Chapter 5

Predictive Analytics II: Text, Web, and Social Media Analytics

Learning Objectives (1 of 2)

5.1 Describe text mining and understand the need for text mining

5.2 Differentiate among text analytics, text mining, and data mining

5.3 Understand the different application areas for text mining

5.4 Know the process of carrying out a text mining project

5.5 Appreciate the different methods to introduce structure to text-based data

Learning Objectives (2 of 2)

5.6 Describe sentiment analysis

5.7 Develop familiarity with popular applications of sentiment analysis

5.8 Learn the common methods for sentiment analysis

5.9 Become familiar with speech analytics as it relates to sentiment analysis

Data Mining vs Text Mining

Text Analytics and Text Mining (1 of 2)

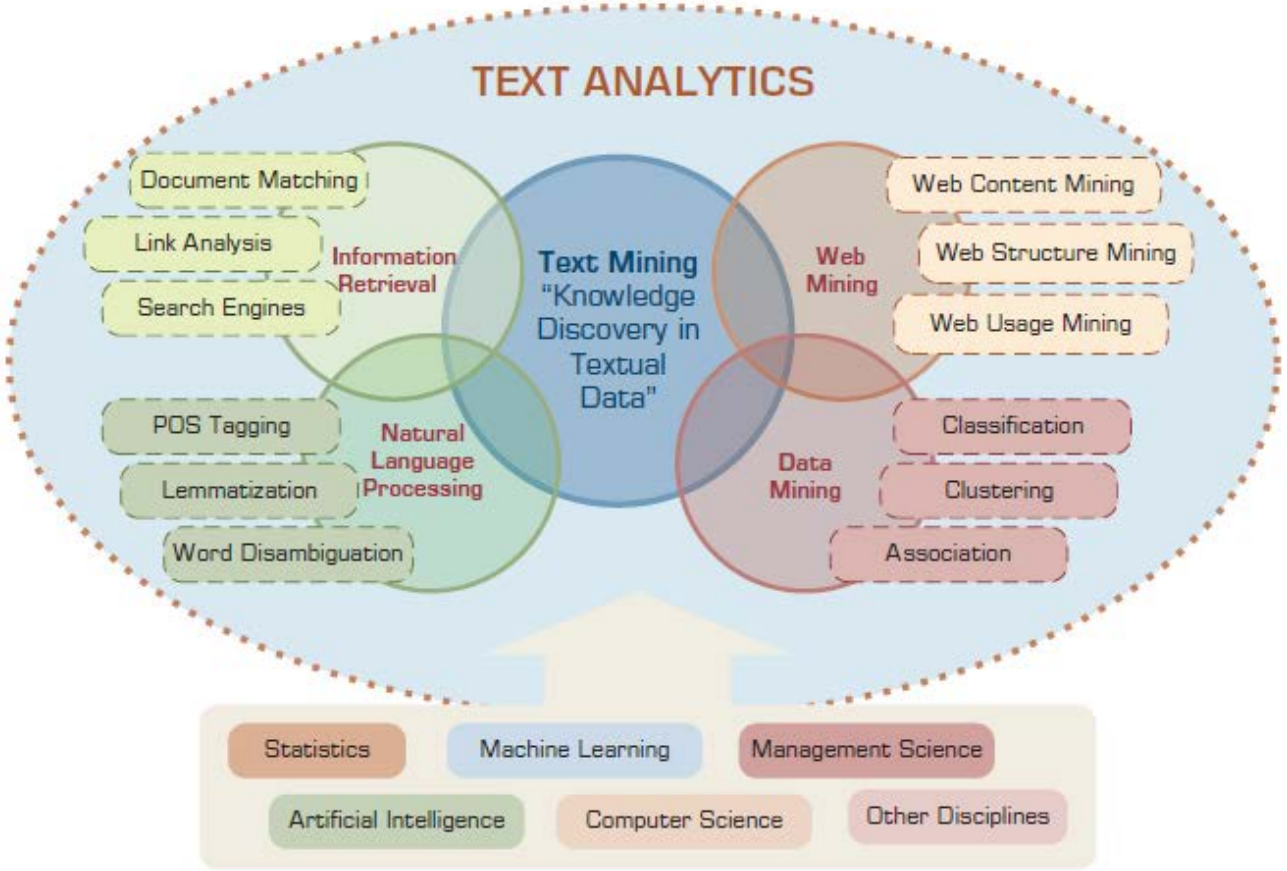
- Text Analytics versus Text Mining
- Text Analytics =
  - Information Retrieval +
  - Information Extraction +
  - Data Mining +
  - Web Mining

or simply

Text Analytics = Information Retrieval + Text Mining

Text Analytics and Text Mining (2 of 2)

- **Figure 5.2** Text Analytics, Related Application Areas, and Enabling Disciplines



Text Mining Concepts (1 of 2)

- 85-90 percent of all corporate data is in some kind of unstructured form (e.g., text)
- Unstructured corporate data is doubling in size every 18 months
- Tapping into these information sources is not an option, but a need to stay competitive
- Answer: text mining
  - A semi-automated process of extracting knowledge from unstructured data sources
  - a.k.a. text data mining or knowledge discovery in textual databases

Data Mining Versus 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
- **To perform text mining** – first, impose structure to the data, then mine the structured data

Text Mining Concepts (2 of 2)

- Benefits of text mining are obvious especially in text-rich data environments
  - e.g., law (court orders), academic research (research articles), finance (quarterly reports), medicine (discharge summaries), biology (molecular interactions), technology (patent files), marketing (customer comments), etc.
- Electronic communization records (e.g., e-mail)
  - Spam filtering
  - E-mail prioritization and categorization
  - Automatic response generation

Text Mining Application Area

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

Text Mining Terminology (1 of 2)

- Unstructured or semistructured data
- Corpus (and corpora)
- Terms
- Concepts
- Stemming
- Stop words (and include words)
- Synonyms (and polysemes)
- Tokenizing

Text Mining Terminology (2 of 2)

- Term dictionary
- Word frequency
- Part-of-speech tagging
- Morphology
- Term-by-document matrix
  - Occurrence matrix
- Singular value decomposition
  - Latent semantic indexing

Application Case 5.1

Insurance Group Strengthens Risk Management with Text Mining Solution

Questions for Discussion

1. How can text analytics and mining be used to keep up with changing business needs of insurance companies?
2. What were the challenges, the proposed solution, and the obtained results?
3. Can you think of other uses of text analytics and text mining for insurance companies?

Natural Language Processing (NLP) (1 of 4)

- Structuring a collection of text
  - **Old approach:** bag-of-words
  - **New approach:** natural language processing
- NLP is …
  - a very important concept in text mining
  - a subfield of artificial intelligence and computational linguistics
  - the studies of "understanding" the natural human language
- Syntax versus semantics-based text mining

Natural Language Processing (NLP) (2 of 4)

- What is “Understanding”?
  - Human understands, what about computers?
  - Natural language is vague, context driven
  - True understanding requires extensive knowledge of a topic
  - **Can/will computers ever understand natural language the same/accurate way we do?**

Natural Language Processing (NLP) (3 of 4)

- Challenges in NLP
  - Part-of-speech tagging
  - Text segmentation
  - Word sense disambiguation
  - Syntax ambiguity
  - Imperfect or irregular input
  - Speech acts
- Dream of AI community
  - to have algorithms that are capable of automatically reading and obtaining knowledge from text

Natural Language Processing (NLP) (4 of 4)

- WordNet
  - A laboriously hand-coded database of English words, their definitions, sets of synonyms, and various semantic relations between synonym sets
  - A major resource for NLP
  - Need automation to be completed
- Sentiment Analysis
  - A technique used to detect favorable and unfavorable opinions toward specific products and services
  - SentiWordNet

Application Case 5.2 (1 of 2)

AMC Networks Is Using Analytics to Capture New Viewers, Predict Ratings, and Add Value for Advertisers in a Multichannel World

A Web-Based Dashboard Used by AMC Networks

[**Source:** AMC Networks]

Application Case 5.2 (2 of 2)

Questions for Discussion

1. What are the common challenges broadcasting companies are facing nowadays? How can analytics help to alleviate these challenges?
2. How did AMC leverage analytics to enhance their business performance?
3. What were the types of text analytics and text mining solutions developed by AMC networks? Can you think of other potential uses of text mining applications in the broadcasting industry?

NLP Task Categories

- Question answering
- Automatic summarization
- Natural language generation & understanding
- Machine translation
- Foreign language reading & writing
- Speech recognition
- Text proofing, optical character recognition
- Optical character recognition

Text Mining Applications

- Marketing applications
  - Enables better CRM
- Security applications
  - ECHELON, OASIS
  - Deception detection (…)
- Medicine and biology
  - Literature-based gene identification (…)
- Academic applications
  - Research stream analysis

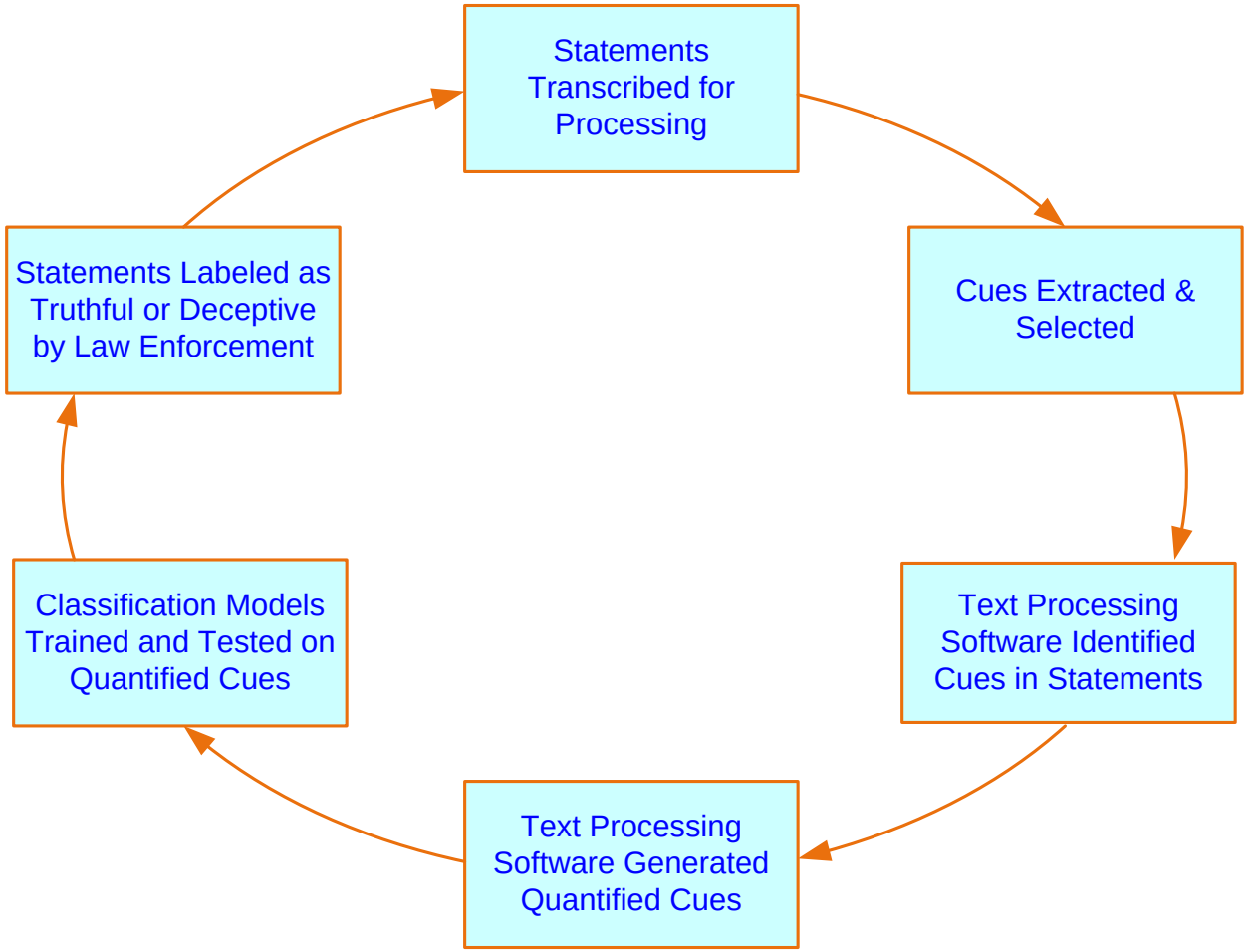
Application Case 5.3 (1 of 4)

Mining for Lies

- Deception detection
  - A difficult problem
  - If detection is limited to only text, then the problem is even more difficult
- The study
  - analyzed text-based testimonies of person of interests at military bases
  - used only text-based features (cues)

Application Case 5.3 (2 of 4)

- **Figure 5.3** Text-Based Deception-Detection Process



Application Case 5.3 (3 of 4)

- **Table 5.1** Categories and Examples of Linguistic Features Used in Deception Detection

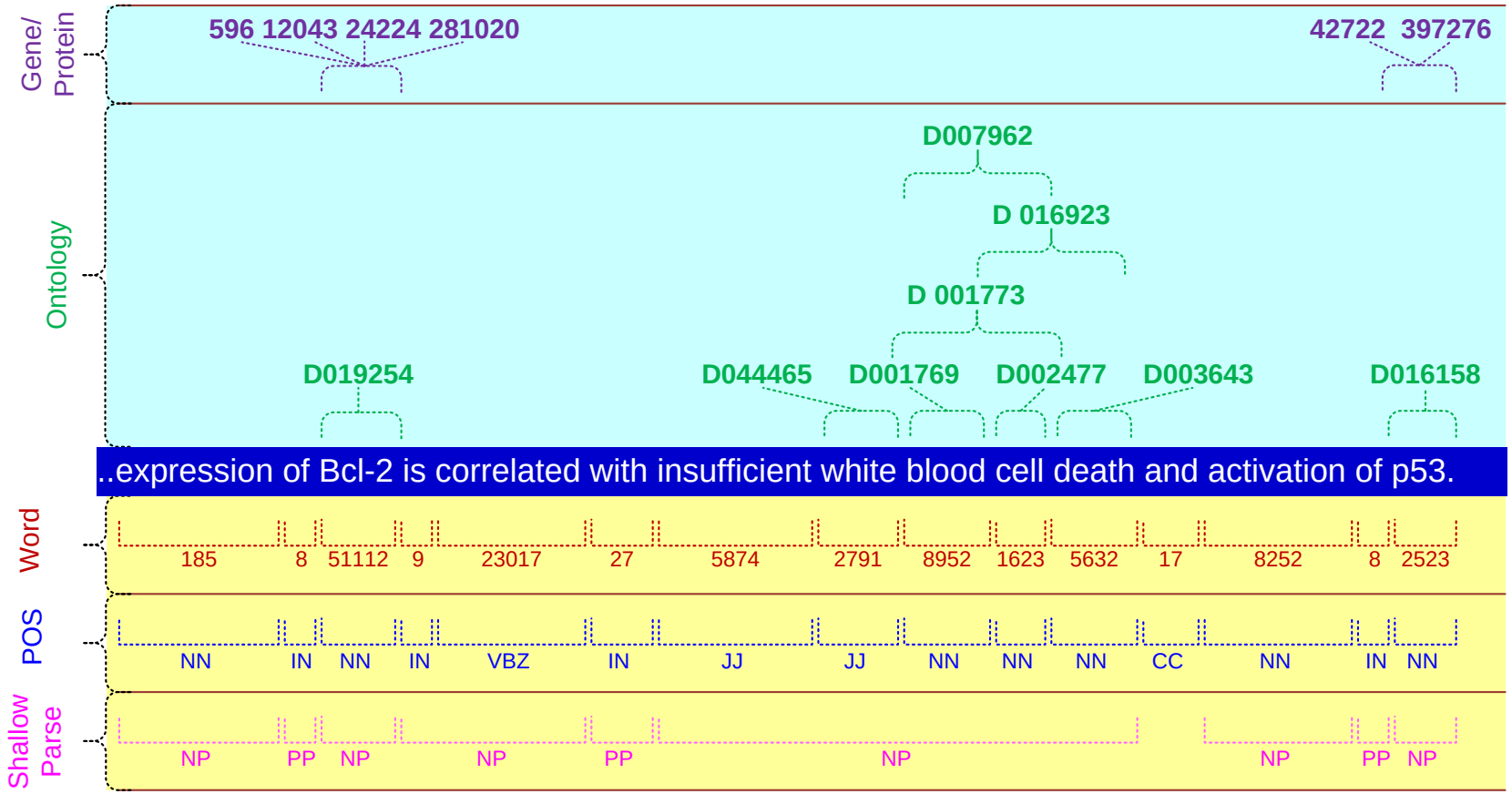
Category	Example Cues
Quantity	Verb count, noun-phrase count, etc.
Complexity	Average number of clauses, average sentence length, etc.
Uncertainty	Modifiers, modal verbs, etc.
Nonimmediacy	Passive voice, objectification, etc.
Expressivity	Emotiveness
Diversity	Lexical diversity, redundancy, etc.
Informality	Typographical error ratio
Specificity	Spatiotemporal information, perceptual information, etc.
Affect	Positive affect, negative affect, etc.

Application Case 5.3 (4 of 4)

- 371 usable statements are generated
- 31 features are used
- Different feature selection methods used
- 10-fold cross validation is used
- Results (overall % accuracy)

Logistic regression	67.28
Decision trees	71.60
Neural networks	73.46

Text Mining Applications (Gene/Protein Interaction Identification)



Application Case 5.4

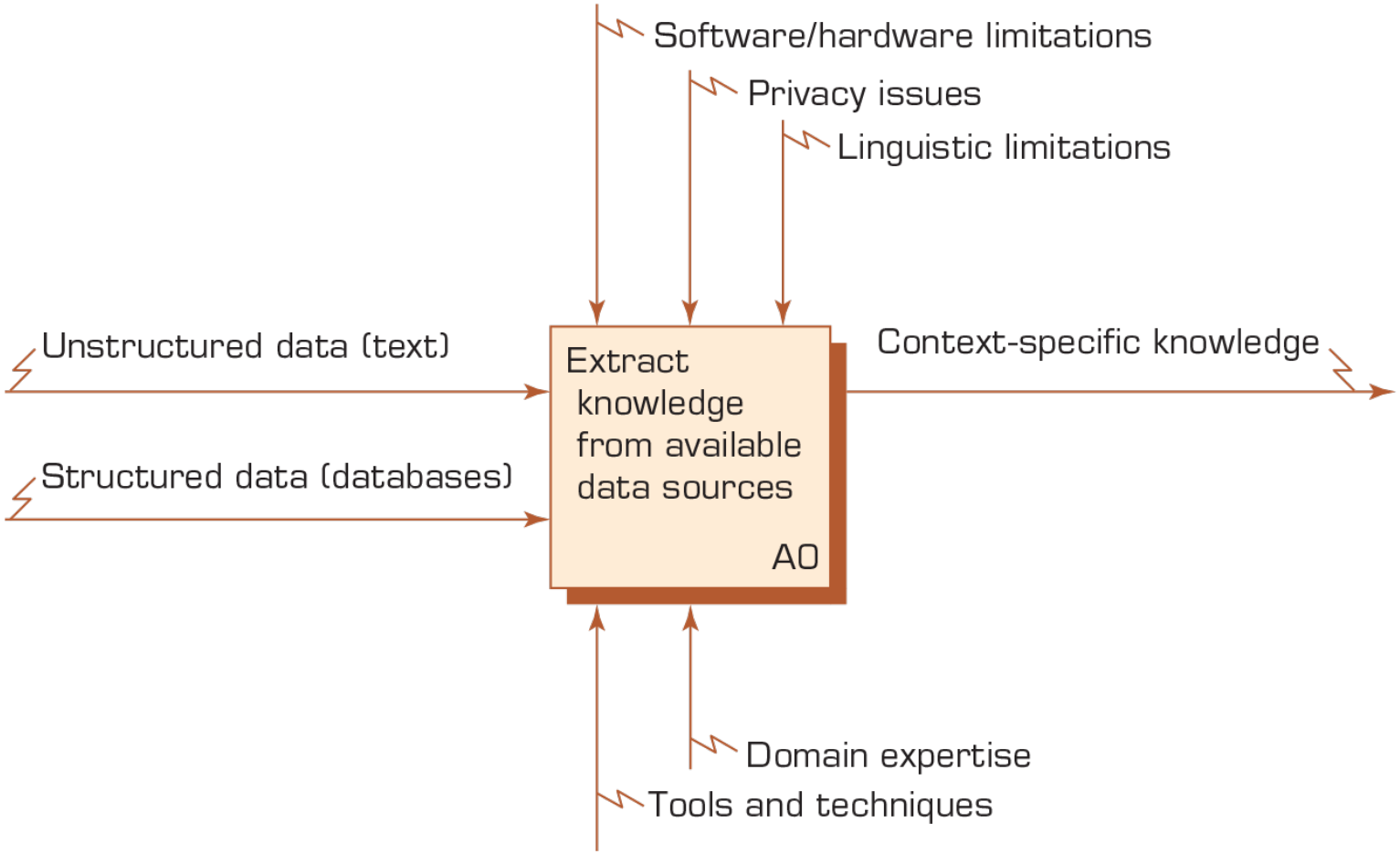
Bringing the Customer into the Quality Equation: Lenovo Uses Analytics to Rethink Its Redesign

Questions for Discussion

1. How did Lenovo use text analytics and text mining to improve quality and design of their products and ultimately improve customer satisfaction?
2. What were the challenges, the proposed solution, and the obtained results?

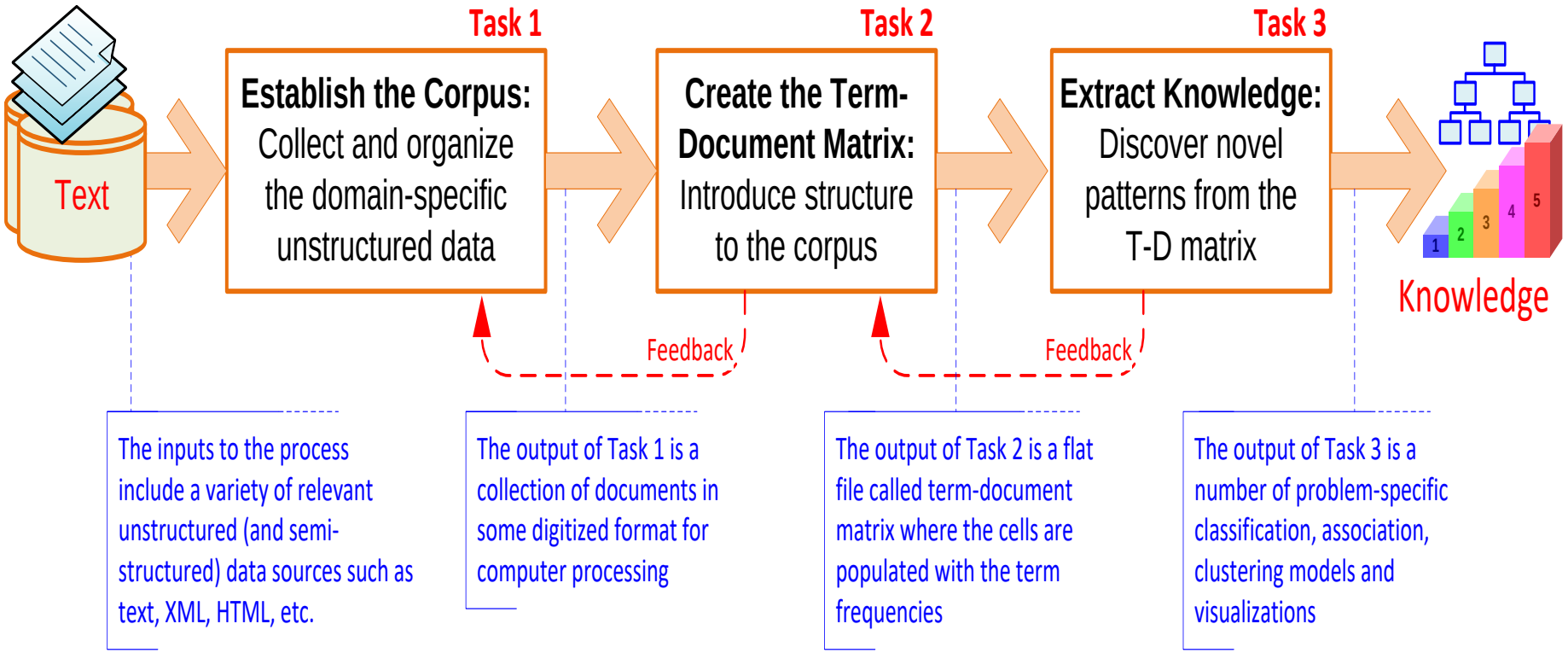
Text Mining Process (1 of 7)

- A Context Diagram for Text Mining Process



Text Mining Process (2 of 7)

- **Figure 5.6** The Three-Step/Task Text Mining Process



Text Mining Process (3 of 7)

- **Step 1:** Establish the corpus
  - Collect all relevant unstructured data (e.g., textual documents, XML files, e-mails, Web pages, short notes, voice recordings…)
  - Digitize, standardize the collection (e.g., all in ASCII text files)
  - Place the collection in a common place (e.g., in a flat file, or in a directory as separate files)

Text Mining Process (4 of 7)

- **Step 2:** Create the Term–by–Document Matrix

Terms / Documents	investment risk	project management	software engineering	development	SAP	...	
Document 1	1			1			
Document 2		1					
Document 3			3		1		
Document 4		1					
Document 5			2	1			
Document 6	1			1			
...							

Text Mining Process (5 of 7)

- Should all terms be included?
  - Stop words, include words
  - Synonyms, homonyms
  - Stemming
- What is the best representation of the indices (values in cells)?
  - Row counts; binary frequencies; log frequencies;
  - Inverse document frequency

Text Mining Process (6 of 7)

- TDM is a sparse matrix. 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 Process (7 of 7)

- **Step 3:** Extract patterns/knowledge
  - Classification (text categorization)
  - Clustering (natural groupings of text)
    - Improve search recall
    - Improve search precision
    - Scatter/gather
    - Query-specific clustering
  - Association
  - Trend Analysis (…)

Application Case 5.5 (1 of 5)

Research Literature Survey with Text Mining

- Mining the published IS literature
  - MIS Quarterly (MISQ)
  - Journal of MIS (JMIS)
  - Information Systems Research (ISR)
  - Covers 12-year period (1994-2005)
  - 901 papers are included in the study
  - Only the paper abstracts are used
  - 9 clusters are generated for further analysis

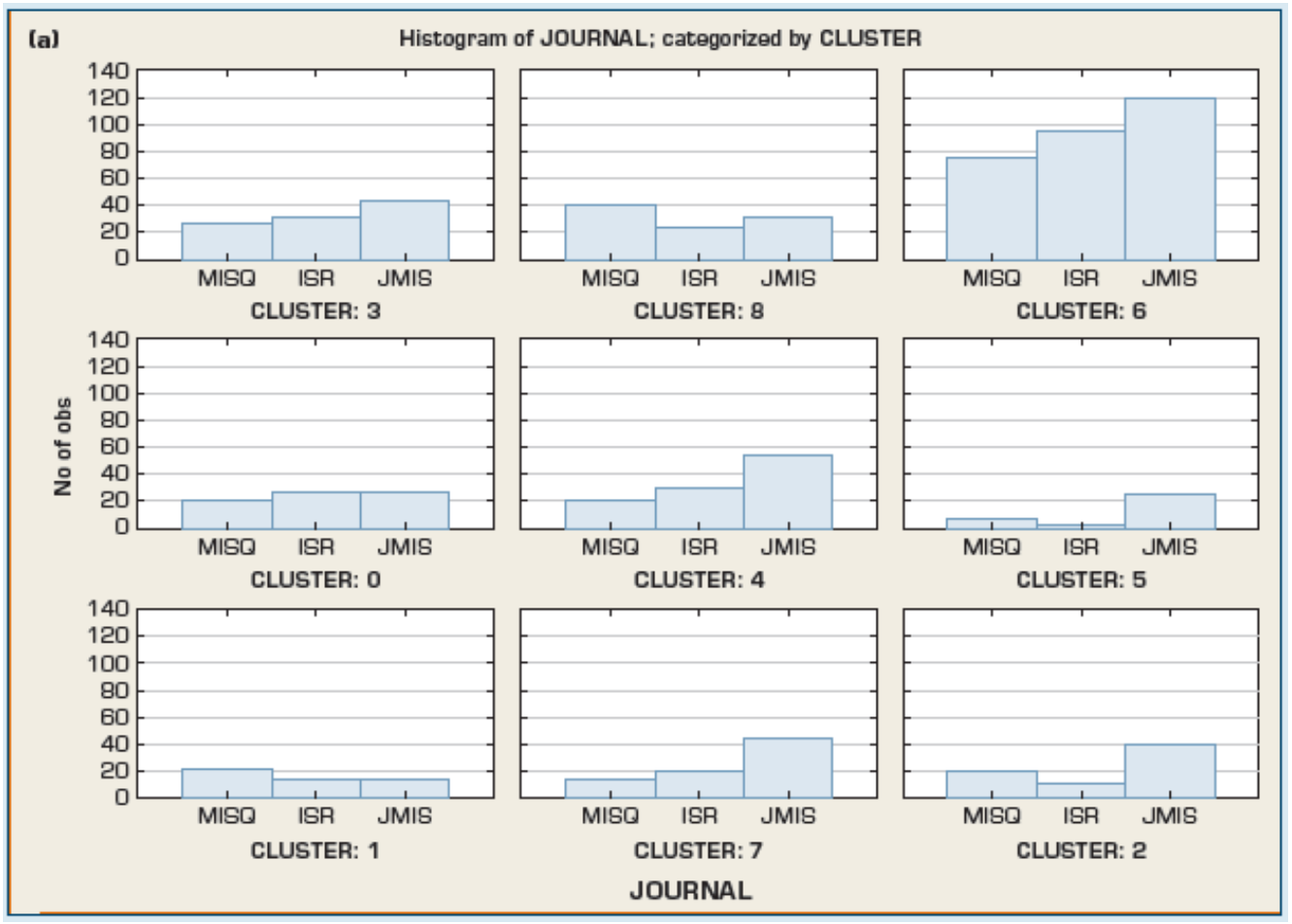
Application Case 5.5 (2 of 5)

Journal	Year	Author(s)	Title	Vol\No	Pages	Keywords	Abstract
MISQ	2005	A. Malhotra, S. Gosain and O.A. El sawy	Absorptive capacity configurations in supply chains: Gearing for partner-enabled market knowledge creation	29/1	145-197	knowledge management supply chain absorptive capacity interorganizational information systems configuration approaches	The need for continual value innovation is driving supply chains to evolve from a pure transactional focus to leveraging interorganizational partner ships for sharing
ISR	1999	D. Robey and M.C Boudreau	Accounting for the contradictory organizational consequences of information technology Theoretical directions and methodological Implications	2-Oct	167-185	organizational transformation impacts of technology organization theory research methodology intraorganizational power electronic communication mis implementation culture system	Although much contemporary thought considers advanced information technologies as either determinants or enablers of radical organizational change, empirical studies have revealed inconsistent findings to support the deterministic logic implicit in such arguments. This paper reviews the contradictory

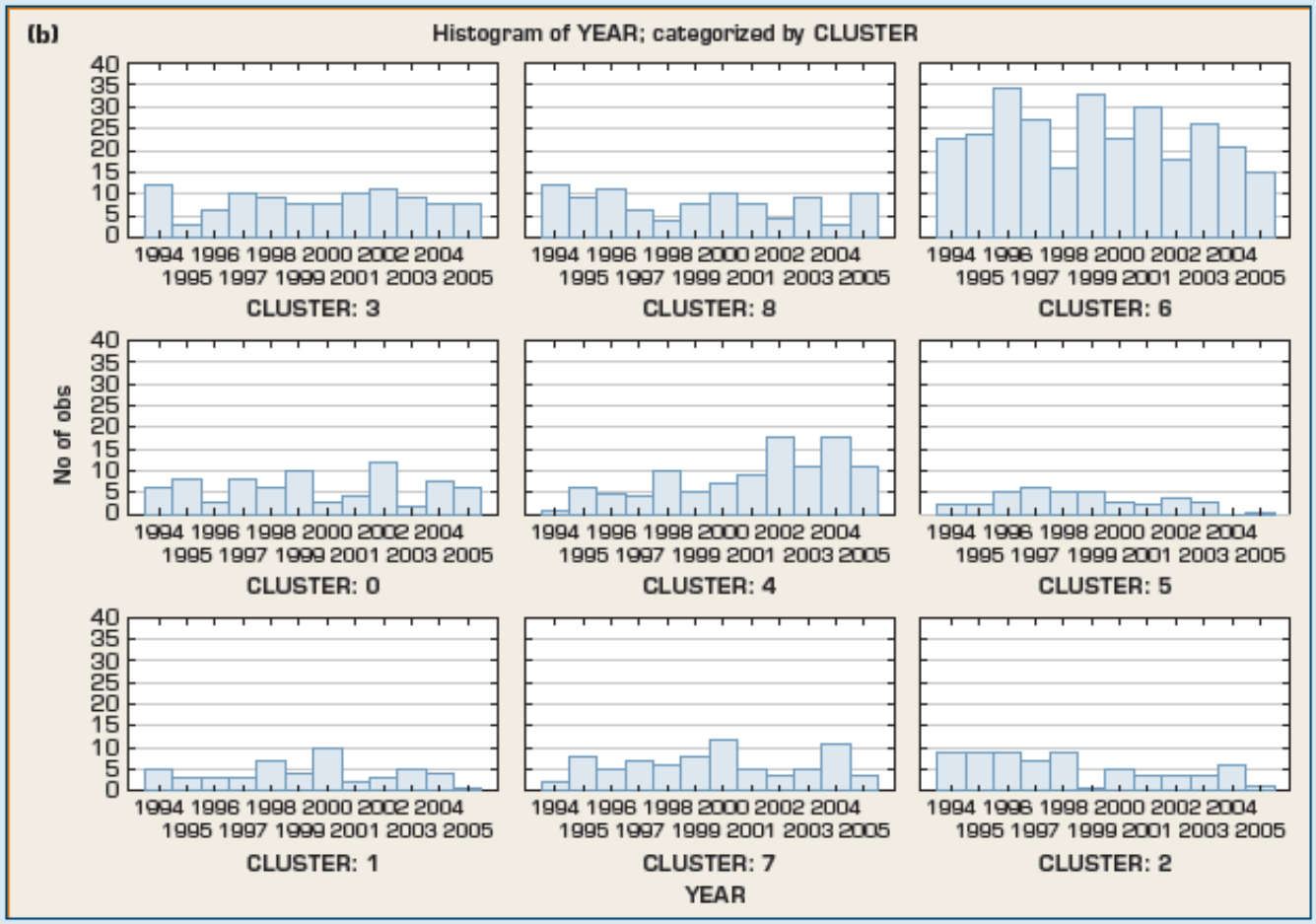
Application Case 5.5 (3 of 5)

Journal	Year	Author(s)	Title	Vol\No	Pages	Keywords	Abstract
JMIS	2001	R. Aron and E.K. Clemons	Achieving the optimal balance between investment in quality and investment in self-promotion for information products	18/2	65-88	information products internet advertising product positioning signaling signaling games	When producers of goods (or services) are confronted by a situation in which their offerings no longer perfectly match consumer preferences, they must determine the extend to which the advertised features of
…	…	…	…	…	…		

Application Case 5.5 (4 of 5)



Application Case 5.5 (5 of 5)



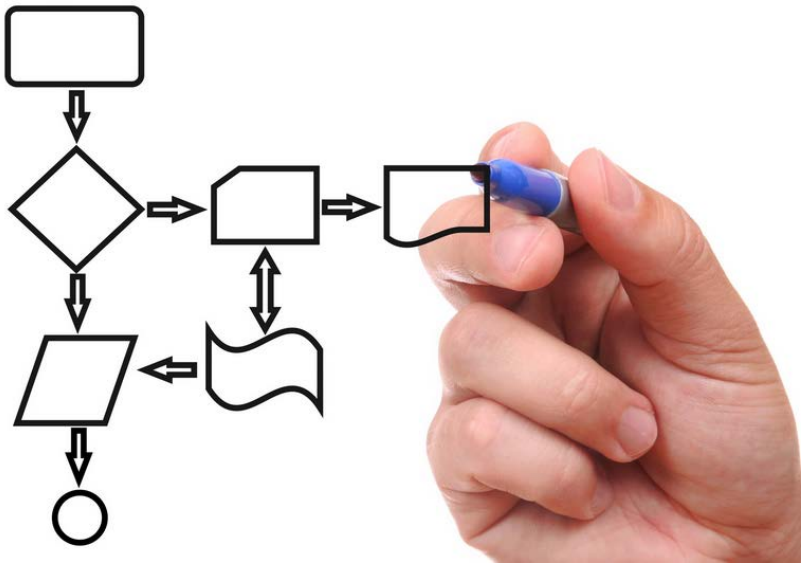
Sentiment Analysis

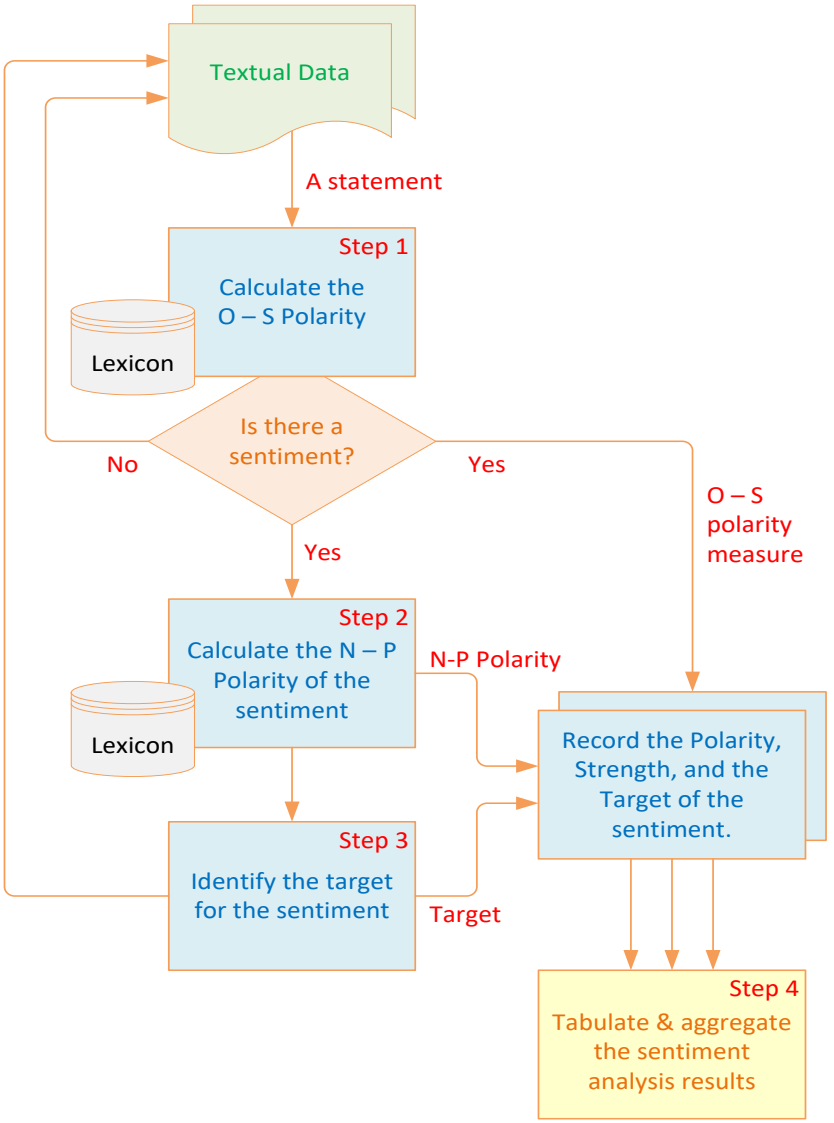
- Sentiment → belief, view, opinion, and conviction
- Sentiment analysis is trying to **answer** the question "What do people feel about a certain topic?"
- By analyzing data related to opinions of many using a variety of automated tools
- Used in variety of domains, but its applications in CRM are especially noteworthy (which related to customers/consumers' opinions)

Sentiment Analysis Applications

- Voice of the customer (VOC)
- Voice of the Market (VOM)
- Voice of the Employee (VOE)
- Brand Management
- Financial Markets
- Politics
- Government Intelligence
- … others

Sentiment Analysis Process (1 of 3)





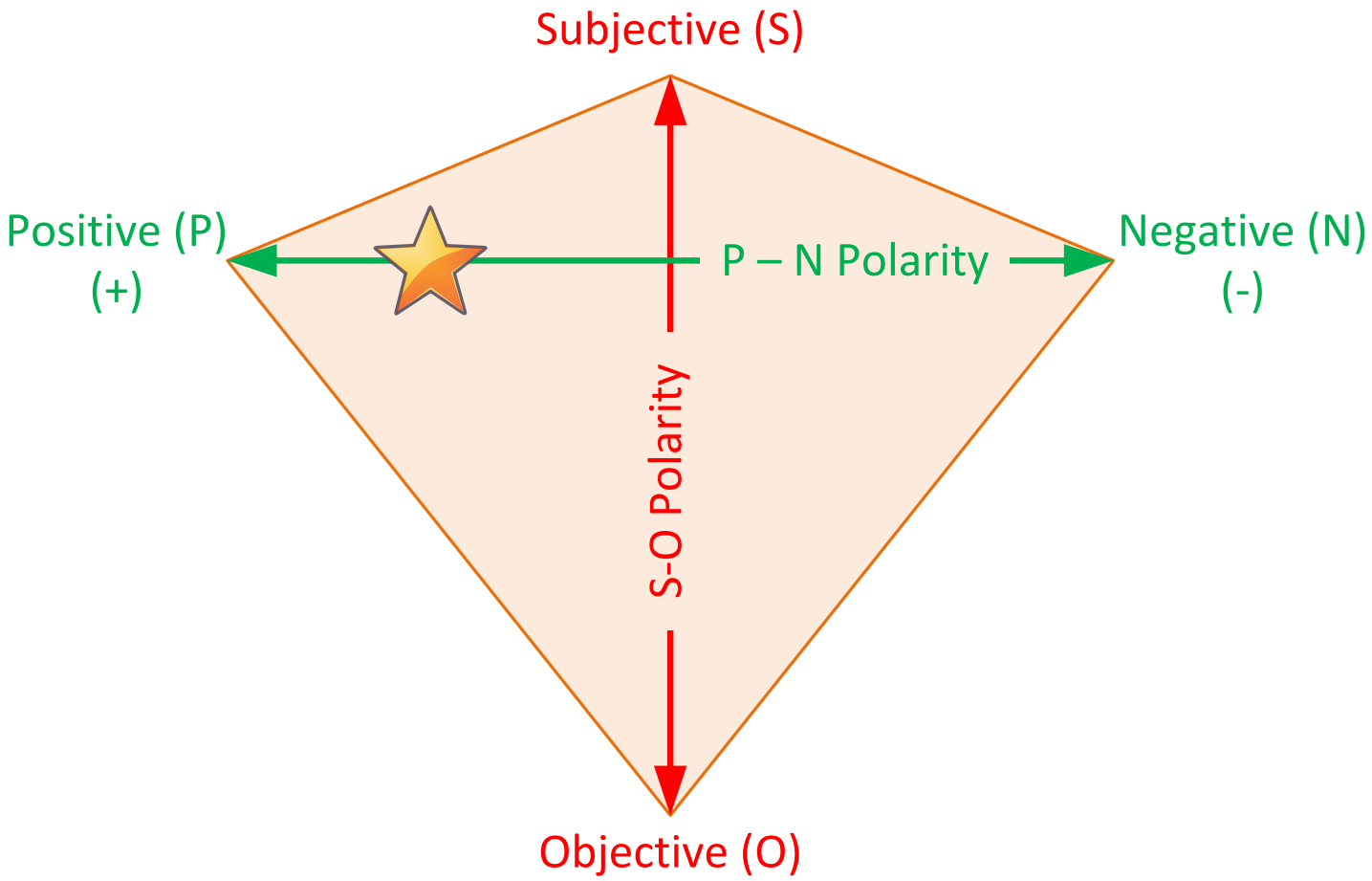
Sentiment Analysis Process (2 of 3)

- **Step 1** – Sentiment Detection
  - Comes right after the retrieval and preparation of the text documents
  - It is also called detection of objectivity
    - **Fact [= objectivity]** versus **Opinion [= subjectivity]**
- **Step 2** – **N-P Polarity Classification**
  - Given an opinionated piece of text, the goal is to classify the opinion as falling under one of two opposing sentiment polarities
    - **N [= negative]** versus **P [= positive]**

Sentiment Analysis Process (3 of 3)

- **Step 3** – Target Identification
  - The goal of this step is to accurately identify the target of the expressed sentiment (e.g., a person, a product, an event, etc.)
    - Level of difficulty → the application domain
- **Step 4** – Collection and Aggregation
  - Once the sentiments of all text data points in the document are identified and calculated, they are to be aggregated
    - Word → Statement → Paragraph → Document

P-N Polarity and S-O Polarity

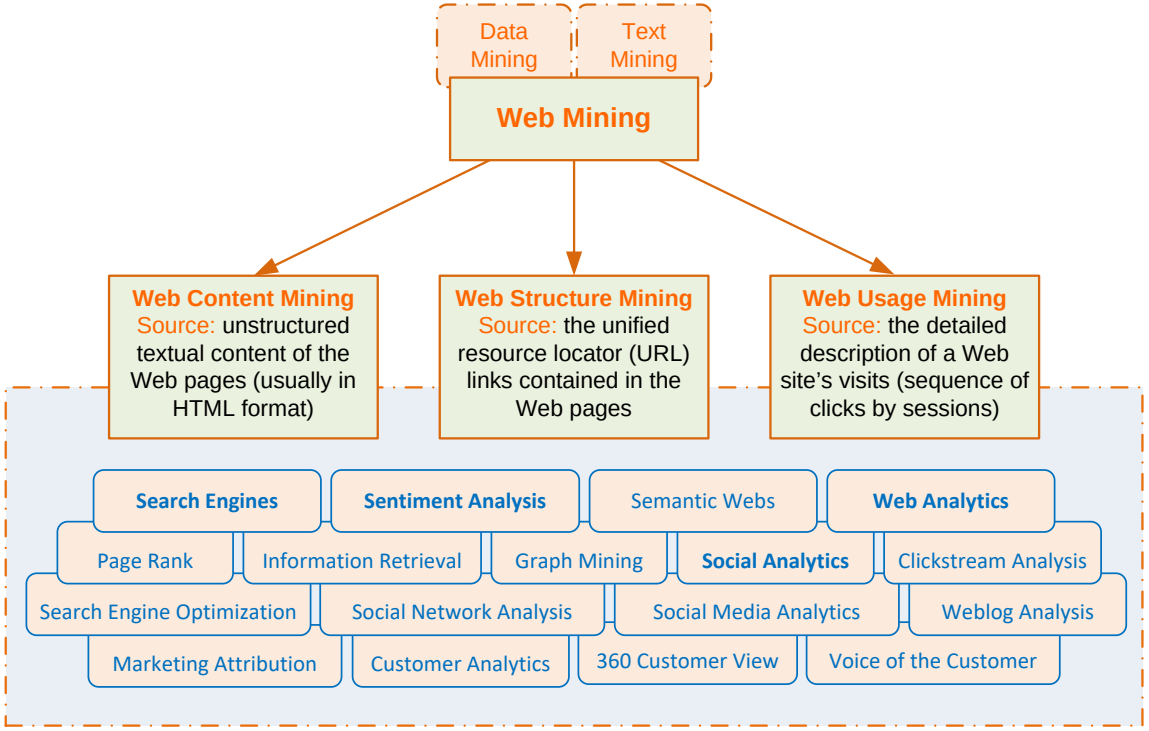


Web Mining Overview

- Web is the largest repository of data
- Data is in HTML, XML, text format
- Challenges (of processing Web data)
  - The Web is too big for effective data mining
  - The Web is too complex
  - The Web is too dynamic
  - The Web is not specific to a domain
  - The Web has everything
- **Opportunities and challenges are great!**

Web Mining

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



Web Content/Structure Mining

- Mining 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 (1 of 2)

- 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

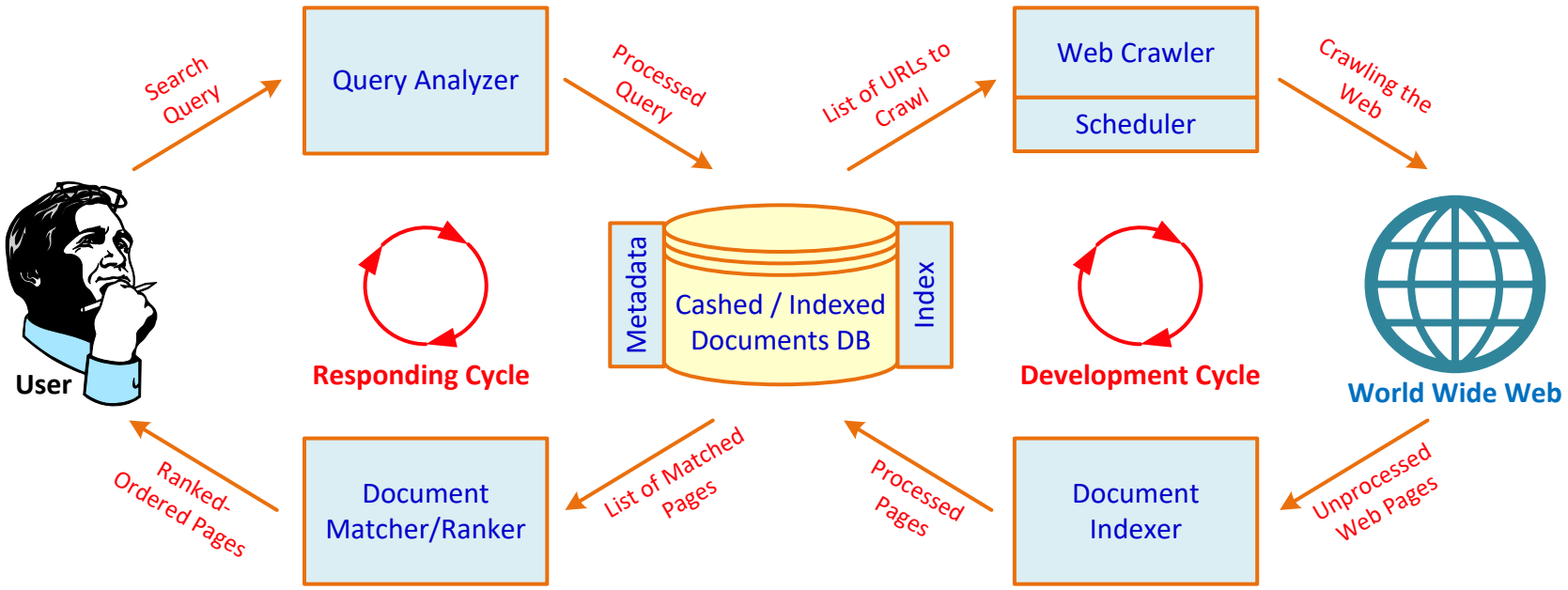
Web Usage Mining (2 of 2)

- 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
  - …

Search Engines

- Google, Bing, Yahoo, …
- For what reason do you use search engines?
- **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
- They are the workhorses of the Internet

Structure of a Typical Internet Search Engine



Anatomy of a Search Engine

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

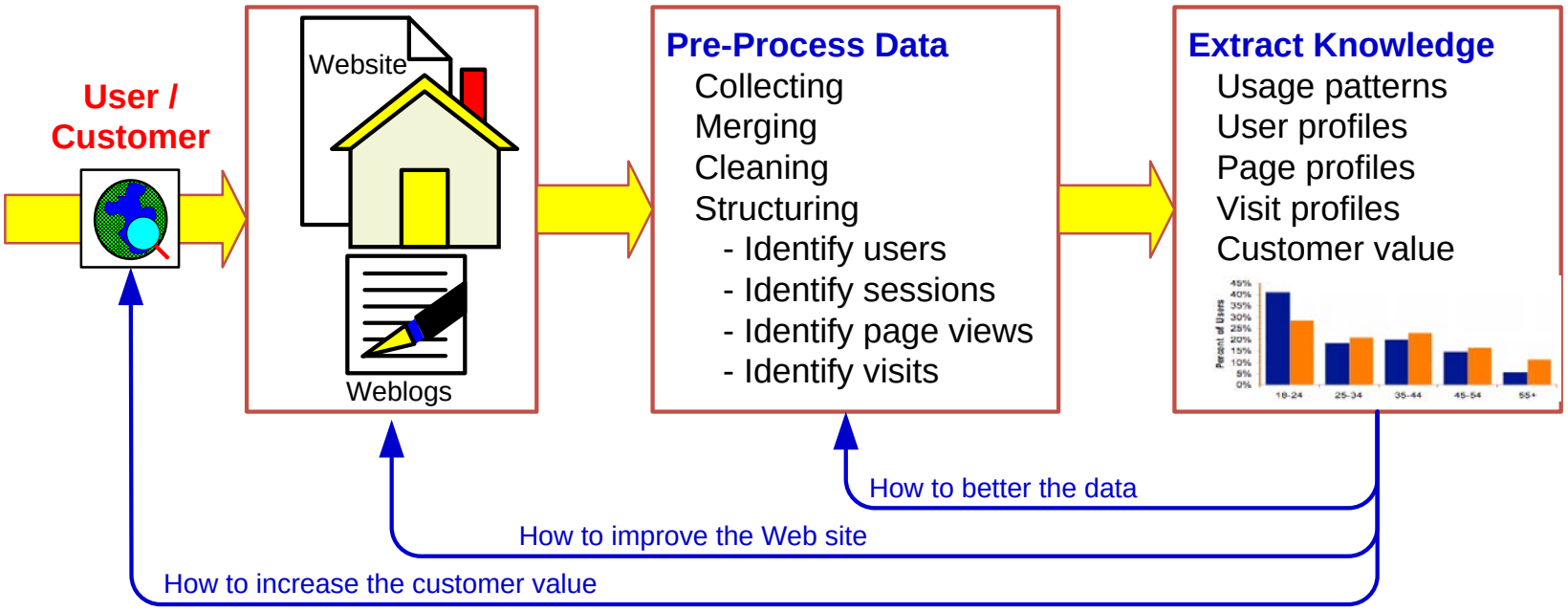
Search Engine Optimization

- 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
- Part of an Internet marketing strategy
- Based on knowing how a Search Engine works
  - Content, HTML, keywords, external links, …
- Indexing based on …
  - Webmaster submission of URL
  - Proactively and continuously crawling the Web

Top 15 Most Popular Search Engines (by eBizMBA)

Rank	Name	Estimated Unique Monthly Visitors
1	Google	1,600,000,000
2	Bing	400,000,000
3	Yahoo! Search	300,000,000
4	Ask	245,000,000
5	AOL Search	125,000,000
6	Wow	100,000,000
7	WebCrawler	65,000,000
8	MyWebSearch	60,000,000
9	Infospace	24,000,000
10	Info	13,500,000
11	DuckDuckGo	11,000,000
12	Contenko	10,500,000
13	Dogpile	7,500,000
14	Alhea	4,000,000
15	ixQuick	1,000,000

Web Usage Mining (Clickstream Analysis)



Web Analytics Metrics (1 of 3)

- **Web site usability**
  - How were the visitors using my Web site?
- **Traffic sources**
  - Where did they come from?
- **Visitor profiles**
  - What do my visitors look like?
- **Conversion statistics**
  - What does it all mean for the business?

Web Analytics Metrics (2 of 3)

Web Site Usability

- Page views
- Time on site
- Downloads
- Click map
- Click paths

Traffic Source

- Referral Web sites
- Search engines
- Direct
- Offline campaigns
- Online campaigns

Web Analytics Metrics (3 of 3)

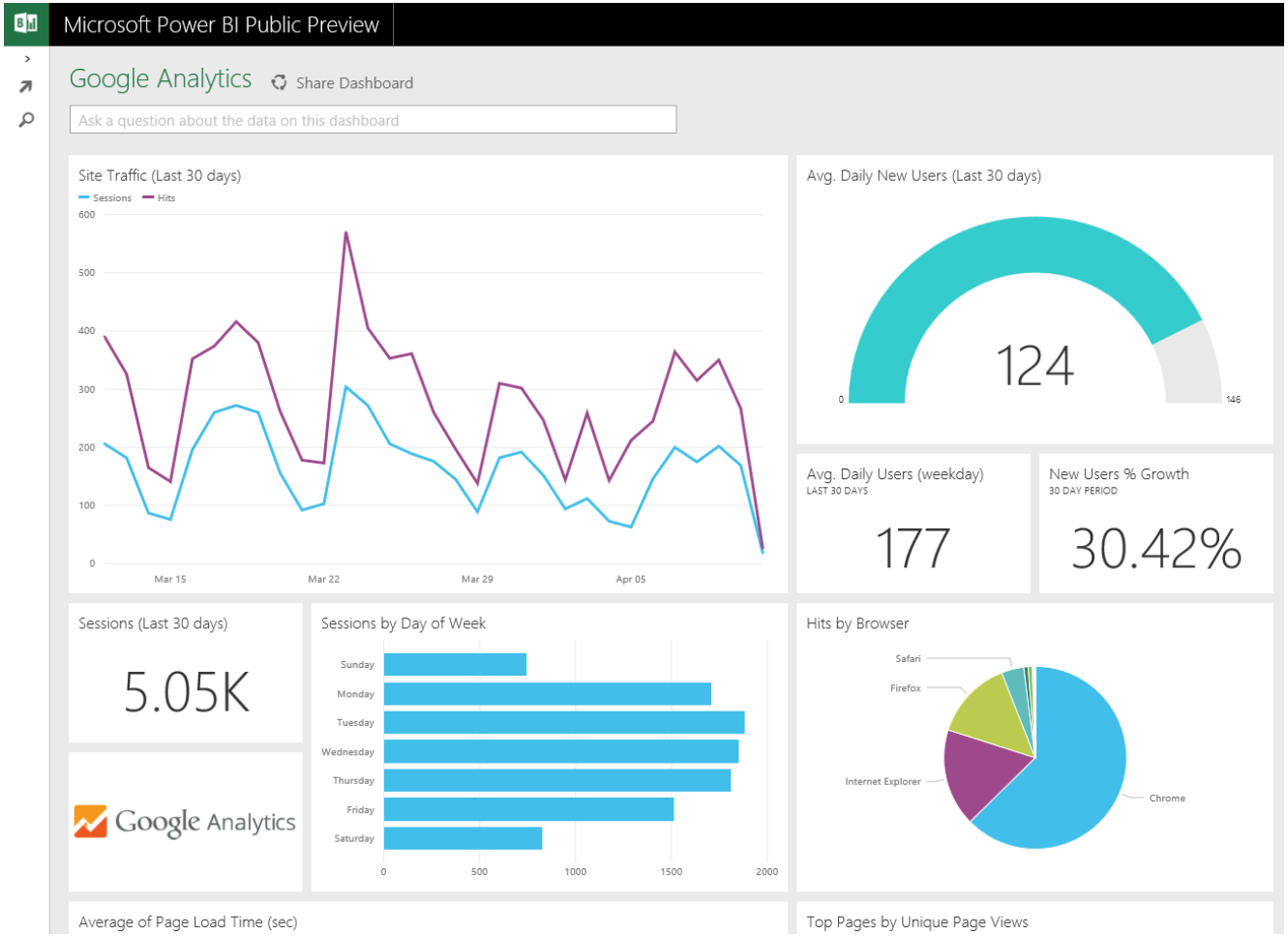
Visitor Profiles

- Keywords
- Content groupings
- Geography
- Time of day
- Landing page profiles

Conversion Statistics

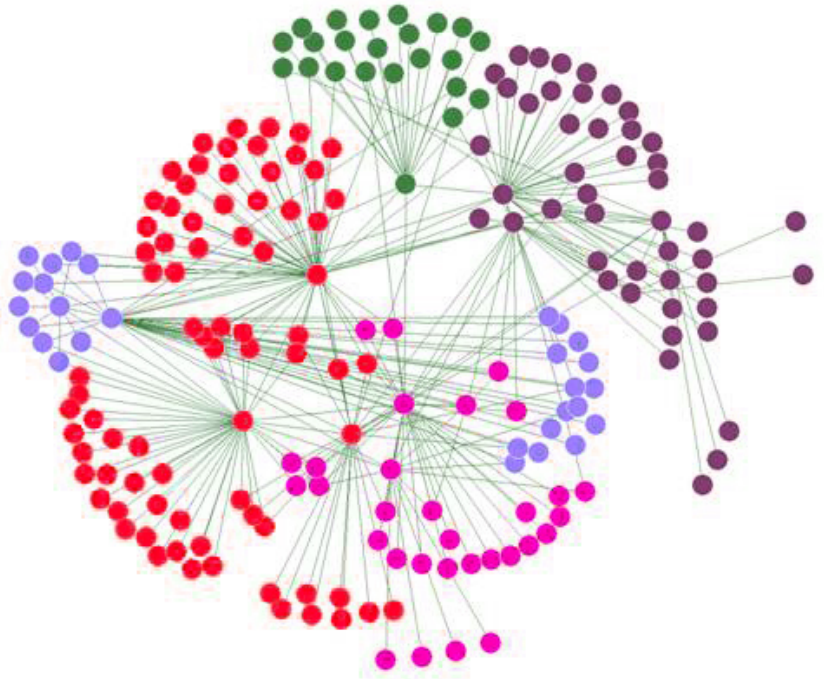
- New visitors
- Returning visitors
- Leads
- Sales/conversions
- Abandonment/exit rate

A Sample Web Analytics Dashboard



Social Analytics Social Network Analysis (1 of 2)

- **Social Network** - social structure composed of individuals linking to each other
- Analysis of social dynamics
- Interdisciplinary field
  - Social psychology
  - Sociology
  - Statistics
  - Graph theory



Social Analytics Social Network Analysis (2 of 2)

- **Social Networks help study** relationships between individuals, groups, organizations, societies
  - **Self organizing**
  - **Emergent**
  - **Complex**
- Typical social network types
  - Communication networks, community networks, criminal networks, innovation networks, …

Application Case 5.8

Tito's Vodka Establishes Brand Loyalty with an Authentic Social Strategy

Discussion Questions

1. How can social media analytics be used in the consumer products industry?
2. What do you think are the key challenges, potential solutions, and probable results in applying social media analytics in consumer products and services firms?

Social Analytics Social Network Analysis Metrics

- **Connections**
  - Homophily
  - Multiplexity
  - Mutuality/reciprocity
  - Network closure
  - Propinquity
- **Segmentation**
  - Cliques and social circles
  - Clustering coefficient
  - Cohesion
- **Distribution**
  - Bridge
  - Centrality
  - Density
  - Distance
  - Structural holes

Social Media Definitions and Concepts

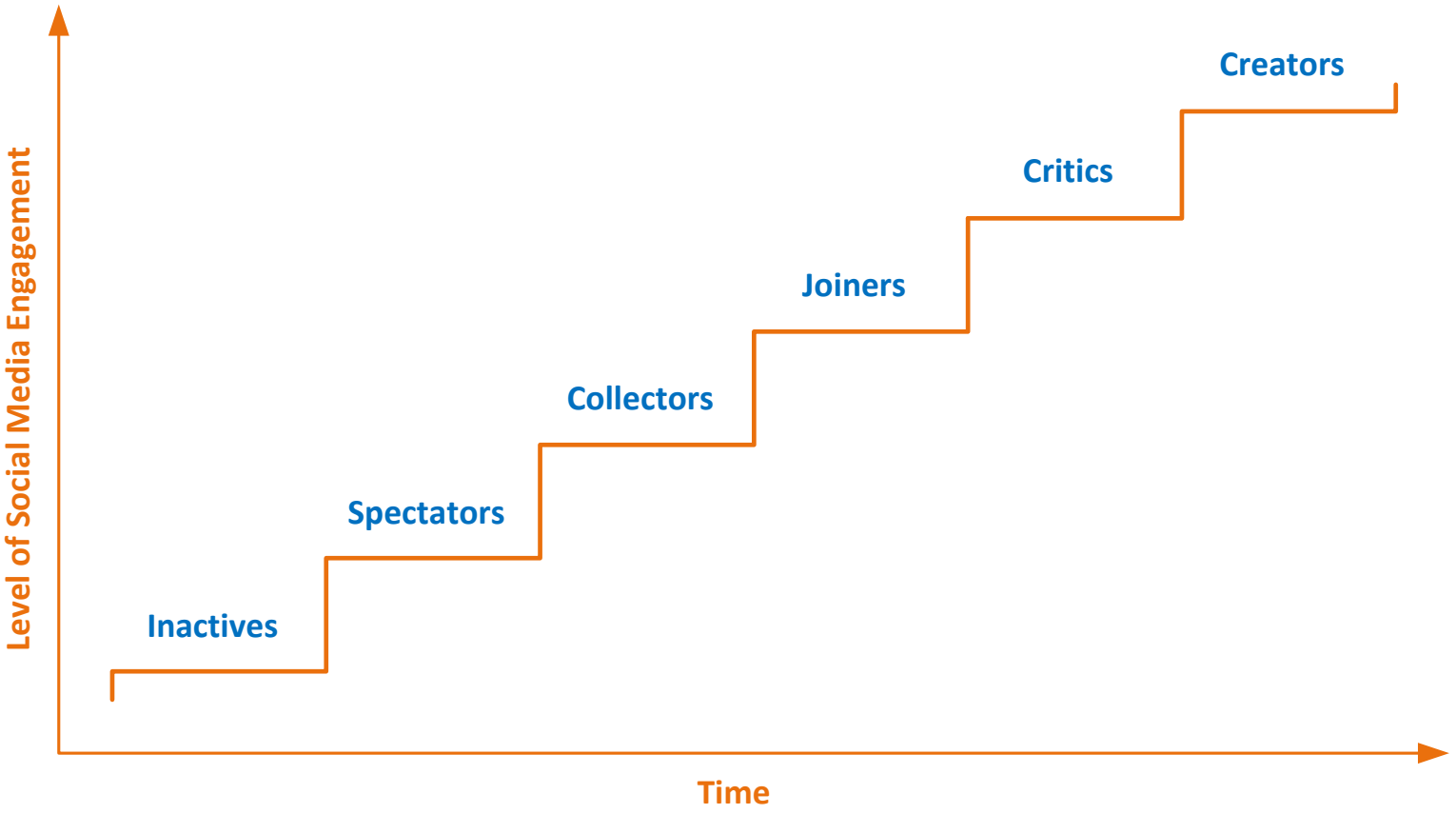
- Enabling technologies of social interactions among people
- Relies on enabling technologies of Web 2.0
- Takes on many different forms
  - Internet forums, Web logs, social blogs, microblogging, wikis, social networks, podcasts, pictures, video, and product reviews
- Different types of social media
  - Based on **media research** and **social process**

Social Versus 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

How Do People Use Social Media?

- Different engagement levels



Social Media Analytics

- It is the systematic and scientific ways to consume the vast amount of content created by Web-based social media outlets, tools, and techniques for the betterment of an organization’s competitiveness
- Tools to measure social media impact:
  - Descriptive analytics
  - Social network analysis
  - Advanced analytics

Best Practices in Social Media Analytics

- Think of measurement as a guidance system, not a rating system
- Track the elusive sentiment
- Continuously improve the accuracy of text analysis
- Look at the ripple effect
- Look beyond the brand
- Identify your most powerful influencers
- Look closely at the accuracy of your analytic tool
- Incorporate social media intelligence into planning