Lambton College

School of Computer Studies

CBD-3335 Data Mining and Analysis

Required Materials

- Data Mining: The Textbook, 1st Edition

  Author: Aggarwal
  ISBN: 978-3319141411
  Publisher: Springer
  Published: April
  2015

- R and Data Mining: Examples and Case Studies, 1st Edition

  Author: Zhao
  ISBN: 978-0123969637
  Publisher: Academic Press
  Published: December 2012

Learning Outcomes

- Data sources, data interpretation and methods of relating data to observations.

- Association pattern mining utilizing a variety of algorithms.

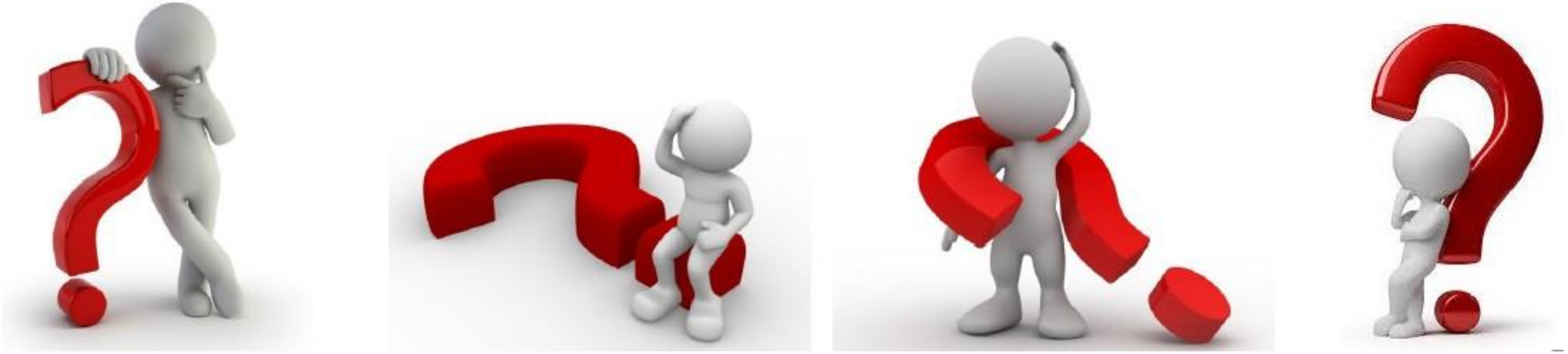
- Analyze clusters and outliers using a variety of methods.

- Perform mining for data streams and text data using algorithms.

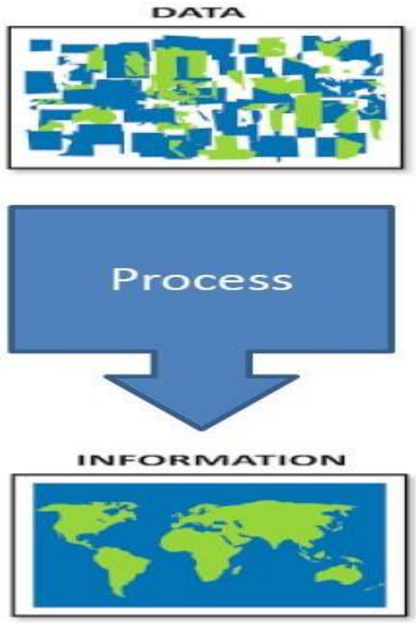
- Apply data mining techniques for time series, spatial data and discrete sequences.

What is Data Mining?

Generally, data mining (sometimes called data or knowledge discovery) is the **process** of analyzing **data** from different perspectives and summarizing it into useful **information** - information that can be used to increase revenue, cuts costs, or both.



What is Data Mining?



DATA

Process

INFORMATION

Which one is a data mining task?

- Look up phone number in phone directory ✗

- Group together similar documents returned by search engine according to their context ✓

- Query a Web search engine for information about “Amazon” ✗

What is Data?

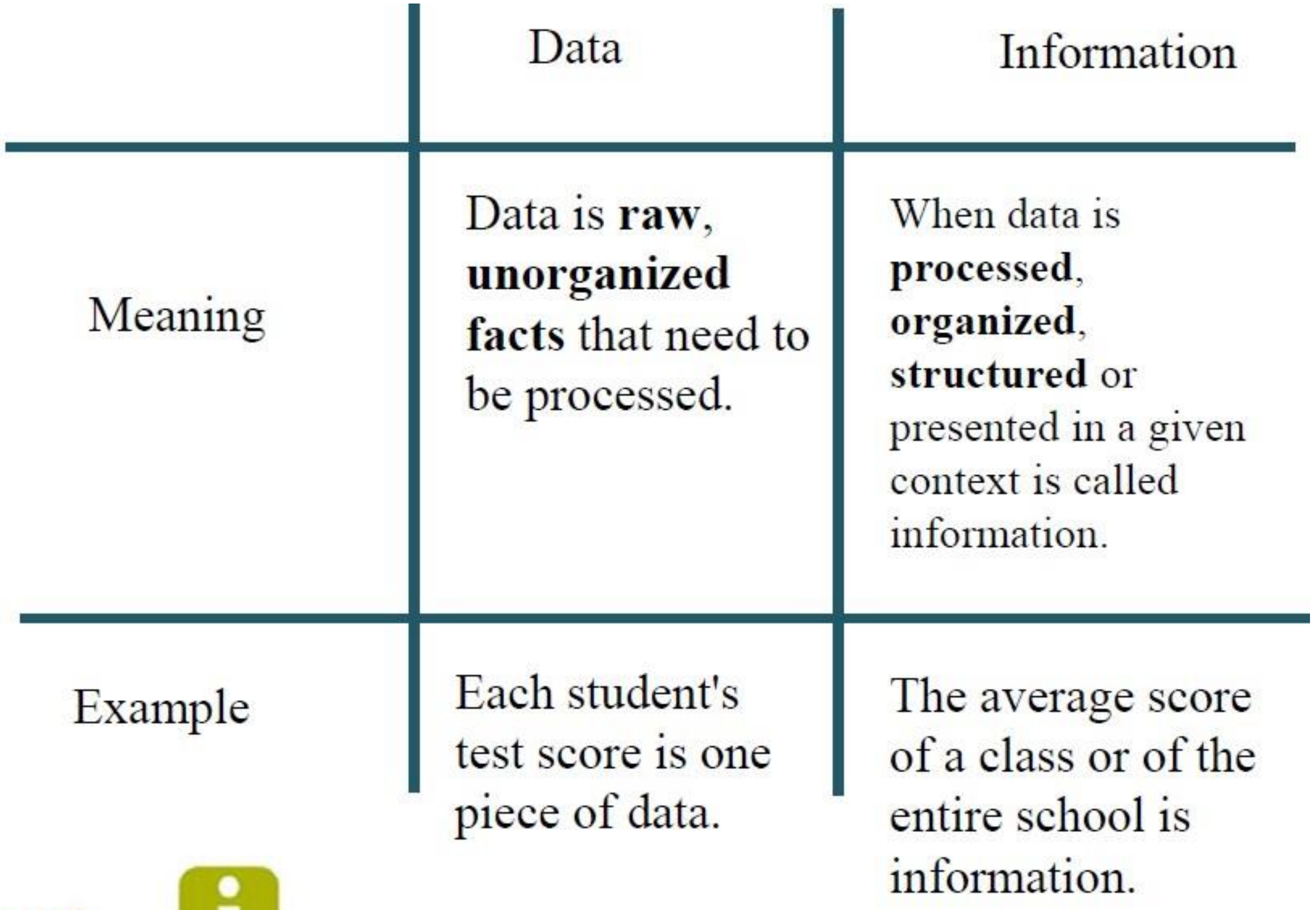
- 100117
  - First day of our course (10/01/17)
  - Average salary of data scientist (100,117)
  - Zip code of a neighborhood in SF

There is story behind every data. Story could be structure, semantic, relations, and so on.

Information from Data

- Summarizing the data
- Averaging the data
- Selecting part of the data
- Graphing the data
- Adding context
- Adding value
- Relationship between data

Data vs. Information



	Data	Information
Meaning	Data is **raw**, **unorganized facts** that need to be processed.	When data is **processed**, **organized**, **structured** or presented in a given context is called information.
Example	Each student's test score is one piece of data.	The average score of a class or of the entire school is information.

Quiz

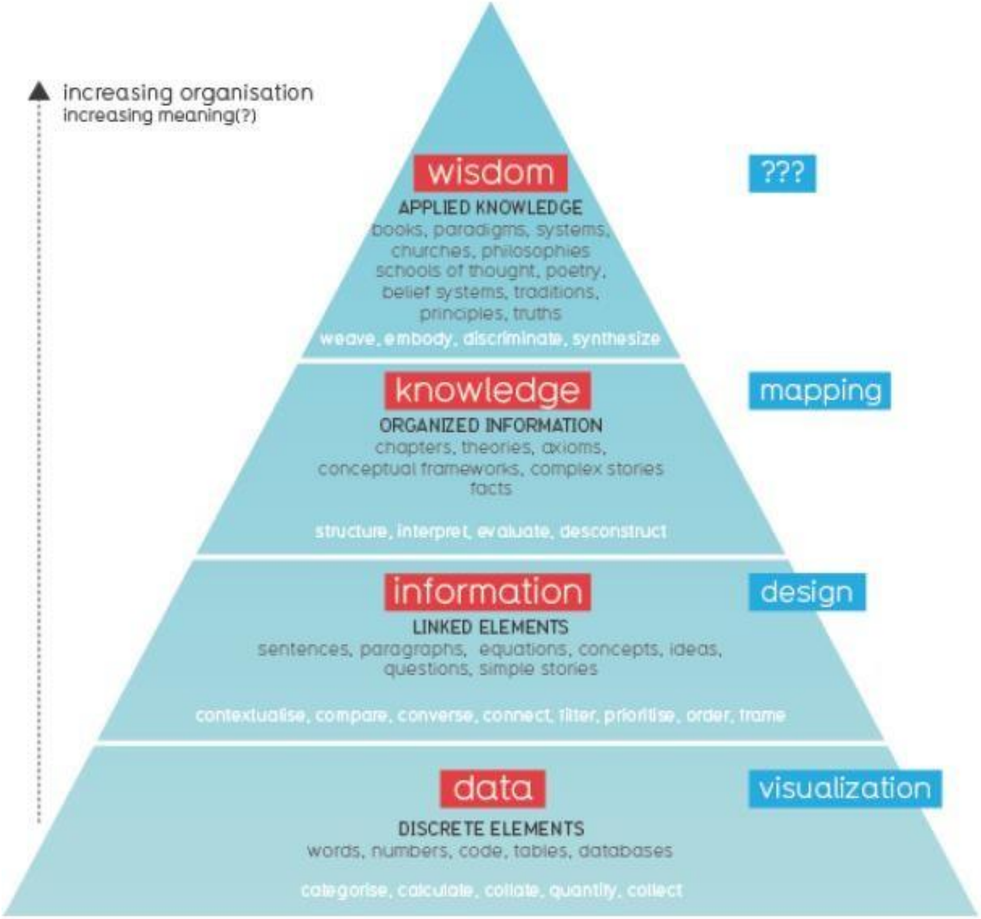
Which of the following is Information?

- Winning time for a race
- [✓ Transcriptionist accuracy]
- Allergies
- Date of birth

Knowledge from Information

- How is the information tied to outcomes?
- Are there any patterns in the information?
- What information is relevant to the problem?
- How does this information affect the system?
- What is the best way to use the information?
- How can we add more value to the information?

Data, Information, Knowledge & Wisdom


increasing organisation
increasing meaning(?)
wisdom
APPLIED KNOWLEDGE
books, paradigms, systems, churches, philosophies schools of thought, poetry, belief systems, traditions, principles, truths
weave, embody, discriminate, synthesize
???
knowledge
ORGANIZED INFORMATION
chapters, theories, axioms, conceptual frameworks, complex stories facts
structure, interpret, evaluate, desconstruct
mapping
information
LINKED ELEMENTS
sentences, paragraphs, equations, concepts, ideas, questions, simple stories
contextualise, compare, converse, connect, filter, prioritise, order, frame
design
data
DISCRETE ELEMENTS
words, numbers, code, tables, databases
categorise, calculate, collate, quantify, collect
visualization

David McCandless // v 0.1 // work in progress
InformationIsBeautiful.net

Data Mining Definition

- Data mining is the process of discovering interesting patterns (or knowledge) from large amount of data.
- The data sources can include
  - Databases and data warehouses
  - Web data, social and traditional media
  - Demographic, financial, political data, purchases at department/grocery stores, Bank/Credit Card transactions
  - Health records, gene expression data, scientific and environmental data
  - Or any data that are streamed into system dynamically.

Streaming data mining

- Extracting knowledge structures from continuous, rapid data.
- A data stream is an ordered sequence of instances that in many applications of data stream mining can be read only once or a small number of times using limited computing and storage capabilities.

Data analytics

There are four general categories of analytics that are distinguish by the results they produce:

- Descriptive analytics
- Diagnostic analytics
- Predictive analytics
- Prescriptive analytics

Value and Complexity

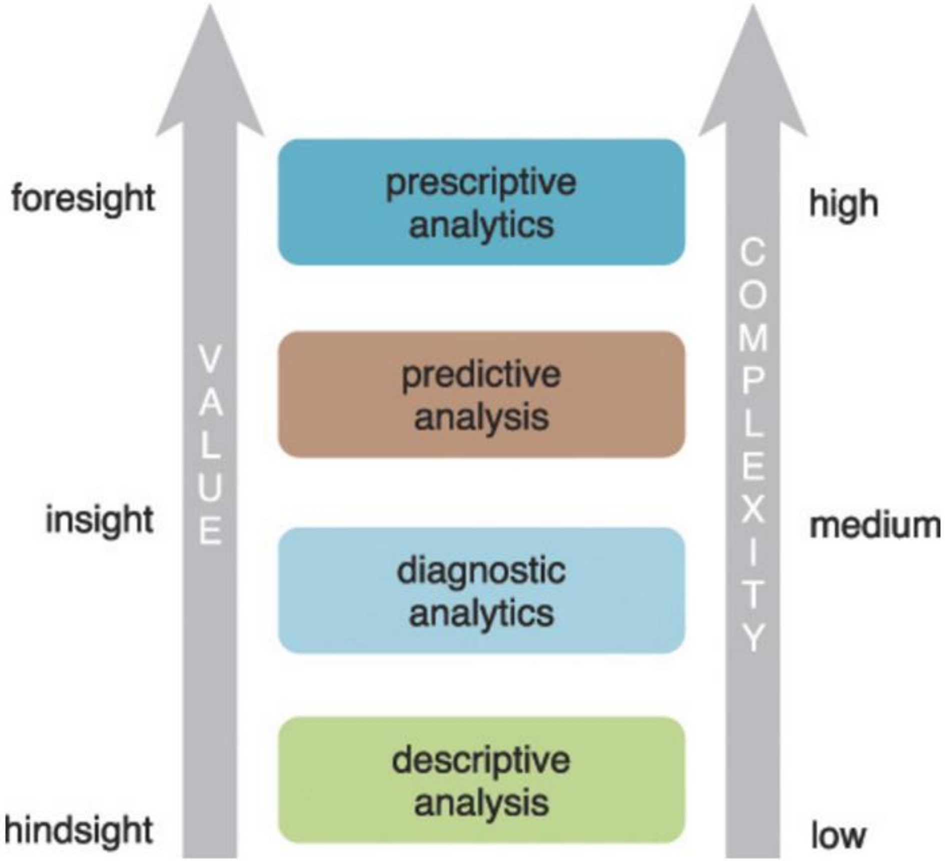


Figure 1.4 Value and complexity increase from descriptive to prescriptive analytics.

Descriptive Analytics

Descriptive analytics are carried out to answer questions about events that have already occurred. This form of analytics contextualizes data to generate information.

Sample questions can include:

- What was the sales volume over the past 12 months?
- What is the number of support calls received as categorized by severity and geographic location?
- What is the monthly commission earned by each sales agent?

Descriptive Analysis

It is estimated that 80% of generated analytics results are descriptive in nature. Value-wise, descriptive analytics provide the least worth and require a relatively basic skillset.

Diagnostic Analytics

Diagnostic analytics aim to determine the cause of a phenomenon that occurred in the past using questions that focus on the reason behind the event. The goal of this type of analytics is to determine what information is related to the phenomenon in order to enable answering questions that seek to determine why something has occurred.

Such questions include:

- Why were a specific sales less than the other sales?
- Why have there been more support calls originating from the Eastern region than from the Western region?
- Why was there an increase in patient re-admission rates over the past three months?

Diagnostic analytics results are viewed via interactive visualization tools that enable users to identify trends and patterns. The executed queries are more complex compared to those of descriptive analytics and are performed on multi-dimensional data held in analytic processing systems.

Predictive Analytics

Predictive analytics are carried out in an attempt to determine the outcome of an event that might occur in the future. With predictive analytics, information is enhanced with meaning to generate knowledge that conveys how that information is related. The strength and magnitude of the associations form the basis of models that are used to generate future predictions based upon past events. It is important to understand that the models used for predictive analytics have implicit dependencies on the conditions under which the past events occurred. If these underlying conditions change, then the models that make predictions need to be updated.

Questions are usually formulated using a what-if rationale, such as the following:
What are the chances that a customer will default on a loan if they have missed a monthly payment?
What will be the patient survival rate if Drug B is administered instead of Drug A?
If a customer has purchased Products A and B, what are the chances that they will also purchase Product C?

Predictive analytics try to predict the outcomes of events, and predictions are made based on patterns, trends and exceptions found in historical and current data. This can lead to the identification of both risks and opportunities.

Prescriptive Analysis

Prescriptive analytics build upon the results of predictive analytics by prescribing actions that should be taken. The focus is not only on which prescribed option is best to follow, but why. In other words, prescriptive analytics provide results that can be reasoned about because they embed elements of situational understanding. Thus, this kind of analytics can be used to gain an advantage or mitigate a risk.

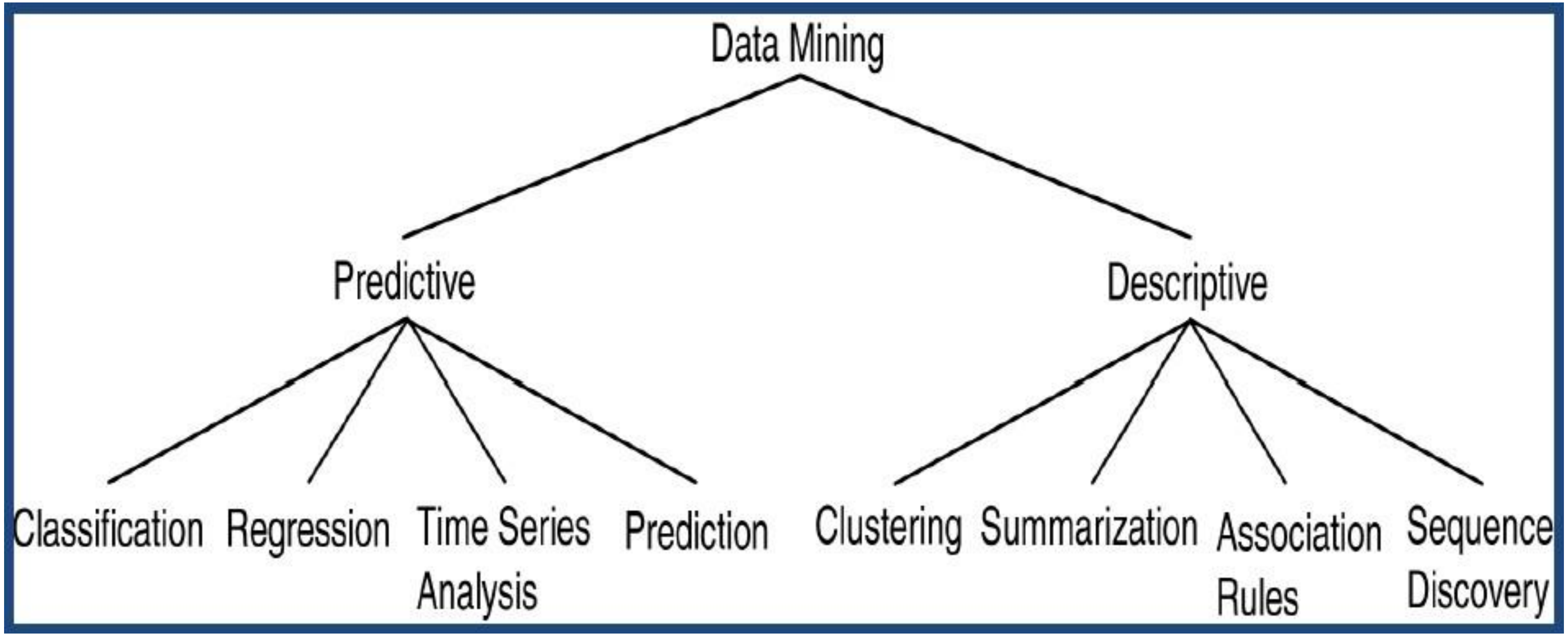
Prescriptive Analysis

Sample questions may include:

- Among three drugs, which one provides the best results?
- When is the best time to trade a particular stock?

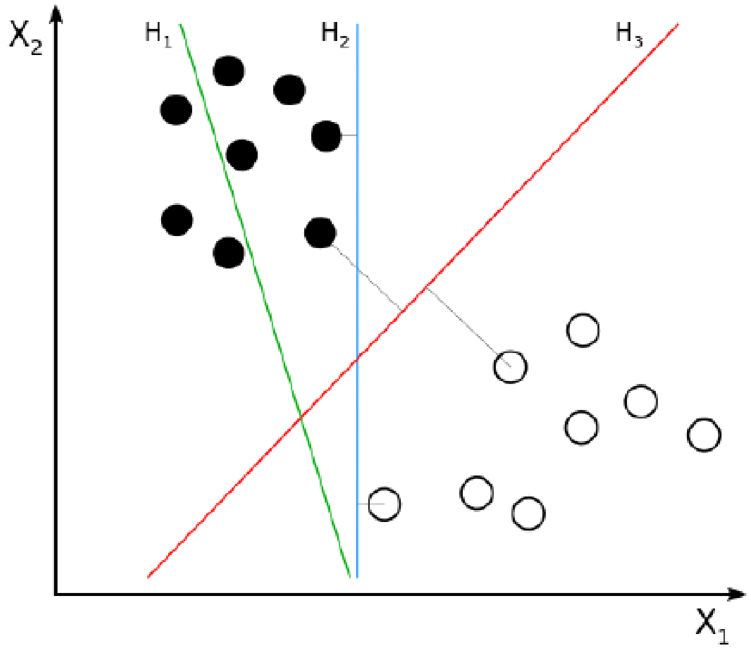
Prescriptive analytics provide more value than any other type of analytics and correspondingly require the most advanced skillset, as well as specialized software and tools. Various outcomes are calculated, and the best course of action for each outcome is suggested. The approach shifts from explanatory to advisory and can include the simulation of various scenarios.

Data Mining Tasks



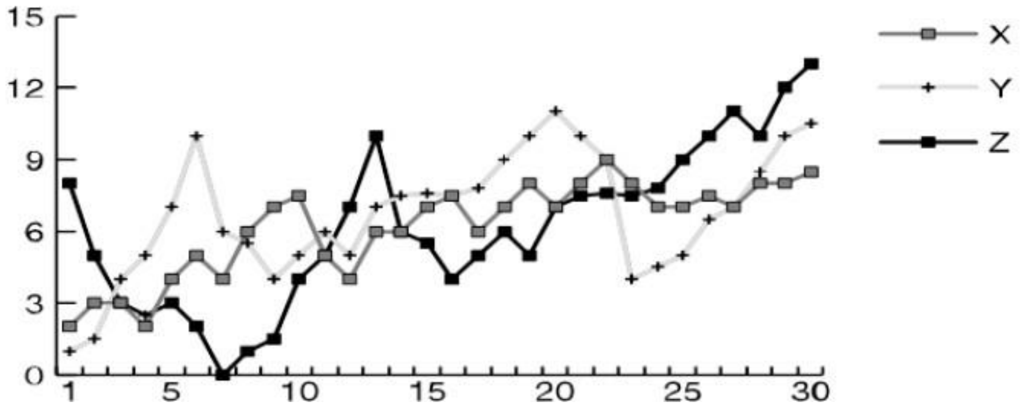
Classification

- Maps data into predefined groups or classes
  - (Semi)supervised learning
  - Pattern recognition
  - Prediction

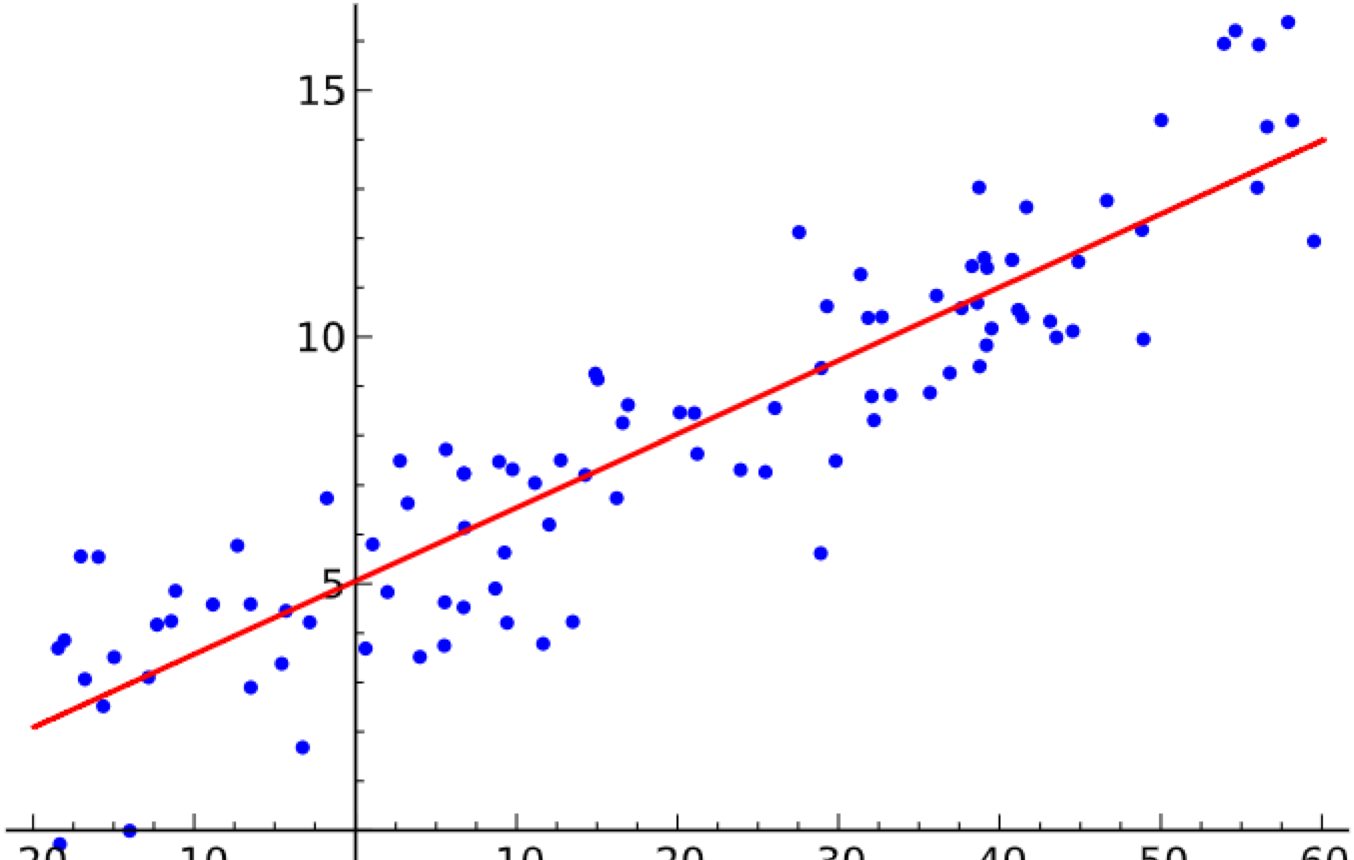


Regression

- Maps a data item to a real valued prediction variable.
- Example: Time series analysis (Stock Market)
  - Predict future values
  - Determine similar patterns over time
  - Classify behavior

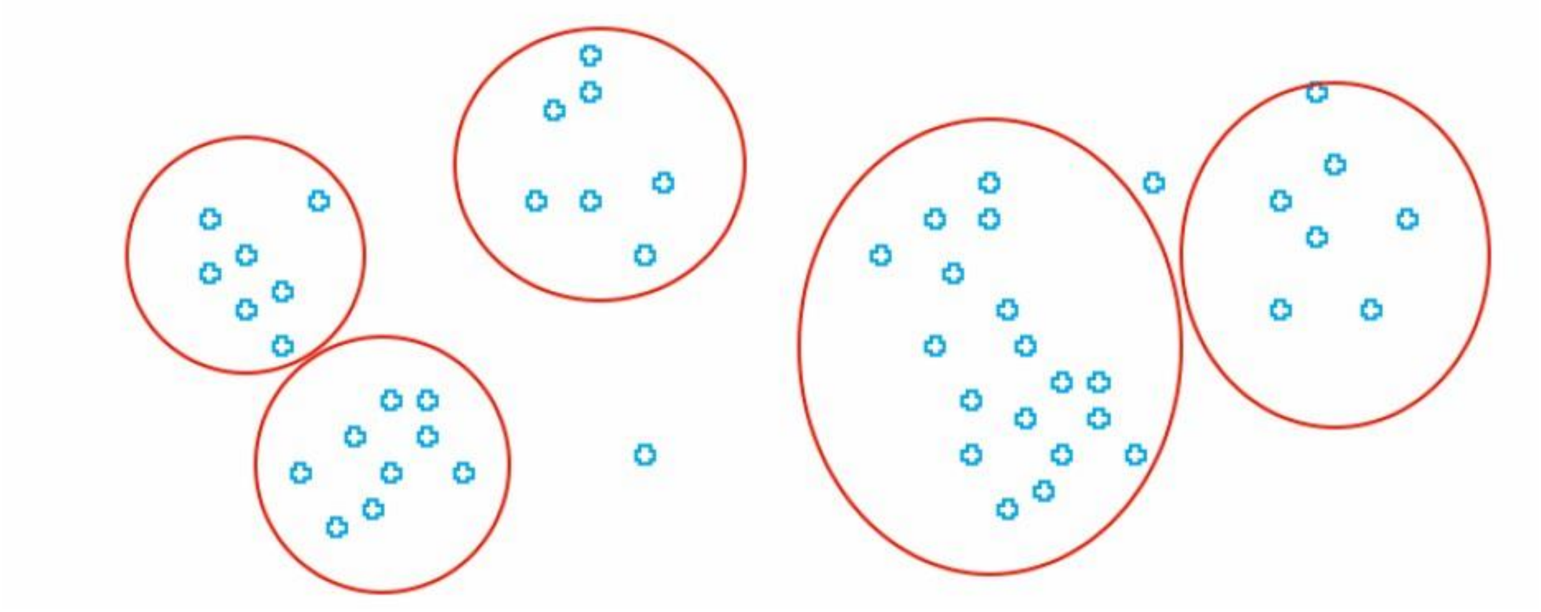


Regression



Clustering

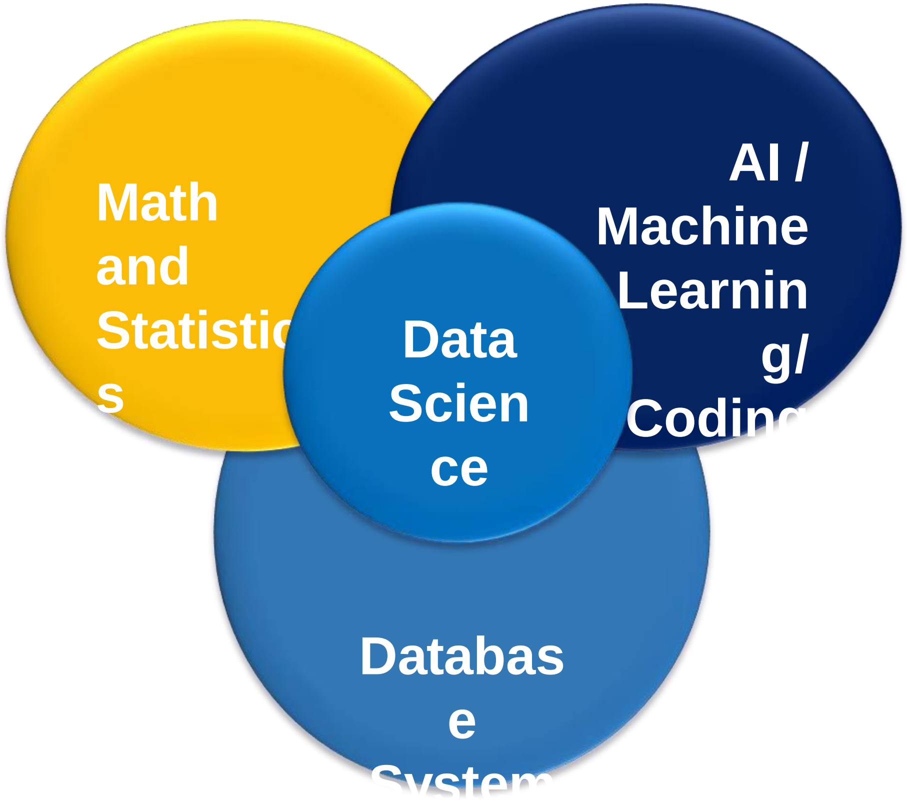
- Groups similar data together into clusters.
  - Unsupervised learning
  - Segmentation
  - Partitioning



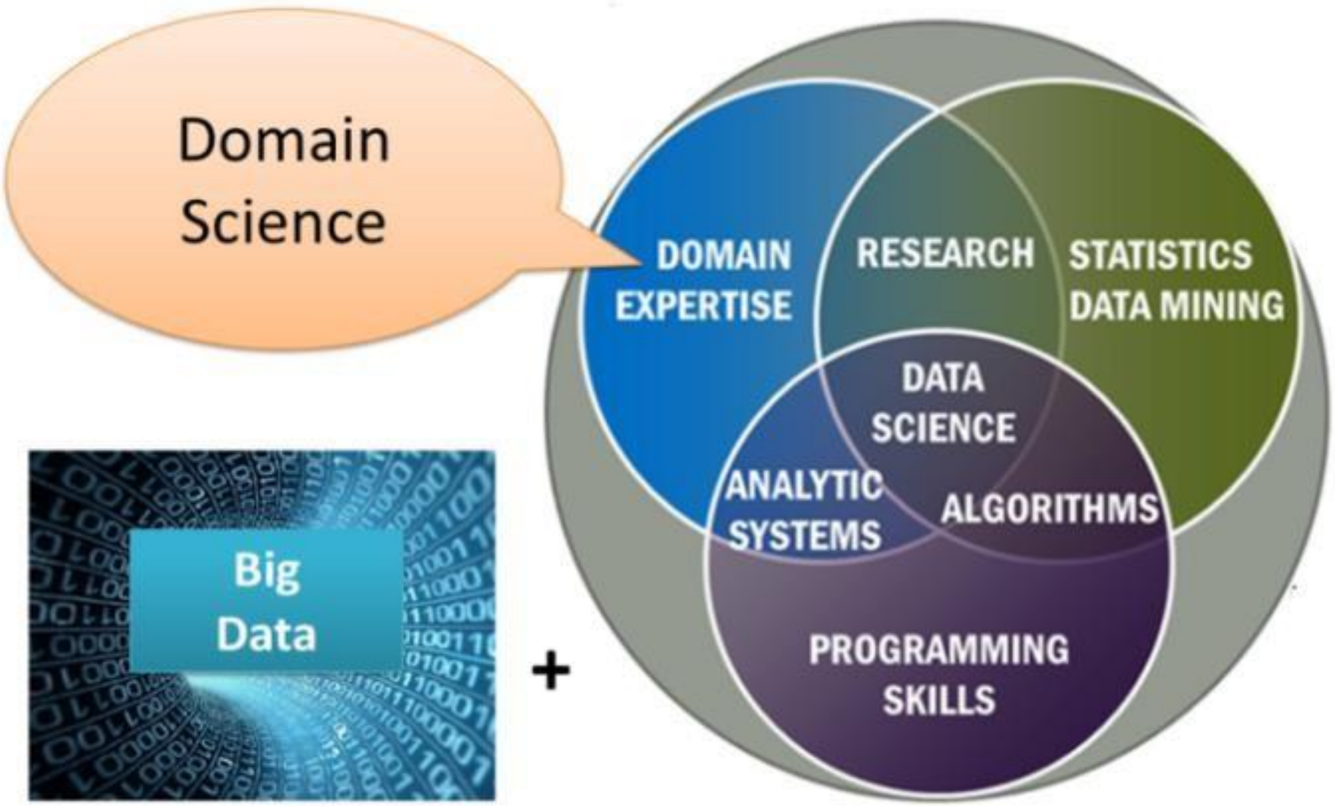
Are all patterns interesting?

- What makes a pattern interesting?
  - valid: hold on new data with some certainty
  - novel: non-obvious to the system
  - useful: should be possible to act on the item
  - understandable: humans should be able to interpret the pattern

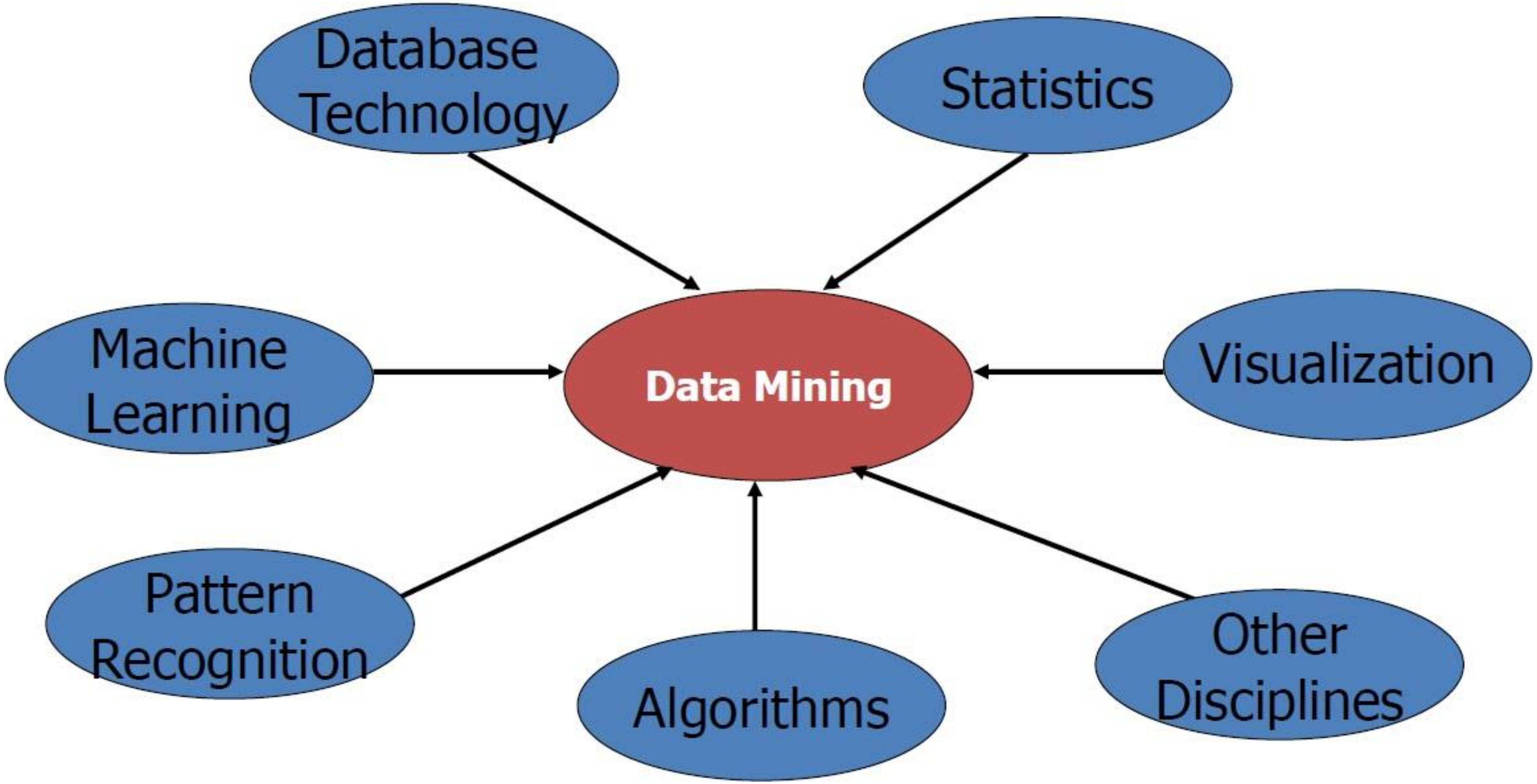
Data Science Concepts



Data Science Concepts



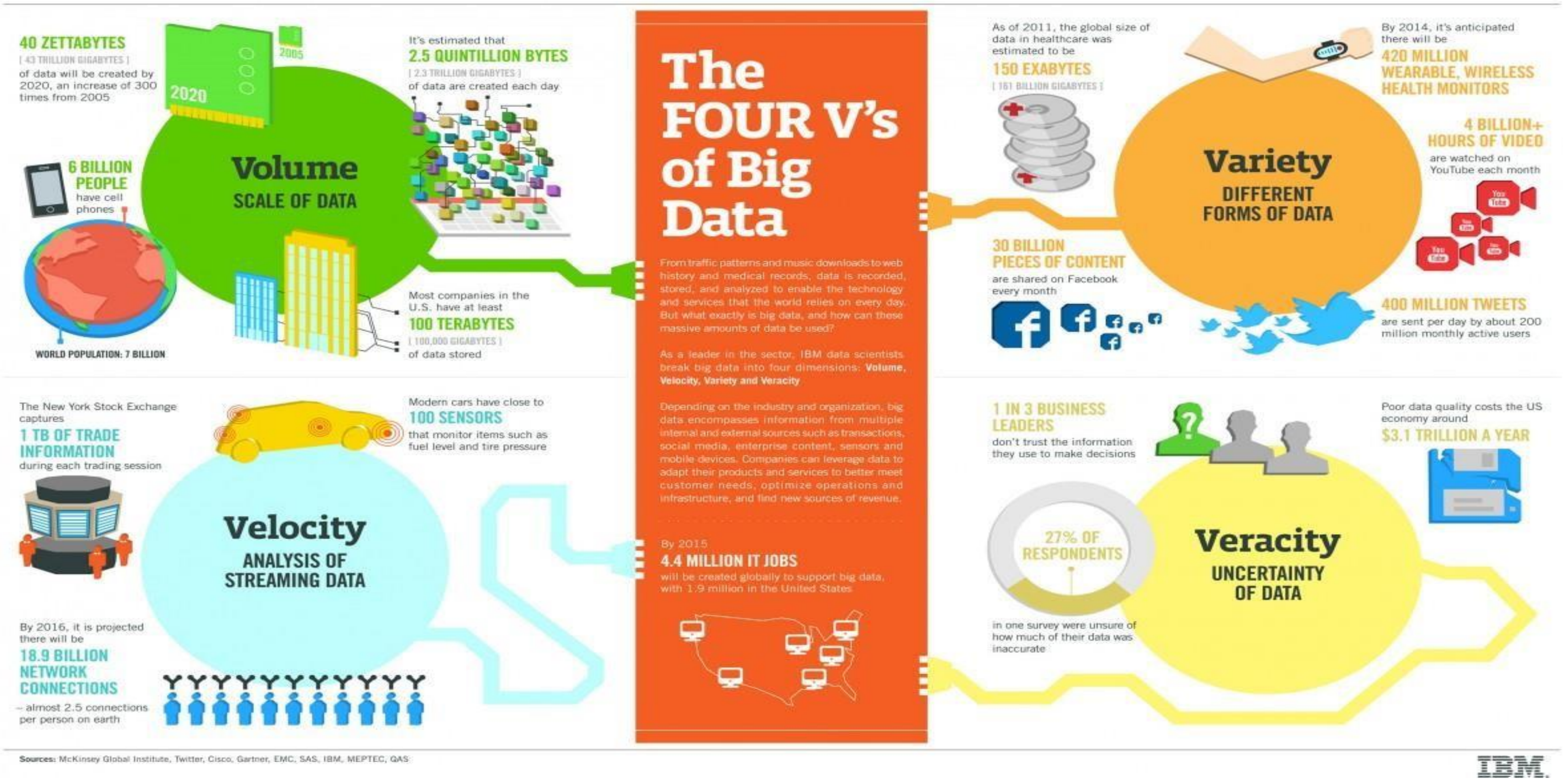
Related Disciplines



Other Related or Similar Fields

- Data science
- Data analytics
- Artificial intelligence
- Information extraction
- Natural language processing
- Computational linguistics
- Text and web mining
- Search and information retrieval
- Social network analysis
- Graph theory and network science
- Recommender systems
- Link mining

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

Big Data

- A collection of very large and complex data sets very difficult to be handled by conventional database management systems and classical data processing and mining techniques.
- Four characteristics of big data
  - Volume: Constantly increasing data volume
  - Velocity: High speed data in and out (sensors)
  - Variety: Very diverse range of data types and resources
  - Veracity: Big data as honest data
- Every day more than 2.5 quintillion (2.5×10^{18}) bytes of data are created (in 2013)

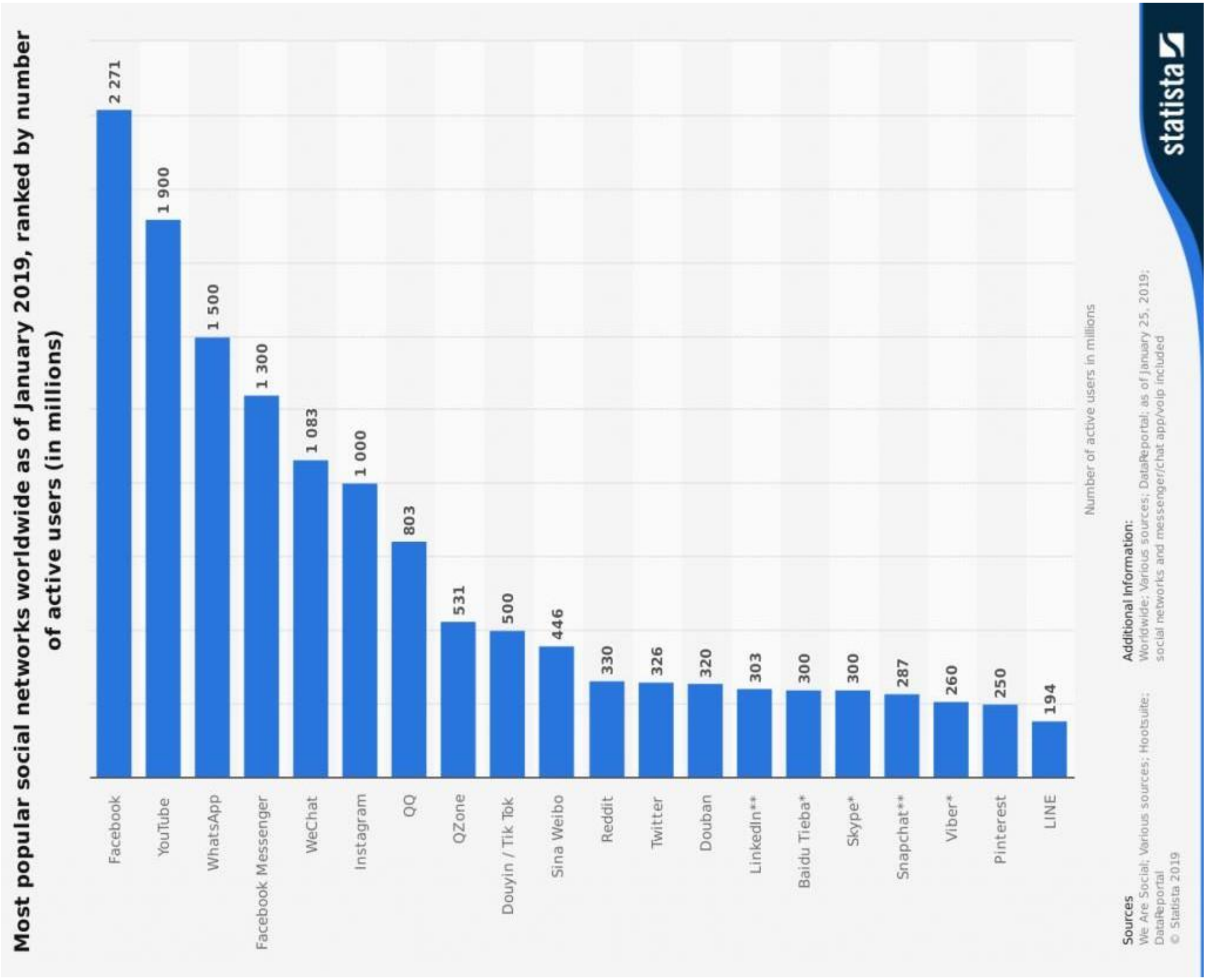
Big Data

- Facebook
  - Daily active users: 968 million
  - Monthly active users: 1.393 billion
  - 751 million mobile users access Facebook every month
- Twitter
  - 44% growth from June 2012 to March 2013
  - 700 million monthly active users
  - 21% of the world's internet population are using Twitter every month (in 2013)
  - Over 500 million registered accounts

Big Data

- YouTube
  - 1 billion unique monthly visitors
  - 6 billion hours of videos are watched every month
- Google+
  - 359 million monthly active users
- LinkedIn
  - Over 200 million users
  - 2 new users join it every second
  - 64% of users are outside the USA

Big Data


Most popular social networks worldwide as of January 2019, ranked by number of active users (in millions)
Facebook
2 271
YouTube
1 900
WhatsApp
1 500
Facebook Messenger
1 300
WeChat
1 083
Instagram
1 000
QQ
803
QZone
531
Douyin / Tik Tok
500
Sina Weibo
446
Reddit
330
Twitter
326
Douban
320
LinkedIn**
303
Baidu Tieba*
300
Skype*
300
Snapchat**
287
Viber*
260
Pinterest
250
LINE
194
Number of active users in millions
Sources
We Are Social; Various sources; Hootsuite; DataReportal
© Statista 2019
Additional Information:
Worldwide; Various sources; DataReportal; as of January 25, 2019; social networks and messenger/chat app/voip included
statista

Big Data

- The Large Hadron Collider (LHC)
  - 150 million sensors delivering data 40 million times per second.
  - There are nearly 600 million collisions per second. After filtering and refraining from recording more than 99.999% of these streams, there are 100 collisions of interest per second (less than 0.001%).
  - The data flow would exceed 150 million petabytes annual rate.

Big Data

- As a result, only working with less than 0.001% of the sensor stream data, the data flow from all four LHC experiments represents 25 petabytes annual rate before replication (as of 2012). This becomes nearly 200 petabytes after replication.

LHC:
https://www.youtube.com/watch?v=bTHzB4h0po4

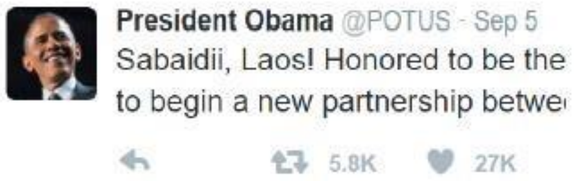
Big Data

- If all sensor data were to be recorded in LHC, the data flow would be extremely hard to work with. The data flow would exceed 150 million petabytes annual rate, or nearly 500 exabytes (500×10^{18}) per day, before replication. To put the number in perspective, this is equivalent to 500 quanitilion (5×10^{20}) bytes per day, almost 200 times higher than all the other sources combined in the world.

Big Data

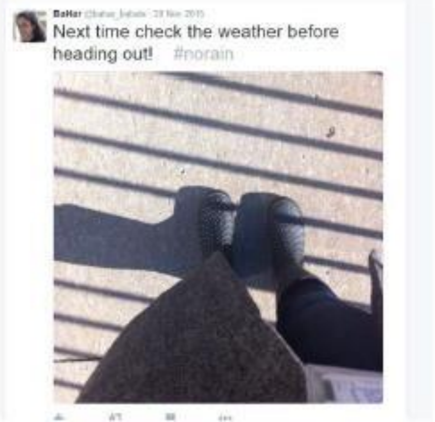
Twitter...

Variety of People

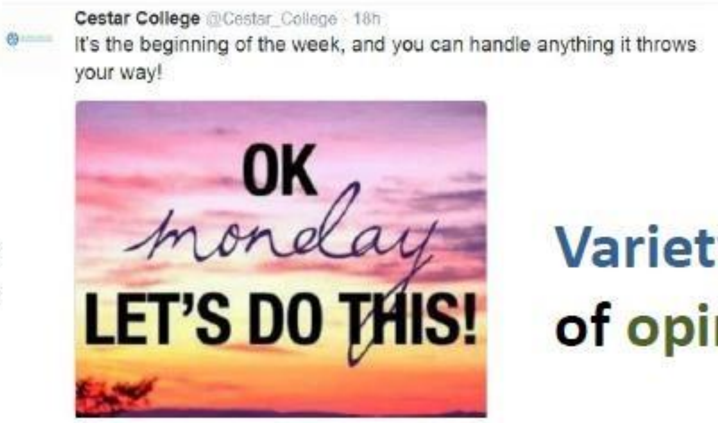


President Obama @POTUS · Sep 5
Sabaidii, Laos! Honored to be the
to begin a new partnership betwe

5.8K 27K

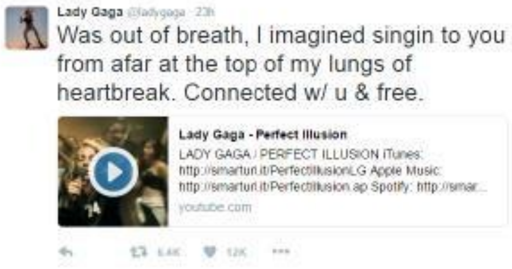


BaHar
Next time check the weather before heading out! #norain



Cestar College @Cestar_College · 18h
It's the beginning of the week, and you can handle anything it throws your way!

Variety of opinions

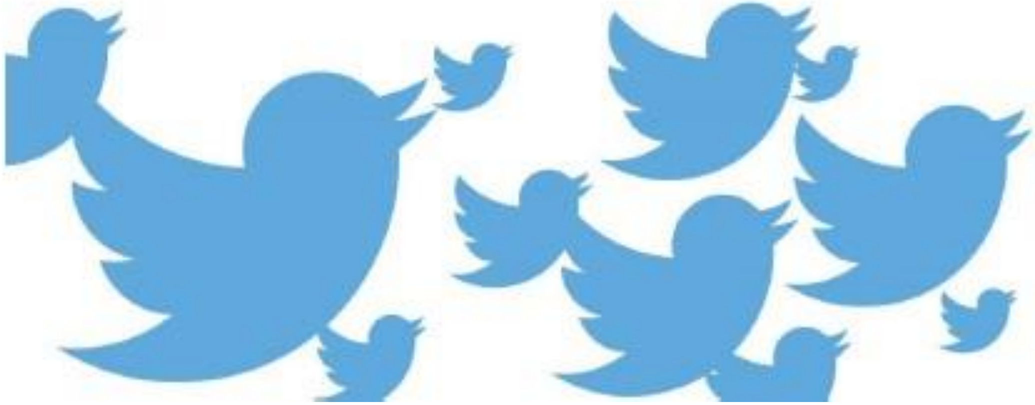


Lady Gaga @ladygaga · 23h
Was out of breath, I imagined singin to you from afar at the top of my lungs of heartbreak. Connected w/ u & free.

Lady Gaga - Perfect Illusion
LADY GAGA / PERFECT ILLUSION iTunes: http://smarturl.it/PerfectIllusionLG Apple Music: http://smarturl.it/PerfectIllusion.ap Spotify: http://smar...
youtube.com

Twitter is useful...

Stock market
Customer satisfaction
Outbreak of diseases
Crime prediction
Politics
Social science
Legal documents
...



Data Mining Challenges

- Scalability
- Dimensionality (curse of dimensionality)
- Complex and Heterogeneous Data
- Data Quality (noise, outliers, missing data, lack of gold or training data)
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data and real-time data
- Ethical issues

Data mining project approach

Don’t be afraid to explore Web. It has all you need to learn

Web-based documentations, Github, Stackoverflow,…

They are your friends and enemies, **You choose**.

Noise vs. Outliers

- Noise is anything that is not the "true" signal.
  - It may have values close to the true signal.

Noise vs. Outliers

- An outlier is something that is much different than the other values.
  - Extreme feature values in one or more dimensions
  - Examples with the same feature values but different labels

- The vast majority of the time outliers are noise bu times a data point that is true signal can be an outli

Noise vs. Outliers

Example: IQ of our class plus Stephen Hawking.

- Hawking would be an outlier even though we accurately measured his IQ.
- If we measured Stephen's IQ as 90, then that would be noise since his real IQ is much higher than that.

Python for Data Mining

Installing Python and Pycharm

- https://www.python.org/downloads/
- https://www.jetbrains.com/pycharm/download/#section=windows (download the free version)

Installing Python and Pycharm

- Install with pip https://packaging.python.org/installing/
  - –python -m pip install -U pip setuptools(installing pip)
  - –pip install -U pip (updating pip)
  - –Pip install “library_name”

- Install with easy_install
  - –easy_install“library_name”

- Install with wheel (.whlfiles)
  - Go to binaries for python packages http://www.lfd.uci.edu/~gohlke/pythonlibs/
  - Pip install wheel
  - pip install PATH+“library_name”.whl

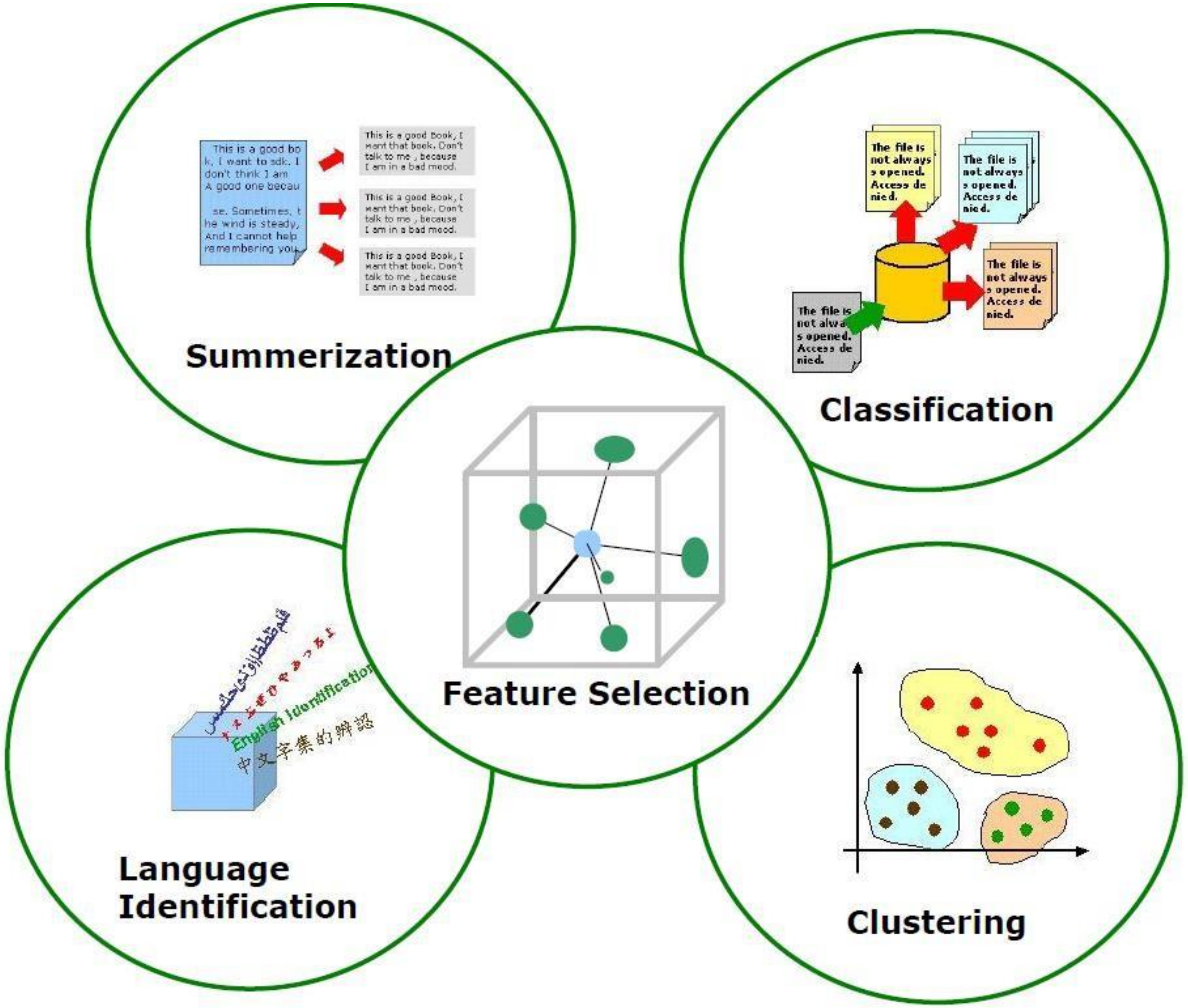
Twitter data collection

- Create your Twitter account
- Got to https://apps.twitter.com/app/new
- Create new app
- Save your credentials

Twitter data collection

- Install tweepy (follow pervious slide)

Text Analysis


This is a good bo k, I want to sdk. I don't think I am A good one becau se. Sometimes, t he wind is steady, And I cannot help remembering you.
This is a good Book, I want that book. Don't talk to me , because I am in a bad mood.
This is a good Book, I want that book. Don't talk to me , because I am in a bad mood.
This is a good Book, I want that book. Don't talk to me , because I am in a bad mood.
Summerization
The file is not always opened. Access de nied.
The file is not always opened. Access de nied.
The file is not always opened. Access de nied.
The file is not always opened. Access de nied.
Classification
Feature Selection
English Identification
中文字集的辨認
Language Identification
Clustering

Further Readings

- Data-Mining Boosts Netflix's Subscriber Base
- The Secret Sauce Behind Netflix's Hit, "House Of Cards": Big Data
- Everything You Wanted to Know About Data Mining but Were Afraid to Ask

- Presentation reference: Cestar College course presentation by Somayyeh (Bahar) Aghababaei