University of Reading

CSMDM21 - Data Analytics and Mining

Introduction to Data Analytics and Mining

Module convenor
Dr. Carmen Lam
carmen.lam@reading.ac.uk
Department of Computer Science

Lecture notes and videos powered by
Prof. Giuseppe Di Fatta

Why Mining Data? Commercial Viewpoint

Drowning in Data
but
Starving for Knowledge!

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

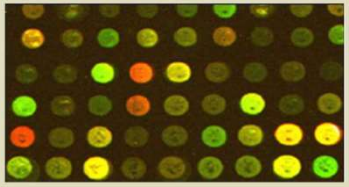
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

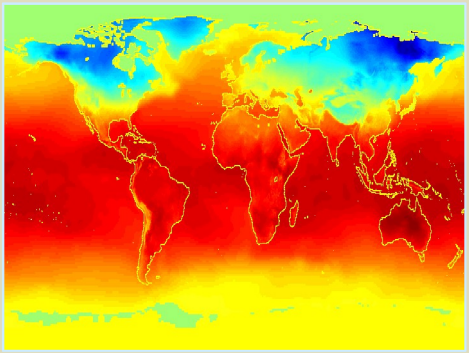
Why Mining Data? Scientific Viewpoint

Drowning in Data
but
Starving for Knowledge!

- 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 tera/petabytes of data
- Traditional techniques infeasible for raw data
- Data Mining may help scientists
  - in classifying and segmenting data
  - in hypothesis formation

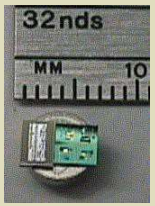
In 2017 the CERN Data Centre in Geneva passed the 200-petabyte milestone.

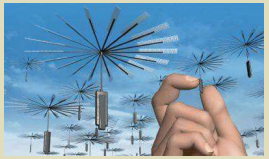




The ECMWF Data Centre in Reading holds several hundreds petabytes of weather data.

The Dawn of the Information Age





- From Smart Dust to Smart Cities:
  - RF Tags: radiofrequency tags require no battery to read and operate.
  - Smart Dust: miniature machines, each the size of a dust mote, may eventually saturate the environment, invisibly performing countless tasks.
  - IoT, Smart Cities, Cloud/Edge Computing
- Evolution of digital technologies generating huge amounts of data:
  - Online Social Media
  - Smart Cities, Smart Buildings, Smart objects
  - Augmented reality (e.g. Ingress/Pokemon Go)

What is Data Mining?

- Data Mining:
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases
- Learning and describing concepts from data
- Alternative names:
  - Data Mining: a misnomer?
  - Knowledge discovery in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, business intelligence, etc.

Data Mining: Basic Design Decisions

- Design space in Data Mining
  - Kinds of data and databases available/needed
  - Kinds of knowledge to be discovered
  - Kinds of algorithms/techniques utilised
  - Kinds of applications
- Data Mining tasks
  - Descriptive Data Mining
  - Predictive Data Mining

Data Mining Tasks

- **Predictive Tasks**

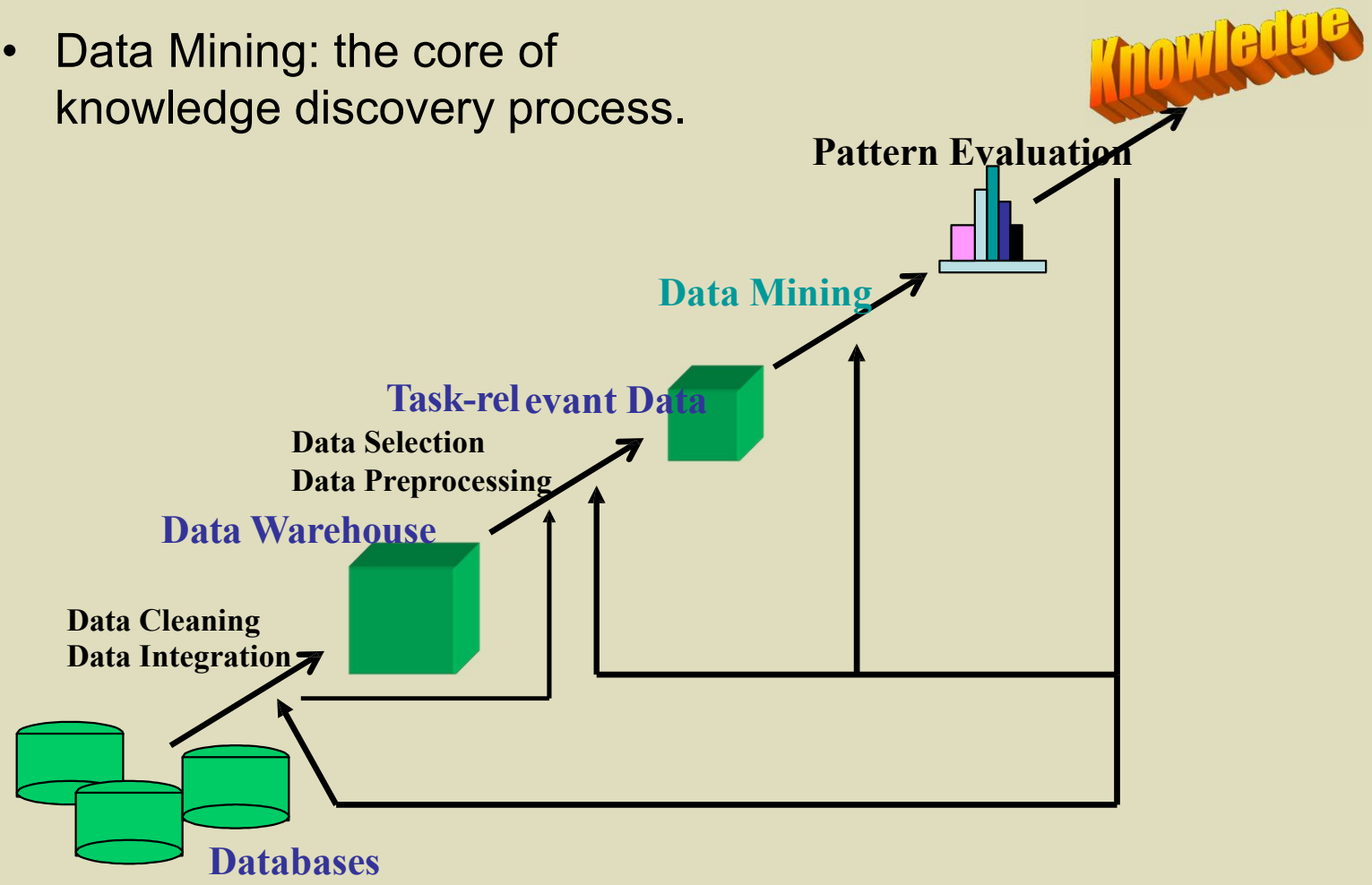
  Use some variables to predict unknown or future values of other variables
- **Descriptive Tasks**
  - Find human-interpretable patterns that describe the data.

Common Data Mining tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Data Mining: A KDD Process

- Data Mining: the core of knowledge discovery process.



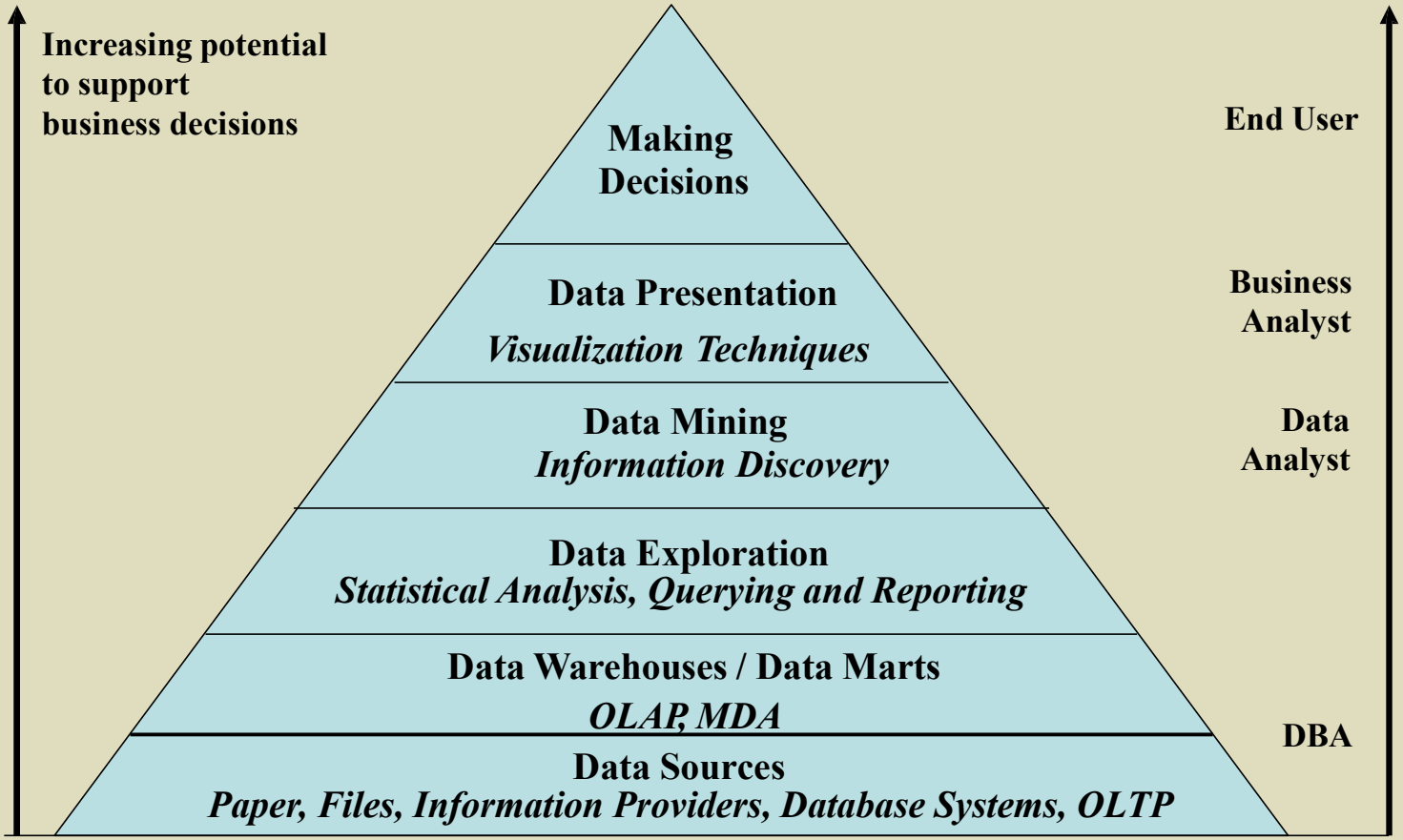
Steps of a KDD Process

1. data gathering
2. data cleansing
3. data transformation
4. selecting techniques
5. applying Data Mining
6. processing results

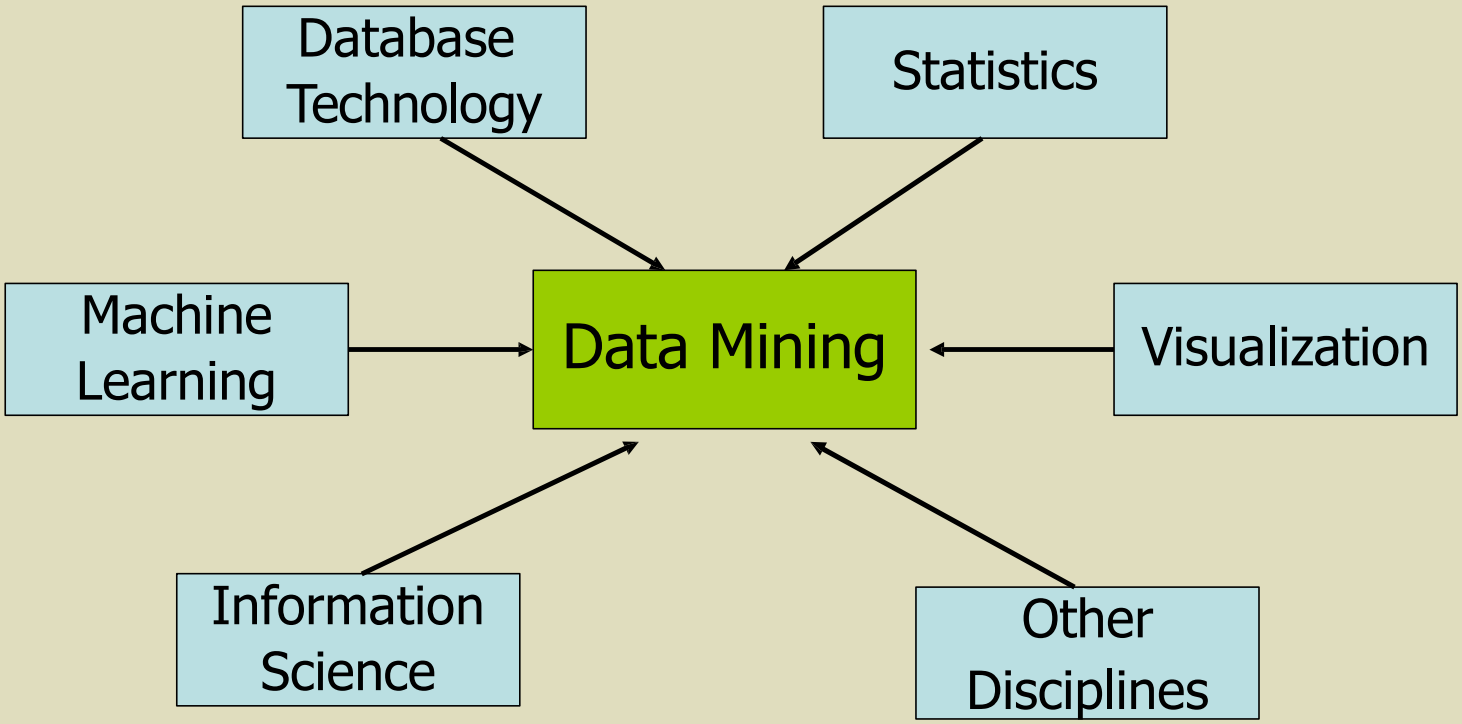
Steps of a KDD Process

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

Data Mining and Business Intelligence



Data Mining: Confluence of Multiple Disciplines



Next video lecture:

➢Introduction to Data Science Platforms