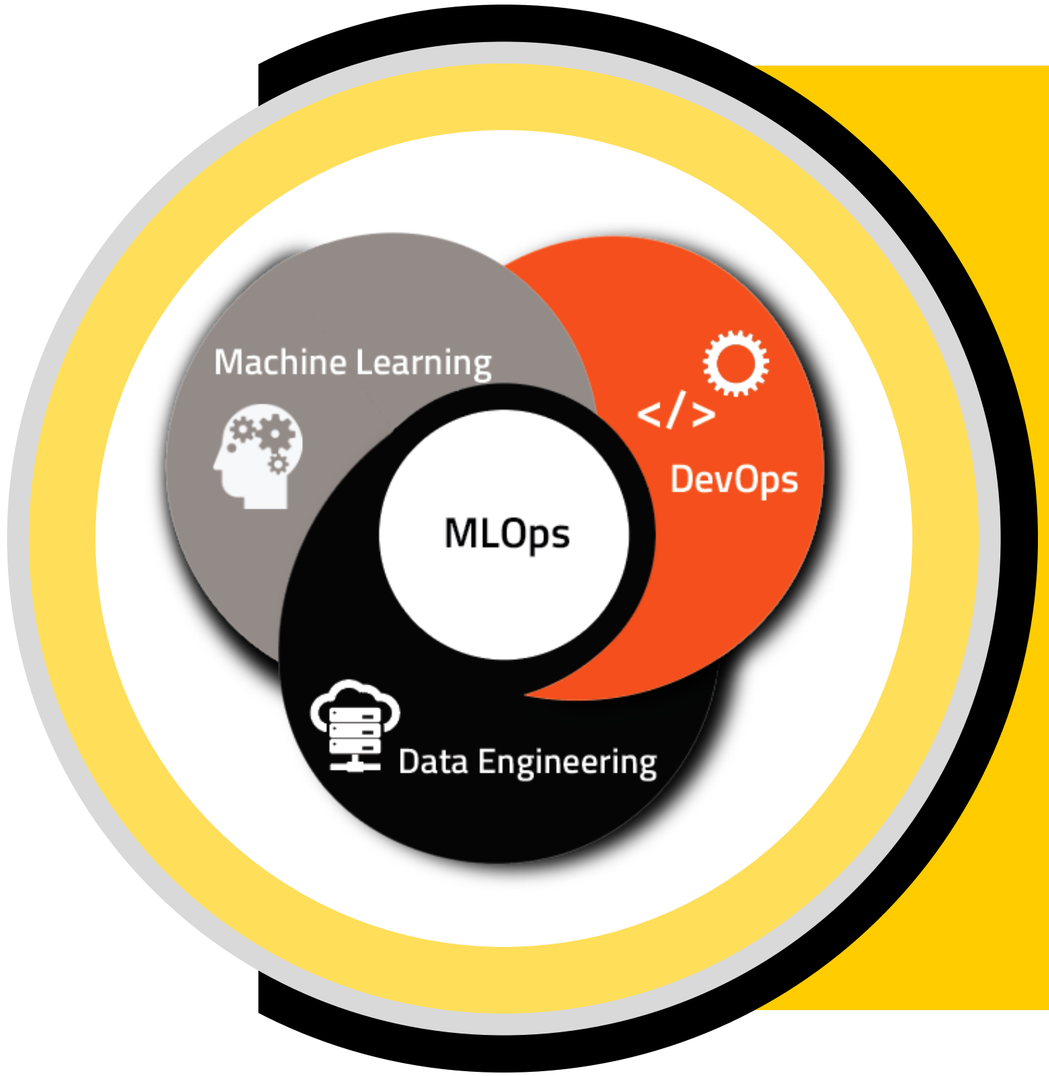
StatusNeo

statusneo.com



MLOps

Operationalizing Machine learning model to production

SESSION 1

AGENDA

- Deployment of ML Models
- Deployment of Machine Learning Pipelines
- Research and Production Environment
- Building Reproducible Machine Learning Pipelines
- Challenges to Reproducibility
- Streamlining Model Deployment with Open-Source
- Additional Resources

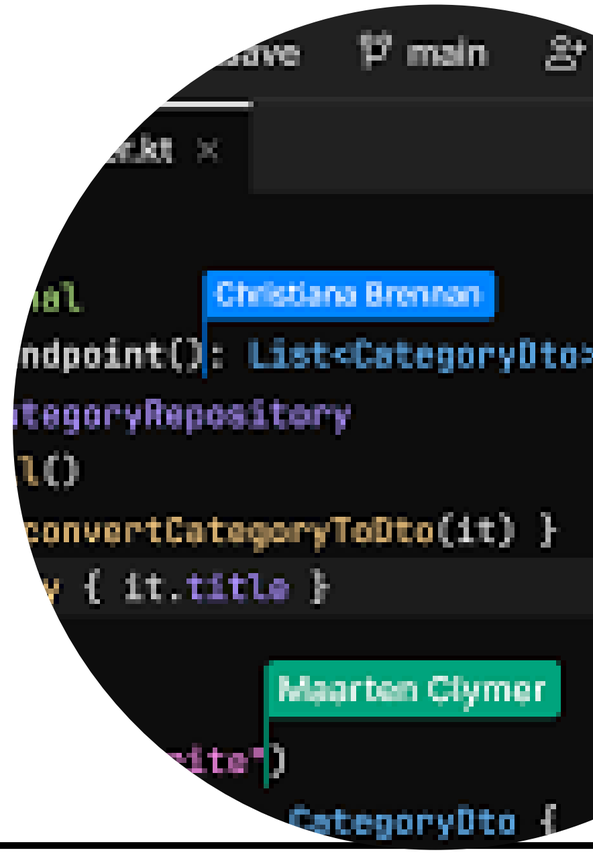
StatusNeo

What is Model Deployment

- The deployment of machine learning models is the process of making our models available in production environments where they can provide predictions to other software systems.

- Model deployment is one of the last stages in the machine learning lifecycle

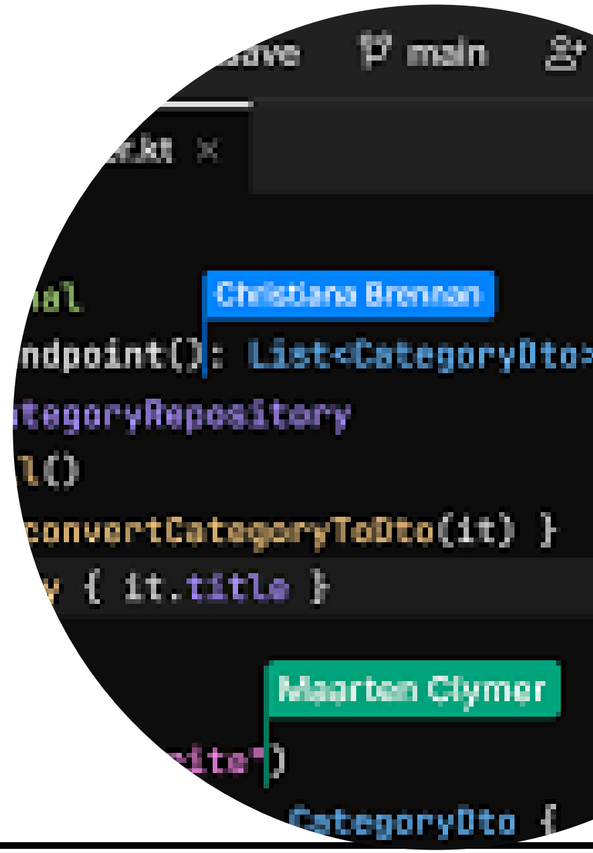
- Potentially the most challenging one.



StatusNeo

Why is Model Deployme Challenging

- Challenges of traditional software
  - Reliability
  - Reusability
  - Maintainability
  - Flexibility
- Additional Challenges specific to Machine Learning
  - Reproducibility

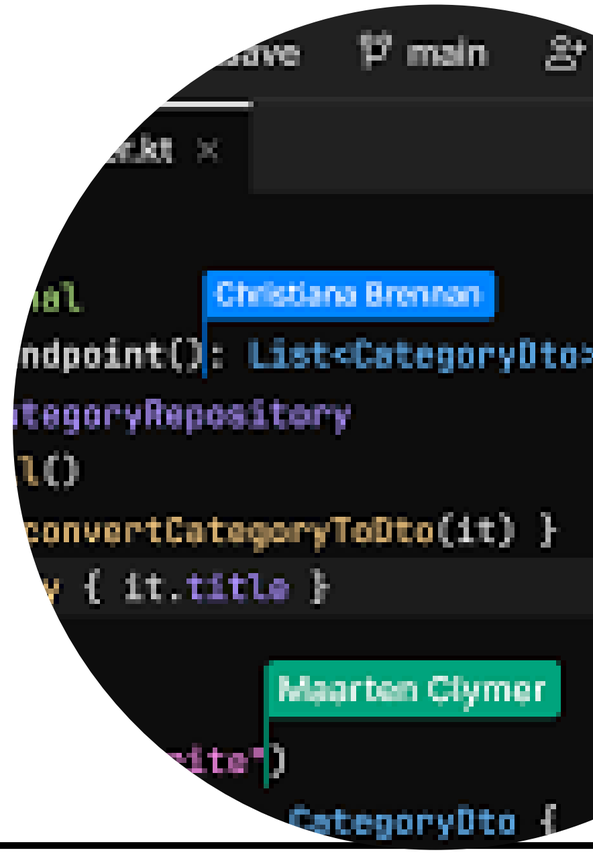


StatusNeo

Why is Model Deployment Challenging

- Needs the coordination of data scientists who develop the machine learning model I.T. teams , software developers and business professionals
  - Ensure model works reliably
  - Ensure model delivers the intended results.

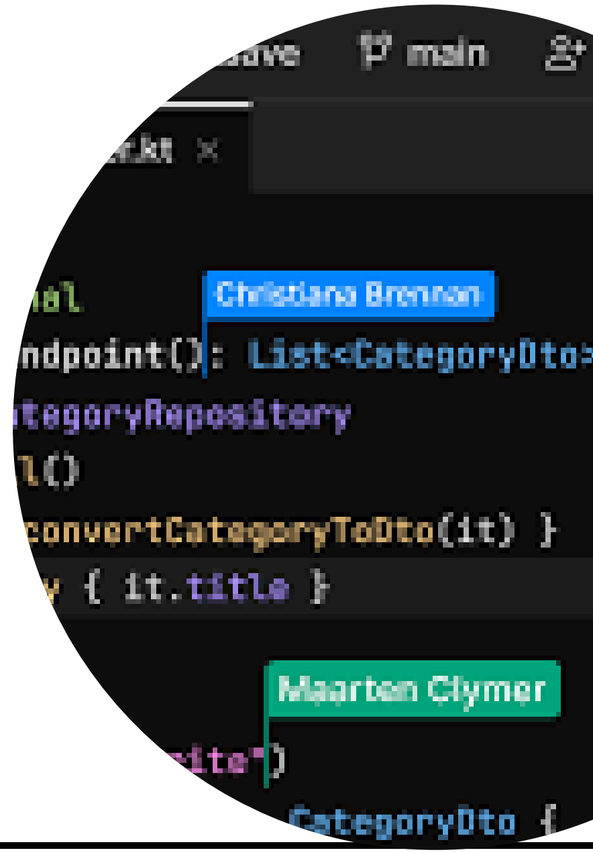
- Potential discrepancy between programming language in which the model is devleoped and the production sysem labguage
  - Re-coding the model extends the project timeline and risks lack of reporducibility



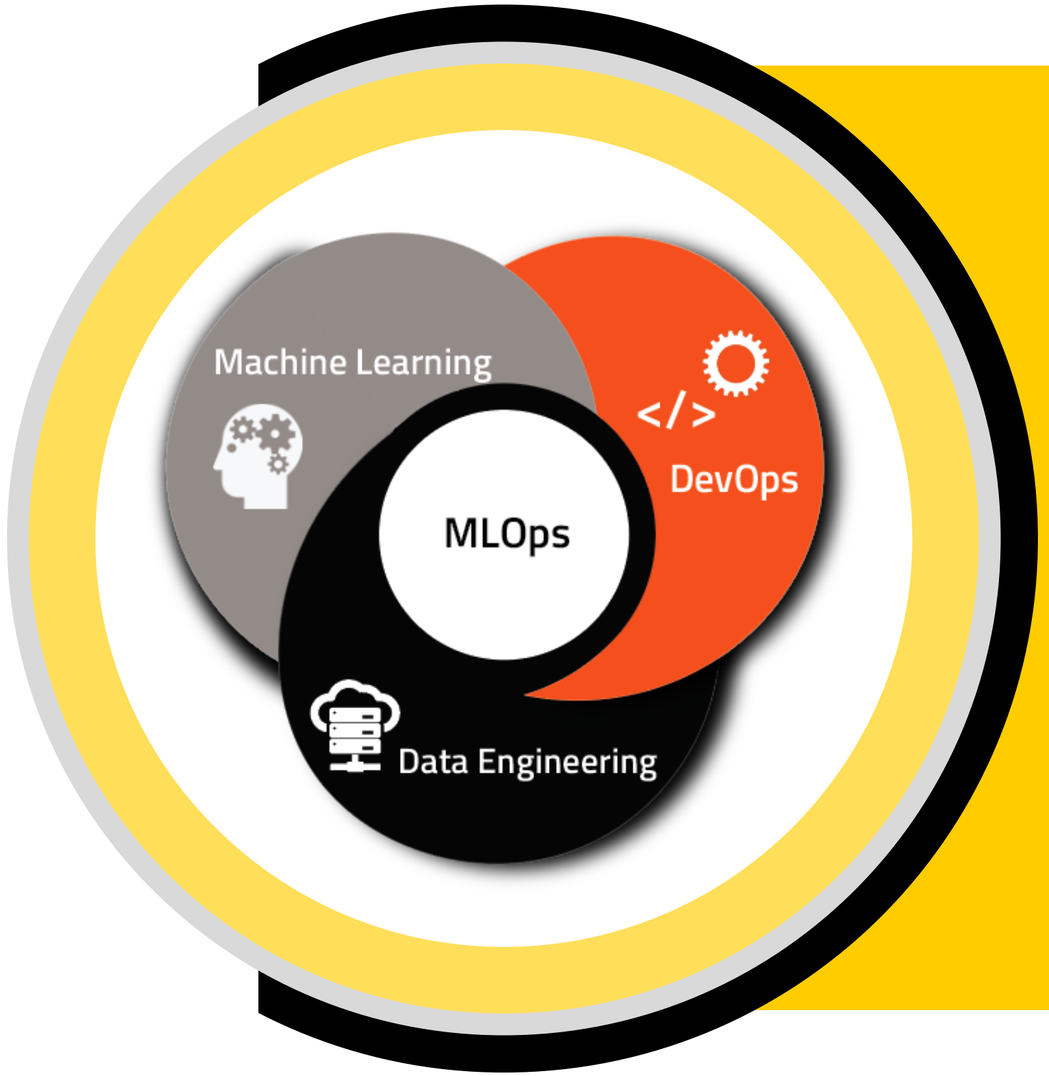
Why is Model Deployment Important

- To start using the machine learning model, it needs to be effectively deployed into production so that it can provide its predictions to other software systems.

- To maximize the value of the machine learning model we create, we need to be able to reliably extract the predictions from the model and share them with other systems.



StatusNeo

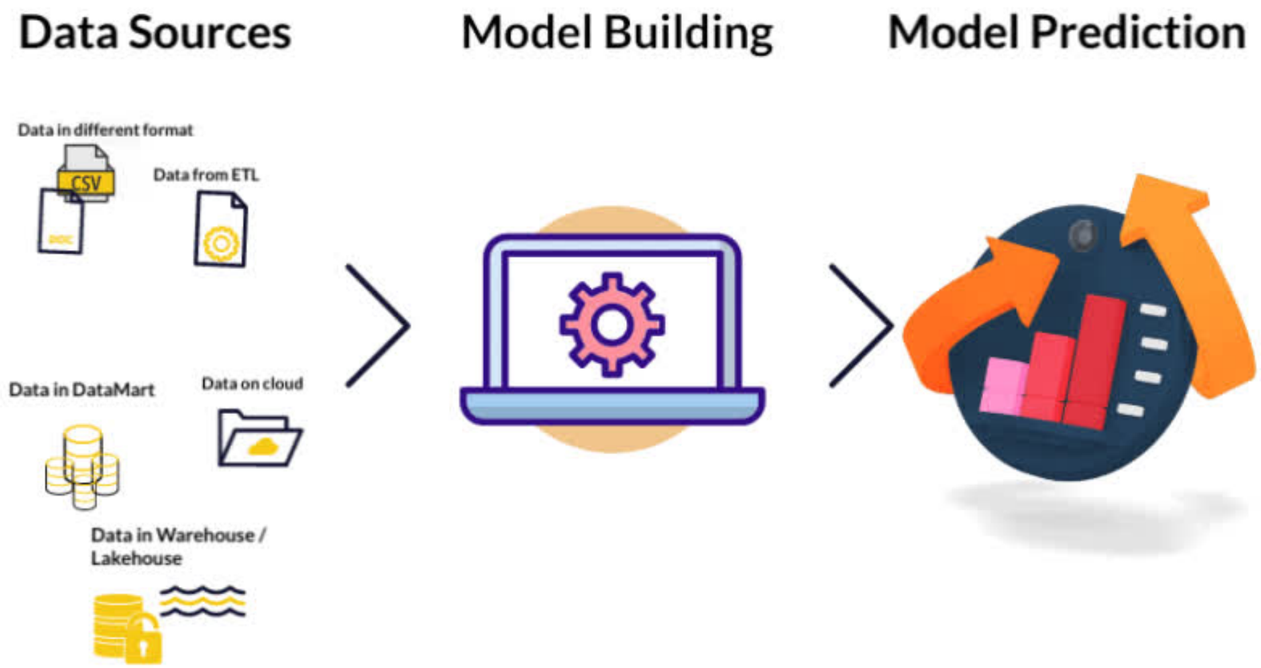


Deployment of ML Pipelines

SESSION 1 - PART 2

statusneo.com

StatusNeo



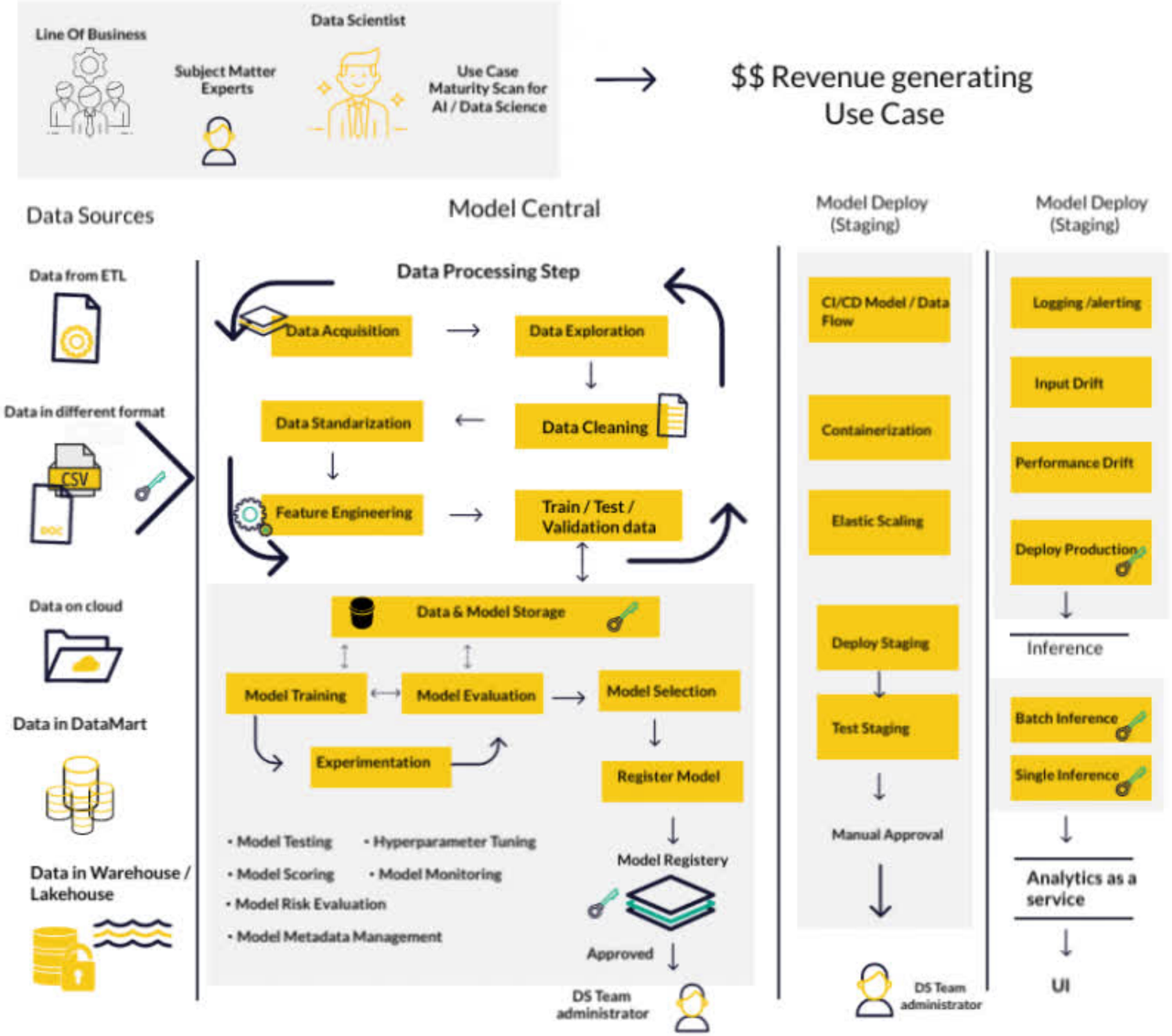
Machine Learning Models OR Pipeline

StatusNeo

Data Format & Quality

- Transform Variables
- Feature Extraction
- Create New Features
- Scaling , Rescaling of Data

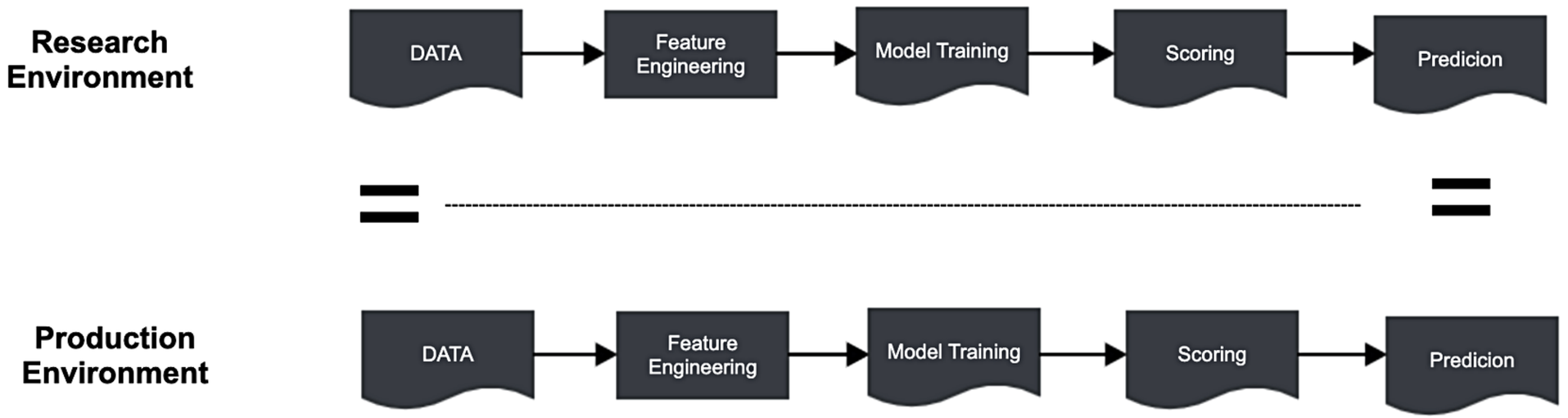
StatusNeo



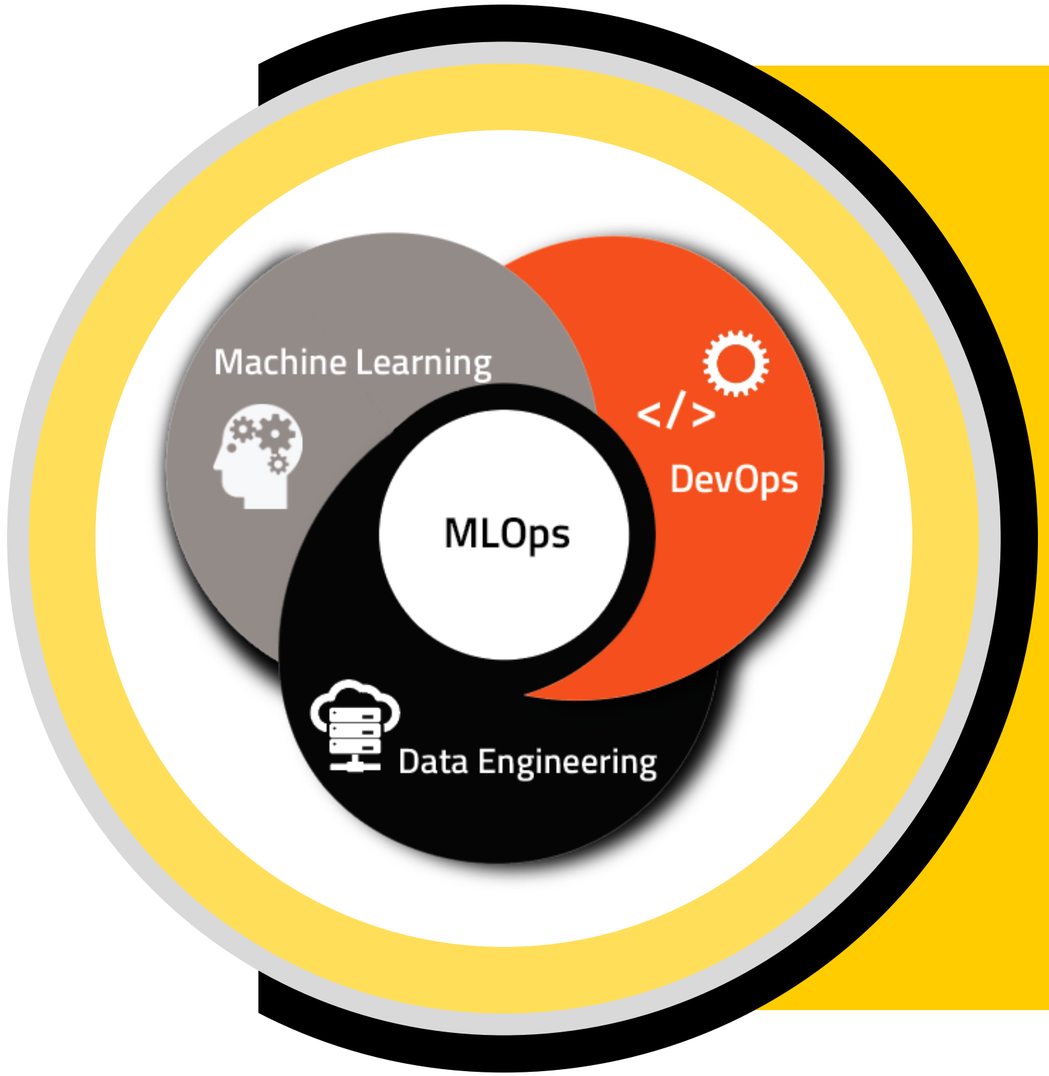
Machine Learning Models OR Pipeline

StatusNeo

Deployment of ML Pipeline



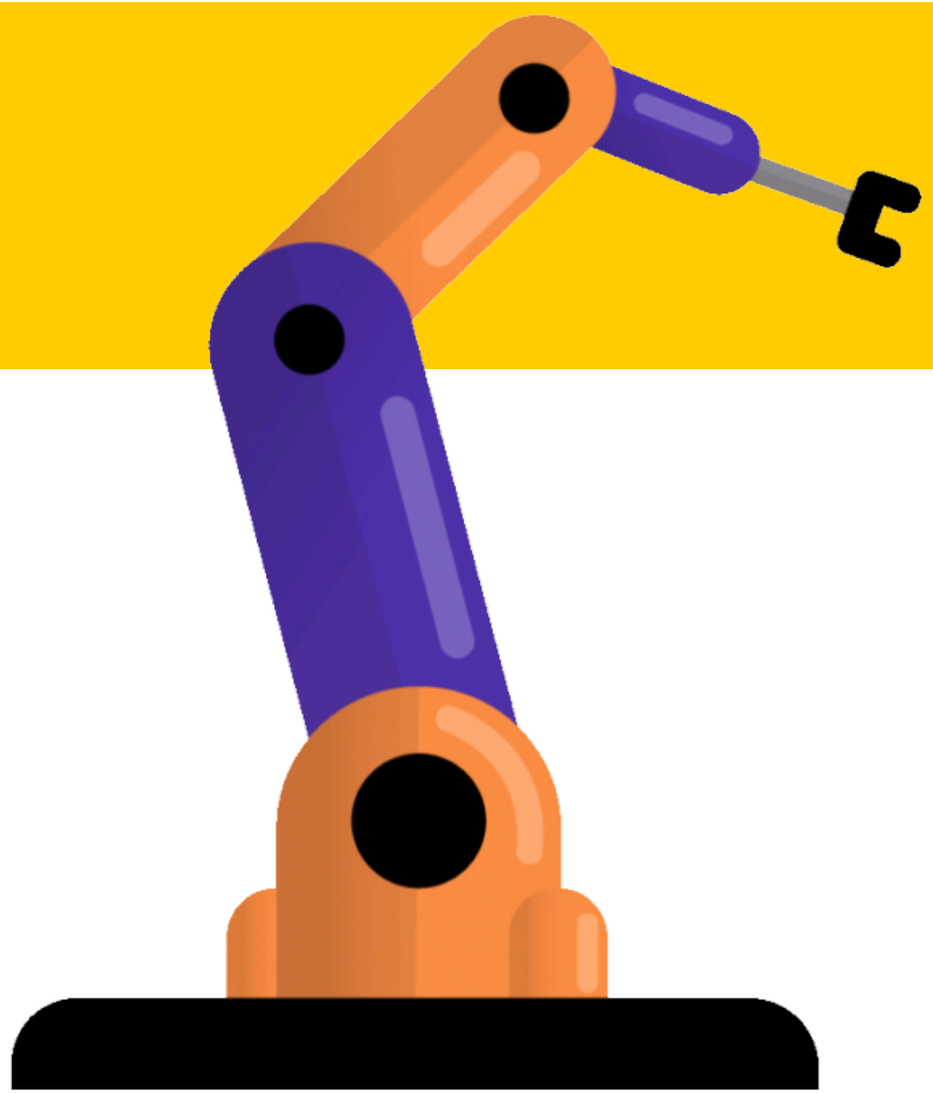
StatusNeo



Research & Production environment

SESSION 1: PART 3

statusneo.com



Environments

- The term environment refers to the setting or the state of a computer where software or other products are developed or put in operation.
- This setting usually includes
  - software
  - hardware programs &
  - operating systems.

Research Environment

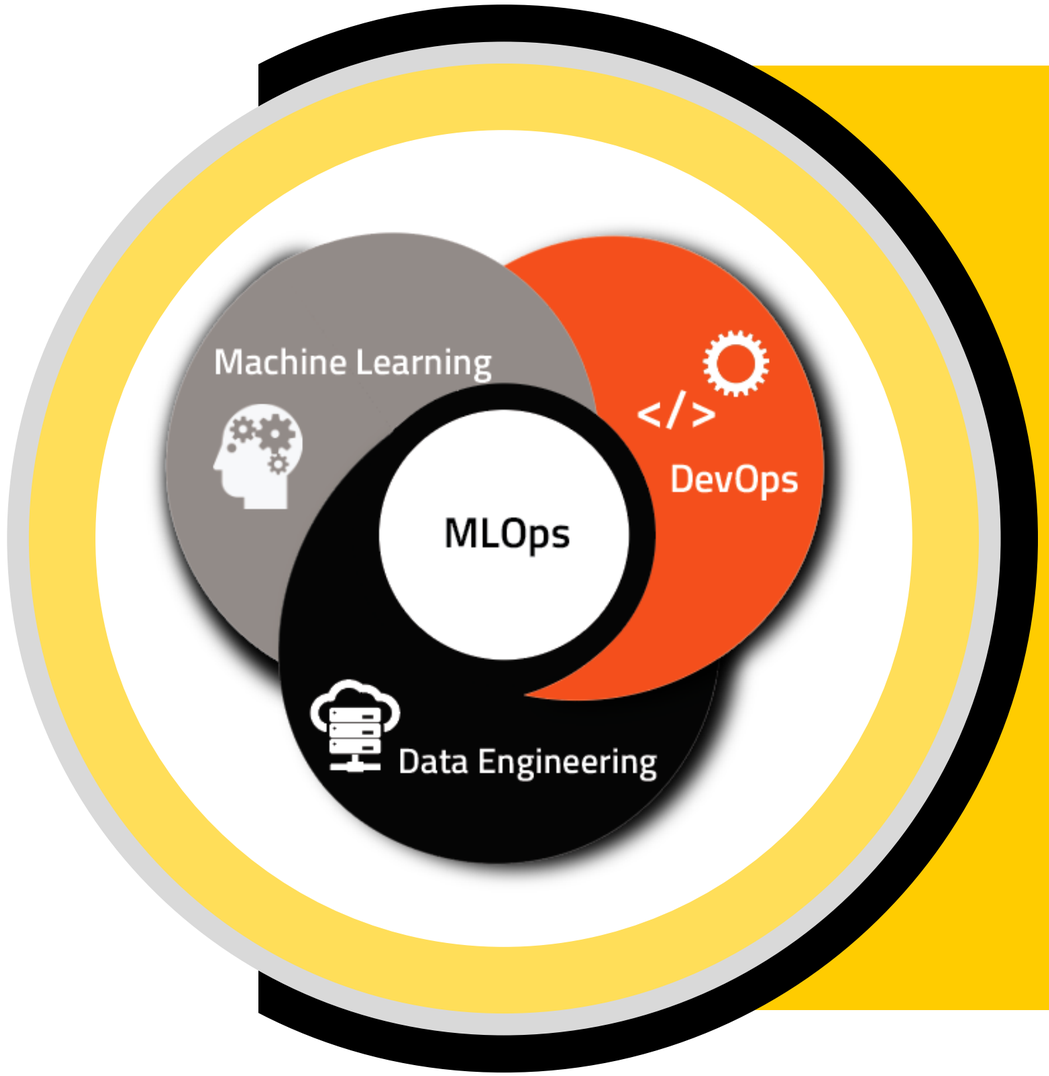
- The research environment is a setting that contains the tools, programs and software that are suitable for data analysis, pre processing and the development of machine learning models.

- Here, the data scientists usually develop the machine learning models and identify the potential value for the organization.

Production Environment

- The production environment is a real time setting, with running programs and hardware setups that allow the organisations daily operations.
- Production environment is where the machine learning models are actually available for business use.
- It help the organisation provide client "live" service of the ML Model

StatusNeo

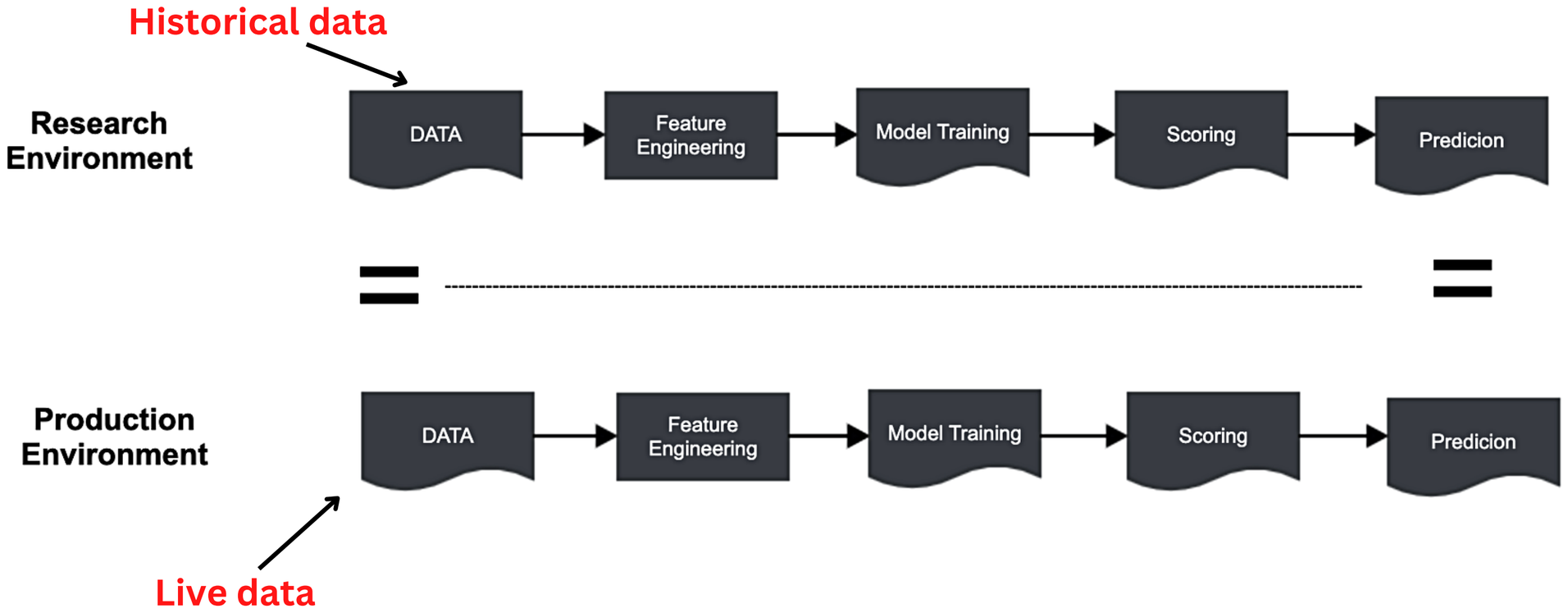


Building Reproducible pipeline

SESSION 1 : PART 4

statusneo.com

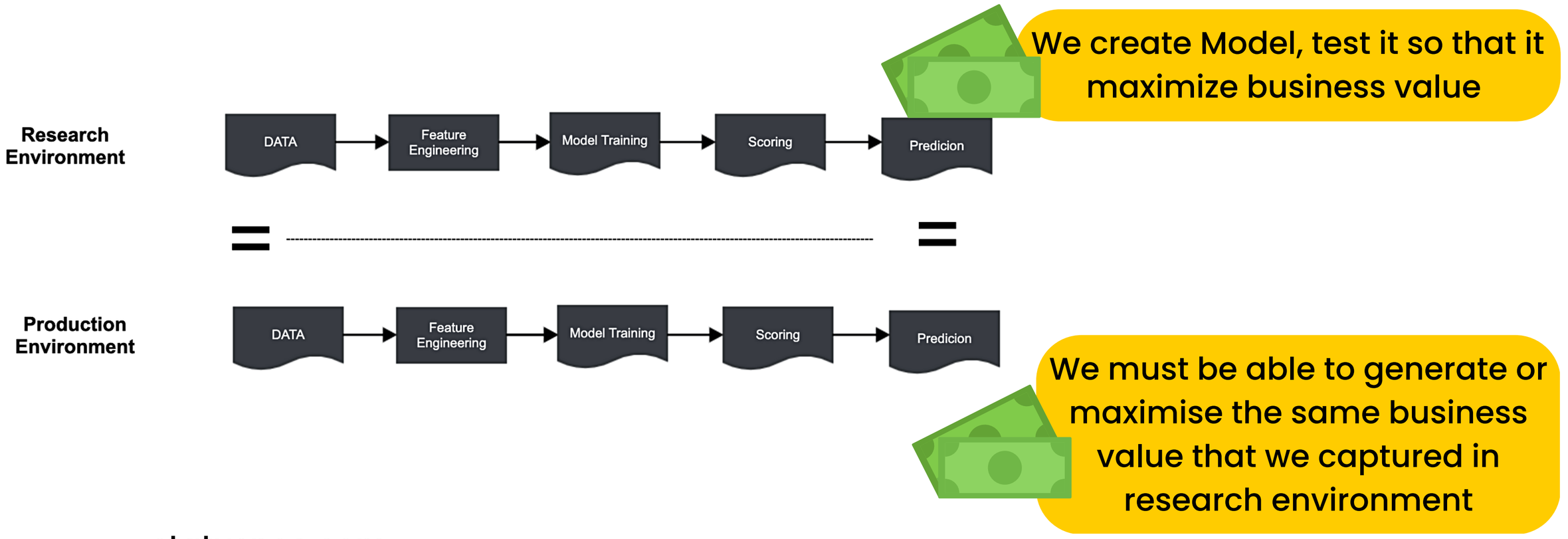
Deployment of ML Model



We often talk about deployment of machine learning models when, in fact, we mean deployment of a machine learning pipeline

- It is the series of steps that need to occur from the moment we received the data up to the moment we obtain a prediction.
    - A typical machine learning pipeline includes a big proportion of feature transformation steps. Perhaps this is even the biggest part of the pipeline.
    - Then it includes steps to train the machine learning mother and steps to output a prediction would create an entire pipeline in a research environment.
    - I would need to deploy as well the entire pipeline to the production environment. Because in the light environment, we're also going to receive raw data as input and we need to transform it to create the necessary features so that the model can interpret them and return the prediction.
    - When we deploy our pipeline to production, we need to do it in a way so that the pipelines are reproducible

Deployment of ML Model



StatusNeo

Why Reproducibility Matters

If we cannot reproduce our machine learning pipelines within the research and production environments we may incur in financial costs

Financial cost	We may measure the increased revenue that our model will create or the increased customer satisfaction or any other measure that we're interested in, and that will vary with the organisation. If we want to translate this value, this revenue increase from a research environment to the real life that is to the environment, we need to make sure that the like model reproduces.
Time Cost	We may not only incur financila loss but we will incur a substantial time loss because we need to spend a lot of time trying to figure out why the models are not identical or atleast similar
Lost reputation	All this will end up in loss of reputation of the organization

What is reproducibility?

Reproducibility is the real problem that exists both in businesses and in academia.

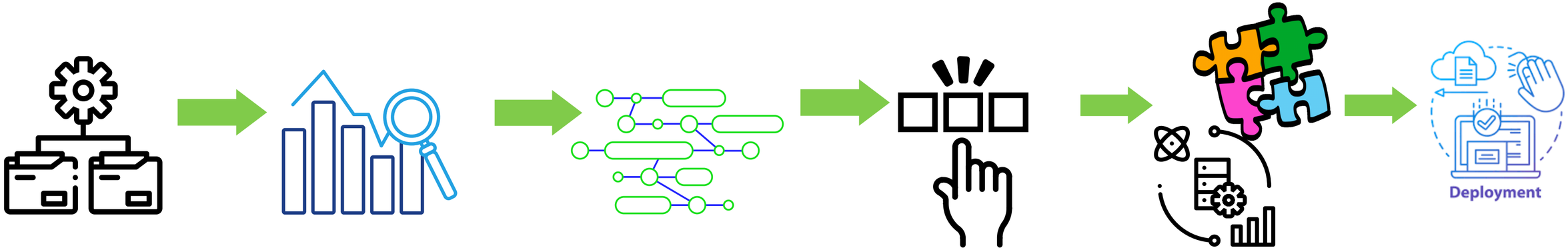
It is important to try and get this right right from the beginning.

Reporducibility is the ability to duplicate a machine learning model exactly, such that given the same raw data as input both models return the same output.

StatusNeo

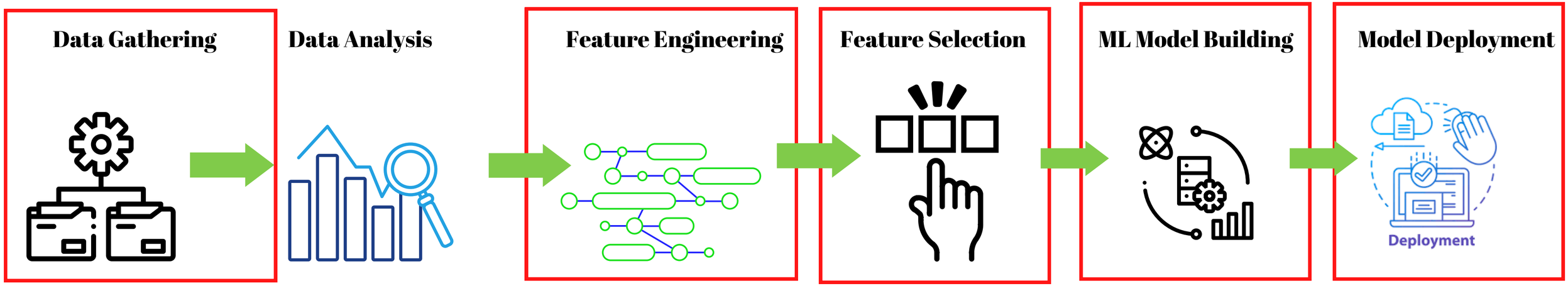
At which step of the pipeline do we need to ensure reproducibility?

Data Gathering → Data Analysis → Feature Engineering → Feature Selection → ML Model Building → Model Deployment



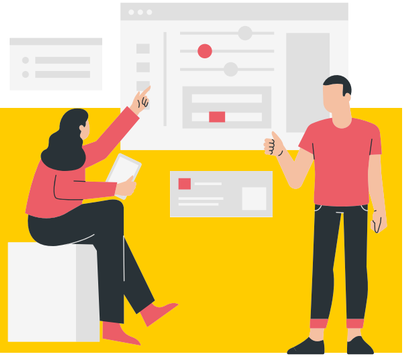
StatusNeo

Well, pretty much at every step of the pipeline!



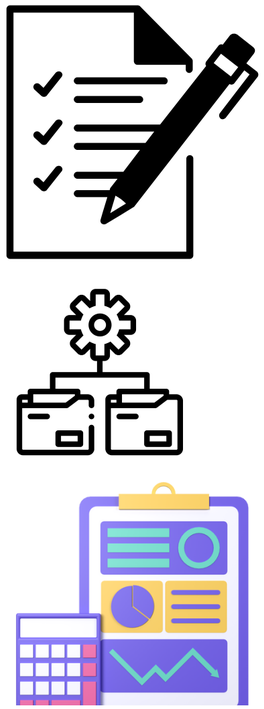
So we need to make sure that all the steps in red are reproducible. This means that all these steps should produce identical results, given the same data, both in a research environment and in our production and deployed models.

StatusNeo



Reproducibility During Data Gathering Stage

This stage represents the most difficult challenge to ensure reproducibility.



Challenges

Training Data can't be reproduced

Why?

→ Database are constantly updated and overwritten

→ Order of data while loading is random (SQL)

Solution

Save a snapshot of data

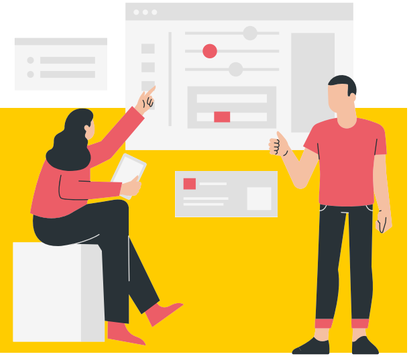
✓ Simple
- Potential Conflict with GDPR
- Not suitable for Big data

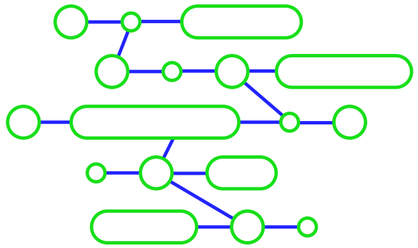
Design Data Sources with accurate timestamp

✓ Ideal Solution
- Big effort to re-design the data sources

StatusNeo

Reproducibility During Feature Creation (feature engineering)





Lack of Reproducibility may arise

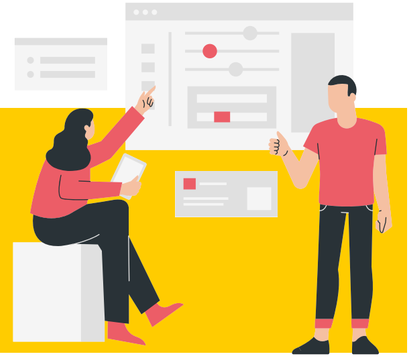
- Replacing Missing data with random extracted values
- Removing labels based on percentages of observation
- Calculating Statistical values like the mean to use for missing value replacement
- More complex equations to extract features, e.g., aggregating over time

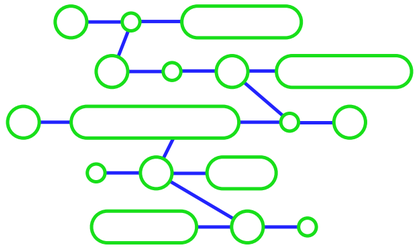
Solution

- ✓ Code on how a feature is generated should be tracked under a version control and published with auto-incremented or timestamp hashed version
- ✓ Many of the parameters extracted from feature engineering depends on the data used for training -> ensure data is reproducible
- ✓ If replacing by extracting random samples, always set a seed

StatusNeo

Reproducibility During Feature Creation (feature engineering)





Lack of Reproducibility may arise

- Machine learning models reply on randomness for training
  - Data & feature extraction for trees
  - Weight Initialisation for neural nets, etc

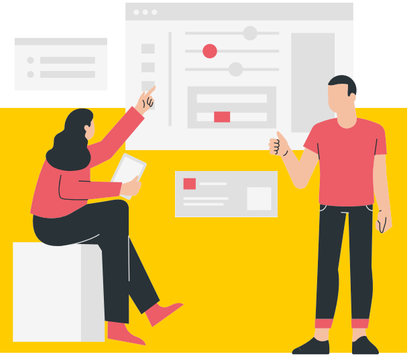
- Machine Learning Models implementation work with arrays agnostic to feature name
  - Need to be careful to feed data in the correct order

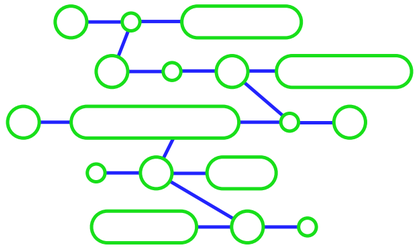
Solution

- ✓ Record the error of feature
- ✓ Record applied feature transformation
- ✓ Record hyperparameters
- ✓ For models that require randomness always set the seed
- ✓ If the final model is the stack of models, record the structure of the ensemble

StatusNeo

Reproducibility During Feature Creation (feature engineering)





Challenges

- A feature is not available in the live environment
- There are different softwares and sometimes the programming language that we obtain through the live production environment do not match.
- The population used to train the machine learning models to ensure reproducibility during the integration of the model.
- Different Software version

Solution

- ✓ In the production environment, software versions need to match exactly those used in the research environment and the applications or the model API should list all third party library dependencies under versions.
- ✓ Use a container and rack software specifications
- ✓ Research, develop and deploy using the same language ; eg ; python
- ✓ Prior to building the model think about how this model will be integrated to the different system

Reproducibility During Feature Creation (feature engineering)

Building Reproducible Pipelines:

- Building a Reproducible Machine Learning Pipeline - https://arxiv.org/ftp/arxiv/papers/1810/1810.04570.pdf
- Reproducible Machine Learning - presentation, Kaggle - https://www.rctatman.com/files/Tatman_2018_ReproducibleML.pdf
- The Machine Learning Reproducibility Crisis - article, by Google developer
- Streamlining Deployment with Open Source
    - Six motivations for using open source - https://opensource.com/life/15/12/why-open-source

Download DataSet

- Download House Prices data set from Kaggle
  - Create a Kaggle account
  - If you have a Kaggle account already, go straight to Download data set.
- Visit Kaggle's website
- Click on the "Create an account" button and follow the instructions to set up your account
- Download data set
- Visit the House Sale Price competition website, scroll down and click on train.csv and test.csv files (see image below) and then on the download button on the right.
- Store the datasets somewhere safe, and we will tell you were to move them as we go along in the course.