भारतीय भूवैज्ञानिक सर्वेक्षण

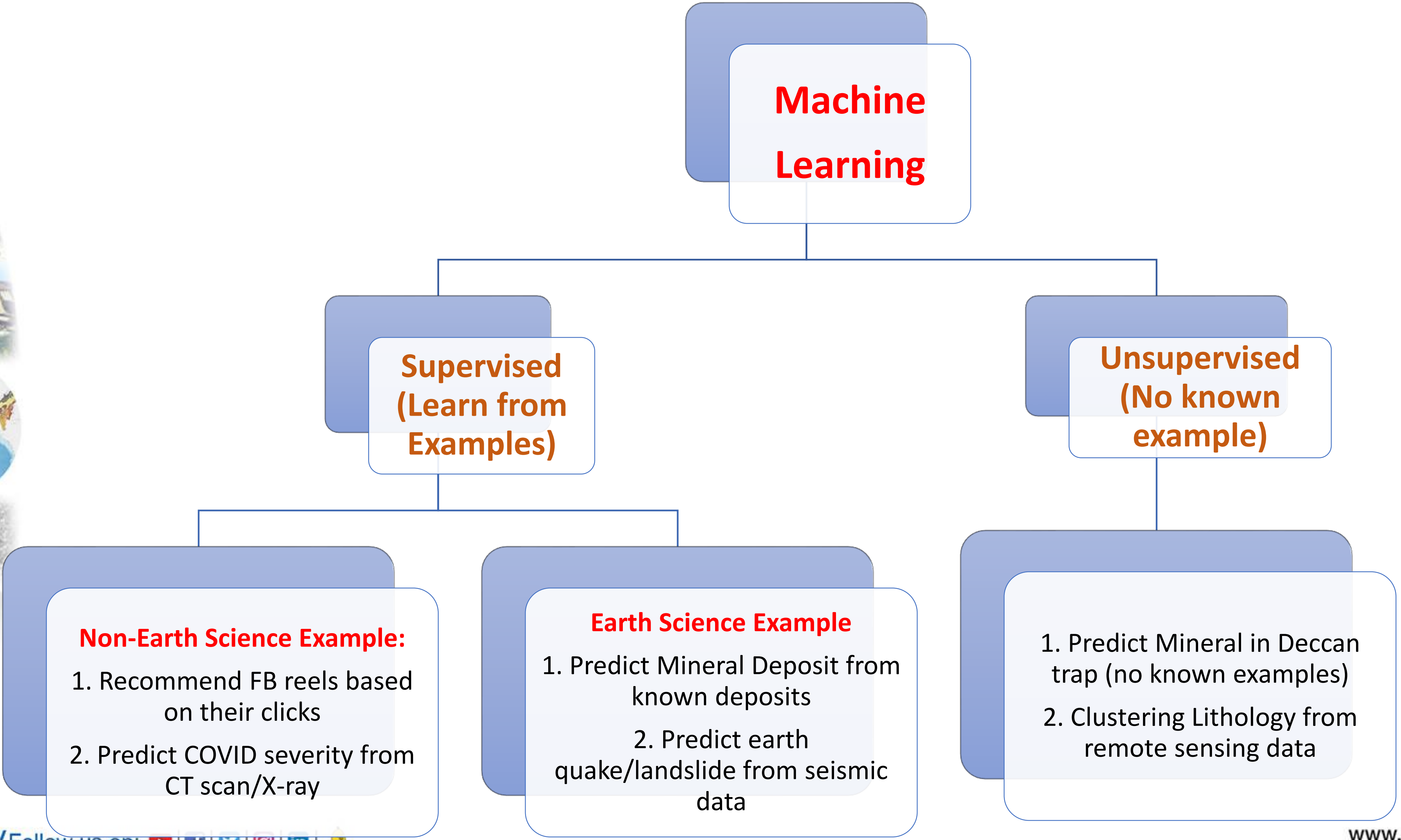
GEOLOGICAL SURVEY OF INDIA

Introduction to Machine Learning - II

Sabyasachi Nag

Sr. Geophysicist, NM1B-CHQ, GSI

Types of Machine Learning



Machine Learning

Supervised (Learn from Examples)

Unsupervised (No known example)

Non-Earth Science Example:

1. Recommend FB reels based on their clicks
2. Predict COVID severity from CT scan/X-ray

Earth Science Example

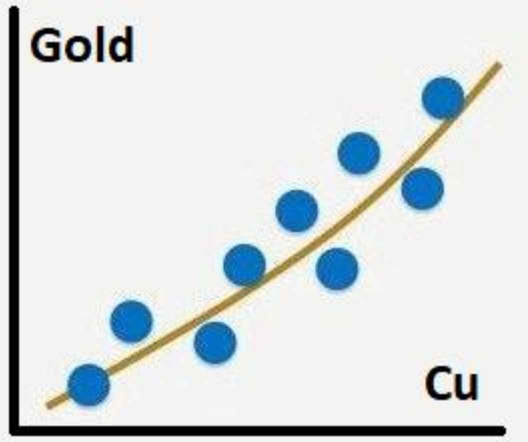
1. Predict Mineral Deposit from known deposits
2. Predict earth quake/landslide from seismic data

1. Predict Mineral in Deccan trap (no known examples)
2. Clustering Lithology from remote sensing data

Supervised Regression vs Supervised Classification

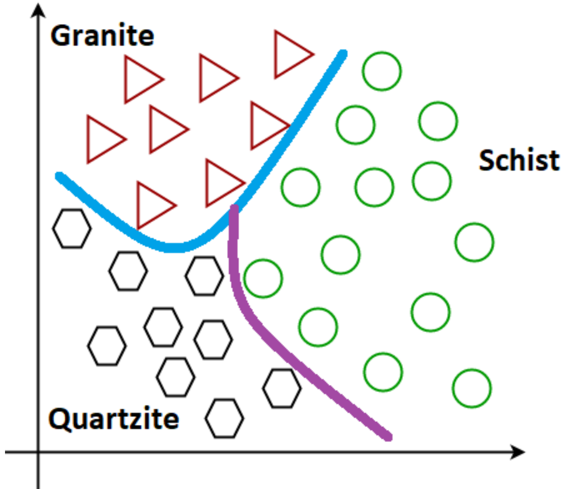
- **Regression**
  - **Output is a numerical variable**
  - Input may be also be numbers/ categories (e.g. name of rock) or combination of both

  - Input(Features):
    1. Cu concentration
    2. Gravity Anomaly
    3. Rock types
  - Output(Target)
    - Calculate Gold concentration



- **Classification**
  - **Output is a categorical variable**
  - Input may be numbers/ categories or combination

  - Input(Features):
    1. Remote sensing pixel intensity
    2. Radiometric U map
    3. Magnetic anomaly
  - Output(Target)
    - Identify Lithology



Use Cases

Non Earth Science

1. Predict Age of a Person from photo - R
2. Predict Gender of a person from photo - C

3. Agriculture produce forecast from rainfall/weather data - R
4. Salary calculator from Graduation marks and Work Experience R
5. Spam email detection from mail text C
6. Predicting sales of AC from weekly temperature and humidity data R
7. Auto-tagging your friend on Facebook C
8. Fraud detection in banking transactions C
9. Recommending green peas, paneer masala if a person adds Paneer to cart C
10. Digitize old hard copy book to soft copy C

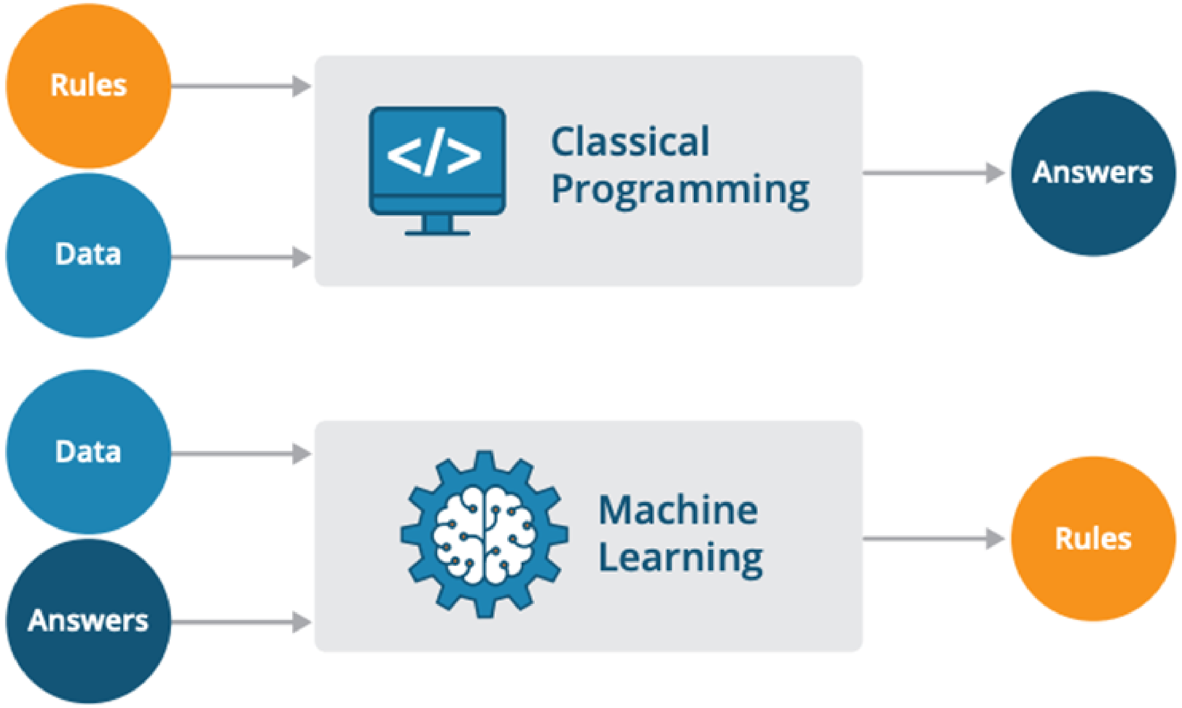
Earth Science

1. Identify fault/fold from geophysical map/seismic profile - C
2. Predict Zinc concentration from known Pb concentration - R

3. Automatic lithology identification from pictures of litho-logs C
4. Estimate earthquake intensity from seismograph data R
5. Mineral type identification from thin sections C
6. Landslide warning system from continuous GPS station data
7. . Flood extent - R
8. . Groundwater contamination spread R
9. . Cyclone path prediction R
10. .

But Why do Machine Learning in Earth Science?

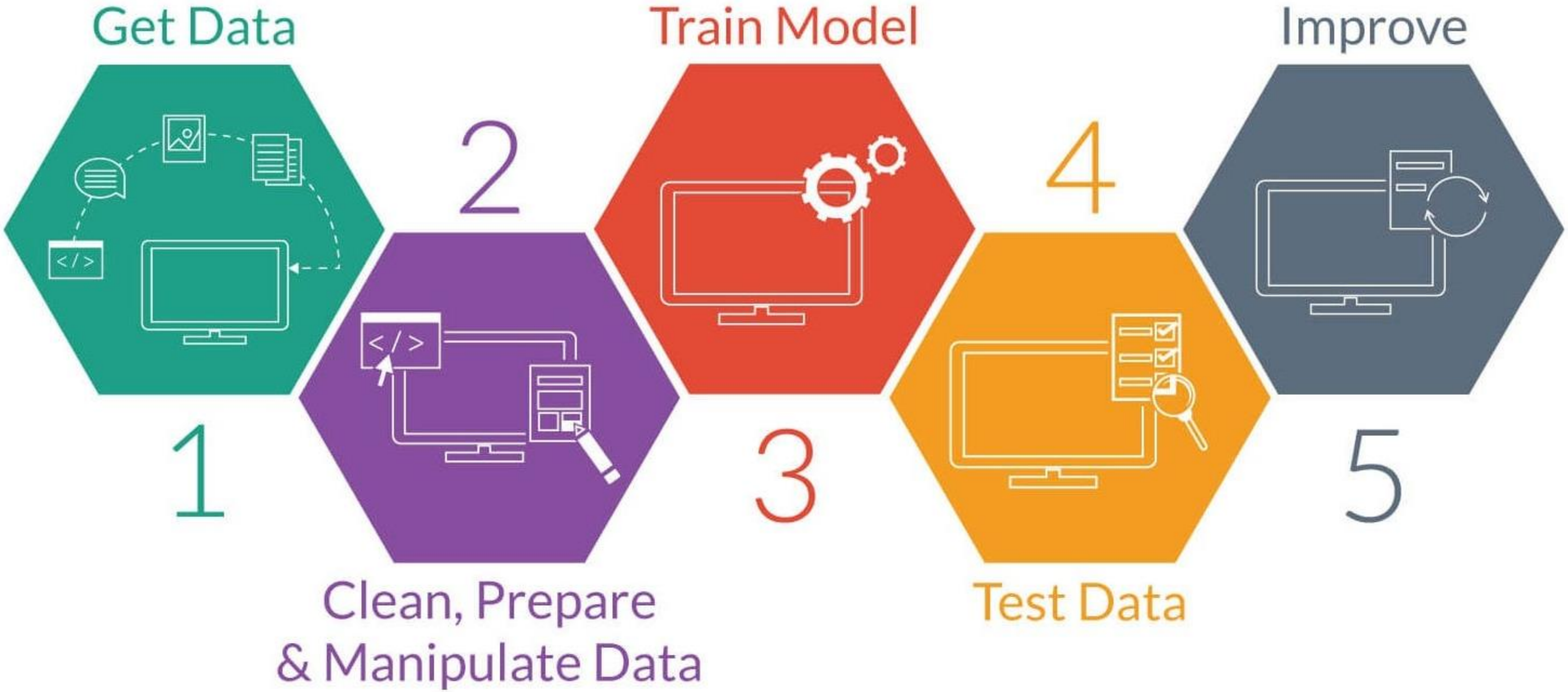
- Geological Processes are inherently complex and often simple rules do not work
- Geological/Geophysical data are huge with multiple layers
  - Nearly impossible for humans to detect patterns/ discover rules/correlation among them



- **Conventional approach**
  - We(humans) ‘think’ that some rules work
  - Apply the rules on data to predict unknowns
  - Our rules may be flawed (limited success/ high failure rate/ resource wastage in exploration)

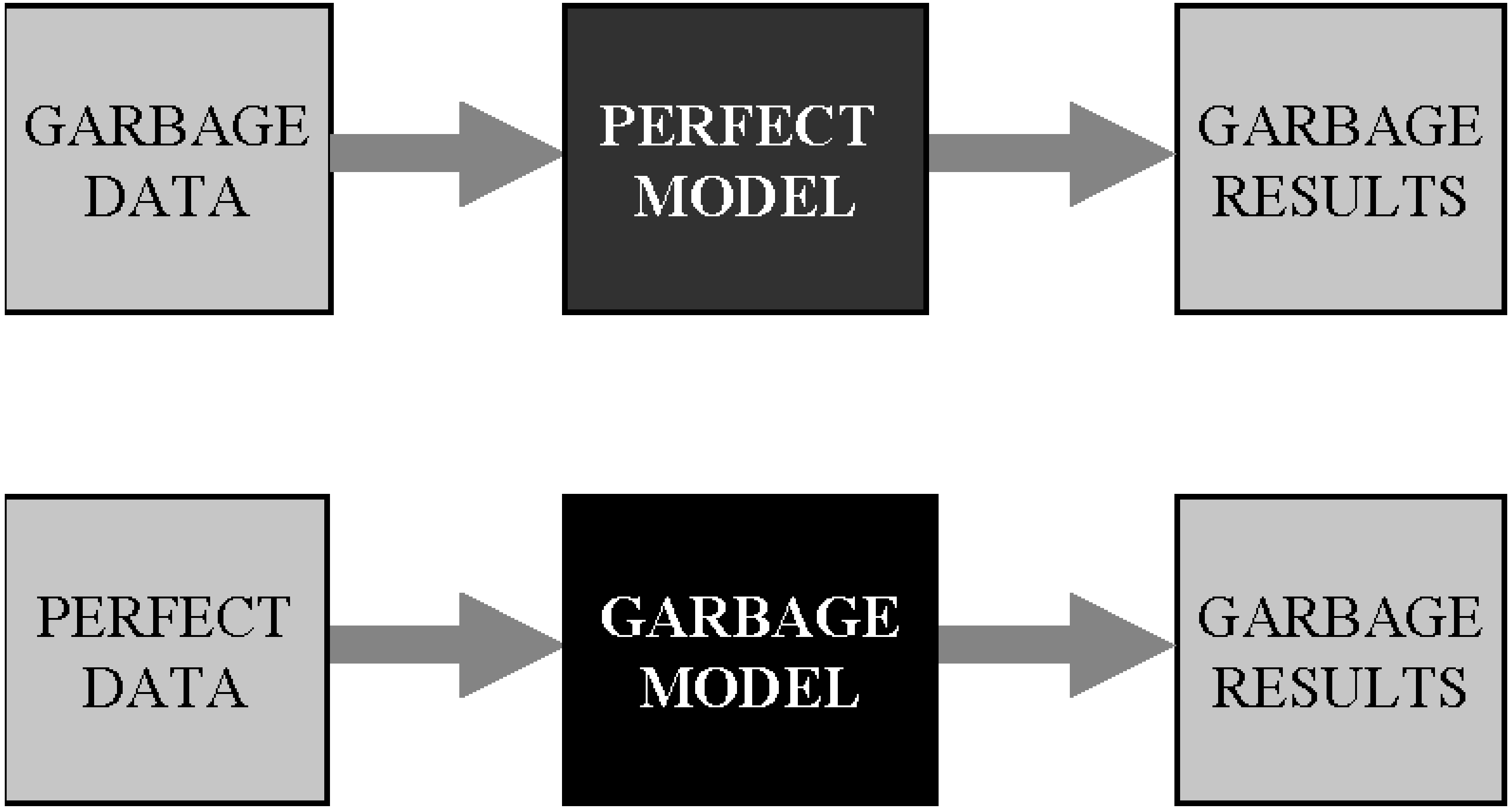
- **Machine Learning approach**
  - We do not set any prior rules
  - Data is supplied to machine with known examples(targets)
  - Machine determines the rules that inter-connects the data and the known examples
  - Machine apply the rule to predict unknown/new data
  - Rules understood by machine is in mathematical, may not be easily readable by humans but they can be applied to find new results which is often all that matters

Steps of Implementing Machine Learning Algorithm



Data Cleaning, Feature Engineering

- Data must be expressed in manner that makes sense for the problem
  - Bad data must be removed
  - Data may need to be transformed/rescaled
- Model should be constructed in fashion that can generalize the overall picture



GARBAGE DATA → PERFECT MODEL → GARBAGE RESULTS

PERFECT DATA → GARBAGE MODEL → GARBAGE RESULTS

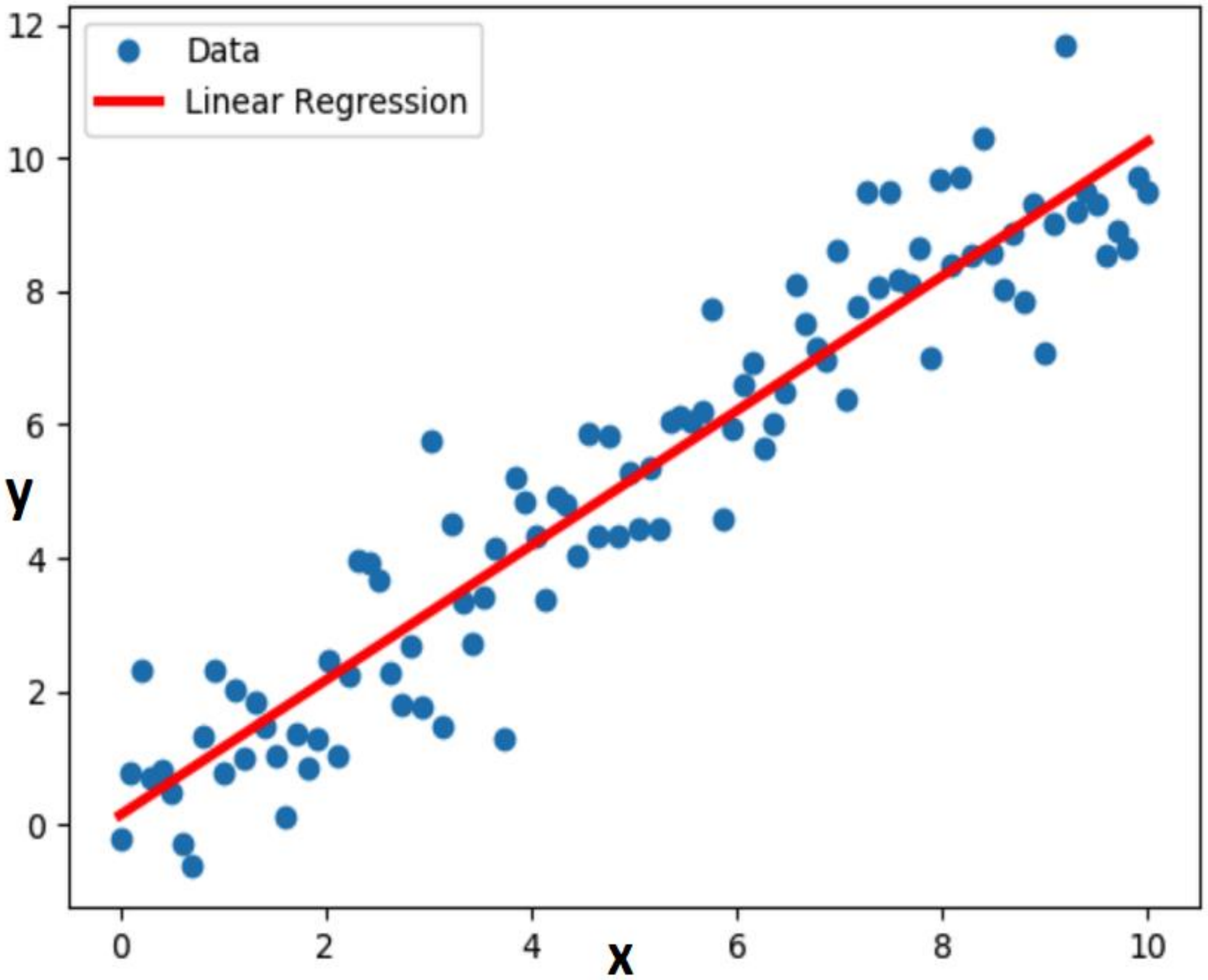
Lists of Common Machine Learning Models

1. Linear Regression
2. K-Nearest Neighbour(KNN) Regression/classification
3. Decision Tree Classification/Regression
4. Random Forest Classification/Regression

5. Support Vector Machines (SVM)
6. Naïve Bayes

7. Artificial Neural Network (ANN)
8. Convolutional Neural Network(CNN)

Linear Regression



- Suppose the data points are represented by y_true

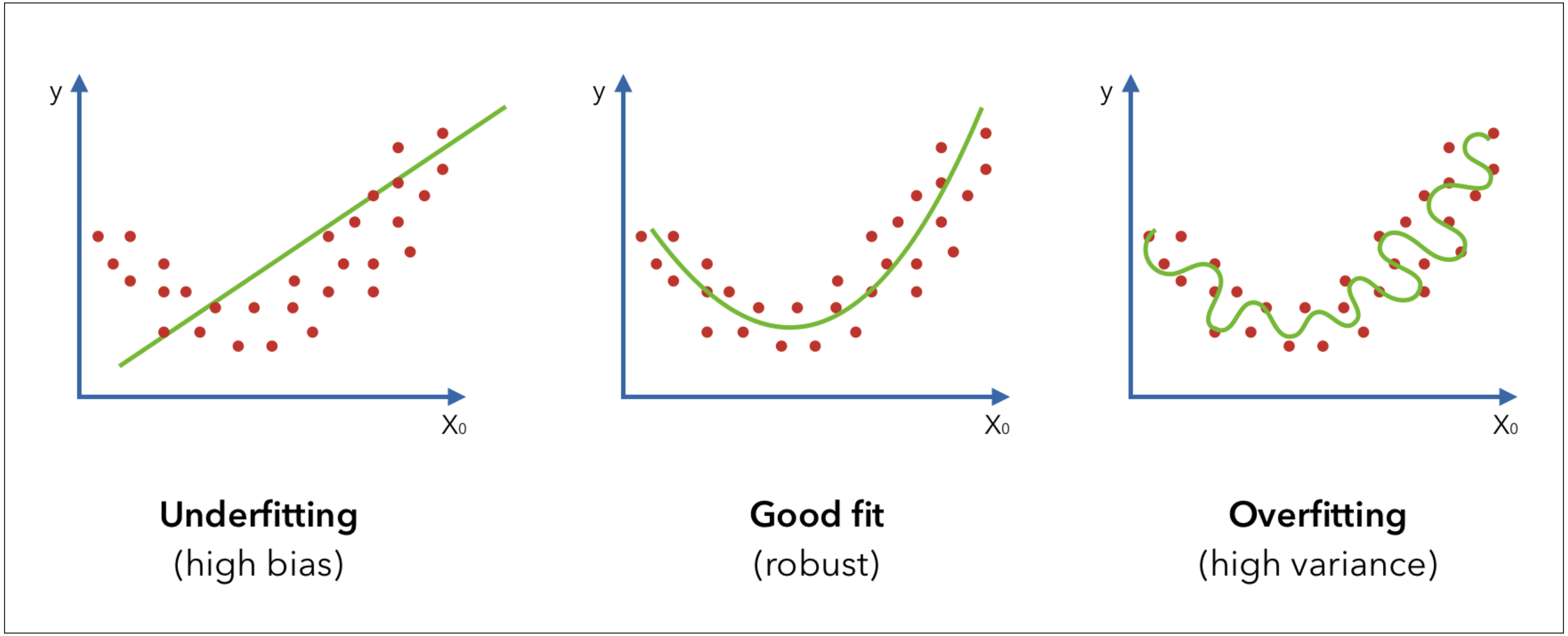
- Assume Linear regression model = y
- We have to minimise the difference of y and y_true (for every point)
- Assume that y_true depends on x linearly
  - y_true = a + b*x
  - a = intercept
  - b = slope

- Through ML, try to reduce the error by summing over all points
  - e = (1/N_pts) sum (y – y_true)^2

- The red line on left represents the final model that best fits the data

Model Selection

- A perfect fit may not always be the best solution
- Model should be able to capture the general trend

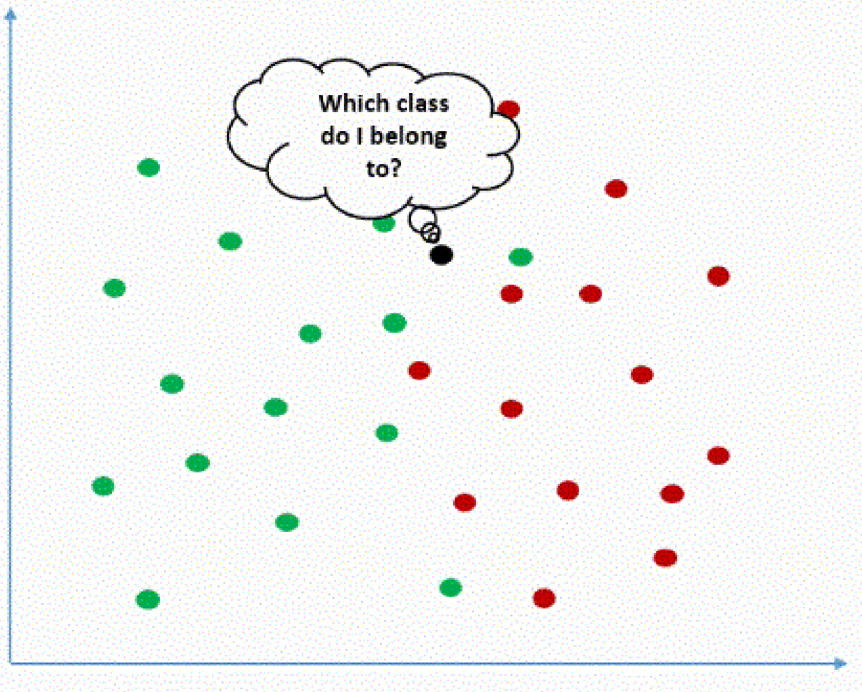


Underfitting
(high bias)

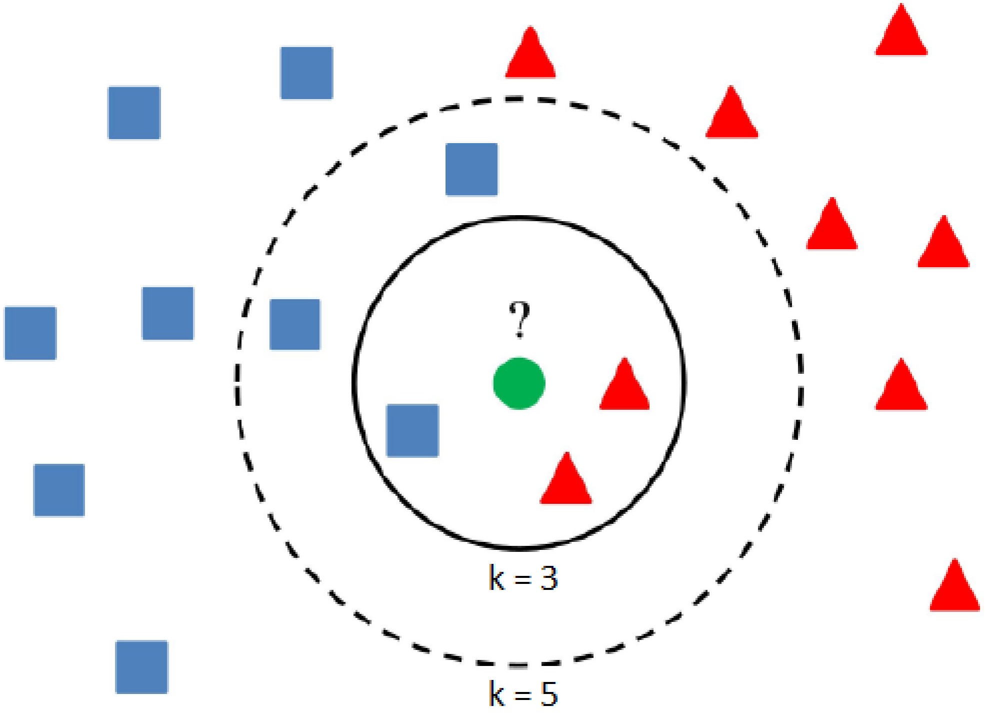
Good fit
(robust)

Overfitting
(high variance)

K-Nearest Neighbour



- Assume some K (say 5)
- Calculate distance of target point to every other point
- From least 5 distances, select K=5 nearest point from the target
- Check to which category most of the neighbours belong
- Assign the target point to that category



Why Python

1. Great set of built-in ML libraries (no need to rewrite/ ‘reinvent the wheel’)
    - Scikit-Learn
    - Tensorflow
    - Pytorch
2. Built-in libraries for handling big (earth science) data
    - Numpy
    - Scipy
    - Pandas
    - Geopandas
3. Great community support
    - Only remember basic few syntaxes/commands
    - Look up internet for almost everything (most people have encountered same problem and solution is already there. Modify according to your own problem
        - Stack exchange/stack overflow
        - Geek for geeks
        - Towards data science
4. Free, open source and cross platform supports(Compatible with Linux)
    - Only internet is needed (which is not free!)

Thank You