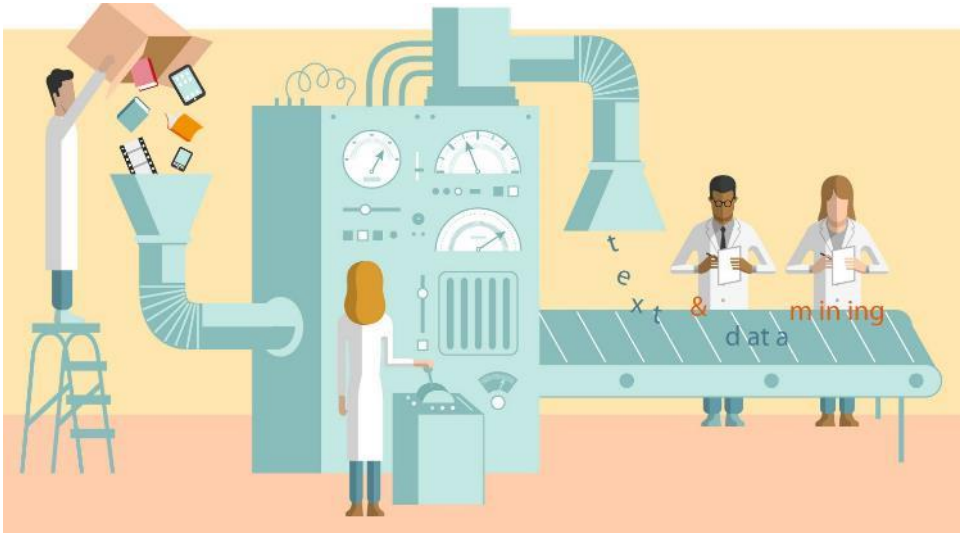
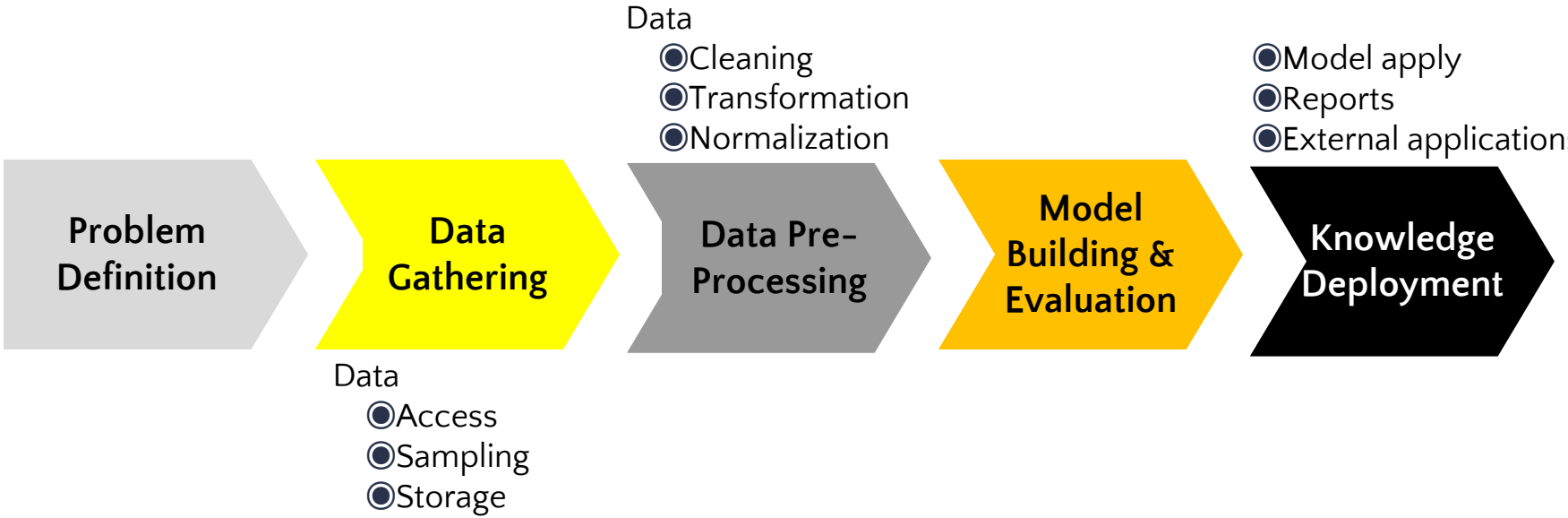
Topic 4: Basic representation of unstructured data



Juan Carlos Gómez
PhD in Computer Science
jc.gomez@ugto.mx
http://jcgcarranza.wix.com/juancarlosgomez
Office 314

General process of Data Mining



Group activity 2

Data Gathering

Data
- Access
- Sampling
- Storage

Individual work: **Collect** from 4 of your friends in Facebook 30 text posts (each, in total 120 posts), two men and two women. Save them in 4 files, with one post per line.

- Access → Personal/private/free
- Sampling → Friend
- Storage → File

File names:

myinitials_male_1.txt
myinitials_male_2.txt
myinitials_female_1.txt
myinitials_female_2.txt

For example:

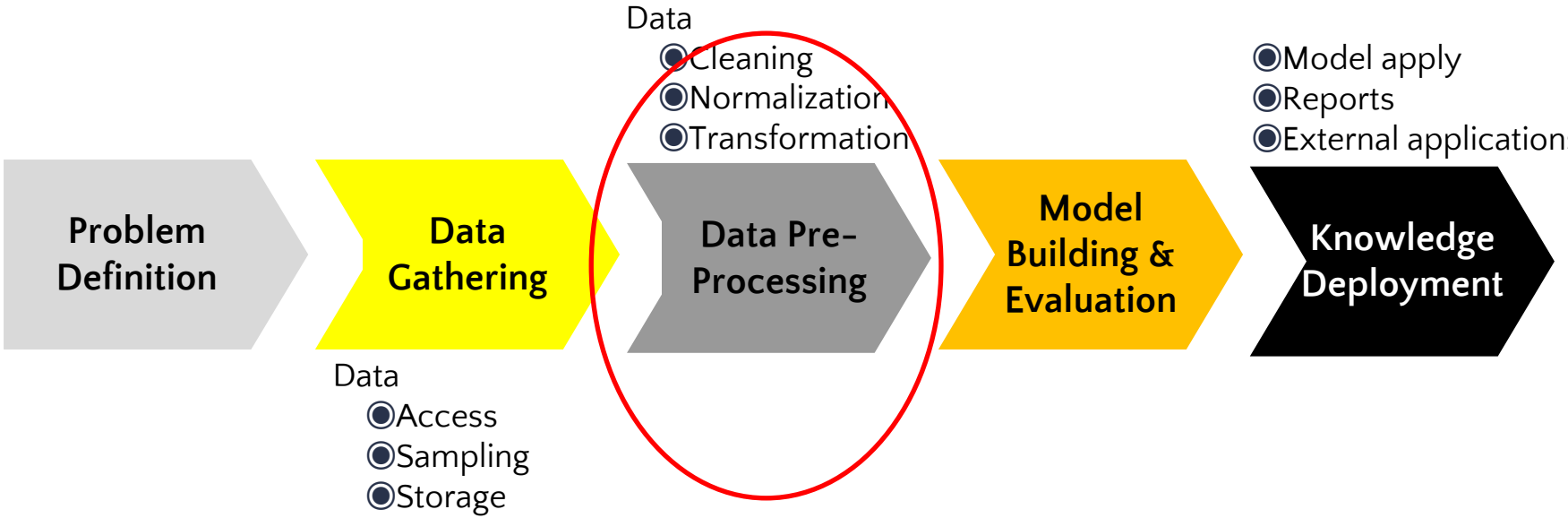
jcgc_male_1.txt
jcgc_male_2.txt

Problem definition

Problem Definition

Group users of social media based on their publications.

General process of Data Mining

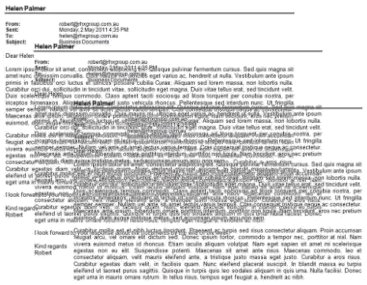


Structured Data

Name	Sex	Age	Height
Juan	H	19	1.70
Ana	M	20	1.65
Maria	M	18	1.56
Pedro	H	21	1.75

The **structure** is separated from the content. That means, each observation is an **instance** of a previously defined structure

Unstructured Data



Text

The **structure** is merged with the content. That means, each observation has its own structure

Data pre-processing: text data

1. Lowercase text
   - Be careful, sometimes uppercase has a meaning

   E.g. *A bite apple is the logo of Apple*

2. **Tokenization**: Split text in words
   - Define what is a “word”

   E.g. “*'When I'M a Duchess,' she said to herself 'I won't have any pepper in my kitchen AT ALL'. Maybe it's! 12$ 82% pepper $10.2 U.S.A.*”

Data pre-processing: text data

Commonly to tokenize:

- Split by space
- Split by regular expressions. E.g. [a-z]+
  \w+

Data pre-processing: text data

3. Remove **stop words**
   - Words that appear very frequently across almost any type of documents

   E.g. *of, the, from, to, …*

   - Be careful, sometimes they are useful to understand a sentence
   - **Commonly**: Use pre-compiled lists of stop words

Data pre-processing: text data

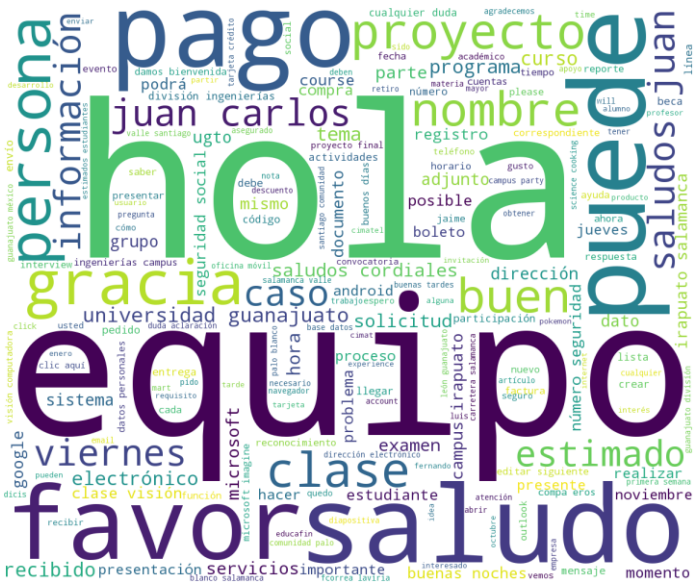
4. Remove other unimportant words
   E.g. numbers, mixed numbers of characters, words with punctuations or symbols, very long/short words, etc.

   - Be careful, some words could be meaningful

   E.g. "*C2HF2O2 is the formula of the trifluoroacetic acid*"

Data pre-processing: text data

- Example: Wordcloud



Credits

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival