**123…**

**Hello everyone,** I am Carlos Carrasco and today I'll take you on a tour through the data analytics process. Before starting though, we must define what is a workflow!

**If we just look for workflow at the dictionary** , this is defined as the sequence of steps to complete a working process from beginning to end. Cain and Haque, 2008 extend this definition to include also the people and resources needed to complete that goal.

**The objective of a workflow** is to define the main components of a project cycle in a logical order so the waiting time between tasks is reduced to a minimum. Some of the benefits of following a workflow are:

*Improving operations* because a workflow ensures that a job will be completed in order from beginning to end.

The analysis of the working process help to identify redundant tasks.

It also helps to streamline a process, because the knowledge of the work that has been done and what is yet to be done helps to reduce the time between tasks.

Similarly, this helps to minimize mistakes, especially by forgetting to do something, but also if something goes wrong it is easier to identify where a problem might have happened.

Finally, keeping a consistent working methodology helps to assure reproducibility between project cycles.

**The Scientific Method** is a well-known example of a workflow. It describes a research process from beginning to end, startingwith a question or observation from which we create a hypothesis and the design an experiment to test it. Then, the data obtained during the experimentation cycle is analysed and evaluated to determine if the hypothesis is accepted or not.

Of course, a workflow is domain-specific and might describe the logistic chain to deliver a product or a mineral exploration cycle. However, in the rest of this video, we will focus on data analytics workflows.

**Wickham and Grolemund (2016)** summarize the data analytics cycle as Import or obtain the required data for a project followed by tidy which consider cleaning and wrangle of the datasets into a machine friendly format before a continuous cycle of exploration that considers transformation or feature engineering of the existing variables, visualization and modelling methods to extract knowledge from the data prior communication of the findings.

The exploration cycle makes a difference to other disciplines since this is an iterative process that might be repeated many times.

**Several workflows have been developed for data analytics.** The 2014 KDnuggets poll on methodologies shows that nearly 60% of the respondents use standardized methodologies like CRISP-DM, SEMMA, or KDD process. So, in the next few slides, I will give you an overview of the most used standardized workflows.

**In an increasing complexity,** they are SEMMA, OSEMN, Blitzstein and Pfister, KDD, and CRISP-DM workflow. Despite being different methodologies, you’ll notice that all of them share the same core steps.

**The SEMMA workflow** was developed by the SAS institute. It stands for Sample or collecting the required data, Exploring as the exploratory data analysis, Modifying for feature engineering, Model for fitting different models, and Assessing to evaluate the results of the models.

You can see that this workflow focuses on the modelling process rather than the knowledge that can be obtained from the data.

**The OSEMN workflow** is similar to the SEMMA. It was proposed by Mason and Wiggins in 2010. It stands for Obtain the data, Scrub as the data cleaning and preparation process, explore, model, and interpret. In contrast with the previous methodology, it includes a interpretation phase that attempts to get insights to make decisions or report findings. Mason and Wiggins mention that the scrub phase, is by far the most time consuming, but that the interpretation, despite being the least technical, is the most important.

**The next workflows are taught in the CS109 c**ourse by Blitzstein and Pfister at Harvard University. The core of this workflow consider getting the data, exploring it, and modelling it, however, this workflow start by asking a question that can be answered through data, in other words, it set objectives changing the focus from the modelling process to the problem for which you want to acquire knowledge to solve.

**The Knowledge Discovery in Databases process was proposed by Fayyad and collaborators in 1996.** It is considered a predecessor of the CRISP-DM. In contrast with the previous methods, KDD starts with the understanding of the domain-specific information required to solve a problem and setting objectives of what needs to be achieved. However, the core is still the same, including importing, preparing, exploring, and modelling the data, specified as data mining in the diagram. In KDD the feature engineering, depicted as transformation in this diagram, is a defined step, whereas previous models, except SEMMA, do not consider this important stage. The final objective of KDD is to obtain knowledge that can be used for decision-making or just be documented and reported.

**Finally, the Cross Industry Standard Process for** Data Mining or crisp DM the most used workflow in data analytics. It considers 6 stages, and it is very similar to KDD, including the understanding of domain-specific information and goal-setting at the beginning of the project. The main difference with the previous workflow is that CRISP-DM includes data understanding stage which is an exploratory data analysis before preparing the data, which helps to assess the quality of the data, if it is adequate for the question to answer or decide if more data is required. The core of the workflow is still the same, and include a continuous interaction between feature engineering and modelling.

During the evaluation, the goals assessed and depending on the results, They can be redefined by starting the cycle again or moving into the deployment stage that considers how obtained knowledge will be transferred or communicated.

**If we go back to the process described by Wickham and Grolemund**, we can observe that all these workflows have the same core structure, that considers a continuous exploration Cycle. Some of them like the OSEMN or SEMMA focus on the modelling process while KDD, CRISP-DM, and Blitzstein and Pfister add domain-specific knowledge into their workflows. I would like to highlight the non-linearity of all these workflows, and many stages need to be revisited, every time with new insights.

**Finally,** it is overwhelming the number of different methodologies available for data analytics, and it is not possible to say that there is a workflow better than others, but remember they all share the same structure at their core. We are all different people and there might be methodologies that work better for some than for others. If you remember the KDD nuggets poll, many people prefer to use their workflow! But something we could agree on is that having any workflow is better than not having one at all.

If you use one, your future self will thank you.

**If you are interested in how this presentation** was rendered you can access it at the following GitHub repository, including notes, script and references.