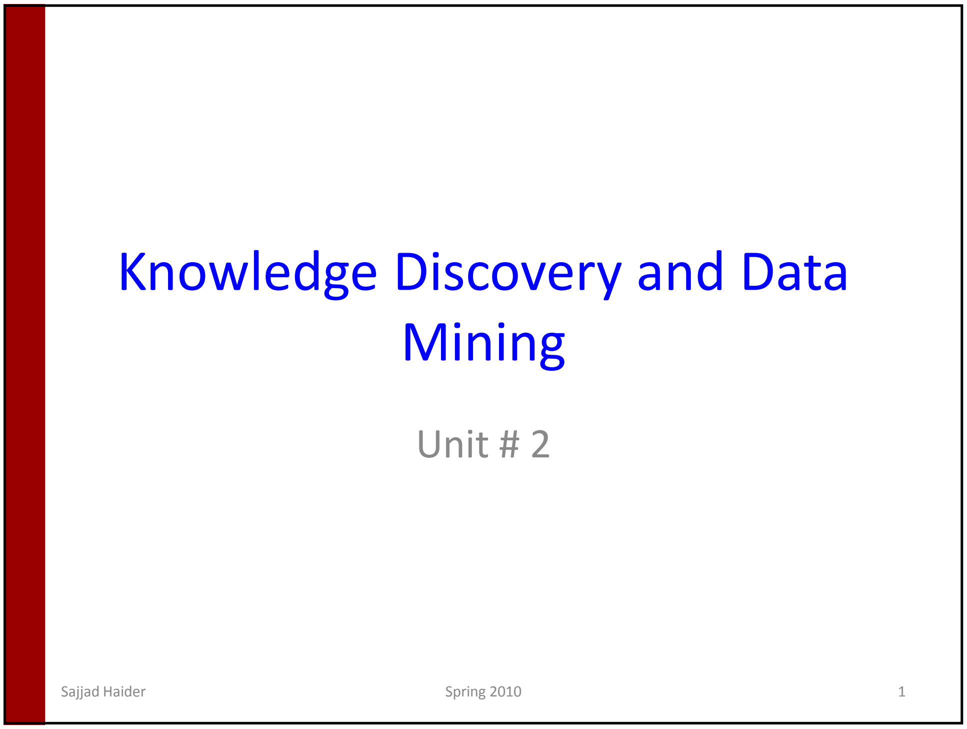
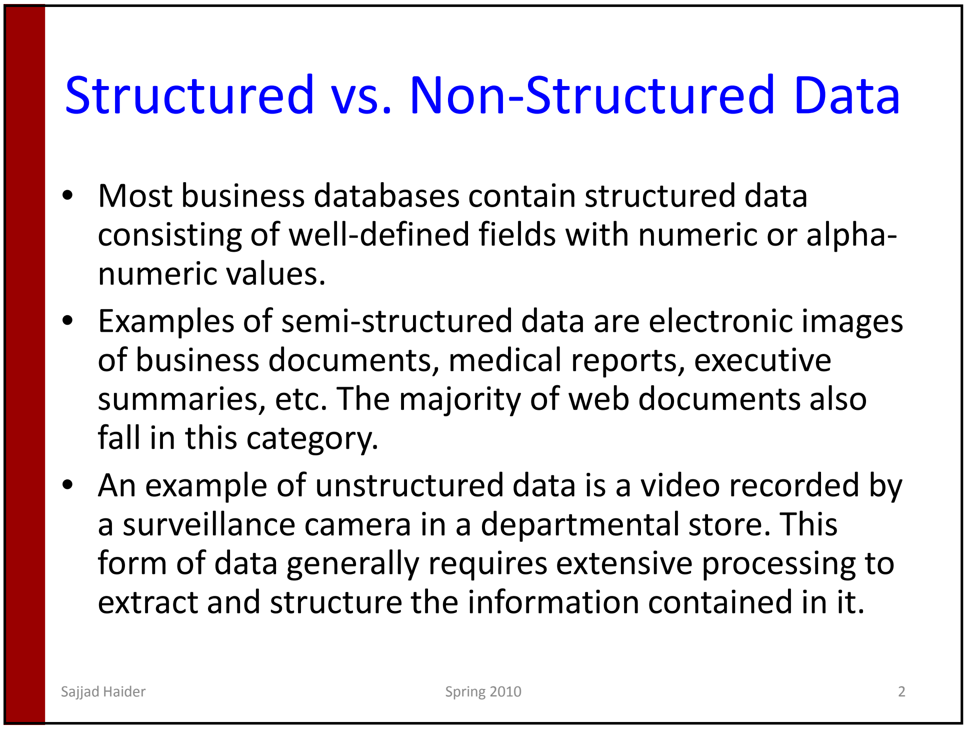


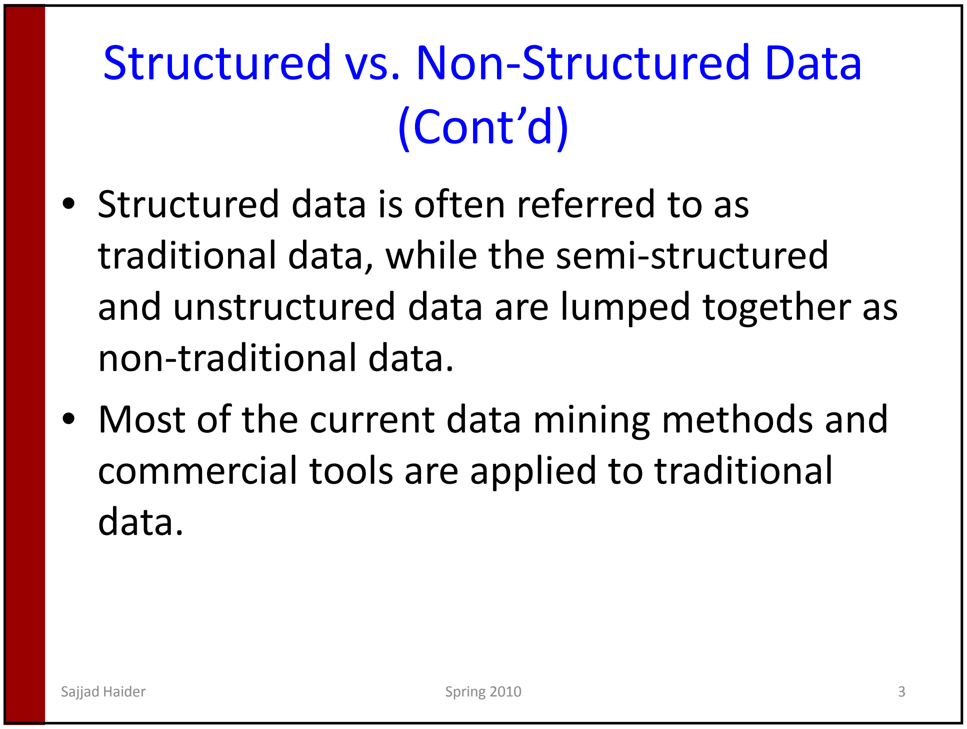
# Knowledge Discovery and Data Mining

Unit # 2



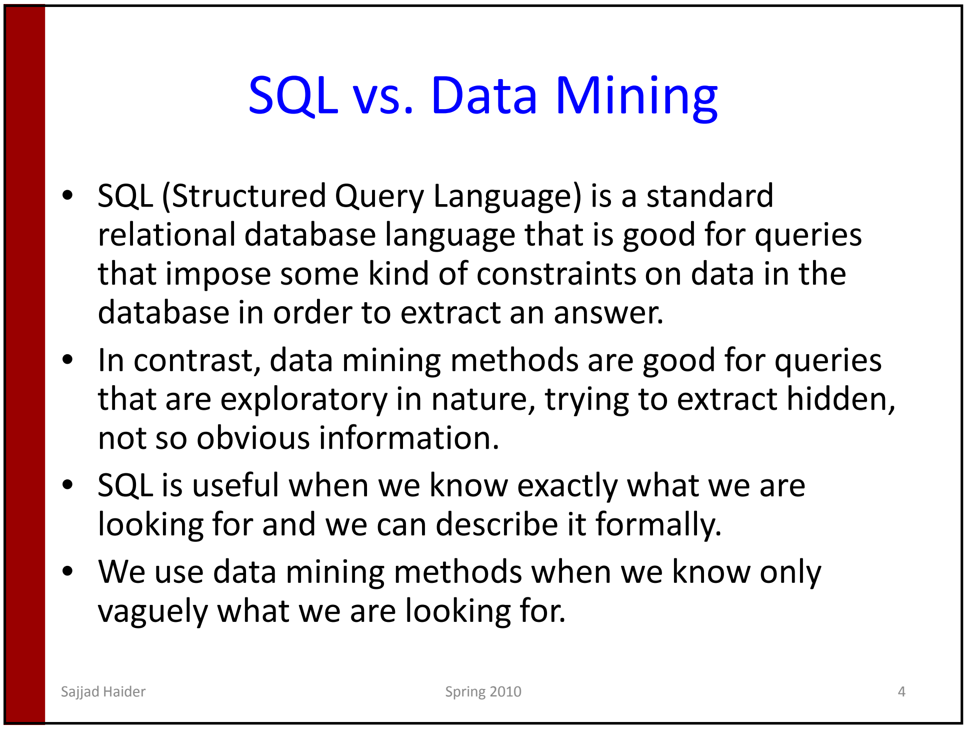
## Structured vs. Non-Structured Data

- Most business databases contain structured data consisting of well-defined fields with numeric or alpha-numeric values.
- Examples of semi-structured data are electronic images of business documents, medical reports, executive summaries, etc. The majority of web documents also fall in this category.
- An example of unstructured data is a video recorded by a surveillance camera in a departmental store. This form of data generally requires extensive processing to extract and structure the information contained in it.



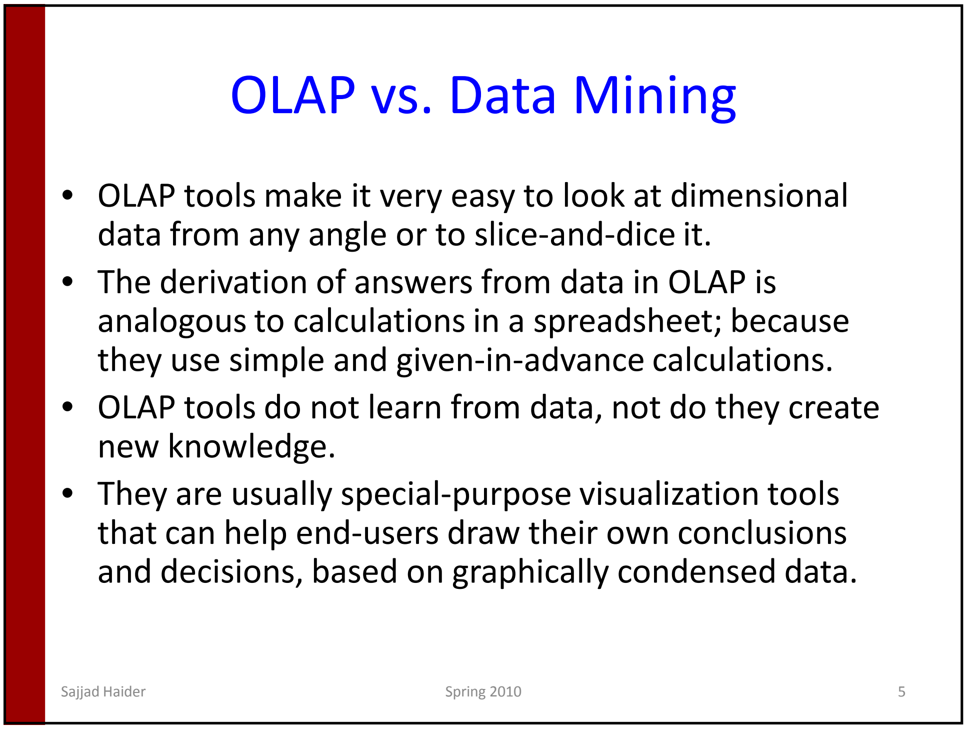
## Structured vs. Non-Structured Data (Cont’d)

- Structured data is often referred to as traditional data, while the semi-structured and unstructured data are lumped together as non-traditional data.
- Most of the current data mining methods and commercial tools are applied to traditional data.



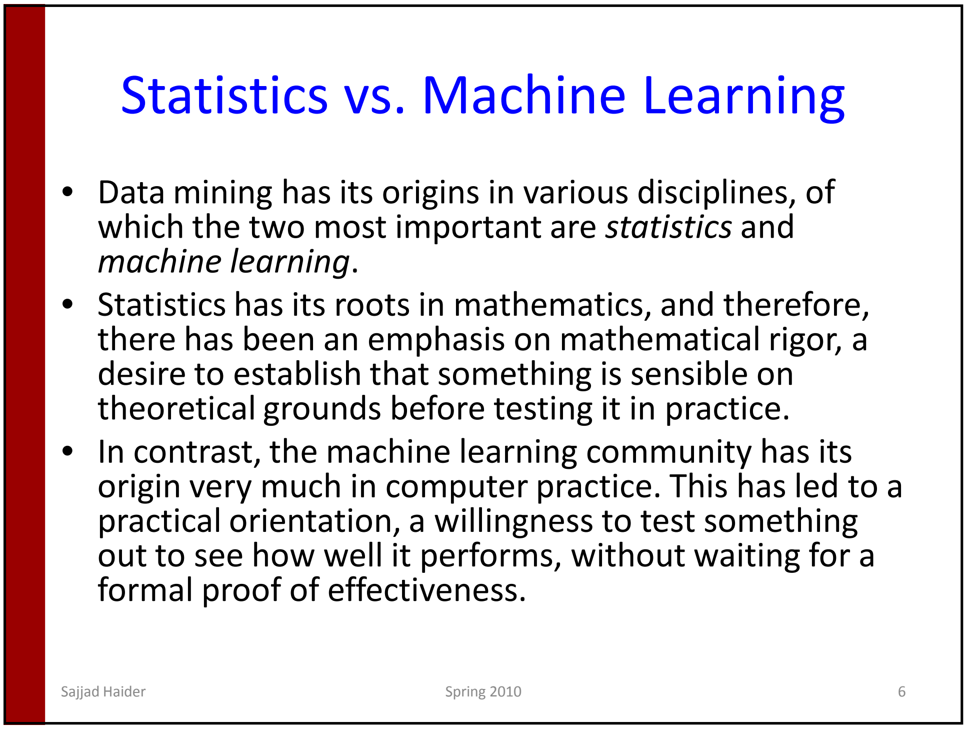
## SQL vs. Data Mining

- SQL (Structured Query Language) is a standard relational database language that is good for queries that impose some kind of constraints on data in the database in order to extract an answer.
- In contrast, data mining methods are good for queries that are exploratory in nature, trying to extract hidden, not so obvious information.
- SQL is useful when we know exactly what we are looking for and we can describe it formally.
- We use data mining methods when we know only vaguely what we are looking for.



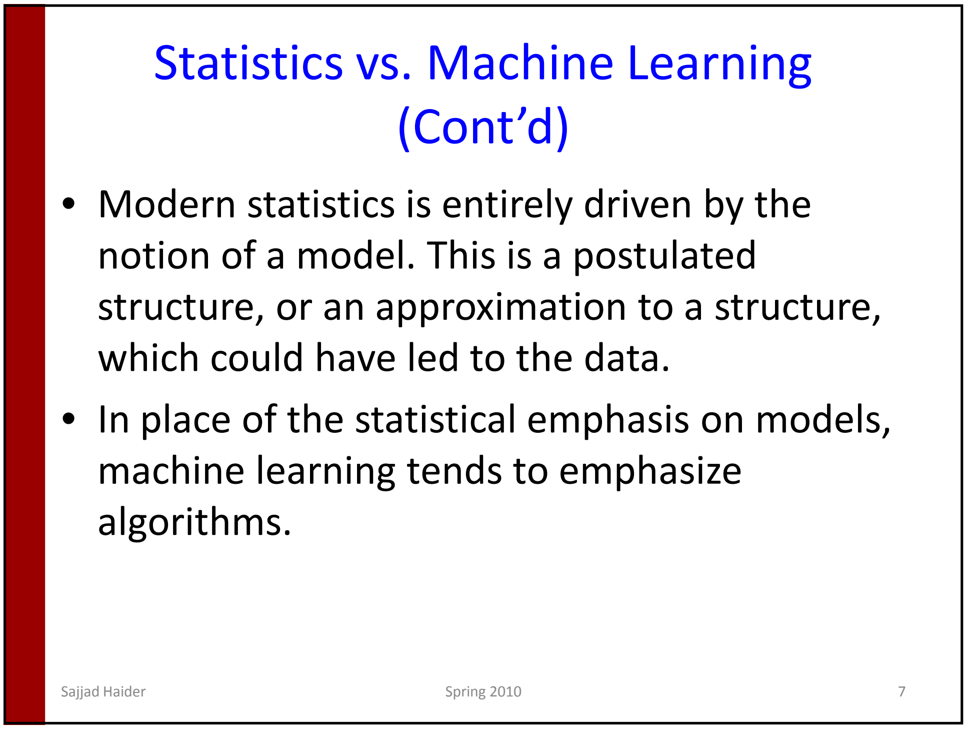
## OLAP vs. Data Mining

- OLAP tools make it very easy to look at dimensional data from any angle or to slice-and-dice it.
- The derivation of answers from data in OLAP is analogous to calculations in a spreadsheet; because they use simple and given-in-advance calculations.
- OLAP tools do not learn from data, not do they create new knowledge.
- They are usually special-purpose visualization tools that can help end-users draw their own conclusions and decisions, based on graphically condensed data.



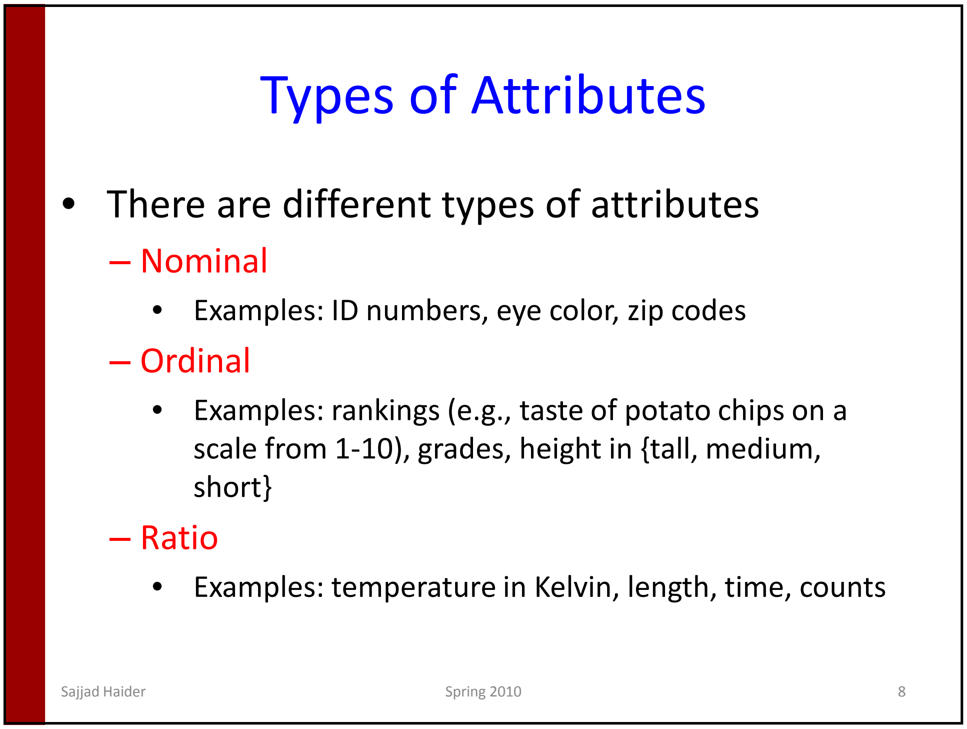
## Statistics vs. Machine Learning

- Data mining has its origins in various disciplines, of which the two most important are *statistics* and *machine learning*.
- Statistics has its roots in mathematics, and therefore, there has been an emphasis on mathematical rigor, a desire to establish that something is sensible on theoretical grounds before testing it in practice.
- In contrast, the machine learning community has its origin very much in computer practice. This has led to a practical orientation, a willingness to test something out to see how well it performs, without waiting for a formal proof of effectiveness.



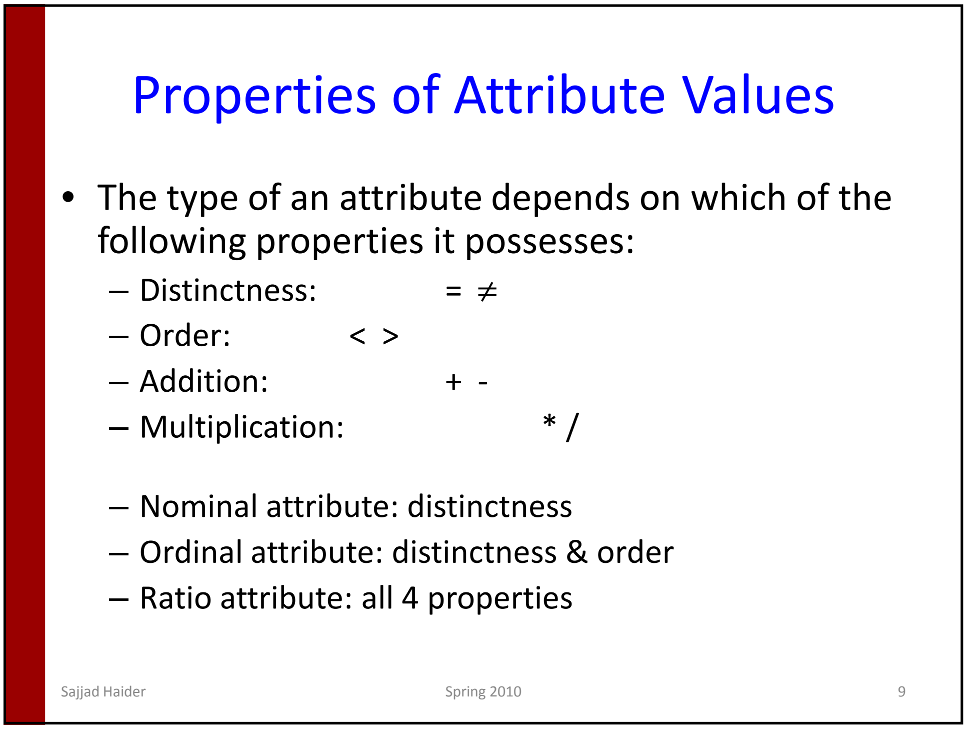
## Statistics vs. Machine Learning (Cont’d)

- Modern statistics is entirely driven by the notion of a model. This is a postulated structure, or an approximation to a structure, which could have led to the data.
- In place of the statistical emphasis on models, machine learning tends to emphasize algorithms.



## Types of Attributes

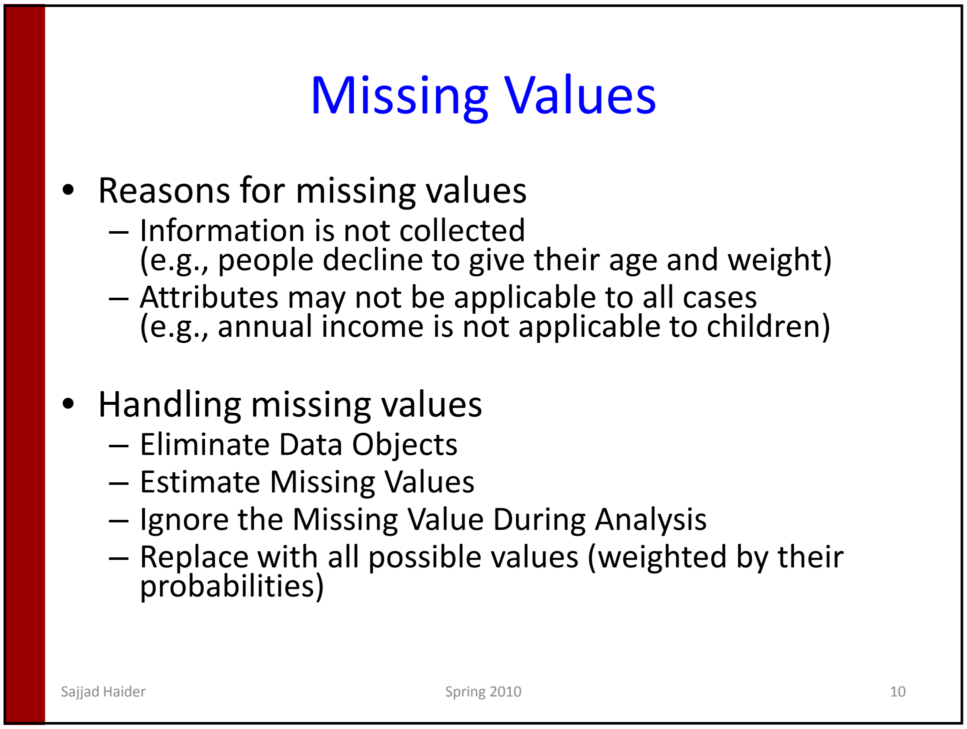
- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts



## Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Addition: + -
  - Multiplication: * /

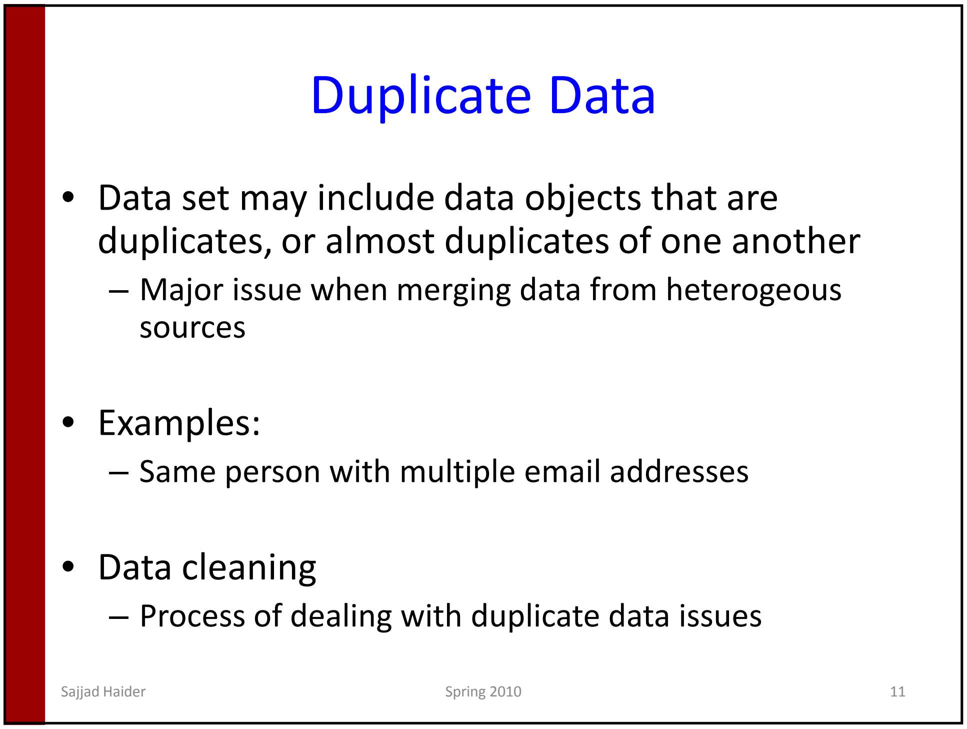
  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Ratio attribute: all 4 properties



## Missing Values

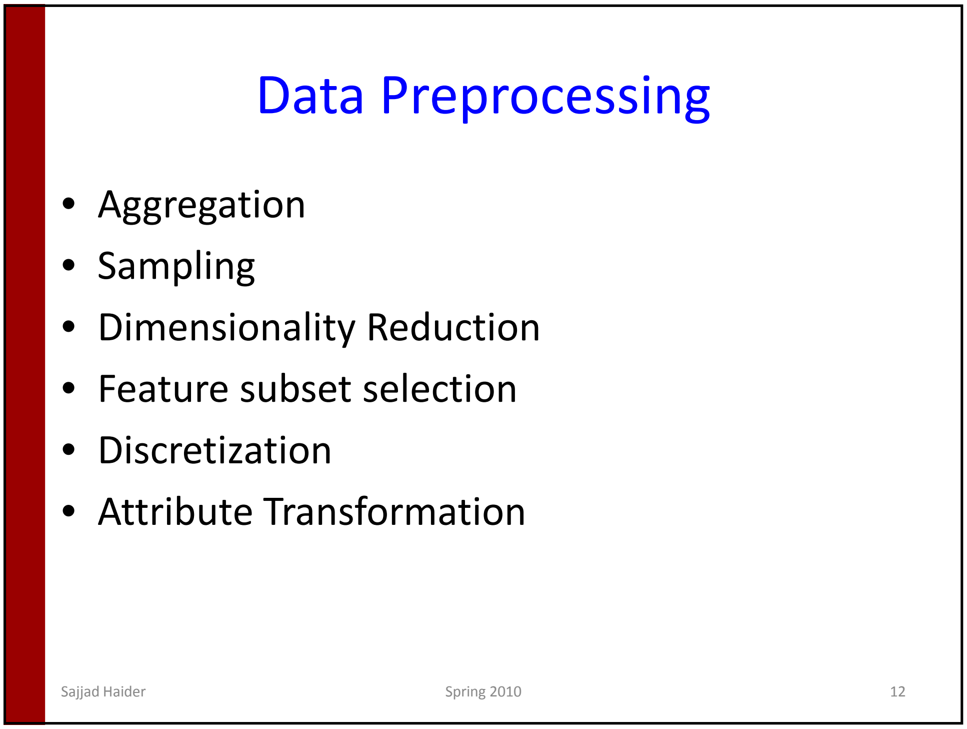
- Reasons for missing values
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)



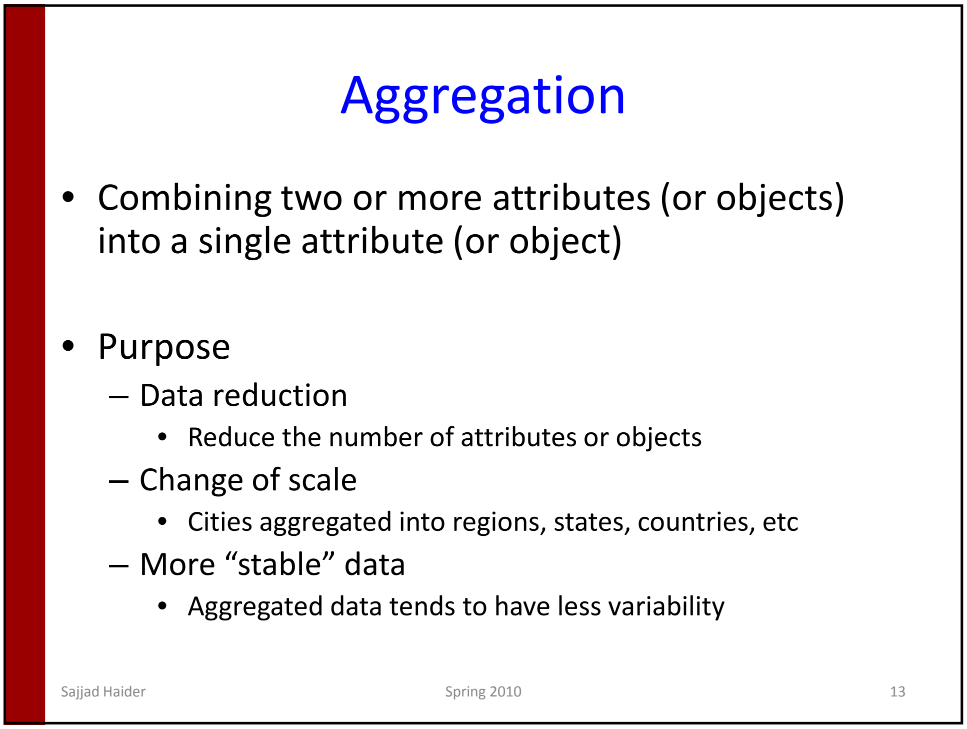
## Duplicate Data

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



## Data Preprocessing

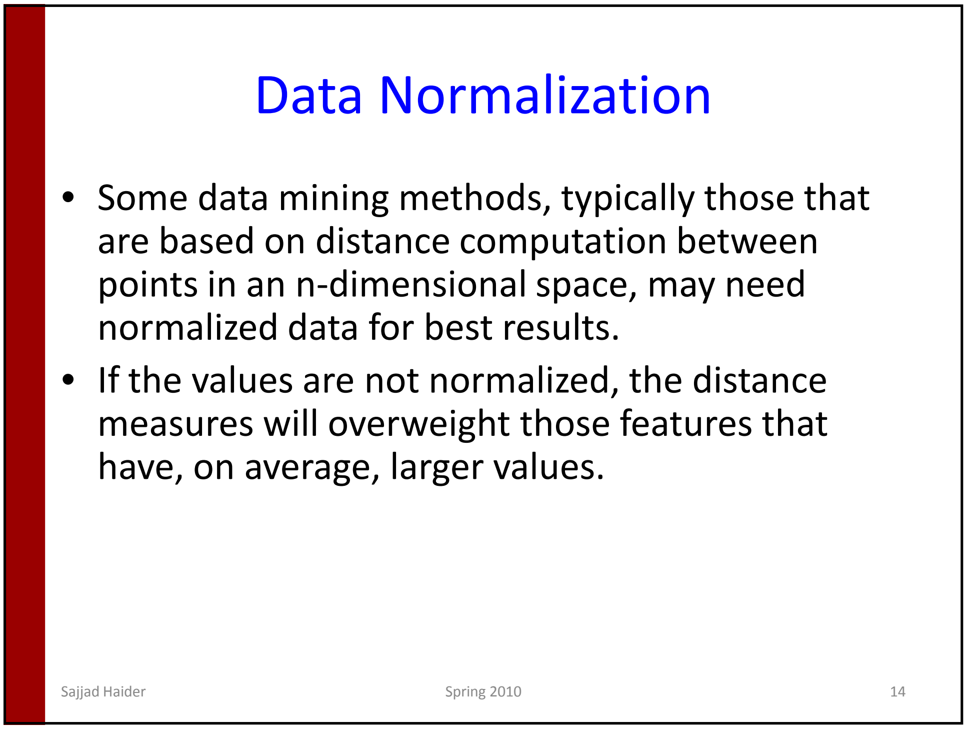
- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Discretization
- Attribute Transformation



## Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)

- Purpose
  – Data reduction
    - Reduce the number of attributes or objects
  – Change of scale
    - Cities aggregated into regions, states, countries, etc
  – More “stable” data
    - Aggregated data tends to have less variability



## Data Normalization

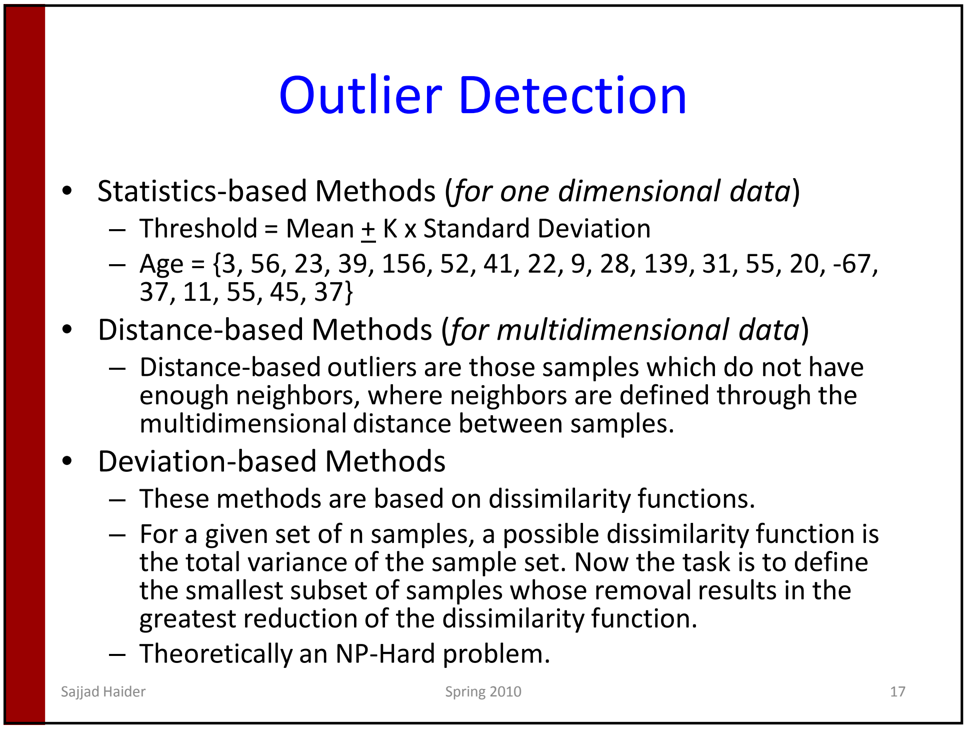
- Some data mining methods, typically those that are based on distance computation between points in an n-dimensional space, may need normalized data for best results.
- If the values are not normalized, the distance measures will overweight those features that have, on average, larger values.

# Normalization Techniques

- Decimal Scaling
  - $v'(i) = v(i) / 10^k$
  - For the smallest k such that max $|v'(i)| < 1$.
- Min-Max Normalization
  - $v'(i) = [v(i) – \min(v(i))]/[\max(v(i)) – \min(v(i))]$
- Standard Deviation Normalization
  - $v'(i) = [v(i) – \text{mean}(v)]/\text{sd}(v)$

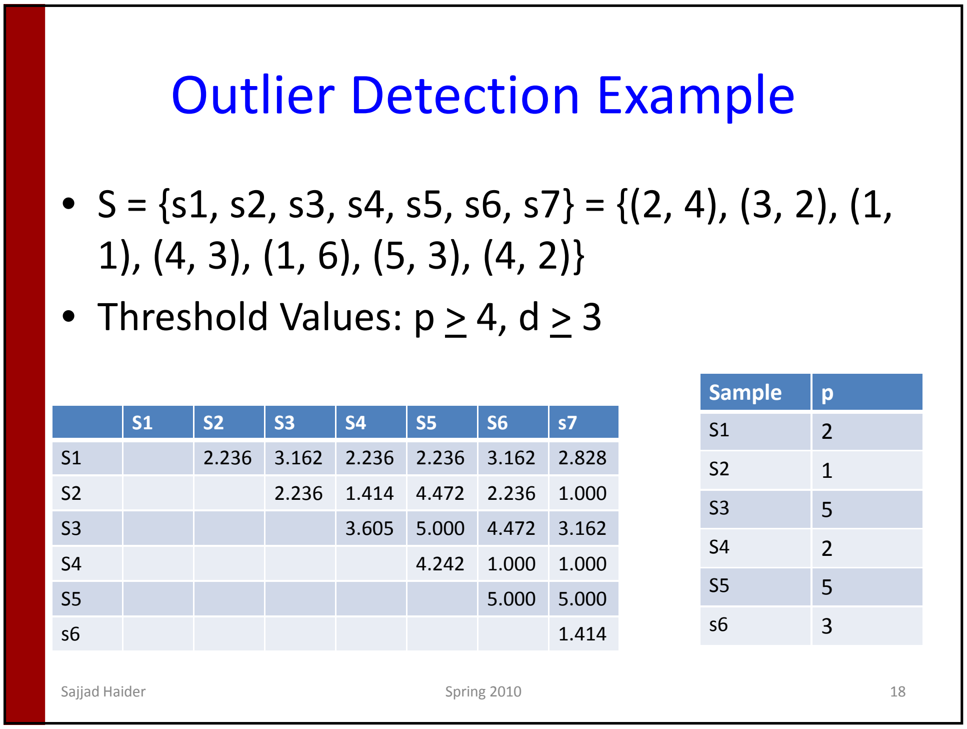
# Normalization Example

- Given one-dimensional data set X = {-5.0, 23.0, 17.6, 7.23, 1.11}, normalize the data set using
  - Decimal scaling on interval [-1, 1].
  - Min-max normalization on interval [0, 1].
  - Min-max normalization on interval [-1, 1].
  - Standard deviation normalization.



# Outlier Detection

- Statistics-based Methods (*for one dimensional data*)
  - Threshold = Mean ± K x Standard Deviation
  - Age = {3, 56, 23, 39, 156, 52, 41, 22, 9, 28, 139, 31, 55, 20, -67, 37, 11, 55, 45, 37}
- Distance-based Methods (*for multidimensional data*)
  - Distance-based outliers are those samples which do not have enough neighbors, where neighbors are defined through the multidimensional distance between samples.
- Deviation-based Methods
  - These methods are based on dissimilarity functions.
  - For a given set of n samples, a possible dissimilarity function is the total variance of the sample set. Now the task is to define the smallest subset of samples whose removal results in the greatest reduction of the dissimilarity function.
  - Theoretically an NP-Hard problem.

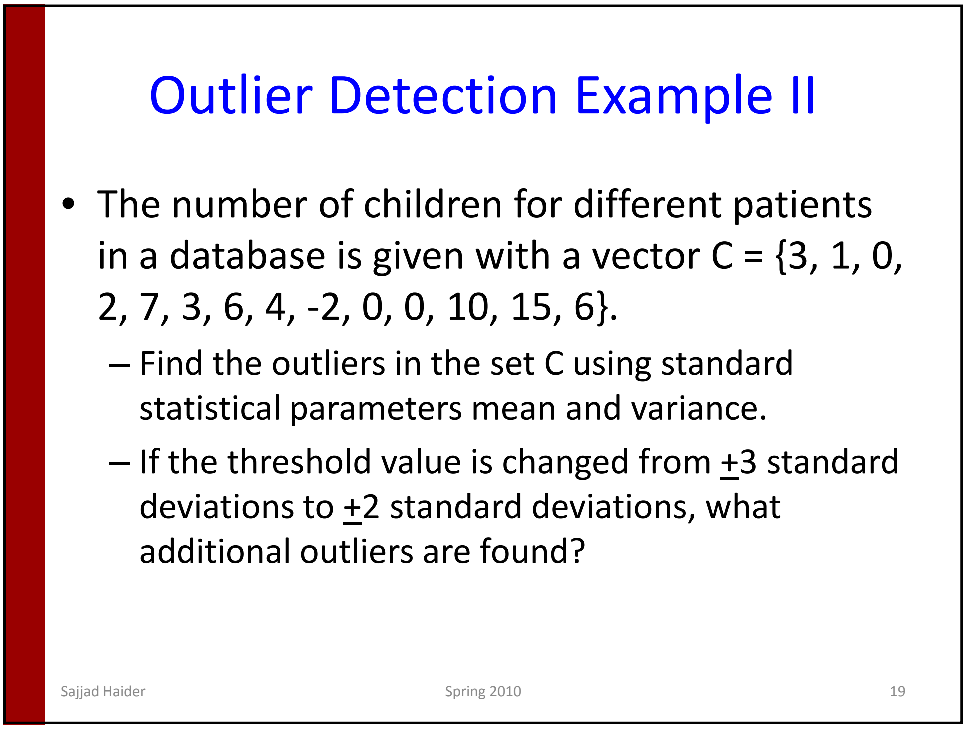


# Outlier Detection Example

- S = {s1, s2, s3, s4, s5, s6, s7} = {(2, 4), (3, 2), (1, 1), (4, 3), (1, 6), (5, 3), (4, 2)}
- Threshold Values: p ≥ 4, d ≥ 3

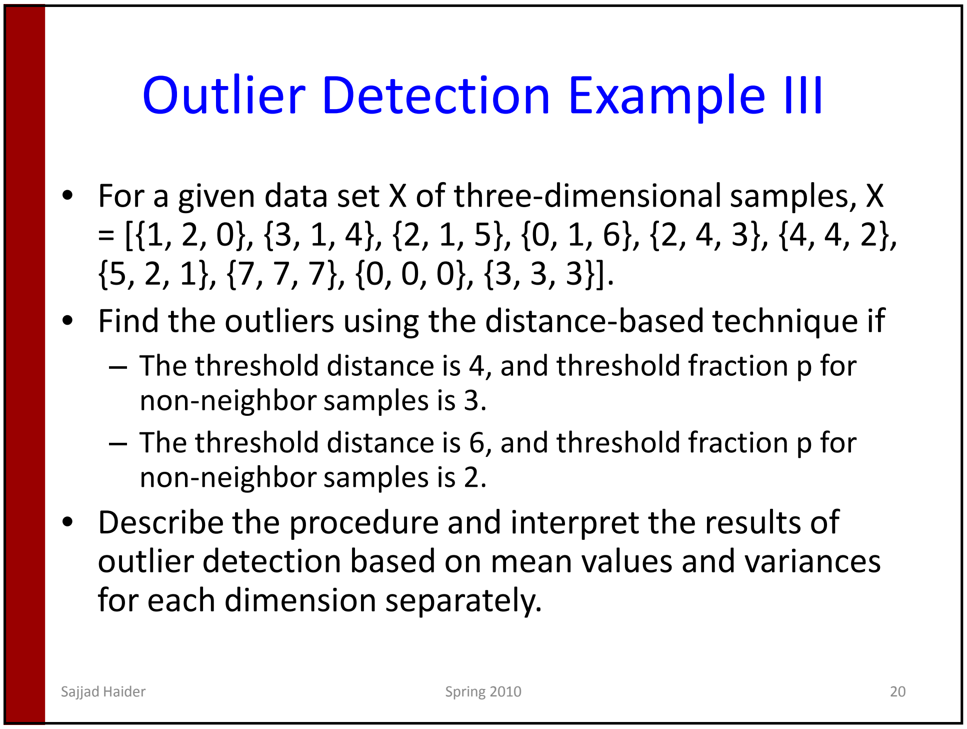
| | S1 | S2 | S3 | S4 | S5 | S6 | s7 |
|---|---|---|---|---|---|---|---|
| S1 | | 2.236 | 3.162 | 2.236 | 2.236 | 3.162 | 2.828 |
| S2 | | | 2.236 | 1.414 | 4.472 | 2.236 | 1.000 |
| S3 | | | | 3.605 | 5.000 | 4.472 | 3.162 |
| S4 | | | | | 4.242 | 1.000 | 1.000 |
| S5 | | | | | | 5.000 | 5.000 |
| s6 | | | | | | | 1.414 |

| Sample | p |
|---|---|
| S1 | 2 |
| S2 | 1 |
| S3 | 5 |
| S4 | 2 |
| S5 | 5 |
| s6 | 3 |



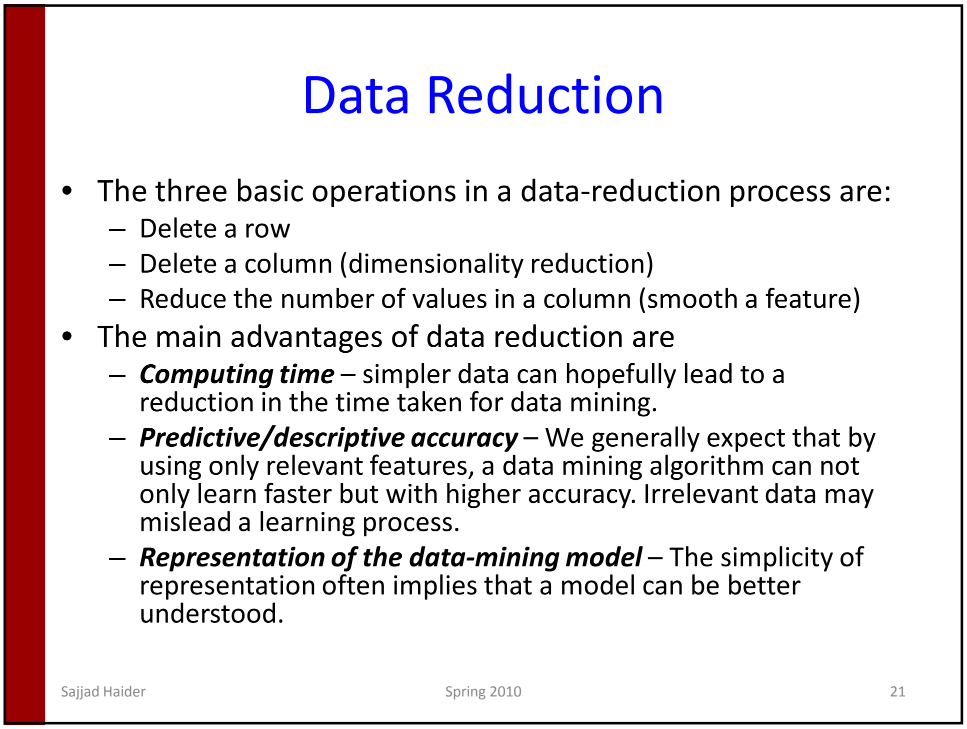
## Outlier Detection Example II

- The number of children for different patients in a database is given with a vector C = {3, 1, 0, 2, 7, 3, 6, 4, -2, 0, 0, 10, 15, 6}.
  – Find the outliers in the set C using standard statistical parameters mean and variance.
  – If the threshold value is changed from ±3 standard deviations to ±2 standard deviations, what additional outliers are found?



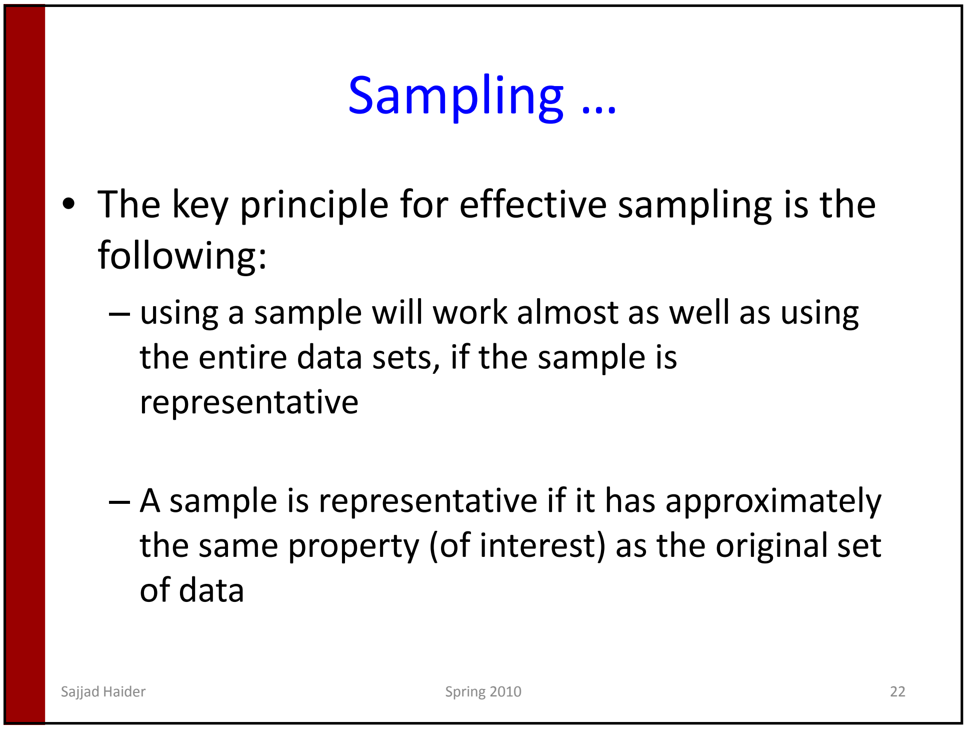
## Outlier Detection Example III

- For a given data set X of three-dimensional samples, X = [{1, 2, 0}, {3, 1, 4}, {2, 1, 5}, {0, 1, 6}, {2, 4, 3}, {4, 4, 2}, {5, 2, 1}, {7, 7, 7}, {0, 0, 0}, {3, 3, 3}].
- Find the outliers using the distance-based technique if
  – The threshold distance is 4, and threshold fraction p for non-neighbor samples is 3.
  – The threshold distance is 6, and threshold fraction p for non-neighbor samples is 2.
- Describe the procedure and interpret the results of outlier detection based on mean values and variances for each dimension separately.



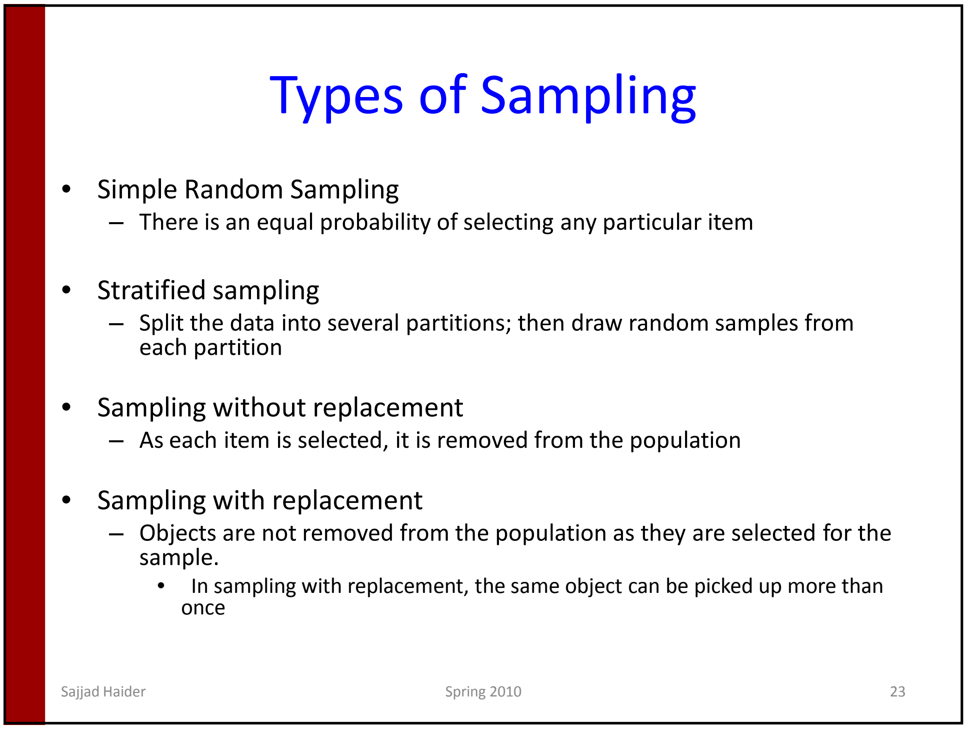
# Data Reduction

- The three basic operations in a data-reduction process are:
  - Delete a row
  - Delete a column (dimensionality reduction)
  - Reduce the number of values in a column (smooth a feature)
- The main advantages of data reduction are
  - ***Computing time*** – simpler data can hopefully lead to a reduction in the time taken for data mining.
  - ***Predictive/descriptive accuracy*** – We generally expect that by using only relevant features, a data mining algorithm can not only learn faster but with higher accuracy. Irrelevant data may mislead a learning process.
  - ***Representation of the data-mining model*** – The simplicity of representation often implies that a model can be better understood.



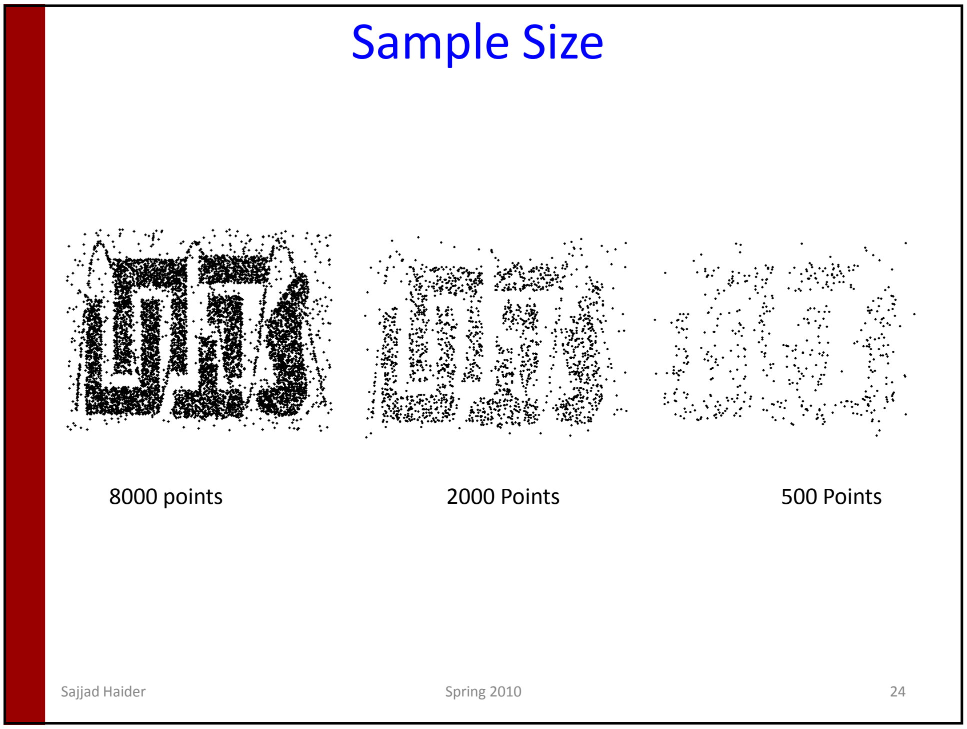
# Sampling …

- The key principle for effective sampling is the following:
  - using a sample will work almost as well as using the entire data sets, if the sample is representative
  - A sample is representative if it has approximately the same property (of interest) as the original set of data



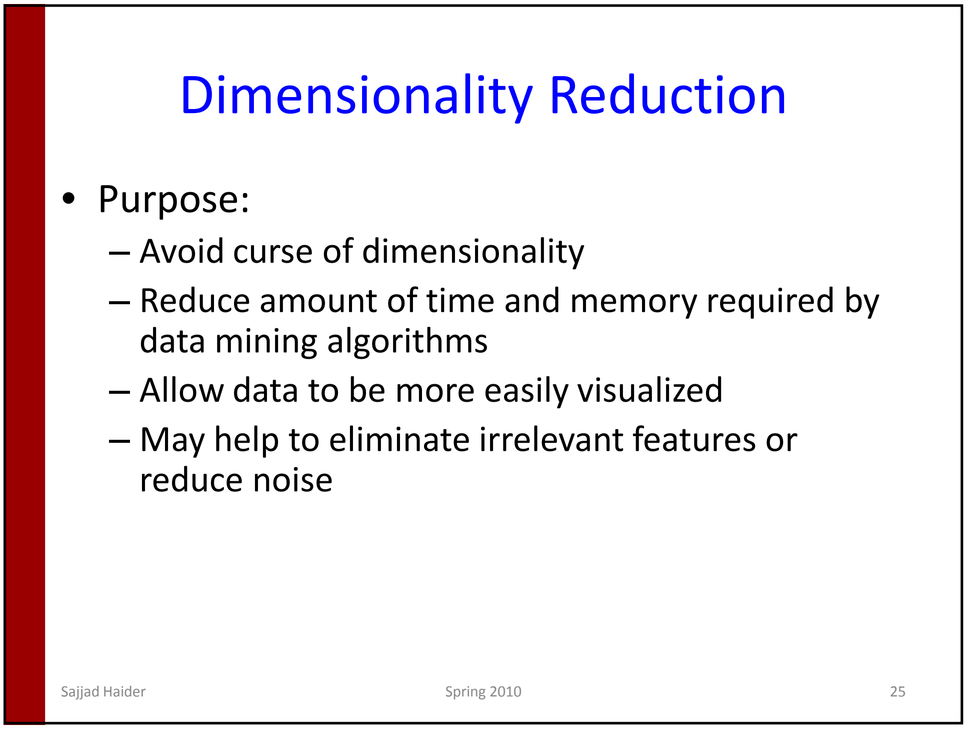
# Types of Sampling

- Simple Random Sampling
  - There is an equal probability of selecting any particular item
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition
- Sampling without replacement
  - As each item is selected, it is removed from the population
- Sampling with replacement
  - Objects are not removed from the population as they are selected for the sample.
    - In sampling with replacement, the same object can be picked up more than once



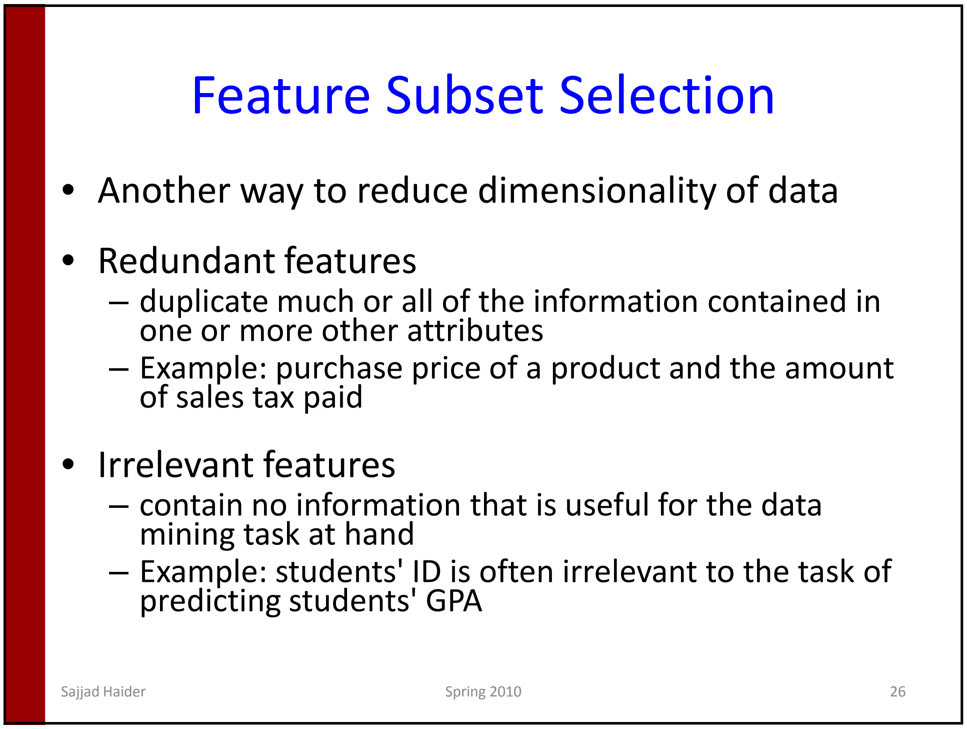
# Sample Size

8000 points 2000 Points 500 Points



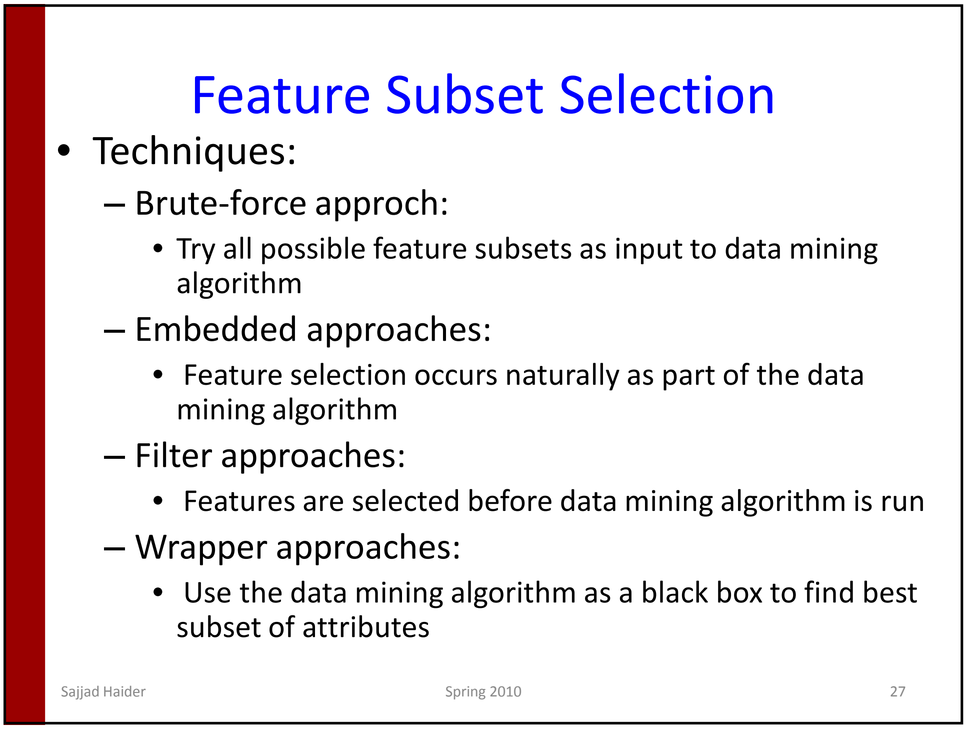
## Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise



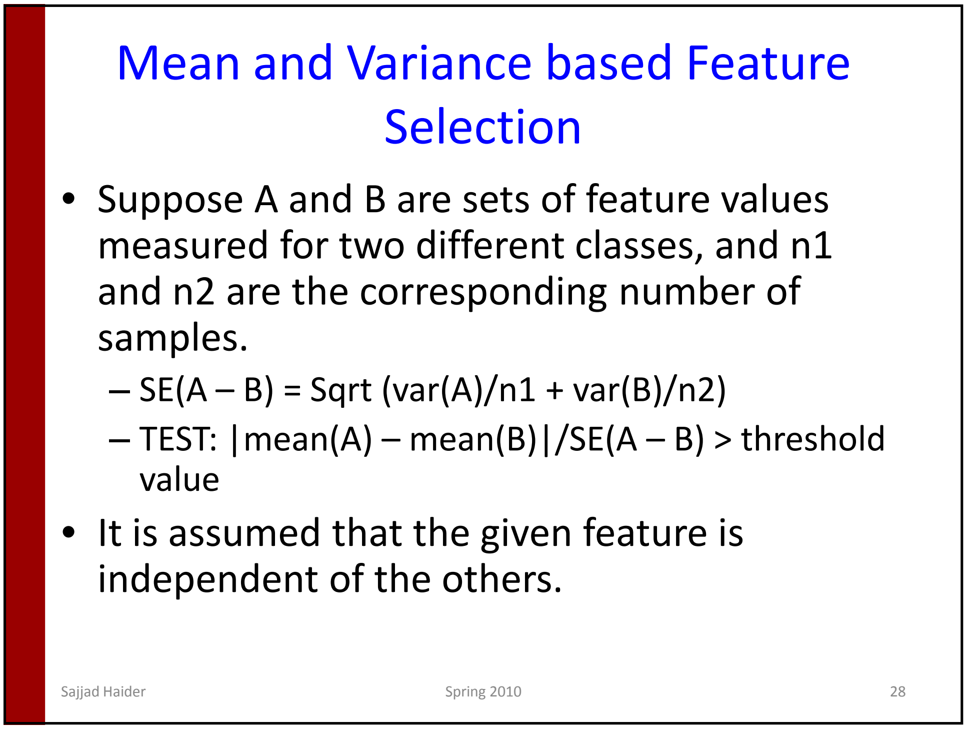
## Feature Subset Selection

- Another way to reduce dimensionality of data
- Redundant features
  - duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA



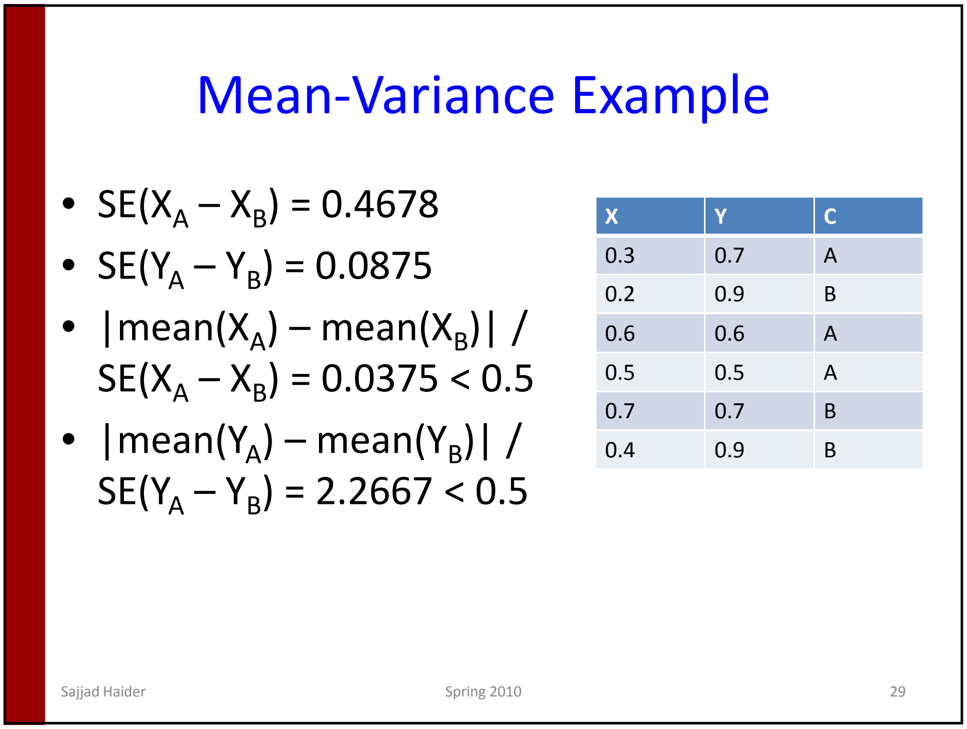
## Feature Subset Selection

- Techniques:
  - Brute-force approch:
    - Try all possible feature subsets as input to data mining algorithm
  - Embedded approaches:
    - Feature selection occurs naturally as part of the data mining algorithm
  - Filter approaches:
    - Features are selected before data mining algorithm is run
  - Wrapper approaches:
    - Use the data mining algorithm as a black box to find best subset of attributes



## Mean and Variance based Feature Selection

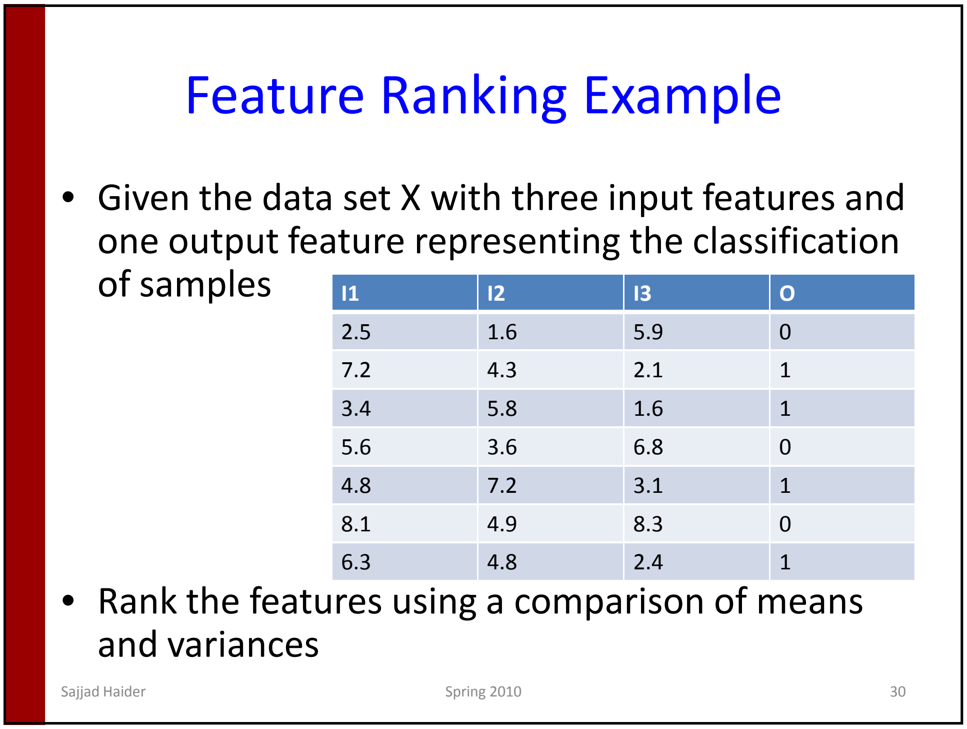
- Suppose A and B are sets of feature values measured for two different classes, and n1 and n2 are the corresponding number of samples.
  - SE(A – B) = Sqrt (var(A)/n1 + var(B)/n2)
  - TEST: |mean(A) – mean(B)|/SE(A – B) > threshold value
- It is assumed that the given feature is independent of the others.



## Mean-Variance Example

- $SE(X_A – X_B) = 0.4678$
- $SE(Y_A – Y_B) = 0.0875$
- $|mean(X_A) – mean(X_B)| / SE(X_A – X_B) = 0.0375 < 0.5$
- $|mean(Y_A) – mean(Y_B)| / SE(Y_A – Y_B) = 2.2667 < 0.5$

| X | Y | C |
|---|---|---|
| 0.3 | 0.7 | A |
| 0.2 | 0.9 | B |
| 0.6 | 0.6 | A |
| 0.5 | 0.5 | A |
| 0.7 | 0.7 | B |
| 0.4 | 0.9 | B |



## Feature Ranking Example

- Given the data set X with three input features and one output feature representing the classification of samples

| I1 | I2 | I3 | O |
|---|---|---|---|
| 2.5 | 1.6 | 5.9 | 0 |
| 7.2 | 4.3 | 2.1 | 1 |
| 3.4 | 5.8 | 1.6 | 1 |
| 5.6 | 3.6 | 6.8 | 0 |
| 4.8 | 7.2 | 3.1 | 1 |
| 8.1 | 4.9 | 8.3 | 0 |
| 6.3 | 4.8 | 2.4 | 1 |

- Rank the features using a comparison of means and variances