Tamkang University

大數據分析

Big Data Analysis (IM EMBA, TKU)

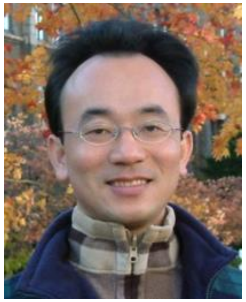
鄭啟斌 教授

資料探勘介紹

(Introduction to Data Mining)

Time: 2015/10/12 (19:20-22:10)

Place: D325



Min-Yuh Day

戴敏育

Assistant Professor

專任助理教授

Dept. of Information Management, Tamkang University

淡江大學 資訊管理學系

http://mail. tku.edu.tw/myday/

2015-10-12

Outline

- Data Mining and Big Data Analytics
- Data Mining Process
- Data Mining Tasks
- Data Mining Evaluation
- Social Network Analysis

Source: Han & Kamber (2006)

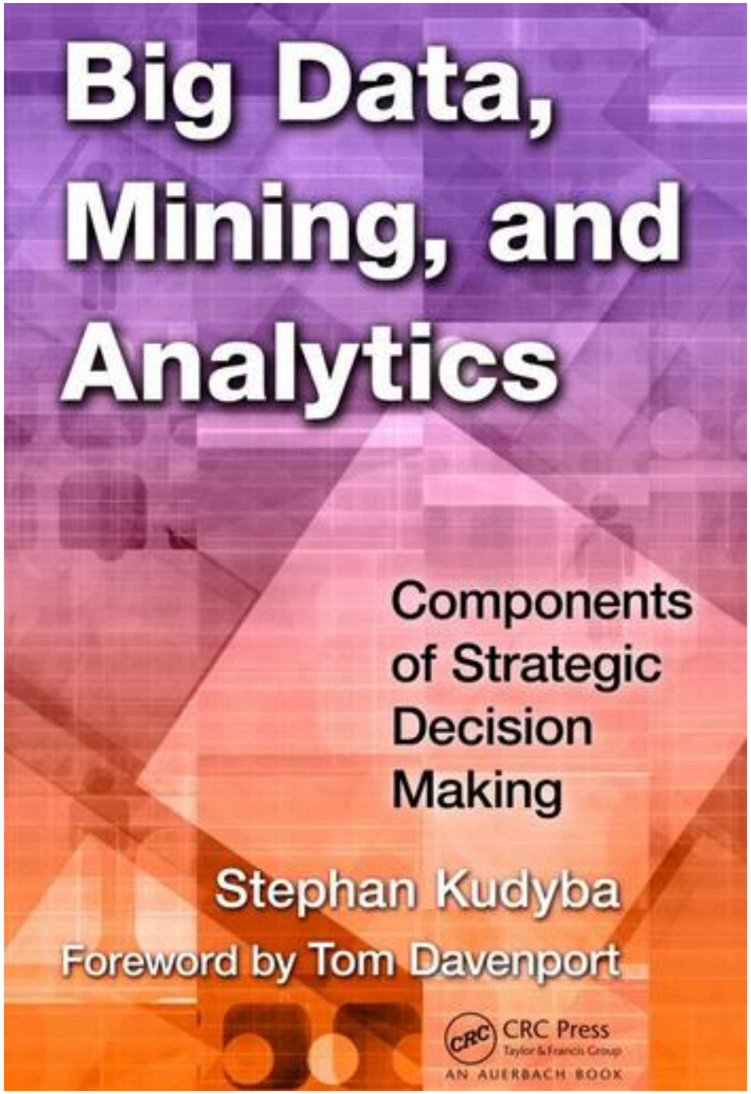
Data Mining

and

Big Data Analytics

Stephan Kudyba (2014),

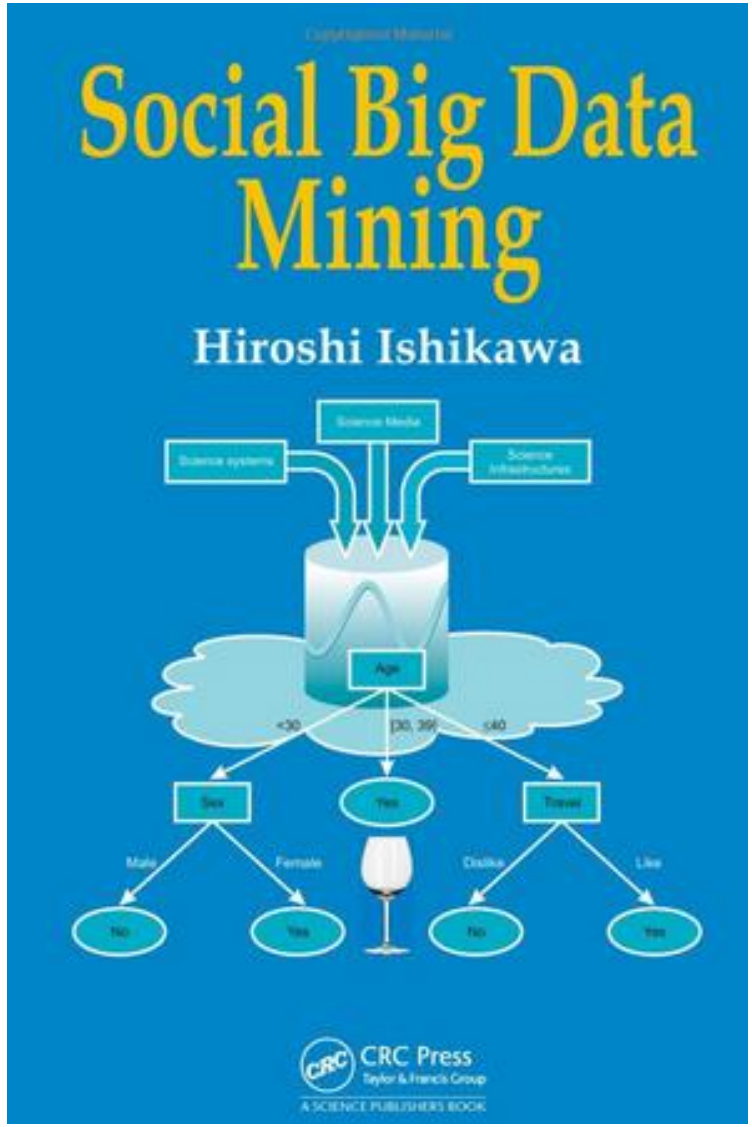
Big Data, Mining, and Analytics:
Components of Strategic Decision Making, Auerbach Publications



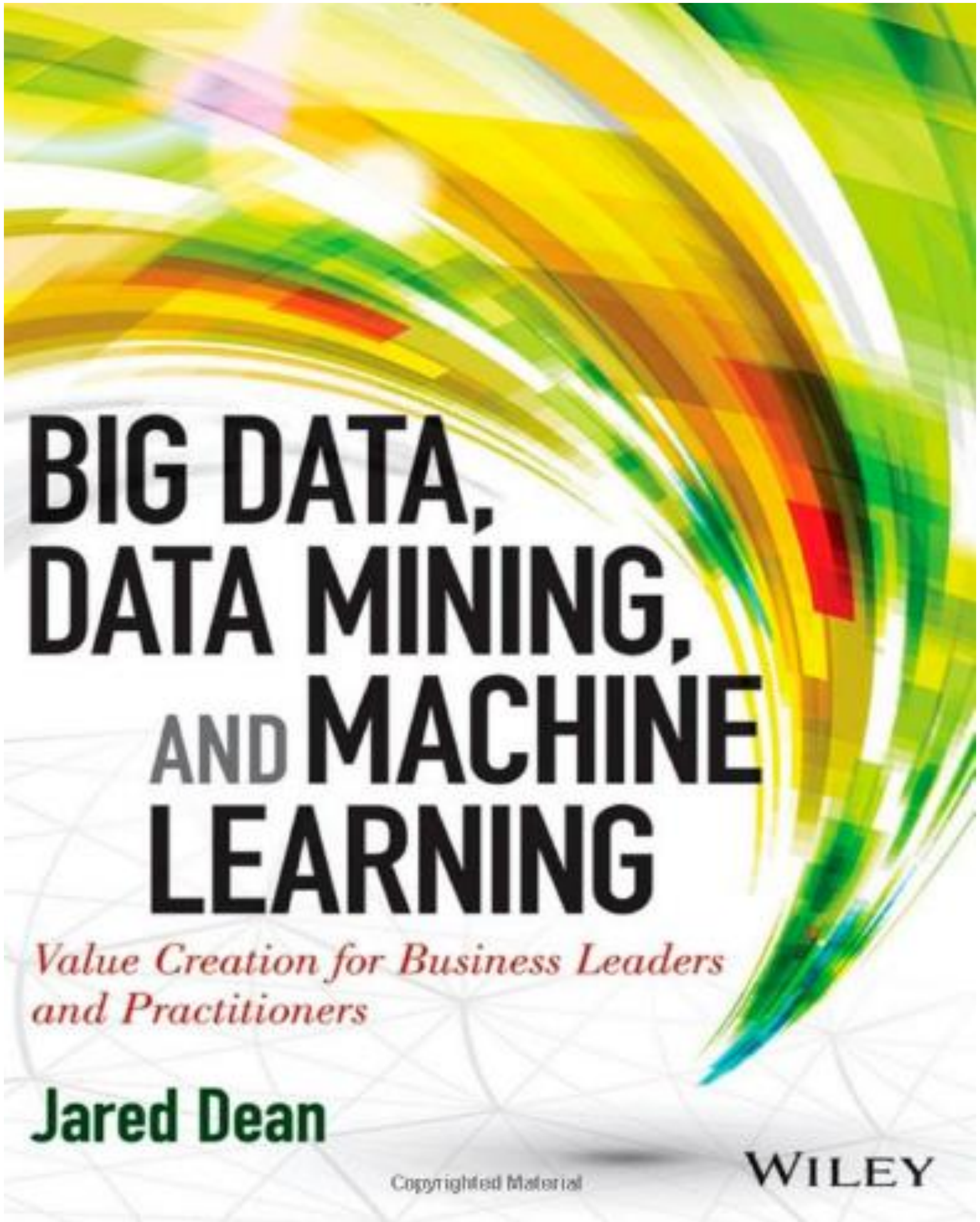
Source: http://www.amazon.com/gp/product/1466568704

Social Big Data Mining

(Hiroshi Ishikawa, 2015)

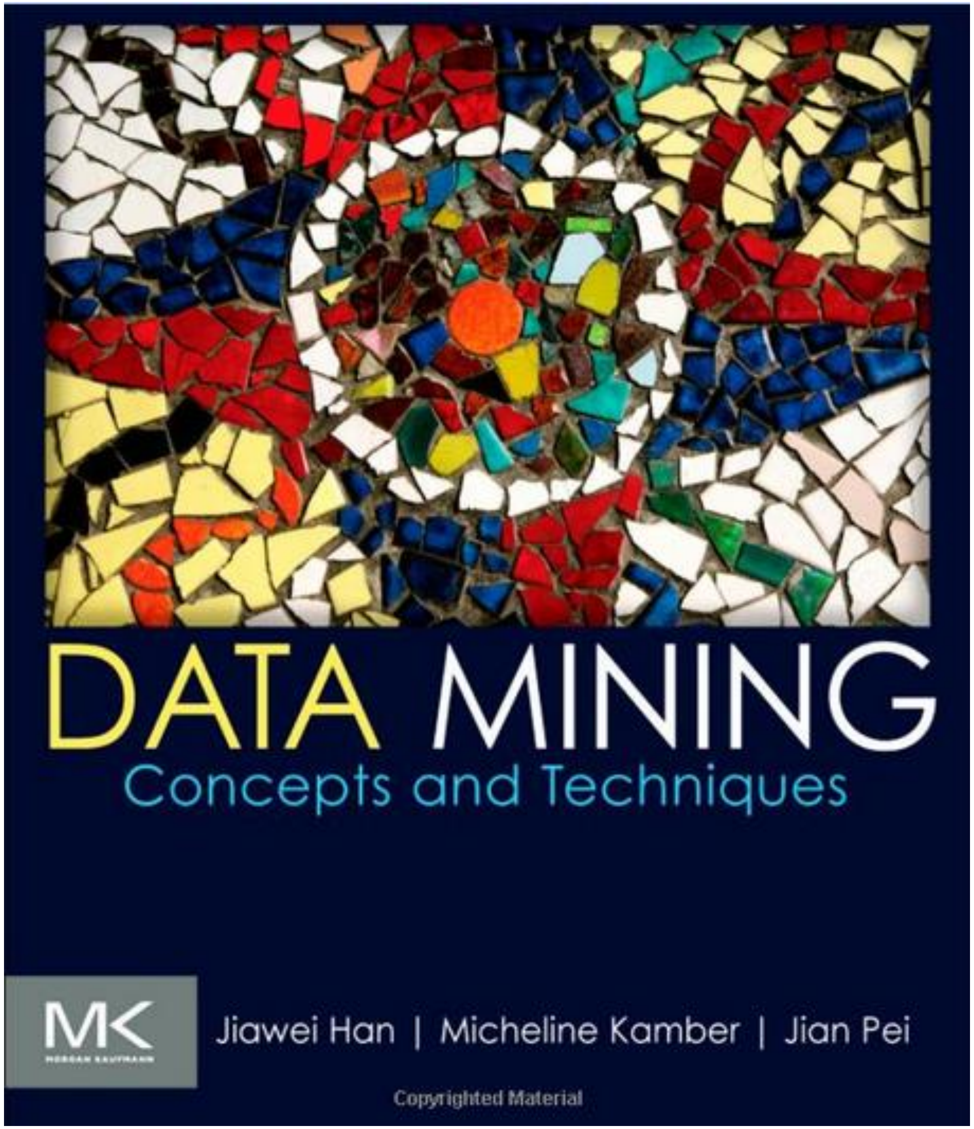


Source: http://www.amazon.com/Social-Data-Mining-Hiroshi-Ishikawa/dp/149871093X



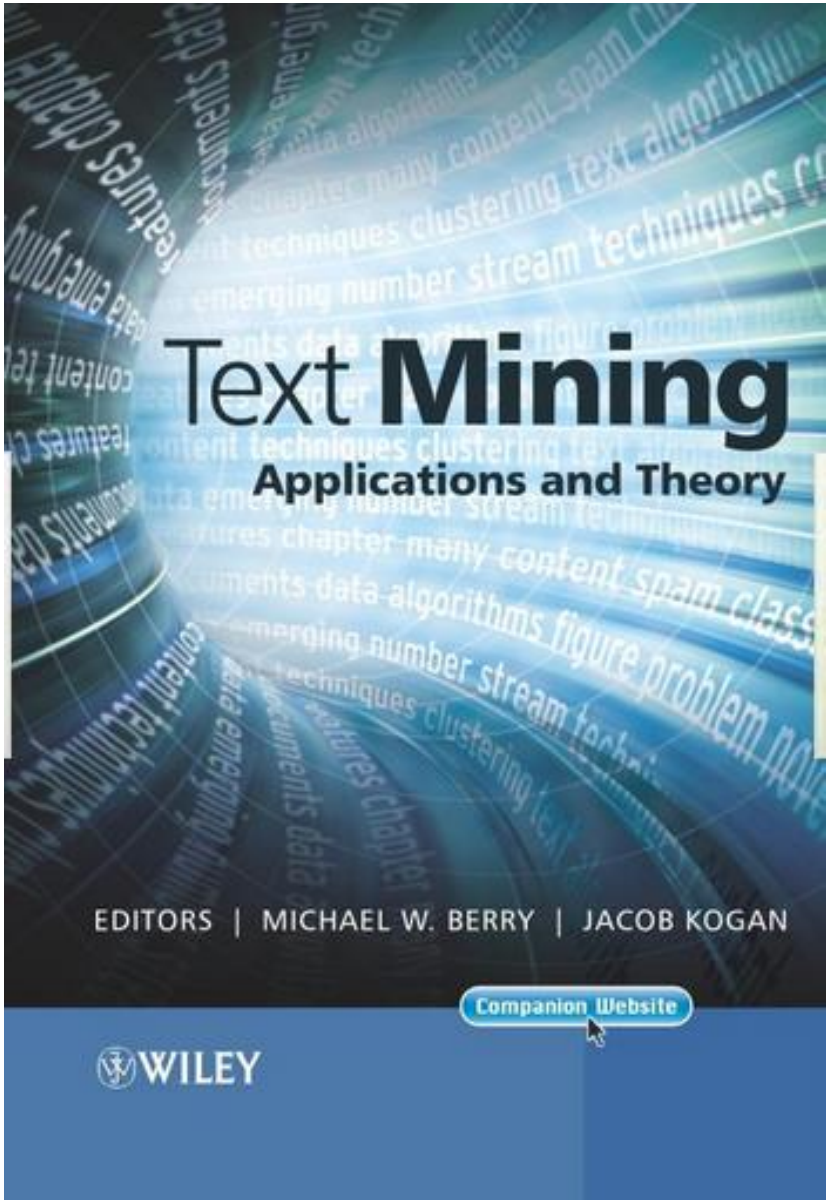
Source: http://www.amazon.com/Data-Mining-Machine-Learning-Practitioners/dp/1118618041

Data Mining



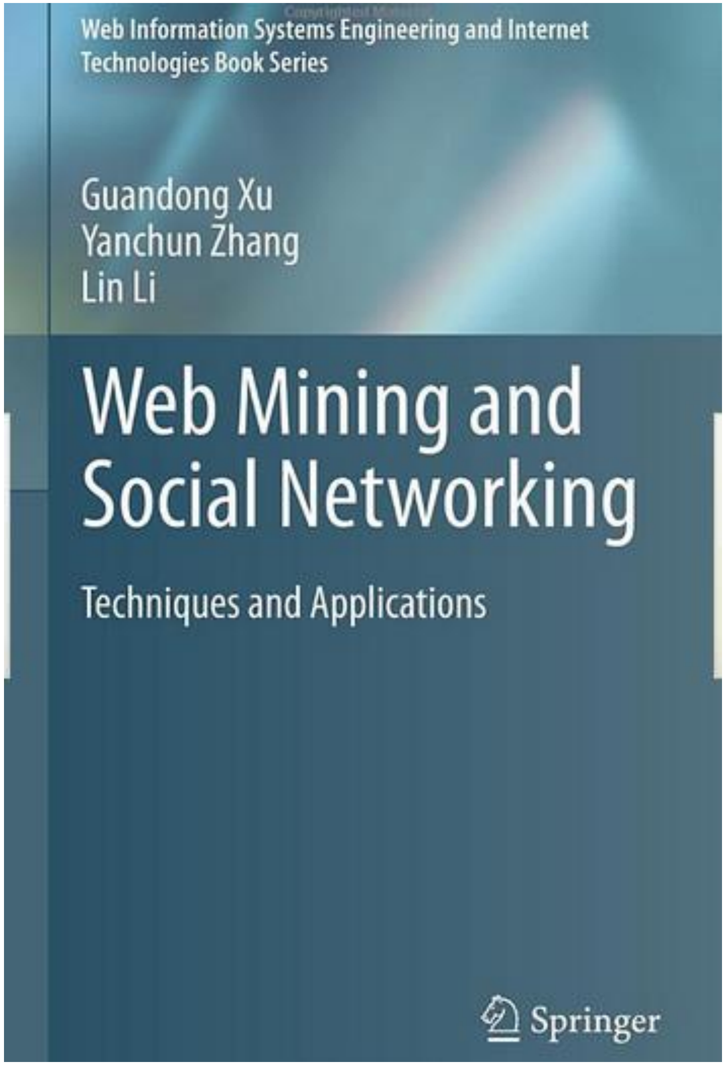
Source: http://www.amazon.com/Data-Mining-Concepts-Techniques-Management/dp/0123814790

Text Mining



Source: http://www.amazon.com/Text-Mining-Applications-Michael-Berry/dp/0470749822/

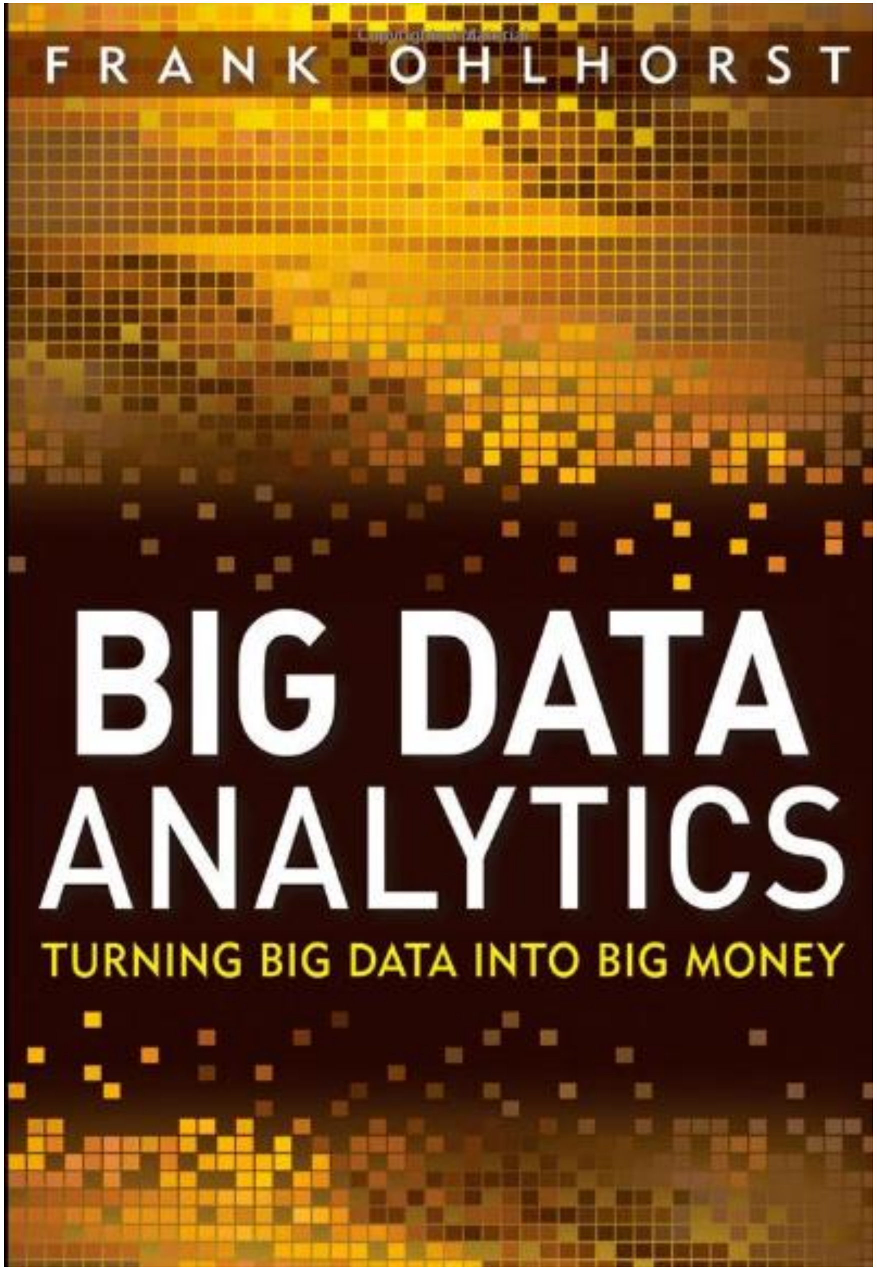
Web Mining and Social Networking



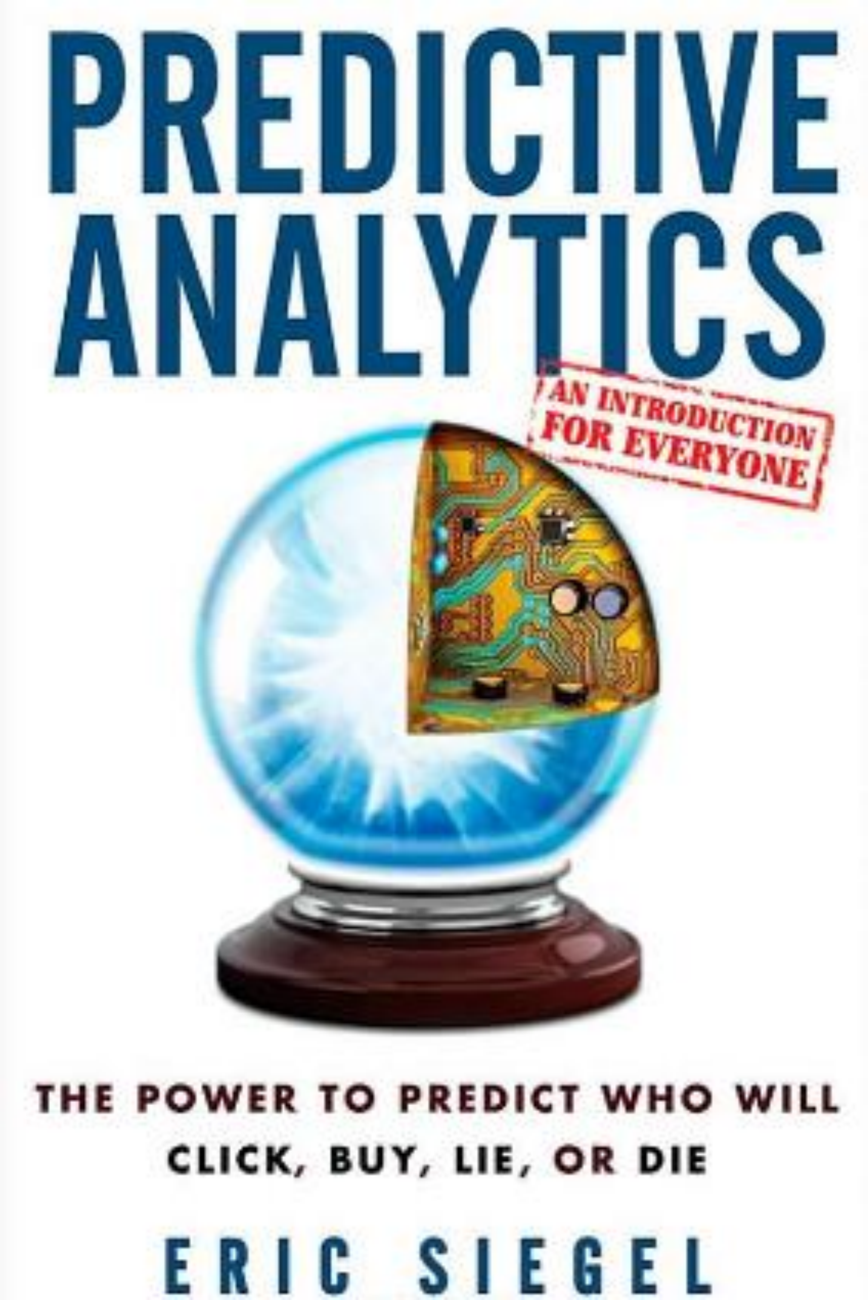
Source: http://www.amazon.com/Web-Mining-Social-Networking-Applications/dp/1441977341

Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites

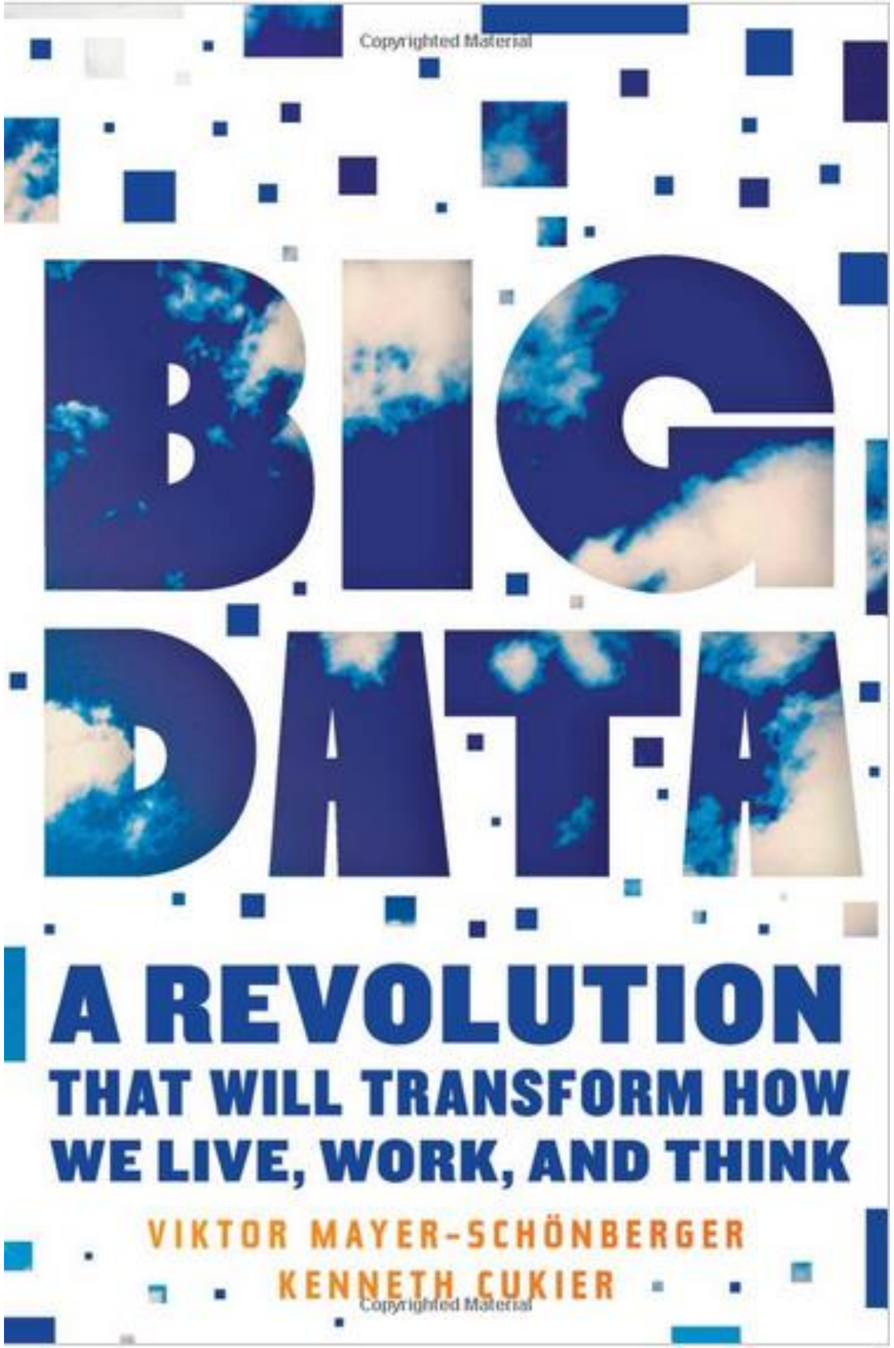
Source: http://www.amazon.com/Mining-Social-Web-Analyzing-Facebook/dp/1449388345



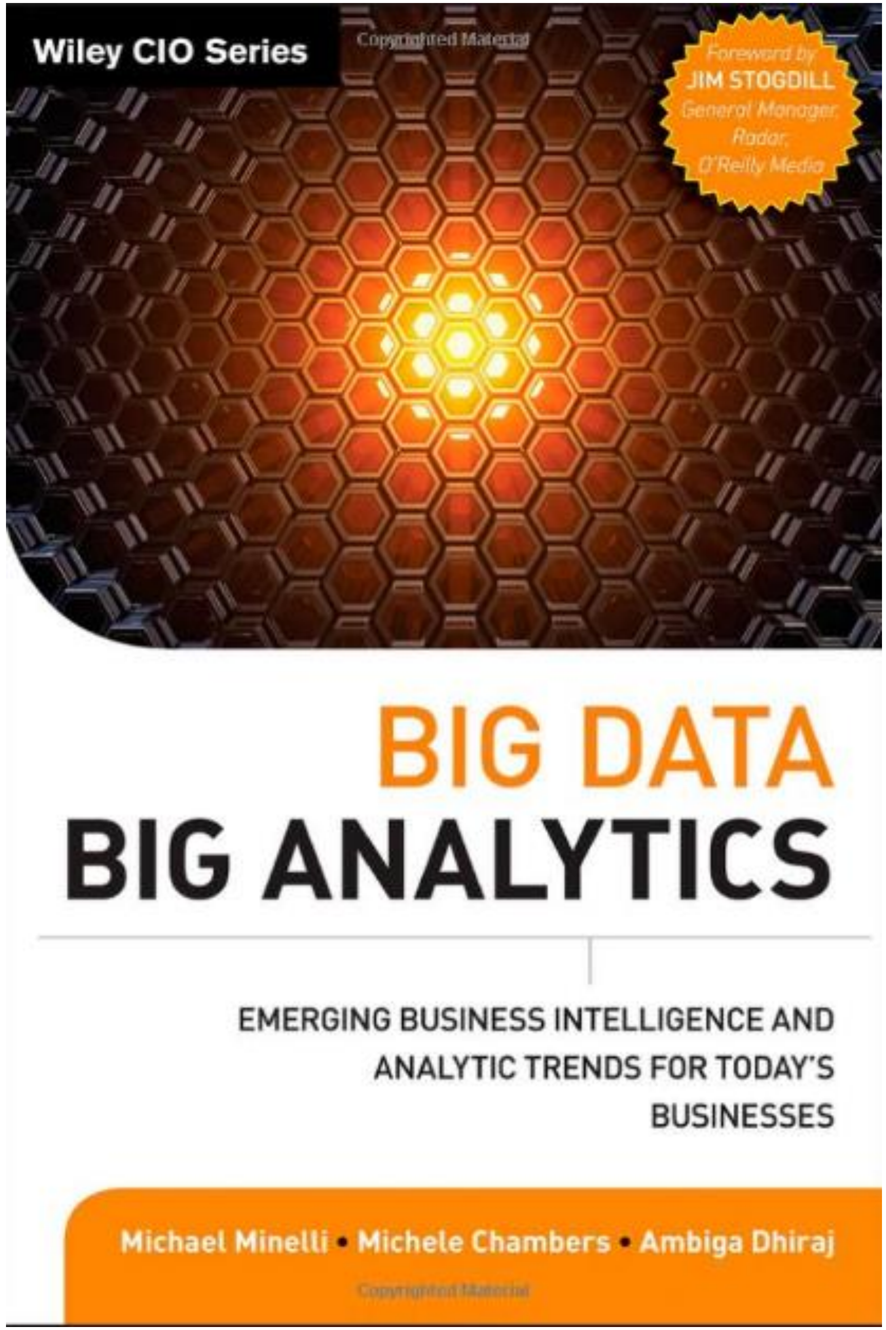
Source: http://www.amazon.com/Big-Data-Analytics-Turning-Money/dp/1118147596



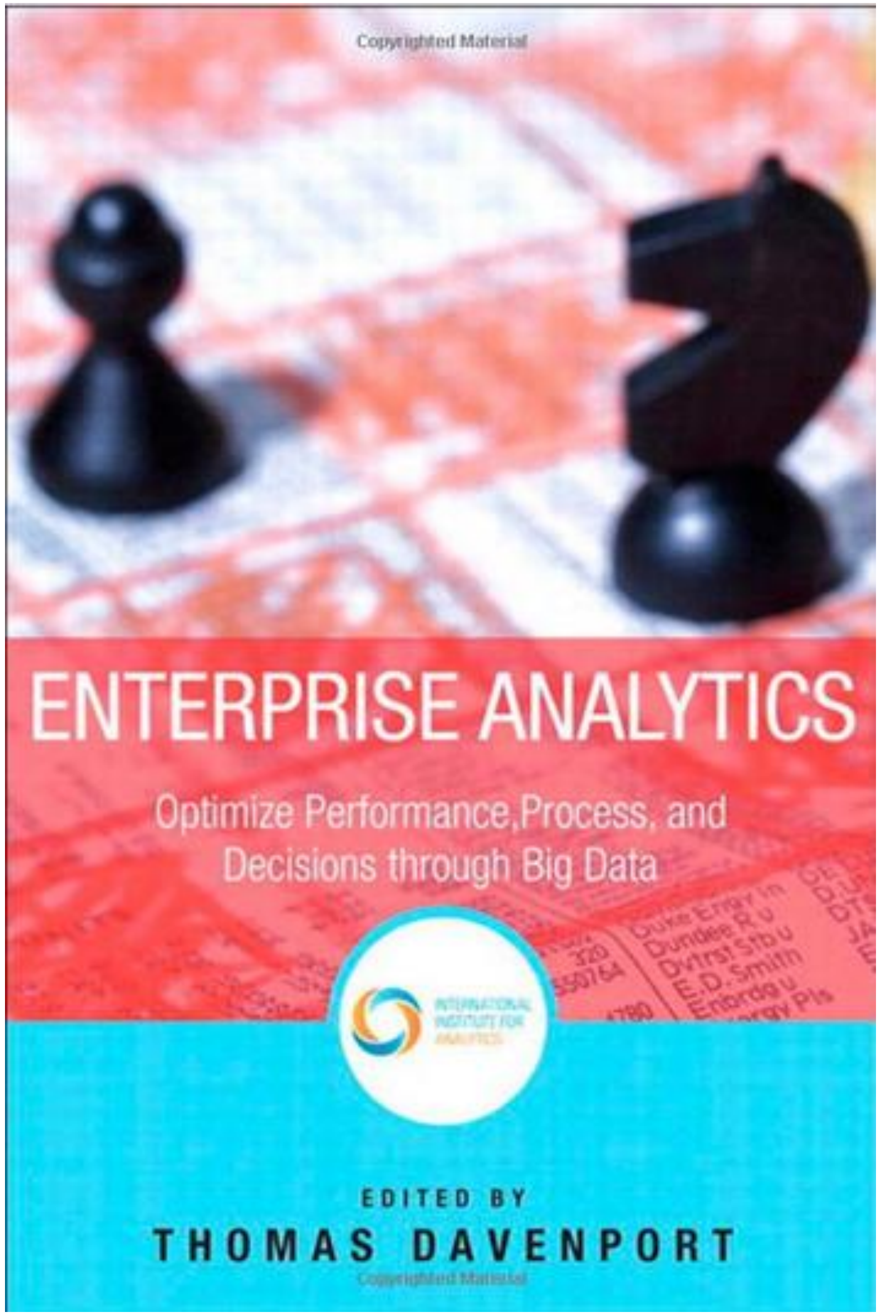
Source: http://www.amazon.com/Predictive-Analytics-Power-Predict-Click/dp/1118356853



Source: http://www.amazon.com/Big-Data-Revolution-Transform-Mayer-Schonberger/dp/B00D81X2YE



Source: http://www.amazon.com/Big-Data-Analytics-Intelligence-Businesses/dp/111814760X



Source: http://www.amazon.com/Enterprise-Analytics-Performance-Operations-Management/dp/0133039439

HBR.ORG

Harvard Business Review

OCTOBER 2012
REPRINT R1210C

SPOTLIGHT ON BIG DATA

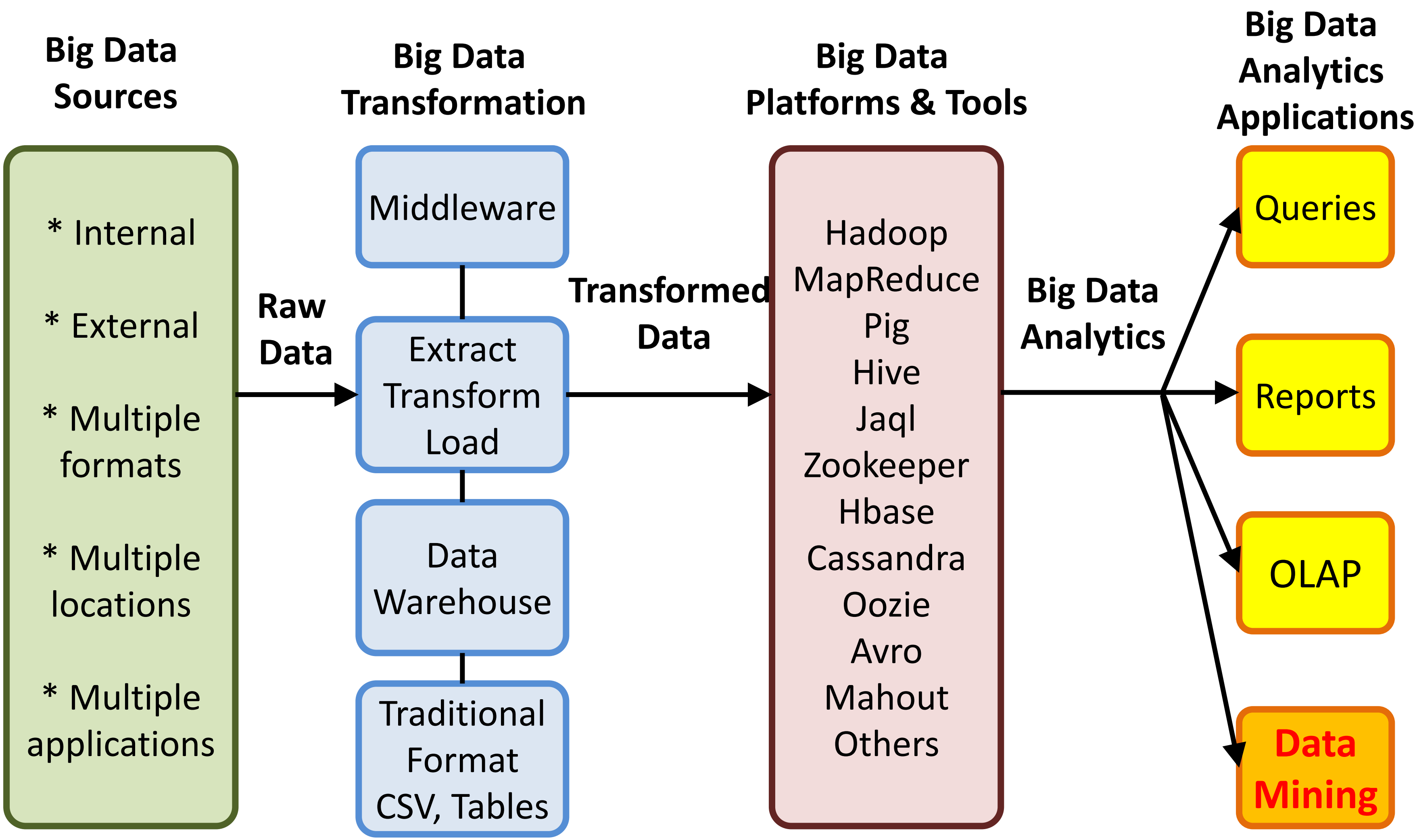
Big Data: The Management Revolution

Exploiting vast new flows of information can radically improve your company's performance. But first you'll have to change your decision-making culture.

by Andrew McAfee and Erik Brynjolfsson

Source: McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution.*Harvard business review*.

Architecture of Big Data Analytics



Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications

Architecture of Big Data Analytics

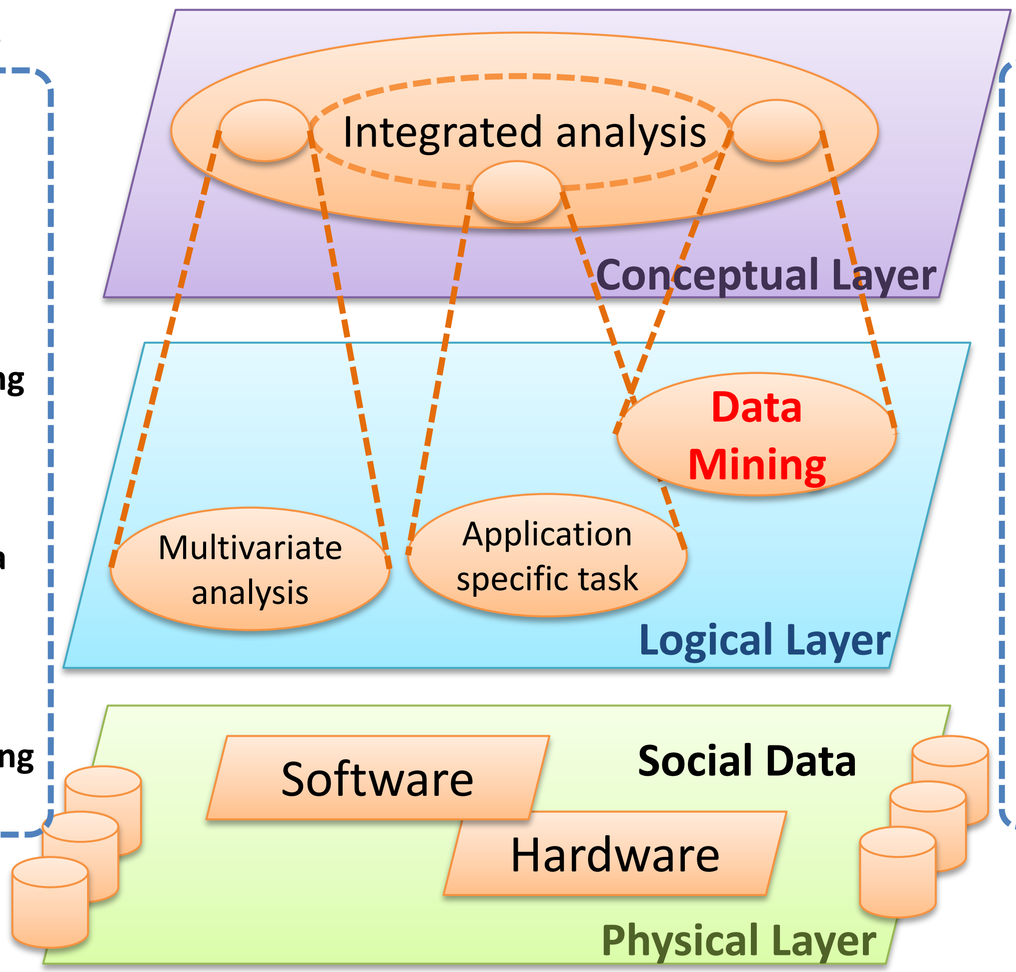
Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications

Architecture for Social Big Data Mining

(Hiroshi Ishikawa, 2015)

Enabling Technologies

- Integrated analysis model
- Natural Language Processing
- Information Extraction
- Anomaly Detection
- Discovery of relationships among heterogeneous data
- Large-scale visualization
- Parallel distrusted processing

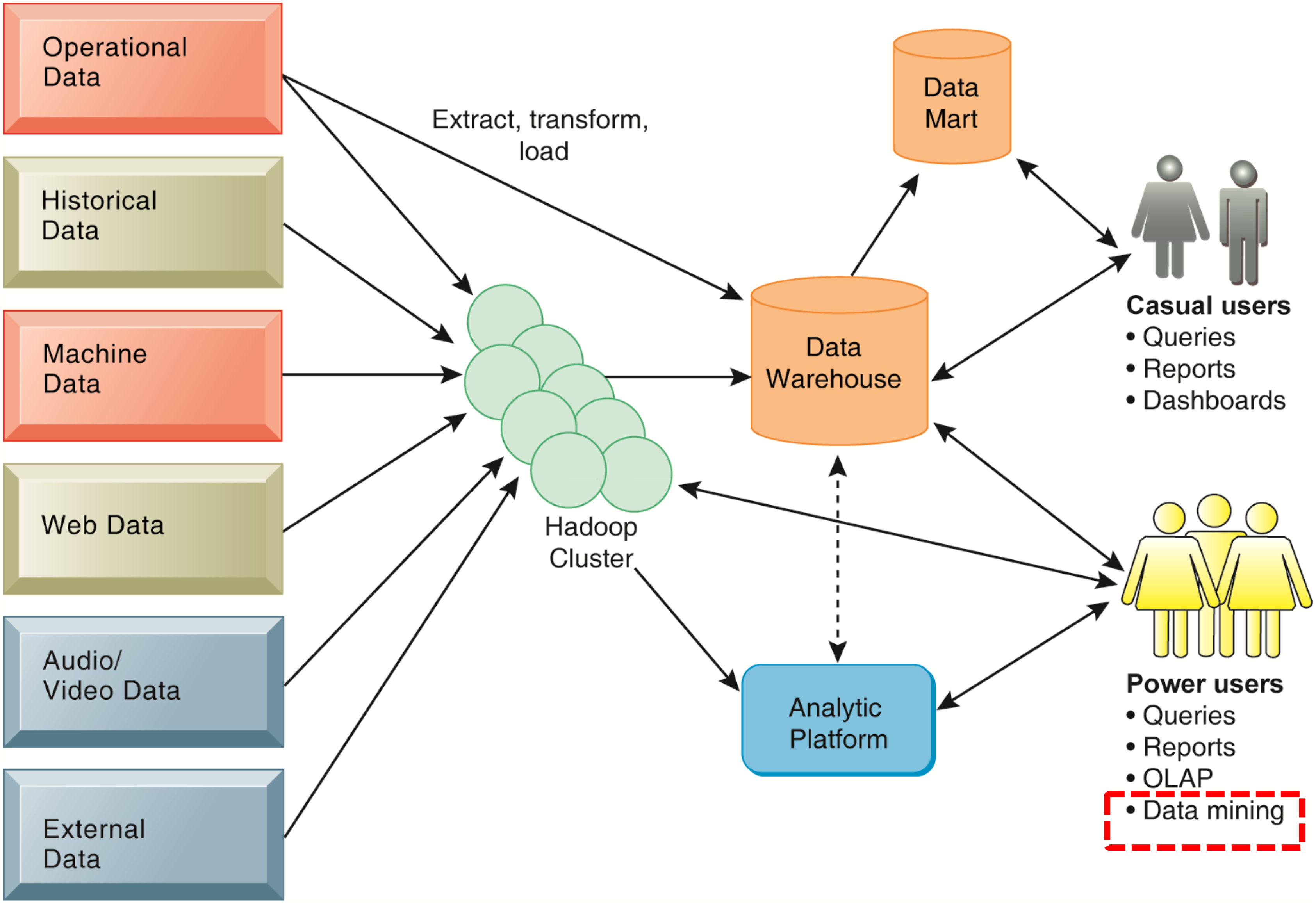


Analysts

- Model Construction
- Explanation by Model
- Construction and confirmation of individual hypothesis
- Description and execution of application-specific task

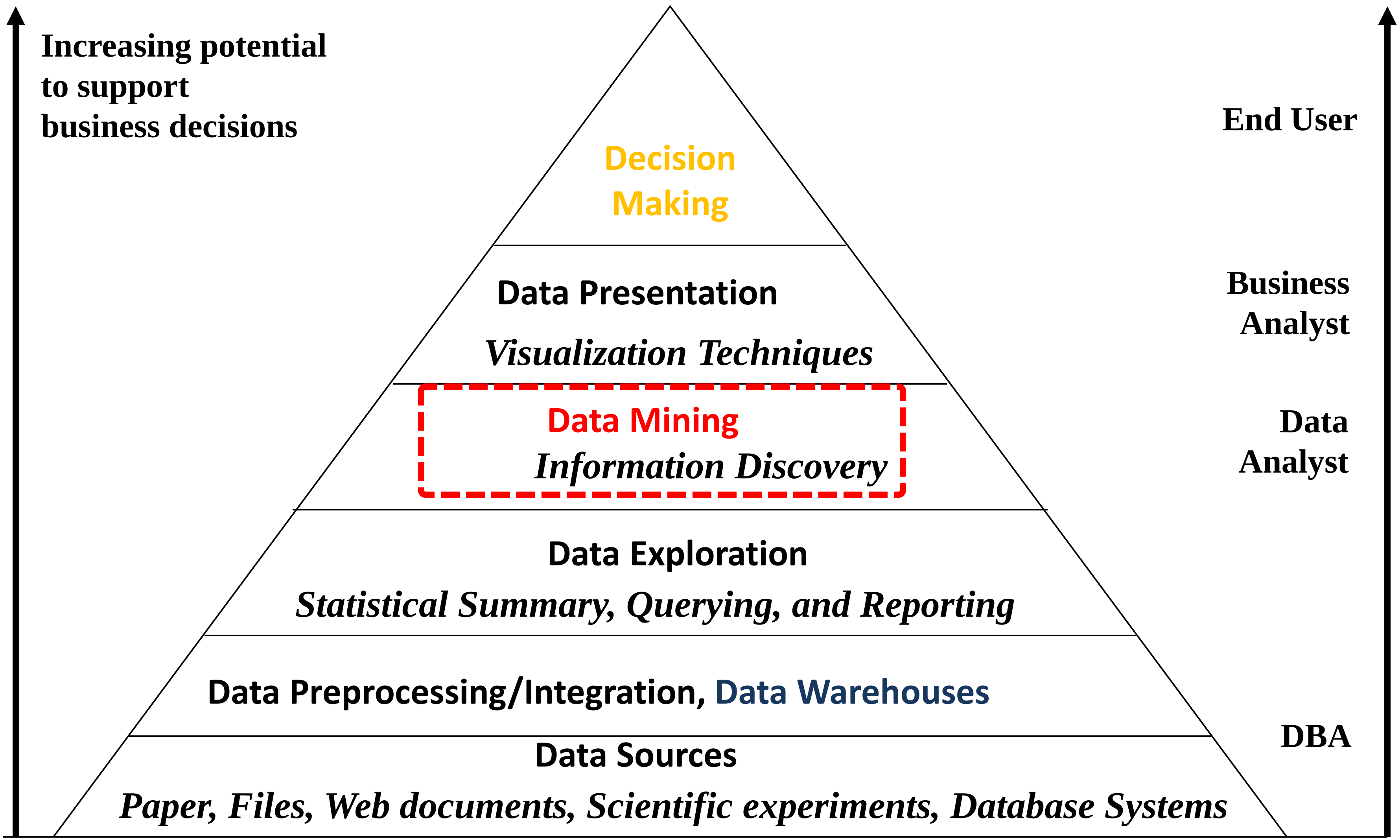
Source: Hiroshi Ishikawa (2015), Social Big Data Mining, CRC Press

Business Intelligence (BI) Infrastructure



Source: Kenneth C. Laudon & Jane P. Laudon (2014), Management Information Systems: Managing the Digital Firm, Thirteenth Edition, Pearson.

Data Warehouse
Data Mining and Business Intelligence



Source: Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Second Edition, Elsevier

The Evolution of BI Capabilities

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Business Intelligence and Analytics

- Business Intelligence 2.0 (BI 2.0)
  – Web Intelligence
  – Web Analytics
  – Web 2.0
  – Social Networking and Microblogging sites
- Data Trends
  – Big Data
- Platform Technology Trends
  – Cloud computing platform

Source: Lim, E. P., Chen, H., & Chen, G. (2013). Business Intelligence and Analytics: Research Directions. *ACM Transactions on Management Information Systems (TMIS)*, *3*(4), 17

Business Intelligence and Analytics: Research Directions

1. Big Data Analytics
   – Data analytics using Hadoop / MapReduce framework
2. Text Analytics
   – From Information Extraction to Question Answering
   – From Sentiment Analysis to Opinion Mining
3. Network Analysis
   – Link mining
   – Community Detection
   – Social Recommendation

Source: Lim, E. P., Chen, H., & Chen, G. (2013). Business Intelligence and Analytics: Research Directions. *ACM Transactions on Management Information Systems (TMIS)*, *3*(4), 17

Big Data, Big Analytics:

Emerging Business Intelligence and Analytic Trends for Today's Businesses

Big Data,

Prediction

vs.

Explanation

Source: Agarwal, R., & Dhar, V. (2014). Editorial—Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. Information Systems Research, 25(3), 443-448.

Big Data:

The Management Revolution

Source: McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution.*Harvard business review*.

Business Intelligence and Enterprise Analytics

- Predictive analytics
- Data mining
- Business analytics
- Web analytics
- Big-data analytics

Source: Thomas H. Davenport, "Enterprise Analytics: Optimize Performance, Process, and Decisions Through Big Data", FT Press, 2012

Three Types of Business Analytics

- Prescriptive Analytics
- Predictive Analytics
- Descriptive Analytics

Source: Thomas H. Davenport, "Enterprise Analytics: Optimize Performance, Process, and Decisions Through Big Data", FT Press, 2012

Three Types of Business Analytics

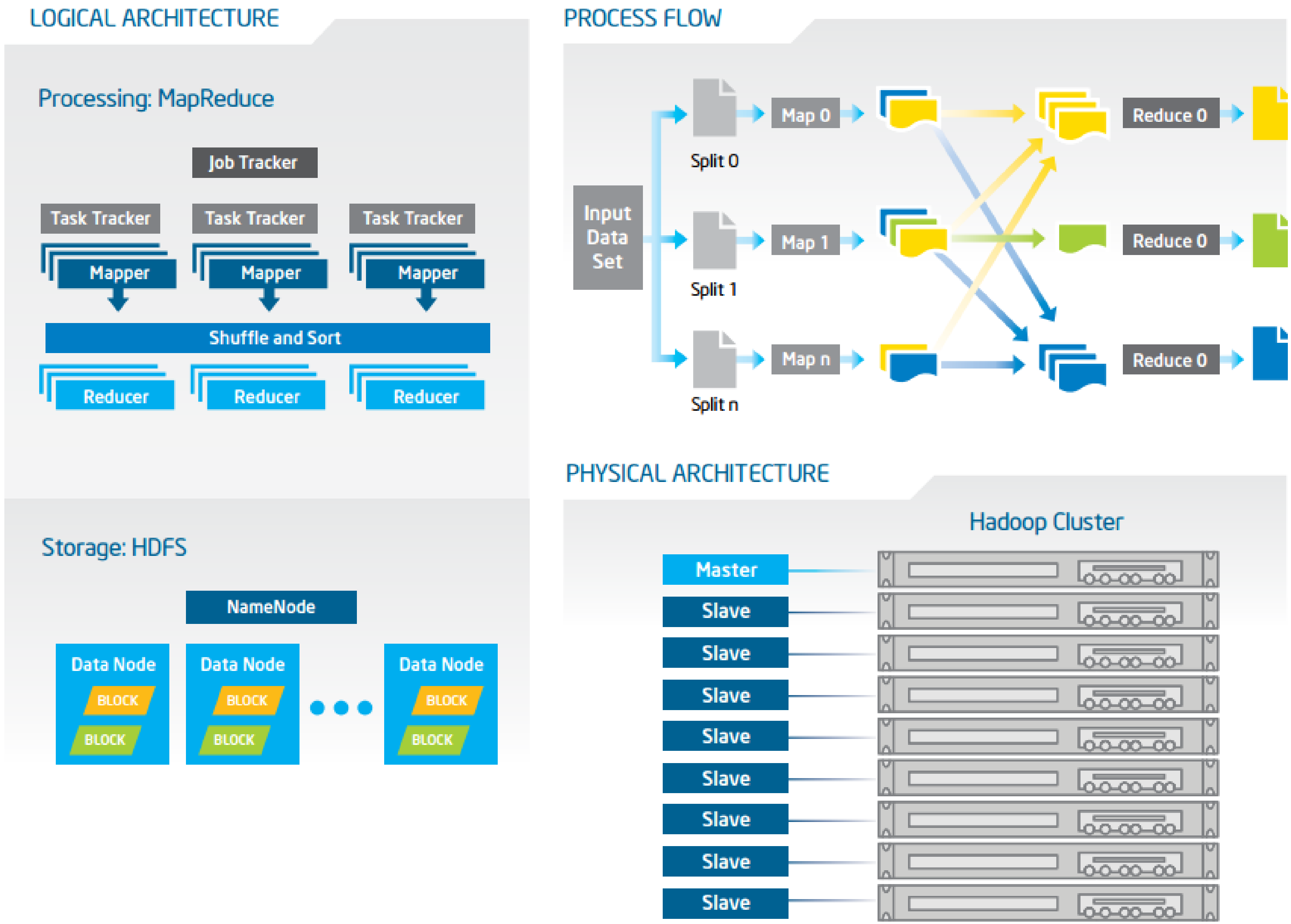
Optimization	"What's the best that can happen?"	**Prescriptive Analytics**
Randomized Testing	"What if we try this?"	
Predictive Modeling / Forecasting	"What will happen next?"	**Predictive Analytics**
Statistical Modeling	"Why is this happening?"	
Alerts	"What actions are needed?"	**Descriptive Analytics**
Query / Drill Down	"What exactly is the problem?"	
Ad hoc Reports / Scorecards	"How many, how often, where?"	
Standard Report	"What happened?"	

Source: Thomas H. Davenport, "Enterprise Analytics: Optimize Performance, Process, and Decisions Through Big Data", FT Press, 2012

Big-Data Analysis

- **Too Big,**
  too Unstructured,
  too many different source
  to be manageable through traditional databases

Big Data with Hadoop Architecture

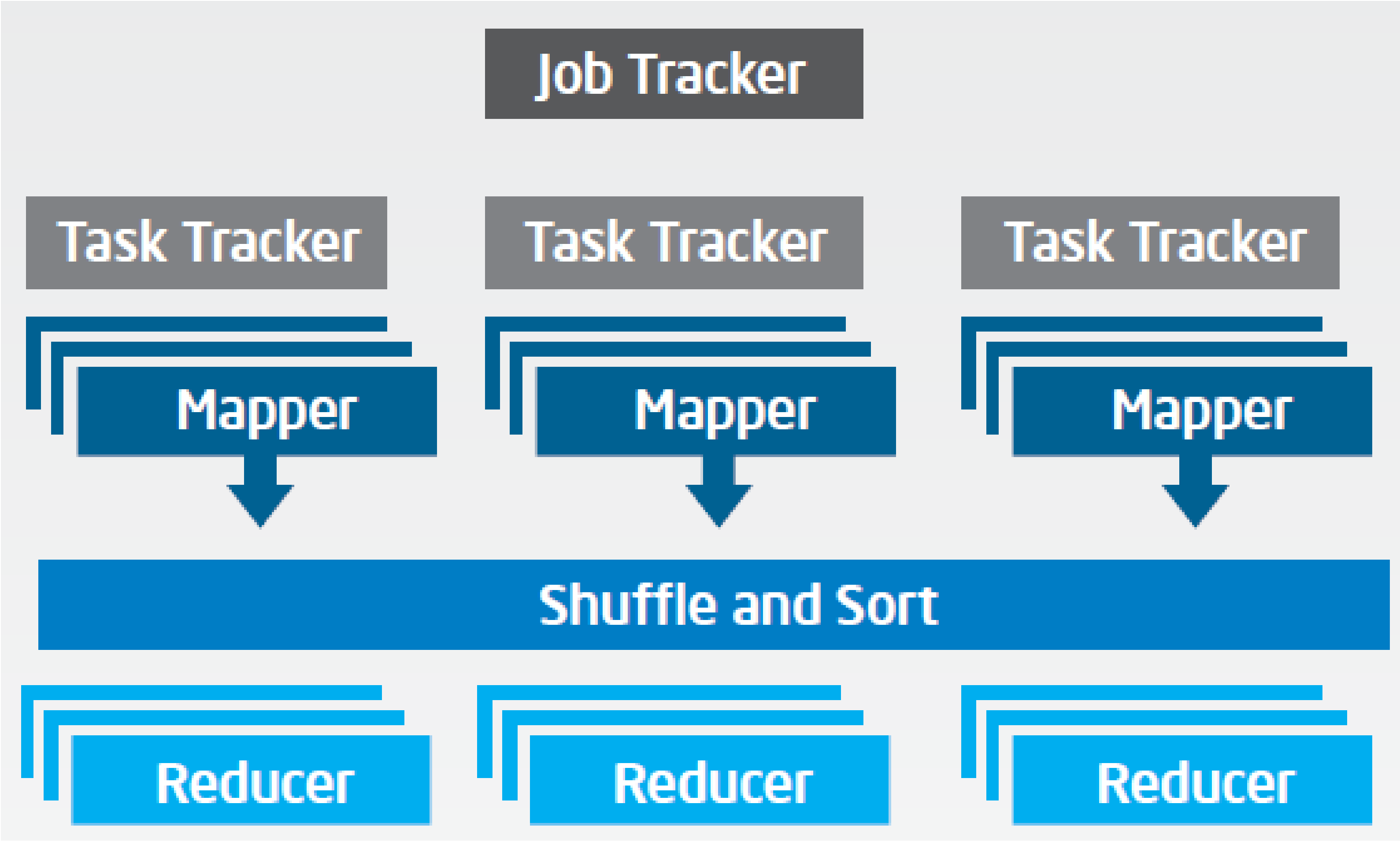


Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Big Data with Hadoop Architecture

Logical Architecture

Processing: MapReduce

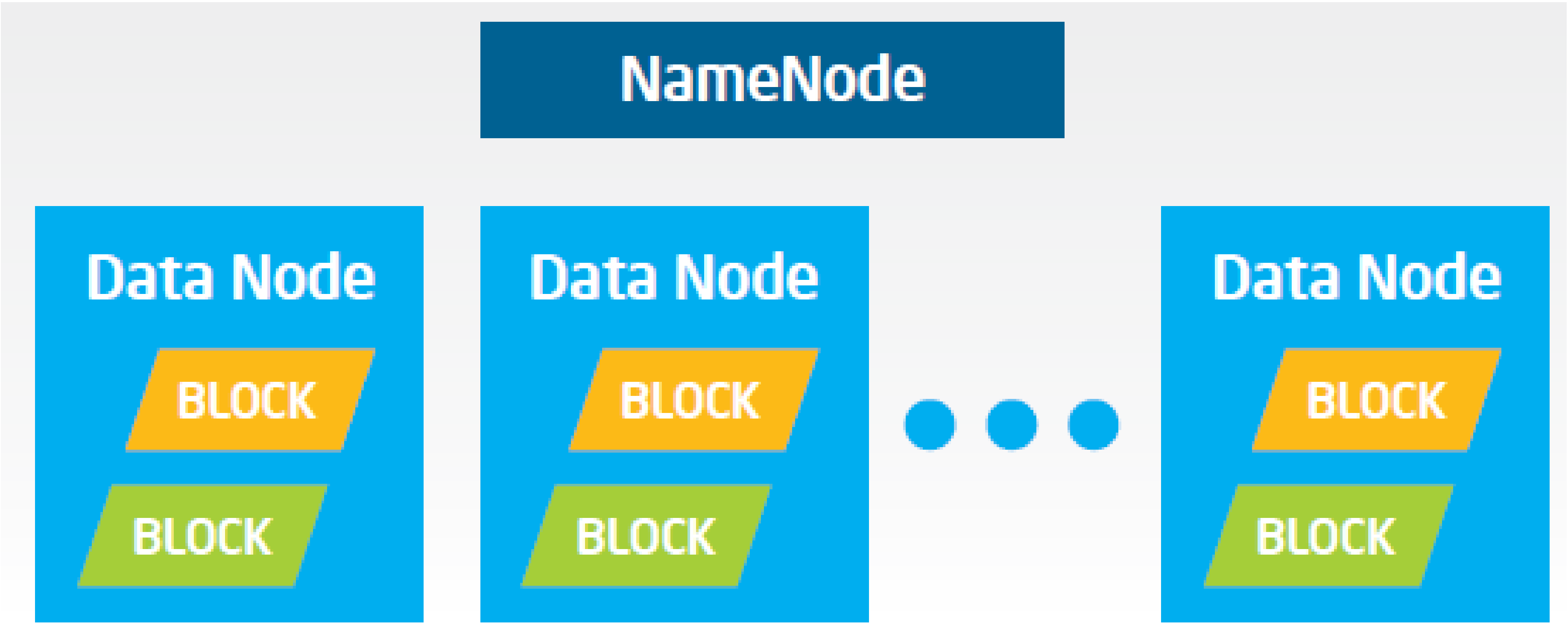


Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Big Data with Hadoop Architecture

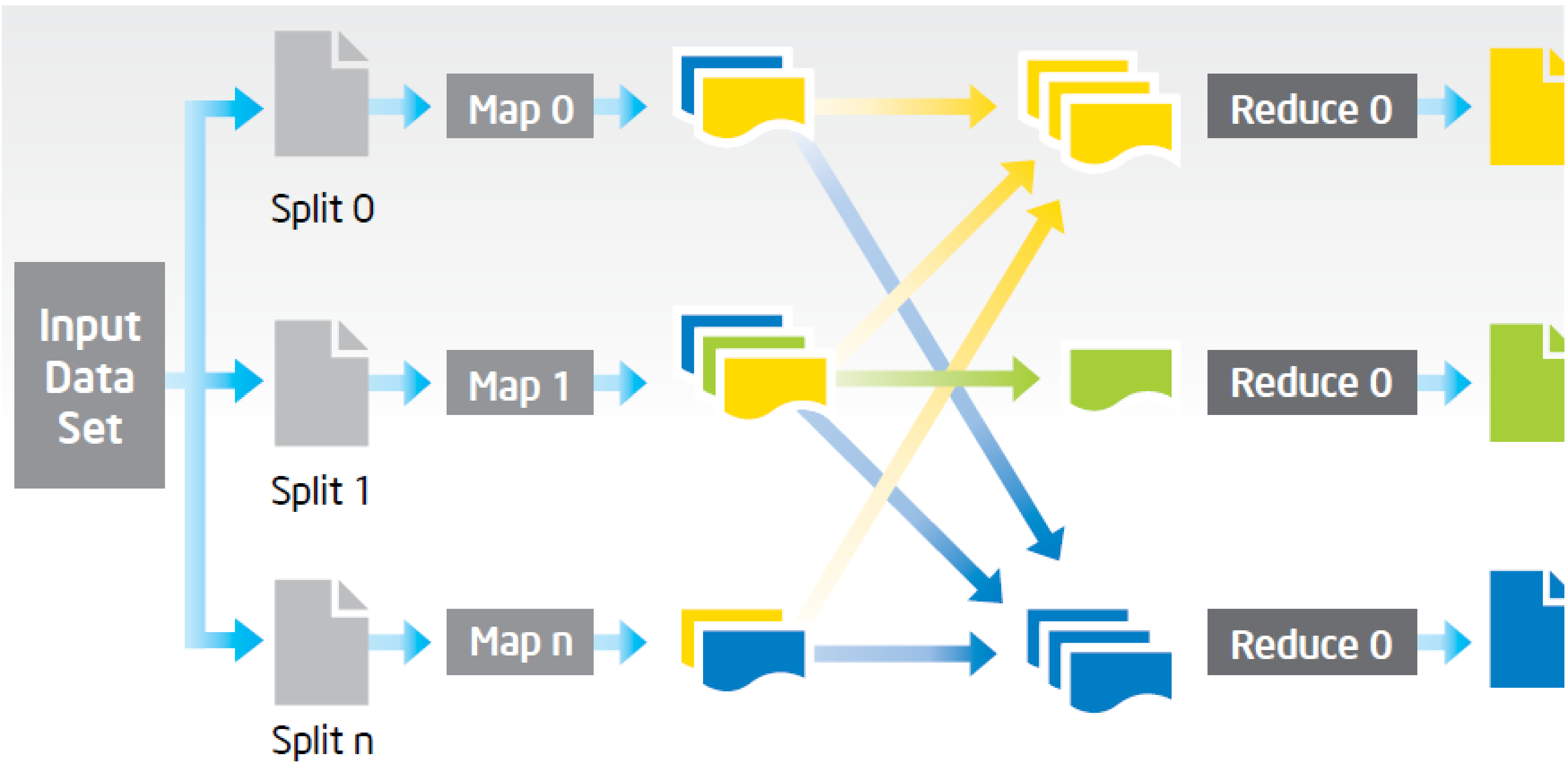
Logical Architecture

Storage: HDFS



Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

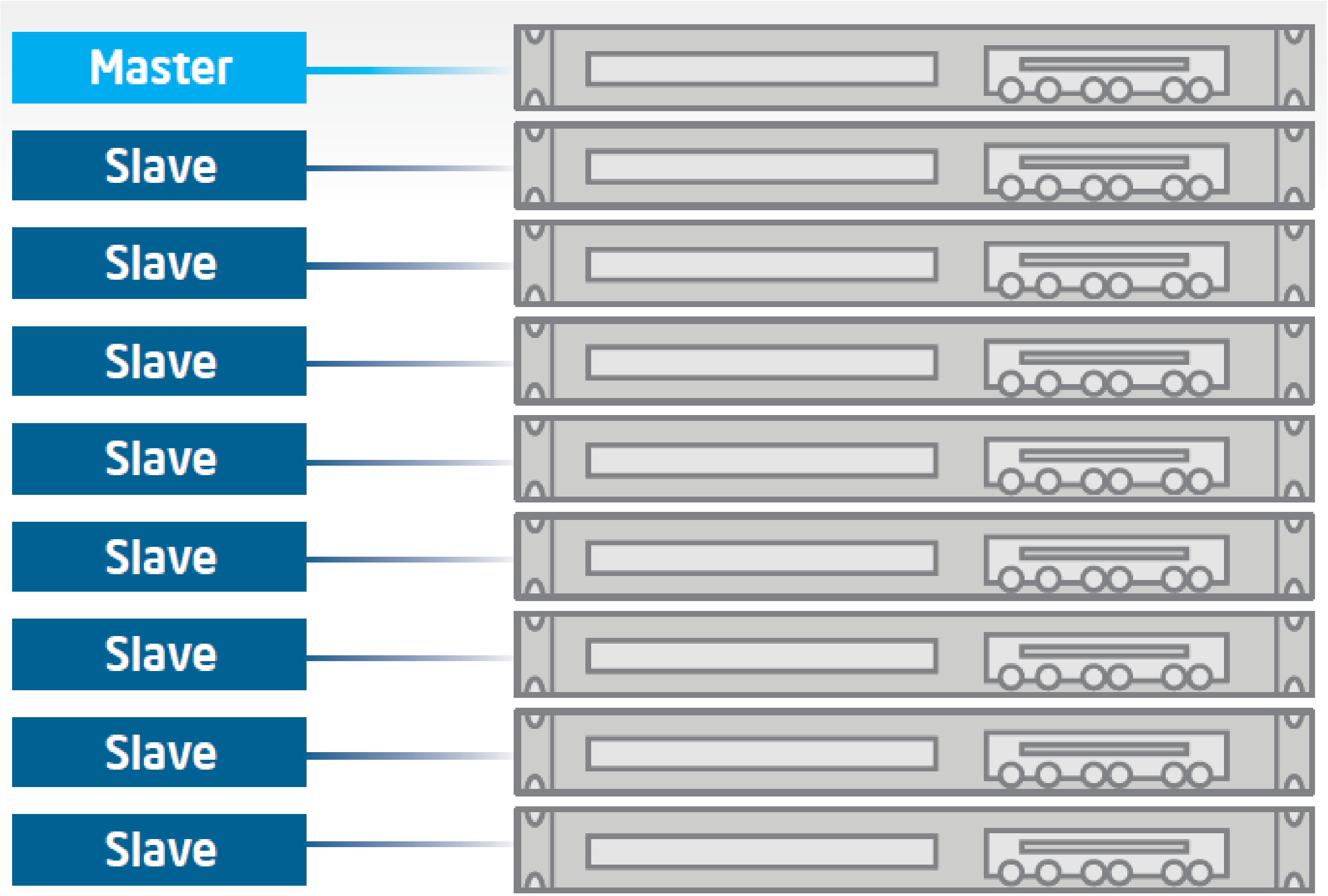
Big Data with Hadoop Architecture
Process Flow



Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

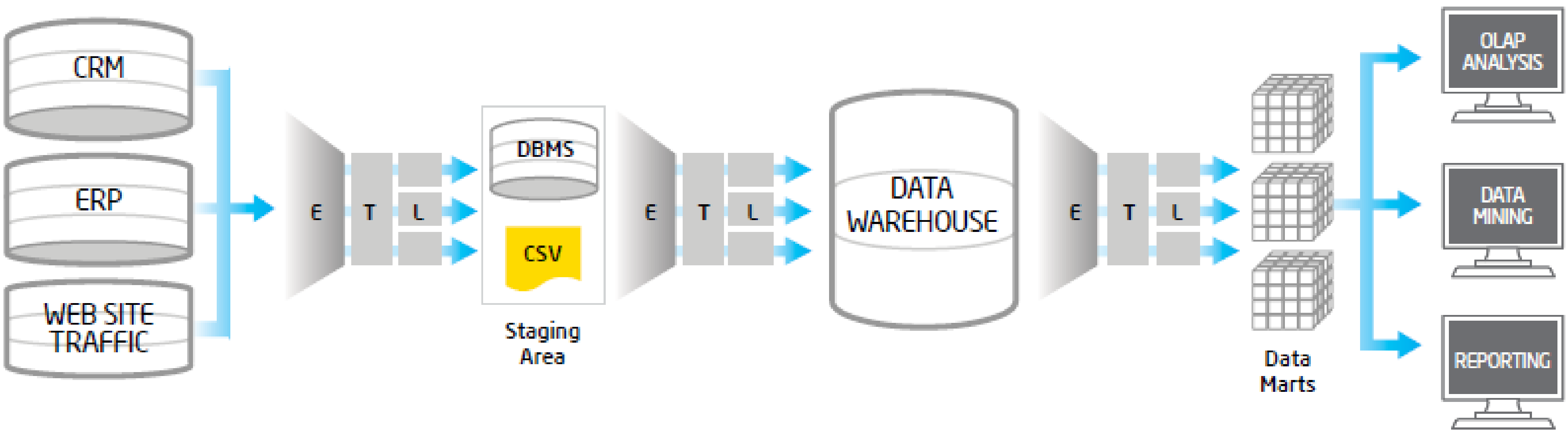
Big Data with Hadoop Architecture

Hadoop Cluster



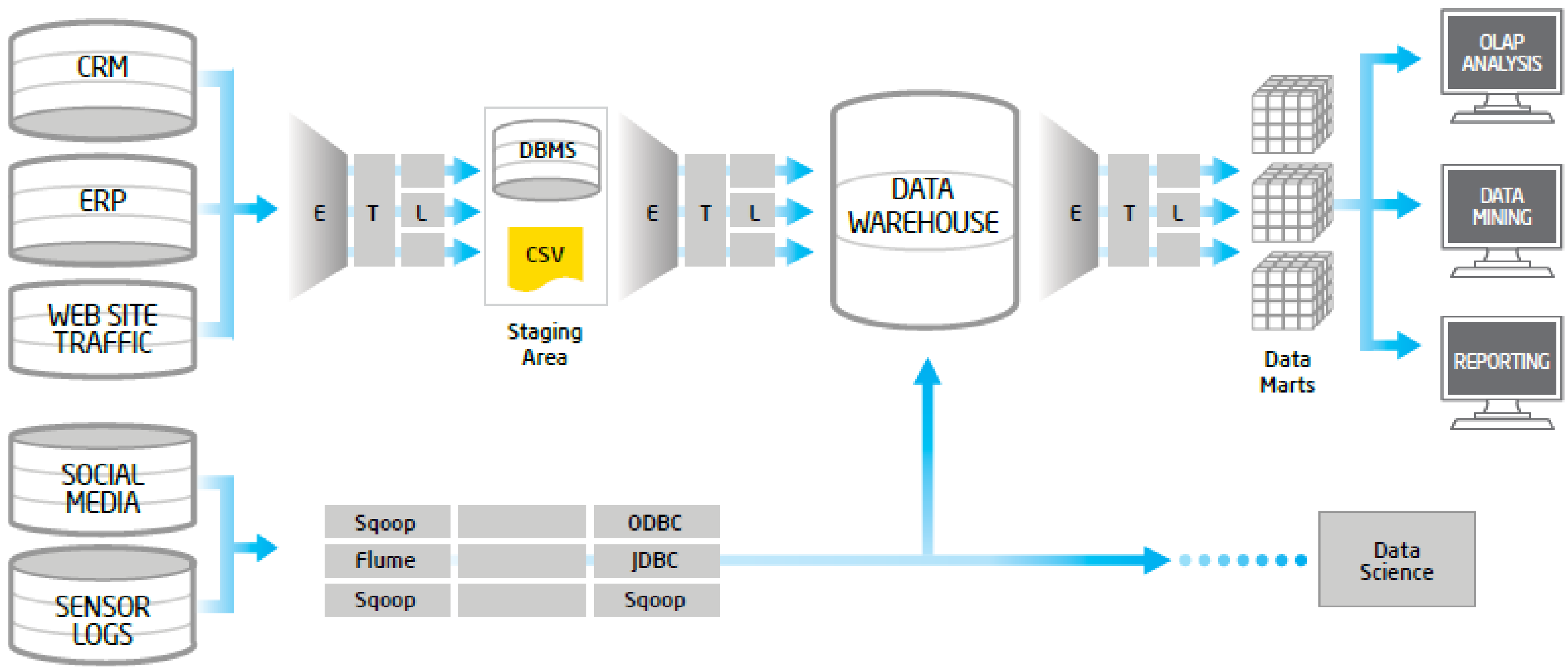
Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Traditional ETL Architecture



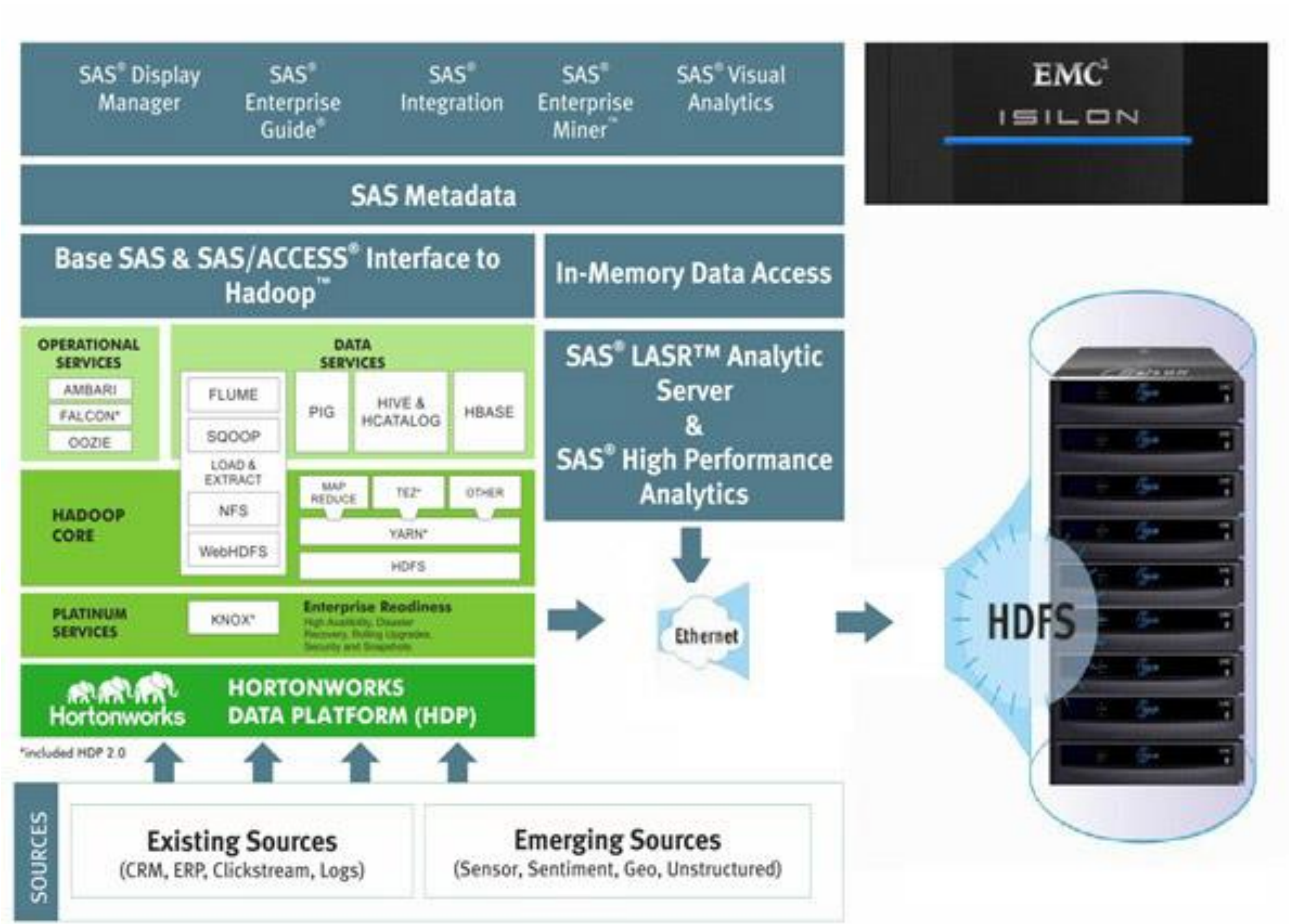
Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Offload ETL with Hadoop (Big Data Architecture)



Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

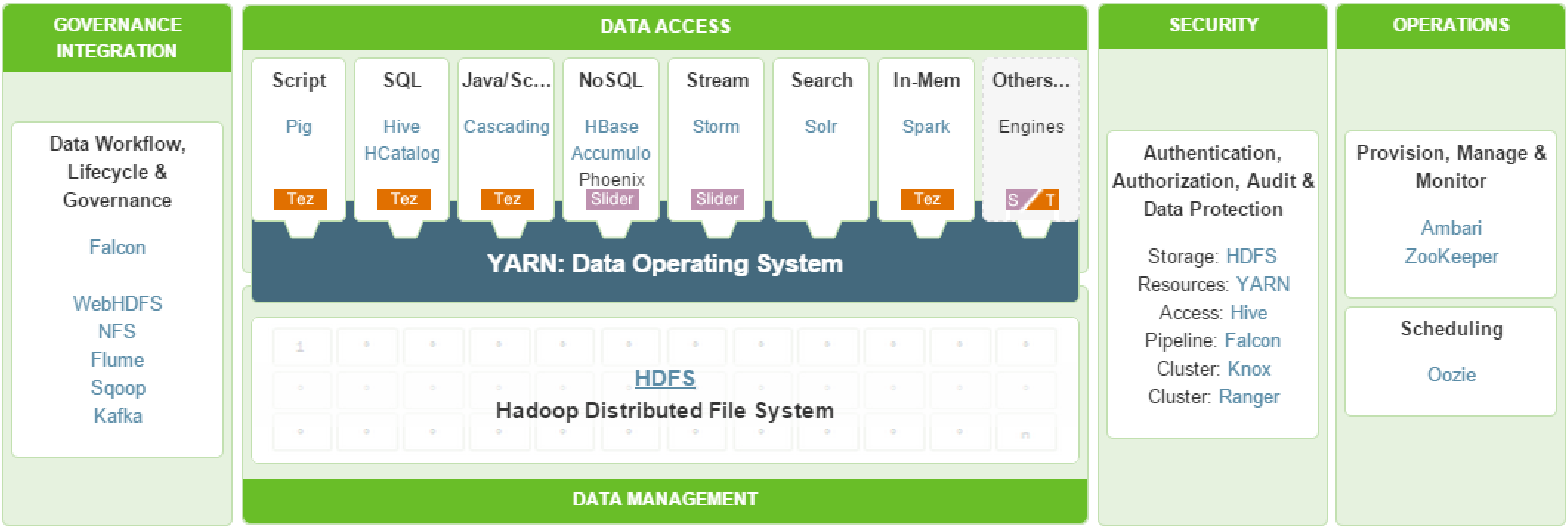
Big Data Solution



Source: http://www.newera-technologies.com/big-data-solution.html

HDP

A Complete Enterprise Hadoop Data Platform



Source: http://hortonworks.com/hdp/

Data Mining

Advanced Data Analysis

Evolution of
Database System Technology

Evolution of Database System Technology

Data Collection and Database Creation

(1960s and earlier)

- Primitive file processing

Database Management Systems

(1970s–early 1980s)

- Hierarchical and network database systems
- Relational database systems
- Query languages: SQL, etc.
- Transactions, concurrency control and recovery
- On-line transaction processing (OLTP)

Advanced Database Systems

(mid-1980s–present)

- Advanced data models: extended relational, object-relational, etc.
- Advanced applications: spatial, temporal, multimedia, active, stream and sensor, scientific and engineering, knowledge-based
- XML-based database systems
- Integration with information retrieval
- Data and information integration

Advanced Data Analysis:

(late 1980s–present)

- Data warehouse and OLAP
- Data mining and knowledge discovery: generalization, classification, association, clustering
- Advanced data mining applications: stream data mining, bio-data mining, time-series analysis, text mining, Web mining, intrusion detection, etc.
- Data mining applications
- Data mining and society

New Generation of Information Systems

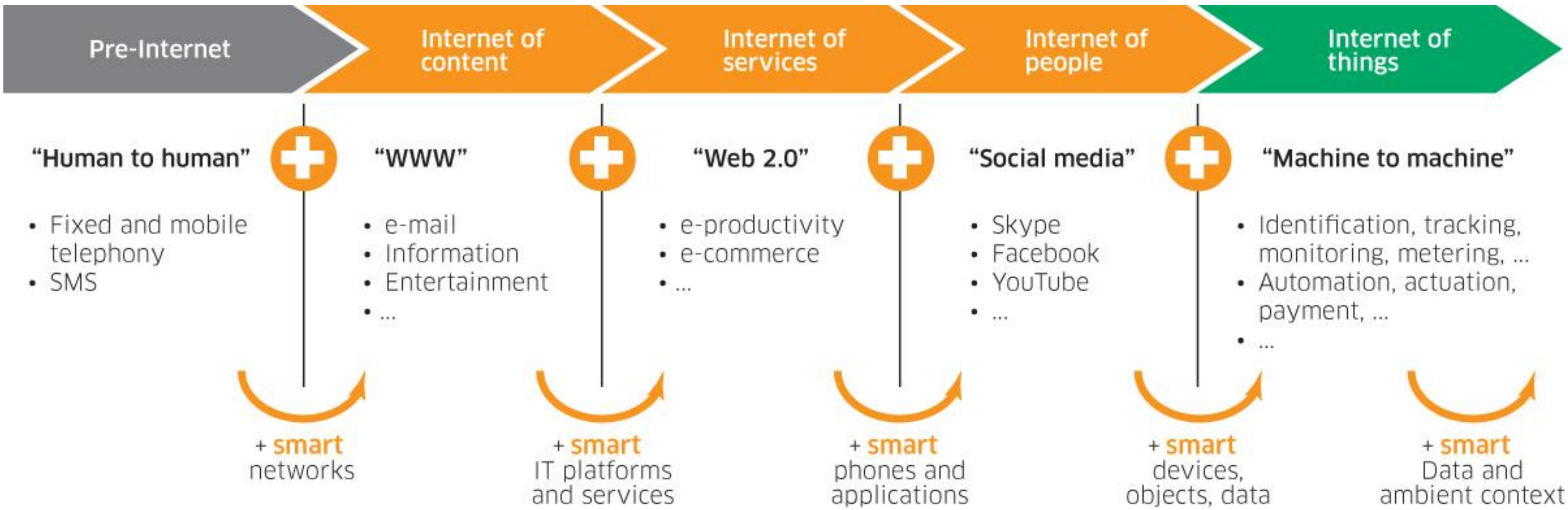
(present–future)

Source: Jiawei Han and Micheline Kamber (2011), Data Mining: Concepts and Techniques, Third Edition, Elsevier

Internet Evolution

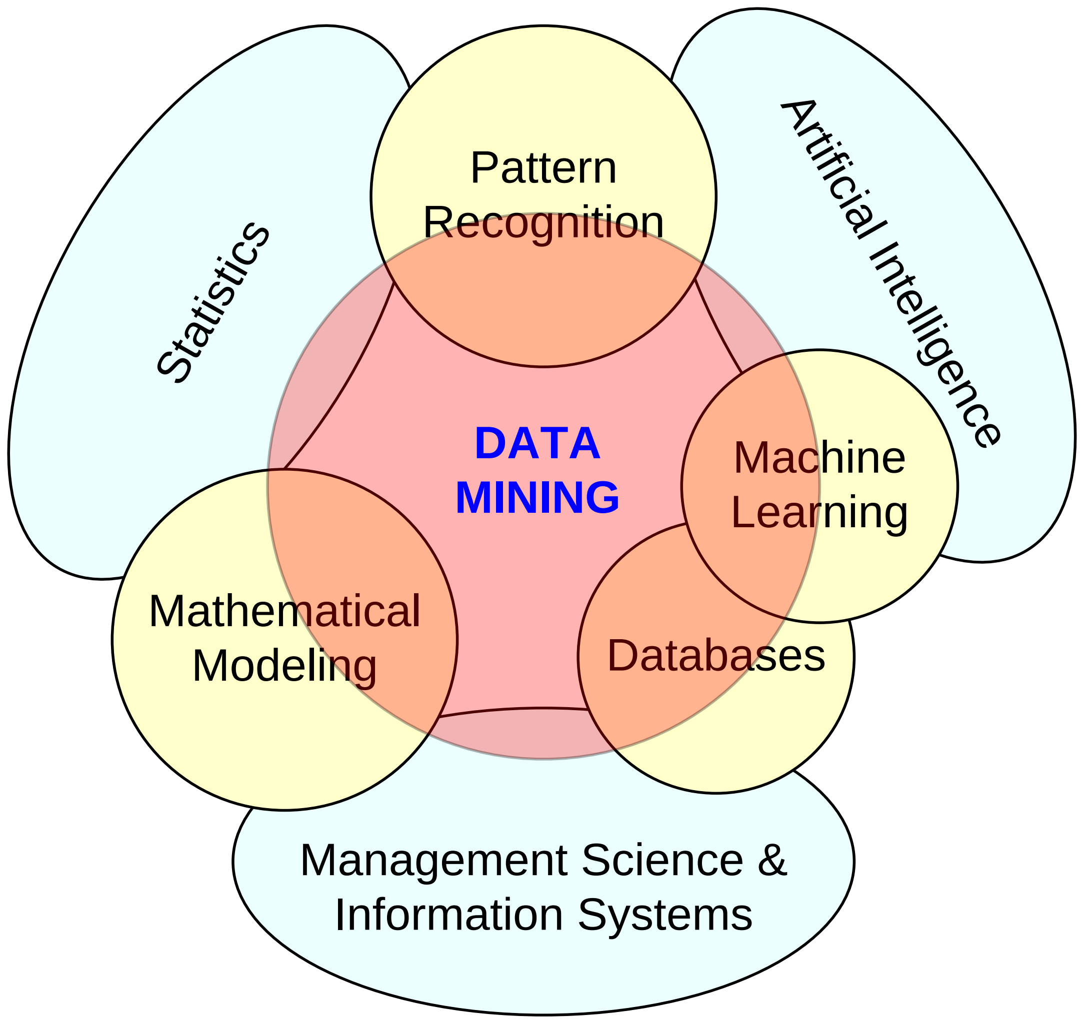
Internet of People (IoP): Social Media

Internet of Things (IoT): Machine to Machine



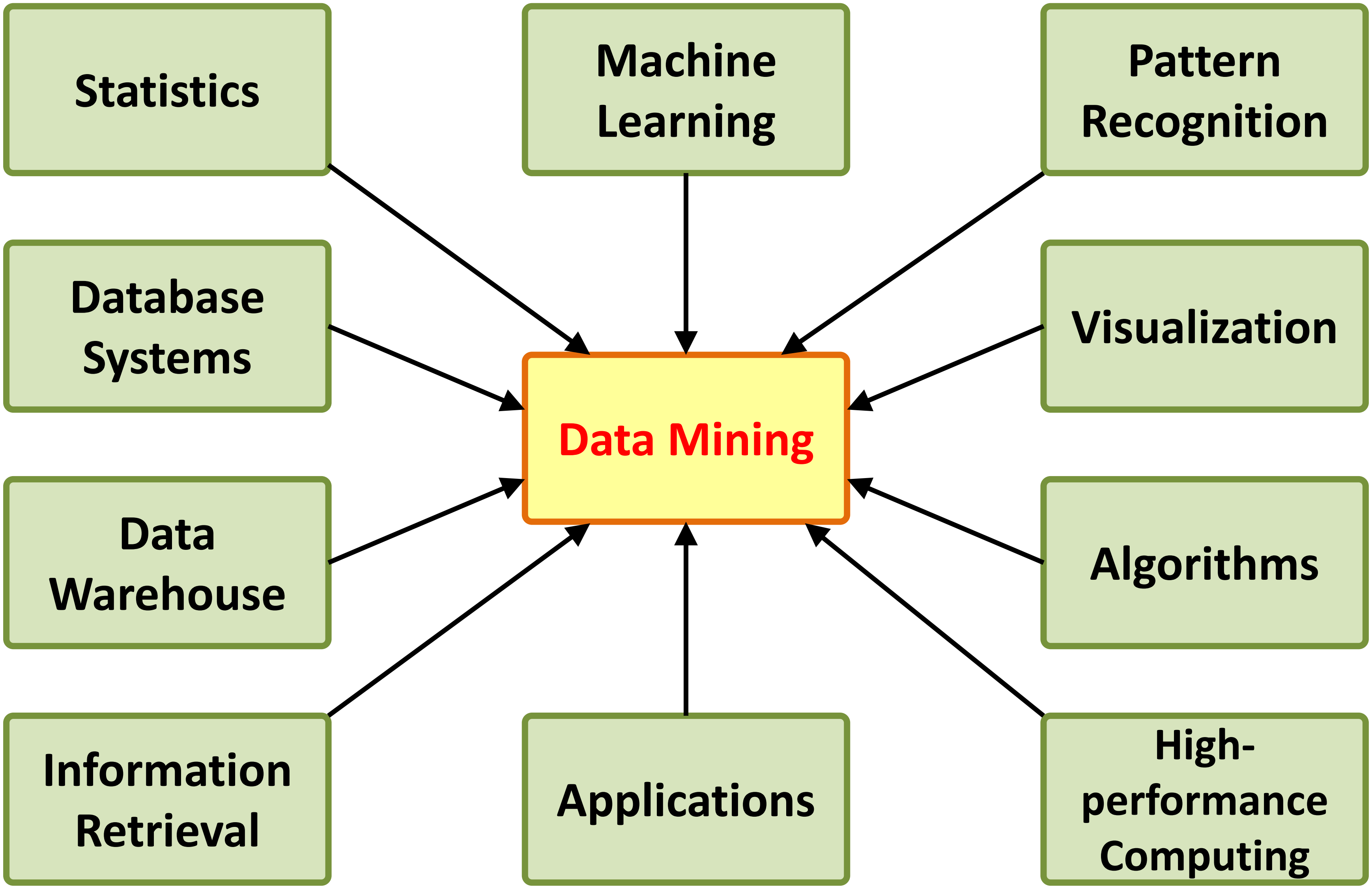
Source: Marc Jadoul (2015), The IoT: The next step in internet evolution, March 11, 2015
http://www2.alcatel-lucent.com/techzine/iot-internet-of-things-next-step-evolution/

Data Mining at the Intersection of Many Disciplines



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Mining Technologies



Source: Jiawei Han and Micheline Kamber (2011), Data Mining: Concepts and Techniques, Third Edition, Elsevier

Data Mining Process

Data Mining Process

- A manifestation of best practices
- A systematic way to conduct DM projects
- Different groups has different versions
- Most common standard processes:
  - CRISP-DM
    (Cross-Industry Standard Process for Data Mining)
  - SEMMA
    (Sample, Explore, Modify, Model, and Assess)
  - KDD
    (Knowledge Discovery in Databases)

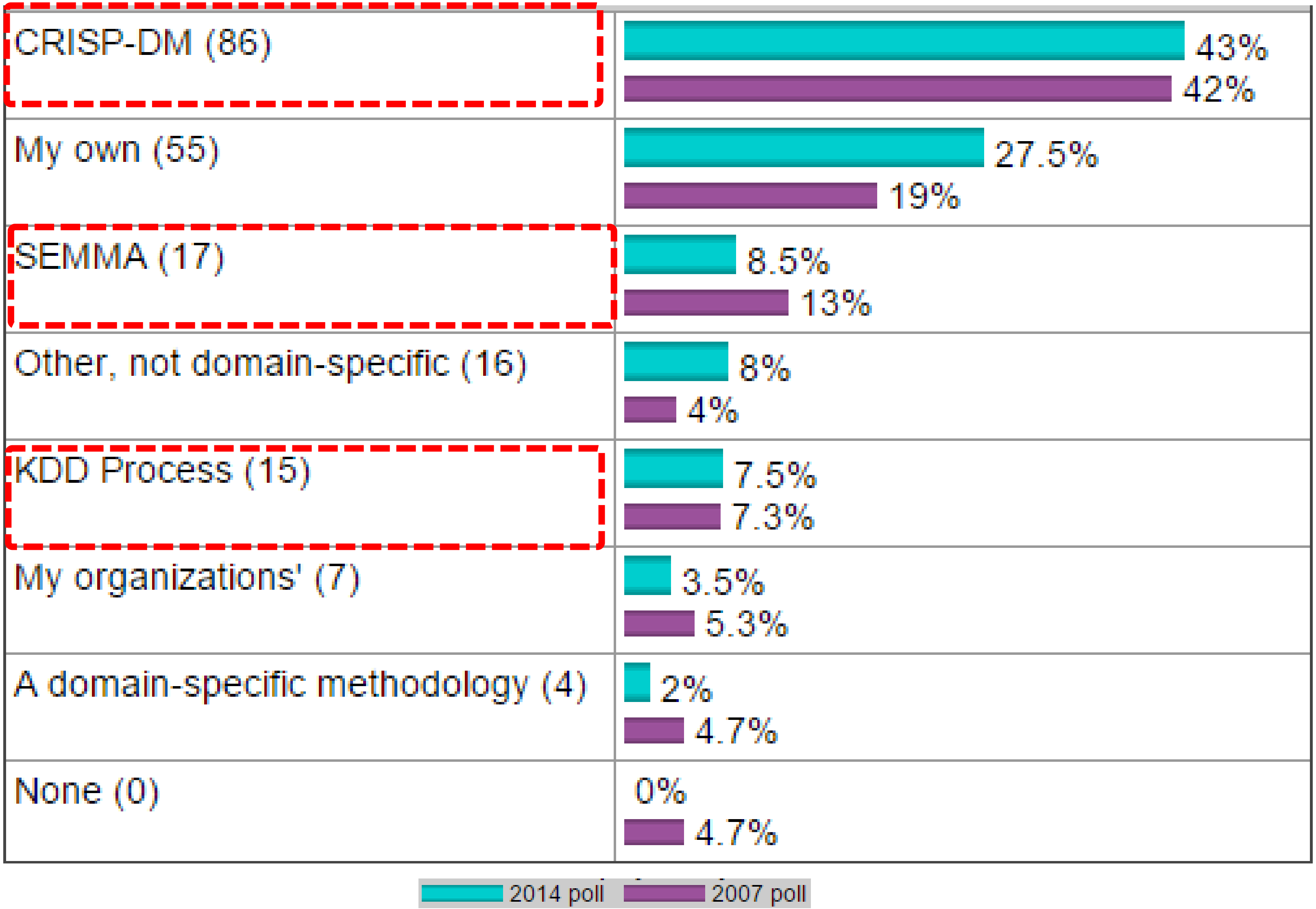
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Mining Process (SOP of DM)

What main methodology are you using for your analytics,
data mining,
or data science projects ?

Source: http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

Data Mining Process



Source: http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

Data Mining:

Core Analytics Process

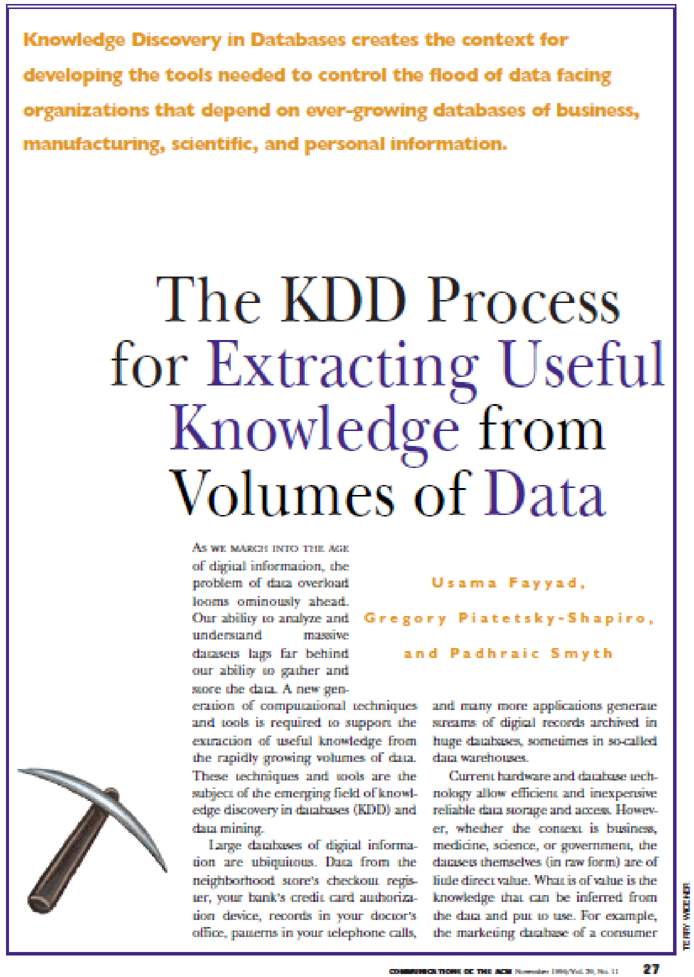
The KDD Process for Extracting Useful Knowledge from Volumes of Data

Source: Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM, 39(11), 27-34.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996).

The KDD Process for Extracting Useful Knowledge from Volumes of Data.

Communications of the ACM, 39(11), 27-34.



Knowledge Discovery in Databases creates the context for developing the tools needed to control the flood of data facing organizations that depend on ever-growing databases of business, manufacturing, scientific, and personal information.

The KDD Process for Extracting Useful Knowledge from Volumes of Data

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth

As we march into the age of digital information, the problem of data overload looms ominously ahead. Our ability to analyze and understand massive datasets lags far behind our ability to gather and store the data. A new generation of computational techniques and tools is required to support the extraction of useful knowledge from the rapidly growing volumes of data. These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD) and data mining.

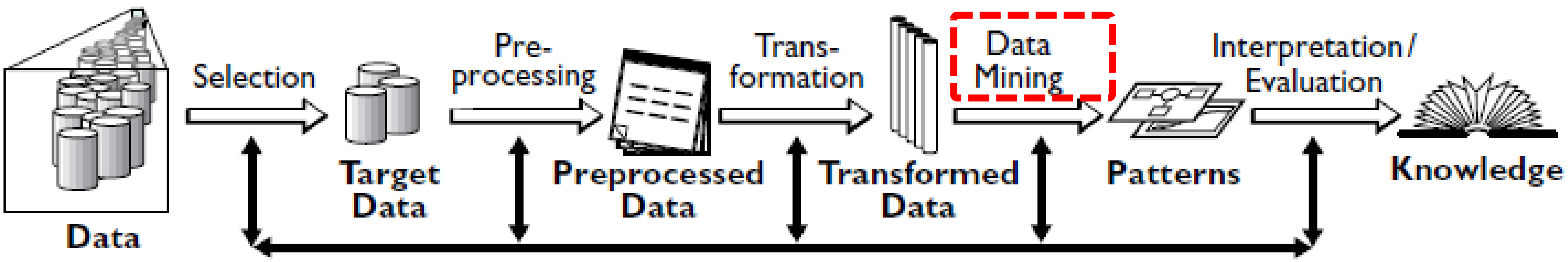
Large databases of digital information are ubiquitous. Data from the neighborhood store's checkout register, your bank's credit card authorization device, records in your doctor's office, patterns in your telephone calls, and many more applications generate streams of digital records archived in huge databases, sometimes in so-called data warehouses.

Current hardware and database technology allow efficient and inexpensive reliable data storage and access. However, whether the context is business, medicine, science, or government, the datasets themselves (in raw form) are of little direct value. What is of value is the knowledge that can be inferred from the data and put to use. For example, the marketing database of a consumer

Data Mining

Knowledge Discovery in Databases (KDD) Process

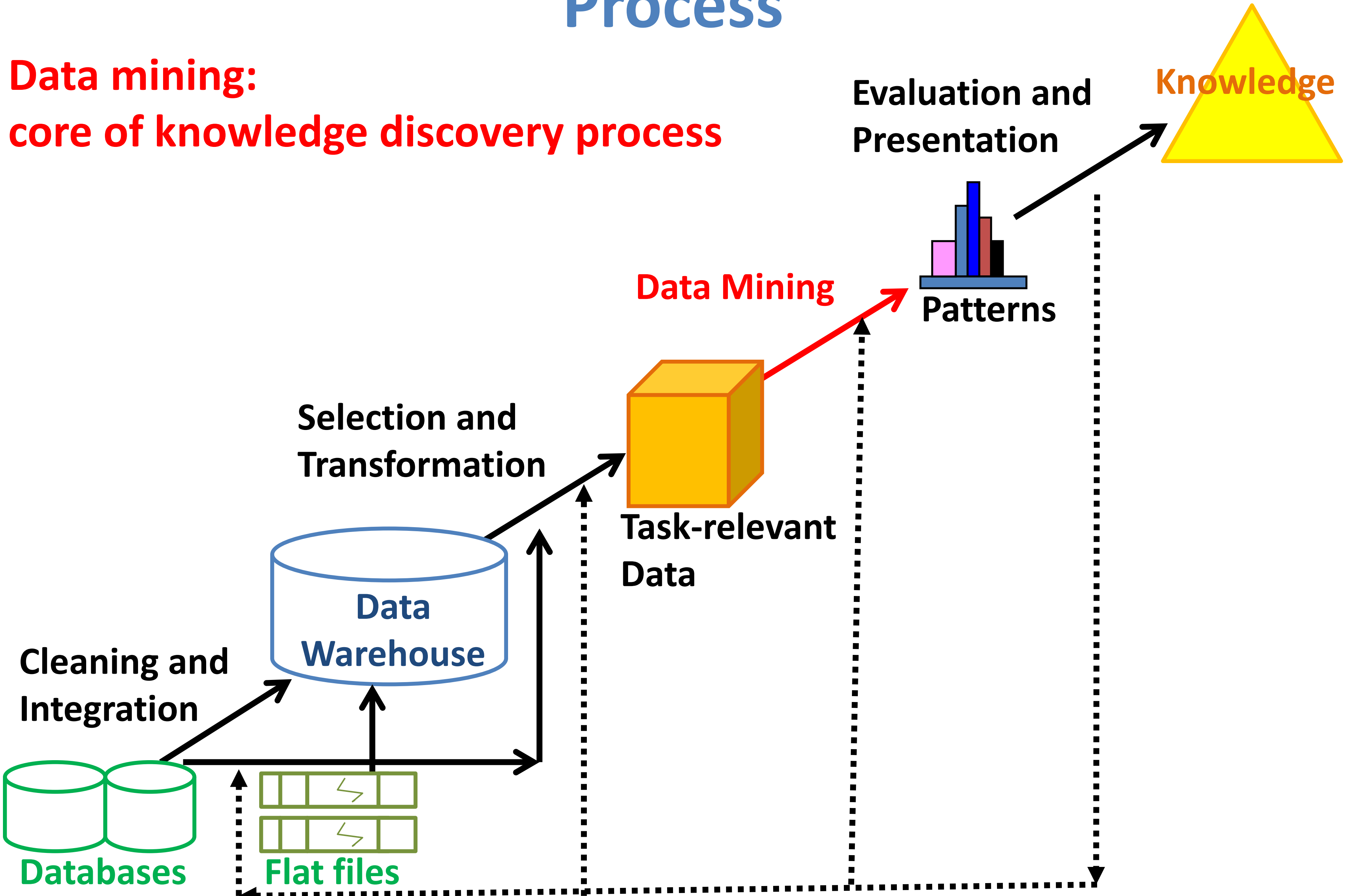
(Fayyad et al., 1996)



Source: Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM, 39(11), 27-34.

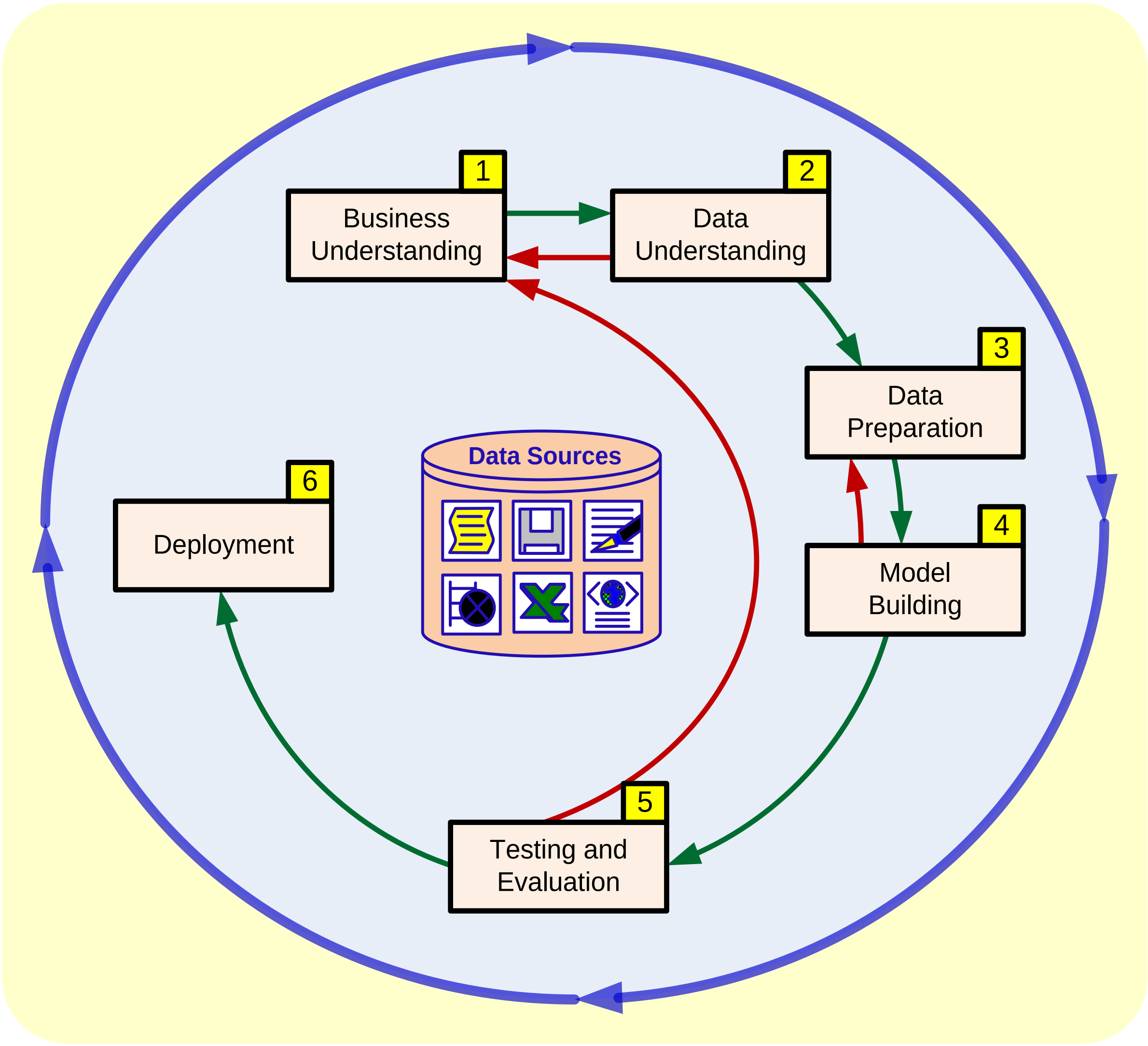
Knowledge Discovery in Databases (KDD) Process

Data mining:
core of knowledge discovery process



Source: Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Second Edition, Elsevier

Data Mining Process: CRISP-DM



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Mining Process: CRISP-DM

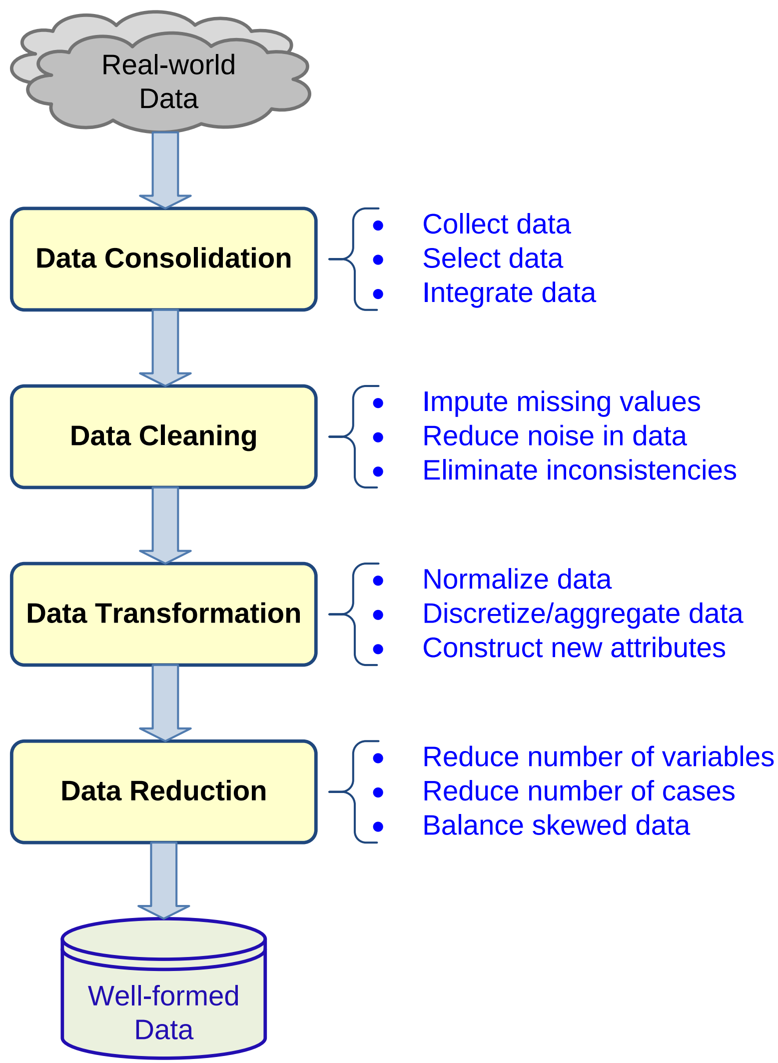
Step 1: Business Understanding
Step 2: Data Understanding
Step 3: Data Preparation (!)
Step 4: Model Building
Step 5: Testing and Evaluation
Step 6: Deployment

Accounts for ~85% of total project time

- The process is highly repetitive and experimental (DM: art versus science?)

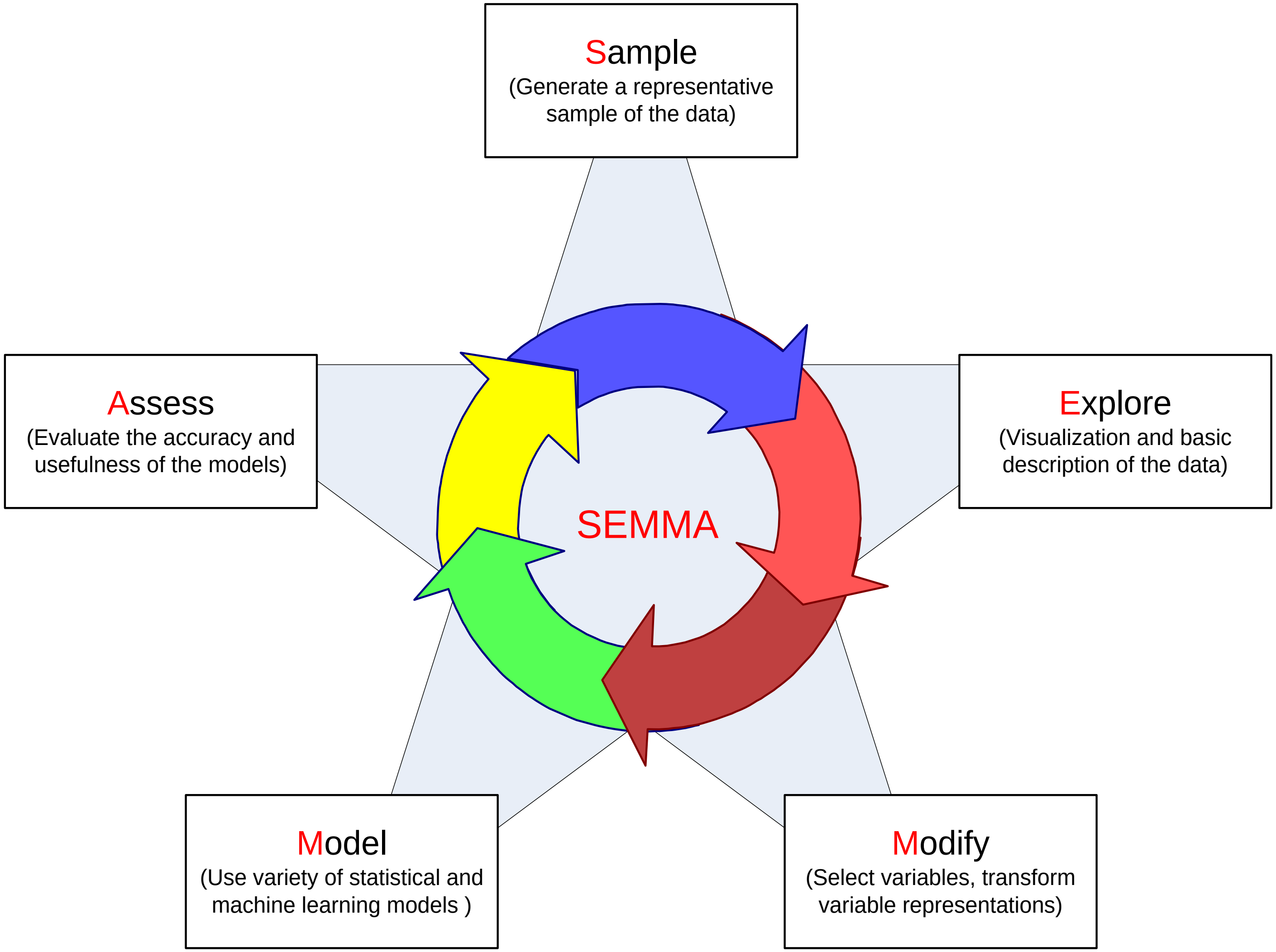
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Preparation – A Critical DM Task



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

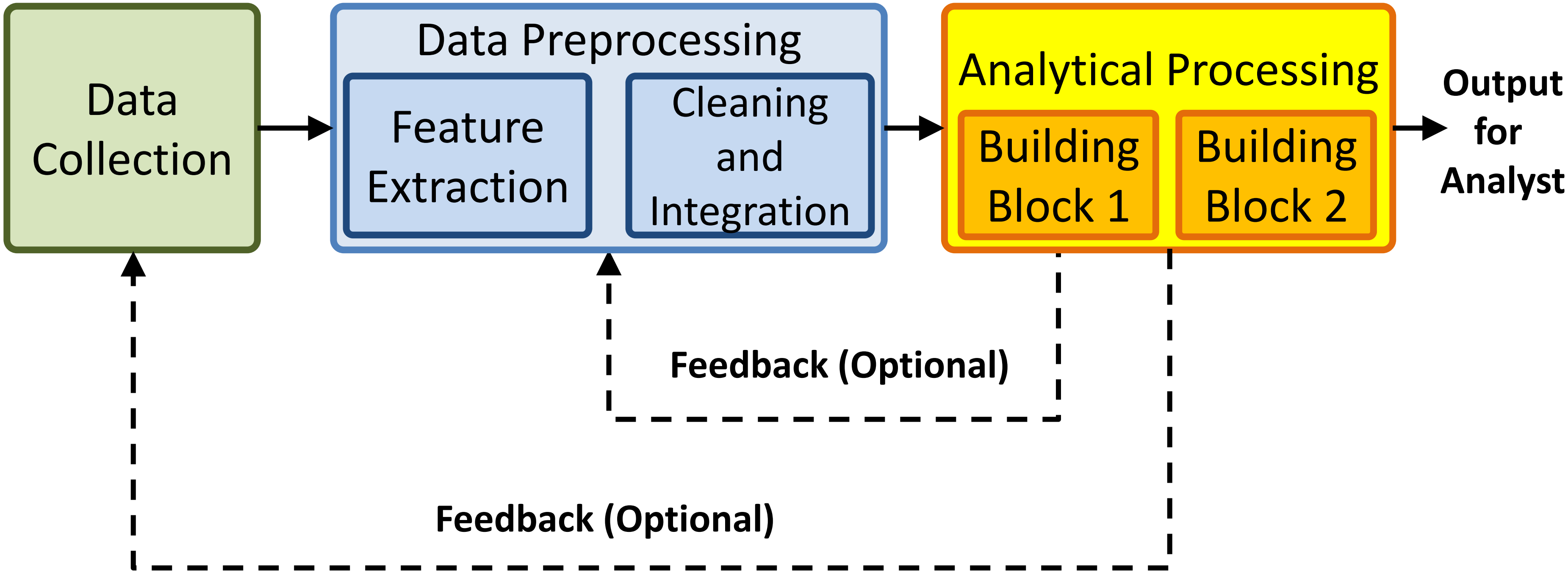
Data Mining Process: SEMMA



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Mining Processing Pipeline

(Charu Aggarwal, 2015)



Source: Charu Aggarwal (2015), Data Mining: The Textbook Hardcover, Springer

Using Databases to Improve Business Performance and Decision Making

- **Big data**
  - **Massive sets of unstructured/semi-structured data from Web traffic, social media, sensors, and so on**
  - **Petabytes, exabytes of data**
    - Volumes too great for typical DBMS
  - **Can reveal more patterns and anomalies**

Source: Kenneth C. Laudon & Jane P. Laudon (2014), Management Information Systems: Managing the Digital Firm, Thirteenth Edition, Pearson.

Using Databases to Improve Business Performance and Decision Making

- **Business intelligence infrastructure**
  - Today includes an array of tools for separate systems, and big data
- **Contemporary tools**:
  - Data warehouses
  - Data marts
  - Hadoop
  - In-memory computing
  - Analytical platforms

Source: Kenneth C. Laudon & Jane P. Laudon (2014), Management Information Systems: Managing the Digital Firm, Thirteenth Edition, Pearson.

Data Warehouse vs. Data Marts

- **Data warehouse:**
  - Stores current and historical data from many core operational transaction systems
  - Consolidates and standardizes information for use across enterprise, but data cannot be altered
  - Provides analysis and reporting tools
- **Data marts:**
  - Subset of data warehouse
  - Summarized or focused portion of data for use by specific population of users
  - Typically focuses on single subject or line of business

Source: Kenneth C. Laudon & Jane P. Laudon (2014), Management Information Systems: Managing the Digital Firm, Thirteenth Edition, Pearson.

Hadoop

- Enables distributed parallel processing of big data across inexpensive computers
- Key services
  - Hadoop Distributed File System (HDFS): data storage
  - MapReduce: breaks data into clusters for work
  - Hbase: NoSQL database
- Used by Facebook, Yahoo, NextBio

Source: Kenneth C. Laudon & Jane P. Laudon (2014), Management Information Systems: Managing the Digital Firm, Thirteenth Edition, Pearson.

In-memory computing

- Used in big data analysis
- Use computers main memory (RAM) for data storage to avoid delays in retrieving data from disk storage
- Can reduce hours/days of processing to seconds
- Requires optimized hardware

Source: Kenneth C. Laudon & Jane P. Laudon (2014), Management Information Systems: Managing the Digital Firm, Thirteenth Edition, Pearson.

Analytic platforms

- High-speed platforms using both relational and non-relational tools optimized for large datasets
- Examples:
  - IBM Netezza
  - Oracle Exadata

Source: Kenneth C. Laudon & Jane P. Laudon (2014), Management Information Systems: Managing the Digital Firm, Thirteenth Edition, Pearson.

Analytical tools: Relationships, patterns, trends

- Business Intelligence Analytics and Applications
- Tools for consolidating, analyzing, and providing access to vast amounts of data to help users make better business decisions
  - Multidimensional data analysis (OLAP)
  - Data mining
  - Text mining
  - Web mining

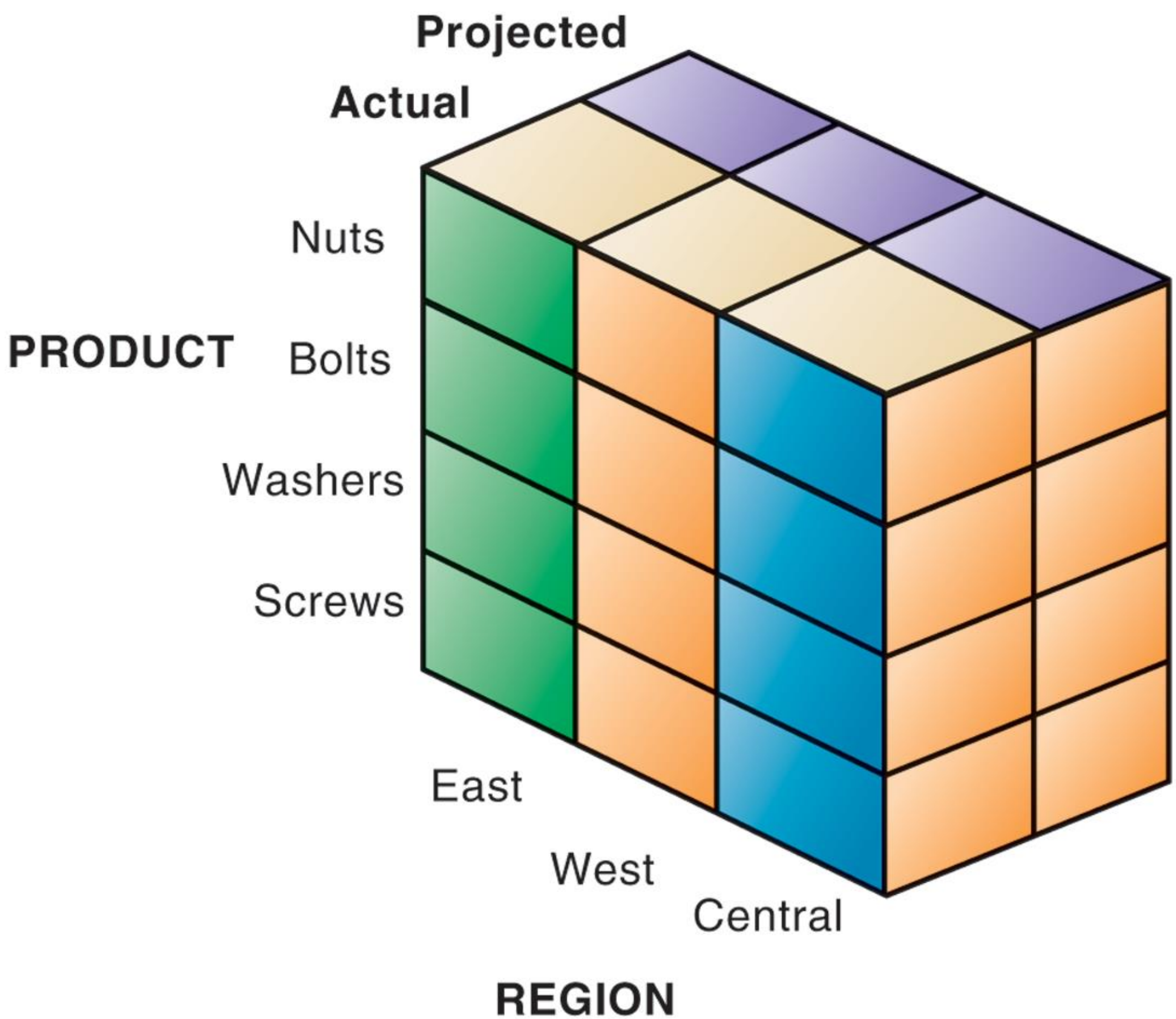
Source: Kenneth C. Laudon & Jane P. Laudon (2014), Management Information Systems: Managing the Digital Firm, Thirteenth Edition, Pearson.

Online analytical processing (OLAP)

- Supports multidimensional data analysis
  - Viewing data using multiple dimensions
  - Each aspect of information (product, pricing, cost, region, time period) is different dimension
  - Example: How many washers sold in East in June compared with other regions?
- OLAP enables rapid, online answers to ad hoc queries

Source: Kenneth C. Laudon & Jane P. Laudon (2014), Management Information Systems: Managing the Digital Firm, Thirteenth Edition, Pearson.

MULTIDIMENSIONAL DATA MODEL



Source: Kenneth C. Laudon & Jane P. Laudon (2014), Management Information Systems: Managing the Digital Firm, Thirteenth Edition, Pearson.

Data mining

- Finds hidden patterns, relationships in datasets
  - Example: customer buying patterns
- Infers rules to predict future behavior
  - Data mining provides insights into data that cannot be discovered through OLAP, by inferring rules from patterns in data.

Source: Kenneth C. Laudon & Jane P. Laudon (2014), Management Information Systems: Managing the Digital Firm, Thirteenth Edition, Pearson.

Types of Information Obtained from Data Mining

- **Associations**: Occurrences linked to single event
- **Sequences**: Events linked over time
- **Classification**: Recognizes patterns that describe group to which item belongs
- **Clustering**: Similar to classification when no groups have been defined; finds groupings within data
- **Forecasting**: Uses series of existing values to forecast what other values will be

Source: Kenneth C. Laudon & Jane P. Laudon (2014), Management Information Systems: Managing the Digital Firm, Thirteenth Edition, Pearson.

Text mining

- Extracts key elements from large unstructured data sets
  - Stored e-mails
  - Call center transcripts
  - Legal cases
  - Patent descriptions
  - Service reports, and so on
- Sentiment analysis software
  - Mines e-mails, blogs, social media to detect opinions

Source: Kenneth C. Laudon & Jane P. Laudon (2014), Management Information Systems: Managing the Digital Firm, Thirteenth Edition, Pearson.

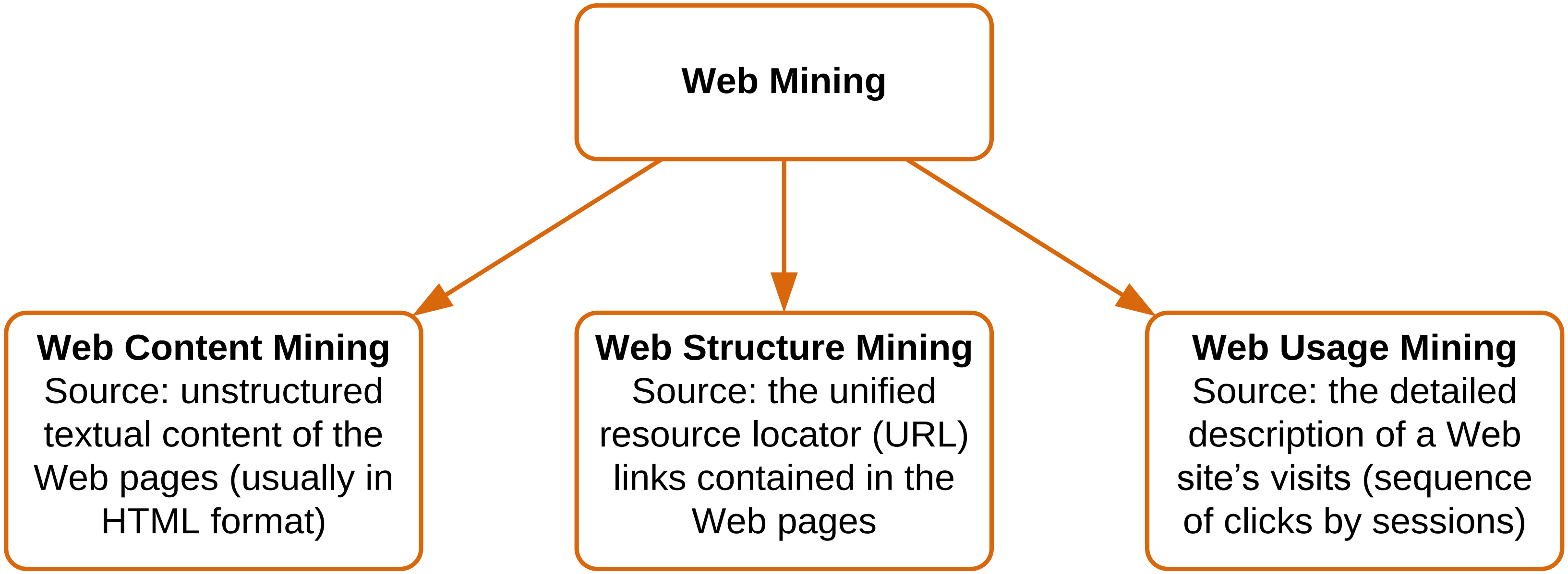
Web mining

- Discovery and analysis of useful patterns and information from Web
  - Understand customer behavior
  - Evaluate effectiveness of Web site, and so on
- 3 Tasks of Web Mining
  - Web content mining
    - Mines content of Web pages
  - Web structure mining
    - Analyzes links to and from Web page
  - Web usage mining
    - Mines user interaction data recorded by Web server

Source: Kenneth C. Laudon & Jane P. Laudon (2014), Management Information Systems: Managing the Digital Firm, Thirteenth Edition, Pearson.

Web Mining

- Web mining (or Web data mining) is the <u>process</u> of discovering intrinsic relationships from Web data (textual, linkage, or usage)



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Databases and the Web

- Many companies use Web to make some internal databases available to customers or partners
- Typical configuration includes:
  - Web server
  - Application server/middleware/CGI scripts
  - Database server (hosting DBMS)
- Advantages of using Web for database access:
  - Ease of use of browser software
  - Web interface requires few or no changes to database
  - Inexpensive to add Web interface to system

Source: Kenneth C. Laudon & Jane P. Laudon (2014), Management Information Systems: Managing the Digital Firm, Thirteenth Edition, Pearson.

Web Content/Structure Mining

- Mining of the textual content on the Web
- Data collection via Web crawlers

- Web pages include hyperlinks
  - Authoritative pages
  - Hubs
  - hyperlink-induced topic search (HITS) alg

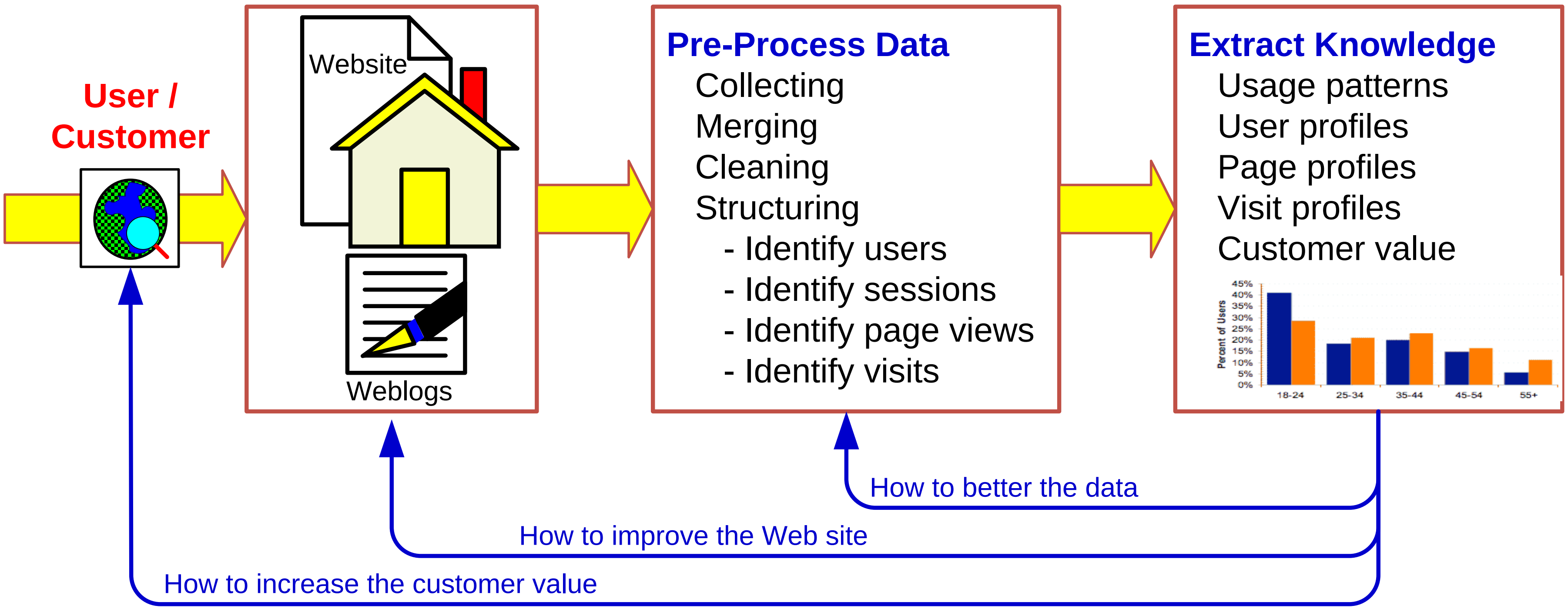
Web Usage Mining

- Extraction of information from data generated through Web page visits and transactions…
  - data stored in server access logs, referrer logs, agent logs, and client-side cookies
  - user characteristics and usage profiles
  - metadata, such as page attributes, content attributes, and usage data
- Clickstream data
- Clickstream analysis

Web Usage Mining

- Web usage mining applications
  - Determine the lifetime value of clients
  - Design cross-marketing strategies across products.
  - Evaluate promotional campaigns
  - Target electronic ads and coupons at user groups based on user access patterns
  - Predict user behavior based on previously learned rules and users' profiles
  - Present dynamic information to users based on their interests and profiles…

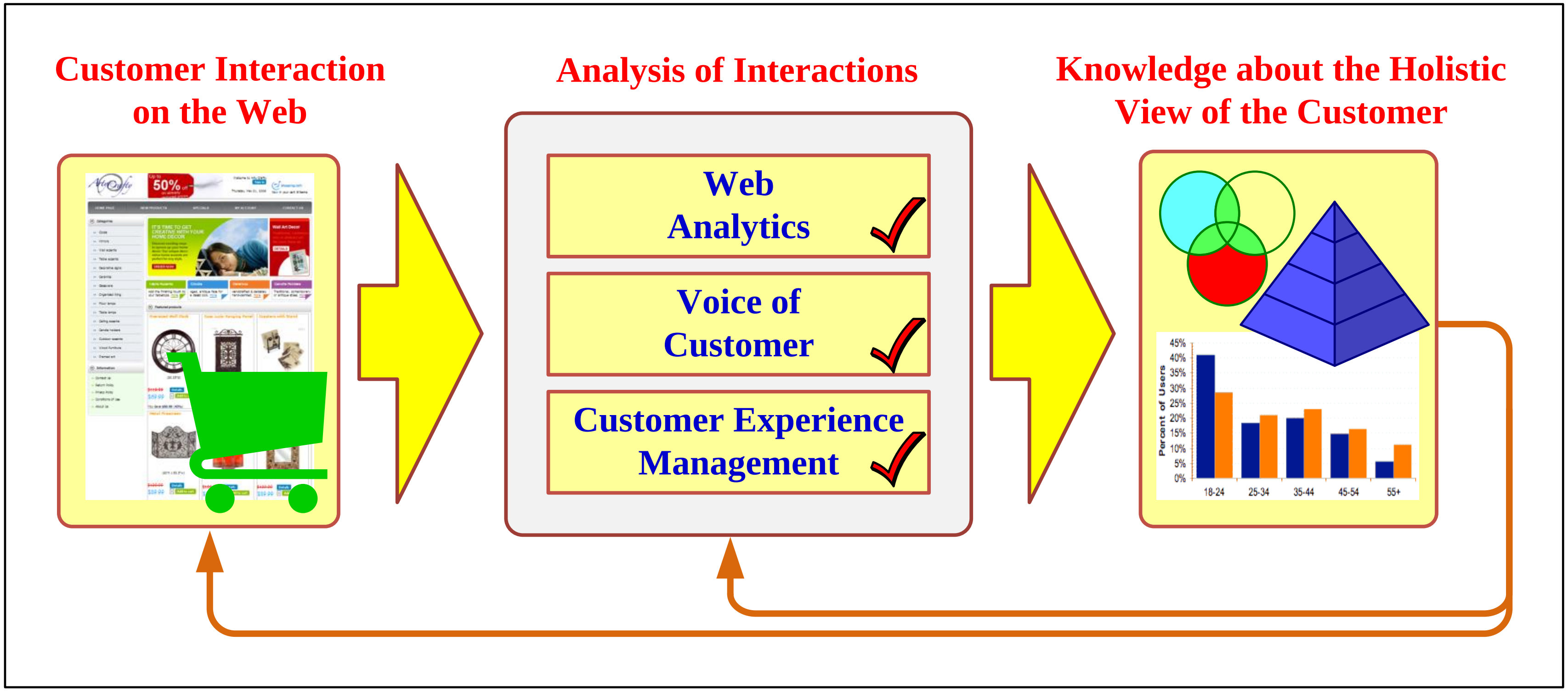
Web Usage Mining
(clickstream analysis)



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Web Mining Success Stories

- Amazon.com, Ask.com, Scholastic.com, …
- Website Optimization Ecosystem



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Mining Tasks

A Taxonomy for Data Mining Tasks

Data Mining	Learning Method	Popular Algorithms
Prediction	Supervised	Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Classification	Supervised	Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms
Regression	Supervised	Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM
Association	Unsupervised	Apriory, OneR, ZeroR, Eclat
Link analysis	Unsupervised	Expectation Maximization, Apriory Algorithm, Graph-based Matching
Sequence analysis	Unsupervised	Apriory Algorithm, FP-Growth technique
Clustering	Unsupervised	K-means, ANN/SOM
Outlier analysis	Unsupervised	K-means, Expectation Maximization (EM)

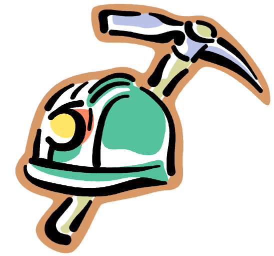
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Why Data Mining?

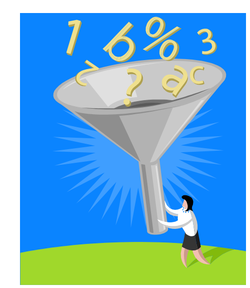
- More intense competition at the global scale
- Recognition of the value in data sources
- Availability of quality data on customers, vendors, transactions, Web, etc.
- Consolidation and integration of data repositories into data warehouses
- The exponential increase in data processing and storage capabilities; and decrease in cost
- Movement toward conversion of information resources into nonphysical form

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Definition of Data Mining



- The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases.
  - Fayyad et al., (1996)
- Keywords in this definition: Process, nontrivial, valid, novel, potentially useful, understandable.
- Data mining: a misnomer?
- Other names:
  - knowledge extraction, pattern analysis, knowledge discovery, information harvesting, pattern searching, data dredging,...



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

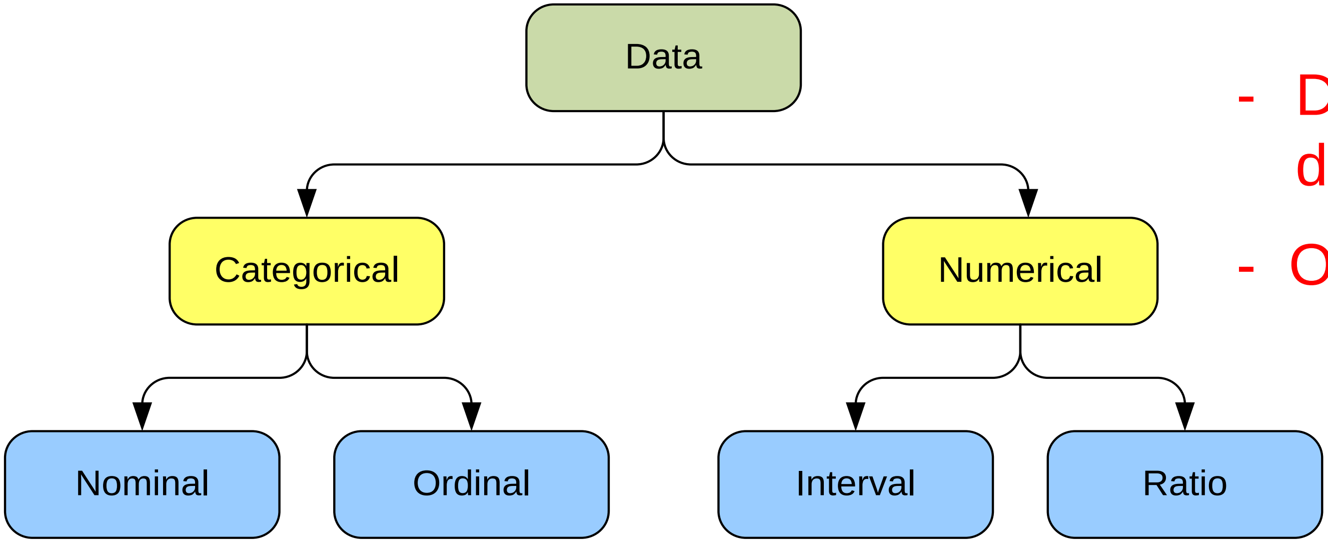
Data Mining Characteristics/Objectives

- Source of data for DM is often a consolidated data warehouse (not always!)
- DM environment is usually a client-server or a Web-based information systems architecture
- Data is the most critical ingredient for DM which may include soft/unstructured data
- The miner is often an end user
- Striking it rich requires creative thinking
- Data mining tools' capabilities and ease of use are essential (Web, Parallel processing, etc.)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data in Data Mining

- Data: a collection of facts usually obtained as the result of experiences, observations, or experiments
- Data may consist of numbers, words, images, …
- Data: lowest level of abstraction (from which information and knowledge are derived)



- DM with different data types?
- Other data types?

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

What Does DM Do?

- DM extract patterns from data
  - Pattern?
    A mathematical (numeric and/or symbolic) relationship among data items

- Types of patterns
  - Association
  - Prediction
  - Cluster (segmentation)
  - Sequential (or time series) relationships

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Mining Applications

- Customer Relationship Management
  - Maximize return on marketing campaigns
  - Improve customer retention (churn analysis)
  - Maximize customer value (cross-, up-selling)
  - Identify and treat most valued customers
- Banking and Other Financial
  - Automate the loan application process
  - Detecting fraudulent transactions
  - Optimizing cash reserves with forecasting

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Mining Applications (cont.)

- Retailing and Logistics
  - Optimize inventory levels at different locations
  - Improve the store layout and sales promotions
  - Optimize logistics by predicting seasonal effects
  - Minimize losses due to limited shelf life
- Manufacturing and Maintenance
  - Predict/prevent machinery failures
  - Identify anomalies in production systems to optimize the use manufacturing capacity
  - Discover novel patterns to improve product quality

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Mining Applications (cont.)

- Brokerage and Securities Trading
  - Predict changes on certain bond prices
  - Forecast the direction of stock fluctuations
  - Assess the effect of events on market movements
  - Identify and prevent fraudulent activities in trading
- Insurance
  - Forecast claim costs for better business planning
  - Determine optimal rate plans
  - Optimize marketing to specific customers
  - Identify and prevent fraudulent claim activities

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Mining Applications (cont.)

- Computer hardware and software
- Science and engineering
- Government and defense
- Homeland security and law enforcement
- Travel industry
- Healthcare
- Medicine
- Entertainment industry
- Sports
- Etc.

Highly popular application areas for data mining

A Taxonomy for Data Mining Tasks

Data Mining	Learning Method	Popular Algorithms
Prediction	Supervised	Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Classification	Supervised	Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms
Regression	Supervised	Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM
Association	Unsupervised	Apriory, OneR, ZeroR, Eclat
Link analysis	Unsupervised	Expectation Maximization, Apriory Algorithm, Graph-based Matching
Sequence analysis	Unsupervised	Apriory Algorithm, FP-Growth technique
Clustering	Unsupervised	K-means, ANN/SOM
Outlier analysis	Unsupervised	K-means, Expectation Maximization (EM)

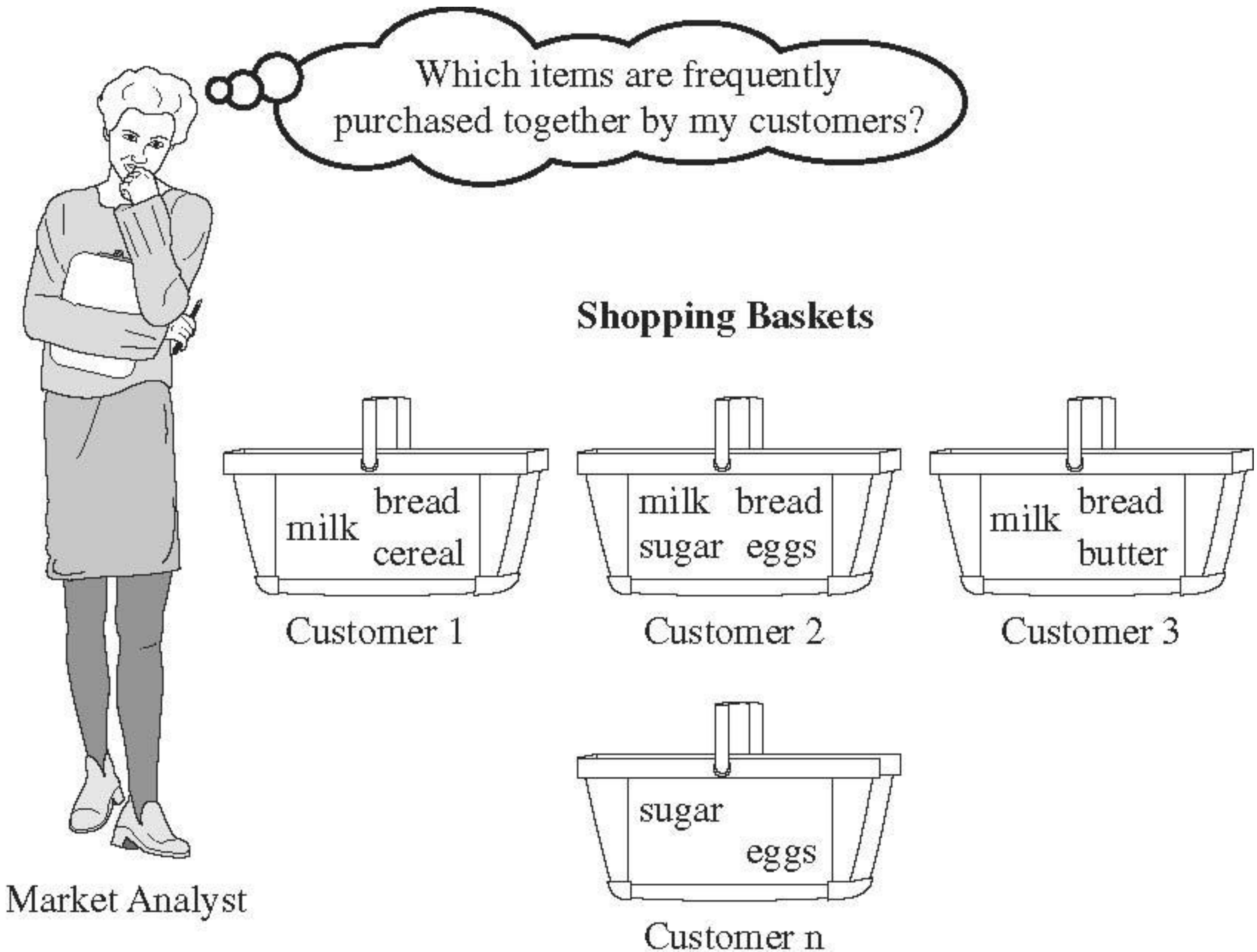
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Association Analysis: Mining Frequent Patterns, Association and Correlations

- Association Analysis
- Mining Frequent Patterns
- Association and Correlations
- Apriori Algorithm

Source: Han & Kamber (2006)

Market Basket Analysis



Source: Han & Kamber (2006)

Association Rule Mining

- Apriori Algorithm

Raw Transaction Data

Transaction No	SKUs (Item No)
1	1, 2, 3, 4
1	2, 3, 4
1	2, 3
1	1, 2, 4
1	1, 2, 3, 4
1	2, 4

One-item Itemsets

Itemset (SKUs)	Support
1	3
2	6
3	4
4	5

Two-item Itemsets

Itemset (SKUs)	Support
1, 2	3
1, 3	2
1, 4	3
2, 3	4
2, 4	5
3, 4	3

Three-item Itemsets

Itemset (SKUs)	Support
1, 2, 4	3
2, 3, 4	3

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Association Rule Mining

- A very popular DM method in business
- Finds interesting relationships (affinities) between variables (items or events)
- Part of machine learning family
- Employs unsupervised learning
- There is no output variable
- Also known as market basket analysis
- Often used as an example to describe DM to ordinary people, such as the famous "relationship between diapers and beers!"

Association Rule Mining

- Input: the simple point-of-sale transaction data
- Output: Most frequent affinities among items
- Example: according to the transaction data…

  "Customer who bought a laptop computer and a virus protection software, also bought extended service plan 70 percent of the time."
- How do you use such a pattern/knowledge?
  - Put the items next to each other for ease of finding
  - Promote the items as a package (do not put one on sale if the other(s) are on sale)
  - Place items far apart from each other so that the customer has to walk the aisles to search for it, and by doing so potentially seeing and buying other items

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Association Rule Mining

- A representative applications of association rule mining include
  - In business: cross-marketing, cross-selling, store design, catalog design, e-commerce site design, optimization of online advertising, product pricing, and sales/promotion configuration
  - In medicine: relationships between symptoms and illnesses; diagnosis and patient characteristics and treatments (to be used in medical DSS); and genes and their functions (to be used in genomics projects)…

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Association Rule Mining

- Are all association rules interesting and useful?

A Generic Rule: **X $\Rightarrow$ Y [S%, C%]**

X, Y: products and/or services

X: Left-hand-side (LHS)

Y: Right-hand-side (RHS)

S: Support: how often **X** and **Y** go together

C: Confidence: how often **Y** go together with the **X**

Example: {Laptop Computer, Antivirus Software} $\Rightarrow$ {Extended Service Plan} [30%, 70%]

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Association Rule Mining

- Algorithms are available for generating association rules
  - Apriori
  - Eclat
  - FP-Growth
  - + Derivatives and hybrids of the three
- The algorithms help identify the frequent item sets, which are, then converted to association rules

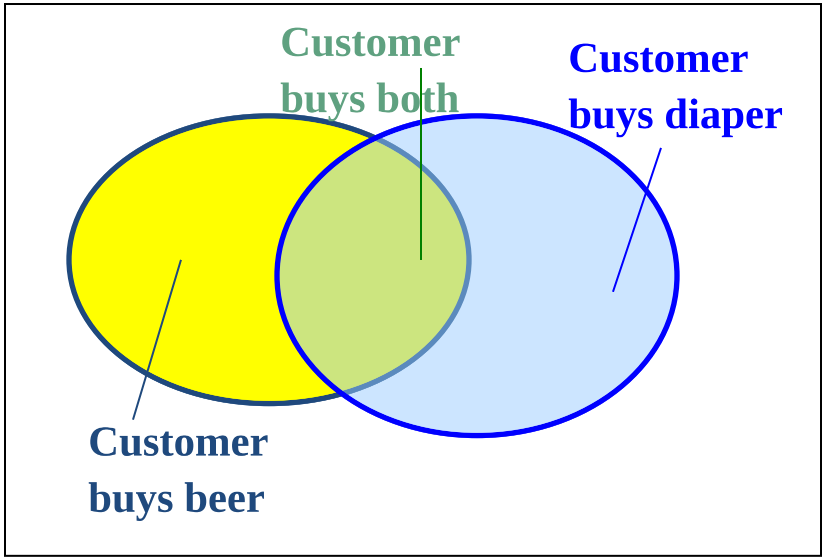
Association Rule Mining

- Apriori Algorithm
  - Finds subsets that are common to at least a minimum number of the itemsets
  - uses a bottom-up approach
    - frequent subsets are extended one item at a time (the size of frequent subsets increases from one-item subsets to two-item subsets, then three-item subsets, and so on), and
    - groups of candidates at each level are tested against the data for minimum

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Basic Concepts: Frequent Patterns and Association Rules

Transaction-id	Items bought
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F



- Itemset X = $\{x_1, \ldots, x_k\}$
- Find all the rules $X \rightarrow Y$ with minimum support and confidence
  - support, *s*, probability that a transaction contains $X \cup Y$
  - confidence, *c,* conditional probability that a transaction having X also contains *Y*

Let sup_{min} = 50%, $conf_{min}$ = 50%
Freq. Pat.: {*A:3, B:3, D:4, E:3, AD:3*}
Association rules:

A → *D* (60%, 100%)
D → *A* (60%, 75%)

A → *D* (support = 3/5 = 60%, confidence = 3/3 =100%)
D → *A* (support = 3/5 = 60%, confidence = 3/4 = 75%)

Source: Han & Kamber (2006)

Market basket analysis

- Example
  - Which groups or sets of items are customers likely to purchase on a given trip to the store?
- Association Rule
  - *Computer → antivirus_software [support = 2%; confidence = 60%]*
    - A support of 2% means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together.
    - A confidence of 60% means that 60% of the customers who purchased a computer also bought the software.

Source: Han & Kamber (2006)

Association rules

- Association rules are considered interesting if they satisfy both
  - a minimum support threshold and
  - a minimum confidence threshold.

Source: Han & Kamber (2006)

Frequent Itemsets, Closed Itemsets, and Association Rules

Let $I = \{I_1, I_2, \ldots, I_m\}$ be a set of items. Let D, the task-relevant data, be a set of database transactions where each transaction T is a set of items such that $T \subseteq I$. Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset I$, $B \subset I$, and $A \cap B = \phi$. The rule $A \Rightarrow B$ holds in the transaction set D with **support** s, where s is the percentage of transactions in D that contain $A \cup B$ (i.e., the *union* of sets A and B, or say, both A and B). This is taken to be the probability, $P(A \cup B)$.[1] The rule $A \Rightarrow B$ has **confidence** c in the transaction set D, where c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability, $P(B|A)$. That is,

$$Support\ (A \rightarrow B) = P(A \cup B)$$

$$Confidence\ (A \rightarrow B) = P(B|A)$$

Source: Han & Kamber (2006)

Support (A → B) = P(A ∪ B)
Confidence (A → B) = P(B|A)

- The notation *P(A ∪ B) indicates the probability that a transaction contains the union of set A and set B*
  - *(i.e., it contains every item in A and in B).*
- *This should not be confused with P(A or B),* which indicates the probability that a transaction contains either *A or B.*

Source: Han & Kamber (2006)

Does diaper purchase predict beer purchase?

- Contingency tables

	Beer: Yes	Beer: No	
No diapers	6	94	100
diapers	40	60	100

DEPENDENT (yes)

	Beer: Yes	Beer: No
No diapers	23	77
diapers	23	77

INDEPENDENT (no predictability)

Source: Dickey (2012) http://www4.stat.ncsu.edu/~dickey/SAScode/Encore_2012.ppt

$$Support\ (A \rightarrow B) = P(A \cup B)$$

$$Confidence\ (A \rightarrow B) = P(B|A)$$

$$Conf\ (A \rightarrow B) = Supp\ (A \cup B) / Supp\ (A)$$

$$Lift\ (A \rightarrow B) = Supp\ (A \cup B) / (Supp\ (A) \times Supp\ (B))$$

Lift (Correlation)

$$Lift\ (A \rightarrow B) = Confidence\ (A \rightarrow B) / Support(B)$$

Lift

Lift = Confidence / Expected Confidence if Independent

Checking → Saving ↓	No (1500)	Yes (8500)	(10000)
No	500	3500	4000
Yes	1000	5000	6000

SVG=>CHKG Expect 8500/10000 = 85% if independent
Observed Confidence is 5000/6000 = 83%
Lift = 83/85 < 1.
Savings account holders actually LESS likely than others to have checking account !!!

- Rules that satisfy both a minimum support threshold (*min_sup*) and a minimum confidence threshold (*min_conf*) are called **strong**.
- By convention, we write support and confidence values so as to occur between 0% and 100%, rather than 0 to 1.0.

Source: Han & Kamber (2006)

- itemset
  - A set of items is referred to as an itemset.
- K-itemset
  - An itemset that contains *k items is a k*-itemset*.*
- Example:
  - The set {*computer, antivirus software*} is a 2-itemset.

Source: Han & Kamber (2006)

Absolute Support and Relative Support

- Absolute Support
  - The occurrence frequency of an itemset is the number of transactions that contain the itemset
    - frequency, support count, or count of the itemset
  - Ex: 3
- Relative support
  - Ex: 60%

- If the relative support of an itemset *I satisfies a prespecified minimum support threshold, then I is a* frequent itemset.
  - *i.e., the absolute* support of *I satisfies the corresponding minimum support count threshold*
- The set of frequent *k-itemsets is commonly denoted by* L_K

Source: Han & Kamber (2006)

$$confidence(A \Rightarrow B) = P(B|A) = \frac{support(A \cup B)}{support(A)} = \frac{support\_count(A \cup B)}{support\_count(A)}$$

- the confidence of rule *A* → *B can be easily derived from the* support counts of *A and A ∪ B.*
- once the support counts of *A, B, and A ∪ B are* found, it is straightforward to derive the corresponding association rules *A* →*B and B* →*A* and check whether they are strong.
- Thus the problem of mining association rules can be reduced to that of mining frequent itemsets.

Source: Han & Kamber (2006)

Association rule mining: Two-step process

1. Find all frequent itemsets
   - By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count, *min_sup.*
2. Generate strong association rules from the frequent itemsets
   - By definition, these rules must satisfy minimum support and minimum confidence.

Source: Han & Kamber (2006)

Efficient and Scalable Frequent Itemset Mining Methods

- The Apriori Algorithm
  - Finding Frequent Itemsets Using Candidate Generation

Source: Han & Kamber (2006)

Apriori Algorithm

- **Apriori** is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules.
- The name of the algorithm is based on the fact that the algorithm uses *prior knowledge of frequent itemset properties, as we shall* see following.

Source: Han & Kamber (2006)

Apriori Algorithm

- Apriori employs an iterative approach known as a *level-wise search, where k-itemsets are used to explore (k+1)-itemsets.*
- *First, the set of frequent 1-itemsets is found* by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted L_1.
- *Next, L_1 is used to find L_2,* the set of frequent 2-itemsets, which is used to find L_3, *and so on, until no more frequent k-itemsets can be found.*
- *The finding of each L_k requires one full scan of the database.*

Source: Han & Kamber (2006)

Apriori Algorithm

- To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property.
- Apriori property
  - *All nonempty subsets of a frequent itemset must also be frequent.*

Source: Han & Kamber (2006)

Apriori algorithm
(1) Frequent Itemsets
(2) Association Rules

Transaction Database

Transaction ID	Items bought
T01	A, B, D
T02	A, C, D
T03	B, C, D, E
T04	A, B, D
T05	A, B, C, E
T06	A, C
T07	B, C, D
T08	B, D
T09	A, C, E
T10	B, D

Table 1 shows a database with 10 transactions.
Let *minimum support = 20%* and *minimum confidence = 80%*.
Please use **Apriori algorithm** for generating **association rules** from frequent itemsets.

Table 1: Transaction Database

Transaction ID	Items bought
T01	A, B, D
T02	A, C, D
T03	B, C, D, E
T04	A, B, D
T05	A, B, C, E
T06	A, C
T07	B, C, D
T08	B, D
T09	A, C, E
T10	B, D

Transaction ID	Items bought
T01	A, B, D
T02	A, C, D
T03	B, C, D, E
T04	A, B, D
T05	A, B, C, E
T06	A, C
T07	B, C, D
T08	B, D
T09	A, C, E
T10	B, D

Apriori Algorithm
$C_1 \rightarrow L_1$

Step 1-1

C_1

Itemset	Support Count
A	6
B	7
C	6
D	7
E	3

minimum support = 20% = 2 / 10
Min. Support Count = 2

→

L_1

Itemset	Support Count
A	6
B	7
C	6
D	7
E	3

Apriori Algorithm
$C_2 \rightarrow L_2$

Step **1-2**

Transaction ID	Items bought
T01	A, B, D
T02	A, C, D
T03	B, C, D, E
T04	A, B, D
T05	A, B, C, E
T06	A, C
T07	B, C, D
T08	B, D
T09	A, C, E
T10	B, D

L_1

Itemset	Support Count
A	6
B	7
C	6
D	7
E	3

C_2

Itemset	Support Count
A, B	3
A, C	4
A, D	3
A, E	2
B, C	3
B, D	6
B, E	2
C, D	3
C, E	3
D, E	1

minimum support = 20% = 2 / 10

Min. Support Count = 2

L_2

Itemset	Support Count
A, B	3
A, C	4
A, D	3
A, E	2
B, C	3
B, D	6
B, E	2
C, D	3
C, E	3

Transaction ID	Items bought
T01	A, B, D
T02	A, C, D
T03	B, C, D, E
T04	A, B, D
T05	A, B, C, E
T06	A, C
T07	B, C, D
T08	B, D
T09	A, C, E
T10	B, D

Apriori Algorithm
$C_3 \rightarrow L_3$

Step **1-3**

L_2

Itemset	Support Count
A, B	3
A, C	4
A, D	3
A, E	2
B, C	3
B, D	6
B, E	2
C, D	3
C, E	3

C_3

Itemset	Support Count
A, B, C	1
A, B, D	2
A, B, E	1
A, C, D	1
A, C, E	2
B, C, D	2
B, C, E	2

minimum support = 20% = 2 / 10
Min. Support Count = 2

L_3

Itemset	Support Count
A, B, D	2
A, C, E	2
B, C, D	2
B, C, E	2

Transaction ID	Items bought
T01	A, B, D
T02	A, C, D
T03	B, C, D, E
T04	A, B, D
T05	A, B, C, E
T06	A, C
T07	B, C, D
T08	B, D
T09	A, C, E
T10	B, D

Generating Association Rules

Step 2-1

minimum confidence = 80%

L_1

Itemset	Support Count
A	6
B	7
C	6
D	7
E	3

L_2

Itemset	Support Count
A, B	3
A, C	4
A, D	3
A, E	2
B, C	3
B, D	6
B, E	2
C, D	3
C, E	3

Association Rules Generated from L_2

A→B: 3/6	B→A: 3/7
A→C: 4/6	C→A: 4/6
A→D: 3/6	D→A: 3/7
A→E: 2/6	E→A: 2/3
B→C: 3/7	C→B: 3/6
B→D: 6/7=85.7% *	D→B: 6/7=85.7% *
B→E: 2/7	E→B: 2/3
C→D: 3/6	D→C: 2/7
C→E: 3/6	E→C: 3/3=100% *

Transaction ID	Items bought
T01	A, B, D
T02	A, C, D
T03	B, C, D, E
T04	A, B, D
T05	A, B, C, E
T06	A, C
T07	B, C, D
T08	B, D
T09	A, C, E
T10	B, D

Generating Association Rules

Step 2-2

minimum confidence = 80%

Association Rules Generated from L_3

L_1

Itemset	Support Count
A	6
B	7
C	6
D	7
E	3

L_2

Itemset	Support Count
A, B	3
A, C	4
A, D	3
A, E	2
B, C	3
B, D	6
B, E	2
C, D	3
C, E	3

L_3

Itemset	Support Count
A, B, D	2
A, C, E	2
B, C, D	2
B, C, E	2

A→BD: 2/6	B→CD: 2/7
B→AD: 2/7	C→BD: 2/6
D→AB: 2/7	D→BC: 2/7
AB→D: 2/3	BC→D: 2/3
AD→B: 2/3	BD→C: 2/6
BD→A: 2/6	CD→B: 2/3
A→CE: 2/6	B→CE: 2/7
C→AE: 2/6	C→BE: 2/6
E→AC: 2/3	E→BC: 2/3
AC→E: 2/4	BC→E: 2/3
AE→C: 2/2=100%*	BE→C: 2/2=100%*
CE→A: 2/3	CE→B: 2/3

Frequent Itemsets and Association Rules

Transaction ID	Items bought
T01	A, B, D
T02	A, C, D
T03	B, C, D, E
T04	A, B, D
T05	A, B, C, E
T06	A, C
T07	B, C, D
T08	B, D
T09	A, C, E
T10	B, D

L_1

Itemset	Support Count
A	6
B	7
C	6
D	7
E	3

L_2

Itemset	Support Count
A, B	3
A, C	4
A, D	3
A, E	2
B, C	3
B, D	6
B, E	2
C, D	3
C, E	3

L_3

Itemset	Support Count
A, B, D	2
A, C, E	2
B, C, D	2
B, C, E	2

minimum support = 20%
minimum confidence = 80%

Association Rules:

B→D (60%, 85.7%) (Sup.: 6/10, Conf.: 6/7)
D→B (60%, 85.7%) (Sup.: 6/10, Conf.: 6/7)
E→C (30%, 100%) (Sup.: 3/10, Conf.: 3/3)
AE→C (20%, 100%) (Sup.: 2/10, Conf.: 2/2)
BE→C (20%, 100%) (Sup.: 2/10, Conf.: 2/2)

Table 1 shows a database with 10 transactions.
Let *minimum support = 20%* and *minimum confidence = 80%*.
Please use **Apriori algorithm** for generating **association rules** from frequent itemsets.

Transaction ID	Items bought
T01	A, B, D
T02	A, C, D
T03	B, C, D, E
T04	A, B, D
T05	A, B, C, E
T06	A, C
T07	B, C, D
T08	B, D
T09	A, C, E
T10	B, D

Association Rules:

B→D (60%, 85.7%) (Sup.: 6/10, Conf.: 6/7)
D→B (60%, 85.7%) (Sup.: 6/10, Conf.: 6/7)
E→C (30%, 100%) (Sup.: 3/10, Conf.: 3/3)
AE→C (20%, 100%) (Sup.: 2/10, Conf.: 2/2)
BE→C (20%, 100%) (Sup.: 2/10, Conf.: 2/2)

A Taxonomy for Data Mining Tasks

Data Mining	Learning Method	Popular Algorithms
Prediction	Supervised	Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Classification	Supervised	Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms
Regression	Supervised	Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM
Association	Unsupervised	Apriory, OneR, ZeroR, Eclat
Link analysis	Unsupervised	Expectation Maximization, Apriory Algorithm, Graph-based Matching
Sequence analysis	Unsupervised	Apriory Algorithm, FP-Growth technique
Clustering	Unsupervised	K-means, ANN/SOM
Outlier analysis	Unsupervised	K-means, Expectation Maximization (EM)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Classification vs. Prediction

- Classification
  - predicts categorical class labels (discrete or nominal)
  - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
- Prediction
  - models continuous-valued functions
    - i.e., predicts unknown or missing values
- Typical applications
  - Credit approval
  - Target marketing
  - Medical diagnosis
  - Fraud detection

Source: Han & Kamber (2006)

Data Mining Methods: Classification

- Most frequently used DM method
- Part of the machine-learning family
- Employ supervised learning
- Learn from past data, classify new data
- The output variable is categorical (nominal or ordinal) in nature
- Classification versus regression?
- Classification versus clustering?

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Classification Techniques

- Decision tree analysis
- Statistical analysis
- Neural networks
- Support vector machines
- Case-based reasoning
- Bayesian classifiers
- Genetic algorithms
- Rough sets

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Example of Classification

- Loan Application Data
  - Which loan applicants are "safe" and which are "risky" for the bank?
  - "Safe" or "risky" for load application data
- Marketing Data
  - Whether a customer with a given profile will buy a new computer?
  - "yes" or "no" for marketing data
- **Classification**
  - Data analysis task
  - A model or **Classifier** is constructed to predict categorical labels
    - Labels: "safe" or "risky"; "yes" or "no"; "treatment A", "treatment B", "treatment C"

Source: Han & Kamber (2006)

What Is Prediction?

- (Numerical) prediction is similar to classification
  - construct a model
  - use model to predict continuous or ordered value for a given input
- Prediction is different from classification
  - Classification refers to predict categorical class label
  - Prediction models continuous-valued functions
- Major method for prediction: regression
  - model the relationship between one or more *independent* or **predictor** variables and a *dependent* or **response** variable
- Regression analysis
  - Linear and multiple regression
  - Non-linear regression
  - Other regression methods: generalized linear model, Poisson regression, log-linear models, regression trees

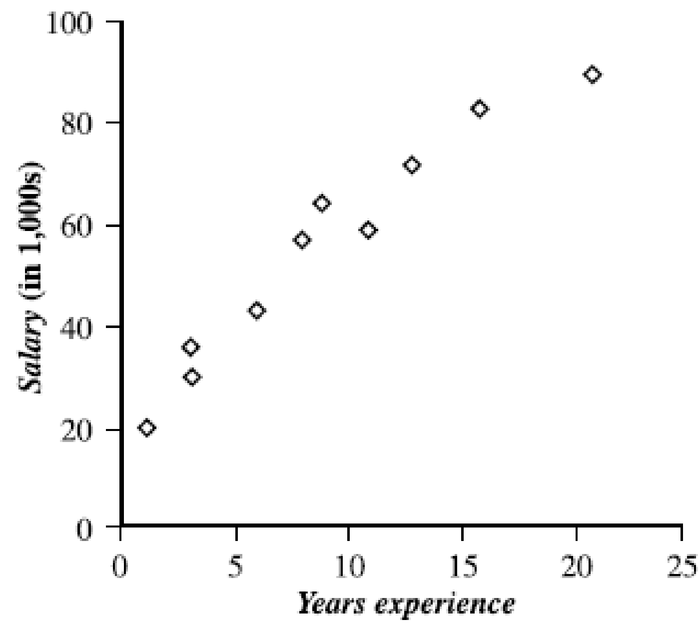
Source: Han & Kamber (2006)

Prediction Methods

- Linear Regression
- Nonlinear Regression
- Other Regression Methods

Salary data.

x *years experience*	y *salary* (*in* $*1000s*)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83



Classification and Prediction

- Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends.
- Classification
  - Effective and scalable methods have been developed for decision trees induction, Naive Bayesian classification, Bayesian belief network, rule-based classifier, Backpropagation, Support Vector Machine (SVM), associative classification, nearest neighbor classifiers, and case-based reasoning, and other classification methods such as genetic algorithms, rough set and fuzzy set approaches.
- Prediction
  - Linear, nonlinear, and generalized linear models of regression can be used for prediction. Many nonlinear problems can be converted to linear problems by performing transformations on the predictor variables. Regression trees and model trees are also used for prediction.

Source: Han & Kamber (2006)

Classification—A Two-Step Process

1. Model construction: describing a set of predetermined classes
   – Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
   – The set of tuples used for model construction is training set
   – The model is represented as classification rules, decision trees, or mathematical formulae
2. Model usage: for classifying future or unknown objects
   – Estimate accuracy of the model
     - The known label of test sample is compared with the classified result from the model
     - Accuracy rate is the percentage of test set samples that are correctly classified by the model
     - Test set is independent of training set, otherwise over-fitting will occur
   – If the accuracy is acceptable, use the model to classify data tuples whose class labels are not known

Source: Han & Kamber (2006)

Supervised vs. Unsupervised Learning

- Supervised learning (classification)
  - Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations
  - New data is classified based on the training set
- Unsupervised learning (clustering)
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Source: Han & Kamber (2006)

Issues Regarding Classification and Prediction: Data Preparation

- Data cleaning
  - Preprocess data in order to reduce noise and handle missing values
- Relevance analysis (feature selection)
  - Remove the irrelevant or redundant attributes
  - Attribute subset selection
    - Feature Selection in machine learning
- Data transformation
  - Generalize and/or normalize data
  - Example
    - Income: low, medium, high

Source: Han & Kamber (2006)

Issues:
Evaluating Classification and Prediction Methods

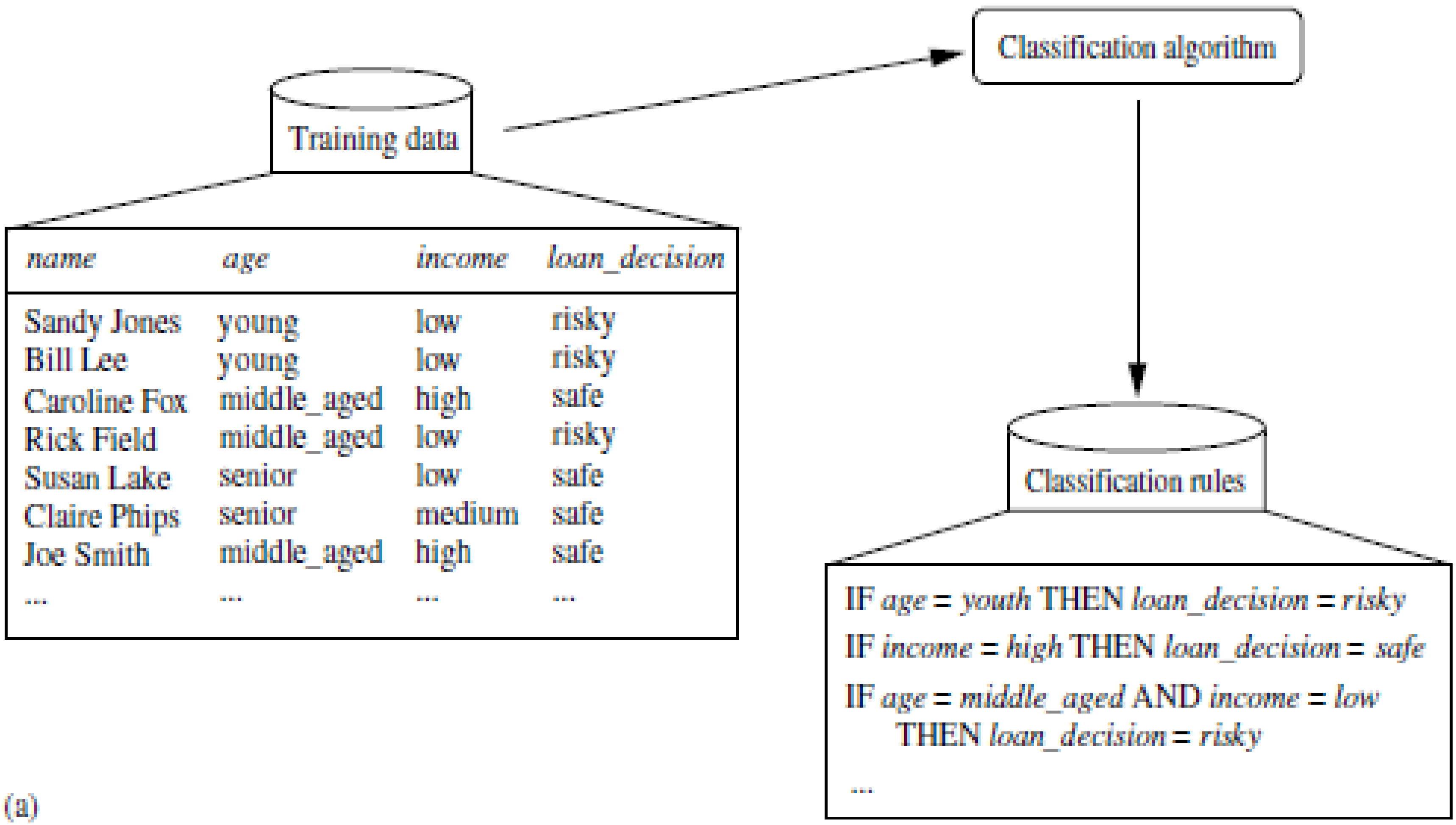
- **Accuracy**
  - classifier accuracy: predicting class label
  - predictor accuracy: guessing value of predicted attributes
  - estimation techniques: cross-validation and bootstrapping
- Speed
  - time to construct the model (training time)
  - time to use the model (classification/prediction time)
- Robustness
  - handling noise and missing values
- Scalability
  - ability to construct the classifier or predictor efficiently given large amounts of data
- Interpretability
  - understanding and insight provided by the model

Source: Han & Kamber (2006)

Data Classification Process 1: Learning (Training) Step

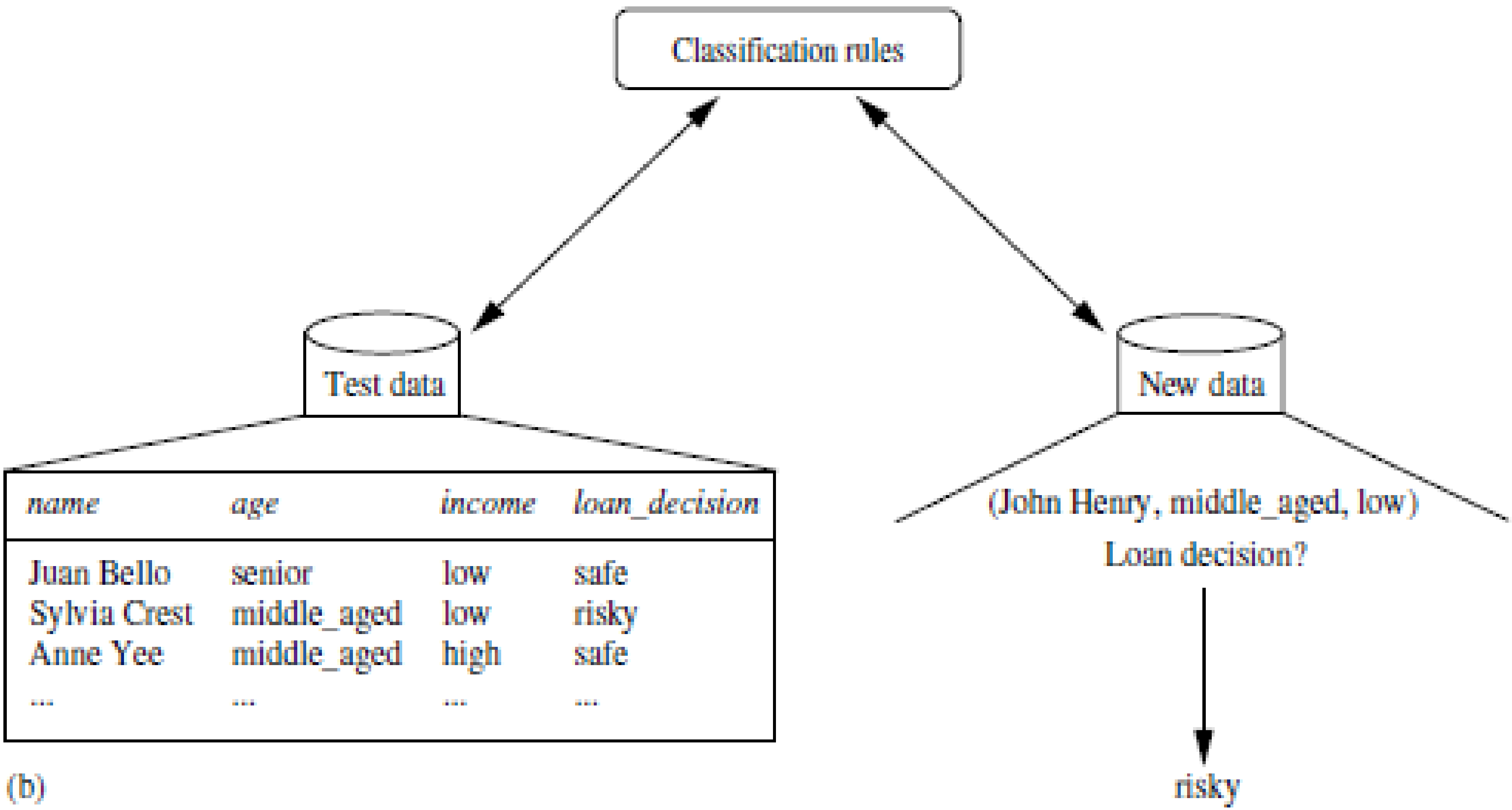
(a) Learning: Training data are analyzed by classification algorithm

$y = f(X)$



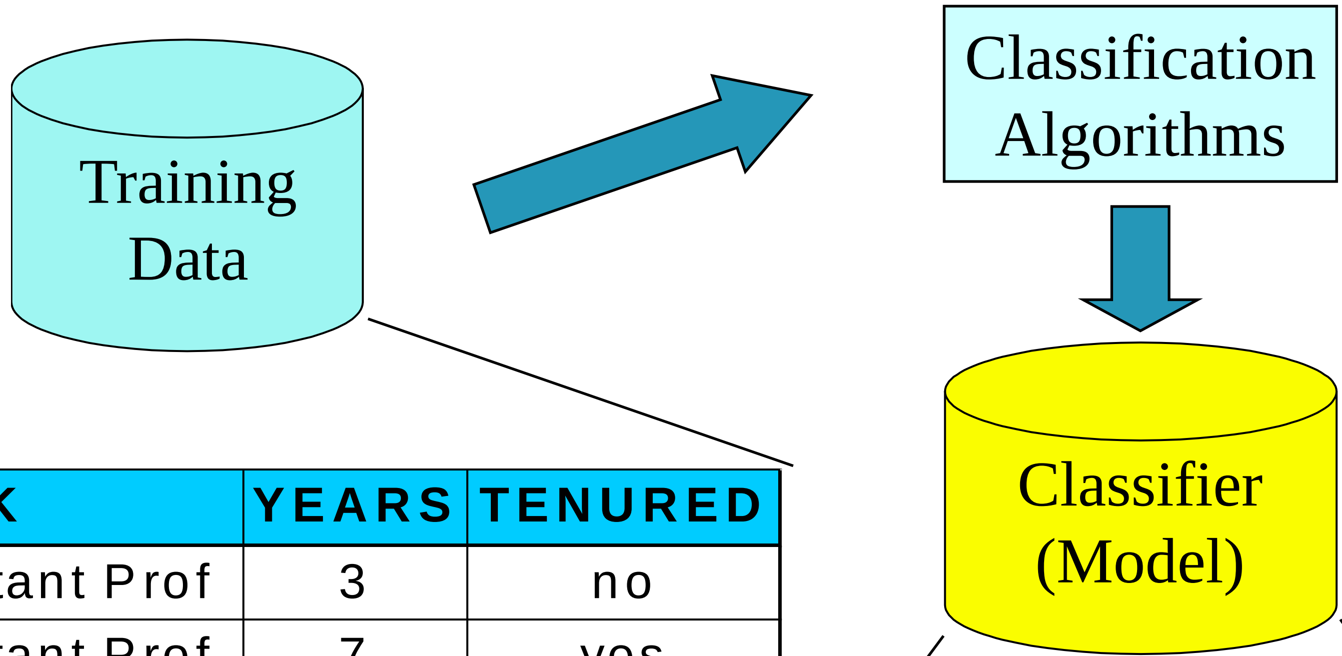
name	*age*	*income*	*loan_decision*
Sandy Jones	young	low	risky
Bill Lee	young	low	risky
Caroline Fox	middle_aged	high	safe
Rick Field	middle_aged	low	risky
Susan Lake	senior	low	safe
Claire Phips	senior	medium	safe
Joe Smith	middle_aged	high	safe
...	...	...	...

Data Classification Process 2
(b) Classification: Test data are used to estimate the accuracy of the classification rules.



Source: Han & Kamber (2006)

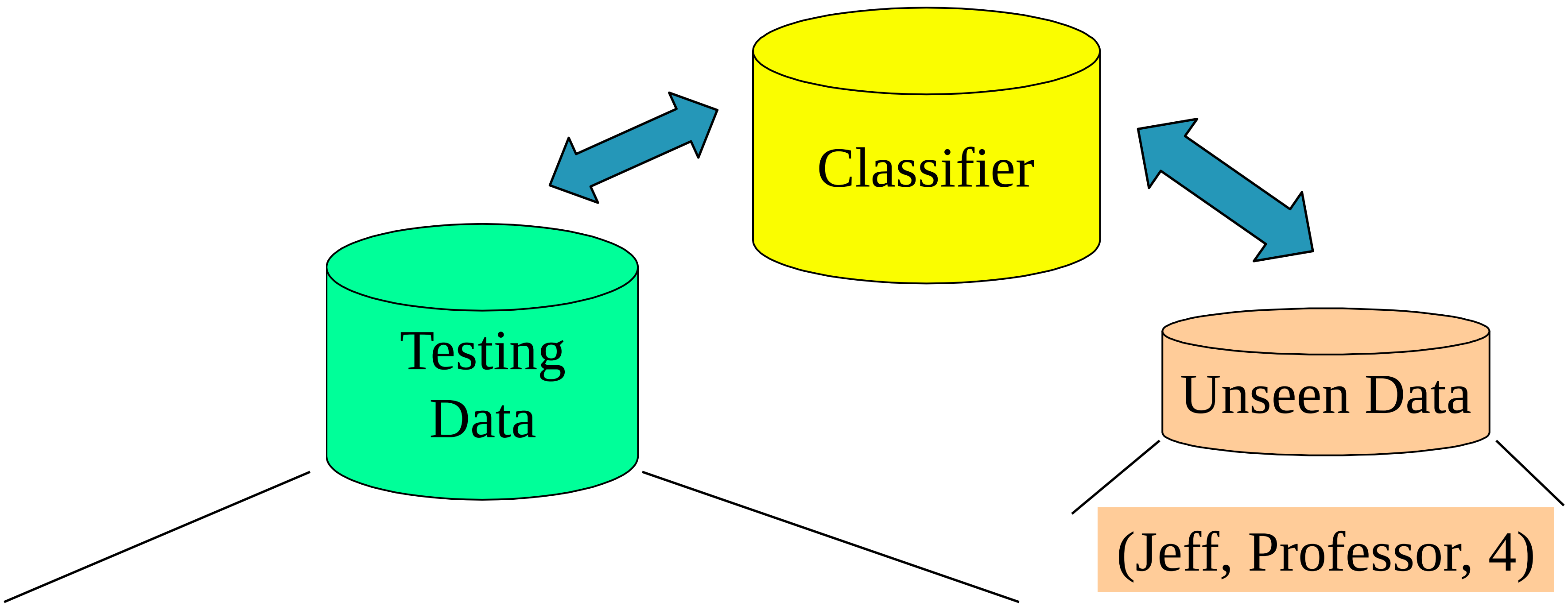
Process (1): Model Construction



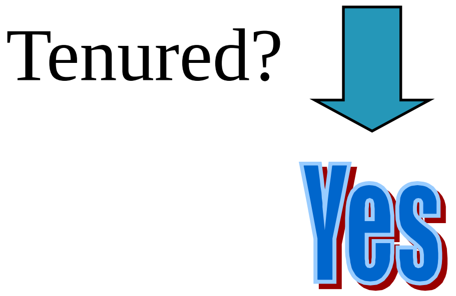
NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’

Process (2): Using the Model in Prediction



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes



Source: Han & Kamber (2006)

Decision Trees

Decision Trees

A general algorithm for decision tree building

- Employs the divide and conquer method
- Recursively divides a training set until each division consists of examples from one class
  1. Create a root node and assign all of the training data to it
  2. Select the best splitting attribute
  3. Add a branch to the root node for each value of the split. Split the data into mutually exclusive subsets along the lines of the specific split
  4. Repeat the steps 2 and 3 for each and every leaf node until the stopping criteria is reached

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Decision Trees

- DT algorithms mainly differ on
  - Splitting criteria
    - Which variable to split first?
    - What values to use to split?
    - How many splits to form for each node?
  - Stopping criteria
    - When to stop building the tree
  - Pruning (generalization method)
    - Pre-pruning versus post-pruning
- Most popular DT algorithms include
  - ID3, C4.5, C5; CART; CHAID; M5

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Decision Trees

- Alternative splitting criteria
  - Gini index determines the purity of a specific class as a result of a decision to branch along a particular attribute/value
    - Used in CART
  - Information gain uses entropy to measure the extent of uncertainty or randomness of a particular attribute/value split
    - Used in ID3, C4.5, C5
  - Chi-square statistics (used in CHAID)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Classification by Decision Tree Induction Training Dataset

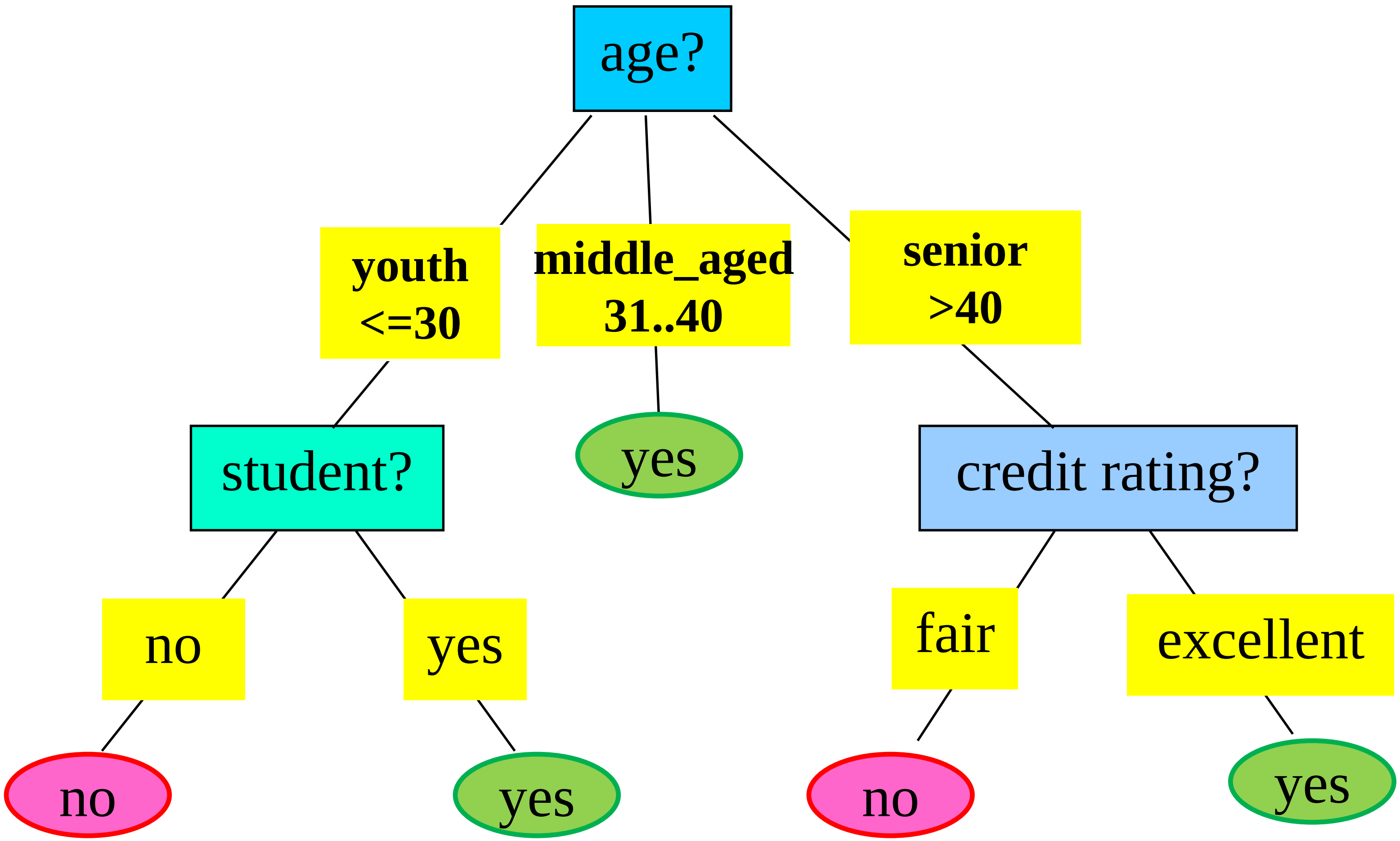
age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

This follows an example of Quinlan's ID3 (Playing Tennis)

Source: Han & Kamber (2006)

Classification by Decision Tree Induction

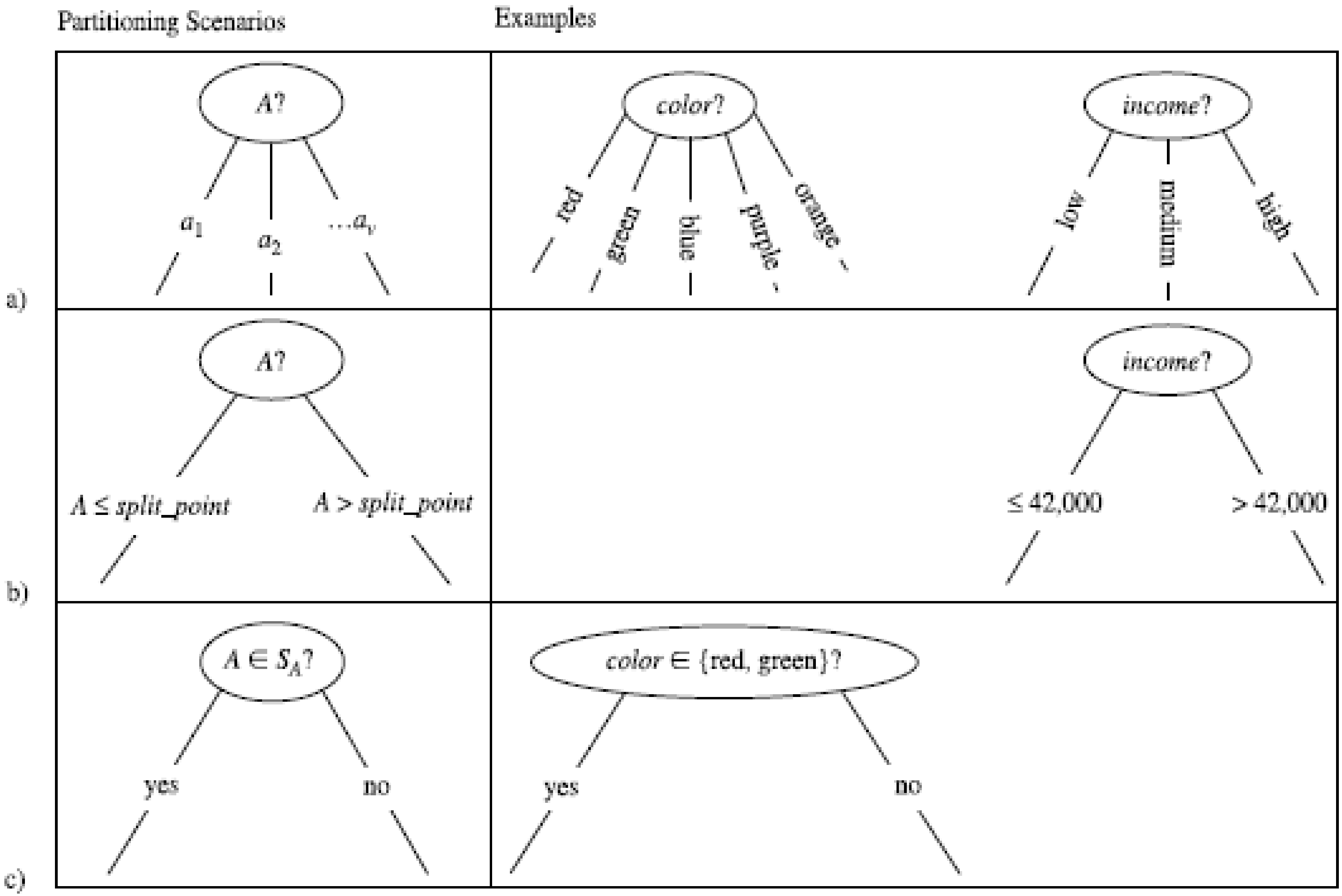
Output: A Decision Tree for "*buys_computer*"



buys_computer="yes" or buys_computer="no"

Source: Han & Kamber (2006)

Three possibilities for partitioning tuples based on the splitting Criterion



Source: Han & Kamber (2006)

Algorithm for Decision Tree Induction

- Basic algorithm (a greedy algorithm)
  - Tree is constructed in a top-down recursive divide-and-conquer manner
  - At start, all the training examples are at the root
  - Attributes are categorical (if continuous-valued, they are discretized in advance)
  - Examples are partitioned recursively based on selected attributes
  - Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)
- Conditions for stopping partitioning
  - All samples for a given node belong to the same class
  - There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
  - There are no samples left

Source: Han & Kamber (2006)

Attribute Selection Measure

- Notation: Let *D, the data partition, be a training set of* class-labeled tuples.
  Suppose the class label attribute has *m distinct values defining m* distinct classes, C_i *(for i = 1, … , m).*
  Let $C_{i,D}$ *be the set of tuples of class* C_i *in D.*
  Let $|D|$ and $|C_{i,D}|$ *denote the number of tuples in D and* $C_{i,D}$, *respectively.*
- *Example:*
  - *Class: buys_computer= "yes" or "no"*
  - *Two distinct classes (m=2)*
    - *Class* C_i *(i=1,2):*
      C_1 = *"yes",*
      C_2 = *"no"*

Attribute Selection Measure: Information Gain (ID3/C4.5)

- Select the attribute with the highest information gain
- Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$
- Expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- Information needed (after using A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

- Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

Source: Han & Kamber (2006)

Decision Tree
Information Gain

Customer database

ID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	middle_aged	high	no	fair	yes
3	youth	high	no	excellent	no
4	senior	medium	no	fair	yes
5	senior	high	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	excellent	yes

Table 2 shows the class-labeled training tuples from customer database. Please calculate and illustrate the final **decision tree** returned by decision tree induction using **information gain**.

(1) What is the Information Gain of “age”?
(2) What is the Information Gain of “income”?
(3) What is the Information Gain of “student”?
(4) What is the Information Gain of “credit_rating”?
(5) What is the class (buys_computer = “yes” or buys_computer = “no”) for a customer (age=youth, income=low, student =yes, credit= fair) based on the classification result by decision three induction?

ID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	middle_aged	high	no	fair	yes
3	youth	high	no	excellent	no
4	senior	medium	no	fair	yes
5	senior	high	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	excellent	yes

Attribute Selection Measure: Information Gain (ID3/C4.5)

- Select the attribute with the highest information gain
- Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$
- Expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- Information needed (after using A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

- Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

Source: Han & Kamber (2006)

ID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	middle_aged	high	no	fair	yes
3	youth	high	no	excellent	no
4	senior	medium	no	fair	yes
5	senior	high	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	excellent	yes

Class P (Positive): buys_computer = “yes”

Class N (Negative): buys_computer = “no”

$P(buys = yes) = P_{i=1} = P_1$ = 6/10 = 0.6

$P(buys = no) = P_{i=2} = P_2$ = 4/10 = 0.4

$\log_2 (0.1) = -3.3219$
$\log_2 (0.2) = -2.3219$
$\log_2 (0.3) = -1.7370$
$\log_2 (0.4) = -1.3219$
$\log_2 (0.5) = -1$
$\log_2 (0.6) = -0.7370$
$\log_2 (0.7) = -0.5146$
$\log_2 (0.8) = -0.3219$
$\log_2 (0.9) = -0.1520$
$\log_2 (1) = 0$

$\log_2 (1) = 0$
$\log_2 (2) = 1$
$\log_2 (3) = 1.5850$
$\log_2 (4) = 2$
$\log_2 (5) = 2.3219$
$\log_2 (6) = 2.5850$
$\log_2 (7) = 2.8074$
$\log_2 (8) = 3$
$\log_2 (9) = 3.1699$
$\log_2 (10) = 3.3219$

Step 1: Expected information

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$Info(D) = I(6,4) = -\frac{6}{10}\log_2(\frac{6}{10}) + (-\frac{4}{10}\log_2(\frac{4}{10}))$$

$$= -0.6 \times \log_2(0.6) - 0.4 \times \log_2(0.4)$$

$$= -0.6 \times (-0.737) - 0.4 \times (-1.3219)$$

$$= 0.4422 + 0.5288$$

$$= 0.971$$

Info(D) = I(6,4) = 0.971

ID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	middle_aged	high	no	fair	yes
3	youth	high	no	excellent	no
4	senior	medium	no	fair	yes
5	senior	high	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	excellent	yes

age	p_i	n_i	***total***
youth	1	3	4
middle_ aged	2	0	2
senior	3	1	4

income	p_i	n_i	***total***
high	2	2	4
midium	2	1	3
low	2	1	3

student	p_i	n_i	***total***
yes	4	1	5
no	2	3	5

credit_ rating	p_i	n_i	***total***
excellent	2	2	4
fair	4	2	6

age	p_i	n_i	*total*	$I(p_i, n_i)$	$I(p_i, n_i)$
youth	1	3	4	*I(1,3)*	*0.8112*
middle_aged	2	0	2	*I(2,0)*	*0*
senior	3	1	4	*I(3,1)*	*0.8112*

Step 2: Information
Step 3: Information Gain

$$I(1,3) = -\frac{1}{4}\log_2(\frac{1}{4}) + (-\frac{3}{4}\log_2(\frac{3}{4}))$$
$$= -0.25\times[\log_2 1 - \log_2 4] + (-0.75\times[\log_2 3 - \log_2 4])$$
$$= -0.25\times[0-2] - 0.75\times[1.585-2]$$
$$= -0.25\times[-2] - 0.75\times[-0.415]$$
$$= 0.5 + 0.3112 = 0.8112$$

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Info(D) = I(6,4) = 0.971

$$Info_A(D) = \sum_{j=1}^{v}\frac{|D_j|}{|D|}\times I(D_j)$$

$$I(2,0) = -\frac{2}{2}\log_2(\frac{2}{2}) + (-\frac{0}{2}\log_2(\frac{0}{2}))$$
$$= -1\times\log_2 1 + (-0\times\log_2 0)$$
$$= -1\times 0 + (-0\times -\infty)$$
$$= 0 + 0 = 0$$

$$Info_{age}(D) = \frac{4}{10}I(1,3) + \frac{2}{10}I(2,0) + \frac{4}{10}I(3,1)$$
$$= \frac{4}{10}\times 0.8112 + \frac{2}{10}\times 0 + \frac{4}{10}\times 0.8112$$
$$= 0.3244 + 0 + 0.3244 = 0.6488$$

$$I(3,1) = -\frac{3}{4}\log_2(\frac{3}{4}) + (-\frac{1}{4}\log_2(\frac{1}{4}))$$
$$= -0.75\times[\log_2 3 - \log_2 4] + (-0.25\times[\log_2 1 - \log_2 4])$$
$$= -0.75\times[1.585-2] - 0.25\times[0-2]$$
$$= -0.75\times[-0.415] - 0.25\times[-2]$$
$$= 0.3112 + 0.5 = 0.8112$$

$$Gain(A) = Info(D) - Info_A(D)$$

$$Gain(age) = Info(D) - Info_{age}(D)$$
$$= 0.971 - 0.6488 = 0.3221$$

(1) Gain(age)= 0.3221

income	p_i	n_i	*total*	$I(p_i, n_i)$	$I(p_i, n_i)$
high	2	2	4	*I(2,2)*	*1*
midium	2	1	3	*I(2,1)*	*0.9182*
low	2	1	3	*I(2,1)*	*0.9182*

$$
\begin{aligned}
I(2,2) &= -\frac{2}{4}\log_2(\frac{2}{4}) + (-\frac{2}{4}\log_2(\frac{2}{4})) \\
&= -0.5\times[\log_2 2 - \log_2 4] + (-0.5\times[\log_2 2 - \log_2 4]) \\
&= -0.5\times[1-2] - 0.5\times[1-2] \\
&= -0.5\times[-1] - 0.5\times[-1] \\
&= 0.5 + 0.5 = 1
\end{aligned}
$$

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Info(D) = I(6,4) = 0.971

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

$$
\begin{aligned}
I(2,1) &= -\frac{2}{3}\log_2(\frac{2}{3}) + (-\frac{1}{3}\log_2(\frac{1}{3})) \\
&= -0.67\times[\log_2 2 - \log_2 3] + (-0.33\times[\log_2 1 - \log_2 3]) \\
&= -0.67\times[1-1.585] - 0.33\times[0-1.585] \\
&= -0.67\times[-0.585] - 0.33\times[-1.585] \\
&= 0.9182
\end{aligned}
$$

$$
\begin{aligned}
Info_{income}(D) &= \frac{4}{10} I(2,2) + \frac{3}{10} I(2,1) + \frac{3}{10} I(2,1) \\
&= \frac{4}{10}\times 1 + \frac{3}{10}\times 0.9182 + \frac{3}{10}\times 0.9182 \\
&= 0.4 + 0.2755 + 0.2755 = 0.951
\end{aligned}
$$

$$Gain(A) = Info(D) - Info_A(D)$$

$$
\begin{aligned}
Gain(income) &= Info(D) - Info_{income}(D) \\
&= 0.971 - 0.951 = 0.02
\end{aligned}
$$

(2) Gain($_{income}$)= 0.02

student	p_i	n_i	total	$I(p_i, n_i)$	$I(p_i, n_i)$
yes	4	1	5	$I(4,1)$	0.7219
no	2	3	5	$I(2,3)$	0.971

$$I(4,1) = -\frac{4}{5}\log_2(\frac{4}{5}) + (-\frac{1}{5}\log_2(\frac{1}{5}))$$
$$= -0.8\times[\log_2 4 - \log_2 5] + (-0.2\times[\log_2 1 - \log_2 5)$$
$$= -0.8\times[2-2.3219] - 0.2\times[0-2.3219]$$
$$= -0.8\times[-0.3219] - 0.2\times[-2.3219]$$
$$= 0.25752 + 0.46438 = 0.7219$$

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Info(D) = I(6,4) = 0.971

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

$$I(2,3) = -\frac{2}{5}\log_2(\frac{2}{5}) + (-\frac{3}{5}\log_2(\frac{3}{5}))$$
$$= -0.4\times[\log_2 0.4] + (-0.6\times[\log_2 0.6)$$
$$= -0.4\times[-1.3219] - 0.6\times[-0.737]$$
$$= 0.5288 + 0.4422 = 0.971$$

$$Info_{student}(D) = \frac{5}{10}I(4,1) + \frac{5}{10}I(2,3)$$
$$= 0.5\times 0.7219 + 0.5\times 0.971$$
$$= 0.36095 + 0.48545 = 0.8464$$

$$Gain(A) = Info(D) - Info_A(D)$$

$$Gain(student) = Info(D) - Info_{student}(D)$$
$$= 0.971 - 0.8464 = 0.1245$$

(3) Gain($_{student}$)= 0.1245

credit	p_i	n_i	total	$I(p_i, n_i)$	$I(p_i, n_i)$
excellent	2	2	4	$I(2,2)$	1
fair	4	2	6	$I(4,2)$	0.9183

$$I(2,2) = -\frac{2}{4}\log_2(\frac{2}{4}) + (-\frac{2}{4}\log_2(\frac{2}{4}))$$
$$= -0.5\times[\log_2 2 - \log_2 4] + (-0.5\times[\log_2 2 - \log_2 4])$$
$$= -0.5\times[1-2] - 0.5\times[1-2]$$
$$= -0.5\times[-1] - 0.5\times[-1]$$
$$= 0.5 + 0.5 = 1$$

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Info(D) = I(6,4) = 0.971

$$Info_A(D) = \sum_{j=1}^{v}\frac{|D_j|}{|D|}\times I(D_j)$$

$$I(4,2) = -\frac{4}{6}\log_2(\frac{4}{6}) + (-\frac{2}{6}\log_2(\frac{2}{6}))$$
$$= -0.67\times[\log_2 2 - \log_2 3] + (-0.33\times[\log_2 1 - \log_2 3])$$
$$= -0.67\times[1-1.585] - 0.33\times[0-1.585]$$
$$= -0.67\times[-0.585] - 0.33\times[-1.585]$$
$$= 0.9182$$

$$Info_{credit}(D) = \frac{4}{10}I(2,2) + \frac{6}{10}I(4,2)$$
$$= \frac{4}{10}\times 1 + \frac{6}{10}\times 0.9182$$
$$= 0.4 + 0.5509 = 0.9509$$

$$Gain(A) = Info(D) - Info_A(D)$$

$$Gain(credit) = Info(D) - Info_{credit}(D)$$
$$= 0.971 - 0.9509 = 0.019$$

(4) Gain($_{\text{credit}}$)= 0.019

age	p_i	n_i	*total*
youth	**1**	**3**	**4**
middle_ aged	2	0	2
senior	3	1	4

student	p_i	n_i	*total*
yes	**4**	**1**	**5**
no	2	3	5

income	p_i	n_i	*total*
high	2	2	4
midium	2	1	3
low	2	1	3

credit_ rating	p_i	n_i	*total*
excellent	2	2	4
fair	4	2	6

(5) What is the class (buys_computer = “yes” or buys_computer = “no”) for a customer (age=youth, income=low, student =yes, credit= fair) based on the classification result by decision three induction?

(5) Yes =0.0889 (No=0.0167)

age (0.3221) > student (0.1245) > income (0.02) > credit (0.019)

buys_computer = “yes”

age:youth (1/4) x student:yes (4/5) x income:low (2/3) x credit:fair (4/6)

Yes: 1/4 x 4/5 x 2/3 x 4/6 = 4/45 = 0.0889

buys_computer = “no”

age:youth (3/4) x student:yes (1/5) x income:low (1/3) x credit:fair (2/6)

No: 3/4 x 1/5 x 1/3 x 2/6 = 0.01667

A Taxonomy for Data Mining Tasks

Data Mining	Learning Method	Popular Algorithms
Prediction	Supervised	Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Classification	Supervised	Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms
Regression	Supervised	Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM
Association	Unsupervised	Apriory, OneR, ZeroR, Eclat
Link analysis	Unsupervised	Expectation Maximization, Apriory Algorithm, Graph-based Matching
Sequence analysis	Unsupervised	Apriory Algorithm, FP-Growth technique
Clustering	Unsupervised	K-means, ANN/SOM
Outlier analysis	Unsupervised	K-means, Expectation Maximization (EM)

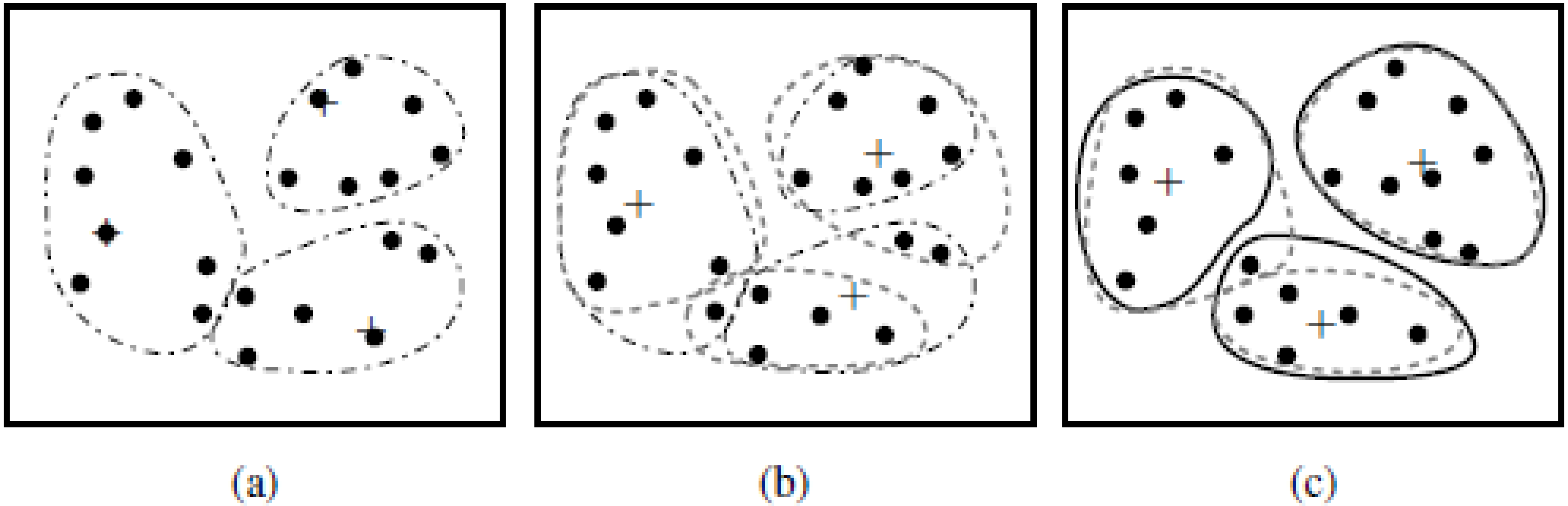
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Cluster Analysis

- Used for automatic identification of natural groupings of things
- Part of the machine-learning family
- Employ unsupervised learning
- Learns the clusters of things from past data, then assigns new instances
- There is not an output variable
- Also known as segmentation

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Cluster Analysis



Clustering of a set of objects based on the *k-means method.*
(The mean of each cluster is marked by a "+".)

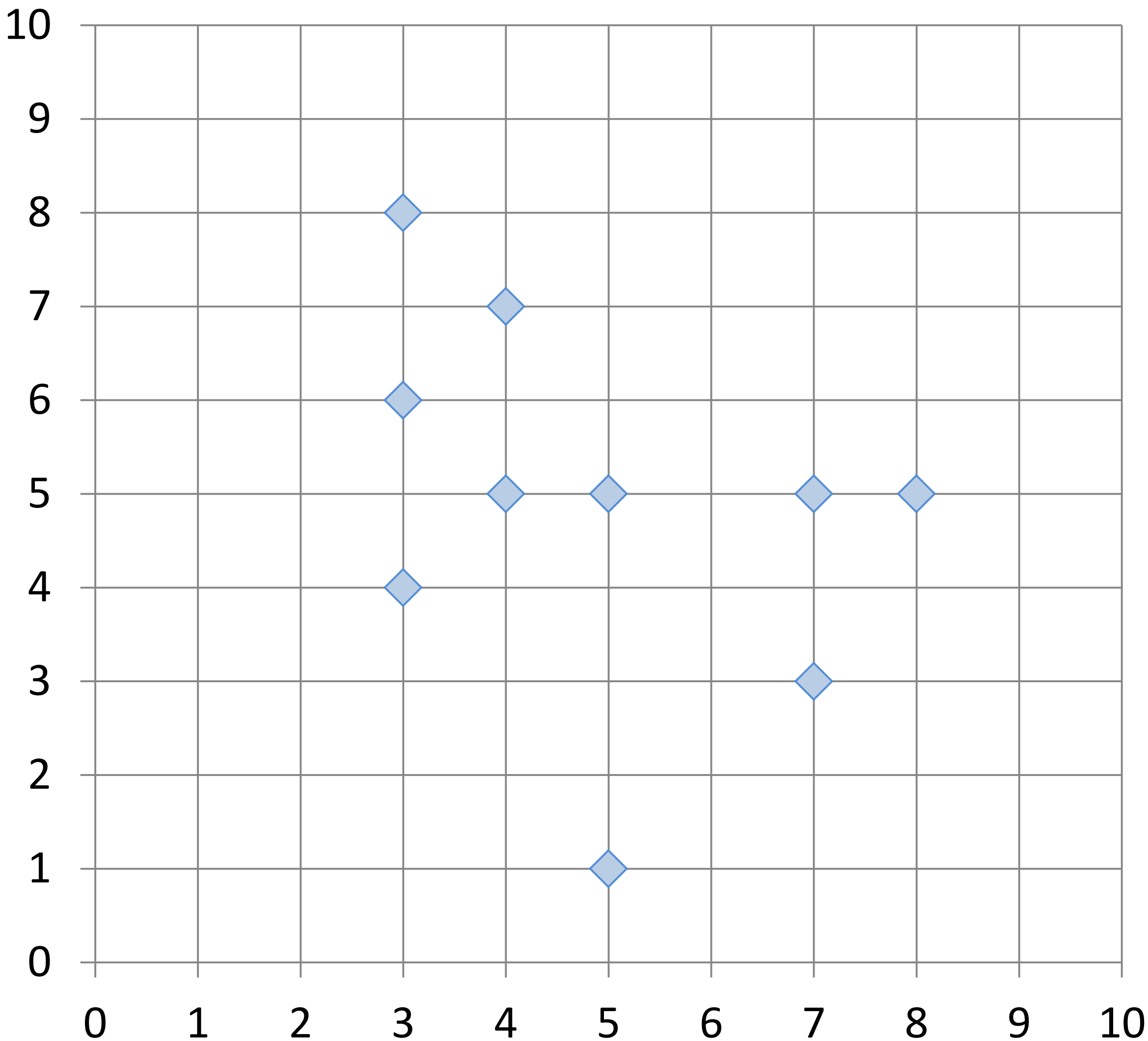
Source: Han & Kamber (2006)

Cluster Analysis

- Clustering results may be used to
  - Identify natural groupings of customers
  - Identify rules for assigning new cases to classes for targeting/diagnostic purposes
  - Provide characterization, definition, labeling of populations
  - Decrease the size and complexity of problems for other data mining methods
  - Identify outliers in a specific domain (e.g., rare-event detection)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Example of Cluster Analysis



Point	P	P(x,y)
p01	a	(3, 4)
p02	b	(3, 6)
p03	c	(3, 8)
p04	d	(4, 5)
p05	e	(4, 7)
p06	f	(5, 1)
p07	g	(5, 5)
p08	h	(7, 3)
p09	i	(7, 5)
p10	j	(8, 5)

Cluster Analysis for Data Mining

- Analysis methods
  - Statistical methods
    (including both hierarchical and nonhierarchical), such as *k*-means, *k*-modes, and so on
  - Neural networks
    (adaptive resonance theory [ART], self-organizing map [SOM])
  - Fuzzy logic (e.g., fuzzy c-means algorithm)
  - Genetic algorithms
- Divisive versus Agglomerative methods

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Cluster Analysis for Data Mining

- How many clusters?
  - There is not a “truly optimal” way to calculate it
  - Heuristics are often used
    1. Look at the sparseness of clusters
    2. Number of clusters = $(n/2)^{1/2}$ (n: no of data points)
    3. Use Akaike information criterion (AIC)
    4. Use Bayesian information criterion (BIC)
- Most cluster analysis methods involve the use of a distance measure to calculate the closeness between pairs of items
  - Euclidian versus Manhattan (rectilinear) distance

k-Means Clustering Algorithm

- *k* : pre-determined number of clusters
- Algorithm (Step 0: determine value of *k*)

Step 1: Randomly generate *k* random points as initial cluster centers

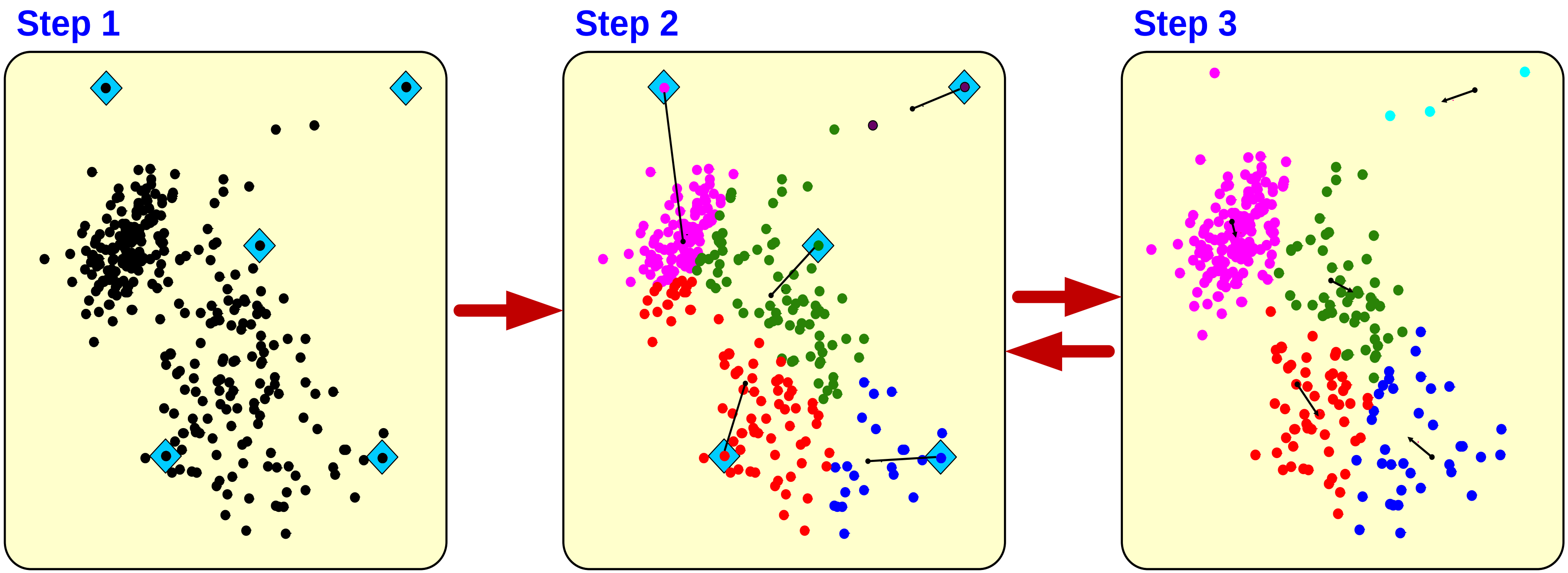
Step 2: Assign each point to the nearest cluster center

Step 3: Re-compute the new cluster centers

Repetition step: Repeat steps 2 and 3 until some convergence criterion is met (usually that the assignment of points to clusters becomes stable)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Cluster Analysis for Data Mining - *k*-Means Clustering Algorithm



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Similarity and Dissimilarity Between Objects

- Distances are normally used to measure the similarity or dissimilarity between two data objects
- Some popular ones include: *Minkowski distance*:

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + \ldots + |x_{i_p} - x_{j_p}|^q)}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and q is a positive integer

- If $q = 1$, d is Manhattan distance

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$$

Source: Han & Kamber (2006)

Similarity and Dissimilarity Between Objects (Cont.)

- *If q = 2, d* is Euclidean distance:

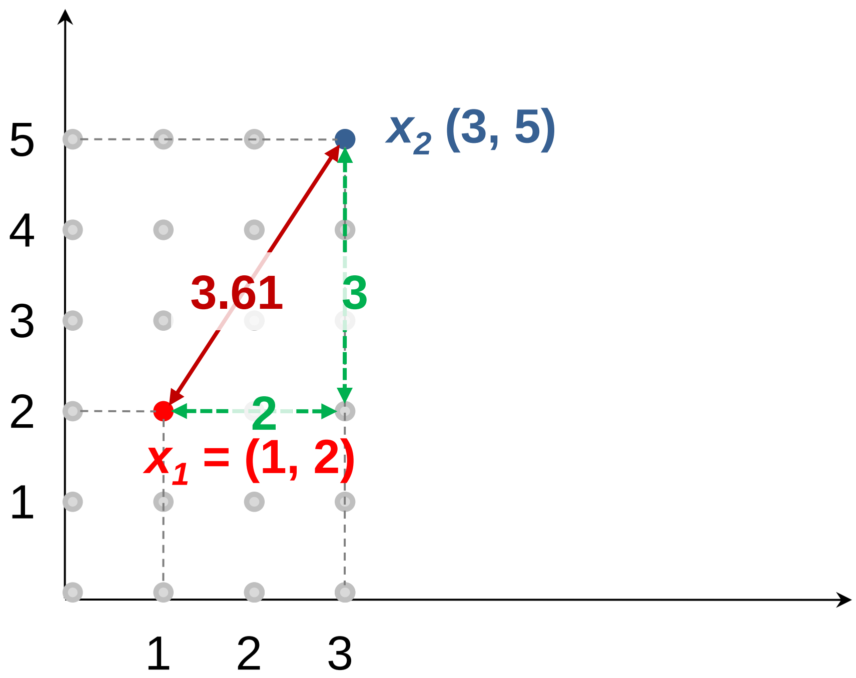
$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + ... + |x_{i_p} - x_{j_p}|^2)}$$

  - Properties
    - $d(i,j) \geq 0$
    - $d(i,i) = 0$
    - $d(i,j) = d(j,i)$
    - $d(i,j) \leq d(i,k) + d(k,j)$
- Also, one can use weighted distance, parametric Pearson product moment correlation, or other disimilarity measures

Source: Han & Kamber (2006)

Euclidean distance vs Manhattan distance

- Distance of two point $x_1 = (1, 2)$ and x_2 $(3, 5)$



Euclidean distance:

$= ((3-1)^2 + (5-2)^2)^{1/2}$

$= (2^2 + 3^2)^{1/2}$

$= (4 + 9)^{1/2}$

$= (13)^{1/2}$

$= 3.61$

Manhattan distance:

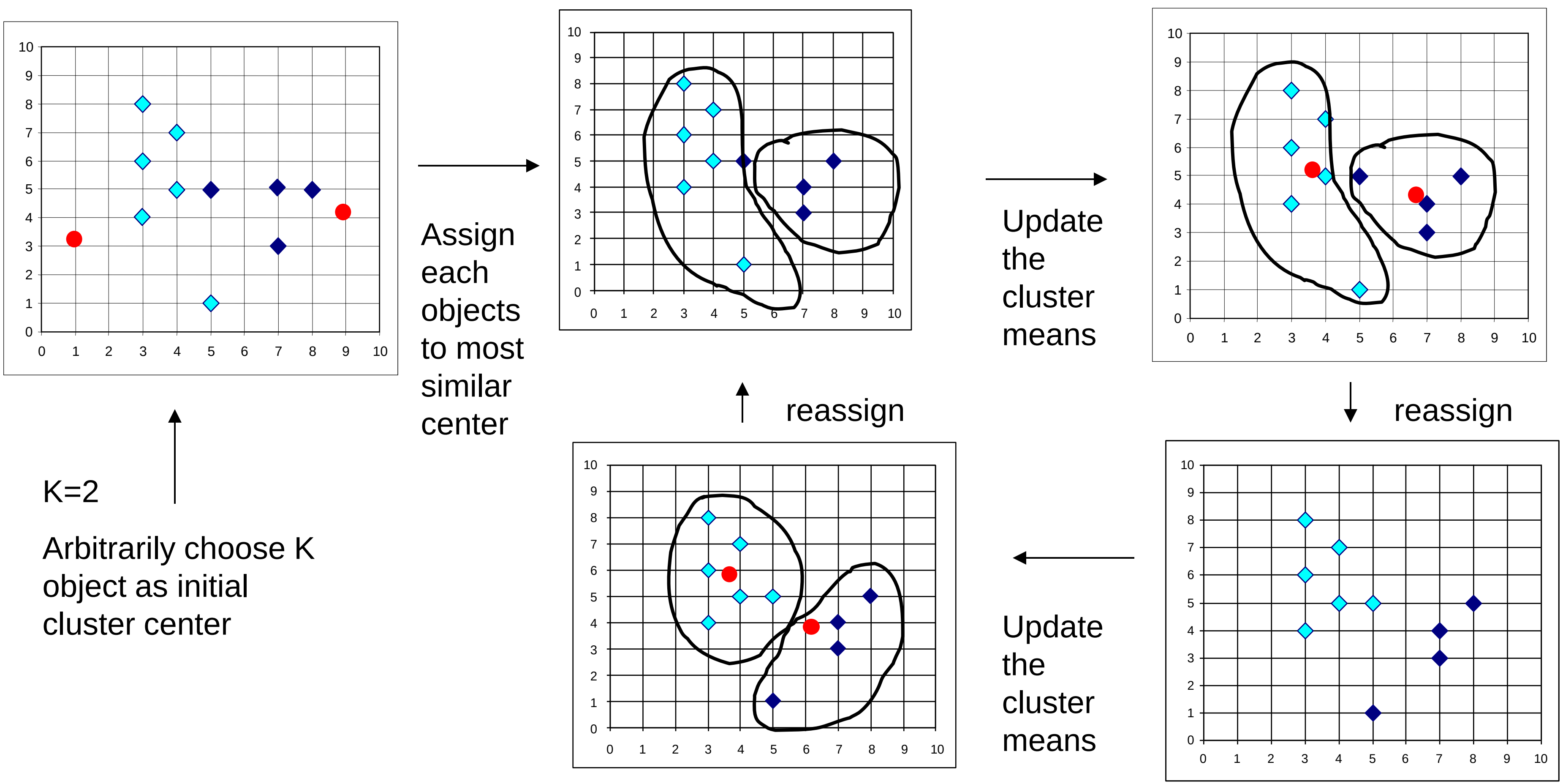
$= (3-1) + (5-2)$

$= 2 + 3$

$= 5$

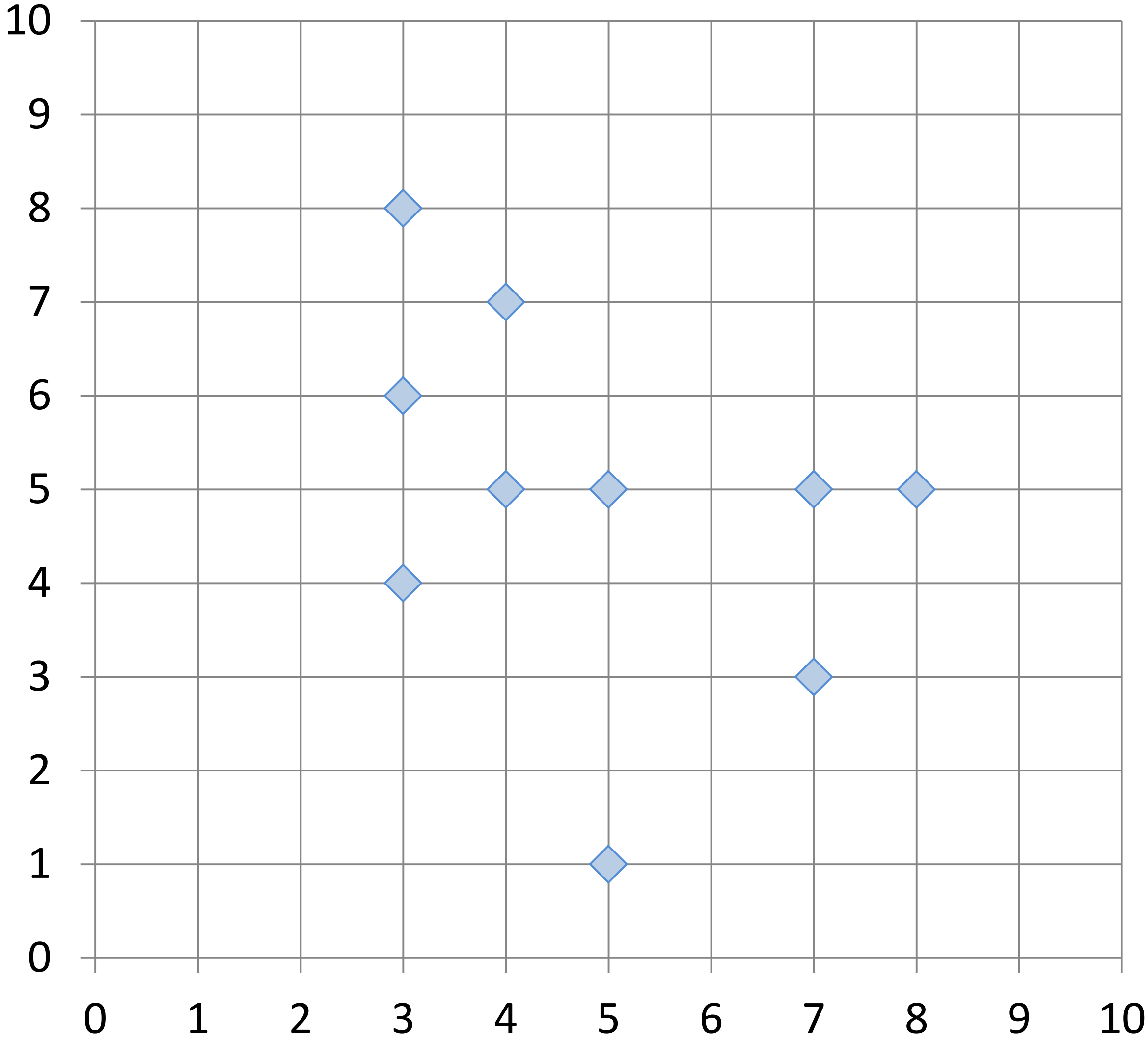
The *K-Means* Clustering Method

- Example



Source: Han & Kamber (2006)

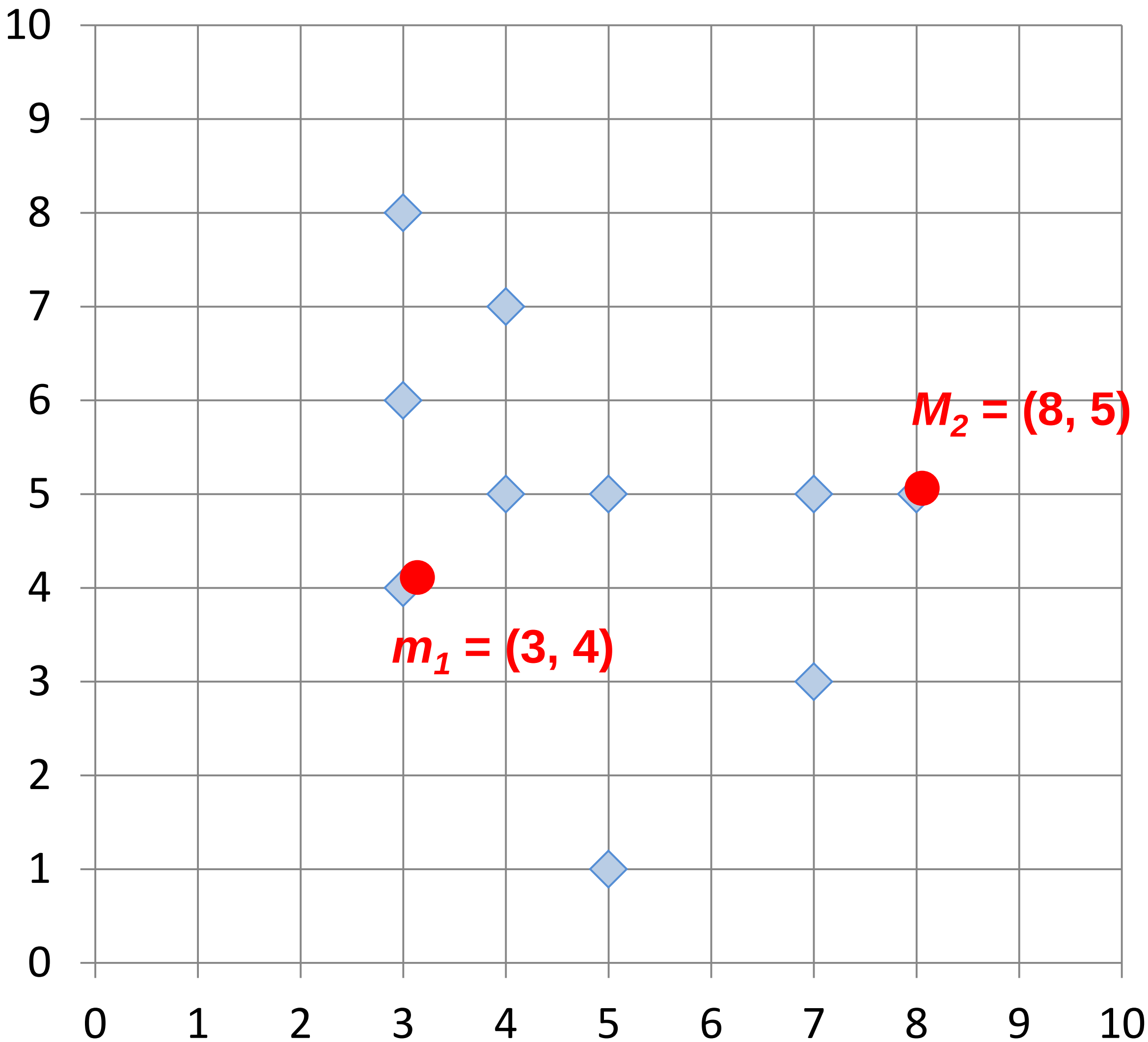
K-Means Clustering Step by Step



Point	P	P(x,y)
p01	a	(3, 4)
p02	b	(3, 6)
p03	c	(3, 8)
p04	d	(4, 5)
p05	e	(4, 7)
p06	f	(5, 1)
p07	g	(5, 5)
p08	h	(7, 3)
p09	i	(7, 5)
p10	j	(8, 5)

K-Means Clustering

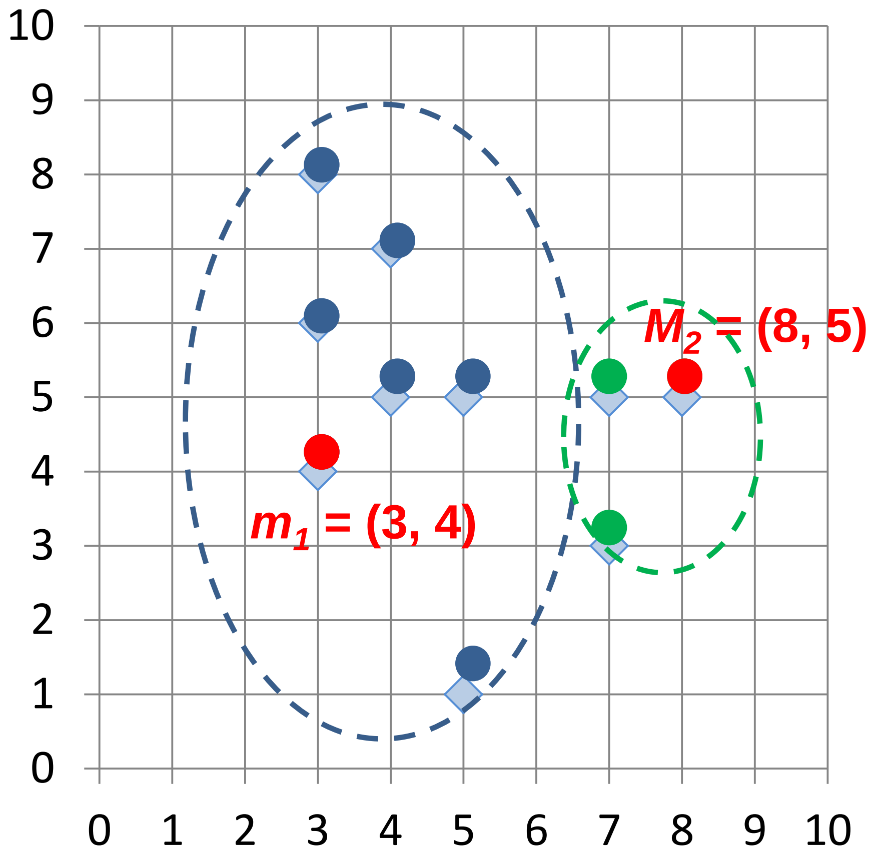
Step 1: K=2, Arbitrarily choose K object as initial cluster center



Point	P	P(x,y)
p01	a	(3, 4)
p02	b	(3, 6)
p03	c	(3, 8)
p04	d	(4, 5)
p05	e	(4, 7)
p06	f	(5, 1)
p07	g	(5, 5)
p08	h	(7, 3)
p09	i	(7, 5)
p10	j	(8, 5)
Initial	m1	(3, 4)
Initial	m2	(8, 5)

Step 2: Compute seed points as the centroids of the clusters of the current partition

Step 3: Assign each objects to most similar center



Point	P	P(x,y)	m1 distance	m2 distance	Cluster
p01	a	(3, 4)	0.00	5.10	Cluster1
p02	b	(3, 6)	2.00	5.10	Cluster1
p03	c	(3, 8)	4.00	5.83	Cluster1
p04	d	(4, 5)	1.41	4.00	Cluster1
p05	e	(4, 7)	3.16	4.47	Cluster1
p06	f	(5, 1)	3.61	5.00	Cluster1
p07	g	(5, 5)	2.24	3.00	Cluster1
p08	h	(7, 3)	4.12	2.24	Cluster2
p09	i	(7, 5)	4.12	1.00	Cluster2
p10	j	(8, 5)	5.10	0.00	Cluster2

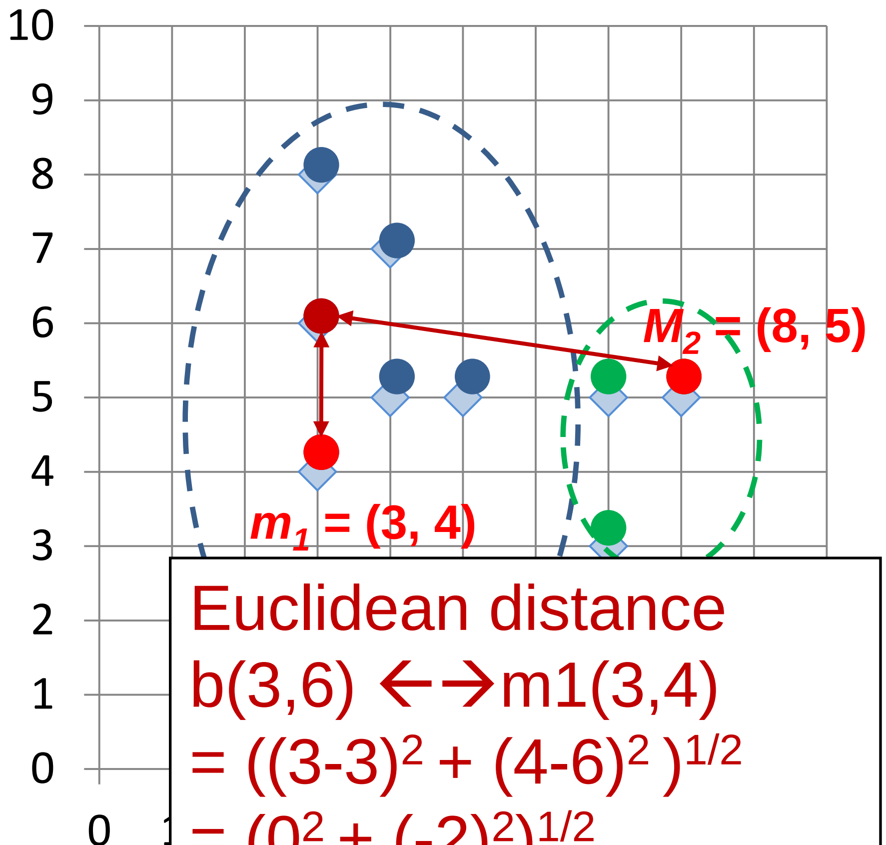
Initial m1 (3, 4)

Initial m2 (8, 5)

K-Means **Clustering**

Step 2: Compute seed points as the centroids of the clusters of the current partition

Step 3: Assign each objects to most similar center



Point	P	P(x,y)	m1 distance	m2 distance	Cluster
p01	a	(3, 4)	0.00	5.10	Cluster1
p02	b	(3, 6)	2.00	5.10	Cluster1
p03	c	(3, 8)	4.00	5.83	Cluster1
p04	d	(4, 5)	1.41	4.00	Cluster1
p05					ster1
p06					ster1
p07					ster1
p08					ster2
p09					ster2
p10					ster2

Euclidean distance

b(3,6) ←→m1(3,4)

$= ((3-3)^2 + (4-6)^2)^{1/2}$

$= (0^2 + (-2)^2)^{1/2}$

$= (0 + 4)^{1/2}$

$= (4)^{1/2}$

$= 2.00$

Euclidean distance

b(3,6) ←→m2(8,5)

$= ((8-3)^2 + (5-6)^2)^{1/2}$

$= (5^2 + (-1)^2)^{1/2}$

$= (25 + 1)^{1/2}$

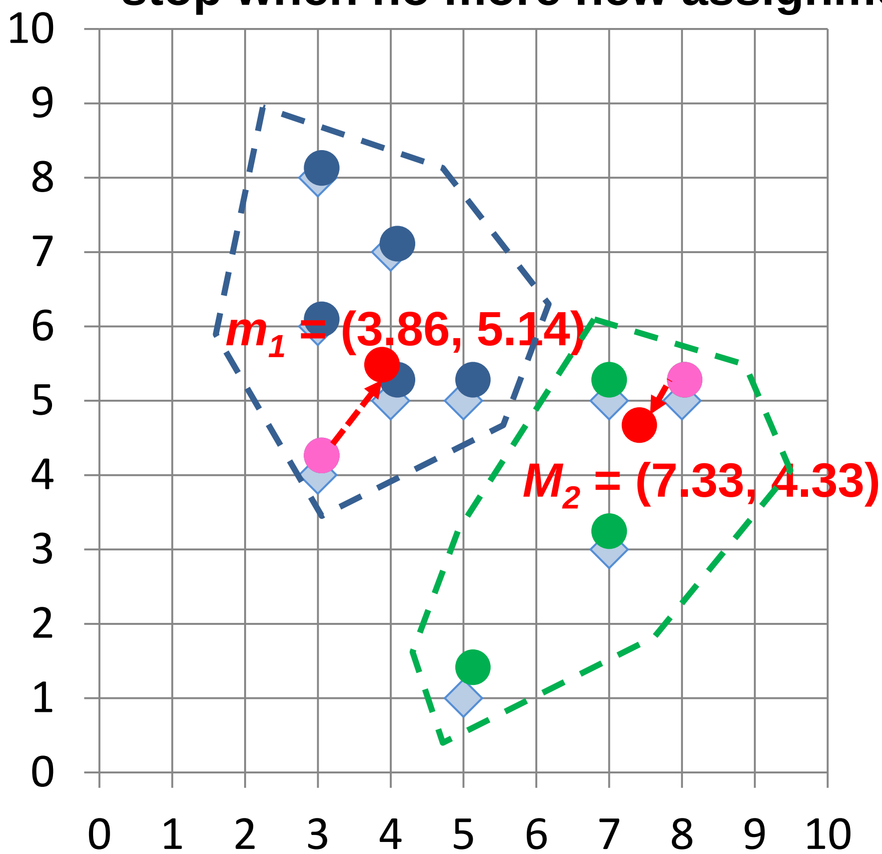
$= (26)^{1/2}$

$= 5.10$

Initial m1 (3, 4)

Initial m2 (8, 5)

Step 4: Update the cluster means,
Repeat Step 2, 3,
stop when no more new assignment



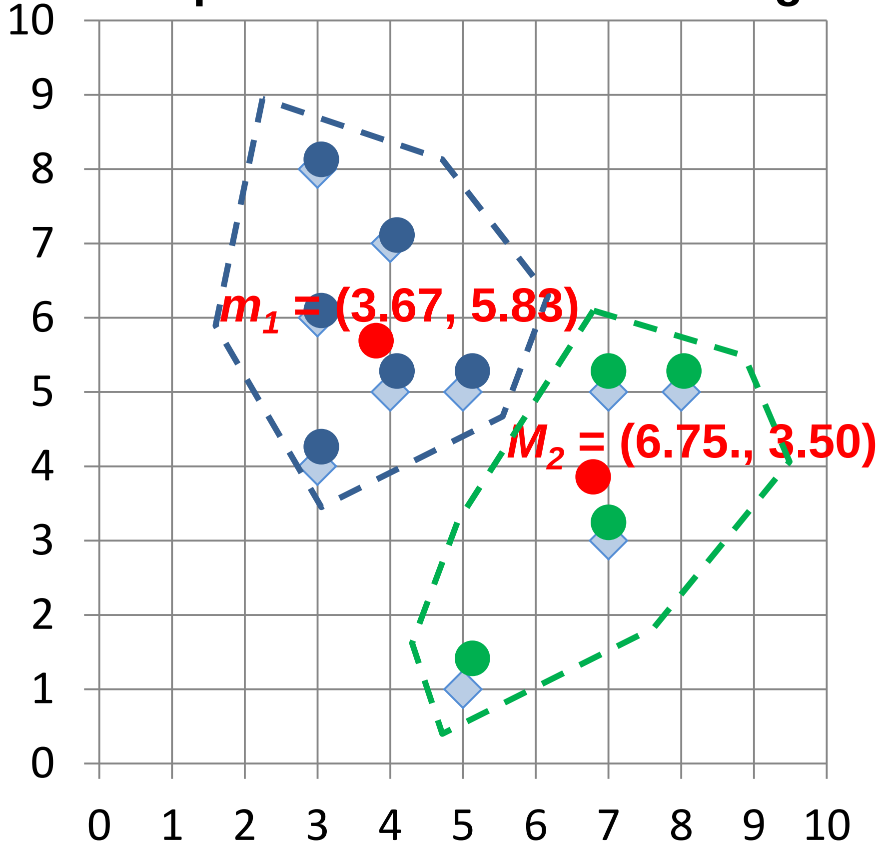
Point	P	P(x,y)	m1 distance	m2 distance	Cluster
p01	a	(3, 4)	1.43	4.34	Cluster1
p02	b	(3, 6)	1.22	4.64	Cluster1
p03	c	(3, 8)	2.99	5.68	Cluster1
p04	d	(4, 5)	0.20	3.40	Cluster1
p05	e	(4, 7)	1.87	4.27	Cluster1
p06	f	(5, 1)	4.29	4.06	Cluster2
p07	g	(5, 5)	1.15	2.42	Cluster1
p08	h	(7, 3)	3.80	1.37	Cluster2
p09	i	(7, 5)	3.14	0.75	Cluster2
p10	j	(8, 5)	4.14	0.95	Cluster2

m1 (3.86, 5.14)

m2 (7.33, 4.33)

K-Means Clustering

Step 4: Update the cluster means,
Repeat Step 2, 3,
stop when no more new assignment



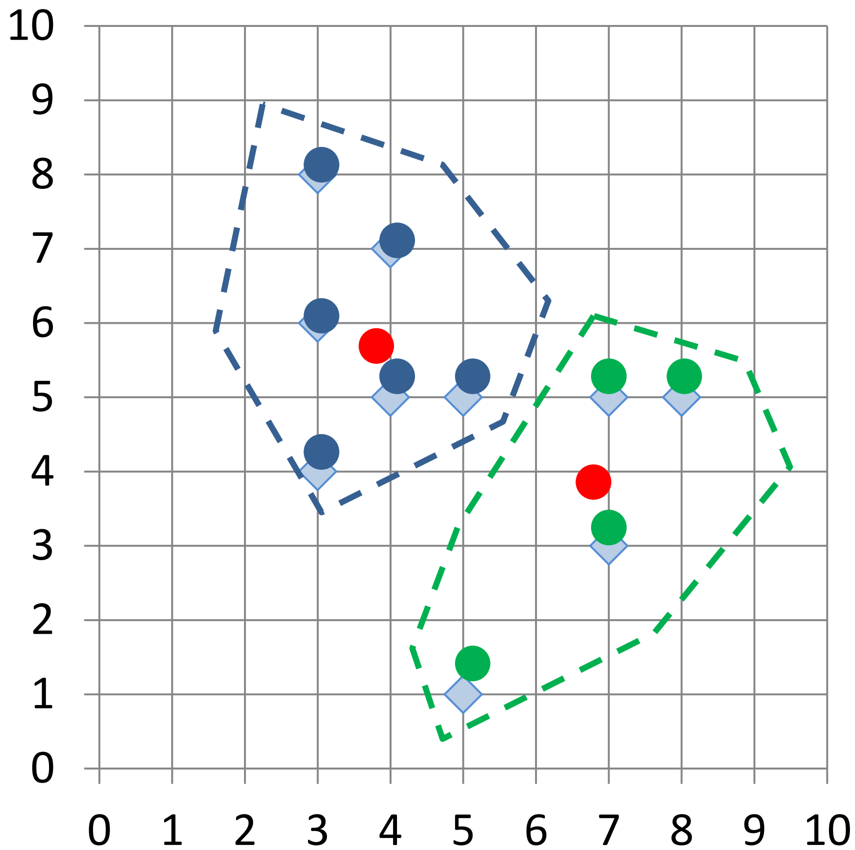
Point	P	P(x,y)	m1 distance	m2 distance	Cluster
p01	a	(3, 4)	1.95	3.78	Cluster1
p02	b	(3, 6)	0.69	4.51	Cluster1
p03	c	(3, 8)	2.27	5.86	Cluster1
p04	d	(4, 5)	0.89	3.13	Cluster1
p05	e	(4, 7)	1.22	4.45	Cluster1
p06	f	(5, 1)	5.01	3.05	Cluster2
p07	g	(5, 5)	1.57	2.30	Cluster1
p08	h	(7, 3)	4.37	0.56	Cluster2
p09	i	(7, 5)	3.43	1.52	Cluster2
p10	j	(8, 5)	4.41	1.95	Cluster2

m1 (3.67, 5.83)

m2 (6.75, 3.50)

K-Means Clustering

stop when no more new assignment



Point	P	P(x,y)	m1 distance	m2 distance	Cluster
p01	a	(3, 4)	1.95	3.78	Cluster1
p02	b	(3, 6)	0.69	4.51	Cluster1
p03	c	(3, 8)	2.27	5.86	Cluster1
p04	d	(4, 5)	0.89	3.13	Cluster1
p05	e	(4, 7)	1.22	4.45	Cluster1
p06	f	(5, 1)	5.01	3.05	Cluster2
p07	g	(5, 5)	1.57	2.30	Cluster1
p08	h	(7, 3)	4.37	0.56	Cluster2
p09	i	(7, 5)	3.43	1.52	Cluster2
p10	j	(8, 5)	4.41	1.95	Cluster2

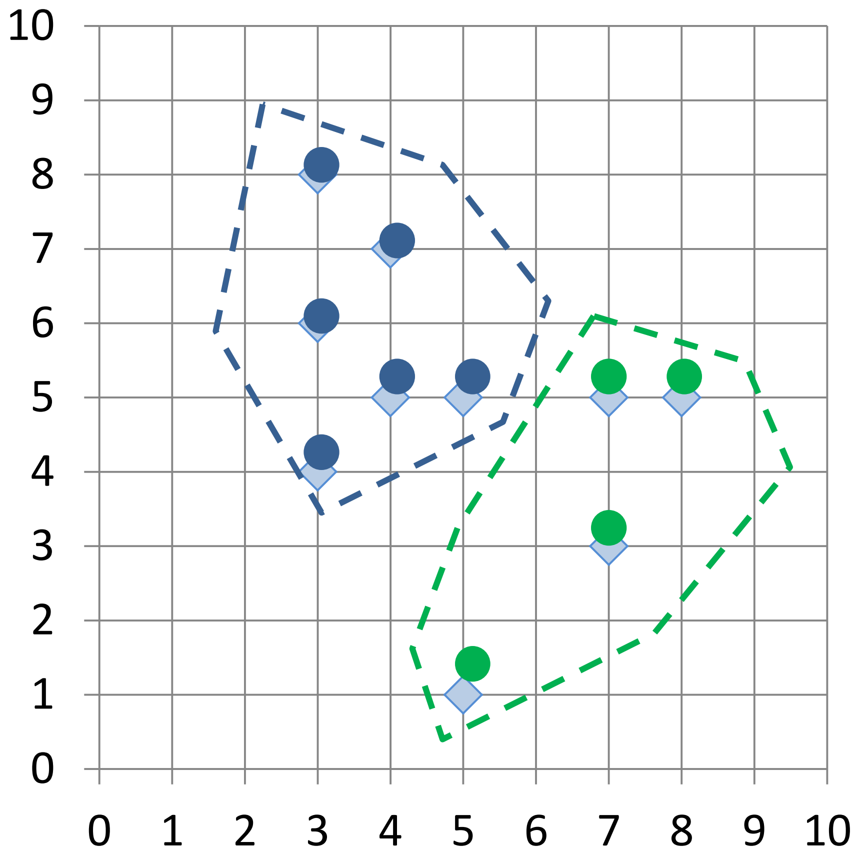
m1 (3.67, 5.83)

m2 (6.75, 3.50)

K-Means Clustering

K-Means Clustering (*K*=2, two clusters)

stop when no more new assignment



Point	P	P(x,y)	m1 distance	m2 distance	Cluster
p01	a	(3, 4)	1.95	3.78	Cluster1
p02	b	(3, 6)	0.69	4.51	Cluster1
p03	c	(3, 8)	2.27	5.86	Cluster1
p04	d	(4, 5)	0.89	3.13	Cluster1
p05	e	(4, 7)	1.22	4.45	Cluster1
p06	f	(5, 1)	5.01	3.05	Cluster2
p07	g	(5, 5)	1.57	2.30	Cluster1
p08	h	(7, 3)	4.37	0.56	Cluster2
p09	i	(7, 5)	3.43	1.52	Cluster2
p10	j	(8, 5)	4.41	1.95	Cluster2

m1 (3.67, 5.83)

m2 (6.75, 3.50)

K-Means Clustering

Data Mining Evaluation

Evaluation

(Accuracy of Classification Model)

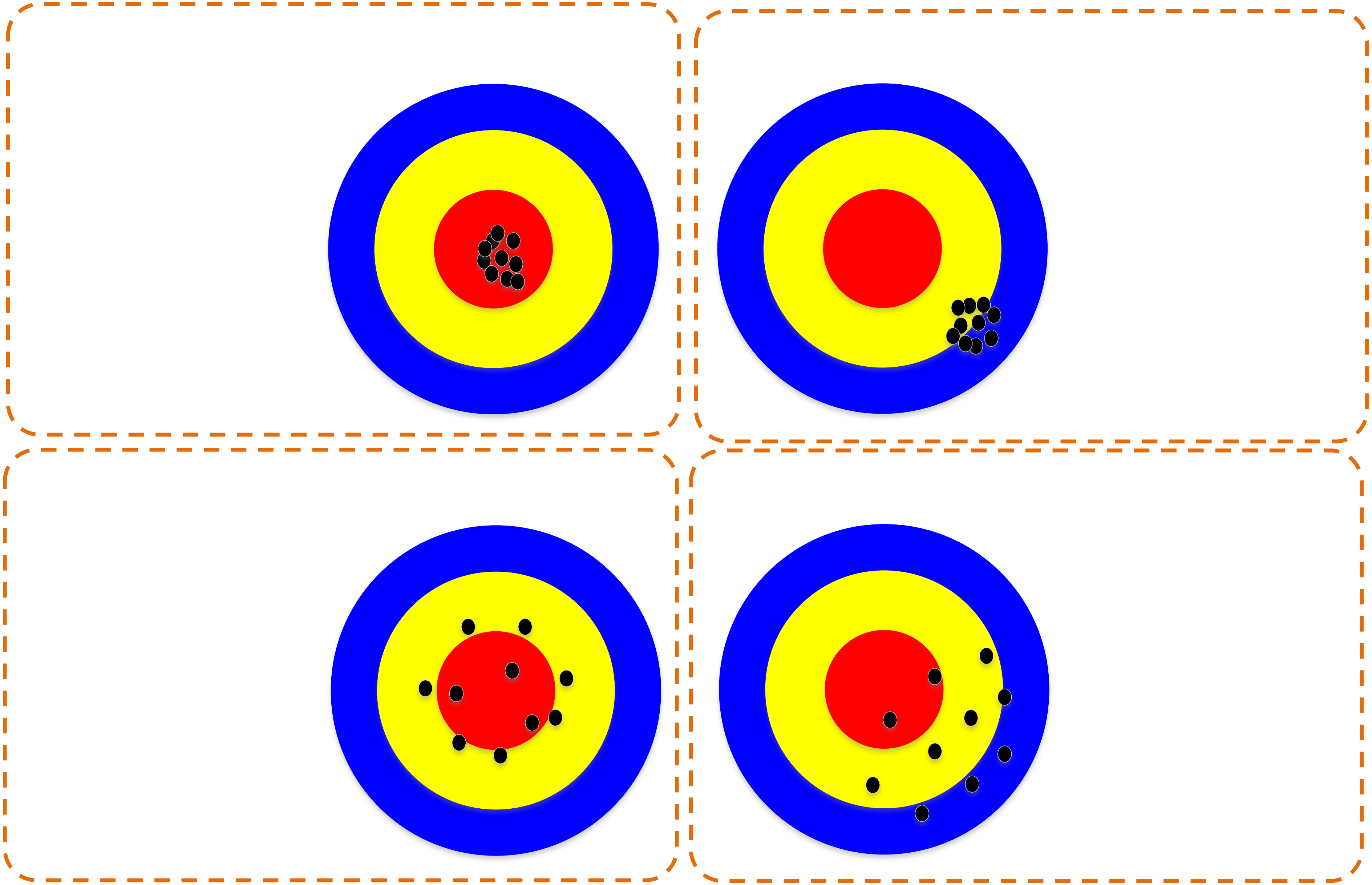
Assessment Methods for Classification

- Predictive accuracy
  - Hit rate
- Speed
  - Model building; predicting
- Robustness
- Scalability
- Interpretability
  - Transparency, explainability

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Accuracy **Validity**

Precision **Reliability**

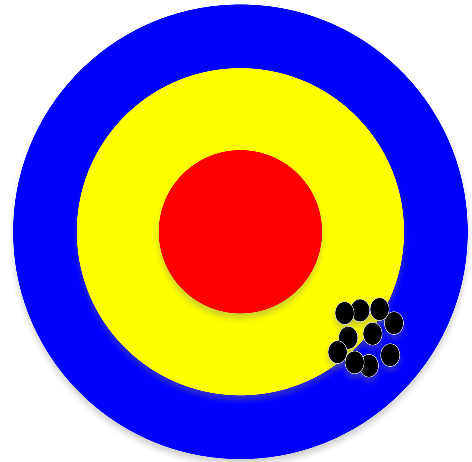


Accuracy vs. Precision

A

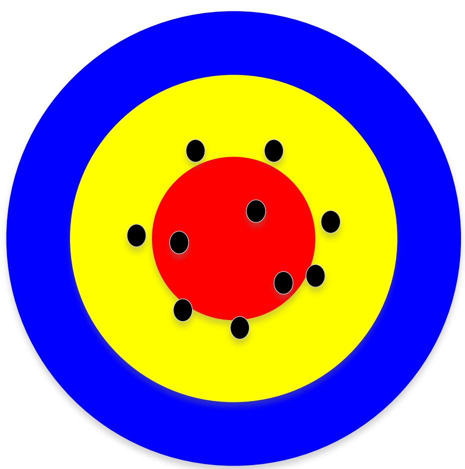
High Accuracy
High Precision

B



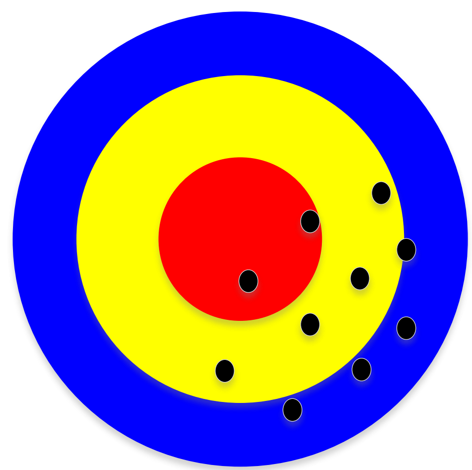
Low Accuracy
High Precision

C



High Accuracy
Low Precision

D



Low Accuracy
Low Precision

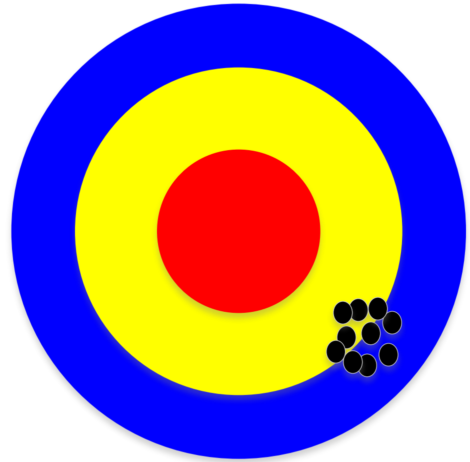
Accuracy vs. Precision

A

High Accuracy
High Precision

High Validity
High Reliability

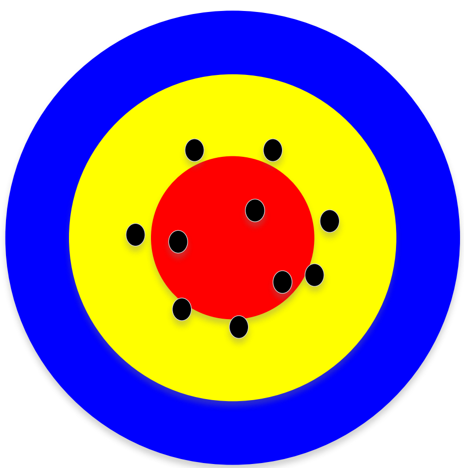
B



Low Accuracy
High Precision

Low Validity
High Reliability

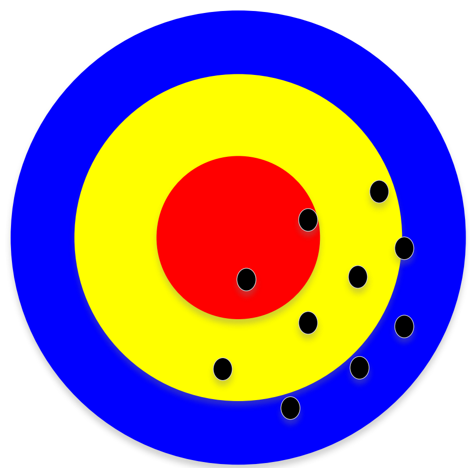
C



High Accuracy
Low Precision

High Validity
Low Reliability

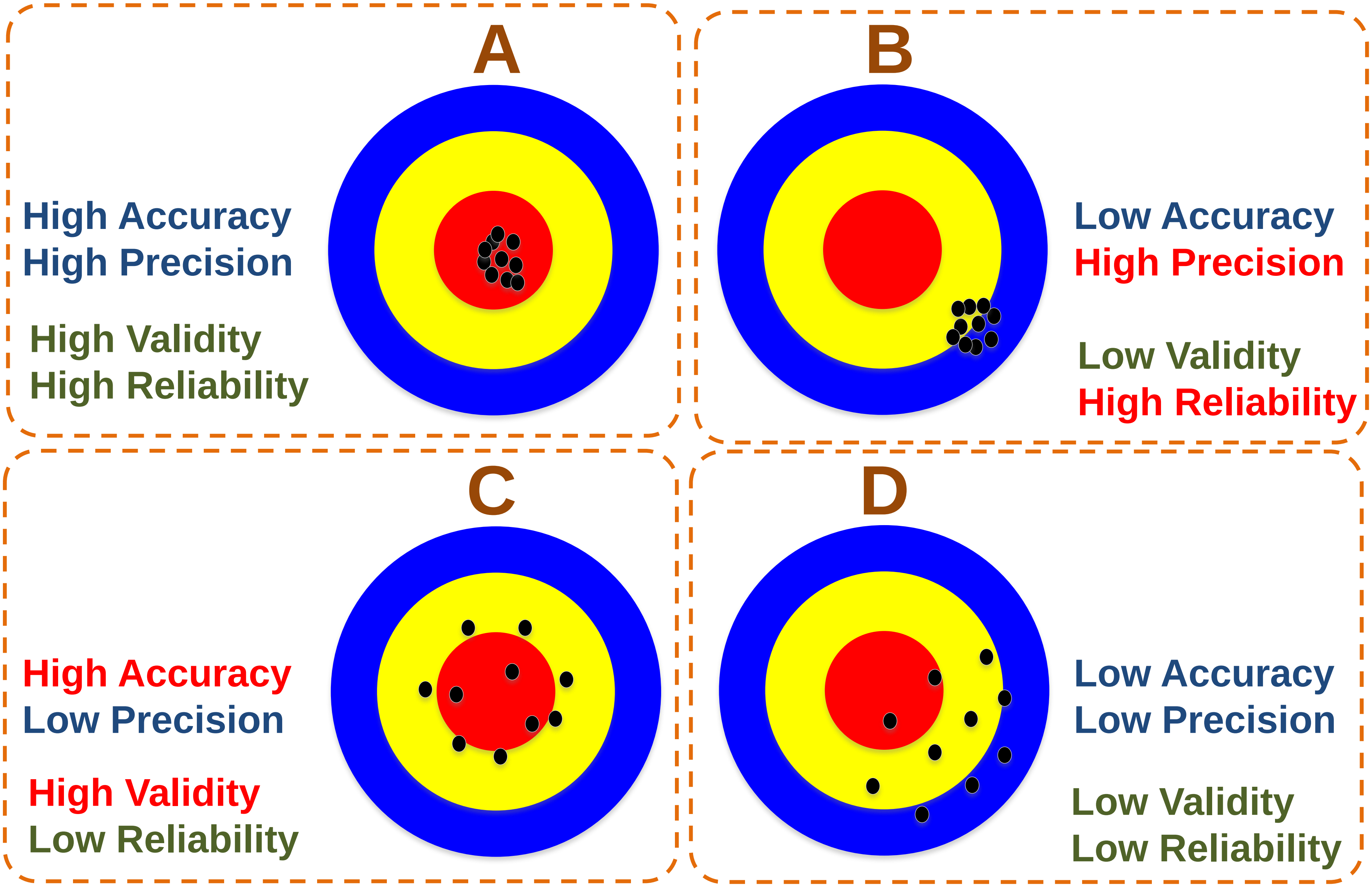
D



Low Accuracy
Low Precision

Low Validity
Low Reliability

Accuracy vs. Precision



Accuracy of Classification Models

- In classification problems, the primary source for accuracy estimation is the confusion matrix

		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive Count (TP)	False Positive Count (FP)
	Negative	False Negative Count (FN)	True Negative Count (TN)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

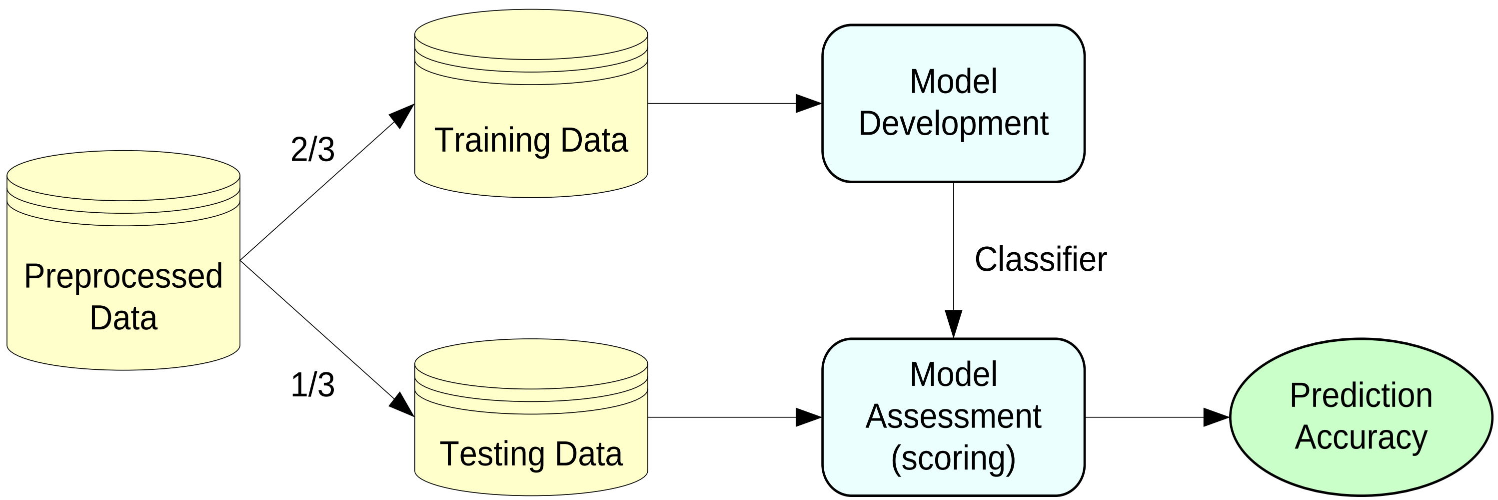
$$True\ Negative\ Rate = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Estimation Methodologies for Classification

- Simple split (or holdout or test sample estimation)
  - Split the data into 2 mutually exclusive sets training (~70%) and testing (30%)



  - For ANN, the data is split into three sub-sets (training [~60%], validation [~20%], testing [~20%])

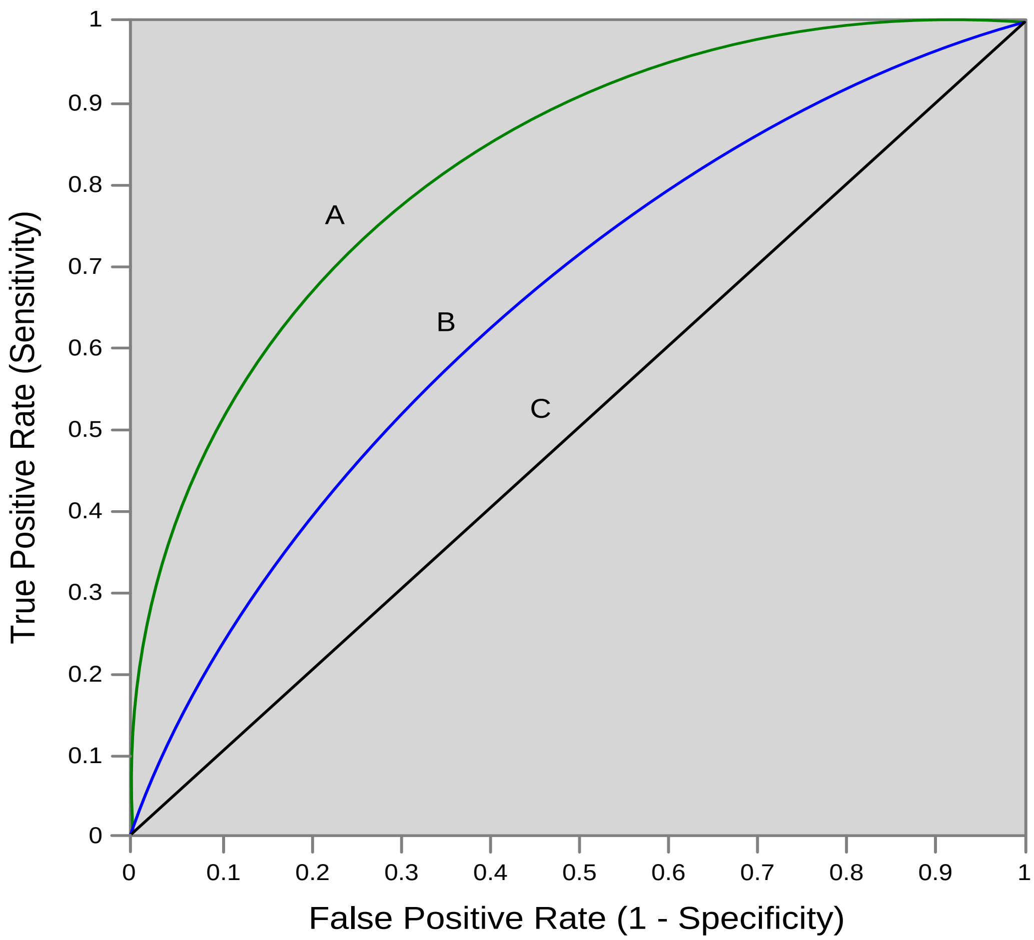
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Estimation Methodologies for Classification

- *k*-Fold Cross Validation (rotation estimation)
  - Split the data into *k* mutually exclusive subsets
  - Use each subset as testing while using the rest of the subsets as training
  - Repeat the experimentation for *k* times
  - Aggregate the test results for true estimation of prediction accuracy training
- Other estimation methodologies
  - Leave-one-out, bootstrapping, jackknifing
  - Area under the ROC curve

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Estimation Methodologies for Classification – ROC Curve



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Sensitivity =True Positive Rate

Specificity =True Negative Rate

		True Class (actual value)		
		Positive	Negative	**total**
Predictive Class (prediction outcome)	Positive	True Positive (TP)	False Positive (FP)	P'
	Negative	False Negative (FN)	True Negative (TN)	N'
total		P	N	

$$True\ Positive\ Rate\ (\text{Sensitivity}) = \frac{TP}{TP + FN}$$

$$True\ Negative\ Rate\ (\text{Specificity}) = \frac{TN}{TN + FP}$$

$$False\ Positive\ Rate = \frac{FP}{FP + TN}$$

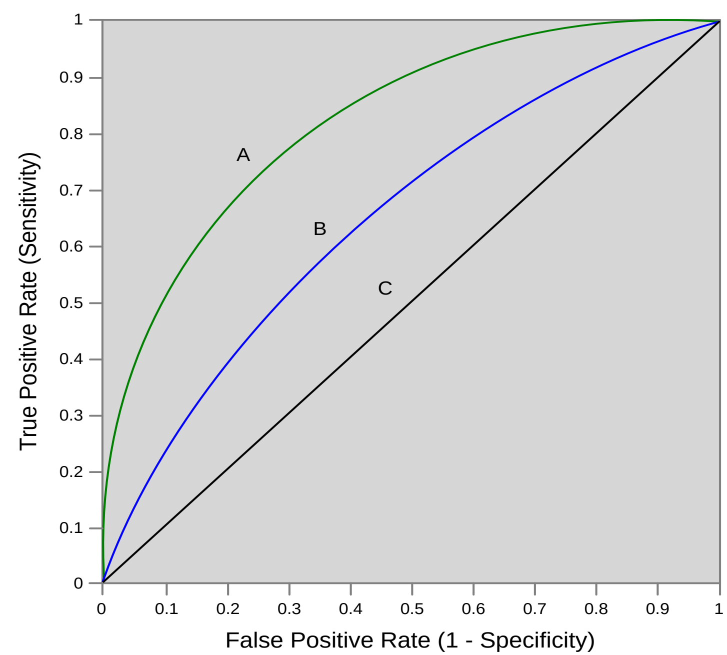
$$False\ Positive\ Rate\ (1 - \text{Specificity}) = \frac{FP}{FP + TN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

$$True\ Negative\ Rate = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$



Source: http://en.wikipedia.org/wiki/Receiver_operating_characteristic

True Class (actual value)

Predictive Class (prediction outcome)	Positive	Negative	**total**
Positive	True Positive (TP)	False Positive (FP)	P'
Negative	False Negative (FN)	True Negative (TN)	N'
total	P	N	

$$True\ Positive\ \ Rate\ (\text{Sensitivity}) = \frac{TP}{TP + FN}$$

Sensitivity

= True Positive Rate

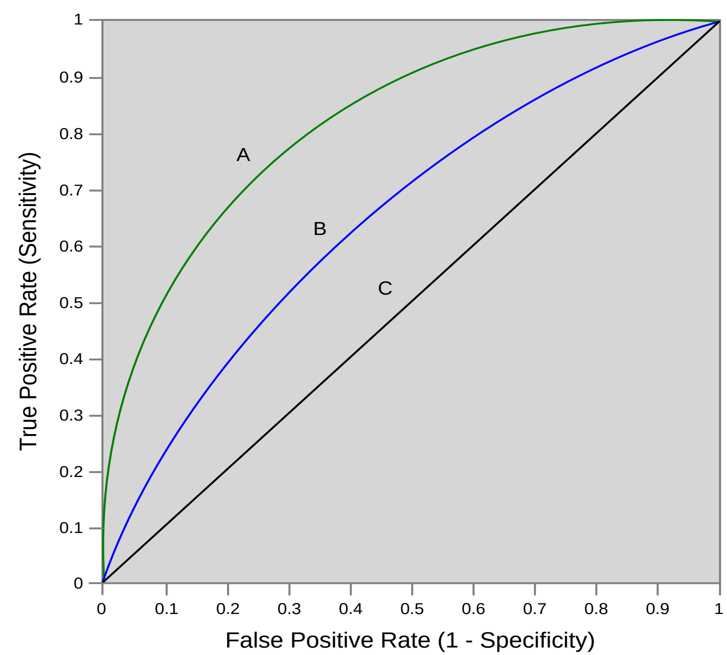
= Recall

= Hit rate

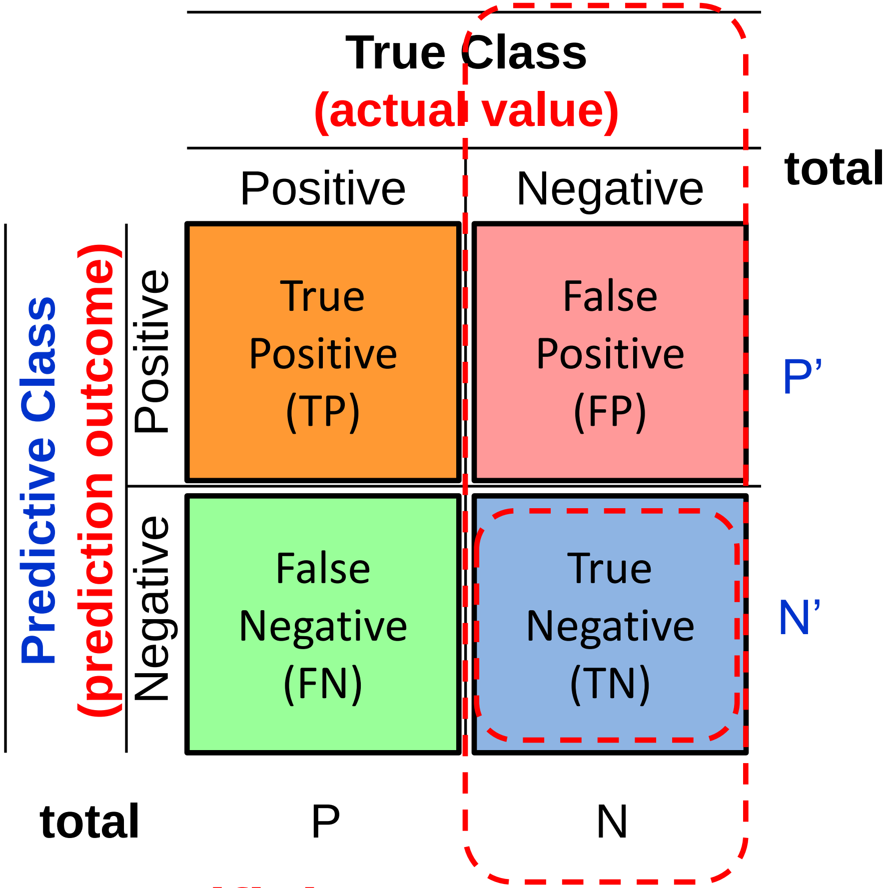
= TP / (TP + FN)

$$True\ Positive\ \ Rate = \frac{TP}{TP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$



Source: http://en.wikipedia.org/wiki/Receiver_operating_characteristic



$$True\ Negative\ Rate = \frac{TN}{TN + FP}$$

Specificity

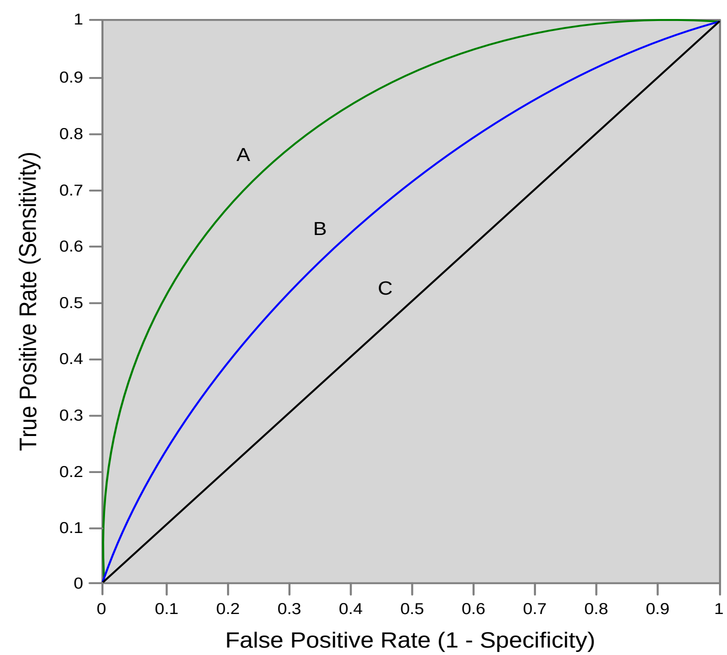
= True Negative Rate

= TN / N

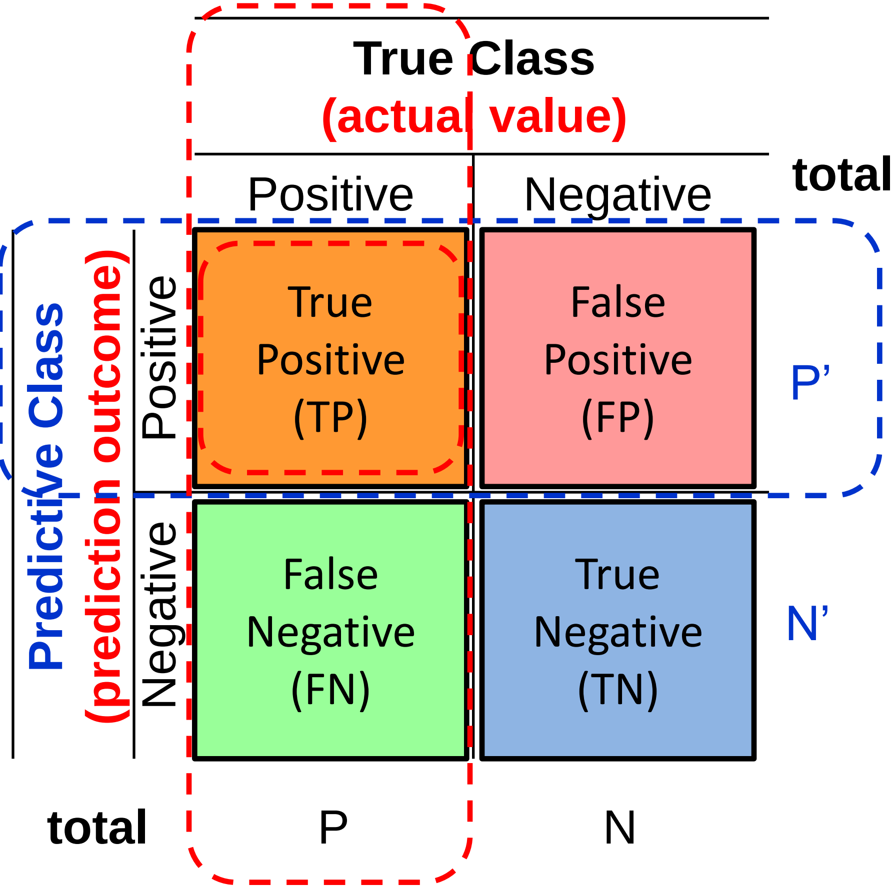
= TN / (TN+ FP)

$$True\ Negative\ Rate\ (\text{Specificity}) = \frac{TN}{TN + FP}$$

$$False\ Positive\ Rate\ (1 - \text{Specificity}) = \frac{FP}{FP + TN}$$



Source: http://en.wikipedia.org/wiki/Receiver_operating_characteristic



Precision

= Positive Predictive Value (PPV)

$$Precision = \frac{TP}{TP + FP}$$

Recall

= True Positive Rate (TPR)

= Sensitivity

= Hit Rate

$$Recall = \frac{TP}{TP + FN}$$

F1 score (F-score)(F-measure)

is the harmonic mean of precision and recall

= 2TP / (P + P')

= 2TP / (2TP + FP + FN)

$$F = 2 * \frac{precision * recall}{precision + recall}$$

Source: http://en.wikipedia.org/wiki/Receiver_operating_characteristic

A

63 (TP)	28 (FP)	91
37 (FN)	72 (TN)	109
100	100	200

Recall
= True Positive Rate (TPR)
= Sensitivity
= Hit Rate
= TP / (TP + FN)

Specificity
= True Negative Rate
= TN / N
= TN / (TN + FP)

TPR = 0.63

$$Recall = \frac{TP}{TP + FN}$$

$$True\ Negative\ Rate\ (\text{Specificity}) = \frac{TN}{TN + FP}$$

FPR = 0.28

$$False\ Positive\ Rate\ (1 - \text{Specificity}) = \frac{FP}{FP + TN}$$

PPV = 0.69
=63/(63+28)
=63/91

$$Precision = \frac{TP}{TP + FP}$$

Precision
= Positive Predictive Value (PPV)

F1 = 0.66
= 2*(0.63*0.69)/(0.63+0.69)
= (2 * 63) /(100 + 91)
= (0.63 + 0.69) / 2 =1.32 / 2 =0.66

$$F = 2 * \frac{precision * recall}{precision + recall}$$

F1 score (F-score) (F-measure)
is the harmonic mean of precision and recall
= 2TP / (P + P')
= 2TP / (2TP + FP + FN)

ACC = 0.68
= (63 + 72) / 200
= 135/200 = 67.5

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Source: http://en.wikipedia.org/wiki/Receiver_operating_characteristic

A

63 (TP)	28 (FP)	91
37 (FN)	72 (TN)	109
100	100	200

TPR = 0.63

FPR = 0.28

PPV = 0.69
=63/(63+28)
=63/91

F1 = 0.66
= 2*(0.63*0.69)/(0.63+0.69)
= (2 * 63) /(100 + 91)
= (0.63 + 0.69) / 2 =1.32 / 2 =0.66

ACC = 0.68
= (63 + 72) / 200
= 135/200 = 67.5

B

77 (TP)	77 (FP)	154
23 (FN)	23 (TN)	46
100	100	200

TPR = 0.77

FPR = 0.77

PPV = 0.50

F1 = 0.61

ACC = 0.50

Recall
= True Positive Rate (TPR)
= Sensitivity
= Hit Rate

$$Recall = \frac{TP}{TP + FN}$$

Precision
= Positive Predictive Value (PPV)

$$Precision = \frac{TP}{TP + FP}$$

Source: http://en.wikipedia.org/wiki/Receiver_operating_characteristic

C

24 (TP)	88 (FP)	112
76 (FN)	12 (TN)	88
100	100	200

TPR = 0.24
FPR = 0.88
PPV = 0.21
F1 = 0.22
ACC = 0.18

C'

76 (TP)	12 (FP)	88
24 (FN)	88 (TN)	112
100	100	200

TPR = 0.76
FPR = 0.12
PPV = 0.86
F1 = 0.81
ACC = 0.82

Recall
= True Positive Rate (TPR)
= Sensitivity
= Hit Rate

$$Recall = \frac{TP}{TP + FN}$$

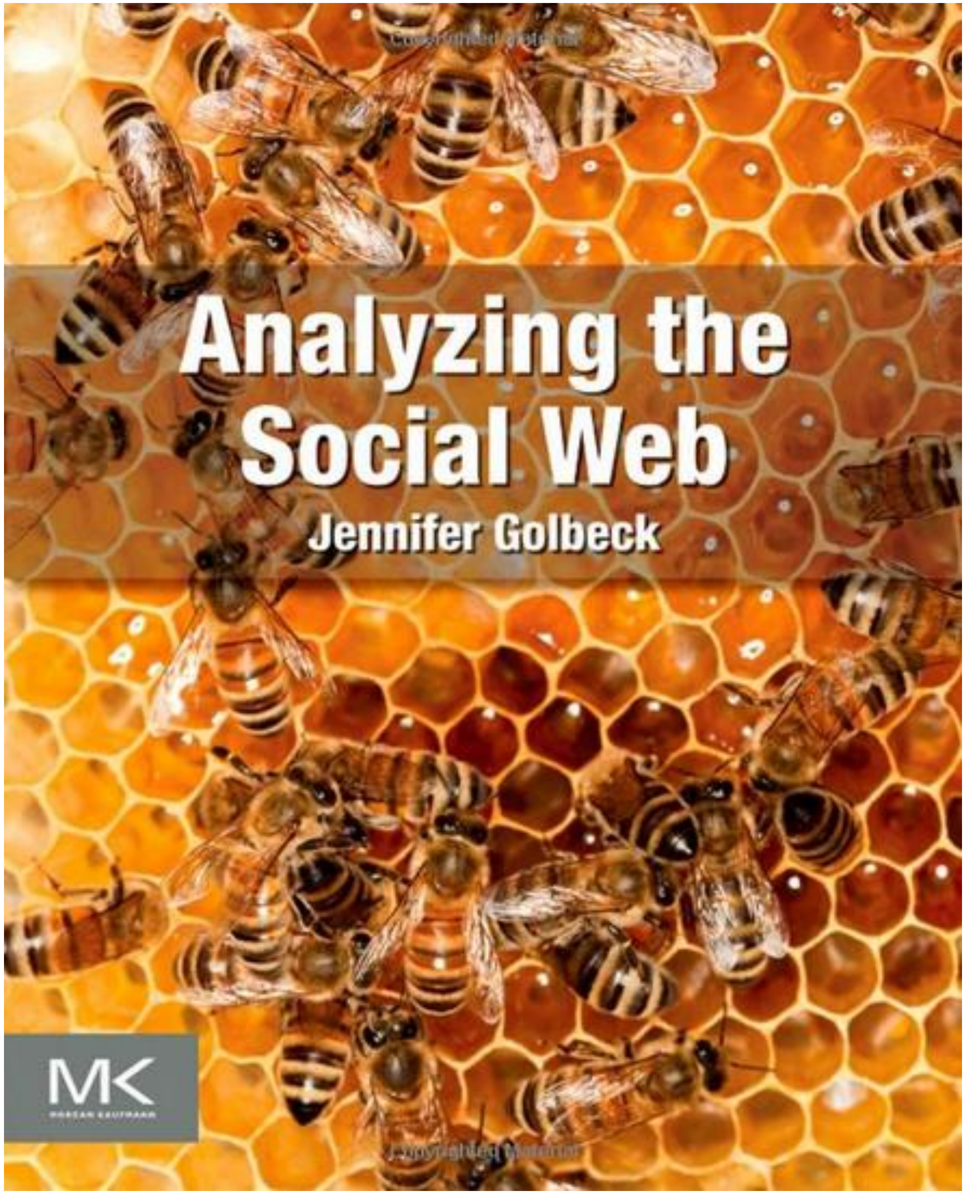
Precision
= Positive Predictive Value (PPV)

$$Precision = \frac{TP}{TP + FP}$$

Source: http://en.wikipedia.org/wiki/Receiver_operating_characteristic

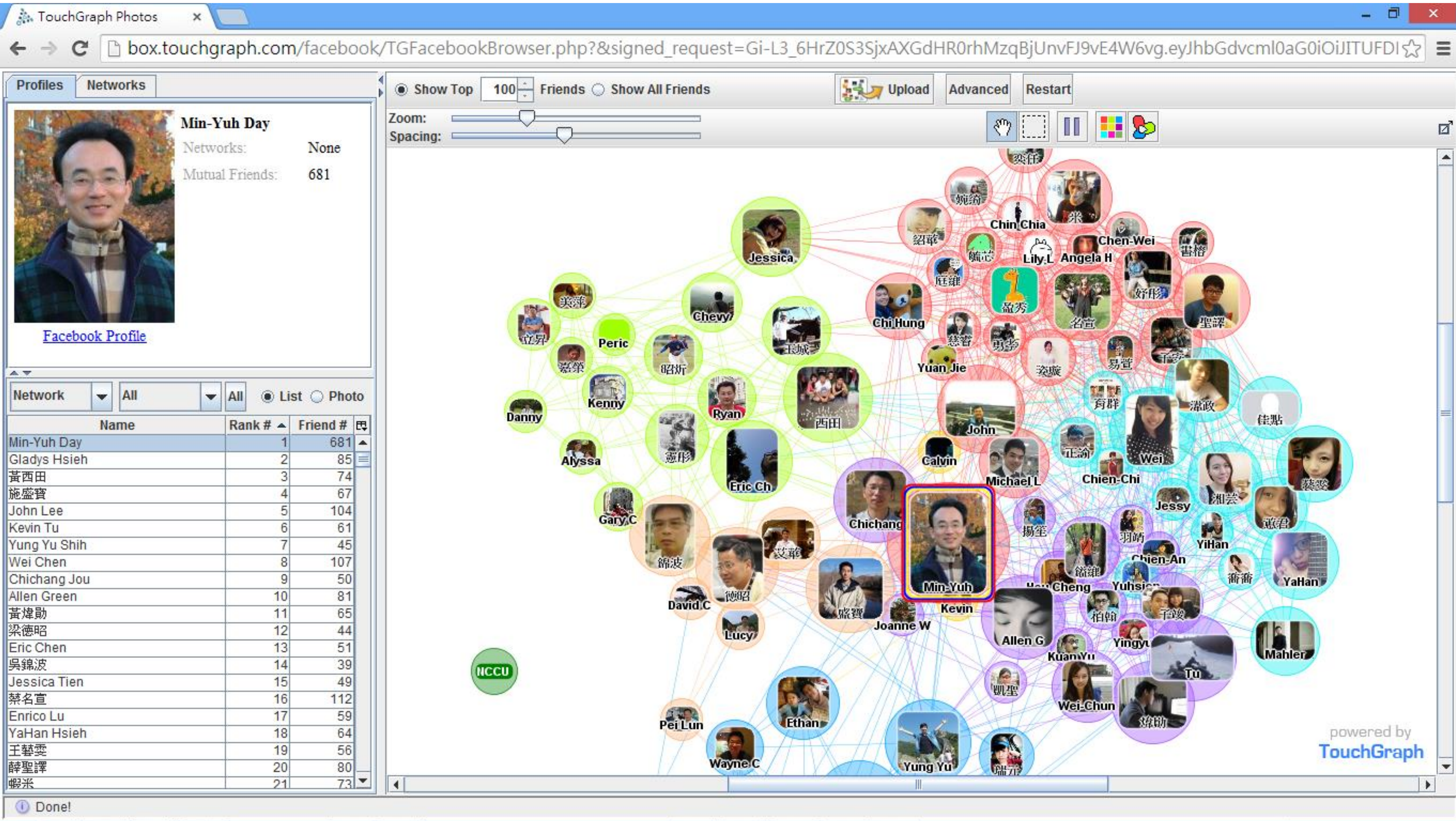
Social Network Analysis

Jennifer Golbeck (2013), **Analyzing the Social Web**, Morgan Kaufmann

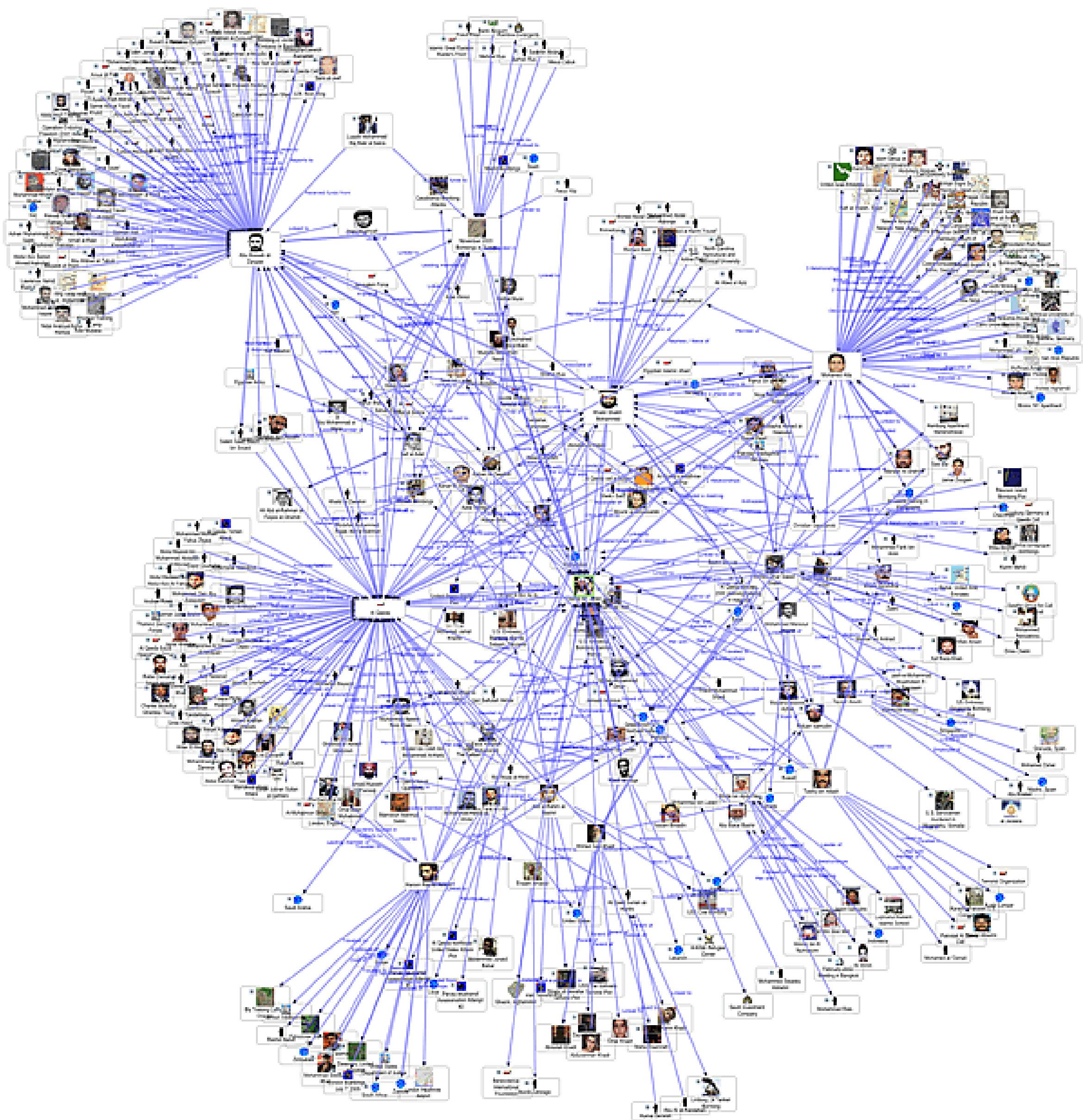


Source: http://www.amazon.com/Analyzing-Social-Web-Jennifer-Golbeck/dp/0124055311

Social Network Analysis (SNA)
Facebook TouchGraph



Social Network Analysis



Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Social Network Analysis

- A **social network** is a social structure of people, related (directly or indirectly) to each other through a common relation or interest
- **Social network analysis (SNA)** is the study of social networks to understand their structure and behavior

Source: (c) Jaideep Srivastava, srivasta@cs.umn.edu, Data Mining for Social Network Analysis

Social Network Analysis

- Using Social Network Analysis, you can get answers to questions like:
  - How highly connected is an entity within a network?
  - What is an entity's overall importance in a network?
  - How central is an entity within a network?
  - How does information flow within a network?

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Social Network Analysis

- Social network is the study of social entities (people in an organization, called **actors**), and their interactions and relationships.
- The interactions and relationships can be represented with a network or graph,
  - each vertex (or node) represents an actor and
  - each link represents a relationship.
- From the network, we can study the properties of its structure, and the role, position and prestige of each social actor.
- We can also find various kinds of sub-graphs, e.g., **communities** formed by groups of actors.

Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data”

Social Network and the Web

- Social network analysis is useful for the Web because the Web is essentially a virtual society, and thus a virtual social network,
  - Each page: a social actor and
  - each hyperlink: a relationship.
- Many results from social network can be adapted and extended for use in the Web context.
- Two types of social network analysis,
  - **Centrality**
  - **Prestige**

  closely related to hyperlink analysis and search on the Web

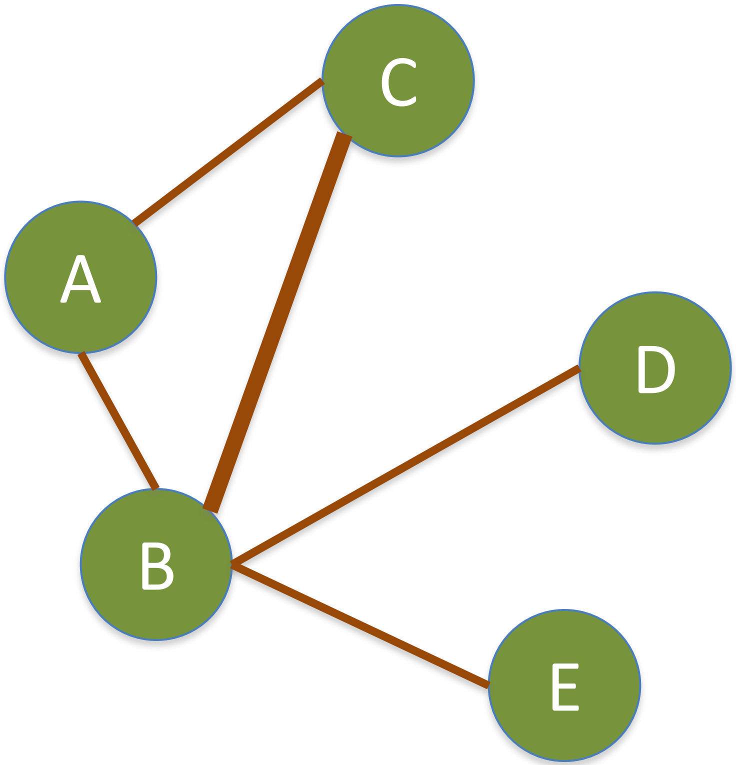
Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data”

Social Network Analysis (SNA)

Centrality

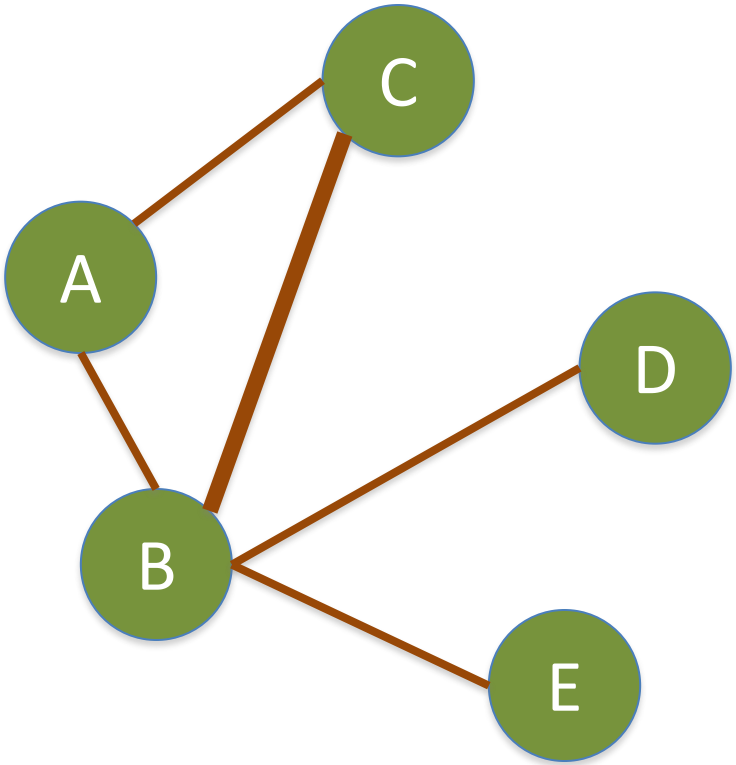
Prestige

Degree



Source: https://www.youtube.com/watch?v=89mxOdwPfxA

Degree



A: 2

<u>B: 4</u>

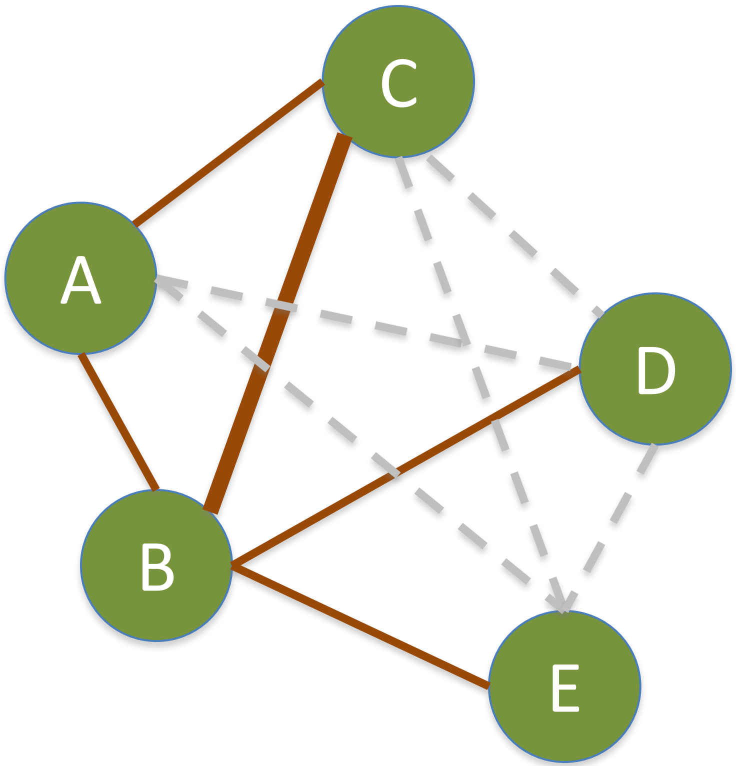
C: 2

D:1

E: 1

Source: https://www.youtube.com/watch?v=89mxOdwPfxA

Density



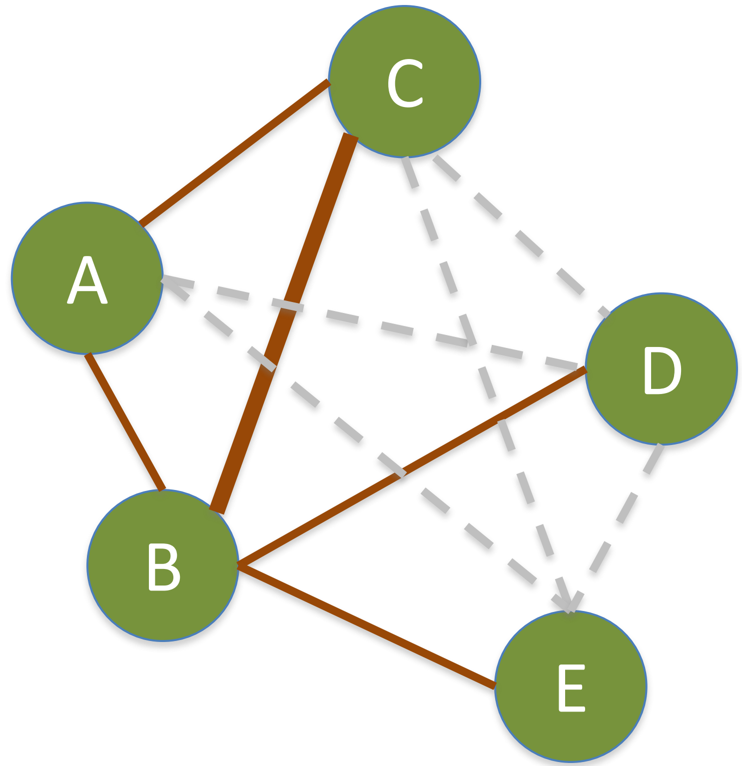
Source: https://www.youtube.com/watch?v=89mxOdwPfxA

Density

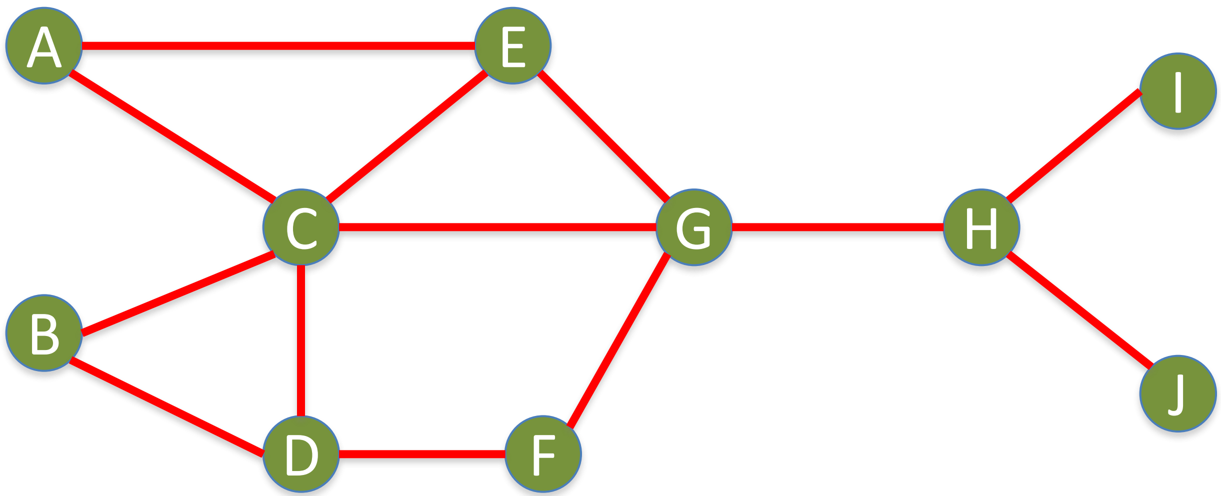
Edges (Links): 5

Total Possible Edges: 10

Density: 5/10 = 0.5

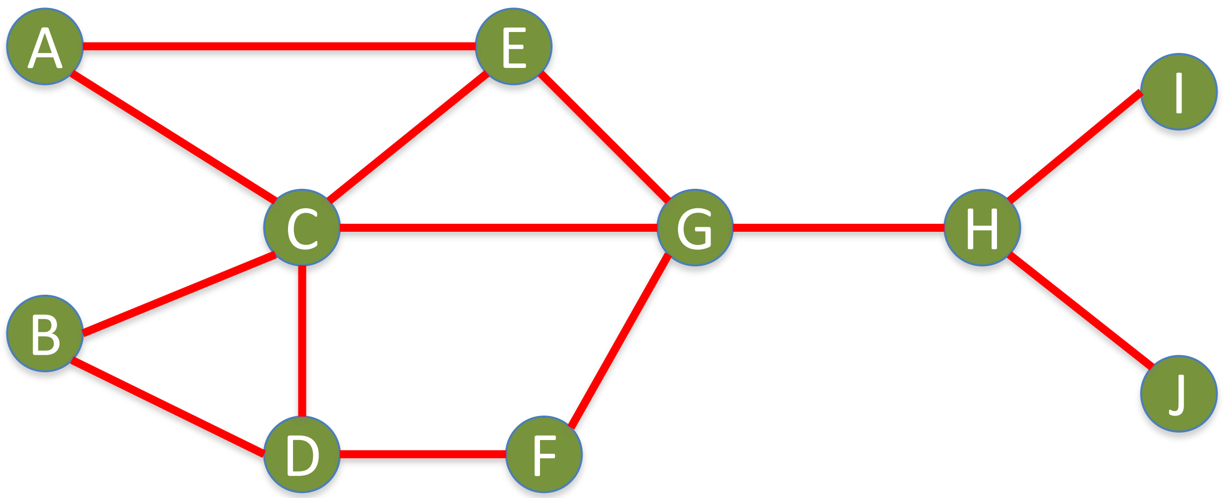


Density



Nodes (n): 10
Edges (Links): 13
Total Possible Edges: (n * (n-1)) / 2 = (10 * 9) / 2 = 45
Density: 13/45 = 0.29

Which Node is Most **Important**?



Centrality

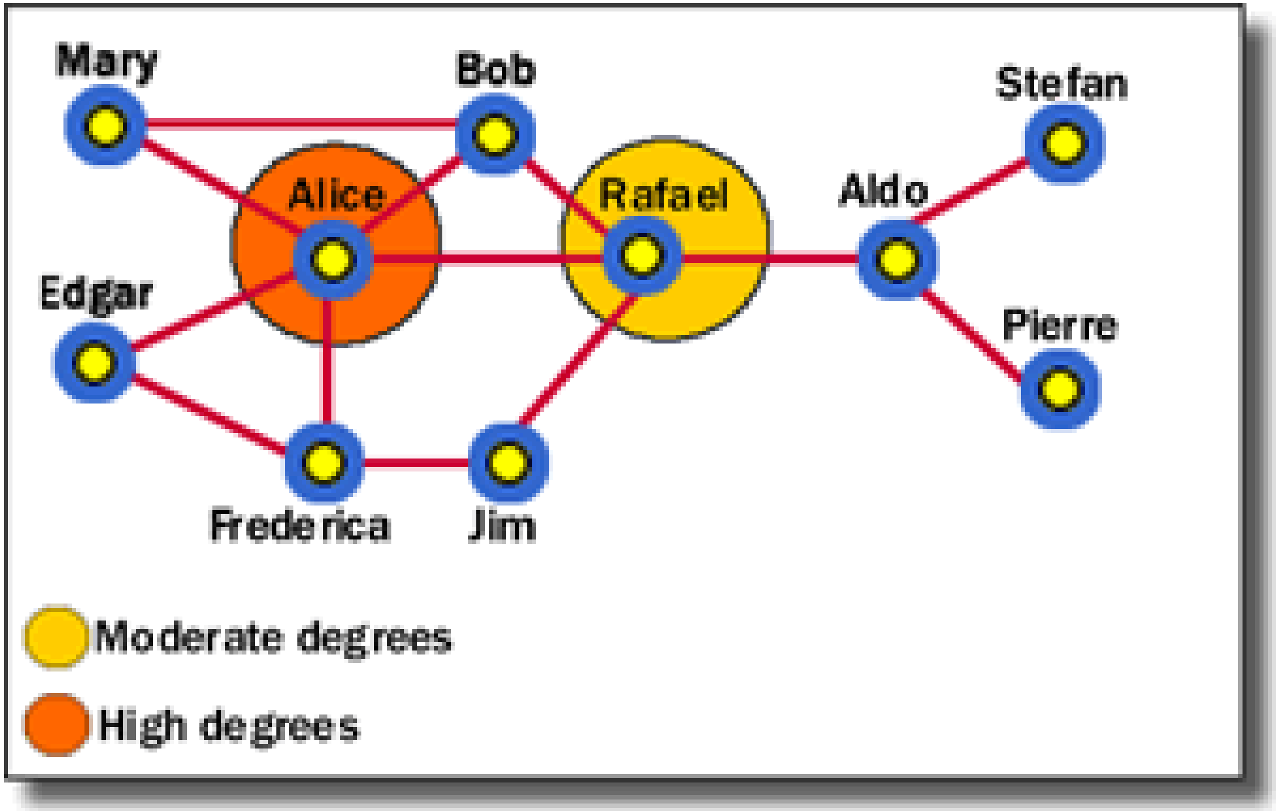
- Important or prominent actors are those that are linked or involved with other actors extensively.
- A person with extensive contacts (links) or communications with many other people in the organization is considered more important than a person with relatively fewer contacts.
- The links can also be called **ties**.
  A central actor is one involved in many ties.

Source: Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data"

Social Network Analysis (SNA)

- Degree Centrality
- Betweenness Centrality
- Closeness Centrality

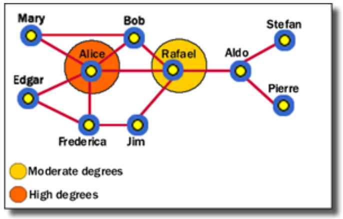
Social Network Analysis: Degree Centrality



Alice has the highest degree centrality, which means that she is quite active in the network. However, she is not necessarily the most powerful person because she is only directly connected within one degree to people in her clique—she has to go through Rafael to get to other cliques.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

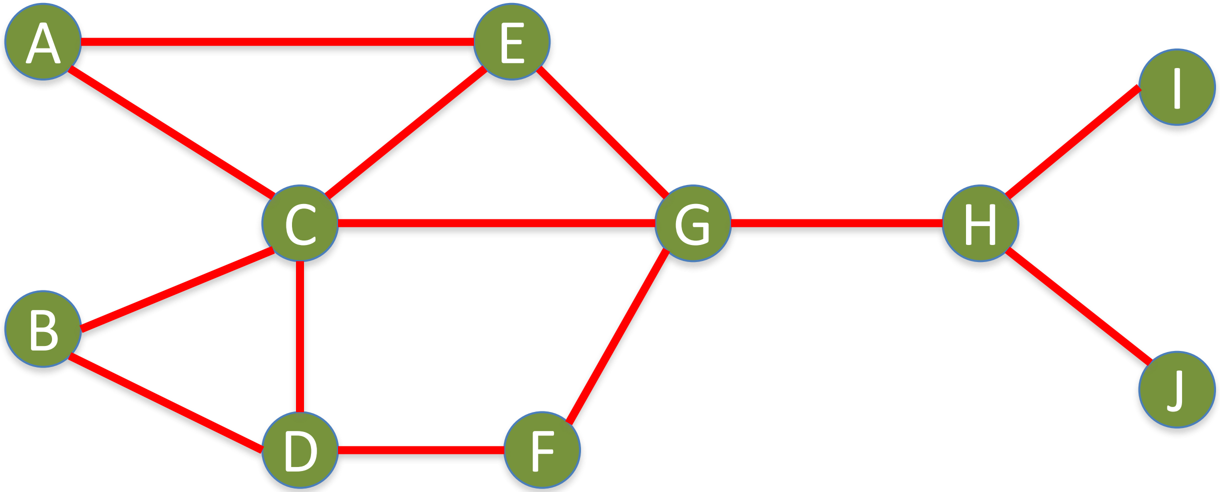
Social Network Analysis: Degree Centrality



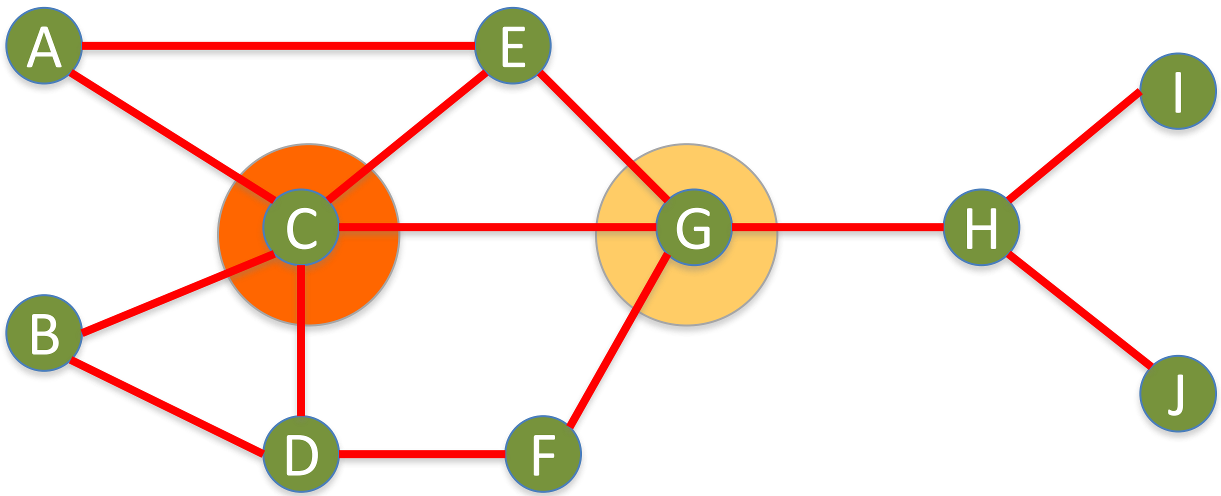
- Degree centrality is simply the number of direct relationships that an entity has.
- An entity with high degree centrality:
  - Is generally an active player in the network.
  - Is often a connector or hub in the network.
  - s not necessarily the most connected entity in the network (an entity may have a large number of relationships, the majority of which point to low-level entities).
  - May be in an advantaged position in the network.
  - May have alternative avenues to satisfy organizational needs, and consequently may be less dependent on other individuals.
  - Can often be identified as third parties or deal makers.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Social Network Analysis: Degree Centrality

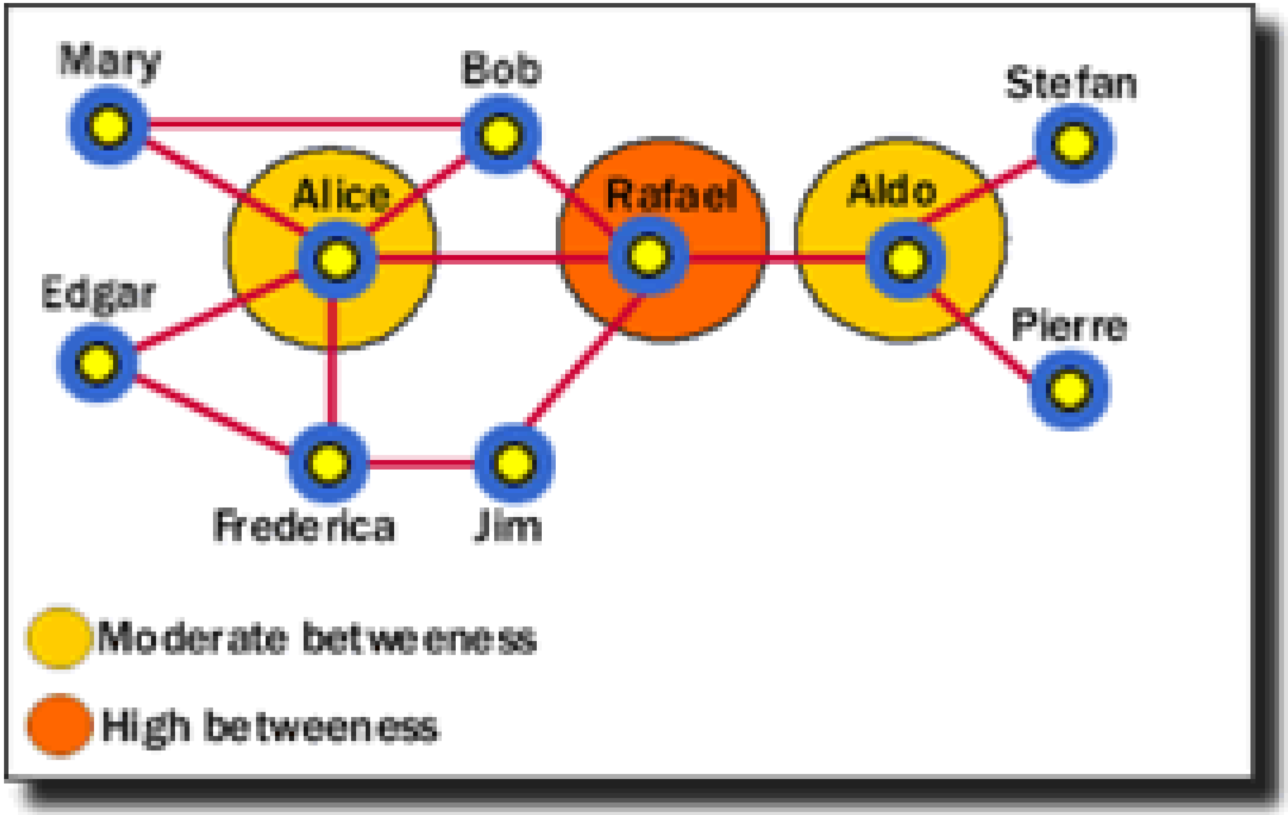


Social Network Analysis: Degree Centrality



Node	Score	Standardized Score
A	2	2/10 = 0.2
B	2	2/10 = 0.2
C	**5**	**5/10 = 0.5**
D	3	3/10 = 0.3
E	3	3/10 = 0.3
F	2	2/10 = 0.2
G	**4**	**4/10 = 0.4**
H	3	3/10 = 0.3
I	1	1/10 = 0.1
J	1	1/10 = 0.1

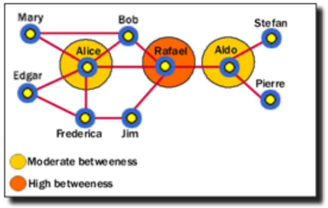
Social Network Analysis: Betweenness Centrality



Rafael has the highest betweenness because he is between Alice and Aldo, who are between other entities. Alice and Aldo have a slightly lower betweenness because they are essentially only between their own cliques. Therefore, although Alice has a higher degree centrality, Rafael has more importance in the network in certain respects.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Social Network Analysis: Betweenness Centrality



- Betweenness centrality identifies an entity's position within a network in terms of its ability to make connections to other pairs or groups in a network.
- An entity with a high betweenness centrality generally:
  - Holds a favored or powerful position in the network.
  - Represents a single point of failure—take the single betweenness spanner out of a network and you sever ties between cliques.
  - Has a greater amount of influence over what happens in a network.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Betweenness centrality:

Connectivity

Number of shortest paths going through the actor

Betweenness Centrality

$$C_B(i) = \sum_{j<k} g_{ik}(i) / g_{jk}$$

Where g_{jk} = the number of shortest paths connecting jk
$g_{jk}(i)$ = the number of that actor i is on.

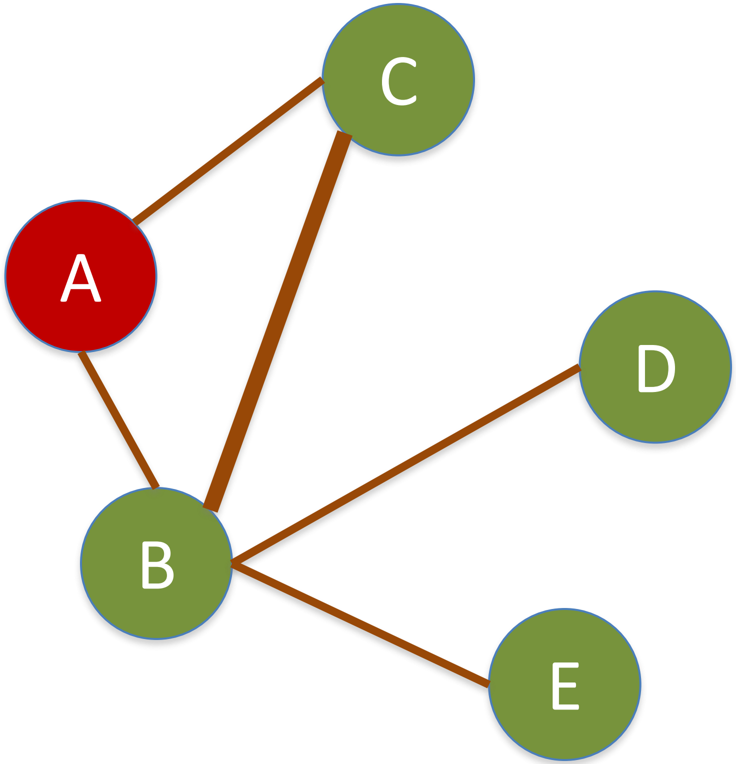
Normalized Betweenness Centrality

$$C'_B(i) = C_B(i) / \boxed{[(n-1)(n-2)/2]}$$

Number of pairs of vertices excluding the vertex itself

Source: https://www.youtube.com/watch?v=RXohUeNCJiU

Betweenness Centrality

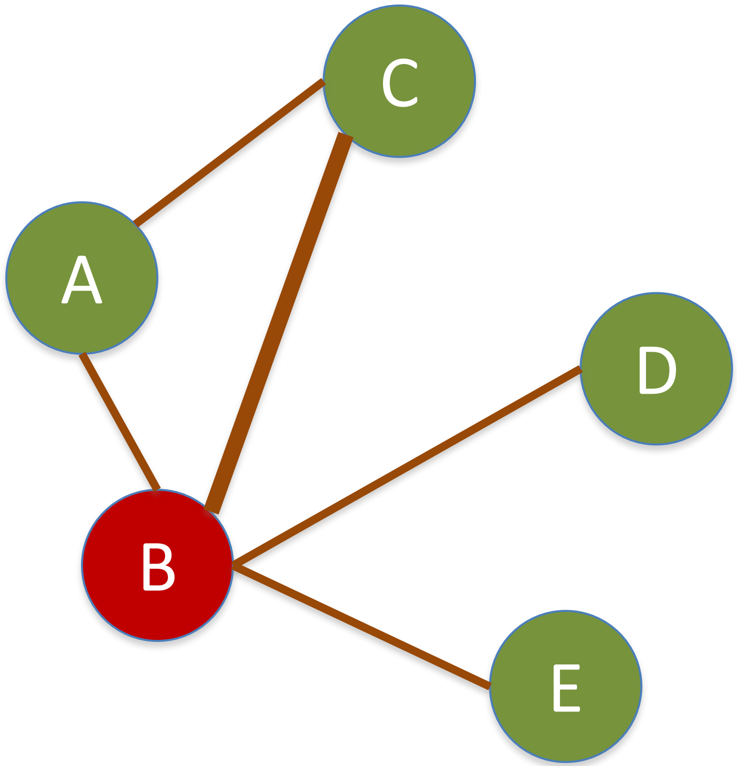


A:
B→C: 0/1 = 0
B→D: 0/1 = 0
B→E: 0/1 = 0
C→D: 0/1 = 0
C→E: 0/1 = 0
D→E: 0/1 = 0

Total: 0

A: Betweenness Centrality = 0

Betweenness Centrality

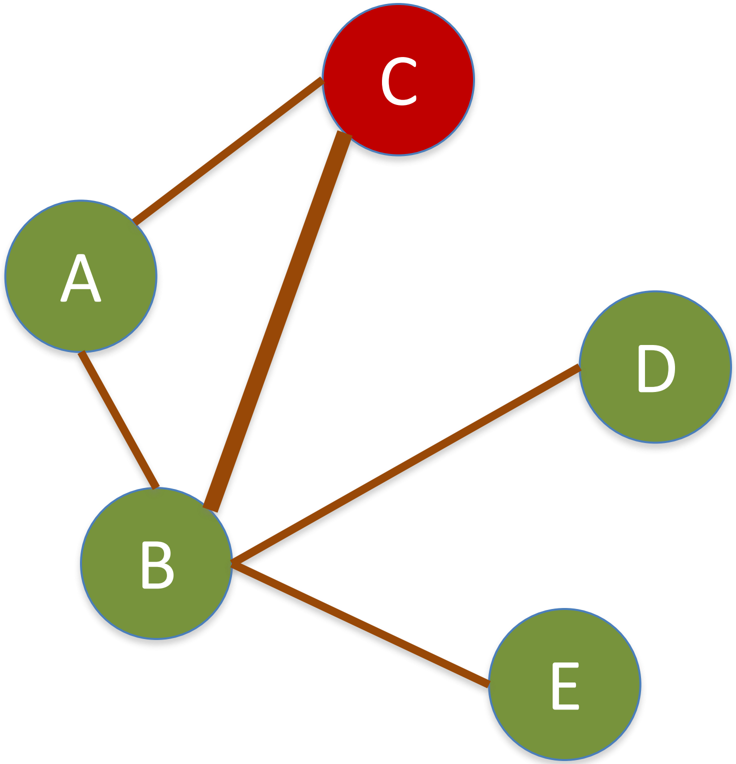


B:
A→C: 0/1 = 0
A→D: 1/1 = 1
A→E: 1/1 = 1
C→D: 1/1 = 1
C→E: 1/1 = 1
D→E: 1/1 = 1

Total: 5

B: Betweenness Centrality = 5

Betweenness Centrality

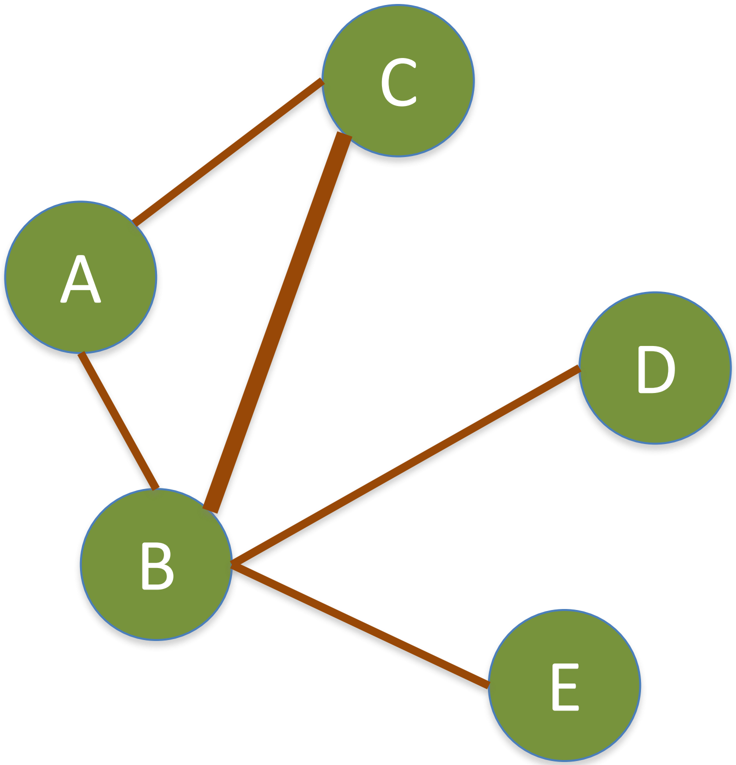


C:
A→B: 0/1 = 0
A→D: 0/1 = 0
A→E: 0/1 = 0
B→D: 0/1 = 0
B→E: 0/1 = 0
D→E: 0/1 = 0

Total: 0

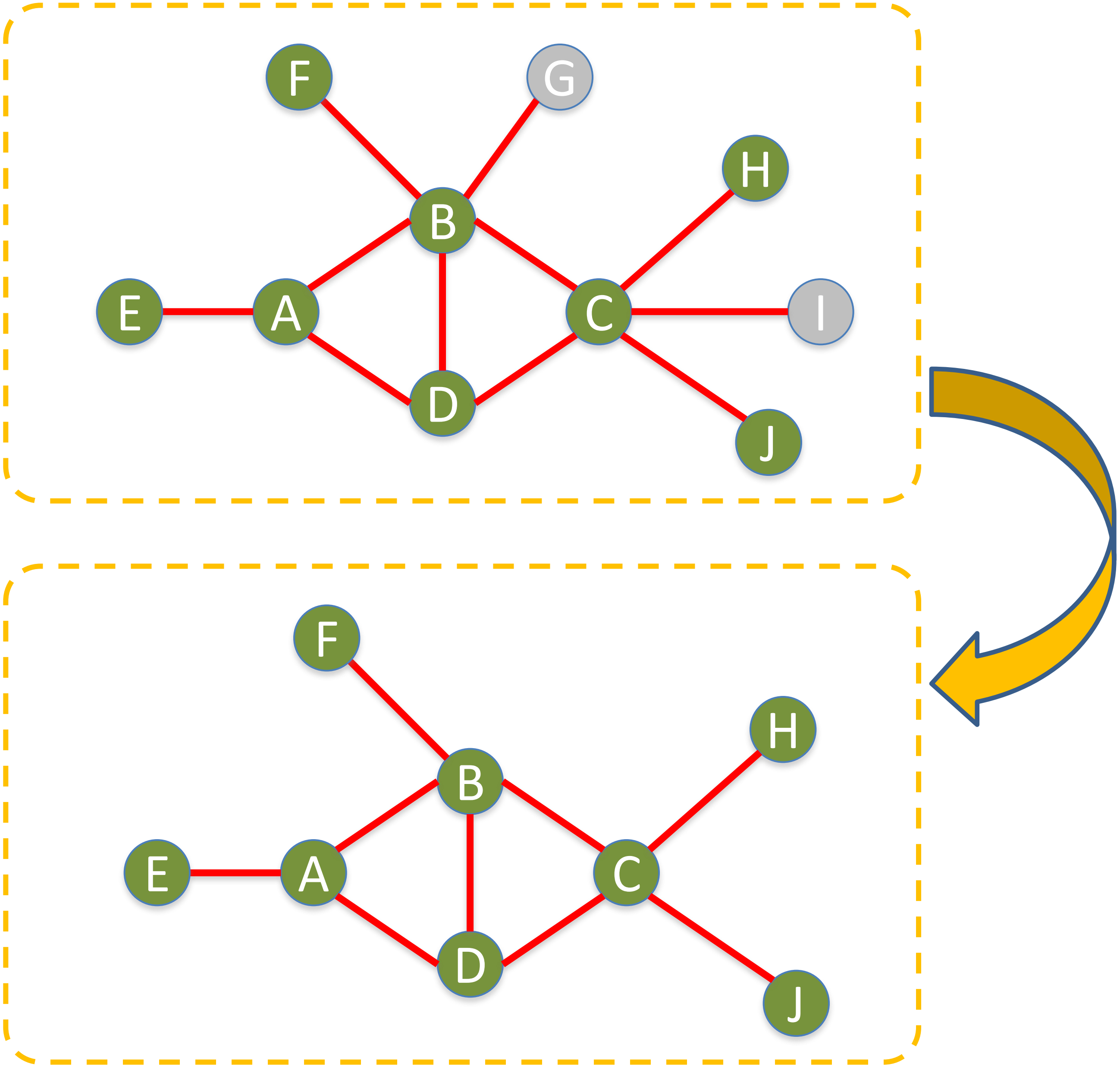
C: Betweenness Centrality = 0

Betweenness Centrality

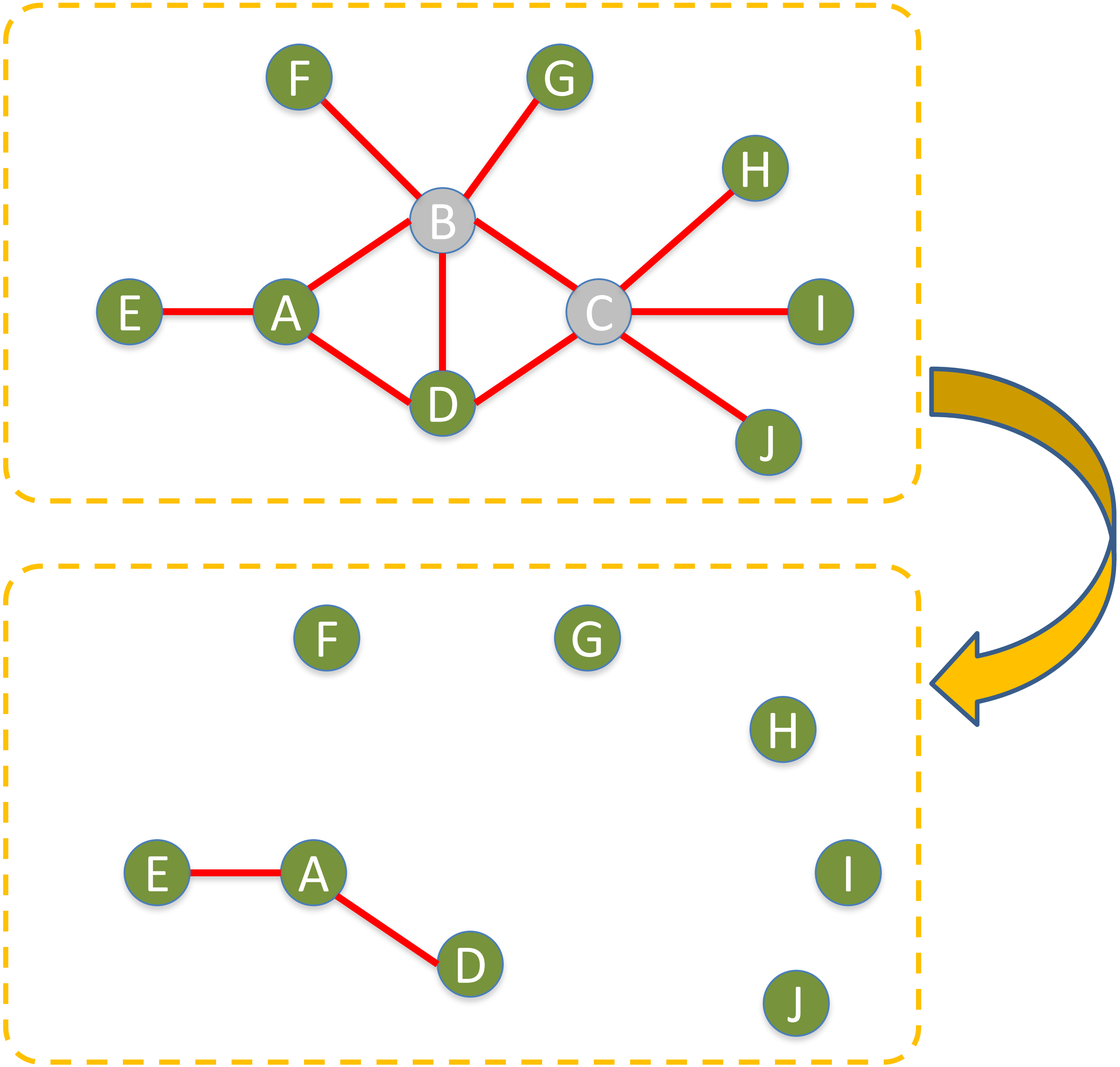


A: 0
<u>B: 5</u>
C: 0
D: 0
E: 0

Which Node is Most Important?

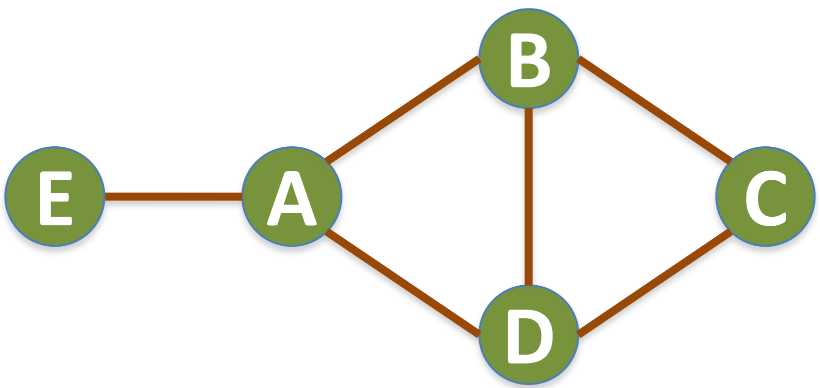


Which Node is Most Important?

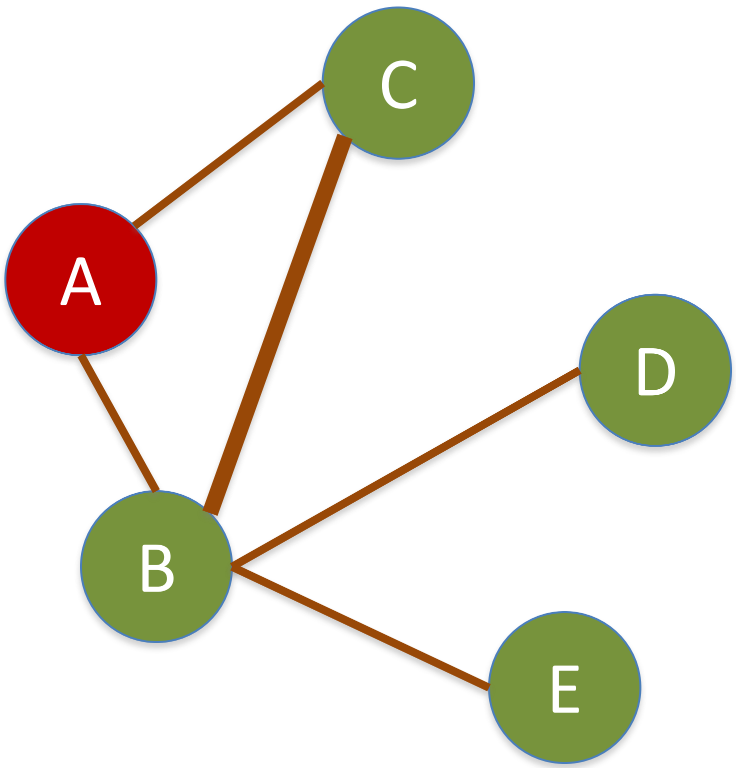


Betweenness Centrality

$$C_B(i) = \sum_{j<k} g_{jk}(i) / g_{jk}$$



Betweenness Centrality

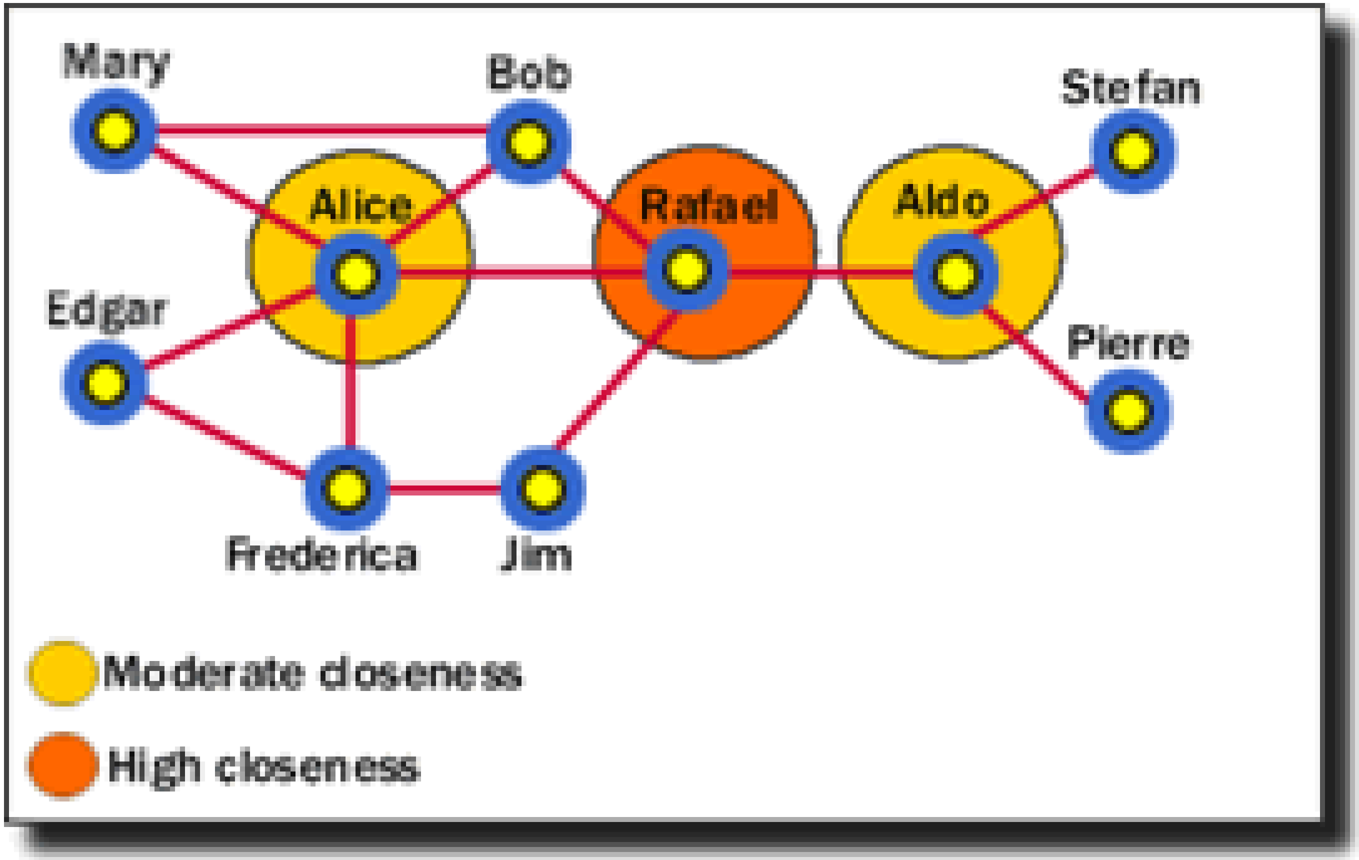


A:
B→C: 0/1 = 0
B→D: 0/1 = 0
B→E: 0/1 = 0
C→D: 0/1 = 0
C→E: 0/1 = 0
D→E: 0/1 = 0

Total: 0

A: Betweenness Centrality = 0

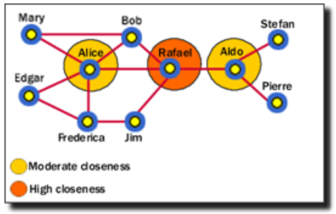
Social Network Analysis: Closeness Centrality



Rafael has the highest closeness centrality because he can reach more entities through shorter paths. As such, Rafael's placement allows him to connect to entities in his own clique, and to entities that span cliques.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

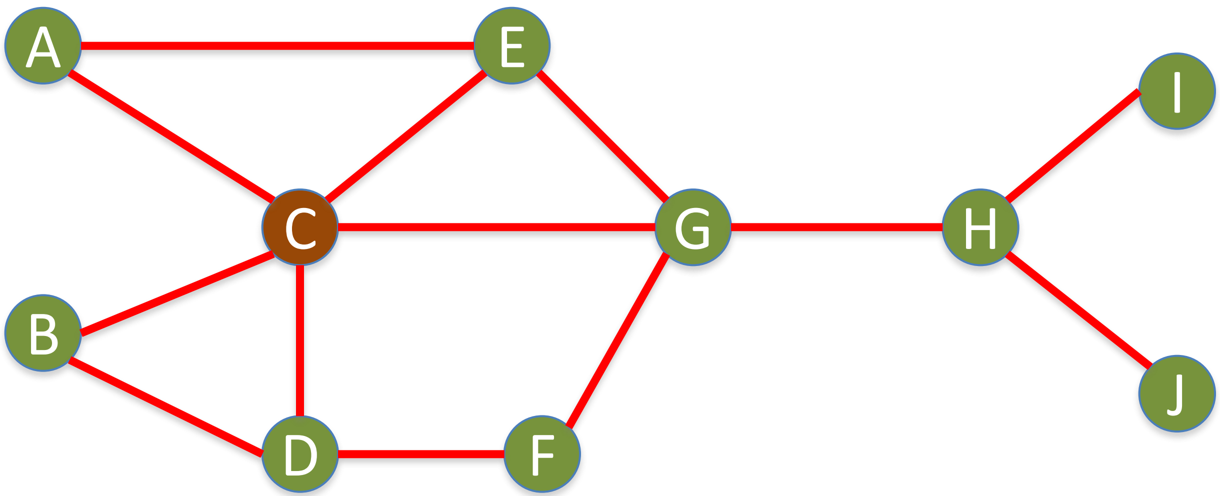
Social Network Analysis: Closeness Centrality



- Closeness centrality measures how quickly an entity can access more entities in a network.
- An entity with a high closeness centrality generally:
  - Has quick access to other entities in a network.
  - Has a short path to other entities.
  - Is close to other entities.
  - Has high visibility as to what is happening in the network.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Social Network Analysis: Closeness Centrality

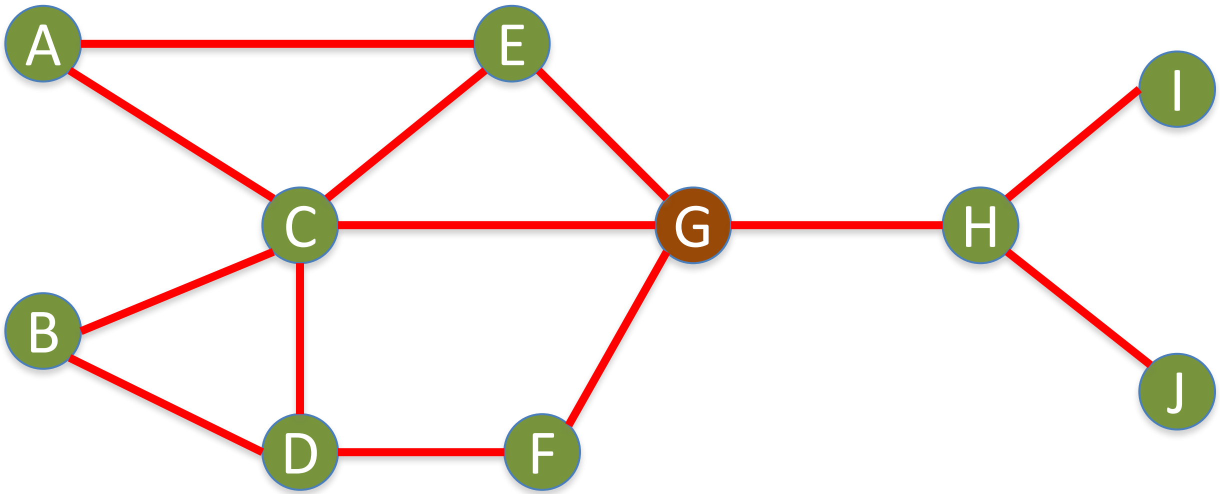


C→A: 1
C→B: 1
C→D: 1
C→E: 1
C→F: 2
C→G: 1
C→H: 2
C→I: 3
C→J: 3

Total=15

C: Closeness Centrality = 15/9 = 1.67

Social Network Analysis: Closeness Centrality

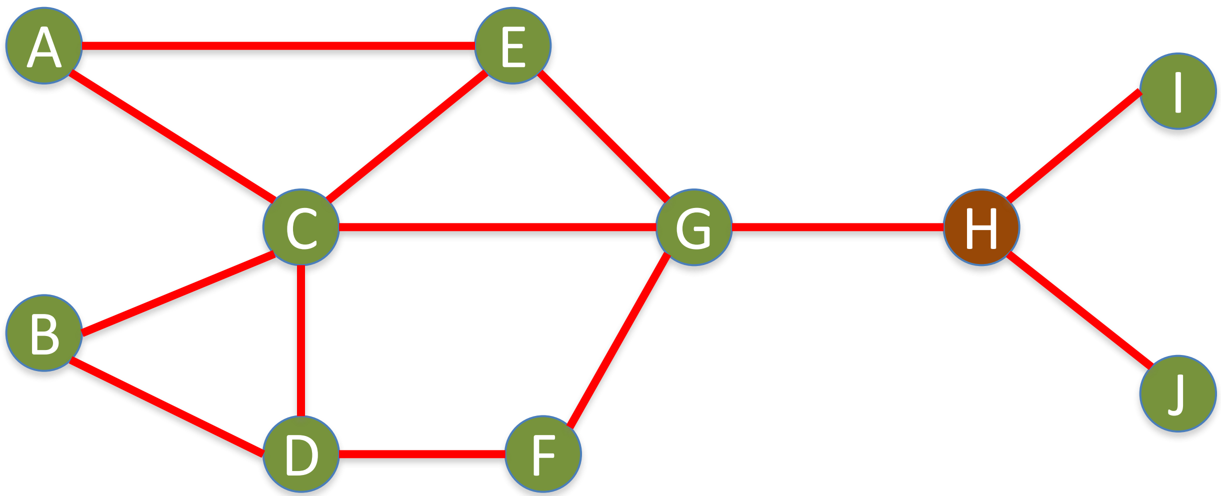


G→A: 2
G→B: 2
G→C: 1
G→D: 2
G→E: 1
G→F: 1
G→H: 1
G→I: 2
G→J: 2

Total=14

G: Closeness Centrality = 14/9 = 1.56

Social Network Analysis: Closeness Centrality

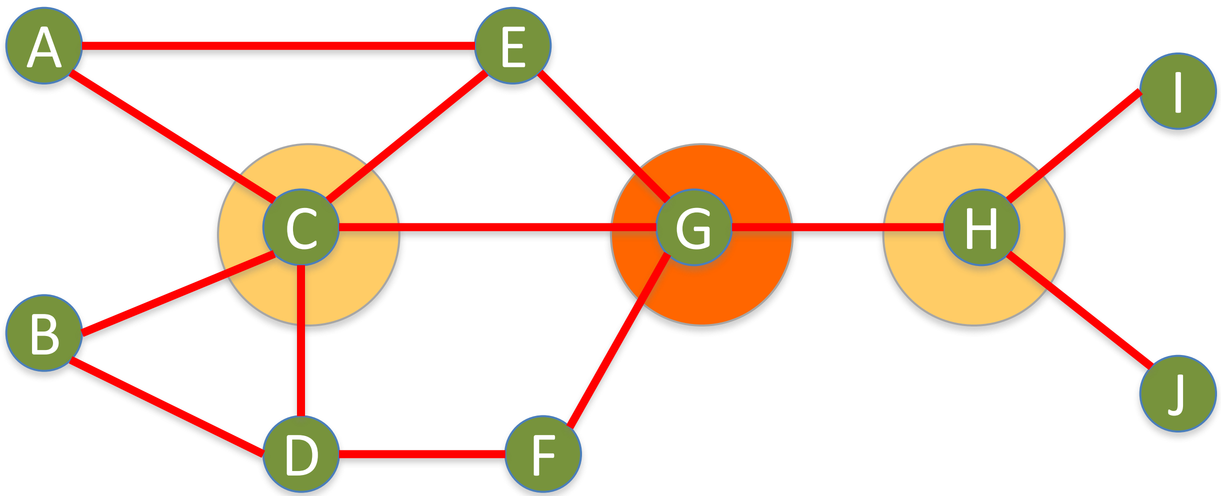


H→A: 3
H→B: 3
H→C: 2
H→D: 2
H→E: 2
H→F: 2
H→G: 1
H→I: 1
H→J: 1

Total=17

H: Closeness Centrality = 17/9 = 1.89

Social Network Analysis: Closeness Centrality



G: Closeness Centrality = 14/9 = 1.56 1

C: Closeness Centrality = 15/9 = 1.67 2

H: Closeness Centrality = 17/9 = 1.89 3

Social Network Analysis: Closeness Centrality

Sum of the reciprocal distances

$$C_C(p_k) = \sum_{i=1}^{n} d(p_i, p_k)^{-1}$$

where $d(p_j, p_k)$ is the geodesic distance (shortest paths) linking p_j, p_k

Social Network Analysis: Betweenness Centrality

$$C_B(p_k) = \sum_{i<j}^{n} \frac{g_{ij}(p_k)}{g_{ij}}; \qquad i \neq j \neq k$$

where g_{ij} is the geodesic distance (shortest paths) linking p_i and p_j and $g_{ij}(p_k)$ is the geodesic distance linking p_i and p_j that contains p_k.

Source: Abbasi, A., Hossain, L., & Leydesdorff, L. (2012). Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. *Journal of Informetrics*, *6*(3), 403-412.

Social Network Analysis: Degree Centrality

$$C_D(p_k) = \sum_{i=1}^{n} a(p_i, p_k)$$

where $a(p_i, p_k) = 1$ if and only if p_i and p_k are connected by a line
0 otherwise

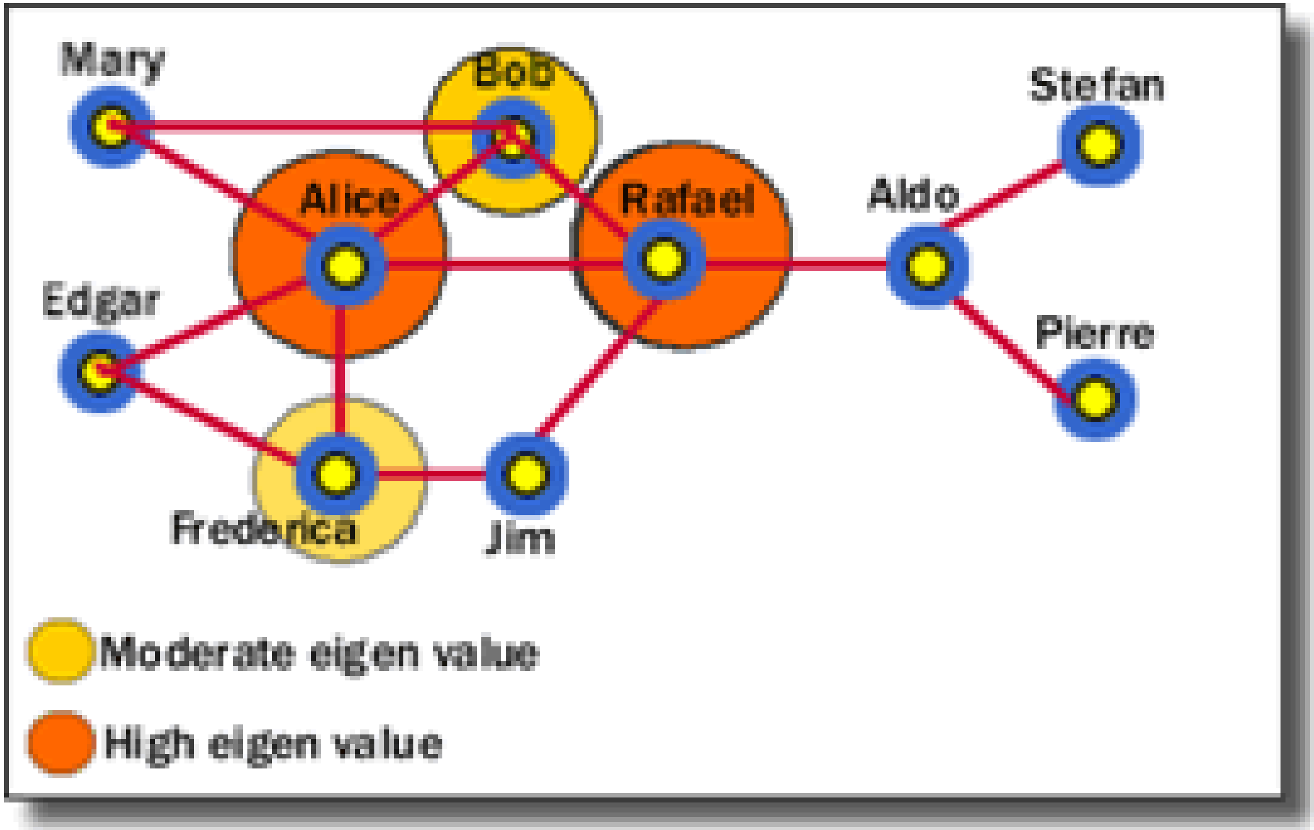
$$C'_D(p_k) = \frac{\sum_{i=1}^{n} a(p_i, p_k)}{n-1}$$

Social Networks, 1 (1978/79) 215–239
215

Centrality in Social Networks Conceptual Clarification

Linton C. Freeman
*Lehigh University**

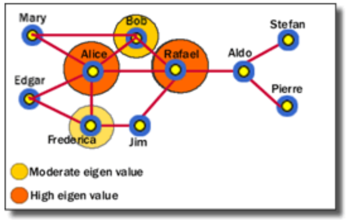
Social Network Analysis: Eigenvalue



Alice and Rafael are closer to other highly close entities in the network. Bob and Frederica are also highly close, but to a lesser value.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Social Network Analysis: Eigenvalue



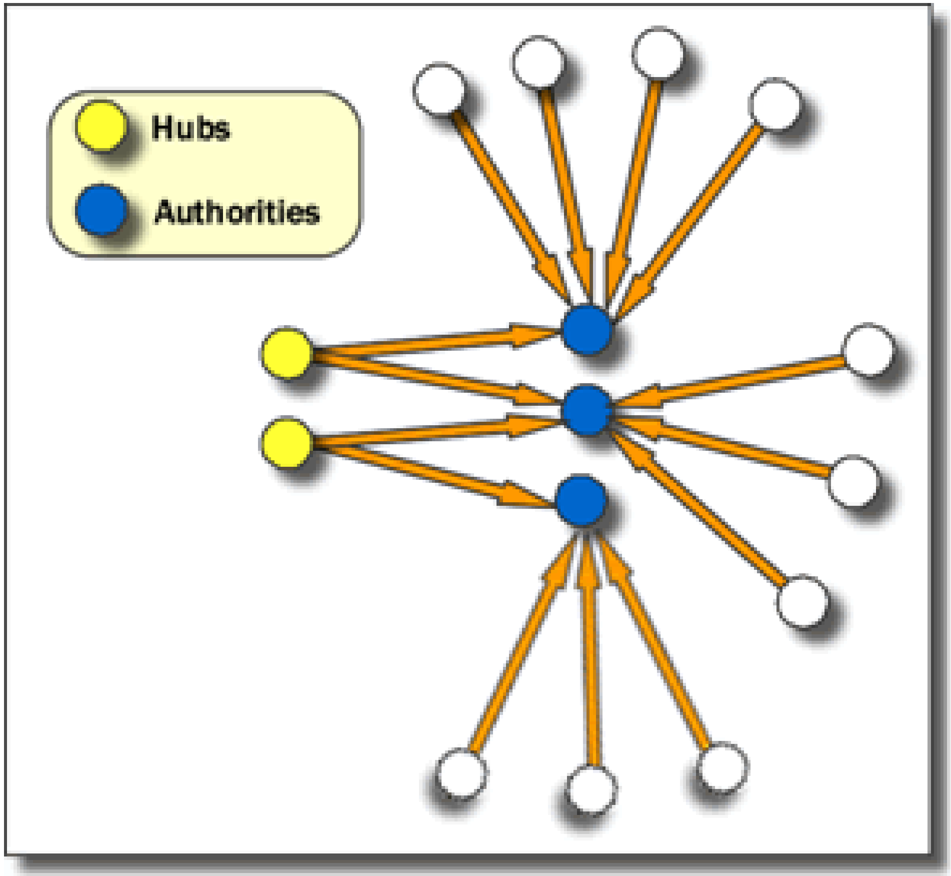
- Eigenvalue measures how close an entity is to other highly close entities within a network. In other words, Eigenvalue identifies the most central entities in terms of the global or overall makeup of the network.
- A high Eigenvalue generally:
  - Indicates an actor that is more central to the main pattern of distances among all entities.
  - Is a reasonable measure of one aspect of centrality in terms of positional advantage.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Eigenvector centrality:

Importance of a node
depends on
the importance of its neighbors

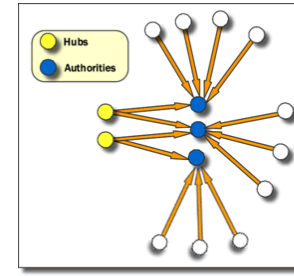
Social Network Analysis: Hub and Authority



Hubs are entities that point to a relatively large number of authorities. They are essentially the mutually reinforcing analogues to authorities. Authorities point to high hubs. Hubs point to high authorities. You cannot have one without the other.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Social Network Analysis: Hub and Authority



- Entities that many other entities point to are called Authorities. In Sentinel Visualizer, relationships are directional—they point from one entity to another.
- If an entity has a high number of relationships pointing to it, it has a high authority value, and generally:
  – Is a knowledge or organizational authority within a domain.
  – Acts as definitive source of information.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Social Network Analysis

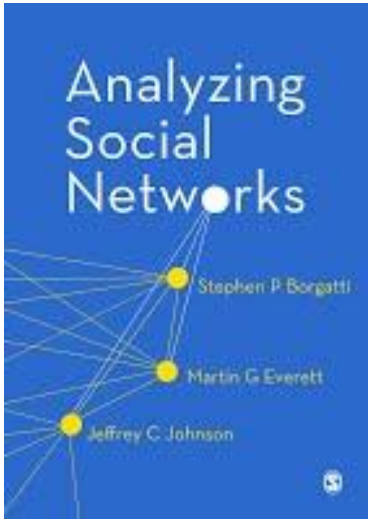
Network Metrics

Calculate | CardView | TableView | Group area | Expand groups | Collapse groups

Name	Type	Degree	Betweenness	Closeness	Eigenvalue	Hub	Authority
Osama bin Laden	Person	44	0.920492092358...	1	0.0271	0	0.011
Abdallah Al-Halabi	Person	2	0	0.654867256637...	0.0001	0	0
Abu Mussab al-Zarqawi	Person	84	0.934887847326...	0.869451697127...	0.7028	0.6572	0.1076
Al Qaeda	Terrorist Organiz...	85	1	0.962427745664...	0.0416	0.3941	0.0166
Ayman Al-Zawahiri	Person	14	0.045794908783...	0.716129032258...	0	0	0.0173
Enaam Arnaout	Person	4	0.031189325814...	0.656804733727...	0.0001	0	0
Imad Eddin Barakat Yarbas	Person	11	0.065049589038...	0.704016913319...	0.0015	0	0.0025
Khalid Shaikh Mohammed	Person	32	0.339916464724...	0.866059817945...	0.002	0	0.1528
Mohamed Atta	Person	61	0.666268740074...	0.820197044334...	0.0015	0	0.6816

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Social Network Analysis (SNA) Tools



- **UCINet**
- **Pajek**

Summary

- Data Mining and Big Data Analytics
- Data Mining Process
- Tasks of Data Mining
- Evaluation of Data Mining
- Social Network Analysis

Source: Han & Kamber (2006)

References

- Jiawei Han and Micheline Kamber (2011),
  Data Mining: Concepts and Techniques, Third Edition, Elsevier
- Jennifer Golbeck (2013),
  Analyzing the Social Web, Morgan Kaufmann
- Stephan Kudyba (2014),
  Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications
- Hiroshi Ishikawa (2015),
  Social Big Data Mining, CRC Press