AI/ML for Environmental Data Analytics

DS 392

Deepak Subramani
Assistant Professor
Dept. of Computational and Data Science
Indian Institute of Science Bengaluru

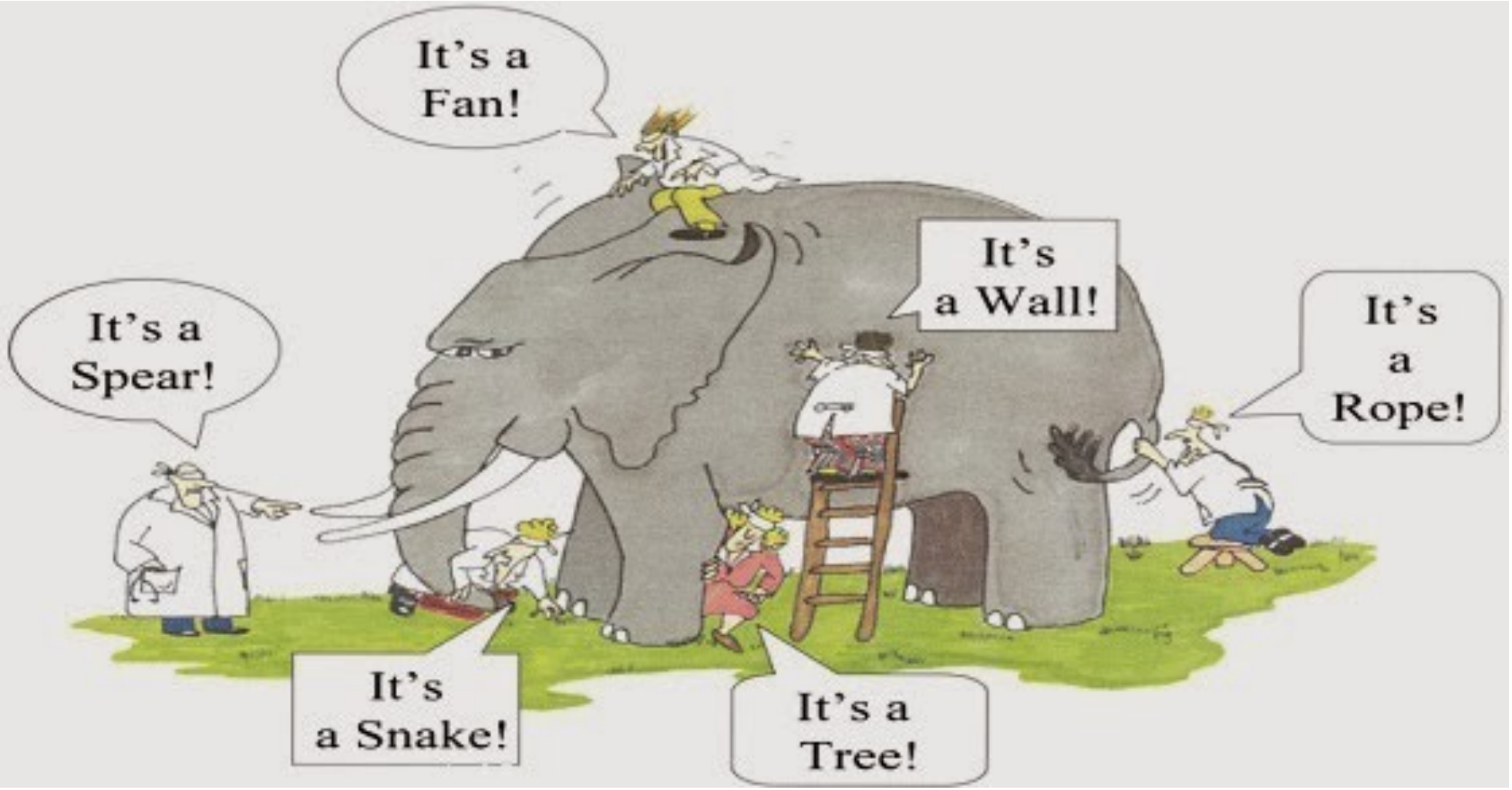
Department of Computational and Data Sciences

Summary of Lecture 01

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

Deepak Subramani, deepakns@iisc.ac.in

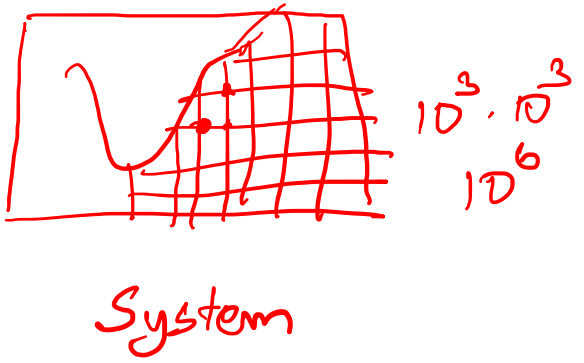
The Proverbial Elephant



Deepak Subramani, deepakns@iisc.ac.in

Geosciences: Data and Models

- The science of Earth – Geology, Meteorology, Oceanography, Astronomy
- Characteristics: Large spatial and temporal scales
- Let us take Oceanography
  - 7 equations – The Primitive Equations (momentum, mass, energy, height, state)
  - Bay of Bengal – 2 million sq. km. Avg Depth : 2 km
  - Data on Primitive Variables
    - If 1 km grid and 100 vertical levels $\sim 10^9$ at a time
    - 365 days a year – 4 times a day – Add 3 orders of magnitude
  - Really Big Data is needed to study
- Is data available?
  - Satellites – temporally sparse, medium spatial coverage
  - In-situ – spatially sparse, temporally dense
- Numerical models – Big Compute

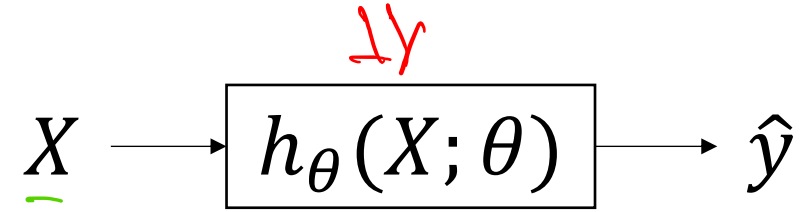


ML: Mental Model

Data that can be collected

Quantity that must be predicted to make money

Data that can be collected



Machine's Prediction

h_θ : Machine

Model Selection
L.R ; NN

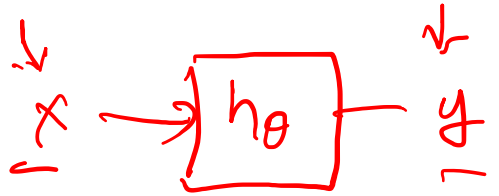
$\hat{y} = \theta_0 + \theta_1 x$

① What is h_θ ?

② What is θ given some h_θ, X, y ?

The Machine Learning Workflow

1. Frame the ML problem by looking at the science need
   a. Identify subproblems →
2. Gather the data and do Data Munging/Wrangling
   a. Explore the data EDA xarray pandas
   b. Clean data and prepare for the downstream ML models
3. Explore different models, perform V&V and shortlist promising candidates h_θ → train, validate, test
4. Fine-tune shortlisted models, draw insights, and combine them together to form the final solution Ensemble
5. Present your solution
   a. Say a story with the data
6. Deploy and monitor: Write a Paper; Publish Code etc



Deepak Subramani, deepakns@iisc.ac.in

ML Workflow: Tech Stack

1. Frame the ML problem by looking at the science need
2. Gather the data and do Data Munging/Wrangling for each subproblem
   a. **Xarray, Pandas, Numpy, Seaborn, Matplotlib, pyferret**
   b. **sklearn.preprocessing (scaler, OneHotEncoder), sklearn.impute (data cleaning, drop nan etc), custom transformers**
3. Explore different models, perform V&V and shortlist promising candidates
   a. **sklearn.pipeline, sklearn.model_selection, sklearn.xxx (where xxx is a model), XGBoost, TF2, Keras**
4. Fine-tune shortlisted models and combine them together to form the final solution
   a. **sklearn.ensemble.VotingClassifier etc,**
5. Present your solution
   a. **PowerPoint, Seaborn, matplotlib, plotly, dash, javascript (fusion charts, react, d3), overleaf**
6. Deploy and monitor
   a. **As a journal paper with code, A user service on cloud (either local or cloud)** APT

Audience Poll

1. What are the supervised learning tasks

   (i) Clustering, (ii) Regression, (iii) Anomaly Detection, (iv) Density Estimation

2. What are the unsupervised learning tasks

   (i) Classification, (ii) Regression, (iii) SVM, (iv) Clustering

3. What is the tool used for validation while exploring models

   (i) sklearn.linear_model, (ii) sklearn.model_selection, (iii) sklearn.ensemble, (iv) pandas

   cv

4. How many overall steps did we discuss was in the ML Workflow?

   (i) 1, (ii) 3, (iii) 6, (iv) 7

Human-Elephant Conflict

Science News

Asian elephants may lose up to 42 percent of habitats in India and Nepal

Date: February 28, 2019

Source: Forschungsverbund Berlin

Summary: Protecting and expanding suitable habitats for wildlife is key to the ... dangered species, but owing to climate and land use chan... day may not be fitting in 30 or 50 years. An interna... predicted range shifts of Asian elephan... models based on distributio...

Home / Bangladesh / Nation

Farmer trampled to death by elephant in Chittagong

Pimple Barua, Chittagong

Published at 09:49 am November 14th, 2021

WILDLIFE & BIODIVERSITY

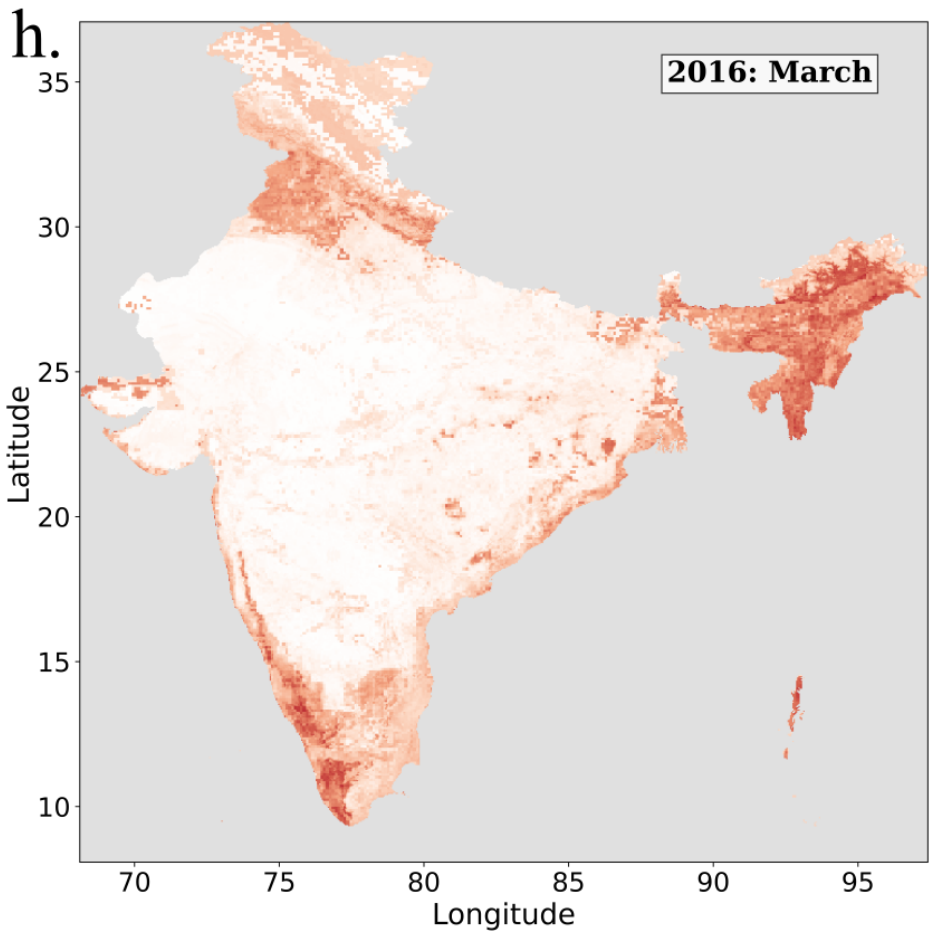
On same day, 3 elephants die in 2 Odisha districts

The development has raised concerns among forest officials and animal conservationists

By Ashis Senapati
Published: Monday 25 October 2021

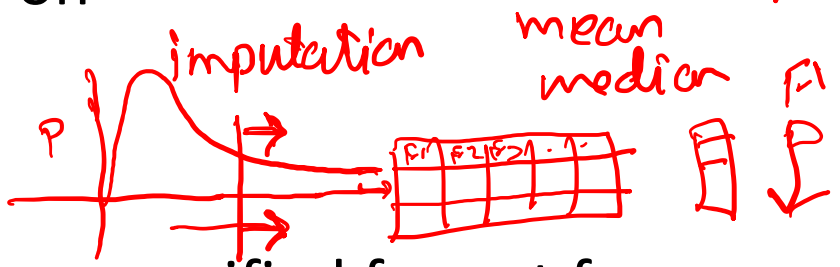
Data Science Problem: Predict Habitat Suitability

Machine Learning for Asian Elephant Habitat Suitability Estimation in India



Data Munging/Wrangling

1. Data Integration → Gather from various data sources and combine them at one place
2. Data Cleaning → In this step the primary focus is on
   1. Handling missing data
   2. Handling noisy data
   3. Detection and removal of outliers



3. Data Transformation → Convert the raw data into a specified format for feeding to downstream ML models.
   1. Normalization
   2. Aggregation
   3. Standardization

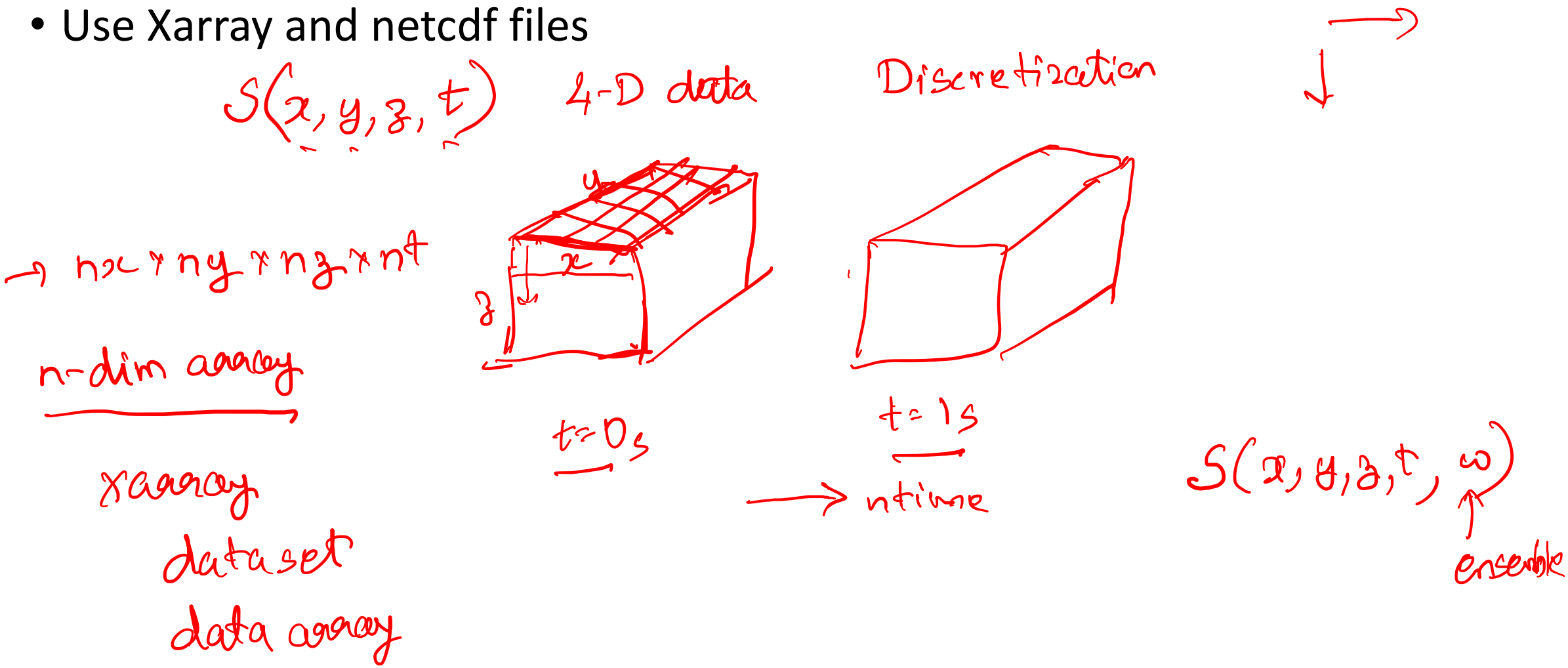
Scaling
0 - 1
Min Max Scaler
$z = \frac{x - \mu}{\sigma}$
Standard Scaler
X, y

4. Data Reduction → Following data transformation and scaling, the redundancy within the data is removed and is organized efficiently.

P.C.A S.V.D

Data Munging

- Use Xarray and netcdf files



ML Workflow Step 2: Data Munging

- Needs practice
- There are some guidelines – which are abstract ideas
- In all labs, spend time to do the data munging

Predictors and Targets

Category	Variable	Source	Unit	Spatial resolution
Climatic	Monthly precipitation	https://www.worldclim.org	mm	2.5 minute
	Monthly minimum temperature	https://www.worldclim.org	°C	2.5 minute
	Monthly maximum temperature	https://www.worldclim.org	°C	2.5 minute
Topographic	Elevation above sea level	https://www.worldclim.org	m	30 arc-seconds
	Distance to rivers and water-bodies	Derived using QGIS with data downloaded from https://www.openstreetmap.org	m	
	Distance to roads	Derived using QGIS with data downloaded from https://www.openstreetmap.org	m	
	Land Use Land Cover (LULC)	https://bhuvan.nrsc.gov.in	categorical	30m
Vegetation related	Net Primary Productivity (NPP)	https://neo.sci.gsfc.nasa.gov	gC/m²/day	0.1 degrees
	Leaf Area Index (LAI)	https://neo.sci.gsfc.nasa.gov	m²/m²	0.1 degrees
	Normalized Difference Vegetation Index (NDVI)	https://neo.sci.gsfc.nasa.gov	Dimensionless	0.1 degrees

Input to the ML

X

Target of the ML

y

GBIF
Global Biodiversity Information Facility

Asian Elephant Presence Data

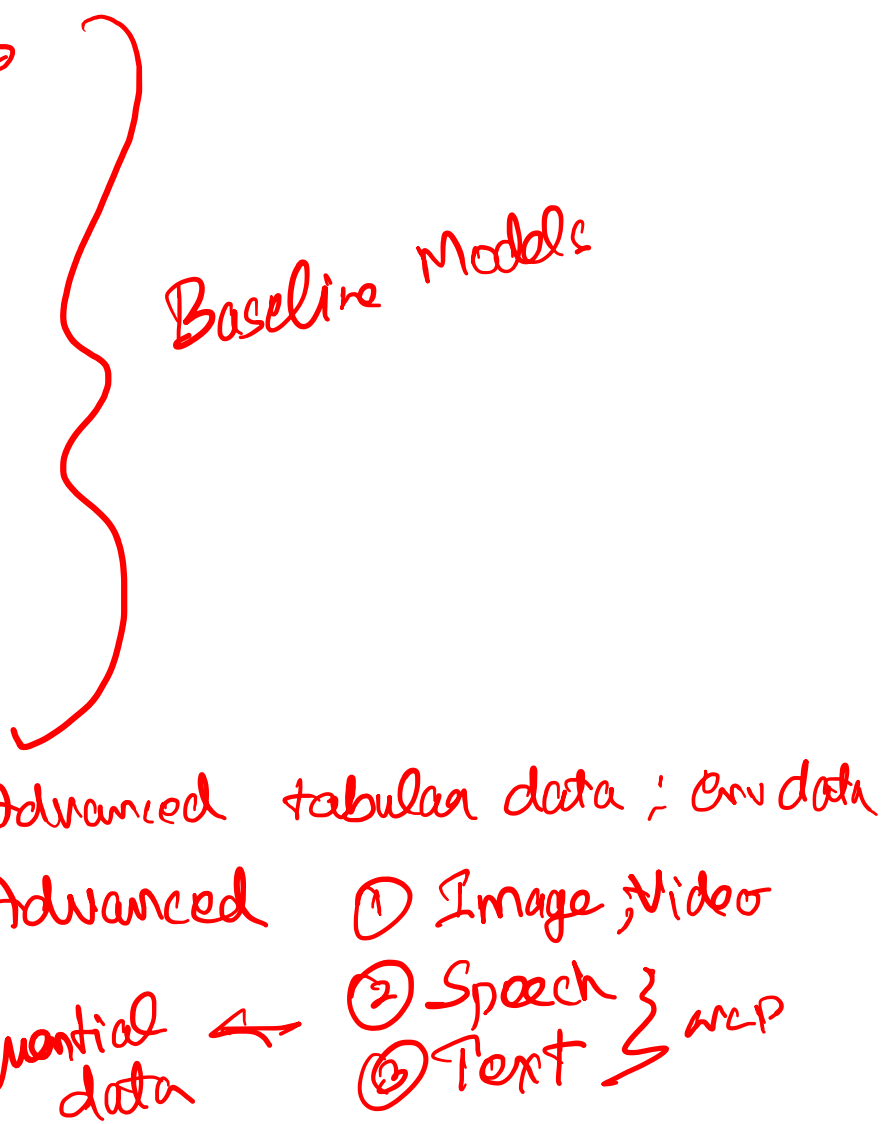
Paper: arXiv:2107.10478
2nd Place Award in InGARSS 2021

ML Problem – The view

θ

Step 3: Explore different ML Models

1. Linear Regression with Regularization
   - For regression task – predicting a continuous variable
2. Logistic Regression
   - For classification tasks alone
3. Support Vector Machines – Linear and Kernel
4. K-Nearest Neighbours
   - Instance-Based Method
5. Naïve Bayes
   - Simplest Bayesian Network Model
6. Decision Tree, Random Forests, XGBoost
7. Neural Networks



Deepak Subramani, deepakns@iisc.ac.in