

QRE<sub>Lab</sub>

Quality & Reliability Engineering Laboratory

آزمایشگاه مهندسی کیفیت و پایداری

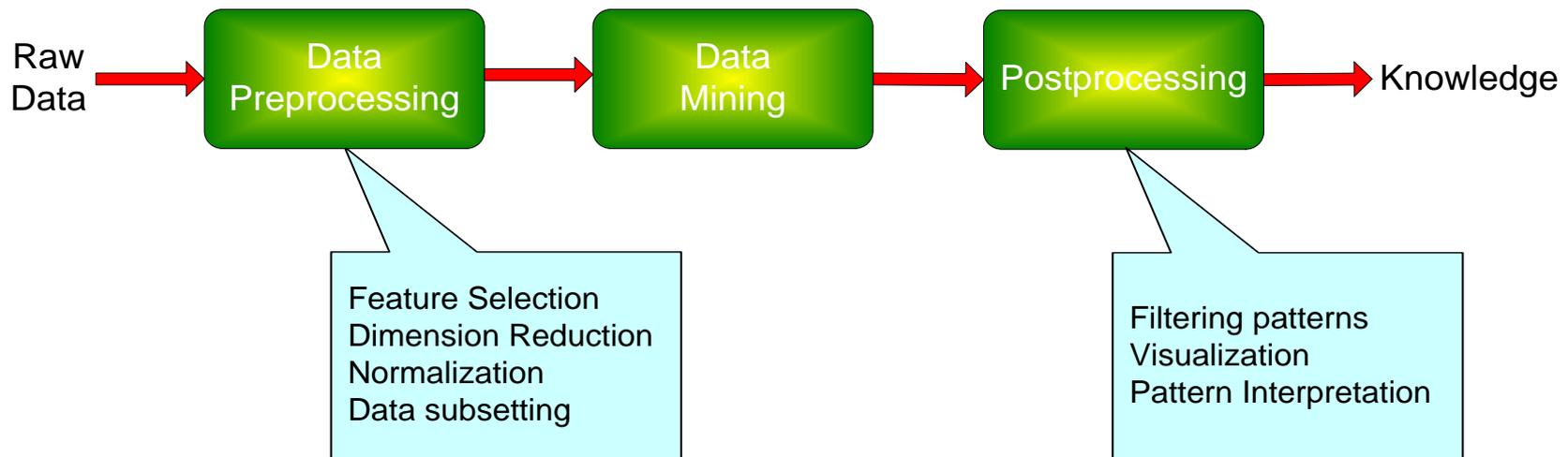


# Data Mining

## Lecture 1: Introduction

# What is Data Mining?

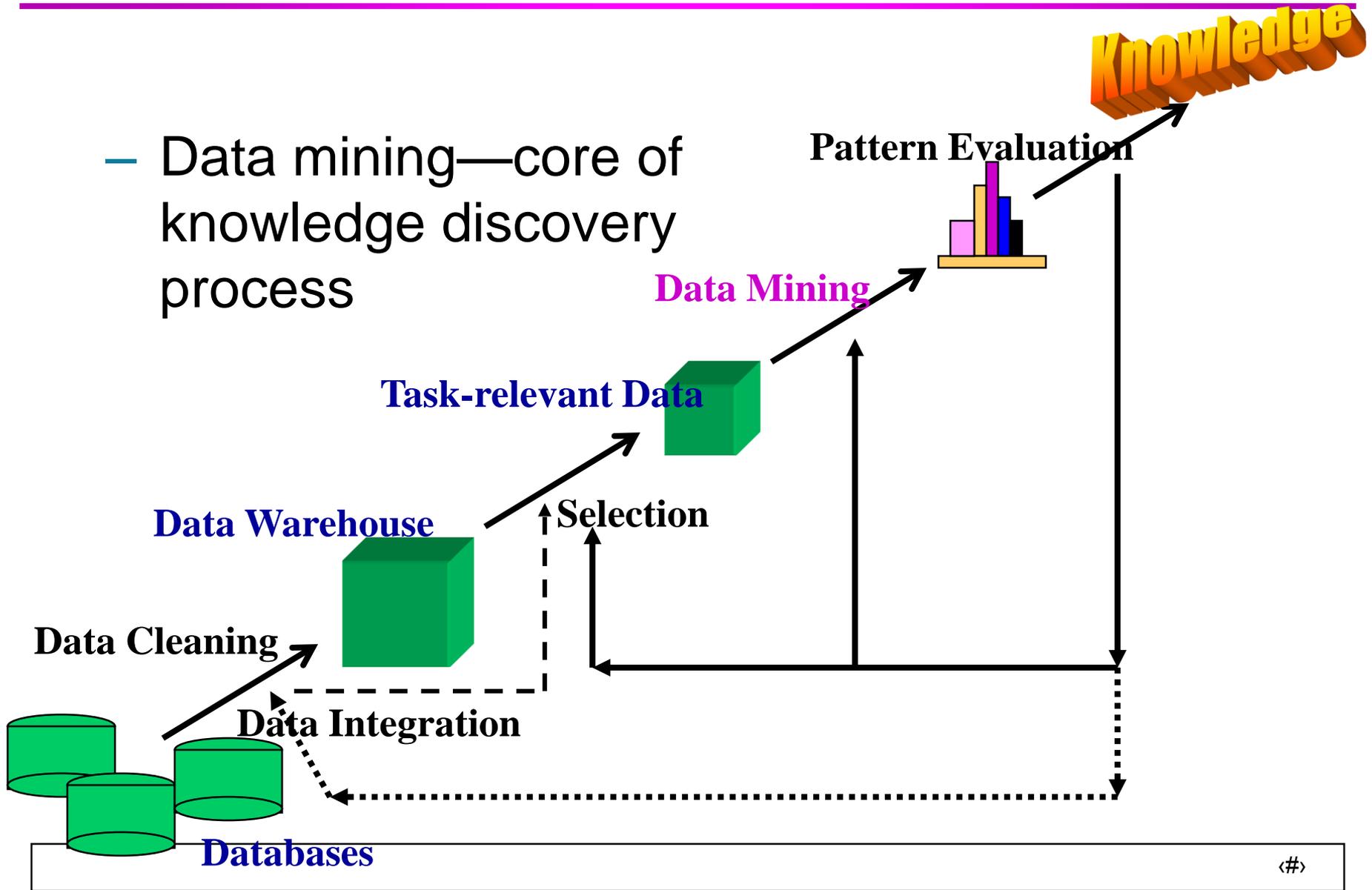
- Non-trivial extraction of implicit, previously unknown and potentially useful information from data



**Knowledge Discovery in Databases (KDD)**

# Data Mining: A KDD Process

- Data mining—core of knowledge discovery process



# Steps of a KDD Process

---

- Learning the application domain
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- **Data cleaning** and preprocessing: (may take 60% of effort!)
- **Data reduction and transformation**
  - Find useful features, dimensionality/variable reduction, invariant representation.
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering.
- Choosing the mining algorithm(s)
- **Data mining**: search for patterns of interest
- **Pattern evaluation and knowledge presentation**
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

# Data Mining: On What Kinds of Data?

---

- Relational database
- Data warehouse
- Transactional database
- Advanced database and information repository
  - Object-relational database
  - Spatial and temporal data
  - Time-series data
  - Stream data
  - Multimedia database
  - Heterogeneous and legacy database
  - Text databases & WWW

# 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’Rourke, O’Reilly... in Boston area)
- Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

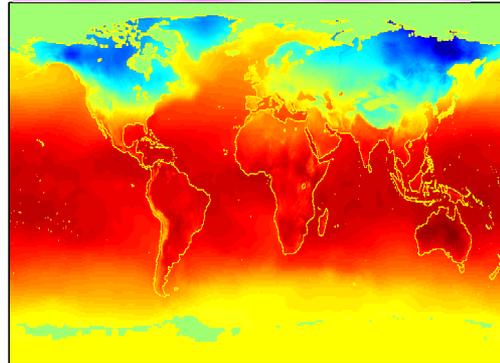
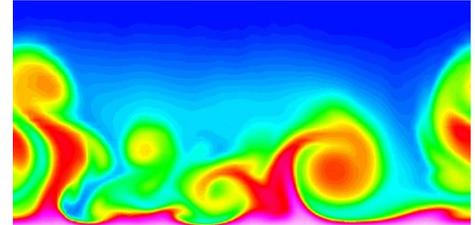
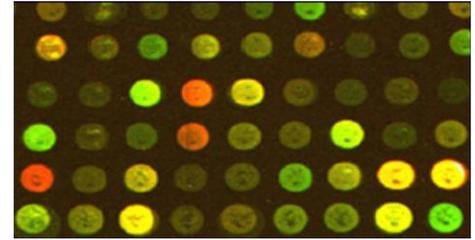
# Why Mine Data? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Walmart: ~20 million txn/day
  - Google: > 3 billion Web pages
  - Yahoo: ~10 GB Web data/hr
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)



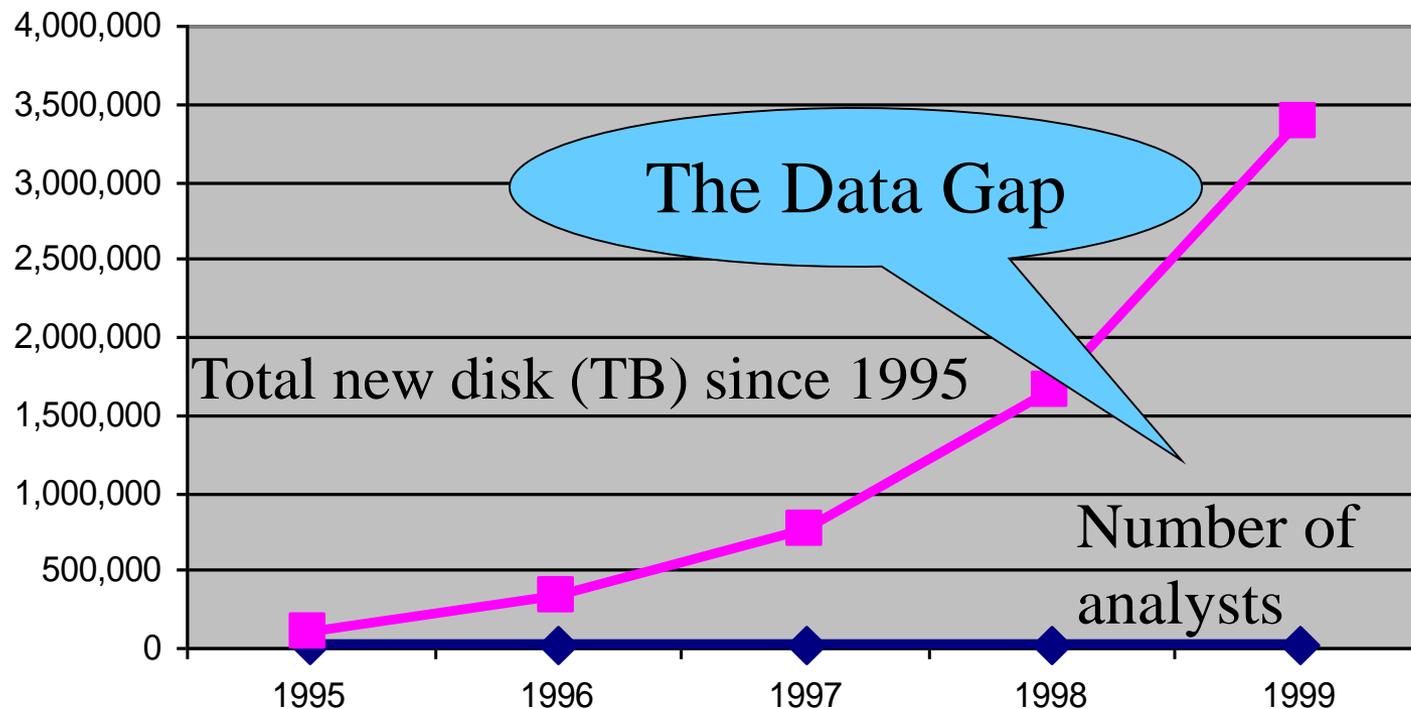
# Why Mine Data? Scientific Viewpoint

- 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
- Need techniques that can automatically analyze the data and form new hypotheses for further evaluation by scientists



# Mining Large Data Sets - Motivation

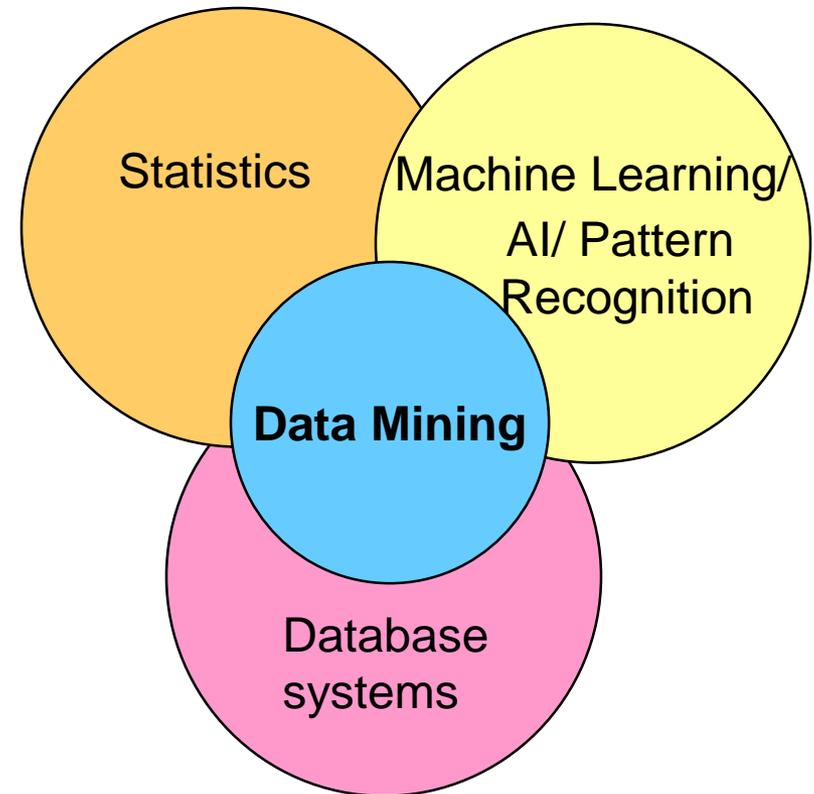
- There is often information “hidden” in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all



# 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

---

## □ Predictive Methods

- Use some variables to predict unknown or values of other variables.

## □ Descriptive 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]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Anomaly Detection [Predictive]

# Classification: Definition

---

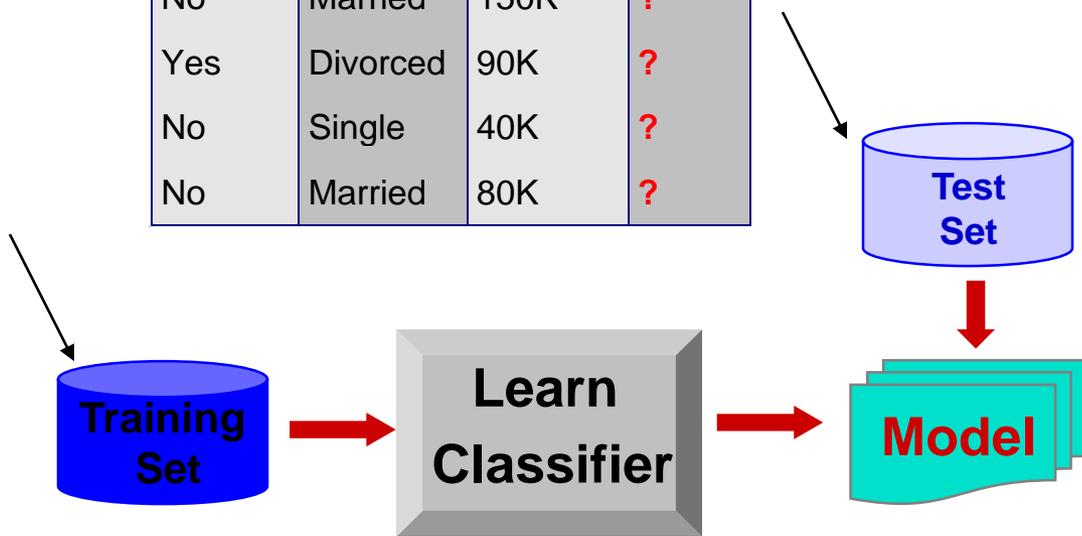
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Task: Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately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

<i>Tid</i>	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 1

---

## □ 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 2

---

## □ 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.

# Classification: Application 3

---

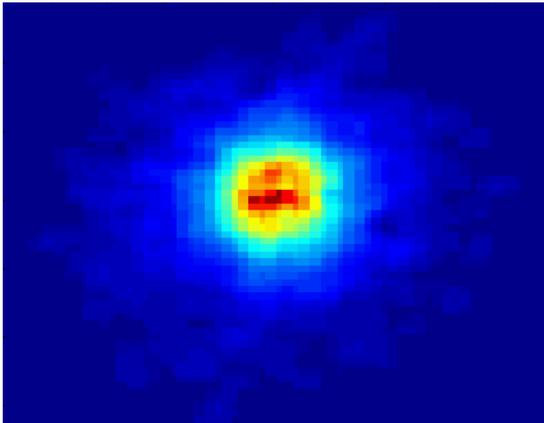
- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - ◆ Segment the image.
    - ◆ Measure image 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>

*Early*



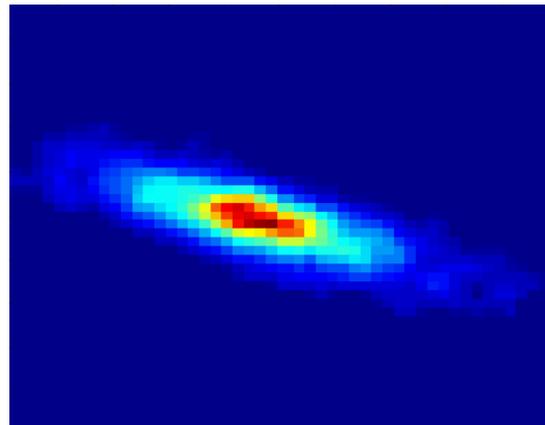
**Class:**

- Stages of Formation

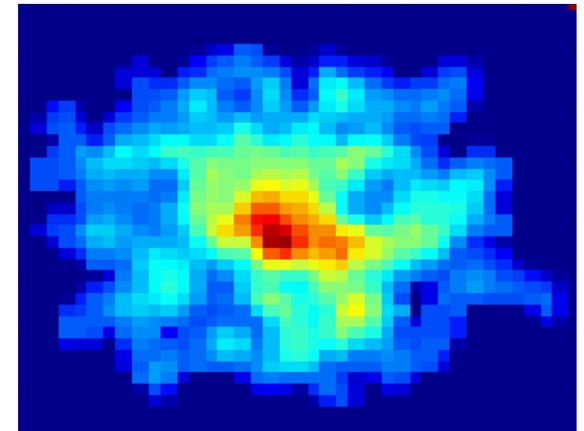
**Attributes:**

- 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 Definition

---

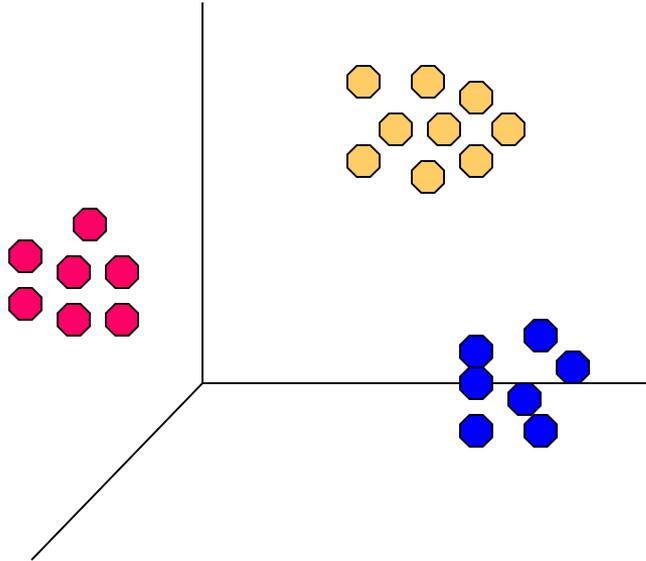
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance 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 1

---

- 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).

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

# 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.

<i>TID</i>	<i>Items</i>
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}**

**{Diaper, Milk} --> {Beer}**

# 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.
  - A classic rule --
    - ◆ If a customer buys diaper and milk, then he is very likely to buy beer.
    - ◆ So, don't be surprised if you find six-packs stacked next to diapers!

# 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.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

# What is Data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

## Attributes

<i>Tid</i>	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

## Objects

# Attribute Values

---

- Attribute values are numbers or symbols assigned to an attribute
- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - ◆ Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - ◆ Example: Attribute values for ID and age are integers
    - ◆ But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value ⟨#⟩

# Types of Attributes

---

- There are different types of attributes
  - **Nominal**
    - ◆ Examples: ID numbers, eye color, zip codes
  - **Ordinal**
    - ◆ Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - **Interval**
    - ◆ Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - **Ratio**
    - ◆ Examples: temperature in Kelvin, length, time, counts

# Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - **Distinctness:** = ≠
  - **Order:** < >
  - **Addition:** + -
  - **Multiplication:** \* /
  
- Nominal attribute (ID numbers, eye color, zip codes): **Distinctness**
- Ordinal attribute (rankings, grades): **Distinctness** & **Order**
- Interval attribute (calendar): **Distinctness**, **Order** & **Addition**
- Ratio attribute: all 4 properties

# Discrete and Continuous Attributes

---

## □ Discrete Attribute

- Has only a finite or countably infinite set of values
- Examples: zip codes, counts, or the set of words in a collection of documents
- Often represented as integer variables.
- Note: binary attributes are a special case of discrete attributes

## □ Continuous Attribute

- Has real numbers as attribute values
- Examples: temperature, height, or weight.
- Practically, real values can only be measured and represented using a finite number of digits.
- Continuous attributes are typically represented as floating-point variables.

# Types of data sets

---

## □ Record

- Data Matrix
- Document Data
- Transaction Data

## □ Graph

- World Wide Web
- Molecular Structures

## □ Ordered

- Spatial Data
- Temporal Data
- Sequential Data
- Genetic Sequence Data

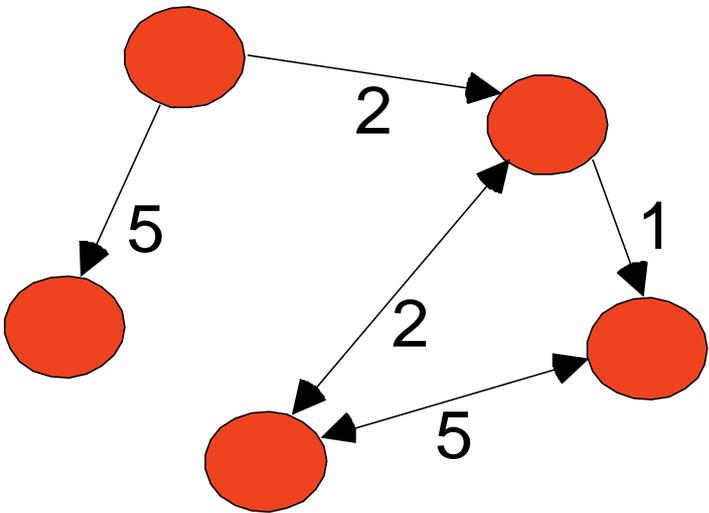
# Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

<i>Tid</i>	<i>Refund</i>	<i>Marital Status</i>	<i>Taxable Income</i>	<i>Cheat</i>
1	Yes	Single	125K	<b>No</b>
2	No	Married	100K	<b>No</b>
3	No	Single	70K	<b>No</b>
4	Yes	Married	120K	<b>No</b>
5	No	Divorced	95K	<b>Yes</b>
6	No	Married	60K	<b>No</b>
7	Yes	Divorced	220K	<b>No</b>
8	No	Single	85K	<b>Yes</b>
9	No	Married	75K	<b>No</b>
10	No	Single	90K	<b>Yes</b>

# Graph Data

## □ Examples: Generic graph and HTML Links



```
<a href="papers/papers.html#bbbb">  
Data Mining </a>
```

```
<li>
```

```
<a href="papers/papers.html#aaaa">  
Graph Partitioning </a>
```

```
<li>
```

```
<a href="papers/papers.html#aaaa">  
Parallel Solution of Sparse Linear System of Equations </a>
```

```
<li>
```

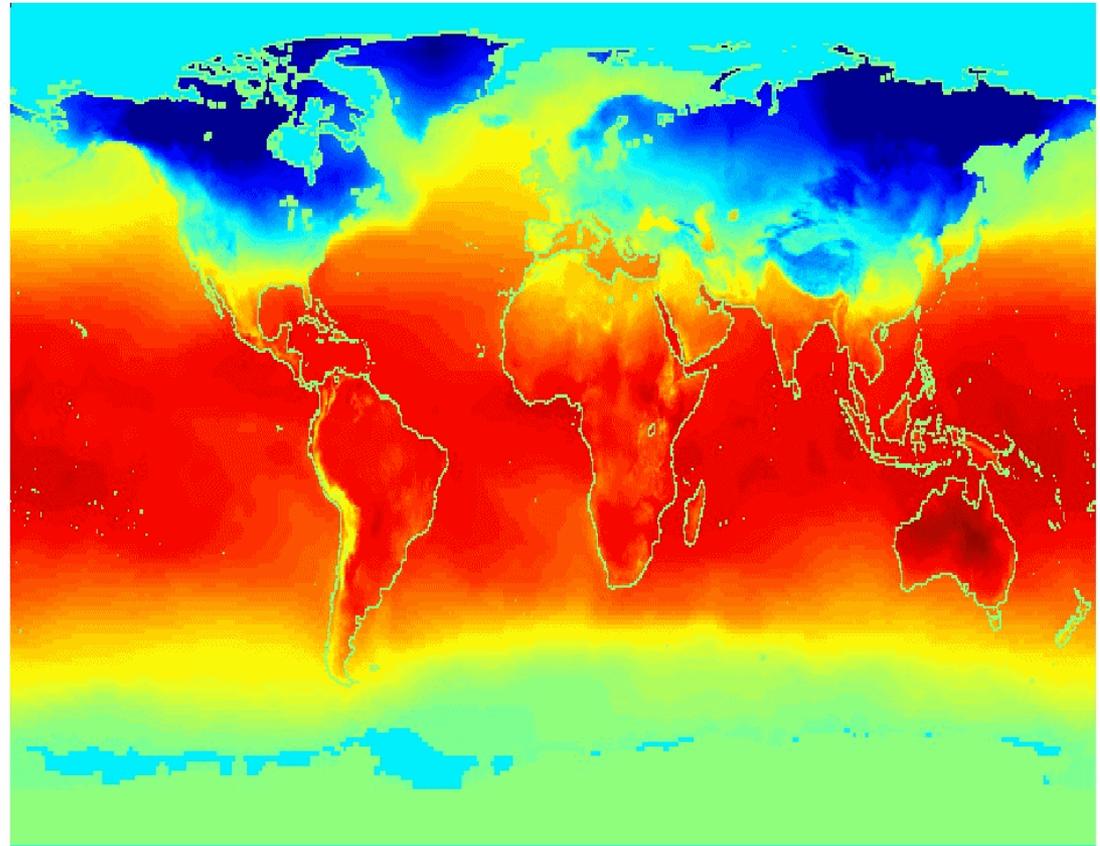
```
<a href="papers/papers.html#ffff">  
N-Body Computation and Dense Linear System Solvers
```

# Ordered Data

## □ Spatio-Temporal Data

**Average Monthly  
Temperature of  
land and ocean**

Jan



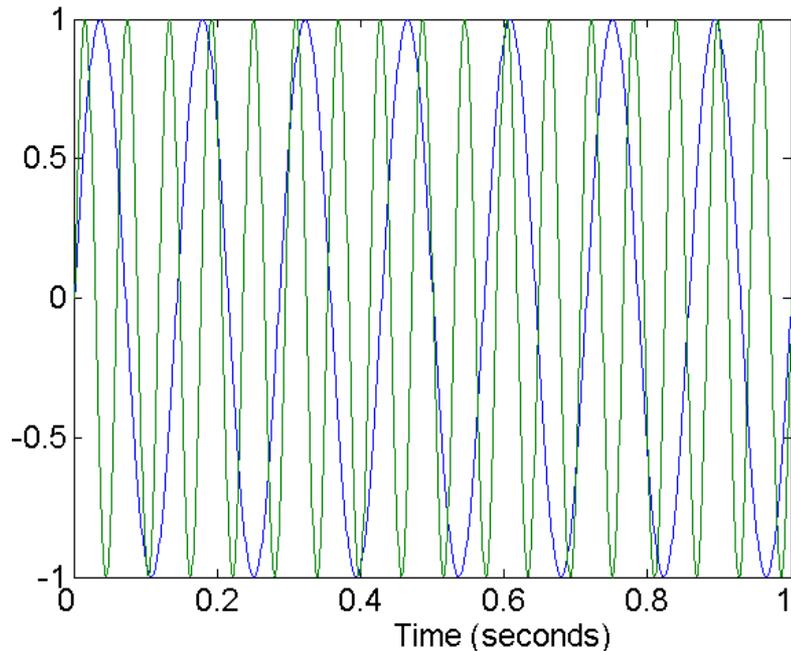
# Data Quality

---

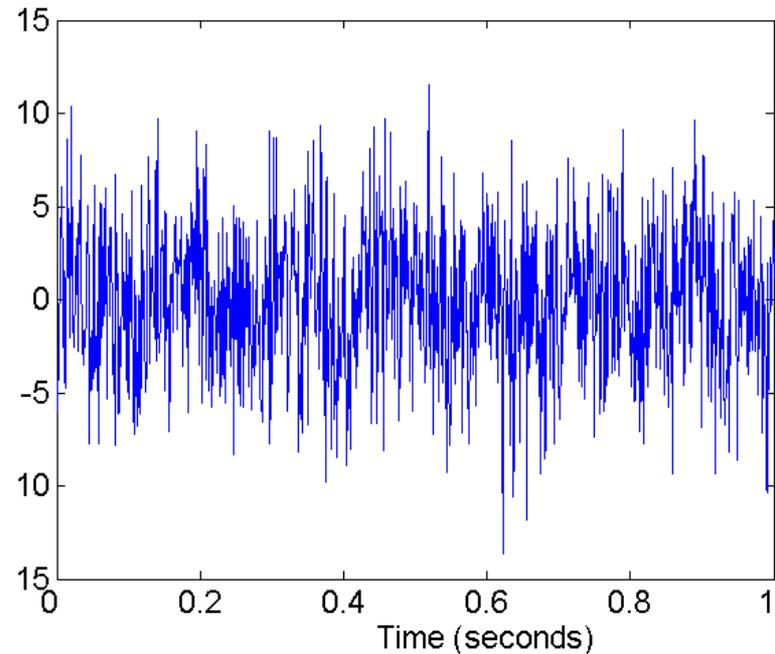
- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?
  
- Examples of data quality problems:
  - Noise and outliers
  - missing values
  - duplicate data

# Noise

- Noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and “snow” on



**Two Sine Waves**

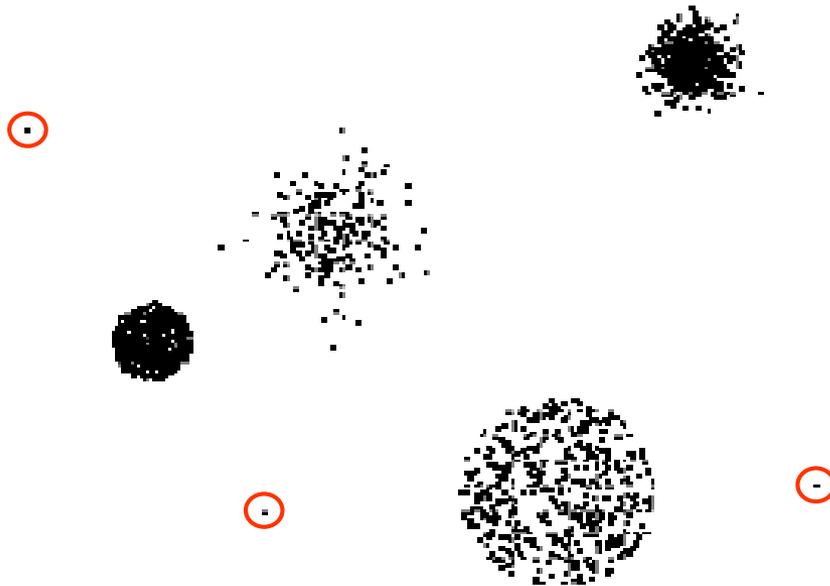


**Two Sine Waves + Noise**

# Outliers

---

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



# Missing Values

---

- Reasons for missing values
  - Information is not collected (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
  
- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

# Duplicate Data

---

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning
  - Process of dealing with duplicate data issues