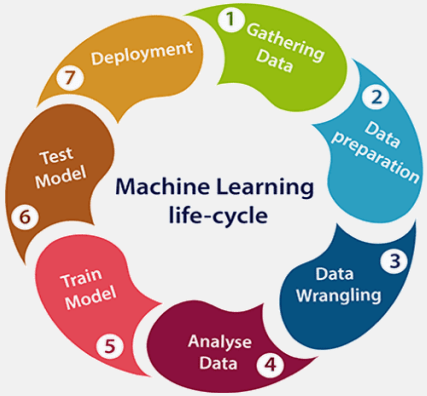
CLOUD COMPUTING APPLICATIONS

Cloud Machine Learning: Workflow

Prof. Reza Farivar

Machine Learning Workflow

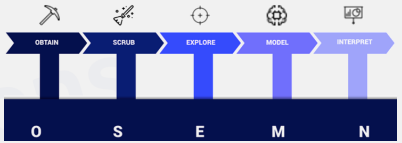
- AI/ML Life Cycle Workflow
  - 7-step model
- The OSEMN Data Science model
  - Obtain
  - Scrub
  - Explore
  - Model
  - Interpret

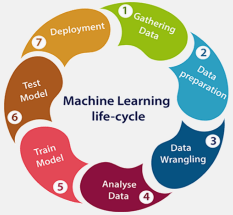


- *A Taxonomy of Data Science* by *Hillary Mason and Chris Wiggins*
- Most providers offer PaaS solutions for the workflow steps

OSEMN: Obtain

- Data Sources on the cloud
  - Amazon: AES Open Data Registry
  - Azure: Open Datasets
  - Google: Cloud Public Datasets

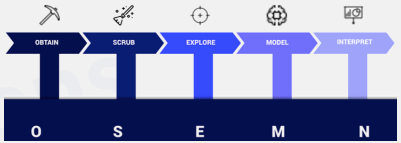


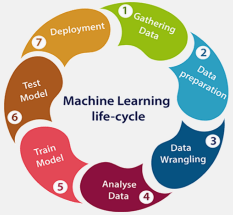


- Command line
- APIs (REST, etc.)
- Jupyter notebooks
- Spreadsheets

OSEMN: Obtain

- Structured Data: Rows and Columns
- Tools:
  - Cloud Storage
  - Cloud Databases
    - SQL
    - NoSQL
      - e.g. MongoDB
  - Big Data
    - Parquet
    - HDFS
    - HDF
    - Pig, Hive



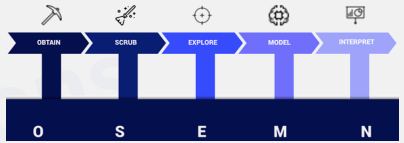


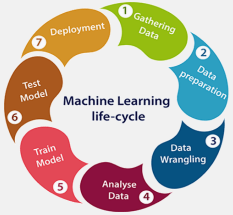
OSEMN: Scrub Data

- Data Preparation and Data Wrangling (2 and 3)
- Clean and filter data
- Consolidate multiple files
- Extracting and replacing values
- Split, merge and extract columns

- Jupyter notebooks

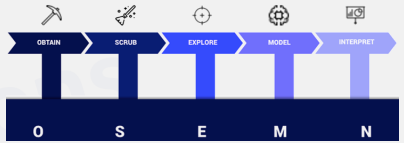
- Python, R for data that can fit in one machine
- Spark, MapReduce for Big Data

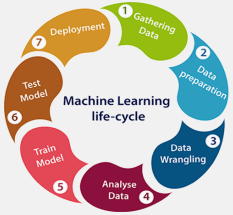




OSEMN: Explore Data

- Inspect data, data wrangling, analyze data (3 and 4)
- Descriptive statistics
- Test significant variables
  - Correlation
- Feature selection
- Data visualization



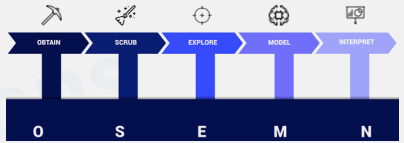


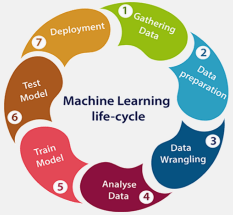
- Jupyter notebooks

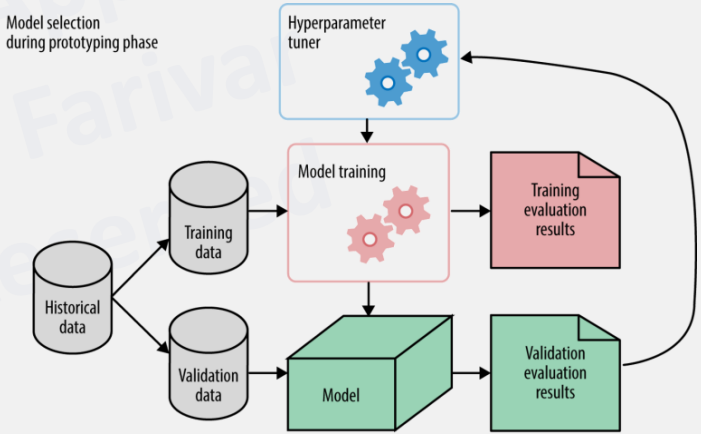
- If data is small → Python, R
  - Numpy
  - Matplotlib
  - Pandas
  - Scipy
- For Big Data
  - Spark
  - EMR

OSEMN: Model Data (1)

- *"Where the magic happens"*
  - Train and test model (5 and 6)
- Feature Engineering
  - Dimensionality reduction
- Model training
  - Regression
  - Classification
  - Clustering
  - Frequent Pattern Mining
  - Decision Trees, Random Forests
  - XGBoost
  - Deep Learning

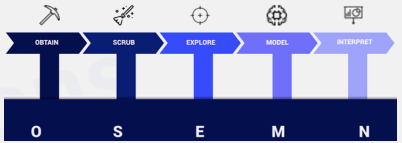


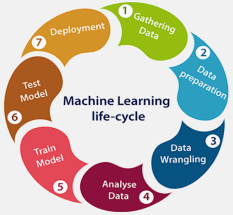




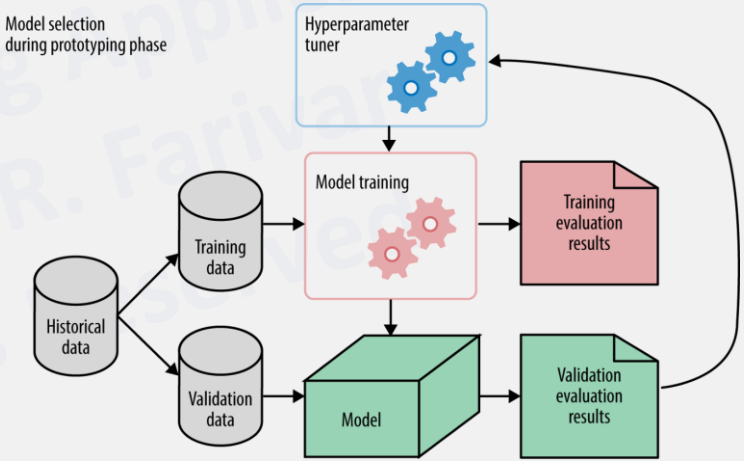
OSEMN: Model Data (2)

- Evaluation
  - Precision
  - Recall
  - F1 Scores
  - Regression
    - MAE (Mean Average Error)
    - RMSE (Root Mean Square Error)



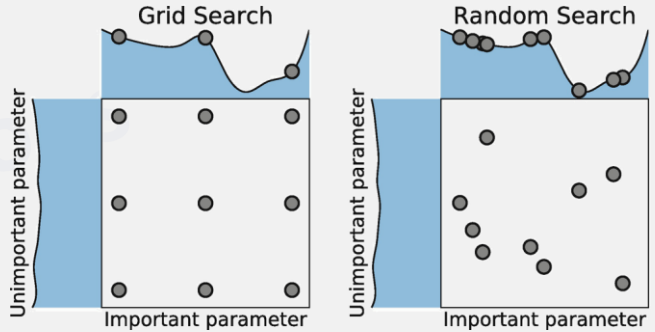


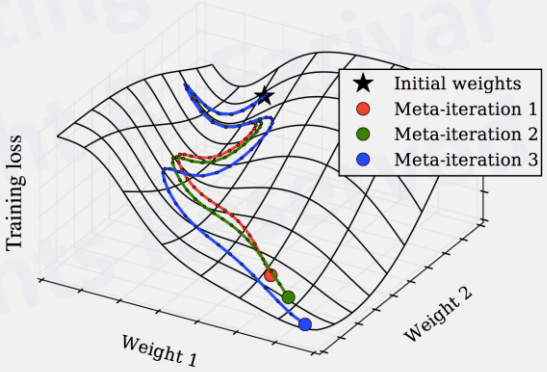
- Small Data: Python, R
  - Scikit Learn
  - H2O
- Big Data
  - Spark Mllib
  - Mahout
  - Google Cloud Dataproc
    - Managed Apache Spark and Hadoop clusters



Hyper Parameter Optimization and AutoML

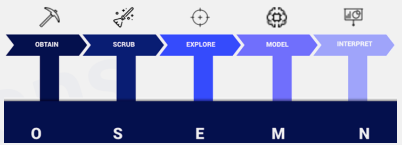
- Hyperparameters: parameters about the training of the model
  - Number of iterations
  - Topology and Size of a neural network
  - Learning rate
- Very time consuming to do manually
- The search space can be huge
- AutoML strategies
  - Grid search
  - Random search
  - Gradient descent
- AutoML vs. Hyperparameter optimization
- Hot competition
  - Azure ML
  - Google AutoML
  - AWS Sagemaker autopilot
  - H2O driverless AI
  - DataRobot

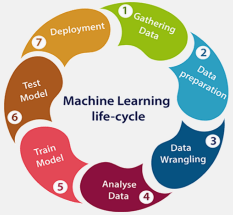




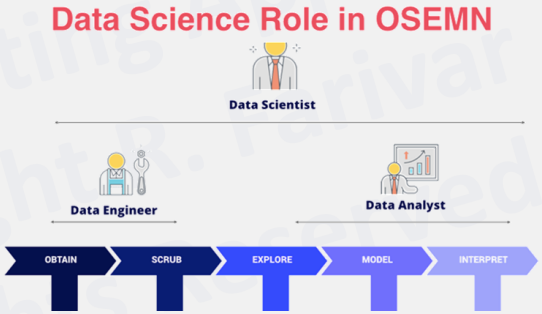
OSEMN: Interpreting Data

- Presentation of the model to a non-technical layman
- Visualizations





- Matplotlib
- Tableau
- D3.js
- Seaborn



Model Deployment

- Model Artifacts
  - Program
  - Parameters
- Keeping the model up to date
  - Data drift detection
  - Model drift detection
  - Version management
- Example: Google AI Platform Prediction