



CEFET/RJ

Data Mining Introduction



Eduardo Ogasawara

eduardo.ogasawara@cefet-rj.br

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

Outline / Roadmap

- Why Data Mining?
- What is Data Mining?
- KDD vs CRISP-DM
- Types of Data & Functions
- Evaluation and Applications
- Trends, Tools & Publishing

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



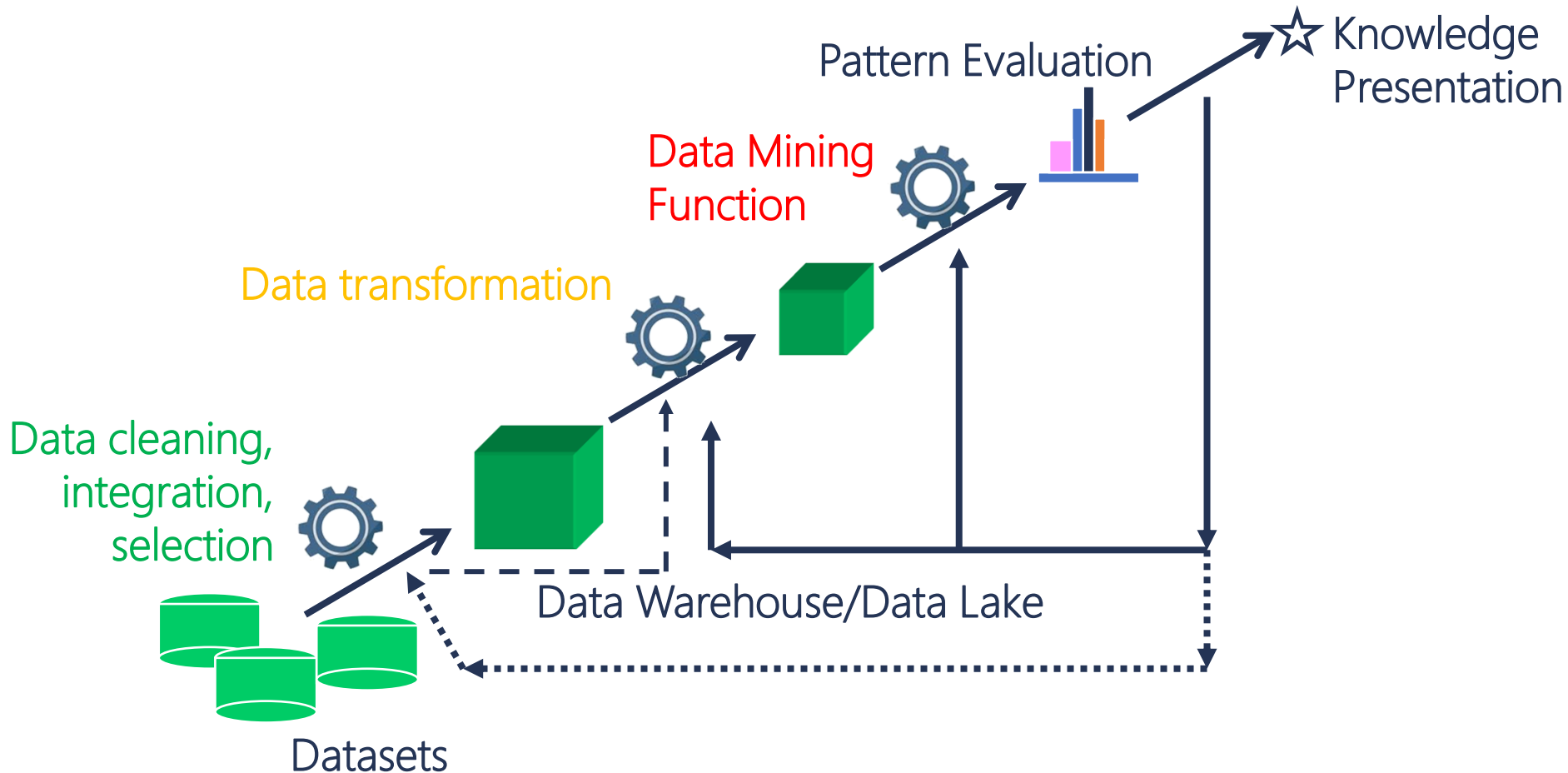
What is Data Mining?

- Data mining (knowledge discovery from data)
 - Extraction of interesting patterns or knowledge from a massive amount of data
 - non-trivial: not obvious
 - Implicit: hidden in data
 - previously unknown: not explicitly stored
 - potentially useful: actionable
- Alternative names
 - Knowledge Discovery in Databases (KDD)
 - Knowledge Extraction
 - Business Intelligence
 - Data Analysis



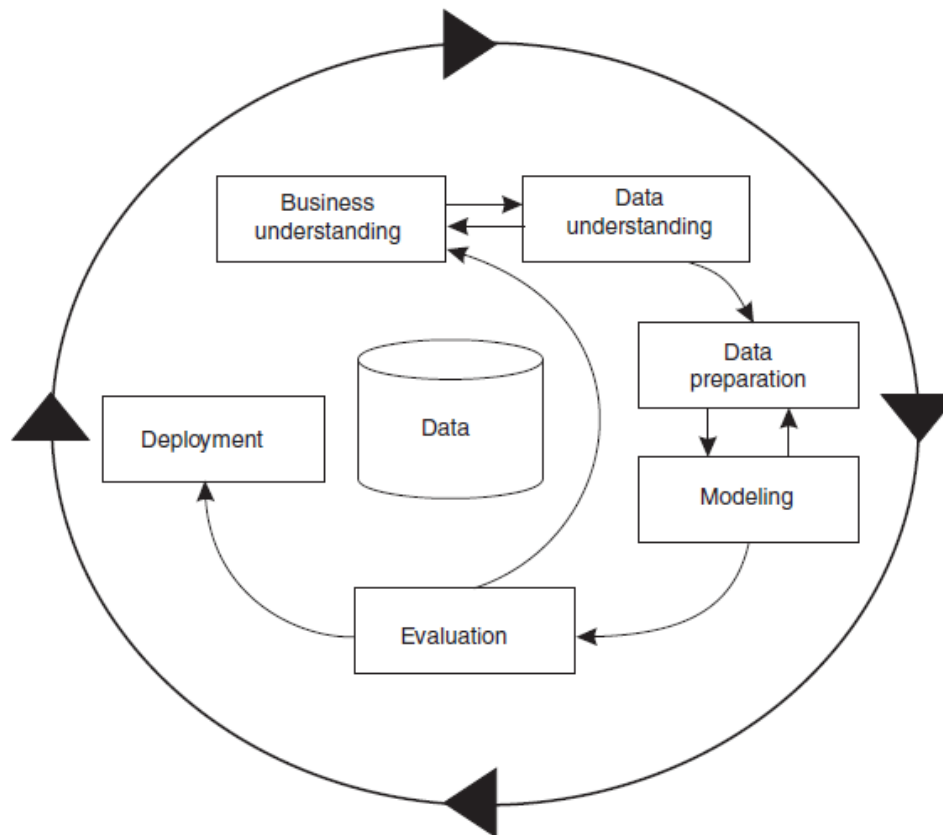
The KDD Process

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



CRISP-DM Overview

- CRISP-DM (Cross-Industry Standard Process for Data Mining)
- A widely used, industry-standard methodology for structured data mining projects

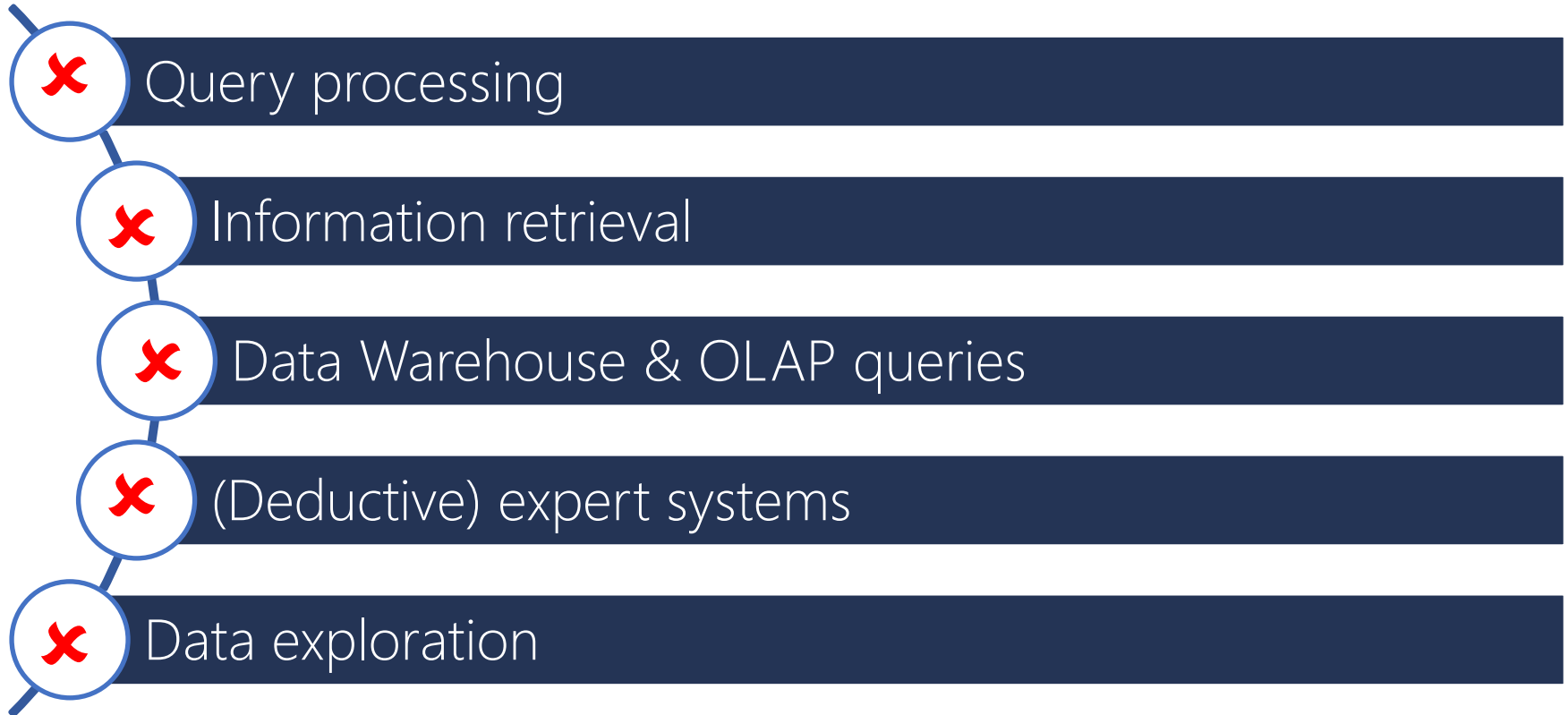


Knowledge discovery in databases (KDD) process vs. Cross-Industry Standard Process for Data Mining (CRISP-DM)

- KDD is the broader, conceptual process of discovering knowledge in data—data mining is one of its steps
- CRISP-DM provides a practical, cyclical guide for managing end-to-end data mining projects in real-world settings
- While KDD is more academic and linear, CRISP-DM is iterative and grounded in real applications

Feature	KDD	CRISP-DM
Origin	Academia	Industry
Process	Linear	Cyclical
Emphasis	Knowledge Extraction	Business Problem Solving

Is everything "data mining"?



Multi-Dimensional View of Data Mining

Data models

Knowledge to be mined

Data mining functions

Methods/Techniques used

Applications

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

Data Mining: on what kinds of data models?

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

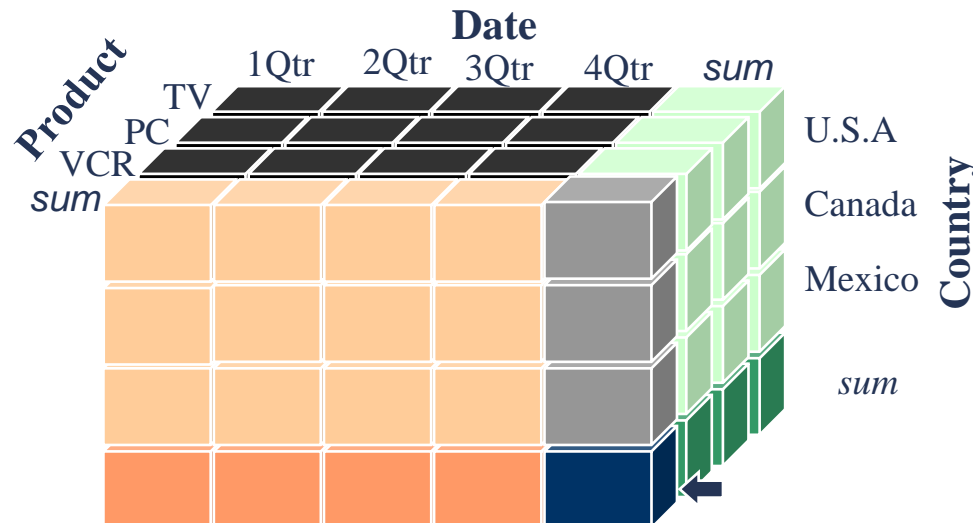
Core Data Mining Functions

1. Generalization
2. Association/Correlation
3. Prediction
4. Clustering
5. Outlier Detection
6. Event Detection
7. Sequential Pattern Mining
8. Structure & Network Analysis

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)



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 \rightarrow 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?

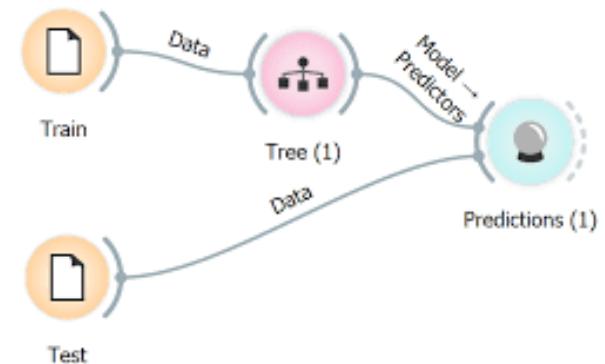
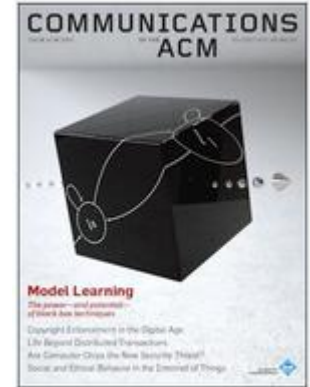
Rakesh Agrawal Tomasz Imielinski* Arun Swami
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120

[illegible]

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

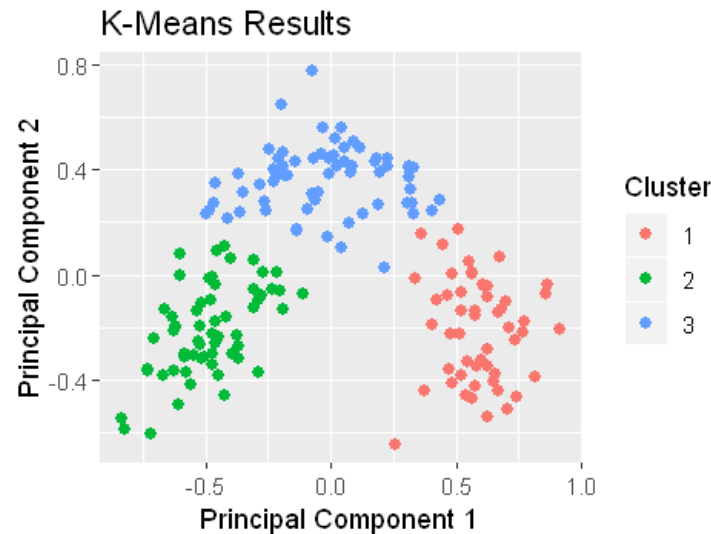


[1] F. Vaandrager, 2017, Model learning, *Communications of the ACM*, v. 60, n. 2, p. 86–95.

[2] <https://towardsdatascience.com/data-science-made-easy-data-modeling-and-prediction-using-orange-f451f17061fa>

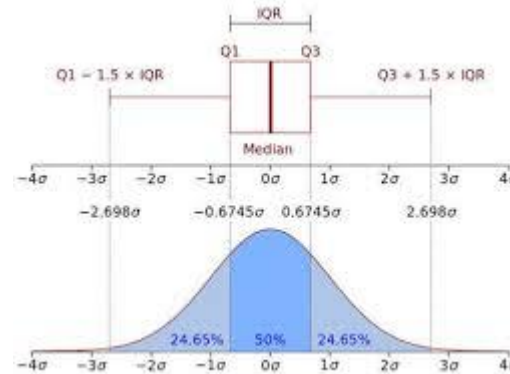
Data Mining Function: (4) Cluster Analysis

- 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

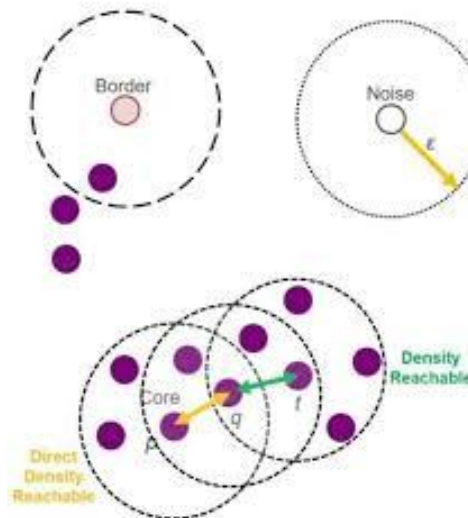


Data Mining Function: (5) Outlier Analysis

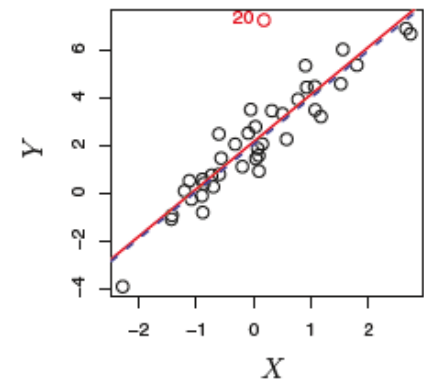
- Outlier: A data object that does not comply with the general behavior of the data
- Noise or exception? — One person's garbage could be another person's treasure
- Methods: statistical, clustering or regression analysis
- Useful in fraud detection, rare events analysis



(distribution)



(density)



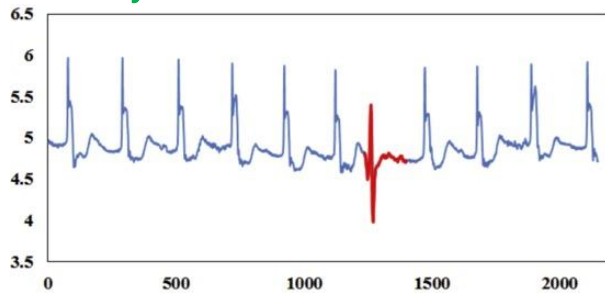
(model)

Data Mining Function:

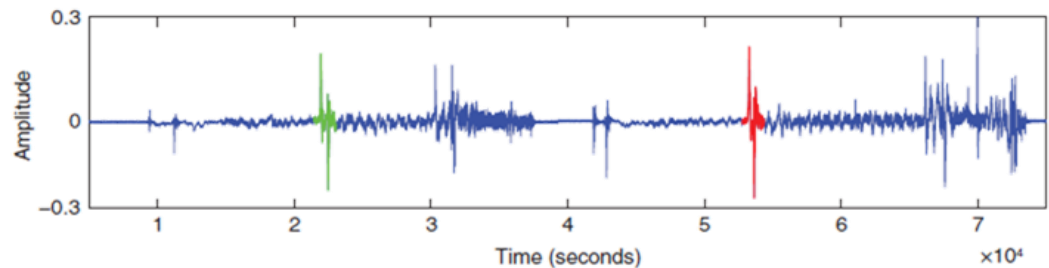
(6) Event detection

- 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

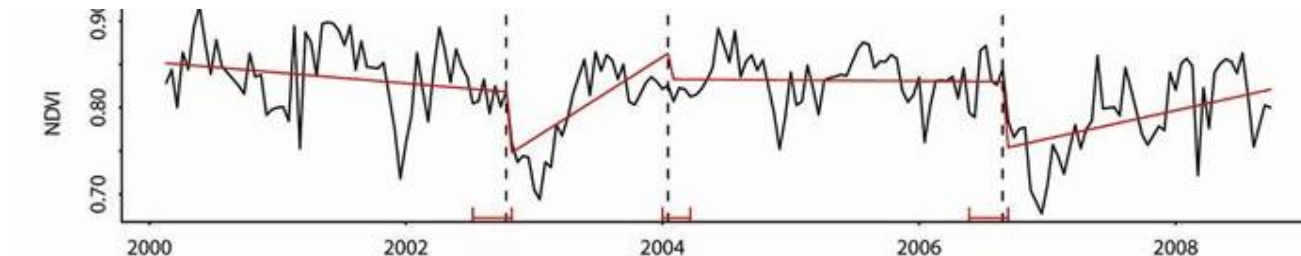
(anomaly)



(motif)



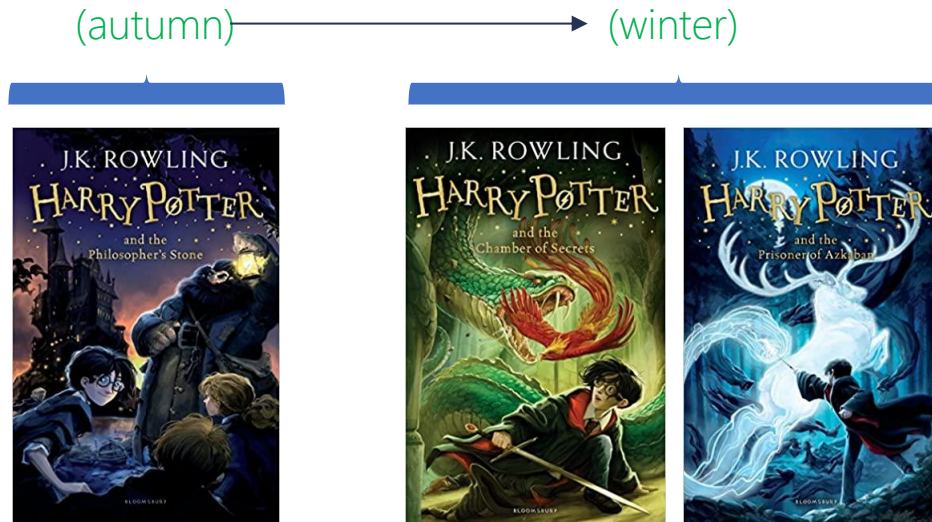
(change point)



[1]

Data Mining Function: (7) Sequential Pattern

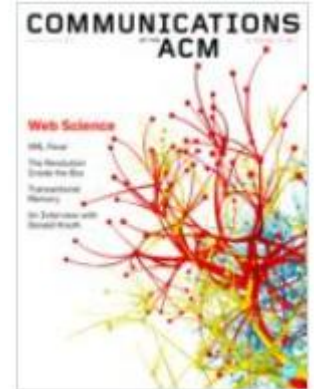
- Sequential pattern mining
 - Detect and analyze frequent subsequences of events, items, or tokens occurring in an ordered metric space



Data Mining Function:

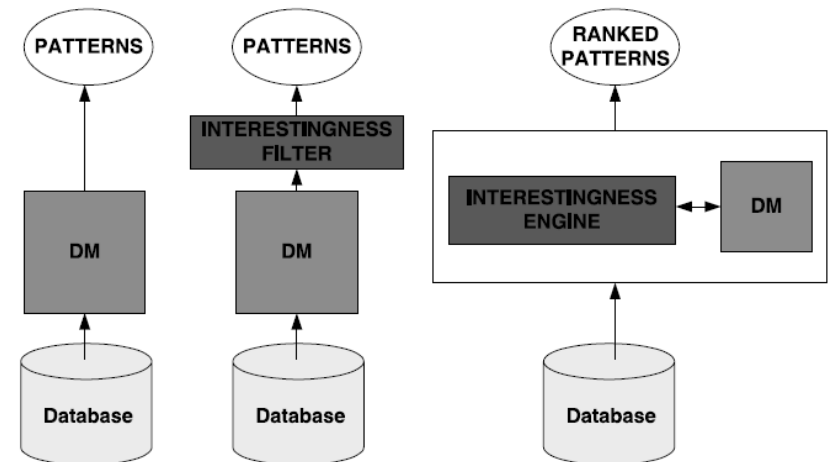
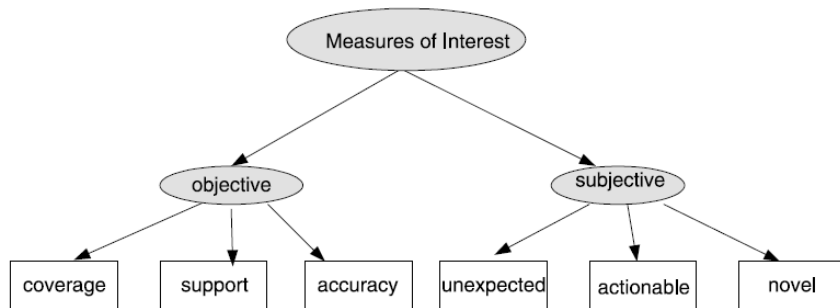
(8) Structure and Network Analysis

- Structure Mining
 - Extract patterns from structured data (e.g., XML trees, molecule graphs)
 - Applications: bioinformatics, chemistry, web page layouts
- Network Analysis
 - Mine nodes and edges in networks (e.g., social, citation, communication)
 - Tasks: community detection, centrality, influence, link prediction
- 🔍 Examples:
 - Web community detection
 - Finding similar proteins in biological networks
 - Discovering influencers in social media

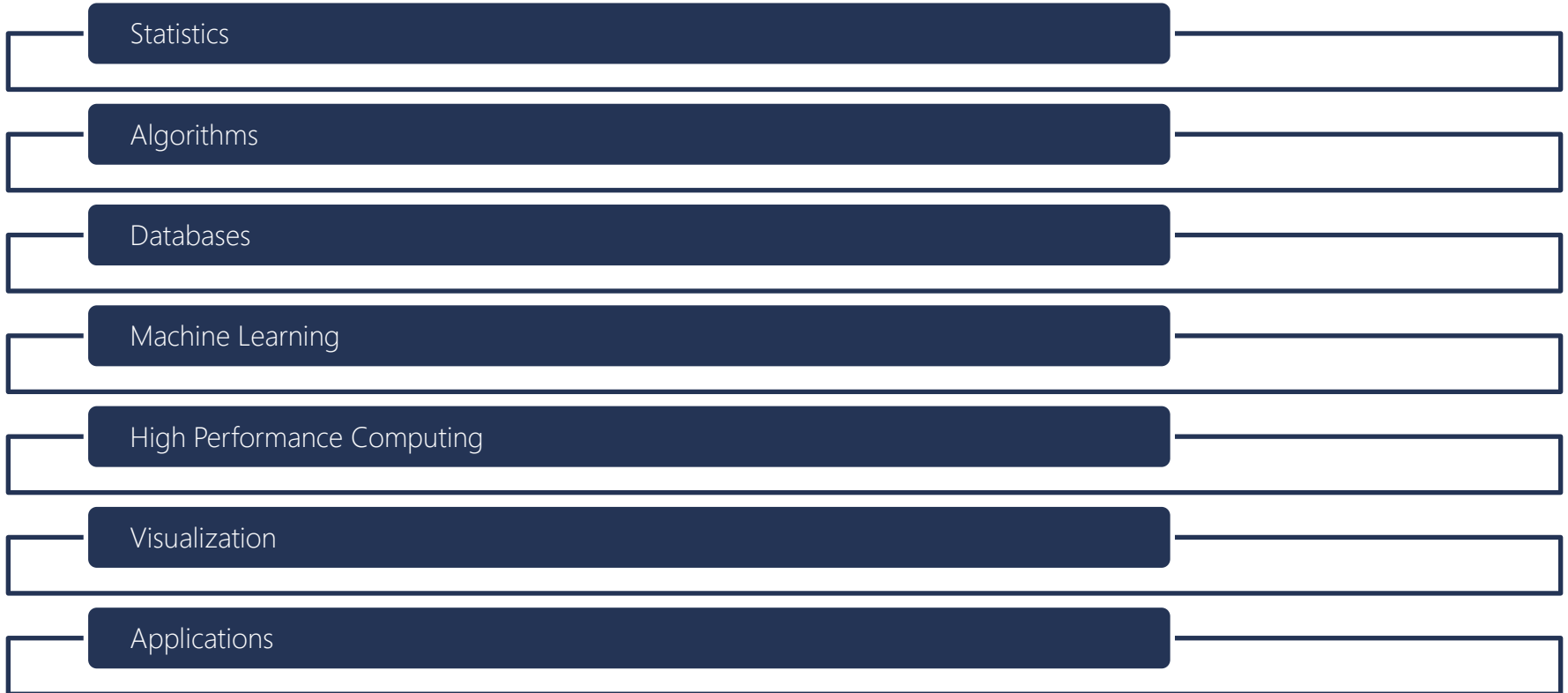


What Makes Patterns Interesting?

- Objective Measures: Support, confidence, lift, coverage, etc.
- Subjective Measures: Novelty, unexpectedness, usefulness, actionability
- Not all discovered patterns are useful
 - Evaluation helps focus on valuable insights
- 🧠 Interestingness helps separate noise from knowledge



Multidisciplinary Nature of Data Mining



Why So Many 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

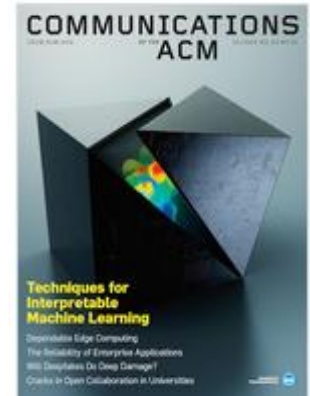
Availability of data

- 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

Major Issues in Data Mining

- Mining Methodology and User Interaction
 - Mining in multidimensional space
 - Mining knowledge at multiple abstraction levels
 - Incorporation of domain knowledge
 - Interactive and exploratory mining
- 2. Performance and Scalability
 - Efficiency and scalability of algorithms
 - Parallel, distributed, and incremental mining
- 3. Diversity of Data Types
 - Handling structured, semi-structured, unstructured, complex data (e.g., graphs, sequences, spatial, multimedia, streams)
- 4. Data Mining and Society
 - Data ownership and privacy
 - Security and data sensitivity
 - Ethical issues and social impact



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



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

Trending Data Mining Languages


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

Data Mining Tools

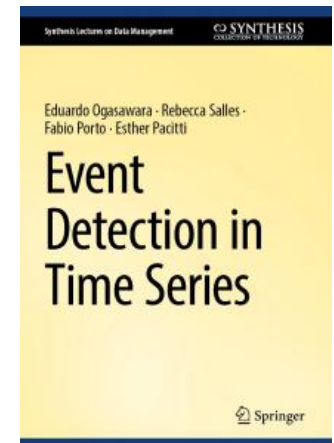
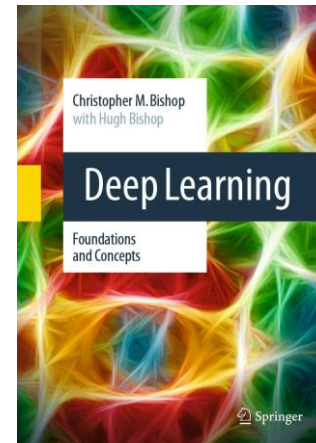
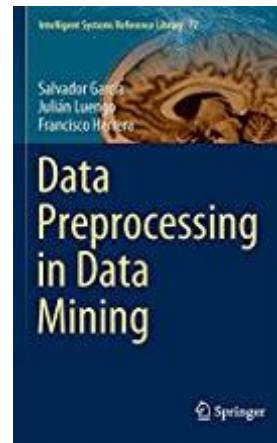
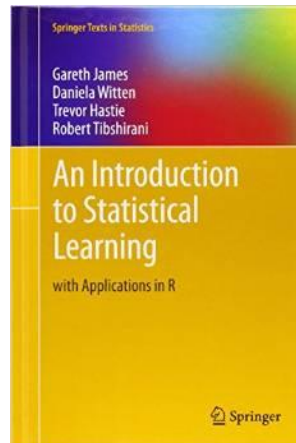
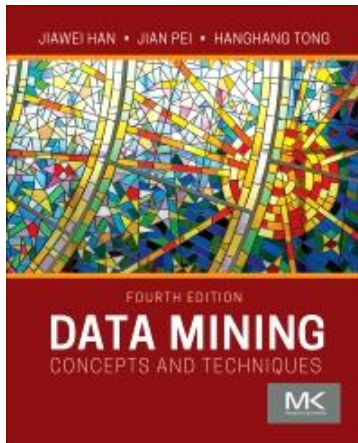
- Rapid Miner (open source)
 - <https://rapidminer.com>
- Orange (open source)
 - <https://orange.biolab.si>
- Weka (open source)
 - <https://www.cs.waikato.ac.nz/ml/weka>
- Knime (open source)
 - <https://www.knime.com>
- Apache Mahout (open source)
 - <http://mahout.apache.org>
- Rattle (open source)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



Key Takeaways

- Data mining is essential to transform vast data into knowledge
- It is part of a larger process (KDD or CRISP-DM)
- Various data types and functions are supported
- Challenges include scalability, complexity, ethics, and privacy
- Tools, languages, and communities support real-world application
-  "Data mining is not magic—it's a process built on sound principles."

Main References



- [1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.
- [2] G. M. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: With Applications in R. Springer Nature, 2021.
- [3] S. Garcia, J. Luengo, and F. Herrera, Data Preprocessing in Data Mining. Springer, 2014.
- [4] C. M. Bishop and H. Bishop, Deep Learning: Foundations and Concepts. Springer Nature, 2023.
- [5] E. Ogasawara, R. Salles, F. Porto, and E. Pacitti, Event Detection in Time Series, 1st ed. in Synthesis Lectures on Data Management. Cham: Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-75941-3.

Slides and videos at: <https://eic.cefet-rj.br/~eogasawara/data-mining/>

