PROGRAMA DE VERÃO DO LNCC

JORNADA DE CIÊNCIA DE DADOS

Introduction to Data Mining

CEFET/RJ

Eduardo Ogasawara

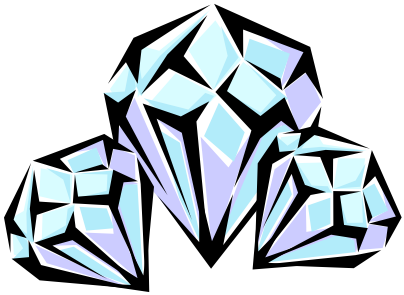
http://eic.cefet-rj.br/~eogasawara

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

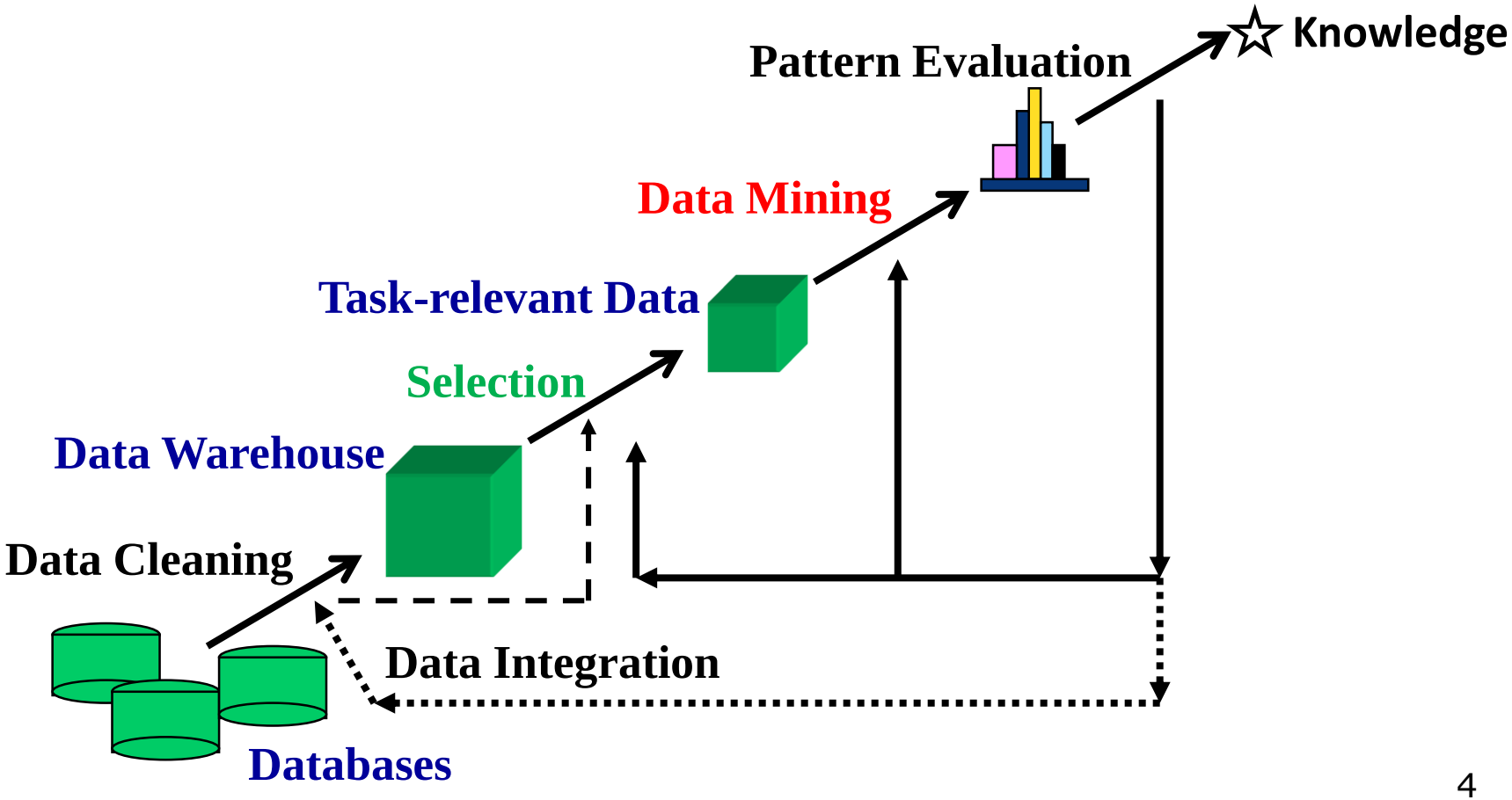
What is Data Mining?

- 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

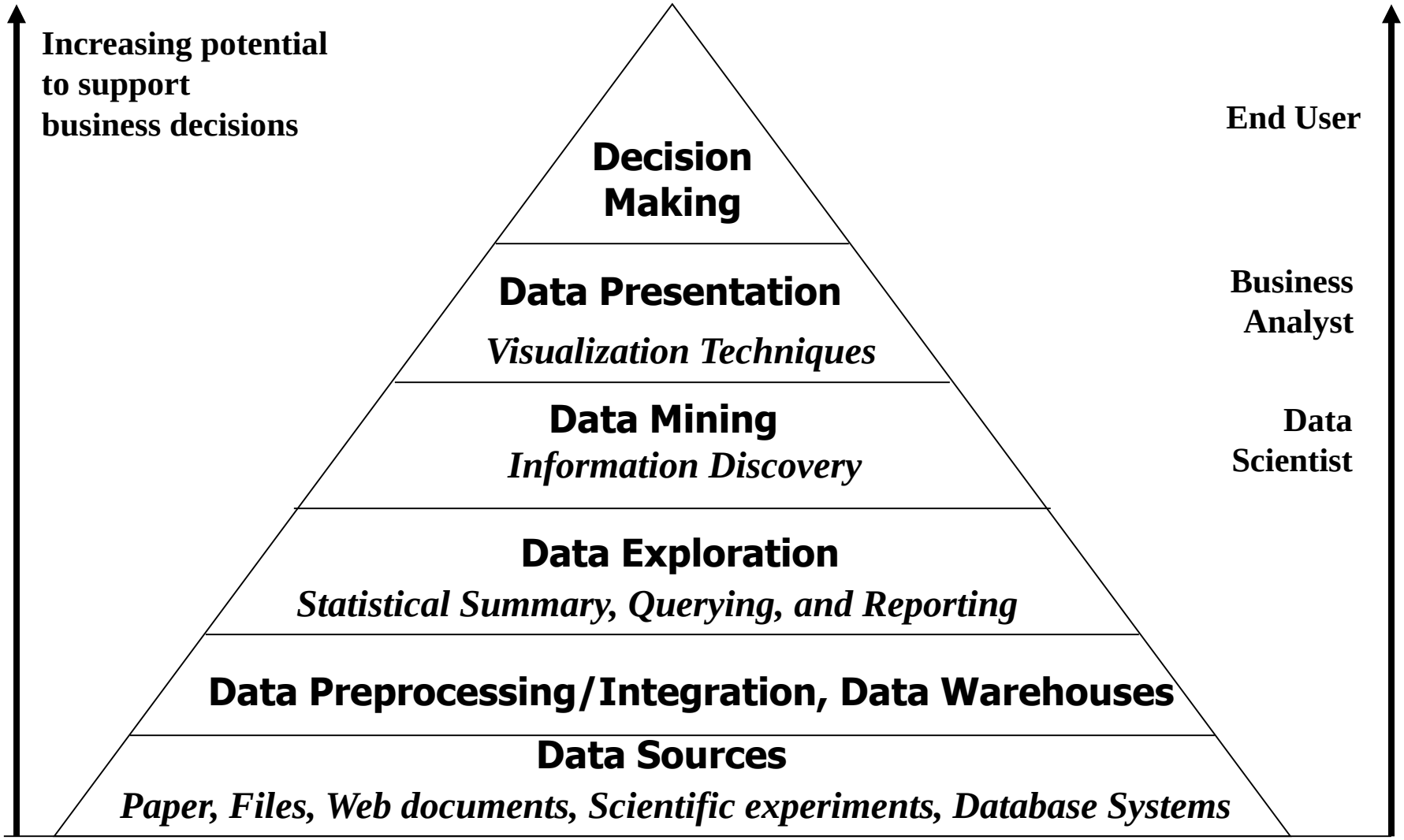


Knowledge discovery from data (KDD) process (Database perspective)

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

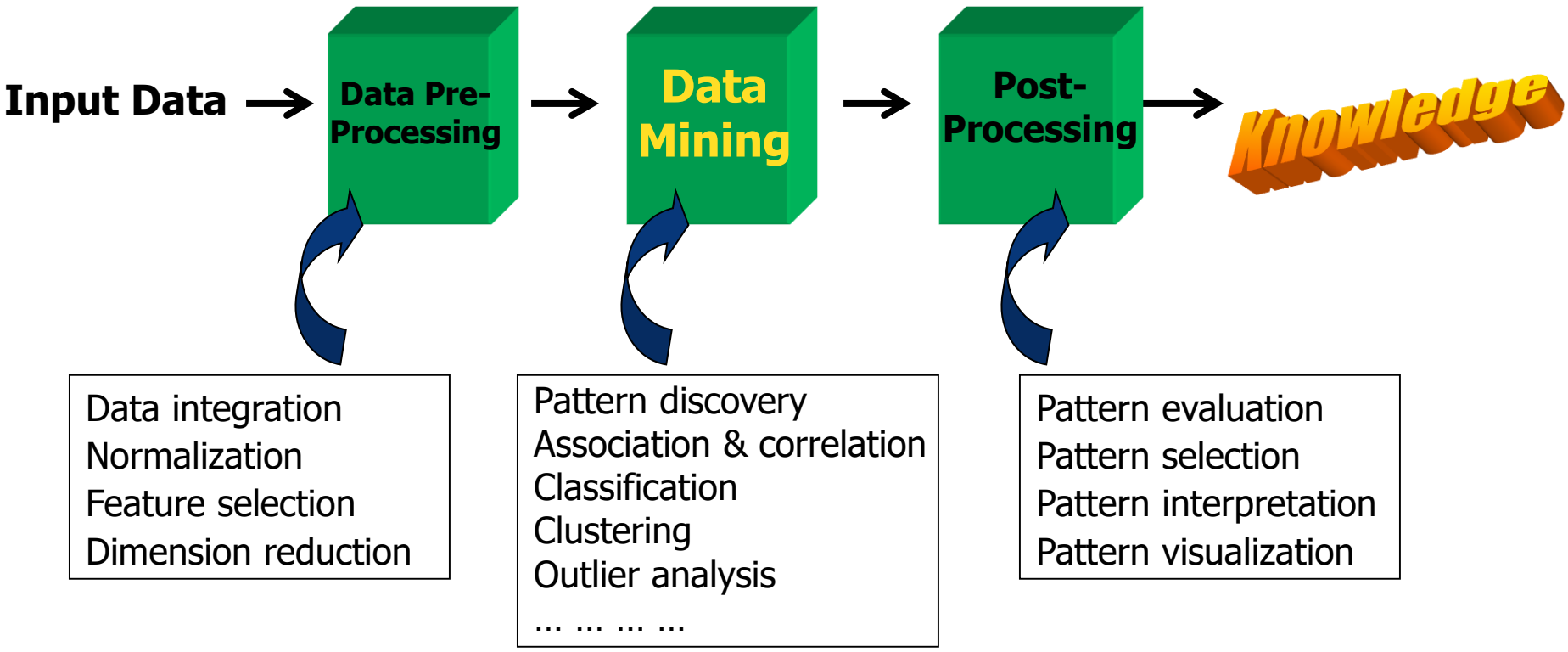


Data Mining (Business Intelligence perspective)



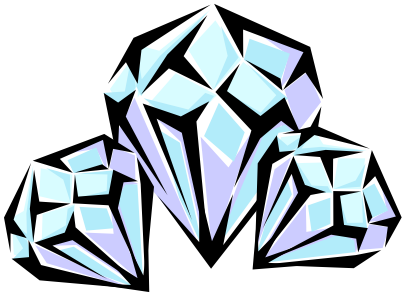
KDD process
Machine/Statistical Learning perspective

- This is a view from typical machine learning and statistics communities



Is everything "data mining"?

- Watch out:
  - ✘ Simple search and query processing
  - ✘ (Deductive) expert systems
  - ✘ Data Mining vs. data exploration
  - ✘ Mining vs. OLAP vs. presentation tools
  - ✘ Warehouse, data cube, reporting but not much mining



Multi-Dimensional View of Data Mining

- Data models to be mined
- Knowledge to be mined
  - Data mining functions
- 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
- 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

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
  - A typical association rule
    - 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?

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, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression
- Typical applications:
  - classifying stars, predicting delays, forecasting diseases

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 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: by-product of clustering or regression analysis
  - Useful in fraud detection, rare events analysis

Data Mining Function: (6) Sequential Pattern, Trend and Evolution Analysis

- Sequential pattern mining
  - e.g., buy a digital camera, then buy a large SD memory cards
- Time-series, trend, and deviation analysis:
  - e.g., regression and value prediction
- Periodicity analysis
- Motifs and biological sequence analysis
  - Approximate and consecutive motifs
- Similarity-based analysis

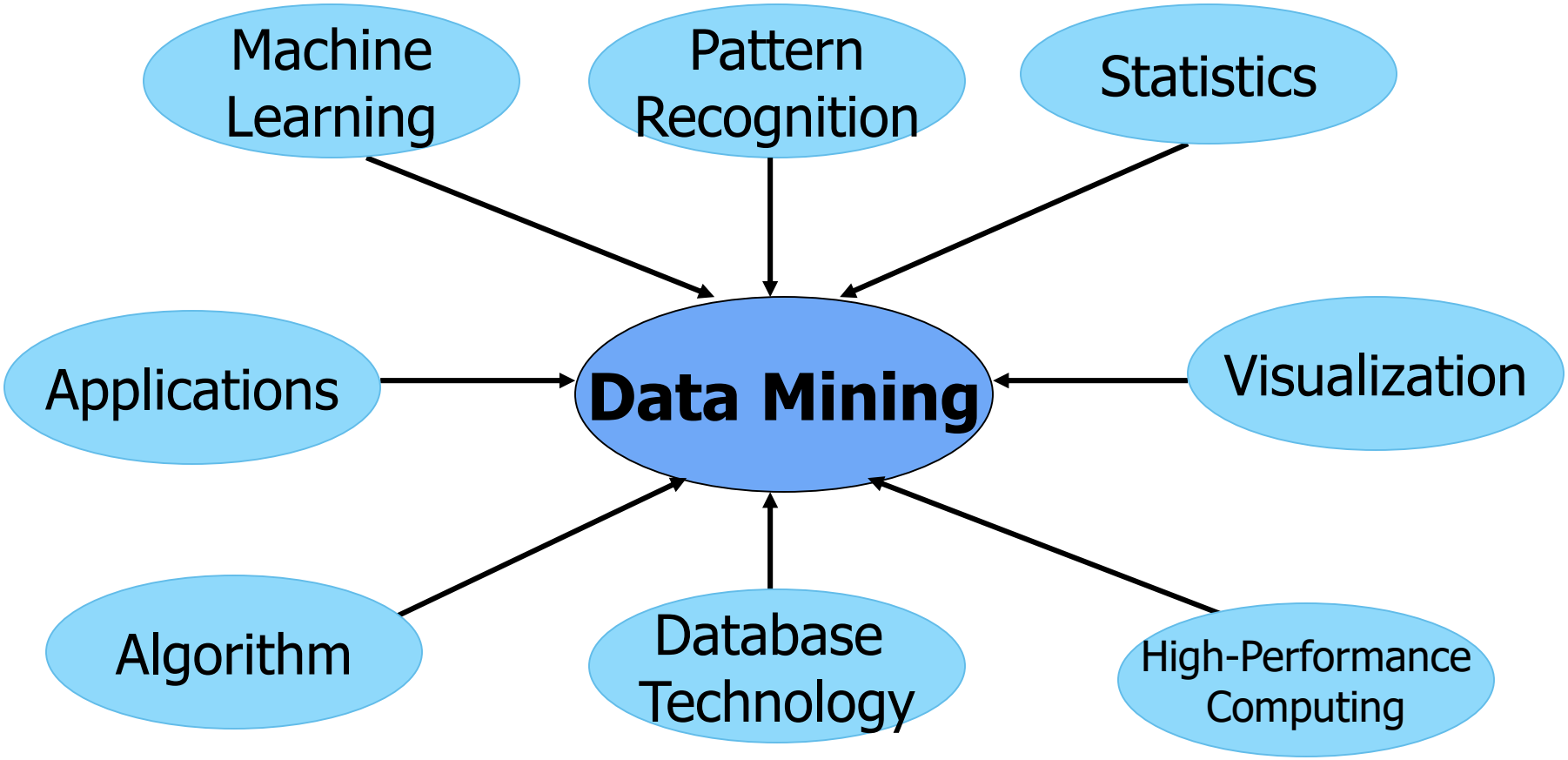
Data Mining Function: (7) Structure and Network Analysis

- 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

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?
  - Descriptive vs. Predictive vs. Prescriptive
  - Accuracy
  - Coverage
  - Typicality vs. novelty vs. rarity

Data Mining: Confluence of Multiple Disciplines



Why Confluence of Multiple Disciplines?

- Tremendous amount of data
  - Algorithms must be scalable to handle big data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- 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 enough to be considered data mining?

Applications of Data Mining

- Consider all data mining functions
- The question might be:
  - What areas are difficult to do Data Mining?

Major Issues in Data Mining (1)

- Mining Methodology
  - Mining knowledge in multi-dimensional space
  - Boosting the power of discovery in a networked environment
  - 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

Major Issues in Data Mining (2)

- 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

Where to publish?

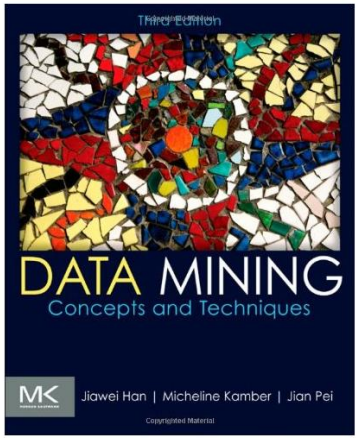
- Data mining and KDD
  - Conferences: SIGKDD, IEEE-ICDM, SIAM-DM, PKDD.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems
  - Conferences: SIGMOD, PODS, VLDB, IEEE-ICDE, EDBT, ICDT, SSDBM
  - Journals: IEEE-TKDE, VLDB J., Info. Sys.
- AI & Machine Learning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

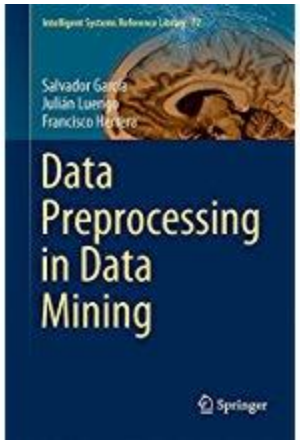
Consider domain-driven targets

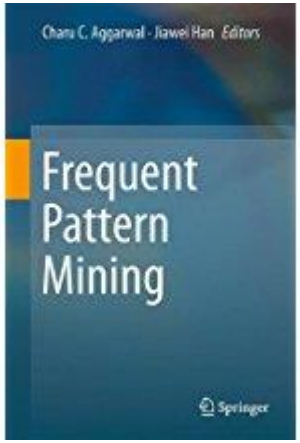
Trending Data Mining Languages/Frameworks

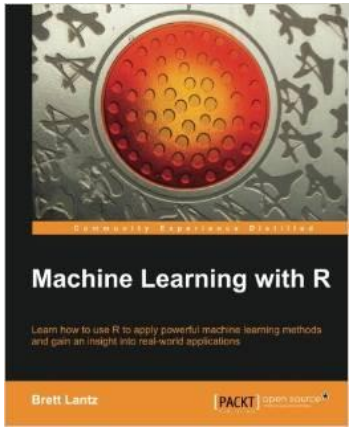
- Python (Machine Learning Course)
- R (Data Mining Course)
- Spark (Parallel and Distributed Computing)

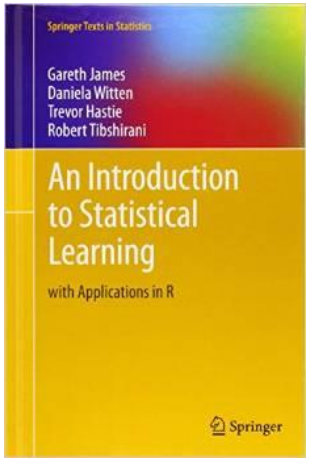
Main References

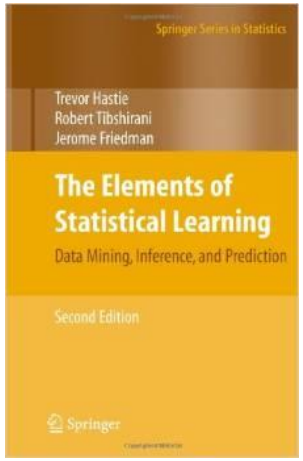


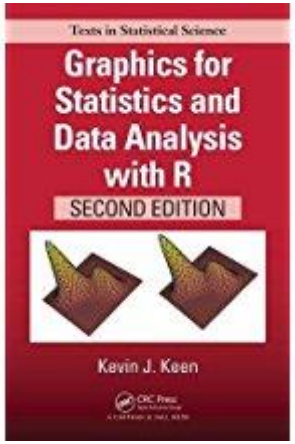


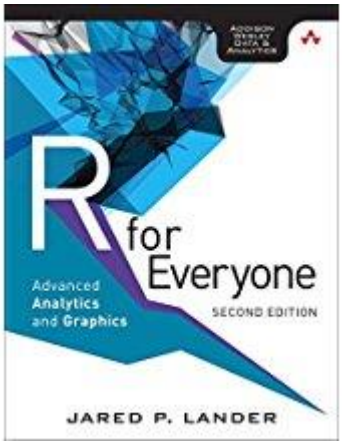












Most of the slides were extracted from
Data Mining Concepts and Techniques