Welcome to the CIT’s Data Cell !

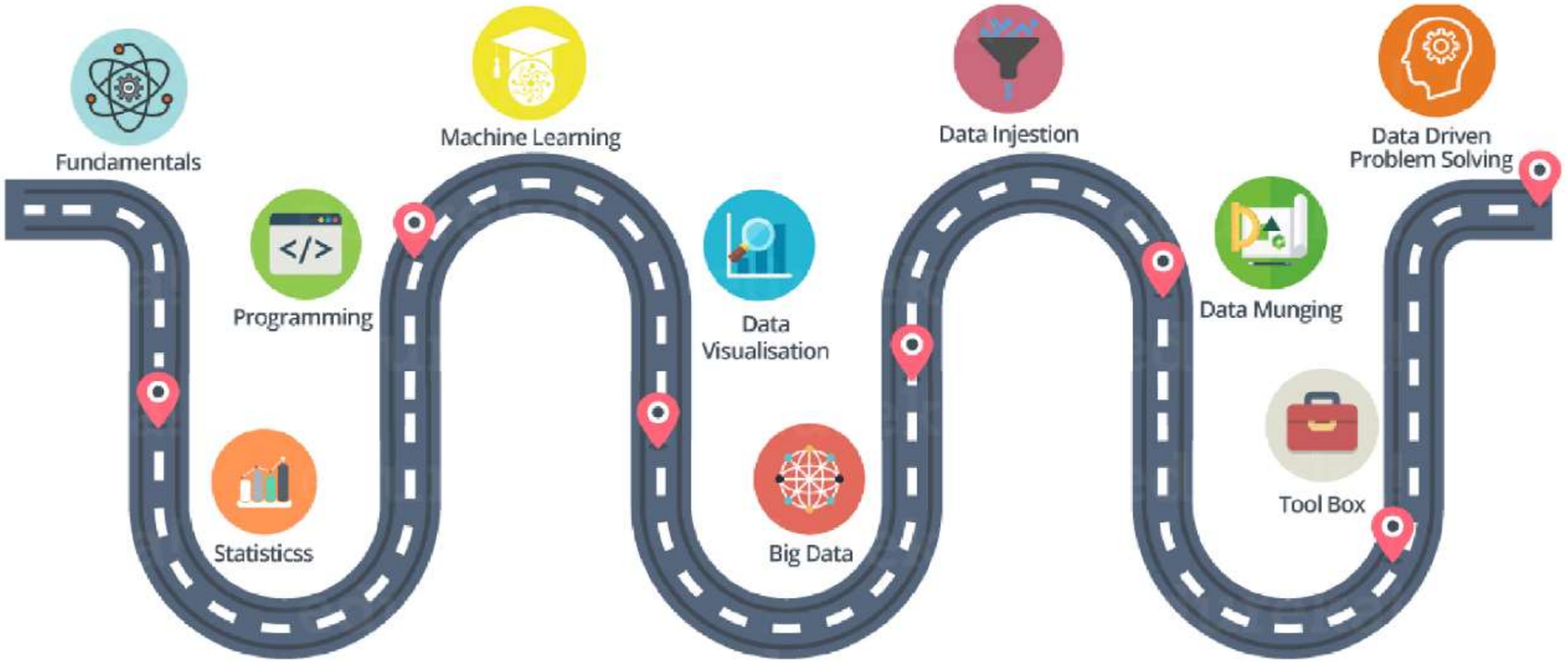
A STEP INTO DATA WORLD

Club Informatique & Télécom
Data Cell

Data Engineer

Collection + Integration	rudderstack	JITSU	Airbyte	SNOWPLOW	nifi	
Warehousing	trino (Formerly PrestoSQL)	Apache Spark	pinot	druid	ClickHouse	
Transformation	dbt	Apache Airflow	PREFECT	beam	Flink	DAGSTER
Data Cataloging	Amundsen	DataHub				
Analysis	Metabase	Lightdash	Apache Superset	jupyter	PostHog	
	re:dash	Plausible	Querybook			

Data Scientist & Analyst

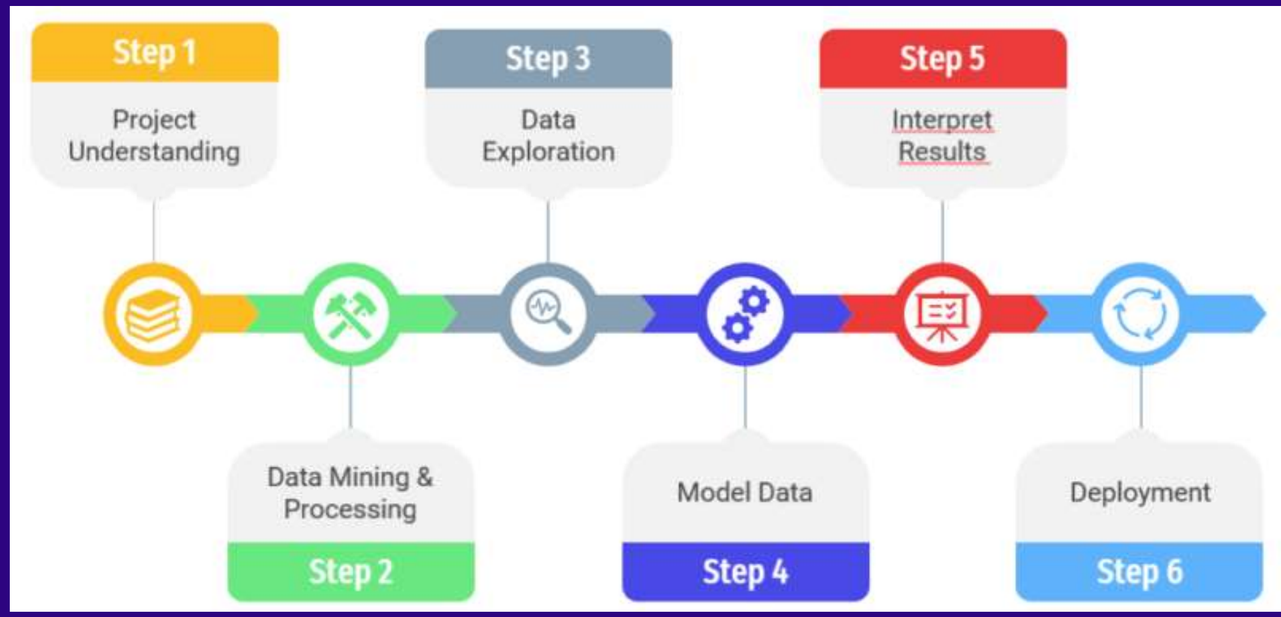


SOME DATA SOURCES:

- Kaggle
- UCI Machine Learning Repository
- Google Dataset Search
- Data.gov
- ...

DS project steps

- Understand goals
- Get data
- Understand data
- Build ML models
- Deploy

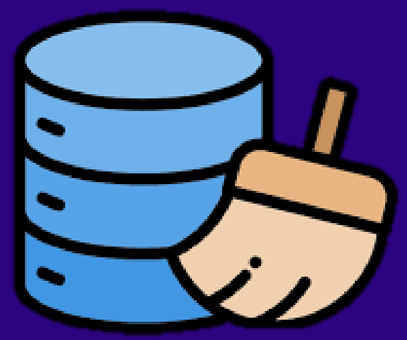


Step 1: Define the Problem

What Question are we trying to answer ?

- Will customers buy this product?
- Which patients are at risk??
- How many cars will be rented tomorrow ?

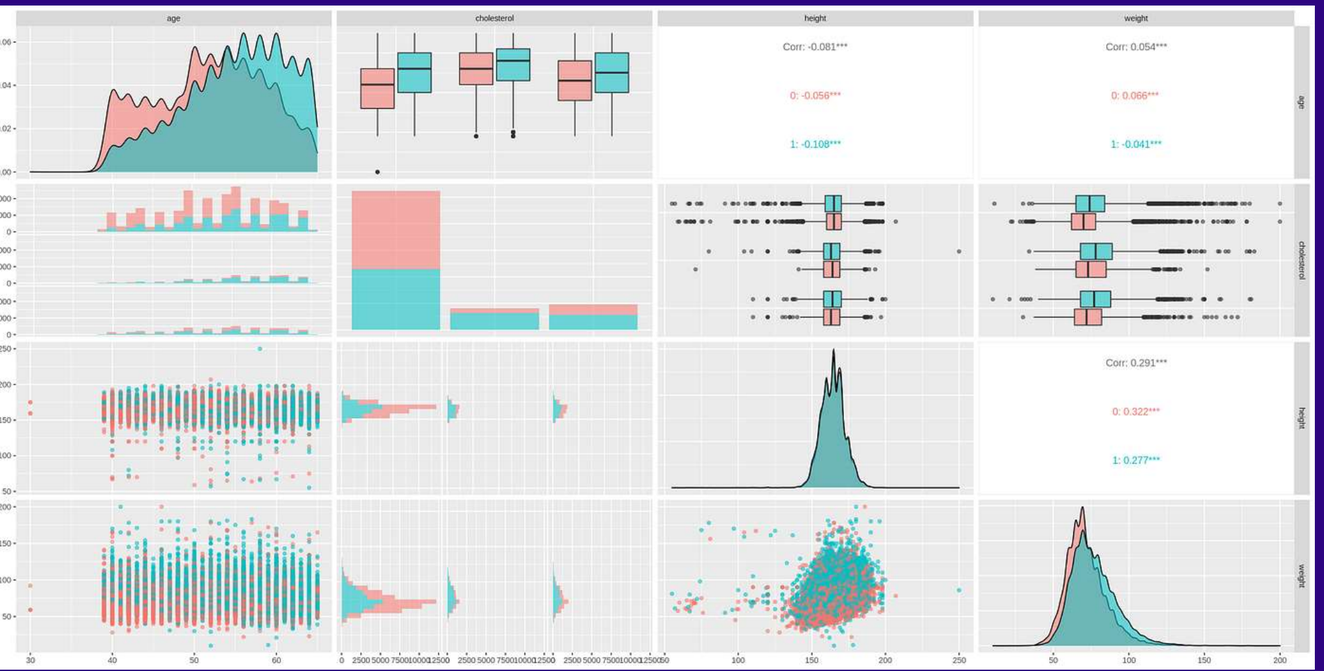
Step 2: Clean & Prepare the Data



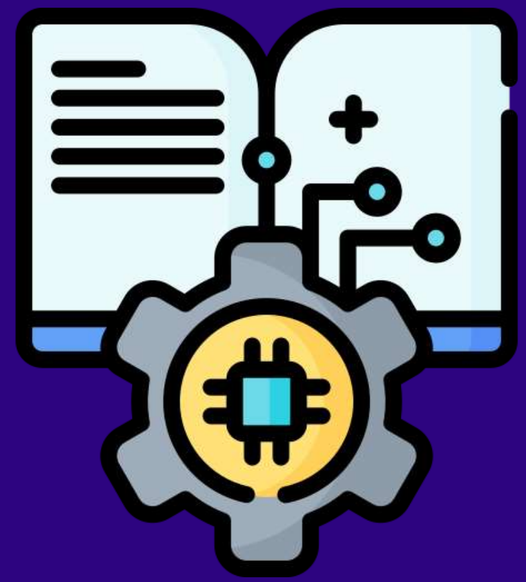
- Handle missing values
- Remove duplicates/errors
- Create useful features (feature eng)

Step 3: Explore the Data (EDA)

- Use statistics and visualizations
- Look for patterns,outliers,trends..
- Create useful features (feature eng)

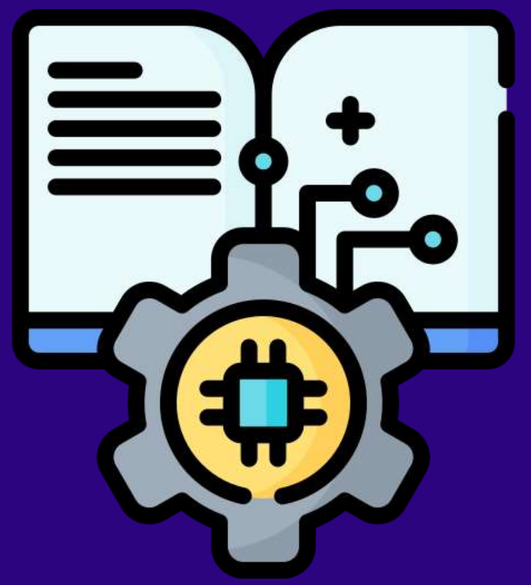


Step 4: Build the Model



- Choose an algorithm
- Train it on train data
- Test it on unseen data

Step 5 : Evaluate



- Check accuracy, other metrics..
- Deploy the model IRL(app,dashboard...)
- Monitor and improve over time


CO
lab

Presenting our partner

Ready, Set, Code!