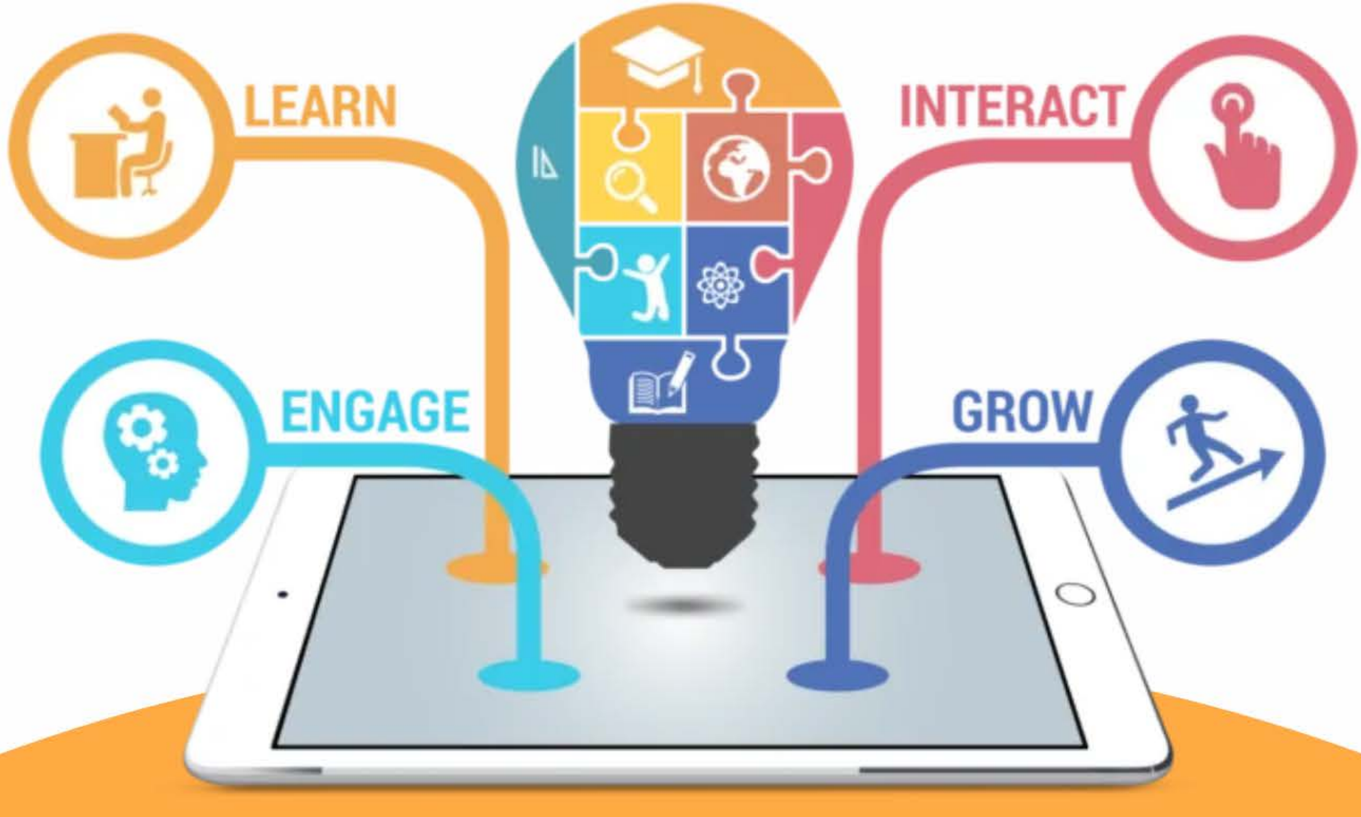
BSC-IT SEM 6

BUSINESS INTELLIGENCE



With lots of efforts, research, reviews we have launched the prerecorded series of academics for multiple universities, bundled in userfriendly application "The Shikshak"

The Shikshak EdTech

Students should do topic-wise study rather than question-wise study for several reasons:

1. **Comprehensive understanding**: Topic-wise study allows students to have a thorough understanding of a particular topic. It helps in building a strong foundation of knowledge on a subject. Once they have a good understanding of a particular topic, they can answer any question related to it.
2. **Efficient use of time**: When students study topic-wise, they can cover a range of questions related to a particular topic in one go. This way, they can utilize their time more efficiently instead of jumping from one question to another and losing focus.
3. **Better retention**: Studying a topic in-depth helps students retain the information for a longer time. It is because they learn the concepts in a logical sequence, making it easier for them to remember.
4. **Effective exam preparation**: Most exams are organized based on topics or units, so studying topic-wise will enable students to be well-prepared for the exam. They will have a good grasp of all the topics that will appear on the exam.
5. **Build analytical skills**: When students study topic-wise, they develop their analytical skills by understanding how various concepts in a subject connect with each other. This helps them develop a deeper understanding of the subject, making them better problem solvers.

In conclusion, studying topic-wise is more beneficial for students as it enables them to develop a better understanding of a subject, retain information better, utilize their time more efficiently, and be well-prepared for exams.

TheShikshak Edu App is an online learning platform that offers a range of resources and tools to help students pursuing BScIT and BScCS programs. Here are some ways in which TheShikshak Edu App can benefit BScIT and BScCS students:

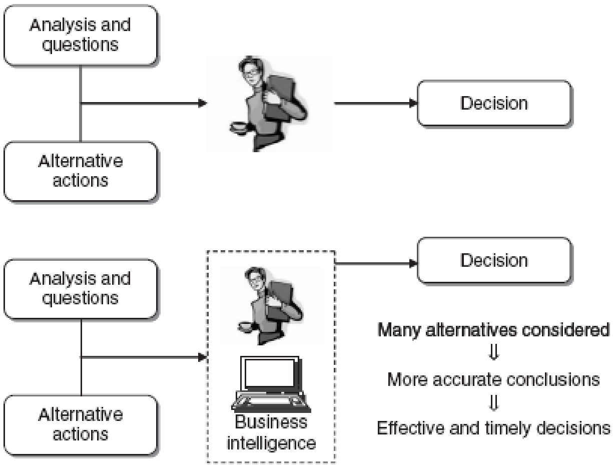
1. **Comprehensive course material**: TheShikshak Edu App offers comprehensive course material for BScIT and BScCS students, covering all topics and concepts required in these programs.
2. **Track their progress** : analytics program helps student to know which topics are remaining and which are lowest watched lectures
3. **Expert guidance**: TheShikshak Edu App has a team of experienced instructors who provide expert guidance and support to students. Students can get their doubts clarified and receive personalized feedback on their performance.

UNIT 1 : CHAPTER 1 : Business intelligence

- Business intelligence may be 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**.

Effective and timely decisions

- In complex organizations, public or private, **decisions are made on a continual basis**.
- The ability of these *knowledge workers* to make decisions, both as individuals and as a community, is one of the primary factors that influence the **performance and competitive strength** of a given organization.
- require a more rigorous **attitude based on analytical methodologies and mathematical models**
- **Retention in the mobile phone industry**
  - **low customer loyalty**, also known as customer *attrition* or *churn*
  - **rely on a budget adequate** to pursue a customer retention
  - **choosing those customers** to be contacted so as **to optimize** the **effectiveness** of the campaign
  - **target the best group of customers** and thus reduce churning and maximize customer retention
- 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 decisions.**
  - rigorous analytical methods allows decision makers to rely on information and knowledge
  - ensuing in-depth examination and thought lead to a deeper awareness and comprehension of the underlying logic of the decision-making process
- **Timely decisions.**
  - If decision makers can rely on a business intelligence system facilitating their activity, we can expect that the overall quality of the decision-making process will be greatly improved.



Benefits of a business intelligence system

  - 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.**

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. (eg sales receipt , sales items, and transaction)
- **Information**. Information 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. (monthly sales, bonus declared)
- **Knowledge**. Information is transformed into knowledge when it is used to make decisions and develop the corresponding actions. (sales analysis)

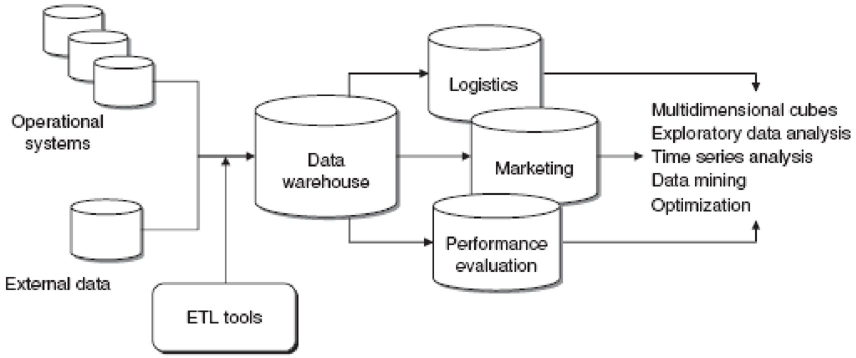
The role of mathematical models

- A business intelligence system provides decision makers with information and knowledge extracted from data, through the application of mathematical models and algorithms.
- rational approach typical of a business intelligence analysis can be summarized schematically in the following main characteristics
  - **First**, the objectives of the analysis are identified and the performance indicators that will be used to evaluate alternative options are defined.
  - **Mathematical models** are then developed by exploiting the relationships among system control variables, parameters and evaluation metrics.
  - **Finally, *what-if* analyses** are carried out to evaluate the effects on the performance determined by variations in the control variables and changes in the parameters.

Business intelligence architectures

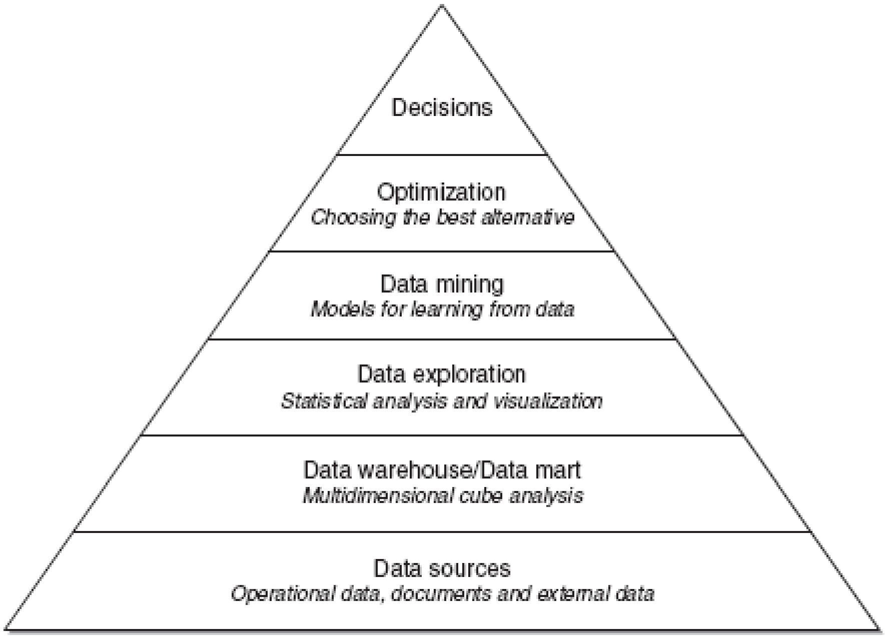
business intelligence system includes three major components.

- **Data sources.** In a first stage, it is necessary to gather and integrate the data stored in the various primary and secondary sources, which are heterogeneous in origin and type.
- **Data warehouses and data marts**. Using extraction and transformation tools known as extract, transform, load (ETL), the data originating from the different sources are stored in databases intended to support business intelligence analyses.
- **Business intelligence methodologies.** Data are finally extracted and used to feed mathematical models and analysis methodologies intended to support decision makers.
- **In a business intelligence system, several decision support applications may be implemented.**
  - multidimensional cube analysis
  - exploratory data analysis
  - time series analysis
  - inductive learning models for data mining
  - optimization models



A typical business intelligence architecture

- **The main components of a business intelligence system**

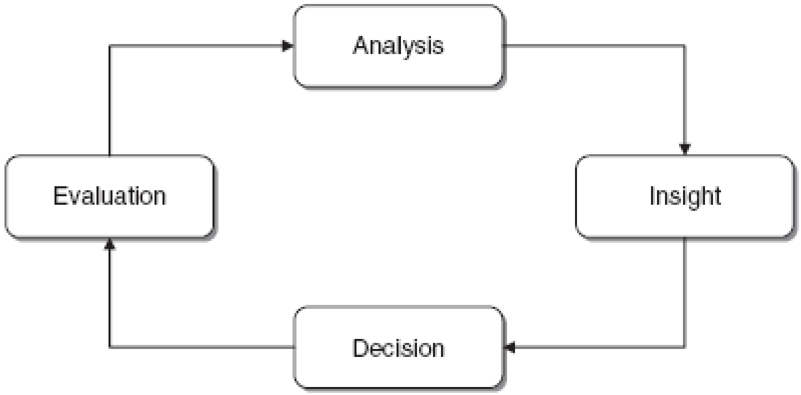

Decisions
Optimization
Choosing the best alternative
Data mining
Models for learning from data
Data exploration
Statistical analysis and visualization
Data warehouse/Data mart
Multidimensional cube analysis
Data sources
Operational data, documents and external data

- **Data exploration**. At the third level of the pyramid we find the tools for performing a passive business intelligence analysis, which **consist of query and reporting systems**, as well as statistical methods.
- **Data mining.** The fourth level includes active business intelligence methodologies, whose purpose is the **extraction of information and knowledge** from data.
- **Optimization.** By moving up one level in the pyramid we find optimization models that allow us to determine **the best solution** out of a set of alternative actions, which is usually fairly extensive and sometimes even infinite.
- **Decisions.** Finally, the top of the pyramid corresponds to the choice and the **actual adoption of a specific decision**, and in some way represents the natural conclusion of the decision-making process.

Cycle of a business intelligence analysis

- business intelligence analysis follows its own path according to the application domain, the personal attitude of the decision makers and the available analytical methodologies.

Cycle of a business intelligence analysis



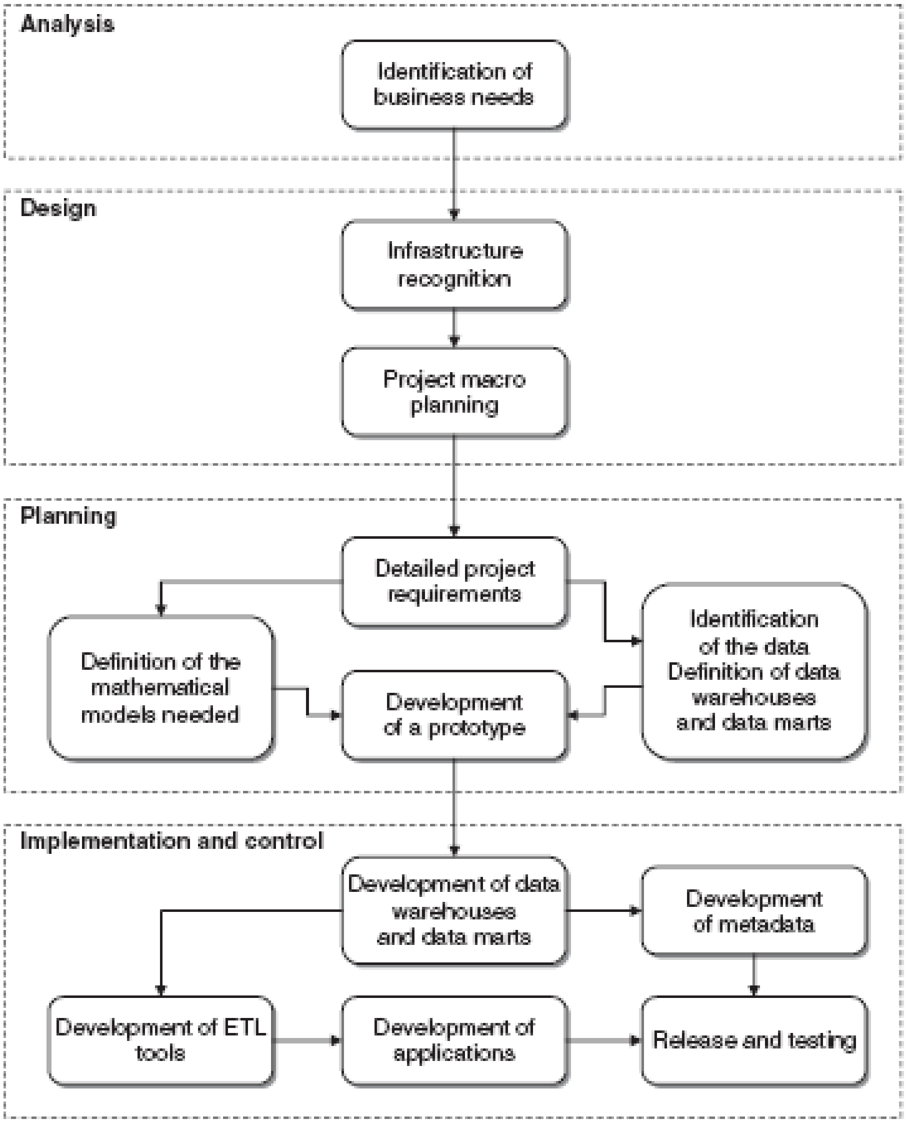
- **Analysis.** During the analysis phase, it is necessary to recognize and accurately spell out the problem at hand.
- **Insight.** The second phase allows decision makers to better and more deeply understand the problem at hand, often at a causal level.
- **Decision.** During the third phase, knowledge obtained as a result of the insight phase is converted into decisions and subsequently into actions.
- **Evaluation.** Finally, the fourth phase of the business intelligence cycle involves performance measurement and evaluation.

Enabling factors in business intelligence projects

- **Technologies.** Hardware and software technologies are significant enabling factors that have facilitated the development of business intelligence systems within enterprises and complex organizations.
- **Analytics**. As stated above, mathematical models and analytical methodologies play a key role in information enhancement and knowledge extraction from the data available inside most organizations.
- **Human resources**. The human assets of an organization are built up by the competencies of those who operate within its boundaries, whether as individuals or collectively. The overall knowledge possessed and shared by these individuals constitutes the organizational culture.

Development of a business intelligence system

- The development of a business intelligence system can be assimilated to a project, with a specific final objective, expected development times and costs, and the usage and coordination of the resources needed to perform planned activities.

- **Analysis**. During the first phase, the needs of the organization relative to the development of a business intelligence system should be carefully identified.
- **Design.** The second phase includes two sub-phases and is aimed at deriving a provisional plan of the overall architecture, taking into account any development in the near future and the evolution of the system in the mid term.
- **Planning.** The planning stage includes a sub-phase where the functions of the business intelligence system are defined and described in greater detail.
- **Implementation and control.** The last phase consists of five main sub-phases. First, the data warehouse and each specific data mart are developed. These represent the information infrastructures that will feed the business intelligence system.
- 
Analysis
Identification of business needs
Design
Infrastructure recognition
Project macro planning
Planning
Detailed project requirements
Definition of the mathematical models needed
Development of a prototype
Identification of the data Definition of data warehouses and data marts
Implementation and control
Development of data warehouses and data marts
Development of metadata
Development of ETL tools
Development of applications
Release and testing

Ethics and business intelligence

- Usage of data by public and private organizations that is improper and does not respect the individuals' right to privacy should not be tolerated.

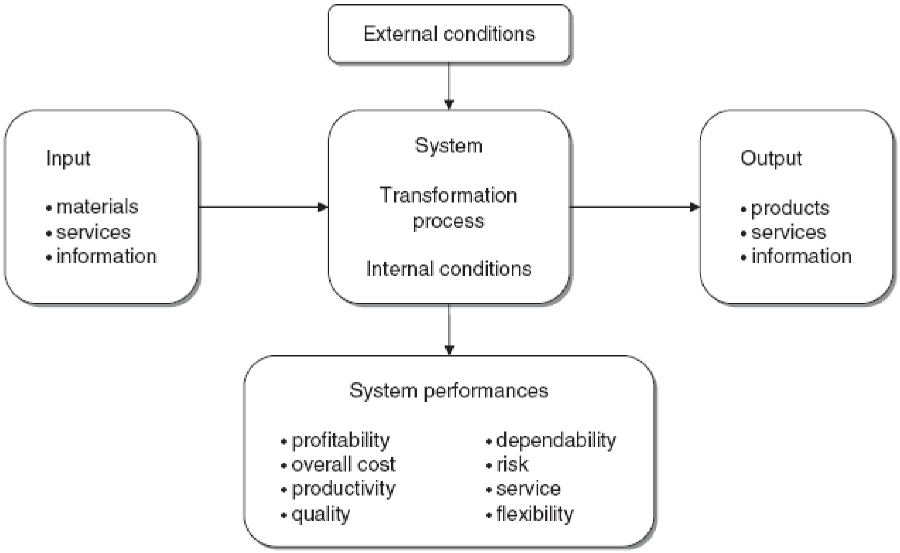
- guard against the excessive growth of the political and economic power of enterprises allowing the transformation processes outlined above to exclusively and unilaterally benefit
- it is essential that business intelligence analysts and decision makers abide by the ethical principle
- analysts developing a mathematical model and those who make the decisions cannot remain neutral,
- but have the moral obligation to take an ethical stance.

CHAPTER 2 : Decision support systems

A **decision support system (DSS)** is an interactive computer-based application that combines data and mathematical models to help **decision makers solve complex problems faced** in managing the public and private enterprises and organizations.

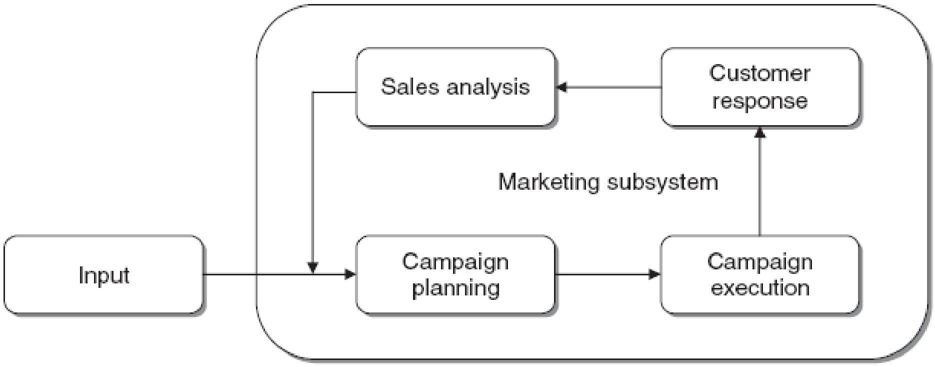
Definition of system

- The entities that we intuitively denominate *systems* share a common characteristic, which we will adopt as an abstract definition of the notion of system: each of them is **made up of a set of components that are in some way connected to each other** so as to provide a single collective result and a common purpose.
- A system receives a set **of *input* flows** and returns a set **of *output* flows** through **a *transformation*** process regulated by ***internal conditions* and *external conditions***. The effectiveness and efficiency of a system are assessed using measurable performance indicators that can be classified into different categories.



Abstract representation of a system

- A **system will often incorporate a *feedback* mechanism**. Feedback occurs when a system component generates an output flow that is fed back into the system itself as an input flow, possibly as a result of a further transformation.
- Systems that are able to **modify their own output flows based on feedback are called *closed cycle systems*.**



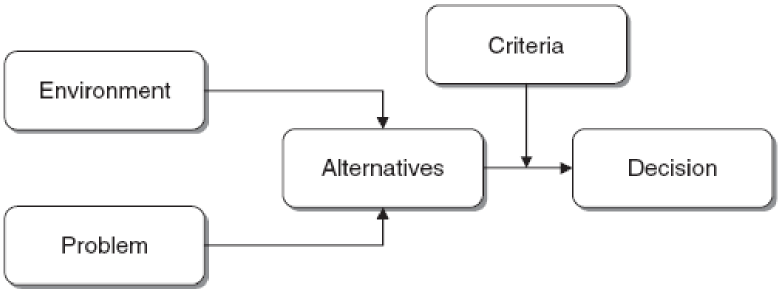
A closed cycle marketing system with feedback effects

- The sales results for each campaign are gathered and become available as feedback input so as to design subsequent marketing promotions.
- **To assess the performance** of a system it is appropriate to **categorize the evaluation metrics into two main classes: *effectiveness* and *efficiency***.
  - **Effectiveness.** Effectiveness measurements express the level of conformity of a given system to the objectives for which it was designed.(EG volumes, weekly sales and yield per share.)
  - **Efficiency.** Efficiency measurements highlight the relationship between input flows used by the system and the corresponding output flows. (EG amount of resources needed to achieve a given sales volume.)

Representation of the decision-making process

Rationality and problem solving

- A decision is a choice from multiple alternatives, usually made with a fair degree of rationality.
- The decision-making process is part of a broader subject usually referred to as problem solving, which refers to the process through which individuals try to **bridge the gap between the current operating conditions of a system (as is) and the supposedly better conditions to be achieved in the future (to be).**
- The **alternatives** represent the **possible actions aimed** at solving the given problem and helping to achieve the planned objective.
- ***Criteria*** are the measurements of effectiveness of the **various alternatives** and correspond to the different kinds of system performance



Logical flow of a problem-solving process

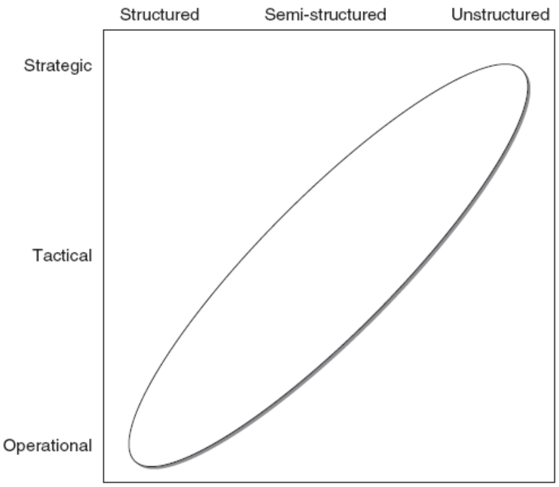
- **A *rational* approach** to decision making implies that the option fulfilling the best performance criteria is selected out of all possible alternatives.
- **factors influencing a rational choice.**
  - Economic(minimization of costs),
  - Technical(actual problems),
  - Legal(compatible with the legislation),
  - Ethical(should abide by the ethical principles and social rules),
  - Procedural(must follow standard operating procedure),
  - Political(political consequences of a specific decision)

The decision-making process

- The model includes three phases, termed intelligence, design, choice. extended phases implementation and control
- **Intelligence.** In the intelligence phase the task of the decision maker is to identify, circumscribe and explicitly define the problem that emerges in the system under study. Is a action to improve the system performance
- **Design.** In the design phase actions aimed at solving the identified problem should be developed and planned.
- **Choice.** Once the alternative actions have been identified, it is necessary to evaluate them on the basis of the performance criteria deemed significant.
- **Implementation.** When the best alternative has been selected by the decision maker, it is transformed into actions by means of an implementation plan.
- **Control.** Once the action has been implemented, it is finally necessary to verify and check that the original expectations have been satisfied and the effects of the action match the original intentions.

Types of decisions

- Decisions can be classified in terms of two main dimensions, according to their ***nature* and *scope*.**
- Each dimension will be subdivided into three classes, giving a total of **nine possible combinations**
  - **According to their nature**, decisions can be classified as ***structured*, *unstructured* or *semi-structured*.**



- **Structured decisions**. A decision is structured if it is based on **a well-defined** and recurring decision-making procedure.

  - **Unstructured decisions.** least **one element in the system** (input flows, output flows and the transformation processes) **that cannot be described in detail** and reduced to a predefined sequence of steps.
  - **Semi-structured decisions.** A decision is semi-structured when **some phases are structured and others are not.**
- **Depending on their scope**, decisions can be classified as ***strategic***, ***tactical*** **and** ***operational***.
  - Decisions are **strategic** when they **affect the entire organization** or at least a substantial part of it for a long period of time.
  - **Tactical decisions affect only parts** of an enterprise and are usually restricted to a single department.
  - **Operational decisions.** Operational decisions refer to specific activities carried out within an organization and **have a modest impact on the future**.

Approaches to the decision-making process

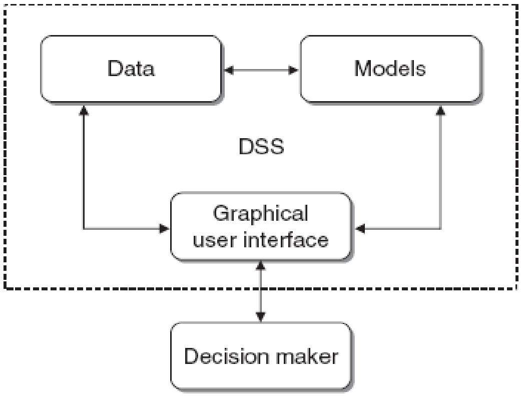
- A preliminary distinction should be made between a rational approach and a political-organizational approach.
- **Rational approach.** When a rational approach is followed, a decision maker considers major factors, such as economic, technical, legal, ethical, procedural and political.
- **Political-organizational approach.** Decisions are not based on clearly defined alternatives and selection criteria. DSS can only help in a passive way, providing timely and versatile access to information.
- **Rational approach we can further distinguish between two alternative ways** in which the actual decision-making process influences decisions: **absolute rationality and bounded rationality.**
  - **Absolute rationality**. The term 'absolute rationality' refers to a decision-making process for which multiple performance indicators can be reduced to a single criterion, which therefore naturally lends itself to an optimization model.
  - **Bounded rationality**. Bounded rationality occurs whenever it is not possible to meaningfully reduce multiple criteria into a single objective, so that the decision maker considers an option to be satisfactory when the corresponding performance indicators fall above or below prefixed threshold values.

Evolution of information systems

- **data processing**.
- **management information systems(MIS),** in order to ease access to useful and timely information for decision makers.
- The increase in independent processing capabilities held by users, usually referred to as **end user computing.**
- **executive information systems and strategic information systems**.
- client–server computing brought the concepts of **data warehouses and data marts**.
- **business intelligence.**

Definition of decision support system

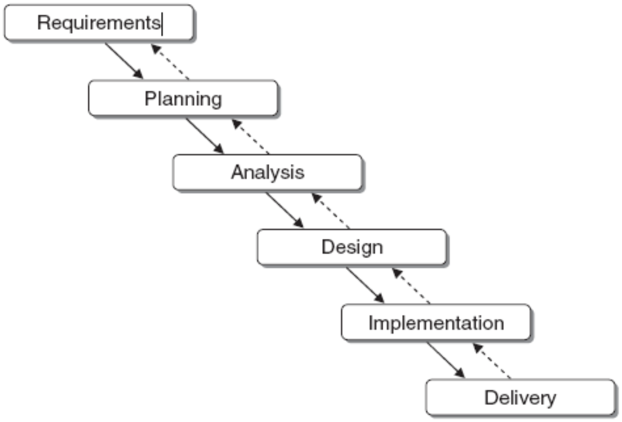
- **Three main elements of a DSS**
- **database**, a repository of **mathematical models** and a **module for handling the dialogue between the system and the users.**



Structure of a decision support system

- It thus highlights the role of DSSs as the **focal point of evolution trends in two distinct areas**:
  - on the one hand, **data processing and information technologies**; and
  - on the other hand, the disciplines addressing the **study of mathematical models** and methods, such as **operations research and statistics**.
- **Relevant features of a DSS**
  - **Effectiveness** (to reach more effective decisions.)
  - **Mathematical models** (transform data into knowledge and provide active support)
  - **Integration in the decision-making process** (adapting needs)
  - **Organizational role** (communication between the various parts)
  - **Flexibility** (flexible and adaptable in order to incorporate the changes)
  - **Data management** (data management module of a DSS is usually connected with a company data warehouse)
  - **Model management** (collection of mathematical models)
  - **Interactions** (receive input data and return the extracted information and the knowledge)
  - **Knowledge management** (interconnected with the company knowledge management)

Development of a decision support system

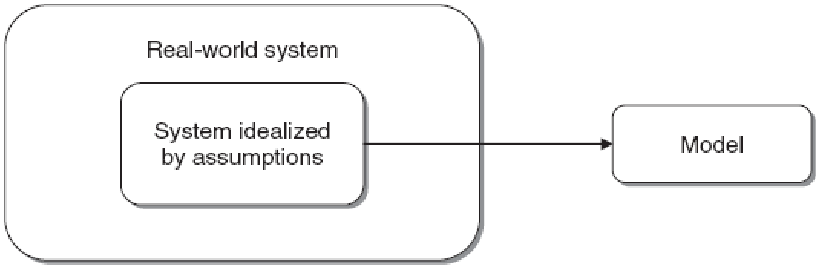


- **Planning** usually involves a feasibility study to address the question: Why do we wish to develop a DSS? During the feasibility analysis, general and specific objectives of the system, recipients, possible benefits, execution times and costs are laid down.
- **Analysis**. In the analysis phase, it is necessary to define in detail the functions of the DSS to be developed, by further developing and elaborating the preliminary conclusions achieved during the feasibility study.
- **Design.** During the design phase the main question is: How will the DSS work? The entire architecture of the system is therefore defined
- **Implementation.** Once the specifications have been laid down, it is time for implementation, testing and the actual installation, when the DSS is rolled out and put to work.
- Integration. The design and development of a DSS require a significant number of methodologies, tools, models, individuals and organizational processes to work in harmony (WORK TOGETHOR)
- **Involvement:** involvement of decision makers and users during the development process is of primary importance to avert their inclination to reject a tool that they perceive as alien.

UNIT 2 : CHAPTER 1 : Mathematical models for decision making

Structure of mathematical models

- Scientific and technological development has turned to mathematical models of various types for the **abstract representation of real systems.**



- According to their characteristics, models can be divided into **iconic, analogical and symbolic.**
  - **Iconic.** An iconic model is imitated for **the purpose of the analysis**
  - **Analogical.** An analogical model imitates the **real behavior**
  - **Symbolic.** A symbolic model It is intended to describe the behavior of the system through a **series of symbolic variables, numerical parameters and mathematical relationships**.
  - The **probabilistic nature of models**, which can be either **stochastic** or **deterministic.**
    - **Stochastic.** In a stochastic model some input information represents **random events**.
    - **Deterministic**. A model is called deterministic when all input data are supposed **to be known a priori and with certainty**
  - **Temporal dimension** in a mathematical model, which can be either **static or dynamic**.
    - **Static.** Static models determine an **optimal plan** for the distribution of goods in a specific time frame.
    - **Dynamic.** Dynamic models consider a given system through several temporal stages, corresponding to a **sequence of decisions**.

Development of a model

- It is possible to break down the development of a mathematical model for decision making into **four primary phases**.
- **Includes a feedback mechanism** which takes into account the possibility of changes and revisions of the model.
    - **Problem identification**
        - The observed critical symptoms must be analyzed and interpreted in order to formulate hypotheses for investigation
    - **Model formulation**
        - defining an **appropriate** mathematical **model** to represent the system**. number of factors affect and influence the choice of model.**
            - **Time horizon**.(specify the production rate for each week in a year)
            - **Evaluation criteria**.(Appropriate measurable performance indicators, evaluation and comparison of the alternative decisions) eg quality vs service, flexibility vs condition
            - **Decision variables**.(decision variables should express production volumes for each product, for each process and for each period of the planning horizon.)
            - **Numerical parameters**(available capacity should be known in advance for each process)
    - **Development of algorithms**
        - Once a mathematical model has been defined, one will naturally wish to proceed with its solution to assess decisions and to select the best alternative.
    - **Implementation and test**
        - When a model is fully developed, then it is finally implemented, tested and utilized in the application domain.

Classes of models

main categories of mathematical models for decision making

- predictive models;
- pattern recognition and learning models;
- optimization models;
- project management models;
- risk analysis models;
- waiting line models.

- **predictive models**
    - The results of random events determine the future demand for a product or service, the development of new scenarios of technological innovation and the level of prices and costs.
    - predictive models play a primary role in business intelligence systems
    - *Predictions allow input information to be fed into different decision-making processes, arising in strategy, research and development, administration and control, marketing, production and logistics.*
    - Predictive models can be subdivided into **two main categories**.

  - **Regression models,**
  - **classification models**,

- **Pattern recognition and machine learning models**
  - **ability to extract knowledge** from past experience in order to apply it in the future.
  - **just like the human mind** is able to do with great effectiveness due to the sophisticated mechanisms developed and fine-tuned in the course of evolution.
  - Besides an intrinsic theoretical interest, mathematical methods for learning are applied in several domains, such as **recognition of images, sounds and texts**
  - **Mathematical models for learning have two primary objectives.**
    - **The purpose of interpretation models** is **to identify regular patterns** in the data and to express them through easily understandable rules and criteria.
    - **Prediction models help to forecast** the value that a given random variable will assume in the future, based on the values of some variables associated with the entities of a database

- **Optimization models**
  - optimization models arise naturally in decision-making processes where a **set of limited resources must be allocated in the most effective way to different entities**.
  - **domains requiring an optimal allocation of the resources**
    - logistics and production planning;
    - financial planning;

- **Project management models**
  - Project management methods are based on the *contributions of various disciplines*, such as *business organization, behavioral psychology and operations research*.
  - Mathematical models for decision making play an important role in project management methods.
  - project evaluation and review techniques *(PERT)*, are used to derive the execution times when stochastic assumptions are made regarding the duration of the activities, represented by random variables.

- **Risk analysis models**
  - The decision maker is required to choose among a *number of available alternatives*, having uncertain information regarding the effects that these options may have in the future.
  - The decision maker is forced to make a choice before knowing with absolute certainty the level of future demand

- **Waiting line models**
  - If the arrival times of the customers and the duration of the service are not known beforehand in a deterministic way, conflicts may arise between customers in the use of limited shared resources. Consequently, some customers are forced to wait in a line.
  - The main components of a *waiting line system are the population, the arrivals process, the service process, the number of stations, and the waiting line rules*.

CHAPTER 2 : Data mining

the term *data mining* indicates the **process of exploration and analysis** of a dataset, usually of large size, in order to find regular patterns, to extract relevant knowledge and to obtain meaningful recurring rules.

Definition of data mining

- Data mining activities constitute an iterative process aimed at the analysis of large databases, with the **purpose of extracting information and knowledge** that may prove accurate and potentially **useful for knowledge workers engaged in decision making and problem solving.**
- The term *data mining* refers therefore to the overall **process consisting of data gathering and analysis**, development of inductive learning models and adoption of practical decisions and consequent actions based on the knowledge acquired.
- **Data mining activities can be subdivided into two major investigation Streams** *interpretation* **and** *prediction*.
  - **Interpretation.** The purpose of interpretation is to identify regular patterns in the data and to express them through rules and criteria that can be easily understood by experts in the application domain.
  - **Prediction.** The purpose of prediction is to anticipate the value that a random variable will assume in the future or to estimate the likelihood of future events.

- **Models and methods for data mining**
  - There are several learning methods that are available to perform the different data mining tasks. A number of techniques originated in the field of computer science, such as **classification trees or association rules**, and are referred to as *machine learning* or *knowledge discovery in databases*.
- **Data mining, classical statistics and OLAP**

OLAP	statistics	data mining
extraction of details and aggregate totals from data information distribution of incomes of home loan applicants	verification of hypotheses formulated by analysts validation analysis of variance of incomes of home loan applicants	identification of patterns and recurrences in data knowledge characterization of home loan applicants and prediction of future applicants

- **Applications of data mining**
  - Data mining methodologies can be applied to a variety of domains, from marketing and manufacturing process control to the study of risk factors in medical diagnosis, from the evaluation of the effectiveness of new drugs to fraud detection.
  - **Relational marketing (**identification of customer segments, prediction of the rate of positive responses)
  - **Fraud detection**(illegal use of credit cards and bank checks)
  - **Risk evaluation.(** estimate the risk connected with future decisions)
  - **Text mining(**represent unstructured data, in order to classify articles, books, documents)
  - **Image recognition.** The treatment and classification of digital images It is useful to recognize written characters, compare and identify human faces, apply correction filters to photographic equipment and detect suspicious behaviors through surveillance video cameras.

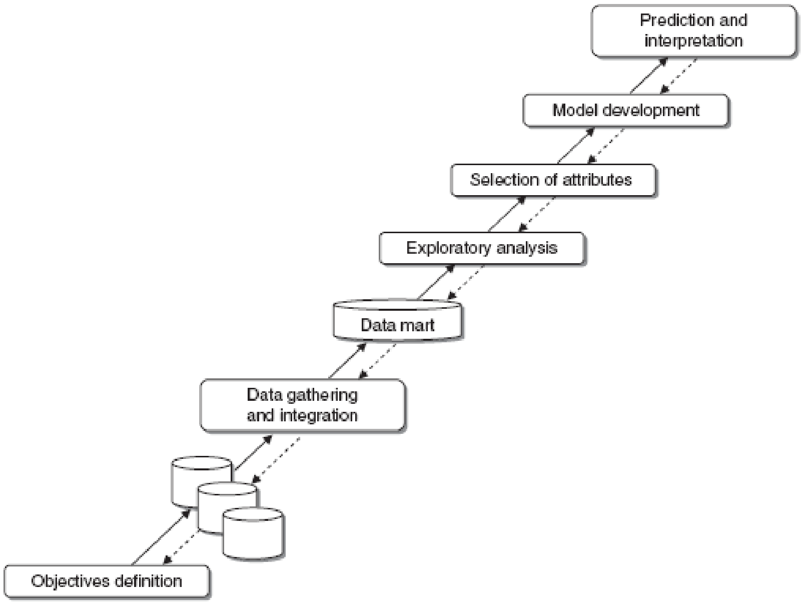
- **Web mining.** intended for the analysis of so-called *clickstreams* – the sequences of pages visited and the choices made by a web surfer
- **Medical diagnosis.** Learning models are an invaluable tool within the medical field for the early detection of diseases using clinical test results.

Representation of input data

- input to a data mining analysis takes the form of a **two dimensional table, called a *dataset***, irrespective of the actual logic and material representation adopted to store the information in files, databases, data warehouses and data marts used as data sources.
- The **rows** in the dataset **correspond to the *observations*** recorded in the past and are also called *examples*, *cases*, *instances* or *records*.
- The **columns** represent the **information available for each observation** and are termed *attributes*, *variables*, *characteristics* or *features*.
- Attributes contained in a dataset can be categorized as ***categorical* or *numerical***
  - **Categorical.** Categorical attributes assume a finite number of distinct values (product name, location names)
  - **Numerical.** Numerical attributes assume a finite or infinite number of values and lend themselves to subtraction or division operations(price, commission)
- **Taxonomy of attributes**
  - **Counts.** Counts are categorical attributes in relation to which a specific property can be true or false
  - **Nominal.** Nominal attributes are categorical attributes without a natural ordering, such as the **province of residence.**
  - **Ordinal.** Ordinal attributes, such as education level, are categorical attributes that lend themselves to a natural ordering
  - **Discrete.** Discrete attributes are numerical attributes that assume a finite number or a countable infinity of values(not same every time)
  - **Continuous.** Continuous attributes are numerical attributes that assume an uncountable infinity of values.

Data mining process

Definition of objectives. Data mining analyses are carried out in specific application domains and are intended to provide decision makers with useful knowledge.



Data mining process

- **Data gathering and integration.** Once the objectives of the investigation have been identified, the gathering of data begins. Data may come from different sources and therefore may require integration.
- **Exploratory analysis.** In the third phase of the data mining process, a preliminary analysis of the data is carried out with the purpose of getting acquainted with the available information and carrying out *data cleansing*.
- **Attribute Selection.** In the subsequent phase, the relevance of the different attributes is evaluated in relation to the goals of the analysis.(selecting appropriate attribute means columns rollno, rank, name, state)
- **Model development and validation.** Once a high quality dataset has been assembled and possibly enriched with newly defined attributes, pattern recognition and predictive models can be developed. (validation against training sets)
- **Prediction and interpretation.** knowledge workers may be able to use it to draw predictions and acquire a more in-depth knowledge of the phenomenon of interest.

Analysis methodologies

draw a first fundamental distinction between *supervised* and *unsupervised* learning processes

- **Supervised learning.** In a supervised **(or *direct*) learning analysis**, a target attribute either represents the class to which each record belongs. (set of rules are defined and observed)
- **Unsupervised learning.** Unsupervised **(or *indirect*) learning analyses** are not guided by a target attribute.
- unsupervised learning analyses, one is interested in identifying *clusters* of records that are similar within each cluster and different from members of other clusters.
- **Seven basic data mining tasks**

  • **characterization and discrimination**;(arrange data according to characterization, and divide as per set of rules, students or professor, divide as per class or subjects)

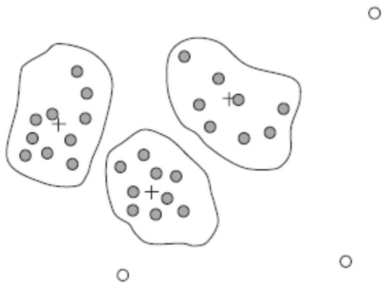
  • **classification**;(Each observation is described by a given number of attributes whose value is known)

• **regression**;(regression is used when the target variable takes on continuous values, predict the sale of product before launch)
• **time series analysis**;(predicting the value of the target variable for one or more future periods, on history of that variable)
• **association rules**;(identify interesting and recurring associations between groups of records of a dataset, who buy what how many times)
• **clustering**;(The term *cluster* refers to a homogeneous subgroup existing within a population.)
• **description and visualization**.(representation is justified by the remarkable conciseness of the information achieved through a well-designed chart)

Chapter 3 :Data preparation

Data validation

- **Incompleteness.** Some records may contain missing values corresponding to one or more attributes, and there may be a variety of reasons for this.
- **Noise.** Data may contain erroneous or anomalous values, which are usually referred to as *outliers*.
- **Inconsistency.** Sometimes data contain discrepancies due to changes in the coding system used for their representation, and therefore may appear inconsistent.
- **Incomplete data**
  - **Elimination.** It is possible to discard all records for which the values of one or more attributes are missing.
  - **Inspection.** Alternatively, one may opt for an inspection of each missing value in order to obtain recommendations on possible substitute values.
  - **Identification.** As a third possibility, a conventional value might be used to encode and identify missing values, making it unnecessary to remove entire records from the given dataset.
  - categorical attribute one might replace missing values with a new value
  - **Substitution.** Several criteria exist for the automatic replacement of missing data, although most of them appear somehow arbitrary
- **Data affected by noise**
  - *noise* refers to a random perturbation within the values of a numerical attribute, usually resulting in noticeable anomalies
  - **The easiest way to identify outliers is based on the statistical concept of *dispersion*.**
  - If the attribute follows a distribution that is not too far from normal, the values falling outside an appropriate interval centered around the mean value are identified as outliers



Identification of outliers using cluster analysis

Data transformation

data mining analyses it is appropriate to apply a few transformations to the dataset in order to improve the accuracy of the learning models subsequently developed.

- **Standardization**
  - Most learning models benefit from a preventive *standardization* of the data, also called *normalization*.
  - **The most popular standardization techniques include the *decimal scaling* method, the *min-max* method and the *z-index* method.**
    - **Decimal scaling.** Decimal scaling is based on the transformation

$$x'_{ij} = \frac{x_{ij}}{10^h},$$

    - where *h* is a given parameter which determines the scaling intensity.
    - **Min-max.** Min-max standardization is achieved through the transformation

$$x'_{ij} = \frac{x_{ij} - x_{\min,j}}{x_{\max,j} - x_{\min,j}}(x'_{\max,j} - x'_{\min,j}) + x'_{\min,j},$$

    - The minimum and maximum values of the attribute **a***j* before transformation, while *x*`min,*j* and *x*`max,*j* are the minimum and maximum values that we wish to obtain after transformation. In general, the extreme values of the range are defined so that *x*`min,*j* = −1 and *x*`max,*j* = 1 or *x*`min,*j* = 0 and *x*`max,*j* = 1.
    - ***z* -index.** *z*-index based standardization uses the transformation

$$x'_{ij} = \frac{x_{ij} - \bar{\mu}_j}{\bar{\sigma}_j},$$

    - where *μ*¯ *j* and *σ*¯*j* are respectively the sample mean and sample standard deviation of the attribute **a***j* . If the distribution of values of the attribute **a***j* is roughly normal, the *z*-index based transformation generates values that are almost certainly within the range *(*−3*,* 3*)*.
- **Feature extraction**
    - situations in which more complex transformations are used to generate new attributes that represent a set of additional columns in the matrix **X** representing the dataset *D*. Transformations of this kind are usually referred to as *feature extraction*. (customer spending capacity, customer visit intervals)

Data reduction

- when facing a large dataset it is also appropriate to **reduce its size**, in order to make learning algorithms more efficient, **without sacrificing the quality** of the results obtained.
- **There are three main criteria to determine whether a data reduction technique should be used: *efficiency*, *accuracy* and *simplicity* of the models generated.**
    - **Efficiency.** The application of learning algorithms to a dataset smaller than the original one usually means a shorter computation time.
    - **Accuracy.** Data reduction techniques should not significantly compromise the accuracy of the model generated.
    - **Simplicity.** it is important that the models generated be easily translated into simple rules that can be understood by experts in the application domain.
- **Data reduction can be pursued in three distinct directions**
    - reduction in the number of observations through ***sampling***

  - reduction in the number of attributes through ***selection*** **and** ***projection***
  - reduction in the number of values through ***discretization*** **and** ***aggregation***
- **Sampling**
  - reduction in the size of the original dataset can be achieved by extracting a sample of observations that is significant from a statistical standpoint.
- **Feature selection**
  - The purpose of *feature selection*, also called *feature reduction*, is to eliminate from the dataset a subset of variables which are not deemed relevant for the purpose of the data mining activities.
  - Feature selection methods can be classified into **three main categories: *filter* methods, *wrapper* methods and *embedded* methods**
    - **Filter methods.** Filter methods select the relevant attributes before moving on to the subsequent learning phase, and are therefore independent of the specific algorithm being used.
    - **Wrapper methods** : use of a wrapper method is for attribute selection.
    - **Embedded methods.** For the embedded methods, the attribute selection process lies *inside* the learning algorithm, so that the selection of the optimal set of attributes is directly made during the phase of model generation.
  - **In particular, three distinct myopic search schemes can be followed: *forward*, *backward* and *forward–backward* search.**
    - **Forward.** According to the forward search scheme, also referred to as *bottom-up* search(from start index to last index)
    - **Backward.** The backward search scheme, also referred to as *top-down* search (from last indexes to first index)
    - **Forward–backward.** The forward–backward method represents a trade-off between forward and backward search(stops when appropriate value is found)
- **Principal component analysis**
  - *Principal component analysis* (PCA) is the most widely known technique of attribute reduction by means of projection.
  - Generally speaking, the purpose of this method is to obtain a projective transformation that replaces a subset of the original numerical attributes with a lower number of new attributes obtained as their linear combination, without this change causing a loss of information.
  - PCA is mostly used as a tool in exploratory data analysis and for making predictive models. It is often used to visualize genetic distance and relatedness between populations.
  - **Application of PCA**
    - o Quantitative finance (principal component analysis can be directly applied to the risk management of interest rate derivative portfolios)
    - o Neuroscience(principal components analysis is used in neuroscience to identify the specific properties of a stimulus that increase a neuron's probability)
- **Data discretization**
  - Data discretization is the primary reduction method. On the one hand, it reduces continuous attributes to categorical attributes characterized by a limited number of distinct values.
  - the **weekly spending of a mobile phone** customer is a continuous numerical value, which might be discretized into, say, five classes:
  - low, [0 – 10*)* euros; medium low, [10 – 20*)* euros; medium, [20 – 30*)* euros; medium high, [30 – 40*)* euros; and high, over 40 euros.
- **most popular discretization techniques are *subjective subdivision*, *subdivision into classes* and *hierarchical discretization*.**

- **Subjective subdivision.** Subjective subdivision is the most popular and intuitive method. Classes are defined based on the experience and judgment of experts in the application domain.
- **Subdivision into classes.** subdivision can be based on classes of equal size or equal width.
- **Hierarchical discretization.** discretization is based on hierarchical relationships between concepts and may be applied to categorical attributes, just as for the hierarchical relationships between provinces and regions.

UNIT 3 : CHAPTER 1 : Classification

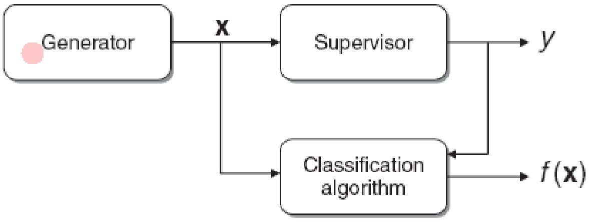
Classification models are supervised learning methods for predicting the value of a categorical target attribute, unlike regression models which deal with numerical attributes.

Starting from a set of past observations whose target class is known, classification models are used to generate a set of rules that allow the target class of future examples to be predicted.

the opportunities afforded by classification extend into several different application domains: selection of the target customers for a marketing campaign, fraud detection, image recognition, early diagnosis of diseases, text cataloguing and spam email recognition

Classification problems

- have a **dataset *D* containing *m* observations** described in terms **of *n explanatory* attributes** and a **categorical *target* attribute**.
- The explanatory attributes, also called *predictive variables*, may be partly categorical and partly numerical.
- The target attribute is also called a class or label
- while the observations are also termed *examples* or *instances*.
- for classification models the target variable takes a finite number of values. In particular, we have a binary classification problem if the instances belong to two classes only, and a multiclass or multicategory classification if there are more than two classes.
- The purpose of a classification model is to *identify recurring relationships among the explanatory variables* which describe the examples belonging to the same class.
- Such relationships are then translated into *classification rules*
- **three components of a classification problem**
  - the probability assumptions concerning **the three components of a classification problem**: a *generator* of observations, a *supervisor* of the target class and a *classification algorithm*.
  - **Generator.** The task of the generator is to **extract random vectors x** of examples according to an unknown probability distribution $P_{\mathbf{x}}(\mathbf{x})$.
  - **Supervisor.** The **supervisor returns for each vector x** of examples the value of the target class according to a conditional distribution $P_{y|\mathbf{x}}(y|\mathbf{x})$ which is also unknown.
  - **Algorithm.** A classification algorithm A_F, also called a *classifier*, chooses a function $f^* \in F$ in the hypothesis space so as **to minimize a suitably defined loss function**.



- A portion of the examples in the dataset *D* is used for *training* a classification model, that is, for deriving the functional relationship between the target variable and the explicative variables expressed by means of the hypothesis $f* \in F$.
- **What remains of the available data is used later to evaluate the accuracy** of the generated model and to select the best model out of those developed using alternative classification methods.
- **development of a classification model consists therefore of three main phases.**
  - **Training phase.** During the *training* phase, the classification algorithm is applied to the examples belonging to a subset *T* of the dataset *D*, called the *training set* , in order to derive classification rules that allow the corresponding target class *y* to be attached to each observation **x**.
  - **Test phase.** In the *test* phase, the rules generated during the training phase are used to classify the observations of *D* not included in the training set, for which the target class value is already known.
  - **Prediction phase.** The *prediction* phase represents the actual use of the classification model to assign the target class to new observations that will be recorded in the future.

• Taxonomy of classification models

- **four main categories of classification models.**
  - **Heuristic models.** Heuristic methods make use of classification procedures based on simple and intuitive algorithms. This category **includes *nearest neighbor* methods**, based on the **concept of distance between observations, and *classification trees***, which use ***divide-and-conquer* schemes** to derive groups of observations that are as homogeneous as possible with respect to the target class.
  - **Separation models.** Separation models **divide the attribute space R*n* into *H* disjoint(** if they have no element in common**) regions** {*S*1, *S*2, . . . , *SH* }, separating the observations based on the target class. popular separation techniques include *discriminant analysis*, *perceptron methods*, *neural networks* and ***support vector machines***.
  - **Regression models.** Regression models is used for the prediction of continuous target variables, make an explicit assumption concerning the functional form of the conditional probabilities *Py*|**x**(*y*|**x**), which correspond to the **assignment of the target class by the supervisor**.
  - **Probabilistic models.** In probabilistic models, a hypothesis is formulated regarding the functional form of the conditional probabilities *P***x**|*y* (**x**|*y)* of the observations given the target class, known as *class-conditional probabilities*.

Evaluation of classification models

classification analysis it is usually advisable to develop alternative models and then select the method affording the best prediction accuracy.

- **Accuracy.** Evaluating the accuracy of a classification model is crucial, the accuracy of a model is an indicator of its ability to predict the target class for future observations. Based on their accuracy values, it is also possible to compare different models in order to select the classifier associated with the best performance.
- **Speed.** Some methods require shorter computation times than others and can handle larger problems. However, classification methods characterized by longer computation times may be applied to a small-size training set obtained from a large number of observations by means of random sampling schemes.
- **Robustness.** A classification method is *robust* if the classification rules generated, as well as the corresponding accuracy, do not vary significantly as the choice of the training set and the test set varies, and if it **is able to handle missing data and outliers**.
- **Scalability.** The scalability of a classifier refers to its ability to learn from large datasets, and it is inevitably related to its computation speed.
- **Interpretability.** If the aim of a classification analysis is to interpret as well as predict, then the rules generated should be simple and easily understood by knowledge workers and experts in the application domain.

Holdout method

- The *holdout* estimation method involves **subdividing the *m* observations available into two disjoint subsets *T* and *V*, for training and testing purposes** respectively, and then evaluating the accuracy of the model through the accuracy acc*A(V)* on the test set.
- *T* is obtained through a simple sampling procedure, which randomly extracts *t* observations from *D*, leaving the remaining examples for the test set *V*. The portion of data used for training may vary based on the size of the dataset *D*.
- The accuracy of a classification algorithm evaluated via the holdout method depends on the test set *V* selected, and therefore it may over- or underestimate the actual accuracy of the classifier as *V* varies.

Repeated random sampling

- The *repeated random sampling* method involves replicating the holdout method a number *r* of times. For each repetition a random independent sample *Tk* is extracted, which includes *t* observations, and the corresponding accuracy acc*A(Vk)* is evaluated, where *Vk* = *D* − *Tk*.

Cross-validation

- The method of *cross-validation* offers an alternative to repeated random sampling techniques and guarantees that each observation of the dataset *D* appears the same number of times in the training sets and exactly once in the test sets.
- The cross-validation scheme is based on a partition of the dataset *D* into *r* disjoint subsets *L*1,*L*2, . . . ,*Lr* , and requires *r* iterations. At the *k*th iteration of the procedure, subset *Lk* is selected as the test set and the union of all the other subsets in the partition as the training set.

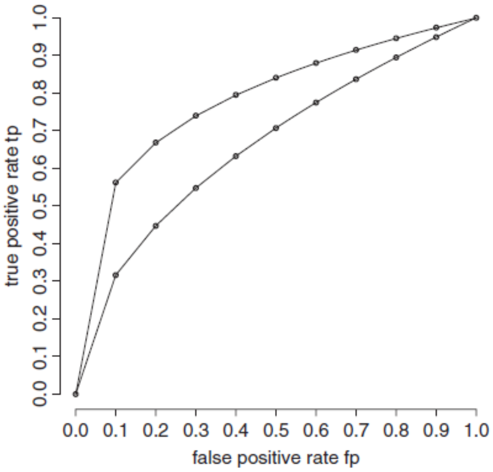
Confusion matrices

- confusion matrix have the following meanings:
- *p* is the number of correct predictions for the negative examples, called ***true negatives***;

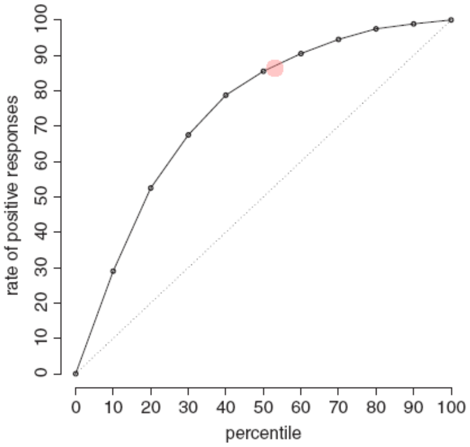
- *u* is the number of incorrect predictions for the positive examples, called ***false negatives***;
- *q* is the number of incorrect predictions for the negative examples, called ***false positives***; and
- *v* is the number of correct predictions for the positive examples, called ***true positives***.
- **Accuracy.** The accuracy of a classifier may be expressed as
  - $$\mathrm{acc} = \frac{p+v}{p+q+u+v} = \frac{p+v}{m}.$$
- **True negatives rate.** The true negatives rate is defined as
  - $$\mathrm{tn} = \frac{p}{p+q}.$$
- **False negatives rate.** The false negatives rate is defined as
  - $$\mathrm{fn} = \frac{u}{u+v}.$$
- **False positives rate.** The false positives rate is defined as
  - $$\mathrm{fp} = \frac{q}{p+q}.$$
- **True positives rate.** The true positives rate, also known as *recall*
  - $$\mathrm{tp} = \frac{v}{u+v}.$$
- **Precision.** The precision is the proportion of correctly classified positive examples, and is given by
  - $$\mathrm{prc} = \frac{v}{q+v}.$$

- ## ROC curve charts

- *Receiver operating characteristic* (ROC) curve charts **allow the user to visually evaluate the accuracy of a classifier and to compare different classification models.** They visually express the information content of a sequence of confusion matrices and allow the ideal trade-off between the number of correctly classified positive observations and the number of incorrectly classified negative observations to be assessed.
- ROC chart is a two-dimensional plot with the proportion of false positives fp on the horizontal axis and the proportion of true positives tp on the vertical axis.
- The point (0,1) represents the ideal classifier, which makes no prediction error since its proportion of false positives is null (fp = 0) and its proportion of true positives is maximum (tp = 1). The point (0,0) corresponds to a classifier that predicts the class {−1} for all the observations, while the point (1,1) corresponds to a classifier predicting the class {1} for all the observations.

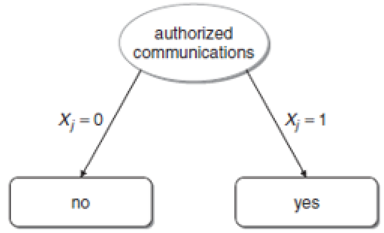


- ## Cumulative gain and lift charts
  - The *lift* measure corresponds to the intuitive idea of evaluating the accuracy of a classifier based on the density of positive observations inside the set that has been identified based on model predictions.
  - Cumulative gain and lift diagrams allow the user to visually evaluate the effectiveness of a classifier. To describe the procedure for constructing the charts and to understand their meaning, consider an example of binary classification arising in a relational marketing analysis, in which we are interested in identifying a subset of customers to be recipients of a cross-selling campaign aimed at promoting a new service.
  - criteria based on cumulative gain and lift curves allow different classifiers to be compared. On the one hand, the area between the cumulative gain curve and the line corresponding to random sampling is evaluated: a greater area corresponds to a classification method that is more effective overall.

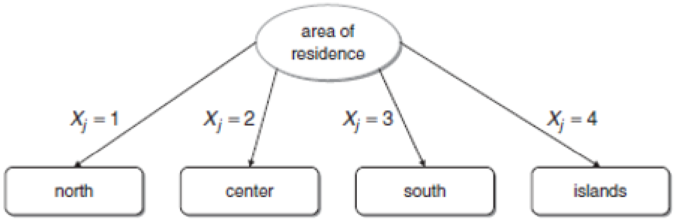


- ## Classification trees
  - *Classification trees* are perhaps the best-known and most widely used learning methods in data mining applications. The reasons for their popularity lie in their conceptual simplicity, ease of usage, computational speed, robustness with respect to missing data and outliers and, most of all, the interpretability of the rules they generate.

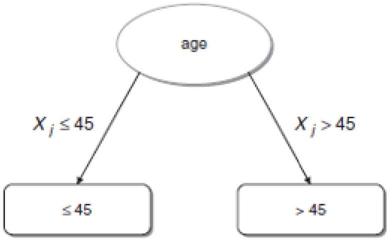
- To separate the observations belonging to different classes, methods based on trees obtain simple and explanatory rules for the relationship existing between the target variable and predictive variables.
- The development of a classification tree corresponds to the training phase of the model and is regulated by a recursive procedure of heuristic nature, based on a divide-and-conquer partitioning scheme referred to as *top-down induction of decision trees*
- *root node* of the tree are divided into disjoint subsets that are tentatively placed in two or more descendant nodes (*branching*).
- The subdivision of the examples in each node is carried out by means of a *splitting rule*, also termed a *separating rule*, to be selected based upon a specific evaluation function.
- set of splitting rules that can be found along the path connecting the tree root to a leaf node constitutes a *classification rule*
    - **components of the top-down induction of decision trees procedure.**
    - **Splitting rules.** For each node of the tree it is necessary to specify the criteria used to identify the optimal rule for splitting the observations and for creating the descendant nodes.
        - **Binary trees.** A tree is said to be *binary* if each node has at most two branches. Binary trees represent in a natural way the subdivision of the observations contained at a node based on the value of a binary explanatory attribute.
        - **Multi-split classification trees.** A tree is said to be *multi-split* if each node has an arbitrary number of branches. This allows multi-valued categorical explanatory attributes to be handled more easily.
        - **Univariate trees.** For *univariate* trees the splitting rule is based on the value assumed by a single explanatory attribute *Xj*



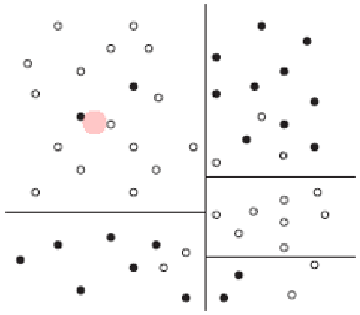
Univariate split for a binary attribute



Univariate split for a nominal attribute



Univariate split for a numerical attribute



Classification by means of univariate splitting rules (axis-parallel).

- **Stopping criteria.** At each node of the tree different *stopping* criteria are applied to establish whether the development should be continued recursively or the node should be considered as a leaf.

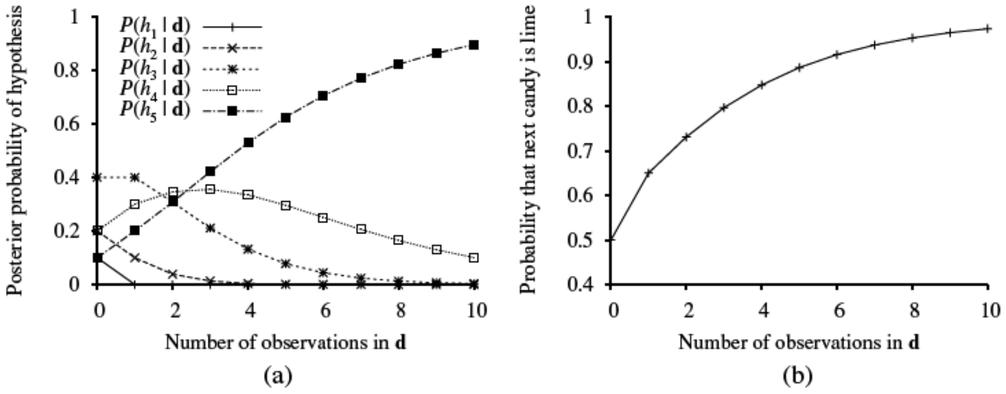
- **Pruning criteria.** Finally, it is appropriate to apply a few *pruning* criteria, first to avoid excessive growth of the tree during the development phase (*pre-pruning*), and then to reduce the number of nodes after the tree has been generated (*post-pruning*).

Bayesian methods

- Bayesian methods belong to the family of probabilistic classification models. They explicitly calculate the *posterior* probability *P(y|***x***)* that a given observation belongs to a specific target class by means of Bayes' theorem, once the *prior* probability *P(y)* and the class conditional probabilities *P(***x***|y)* are known.
- Bayesian classifiers require the user to estimate the probability *P(***x***|y)* that a given observation may occur, provided it belongs to a specific class. The learning phase of a Bayesian classifier may therefore be identified with a preliminary analysis of the observations in the training set, to derive an estimate of the probability values required to perform the classification task.
- Bayesian statistics uses both prior and sample information. Usually something is known about possible parameter values before the experiment is performed.
- The Bayesian approach allows direct probability interpretations of the parameters, given the observed data.

Bayes Theorem

- P(hi | d) = αP(d | hi)P(hi) .
- P(hi | d) is probability of hypothesis given Data
- P(hi) is prior probability
- P(d | hi) is probability of data
- The key quantities in the Bayesian approach are the **hypothesis prior**, P(hi), and the **likelihood** of the data under each hypothesis, P(**d** | hi).



(a) (b)

a) Posterior probabilities P(hi | d1, . . . , dN) The number of observations N ranges from 1 to 10, and each observation is of a lime candy.
b) Bayesian prediction P(dN+1 =lime | d1, . . . , dN)

- *the Bayesian prediction eventually agrees with the true hypothesis.* This is characteristic of Bayesian learning.

- For any fixed prior that does not rule out the true hypothesis, the posterior probability of any false hypothesis will, under certain technical conditions, eventually vanish.
- This happens simply because the probability of generating “uncharacteristic” data indefinitely is vanishingly small.
- A very common approximation—one that is usually adopted in science—is to make predictions based on a single *most probable* hypothesis—that is, an hi that maximizes P(hi | **d**). This is often called a **maximum a posteriori** or MAP (pronounced “em-ay-pee”) hypothesis.
- Predictions made according to an MAP hypothesis hMAP are approximately Bayesian to the extent that **P**(X | **d**) ≈ **P**(X | hMAP).
- A final simplification is provided by assuming a uniform prior over the space of hypotheses. In that case, MAP learning reduces to choosing an hi that maximizes P(d | hi).
- This is called a maximum-likelihood (ML) hypothesis, hML. Maximum-likelihood learning is very common in statistics, a discipline in which many researchers distrust the subjective nature of hypothesis priors.

Bayesian networks

- *Bayesian networks*, also called *belief networks*, allow the hypothesis of conditional independence of the attributes to be relaxed, by introducing some reticular hierarchical links through which it is possible to assign selected stochastic dependencies that experts of the application domain deem relevant.
- **A Bayesian network comprises two main components.**
    - The first is an *acyclic oriented graph* in which the nodes correspond to the predictive variables and the arcs indicate relationships of stochastic dependence.
    - The second component consists of a table of conditional probabilities assigned for each variable.

Logistic regression

- *Logistic regression* is a technique for converting binary classification problems into linear regression ones
- Suppose that the response variable *y* takes the values {0,1}, as in a binary classification problem. The logistic regression model postulates that the posterior probability *P(y*|**x***)* of the response variable conditioned on the vector **x** follows a *logistic function*, given by

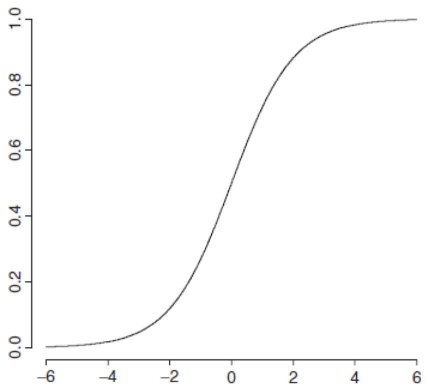
$$P(y=0|\mathbf{x}) = \frac{1}{1+e^{\mathbf{w}'\mathbf{x}}},$$

$$P(y=1|\mathbf{x}) = \frac{e^{\mathbf{w}'\mathbf{x}}}{1+e^{\mathbf{w}'\mathbf{x}}}.$$

- The *standard logistic function S(t)*, also known as the *sigmoid* function, can be found in many applications of statistics in the economic and biological fields and is defined as

$$S(t) = \frac{1}{1+e^{-t}}.$$

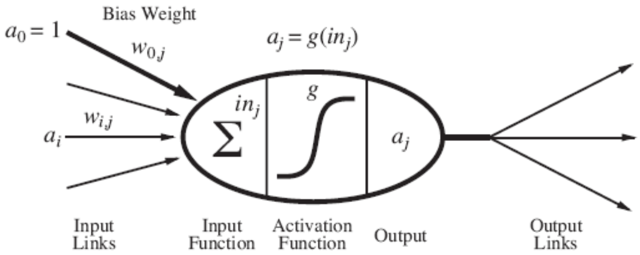
- The function *S(t)* has the graphical shape



Graph of the standard logistic function (sigmoid)

Neural networks

- Neural networks are intended to simulate the behavior of biological systems composed of neurons.
- neural networks have been used for predictive purposes, not only for classification but also for regression of continuous target attributes.
- A neural network is an oriented graph consisting of nodes, which in the biological analogy represent neurons, connected by arcs, which correspond to dendrites and synapses. Each arc is associated with a *weight*, while at each node an *activation function* is defined which is applied to the values received as input by the node along the incoming arcs, adjusted by the weights of the arcs. The training stage is performed by analyzing in sequence the observations contained in the training set one after the other and by modifying at each iteration the weights associated with the arcs.
- the hypothesis that mental activity consists primarily of electrochemical activity in networks of brain cells called **neurons**.
- Inspired by this hypothesis, some of the earliest AI work aimed to create artificial **neural networks**.



A simple mathematical model for a neuron.

Neural network structures

- Neural networks are composed of nodes or **units** connected by directed **links**.
- A link from unit i to unit j serves to propagate the **activation** a_i from i to j.
- Each link also has a numeric **weight** $w_{i,j}$ associated with it, which determines the strength and sign of the connection.

$$in_j = \sum_{i=0}^{n} w_{i,j} a_i \ .$$

- Then it applies an **activation function** g to this sum to derive the output:

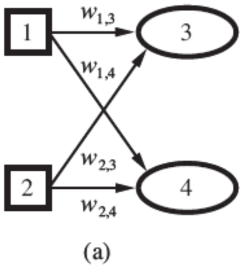
$$a_j = g(in_j) = g\left(\sum_{i=0}^{n} w_{i,j} a_i\right)$$

- The activation function g is typically either a hard threshold in which case the unit is called a **perceptron.**
- Having decided on the mathematical model for individual "neurons," the next task is to connect them together to form a network.

  There are two fundamentally distinct ways to do this.
  1. A **feed-forward network** has connections only in one direction—that is, it forms a directed acyclic graph.
  2. A **recurrent network**, on the other hand, feeds its outputs back into its own inputs.
- Feed-forward networks are usually arranged in **layers**, such that each unit receives input only from units in the immediately preceding layer.

Single-layer feed-forward neural networks (perceptrons)

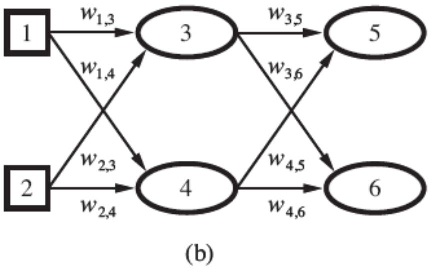
- A network with all the inputs connected directly to the outputs is called a **single-layer neural network**, or a **perceptron network**.



A perceptron network with two inputs and two output units.

Multilayer feed-forward neural networks

- A network with all the inputs connected through hidden layer or unit to the outputs is called a **multi-layer neural network**.



A neural network with two inputs, one hidden layer of two units, and one output unit. Not shown are the dummy inputs and their associated weights.

Learning neural network structures

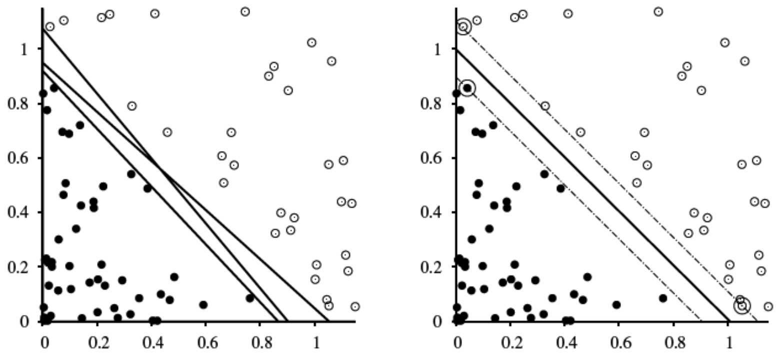
- all statistical models, neural networks are subject to **overfitting** when there are too many parameters in the model.
- If we stick to fully connected networks, the only choices to be made concern the number of hidden layers and their sizes. The usual approach is to **try several and keep the best**.
- The **optimal brain damage** algorithm begins with a fully connected network and removes connections from it. After the network is trained for the first time, an information-theoretic approach identifies an optimal selection of connections that can be dropped.

- the **tiling** algorithm, resembles decision-list learning. The idea is to start with a single unit that does its best to produce the correct output on as many of the training examples as possible.

SUPPORT VECTOR MACHINES

- The **support vector machine** or SVM framework is currently the most popular approach for "off-the-shelf" **supervised learning**: if you don't have any specialized prior knowledge about a domain, then the SVM is an excellent method to try first. There are three properties that make SVMs attractive:

1. SVMs construct a **maximum margin separator**—a decision boundary with the largest possible distance to example points. This helps them generalize well.
2. SVMs create a linear separating hyperplane, but they have the ability to embed the data into a higher-dimensional space, using the so-called **kernel trick(smoothing)(no separable data to separable dataset)**.
3. **SVMs are a nonparametric method**—they retain training examples and potentially need to store them all. On the other hand, in practice they often end up retaining only a small fraction of the number of examples—sometimes as few as a small constant times the number of dimensions.
    - Thus SVMs combine the advantages of nonparametric and parametric models: they have the flexibility to represent complex functions, but they are resistant to overfitting.



(a) (b)

Support vector machine classification:

a) Two classes of points (black and white circles) and three candidate linear separators.
b) The maximum margin separator (heavy line), is at the midpoint of the **margin** (area between dashed lines). The **support vectors** (points with large circles) are the examples closest to the separator.

- Instead of minimizing expected *empirical loss* on the training data, SVMs attempt to minimize expected *generalization* loss. We call this separator the **maximum margin separator(**width of the area bounded by dashed lines)

CHAPTER 2 : Clustering

- The second class of models for unsupervised learning is represented by *clustering* Methods
- By defining appropriate metrics and the induced notions of distance and similarity between pairs of observations, the purpose of clustering methods is the identification of homogeneous groups of records called *clusters*. With respect to the specific distance selected, the observations belonging to each cluster must be close to one another and far from those included in other clusters.

Clustering methods

- The aim of clustering models is to subdivide the records of a dataset into homogeneous groups of observations, called *clusters*, so that observations belonging to one group are similar to one another and dissimilar from observations included in other groups.
- **Clustering methods must fulfill a few general requirements, as indicated below.**
  - **Flexibility.** Some clustering methods can be applied to numerical attributes only, for which it is possible to use the Euclidean metrics to calculate the distances between observations. However, **a flexible clustering algorithm should also be able to analyze datasets containing categorical attributes**. Algorithms based on the Euclidean metrics tend to generate spherical clusters and have difficulty in identifying more complex geometrical forms.
  - **Robustness.** The robustness of an algorithm manifests itself through the stability of the clusters generated with respect to small changes in the values of the attributes of each observation. **This property ensures that the given clustering method is basically unaffected by the noise possibly existing in the data.** Moreover, the clusters generated must be stable with respect to the order of appearance of the observations in the dataset.
  - **Efficiency.** In some applications the number of observations is quite large and therefore clustering algorithms must generate clusters efficiently in order to guarantee reasonable computing times for large problems. In the case of massive datasets, one may also resort to the extraction of samples of reduced size in order to generate clusters more efficiently. However, this approach inevitably implies a lower robustness for the clusters so generated. **Clustering algorithms must also prove efficient with respect to the number of attributes existing in the dataset.**

- **Taxonomy of clustering methods**
  - Clustering methods can be classified into a few main types based on the logic used for deriving the clusters: *partition* methods, *hierarchical* methods, *density based* methods and *grid* methods.
    - **Partition methods.** develop a subdivision of the given dataset into a predetermined number *K* of non-empty subsets. They are suited to obtaining groupings of a spherical or at most convex shape, and can be applied to datasets of small or medium size.
    - **Hierarchical methods.** Carry out multiple subdivisions into subsets, based on a tree structure and characterized by different homogeneity thresholds within each cluster and inhomogeneity thresholds between distinct clusters. Unlike partition methods, hierarchical algorithms do not require the number of clusters to be predetermined.
    - **Density-based methods.** Whereas the two previous classes of algorithms are founded on the notion of distance between observations and between clusters, density-based methods derive clusters from the number of observations locally falling in a neighborhood of each observation. More precisely, for each record belonging to a specific cluster, a neighborhood with a specified diameter must contain a number of observations which should not be lower than a minimum threshold value. Density-based methods can identify clusters of non-convex shape and effectively isolate any possible outliers.
    - **Grid methods.** Grid methods first derive a discretization of the space of the observations, obtaining a grid structure consisting of cells. Subsequent clustering

operations are developed with respect to the grid structure and generally achieve reduced computing times, despite a lower accuracy in the clusters generated.

Affinity measures

- Clustering models are usually based on a measure of similarity between observations. In many instances this can be obtained by defining an appropriate notion of distance between each pair of observations.
    - **Numerical attributes**
        - If all n attributes in a dataset are numerical, we may turn to the *Euclidean distance* between the vectors associated with the pair of observations
    - **Binary attributes**
        - Suppose that a given attribute $\mathbf{a}j = (x1j, x2j, \dots, xmj)$ is binary, so that it assumes only one of the two values 0 or 1. Even if it is possible to formally calculate the difference $xij - xkj$ for every two observations of the dataset, it is clear that this quantity does not represent a distance that can be meaningfully associated with the metrics defined for numerical attributes, since the values 0 and 1 are purely conventional, and their meanings could be interchanged.
    - **Nominal categorical attributes**
        - 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. For example, gender is a categorical variable having two categories (male and female) and there is no intrinsic ordering to the categories.
    - **Ordinal categorical attributes**
        - An ordinal variable is similar to a categorical variable. The difference between the two is that there is a clear ordering of the variables. For example, suppose you have a variable, economic status, with three categories (low, medium and high). In addition to being able to classify people into these three categories, you can order the categories as low, medium and high. Now consider a variable like educational experience (with values such as elementary school graduate, high school graduate, some college and college graduate).

Partition methods

- Given a dataset D of m observations, each represented by a vector in n-dimensional space, partition methods construct a subdivision of D into a collection of non-empty subsets $C = \{C1, C2, \dots, CK\}$, where $K \le m$. In general, the number K of clusters is predetermined and assigned as an input to partition algorithms.
- ***K*-means algorithm**
    a) During the initialization phase, K observations are arbitrarily chosen in D as the **centroids** of the clusters.
    b) Each observation is iteratively assigned to the cluster whose centroid is the most similar to the observation, in the sense that it minimizes the distance from the record.
    c) If no observation is assigned to a different cluster with respect to the previous iteration, the algorithm stops.
    d) For each cluster, the new centroid is computed as the mean of the values of the observations belonging to the cluster, and then the algorithm returns to step 2.
- ***K*-medoids algorithm**
    - The *K-medoids* algorithm, also known as *partitioning around medoids*, is a variant of the *K*-means method. It is based on the use of *medoids* instead of the means of the

observations belonging to each cluster, with the purpose of mitigating the sensitivity of the partitions generated with respect to the extreme values in the dataset.

- Given a cluster *Ch*, a *medoid* **u***h* is the most *central* observation, in a sense that will be formally defined, among those that are assigned to *Ch*. Once the **medoid** representing each cluster has been identified, the *K*-medoids algorithm proceeds like the *K*-means method, using medoids instead of centroids.
- Medoids are most commonly used on data when a mean or centroid cannot be defined, such as graphs. They are also used in contexts where the centroid is not representative of the dataset
- similar to the k-means algorithm but works when a mean or centroid is not definable

Hierarchical methods

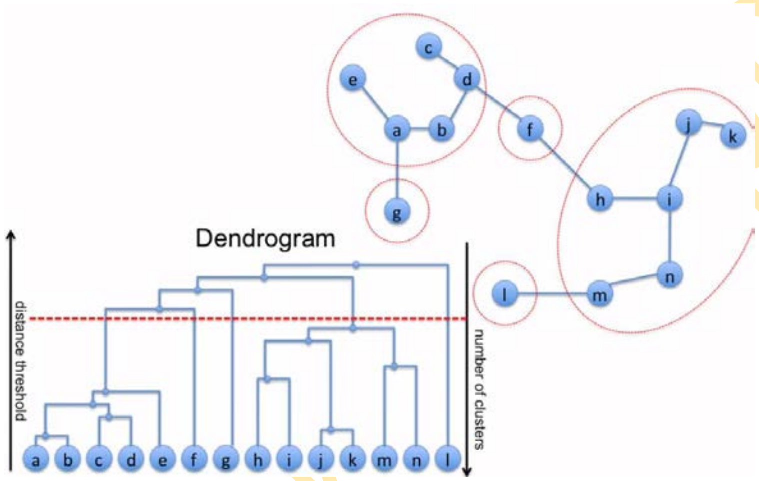
- Hierarchical clustering methods are based on a tree structure. Unlike partition methods, they do not require the number of clusters to be determined in advance. Hence, they receive as input a dataset *D* containing *m* observations and a matrix of distances dist(**x***i* , **x***k)* between all pairs of observations.
- In order to evaluate the distance between two clusters, most hierarchical algorithms resort to one of five alternative measures: minimum distance, maximum distance, mean distance, distance between centroids, and Ward distance.
  - **Minimum distance.** According to the criterion of *minimum distance*, also called the *single linkage* criterion, the dissimilarity between two clusters is given by the minimum distance among all pairs of observations such that one belongs to the first cluster and the other to the second cluster,
  - **Maximum distance.** According to the criterion of *maximum distance*, also called the *complete linkage* criterion, the dissimilarity between two clusters is given by the maximum distance among all pairs of observations such that one belongs to the first cluster and the other to the second cluster
  - **Mean distance.** The *mean distance* criterion expresses the dissimilarity between two clusters via the mean of the distances between all pairs of observations belonging to the two clusters,
  - **Distance between centroids.** The criterion based on the *distance between centroids* determines the dissimilarity between two clusters through the distance between the centroids representing the two clusters
  - **Ward distance.** The criterion of *Ward distance*, based on the analysis of the variance of the Euclidean distances between the observations
  - **Hierarchical methods can be subdivided into two main groups: *agglomerative* and *divisive* methods**
    - **Agglomerative hierarchical methods**
    - Agglomerative methods are *bottom-up* techniques in which each single observation initially represents a distinct cluster. These clusters are then aggregated during subsequent iterations, deriving clusters of increasingly larger cardinalities. The algorithm is stopped when a single cluster including all the observations has been reached.
    - **Agglomerative algorithm**

      a) In the initialization phase, each observation constitutes a cluster. The distance between clusters therefore corresponds to the matrix **D** of the distances between all pairs of observations.

b) The minimum distance between the clusters is then computed, and the two clusters *Ch* and *Cf* with the minimum distance are merged, thus deriving a new cluster *Ce*. The corresponding minimum distance dist*(Ch,Cf)* originating the merger is recorded.

c) The distance between the new cluster *Ce*, resulting from the merger between *Ch* and *Cf* , and the preexisting clusters is computed.

d) If all the observations are included into a single cluster, the procedure stops. Otherwise it is repeated from step 2.



- **Divisive hierarchical methods**
  - Divisive algorithms are the opposite of agglomerative methods, in that they are based on a *top-down* technique, which initially places all the observations in a single cluster. This is then subdivided into clusters of smaller size, so that the distances between the generated subgroups are minimized. The procedure is repeated until clusters containing a single observation are obtained, or until an analogous stopping condition is met.

Evaluation of clustering models

- To evaluate a clustering method it is first necessary to verify that the clusters generated correspond to an actual regular pattern in the data. It is therefore appropriate to apply other clustering algorithms and to compare the results obtained by different methods. In this way it is also possible to evaluate if the number of identified clusters is robust with respect to the different techniques applied.

UNIT 4: CHAPTER 1: Marketing Models

Marketing Decision Process

- Marketing decision processes are characterized by a high level of complexity due to the simultaneous presence of multiple objectives and countless alternatives actions resulting from the combination of the major choice options available to decision makers.

Relational marketing

- The aim of a *relational marketing* strategy is to initiate, strengthen, intensify and preserve over time the relationships between a company and its stakeholders, represented primarily by its customers, and involves the analysis, planning, execution and evaluation of the activities carried out to pursue these objectives.

- **Motivations and objectives**
  - The increasing concentration of companies in large enterprises and the resulting growth in the number of customers have led to greater complexity in the markets.
  - Since the 1980s, the innovation-production-obsolescence cycle has progressively shortened, causing a growth in the number of customized options on the part of customers, and an acceleration of marketing activities by enterprises.
  - The increased flow of information and the introduction of e-commerce have enabled global comparisons.
  - Customer loyalty has become more uncertain, primarily in the service industries, where often filling out an on-line form is all one has to do to change service provider.
- **Relational marketing strategies**
  - Relational marketing strategies revolve around the choices shown in Figure 1, which can be effectively summarized as formulating for each segment, ideally for each customer, the appropriate offer through the most suitable channel, at the right time and at the best price.

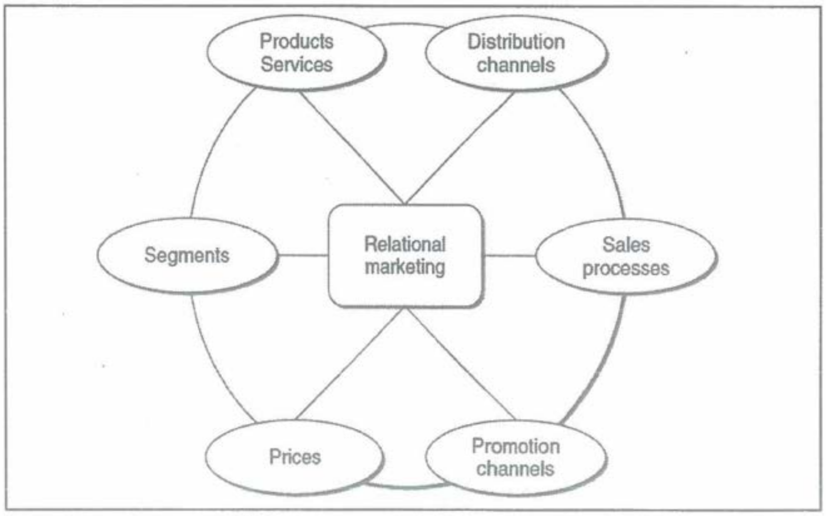


Fig 1 : Decision making options for a relational marketing strategy

  - In particular, it is advisable to stress the distinction between a relational marketing vision and the software tools usually referred to as *customer relationship management* (CRM).
  - As shown in Figure 2, relational marketing is not merely a collection of software applications, but rather a coherent project where the various company departments are called upon to cooperate and integrate the managerial culture and human resources, with a high impact on the organizational structures.

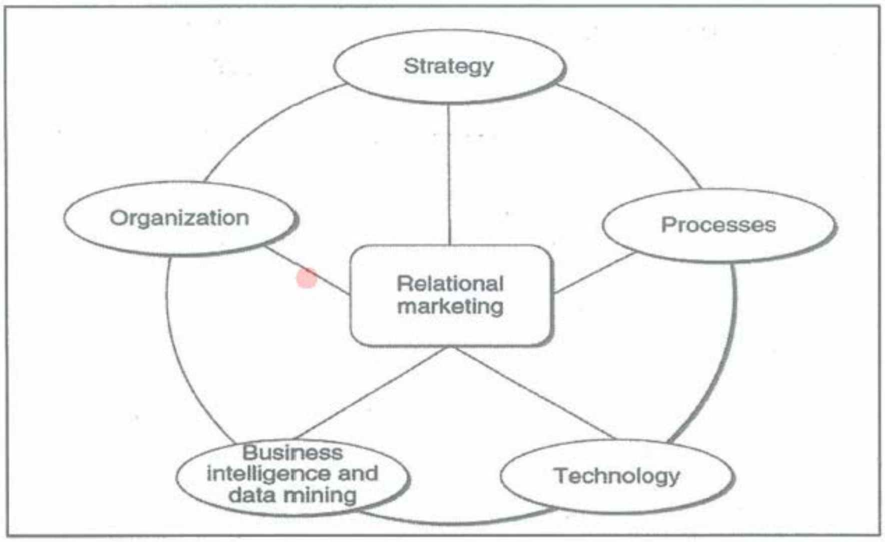


Fig 2 : Components of a relational marketing strategy

- **Network of relationships**
  - The relationship system of an enterprise is not limited to the dyadic (binary relations) relationship with its customers, represented by individuals and companies that purchase the products and services offered, but also includes other actors, such as the employees, the suppliers and the sales network.

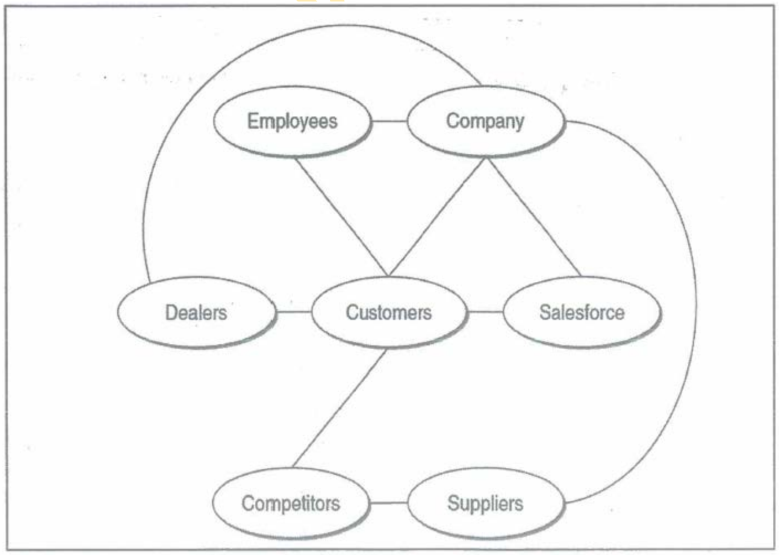


Fig 3 : Network of relationship involved in a relational marketing strategy

- **Intensity of customer**
  - For most relationships shown in Figure 3, a mutually beneficial exchange occurs between the different subjects involved.
  - More generally, we can widen the boundaries of relational marketing systems to include the stakeholders of an enterprise.
  - The number of customers and their characteristics strongly influence the nature and intensity of the relationship with an enterprise, as shown in Figure 4.

- At the opposite extreme of the diagonal are the relationships typical of consumer goods and *business-to-consumer* (B2C) activities, for which a high number of low-value customers get in contact with the company in an impersonal way, through websites, call centers and points of sale.

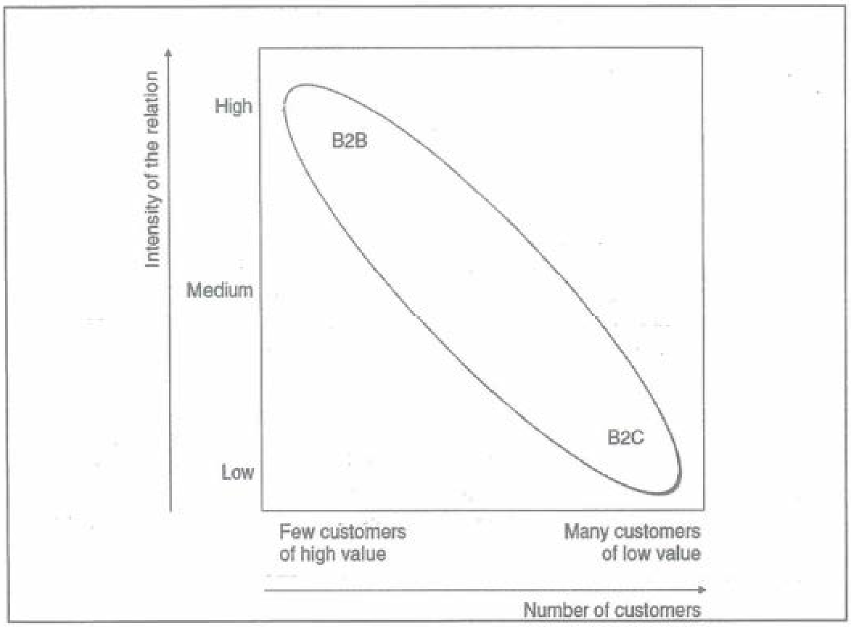

Intensity of the relation
High
Medium
Low
B2B
B2C
Few customers of high value
Many customers of low value
Number of customers

Fig: 4 : intensity of customer relationship as a function of members of customer

- **Efficiency of sales action**
  - Figure 5 contrasts the cost of sales actions and the corresponding revenues.
  - Where transactions earn a low revenue per unit, it is necessary to implement low-cost actions, as in the case of mass-marketing activities.
  - Moving down along the diagonal in the figure, more evolved and intense relationships with the customers can be found.

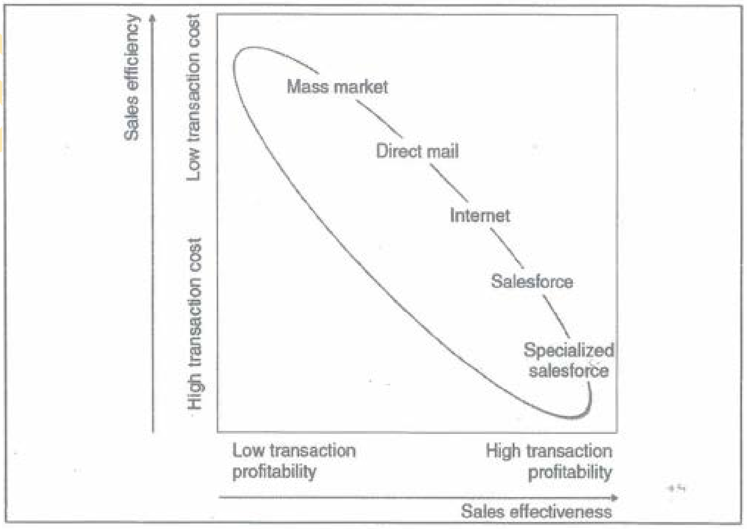


Fig: 5 : Efficiency of sales action as a function of their effectiveness

- **Level of customization as a function of complexity**

- Figure 6 shows the ideal path that a company should follow so as to be able to offer customized products and services at low cost and in a short time.

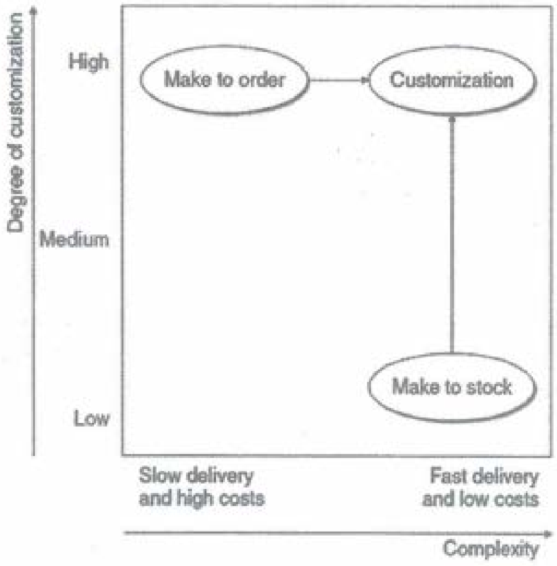


Fig : 6 : Level of customization as a function of complexity of products and services

- **An environment for relational marketing analysis**
  - Figure 7 shows the main elements that make up an environment for relational marketing analysis.
  - Information infrastructures include the company's data warehouse, obtained from the integration of the various internal and external data sources, and a marketing data mart that feeds business intelligence and data mining analyses for profiling potential and actual customers.

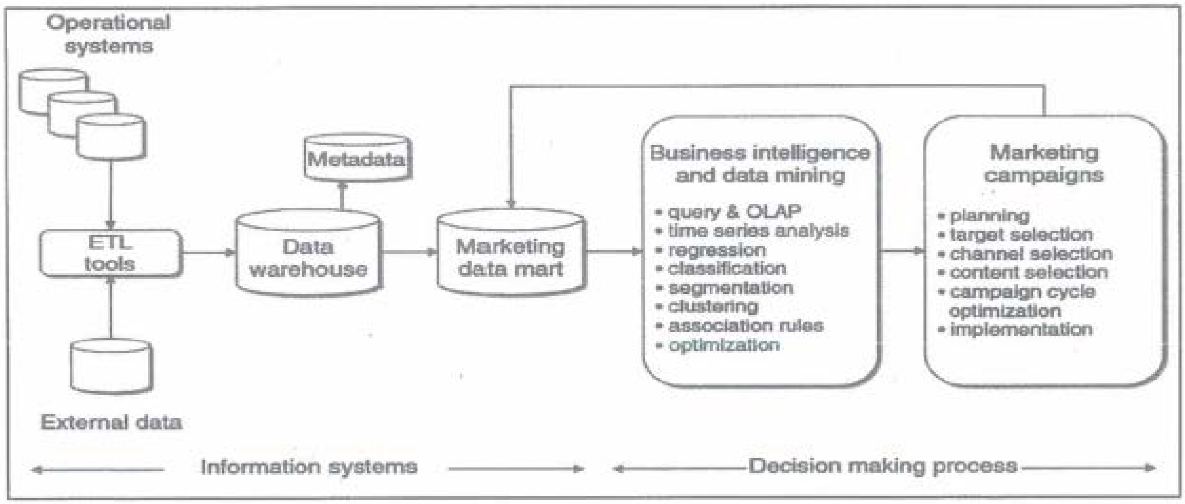


Fig : 7 :Components of an environment for relational marketing analysis

- **Types of data feeding**
  - Figure 8 describes the main types of data stored in a data mart for relational marketing analyses.

- A company data warehouse provides demographic and administrative information on each customer and the transactions carried out for purchasing products and using services.

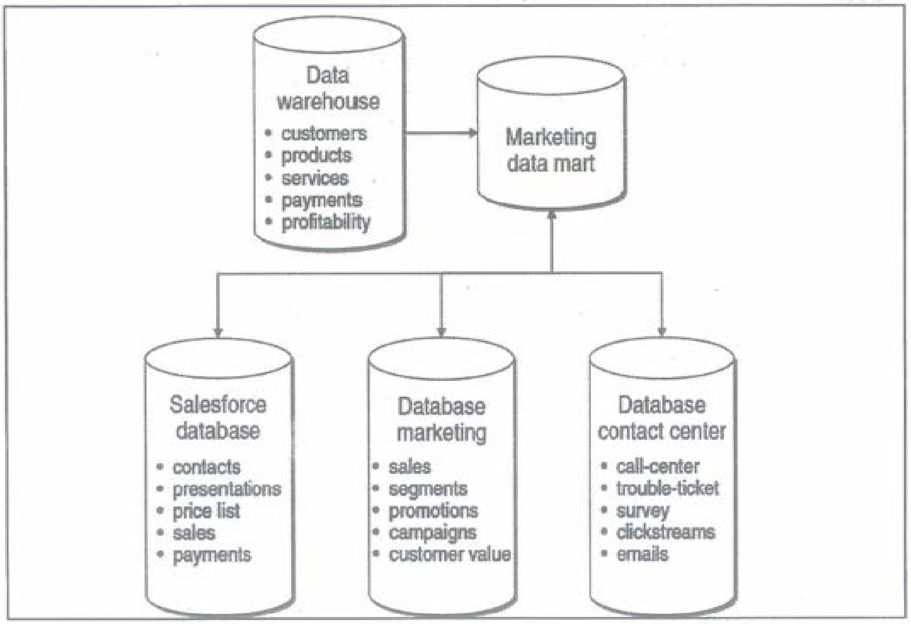


Fig.8 : Types of data feeding a data mart of relational marketing analysis

- **Life cycle of relational marketing**
  - The main phases of a relational marketing analysis proceeds as shown in Figure 9.
  - The first step is the exploration of the data available for each customer.
  - At a later time, by using inductive learning models, it is possible to extract from those data the insights and the rules that allow market segments characterized by similar behaviors to be identified. Knowledge of customer profiles is used to design marketing actions which are then translated into promotional campaigns and generate in turn new information to be used in the course of subsequent analyses.

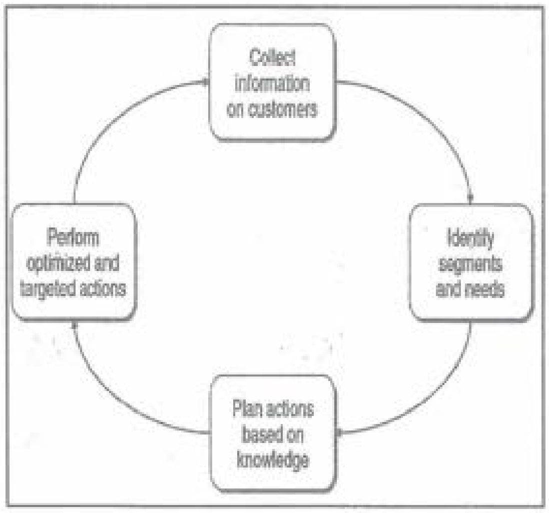

Collect information on customers
Identify segments and needs
Plan actions based on knowledge
Perform optimized and targeted actions

Fig.: 9: Cycle of relational marketing analysis

- **Lifetime value**

  - Figure 10 shows the main stages during the customer *lifetime,* showing the cumulative value of a customer over time. The figure also illustrates the different actions that can be undertaken toward a customer by an enterprise.
  - In the initial phase, an individual is a *prospect,* or potential customer, who has not yet begun to purchase the products or to use the services of the enterprise.
  - Toward potential customers, *acquisition* actions are carried out, both directly (telephone contacts, emails, talks with sales agents) and indirectly (advertising, notices on the enterprise website).

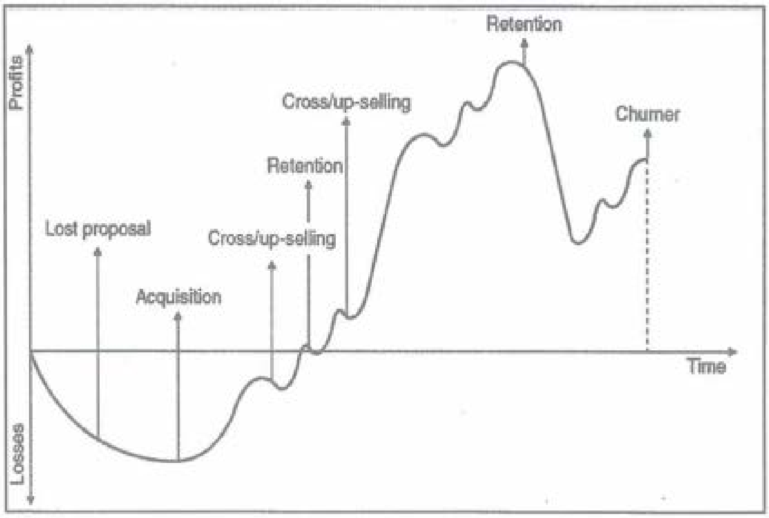


Fig.:10: Lifetime of a customer

CHAPTER 2: Logistic and production models

- **<u>SUPPLY CHAIN OPTIMIZATION</u>**
- In a broad sense, a supply chain may be defined as a network of connected and interdependent organizational units that operate in a coordinated way to manage, control and improve the flow of materials and information originating from the suppliers and reaching the end customers, after going through the procurement, processing and distribution subsystems of a company, as shown in Figure 1.
- The aim of the integrated planning and operations of the supply chain is to combine and evaluate from a systemic perspective the decisions made and the actions undertaken within the various sub processes that compose the logistic system of a company.
- Many manufacturing companies, such as those operating in the consumer goods industry, have concentrated their efforts on the integrated operations of the supply chain, even to the point of incorporating parts of the logistic chain that are outside the company, both upstream and downstream.
- The major purpose of an integrated logistic process is to minimize a function expressing the total cost, which comprises processing costs, transportation costs for procurement and distribution, inventory costs and equipment costs.
- Optimization models represent a powerful and versatile conceptual paradigm for analyzing and solving problems arising within integrated supply chain planning, and for developing the necessary software.

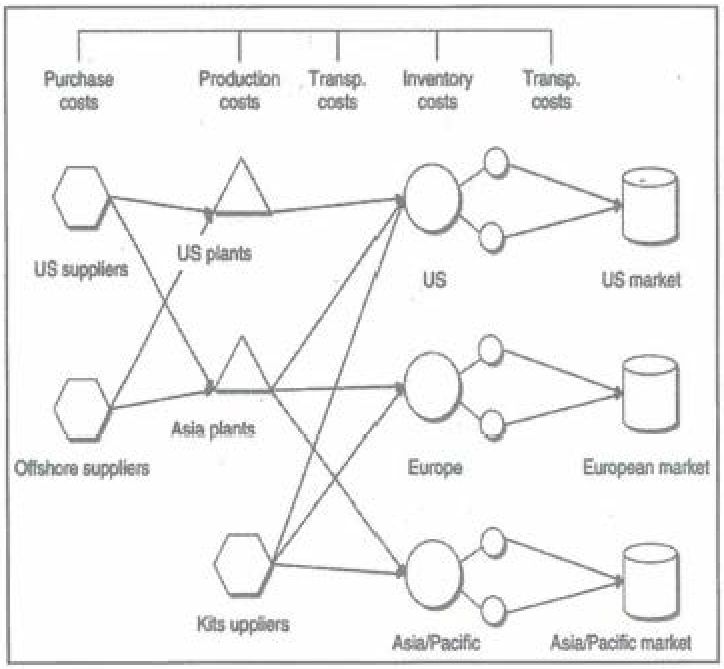
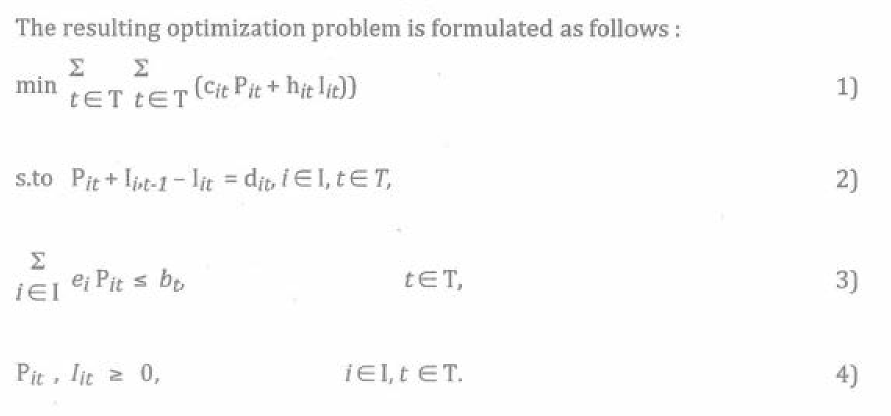


Fig.:1: An example of global supply chain

- **OPTIMIZATION MODELS FOR LOGISTICS PLANNING**
- **Tactical Planning**
- The aim of tactical planning is to determine the production volumes for each product over the T periods included in the medium-term planning horizon in such a way as to satisfy the given demand and capacity limits for a single resource, and also to minimize the total cost; defined as the sum of manufacturing production costs and inventory costs.
- We therefore consider the decision variables:
    - P_{it} = units of product i to be manufactured in period t,
    - I_{it} = units of product i in inventory at the end of period t, and the parameters
    - d_{it} = demand for product i in period t,
    - c_{it} = unit manufacturing cost for product i in period t,
    - h_{it} = unit inventory cost for product i in period t,
    - e_i = capacity absorption to manufacture a unit of product i,
    - b_t = capacity available in period t.



The resulting optimization problem is formulated as follows :

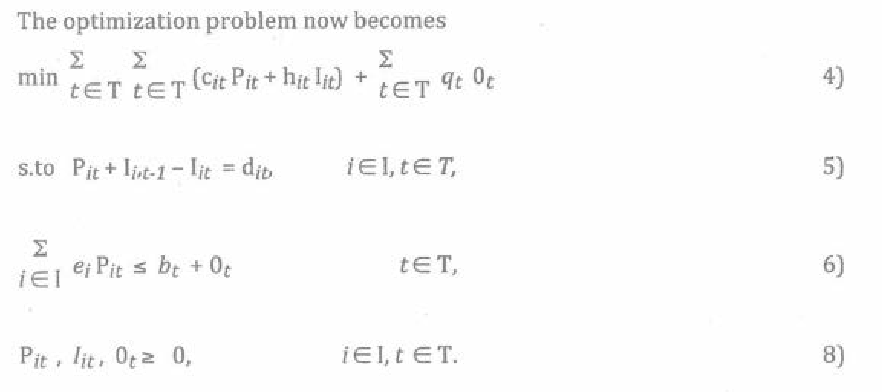
$$\min \sum_{t\in T}\sum_{t\in T}(c_{it}P_{it}+h_{it}I_{it})) \qquad 1)$$

$$\text{s.to}\quad P_{it}+I_{i,t-1}-I_{it}=d_{it},\ i\in I,\ t\in T, \qquad 2)$$

$$\sum_{i\in I} e_i P_{it}\le b_t, \qquad t\in T, \qquad 3)$$

$$P_{it},\ I_{it}\ \ge 0, \qquad i\in I,\ t\in T. \qquad 4)$$

- Constraints 2) express the balance conditions among production, inventory and demand, by establishing a connection between successive periods along the planning horizon.
- Inequalities 3) constrain the absorbed capacity not to exceed the available capacity for each period.
- A first extension of the basic model shown in Figure 9.1 deals with the possibility of resorting to extra capacity, perhaps in the form of overtime, part-time or third-party capacity.
- In addition to the decision variables already included in model (9.1), we define the variables O_t = extra capacity used in period t, and the parameters qt = unit cost of extra capacity in period t.
- q_t = unit cost of extra capacity in period t.
- The optimization problem now becomes



The optimization problem now becomes

$$\min \sum_{t\in T}\sum_{t\in T}(c_{it}P_{it}+h_{it}I_{it}) + \sum_{t\in T} q_t O_t \qquad 4)$$

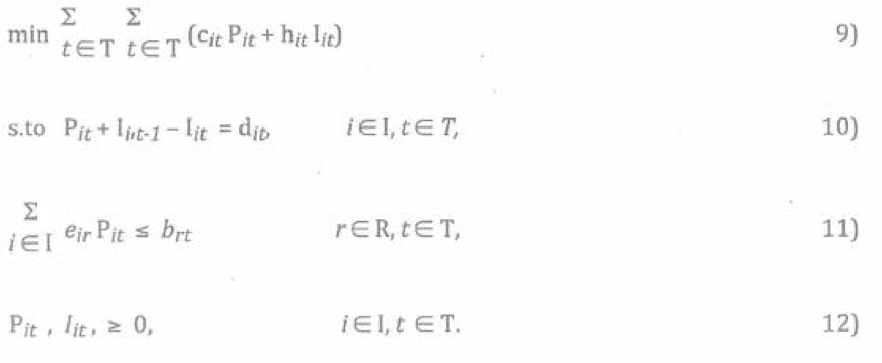
$$\text{s.to}\quad P_{it}+I_{i,t-1}-I_{it}=d_{it}, \qquad i\in I, t\in T, \qquad 5)$$

$$\sum_{i\in I} e_i P_{it} \le b_t + O_t \qquad t\in T, \qquad 6)$$

$$P_{it}\,,\ I_{it}\,,\ O_t \ge 0, \qquad i\in I, t\in T. \qquad 8)$$

- Constraints 7) have been modified to include the available extra capacity.
- The extended model 5) is still a linear optimization problem which can be therefore solved efficiently.

- **Multiple Resources**
- If the manufacturing system requires R critical resources, a further extension of model shown in figure 9.1 can be devised by considering multiple capacity constraints.
- The decision variables already included in model figure 1 remain unchanged, though it is necessary to consider the additional parameters.
- b_{rt} = quantity of resource r available in period t,
- e_{ir} = quantity of resource r absorbed to manufacture one unit of product z.
- The resulting optimization problem is given by



$$\min \sum_{t\in T}\sum_{t\in T}(c_{it}P_{it}+h_{it}I_{it}) \qquad 9)$$

$$\text{s.to}\quad P_{it}+I_{i,t-1}-I_{it}=d_{it}, \qquad i\in I, t\in T, \qquad 10)$$

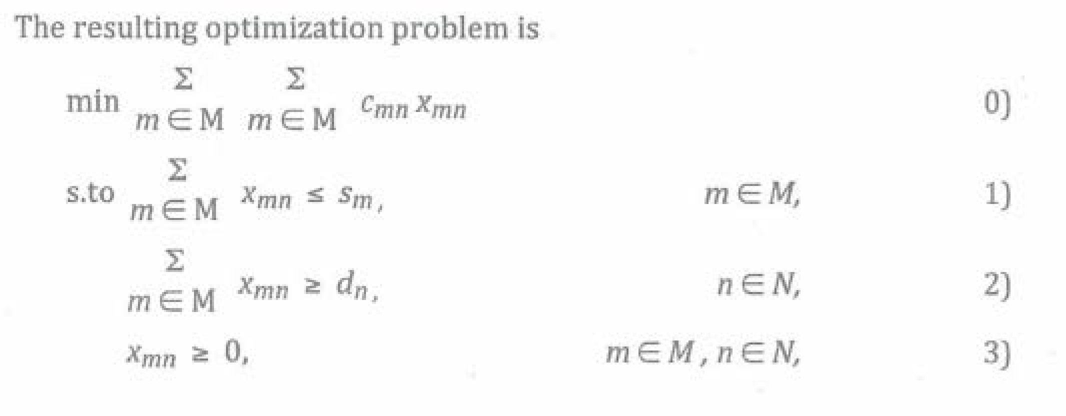
$$\sum_{i\in I} e_{ir} P_{it} \le b_{rt} \qquad r\in R, t\in T, \qquad 11)$$

$$P_{it}\,,\ I_{it}\,, \ge 0, \qquad i\in I, t\in T. \qquad 12)$$

- Constraints 11) have been modified to take into account the upper limits on the capacity of the R resources in the system.

- Model 9) remains a linear optimization problem which can be solved efficiently.

- **Backlogging**
  - The term backlog refers to the possibility that a portion of the demand due in a given period may be satisfied in a subsequent period, incurring an additional penalty cost.
  - Backlogs are a feature of production systems more likely to occur in B2B or make-to order manufacturing contexts.
- **Minimum Lots and Fixed Costs**
  - A further feature often appearing in manufacturing systems is represented by minimum lot conditions: for technical or scale economy reasons, it is sometimes necessary that the production volume for one or more products be either equal to 0 (i.e. the product is not manufactured in a specific period) or not less than a given threshold value, the minimum lot.
- **Bill Of Materials**
  - A further extension of the basic planning model deals with the representation of products with a complex structure, described via the so-called bill of materials, where end-items are made by components that in turn may include other components.
- **Multiple Plants**
  - The logistic system is responsible for supplying N peripheral depots, located in turn at distinct sites.
  - Each production plant m E M = (1, 2,...,M} is characterized by a maximum availability of product, denoted by sm, while each plant n G N = (1, 2,...,N} has a demand dn.
  - We further assume that a transportation cost cmn is incurred by sending a unit of product from plant m to depot n, for each pair (m, n) of origins and destinations in the logistic network.
  - The decision variables needed to model the problem described represent the quantity to be transported for each plant-depot pair, x_{mn} = unit of product to be transported from *m* to *n.*
  - The resulting optimization problem is



The resulting optimization problem is

$$\min \sum_{m \in \mathrm{M}} \sum_{m \in \mathrm{M}} c_{mn} x_{mn} \qquad 0)$$

$$\text{s.to} \sum_{m \in \mathrm{M}} x_{mn} \leq s_m, \qquad m \in M, \qquad 1)$$

$$\sum_{m \in \mathrm{M}} x_{mn} \geq d_n, \qquad n \in N, \qquad 2)$$

$$x_{mn} \geq 0, \qquad m \in M, n \in N, \qquad 3)$$

  - Constraints 1) ensure that the availability of each plant is not exceeded, whereas constraints 2) establish that the demand of each depot be satisfied. Model 0) is a linear optimization problem, and can be therefore solved efficiently.
- **REVENUE MANAGEMENT SYSTEMS**
  - Revenue management is a managerial policy whose purpose is to maximize profits through an optimal balance between demand and supply.

- Despite the potential advantages that revenue management initiatives may offer for enterprises, there are certain difficulties that hamper the actual implementation of practical projects and actions aimed at adopting revenue management methodologies and tools.

- **Decision Processes In Revenue Management Systems:**
  - Revenue management involves the *application of mathematical models to predict the behavior of customers at a micro-segmentation level and to optimize the availability and price of products in order to maximize profits*.
  - Revenue management affects some highly complex decision-making processes of strategic relevance, as shown in Figure 2:

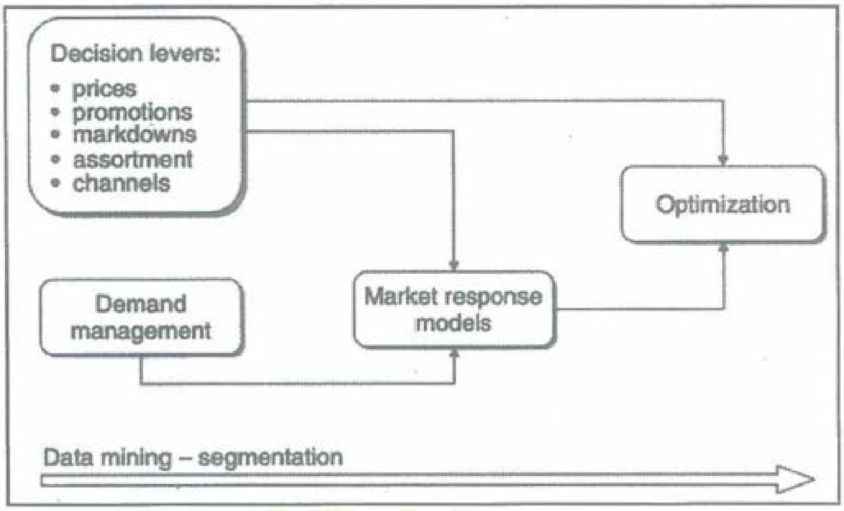


Fig.: 2 : Decision processes in revenue management

- **Decision Processes In Revenue Management Systems**
  - Market segmentation, by product, distribution channel, consumer type and geographic area, performed using data mining models;
  - Prediction of future demand, using time series and regression models; Identification of the optimal assortment, i.e. the mix of products to be allocated to each point of sale;
  - Definition of the market response function, obtained by identifying models and rules that explain the demand based on company actions, the initiatives of competitors and other exogenous contextual events;
  - Management of activities aimed at determining the price of each product (pricing) as well as the timing and the amount of markdowns;
  - Planning, management and monitoring of sales promotions, and assessment of their effectiveness;
  - Sales analysis and control, and use of the information gathered to evaluate market trends;
  - Material procurement and stock management policies, such as control policy, frequency of issued orders, reorder quantities;
  - Integrated management of the different sales and distribution channels.

- **Revenue Management Relies On The Following Basic Principles**
  - To address sales to micro-segments: segmentation carried out by means of business intelligence and data mining models is critical to achieve an adequate knowledge of the market.

- To exploit the product value cycle: to generate the highest revenues, it is required to grasp the value cycle of products and services, in order to optimally synchronize their availability over time and to determine the price for each market micro-segment.
- To have a price-oriented rather than cost-oriented approach in balancing supply and demand : when supply and demand are out of balance, most enterprises tend to react by increasing or decreasing capacity.
- To make informed and knowledge-based decisions : a consistent use of prediction models tends to mean that decisions rest on a more robust knowledge basis.
- To regularly examine new opportunities to increase revenues and profits : the possibility of timely access to the available information, combined with the possibility of considering alternative scenarios, strengthens the competencies of marketing analysts and increases the effectiveness of their activity.

Chapter 3: Data envelopment analysis

- **EFFICIENCY MEASURES**
  - In data envelopment analysis the units being compared are called decision making units (DMUs], since they enjoy a certain decisional autonomy.
  - Assuming that we wish to evaluate the efficiency of n units, let N = {1, 2,...,n} denote the set of units being compared.
  - If the units produce a single output using a single input only, the efficiency of the j th decision making unit DMUj, j 6 N, is defined as, in which yj is the output value produced by DMUj and xj the input value used.

$$\theta_j = \frac{y_j}{x_j},$$

- **EFFICIENT FRONTIER**
  - The efficient frontier, also known as production function, expresses the relationship between the inputs utilized and the outputs produced.
  - It indicates the maximum quantity of outputs that can be obtained from a given combination of inputs.
  - At the same time, it also expresses the minimum quantity of inputs that must be used to achieve a given output level.
  - Hence, the efficient frontier corresponds to technically efficient operating methods.
  - The efficient frontier may be empirically obtained based on a set of observations that express the output level obtained by applying a specific combination of input production factors.
  - The production possibility set is defined as the region delimited by the efficient frontier where the observed units being compared are found.

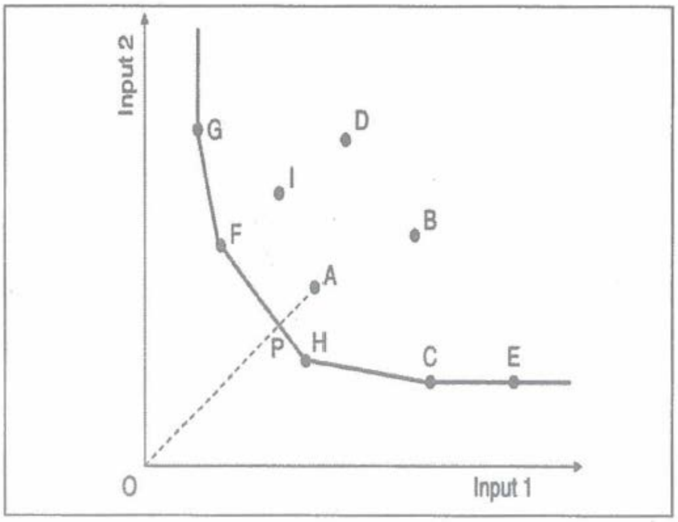


FIG :1: Efficient frontier with two inputs and one output

- **THE CCR MODEL**
  - Using data envelopment analysis, the choice of the optimal system of weights for a generic DMUj, involves solving a mathematical optimization model whose decision variables are represented by the weights ur, r £ K, and vi, i G H, associated with each output and input.
  - The CCR (Charnes-Cooper-Rhodes) model formulated for decision making units, DMUj, takes the form.

$$\max v = \frac{\Sigma r \in K\, u_r y_{rj}}{\Sigma i \in \mathrm{H}\, v_i x_{ij}}$$

Definition Of Target Objectives

- In real-world applications it is often desirable to set improvement objectives for inefficient units, in terms of both outputs produced and inputs utilized.
- Data envelopment analysis provides important recommendations in this respect, since it identifies the output and input levels at which a given inefficient unit may become efficient.
- The efficiency score of a unit expresses the maximum proportion of the actually utilized inputs that the unit should use in conditions of efficiency, in order to guarantee its current output levels.
- **Performance Improvement Strategies**
  - o Other performance improvement strategies may be preferred over the proportional reduction in the quantities of inputs used or the proportional **increase in the output quantities produced**
- **Priority order for the production factors**
  - o The target values for the inputs are set in such a way as to **minimize the quantity** used of the resources to which the highest priority has been assigned, without allowing variations in the level of other inputs or in the outputs produced;
- **Priority order for the outputs**
  - o The target values for the outputs are set in such a way as to **maximize the quantity** produced of the outputs to which highest priority has been assigned, without allowing variations in the level of other outputs or inputs used;
  - o Preferences expressed by the decision makers with respect to a decrease in some inputs or an increase in specific outputs.

- **Peer Groups**
  - Data envelopment analysis identifies for each inefficient unit a set of excellent units, called a peer group, which includes those **units that are efficient if evaluated with the optimal system of weights of an inefficient unit**.
  - The peer group, made up of DMUs which are characterized by operating methods similar to the inefficient unit being examined, is a realistic term of comparison which the unit should **aim to imitate in order to improve its performance.**

IDENTIFICATION OF GOOD OPERATING PRACTICES

- The need to identify the efficient units, for the purpose of defining the best operating practices, stems from the principle itself on which data envelopment analysis is grounded, since it allows each unit to evaluate its own degree of efficiency by choosing the most advantageous structure of weights for inputs and outputs.
- We may resort to a combination of different methods: cross-efficiency analysis, evaluation of virtual inputs and virtual outputs, and weight restrictions.
- **Cross-efficiency analysis**
  - The analysis of cross-efficiency is based on the definition of the efficiency matrix, which provides information on the nature of the weights system adopted by the units for their own efficiency evaluation.
  - The square efficiency matrix contains as many rows and columns as there **are units being compared.**
- **Virtual Inputs And Virtual Outputs**
  - Virtual inputs and virtual outputs provide information on the relative importance that each unit attributes to each individual input and output, for the purpose of maximizing its own efficiency score. The virtual inputs of a DMU are defined as the product of the inputs used by the unit and the corresponding optimal weights. Similarly, virtual outputs are given by the product of the outputs of the unit and the associated optimal weights.
  - Inputs and outputs for which the unit shows high virtual scores provide an indication of the activities in which the unit being analyzed appears particularly efficient.
  - Two efficient units may yield high virtual values corresponding to different combinations of inputs and outputs, showing good operating practices in different contexts.
- **Weight Restrictions**
  - To separate the units that are really efficient from those whose efficiency score largely depends on the selected weights system, we may impose some restrictions on the value of the weights to be associated with inputs and outputs.
  - In general, these restrictions translate into the definition of maximum thresholds for the weight of specific outputs or minimum thresholds for the weight of specific inputs

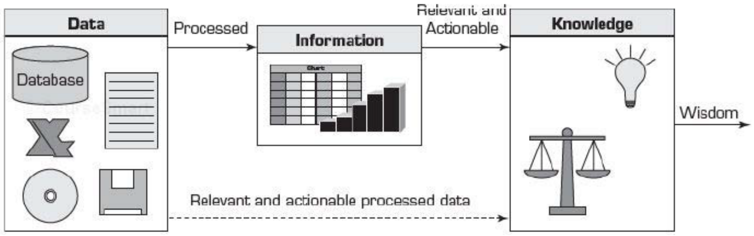
UNIT 5 : CHAPTER 1 : Knowledge Management

INTRODUCTION TO KNOWLEDGE MANAGEMENT

- **What is Knowledge?**
  - A frequently adopted definition of knowledge is that of "**justified true belief'**, That definition incorporates three basic conditions for knowledge.
  - **The truth conditions**
    - It requires that if one knows a proposition then that proposition must be true.

    - If the proposition is not true, then that person does not know what he claims to know
    - The truth condition makes the difference between opinion and knowledge.
  - **The belief conditions**
    - That condition demands that if one knows proposition then he believes that proposition.
  - **The justification conditions**
    - That condition requires a practical way of justifying that the belief one has is true.

- **What is Data, Information, Knowledge, Wisdom**
  - **Data**
    - Information, often in the form of facts or figures obtained from experiments or surveys, used as a basis for making calculations or drawing conclusions Information, for example, numbers, text, images, and sounds, in a form that is suitable for storage in or processing by a computer
  - **Information**
    - Definite knowledge acquired or supplied about something or somebody the collected facts and data about a particular subject a telephone service that supplies telephone numbers to the public on request.
  - **Knowldege**
    - General awareness or possession of information, facts, ideas, truths, or principles clear awareness or explicit information, for example, of a situation or fact all the information, facts, truths, and principles learned throughout time familiarity or understanding gained through experience or study.
  - **Wisdom**
    - The knowledge and experience needed to make sensible decisions and judgments, or the good sense shown by the decisions and judgments made accumulated knowledge of life or in a particular sphere of activity that has been gained through experience an opinion that almost everyone seems to share or express ancient teachings or sayings.



- **knowledge has the following characteristics**
  - **Extraordinary leverage and increasing returns.** Knowledge is not subject to diminishing returns. When it is used, it is not decreased (or depleted), rather it is increased (or improved). Its consumers can add to it, thus increasing its value.
  - **Fragmentation, leakage, and the need to refresh.** As knowledge grows, it branches and fragments. Knowledge is dynamic; it is information in action. Thus, an organization must continually refresh its knowledge base to maintain it as a source of competitive advantage.

- **Uncertain value.** It is difficult to estimate the impact of an investment in knowledge. There are too many intangible aspects that cannot be easily quantified.
- **Value of sharing.** It is difficult to estimate the value of sharing one's knowledge or even who will benefit most from it.

- **The knowledge-based economy**
  - is a reality Rapid changes in the business environment cannot be handled in traditional ways. Firms are much larger today than they used to be, and in some areas, turnover is extremely high, fueling the need for better tools for collaboration, communication, and knowledge sharing. Firms must develop strategies to sustain competitive advantage by leveraging their intellectual assets for optimal performance

- **Explicit and Tacit Knowledge**
  - **Explicit knowledge** deals with more objective, rational, and technical knowledge (e.g., data, policies, procedures, software, documents).
  - **Tacit knowledge** is usually in the domain of subjective, cognitive, and experiential learning; it is highly personal and difficult to formalize.

ORGANIZATIONAL LEARNING AND TRANSFORMATION

- Organizational learning is the development of new knowledge and insights that have the potential to influence an organization's behavior.
- It occurs when associations, cognitive systems, and memories are shared by members of an organization.
- **Organization Learning includes**
  1. Openness to new perspectives
  2. Awareness of personal biases
  3. Exposure to unfiltered data
  4. A sense of humility.
- Establishing a corporate memory is critical for success, Information Technology plays a critical role in organizational learning and management must place emphasis on this area to foster it.
- Because organizations are becoming more virtual in their operations, they must develop methods for effective organizational learning.
- Modern collaborative technologies can help in Knowledge Management initiatives.
- Organizational learning and memory depend less on technology than on people issues.
- **Organizational Culture**
  - An organization's ability to learn, develop memory and share knowledge is dependent on its culture.
  - Culture is a pattern of shared basic assumptions.
  - Over time, organizations learn what works and what does not work, as a lessons become secondary, they become part of the organizational culture.
  - The impact of organizational culture is difficult to measure, However strong culture generally produce strong, measurable bottom-line results such as net income, return on invested capital and yearly increases in stock price.
  - Encouraging employees to use a KMS, both for contributing knowledge and for seeking knowledge, can be difficult.
  - **Many Reasons why people may not share their knowledge are as follows:**

1. They don't have time.
2. They don't trust others.
3. They think that knowledge is power.
4. They don't know why they should do it.
5. They don't know how to do it.
6. They don't know what they are supposed to do
7. They think the recommended way will not work.
8. They think their way is better.
9. They think something else is more important.
10. There is no positive consequence to them for doing it.

KNOWLEDGE MANAGEMENT ACTIVITIES

- Most KM initiatives have one of three aims :
  1. To Make knowledge visible
  2. To develop knowledge-intensive culture
  3. To build knowledge infrastructure
- Several activities or processes surround the management of knowledge.
- These include the creation of knowledge, the sharing of knowledge and the seeking and use of knowledge
- KM Activities involves three activities as follows:
  1. Knowledge Creation
  2. Knowledge Sharing
  3. Knowledge Seeking

- **Knowledge Creation**
  - To create new knowledge means quite literally to re-create the company and everyone in it in a nonstop process of personal and organizational self-renewal.
  - In the knowledge-creating company, inventing new knowledge is not a specialized activity–the province of the R&D department or marketing or strategic planning.
  - Creating new knowledge is as much about ideals as it is about ideas.
  - It is a way of behaving, indeed a way of being, in which everyone is a knowledge worker–that is to say, an entrepreneur.
  - Four modes of knowledge creation are socialization, externalization, internalization and combination.
  - The socialization mode refers to the conversion of tacit knowledge to new tacit knowledge through social interactions and shared experience among organization members (eg. mentoring).
  - The combination mode refers to the creation of new explicit knowledge by merging, categorizing, reclassifying and synthesizing existing explicit knowledge (eg. Statistical analyses of market data).
  - The other two modes involve interactions and conversion between tacit and explicit knowledge.
  - Externalization refers to converting tacit knowledge to new explicit knowledge (eg. producing a written document describing the procedures used in solving a particular client's problem).

- Internalization refers to the creation of new tacit knowledge from explicit knowledge (eg. obtaining a novel insight through reading a document).

- **Knowledge Sharing**
  - Knowledge sharing is the willful explication of one person's ideas, insights, solutions, experiences (ie. knowledge) to another individual either via an intermediary, such as a computer-based system, or directly.
  - However, in many organizations, information and knowledge are not considered organizational resources to be shared but individual competitive weapons to be kept private.
  - Organizational members may share personal knowledge with trepidation, they perceive that they are of less value if their knowledge is a part of the organizational public domain.

- **Knowledge Seeking**
  - Knowledge Seeking/sourcing is the search for and use of internal organizational knowledge.
  - Lack of time or lack of reward may hinder the sharing of knowledge, and the same is true of knowledge seeking.
  - Individuals may sometimes prefer to not reuse knowledge if they feel that their own performance review is based on the originality or creativity of their ideas.
  - Note : Individual may engage in knowledge creation, sharing and seeking with or without using I.T. Tools.

APPROACHES TO KNOWLEDGE MANAGEMENT

- Two fundamental approaches to KM are Process approach and the practice approach.
- **Process approach to KM**
  - Process approach to knowledge management attempts to codify organizational knowledge through formalized controls, processes and technologies.
  - Organizations that adopt the process approach may implement explicit policies governing how knowledge is to be collected, stored and disseminated throughout the organization.
- **Practice approach to KM**
  - Practice approach to knowledge management assumes that a great deal a organizational knowledge is tacit in nature and that formal controls, processes and technologies are not suitable for transmitting this type of understanding.
  - Rather than build formal systems to manage knowledge, the focus of this approach is to build the social environments or communities of practice necessary to facilitate the sharing of tacit understanding.

The Process and Practice Approaches to Knowledge Management

	Process Approach	**Practice Approach**
Type of knowledge supported	Explicit knowledge - codified in rules, tools, and processes	Mostly tacit knowledge - unarticulated knowledge not easily captured or codified
Means of transmission	Formal controls, procedures, and standard operating procedures, with heavy	Informal social groups that engage in storytelling and improvisation.

	emphasis on information technologies to support knowledge creation, codification and transfer of knowledge.	
Benefit	Provides structure to harness generated ideas and knowledge Achieves scale in knowledge reuse Provides spark for fresh ideas and responsiveness to changing environment	Provides an environment to generate and transfer highvalue tacit knowledge
Disadvantages	Fails to tap into tacit knowledge May limit innovation and forces participants into fixed patterns of thinking	Can result in inefficiency Abundance of ideas with no structure to implement them.
Role of information technology (IT)	Requires heavy investment in IT to connect people with reusable codified knowledge	Requires moderate investment in IT to facilitate conversations and transfer of tacit knowledge.

INFORMATION TECHNOLOGY (IT) IN KNOWLEDGE MANAGEMENT (COMPONENTS OF KMS)

- The two primary functions of I.T. in knowledge management are retrieval and communication.
- I.T. also extends the reach and range of knowledge use and enhances the speed of knowledge transfer.
- Network facilitate collaboration in KM.
- Storage and retrieval technologies originally meant using a dbms to store and manage knowledge.
- However capturing, storing, and managing tacit knowledge usually requires a different set of tools.
- E-document management system and specialized storage systems that are part of collaborative computing system fill this void.
- This systems are known as knowledge repositories.

Knowledge Management Technologies and Web Impacts

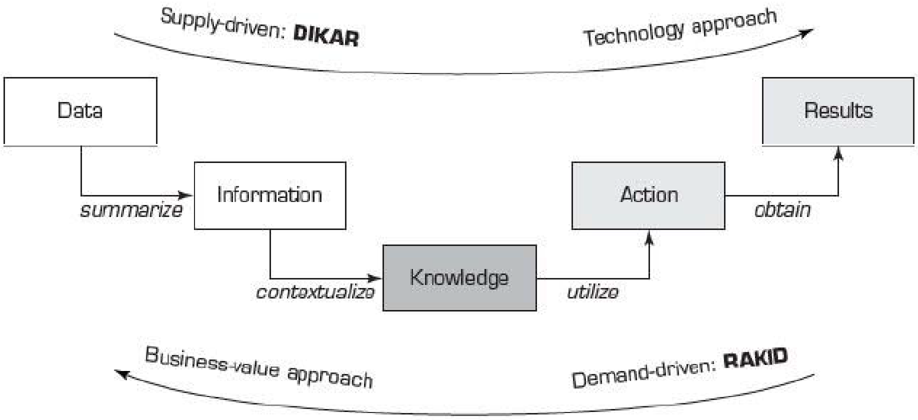
Knowledge Management	Web Impacts	Impacts on the Web
Communication	Consistent, friendly graphical user interface (GUI) for client units Improved communication tools Convenient, fast access to knowledge and knowledgeable individuals Direct access to knowledge on servers	K now ledge captured and shared is used in improving communication, communication management, and communication technologies
Collaboration	Improve collaboration tools Enables anywhere/anytime collaboration Enables collaboration between companies, customers, and vendors.	K now ledge captured and shared is used in improving collaboration, collaboration management, and collaboration technologies (i.e., GSS)

	Enables document sharing improved, fast collaboration and links to knowledge source Makes audio-and videoconferencing a reality, especially for individuals not using a local area network.	
Storage and retrieval	Consistent, friendly GUI for clients Servers provide for efficient and effective storage and retrieval of knowledge	Knowledge captured and shared is utilized in improving data storage and retrieval systems, database management/knowledge repository management, and database and knowledge repository technologies.

- **Hybrid Approaches to Knowledge Management**
  - Many organizations use a hybrid of the process and practice approaches. Early in the development process, when it may not be clear how to extract tacit knowledge from its sources, the practice approach is used so that a repository stores only explicit knowledge that is relatively easy to document.
  - The tacit knowledge initially stored in the repository is contact information about experts and their areas of expertise. Such information is listed so that people in the organization can find sources of expertise (e.g., the process approach).
  - From this start, best practices can eventually be captured and managed so that the knowledge repository will contain an increasing amount of tacit knowledge over time. Eventually, a true process approach may be attained. But if the environment changes rapidly, only some of the best practices will prove useful.
  - Regardless of the type of KMS developed, a storage location for the knowledge (i.e., a knowledge repository') of some kind is needed.

The DIKAR (Data-Information-Knowledge-Action-Results)

- The DIKAR (Data-Information-Knowledge-Action-Results) model in the left to right mode is useful in identifying, collecting and storing data and information assets of the organization in a systematic manner. It is a technology-driven approach to automate the process of knowledge accumulation from what is available in organizational memory.
- The assumption is that once you have compiled the data, information, and knowledge (because you are able to do so with automated means), you will be able to figure out a way to productively use it for specific business actions. Unfortunately, establishing the seemingly simple connection between the accumulated knowledge assets and the necessary business actions is not a trivial task. Because of the vast variety of business situations, the knowledge nuggets that the decision maker needs to initiate the right action for a specific situation may not be in the right form to be recognized or often may not even exist in the knowledge repository.



KNOWLEDGE MANAGEMENT SYSTEMS IMPLEMENTATION

- The two primary functions of IT in knowledge management are retrieval and communication.
- IT also extends the reach and range of knowledge use and enhances the speed of knowledge transfer. Networks facilitate collaboration in KM.
- **The KMS Cycle**
  - A functioning KMS follows six steps in a cycle. The reason for the cycle is that knowledge is dynamically refined over time. The knowledge in a good KMS is never finished because the environment changes over time, and the knowledge must be updated to reflect the changes. The cycle works as follows:
    1. **Create knowledge.** Knowledge is created as people determine new ways of doing things or develop knowhow. Sometimes external knowledge is brought in. Some of these new ways may become best practices.
    2. **Capture knowledge**. New knowledge must be identified as valuable and be represented in a reasonable way.
    3. **Refine knowledge**. New knowledge must be placed in context so that it is actionable. This is where human insights (i.e., tacit qualities) must be captured along with explicit facts.
    4. **Store knowledge**. Useful knowledge must be stored in a reasonable format in a knowledge repository so that others in die organization can access it.
    5. **Manage knowledge**. Like a library, a repository must be kept current. It must be reviewed to verify that it is relevant and accurate.
    6. **Disseminate knowledge**. Knowledge must be made available in a useful format to anyone in the organization who needs it, anywhere and anytime.

- **Technologies That Support Knowledge Management**
  - o Several technologies have contributed to significant advances in knowledge management tools. Artificial intelligence, intelligent agents, knowledge discovery in databases, extensible Markup Language (XML), and Web 2.0 are examples of technologies.
    - ▪ **ARTIFICIAL INTELLIGENCE** In the definition of knowledge management, artificial intelligence (AI) is rarely mentioned. Howrever, practically speaking, AI methods and tools are embedded in a number of KMS, either by vendors or by system developers. AI methods can assist in identifying expertise, eliciting knowledge

automatically and semi-automatically, interfacing through natural language processing, and intelligently searching through intelligent agents.

- **INTELLIGENT AGENTS** Intelligent agents are software systems that learn how users work and provide assistance in their daily tasks. Intelligent agents can help in KMS in a number of ways. Typically, they are used to elicit and identify knowledge.
- **KNOWLEDGE DISCOVERY IN DATABASES** Knowledge discovery in databases (KDD) is a process used to search for and extract useful information from volumes of documents and data. It includes tasks such as knowledge extraction, data archaeology, data exploration, data pattern processing, data dredging, and information harvesting.
- **EXTENSIBLE MARKUP LANGUAGE (XML)** extensible Markup Language (XML) enables standardized representations of data structures so that data can be processed appropriately by heterogeneous systems without case-by-case programming. This method suits e-commerce applications and supply-chain management (SCM) systems that operate across enterprise boundaries. XML can not only automate processes and reduce paperwork, it can also unite business partners and supply chains for better collaboration and knowledge transfer
- **WEB 2.0** Recent years have seen a shift in how people use the World Wide Web. The Web has evolved from a tool for disseminating information and conducting business to a platform for facilitating new ways of information sharing, collaboration, and communication in the digital age. A new vocabulary has emerged, as mashup, social networks, mediasharing sites, RSS, blogs, and wikis have come to characterize the genre of interactive applications collectively known as Web 2.0.

KNOWLEDGE MANAGEMENT SYSTEMS IMPLEMENTATION

- The challenge with KMS is to identify and integrate the three essential components communication technologies, collaboration technologies, and storage and retrieval technologies—to meet the knowledge management needs of an organization. The earliest
- **Knowledge Management Products and Vendors**
  - Technology tools that support KM are called know ware. Most knowledge management includes collaborative tools, knowledge servers, enterprise knowledge portals, electronic document management systems, knowledge harvesting tools, search engines, and knowledge management suites.
- **SOFTWARE DEVELOPMENT COMPANIES AND EIS VENDORS**
  - Software development companies and EIS vendors offer numerous knowledge management packages, from individual tools to comprehensive knowledge management suites.
- **KNOWLEDGE SERVERS**
  - A knowledge server contains the main knowledge management software, including the knowledge repository, and provides access to other knowledge, information, and data.
- **Electronic document management (EDM)**
  - systems use the document in electronic form as the collaborative focus of work. EDM systems allow users to access needed documents, generally via a Web browser over a corporate intranet.

    - A new approach to EDM, called content management systems (CMS), is changing the way documents and their content are managed. A CMS produces dynamic versions of documents and automatically maintains the "current" set for use at the enterprise level.
    - **A subset of CMS is business rules management. New software tools and systems, such as Ilog JRules and Blaze Advisor, have been developed to handle these smaller chunks of content.**
        - Knowledge Harvesting Tools
            - Tools for capturing knowledge unobtrusively are helpful because they allow a knowledge contributor to be minimally (or not at all) involved in the knowledge-harvesting efforts.
        - Search Engines
            - Search engines perform one of the essential functions of knowledge management—locating and retrieving necessary documents from vast collections accumulated in corporate repositories.
        - Knowledge Management
            - Suites Knowledge management suites are complete out-of-the-box knowledge management solutions. They integrate the communications, collaboration, and storage technologies into a single convenient package.
- **KNOWLEDGE MANAGEMENT CONSULTING FIRMS**
    - All the major consulting firms (e.g., Accenture, Cap Gemini Ernst & Young, Deloitte & Touche, KPMG, PWC) have massive internal knowledge management initiatives.
- **KNOWLEDGE MANAGEMENT ASPS**
    - ASPs have evolved as a form of KMS outsourcing on the Web. There are many ASPs for e-commerce on the market.
- **Integration of KMS with Other Business Information Systems**
    - Because a KMS is an enterprise system, it must be integrated with other enterprise and information systems in an organization. Obviously, when it is designed and developed it.
    - **INTEGRATION OF KMS WITH DSS/BI SYSTEMS**
        - KMS typically do not involve running models to solve problems. This is typically done in DSS/BI systems.
    - **INTEGRATION OF KMS WITH AI**
        - KM has a natural relationship with AI methods and software, although knowledge management, strictly speaking, is not an AI method.
    - **INTEGRATION OF KMS WITH DATABASES AND INFORMATION SYSTEMS**
        - Because a KMS uses a knowledge repository, sometimes constructed out of a database system or an EDM system, it can automatically integrate to this part of the firm's information system.
    - **INTEGRATION OF KMS WITH CRM SYSTEMS**
        - CRM systems help users in dealing with customers. One aspect is the help-desk notion described earlier. But CRM goes much
    - **INTEGRATION WITH SCM SYSTEMS**
        - The supply chain is often considered to be the logistics end of a business.

Roles of People in Knowledge Management

- They include chief knowledge officer (CKO), the CEO, the other officers and managers of an organisation, members and leaders of communities of the practice, KMS developers and KMS staff.

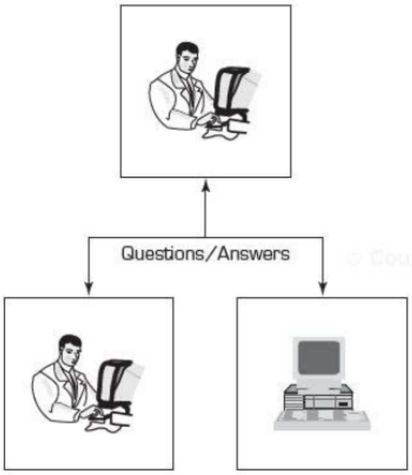
- Each person or group has an important role in either the development, management, or use of a KMS
- **Role of CKO as follows**
    - Set knowledge management strategic priorities.
    - Establish a knowledge repository of best practices.
    - Gain a commitment from senior executives to support a learning environment.
    - Teach information seekers how to ask better and smarter questions.
    - Establish a process for managing intellectual assets.
    - Obtain customer satisfaction information in near real time.
    - Globalize knowledge management
- **CKO needs following skills in KMS:**
    - Interpersonal communication skills to convince employees to adopt cultural changes
    - Leadership skills to convey the knowledge management vision and passion for it
    - Business acumen to relate knowledge management efforts to efficiency and profitability
    - Strategic thinking skills to relate knowledge management efforts to efficiency and profitability
    - Collaboration skills to work with various departments and persuade them to work together
    - The ability to institute effective educational programs
    - An understanding of IT and its role in advancing knowledge management
- **Managers in KMS**
    - In many KMS, Managers are the part of the communities of practice.
- **Communities Of Practice (COP)**
    - COP is a group of people in organisation with common professional interest.
    - Ideally all KMS users should each be in one COP.
    - Properly creating and nurturing COP is one of the key to KMS success.
- **KMS Developers**
    - KMS developers are the team members who actually develop the system. They work for the CKO.
- **Seven Principles for Designing Successful COP**
    1. Design for evolution.
    2. Open a dialogue between inside and outside.
    3. Invite different levels of participation
    4. Develop public and private spaces.
    5. Focus on value.
    6. Combine familiarity and excitement.
    7. Create a rhythm for the community.

CHAPTER 2 : Artificial Intelligence and Expert Systems

Concepts and Definitions of Artificial Intelligence

- In addition to rule-based expert systems, several other technologies can be used to support decision situations where expertise is required. Most of these technologies use qualitative (or symbolic) knowledge rather than numeric and/or mathematical models to provide the needed support; hence, they are referred to as knowledge-based systems (KBS). The overarching field of study that encompasses these technologies and underlying applications is called artificial intelligence.

- **Artificial Intelligence (AI) Definitions**
    - Artificial intelligence (AI) is an area of computer science. Even though the term has many different definitions, most experts agree that AI is concerned with two basic ideas: (1) the study of human thought processes (to understand what intelligence is) and (2) the representation and duplication of those thought processes in machines (e.g., computers, robots).
- To understand what artificial intelligence is, we need to examine those abilities that are considered to be signs of intelligence:
    - Learning or understanding from experience
    - Making sense out of ambiguous or contradictory messages
    - Responding quickly and successfully to a new situation (i.e., different responses, flexibility)
    - Using reasoning in solving problems and directing conduct effectively
    - Dealing with perplexing situations
    - Understanding and inferring in a rational way
    - Applying knowledge to manipulate the environment
    - Thinking and reasoning
    - Recognizing and judging the relative importance of different elements in a situation



- A Pictorial Representation of the Turing Test

- Alan Turing designed an interesting test to determine whether a computer exhibits intelligent behavior; the test is called the Turing test. According to this test, a computer can be considered smart only when a human interviewer cannot identify the computer while conversing with both an unseen human being and an unseen computer

- **Characteristics of Artificial Intelligence**
- **SYMBOLIC PROCESSING** Symbolic processing is an essential characteristic of AI, as reflected in the following definition: Artificial intelligence (AI) is the branch of computer science that deals primarily with symbolic, non-algorithmic methods of problem solving. This definition focuses on two characteristics:

- **Numeric versus symbolic.** Computers were originally designed specifically to process numbers (i.e., numeric processing). However, people tend to think symbolically; our intelligence is based, in part, on our mental ability to manipulate symbols rather than just numbers.
- **Algorithmic versus heuristic.** An algorithm is a step-by-step procedure that has well-defined starting and ending points and is guaranteed to find the same solution to a specific problem. Most computer architectures readily lend themselves to this type of step-by-step approach. Many human reasoning processes however tend to be non-algorithmic; in other words, our mental activities consist of more than just following logical, step-by-step procedures. Rather, human thinking relies more on rules, opinions, and gut feelings, learned from previous experiences.

- **HEURISTICS Heuristics** are intuitive knowledge, or rules of thumb, learned from experience.
- **INFERENCING** As an alternative to merely using individual heuristics, AI also includes reasoning (or inferencing) capabilities that can build higher-level knowledge using existing knowledge represented as heuristics in the form of rules. Inference is the process of deriving a logical outcome from a given set of facts and rules.
- **MACHINE LEARNING** Learning is an important capability for human beings; it is one of the features that separate humans from other creatures. AI systems do not have the same learning capabilities that humans have; rather, they have simplistic learning capabilities (modeled after die human learning methods) called machine learning.

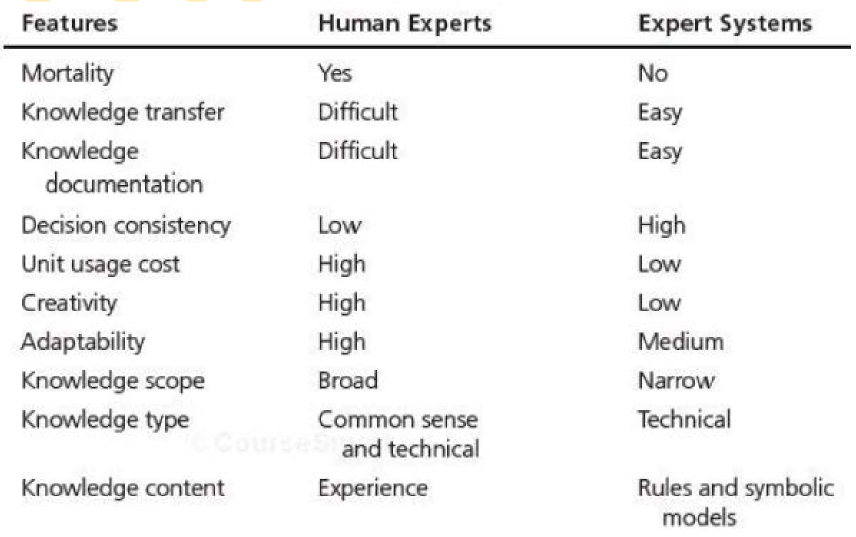
Artificial Intelligence Versus Natural Intelligence

- The potential value of artificial intelligence can be better understood by contrasting it with natural, or human, intelligence. AI has several important advantages over natural intelligence:
  - AI is more permanent. Natural intelligence is perishable from a commercial standpoint, in that Workers can change their place of employment or forget information.
  - AI offers ease of duplication and dissemination. when knowledge is embedded in a computer system, it can easily be transferred from that computer to any other computer on the Internet or on an intranet.
  - AI can be less expensive than natural intelligence. There are many circumstances in which buying computer services costs less than having corresponding human power carry out the same tasks.
  - AI. being a computer technology, is consistent and thorough.
  - AI can be documented. Decisions made by a computer can be easily documented by tracing the activities of the system. Natural intelligence is difficult to document.
  - AI can execute certain tasks much faster than a human can.
  - AI can perform certain tasks better than many or even most people.
  - **Natural intelligence does have some advantages over AI, such as the following:**
    - Natural intelligence is truly creative, whereas AI is uninspired. The ability to acquire knowledge is inherent in human beings, but with AI knowledge must be built into a carefully constructed system constrained by a large number of assumptions.
    - Natural intelligence enables people to benefit from and use sensory experience directly in a synergistic way, whereas most AI systems must work with numeric and/or symbolic inputs in a sequential manner with predetermined representational forms.

BASIC CONCEPTS OF EXPERT SYSTEMS

- **Expert systems (ES)** are computer-based information systems that use expert knowledge to attain high-level decision performance in a narrowly defined problem domain. ES has also been used in taxation, credit analysis, equipment maintenance. help desk automation, environmental monitoring, and fault diagnosis. ES has been popular in large and medium-sized organizations as a sophisticated tool for improving productivity and quality.

- **Experts**
  - An expert is a person who has the special knowledge, judgment, experience, and skills to put his or her knowledge in action to provide sound advice and to solve complex problems in a narrowly defined area.
- **human experts are capable of doing the following:**
  - Recognizing and formulating a problem
  - Solving a problem quickly and correctly
  - Explaining a solution
  - Learning from experience
  - Restructuring knowledge
  - Breaking rules (i.e., going outside the general norms), if necessary
  - Determining relevance and associations
  - Declining gracefully (i.e., being aware of one's limitations)
- **Expertise**
  - Expertise is the extensive, task-specific knowledge that experts possess. The level of expertise determines the performance of a decision. Expertise is often acquired through training, reading, and experience in practice.
- **Features of ES**
  - **ES must have the following features:**
    - An ES must possess expertise that enables it to make expert-level decisions. The system must exhibit expert performance with adequate robustness.
    - Symbolic reasoning. The basic rationale of artificial intelligence is to use symbolic reasoning rather than mathematical calculation.
- **Deep knowledge.** Deep knowledge concerns the level of expertise in a knowledge base.
- **Self-knowledge.** ES must be able to examine their own reasoning and provide proper explanations as to why a particular conclusion was reached.
- **Differences Between Human Experts and Expert Systems**



Features	Human Experts	Expert Systems
Mortality	Yes	No
Knowledge transfer	Difficult	Easy
Knowledge documentation	Difficult	Easy
Decision consistency	Low	High
Unit usage cost	High	Low
Creativity	High	Low
Adaptability	High	Medium
Knowledge scope	Broad	Narrow
Knowledge type	Common sense and technical	Technical
Knowledge content	Experience	Rules and symbolic models

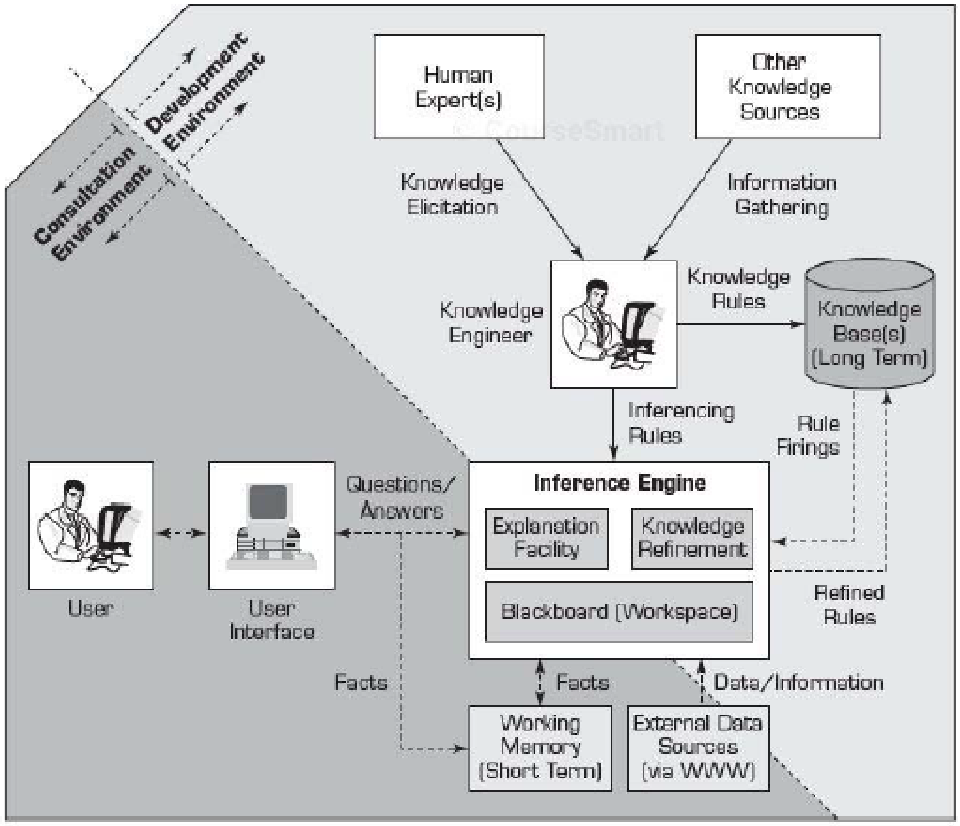
APPLICATIONS OF EXPERT SYSTEMS

- **DENDRAL:** It uses a set of knowledge- or rule-based reasoning commands to deduce the likely molecular structure of organic chemical compounds from known chemical analyses and mass spectrometry data.
- **MYCIN** is a rule-based ES that diagnoses bacterial infections of the blood.
- **XCON** a rule-based system developed at Digital Equipment Corp., used rules to help determine the optimal system configuration that fit customer requirements.
- **CREDIT ANALYSIS SYSTEMS ES** have been developed to support the needs of commercial lending institutions.
- **PENSION FUND ADVISORS** Nestle Foods Corporation has developed an ES that provides information on an employee's pension fund status. The system maintains an up-to-date knowledge base to give participants advice concerning the impact of regulation changes and conformance with new standards.
- **AUTOMATED HELP DESKS** This browser-based tool enables small businesses to deal with customer requests more efficiently. Incoming e-mails automatically pass into HelpDesk's business rule engine. The messages are sent to the proper technician, based on defined priority and status.
- **Areas for ES Applications**
- As indicated in the preceding examples, ES have been applied commercially in a number of areas, including the following:
    - **Finance.** Finance ES include insurance evaluation, credit analysis, tax planning, fraud prevention, financial report analysis, financial planning, and performance evaluation.
    - **Data processing.** Data processing ES include system planning, equipment selection, equipment maintenance, vendor evaluation, and network management.
    - **Marketing.** Marketing ES include customer relationship management, market analysis, product planning, and market planning.
    - **Human resources.** Examples of human resources ES are human resources planning, performance evaluation, staff scheduling, pension management, and legal advising.
    - **Manufacturing.** Manufacturing ES include production planning, quality management, product design, plant site selection, and equipment maintenance and repair.
    - **Homeland security.** Homeland security ES include terrorist threat assessment and terrorist finance detection.
    - **Business process automation.** ES have been developed for help desk automation, call center management, and regulation enforcement.
    - **Health care management**. ES have been developed for bioinformatics and other health care management issues.

STRUCTURE OF EXPERT SYSTEMS

- **ES can be viewed as having two environments:**
    - **the development environment and the consultation environment**
    - An ES builder uses the **development environment** to build the necessary components of the ES and to populate the knowledge base with appropriate representation of the expert knowledge.
    - A nonexpert uses the **consultation environment** to obtain advice and to solve problems using the expert knowledge embedded into the system.
- The **three major components** that appear in virtually every ES are the knowledge base, the inference engine, and the user interface. In general, though, an ES that interacts with the user can contain the following additional components:
    - **Knowledge acquisition subsystem**

    - Knowledge acquisition is the accumulation, transfer, and transformation of problem solving expertise from experts or documented knowledge sources to a computer program for constructing or expanding the knowledge base. The knowledge base is the foundation of an ES. It contains the relevant knowledge necessary for understanding, formulating, and solving problems A typical knowledge base.
- **Blackboard (workplace)**
    - The blackboard is an area of working memory set aside as a database for description of the current problem, as characterized by the input data. It is also used for recording intermediate results, hypotheses, and decisions. Three types of decisions can be recorded on die blackboard: a plan (i.e., how to attack the problem), an agenda (i.e., potential actions awaiting execution), and a solution (i.e., candidate hypotheses and alternative courses of action that the system has generated thus far).
- **Explanation subsystem (justifier)**
    - Explanation Subsystem (Justifier) The ability to trace responsibility for conclusions to their sources is crucial both in the transfer of expertise and in problem solving. The explanation subsystem can trace such responsibility and explain the ES behavior by interactively answering questions.
- **Knowledge-refining system**
    - Human experts have a knowledge-refining system: that is, they can analyze their own knowledge and its effectiveness, learn from it, and improve on it for future consultations.



-

KNOWLEDGE ENGINEERING

- The collection of intensive activities encompassing the acquisition of knowledge from human experts (and other information sources) and conversion of this knowledge into a repository (commonly called a knowledge base) are called knowledge engineering.
- **five major activities in knowledge engineering:**
  - Knowledge acquisition. Knowledge acquisition involves the acquisition of knowledge from human experts, books, documents, sensors, or computer files.
  - Knowledge representation. Acquired knowledge is organized so that it will be ready for use, in an activity called knowledge representation. This activity involves.
  - Knowledge validation. Knowledge validation (or verification) involves validating and verifying the knowledge (e.g., by using test cases) until its quality is acceptable.
  - Inferencing. This activity involves the design of software to enable the computer to make inferences based on the stored knowledge and the specifics of a problem.
  - Explanation and justification. This step involves the design and programming of an explanation capability (e.g., programming the ability to answer questions such as why a specific piece of information is needed by the computer or how a certain conclusion was derived by the computer).

DEVELOPMENT OF EXPERT SYSTEMS

- The development of ES is a tedious process and typically includes defining the nature and scope of the problem, identifying proper experts, acquiring knowledge, selecting the building tools, coding the system, and evaluating the system.
- **Identifying Proper Experts**
  - After the nature and scope of the problem have been clearly defined, the next step is to find proper experts who have the knowledge and are willing to assist in developing the knowledge base.
- **Acquiring Knowledge**
  - After identifying helpful experts, it is necessary to start acquiring decision knowledge from them. The process of eliciting knowledge is called knowledge engineering. The person who is interacting with experts to document the knowledge is called a knowledge engineer.
- **Selecting the Building Tools**
  - After the knowledge base is built, the next step is to choose a proper tool for implementing the system.