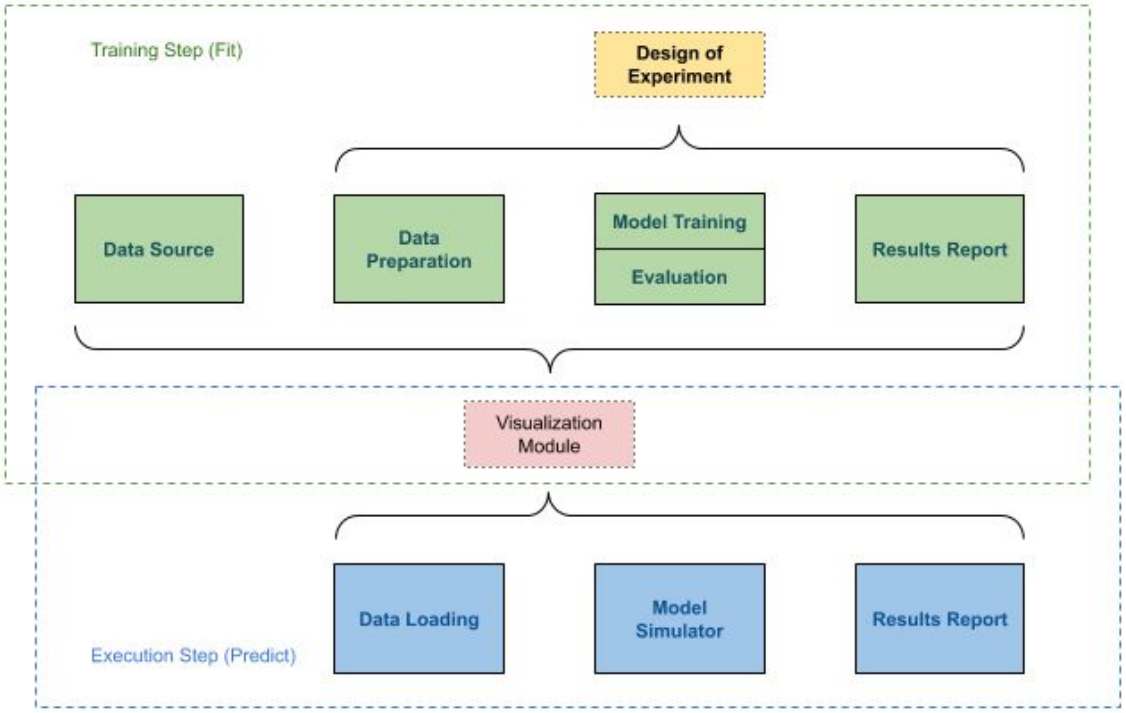
Description of the MLDD Main Workflow (rsmt)

The context of this project is to accomplish the lower time demand answers in the field of oil/gas reservoir simulation, especially in the tasks of planning and prediction of the reservoir state and production. This document intends to describe the main purpose of the Reservoir Surrogate Modeling Toolbox (rsmt). A tool that follows the principles of a Machine Learning Data-Driven (MLDD) problem approach and solution.

As an ordinary machine learning (ML) problem usually it is necessary a set of data (as much as possible) to represent all the space of the problem and its nuances, a model that relies on a data structure and associated algorithm to adjust it and later to run different prediction instances, and to finalize a way to visualize the obtained results and some metrics to evaluate all the system.

The rsmt approach split this process into two main steps namely “Training Step (Fit)” and “Execution Step (Predict)” as displayed in the following diagram.



Surrounded by the green dashed line, the first step aims to manage the base process of define, structure, train and evaluate one machine learning model. It is done based on a set of data, respecting defined constrain or special metrics, and give as a result a useful piece of code ready to be incorporated as a data-driven (DD) solution.

This first step is composed of four sub-steps and one connected tool to manage the set of experiments in order to achieve some result objective, usually implemented as a Design of Experiment (DoE) tool (depicted in yellow in the diagram). The sub-steps are:

- Data Source - related to the set of tools or modules that give the user the essential resources to define, demand, retrieve and make accessible enough data to represent on reservoir environment, with all necessary related data like reservoir characteristic parameter (saturation, permeability, porosity), plan of development, well positions, constraints, etc.
- Data Preparation - a sub-step related to the selection and structuration of a subset from a specific set of data. It needs to define and deliver a set of datasets able to be used by an ML model, including a training set, test set, and when required a validation set.
- Model Training / Evaluation - As the central point of the step, this sub-step intends to define, training and evaluate a machine learning structure guided by a learning algorithm, using the data provided, until reach some defined objective in relation to a metric of accuracy. It needs to be stored in order to be useful in the next step of execution.
- Results Report - Implemented as a tool, it needs to be connected with all preceding sub-steps in order to present the important and demanded data, related to each individual task, essentially in text way and to be accessible as independently and at any time.

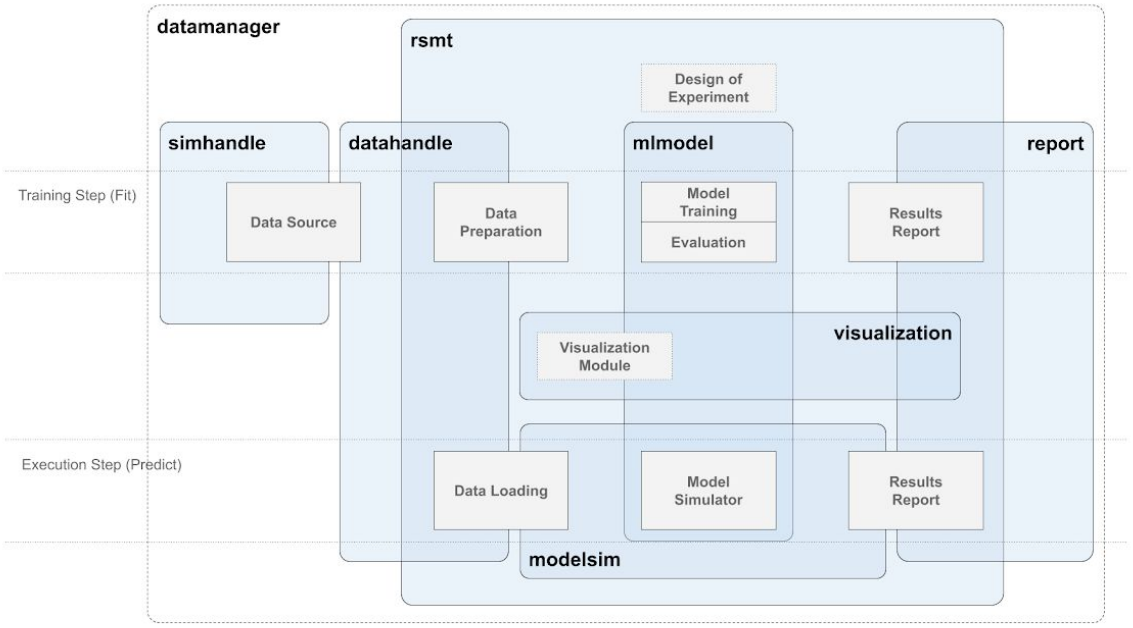
As a connected intermediaries utility, the “Visualization” Tool (depicted in pink in the diagram), is supposed to give the user all possibilities of visually access the state of any sub-steps belonging to both steps of the rsmt, including reservoir visualization, dataset view, training curves, error-value caparison, and so on. To do that it has also an independent module to generate all kinds of images, animations, and output flow related to doing that.

The second step, included in the blue dashed line, is related to the process of perform the simulation process in substitution of the real simulator, fed by the data description of the initial state of the environment (reservoir), the proposed development plan, and of course all the necessary parameters like wells and flow controls. The sub-steps within this part are:

- Data Load - A set of functions or parts of a module able to load, from a specific dataset, a chunk of data used to run on reservoir simulation, its development plan or part of it.
- Model Simulator - a module capable of rum an ML model with a predefined set of data (representing the reservoir state and the plan of development) and execute the step simulation in order to obtain the future reservoir state and production level measures.
- Results Report - the same as aforementioned, but in respect to the execution-only sub-step.

With this description in mind, a package with a set of modules was designed and planned, in order to achieve this workflow objective. The mindset behind the design was mainly based on the component view of each workflow step, in a way that one could go and back through the modules, not necessarily in a given order. The execution of each component just needs to accomplish its own constraints, for e.g., one couldn’t execute a model training sub-step if he hadn’t loaded the data in advance, and so on.

Eight (8) modules were projected as depicted in the following diagram comprehending the functionalities of some part in the main rsmt workflow.



The modules are intrinsically intersected within each other but with their own unique purpose. The **rsmt** module is the principal in the package, it connects with and demands data and functions from/to the others, and will be detailed in the end.

The **simhandle** module is responsible to interact with the oil and gas reservoir simulator. Through it, it is possible to define a generic reservoir specification (dimension, actual state, and petrophysical properties), a set of wells and its attributes and a plan of development (duration, step size, etc.). With that, it is possible to generate the reservoir simulator inputs, manage the execution process, retrieve the output and incorporate it. With all this data the module is able to feed the Data Source storage container (data and metadata).

Following the training step flow, the **datahandle** module has the objective to create, using the output of the simulator, data structures to feed both training and execution steps. To the first, deliver a dataset training with input/output samples. To the second, a data input well structured to feed an already trained ML model.

The **mlmodel** relies on to create, training and deliver ML models able to execute the data-driven surrogate model (proxy model) for the reservoir, based on the input and output given. It delivers the final result objective of this package. These models could be used later by the **modelsim** module and shared with other users. The modelsim module has the ability to execute ML models created by the mlmodel package, through the toolbox, in order to predict a reservoir state over some steps of simulation.

Besides it is not directly pictured with intersections with all, the objective of the **report** module is to serve as a general reporter to all main actions executed by all other modules.

And more focus on the space of the rsmt, the **visualization** module is aimed to provide graphical and plotting outputs to the user, e.g., reservoir state for each property, production curves, etc.

The two last modules are the main module **rsmt** and the **datamanager**. The rsmt is the point of access to all main tasks in the package, excepts the simhandle module, everything could be done via rsmt including dataset chose, defining the experiment design using some DoE strategy, training and evaluating the ML models, call reports, visualizations, manage data, etc.

The datamanager acts like a general data storage manager. After start the use of the package the user has the option to create a data space (composed by a folder and an *SQLite* database) to follow every action did and all generated files.