Module 3
Business Intelligence, Data Warehousing, Data Mining, Data Visualization

Business Intelligence

- Business Intelligence (BI) is an umbrella term that includes a variety of IT applications that are used to analyze an organization's data and communicate the information to relevant users.

BIDM Life Cycle

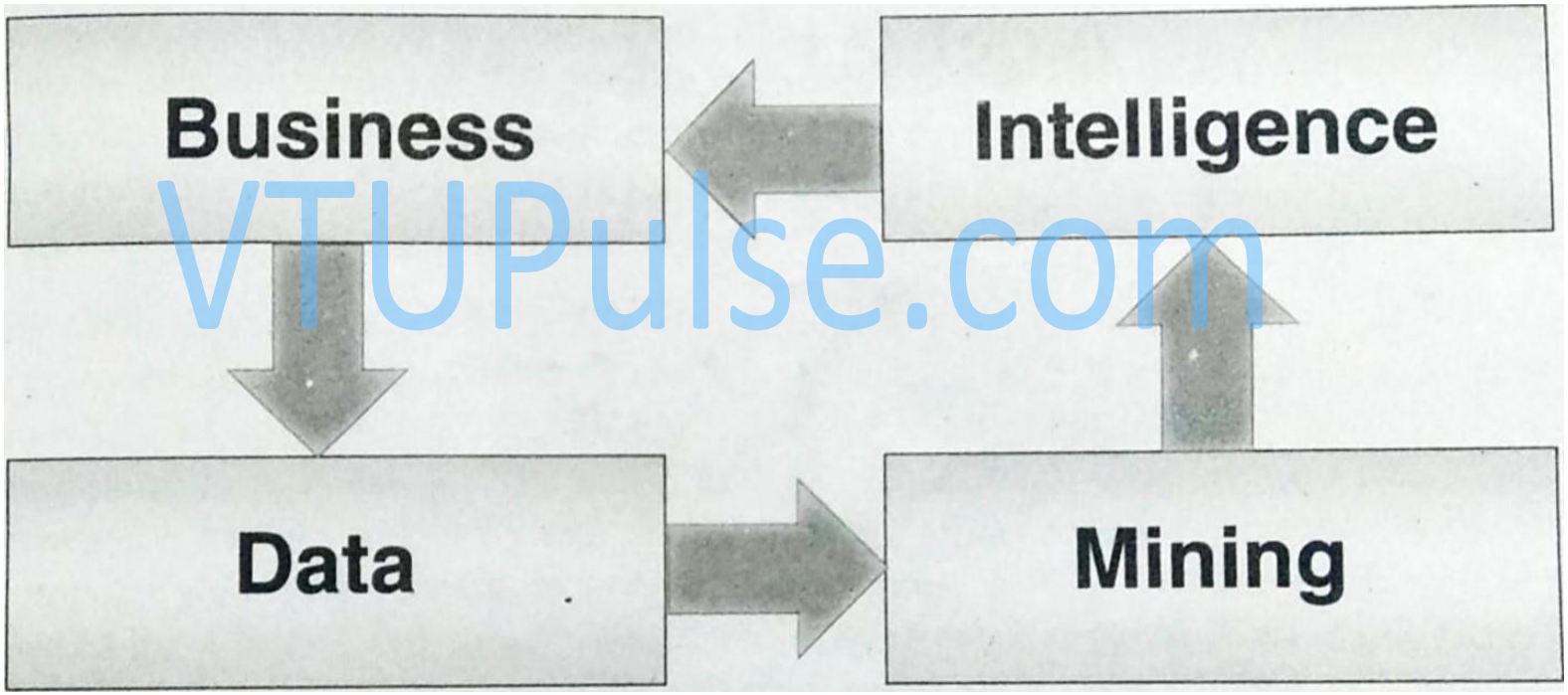


FIGURE 2.1 BIDM cycle

BIDM Life Cycle

- The nature of life and businesses is to grow.
- Information is the life-blood of business.
- They use many techniques for understanding their environment and predicting the future for their own benefit and growth.

BIDM Life Cycle

- One's own data can be the most effective teacher.
- Therefore, organizations should gather data, sift through it, analyze and mine it, find insights, and then embed those insights into their operating procedures.

DECISION TYPES

- There are two main kinds of decisions **strategic decisions** and **operational decisions**.
- BI can help make both better.
- Strategic decisions are those that the direction of the company. The decision to reach out to a new customer set would be a strategic decision.
- In strategic decision-making, the goal itself may or may not be clear, and the same is true for the path to reach the goal.
- BI can help with what-if analysis of many possible scenarios.
- BI can also help create new ideas based on new patterns found from data mining.

DECISION TYPES

- Operational decisions are more routine and tactical decisions, focused on developing greater efficiency.
- Updating an old website with new features will be an operational decision.
- Operational decisions can be made more efficient using an analysis of past data.
- A classification system can be created and modeled using the data of past instances to develop a good model of the domain. This model can help improve operational decisions in the future.

BI TOOLS

- BI includes a variety of software tools and techniques to provide the managers with the information and insights needed to run the business.
- Information can be provided about the current state of affairs with the capability to drill clown into details, and also insights about emerging patterns which lead to projections into the future.
- BI tools include data warehousing, online analytical processing social media analytics, reporting, dashboards, querying, and data mining.

BI APPLICATIONS

- BI tools are required in almost all industries and functions.
- The nature of the information and the speed of action may be different across businesses, but every manager today needs access to BI tools to have up-to-date metrics about business performance.
- Businesses need to embed new insights into their operating processes to ensure that their activities continue to evolve with more efficient practices.

BI APPLICATIONS

The following are some areas of applications of BI and data mining.

1. Retail
2. Telecom
3. Customer Relationship Management
4. Healthcare and Wellness
5. Education
6. Banking
7. Financial Services
8. Insurance
9. Manufacturing
10. Public Sector

Retail

- Retail organizations grow by meeting customer needs with quality products in a convenient, timely, and cost-effective manner. Understanding emerging customer shopping patterns can help retailers organize their products, inventory, store layout, and web presence in order to delight their customers, which in turn would help increase revenue and profits. Retailers generate a lot of transaction and logistics data that can be used to diagnose and solve problems.

Retail

Optimize Inventory Levels

- At different Locations Retailers need to manage their inventories carefully. Carrying too much inventory imposes carrying costs, while carrying too little inventory can cause stock-outs and lost sales opportunities. Predicting sales trends dynamically help retailers move inventory to where it is most in demand. Retail organizations can provide their suppliers with real time information about sales of their items, so the suppliers can deliver their product to the right locations and minimize stock-outs.

Improve Store Layout and Sales Promotions

- A market basket analysis can develop predictive models of the products often sold together. This knowledge of affinities between products can help retailers co-locate those products. Alternatively, those affinity products could be located farther apart to make the customer walk the length and breadth of the store, and thus be exposed to other products. Promotional discounted product bundles can be created to push a nonselling item along with a set of products that sell well together.

Optimize Logistics for Seasonal Effects

- Seasonal products offer tremendously profitable short-term sales opportunities, yet they also offer the risk of unsold inventories at the end of the season. Understanding the products that are in season in which market can help retailers dynamically manage prices to ensure their inventory is sold during the season. If it is raining in a certain area, then the inventory of umbrella and ponchos could be rapidly moved there from nonrainy areas to help increase sales.

Minimize Losses due to Limited Shelf Life

- Perishable goods offer challenges in terms of disposing off the inventory in time. By tracking sales trends, the perishable products at risk of not selling before the sell-by date, can be suitably discounted and promoted.

Telecom

BI in telecom can help the customer side as well as network side of the operations. Key BI applications include churn management, marketing/customer profiling, network failure, and fraud detection.

Churn Management

- Telecom customers have shown a tendency to switch their providers in search for better deals. Telecom companies tend to respond with many incentives and discounts to hold on to customers. However, they need to determine which customers are at a real risk of switching and which others are just negotiating for a better deal. The level of risk should be factored into the kind of deals and discounts that should be given. Millions of such customer calls happen every month. The telecom companies need to provide a consistent and data-based way to predict the risk of the customer switching, and then make an operational decision in real time while the customer call is taking place. A decision-tree or a neural network-based system can be used to guide the customer service call operator to make the right decisions for the company, in a consistent manner.

Telecom

Marketing and Product Creation

- In addition to customer data, telecom companies also store call detail records (CDRs), which can be analyzed to precisely describe the calling behavior of each customer. This unique data can be used to profile customers and then can be used for creating new product/service bundles for marketing purposes. An American telecom company, MCI, created a program called Friends & Family that allowed free calls with one's friends and family on that network, and thus, effectively locked many people into their network.

Network Failure Management

- Failure of telecom networks for technical failures or malicious attacks can have devastating impacts on people, businesses, and society. In telecom infrastructure, some equipment will likely fail with certain mean time between failures. Modeling the failure pattern of various componen[ts] of the network can help with preventive maintenance and capacity planning.

Fraud Management

- There are many kinds of fraud in consumer transactions. subscription fraud occurs when a customer opens an account with the intention Of paying for the services. Superimposition fraud involves illegitimate activity, a person other than the legitimate account holder. Decision rules can developed to analyze each CDR in real time to identify chances of fraud and take effective action.

Customer Relationship Management

A business exists to serve a customer. A happy customer becomes a repeat customer. A business should understand the needs and sentiments of the customer' sell more of its offerings to the existing customers, and also expand the pool Of customers it serves. BI applications can impact many aspects of marketing.

Customer Relationship Management

Maximize the Return on Marketing Campaigns

- Understanding the customer's pain points from data-based analysis can ensure that the marketing messages are fine-tuned to better resonate with customers.

Improve Customer Retention (Churn Analysis)

- It is more difficult and expensive to win new customers than it is to retain existing customers. Scoring each customer on their likelihood to quit can help the business design effective interventions, such as discounts or free services to retain profitable customers in a cost-effective manner.

Maximize Customer Value

- Every contact with the customer should be seen as an opportunity to gauge their current needs. Offering a customer new products and solutions based on those imputed needs can help increase revenue per customer. Even a customer complaint can be seen as an opportunity to wow the customer. Using the knowledge of the customer's history and value, the business can choose to sell a premium service to the customer.

Identify and Delight Highly-Valued Customers

- By segmenting the customers, the best customers can be identified. They can be proactively contacted, and delighted, with greater attention and better service. Loyalty programs can be managed more effectively.

Manage Brand Image

- A business can create a listening post to listen to social media chatter about itself. It can then do sentiment analysis of the text to understand the nature of comments, and respond appropriately to the prospects and customers.

Financial Services

Predict Changes in Bond and Stock Prices

- Forecasting the price of stocks and bonds is a favorite pastime of financial experts as well as lay people. Stock transaction data from the past. along with other variables. can be used to predict future price patterns. This can help traders develop long-term trading strategies.

Assess the Effect of Events on Market Movements

- Decision models using decision trees can be created to assess the impact of events on changes in market volume and prices. Monetary policy changes (such as Federal Reserve interest rate change) or geopolitical changes (such as war in a part of the world) can be factored into the predictive model to help take action with greater confidence and Jess risk.

Identify and Prevent Fraudulent Activities in Trading

- There have unfortunately been many cases of insider trading, leading to many prominent financial industry stalwarts going to jail. Fraud detection models seek out-of-the-ordinary activities, and help identify and flag fraudulent activity patterns.

Manufacturing

- Manufacturing operations are complex systems with interrelated subsystems. From machines working right, to workers having the right skills, to the right components arriving with the right quality at the right time, to money to source the components, many things have to go right.
- Toyota's famous lean manufacturing company works on just-in-time inventory systems to optimize investments in inventory and to improve flexibility in their product-mix.

Manufacturing

Discover Novel Patterns to Improve Product Quality

- Quality of a product can also be tracked, and this data can be used to create a predictive model of product quality deteriorating. Many companies, such as automobile companies, have to recall their products if they have found defects that have a public safety implication. Data mining can help with root cause analysis that can be used to identify sources of errors and helps improve product quality in the future.

Predict/Prevent Machinery Failures

- Statistically, all equipment is likely to break down at some point in time. Predicting which machine is likely to shut down is a complex process. Decision models to forecast machinery failures could be constructed using past data. Preventive maintenance can be planned, and manufacturing capacity can be adjusted, to account for such maintenance activities.

Healthcare and Wellness

Healthcare is one of the biggest sectors in advanced economies. Evidence-based medicine is the newest trend in data-based healthcare management. BI applications can help apply the most effective diagnoses and prescriptions for various ailments. They can also help manage public health issues, and reduce waste and fraud.

Healthcare and Wellness

Diagnose Disease in Patients

- Diagnosing the cause of a medical condition is the critical first step in a medical engagement. Accurately diagnosing cases of cancer or diabetes can be a matter of life and death for the patient. In addition to the patient's own current situation, many other factors can be considered, including the patient's health history, medication history, family history, and Other environmental factors. This makes diagnosis as much of an art form as it is science. Systems, such as IBM Watson, absorb all the medical research to date and make probabilistic diagnoses.

Treatment Effectiveness

- The prescription of medication and treatment is also a difficult choice out of so many possibilities. For example, there are more than 100 medications for hypertension (high blood pressure) alone. There are also interactions in terms of which drugs work well with others and which drugs do not.

Wellness Management

- This includes keeping a track of patient's health records, analyzing customer health trends and proactively advising them to take any needed precautions.

Education

As higher education becomes more expensive and competitive, it becomes a great user of data-based decision-making. There is a strong need for efficiency, increasing revenue, and improving the quality of student experience at all levels of education.

Education

Student Enrollment (Recruitment and Retention)

- Marketing to new potential students requires schools to develop profiles of the students that are most likely to attend. Schools can develop models of what kinds of students are attracted to the school, and then reach out to those students. The students at risk of not returning can be flagged, and corrective measures can be taken in time.

Course Offerings

- Schools can use the class enrolment data to develop models of which new courses are likely to be more popular with students. This can h[el]p increase class size, reduce costs, and improve student satisfaction.

Banking

- Banks make loans and offer credit cards to millions of customers. They are interested in improving the quality of loans and reducing bad debts. They want to retain more good customers, and sell more services to them.

Banking

Automate the Loan Application Process

- Decision models can be generated from past data that predict the likelihood of a loan proving successful. These be inserted in business processes to automate the financial loan approval process.

Detect Fraudulent Transactions

- Billions of financial transactions happen around the world every day. Exception-seeking models can identify Patterns of fraudulent transactions. For example, if money is being transferred to an unrelated account for the first time, it could be a fraudulent transact

Maximize Customer Value

- Selling more products and services to existing customers is often the easiest way to increase revenue. A checking account customer in good standing could be offered home, auto, educational loans on more favorable terms than other customers, and thus. the value generated from that customer could be increased.

Public Sector

- Government gathers a large amount of data by virtue of their regulatory function.
- That data could be analyzed for developing models of effective functioning. There are innumerable applications that can benefit from mining that data. A couple of sample applications are shown here.

Public Sector

Law Enforcement

- Social behavior is a lot more patterned and predictable than one would imagine. For example, Los Angeles Police Department (LAPD) mined the data from its 13 million crime records over 80 years and developed models of what kind of crime going to happen when and where. By increasing patrolling in those particular areas, LAPD was able to reduce property crime by 27 percent. Internet chatter can be analyzed to learn about and prevent any evil designs.

Scientific Research

- Any large collection of research data is amenable to being mined for patterns and insights. Protein folding (microbiology), nuclear reaction analysis (sub-atomic physics), disease control (public health) are some examples where data mining can yield powerful new insights.

Data Warehousing

A data warehouse (DW) is an organized collection of integrated, subject oriented databases designed to support decision support functions. DW is organized at the right level of granularity to provide clean enterprise wide data in a standardized format for reports, queries, and analysis.

DW offers many business and technical benefits.

1. DW supports business reporting and data mining activities. It can facilitate distributed access to up-to-date business knowledge for departments and functions, thus improving business efficiency and customer service.
2. DW can present a competitive advantage by facilitating decision making and helping reform business processes.
3. DW enables a consolidated view of corporate data, all cleaned and organized. Thus, the entire organization can see an integrated view of itself.
4. DW thus provides better and timely information. It simplifies data access and allows end users to perform extensive analysis.

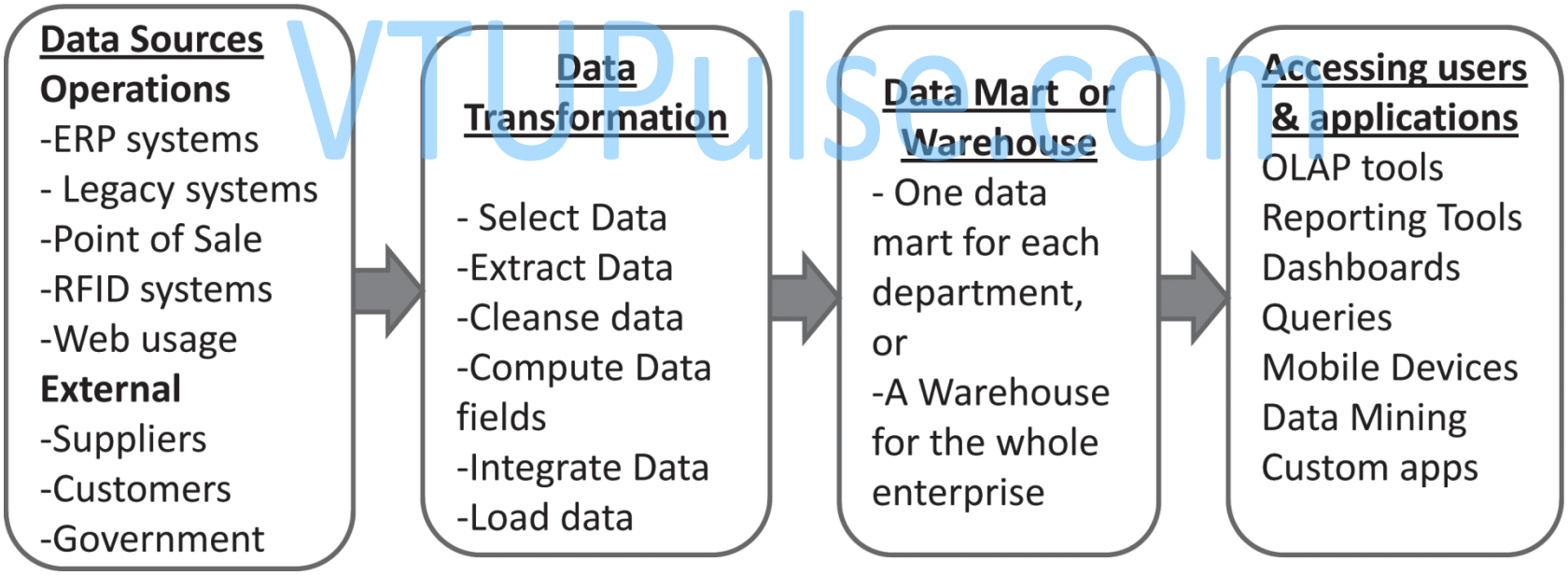
Design Considerations for DW

The objective of DW is to provide business knowledge to support decision making. For DW to serve its objective, it should be aligned around those decisions. It should be comprehensive, easy to access, and up-to-date. Here are some requirements for a good DW:

1. ***Subject-oriented:*** To be effective, DW should be designed around a subject domain, that is, to help solve a certain category of problems.
2. ***Integrated:*** DW should include data from many functions that can shed light on a particular subject area. Thus, the organization can benefit from a comprehensive view of the subject area.
3. ***Time-variant (time series):*** The data in DW should grow at daily or other chosen intervals. That allows latest comparisons over time.
4. ***Nonvolatile:*** DW should be persistent, that is, it should not be created on the fly from the operations databases. Thus, DW is consistently available for analysis, across the organization and over time.
5. ***Summarized:*** DW contains rolled-up data at the right level for queries and analysis. The rolling up helps create consistent granularity for effective comparisons. It helps reduces the number of variables or dimensions of the data to make them more meaningful for the decision makers.
6. ***Not normalized:*** DW often uses a star schema, which is a rectangular central table, surrounded by some lookup tables. The single-table view significantly enhances speed of queries.
7. ***Metadata:*** Many of the variables in the database are computed from other variables in the operational database. For example, total daily sales may be a computed field. The method of its calculation for each variable should be effectively documented. Every element in DW should be sufficiently well-defined.
8. ***Near real-time and/or right-time (active):*** DWs should be updated in near real-time in many high-transaction volume industries, such as airlines. The cost of implementing and updating DW in real time could discourage others. Another downside of real-time DW is the possibilities of inconsistencies in reports drawn just a few minutes apart.

DW Architecture

- DW has four key elements shown in below figure



DW Architecture

- The first element is the data sources that provide the raw data.
- The second element is the process of transforming that data to meet the decision needs.
- The third element is the methods of regularly and accurately loading of that data into EDW or data marts.
- The fourth element is the data access and analysis part, where devices and applications use the data from DW to deliver insights and other benefits to users.

Data Sources

DWs are created from structured data sources. Unstructured data, such as text data, would need to be structured before inserted into DW.

1. Operations data include data from all business applications, including from ERPs systems that form the backbone of an organization's IT systems. The data to be extracted will depend upon the subject matter of DW. For example, for a sales/marketing DW, only the data about customers, orders, customer service, and so on would be extracted.
2. Other applications, such as point-of-sale (POS) terminals and e-commerce applications, provide customer-facing data. Supplier data could come from supply chain management systems. Planning and budget data should also be added as needed for making comparisons against targets.
3. External syndicated data, such as weather or economic activity data, could also be added to DW, as needed, to provide good contextual information to decision makers.

Data Transformation Processes

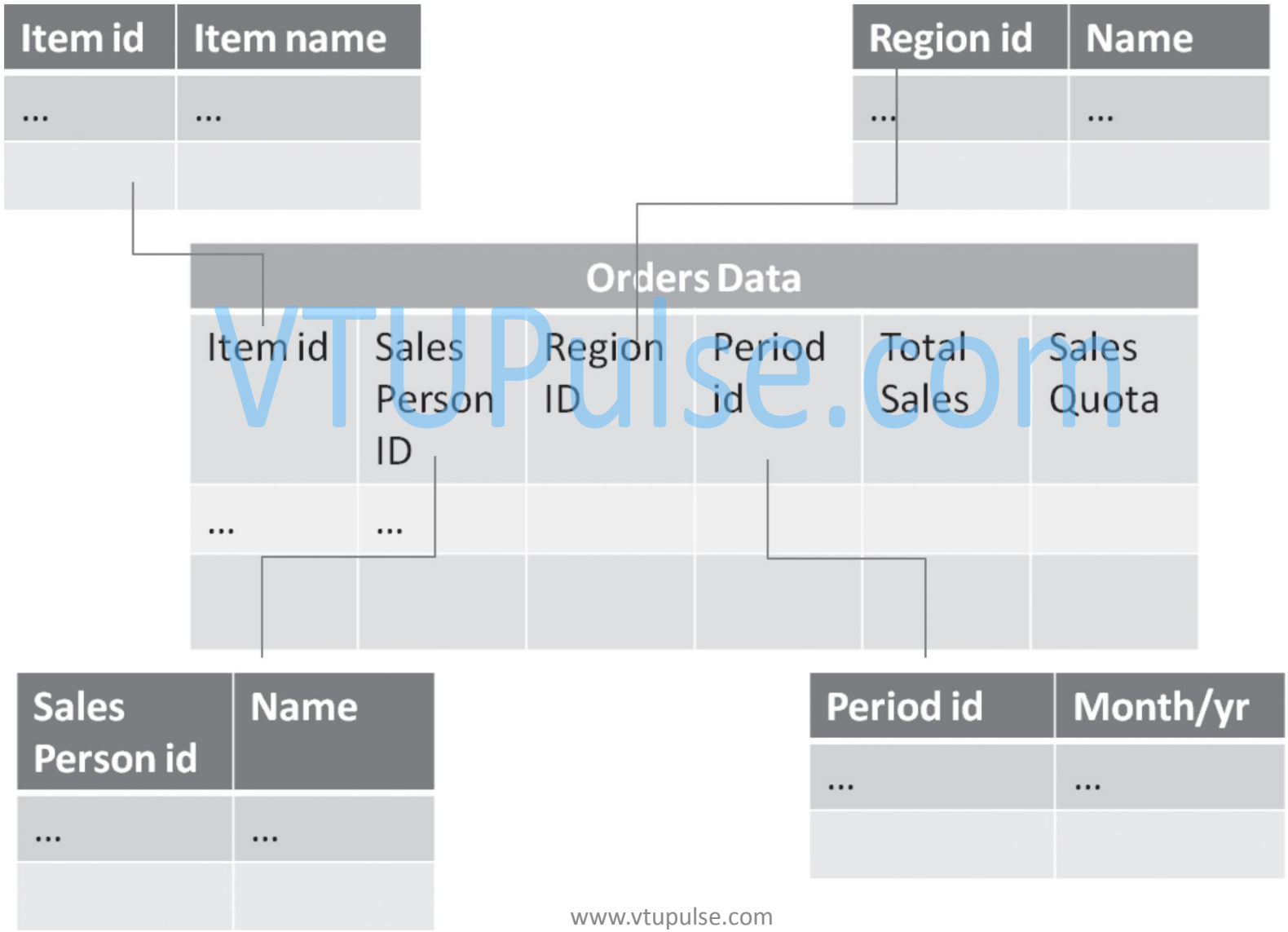
The heart of a useful DW is the processes to populate the DW with good quality data. This is called the extract-transform-load (ETL) cycle.

1. Data should be extracted from many operational (transactional) database sources on a regular basis.
2. Extracted data should be aligned together by key fields. It should be cleansed of any irregularities or missing values. It should be rolled up together to the same level of granularity. Desired fields, such as daily sales totals, should be computed. The entire data should then be brought to the same format as the central table of DW.
3. The transformed data should then be uploaded into DW.

DW Design

- Star schema is the preferred data architecture for most DWs. There is a central fact table that provides most of the information of interest. There are lookup tables that provide detailed values for codes used in the central table. For example, the central table may use digits to represent a sales person. The lookup table will help provide the name for that sales person code. Here is an example of a star schema for a data mart for monitoring sales performance (Figure).
- Other schemas include the snowflake architecture. The difference between a star and snowflake is that in the latter, the lookup tables can have their own further lookup tables.
- There are many technology choices for developing DW. This includes selecting the right database management system and the right set of data management tools. There are a few big and reliable providers of DW systems. The provider of the operational DBMS may be chosen for DW also.

DW Design - *Star schema architecture*



DW Access

Data from DW could be accessed for many purposes, through many devices.

1. A primary use of DW is to produce routine management and monitoring reports. For example, a sales performance report would show sales by many dimensions, and compared with plan. A dashboarding system will use data from the warehouse and present analysis to users. The data from DW can be used to populate customized performance dashboards for executives. The dashboard could include drill-down capabilities to analyze the performance data for root cause analysis.
2. The data from the warehouse could be used for ad hoc queries and any other applications that make use of the internal data.
3. Data from DW is used to provide data for mining purposes. Parts of the data would be extracted, and then combined with other relevant data, for data mining.

Data Mining

- Data mining is the art and science of discovering knowledge, insights, and patterns in data.
- Patterns must be valid, novel, potentially useful, and understandable.
- The implicit assumption is that data about the past can reveal patterns of activity that can be projected into the future.
- Data mining is a multidisciplinary field that borrows techniques from a variety of fields.
- It utilizes the knowledge of data quality and data organizing from the databases area. It draws modeling and analytical techniques from statistics and computer science (artificial intelligence) areas. It also draws the knowledge of decision-making from the field of business management.

Data Mining

- For example, “customers who buy *cheese* and *milk* also buy *bread* 90 percent of the time” would be a useful pattern for a grocery store, which can then stock the products appropriately.
- Similarly, “people with blood pressure greater than 160 and an age greater than 65 were at a high risk of dying from a heart stroke” is of great diagnostic value for doctors, who can then focus on treating such patients with urgent care and great sensitivity.

Gathering and Selecting Data

- The total amount of data in the world is doubling every 18 months. There is an ever-growing avalanche of data coming with higher velocity, volume, and variety. One has to quickly use it or lose it.
- To learn from data, one needs to effectively gather quality data, clean and organize it, and then efficiently process it.
- One requires the skills and technologies for consolidation and integration of data elements from many sources.

Gathering and Selecting Data

- Gathering and curating data takes time and effort, particularly when it is unstructured or semistructured.
- Unstructured data can come in many forms like databases, blogs, images, videos, and chats. There are streams of unstructured social media data from blogs, chats, and tweets.
- There are also streams of machine-generated data from connected machines, RFID tags, the internet of things, and so on. The data should be put in rectangular data shapes with clear columns and rows before submitting it to data mining.
- Knowledge of the business domain helps select the right streams of data for pursuing new insights. Data that suits the nature of the problem being solved should be gathered. The data elements should be relevant, and suitably address the problem being solved.

Data Cleansing and Preparation

- The quality of data is critical to the success and value of the data mining project.
- Otherwise, the situation will be of the kind of garbage in and garbage out (GIGO).
- The quality of incoming data varies by the source and nature of data. Data from internal operations is likely to be of higher quality, as it will be accurate and consistent.
- Data from social media and other public sources is less under the control of business, and is less likely to be reliable.
- Data almost certainly needs to be cleansed and transformed before it can be used for data mining.
- There are many ways in what data may need to be cleansed—filling missing values, reigning in the effects of outliers, transforming fields and many more. Data cleansing and preparation is a labor-intensive or semi-automated activity that can take up to 60 to 70 percent of the time needed for a data mining project.

Data Cleansing and Preparation

1. Duplicate data needs to be removed.
2. Missing values need to be filled in, or those rows should be removed from analysis. Missing values can be filled in with average or modal or default values.
3. Data elements may need to be transformed from one unit to another. For example, total costs of health care and the total number of patients may need to be reduced to cost/patient to allow comparability of that value.
4. Continuous values may need to be binned into a few buckets to help with some analyses. For example, work experience could be binned as low, medium, and high.
5. Data elements may need to be adjusted to make them comparable over time. For example, currency values may need to be adjusted for inflation; they would need to be converted to the same base year for comparability. They may need to be converted to a common currency.

Data Cleansing and Preparation

6. Outlier data elements need to be removed after careful review, to avoid the skewing of results. For example, one big donor could skew the analysis of alumni donors in an educational setting.
7. Any biases in the selection of data should be corrected to ensure the data is representative of the phenomena under analysis. If the data includes many more members of one gender than is typical of the population of interest, then adjustments need to be applied to the data.
8. Data should be brought to the same granularity to ensure comparability. Sales data may be available daily, but the sales person compensation data may only be available monthly. To relate these variables, the data must be brought to the lowest common denominator, in this case, monthly.
9. Data may need to be selected to increase information density. Some data may not show much variability, because it was not properly recorded or for any other reasons. This data may dull the effects of other differences in the data and should be removed to improve the information density of the data.

Outputs of Data Mining

- Data mining techniques can serve different types of objectives. The outputs of data mining will reflect the objective being served. There are many representations of the outputs of data mining.
- One popular form of data mining output is a decision tree. It is a hierarchically branched structure that helps visually follow the steps to make a model-based decision. The tree may have certain attributes, such as probabilities assigned to each branch. A related format is a set of business rules, which are if-then statements that show causality. A decision tree can be mapped to business rules. If the objective function is prediction, then a decision tree or business rules are the most appropriate mode of representing the output.
- The output can be in the form of a regression equation or mathematical function that represents the best fitting curve to represent the data. This equation may include linear and nonlinear terms. Regression equations are a good way of representing the output of classification exercises. These are also a good representation of forecasting formulae.

Evaluating Data Mining Results

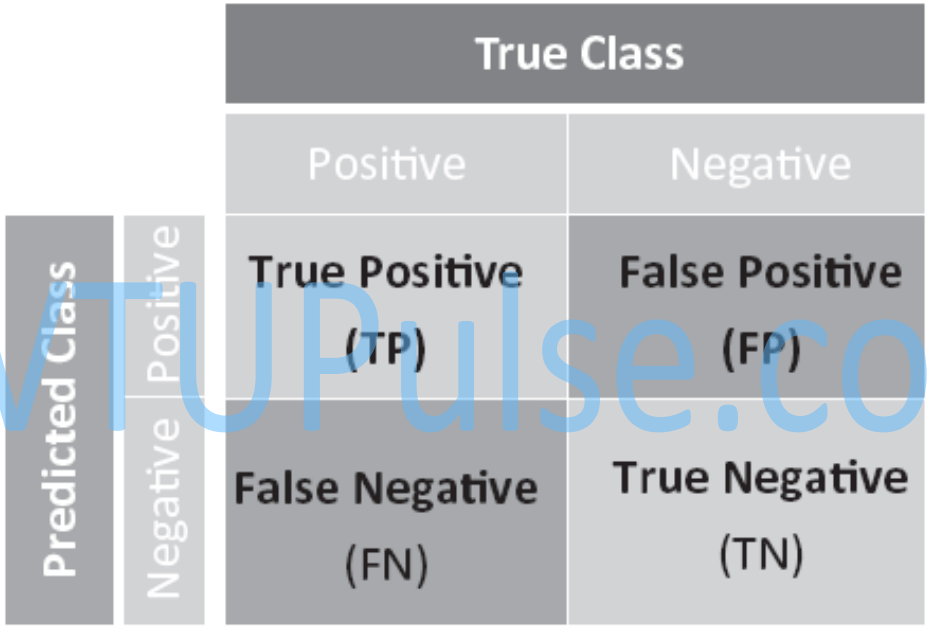
- There are two primary kinds of data mining processes: supervised learning and unsupervised learning. In supervised learning, a decision model can be created using past data, and the model can then be used to predict the correct answer for future data instances. Classification is the main category of supervised learning activity.
- There are many techniques for classification, decision trees being the most popular one. Each of these techniques can be implemented with many algorithms.
- A common metric for all of classification techniques is predictive accuracy.

Predictive Accuracy = (Correct Predictions) / Total Predictions

Evaluating Data Mining Results

- Suppose a data mining project has been initiated to develop a predictive model for cancer patients using a decision tree. Using a relevant set of variables and data instances, a decision tree model has been created.
- The model is then used to predict other data instances.
- When a true positive data point is positive, that is a correct prediction, called a true positive (TP).
- When a true negative data point is classified as negative, that is a true negative (TN).
- When a true-positive data point is classified by the model as negative, that is an incorrect prediction, called a false negative (FN).
- When a true-negative data point is classified as positive, that is classified as a false positive (FP).
- This is called the confusion matrix.

Evaluating Data Mining Results



- Thus, the predictive accuracy can be specified by the following formula.

  Predictive Accuracy = (TP + TN) / (TP + TN + FP + FN).

- All classification techniques have a predictive accuracy associated with a predictive model. The highest value can be 100 percent. In practice, predictive models with more than 70 percent accuracy can be considered usable in business domains, depending upon the nature of the business.

Data Mining Techniques

- Data may be mined to help make more efficient decisions in the future. Or it may be used to explore the data to find interesting associative patterns. The right technique depends upon the kind of problem being solved

Important Data Mining Techniques		
Supervised Learning: Classification	Machine Learning Techniques	Decision Trees
		Artificial Neural Networks
	Statistical Techniques	Regression
Unsupervised Learning: Exploration	Machine Learning Techniques	Cluster Analysis
		Association Rule Mining

Data Mining Techniques

- The most important class of problems solved using data mining are classification problems.
- These are problems where data from past decisions is mined to extract the few rules and patterns that would improve the accuracy of the decision-making process in the future.
- The data of past decisions is organized and mined for decision rules or equations, which are then codified to produce more accurate decisions.
- Classification techniques are called supervised learning as there is a way to supervise whether the model's prediction is right or wrong.

Decision Trees

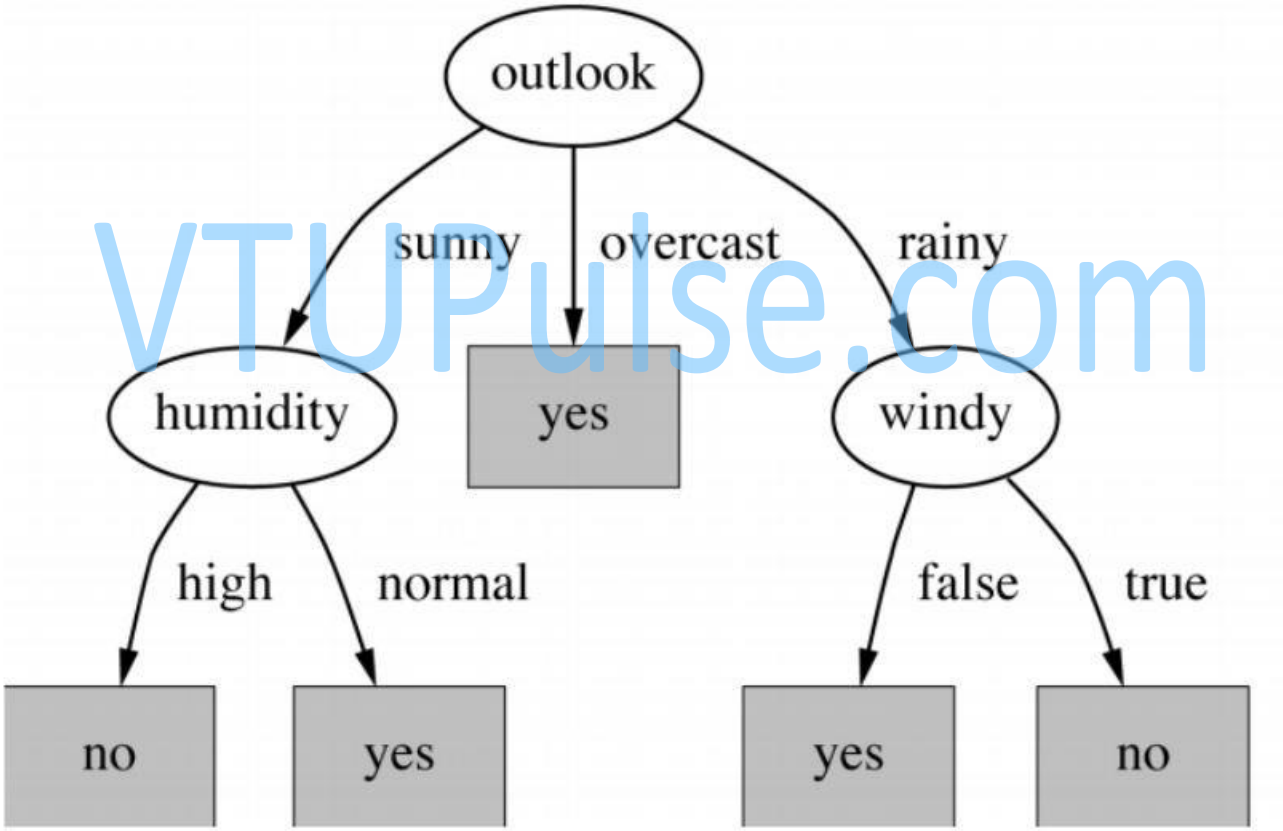
A decision tree is a hierarchically organized branched, structured to help make decision in an easy and logical manner.

Decision trees are the most popular data mining technique, for many reasons.

1. Decision trees are easy to understand and easy to use, by analysts as well as executives. They also show a high predictive accuracy.
2. They select the most relevant variables automatically out of all the available variables for decision-making.
3. Decision trees are tolerant of data quality issues and do not require much data preparation from the users.
4. Even nonlinear relationships can be handled well by decision trees.

There are many algorithms to implement decision trees. Some of the popular ones are C5, CART, and CHAID.

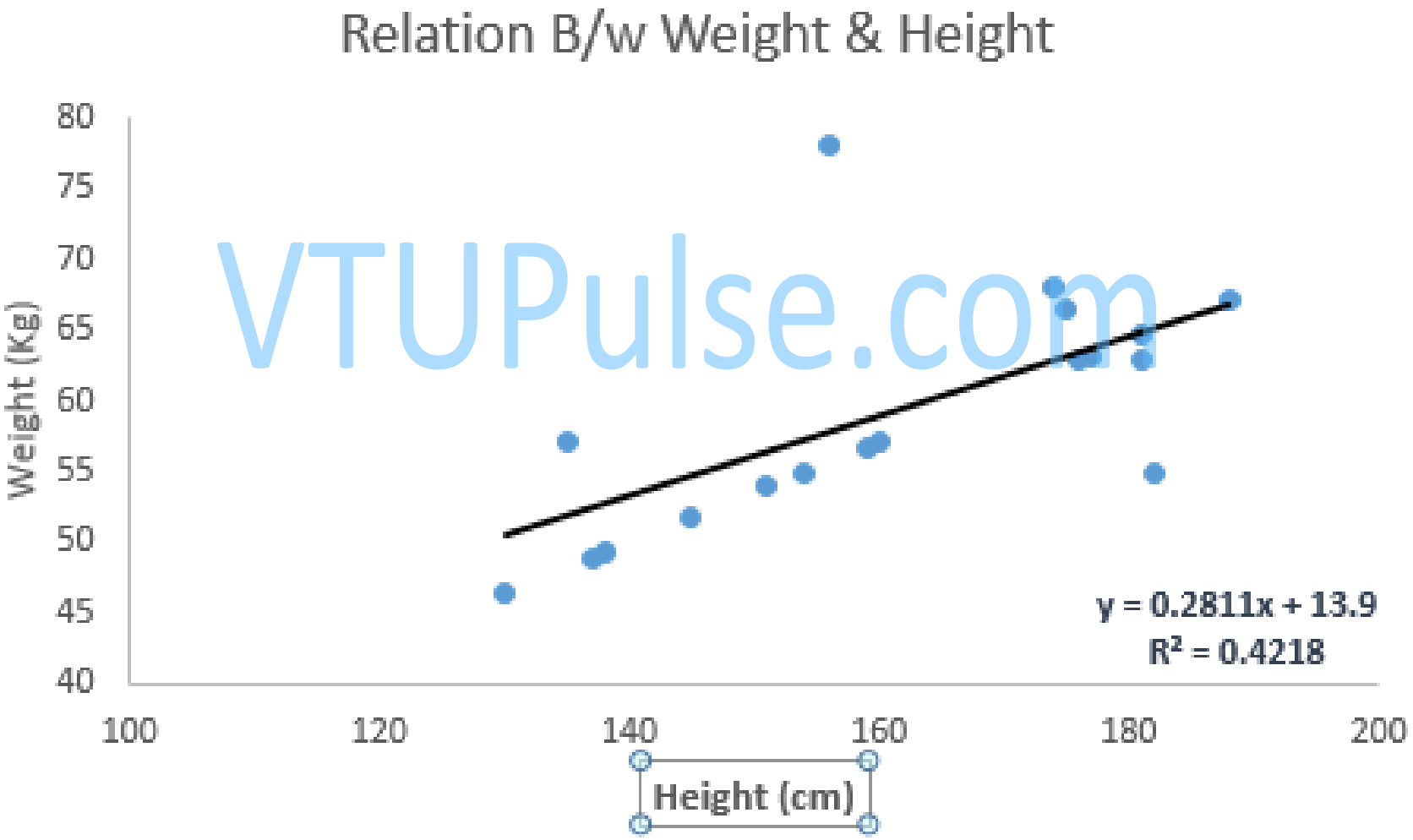
Final decision tree



Regression

- *Regression* is a relatively simple and the most popular statistical data mining technique. The goal is to fit a smooth well-defined curve to the data.
- Regression analysis techniques, for example, can be used to model and predict the energy consumption as a function of daily temperature.
- Simply plotting the data shows a nonlinear curve. Applying a nonlinear regression equation will fit the data very well with high accuracy.
- Thus, the energy consumption on any future day can be predicted using this equation.

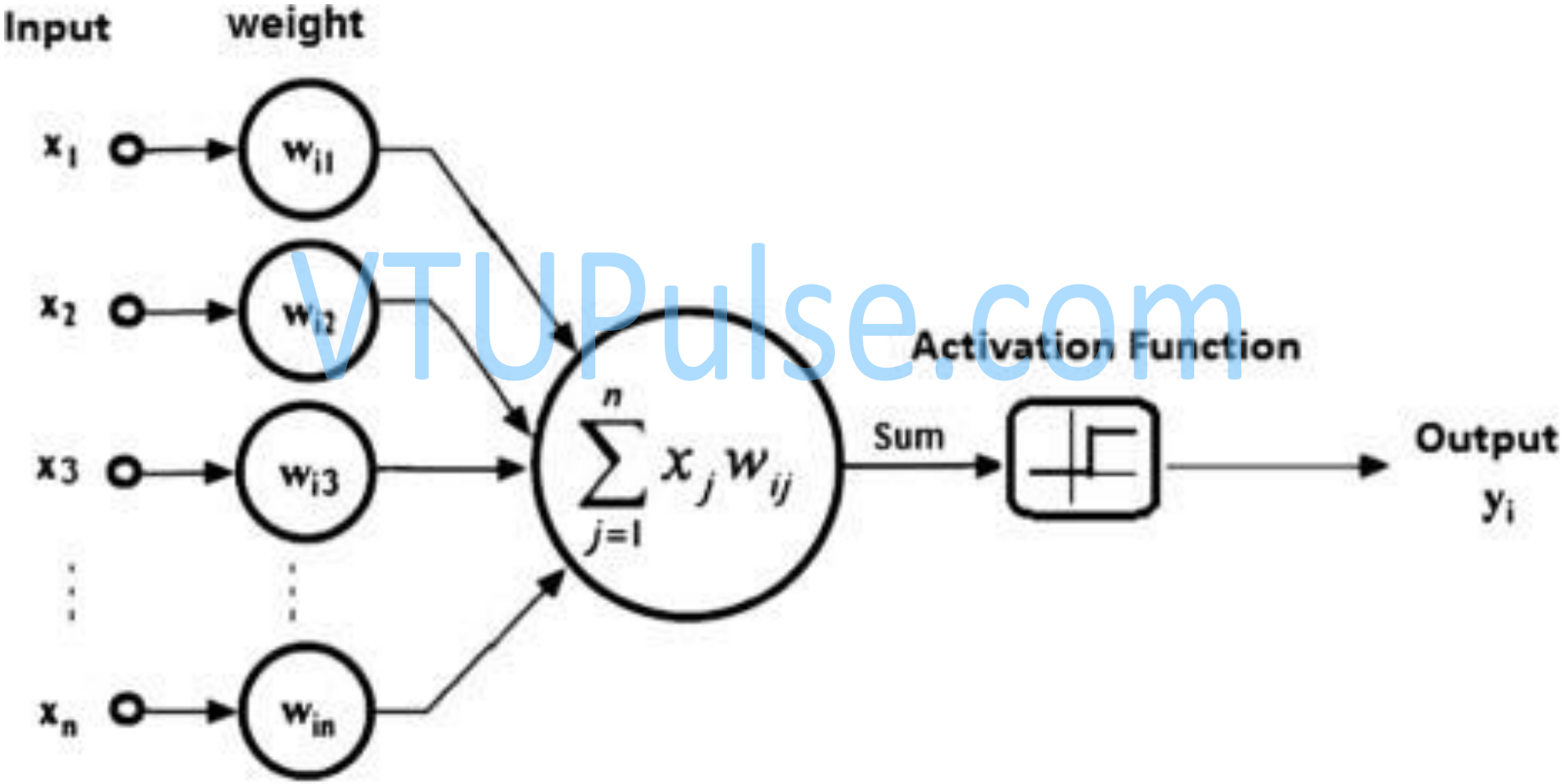
Regression



Artificial neural network (ANN)

- *Artificial neural network* (ANN) is a sophisticated data mining technique from the Artificial Intelligence stream in Computer Science.
- It mimics the behavior of human neural structure: Neurons receive stimuli, process them, and communicate their results to other neurons successively, and eventually a neuron outputs a decision.
- A decision task may be processed by just one neuron and the result may be communicated soon.
- The neural network can be trained by making a decision over and over again with many data points. It will continue to learn by adjusting its internal computation and communication parameters based on feedback received on its previous decisions. The intermediate values passed within the layers of neurons may not make intuitive sense to an observer. Thus, the neural networks are considered a black-box system.

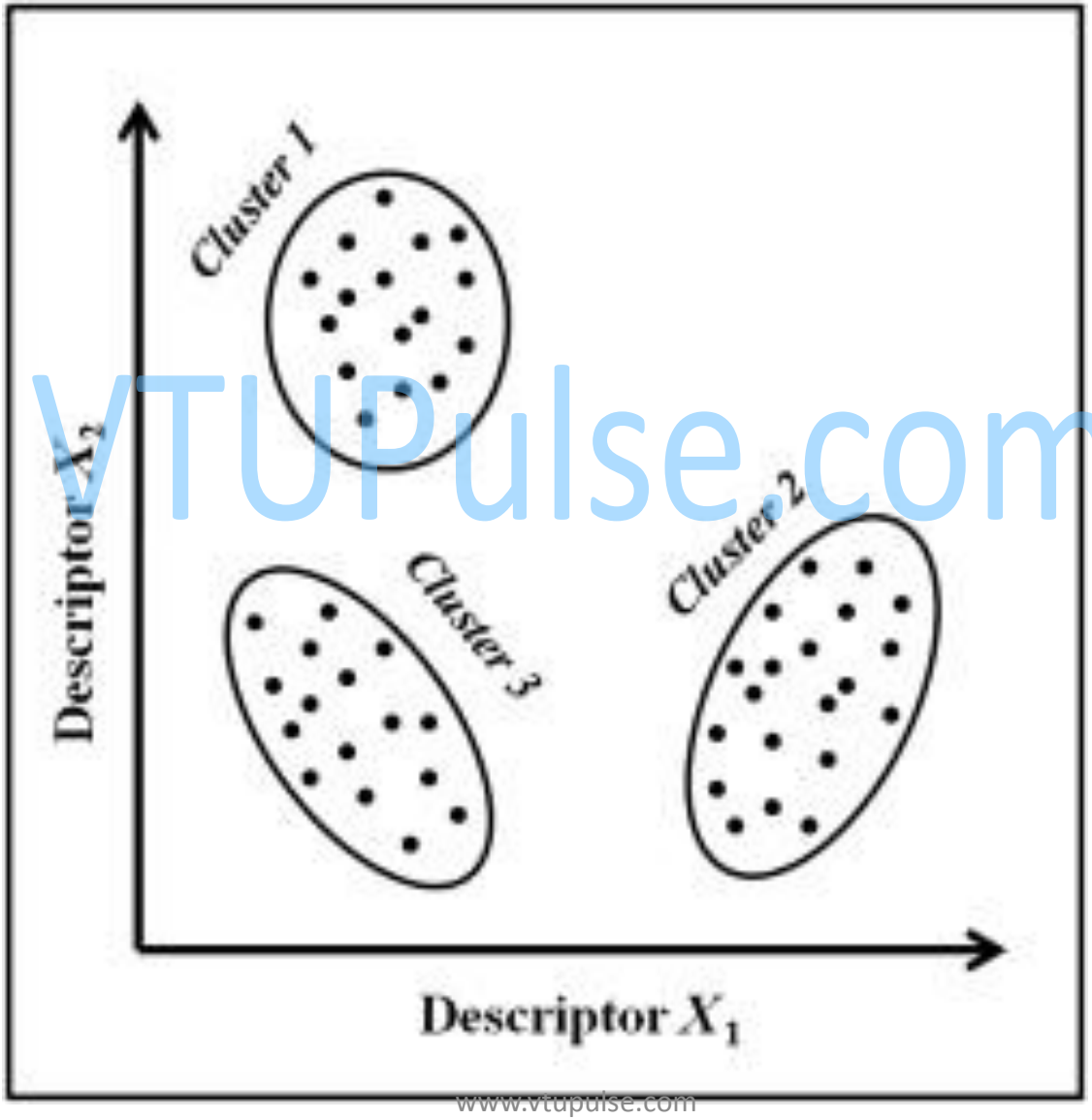
Artificial neural network (ANN)



Cluster analysis

- *Cluster analysis* is an exploratory learning technique that helps in identifying a set of similar groups in the data.
- It is a technique used for automatic identification of natural groupings of things. Data instances that are similar to (or near) each other are categorized into one cluster, while data instances that are very different (or far away) from each other are categorized into separate clusters.
- There can be any number of clusters that could be produced by the data. The K-means technique is a popular technique and allows the user guidance in selecting the right number (K) of clusters from the data.
- Clustering is also known as the segmentation technique. The technique shows the clusters of things from past data. The output is the centroids for each cluster and the allocation of data points to their cluster. The centroid definition is used to assign new data instances that can be assigned to their cluster homes.

Cluster analysis



Association rules

- *Association rules* are a popular data mining method in business, especially where selling is involved. Also known as market basket analysis, it helps in answering questions about cross-selling opportunities.
- This is the heart of the personalization engine used by e-commerce sites like Amazon.com and streaming movie sites like Netflix.com.
- The technique helps find interesting relationships (affinities) between variables (items or events). These are represented as rules of the form $X \Rightarrow Y$, where X and Y are sets of data items.
- A form of unsupervised learning, it has no dependent variable; and there are no right or wrong answers. There are just stronger and weaker affinities. Thus, each rule has a confidence level assigned to it.

Association rules

Transaction ID	Items
1	Bread, Milk, Cheese, Butter
2	Bread, Milk, Cheese, Butter, Eggs, sugar
3	Milk, Butter, cocoa
4	Bread, milk, Cheese, Eggs

Tools and Platforms for Data Mining

Data mining tools have existed for many decades. However, they have recently become more important as the values of data have grown and the field of big data analytics has come into prominence. There are a wide range of data mining platforms available in the market today.

1. There are simple end-user data mining tools, such as MS Excel, and there are more sophisticated tools, such as IBM SPSS Modeler.
2. There are stand-alone tools, and there are tools embedded in an existing transaction processing or data warehousing or ERP system.
3. There are open-source and freely available tools, such as Weka, and there are commercial products.
4. There are text-based tools that require some programing skills, and there are Graphical User Interface (GUI)-based drag-and-drop format tools.
5. There are tools that work only on proprietary data formats, and there are those directly accept data from a host of popular data management tools formats.

Comparison of popular data mining platforms

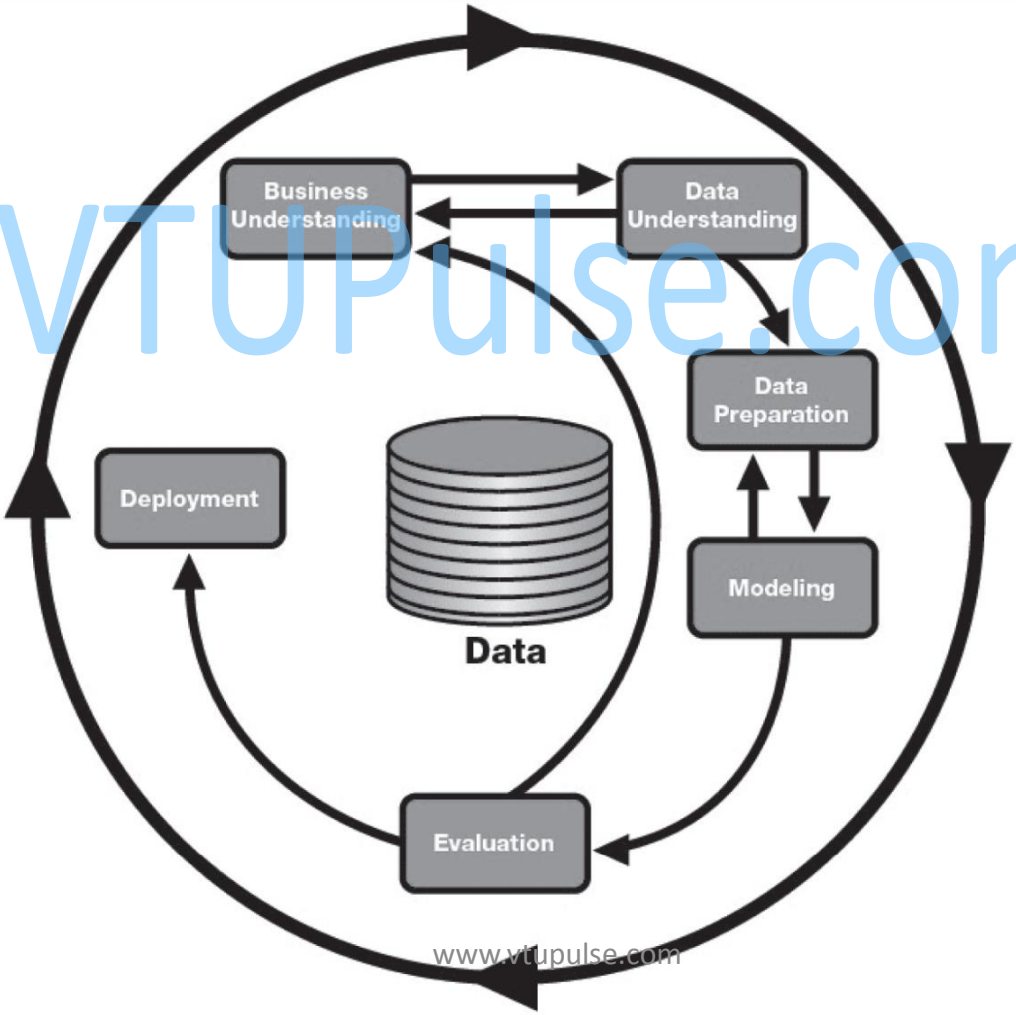
Feature	Excel	IBM SPSS Modeler	Weka
Ownership	Commercial	Commercial, expensive	Open-source, free
Data mining features	Limited, extensible with add-on modules	Extensive features, unlimited data sizes	Extensive, performance issues with large data
Stand-alone	Stand-alone	Embedded in BI software suites	Stand-alone
User skills needed	End users	Skilled BI analysts	Skilled BI analysts
User interface	Select and click, easy	Drag-and-drop use, colorful, beautiful GUI	GUI, mostly b&w text output
Data formats	Industry standard	Variety of data sources accepted	Proprietary

Data Mining Best Practices

- Effective and successful use of data mining activity requires both business and technology skills. The business aspects help understand the domain and the key questions. It also helps one imagine possible relationships in the data and create hypotheses to test it. The IT aspects help fetch the data from many sources, clean up the data, assemble it to meet the needs of the business problem, and then run the data mining techniques on the platform.
- An important element is to go after the problem iteratively. It is better to divide and conquer the problem with smaller amounts of data, and get closer to the heart of the solution in an iterative sequence of steps. There are several best practices learned from the use of data mining techniques over a long period of time.
- The data mining industry has proposed a Cross-Industry Standard Process for Data Mining (CRISP-DM).
- It has six essential steps:

Data Mining Best Practices

- ***CRISP-DM data mining cycle***



Data Mining Best Practices

1. The first and most important step in data mining is business understanding, that is, asking the right business questions. A question is a good one if answering it would lead to large payoffs for the organization, financially and otherwise. In other words, selecting a data mining project is like any other project, in which it should show strong payoffs if the project is successful. There should be strong executive support for the data mining project, which means that the project aligns well with the business strategy.
2. A second important step is to be creative and open in proposing imaginative hypotheses for the solution. Thinking outside the box is important, both in terms of a proposed model as well in the data sets available and required.
3. The data should be clean and of high quality. It is important to assemble a team that has a mix of technical and business skills, who understand the domain and the data. Data cleaning can take 60 to 70 percent of the time in a data mining project. It may be desirable to add new data elements from external sources of data that could help improve predictive accuracy.
4. Patience is required in continuously engaging with the data until the data yields some good insights. A host of modeling tools and algorithms should be used. A tool could be tried with different options, such as running different decision tree algorithms.
5. One should not accept what the data says at first. It is better to triangulate the analysis by applying multiple data mining techniques and conducting many what-if scenarios, to build confidence in the solution. Evaluate the model's predictive accuracy with more test data.
6. The dissemination and rollout of the solution is the key to project success. Otherwise the project will be a waste of time and will be a setback for establishing and supporting a data-based decision-process culture in the organization. The model should be embedded in the organization's business processes.

Myths about Data Mining

There are many myths about this area, scaring away many business executives from using data mining.

***Myth #1*:** Data mining is about algorithms: Data mining is used by business to answer important and practical business questions. Formulating the problem statement correctly and identifying imaginative solutions for testing are far more important before the data mining algorithms get called in.

***Myth #2*:** Data mining is about predictive accuracy: While important, predictive accuracy is a feature of the algorithm. As in myth #1, the quality of output is a strong function of the right problem, right hypothesis, and the right data.

***Myth #3*:** Data mining requires a data warehouse: While the presence of a data warehouse assists in the gathering of information, sometimes the creation of the data warehouse itself can benefit from some exploratory data mining.

***Myth #4*:** Data mining requires large quantities of data: Many interesting data mining exercises are done using small- or medium-sized data sets.

***Myth #5*:** Data mining requires a technology expert: Many interesting data mining exercises are done by end users and executives using simple everyday tools like spreadsheets.

Data Mining Mistakes

Data mining is an exercise in extracting nontrivial useful patterns in the data. It requires a lot of preparation and patience to pursue the many leads that data may provide. Much domain knowledge, tools, and skill are required to find such patterns. Here are some of the more common mistakes in doing data mining, and should be avoided.

***Mistake #1*:** Selecting the wrong problem for data mining: Without the right goals or having no goals, data mining leads to a waste of time. Getting the right answer to an irrelevant question could be interesting, but it would be pointless.

***Mistake #2*:** Buried under mountains of data without clear metadata: It is more important to be engaged with the data, than to have lots of data. The relevant data required may be much less than initially thought. There may be insufficient knowledge about the data or metadata.

***Mistake #3*:** Disorganized data mining: Without clear goals, much time is wasted. Doing the same tests using the same mining algorithms repeatedly and blindly, without thinking about the next stage, without a plan, would lead to wasted time and energy. This can come from being sloppy about keeping track of the data mining procedure and results.

Data Mining Mistakes

***Mistake #4*:** Insufficient business knowledge: Without a deep understanding of the business domain, the results would be gibberish and meaningless. Do not make erroneous assumptions, courtesy of experts. Do not rule out anything when observing data analysis results. Do not ignore suspicious (good or bad) findings and quickly move on. Be open to surprises. Even when insights emerge at one level, it is important to sliced and dice the data at other levels to see if more powerful insights can be extracted.

***Mistake #5*:** Incompatibility of data mining tools: All the tools from data gathering, preparation, mining, and visualization should work together.

***Mistake #6*:** Locked in the data jailhouse: Use tools that can work with data from multiple sources in multiple industry standard formats.

***Mistake #7*:** Looking only at aggregated results and not at individual records/predictions. It is possible that the right results at the aggregate level provide absurd conclusions at an individual record level.

***Mistake #8*:** Running out of time: Not leaving sufficient time for data acquisition, selection, and preparation can lead to data quality issues and GIGO. Similarly not providing enough time for testing the model, training the users and deploying the system can make the project a failure.

***Mistake #9*:** Measuring your results differently from the way your sponsor measures them: This comes from losing a sense of business objectives and beginning to mine data for its own sake.

***Mistake #10*:** Naively believing everything you are told about the data: Also naively believing everything you are told about your own data mining analysis.

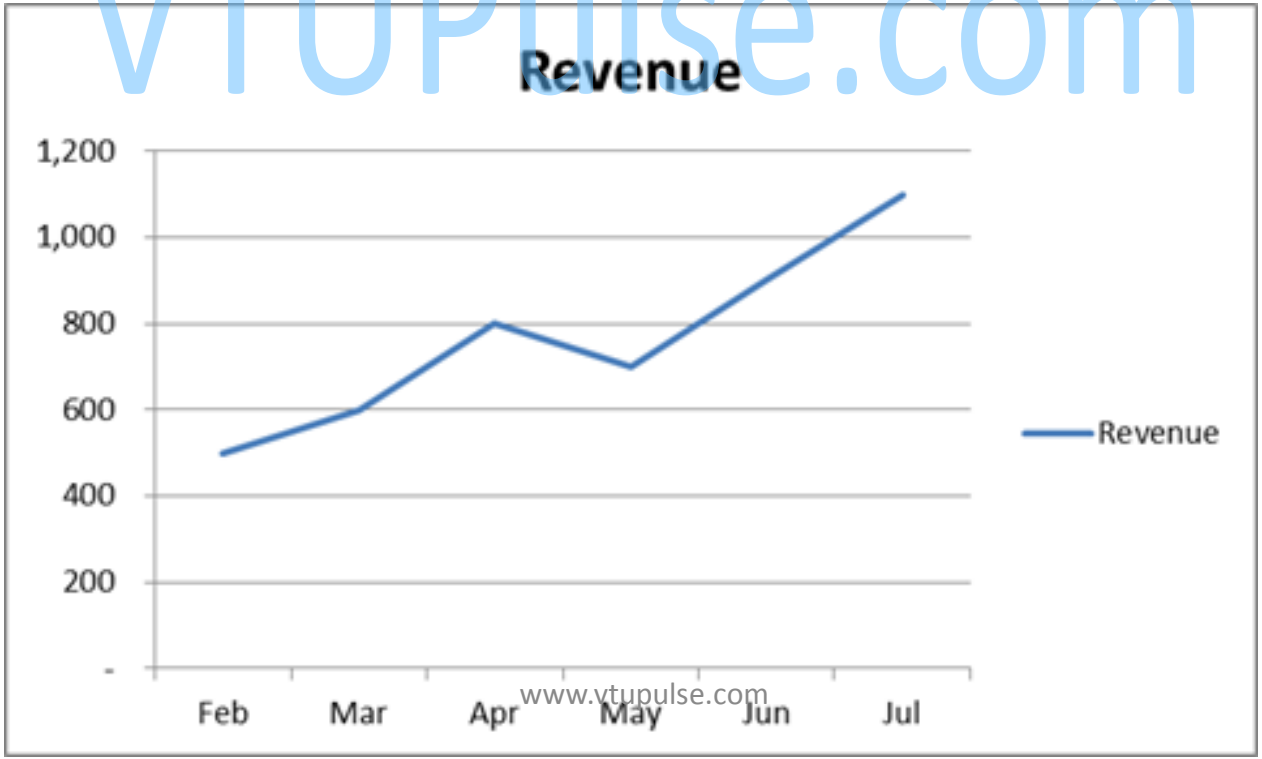
Data Visualization

Objectives for graphical excellence (Data Visualization)

1. *Show, and even reveal, the data*: The data should tell a story, especially story hidden in large masses of data. However, reveal the data in context, so the story is correctly told.
2. *Induce the viewer to think of the substance of the data*: The format of the graph should be so natural to the data, that it hides itself and lets data shine.
3. *Avoid distorting what the data have to say*: Statistics can be used to lie. In the name of simplifying, some crucial context could be removed leading to distorted communication.
4. *Make large data sets coherent*: By giving shape to data, visualizations can help bring the data together to tell a comprehensive story.
5. *Encourage the eyes to compare different pieces of data*: Organize the chart in ways the eyes would naturally move to derive insights from the graph.
6. *Reveal the data at several levels of detail*: Graphs leads to insights, which raise further curiosity, and thus presentations should help get to the root cause.

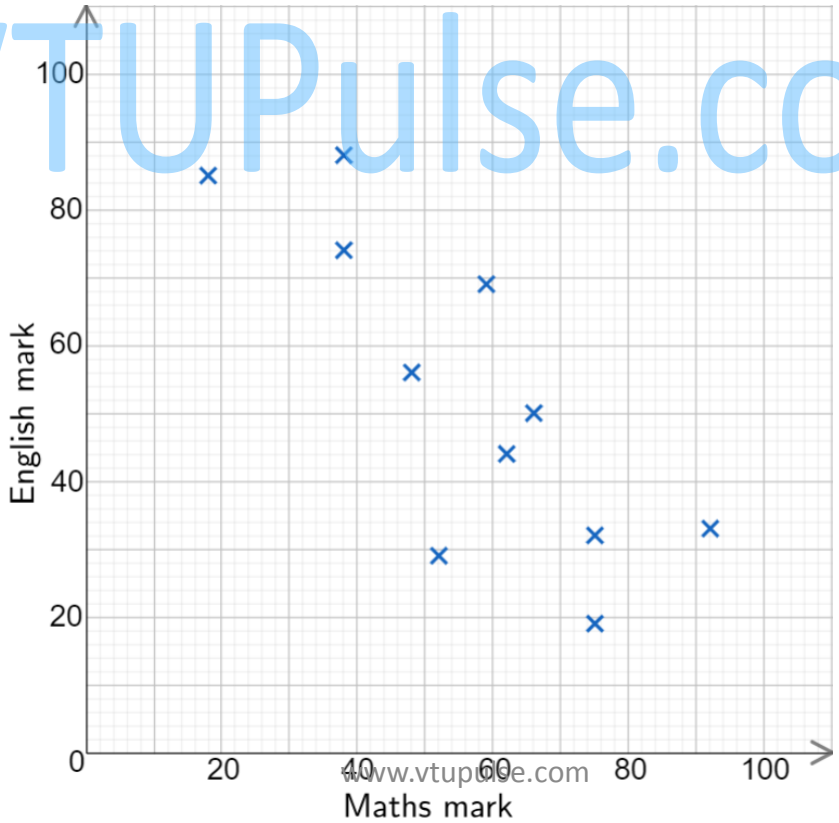
Types of Charts

- *Line graph*. This is a basic and most popular type of displaying information. It shows data as a series of points connected by straight line segments. If mining with time-series data, time is usually shown on the x-axis. Multiple variables can be represented on the same scale on y-axis to compare of the line graphs of all the variables.



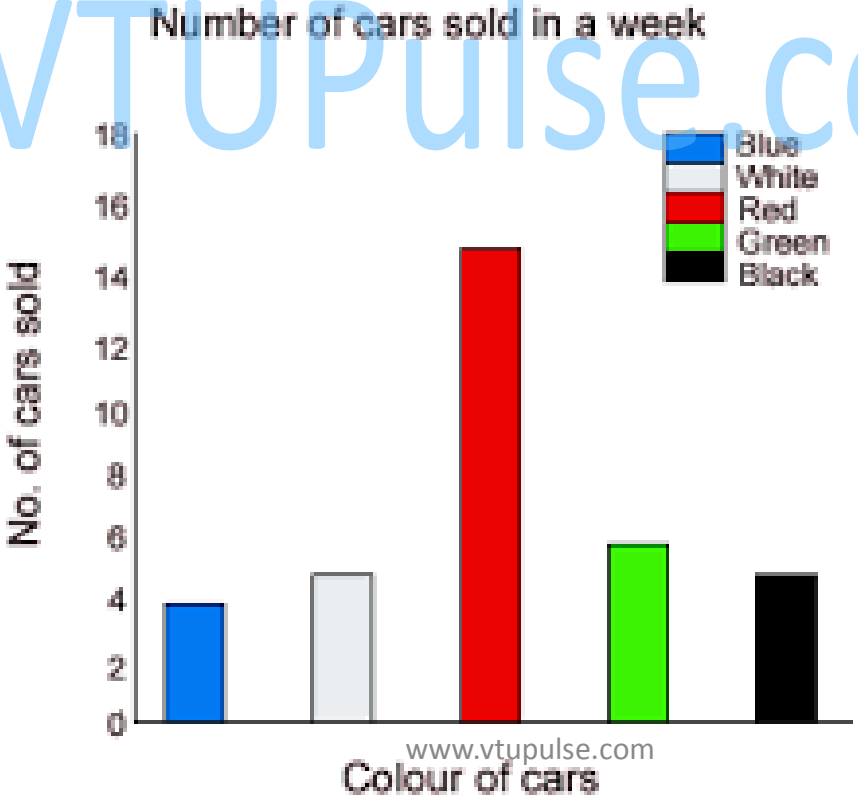
Types of Charts

- *Scatter plot:* This is another very basic and useful graphic form. It helps reveal the relationship between two variables. In the above caselet, it shows two dimensions: Life Expectancy and Fertility Rate. Unlike in a line graph, there are no line segments connecting the points.



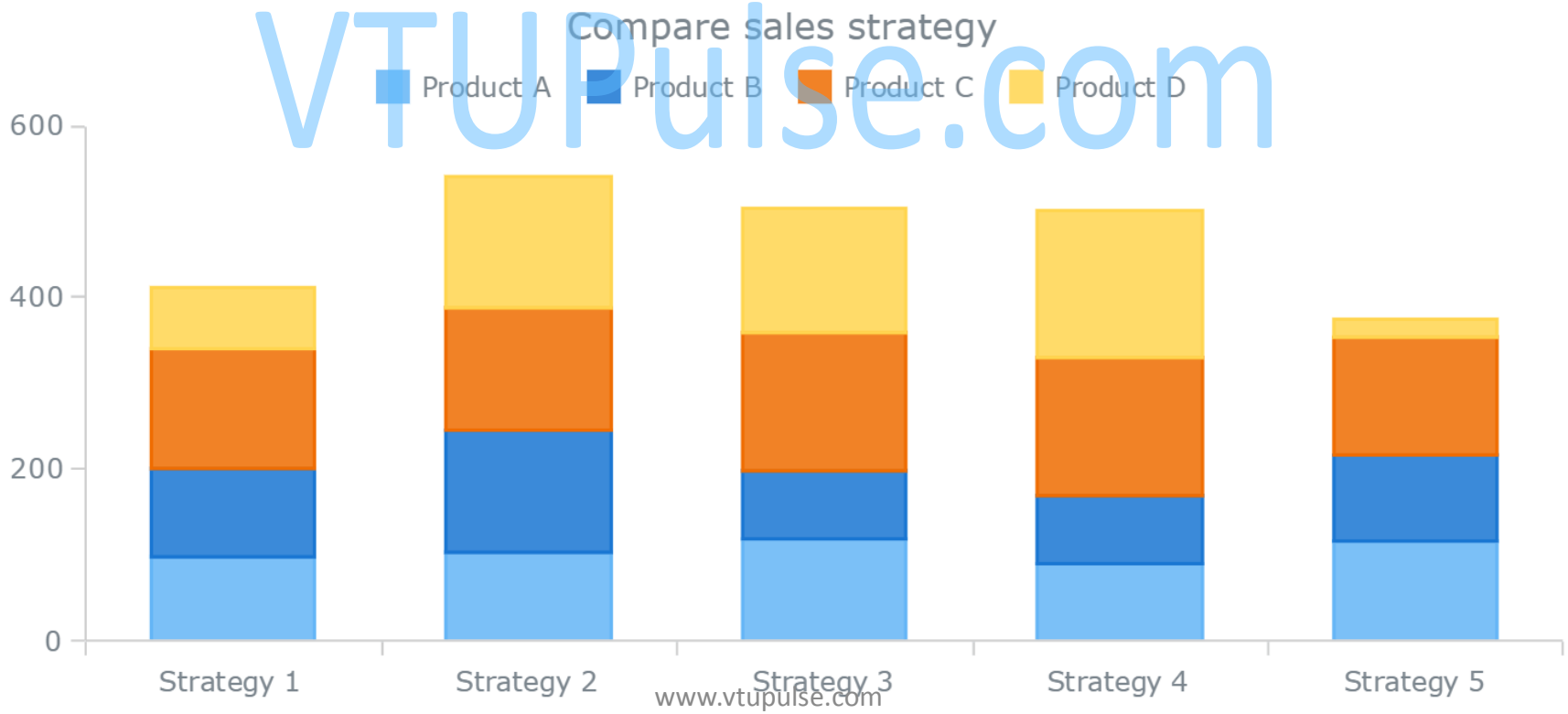
Types of Charts

- *Bar graph:* A bar graph shows thin colorful rectangular bars with their lengths being proportional to the values represented. The bars can be plotted vertically or horizontally. The bar graphs use a lot of more ink than the line graph and should be used when line graphs are inadequate.



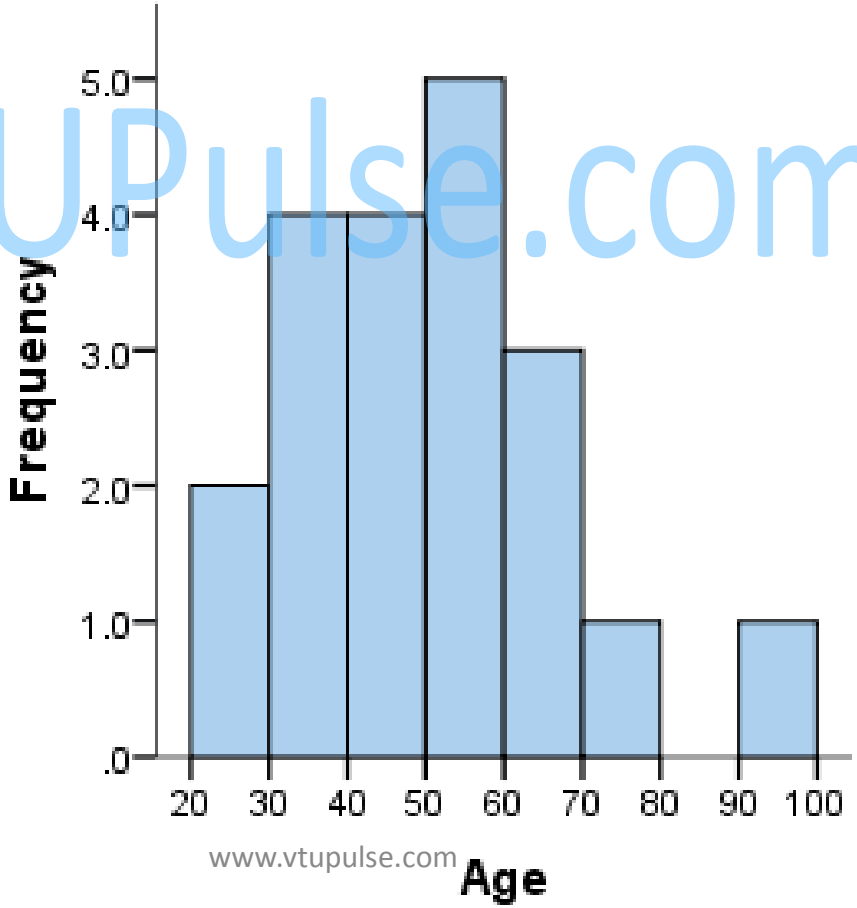
Types of Charts

- *Stacked Bar graphs:* These are a particular method of doing bar graphs. Values of multiple variables are stacked one on top of the other to tell an interesting story. Bars can also be normalized such as the total height of every bar is equal, so it can show the relative composition of each bar.



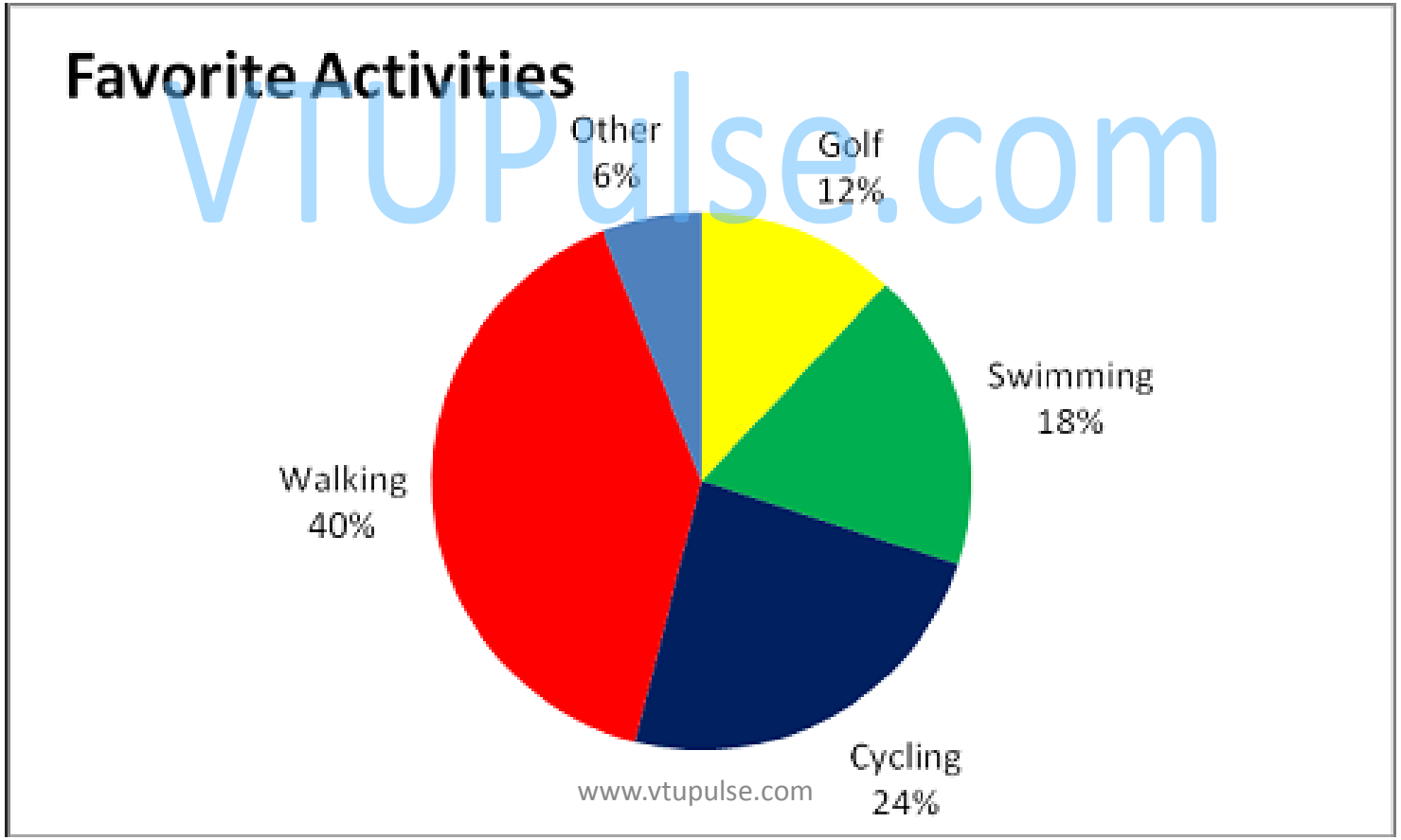
Types of Charts

- *Histograms*: These are like bar graphs, except that they are useful in showing data frequencies or data values on classes (or ranges) of a numerical variable.



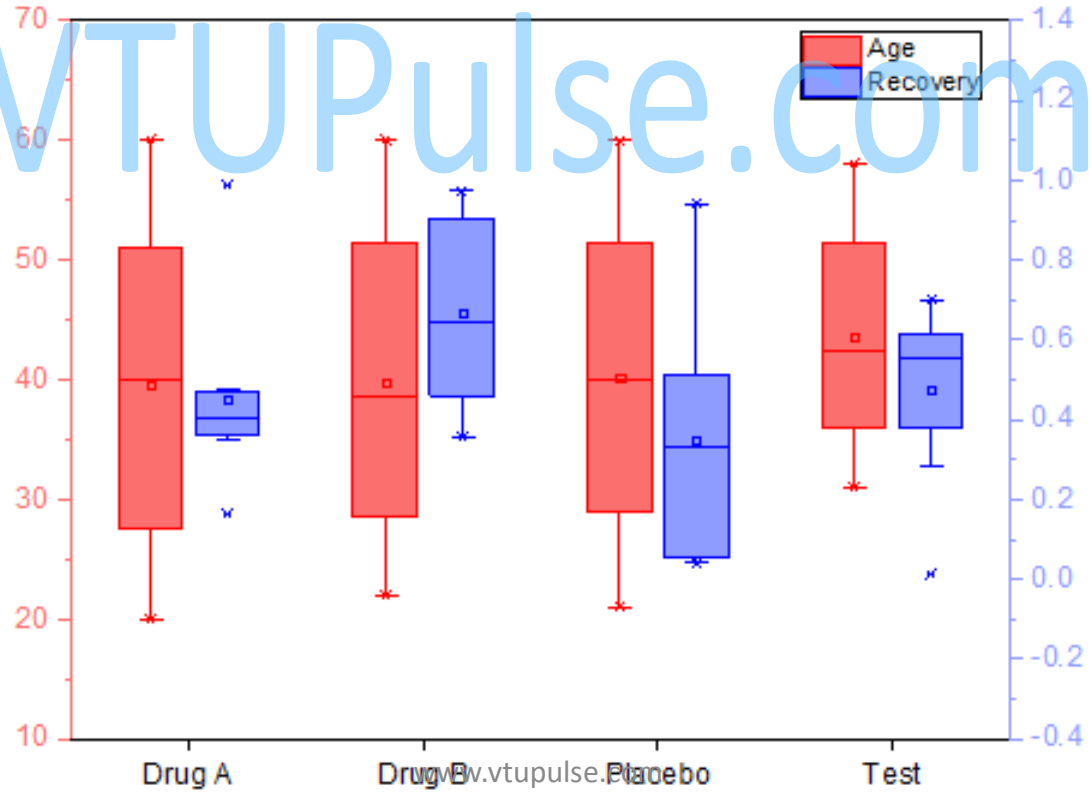
Types of Charts

- *Pie charts*: These are very popular to show the distribution of a variable, such as sales by region. The size of a slice is representative of the relative strengths of each value.



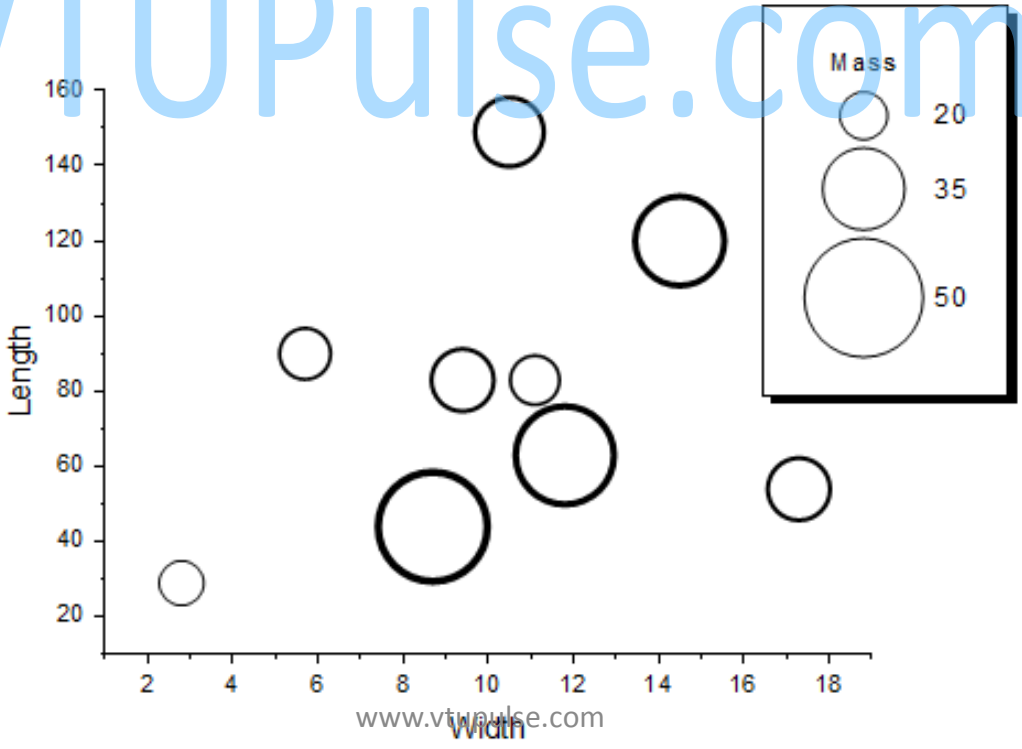
Types of Charts

- *Box charts:* These are special form of charts to show the distribution of variables. The box shows the middle half of the values, while whiskers on both sides extend to the extreme values in either direction.



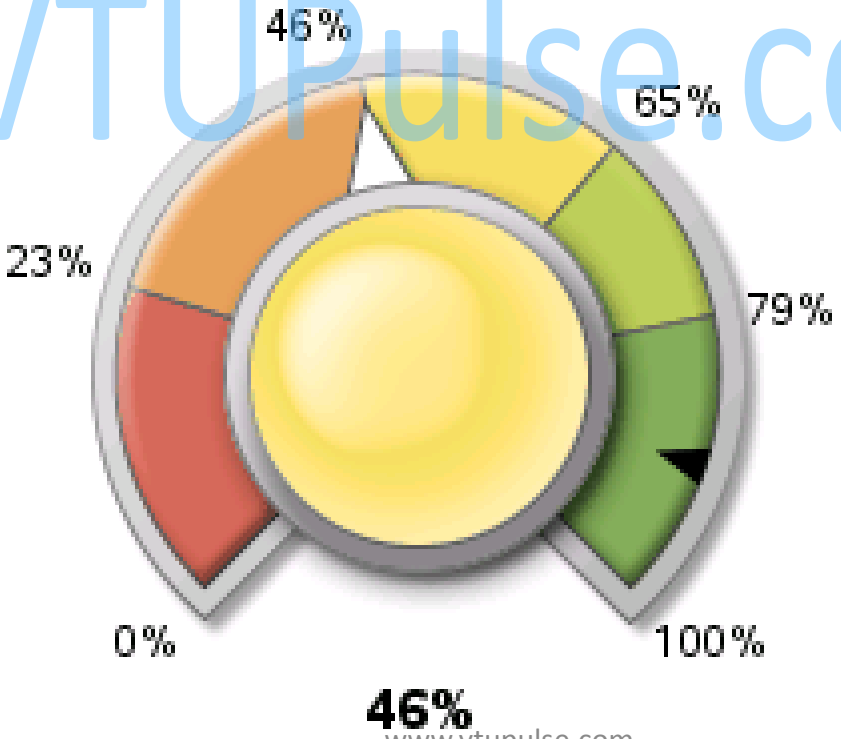
Types of Charts

- *Bubble Graph*: This is an interesting way of displaying multiple dimensions in one chart. It is a variant of a scatter plot with many data points marked on two dimensions. Now imagine that each data point on the graph is a bubble (or a circle) … the size of the circle and the color fill in the circle could represent two additional dimensions.



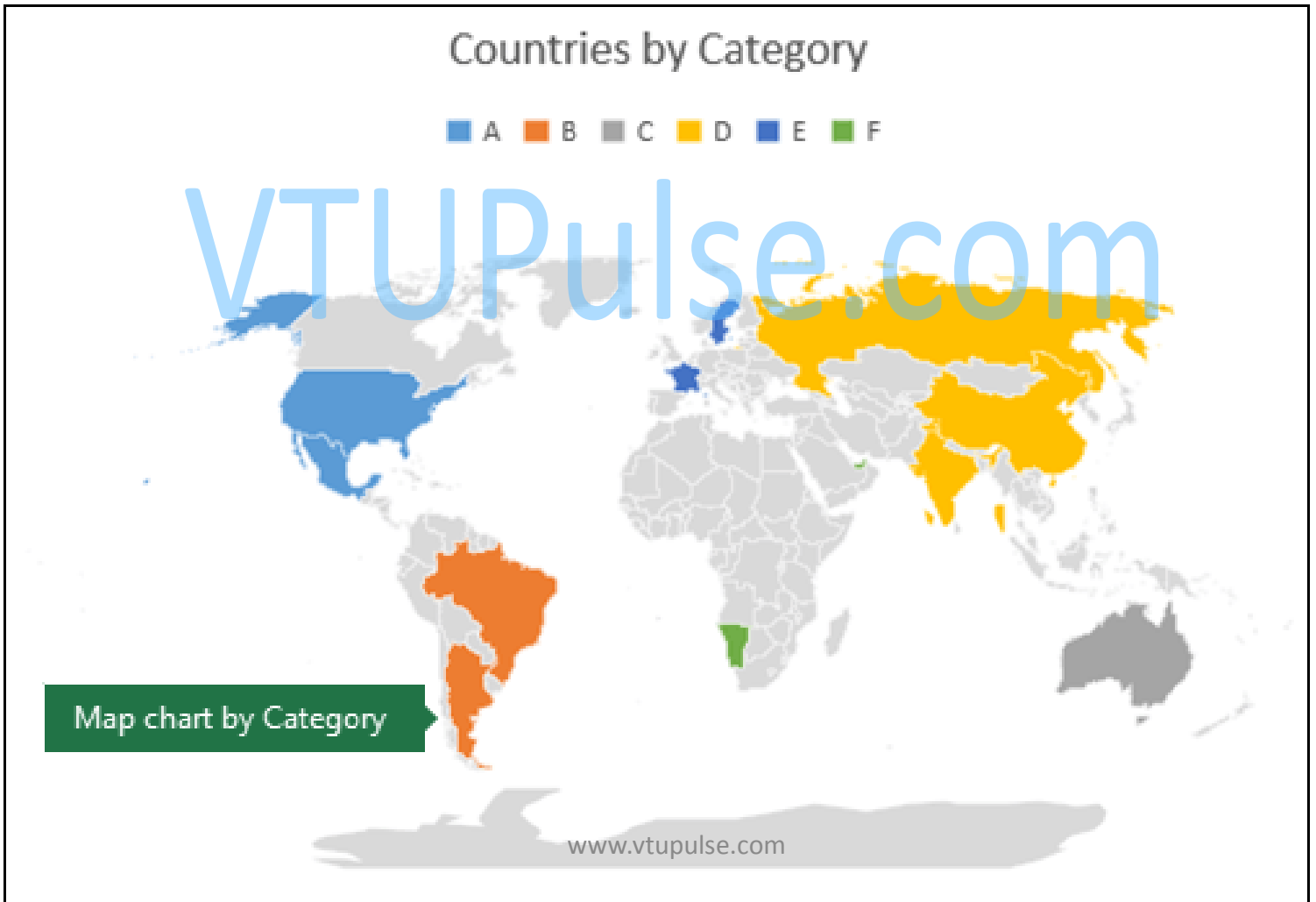
Types of Charts

- *Dials*: These are charts like the speed dial in the car, that shows whether the variable value (such as sales number) is in the low range, medium range, or high range. These ranges could be colored red, yellow and gray to give an instant view of the data.



Types of Charts

- *Geographical* Data maps are particularly useful maps to denote statistics.
-



Types of Charts