DATA MINING

Lecture 1: Introduction

Dr. Doaa Elzanfaly

Course Info.

- **Lectures:** Monday – 10:00-12:00 / 12:00-2:00
- **Instructor:** Dr. Doaa Elzanfaly
  - email: doaa.saad@fci.helwan.edu.eg
  - Contact: Teams
- **Textbook:** Jiawei Han, Micheline Kamber, and Jian Pei. 2011. Data Mining: Concepts and Techniques (3rd. ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- **Reference book:** Charu C. Aggarwal. 2015. Data Mining: The Textbook. Springer Publishing Company, Incorporated.

Course Objectives

- To introduce students to various data mining concepts and technologies.
  - Understanding the foundation concepts of data mining.
  - Exploring algorithms commonly used in data mining tools.
  - Ability to apply data mining tools to real-world problems.

Tentative Syllabus

Week	Lecture	Lab
1	Introduction	Environment Installation Data Mining Introduction
2	Getting to Know Your Data	Data Preprocessing Techniques
3	Data Preprocessing I	Data Preprocessing Techniques Continued.
4	Data Preprocessing II	Data Visualization
5	Association Analysis I	Apriori Algorithm Implementation
6	Association Analysis II	A Frequent Pattern Growth Approach
7	Midterm Exam	
8	Classification	Classification Algorithm Implementation
9	Clustering	Regression Algorithm
10	Outlier Detection	Outlier Detection

Assessment Scheme

Midterm:	30 marks
Lab Assignments:	10 marks
Practical Exam:	10 marks
Final Exam:	50 marks
Bonus points:	5 marks - Based on participation.

What I expect from you …

- Attend the lectures and lab regularly. (70% to pass)
- Study and learn the material presented in the teaching sessions *and in the textbook.*
- Perform well in the exams.
- Don’t cheat.

Lecture Outline

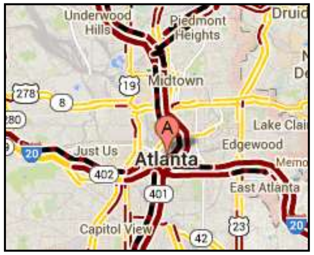
- Why Data Mining?
- What Is Data Mining?
- What Kind of Data Can Be Mined?
- Data Mining Tasks
- What Kind of Applications Are Targeted?
- Major Issues in Data Mining

Large-scale Data is Everywhere!

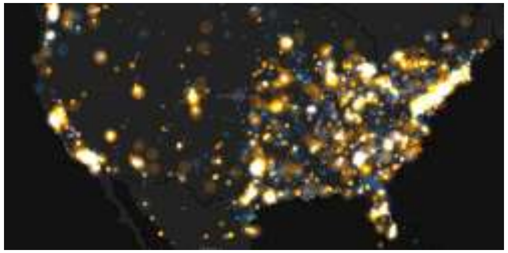
- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- New mantra
  - Gather whatever data you can whenever and wherever possible.
- Expectations
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.

Cyber Security

E-Commerce

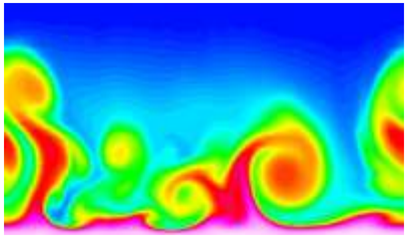


Traffic Patterns



Social Networking: Twitter

Sensor Networks

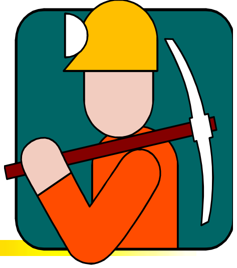


Computational Simulations

Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data but starving for knowledge!
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

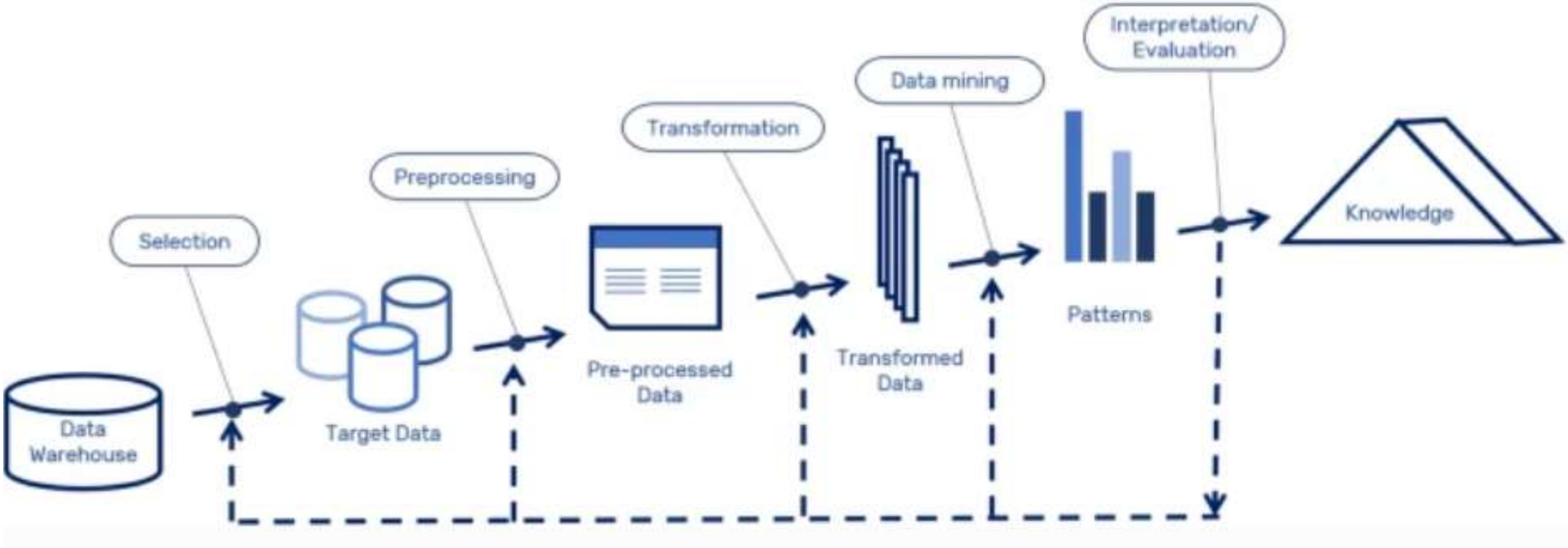
What Is Data Mining?



- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data to predict future trends.
- Alternative names
  - Knowledge Discovery (mining) in Databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

What is the difference between mining and Querying??

Knowledge Discovery (KDD) Process



https://link.springer.com/article/10.1007/s42979-020-0117-6

What Kind of Data Can Be Mined?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- 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 multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

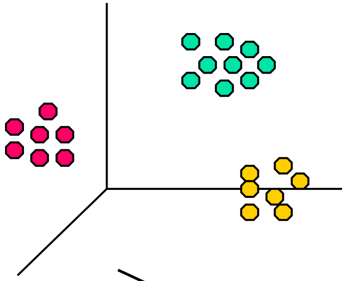
Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.

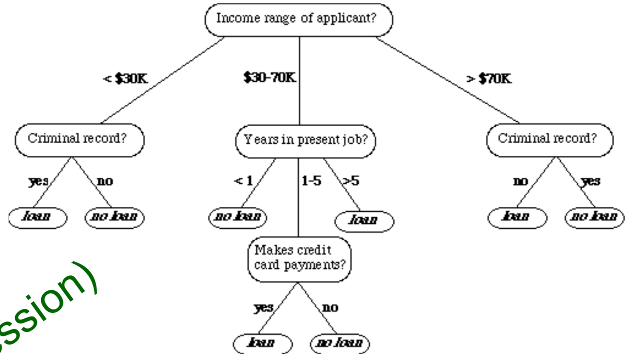
- Description Methods
  - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Data Mining Tasks



Clustering



Predictive Modeling

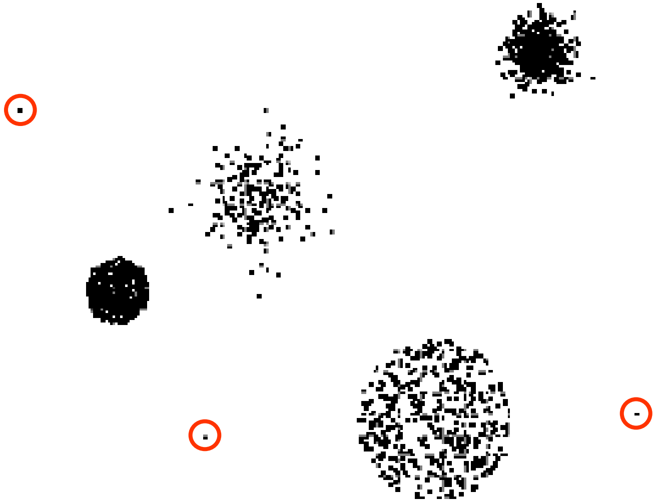
(Classification & Regression)

Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Association Rules

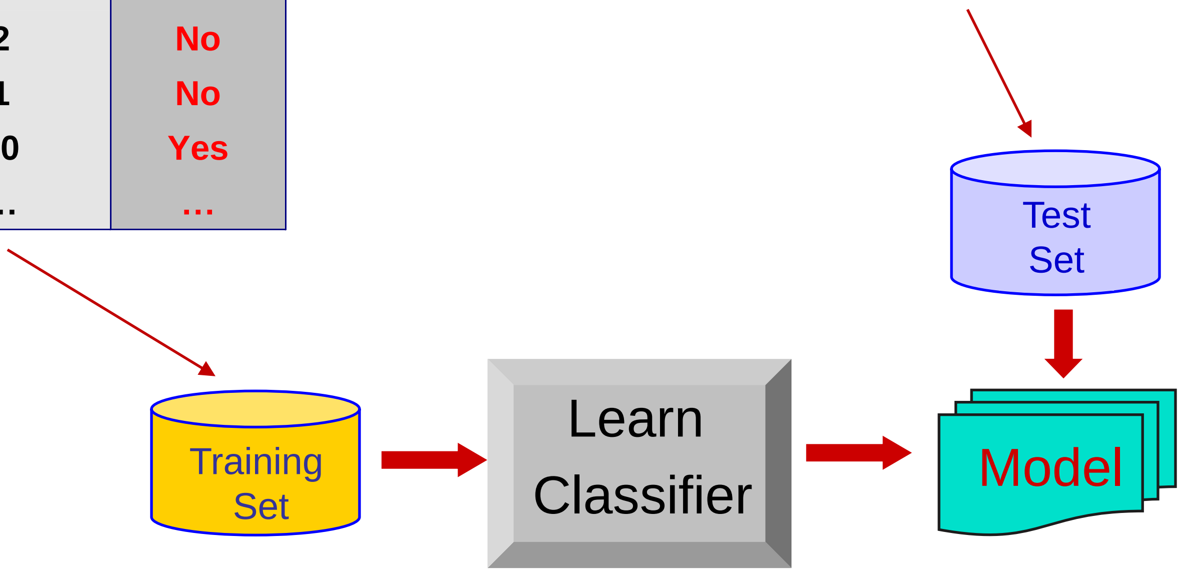
Anomaly Detection



Predictive Modeling: Classification

	categorical	categorical	quantitative	class
Tid	**Employed**	**Level of Education**	**# years at present address**	**Credit Worthy**
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	**Employed**	**Level of Education**	**# years at present address**	**Credit Worthy**
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…

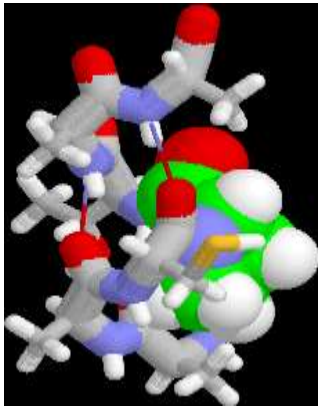


Classification

- Classification and label prediction – Supervised Learning
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …

Examples of Classification Task

- Classifying credit card transactions as legitimate or fraudulent.
- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data
- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace
- Predicting tumor cells as benign or malignant
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil

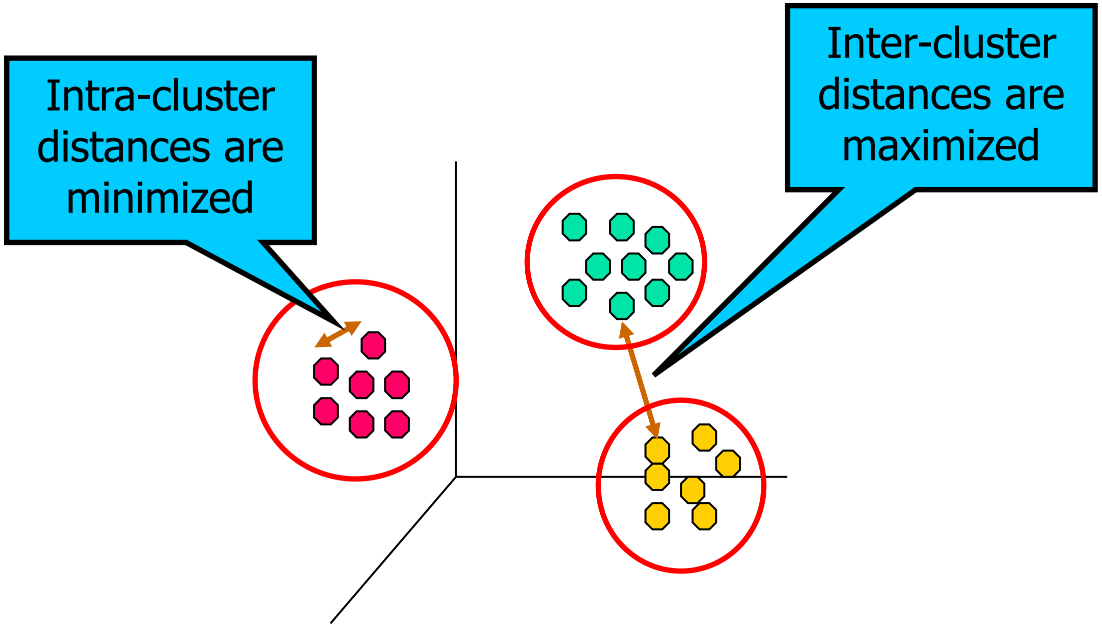


Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Extensively studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups
- Unsupervised learning (i.e., Class label is unknown)
- Principle: Maximizing intra-class similarity & minimizing interclass similarity



Intra-cluster distances are minimized

Inter-cluster distances are maximized

Applications of Cluster Analysis

- **Market Segmentation:**
  - **Goal:** subdivide a market into distinct subsets of customers where any subset may possibly be selected as a market target to be reached with a distinct marketing mix.
  - **Approach:**
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Applications of Cluster Analysis

- **Document Clustering:**
  - **Goal:** To find groups of documents that are similar to each other based on the important terms appearing in them.
  - **Approach:**
    - To identify frequently occurring terms in each document.
    - Form a similarity measure based on the frequencies of different terms and use it to cluster.

Association and Correlation Analysis

- Given a set of records each of which contain some number of items from a given collection
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	Bread, Coke, Milk
2	Tea, Bread
3	Tea, Coke, Sugar, Milk
4	Tea, Bread, Sugar, Milk
5	Coke, Sugar, Milk

Rules Discovered:
{Milk} --> {Coke}
{Sugar, Milk} --> {Tea}

Association and Correlation Analysis

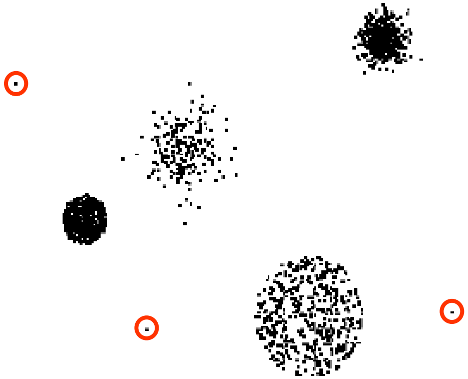
- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together?
- Association, correlation vs. causality
  - A typical association rule
    - Tea → Sugar [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?
- How to mine such patterns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

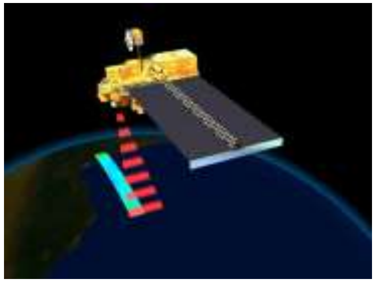
Applications of Association Analysis

- Market-basket analysis
  - Rules are used for sales promotion, shelf management, and inventory management
- Telecommunication alarm diagnosis
  - Rules are used to find combination of alarms that occur together frequently in the same time period
- Medical Informatics
  - Rules are used to find combination of patient symptoms and test results associated with certain diseases

Deviation/Anomaly/Change Detection

- Also known as outlier analysis
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Identify anomalous behavior from sensor networks for monitoring and surveillance.
  - Detecting changes in the global forest cover.

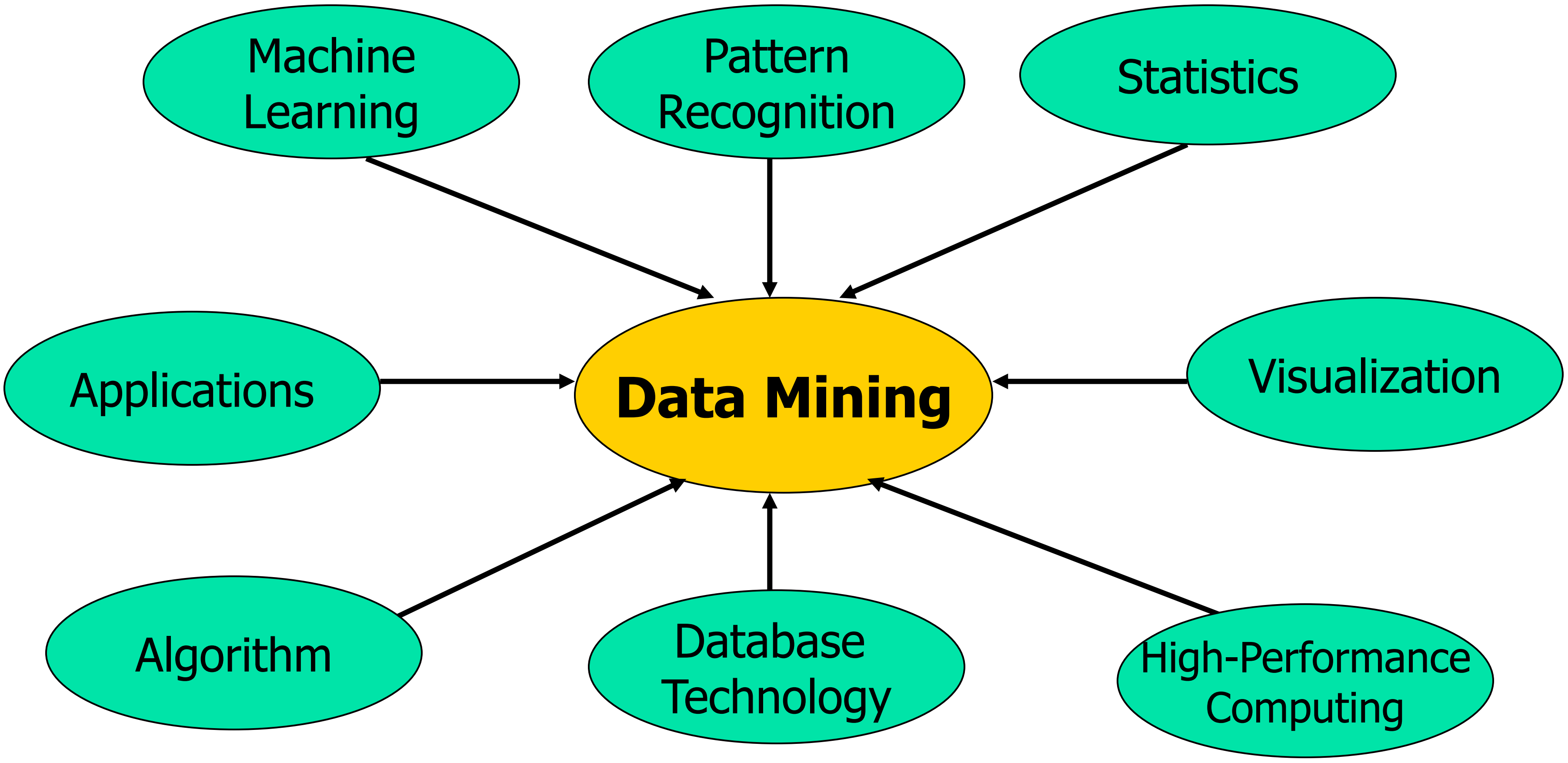




Outlier Analysis

- Outlier analysis
  - Outlier: A data object that does not comply with the general behavior of the data.
  - Outlier analysis is to detect significant deviations from normal behavior
    - 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, Network Intrusion Detection, Identify anomalous behavior from sensor networks for monitoring and surveillance

Data Mining: Confluence of Multiple Disciplines



Major Issues in Data Mining

- 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
  - Invisible data mining

Where to Find References? DBLP, CiteSeer, Google

- Data mining and KDD (SIGKDD: CDROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- Web and IR
  - Conferences: SIGIR, WWW, CIKM, etc.
  - Journals: WWW: Internet and Web Information Systems,
- Statistics
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization
  - Conference proceedings: CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

Recommended Reference Books

- S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data. Morgan Kaufmann, 2002
- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2011
- D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer-Verlag, 2009
- B. Liu, Web Data Mining, Springer 2006.
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005
- S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005