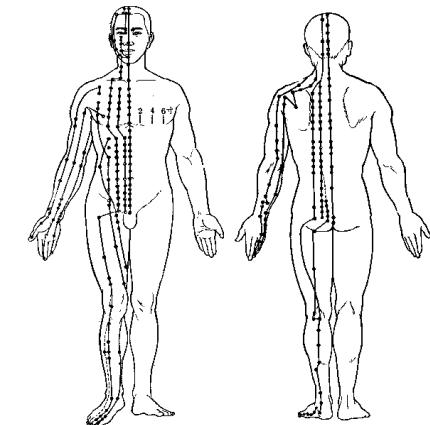
My Other Research Topics

- 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

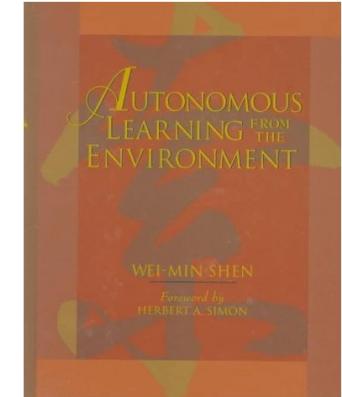
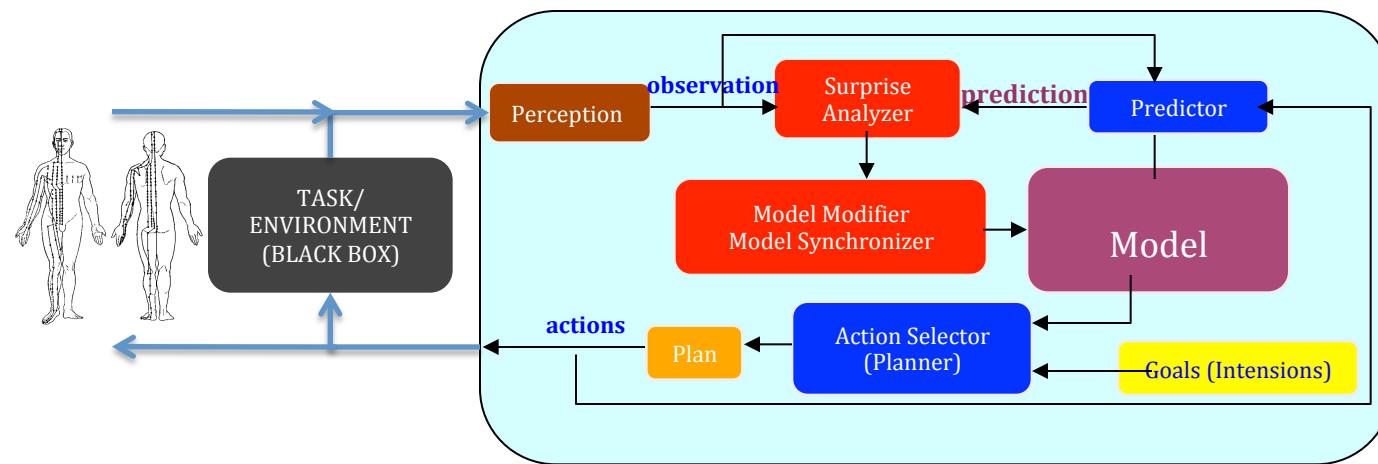


SBL Objectives

- 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 uses 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