i296A:Thought Leaders in Data Science and Analytics

Ram Akella
University of California, Berkeley
iSchool

akella@ischool.berkeley.edu
650-279-3078 (cell)

Lecture 1
January 18, 2012

Co-Instructors

- □ Prof Ray Larson – iSchool – IR

- □ Industry Expert
  Dr. Jimi Shanahan –Adobe+++
  james.shanahan@gmail.com
  415-630-0890

Session Outline

- Introduction – Data Analytics at iSchool
- Introduction to Data Analytics and Trends
- Introduction to seminar philosophy and organization
- Introduction to Online Advertising and Computational Marketing

Data Mining and IR

- At the iSchool and Beyond

- i290: Basic Data Mining course
  - Hands on project
  - Top 10 algorithms
  - Use and software
- i296A-2: Thought Leaders in Data Science and Analytics
  - Landscape and business persective
    - Primarily, leading industry Executives, Researchers, Entrepreneurs, Venture capitalists

    - Some academic leaders

Data Mining and IR

- At the iSchool and Beyond (continued)

- i296A-3: Advanced Analytics projects
  - Presumes prior exposure to one or more of
    - Data mining, machine learning, optimization, deeper probability/statistics including possibly inference, Bayesian statistics etc.
  - Objectives of students either
    - Important new theory from practical real world need or
    - Solve real world problems with deep theory, as required

Data Mining and IR

- At the iSchool and Beyond (continued)

- I290: Social Computing
- IR: Prof. Larson will expand
  - E.g. i202, i240 etc.

Seminar Outline

□ Knowledge Services, Data Mining, and Business Analytics

- Internet marketing and online ads
- Financial services
- Health services
- Service center analytics
  - Future of all enterprises
- Social networks and recommenders
- Data Science and Big Data

Who?

- Who Should Take This Course?
  - Graduate Students
  - Engineers and Managers who wish to
    - Gain perspective
    - Move into this area
  - (Potential) Entrepreneurs who wish to
    - Brainstorm new ideas
    - Create process for startup

What?

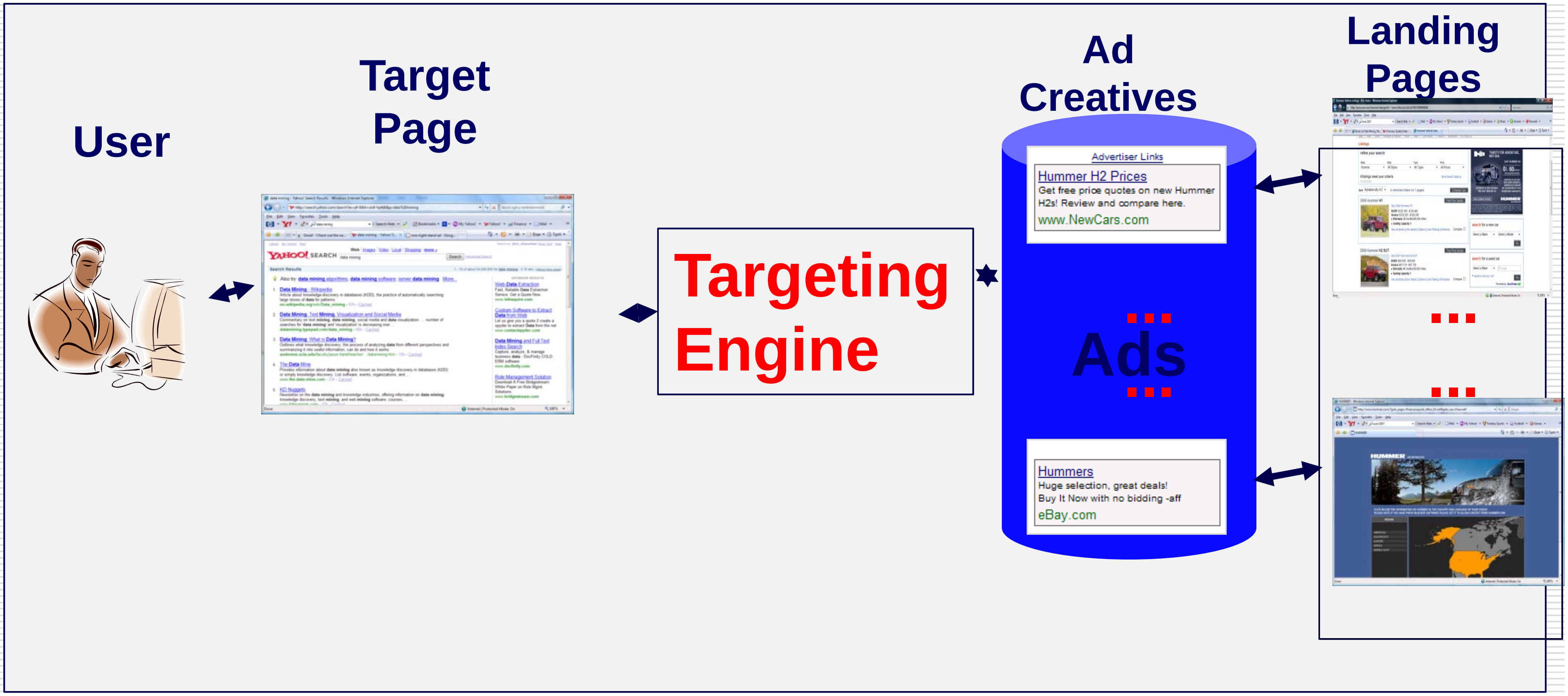
- What will you learn in this course?

  Perspectives and landscape in:

  - Statistics, Data Mining, and Business Analytics
  - Online marketing, computational advertising, healthcare services, financial services, service/call centers, social networks, recommenders and text mining
  - Distinction between combining “commoditized” algorithms for real world problems vs. creating new ones

Knowledge Services Examples

Online Marketing (Ranking Ads)



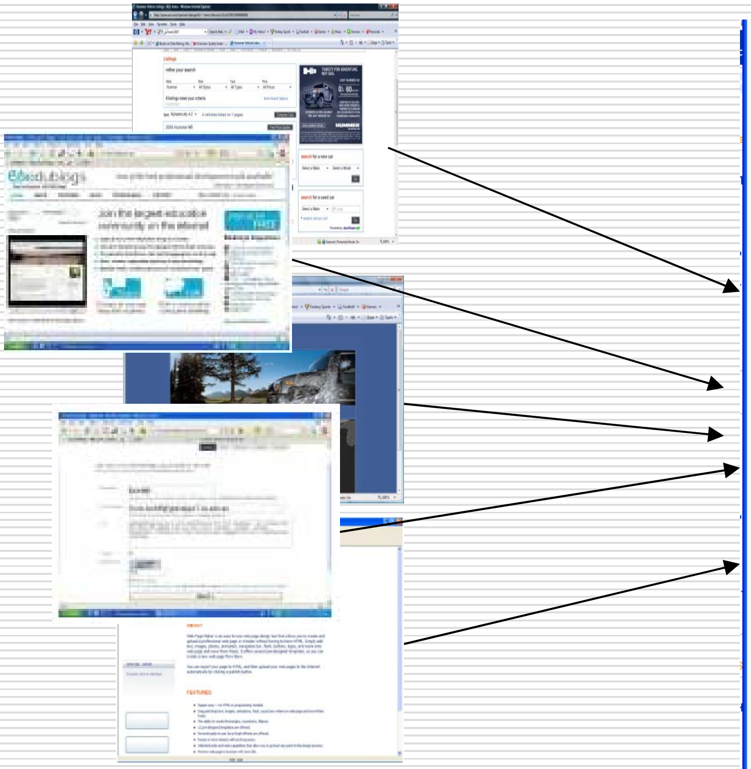
Advertiser Links

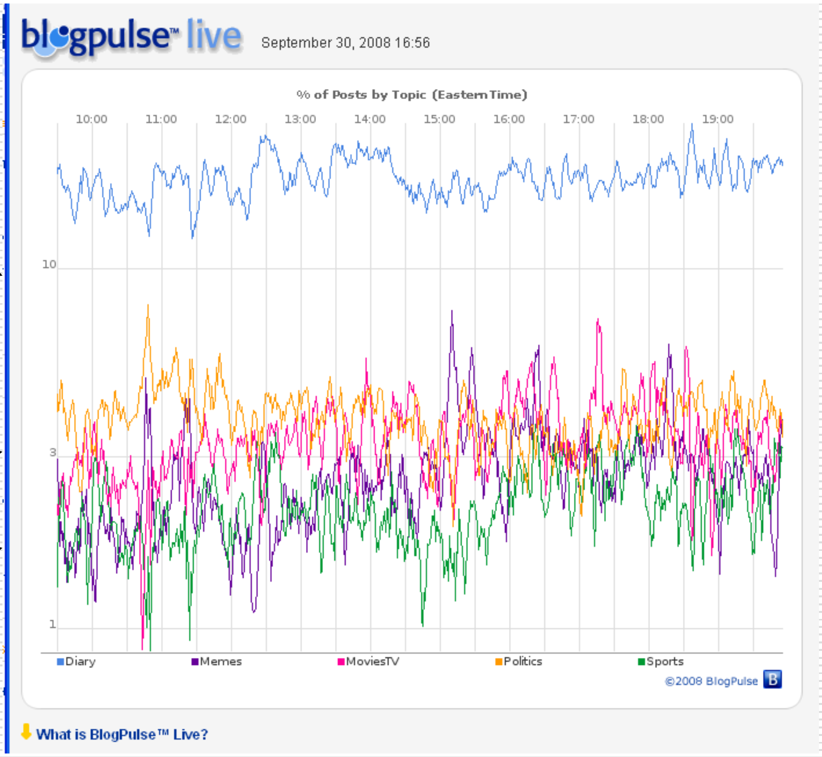
Hummer H2 Prices
Get free price quotes on new Hummer H2s! Review and compare here.
www.NewCars.com

Hummers
Huge selection, great deals!
Buy It Now with no bidding -aff
eBay.com

Knowledge Services Examples

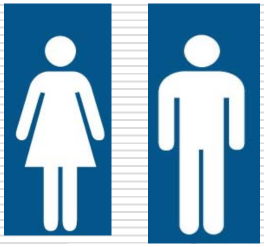
Opinion Mining (Blog Trend)

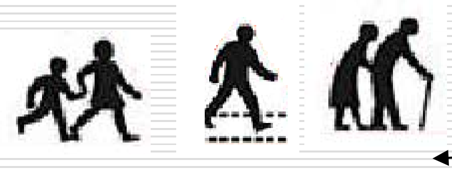


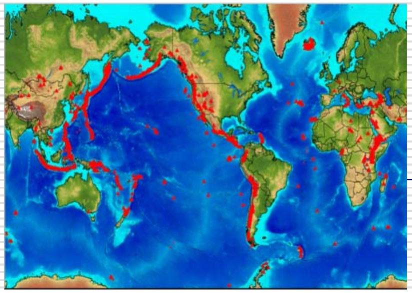


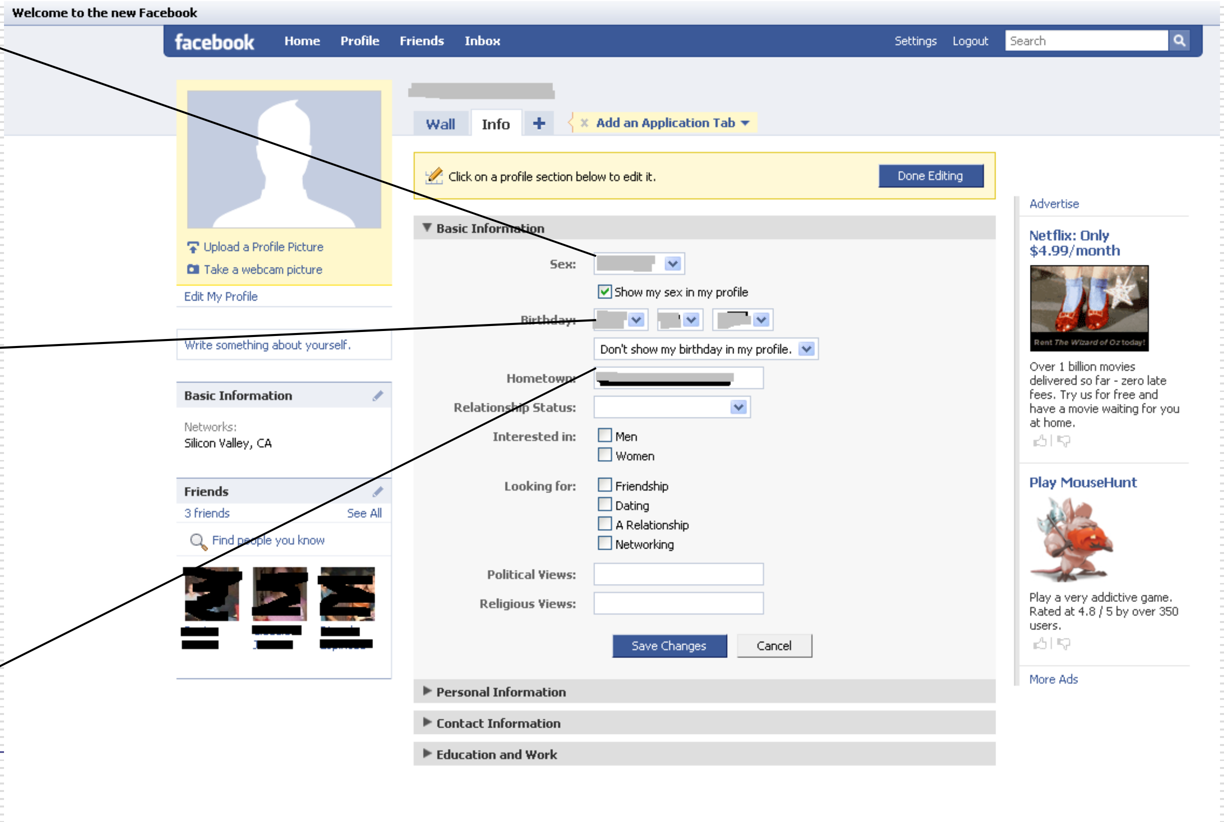
Knowledge Services Examples

- Social Networks









Welcome to the new Facebook

[redacted]

Upload a Profile Picture

Take a webcam picture

Edit My Profile

Write something about yourself.

Basic Information

Networks:
Silicon Valley, CA

Friends

3 friends See All

Find people you know

[redacted]

Click on a profile section below to edit it. Done Editing

▼ **Basic Information**

Sex: [redacted]

☑ Show my sex in my profile

Birthday: [redacted]

Don't show my birthday in my profile.

Hometown: [redacted]

Relationship Status:

Interested in: □ Men □ Women

Looking for: □ Friendship □ Dating □ A Relationship □ Networking

Political Views:

Religious Views:

Save Changes Cancel

▶ **Personal Information**

▶ **Contact Information**

▶ **Education and Work**

Knowledge Services and Data Mining

□ What are Knowledge Services?

□ What is Data Mining? Business Analytics?

□ What is the connection between all three?

Services

- What is a service?
  - http://en.wikipedia.org/wiki/Service_(economics)
- A **service** is the non-material equivalent of a good. A service provision is an economic activity that does not result in ownership
- Service professions
  - http://www.bls.gov/oco/oco1006.htm
- Management and Business Professionals
  - http://www.bls.gov/oco/oco1001.htm

Knowledge Services

- Marketing
  - Internet and other marketing campaigns
  - Online (computational) advertising
  - Customer identification and churn
- Financial Services
  - How should you invest?
  - What are stock and industry trends?
  - Fraud detection
  - Banks, investment, and risks

Knowledge Services (Continued)

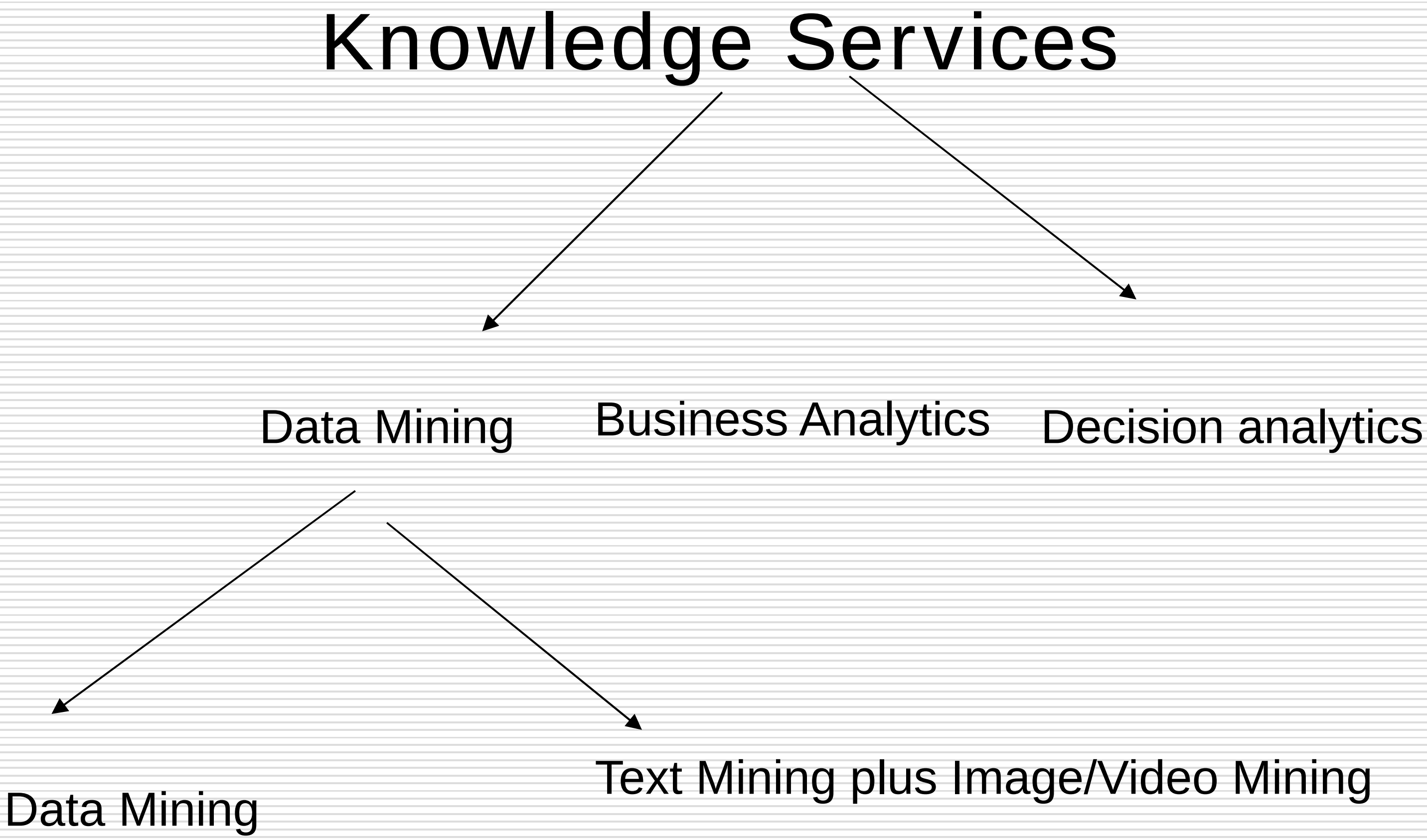
- Health Services
  - Body fat profile and weight prediction
  - Cancer identification
  - Social networks for diabetes knowledge sharing
- Service Centers
  - Call center management
  - Network prognostics and diagnostics
    - Anomaly detection

Data Mining and Business Analytics

- Data Mining and Business Analytics
  - Techniques to model and solve Knowledge Services problems

- Decision Theory is an aspect of business analytics
  - Techniques to solve business management decision making
  - E.g. How many experts and technicians of each type in a service center

Data Mining and Text Mining



Knowledge Services

Data Mining

Business Analytics

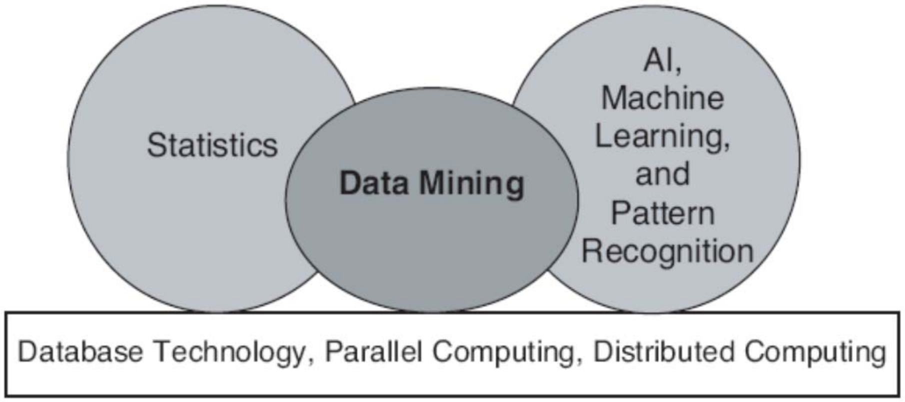
Decision analytics

Data Mining

Text Mining plus Image/Video Mining

Statistics and Data Mining - 1

- How are statistics and data mining related?
- Or are they not?


Statistics
Data Mining
AI, Machine Learning, and Pattern Recognition
Database Technology, Parallel Computing, Distributed Computing

Data Mining: Definitions

- □ Data mining is the nontrivial process of identifying, novel, potentially useful, and ultimately understandable patterns in data. - Fayyad.
- □ Data mining is the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions. - Zekulin.
- □ Data Mining is a set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data. - Ferruzza.
- □ Data mining is the process of discovering advantageous patterns in data. - John

Statistics

- Hypothesis testing
- Experimental design
- Response surface modeling
- ANOVA, MANOVA, etc.
- Linear regression
- Discriminant analysis
- Logistic regression
- GLM
- Canonical correlation
- Principal components
- Factor analysis

Data Mining

- □ Decision tree induction (C4.5, CART, CHAID)
- □ Rule induction (AQ, CN2, Recon, etc.)
- □ Nearest neighbors (case based reasoning)
- □ Clustering methods (data segmentation)
- □ Association rules (market basket analysis)
- □ Feature extraction
- □ Visualization
- □ In addition, some include:
- □ Neural networks
- □ Bayesian belief networks (graphical models)
- □ Genetic algorithms
- □ Self-organizing maps

Statistics to Data Mining Transition

- DM packages implement well known procedures from machine learning, pattern recognition, neural networks and data visualization.
- Statistics concentrate on probabilistic inference in information science while DM also finds patterns in the data.
- Dimensionality reduction with statistical assumptions can be applied in DM (PCA).
- Assessing data quality.

Machine Learning and Data Mining Algorithms and Approaches

- Unsupervised
  - Only machine, no human inputs, labeling etc. (Only X(i) s)
- Supervised
  - Human inputs, labeling etc. included Relating Y(i)s and X(i)s
- Reinforcement Learning
  - Humans provide rewards (no labels etc.)

Basic concepts -1

- Given a set of data X(i), i=1, .., N
  - X(i) multi -dimensional
  - E.g. documents and terms, health vitals
- N very large
- Questions:
  - How do we group "similar" points
    => Clustering
  - How do we reduce dimensionality?
    => Principal Components – combine similarly "behaving" components

Basic concepts - 2

- Y= f(X)+ noise
- If Y is discrete, we have classification
- If Y is continuous, we have prediction

Prediction and Classification

- Classification
  - Classification is the task of assigning objects to one of several predefined categories.
  - E.g. Is a borrower a high risk or not (in terms of defaulting on payments)?
- Prediction
  - A prediction is a statement or claim that a particular event or value will occur in the future in more certain terms than a forecast.
  - Real estate prices based on features of home and location

In DM, typically these tasks are performed based on a set of attributes which describe the object to classify or the variable to predict.

Class Administration

- **Office Hours:**
  By appointment in the 4-6 pm window (or variants) on Wednesday (or by phone).
  Location: South Hall 205; on January 25 and March 7, SDH 422

- Usually, we are hard task masters
- This time, for the seminar 296A-2, we will phase in nice and easy
- 296A-3, Advanced Project course, will be demanding

- Please look at website for Course Objectives, Philosophy, separation of doctoral and masters groups from study and assignment perspective
- http://courses.ischool.berkeley.edu/i296a-dsa/s12/

- Please review Speaker Schedule
- Please review Assignment Schedule
  - Weekly, monthly, and final submission

Basic Exposure to Data Mining

□ Best approach: First half of i290 – Data Mining and Analytics

- Clustering
- Classification
- Prediction
- Mining frequent patterns

□ http://courses.ischool.berkeley.edu/i290-dma/s12/doku.php

For Basic Help with Data Mining

- Contact Jimi Shanahan

james.shanahan@gmail.com