داده‌کاوی محاسباتی
Computational Data Mining

محمدرضا اصغری اسکوئی
استادیار دانشکده علوم ریاضی و رایانه
دانشگاه علامه طباطبائی

Mohammadreza Asghari Oskoei
Allameh Tabataba’i University (ATU)
2018-19

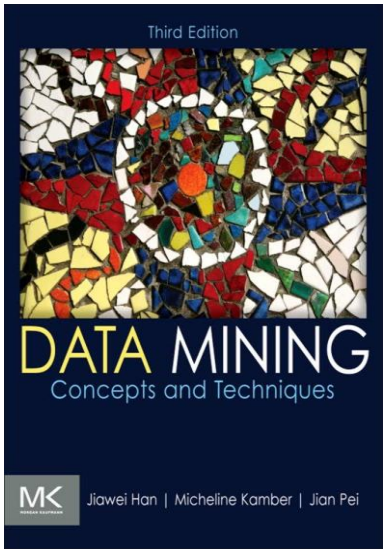
معرفی دوره آموزشی داده کاوی

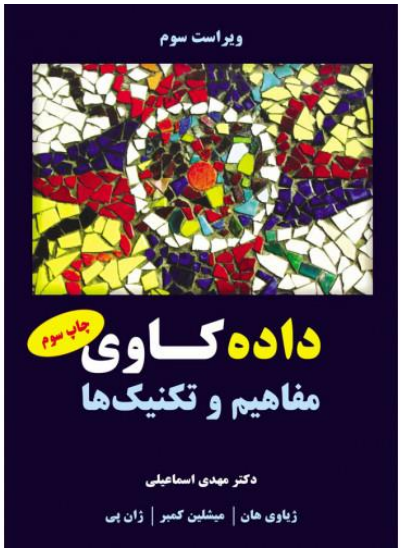
- عنوان: دوره آموزشی داده کاوی (۱)
- سرفصل و عناوین:
  - مقدمه بر داده کاوی، یادگیری ماشین، هوش مصنوعی
  - آشنائی با زبان پایتون، ساختارهای داده و کتابخانه‌های NumPy, Matlibplot
  - روش‌های آماده سازی، پیش‌پردازش داده و نمایش داده
  - روش‌های خوشه‌بندی داده و ارزیابی آن با استفاده از کتابخانه SciKit-Learn
- مدرس: دکتر محمدرضا اصغری اسکوئی (استادیار دانشگاه علامه طباطبائی)
- زمان: روزهای چهارشنبه ساعت ۱۳/۳۰ الی ۱۵/۳۰
- مکان: پژوهشکده بیمه

مرجع درس

- J. Han, M. Kamber, and Jian Pei, *Data Mining: Concepts and Techniques*, Morgan, Kaufmann, 2012.

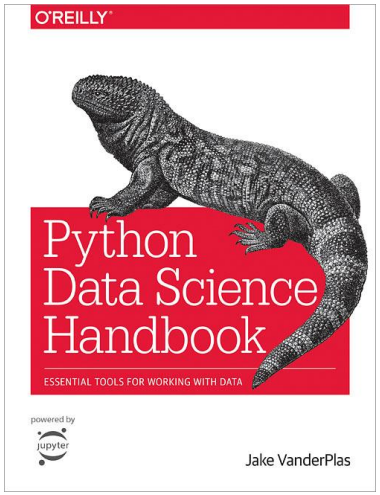
- داده‌کاوی مفاهیم و تکنیک‌ها، ترجمه مهدی اسماعیلی، انتشارات نیاز دانش، ۱۳۹۳ (در کتابخانه دانشکده و کتابفروشیها موجود است)





منابع کمکی

- *VanderPlas, Jake. Python data science handbook: essential tools for working with data. " O'Reilly Media, Inc.", 2016.*



- Some notes on Linear Algebra, Mark Schmidt, September 10, 2009
- *QR Matrix Factorization, The Singular Value Decomposition (SVD)*

سرفصل‌های کتاب داده کاوی مفاهیم و تکنیک ها

- مقدمه‌ای بر داده کاوی
- پیش‌پردازش داده‌ها
- انبار داده و تکنولوژیOLAP (پردازش تحلیلی برخط)
- فناوری مکعب داده و تعمیم داده ها
- کاوش الگوهای متوالی، وابستگی‌ها و همبستگی‌ها
- دسته‌بندی و پیش بینی
- خوشه‌بندی یا تحلیل خوشه‌ای
- کاوش جریان داده، سری های زمانی و داده های متوالی
- کاوش نموداری، تحلیل شبکه های اجتماعی و داده کاوی چند بعدی
- کاوش داده های فضایی، چند رسانه ای، متنی و داده های وب
- کاربردها و روندهای موجود در داده کاوی

مجموعه اسلایدهای کتاب داده کاوی مفاهیم و تکنیک ها

- Chapter 1. Introduction
- Chapter 2. Know Your Data
- Chapter 3. Data Preprocessing
- Chapter 4. Data Warehousing and On-Line Analytical Processing
- Chapter 5. Data Cube Technology
- Chapter 6. Mining Frequent Patterns, Associations and Correlations
- Chapter 7. Advanced Frequent Pattern Mining
- Chapter 8. Classification: Basic Concepts
- Chapter 9. Classification: Advanced Methods
- Chapter 10. Cluster Analysis: Basic Concepts and Methods
- Chapter 11. Cluster Analysis: Advanced Methods
- Chapter 12. Outlier Detection
- Chapter 13. Trends and Research Frontiers in Data Mining

سرفصل‌های دوره آشنائی با داده کاوی

- مقدمه بر علم داده کاوی
- آشنائی با انواع داده و ساختارهای داده
- آماده سازی و پیش پردازش داده ها
- برنامه سازی پایتون و بسته‌های محاسباتی Numpy و Matlibplot و Pandas
- فضای ویژگی‌ها و کاهش بعد فضای ویژگی
- آشنائی با خوشه‌بندی، رگرسیون و طبقه‌بندی
- خوشه بندی پیشرفته

مقدمه بر داده کاوی

مقدمه بر داده کاوی

- انگیزه و دلیل داده کاوی
- تعریف دقیق داده کاوی
- داده کاوی از ابعاد مختلف
- انواع الگوهای قابل کاوش
- فناوری مورد استفاده در داده کاوی
- نمونه های از کاربرد داده کاوی
- چالش های اصلی در داده کاوی
- تاریخچه داده کاوی

انگیزه و دلیل داده کاوی

- رشد انفجاری داده ها از ابعاد ترابایت تا پتابایت
  - حجم گسترده (Volume)
  - سرعت زیاد تولید داده (Velocity)
  - تنوع زیاد داده (Variety)
- روند مستمر تولید و ثبت داده: بانک های اطلاعاتی، حسگرها، وب، فرایندهای رایانه ای
  - کسب وکارها، تراکنش ها، وب، سهام، بیمه و حوادث
  - علوم: سنجش از راه دور، بیوانفورماتیک، شبیه سازی ها
  - افراد و جامعه، اخبار، تصاویر، اصوات، یوتیوب، ...
- در اقیانوسی از داده غرق هستیم اما نیازمند اطلاعات و دانش مفید و کاربردی
- نیاز دلیل اختراع است: داده کاوی: تحلیل داده و استخراج دانش

تعریف داده کاوی

- داده کاوی: کشف دانش از داده
- استخراج الگو یا دانش ارزشمند، ضمنی، ناشناخته، ناپیدا و احیانا مفید

- داده، اطلاعات، نظریه یا مدل، دانش

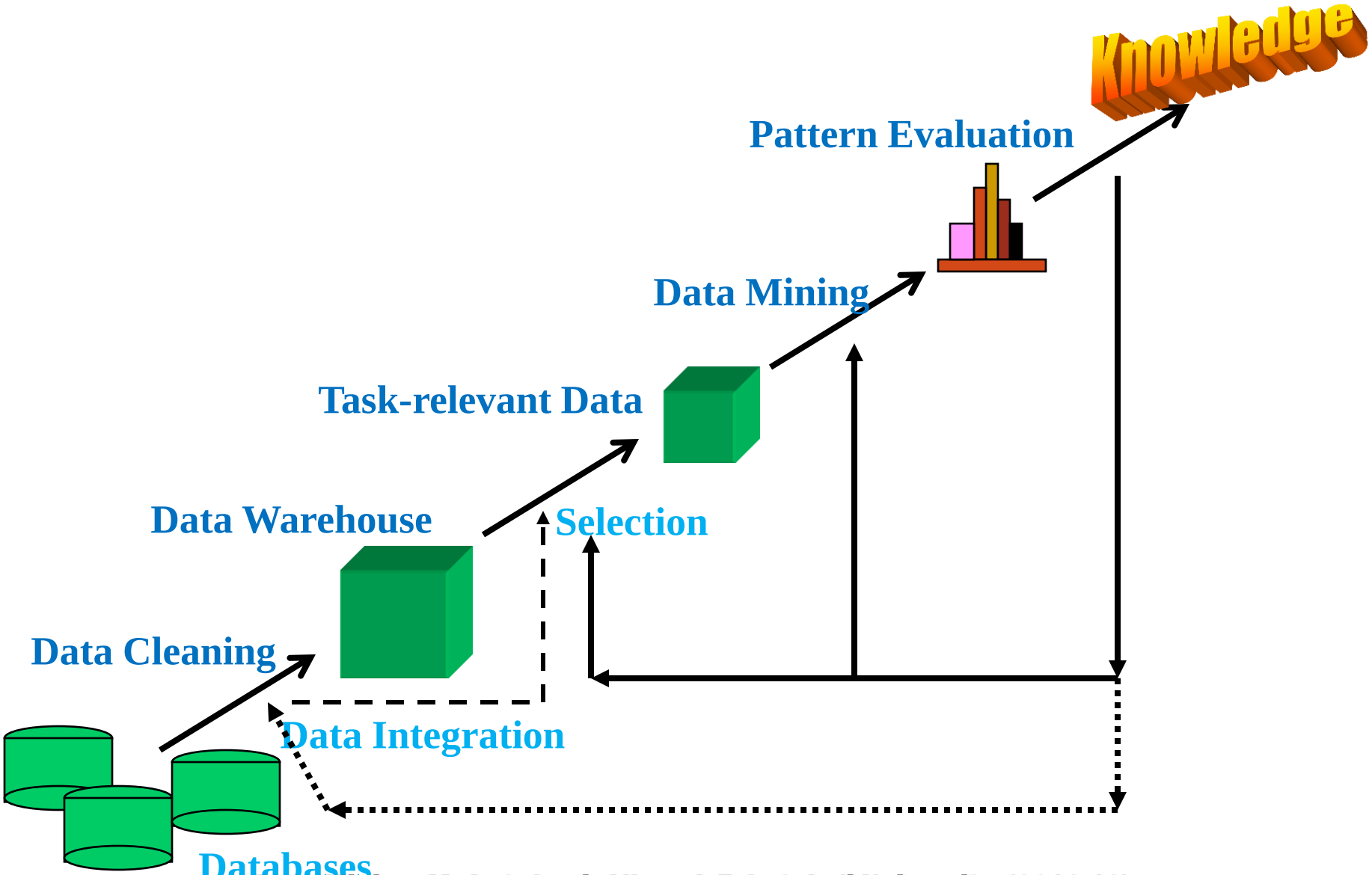
- کشف دانش، استخراج دانش، شناخت الگو، معرفت شناسی، کسب و کار هوشمند
- گزارش گیری یا استنتاج منطقی لزوما به معنی داده کاوی نیست

سیر تحول علوم داده

- علوم تجربی Before 1600, empirical science
- علوم نظری 1600-1950s, theoretical science
- علوم محاسباتی 1950s-1990s, computational science
- علوم داده 1990-now, data science

- ثبت داده، بانک اطلاعاتی، سامانه های مدیریت اطلاعات 1960s
- بانک اطلاعات رابطه ای، 1970s
- بانک های اطلاعاتی پیشرفته، شئ گرا، استنتاجی و کاربرد محور 1980s
- داده کاوی، انباره های بزرگ داده، چندرسانه و وب 1990s
- داده کاوی جریان داده های برخط، داده‌های برچسب‌دار و بدون ساختار 2000s

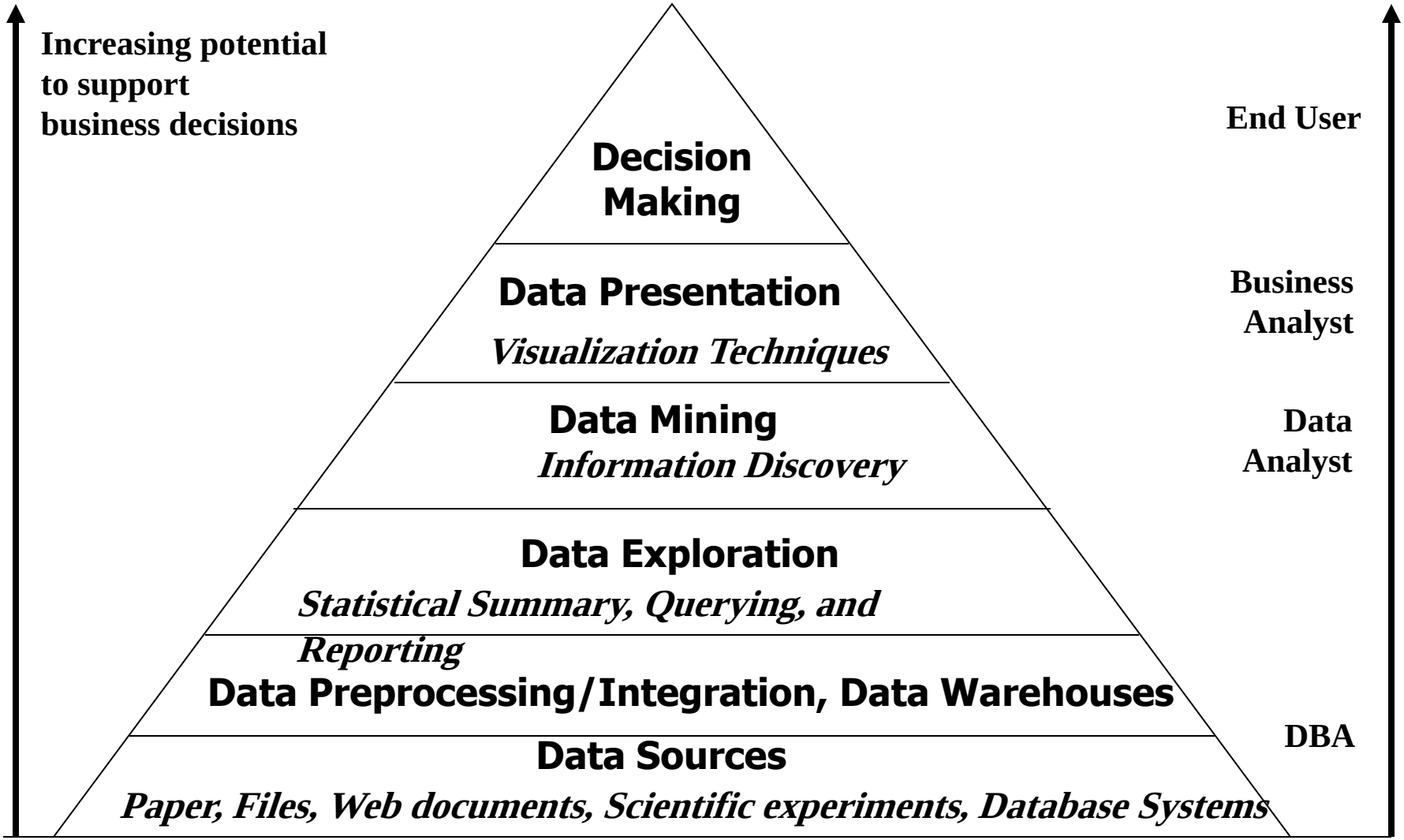
فرایند یا چرخه کشف دانش



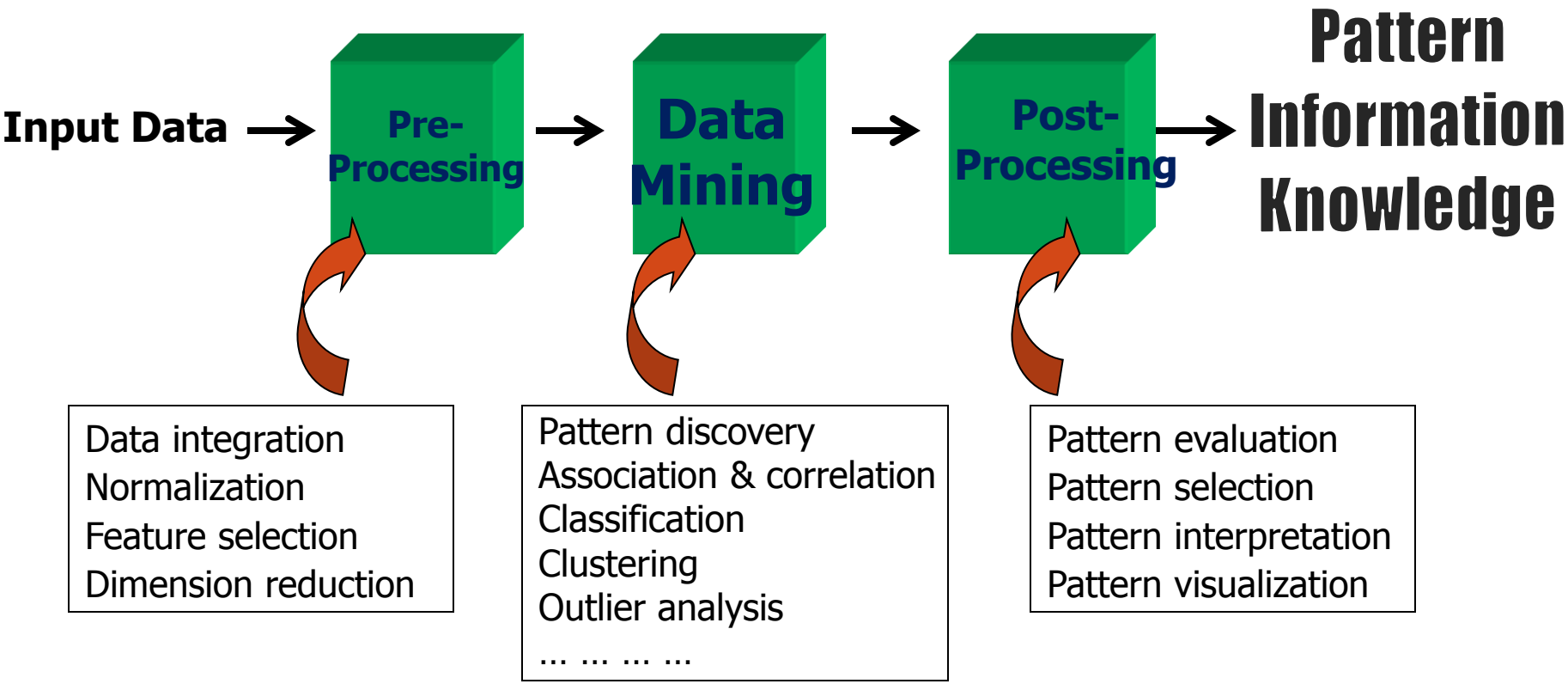
مراحل انجام داده کاوی

- پاکسازی، یکسان سازی داده ها
- یکپارچه سازی داده های به دست امده از منابع مختلف
- تشکیل انباره مجتمع داده ها
- ساخت ترکیب های متنوع سه بعدی از داده ها
- انتخاب گروه داده برای داده کاوی
- داده کاوی (تحلیل فراوانی، دسته بندی، برازش، خوشه بندی ...)
- نمایش نتایج داده کاوی
- استحصال دانش یا الگو و ثبت در پایگاه دانش

Data Mining in Business Intelligence



مراحل انجام داده کاوی



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

تقسیم بندی انواع داده کاوی از جنبه های متفاوت

- **Data to be mined**
  - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined (or: Data mining functions)**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

داده کاوی از لحاظ انواع داده ها

- 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 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,

DM Function: (2) Association, Correlation Analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together?
- Association, correlation vs. causality
  - A typical association rule
  - 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?

Data Mining Function: (3) Classification

- Classification and label prediction
  - Construct models (functions) based on training examples
  - Describe and distinguish classes for future prediction
  - 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, …

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

Time and Ordering: Sequential Pattern, Trend

- Sequence, trend and evolution analysis
  - regression and value prediction
  - Sequential pattern mining
  - Periodicity analysis
  - Motifs and biological sequence analysis
    - Approximate and consecutive motifs
  - Similarity-based analysis
- Mining data streams
  - Ordered, time-varying, potentially infinite, data streams

Structure and Network Analysis

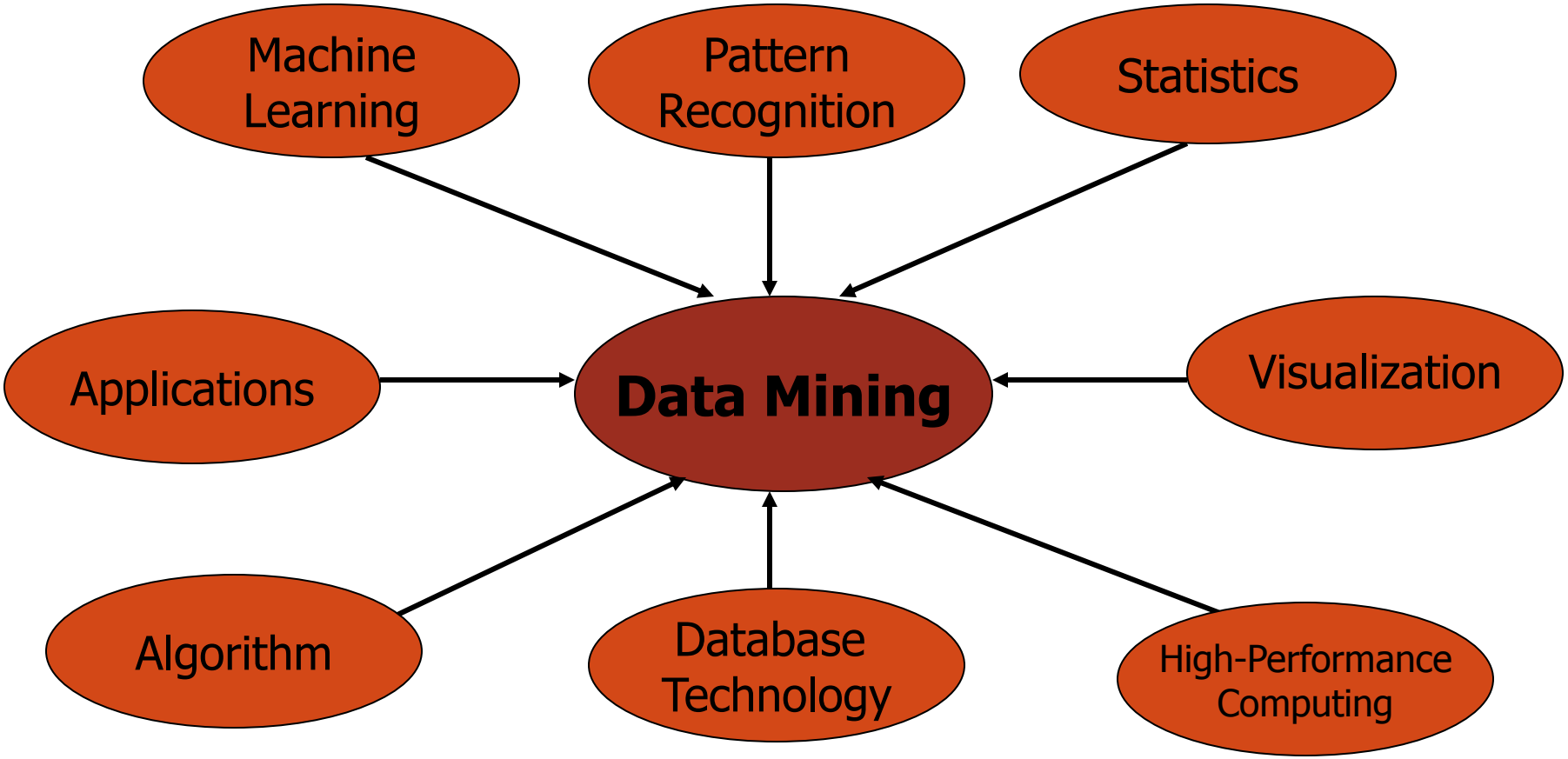
- Graph mining
  - Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- Information network analysis
  - Social networks
  - Multiple heterogeneous networks
  - Links carry a lot of semantic information
- Web mining
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
    - Web community discovery, opinion mining, usage mining, …

Evaluation of Knowledge

- Are all mined knowledge interesting?
  - One can mine tremendous amount of “patterns” and knowledge
  - 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, Coverage, Typicality vs. novelty, Accuracy, Timeliness…

Data Mining: Confluence of Multiple Disciplines



Applications of Data Mining

- Web page analysis: from web page classification, clustering to PageRank & HITS algorithms
- Collaborative analysis & recommender systems
- Basket data analysis to targeted marketing
- Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis
- Data mining and software engineering (e.g., IEEE Computer, Aug. 2009 issue)
- From major dedicated data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) to invisible data mining

Major Issues in Data Mining (1)

- Mining Methodology
  - Mining various and new kinds of knowledge
  - Mining knowledge in multi-dimensional space
  - Data mining: An interdisciplinary effort
  - Boosting the power of discovery in a networked environment
  - Handling noise, uncertainty, and incompleteness of data
  - Pattern evaluation and pattern- or constraint-guided 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
  - Invisible data mining

Summary

- Data mining: Discovering interesting patterns and knowledge from massive amount of data
- A natural evolution of database technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed in a variety of data
- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- Data mining technologies and applications
- Major issues in data mining