CS 412 Intro. to Data Mining

Chapter 3. Data Preprocessing

Qi Li Computer Science, Univ. Illinois at Urbana-Champaign, 2018

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
- Data Cleaning
- Data Integration
- Data Reduction and Transformation
- Dimensionality Reduction
- Summary

Why Preprocess Data?

- Raw data not ready to analyze
- Issues of data quality
- Conclusions drawn may be questionable or unreliable

Measures for data quality

- Accuracy: is the data correct or wrong, accurate or not?
- Completeness: is there missing data?
- Consistency: are there conflicts in the data?
- Timeliness: is data old or recently updated?
- Believability: can you trust that the data is correct?
- Interpretability: how easily can the data be understood?

Major Data Preprocessing Tasks

- **Data cleaning**
  - Handle missing data, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
  - Often involves resolving conflicts between data sources
- **Data reduction and transformation**
  - Speeds up analysis when data is *too* big
  - E.g., can reduce rows (data points) or columns (attributes) of matrices

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
- Data Cleaning
- Data Integration
- Data Reduction and Transformation
- Dimensionality Reduction
- Summary

Data Cleaning

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

Data Cleaning, continued

- Inconsistent: containing discrepancies in codes or names, e.g.,
  - *Age* = "42", *Birthday* = "03/07/2010"
  - Different data formats, e.g., rating "1, 2, 3" is now "A, B, C"
  - Discrepancy between duplicate records
- Intentional: (e.g., *disguised missing* data)
  - Defaults: Jan. 1 as everyone's birthday?

Incomplete (Missing) 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 were not entered due to misunderstanding
  - Certain data may not be considered important at the time of entry
  - Did not register history or changes of the data
- Missing data may need to be inferred

How to Handle Missing Data?

- Ignore the tuple
  - Often not desirable, can cause data set to shrink dramatically
- Fill in the missing value manually
  - Tedious + infeasible?
- Fill in it automatically with
  - 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 such as Bayesian formula or decision tree**

Handling Missing Data: Example

- Want to predict likely value for missing data
- Example: Student missing data for final course grade
  - This student is male, age 33, 4.0 GPA
  - Find similar people in the data and see what their value for final grade is
  - Fill missing spot with most likely final grade based on the other data

Noisy Data

- **Noise:** random error or variance in a measured variable
- **Incorrect attribute values** may be due to various reasons
  - Faulty data collection instruments, Data entry problems, Data transmission problems, Technology limitation, Inconsistency in naming convention, …
- **Other data problems**
  - Outliers
  - Duplicate records
  - Incomplete data
  - Inconsistent data

How to Handle Noisy Data?

- Want to detect and (possibly) remove outliers
  - **Binning**
    - Sort data and partition into bins
    - Can smooth by bin means, bin median, bin boundaries, etc.
  - **Regression**
    - Smooth by fitting the data into regression functions
  - **Clustering**
    - Group data so that that points in the same cluster are more similar to each other than to those in other clusters
  - **Semi-supervised:** Combined computer and human inspection
    - Detect suspicious values and have humans check

Data Cleaning as a Process

- Tools and guidelines exist to help with data cleaning
- **Not a one-pass task**
  - Often requires multiple rounds of identifying problems and resolving them

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
- Data Cleaning
- Data Integration
- Data Reduction and Transformation
- Dimensionality Reduction
- Summary

Data Integration

- Data integration – What is it?
  - Combining data from multiple sources into a coherent store
- **Schema integration**:
  - e.g., A.cust-id $\equiv$ B.cust-#
  - Integrate metadata from different sources
- **Entity identification:**
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton

Data Integration – Why?

- Why data integration?
  - Clarifies data inconsistencies/Noise
    - Example: Age and Date of Birth.
      - **Database 1 (Google)**: 02/26/19**08**; Age 38,
      - **Database 2(Wikipedia)**: 02/26/1980; Age 38
        - Data from Database 2 clarifies the error in Year of Birth
  - Fills in Important Attributes for Analysis
    - Merging from more than 1 dataset provides more important information.
  - Speeds up Data Mining
    - One Master Schema can be mined rather than each of the 10 one-by-one

Data Integration- Challenges

- What problems will you face?
  - Schema differences
    - Column is called "PersonAge" from Customer Table
    - Column is called "CustomerAge" from Person Table
  - Data Value Representation Conflicts
    - Database 1 -> "William Clinton"
    - Database 2 -> "Bill Clinton"
  - Bad Data
    - Typo; Wrong recording
    - Different Scales/Units for Data Type (£, $, or €)

Data Integration - Handling Noise

- Detecting data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: no reason, different representations, different scales, e.g., metric vs. British units
- Resolving conflict information
  - Take the mean/median/mode/max/min
  - Take the most recent
  - Truth finding (Advanced): consider the source quality
- Data cleaning should happen <u>again</u> after data integration

Data Integration - Handling Redundancy

- Redundant data often occurs when multiple databases are integrated
  - *Object identification / Entity Matching*: The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a "derived" attribute in another table, e.g., annual revenue
- What's the problem?
  - $Y = 2X \rightarrow Y = X_1 + X_2 \quad Y = 3X_1 - X_2 \quad Y = -1291X_1 + 1293X_2$
    - Y equal to 2X in one DB, Y equal to sum of > 1 variable in another.
- Redundant attributes may be detected by correlation analysis and covariance analysis

Example: stock market

Yahoo! Finance

Day's Range: 93.80-95.71

Nasdaq

Green Mountain Coffee Roasters, (NasdaqGS: GMCR)

After Hours: 95.13 ↓ -0.01 (-0.02%) 4:07PM EDT

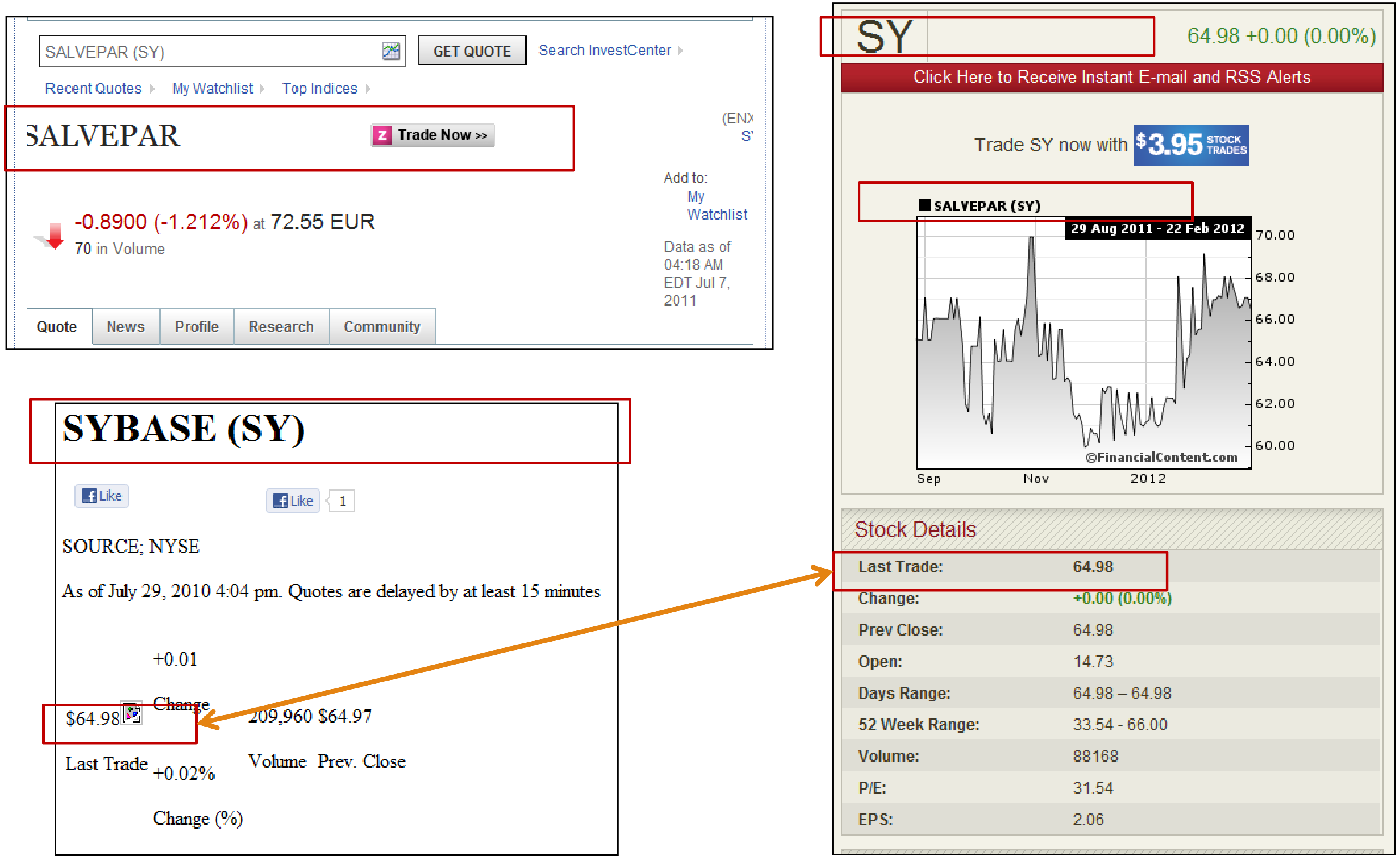
Last Trade:	**95.14**	Day's Range:	93.80 - 95.71
Trade Time:	4:00PM EDT	52wk Range:	25.38 - 95.71
Change:	↑ 1.69 (1.81%)	Volume:	2,384,075
Prev Close:	93.45	Avg Vol (3m):	2,512,070
Open:	94.01	Market Cap:	13.51B
Bid:	95.03 x 100	P/E (ttm):	119.82
Ask:	95.94 x 100	EPS (tt	0.79
1y Target Est:			N/A (N/A)

52wk Range: 25.38-95.71

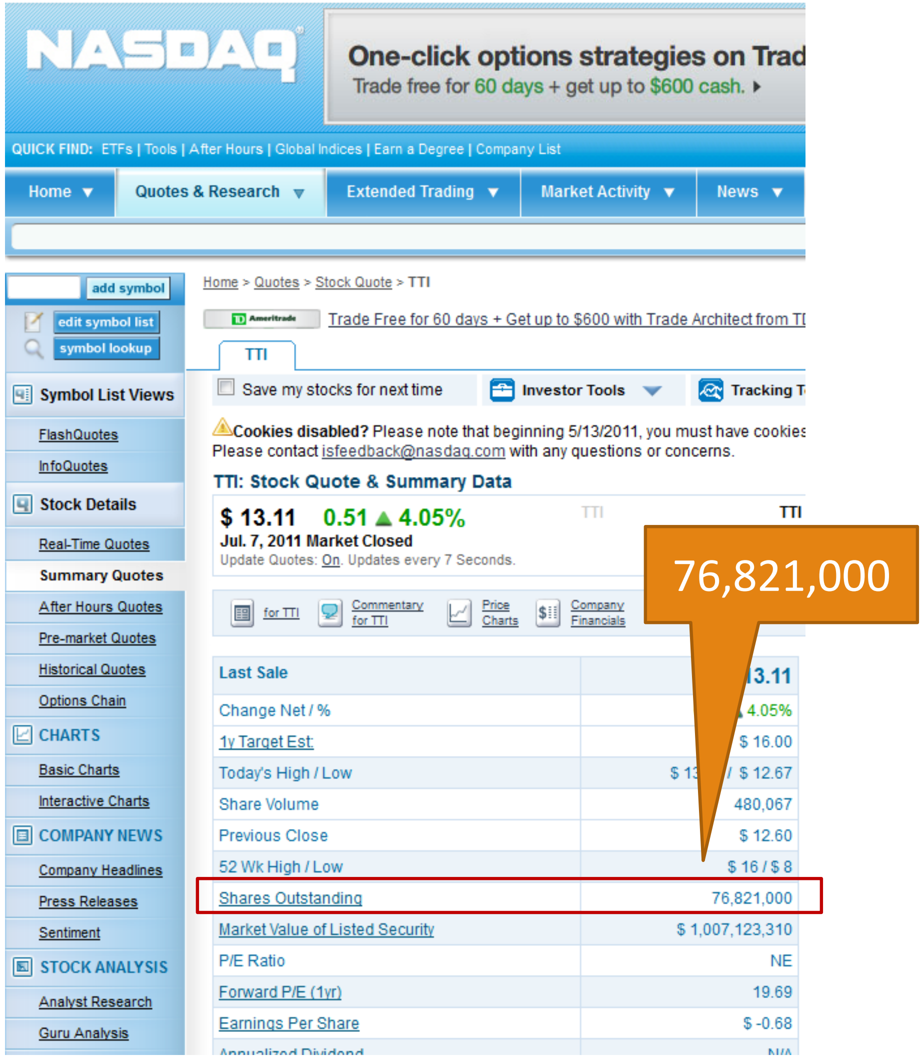
52 Wk: 25.38-93.72

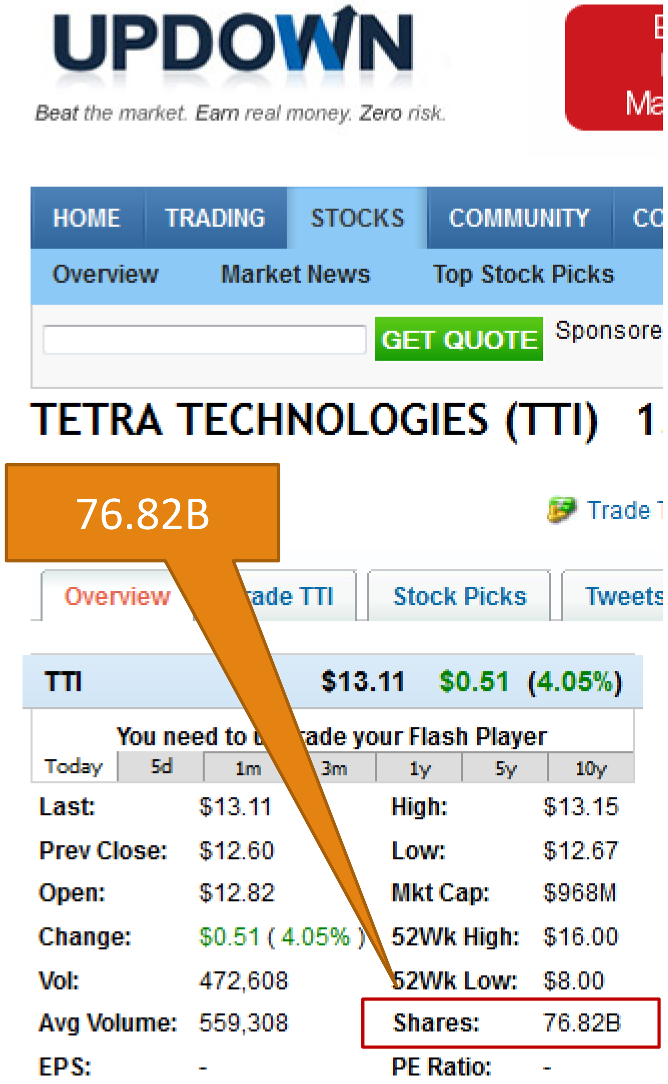
Last Sale	**$ 95.14**
Change Net / %	1.69 ▲ 1.81%
Best Bid / Ask	$ 95.03 / $ 95.94
1y Target Est:	$ 95.00
Today's High / Low	$ 95.71 / $ 93.80
Share Volume	2,384,175
50 Day Avg. Daily Volume	2,751,062
Previous Close	$ 93.45
52 Wk High / Low	$ 93.72 / $ 25.38
Shares Outstanding	152,785,000
Market Value of Listed Security	$ 14,535,964,900
P/E Ratio	120.43
Forward P/E	63.57
Ea er Share	$ 0.79
nualized Dividend	N/A
Ex Dividend Date	N/A
Dividend Payment Date	N/A
Current Yield	N/A
Beta	0.82
NASDAQ Official Open Price:	$ 94.01
Date of NASDAQ Official Open Price:	Jul. 7, 2011
NASDAQ Official Close Price:	$ 95.14
Date of NASDAQ Official Close Price:	Jul. 7, 2011

Example: stock market

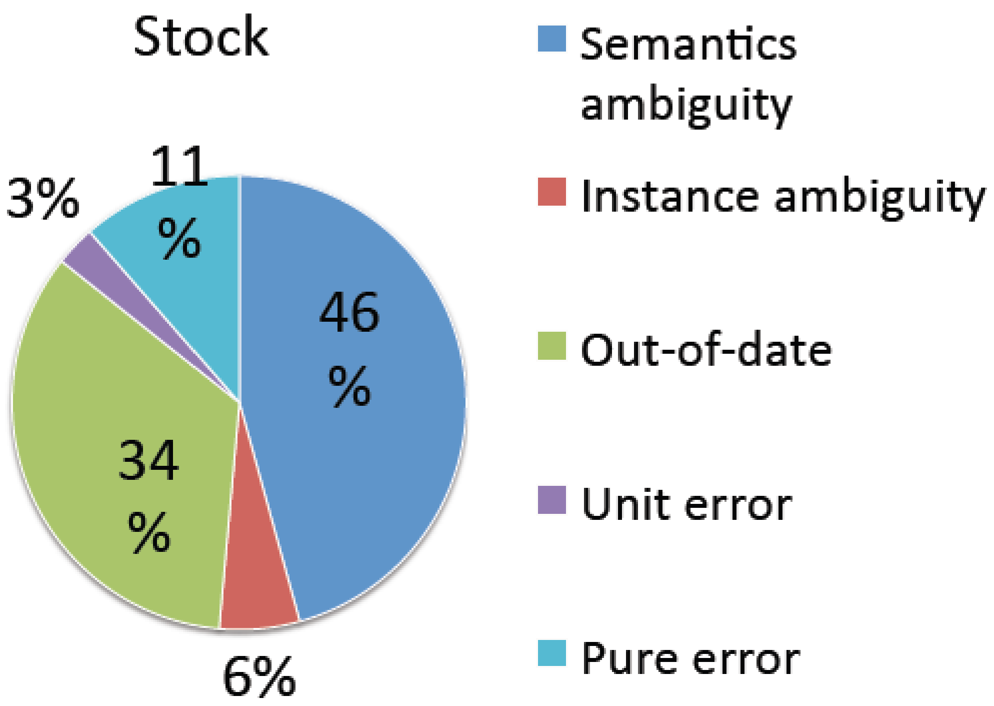


Example: stock market





Example: stock market



Source	Accuracy	Coverage
Google Finance	.94	.82
Yahoo! Finance	.93	.81
NASDAQ	.92	.84
MSN Money	.91	.89
Bloomberg	.83	.81

Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. Truth finding on the Deep Web: Is the problem solved? In *VLDB,* 2013.

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
- Data Cleaning
- Data Integration
- Data Reduction and Transformation
- Dimensionality Reduction
- Summary

Data Reduction

- Data reduction (https://en.wikipedia.org/wiki/Data_reduction):
  - The transformation of data into a simplified or more meaningful form.
- Why data reduction?
  - Data is too big, which causes analysis to be a pain.
    - Example: Large Gigabytes of Data: Forced to chunk/analyze 1 chunk at a time, which is time consuming.

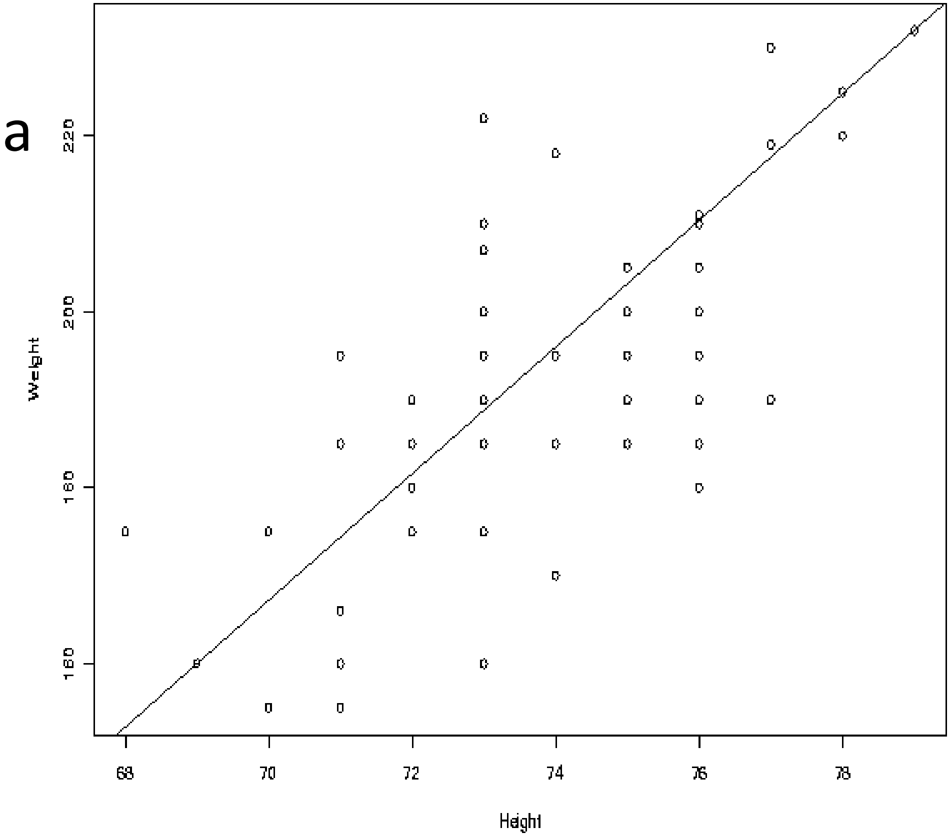
Data Reduction – Row and Column

- Smart Data Reduction
  - Attribute Elimination (Column Reduction)
    - Throw away useless attributes, **not** random ones.
      - Example: Predict if patient will get Disease A
        - Throw out “hasASibling” attribute, and keep “siblingHasDiseaseA” attribute.
  - Entity Elimination (Row Reduction)
    - Example: Find citizens income. Do you need everyone’s income to do this analysis? No
      - Smart Reduction

Data Reduction: Parametric

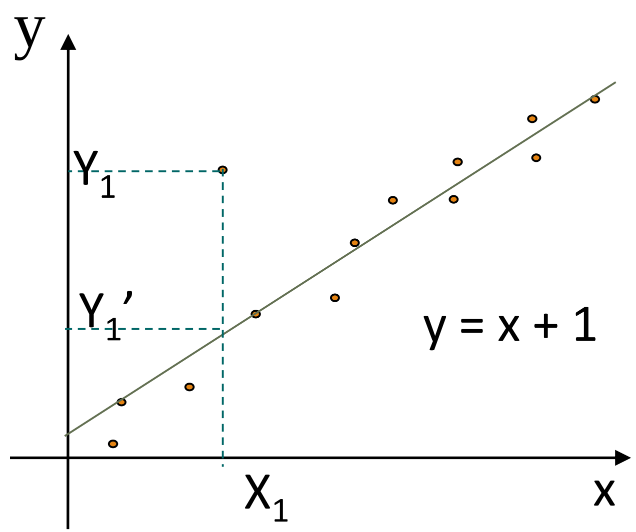
- **Parametric Method (https://en.wikipedia.org/wiki/Parametric_model)**
  - Applying an ***assumed model*** onto data in order to simplify and add meaning.
  - Example: If you want to analyze data related to a population's height and weight.
  - Assumption: Linear Regression Model
    - Estimate the parameters to model the data:
      - y=mx+b
    - The equation replaces and simplifies the data

Height vs. Weight



Terminology: Regression

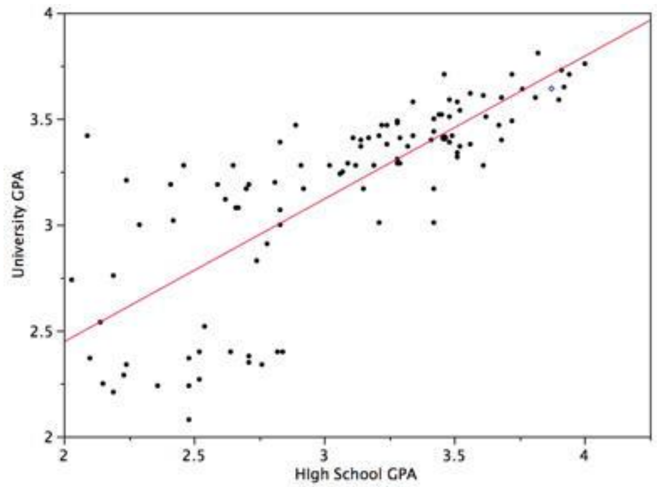
- ***Dependent variable*** (also called ***response variable***)
  - *Plotted on "Y" Axis*
- **Independent variables** (also known as ***explanatory variables*** or ***predictors***)
  - Plotted on "X" axis. The variable that is manipulated.
- The parameters are estimated so as to give a "**best fit**" of the data
- Most commonly the best fit is evaluated by the ***least squares method***, but other criteria have also been used

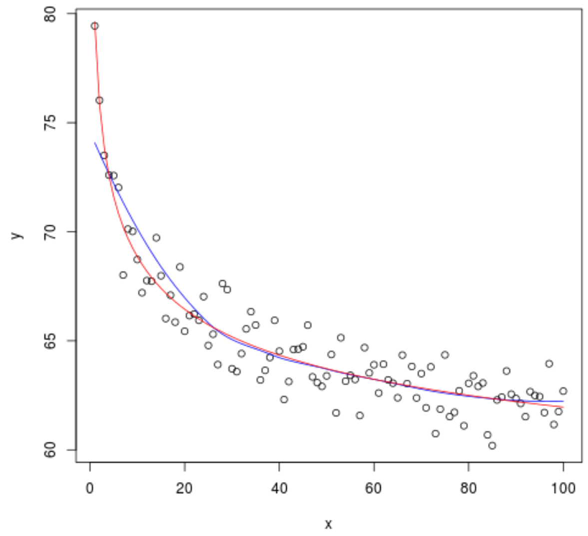


- Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

Parametric Regression Types

- Linear regression:
  - Y = wX + b is one form of linear regression
  - Find w and b to minimize the least squared of the errors
  - Data modeled to fit a straight line
- Nonlinear regression:
  - The function you want to fit is nonlinear
  - The data are fitted by a method of successive approximations. Example: Y = exp(wx + b)
- Linear and Nonlinear
  - Both are parametric methods and both need to have an assumption

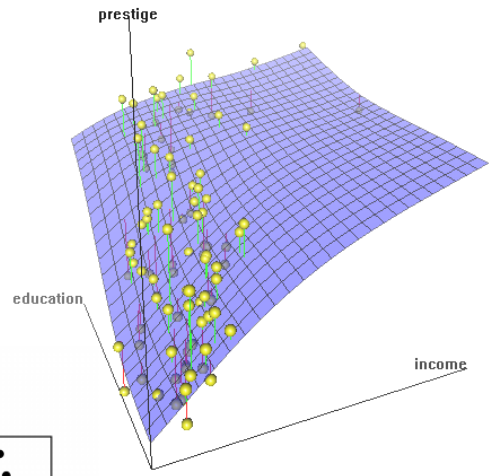




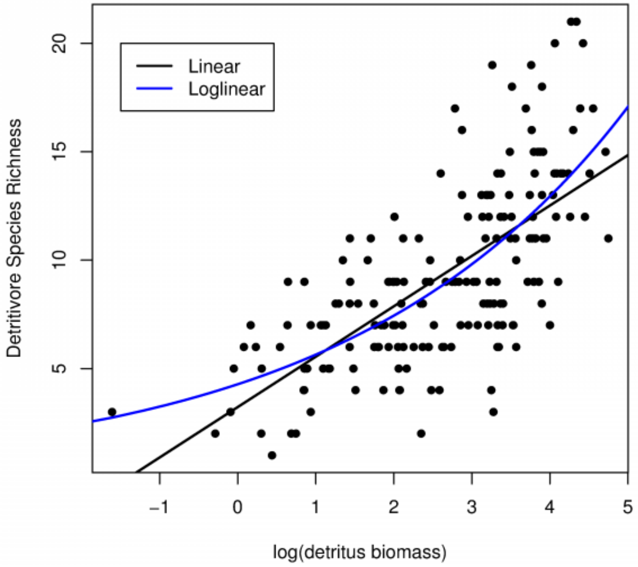
Multiple Regression and Log-Linear Models

- Multiple regression: $Y = b_0 + b_1 X_1 + b_2 X_2$
  - More than one dependent variable.
  - Example: House Price is dependent on location **and** size **and** the number of bedrooms.
  - Still a linear combination /model.
- Log-linear model:
  - Another very popular category of regression
  - Will talk about this in future lectures.

Multiple regression



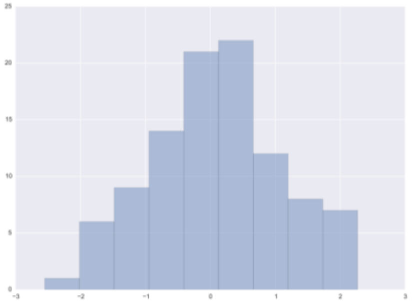
Log-linear model



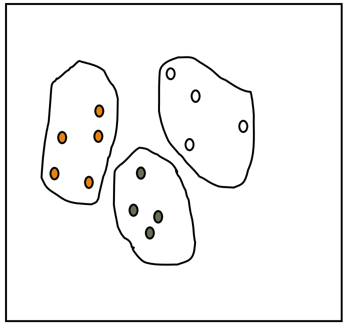
Data Reduction: Non-Parametric

- **Non-Parametric Method**
  - Do not assume what kind of model is best.
  - Major families: histograms, clustering, sampling, etc.
  - Good for when you don't know much about the data.
  - Algorithms: k-nearest neighbors, decision tree,…

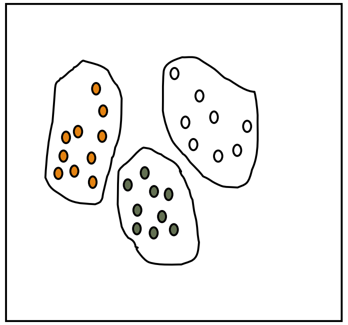
Histogram



Clustering on the Raw Data

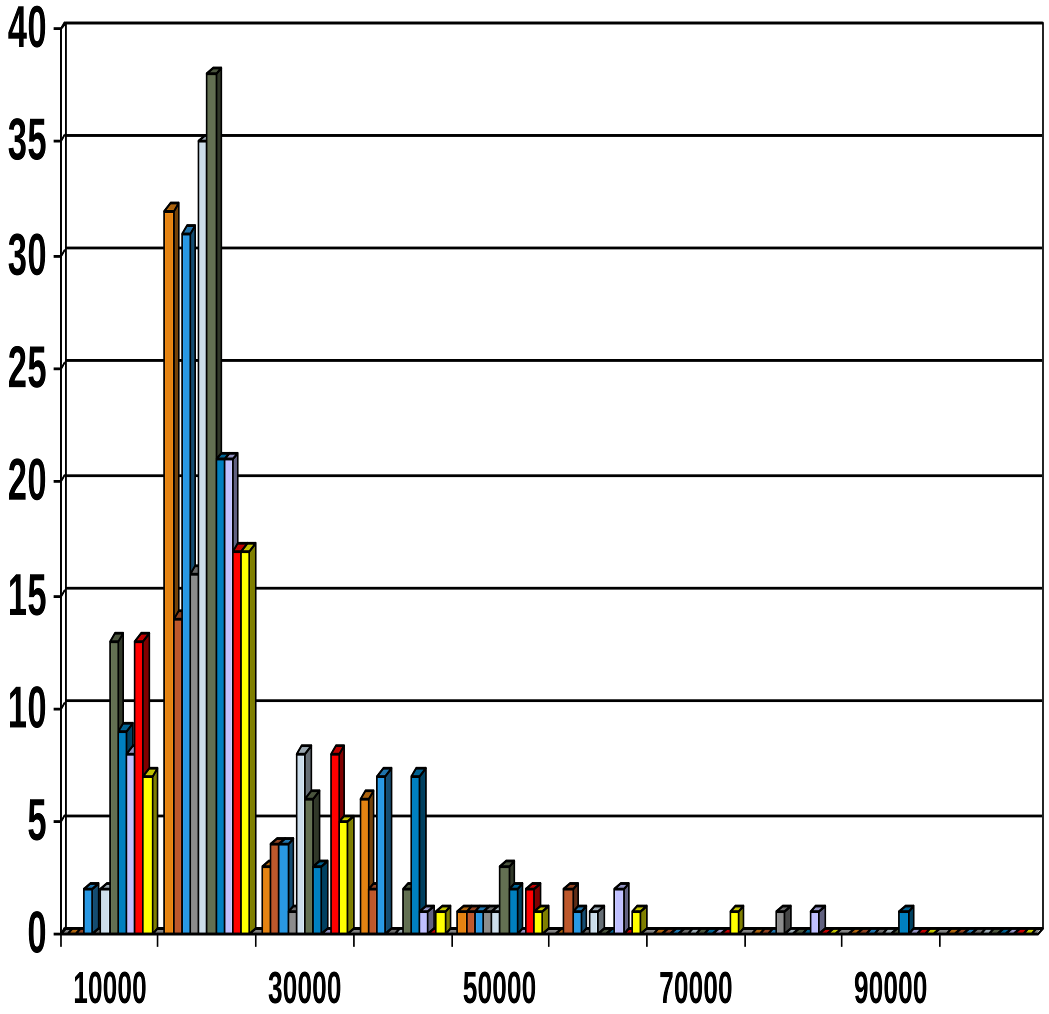


Stratified Sampling



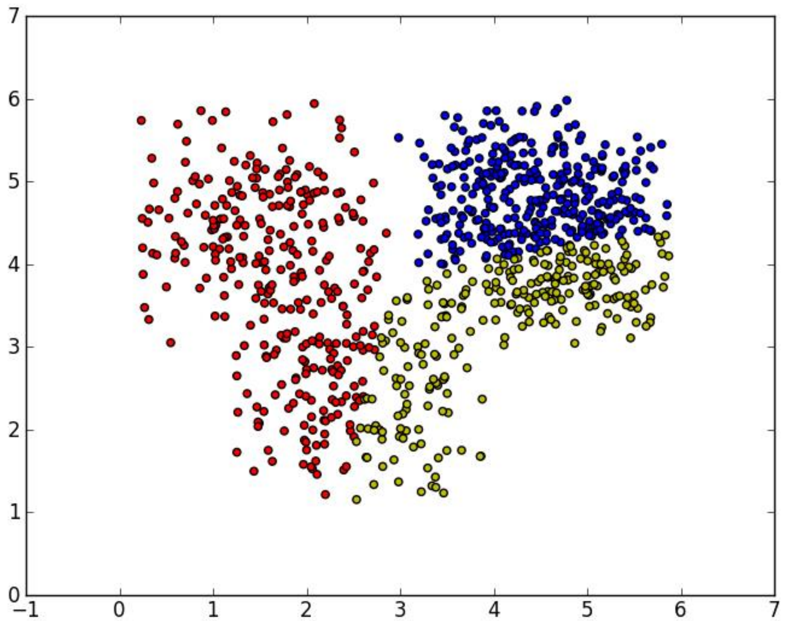
Non-Parametric: Histogram

- This is a typical Non-Parametric Method
  - i.e: it doesn’t assume how the data is distributed.
- Divide data into buckets.
- Several ways to do binning:
  - Partitioning rules:
    - Equal-width: equal bucket range
    - Equal-frequency (or equal-depth)



Clustering

- Can be parametric (e.g. GMM) or non-parametric (e.g. hierarchical clustering)
  - Based on the model chosen
- Key idea: put the data into different groups.
- How: data within a group should be close to one another, and data from different groups to be more different.
- If the goal is met, then when you do sampling it is easier to choose what would be the representative items to represent the distribution of the data.

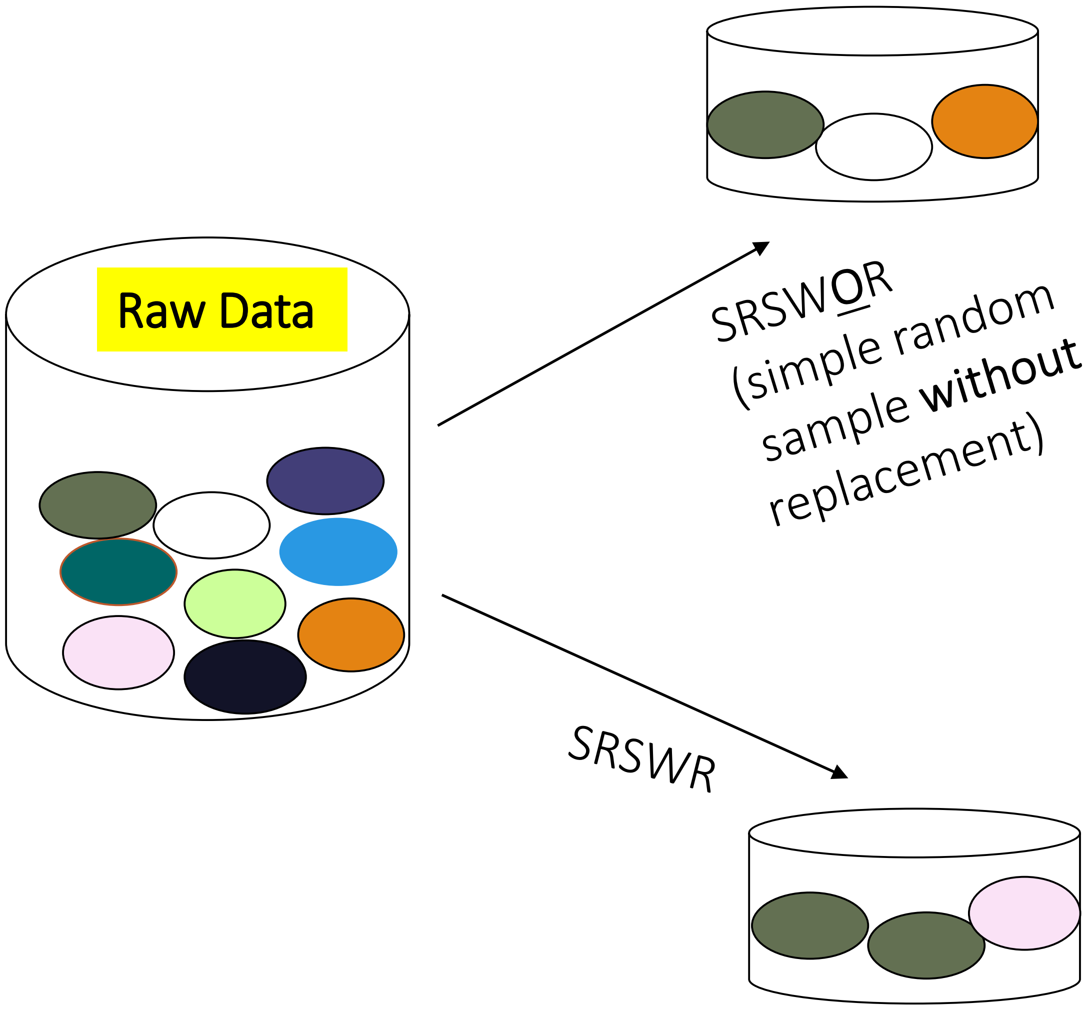


Data Reduction: Sampling

- Sampling: obtaining a small sample ***s*** to represent the whole data set ***N***
- Key principles:
  - Choose a **representative** subset and of the data
  - Choose an appropriate size of the sample
- Example: Presidential election
  - Predicting which candidate will win the final election
  - Can not pick 1000s, you need more
  - Can not only sample people from NY. You need multiple regions

Data Reduction: Random Sampling

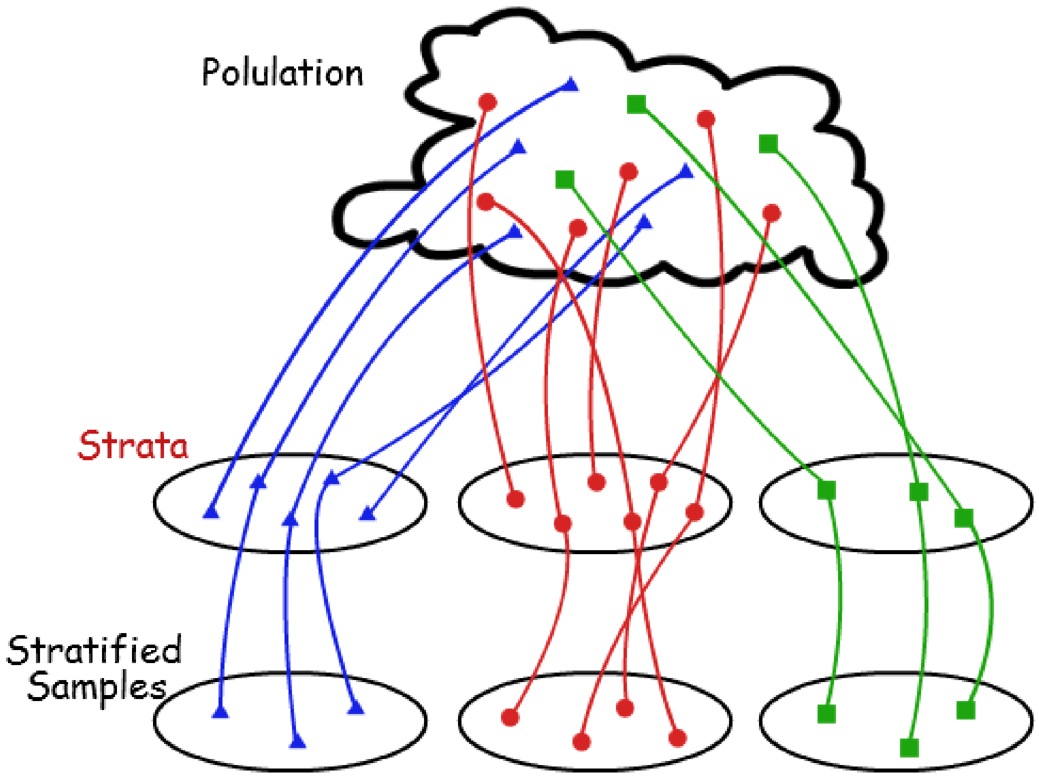
- **Random sampling**
  - Pick sample population at random.
  - Example: Go to the street, asking people, "Do you like Starbucks coffee?"
- **Random Sampling - Types**
  - Without replacement
    - Once an object is selected, it is removed from the population
  - With replacement
    - A selected object is not removed from the population



Random Sampling (Continued…)

- **Stratified Random Sampling**
  - Used if you have a basic understanding of the data.
  - Example: Presidential Election
    - I don’t need to spend time in NY already know what will happen in that state.
    - Only spend time in the “Swing” States.
- Need the same percentage of the original data
  - Example: Male and Female in CS department.
  - Know that there are more males than females. Want to keep same percentage. The sample should have the same ratio.

Stratified sampling

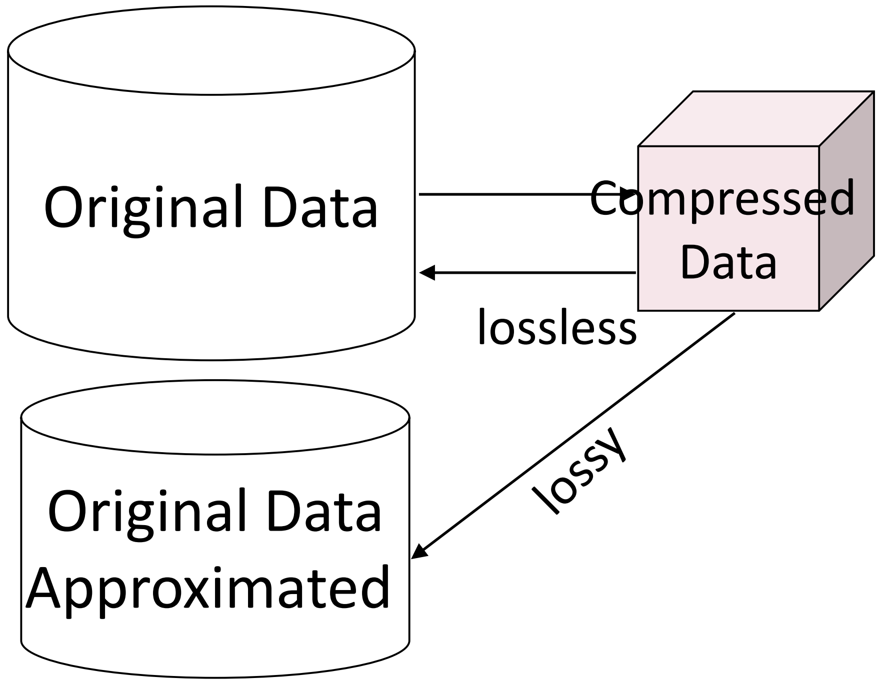


Data Reduction: Aggregation/Summation/Data Cube

- Data aggregation is any process in which information is gathered and expressed in a summary form.
- Also known as Summation or a Data Cube.
- Example: Handling transactional data from Walmarts
  - You get records for every customer's visit, and all the details of those transactions.
  - If want to find out what type of beer is more popular in the market, then you ***<u>don't</u>*** need to go to ***each individual's transaction*** one-by-one.
  - A summarization is good enough.
  - **NOTE**: Future lectures go into aggregation in very much more detail.

Data Compression : Terminology

- Very effective way to reduce the size of your data
- Types of Compression:
  - Lossy
    - Original Data is Approximated
  - Lossless
    - Original Data can be recovered
- Next Slide…(Examples)



Lossy vs. lossless compression

Data Compression : Lossless vs Lossy

- Lossy
  - Example: Phone Photo Sample Image
    - A small photo, less space, gives basic information.
    - If you think its important, you click download and you get all the data from the web for this image.
    - The small photo is lossy compression of the bigger image.
- Lossless
  - Example: Zip file.
    - On unzip, you get the original data. No Data is lost, its just in a different format.
  - Example: 10k x 10k matrix of 1s and 0s
    - Of only 100 are "1", then only store the location of 1s.
    - Can easily reconstruct the matrix.   No information is lost.

Data Transformation

- What is it?
  - A function that maps the entire set of values of a given attribute to a new set of replacement values
  - Each old value can be identified with one of the new values
- When do I normalize?
  - If your data already has the same meaning, there there is no need to normalize.
    - Example: Two data sets: High Temperature, Low Temperature
  - If they have different meanings, then you should normalize to compare them.
    - Example: Age vs Income.
      - These are hard to compare because of the range of values in Age is 0-99 , and income maybe is 12k – 100k, etc.

Normalization – Min/Max

- **Maps data into a bounded range of your choice**
  - Standardizes /Smooths the data value range
  - Smoothed ranges can then be easily compared to each other.
- **Min-max normalization**: 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 range \$12,000 to \$98,000 normalized to [0.0, 1.0]
    - Then \$73,000 is mapped to

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

Normalization – Z-Score

- Used to compare how each data point in a data set compares to the mean
  - Maps data to a standard normal curve
  - Answers the questions: Where is the center? / Where are the most values located? / Where are outliers?
  - If value is < -3 or > 3, then it may be an outlier in the data set
    - More analysis can then be done to find out *why* it is an outlier.
- **Z-score normalization** (μ: mean, σ: standard deviation):

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

Z-score: The distance between the raw score and the population mean in the unit of the standard deviation

  - Ex. Let μ = 54,000, σ = 16,000. Then

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

Normalization – Decimal Scaling

- Very simple technique to **scale data points** to make them comparable to other data points.
- Each data point is scaled by a factor of 10
- **Normalization by decimal scaling**

$$v' = \frac{v_n}{10^j}, j = 4$$

→ Where j is the smallest integer such that Max(|v'|) <=1

$$A_{dataset} = \{v_1, v_2, v_3\} = \{1000, 2000, 10000\}$$
$$A'_{dataset} = \{0.1, 0.2, 1.0\}$$

Discretization

- One type of transformation to reduce data for most popular data types (many more):
  - Nominal
  - Ordinal
  - Numeric
- Discretization is most useful when reducing **Numeric** attribute types
  - Example: A large data set with many employee salaries ($)
    - Bucket the salaries into ranges/***categories***:
      - < 60k; [60k, 90k]; 90k to 1 Million
      - The size of data is reduced dramatically
      - Often, it is the ***category*** that really matters in the analysis anyway, not the individual values.

Data Discretization Methods

- Unsupervised / Top-down Split
  - Binning
  - Histogram analysis
  - Clustering analysis
- Unsupervised / bottom-up merge
  - Clustering analysis
  - Correlation (e.g., χ^2) analysis
- Supervised / top-down split
  - Decision-tree analysis
- Note: All the methods can be applied recursively

- **Top-down Split**
  - Take all the data at once, and bin into smaller groups
- **Bottom-up Merge**
  - Start with a small group of items, merge each individual with nearest neighbor.
  - Then merge the new groups with their nearest neighbor, etc.

Simple Discretization: Binning

- **Equal-width Binning**
  - Bin intervals are the same
  - Example: If the final grades in class range from 60 to 100
    - Interval for each grade will be the same (10 points each)
    - 100-90 (A), 90-80 (B), 80-70 (C), 70-60(D)
- **Equal-depth Binning**
  - The number of the items in each bin is approximately the same
  - Example:
    - 240 students, each grade will have 60 students no matter what
      - A possible result is:
      - 100-75 is A, 75-72 is B, 72-65 is C, 65-60 is D

Example: 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 (**equal-depth**) bins:

- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

* Smoothing by **bin means**:

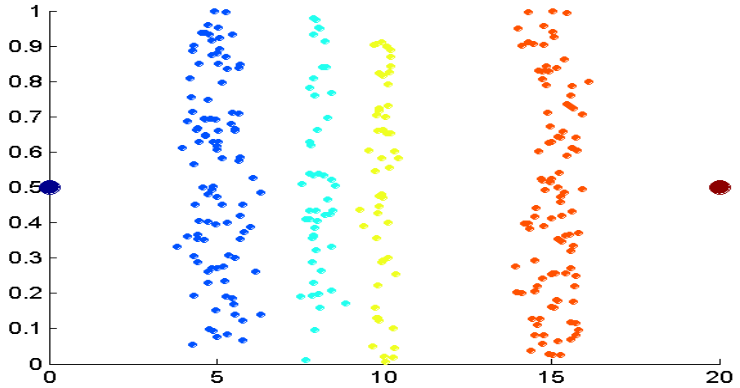
- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

* Smoothing by **bin boundaries** (Specific rules. What rule is applied below?):

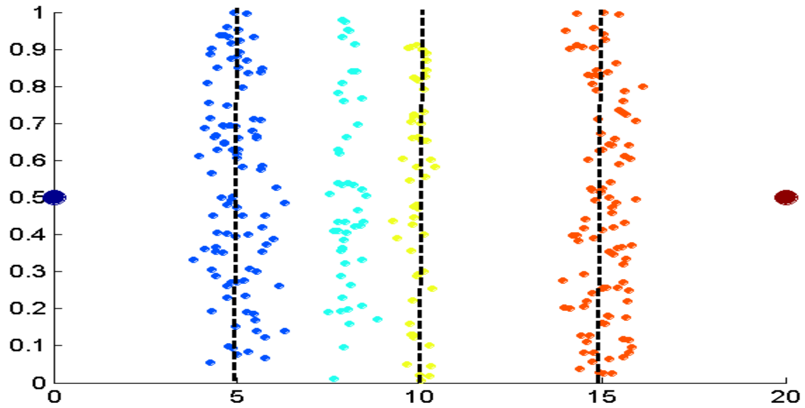
- Bin 1: 4, **4, 4,** 15
- Bin 2: 21, **21, 25,** 25
- Bin 3: 26, **26, 26,** 34

- Challenge:
  - How would you apply **equal-width** binning in the example?

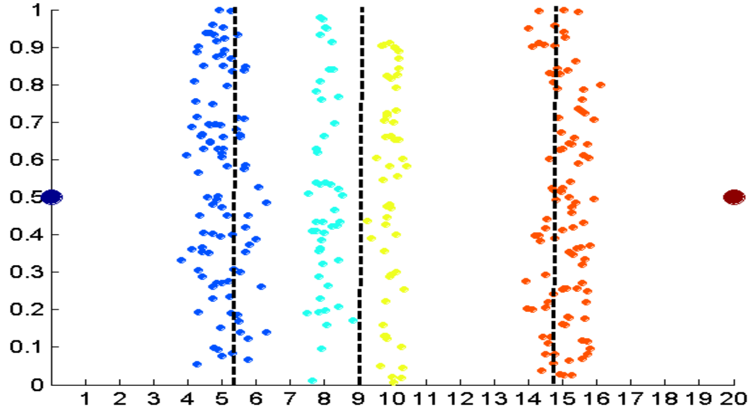
Discretization Without Supervision: Binning vs. Clustering



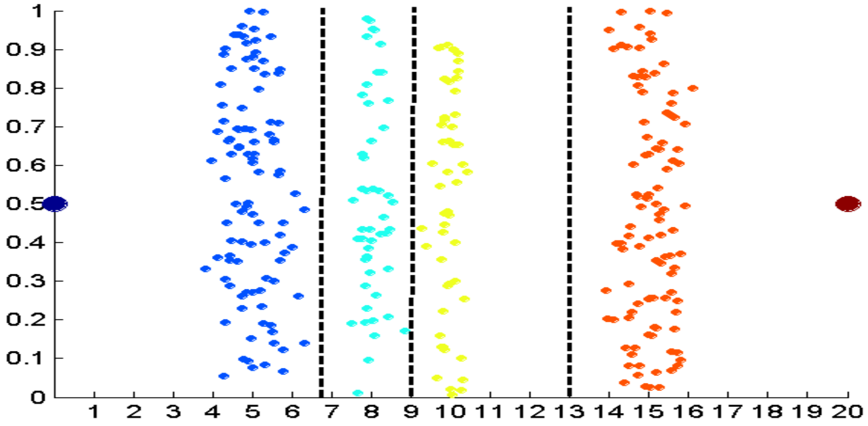
Data



Equal width (distance) binning



Equal depth (frequency) binning



K-means clustering leads to better results

Concept Hierarchy Generation

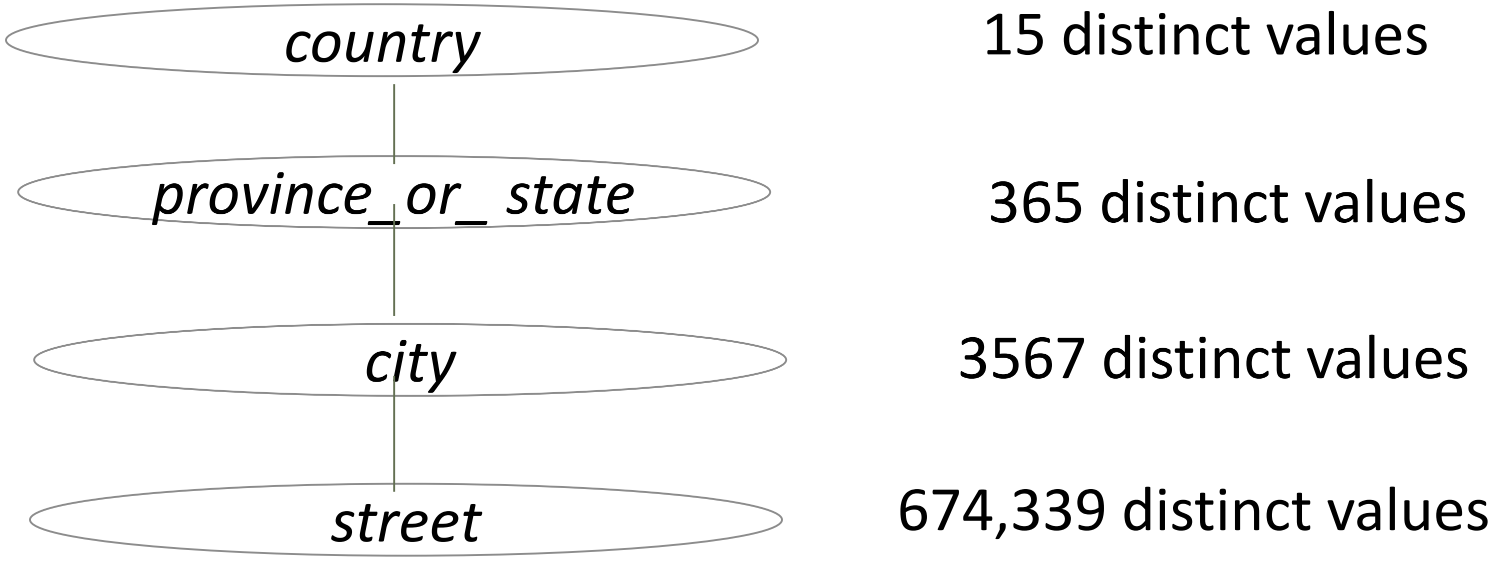
- A lot of things in life follow hierarchical structure:
  - Hour, day, week, month, year
- Once you put data in a hierarchy you get a better understanding of the data
- Organization like this is used a lot in data warehouse construction
- Example: Reports about Walmart Sales Activity
  - Start reports about local Walmart, **then** State Walmart, **then** country Walmart, and **then** the global Walmart.
  - A manager might want a summarizations on different levels of **time** too, for reports.

Concept Hierarchy Generation for Nominal Data

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
  - *street* < *city* < *state* < *country*
- Specification of a hierarchy for a set of values by explicit data grouping
  - {Urbana, Champaign, Chicago} < Illinois
- Specification of only a partial set of attributes
  - E.g., only *street* < *city*, not others
- Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
  - E.g., for a set of attributes: {*street, city, state, country*}

Automatic Concept Hierarchy Generation

- Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy
  - Exceptions, e.g., weekday, month, quarter, year



Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
- Data Cleaning
- Data Integration
- Data Reduction and Transformation
- Dimensionality Reduction
- Summary

Dimensionality Reduction

- **Dimensionality reduction**
  - Obtain principal variables, get rid of the others

- **Why?**
  - Combinations will grow exponentially
  - Data becomes sparse
  - Density and distance between points becomes less meaningful
  - **“Curse of Dimensionality”**

$$\begin{pmatrix} 1.0 & 0 & 5.0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3.0 & 0 & 0 & 0 & 0 & 11.0 & 0 \\ 0 & 0 & 0 & 0 & 9.0 & 0 & 0 & 0 \\ 0 & 0 & 6.0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 7.0 & 0 & 0 & 0 & 0 \\ 2.0 & 0 & 0 & 0 & 0 & 10.0 & 0 & 0 \\ 0 & 0 & 0 & 8.0 & 0 & 0 & 0 & 0 \\ 0 & 4.0 & 0 & 0 & 0 & 0 & 0 & 12.0 \end{pmatrix}$$

Dimensionality Reduction

- **Advantages of 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

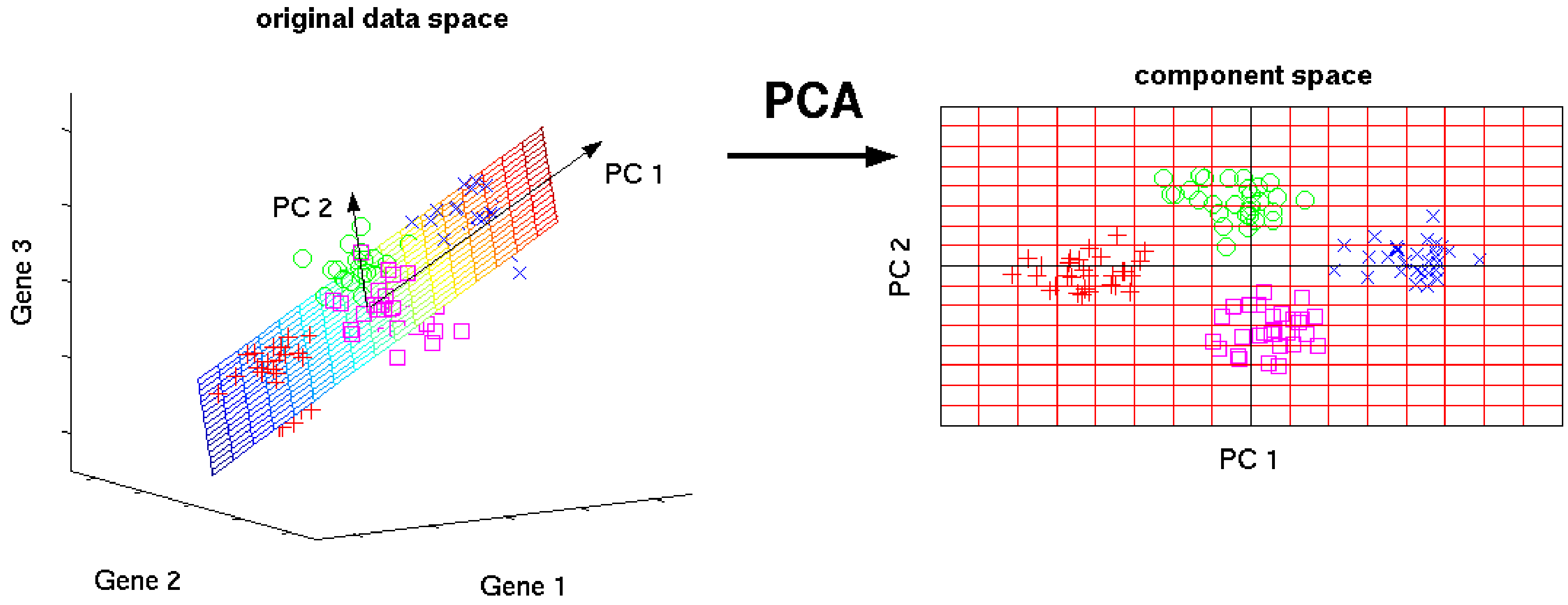
- Dimensionality reduction methodologies
  - **Feature selection**:
    - Example: choose TAs for next semester
    - (…, name, netid, major, $\boxed{\text{midterm\_score, final\_score, assignment\_score,}}$ standing, age, birthday, …)
  - **Feature extraction**:
    - Example: analyze iPhone's **annual** sales in different stores
    - (store_id, address, city, state$\boxed{\text{, sales_Q1, sales_Q2, sales_Q3, sales_Q4}}$,…)
    - (store_id, address, city, state, annual_sales,…)

Dimensionality Reduction Techniques

- Some typical dimensionality methods
  - Principal Component Analysis
  - Supervised and nonlinear techniques
    - Feature subset selection
    - Feature creation

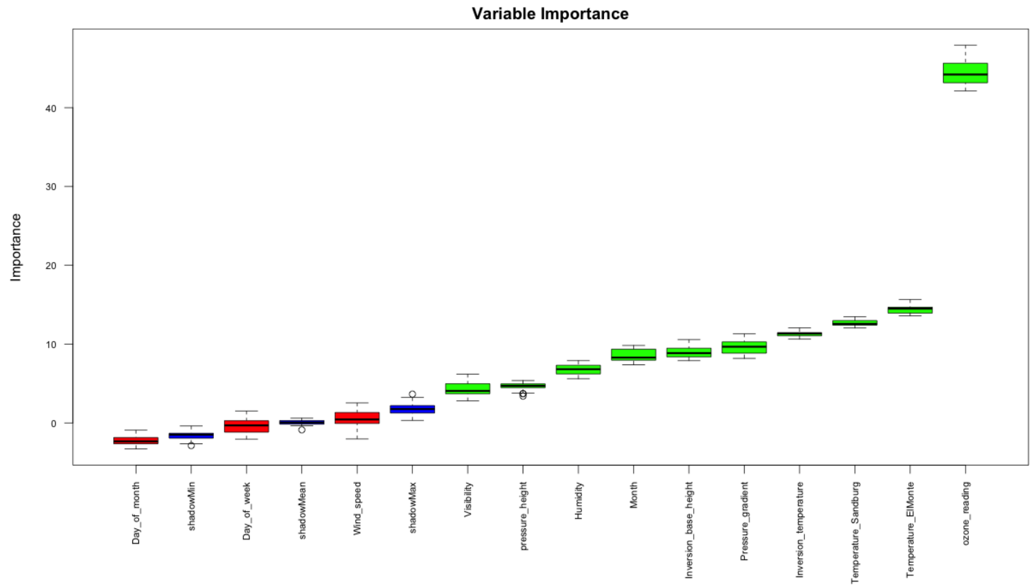
Principal Component Analysis (PCA)

- Basic idea*:*
  - Search for *k* *n*-dimensional orthogonal vectors that can best be used to represent the data, where $k \leq n$
  - The original data can be projected onto smaller space
  - Only work on numerical data



Attribute Subset Selection

- **Goal**
  - Minimal set of attributes
  - Similar probability distribution
- **Redundant attributes**
  - E.g., purchase price of a product and the amount of sales tax paid
- **Irrelevant attributes**
  - Ex. A student's ID is often irrelevant to the task of predicting his/her GPA



Heuristic Search in Attribute Selection

- Strategy: make locally optimal choice in the hope that this will lead to a globally optimal solution
- There are *2^d^* possible attribute combinations of *d* attributes (exponential)
- Typical heuristic attribute selection methods:
  - Best single attribute under the attribute independence assumption: choose by significance tests

Forward selection	Backward elimination
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$

Greedy (heuristic) methods for attribute subset selection

Attribute Creation (Feature Generation)

- Create new attributes (features) that can capture the important information in a data set more effectively than the original ones
- Three general methodologies
  - Attribute extraction
    - Domain-specific
  - Mapping data to new space (see: data reduction)
    - E.g., Fourier transformation, wavelet transformation, manifold approaches (not covered)
  - Attribute construction
    - Combining features (see: discriminative frequent patterns in Chapter on "Advanced Classification")
    - Data discretization

Summary

- **Data quality**: accuracy, completeness, consistency, timeliness, believability, interpretability
- **Data cleaning**: e.g. missing/noisy values, outliers
- **Data integration** from multiple sources:
  - Entity identification problem; Remove redundancies; Detect inconsistencies
- **Data reduction, data transformation and data discretization**
  - Numerosity reduction; Data compression
  - Normalization; Concept hierarchy generation
- **Dimensionality reduction**
  - Feature selection and feature extraction
  - PCA; attribute subset selection (heuristic search); attribute creation

References

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. Comm. of ACM, 42:73-78, 1999
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- T. Dasu, T. Johnson, S. Muthukrishnan, V. Shkapenyuk. Mining Database Structure; Or, How to Build a Data Quality Browser. SIGMOD'02
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4), Dec. 1997
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- E. Rahm and H. H. Do. Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering. Vol.23, No.4*
- V. Raman and J. Hellerstein. Potters Wheel: An Interactive Framework for Data Cleaning and Transformation, VLDB'2001
- T. Redman. Data Quality: Management and Technology. Bantam Books, 1992
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. IEEE Trans. Knowledge and Data Engineering, 7:623-640, 1995