CEFET/RJ

Laboratório Nacional de Computação Científica

UNIDADE DE PESQUISA DO MCTI

INTRODUCTION TO DATA MINING

Eduardo Ogasawara

eogasawara@ieee.org

https://eic.cefet-rj.br/~eogasawara

Why Data Mining?

- *Big Data scenario:*
  - *The explosive growth of data: from terabytes to petabytes*
  - *Data collection and data availability*
    - *Automated data collection tools, database systems, Web*
  - *Major sources of abundant and diverse data*
    - *Business: Web, e-commerce, transactions*
    - *Science: sensors, astronomy, bioinformatics, simulation*
    - *Society and everyone: news, photos, videos, open data, IoT*
- *We are drowning in data but starving for knowledge!*
- *"Need is the mother of invention"*
  - *Data mining - Automated analysis of massive data sets*

[1] F. Berman, 2008, Got data?: A guide to data preservation in the information age, Communications of the ACM, v. 51, n. 12, p. 50–56.

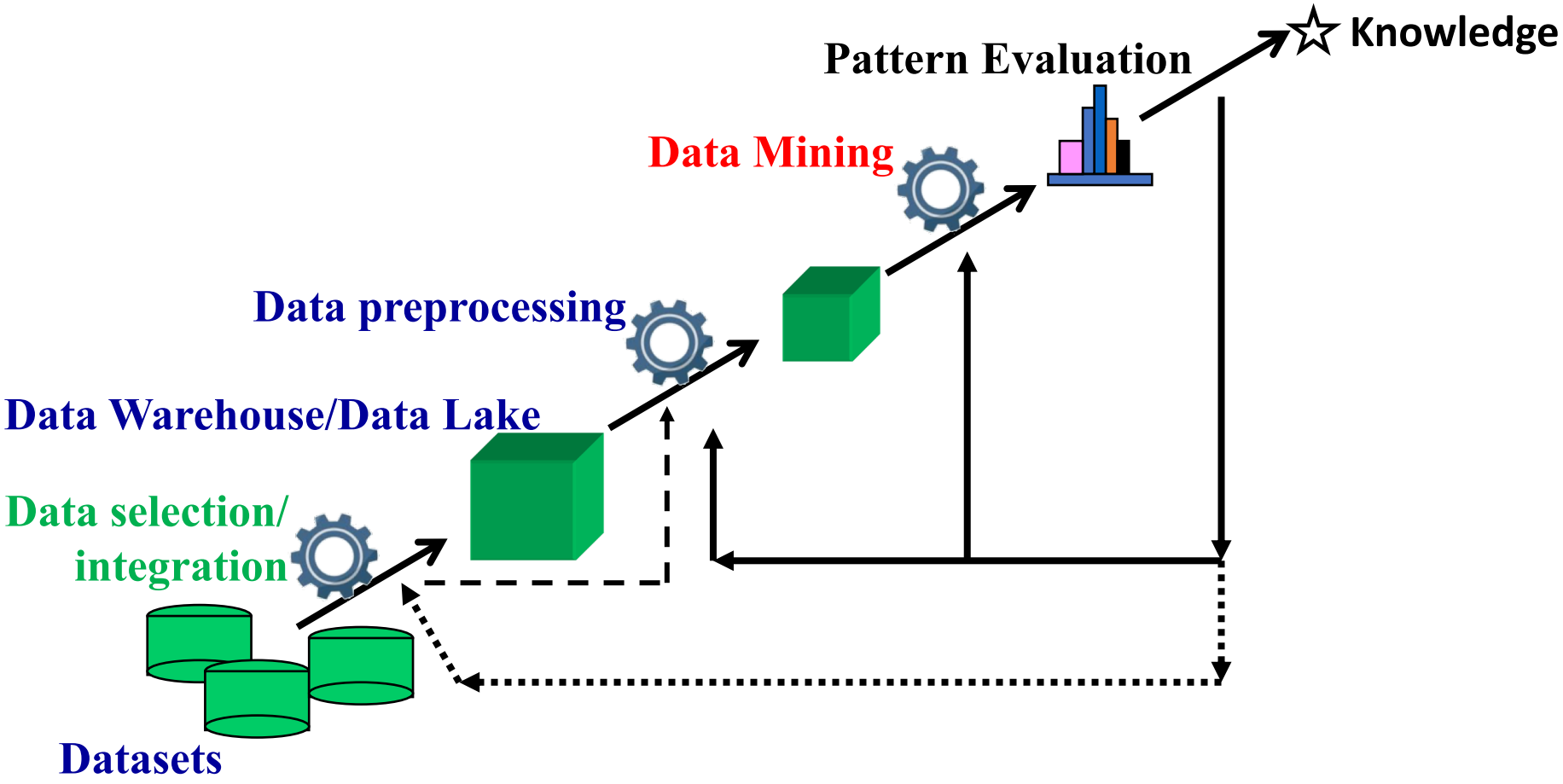
What is Data Mining?

- *Data mining (knowledge discovery from data)*
  - *Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from a ~~massive~~ amount of data*
- *Alternative names*
  - *Knowledge Discovery in Databases (KDD)*
  - *Knowledge Extraction*
  - *Business Intelligence*
  - *Data Analysis*

[1] U. Fayyad e R. Uthurusamy, 2002, Evolving data mining into solutions for insights, Communications of the ACM, v. 45, n. 8, p. 28–31.

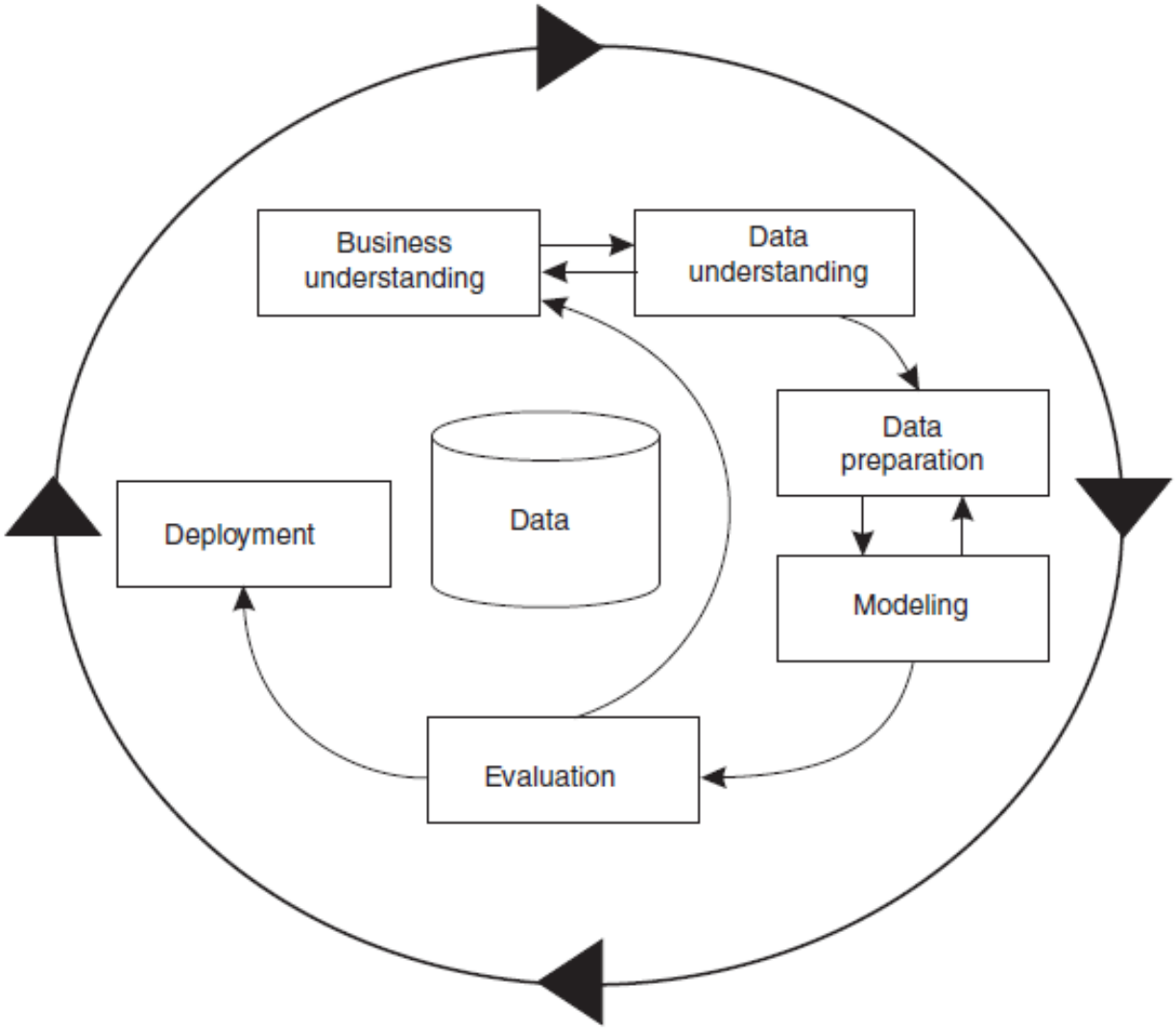
Knowledge discovery from data (KDD) process (Database perspective)

- *This is a view from typical database systems*
- *Data mining plays an essential role in the KDD process*



[1] J. Han, J. Pei, e M. Kamber, 2011, Data Mining: Concepts and Techniques. Elsevier.

Cross-Industry Standard Process for Data Mining (CRISP-DM)



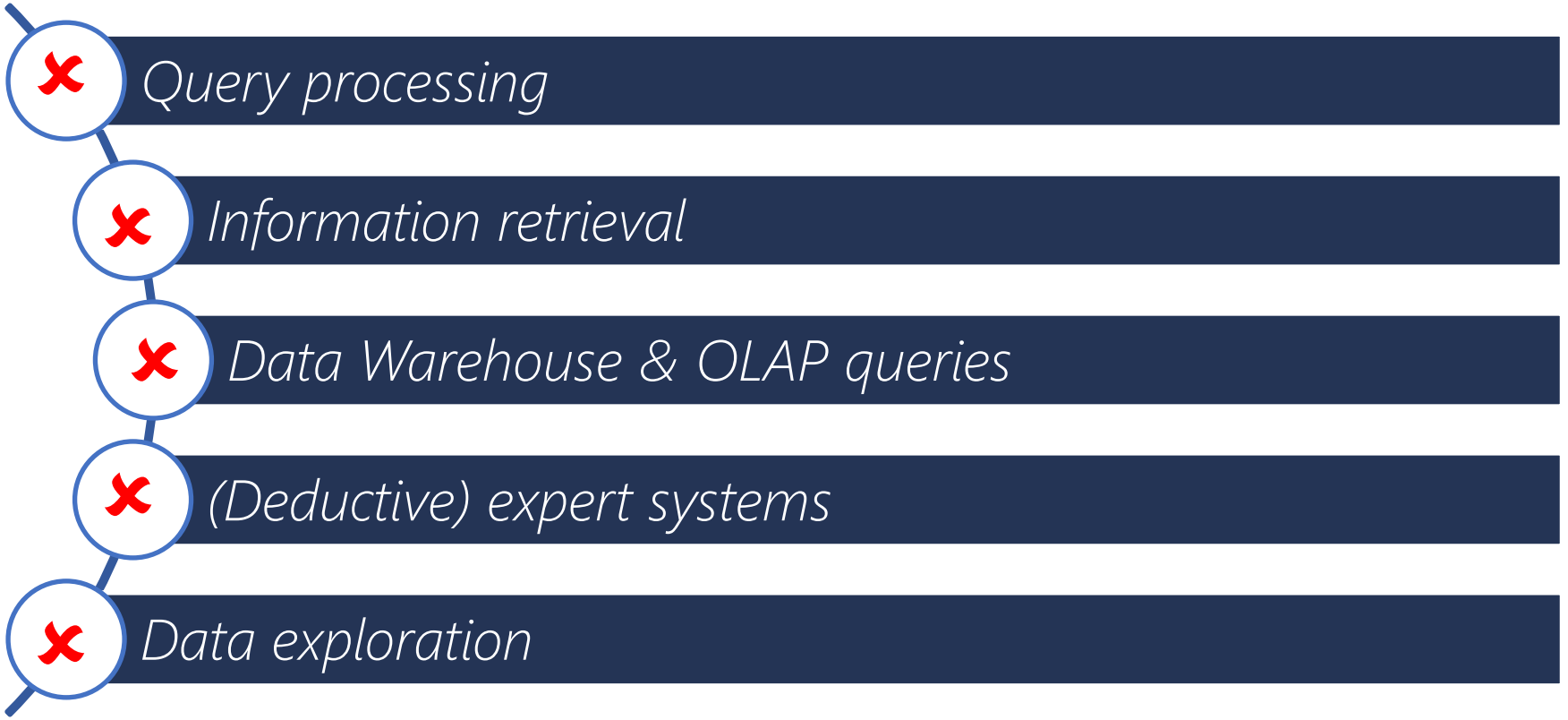
[1] G. Mariscal, Ó. Marbán, e C. Fernández, 2010, A survey of data mining and knowledge discovery process models and methodologies, Knowledge Engineering Review, v. 25, n. 2, p. 137–166.

Knowledge discovery in databases (KDD) process vs. Cross-Industry Standard Process for Data Mining (CRISP-DM)

<table>
<tr><th>Methodology</th><th colspan="6">Phases</th></tr>
<tr><td>CRISP-DM</td><td>Business understanding</td><td>Data understanding</td><td>Data preparation</td><td>Modeling</td><td>Evalution</td><td>Deployment</td></tr>
<tr><td rowspan="3">KDD-Detailed</td><td rowspan="3">Learning the application domain</td><td rowspan="3">Creating a target data set</td><td>Data cleaning and pre-processing</td><td>Choosing the function of DM</td><td rowspan="3">Interpretation</td><td rowspan="3">Using discovered knowledge</td></tr>
<tr><td rowspan="2">Data reduction and projection</td><td>Choosing the DM algorithm</td></tr>
<tr><td>Data mining</td></tr>
<tr><td rowspan="2">KDD-Outlined</td><td rowspan="2"></td><td>Selection</td><td>Pre-processing</td><td>Modeling</td><td rowspan="2">Interpretation / Evalution</td><td rowspan="2"></td></tr>
<tr><td>Pre-processing</td><td>Transformation</td><td>Data mining</td></tr>
</table>

[1] G. Mariscal, Ó. Marbán, e C. Fernández, 2010, A survey of data mining and knowledge discovery process models and methodologies, Knowledge Engineering Review, v. 25, n. 2, p. 137–166.

Is everything "data mining"?



[1] J. Han, J. Pei, e M. Kamber, 2011, Data Mining: Concepts and Techniques. Elsevier.

Multi-Dimensional View of Data Mining

- *Data models*
- *Knowledge to be mined*
- *Data mining functions*
- *Methods/Techniques used*
- *Applications*

[1] J. Han, J. Pei, e M. Kamber, 2011, Data Mining: Concepts and Techniques. Elsevier.

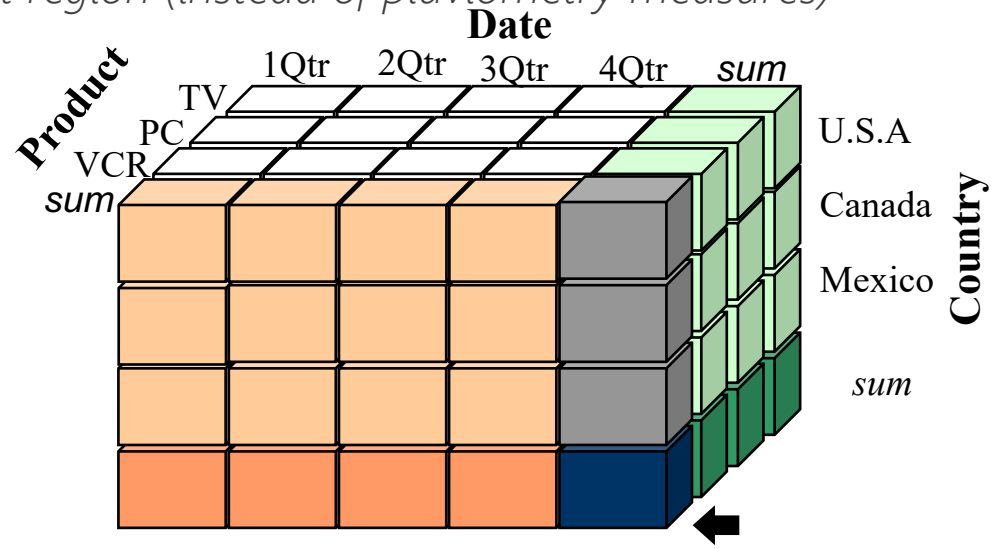
Data Mining: on what kinds of data models?

- *Database-oriented data sets and applications*
  - *A Relational database, data warehouse, transactional database*
  - *Object-relational databases, Heterogeneous databases, and legacy databases*
- *Advanced data sets and advanced applications*
  - *Data streams 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*

[1] J. Han, J. Pei, e M. Kamber, 2011, Data Mining: Concepts and Techniques. Elsevier.

Data Mining Function: (1) Generalization

- *Information integration and data warehouse construction*
  - *Data cleaning, transformation, integration, and multidimensional data model*
- *Data cube technology*
  - *Scalable methods for computing (i.e., materializing) multidimensional aggregates*
  - *OLAP (online analytical processing)*
- *Multidimensional concept description: characterization and discrimination*
  - *Generalize, summarize, and contrast data characteristics,*
    - *E.g.: dry vs. wet region (instead of pluviometry measures)*



[1] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, e H. Pirahesh, 1997, Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals, Data Mining and Knowledge Discovery, v. 1, n. 1, p. 29–53.

Data Mining Function: (2) Association and Correlation Analysis

- *Frequent patterns (or frequent item sets)*
  - *What items are frequently purchased together in your supermarket?*
- *Association, correlation vs. causality*
  - *Typical association rules*
    - *Diaper → Beer [0.5%, 75%] (support, confidence)*
  - *Are strongly associated items also strongly correlated?*
- *How to mine such patterns and identify such rules efficiently in large datasets?*
- *How to use (rank) such patterns?*

Mining Association Rules between Sets of Items in Large Databases

Rakesh Agrawal Tomasz Imieliński* Arun Swami

IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120

Abstract

We are given a large database of customer transactions. Each transaction consists of items purchased by a customer in a visit. We present an efficient algorithm that generates all significant association rules between items in the database. The algorithm incorporates buffer management and novel estimation and pruning techniques. We also present results of applying this algorithm to sales data obtained from a large retailing company, which shows the effectiveness of the algorithm.

1 Introduction

Consider a supermarket with a large collection of items. Typical business decisions that the management of the supermarket has to make include what to put on sale, how to design coupons, how to place merchandise on shelves in order to maximize the profit, etc. Analysis of past transaction data is a commonly used approach in order to improve the quality of such decisions. Until recently, however, only global data about the cumulative sales during some time period (a day, a week, a month, etc.) was available on the computer. Progress in bar-code technology has made it possible to store the details of each transaction. A detail record of sale is the point-of-transaction basis. Basket data type transactions do not necessarily consist of items bought together at the same point of time. It may consist of items bought by a customer over a period of time. Examples include monthly purchases by members of a book club or a music club.

Several organizations have collected massive amounts of such data. These data sets are usually stored

on tertiary storage and are very slowly migrating to database systems. One of the main reasons for the limited success of database systems in this area is that current database systems do not provide necessary functionality for a user interested in taking advantage of this information.

This paper introduces the problem of "mining" a large collection of basket data type transactions for association rules between sets of items with some minimum specified confidence, and presents an efficient algorithm for this purpose. An example of such an association rule is the statement that 90% of transactions that purchase bread and butter also purchase milk. The antecedent of this rule consists of bread and butter and the consequent consists of milk alone. The number 90% is the confidence factor of the rule.

The work reported in this paper could be viewed as a step towards enhancing databases with functionalities to process queries such as (we have omitted the confidence factor specifications):

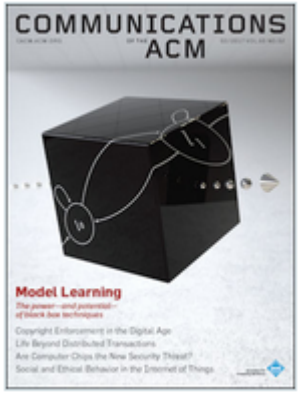
- Find all rules that have "Diet Coke" as consequent. These rules may help plan what the store should do to boost the sale of Diet Coke.
- Find all rules that have "bagels" in the antecedent. These rules may help determine what products may be impacted if the store discontinues selling bagels.
- Find all rules that have "sausage" in the antecedent and "mustard" in the consequent. This query can be phrased alternatively as a request for the additional items that have to be sold together with sausage in order to make it highly likely that mustard will also be sold.
- Find all the rules relating items located on shelves A and B in the store. These rules may help shelf planning by determining if the sale of items on shelf A is related to the sale of items on shelf B.
- Find the "best" k rules that have "bagels" in the consequent. Here, "best" can be formalized in terms of the confidence factors of the rules, or in terms

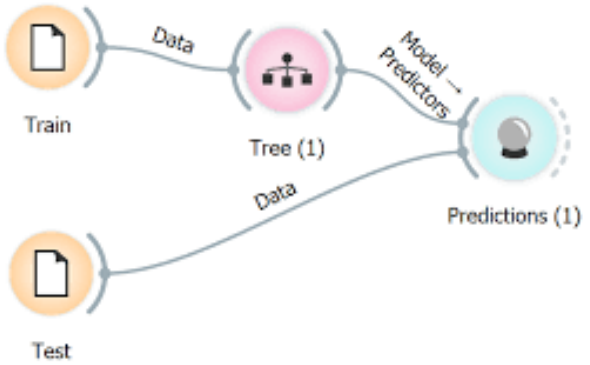
207

[1] R. Agrawal, T. Imieliński, e A. Swami, 1993, Mining Association Rules Between Sets of Items in Large Databases, ACM SIGMOD Record, v. 22, n. 2, p. 207–216.

Data Mining Function: (3) Prediction

- *Classification and label prediction*
  - *Construct models based on some training examples*
    - *Data-driven models*
  - *Describe and distinguish classes or concepts for future prediction*
    - *E.g., classify countries based on (climate)*
- *Typical methods:*
  - *Decision trees*
  - *Naive Bayes classifier*
  - *Support vector machines*
  - *Neural networks & deep learning*
  - *Random Forest*
  - *Linear/Logistic regression*
- *Typical applications*
  - *Scientific*
  - *Industry & enterprises*
  - *Government & society*

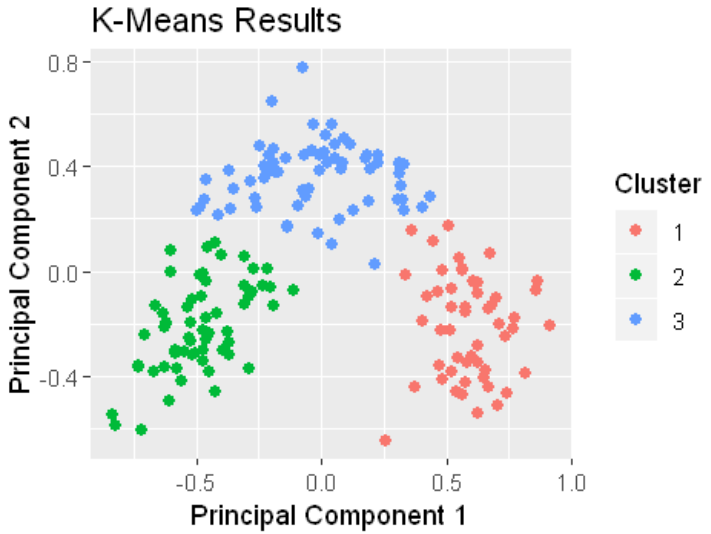




[1] F. Vaandrager, 2017, Model learning, Communications of the ACM, v. 60, n. 2, p. 86–95.
[2] https://towardsdatascience.com/data-science-made-easy-data-modeling-and-prediction-using-orange-f451f17061fa

Data Mining Function: (4) Cluster Analysis

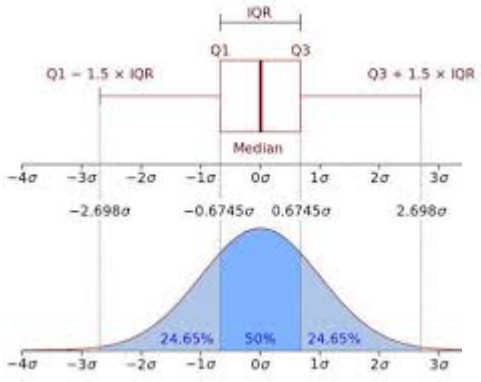
- *Unsupervised learning (i.e., Class label is unknown)*
- *Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns*
- *Principle: Maximizing intra-class similarity & minimizing interclass similarity*
- *Many methods and applications*



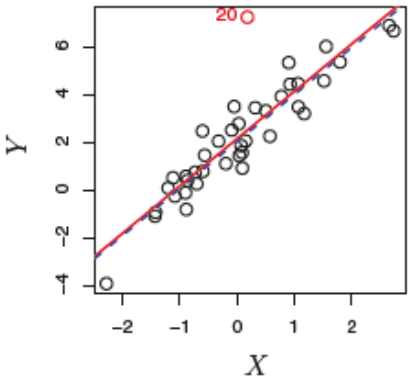
[1] A.K. Jain, 2010, Data clustering: 50 years beyond K-means, Pattern Recognition Letters, v. 31, n. 8, p. 651–666.

Data Mining Function: (5) Outlier Analysis

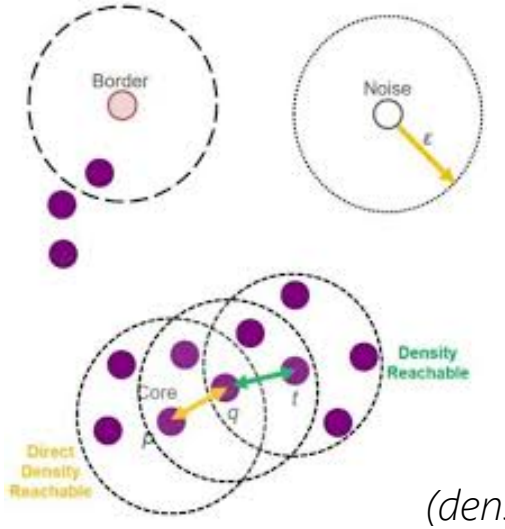
- *Outlier: A data object that does not comply with the general behavior of the data*
- *Noise or exception? — One person's garbage could be another person's treasure*
- *Methods: statistical, clustering or regression analysis*
- *Useful in fraud detection, rare events analysis*



(distribution)



(model)

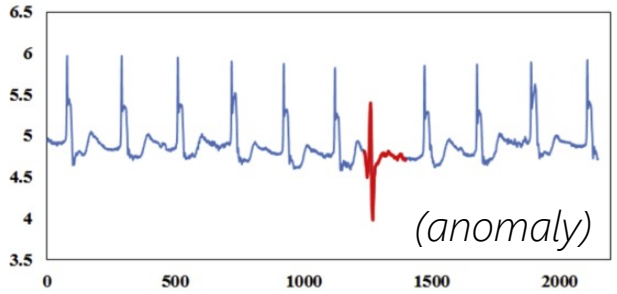


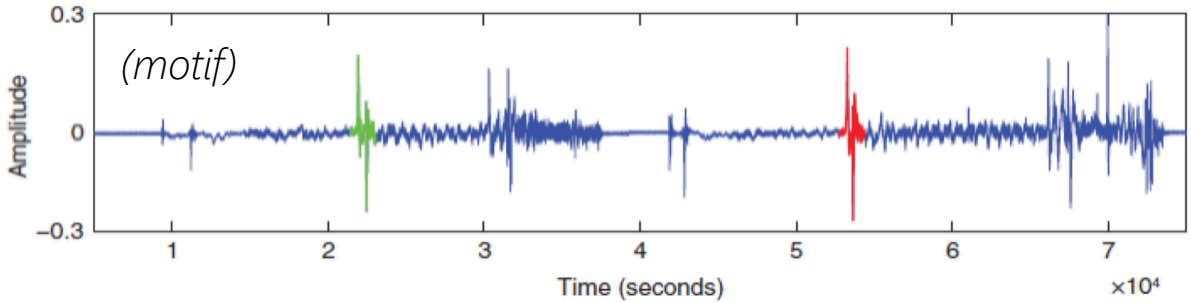
(density)

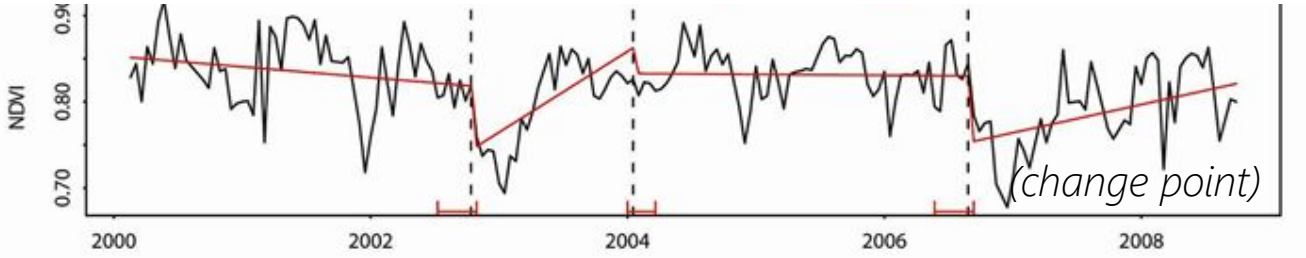
[1] R.J. Larsen e M.L. Marx, 2017, An Introduction to Mathematical Statistics and Its Applications. Pearson Education.
[2] C.C. Aggarwal, 2013, Outlier Analysis. Springer Science & Business Media.

Data Mining Function: (6) Event detection

- Anomalies (distribution, distance from a model, volatility)
  - A pattern or observation that do not conform to expected behavior
    - Build from another process
- Motifs
  - A pattern (unknown) that occurs a significant number of times in a dataset
- Change points / concept drifts
  - Points (or time intervals) that mark significant change in dataset behavior
  - They separate different states in the process that generates the dataset







V. Chandola, A. Banerjee, e V. Kumar, 2009, Anomaly detection: A survey, ACM Computing Surveys, v. 41, n. 3

S. Torkamani and V. Lohweg, "Survey on time series motif discovery," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 7, no. 2, 2017.

J.-I. Takeuchi e K. Yamanishi, 2006, A unifying framework for detecting outliers and change points from time series, IEEE Transactions on Knowledge and Data Engineering, v. 18, n. 4, p. 482–492.

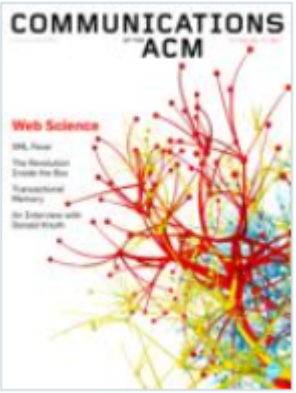
Data Mining Function: (7) Sequential Pattern

- *Sequential pattern mining*
  - *Detect and analyze frequent subsequences of events, items, or tokens occurring in an ordered metric space*

[1] C.H. Mooney e J.F. Roddick, 2013, Sequential pattern mining - Approaches and algorithms, ACM Computing Surveys, v. 45, n. 2

Data Mining Function: (8) Structure and Network Analysis

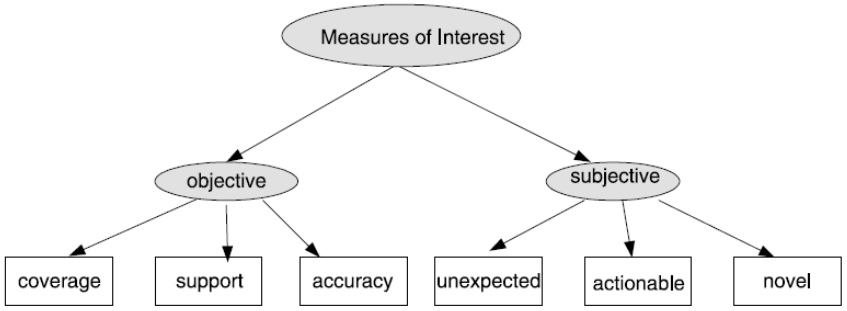
- *Graph mining*
  - *Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)*
- *Information network analysis*
  - *Social networks: actors (objects, nodes) and relationships (edges)*
    - *e.g., Web community discovery*
- *Web mining*
  - *Opinion mining, topic detection, sentiment analysis*

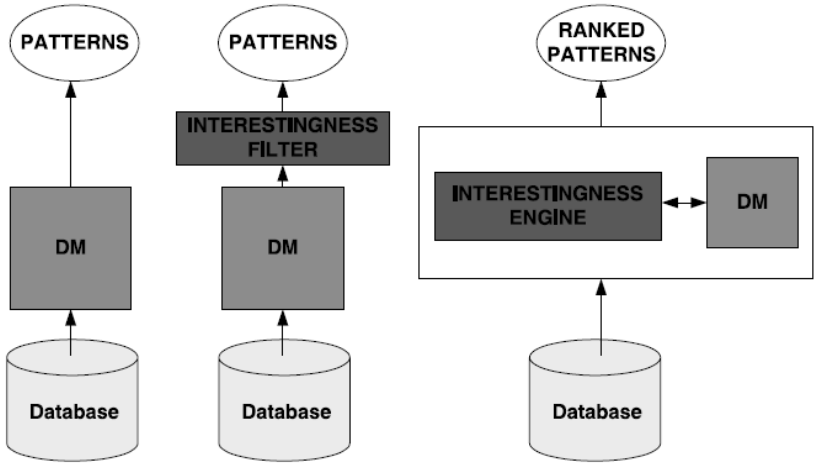


[1] T. Tiropanis, W. Hall, J. Crowcroft, N. Contractor, e L. Tassiulas, 2015, Network science, web science, and internet science, Communications of the ACM, v. 58, n. 8, p. 76–82.

Evaluation of Knowledge

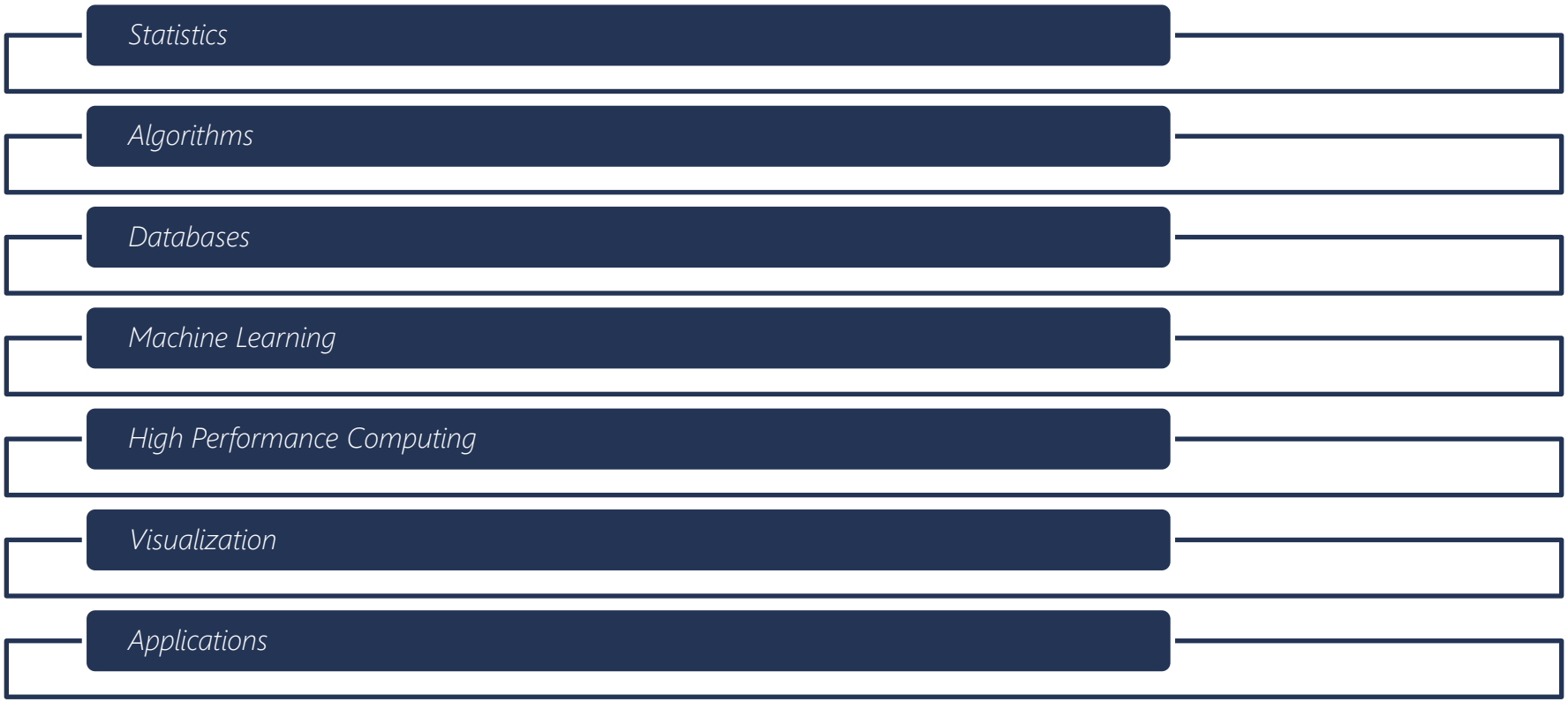
- *Are all mined knowledge interesting?*
  - *One can mine the tremendous amount of "patterns"*
  - *Some may fit only certain dimension space (time, location)*
  - *Some may not be representative, may be transient*
- *Evaluation of mined knowledge → directly mine only interesting knowledge?*





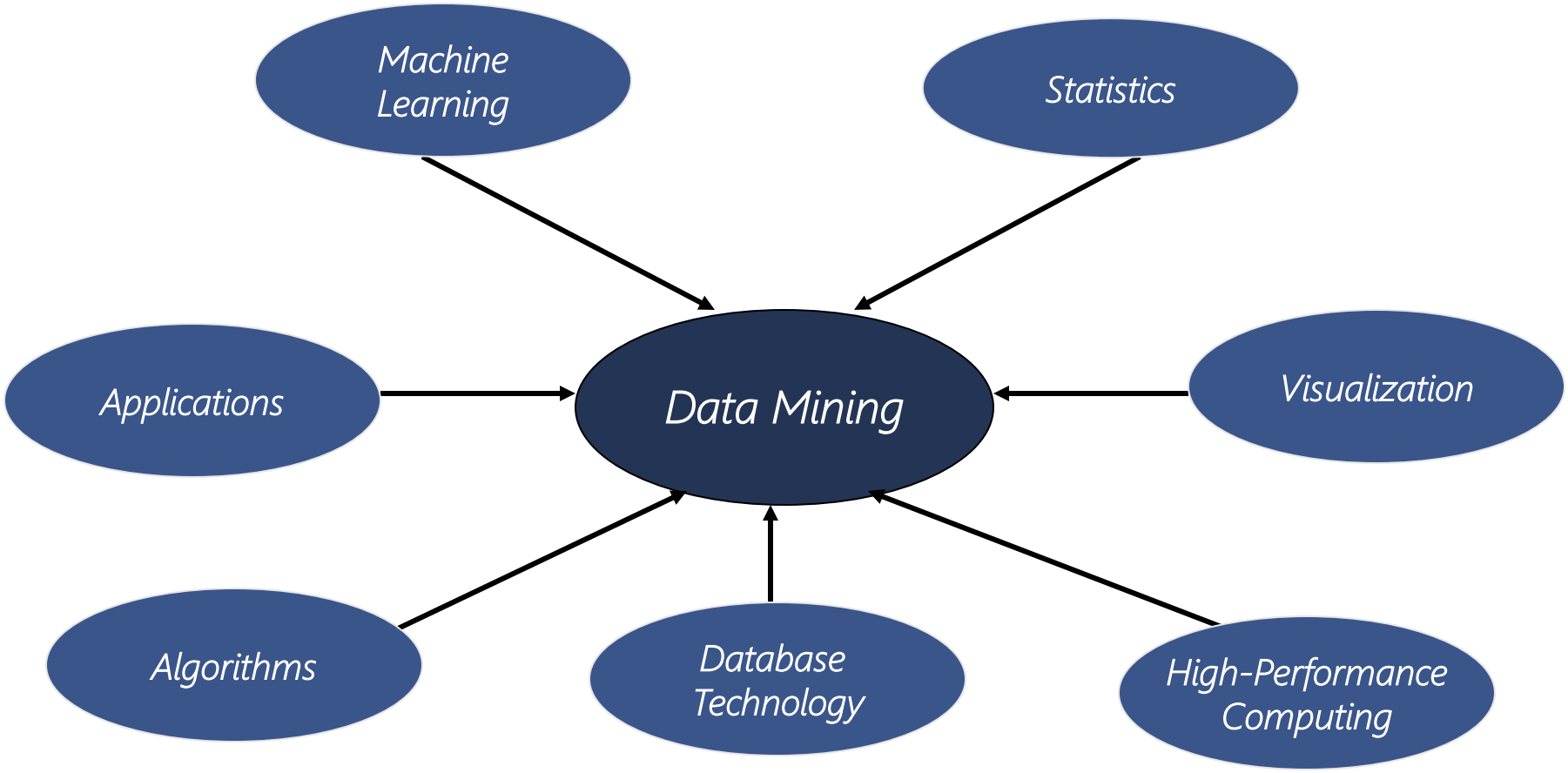
[1] L. Geng e H.J. Hamilton, 2006, Interestingness measures for data mining: A survey, ACM Computing Surveys, v. 38, n. 3, p. 3.

Data Mining: Confluence of Multiple Disciplines



[1] J. Han, J. Pei, e M. Kamber, 2011, Data Mining: Concepts and Techniques. Elsevier.

Data Mining: Confluence of Multiple Disciplines



[1] J. Han, J. Pei, e M. Kamber, 2011, Data Mining: Concepts and Techniques. Elsevier.

Why Confluence of Multiple Disciplines?

- *Tremendous amount of data*
  - *Algorithms must be scalable to handle big data*
- *High-dimensionality of data*
- *High complexity of data*
  - *Data streams, sensor data, spatial-temporal, text, multimedia*
- *New and sophisticated applications*

[1] J. Han, J. Pei, e M. Kamber, 2011, Data Mining: Concepts and Techniques. Elsevier.

Availability of data

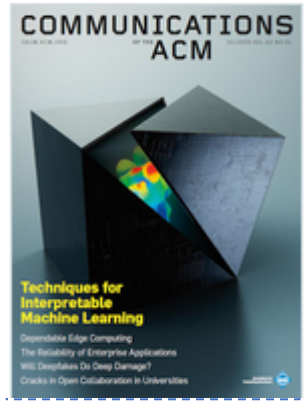
- *Do you have access to the data?*
- *Can you use the data?*
- *Can you publish your results?*
- *Is it big or small data?*
  - *Is enough to be considered data mining?*

Characteristic	Small data	Big data
Volume	Limited to large	Very large
Exhaustivity	Samples	Entire populations
Resolution and indexicality	Coarse and weak to tight and strong	Tight and strong
Relationality	Weak to strong	Strong
Velocity	Slow, freeze-framed	Fast
Variety	Limited to wide	Wide
Flexible and scalable	Low to middling	High

[1] R. Kitchin e T.P. Lauriault, 2015, Small data in the era of big data, GeoJournal, v. 80, n. 4, p. 463–475.

Major Issues in Data Mining

- *Mining Methodology*
  - *Mining knowledge in multi-dimensional space*
  - *Handling noise, uncertainty, and incompleteness of data*
  - *Pattern evaluation and constraint-guided pattern mining*
- *User Interaction*
  - *Interactive mining*
  - *Incorporation of background knowledge*
  - *Presentation and visualization of data mining results*
- *Efficiency and Scalability*
  - *Efficiency and scalability of data mining algorithms*
  - *Parallel, distributed, stream, and incremental mining methods*
- *Diversity of data types*
  - *Handling complex types of data*
  - *Mining dynamic, networked, and global data repositories*
- *Data mining and society*
  - *Social impacts of data mining*
  - *Privacy-preserving data mining*



[1] J. Han, J. Pei, e M. Kamber, 2011, Data Mining: Concepts and Techniques. Elsevier.

Data Mining & Data Science

- *Overview*
  - *Science of data or the study of data*
- *Disciplinary view*
  - *Data Science is a new interdisciplinary field that synthesizes and builds on Statistics, Computer Science, Communication, Management and Sociology to study data and its environments (including domain and contextual aspects) in a way to transform data into knowledge and decisions*

[1] V. Dhar, 2013, Data science and prediction, Communications of the ACM, v. 56, n. 12, p. 64–73.
[2] L. Cao, 2017, Data science: Challenges and directions, Communications of the ACM, v. 60, n. 8, p. 59–68.

Data Mining & Data Analytics

- *Data Analytics*
  - *Theories, technologies, tools and processes that allow the understanding and discovery from data*
  - *Entire process of knowledge discovery: selection, pre-processing, transformation, data mining and interpretation*
- *Descriptive Analytics*
  - *Refers to the type of data analysis that normally uses statistics to describe the data used*
- *Predictive Analytics*
  - *Refers to the type of data analysis that makes predictions about unknown future events and reveals the reasons behind them, usually through advanced analysis*
- *Prescriptive Analytics*
  - *Refers to the type of data analysis that optimizes referrals and recommends actions for smart decision making*
- *Business Intelligence*
  - *Application of Data Analytics to support business decisions*

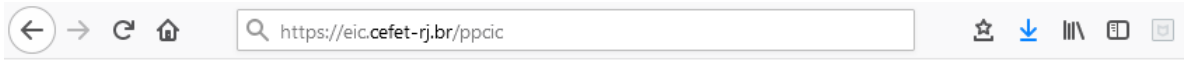
[1] C.-W. Tsai, C.-F. Lai, H.-C. Chao, e A.V. Vasilakos, 2015, Big data analytics: a survey, Journal of Big Data, v. 2, n. 1
[2] L. Cao, 2017, Data science: A comprehensive overview, ACM Computing Surveys, v. 50, n. 3

Where to publish?

- *Data mining, KDD, Data Science*
  - *Conferences: SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, IEEE-DSAA*
  - *Journal: Data Mining and Knowledge Discovery, Statistical Analysis and Data Mining, ACM Transactions on Knowledge Discovery from Data*
- *Database systems*
  - *Conferences: SIGMOD, PODS, VLDB, IEEE-ICDE, EDBT, ICDT, SSDBM*
  - *Journals: IEEE-TKDE, VLDB J., Info. Sys.*
- *AI & Machine Learning*
  - *Conferences: AAAI, ML, IJCNN, IJCAI, NeurIPS*
  - *Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems*
- *Statistics*
  - *Journals: Journal of Applied Statistics, Annals of Data Science*
- *Applications*
  - *Journals*

Onde submeter os artigos? E o Qualis? https://youtu.be/SP_JaNTyRic

PPCIC: Master Degree on Computer Science @ CEFET/RJ
Focus on Data Science



Disciplinas do PPCIC

Disciplina	Núcleo	Créditos
Álgebra Linear Computacional	Específico	3
Álgebra Linear e Grafos	Específico	3
Algoritmos em Grafos	Específico	3
Análise e Projeto de Algoritmos	Básico	3
Aplicações de Robótica	Específico	3
Aprendizado de Máquina	Específico	3
Arquitetura de Computadores	Básico	3
Banco de Dados	Básico	3
Computação Paralela e Distribuída	Básico	3
Fundamentos de Sistemas Multimídia	Específico	3
Gerência de Dados em Larga Escala	Específico	3
Metodologia Científica em Computação	Básico	3
Métodos Estatísticos	Básico	3
Mineração de Dados	Específico	3
Mineração de Processos	Específico	3
Mineração de Textos	Específico	3
Otimização por Metaheurísticas	Específico	3
Pesquisa Operacional	Específico	3

Trending Data Mining Languages

- *Python (Machine Learning Course)*
  - *Scikit learning*
- *R (Data Mining Course)*
  - *Myriad of packages*
- *Spark (Parallel and Distributed Computing)*
  - *Mlib*

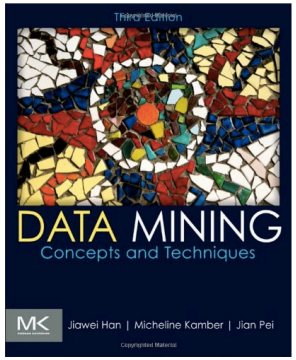
https://www.anaconda.com/

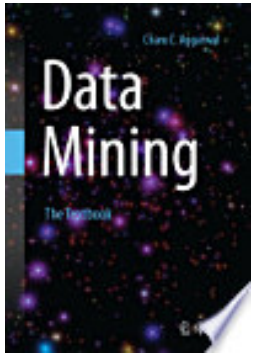
Data Mining Tools

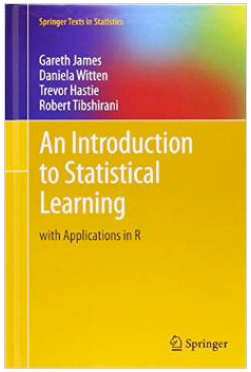
- Rapid Miner (open source)
  - https://rapidminer.com
- Orange (open source)
  - https://orange.biolab.si
- Weka (open source)
  - https://www.cs.waikato.ac.nz/ml/weka
- Knime (open source)
  - https://www.knime.com
- Apache Mahout (open source)
  - http://mahout.apache.org
- Rattle (open source)
  - https://rattle.togaware.com

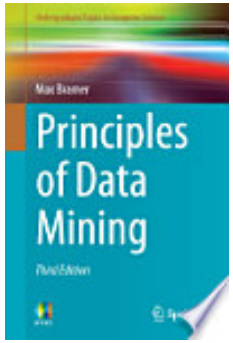
https://www.analyticsinsight.net/top-10-data-mining-tools-to-dominate-2020/
https://www.softwaretestinghelp.com/data-mining-tools/

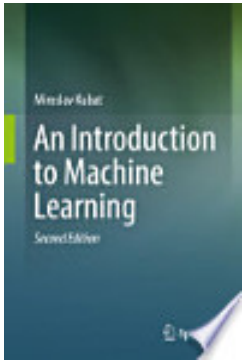
Main References

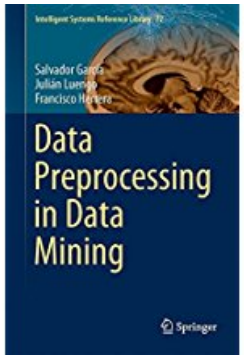


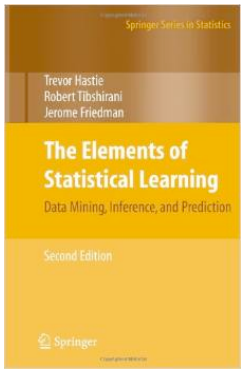


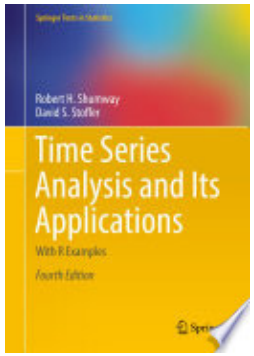


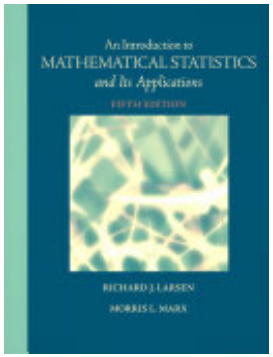












Most of the slides were extracted from Data Mining Concepts and Techniques

Next Topics

- *Next classes*
  - *R programming language*
  - *Exploratory data analysis*
  - *Data preprocessing*
- *Slides and videos*
  - *https://eic.cefet-rj.br/~eogasawara/*
  - *Follow up YouTube channel*
    - *https://www.youtube.com/channel/UCAm1hAXWEqYJfXz4EzzBhVg*