Introduction to Data Mining

CS 584
Data Mining
(Fall 2016)

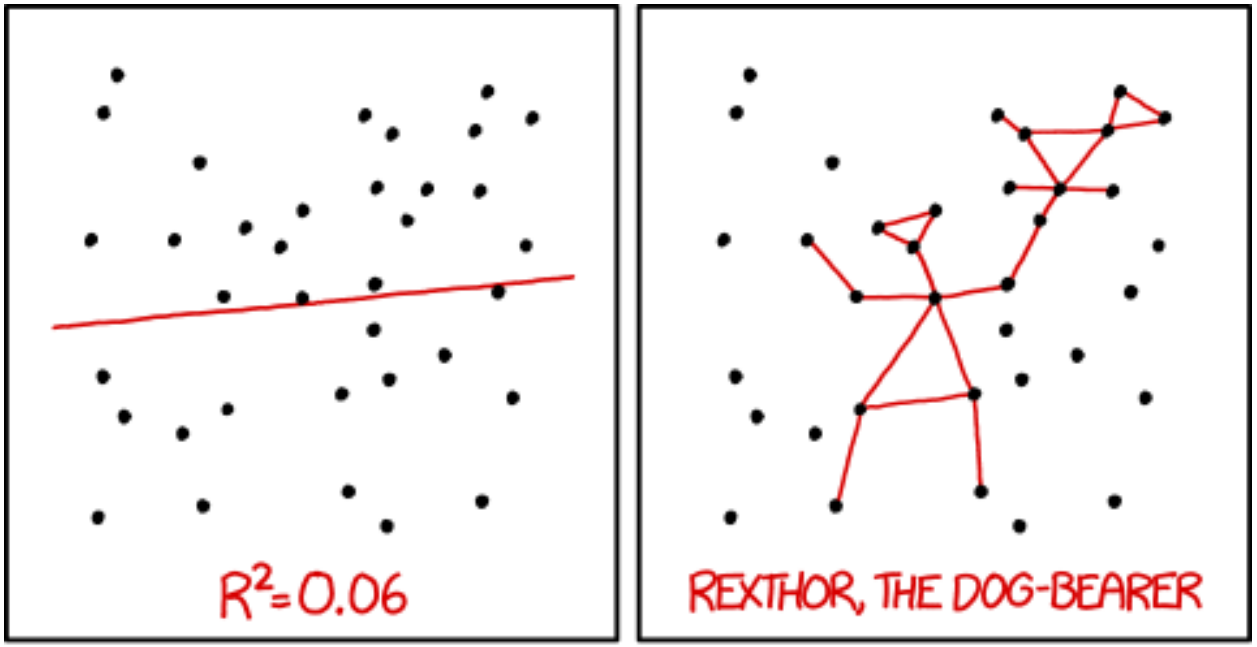
Huzefa Rangwala
AssociateProfessor,
Computer Science
George Mason University
Email: rangwala@cs.gmu.edu
Website: www.cs.gmu.edu/~hrangwal

Slides are adapted from the available book slides developed by Tan, Steinbach and Kumar

Roadmap for Today

- Welcome & Introduction
  - Survey (Show of hands)
- Introduction to Data Mining
  - Examples, Motivation, Definition, Methods
- Administrative/ Class Policies & Syllabus
  - Grading, Assignments, Exams, Policies
- 10-15 minute break.
- Data
  - Lets begin!

What do you think of data mining?



- **Please could you write down examples that you know of or have heard of on the provided index card.**

- **Also write down your own definition.**

Source: xkcd.com

Data Deluge

http://www.economist.com/node/15579717

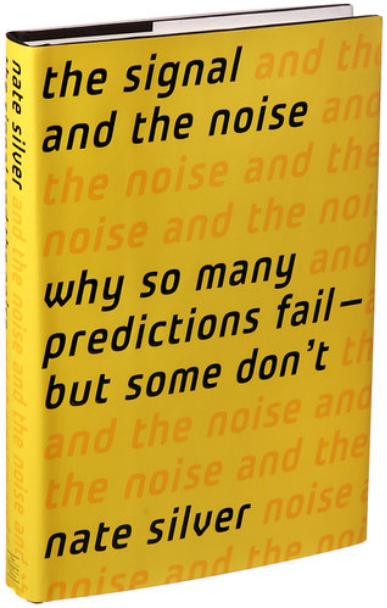
Political Data Mining

Inside the Secret World of the Data Crunchers Who Helped Obama Win

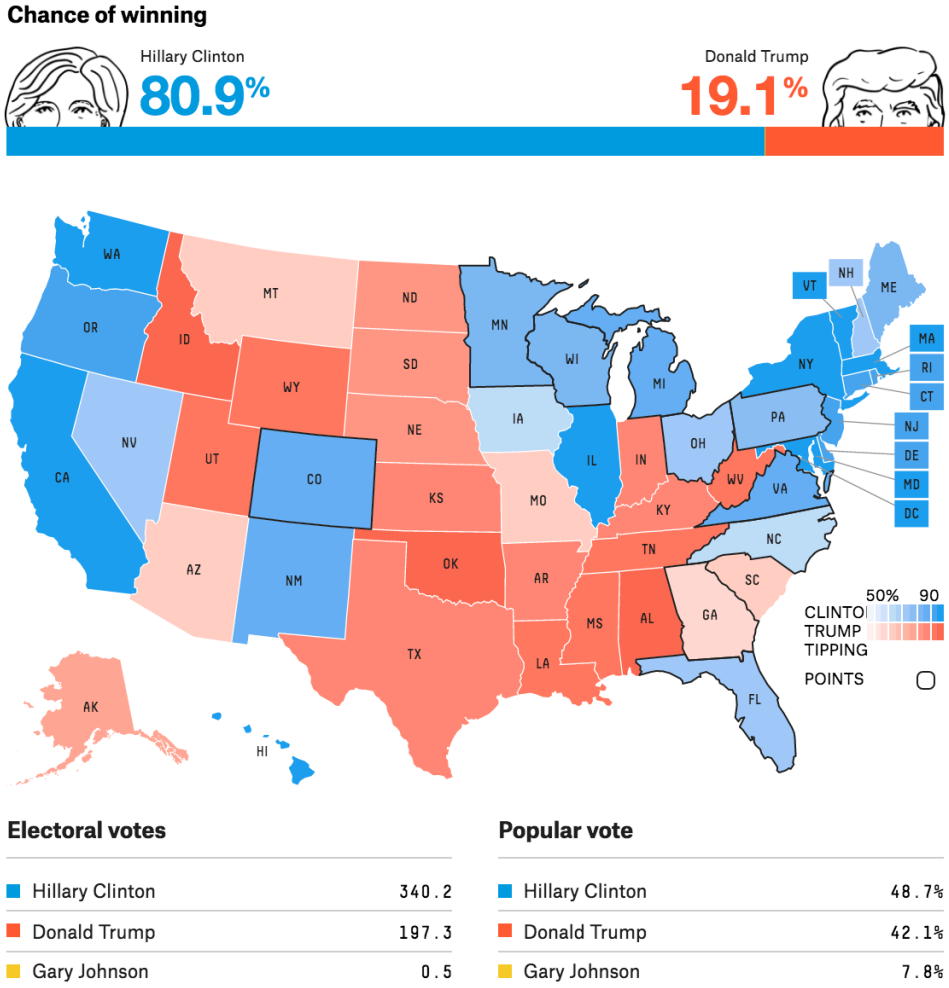
Read more:
http://swampland.time.com/2012/11/07/inside-the-secret-world-of-quants-and-data-crunchers-who-helped-obama-win/#ixzz2luhEmNcB

Mining Truth From Data Babel --- Nate Silver

http://www.nytimes.com/2012/10/24/books/nate-silvers-signal-and-the-noise-examines-predictions.html?_r=0



Today (08/29)



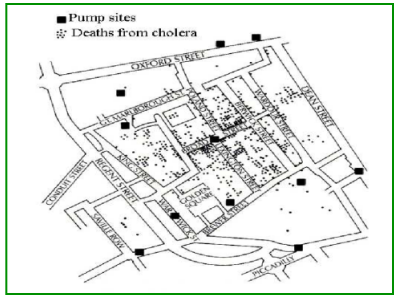
Source:http://projects.fivethirtyeight.com/2016-election-forecast/?ex_cid=rrpromo

Large-scale Data is Everywhere!

- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- New mantra
  - Gather whatever data you can whenever and wherever possible.
- Expectations
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.

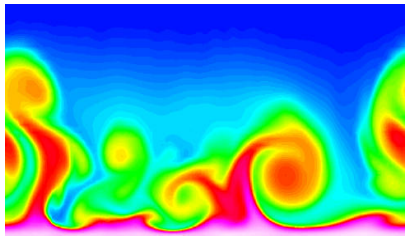
Homeland Security

Business Data



Geo-spatial data

Sensor Networks



Computational Simulations

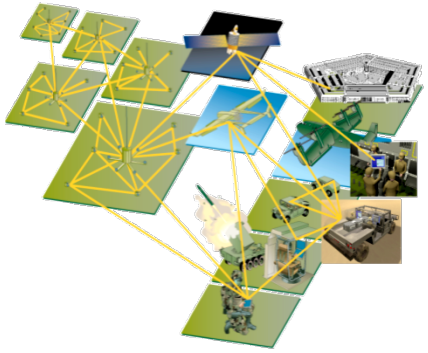
Why Data Mining? Commercial Viewpoint

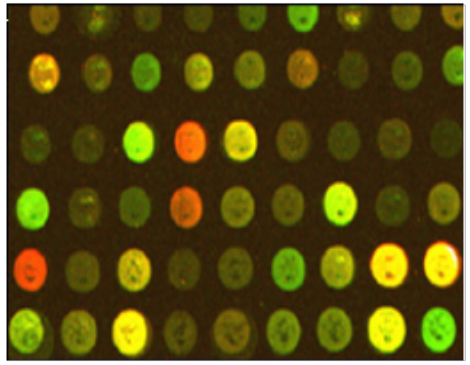
- Lots of data is being collected and warehoused
  - Web data
    - Yahoo has PBs of web data
    - Facebook has 1.7 Billion Active users
  - purchases at department/ grocery stores, e-commerce
    - Amazon records several million items/day
  - Bank/Credit Card transactions

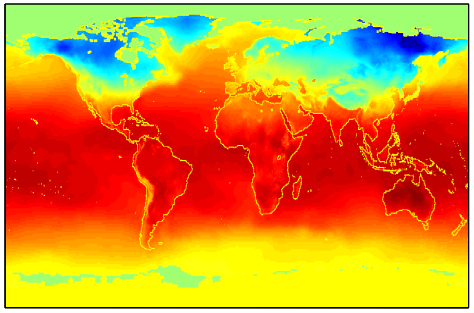
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

Why Data Mining? Scientific Viewpoint

- Data collected and stored at enormous speeds
  - remote sensors on a satellite
    - NASA EOSDIS archives over 1-petabytes of earth science data / year
  - telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - scientific simulations
    - terabytes of data generated in a few hours
- Data mining helps scientists
  - in automated analysis of massive datasets
  - In hypothesis formation







Mining Scientific Data - Fields

- Past decade has seen a huge growth of interest in mining data in a variety of scientific domains

- **Astroinformatics**
- **Neuroinformatics**
- **Quantum Informatics**
- **Health Informatics**
- **Evolutionary Informatics**
- **Veterinary Informatics**
- **Organizational Informatics**
- **Pharmacy Informatics**

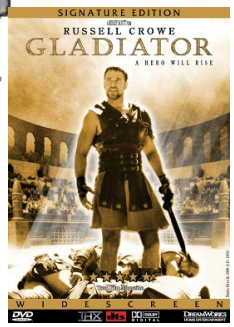
- Social Informatics
- Ecoinformatics
- Geoinformatics
- Chemo Informatics

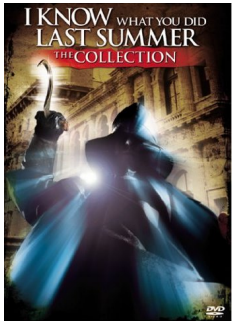
My Favorite Data Mining Examples

"Remember, the other team is counting on Big Data insights based on previous games. So, kick the ball with your other foot."

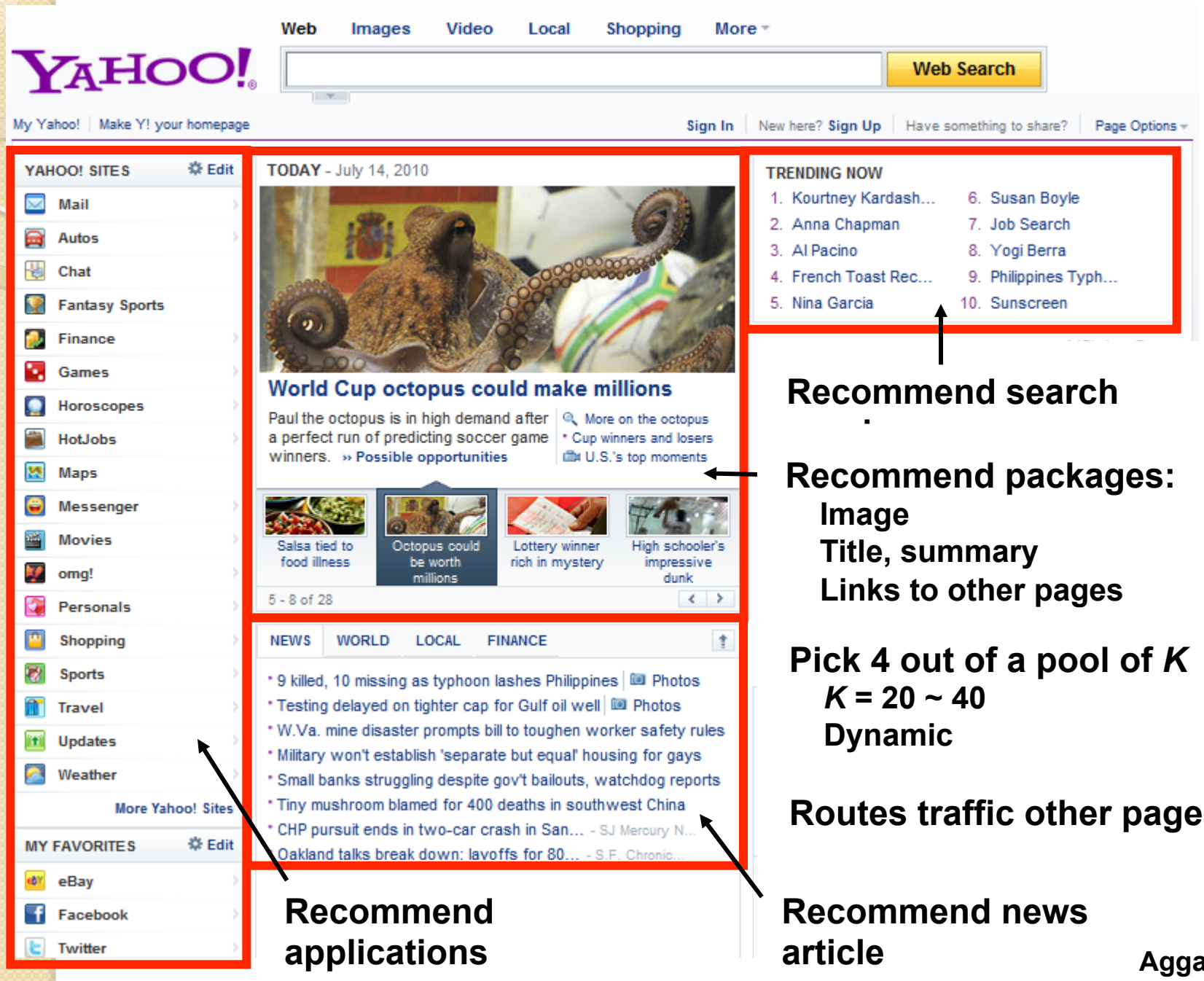
"Sweetheart, my neural net predicts that you and I are 98.9% compatible. Will you be my Valentine?"

Recommender systems






Netflix Prize
Home
Rules
Leaderboard
Register
Update
Submit
Download
NETFLIX
Browse
Recommendations
Friends
Queue
Buy DVDs
Home
Genres
New Releases
Previews
Netflix Top 100
Movies For You
Randy, the following movies were chosen based on your interest in:
Bowling for Columbine
Carnivale: Season 1
Fahrenheit 9/11
The Big One
All Discs Guaranteed!
You really liked it...
Now o for just $5.99
Shop
titles
as lo
Original art
Lewis Black: Re and Scre
Add
Not Interested
Not Interested
Guides:
Member Favorites
Easter Eggs
By Decade
By Studio
Movies You've Seen
Red Eye
Rear Window
Give a friend
Welcome!
The Netflix Prize seeks to substantially improve the accuracy of predictions about how much someone is going to love a movie based on their movie preferences. Improve it enough and you win one (or more) Prizes. Winning the Netflix Prize improves our ability to connect people to the movies they love.
Read the Rules to see what is required to win the Prizes. If you are interested in joining the quest, you should register a team.
You should also read the frequently-asked questions about the Prize. And check out how various teams are doing on the Leaderboard.
Good luck and thanks for helping!



Recommend search

Recommend packages:
- Image
- Title, summary
- Links to other pages

Pick 4 out of a pool of *K*
- ***K*** **= 20 ~ 40**
- **Dynamic**

Routes traffic other pages

Recommend applications

Recommend news article

Aggarwal

Collaborative filtering

- Recommend items based on past transactions of users
- Analyze relations between users and/or items
- Specific data characteristics are irrelevant
  - Domain-free: user/item attributes are not necessary
  - Can identify elusive aspects

amazon.com

Customers who bought items in your Recent History also bought:

I Own It Not interested

Rate it

Add to Cart Add to Wish List

I Own It Not interested

Rate it

Add to Cart Add to Wish List

I Own It Not interested

Rate it

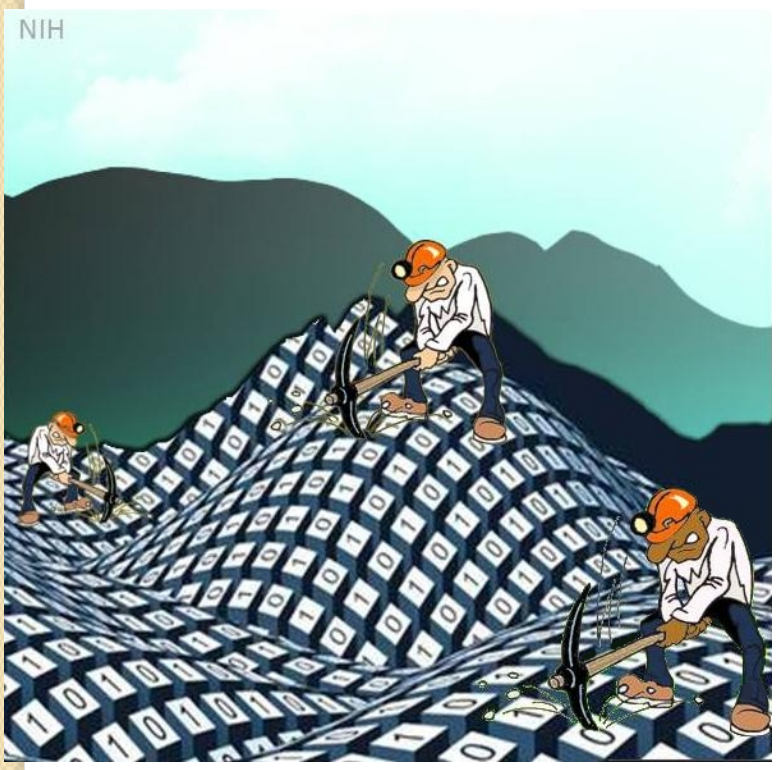
Add to Cart Add to Wish List

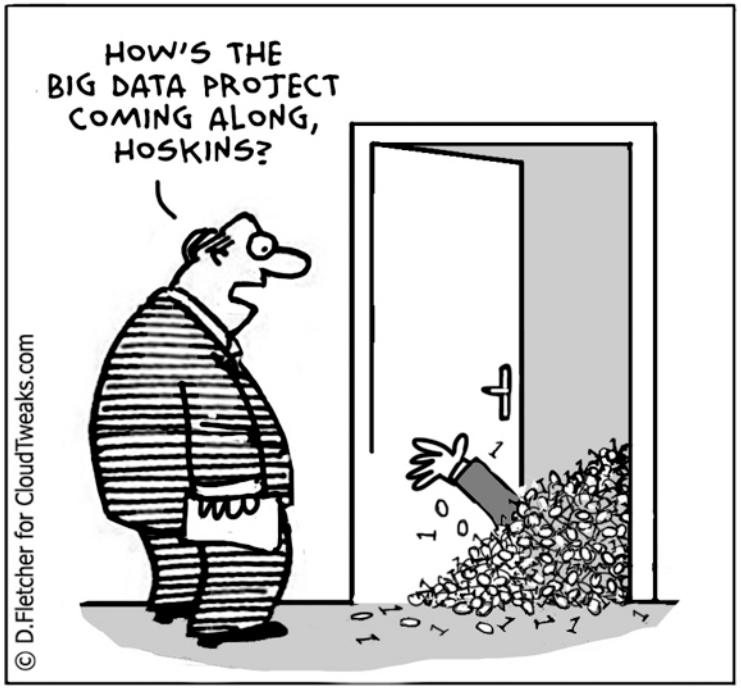
My Favorite Data Mining Examples

- Amazon.com, Google, Netflix
  - Personal Recommendations.
  - Profile-based advertisements.
- Spam Filters/Priority Inbox
  - Keep those efforts to pay us millions of dollars at bay.
- Scientific Discovery
  - Grouping patterns in sky.
  - Inferring complex life science processes.
  - Forecasting weather.
- Security
  - Phone Conversations, Network Traffic

Mining Large Data Sets - Motivation

- There is often information "hidden" in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all



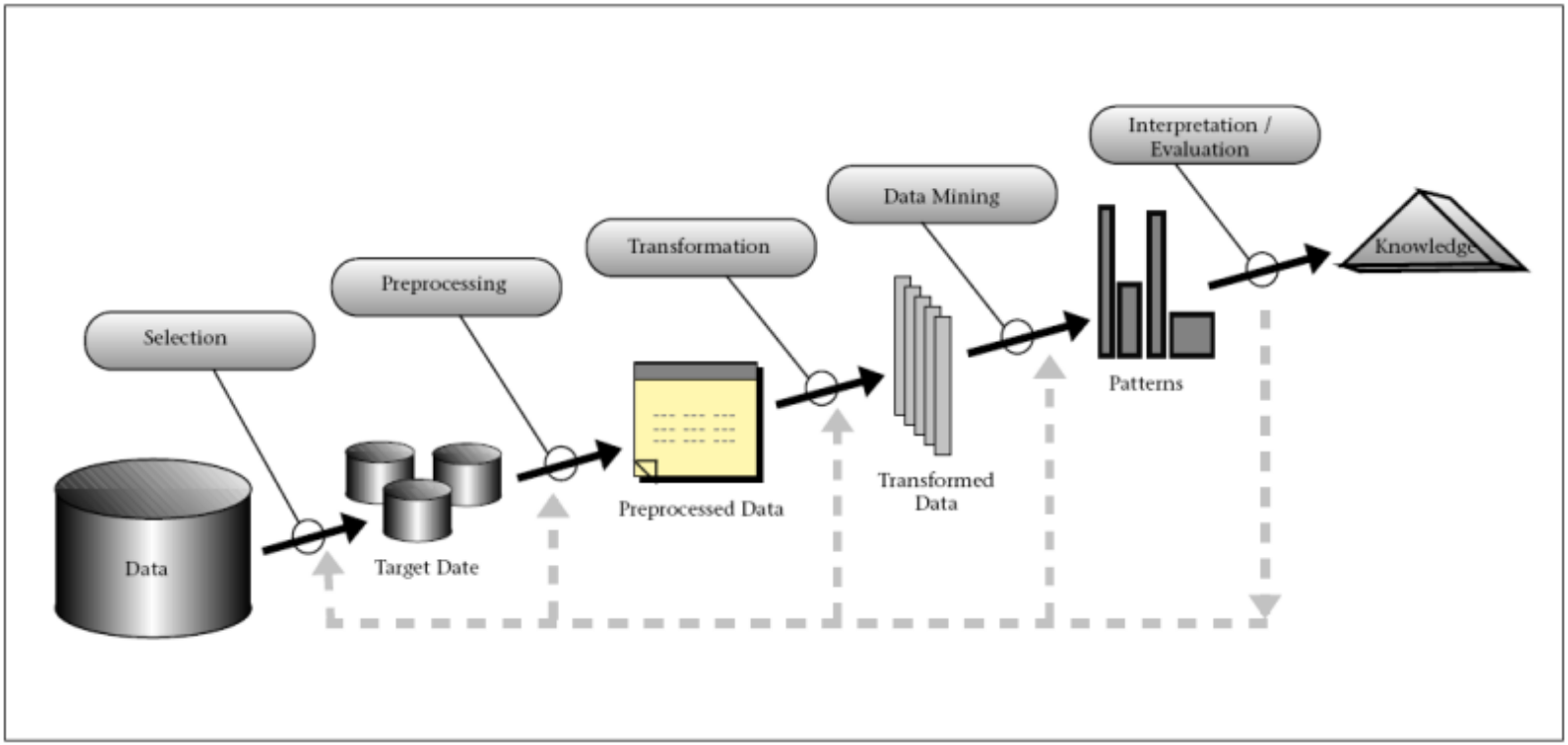


Data Mining Definitions

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data (normally large databases)
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.
- Part of the Knowledge Discovery in Databases Process.

KDD Process

CONVERTING RAW DATA TO USEFUL INFORMATION.



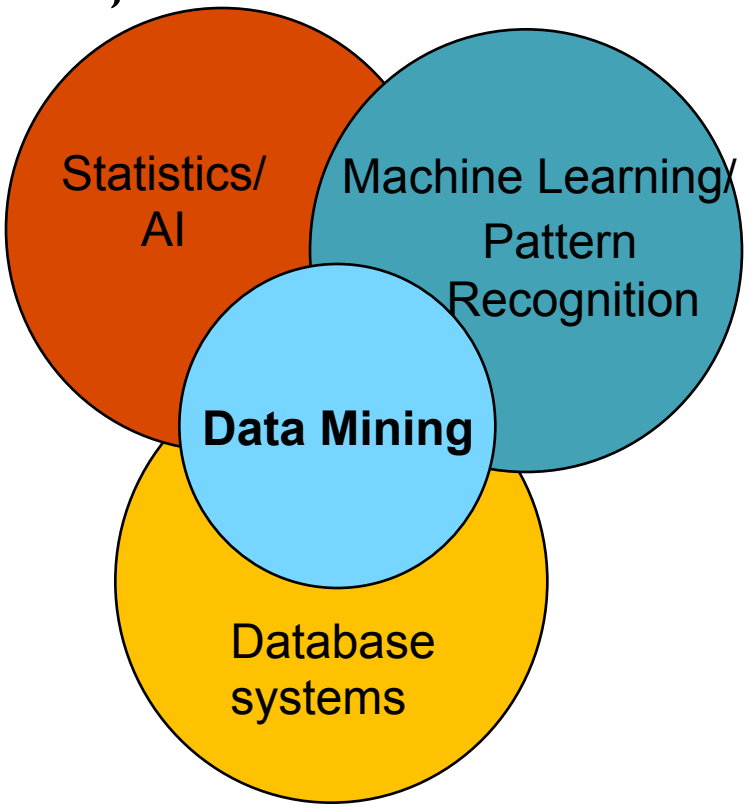
Fayyad 1996
http://liris.cnrs.fr/abstract/abstract.html

What is (not) Data Mining?

- What is not Data Mining?
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"
- What is Data Mining
  - Certain names are more prevalent in certain US locations (O' Brien, O' Rurke, O' Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
- Description Methods
  - Find human-interpretable patterns that describe the data. Make good inferences from the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

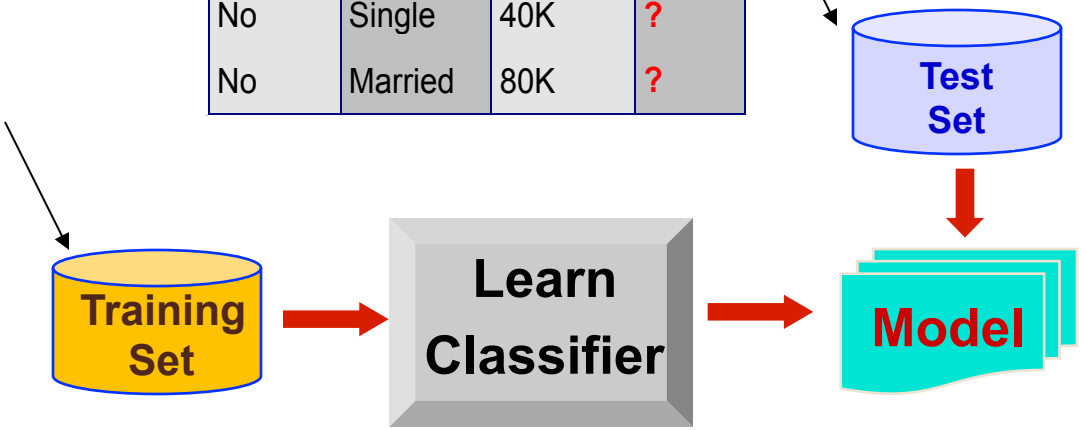
Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Classification Example

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification: Campaign Targeting

- Direct Marketing
  - Goal: Reduce cost of campaigning by *targeting* a set of voters likely to vote for candidate.
  - Approach:
    - Use the data for a similar candidate from history introduced before.
    - We know which voters decided to vote for and which decided otherwise. This {*yes, no*} decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and candidate-interaction related information about all such voters.
      - Type of donation, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Your Turn

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - What kind of data will you try to get ?
    - Can you say something about the characteristics of the data ?
    - Estimate the size of the data.
    - What kind of pitfalls you might run into ?

- You have 5 minutes to think and discuss.

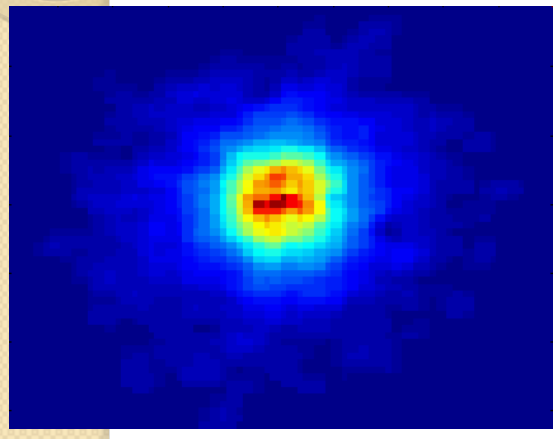
Classification: Fraud Detection

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
    - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Classifying Galaxies

Courtesy: http://aps.umn.edu

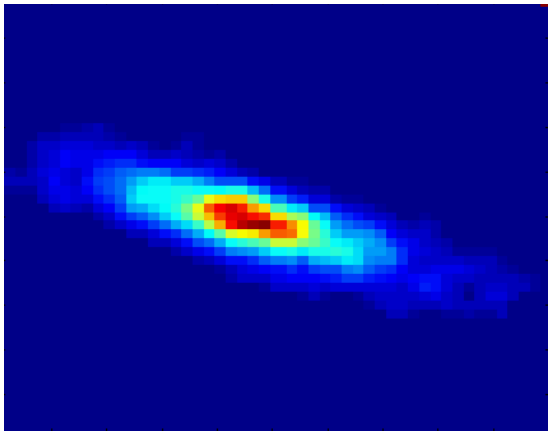
Early



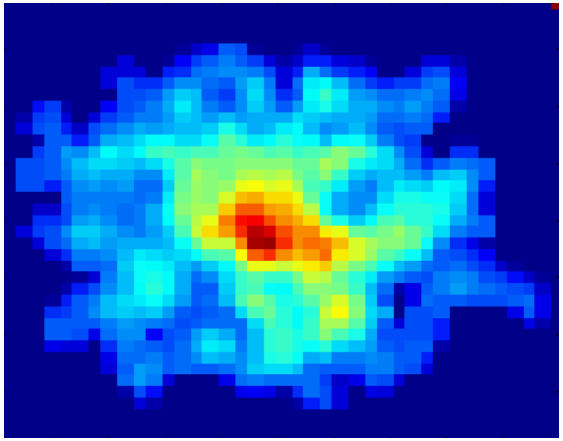
Class:
- **Stages of Formation**

Attributes:
- **Image features,**
- **Characteristics of light waves received, etc.**

Intermediate



Late



Data Size:
- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

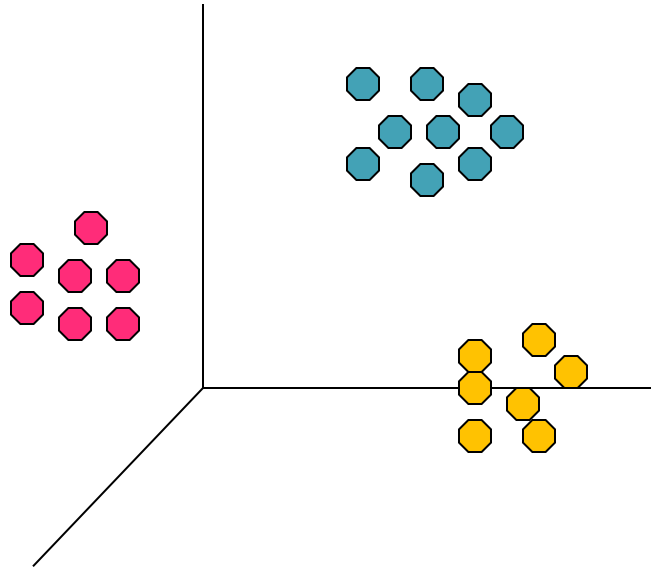
Clustering

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustrating Clustering

Intracluster distances are minimized

Intercluster distances are maximized



- Euclidean Distance Based Clustering in 3-D space.

Clustering: Document

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Illustrating Document Clustering

- Clustering Points: 3204 Articles of Los Angeles Times.
- Similarity Measure: How many words are common in these documents (after some word filtering).

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Think point ?

- Differences between classification and clustering?

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}

Urban Legend ….

- Classic Association Rule Example:
  - If a customer buys diaper and milk, then he is very likely to buy beer.
  - Any plausible explanations ? ☺

Association Rule Discovery: Application 1

- Marketing and Sales Promotion:
  - Let the rule discovered be

    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent and Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rule Discovery: Application 2

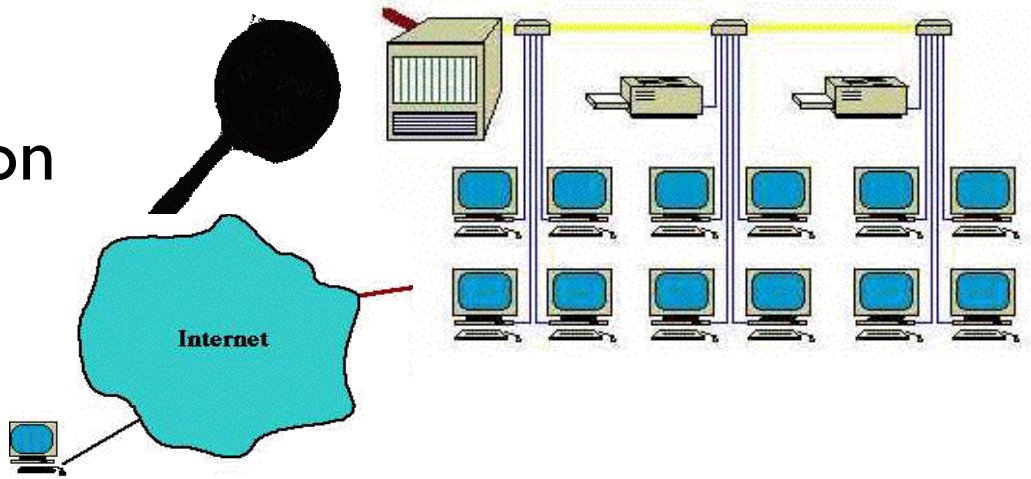
- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - Wal-mart, Target, and departmental store managers are big into this.
  - All your ticket gets processed & analyzed in a warehouse.

Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Also called density estimation.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection



What else can Data Mining do ?

Dilbert

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

CS 584 Class Semantics

CLASS Syllabus

Go To Piazza

Go To Miner