

資料探勘簡介

Data Mining

吳漢銘

國立政治大學 統計學系



<https://hmwu.idv.tw>



為什麼要使用R做為資料探勘工具?^{2/30}

- Why R?
 - R is a high-quality, cross-platform, flexible, widely used open source, free language for statistics, graphics, mathematics, and data science
 - R contains more than 5,000 algorithms and millions of users with domain knowledge worldwide.
- There are three shortages of R:
 - One is that it is **memory bound**, so it requires the entire dataset store in memory (RAM) to achieve high performance, which is also called in-memory analytics.
 - packages contributing to R communities are **bug-prone** and need more testing to ensure the quality of codes.
 - R seems **slow** than some other commercial languages.
 - Fortunately, there are packages available to overcome these problems.

Six of the Best Open Source Data Mining Tools

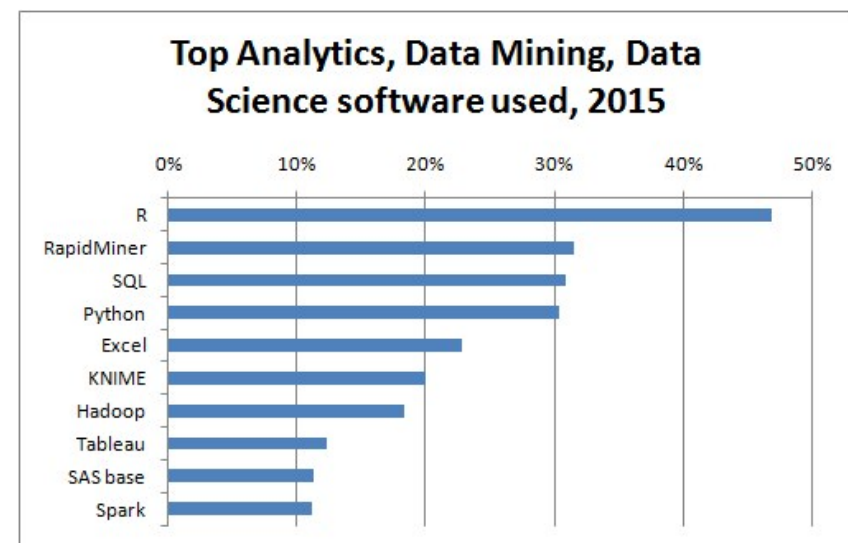
<http://thenewstack.io/six-of-the-best-open-source-data-mining-tools/>

40 Top Free Data Mining Software

<http://www.predictiveanalyticstoday.com/top-free-data-mining-software/>

10 Best Data Mining Software For Better Analysis

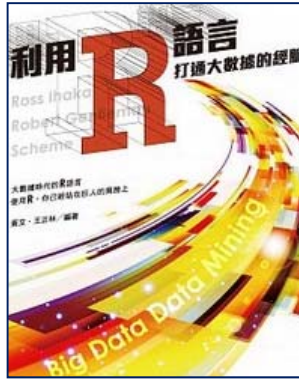
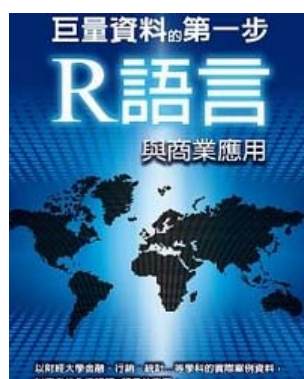
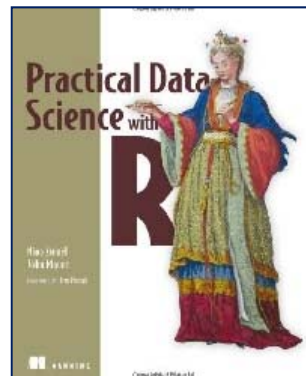
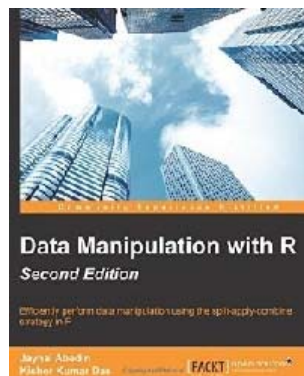
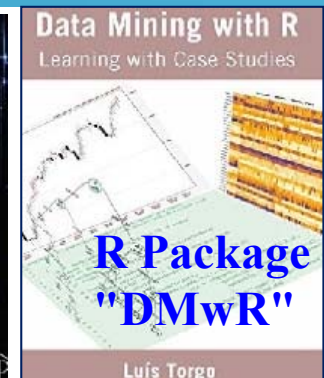
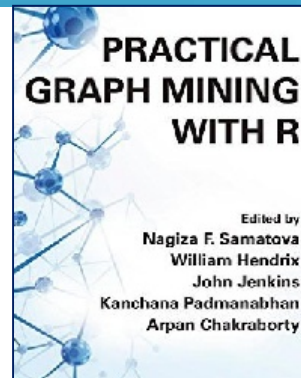
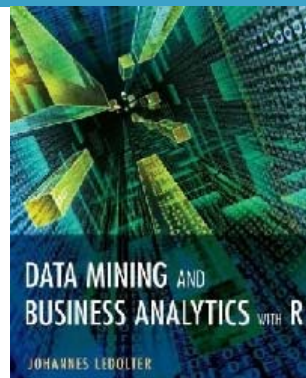
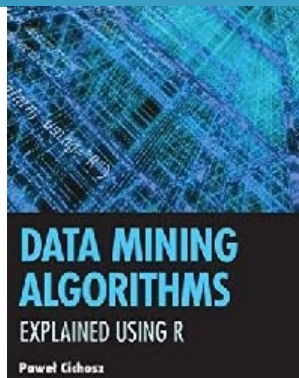
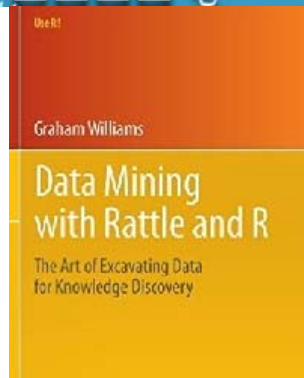
<http://digital.guide/10-best-data-mining-software-better-analysis/14362/?v=1>




Source: <http://www.kdnuggets.com/polls/2015/analytics-data-mining-data-science-software-used.html>

參考書目/學習網站

3/30



<http://www.rdatamining.com/>

**RDataMining.com: R and Data Mining**

Home

News

▼ Training

Past Trainings

▼ Documents

Introduction to Data Mining with R

R Reference Card for Data Mining

R and Data Mining: Examples and Case Studies

Introduction to Data Mining with R

Introduction to Data Mining with R and Data Import/Export in R

Data Exploration and Visualization with R

Regression and Classification with R

Data Clustering with R

Association Rule Mining with R

Text Mining with R: Twitter Data Analysis

Time Series Analysis and Mining with R

▼ Examples

Data Exploration

This website presents documents, examples, tutorials and resources on R and data mining.

[Documents on Data Mining with R](#)

- [R Reference Card for Data Mining](#)
- [R and Data Mining: Examples and Case Studies](#)
- [Introduction to Data Mining with R](#)
- RDataMining slides series on
 - [Introduction to Data Mining with R and Data Import/Export in R](#)
 - [Data Exploration and Visualization with R](#),
 - [Regression and Classification with R](#),
 - [Data Clustering with R](#),
 - [Association Rule Mining with R](#),
 - [Text Mining with R: Twitter Data Analysis](#), and
 - [Time Series Analysis and Mining with R](#)

[Examples on Data Mining with R](#)

■ 27 Free Data Mining Books

<http://www.dataonfocus.com/21-free-data-mining-books/>

■ Introduction to Data Mining

<http://www-users.cs.umn.edu/~kumar/dmbook/index.php>

■ Data Mining, Analytics, Big Data, and Data Science

<http://www.kdnuggets.com/>

KDnuggets™ Data Mining, Analytics, Big Data, and Data Science
Subscribe to [KDnuggets News](#) | Follow [Twitter](#) [Facebook](#) [LinkedIn](#) | [Contact](#)

[Data Mining Software](#) | [News](#) | [Top stories](#) | [Opinions](#) | [Jobs](#) | [Academic](#) | [Companies](#) | [Courses](#) | [Datasets](#) | [Education](#)

Data Mining, Analytics, Big Data, Data Science

[Software](#) (Suites, Text, Visualization)
[Jobs - Industry](#) | [Academic](#)
[Meetings, Conferences](#)
[Companies](#) (Consulting, Products)
[Courses in Big Data, Data Science](#)
[Datasets](#) (APIs/Markets, Gov)
[Data Mining Course](#) | [Gregory Piatetsky](#)
[Education](#) (online, USA, Europe, cert)
[FAQ](#) | [Polls](#) | [Publications](#) (Books)
[Solutions](#) (Fraud, Data Cleaning)
[Webcasts](#) | [Websites](#) (Blogs, Cartoons, Podcasts)

Most Recent

- [Top /r/MachineLearning Posts, February: AlphaGo, Distr...](#)
- [Top Spark Ecosystem Projects](#)
- [KDnuggets 16:n08, Mar 2: Citizen Data Scientist Mirage; Spa...](#)
- [2nd Annual Global Big Data For Executives Conference, ...](#)
- [New Salford Predictive Modeler 8](#)

Last poll results: [Deep Learning is not Enough](#)

Latest
[News](#) | [Software](#) | [Tutorials](#)

- [Top /r/MachineLearning Posts, February: AlphaGo, Distr...](#)
- [Top Spark Ecosystem Projects](#)

淡江大學覺生紀念圖書館館藏目錄

查詢結果: data mining

紀錄 1 - 20 of 965

[如何訂閱](#)

[Add Page To Cart](#) Search results too large to add all to cart.

- Data mining and data visualization / edited by C.R. Ra...**

索書號 QA76.9.D343 D3814 2005
出版項 Elsevier North-Holland, Amsterdam ; San Diego, CA : 2005.
版次 1st ed.
預約人數: 0
[加入標註清單](#)
- Exploratory data mining and data cleaning / Tamrapar...**

索書號 QA76.9.D343 D34 2003
出版項 New York : Wiley-Interscience, 2003.
電子版 [Table of contents](#)
1 copy is available
總館外圖書館 (1 available)
預約人數: 0
[加入標註清單](#)
- Data mining for social network data [electronic resour...**

索書號 QA76.9.D343 D38 2010
QA76.9.D343 D232 2010
出版項 Boston, MA : Springer Science+Business Media, LLC, 2010.
電子版 <http://info.lib.tku.edu.tw/ebook/redirect.asp?bibid=1329305>
預約人數: 0



Some free online data sources

Some free online data sources particularly helpful to learn about data mining

- Frequent Itemset Mining Dataset Repository:

<http://fimi.ua.ac.be/data/>

- UCI Machine Learning Repository:

<http://archive.ics.uci.edu/ml/>

- The Data and Story Library at statlib:

<http://lib.stat.cmu.edu/DASL/>

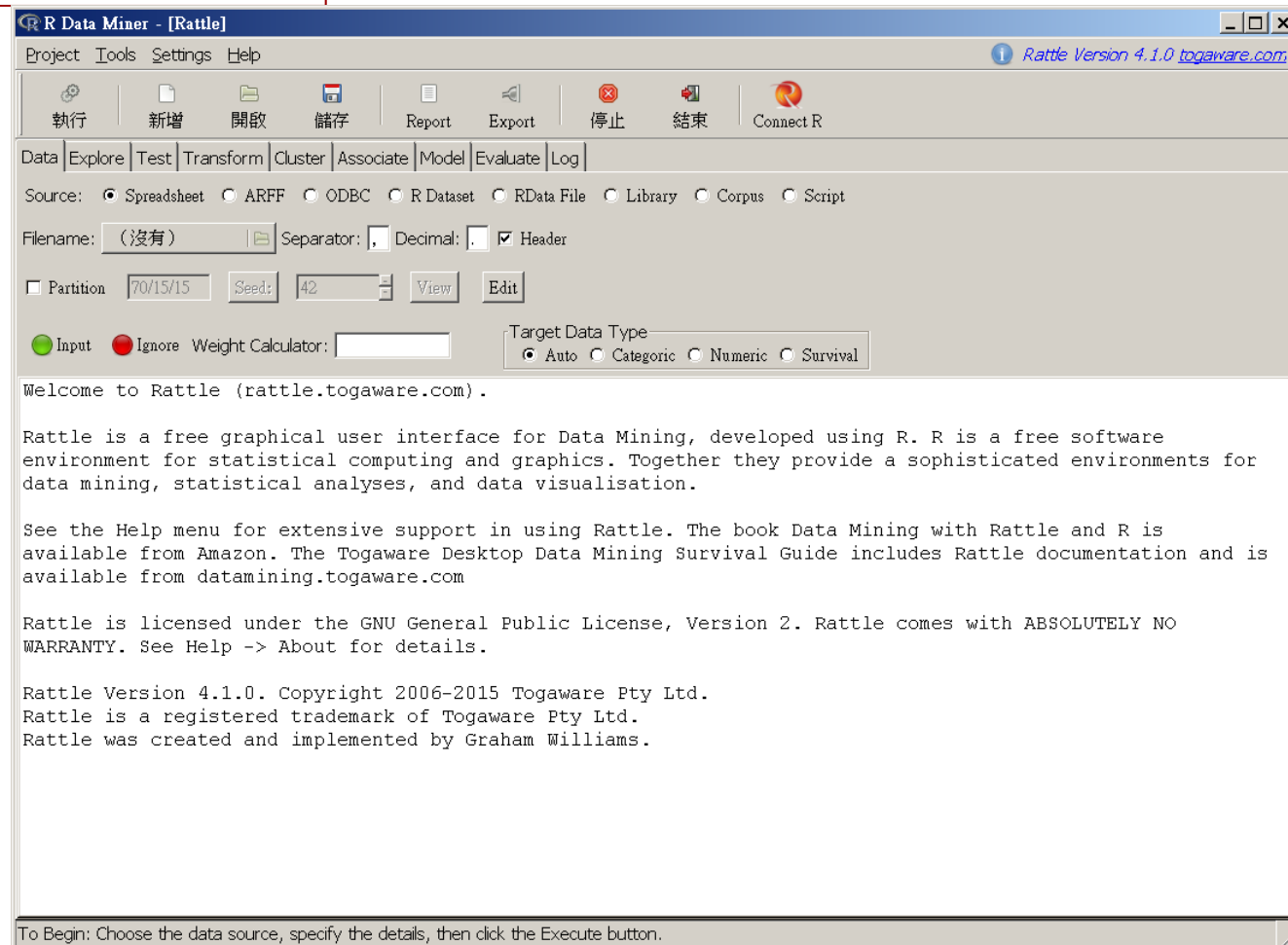
- WordNet:

<http://wordnet.princeton.edu>

Rattle: A Graphical User Interface for Data Mining using R 6/30

```
> install.packages("rattle")
> library(rattle)
> rattleInfo()
> nstall.packages(rattleInfo())
> rattle()
```

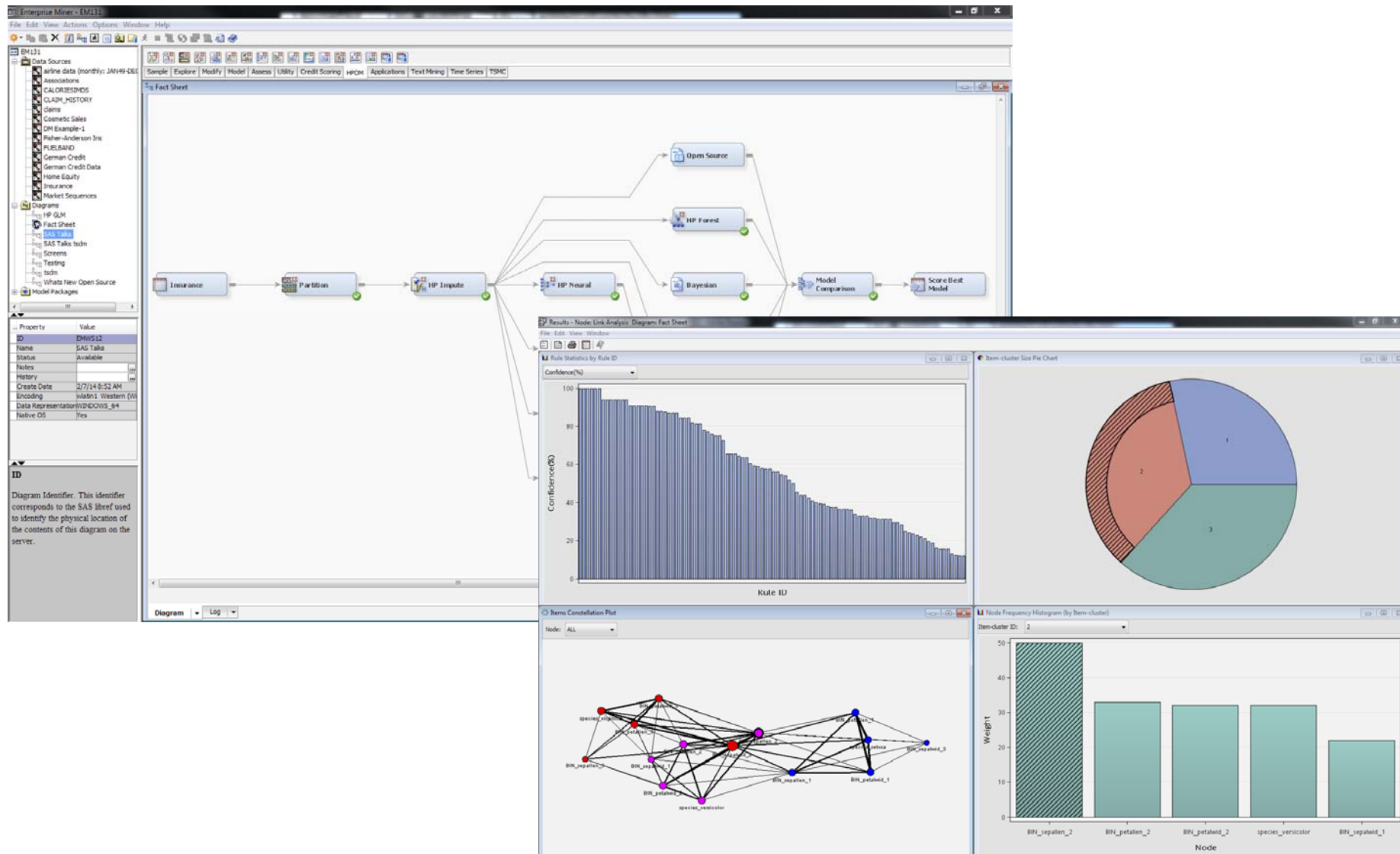
<http://rattle.togaware.com/>



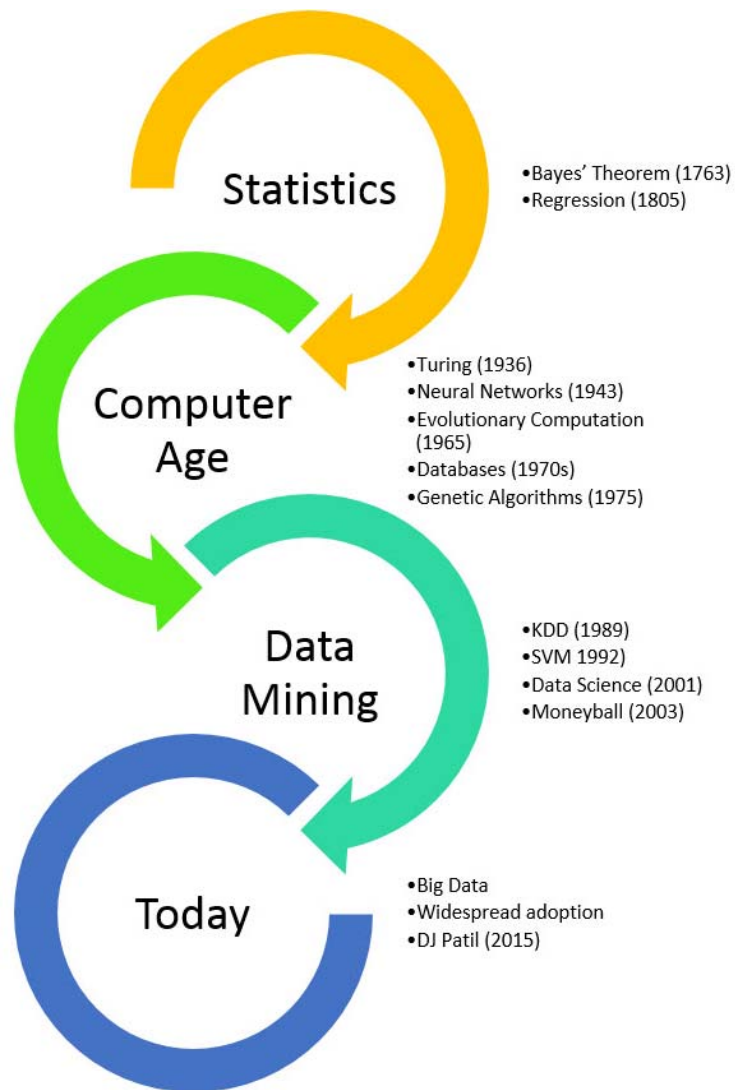
SAS Enterprise Miner

7/30

http://www.sas.com/zh_tw/software/analytics/enterprise-miner.html



History



- The term "Data Mining" appeared around **1990** in the **database** community.
- Gregory Piatetsky-Shapiro coined the term "**Knowledge Discovery in Databases**" for the first workshop on the same topic (KDD-1989) and this term became more popular in **AI** and **Machine Learning Community**.
- However, the term data mining became more popular in the **business and press** communities.
- Currently, **Data Mining** and **Knowledge Discovery** are used interchangeably.
- Since about 2007, "**Predictive Analytics**" and since 2011, "**Data Science**" terms were also used to describe this field.

FRANS COENEN, **Data Mining: Past, Present and Future**, The Knowledge Engineering Review, Vol. 00:0, 1–24. 2004, Cambridge University Press

Source: History of data mining
<http://rayli.net/blog/data/history-of-data-mining/>



What is Data Mining? (1/4)

- "Data mining is the process of exploration and **analysis**, by automatic or semiautomatic means, of large quantities of data in order to **discover** meaningful **patterns and rules**." (M. J. A. Berry and G. S. Linoff)
- "Data mining is **finding** interesting **structure** (patterns, statistical models, relationships) in **databases**." (U. Fayyad, S. Chaudhuri and P. Bradley)
- "Data mining is the application of statistics in the form of exploratory data analysis and predictive models to **reveal patterns and trends** in very large data sets." ("Insightful Miner 3.0 User Guide")
- Data mining is also known as **Knowledge Discovery in Data** (KDD).



What is Data Mining? (2/4)

10/30

- Non-trivial **extraction** of implicit, previously unknown and potentially **useful information** from data.
- Generally, data mining is the process of **analyzing data** from different perspectives and summarizing it into **useful** information.
- Technically, data mining is the process of **finding** correlations or **patterns** among dozens of fields in large relational databases.
- Data mining is the practice of automatically searching large stores of data to **discover patterns** and trends that go beyond simple analysis.



What is Data Mining? (3/4)

11/30

- Data mining is a process used by companies to turn raw data into **useful information**.
- Data mining depends on effective data collection and warehousing as well as computer **processing**.
- Data mining is the computational process of **discovering patterns** in large data sets ("big data") involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.
- The overall goal of the data mining process is to **extract information** from a data set and transform it into an understandable structure for further use.



What is Data Mining? (4/4)

- Data mining is **sifting** through very large amounts of data for useful information.
- Data mining uses artificial intelligence techniques, neural networks, and advanced statistical tools (such as cluster analysis) to **reveal trends**, patterns, and relationships, which might otherwise have remained undetected.
- Data mining requires a class of database applications that look for hidden patterns in a group of data that can be used to **predict** future behavior.

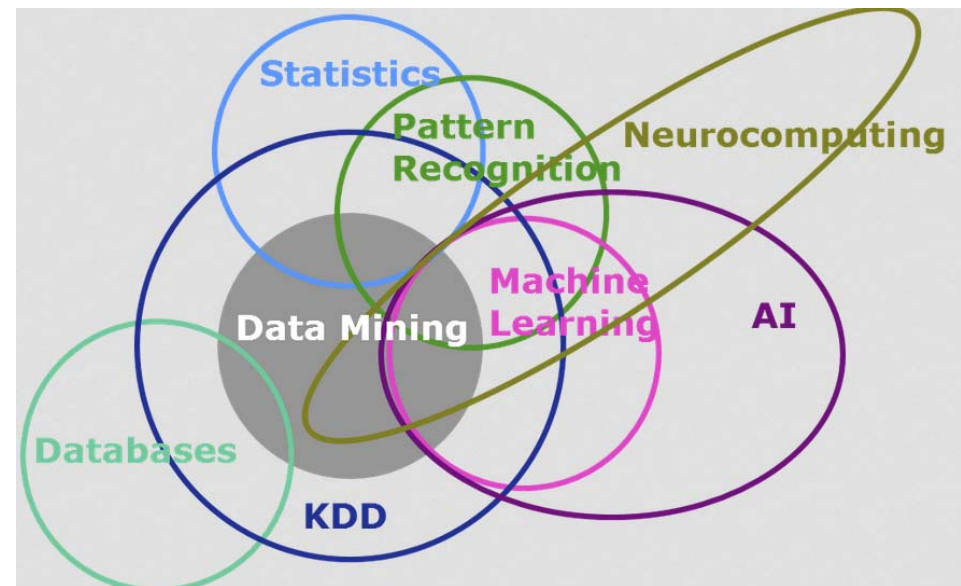
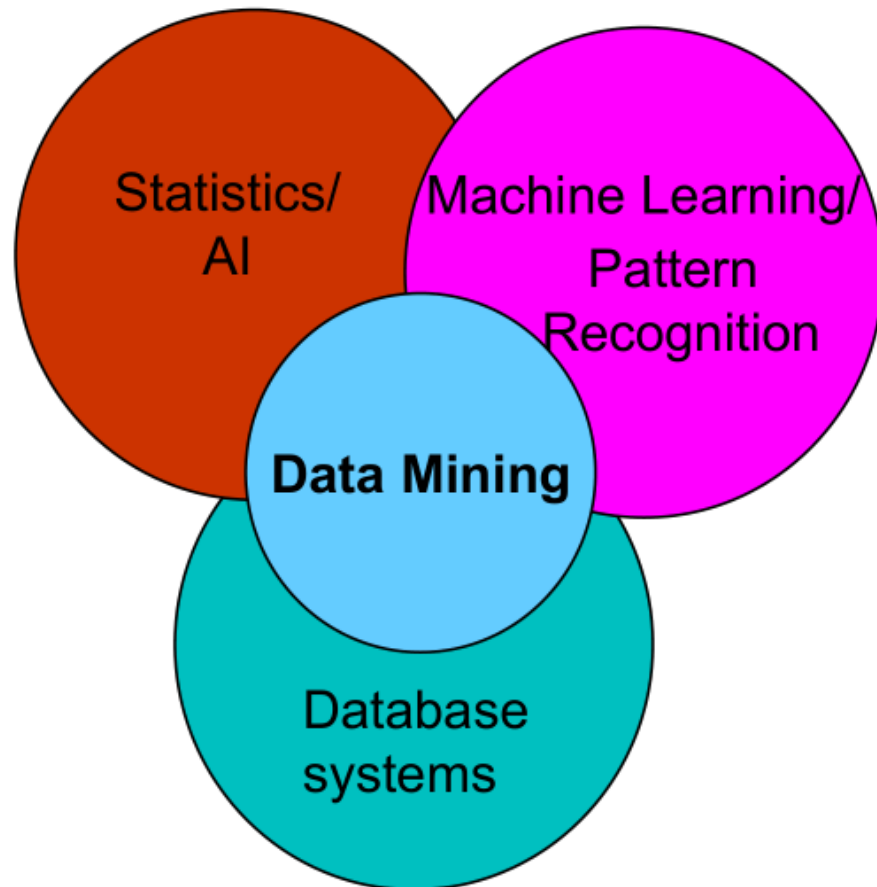


Data Mining

- Data mining derives its name from the similarities between **searching for valuable information in a large database** and **mining rocks for a vein (礦脈) of valuable ore (礦石)**.
 - mining for gold in rocks: "gold mining" (not "rock mining")
 - data mining should have been called "knowledge mining" .
- Other similar terms referring to data mining are: data dredging, knowledge extraction and pattern discovery.
- Data mining is the analysis step of the "knowledge discovery in databases" process, or **KDD**.

Data Mining Diagrams

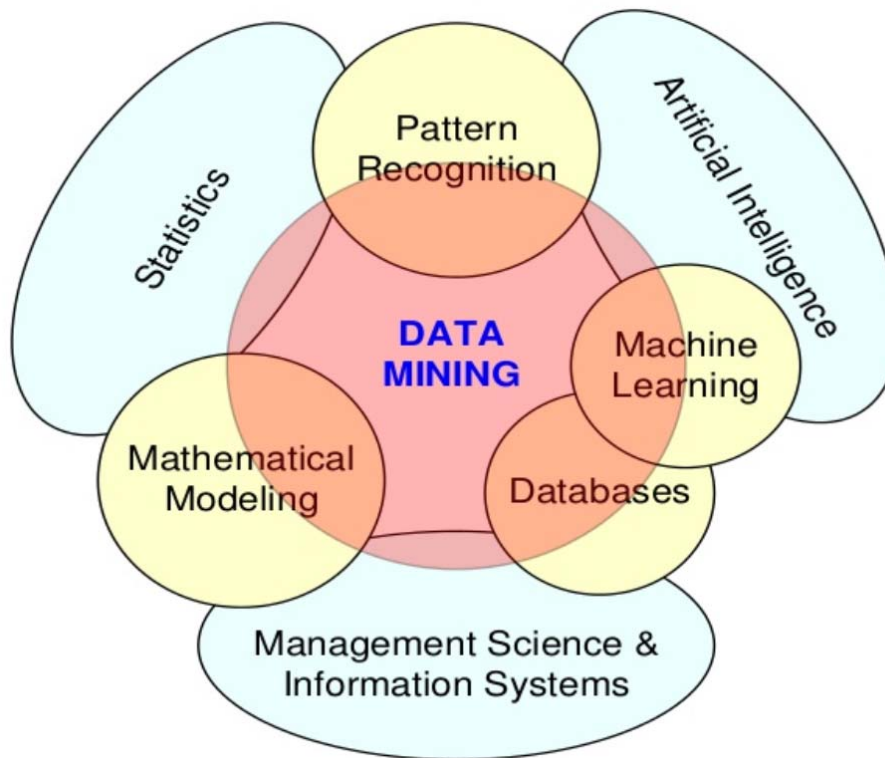
14/30



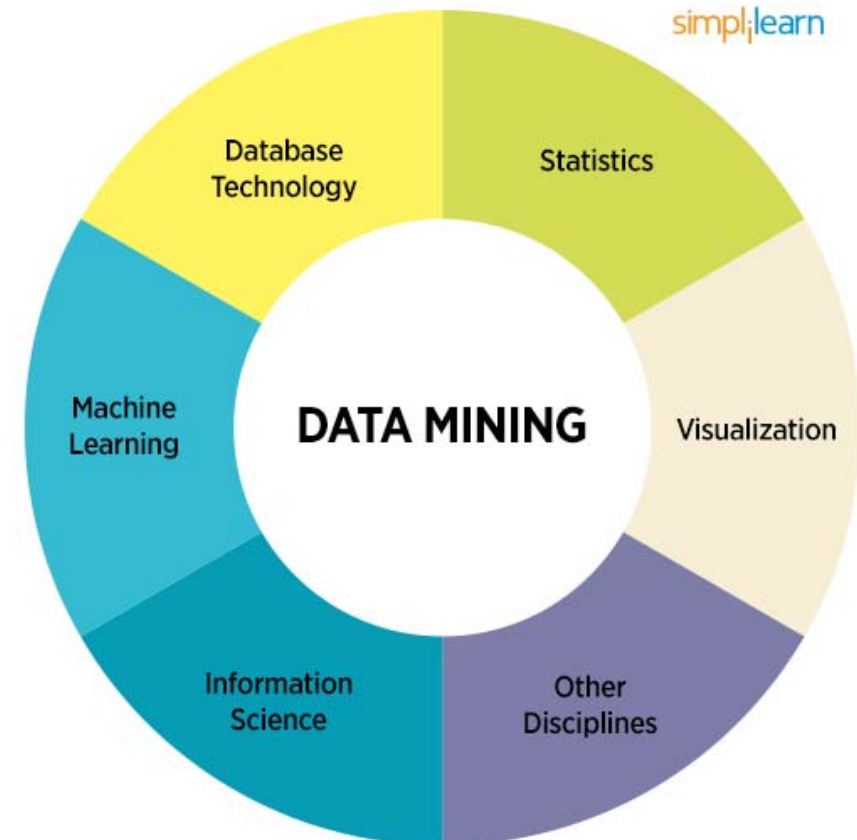
Source:

<http://blogs.sas.com/content/subconsciousmusings/files/2014/08/data-mining-Venn-diagram.png>

Source: Tan, Steinbach, Kumar Introduction to Data Mining 4/18/2004



Source: Published on Nov 26, 2014
 Language Technologies for Geomatics: From Intelligence to Agility
 Published in: Technology
<http://www.slideshare.net/VisionGEOMATIQUE2014/gagnon-20141112vision>



Source:
<http://www.simplilearn.com/data-mining-vs-statistics-article>

What are Data Mining and Knowledge Discovery?

16/30

Examples of data mining

https://en.wikipedia.org/wiki/Examples_of_data_mining

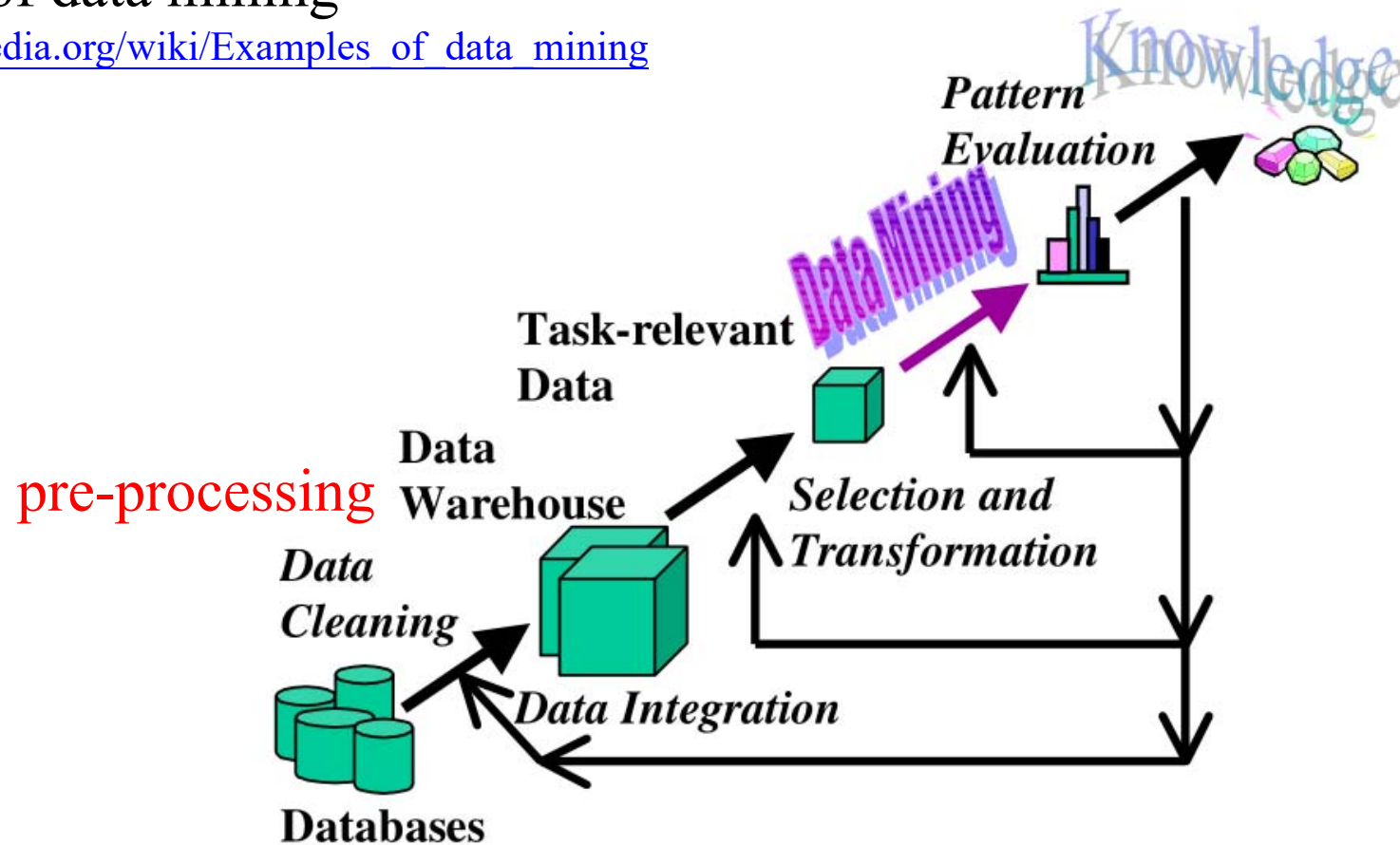


Figure 1.1: Data Mining is the core of Knowledge Discovery process

Source: Osmar R. Zaïane, 1999 CMPUT690 Principles of Knowledge Discovery in Databases

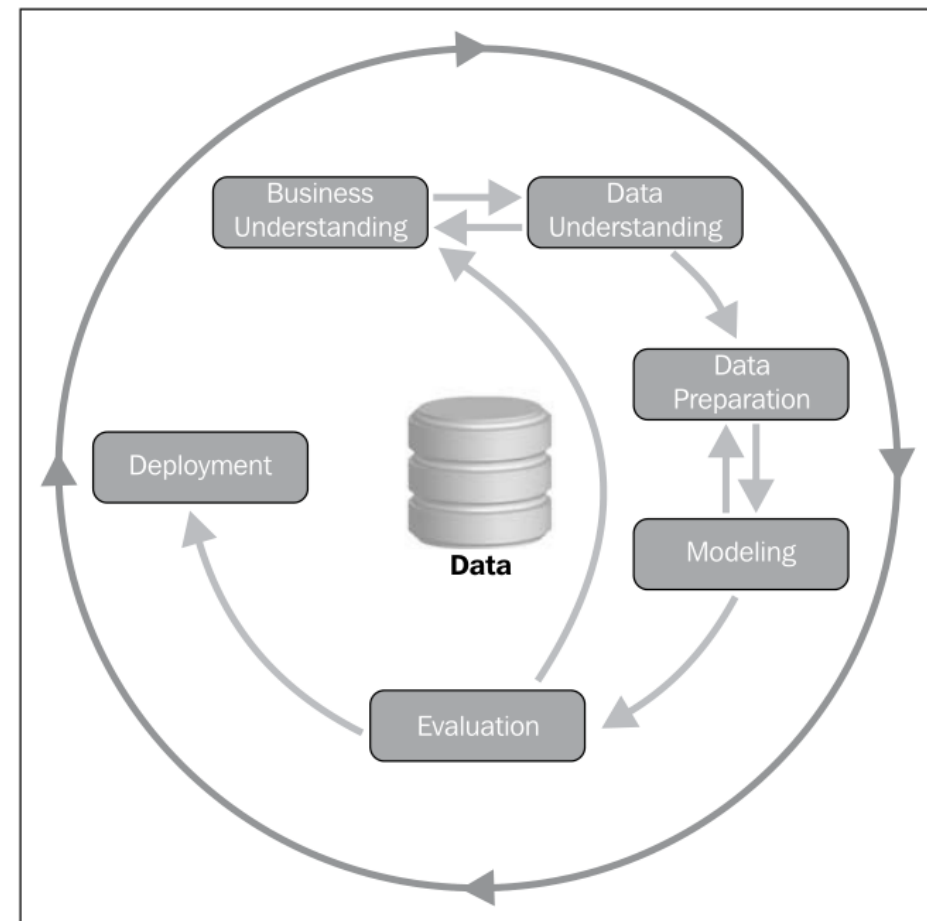


The Data Mining Process: CRISP-DM

17/30

- **Business understanding**: determining business objectives, establishing data mining goals, and developing a plan.
- **Data understanding**: initial data collection, data description, data exploration, and the verification of data quality.
- **Data preparation**: the data needs to be selected, cleaned, and then built into the desired form and format.
- **Modeling**: visualization and cluster analysis, association rules. to discover knowledge represented as rules.
- **Evaluation** : the results should be evaluated in the context specified by the business objectives in the first step. This leads to the identification of new needs and in turn reverts to the prior phases in most cases.
- **Deployment**: data mining can be used to both verify previously held hypotheses or for knowledge.

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

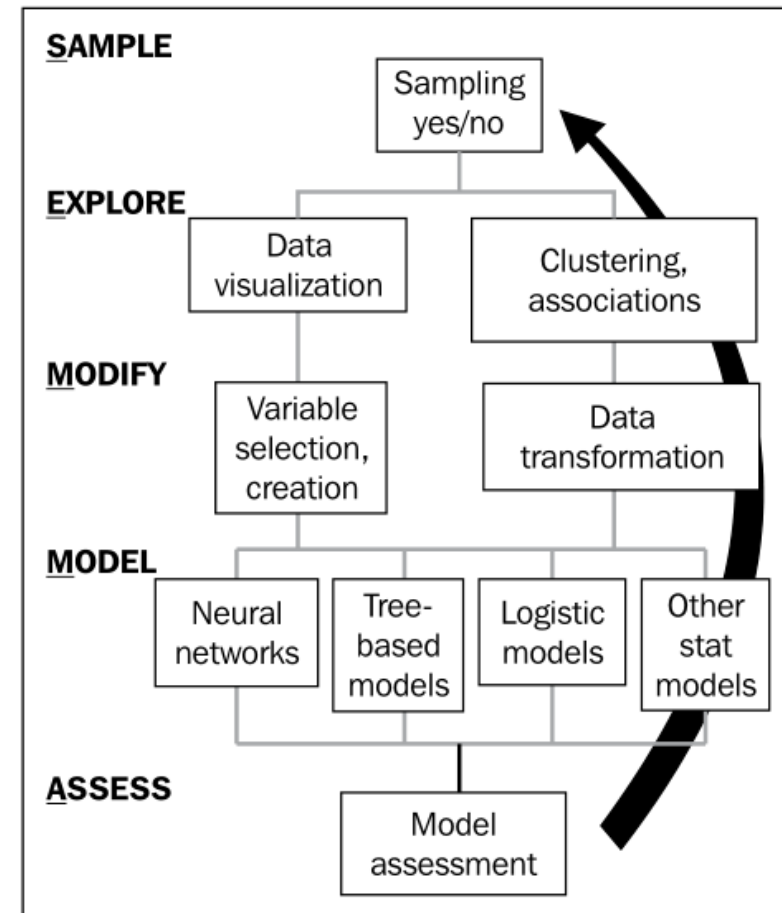


Source: Biter Makhabel, 2014, Learning Data Mining with R, Packt Publishing, (December 22, 2014).

The Data Mining Process: SEMMA

- **Sample:** In this step, a portion of a large dataset is extracted.
- **Explore:** To gain a better understanding of the dataset, unanticipated trends and anomalies are searched in this step.
- **Modify:** The variables are created, selected, and transformed to focus on the model construction process.
- **Model:** A variable combination of models is searched to predict a desired outcome.
- **Assess:** The findings from the data mining process are evaluated by its usefulness and reliability.

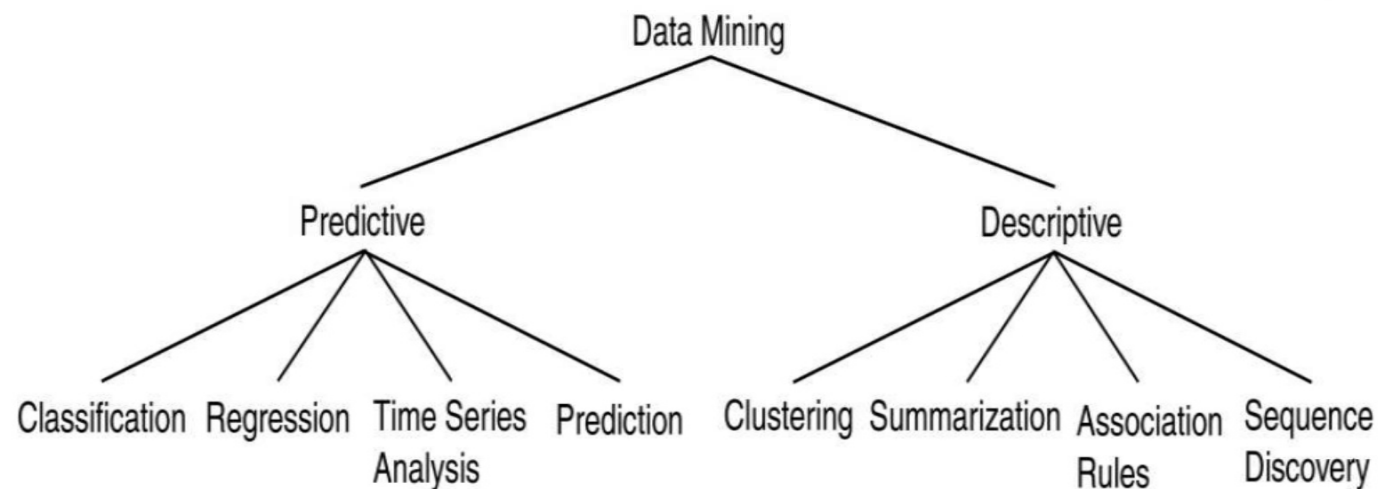
Sample, Explore, Modify, Model, Assess (SEMMA), which was developed by the SAS Institute, USA



Source: Biter Makhabel, 2014, Learning Data Mining with R, Packt Publishing, (December 22, 2014).

Data Mining Tasks

- There are two types of data mining tasks:
 - **descriptive** data mining tasks that describe the general properties of the existing data, and
 - **predictive** data mining tasks that attempt to do predictions based on inference on available data.
- **The 6 Common DM Tasks:** Description, Estimation, Prediction, Classification, Clustering, Association





Data Mining Tasks

- **Data processing:** cleaning, integration, reduction, transformation, feature extraction (dimension reduction), discretization, missing values, outliers detection.
- **Data exploration:** summary, visualization, correlation analysis
- **Regression:** linear, polynomial, lasso, logistic regression, nonlinear regression, regression tree.
- **Classification:** nearest neighbor, linear discriminant analysis, decision tree, naive Bayes classifier, CART, random forest, SVM, artificial neural networks, ...
- **Cluster analysis:** k-means, hierarchical clustering, PAM, model-based, ...
- **Link analysis:** associations rules discovery
- **Model evaluation:** variables selections, cross-validation
- **Better Modelling:** ensembles (bagging, boosting), kernel methods, ...
- **Applications and Beyond:**
 - Data types: text and web data, stream, time series and sequence data, network data.
 - Big data



Six Common Classes of Tasks

21/30

- **Summarization** : providing a more compact representation of the data set, including visualization and report generation.
- **Anomaly detection** (Outlier/change/deviation detection): The identification of unusual data records, that might be interesting or data errors.
- **Regression** : find a function which models the data with the least error.
- **Classification** : is the task of generalizing known structure to apply to new data.
- **Clustering**: is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- **Association rule learning** (Dependency modelling): Searches for relationships between variables.

Source: https://en.wikipedia.org/wiki/Data_mining

- 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.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Source: Tan, Steinbach, Kumar Introduction to Data Mining 4/18/2004 23



What is Statistics?

- Statistics is a component of data mining that provides the tools and analytics techniques for dealing with large amounts of data.
 - It is the science of learning from data and includes everything from collecting and organizing to analyzing and presenting data.
 - It is concerned with **probabilistic models**, specifically inference, using data.
- While the aims of statistics and data mining are **similar**, it is estimated that there are **very few statisticians to deal with the demands of data analysts**.
- Statisticians were the first to use the term data mining.

Source: <http://www.simplilearn.com/data-mining-vs-statistics-article>



Statistics and Data Mining (1/4)

24/30

- Statistics studies the collection, analysis, interpretation or explanation, and presentation of data. It serves as the **foundation** of data mining.
- Originally, data mining was a derogatory (貶低的) term referring to attempts to extract information that was not supported by the data.
- To some extent, data mining constructs **statistical models**, which is an underlying distribution, used to visualize data.
- Data mining has an inherent relationship with statistics; one of the mathematical foundations of data mining is statistics, and **many statistics models are used in data mining**.
- Statistical methods can be used to summarize a collection of data and can also be used to **verify data mining results**.



Statistics and Data Mining (2/4)

25/30

- There is a great deal of **overlap** between data mining and statistics.
- Most of the techniques used in data mining can be placed in a statistical framework. However, data mining techniques are not the same as traditional statistical techniques.
 - Traditional statistical methods, in general, require a great deal of user interaction in order to validate the correctness of a model. As a result, statistical methods can be **difficult to automate**.
 - Moreover, statistical methods typically **do not scale well** to very large data sets.
 - Statistical methods rely on testing hypotheses or finding correlations based on smaller, **representative samples** of a larger population.
- Data mining methods are suitable for large data sets and can be more readily automated.
- In fact, data mining algorithms often require large data sets for the creation of quality models.



Statistics and Data Mining (3/4)

26/30

- Both data mining and statistics are related to **learning from data**. They are all about discovering and identifying structures in them, thus **aimed** at turning data to information.
- And although the aims of both these techniques overlap, they have **different approaches**.
- Statistics is only about quantifying data. While it uses tools to find relevant properties of data, it is a lot like math. It provides the tools necessary for data mining.
- Data mining, on the other hand, builds **models** to detect patterns and relationships in data, **particularly from large data bases**.



Statistics and Data Mining (4/4)

27/30

- Gregory Piatetsky-Shapiro: Statistics is at the core of data mining - helping to distinguish between **random noise** and **significant findings**, and providing a theory for estimating probabilities of predictions, etc.
- However Data Mining is **more than** Statistics.
- DM covers the entire process of data analysis, including data cleaning and preparation and visualization of the results, and how to produce predictions in real-time, etc.

Source: <http://www.kdnuggets.com/faq/difference-data-mining-statistics.html>

[Wiki] Gregory I. Piatetsky-Shapiro (born 7 April 1958) is a data scientist, co-founder of KDD conferences and ACM SIGKDD association for Knowledge Discovery and Data Mining, and President of KDnuggets, a leading site on Business Analytics, Data Mining, and Data Science. For simplicity, he usually abbreviates his name as Gregory Piatetsky.

Further reading: Friedman, J. H. "Data Mining and Statistics: What's the Connection?" (Nov. 1997b).

<http://www-stat.stanford.edu/~jhf/ftp/dm-stat.pdf>



Statistics, Data Mining and Big Data

	Statistics	Data Mining	Big Data
Structure	structured	structured	unstructured
Size	small	large	very large
Generation	planned	transactional	behavioral
Aim	understand	optimize business	generate business
Privacy Issues	non	minor	huge
Founded On	concepts & theory	technology & tool	technology & tools
Marketing	bad	good	perfect

Source: <http://www.theusrus.de/blog/some-truth-about-big-data/>



Data Mining and Statistics: What is the Connection?

29/30

- The field of data mining, like statistics, concerns itself with "learning from data" or "turning data into information".
- Rather, it is important to note that data mining can learn from statistics – that, to a large extent, statistics is fundamental to what data mining is really trying to achieve.

However,

- most **data miners** tend to be ignorant of statistics and client's domain;
- **statisticians** tend to be ignorant of data mining and client's domain; and
- **clients** tend to be ignorant of data mining and statistics.

- **computer scientists** focus upon database manipulations and processing algorithms;
- **statisticians** focus upon identifying and handling uncertainties; and
- **clients** focus upon integrating knowledge into the knowledge domain.
- Moreover, most **data miners and statisticians** continue to sarcastically criticise each other.

- Data mining and statistics will inevitably grow toward each other in the near future because data mining will not become knowledge discovery without **statistical thinking**, statistics will not be able to succeed on massive and complex datasets without data mining approaches.

Source: Data Mining and Statistics: What is the Connection? Posted on October 1, 2004 by Diego Kuonen, PhD
<http://tdan.com/data-mining-and-statistics-what-is-the-connection/5226>



Challenges/Trends of Data Mining

Challenges

- Scalability, Dimensionality, Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution (eg, skewed distribution)
- Privacy Preservation
- New application-based requirements: Streaming Data

Trends

- Application Exploration
- Scalable and interactive data mining methods
- Visual data mining
- New methods of mining complex types of data
- Biological data mining
- Data mining and software engineering
- Web mining, real-time data mining
- Distributed data mining
- Real time data mining
- Multi database data mining
- Privacy protection and information security in data mining.

Source: <http://www.simplilearn.com/data-mining-vs-statistics-article>