Methodology for Data Science: Mastering DataOps and MLOps

Feda ALMUHISEN

Email: feda.almuhisen@univ-amu.fr

Méthodologie pour la science des données

Course details:

27 Hours (9 CM, 9 Hours TD, 9TP)

- **Tools:**
  - Python (Version 3.9 or higher)
  - GitHub / Git for Version Control
  - MLflow for Model and Data Versioning
  - Automated Machine Learning Libraries
  - Flask for Model Deployment
- **Prerequisites**:
  - Proficiency in Python Programming
  - Basic Understanding of Machine Learning Algorithms
- **Assessment Methodology**:
  - Practical Assignments to Reinforce Learning
  - Project for Real-world Application

Data Science






Méthodologie pour la science des données

Data science Pipeline (OSEMI):

O Obtaining data

S Scrubbing / Cleaning data

E Exploring patterns and trends

M Modeling data

I Interpret events

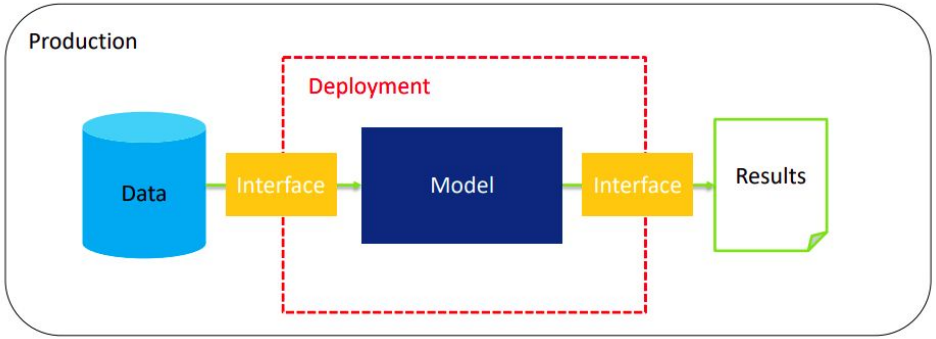
Méthodologie pour la science des données

Data science:

Manipulate and interpret raw collected data from different resources to extract useful intelligence(knowledge) results from noisy, structured or unstructured data. Using scientific methods, processes, algorithms ..

Data Science Challenges:

- Understanding the business problem.
- Effective communication with non-technical stakeholders.
- Data preparation.
- Multiple data sources.
- Data security.
- Collaboration with data engineers.
- Misconceptions about the role.
- Undefined KPIs and metrics.



Méthodologie pour la science des données

Questions: How can we overcome these challenges?

Data is more dense, distributed, diverse today!

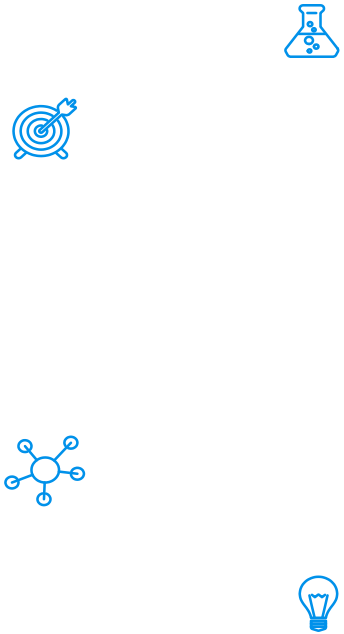
- **Full value from data → Shift the data management strategies to be more collaborative, unified and automated.**
- **Getting the right data in the hands of the right people at the right time !**
- **DataOps Methodology!**

Méthodologie pour la science des données

Overcome Data Science challenges:

- Agile mindset (collaborate with customers), respond to change..etc.
- Having a dedicated team.
- Focus on the business requirements and domain.
- Setting up data governance.
- Setting up data quality KPI’s and metrics.
- Apply DataOps, DevOps concept reusable code, unit testing, documentation, version control.

DataOps
(Data Operations)



Méthodologie pour la science des données

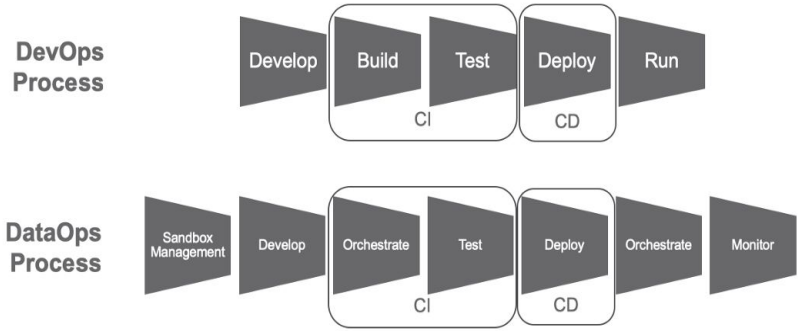
DevOps DataOps … What's the difference?!

Software development process:

DataOps is NOT Just DevOps for Data!

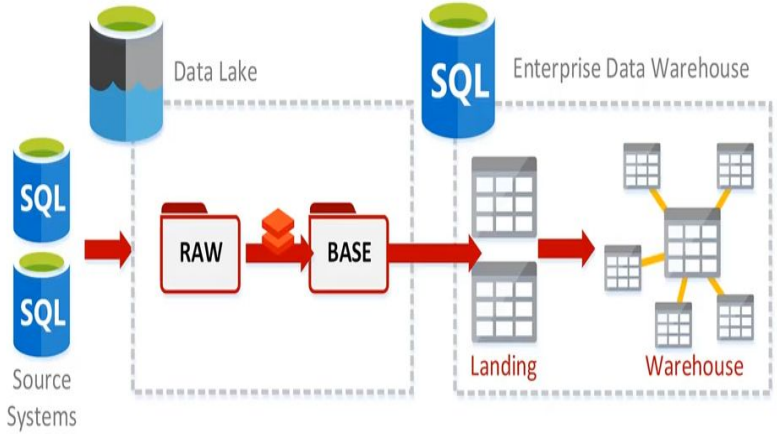
DevOps: *You build it you run it !*

- **Extends the Agile mindset** of (Requirements, Design,Implementation,Testing and verification,Release maintenance by **including the concept of CI, CD, continuous deployment, continuous monitoring and feedback**.



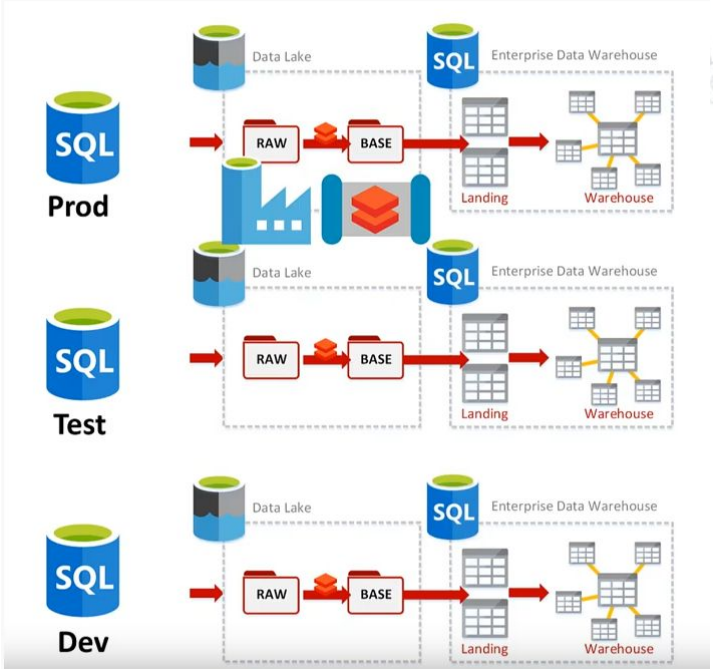
Méthodologie pour la science des données

Data platform:



Framework Development Workflow(DevOps)

Develop area (change and growing) -->**test** (validation) →

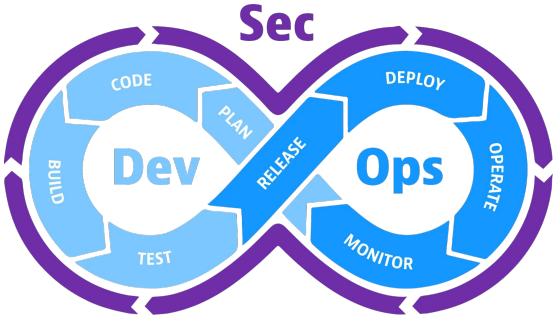


Méthodologie pour la science des données

DevSecOps: (Development - Security - Operations)

Building security into app development from end to end. The combination of DevOps with security teams.

You build it you secure it !

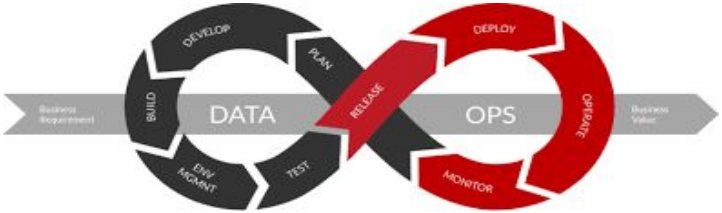


- Automate, monitor and apply security at all phases of the software lifecycle.
- Deploy application within security configurations.
- Using test automation,

DataOps:

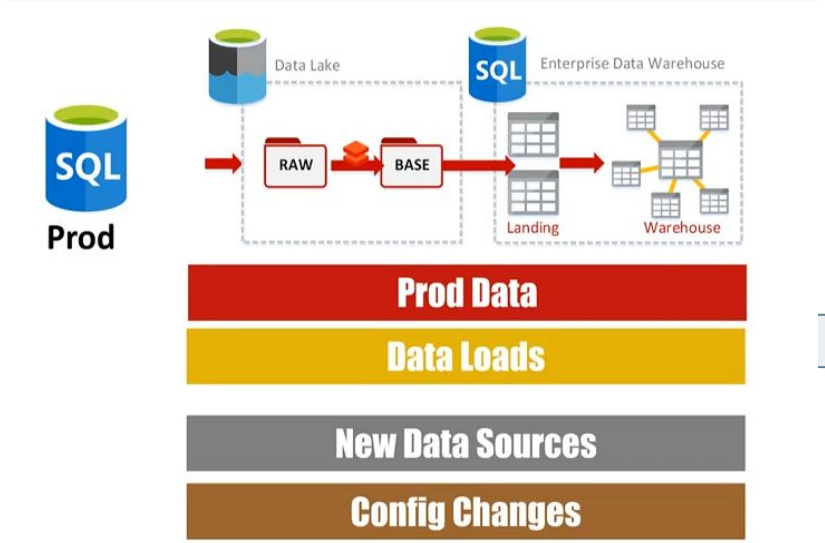
Leveraging DevSecOps Principles for Secure Data Analytics ,focuses on data quality improvement.

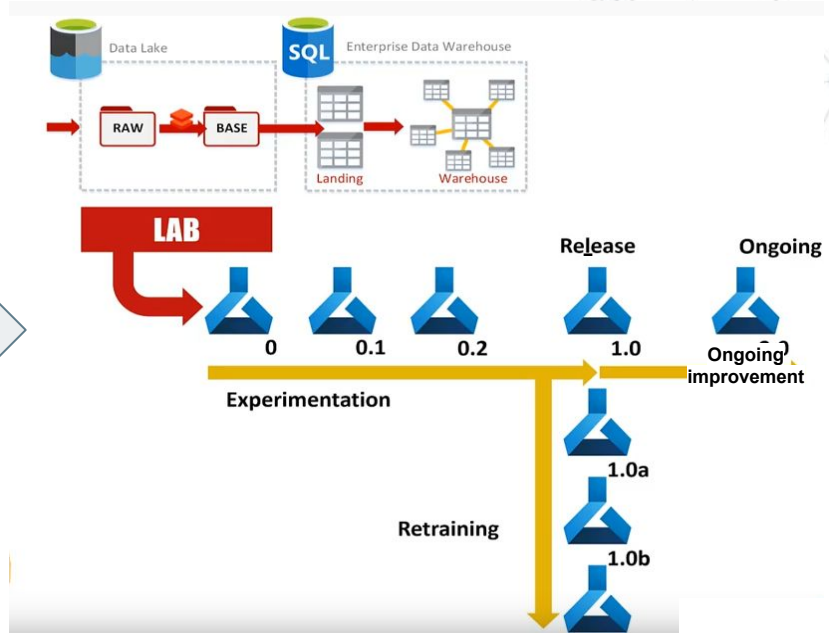
(From Data science to Data operations).



Méthodologie pour la science des données

Framework(DataOPs)





1. Implementing automatic testing
2. Use version controls
3. Branch and merge
4. Provide isolated environments
5. Reuse code
6. Use parameters in the pipeline

Iterative improvement over Traditional development
Experiment tracking

Méthodologie pour la science des données

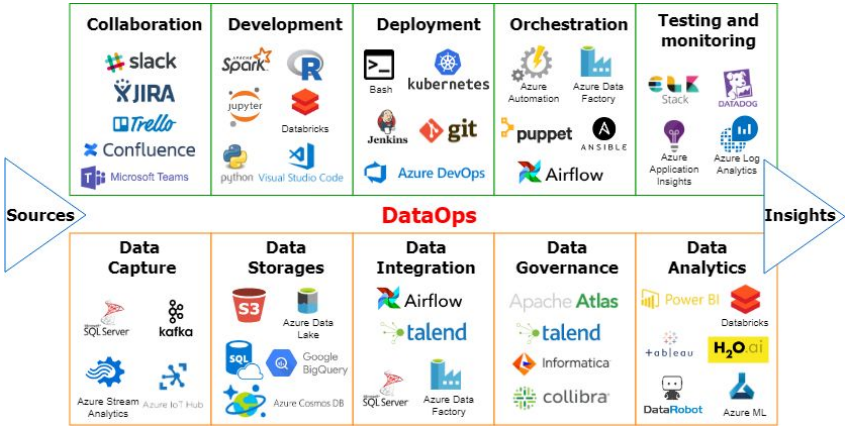
What Is DataOps?

Collaborative data management practice focused on improving the communication, integration, and automation of data flow for analytics.

DataOps objectives:

- **End to End Efficiency:** Complete control over the data lifecycle (fast delivery high quality data).
- Analytic collaborations:
  Brings together **DevOps** teams with ***data engineers*** and **data scientists** to provide the tools, processes and organizational structures to support the data-focused enterprise.

-



DataOps provides:

- Data integration
- Data validation
- Metadata management
- Observability

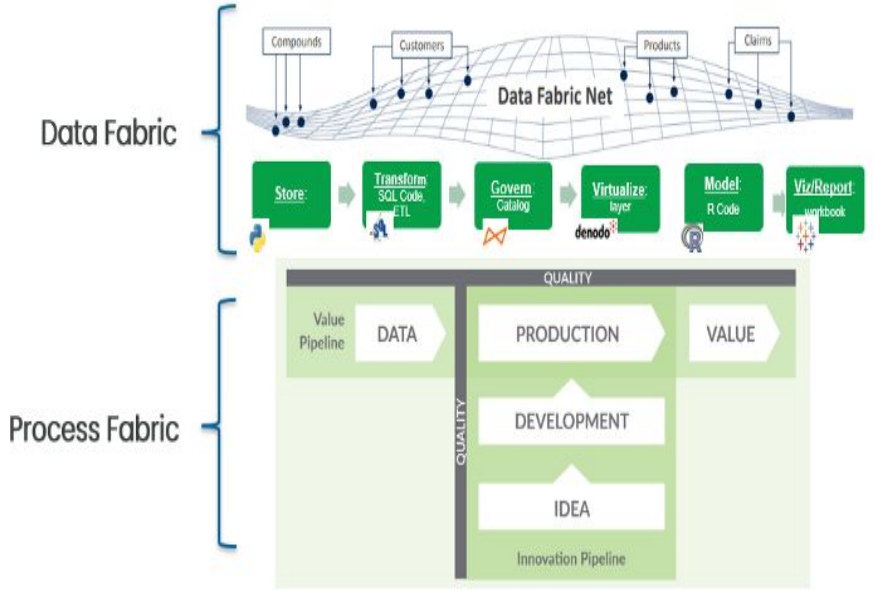
Méthodologie pour la science des données

DataOps Methodology:

The data fabric takes raw data sources as input and through a series of orchestrated steps produces analytic insights that create "*value*" for the organization.

DataOps methodology: allows us to ensure that the data used in problem-solving is relevant, reliable, and traceable to address the question at hand.

- It involves building and deploying the data pipelines and analytics, model management and data governance
- Reduce cost
- Driving Innovation

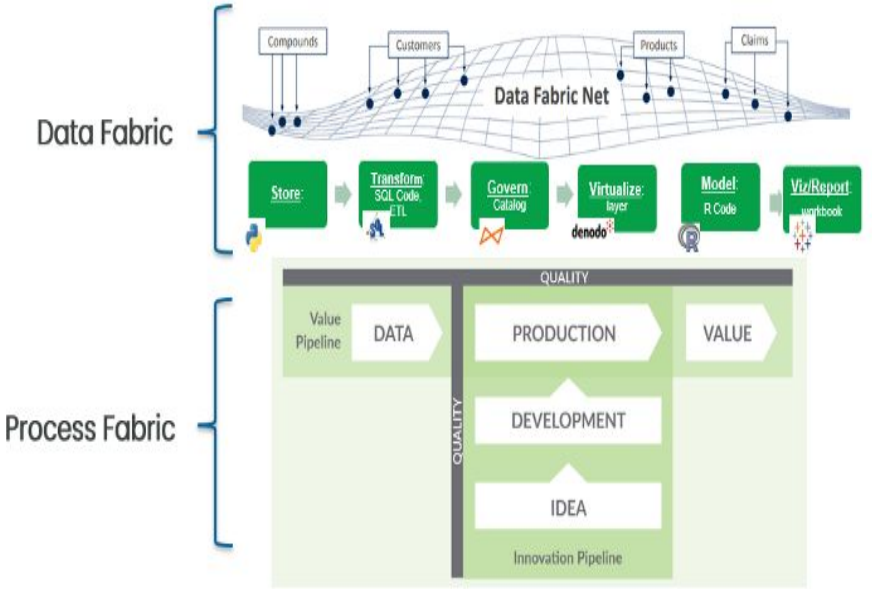


Inputs (data sources)-->processes (transformations)--> outputs (analytics)

Méthodologie pour la science des données

Key Steps to Implementing a Successful DataOps Practice:

- Enable technology: such as (IT automation tools , data management tools, machine learning..etc)
- Architecture based on major technologies capable of continuous change.
- Automated tools
- Employ the DataOps methodology to build and deploy analytics data pipelines
- Culture get people from different departments work together



Inputs (data sources)-->processes (transformations)--> outputs (analytics)

Méthodologie pour la science des données

DataOps ecosystem (Diverse team) :

Data engineers
Can operate in any data environment and create the master table with data scientists

Data Scientist
Leverage advanced analytical techniques to extract insights from master table

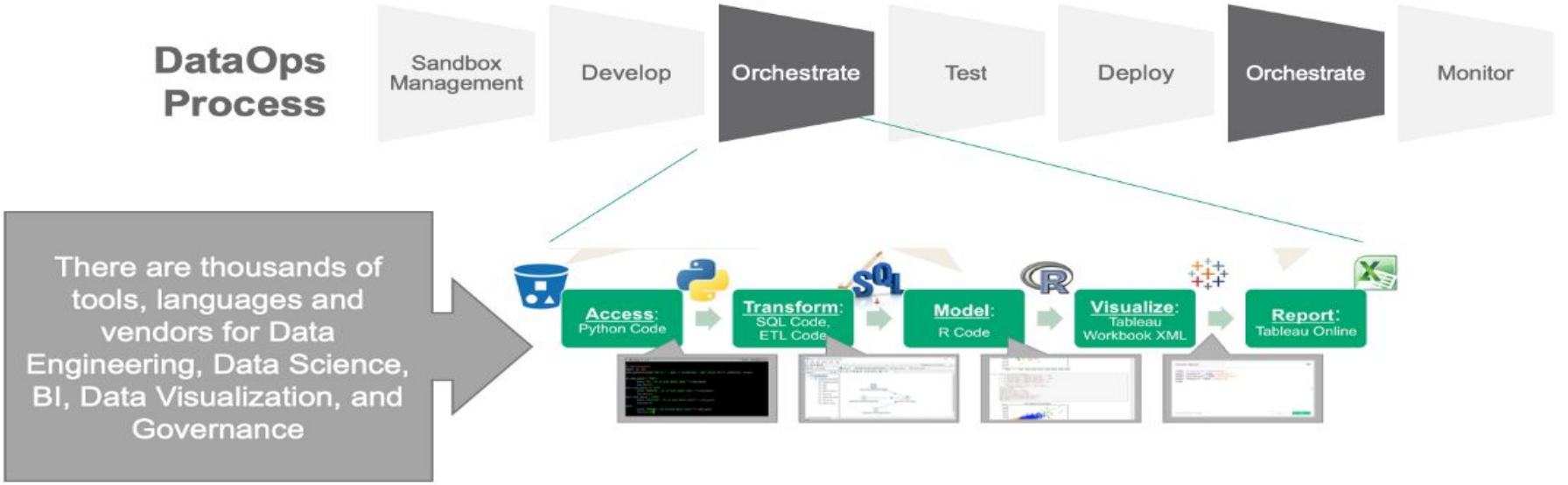
Data Science

Machine learning engineers
Deliver data science code into production

Méthodologie pour la science des données

The Duality of Orchestration in DataOps:

Creating a Pipeline for High-Quality Data and Actionable Information



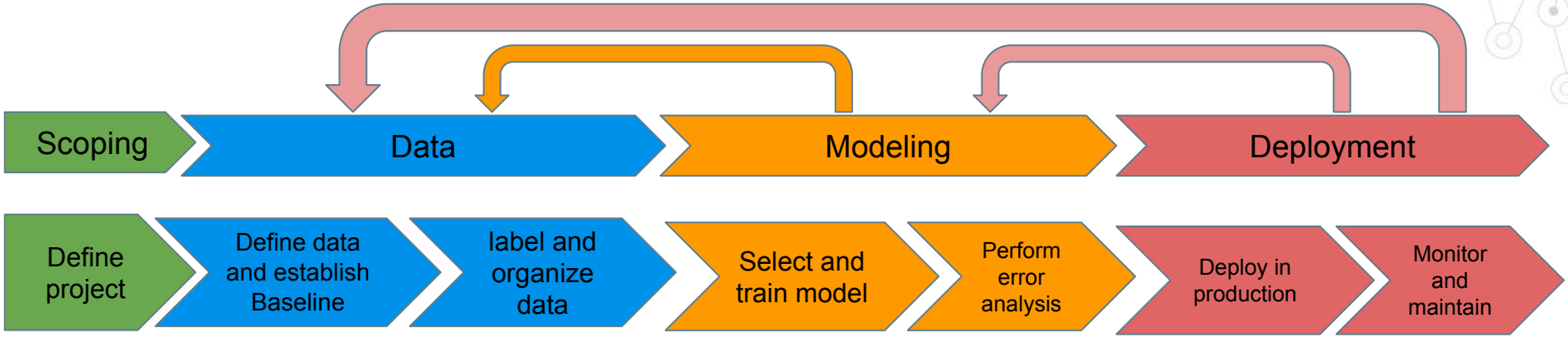
Data Science in practice






Méthodologie pour la science des données

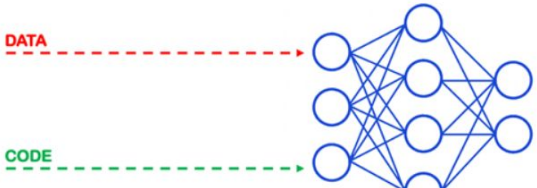
The ML project lifecycle



MLOps :is an emerging discipline and comprises a set of tools and principles to support progress throught the ML project lifecycle

- Code (Algorithm/model)
- Hyperparameters
- Data

Machine learning system = data + code



Méthodologie pour la science des données

Model Centric view

Collect what data you can, and develop a model good enough to deal with the noise in the data.

Hold the data fixed and iteratively improve the code/ model.

Data Centric view

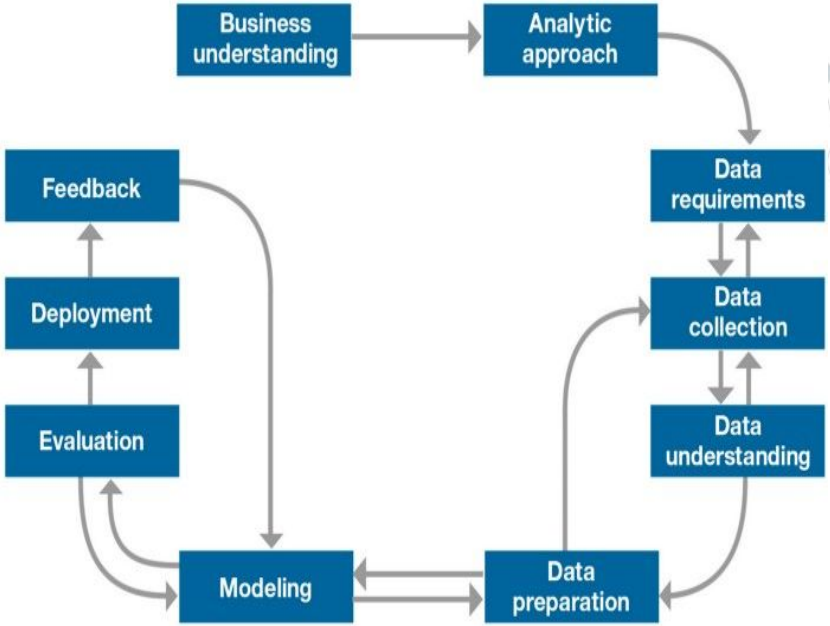
Use tools to improve the data quality; this will allow multiple models to do well.

Hold the code fixed and iteratively improve the data.

Méthodologie pour la science des données

Solving Data science problems methodology:

1. From Problem to Approach
2. From Requirements to Collection
3. From Understanding to Preparation
4. From Modeling to Evaluation
5. From Deployment to Feedback



Méthodologie pour la science des données

1. **From Problem to Approach :**

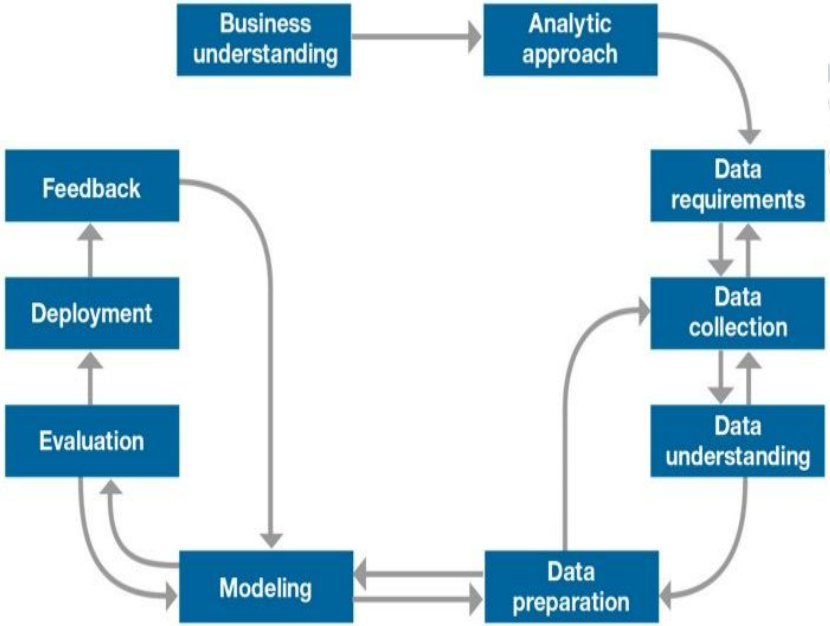
   a. **Business understanding:**

      i. What is the customer need?
      ii. What is the expected outcome?
      iii. What is the level of service?

   b. **Define analytic approach:**

Examples:

   i. Determine probability -->predictive model
   ii. Show relationship-->descriptive model
   iii. counts → statistical analysis

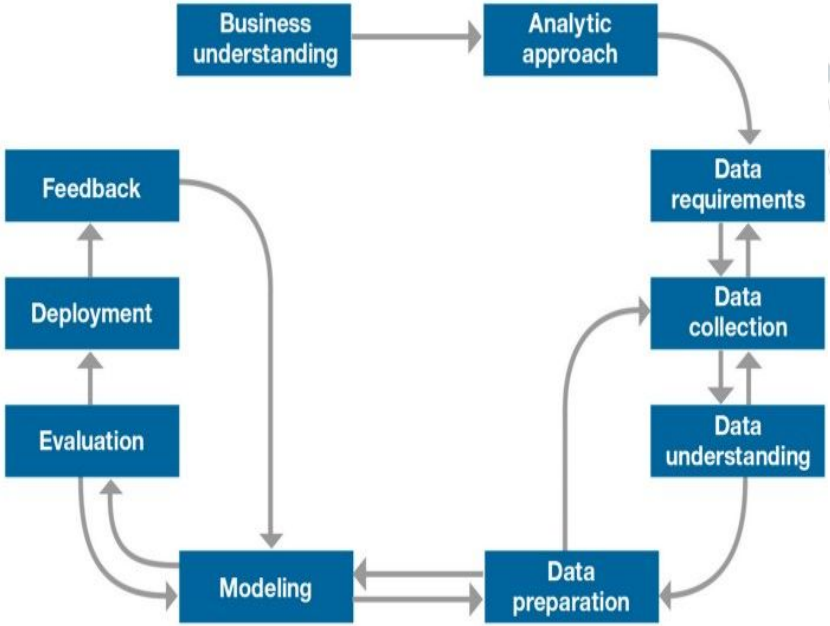


Méthodologie pour la science des données

2. **From Requirements to Collection :**

   a. **Data Requirements**:

   identify data content, formats, and sources for initial data collection, and use this data inside the algorithm of the approach we chose.



   a. **Data Collection** identify the available data resources relevant to the problem domain.

      i. Web scraping

      ii. Public Datasets

      iii. API Rest

Build a dataset of “reasonable” size representative of reality in order to be able to analyze it

Méthodologie pour la science des données

3. From Understanding to Preparation(EDA) :

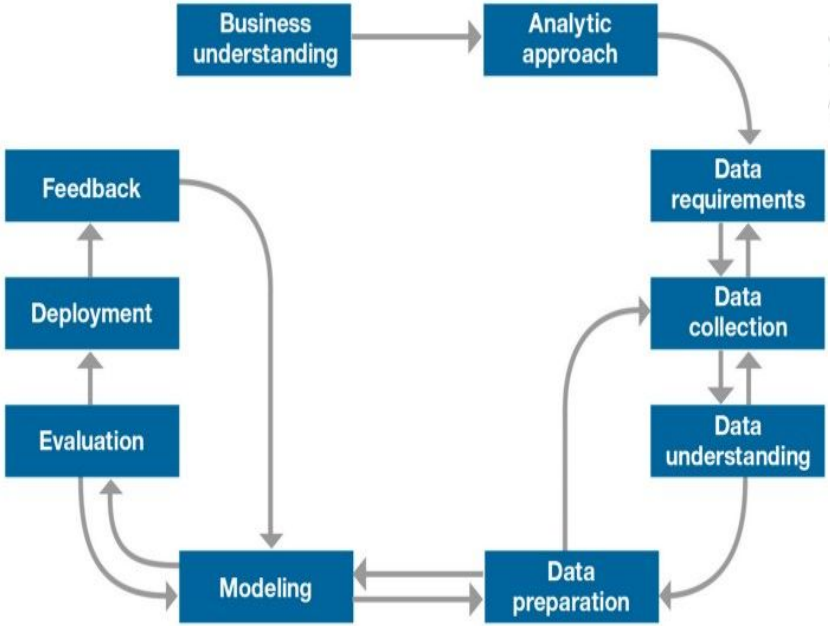
a. Data understanding :

Check the type of each data and learn more about the attributes and their names. and test hypotheses, detect outliers.

Statistics +dataViz

a. Data preparation :

Be sure that data are in the correct format for the machine learning algorithm we chose in the analytic approach stage.

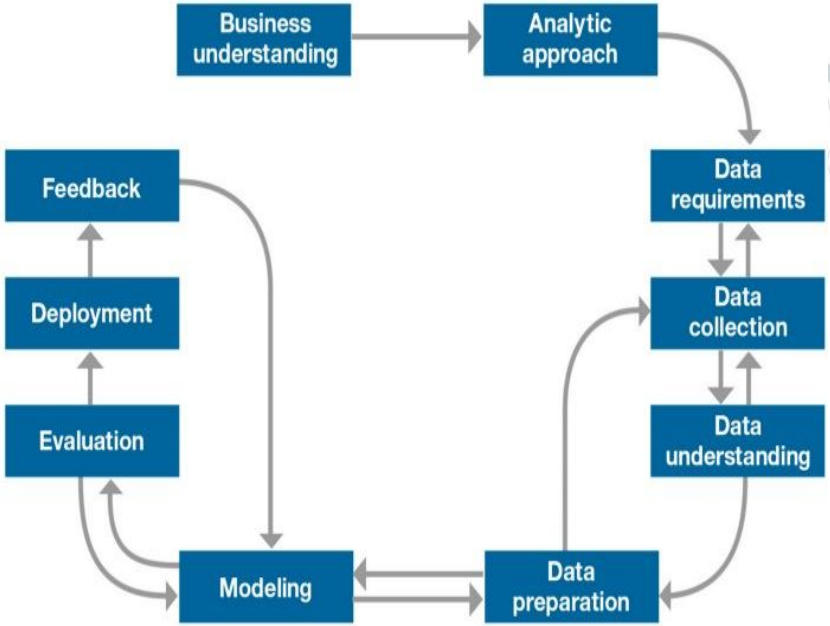


Méthodologie pour la science des données

4. From Modeling to Evaluation:

a. **Modeling** : based on analytics phase developing models either descriptive or predictive.

a. **Evaluation :**Hold-Out and Cross-Validation. **training set,validation set** , **test set**



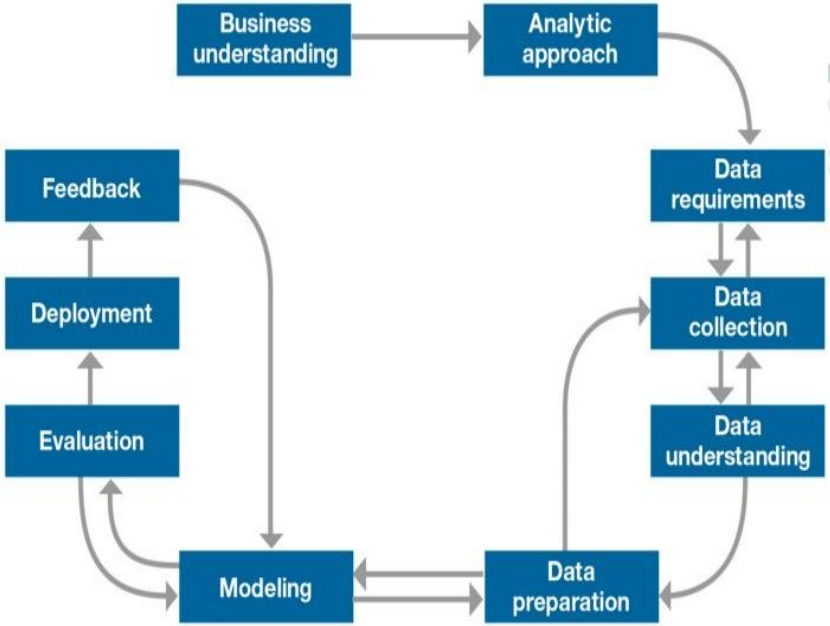
Méthodologie pour la science des données

5. From Deployment to Feedback:

a. **Deployment** :
Rolled out to small group of users to test.

a. **Feedback:**
Get feedback and decide if mode need improvements.

Process from modeling to feedback is highly iterative.



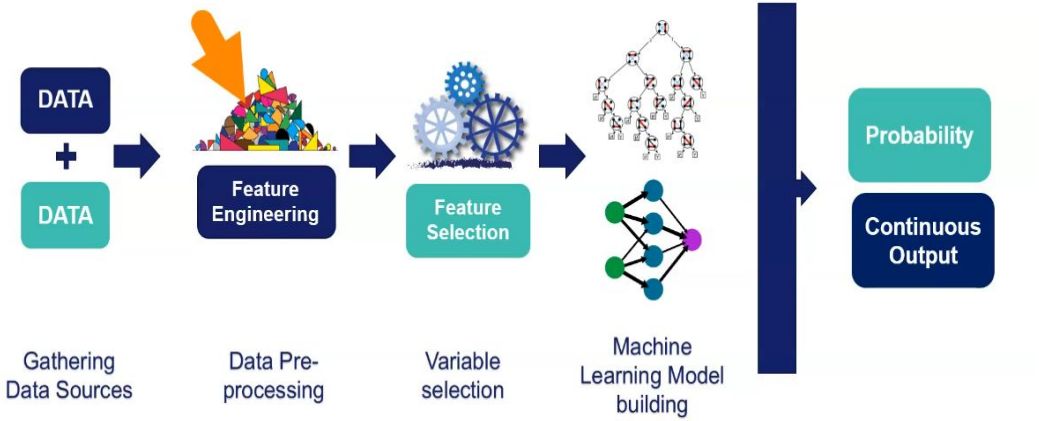
ML Pipeline



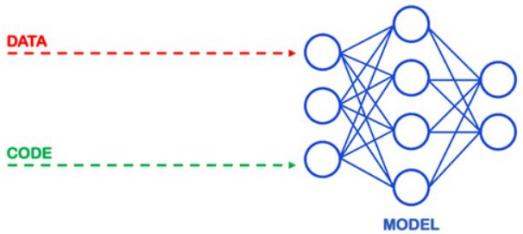



Méthodologie pour la science des données

Machine learning Pipeline



Machine learning = data + code



Méthodologie pour la science des données

Feature Engineering:

- **Missing data** : missing values for certain observations
- **Labels in categorical variables:**
  - Cardinality (high number of labels)
    - lead to overfitting in tree based algorithms
  - Rare labels : infrequent catégories
  - Categories: string
- **Distribution:** Better spread of values may benefit performance
- **Outliers:** The presence of outlier may affect certain models such as linear models and Adaboost..etc)
- **Feature Magnitude - Scale:**
  some Machine learning models are sensitive to scale such as (K means,SVM, KNN , and LDA)

Méthodologie pour la science des données

Feature Engineering:

Technique for missing Data:

- Numerical Variables: Mean / Median Imputation, Arbitrary value imputation, End of tail imputation
- Categorical Variables: Frequent category imputation.Adding a “missing” category

For Both type: Complete Case Analysis, Adding a “Missing” indicator, Random sample imputation

Méthodologie pour la science des données

What makes ML challenging in production?

1. **Dataset dependency :**
   - Many inputs (algorithmic, human, dataset etc.) going to provide output.
   - Difficult to have reproducible, deterministically ‘correct’ result as input data changes
   - ML in production may behave differently than in developer sandbox

     because live data ≠ training data

1. **Heterogeneity and scale**
   - Possibly different engines (Spark, Tensorflow, Scikit Learn, etc.)
   - Different languages (Python, Java, Scala, R ..)
   - Inference vs Training engines
     - Training can be frequently batch
     - Inference (Prediction, Model Serving) can be REST endpoint/custom code, streaming engine,micro-batch, etc.
     - Feature manipulation done at training needs to be replicated (or factored in) at inference
   - Each engine presents its own scale opportunities/issues

3. Collaboration , Process

- Many objects to be tracked and managed (algorithms, models,pipelines, versions etc.)
- ML pipelines are code. Some approach them as code, some not
- Some ML objects (like Models and Human approvals) are not best handled in source control repositories

Méthodologie pour la science des données

Limitations while dealing with ML model in productions:

1. **Data Quality :** better data better business problem solution

1. **Model decay:** real life data changes with the flow of time (should have a continuous management).

1. **Data Locality :** the model pre trained cannot be reused in different market.

Solution :continuous development & continuous integration

Machine learning operations (MLOps)

Key challenges in deployment

a. **Concept Drift:**

   if the desired mapping. From x to y changes x->y

b. **Data Drift:** describe if the input distribution x changes.

c. **Software engineering issues :**

   i. Realtime or batch

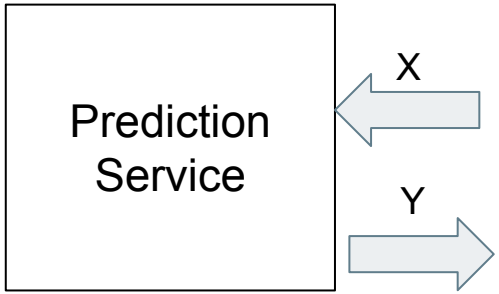
   ii. Cloud vs Edge/Browser

d. **Compute resources (CPU, GPU, memory)**

e. **Latency (QPS)**

f. **Logging**

g. **Security and privacy**



Méthodologie pour la science des données

MLOPS:

Objective: Remove all the manual and time-consuming tasks required for our model creation (such as finding out the required hyperparameter values) with the automated power of DevOps.

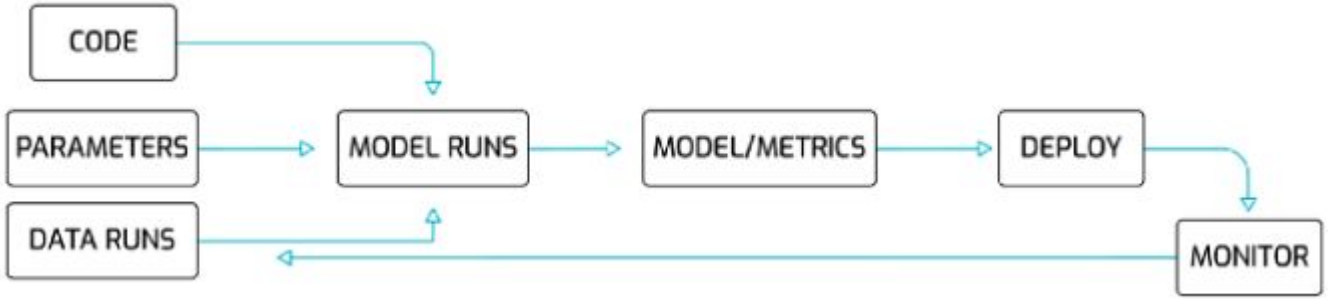
Technologies Used: Containerization (Dockers), Jenkins/rancher , Shell Scripting, Git and GitHub.

Example :

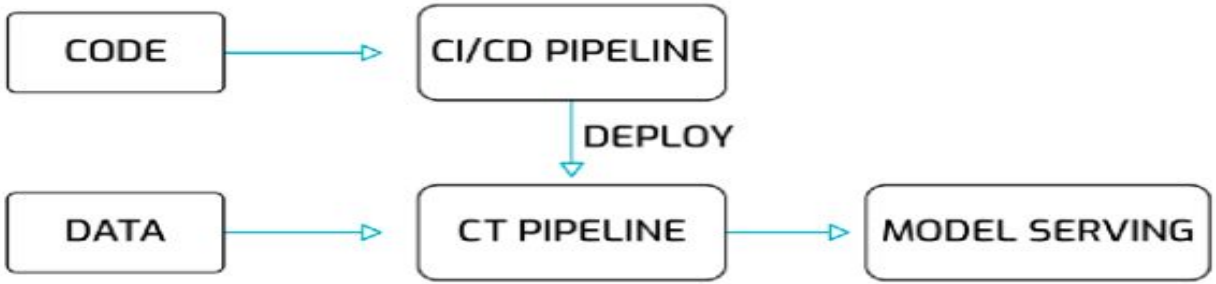
- Create a Dockerfile for setup machine learning environment(scikit-learn, panda..etc) and deep learning(tensorflow,keras ,pillow ..)
- Build the docker image, contonarise the model.
- Set up the github repository
- Create jenkins/rancher jobs
- Add the link to github

Méthodologie pour la science des données

Data Science workflow:



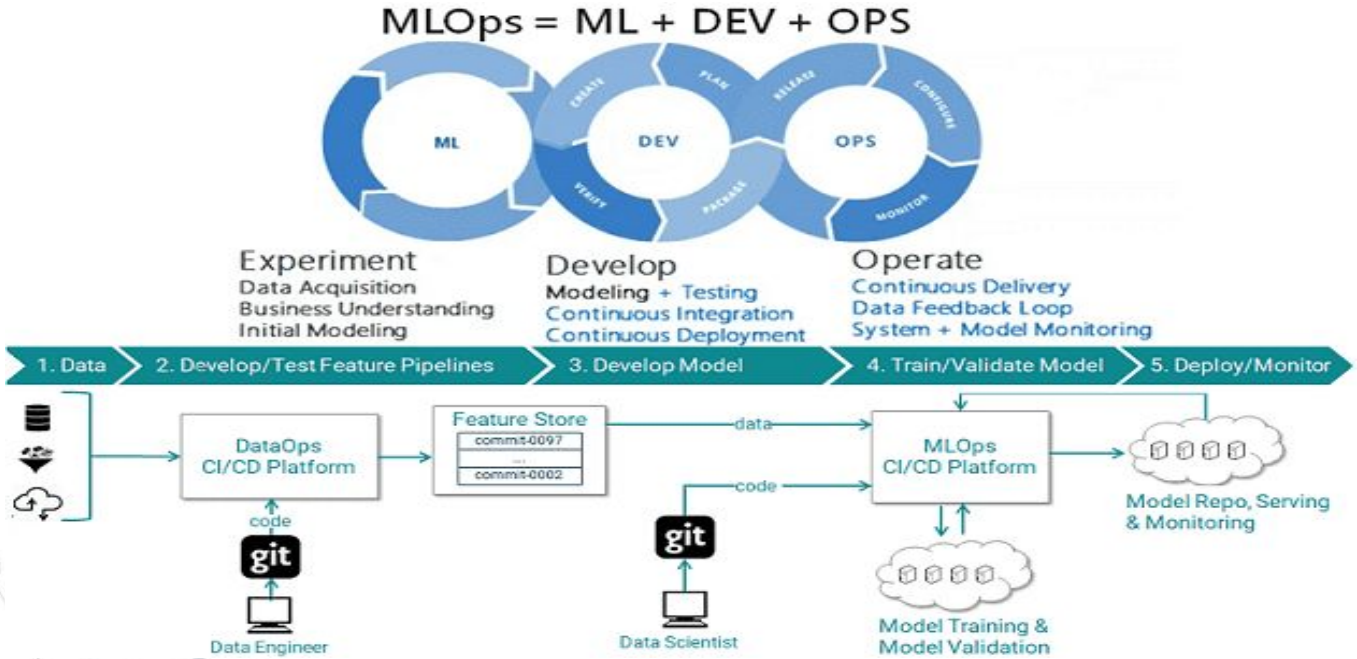
ML pipeline must include Continuous Training:



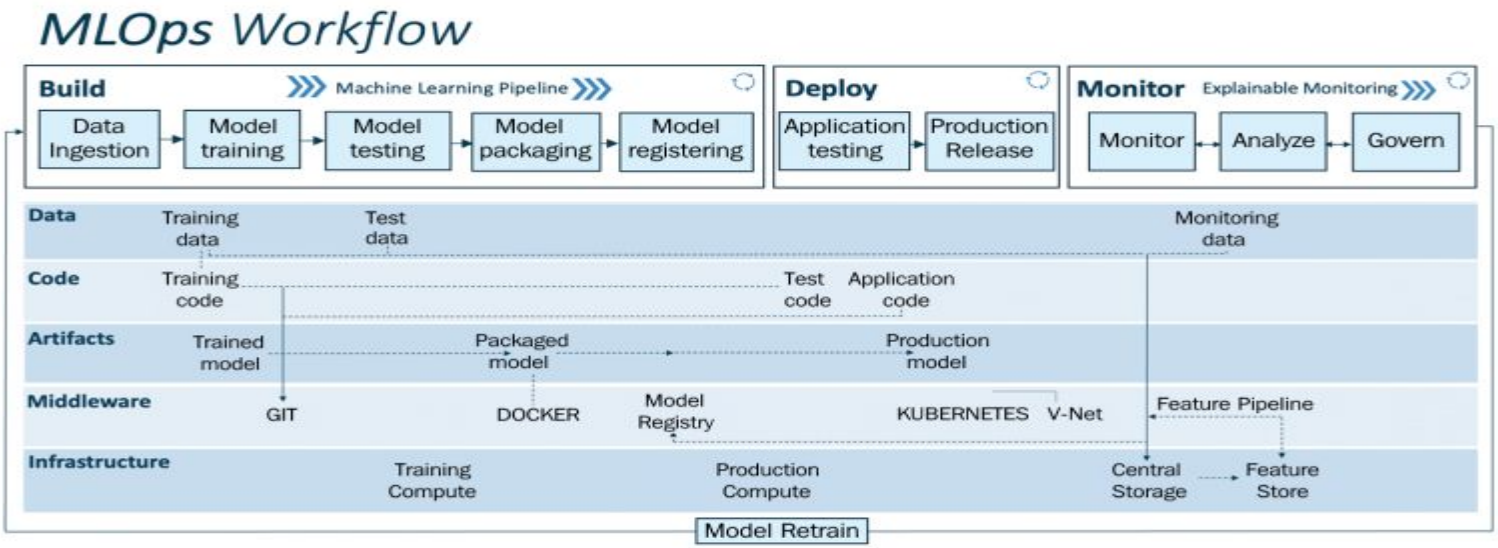
Méthodologie pour la science des données

MI Ops:

Integrate the concepts of DevOps (continuous development & continuous integration) with Machine Learning for automated model creation and its testing.



Data Science in practice:



This workflow is segmented into two modules:

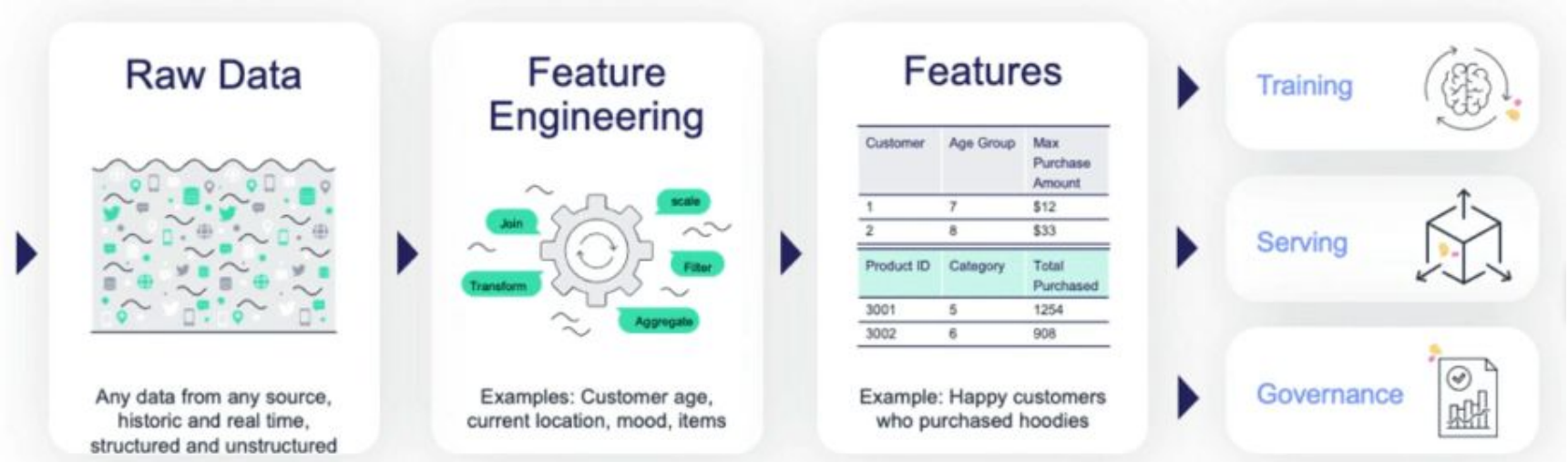
- MLOps pipeline (build, deploy, and monitor) – the upper layer
- Drivers: Data, code, artifacts, middleware, and infrastructure – mid and lower layers

MLOps Stages:

- Stage 1: Model and Data **Version Control**
- Stage2: **AutoML** + Model and Data version control
- Stage 3: AutoML + Model + Data version control **Model Serving**
- Stage 4: AutoML + Model + Data version control Model Serving+**Monitoring , Governance and Retraining**

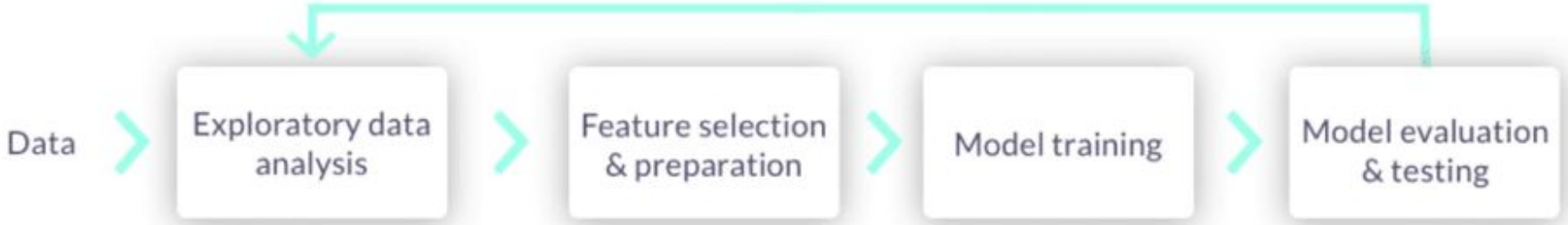
MLOps Stages:

Stage 1: Need access to historical /online data from multiple sources this data need to be in catalog and organized.



MLOps Stages:

Stage 2: AutoML (automation of machine learning process)



All the runs within the data , metadata,code, and results must be versioned and logged.

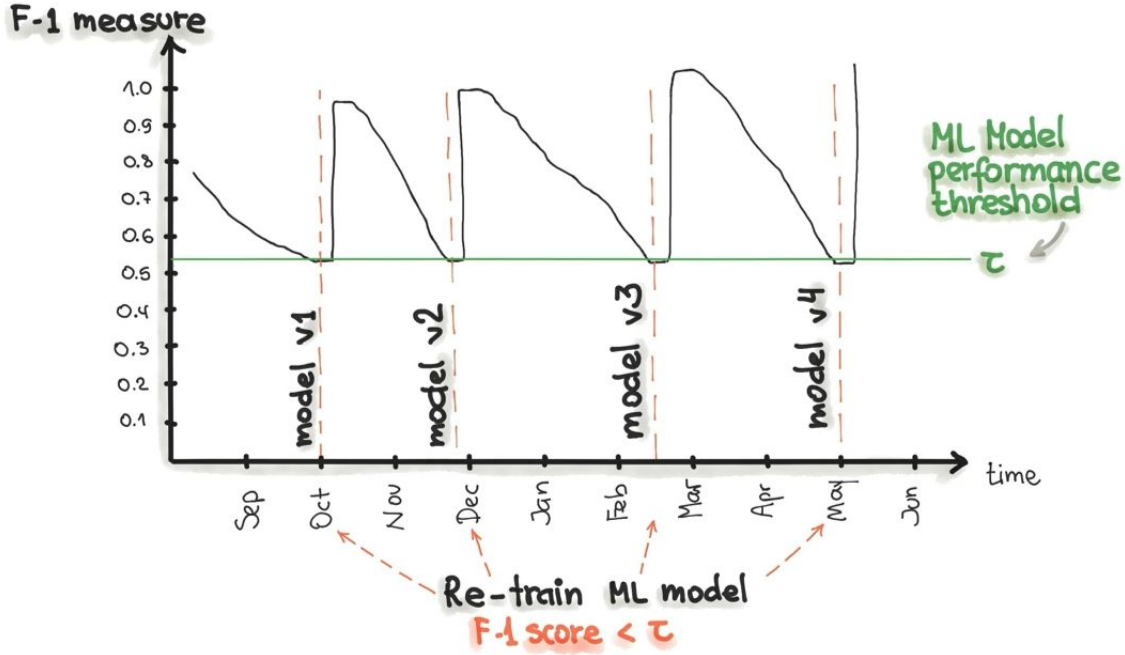
MLOps Stages:

Stage 3: Model serving (create ML services)
integration the model with the business application or front end services

MLOps Stages:

Stage 4: Model Monitoring

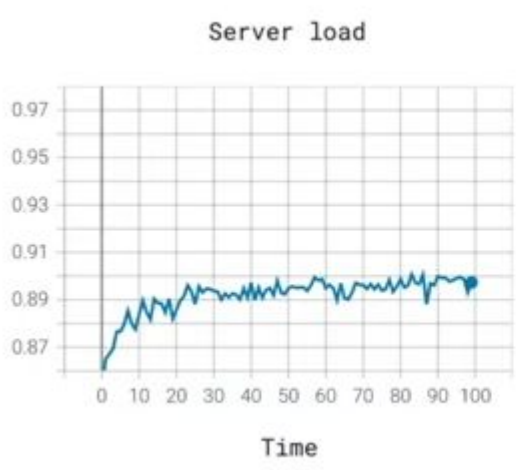
Keep models upto date and predicting with maximum accuracy

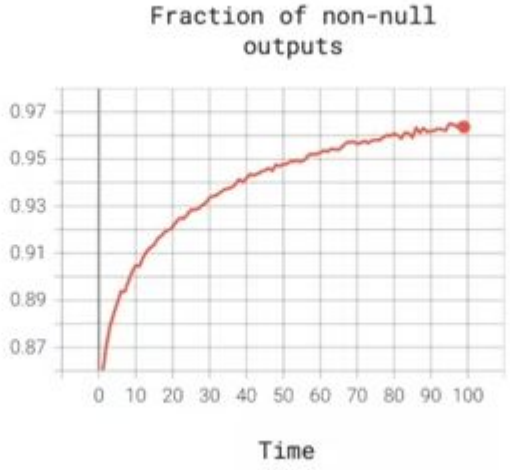


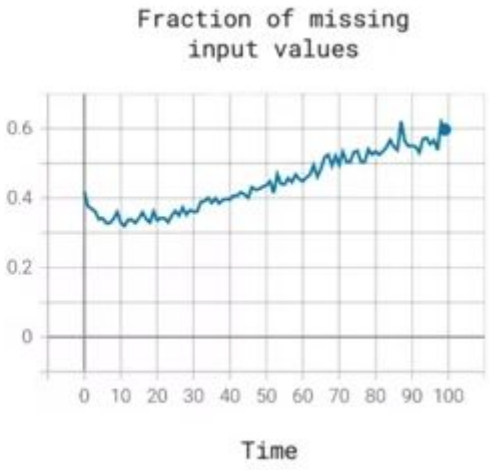
Méthodologie pour la science des données

Monitoring ML systems:

Monitoring Dashboard :







Data Science Tools

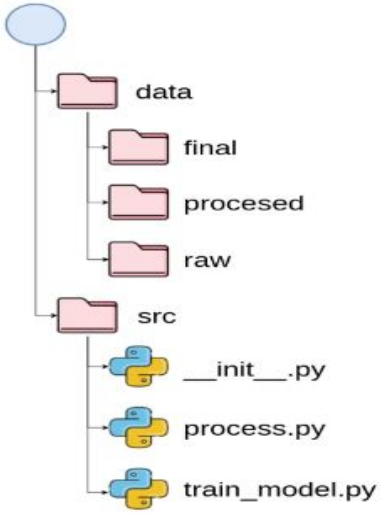


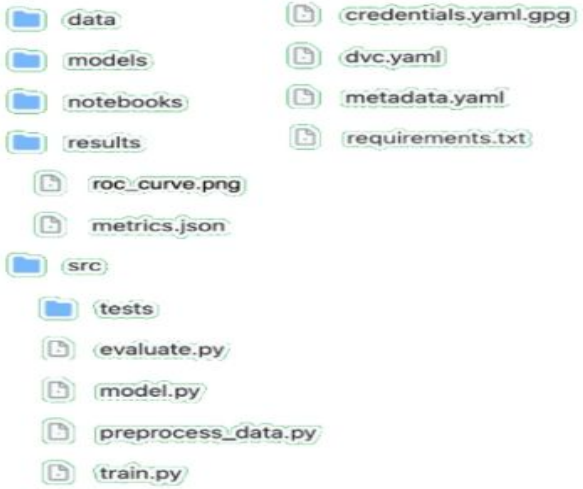



Méthodologie pour la science des données

Structuring ML project:

ML projects should be transferable and documented ! It is important to structure the project according to a standard.






```
.
├── LICENSE
├── README.md
├── data
│   └── README.md
├── metadata.yaml
├── models
│   └── README.md
├── notebooks
│   └── README.md
├── requirements.txt
├── results
│   └── README.md
└── src
    ├── scripts
    │   └── README.md
    └── tests
        ├── README.md
        └── test_australia_weather_predic

7 directories, 11 files
```

Data Science Tools:

- **Project structure tool:**

Cookiecutter: is one of the tools for creating projects folder structure automatically using templates.You can create static file and folder structures based on input information.

https://github.com/cookiecutter/cookiecutter

example : cookiecutter

https://github.com/khuyentran1401/data-science-template

Data Science Tools:

- **Poetry:** Dependency management
- **Hydra:** To manage configuration files
- **Pre commit plugins:** Automate code review and formatting
- **DVC:** Data Version Control
- **pdoc:** automatically project documentation.

Data Science Tools:

- **Anaconda :**
  https://www.anaconda.com/products/distribution

- **Docker according to your system :**
  https://docs.docker.com/desktop/

Data Science Tools:

- **Poetry:** Dependency management. Alternative to installing libraries with pip

Advantages:

- It allows to separate main dependencies and sub dependencies into two separate files (vs requirment.txt)
- Create readable dependency files
- Remove all unused sub dependencies when removing a library
- Avoid installing new libraries in conflict with existing libraries
- Package the project with few lines of code

All the dependencies of the project are specified in **pyproject.toml**

Commands:

Generate project	Install dependencies	To add a new PyPI library	To delete a library
poetry new <proyect-name>	poetry install	poetry add <library-name>	poetry remove <library-name>

MLOps Stages:

PyCaret: is an open source, low code machine learning library.

https://pycaret.org/

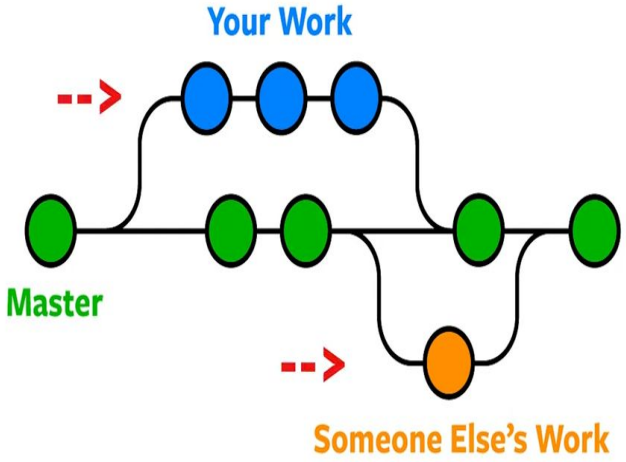
https://github.com/pycaret/pycaret

Méthodologie pour la science des données

Git and GitHub:

Benefits:

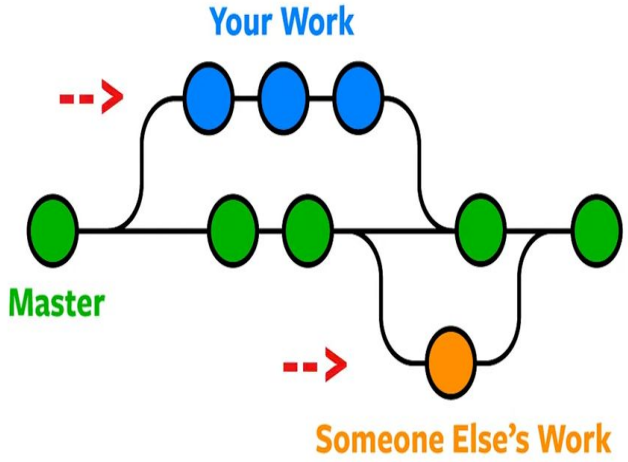
- Track changes and who makes them
- Limit bugs
- Manage concurrent workflows
- Documentation.



Méthodologie pour la science des données

Git and GitHub:

Data Pipeline:
The process of taking data from a source , or many different sources, then process this data and save somewhere else



Data Science Tools:

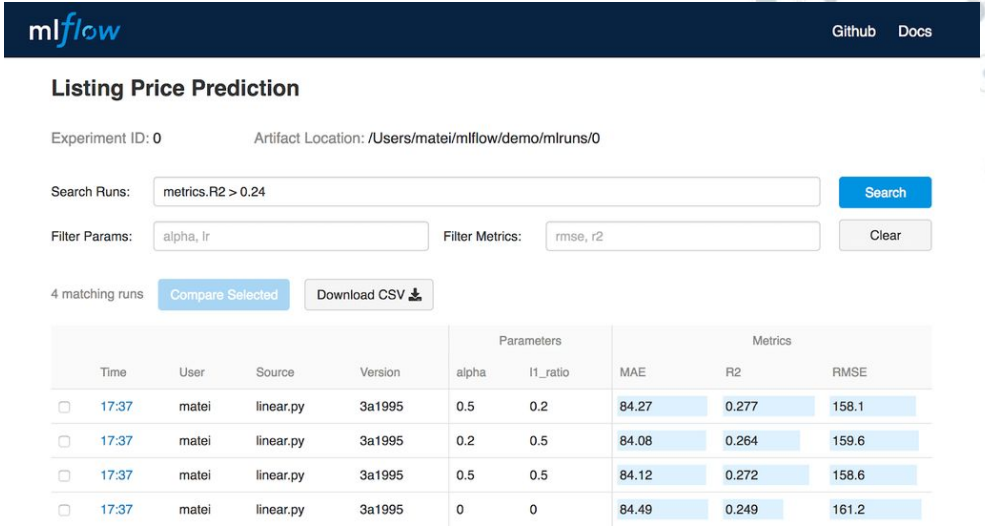
Data Version Control (DVC)

- python written open source tool for Data Science and Machine Learning projects.
- It takes on a Git-like model to provide management and versioning of datasets and machine learning models.
- DVC is a simple command-line tool that makes machine learning projects shareable and reproducible.

Méthodologie pour la science des données

MLFlow

MLFlow is an open source Machine Learning lifecycle management platform that offers various components in experiments tracking, project packaging, model deployment, and registry. MLFlow integrates with various Machine Learning libraries including TensorFlow and Pytorch, to streamline the training, deployment, and management of Machine Learning applications.



Flask App:

Is a micro web framework written in Python. Remote call for your data pipeline

pip install flask

55

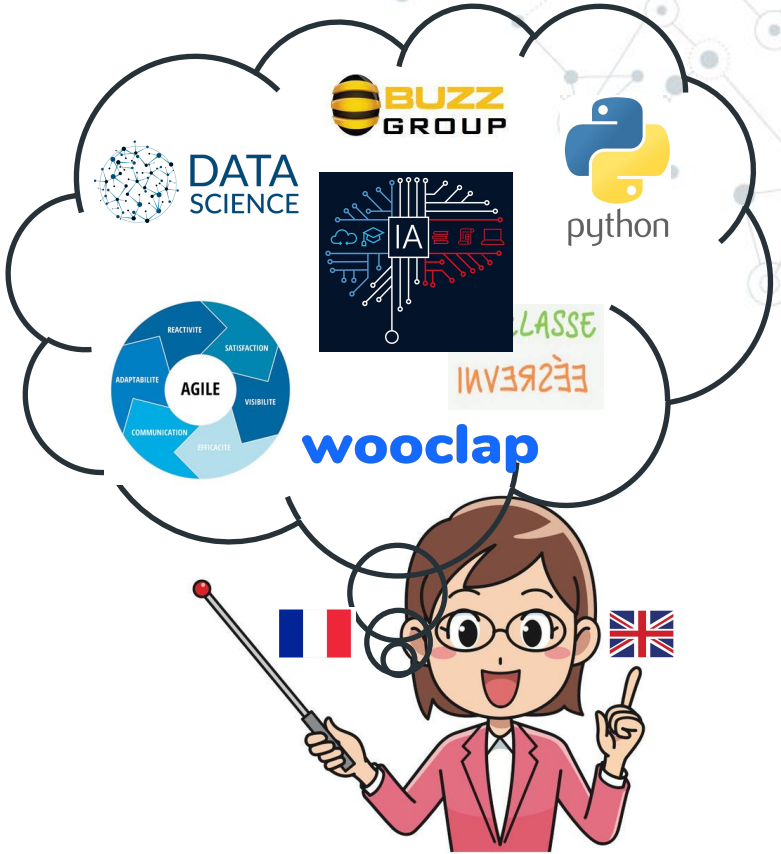
Resources:

1. Introducing MLOps : How to scale Machine learning in the enterprise:

   https://www.oreilly.com/library/view/introducing-mlops/9781492083283/

56

Questions


BUZZ GROUP
DATA SCIENCE
IA
python
REACTIVITE
SATISFACTION
ADAPTABILITE
AGILE
VISIBILITE
COMMUNICATION
EFFICACITE
wooclap