**AMBO UNIVERSITY**



**WOLISo CAMPUS**

DEPARTMENT OF COMPUTER SCIENCE

SELECTED TOPICS IN COMPUTER SCIENCE HAND OUT FOR 4RTH  YEAR REGULAR STUDENTS

PREPARED BY CHALTU M.

MAY 25,2020

AMBO,ETHIOPAI

Table of Contents

[Chapter 1: Natural Language Processing (NLP) 1](#_Toc41717415)

[1.1 Definition of NLP 1](#_Toc41717416)

[1.2 Natural Language Processing 1](#_Toc41717417)

[1.3 Importance of NLP 3](#_Toc41717418)

[1.4 Difficulty of NLP 4](#_Toc41717419)

[1.5 Levels of Linguistic Analysis 4](#_Toc41717420)

[1.5.1 Morphology 4](#_Toc41717421)

[1.5.2 Roots, Affixes versus Combining Forms 6](#_Toc41717422)

[1.5.3 Syntactic Analysis- analyzes: 8](#_Toc41717423)

[1.5.3 Parsing 9](#_Toc41717424)

[1.5.4 Semantic Analysis 12](#_Toc41717425)

[1.5.5 Discourse 13](#_Toc41717426)

[1.5.6 Disambiguation 15](#_Toc41717427)

[1.6 Approaches to NLP 16](#_Toc41717428)

[1.7 Application of NLP 17](#_Toc41717429)

[Chapter 2: Data Mining 20](#_Toc41717430)

[2.1 What is Data Mining? 20](#_Toc41717431)

[2.2 Statistics vs. Data Mining 24](#_Toc41717432)

[2.3 Goal of Data Mining 24](#_Toc41717433)

[2.4 Application of Data Mining 27](#_Toc41717434)

[2.6 Data source for DM applications 28](#_Toc41717435)

[2.7 Data Mining Functionalities 29](#_Toc41717436)

[2.8 A Multi-Dimensional View of Data Mining Classification 33](#_Toc41717437)

[Chapter 3: Data Warehouse 34](#_Toc41717438)

[3.1 What is Data Warehouse? 34](#_Toc41717439)

[3.2 Data Warehouse vs. Operational DBMS 36](#_Toc41717440)

[3.3 OLTP (On-Line Transaction Processing) vs OLAP (On-Line Analytical processing) 36](#_Toc41717441)

[3.4 Design of a Data Warehouse: 37](#_Toc41717442)

[3.5 Typical OLAP Operations 43](#_Toc41717443)

[3.6 Data Warehouse Design Process 44](#_Toc41717444)

[3.7 Three Data Warehouse Models 44](#_Toc41717445)

[3.8 Characteristics of Data Warehouse 45](#_Toc41717446)

[3.9 Difficulties of Implementing Data warehouse: 45](#_Toc41717447)

[Chapter 4 Geographical Information system (GIS) 46](#_Toc41717448)

[4.1 Introduction 46](#_Toc41717449)

[4.2 Defining GIS 47](#_Toc41717450)

[4.3 Types of GIS 47](#_Toc41717451)

[4.3.1 Four dimensional GIS 47](#_Toc41717452)

[4.3.2 Multimedia/hypermedia GIS 48](#_Toc41717453)

[4.3.3 Web GIS 48](#_Toc41717454)

[4.3.4 Virtual Reality GIS 48](#_Toc41717455)

[4.4 Multimedia and Geographical Information System (GIS) 49](#_Toc41717456)

[4.4.1Multimedia 49](#_Toc41717457)

[4.5 GIS subsystem 50](#_Toc41717458)

[2. Data storage editing and retrieval subsystem 51](#_Toc41717459)

[4.6 Application of multimedia GIS 52](#_Toc41717460)

[4.6.1 Education 52](#_Toc41717461)

[4.6.2 Mapmaking 52](#_Toc41717462)

[4.6.3 Land Information 53](#_Toc41717463)

[4.7 Infrastructure and Utilities 53](#_Toc41717464)

[4.7.1 Environmental 53](#_Toc41717465)

[4.7.2 Archaeology 54](#_Toc41717466)

[4.7.3 Natural Hazards 54](#_Toc41717467)

[4.7.4 Forestry 54](#_Toc41717468)

[4.7.5 Military GIS 54](#_Toc41717469)

[4.8 Currently Available GIS software 55](#_Toc41717470)

[4.9 Futures of GIS 55](#_Toc41717471)

[Chapter 5: cloud computing 57](#_Toc41717472)

[5.1 An over view 57](#_Toc41717473)

[5.2 cloud computing model 59](#_Toc41717474)

[60](#_Toc41717475)

[5.3 Understanding Public and Private Clouds 61](#_Toc41717476)

[5.4 Cloud computing benefits 62](#_Toc41717477)

[5.5 Cloud computing challenges 63](#_Toc41717478)

# Chapter 1: Natural Language Processing (NLP)

## 1.1 Definition of NLP

**Natural Language**

**Natural language:** refers to human languages (Amharic, Afaan Oromo, Tigrigna, English, Arabic, Chinese, etc.), as opposed to artificial/programming languages such as C++,Java, Pascal, etc.

* **Natural language:** is represented using texts in spoken or written forms.

**Natural Language Processing**

* NLP is the computerized approach to analyzing text that is based on both a set of theories and a set of technologies.

## 1.2 Natural Language Processing

**A more comprehensive definition of NLP is given as:**

An interdisciplinary field of study dealing with computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.

* …...interdisciplinary field...

Several fields including linguistics, psycholinguistics, mathematics, computer science, and electrical engineering contribute to the research and development of NLP.

* .....computational techniques...

Multiple models, methods and algorithms are employed to accomplish a particular type of language analysis.

* .....naturally occurring texts...

Texts can be in spoken or written forms representing natural languages used by humans to communicate to one another.

* .....levels of linguistic analysis...

Multiple types of language processing are known to be at work when humans produce or comprehend language.

* ...human-like language processing...

NLP strives for human-like performance, and thus considered as a discipline with in Artificial Intelligence.

* ...tasks or applications...

The goal of NLP is to accomplish human-like language processing for various tasks and applications such as **machine translation**, **information retrieval**, **question-answering**, etc.

Closely related (and overlapping) fields are Natural Language Understanding and Computational Linguistics.

The field of NLP was originally referred to as Natural Language Understanding (NLU) in the early days of Artificial Intelligence. A full NLU system would be able to:

* + paraphrase an input text
  + translate the text into another language
  + answer questions about the contents of the text
  + Draw inferences from the text.

Computational Linguistics emerged as a field of study in linguistics with the purpose of providing computational models for various linguistics phenomena.

An alternative view on NLP is that it is a computer system which uses natural language as input and/or output. In this view, NLP is considered to have two distinct focuses-Natural Language Understanding and Natural Language Generation.

The task of Natural Language Understanding is equivalent to the role of reader/listener, whereas the task of Natural Language Generation is that of the writer/speaker.

## 1.3 Importance of NLP

* Natural language is the preferred medium of communication for people.
  + People communicate with each other in natural languages.
  + Scientific articles, magazines etc. are all in natural languages.
  + Billions of web pages are also in natural languages
* Computers can do useful things for us if:
  + ™Data is in structured form, e.g. databases, knowledge bases.
  + Specifications are in formal language, e.g. programming languages.
* NLP bridges the communication gap between people and computers.
  + lead to a better and a more natural communication with computers.

process an ever increasing amount of natural language data generated by people  
e.g. extract required information from web.

## 1.4 Difficulty of NLP

* People generally don’t appreciate how intelligent they are as natural language processors.
* For them natural language processing is deceptively simple because no conscious effort is required.
* Since computers are orders of magnitude faster, many find it hard to believe that computers are not good at processing natural languages.

NLP is hard because of:  
**Ambiguity**

* A word, term, phrase or sentence could mean several possible things .  
   - Computer languages are designed to be unambiguous.  
  **Variability**
* Lots of ways to express the same thing.  
   - Computer languages have variability but the equivalence of expressions can be automatically detected.

## 1.5 Levels of Linguistic Analysis

### 1.5.1 Morphology

is the study of the structure of words.

* is the study of word formation from smallest unit of words.
* At morphological level, the smallest parts of words that carry meanings, affixes are analyzed.

**Morphology** is important in NLP because language is **productive**: in any given text we will encounter words and word forms that we haven’t seen before and that are not in our precompiled dictionary.

**Morpheme**: The minimal units of morphology. e.g. helpfulness.

**Stem:** part of the word that never changes even when morphologically inflected.

**For example**, walk is the stem for the words walk, walks, walking, and walked.

**Root/Lemma** is citation form of a set of words, e.g. *break* is the root form for the words *break*, *breaks*, *breaking*, *broke*, and *broken*.

**Part-of-Speech/Lexical Category/Word Class**

is a linguistic category of words that explains how the word is used in a sentence.

* Although different languages may have different classification schemes, English and Amharic words are usually classified into eight lexical categories: noun, pronoun, adjective, verb, adverb, preposition, conjunction and interjection.
* Morphologically important parts-of-speech in English and Amharic include: **nouns, adjectives and verbs**.

**Morphological Analysis** –

the process of finding morphemes of a word.

* It is an important component of Spelling Correction, Machine Translation, Information Retrieval, Text Generation and other natural language systems.

**Morphological Generation:**

the process of generating different words from a morpheme.

**Lemmatization:** the process of finding the root/lemma of a word.

**Stemming:** the process of finding the stems of a word.

**Morphemes** can be classified in two ways:

* + **Free** versus **Bound**

### 1.5.2 Roots, Affixes versus Combining Forms

1. **Free versus Bound** 
   1. **Free morphemes** - morphemes that can stand on their own to give meaning.

e.g. Friend in friendly

Large in enlarge  
 help in helpfulness  
 perform in performance

**2. Bound morphemes** - morphemes that cannot stand on their own as a word.

e.g. -ly in friendly  
 en- in enlarge  
 -ful and -ness in helpfulness

**2. Roots, Affixes versus Combining Forms**

* **Roots** - morphemes (within a non-compound word) that makes the most precise and concrete contribution to the word’s meaning, and is either the sole morpheme or else the only one that is not an affix.  
   e.g. break in breaks  
   help in unhelpfulness
* **Affixes -** bound morphemes that either precede, follow or are inserted inside the root or stem.  
  e.g. **Prefix**: en- in enlarge is an affix that precedes the root *large* **Suffix:** -ly in *l*argely is an affix that follows the root *large* **Infix:** is an affix that is inserted inside the root.
* **Circumfix:** is an affix that precedes and follows the stem.
* **Combining Forms** - morphemes that are formed from two bound or free-like roots.

e.g. two free roots: photo and graph in photograph

two bound roots: electro- and -lysis in electrolysis

bound and free roots: Ethio- and America in Ethio-American

**The major types of morphological process:**

1. ***Inflections*** are the systematic modifications of a ***root form*** by means of prefixes and suffixes to indicate grammatical distinctions like singular and plural.

* Inflection does not change word class or meaning significantly, but varies features such as **tense**, **number**, and **plurality.**
* All the inflectional forms of a word are often grouped as manifestations of a single lexeme.

**2. Derivation** is less systematic.

* This derivation process usually changes the **part-of-speech category**.
* An example is the derivation of the adverb **widely** from the adjective **wide**.

***3. Compounding*** refers to the merging of two or more words into a new word.

* English has many noun-noun compounds, nouns that are carbonations of two other nouns.

Examples are disk drive, college degree, high school, etc.….

1. **Syntax-** refers to the way words are related to each other in a sentence.

### 1.5.3 Syntactic Analysis- analyzes:

**Syntactic Analysis-** analyzes:

* + - how words are grouped together into phrases;
    - what words modify other words;
    - what words are of central importance to the sentence.
* **Syntactic Analysis** is used in many NLP applications such as:
  + - Grammar Checking
    - Question Answering
    - Information Extraction
    - Machine Translation
* **English Noun Phrases**
  + - Student, the student, that student, two students, many students, Clever student, A student of computer science
    - **English Verb Phrases**
    - turn, turn on, is turning on, have been working
    - threatened to throw himself into the window
    - was an understandable reaction by the visitors
    - is amazingly rich in minerals

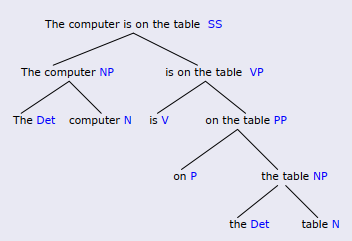
### 1.5.3 Parsing

is a derivation process which identifies the structure of sentences using a given grammar. Considered as a special case of a search problem.

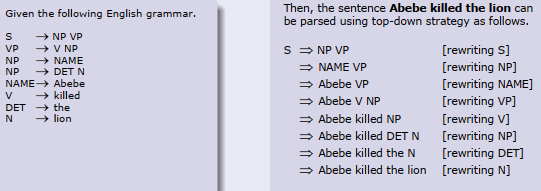
* + two basic methods of searching are used
    - * **top-down strategy**
      * **bottom-up strategy**
  + methods of improving efficiency
    - * **storing lexical rules separately**
      * **chunking**

**Simple Sentences (English)**

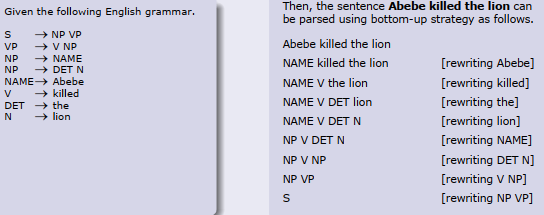
* The computer is on the table …represent using parse tree



* **Top-down parsing** starts with the symbol **S** and then searches through different ways to rewrite the symbols until the input sentence is generated.



* **Bottom-up parsing** starts with words in a sentence and uses production rules backward to reduce the sequence of symbols until it consists solely of **S.**



* **Chunking:** also called **partial parsing**, is a technique which attempts to model human parsing by breaking the text up into small pieces, each parsed separately.
* Chunk boundaries correspond roughly to the pauses in everyday speech.
* For example, consider the following sentence.

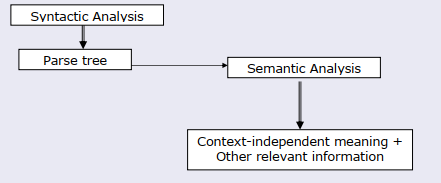
When I read a sentence, I read it a chunk at a time.

* Then, the following chunks can be identified.

[**When I read**] [**a sentence**], [**I read it**] [**a chunk**] [**at a time**].

### 1.5.4 Semantic Analysis

Involves extraction of context-independent aspects of a sentence’s meaning, including the semantic roles of entities mentioned in the sentence, and quantification information, such as cardinality, iteration, and dependency.



* **Semantic Analysis:** is an important component for many NLP applications.
* **A semantic role:** is the underlying relationship that a participant has with the main verb in a clause.Semantic roles are identified from the grammatical relations

### 1.5.5 Discourse

Deals with the properties of the text as a whole that convey meaning by making connections between component sentences.

* Much of language interpretation is dependent on the preceding discourse/dialogue.
* Discourse imposes meaning and structure on individual sentences (or utterances) that go well beyond the compositional meaning of sentences in isolation.
* The most common methods applied for discourse processing are discourse  
  segmentation and reference resolution.
* Discourse processing is required for:
* Natural language understanding , Text summarization, Machine translation, Natural language generation
* **References:** in a given text can be anaphoric or coreferential noun phrases.
* The task of reference resolution is to determine which noun phrases refer to each real world entity mentioned in the text.

**Anaphora:**

* + - An expression α1 is in an anaphoric relation with expression α2 if and only if the interpretation of α1 depends on α2.
    - The relation holds within a text.

**Reference:**

* + - Two expressions α1 and α2 are referential if and only if  
      Referent (α1) = Referent (α2).
    - The expressions can be in the same text or different texts, in the same  
      language or different language.
* Some expressions are both referential and anaphoric.

*Example: A bus had to divert to the local hospital when one of the  
passengers had a heart attack. It go to the hospital in time and the man’s life was saved.*

**Coreferential**: {*the local hospital*, *the hospita*l} , {*bus, it*},  
{*one of the passengers, the man*}

**Anaphoric**: {it}

**Pragmatics:**

is the study of how linguistic properties and contextual factors interact in the interpretation of utterances, enabling hearers to bridge the gap between sentence meaning and speaker’s meaning.

### 1.5.6 Disambiguation

* A text is said to be **ambiguous** if multiple or alternative linguistic structures can be built for it.

For example, given the following lexical entry in a lexicon

* Ambiguity may occur at:
  + - * Phonological level - multiple orthographic representations
      * Morphological level - multiple word classes
      * Syntactic level - different ways to parse the tree
      * Semantic level - different meanings of the same parse tree
      * Discourse level - different references of the same anaphora
      * Pragmatic level - cannot be clearly interpreted

## 1.6 Approaches to NLP

1. **Rule-based Approach**

* Rule-based systems are based on explicit representation of facts about language through well-understood knowledge representation schemes and associated algorithms.
* Rule-based systems usually consist of a set of rules, an inference engine, and a workspace or working memory.
* Knowledge is represented as facts or rules in the rule-based approach.
* The inference engine repeatedly selects a rule whose condition is satisfied and  
  executes the rule.
* The primary source of evidence in rule-based systems comes from human-developed rules (e.g. grammatical rules) and lexicons.
* Rule-based approaches have been used tasks such as information extraction, text  
  categorization, ambiguity resolution, and so on.

1. **Statistical Approach**

* Statistical approaches employ various mathematical techniques and often use large text corpora to develop approximate generalized models of linguistic phenomena based on actual examples of these phenomena provided by the text corpora without adding significant linguistic or world knowledge.
* The primary source of evidence in statistical systems comes from observable data (e.g. Large text corpora).
* Statistical approaches have typically been used in tasks such as speech recognition, parsing, part-of-speech tagging, statistical machine translation, statistical grammar learning, and so on.

**3. Connectionist Approach**

* **A connectionist model** is a network of interconnected simple processing units with knowledge stored in the weights of the connections between units.
* Similar to the statistical approaches, connectionist approaches also develop  
  generalized models from examples of linguistic phenomena.
* What separates connectionism from other statistical methods is that connectionist models combine statistical learning with various theories of representation.

Connectionist approaches have been used in tasks such as word-sense disambiguation, language generation, syntactic parsing, limited domain translation tasks, and so on.

## 1.7 Application of NLP

**1. Spelling Correction and Grammar Checking**

* + **Spelling Correction** is a process of detecting and sometimes providing suggestions for incorrectly spelled words in a text.
  + **Spell Checker** is an application program that flags words in a  
    document that may not be spelled correctly.
  + **Grammar Checking is** an application program that checks whether the sentence is constructed correctly or not.. subject-Verb agreement and others..

**2. Information retrieval**

* provides a list of potentially relevant documents in response to a  
  user’s query.

1. **Information Extraction**

* focuses on the recognition, tagging, and extraction of certain key elements of information (e.g. persons, companies, locations, organizations, etc.) from large collections of text into a structured representation. It has the following subtasks:
  + Named Entity Recognition: recognition of entity names.
  + Relation Detection and Classification: identification of relations between entities.
  + Coreference and Anaphoric Resolution: resolving links to previously named entities.
  + Temporal and Event Processing: recognizing temporal expressions and analyzing events.
  + Template Filling: filling in the extracted information.

**4. Machine Translation** is an automatic translation of text from one language to another.

**5. Question-Answering** provides the user with either just the text of the answer itself or answer-providing passages.

**6**. **Dialogue Systems** are agents that converse with human beings in a coherent structure using several modes of communication such as text, speech, gesture, etc.

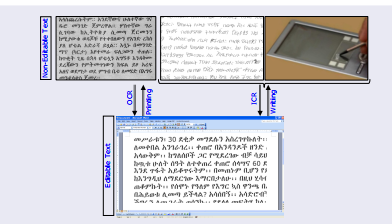
**7. Text Summarization:** reduces a larger text into a shorter, yet richly constituted representation of the original document.

**8. Speech Recognition** is the process of converting spoken words (acoustic signals) into equivalent text.

**9. Speech Synthesis,** also known as Text-to-Speech system, performs the reverse process, i.e. artificially produces human speech from a given text.

**10. Optical Character Recognition (OCR)** is a computerized system that converts non-editable text to machine-encoded text.

* If the text to be converted is handwritten, the system is also known as Intelligent Character Recognition (ICR).



# Chapter 2: Data Mining

## 2.1 What is Data Mining?

**Data mining** is extraction of interesting (*non-trivial, implicit, previously unknown and potentially useful*) information or patterns from data source. **Data mining** refers to the mining or discovery of new information in terms of patterns or rules from vast amounts of data. It is exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. Data mining can be applied to operational databases with individual transactions. To make data mining more efficient, the data warehouse should have an aggregated or summarized collection of data. Data mining helps in extracting meaningful new patterns that cannot necessarily be found by merely querying or processing data or metadata in the data warehouse.

Therefore, data mining applications should be strongly considered early, during the design of a data warehouse.

Also, data mining tools should be designed to facilitate their use in conjunction with data warehouses. In fact, for very large databases running into terabytes and even petabytes of data, successful use of data mining applications will depend first on the construction of a data warehouse. Focused on **hypothesis generation**, not on **hypothesis testing**

Alternative names:

* Knowledge discovery(mining) from databases (KDD),
* knowledge extraction,
* data/pattern analysis,
* data archeology,
* information harvesting,
* business intelligence, etc.
* Note that: query processing systems, Expert statistical data analysis or Information retrieval systems are not data mining tasks.
* Sample pattern you might find
* **Supermarket data**
  + On Thursday nights people who buy diapers also tend to buy beer
* **Insurance company data**
  + People with good credit ratings are less likely to have accidents
* **Telecom data**
  + Government lines are busy than private line

The result of mining may be to discover the following type of new information:

* **Association rules:**

for example, whenever a customer buys video equipment, he or she also buys another electronic device.

**Sequential patterns:**

for example, suppose a customer buys a camera, and within three months he or she buys photographic supplies, then within six months he is likely to buy an accessory item. This defines a sequential pattern of transactions. A customer who buys more than twice in lean periods may be likely to buy at least once during the Christmas period.

* **Classification trees:**

for example, customers may be classified by frequency of visits, types of financing used, amount of purchase, or affinity for types of items; some revealing statistics may be generated for such classes.

The results of data mining may be reported in a variety of formats, such as **listings,** **graphic**

**outputs**, **summary tables**, or **visualizations.**

* **Knowledge Discovery in Databases (**KDD),

Typically encompasses more than data mining. The knowledge discovery process comprises six phases: data selection, data cleansing, enrichment, data transformation or encoding, data mining, and the reporting and display of the discovered information.

**Knowledge**

* ***Data mining: the core of knowledge discovery in Database***

**Pattern Evaluation**

**Data Mining**

**Task-relevant Data**

**Selection & Transformation**

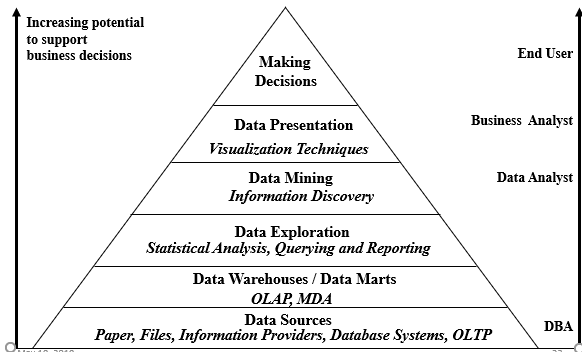
**Data Warehouse**

**Data Cleaning**

**Data Integration**

**Databases**

* **Data Mining and Business Intelligence**



## 2.2 Statistics vs. Data Mining

|  |  |
| --- | --- |
| **Statistics** | **Data Mining** |
| Confirmative | Explorative |
| Small data sets/File-based | Large data sets/Databases |
| Small number of variables | Large number of variables |
| Deductive | Inductive |
| Numeric data | Numeric and non-numeric (including txt, networks) |
| Clean data | Data cleaning |

## 2.3 Goal of Data Mining

Data mining is typically carried out with some end goals or applications. Broadly speaking, these goals fall into the following classes: **prediction, identification, classification, and optimization.**

**1 Prediction.**

Data mining can show how certain attributes within the data will behave in the future.

* Examples of predictive data mining include the analysis of buying transactions to predict what consumers will buy under certain discounts, how much sales volume a store will generate in a given period, and whether deleting a product line will yield more profits.
* In such applications, business logic is used coupled with data mining

**2.Identification.**

* Data patterns can be used to identify the existence of an item,  
  an event, or an activity.
* For example, intruders trying to break a system may be identified by the programs executed, files accessed, and CPU time per session.
* In biological applications, existence of a gene may be identified by certain sequences of nucleotide symbols in the DNA sequence.
* The area known as authentication is a form of identification.
* It ascertains whether a user is indeed a specific user or one from an authorized class, and involves a comparison of parameters or images or signals against a database.

**3. Classification.** Data mining can partition the data so that different classes or categories can be identified based on combinations of parameters.

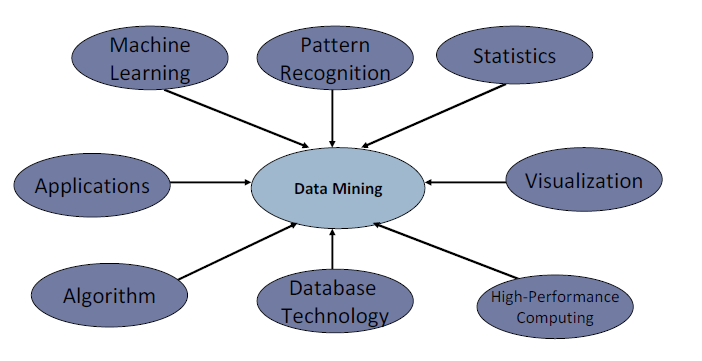
* For example, customers in a supermarket can be categorized into discount seeking shoppers, shoppers in a rush, loyal regular shoppers, shoppers attached to name brands, and infrequent shoppers.

This classification may be used in different analyses of customer buying transactions as a post mining activity. Sometimes classification based on common domain knowledge is used as an input to decompose the mining problem and make it simpler.

* For instance, health foods, party foods, or school lunch foods are distinct categories in the supermarket business.

It makes sense to analyze relationships within and across categories as separate problems. Such categorization may be used to encode the data appropriately before subjecting it to further data mining.

1. **Optimization.**

* One eventual goal of data mining may be to optimize the use of limited resources such as time, space, money, or materials and to maximize output variables such as sales or profits under a given set of constraints. As such, this goal of data mining resembles the objective function used in operations research problems that deals with optimization under constraints. The term data mining is popularly used in a very broad sense.
* ****In some situations, it includes **statistical analysis** and **constrained optimization** as well as **machine learning**. There is no sharp line separating data mining from these disciplines.

## 2.4 Application of Data Mining

Data mining technologies can be applied to a large variety of decision-making contexts in business. In particular, areas of significant payoffs are expected to include the following:

* **Marketing.**

Applications include analysis of consumer behavior based on buying patterns; determination of marketing strategies including advertising, store location, and targeted mailing; segmentation of customers, stores, or products; and design of catalogs, store layouts, and advertising campaigns.

* **Finance.**

Applications include analysis of creditworthiness of clients, segmentation of account receivables, performance analysis of finance investments like stocks, bonds, and mutual funds; evaluation of financing options; and fraud detection.

* **Manufacturing.**

Applications involve optimization of resources like machines, manpower, and materials; and optimal design of manufacturing processes, shop-floor layouts, and product design, such as for automobiles based on customer requirements.

* **Health Care**

Applications include discovery of patterns in radiological images, analysis of microarray (gene-chip) experimental data to cluster genes and to relate to symptoms or diseases, analysis of side effects of drugs and effectiveness of certain treatments, optimization of processes within a hospital, and the relationship of patient wellness data with doctor qualifications.

**2.5 Challenges in Data Mining**

* Efficiency and scalability of data mining algorithms
* Parallel, distributed, stream, and incremental mining methods
* Handling high-dimensionality
* Handling noise, uncertainty, and incompleteness of data
* Incorporation of constraints, expert knowledge, and background knowledge
* Pattern evaluation and knowledge integration
* Mining diverse and heterogeneous kinds of data: e.g., bioinformatics, Web,
* Application-oriented and domain-specific data mining
* Invisible data mining (embedded in other functional modules)
* Protection of security, integrity, and privacy in data mining

## 2.6 Data source for DM applications

* Where are the data sources for analysis?
  + - Credit card transactions,
    - loyalty cards,
    - discount coupons,
    - customer complaint calls,
    - Customer calls
    - Log files
    - Transaction files etc.

Currently, commercial data mining tools use several common techniques to extract  
knowledge. These include association rules, clustering, neural networks, sequencing, and statistical analysis.

* The best-known tool for data mining applications is **Weka**

## 2.7 Data Mining Functionalities

Data mining functionalities are used to specify the kind of patterns to be found in data mining task. Data mining task can be broadly classified into two as:

* Descriptive
* Predictive

1. **Descriptive data mining**

task characterize the general properties of the data in a database. For example, one can say:

* Ethiopia’s weather is selected to leave in for many birds
* The past 10 years rainfall of Ethiopia is appropriate for the agriculturalist in southern Shewa
* All mobile callers make few calls to wired lines than mobile recipients

1. **Predictive data mining**

task perform interpretation on the current data in order to make prediction to the future reference. For example, one can say

* A person loves to leave in Ethiopia if he/she was in ASIA for the last two years
* It will rain in Ambo within two days if there is a wind from Mediterranean see in west - east direction and average current temperature at Ambo is bellow 20oc

The kind of pattern to be mined form a given data is not known for the user (hence it is hypothesis generation not hypothesis proving). Techniques should be implemented to extract various pattern from the available data so that user can choose what they need to use.

There are different kinds of data mining functionalities that can be used to extract various types of pattern from data. This are:

* Concept /class description: Characterization and discrimination
* Association Analysis
* Classification and prediction
* Clustering analysis
* Outlier analysis
* Evolution analysis

1. **Concept/class description: Characterization and discrimination**

* Given a class/classes with data that belongs to the class, describe the class by making observation of its members.
* Hence one can describe individual classes or concepts in a **summarized, concise and yet precise** terms which is called **class/concept description.**
* These descriptions can be derived via data characterization or data discrimination or both

**Data characterization**

refers to summarizing the data of the class under consideration (target class) in general term.

* For example, one may characterize the item class as a class in which 90% of the objects are computer and its peripheral.

**Data discrimination**

is description made by making comparative analysis between the target class with the other comparative class (contrasting classes)

* For example, one may discriminate item class from other class like customer and order class by saying the item class attributes get modified more frequently than others.

1. **Association Analysis**

Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data.

Association rules are of the form X🡺Y [Support = s%, confidence = c%] where X is conjunctions of attributes and Y is conjunctions of values and interpreted as if X then it is likely to happen Y with support s% and confidence c%.

* For example
  + age(X, “20..29”) ^ income(X, “20..29K”) 🡺 buys(X, “PC”) [support = 2%, confidence = 60%]
  + Interpreted as anyone whose age ranges from 20 to 29 and income range is from 20 to 29K likely buy PC with support 2% and confidence of 60%
* **Support** shows the probability that all the predicates in X and Y fulfill together. i.e. P(X U Y)
* **Confidence** shows if predicates in X fulfilled then the predicate in Y is also fulfilled with the stated percentage. i.e. P(Y | X).

1. **Classification and Prediction**

**Classification**

is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts for the purpose of being able to use the model to predict the class of an object whose class is unknown.

* The derived class is based on training data set and can be represented in various forms such as classification IF—THEN rule, decision tree, mathematical formulae or neural networks.

**Prediction** is the process of predicting some missing or unavailable *data values* rather than class labels.

* Finding models (functions) that describe and distinguish classes or concepts for future prediction

1. **Cluster analysis**

In cluster Analysis, class labels are unknown and a group of data is given to be classified. Cluster analysis group data to form new classes, e.g., cluster houses to find distribution patterns.

Clustering based on the principle:

* + maximizing the intra-class similarity and minimizing the inter-class similarity

1. **Outlier analysis**

Database may contain data object that do not comply with the general behavior or model of the data. These data objects are **outliers.** Usually outlier data items are considered as noise or exception in many data mining applications

* + However, in some application such as fraud detection, the rare events can be more interesting than the more regularly occurring ones.
  + The analysis of outlier data is referred to as **outlier mining**

## A Multi-Dimensional View of Data Mining Classification

Different views, different classifications:

* Kinds of Databases to be mined
* Kinds of Knowledge to be mined
* Kinds of Techniques utilized
* Kinds of Applications adapted

# Chapter 3: Data Warehouse

## 3.1 What is Data Warehouse?

**Data warehouse**

is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.

A decision support ***database*** that is maintained separately from the organization’s operational database and supports information processing by providing a solid platform of consolidated, historical data for analysis.

It is a repository of multiple heterogeneous data sources organized under a unified schema at a single site in order to facilitate management decision making.

**Data warehouse** allows “**knowledge workers**” (such as **managers**, **analysts**, and **executives**) to use the warehouse to quickly and conveniently obtain an overview of the data and to make **sound decision** based on information in the warehouse. **Data warehousing** is the process of constructing and using data warehouses. The **goal** of a data warehouse is **to support decision making with data.**

**Data Warehouse: Subject-Oriented**

* Organized around major subjects, such as customer, product, sales.
* Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
* Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.
* Constructed by integrating multiple, ***heterogeneous data sources***
  + relational databases, flat files, on-line transaction records
* Data cleaning and data integration techniques are applied.
  + Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.

**Data Warehouse: Time Variant**

* The time horizon for the data warehouse is significantly longer than that of operational systems.
  + Operational database: current value data.
  + Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
* Every key structure in the data warehouse
  + Contains an element of time, explicitly or implicitly
  + But the key operational data may or may not contain “time element”.

**Data Warehouse: Non-Volatile**

* A physically separate store of data transformed from the operational environment.
* Operational update of data does not occur in the data warehouse environment.
  + Does not require transaction processing, recovery, and concurrency control mechanisms
  + Requires only two operations in data accessing:
    - **initial loading of data** and **access of data**

## 3.2 Data Warehouse vs. Operational DBMS

* **DBMS - tuned for OLTP (Online Transactional Processing):** 
  + access methods, indexing, concurrency control, recovery mechanism are desirable
* **Warehouse - tuned for OLAP:** 
  + Complex OLAP queries, multidimensional view, consolidation are desirable.
  + Indexing, concurrency control, recovery mechanism are not desirable in warehouse

## 3.3 OLTP (On-Line Transaction Processing) vs OLAP (On-Line Analytical processing)

* **OLTP (On-Line Transaction Processing)**
  + Major task of traditional relational DBMS
  + Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
* **OLAP (On-Line Analytical Processing)**
  + Major task of data warehouse system
  + Data analysis and decision making
* OLTP and OLAP differs in
  + User and system orientation
  + Data contents they operate
  + Database design used
  + View
  + Data Access patterns

|  |  |  |
| --- | --- | --- |
|  | OLTP | OLAP |
| Users | Clerk, IT professionals | Knowledge worker |
| Function | Day to day operations | Decision support |
| DB Design | Application-oriented | Subject-oriented |
| Data | Current, up-to-date detailed, flat relational isolated | Historical, summarized, multidimensional integrated, consolidated |
| Usage | Repetitive | Ad-hoc |
| Access | Read/write index/hash on pri. key | Lots of scan |
| Unit of Work | Short, simple transaction | Complex query |
| #records accessed | Tens | Millions |
| #users | Thousands | Hundreds |
| DB size | 100MB-GB | 100GB-MG |
| Metric | Transaction throughput | Query throughput, response |

## 3.4 Design of a Data Warehouse:

A Business Analysis Framework

The basic steps involved in the design process of data warehouse mainly involves business analysis framework which give clear understanding of what can a business analyst gain from having a data warehouse?

Some of the gains may include:

* Provide a competitive advantage by presenting relevant information
* Enhance business productivity as it enables to quickly and efficiently gather information that accurately describe the organization
* Facilitate customer relationship management by providing consistent view of customers and items across all lines of business, all departments and all markets

**From Tables and Spreadsheets to Data Cubes**

A data warehouse is based on a multidimensional data model which views data in the form of a **data cube.** **A data cube** allows data to be modeled and viewed in multiple dimensions. A data cube is modeled around a central team like sales which is maintained by a table called **fact table. Dimensions** are the perspective of entities with respect to which an organization wants to keep records.

**For example:**

* Records of **store sales** can be maintained with respect to the dimension time(day, week, month, quarter, year), item(item\_name, brand, type), branch, and location
* **Fact table** contains measures (such as dollars\_sold, unit sold, amount\_budgeted) and keys to each of the related dimension tables where
  + - Dollar sold refers to the amount of money sold
    - Unit sold refers to the number of items sold
    - Amount budgeted refers to the amount of money planned
* Consider the amount of money collected in Birr at Bontu Supermarket at different branches
* Branch: **Ambo**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Time** | | | | | | | |
|  |  | Mon | Tue | Wed | Thu | Fri | Sat | Sun |
| **Item** | Chocolate | 20 | 19 | 21 | 34 | 30 | 35 | 28 |
| Alcoholic Drink | 80 | 74 | 45 | 87 | 90 | 99 | 91 |
| canned foods | 67 | 68 | 63 | 55 | 64 | 52 | 55 |
| Soft drink | 44 | 60 | 63 | 54 | 64 | 45 | 54 |
| Baby diaper | 45 | 54 | 55 | 65 | 65 | 54 | 67 |

* Branch = **Guder**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Time** |  |  |  |  |  |  |
|  |  | **Mon** | **Tue** | **Wed** | **Thu** | **Fri** | **Sat** | **Sun** |
| **Item** | Chocolate | 43 | 45 | 34 | 78 | 54 | 34 | 19 |
| Alcoholic Drink | 45 | 43 | 26 | 33 | 54 | 71 | 31 |
| canned foods | 22 | 76 | 34 | 34 | 91 | 42 | 21 |
| Soft drink | 41 | 53 | 94 | 54 | 29 | 61 | 42 |
| Baby diaper | 76 | 34 | 89 | 67 | 18 | 27 | 53 |

* Branch= **Ginchi**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Time** |  |  |  |  |  |  |
|  |  | Mon | Tue | Wed | Thu | Fri | Sat | Sun |
| **Item** | Chocolate | 34 | 54 | 43 | 87 | 45 | 43 | 91 |
| Alcoholic Drink | 54 | 34 | 62 | 33 | 45 | 27 | 13 |
| canned foods | 22 | 67 | 43 | 43 | 19 | 24 | 12 |
| Soft drink | 14 | 35 | 49 | 45 | 92 | 16 | 24 |
| Baby diper | 67 | 43 | 98 | 76 | 81 | 72 | 35 |

* This data can be seen at various granularity such as amount of money per day, per week, for coca cola, sprite, biscuits, etc.
* The above three tables can be seen as sub-cuboids of the cube shown below.

55

21

12

44

63

54

64

45

54

60

45

55

65

65

54

67

54

67

63

55

64

52

55

68

80

45

87

90

99

91

74

20

21

34

30

35

28

19

20

19

21

34

30

35

28

43

45

34

54

34

19

34

54

43

87

45

43

91

28

19

91

67

53

35

54

42

24

91

31

13

**Item**

**Branch**

checholet

Alcoholic drink

Canned food

Ginchi

Soft drink

Guder

Bay Dipper

Ambo

**Time**

Sun

Thur

Sat

Tue

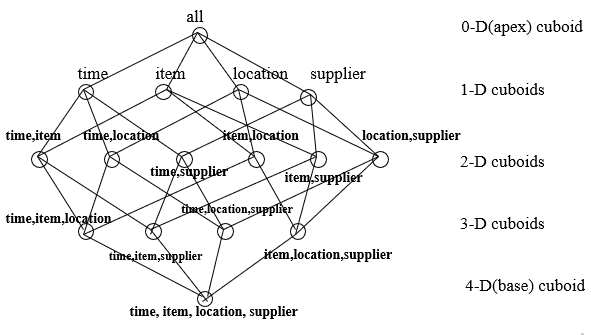
Wed

Fri

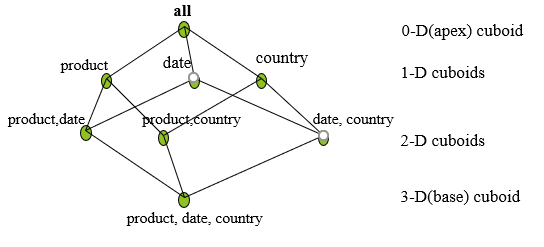
Mon

* As data warehouse can be seen from various views. In data warehousing literature, an n dimensional (n-D) cube is called **a base cuboid.** Base cuboid shows some information about every attribute at most refined granularity. The top most 0-D cuboid, which holds the highest-level of summarization, is called **the apex cuboid.** This shows the most summarized information which is free from any attribute. Lattice is formed by systematically arranging the possible cuboid and their relationship. The lattice of cuboids forms a data cube.
* Example of a lattice with four dimensions *(time, item, location, supplier).* The fact and dimension table model will be discussed soon.

**Cube:** A Lattice of Cuboids with four dimensions



**Example2:** lattice of Cuboids with three dimensions

****

**Concept Hierarchy**

**Dimensions** are organized into concept hierarchies. **A concept hierarchy** defines a sequence of mappings from a set of low-level concepts to higher-level and more general concepts. As shown in the concept hierarchy, each level refers to values of some type. The type of hierarchy define ordering which can be partial ordering or total ordering.

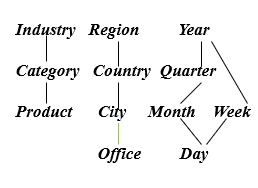
Location dimension can be seen as a total ordering

***continent 🡪 country 🡪 Region 🡪 Zone 🡪 city 🡪 kifle ketema 🡪 kebele***

Time dimension shows partial ordering

***second🡪 minute 🡪 hour 🡪 day 🡪 {month🡪quarter, week}🡪year***

More ordering



Many concept hierarchies are implicit within the database schema as location and time are described by the fields shown above.

define dimension time as (time\_key, day, day\_of\_week, month, quarter, year)

These are called **schema hierarchy.** The concept hierarchy for location is schema hierarchy whereas for annual income concept hierarchy may be set as grouping hierarchy. Concept hierarchies may also be defined by discretizing or grouping values for a given dimension or attribute resulting in a **set-grouping hierarchy**

## 3.5 Typical OLAP Operations

In multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple level of abstraction defined by concept hierarchies. This organization provides users with flexibility to view data from different perspectives. Different OLAP data cube operations exist to materialize these views:

* Roll up (drill-up)
* Drill down (roll down)
* Slice and dice
* Pivot (rotate)

**Roll up (drill-up):** Data is summarized with increasing generalization (for example, weekly to quarterly to annually). For example: from cities to countries, form second to minute.

**Drill down (roll down):** reverse of roll-up: from higher level summary to lower level summary or detailed data, or introducing new dimensions. For example: from region to town, from year to month.

**Slice:** performs selection on one dimension of a given cube resulting in a sub-cube (say time = Q1 from different dimensions like time, location and item).

**Dice:** performs defines a sub cube by performing a selection on two or more dimension.

Dice for(location=”Ambo” or “ Wolliso” and Time =”Q1” or “Q2”, Item=”Mobile”, or “Compuer”)

From (location (Ambo, Wolliso, Ginch, Guder, Bako), Time=(quarters(Q1,Q2,Q3,Q4)),Item(mobile, computer, stabilizer, divider, cable)).

**Pivot (rotate):** *reorient the cube, visualization, 3D to series of 2D planes.*

## 3.6 Data Warehouse Design Process

Data warehouse design process consists of 4 steps

1. Choosing a **business process** to model, e.g., orders, invoices, sales, shipment, inventory, account administration, general ledger etc.
2. Choosing **the dimensions** that will apply to each fact table record
3. Choosing the **grain (atomic level of data)** of the business process that will be represented in the fact table
4. Choosing the **measure** that will populate each fact table record

## 3.7 Three Data Warehouse Models

From the architecture point of view, there are three data warehouse models described as

**Enterprise warehouse, Data Mart, or Virtual warehouse**

1. **Enterprise warehouse:** collects all information about subjects that span the entire organization (*customers, products, sales, assets, personnel*). Requires extensive business modeling (may take years to design and build).
2. **Data Mart:** a subset of corporate-wide data that is of value to a specific group of users. Its scope is confined to specific, selected groups. For example, a marketing data mart my confine its subject to customer, product and sales. Data marts depending on the data source can be dependent or independent. **Dependent data marts** are sourced directly from the enterprise data warehouse. **Independent data marts** source can be from some operational data sources, external information providers, from data generated locally within a particular department or geographic area.
3. **Virtual warehouse:** a set of views over operational databases. Only some of the possible summary views may be materialized. Easy to build but requires excess capacity on operational database servers.

## 3.8 Characteristics of Data Warehouse

Data warehouses have the following distinctive characteristics:

* Multidimensional conceptual view
* Generic dimensionality
* Unlimited dimensions and aggregation levels
* Unrestricted cross-dimensional operations
* Dynamic sparse matrix handling
* Client-server architecture
* Multiuser support
* Accessibility
* Transparency
* Intuitive data manipulation
* Consistent reporting performance
* Flexible reporting

## 3.9 Difficulties of Implementing Data warehouse:

Construction, administration, and quality control.

* **Project management**: the design, construction, and implementation of the warehouse-is an important and challenging consideration that should not be underestimated.
* The **administration of a data warehouse** is an intensive enterprise, proportional to the size and complexity of the warehouse
* A significant issue in data warehousing is the **quality control of data**. Both quality and consistency of data are major concerns.

# Chapter 4 Geographical Information system (GIS)

## 4.1 Introduction

Geographical Information system (GIS) is computer based system that enables users to collect, store, process and analyses and present spatial data.

It provides an electronic representation of information called spatial data, about the earth’s natural and man-made features

A GIS references these real world spatial data elements to a coordinate system these features can be separated in to different layers

A GIS system stores each category of information in a separate layer for ease of maintenance, visualization and analysis

For example, layers can represent terrain characteristics, census data, demographics information, environmental and ecological data,roads,land users, rivers drainage and flood paints, and rare wildlife habitats. Different applications creates and use different layers.

A GIS can also store attribute data, which is descriptive information of the map features.

This attribute information is placed in a database separate from

## 4.2 Defining GIS

A GIS is a computer-based system that provides the following four sets of capabilities to handle georeferenced data:

1 data capture and preparation

2. Data management including store and maintenance

3. Data manipulation and analysis

4. Data representation

## 4.3 Types of GIS

The following GIS types are not necessarily manually exclusive and a GIS application can be always classified under more than one type.

### 4.3.1 Four dimensional GIS

### 

while patio-temporal geo representation can handle two dimensions space and one of time, four-dimensional GIS are designed for three dimensions of space and one of time

### 4.3.2 Multimedia/hypermedia GIS

Multimedia/hypermedia GIS allows the user to access a wide range of georeferenced multimedia data(eg., simulations, sound and videos) by selecting resources from a georeferenced image map base. A map servicing as a primary in index to multimedia data in a multimedia georepresentation is termed a hyper map. Multimedia and virtual geo-representations can be stored either in extended relational database, object database or in application-specific data stores.

### 4.3.3 Web GIS

Widespread access to the internet, the ubiquity of browsers and the explosion of commodified geographic information has made it possible to develop new forms of multimedia geo-representation on the web.

### 4.3.4 Virtual Reality GIS

Virtual Reality GIS has been developed to allow the creation, manipulation and exploration of geo-referenced virtual environments ,eg.using VRML modeling (Virtual Reality Modeling Language).

Virtual Reality GIS can be also web-based. Application includes 3D simulation for planning (to experiment with different scenarios).

## Multimedia and Geographical Information System (GIS)

### 4.4.1Multimedia

**Multimedia** is a technology that encompasses various types of data and present them in an integrated form. There are several types of data that are used by the technology. Including text, graphics, hyperlinks, images sound digital and analogue video and animation. Although many GIS have been successfully implemented. It has become quite clear that two-dimensional maps cannot precisely present multidimensional and dynamic spatial data. It seems that merging GIS and multimedia is a way growing towards accessing spatial data. It seems that GIS merging and multimedia is a way to with these issues. The latest advances in computer industry especially in hardware have led to the development of the multimedia and geographical information system(GIS) technology

Multimedia provides communication using texts, graphics, animation and video. Multimedia GIS systems is a way to overcome the limitations displayed by the technologies when they are used separately.

Multimedia can extend GIS capabilities of presenting geographic and other information.

The combination of several media often results in a powerful and richer presentation of information and ideas to stimulate interest and enhance information retention. They can also make GIS friendlier and easier to use.

## 4.5 GIS subsystem

A GIS has the following main functional subsystems these are,

1. **data input subsystem**

A data input subsystem allows the user to capture, collect and transform spatial and thematic data in to digital form. The data inputs are usually derived from a combination of hardcopy maps, aerial photographs, remotely sensed images, reports, survey documents, etc.

**GIS data types**

These data will contains maps of different details levels(maps of the country, its main cities, and villages, maps of the archaeological and historical sites etc.) photos of places and moments, video images, text(in many language),music and sound.

For more complex applications, multimedia data can be remotely sensed imagery scanned maps, digitized video clips,DTMs, one or more dimensional measurements, simulations model output and other. Most of them are complicated objects which have large data volumes, intensive processing requirements and rich semantic. The basic data types in a GIS reflect traditional data found on a map .Accordingly, GIS technology utilizes two basic types of data. These datas are:

* Spatial data
* Attribute data

**Source of data**

A wide variety of data sources exist for both spatial and attribute data.

The most common digital data sources for spatial data are:

* Hard copy map
* Aerial photo graphs
* Remotely sensed imagery
* Point data samples from surveys
* Existing data files

**Data Editing and quality Assurance**

Data editing and verification is in response to the errors that arise during the encoding of spatial and non-spatial data.

## 2. Data storage editing and retrieval subsystem

The second necessary component for a GIS is data storage and retrieval subsystem

the data storage and retrieval subsystem organize the data, spatial and attribute ,in a form ,which permits it to be quickly retrieved by the user for analysis and permits rapid and accurate update to made to the database.

* Organizing data for Analysis
* Editing and updating Data
* Data Retrieval and Querying

## 4.6 Application of multimedia GIS

### 4.6.1 Education

Education is a field where integration of multimedia and GIS can bring enormous benefits.

In addition, it will possible to individualize learning and tune it to particular preference of each student.

In this model, a teacher becomes a guide rather than repository of fact.it is a computer that takes on a role of “an infinitely patient teacher”.

### 4.6.2 Mapmaking

GIS can use and combine all layers that are available for an area, in order to produce an overlay that can be analyzed by using the same GIS.

### 4.6.3 Land Information

GIS has aided management of land information by enabling easy creation and maintenance of data for land records, land planning and land use.

## 4.7 Infrastructure and Utilities

GIS technologies are also widely applied to the planning and management of public utilities. Typical uses include management of the following services.

Electric gas, water roads, telecommunication, storm sewers, TV/FM transmitting facilities, hazard analysis and dispatch and emergency services.

### 4.7.1 Environmental

The Environmental field has long used GIS for a variety of applications that range from simple inventory and query, to map analysis and overlay, to complex spatial decision making systems.

Examples includes: Forest modeling, air /water quality modeling and monitoring, environmentally sensitive zone mapping analysis of interaction between Economic, meteorological and hydrological and Ecological change. Typical data input in to an environmental GIS include: elevation. Forest cover, and soil quality and hydrogeology coverage.

### 4.7.2 Archaeology

Archaeology, as a spatial discipline ,has used GIS in a variety of ways. At the simplest level, GIS has found applications as database management for archaeological records, with the added benefits of being able to create instant maps.

### 4.7.3 Natural Hazards

Areas vulnerable to earth quakes, floods,cyclones,storms,droughts,fire,volcano,landslides,soils erosions can be used to accurately predict feature disasters.

### 4.7.4 **Forestry**

GIS has been emerging as a strong tool for many areas for forestry, from harvesting schedules to urban forestry.

### 4.7.5 **Military GIS**

GIS offers a virtually unique ability to aggregate, automate, integrate and analyze geographical data, which further enhance the intelligence base for defense operations.

## 4.8 Currently Available GIS software

Some of the big players providing GIS software are:

* ESRI’s Arc GIS
* Autodesk’s AutoCAD map
* Autodesk’s GIS design overlay
* Intergraph’s GeoMedia Transportation

## 4.9 Futures of GIS

As discussed above many disciplines can benefit from GIS techniques.an active GIS market has resulted in lower cost and continual improvements in the hardware and software of the components of GIS. These developments will, in turn, result in a much wider application of the technology throughout government, business and industry.

It is quite likely that the future GIS system of the future will include the additional dimension of time, giving researchers the ability to examine the variation in earth processes over days, months and years. The advances in computer hardware ,software and remote sensing technology will lead to more band more GIS adopting multimedia to represent data,. These GIS systems coupled with with the multimedia technologies will result in a power full and richer presentation of information and ideas to stimulate interest and enhance information retention. The GIS of the future will also be more user friendly and accessible to the common man.

It is quite likely that the future GIS system of the future will include the additional dimension of time, giving researchers the ability to examine the variations in earth processes over days, months and years. The advances in computer hardware, software and remote sensing technology will lead to more and more GIS adopting multimedia to represent data .

these GIS systems coupled with the multimedia technologies will result in a powerful and richer presentation of information and ideas to stimulate interest and enhance information retention. The GIS of future will also be more user friendly and accessible to the common man

# Chapter 5: cloud computing

## An over view



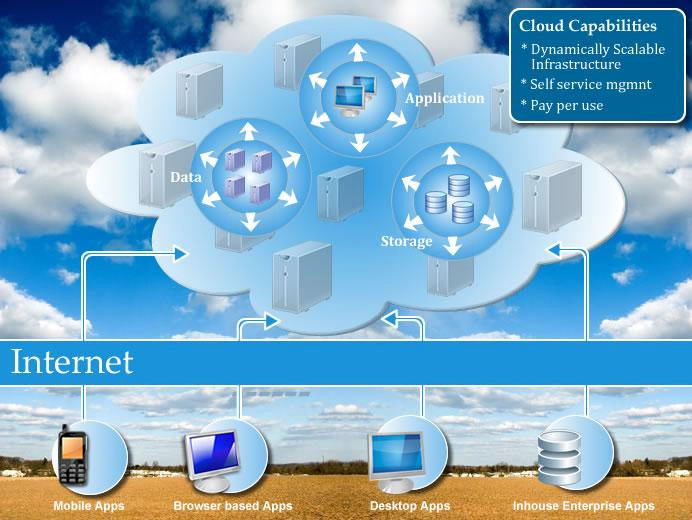
A cloud computing is a computing paradigm where a large pool of systems are connected in private or public networks, to provide dynamically scalable infrastructure for application, data and file storage. With the advent of this technology, the cost of computation, application hosting, content storage and delivery is reduced significantly.

Cloud computing is a practical approach to experience direct cost benefits and it has a potential to transform a data center from a capital-intensive set up to a variable priced environment.

The idea of cloud computing is based on a fundamental principal of “reusability of IT capabilities “. the difference that cloud computing brings compared to traditional concepts or “grid computing”, “distributed computing:, “utility computing”, or autonomic computing” is to broaden horizons across organizational boundaries.

**Forrester defines cloud computing as:**

“a pool of abstructed,highly scalable, and managed compute infrastructure capable of hosting end-customer applications and billed by consumption.



## 

Figure 1: conceptual view of cloud computing

## 5.2 cloud computing model

Cloud providers offer services that can be grouped in to three categories

1. Software as a service (SaaS): in this model a complete application is offered to the customer, as a service on demand. A single instance of a service runs on the cloud and multiple end users are serviced. On the customers’ side, there is no need for upfront investment in servers or software licenses, while for the provider, the costs are lowered, since only a single application needs to be hosted and maintained.today,SaaS is offered by companies such as Google,salesforce,Microsoft,zoho,etc.
2. Platform as a service(PaaS): here, a layer of software, or development environment is encapsulated and offered as a service, up on which other higher levels of service can be built .the customer has the freedom to build its own application, which run on the provider’s infrastructure.to meet manageability and scalability requirements of the applications,PaaS providers offer predefined combinations of OS and application servers, such as LAMP platform(Linux, Apache, MySQL PHP),restricted JEE,Ruby etc.
3. Infrastructure as a service(IaaS): IaaS provides a basic storage and computing capabilities as standardized services over the network . servers, storage system, networking equipments,data center space etc are pooled and made available to handle workloads.

The customer would typically deploy his own software on the infrastructure. Some Examples are Amazon,GoGrid,3 Tera,etc.

## 

## 

Figure 2: cloud model

## 5.3 Understanding Public and Private Clouds

Enterprises choose to deploy applications on public, private or hybrid clouds. Cloud integrators can play a vital part in determining the right cloud path for each organization.

**Public cloud**

Public clouds are, owned and operated by third parties they deliver superior economies of scale to customers, as the infrastructure cost are spread among a mix of users, giving each individual client an attractive low cost ,’pay as you go’ model. All customer share the same infrastructure pool with limited configuration, security protections, and availability variance.

These are managed and supported by the cloud providers. One of the advantage of a public cloud is that they may be larger than an enterprises cloud ,thus providing the ability to scale seamless ,on demand.

**Private cloud**

Private clouds are built exclusively for single enterprise. They aim to address concerns on data security and offer greater control, which is typically lacking in a public cloud,

Two variations to private cloud are:

* **On-premises private cloud**

It also known as Internal clouds are hosted within one’s own data center. This model provides a more standardized process and protection,but is limited in aspect of size and scalability.

* **External hosted private cloud**

This type of private cloud is hosted externally with a cloud provider,where the provider facilitate an exclusive cloud environment with full guarantee of privacy.

**Hybrid cloud**

Hybrid cloud combines both public and private cloud models. With a hybrid model, service providers can utilize 3rd party cloud providers in a full or partial manner thus increasing the flexibility of computing .the hybrid cloud environment is capable of providing on-demand, externally provisioned scale. The ability to augment a private cloud with the resources of a public cloud due to sharing of physical resources

## 5.4 Cloud computing benefits

Some of the typical benefits are:

1. **Reduced cost**

There are a number of reasons to attribute cloud technology with lower cost the billing model is pay as per usage. The Infrastructure is not purchased thus lowering maintenance. Initial expense and recurring expenses are much lower than traditional computing.

1. **Increased Storage**

With the massive Infrastructure that is offered by cloud providers today, storage and maintenance of large volume of data is a reality.

1. **Flexibility**

This is an extremely important characteristic. with enterprises having to adapt ,even more rapidly to changing business conditions ,speed to deliver is critical.

## 5.5 Cloud computing challenges

Despite its growing influence, concerns regarding cloud computing still remain. In our opinion, the benefits outweigh the drawback and the model is worth exploring. Some common challenges are:

1. **Data protection**

Data security is a crucial element that warrents scrutiny. Enterprises are reluctant to to buy an assurance of business data security from vendors. They fear losing data to competition and the data confidentiality of consumers.in many instances, the actual storage location is not disclosed,adding onto the security concerns of enterprises.in the existing model, firewalls across data centers (owned by enterprise ) protect this sensitive information. In the cloud model service providers are responsible for maintaining data security and enterprises would have to rely on them

1. **Data Recovery and Availability**

All business applications have service level agreements that are stringently followed. Operational teams play a key role in management of service level agreement and runtime governance of application. In production environments,operational teams support

In production environment ,operational teams support.

* Appropriate clustering and fall over
* Data Replication
* System monitoring(Transaction monitoring ,Logs monitoring and others)
* Maintenance(Runtime Governance)
* Disaster Recovery
* Capacity and performance management

If, any of the above mentioned services is under-served by a cloud provider, the damage and impact could be severe

1. **Management capabilities**

Despite there being multiple cloud providers, the management of platform and infrastructure is still in Infancy, Features like ‘Auto scaling’ are crucial requirements for many enterprises. There is huge potential to improve on the scalability and load balancing features provided today.

1. **Regulatory and compliance Restrictions**

In some of European countries, government regulations do not allow customers

With cloud computing, the action moves to the interface that is so, the interface between service suppliers and multiple groups of service consumers. Cloud services will demand expertise in distributed services ,procurement, risk assessment and service negotiation areas that many enterprises are only modestly equipped to handle