Business Intelligent
Introduction

Munawar, Ir. MMSI. M.Com, PhD

About the course

- **Instructor:**
  - Munawar, an_moenawar@yahoo.com
- **Home Page:**
  - http://moenawar.web.id

    Check frequently! Syllabus, schedule, assignments, announcements…

Grading

❖Assignment

- Mid Test (45%) →design DW for data mining
- Final Test (55%) → designing data mining from many sources : text, web, data

Course Syllabus

Textbook:

Kandasami & Benson. 2012. Making the Most of Big Data: Manager's Guide to Business Intelligent Success. Bookboon.com

Munawar, 2016. Data Warehouse Development Framework with Data Quality Consideration

Topics:

- Overview of data warehousing and mining
- Data warehouse and OLAP technology for data mining
- Data preprocessing
- Mining association rules
- Classification and prediction
- Cluster analysis
- Mining complex types of data

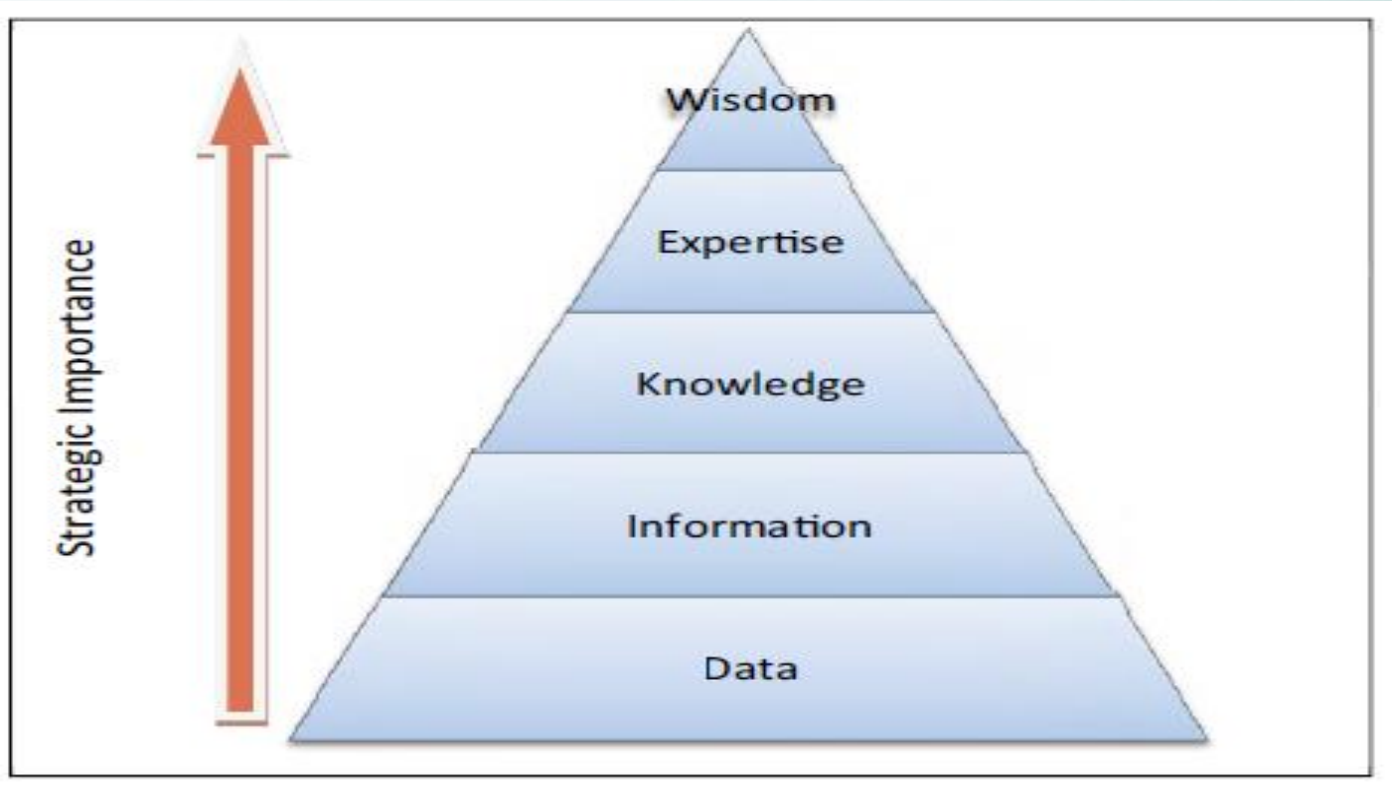
BI definition

- Business Intelligent is the process, technologies and tools needed to turn data into information, information into knowledge, and knowledge into plans that drive profitable business action (The DW institute)
- Business Intelligent is active, model-based, and prospective approach to discover and explain hidden decision-relevant aspects in large amount of business data to better inform business decision process (KMBI, 2005)

Aim of BI

- To collect data from heterogeneous sources, maintain, and organise knowledge.
- Analytical tools present this information to users in order to support decision making process within the organisation
- The objective is to improve the quality and timeliness of inputs to decision process
- The potential to maximize the use of information by improving company's capacity to structure large volume of information and make it accessible, thereby creating competitive advantage (Davenport, 2005)

Traditioanl Intelligence Hierarchy



Source: Liebowitz, 2006

The Importance of BI initiatives

Gartner Research says

- 70-80 % of corporate BI project fail due to a poor communication between IT & business
- The failure to ask right questions or think about the real need of the business (Goodwin, 2010)
- 60-70% of BI application fail due to the technology, organisation culture and infrastructure isses (Lupu, et al, 2007)

Motivation: "Necessity is the Mother of Invention"

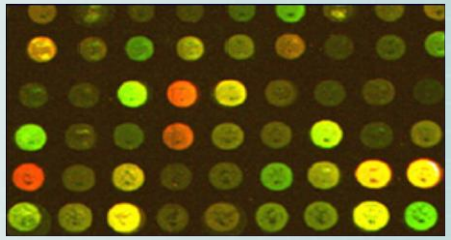
- **Data explosion problem**
  - Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories
- **We are drowning in data, but starving for knowledge!**
- **Solution: Data warehousing and data mining**
  - Data warehousing and on-line analytical processing
  - Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

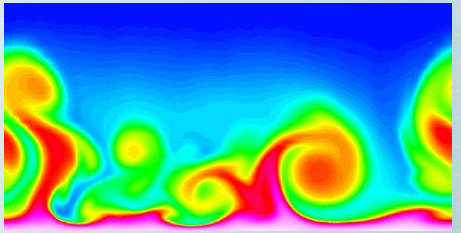
Why Mine Data? Commercial Viewpoint

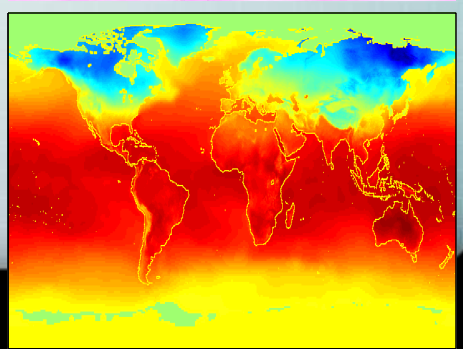
- **Lots of data is being collected and warehoused**
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

- **Computers have become cheaper and more powerful**
- **Competitive Pressure is Strong**
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

Why Mine Data? Scientific Viewpoint





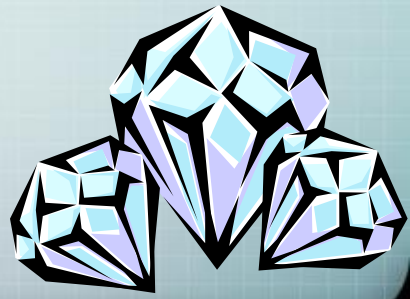


- **Data collected and stored at enormous speeds (GB/hour)**
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- **Traditional techniques infeasible for raw data**
- **Data mining may help scientists**
  - in classifying and segmenting data
  - in Hypothesis Formation

What Is Data Mining?

- **Data mining (knowledge discovery in databases):**
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases

- **Alternative names and their "inside stories":**
  - Data mining: a misnomer?
  - Knowledge discovery(mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, business intelligence, etc.



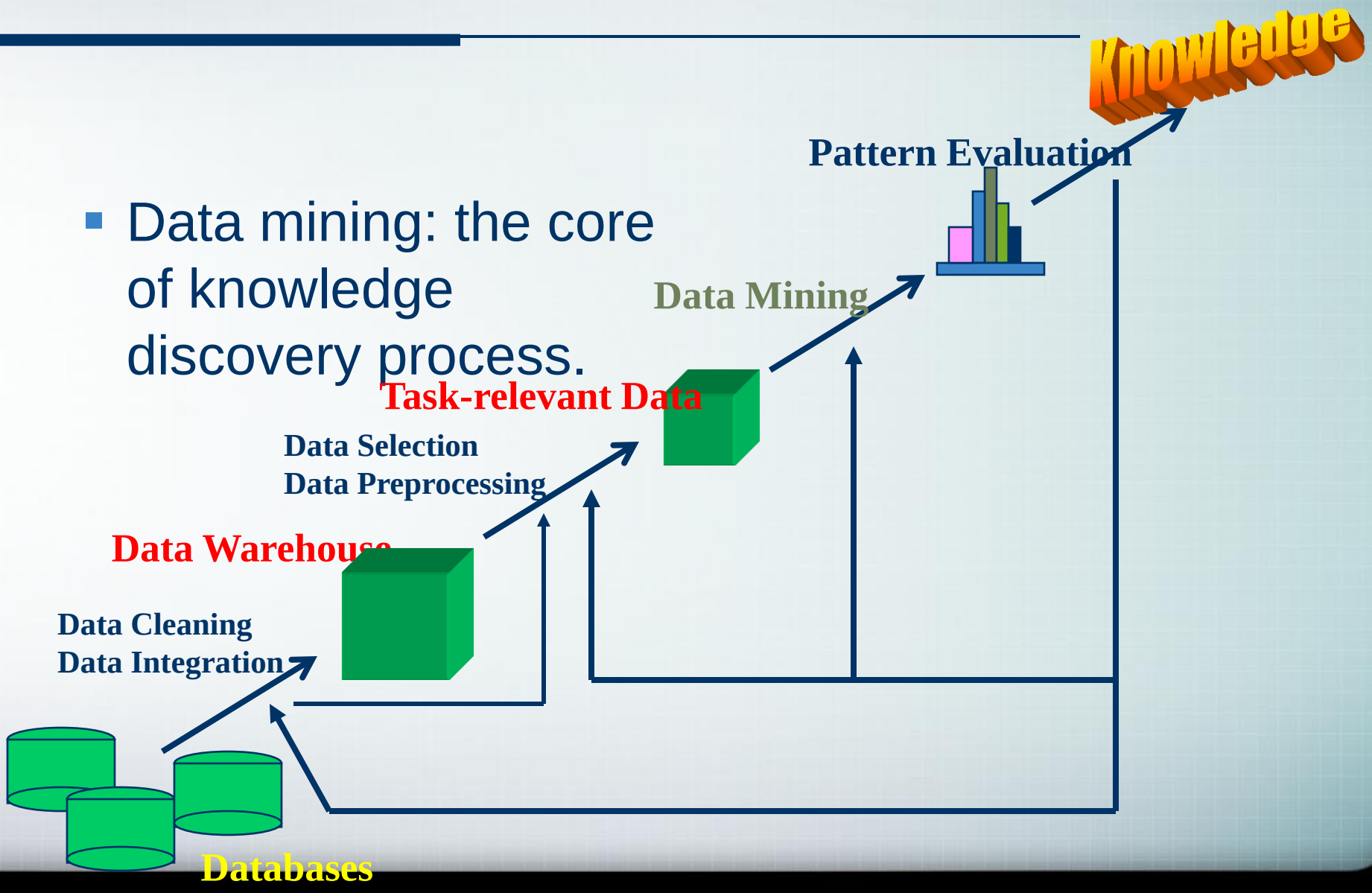
Examples: What is (not) Data Mining?

- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"

- **What is Data Mining?**
  - Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Data Mining: A KDD Process

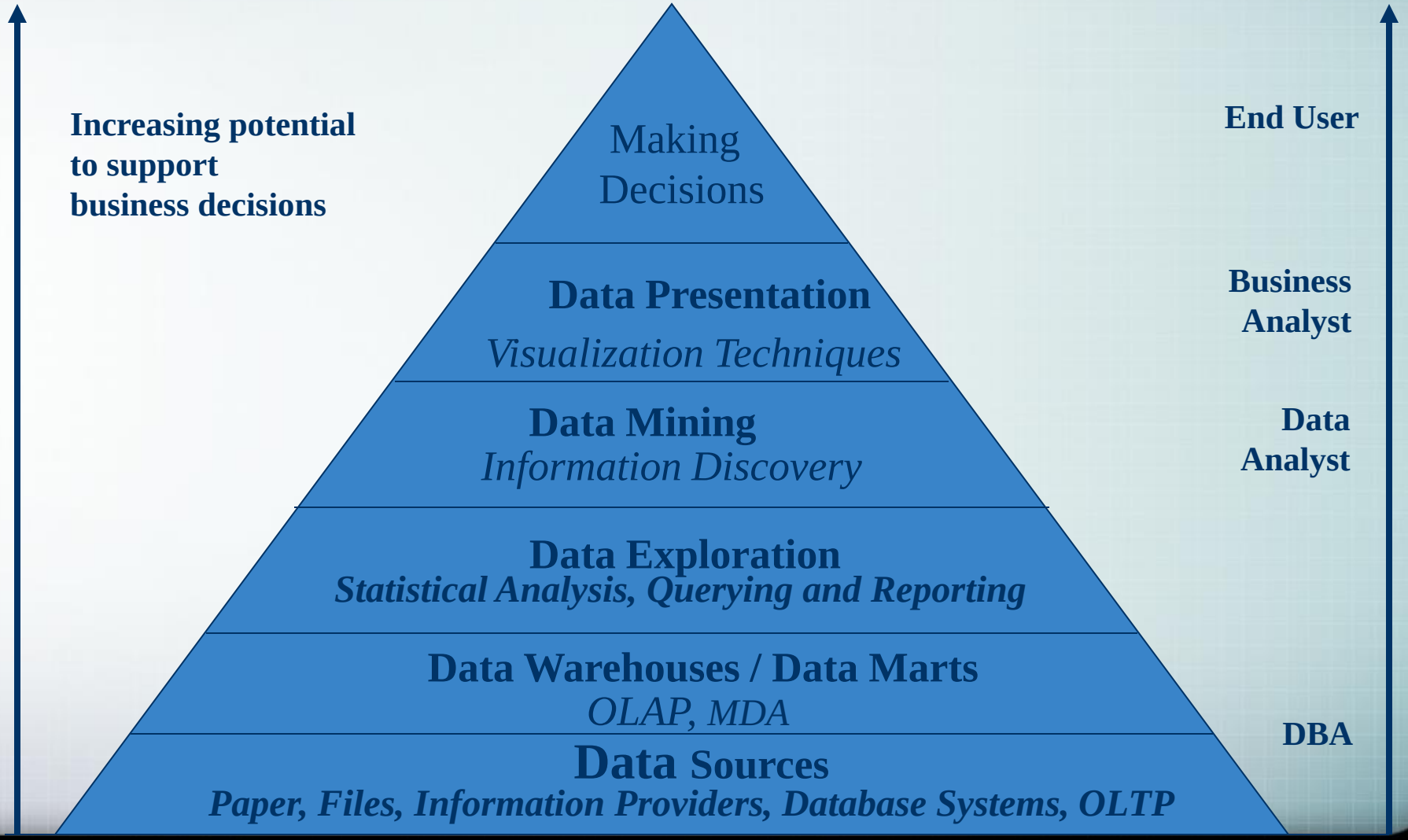
- Data mining: the core of knowledge discovery process.



Steps of a KDD Process

- **Learning the application domain:**
  - relevant prior knowledge and goals of application
- **Creating a target data set: data selection**
- **Data cleaning and preprocessing: (may take 60% of effort!)**
- **Data reduction and transformation:**
  - Find useful features, dimensionality/variable reduction, invariant representation.
- **Choosing functions of data mining**
  - summarization, classification, regression, association, clustering.
- **Choosing the mining algorithm(s)**
- **Data mining: search for patterns of interest**
- **Pattern evaluation and knowledge presentation**
  - visualization, transformation, removing redundant patterns, etc.
- **Use of discovered knowledge**

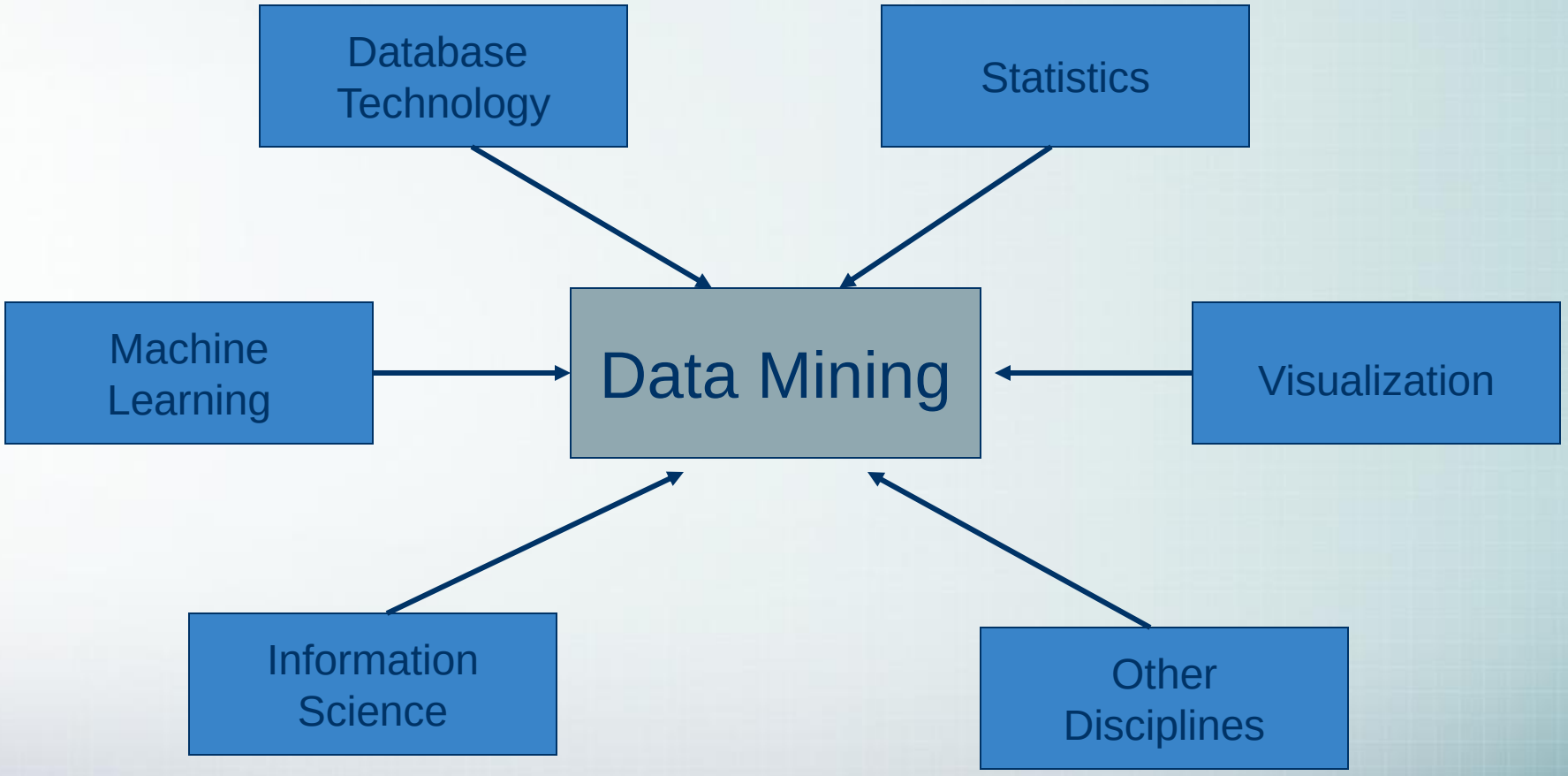
Data Mining and Business Intelligence



Data Mining: On What Kind of Data?

- **Relational databases**
- **Data warehouses**
- **Transactional databases**
- **Advanced DB and information repositories**
  - Object-oriented and object-relational databases
  - Spatial databases
  - Time-series data and temporal data
  - Text databases and multimedia databases
  - Heterogeneous and legacy databases
  - WWW

Data Mining: Confluence of Multiple Disciplines



Data Mining vs. Statistical Analysis

Statistical Analysis:

- Ill-suited for Nominal and Structured Data Types
- Completely data driven - incorporation of domain knowledge not possible
- Interpretation of results is difficult and daunting
- Requires expert user guidance

Data Mining:

- Large Data sets
- Efficiency of Algorithms is important
- Scalability of Algorithms is important
- Real World Data
- Lots of Missing Values
- Pre-existing data - not user generated
- Data not static - prone to updates
- Efficient methods for data retrieval available for use

Data Mining vs. DBMS

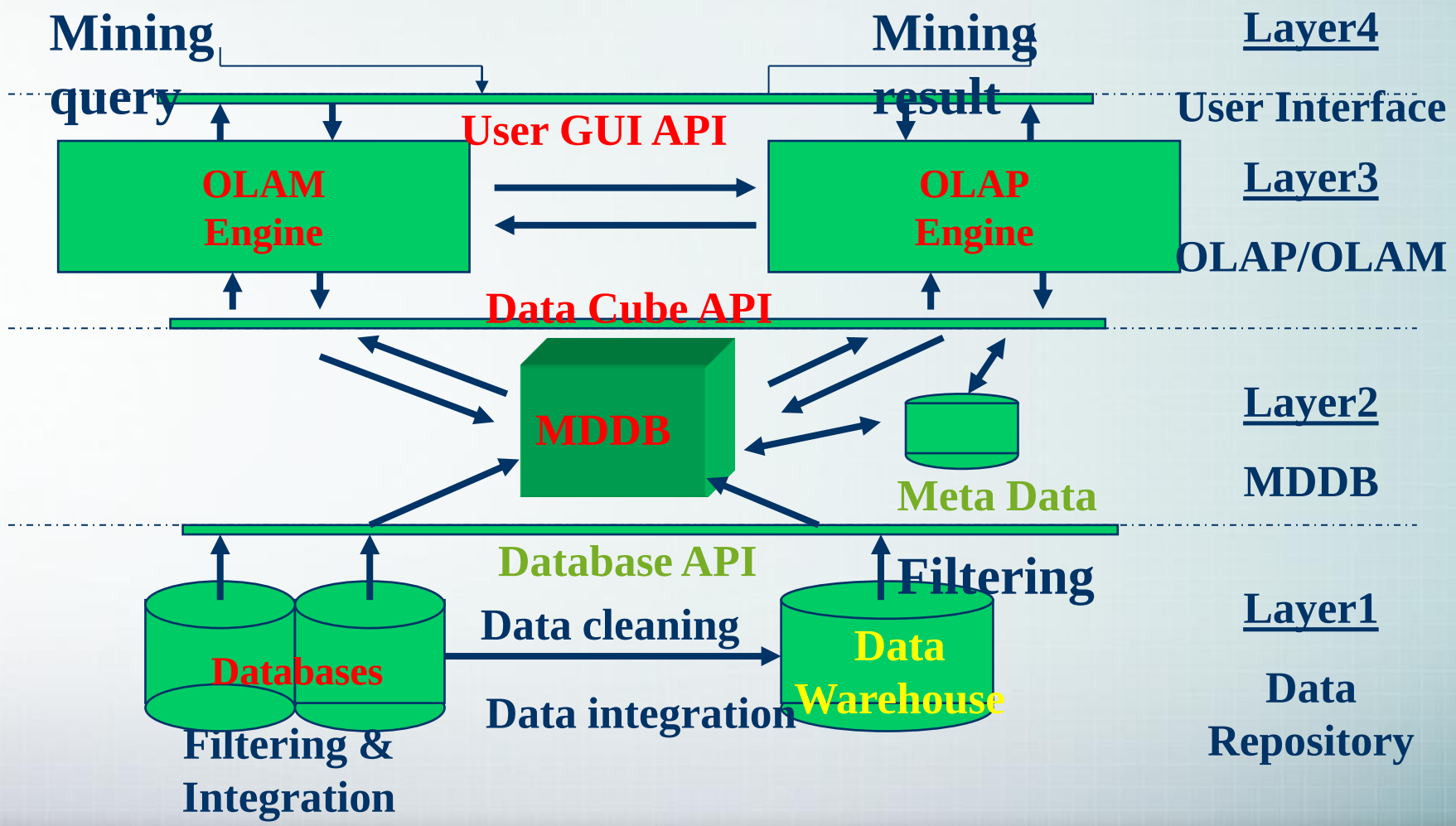
- **Example DBMS Reports**
  - Last months sales for each service type
  - Sales per service grouped by customer sex or age bracket
  - List of customers who lapsed their policy

- **Questions answered using Data Mining**
  - What characteristics do customers that lapse their policy have in common and how do they differ from customers who renew their policy?
  - Which motor insurance policy holders would be potential customers for my House Content Insurance policy?

Data Mining and Data Warehousing

- **Data Warehouse: a centralized data repository which can be queried for business benefit.**
- **Data Warehousing makes it possible to**
  - extract archived operational data
  - overcome inconsistencies between different legacy data formats
  - integrate data throughout an enterprise, regardless of location, format, or communication requirements
  - incorporate additional or expert information
- **OLAP: On-line Analytical Processing**
- **Multi-Dimensional Data Model (Data Cube)**
- **Operations:**
  - Roll-up
  - Drill-down
  - Slice and dice
  - Rotate

An OLAM Architecture



DBMS, OLAP, and Data Mining

	DBMS	OLAP	Data Mining
Task	Extraction of detailed and summary data	Summaries, trends and forecasts	Knowledge discovery of hidden patterns and insights
Type of result	Information	Analysis	Insight and Prediction
Method	Deduction (Ask the question, verify with data)	Multidimensional data modeling, Aggregation, Statistics	Induction (Build the model, apply it to new data, get the result)
Example question	Who purchased mutual funds in the last 3 years?	What is the average income of mutual fund buyers by region by year?	Who will buy a mutual fund in the next 6 months and why?

Example of DBMS, OLAP and Data Mining: Weather Data

DBMS:

Day	outlook	temperature	humidity	windy	play
1	sunny	85	85	false	no
2	sunny	80	90	true	no
3	overcast	83	86	false	yes
4	rainy	70	96	false	yes
5	rainy	68	80	false	yes
6	rainy	65	70	true	no
7	overcast	64	65	true	yes
8	sunny	72	95	false	no
9	sunny	69	70	false	yes
10	rainy	75	80	false	yes
11	sunny	75	70	true	yes
12	overcast	72	90	true	yes
13	overcast	81	75	false	yes
14	rainy	71	91	true	no

Example of DBMS, OLAP and Data Mining: Weather Data

- By querying a DBMS containing the above table we may answer questions like:
- What was the temperature in the sunny days? {85, 80, 72, 69, 75}
- Which days the humidity was less than 75? {6, 7, 9, 11}
- Which days the temperature was greater than 70? {1, 2, 3, 8, 10, 11, 12, 13, 14}
- Which days the temperature was greater than 70 and the humidity was less than 75? The intersection of the above two: {11}

Example of DBMS, OLAP and Data Mining: Weather Data

OLAP:

- Using OLAP we can create a Multidimensional Model of our data (Data Cube).
- For example using the dimensions: time, outlook and play we can create the following model.

9 / 5	sunny	rainy	overcast
Week 1	0 / 2	2 / 1	2 / 0
Week 2	2 / 1	1 / 1	2 / 0

Example of DBMS, OLAP and Data Mining: Weather Data

Data Mining:

- Using the ID3 algorithm we can produce the following decision tree:
  - outlook = sunny
    - humidity = high: no
    - humidity = normal: yes
  - outlook = overcast: yes
  - outlook = rainy
    - windy = true: no
    - windy = false: yes

Major Issues in Data Warehousing and Mining

- **Mining methodology and user interaction**
  - Mining different kinds of knowledge in databases
  - Interactive mining of knowledge at multiple levels of abstraction
  - Incorporation of background knowledge
  - Data mining query languages and ad-hoc data mining
  - Expression and visualization of data mining results
  - Handling noise and incomplete data
  - Pattern evaluation: the interestingness problem
- **Performance and scalability**
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed and incremental mining methods

Major Issues in Data Warehousing and Mining

- **Issues relating to the diversity of data types**
  - Handling relational and complex types of data
  - Mining information from heterogeneous databases and global information systems (WWW)
- **Issues related to applications and social impacts**
  - Application of discovered knowledge
    - Domain-specific data mining tools
    - Intelligent query answering
    - Process control and decision making
  - Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem
  - Protection of data security, integrity, and privacy

Thank You !

Munawar, PhD