Data Mining

— Chapter 3 Data Preprocessing—

Adapted from Materials by

Jiawei Han, Micheline Kamber, and Jian Pei

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Why Preprocess the Data?

- Measures for data quality
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?

Major Tasks in Data Preprocessing

- **Data cleaning**
  - Fill in missing values
  - Smooth noisy data
  - Identify or remove outliers
- **Data integration**
  - Integration of multiple data sources (databases)
  - Resolve inconsistencies
- **Data reduction**
  - Dimensionality reduction (More on the next chapter)
- **Data transformation and data discretisation**
  - Normalisation, Standardisation

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Why data cleaning?

- Data in the Real World Is Dirty:
- Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
- Two major dirtiness in data
  - Incomplete (missing): lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation* = “ ” (missing data)
  - Incorrect (noise): containing noise, errors, or outliers
    - e.g., *Salary* = “−10” (an error)

Incomplete Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

How to Handle Missing Data?

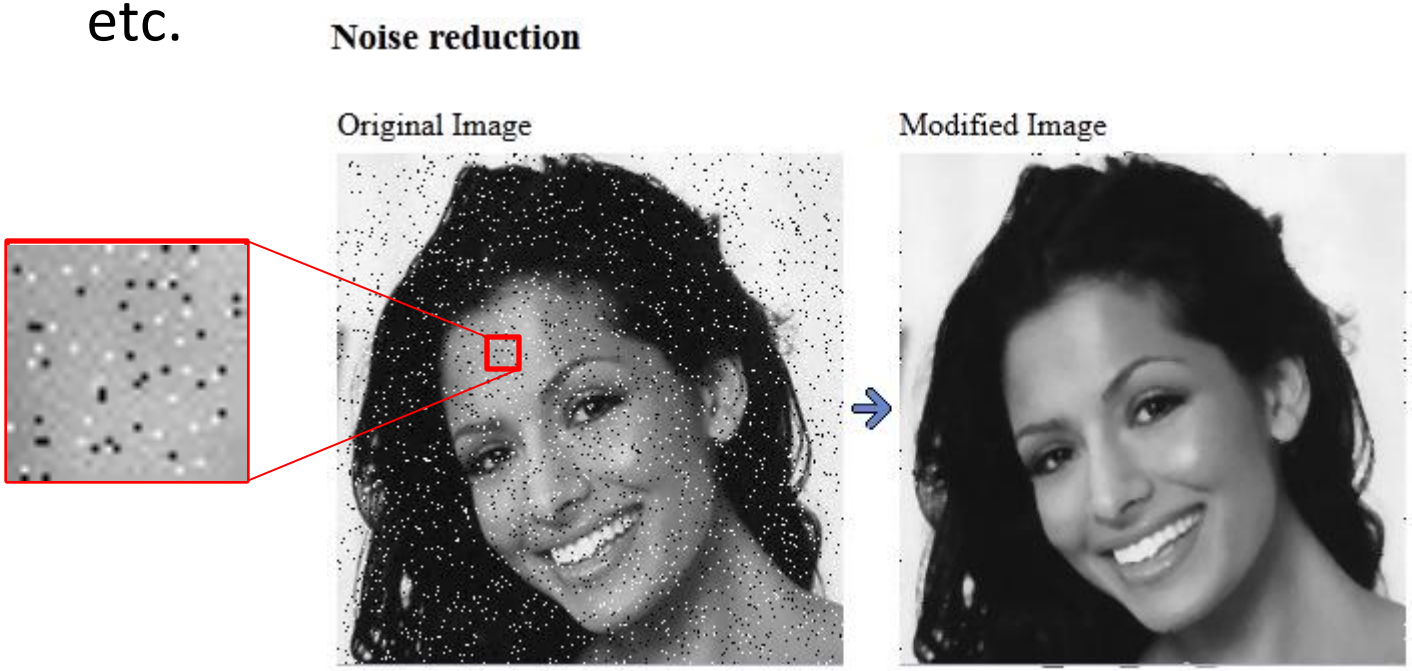
- Ignore the data point: usually done when class label is missing (when doing classification)—not effective why?
- Fill in the missing value manually: problem?
- Fill in automatically with
  - a global constant : e.g., “unknown”, “0”.
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter

Incorrect (Noisy) Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
  - Faulty data collection instruments
  - Disagreement between data experts
  - Data transmission problems
  - Technology limitation

How to Handle Noisy Data?

- Binning
  - Used to reduce the effects of minor observation error by consulting neighbours.
  - then one can smooth by bin means, smooth by bin median, etc.



More on Handling Noise

- Clustering
  - detect and remove outliers

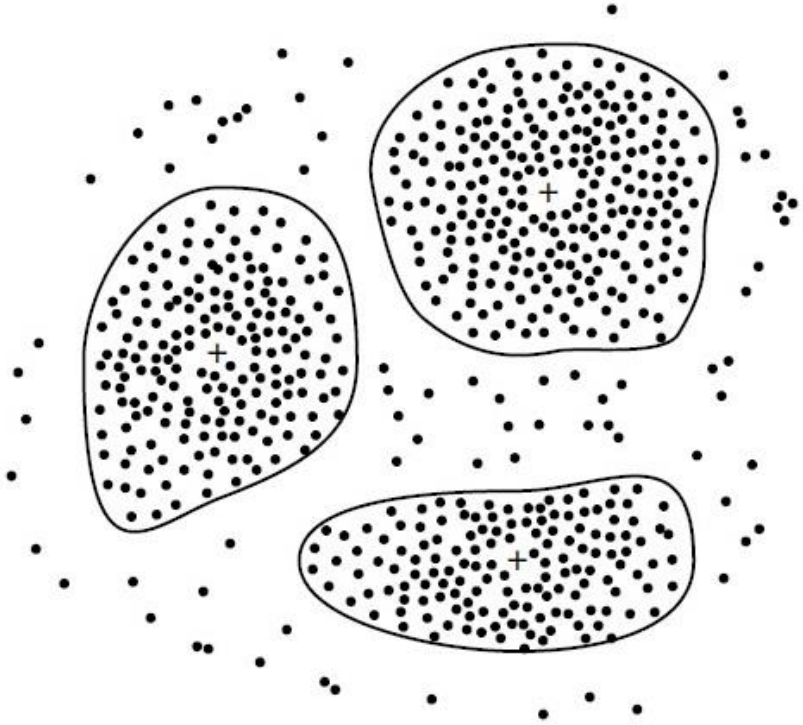
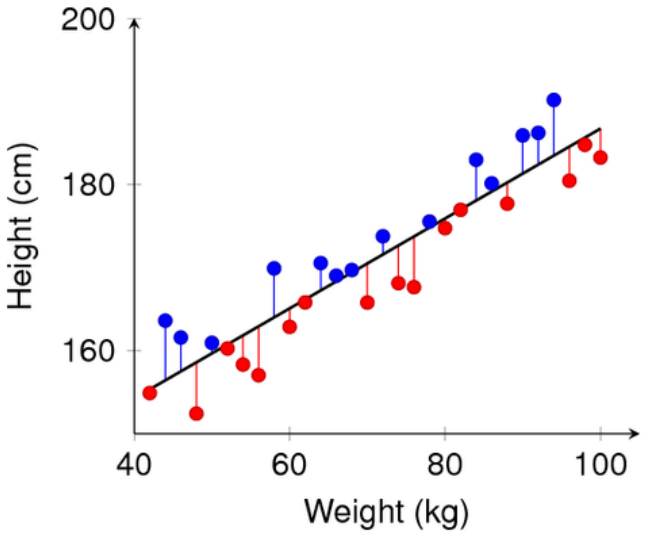


Figure 1.11 A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters. Each cluster "center" is marked with a "+".

Even more on handling noise

- Regression
  - smooth by fitting the data into regression functions



- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Data Integration

- **Data integration**:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton (same person)
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
- Redundant attributes may be able to be detected by *correlation analysis* and *covariance analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Correlation Analysis (Nominal Data)

- X^2 (chi-square) test
- A form of hypothesis testing
- Null hypothesis: Two variables are independent.
- Formula is given by

$$X^2 = \sum_{j=1}^{R} \sum_{i=1}^{C} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

- $E_{i,j} = \frac{(\sum_{l=1}^{L} O_{i,l}) \cdot (\sum_{r=1}^{R} O_{r,j})}{N}$ is theoretical frequency given null hypothesis, that is $P(X, Y) = P(X) \times P(Y)$
- The larger the X^2 value, the more likely we can reject the null hypothesis (two variables are not independent)

Chi-Square Calculation: An Example

Contingency Table

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the null hypothesis of the 2 classes)

$$\chi^2 = \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} = 507.93$$

- It shows that like_science_fiction and play_chess are correlated in the group

Covariance (Numeric Data)

$$cov(X, Y) = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{N}$$

- N is the number of data points, $\overline{x}$ and $\overline{y}$ are the respective mean of feature X and feature Y.
- If the two random variables are identical then this reduces to variance.

- It measures of how much two random variables change together.

- **Independence**: $cov(X, Y) = 0$
  - BUT the converse is not true:
  - Some pairs of random variables may have a covariance = 0 but are not independent.
  - Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence

Covariance: An Example

- Suppose two stocks X and Y have the following values in one week:
  - (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
  - $\overline{x}$ = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4
  - $\overline{y}$ = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6
  - cov(X,Y) = ((2-4)(5-9.6)+ (3-4)(8-9.6) + (5-4)(10-9.6) + (6-4)(14-9.6))/5 = 5.92
- Thus, X and Y rise together

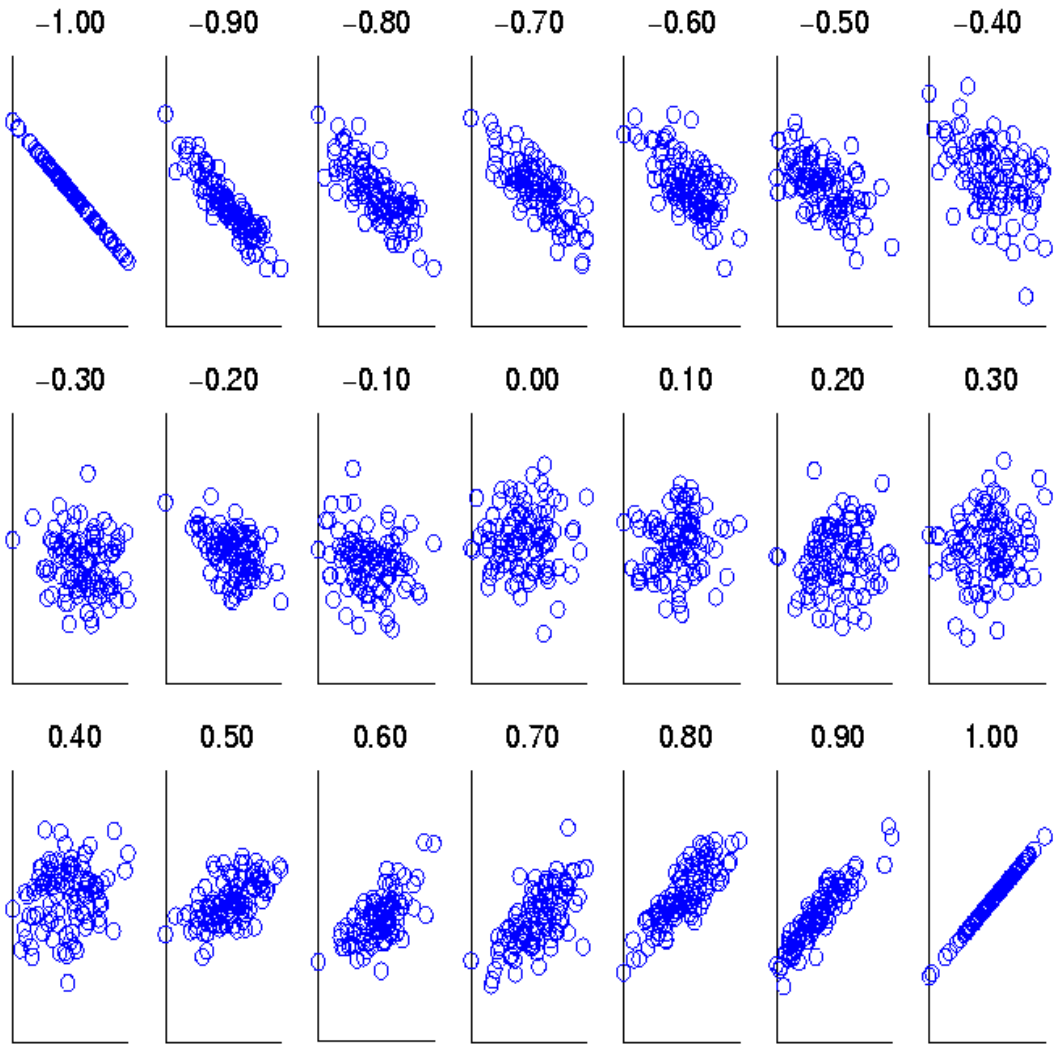
Correlation Analysis (Numeric Data)

- Correlation coefficient (Pearson's product moment coefficient).
- $R(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$
- Measures linear correlation

- If $R(X,Y) > 0$, X and Y are positively correlated (A's values increase as Y's). The higher, the stronger correlation.
- $R(X,Y) = 0$: independent.

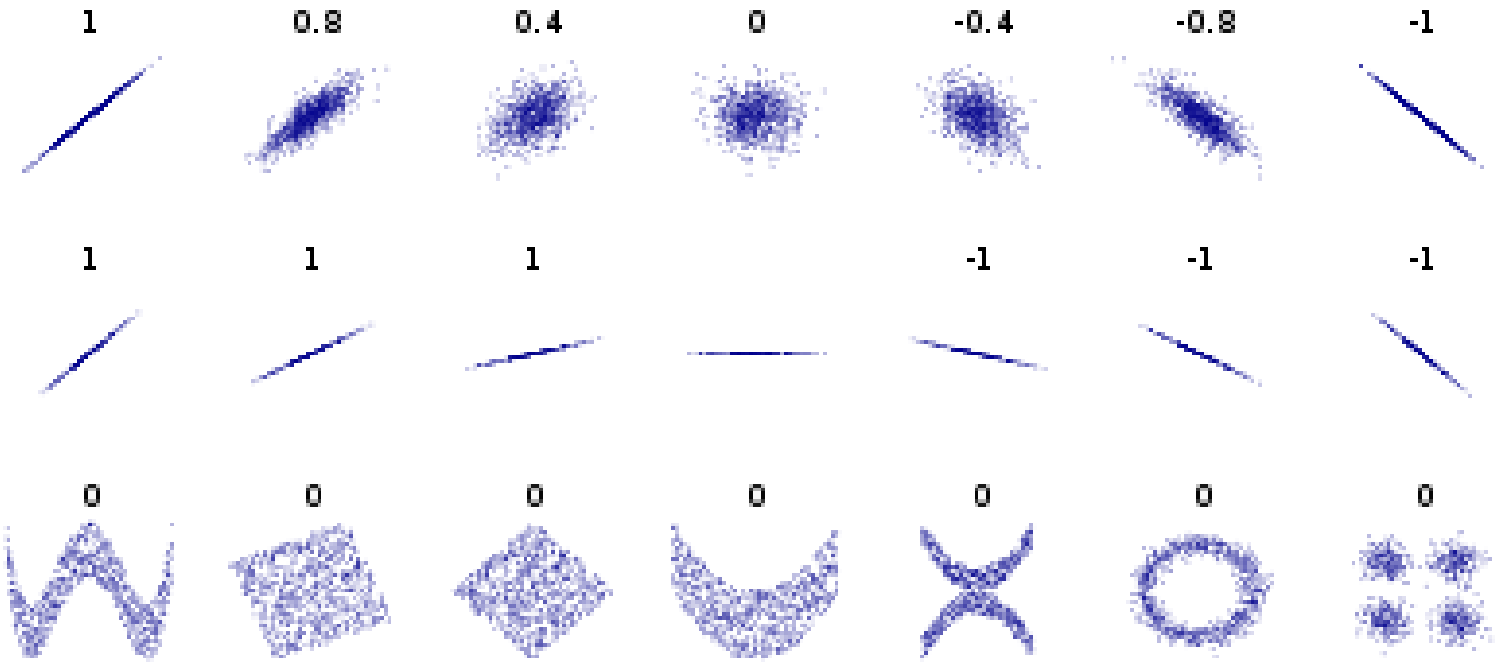
- $R(X,Y) < 0$: negatively correlated.

Visually Evaluating Correlation



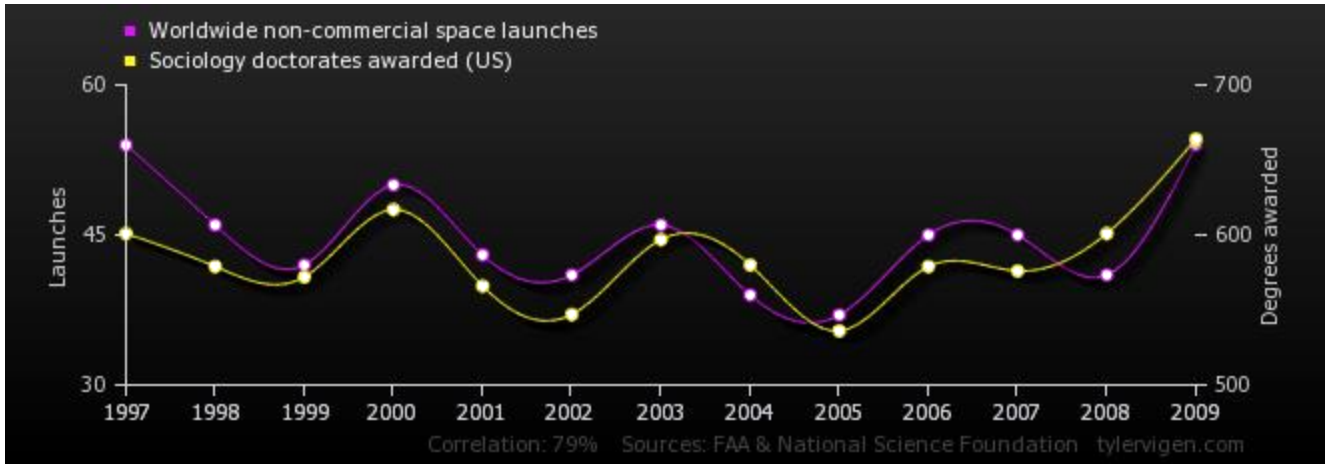
Scatter plots showing the correlation from –1 to 1.

Examples on Capturing Linear Relations

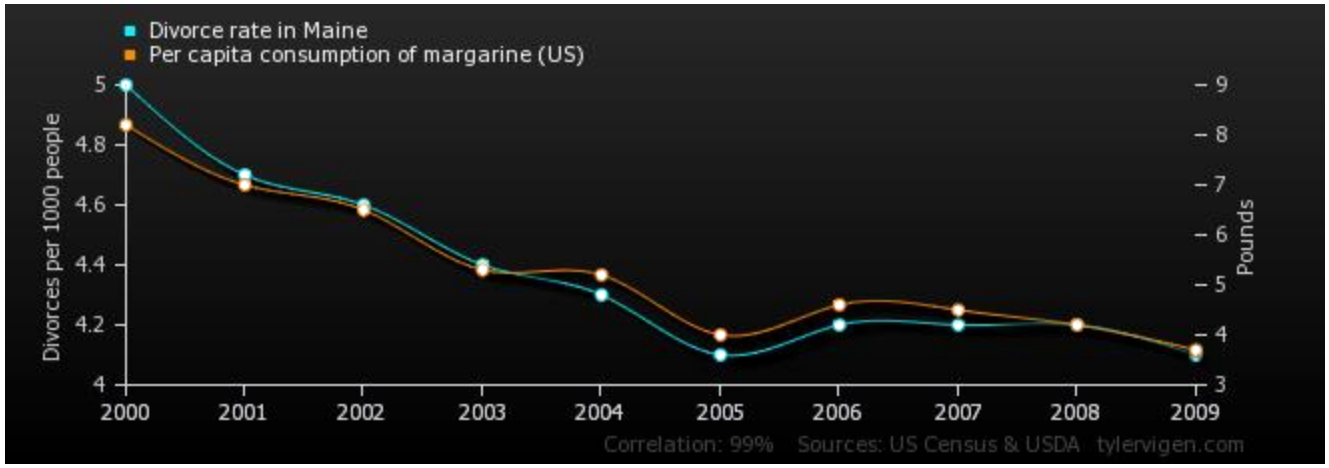


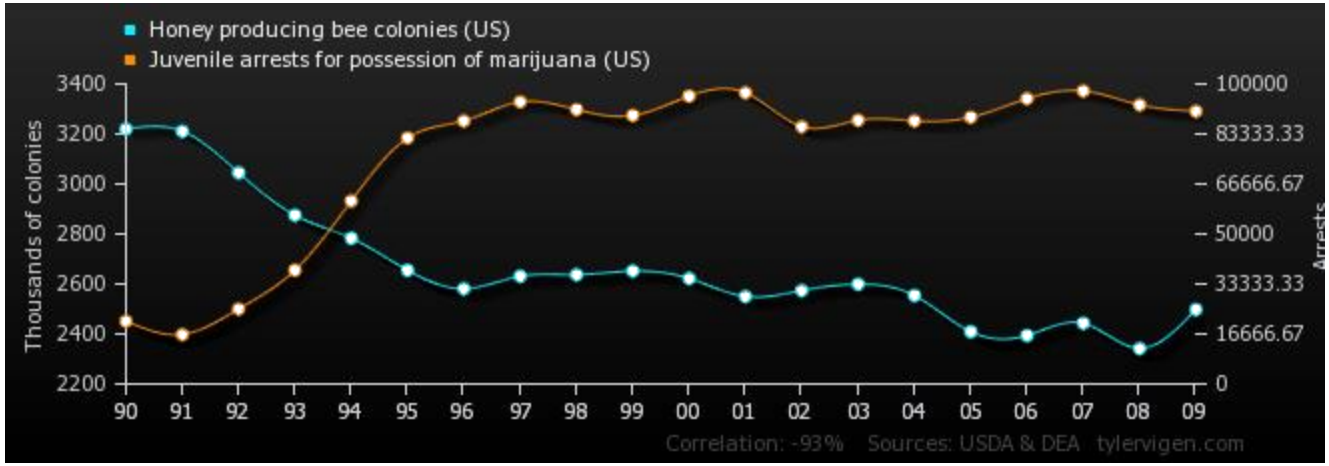
Example of Spurious Correlation

- Correlation does not imply causality



More examples





Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction 
- Data Transformation and Data Discretization
- Summary

Data Reduction Strategies

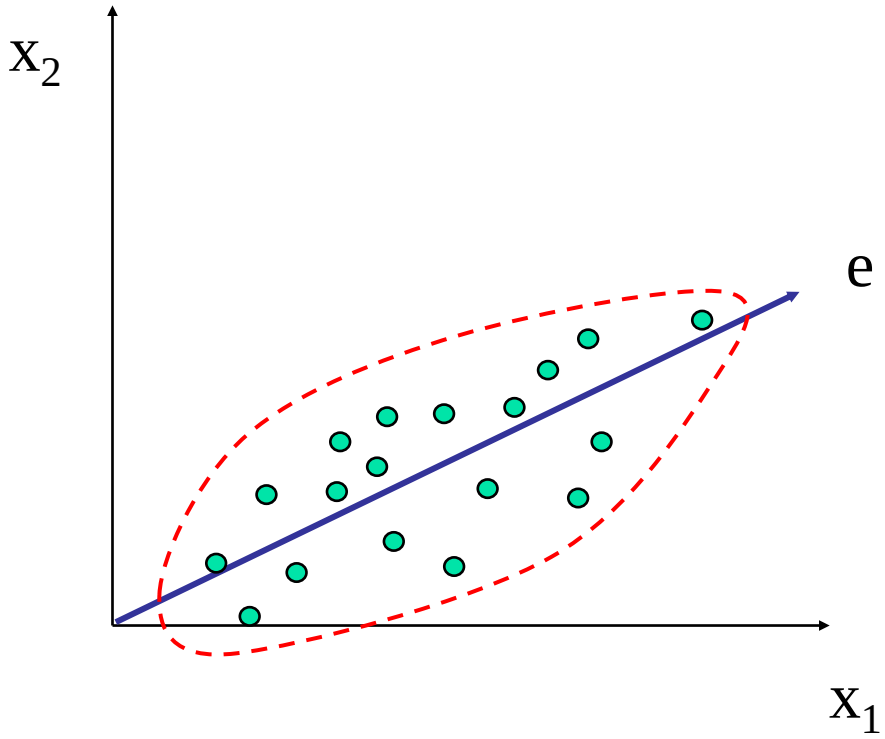
- **Data reduction**: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- Data reduction strategies
  - Dimensionality reduction, e.g., remove unimportant features
    - Principal Components Analysis (PCA)
    - Feature subset selection
  - Data reduction
    - Histograms
    - Clustering
    - Sampling
    - Model-based (Gaussian Mixture Model)

Dimensionality Reduction

- **Curse of dimensionality**
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially
- **Dimensionality reduction**
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization
- **Dimensionality reduction techniques**
  - Principal Component Analysis
  - Feature subset selection

Principal Component Analysis (PCA)

- Find a projection that captures the largest amount of variation in data
- The original data are projected onto a much smaller space, resulting in dimensionality reduction. We find the eigenvectors of the covariance matrix, and these eigenvectors define the new space



Principal Component Analysis (Steps)

- Given *N* data vectors from M-dimensions, find $k \leq M$ orthogonal vectors (*principal components*) that can be best used to represent data
  - Normalize input data: Each attribute falls within the same range
  - Compute *k* orthonormal (unit) vectors, i.e., *principal components*
  - Each input data (vector) is a linear combination of the *k* principal component vectors
  - The principal components are sorted in order of decreasing "significance" or strength
  - Since the components are sorted, the size of the data can be reduced by eliminating the *weak components*, i.e., those with low variance (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data)
- Works for numeric data only

Feature Selection

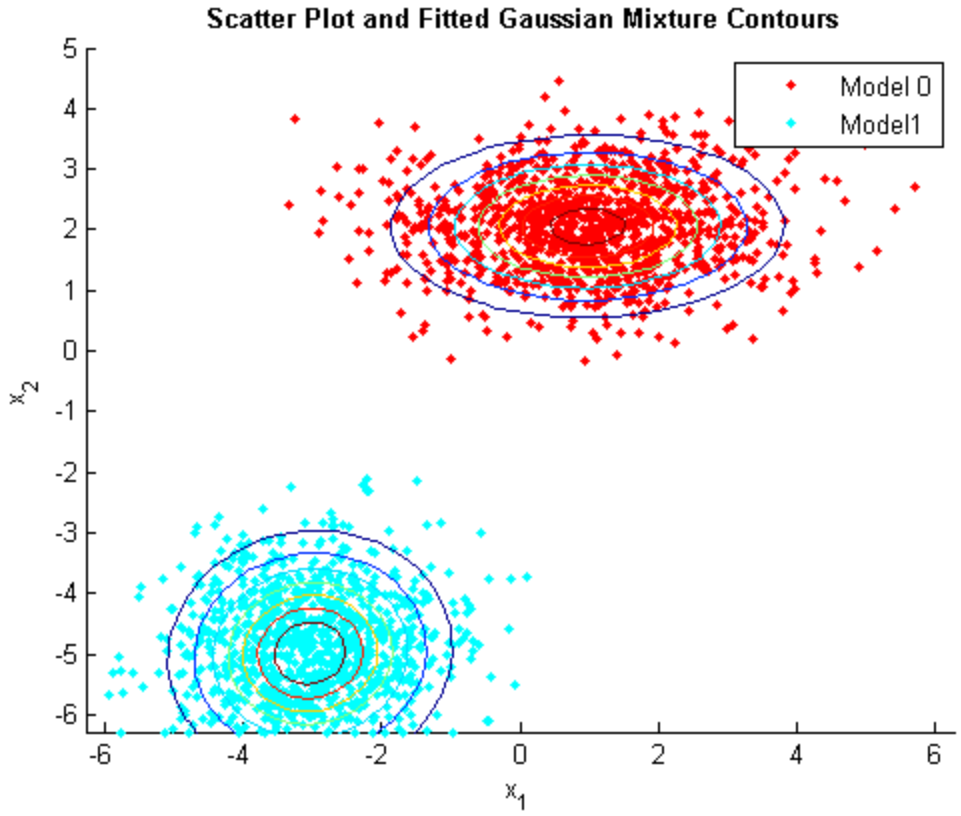
- Select relevant subset of features for the task.
- Can be divided into 3 categories.
  - Wrapper
    - Find all combinations of features and evaluate the usefulness of the subset using task's performance criteria. (e.g., the predictive performance)
  - Filter
    - Use mutual Information or correlation with target class.
  - Embedded
    - Feature selection is part of the learning machine.
    - LASSO, SVM.

Data Reduction

- Reduce data volume by choosing alternative, *smaller forms* of data representation
- **Parametric methods**
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
  - Ex.: Mixture Model
- **Non-parametric methods**
  - Do not assume models
  - Major families: histograms, clustering, sampling

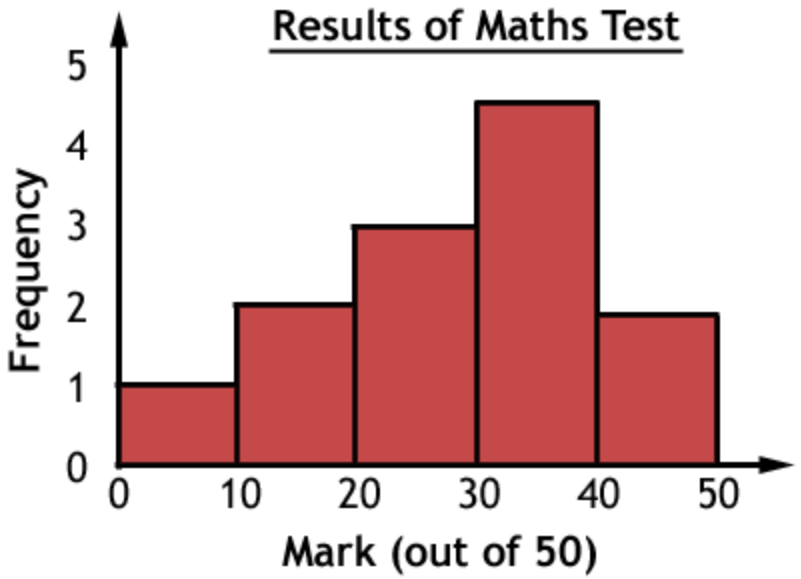
Gaussian Mixture Model

Assume the data obeys Gaussian Distribution, Estimate the model and store only the model's parameters.



Histogram Analysis

- Divide data into buckets and store average for each bucket
- Partitioning rules:
  - Equal-width: equal bucket range
  - Equal-frequency (or equal-depth)



Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Can be very effective if data is clustered but not if data is "smeared"
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms
- Cluster analysis will be studied in depth later.

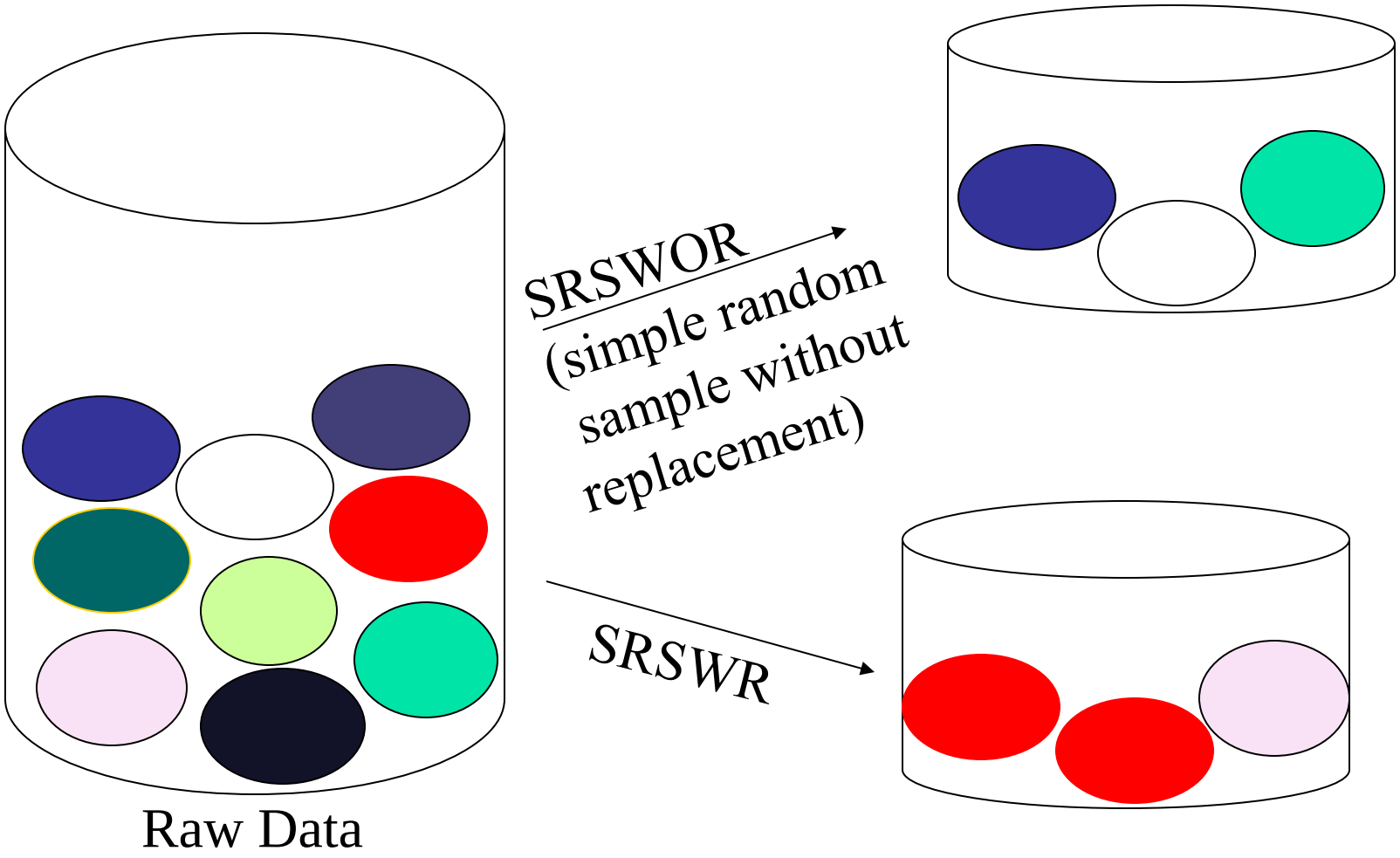
Sampling

- Sampling: obtaining a small sample *s* to represent the whole data set *N*
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Key principle: Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
  - Develop adaptive sampling methods, e.g., stratified sampling:

Types of Sampling

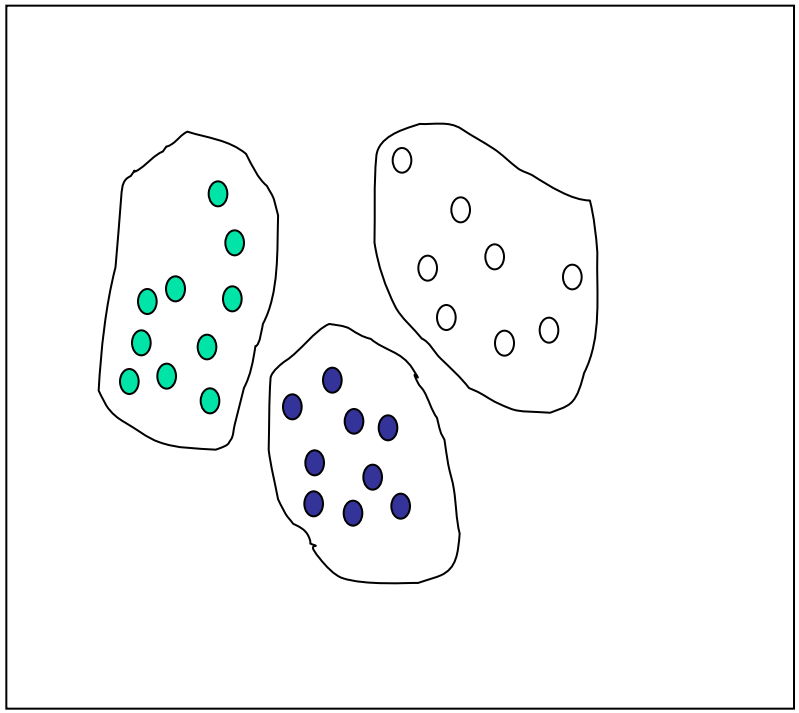
- **Simple random sampling**
  - There is an equal probability of selecting any particular item
- **Sampling without replacement**
  - Once an object is selected, it is removed from the population
- **Sampling with replacement**
  - A selected object is not removed from the population
- **Stratified sampling:**
  - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
  - Used in conjunction with skewed data

Sampling: With or without Replacement

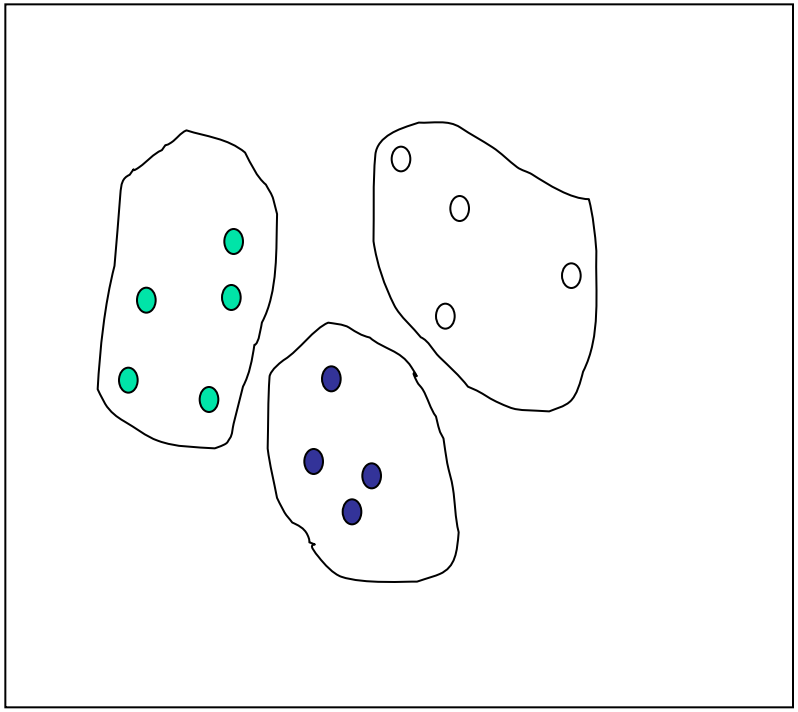


Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample



Chapter 3: Data Preprocessing

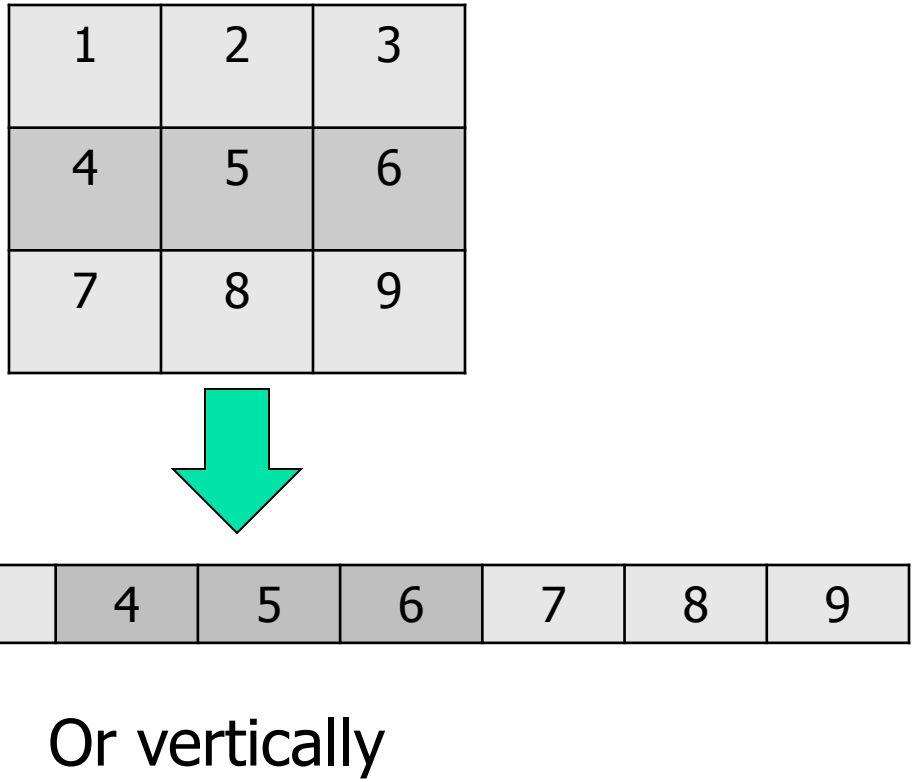
- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization 
- Summary

Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values
- Methods
  - Attribute/feature construction
    - New attributes constructed from the given ones
    - Image → data point, Signal → data point
  - Normalization: Scaled to fall within a smaller, specified range
    - min-max normalisation
    - z-score normalisation
    - normalisation by decimal scaling
  - Discretisation: Transform continuous variable to discrete ones.

Image → Data vector

- Reshaping an image into a vector of length (Width x Height)



Normalisation

- **Min-max normalisation**: to [new_min$_X$, new_max$_X$]

$$v' = \frac{v - min_X}{max_X - min_X}(new\_max_X - new\_min_X) + new\_min_X$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73,600 - 12,000}{98,000 - 12,000}(1.0 - 0) + 0 = 0.716$

- **Z-score normalisation** (μ: mean, σ: std dev): a.k.a standardisation

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73,600 - 54,000}{16,000} = 1.225$

- **Normalisation by decimal scaling**

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max(|v'|) < 1

Discretisation

- Three types of attributes
  - Nominal—values from an unordered set, e.g., color, profession
  - Ordinal—values from an ordered set, e.g., military or academic rank
  - Numeric—real numbers, e.g., integer or real numbers
- Discretisation: Divide the range of a continuous attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Reduce data size by discretization
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute

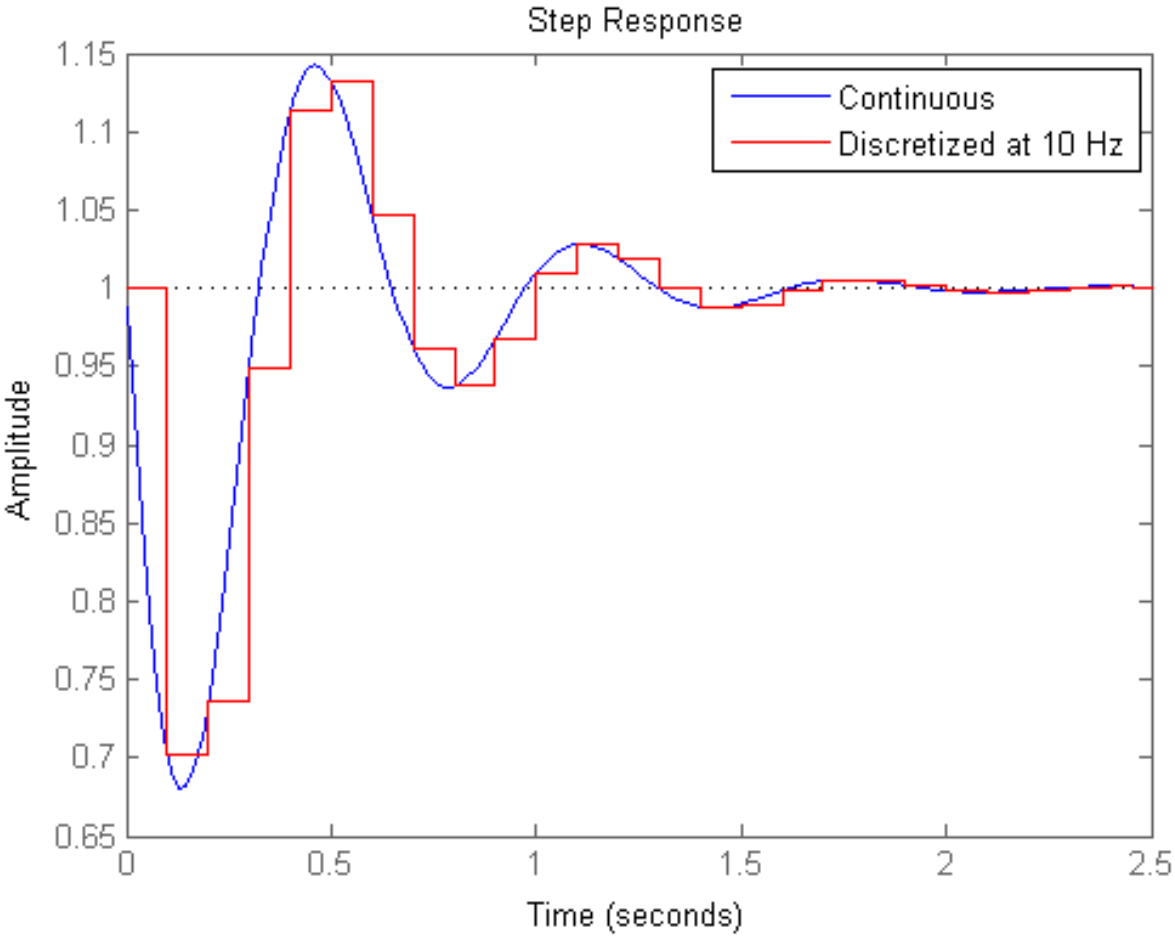
Data Discretisation Methods

- Typical methods: All the methods can be applied recursively
  - Binning
    - Top-down split, unsupervised
  - Histogram analysis
    - Top-down split, unsupervised
  - Clustering analysis (unsupervised, top-down split or bottom-up merge)
  - Correlation (e.g., χ^2) analysis (unsupervised, bottom-up merge)

Simple Discretization: Binning

- Binning Image's pixel to get a more compact representation.
- Bin size can vary. Smaller in the region with high variations.

Sound → Data Vector



Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Summary

- **Data quality**:
  - accuracy, completeness, consistency, timeliness
- **Data cleaning**: e.g. missing/noisy values, outliers
- **Data integration** from multiple sources:
  - Remove redundancies; Detect inconsistencies
- **Data reduction**
  - Dimensionality reduction; Dataset size reduction
- **Data transformation and data discretization**