Introduction to Data Science & Machine Learning for Geoscience

Geoscience Training Initiative

Amr.Moslim

Feb.2022

Today's Agenda

- What is Machine Learning?
- Machine Learning Vs Coding ?
- Machine Learning Algorithm Classes
- Machine Learning Algorithm
- Machine Learning Classifications
- Machine Learning Workflow
- How does Machine Learning work?
- Machine Learning Models Evaluation
- Machine Learning Algorithms
- Machine Learning Applications in Geoscience
- The Road map for start learning Machine Learning

What's Machine Learning?

Machine learning is the field of AI that allows systems to learn from past data and make intelligent decisions on their own using algorithms without explicitly programed and improve its experience

Machine Learning vs Coding

Characteristics	Machine Learning Algorithms	Common coding
Objective	To teach the machine to create models to solve the problem without hard coding using data patterns	To use programming language to explicitly code the solution to the problem
Example: v = d/t	Data = (mass, height, width, velocity) Lm = linearregression() Lm.fit() Lm.predict()	Data = (d, t) def velocity(d,t): v = d/t return (v)
Tools	Python, R, Scikit learn, Tensorflow, etc…	Python, R, Visual Basic, Java, Go, Excel
Running time	Most of time in data wrangling and model evaluation	Most of the time in coding the problem and solution
Output	ML model and forecast	Data table, graphs, dashboards
Reproducibility	Yes with the same data formats	Yes with the same data formats
Domain knowledge	It is very important and highly recommended	a must

Machine Learning Algorithm Classes

Supervised Learning

Labeled data

- *Data has labels (reference) model should learn.*
- *Model should be continuously test based on the label prediction or classification.*

Unsupervised Learning

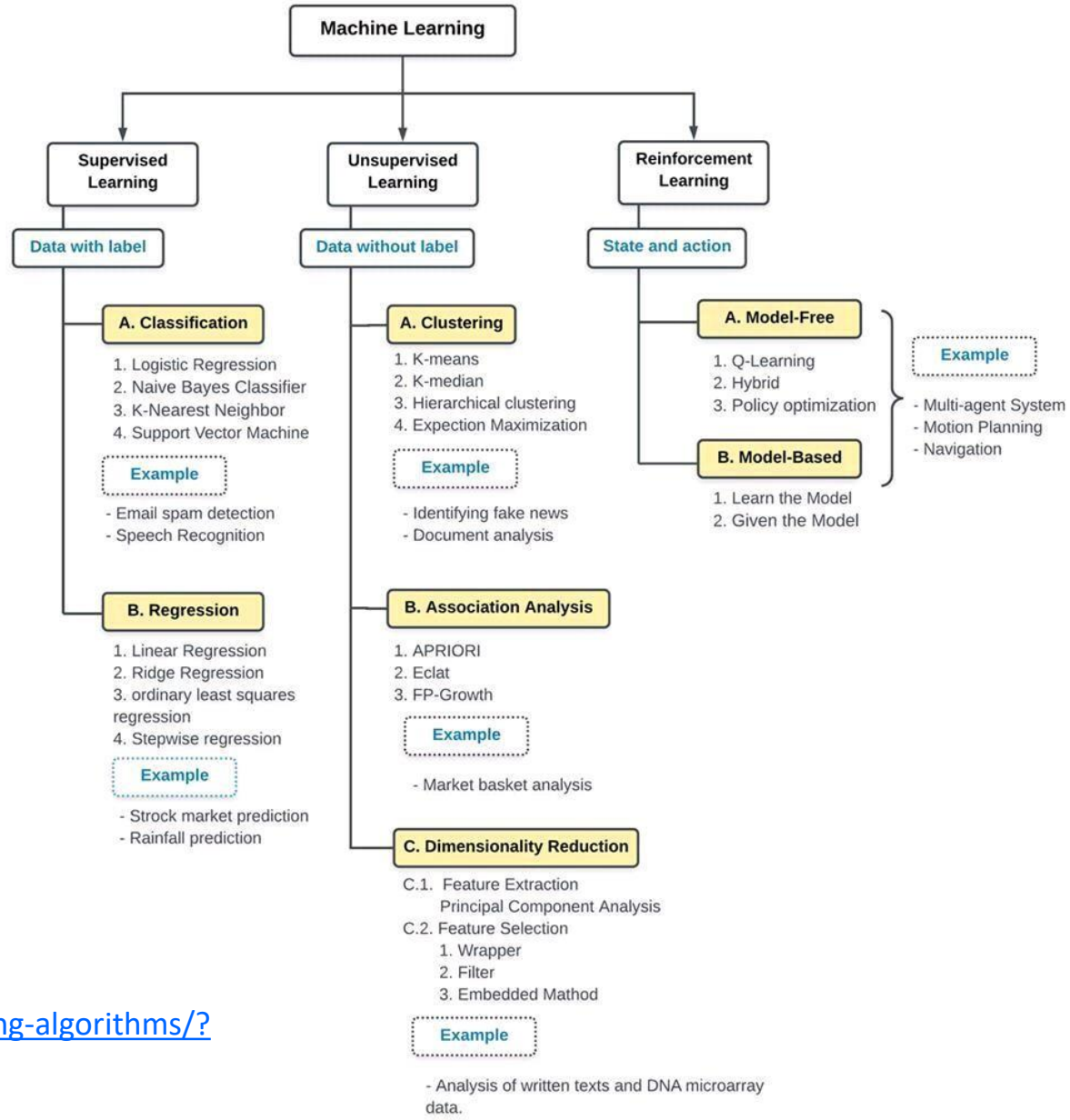
unlabeled data

- *Data has NO labels. Data learn from itself.*
- *Model should be judged based on certain criteria.*

Machine Learning Algorithms

Most commonly used Machine learning algorithms:

1. Linear Regression
2. Logistic Regression
3. Decision Tree
4. SVM
5. Naive Bayes
6. kNN
7. K-Means
8. Random Forest
9. Dimensionality Reduction Algorithms PCA
10. Gradient Boosting algorithms
    1. GBM
    2. XGBoost
    3. LightGBM
    4. CatBoost



https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/?

Machine Learning Algorithm Classification

Supervised Learning

Labeled data prediction

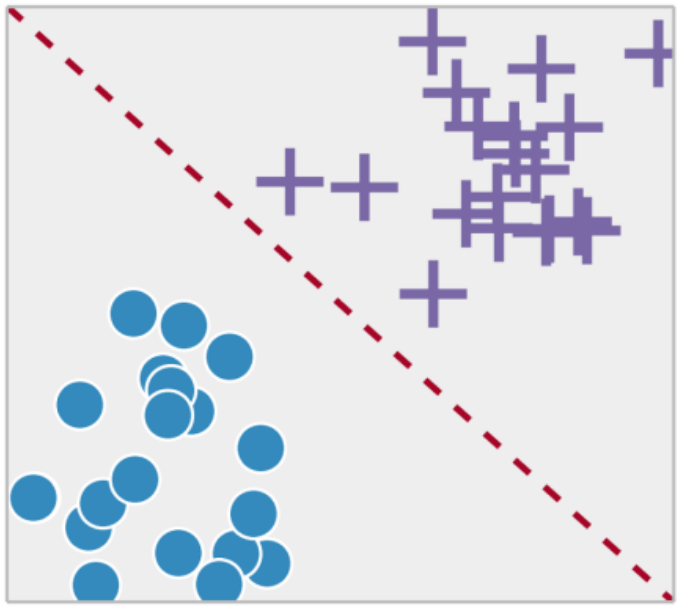
- Regression
- Classification

Unsupervised Learning

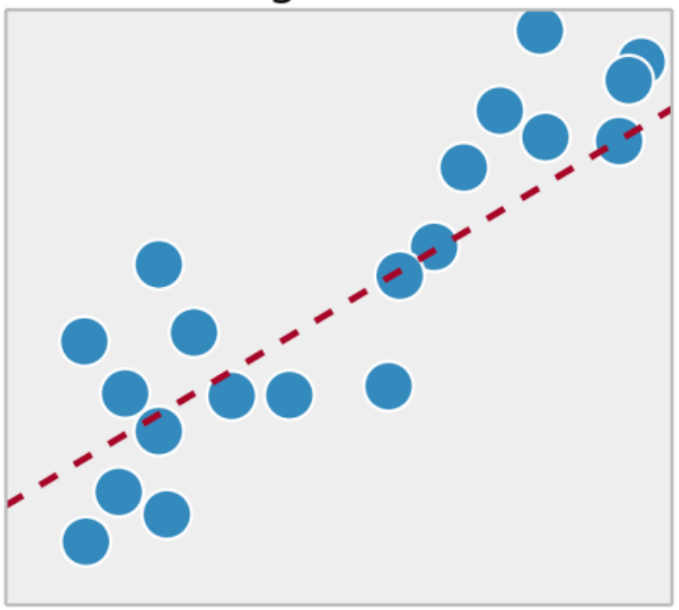
unlabeled data

- Dimensionality reduction
- Clustering

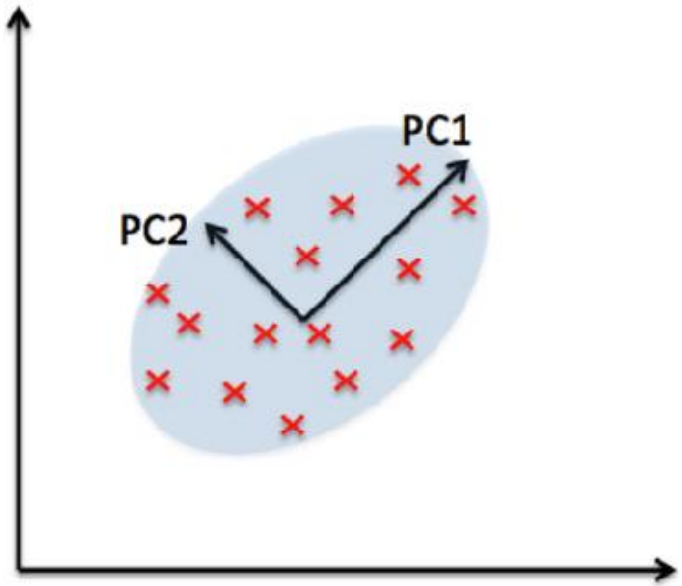
Classification



Regression



Dimensionality reduction



Machine Learning Algorithm Classification

Supervised Learning

Labeled data prediction

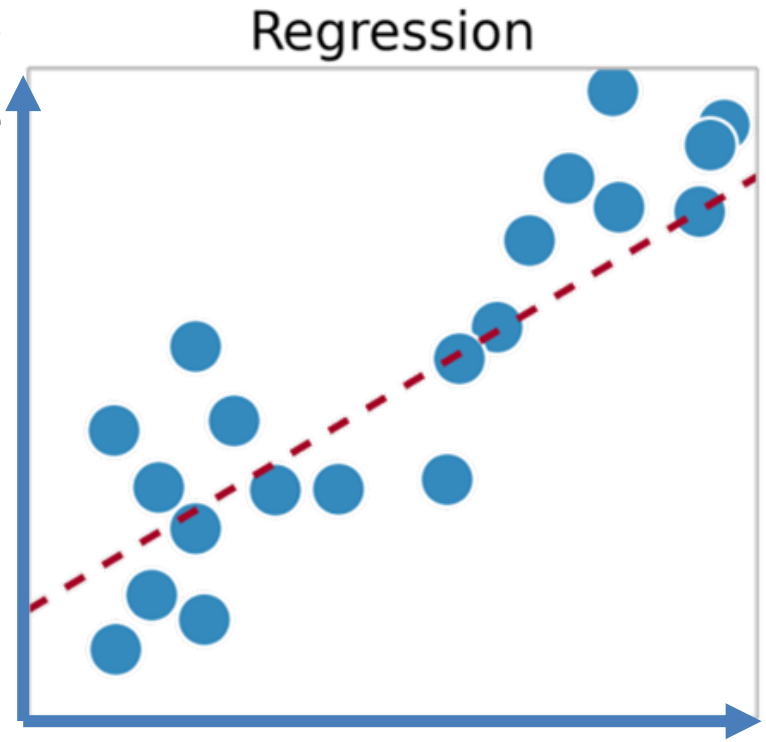
- Regression
- Classification

Regression:

is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables.

- Statistical Modeling Technique
- Types (Linear, Logistic, Polynomial, …)
- Data is numerical values (Not Categorical)

Example : missing logs predication



Machine Learning Algorithm Classification

Supervised Learning

Labeled data prediction

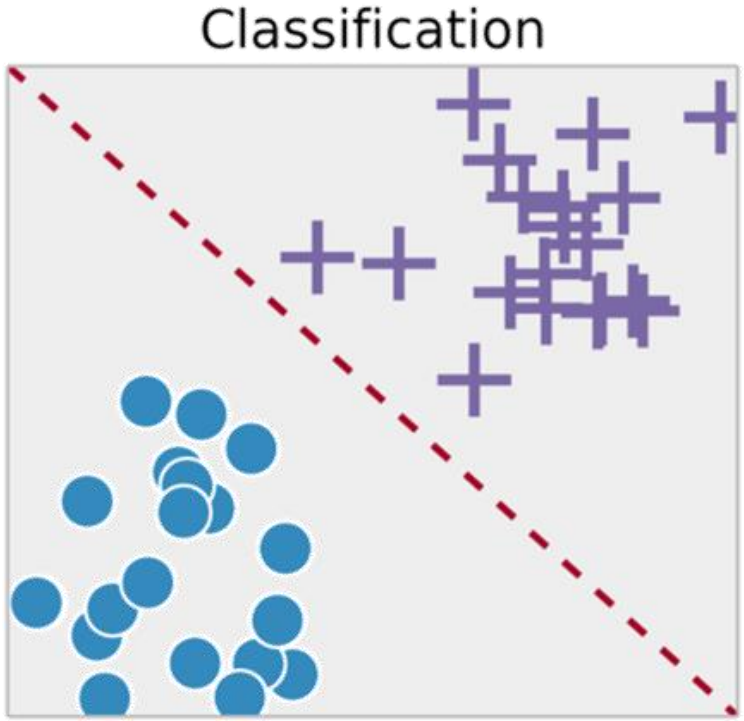
- Regression
- Classification

Classification: (Categorization)

systematic arrangement in groups or categories according to established criteria

- Uses predefined classes
- Belongs to which class

Example : Fraud Detection (Spam / No Spam)
Facies Classification



Machine Learning Algorithm Classification

Unsupervised Learning

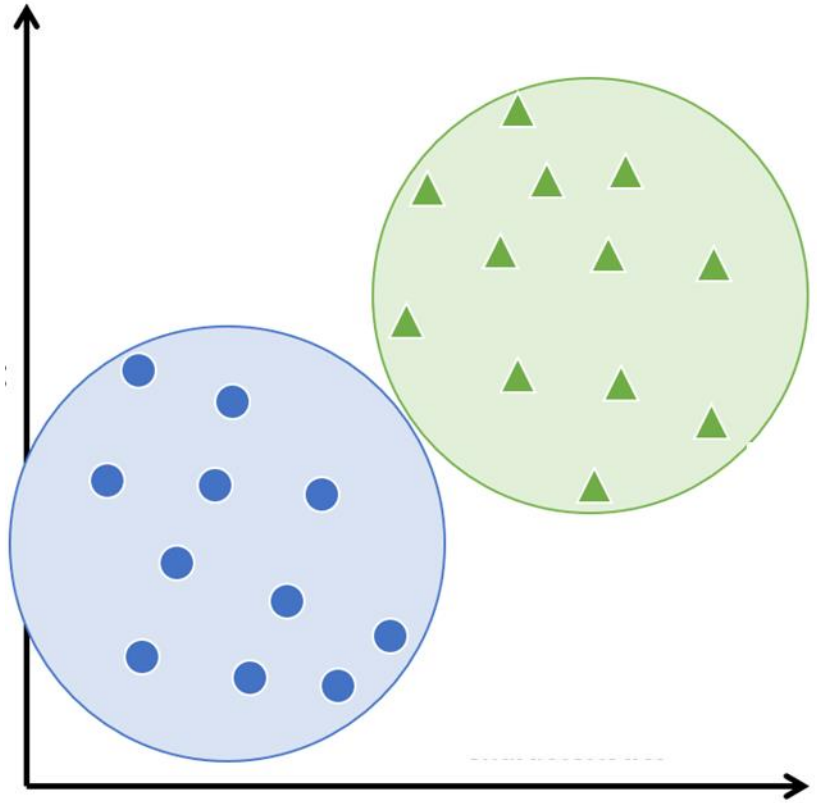
unlabeled data

- Dimensionality reduction
- Clustering

Clustering:

identifies similarities between objects, which it groups according to those characteristics in common and which differentiate them from other groups of objects. These groups are known as "clusters".

- NO predefined classes
- Similar data points properties clusters together



Example : Customer Segmentation

Facies Classification (first time ☺)

Machine Learning Algorithm Classification

Unsupervised Learning

unlabeled data

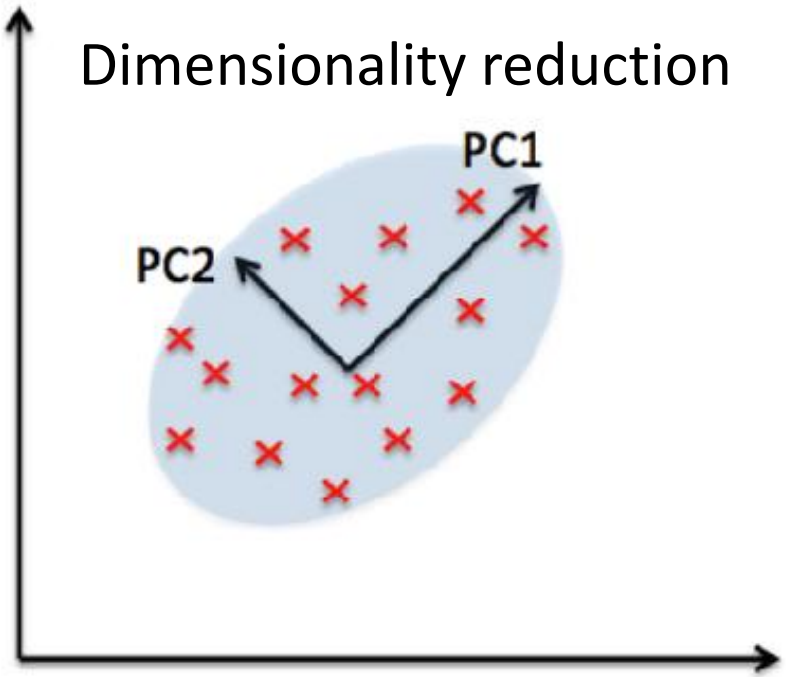
- Dimensionality reduction
- Clustering

Dimensionality Reduction:

Analyzing the datasets with an extremely high number of features
is often performed to obtain better input features for machine learning algorithms.

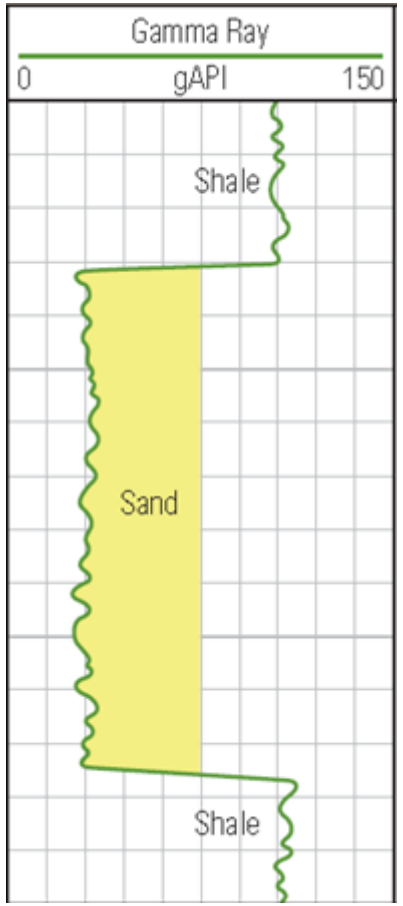
- It improves computational efficiency without sacrificing much on the prediction capability
- removes the collinearity

Dimensionality reduction

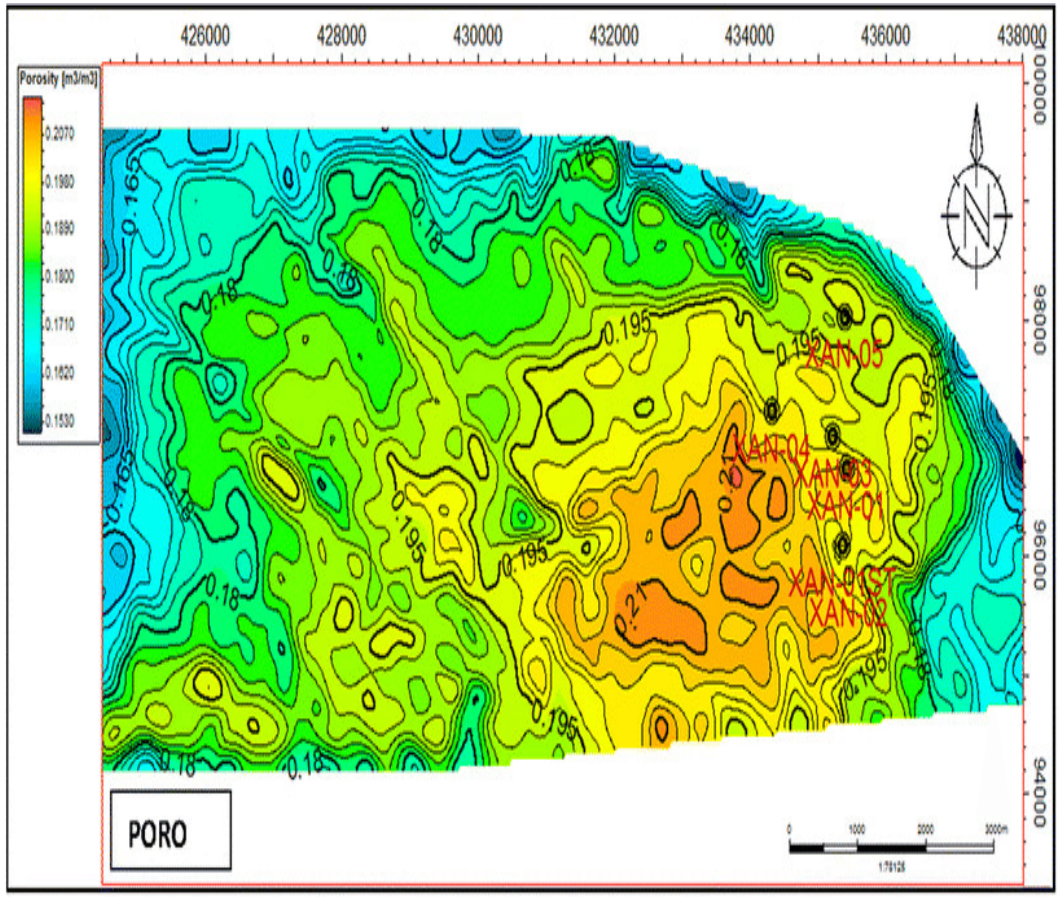


Machine Learning Algorithm Classification

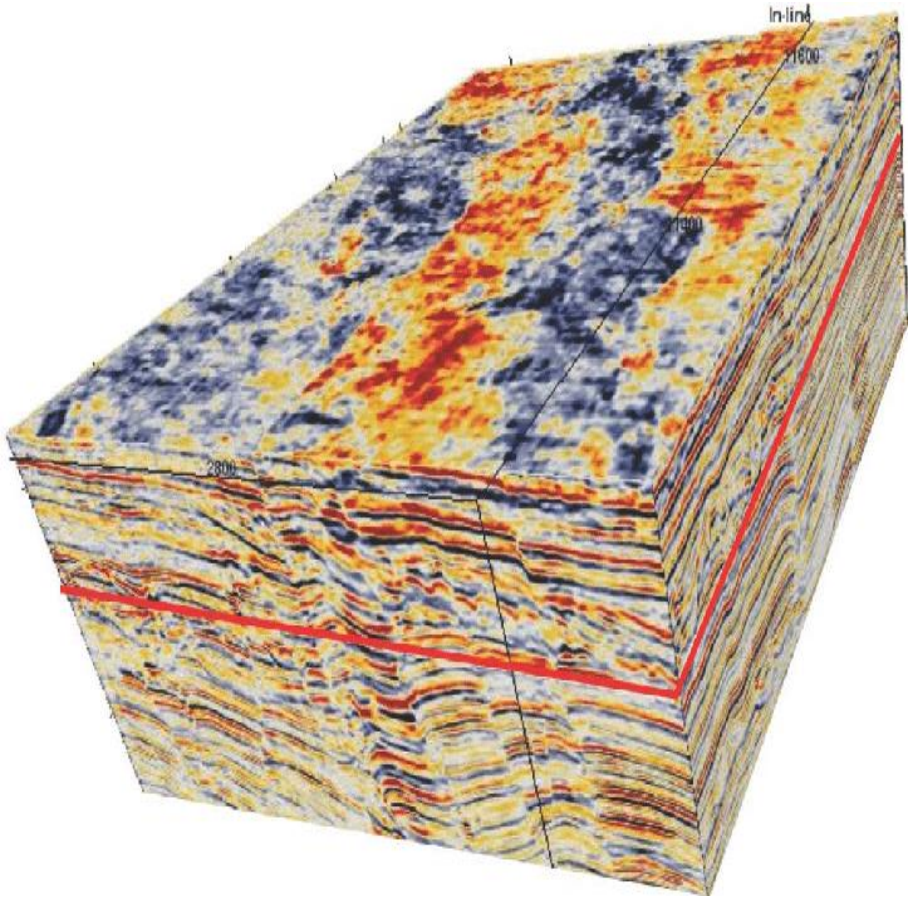
1D Graph



2D Maps



3D Cubes

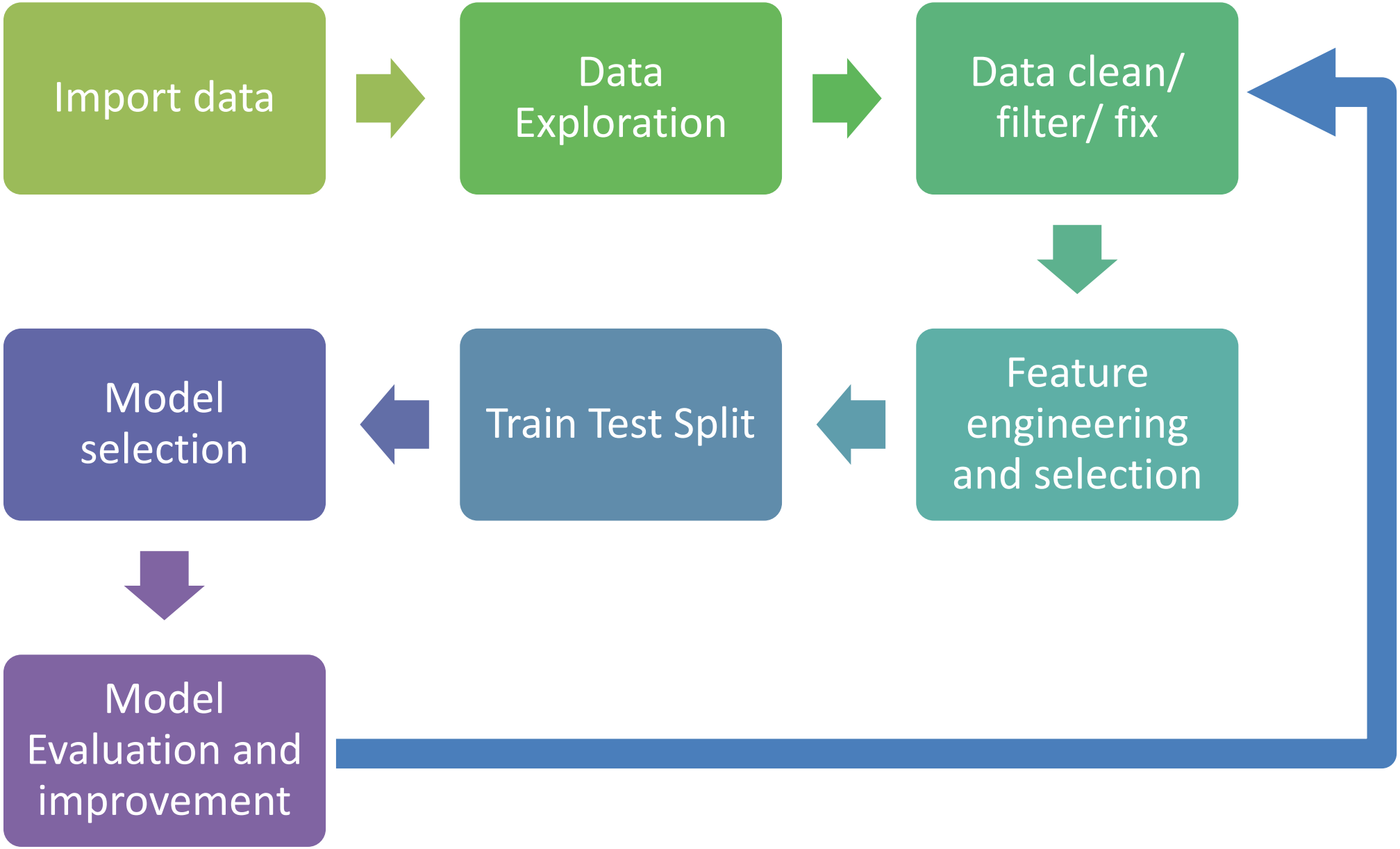


The curse of dimensionality

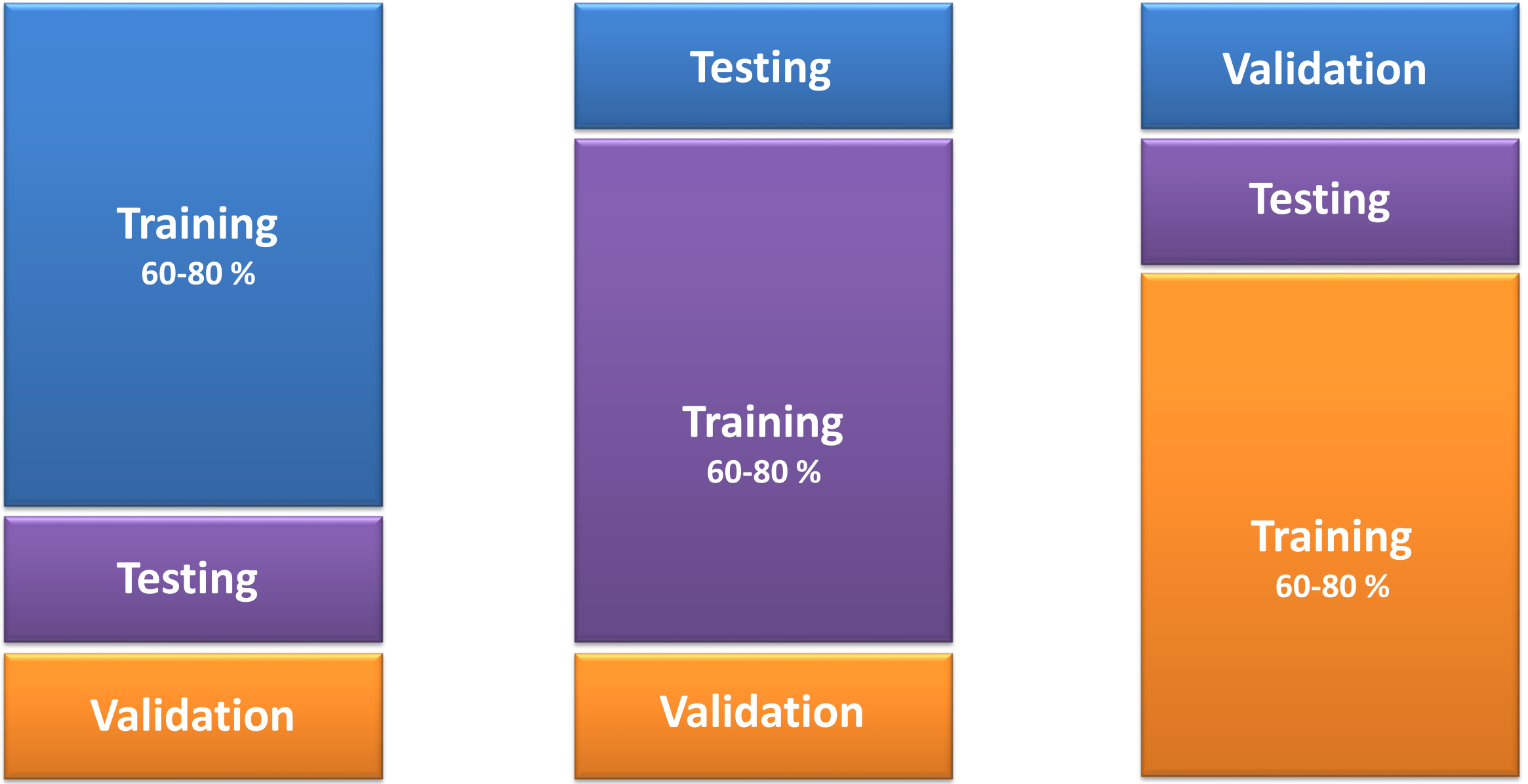
[Richard E. Bellman]

MACHINE LEARNING WORKFLOW

Machine Learning Work flow



Train – Test - Split



ML MODELS EVALUATION

Model Evaluation – Loss Function

Cost Function:

“It is a function that measures the performance of a model for any given data. Cost Function quantifies the error between predicted values and expected values and presents it in the form of a single real number”

Types of Cost functions:

- **MSE**

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- **RMSE**

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}}$$

- **R2**

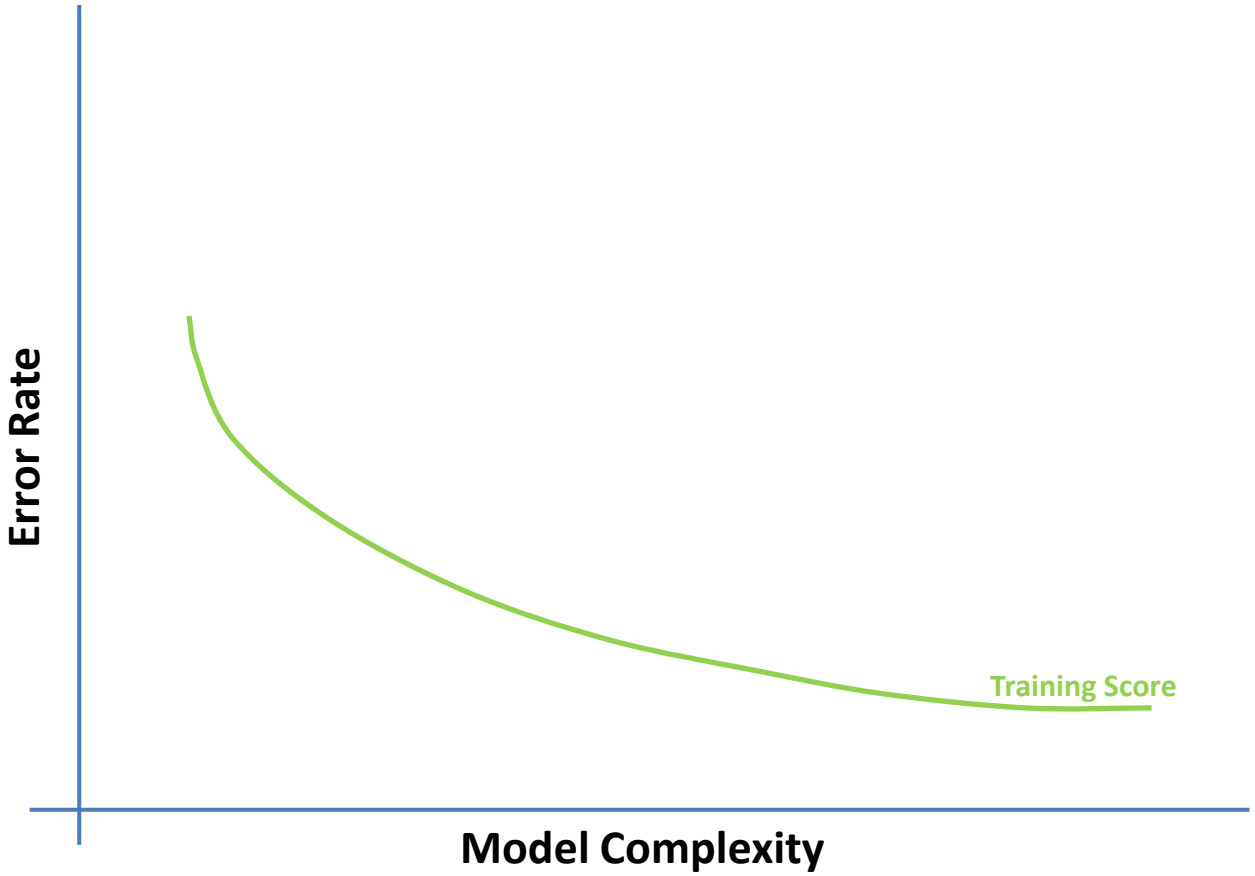
$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 = coefficient of determination

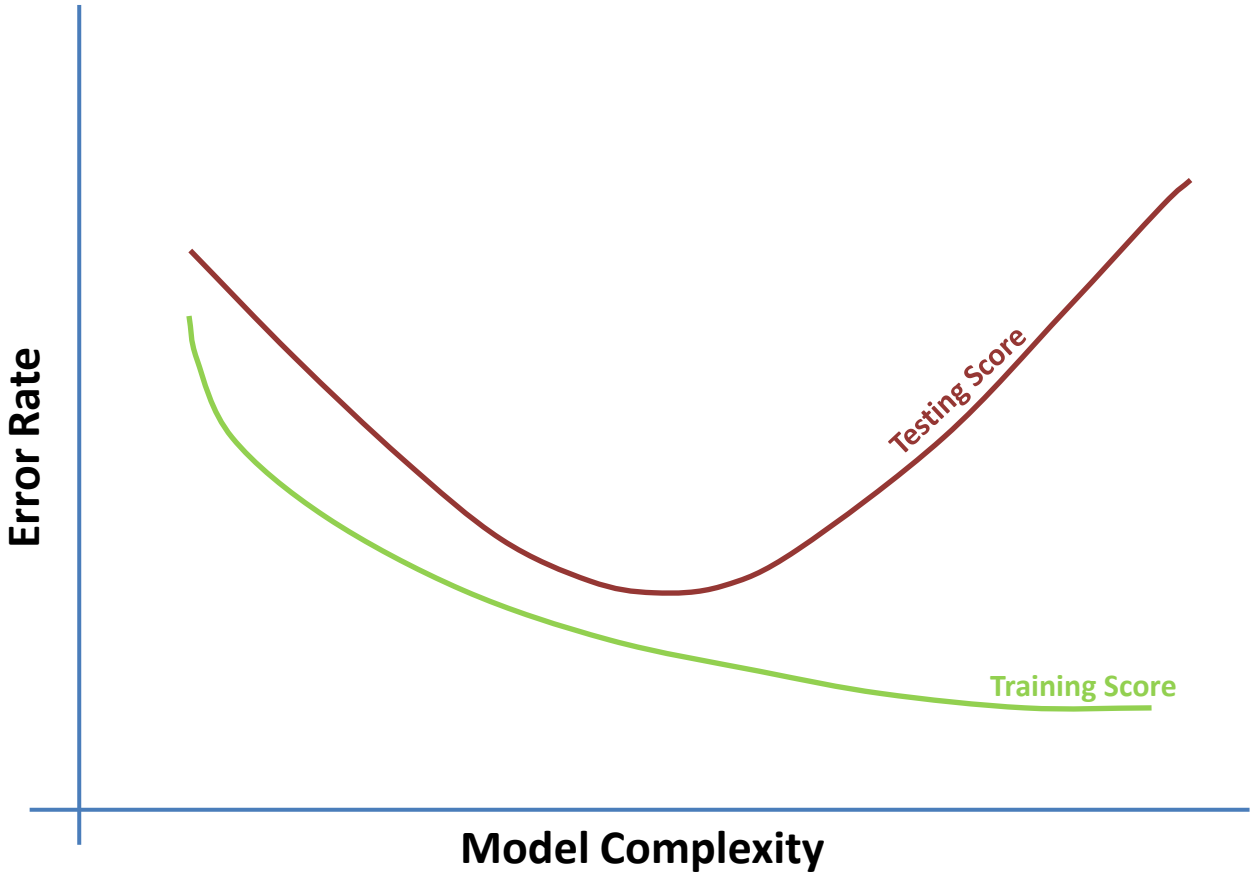
RSS = sum of squares of residuals

TSS = total sum of squares

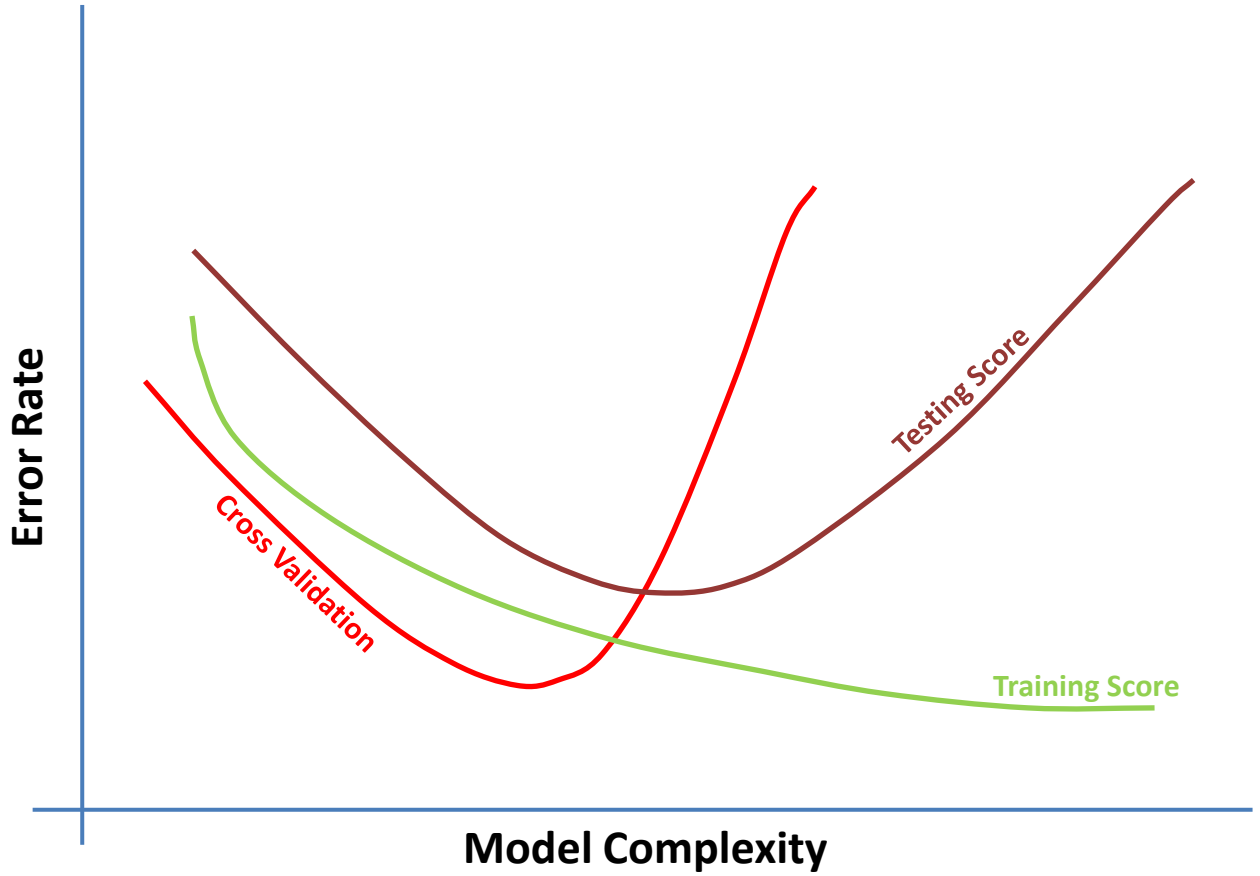
Model Evaluation - Error vs Model Complexity



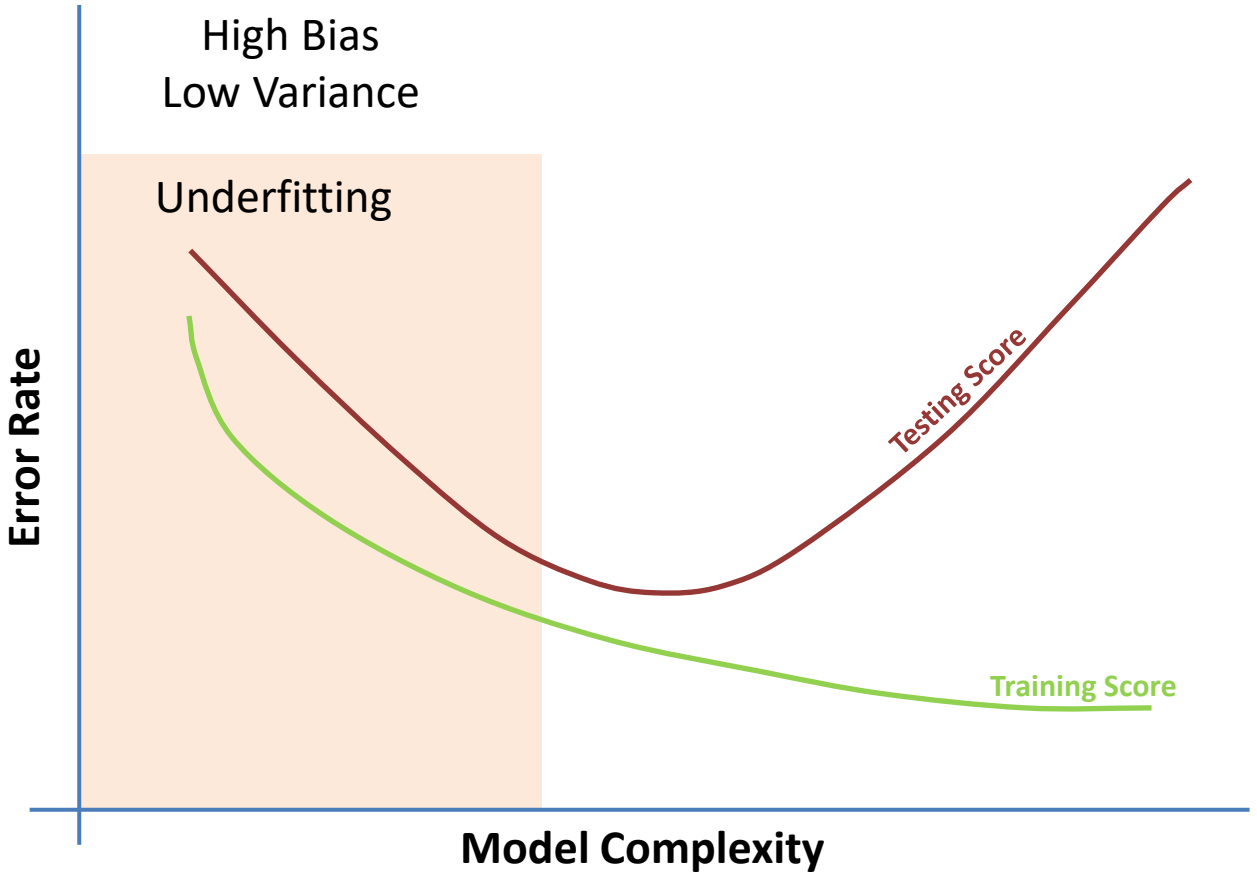
Model Evaluation - Error vs Model Complexity

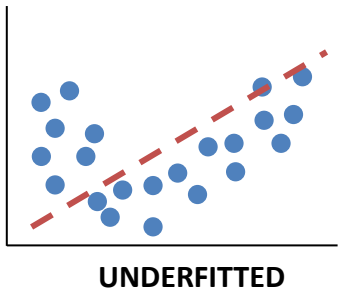


Model Evaluation - Error vs Model Complexity

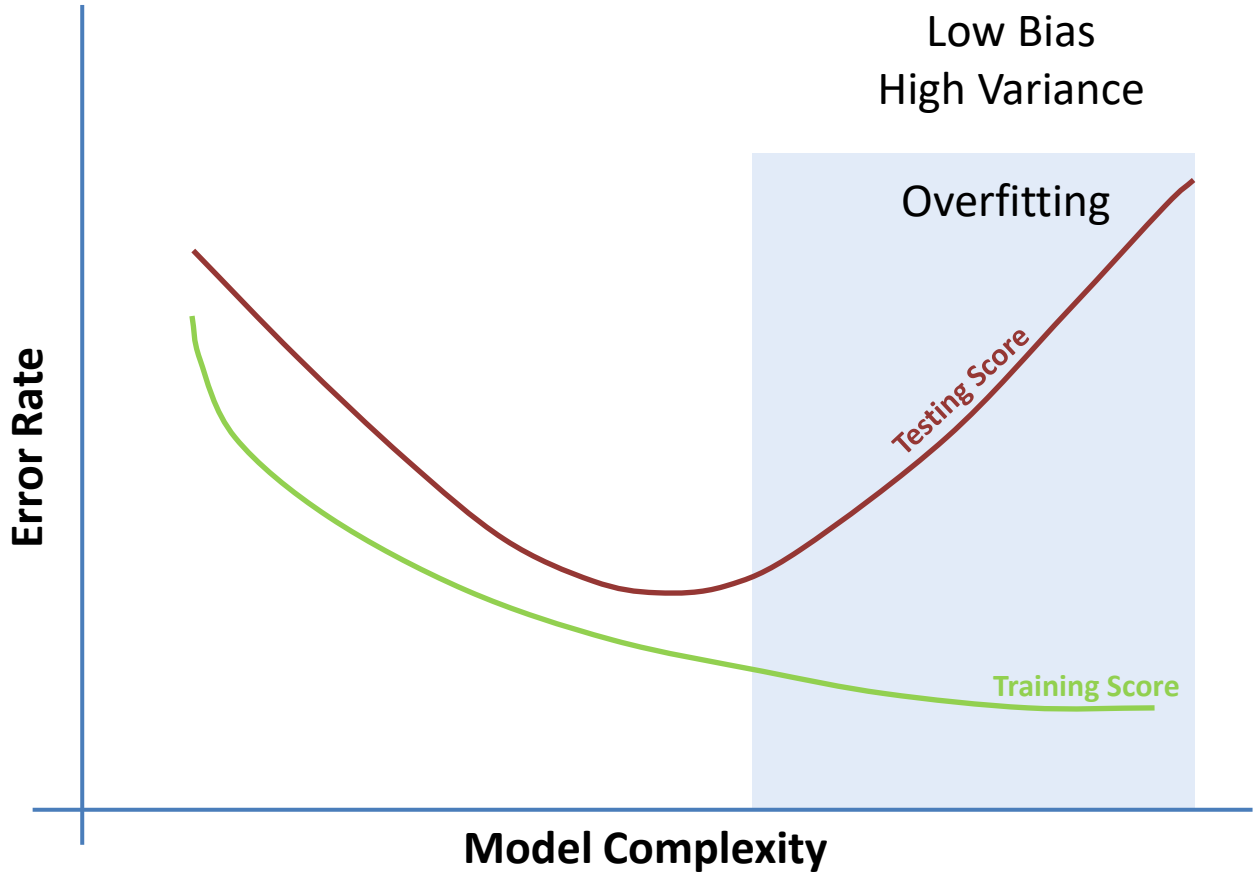


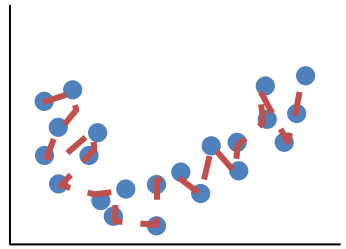
Model Evaluation – Underfitting



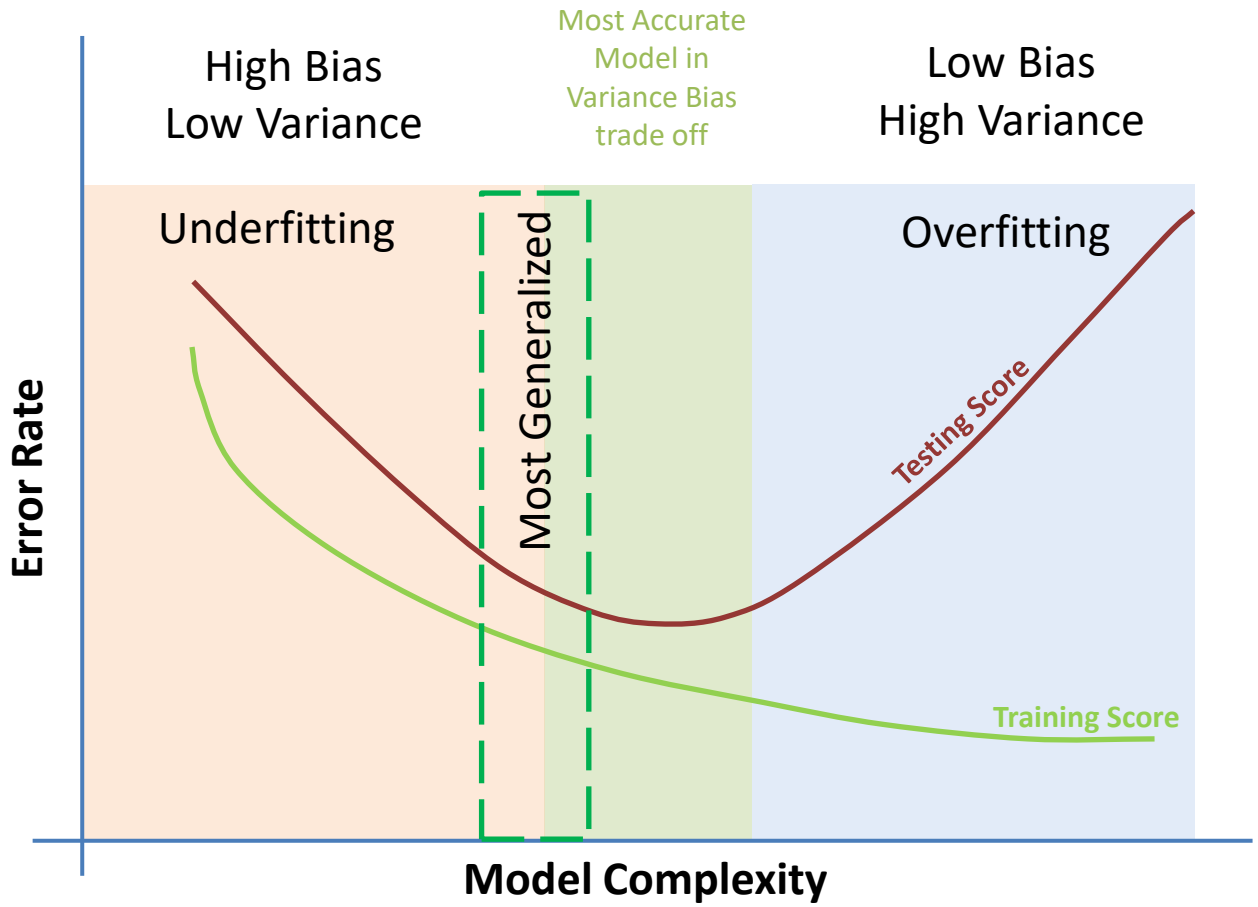


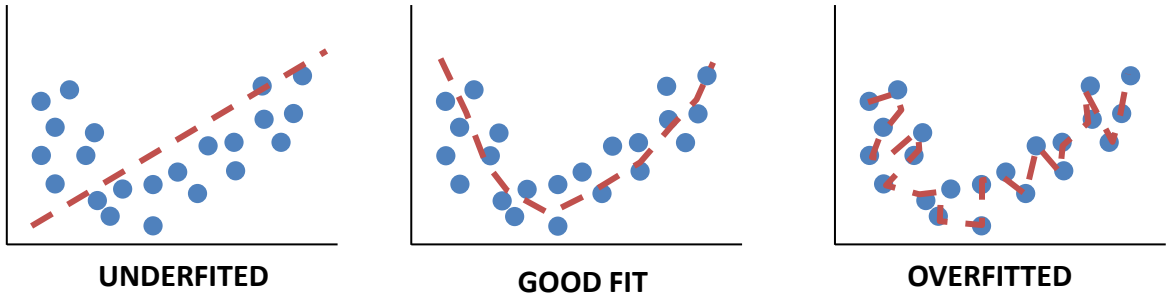
Model Evaluation - Overfitting





Model Evaluation - Good Fit





Model Evaluation – Classification Problems

Confusion Matrix :

- ***Precision:*** *true positive rate*

$$\frac{TP}{TP + FP}$$

- ***Recall:*** *true positive over the 1 class predict*

$$\frac{TP}{TP + FN}$$

- ***Accuracy:***

$$\frac{TP + TN}{TP + TN + FP + FN}$$

- ***F1 Score:***

$$\frac{2{*}precision\ {*}Recall}{Precision{+}Recall}$$

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP + FN)}$
Actual Class	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	**Negative Predictive Value** $\frac{TN}{(TN + FN)}$	**Accuracy** $\frac{TP + TN}{(TP + TN + FP + FN)}$

HOW DOES ML WORK?

Linear Regression

- ***Objective:***

  model the expected value of a continuous variable, ***Y***, as a linear function of the continuous predictor, ***X***

- ***Model structure:***

  $Y = Ax + B$

- ***Model assumptions:***

  Y is normally distributed, errors are normally distributed, and independent

- ***Parameter estimates and interpretation:***

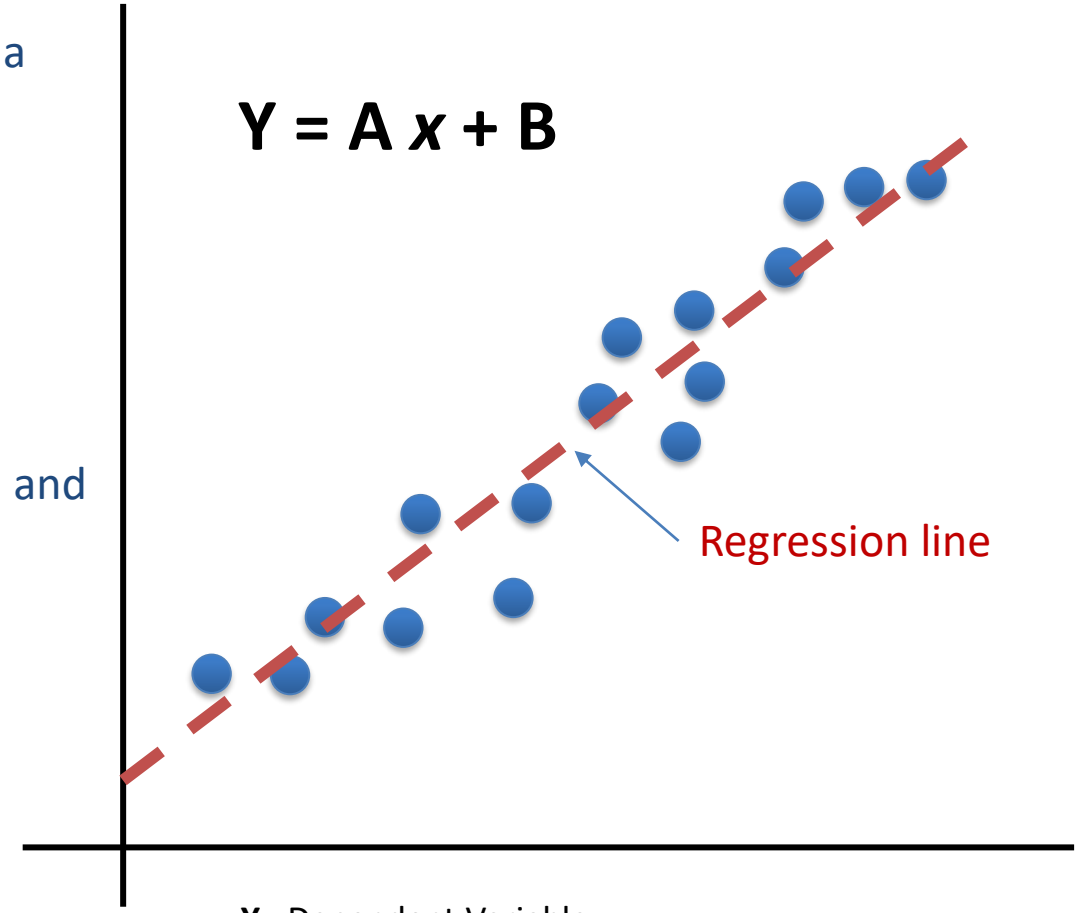
  B the intercept, and **A** is estimate of the slope

- ***Model fit:***

  R^2, residual analysis

- ***Model selection:***

  possible predictors, which variables to include?



Y : Dependent Variable
A : Slope
***x* :** Independent variable
B : Intercept

Linear Regression - Gradient Descent

- ***Objective:***

  To minimize the error function to close to zero (Cost Function) If possible.

- ***Function structure:***

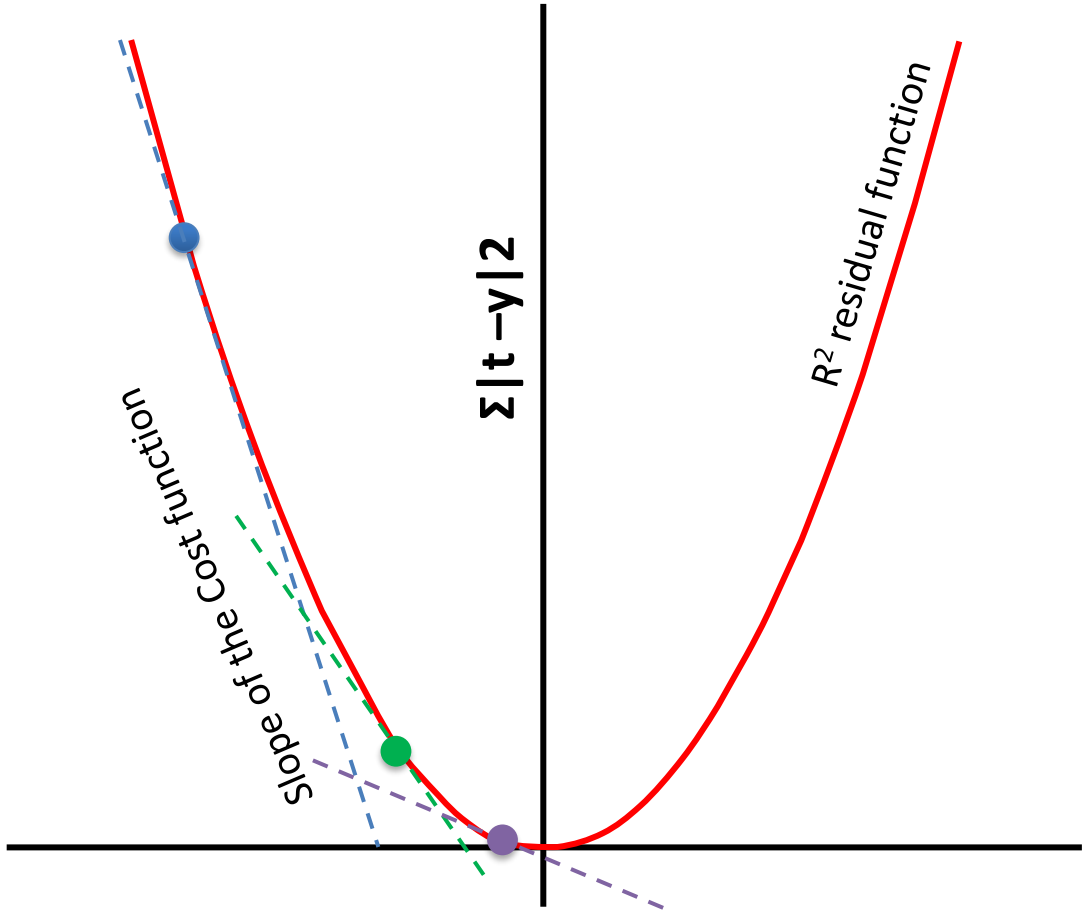
  Cost function : $\Sigma \mid t - y \mid^2$

- ***Model assumptions:***

  Slope of the *cost function* ~= Zero, then it is the best prediction

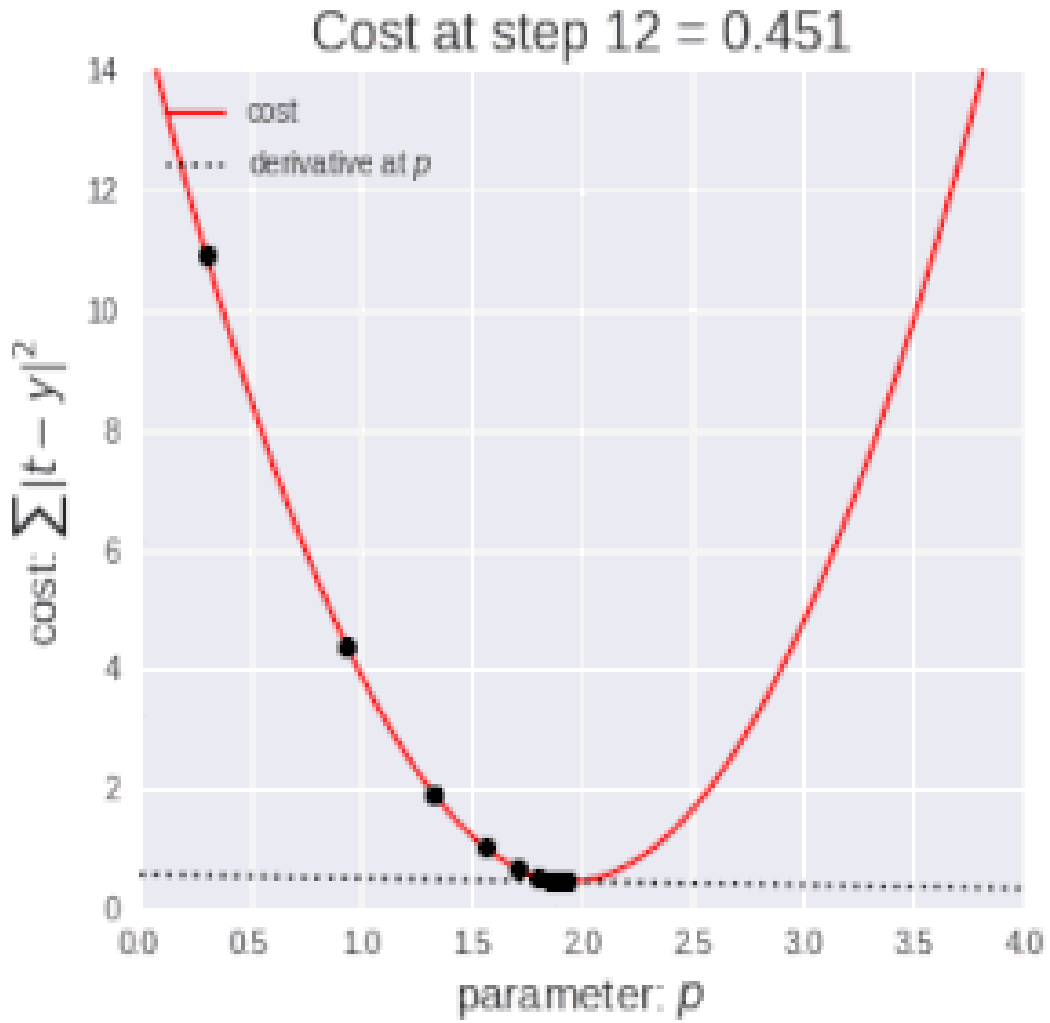
- ***Parameter estimates and interpretation:***

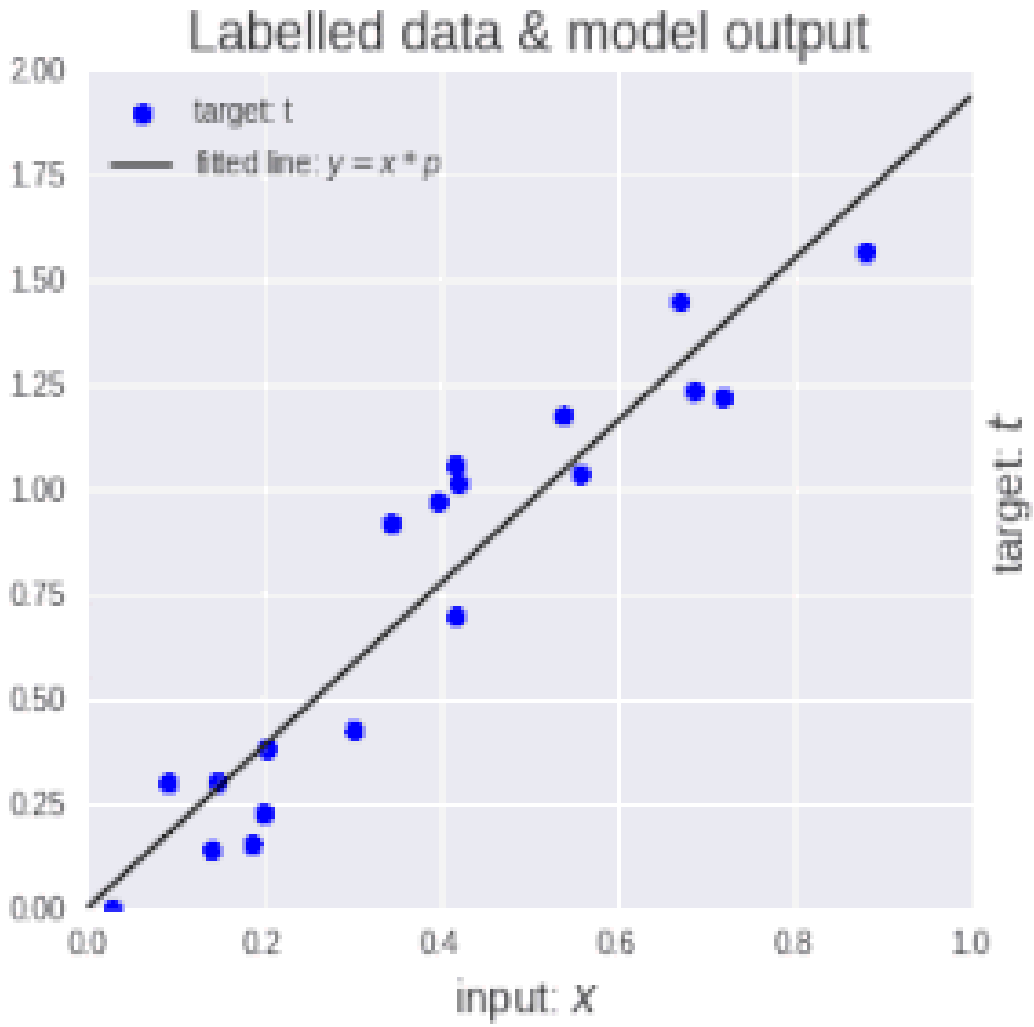
  - Slope first derivative over certain iterations,
  - Learning rate



Y : Cost Function (Loss function, Error)
A : Slope
x **:** N# of iterations

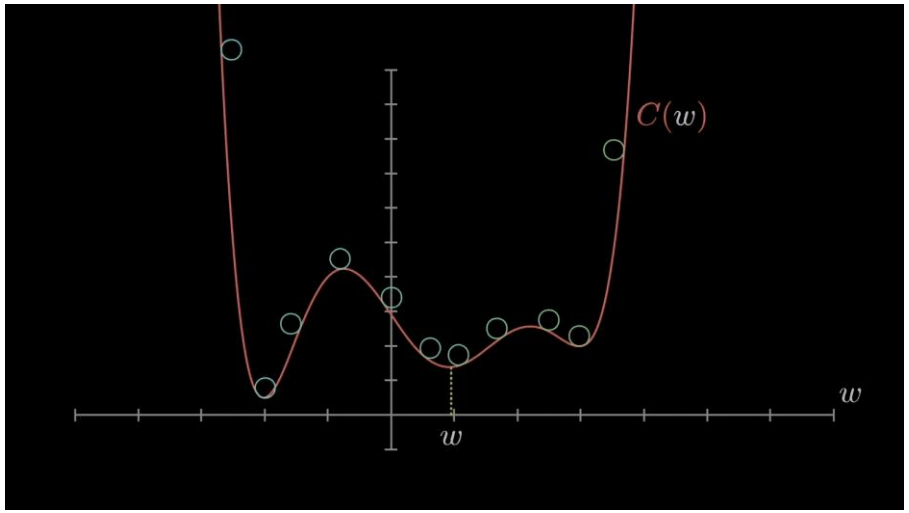
Linear Regression - Gradient Descent

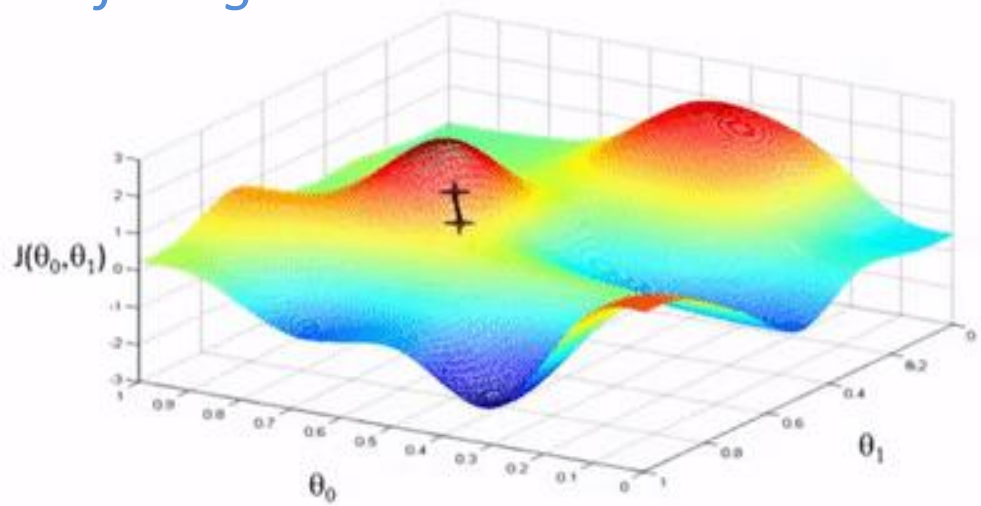




Gradient Descent

- *Gradient descent is based on calculus.*
- *Gradient descent is different from one algorithm to another based on the complexity of the algorithm and no# of variables (dimensions)*
- *It always has local minima.*
- *Learning rate is the essential step to reach a healthy GD*
- *Learning rate can be cause of overfitting or underfitting*





Machine Learning Applications in Oil and Gas Industry

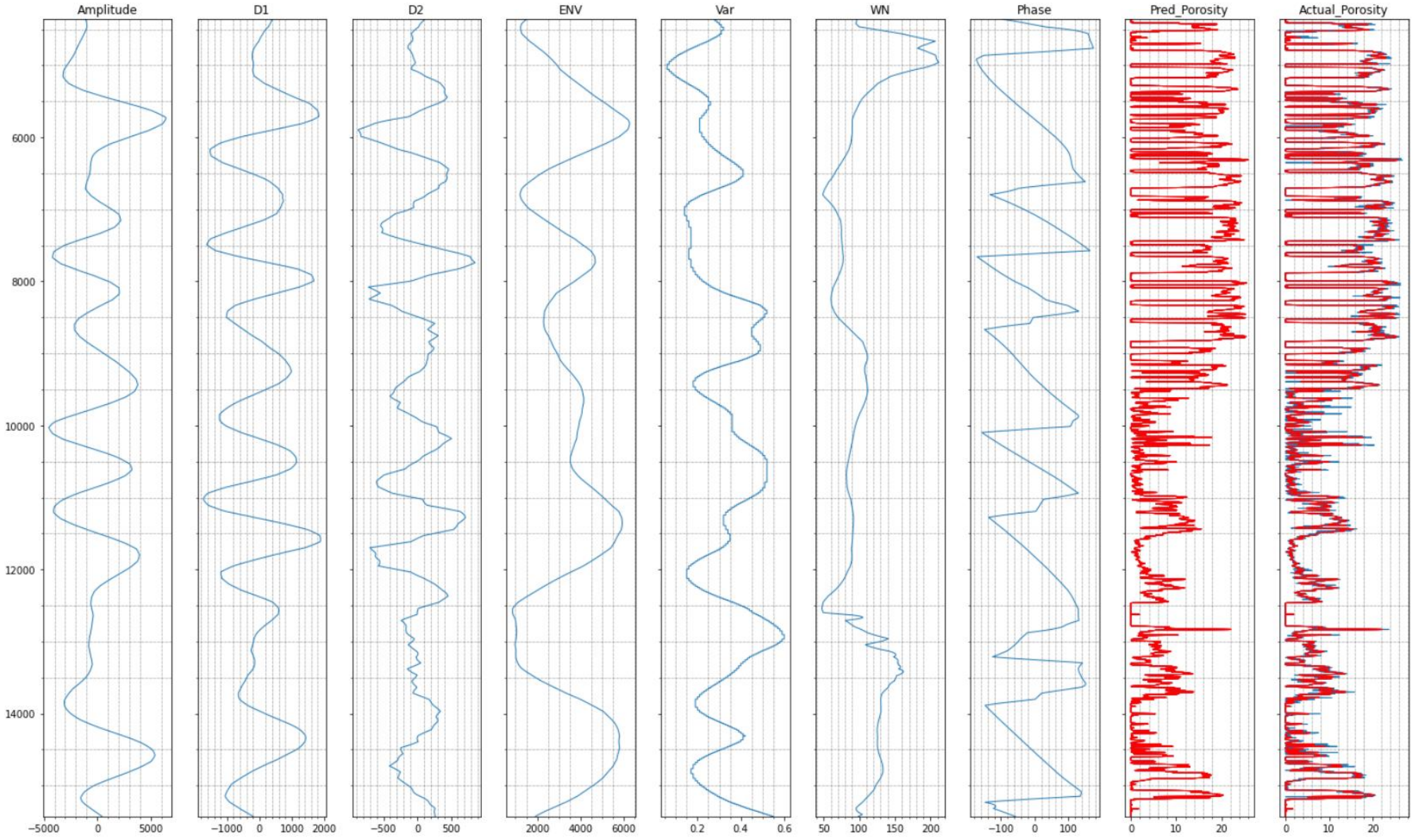
Amr.Moslim

Machine Learning Application Examples

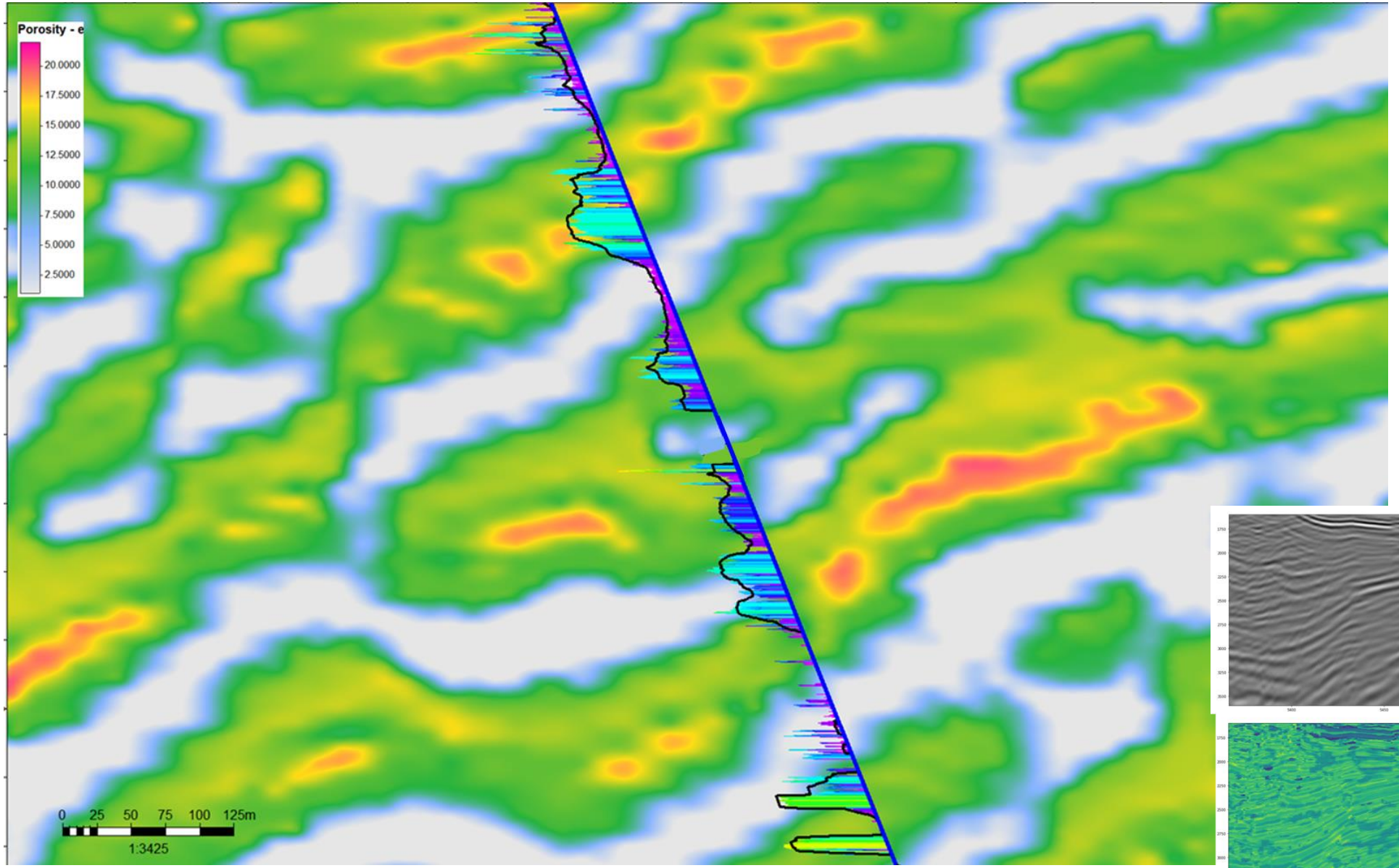
- Facies Classification using well log data
- Porosity Prediction using seismic attributes
- Permeability Prediction using Petrophysical volumes
- Facies Classification using seismic attributes
- Seismic Data inversion using Multi solver algorithms.

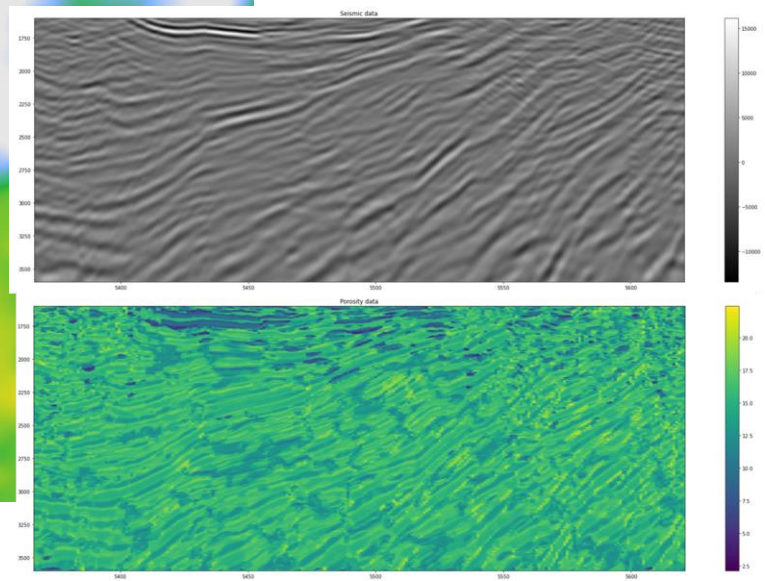
Using Seismic Attributes to predict Porosity logs . . .

Comparison between the Predicted and Actual

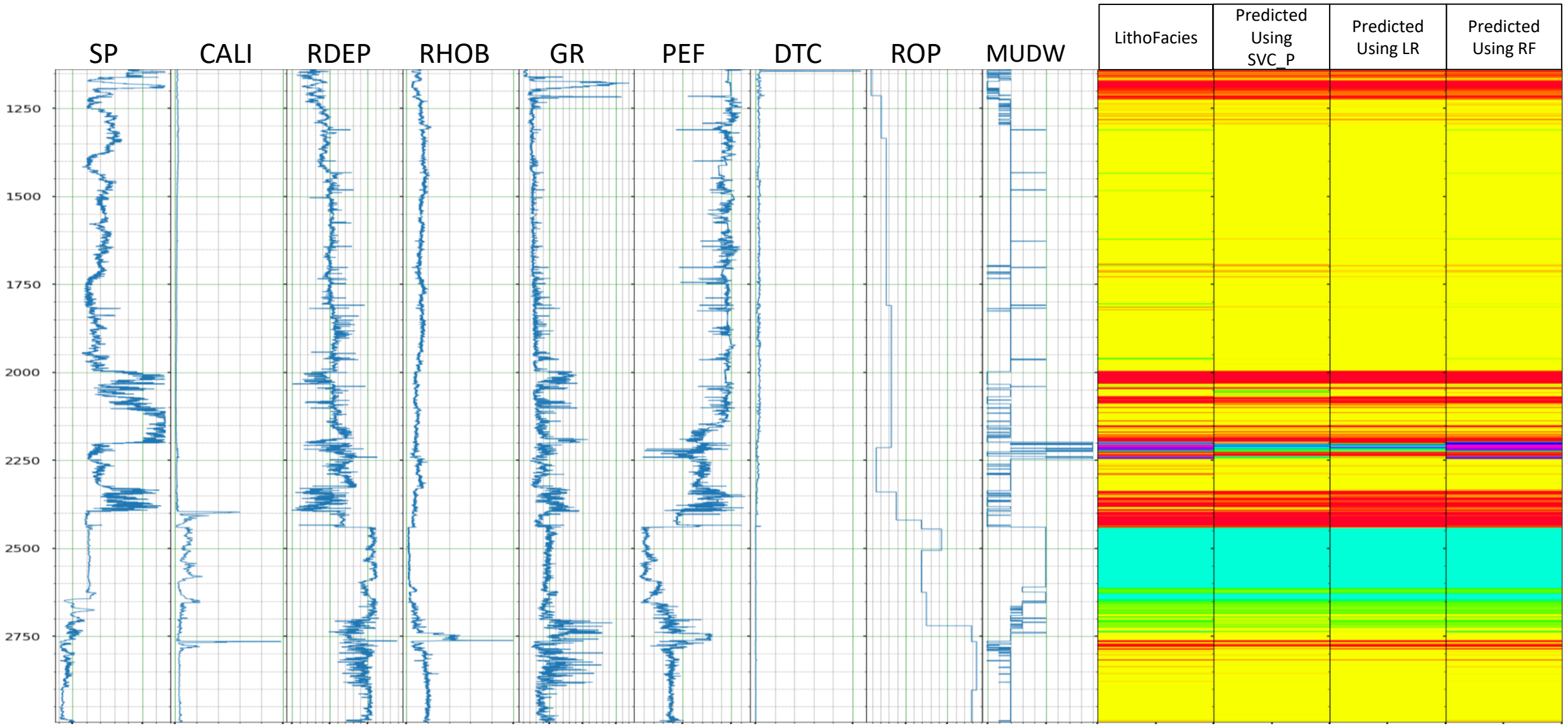


Porosity log vs 3D Predicted Porosity Cube

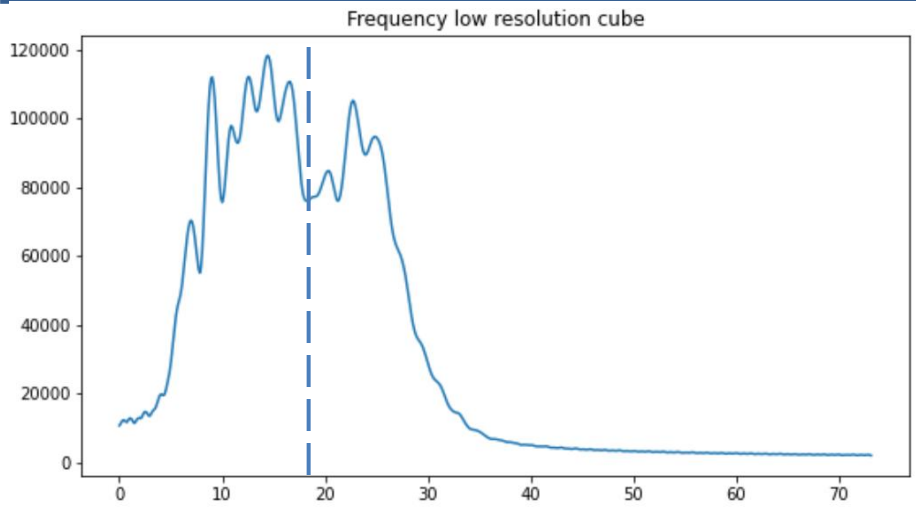


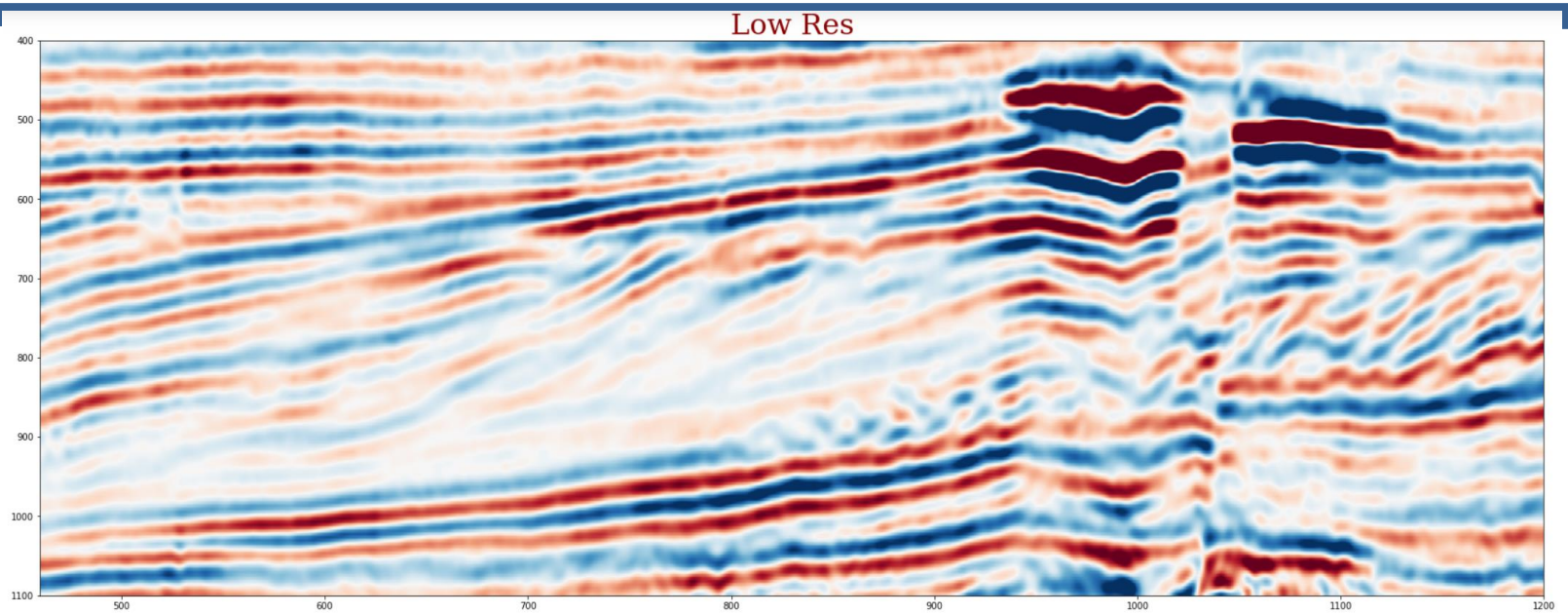


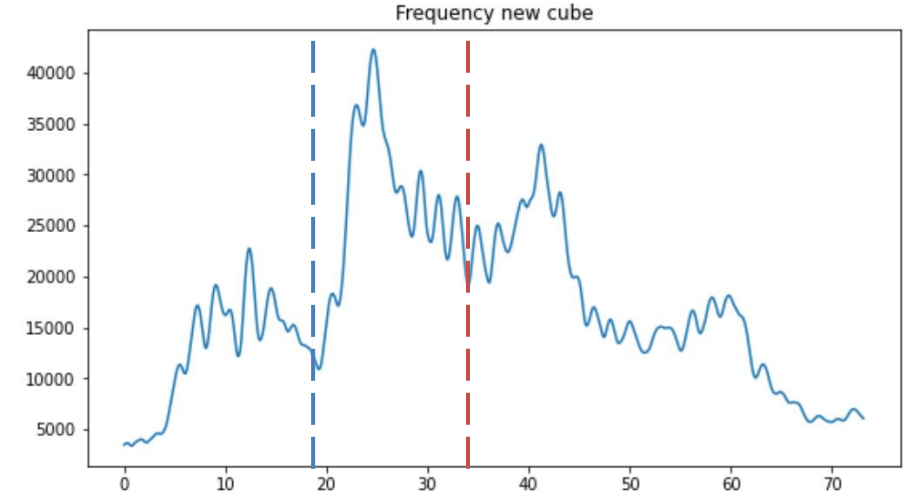
Facies Classification Using Well Logs . . .

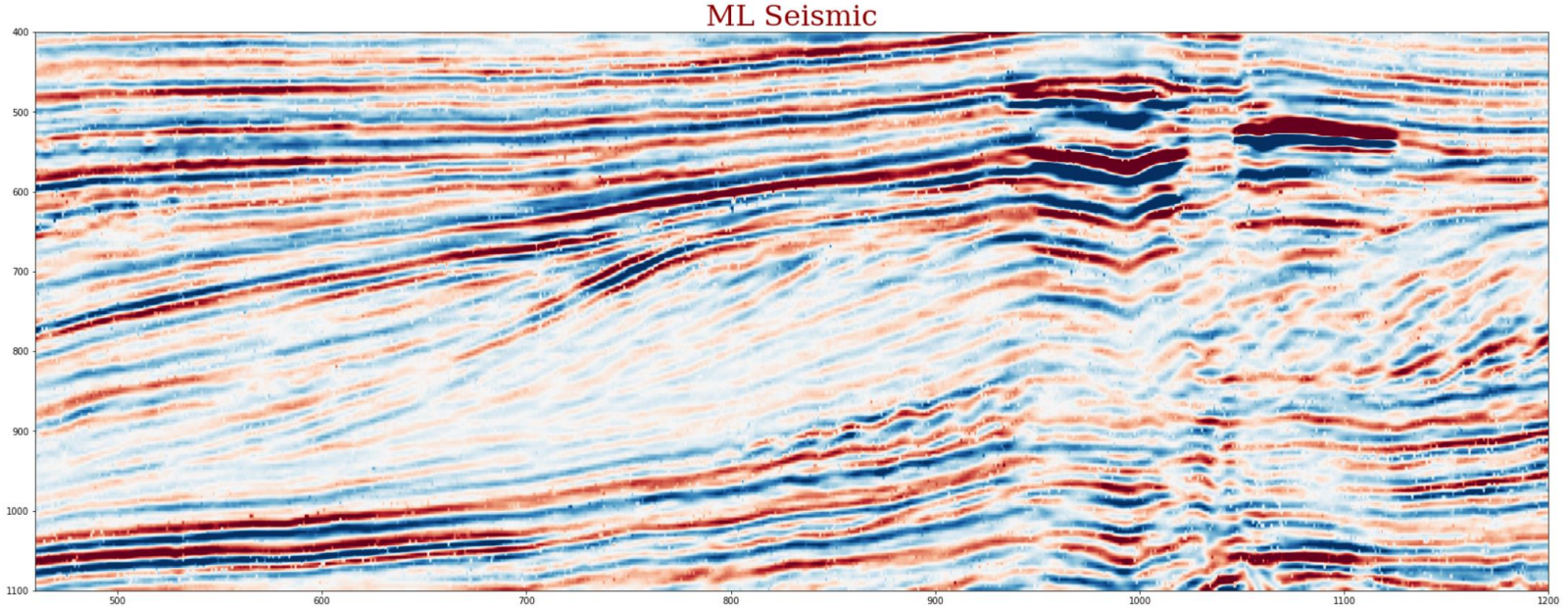


Seismic Data Enhanced Resolution









Amr.Moslim

The Road Map

Python: Numpy, Matplotlib, pandas
R: Tidyverse- tidymodels - ggplot

Phase 1 : Learn a Programming Language

(Python, R,SQL)

Python: Scipy, Obspy, statmodels, Plotly,
R: Tidyverse- tidymodels - ggplot

Phase 2 : Data Wrangling Techniques & Database

Probabilities, center measures, variation

Phase 3 : Mathematics

(Statistics, Linear Algebra, Calculus)

Phase 4 : Machine Learning Algorithms

Machine Learning packages

- Tensorflow : Neural NetWork and Deep learning
- Keras: ML algorithms
- Scikit Learn: ML algorithms and model evaluations

Geoscience Package

- Welly : reading / write well logs las files
- Lasio : reading / write well logs las files
- Segyio : seismic Segy files reading / writing and manipulation.
- Petropy : Petrophysical evaluation

Thank You for Your Attention

Amr Moslim