Data Analysis and Decision Making

Lecture 1: Introduction

Dr.Ramez Alkhatib

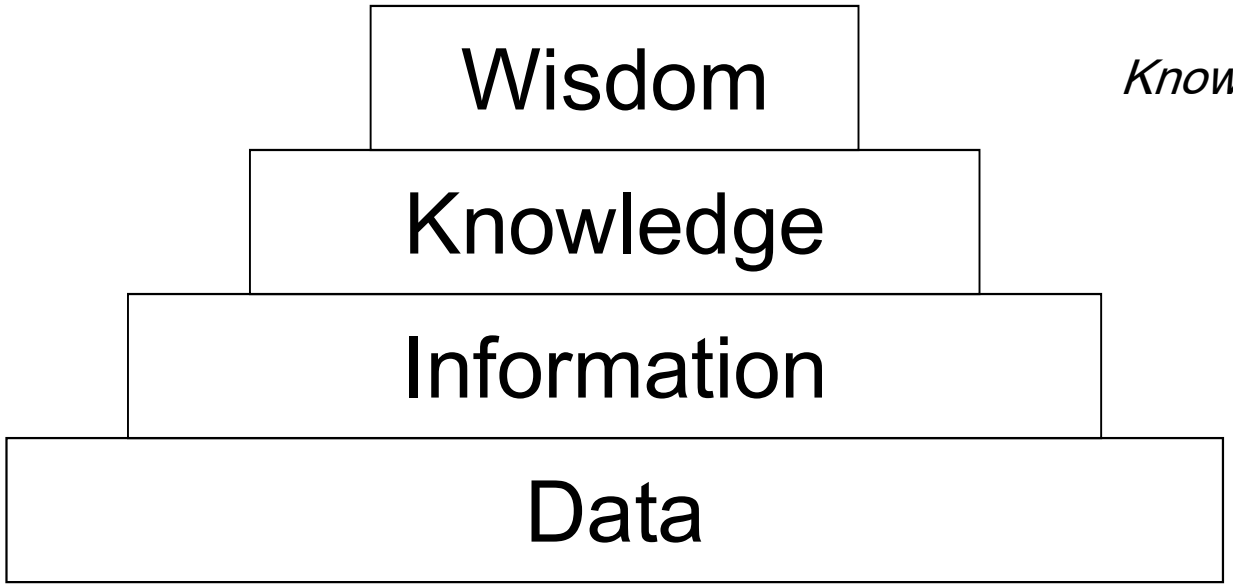
Out Lines

- Definition of Data Mining
- Need for Data Mining
- Data Mining Tasks/Challenges
- Data Mining as an Interdisciplinary field
- Process of Data Mining

Definition of Data Mining

- What is KDD or "Knowledge discovery from databases"?
- "A non-trivial process of identifying valid, novel, useful and ultimately understandable patterns in data".

Data pyramid



Definition of Data Mining (Example)

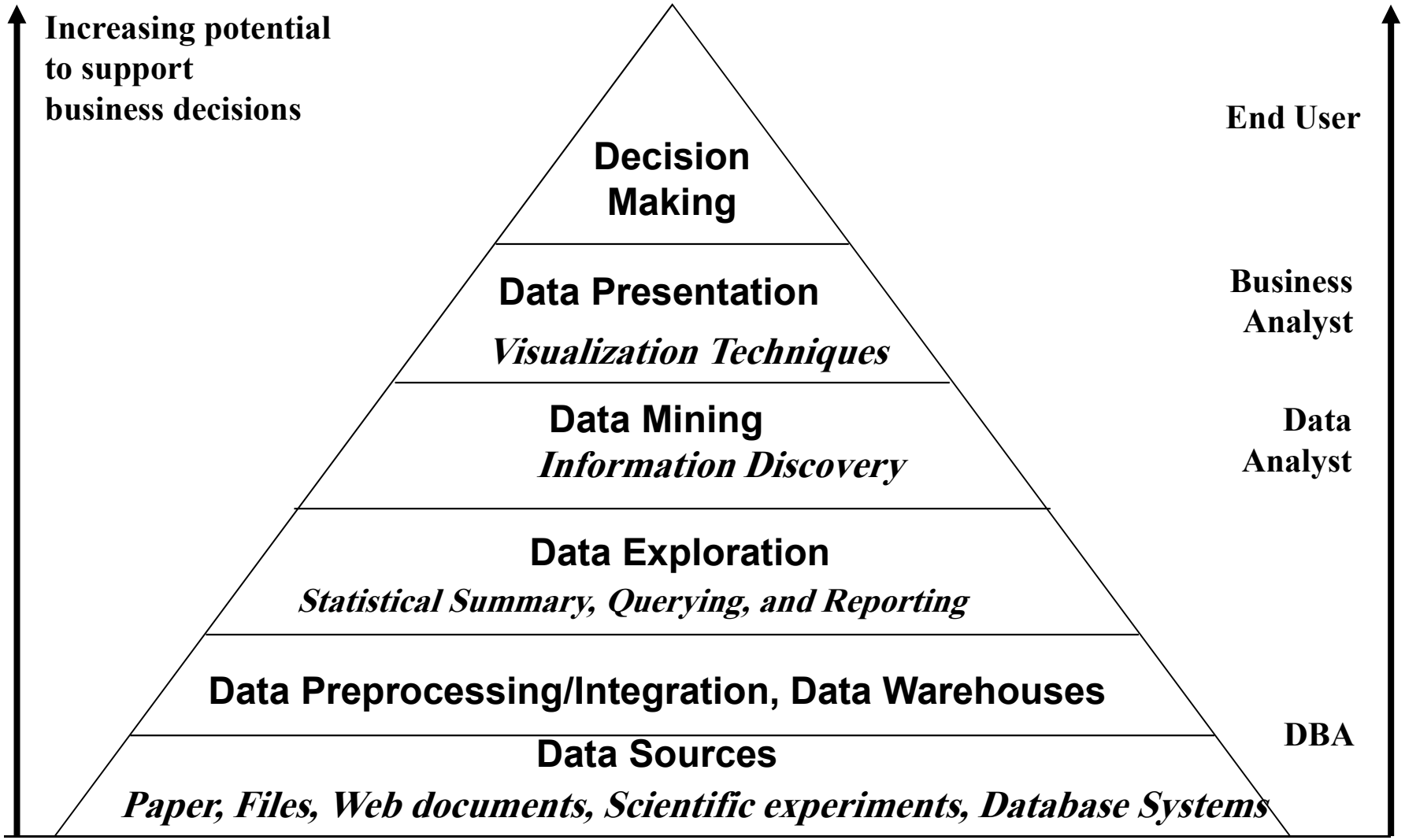
- Consider for example, the following table that contains data about objects; shape, color, and weight.

- Pattern
  - **Most Boxes are Red.**
- We can represent Pattern as rule:
- If Shape = Box

  then Color = Red.

Row #	Shape	Color	Weight
1->	Box	Red	100
2->	Box	Red	200
3->	Box	Red	300
4	Box	Blue	400
5	Cone	Blue	400

Data Mining and Business Intelligence



Need for Data Mining(1)

- Large **quantities** of data is being accumulated.
- Data could be large in two senses.
  - In terms of size, e.g. for Image Data
  - or in terms of dimensionality, e.g. for Gene expression data.

Need for Data Mining(2)

- A **huge gap** from the stored data to the knowledge that could be construed from the data.
- Data **analysis** for large data analysis.
- **New demands**, Data Mining techniques are now being applied to all kinds of domains.

Data Mining Tasks

- Data mining tasks are the kind of data **patterns that can be mined.**
- Data Mining functionalities are used to specify the kind of patterns to be found in the data mining tasks.

Data Mining Tasks

- In general data mining tasks can be classified into two categories:
  - **Descriptive mining** tasks characterize the **general properties** of the data.
  - **Predictive mining** tasks perform **inferences** on the current data in order to make **predictions**.

Data Mining Tasks

- Most famous data mining tasks:
  - Classification [Predictive]
  - Prediction [Predictive]
  - Association Rules [Descriptive]
  - Clustering [Descriptive]
  - Outlier Analysis [Descriptive]

Data Mining challenges

- **Scalability**: Scalable techniques are needed to handle the massive **scale of data**
- **Dimensionality**: Many applications may involves a large number of dimensions (e.g. features or attributes of data)

Data Mining challenges

- **Heterogeneous and Complex Data:** In recent years complicated data types such as graph-based, text-free and structured data types are introduced. Techniques developed for data mining must be able to handle the **heterogeneity of the data.**

Data Mining challenges

- **Data Quality:** Many data sets are imperfect due to present of missing values and noise un the data. To handle the imperfection, robust data mining algorithms must be developed.

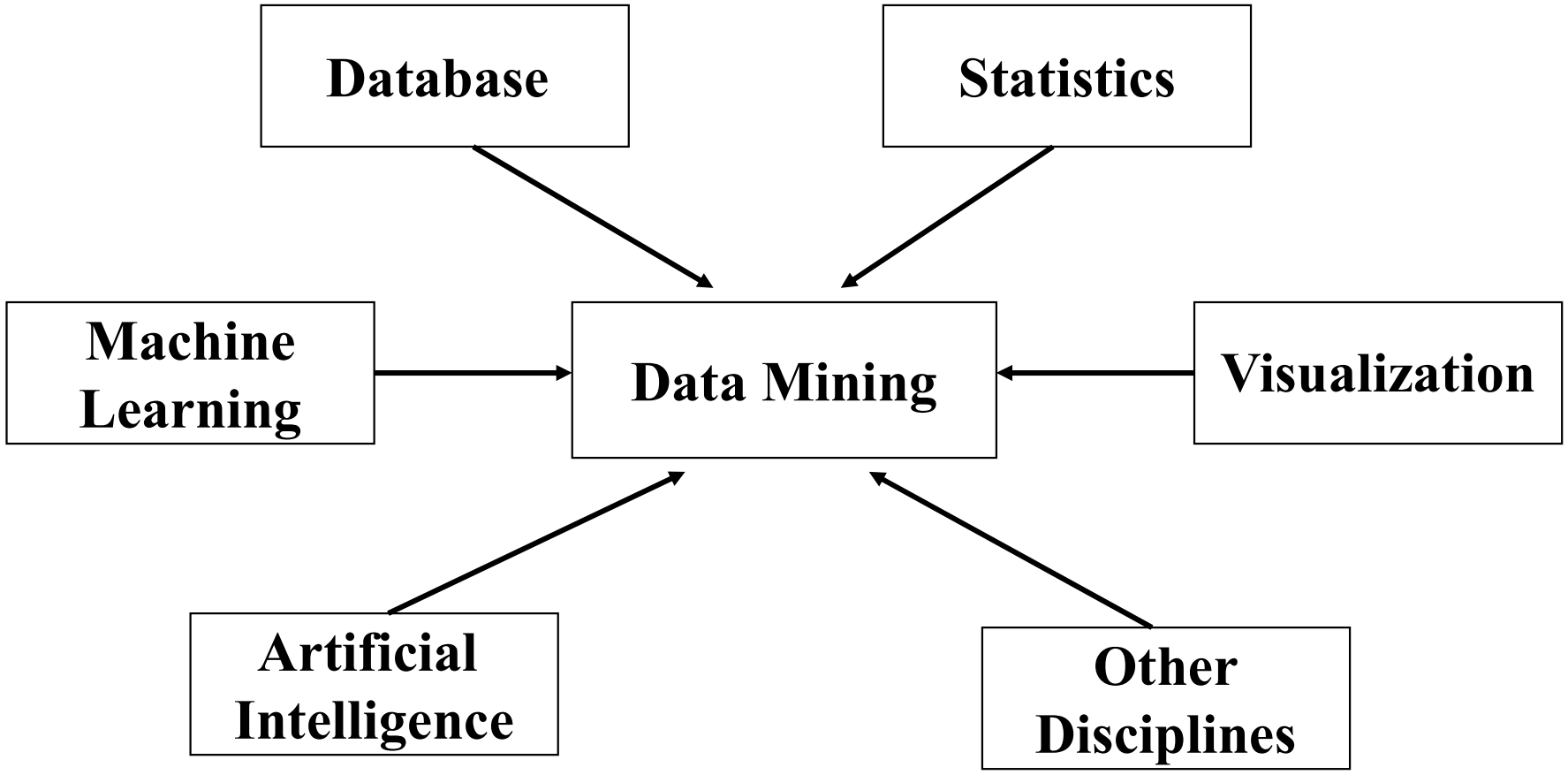
Data Mining challenges

- **Data Distribution:** As the volume of data increases , it is no longer possible or safe to keep all the data in the same place. As a result , the need for **distributed data mining techniques** has increased over the years.

Data Mining challenges

- **Privacy Preservation:** While privacy intends to prevent the disclosure of information, data mining attempts to revel interesting knowledge about data. As a result, there is growing interest in developing privacy-preserving data mining algorithms

Data Mining as an Interdisciplinary field



Data Mining as an Interdisciplinary field

- **Statistics**: Data Mining in Statistics deals with finding useful patterns in data sets.

- **Relational Databases:** Database part of data Mining that provide the fast and reliable access to data.
  - It used for data operation (Storing and retrieving data), Data Mining for Decision making.

Data Mining as an Interdisciplinary field

- **Artificial Intelligence:** Knowledge acquisition, maintenance and application are other branches of Artificial Intelligence, which are highly related with Databases and also with Data Mining.

Data Mining as an Interdisciplinary field

- **Machine Learning:** focuses on complex representations and search methods for specialized data-intensive problems.
  - Data Mining uses methods from Machine Language such as **decision tree and neural nets.**

Data Mining as an Interdisciplinary field

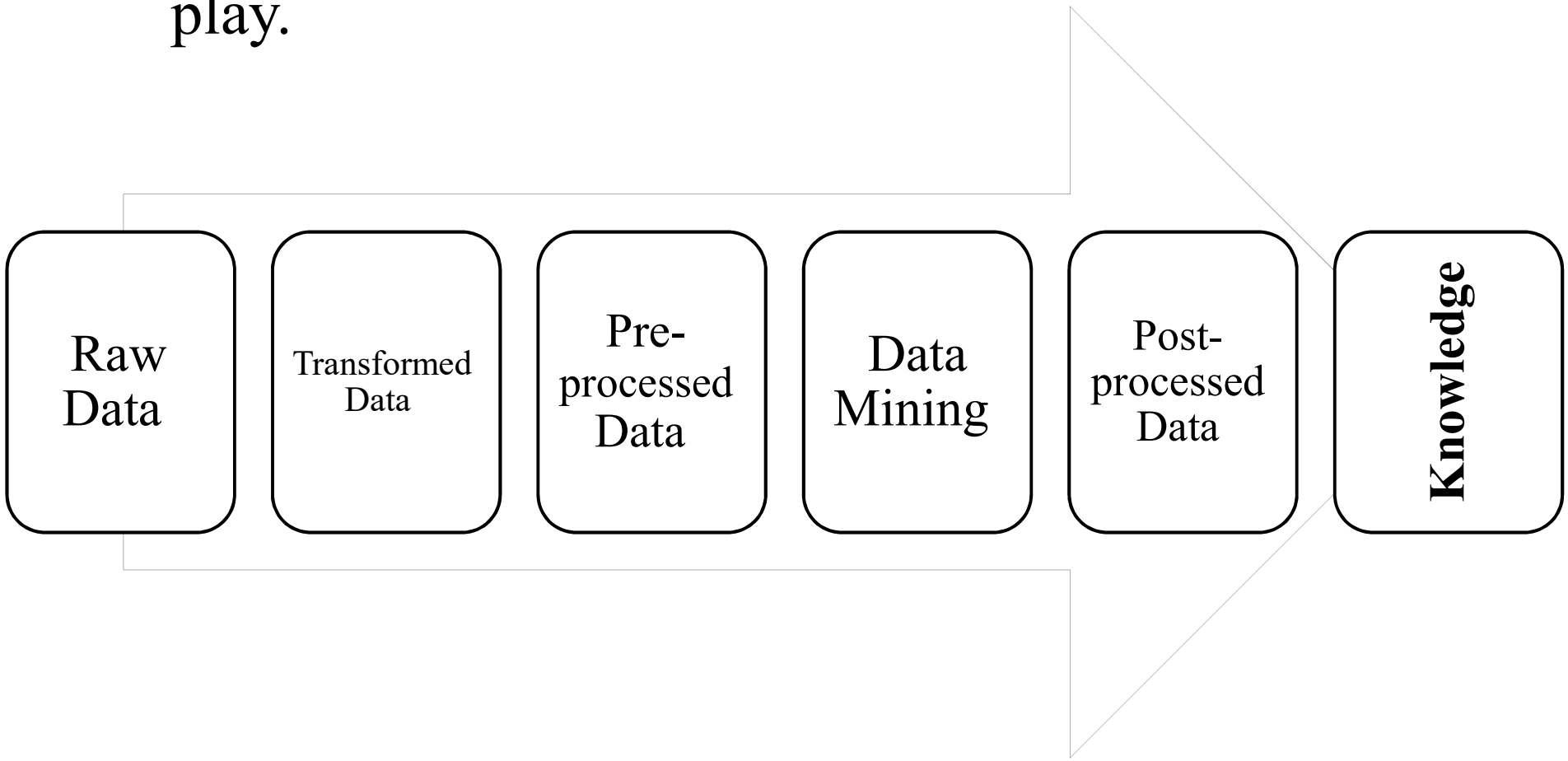
- **Visualization :** is used to gain visual insights into the structure of the data.
  - Visualization is abundantly used as a pre- and post-processing tool for data mining.

Data Mining as an Interdisciplinary field

- **Knowledge Representation**
  - Knowledge presentation is the framework that converts a large amount of data into a particular data or procedure that human being can figure out based on an intention.
  - In Knowledge representation visualization tools and knowledge representation techniques are used to present the mined knowledge to the user.

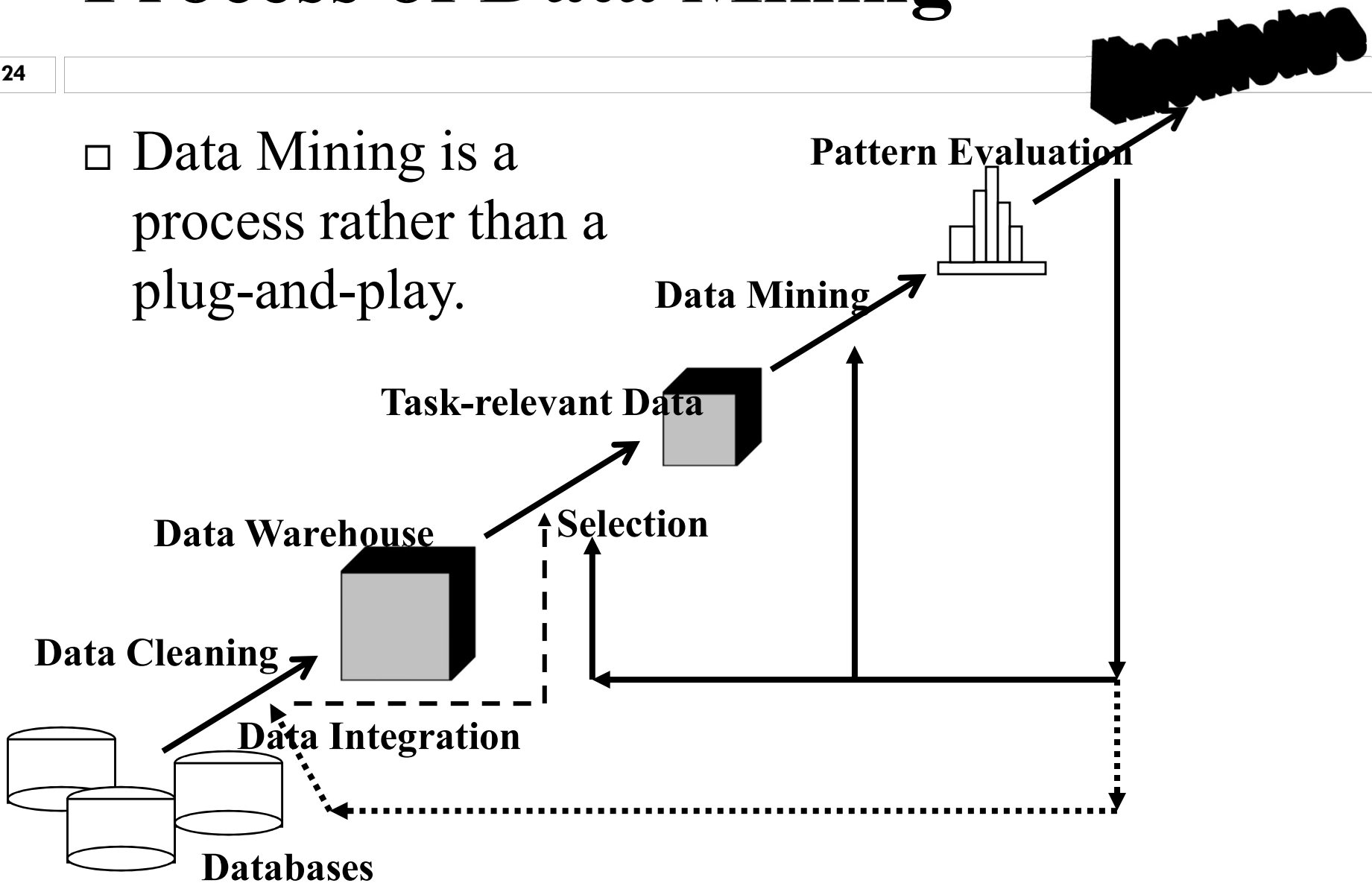
Process of Data Mining

- Data Mining is a process rather than a plug-and-play.



Process of Data Mining

- Data Mining is a process rather than a plug-and-play.



Process of Data Mining

1- Data cleaning:

Real-world data tends to be *incomplete*, *noisy* and *inconsistent*.

- incomplete: lacking attribute values, lacking certain attributes of interest,
  - e.g., occupation=“ ” (missing data)
- noisy: containing noise, errors, or outliers
  - e.g., Salary=“−10” (an error)
- inconsistent: containing discrepancies in codes or names, e.g.,
  - e.g., Age=“42” Birthday=“03/07/1997”

Process of Data Mining

2- Data Integration:

Data integration is the merging of data from multiple sources. These sources may include multiple databases, data cubes, or flat files.

Process of Data Mining

3- Data Selection:

Where data relevant to the analysis task are retrieved from the database. Therefore, irrelevant, weakly relevant or redundant attributes may be detected and removed.

Process of Data Mining

4- Data Transformation

Where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operation (for example daily sales may be aggregated to monthly sales or annual sales), Generalization (for example, city may be generalized to country or age may generalized to young , middle- age, senior) .

Process of Data Mining

5- Data Mining:

An essential process where intelligent methods are applied on data to covert it to knowledge in for decision making. Wide range of methods can be used in data mining such neural nets, decision tree and Association.

Process of Data Mining

6- Pattern evaluation :

To identify the truly interesting pattern based on some interestingness measures. A pattern consider interesting if it is:

- *Valid*
- *Novel*
- *Actionable*
- *Understandable*