Data Science – Semester 5 – Fall 2020/2021

INTRODUCTION TO DATA MINING & WAREHOUSING

Lecture 1 - Introduction

Course Format & Delivery method

- Live sessions (1h)
  - Monday: 10:00 – 11:00
  - Wednesday: 11:30 – 12:20
- Live sessions will be recorded and posted on MS teams
- Course consists of:
  - **Theoretical lectures** (PowerPoint presentations)
  - **Practical sessions** (Practice exercises + Python tutorials)
  - **Online forums & Discussions**: You will have to actively participate in a set of discussion forums throughout the semester: ask questions, answer other's questions

Assessments Methods

- Assignments : 40%
- Final exam: 60%

- 4 Assignments (10% each)
  - **3 programming assignments** with Python **(individual)**
  - One of the following two assignments:
    - **Read and present a research paper (individual)**
    - **Write a project proposal**: if you have a novel idea of an application of data mining or you want to develop a new software/product for data mining, you can write a report that consists of 2-3 pages where you describe your proposal **(groups of maximum 3 students)**

Specific Requirements

- Assignments submitted after the deadline will be **refused**.

- In case of **plagiarism/cheating**, all students who are involved in the activity will get the **mark 0** on the corresponding assignment.

Covered Topics

- **Week 1: Introduction to Data Mining**
- **Week 1,2: Python Programming Basics**

- **Week 3: Data Exploration**
- **Week 4: Data preprocessing (Assignment 1)**

- **Week 5,6: Data warehousing and OLAP (Assignment 2)**

- **Week 7,8: Association Rule Mining**
- **Week 9,10: Data Classification (Assignment 3)**
- **Week 11,12: Cluster Analysis**

Textbooks

- Data mining:
  - Data Mining: Concepts and Techniques: A volume in The Morgan Kaufmann Series in Data Management Systems Book, 3rd Edition, 2012, by Jiawei Han
  - Introduction to Data Mining: by Pang-Ning Tan, Michael Steinbach, Vipin Kumar
- Python:
  - Python for Data Analysis by Wes McKinney
  - Python for Data Mining Quick Syntax Reference 1st ed. Edition

Required Software

- Microsoft SQL Server 2016 with Integration Services installed
- Python IDE: Jupyter Notebook from Anaconda
- Download and installation will be explained during live sessions

Data Mining

What data mining is and why we need it

Need for Data Mining

- Data are being gathered and stored extremely fast

http://www.internetlivestats.com/one-second/

"In 1 second, each and every second there are …

7,998 Tweets sent in 1 second

839 Instagram photos uploaded in 1 second

1,364 Tumblr posts in 1 second

3,083 Skype calls in 1 second

55,560GB of Internet traffic in 1 second

66,335 Google searches in 1 second

73,391 YouTube videos viewed in 1 second

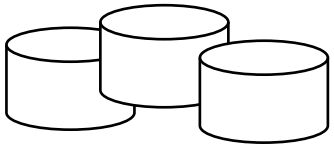
2,681,874 Emails sent in 1 second"

- Computational tools and techniques are needed to help humans summarize, understand, and extract knowledge

Data vs Information

- Society produces huge amounts of data
  - Sources: business, science, medicine, economics, geography, environment, sports, …
- Potentially valuable resource
- Raw data is **useless**: need techniques to automatically **extract information** from it
- Data: recorded facts
- Information: patterns underlying the data

Data Mining Examples



Raw Data

Data Mining

Patterns, knowledge

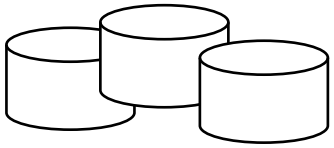
Example 1: Recommender Systems

Data on library books and users' past reading history

Data Mining

What book to recommend next to given user such that there is a high likelihood that the user will like it?

Data Mining Examples



Raw Data

Data Mining

Patterns, knowledge

Example 2: Resource Allocation

Data on library books and users' past reading history

Data Mining

Given a newly acquired book, what is an accurate estimate of the number of users who will read it in the next 12 months?

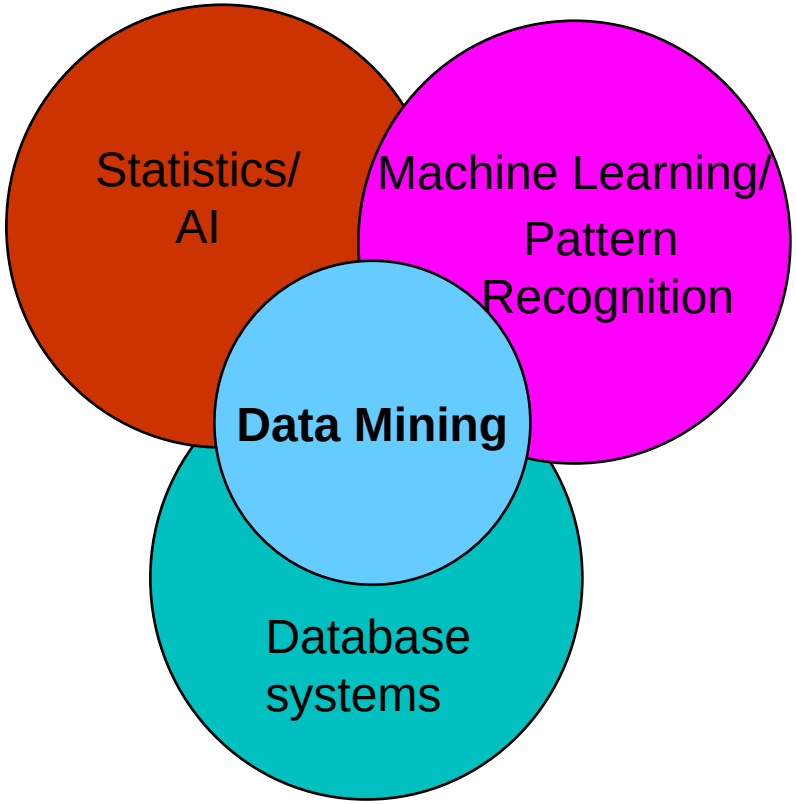
What is Data Mining

"Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data."

- Valid: The patterns hold in general.
- Novel: We did not know the pattern beforehand.
- Useful: We can devise actions from the patterns.
- Understandable: We can interpret and comprehend the patterns.

Alternative names

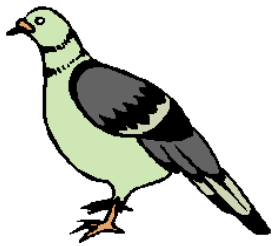
- **Knowledge discovery (mining) in databases (KDD)**
- Knowledge extraction
- Knowledge engineering
- Data/pattern analysis
- Data archeology
- Data dredging
- Information harvesting
- Business intelligence
- etc.



We will return to the actual topic in two minutes. In the meantime, we are going to play a quick game.

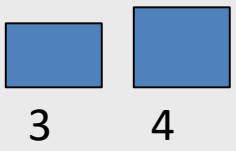
I am going to show you some problems which were shown to pigeons!

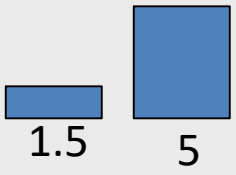
Let's see if you are as smart as a pigeon!

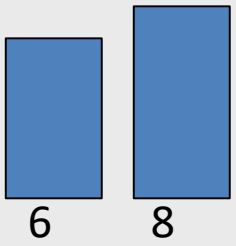


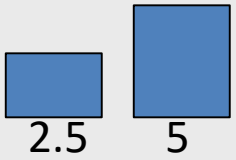
Pigeon Problem 1

Examples of class A

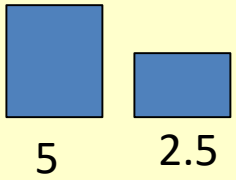


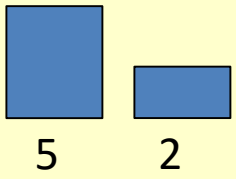


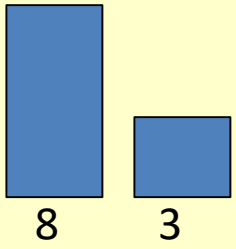


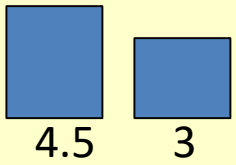


Examples of class B



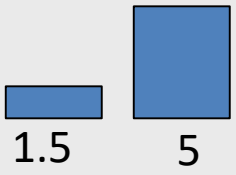


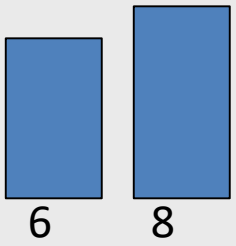


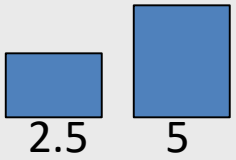


Pigeon Problem 1

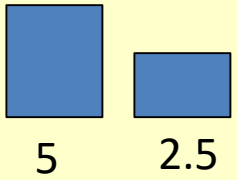
Examples of class A

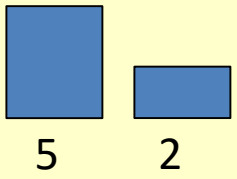


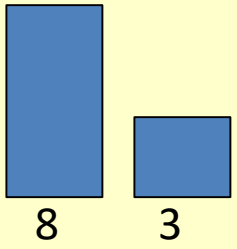


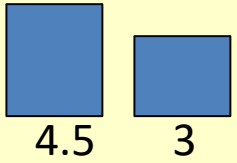


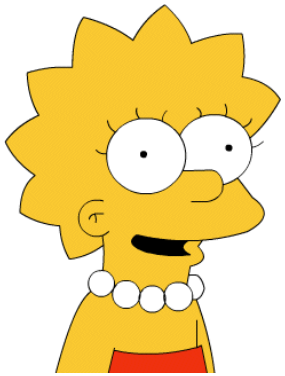
Examples of class B



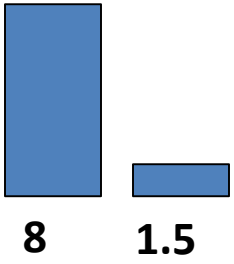




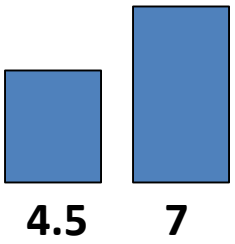




What class is this object?

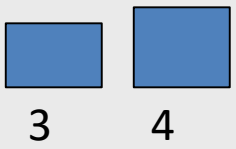


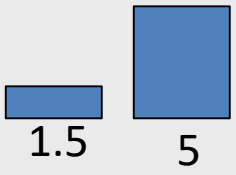
What about this one, A or B?

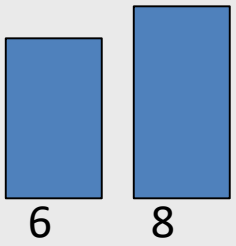


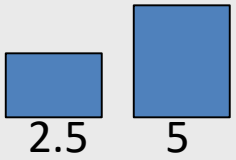
Pigeon Problem 1

Examples of class A

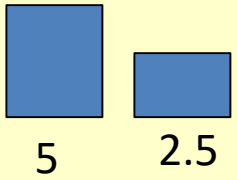


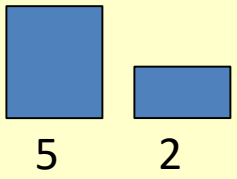


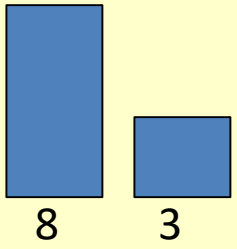


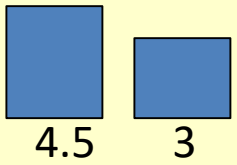


Examples of class B

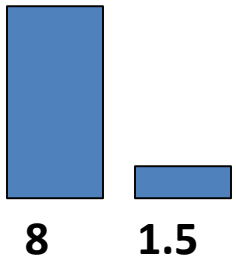








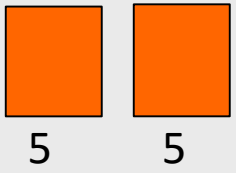
This is a B!

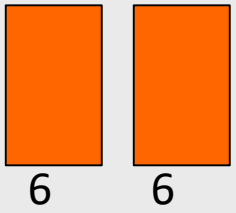


Here is the rule.
If the left bar is smaller than the right bar, it is an A, otherwise it is a B.

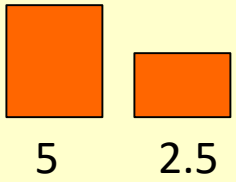
Pigeon Problem 2

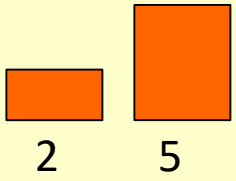
Examples of class A

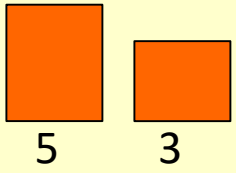


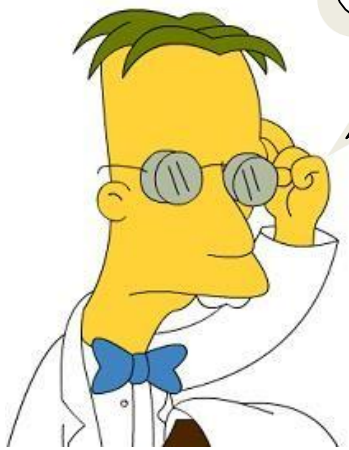


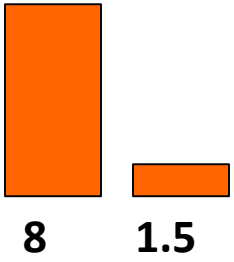
Examples of class B

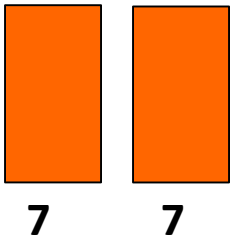






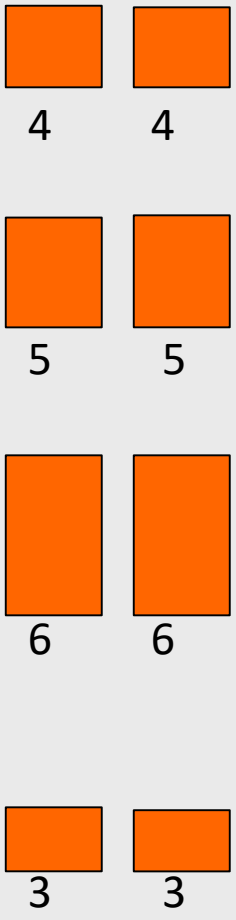




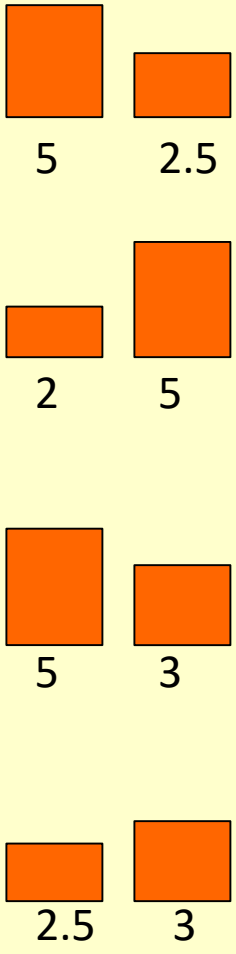


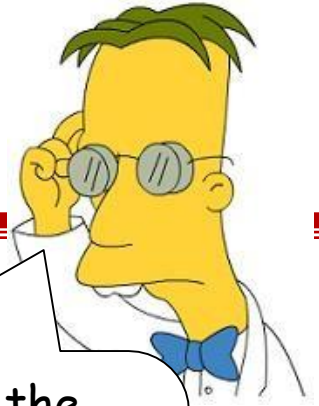
Pigeon Problem 2

Examples of class A



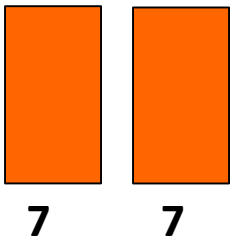
Examples of class B





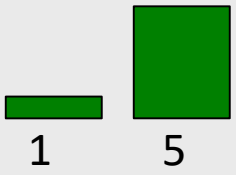
The rule is as follows, if the two bars are equal sizes, it is an A. Otherwise it is a B.

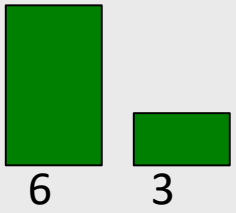
So this one is an A.

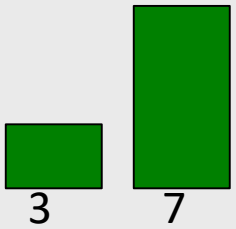


Pigeon Problem 3

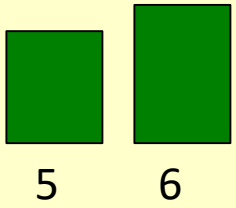
Examples of class A

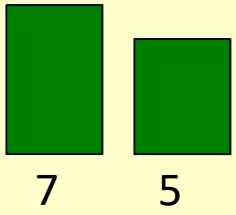


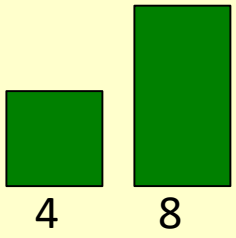


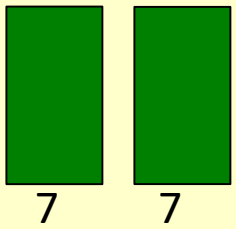


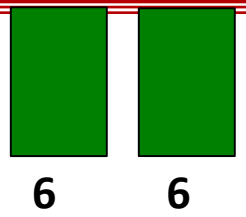
Examples of class B







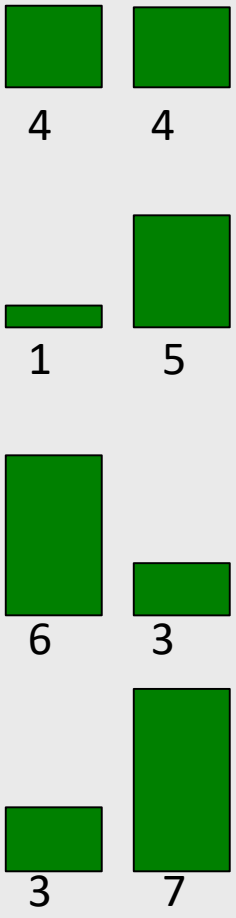




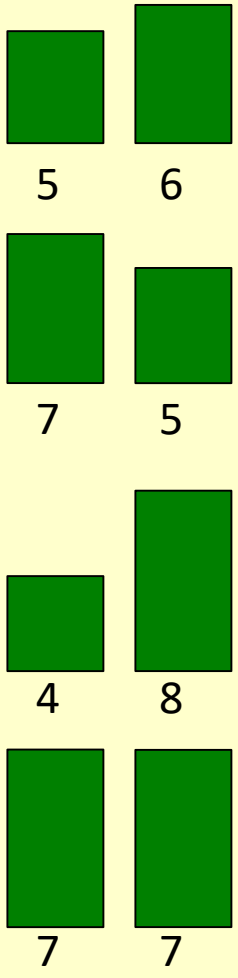
This one is really hard!
What is this, A or B?

Pigeon Problem 3

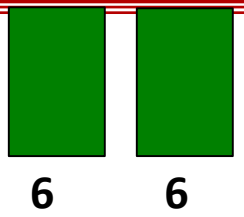
Examples of class A



Examples of class B



It is a B!



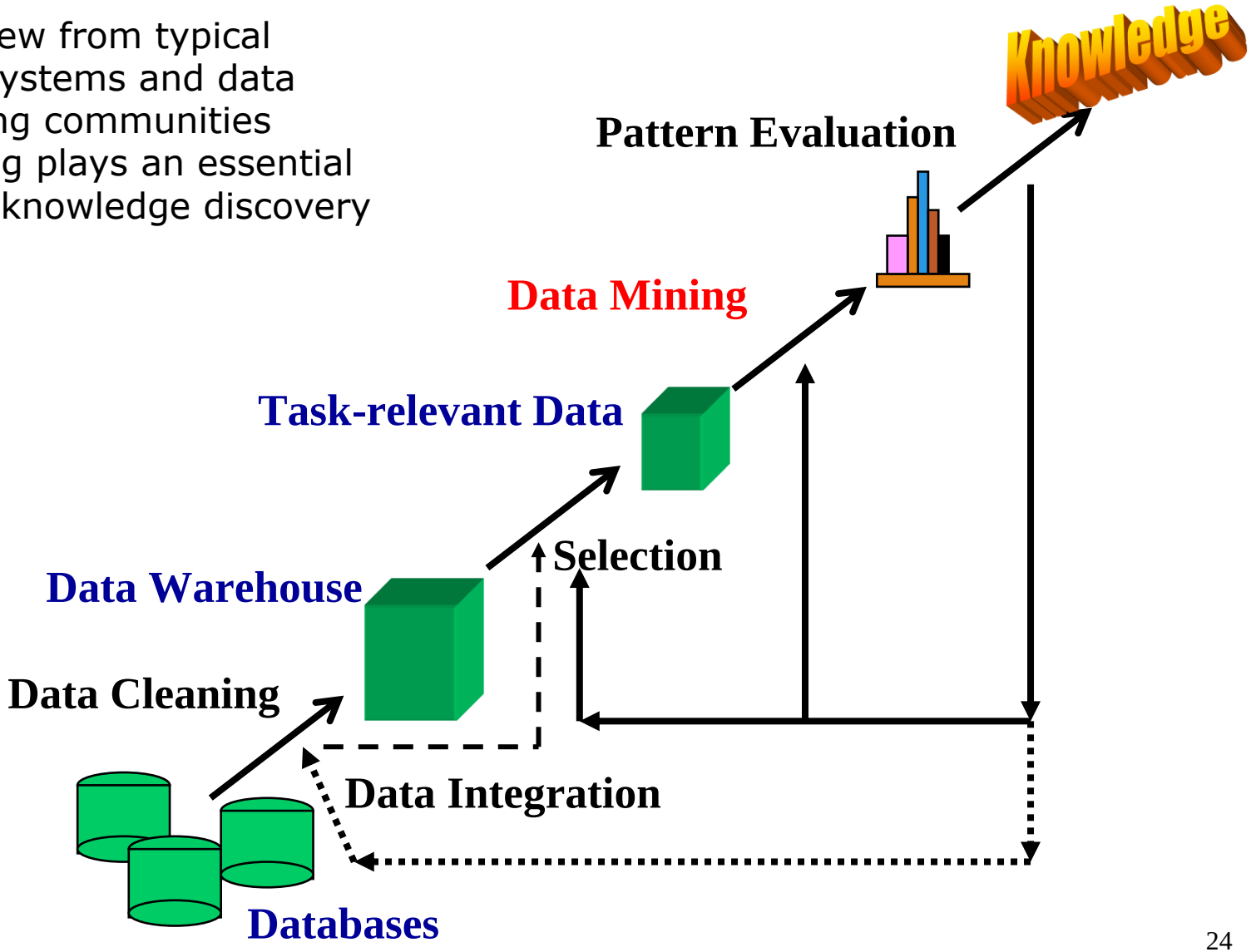
The rule is as follows, if the sum of the two bars is less than or equal to 10, it is an A. Otherwise it is a B.

Data Mining Process

From database & Business Intelligence Perspective

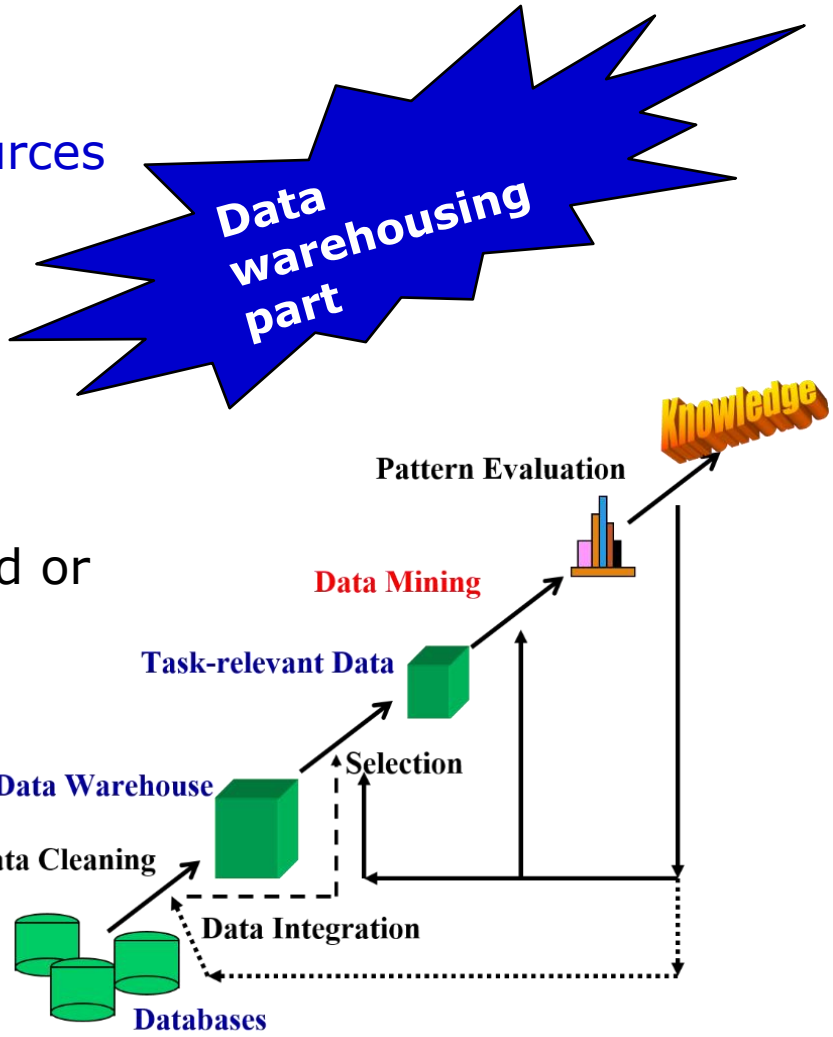
Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process

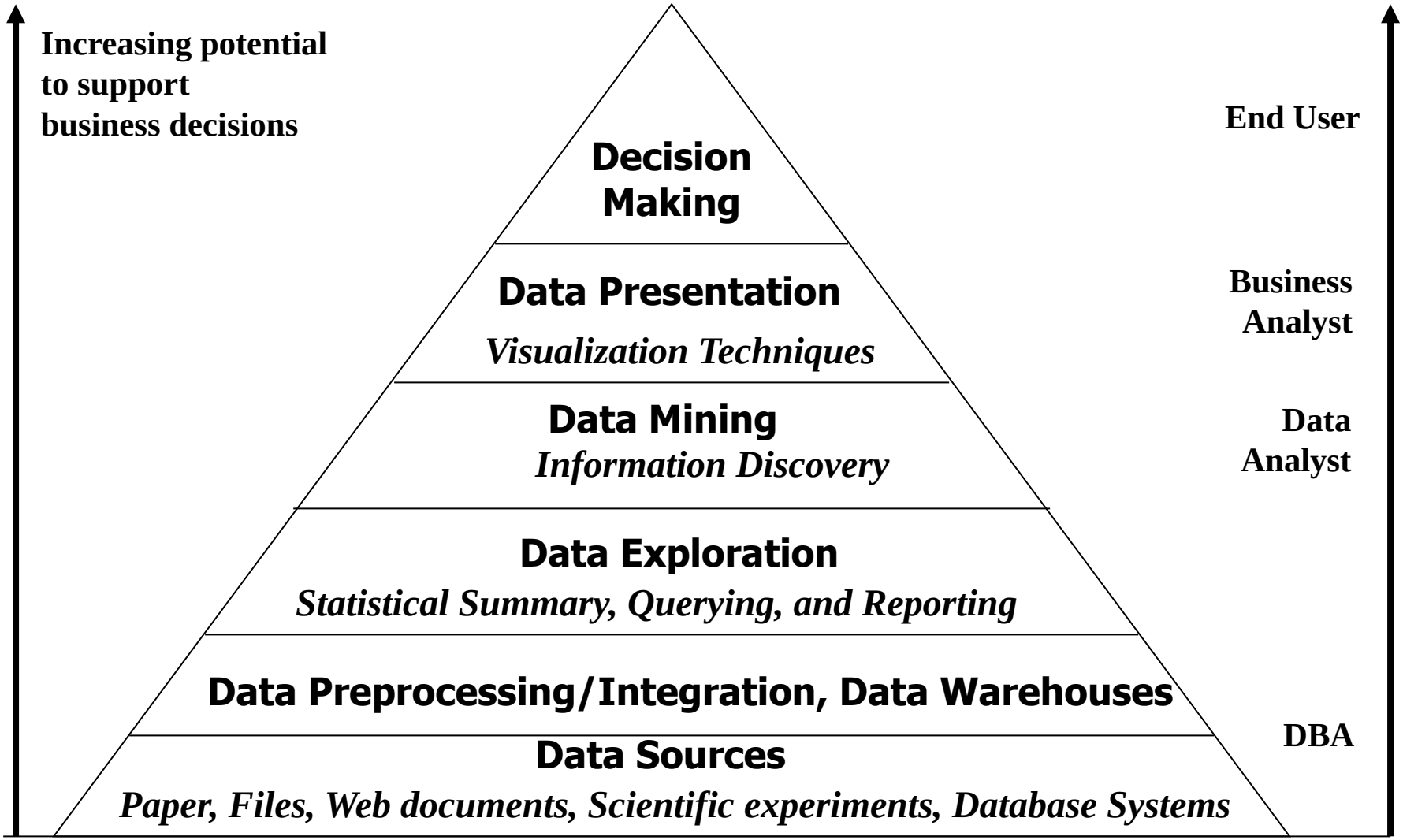


Mining Framework

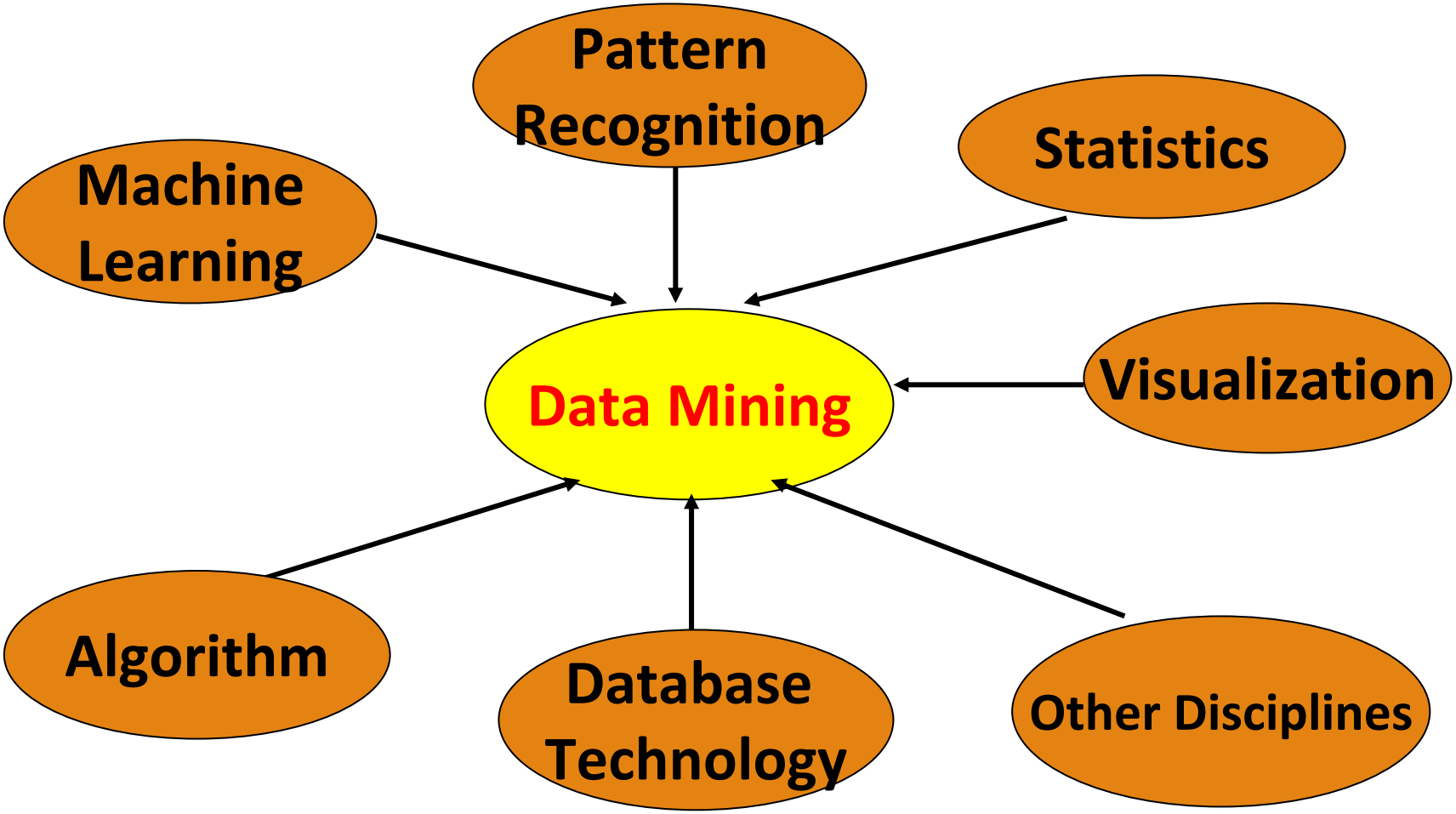
- Mining usually involves
  - Data cleaning
  - Data integration from multiple sources
  - Warehousing the data
  - Data cube construction
  - Data selection for data mining
  - Data mining
  - Presentation of the mining results
  - Patterns and knowledge to be used or stored into knowledge-base



Data Mining in Business Intelligence



Confluence of Multiple Disciplines



Data Mining sources

What kinds of data can be mined?

Data Mining: on what kinds of data

- **Database-oriented data sets and applications**
  - Relational database, data warehouse, transactional database
  - Object-relational databases, Heterogeneous databases and legacy databases
- **Advanced data sets and advanced applications**
  - Data streams (**data stream mining**) and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and information networks
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases (**Text mining**)
  - The World-Wide Web (**Web mining**)

Data Mining Approaches and Techniques

What kinds of patterns can be mined from data?

1. **Concept/Class Description: Characterization and Discrimination**
2. **Mining frequent patterns, associations, and Correlations**
3. **Classification and Regression**
4. **Cluster Analysis**
5. **Outlier Detection**

1. Characterization/Discrimination (Data warehousing part)

- **Data characterization:** a **summarization of the general characteristics** or features of a target class of data.
  - Typically collected by query

- Example: to study the characteristics of software products with sales that increased by 10% in the previous year, the data related to such products can be collected by executing an SQL query on the sales database.

- Examples of techniques:
  - Simple data summaries based on statistical measures and plots
  - data cube-based OLAP roll-up operation (data warehousing part)

1. Characterization/Discrimination (Data warehousing part)

- **Data discrimination:** a **comparison** of the general features of the **target class** data objects against the general features of objects from one or multiple **contrasting classes**

- Example: a user may want to compare the general features of software products with sales that increased by 10% last year against those with sales that decreased by at least 30% during the same period.

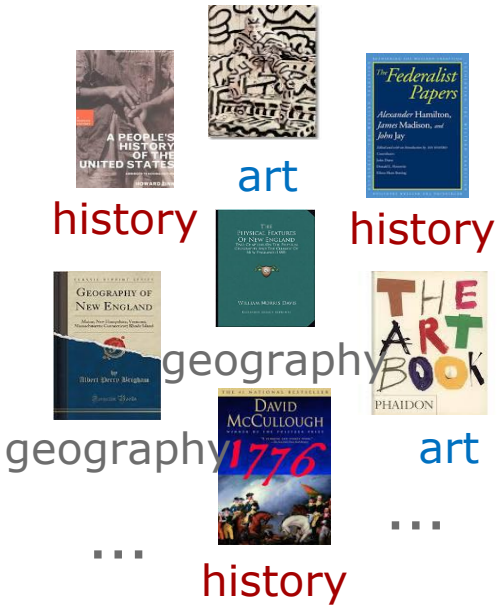
- The methods used for data discrimination are similar to those used for data characterization

2. Pattern Discovery

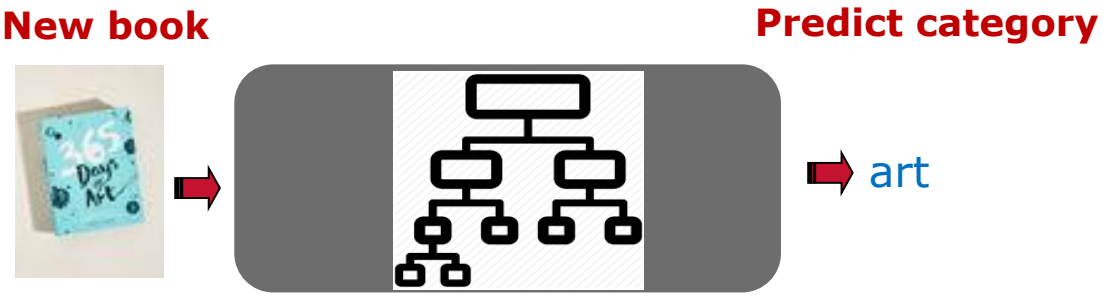
- **Frequent patterns (or frequent itemsets)**
  - What items are frequently purchased together in your store?
- **Association, correlation vs. causality**
  - A typical association rule
    - Diaper → coffee [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

3. Classification

- **Data**: Large collection of books. For each book: **title, info, full text and a category**



Automatically derive from these data a **Classification Model**: A collection of patterns that map books to their categories



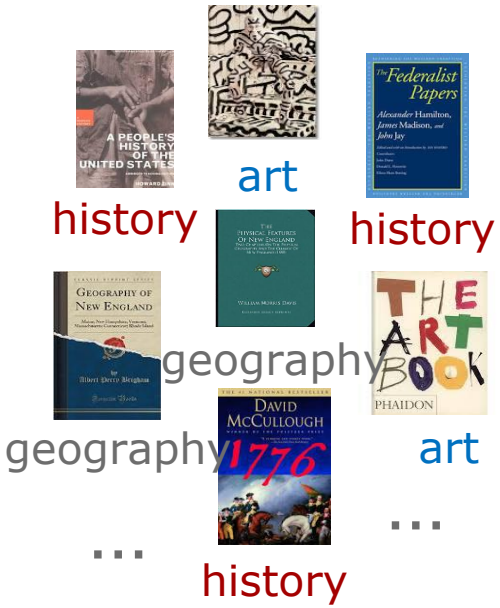
This model can be used for

Prediction: given a new book, predict its category

Description: provide insights into the data

3. Regression

- **Data**: Large collection of books. For each book: **title, info, full text and number of users that accessed the books in the past 12 months**



Automatically derive from these data a **Regression Model**: A collection of patterns that map books to their expected number of readers

This model can be used for

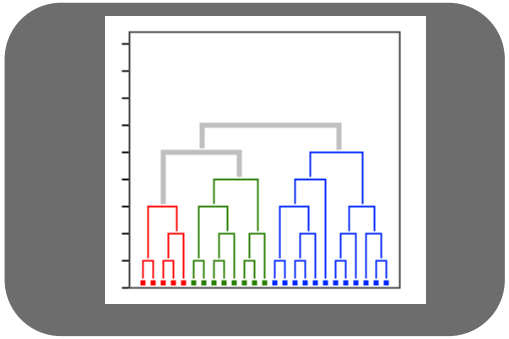
Prediction: given a new book, predict its expected number of readers in the next 12 months

Description: provide insights into the data

4. Clustering

- **Data**: Large collection of books. For each book: **title, info, full text, ..**

Automatically derive from these data a **a set of clusters**: that group books by similarity



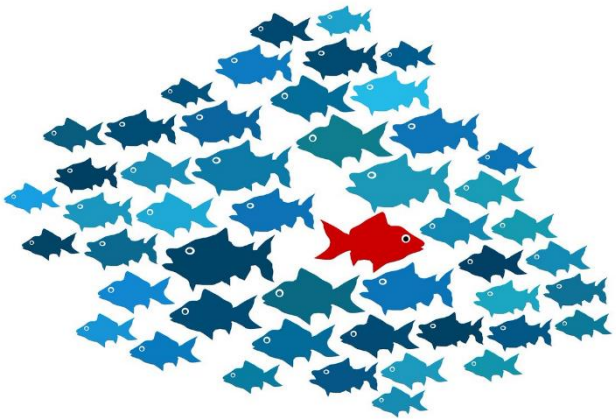
This model can be used for

Description: provide insights into the data

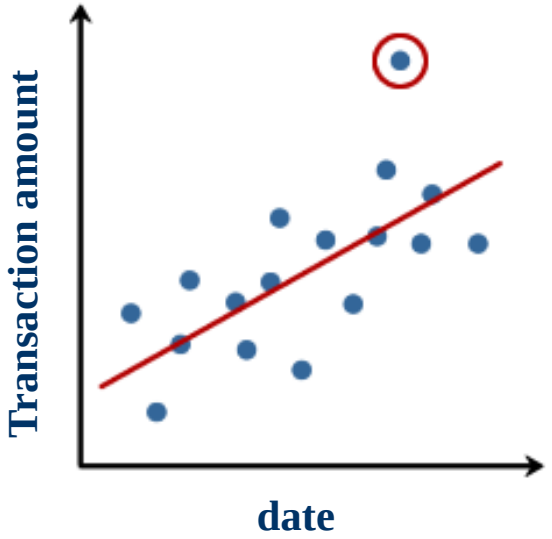
Useful for example to recommend books to users or to organize books in (virtual) library shelves

5. Outlier Analysis

- **Data**: customers' transactions. For each customer: **transaction amount, date, …**



Automatically identify observations (data points) which **deviate so much** from the other observations



Data Mining Tasks

- **Descriptive Data Mining:** Patterns for presenting the behavior of observed entities in a human understandable format
  - Concept/Class description, pattern discovery, cluster analysis
- **Predictive Data Mining:** Patterns for predicting the behavior of newly encountered entities
  - Classification and regression, outlier detection
- in some cases you follow a hybrid approach: Descriptive then predictive
- Some patterns can be descriptive and predictive

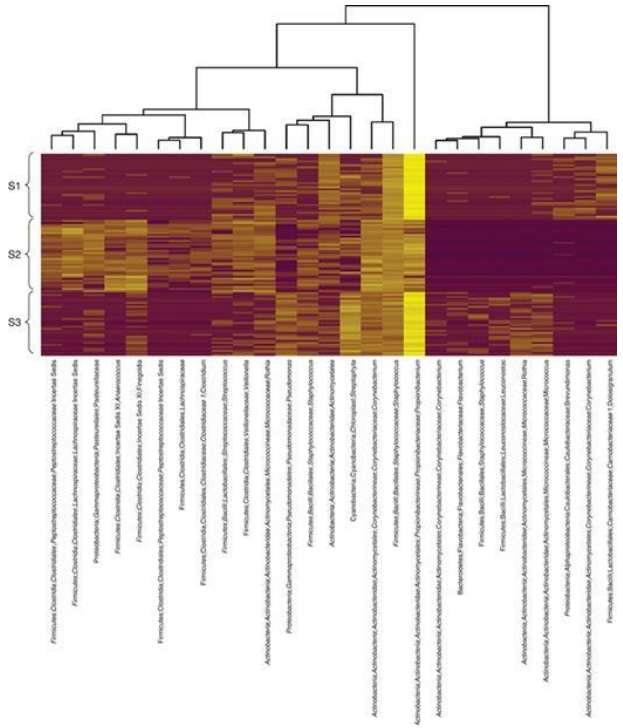
Evaluation of discovered knowledge

- **Are all mined knowledge interesting?**
  - One can mine tremendous amount of "patterns" and knowledge
  - Some may fit only certain dimension space (time, location,…)
  - Some may not be representative, may be transient, …
- **Can a data mining systems directly mine only interesting knowledge?**
  - An optimization problem that remains a challenging issue in data mining

Data Mining Applications

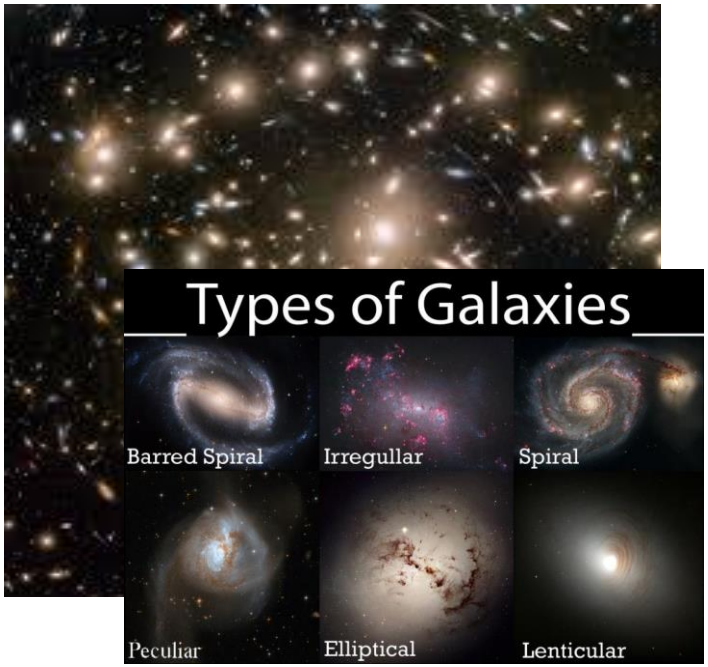
Data Mining Applications

Identifying important groups of microorganisms in the human body



Dan Knights Elizabeth K. Costello Rob Knight "Supervised classification of human microbiota" FEMS Microbiology Reviews, Volume 35, Issue 2, 1 March 2011, Pages 343–359

Classifying galaxies in the universe



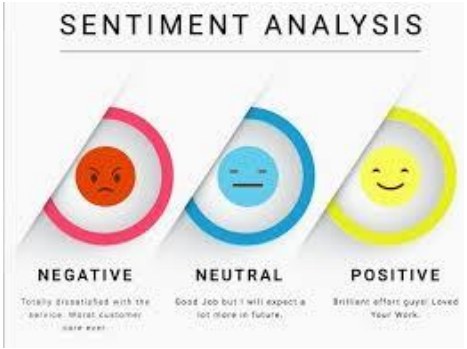
Fowler, L., Schawinski, K., & Brandt, B.-E. Galaxy Classification using Machine Learning. Paper presented at the American Astronomical Society Meeting Abstracts. 2017

Data Mining Applications

Email spam filtering

Blanzieri, E. & A. Bryl. "A survey of learning-based techniques of email spam filtering" Artificial Intelligence Review March 2008, Vol. 29, Issue 1, pp 63–92

Document sentiment analysis



Liu B., Zhang L. "A Survey of Opinion Mining and Sentiment Analysis." In: Aggarwal C., Zhai C. (eds) Mining Text Data. Springer, Boston, MA. 2012

Data Mining Applications

image and video processing

https://www.classaction.org/blog/facebook-sued-over-face-recognition-feature

audio and voice processing

Personal assistants

Bgr.com/tag/siri

recommender systems

Data Mining Applications

black and white image colorization

Zhang, Isola, Efros.
Colorful Image Colorization.
In ECCV, 2016.

http://richzhang.github.io/colorization/

See also https://machinelearningmastery.com/inspirational-applications-deep-learning/

Data Mining Applications

image classification, object recognition, description generation using deep neural networks

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"girl in pink dress is jumping in air."

"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

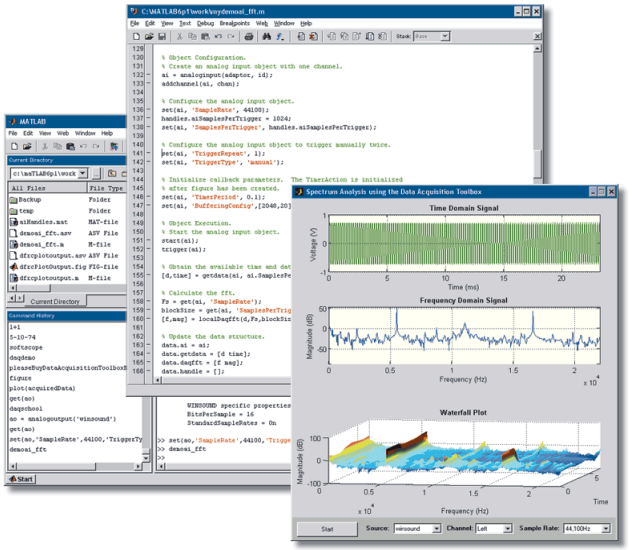
Andrej Karpathy & Li Fei-Fei
"Deep Visual-Semantic Alignments for Generating Image Descriptions"
CVPR 2015

https://cs.stanford.edu/people/karpathy/deepimagesent/

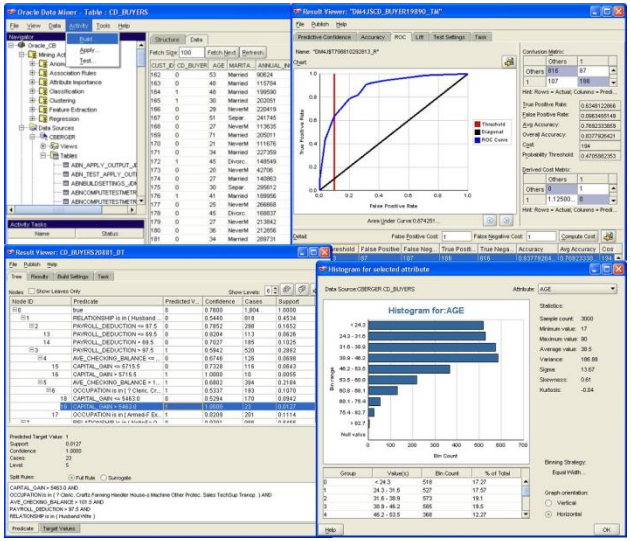
Data Mining Packages and Platforms

Commercial Data Mining Systems

Matlab



Oracle data mining



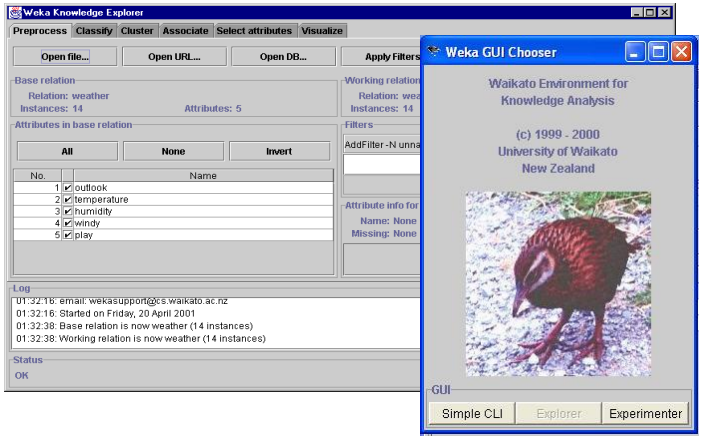
sas®
THE POWER TO KNOW®

and lots more ….

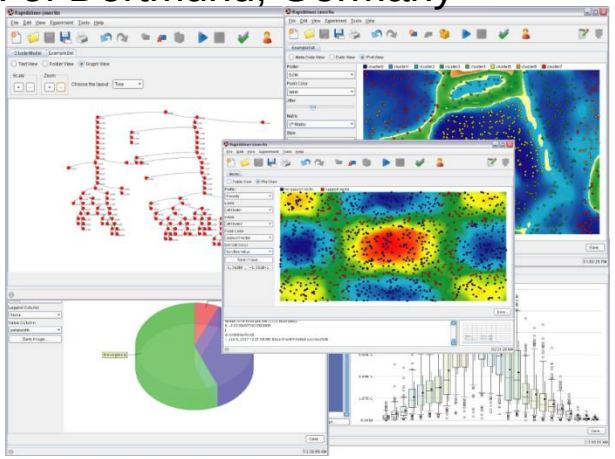
Open Source Data Mining Tools

WEKA

Frank et al., University of Waikato, New Zealand



RapidMiner Klinkenberg et al., Univ. of Dortmund, Germany



R Programming Language

Ross Ihaka and Robert Gentleman

Univ. of Auckland, New Zealand

Python

Data Mining Libraries

and many more ….

Data Mining Resources

Data Mining Books

- "Data Mining: Concepts and Techniques (3rd Edition)". J. Han and M. Kamber. Morgan Kaufmann Publishers. 2012. **(Textbook)**
- Introduction to Data Mining (2nd edition) P.-N. Tan, M. Steinbach, A. Karpatne, V. Kumar. Pearson, 2018.
- "Data Mining: Practical Machine Learning Tools and Techniques (4th Edition)" I.H. Witten, E. Frank, M. Hall, C. Pal. Morgan Kaufmann Publishers. 2017.
- "Advances in Knowledge Discovery and Data Mining". Eds.: Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy. The MIT Press, 1995.
- …

Data Mining Journals

- Data Mining and Knowledge Discovery Journal
- ACM SIGKDD Explorations Newsletter
- TKDE: IEEE Transactions in Knowledge and Data Engineering
- TODS: ACM Transactions on Database Systems
- JACM: Journal of ACM
- Data and Knowledge Engineering
- JIIS: Intl. Journal of Intelligent Information Systems
- …

Data Mining Conferences

- KDD: ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining
- ICDM: IEEE International Conference on Data Mining,
- SIAM International Conference on Data Mining
- PKDD: European Conference on Principles and Practice of Knowledge Discovery in Databases
- PAKDD Pacific-Asia Conference on Knowledge Discovery and Data Mining
- DaWak: Intl. Conference on Data Warehousing and Knowledge Discovery

Other related Conferences:

- ICML: Intl. Conf. On Machine Learning
- IDEAL: Intl. Conf. On Intelligent Data Engineering and Automated Learning
- IJCAI: International Joint Conference on Artificial Intelligence
- AAAI: American Association for Artificial Intelligence Conference
- SIGMOD/PODS: ACM Intl. Conference on Data Management
- ICDE: International Conference on Data Engineering
- VLDB: International Conference on Very Large Data Bases

Data Mining Datasets

- Univ. of California Irvine Machine Learning Data Repository.
- Univ. of California Irvine KDD Data Repository.
- Datasets for Data Mining
- Datamob - Public data put to good use.
- Time Series Data Library
- CMU's StatLib-Datasets Archive
- Stanford Large Network Dataset Collection (SNAP)
- 100+ Interesting Data Sets for Statistics
- ...

Summary

- Data mining is the "non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"
- The KDD process includes data collection and pre-processing, data mining, and evaluation and validation of those patterns
- Data mining is the discovery and extraction of patterns from data, not the extraction of data
- Important challenges in data mining: privacy, security, scalability, real-time, and handling non-conventional data