CS378 Introduction to Data Mining

Data Exploration and Data Preprocessing

Li Xiong

Data Exploration and Data Preprocessing

- Data and Attributes
- Data exploration
- Data pre-processing

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

Objects

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

Types of Attributes

- Categorical (qualitative)
  - 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}
- Numeric (quantitative)
  - 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: $=\ \neq$
  - Order: $<\ >$
  - Addition: $+\ -$
  - Multiplication: $*\ /$

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

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

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.
  - Continuous attributes are typically represented as floating-point variables.

- Typically, nominal and ordinal attributes are discrete attributes, while interval and ratio attributes are continuous

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
- Points in a multi-dimensional space, where each dimension represents a distinct attribute
- Represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

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**

Document Data

- Document-term matrix
  - Each document is a `term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

	team	coach	pla y	ball	score	game	wi n	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Transaction Data

- A special type of record data, where
  - each record (transaction) has a set of items
  - transaction-item matrix vs transaction list

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

Data Exploration and Data Preprocessing

- Data and Attributes
- Data exploration/summarization
  - Summary statistics
  - Graphical description (visualization)
- Data pre-processing

Summary Statistics

- Summary statistics are quantities, such as mean, that capture various characteristics of a potentially large set of values.
  - Measuring central tendency – how data seem similar, location of data
  - Measuring statistical variability or dispersion of data – how data differ, spread

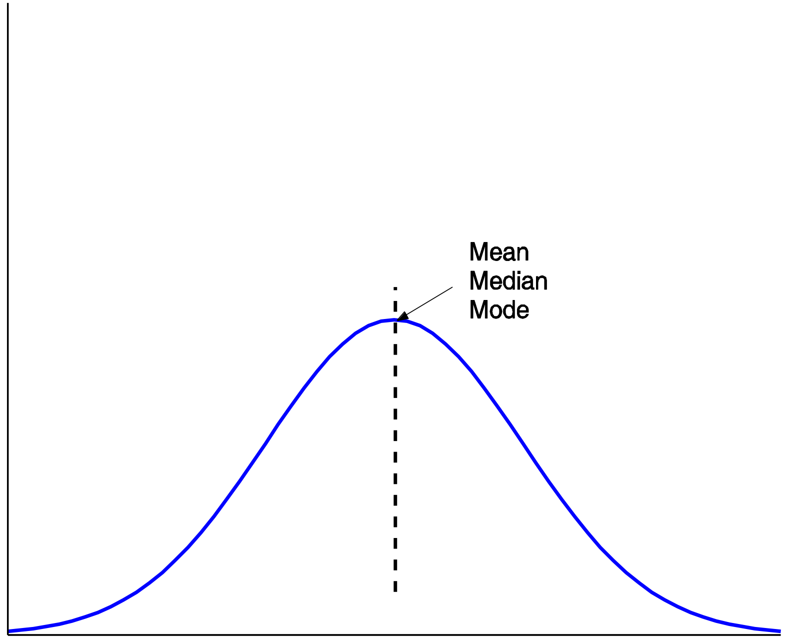
Measuring the Central Tendency

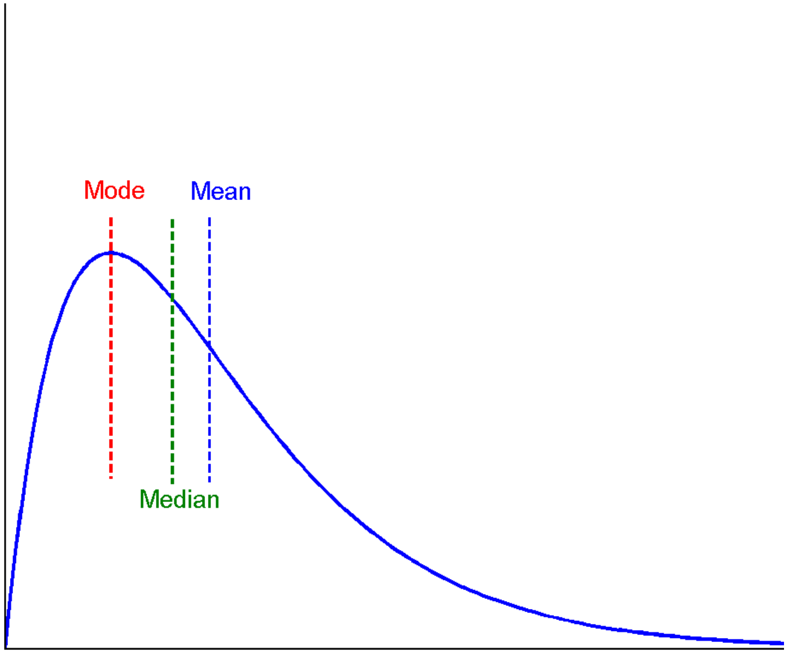
- Mean (sample vs. population): $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\mu = \frac{\sum x}{N}$
  - Weighted arithmetic mean: $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
  - Trimmed mean: chopping extreme values
- Median
  - Middle value if odd number of values, or average of the middle two values otherwise
- Mode
  - Value that occurs most frequently in the data
  - Mode may not be unique
  - Unimodal, bimodal, trimodal

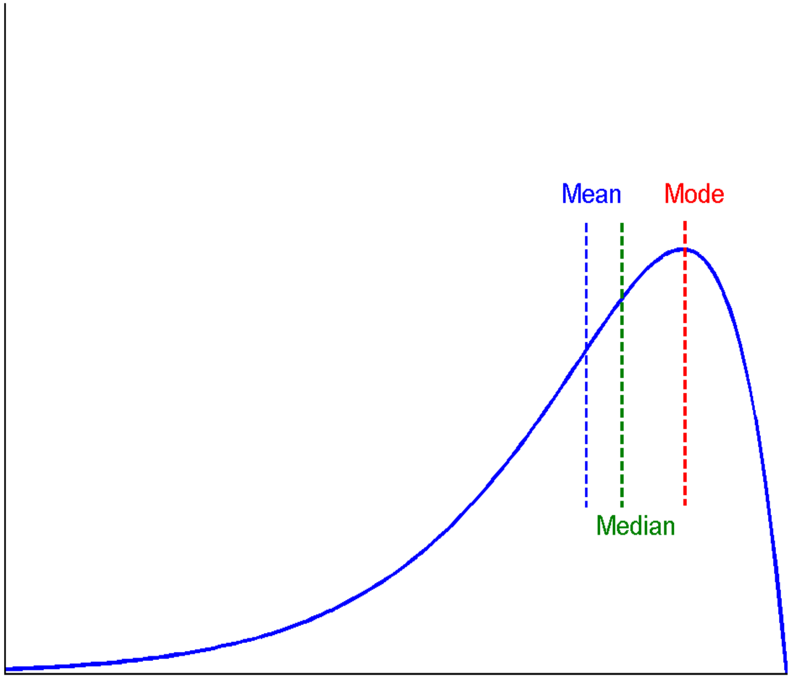
- Which ones make sense for nominal, ordinal, interval, ratio attributes respectively?

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data

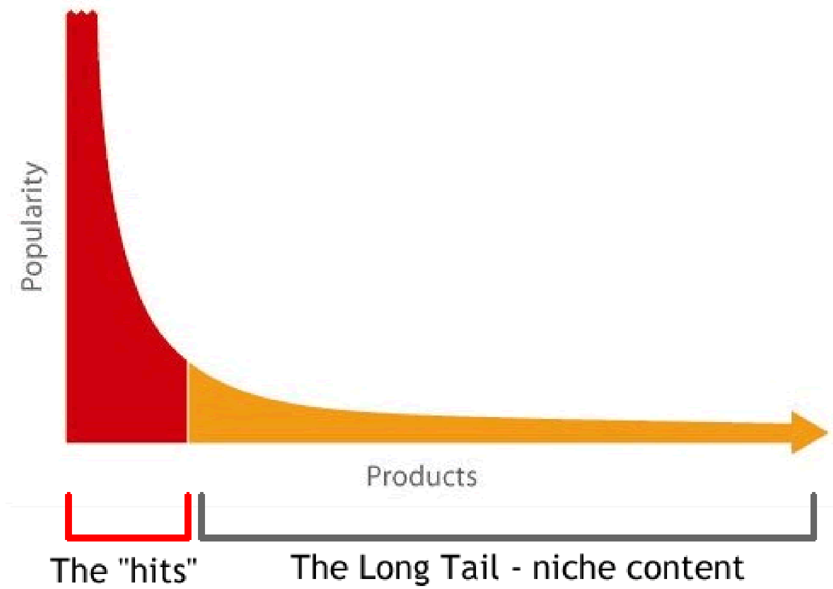






The Long Tail

- Long tail: low-frequency population (e.g. wealth distribution)
- The Long Tail [Anderson]: the current and future business and economic models
  - Empirical studies: Amazon, Netflix
  - Products that are in low demand or have low sales volume can collectively make up a market share that rivals or exceeds the relatively few bestsellers and blockbusters



- The Long Tail. Chris Anderson, Wired, Oct. 2004
- The Long Tail: Why the Future of Business is Selling Less of More. Chris Anderson. 2006

Computational Issues

- Different types of measures
  - Distributed measure – can be computed by partitioning the data into smaller subsets. E.g. sum, count
  - Algebraic measure – can be computed by applying an algebraic function to one or more distributed measures. E.g. ?
  - Holistic measure – must be computed on the entire dataset as a whole. E.g. ?
- Ordered statistics (selection algorithm): finding *k*th smallest number in a list. E.g. min, max, median
  - Selection by sorting: $O(n * \log n)$
  - Linear algorithms based on quicksort: $O(n)$

Measuring the Dispersion of Data

- Dispersion or variance: the degree to which numerical data tend to spread
- Range and Quartiles
  - Range: difference between the largest and smallest values
  - Percentile: the value of a variable below which a certain percent of data fall
  - Quartiles: Q_1 (25th percentile), Median (50th percentile), Q_3 (75th percentile)
  - Inter-quartile range: IQR = $Q_3 - Q_1$
  - Five number summary: min, Q_1, M, Q_3, max (Boxplot)
  - Outlier: usually, a value at least 1.5 x IQR higher/lower than Q3/Q1
- Variance and standard deviation (*sample: s, population: σ)*
  - Variance: sample vs. population (algebraic or holistic?)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}x_i^{\ 2} - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2] \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^{\ 2} - \mu^2$$

  - Standard deviation *s (or σ)* is the square root of variance *s^2 (or σ^2)*

Data Exploration and Data Preprocessing

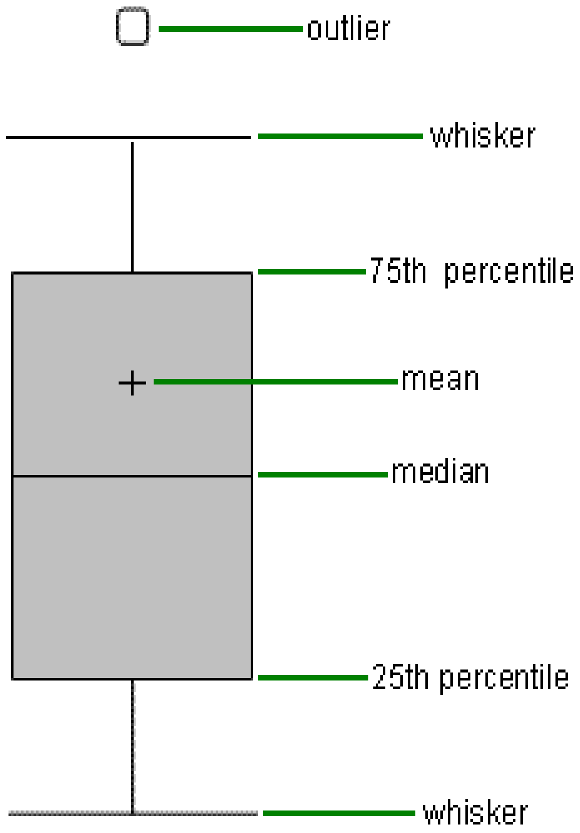
- Data and Attributes
- Data exploration
  - Summary statistics
  - Visualization
  - Online Analytical Processing (OLAP)
- Data pre-processing

Graphic Displays of Basic Statistical Descriptions

- Boxplot
- Histogram
- Scatter plot

Boxplot Analysis

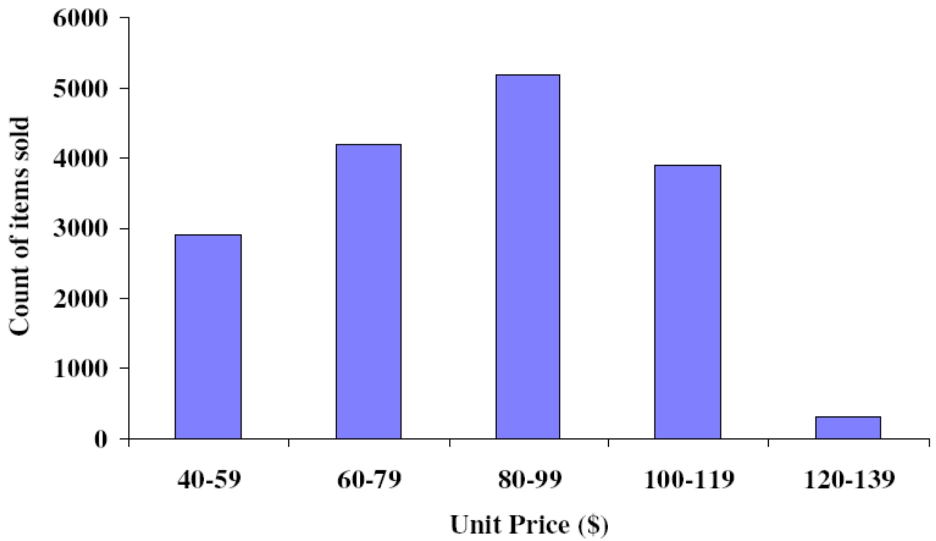
- The ends of the box are first and third quartiles (Q1 and Q3), i.e., the height of the box is IRQ
- The median (M) is marked by a line within the box
- Whiskers: two lines outside the box extend to Minimum and Maximum



Demo:
http://www.shodor.org/interactivate/activities/BoxPlot/

Histogram Analysis

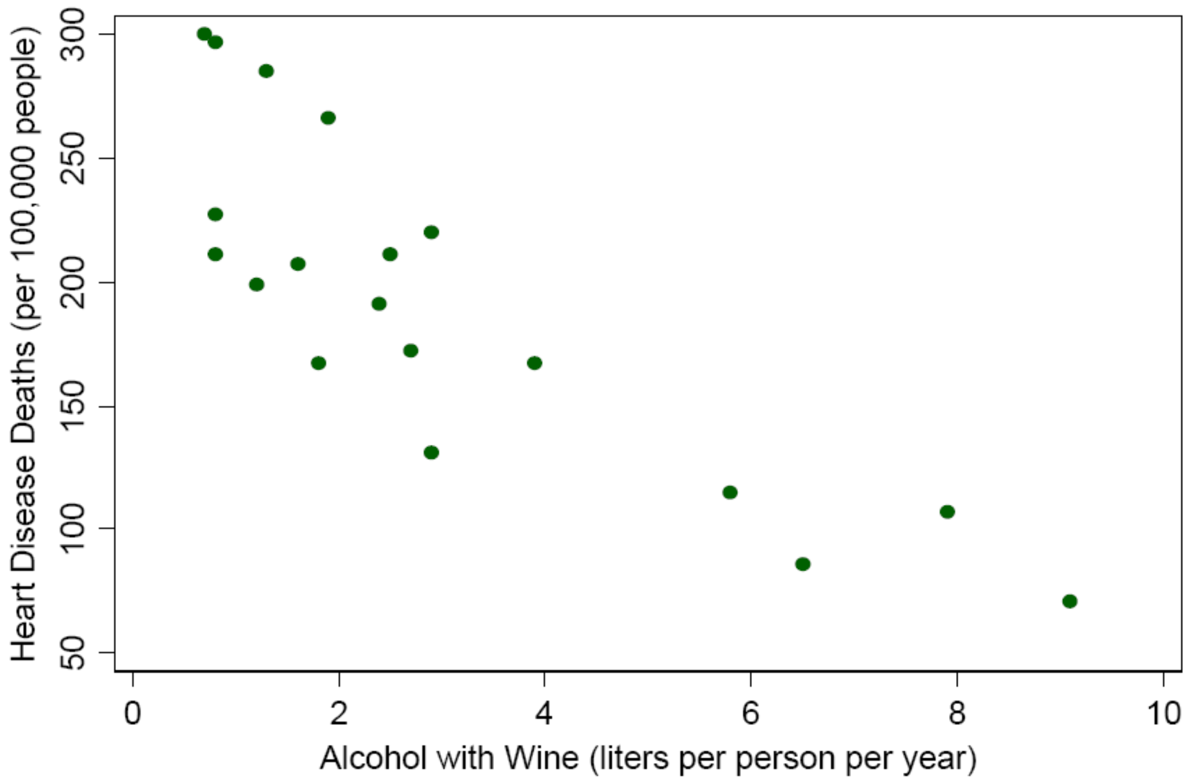
- Univariate (one attribute) vs multivariate
- Data partitioned into disjoint *buckets*
  - Unsupervised (typically equal-width)
  - Supervised
- A set of rectangles that reflect the counts or frequencies of values at the bucket (bar chart)



Demo:
http://www.shodor.org/interactivate/activities/Histogram/

Scatter plot

- Displays values for two numerical attributes (bivariate data)
- Each pair of values plotted as a point in the plane
- can suggest correlations between variables with a certain confidence level: positive (rising), negative (falling), or null (uncorrelated).



Data Exploration and Data Preprocessing

- Data and Attributes
- Data exploration
- Data pre-processing
  - Data cleaning
  - Data integration
  - Data transformation
  - Data reduction

Data Quality Issues

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=" "
  - noisy: containing errors or outliers
    - e.g., Salary="-10"
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age="42" Birthday="03/07/1997"
    - e.g., Was rating "1,2,3", now rating "A, B, C"
    - e.g., discrepancy between duplicate records
  - duplicate: containing duplicate records

How to Handle Missing Values?

- Missing data mechanism
  - Missing completely at random
  - Missing at random
  - Missing not at random
- Techniques to handle missing data
  - Ignore the tuple (deletion)
  - Fill in the missing value (imputation)
    - a global constant : e.g., "unknown", a new class?!
    - the attribute mean
    - the attribute mean for all samples belonging to the same class: smarter
    - the most probable value: inference-based prediction methods (discussed later)

How to Handle Noisy Data?

- Noise: random error or variance in a measured variable
- Binning and smoothing
  - sort data and partition into bins (equi-width, equi-depth)
  - then smooth by bin mean, bin median, bin boundaries, etc.
- Regression (discussed later)
  - smooth by fitting the data into a function with regression
- Clustering (discussed later)
  - detect and remove outliers that fall outside clusters
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Simple Discretization Methods: Binning

- Equal-width (distance) partitioning
  - Divides the range into N intervals of equal size: uniform grid
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well
- Equal-depth (frequency) partitioning
  - Divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

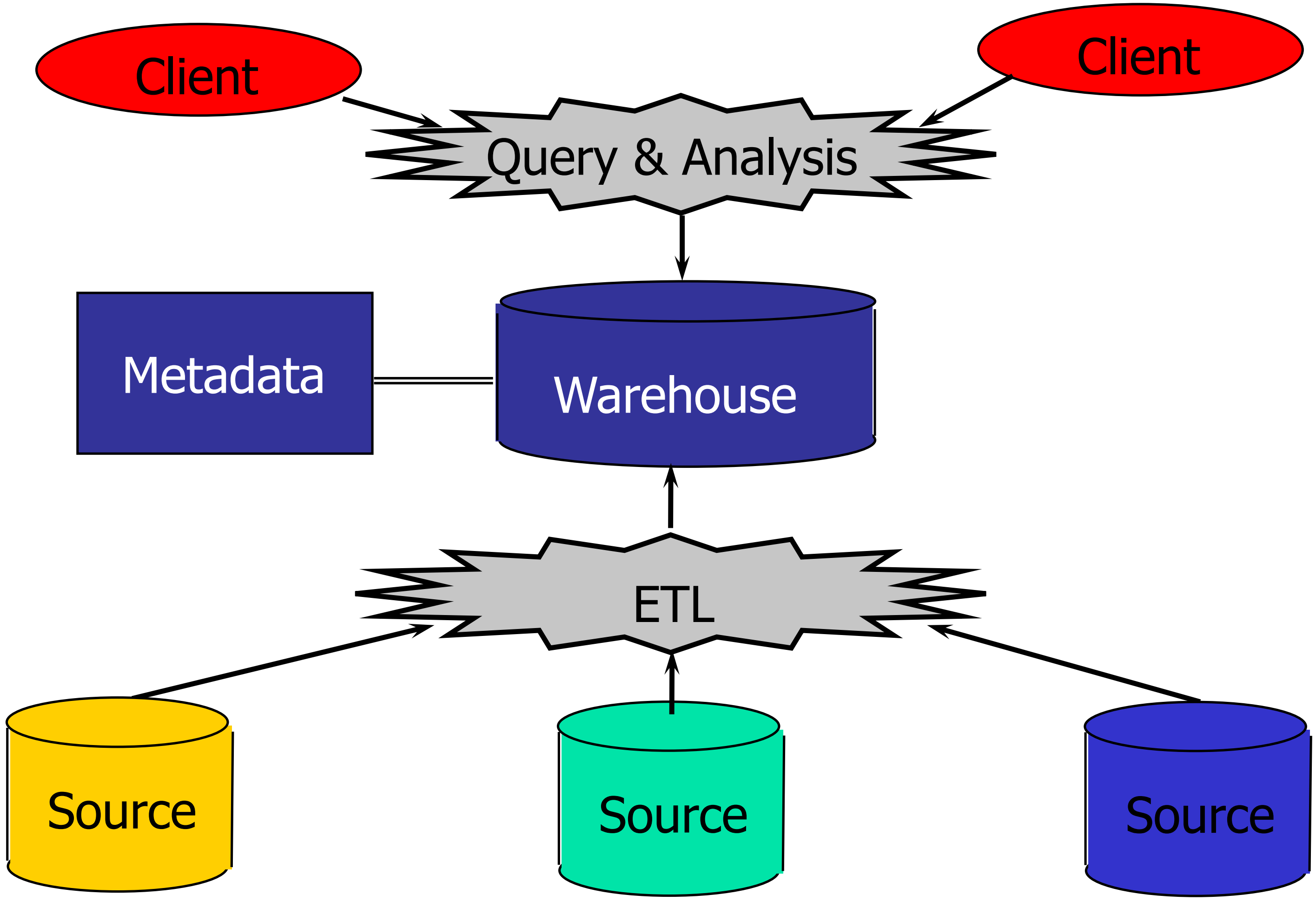
Data Exploration and Data Preprocessing

- Data and Attributes
- Data exploration
- Data pre-processing
  - Data cleaning
  - **Data integration**
  - Data transformation
  - Data reduction

Data Integration

- Data integration: combines data from multiple sources into a unified view
- Architectures
  - Data warehouse (tightly coupled)
  - Federated database systems (loosely coupled)
- Database heterogeneity
  - Semantic integration

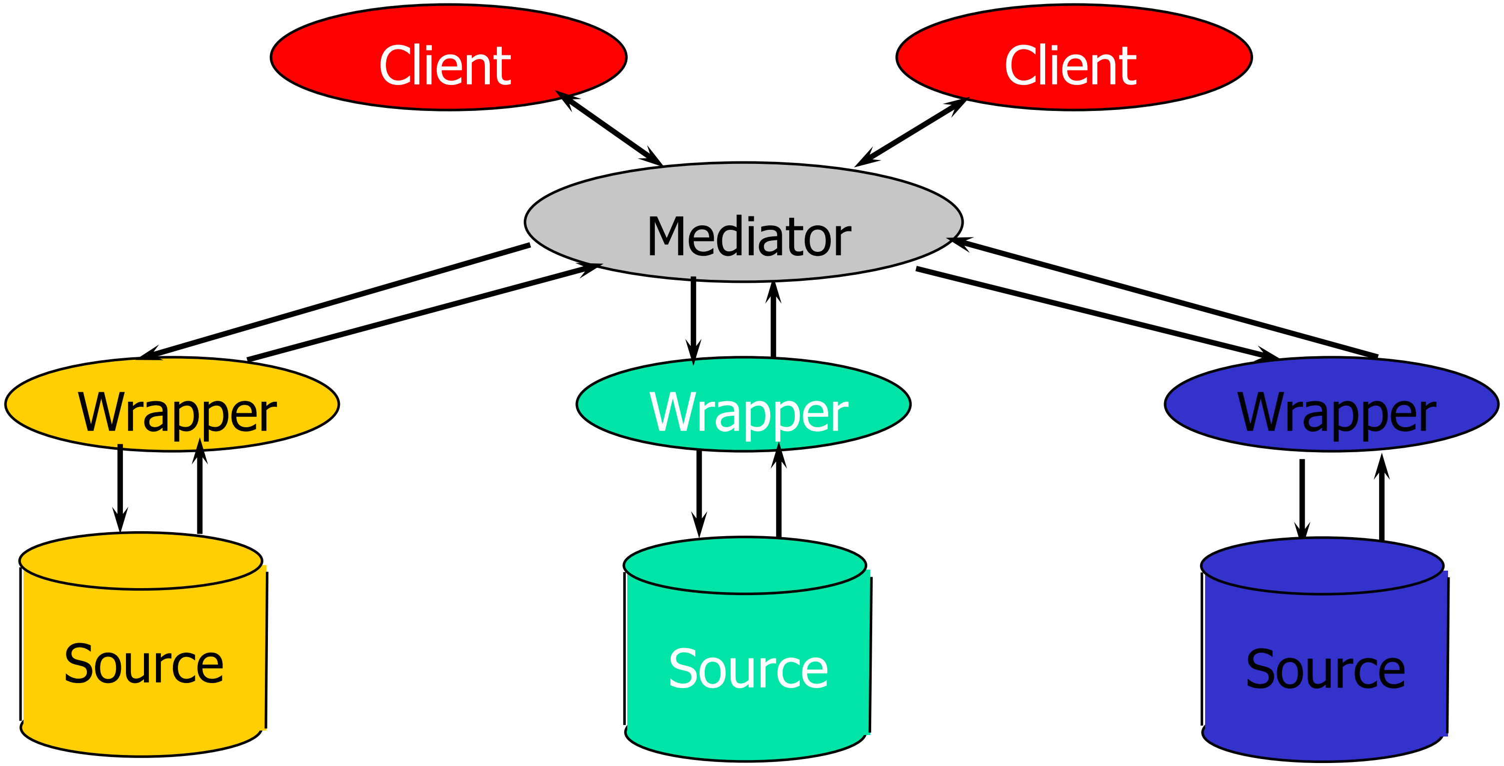
Data Warehouse Approach



Advantages and Disadvantages of Data Warehouse

- Advantages
  - High performance
  - Can operate when sources unavailable
  - Extra information at warehouse
    - Modification, summarization (aggregates), historical information
  - Local processing at sources unaffected
- Disadvantages
  - Data freshness
  - Difficult to construct when only having access to query interface of local sources
  - Privacy/security constraints

Federated Systems / Federated learning



Advantages and Disadvantages of Federated Systems

- Advantage
  - No need to copy and store data at mediator
  - More up-to-date data
  - Privacy/security advantage
- Disadvantage
  - Performance
  - Source availability
  - Convergence

Semantic Integration

- Problem: reconciling semantic heterogeneity
- Levels
  - Schema matching (schema mapping)
    - e.g., A.cust-id $\equiv$ B.cust-#
  - Data matching
    - e.g., Bill Clinton = William Clinton
- In practice, 60-80% of resources spent on reconciling semantic heterogeneity in data sharing project

Schema Matching

- Techniques
  - Rule based
  - Learning based
- Type of matches
  - 1-1 matches vs. complex matches (e.g. list-price = price *(1+tax_rate))
- Information used
  - Schema information: element names, data types, structures, number of sub-elements, integrity constraints
  - Data information: value distributions, frequency of words
  - External evidence: past matches, corpora of schemas
  - Ontologies. E.g. Gene Ontology
- Multi-matcher architecture

Data Matching (entity resolution, record linkage)

- Techniques
  - Rule based
  - Probabilistic Record Linkage (Fellegi and Sunter, 1969)
    - Similarity between pairs of attributes
    - Combined scores representing probability of matching
    - Threshold based decision
  - Machine learning approaches
- New challenges
  - Complex information spaces
  - Multiple classes

Data Exploration and Data Preprocessing

- Data and Attributes
- Data exploration
- Data pre-processing
  - Data cleaning
  - Data integration
  - **Data transformation**
  - Data reduction

Data Transformation

- Aggregation: sum/count/average
  - E.g. Daily sales -> monthly sales
- Discretization (continuous -> discrete)
  - E.g. age -> youth, middle-aged, senior
- (Statistical) Normalization: scaled to fall within a small, specified range
  - E.g. income vs. age
  - Not to be confused with database normalization and text normalization
- Attribute construction: construct new attributes from given ones
  - E.g. birthday -> age

Normalization

- scaled to fall within a small, specified range
- Min-max normalization: [min_A, max_A] to [new_min_A, new_max_A]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income [\$12,000, \$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 normalization (μ: mean, σ: standard deviation):

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

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

Data Exploration and Data Preprocessing

- Data and Attributes
- Data exploration
- Data pre-processing
  - Data cleaning
  - Data integration
  - Data transformation
  - Data reduction

Data Reduction

- Why data reduction?
  - A database/data warehouse may store terabytes of data
    - Number of data points
    - Number of dimensions
  - Complex data analysis/mining may take a very long time to run on the complete data set
- Data reduction
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results

Data Reduction

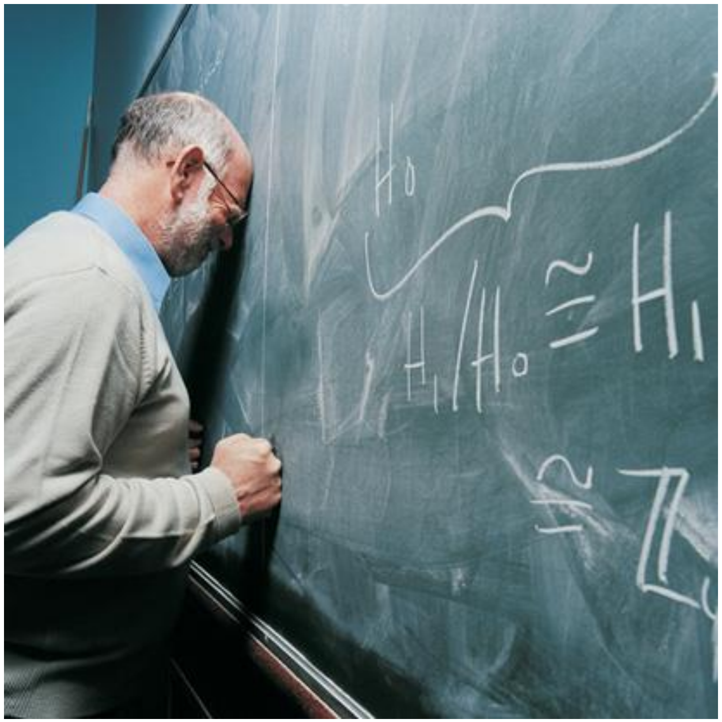
- Instance reduction
  - Sampling (instance selection)
  - Numerocity reduction
- Dimension reduction
  - Feature selection
  - Feature extraction
- Data compression

Instance Reduction: Sampling

- Sampling: obtaining a small representative sample *s* to represent the whole data set *N*
  - A sample is representative if it has approximately the same property (of interest) as the original set of data
- Statisticians sample because obtaining the entire set of data is too expensive or time consuming.
- Data miners sample because processing the entire set of data is too expensive or time consuming

- Sampling method
- Sampling size

Why sampling



A statistics professor was describing sampling theory

Student: I don't believe it, why not study the whole population in the first place?

The professor continued explaining sampling methods, the central limit theorem, etc.

Student: Too much theory, too risky, I couldn't trust just a few numbers in place of ALL of them.

The professor explained the Nielsen television ratings

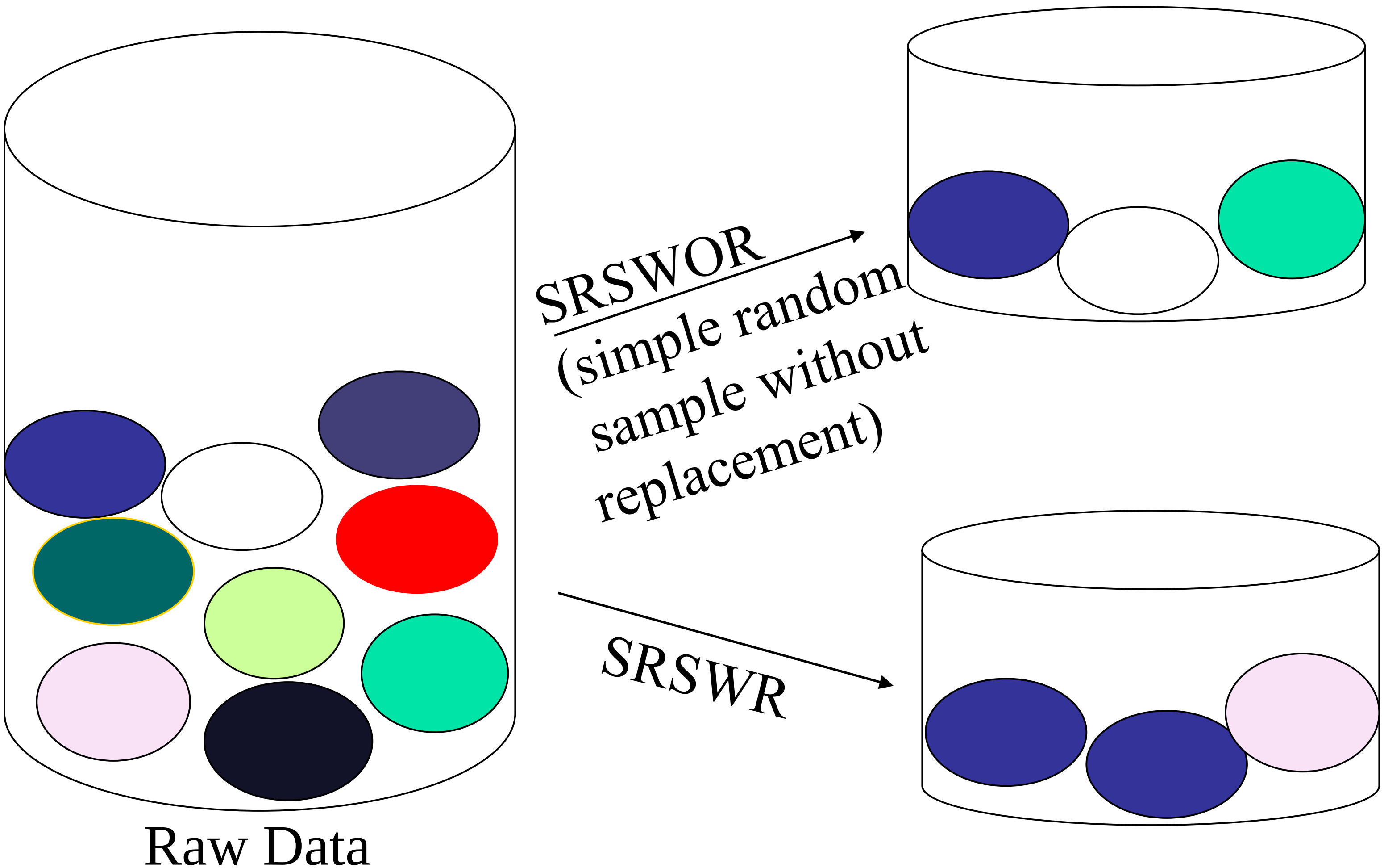
Student: You mean that just a sample of a few thousand can tell us exactly what over 250 MILLION people are doing?

Professor: Well, the next time you go to the campus clinic and they want to do a blood test…tell them that's not good enough …tell them to TAKE IT ALL!!"

Sampling Methods

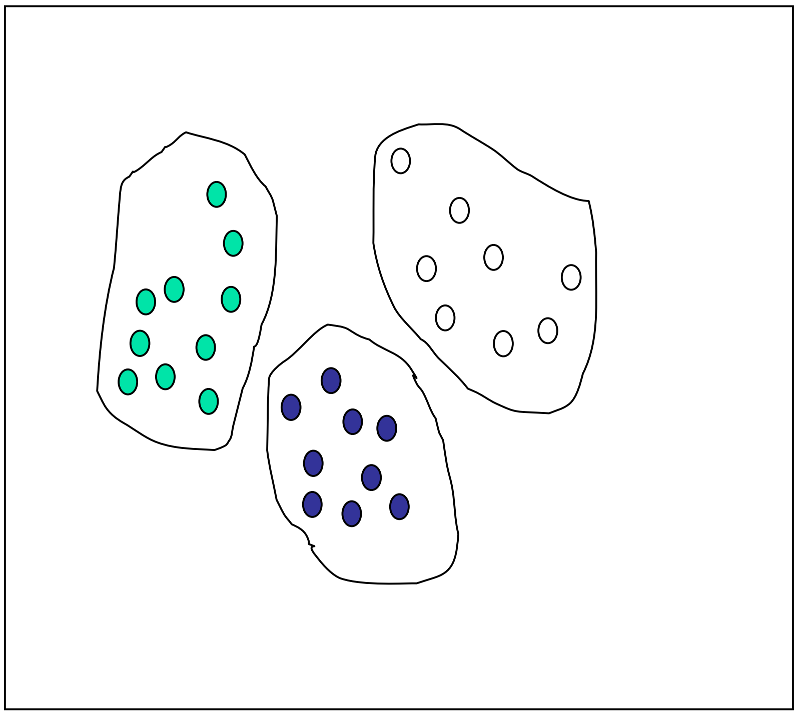
- Simple Random Sampling
  - There is an equal probability of selecting any particular item
- Sampling without replacement
  - As each item is selected, it is removed from the population
- Sampling with replacement
  - Objects are not removed from the population as they are selected for the sample - the same object can be picked up more than once
- Stratified sampling
  - Split the data into several partitions (stratum); then draw random samples from each partition
- Cluster sampling
  - When "natural" groupings are evident in a statistical population; sample a small number of clusters

Simple random sampling without or with replacement

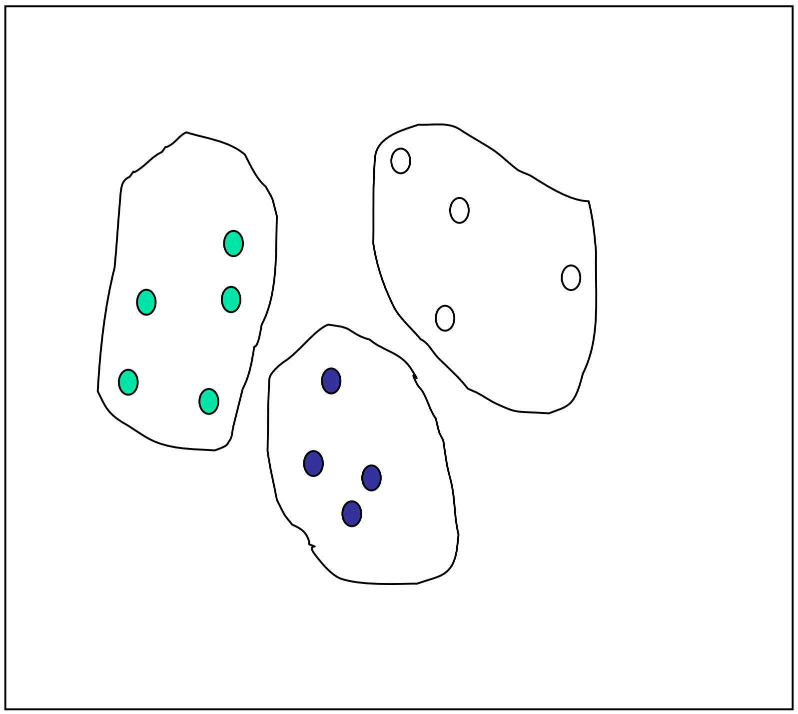


Stratified Sampling Illustration

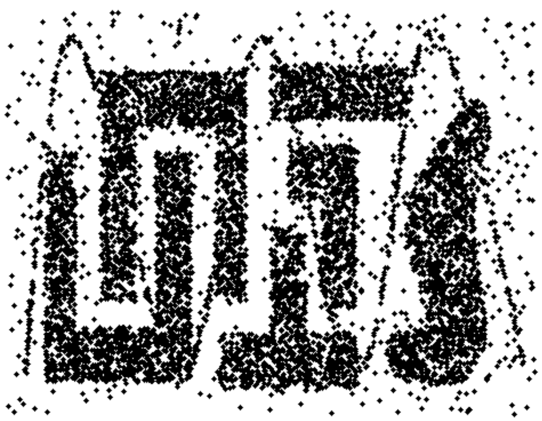
Raw Data



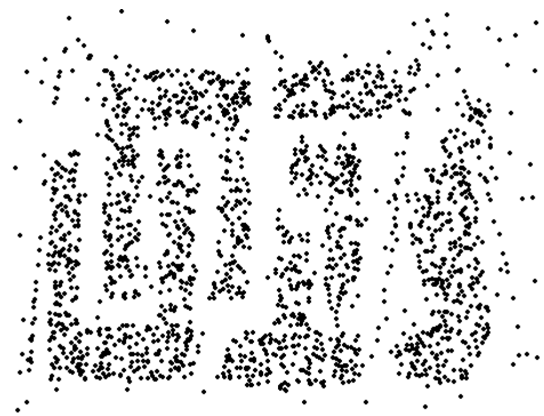
Stratified Sample



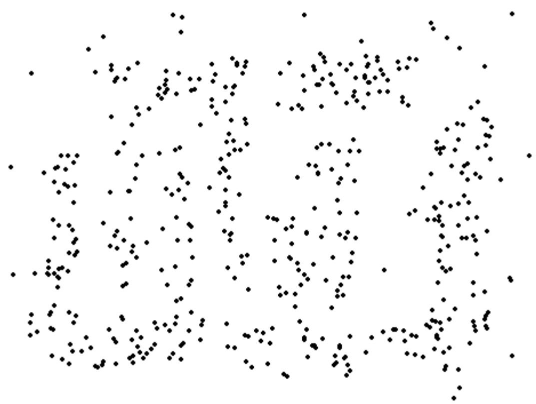
Sampling Size



8000 points



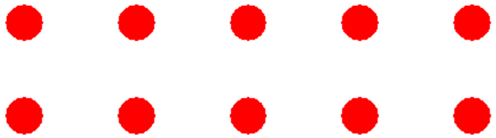
2000 Points

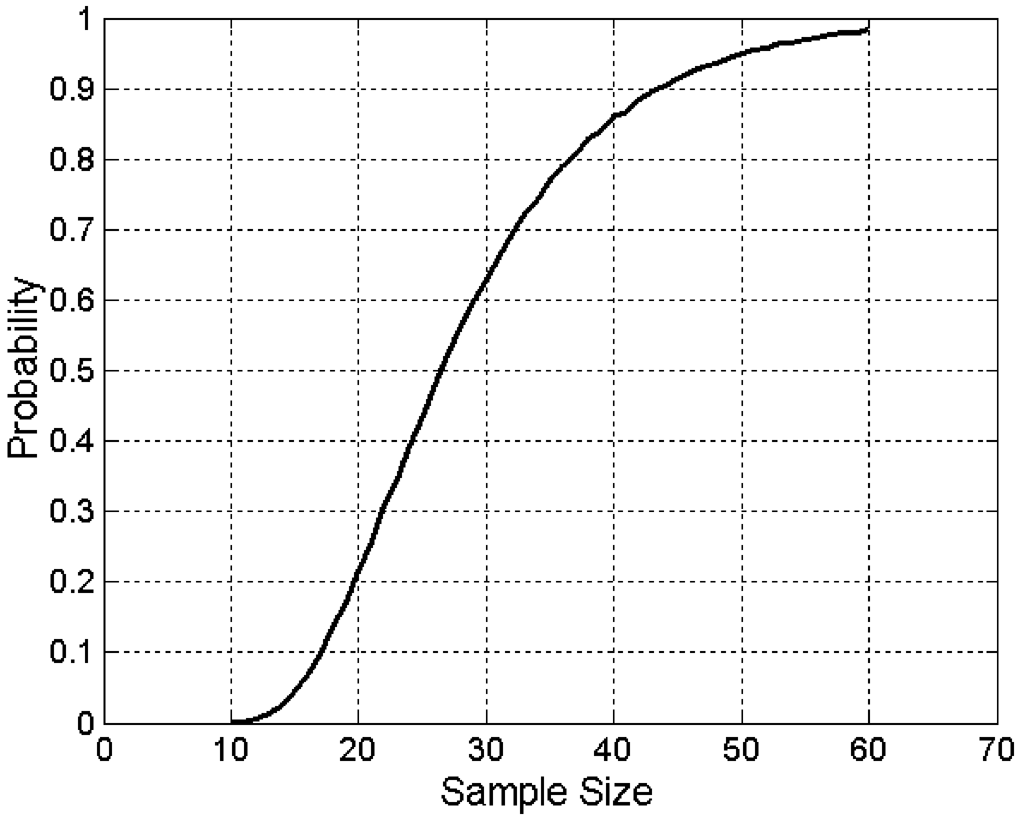


500 Points

Sample Size

- **What sample size is necessary to get at least one object from each of 10 groups.**





Data Reduction

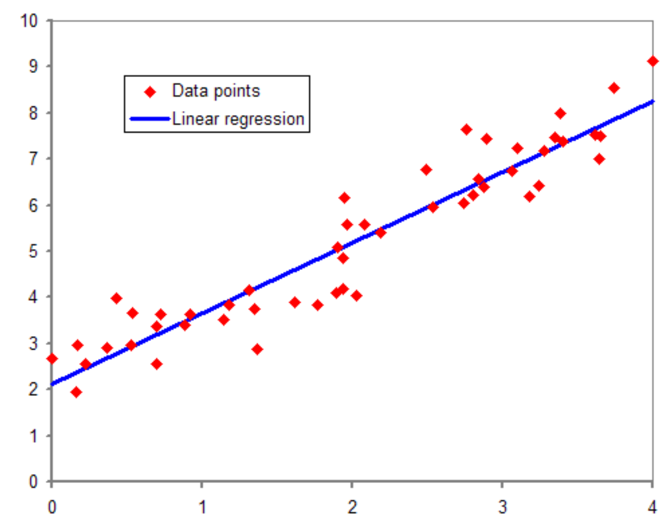
- Instance reduction
  - Sampling (instance selection)
  - Numerosity reduction
- Dimension reduction
  - Feature selection
  - Feature extraction

Numerosity 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)
  - Regression
- Non-parametric methods
  - Do not assume models
  - Histograms, clustering

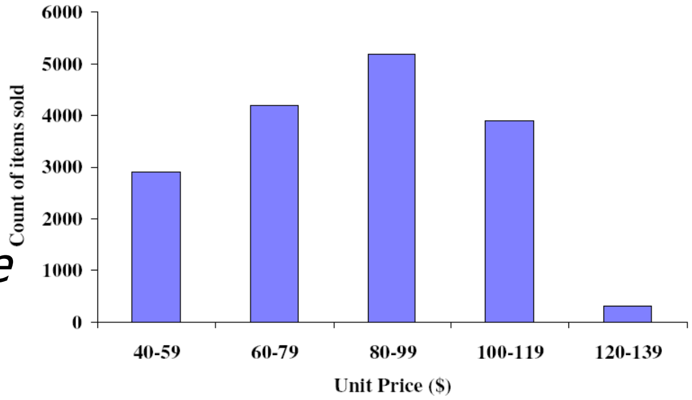
Regress Analysis

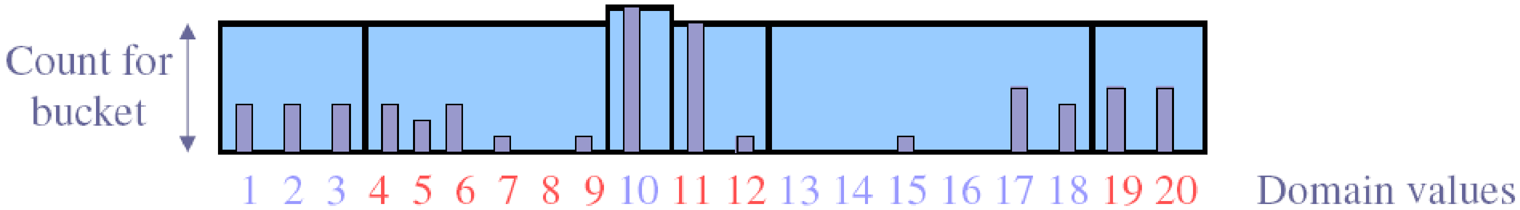
- Assume the data fits some model and estimate model parameters
- Linear regression: $Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_PX_P$
  - Line fitting: $Y = b_1X + b_0$
  - Polynomial fitting: $Y = b_2x^2 + b_1x + b_0$
- Regression techniques
  - Least square fitting
- Regression analysis will be studied in depth later for prediction

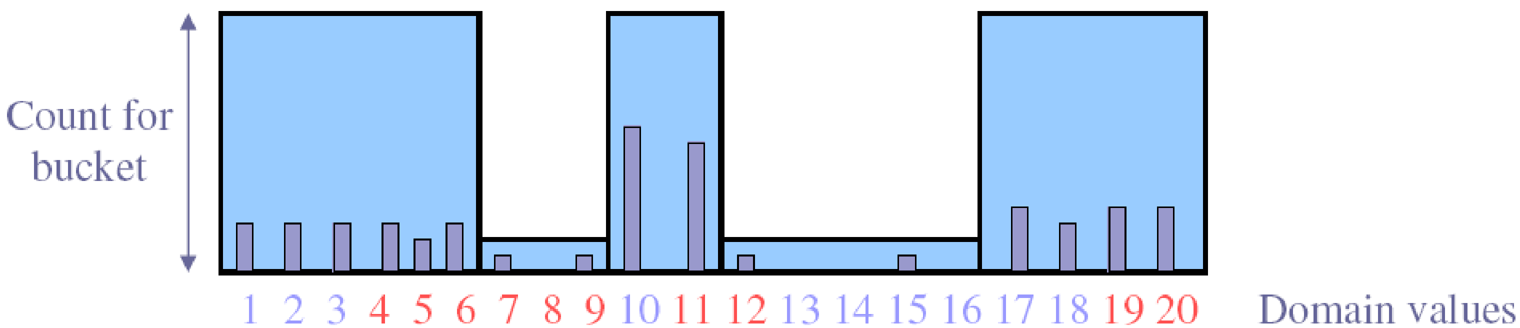


Instance Reduction: Histograms

- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
  - Equi-width: equal bucket range
  - Equi-depth: equal frequency
  - V-optimal: with the least *frequency variance*

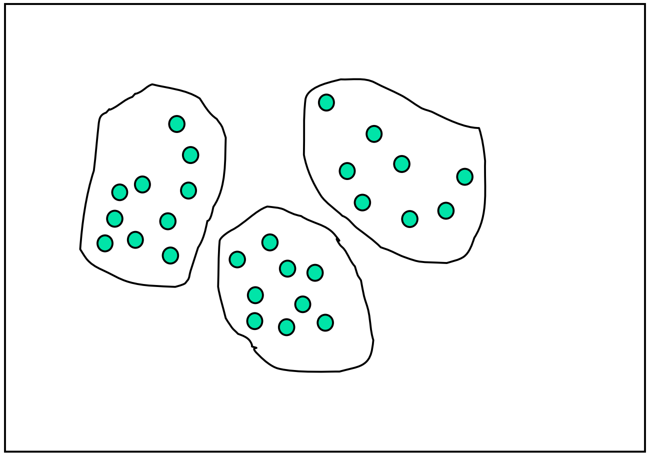






Instance Reduction: Clustering

- 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
- Cluster analysis will be studied in depth later



Data Reduction

- Instance reduction
  - Sampling (instance selection)
  - Numerosity reduction
- Dimension reduction
  - Feature subset selection
  - Feature extraction/transformation

Feature Subset Selection

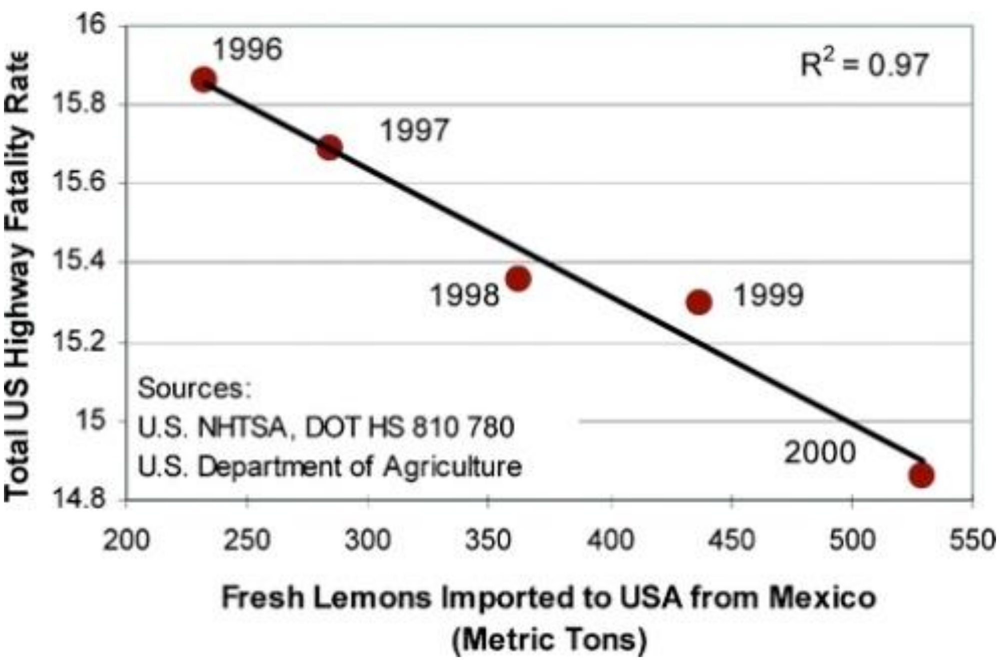
- Select a subset of features by removing irrelevant, redundant, or correlated features such that mining result is not affected

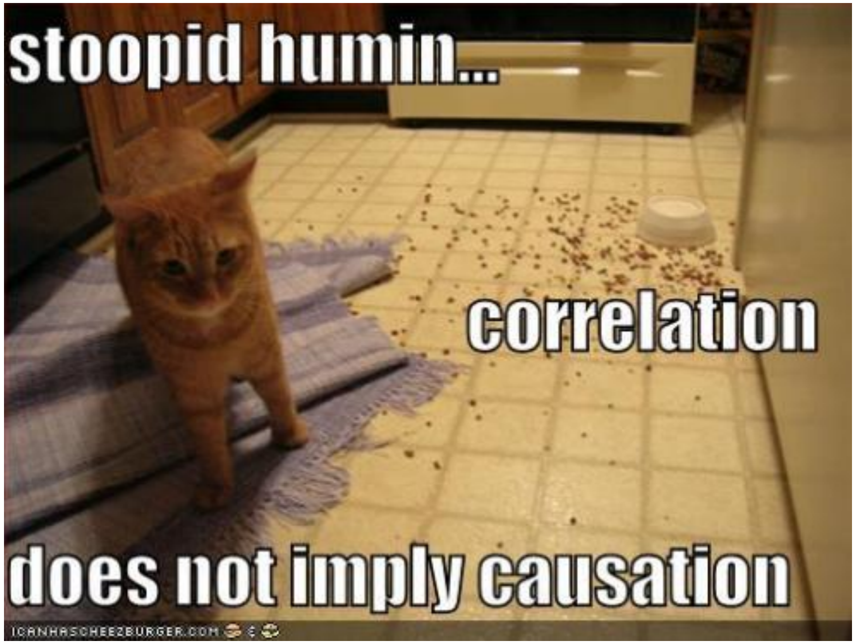
- Irrelevant features
  - contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA

- Redundant or correlated 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
  - Correlation analysis

Correlation between attributes

- Correlation measures the linear relationship between variables
  - Does not necessarily imply causality





Correlation Analysis (Numerical Data)

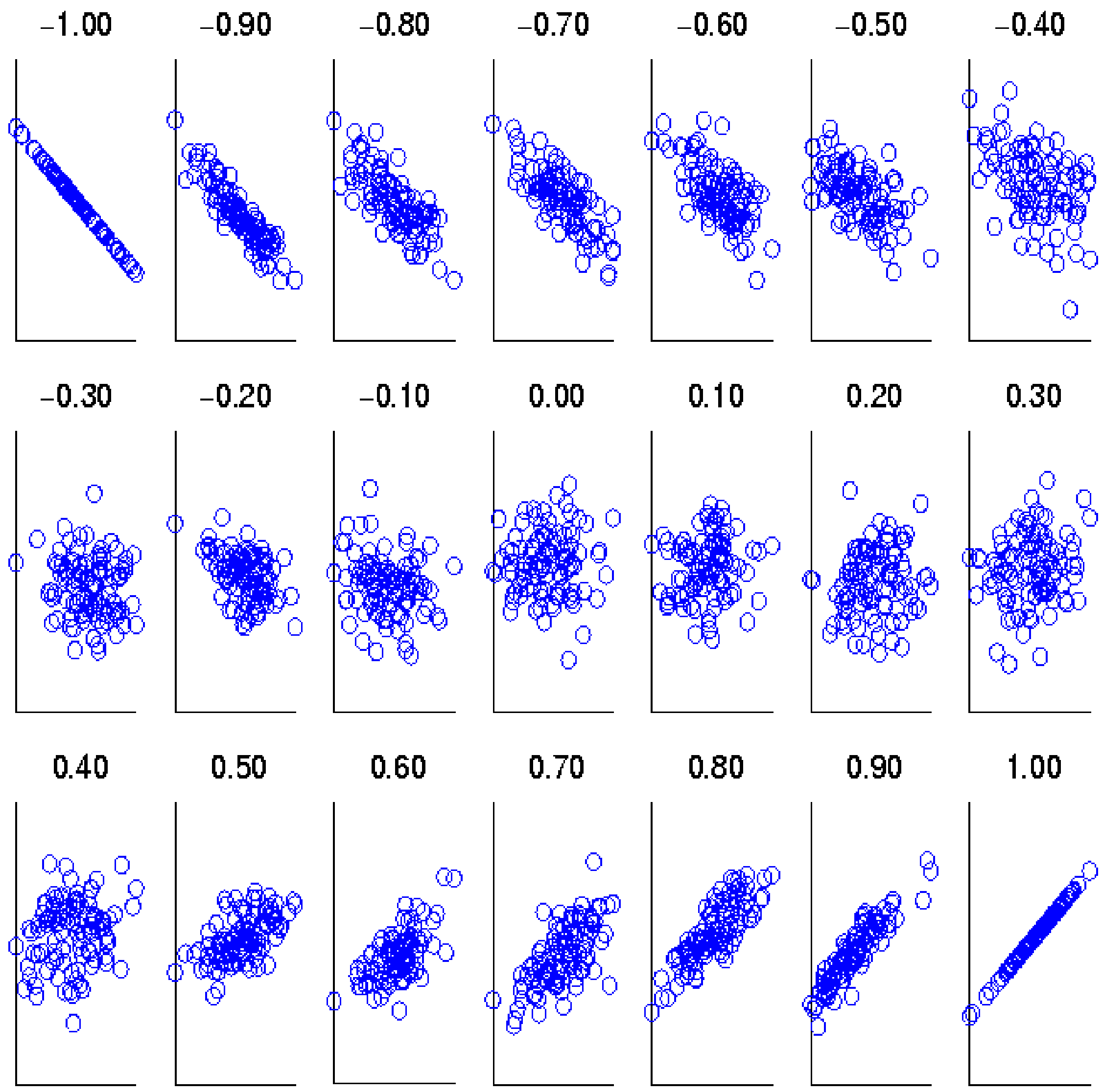
- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum (A-\bar{A})(B-\bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum (AB) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and Σ(AB) is the sum of the AB dot-product.

- $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's)
- $r_{A,B} = 0$: independent
- $r_{A,B} < 0$: negatively correlated

Visually Evaluating Correlation



Scatter plots showing the Pearson correlation from −1 to 1.

Correlation Analysis (Categorical Data)

- Contingency table of two attributes A and B
- X^2 (chi-square) statistic tests the hypothesis that A and B are *independent*

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The larger the X^2 value, the more likely the variables are related
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count

Chi-Square Calculation: An Example

	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 data distribution in the two categories)

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

- It shows that like_science_fiction and play_chess are correlated in the group (10.828 needed to reject the independence hypothesis at 0.0001 significance level)

Feature Selection

- Filter approaches:
  - Features are selected independent of data mining algorithm
  - E.g. Minimal pair-wise correlation/dependence, top k information entropy
- Wrapper approaches:
  - Use the data mining algorithm as a black box to find best subset
  - E.g. best classification accuracy
- Embedded approaches:
  - Feature selection occurs naturally as part of the data mining algorithm
  - E.g. Decision tree classification

Data Reduction

- Instance reduction
  - Sampling
  - Aggregation
- Dimension reduction
  - Feature selection
  - Feature extraction/creation

Feature Extraction

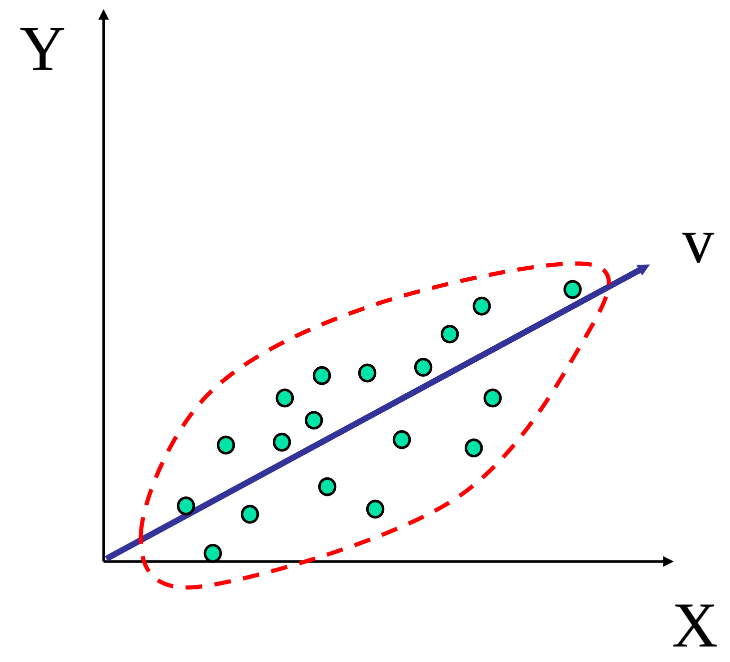
- Create new features (attributes) by combining/mapping existing ones
- Common methods
  - Principle Component Analysis
  - Singular Value Decomposition
- Other compression methods (time-frequency analysis)
  - Fourier transform (e.g. time series)
  - Discrete Wavelet Transform (e.g. 2D images)

Principal Component Analysis (PCA)

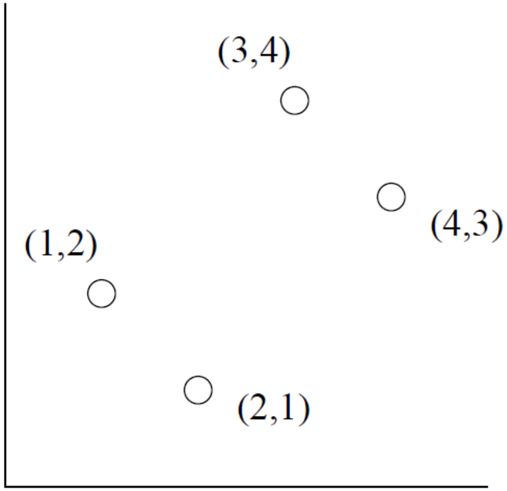
- Principle component analysis: find the dimensions that capture the most variance
  - A linear mapping of the data to a new coordinate system such that the greatest variance lies on the first coordinate (the first principal component), the second greatest variance on the second coordinate, and so on.
- Steps
  - 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"
  - Weak components can be eliminated, i.e., those with low variance

Dimensionality Reduction: PCA

- Mathematically
  - Compute the covariance matrix
    $$\mathrm{Cov}(X,Y) = \mathrm{E}\big[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\big],$$
  - Find the eigenvectors of the covariance matrix correspond to large eigenvalues $A\mathbf{v} = \lambda\mathbf{v}$.

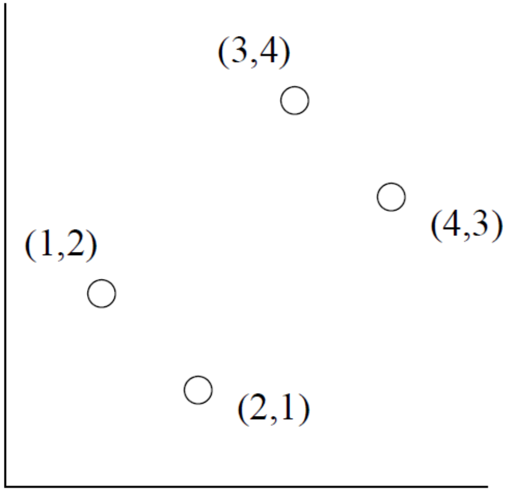


PCA: Illustrative Example



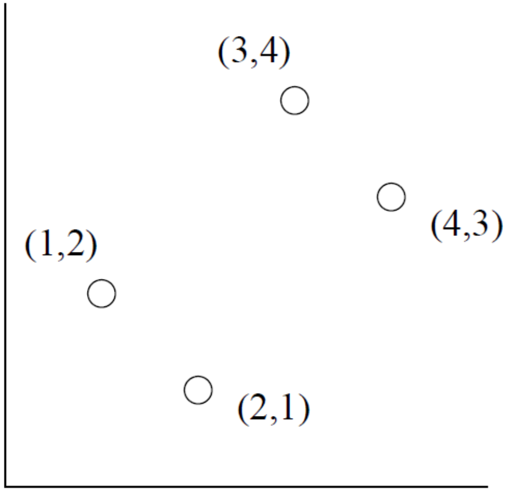
$$M = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix}$$

PCA: Illustrative Example



$$M = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \qquad M^{\mathrm{T}}M = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} 30 & 28 \\ 28 & 30 \end{bmatrix}$$

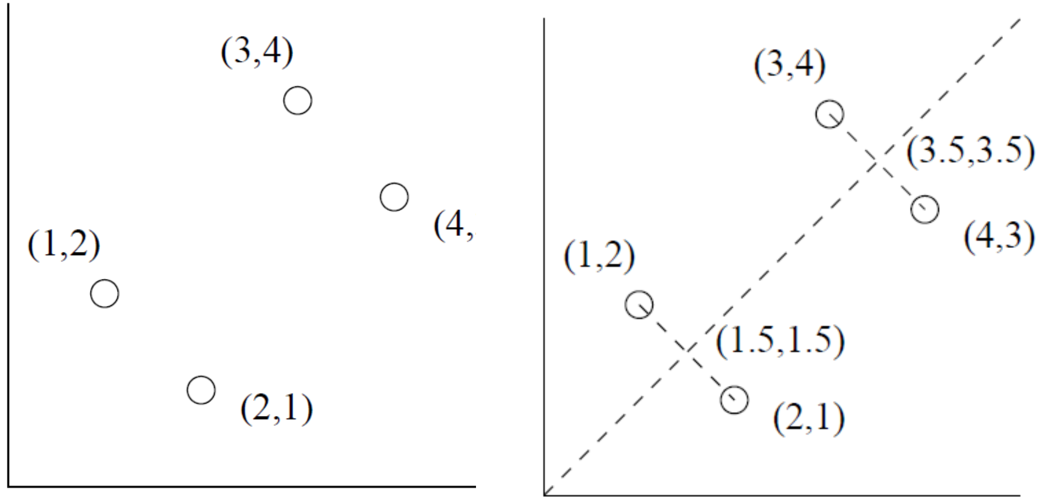
PCA: Illustrative Example



$$M = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \qquad M^{\mathrm{T}}M = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} 30 & 28 \\ 28 & 30 \end{bmatrix}$$

$$\lambda = 58 \text{ and } \lambda = 2 \qquad E = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

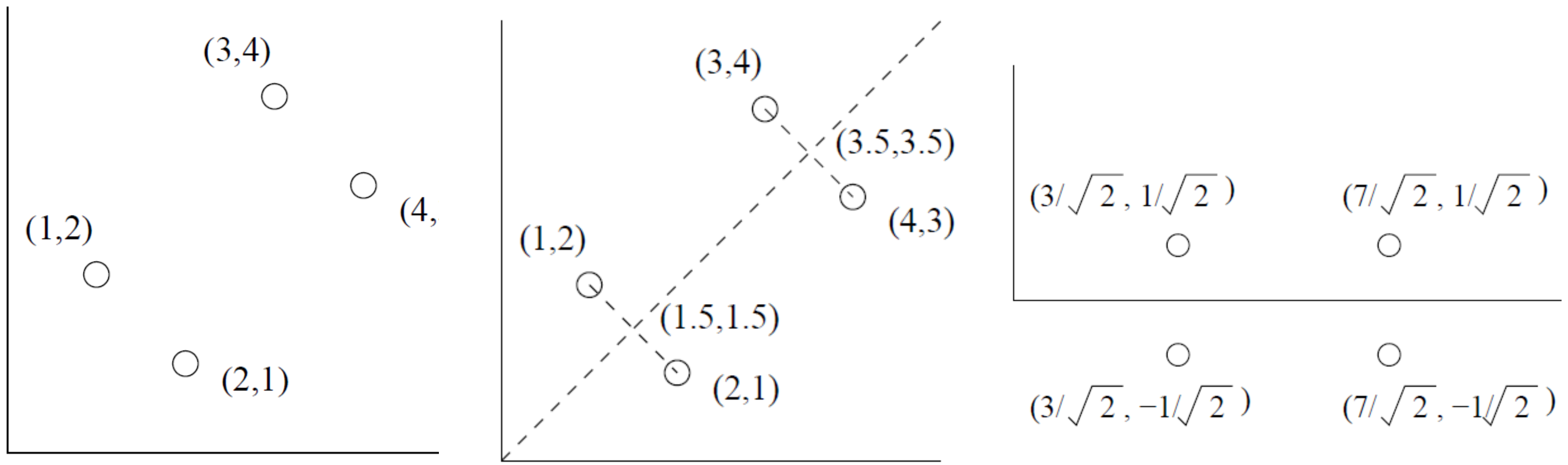
PCA: Illustrative Example



$$M = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \qquad M^{\mathrm{T}}M = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} 30 & 28 \\ 28 & 30 \end{bmatrix}$$

$$\lambda = 58 \text{ and } \lambda = 2 \qquad E = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

PCA: Illustrative Example



$$M = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix}$$

$$ME = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 3/\sqrt{2} & 1/\sqrt{2} \\ 3/\sqrt{2} & -1/\sqrt{2} \\ 7/\sqrt{2} & 1/\sqrt{2} \\ 7/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

Feature Extraction

- Create new features (attributes) by combining/mapping existing ones
- Common method
  - Principle Component Analysis
- Other compression methods (time-frequency analysis)
  - Fourier transform (e.g. time series)
  - Discrete Wavelet Transform (e.g. 2D images)