



Birla Institute of Technology and Science Pilani, Hyderabad Campus

08.08.2025

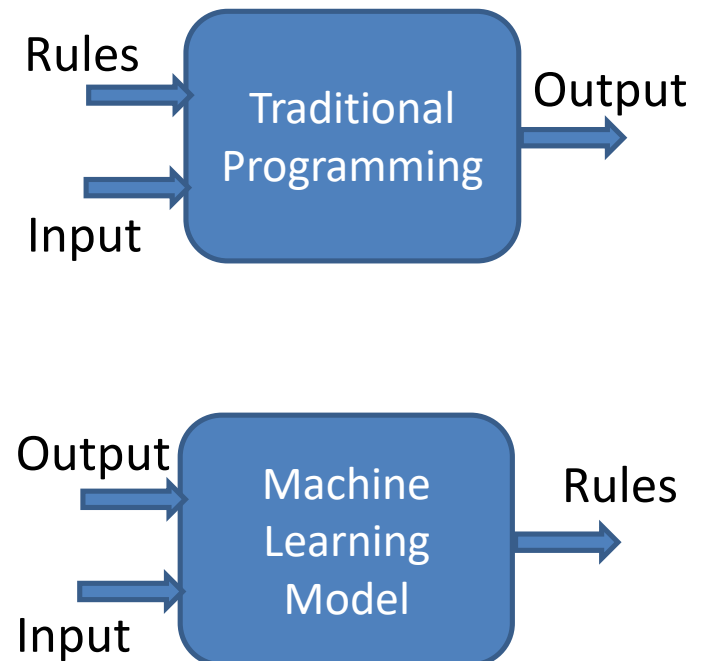
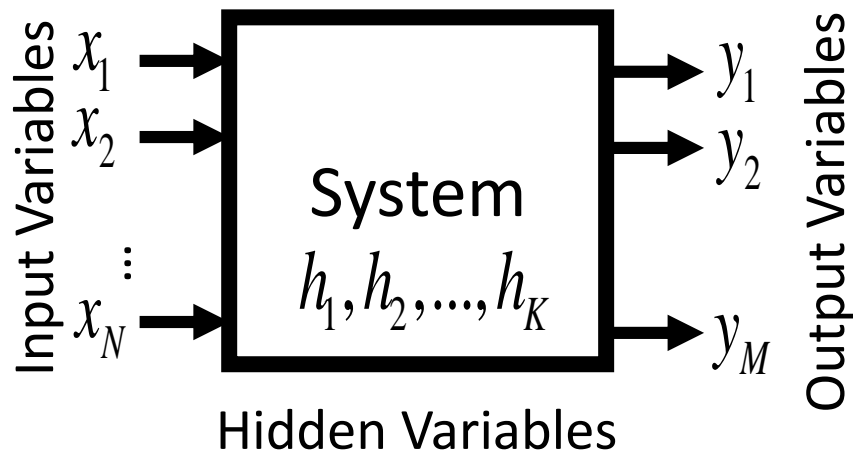
BITS F464: Machine Learning (1st Sem 2025-26)

MACHINE LEARNING OVERVIEW

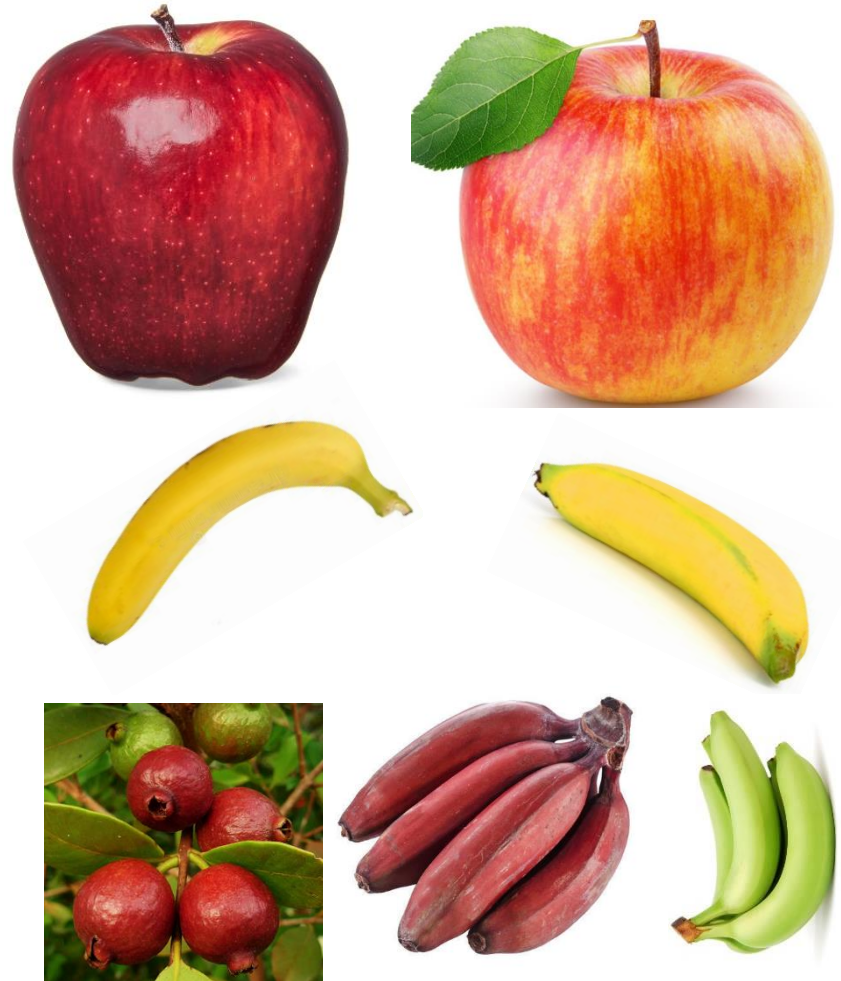
Chittaranjan Hota, Sr. Professor
Dept. of Computer Sc. and Information Systems
hota@hyderabad.bits-pilani.ac.in

What is Machine Learning?

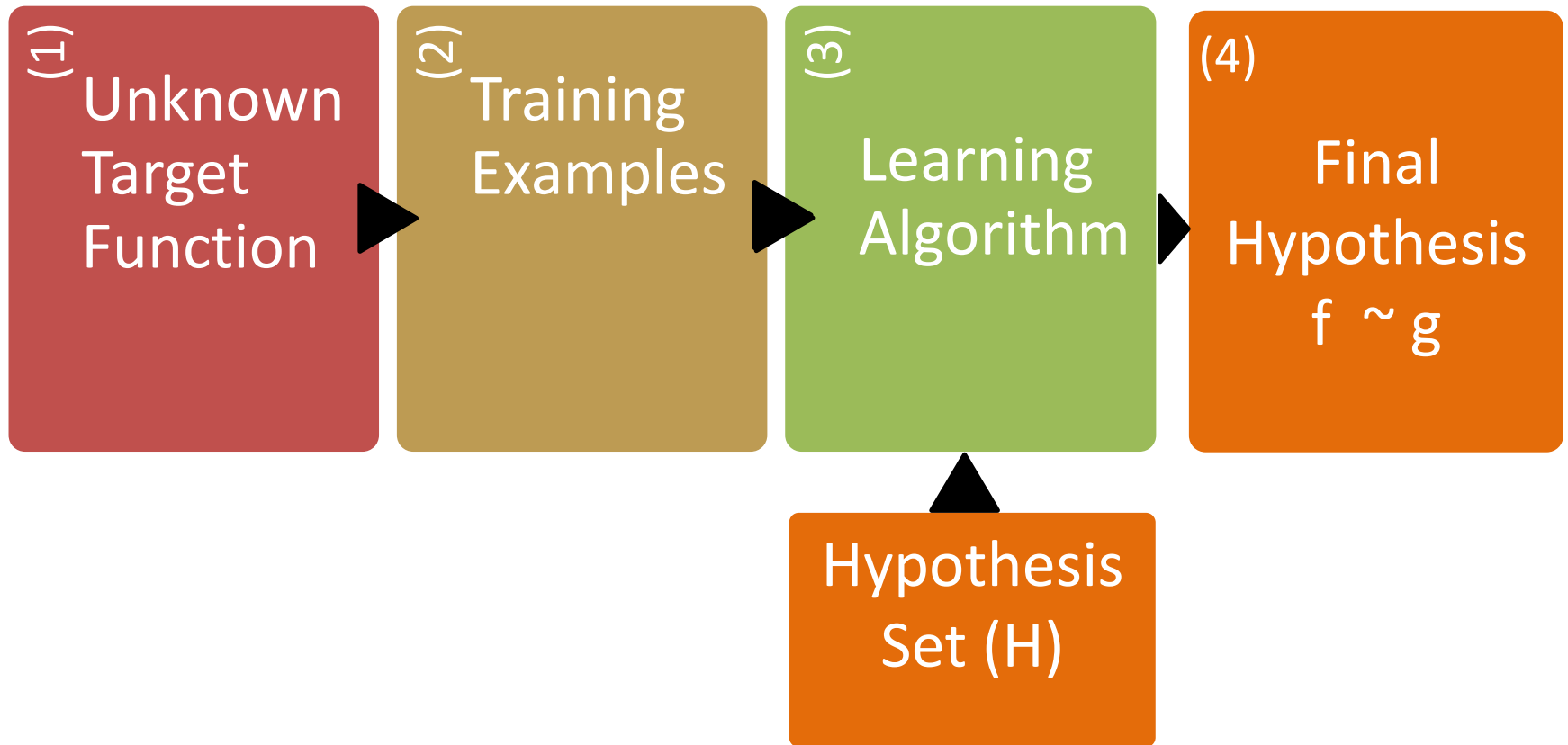
- Optimize a performance criterion using example data or past experience.



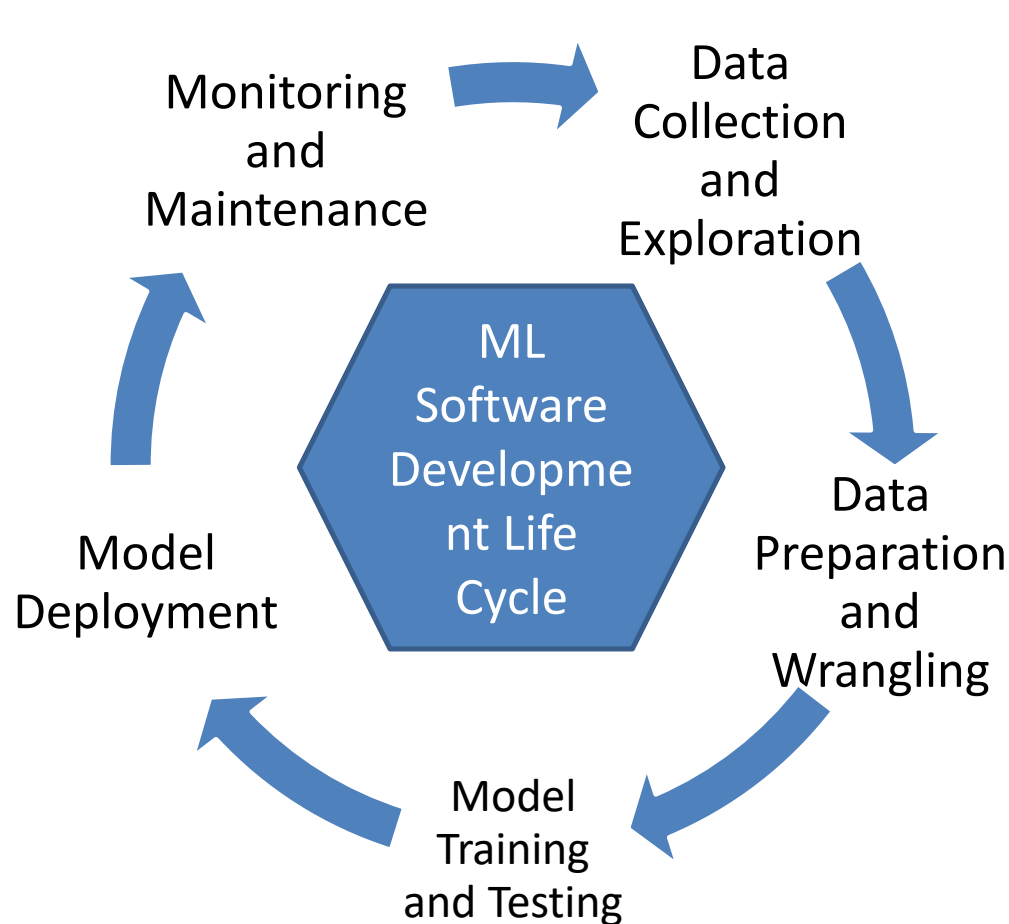
How do we Learn?



Simple Learning Process



ML Software Development Life Cycle

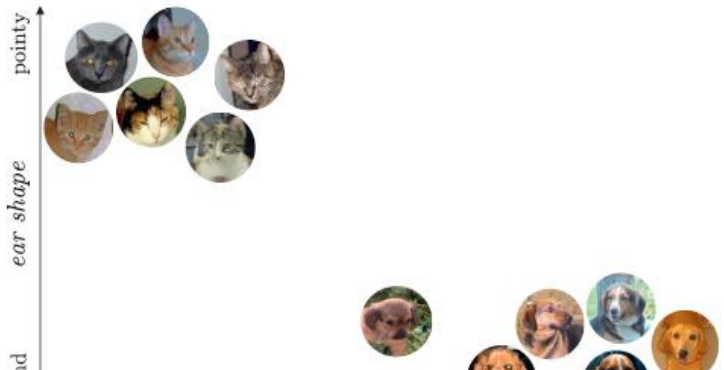


An Example: Step 1: Collecting the data

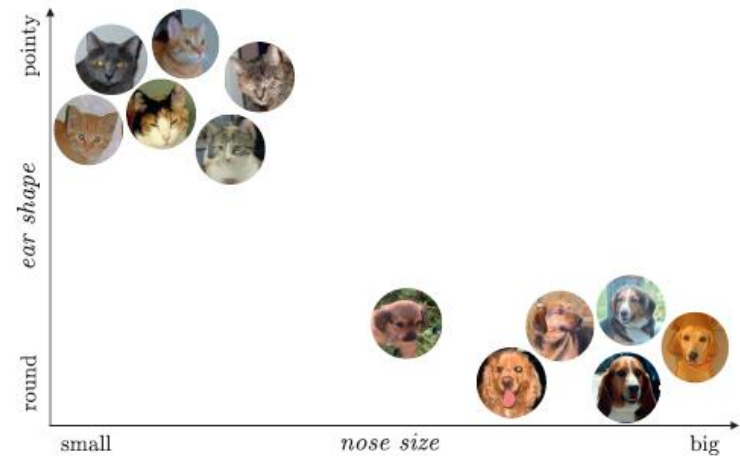


(Training Set)

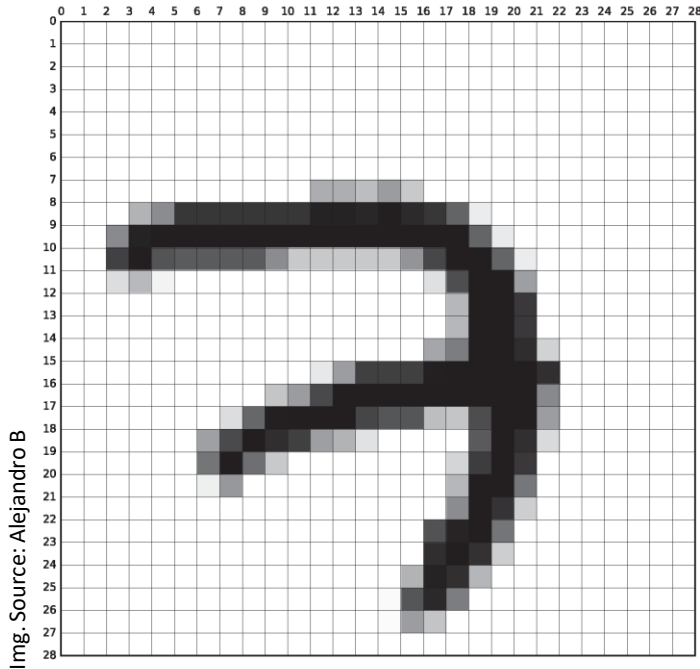
Step 2: Designing the features

- Not a trivial task. How should you go about taking a set of quality features?
 - For ex: Would you like to take “**number of legs**” as one feature to distinguish cats from dogs?
 - A good one for our example:
 - size of nose, relative to the size of the head (ranging from small to big);
 - shape of ears (ranging from round to pointy).
- 

➤ Called as ?



Another ex. Step 2: handwritten digits



(a) MNIST sample belonging to the digit '7'.



(b) 100 samples from the MNIST training set.

```
tfds  
tfds.load(  

```

set 11.06 MiB

Features: Pixel values, Image size,
Aspect Ratio, Normalized Pixel
values, edges, ... (manual)



Automatic: CNN: texture,
shape, corners,...

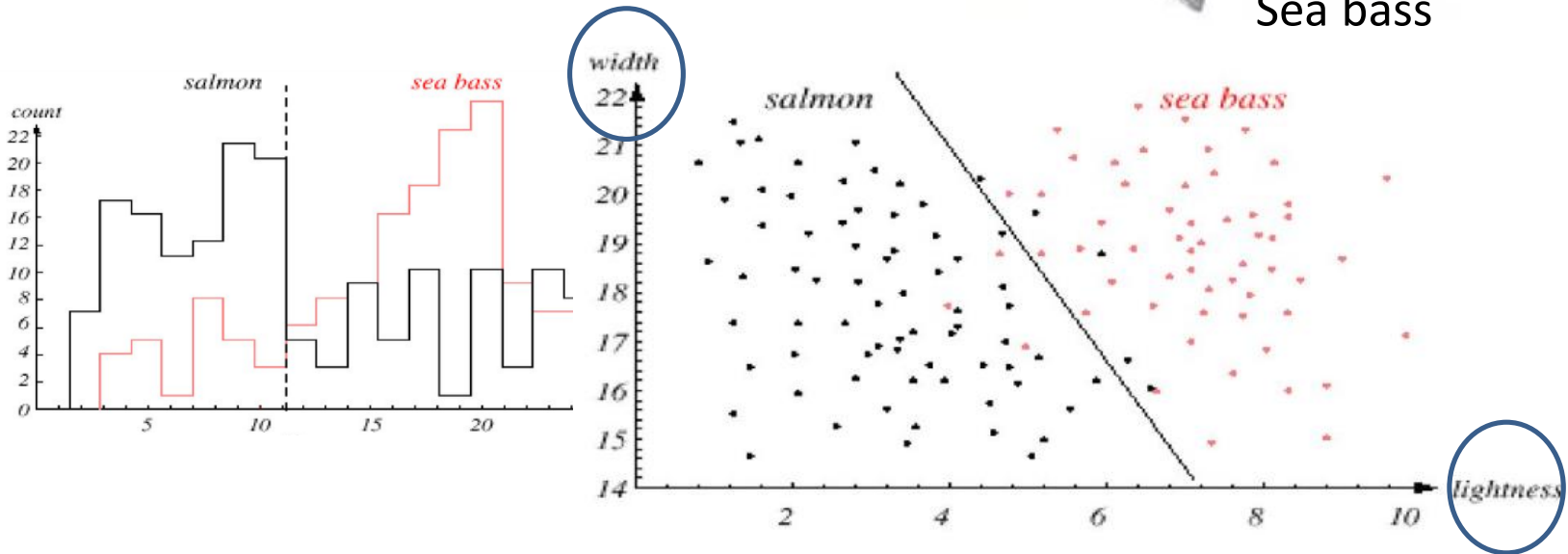
Step 2: Designing features (Another Ex.)



Salmon



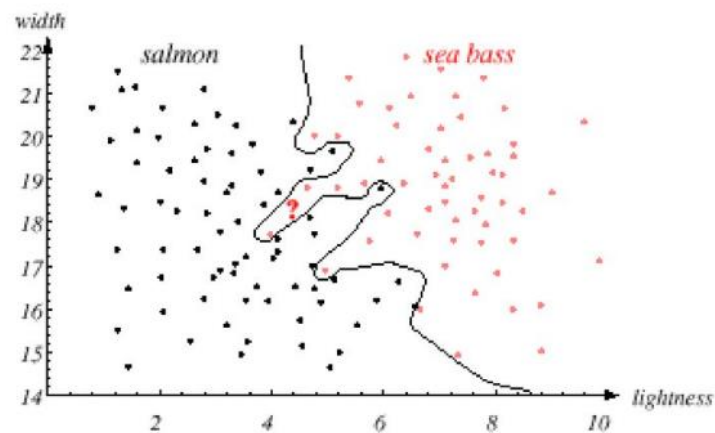
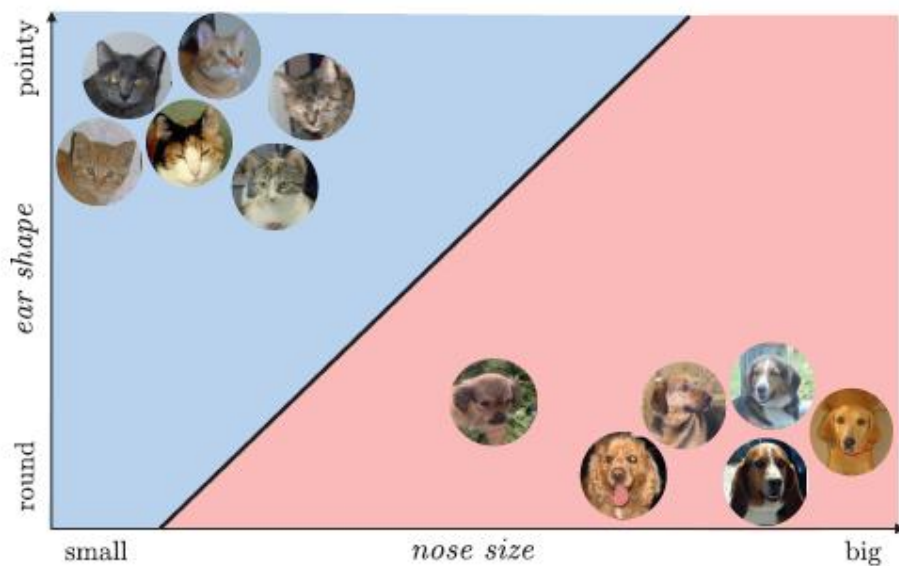
Sea bass



Earlier Examples: Cats Vs Dogs, Handwritten digits

Step 3: Training the Model (Cats Vs Dogs)

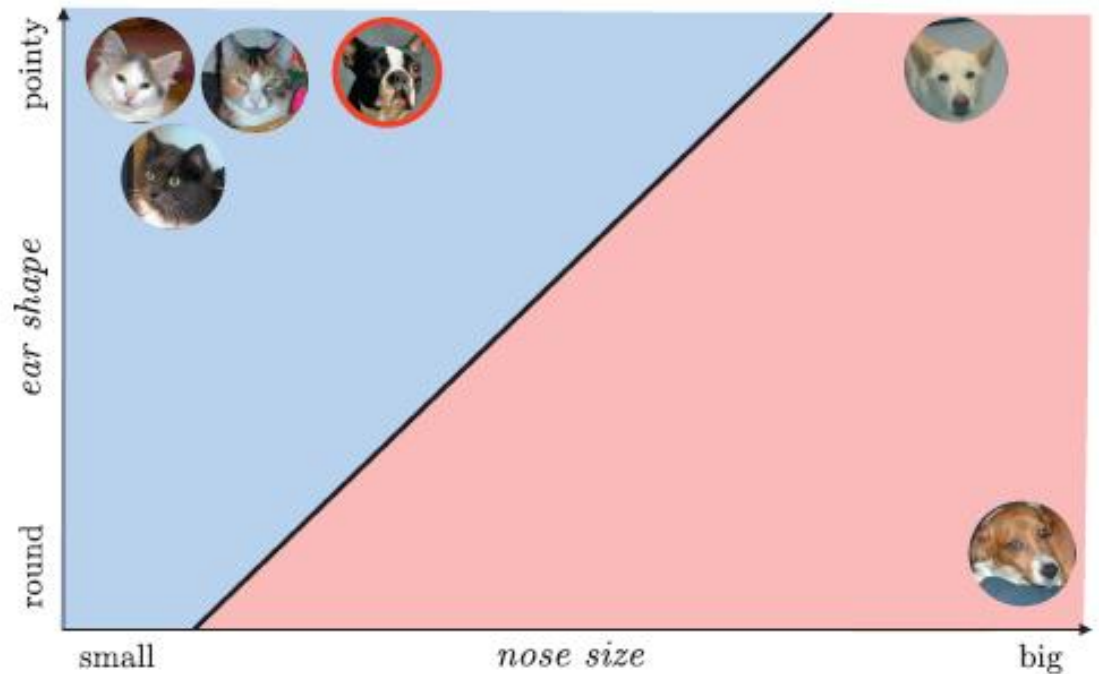
- Now it is a simple geometric problem. Let the computer find out a Line (linear model) that separates cats from dogs.



How good this model would be?

We could instead find a curve or **nonlinear model** that separates the data. In general, linear models are by far the most common choice in practice when features are designed properly.

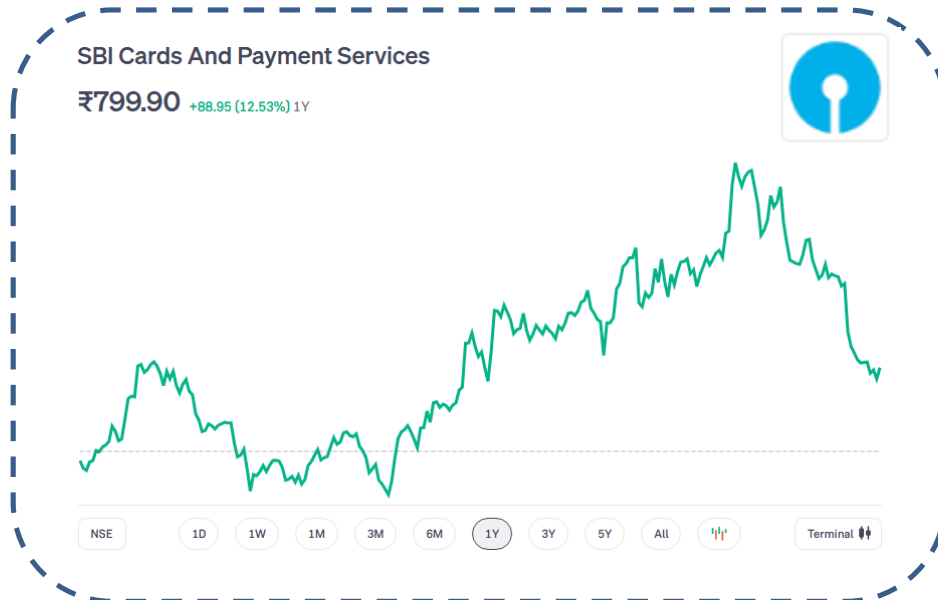
Step 4: Testing the Model



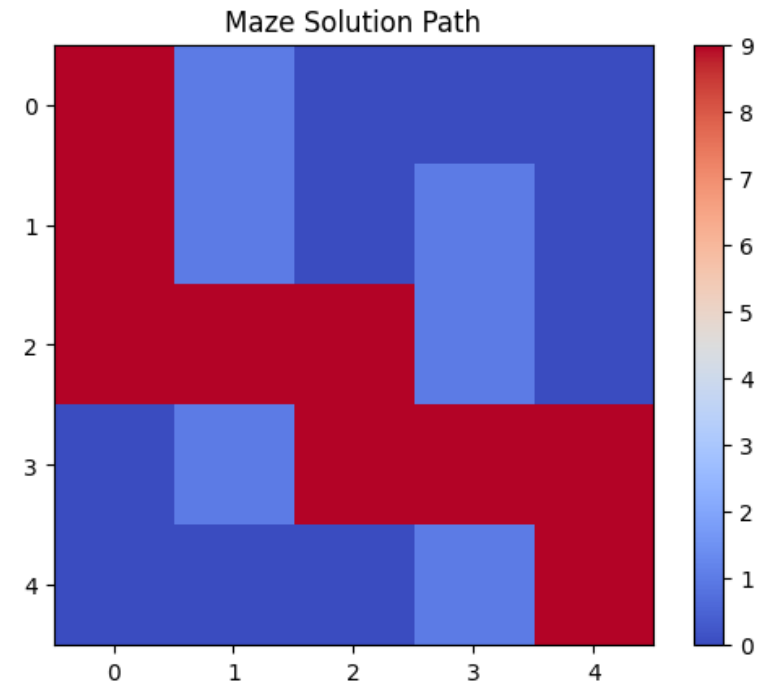
- What is the problem here?
- Can you list down few more discriminating features?

Types of Machine Learning

Shopper stop(1) Big Bazar(3) Tata Trent(2) Life style(1)



(Supervised: Stock prediction) 11.08.2025



(Reinforcement: Robotics or Gaming)

Subject: Urgent: Verify Your Account Immediately!
 From: "Security " <support@secure-banking-alerts.com>
 To: hota@hyderabad.bits-pilani.ac.in

Dear Customer,

We have detected unusual login activity on your online banking account. For your protection, your account access has been temporarily limited.

To restore full access, please verify your account within the next 24 hours:

[Click here to verify your account](#)

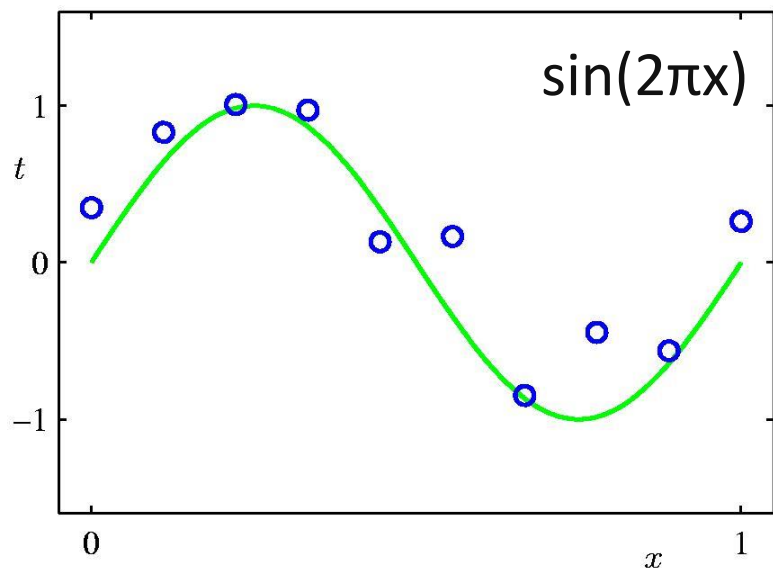
Failure to complete verification may result in permanent account suspension.

Account Security Department, SBI

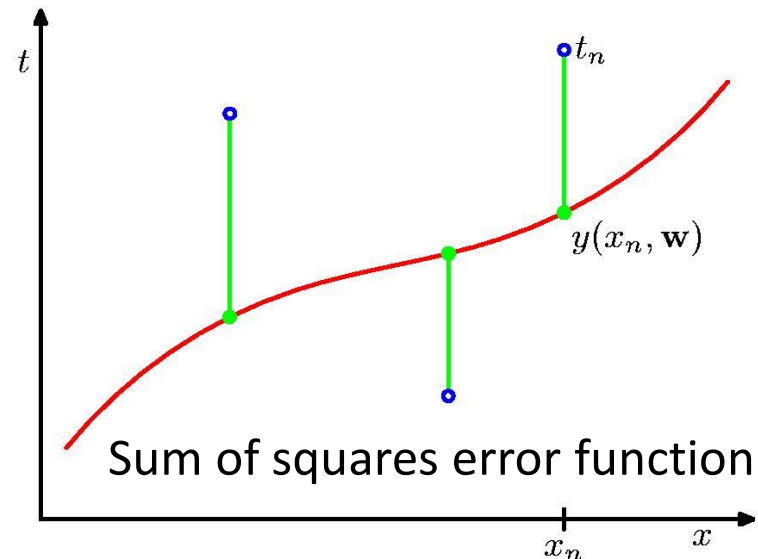
(Semi-Supervised: E-mail spam)

Supervised Learning

- Correct Output known for each training example.
 - **Classification:** 1-of-N output (whether it is a Cat or a Dog?)
 - **Regression:** Real valued output (how many students will enroll into ML course next semester?)

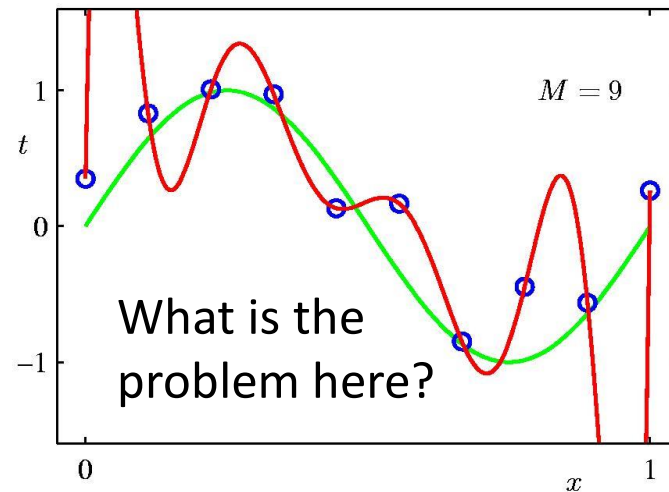
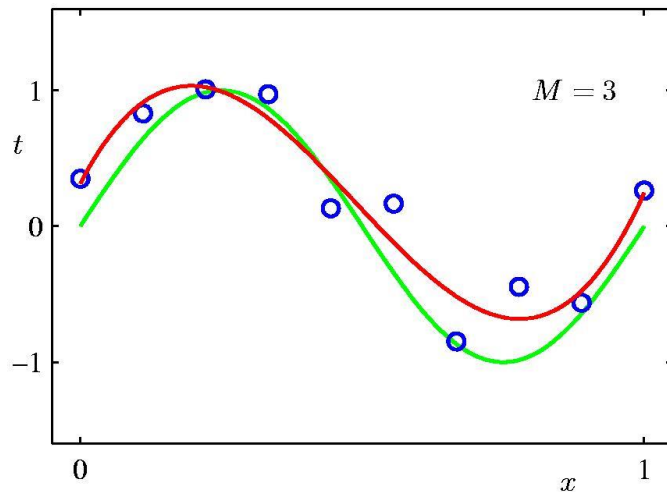
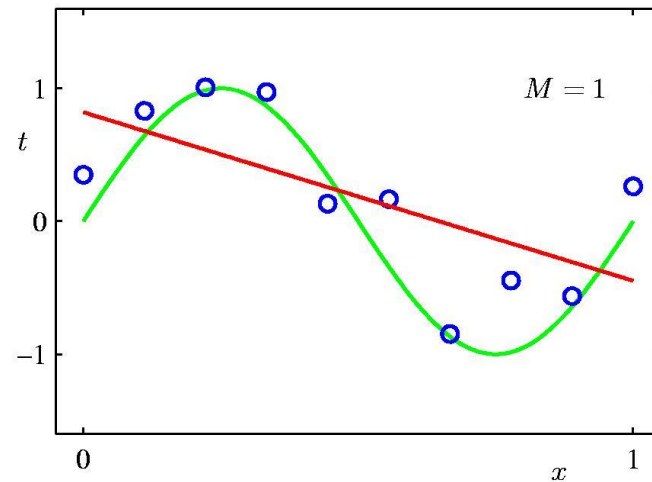
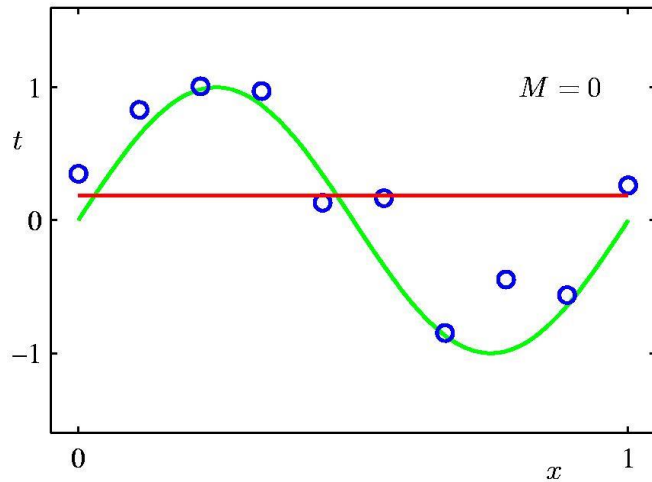


$$y(x, \mathbf{w}) = w_0 + w_1x + w_2x^2 + \dots + w_Mx^M = \sum_{j=0}^M w_jx^j$$

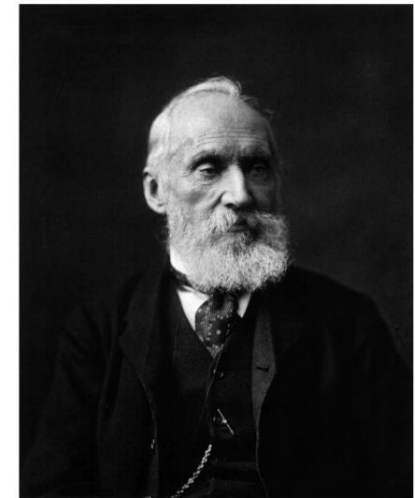
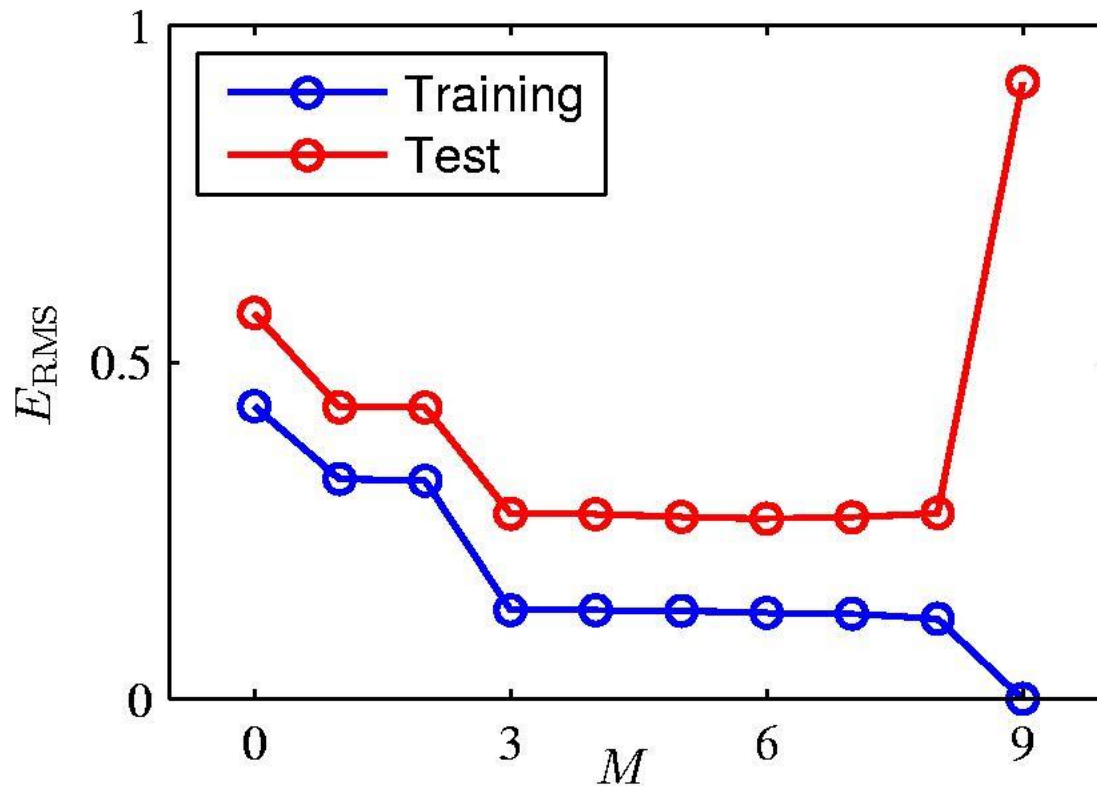


$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2$$

Model selection: What should be M ?



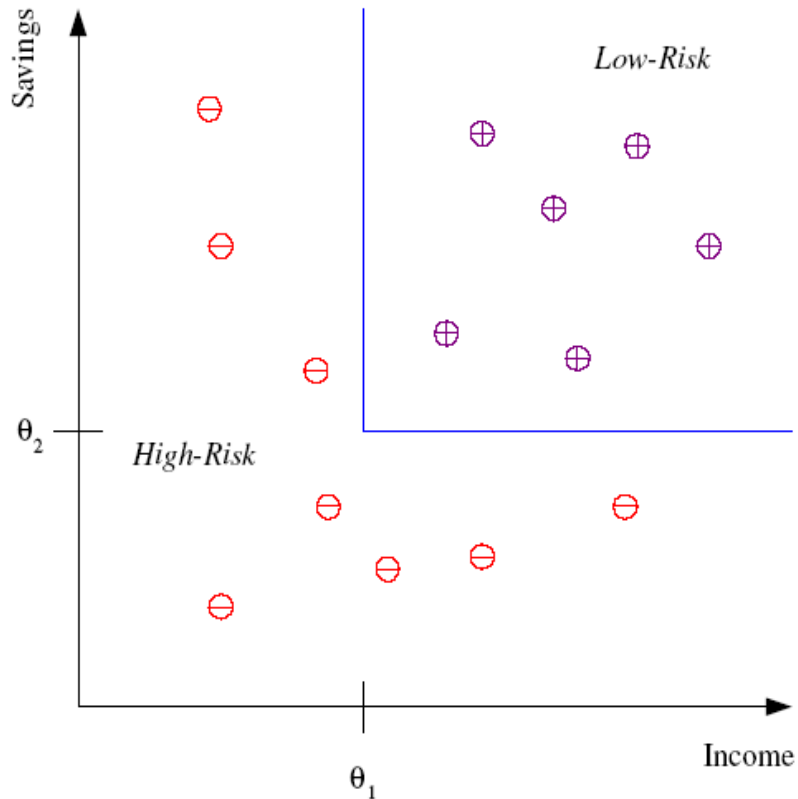
Over-fitting



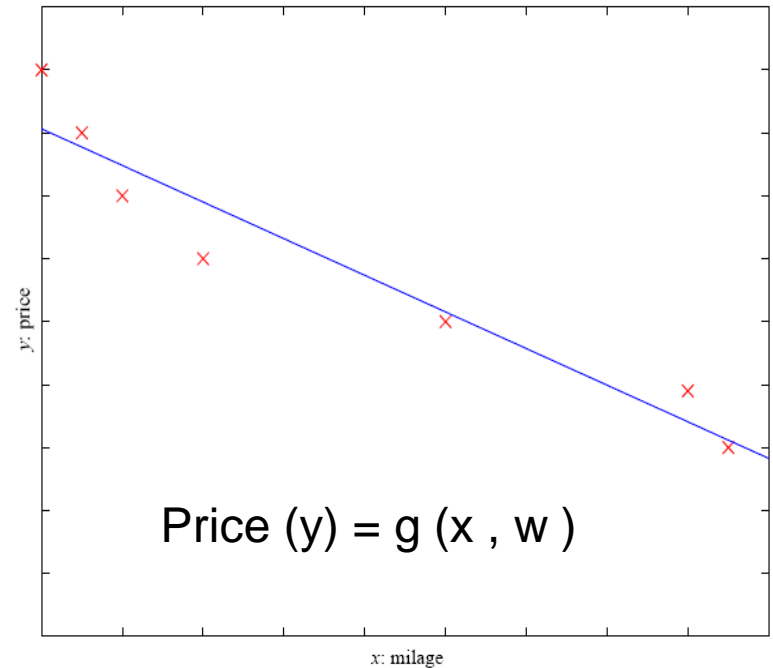
Solutions: later

Root-Mean-Square (RMS) Error: $E_{\text{RMS}} = \sqrt{2E(\mathbf{w}^*)/N}$

Classification Vs. Regression



Rule: IF *income* $> \theta_1$ AND *savings* $> \theta_2$
THEN **low-risk** ELSE **high-risk**



Data Wrangling

1. Data Cleaning:

1. Handling missing data: Imputation (filling with mean/median/mode), drop missing rows/cols
2. Removing duplicates (`df.drop_duplicates()` in **pandas**)
3. Correcting data types (convert strings to dates, integers to float etc.)
4. Fixing inconsistent formatting (Trim whitespace, Remove special characters etc.)

[Let us see these in Colab...](#)

2. Data Transformation:

1. Normalization/ scaling (**min-max**, **Z-score**, **Log scale** etc.)

$$X_{\text{normalized}} = (X - X_{\min}) / (X_{\max} - X_{\min}) \quad Z = \frac{x - \mu}{\sigma} \quad X_{\log} = \log(X)$$

2. Encoding categorical values (one-hot, label etc.)

3. Feature engineering: Creating new features, or aggregating a few

4. Handling Outliers: Remove, or Impute with mean, median, mode

Data Wrangling & Feature Engineering

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
# Simulated dataset
data = {
    'House Size (sqft)': [1400, 1600, 1700, 1875, np.nan, 2100, 2300, 2450, 2700, 3000],
    'Number of Rooms': [3, 3, 3, 4, 4, 4, 5, 5, 5, np.nan],
    'Age of House (years)': [10, 15, 10, 20, 8, 5, 5, np.nan, 3, 1],
    'Price ($)': [300000, 320000, 340000, 360000, 400000, 420000, 450000, 470000, 500000, 520000]
}
df = pd.DataFrame(data)
# Handling missing values
df['House Size (sqft)'].fillna(df['House Size (sqft)'].mean(), inplace=True)
df['Number of Rooms'].fillna(df['Number of Rooms'].mean(), inplace=True)
df['Age of House (years)'].fillna(df['Age of House (years)'].mean(), inplace=True)
```

Continued...

```
df['Price per Sqft'] = df['Price ($)'] / df['House Size (sqft)']
```

Visualization

```
plt.figure(figsize=(10, 6))
```

Scatter plot of House Size vs Price

```
plt.subplot(1, 2, 1)
```

```
sns.scatterplot(x='House Size (sqft)', y='Price ($)', data=df)
```

```
plt.title('House Size vs Price')
```

Line plot of House Age vs Price per Sqft

```
plt.subplot(1, 2, 2)
```

```
sns.lineplot(x='Age of House (years)', y='Price per Sqft', data=df)
```

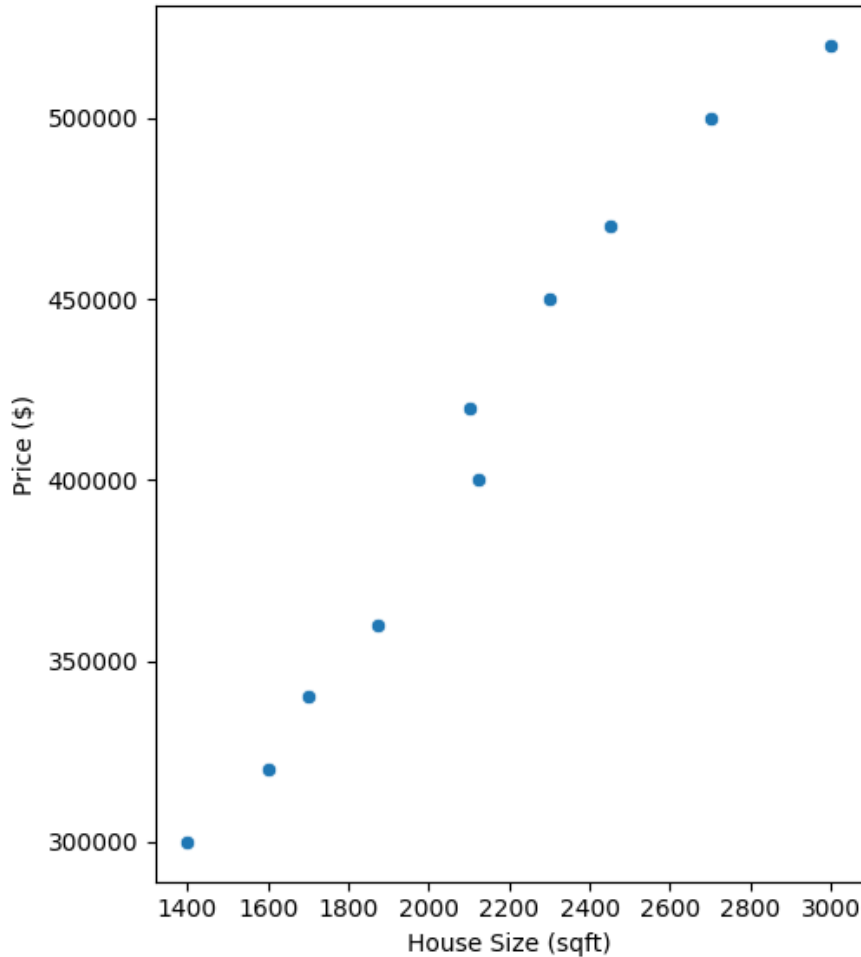
```
plt.title('House Age vs Price per Sqft')
```

```
plt.tight_layout()
```

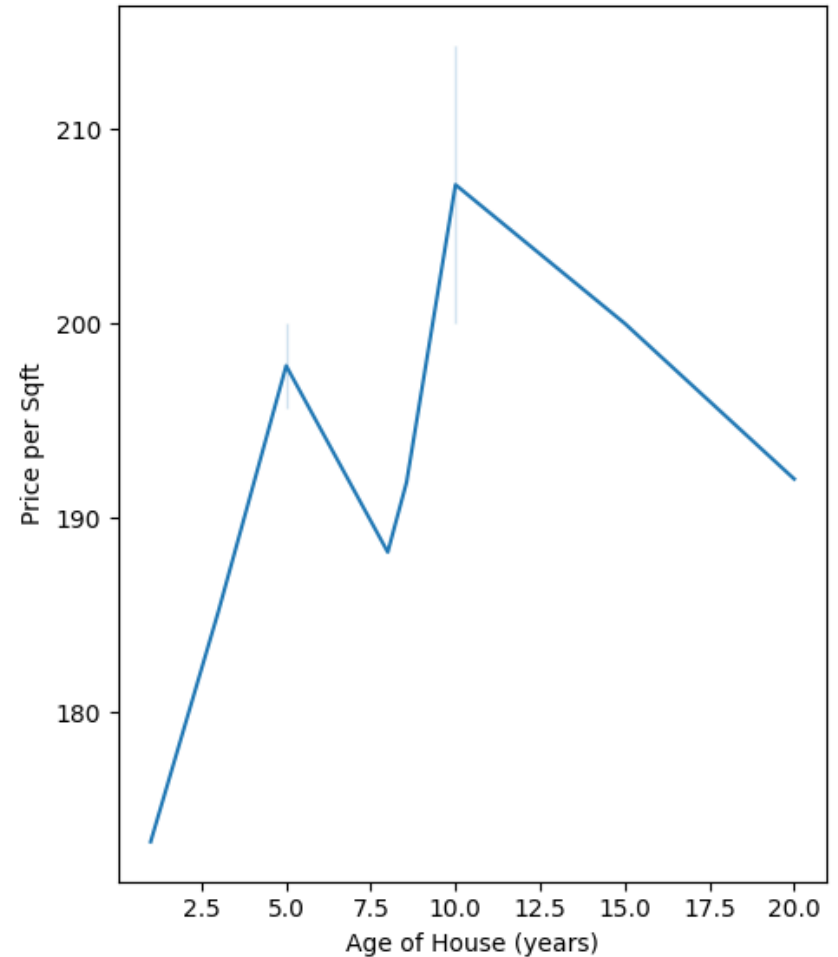
```
plt.show()
```

Example continued...

House Size vs Price



House Age vs Price per Sqft



Exploratory Data Analysis Workflow

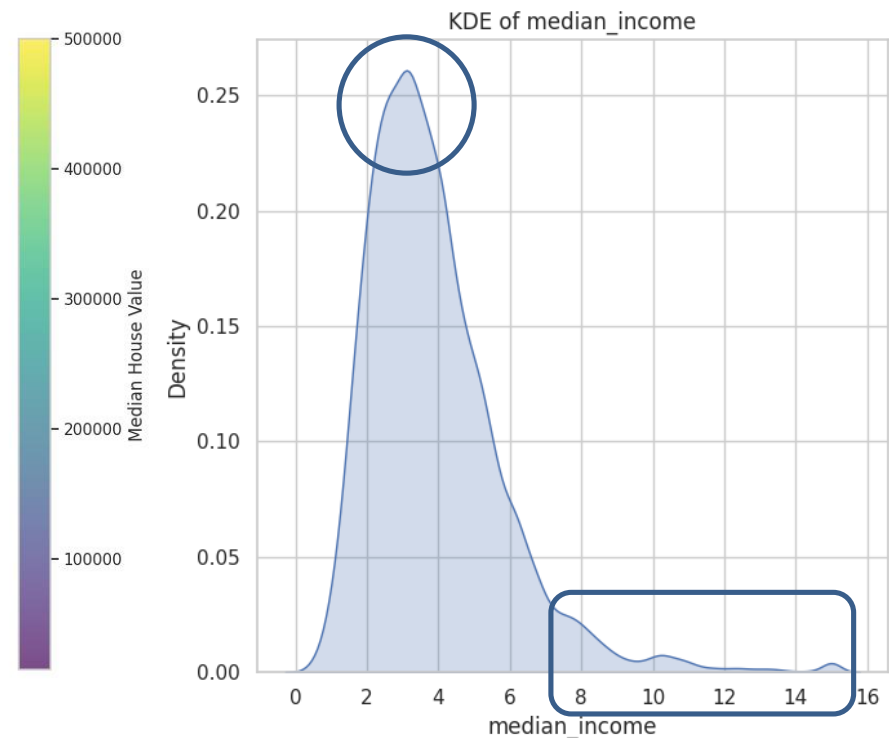
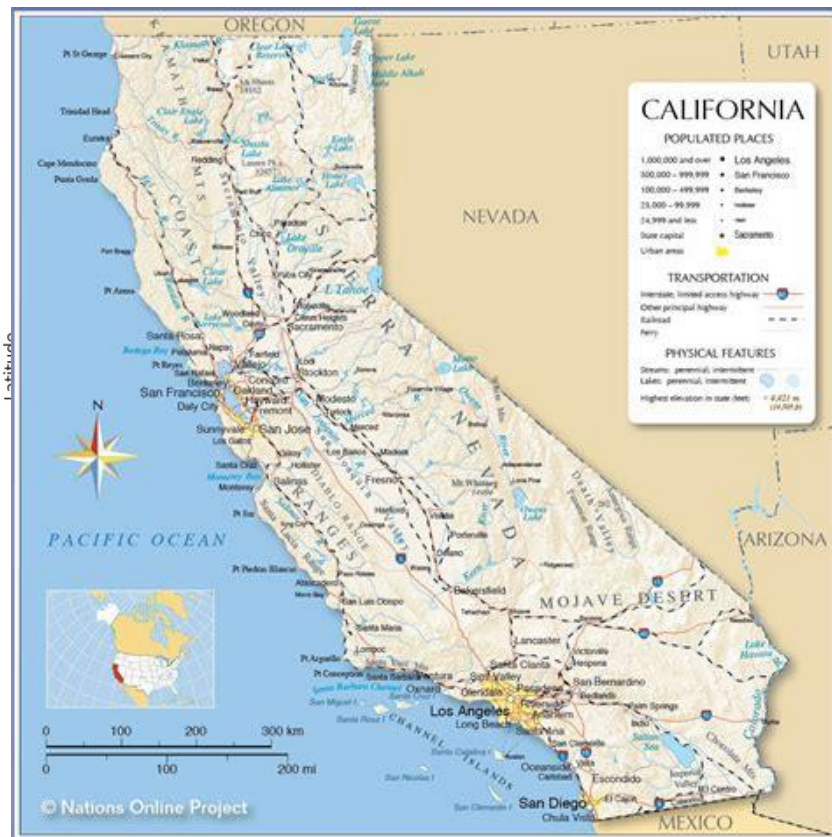
- Examining and summarizing a dataset to understand its' key characteristics before applying a model.
- What hides in the data, spot patterns, find relationships, and detect any strange anomalies or errors.
- **Tasks:** Load, Inspect, Clean, Univariate and Multivariate Analysis, Detect patterns/trends/outliers.



Let us see this happening with “tips” dataset built into Seaborn!

total_bill	tip	sex	smoker	day	time	size
16.99	1.01	Female	No	Sun	Dinner	2
10.34	1.66	Male	No	Sun	Dinner	3
21.01	3.50	Male	No	Sun	Dinner	3

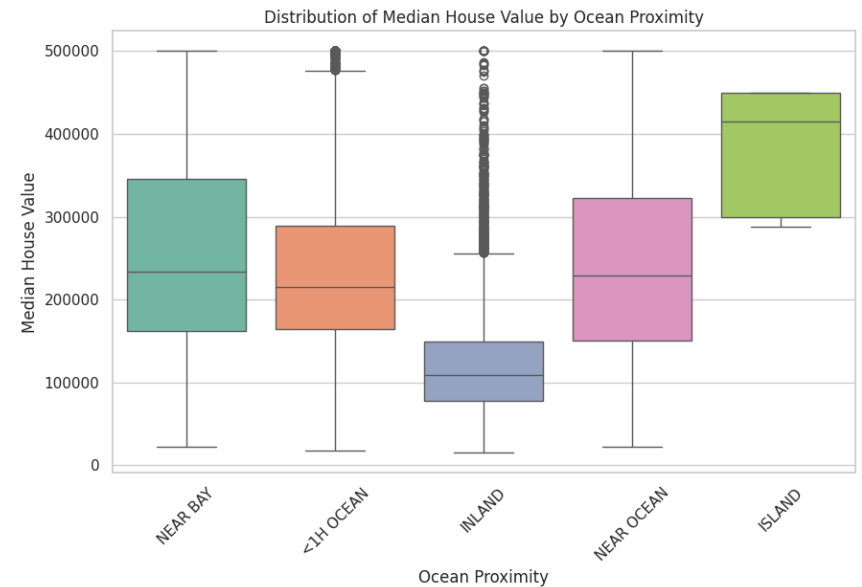
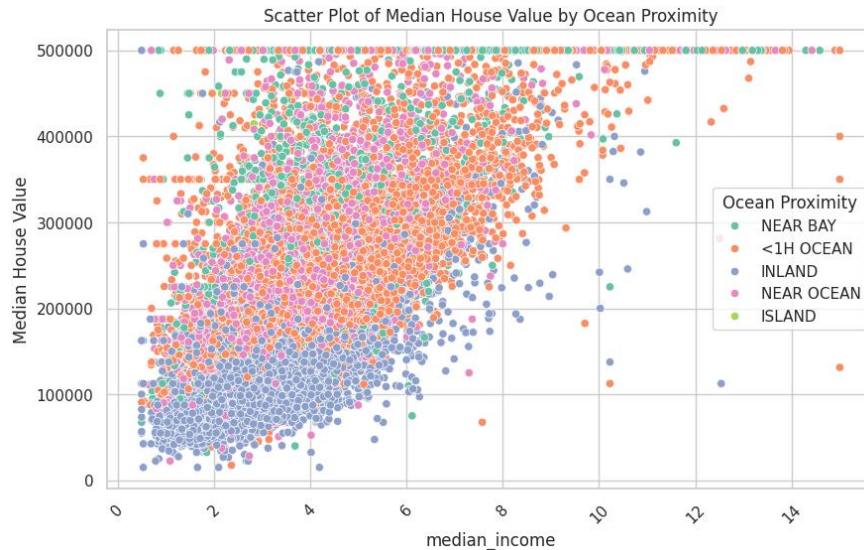
A More Bigger Example: Data Exploration



Most California districts had middle-class income.

- Fewer very-rich districts.

Continued...



Assignment 1: Scatter, Histogram, Heatmap, Box, KDE: Submission deadline: 31.08.2025

What are some Issues in Machine Learning?

- What algorithms are available for learning a concept? How well do they perform? [For example: Classifying spam emails.](#)
 - How much training data is sufficient to learn a concept with high confidence?
 - How are the features generated?
 - Are some training examples more useful than others?
 - What are the best tasks for a system to learn?
-

ML Dataset Resources

- UCI ML Repository: <https://archive.ics.uci.edu/>
 - Kaggle ML Datasets: <https://www.kaggle.com/datasets>
 - Google ML Datasets: <https://research.google/resources/datasets/>
 - Open Data on AWS: <https://aws.amazon.com/opendata/>
 - Image Datasets: MNIST, ImageNet, CIFAR-10, CIFAR-100
 - NLP Datasets: GLUE, SQuAD, HuggingFace
 - Clinical Dataset: MIMIC-IV
 - Wikipedia's ML Datasets
-

Thank you!
