NOM BOUKHELOUA

Prénom Rémi

Promo Data

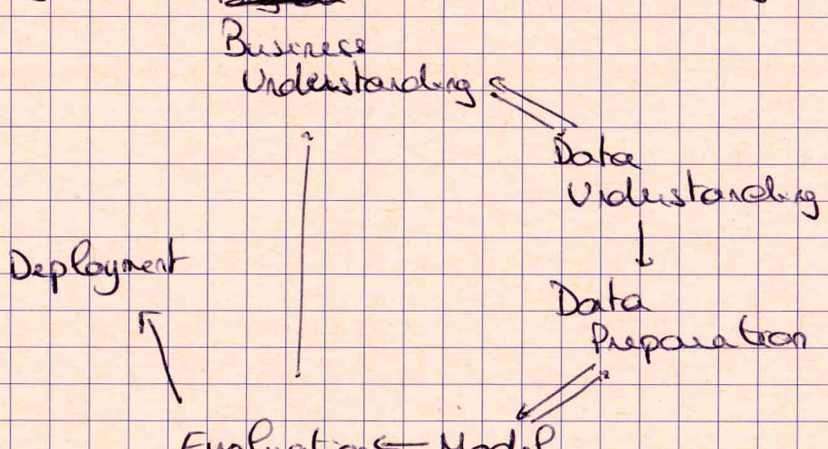
Date 16.04.

20150257: BOUKHELOUA Rémi
M1:
ST2DVA-DE (16/04/2019)
Amphi jaune

16

MATIÈRE Data Visualisation (1/2).

We followed CRISP DM method (Cross Industrial Standard Process for Data Mining (CRISP DM) method to do data mining. It follows this graph:



- Business Understanding is knowing what we need and why we need it.
- Data Understanding is knowing what data we acquired and what are the trends.
- Data Preparation is creating new features or arranging data according to the modelf we want to do.
- Modelf is creation of data mining model.
- Evaluation is testing our modelf to see if it fits business understanding
- And if it fits, we deploy it to make it usable by everyone

Other wise we have the SEMMA (Sample Explore Modify Modelf Asses) modelf which consists in:

- Sample: select useful data regarding our needs.
- Explore: looking for trends, visualize data, basics analytics.
- Modify: transform the data to make it usable for our model
- Modelf: create modelf regarding our needs
- Asses: Evaluate it.

Q3. Comparaison between Tableau, PowerBI, QlikView:

	Tableau	PowerBI	QlikView
Reputation	Mostly Used Tool.	2nd most popular	
Usage	Good Visualization tool	Good Data manipulation tool.	~~Good~~ to ok in both.
Price	Expensive over 35$/month/user	Less expensive (arround 10$/month)	Mid Range Price
Sources	Huge number of sources.	Most restricted number of sources.	
Limitations	Server's limit	10GB	Server's limit.
Speed	Efficient (RAM used)	Slower.	Efficient (RAM) used.
Others		• Produced By Microsoft • Lots of updates. • Good compatibility with Excel.	

Q5. It exists 4 main point of views bias that ~~could~~ we have to ~~check~~ avoid if we want to present the data rightly.

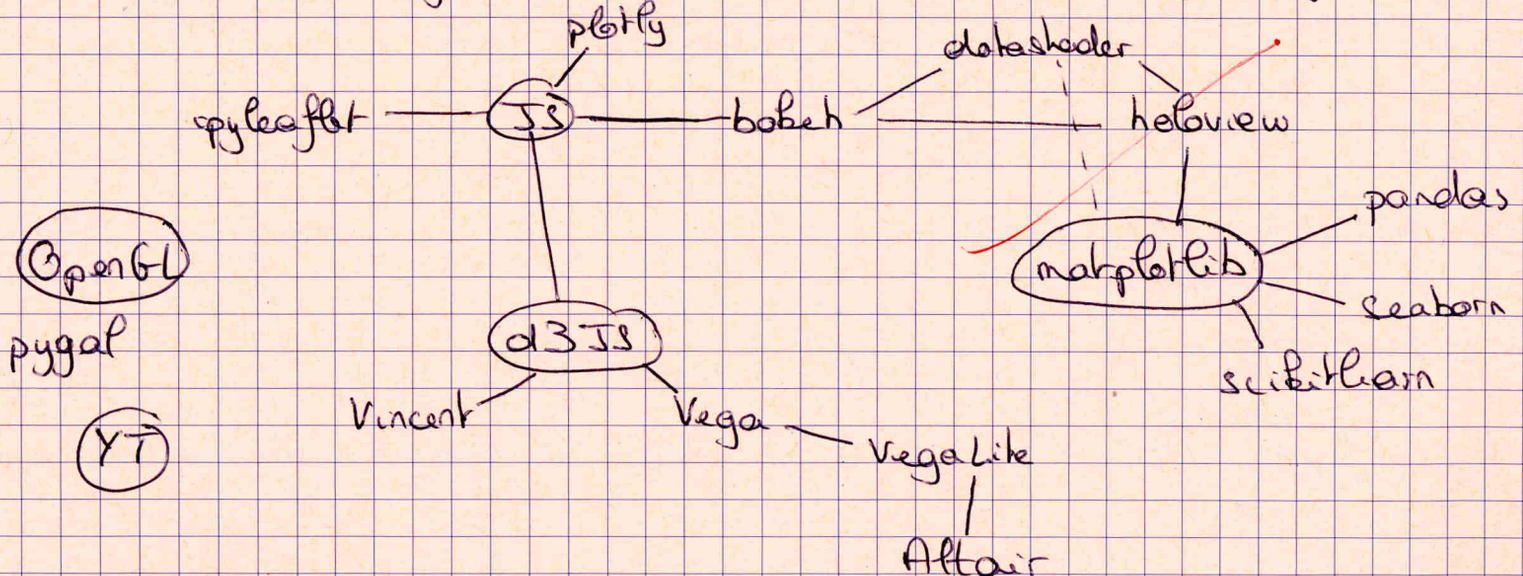
- Conceptual POV: if you find a correlation between two features, it doesn't mean they are correlated
- Perceptual POV: eventhough the dataset is complex, we have to make sure our representations are clear, not ~~overcharged~~ overloaded and straight to what we want to say.

- Graphical POV: we have to make sure our representations fits the reality. To do so, we can use the Lie Factor f to estimate proportions. Math sure of the size of objects or person such as the electors that get smaller and smaller (in % ages)

- ~~Political POV~~:

- Political POV: we have to be careful of any political orientation. For example, for one election, we could represent the winner ~~each~~ in every city over a map. But not every city has the same number of electors or senators. So it would be more accurate to represent ~~each~~ the winner per city when each city is sized regarding number of electors.

Q7. Some of most famous Python Landscape libraries ~~follows~~ for Data Visualization



Q2. ~~We used to follow~~ We follow this cycle:

Business Process Definition: where we state on our needs

↓

Research: where we observe what concurrents do and ~~what~~ decide what we want to do.

↓

Human Ressources Assessment: can we ~~support the~~ do this project? or do we need to outsource it?

↓

Data Acquisition: recuperating datas thanks to a crawler for example

↓

Data Munging: transforming unstructured data to structured ones.

↓

Data Storage: storing almost raw data

↓

Exploratory Data Analysis: exploring datas, looking for trends.

↓

Data preparation: preparation for modeling, selecting features, creating new ones.

↓

Model: creating the Model

↓

Implement: deploy the model.

Q4. Nowadays, Big Data can be described through 3V: Volume, Velocity and ~~Vaeu~~ Variety. Data Visualization faces those three V as challenges: We need to process more and more data ~~from~~ with different format but ~~need~~ also need to represent them as we used to do. Having more and more ~~axes~~ parameters, but the same number of axes. So we need ~~way~~ libraries that can support a lot of data and a lot of formats.

At the same time, we ~~need~~ want to have a view of how our product is ~~developping~~ used. So we need realtime view such as the Singapour realtime representation. Tools like however Spark notebook allows us to do so but still not very confortable to use for majority.

NOM BOUKHELOUA

Prénom Rimi

Promo Loko

Date 16.04.

MATIÈRE Data Visualization (2/2)

Q6 Big Data Visualization contains various styles such as Data reduction where you try to aggregate many datas in a simple manner and a simple view. For example, ~~nothing~~ you could make a network between main caracters of Bibli and mostly used words.

Moreover, there is one where you write try to not only ~~expl~~ show facts but also explain it. E. using them heatmaps and PCA. Be careful of Conceptual POV!