2612
Introduction to
Machine Learning and Data Mining

Patricia Hoffman, PhD

Machine Learning and Data Mining

Patricia Hoffman, PhD

Class Web Page:

https://online.ucsc-extension.edu/xsl-portal/site/a635060e-7190-452d-9c51-fd06b3a33e65

R References:

http://patriciahoffmanphd.com/statictical languager.php

Homework 01 Due Sunday before next class:

https://online.ucsc-extension.edu/xsl-portal/site/a635060e-7190-452d-9c51-fd06b3a33e65

UCSC Extension Silicon Valley

Upcoming Data Mining Events

Sign Up to Receive Announcements

Announcing Many Data Mining Events through Google Group:

http://groups.google.com/group/machine-learning-class

UCSC Extension Silicon Valley

Download Statistical Language R

Download the statistical R language:

http://cran.r-project.org/

Directions:

https://online.ucsc-extension.edu/xsl-portal/site/2a4f48bb-81ce-4236-b22b-758d12720b22/page/bf0a82c3-ae95-4e73-a465-1d5acbfb661e

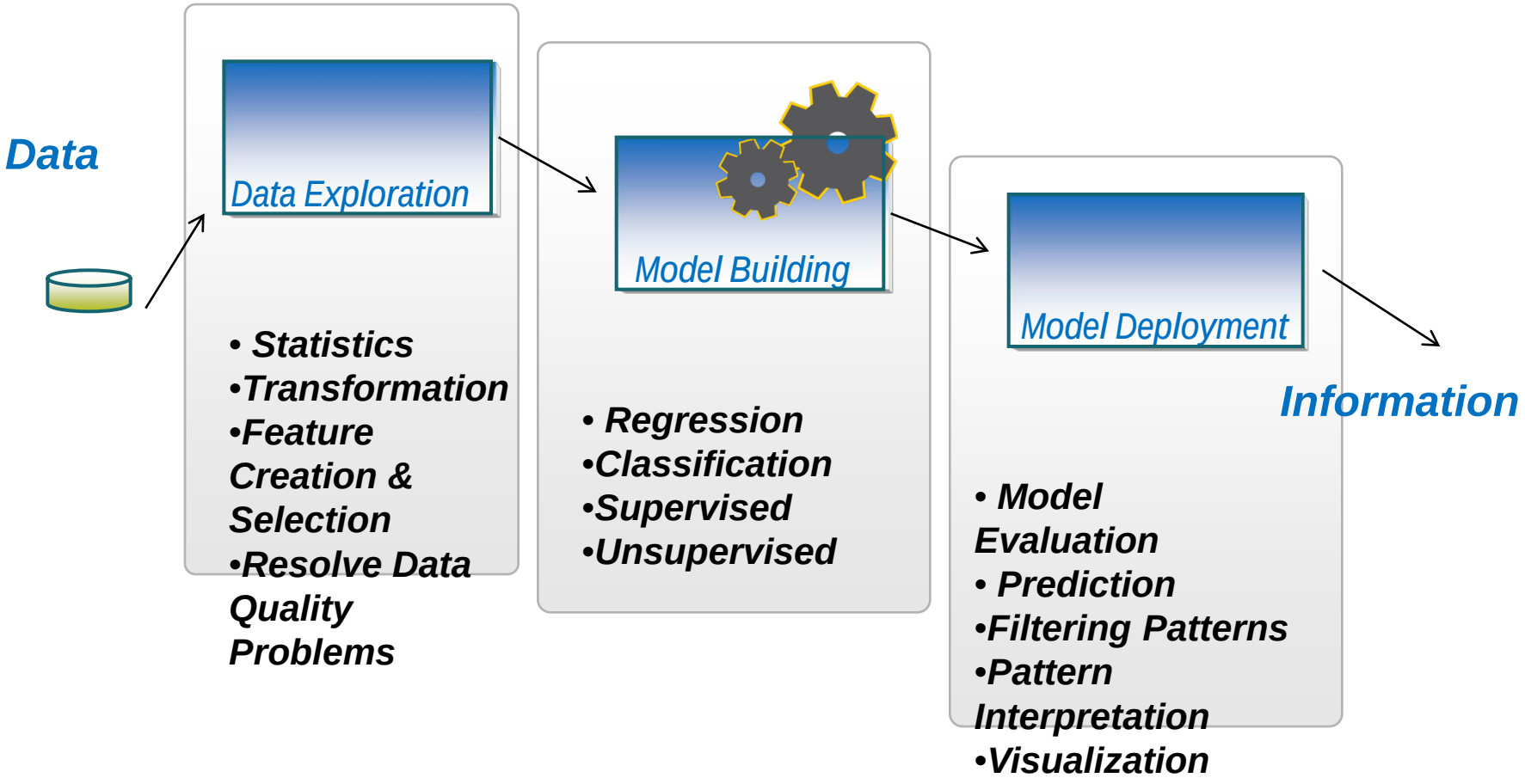
IDE for R Visual Studio: http://rstudio.org/

Integrate with statET:

http://www.splusbook.com/RIntro/R_Eclipse_StatET.pdf

UCSC Extension Silicon Valley

Analytic Model Building Process
Data Exploration

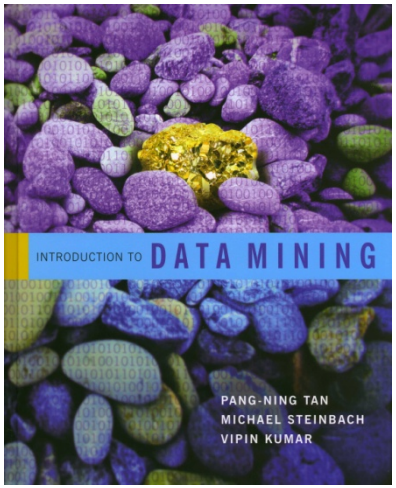


Data Mining: Introduction
Using Slides from

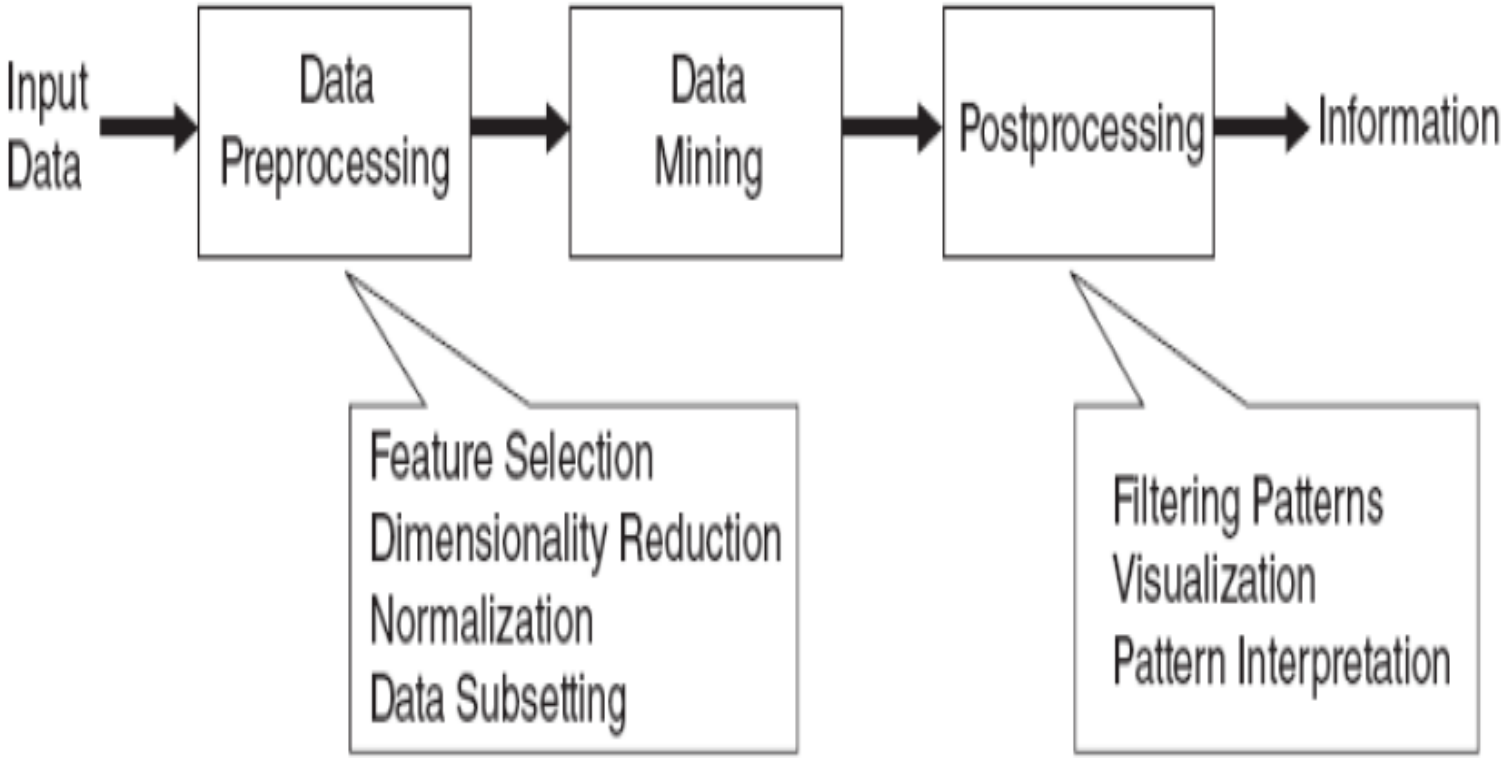
Introduction to Data Mining

by

Tan, Steinbach, Kumar



Knowledge Discovery Process



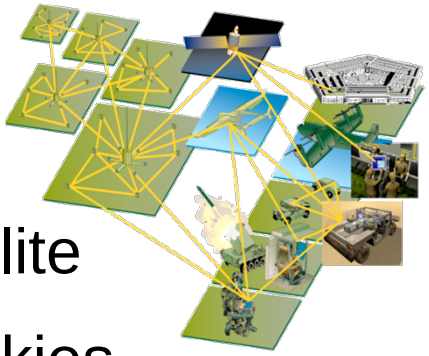
Why Mine Data? Commercial Viewpoint

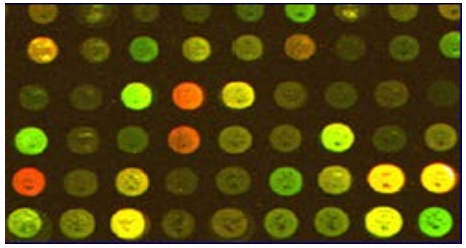
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

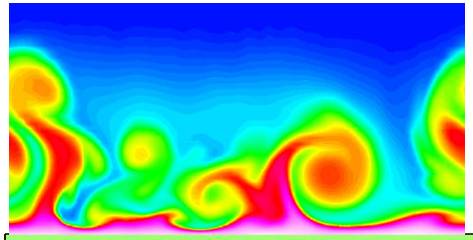
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

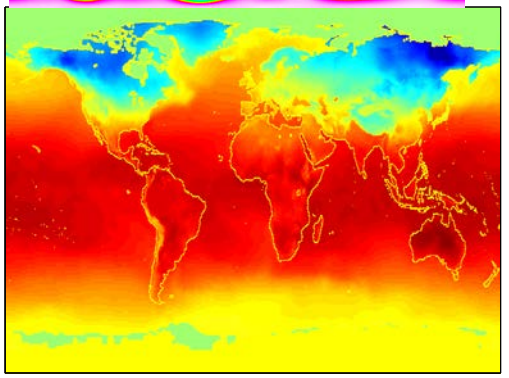
Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation



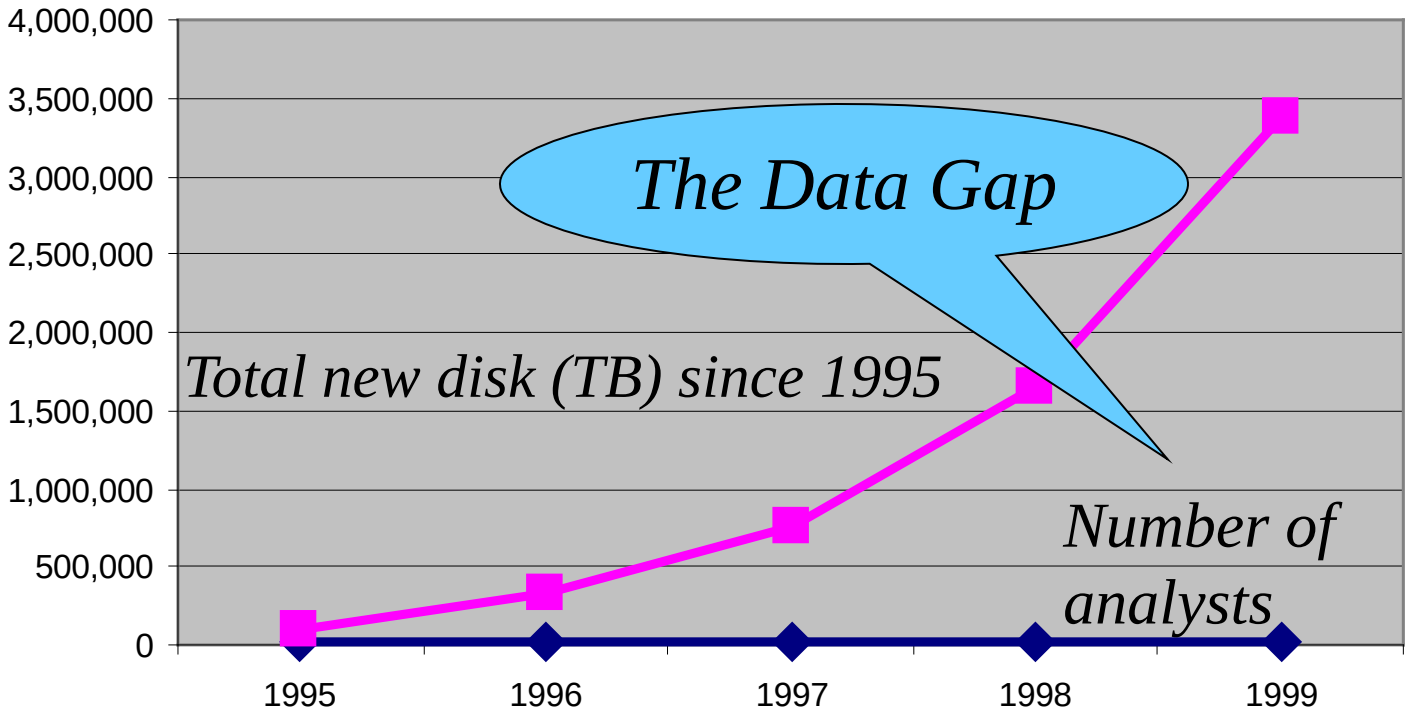






Mining Large Data Sets - Motivation

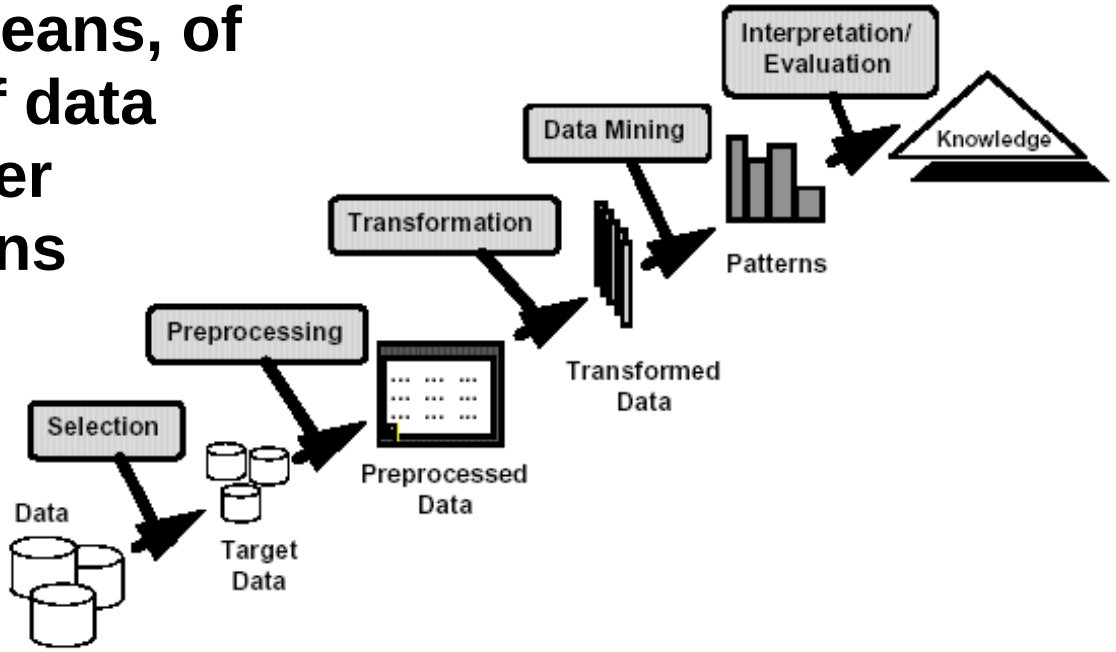
- **There is often information "hidden" in the data that is not readily evident**
- **Human analysts may take weeks to discover useful information**
- **Much of the data is never analyzed at all**



From: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

What is Data Mining?

- **Many Definitions**
  - **Non-trivial extraction of implicit, previously unknown and potentially useful information from data**
  - **Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns**



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,)*

Machine Learning

- Networks – Graphs
- Weights
- Learning
- Generalization
- Supervised Learning
- Unsupervised Learning

- Large Grant = $1,000,000
- Conference: French Alps

  Snowbird, Utah

Statistics

- Models
- Parameters
- Fitting
- Test Set Performance
- Regression / Classification
- Density Estimate/Clustering

- Large Grant = $50,000
- Conference:

  Los Vegas in August

Quote from Professor Robert Tibshirani, PhD

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.

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

Data Mining Tasks...

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

Group Discussion

1) Discuss whether or not each of the following activities is a data mining task:
   a. Dividing the customers of a company according to their gender
   b. Diving the customers of a company according to their profitability
   c. Computing the total sales of a company
   d. Sorting a student database based on student identification numbers

2) What are some data mining tasks that Netflix and Amazon have in common?

Classification: Definition

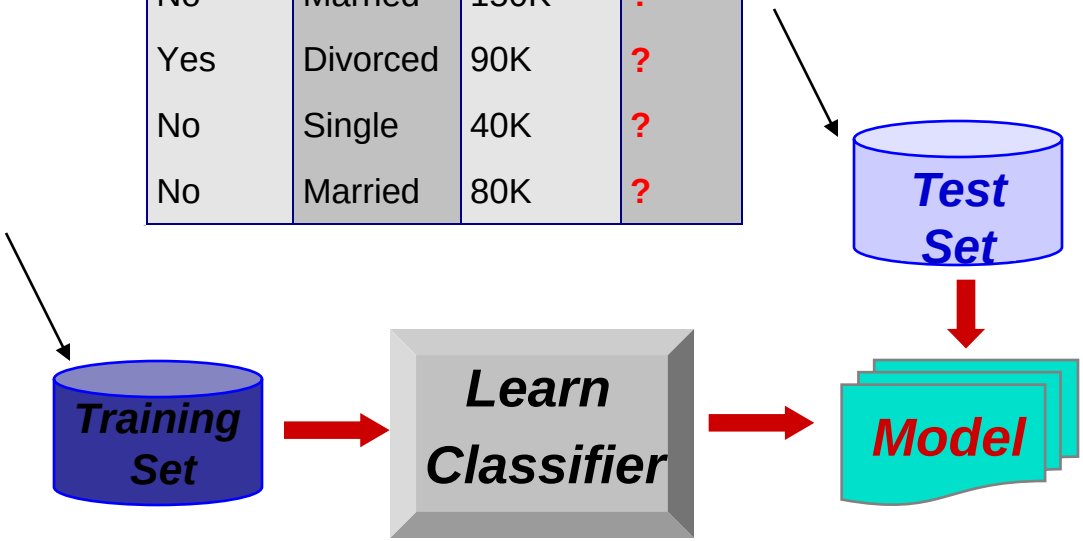
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - 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 Example

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
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: Application

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, 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: Application

- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

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

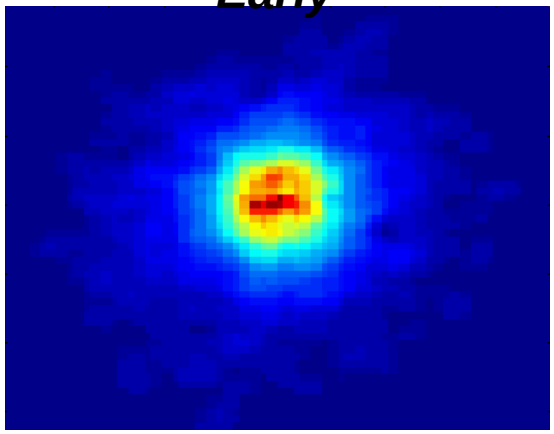
Classifying Galaxies

Courtesy: http://aps.umn.edu

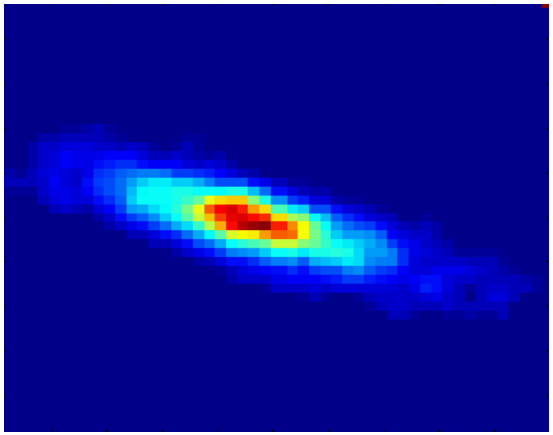
Class:
- *Stages of Formation*

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

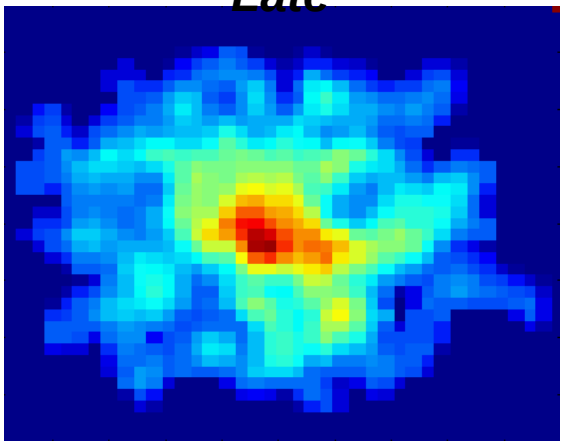
Early



Intermediate



Late



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

Group Discussion

1. Discuss which are data mining tasks:
   a. Predicting the future stock price of a company using historical records
   b. Monitoring the heart rate of a patient for abnormalities
   c. Monitoring seismic waves for earthquake activities
   d. Extracting the frequencies of a sound wave
2. Assume you are working at IBM and the decision makers asked, “What will keep our most important customers loyal?” How would you answer their question?

Clustering Definition

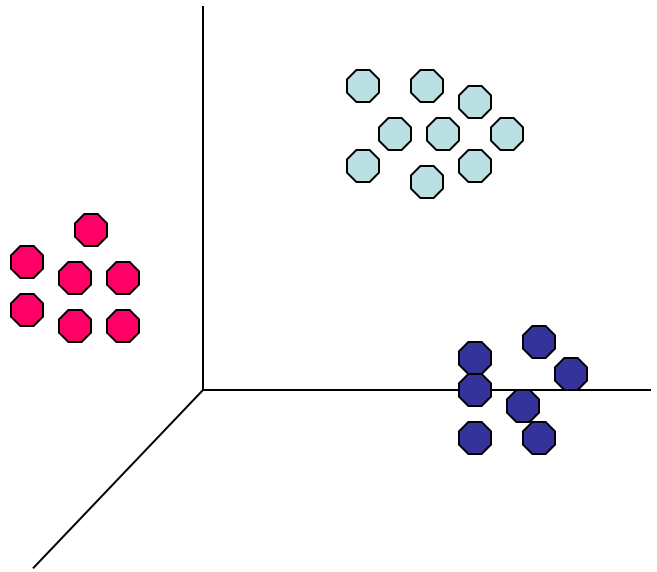
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance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

Illustrating Clustering

☒ *Euclidean Distance Based Clustering in 3-D space.*

Intracluster distances are minimized

Intercluster distances are maximized



Clustering: Application

- 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

Group Discussion

Suppose that you are employed as a data mining consultant for an Internet search engine company.

Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied.

Challenges of Data Mining

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

Resources: Datasets

- UCI Repository:
- http://www.ics.uci.edu/~mlearn/MLRepository.html
- UCI KDD Archive:
- http://kdd.ics.uci.edu/summary.data.application.html
- Statlib: http://lib.stat.cmu.edu/
- Delve: http://www.cs.utoronto.ca/~delve/

Resources: Datasets

- UCI Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html
- UCI KDD Archive: http://kdd.ics.uci.edu/summary.data.application.html
- Statlib: http://lib.stat.cmu.edu/
- Delve: http://www.cs.utoronto.ca/~delve/

Resources: Journals

- Journal of Machine Learning Research www.jmlr.org
- Machine Learning
- Neural Computation also Neural Networks
- IEEE Transactions on Neural Networks
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- Annals of Statistics
- Journal of the American Statistical Association

Group Discussion

Discuss whether or not each of the following activities is a data mining task:

a. Dividing the customers of a company according to their gender

b. Diving the customers of a company according to their profitability

c. Computing the total sales of a company

d. Sorting a student database based on student identification numbers

e. Predicting the outcomes of tossing a (fair) pair of dice

Group Discussion

Discuss whether or not each of the following activities is a data mining task:

a. Predicting the future stock price of a company using historical records

b. Monitoring the heart rate of a patient for abnormalities

c. Monitoring seismic waves for earthquake activities

d. Extracting the frequencies of a sound wave

Group Discussion

1) Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied.

2) How do eBay's data mining challenges differ from Netflix's data mining problems?


ucsc Extension
Silicon Valley