



# Introduction to Data Mining

*Day 1 – 26 February, 2023*

*13:00 – 13:45*

# What is data mining?

Data mining (or *knowledge discovery*) is the exploration and analysis of large quantities of data in order to discover hidden patterns in data.

❑ **Valid:** The patterns hold in general.

❑ **Novel:** We did not know the pattern beforehand.

❑ **Potentially useful:** We can devise **actions** from the patterns.

❑ **Understandable:** We can interpret and comprehend the patterns.

**def·i·ni·tion**

/ defə|niSH(ə)n /

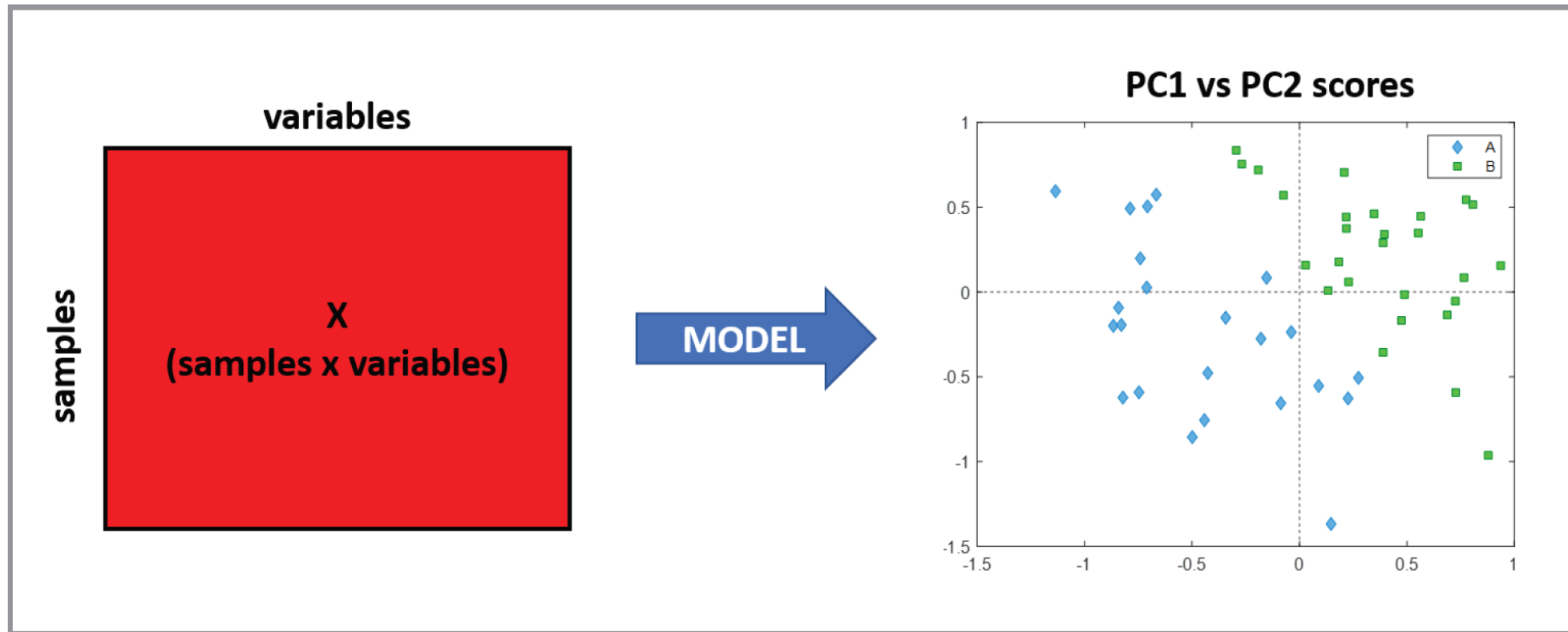
*noun:* a statement of the exact meaning of a word.

# What is data mining? (cont'd)

- ❑ A combination of **statistics, probability analysis and database**
- ❑ Is performed on **huge amount** of data
- ❑ Reduces the **time of analysis**
- ❑ Is a **standardized** operation process
- ❑ Is a young process (less than 20 years) and not fully mature yet
  - Next step will be using IA



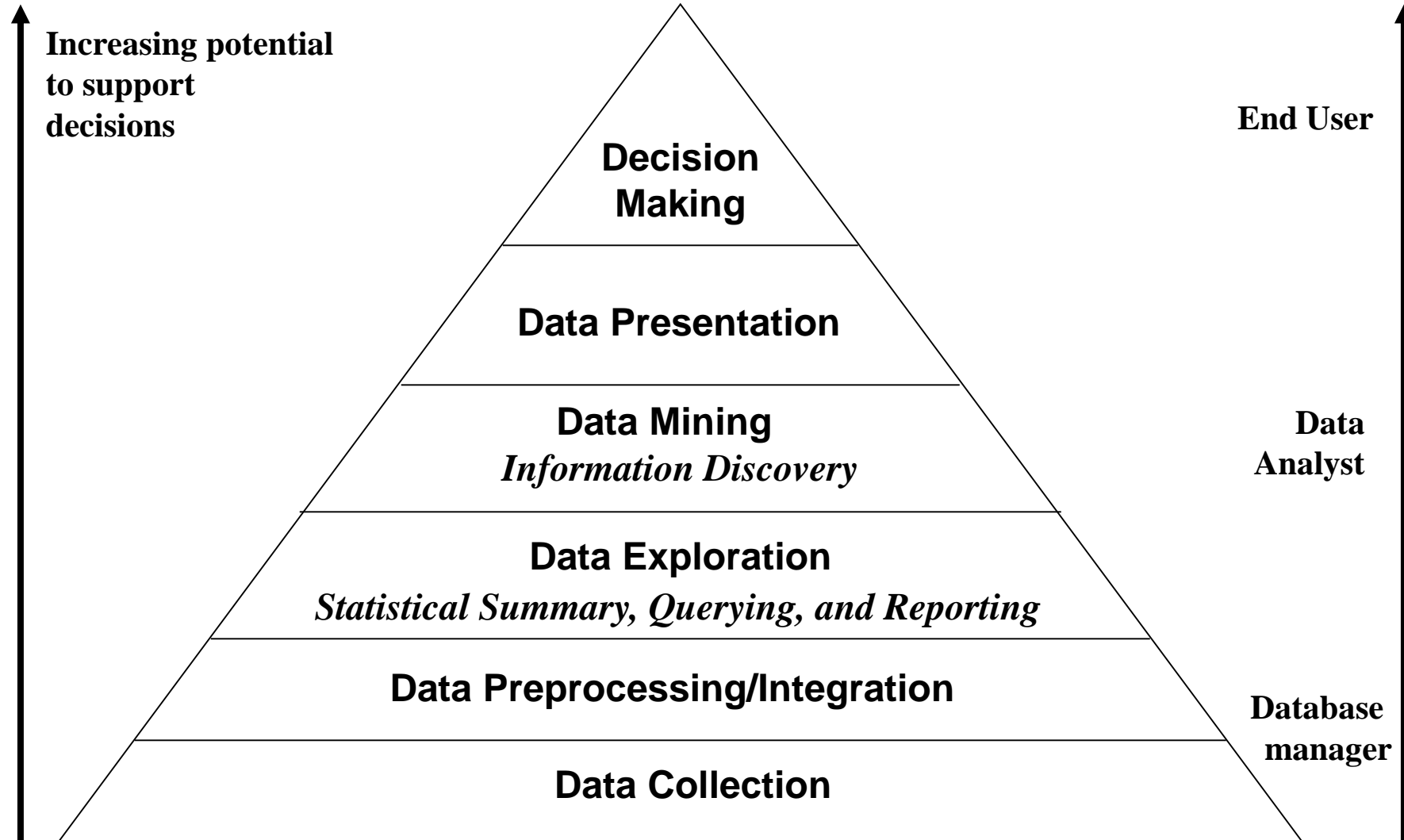
# What is data mining? (cont'd)



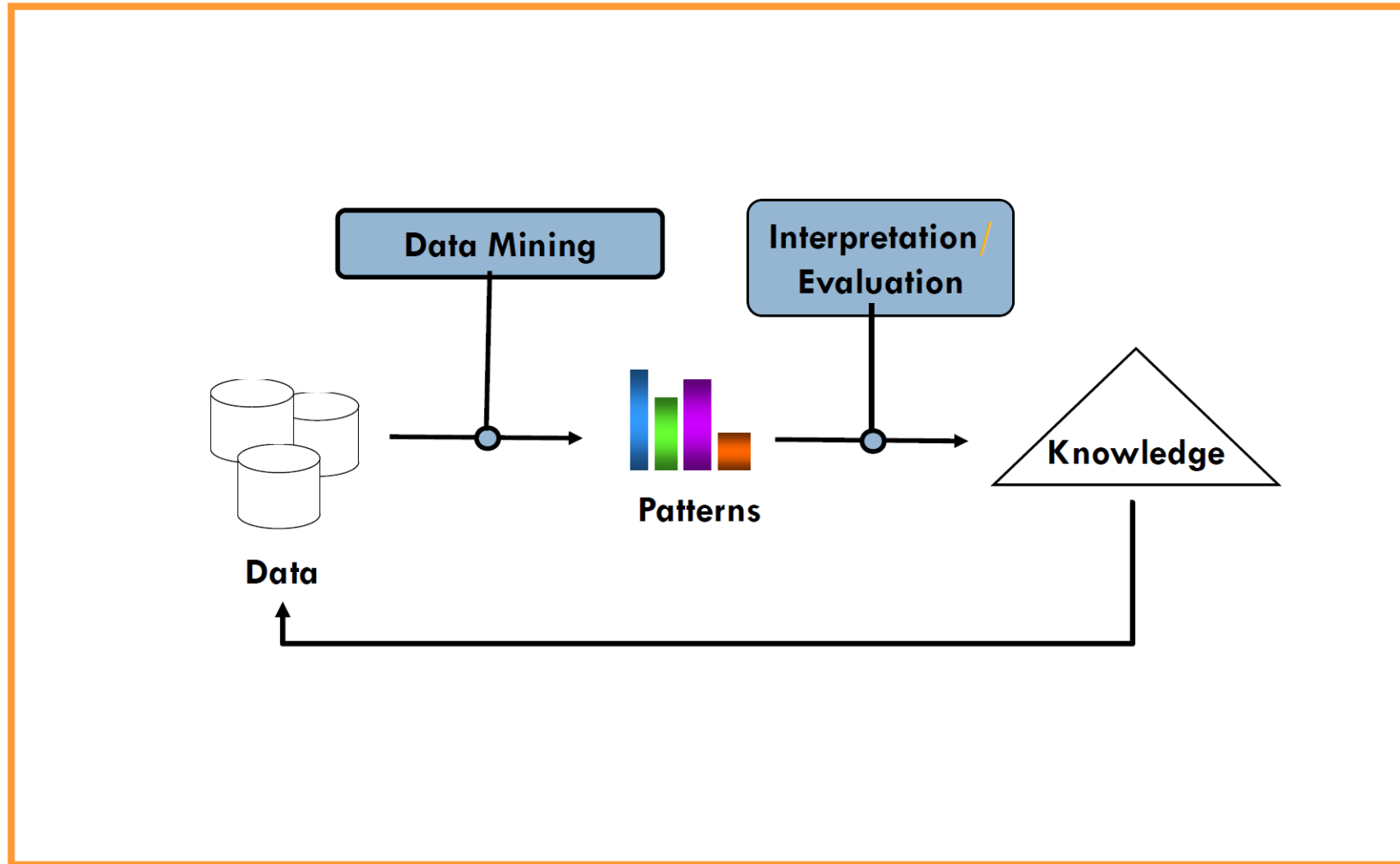
**Figure 3.** Data Mining applied to a matrix  $X$ . A PCA model has been applied, and the score scatter plot between PC1 and PC2 is shown.

*Amigo, 2021*

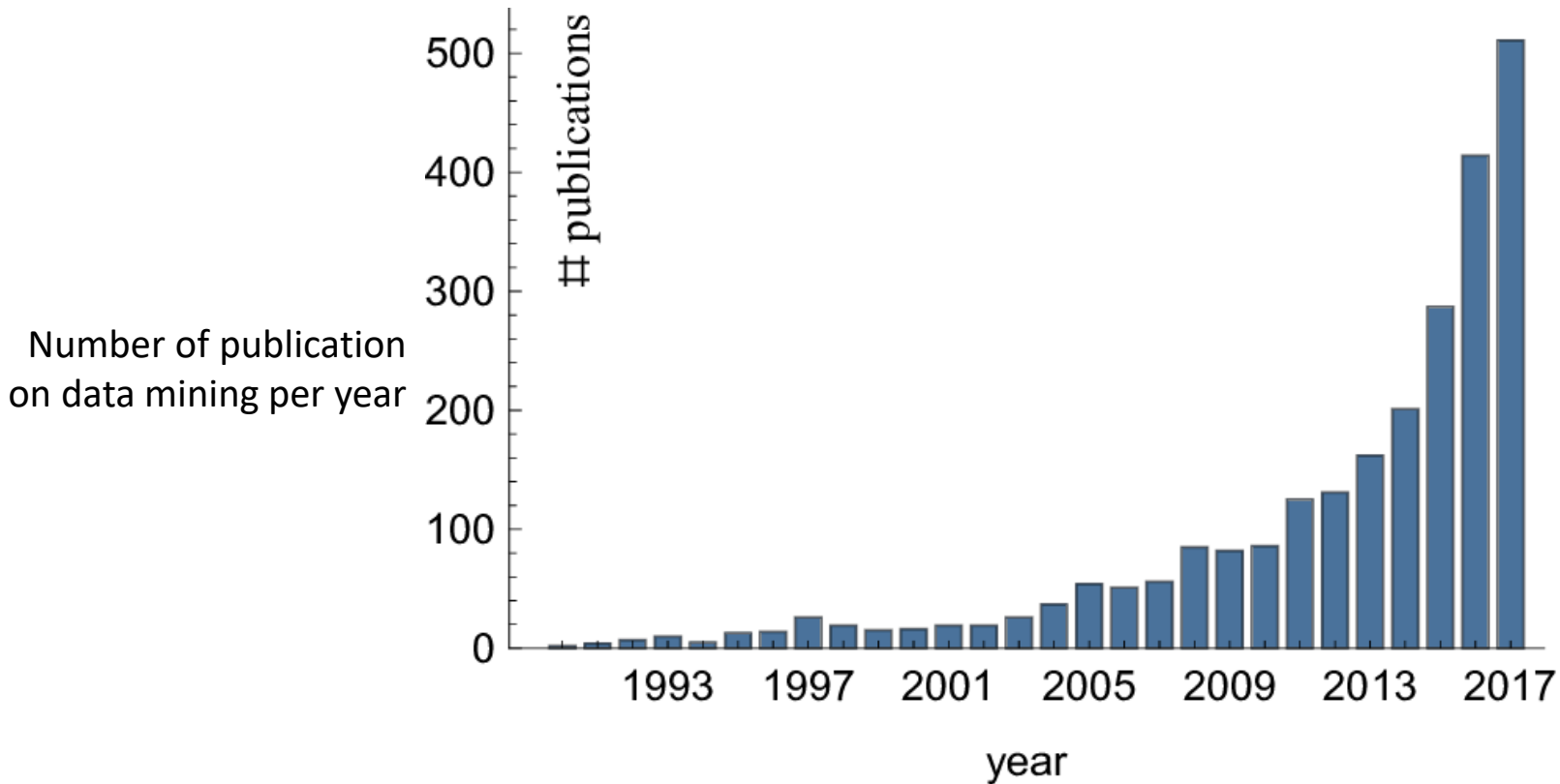
# Data Mining Process



# Knowledge Discovery: Process

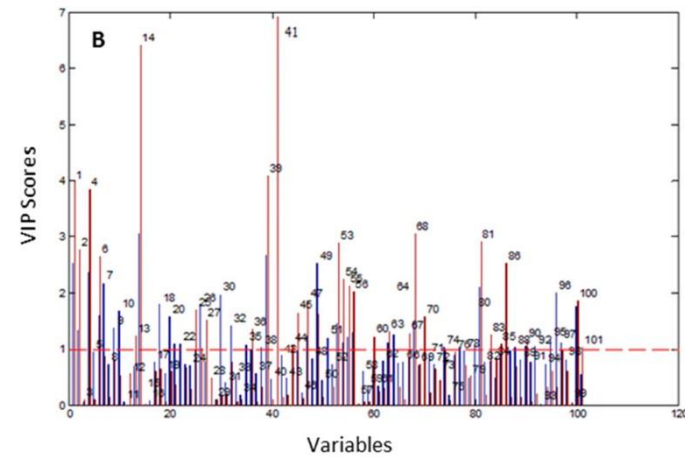
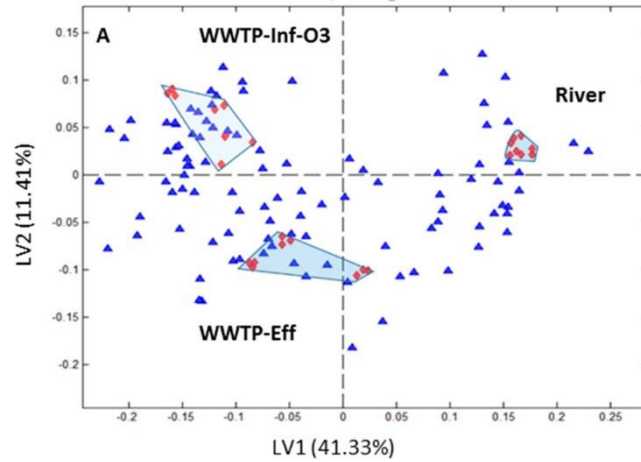
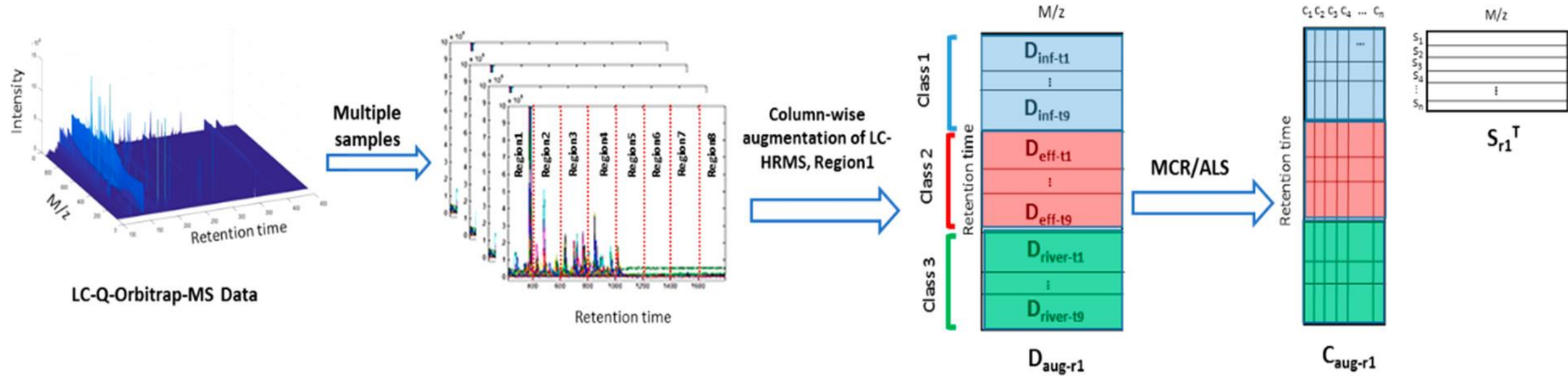


# Why data mining?



- Data are produced too quickly
- Or there are already too much data
- Most of the published data are never analyzed at all
- Waste of information

# Example Application: For LC-MSMS Data





# Example Application: Food Fraud Detection

## Objectives:

- Discrimination of fraudulent honey based on floral and geographic origins

## Approach:

- Comparative untargeted metabolomics using an analytical method (UHPLC-Q-Orbitrap) and statistical tools (principal component analysis and a three-stage approach (t-test, volcano plot and variable importance in projection plot))

## Outcome:

- Identification of key components per geographic origins that will help for the identification of outliers

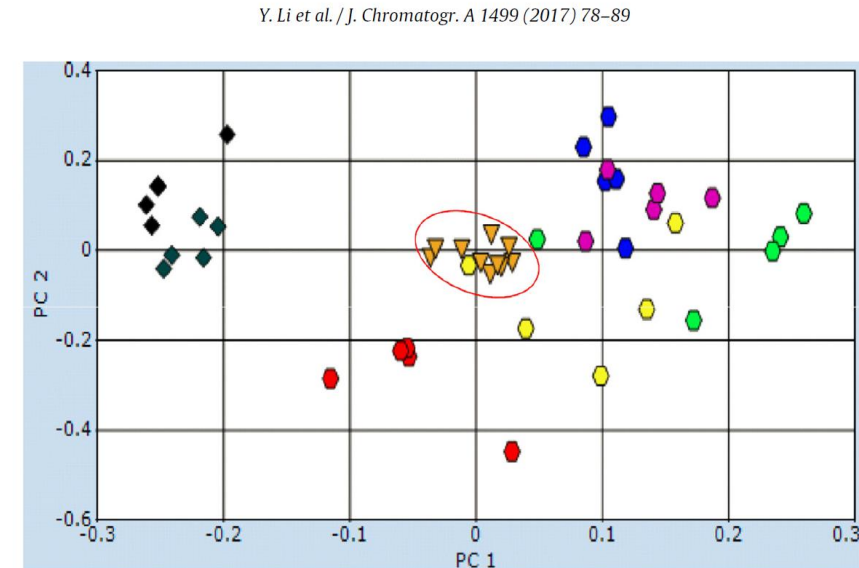


Fig. 5. PCA clustering plot of honey samples from five geographic areas and two nectariferous plants, as well as a QC sample (color codes: red, Litchi honey from Yunnan Province; yellow, Litchi honey from Hainan Province; green, Litchi honey from Fujian Province; purple, Litchi honey from Guangdong Province; blue, Litchi honey from Guangxi Province; cyan, acacia honey from Liaoning Province; black, acacia honey from Shaanxi Province; dark yellow inverted triangles, QC sample). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

# Example Application: Metagenomics

## Objectives:

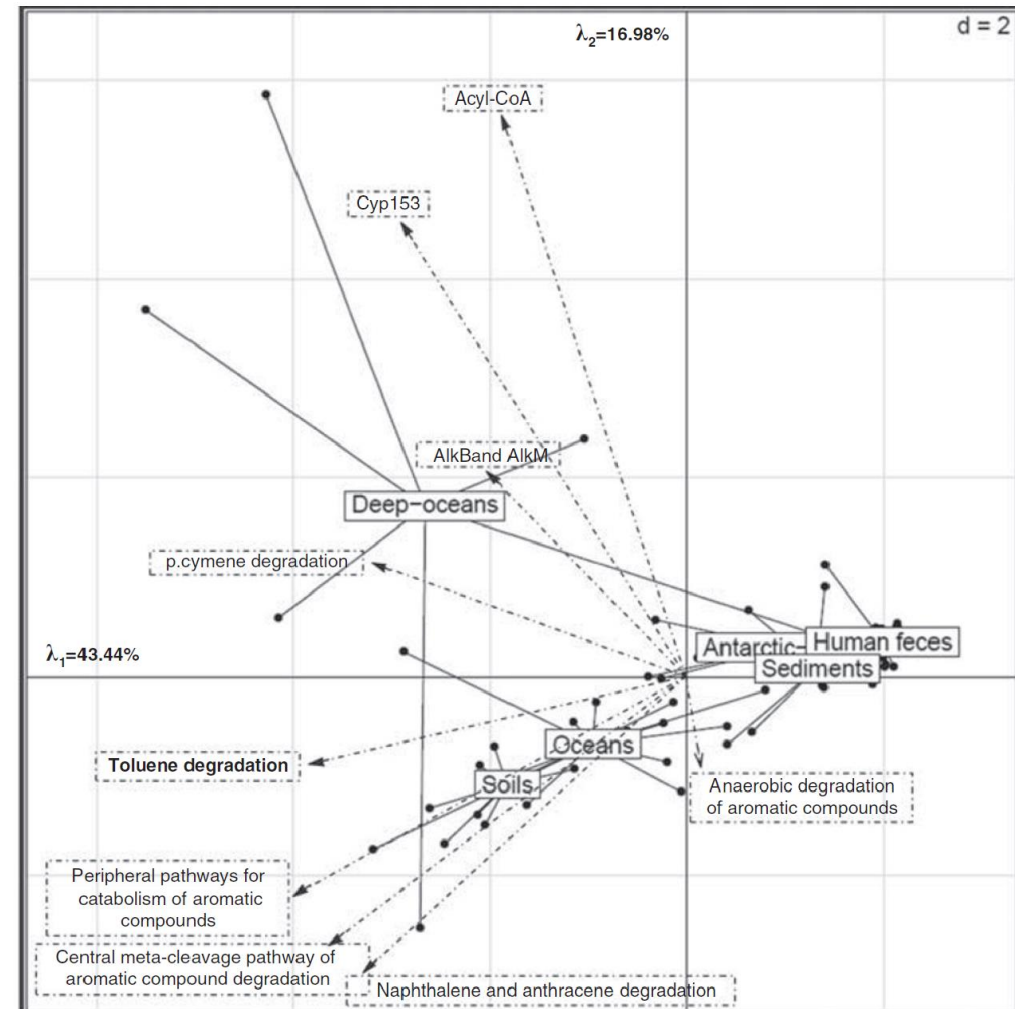
- Detection of the most interesting ecosystems for oil degradation

## Approach:

- Comparison of metagenomes from different ecosystems using functional subsystems associated with hydrocarbon degradation using a PCA approach

## Outcome:

- This strategy will help guide standardized, comparative untargeted metabolomics studies of honey and other agro-products from different floral and geographic origins.



**Figure 3** PCA of six selected ecosystems based on their number of sequences associated with petroleum hydrocarbon degradation functions ( $E$ -value  $< 10^{-5}$ ). The functional classes as provided by MG-RAST and the local blasts are plotted on the same PCA as the samples in order to observe relationships between function and environment.

# Applications in the Agri-food sector

Biodivers Conserv (2015) 24:3009–3031  
DOI 10.1007/s10531-015-0994-5



ORIGINAL PAPER

**Linking traditional tree-crop landscapes and agro-biodiversity in central Italy using a database of typical and traditional products: a multiple risk assessment through a data mining analysis**

Rita Biasi<sup>1</sup> · Elena Brunori<sup>1</sup> · Daniela Smiraglia<sup>2</sup> · Luca Salvati<sup>3</sup>

Received: 21 March 2015 / Revised: 11 August 2015 / Accepted: 17 August 2015 /  
Published online: 29 August 2015  
© Springer Science+Business Media Dordrecht 2015

**The Application of Big data Mining in Risk Warning for Food Safety**

Yajie WANG, Bing YANG, Yan LUO, Jinlin HE, Hong TAN<sup>\*</sup>  
Guizhou Academy of Testing and Analysis, Guiyang 550002, China

*Research Article*

**Deep-Stacking Network Approach by Multisource Data Mining for Hazardous Risk Identification in IoT-Based Intelligent Food Management Systems**

Jianlei Kong<sup>1,2</sup>, Chengcai Yang<sup>1</sup>, Jianli Wang<sup>1</sup>, Xiaoyi Wang<sup>1</sup>, Min Zuo<sup>1,2</sup>,  
Xuebo Jin<sup>1</sup> and Sen Lin<sup>3</sup>

The Journal of Antibiotics (2021) 74:838–849  
<https://doi.org/10.1038/s41429-021-00471-w>

JARA  
Japan Antibiotics  
Research Association

The Society for  
Actinomycetes Japan

REVIEW ARTICLE



**A review: antimicrobial resistance data mining models and prediction methods study for pathogenic bacteria**

Xinxing Li<sup>1</sup> · Ziyi Zhang<sup>1</sup> · Buwen Liang<sup>1</sup> · Fei Ye<sup>2</sup> · Weiwei Gong<sup>2</sup>

Received: 11 May 2021 / Revised: 27 May 2021 / Accepted: 16 July 2021 / Published online: 14 September 2021  
© The Author(s), under exclusive licence to the Japan Antibiotics Research Association 2021

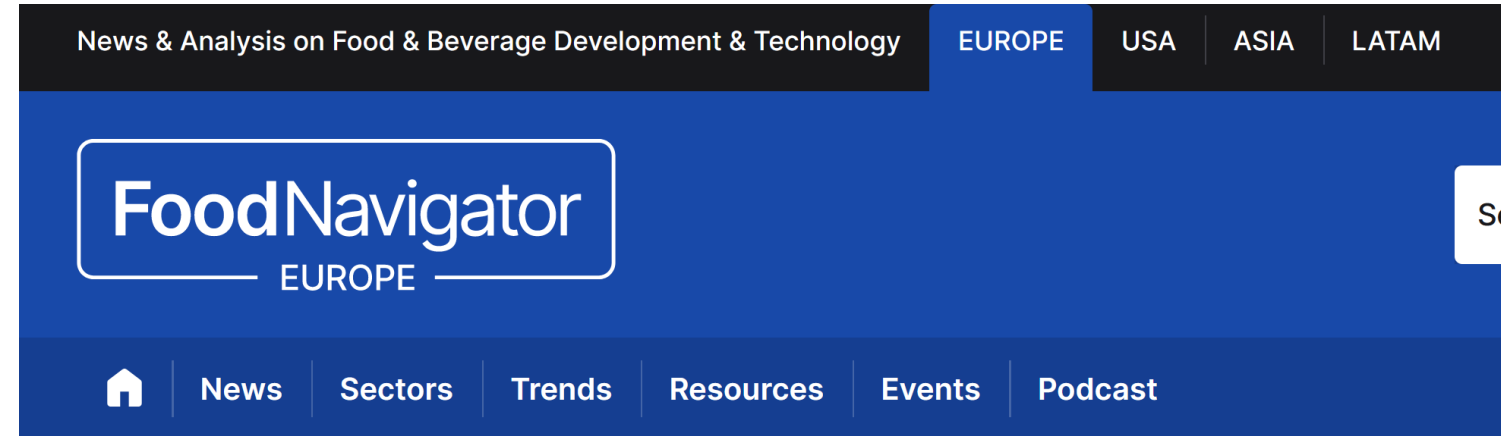
# Potential Other Applications in food

## □ Data analysis and decision support

- Market analysis and management
- Consumption patterns
- Risk analysis and management
  - Forecasting, quality control, customer retention, competitive analysis
- Fraud detection and detection of unusual patterns (outliers)

## □ Other Applications

- Bioinformatics and bio-data analysis



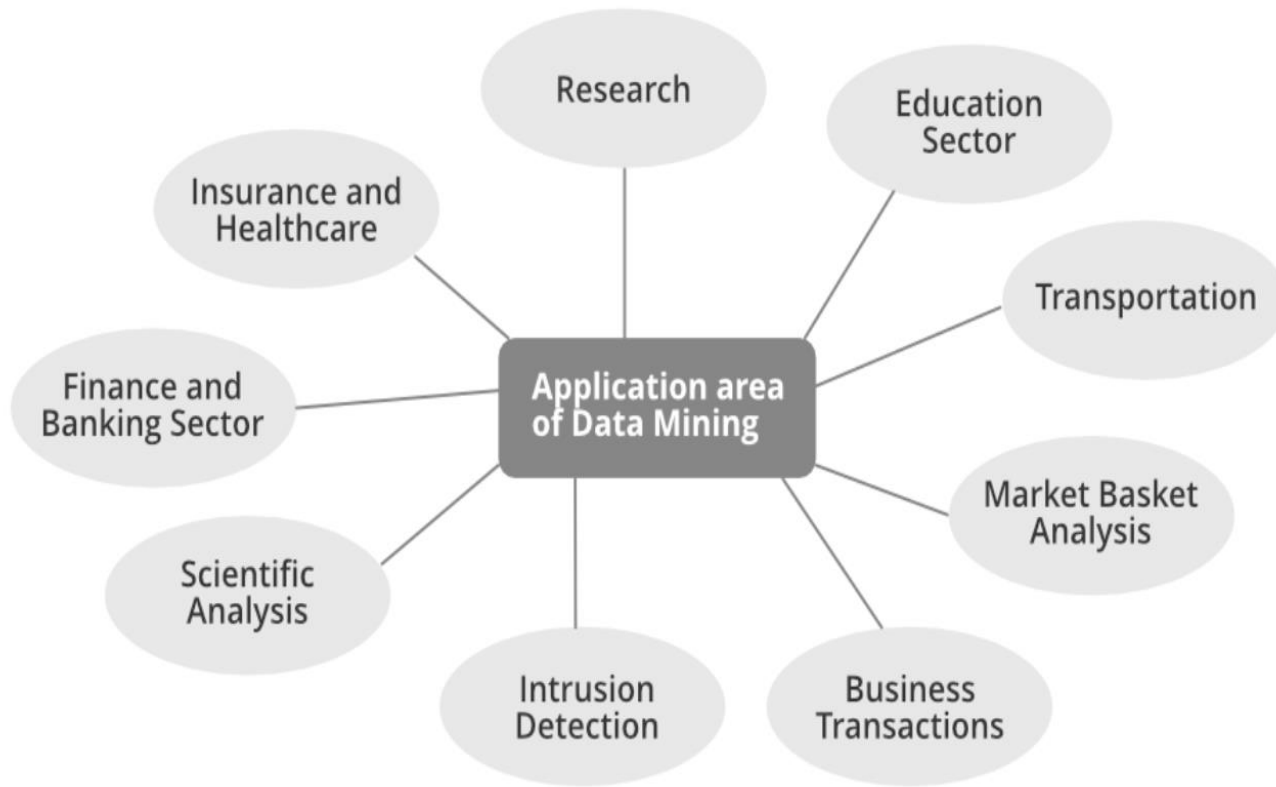
## Big data: 'Exhaustive review' pulls together evidence on food groups and diet-related disease

By Nathan Gray

11-Dec-2014 - Last updated on 11-Dec-2014 at 11:40 GMT

SHARE

# Applications



- Customer behaviors
  - For purchase of goods
  - For the consumption of food commodities
  - Learning behavior
- Prediction of natural disasters
- Energy consumption
- Biological data applications
- Counter-terrorism
- Intrusion detection
- Etc...

# Goals of Data Mining

- ❑ **Prediction:** How certain variables within the data will behave in the future.
- ❑ **Identification:** Identify the existence of an item, an event, an activity or a category.
- ❑ **Classification:** Partition of data into categories.
- ❑ **Optimization:** Optimize the use of limited resources.





# On what kind of data? - Data collection

- ❑ Reliable data
- ❑ In our case (Arab Food Monitoring Database)
  - Scientific papers
  - Governmental reports
  - International agency reports
  - Others...
- ❑ Spatiotemporal Datapoints (when and where?)
- ❑ Static or dynamic data (updates needed)



# On what kind of data? - Data reprocessing

- ❑ **Choice of variables** — identification, extraction, accessibility.
- ❑ **Standardised Process** — because data mining is standardised, the outputs produced are systematic and statistically “objective,” given the limitations of the data.
- ❑ **Data cleaning** — because of the large amount of data, we need to make sure that we only have the necessary data and remove the unwanted. Otherwise, they may lead us to false conclusions.



# What kinds of patterns can be mined?

- ❑ **Descriptive** mining tasks characterize the general properties of the data in the database
- ❑ **Predictive** mining tasks perform inference on the current data in order to make predictions

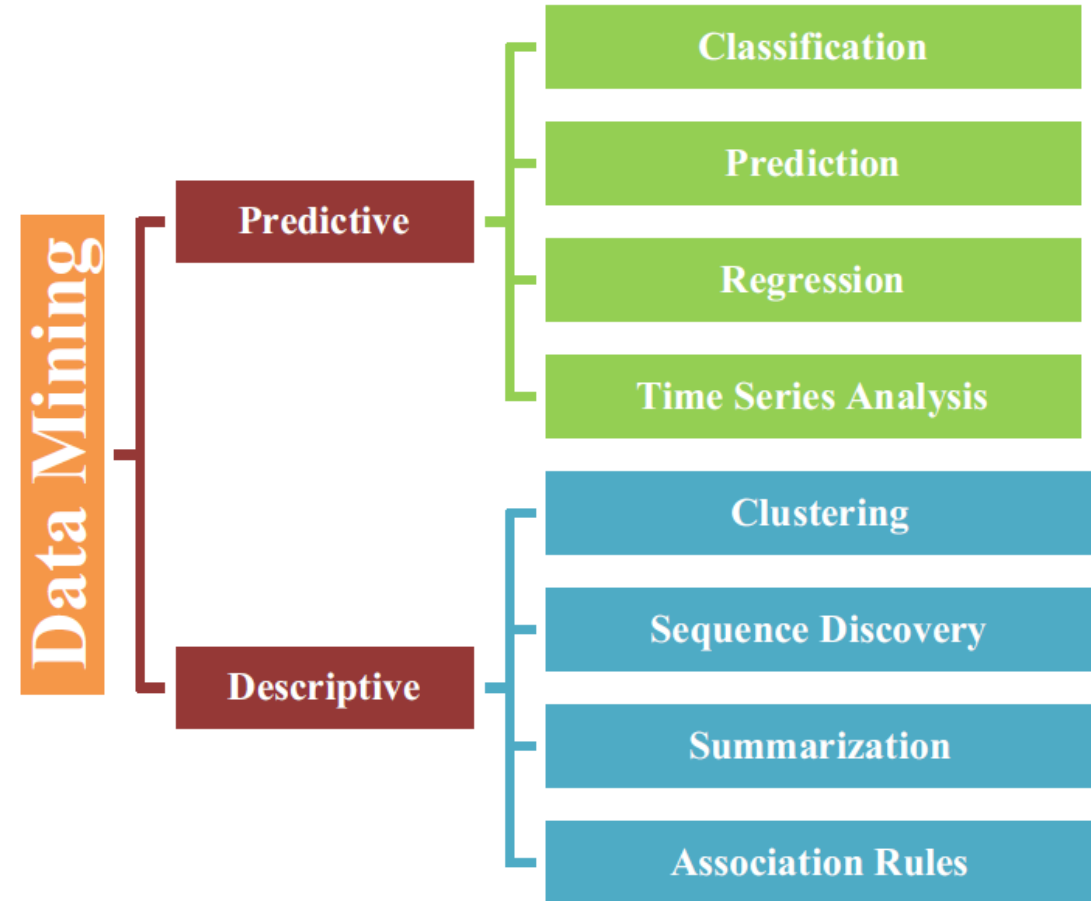
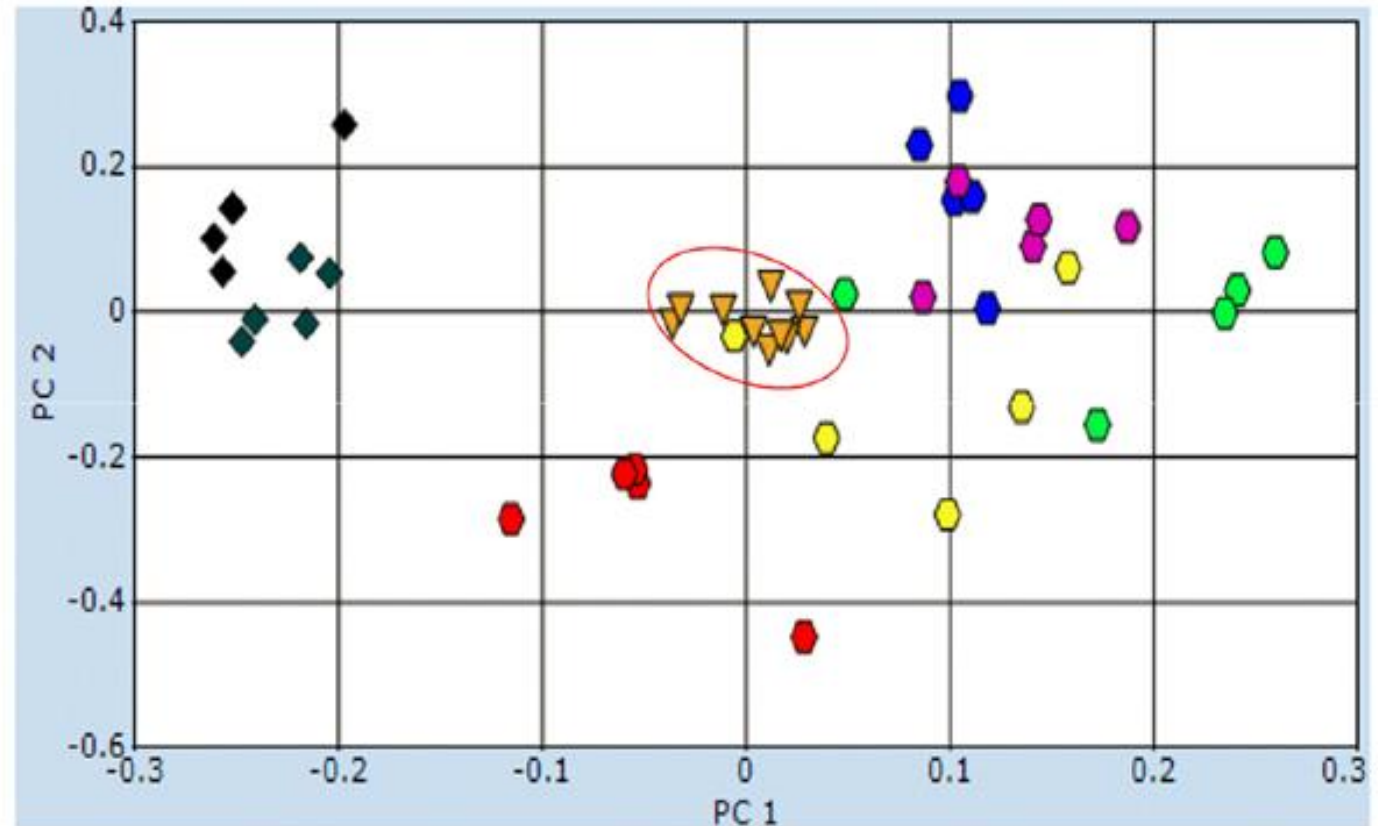


Figure: Data mining models. From Mustafa et al. (2022)

# What kinds of patterns can be mined? (cont'd)

## Concept/Class Description:

- Data can be associated with items/categories of items or concept.
  - E.g. items - tomato, cucumber, shrimp... or categories of items – fruits and vegetables, dairy products, etc ...
  - E.g. concepts – highly/frequently contaminated food commodities

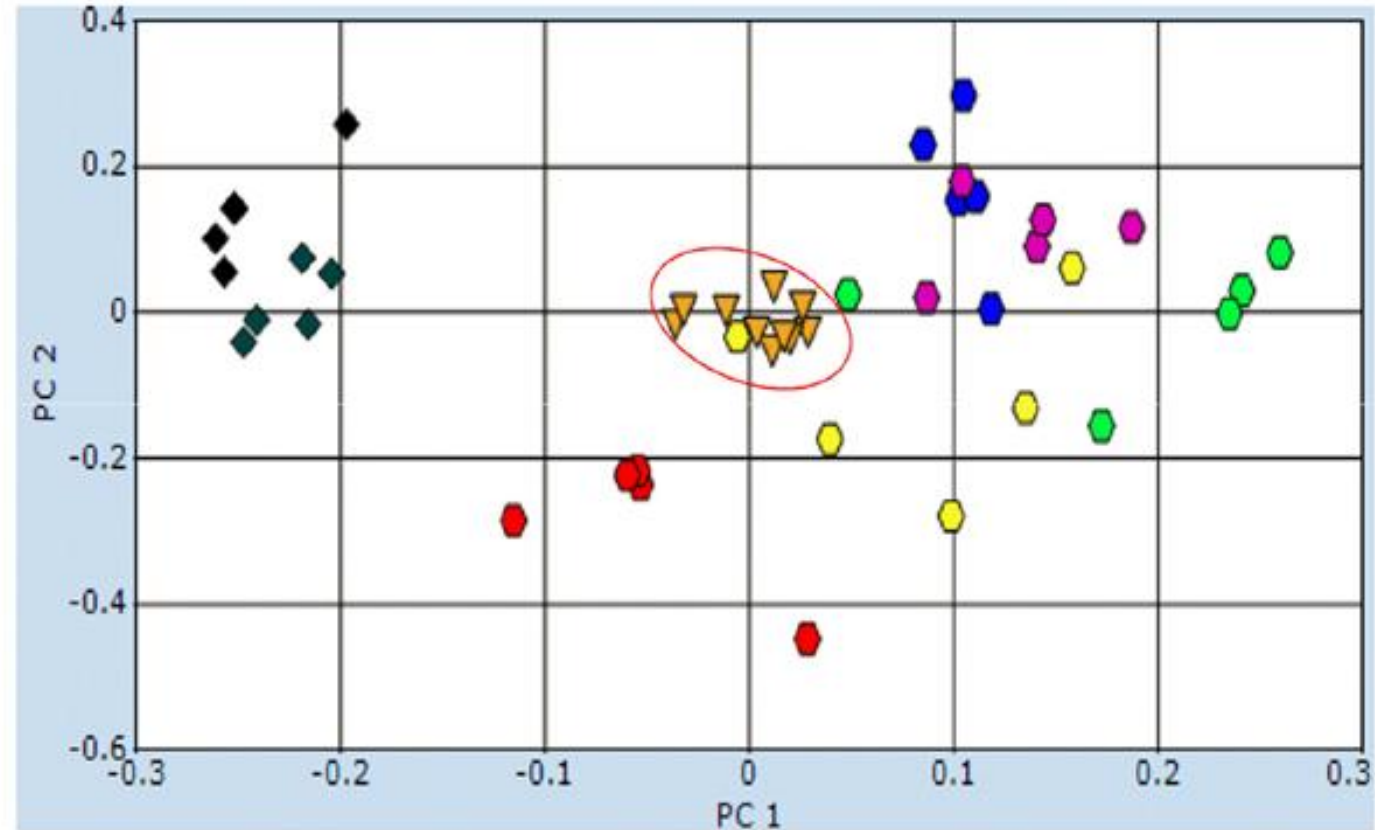


# What kinds of patterns can be mined? (cont'd)

## Concept/Class Description:

### ☐ Descriptions can be derived via

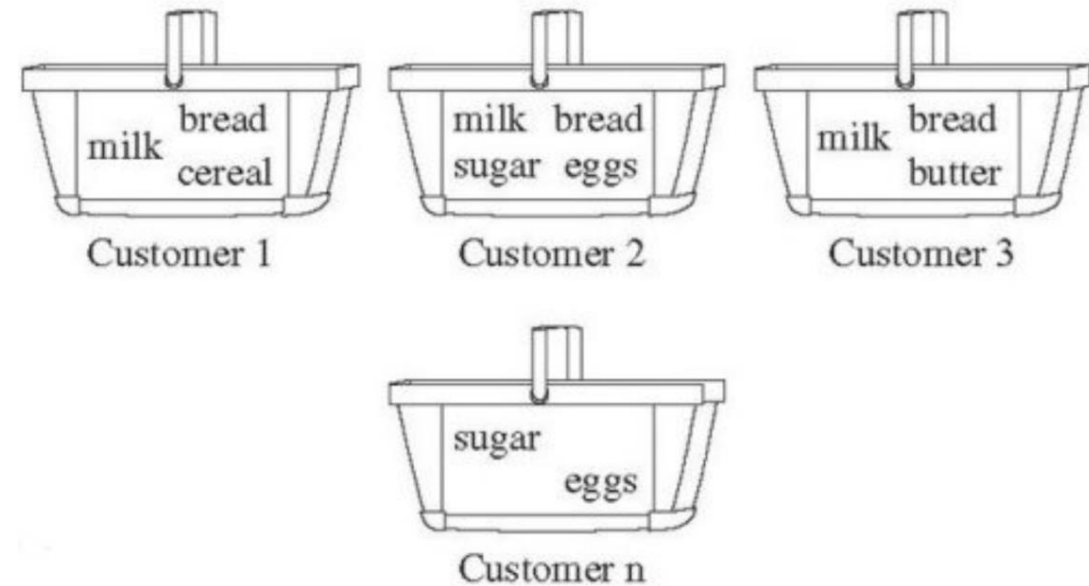
- Data characterization – summarizing the general characteristics of a target class of data.
  - E.g. summarizing the characteristics of certain food by categories
- Data discrimination – comparing the target category with one or a set of comparative category
  - E.g. Compare the contamination of fruits with dairy products during the same period for mycotoxins
- Or both of the above



# What kinds of patterns can be mined? (cont'd)

## Mining Frequent Patterns, Associations and Correlations

- ❑ Frequent itemset: a set of items that frequently appear together in a data set (e.g. consumption of milk and bread)
- ❑ Frequent subsequence: a sequential pattern like if wheat flour contains mycotoxins, then bread will probably contain mycotoxins



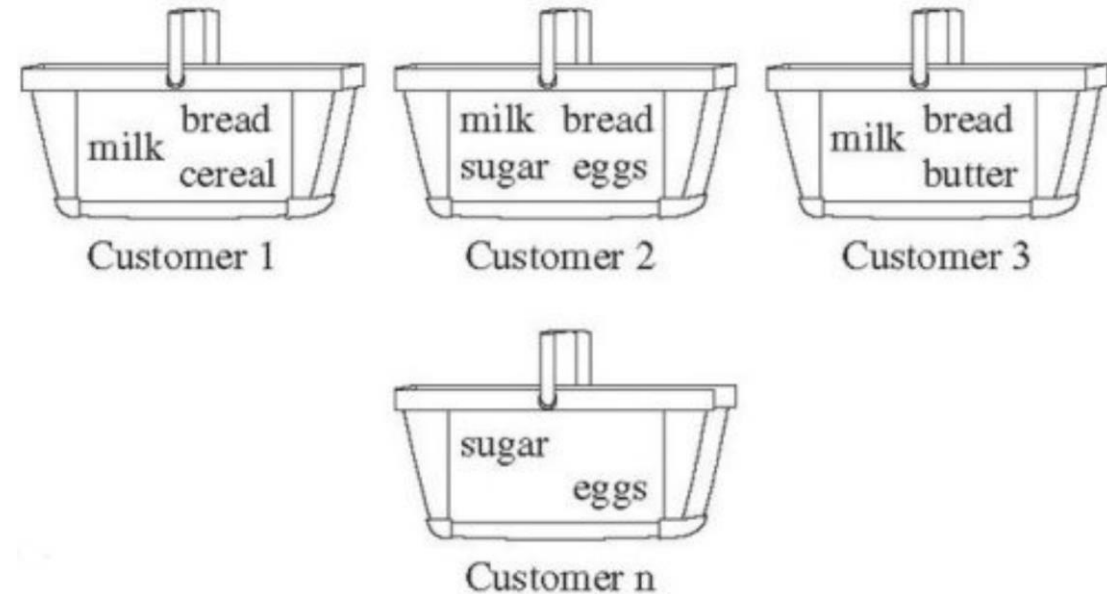
# What kinds of patterns can be mined? (cont'd)

## Mining Frequent Patterns, Associations and Correlations

### ❑ Association Analysis: find frequent patterns

- E.g. a sample analysis result – an association rule: if wheat flour is contaminated by pesticides, there is a 50% chance wheat flour does not contain mycotoxins. 1% of all of the data point under analysis showed that wheat flour are contaminated by pesticides and mycotoxins at the same time.

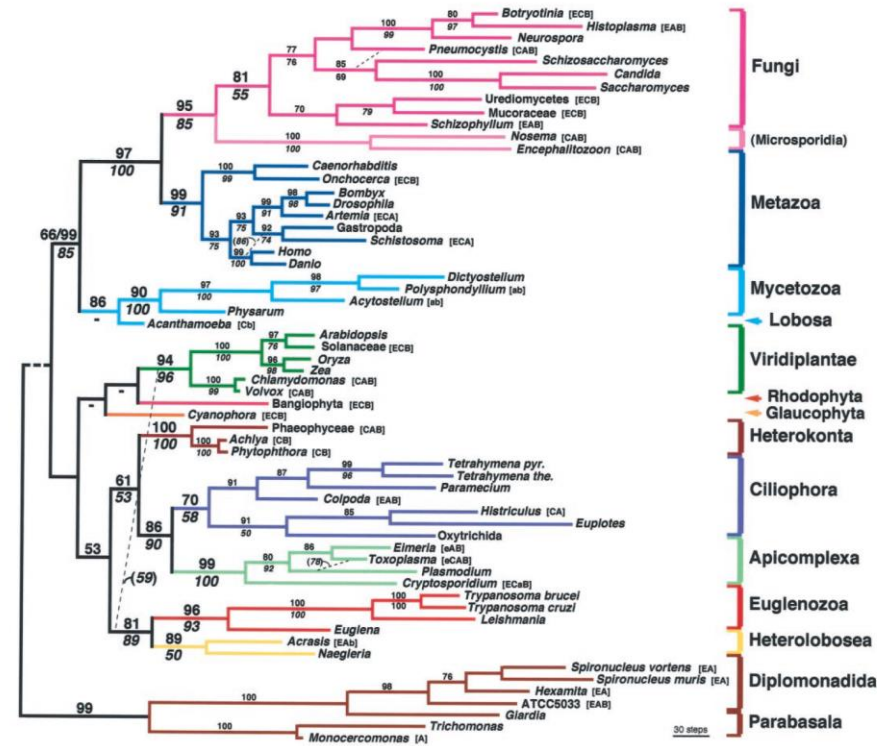
### ❑ Correlation Analysis: additional analysis to find statistical correlations between associated pairs



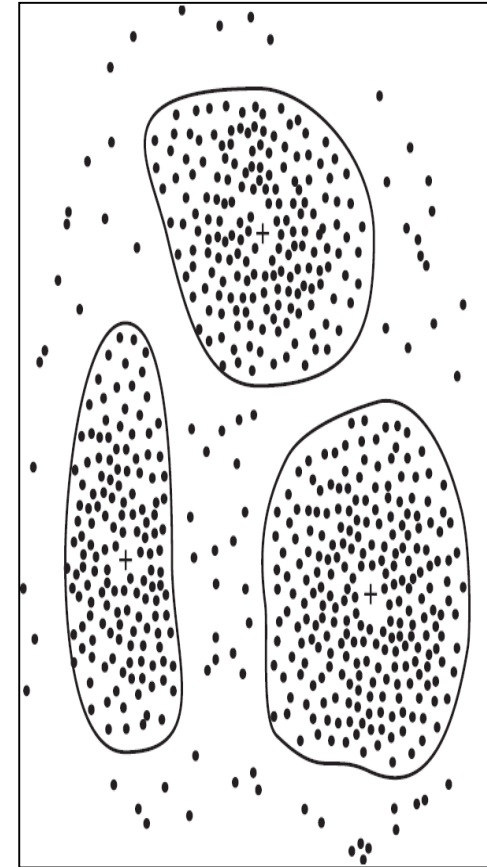
# What kinds of patterns can be mined? (cont'd)

## Cluster Analysis

- ❑ Class label is unknown: group data to form new classes
- ❑ Clusters of objects are formed based on the principle of *maximizing intra-class similarity & minimizing interclass similarity*
  - E.g. Identify homogeneous subpopulations of customers. These clusters may represent individual target groups for risk assessment.



Downloaded from www.sciencemag.org on July 27, 2009



# What kinds of patterns can be mined? (cont'd)

## ❑ Outlier Analysis

- Data that do not comply with the general behavior or model.
- Outliers are usually discarded as noise or exceptions.
- Useful for fraud detection or bad practices.
  - E.g. Detect large amount of pesticides in particular products, addition of sugar in honey, etc...

## ❑ Evolution Analysis

- Describes and models regularities or trends for objects whose behavior changes over time.
  - E.g. Correction in the risk assessment after the setup of an MRL.



# What kinds of patterns can be mined? (cont'd)

## □ Classification and Prediction (predictive approach)

### ■ Classification

- The process of finding a model that describes and distinguishes the data categories or items, for the purpose of being able to use the model to predict the category of items whose category label is still unknown.
- The derived model is based on the analysis of a set of training data (data objects whose category label is known).
- The model can be represented in *classification (IF-THEN) rules*, decision trees, *neural networks*, etc.

### ■ Prediction

- Predict missing or unavailable numerical data values



# Are All of the Patterns Interesting?

- ❑ Data mining may generate thousands of patterns: Not all of them are interesting
- ❑ A pattern is **interesting** if it is
  - Easily understood by humans (Ex: Shrimps are contaminated by pesticides if dates are contaminated by mycotoxins)
  - Valid on new or test data with some degree of certainty
  - Potentially useful
  - Novel (bread and cheese are eaten together... yes)
  - Validates some hypothesis that a user seeks to confirm



# Are All of the Patterns Interesting? (cont'd)

## ❑ Objective measures

- Based on statistics and structures of patterns, e.g., support, confidence, etc. (Rules that do not satisfy a threshold are considered uninteresting.)

## ❑ Subjective measures

- Reflect the needs and interests of a particular user.
  - E.g. General population, Subgroup: babies, elderly, allergic people, etc...
- Based on user's belief in the data.
  - e.g., Patterns are interesting if they are unexpected, or can be used for strategic planning, etc

## ❑ Objective and subjective measures need to be combined.

# Are All of the Patterns Interesting? (cont'd)

## ❑ Find all the interesting patterns: Completeness

- Unrealistic and inefficient
- User-provided constraints and interestingness measures should be used

## ❑ Search for only interesting patterns: An optimization problem

- Highly desirable
- No need to search through the generated patterns to identify truly interesting ones.
- Measures can be used to rank the discovered patterns according their interestingness.
- Prioritization is the key

# Major Issues in Data Mining

## ❑ Conceptual issues:

- Data mining is based on reliable data
  - Data type
  - Choice of variables
  - Reliable data sources
  - Problem with duplicate data point
  - Lack of representativeness for some items/categories
- Handling noisy or incomplete data
  - Require data cleaning methods and data analysis methods that can handle noise
- Types of patterns
  - Negative
  - Dependence
  - Causal
  - Spatial
  - Temporal



# Major Issues in Data Mining (cont'd)

## □ Implementation issues:

- Data issues
  - Optimization
  - Multi database
- Pattern discovery
  - Computational problems
  - Algorithms and data structure
- Interactive mining of knowledge at multiple levels of abstraction
  - Require the development of numerous data mining techniques
  - High-level query languages need to be developed
  - Difficult to know exactly what will be discovered
  - Allow users to focus the search, refine data mining requests

# Major Issues in Data Mining (cont'd)

## □ Performance Issues

- Efficiency and scalability
  - Huge amount of data
  - Running time must be predictable and acceptable
- Parallel, distributed and incremental mining algorithms
  - Divide the data into partitions and processed in parallel
  - Incorporate database updates without having to mine the entire data again from scratch

# Major Issues in Data Mining (cont'd)

## □ Application issues

- Incorporation of background knowledge
  - Guide the discovery process
  - Allow discovered patterns to be expressed in concise terms and different levels of abstraction
- Presentation and visualization of results
  - Knowledge should be easily understood and directly usable
  - High level languages, visual representations or other expressive forms
  - Require the DM system to adopt the above techniques
- Pattern evaluation – the interestingness problem
  - How to develop techniques to access the interestingness of discovered patterns, especially with subjective measures bases on user beliefs or expectations

