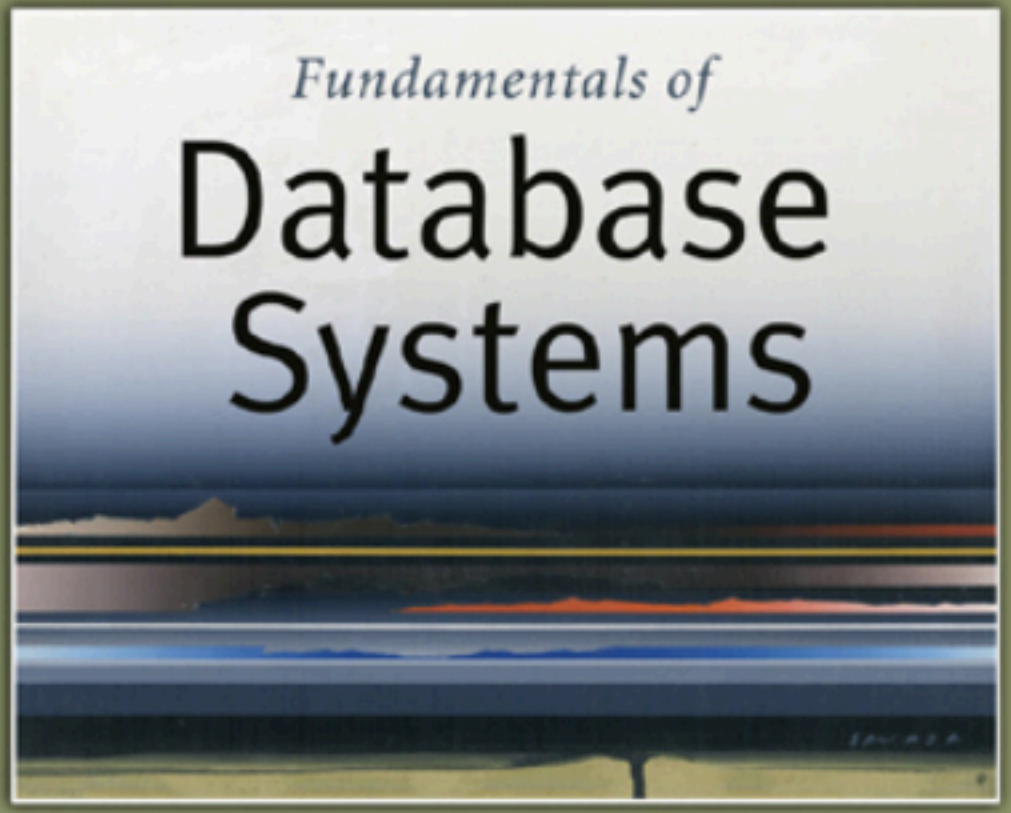
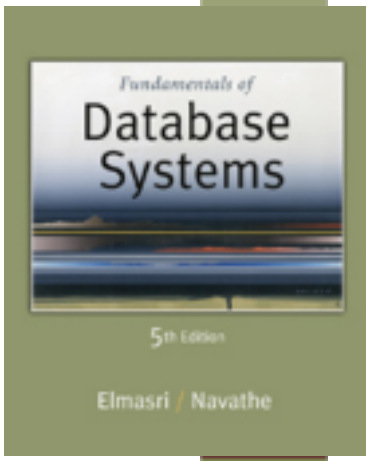


5th Edition

Elmasri / Navathe

Chapter 30

Emerging Database Technologies and Applications



Chapter Outline

- 1 Mobile Databases
- 2 Multimedia Databases
- 3 Geographic Information Systems
- 4 GENOME Data Management

Chapter Outline

- 1 Mobile Databases
  - 1.1 Mobile Computing Architecture
  - 1.2 Characteristics of Mobile Environments
  - 1.3 Data Management Issues
  - 1.4 Application: Intermittently Synchronized Databases

- 2 Multimedia Databases
  - 2.1 The Nature of Multimedia Data and Applications
  - 2.2 Data Management Issues
  - 2.3 Open Research Problems
  - 2.4 Multimedia Database Applications

Chapter Outline(contd.)

- 3 Geographic Information Systems
  - 3.1 GIS Applications
  - 3.2 Data Management Requirements of GIS
  - 3.3 Specific GIS Data Operations
  - 3.4 An Example of GIS Software: ARC-INFO
  - 3.5 Problems and Future issues in GIS

- 4 GENOME Data Management
  - 4.1 Biological Sciences and Genetics
  - 4.2 Characteristics of Biological Data
  - 4.3 The Human Genome Project and Existing Biological Database

Emerging Database Technologies and Applications

- Emerging database technologies
- The major application domains

1 Mobile Databases

- Recent advances in portable and wireless technology led to **mobile computing**, a new dimension in data communication and processing.

- Portable computing devices coupled with wireless communications allow clients to access data from virtually anywhere and at any time.
- There are a number of hardware and software problems that must be resolved before the capabilities of mobile computing can be fully utilized.
- Some of the software problems – which may involve data management, transaction management, and database recovery – have their origins in distributed database systems.

1 Mobile Databases(2)

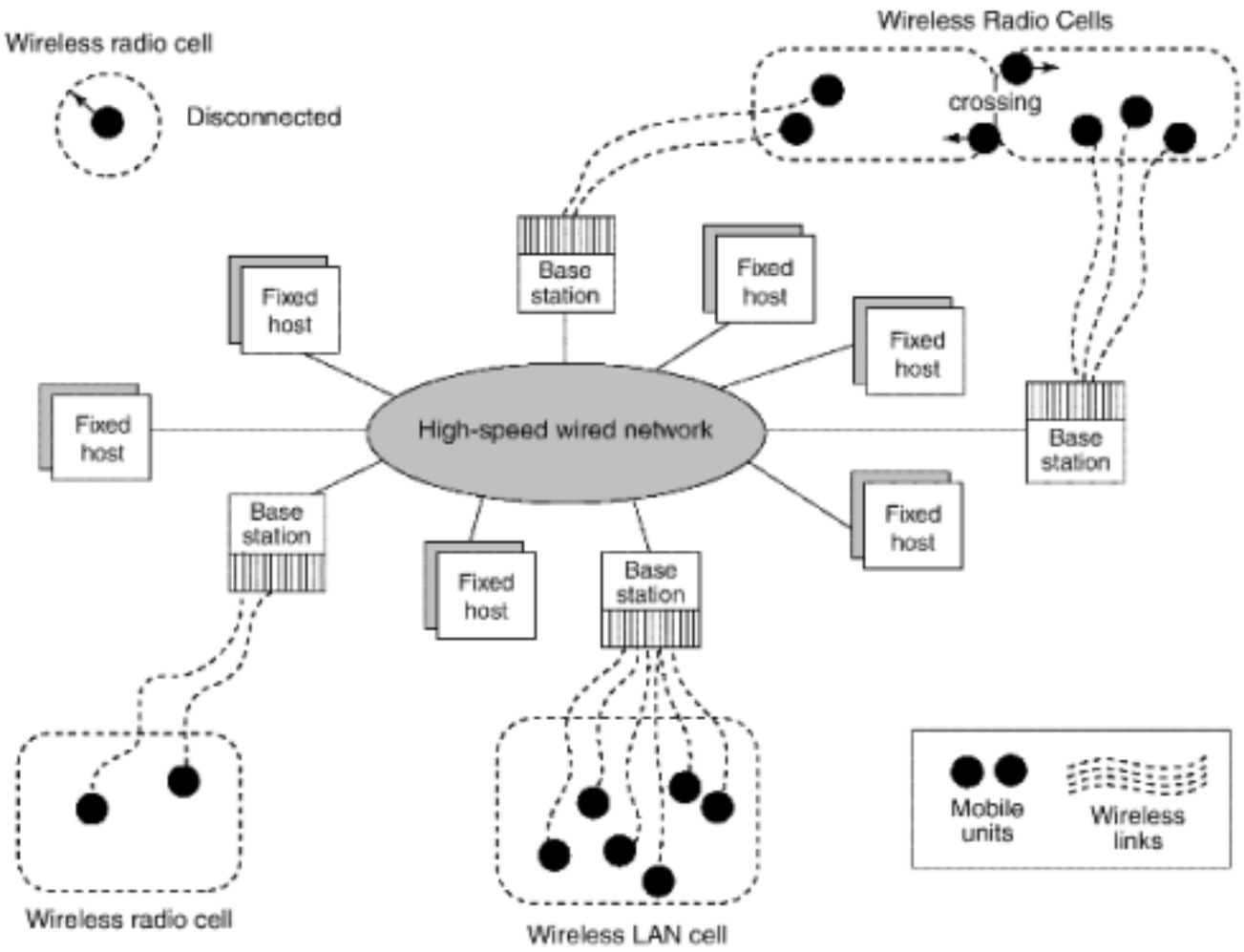
- In mobile computing, the problems are more difficult, mainly:
  - The limited and intermittent connectivity afforded by wireless

communications.

- The limited life of the power supply(battery).
- The changing topology of the network.
- In addition, mobile computing introduces new architectural possibilities and challenges.

1.1 Mobile Computing Architecture

- The general architecture of a mobile platform is illustrated in Fig 30.1.



1.1 Mobile Computing Architecture(2)

- It is distributed architecture where a number of computers, generally referred to as **Fixed Hosts** and **Base Stations** are interconnected

through a high-speed wired network.

- Fixed hosts are general purpose computers configured to manage mobile units.
- Base stations function as gateways to the fixed network for the **Mobile Units**.

1.1 Mobile Computing Architecture(3)

- **Wireless Communications –**
  - The wireless medium have bandwidth significantly lower than those of a wired network.

- The current generation of wireless technology has data rates range from the tens to hundreds of kilobits per second (2G cellular telephony) to tens of megabits per second (wireless Ethernet, popularly known as WiFi).
- Modern (wired) Ethernet, by comparison, provides data rates on the order of hundreds of megabits per second.

1.1 Mobile Computing Architecture(4)

- **Wireless Communications –**
  - The other characteristics distinguish wireless connectivity options:
    - interference,
    - locality of access,

- range,
- support for packet switching,
- seamless roaming throughout a geographical region.

1.1 Mobile Computing Architecture(5)

- **Wireless Communications –**
  - Some wireless networks, such as WiFi and Bluetooth, use unlicensed areas of the frequency spectrum, which may cause interference with other appliances, such as cordless telephones.
  - Modern wireless networks can transfer data in units called

packets, that are used in wired networks in order to conserve bandwidth.

1.1 Mobile Computing Architecture(6)

- **Client/Network Relationships –**
  - Mobile units can move freely in a **geographic mobility domain**, an area that is circumscribed by wireless network coverage.
  - To manage entire mobility domain is divided into one or more smaller domains, called **cells**, each of which is supported by at least one base station.

- Mobile units be unrestricted throughout the cells of domain, while maintaining information **access contiguity**.

1.1 Mobile Computing Architecture(7)

- **Client/Network Relationships –**
  - The communication architecture described earlier is designed to give the mobile unit the impression that it is attached to a fixed network, emulating a traditional client-server architecture.
  - Wireless communications, however, make other architectures possible. One alternative is a mobile ad-hoc network (**MANET**), illustrated in 29.2.

1.1 Mobile Computing Architecture(9)

- **Client/Network Relationships –**
  - In a **MANET**, co-located mobile units do not need to communicate via a fixed network, but instead, form their own using cost effective technologies such as Bluetooth.
  - In a **MANET**, mobile units are responsible for routing their own data, effectively acting as base stations as well as clients.
    - Moreover, they must be robust enough to handle changes in the network topology, such as the arrival or departure of other mobile units.

1.1 Mobile Computing Architecture(10)

- **Client/Network Relationships –**
  - MANET applications can be considered as peer-to-peer, meaning that a mobile unit is simultaneously a client and a server.
    - Transaction processing and data consistency control become more difficult since there is no central control in this architecture.
    - Resource discovery and data routing by mobile units make computing in a MANET even more complicated.
    - Sample MANET applications are multi-user games, shared

whiteboard, distributed calendars, and battle information sharing.

1.2 Characteristics of Mobile Environments

- The characteristics of mobile computing include:
  - **Communication latency**
  - **Intermittent connectivity**
  - **Limited battery life**
  - **Changing client location**

1.2 Characteristics of Mobile Environments(2)

- The server may not be able to reach a client.
  - A client may be unreachable because it is **dozing** – in an energy conserving state in which many subsystems are shut down – or because it is out of range of a base station.
    - In either case, neither client nor server can reach the other, and modifications must be made to the

architecture in order to compensate for this case.
- **Proxies** for unreachable components are added to the architecture.
  - For a client (and symmetrically for a server), the proxy can cache updates intended for the server.

1.2 Characteristics of Mobile Environments(3)

- Mobile computing poses challenges for servers as well as clients.
  - The latency involved in wireless communication makes scalability a problem.
    - Since latency due to wireless communications increases the

time to service each client request, the server can handle fewer clients.

- One way servers relieve this problem is by broadcasting data whenever possible.
- A server can simply broadcast data periodically. Broadcast also reduces the load on the server, as clients do not have to maintain active connections to it.

1.2 Characteristics of Mobile Environments(4)

- Client mobility also poses many data management challenges.
  - Servers must keep track of client locations in order to efficiently route messages to them.
  - Client data should be stored in the network location that

minimizes the traffic necessary to access it.

- The act of moving between cells must be transparent to the client.
- The server must be able to gracefully divert the shipment of data from one base to another, without the client noticing.
- Client mobility also allows new applications that are **location-based**.

1.3 Data Management Issues

- From a data management standpoint, mobile computing may be considered a variation of distributed computing. Mobile databases can be distributed under two possible scenarios:
  - The entire database is distributed mainly among the wired components, possibly with full or partial replication.
  - A base station or fixed host manages its own database with a

DBMS-like functionality, with additional functionality for locating mobile units and additional query and transaction management features to meet the requirements of mobile environments.

- The database is distributed among wired and wireless components.
  - Data management responsibility is shared among base stations or fixed hosts and mobile units.

1.3 Data Management Issues(2)

- Data management issues as it is applied to mobile databases:
  - Data distribution and replication
  - Transactions models
  - Query processing
  - Recovery and fault tolerance
  - Mobile database design
  - Location-based service

- Division of labor
- Security

1.4 Application: Intermittently Synchronized Databases

- Whenever clients connect – through a process known in industry as **synchronization** of a client with a server – they receive a batch of updates to be installed on their local database.
  - The primary characteristic of this scenario is that the clients are mostly disconnected; the server is not necessarily able reach them.

- This environment has problems similar to those in distributed and client-server databases, and some from mobile databases.

- This environment is referred to as **Intermittently Synchronized Database Environment** (**ISDBE**).

1.4 Application: Intermittently Synchronized Databases(2)

- The characteristics of **Intermittently Synchronized Databases** (**ISDBs**) make them distinct from the mobile databases are:
  - A client connects to the server when it wants to exchange updates.
    - The communication can be unicast –**one-on-one** communication between the server and the client– or **multicast**– one sender or server may periodically

communicate to a set of receivers or update a group of clients.

- A server cannot connect to a client at will.

1.4 Application: Intermittently Synchronized Databases(3)

- The characteristics of **ISDBs** (contd.) :
  - Issues of wireless versus wired client connections and power conservation are generally immaterial.
  - A client is free to manage its own data and transactions while it is disconnected. It can also perform its own recovery to some extent.
  - A client has multiple ways connecting to a server and, in case of

many servers, may choose a particular server to connect to based on proximity, communication nodes available, resources available, etc.

2 Multimedia Databases

- In the years ahead multimedia information systems are expected to dominate our daily lives.
  - Our houses will be wired for bandwidth to handle interactive multimedia applications.
  - Our high-definition TV/computer workstations will have access to a large number of databases, including digital libraries, image and video databases that will distribute vast amounts of multisource multimedia content.

2.1 Multimedia Databases

- DBMSs have been constantly adding to the types of data they support.
- Today many types of multimedia data are available in current systems.

2.1 Multimedia Databases(2)

- Types of multimedia data are available in current systems
  - **Text**: May be formatted or unformatted. For ease of parsing structured documents, standards like SGML and variations such as HTML are being used.
  - **Graphics**: Examples include drawings and illustrations that are encoded using some descriptive standards (e.g. CGM, PICT,

postscript).

2.1 Multimedia Databases(3)

- Types of multimedia data are available in current systems (contd.)
  - **Images**: Includes drawings, photographs, and so forth, encoded in standard formats such as bitmap, JPEG, and MPEG. Compression is built into JPEG and MPEG.
    - These images are not subdivided into components. Hence querying them by content (e.g., find all images containing circles) is nontrivial.
  - **Animations**: Temporal sequences of image or graphic data.

2.1 Multimedia Databases(4)

- Types of multimedia data are available in current systems (contd.)
  - **Video**: A set of temporally sequenced photographic data for presentation at specified rates– for example, 30 frames per second.
  - **Structured audio**: A sequence of audio components comprising note, tone, duration, and so forth.

2.1 Multimedia Databases(5)

- Types of multimedia data are available in current systems (contd.)
  - **Audio**: Sample data generated from aural recordings in a string of bits in digitized form. Analog recordings are typically converted into digital form before storage.

2.1 Multimedia Databases(6)

- Types of multimedia data are available in current systems (contd.)
  - **Composite** or mixed multimedia data: A combination of multimedia data types such as audio and video which may be physically mixed to yield a new storage format or logically mixed while retaining original types and formats. Composite data also contains additional control information describing how the

information should be rendered.

2.1 Multimedia Databases(7)

- Nature of Multimedia Applications:
  - Multimedia data may be stored, delivered, and utilized in many different ways.
  - Applications may be categorized based on their data management characteristics.

2.1 Multimedia Databases(8)

- Characterization of applications based on their data management characteristics:
  - **Repository** applications: A large amount of multimedia data as well as metadata is stored for retrieval purposes. Examples include repositories of satellite images, engineering drawings and designs, space photographs, and radiology scanned pictures.

2.1 Multimedia Databases(9)

- Characterization of applications based on their data management characteristics (contd.):
  - **Presentation** applications: A large amount of applications involve delivery of multimedia data subject to temporal constraints; simple multimedia viewing of video data, for example, requires a system to simulate VCR-like functionality. Complex and interactive multimedia

presentations involve orchestration directions to control the retrieval order of components in a series or in parallel. Interactive environments must support capabilities such as real-time editing analysis or annotating of video and audio data.

2.1 Multimedia Databases(10)

- Characterization of applications based on their data management characteristics:
  - **Collaborative** work using multimedia information: This is a new category of applications in which engineers may execute a complex design task by merging drawings, fitting subjects to design constraints, and generating new documentation, change notifications, and so forth. Intelligent healthcare networks as well as telemedicine will involve doctors collaborating among themselves, analyzing multimedia patient data and information in real time as it is generated.

2.2 Data Management Issues

- Multimedia applications dealing with thousands of images, documents, audio and video segments, and free text data depend critically on
  - Appropriate modeling of the structure and content of data
  - Designing appropriate database schemas for storing and retrieving multimedia information.

2.2 Data Management Issues(2)

- Multimedia information systems are very complex and embrace a large set of issues:
  - Modeling
    - Complex objects
  - Design
    - Conceptual, logical, and physical design of multimedia has not been addressed fully.

2.2 Data Management Issues(3)

- Multimedia information systems are very complex and embrace a large set of issues (contd.):
  - Storage
    - Multimedia data on standard disklike devices presents problems of representation, compression, mapping to device hierarchies, archiving, and buffering during the input/output operation.
  - Queries and retrieval
    - “Database” way of retrieving information is based on

query languages and internal index structures.

2.2 Data Management Issues(4)

- Multimedia information systems are very complex and embrace a large set of issues (contd.):
  - Performance
    - Multimedia applications involving only documents and text, performance constraints are subjectively determined by the user.
    - Applications involving video playback or audio-video synchronization, physical limitations dominate.

2.3 Multimedia Database Applications

- Large-scale applications of multimedia databases can be expected encompasses a large number of disciplines and enhance existing capabilities.
  - Documents and records management
  - Knowledge dissemination
  - Education and training
  - Marketing, advertising, retailing, entertainment, and travel
  - Real-time control and monitoring

3 Geographic Information Systems

- Geographic information systems(GIS) are used to collect, model, and analyze information describing physical properties of the geographical world.

3 Geographic Information Systems(2)

- The scope of GIS broadly encompasses two types of data:
  - **Spatial** data, originating from maps, digital images, administrative and political boundaries, roads,

transportation networks, physical data, such as rivers, soil characteristics, climatic regions, land elevations, and

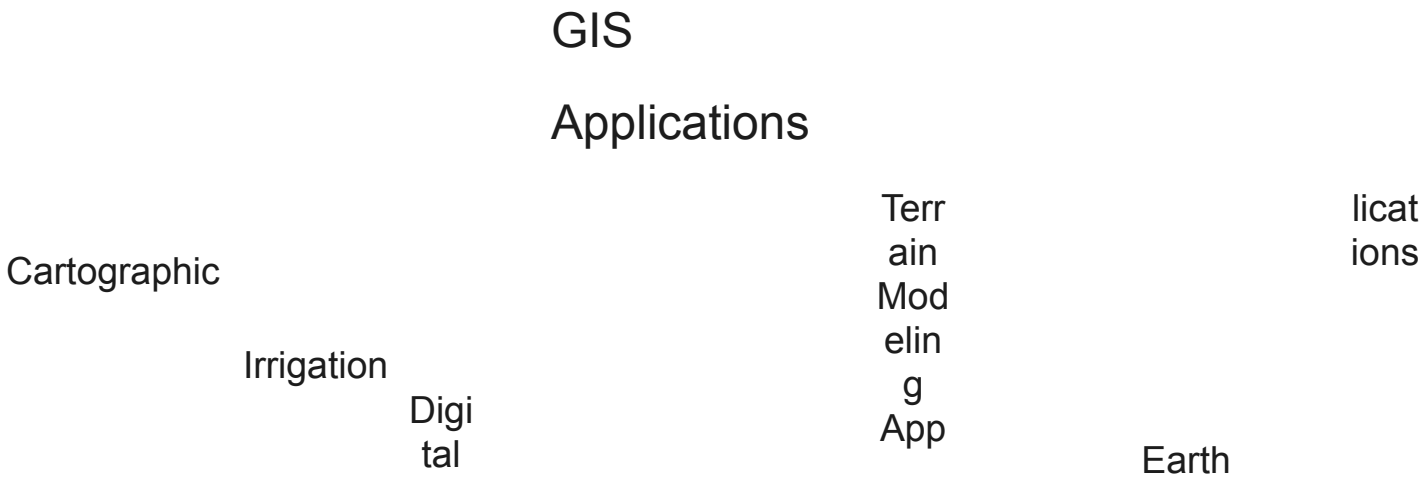
- **Non-spatial** data, such as socio-economic data (like census counts), economic data, and sales or marketing information.

GIS is a rapidly developing domain that offers highly innovative approaches to meet some challenging technical demands.

3.1 GIS Applications

- It is possible to divide GISs into three categories:
  - Cartographic applications
  - Digital terrain modeling applications
  - Geographic objects applications

3.1 GIS Applications(2)



Geographic Objects Applications

Crop yield analysis

Land Evaluation

Planning and Facilities management

Landscape studies

Traffic pattern analysis science

Car navigation

Civil engineering and military evaluation

Soil Surveys

Air and water pollution studies

Flood Control

Water resource

systems

management

Geographic market analysis

Utility distribution and consumption

Consumer product and services – economic analysis

3.2 Data Management Requirements of GIS

- The functional requirements of the GIS applications above translate into the following database requirements.

3.2 Data Management Requirements of GIS (2)

Data Modeling and Representation

- GIS data can be broadly represented in two formats:
  - **Vector** data

represents geometric objects such as points, lines, and polygons.

3.2 Data Management Requirements of GIS (3)

- **Data Modeling and Representation** (contd.):
  - **Raster** data is characterized as an array of points, where each point represents the value of an attribute for a real-world location.
    - Informally, raster images are n-dimensional array where

each entry is a unit of the image and represents an attribute. Two-dimensional units are called pixels, while three-dimensional units are called voxels.

- Three-dimensional elevation data is stored in a raster based digital elevation model (DEM) format.

3.2 Data Management Requirements of GIS (4)

- Another raster format called **triangular irregular network (TIN)** is a topological vector-based approach that models surfaces by connecting sample points as vector of triangles and has a point density that may vary with the

roughness of the terrain.

- Rectangular grids (or elevation matrices) are two dimensional array structures.
  - In **digital terrain modeling (DTM)**, the model also may be used by substituting the elevation with some attribute of interest such as population density or air temperature.
- GIS data often includes a temporal structure in addition to a spatial structure.

3.2 Data Management Requirements of GIS (5)

Data Analysis

- GIS data undergoes various types of analysis.
  - For example, in applications

- such as soil erosion studies, environmental impact studies, or hydrological runoff simulations,
- DTM data may undergo various types of geomorphometric analysis – measurements such as
  - slope values, gradients (the rate of change in altitude), aspect (the compass direction of the gradient), profile convexity (the rate of change of gradient), plan convexity (the convexity of contours and other parameters).

3.2 Data Management Requirements of GIS (6)

Data Integration

- GISs must integrate both vector and raster data from a variety of sources.
  - Sometimes edges and regions are inferred from a raster

image to form a vector model, or conversely, raster images such as aerial photographs are used to update vector models.

- Several **coordinate systems** such as Universal Transverse Mercator (UTM), latitude/longitude, and local cadastral systems are used to identify locations.
- Data originating from different coordinate systems requires appropriate transformations.

3.2 Data Management Requirements of GIS (7)

Data Capture

- The first step in developing a spatial database for cartographic modeling is
  - to capture the two-dimensional or three-dimensional geographical information in digital form

- This process that is sometimes impeded by source map characteristics such as resolution, type of projection, map scales, cartographic licensing, diversity of measurement techniques, and coordinate system differences.

- Spatial data can also be captured from remote sensors in satellites such as Landsat, NORA, and Advanced Very High Resolution Radiometer(AVHRR) as well as SPOT HRV (High Resolution Visible Range Instrument.

3.3 Specific GIS Data Operations

- GIS applications are conducted through the use of special operators such as the following:
  - Interpolation
  - Interpretation
  - Proximity analysis

- Raster image processing
- Analysis of networks

3.3 Specific GIS Data Operations(2)

- The functionality of a GIS database is also subject to other considerations:
  - Extensibility
  - Data quality control
  - Visualization
- Such requirements clearly illustrate that standard

RDBMSs or ODBMSs do not meet the special needs of GIS.

- Therefore it is necessary to design systems that support the vector and raster representations and the spatial functionality as well as the required DBMS features.

4.1 Genome Data Management

- **Biological Sciences and Genetics**:
  - The biological sciences encompass an enormous variety of information.
    - **Environmental** science gives us a view of how species live and interact in a world filled with natural phenomena.
    - **Biology** and **ecology** study particular species.
      - **Anatomy** focuses on the overall structure of an organism, documenting the physical aspects of individual bodies.

Traditional medicine and physiology break the organism into systems and tissues and strive to collect information on the workings of these systems and the organism as a whole.

4.1 Genome Data Management(2)

- Histology and cell biology delve into the tissue and cellular levels and provide knowledge about the inner structure and function of the cell.
  - This wealth of information that has been generated, classified, and stored for centuries has only recently become a major application of database technology.

4.1 Genome Data Management(3)

- **Genetics** has emerged as an ideal field for the application of information technology.
  - In a broad sense, it can be taught of as the construction of models based on information about genes and population and the seeking out of relationships in that information.
    - Genes can be defined as units of heredity

4.1 Genome Data Management(4)

- The study of genetics can be divided into three branches:
  - **Mendelian** genetics is the study of the transmission of traits between generations.
  - **Molecular** genetics is the study of the chemical structure and function of genes at the molecular level.
  - **Population** genetics is the study of how genetic information varies across populations of organisms.

4.1 Genome Data Management(5)

- The origins of **molecular genetics** can be traced to two important discoveries:
  - In 1869 when Friedrich Miescher discovered nuclein and its primary component, deoxyribonucleic acid (DNA).
    - In subsequent research DNA and a related compound, ribonucleic acid, were found to be composed of nucleotides (a sugar, a phosphate, and a base combining to form nucleic acid) linked into long polymers via the sugar and phosphate.

- The second discovery was the demonstration in 1944 by Oswald Avery that DNA was indeed the molecular substance carrying genetic information.

4.1 Genome Data Management(6)

- Genes were shown to be composed of chains of nucleic acids arranged linearly on chromosomes and to serve three primary functions:
  - Replicating genetic information between generations,
  - Providing blueprints for the creation of polypeptides, and
  - Accumulating changes– thereby allowing evolution to occur.
- Watson and Crick found the double-helix structure of the DNA in 1953, which gave molecular biology a new

direction.

4.2 Characteristics of Biological Data

- Biological data exhibits many special characteristics that make management of biological information a particularly challenging problem.
- The characteristics related to biological information, and focusing on a multidisciplinary field called **bioinformatics** that has emerged.
- Bioinformatics addresses information management of genetic information with special emphasis on DNA sequence analysis.

4.2 Characteristics of Biological Data(2)

- Applications of bioinformatics span design of targets for drugs, study of mutations and related diseases, anthropological investigations on migration patterns of tribes and therapeutic treatments.
  - Characteristic 1: Biological data is highly complex when compared with most other domains or applications.
  - Characteristic 2: The amount and range of variability in data is high.

4.2 Characteristics of Biological Data(3)

- Characteristics of Biological Data (contd.)
  - Characteristic 3: Schemas in biological databases change at a rapid pace.
  - Characteristic 4: Representations of the same data by different biologists will likely be different (even using the same system).
  - Characteristic 5: Most users of biological data do not require write access to the database; read-only access is adequate.

4.2 Characteristics of Biological Data(4)

- Characteristics of Biological Data (contd.)
  - Characteristic 6: Most biologists are not likely to have knowledge of the internal structure of the database or about schema design.
  - Characteristic 7: The context of data gives added meaning for its use in biological applications.

4.2 Characteristics of Biological Data(5)

- Characteristics of Biological Data (contd.)
  - Characteristic 8: Defining and representing complex queries is extremely important to the biologist.
  - Characteristic 9: Users of biological information often require access to “old” values of the data – particularly when verifying previously reported results.

4.3 The Human Genome Project and Existing Biological Databases

- The term **genome** is defined as the total genetic information that can be obtained about an entity.
  - E.g., the **human genome** generally refers to the complete set of genes required to create a human being
    - The number is estimated to be more than 30,000 genes spread over 23 pairs of chromosomes, with an estimated 3 to 4 billion nucleotides.

- The goal of the Human Genome Project (HGP) has been to obtain the complete sequence – the ordering of the bases – of those nucleotides.

4.3 The Human Genome Project and Existing Biological Databases(2)

- Some of the existing database systems that are supporting or have grown out of the Human Genome Project.
- **GenBank**
  - The preeminent DNA sequence database in the world today is GenBank, maintained by the National Center for Biotechnology Information (**NCBI**) of the National Library of Medicine (**NLM**).

4.3 The Human Genome Project and Existing Biological Databases(3)

- **GenBank** (contd.)
  - Established in 1978 as a repository for DNA sequence data.
  - Since 1978 expanded to include sequence tag data, protein sequence data, three-dimensional protein structure, taxonomy, and links to the medical literature (MEDLINE).

4.3 The Human Genome Project and Existing Biological Databases(4)

- **GenBank** (contd.)
  - As of release 135.0 in April 2003, GenBank contains over 31 billion nucleotide bases of more than 24 million sequences from over 100,000 species with roughly 1400 new organisms being added each month.
  - The database size in flat file format is over 100 GB uncompressed

and has been doubling every 15 months.

4.3 The Human Genome Project and Existing Biological Databases(5)

- **GenBank** (contd.)
  - International collaboration with the European Molecular Biology Laboratory (EMBL) in the U.K. and the DNA Data Bank of Japan (DDBJ) on daily basis.
  - Other limited data sources (e.g. three-dimensional structure and Online Mendelian Inheritance in Man (OMIM), have been added recently by reformatting the existing OMIM and PDB databases

and redesigning the structure of the GenBank system to accommodate these new data sets.

4.3 The Human Genome Project and Existing Biological Databases(6)

- **GenBank** (contd.)
  - The system is maintained as a combination of flat files, relational databases, and files containing Abstract Syntax Notation One (ASN.1)

4.3 The Human Genome Project and Existing Biological Databases(7)

- **GenBank** (contd.)
  - The average user of the database is not able to access the structure of the data directly for querying or other functions, although complete snapshots of the database are available for export in a number of formats, including ASN.1. The query

mechanism provided is via the Entrez application (or its www version), which allows keyword, sequence, and GenBank UID searching through a static interface.

4.3 The Human Genome Project and Existing Biological Databases(8)

- **The Genome Database (GDB)**
  - Created in 1989, GDB is a catalog of human gene mapping data, a process that associates a piece of information with a particular location on the human genome.
  - GDB data includes data describing primarily map information

(distance and confidence limits), and Polymerase Chain Reaction (PCR) probe data (experimental conditions, PCR primers, and reagents used).

4.3 The Human Genome Project and Existing Biological Databases(9)

- **The Genome Database (GDB)** (contd.)
  - More recently efforts have been made to add data on mutations linked to genetic loci, cell lines used in experiments, DNA probe libraries, and some limited polymorphism and population data.
  - The GDB system is built around Sybase, a commercial relational DBMS, and its data are modeled using standard Entity

Relationship techniques.

- GDB distributes a Database Access Toolkit

4.3 The Human Genome Project and Existing Biological Databases(10)

- **The Genome Database (GDB)** (contd.)
  - As with GenBank, users are given only a very high-level view of the data at the time of searching and thus cannot make use of any knowledge gleaned from the structure of the GDB tables.
    - Search methods are most useful when users are simply looking for an index into map or probe data.

- Exploratory ad hoc searching is not encouraged by present interfaces.
- Integration of the database structures of GDB and OMIM was never fully established.

4.3 The Human Genome Project and Existing Biological Databases(11)

- **Online Mendelian Inheritance in Man**
  - Online Mandelian Inheritance in Man (**OMIM**) is an electronic compendium of information on the genetic basis of human disease.
  - Begun in hard-copy form by Victor McCusick in 1966 with 1500 entries, it was converted to a full-text electronic form between

1987 and 1989 by GDB.

- In 1991 its administration was transferred from John Hopkins University to the NCBI, and the entire database was converted to NCBI's GenBank format. Today it contains more than 14,000 entries.

4.3 The Human Genome Project and Existing Biological Databases(12)

- **Online Mendelian Inheritance in Man** (contd.)
  - OMIM covers material on five disease areas based loosely on organs and systems.
    - Any morphological, biochemical, behavioral, or other properties under study are referred to as **phenotype** of

an individual (or a cell).

- Mendel realized that genes can exist in numerous forms known as **alleles**.
- A **genotype** refers to the actual allelic composition of an individual.

4.3 The Human Genome Project and Existing Biological Databases(13)

- **EcoCyc**
  - The Encyclopedia of **Escherichia coli** Genes and Metabolism (**EcoCyc**) is a recent experiment in combining information about the genome and the metabolism of E.coli K-12.
  - The database was created in 1996 as a collaboration between Stanford Research Institute and Marine Biological Laboratory.

4.3 The Human Genome Project and Existing Biological Databases(14)

- **EcoCyc** (contd.)
  - An object-oriented data model was first used to implement the system, with data stored in Ocelot, a frame knowledge representation system. EcoCyc data was arranged in a hierarchy of object classes based on observations that
    - the properties of a reaction are independent of an enzyme that catalyzes it, and

- an enzyme has a number of properties that are “logically distinct” from its reactions.