DATA
WHAT IS DATA?
Data is ...

• Distinct **pieces** of information, usually **formatted** in a special way.

• Data can exist in a variety of forms -- as numbers or **text** on pieces of paper, as **bits** and **bytes stored** in electronic **memory**, or as facts stored in a person's mind.

• **data** is the plural of **datum**, a single piece of information. In practice, however, people use **data** as both the singular and plural form of the word.
Computer science: viewpoint

• **data** is information in a form suitable for use with a **computer**.
• Relational database
• Transactional database
• Data warehouse
• Data stream  Data repositories (ที่เก็บ, โกดัง)
• Flat files
• Advanced databases and information repositories
  – Object-oriented and object-relational databases
  – Time-series and temporal databases
  – Text database
  – Spatial database
  – Multimedia database
  – www
Relational database
Relational Databases

• A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data.

• A relational database is a collection of tables, each of which is assigned a unique name.
  – Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows). Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values.
Example

• A relational database for AllElectronics company.
  – is described by the following relation tables: customer, item, employee, and branch.
Data warehouse
Data Warehouses [1]

• A **data warehouse** is a repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site.

• Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.
Typical framework of a data warehouse for *AllElectronics*.
Data Warehouses [3]

- A data warehouse is usually modeled by a **multidimensional database structure**.
- **Each dimension** corresponds to an attribute or a set of attributes in the schema.
- **Each cell** stores the value of some aggregate measure, such as count or sales amount.
Data Warehouses [4]

• The actual physical structure of a data warehouse may be a relational data store or a multidimensional data cube.
• A data cube provides a multidimensional view of data and allows the pre-computation and fast accessing of summarized data.
• A data cube for summarized sales data of AllElectronics

• The cube has three dimensions: address (with city values Chicago, New York, Toronto, Vancouver), time (with quarter values Q1, Q2, Q3, Q4), and item (with itemtype values home entertainment, computer, phone, security)
Data mart Vs. Data warehouse

• A *data warehouse* collects information about subjects that span an entire organization, and thus its scope is enterprise-wide.

• A *data mart* is a department subset of a data warehouse.
  – It focuses on selected subjects, and thus its scope is department-wide.
Multidimensional data views on data warehouse

• Data warehouse systems are well suited for *on-line analytical processing*, or *OLAP*.

• OLAP operations *use background knowledge regarding the domain of the data* being studied in order to allow the *presentation of data at different levels of abstraction*. 
<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Transactional database
Transactional Databases

• A transactional database consists of a file where each record represents a transaction.

• A transaction typically includes a unique transaction identity number (trans ID) and a list of the items making up the transaction (such as items purchased in a store).
Example

- A transactional database for AllElectronics

<table>
<thead>
<tr>
<th>trans_ID</th>
<th>list of item IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>T100</td>
<td>I1, I3, I8, I16</td>
</tr>
<tr>
<td>T200</td>
<td>I2, I8</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Advanced Data and Information Systems and Advanced Applications
Object-Relational Databases [1]

• Object-relational databases are constructed based on an object-relational data mode.

• This model extends the relational model by providing a rich data type for handling complex objects and object orientation.

• Each entity is considered as an object.
Object-Relational Databases [2]

• AllElectronics example: Objects can be individual employees, customers, or items.

• Each object has associated with it the following:
  – A set of variables that describe the objects.
  – A set of messages that the object can use to communicate with other objects.
  – A set of methods, where each method holds the code to implement a message.
Temporal Databases

• A temporal database typically stores relational data that include time-related attributes.

• These attributes may involve several timestamps, each having different semantics.
Sequence Databases

- A sequence database stores sequences of ordered events, with or without a concrete notion of time. Examples include customer shopping sequences, Web click streams, and biological sequences.
Time-Series Databases

- A time-series database stores sequences of values or events obtained over repeated measurements of time (e.g., hourly, daily, weekly). Examples include data collected from the stock exchange, inventory control, and the observation of natural phenomena (like temperature and wind).
Spatial Databases

- Spatial databases contain spatial-related information.

- Examples include geographic (map) databases, very large-scale integration (VLSI) or computed-aided design databases and medical and satellite image databases.

- Spatial data may be represented in raster format, consisting of n-dimensional bit maps or pixel maps.
  - A 2-D satellite image may be represented as raster data, where each pixel registers the rainfall in a given area.
Spatial Databases [2]

• Maps can be represented in vector format, where roads, bridges, buildings, and lakes are represented as unions or overlays of basic geometric constructs, such as points, lines, polygons, and the partitions and networks formed by these components.
“What kind of data mining can be performed on spatial databases?”

- patterns describing the characteristics of houses located near a specified kind of location, such as a park, for instance.

- patterns may describe the climate of mountainous areas located at various altitudes

- Clusters and outliers can be identified by spatial cluster analysis.
“What kind of data mining can be performed on spatial databases?” [2]

- **Spatial classification** can be performed to construct models for prediction based on the relevant set of features of the spatial objects.

- “**Spatial data cubes**” may be constructed to organize data into multidimensional structures and hierarchies, on which OLAP operations.
Spatiotemporal Databases

- A spatial database that stores spatial objects that change with time is called a spatiotemporal database. For example, we may be able to group the trends of moving objects and identify some strangely moving vehicles, or distinguish a bioterrorist attack from a normal outbreak of the flu based on the geographic spread of a disease with time.
Text Databases [1]

• Contain word descriptions for objects. (These word descriptions are usually not simple keywords but rather long sentences or paragraphs, such as product specifications, error or bug reports, warning messages, summary reports, notes, or other documents)

• May be highly unstructured (such as some Web pages on the WorldWideWeb)
Text Databases [2]

• May be somewhat structured, that is, semistructured (such as e-mail messages and many HTML/XML Web pages), whereas others are relatively well structured (such as library catalogue databases).
“What can data mining on text databases uncover?”

• general and concise descriptions of the text documents, keyword or content associations, as well as the clustering behavior of text objects.

• To do this, standard data mining methods need to be integrated with information retrieval techniques and the construction or use of hierarchies specifically for text data (such as dictionaries and thesauruses), as well as discipline-oriented term classification systems (such as in biochemistry, medicine, law, or economics).
Multimedia Databases

• Multimedia databases store image, audio, and video data.

• Used in applications such as picture content-based retrieval, voice-mail systems, the World Wide Web, and speech-based user interfaces that recognize spoken commands.
“What kind of data mining can be performed on Multimedia databases?”

• For multimedia data mining, storage and search techniques need to be integrated with standard data mining methods.

• Promising approaches include the construction of multimedia data cubes, the extraction of multiple features from multimedia data, and similarity-based pattern matching.
Heterogeneous Databases

• A heterogeneous database consists of a set of interconnected, autonomous component databases.

• The components communicate in order to exchange information and answer queries.
Legacy Databases

• A legacy database is a group of heterogeneous databases that combines different kinds of data systems, such as relational or object-oriented databases, hierarchical databases, network databases, spreadsheets, multimedia databases, or file systems.

• The heterogeneous databases in a legacy database may be connected by intra- or inter-computer networks.
Data Streams

• **Data flow in and out of an observation platform (or window) dynamically**
• Properties: huge or possibly infinite volume,
• dynamically changing, flowing in and out in a fixed order, allowing only one or a small number of scans, and demanding fast (often real-time) response time.
examples of data streams

• include various kinds of scientific and engineering data, time-series data, and data produced in other dynamic environments, such as power supply, network traffic, stock exchange, telecommunications, Web click streams, video surveillance, and weather or environment monitoring
The World Wide Web

• The World Wide Web and its associated distributed information services, such as Yahoo!, Google, America Online, and AltaVista, provide rich, worldwide, on-line information services, where data objects are linked together to facilitate interactive access.

• Capturing user access patterns in such distributed information environments is called Web usage mining (or Weblog mining)
Applications of www database

• **Authoritative Web page analysis** based on linkages among Web pages can help rank Web pages based on their importance, influence, and topics.

• **Automated Web page clustering and classification** help group and arrange Web pages in a multidimensional manner based on their contents.

• **Web community analysis** helps identify hidden Web social networks and communities and observe their evolution.
Integration of a Data Mining System with a Database or Data Warehouse System

*how to integrate or couple the DM system with a database (DB) system and/or a data warehouse (DW) system?*

- If a DM system *works as a stand-alone system* or is embedded in an application program, there are no DB or DW systems with which it has to communicate. This simple scheme is called *no coupling*, where the main focus of the DM design rests on *developing effective and efficient algorithms for mining the available data sets.*
Integration of a Data Mining System with a Database or Data Warehouse System

• When a DM system works in an environment that requires it to communicate with other information system components, such as DB and DW systems, possible integration schemes include no coupling, loose coupling, semitight coupling, and tight coupling.
Types of schemes [1]

- **No coupling**: a DM system will **not utilize any function of a DB or DW system**. It may fetch data from a particular source (such as a file system), process data using some data mining algorithms, and then store the mining results in another file.
Types of schemes [2]

- **No coupling**: a DM system will **not utilize any function of a DB or DW system**.
  - fetch data from a particular source (such as a file system), process data using some data mining algorithms, and then store the mining results in another file.
  - suffer from several drawbacks:
    - Without using a DB/DW system, a DM system may spend a substantial amount of time finding, collecting, cleaning, and transforming data
Types of schemes [3]

- **Loose coupling:** a DM system will use some facilities of a DB or DW system, fetching data from a data repository managed by these systems, performing data mining, and then storing the mining results either in a file or in a designated place in a database or data warehouse.
Types of schemes [4]

• **Semi-tight coupling:** besides linking a DM system to a DB/DW system, efficient implementations of a few essential data mining primitives can be provided in the DB/DW system.

• These primitives can include sorting, indexing, aggregation, histogram analysis, multi-way join, and precomputation of some essential statistical measures, such as sum, count, max, min, standard deviation, and so on.
Types of schemes [5]

• **Tight coupling:** a DM system is smoothly integrated into the DB/DW system. The data mining subsystem is treated as one functional component of an information system. Data mining queries and functions are optimized based on mining query analysis, data structures, indexing schemes, and query processing methods of a DB or DW system.
Q & A