Lecture 6

# End-to-End ML Project

## From Data Collection to Deployment

## Part 1

Suman Samui

# Checklist: 8 Main steps

- **Frame the problem & Look at a bigger picture**
- **Get the data**
- **Discover and visualize data to gain insights**
- **Prepare the data for Machine Learning Algos**
- **Explore different models and short-list the best one**
- **Fine-tune your models and combine them a greater solution**
- **Present your Solution**
- **Launch, Monitor and Maintain your system**

# Frame the Problem

- What is the basic objective of your project??
- How your solutions will be used?
- Supervised/unsupervised/Reinforcement Learning??
- Classification/Regression Task??
- Batch/Online Learning??
- Performance measure???
- Assumptions???

# Get the data

**Without data, there is no machine learning.**

- Where is your data?
- How much data do you need?

# Where from Real data can be collected?

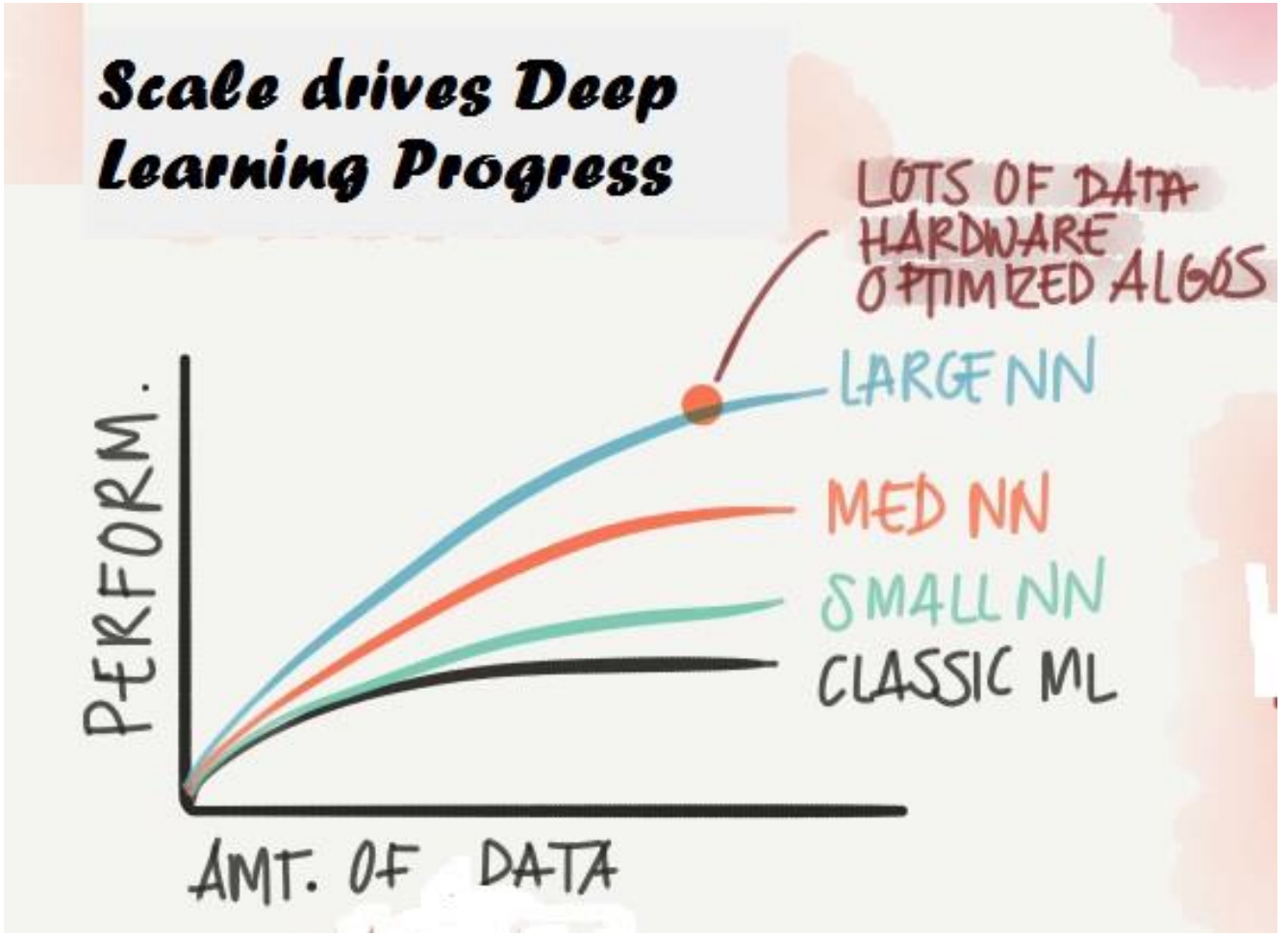
- Provided by the project mentor (database application manager).

- Third-party service providers

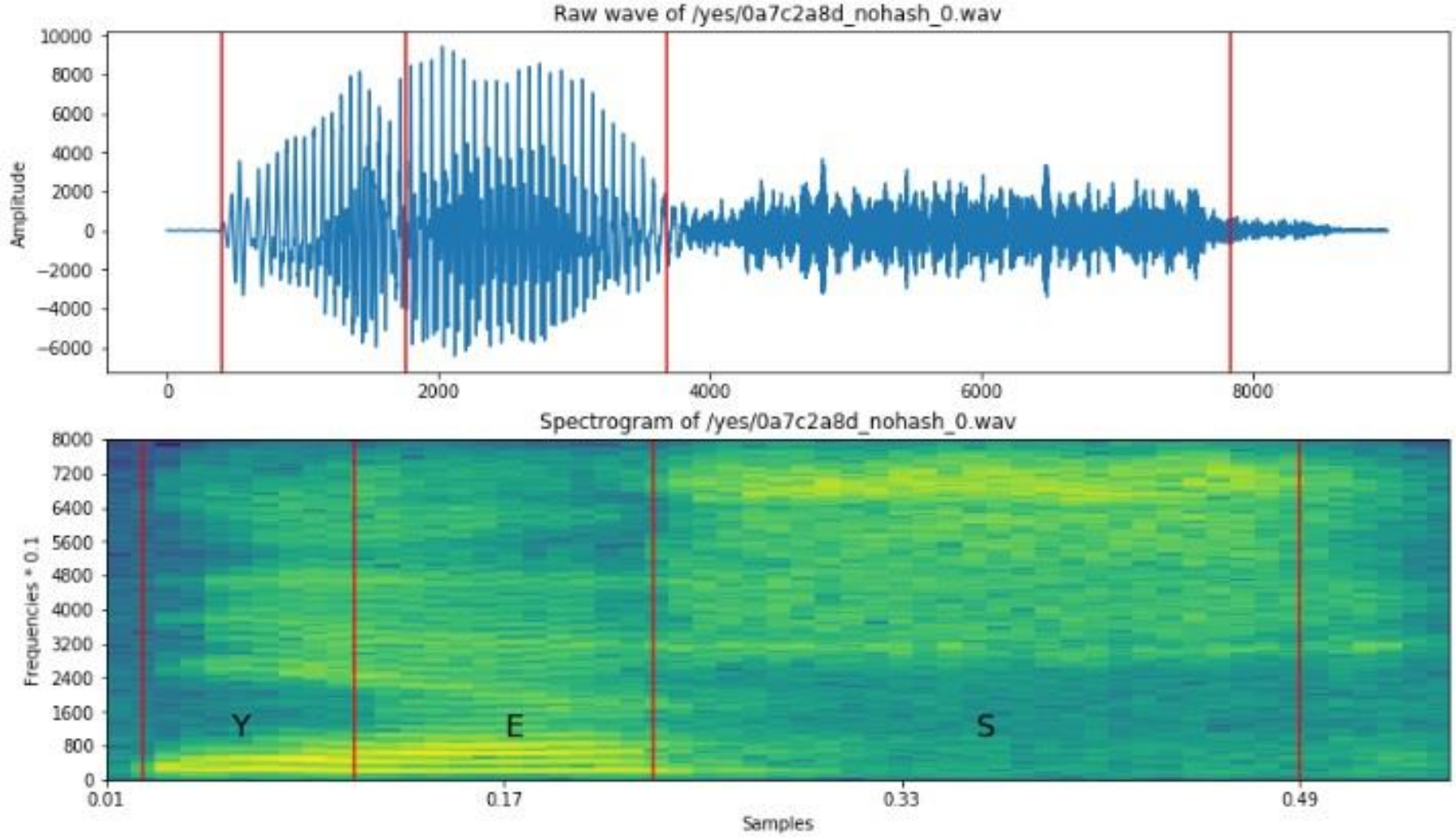
| Data repository | Weblink |
|---|---|
| **Kaggle Datasets** | https://www.kaggle.com/datasets |
| **UCI Machine Learning Repository:** | https://archive.ics.uci.edu/ml/index.php |
| **Amazon Datasets** | https://registry.opendata.aws/ |

# How much data do you need?

- No definite Answer
- Rule of thumb is to collect and use much data as much as possible.



# Data visualization to gain more insights

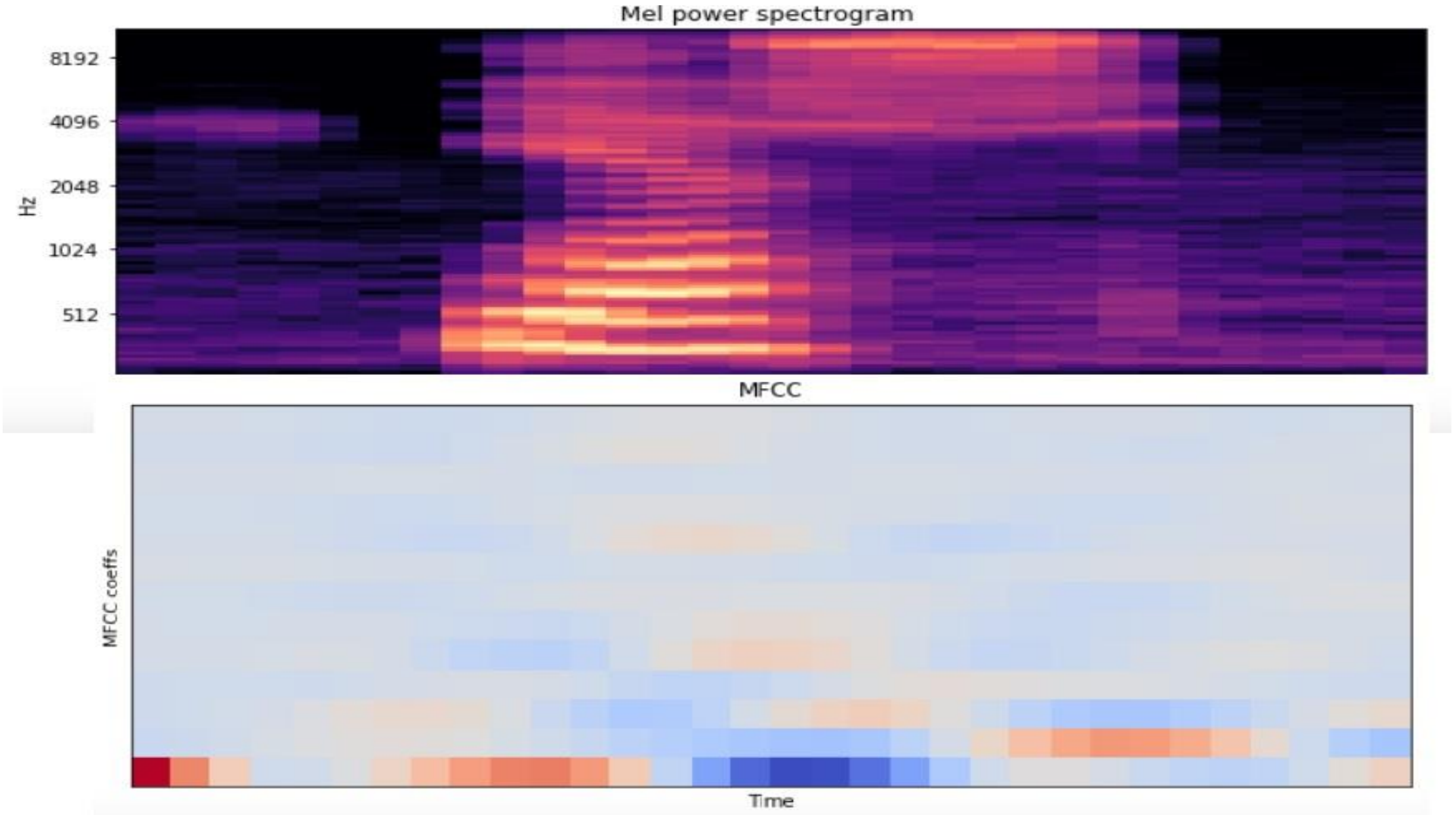



- Silence Removal
- Audio Length
- Noise free or noisy

# Prepare the data

| Structured Dataset | Unstructured Dataset |
| --- | --- |
| (Data given in table format or comma separated numbers)<br>• Data cleaning → Get rid of missing values<br>• Experimenting with Attribute combinations<br>• Feature Selection<br>• Feature Scaling | (Audio, Image, video files, etc.)<br>• Data Cleaning<br>• Feature Extraction<br>• Feature Selection<br>• Feature Scaling |

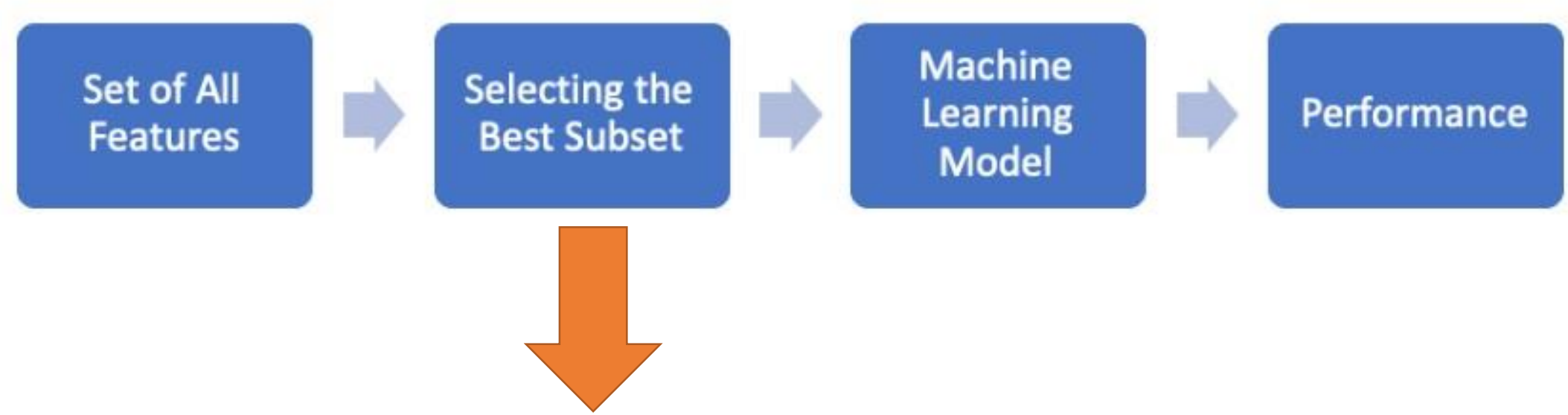
# Audio/Speech → Feature extraction



# Feature Scaling

- Different features have different ranges and statistical distribution
- If you normalize the features this will speed up the training process
- These scaling should be applied to training, dev, and testing sets (but using mean and variance of the train set). Target values don’t need any normalization.
- Two methods: Min-max scaling and Standardization

# Feature Selection Methods




Scikit-learn

- **Univariate feature selection**
- **Recursive feature elimination (RFE)**
- **Recursive feature elimination with cross-validation (RFECV)**