# Database Models

## Relational

* Rows and Columns
* Normalised so data is not repeated more than necessary
* Columns depend on primary key
* Stores data as int or strings in tables

## Object Oriented

* Uses objects that contain stored values found within an object
* Process of designing and preparing what the model's code looks like
* Modelling techniques used during construction
* 3 phases: Analysis, design, and implementation

## Network

* Flexible way of representing objects and their relationships
* Distinguishing feature is that schema is viewed as graph in which object types are nodes and relationship types are arcs
* Not restricted to being hierarchical or lattice

## Spatial

* Stores spatial/locational/geometric data such as point, polygons, lines and 3d shapes
* Models structure of geometric shapes
* Makes use of spatial indexes to access data that supports geometric functions

## Multi-Dimensional

* Classical relations are 1-dimensional – Maps key onto attributes
* Data warehousing needs multiple perspectives
* Cannot be represented with normal relations
* Multi-dimensional data seen and represented as data cubes (Hypercubes)

# Data Warehouses

## Definition

Subject oriented, integrated, time-variant and non-volatile collection of data in support of management’s decision-making process.

Repository of stored data related to specific subject with tools to manage and retrieve metadata.

## Situations

* Large volumes of data
* Different data types
* Poorly integrated data
* Legacy systems work independently

## Time Dependent

* Content in warehouse only valid for a time period
* Data undergoes change dynamically
* Data warehouse’s focus on change over time is time variant

## Methods to Allow Realtime Updates

* Sorting
* Aggregation
* Split column into multiple columns
* Lookup and validate relevant data
* Join data from multiple sources (e.g. merge)
* Deriving new calculated values (E.g. revenue = q\*p)

## Advantages

* Integrate data from multiple sources
* Perform new types of analysis
* Reduce cost to access historical data
* Standardize data across organisation
* Sharing data and allowing others to easily access data

# ETL

* Extract, transform, load
* Useful data discovered and extracted from various sources
* Transformations performed: Repair inaccuracy between data formats/Remove mistakes/Clean data/Encode free-form values (Male 🡪 M)/Sorting
* Load data into data warehouse

# Data Mining

## Applications

* Marketing: Consumer behaviour based on buying patterns
* Finance: Performance analysis of finance investments such as stocks, Banks to identify fraudulent credit card use
* Manufacturing: Applications involve optimization of resources like manpower
* Healthcare: Patterns in radiological images

## Analysis Types

### Associations

* Looks at how entities are connected and finds where one or more events may lead to another
* If-Then rules that predict the occurrence of something based on the presence of other things

### Cluster Analysis

* Finds patterns in customer behaviour by grouping and analysing variables that connect them
* Can find previously unknown links which help in decision making

### Classifications

* Partition data so different classes or categories can be identified based on combination of parameters
* Used in different analysis of customer buying transactions as a post-mining action
* Can be used to encode data appropriately before subjecting it to further data mining

### Sequential Patterns

* Deals with sequence of actions or events
* Detection of sequential patterns equivalent to detecting associations among events with temporal relationships

### Forecasting

* Show how certain attributes within data will behave in future
* E.g. analysis of buying transactions to predict what customers will buy under certain discounts
* Business logic coupled with data mining

## Predictive Modelling

* Process that uses data mining and probability to forecast outcomes
* Each model made of several predictors which are variables that are likely to influence future results
* Once data has been collected for relevant predictors, a statistical model is formulated
* Model may employ a simple linear equation, or it may be a complex neural network that is mapped out by sophisticated software
* As additional data becomes available, the statistical analysis model is validated or revised

## Database Segmentation

* Dividing entities into groups with similar characteristics/common attributes
* Partitions database according to common features in the rows

## Link Analysis

* Visual data analysis technique
* Gain better understanding of connections between objects of interest and problem domain by displaying connections in a form of a network diagram (link chart)
* Purpose
  + Provides flexibility for accessing and manipulating large volumes of data
  + Record-based data models used in DBMS inappropriate for modelling connections between objects

## Deviation Detection

* Involves finding “unusual” records in the dataset
* Example: Used to find unusually large orders or long-periods of time for transport

## Ethical Concerns

* Can reveal connections in personal data which could be misinterpreted or give a misleading data profile used for reasons that the individual does not know or could not know
* Hard to restrict access to data
* Pressure from third parties to share the data such as from insurance companies and government agencies which is unknown and not wanted by the individual
* Data may appear benign but in the hands of a third party leads to unforeseen issues
* When analysed, provides information that may not be accessible via standard query
* Not all customers will be able to foresee the extent to which their data may be misused
* There may be a mission creep and the reason for the original data collection may be super-seeded
* Thus, security of data cannot be guaranteed