Data and Data Mining

CSE 5334 Data Mining, Spring 2020

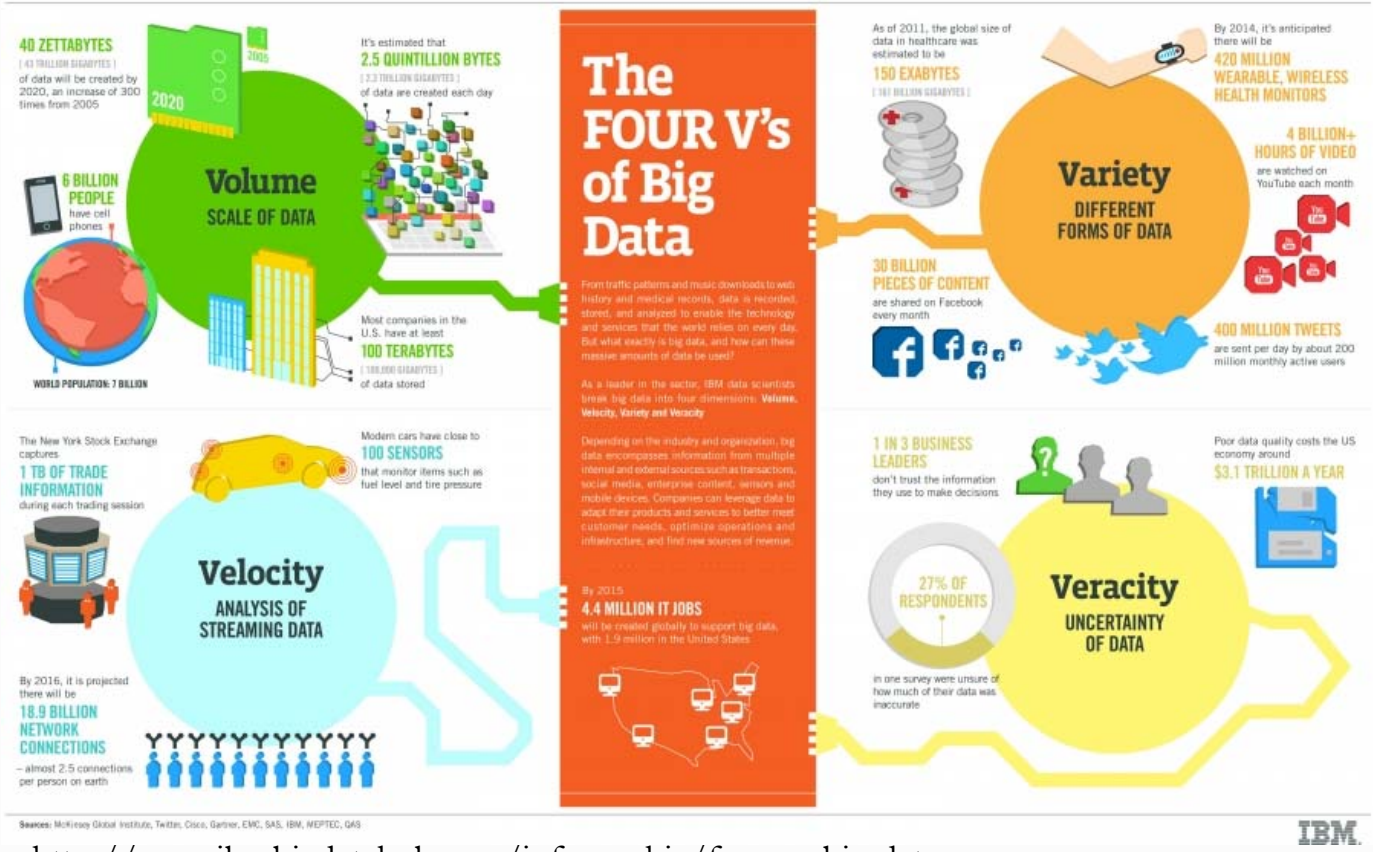
Won Hwa Kim

(Slides courtesy of Pang-Ning Tan, Michael Steinbach and Vipin Kumar)

Big Data

http://dilbert.com/strip/2012-07-29

Big Data



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.

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/

Variety: Types of Data

Structured data

- (relational) database tables
- CSV/TSV files

Name	FName	City	Age	Salary
Smith	John	3	35	$280
Doe	Jane	1	28	$325
Brown	Scott	3	41	$265
Howard	Shemp	4	48	$359
Taylor	Tom	2	22	$250

Semi-structured data

- XML, JSON, RDF

Unstructured data

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

Multimedia data

- images, videos, audios

```
[['The', 'motto', 'originated', 'in', 'the', 'StarSpangled', 'Banner', ''
, 'Tell', 'me', 'that', 'this', 'has'], ['something', 'to', 'do', 'with',
'atheists'], ['The', 'motto', 'oncoins', 'originated', 'as', 'a', 'McCart
hyite', 'smear', 'which', 'equated', 'atheism'], ['with', 'Communism', 'a
nd', 'called', 'both', 'unamerican'], ['No', 'it', "didn't", '', 'The', '
motto', 'has', 'been', 'on', 'various', 'coins', 'since', 'the', 'Civil',
'War'], ['It', 'was', 'just', 'required', 'to', 'be', 'on', 'all', 'curre
ncy', 'in', 'the', "50's"], ['keith']]
```

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

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

Veracity: uncertain and imprecise data

- Quality and origin of data
- Consistent? Complete? Integrity?
- Untrusted and Uncleaned
- Fake stories

- Lots of cost to justify the data…

Datasets

- Amazon Public Data Sets
- Data.gov
- Linked Open Data, Knowledge Bases, Encyclopedia
- Yahoo! Webscope
- Stanford Large Network Dataset Collection
- UCI Machine Learning Repository
- UCR Time Series Classification/Clustering
- Time Series Data Library http://robjhyndman.com/TSDL/
- KDnuggets Dataset List http://www.kdnuggets.com/datasets/index.html
- KDD Cup Datasets http://www.sigkdd.org/kddcup/index.php

Amazon Public Data Sets

http://aws.amazon.com/public-data-sets/

- NASA NEX: A collection of Earth science data sets maintained by NASA, including climate change projections and satellite images of the Earth's surface
- Common Crawl Corpus: A corpus of web crawl data composed of over 5 billion web pages
- 1000 Genomes Project: A detailed map of human genetic variation
- Google Books Ngrams: A data set containing Google Books n-gram corpuses
- US Census Data: US demographic data from 1980, 1990, and 2000 US Censuses
- Freebase Data Dump: A data dump of all the current facts and assertions in the Freebase system, an open database covering millions of topics

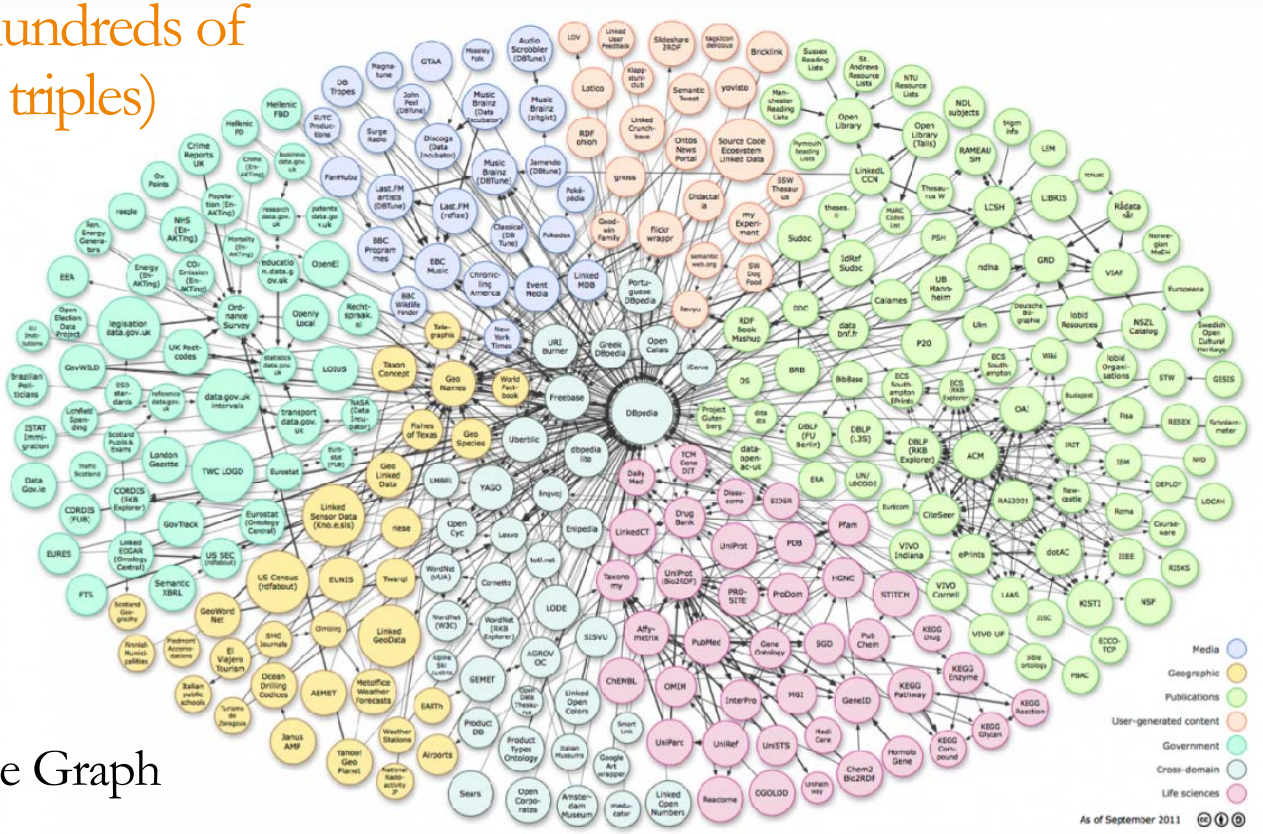
Data.gov

http://www.data.gov/ (137,608 datasets)

- Consumer Complaint Database
- U.S. International Trade in Goods and Services: Monthly report that provides national trade data including imports, exports, and balance of payments for goods and services.
- DTV Reception Maps
- Food Access Research Atlas — presents a spatial overview of food access indicators for low-income and other census tracts using different measures of supermarket...
- U.S. Hourly Precipitation Data
- Great Chile Earthquake of May 22, 1960
- Consumer Expenditure Survey
- Farmers Markets Geographic Data: longitude and latitude, state, address, name, and zip code of Farmers Markets in the United States
- Crimes - 2001 to present (City of Chicago)

Linked Data, Knowledge Bases, Encyclopedia

http://linkeddata.org/ (hundreds of datasets, billions of RDF triples)



IMDB
DBLP
PubMed
Wikipedia, DBpedia
YAGO
Freebase/Google Knowledge Graph

Stanford Large Network Dataset Collection

http://snap.stanford.edu/data/

- Social networks : online social networks, edges represent interactions between people
- Communication networks : email communication networks with edges representing communication
- Citation networks : nodes represent papers, edges represent citations
- Collaboration networks : nodes represent scientists, edges represent collaborations (co-authoring a paper)
- Web graphs : nodes represent webpages and edges are hyperlinks
- Amazon networks : nodes represent products and edges link commonly co-purchased products
- Internet networks : nodes represent computers and edges communication
- Road networks : nodes represent intersections and edges roads connecting the intersections

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

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data

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)

What is Data Mining?

Data mining tasks

- Prediction methods: use variables to make prediction for unknown or future samples

- Description methods: find human-interpretable patterns that describes the data

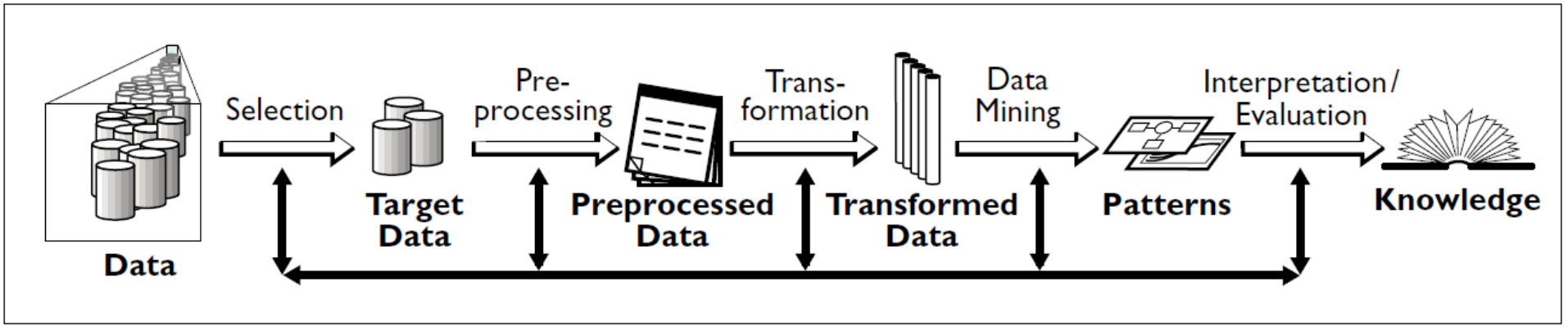
Challenges in Data Mining?

Challenges

- Scalability
- Dimensionality
- Complex and heterogeneousity
- Data quality
- Data ownership and distribution
- Privacy
- Streaming data
- …

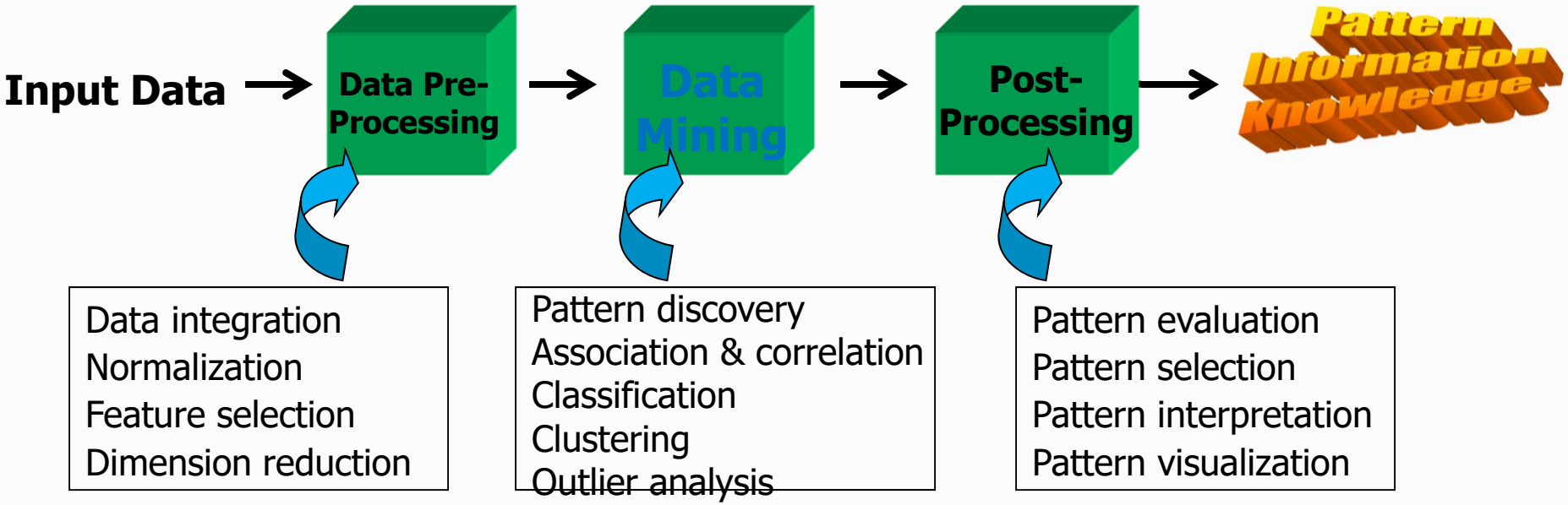
Knowledge Discovery (KDD) Process

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



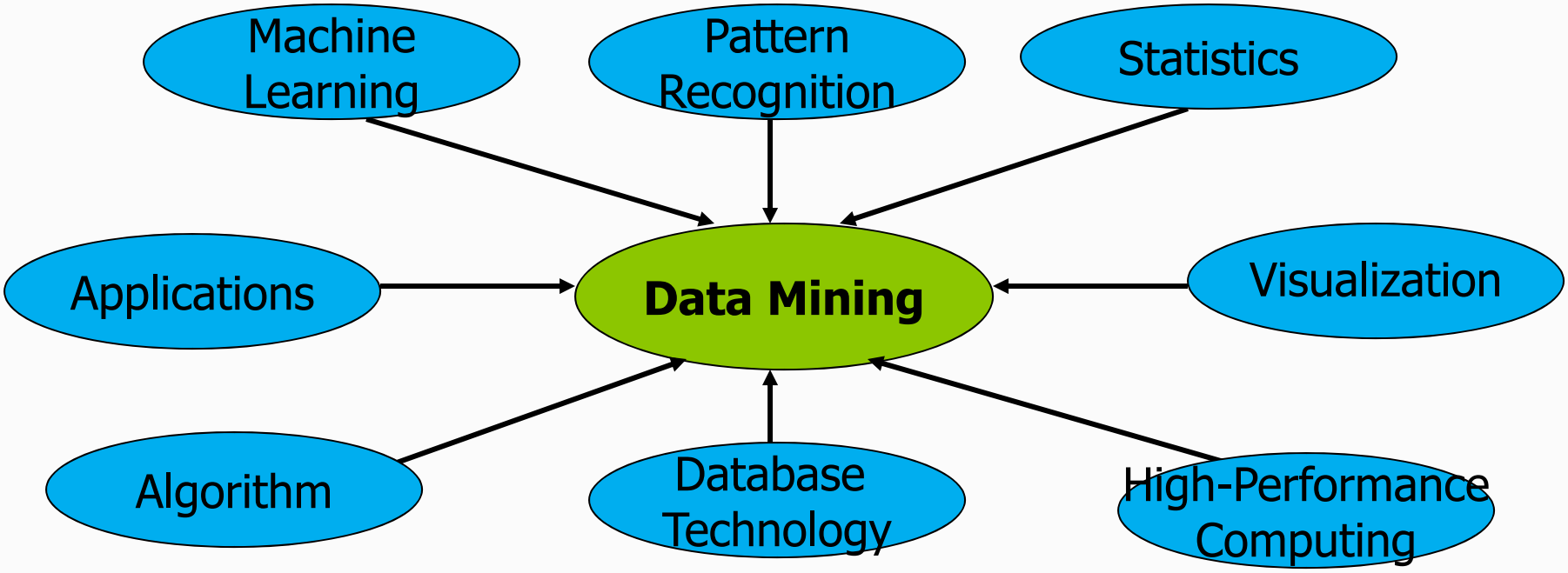
http://cacm.acm.org/magazines/1996/11/8517-the-kdd-process-for-extracting-useful-knowledge-from-volumes-of-data/abstract

KDD Process: A Typical View from ML and Statistics



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