18CSE355T - DATA MINING AND ANALYTICS

18CSE355T - DATA MINING AND ANALYTICS

Course Code	18CSE355T	Course Name	DATA MINING AND ANALYTICS	Course Category	E	Professional Elective	L	T	P	C
							3	0	0	3

Pre-requisite Courses	Nil	Co-requisite Courses	Nil	Progressive Courses	Nil
Course Offering Department	CSE	Data Book / Codes/Standards		Nil	

Course Learning Rationale (CLR):	*The purpose of learning this course is to:*
CLR-1 :	*Understand the concepts of Data Mining*
CLR-2 :	*Familiarize with Association rule mining*
CLR-3 :	*Familiarize with various Classification algorithms*
CLR-4 :	*Understand the concepts of Cluster Analysis*
CLR-5 :	*Familiarize with Outlier analysis techniques*
CLR-6 :	*Familiarize with applications of Data mining in different domains*

Course Learning Outcomes (CLO):	*At the end of this course, learners will be able to:*	Learning			Program Learning Outcomes (PLO)														
		1	2	3	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
		Level of Thinking (Bloom)	Expected Proficiency (%)	Expected Attainment (%)	Engineering Knowledge	Problem Analysis	Design & Development	Analysis, Design, Research	Modern Tool Usage	Society & Culture	Environment & Sustainability	Ethics	Individual & Team Work	Communication	Project Mgt. & Finance	Life Long Learning	PSO - 1	PSO - 2	PSO – 3
CLO-1 :	*Gain knowledge about the concepts of Data Mining*	2	80	85															
CLO-2 :	*Understand and Apply Association rule mining techniques*	2	75	80															
CLO-3 :	*Understand and Apply various Classification algorithms*	2	85	80															
CLO-4 :	*Gain knowledge on the concepts of Cluster Analysis*	2	80	75															
CLO-5 :	*Gain knowledge on Outlier analysis techniques*	2	75	85															
CLO-6 :	*Understand the importance of applying Data mining concepts in different domains*	2	80	85															

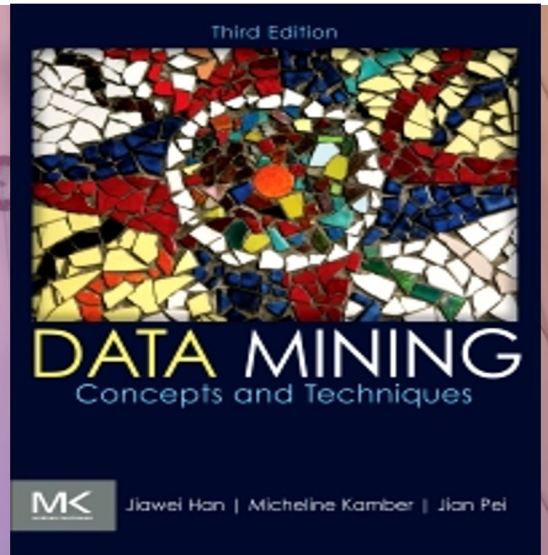
Duration (hour)		9	9	9	9	9
S-1	SLO-1	*Why Data mining? What is Data mining ?*	*Mining frequent patterns: Basic concepts*	*Classification: Basic concepts*	*Cluster Analysis: Introduction*	*Outliers: Introduction*
	SLO-2	*Kinds of data meant for mining*	*Market Basket Analysis*	*General approach to Classification*	*Requirements and overview of different categories*	*Challenges of outlier detection*
S-2	SLO-1	*Kinds of patterns that can be mined*	*Frequent itemsets, Closed itemsets*	*Decision tree induction*	*Partitioning method: Introduction*	*Outlier detection methods: Introduction*
	SLO-2	*Applications suitable for data mining*	*Association rules-Introduction*	*Algorithm for Decision tree induction*	*k-means*	*Supervized and Semi-supervized methods*
S-3	SLO-1	*Issues in Data mining*	*Apriori algorithm-theoritical approach*	*Numerical example for Decision tree induction*	*k-medoids*	*Unsupervized methods*
	SLO-2	*Data objects and Attribute types*	*Apply Apriori algorithm on dataset-1*	*Attribute selection measure*	*Hierarchical method: Introduction*	
S-4	SLO-1	*Statistical descriptions of data*	*Apply Apriori algorithm on dataset-2*	*Tree pruning*	*Agglomerative vs. Divisive method*	*Statistical and Proximity based methods*
	SLO-2		*Generating Association rules from frequent itemsets*	*Scalability and Decision tree induction*	*Distance measures in algorithmic methods*	
S-5	SLO-1	*Need for data preprocessing and data quality*	*Improving efficiency of Apriori*	*Bayes' Theorem*	*BIRCH technique*	*Statistical approaches*
	SLO-2			*Naïve Bayesian Classification*		
S-6	SLO-1	*Data cleaning*	*Pattern growth approach*	*IF-THEN rules for classification*	*DBSCAN technique*	*Statistical data mining*
	SLO-2	*Data integration*		*Rule extraction from a decision tree*		
S-7	SLO-1	*Data reduction*	*Mining frequest itemsets using Vertical data format*	*Metrics for evaluating classifier performance*	*STING technique*	*Data mining and recommender systems*
	SLO-2		*Strong rules vs. weak rules*	*Cross validation*		
S-8	SLO-1	*Data transformation*	*Association analysis to Correlation analysis*	*Bootstrap*	*CLIQUE technique*	*Data mining for financial data analysis*
	SLO-2			*Ensemble methods-Introduction*		
S-9	SLO-1	*Data cube and its usage*	*Comparison of pattern evaluation measures*	*Bagging and Boosting*	*Evaluation of clustering techniques*	*Data mining for Intrusion detection*
	SLO-2			*Random Forests: Introduction*		

Learning Resources	1. *Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", 3rd Edition, Morgan Kauffman Publishers, 2011.*	

Learning Assessment	Bloom's Level of Thinking	Continuous Learning Assessment (50% weightage)								Final Examination (50% weightage)	
		CLA – 1 (10%)		CLA – 2 (15%)		CLA – 3 (15%)		CLA – 4 (10%)#			
		Theory	Practice	Theory	Practice	Theory	Practice	Theory	Practice	Theory	Practice
Level 1	Remember / Understand	40 %	-	30 %	-	30 %	-	30 %	-	30%	-
Level 2	Apply / Analyze	40 %	-	40 %	-	40 %	-	40 %	-	40%	-
Level 3	Evaluate / Create	20 %	-	30 %	-	30 %	-	30 %	-	30%	-
	Total	100 %		100 %		100 %		100 %		100 %	

CLA – 4 can be from any combination of these: Assignments, Seminars, Tech Talks, Mini-Projects, Case-Studies, Self-Study, MOOCs, Certifications, Conf. Paper etc.,

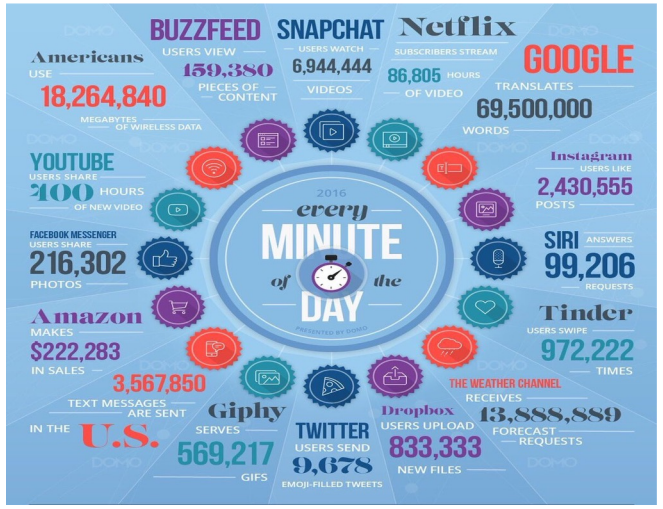
Course Designers		
Experts from Industry	Experts from Higher Technical Institutions	Internal Experts
1. Mr.V.Selvakumar, Hexaware Technologies, selvakumarv@hexaware.com	1. Dr.Latha Parthiba, Pondicherry University, lathaparthiban@yahoo.com	1. Mr.L.N.B.Srinivas, SRMIST
2.	2.	2. Mr.S.Karthick, SRMIST
		3. Dr.V.V.Ramalingam, SRMIST



UNIT - 1

Why Data mining? What is Data mining ?
Kinds of data meant for mining
Kinds of patterns that can be mined
Applications suitable for data mining
Issues in Data mining
Data objects and Attribute types
Statistical descriptions of data
Need for data preprocessing and data quality
Data cleaning
Data integration
Data reduction
Data transformation
Data cube and its usage

Why Data mining?



- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- <u>We are drowning in data, but starving for knowledge!</u>
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

Abbreviation	Unit	Value	Size (in bytes)
b	bit	0 or 1	1/8 of a byte
B	bytes	8 bits	1 byte
KB	kilobytes	1,000 bytes	1,000 bytes
MB	megabyte	$1,000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1,000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1,000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1,000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1,000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1,000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1,000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

What is Data mining ?

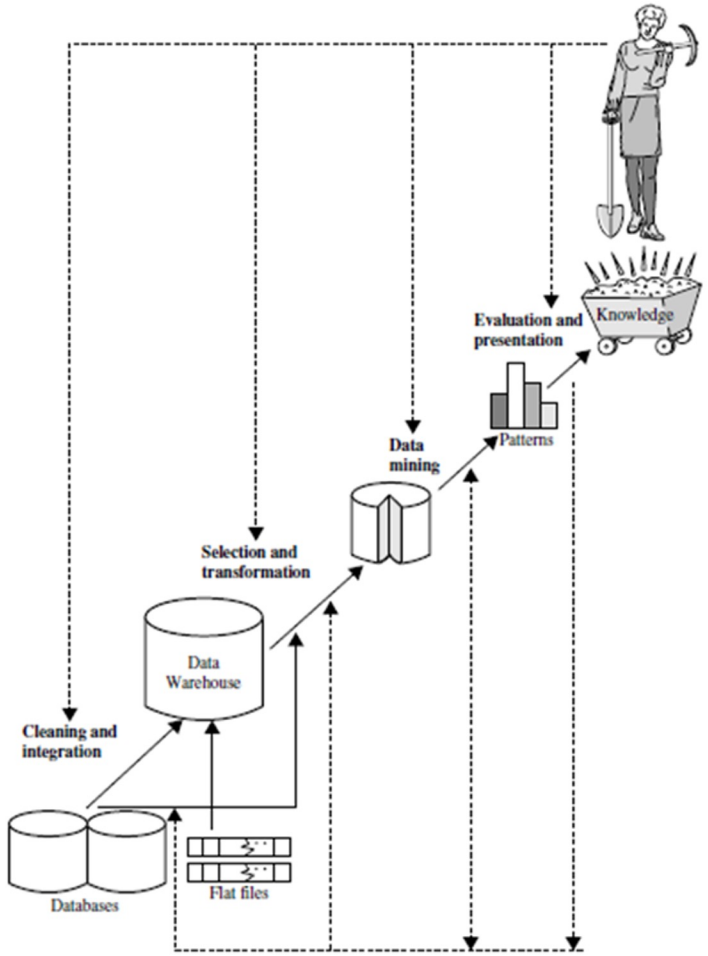
Data mining (knowledge discovery from data)

- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Data mining: a misnomer?

Alternative names

- Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

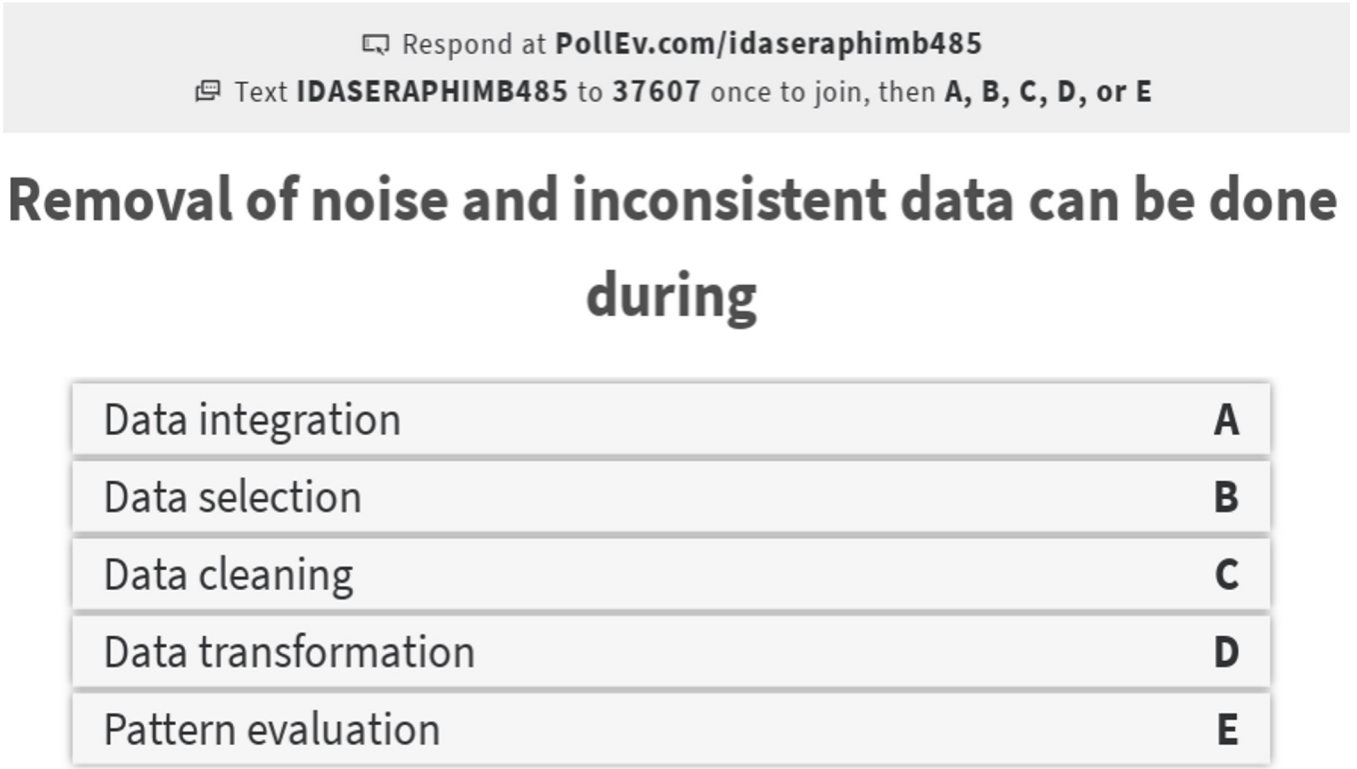
Knowledge discovery process



Data processing

1. Data cleaning (remove noise and inconsistent data)
2. Data integration (multiple data sources maybe combined)
3. Data selection (data relevant to the analysis task are retrieved from database)
4. Data transformation (data transformed or consolidated into forms appropriate for mining) (Done with data preprocessing)
5. Data mining (an essential process where intelligent methods are applied to extract data patterns)
6. Pattern evaluation (identify the truly interesting patterns)
7. Knowledge presentation (mined knowledge is presented to the user with visualization or representation techniques)

Try me…

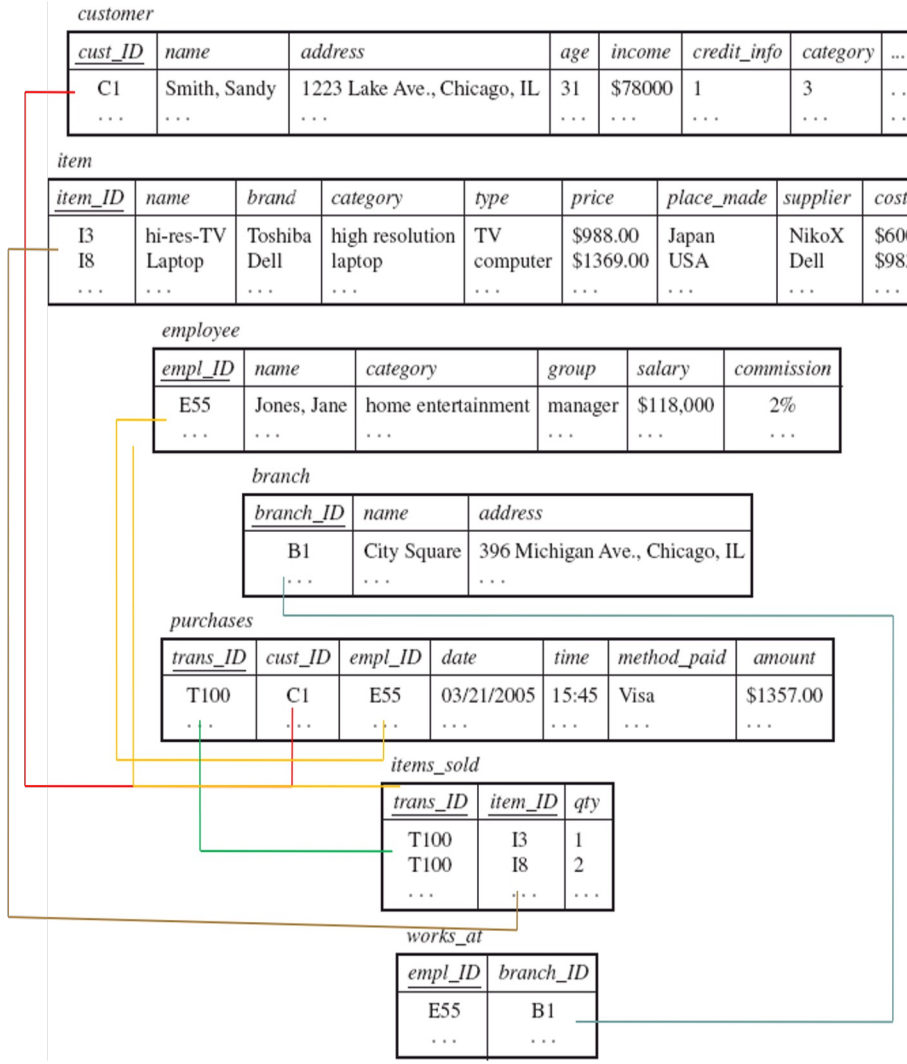


https://PollEv.com/multiple_choice_polls/iVtbiG1lVVqTQ2WsSgXTC/respond

What Kind of Data Can Be mined?

- Database-oriented data sets and applications
- Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications.
- Object-Relational Databases
- Temporal Databases, Sequence Databases, Time-Series databases
- Spatial Databases and Spatiotemporal Databases
- Text databases and Multimedia databases
- Heterogeneous Databases and Legacy Databases
- Data Streams
- The World-Wide Web

Relational Databases



customer

cust_ID	*name*	*address*	*age*	*income*	*credit_info*	*category*	...
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	$78000	1	3	...
...	...	...	...	...	...	...	...

item

item_ID	*name*	*brand*	*category*	*type*	*price*	*place_made*	*supplier*	*cost*
I3	hi-res-TV	Toshiba	high resolution	TV	$988.00	Japan	NikoX	$600.00
I8	Laptop	Dell	laptop	computer	$1369.00	USA	Dell	$983.00
...	...	...	...	...	...	...	...	...

employee

empl_ID	*name*	*category*	*group*	*salary*	*commission*
E55	Jones, Jane	home entertainment	manager	$118,000	2%
...	...	...	...	...	...

branch

branch_ID	*name*	*address*
B1	City Square	396 Michigan Ave., Chicago, IL
...	...	...

purchases

trans_ID	*cust_ID*	*empl_ID*	*date*	*time*	*method_paid*	*amount*
T100	C1	E55	03/21/2005	15:45	Visa	$1357.00
...	...	...	...	...	...	...

items_sold

trans_ID	*item_ID*	*qty*
T100	I3	1
T100	I8	2
...	...	...

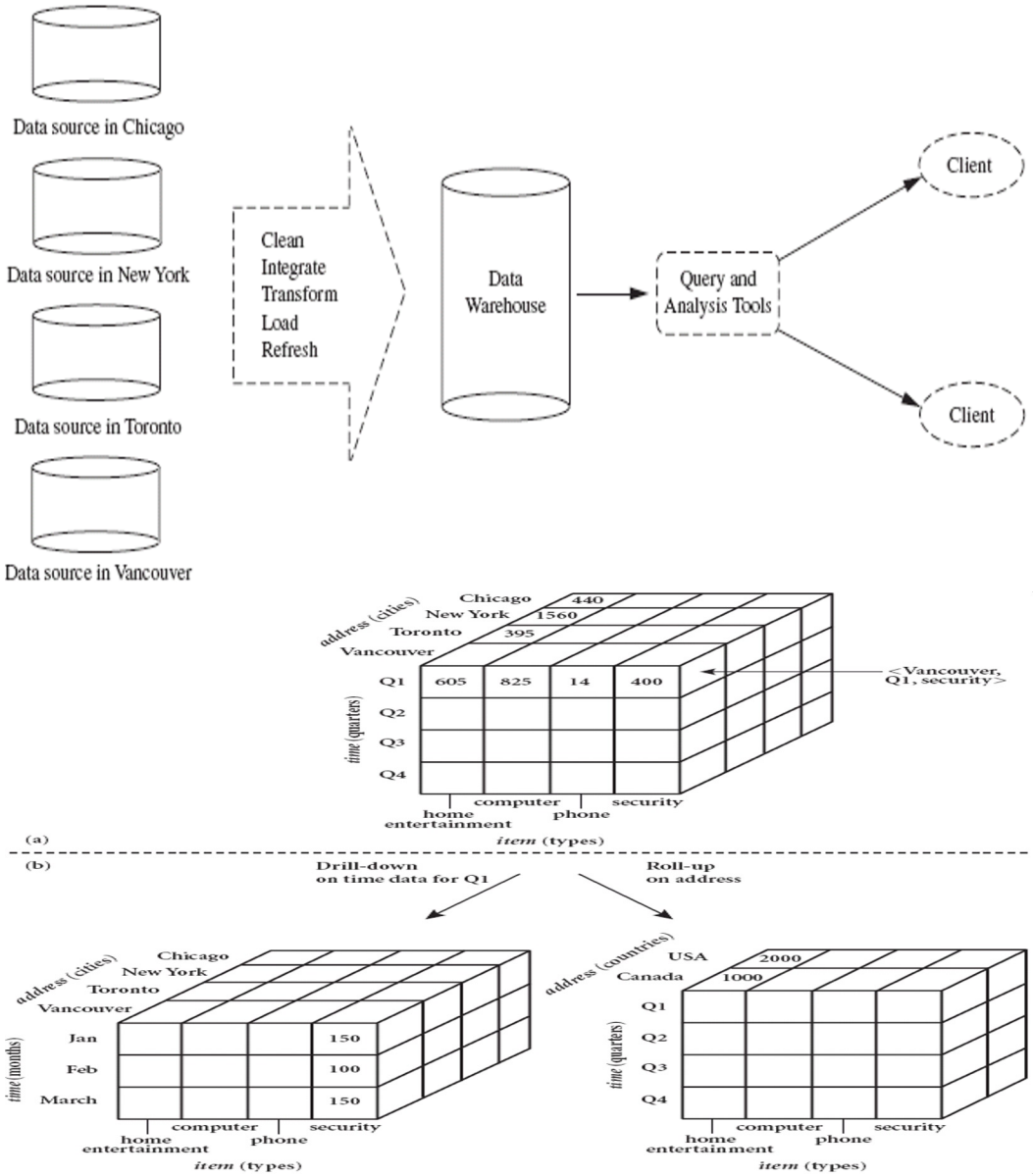
works_at

empl_ID	*branch_ID*
E55	B1
...	...

- DBMS – database management system, contains a collection of

  interrelated data's from databases.
- Each table contains set of attributes(columns) and set of tuples(rows), with columns as attributes of data and rows as records.
- Tables can be used to represent the relationships between or among multiple tables.
- Each tuple in the relational table represents an object identified by a unique key and described by a set of attribute values.
- Relational data can be accessed by database queries (SQL).

Data Warehouses



- A repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site.
- Constructed via a process of data cleaning, data integration, data transformation, data loading and periodic data refreshing.
- User can perform drill-down or roll-up operation to view the data at different degrees of summarization.

Transactional Databases

trans_ID	*list of item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
. . .	. . .

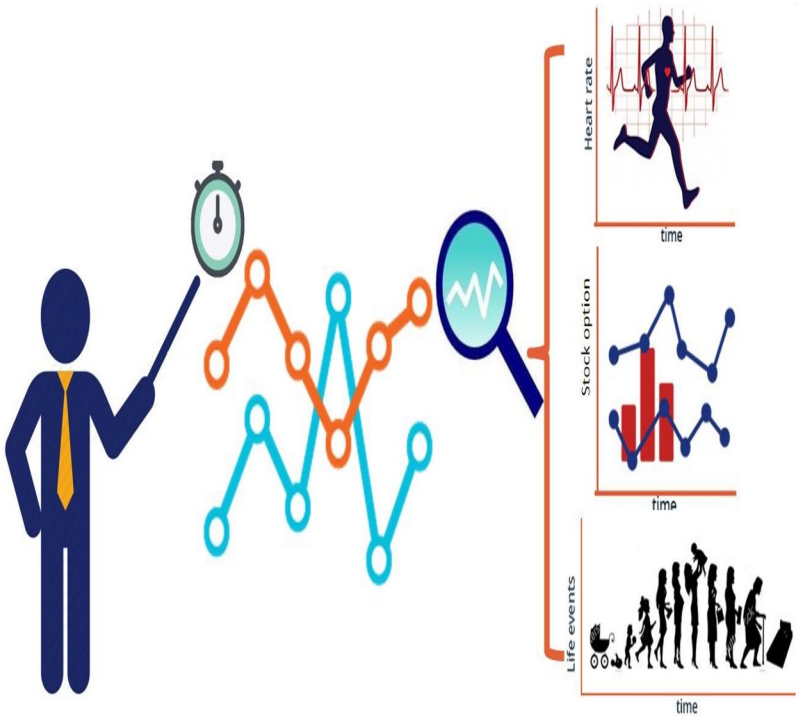
- Consists of a file where each record represents a transaction.
- A transaction typically includes a unique transaction ID and a list of the items making up the transaction.
- Easy to identify items that are frequently sold together.
- Market Basket Analysis enable you to bundle groups of items together. Eg: printers are commonly purchased together with computers.

Object-Relational Databases

- Object-relational databases are constructed based on an object-relational data model.
- This model extends the relational model by providing a rich data type for handling complex objects and object orientation.
- Inherits the essential concepts of object-oriented databases.
- AllElectronics example, objects can be individual employees, customers, or items.
- Data and code relating to an object are encapsulated into a single unit.
- Objects that share a common set of properties can be grouped into an object class.

  - A **set of variables** that describe the objects.
  - A **set of messages** that the object can use to communicate with other objects.
  - A **set of methods**, where each method holds the code to implement a message.

Temporal Databases, Sequence Databases, and Time-Series Databases



- A **temporal database** typically stores relational data that include time-related attributes.
- A **sequence database** stores sequences of ordered events, with or without a concrete notion of time.

  E.g.: customer shopping sequences.
- A **time-series database** stores sequences of values or events obtained over repeated measurements of time (e.g., hourly, daily, weekly).

  E.g.: data collected from the stock exchange, inventory control.

Spatial Databases and Spatiotemporal Databases

- **Spatial databases** contain spatial-related information.

  E.g.: geographic(map) databases commonly used in vehicle navigation and dispatching systems.

- A spatial database that stores spatial objects that change with time is called a **spatiotemporal database.**

Text Databases and Multimedia Databases

- **Text databases** are databases that contain word descriptions for objects.

  E.g.: product specifications, error or bug reports, warning messages etc.

- **Multimedia databases** store image, audio, and video data. They are used in applications such as picture content-based retrieval, voice-mail systems, video-on-demand systems etc.

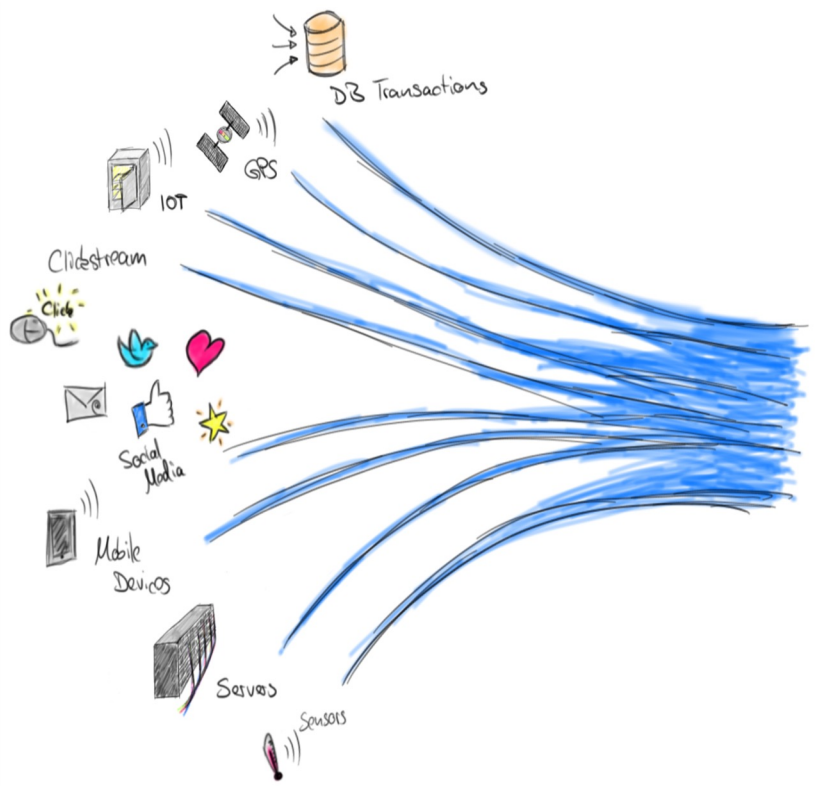
Heterogeneous Databases and Legacy Databases

- A **heterogeneous database** consists of a set of interconnected, autonomous component databases.

- A **legacy database** is a group of *heterogeneous databases* that combines different kinds of data systems.

  E.g.: relational or object-oriented databases, hierarchical databases, network databases, spread sheets, multimedia databases, or file systems.

Data Streams



- Many applications involve the generation and analysis of a new kind of data, called **stream data**.
- where data flow in and out of an observation platform (or window) dynamically.

Unique features

- huge or possibly infinite volume.
- dynamically changing.
- flowing in and out in a fixed order.
- allowing only one or a small number of scans.
- demanding fast (often real-time) response time.

E.g.: scientific and engineering data, time-series data, stock exchange,video surveillance, network traffic, weather or environment monitoring.

The World Wide Web

- Capturing user access patterns in such distributed information environments is called **Web usage mining (or Weblog mining).**
- **Authoritative Web page** analysis based on linkages among Web pages can help rank Web pages based on their importance, influence, and topics.
- **Automated Web page clustering and classification** help group and arrange Web pages in a multidimensional manner based on their contents.
- **Web community analysis** helps identify hidden Web social networks and communities and observe their evolution.

Kinds of pattern that can be mined

Data mining functionalities

1. Characterization and discrimination
2. The mining of frequent patterns, associations, and correlations
3. Classification and regression
4. Clustering analysis
5. Outlier analysis

Data Mining Tasks

- **Descriptive mining** tasks characterize the general properties of the data in the database.
- **Predictive mining** tasks perform inference on the current data in order to make predictions.

Data can be associated with classes or concepts.

E.g. classes of items – computers, printers, …

concepts of customers – bigSpenders, budgetSpenders, …

Data characterization

- summarizing the general characteristics of a target class of data.

  E.g. summarizing the characteristics of customers who spend more than $1,000 a year at AllElectronics.

Data Discrimination

- comparing the target class with one or a set of comparative classes.
  - E.g. Compare the general features of software products whole sales increase by 10% in the last year with those whose sales decrease by 30% during the same period

Mining Frequent Patterns, Associations and Correlations

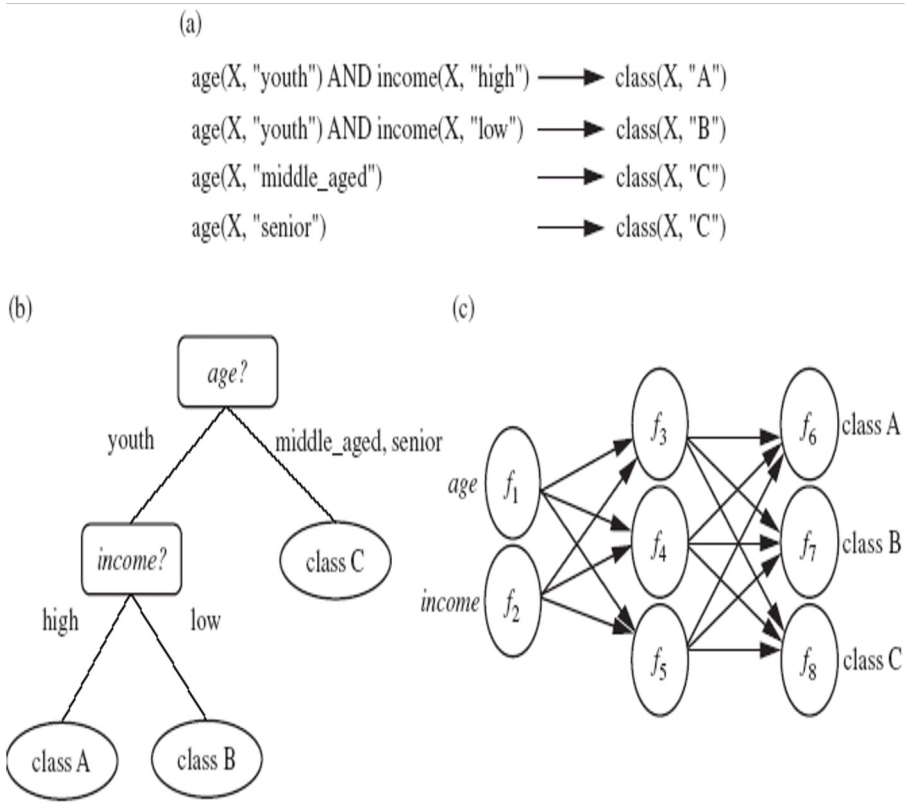
- **Frequent itemset:** a set of items that frequently appear together in a transactional data set. (e.g. milk and bread)

- **Association Analysis:** find frequent patterns
  - E.g. a sample analysis result – an association rule:

    buys(X, “computer”) => buys(X, “software”) [support = 1%, confidence = 50%]

- **Correlation Analysis:** additional analysis to find statistical correlations between associated pairs

Classification and Prediction



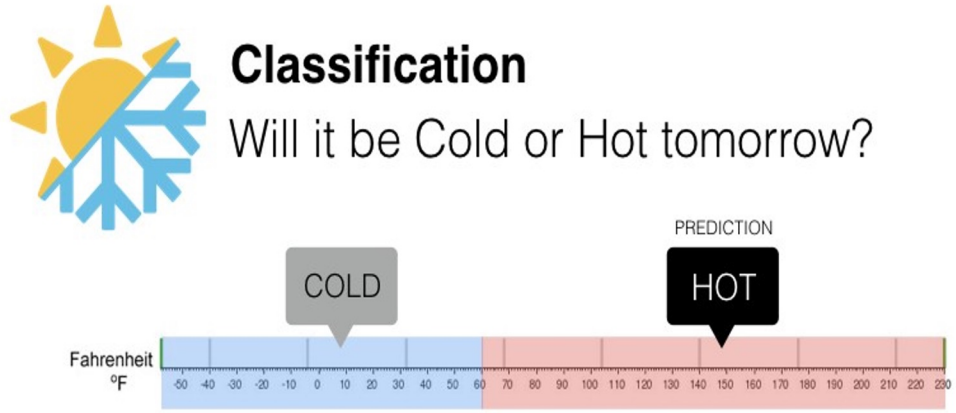
Classification

- The process of finding a model that describes and distinguishes the data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown.

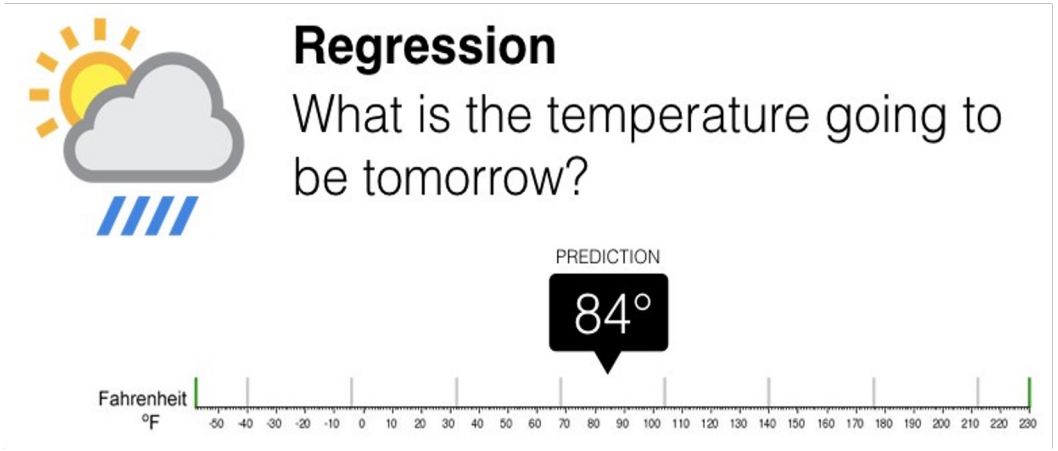
- The model can be represented in classification (IF-THEN) rules, decision trees, neural networks, etc.

Prediction

- Predict missing or unavailable numerical data values.

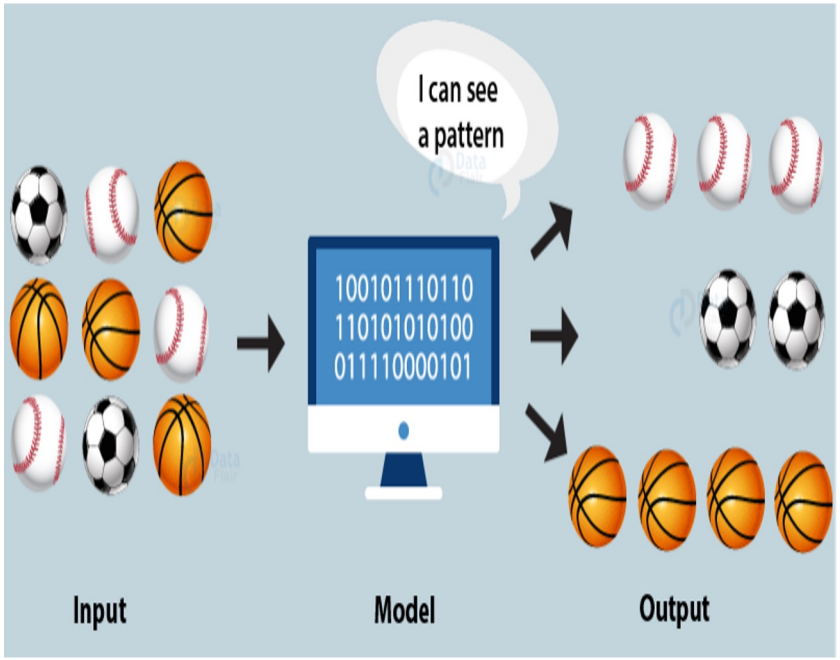


- **Classification**
  - Categorical (discrete, unordered) labels,



- **Regression**
  - Models continuous-valued functions

Cluster Analysis

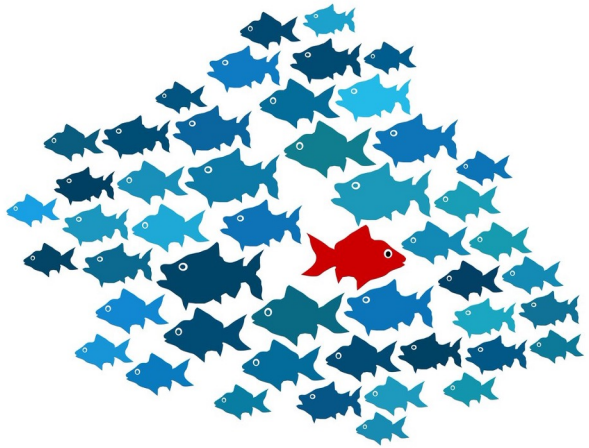


- Class label is unknown: group data to form new classes

- Clusters of objects are formed based on the principle of *maximizing intra-class similarity & minimizing interclass similarity*

  E.g. Identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing.

Outlier Analysis & Evolution Analysis



Outlier Analysis

- Data that do no comply with the general behavior or model.
- Outliers are usually discarded as noise or exceptions.
- Useful for fraud detection.

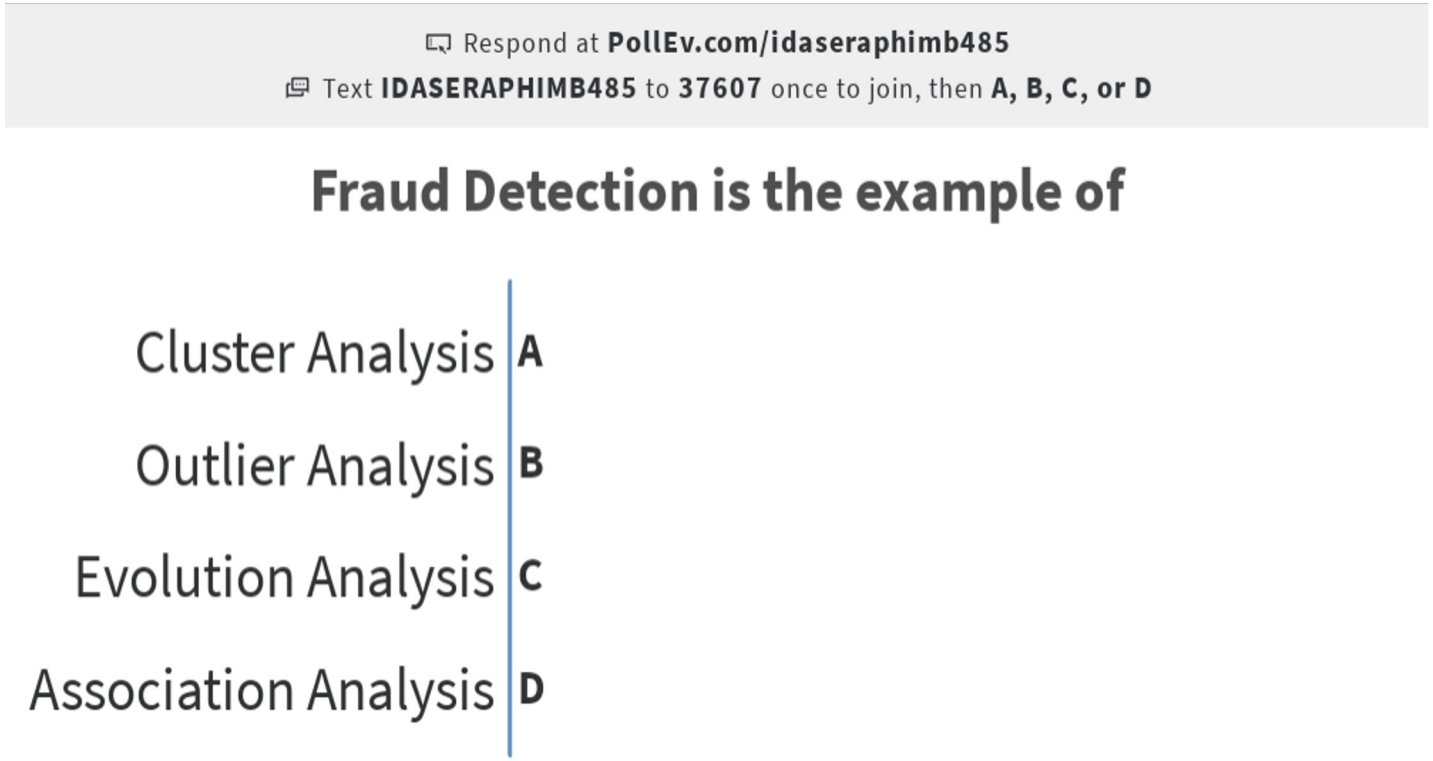
  E.g. Detect purchases of extremely large amounts

Evolution Analysis

- Describes and models regularities or trends for objects whose behavior changes over time.

  E.g. Identify stock evolution regularities for overall stocks and for the stocks of particular companies.

Try me…



https://PollEv.com/multiple_choice_polls/DGVYJIqVVARqUW006IKgU/respond

Applications Suitable for data Mining

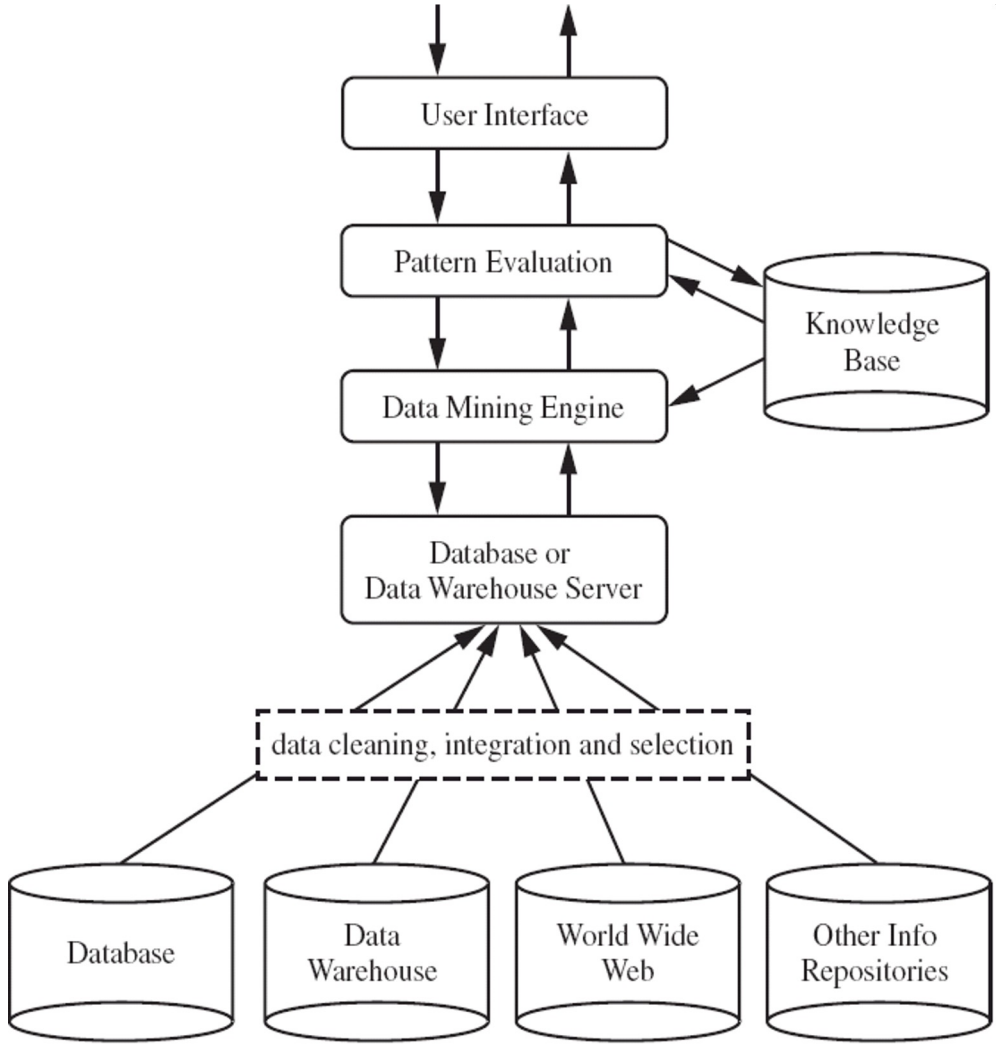
Application Area in Data Mining

- Highly application-driven discipline

Business Intelligence

- **Business intelligence** (**BI**) technologies provide historical, current, and predictive views of business operations.

A typical DM System Architecture



- Data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses or other important repositories.
- Database, data warehouse, WWW or other information repository (store data)
- Database or data warehouse server (fetch and combine data)
- Knowledge base (turn data into meaningful groups according to domain knowledge)
- Data mining engine (perform mining tasks)(like Characterization, association, correlation analysis, classification, prediction, cluster analysis, outlier analysis and evolution analysis.
- Pattern evaluation module (find interesting patterns)(integrated/interacts with Data Mining module)
- User interface (interact with the user)

Respond at **PollEv.com/idaseraphimb485**

Text **IDASERAPHIMB485** to **37607** once to join, then **A, B, C, or D**

Database or Data warehouse server can

A	B	C	D
Fetch and combine the data	Store data	Interact with the user	Mining the tasks

https://PollEv.com/multiple_choice_polls/ocULTRLQ76LFLJcLHP8Yo/respond

Issues in Data Mining

- Mining methodology and User interaction

  Mining different kinds of knowledge

  - DM should cover a wide spectrum of data analysis and knowledge discovery tasks
  - Enable to use the database in different ways
  - Require the development of numerous data mining techniques

  Interactive mining of knowledge at multiple levels of abstraction

  - Difficult to know exactly what will be discovered
  - Allow users to focus the search, refine data mining requests
  - Data cube can be used.
  - Multidimensional data mining can be used.

  Incorporation of background knowledge

  - Guide the discovery process
  - Allow discovered patterns to be expressed in concise terms and different levels of abstraction.

  Domain knowledge related to databases, such as integrity constraints and deduction rules speed up a data mining process.

Issues in Data Mining

Data mining query languages and ad hoc data mining

- High-level query languages need to be developed
- Should be integrated with a DB/DW query language

Presentation and visualization of results

- Knowledge should be easily understood and directly usable.
- High level languages, visual representations or other expressive forms.
- Require the DM system to adopt the above techniques.

Handling noisy or incomplete data

- Require data cleaning methods and data analysis methods that can handle noise.

Pattern evaluation – the interestingness problem

- How to develop techniques to access the interestingness of discovered patterns, especially with subjective measures bases on user beliefs or expectations.

Issues in Data Mining

Performance Issues

- Efficiency and scalability
  - Huge amount of data
  - Running time must be predictable and acceptable
- Parallel, distributed and incremental mining algorithms
  - Divide the data into partitions and processed in parallel
  - Incorporate database updates without having to mine the entire data again from scratch.

Issues relating to Diversity of Database Types

- Other database that contain complex data objects, multimedia data, spatial data, etc.
- Expect to have different DM systems for different kinds of data
- Heterogeneous databases and global information systems
  - Web mining becomes a very challenging and fast-evolving field in data mining

Issues in Data Mining

Diversity of Database Types

- Handling complex types of data (handle structured, semi structured, unstructured data)
- Mining dynamic, networked, and global data repositories

Data Mining and Society

- Social impacts of data mining
- Privacy-preserving data mining
- Invisible data mining

(E.g.: Online shopping)

Data Objects and Attribute Types

- A data object represents an entity.

  E.g.: 1. Sales database - customers, store items, sales

  2. Medical database - patients

  3.University database -students, professors, courses.

- Data sets are made up of data objects.

- Data objects are typically described by attributes. Data objects can also be referred to as samples, examples, instances, data points, or objects.

- If the data objects are stored in a database, they are data tuples.

Attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Data objects are typically described by attributes
- Represents a characteristic or feature of a data object
- Observed values for a given attribute are known as *observations*
- A set of attributes used to describe a given object is called an *attribute vector*

Types Of Attributes

- Nominal
- Binary
- Ordinal
- Numeric
- Discrete
- Continuous

Nominal Attributes

Types of data on the basis of measurement

Scale	True Zero	Equal Intervals	Order	Category	Example
Nominal	No	No	No	Yes	Marital Status, Sex, Gender, Ethnicity
Ordinal	No	No	Yes	Yes	Student Letter Grade, NFL Team Rankings
Interval	No	Yes	Yes	Yes	Temperature in Fahrenheit, SAT Scores, IQ, Year
Ratio	Yes	Yes	Yes	Yes	Age, Height, Weight

Marital status
0 – Single
1 – Married
2 – divorced
3 - widowed

- Nominal means “relating to names.”
- The values of a **nominal attribute** are symbols or *names of things*.
- Also referred to as **categorical**.
- The values do not have any meaningful order.
- values of a nominal attribute are symbols or “names of things,” it is possible to represent such symbols or “names” with numbers.

  E.g.: Marital status, Zip code, Hair colour.

Binary Attributes

- Two categories or states: 0 or 1

  0 – absent

  1 – present

- Binary attributes are referred to as Boolean if the two states correspond to true and false.

Symmetric

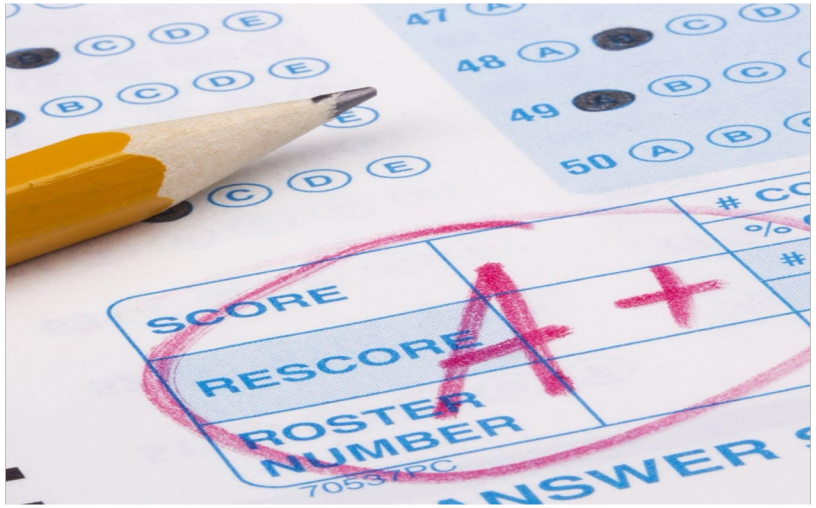
- Both states are equally valuable and carry same weight.

  E.g.: Gender – Male & Female

Asymmetric

- Both states are not equally important
  - rarest one, by 1 (e.g., *Covid19 positive*)
  - other by 0 (e.g., *Covid19 negative*).

Ordinal Attribute

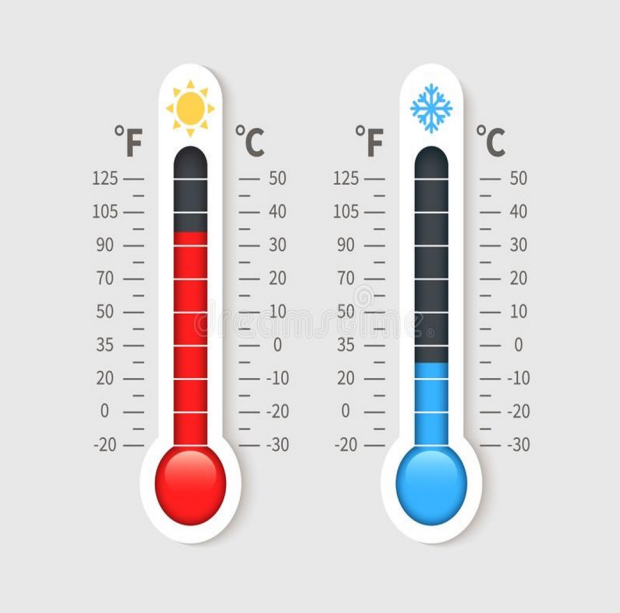


- An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is not known.
- Order is important

  E.g.: Grades {O, A+, A, B+, B, C}

  professional rank {professors, associate, assistant}
- Ordinal attributes are often used in surveys for ratings.

Interval-Scaled Variables



- Interval-scaled attributes are measured on a scale of equal-size units.
- The values of interval-scaled attributes have order and can be positive, 0, or negative.

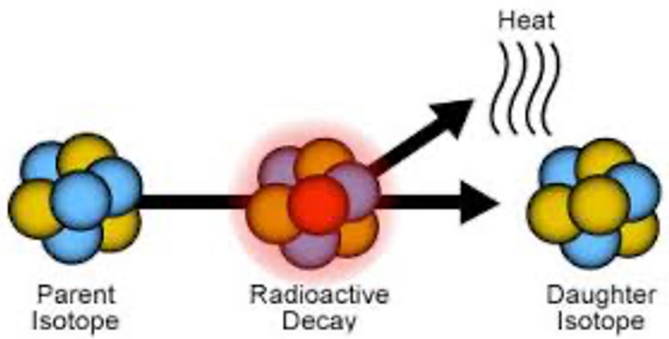
  E.g.: temperature, Calendar dates.

  Temperature of 20◦C is five degrees higher than a temperature of 15◦C.

  years 2002 and 2010 are eight years apart

- Interval-scaled attributes are numeric.

Ratio Scaled Attribute



- A ratio-scaled attribute is a numeric attribute with an inherent zero-point.
- Value as being a multiple (or ratio) of another value.
- Values are ordered.
- Compute the difference between values, as well as the mean, median, and mode.

  E.g.: Growth of a bacteria population

  Decay of a radioactive element

  years of experience

  Temperature in kelvin

  (true zero-point (0◦K = −273.15◦C))

Discrete versus Continuous Attributes

- A discrete attribute has a finite or countably infinite set of values, which may or may not be represented as integers.

E.g.: Attributes countably finite - hair color, smoker, medical test, drink size have a finite number of values

E.g.: Attribute countably infinite - customer ID, Zip codes

- If an attribute is not discrete, it is continuous.

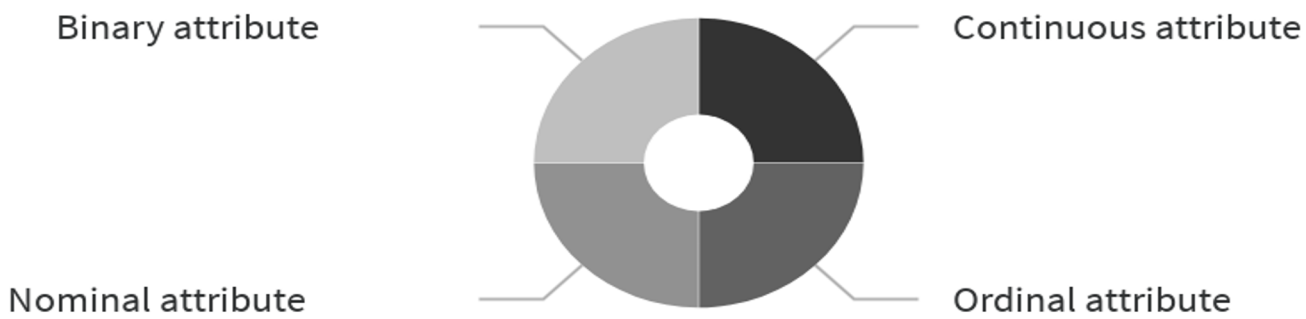
- Continuous attributes are typically represented as floating-point variables.

Respond at **PollEv.com/idaseraphimb485**

Text **IDASERAPHIMB485** to **37607** once to join, then **A, B, C, or D**

Data set {brown, black, blue, green , red} is example of

Continuous attribute **A** Ordinal attribute **B** Nominal attribute **C** Binary attribute **D**



https://PollEv.com/multiple_choice_polls/QXuTdjWOcvaWFu2IiqdYo/respond

Statistical Descriptions of Data

- Statistical descriptions can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.

Example

Suppose we have the following values for salary (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. What is the mean?

$$\bar{x} = \frac{30+36+47+50+52+52+56+60+63+70+70+110}{12}$$

$$= \frac{696}{12} = 58.$$

Mean = $58,000

Measuring the Central Tendency: Mean, Median, and Mode

- The most common and effective numeric measure of the "center" of a set of data is the (arithmetic) mean. Let x1,x2,...,xN be a set of N values or observations, such as for some numeric attribute X, like salary. The mean of this set of values is

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Weighted arithmetic mean

- Some cases each value of xi in a set may be associated with a weight wi for i = 1,...,N.

- The weights reflect the significance, importance, or occurrence frequency attached to their respective values.

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}.$$

- This is called the weighted arithmetic mean or the weighted average.

Drawbacks of Mean Calculation

- Sensitive to extreme values.
- Even a small number of extreme values can corrupt the mean

  E.g.: The mean salary at a company may be substantially pushed up by that of a few highly paid managers.

Solution

- **Trimmed mean**
- Mean obtained after chopping off values at the high and low extremes.
- Sort the values observed for *salary*
- Remove the top and bottom 2% before computing the mean

Median

Example
N is Even (N=12)
Suppose we have the following values for salary (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. What is the median?

52+56/2 = 108/2 = 54

Median = $54,000

N is Odd (N=11)
Suppose we have the following values for salary (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70. What is the median?

Median = $52,000

- Middle value in a set of ordered data values.
- It is the value that separates the higher half of a data set from the lower half.

If *N* is odd

- Then the median is the *middle value* of the ordered set.

If *N* is even

- Then the median is not unique
- It is the two middlemost values and any value in between.

- The median is expensive to compute when we have a large number of observations.

Mode

Example

Suppose we have the following values for salary (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. what is mode? What kind of mode?

The are two modes $52,000 and $70,000

Since there are two modes it is called bimodal

- The **mode -** value that occurs most frequently in the set.
- It is possible for the greatest frequency to correspond to several different values, which results in more than one mode.
- **One mode – Unimodal**
- **Two modes – Bimodal**
- **Three modes – Trimodal**
- Data set with two or more modes is multimodal.
- If each data value occurs only once, then there is no mode.

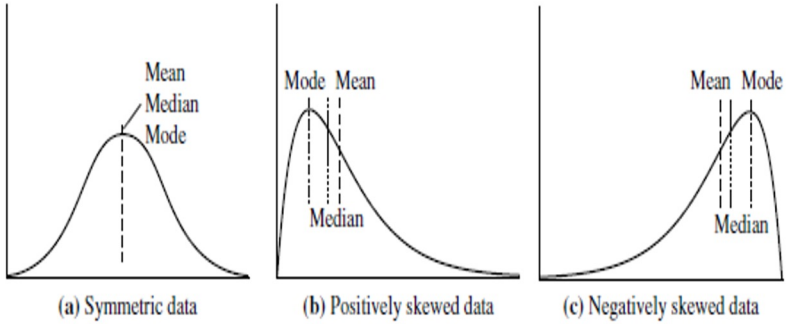
Midrange

Example

Suppose we have the following values for salary (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. What is the midrange of data?

30,000+110,000 / 2 = $70,000

Midrange of data = $70,000



- The midrange can also be used to assess the central tendency of a numeric data set.
- It is the average of the largest and smallest values in the set.

- In a unimodal frequency curve with perfect symmetric data distribution, the mean, median, and mode are all at the same center value

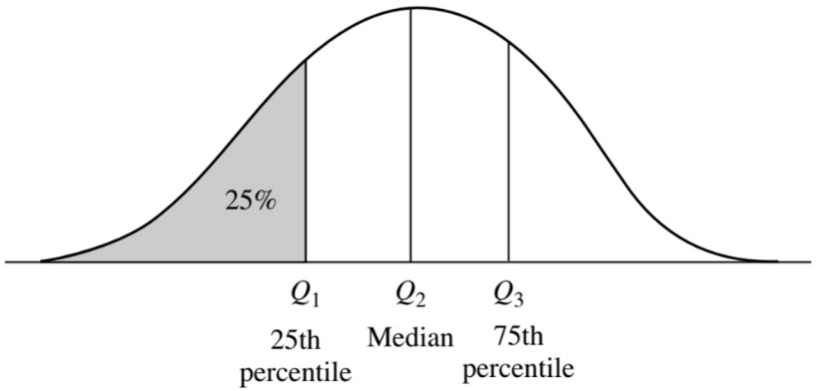
Positively Skewed Data

- The mode occurs at a value that is smaller than the median

Negatively Skewed Data

- The mode occurs at a value greater than the median

Range, Quartiles, Outliers and Boxplots



- The range of the set is the difference between the largest (max()) and smallest (min()) values.
- The most commonly used percentiles other than the median are quartiles.
- The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data.
- This distance is called the interquartile range (IQR) and is defined as

$$IQR = Q3 - Q1$$

Five-Number Summary, Boxplots, and Outliers

Find the lower and upper quartiles for the set.

34, 14, 24, 16, 12, 18, 20, 24, 16, 26, 13, 27

Solution

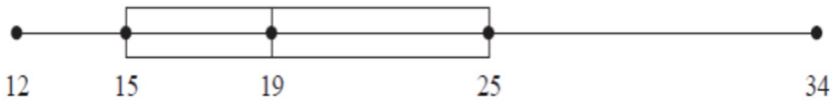
Begin by ordering the set.

12, 13, 14, 16, 16, 18, 20, 24, 24, 26, 27, 34

1st 25% 2nd 25% 3rd 25% 4th 25%

The median of the entire set is 19. The median of the six numbers that are less than 19 is 15. So, the lower quartile is 15. The median of the six numbers that are greater than 19 is 25. So, the upper quartile is 25.

Quartiles are represented graphically by a **box-and-whisker plot,** as shown in Figure A.6. In the plot, notice that five numbers are listed: the smallest number, the lower quartile, the median, the upper quartile, and the largest number. Also notice that the numbers are spaced proportionally, as though they were on a real number line.



Five-Number summary

- The five-number summary of a distribution consists of the median (Q2), the quartiles Q1 and Q3, and the smallest and largest individual observations, written in the order of Minimum, Q1, Median, Q3, Maximum.

Boxplots

- Boxplots are a popular way of visualizing a distribution. A boxplot incorporates the five-number summary as follows:
  - Typically, the ends of the box are at the quartiles so that the box length is the interquartile range.
  - The median is marked by a line within the box.
  - Two lines (called whiskers) outside the box extend to the smallest (Minimum) and largest (Maximum) observations.

Sketch a box-and-whisker plot for each set.

a. 27, 28, 30, 42, 45, 50, 50, 61, 62, 64, 66

b. 82, 82, 83, 85, 87, 89, 90, 94, 95, 95, 96, 98, 99

c. 11, 13, 13, 15, 17, 18, 20, 24, 24, 27

Solution

a. This set has 11 numbers. The median is 50 (the sixth number). The lower quartile is 30 (the median of the first five numbers). The upper quartile is 62 (the median of the last five numbers). See Figure A.7.

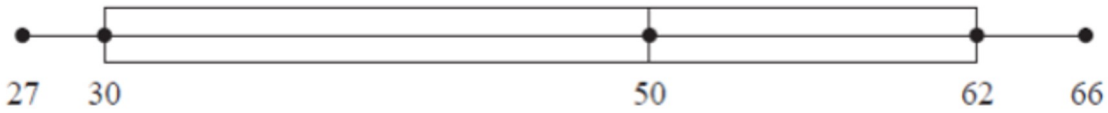


FIGURE A.7

b. This set has 13 numbers. The median is 90 (the seventh number). The lower quartile is 84 (the median of the first six numbers). The upper quartile is 95.5 (the median of the last six numbers). See Figure A.8.

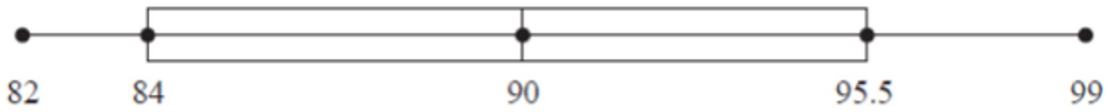
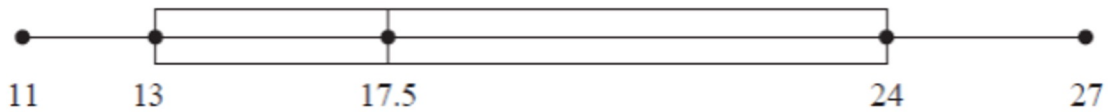


FIGURE A.8

c. This set has 10 numbers. The median is 17.5 (the average of the fifth and sixth numbers). The lower quartile is 13 (the median of the first five numbers). The upper quartile is 24 (the median of the last five numbers). See Figure A.9.



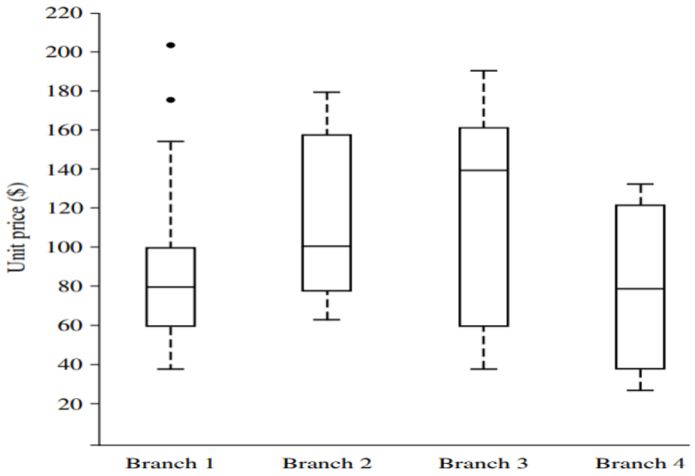
Outliers

A common rule of thumb for identifying suspected outliers is to single out values falling at least 1.5 × IQR above the third quartile or below the first quartile.

Steps to find Outliers

- $IQR = Q_3 - Q_1$
- Test for Outliers:
  1. Find IQR
  2. Multiply 1.5(IQR)
  3. Subtract $Q_1 - 1.5(IQR)$
  4. Add $Q_3 + 1.5(IQR)$
  5. Any value less than the value in step 3 or more than the value in step 4 is an **outlier**.

E.g.: Outlier Observations – 175, 202



E.g.:5, 6,12,13,15,18,22,50

Try…

Respond at **PollEv.com/idaseraphimb485**
Text **IDASERAPHIMB485** to **37607** once to join, then **A, B, C, or D**

Given the values 5, 6, 12, 13, 15, 18, 22, 50. Find Q1, Q2(Median), Q3 and IQR value.

9, 14, 15, 11	**A**
5, 14, 20, 5	**B**
9, 14, 20, 11	**C**
10, 15, 20, 11	**D**

https://PollEv.com/multiple_choice_polls/m4RnVyRZQszR3wM95eHQG/respond

Variance and Standard Deviation

- Variance and standard deviation are measures of data dispersion.
- A low standard deviation - The data observations tend to be very close to the mean
- A high standard deviation - The data are spread out over a large range of values.

The **variance** of N observations, $x_1, x_2, \ldots, x_N$, for a numeric attribute X is

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

- where $\bar{x}$ is the mean value of the observations.
- The standard deviation, σ, of the observations is the square root of the variance, σ^2.

The standard deviation of a set is a measure of how much a typical number in the set differs from the mean. The greater the standard deviation, the more the numbers in the set *vary* from the mean. For instance, each of the following sets has a mean of 5.

{5, 5, 5, 5}, {4, 4, 6, 6}, and {3, 3, 7, 7}

The standard deviations of the sets are 0, 1, and 2.

$$\sigma_1 = \sqrt{\frac{(5-5)^2 + (5-5)^2 + (5-5)^2 + (5-5)^2}{4}}$$

$$= 0$$

$$\sigma_2 = \sqrt{\frac{(4-5)^2 + (4-5)^2 + (6-5)^2 + (6-5)^2}{4}}$$

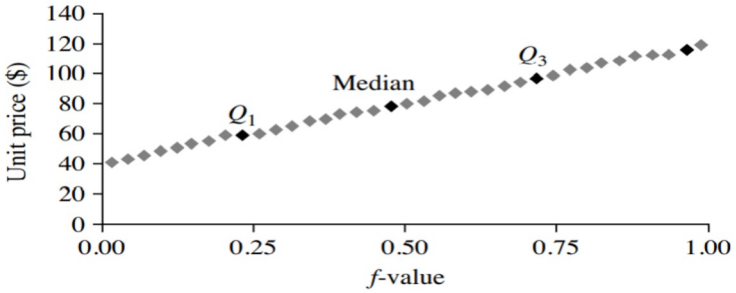
$$= 1$$

$$\sigma_3 = \sqrt{\frac{(3-5)^2 + (3-5)^2 + (7-5)^2 + (7-5)^2}{4}}$$

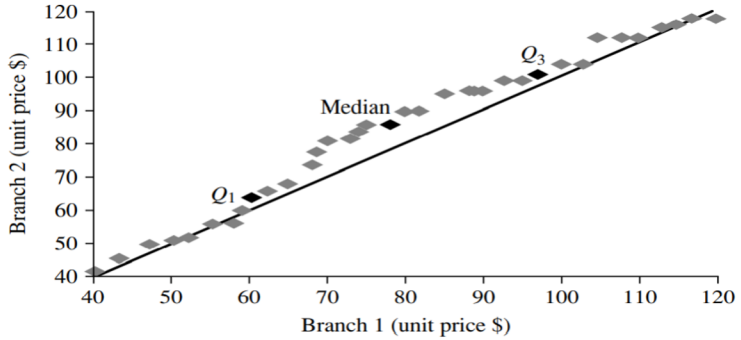
$$= 2$$

Graphic Displays

Quantile Plot



Quantile–Quantile Plot



- Graphic displays include quantile plots, quantile–quantile plots, histograms, and scatter plots.

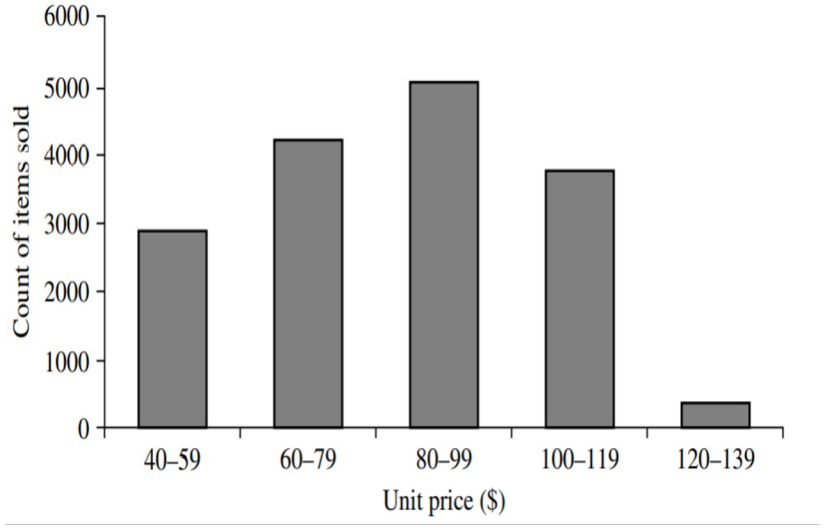
Quantile Plot

- A quantile plot is a simple and effective way to look at a univariate data distribution. First, it displays all of the data for the given attribute. Second, It plots quantile information.

Quantile–Quantile Plot

- A quantile–quantile plot, or q-q plot, graphs the quantiles of one univariate distribution against the corresponding quantiles of another

Graphic Displays

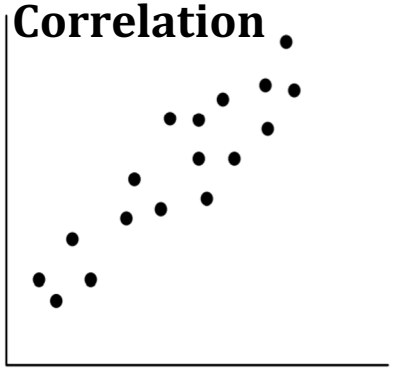


Histograms

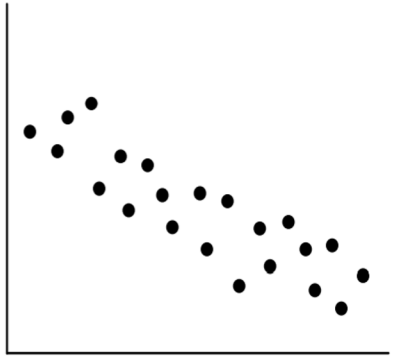
- "Histos" means pole, and "gram" means chart, so a histogram is a chart of poles.
- If X is **nominal**, such as automobile model or item type, then a pole or vertical bar is drawn for each known value of X.
- The resulting graph is more commonly known as a **bar chart**.
- If X is **numeric**, the term histogram is preferred.
- The range of values for X is partitioned into disjoint consecutive subranges. The subranges, referred to as buckets or bins. The range of a bucket is known as the width.

Graphic Displays

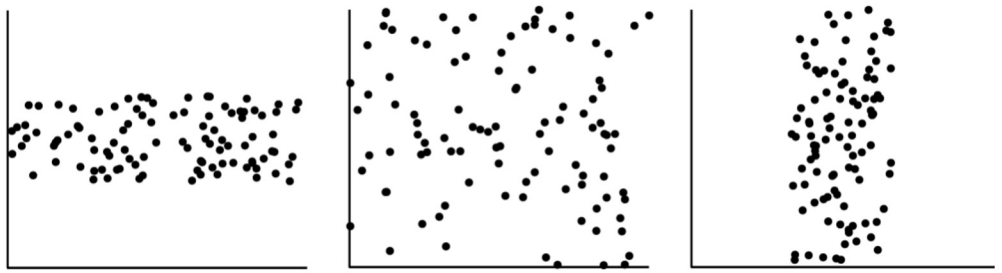
Positive Correlation



Negative Correlation



No Correlation



- A scatter plot is one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numeric attributes.
- To construct a scatter plot, each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane.
- Two attributes, X, and Y, are correlated if one attribute implies the other. Correlations can be positive, negative, or null.

Need for data preprocessing and data quality

- Data have quality if they satisfy the requirements of the intended use.

Factors comprising Data Quality

Accuracy
Completeness
Consistency
Timeliness
Believability
Interpretability

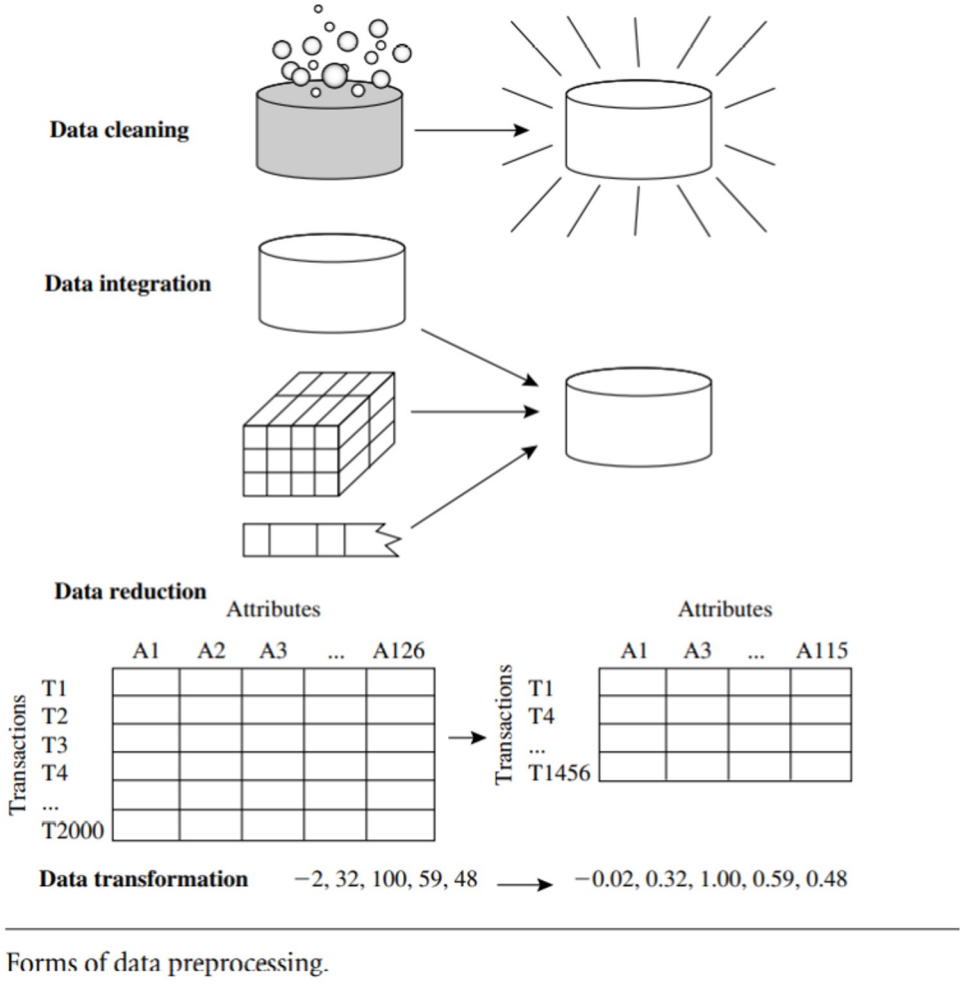
Bad Data Quality

- Incomplete
- Inaccurate or noisy
- Inconsistent

Reasons of Bad Quality Data

The data collection instruments used may be faulty.
Disguised missing data.
Technology limitations.
Inconsistencies in naming conventions

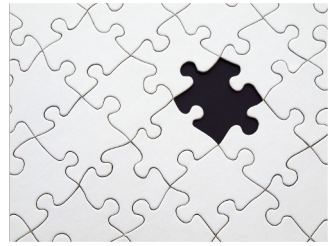
Data Cleaning



Forms of data preprocessing.

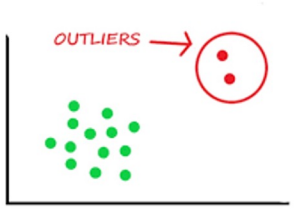
Real-world data

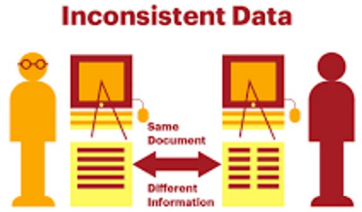
- Incomplete
- Noisy
- Inconsistent



Data cleaning (or data cleansing)

- Fill in missing values
- Smooth out noise while identifying outliers
- Correct inconsistencies in the data.





Handle Missing Values

1. Ignore the tuple

usually done when the class label is missing.

This method is not very effective, unless the tuple contains several attributes with missing values.

2. Fill in the missing value manually

Time consuming

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 (mean or median) to fill in missing value.

5. Use the attribute mean or median for all samples belonging to the same class as the given tuple.

6. Use the most probable value to fill in the missing value

Noisy Data

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

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Binning methods for data smoothing.

- Noise is a random error or variance in a measured variable.

Binning

- Binning methods smooth a sorted data value by consulting its "neighborhood," that is, the values around it.
- The sorted values are distributed into a number of "buckets," or bins.

Smoothing by bin means

- Each value in a bin is replaced by the mean value of the bin.

Smoothing by bin medians

- Each bin value is replaced by the bin median.

Smoothing by bin boundaries

- The minimum and maximum values in a given bin are identified as the bin boundaries.
- Each bin value is then replaced by the closest boundary value.

Binning – Equal depth (frequency)

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

Equal-width (distance) partitioning

Divides the range into N intervals of equal size: uniform grid

if A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B –A)/N.

The most straightforward, but outliers may dominate presentation

Skewed data is not handled well

Equal-depth (frequency) partitioning

Divides the range into N intervals, each containing approximately same number of data points

Good data scaling

Managing categorical attributes can be tricky

Binning – Equal width (distance)

- Sorted data is given in the order - 5,10,11,13,15,35,50,55,72,92,204,215

$$\text{width} = \frac{\max - \min}{\text{number of bins}} : \quad \frac{215 - 5}{3} = 70.$$

1. $70 + 5 = 75$ (from 5 to 75) =Bin 1 : 5, 10, 11, 13, 15, 35, 50, 55, 72
2. $75 + 70 = 145$ (from 75 to 145) =Bin 2 : 92
3. $145 + 70 = 215$ (from 145 to 215) =Bin 3 : 204, 215

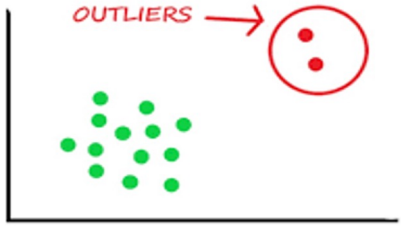
Noisy Data

Regression

- Linear regression - 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.

Outlier analysis

- Outliers may be detected by clustering, for example, where similar values are organized into groups, or “clusters.”



Data Cleaning as process

- Missing values, noise, and inconsistencies contribute to inaccurate data.
- The first step in data cleaning as a process is discrepancy detection.

Factors causes Discrepancies

Poorly designed data entry forms

Human error in data entry, deliberate errors

Data decay

Rules

- Unique Rule - Each value of the given attribute must be different from all other values for that attribute.
- Consecutive Rule - There can be no missing values between the lowest and highest values for the attribute, and that all values must also be unique (e.g., check numbers).
- Null Rule - Use of blanks, question marks, special characters, or other strings that may indicate the null condition.

Data Cleaning as process

Field overloading

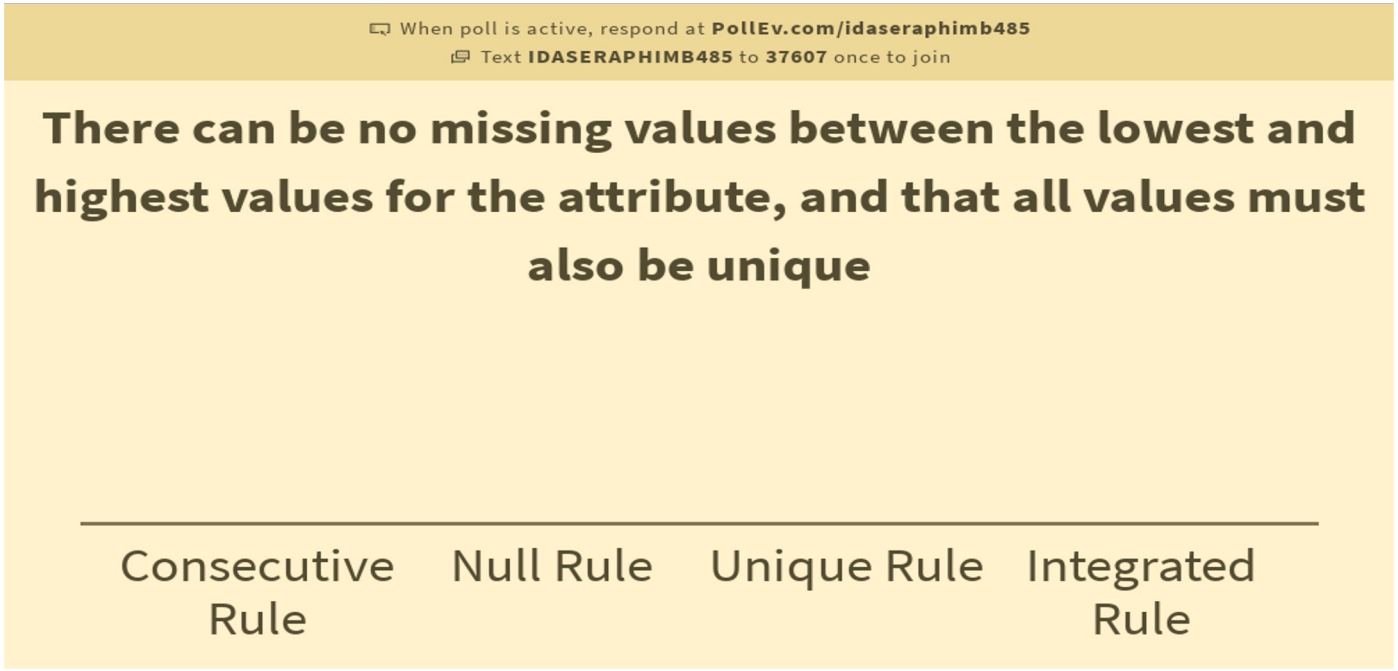
- Another source of errors that typically results when developers squeeze new attribute definitions into unused (bit) portions of already defined attributes.

Discrepancy detection tools

- **Data scrubbing tools -** use simple domain knowledge to detect errors and make corrections in the data.

- **Data auditing tools -** find discrepancies by analyzing the data to discover rules and relationships, and detecting data that violate such conditions.

Give a try…



https://PollEv.com/multiple_choice_polls/TsqJUtyUrMB7007vm8TX5/respond

Data Integration

Data integration

- Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id = B.cust-#
- Integrate metadata from different sources

Entity identification problem

- Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton

Detecting and resolving data value conflicts

- For the same real world entity, attribute values from different sources are different
- Possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases.

  Object identification: The same attribute or object may have different names in different databases.

  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

- For numerical attributes, we can evaluate the correlation between two attributes, A and B, by computing the correlation coefficient.

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B},$$

- If rA,B > 0, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.

- rA,B = 0: uncorrelated;

- rA,B < 0: negatively correlated.

Data Integration

Chi-Square Test

- For categorical (discrete) data, a correlation relationship between two attributes, A and B, can be discovered by a (chi-square) test.

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

- where oij is the observed frequency (i.e., actual count) of the joint event (Ai;Bj) and eij is the expected frequency of (Ai;Bj), which can be computed as

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

Chi-Square Calculation Example

	Male	Female	Total (row)
Like science fiction	250	200	450
Not like science fiction	50	1000	1050
Total (col.)	300	1200	1500

	male	female	Total (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Total (col.)	300	1200	1500

$$e_{11} = \frac{count(male) \times count(fiction)}{N} = \frac{300 \times 450}{1500} = 90,$$

e12 = 1200*450/1500 = 360
e21 = 300*1050/1500 = 210
e22 = 1200*1050/1500 = 840

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

Data Transformation

- **Smoothing:** which works to remove noise from the data. Such techniques include binning, regression, and clustering.

- **Aggregation:** where summary or aggregation operations are applied to the data.

  Eg: the daily sales data may be aggregated so as to compute monthly and annual total amounts.

- **Generalization** of the data, where low-level or "primitive" (raw) data are replaced by higher-level concepts through the use of concept hierarchies.

  Eg: categorical attributes, like street, can be generalized to higher-level concepts, like city or country.

- **Numerical attributes**, like age, may be mapped to higher-level concepts, like youth, middle-aged, and senior.

Data Transformation

- **Normalization:** where the attribute data are scaled so as to fall within a small specified range, such as - 1.0 to 1.0, or 0.0 to 1.0.

- Attribute construction (or feature construction),where new attributes are constructed and added from the given set of attributes to help the mining process.

There are many methods for data normalization.

Min-max normalization

z-score normalization

Normalization by decimal scaling.

Min-max normalization

- Min-max normalization preserves the relationships among the original data values.

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

Example

- Suppose that the minimum and maximum values for the attribute income are \$12,000 and \$98,000, respectively. We would like to map income to the range [0.0;1.0]. By min-max normalization, a value of \$73,600 for income is transformed.

73600 – 12000/98000 – 12000 (1.0 – 0.0) +0 = 0.716

z-score normalization (or zero-mean normalization)

- The values for an attribute, A, are normalized based on the mean and standard deviation of A.

$$v' = \frac{v - \bar{A}}{\sigma_A},$$

- Where $\bar{A}$ and σA are the mean and standard deviation, respectively, of attribute A.

Example

- Suppose that the mean and standard deviation of the values for the attribute income are \$54,000 and \$16,000, respectively. With z-score normalization, a value of \$73,600 for income is transformed

73600 – 54000 / 16000 = 1.225

Normalization by decimal scaling

- 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.

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

where j is the smallest integer such that $Max(|v'|) < 1$.

- Suppose that the recorded 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 1,000 (i.e., j = 3) so that

  - 986 normalizes to – 0.986 and 917 normalizes to 0.917.

Give a try…

Respond at **PollEv.com/idaseraphimb485**

Text **IDASERAPHIMB485** to **37607** once to join, then **A, B, C, or D**

Suppose that the recorded values of Salary Bonus ranges from 310 to 400. Find the maximum absolute value of salary Bonus. Then normalize those values.

Maximum absolute value = 310, … **A**

Maximum absolute value = 400, … **B**

Maximum absolute value = 400, … **C**

Maximum absolute value = 310, … **D**

https://PollEv.com/multiple_choice_polls/8ObNy0zo242lvkilnxsSj/respond

Data Reduction

- Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume.
- closely maintains the integrity of the original data.

Strategies for data reduction

Data cube aggregation: where aggregation operations are applied to the data in the construction of a data cube.

Attribute subset selection: where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed.

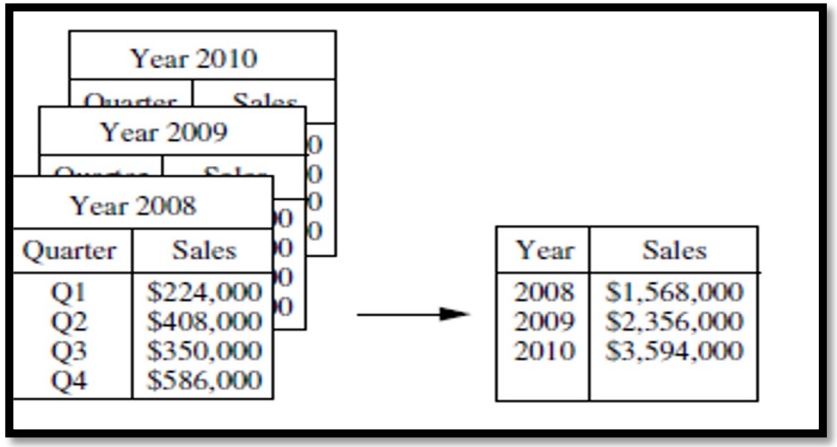
Dimensionality reduction: where encoding mechanisms are used to reduce the data set size.

Numerosity reduction: where the data are replaced or estimated by alternative.

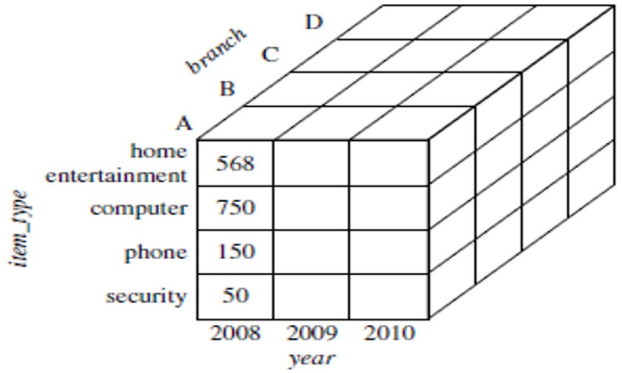
Discretization and concept hierarchy generation: where raw data values for attributes are replaced by ranges or higher conceptual levels.

Data cube and its usage

- Data cubes store multidimensional aggregated information.
- Each cell holds an aggregate data value, corresponding to the data point in multidimensional space.
- Data cubes provide fast access to precomputed summarized data.
- Base Cuboid Vs Apex Cuboid.



Data Cube Aggregation



Data Cube