MSc GFIS

Information Integration

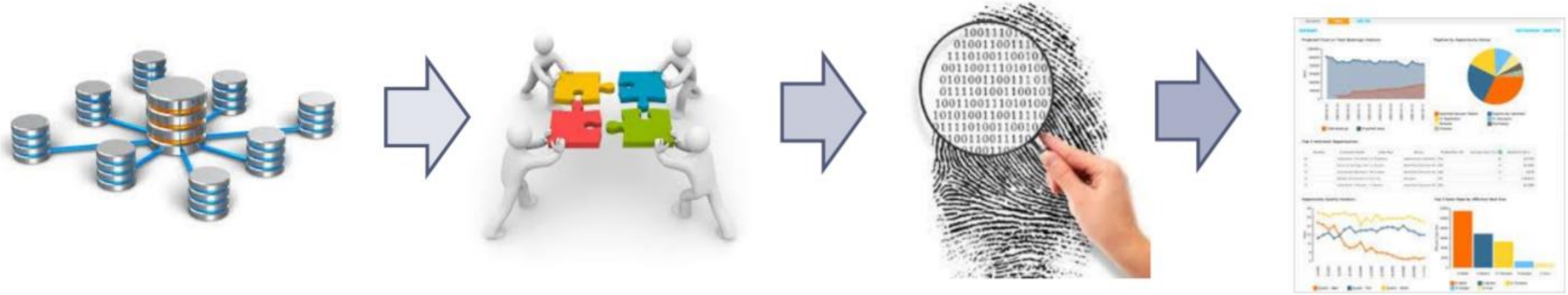
Ruth Barry, rbarry@wit.ie

Objectives

- How is Information integrated?
- Types of Data
- Environmental scanning
- Text mining – concepts, applications, terminology, algorithms and process
- Web mining – content, structure, usage and metrics

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

Capabilities of BI



Organisational Memory

Information Integration

Information Insights

Information Presentation

Storage of structured information in such a form that it can be later accessed and used for BI

Integration of semi-structured and unstructured information so it can be used by BI

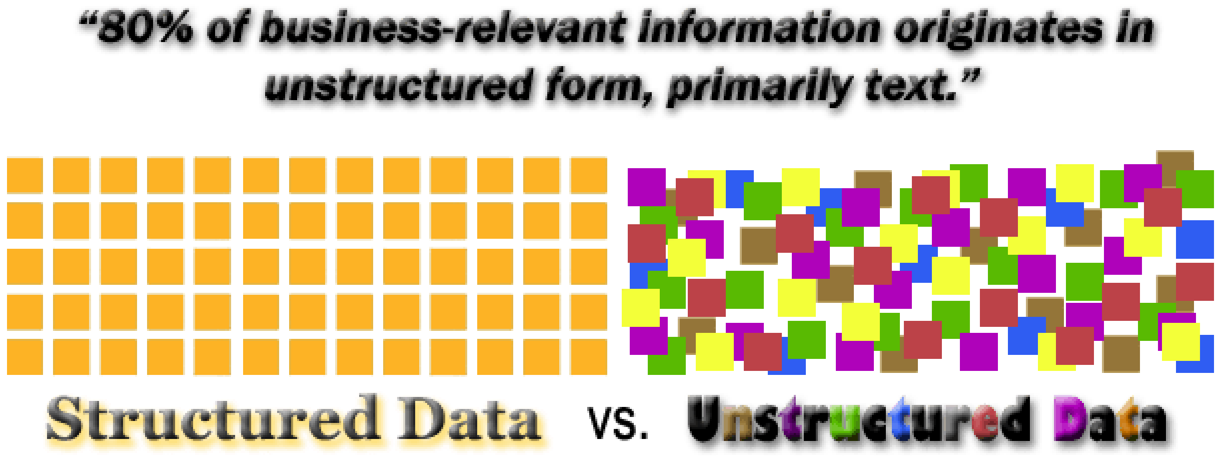
Business Intelligence, Rajiv Sabherwal, Irma Becerra-Fernandez, John Wiley & Sons, 2014

Information Integration

- How is information integrated?
  - Synthesis of new insights from unstructured data residing in the organisation's enterprise systems, such as enterprise portals and document management systems.
  - Creation of new insights via the integration of structured organizational data with external data, such as Web-based unstructured information from customer Web sites or vendor data sources.

Types of Data

1. Structured
2. Semistructured
3. Unstructured

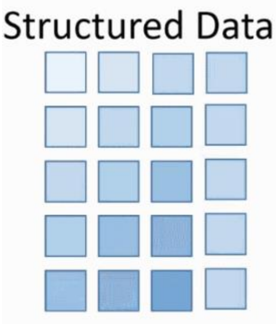


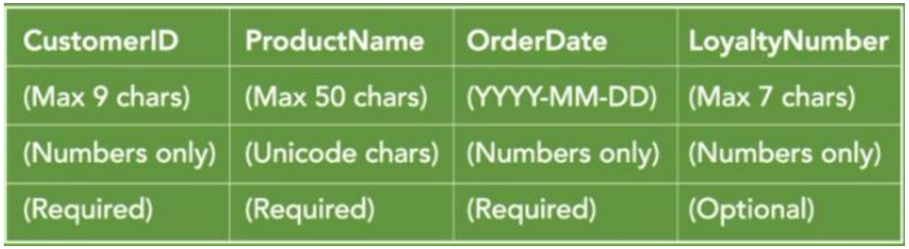
"80% of business-relevant information originates in unstructured form, primarily text."

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

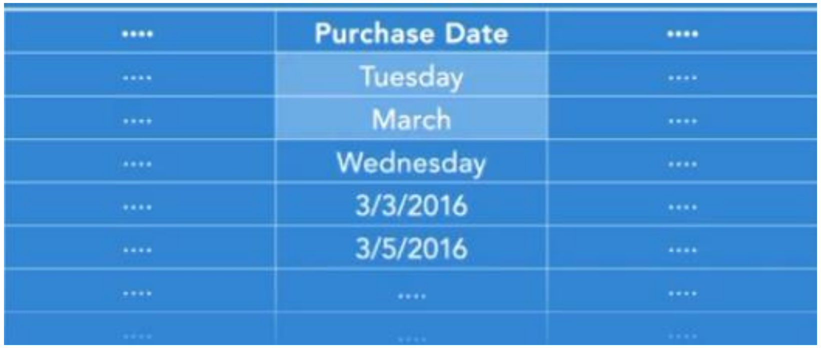
Structured data

- Brick and mortar of the database
- Cheap, inflexible and requires a lot of upfront design.
- Good example is an office Spreadsheet.
- Structured data relies on the data model.
  - Each row must follow same format!





CustomerID	ProductName	OrderDate	LoyaltyNumber
(Max 9 chars)	(Max 50 chars)	(YYYY-MM-DD)	(Max 7 chars)
(Numbers only)	(Unicode chars)	(Numbers only)	(Numbers only)
(Required)	(Required)	(Required)	(Optional)

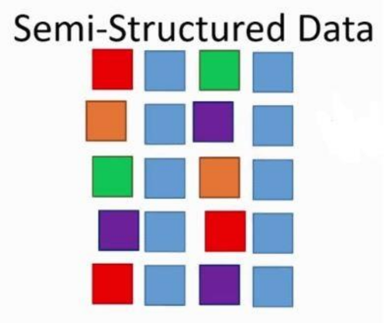


....	Purchase Date	....
....	Tuesday	....
....	March	....
....	Wednesday	....
....	3/3/2016	....
....	3/5/2016	....
....	....	....
....	....	....

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

Semistructured data

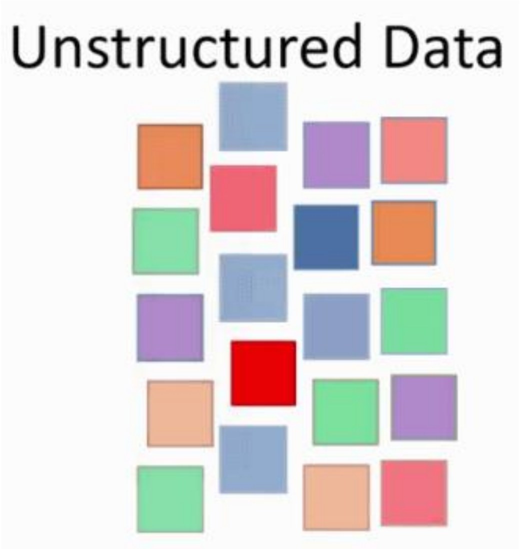
- Semistructured data is even more popular than structured data.
- It has a structure, but that structure depends on the source.
- E-mail is an example of semi-structured data.
- Common ways to work with semistructured data – XML and JSON – JavaScript Object Notation.



Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

Unstructured data

- Some analysts estimate that 80% of your data is unstructured.
- It’s schemaless (no set data model), e.g.
  - Every time you leave a voice mail
  - A picture on Facebook
  - Search on a search engine

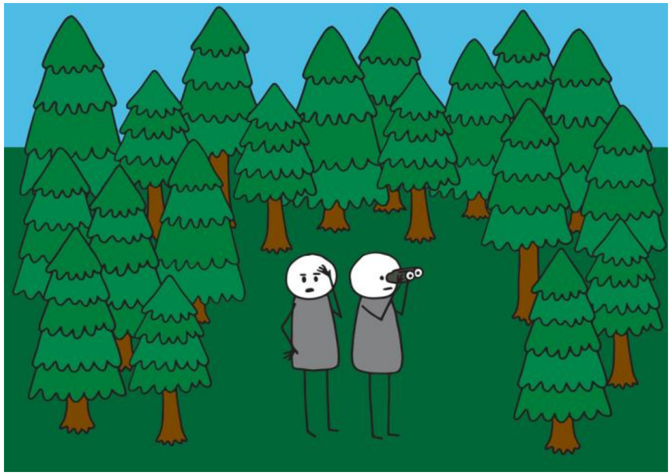


Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

Why integrate with unstructured data?

Environmental Scanning

- Scanning for information about events and relationships in a company's outside environment,
  - the knowledge of which would assist top management in its task of planning the company's future course of action
- Improve organizational performance
- 'looking for' and 'looking at' information



Environmental Scanning

- Brown and Weiner (1985) define environmental scanning as "a kind of radar to scan the world systematically and signal the new, the unexpected, the major and the minor"
- "Searching the environment for important events or issues that might affect an organisation"

Business Intelligence, Rajiv Sabherwal, Irma Becerra-Fernandez, John Wiley & Sons, 2014

Environmental Scanning-Uncertainty

- Industry or market
- Technology
- Regulatory
- Economic
- Social
- Political

Business Intelligence, Rajiv Sabherwal, Irma Becerra-Fernandez, John Wiley & Sons, 2014

How do we scan?

- Technologies and techniques for identifying and extracting external business information.
  - Text mining
  - Web mining

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

Text Mining

- According to a study by Merrill Lynch & Gartner, 85% 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

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

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

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

Text Mining

- Applying software technology to understand volumes of (unstructured) text.
- Analyzing the data to determine which terms are more prevalent than others.
- Learning how terms and phrases are related to one another.
- Understanding what the common themes in the document collection are.

automated data mining survey
responses co ter transcripts
qualatativ oot cause
classificati text analysis insights
ad-hoc an s product
reviews se t vo f the
customer dashboards consume
trends ad-hoc analysis early warning

- https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/discovering-what-you-want-107347.pdf

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

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

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

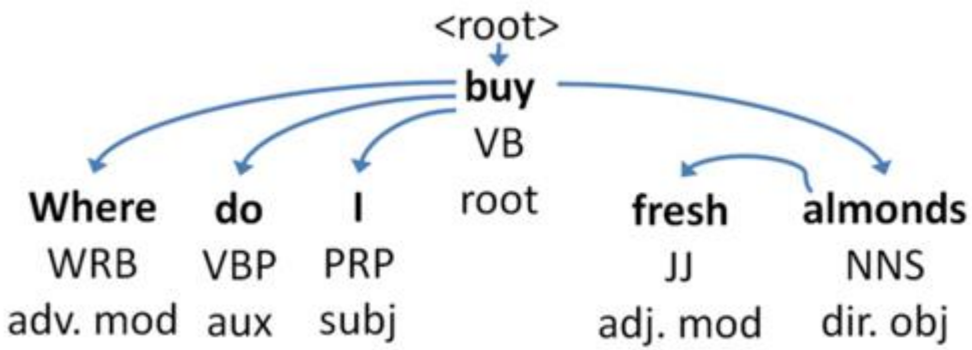
Text Mining Application Area

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

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

Text Mining Terminology

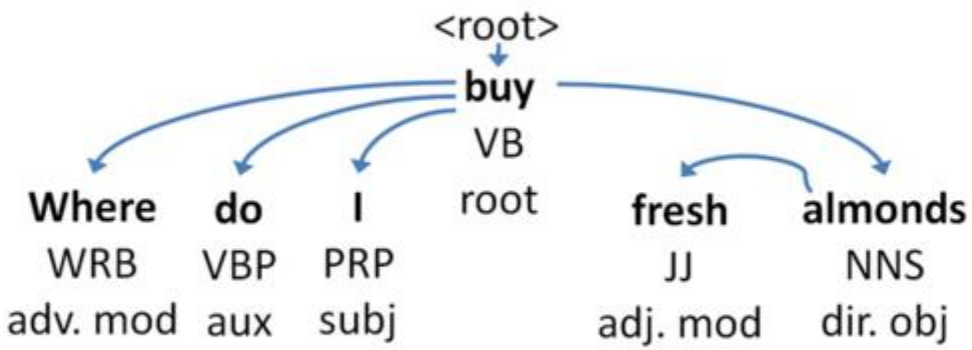
- Unstructured data
- Corpus (and corpora)
- Terms
- Concepts
- Stemming
- Stop words (and include words)
- Synonyms (and polysemes)
- Tokenizing



Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

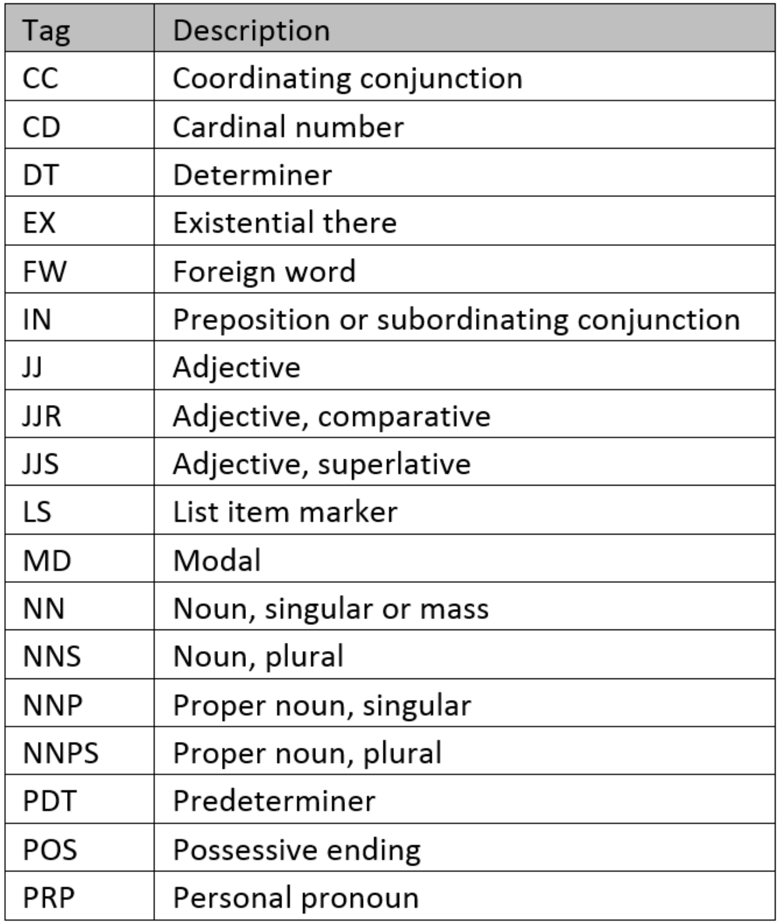
Text Mining Terminology

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

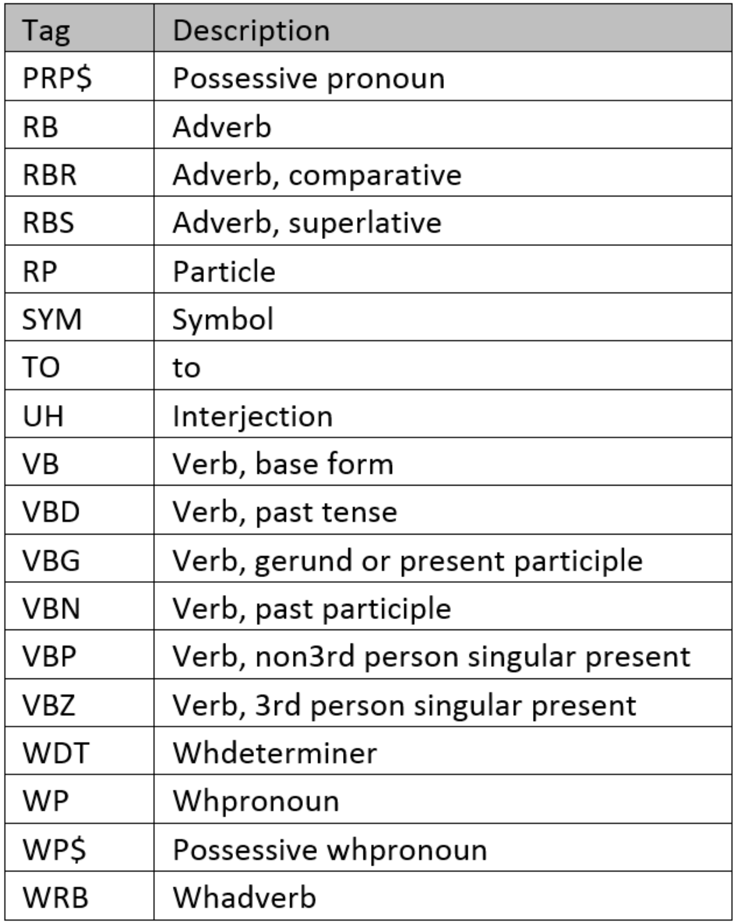


Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

POS Tags and their meanings



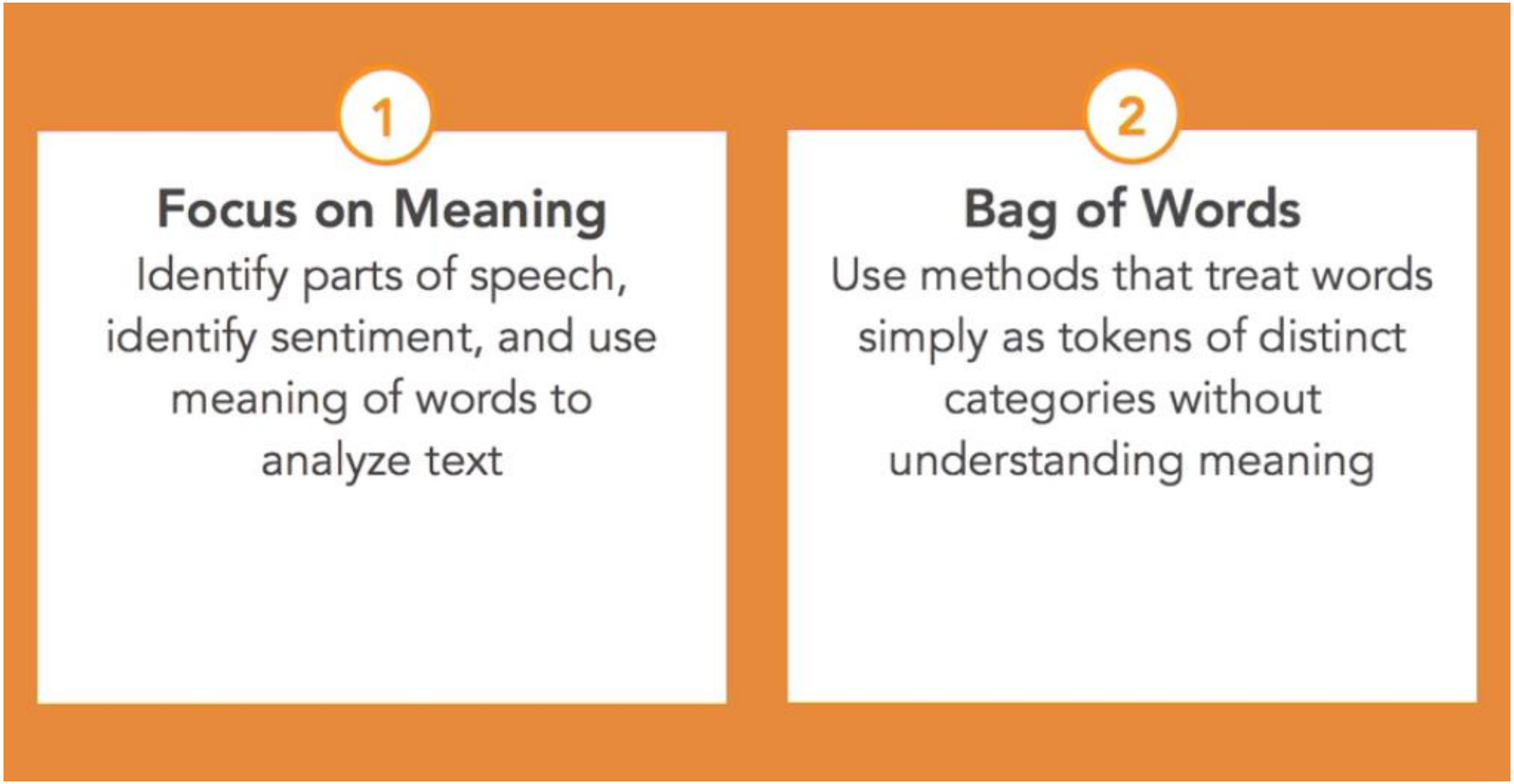
Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun



Tag	Description
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Whdeterminer
WP	Whpronoun
WP$	Possessive whpronoun
WRB	Whadverb

http://blog.thedigitalgroup.com/sagarg/2015/06/18/part-of-speech-tagging-using-opennlp/

Text Mining algorithms



1

Focus on Meaning

Identify parts of speech, identify sentiment, and use meaning of words to analyze text

2

Bag of Words

Use methods that treat words simply as tokens of distinct categories without understanding meaning

Lynda.com https://www.lynda.com/Business-Intelligence-tutorials/Text-mining-algorithms/475936/529757-4.html

Natural Language Processing (NLP)

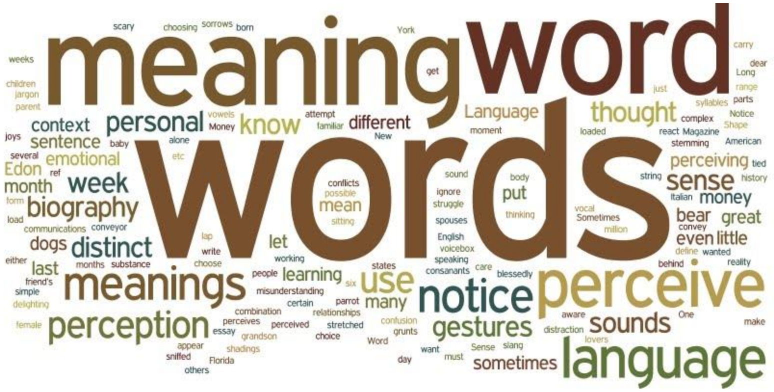
- Structuring a collection of text
  - Old approach: bag-of-words
  - New approach: natural language processing

- What is NLP?
  - Wiki definition: Natural language processing (NLP) is a field o computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages.
  - Goal
    - For computers to process or "understand" natural language in order to preform tasks that are useful e.g.
      - □ Performing tasks, like making appointments, buying things
      - □ Question answering – Siri, Google assistant, Facebook M,
  - Full understanding and representation of language

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

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

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science:
A Managerial Perspective, 4th Ed., Ch 5.

Examples

- The professor said on Monday he would give an exam.
- The chicken is ready to eat.
- Visiting relatives can be boring.
- "A lady with a clipboard stopped me in the street the other day. She said, 'Can you spare a few minutes for cancer research?' I said, 'All right, but we're not going to get much done.'"
  (English comedian Jimmy Carr)
- They are cooking apples.
- “cold” disease, temperature sensation, environmental condition?

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

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

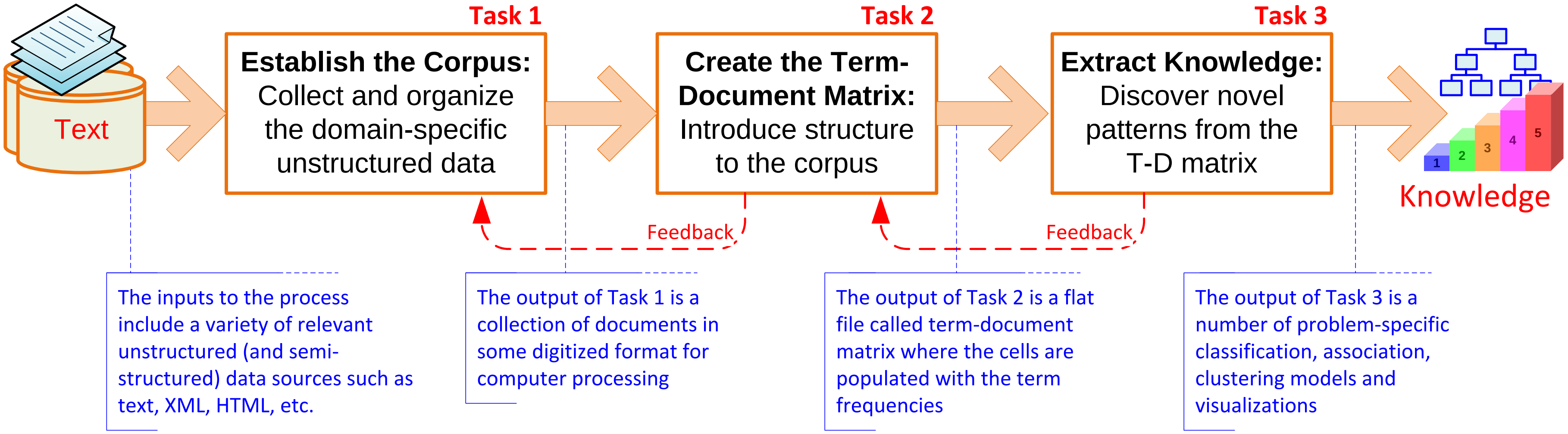
NLP Task Categories

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

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

Text Mining Process

- 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

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 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)

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

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

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

Text Mining Tools

- Commercial Software Tools
  - SPSS PASW Text Miner
  - SAS Enterprise Miner
  - Statistica Data Miner
  - ClearForest
- Free Software Tools
  - RapidMiner
  - GATE
  - Spy-EM

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

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!

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

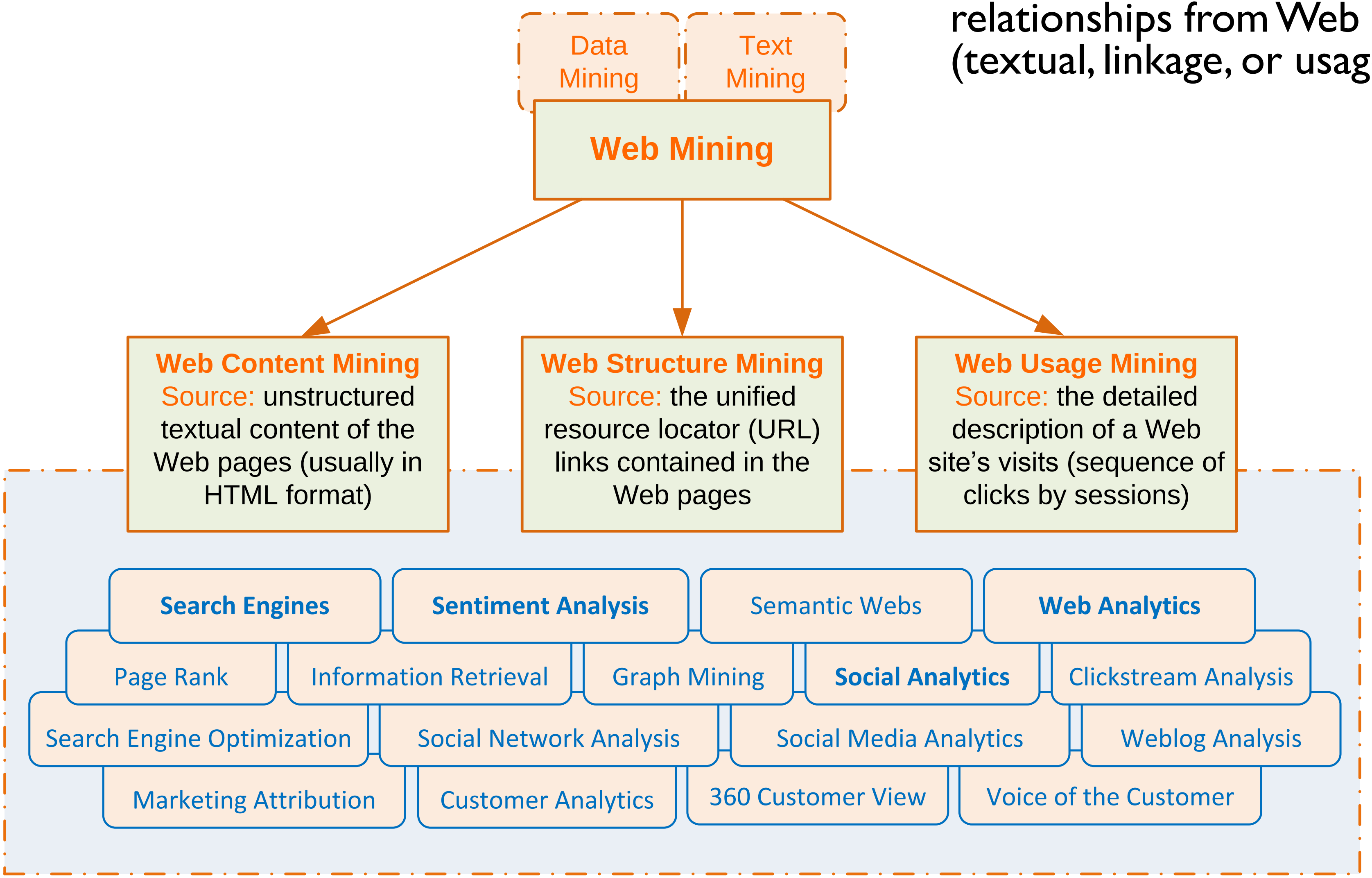
Web Mining- Uses

- There are three types of uses for web mining:
  - *Web structure mining*
  - *Web usage mining*
  - *Web content mining*

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

Web Mining

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

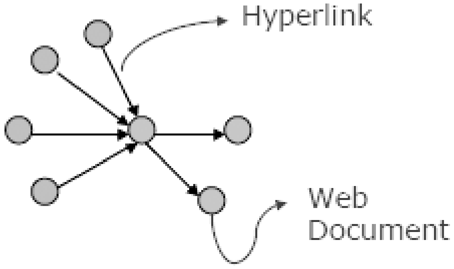


Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

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 Graph Structure

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

Search Engines

- A "search engine" is a key example for information retrieval systems, concepts and techniques in text mining.
- 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

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

Anatomy of a Search Engine

1. Development Cycle

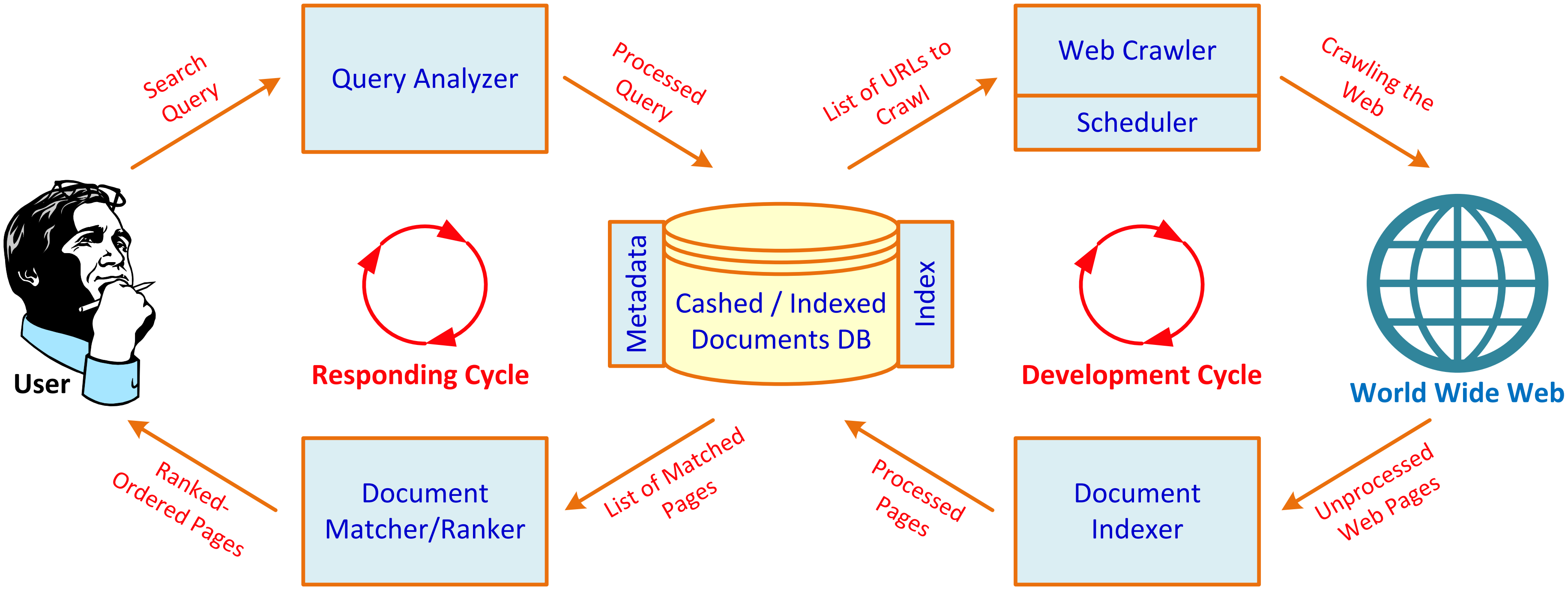
- Web Crawler
- Document Indexer

2. Response Cycle

- Query Analyzer
- Document Matcher/Ranker

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

Structure of a Typical Internet Search Engine



Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

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

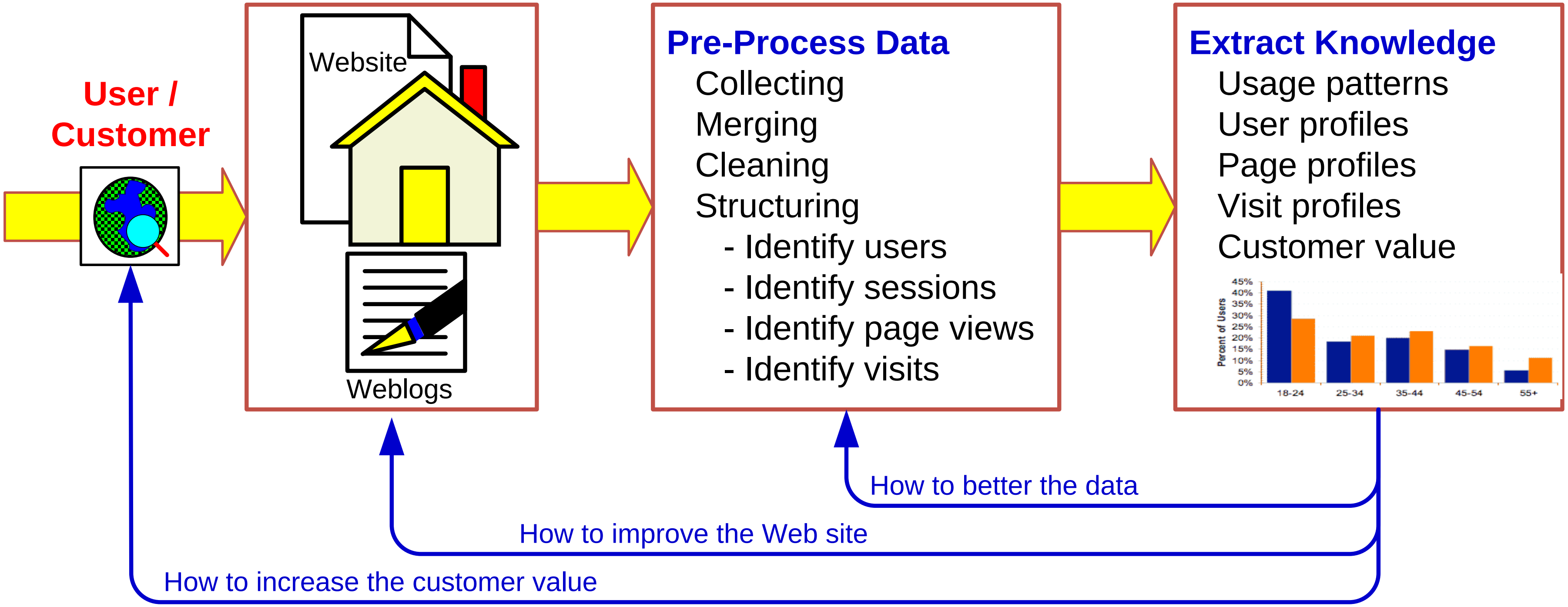
Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

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
  - …

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

Web Usage Mining (Clickstream Analysis)



User / Customer

Pre-Process Data
- Collecting
- Merging
- Cleaning
- Structuring
  - Identify users
  - Identify sessions
  - Identify page views
  - Identify visits

Extract Knowledge
- Usage patterns
- User profiles
- Page profiles
- Visit profiles
- Customer value

How to better the data

How to improve the Web site

How to increase the customer value

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

Web Analytics Metrics

- Web analytics programs can document a marketing campaign or manage the efforts of products and services
- 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?

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

Web Analytics Metrics

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

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

Web Analytics Metrics

Visitor Profiles

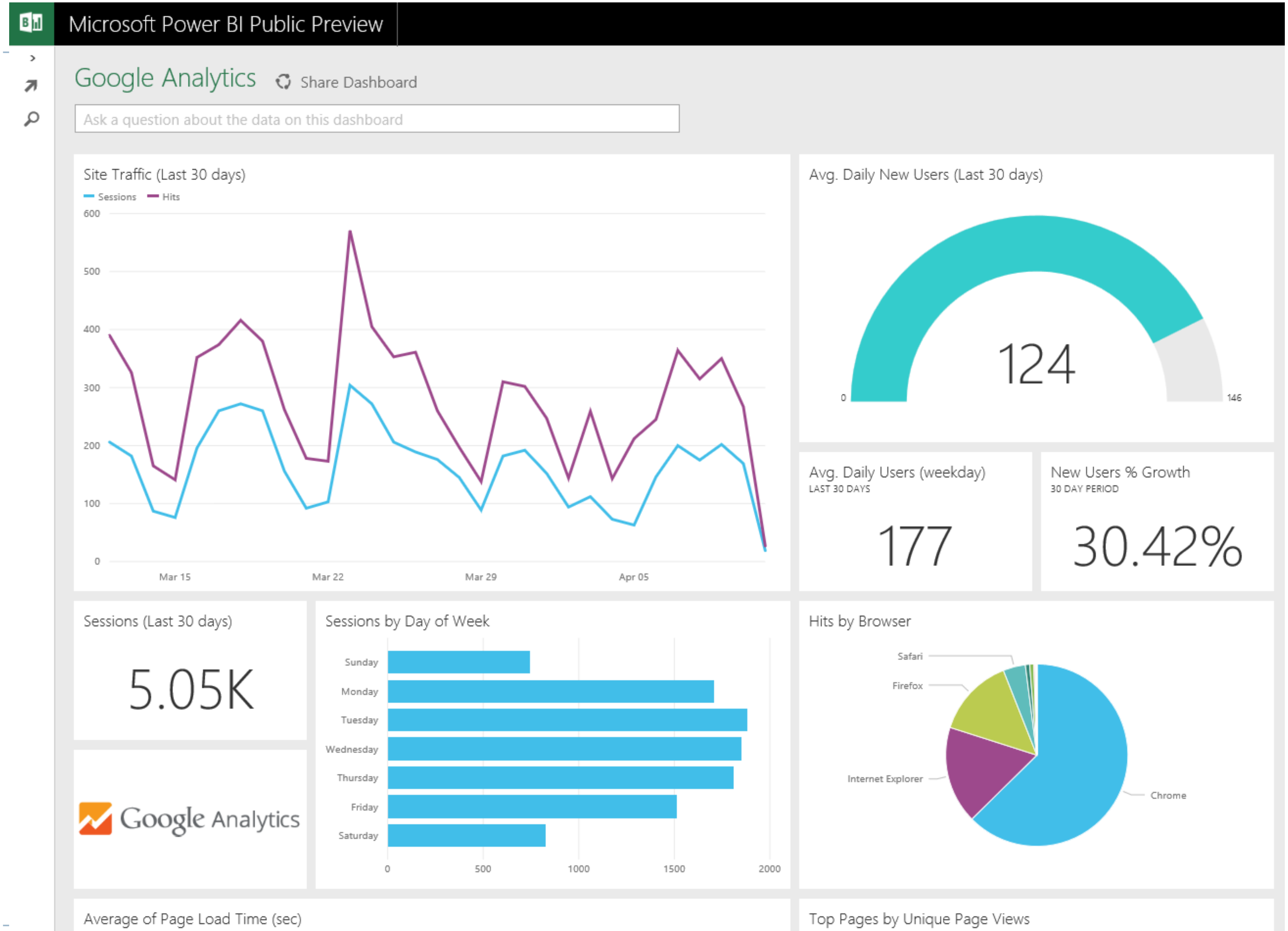
- Keywords
- Content groupings
- Geography
- Time of day
- Landing page profiles

Conversion Statistics

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

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

A Sample Web Analytics Dashboard



Microsoft Power BI Public Preview

Google Analytics   Share Dashboard

Ask a question about the data on this dashboard

Site Traffic (Last 30 days)

Avg. Daily New Users (Last 30 days)

124

Avg. Daily Users (weekday)
LAST 30 DAYS

177

New Users % Growth
30 DAY PERIOD

30.42%

Sessions (Last 30 days)

5.05K

Google Analytics

Sessions by Day of Week

Hits by Browser

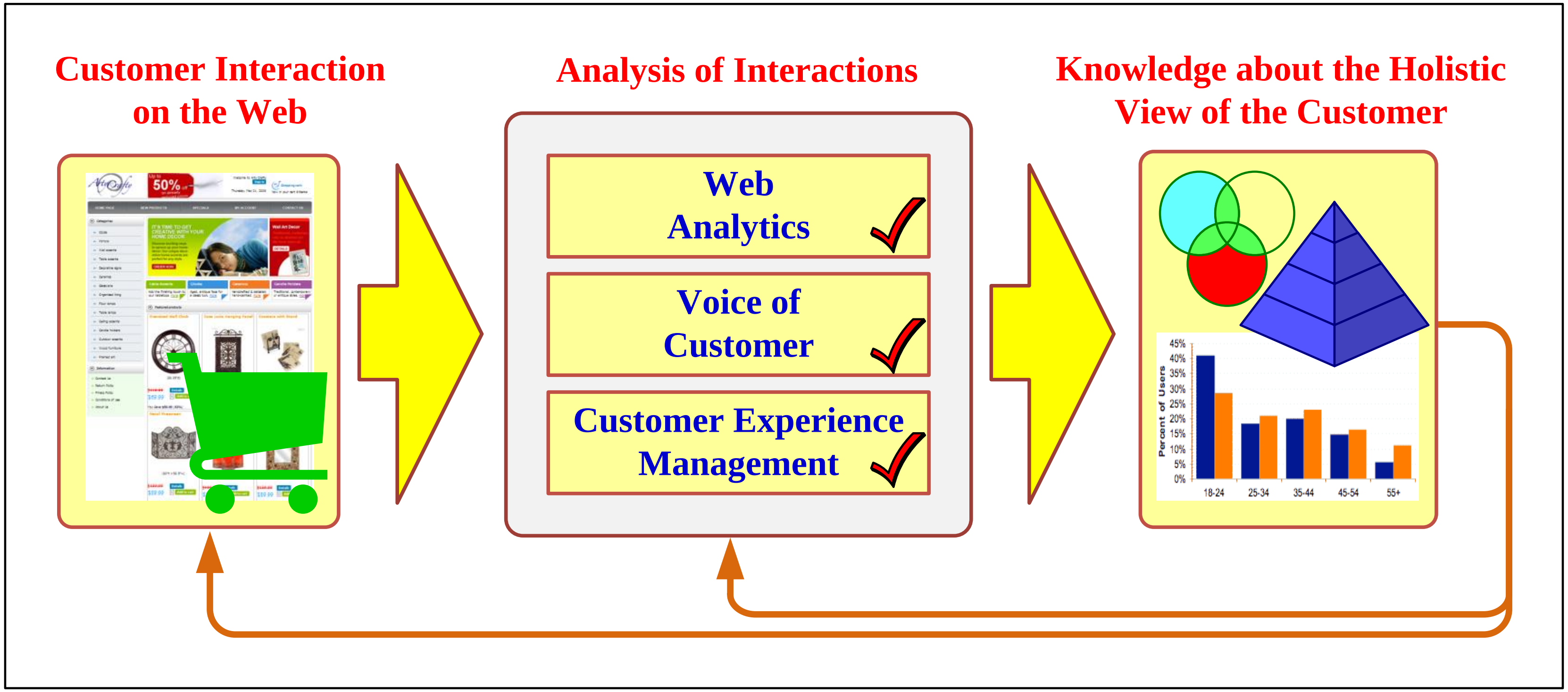
Average of Page Load Time (sec)

Top Pages by Unique Page Views

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.

Web Mining Success Stories

- Amazon, Google, Facebook
- Website Optimization Ecosystem



Customer Interaction on the Web

Analysis of Interactions

Knowledge about the Holistic View of the Customer

Web Analytics ✓

Voice of Customer ✓

Customer Experience Management ✓

Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5