

Hons. Degree in Computing

H4016 Text Analysis

The enrolment key for moodle is **miners**

Lecturer: Geraldine Gray

Contact details: geraldine.gray@itb.ie



office: E020
mail box: E6

Unit 1: Introduction to Text Mining

Objective for today

- 1) Overview of the course.
- 2) Understand what text analytics is.
- 3) How text analytics fits into CRISP-DM.
- 4) Review a publication on text analytics.

Learning outcomes

- Know what 'text analytics' is.
- Have an idea of the steps involved in mining text.
- Have an overview of what this course will entail

Overview of the course

Topics:

- ◆ Intro to Text Mining methodology, recap on CRISP – DM
- ◆ Pre-processing for text mining
- ◆ Mining algorithms
 - ◆ Classification
 - ◆ Clustering
- ◆ Big Data Analytics

Plan for labs

- 1) Work sheets on text analytics with Rapid Miner (4-5 weeks)
- 2) Text Analytics assessment

Assessment

- ◆ Exam – 60%
- ◆ Continuous assessment – 40%
 - ◆ Lab work on Rapid Miner - 10%
 - ◆ Labsheets, to be completed by the end of each lab
 - ◆ Text Mining assessment – 30%

Reading Material

- **Course notes**

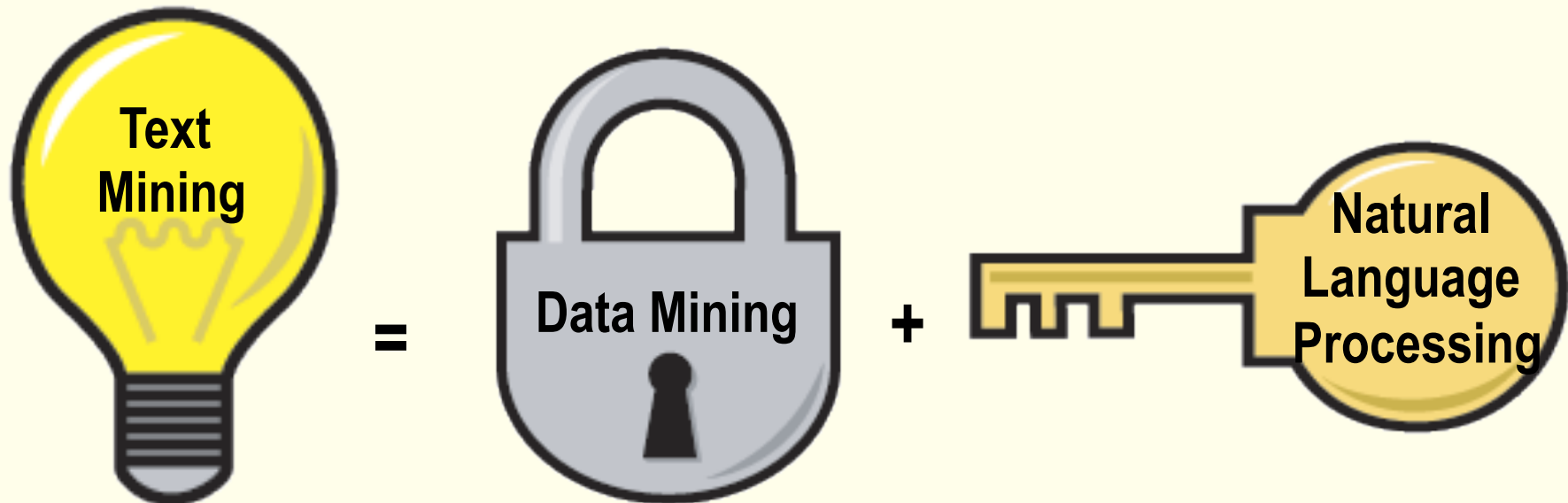
- Berry, Kogan (2010), Text Mining Applications and Theory. ISBN: 0470749822
- Manu Konchady, Text Mining Application Programming, Thomson, 2006. ISBN:1-58450-460-9
- Weiss et al (2004), Sholom Weiss, Nitin Indurkha, Frederick Damerau, Tong Zhang , Text Mining: Methods for Analyzing Unstructured Information, Springer, 2004. ISBN: 0-387-95433-3
- Berry (2004), Michale Berry, Survey of Text Mining, Clustering, Classification and Retrieval, Springer 2004. ISBN: 0-387-95563-1

Overview of text analytics

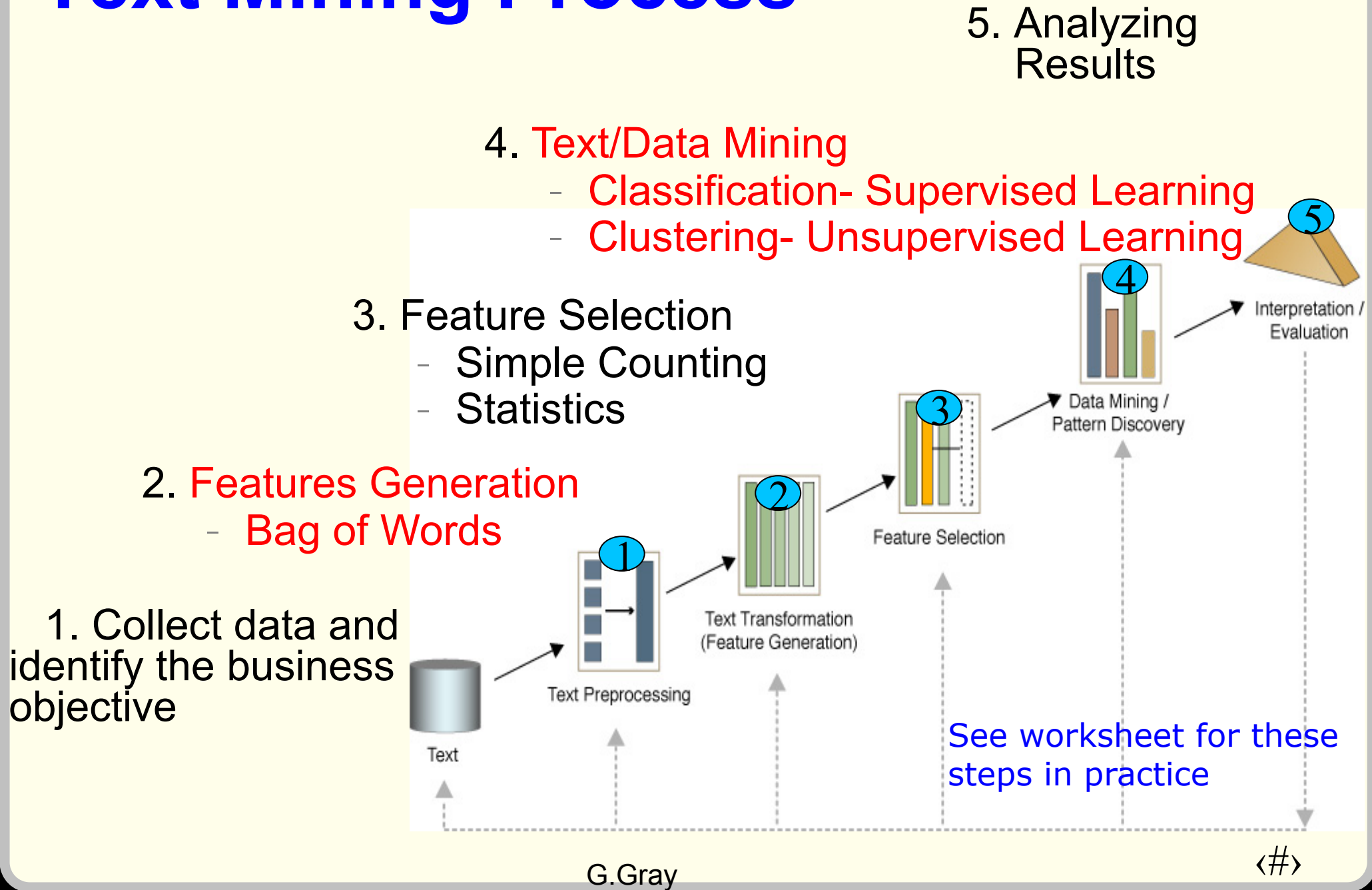
Text Analytics. . . Definitions

Many definitions in the literature:

- The **non trivial extraction** of **implicit**, **previously unknown**, and **potentially useful** information from (large amounts of) **textual data**”
- An exploration and analysis of textual (natural-language) data by automatic and semi automatic means to discover new knowledge



Text Mining Process



WORK SHEET

Text Characteristics

1. Large textual database

- Web is growing
- Publications are electronic

2. High dimensionality

- Consider each word/phrase as a dimension

3. Dependency

- Relevant information can be a complex conjunction of words and phrases.

4. Ambiguity

- Word ambiguity
 - Pronouns (he, she ...)
 - Synonyms (buy, purchase)
 - Words with multiple meanings (**bat** – is it related to baseball or mammal)
- Semantic ambiguity
 - **The king saw the rabbit with his glasses.** (multiple meanings)

Text Characteristics

5. Noisy data

- Spelling mistakes
- Abbreviations
- Acronyms

6. Not well structured text

- Email/Chat rooms
 - “r u available ?”
 - “Hey whazzzzzzz up”
- Speech

7. Authority of the source

- IBM is more likely to be an authorized source than my second cousin

8. Order of words in the query

- hot dog stand in the amusement park
- hot amusement stand in the dog park

Common text mining applications

◆ Analysis of social media

- ◆ There are vast stores in user opinion propagated daily on social media sites that are a powerful way to disseminate information from the ground up. Text analytics can automate the processing of this content into actionable information



◆ E-Mail and Call Center Analysis

- ◆ Need some way to categorise customer communications so that all customer concerns are addressed appropriately.
- ◆ **E-mail Filter**: using automatic classification of texts as 'junk mail' / channel e-mail queries to the correct person.

◆ Efficient and Reliable analysis of Open Ended Responses in Surveys

- ◆ In many surveys, open ended questions are asked. These responses need to be analysed. Text mining can help automate that task.



Common text mining application areas

◆ Discovery search / drug discovery

- ◆ Need some way to quickly review and analyze the information contained in thousands of scientific articles and patents to better understand the relationships existing between a large amount of genes.



◆ Competitive intelligence

- ◆ Need to monitor newsfeeds, trade press, patents and the Internet to watch for new product announcements, mergers, acquisitions, etc.

- ◆ Investigating competitors by crawling their web sites: automatically process the contents of Web pages in a particular domain. For example, you could go to a Web page, and begin "crawling" the links you find there to process all Web pages that are referenced.



Common text mining application areas

Analysis of interviews/conversations, e.g. patient diagnosis, report problems with cars etc. Notes of such interactions are generally recorded electronically. Analysis can give insight into common complaints and solutions.



Government Intelligence analysis

- ◆ Need to monitor huge volumes of unstructured information e.g. Web pages, NewsGroups, e-mails.



Examples from student projects . . .

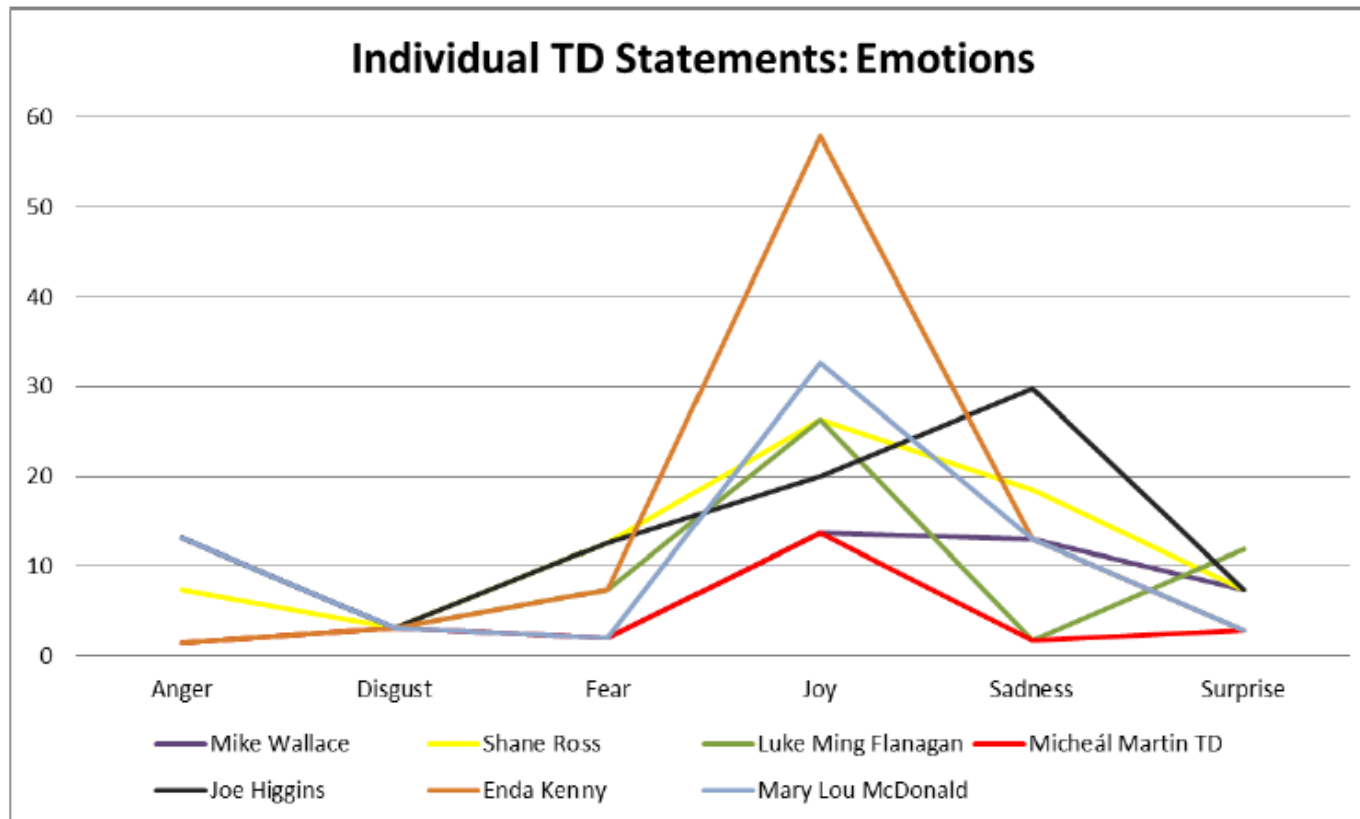
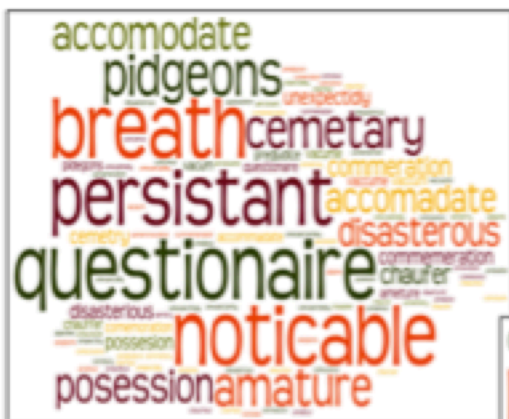
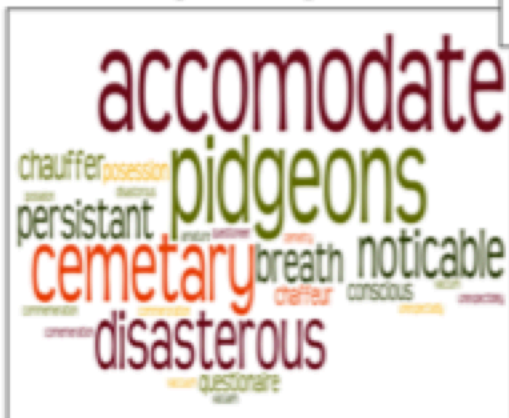


Figure 6.25: TD's Individual Statements: Emotions

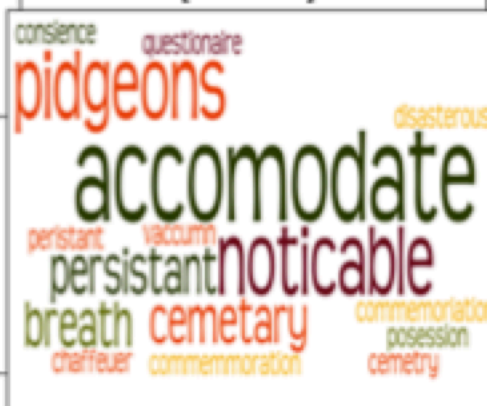




[100th – 120th]



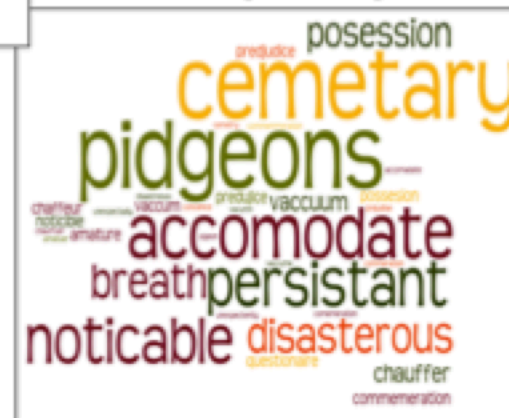
[25th – 50th]



[1st – 25th]



[75th – 100th]

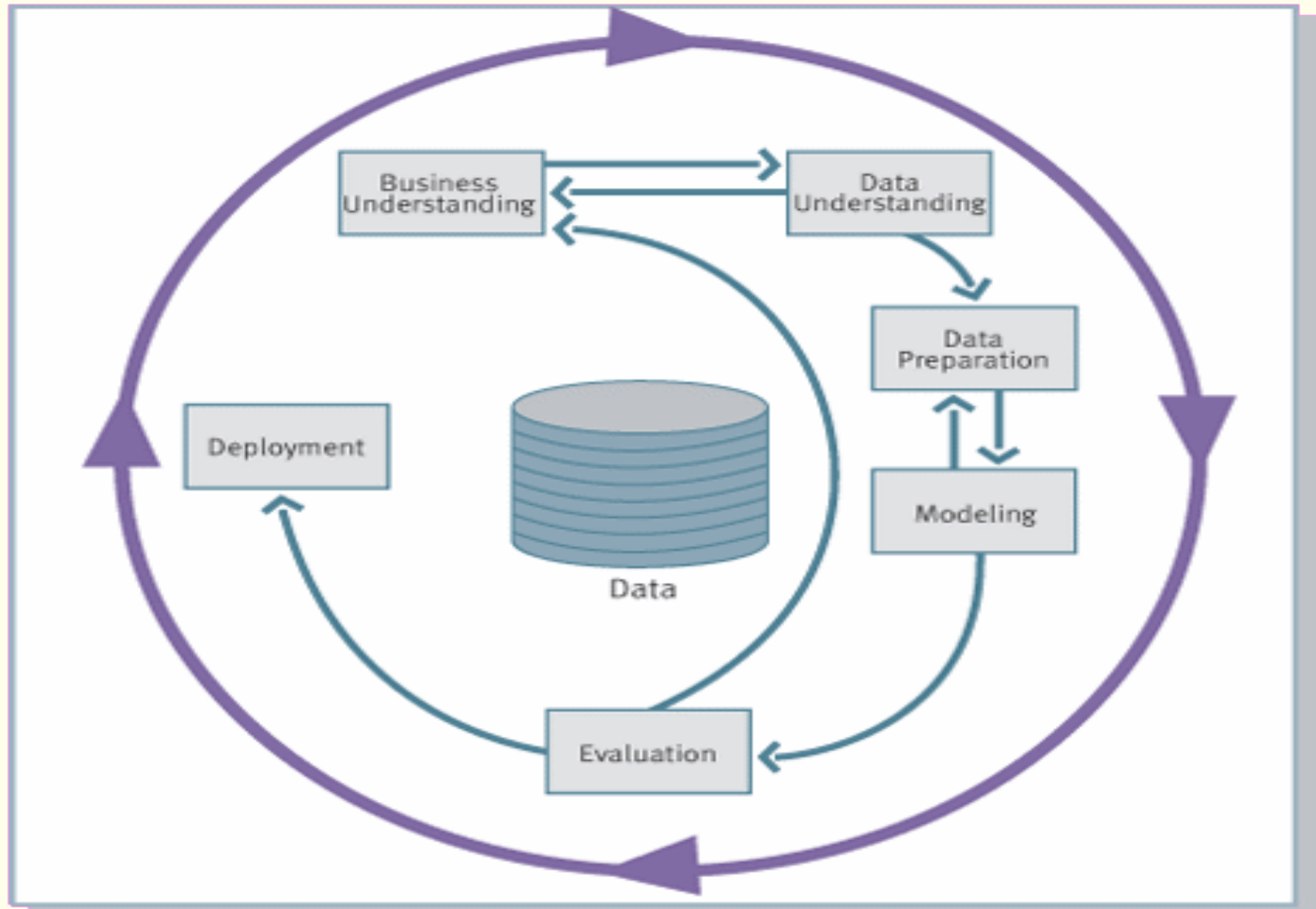


[50th – 75th]

“Hard” word
misspellings

CRISP-DM for text analytics

CRISP – DM for text analytics



Business Understanding

- ◆ As with data mining, this section includes:
 - ◆ Business objectives
 - ◆ Mining objectives
 - ◆ Project analysis (cost benefit analysis; risk assessment; resouces needed; assumptions)
 - ◆ Project plan

Example:

Business objective: Improve SPAM detection by 10%

Mining objective: Generate a predictive model that will classify an email as either SPAM or not SPAM.

Data Understanding

- ◆ Find the key concepts in the collection of document
 - ◆ Eliminate common words & stop words
 - ◆ Eliminate rare words
 - ◆ Identify phrases

Data Preparation

- ◆ Generate a document vector based on selected concepts (terms and phrases)

Mining

- ◆ Select the model
 - ◆ As per data mining; Primarily focus on
 - ◆ Classification
 - ◆ Clustering
 - ◆ Association analysis
- ◆ Generate test design
 - ◆ e.g. split documents into training set and test set
- ◆ Build model
- ◆ Asses model
 - ◆ Evaluate model accuracy

Evaluation & Deployment

- ◆ Assessment of data mining results with respect to original business objectives
- ◆ Project review
- ◆ Project deployment
 - ◆ Generate document vectors for new documents, and run the model.

Other topics on the course. . .

Big Data Analytics

- Introduction to the challenges of working with Big Data and Large scale file systems.
- Adapting standard analytics techniques to run in a distributed environment.

What is big data?

Enabling technologies
Hadoop & MapReduce
Hive
Spark
Mahout

Big Data analytics

Summary

- ◆ Recap on learning outcomes
 - ◆ What is text mining?
 - ◆ What are the stages involved in processing the data?
 - ◆ What will be covered in this module?