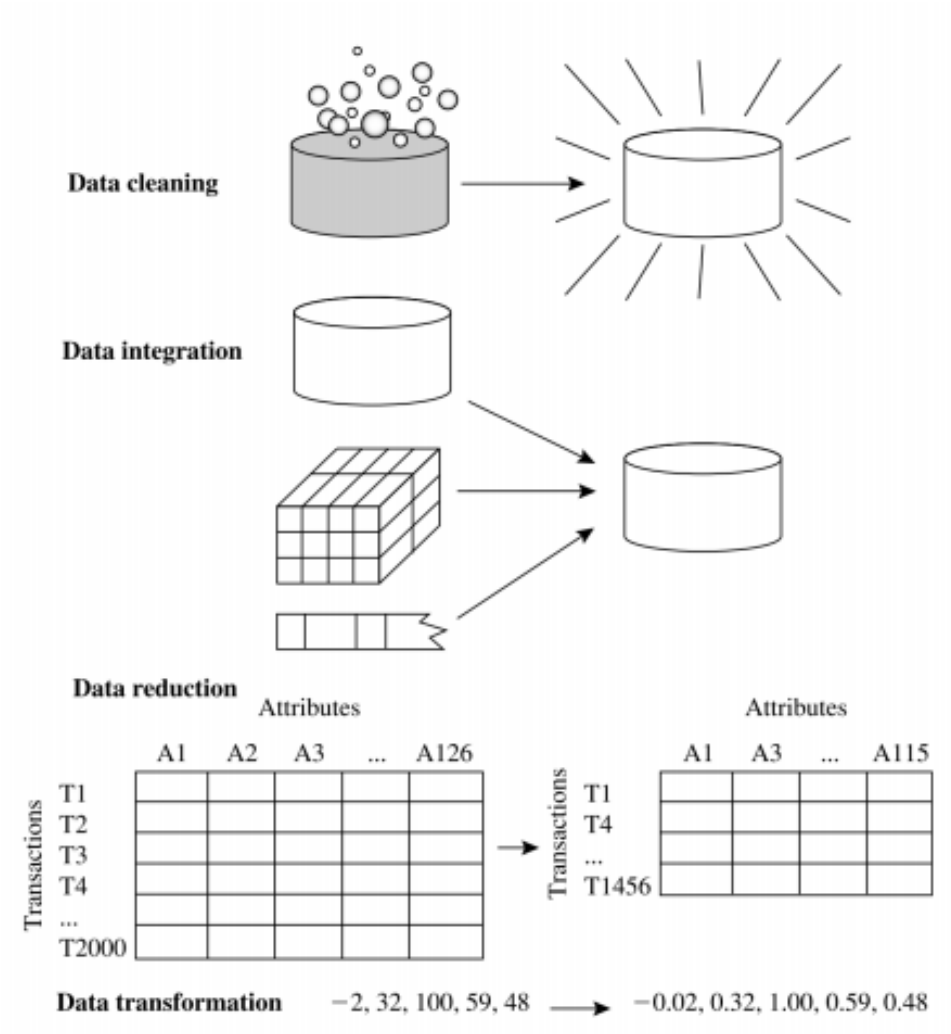
Data Pre-Processing

Data Pre-processing – Major Steps

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data transformation**
  - Normalization and aggregation
- **Data reduction**
  - Obtains reduced representation in volume but produces the same or similar analytical results
- **Data discretization**
  - Part of data reduction but with particular importance, especially for numerical data



Data Collection

- **Data are likely to**
  - originate from multiple heterogeneous sources:
    - *multiple dedicated database, experimentations, observations, web information, past analysis, Government reports, related case studies , etc.*
- **Database are often designed and created**
  - for operational aspects and
  - May not be suitable for strategic decision.
- **Duplicate, noisy or missing data may give incorrect statistics or even misleading decisions.**
- **Quality decision is expected from quality data**

Garbage in => Garbage out

Why data is dirty?

- **Incomplete data may come from**
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
  - Lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
- **Noisy data (incorrect values) may come from**
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
  - Containing errors or outliers
    - e.g., Salary="-10", date as Feb-31st , faulty instrument, transmission error
- **Inconsistent data may come from**
  - Different data sources
  - Violation of functional dependency (e.g., modify some linked data)
  - Containing discrepancies in codes or names, lack of standardization
    - Inconsistent attributes: Age="42" Birthday="03/07/1997" (redundant attributes)
    - Inconsistent rating : "1,2,3" vs "A, B, C"
    - Inconsistent standards: 1.6 km vs 1mile, ASCII vs Unicode

Why data is dirty?

- **Spurious abbreviations**
  - South Bridge Road vs S.B.Road
  - Mahatma Gandhi Road vs M G Road
  - harmonization of short codes (St, rd etc.) to actual words (street, road).
- **Symantec equivalence**
  - Burma and Myanmar, Ceylon and Sri Lanka, Peking and Beijing
- **Duplicate Data**
  - Prof William Sherwin/ Bill Sherwin, American Street or Church Street
- **Inconsistence / Outdated data**
  - Item assigned with a new code for discounted sale.
- **Referential Inconsistency**
  - $1000k sale reported from a branch that has been closed down.
- **Inconsistence association**

How to measure data quality?

- **A well-accepted multidimensional view of data quality:**
  - **Accuracy:** Correct or wrong, degree of accuracy/ precision
  - **Completeness:** The degree to which all required attributes are filled in.
  - **Consistency:** The degree to which set of values are equivalent in across systems, dangling…
  - **Timeliness:** Updated regularly
  - **Believability:** How trustable is the data?
  - **Value added:** Is the data informative and non-redundant
  - **Interpretability:** How easily can the data be understood?
  - **Accessibility:** Ease of availability
  - **Uniformity:** Degree to which a set of data measures are specified using the same units.

Data Cleaning

- **Importance**
  - "Data cleaning is one of the three biggest problems in data warehousing"—Ralph Kimball
  - "Data cleaning is the number one problem in data warehousing"—DCI survey
- **Data cleaning tasks**
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

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 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

1. **Ignore the tuple**: This is usually done when the class label is missing (assuming the mining task involves classification). This method is not very effective, unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.
2. **Fill in the missing value manually**: In general, this approach is time consuming and may not be feasible given a large data set with many missing values.
3. **Use a global constant to fill in the missing value**: Replace all missing attribute values by the same constant such as a label like “Unknown” or $-\infty$.
4. **Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value**: measures of central tendency, which indicate the “middle” value of a data distribution. For normal (symmetric) data distributions, the mean can be used, while skewed data distribution should employ the median.

How to handle missing data

5. **Use the attribute mean or median for all samples belonging to the same class as the given tuple:** For example, if classifying customers according to credit_risk, we may replace the missing value with the mean income value for customers in the same credit risk category as that of the given tuple. If the data distribution for a given class is skewed, the median value is a better choice
6. **Use the most probable value to fill in the missing value:** This may be determined with **regression, inference-based tools using a Bayesian formalism, or decision tree induction**. For example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for income.

Simple Linear Regression

- In **simple linear regression**, we predict scores on one variable from the scores on a second variable.
- The variable we are predicting is called the ***criterion variable*** and is referred to as **Y**. The variable we are basing our predictions on is called the ***predictor variable*** and is referred to as **X**.
- When there is only one predictor variable, the prediction method is called ***simple regression***.
- In simple linear regression, the predictions of **Y** when plotted as a function of **X** form a **straight line**.

Example

- The example data in Table 1 are plotted in Figure 1. You can see that there is a positive relationship between X and Y. If you were going to predict Y from X, the higher the value of X, the higher your prediction of Y.

Table 1. Example data.

X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

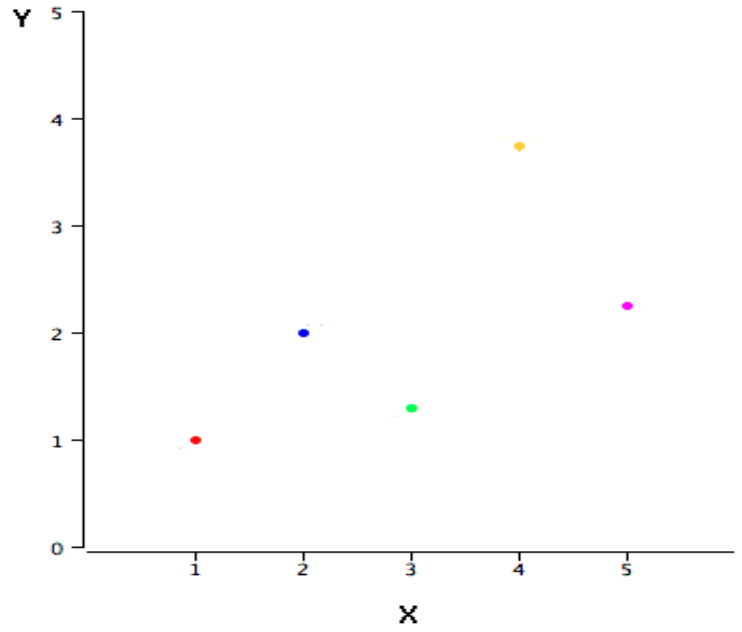
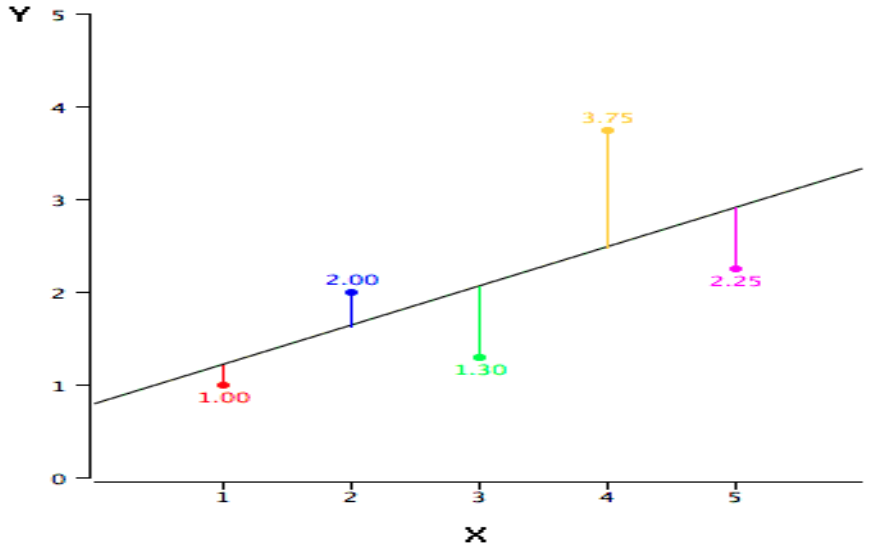


Figure 1. A scatter plot of the example data.

- Linear regression consists of finding the best-fitting straight line through the points.
- The best-fitting line is called a ***regression line***.
- The black diagonal line in Figure 2 is the regression line and consists of the predicted score on Y for each possible value of X. **The vertical lines from the points to the regression line represent the errors of prediction**. As you can see, the red point is very near the regression line; its error of prediction is small. By contrast, the yellow point is much higher than the regression line and therefore its error of prediction is large.

Figure 2



- The error of prediction for a point is the value of the point minus the predicted value (the value on the line). Table 2 shows the predicted values (Y') and the errors of prediction (Y-Y'). For example, the first point has a Y of 1.00 and a predicted Y (called Y') of 1.21. Therefore, its error of prediction is -0.21.

Table 2. Example data.

X	Y	Y'	Y-Y'	$(Y-Y')^2$
1.00	1.00	1.210	-0.210	0.044
2.00	2.00	1.635	0.365	0.133
3.00	1.30	2.060	-0.760	0.578
4.00	3.75	2.485	1.265	1.600
5.00	2.25	2.910	-0.660	0.436

- You may have noticed that we did not specify what is meant by "best-fitting line." By far, the most commonly-used criterion for the **best-fitting line is the line that minimizes the sum of the squared errors of prediction**. That is the criterion that was used to find the line in Figure 2. The last column in Table 2 shows the squared errors of prediction. The sum of the squared errors of prediction shown in Table 2 is lower than it would be for any other regression line.

COMPUTING THE REGRESSION LINE

- The calculations are based on the statistics shown in Table 3. M_X is the mean of X, M_Y is the mean of Y, s_X is the standard deviation of X, s_Y is the *standard deviation* of Y, and r is the *correlation* between X and Y.

Table 3. Statistics for computing the regression line.

M_X	M_Y	s_X	s_Y	r
3	2.06	1.581	1.072	0.627

The slope (b) can be calculated as follows:

$$\mathbf{b = r\, s_Y / s_X}$$

and the intercept (A) can be calculated as

$$\mathbf{A = M_Y - bM_X}$$

For these data,

$$b = (0.627)(1.072)/1.581 = 0.425$$

$$A = 2.06 - (0.425)(3) = 0.785$$

The formula for a regression line is

Y' = bX + A

where Y' is the predicted score, b is the slope of the line, and A is the Y intercept which is 0.425 and 0.785 resp. The equation for the line in Figure 2 is

Y' = 0.425X + 0.785

For X = 1,

Y' = (0.425)(1) + 0.785 = 1.21.

For X = 2,

Y' = (0.425)(2) + 0.785 = 1.64.

S_x Calculation (Standard Deviation)

Scores	*Deviation (X - M)*	*Squared Dev.*
1	-2	4
2	-1	1
3	0	0
4	1	1
5	2	4
		SS: 10
M: 3		

Standard Deviation Calculation

N: 5
M: 3
SS: 10
$s^2 = SS(N - 1) = 10/(5-1) = 2.5$
$s = \sqrt{s^2} = \sqrt{2.5} = 1.58$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2$$

S_y Calculation (Standard Deviation)

Scores	*Deviation (X - M)*	*Squared Dev.*
1	-1.06	1.12
2	-0.06	0.00
1.3	-0.76	0.58
3.75	1.69	2.86
2.25	0.19	0.04
		SS: 4.6
M: 2.06		

Standard Deviation Calculation

N: 5
M: 2.06
SS: 4.6
$s^2 = SS(N - 1) = 4.6/(5-1) = 1.15$
$s = \sqrt{s^2} = \sqrt{1.15} = 1.07$

Pearson Correlation Coefficient Calculator (r)

X Values	Y Values
1	1
2	2
3	1.3
4	3.75
5	2.25

$X - M_x$	$Y - M_y$	$(X - M_x)^2$	$(Y - M_y)^2$	$(X - M_x)(Y - M_y)$
-2.000	-1.060	4.000	1.124	2.120
-1.000	-0.060	1.000	0.004	0.060
0.000	-0.760	0.000	0.578	0.000
1.000	1.690	1.000	2.856	1.690
2.000	0.190	4.000	0.036	0.380
Mx: 3.000	My: 2.060	Sum: 10.000	Sum: 4.597	Sum: 4.250

X Values

$\sum = 15$

Mean = 3

$\sum(X - M_x)^2 = SS_x = 10$

Y Values

$\sum = 10.3$

Mean = 2.06

$\sum(Y - M_y)^2 = SS_y = 4.597$

X and Y Combined

$N = 5$

$\sum(X - M_x)(Y - M_y) = 4.25$

R Calculation

$r = \sum((X - M_y)(Y - M_x)) / \sqrt{((SS_x)(SS_y))}$

$r = 4.25 / \sqrt{((10)(4.597))} = 0.6268$

Meta Numerics (cross-check)

r = 0.6268

Key

X: X Values
Y: Y Values
M_x: Mean of X Values
M_y: Mean of Y Values
$X - M_x$ & $Y - M_y$: Deviation scores
$(X - M_x)^2$ & $(Y - M_y)^2$: Deviation Squared
$(X - M_x)(Y - M_y)$: Product of Deviation Scores

The value of R is 0.6268.

A Real Example

- The case study "SAT and College GPA" contains high school and university grades for 105 computer science majors at a local state school. We now consider how we could predict a student's university GPA if we knew his or her high school GPA.
- Figure 3 shows a scatter plot of University GPA as a function of High School GPA. You can see from the figure that there is a strong positive relationship. The correlation is 0.78. The regression equation is

University GPA' = (0.675)(High School GPA) + 1.097

Therefore, a student with a high school GPA of 3 would be predicted to have a university GPA of

University GPA' = (0.675)(3) + 1.097 = 3.12.

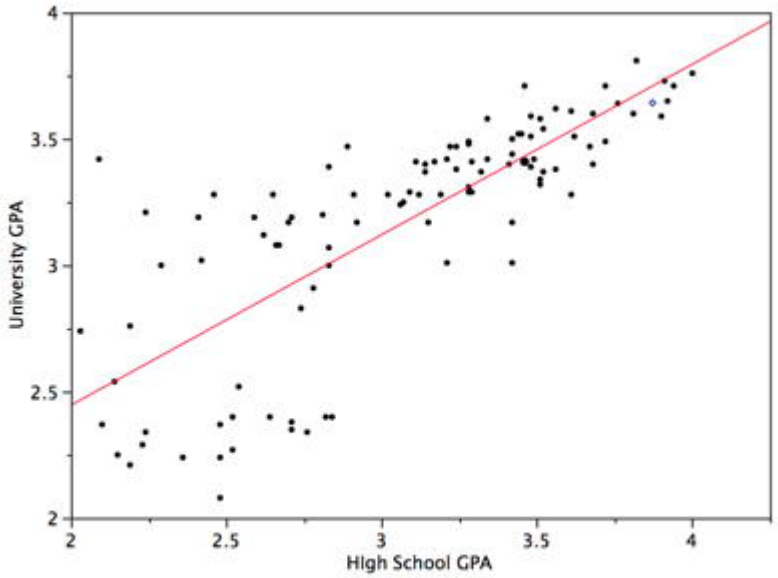


Figure 3. University GPA as a function of High School GPA.

Multiple Regression

- In **simple linear regression**, a *criterion variable (Y)* is predicted from one predictor variable(X).
- In **multiple regression**, the criterion is predicted by two or more variables.

$$UGPA' = b_1 HSGPA + b_2 SAT + A$$

criterion variable

predictor variable 1

Regression Coefficients

predictor variable 2

constant

Noisy Data

- **Noise: random error or variance in a measured variable**
- **Incorrect attribute values may due to**
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- **Other data problems which requires data cleaning**
  - duplicate records
  - incomplete data
  - inconsistent data

How to handle noisy data?

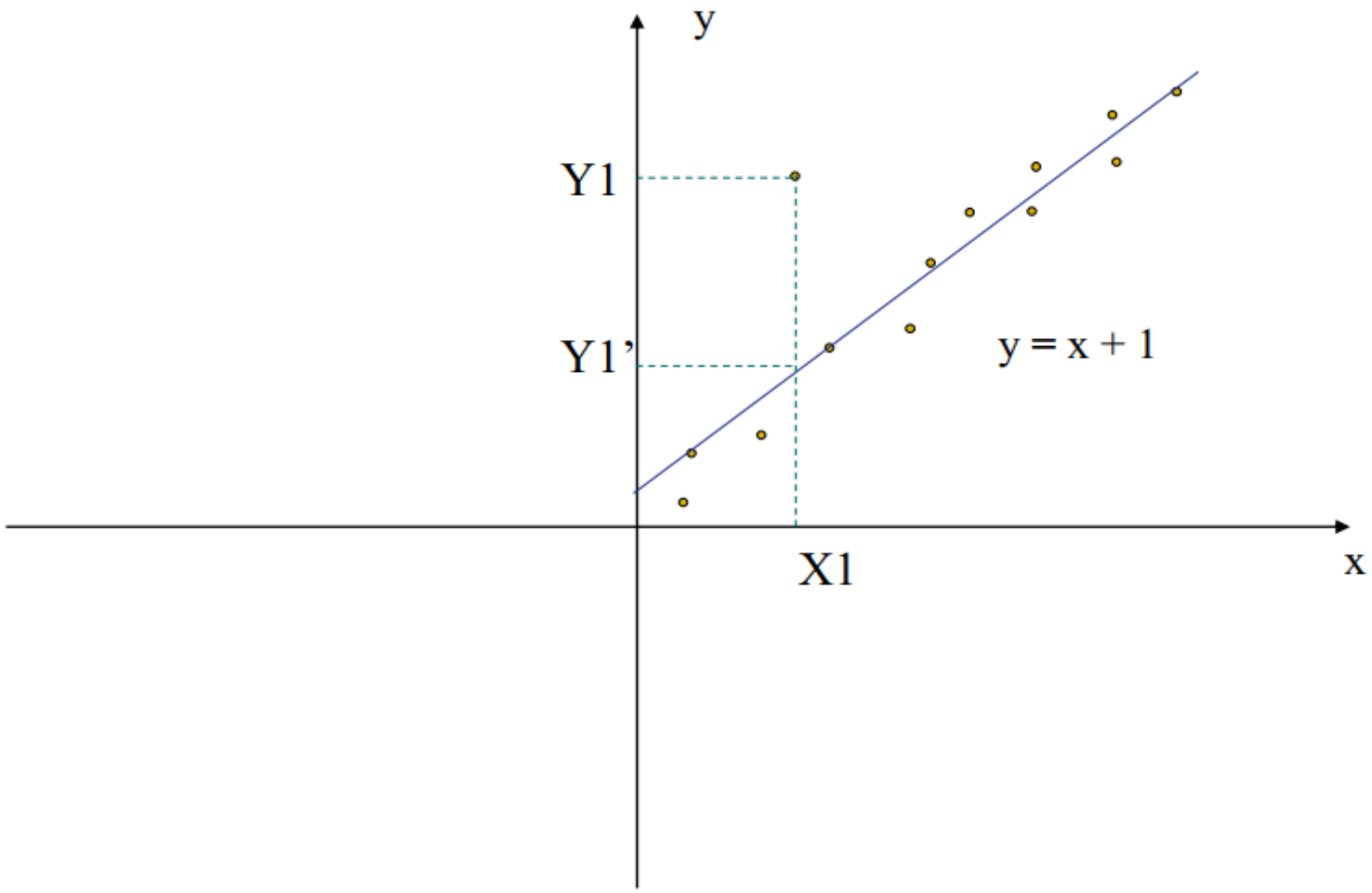
- **Binning**
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- **Regression**
  - smooth by fitting the data into regression functions
- **Clustering**
  - detect and remove outliers
- **Combined computer and human inspection**
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Binning methods for data smoothing

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

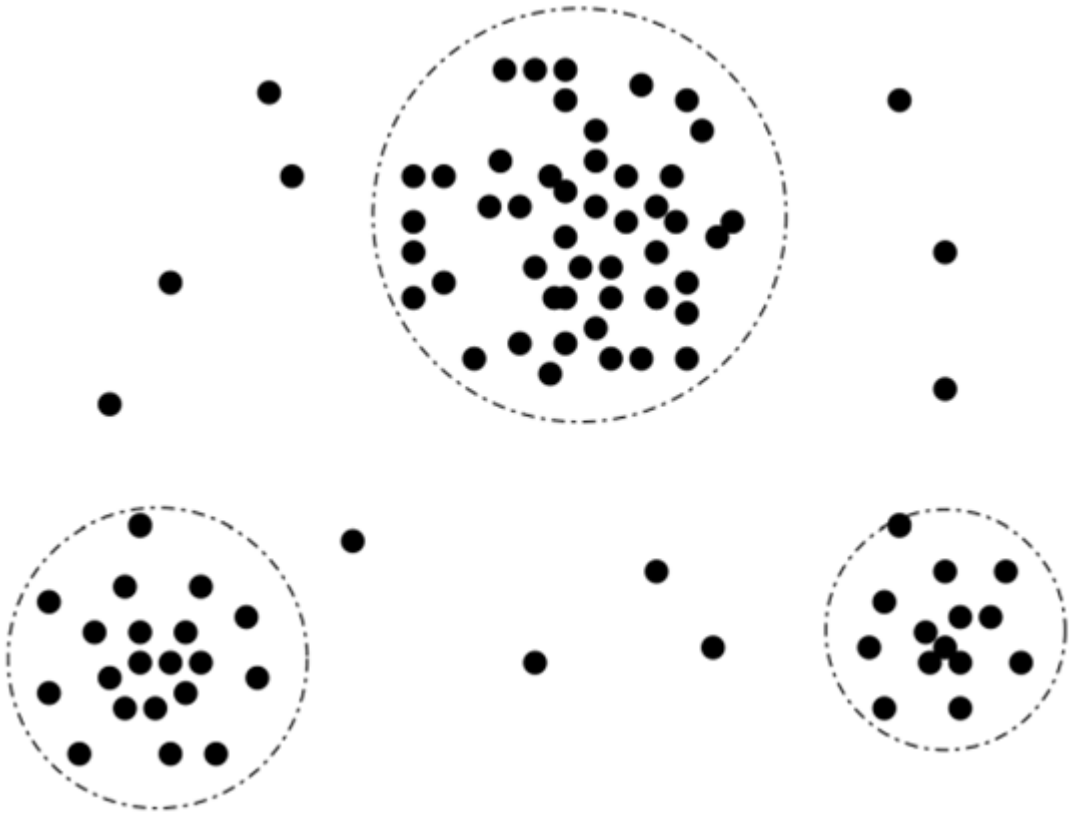
* **Partition into equal-frequency (equi-depth) bins:**
  - **Bin 1: 4, 8, 9, 15**
  - **Bin 2: 21, 21, 24, 25**
  - **Bin 3: 26, 28, 29, 34**
* **Smoothing by bin means:**
  - **Bin 1: 9, 9, 9, 9**
  - **Bin 2: 23, 23, 23, 23**
  - **Bin 3: 29, 29, 29, 29**
* **Smoothing by bin boundaries:**
  - **Bin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

Regression



Regression: Data Smoothing can also be done by regression. Linear regression involves finding the "best" line to fit two attributes (or variables) so that one attribute can be used to predict the other. Multiple linear regression is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.

Clustering Analysis



Outliers may be detected by clustering, for example, where similar values are organized into groups, or "clusters." Intuitively, values that fall outside of the set of clusters may be considered outliers

Data Cleaning as a Process

- **Data discrepancy detection**
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- **Data migration and integration**
  - Data migration tools: allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface
- **Integration of the two processes**
  - Iterative and interactive

Potter's Wheel ABC - http://control.cs.berkeley.edu/

Data Integration

- Data mining often requires **data integration**—the merging of data from multiple data stores.
- Careful integration can help reduce and avoid redundancies and inconsistencies in the resulting data set.
- This can help improve the accuracy and speed of the subsequent data mining process.

Data Integration

- Tasks to do-

1. Entity Identification Problem

   Ex. how can the data analyst or the computer be sure that customer_id in one database and cust_number in another refer to the same attribute? (Metadata)

2. Redundancy and Correlation Analysis
3. Tuple Duplication

   Ex. Purchase dataset and Item dataset contains details of Purchasers (Metadata)

4. Data Value Conflict Detection and Resolution Ex. Weight in one dataset may stored in Pounds and in another as Kgs (Metadata)

Data Integration

Entity Identification Problem- How can equivalent real-world entities from multiple data sources be matched up? This is referred to as the **entity identification problem**.

- For example, how can the data analyst or the computer be sure that customer_id in one database and cust_number in another refer to the same attribute?
- Solution is used **Metadata** of each attribute. When matching attributes from one database to another during integration, special attention must be paid to the structure of the data.

Data Integration

Redundancy and Correlation Analysis-

Redundant data occur often when integration of multiple databases

- *Object identification*: 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

Redundant attributes may be able to be detected by *correlation analysis*

Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Integration

- Some redundancies can be detected using metadata and some are by correlation analysis.
- Given two attributes, such analysis can measure how strongly one attribute implies the other, based on the available data.
- For **nominal data**, we use the **χ^2 (chi-square) test**. It is also used to choose important features for further analysis also.
- For **numeric attributes**, we can use the **correlation coefficient** and **covariance**, both of which access how one attribute's values vary from those of another.

Data Integration

- An attribute (column or feature of data set) is called redundant if it can be derived from any other attribute or set of attributes.
- Inconsistencies in attribute or dimension naming can also lead to the redundancies in data set.

Redundant Feature Example

- **Example –**
  We have a data set having three attributes-
  person_name, is_male, is_female.
- is_male is 1 if the corresponding person is a male else it is 0 .
- is_female is 1 if the corresponding person is a female else it is 0.

person_name	is_male	is_female
Aman	1	0
Abhinav	1	0
Ashutosh	1	0
Dishi	0	1
Abhishek	1	0
Avantika	1	0
Ayushi	0	1

χ^2 (chi-square) test

- Tests whether two **nominal** variables are related or independent.
- The Chi-Squared test is a statistical hypothesis test that assumes (the null hypothesis) that the observed frequencies for a categorical variable match the expected frequencies for the categorical variable.

Interpretation

✓ H0: the two samples are independent.

✓ H1: there is a dependency between the samples.

χ^2 (chi-square) test Interpretation

- We can interpret the test statistic in the context of the **chi-squared distribution Statistic** (χ^2) with the **Critical Value** (***α-value)*** as follows:
- **If Statistic >= Critical Value**: significant result, reject null hypothesis (H0), dependent.
- **If Statistic < Critical Value**: not significant result, fail to reject null hypothesis (H0), independent.

Correlation Analysis (Categorical Data)

- **χ^2 (chi-square) test**

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n}$$

- **Contingency Table**

	a_1	...	a_i	...	a_c	SUM
b_1						
...						
b_j			o_{ij} (e_{ij})			$count(B = b_j)$
...						
b_r						
SUM			$count(A = a_i)$			

χ 2 Test

- **Degree of freedom** $(r-1) \times (c-1)$
  - The test is based on looking up the significance value for degree of freedom using stats table
- **The larger the χ^2 value, the more likely the variables are related (Dependent)**
  - That is, if the tabled χ^2 value for the degree of freedom at a significance level 0.001 is smaller than the calculated, then the independence hypothesis can be rejected.
- **The cells that contribute the most to the χ^2 value are those whose actual count is very different from the expected count**
- **Correlation does not imply causality**
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Chi-Square Calculation: An Example

A- Play Chess/ Not Play Chess

B- Like SF/ Not like SF

	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

Hypo- A & B are independent.

- **χ^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)**

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

- **Degree of freedom is 1, for which significance level is 10.828**
- **It shows that *like_science_fiction* and *play_chess* are correlated in the group**

- The ***p*-value** is the conditional probability of observing the statistic value when the ***null hypothesis*** is true.
- The ***α-value*** is the significance critical value, also called as maximum threshold for ***p-value***.

  It can be 0.1, 0.05, 0.01 or 0.001.
- The critical value for the chi-square statistic is determined by the level of significance (**typically .05**) and the degrees of freedom.
- The degrees of freedom for the chi-square are calculated using the following formula: *df* = (r-1)(c-1) where r is the number of rows and c is the number of columns.
- If the observed chi-square test statistic is greater than the critical value, **the null hypothesis can be rejected**.

probability level (alpha)

Df	0.5	0.10	0.05	0.02	0.01	0.001
1	0.455	2.706	3.841	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.517

Example – Chi-square Test

- EXAMPLE:
  A manger wants to see if geographical region is associated with ownership of a macintosh computer. The manager surveys 100 people and the data breaks down as follows:

	Mac	No Mac	Row total
North East	12	14	26
South West	21	18	39
Mid West	17	18	35
Column Total	50	50	100

	Mac	No Mac	Row total
North East	12 (13)	14 (13)	26
South West	21 (19.5)	18 (19.5)	39
Mid West	17 (17.5)	18 (17.5)	35
Column Total	50	50	100

Hypothesis: the two variables are independent

In this case we have 3 rows and 2 columns, thus df = (3-1)(2-1) = 2. Thus chi-square with 2 df at alpha = 0.001 is 13.815

Calculated value: The sum of (actual-expected)sqrd/expected for each cell. In this case we have 6 cells, so we have to do this formula six times and sum the answers. Cell 1 = (12-13)sqrd/13 = 1/13 = .077. Cell 2 = (14-13)sqrd/13 = .077. Cell 3 =(21-19.5)sqrd/19.5 = .115 Cell 4 = (18-19.5)sqrd/19.5 = .115 cell 5 = (17-17.5)sqrd/17.5 = .014 and Cell 6 = (18-17.5)sqrd/17.5 = .014. Sum up all of these and you get:
.077+.077+.115+.115+.014+.014=.412 This is our calculated value of chi-square statistic.
Compare: Chi-sqr calc is less than critical value so we **do not reject the Hypothesis**.

Conclusion: The Hypothesis was that the two variables were independent and that claim was accepted. The managerial conclusion is that ownership of a mac and geographical region are NOT related (i.e., they are independent).

Chi-square Test for feature selection

- *Chi-square* test is used for categorical features selection in a dataset.
- We calculate Chi-square between **each feature** & **target** and select the desired number of features with best Chi-square scores.

Day	Outlook	Wind	Play Tennis
D1	Sunny	Weak	No
D2	Sunny	Strong	No
D3	Overcast	Weak	Yes
D4	Rain	Weak	Yes
D5	Rain	Weak	Yes
D6	Rain	Strong	No
D7	Overcast	Strong	Yes
D8	Sunny	Weak	No
D9	Sunny	Weak	Yes
D10	Rain	Weak	Yes
D11	Sunny	Strong	Yes
D12	Overcast	Strong	Yes
D13	Overcast	Weak	Yes
D14	Rain	Strong	No

The contingency table for the feature “Outlook” is constructed as below:-

	Yes	No	
Sunny	**2 (3.21)**	**3 (1.79)**	**5**
Overcast	**4 (2.57)**	**0 (1.43)**	**4**
Rain	**3 (3.21)**	**2 (1.79)**	**5**
	9	**5**	**14**

The expected value for the cell (Sunny,Yes) is calculated as $\frac{5}{14} \times 9 = 3.21$ and similarly for others.

The $\chi^2_{outlook}$ value is calculated as below:-

$$\chi^2_{outlook} = \frac{(2-3.21)^2}{3.21} + \frac{(3-1.79)^2}{1.79} + \frac{(4-2.57)^2}{2.57} + \frac{(0-1.43)^2}{1.43} + \frac{(3-3.21)^2}{3.21} + \frac{(2-1.79)^2}{1.79}$$

$$\Rightarrow \chi^2_{outlook} = 3.129$$

The contingency table for the feature "Wind" is constructed as below:-

	Yes	No	
Strong	**3 (3.86)**	**3 (1.14)**	**6**
Weak	**6 (5.14)**	**2 (2.86)**	**8**
	9	**5**	**14**

The χ^2_{wind} value is calculated as below:-

$$\chi^2_{wind} = \frac{(3-3.86)^2}{3.86} + \frac{(3-1.14)^2}{1.14} + \frac{(6-5.14)^2}{5.14} + \frac{(2-2.86)^2}{2.86}$$

$$\Rightarrow \chi^2_{wind} = 3.629$$

On comparing the two scores, we can conclude that the feature "Wind" is more important to determine the output than the feature "Outlook".

Correlation Analysis (Numerical Data)

Stock Prices for *AllElectronics* and *HighTech*

Time point	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

A B

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

$$r_{A,B} = \frac{\sum_{i=1}^{N}(A_i - \overline{A})(B_i - \overline{B})}{N\sigma_A\sigma_B} = \frac{\sum(A_iB_i) - N\overline{A}\,\overline{B}}{N\sigma_A\sigma_B}$$

where N is the number of tuples, $\overline{A}$ and $\overline{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\sum(A_iB_i)$ is the sum of the AB cross-product.

- **If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.**
- **$r_{A,B} = 0$: independent; $r_{A,B} < 0$: negatively correlated**

Covariance of Numerical Data

- **Given two numerical attributes A and B**

$$E(A) = \bar{A} = \frac{\sum_{i=1}^{n} a_i}{n}$$

and

$$E(B) = \bar{B} = \frac{\sum_{i=1}^{n} b_i}{n}.$$

The **covariance** between A and B is defined as

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{n}.$$

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}.$$

Covariance Example

Stock Prices for *AllElectronics* and *HighTech*

Time point	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

A

B

$$Cov(A, B) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}.$$

$$E(AllElectronics) = \frac{6+5+4+3+2}{5} = \frac{20}{5} = \$4$$

and

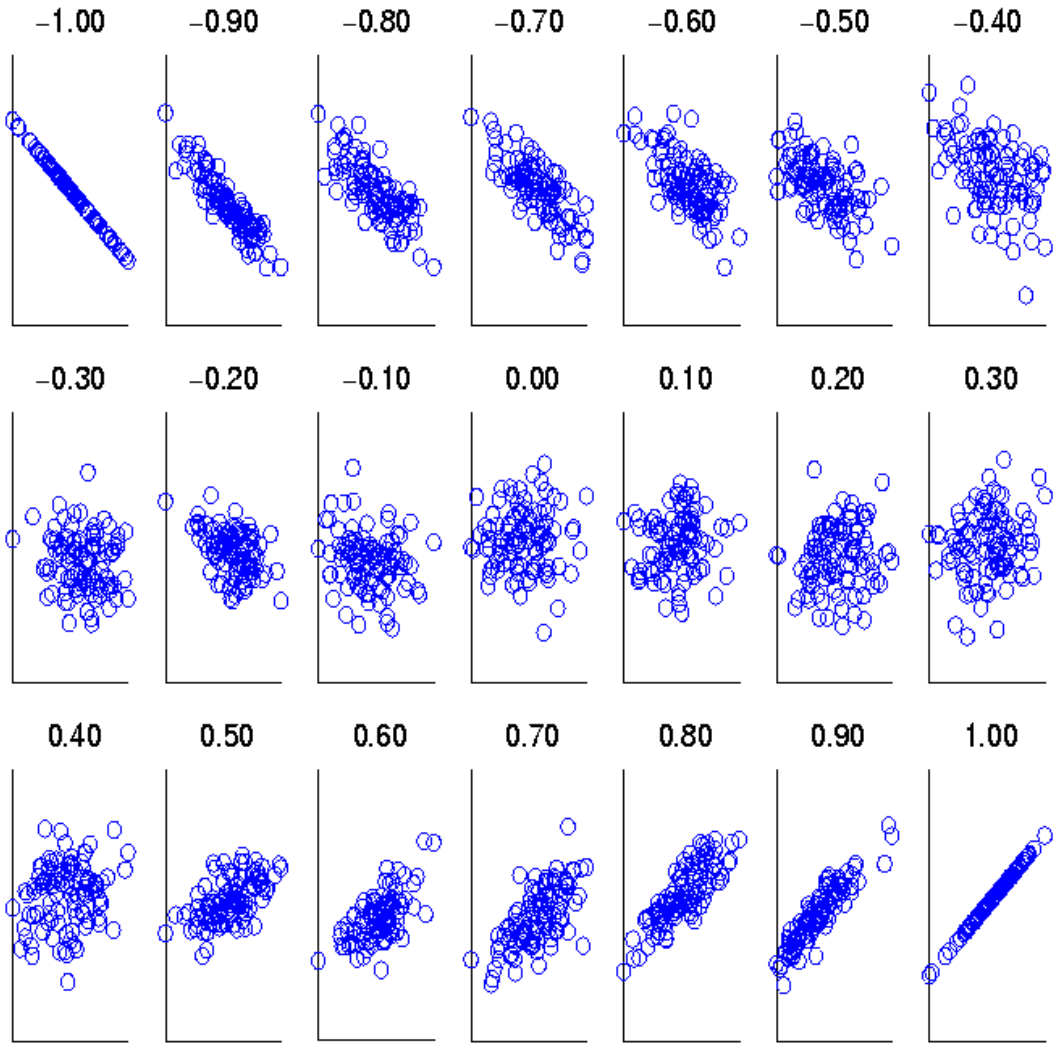
$$E(HighTech) = \frac{20+10+14+5+5}{5} = \frac{54}{5} = \$10.80.$$

Thus, using Eq. (3.4), we compute

$$Cov(AllElectroncis, HighTech) = \frac{6 \times 20 + 5 \times 10 + 4 \times 14 + 3 \times 5 + 2 \times 5}{5} - 4 \times 10.80$$

$$= 50.2 - 43.2 = 7.$$

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

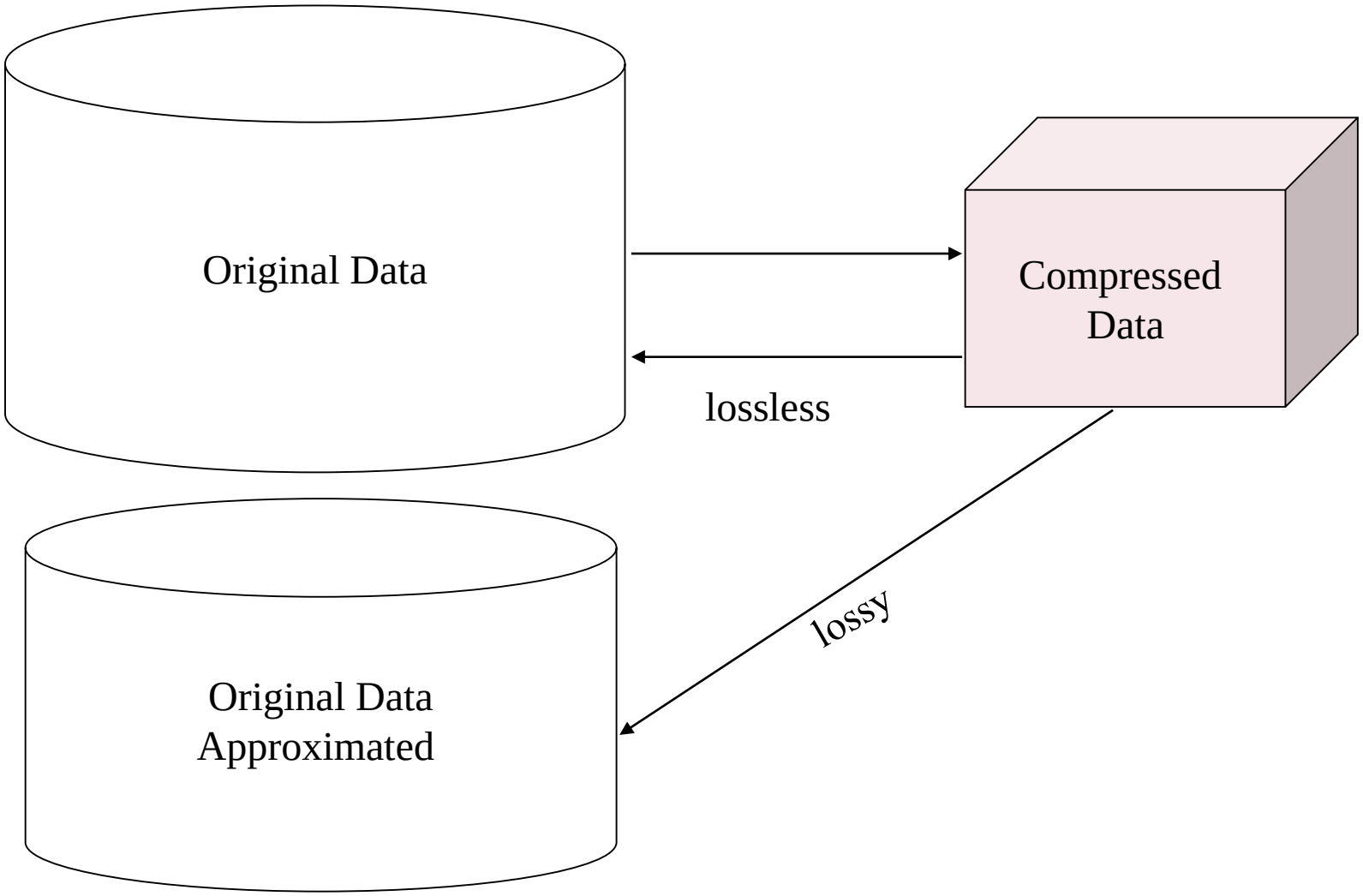
Covariance (Numeric Data)

- **Positive covariance**: If $Cov_{A,B} > 0$, then A and B both tend to be larger than their expected values.
- **Negative covariance**: If $Cov_{A,B} < 0$ then if A is larger than its expected value, B is likely to be smaller than its expected value.
- **Independence**: $Cov_{A,B} = 0$ but the converse is not true:
  - Some pairs of random variables may have a covariance of 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

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 attributes**
    - Wavelet transforms
    - Principal Components Analysis (PCA)
    - Feature subset selection, feature creation
  - Numerosity reduction**, replace original data volume by alternative smaller forms of data representation**
    - Regression and Log-Linear Models
    - Histograms, clustering, sampling
    - Data cube aggregation
  - Data compression

Data Compression



What is & Why Data Reduction

- **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?**
  - Increases storage capacity
  - Easy and efficient Mining, reduces time and memory requirement
  - Easy visualization
  - Help to eliminate irrelevant /redundant features
  - Reduces noise

Curse of Dimensionality: Example

- **53 Blood and urine measurements (wet chemistry) from 65 people (33 alcoholics, 32 non-alcoholics).**
- **Matrix Format (65 X 53): Difficult to see the correlation between the features**

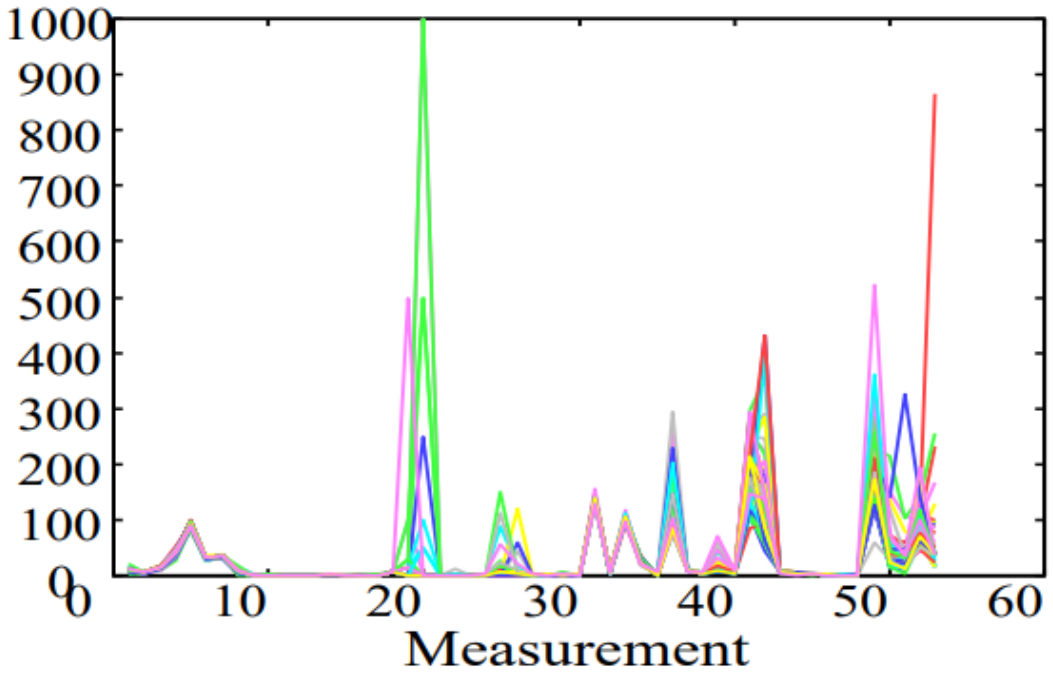
	M1	M2	…	M52	M53	Alcoholics?
P1	-2.053920551	-1.50361144	…	1.02305704	-0.5052951	Yes
P2	-0.004767651	0.04618693	…	-0.07452921	0.8229218	No
P3	0.430102187	1.71553814	…	1.64038150	0.3130619	Yes
P4	-0.817802417	1.56018735	…	-0.21835821	-0.6279286	No
…	…	…	…	…	…	…
…	…	…	…	…	…	…
P53	0.001701727	0.36459985	…	-1.59528279	2.5278118	No

Instances

Features

Data Visualisation

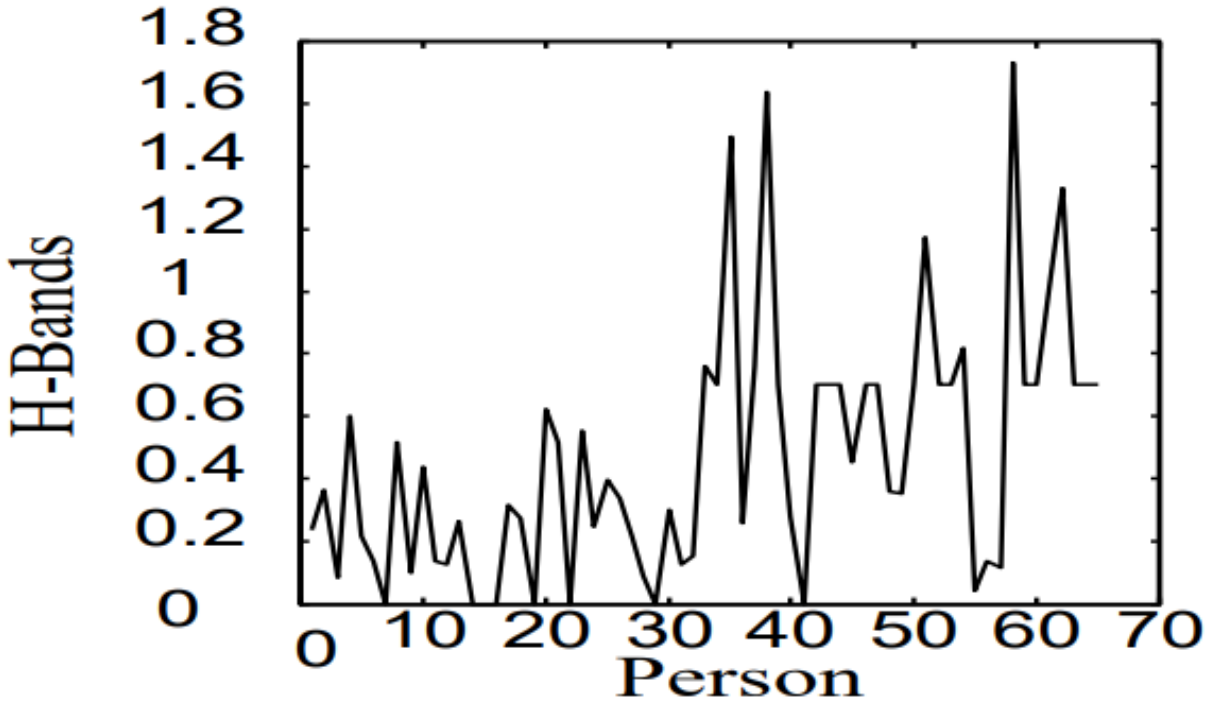
- **Spectral format (65 lines, one for each person)**



- **Difficult to do cross patient analysis**

Data Visualisation

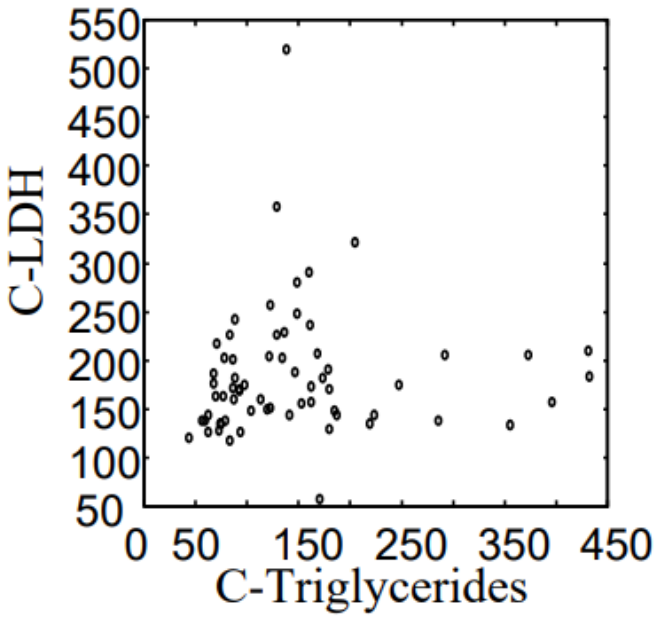
- **Spectral format (53 lines, one for each feature)**



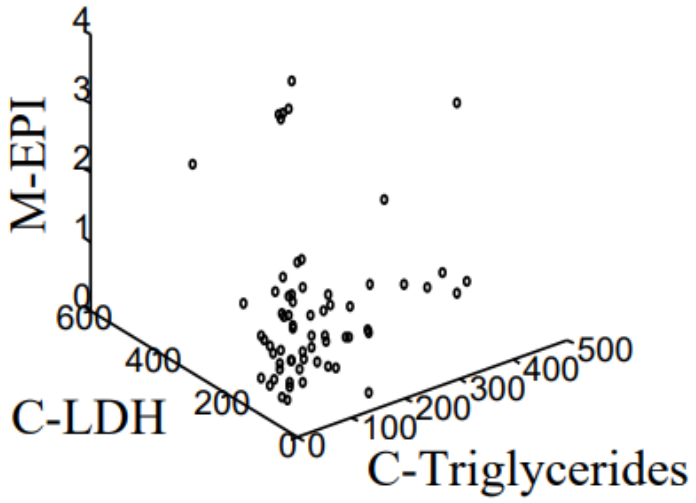
- **Difficult to see the correlations between the features...**

Data Visualisation

Bi - variate



Tri - variate



- **How can we visualize the other variables???**
  - … difficult to see in 4 or higher dimensional spaces...

Data Reduction Problems

- **Do we need a 53 dimension space to view data?**
- **What if there are strong correlation between features?**
- **How to find the 'best' low dimension space that conveys maximum useful information?**

Curse of dimensionality

- **When dimensionality increases, data becomes increasingly sparse**
- **The possible combinations of subspaces will grow exponentially.**
  - Even in the simplest case of 'd' binary variables, the number of possible combinations is 2d , exponential in the dimensionality.
- **Density and distance between points become less significant, which is critical for clustering and outlier analysis**
- **Query accuracy and efficiency degrade rapidly as the dimension increases.**

Data reduction strategies

- Difference between Dimensionality Reduction and Numerosity Reduction :

Dimensionality Reduction	Numerosity Reduction
In dimensionality reduction, data encoding or data transformations are applied to obtain a reduced or compressed for of original data.	In Numerosity reduction, data volume is reduced by choosing suitable alternating forms of data representation.
It can be used to remove irrelevant or redundant attributes.	It is merely a representation technique of original data into smaller form.
In this method, some data can be lost which is irrelevant.	In this method, there is no loss of data.
Methods for dimensionality reduction are: 1. Wavelet transformations. 2.Principal Component Analysis.	Methods for Numerosity reduction are: 1.Regression or log-linear model (parametric). 2.Histograms, clusturing, sampling (non-parametric).
The components of dimensionality reduction are feature selection and feature extraction.	It has no components but methods that ensure reduction of data volume.

Data reduction strategies

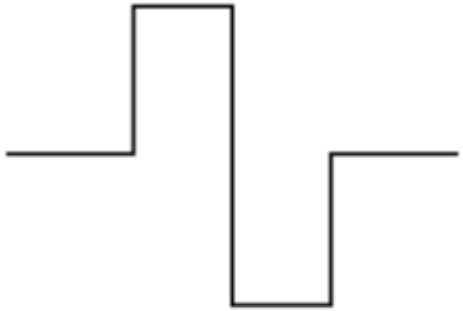
- **Dimensionality Reduction:**
  - Mapping or Projecting on to an efficient feature space.
  - Wavelet Transforms, Principal Component Analysis
  - Attribute Subset Selection
    - Selecting only significant attributes.
    - n-attributes will have 2^n subsets in the simplest binary case, exhaustive search is prohibitively expensive when 'n' is large
    - Irrelevant, weakly relevant or redundant attributes are deleted or removed
- **Numerosity reduction:**
  - Parametric methods: a model is used to estimate the data, so that typically only the data parameters need to be stored, instead of the actual data. (Outliers may also be stored.)
    - Regression and log-linear models are examples.
  - Nonparametric methods for storing reduced representations of the data include
    - histograms, clustering, sampling (Section 3.4.8), and data cube aggregation
  - Regression Models, Histograms, clustering
- **Data Compression**
  - Lossless vs. lossy compression

Dimensionality Reduction: Discrete Wavelet Transforms (DWT)

- **Linear signal processing technique that transforms a data vector to a numerically different vector of wavelet coefficients.**
- **Properties**
  - The original and resulting vectors are of the same size
  - But can be truncated if the coefficients is smaller than user specified value.
  - Closely related to Discrete Fourier Transform (DFT), but achieve better lossy compression, i.e. retain more accurate approximation of the original data.
  - Better than JPEG for lossy image compression.

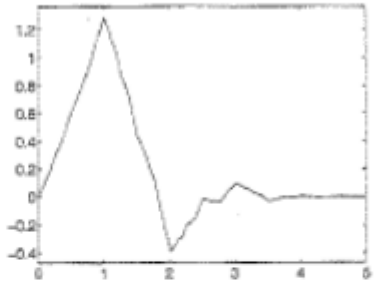
Dimensionality Reduction: Wavelets

- **There are many different wavelets:**



Haar

Morlet

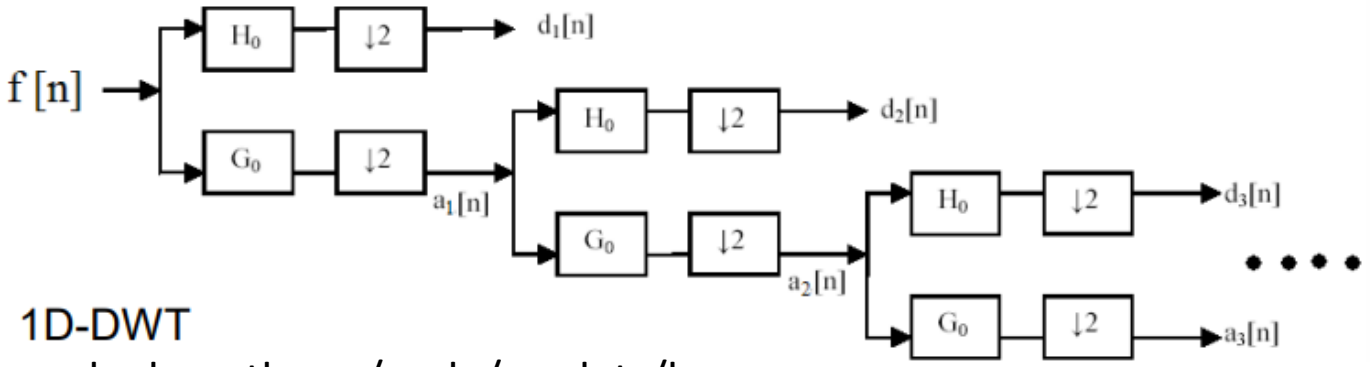


Daubechies

- **Wavelets are functions defined over a finite interval and having an average value of zero.**
- **Wavelet transform**
  - Time –frequency resolution
  - DWT the is successive LP and HP filtering.

A Hierarchical Pyramid Algorithm

1. **The length, L, of the input data vector must be an integer power of 2. This condition can be met by padding the data vector with zeros as necessary ($L \geq n$).**
2. **Each transform involves applying two functions. The first applies some data smoothing, such as a sum or weighted average. The second performs a weighted difference, which acts to bring out the detailed features of the data.**
3. **The two functions are applied to pairs of data points in X, that is, to all pairs of measurements (x_{2i}, x_{2i+1}). This results in two data sets of length L/2. In general, these represent a smoothed or low-frequency version of the input data and the high-frequency content of it, respectively.**
4. **The two functions are recursively applied to the data sets obtained in the previous loop, until the resulting data sets obtained are of length 2.**
5. **Selected values from the data sets obtained in the previous iterations are designated the wavelet coefficients of the transformed data.**



1D-DWT

http://www.whydomath.org/node/wavlets/hwt.html

Dimensionality Reduction: Wavelet Decomposition

- Wavelets: A math tool for space-efficient hierarchical decomposition of functions
- S = [2, 2, 0, 2, 3, 5, 4, 4] can be transformed to $S_{\wedge}$ = [$2^{3}/_{4}$, $-1^{1}/_{4}$, $^{1}/_{2}$, 0, 0, -1, -1, 0]
- Compression: many small detail coefficients can be replaced by 0's, and only the significant coefficients are retained

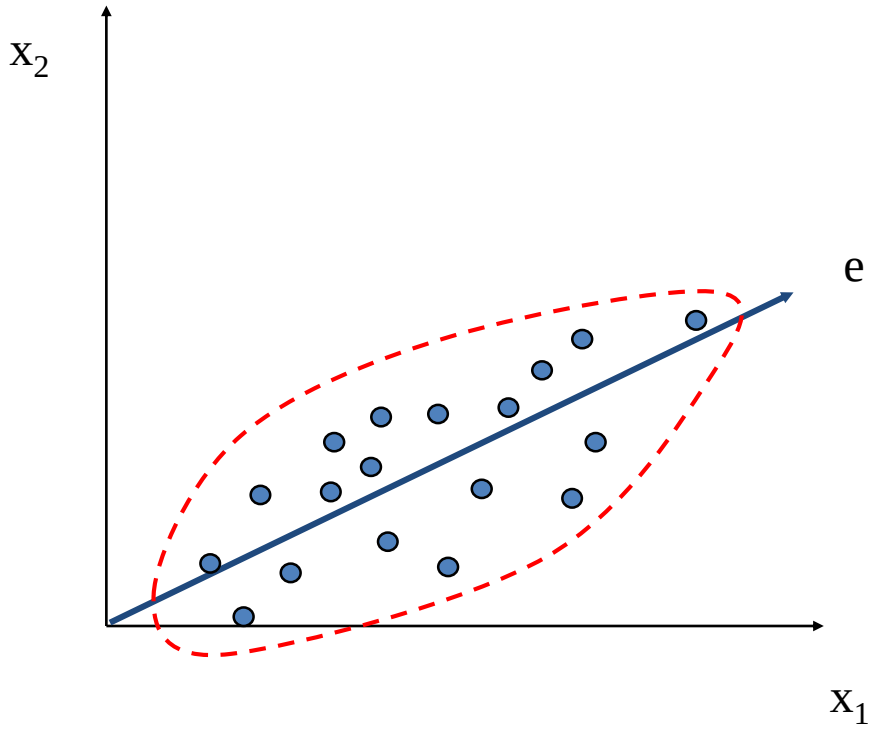
Resolution	Averages	Detail Coefficients
8	[2, 2, 0, 2, 3, 5, 4, 4]	
4	[2, 1, 4, 4]	[0, −1, −1, 0]
2	$[1\frac{1}{2}, 4]$	$[\frac{1}{2}, 0]$
1	$[2\frac{3}{4}]$	$[-1\frac{1}{4}]$

Summary: DWT

- **Compressed approximation: store only a small fraction of the strongest of the wavelet coefficients. Supports truncation.**
- **Effective removal of outliers:**
  - High frequency subbands consist of the details.
  - They might be omitted without substantially affecting the main features of the data set.
  - Additionally, these small details are often those associated with noise; therefore, by setting these coefficients to zero, we can essentially remove the noise.
- **Method:**
  - Length, L, must be an integer power of 2 (padding with 0s)
- **Difficult to apply on high dimensional data**

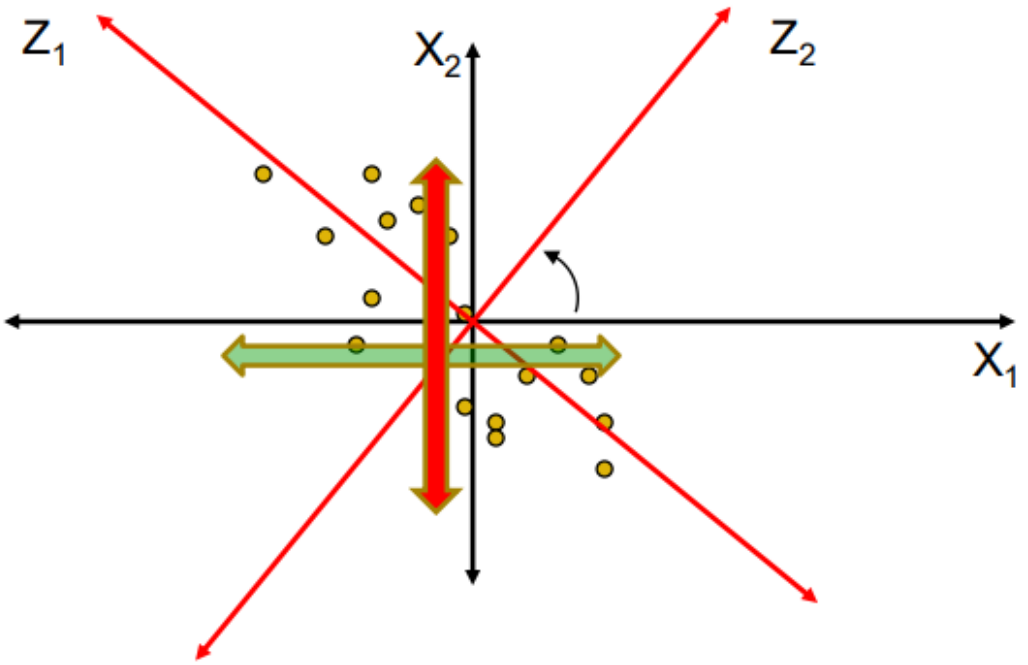
Dimensionality Reduction: Principal Component Analysis (PCA)

- 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



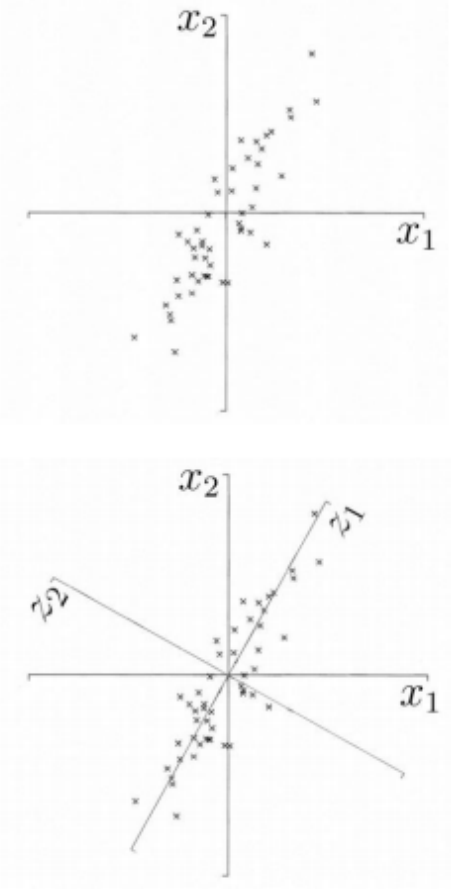
Dimensionality Reduction: Principal Component Analysis

- **Suppose we have attributes measured as p random variables (i.e. attributes)** $X_1, \ldots, X_p$.
  - These random variables represent the p-axes of the Cartesian coordinate system in which the population (data points, instances) resides.
  - Our goal is to develop a new set of p axes (linear combinations of the original p axes) in the directions of greatest variability.



Geometric picture of principal components (PCs)

- **The 1st PC Z_1 is a minimum distance fit to a line in the X space**
- **the 2nd PC Z_1 is a minimum distance fit to a line in the plane perpendicular to the 1st PC**
- **PCs are a series of linear least squares fits to a sample, each orthogonal to all the previous.**



https://en.wikipedia.org/wiki/Principal_component_analysis

Principal Component Analysis (steps)

- **Given *N* data vectors from *n*-dimensions, find $k \leq n$ 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. Complexity increases with size.**

Step By Step Computation Of PCA

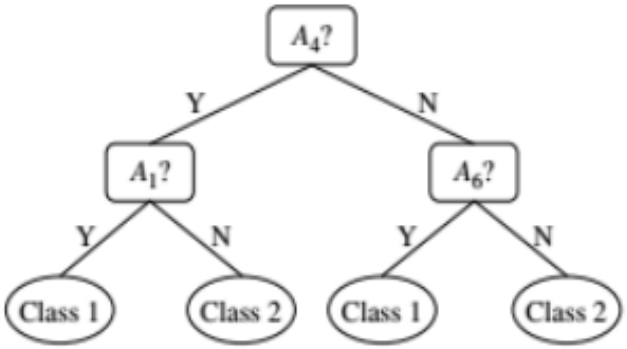
1. Standardization of the data
2. Computing the covariance matrix
3. Calculating the Eigenvectors and Eigenvalues
4. Computing the Principal Components
5. Reducing the dimensions of the data set

Dimensionality Reduction: Attribute Subset Selection

- **Another way to reduce dimensionality of data**
- **Redundant attributes**
  - Duplicate much or all of the information contained in one or more other attributes
  - E.g., purchase price of a product and the amount of sales tax paid
- **Irrelevant attributes**
  - Contain no information that is useful for the data mining task at hand
  - E.g., students' ID is often irrelevant to the task of predicting students' GPA' for performance analysis. Relevant attributes are 'credit hours', 'subject wise grade', ….

Heuristic (Greedy) Search in Attribute Selection

- **There are 2^d possible attribute combinations of d attributes**
- **Typical heuristic attribute selection methods:**
  - **Best single attribute** under the attribute independence assumption: choose by significance tests (*information gain, decision tree, SVM, ..*)
  - Best step-wise **attribute selection**:
    - The best single-attribute is picked first
    - Then next best attribute condition to the first, ...
  - Step-wise **attribute elimination**:
    - Repeatedly eliminate the worst attribute
  - Best **combined attribute selection and elimination**
  - Optimal branch and bound:
    - Use attribute elimination and backtracking

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ $\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ $\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  $\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$

Numerosity Reduction

- **Reduce data volume by choosing alternative, *smaller forms* of data representation**
- **Parametric methods (e.g., regression)**
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
- **Non-parametric methods**
  - Major families: histograms, clustering, sampling, …

Highlights of Parametric methods

- For parametric methods (Regression and log-linear models), a model is used to estimate the data, so that typically only the data parameters need to be stored, instead of the actual data. Outliers may also be stored.

Numerosity Reduction: Regression Analysis and Log-Linear Models

- **Linear regression**: $Y = w X + b$
  - Two regression coefficients, *w* and *b*, specify the line and are to be estimated by using the data at hand
  - Using the least squares criterion to the known values of Y_1, Y_2, …, X_1, X_2, ….
- **Multiple regression (two or more predictors)**:
  - $Y = b_0 + b_1 X_1 + b_2 X_2$
  - Many nonlinear functions can be transformed into the above
- **Log-linear models (log response)**: $\log Y_i = \alpha + \beta X i + \epsilon_i$

-- Approximate discrete multidimensional probability distributions

– Estimate the probability of each point (tuple) in a multi-dimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations

Numerosity Reduction: Histogram

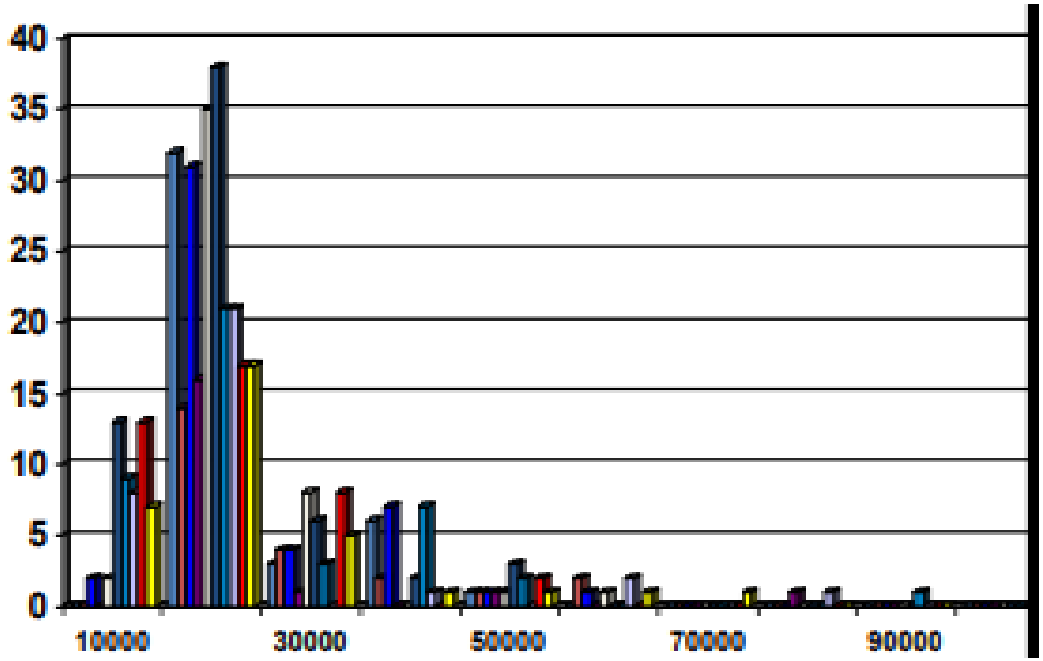
- Histograms use binning to approximate data distributions and are a popular form of data reduction.
- A histogram for an attribute, A, partitions the data distribution of **A** into disjoint subsets, referred to as buckets or bins.
- If each bucket represents only a single attribute–value/frequency pair, the buckets are called singleton buckets.
- Singleton buckets are useful for storing high-frequency outliers.

Numerosity Reduction:Histogram Analysis

- To further reduce the data, it is common to divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
  - ■ Equal-width: In an equal-width histogram, the width of each bucket range is constant
  - ■ Equal-frequency (or equal-depth): In an equal-frequency histogram, the buckets are created so that, roughly, the frequency of each bucket is constant (i.e., each bucket contains roughly the same number of contiguous data samples).

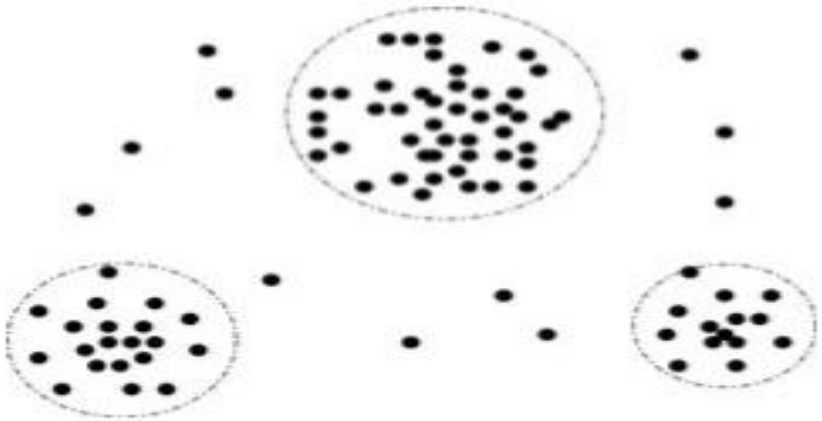
Numerosity Reduction:Histogram Analysis

- Histograms are highly effective at approximating both sparse and dense data, as well as highly skewed and uniform data



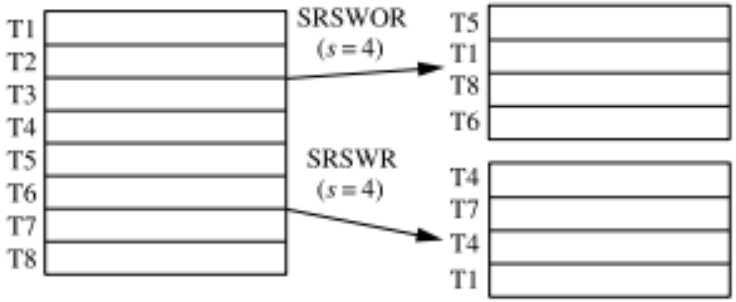
Numerosity Reduction:Clustering

- 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"
- There are many choices of clustering definitions and clustering algorithms
- Ex.- A 2-D plot of customer data with respect to customer locations in a city. Three data clusters are visible.



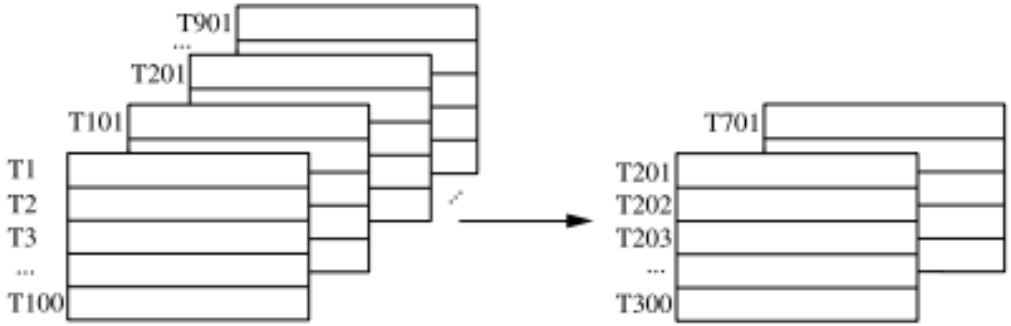
Numerosity Reduction:Sampling

- Sampling can be used as a data reduction technique because it allows a large data set to be represented by a much smaller random data sample (or subset).
- Suppose that a large data set, D, contains N tuples. Let's look at the most common ways that we could sample D for data reduction



Cluster sample

($s = 2$)



Startified sample

(according to *age*)

T38	youth
T256	youth
T307	youth
T391	youth
T96	middle_aged
T117	middle_aged
T138	middle_aged
T263	middle_aged
T290	middle_aged
T308	middle_aged
T326	middle_aged
T387	middle_aged
T69	senior
T284	senior

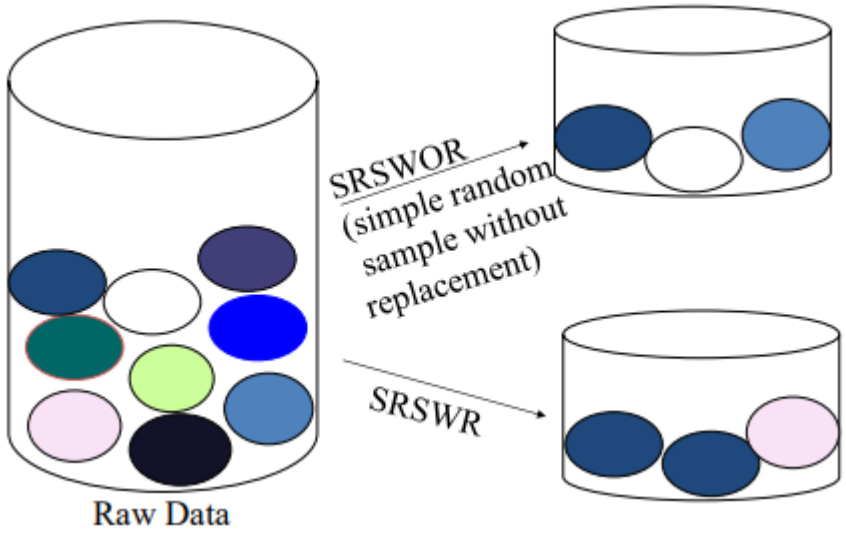
T38	youth
T391	youth
T117	middle_aged
T138	middle_aged
T290	middle_aged
T326	middle_aged
T69	senior

Types of Sampling

- ■ **Simple random sample without replacement (SRSWOR) of size s:** This is created by drawing s of the N tuples from D (s < N), where the probability of drawing any tuple in D is 1/N, that is, all tuples are equally likely to be sampled.

- ■ **Simple random sample with replacement (SRSWR) of size s:** This is similar to SRSWOR, except that each time a tuple is drawn from D, it is recorded and then replaced. That is, after a tuple is drawn, it is placed back in D so that it may be drawn again.

Sampling: With or without Replacement



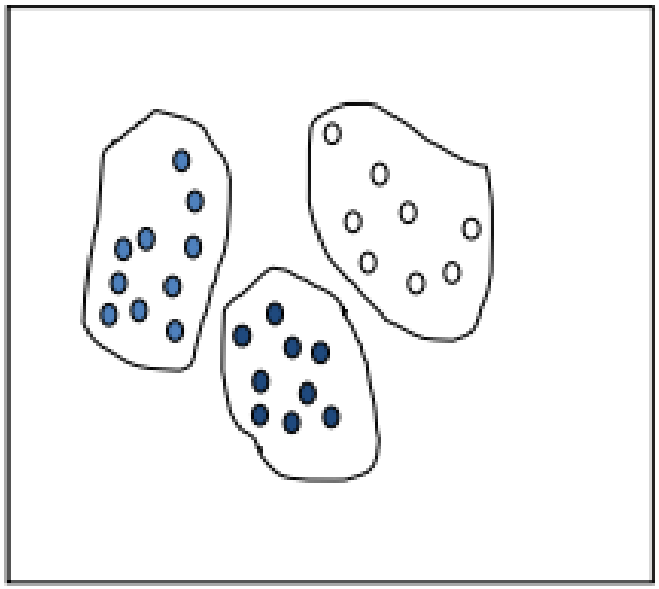
Types of Sampling

Cluster sample: If the tuples in D are grouped into M mutually disjoint "clusters," then an SRS of s clusters can be obtained, where s < M. For example, tuples in a database are usually retrieved a page at a time, so that each page can be considered a cluster. A reduced data representation can be obtained by applying, say, SRSWOR to the pages, resulting in a cluster sample of the tuples. Other clustering criteria conveying rich semantics can also be explored. For example, in a spatial database, we may choose to define clusters geographically based on how closely different areas are located.

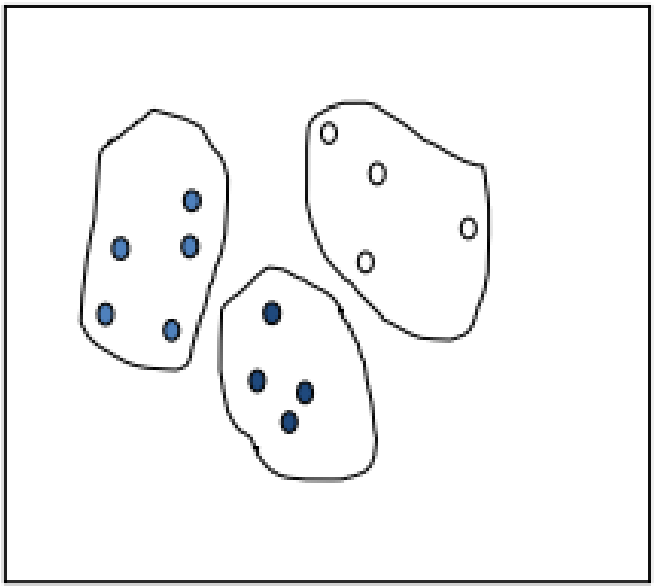
■ **Stratified sample**: If D is divided into mutually disjoint parts called strata, a stratified sample of D is generated by obtaining an SRS at each stratum. This helps ensure a representative sample, especially when the data are skewed. For example, a stratified sample may be obtained from customer data, where a stratum is created for each customer age group. In this way, the age group having the smallest number of customers will be sure to be represented.

Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample



Sampling Example

- The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
- Sketch examples of each of the following sampling techniques: SRSWOR, SRSWR, cluster sampling, stratified sampling. Use samples of size 5 and the strata “youth”, “middle-aged”, and “senior”.

Tuples

T1	13
T2	15
T3	16
T4	16
T5	19
T6	20
T7	20
T8	21
T9	22

T10	22
T11	25
T12	25
T13	25
T14	25
T15	30
T16	33
T17	33
T18	33

T19	33
T20	35
T21	35
T22	36
T23	40
T24	45
T25	46
T26	52
T27	70

SRSWOR vs. SRSWR

SRSWOR	(n = 5)
T4	16
T6	20
T10	22
T11	25
T26	32

SRSWR	(n = 5)
T7	20
T7	20
T20	35
T21	35
T25	46

Clustering sampling: Initial clusters

T1	13
T2	15
T3	16
T4	16
T5	19

T6	20
T7	20
T8	21
T9	22
T10	22

T11	25
T12	25
T13	25
T14	25
T15	30

T16	33
T17	33
T18	33
T19	33
T20	35

T21	35
T22	36
T23	40
T24	45
T25	46

T26	52
T27	70

Cluster sampling (m = 2)

T6	20
T7	20
T8	21
T9	22
T10	22

T21	35
T22	36
T23	40
T24	45
T25	46

Stratified Sampling

T1	13	young
T2	15	young
T3	16	young
T4	16	young
T5	19	young
T6	20	young
T7	20	young
T8	21	young
T9	22	young

T10	22	young
T11	25	young
T12	25	young
T13	25	young
T14	25	young
T15	30	middle age
T16	33	middle age
T17	33	middle age
T18	33	middle age

T19	33	middle age
T20	35	middle age
T21	35	middle age
T22	36	middle age
T23	40	middle age
T24	45	middle age
T25	46	middle age
T26	52	middle age
T27	70	senior

Stratified Sampling (according to age)

T4	16	young
T12	25	young
T17	33	middle age
T25	46	middle age
T27	70	senior

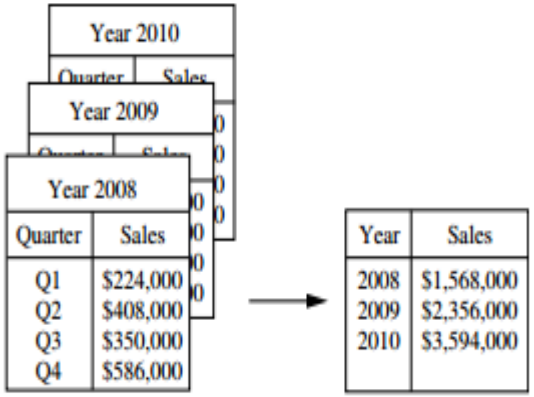
Data Cube Aggregation

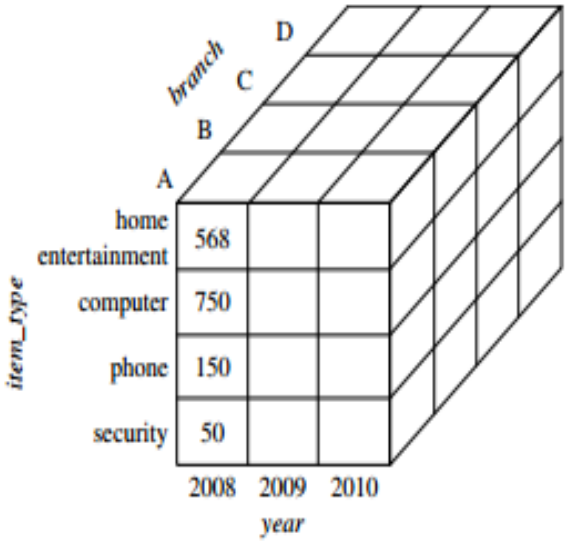
- Data cubes store multidimensional aggregated information.
- A data cube for multidimensional analysis of sales data with respect to annual sales per item type for each AllElectronics branch. Each cell holds an aggregate data value, corresponding to the data point in multidimensional space. (For readability, only some cell values are shown.)
- Concept hierarchies may exist for each attribute, allowing the analysis of data at multiple abstraction levels. For example, a hierarchy for branch could allow branches to be grouped into regions, based on their address. Data cubes provide fast access to precomputed, summarized data, thereby benefiting online analytical processing as well as data mining.

Data Cube Aggregation

- Imagine that you have collected the data for your analysis. These data consist of the AllElectronics sales per quarter, for the years 2008 to 2010. You are, however, interested in the annual sales (total per year), rather than the total per quarter. Thus, the data can be aggregated so that the resulting data summarize the total sales per year instead of per quarter.

Data Cube Aggregation





Data Transformation

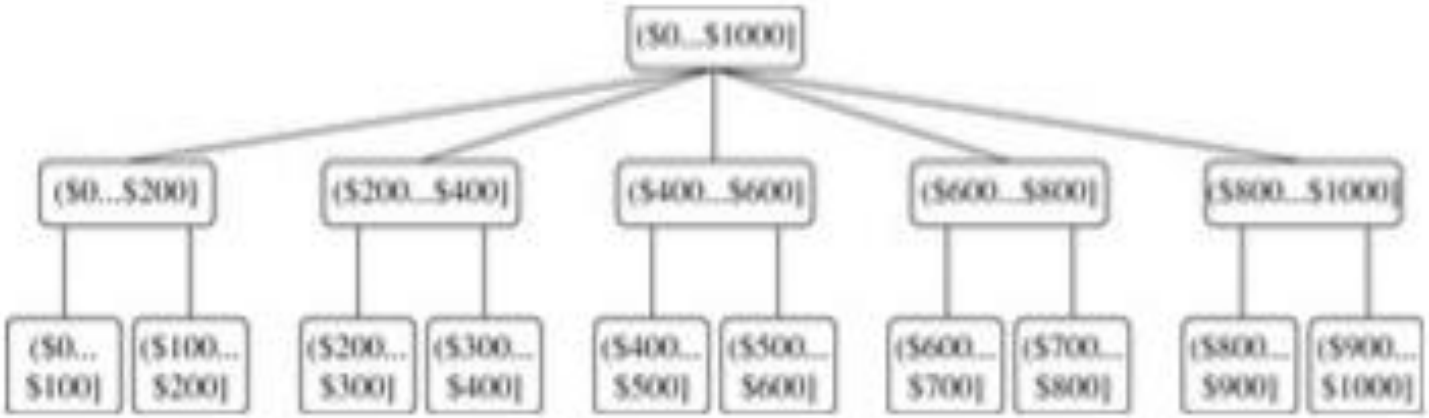
- In data transformation, the data are transformed or consolidated into forms appropriate for mining. Strategies for data transformation include the following:

1. **Smoothing** : works to remove noise from the data. Techniques include binning, regression, and clustering.
2. **Attribute construction** (or feature construction), where new attributes are constructed and added from the given set of attributes to help the mining process.
3. **Aggregation**, where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for data analysis at multiple abstraction levels.
4. **Normalization**, where the attribute data are scaled so as to fall within a smaller range, such as −1.0 to 1.0, or 0.0 to 1.0.

   Min-Max Normalization, Z-score Normalization, Decimal Scaling Normalization
5. **Discretization**, where the raw values of a numeric attribute (e.g., age) are replaced by interval labels (e.g., 0–10, 11–20, etc.) or conceptual labels (e.g., youth, adult, senior). The labels, in turn, can be recursively organized into higher-level concepts, resulting in a concept hierarchy for the numeric attribute.

Data Transformation

6. **Concept hierarchy** generation for nominal data, where attributes such as street can be generalized to higher-level concepts, like city or country. Many hierarchies for nominal attributes are implicit within the database schema and can be automatically defined at the schema definition level.



A concept hierarchy for the attribute price

Normalization

- **Min-max normalization**:Min-max normalization performs a linear transformation on the original data. Suppose that min_A and max_A are the minimum and maximum values of an attribute, A. Min-max normalization maps a value, $\mathbf{v_i}$, of A to $\mathbf{v_i}$'in the range [new_minA, new_maxA] by computing

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

  - 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 normalization** (μ: mean, σ: standard deviation): In z -score normalization (or zero-mean normalization), the values for an attribute, A, are normalized based on the mean (i.e., average) and standard deviation of A. A value, $\mathbf{v_i}$, of A is normalized to $\mathbf{v_i'}$ by computing

$$v_i' = \frac{v_i - \overline{A}}{\sigma_A}$$

Ex. Let $\overline{A}$ = 54,000, σ = 16,000. Then, 73,600 maps to $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

Normalization

- A variation of this z-score normalization replaces the standard deviation by the mean absolute deviation of **A**. The mean absolute deviation of **A**, denoted s_A, is

$$SA = \frac{1}{N}(|\ v1 - \overline{A}| + |\ v2 - \overline{A}| + \ldots\ldots + |\ vN - \overline{A}|)$$

  - Thus, z-score normalization using the mean absolute deviation is

$$v_i' = \frac{v_i - \overline{A}}{s_A}$$

- **The mean absolute deviation, s_A, is more robust to outliers than the standard deviation, σ_A. When computing the mean absolute deviation, the deviations from the mean are not squared; hence, the effect of outliers is somewhat reduced.**

Normalization

- **Normalization by decimal scaling : It normalizes by moving the decimal point of values of attribute A. The number of decimal points moved depends on the maximum absolute value of A. A value, v_i, of A is normalized to v'_i by computing**

$$v' = \frac{v}{10^j}$$

Ex. The values of A range from −986 to 917. The maximum absolute value of A is 986. To normalize by decimal scaling, we therefore divide each value by 1000 (i.e., j = 3) so that −986 normalizes to −0.986 and 917 normalizes to 0.917.

Discretization

- **Data discretization** transforms numeric data by mapping values to interval or concept labels. Such methods can be used to automatically generate concept hierarchies for the data, which allows for mining at multiple levels of granularity.
- **Discretization techniques** include binning, histogram analysis, cluster analysis, decision tree analysis, and correlation analysis.
- For **nominal data**, concept hierarchies may be generated based on schema definitions as well as the number of distinct values per attribute.

Concept Hierarchy generation for Nominal data

- The attributes such as street can be generalized to higher level concepts, like city or country.
- Many hierarchies for nominal attributes are implicit within database schema and can automatically defined at the schema definition level.