Introduction Data Science

Hendro Margono

Data science

- Data Science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.
- Data science is the same concept as data mining and big data: "use the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problems
- Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.
- It employs techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, and information science

Big Data

- Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage and process the data with low latency.
- Big data has one or more of the following characteristics: high volume, high velocity or high variety. Artificial intelligence (AI), mobile, social and the Internet of Things (IoT) are driving data complexity through new forms and sources of data.
- For example, big data comes from sensors, devices, video/audio, networks, log files, transactional applications, web, and social media — much of it generated in real time and at a very large scale.

Big Data

- Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes.
- Analysis of big data allows analysts, researchers and business users to make better and faster decisions using data that was previously inaccessible or unusable.
- Businesses can use advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics and natural language processing to gain new insights from previously untapped data sources independently or together with existing enterprise data.

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
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Why Data Mining ?

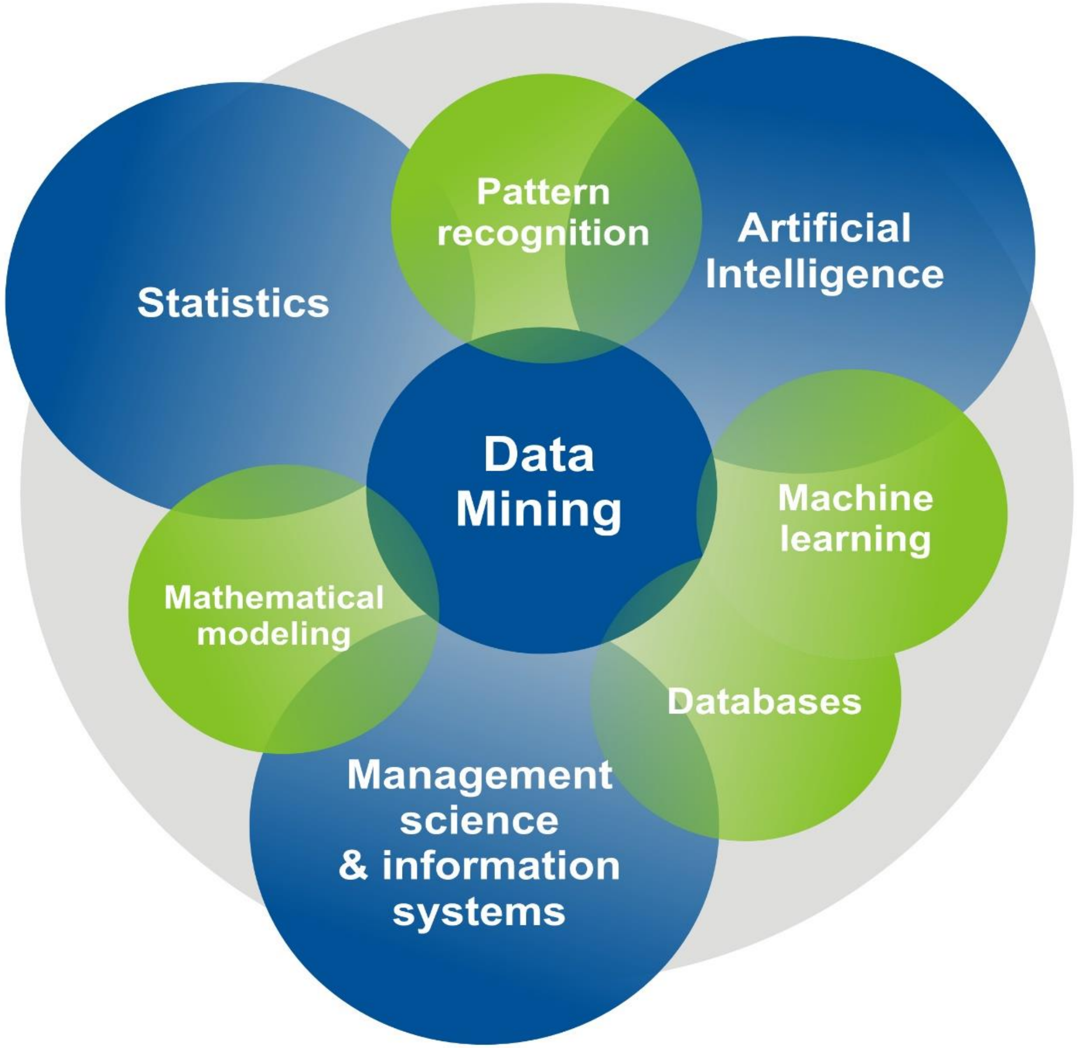
- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
  - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
  - Business: Web, e-commerce, transactions, stocks, …
  - Science: Remote sensing, bioinformatics, scientific simulation, …
  - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets

Why Data Mining ?

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

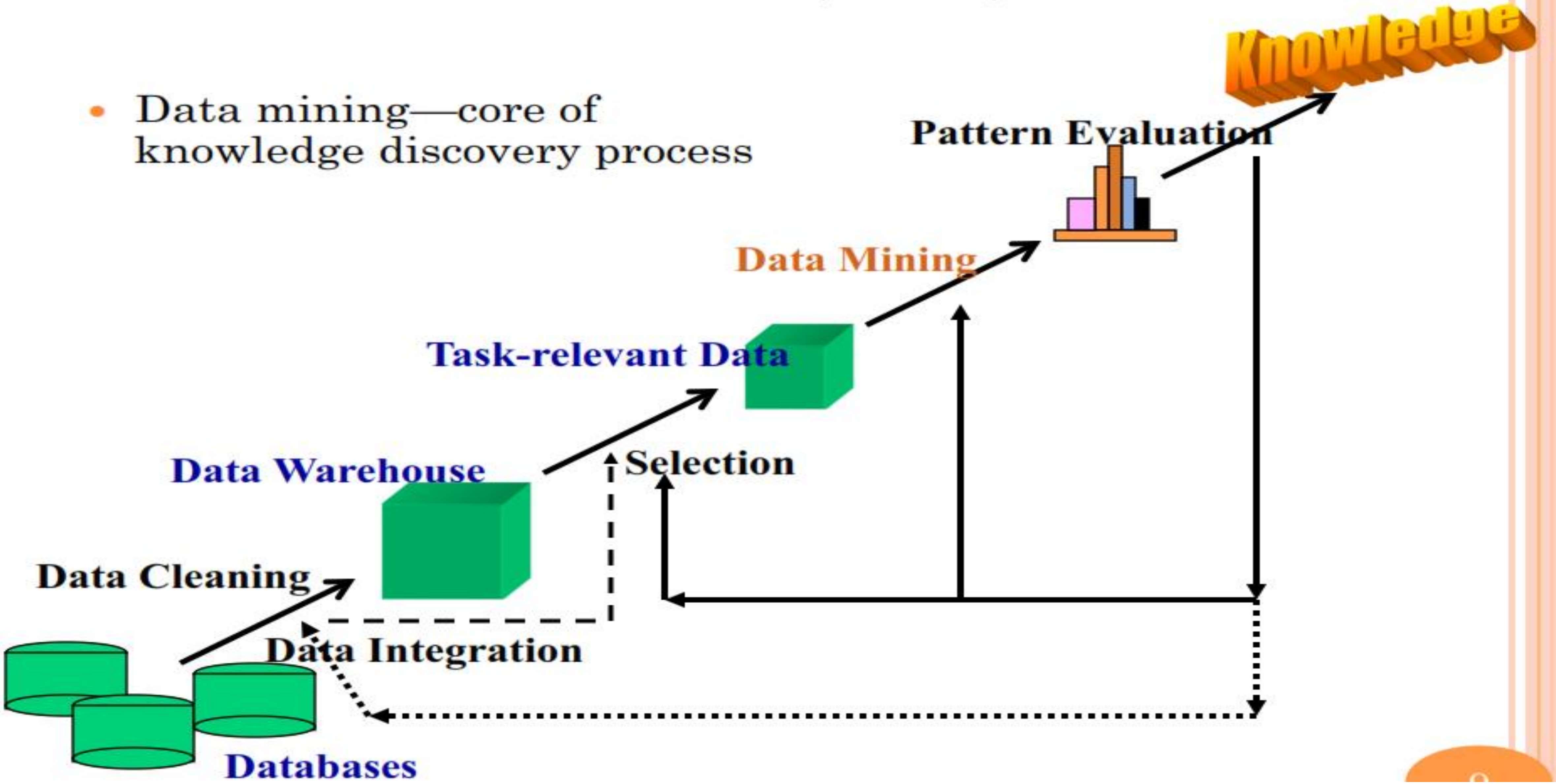
Data Mining

- 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



Knowledge Discovery (KDD) Process

- Data mining—core of knowledge discovery process



Knowledge

Pattern Evaluation

Data Mining

Task-relevant Data

Selection

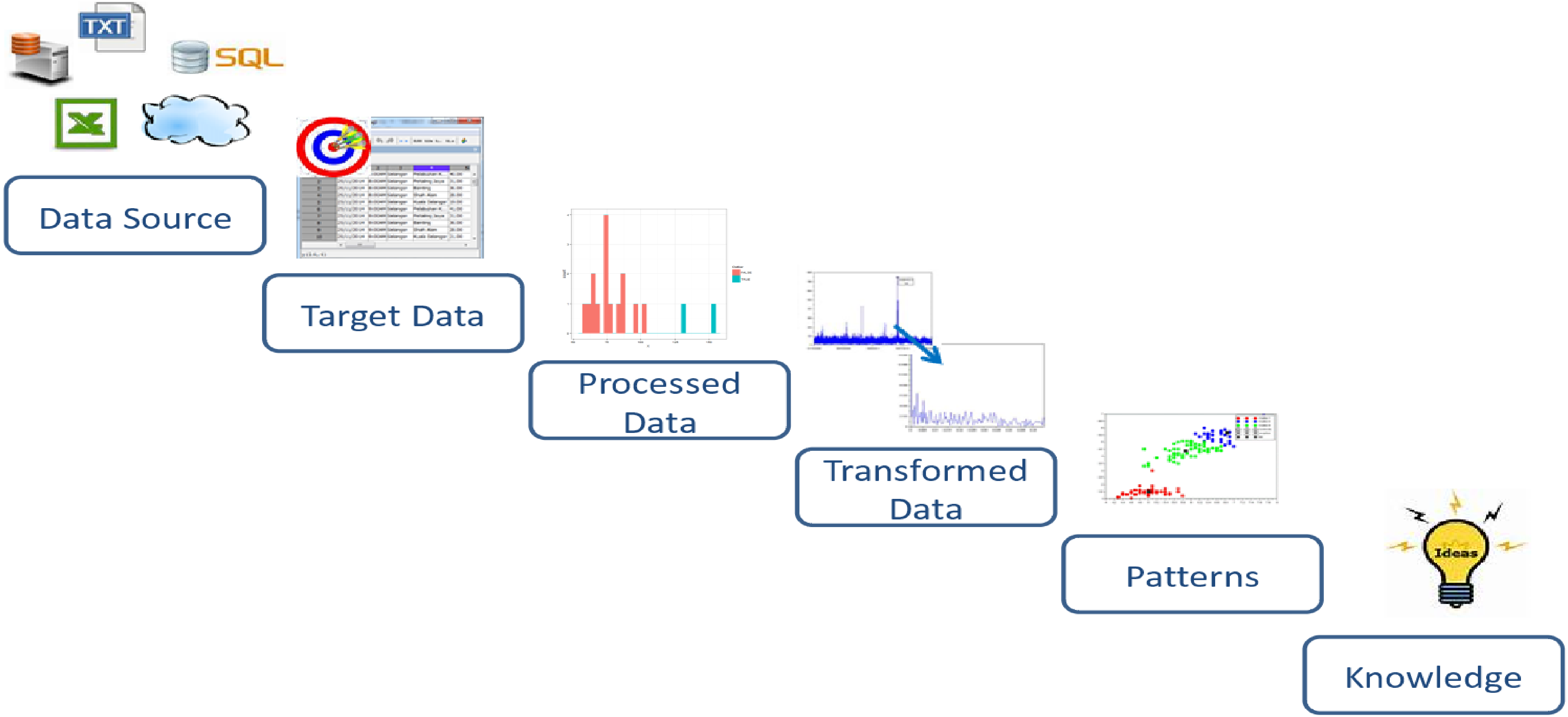
Data Warehouse

Data Cleaning

Data Integration

Databases

Process of Data Mining



Data Source

Target Data

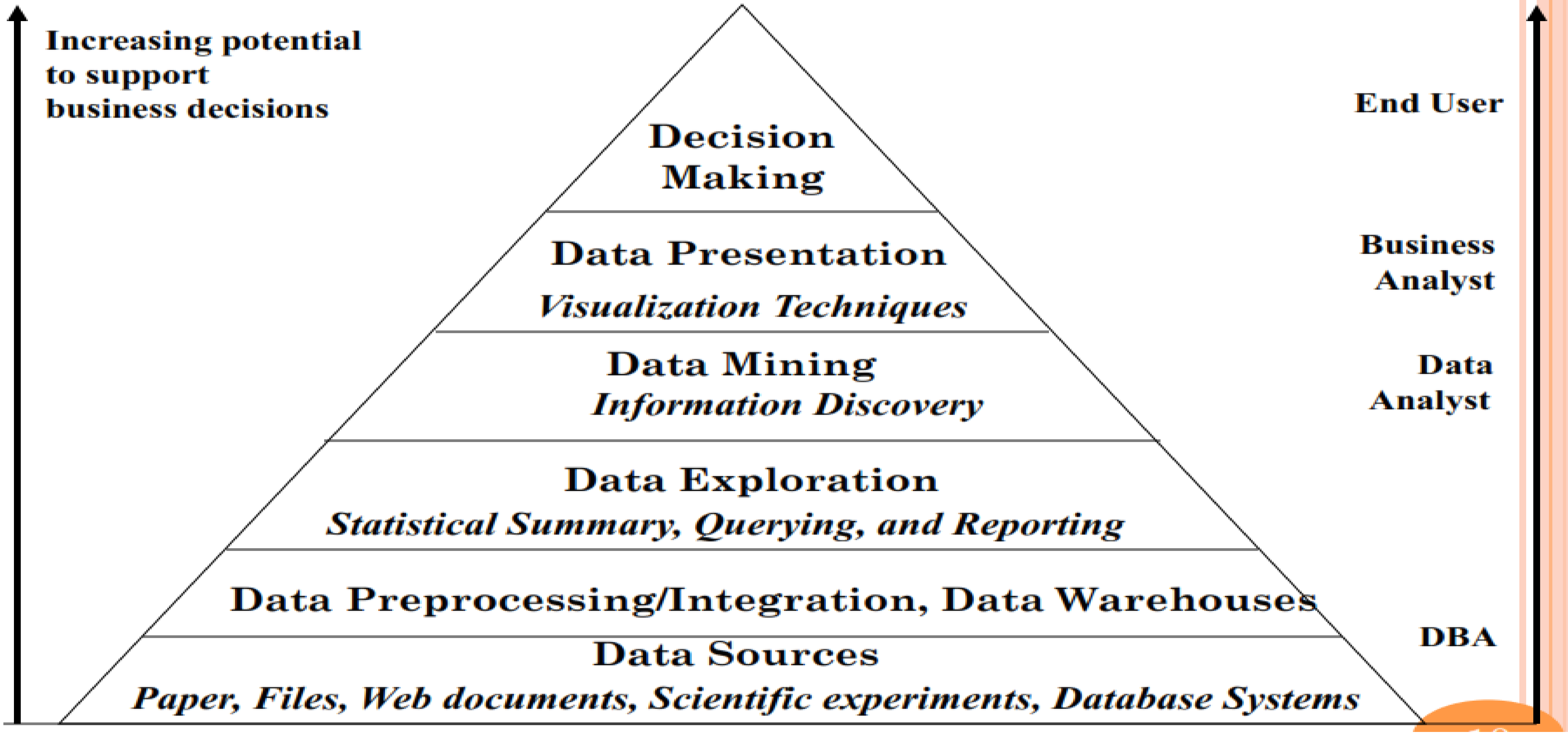
Processed Data

Transformed Data

Patterns

Knowledge

Data Mining and Business Intelligence



Increasing potential to support business decisions

Decision Making

Data Presentation

Visualization Techniques

Data Mining

Information Discovery

Data Exploration

Statistical Summary, Querying, and Reporting

Data Preprocessing/Integration, Data Warehouses

Data Sources

Paper, Files, Web documents, Scientific experiments, Database Systems

End User

Business Analyst

Data Analyst

DBA

Major Issues in Data Mining

- Mining methodology
  - Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web
  - Performance: efficiency, effectiveness, and scalability
  - Pattern evaluation: the interestingness problem
  - Incorporation of background knowledge
  - Handling noise and incomplete data
  - Parallel, distributed and incremental mining methods
  - Integration of the discovered knowledge with existing one: knowledge fusion
- User interaction
  - Data mining query languages and ad-hoc mining
  - Expression and visualization of data mining results
  - Interactive mining of knowledge at multiple levels of abstraction
- Applications and social issue
  - Domain-specific data mining & invisible data mining
  - Protection of data security, integrity, and privacy
  - Research in social issues

Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables. (Classification, Regression, Outlier Detection)
- Description Methods
  - Find human-interpretable patterns that describe the data. (Clustering, Association Rule Mining, Sequential Pattern Discovery)

Data Mining Tasks: Research Issues in social media mining

- Community Analysis
- Sentiment Analysis and Opinion Mining
- Social Recommendation
- Influence Modeling
- Information Diffusion and Provenance
- Privacy, Security and Trust

What is use of data mining in social media mining?

- Social Media data is everywhere.
- Information Overload (blogs, photos, videos, bookmarks) Interaction Overload (friends, taggers, followers, commenters) How to extract data from this chaos?
- Social media captures ‘pulse of humanity’.
- Can directly study opinions and behaviors of millions of users to gain insight into: -Human behavior -Market analytics -Product sentiments

Challenges in social media

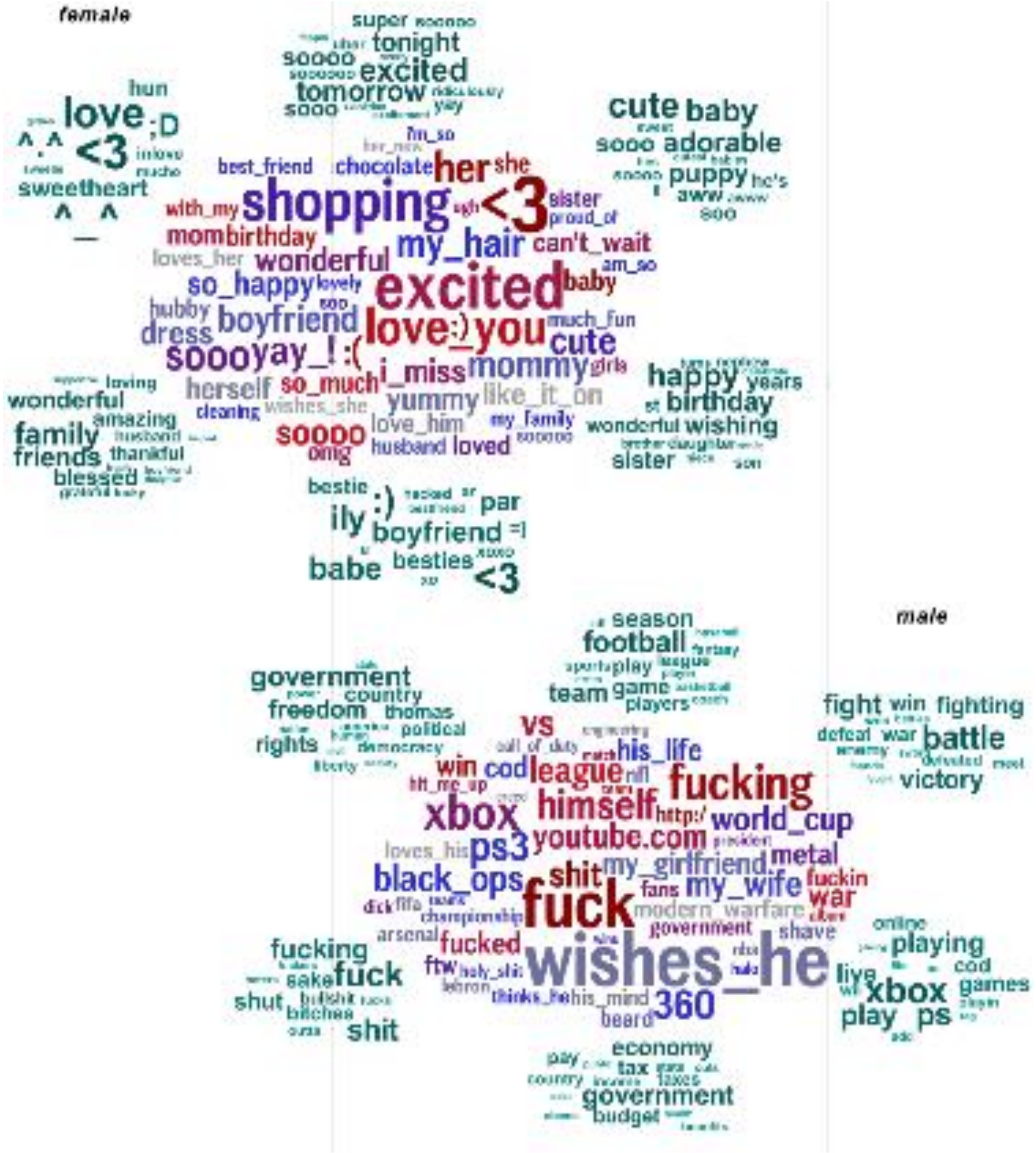
- Mining social media data are vast, noisy, distributed, unstructured, dynamic.
- These characteristics pose challenges to data mining tasks to invent new efficient techniques and algorithms.
- The amount of data! For example, Facebook and Twitter report Web data from approximately 149 million and 90 million unique U.S. visitors per month, respectively.
- According to the video sharing site YouTube,5 more than 4 billion videos are viewed per day, and 60 hours of videos are uploaded every minute.
- The picture sharing site Flickr, as of August 2011, hosts more than 6 billion photo images.
- Web-based, collaborative, and multilingual Wikipedia hosts over 20 million articles attracting over 365 million readers.

Social Media

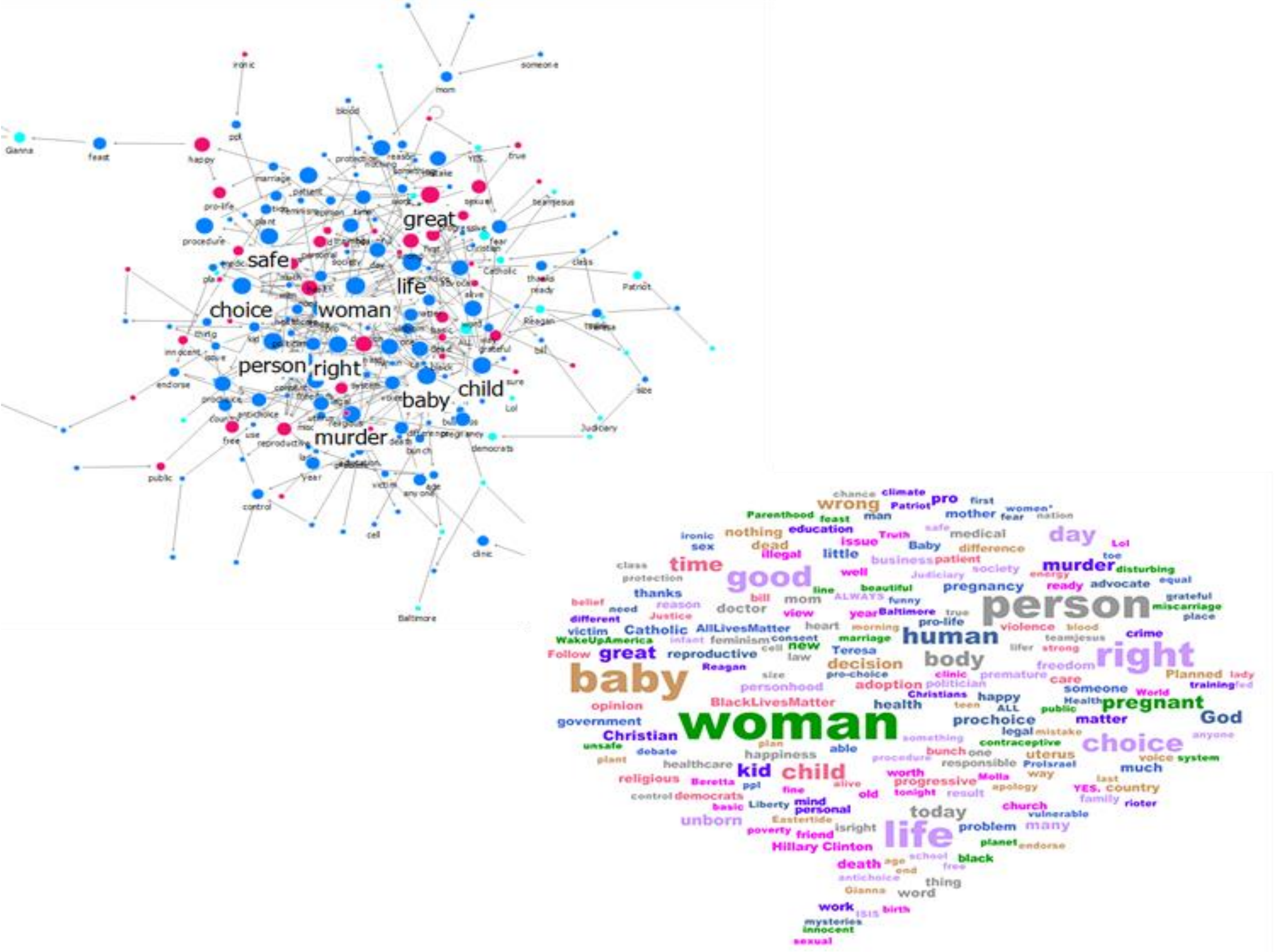
- Social media mining is extracting information from social media. Primary objectives of the data mining process are to effectively handle large-scale data, extract actionable patterns, and gain insightful knowledge. Users on Twitter generate over 400 million Tweets everyday.
- Data mining of social media can expand researchers' capability of understanding new phenomena to provide better services and develop innovative opportunities. Mining social media is a growing multidisciplinary area where researchers of different backgrounds can make important contributions that matter for social media research and development.

Text Mining in Social Media

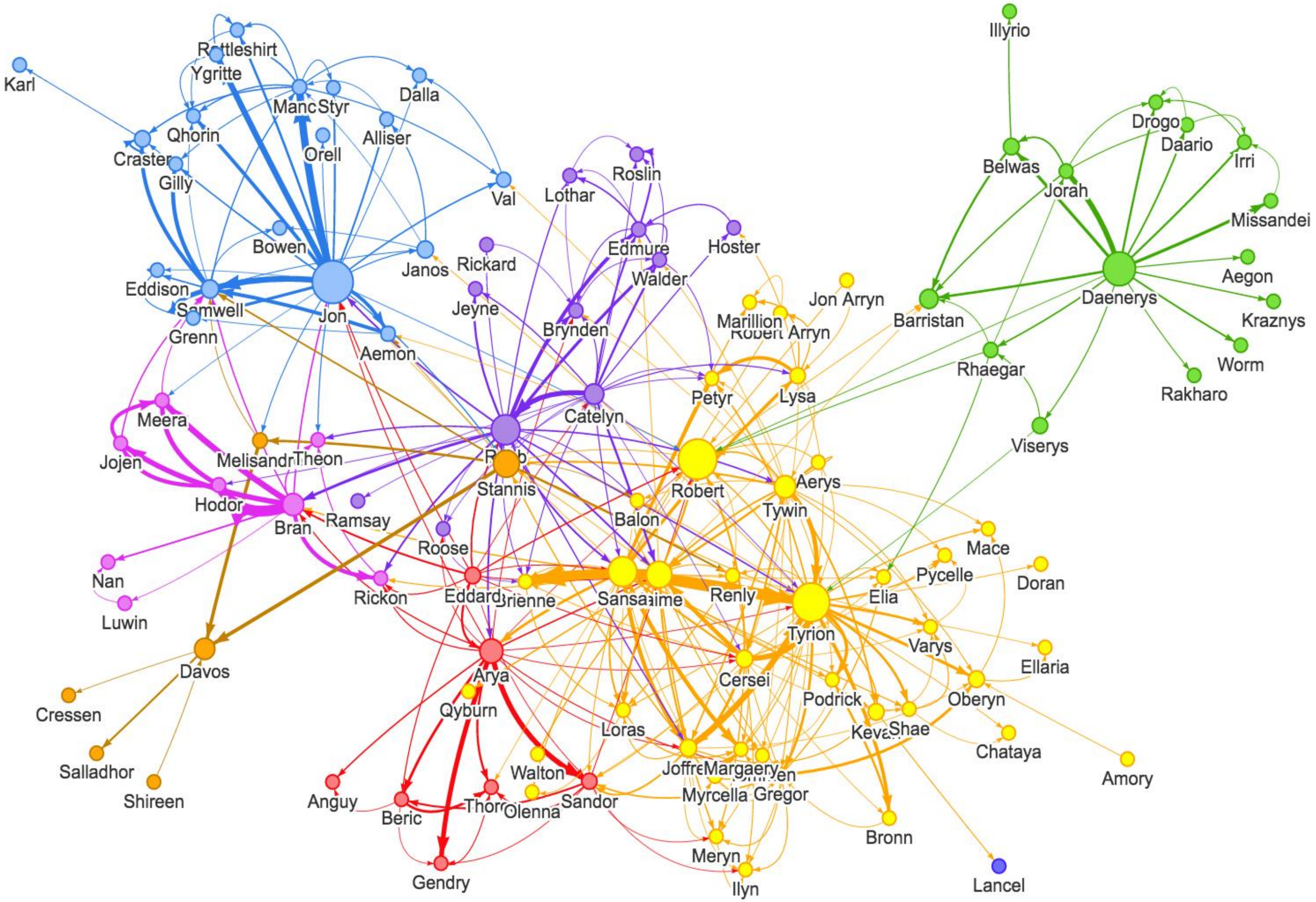
data
experience
business
work
analysis
science
analytics
development
learning
research
team
skills
machine
new
years
disability
must
quantitative
process
provide
insights
environment
techniques
requirements
role
design
degree
develop
results
project
help
ability
modeling
tools
software
required
models
services
company
product
technical
python
teams
solutions
problems
scientist
best
management
strong
knowledge
understanding
statistics
equal
technology
reporting
employer
computer
health
customers
communication
time
systems
analytical
complex
including
engineering
related
people
well
status
working
support
statistical
information
opportunity
job
employment
large
may
build
make
gender
national
predictive
can
apply
quality
customer
products
clients
advanced
field
across
projects
sql
position
using



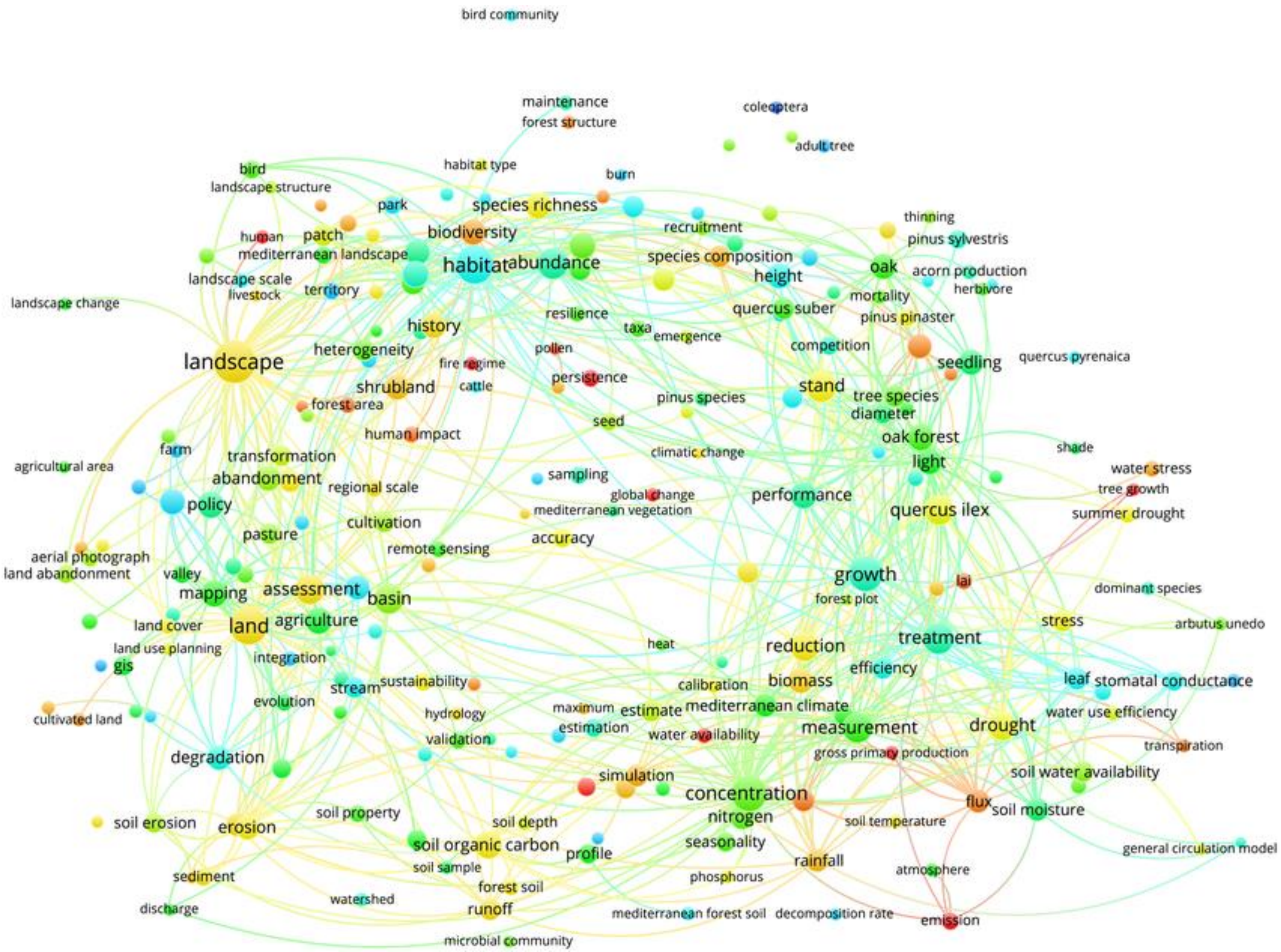
Text Mining


great
safe
life
choice
woman
person
right
baby
child
murder
ironic
someone
mom
blood
ppl
Gianna
feast
happy
true
marriage
patient
pro-life
plant
procedure
sexual
teamjesus
fear
Catholic
class
thanks
ready
Patriot
Reagan
innocent
issue
endorse
bill
grateful
size
Lol
Judiciary
democrats
free
public
reproductive
control
year
victim
any one
cell
clinic
Baltimore
woman
person
baby
right
life
choice
good
time
human
body
child
kid
pregnant
day
murder
wrong
nothing
pregnancy
doctor
decision
adoption
health
prochoice
God
Christian
Catholic
AllLivesMatter
BlackLivesMatter
WakeUpAmerica
reproductive
great
today
unborn
Hillary Clinton
death
democrats
religious
government
opinion
mother
medical
Baby
difference
business
patient
society
advocate
equal
miscarriage
place
violence
crime
freedom
Planned
training
someone
matter
uterus
progressive
church
problem
many
vulnerable
black
word
thing
sexual
innocent
mysteries
work
birth

Mining Network in Social Media



Mining Term in Journals



Data Mining Techniques

- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Clustering [Descriptive]
- Regression [Predictive]
- Classification [Predictive]
- Deviation Detection [Predictive]
- Collaborative Filter [Predictive]

Data Mining Techniques in Analysis data

- Exploratory Data Analysis
- Linear Classification (Perceptron & Logistic Regression)
- Linear Regression
- C4.5 Decision Tree
- Apriori
- K-means Clustering
- EM Algorithm
- PageRank & HITS
- Collaborative Filtering

Association Rule Mining

sales records:

transaction id	customer id	products bought
tran1	cust33	p2, p5, p8
tran2	cust45	p5, p8, p11
tran3	cust12	p1, p9
tran4	cust40	p5, p8, p11
tran5	cust12	p2, p9
tran6	cust12	p9

market-basket data

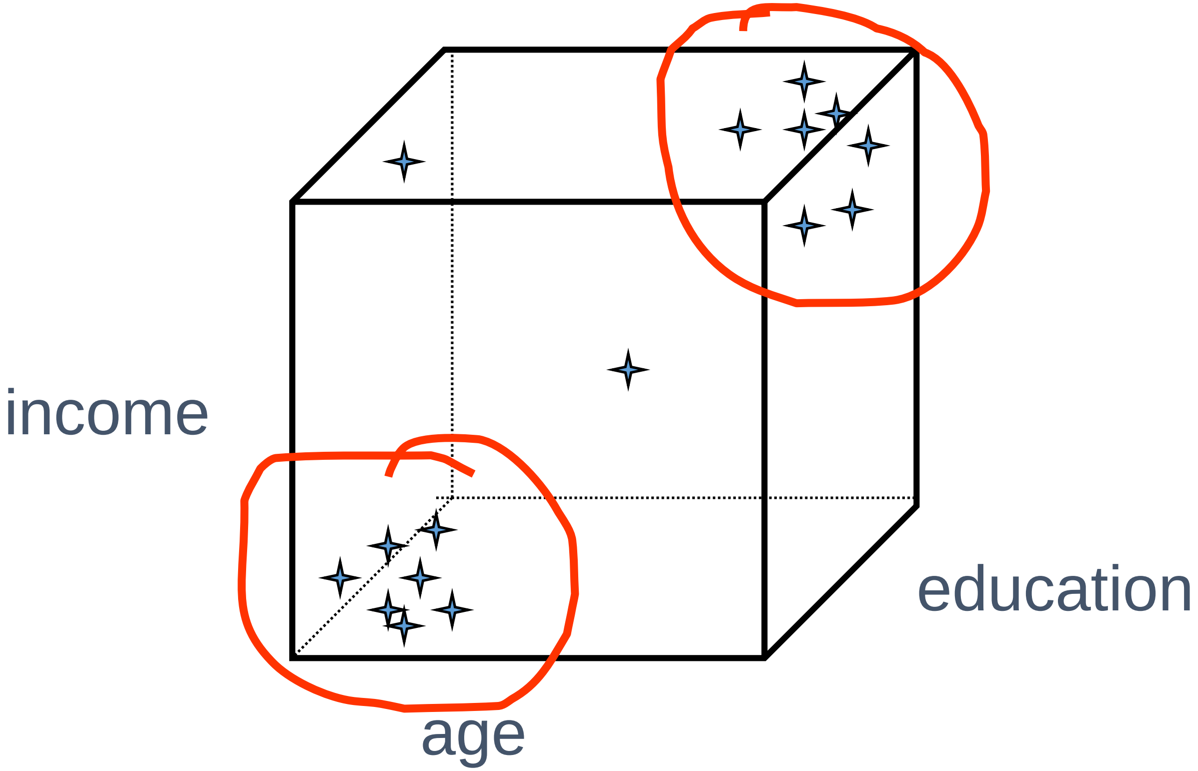
- Trend: Products p5, p8 often bough together
- Trend: Customer 12 likes product p9

Association Rule Discovery

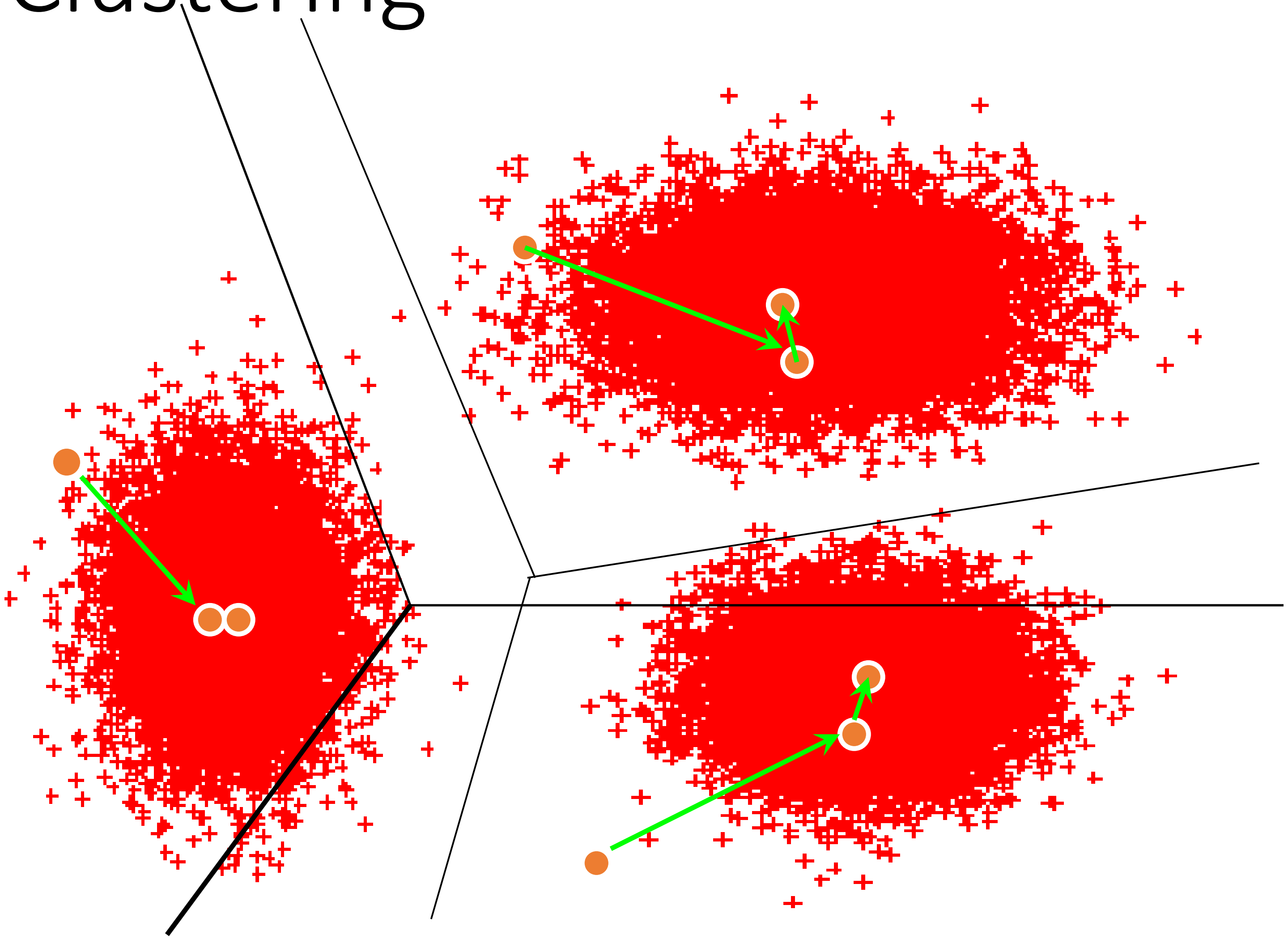
- Marketing and Sales Promotion:
  - Let the rule discovered be

    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!
- Supermarket shelf management.
- Inventory Managemnt

Clustering



K-Means Clustering



Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Decision Trees

Example:

- Conducted survey to see what customers were interested in new model car
- Want to select customers for advertising campaign

sale	custId	car	age	city	newCar
	c1	taurus	27	sf	yes
	c2	van	35	la	yes
	c3	van	40	sf	yes
	c4	taurus	22	sf	yes
	c5	merc	50	la	no
	c6	taurus	25	la	no

training set