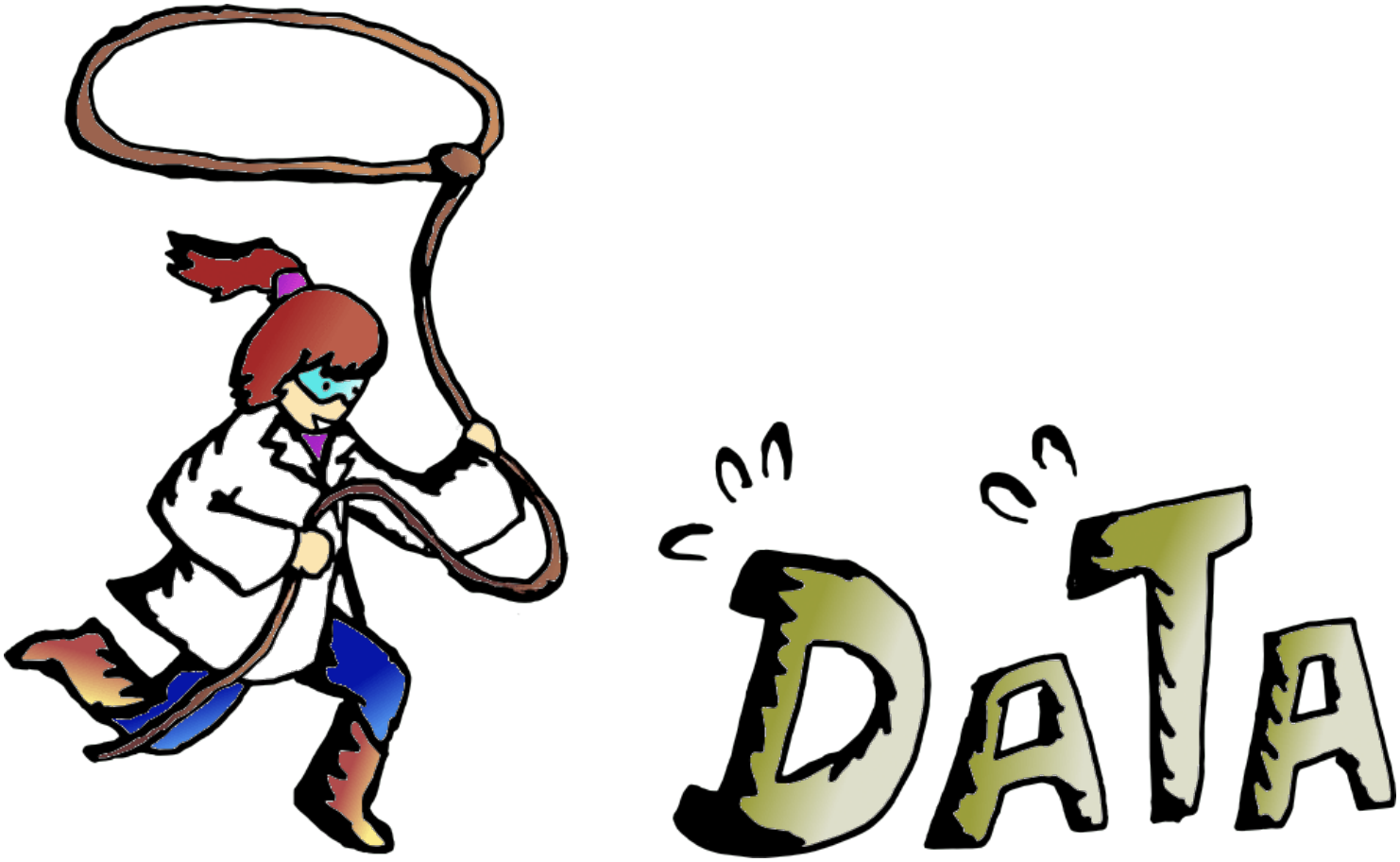
WeRateDogs Twitter data

Data Wrangle Report

Svajune Klimasauskaite - 5 July 2019



“If you torture the data long enough, it will confess”

I’m not a Twitter fan, nor do I follow the tweets, nor do I have an account in Tweeter, therefore, initially, I didn’t have a strong relationship with this dataset, until I started a data torturing process and start receiving some interesting findings … And those findings bring an addiction feeling with never-ending questions “and what if”…

What was stopping me from quickly answering all my “what if” questions were the following skill gaps:

- Business process knowledge, i.e. Twitter,
- Business domain knowledge, i.e. WeRateDogs,
- Pandas library knowledge.

I believe that data analysis without a business value is just a waste of time, therefore it’s very important to understanding the business so that I could focus on the data that matters. Initially, I was feeling lost among all these data without having a clear vision of what I’m searching for and a clear understanding of what is the meaning of the data to the business. Let me elaborate on all these stoppers:

- Process knowledge: is retweeting the same as share? How many users are allowed to tweet? What is dogtionary? What is the rating and if it has a scaling limit? What is the structure of the text? What did it mean a confident level in the prediction algorithm, i.e. a confidant that it is a dog or a confidant about the dog breed? So all these questions need to be clarified before diving into the data.

- Domain knowledge: when I wanted to test manually the outcome of image prediction algorithm against the images, I was not capable to say if it is correct since I don’t recognise dogs by the breed by myself. Domain knowledge gap stops you from having a quick manual check and quick clarification if you are on the right track.

- Pandas library: helps you to have quick wins with little effort. Understanding and knowing Pandas functionality can help you achieve data wrangling process in the most efficient way (by saving time and money). Of course the same result can be achieved in writing much more lines of code, however, I like the saying that IT professionals must be a smart worker and not hard workers, so (in this case) writing a lot of code is not an option.