ALGONQUIN COLLEGE

CST8390 - LAB BUSINESS INTELLIGENCE & DATA ANALYTICS

Week 3

LAB 3 - Data Preparation and Cleaning

Lab 3 – Data Preparation

PART I

- Preparing and Cleaning Data

PART II

- Steps
- Results

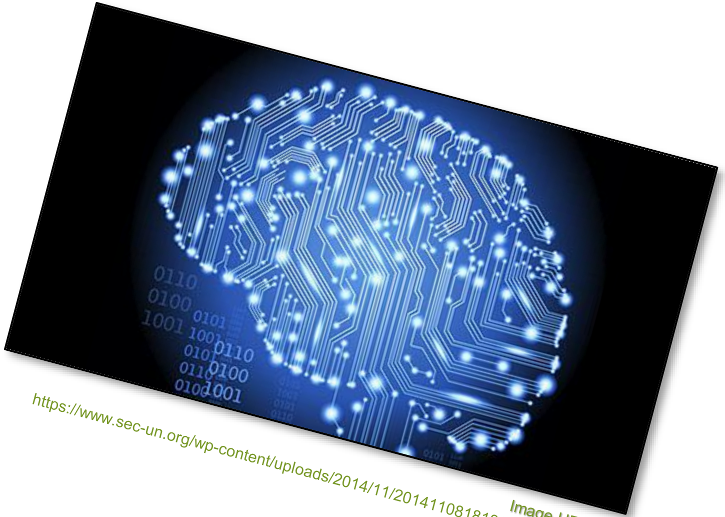


Image URL: https://www.sec-un.org/wp-content/uploads/2014/11/20141108181339_36006.jpg

CST8390 - Lab

Business intelligence & data analytics

Lab 3 – Data Preparation

Part I – Preparing and Cleaning Data

New Economy

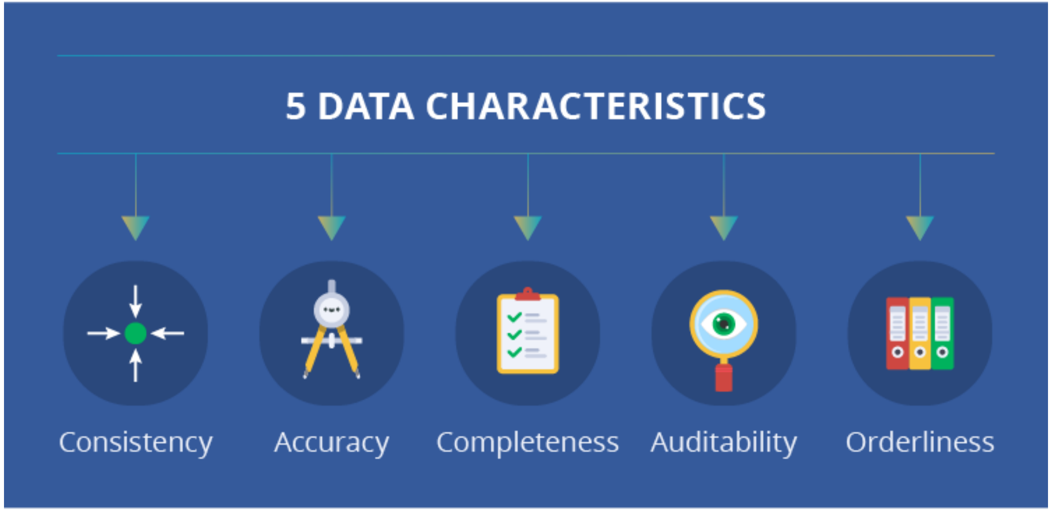
- DATA:
  - More important “value” of an enterprise / company;
  - It is part from the new economy.

So, it is important know how to deal with (process, tools) in order to get its real value.

Image URL: https://images.idgesg.net/images/article/2018/10/analytics_binary_code_network_digital-transformation-100777428-large.jpg

Remember: One important skill for business analysts is to know how to get, prepare and use the date

The “good” data



5 DATA CHARACTERISTICS

Consistency

Accuracy

Completeness

Auditability

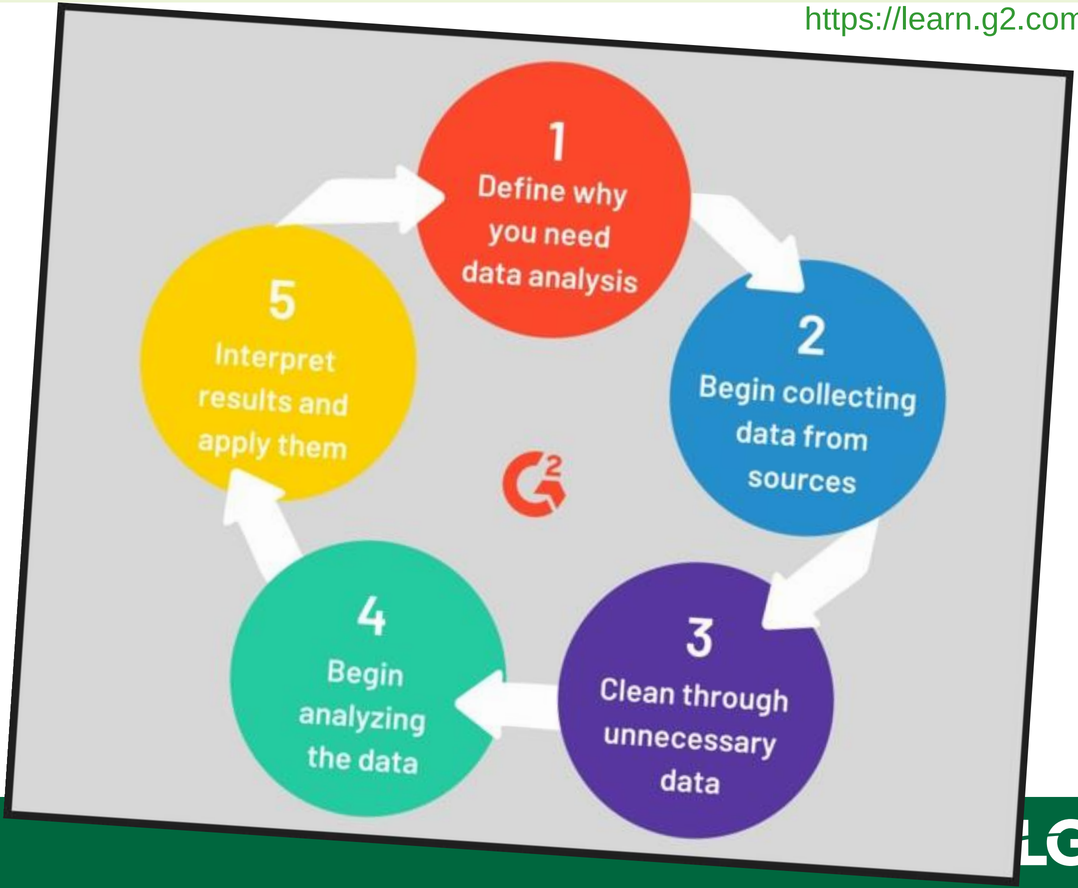
Orderliness

https://www.scnsoft.com/blog/big-data-quality

- Problems:
  - Inconsistences (duplications, contradictions, gaps);
  - Imprecision;
  - Lack of information;
  - Untracking data;
  - Impossibility of relation and comparison.

5 Steps of the Data Analysis Process

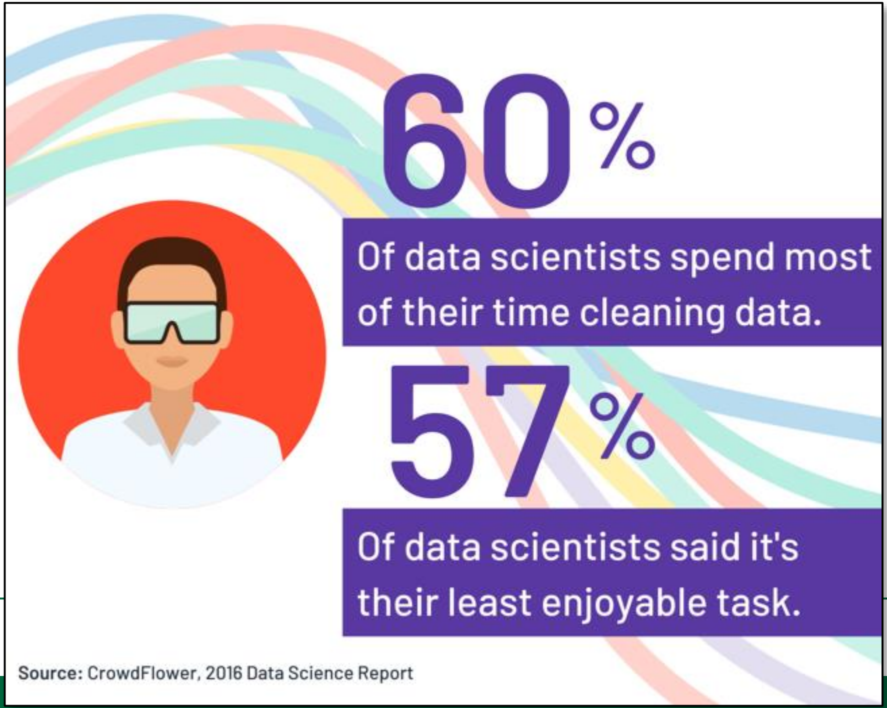
https://learn.g2.com/data-analysis-process


1
Define why you need data analysis
2
Begin collecting data from sources
3
Clean through unnecessary data
4
Begin analyzing the data
5
Interpret results and apply them

Preparing Data

https://learn.g2.com/data-analysis-process

- Some statistics…



60%

Of data scientists spend most of their time cleaning data.

57%

Of data scientists said it's their least enjoyable task.

Source: CrowdFlower, 2016 Data Science Report

Weka Introduction

Demo

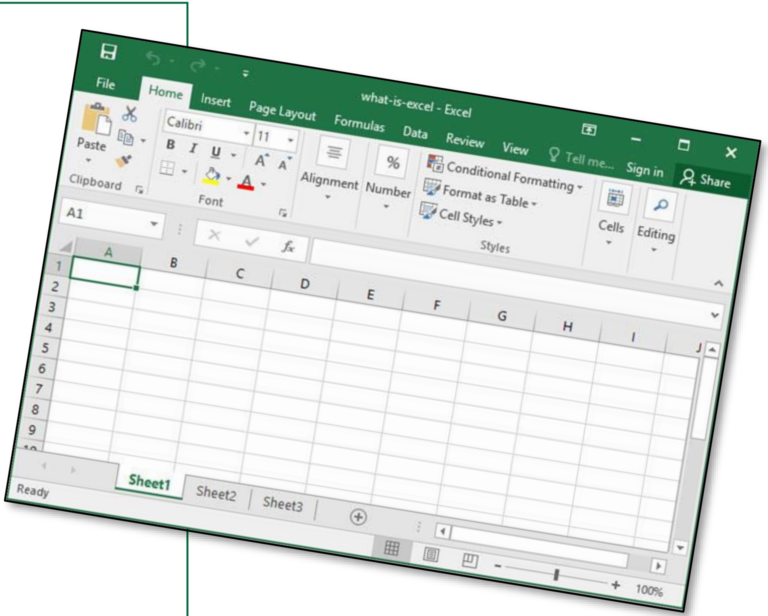
Step-by-step (A)

I. BASIC OPERATIONS

1. Download EmployeesSalaryBigFile.csv file from **Brightspace**;

2. Open EmployeesSalaryBigFile.csv in **Excel** and explore it;

3. Identify the **attributes** of the data. Record the attributes and the type of attribute for the data.



Step-by-step (B)

4. Load the **CSV** file into **Weka** by selecting ‘Open file’ in the ‘Preprocess tab’ (Select CSV data files for the file type).

5. Check different attributes including Branch. Branch is considered as numeric by default. Save the file as **arff** file by clicking on Save on the right corner.

6. Open EmployeesSalaryBigFile.arff file in **Notepad++**. Change the attribute types of first_name, last_name, email, address, Address and Branch with the required types. Save the file. (This can also be done by applying filters).

Step-by-step (C)

7. Open the file again in **Weka**. Check all attributes and their values.

8. How many instances do you have now? __________. ?

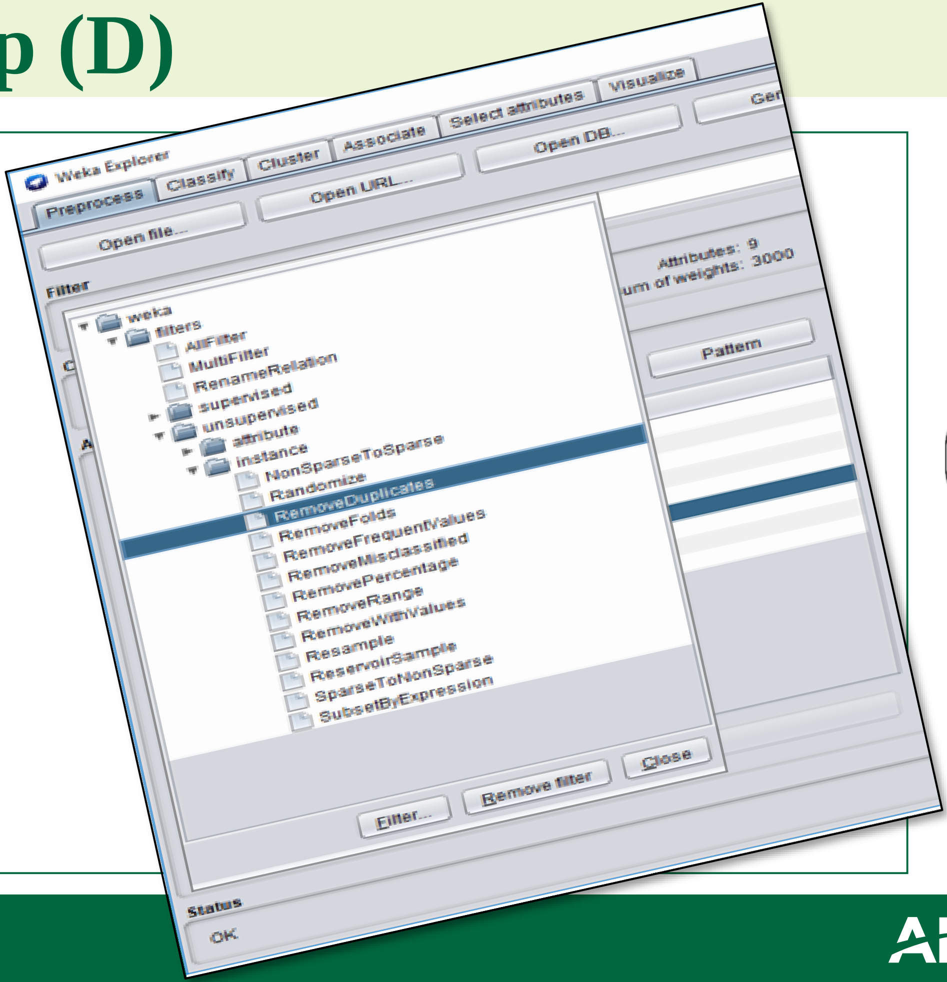
9. Take a **screenshot** and save it in a word document named Lab3.

II. DUPLICATES REMOTION

10. Check manually whether any duplicates exist in the file.

11. Now run RemoveDuplicates filter to remove duplicates. To do this, from ‘Filter’, (Choose weka > filters > unsupervised > instance > Remove Duplicates).

Step-by-step (D)



Step-by-step (E)

12. Select Apply to run the filter operation.

13. How many **instances** do you have now? ____________________ ?
    Duplicates: ____________________

14. Take a **screenshot** and paste it in Lab3 document.

15. Save this new file as EmployeesSalaryBigFileNoDuplicates.arff.

Step-by-step (F)

III. CONVERTING ATRIBUTES

16. How many nominal attributes do you have? ____________________

17. With those nominal values, we cannot apply any of the distance-based classification methods. Convert them into binaries using NominalToBinary filter. For that, from Filter, select Weka > filters > unsupervised > attribute > NominalToBinary, and hit Apply.

18. Take a **screenshot** and paste it in Lab3 document.

19. Save this file to EmployeesSalaryBigFileNoDupBinary.arff

Step-by-step (G)

20. Open the file in **Notepad++** and see the data.

21. Take a screenshot of the file while it is opened in **Notepad++.** Header should be visible.

In order to get the credit for this lab:
1. Show the screenshots of Q14, Q16 & Q21 (2 marks);
2. Show EmployeesSalaryBigFileNoDupBinary.arff in Weka (3 marks).

Remember: Include a minimal analysis in the end.

What about my analysis?

- The previous questions can help you to do your own analysis;
- For instance:
  - About the importance of transforming data (ex: nominal to binary, string to nominal, etc.) or removing data:
  - In which circumstances you should perform these operations and why?
  - Give additional examples

Image URL: https://www.marketingdirecto.com/wp-content/uploads/2018/01/ciencia-datos.jpg

Open questions…

- Before we start, do you have any doubt / question?

Image URL: https://cdn-
az.allevents.in/events3/banners/26b150363d5757da8578fa6a1481585a368f12d4247adfe95837e6ea
6c5ab2af-rimg-w1200-h549-gmir.jpg?v=1569691155

See you…

- Remember:
  - Labs require practice and it is ok committing errors and learning with them.
  - Do not forget to show your results…
  - Any questions, let me know…

sousap@algonquincollege.com

Image URL: https://thumbs.gfycat.com/MaleFrigidBull-size_restricted.gif

Thank you for your attention!