CSE4334/5334 Data Mining

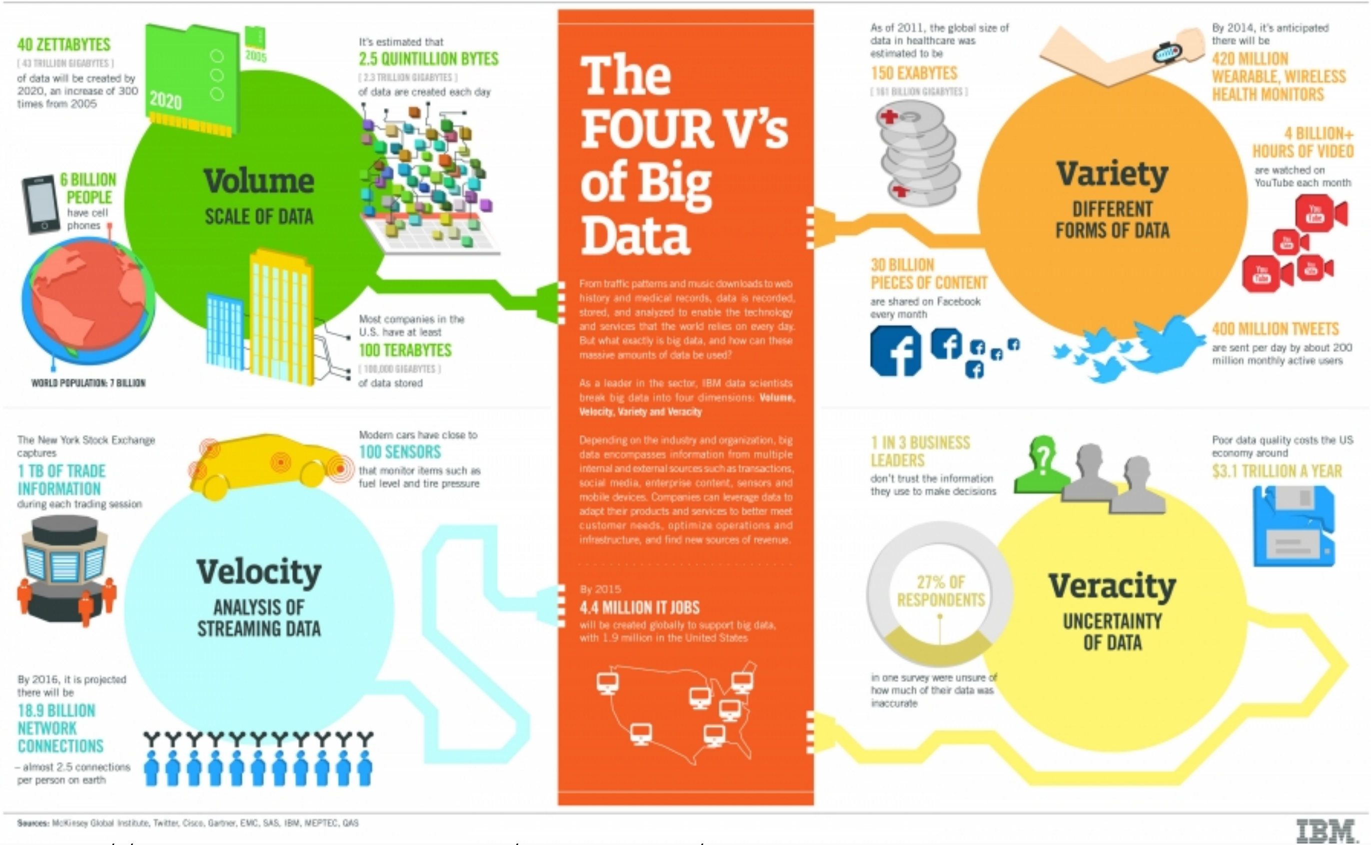
Overview of Data Mining

Chengkai Li

Department of Computer Science and Engineering
University of Texas at Arlington

Fall 2020 (Slides partly courtesy of Pang-Ning Tan, Michael Steinbach and Vipin Kumar, and Jiawei Han, Micheline Kamber and Jian Pei)

Big Data



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM.

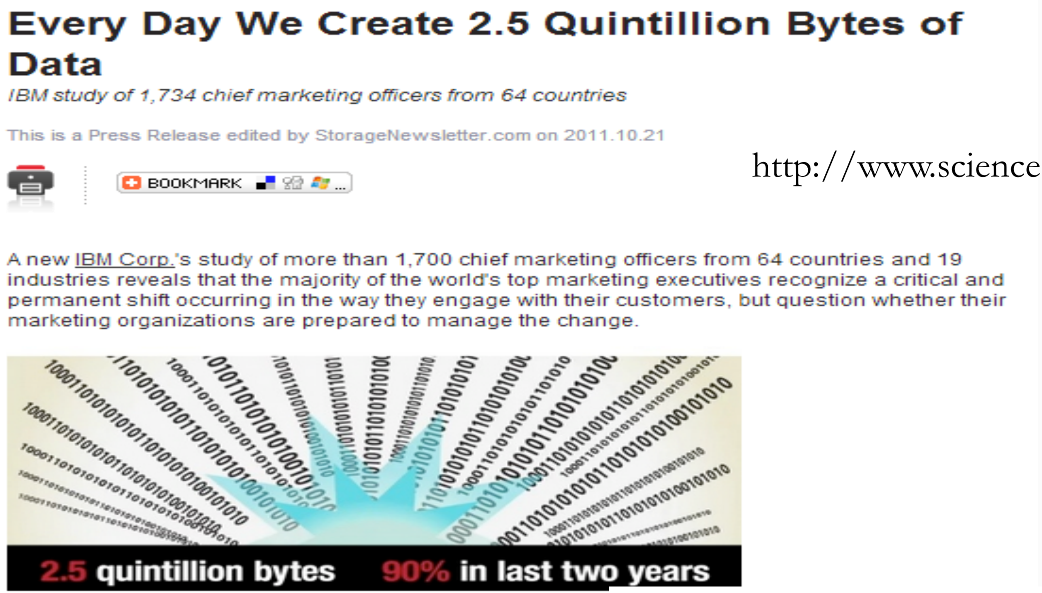
http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Big Data

The 4 Vs

- Volume
- Variety
- Velocity
- Veracity

Volume: How much data is out there?



Every Day We Create 2.5 Quintillion Bytes of Data

IBM study of 1,734 chief marketing officers from 64 countries

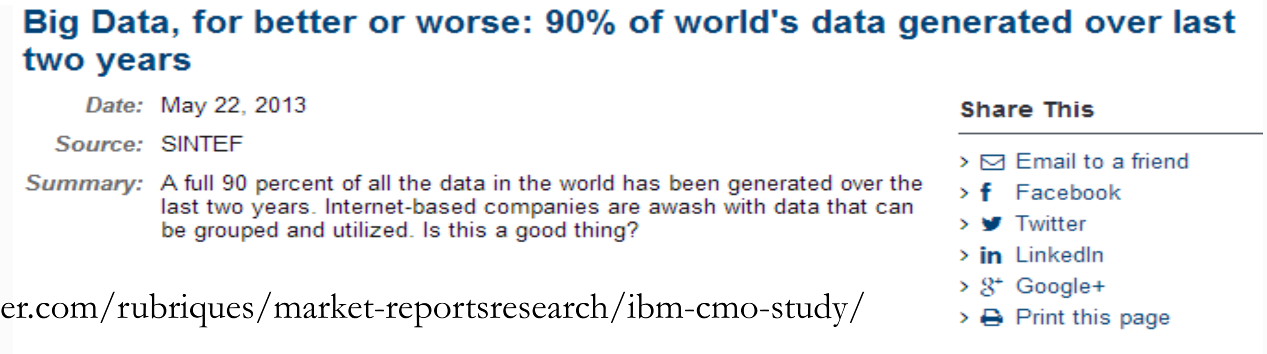
This is a Press Release edited by StorageNewsletter.com on 2011.10.21

BOOKMARK

A new IBM Corp.'s study of more than 1,700 chief marketing officers from 64 countries and 19 industries reveals that the majority of the world's top marketing executives recognize a critical and permanent shift occurring in the way they engage with their customers, but question whether their marketing organizations are prepared to manage the change.

2.5 quintillion bytes 90% in last two years

http://www.sciencedaily.com/releases/2013/05/130522085217.htm



Big Data, for better or worse: 90% of world's data generated over last two years

Date: May 22, 2013

Source: SINTEF

Summary: A full 90 percent of all the data in the world has been generated over the last two years. Internet-based companies are awash with data that can be grouped and utilized. Is this a good thing?

Share This

- Email to a friend
- Facebook
- Twitter
- LinkedIn
- Google+
- Print this page

http://www.storagenewsletter.com/rubriques/market-reportsresearch/ibm-cmo-study/

Volume: How much data is out there?

In total, 2.7 Zettabytes of data exists in our digital universe.

("A terabyte is equal to 1,024 gigabytes. A petabyte is equal to 1,024 terabytes. An exabyte is equal to 1,024 petabytes. A zettabyte is equal to 1,024 exabytes.")

Online Activity

- Every minute:
  - 149, 513 emails are sent
  - 3.3 million Facebook posts are created
  - 3.8 million Google searches are performed
  - 65,972 Instagram photos are uploaded.
  - 448,800 tweets are constructed
  - 500 hours of YouTube videos are uploaded

https://www.nodegraph.se/big-data-facts/

Variety: Types of Data

Structured data

- (relational) database tables
- CSV/TSV files

Semi-structured data

- XML, JSON, RDF

Unstructured data

- text data (documents, Web pages, short texts, e.g., social media)

Multimedia data

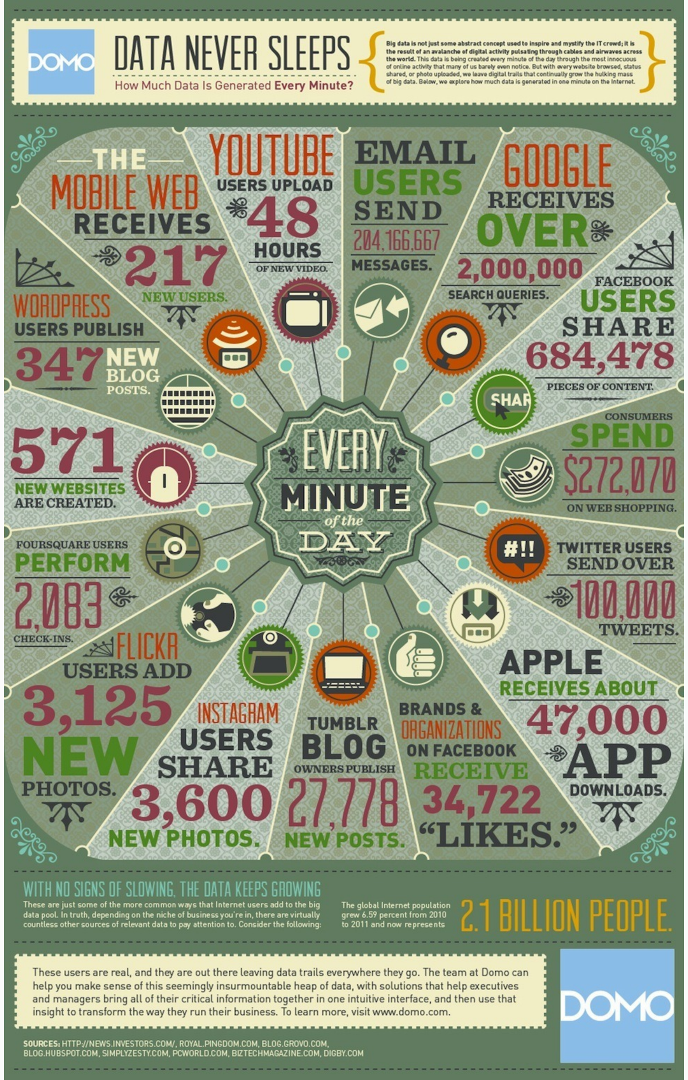
- images, videos, audios

Other types of data

- matrices, graphs, sequences, time-series, spatio-temporal

Velocity: Streaming Data

- Stock trades
- Highway sensors
- Weather data
- Social media
- Telephone calls
- Video streaming



DATA NEVER SLEEPS

How Much Data Is Generated Every Minute?

Big data is not just some abstract concept used to inspire and mystify the IT crowd; it is the result of an avalanche of digital activity pulsating through cables and airwaves across the world. This data is being created every minute of the day through the most innocuous of online activity that many of us barely even notice. But with every website browsed, status shared, or photo uploaded, we leave digital trails that continually grow the hulking mass of big data. Below, we explore how much data is generated in one minute on the Internet.

EVERY MINUTE of the DAY

The mobile web receives 217 new users.

YouTube users upload 48 hours of new video.

Email users send 204,166,667 messages.

Google receives over 2,000,000 search queries.

Facebook users share 684,478 pieces of content.

Consumers spend $272,070 on web shopping.

Twitter users send over 100,000 tweets.

Apple receives about 47,000 app downloads.

Brands & organizations on Facebook receive 34,722 "likes."

Tumblr blog owners publish 27,778 new posts.

Instagram users share 3,600 new photos.

Flickr users add 3,125 new photos.

Foursquare users perform 2,083 check-ins.

571 new websites are created.

WordPress users publish 347 new blog posts.

WITH NO SIGNS OF SLOWING, THE DATA KEEPS GROWING

These are just some of the more common ways that Internet users add to the big data pool. In truth, depending on the niche of business you're in, there are virtually countless other sources of relevant data to pay attention to. Consider the following:

The global Internet population grew 6.59 percent from 2010 to 2011 and now represents 2.1 BILLION PEOPLE.

These users are real, and they are out there leaving data trails everywhere they go. The team at Domo can help you make sense of this seemingly insurmountable heap of data, with solutions that help executives and managers bring all of their critical information together in one intuitive interface, and then use that insight to transform the way they run their business. To learn more, visit www.domo.com.

SOURCES: HTTP://NEWS.INVESTORS.COM/, ROYAL.PINGDOM.COM, BLOG.GROVO.COM, BLOG.HUBSPOT.COM, SIMPLYZESTY.COM, PCWORLD.COM, BIZTECHMAGAZINE.COM, DIGBY.COM

http://mashable.com/2012/06/22/data-created-every-minute/

Veracity: uncertain and imprecise data

- Biases
- Data Lineage
- Bugs
- Noise
- Abnormalities
- Information security
- Unreliable sources
- Falsification
- Uncertainty
- Out of date
- Human error

https://simplicable.com/new/data-veracity

Data in Every Application Area

- Business: e-commerce, transactions (retailers, banking, credit cards), ratings, reviews, stock trading, …
- Web, social media (YouTube, Flickr, …), and social networks (Facebook, Twitter, …)
- News
- Science: bioinformatics, scientific experiments, environment, climate, astronomy
- Logs and measurements
- Personal information: emails, calendars, digital photos, videos
- Transportation
- Telecommunication
- Education
- Entertainment (film, music, gaming, …)
- Sports
- Health care
- Crime, security

What is Data Mining?

Data mining (knowledge discovery from data)

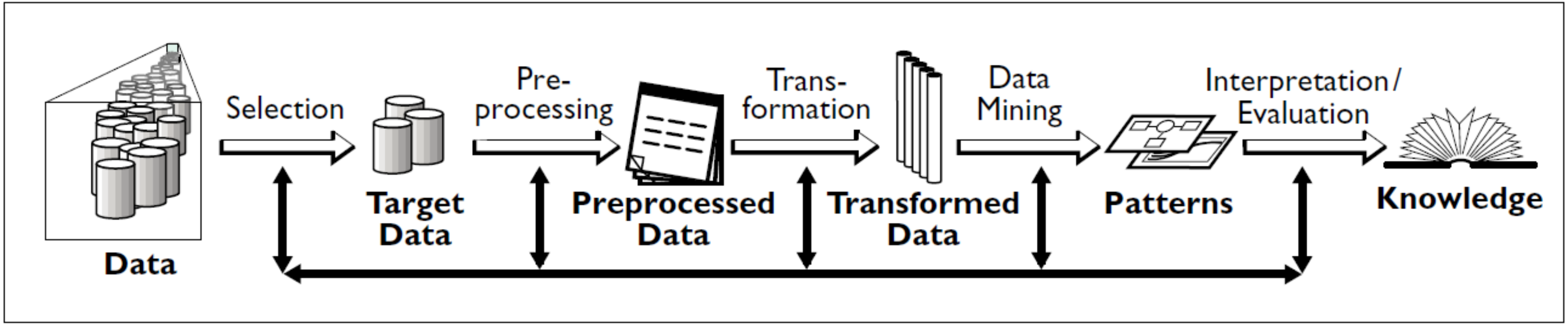
- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data

What is not Data Mining?

- Retrieve data instead of knowledge or pattern
- Not interesting (trivial, explicit, known, useless)

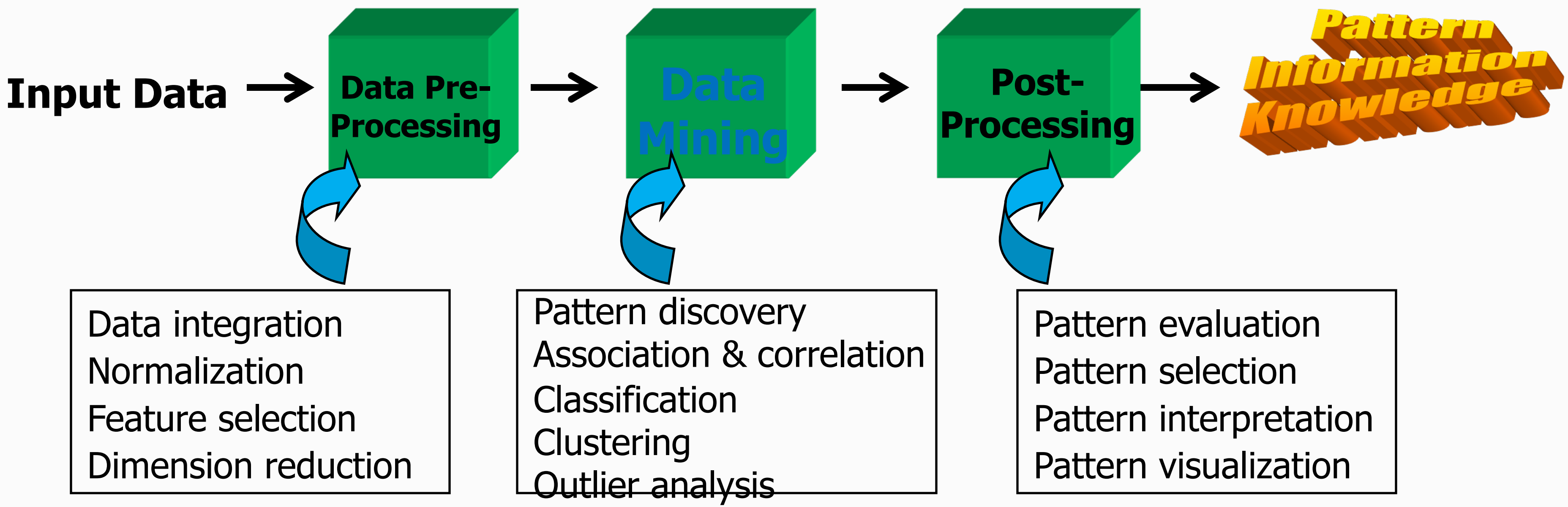
Knowledge Discovery (KDD) Process

- Data mining plays an essential role in the knowledge discovery process



http://cacm.acm.org/magazines/1996/11/8517-the-kdd-process-for-extracting-useful-knowledge-from-volumes-of-data/abstract

KDD Process: A Typical View from ML and Statistics



Input Data → Data Pre-Processing → Data Mining → Post-Processing → Pattern Information Knowledge

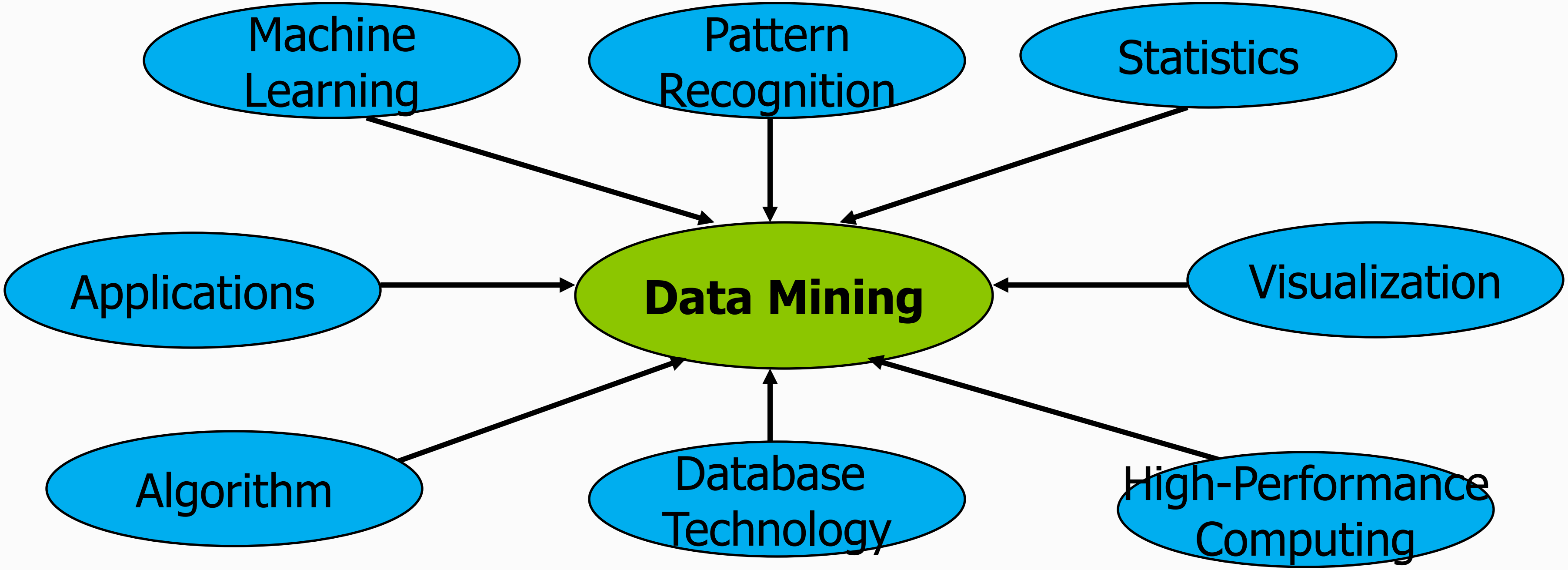
Data integration
Normalization
Feature selection
Dimension reduction

Pattern discovery
Association & correlation
Classification
Clustering
Outlier analysis

Pattern evaluation
Pattern selection
Pattern interpretation
Pattern visualization

This is a view from typical machine learning and statistics communities

Data Mining: Confluence of Multiple Disciplines



Data Mining Software

Free, open-source

- RapidMiner
- Weka: Data mining tool in java
- SCaVis: scientific computation and visualization, Java
- Orange: Python suite
- Scikit-learn: Python machine learning lbirary
- NumPy/SciPy/Ipython/ mlpy (python modules for scientific computing, scientific library, interactive computing, machine learning)
- R: statistical computing and graphic
- RattleGUI: data mining GUI using R
- Octave: numerical analysis
- Shogun: machine learning toolkit in C++

Text Mining Tools

- NLTK (NLP Toolkit): NLP suite for Python
- SenticNet API: sentiment analysis
- Stanford NLP software
- UIMA

Large-Scale Data Processing, Machine Learning

- Apache Mahout
- GraphLab
- MapReduce/Hadoop
- Spark
- Pregel/Giraph

Commercial Products

- Matlab
- Oracle Data Mining
- SAS
- IBM SPSS
- Microsoft SQL Server Analysis Services
- HP Vertica