3.36pt

VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Data Mining and Business Intelligence
ITA5007

Prof. Ramesh Ragala

January 23, 2019

- The Explosive Growth of Data: from petabytes to brontobytes
  - 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,etc
    - Science → Remote sensing, bioinformatics, scientific simulation, etc
    - Society and everyone → news, digital cameras, social networking sites, smart mobile phones, etc
- **We are drowning in data, but starving for knowledge!**
- "Necessity is the mother of invention" → Data mining →Automated analysis of massive data sets.

Why Data Mining

- Credit ratings/targeted marketing:
  - Given a database of 100,000 names, which persons are the least likely to default on their credit cards?
  - Identify likely responders to sales promotions
- Fraud detection
  - Which types of transactions are likely to be fraudulent, given the demographics and transactional history of a particular customer?
- Customer relationship management:
  - Which of my customers are likely to be the most loyal, and which are most likely to leave for a competitor?
- **Data Mining helps extract such information**

- **Process of semi-automatically analyzing large databases to find patterns that are:**
  - **Valid:** hold on new data with some certainity
  - **novel:** non-obvious to the system
  - **useful:** should be possible to act on the item
  - **understandable:** humans should be able to interpret the pattern
- **Also known as Knowledge Discovery in Databases (KDD)**

- There are many definitions are available for Data Mining

Definition - 1

Extracting useful information from large datasets Hand et al., 2001

Definition - 2

Data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules Berry and Linoff: 1997 and 2000

Definition - 3

Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques. Gartner Group

Where Is Data Mining Used

- Data mining is used in a variety of fields and applications.
  - military $\rightarrow$ to learn what roles various factors play in the accuracy of bombs.
  - Intelligence agencies $\rightarrow$ to determine which of a huge quantity of intercepted communications are of interest.
  - Security specialists $\rightarrow$ to determine whether a packet of network data constitutes a threat.
  - Medical Researchers $\rightarrow$ to predict the likelihood of a diseases relapse.
- Data Mining in Business Domain:
  - From a large list of prospective customers, which are most likely to respond? $\rightarrow$ Classification Techniques
  - Using, prediction techniques to forecast how much individual prospects will spend.,
  - Which customers are most likely to commit, fraud?. $\rightarrow$ using, classification methods to identify medical reimbursement applications that have a higher probability of involving fraud, and give them greater attention.
  - Which loan applicants are likely to default?
  - Which customers are more likely to abandon a subscription service?

Origins of Data Mining

- Data mining stands at the confluence of the fields of statistics and machine learning (also known as artificial intelligence)
- Some techniques for exploring data and building models in statistics
  - Linear Regression
  - Logistic Regression
  - Discriminate Analysis
  - Principal Component Analysis
- But the core tenets of classical statistics
  - Computing is difficult and data are scare
  - In data mining applications data and computing power are plentiful.
  - Data mining is "**statistics at scale and speed, and simplicity**"
  - Simplicity in this case refers to **simplicity in the logic of inference**

Origins of Data Mining

- Due to the scarcity of data in the classical statistical setting, the same sample is used to make an estimate, and also to determine how reliable that estimate might be.
- The logic of the confidence intervals and hypothesis tests used for inference
  - Elusive for many
  - Limitations are not well appreciated
- The data mining paradigm is fitting a model with one sample and assessing its performance with another sample is easily understood.
- Computer Science:
  - "machine learning" techniques, such as trees and neural networks.
  - Rely on computational intensity and are less structured than classical statistical models
  - Field of database management is also part of the picture.
- The emphasis that classical statistics places on inference is missing in data mining.

Origins of Data Mining

- Data mining deals with large datasets in open-ended fashion $\rightarrow$ which making it impossible to put the strict limits around the question being addressed that inference would require.
- As a result, the general approach to data mining is vulnerable to the danger of "**overfitting**",
  - Where a model is fit so closely to the available sample of data that it describes not merely structural characteristics of the data, random peculiarities as well.
  - In engineering terms, the model is fitting the noise, not just the signal.

Difference between Statistics and Data Mining

STATISTICS	DATA MINING
Confirmative	Explorative
Small data sets or File-based	Large data sets or Databases
Small number of variables	Large number of variables
Deductive	Inductive
Numeric data	Numeric and non-numeric
Clean data	Data cleaning

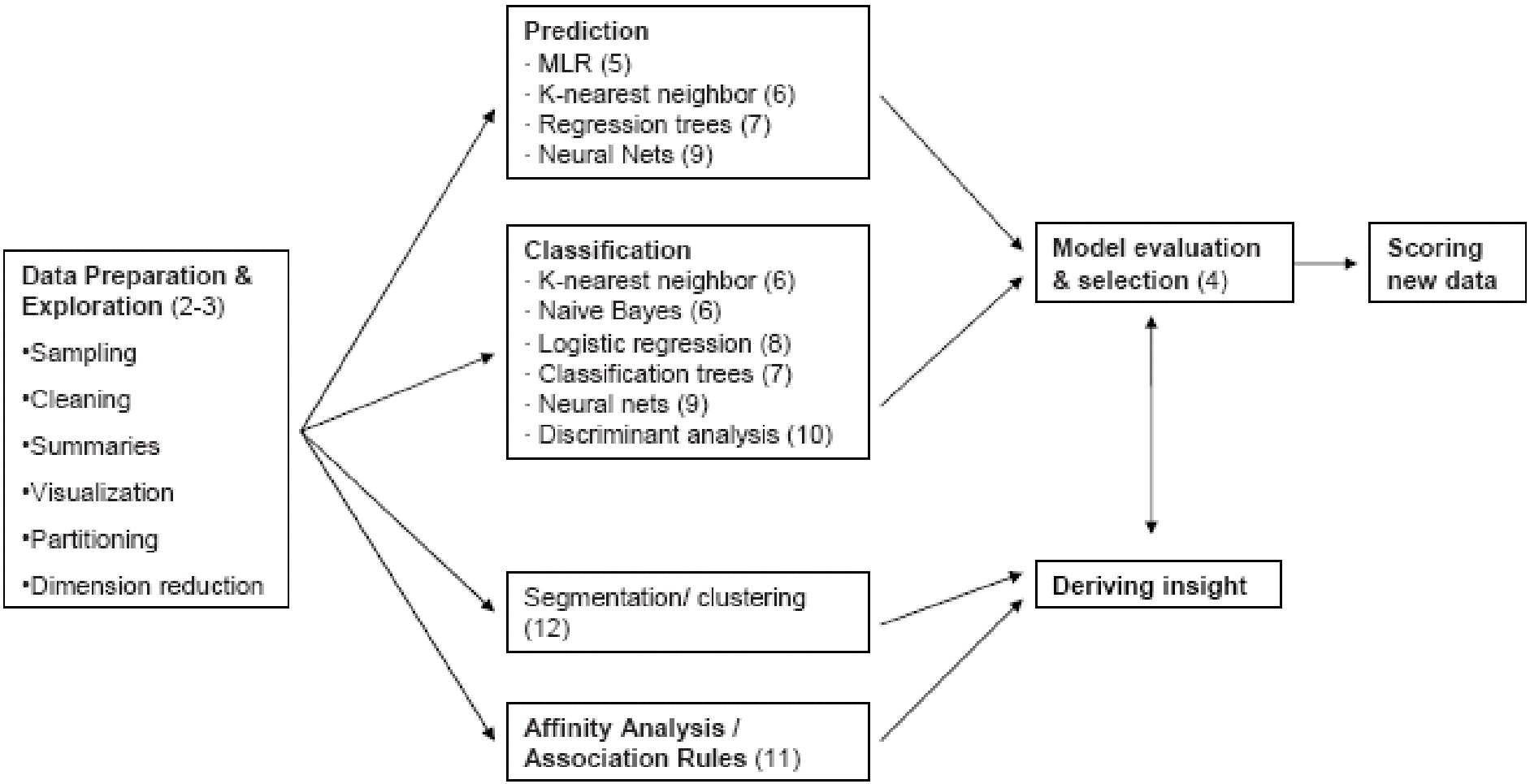
Growth of Data Mining

- Decreasing cost and increasing availability of automatic data capture mechanisms.
- A shift in focus from products and services to a focus on the customer and his or her needs has created a demand for detailed data on customers
- Data from operational databases are extracted, transformed and exported to a data warehouse
- Smaller data marts devoted to a single subject may also be part of the system.
- Data from external sources (e.g. Credit rating data)
- The rapid and continuing improvement in computing capacity is an essential enabler of the growth of data mining

Reason behind many Methods

- Many different methods for prediction and classification
- Each method has its advantages and disadvantages
- Usefulness of a method depends on
  - Size of the dataset
  - The types of patterns that exist in the data
  - Whether the data meet some underlying assumptions of the method
  - How noisy the data are
  - The particular goal of the analysis, etc
- Different methods can lead to different results, and their performance can vary.
- Customary in data mining to apply several different methods and select the one that is most useful for the goal at hand

Taxonomy of Data Mining Methods



Overview of Data Mining Process

The general steps are involved in data mining, starting from a clear goal definition and ending with a model deployment

SCHEMATIC OF THE DATA MODELING PROCESS

Analytical Methods Used in Predictive Analytics

- Classification
  - Most basic form of Data analysis
  - Data can be classified in many ways.
  - Examples:
    - Recipient of an email can respond or ignore
    - Patient suffering from an illness can recover, still be ill or die.
    - A set of student marks can be classified as average, good, satisfactory or excellent
  - Challenges
    - Examine data who's classification is unknown $\rightarrow$ what that classification is?
    - Data will occur in near future $\rightarrow$ what will be?
    - **When similar data is available, we develop rules to help us classify incoming data based on similarity**
  - Used with categorical response variables
- Prediction
  - Predict the numerical value
  - Predict (estimate) value of continuous response variable
  - Prediction used with categorical as well

- Association Rules and Recommended Systems
  - Large databases of customer transactions provide us with information of how the transactions have associations among themselves.
  - Affinity analysis $\rightarrow$ "**what goes with what**"
  - Seeks correlations among data
  - Examples:
    - E.g. Grocery stores and placement of items in a grocery store.
    - Amazon tracks customer preferences using "collaborative filtering"
- Predictive Analytics
  - Classification, prediction, association rules and collaborative filtering put together $\rightarrow$ predictive analytics
  - Data Pattern Identification methods such as clustering.

- Data Reduction and Dimension Reduction
  - Data in large volumes is difficult to handle
  - Impacts performance of data mining algorithms $\rightarrow$ number of variable in the problem
  - Form Homogeneous group
    - Makemytrip groups hotels based on popularity and budget
    - Amazon groups items based on the frequency with which the customer browses his items.
  - Consolidating large number of records into smaller ones is called, Data Reduction
  - Methods for reducing the number of cases is called clustering
  - Reduce variables $\rightarrow$ Group together similar variables $\rightarrow$ Dimension Reduction

- Data Exploration
  - View data as evidence
  - Get "a feel" for the data
  - Exploring Dataset
    - Understanding the global landscape of data
    - Detecting anomalies or unusual values
    - Cleaning unwanted data
  - Data Exploration is used for data cleaning and manipulation as well as for visual discovery and **hypothesis generation**
  - How is exploration done?
    - Look at variables
    - Look at the relationship between variables
    - Find a pattern
- Data Visualization
  - Data Exploration by creating charts and dashboards is called Data visualization or Visual Analytics
  - Locate tends, correlations, etc.
  - Numerical data $\rightarrow$ histograms, boxplots, scatter plots
  - Categorical data $\rightarrow$ bar charts

Supervised Learning

- "Supervised learning" algorithms are those used in classification and prediction.
  - Data (i.e class label) is available in which the value of the outcome of interest is known.
- "Training data" are the data from which the classification or prediction algorithm "learns" or is "trained" about the relationship between predictor variables and the outcome variable. $\rightarrow$ Training Phase
- This process results in a "model" $\rightarrow$ classification model, etc
- Model is then run with another sample of data
  - "validation data" $\rightarrow$ Testing Phase
  - even the outcome is known but we wish to see how well the model performs
  - many different models are being tried out with known outcomes - "test data" $\rightarrow$ finally select a model to predict how well it will do (more accuracy).
- The model can then be used to classify or predict the outcome of interest in new cases where the outcome is unknown.

Unsupervised Learning

- The model is not provided with the correct results during the training.
  $\rightarrow$ No outcome variable to predict or classify
- No "learning" from cases
- Can be used to cluster the input data in classes on the basis of their statistical properties only.
- Unsupervised leaning methods
  - Association Rules
  - Data Reduction Methods
  - Clustering Techniques

- **Training set:** a set of examples used for learning, where the target value is known.
  - The training partition, typically the largest partition, contains the data used to build the various models for examining.
  - The same training partition is generally used to develop multiple models.
- **Validation set:** a set of examples used to tune the architecture of a classifier and estimate the error.
  - The validation partition (sometimes called the test partition) is used to assess the predictive performance of each model so that you can compare models and choose the best one.
  - In some algorithms, the validation partition may be used in an automated fashion to tune and improve the model.
- **Test set:** used only to assess the performances of a classifier.
  - The test partition (sometimes called the holdout or evaluation partition) is used to assess the performance of the chosen model with new data.
  - It is never used during the training process so that the error on the test set provides an unbiased estimate of the generalization error.

Steps in Data Mining

- Poor Understanding of the problems → Errors in Analytics projects
- List of Steps to be taken in a typical Data Mining:
- **Develop an understanding of the purpose of the problem or project**
  - How will the stakeholder use the result?
  - Who will be affected by the results?
  - It is a one-shot effort to answer a question or questions or Application (if it is an ongoing procedure)
- **Obtain the dataset to be used in the analysis**
  - Random sampling from a large database to capture records to be used in an analysis
  - Pulling together data from different databases. (Internal → Past Purchase made by customers and External → Credit Rating)
  - You don't need all the data to perform analysis

Steps in Data Mining

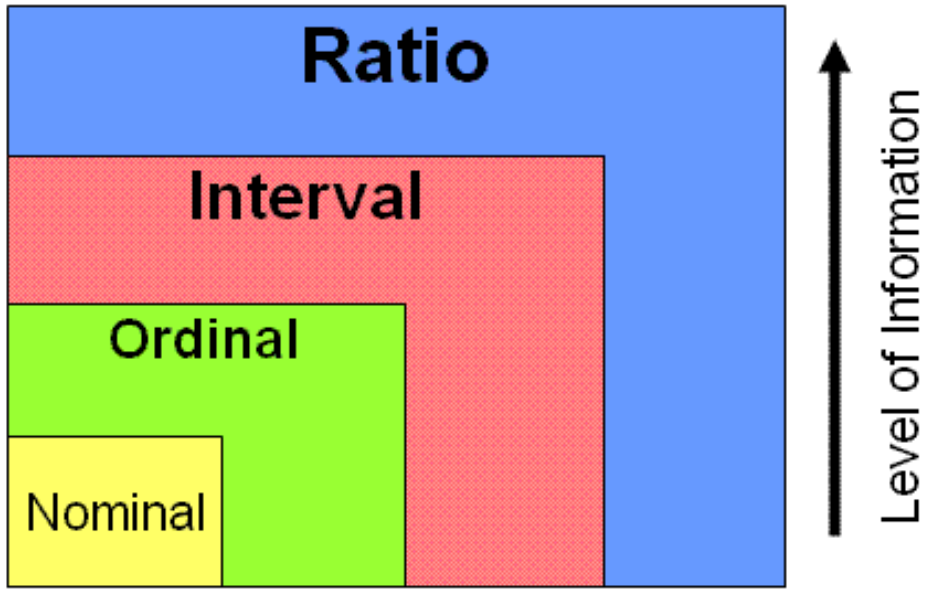
- **Explore, clean, and preprocess the data**
  - Verifying that the data are in reasonable condition.
  - How missing data should be handled?
  - Are the values in a reasonable range, given what you would expect for each variable?
  - Are there obvious "outliers?"
  - Data are reviewed graphically $\rightarrow$ to know the relation among variables
  - Ensure consistency in the definitions of fields, units of measurement, time periods, etc.
- **Reduce Dimension of the Data (if necessary)**
  - Involves the operations like eliminating unneeded variables, transforming variables and creating new variables, etc
  - Make sure you know what each variable means, and whether it is sensible to include it in the model.
- **Determine the data mining task**
  - Classification, prediction, clustering etc..
  - This involves of transforming the general question into specific data mining question.

Steps in Data Mining

- **Partition Data (for supervised learning)**
  - If supervised learning is used, randomly partition the data set into three broad sets $\rightarrow$ Training Set, Validation Set and Test Set
- **Choose Data Mining Techniques to be used**
  - Regression, Neural Networks, Heirarchical Clustering etc..
- **Use algorithms to perform the task**
  - Iterative process - trying multiple variants, and often using multiple variants of the same algorithm
  - When appropriate, feedback from the algorithm's performance on validation data is used to refine the settings.
- **Interpret the results output by the algorithm**
  - Choose the best algorithm to deploy.
  - Use final choice on the test data to get an idea how well it will perform.
- **Deploy the model**
  - Integrate the model into operational systems
  - Run it on real records to produce decisions or actions.
  - For example, the model might be applied to a purchased list of possible customers, and the action might be "include in the mailing if the predicted amount of purchase is $>$ Rs.100."

Different Types of Data or Attribute or Fields in Datasets

- Quantitative Data
  - Continuous Data
  - Discrete Data
    - Interval Data
    - Ratio Data
- Qualitative Data
  - Ordinal Data
  - Nominal Data



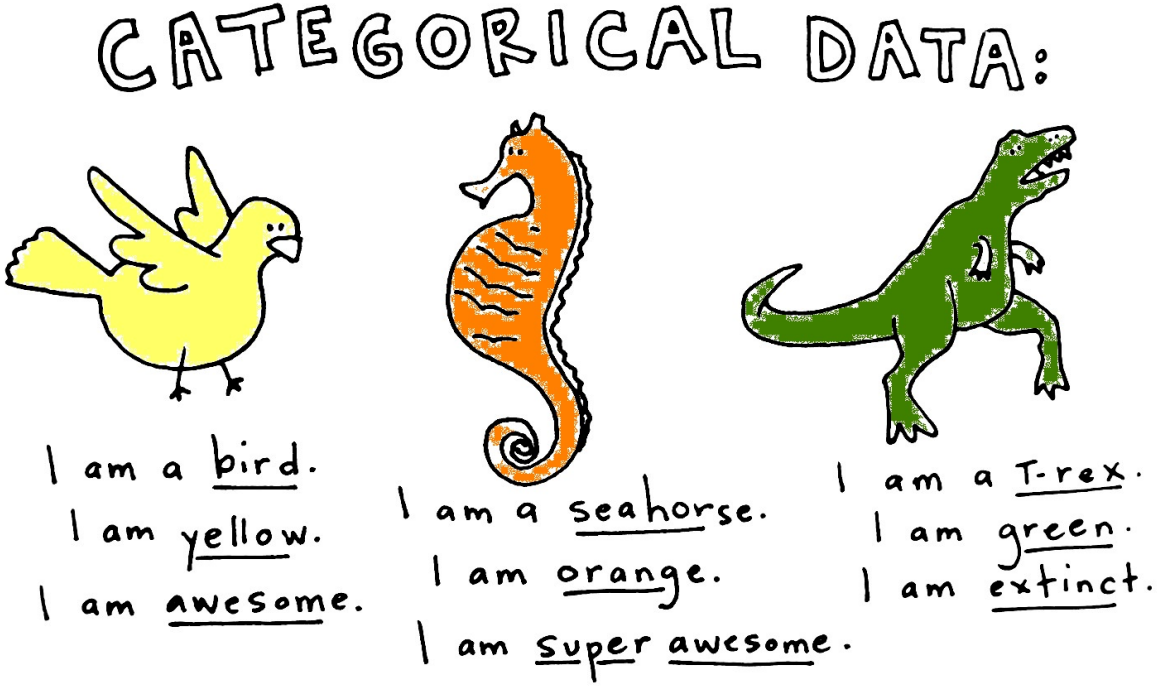
Nonparametric (qualitative data)

Parametric (quantitative data)

***Nonparametric statistics may be used to analyze interval and ratio data measurements.**

- Data variables are classified $\rightarrow$ 4 types $\rightarrow$ based on on the scale by which the values it contains are measured:
  - **Nominal/categorical data:**
    - The data values are categorical and not numeric.
    - A categorical variable is one that has two or more categories or labels or classes, but there is no intrinsic ordering to the categories.
    - simply Categorical variables represent types of data which may be divided into groups.
    - It is completely qualitative measurement.
    - Examples: age, gender, educational levels, countries, people names. operations: == and !=
    - Comparing two observations using the values for the variable, the observations will either be similar or different depending on whether the categorical value matches or not.

- **Example on Categorical Data:**



- if the categorical data has only two outcomes → binary or binomial data
- The Binomial data outcomes may pass/fail, live/dead or extinct/not extinct

- **Examples on Categorical Variables**

	A	B	C	D	E	F	G	H	I
1	Name	Miles Per Gallon	Accceleration	Horsepower	weight	cylinders	year	price	Country
2	Volkswagen Rabbit Dl	43,1	21,5	48	1985	4	78	2400	Germany
3	Ford Fiesta	36,1	14,4	66	1800	4	78	1900	Germany
4	Mazda GLC Deluxe	32,8	19,4	52	1985	4	78	2200	Japan
5	Datsun B210 GX	39,4	18,6	70	2070	4	78	2725	Japan
6	Honda Civic CVCC	36,1	16,4	60	1800	4	78	2250	Japan
7	Oldsmobile Cutlass	19,9	15,5	110	3365	8	78	3300	USA
8	Dodge Diplomat	19,4	13,2	140	3735	8	78	3125	USA
9	Mercury Monarch	20,2	12,8	139	3570	8	78	2850	USA

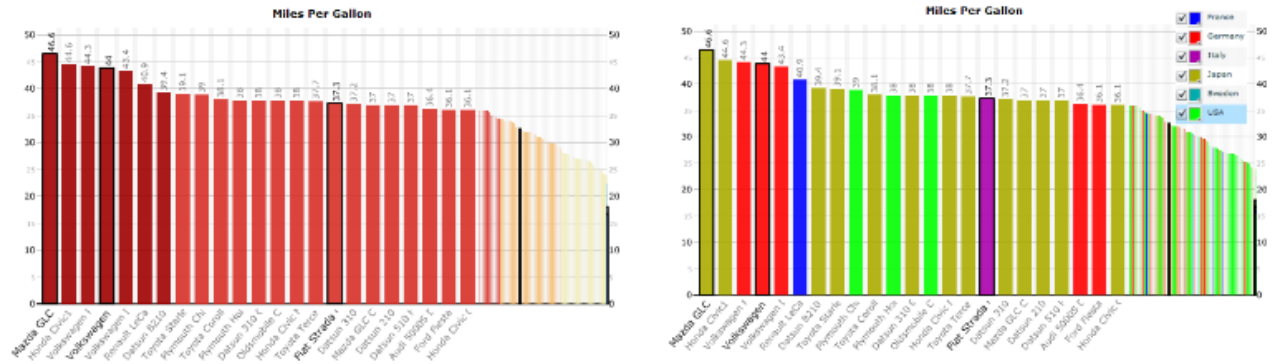


Figure: Classic car data set shown as bar chart for numerical variable "Miles per gallon" and coloured based on categorical variable Country.

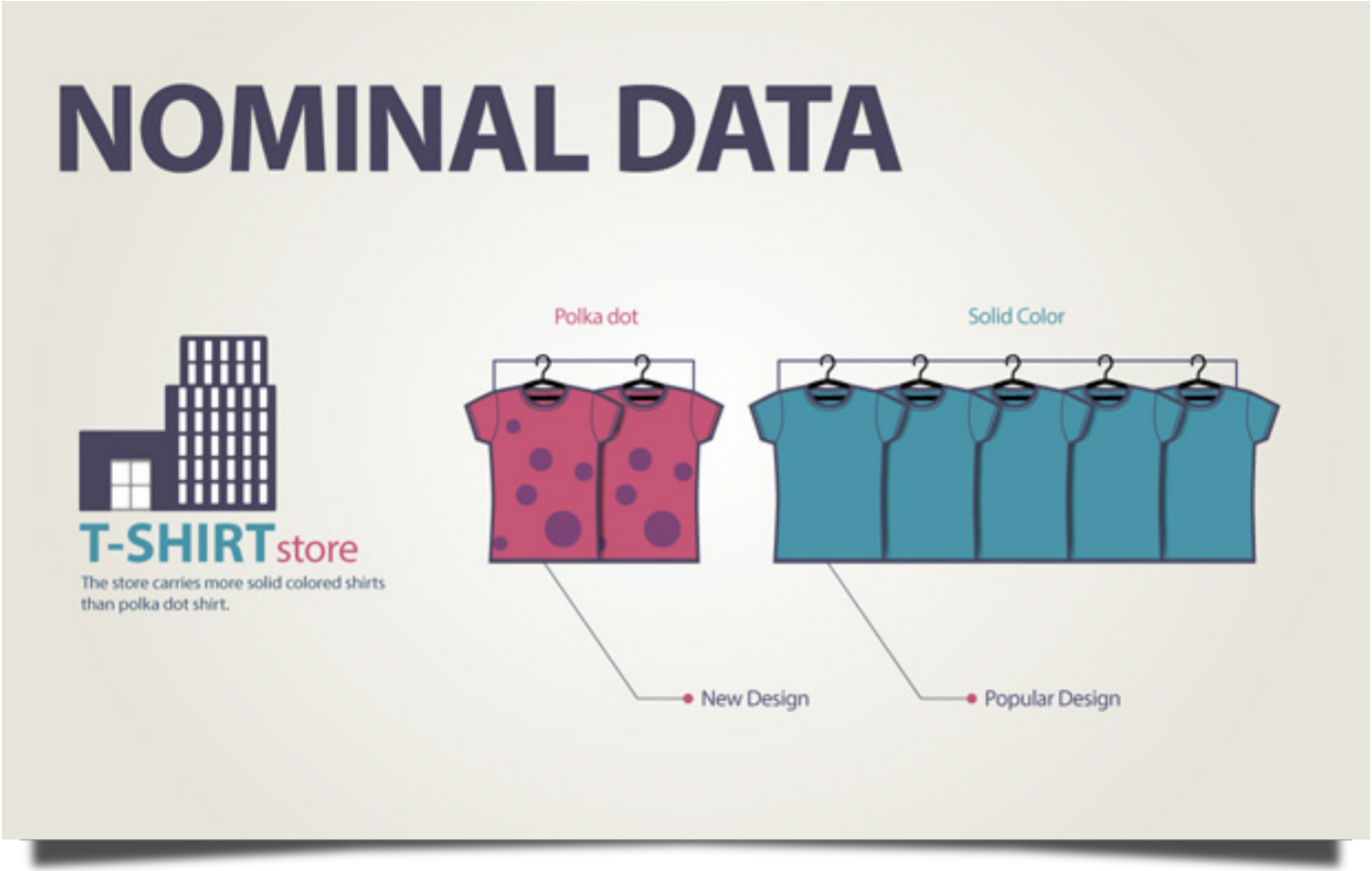


Figure: nominal level of measurement

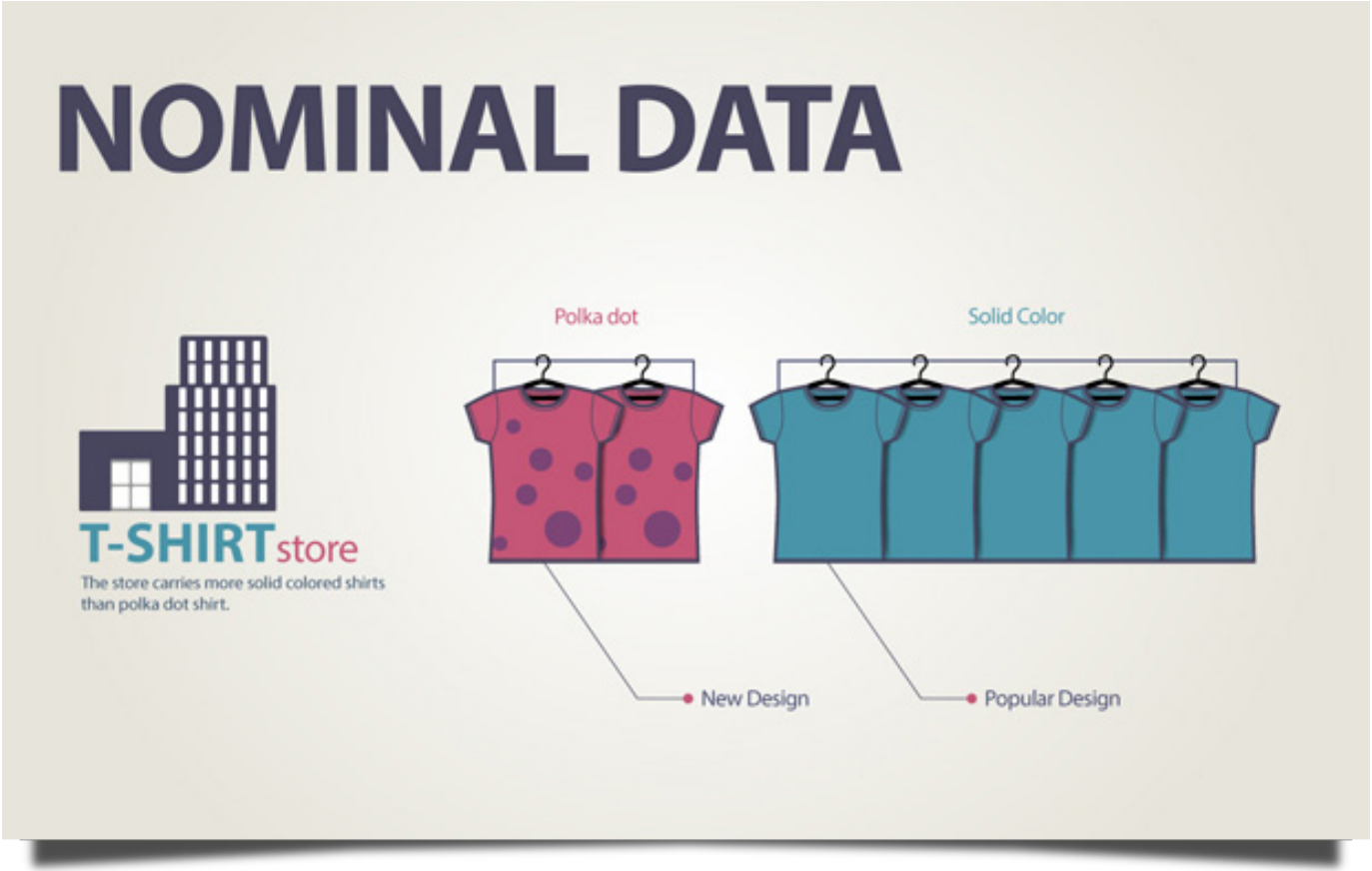


Figure: nominal level of measurement

- A categorical variable (sometimes called a nominal variable) is one that has two or more categories, but there is no intrinsic ordering to the categories.
- A purely categorical variable is one that simply allows you to assign categories but you cannot clearly order the variables.
- If the variable has a clear ordering, then that variable would be an ordinal variable.
- The Nominal or categorical data has only meaning $\rightarrow$ how they are differing from one another.
- **Example:** Country names are Nominal data values $\rightarrow$ putting all country names in alphabetical order is not making any relationship to another.
- Assignment of numbers to categories has no mathematical meaning.
- Nominal categories should be mutually exclusive and exhaustive

- **Where Can We Have Categorical Data:**
  - Social sciences : opinions on issues
  - Health sciences : response to treatments/drugs
  - Behavioral sciences : e.g. diagnose mental illness
  - Public health : AIDS awareness
  - Zoology : animals food preferences
  - Education : student's response to exams
  - Marketing : consumer preferences
  - Almost everywhere
- Distinction in categorical data are: Nominal Data and Ordinal Data

- **Ordinal data values:**
  - The data values are categorical but ordered.
  - Comparing two observations using the values for that variable.
  - Operations: $==, !=, \leq$ and $\geq$
  - it is mainly used for obey ordering relations among data values
  - Ordinal data is that which has inherent order, but no inherent degree of difference between what is being ordered.
  - **Example:** The Ist, IInd and IIIrd place winners in a race are on ordinal scale
  - But we do not know **how much faster** first place was than second place
  - But we know only that one was faster than other.

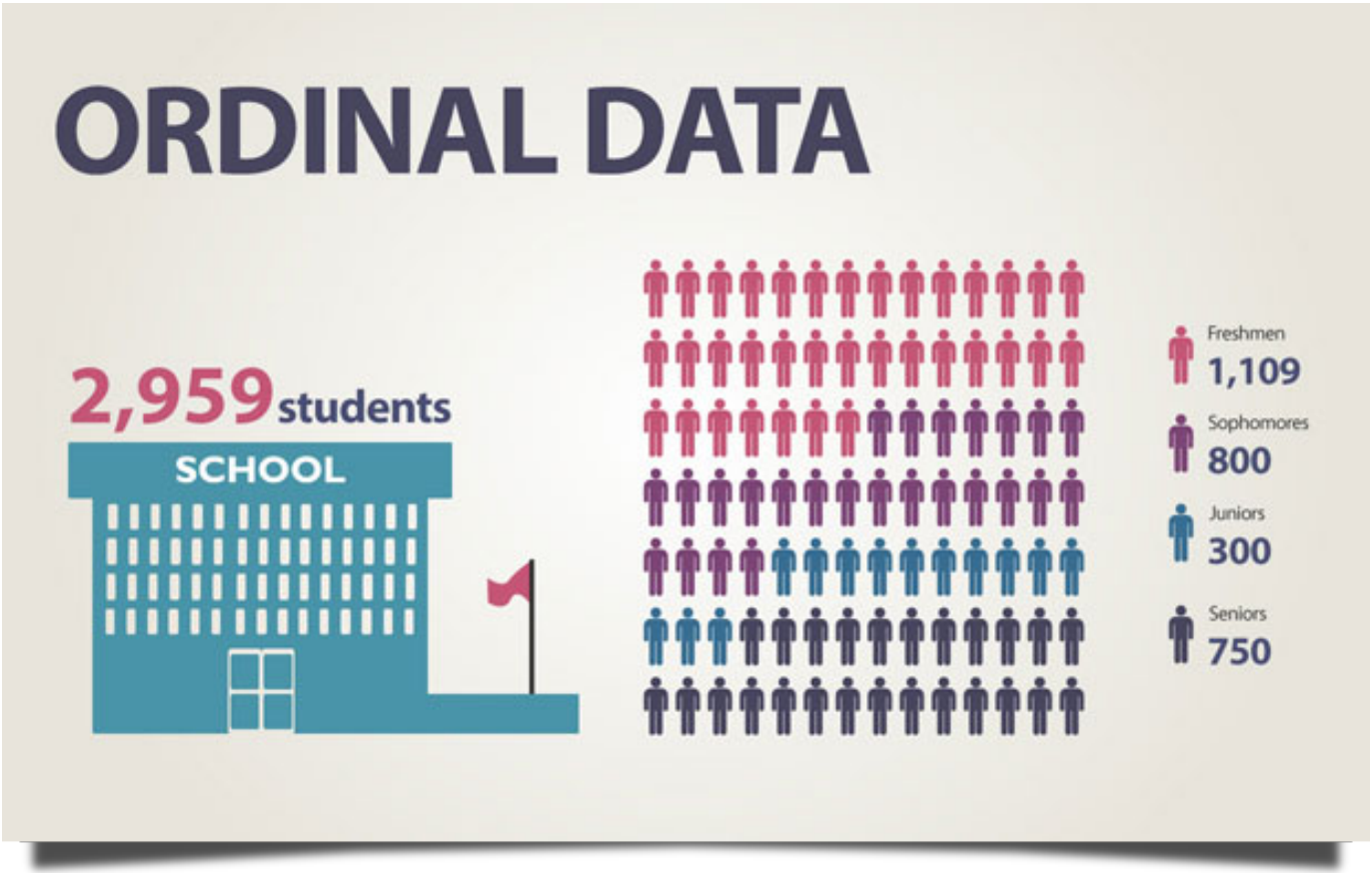
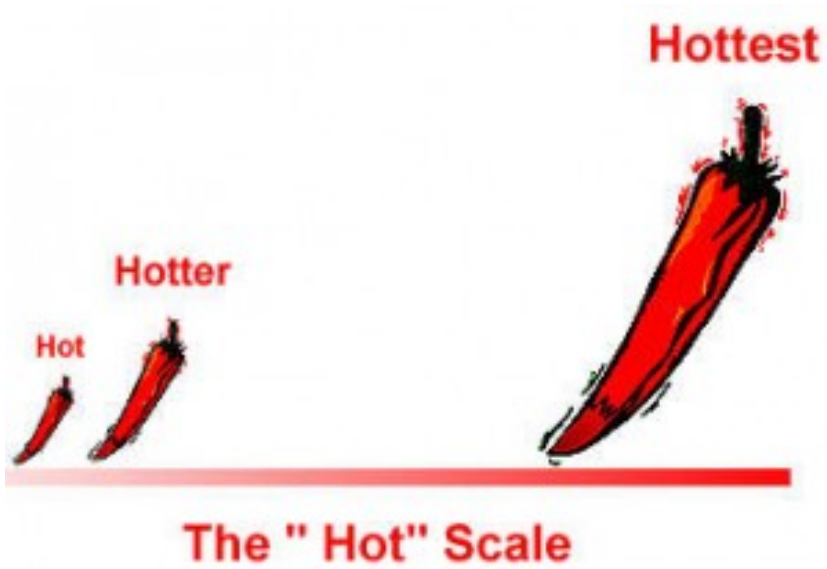


Figure: Ordinal level of measurement

Know your Data



- **Interval Data:**
  - The data values are numeric.
  - It represents the more sensitive type of data or sophisticated form of measurement.
  - simply, Interval data is data which exists on a scale with meaningful quantitative magnitudes between values.
  -
  - Data values can be compared quantitatively using basic arithmetic operations +,-,* and / not the values themselves.
  - The values are ordered. it includes negative numbers and zero. But zero is not absolute reference point.
  - Scale data is usually aggregated or converted to averages.

- **Interval Data:**
  - **Example:1** The dataset does not contain an interval data variable, if there were a variable in a dataset that recorded the measurements of temperature. $\rightarrow$ it would be classified as a interval variable.
  - Temperature variable contains the values 40,60 and 80, we could say that compared with 40^{o}F, 80^{o}F is two times warmer than 60^{o}F (80-40)/(60-40), but not twice as hot because 0^{o}F is an arbitrarily chosen point on the scale.
  - **Example:2** if Sidda Reddy is rated as "6" on attractiveness and Durga Prasad a "3" $\rightarrow$ it does not mean Sidda Reddy is twice as attractive as Durga Prasad.

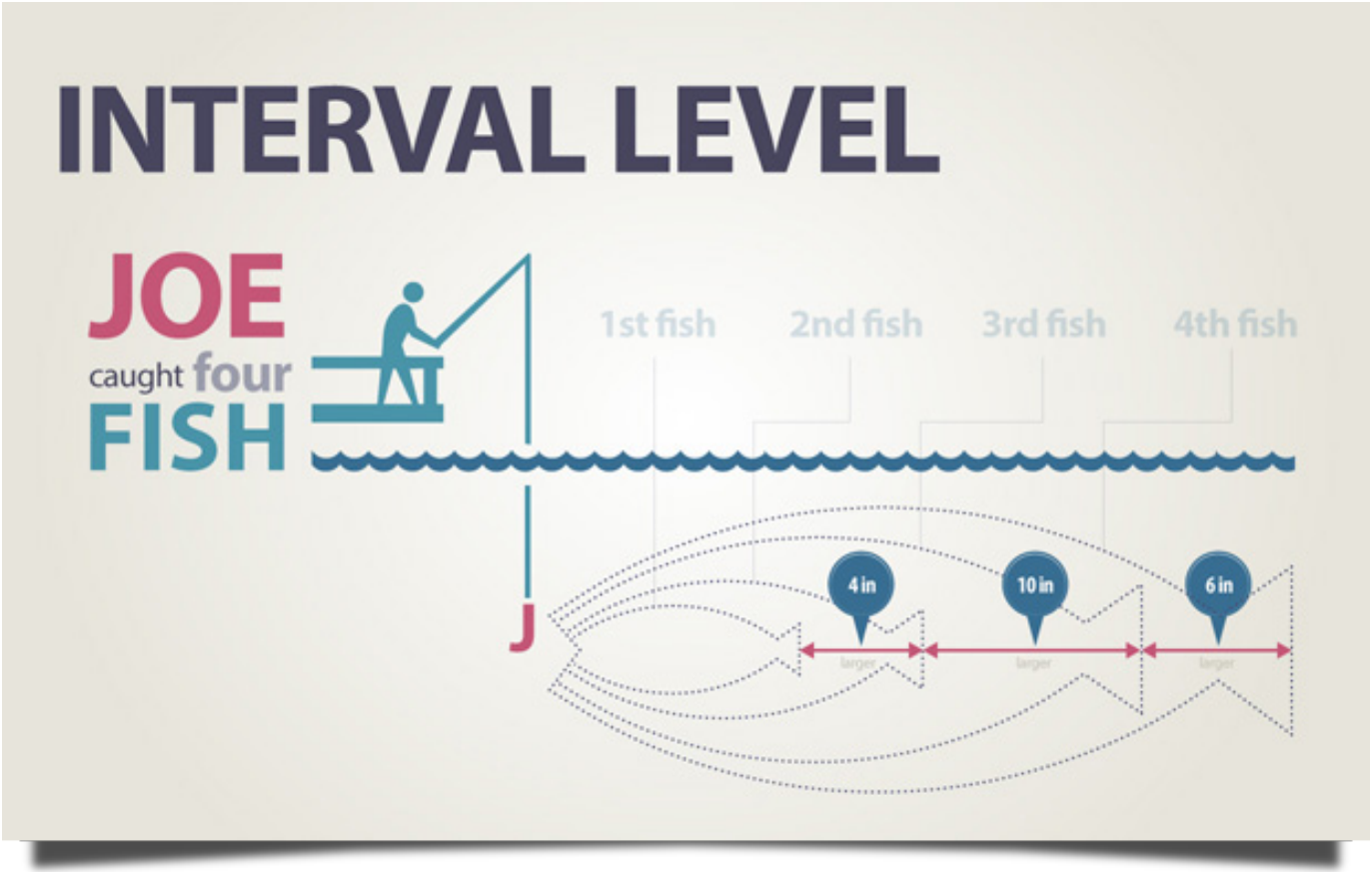


Figure: interval level of measurement

- The measurement between the sizes of the fish Joe caught in order of when he caught them.

- **Ratio Data:**
  - The Data Values are numeric and include an absolute zero.
  - This data values are allowed to compare quantitatively with other using basic arithmetic operations
  - Ratio data is data which, like interval data, has a meaningful order and a constant scale between ordered values, but additionally it has a meaningful zero value.
  - Supported Operations are ==, !=, $\leq$, $\geq$, -, / and *
  - The Ratio level of measurement applies to data that can be arranged in order.
  - In addition, both differences between data values and ratios of data values are meaningful. Data at the ratio level have a true zero.
  - **Example:** If one box weighs 50lbs and another 100lbs $\rightarrow$ the second box weighs twice as much as the first $\rightarrow$ this is not a case in interval data

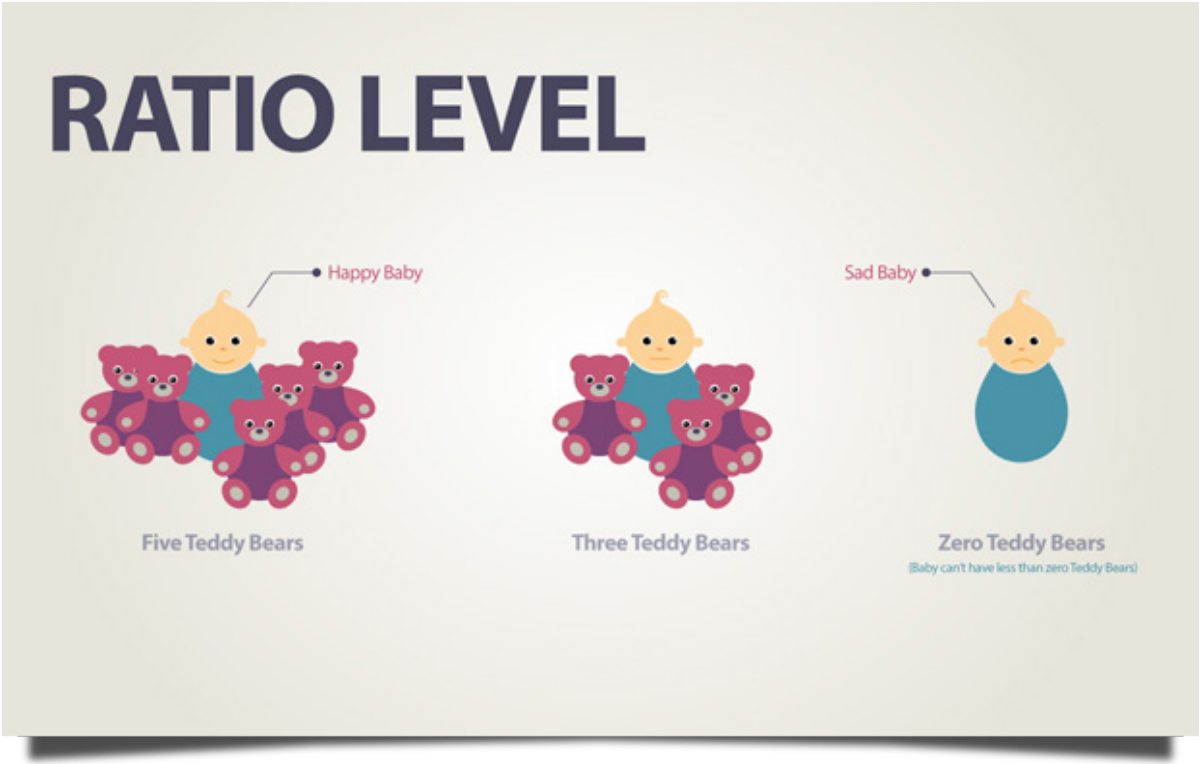


Figure: Ratio level of measurement

- The amounts of teddy bears a certain child has.
- Since we can't have less than zero teddy bears, then the ratio level has a true zero.

- It is defined as a set of mathematical models and analysis methodologies that exploit the available data to generate information and knowledge useful for complex decision-making processes.
- Previously, Knowledge workers are used to take decisions using easy and intuitive methodologies $\rightarrow$ experience, knowledge of the application domain and the available information.
- This approach leads to a **stagnant** decision-making style which is **inappropriate** for the **unstable conditions** $\rightarrow$ frequent and rapid changes in the environment.
- Decision making Process in today's organizations should dynamic, requires rigorous attitude based on analytical methodologies and mathematical models.

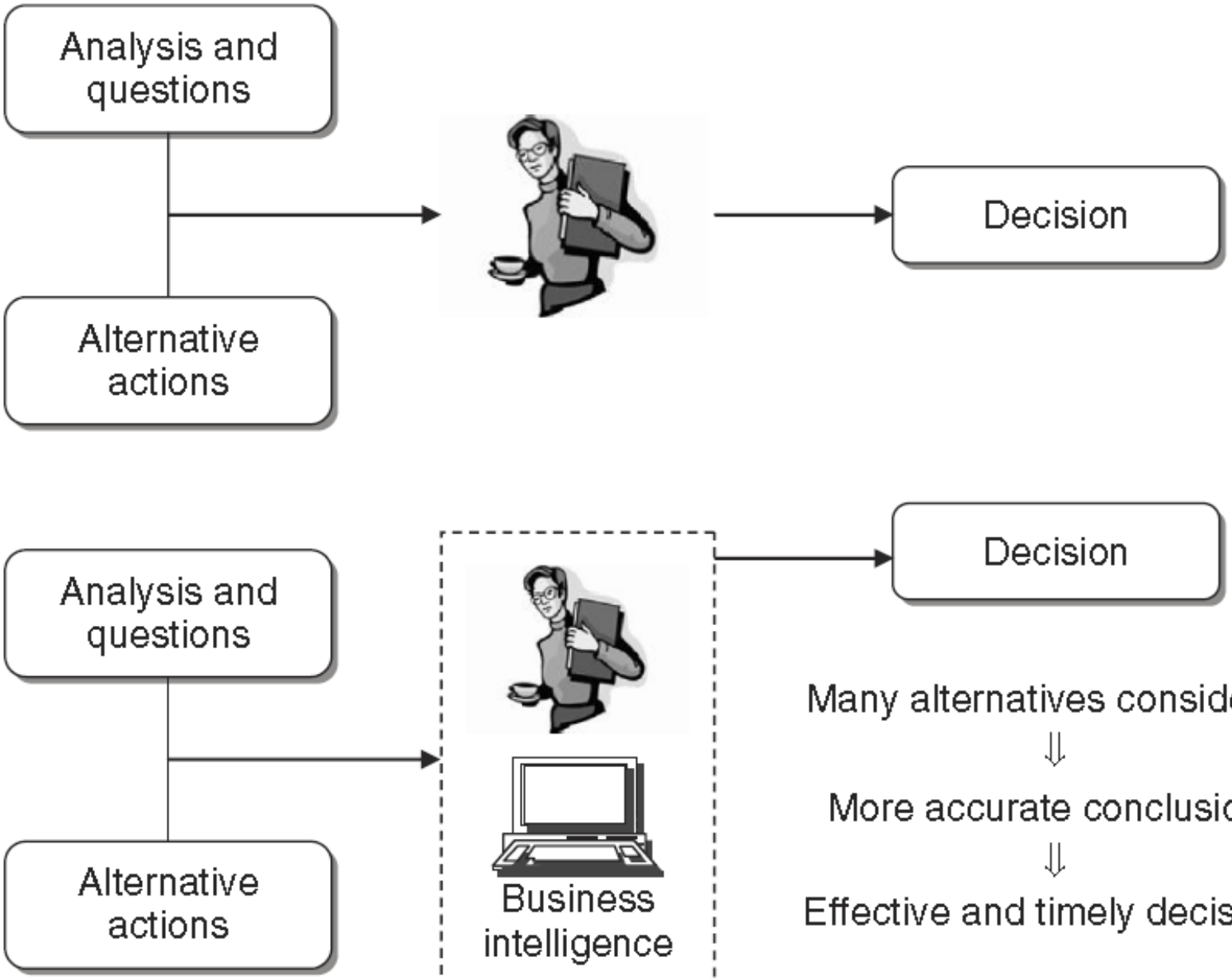
CASE STUDY

Retention in the mobile phone industry:The marketing manager of a mobile phone company realizes that a large number of customers are discontinuing their service, leaving her company in favor of some competing provider. As can be imagined, low customer loyalty, also known as customer attrition or churn, is a critical factor for many companies operating in service industries. Suppose that the marketing manager can rely on a budget adequate to pursue a customer retention campaign aimed at 2000 individuals out of a total customer base of 2 million people. Hence, the question naturally arises of how she should go about choosing those customers to be contacted so as to optimize the effectiveness of the campaign. In other words, how can the probability that each single customer will discontinue the service be estimated so as to target the best group of customers and thus reduce churning and maximize customer retention? By knowing these probabilities, the target group can be chosen as the 2000 people having the highest churn likelihood among the customers of high business value

CASE STUDY

Logistics planning:The logistics manager of a manufacturing company wishes to develop a medium-term logistic-production plan. This is a decision-making process of high complexity which includes, among other choices, the allocation of the demand originating from different market areas to the production sites, the procurement of raw materials and purchased parts from suppliers, the production planning of the plants and the distribution of end products to market areas. In a typical manufacturing company this could well entail tens of facilities, hundreds of suppliers, and thousands of finished goods and components, over a time span of one year divided into weeks. The magnitude and complexity of the problem suggest that advanced optimization models are required to devise the best logistic plan.

- The **main purpose** of business intelligence systems is to provide knowledge workers with tools and methodologies that allow them to make **effective** and **timely decisions**.
- **Effective Decision**:
  - The application of rigorous analytical methods allows decision makers to rely on **information** and **knowledge** which are more dependable.
  - It ensures in-depth examination and thought lead to a deeper awareness and comprehension of the underlying logic of the decision-making process.
- **Timely decisions**:
  - The ability to rapidly react to the actions of competitors and to new market conditions is a critical factor in the success or even the survival of a company.


Analysis and questions
Alternative actions
Decision
Analysis and questions
Alternative actions
Business intelligence
Decision
Many alternatives considered
⇓
More accurate conclusions
⇓
Effective and timely decisions

- If the decision makers depends upon BI system, then overall quality of the decision-making process will be greatly improved.
- With the help of mathematical models and algorithms, it is actually possible to analyze a larger number of alternative actions, achieve more accurate conclusions and reach effective and timely decisions
- Difference between data, information and knowledge
  - **Data**: Generally, data represent a structured codification of single primary entities, as well as of transactions involving two or more primary entities
  - **Information** it is the outcome of extraction and processing activities carried out on data, and it appears **meaningful** for those who receive it in a specific domain.
  - **Knowledge**: Information is transformed into knowledge when it is used to make decisions and develop the corresponding actions.
- The activity of providing support to knowledge workers through the integration of decision-making processes and enabling information technologies is usually referred to as **knowledge management**
- BI and knowledge management share some degree of similarity in

- BI and knowledge management share some degree of similarity in their objectives $\rightarrow$ both are helping knowledge worker for decision making process.
- Boundary between BI and Knowledge-Management Systems
- Knowledge Management system methodologies is focusing on on the treatment of information that is usually unstructured, at times implicit, contained mostly in documents, conversations and past experience.
- BI systems are based on structured information, most often of a quantitative nature and usually organized in a database.