

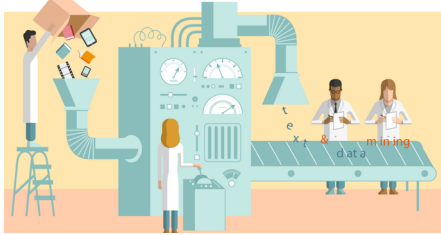
### Compare OLAP & On-Line Transactional Processing (OLTP)

- OLAP application differs from OLTP application in the way that they store data, analyze data & present data to end user.
- OLTP application usually retrieve groups of records & present to end user, use relational DB with tables containing individual transactions linked to meta tables that store customer & product details.  
*E.g. list of computer sold at a store during one day.*
- OLAP present end user with information rather than just data. It is easy for users to identify patterns in data without searching through raw data.  
*E.g. show me our most successful salesman.*

### 3 Limitations of OLTP (Why do we need OLAP?)

- **Increasing data storage:** company need to store large amount of data about their business while retrieving thousands of records for immediate analysis is a time & resource consuming process.
- **Data versus information:** users need both data & information. They make decisions based on happening events that need the information contained within company data. In order to answer users' questions, the procedure must be repeated for multiple days, not in the typical few-seconds response time demanded by users.
- **Data layout:** relational DB model was designed for transactional processing & it is not always the best way to store data when attempting to answer business questions. This type of questions requires large amount of data to be retrieved & that require time & resources to achieve.

**Definition of Data Mining (Knowledge Discovery in DB (KDD)):**  
it is the principle of sorting through large amount of structured data & picking out relevant information. It is also the extraction of interesting knowledge & patterns from huge amount of data & use it for classification & prediction.



**Definition of Text Mining (Discovery in Textual DB):**

it is the process of analyzing textual data which are the unstructured data that requires specific technique to analyze. It involves identifying key words & find relationship between the concepts. It includes automated or partially automated processing of text that involves imposing structure upon text so that relevant information can be extracted.



### 2 Applications of Data Mining

- **Detecting Fraud & other Criminal Activity:** suspicious credit card purchases & potential fraudulent insurance claims can be highlighted using data mining techniques.
- **Provide Early Warning of Supplier Failure:** Ford company use data mining techniques to identify suppliers that are heading for trouble.

### 2 Forms of Data Mining

Unsupervised Data Mining	Supervised Data Mining
There is no model or hypothesis exists before running an analysis. Analysts apply data mining techniques, observe the results and create a hypothesis after analysis. * <b>Cluster Analysis:</b> common technique in this category that group entities together that have similar characteristics.	A model is developed prior to analysis. Analysts apply statistical techniques to estimate parameters of a model. * <b>Regression Analysis:</b> a technique in this category that measures the impact of a set of variables on another variable.

### 3 Applications of Text Mining

- **Information Extraction:** identify key phrases & relationship within text by looking for predefined sequences in text via a process – **Pattern Matching**.
- **Categorization:** identify main theme of document & place the document into a pre-defined set of topic categories.
- **Clustering:** group documents that are similar to each other without having a pre-defined set of categories.

### 3 Activities in Text Mining

- **Text Categorization:** the initial objective is to organize & understand various volumes of textual data.
- **Information Retrieval:** it is relating to searching to find the proverbial 'needle in the haystack'.
- **Measurement:** it defines text measures that may be used to convert contextual information into numeric information.

**Definition of Web Mining:**

it is the process of discovering intrinsic relationship from web data and also the application of data mining techniques to extract knowledge from web data. It includes web documents, hyperlink between documents & usage logs of websites. The data is in HTML, XML & text format.  
*E.g. web mining tools are ClickTracks & WebTrends.*



### Compare Data Mining & Text Mining

- both seek novel & useful patterns.
- both are semi-automated processes.
- the main difference is that in text mining, the patterns are extracted from natural language text rather than from structured DB of facts.

### 3 Categories of Web Mining

- Web Content Mining:** process of extracting useful information from the contents of web documents. The content of data corresponds to collection of facts a webpage was designed to convey to users. There are 4 applications including **Topic Discovery**, **Extract Association Pattern**, **Classification of Webpage** & **Cluster Web Document**.
- Web Structure Mining:** process of discovering structure information from the web and identify relationship between webpages linked by information or direct link connection.  
Two Types of Structure Information Used
  - **Hyperlink:** structural unit that connects a location in a webpage to different location, either within the same webpage or on a different webpage. Intra-document hyperlink is the hyperlink that connects to different part of same page while Inter-document hyperlink is the hyperlink that connect two different pages.
  - **Document Structure:** content within a webpage can be organized in a tree-structured format, based on various HTML tags within the page.
- Web Usage Mining:** extraction of information from data generated through webpage visits & transactions to better serve the needs of web-based application.