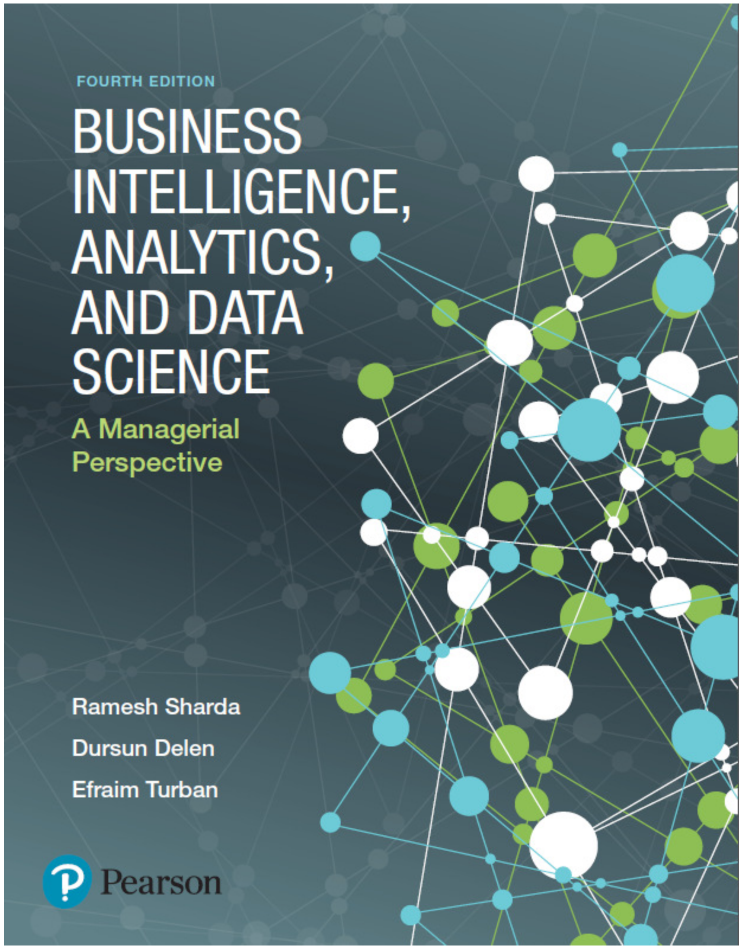
Business Intelligence, Analytics, and Data Science: A Managerial Perspective

Fourth Edition



Chapter 5 – Part A

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

Pearson

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

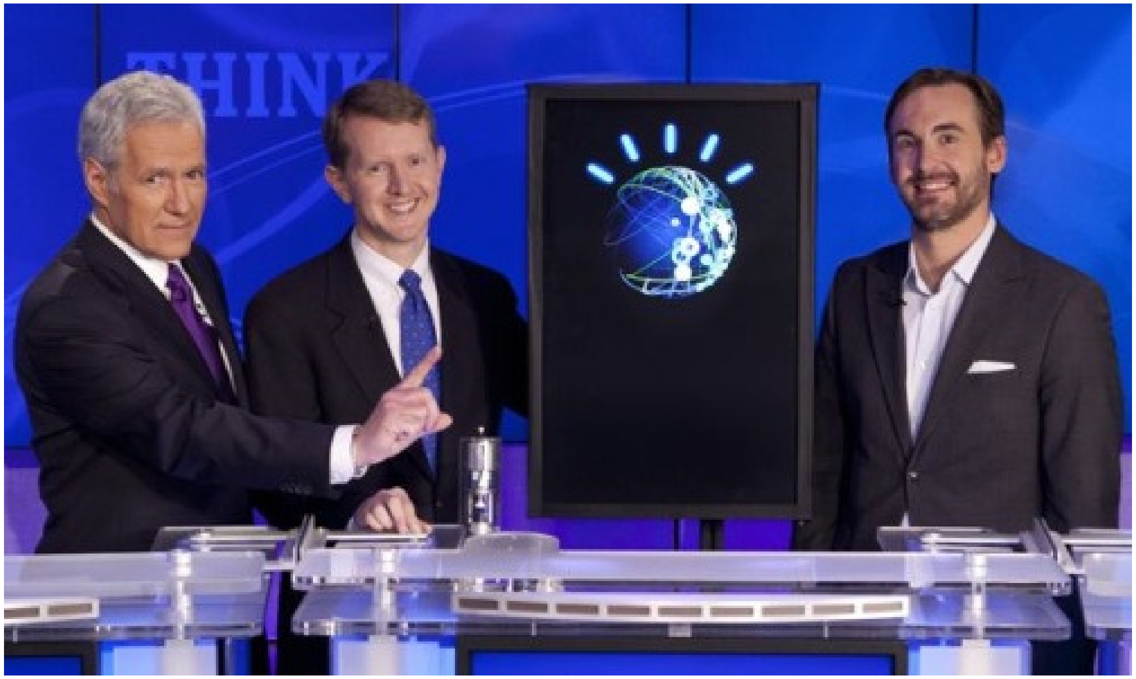
5.7 Develop familiarity with popular applications of sentiment analysis

5.8 Learn the common methods for sentiment analysis

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

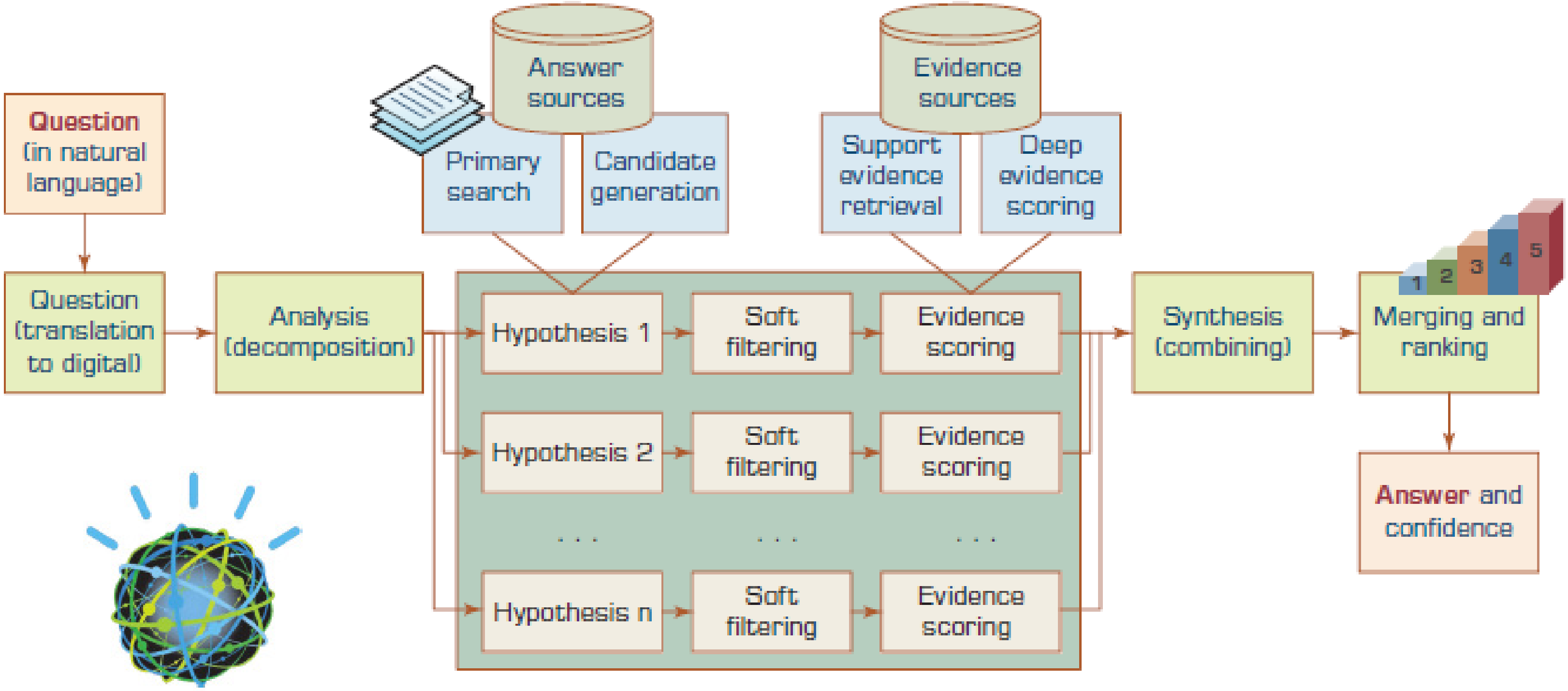
OPENING VIGNETTE Machine versus Men on Jeopardy!: The Story of Watson (1 of 3)

- IBM Watson going head-to-head with the best of the best in *Jeopardy!*



OPENING VIGNETTE Machine versus Men on Jeopardy!: The Story of Watson (2 of 3)

- IBM Watson – How does it do it?



OPENING VIGNETTE Machine versus Men on Jeopardy!: The Story of Watson (3 of 3)

Discussion Questions for the Opening Vignette

1. What is Watson? What is special about it?
2. What technologies were used in building Watson (both hardware and software)?
3. What are the innovative characteristics of DeepQA architecture that made Watson superior?
4. Why did IBM spend all that time and money to build Watson? Where is the return on investment (ROI)?

Text Analytics and Text Mining

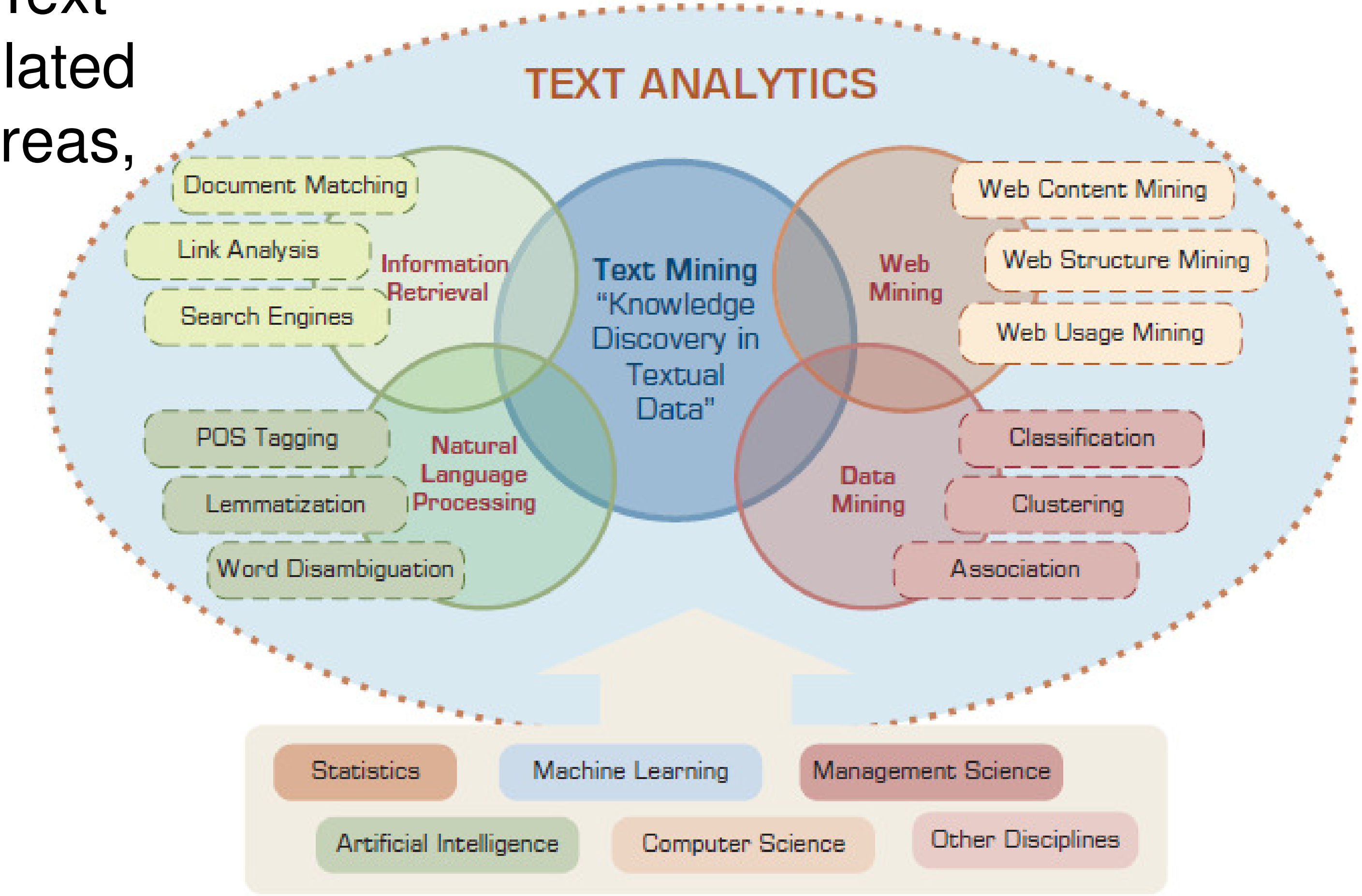
- Text Analytics versus Text Mining
- Text Analytics =
  - Information Retrieval +
  - Information Extraction +
  - Data Mining +
  - Web Mining

or simply

Text Analytics = Information Retrieval + Text Mining

Text Analytics and Text Mining

- FIGURE 5.2 Text Analytics, Related Application Areas, and Enabling Disciplines



Text Mining Concepts

- 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

- 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

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

Text Mining Terminology

- 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)

- 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)

- 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)

- 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

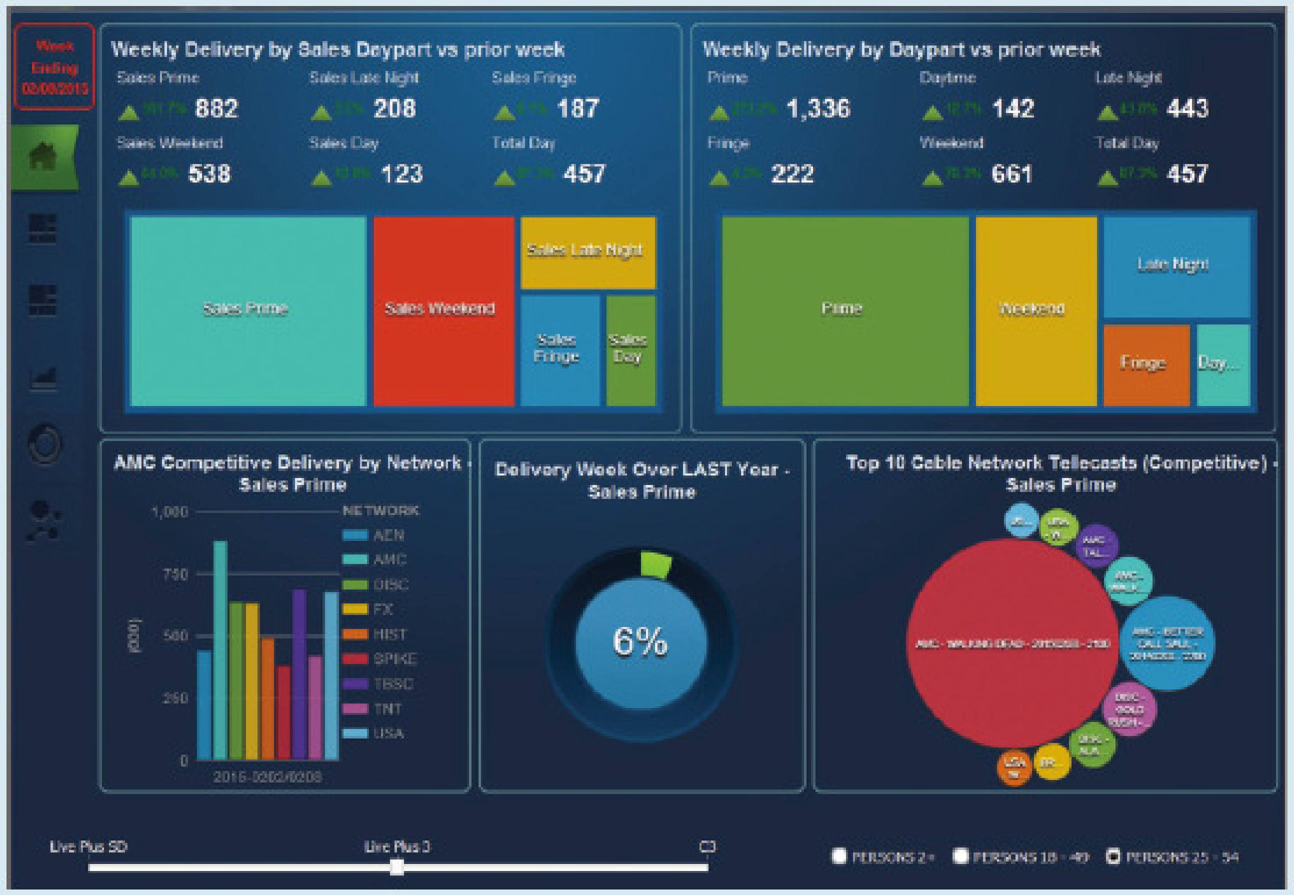
Natural Language Processing (NLP)

- 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

AMC Networks Is Using Analytics to Capture New Viewers, Predict Ratings, and Add Value for Advertisers in a Multichannel World (1 of 2)

A Web-Based Dashboard Used by AMC Networks [*Source:* AMC Networks]



Application Case 5.2

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

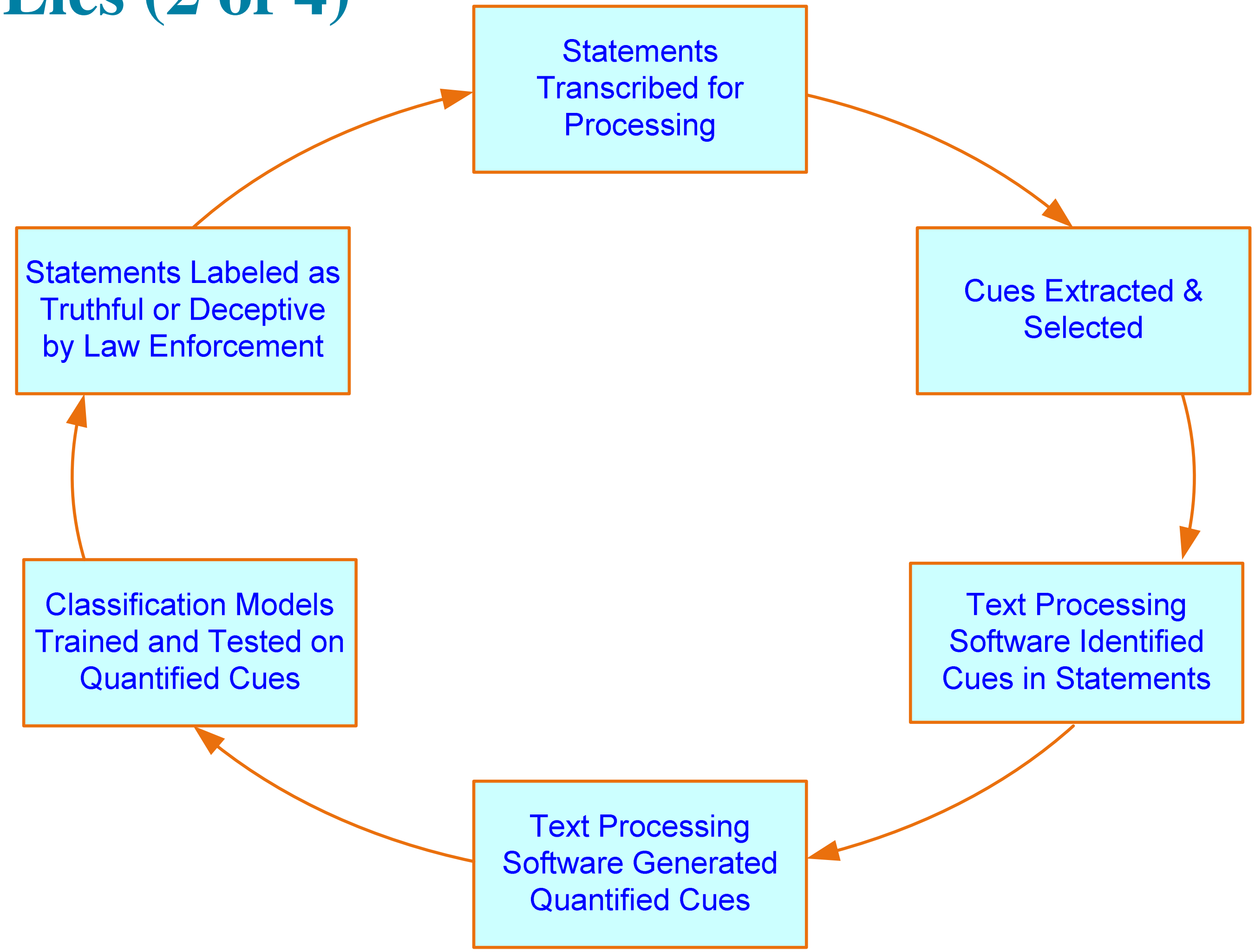
Application Case 5.3
Mining for Lies (1 of 4)

- 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

Mining for Lies (2 of 4)

- FIGURE 5.3 Text-Based Deception-Detection Process


Statements Transcribed for Processing
Cues Extracted & Selected
Text Processing Software Identified Cues in Statements
Text Processing Software Generated Quantified Cues
Classification Models Trained and Tested on Quantified Cues
Statements Labeled as Truthful or Deceptive by Law Enforcement

Application Case 5.3

Mining for Lies (3 of 4)

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

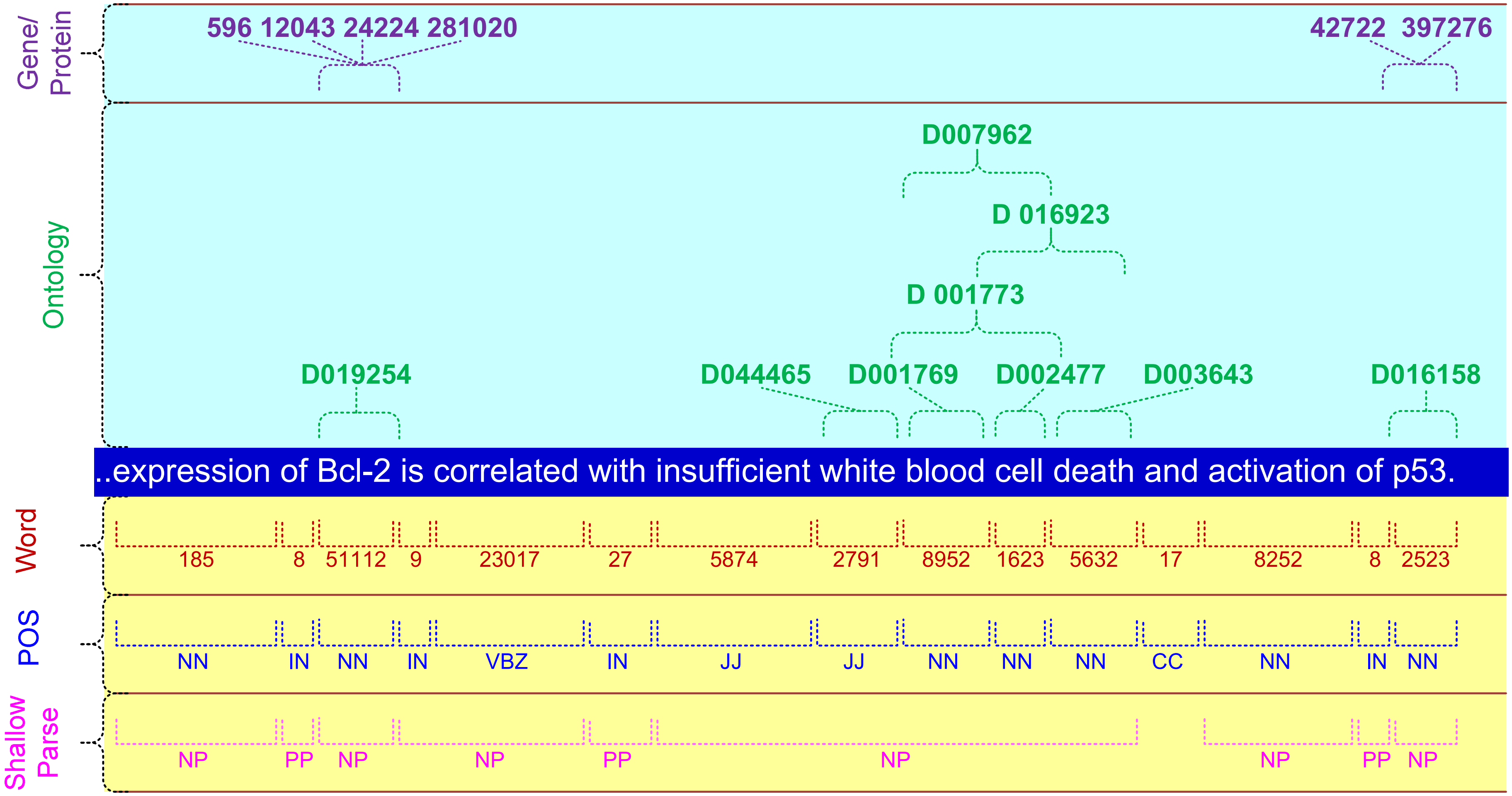
Category	Example Cues
Quantity	Verb count, noun-phrase count, ...
Complexity	Avg. no of clauses, sentence length, …
Uncertainty	Modifiers, modal verbs, ...
Nonimmediacy	Passive voice, objectification, ...
Expressivity	Emotiveness
Diversity	Lexical diversity, redundancy, ...
Informality	Typographical error ratio
Specificity	Spatiotemporal, perceptual information …
Affect	Positive affect, negative affect, etc.

Application Case 5.3
Mining for Lies (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)



Pearson

Application Case 5.4

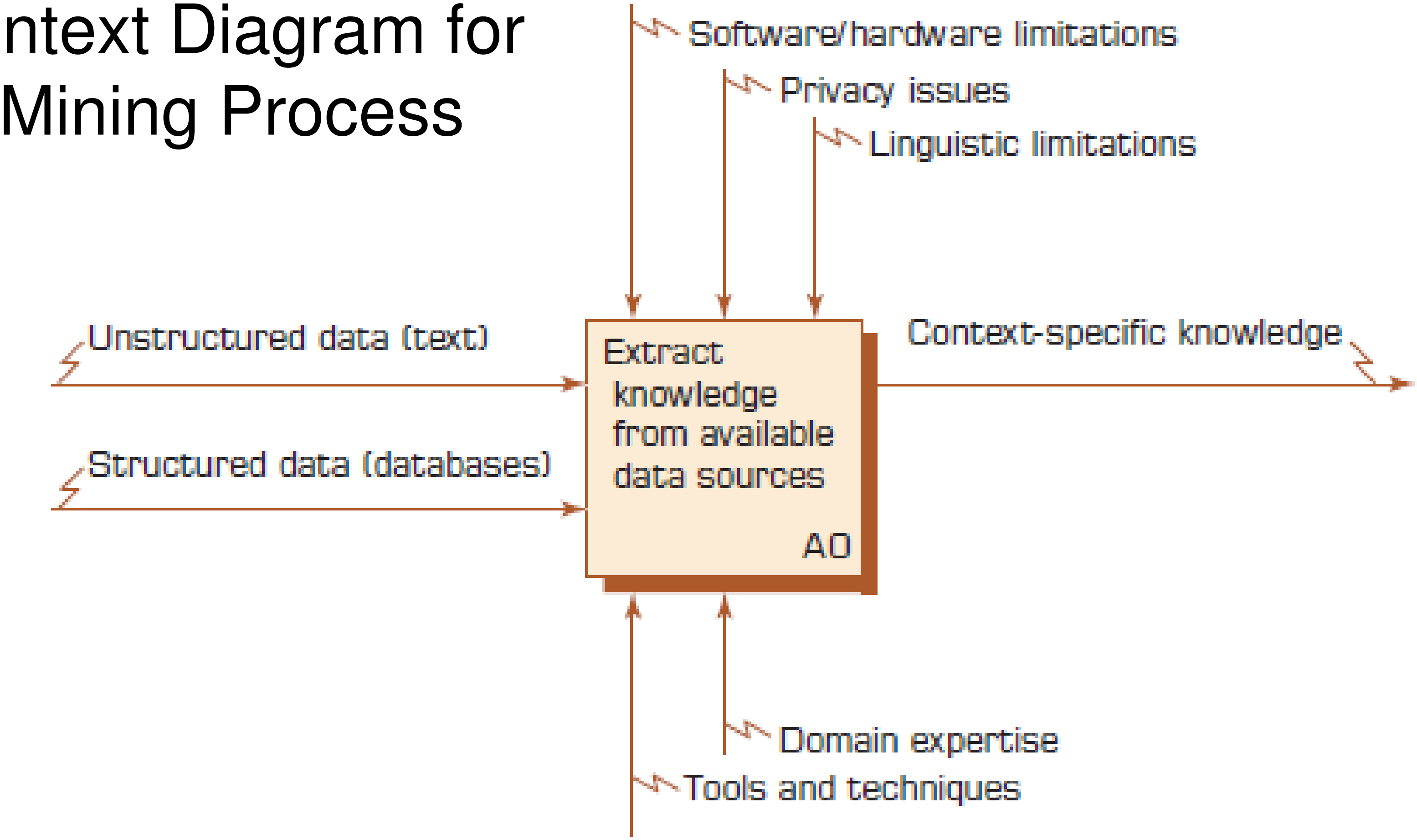
Bringing the Customer into the Quality Equation: Lenovo Uses Analytics to Rethink Its Redesign

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?

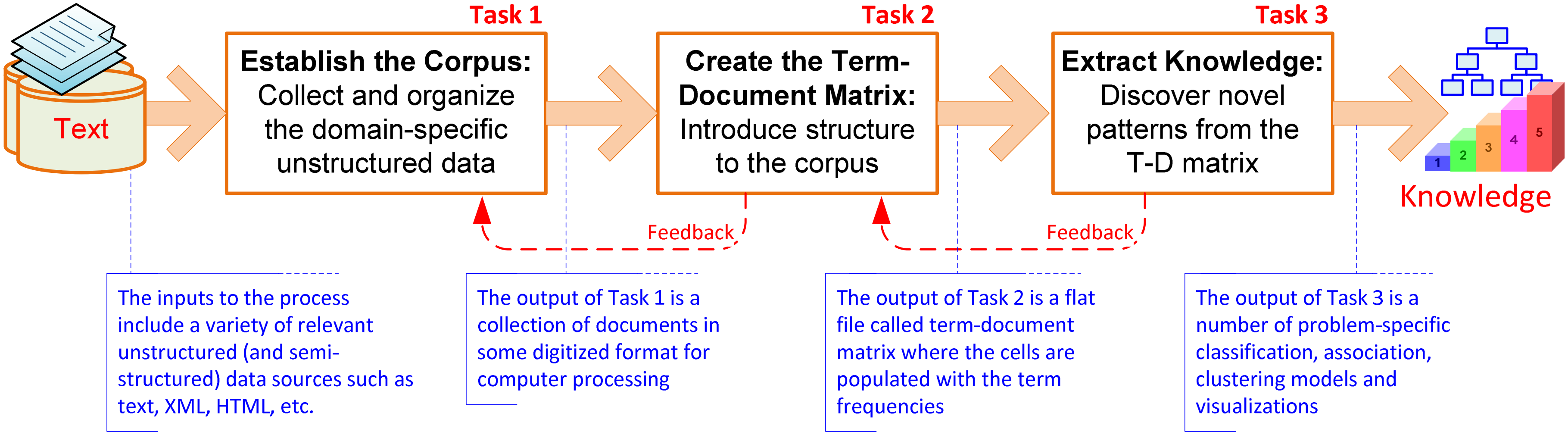
Text Mining Process

- A Context Diagram for Text Mining Process


Software/hardware limitations
Privacy issues
Linguistic limitations
Unstructured data (text)
Structured data (databases)
Extract knowledge from available data sources
A0
Context-specific knowledge
Domain expertise
Tools and techniques

Text Mining Process

- FIGURE 5.6
  The Three-Step/Task Text Mining Process



Text
Task 1
Establish the Corpus: Collect and organize the domain-specific unstructured data
Task 2
Create the Term-Document Matrix: Introduce structure to the corpus
Task 3
Extract Knowledge: Discover novel patterns from the T-D matrix
Knowledge
Feedback
Feedback
1
2
3
4
5

The inputs to the process include a variety of relevant unstructured (and semi-structured) data sources such as text, XML, HTML, etc.

The output of Task 1 is a collection of documents in some digitized format for computer processing

The output of Task 2 is a flat file called term-document matrix where the cells are populated with the term frequencies

The output of Task 3 is a number of problem-specific classification, association, clustering models and visualizations

Text Mining Process

- 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

- 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

- Step 2: Create the Term–by–Document Matrix (TDM) (Cont.)
  - 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

- Step 2: Create the Term–by–Document Matrix (TDM) (Cont.)
  - 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

- 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

Research Literature Survey with Text Mining (1 of 4)

- 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

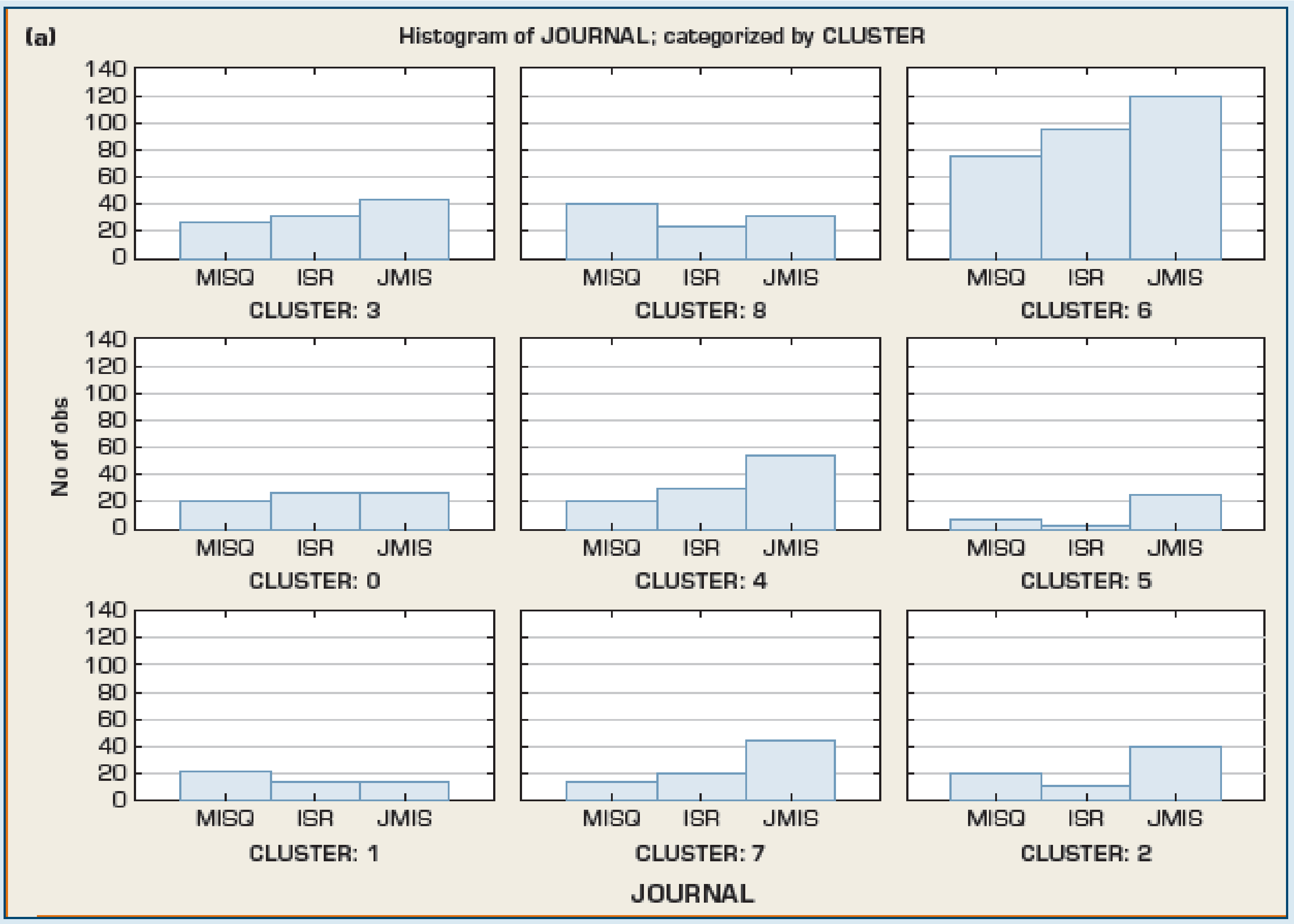
Research Literature Survey with Text Mining (2 of 4)

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	145-187	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 systems	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
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 extent to which the advertised features of
…	…	…	…	…	…	…	…

Pearson

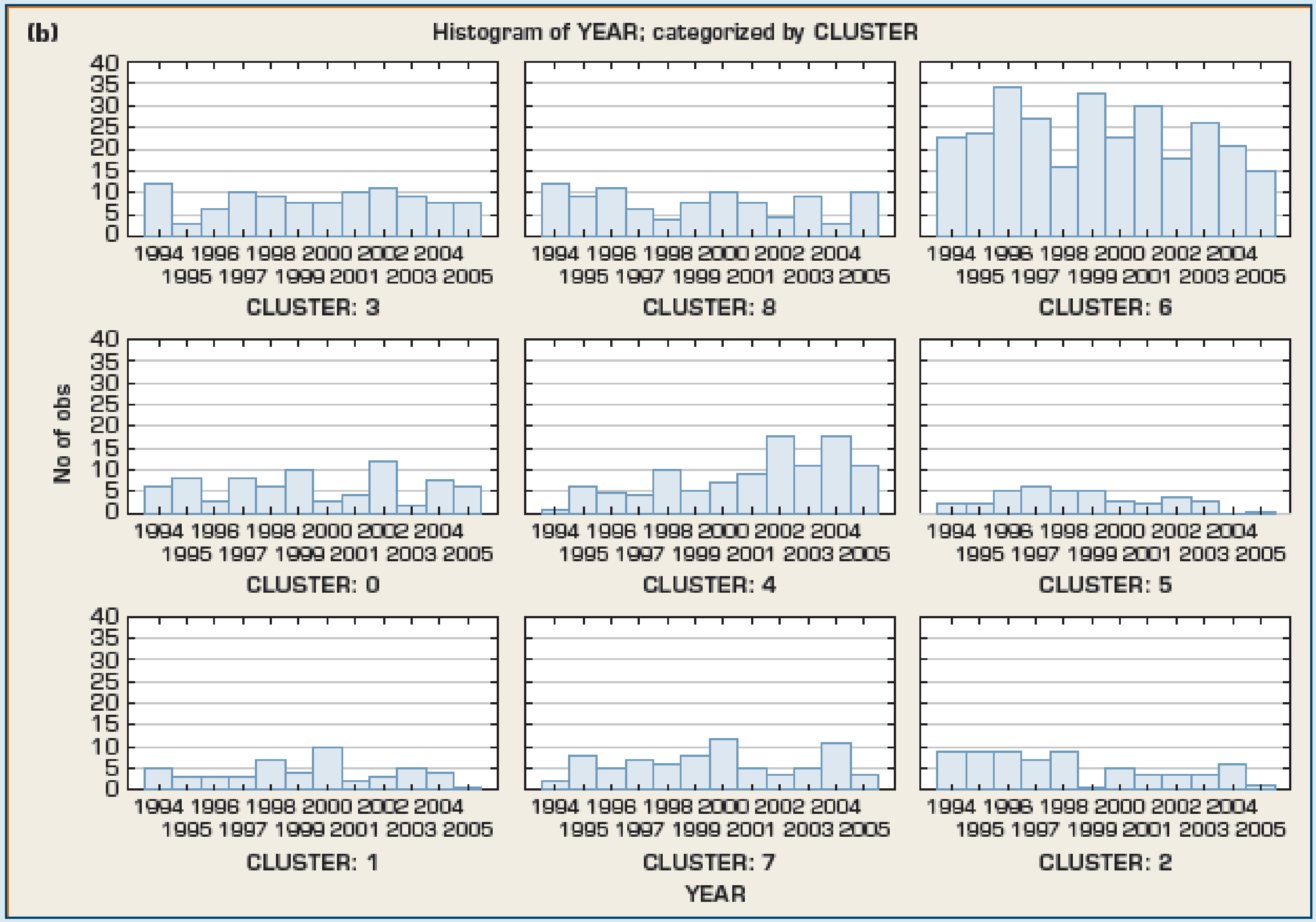
Application Case 5.5

Research Literature Survey with Text Mining (3 of 4)



Application Case 5.5

Research Literature Survey with Text Mining (4 of 4)


(b)
Histogram of YEAR; categorized by CLUSTER
No of obs
CLUSTER: 3
CLUSTER: 8
CLUSTER: 6
CLUSTER: 0
CLUSTER: 4
CLUSTER: 5
CLUSTER: 1
CLUSTER: 7
CLUSTER: 2
YEAR