DM LAB I

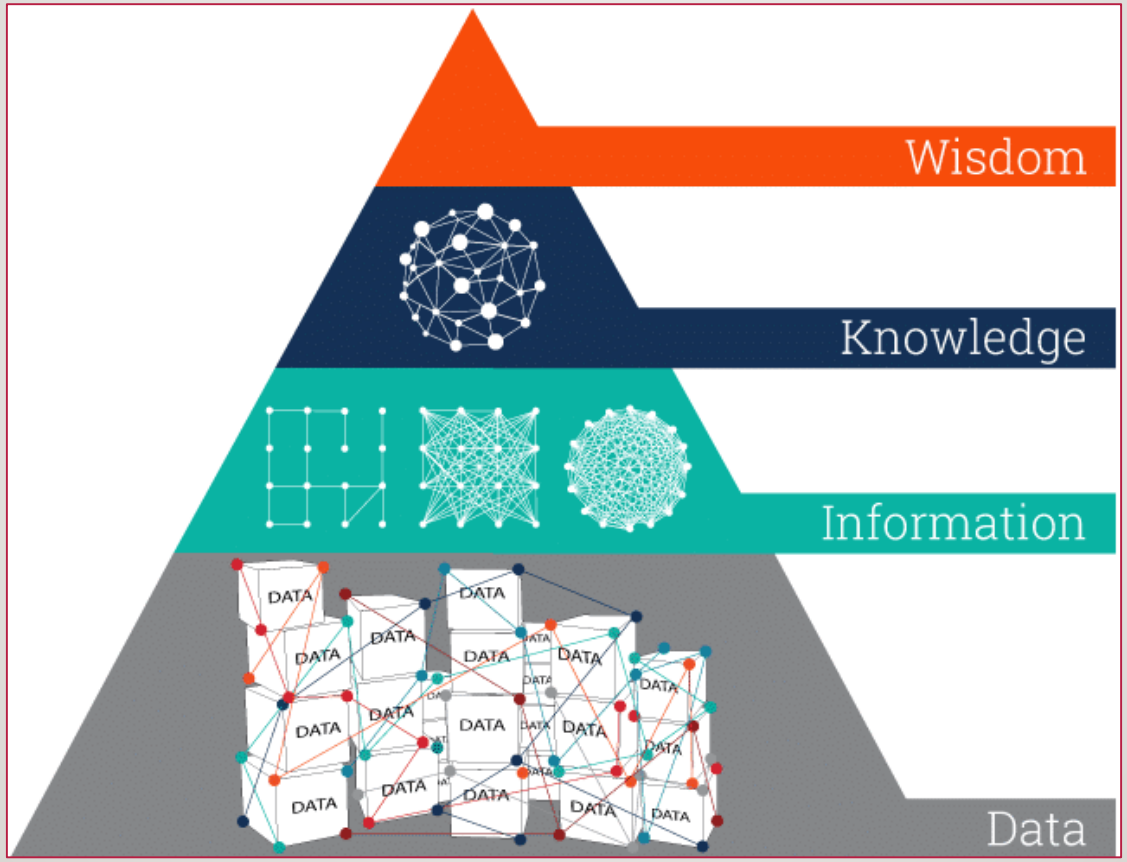
BY

ENG. JOUD KHATTAB

2 INTRODUCTION

- DIKW Pyramid.
- Terminologies.
- Data Warehousing Steps.
- Data Mining Definition.
- Prerequisites.
- Course Topics.
- Lab & Exercises.
- Tools.

3 DIKW PYRAMID



4 TERMINOLOGIES

Big Data

Data Science

Data Mining

Data Analysis

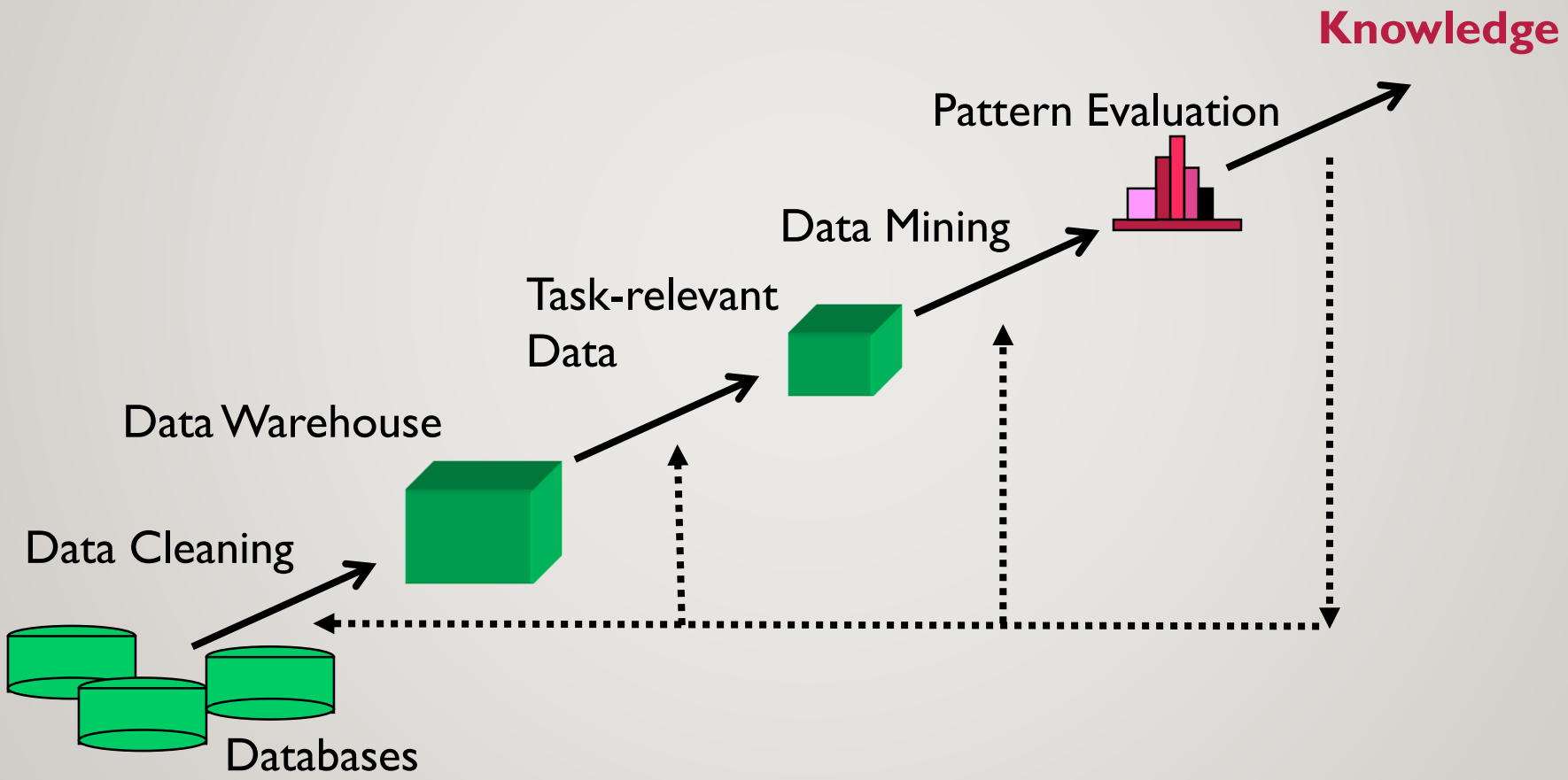
Business Intelligence

Data Warehousing

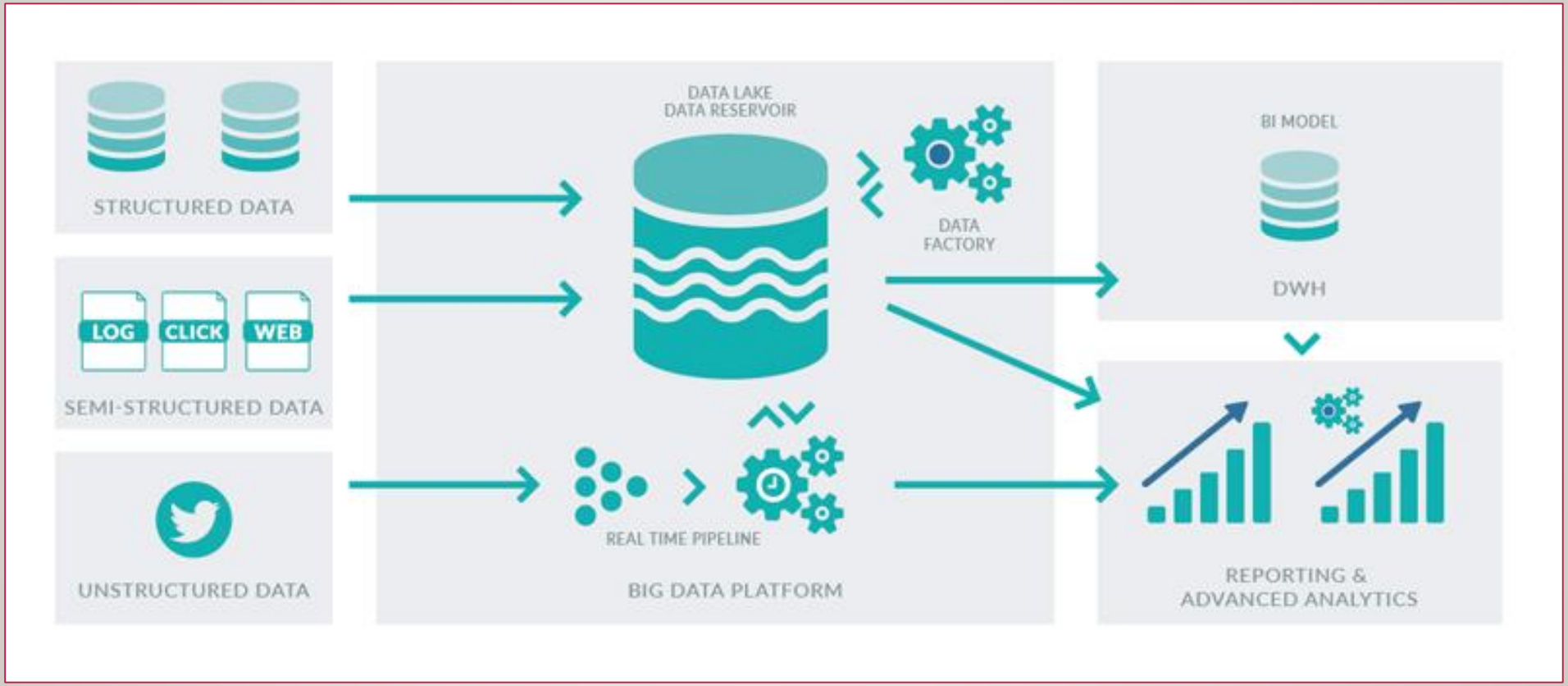
Artificial Intelligence

Statistics

5 KNOWLEDGE DISCOVERY IN DATABASES (KDD) PROCESS



6 DATA WAREHOUSING STEPS



DATA MINING DEFINITION

- The process of collection, searching through, and analyzing a large amount of data in a database, to discover patterns or relationships.

8 PREREQUISITES

- Data mining is a broad field that combines techniques from different areas in computer science and statistics.
- Basic background knowledge in the following areas:
  - Database Systems
    - Data models, query languages, SQL, conceptual database design, transactions
  - Statistics
    - Expectation, basic probability, distributions, hypothesis tests
  - Linear Algebra
    - Vectors and matrices, vector spaces, basis, matrix inversion, solving linear equations
  - Algorithms and Data Structures
    - basic data structures and the understanding of written algorithms in pseudocode

9 COURSE TOPICS

1) Introduction
2) Data Preprocessing
3) Data Warehousing and OLAP
4) Association, correlation, and frequent pattern analysis
5) Classification
6) Cluster and Outlier Analysis
7) Mining Time-Series and Sequence Data
8) Text Mining and Web Mining
9) Visual Data Mining

COURSE TOPICS DESCRIPTION
1) INTRODUCTION

1. Concepts Of Data Mining
2. Knowledge Discovery (KDD) Process
3. Mining On Different Kinds Of Data

   Relational, transactional, object-relational, heterogeneous, spatiotemporal, text, multimedia, Web, stream, mobile, and so on.
4. Mining For Different Kind Of Knowledge

   Classification, regression, clustering, discriminant, outliers, and so on.
5. Evaluation Of Knowledge

   Quality of knowledge, including accuracy, and relevance (such as correlation).
6. Applications Of Data Mining

   market analysis, bioinformatics, homeland security, and so on.

11 COURSE TOPICS DESCRIPTION
2) DATA PREPROCESSING

1. Descriptive Data Summarization

   Computing the measures of: mean, mode, quantiles, boxplots, variances, standard deviation, outliers.

   Graphic statistical display: histogram, scatter plot, boxplot, quantile plot, local regression curves.

2. Data Cleaning Methods

   Techniques for handling missing values, noisy data, and inconsistent data.

3. Data Integration And Transformation Methods

   Data smoothing, data aggregation, data generalization, normalization, feature construction.

4. Basic Data Reduction Methods

   It introduces binning (histograms), sampling, and data cube aggregation.

COURSE TOPICS DESCRIPTION
3) DATA WAREHOUSING AND OLAP

1. Concept And Architecture Of Data Warehouse
2. The Dimensional Data Model

   including dimensions and measures.

   star schema, snowflake schema, and fact constellations.

   data cube concept & concept hierarchies in the cube.
3. OLAP Operations In The Multidimensional Data Model

   drill-down, roll-up, slice and dive, pivot

COURSE TOPICS DESCRIPTION
4) ASSOCIATION

1. Basic Concepts

   frequent patterns, associations, support and confidence of association rules, correlation measure, other objective functions or measures, a typical application scenario (market basket analysis).

2. Frequent Pattern Mining Methods

   The Apriori algorithm, improvements to Apriori, max-patterns, closed patterns, and top-k patterns.

3. Mining Various Kinds Of Frequent Patterns

   Multilevel and multidimensional association rules, Quantitative association rules, Correlation analysis.

4. Applications Of Association Rules

COURSE TOPICS DESCRIPTION
5) CLASSIFICATION

1. Evaluation Of Classification

   evaluation metric, validation for model selection, overfitting

2. Bayesian Classification

   Bayes theorem, Naive Bayesian classification methods

3. Decision Tree And Decision Rule

   attribute selection and reduction, basic top-down classification-tree induction schema, pre/post-pruning uninformative subtrees, extraction of rules from classification trees, decision rule induction

4. Linear models for classification

   linear discriminant analysis, classification by SVM (Support Vector Machine) analysis

5. Basic Concepts Of Nonlinear Classification

   neural network, SVM with nonlinear Kernels

COURSE TOPICS DESCRIPTION
6) CLUSTER AND OUTLIER ANALYSIS

1. Concept Of Cluster Analysis
2. Types Of Data And For Dissimilarity Computation

   Interval-scaled variables, binary variables, nominal, ordinal, and ratio-scaled variables, and variables of mixed types.
3. Categorization Of Major Clustering Methods
   1. Partition-based clustering
   2. Hierarchical clustering
   3. Density-based clustering
   4. Model-based clustering
4. Outlier Analysis

   Concepts and basic outlier detection methods

COURSE TOPICS DESCRIPTION
7) MINING TIME-SERIES AND SEQUENCE DATA

1. Regression Analysis
2. Trend Analysis
3. Sequential Pattern Mining

COURSE TOPICS DESCRIPTION
8) TEXT MINING AND WEB MINING

1. Mining Text Databases

   Text data analysis and information retrieval, keyword-based association analysis, document classification, text clustering analysis

2. Mining The Web

   Mining the Web's link structures, automatic classification of Web documents, mining social networks, Web resource discovery, Web usage mining

COURSE TOPICS DESCRIPTION
9) VISUAL DATA MINING

1. Data Visualization

19 LABORATORIES AND EXERCISES

1. Learn to use data mining systems by using some data mining and data warehousing software's
   - Microsoft SQL Server (Analysis manager), Oracle (data mining part), IBM Intelligent-Miner, and statistics analysis software tools.
2. Implement some data mining functions
   - including association mining, classification, clustering, sequential pattern mining, text-mining, Web mining, spatial data mining
3. Implementation, refinement, and performance comparison of several different data mining methods
4. Proposal, implementation and testing of new data mining algorithms and functions
5. Using some sample data sets to implement and test data mining functions

20 TOOLS


Microsoft
SQL SERVER
Analysis Services


Visual Studio


R


WEKA
The University
of Waikato

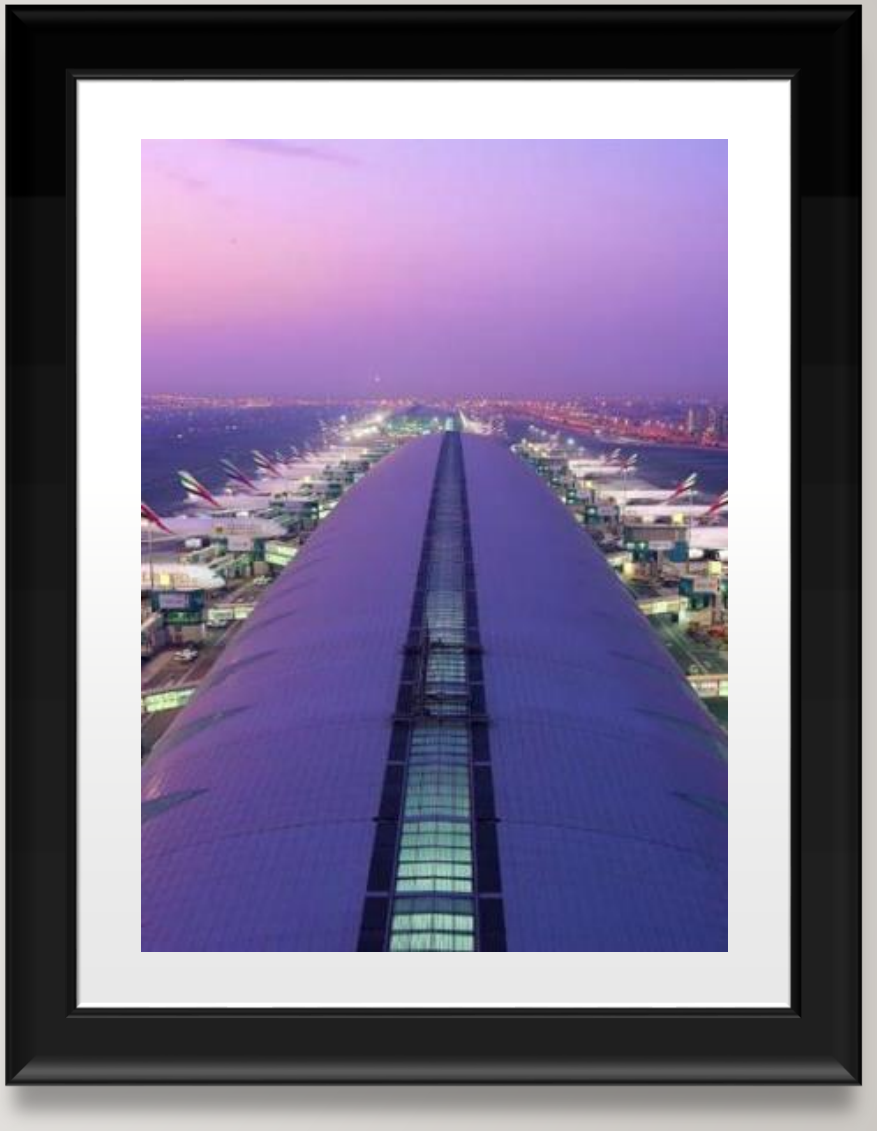

python


Power BI


tableau

CASE STUDY

How Dubai Airports has analyzed real-time data to improve its services



THE END