TU

جامعة الطائف
TAIF UNIVERSITY

DATA MINING

Assoc. Prof. Dr. Salha Alzahrani

College of Computers and Information Technology
Taif University
Saudi Arabia

s.zahrani@tu.edu.sa

Data Mining Intro. | Data mining and knowledge discovery

Recap of Previous Lecture

Content of This Lecture

Summary & Checklist

Recap of Lecture 0

- Data mining course synopsis (summary).
- Course syllabus (code, credits, pre-requisites, instructor’s details, course objectives, learning outcomes, lecture plan, and course policies).
- Assessment of the course.
- Course website.
- Student Portfolio.
- Glossary and Academic Vocabulary List.

twitter

@SalhaAlzahrai

Data Mining Intro. | Data mining and knowledge discovery

Recap of Previous Lecture

Content of This Lecture

Summary & Checklist

Content of Lecture 1

- Black-Box
- Motivation: Why data mining?
- Evolution of sciences
- Evolution of database technology
- What is data mining?
- Knowledge discovery, Knowledge discovery in databases
- Data mining and business intelligence
- Why not traditional data analysis?
- Multi-dimensional view of data mining
- Conferences and Journals on data mining
- Where to Find References? DBLP, CiteSeer, Google
- Applications of data mining
- Summary & Checklist.

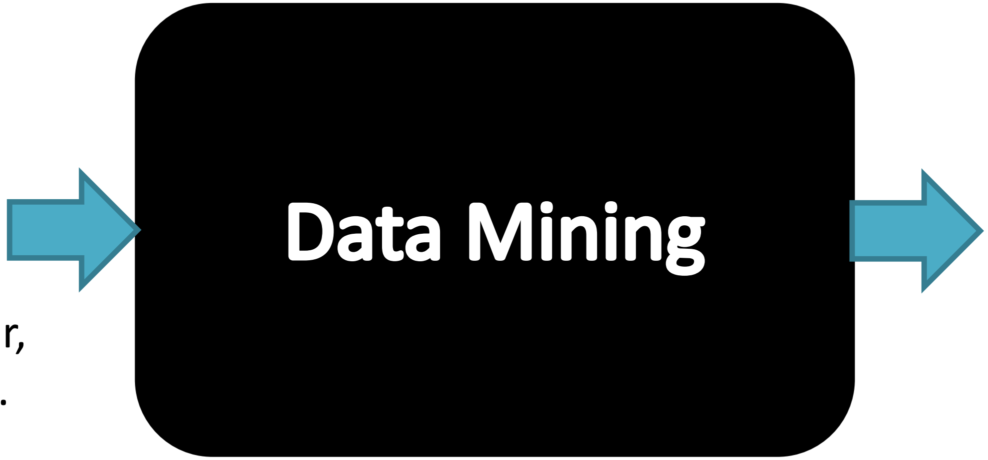
Data Mining Intro. | Black-Box

Input(s)

Output(s)

Data

Types: binary, numbers, character, texts, objects, etc.
From: business, science, society and everyone!



Data Mining

Patterns

Descriptive: e.g. credit card fraud detection.
Predictive: e.g. medical diagnosis!

BLACK-BOX DESIGN OF DATA MINING

Data Mining Intro. | Why Data Mining?

- The Explosive Growth of Data: from gigabytes to terabytes to petabytes.
  - Data collection and data availability:
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data:
    - **Business:** Web, e-commerce, transactions, stocks, …
    - **Science:** Remote sensing, bioinformatics, scientific simulation, …
    - **Society and everyone:** news, digital cameras, YouTube
- We are drowning in **data**, but starving for **knowledge**!
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets.

Data Mining Intro. | Evolution of Sciences

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
  - Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. **Data mining** is a major new challenge!
- Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

Data Mining Intro. | Evolution of Database Technology

1960s:

Data collection, database creation, IMS and network DBMS

1970s:

Relational data model, relational DBMS implementation

1980s:

RDBMS, advanced data models (extended-relational, OO, deductive, etc.)

Application-oriented DBMS (spatial, scientific, engineering, etc.)

1990s:

Data mining, data warehousing, multimedia databases, and Web databases

2000s

Stream data management and mining

Data mining and its applications

Web technology (XML, data integration) and global information systems

Data Mining Intro. | What Is Data Mining?

Data mining (knowledge discovery from data)

Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.

Alternative names

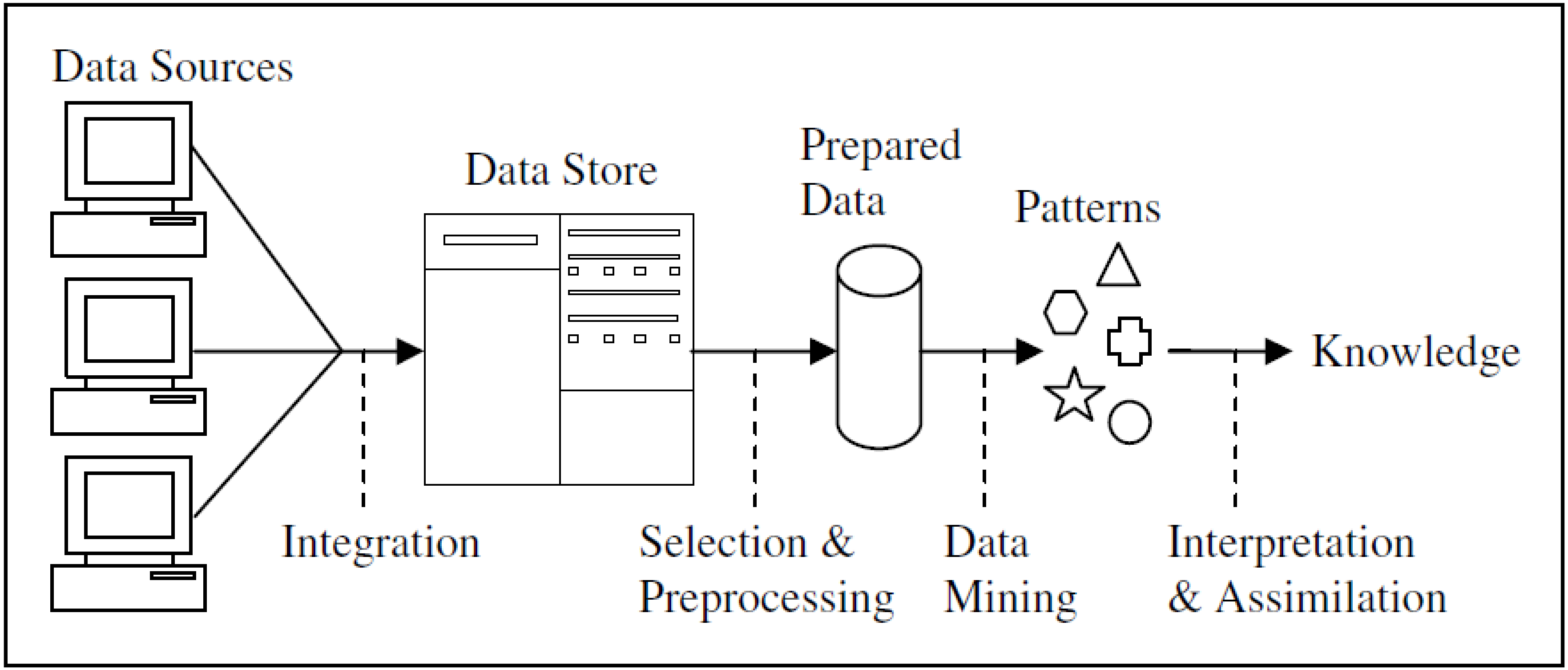
Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Watch out: Is everything "data mining"?

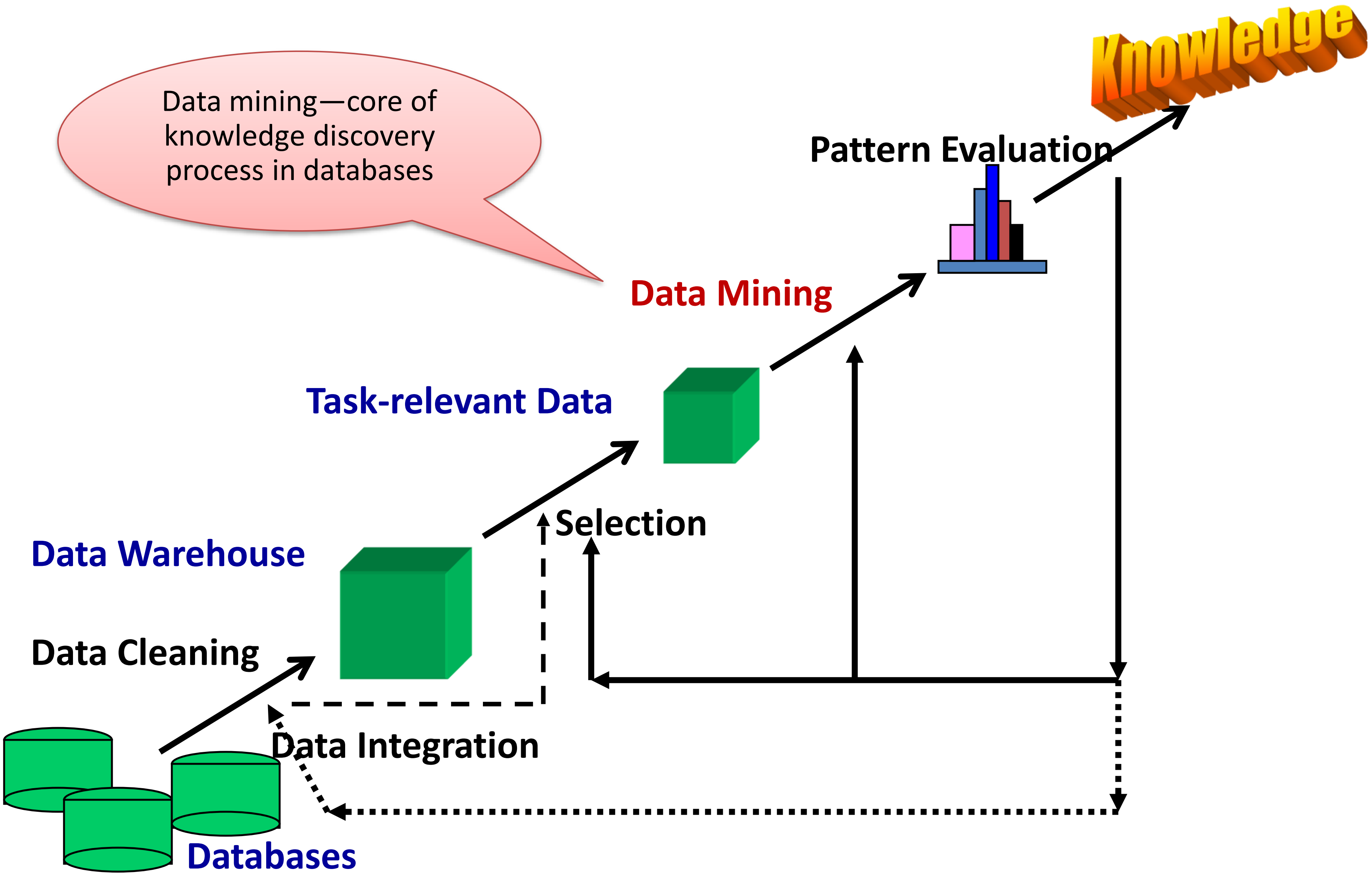
- Simple search and query processing ?
- (Deductive) expert systems ?

Data Mining Intro. | Knowledge Discovery

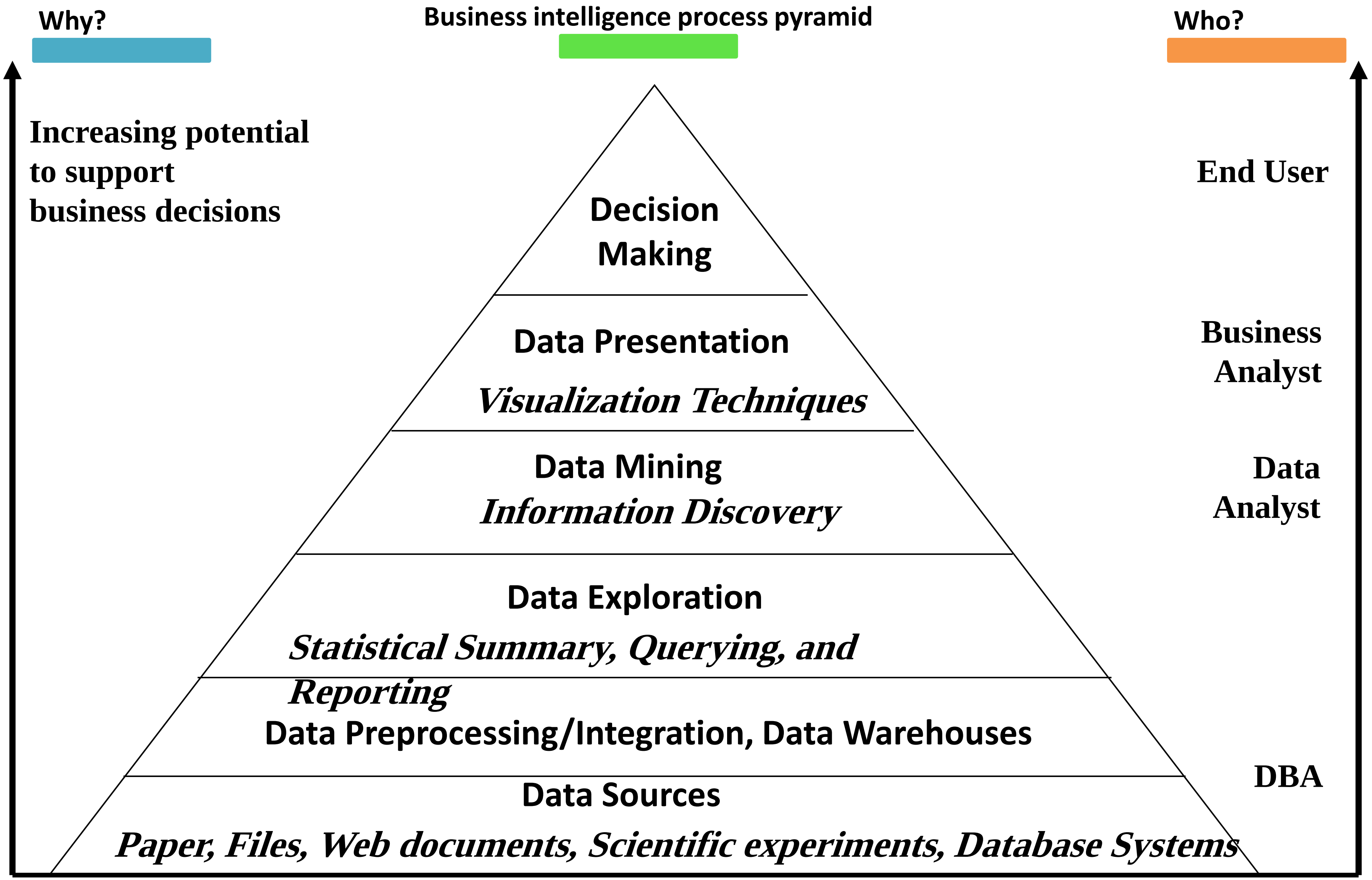
Knowledge Discovery in data is a process of which data mining forms just one part, albeit a central one.



Data Mining Intro. | Knowledge Discovery in Databases (KDD)



Data Mining Intro. | Data Mining and Business Intelligence



Why?

Business intelligence process pyramid

Who?

Increasing potential
to support
business decisions

Decision Making

Data Presentation

Visualization Techniques

Data Mining

Information Discovery

Data Exploration

Statistical Summary, Querying, and Reporting

Data Preprocessing/Integration, Data Warehouses

Data Sources

Paper, Files, Web documents, Scientific experiments, Database Systems

End User

Business Analyst

Data Analyst

DBA

Data Mining Intro. | Why Not Traditional Data Analysis?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as terabytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

Data Mining Intro. | Multi-Dimensional View of Data Mining

Data Knowledge Techniques Applications

- **Data** to be mined
  - Relational, data warehouse, transactional, stream, object-oriented/relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW
- **Knowledge** to be discovered
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques** utilized
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, etc.
- **Applications** adapted
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Data Mining Intro. | Conferences and Journals on Data Mining

- **KDD Conferences**
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - SIAM Data Mining Conf. (SDM)
  - (IEEE) Int. Conf. on Data Mining (ICDM)
  - Conf. on Principles and practices of Knowledge Discovery and Data Mining (PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)

- **Other related conferences**
  - ACM SIGMOD
  - VLDB
  - (IEEE) ICDE
  - WWW, SIGIR
  - ICML, CVPR, NIPS
- Journals
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

Data Mining Intro. | Where to Find References? DBLP, CiteSeer, Google

- Data mining and KDD (SIGKDD: CDROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- Web and IR
  - Conferences: SIGIR, WWW, CIKM, etc.
  - Journals: WWW: Internet and Web Information Systems,
- Statistics
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization
  - Conference proceedings: CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

Data Mining Intro. | Data Mining Applications

There is a rapidly growing body of successful applications in a wide range of areas as diverse as:

- analysis of organic compounds
- automatic abstracting
- credit card fraud detection
- electric load prediction
- financial forecasting
- medical diagnosis
- predicting share of television audiences
- product design
- real estate valuation
- targeted marketing
- thermal power plant optimisation
- toxic hazard analysis
- weather forecasting

and many more.

Data Mining Intro. | Data Mining Applications : Examples

Some examples of applications (potential or actual) are:

- a supermarket chain mines its customer transactions data to optimise targeting of high value customers
- a credit card company can use its data warehouse of customer transactions for fraud detection
- a major hotel chain can use survey databases to identify attributes of a ‘high-value’ prospect
- predicting audience share for television programmes
- to arrange show schedules to maximise market share and increase
- advertising revenues
- predicting the probability that a cancer patient will respond to chemotherapy, thus reducing health-care costs without affecting quality of care.

and many more.

Recap of Previous Lecture

Content of This Lecture

Summary & Checklist

Summary & Checklist

- ☑ Black-Box
- ☑ Motivation: Why data mining?
- ☑ Evolution of sciences
- ☑ Evolution of database technology
- ☑ What is data mining?
- ☑ Knowledge discovery, Knowledge discovery in databases
- ☑ Data mining and business intelligence
- ☑ Why not traditional data analysis?
- ☑ Multi-dimensional view of data mining
- ☑ Conferences and Journals on data mining
- ☑ Where to Find References? DBLP, CiteSeer, Google
- ☑ Applications of data mining
- ☑ Summary & Checklist.

Reminder | Student Portfolio

- **Each student should prepare her own course portfolio!**
- **Portfolios should include the following parts:**
  1) *Course Syllabus*
  2) *Lecture notes (slides)*
  3) *Assignments*
  4) *Quizzes*
  5) *Mid-term exam and answer sheet.*
  6) *Research articles and other supporting materials.*
  7) *Lab lecture notes, exercises, and MATLAB codes.*
  8) *Glossary*
- **Portfolios will be checked regularly by the instructor.**
- **Students who prepare good course portfolios may be given a BONUS +2/+5 on their examinations, if needed.**

Reminder | Next Lecture !

Next Lecture...

Data Processing: Data Cleaning, Preparation, Dealing with Missing Data, and Attributes Reduction (Ch. 1)

- *Be ready!*
- *Prepare your glossary and academic vocabulary lists.*
- *Download & print the lecture notes before your class.*

Thank You !

✉ s.zahrani@tu.edu.sa