CH7

Text Mining Concepts: A semi-automated process of extracting knowledge from unstructured data sources

Data Mining VS Text Mining:

- Both seek for novel and useful patterns
- Both are semi-automated processes
- Difference is the nature of the data:
  - Structured versus unstructured data
  - Structured data: in databases
  - Unstructured data: Word documents, PDF files, text excerpts, XML files, and so on

Benefits of text mining: Use in law, academic research, finance, medicine, technology

Text Mining Application Area:

- Information extraction
- Topic tracking
- Summarization
- Categorization
- Clustering
- Concept linking
- Question answering

Natural Language Processing (NLP): a subfield of artificial intelligence and computational linguistics, the studies of "understanding" the natural human language

Challenges in Natural Language Processing (NLP):

- Part-of-speech tagging
- Text segmentation
- Word sense disambiguation
- Syntax ambiguity
- Imperfect or irregular input
- Speech acts

NLP Task Categories:

- Information retrieval, information extraction
- Named-entity recognition
- Question answering
- Automatic summarization
- Speech recognition

Text Mining Applications:

- Marketing applications
- Security applications
- Medicine and biology
- Academic applications

Text Mining Process:

- Step 1: Establish the corpus
- Step 2: Create the Term-by-Document Matrix (TDM)
- Step 3: Extract patterns/knowledge

How can we reduce the dimensionality of the TDM?

- Manual - a domain expert goes through it
- Eliminate terms with very few occurrences in very few documents (?)
- Transform the matrix using singular value decomposition (SVD)
- SVD is similar to principle component analysis

Text Mining Tools:

- Commercial Software Tools (IBM SPSS, SAS, Statistical Data)
- Free Software Tools (RapidMiner, GATE, Spy-EM)

Sentiment Analysis Applications:

- Voice of the customer (VOC)
- Voice of the Market (VOM)
- Voice of the Employee (VOE)
- Brand Management
- Financial Markets

Sentiment Analysis Process:

- Step 1 – Sentiment Detection
- Step 2 – N-P Polarity Classification
- Step 3 – Target Identification
- Step 4 – Collection and Aggregation

Methods for Polarity Identification:

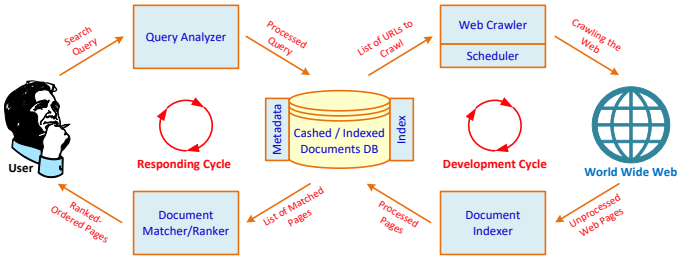
1. Using a lexicon
   - WordNet [wordnet.princeton.edu]
   - SentiWordNet [sentiwordnet.isti.cnr.it]
2. Using pre-classified training documents
   - Data mining / machine learning

CH8

Web mining (or Web data mining) is the process of discovering intrinsic جوهرية relationships from Web data (textual, linkage, or usage)

Search engine is a software program that searches for documents (Internet sites or files) based on the keywords (individual words, multi-word terms, or a complete sentence) that users have provided that have to do with the subject of their inquiry

Structure of a Typical Internet Search Engine:



Anatomy تركيب of a Search Engine:

1. Development Cycle
   - Web Crawler
   - Document Indexer
2. Response Cycle
   - Query Analyzer
   - Document Matcher/Ranker

Search Engine Optimization (SEO): It is the intentional activity of affecting the visibility of an e-commerce site or a Web site in a search engine's natural (unpaid or organic) search results

Methods for Search Engine Optimization:

- Search engine recommended techniques (White-Hat SEO)
- Search engine disapproved مستهجنة/مرفوضة techniques (Black-Hat SEO)

Web Analytics: Extraction of information from data generated through Web page visits and transactions

Web usage mining applications

- Determine the lifetime value of clients
- Design cross-marketing strategies across products
- Evaluate promotional campaigns تقييم الحملات الترويجية
- 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 Analytics Metrics: Provides near-real-time data to deliver invaluable information to:

- Improve site usability
- Manage marketing efforts
- Better document ROI (Rate of return on investment)

Web analytics metric categories:

- Web site usability: How were they using my Web site?
  - Page views, Time on site, Downloads, Click map, Click paths
- Traffic sources: Where did they come from?
  - Referral Web sites, Search engines, Direct, Offline campaigns, Online campaigns
- Visitor profiles: What do my visitors look like?
  - Keywords, Content groupings, Geography, Time of day, Landing page
- Conversion statistics: What does all this mean for the business?
  - New visitors, Returning visitors, Leads, Sales/conversions, Abandonment rates

Web Analytics Maturity Model: degree of proficiency, formality, and optimization of business models

Business Intelligence Maturity Model (TDWI) (6 stages)

- Management Reporting ➔ Spreadmarts ➔ Data Marts ➔ Data Warehouse ➔ Enterprise Data Warehouse ➔ BI Services

Business Analytics Maturity Model (INFORMS) (3 stages)

- Descriptive Analytics ➔ Predictive Analytics ➔ Prescriptive Analytics

Social Network: social structure composed of individuals linked to each other

- Social Networks **help** study relationships between individuals, groups, organizations, societies

Typical social network types: Communication networks, community networks, criminal networks, innovation networks

Social Network Analysis Metrics:

- **Connections**
  - Homophily
  - Multiplexity
  - Network closure
  - Propinquity
- **Segmentation**
  - Cliques and social circles
  - Clustering coefficient
  - Cohesion
- **Distribution**
  - Bridge
  - Centrality
  - Density
  - Structural holes
  - Tie strength

Different Types of Social Media:

1. Collaborative projects (e.g., Wikipedia)
2. Blogs and microblogs (e.g., Twitter)
3. Content communities (e.g., YouTube)
4. Social networking sites (e.g., Facebook)
5. Virtual game worlds (e.g., World of Warcraft), and
6. Virtual social worlds (e.g., Second Life)

Social VS Industrial Media: Web-based social media are different from traditional/industrial media, such as newspapers, television, and film

- **Differentiating characteristics**
  - Quality
  - Reach
  - Frequency
  - Accessibility
  - Usability
  - Immediacy
  - Updatability

CH9

DSS modeling (optimization & simulation) contribute to organizational success.

Major Modeling Issues:

- Problem identification and environmental analysis
- Variable identification
- Forecasting/predicting
- Multiple models
- Model management

Categories of Models:

TABLE 9.1 Categories of Models

Category	Process and Objective	Representative Techniques
Optimization of problems with few alternatives	Find the best solution from a small number of alternatives	Decision tables, decision trees, analytic hierarchy process
Optimization via algorithm	Find the best solution from a large number of alternatives, using a step-by-step improvement process	Linear and other mathematical programming models, network models
Optimization via an analytic formula	Find the best solution in one step, using a formula	Some inventory models
Simulation	Find a good enough solution or the best among the alternatives checked, using experimentation	Several types of simulation
Heuristics	Find a good enough solution, using rules	Heuristic programming, expert systems
Predictive models	Predict the future for a given scenario	Forecasting models, Markov analysis
Other models	Solve a what-if case, using a formula	Financial modeling, waiting lines

Model Categories Static and Dynamic Models:

- Static Analysis
  - Single snapshot of the situation
  - Single interval
  - Steady state
- Dynamic Analysis
  - Dynamic models
  - Evaluate scenarios that change over time
  - Time dependent
  - Represents trends and patterns over time
  - More realistic: Extends static models

Modeling and Decision Making - Under Certainty, Uncertainty, and Risk:

- Certainty
  - Assume complete knowledge
  - All potential outcomes are known
  - May yield optimal solution
- Uncertainty
  - Several outcomes for each decision
  - Probability of each outcome is unknown
  - Knowledge would lead to less uncertainty
- Risk analysis (probabilistic decision making)
  - Probability of each of several outcomes occurring
  - Level of uncertainty => Risk (expected value)

Decision Modeling with Spreadsheets:

- Most popular *end-user modeling tool*
- Flexible and easy to use
- Powerful functions (add-in functions)
- Programmability (via macros)
- What-if analysis and goal seeking
- Simple database management
- Seamless integration of model and data

Optimization via Mathematical Programming:

- **Mathematical Programming:** A family of tools designed to help solve managerial problems in which the decision maker must allocate scarce resources among competing activities to optimize a measurable goal
- **Optimal solution:** The best possible solution to a modeled problem
  - Linear programming (LP): A mathematical model for the optimal solution of resource allocation problems. All the relationships are linear

Linear programming (LP) Problem Characteristics:

1. Limited quantity of economic resources
2. Resources are used in the production of products or services
3. Two or more ways (solutions, programs) to use the resources
4. Each activity (product or service) yields a return in terms of the goal
5. Allocation is usually restricted by constraints

Linear Programming Steps:

1. **Identify the …**
   - Decision variables
   - Objective function
   - Objective function coefficients
   - Constraints (Capacities / Demands /)
2. **Represent the model**
   - LINDO: Write mathematical formulation
   - EXCEL: Input data into specific cells in Excel
3. **Run the model and observe the results**

Multiple Goals, Sensitivity Analysis, What-If Analysis, and Goal Seeking:

- **Multiple Goals**
  - Simple-goal vs. multiple goals
  - Vast majority of managerial problems has multiple goals (objectives) to achieve
  - Methods of handling multiple goals
    - Utility theory
    - Goal programming
    - Expression of goals as constraints, using LP
    - A points system
- **Sensitivity analysis:** It is the process of assessing the impact of change in inputs on outputs
  - Helps to:
    - eliminate (or reduce) variables
    - revise models to eliminate too-large sensitivities
    - adding details about sensitive variables or scenarios
    - obtain better estimates of sensitive variables
    - alter a real-world system to reduce sensitivities
  - Can be automatic or trial and error
- **What-if analysis:** Assesses solutions based on changes in variables or assumptions (scenario analysis)
- **Goal seeking:** Backwards approach, starts with the goal and determines values of inputs needed

Decision Trees:

- Graphical representation of relationships
- Multiple criteria approach
- Demonstrates complex relationships
- Cumbersome, if many alternatives exists
- **Tools include**
  - Mind Tools Ltd., mindtools.com
  - TreeAge Software Inc., treeage.com
  - Palisade Corp., palisade.com

CH10

Search: is the process of identifying the best possible solution / course of action [under limitations such as time, …]

Search techniques in choice phase include:

- analytical techniques,
- algorithms,
- blind searching, and
- heuristic searching

Traveling Salesman Problem: A traveling salesman must visit customers in several cities, visiting each city only once, across the country. Goal: Find the shortest possible route.

When to Use Heuristics?

- Inexact or limited input data
- Complex reality
- Reliable, exact algorithm not available
- Computation time excessive
- For making quick decisions

Limitations of Heuristics: Cannot guarantee an optimal solution

Modern Heuristic Methods:

- **Tabu search:** Intelligent search algorithm
- **Genetic algorithms:** Survival of the fittest
- **Simulated annealing:** Analogy to Thermodynamics
- **Ant colony and other Meta-heuristics**

Genetic Algorithms: Moving toward better and better solutions by letting only the fittest parents create the future generations

- It is a popular **heuristic search** technique
- **Main theme:** Survival of the fittest

Genetic Algorithms- Example:

- The Vector Game
- The Knapsack Problem

Limitations of Genetic Algorithms

- Does not guarantee an optimal solution (often settles in a sub optimal solution / local minimum)
- Not all problems can be put into GA formulation
- Development and interpretation of GA solutions requires both programming and statistical skills
- Relies heavily on the random number generators
- Locating good variables for a particular problem and obtaining the data for the variables is difficult
- Selecting methods by which to evolve the system requires experimentation and experience

Genetic Algorithm Applications:

- Dynamic process control
- Optimization of induction rules
- Discovery of new connectivity topologies (NNs)
- Simulation of biological models of behavior
- Complex design of engineering structures
- Pattern recognition

Simulation:

- Simulation is the “appearance” of reality
- It is often used to conduct what-if analysis on the model of the actual system
- It is a popular DSS technique for conducting experiments with a computer on a comprehensive model of the system to assess its dynamic behavior
- Often used when the system is too complex for other DSS techniques

Major Characteristics of Simulation:

- Imitates reality and captures its richness both in shape and behavior
  - “Represent” versus “Imitate”
- Technique for conducting experiments
- Descriptive, not normative tool
- Often to “solve” [i.e., analyze] very complex systems/problems
- Simulation should be used only when a numerical optimization is not possible

Advantages of Simulation:

- The theory is fairly straightforward
- Great deal of time compression
- Experiment with different alternatives
- The model reflects manager's perspective
- Can handle wide variety of problem types
- Can include the real complexities of problems
- Produces important performance measures
- Often it is the only DSS modeling tool for non-structured problems

Disadvantages of Simulation:

- Cannot guarantee an optimal solution
- Slow and costly construction process
- Cannot transfer solutions and inferences to solve other problems (problem specific)
- So easy to explain/sell to managers, may lead to overlooking analytical solutions
- Software may require special skills

Simulation Methodology (Steps):

1. Define problem
2. Construct the model
3. Test and validate model
4. Design experiments
5. Conduct experiments
6. Evaluate results
7. Implement solution

Simulation Types:

- Probabilistic/Stochastic vs. Deterministic Simulation
- Time-dependent vs. Time-independent Simulation
- Discrete Event vs. Continuous Simulation
- Simulation Implementation

Simulation Software:

- Simio LLC, simio.com
- SAS Simulation, sas.com
- Lumina Decision Systems, lumina.com
- Oracle Crystal Ball, oracle.com
- Palisade Corp., palisade.com
- Rockwell Intl., arenasimulation.com

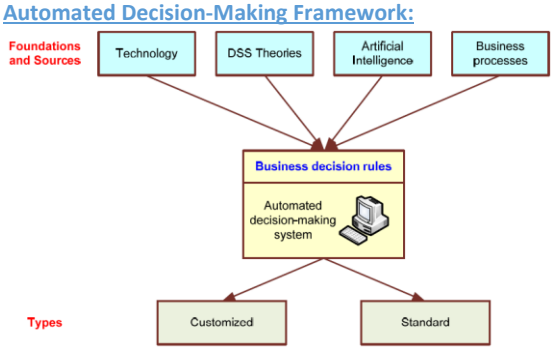
Agent: an autonomous computer program that observes and acts on an environment and directs its activity toward achieving specific goals

Agent-based modeling (ABM): is a simulation modeling technique to support complex decision systems where a system is modeled as a set of autonomous decision-making units called *agents*

CH11

Decision Automation Systems (DAS): Often a rule-based system that provides a solution in a functional area

Automated Decision-Making Framework:



Artificial intelligence (AI): A subfield of computer science, concerned with symbolic reasoning and problem solving.

Artificial intelligence (AI) Objectives:

- **Make machines smarter (primary goal)**
- **Understand what intelligence is**
- **Make machines more intelligent & useful**

Expert Systems (ES): Is a computer program that attempts to imitate expert's reasoning processes and knowledge in solving specific problems

- **Most Popular Applied AI Technology**
  - **Enhance Productivity**
  - **Augment Work Forces**

Expert: A human being who has developed a high level of proficiency in making judgments in a specific domain.

Expertise: The set of capabilities that underlines the performance of human experts.

Features and Concepts in ES:

- **Experts / Expertise**
  - **Degrees or levels of expertise**
  - **Ratio of non-experts to experts → 100 to 1**
- **Transferring Expertise**
  - **From expert to computer to nonexperts via acquisition, representation, inferencing, transfer**
- **Symbolic Reasoning / Inferencing**
- **Deep Knowledge / Self Knowledge**

Conventional vs. Expert Systems:

Execution is done on a step-by-step (algorithmic) basis.	Execution is done by using heuristics and logic.
Large databases can be effectively manipulated.	Large knowledge bases can be effectively manipulated.
Conventional systems represent and use data.	ES represent and use knowledge.
Efficiency is usually a major goal.	
Effectiveness is important only for DSS.	Effectiveness is the major goal.
Conventional systems easily deal with quantitative data.	ES easily deal with qualitative data.
Conventional systems use numeric data representations.	ES use symbolic and numeric knowledge representations.
Conventional systems capture, magnify, and distribute access to numeric data or information.	ES capture, magnify, and distribute access to judgment and knowledge.

Structure of Expert Systems:

- Development Environment
- Consultation Environment
- **Major Components**
  - Knowledge acquisition subsystem (Knowledge Engineer)
  - Knowledge Base
  - Inference Engine
  - User Interface
  - Blackboard (workplace)
  - Explanation subsystem (justifier)
  - Knowledge-refining system

Knowledge Engineering (KE): A set of intensive activities encompassing the acquisition of knowledge from human experts (and other information sources) and converting this knowledge into a repository (commonly called a knowledge base)

The primary goal of KE is to

- help experts articulate *how they do what they do,* and
- to document this knowledge in a reusable form

The Knowledge Engineering Process:

- Knowledge Acquisition
- Knowledge Representation
- Knowledge Validation
- Inferencing (Reasoning)
- Explanation Justification

Difficulties in Knowledge Acquisition:

- Experts may not know how to articulate their knowledge or may be unable to do so.
- Experts may lack time or may be unwilling to cooperate.
- Testing and refining knowledge are complicated.
- Methods for knowledge elicitation may be poorly defined.
- System builders tend to collect knowledge from one source, but the relevant knowledge may be scattered across several sources.
- System builders may attempt to collect documented knowledge rather than use experts. The knowledge collected may be incomplete.
- It is difficult to recognize specific knowledge when it is mixed up with irrelevant data.
- Experts may change their behavior when they are observed or interviewed.
- Problematic interpersonal communication factors may affect the knowledge engineer and the expert.

Validation VS Verification:

- **Validation** is the part of evaluation that deals with the performance of the system
- **Verification** is building the system right or substantiating اثبات that the system is correctly implemented to its specifications

Knowledge Representation in ES: The most common/popular way to represent human knowledge is **Production rules**

Forms of Production Rules:

- IF premise, THEN conclusion
- Conclusion, IF premise
- Inclusion of ELSE

Knowledge and Inference Rules:

- Knowledge rules (declarative rules), state all the facts and relationships about a problem
- Inference rules (procedural rules), advise on how to solve a problem, given that certain facts are known

Inference: is the process of chaining multiple rules together based on available data

Forward chaining VS Backward chaining

- **Forward chaining:** A data-driven search in a rule-based system.
  - If the premise clauses match the situation, then the process attempts to assert the conclusion.
- **Backward chaining:** A goal-driven search in a rule-based system.
  - It begins with the action clause of a rule and works backward through a chain of rules in an attempt to find a verifiable set of condition clauses.

Firing a rule:

- When all of the rule's hypotheses (the "if parts") are satisfied, a rule said to be FIRED
- Inference engine checks every rule in the knowledge base in a forward or backward direction to find rules that can be FIRED
- Continues until no more rules can fire, or until a goal is achieved

How to handle conflicting rules:

1. Establish a goal and stop firing rules when goal is achieved
2. Fire the rule with the highest priority
3. Fire the most specific rule
4. Fire the rule that uses the data most recently entered

How do we choose between Backward chaining and Forward chaining:

- Follow how a domain expert solves the problem
- If the expert **first collect data**, then infer from it **((=> Forward Chaining))**
- If the expert starts with a **hypothetical solution** and then attempts to find facts to prove it **((=> Backward Chaining))**

Development of ES:

- Defining the nature and scope of the problem
- Identifying proper experts
- Acquiring knowledge
  - Knowledge engineer
- Selecting the Building Tools
  - Shells versus Complete Development
- Coding the system
- Evaluating and Launching the System

CH12

Knowledge management: The active management of the expertise in an organization. It involves collecting, categorizing, and disseminating knowledge.

Intellectual capital: The invaluable knowledge of an organization's employees.

Knowledge: is information that is contextual, relevant, and actionable.

- understanding, awareness, or familiarity acquired through education or experience
- anything that has been learned, perceived, discovered, inferred, or understood.
- knowledge is information in action.

Characteristics of knowledge

- Extraordinary leverage and increasing returns
- Fragmentation, leakage, and the need to refresh
- Uncertain value
- Uncertain value of sharing

Knowledge-based economy: The economic shift from natural resources to intellectual assets.

Explicit and tacit knowledge:

- **Explicit (leaky) knowledge:** Knowledge that deals with objective, rational, and technical material (data, policies, procedures, software, documents, etc.).
- **Tacit (embedded) knowledge:** Knowledge that is usually in the domain of subjective, cognitive ادراكية, and experiential learning.

Learning organization: An organization capable of learning from its past experience, implying the existence of an organizational memory and a means to save, represent, and share it through its personnel.

Organizational memory: Repository of what the organization knows.

Organizational culture: The aggregate attitudes in an organization concerning a certain issue (e.g., technology, computers, DSS)

Approaches to Knowledge Management:

- **Process approach:** to knowledge management attempts to codify organizational knowledge through formalized controls, processes and technologies (Focuses on explicit knowledge and IT)
- **Practice approach:** focuses on building the social environments or communities of practice necessary to facilitate the sharing of tacit understanding (Focuses on tacit knowledge and socialization).
- **Hybrid approaches:** Hybrid between Process approach and Practice approach.

Knowledge repository: is the actual storage location of knowledge in a knowledge management system.

Knowledge Management System (KMS) cycle (Step):

1. Create knowledge	2. Capture knowledge	3. Refine knowledge
4. Store knowledge	5. Manage knowledge	6. Disseminate knowledge

Components of Knowledge Management System (KMS):

1. Communication
2. Collaboration
3. Storage and retrieval

Technologies that support Knowledge Management System (KMS):

- Artificial intelligence
- Intelligent agents
- Knowledge discovery in databases
- Web 2.0

Groupwork: the work done by two or more people together

Characteristics of Groupwork:

- A group performs a task
- Members may be located in different places
- Group members may work at different times
- Group members may work for the same organization or for different organizations
- A group can be permanent or temporary
- A group can be at one managerial level or span several levels

Groupwork Process Gains and Losses:

Benefits of Working in Groups (Process Gains)	Dysfunctions of the Group Process (Process Losses)
• It provides learning. Groups are better than individuals at understanding problems.	• Social pressures of conformity may result in **groupthink** (i.e., people begin to think alike and do not tolerate new ideas; they yield to *conformance pressure*).
• People readily take ownership of problems and their solutions. They take responsibility.	• It is a time-consuming, slow process (i.e., only one member can speak at a time).
• Group members have their egos embedded in the decision, so they are committed to the solution.	• There can be lack of coordination of the meeting and poor meeting planning.
• Groups are better than individuals at catching errors.	• Inappropriate influences (e.g., domination of time, topic, or opinion by one or few individuals; fear of contributing because of the possibility of *flaming*).
• A group has more *information* (i.e., knowledge) than any one member. Group members can combine their	• There can be a tendency for group members to either dominate the agenda or rely on others to do most of the work (free-riding).

Traditional methods for Supporting Groupwork:

Nominal Group Technique: Individuals work alone to generate ideas which are pooled under guidance of a trained facilitator

Delphi Method: A structured process for collecting and distilling تقطير knowledge from a group of experts by means of questionnaires.

Groupware products: provide a way for groups to share resources and opinions.

- Synchronous or Asynchronous
- **Examples:** dropbox.com, drive.google.com, office.microsoft.com

Group Decision Support Systems Pros and Cons:

- **Gains:** Parallelism, Anonymity, Triggering, Synergy, Structure, Record keeping
- **Loses:** Free-riding, Flaming

Facilities for Group Decision Support Systems:

- Decision room
- Multiple-use facility
- Web based

CH13

The Vs that define Big Data: Volume, Variety, Velocity, Veracity, Variability, Value

Critical Success Factors for Big Data Analytics:

- A clear business need (alignment with the vision and the strategy)
- Strong, committed sponsorship (executive champion)
- Alignment between the business and IT strategy
- A fact-based decision-making culture
- A strong data infrastructure
- The right analytics tools
- Right people with right skills

Enablers of Big Data Analytics:

- **In-memory analytics:** Storing and processing the complete data set in RAM
- **In-database analytics:** Placing analytic procedures close to where data is stored
- **Grid computing & MPP:** Use of many machines and processors in parallel (MPP- massively parallel processing)
- **Appliances:** Combining hardware, software and storage in a single unit for performance and scalability

Challenges of Big Data Analytics:

- **Data volume:** The ability to capture, store, and process the huge volume of data in a timely manner
- **Data integration:** The ability to combine data quickly/cost effectively
- **Processing capabilities:** The ability to process the data quickly, as it is captured (i.e., stream analytics)
- **Data governance** (… security, privacy, access)
- **Skill availability** (… data scientist)
- **Solution cost** (ROI)

Business Problems Addressed by Big Data Analytics:

- Process efficiency and cost reduction
- Brand management
- Revenue maximization
- Improved customer service
- Risk management

Big Data Technologies:

- **MapReduce:** distributes the processing of very large multi-structured data files across a large cluster of ordinary machines/processors.
  - **Goal** - achieving high performance with "simple" computers
- **Hadoop:** is an open source framework for storing and analyzing massive amounts of distributed, unstructured data
  - **MapReduce + Hadoop =** Big Data core technology
- **Hive**
- **Pig**

How Does Hadoop Work?

- Access unstructured and semi-structured data (e.g., log files, social media feeds, other data sources)
- Break the data up into "parts," which are then loaded into a file system made up of multiple nodes running on commodity hardware using HDFS
- Each "part" is replicated multiple times and loaded into the file system for replication and failsafe processing
- A node acts as the Facilitator and another as Job Tracker
- Jobs are distributed to the clients, and once completed the results are collected and aggregated using MapReduce

Hadoop Technical Components:

- Hadoop Distributed File System (HDFS)
- Name Node (primary facilitator)
- Secondary Node (backup to Name Node)
- Job Tracker
- Slave Nodes (the grunts of any Hadoop cluster)
- Additionally, Hadoop ecosystem is made-up of a number of complementary sub-projects: NoSQL

What is the impact of Big Data on DW? Big Data and RDBMS do not go nicely together

How to Succeed with Big Data?

Simplify	**Coexist**	**Visualize**
Empower	**Integrate**	**Govern**

Why Stream Analytics?

- It may not be feasible to store the data
- It may lose its value if not processed immediately

Stream Analytics Applications:

- e-Commerce
- Telecommunication
- Law Enforcement and Cyber Security
- Power Industry
- Financial Services
- Health Services
- Government

- **CH14**

Geocoding

- Visual maps
- Postal codes
- Latitude & Longitude

Geographic Information System (GIS): Used to capture, store, analyze, and manage the data linked to a location

- Combined with integrated sensor technologies and global positioning systems (GPS)

Location Intelligence (LI): Interactive maps that further drill down to details about any location

Use of Location-Based Analytics:

- Retailers
- Global Intelligence

Geospatial Analytics Examples:

- Sabre Airline Solutions' application
  - Traveler Security
  - Geospatial-enabled dashboard
  - Assess risks across global hotspots
  - Interactive maps
- Telecommunication companies
  - Analysis of failed connections

Real-Time Location Intelligence: Targeting right customer based on their behavior over geographic locations

- Example: Radii app, Cachetown - augmented reality-based game

Two main approaches for recommendation systems:

1. Collaborative filtering:
   - Based on previous users' purchase/view/rating data
   - Collectively deriving **user ⇔ item** profiling
   - Use this knowledge for item recommendations
   - Techniques include user-item rating matrix, kNN, correlation, …
   - Disadvantage – requires huge amount of historic data
2. Content filtering:
   - Based on specifications/characteristics of items (not just ratings)
   - First, characteristics of an item are profiled, and then the content-based individual user profiles are built
   - Recommendations are made if there are similarities found in the item characteristics
   - Techniques include decision trees, ANN, Bayesian classifiers

A social network: is a place where people create their own space, or homepage, on which they write blogs

Using Twitter to Get a Pulse of the Market

- Listening to the public for opinions/sentiments
- Product/service brand management
- Text mining, sentiment analysis
- How – built in-house or outsource

Cloud Computing and BI: A style of computing in which dynamically scalable and often virtualized resources are provided over the Internet.

Major Components of Service-Oriented DSS/BI:

- Data-as-a-Service (DaaS)
- Information-as-a-Service (IaaS)
- Analytics-as-a-Service (AaaS)

Major Components of Service-Oriented DSS/BI:

TABLE 14-1 Major Components of Service-Oriented DSS

	Component	Brief Description
Data sources	Application programming interface	Mechanism to populate source systems with raw data and to pull operational reports.
Data sources	Operational transaction systems	Systems that run day-to-day business operations and provide source data for the data warehouse and DSS environment.
Data sources	Enterprise application integration/staging area	Provides an integrated common data interface and interchange mechanism for real-time and source systems.
Data management	Extract, transform, load (ETL)	The processes to extract, transform, cleanse, reengineer, and load source data into the data warehouse, and move data from one location to another.
Data services	Metadata management	Data that describes the meaning and structure of business data, as well as how it is created, accessed, and used.
Data services	Data warehouse	Subject-oriented, integrated, time-variant, and nonvolatile collection of summary and detailed data used to support the strategic decision-making process for the organization. This is also used for ad hoc and exploratory processing of very large data sets.
Data services	Data marts	Subset of data warehouse to support specific decision and analytical needs and provide business units more flexibility, control, and responsibility.

تم بحمد الله..... دعواتكم