Birla Institute of Technology and Science Pilani, Hyderabad Campus

09.08.2024

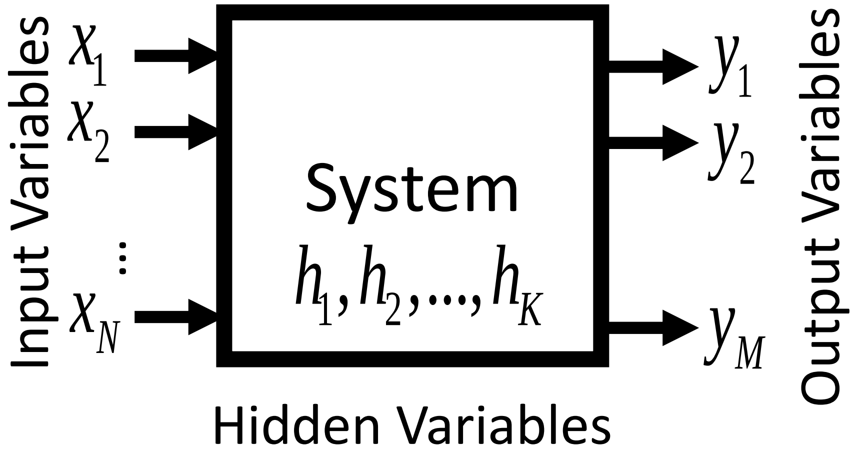
BITS F464: Machine Learning (1st Sem 2024-25)

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.



Input Variables

x_1

x_2

$\vdots$

x_N

System

$h_1, h_2, \ldots, h_K$

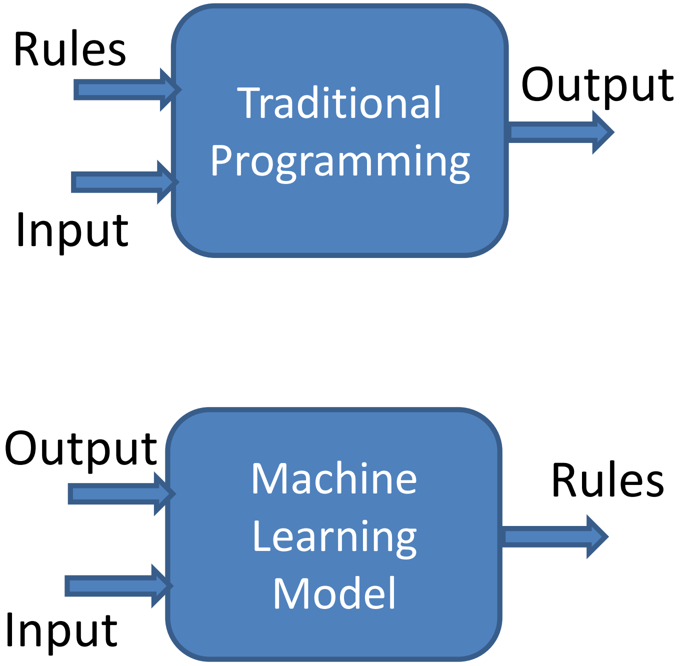
Hidden Variables

y_1

y_2

y_M

Output Variables



Rules

Input

Traditional Programming

Output

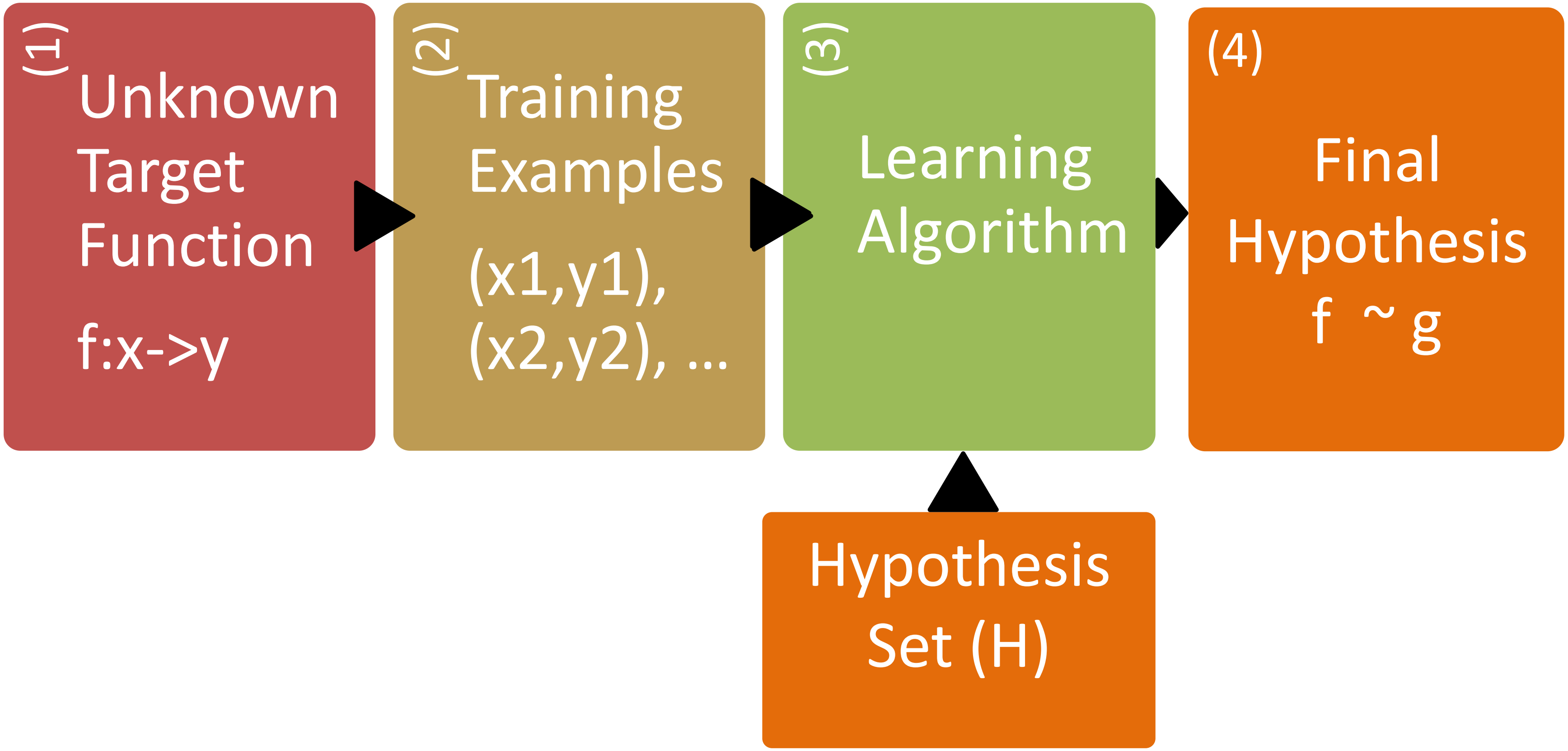
Output

Input

Machine Learning Model

Rules

Simple Learning Process



(1) Unknown Target Function

f:x->y

(2) Training Examples

(x1,y1), (x2,y2), …

(3) Learning Algorithm

(4) Final Hypothesis

f ~ g

Hypothesis Set (H)

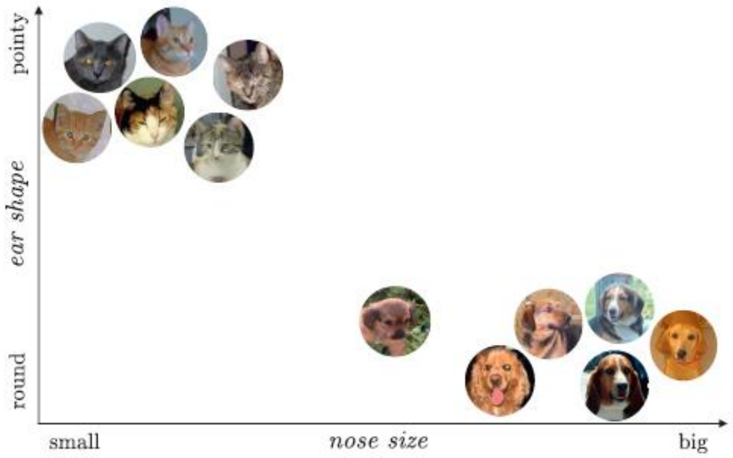
An Example: Step 1: Collecting the data

(Training Set)

Step 2: Designing the features

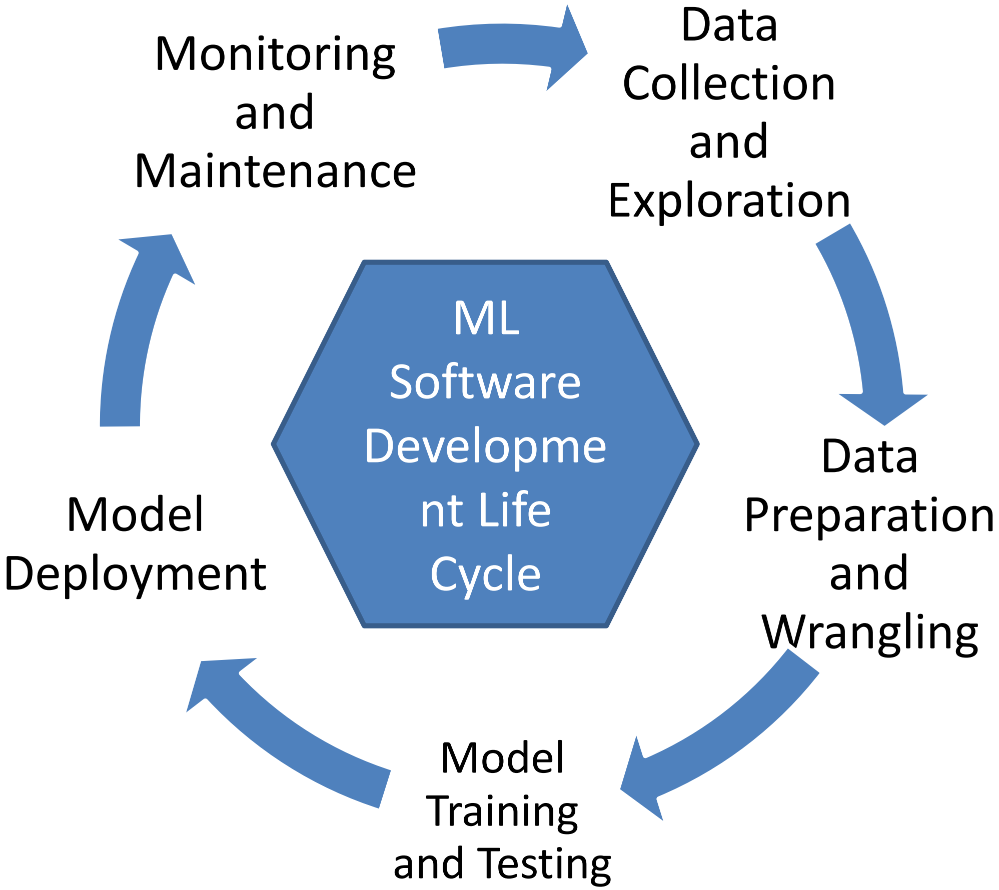
- Not a trivial task. Designing quality features could be very application dependent.
- For ex: Would you like to take "number of legs" as one feature to distinguish cats from dogs?

- 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 a feature vector.

ML Software Development Life Cycle



ML Software Development Life Cycle

Data Collection and Exploration

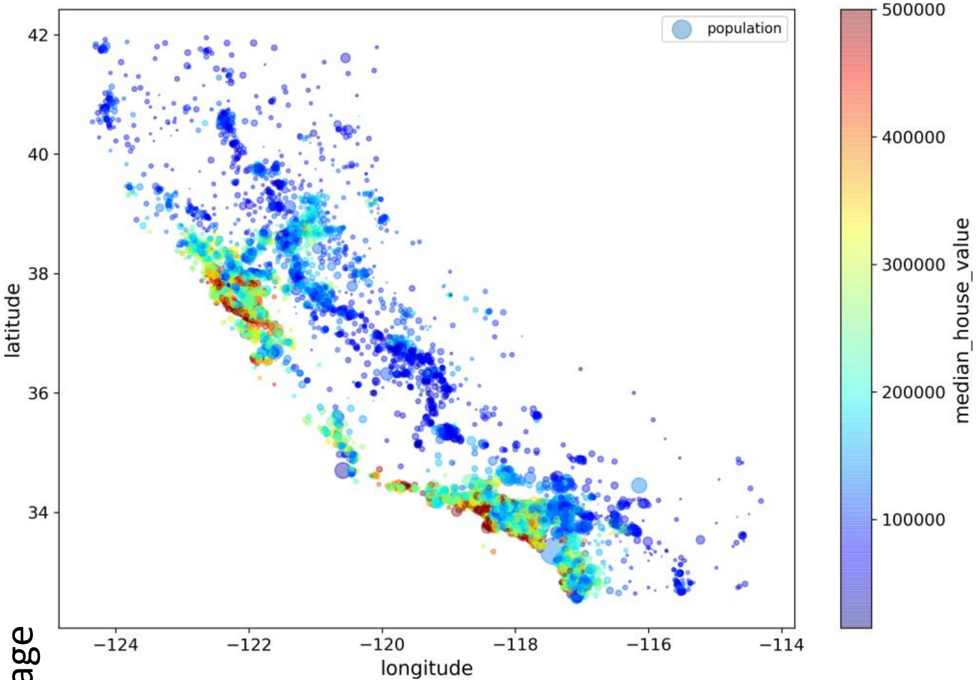
Data Preparation and Wrangling

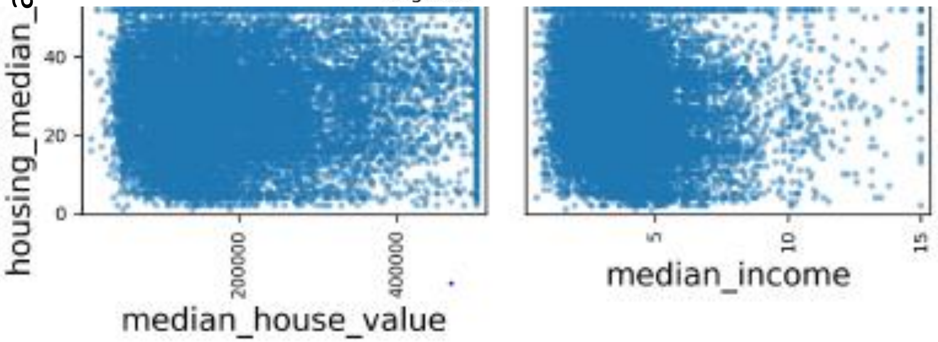
Model Training and Testing

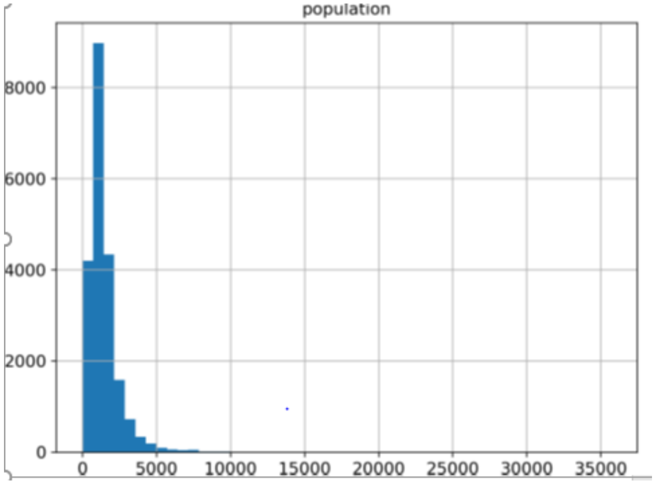
Model Deployment

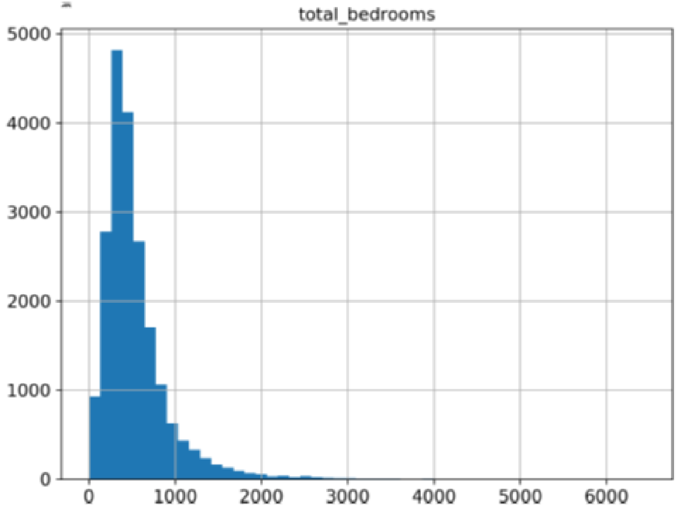
Monitoring and Maintenance

Data Exploration









Data Wrangling

- Handling missing data
  - Imputation (filling with mean/median/mode), Removal (dropping)
- Data Cleaning
  - Removing duplicates, Correcting errors (eg, incorrect data types), Handling outliers (removing or transforming them)
- Data Transformation
  - Normalization/ standardization (min-max, Z-score, Log scale etc.), Encoding categorical values (one-hot, label etc.) etc.

$X_{normalized} = X - X_{min} / X_{max} - X_{min}$ $X_{standardized} = X-\mu / \sigma$ $X_{log} = \log(X)$

- Data Reduction
  - Dimensionality reduction (PCA), Feature engineering etc.

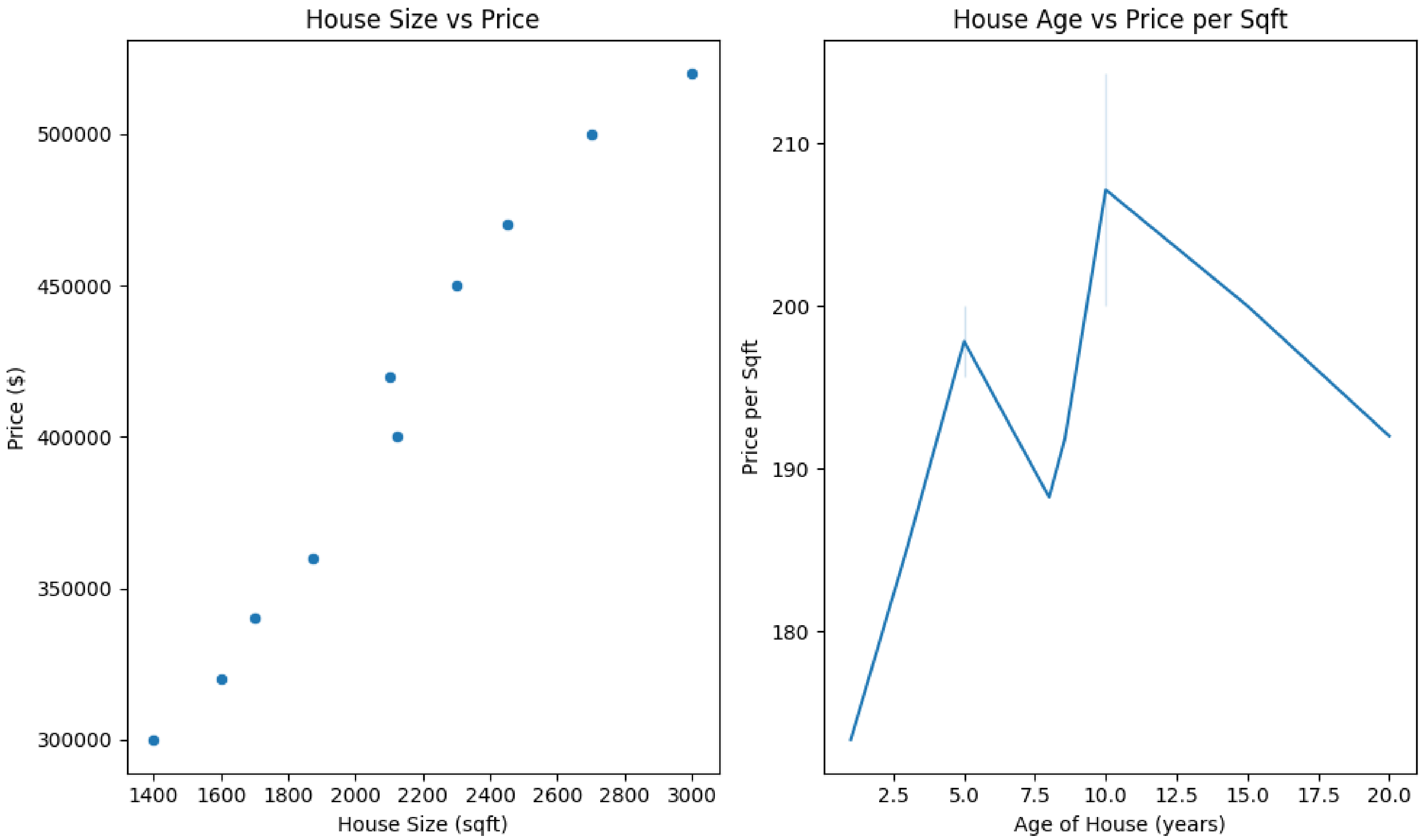
Example Data Wrangling

```
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)
```

Example continued…

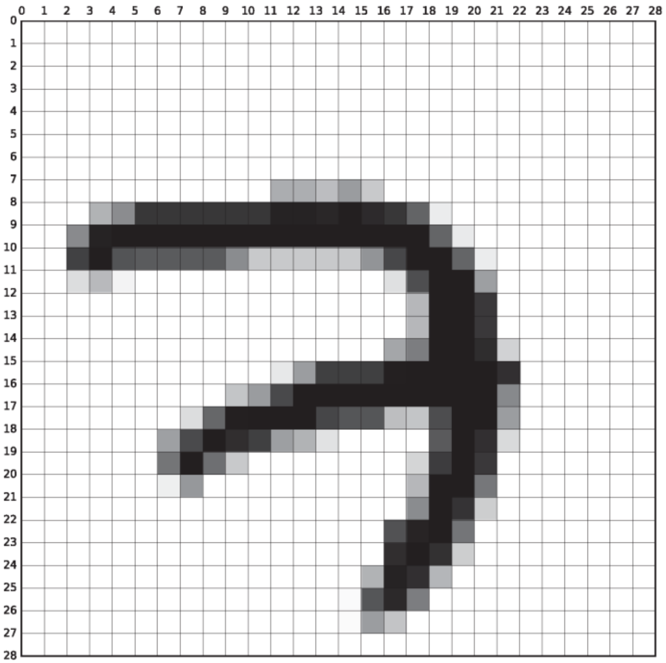
```
# Feature Engineering: Price per Sqft
df['Price per Sqft'] = df['Price ($)'] / df['House Size (sqft)']

# 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…



First Assignment on Data Exploration and Wrangling: Coming week

Step 2: Continued (Handwritten digits)



Img. Source: Alejandro B

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

(b) 100 samples from the MNIST training set.

```
:fds
 tfds.load(
```

How big?

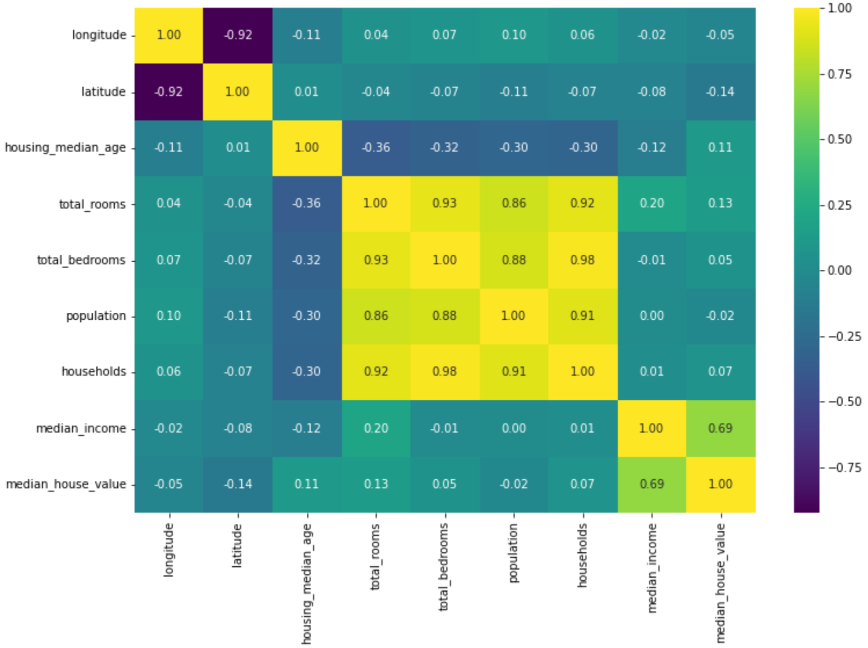
```
set 11.06 MiB
repared to /ro
```

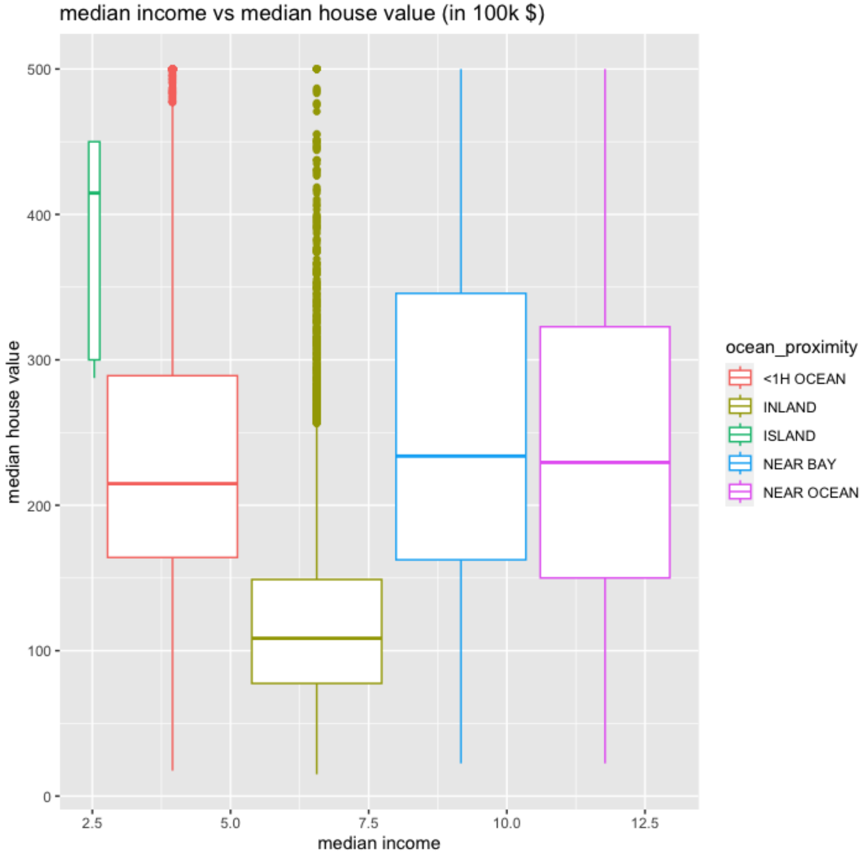
Features: Pixel values, Image size, Aspect Ratio, Normalized Pixel values, edges, ... (manual) →

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

Recap: Data Exploration and Wrangling

Que: If outliers are more in the dataset, will you choose RMSE or MAE to find out the Model error?





```
plt.figure(figsize=(12, 8))
sns.heatmap(df.corr(),annot=True,fmt='.2f',
cmap='viridis')
plt.show()
```

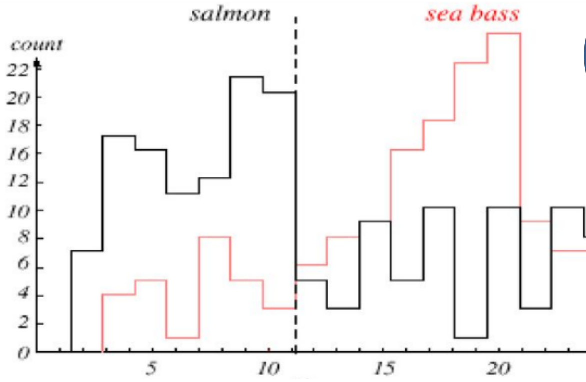
Img. Source: https://rpubs.com/snehganjoo/1018157

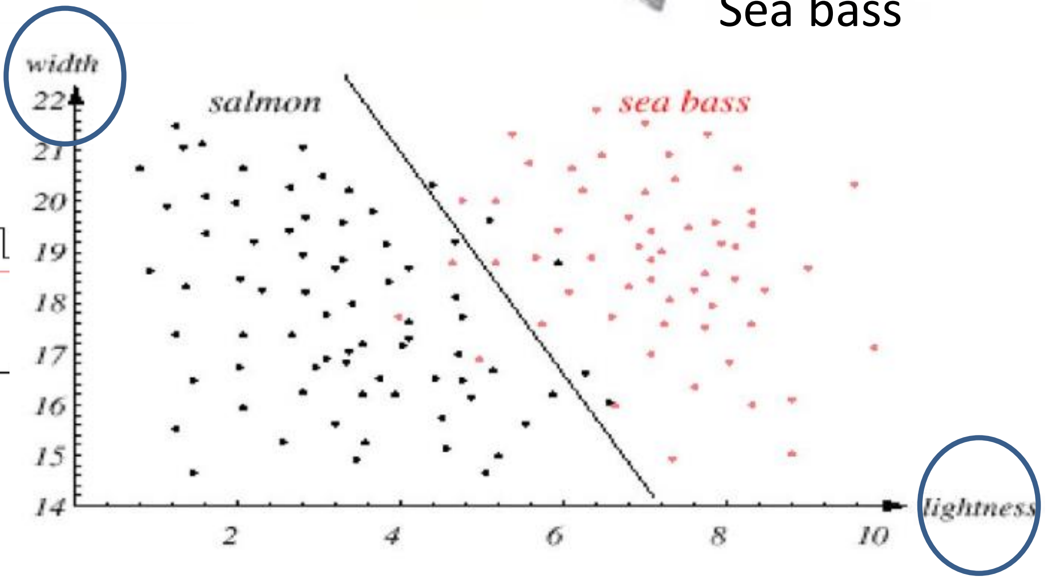
Assignment 1: Scatter, Histogram, Heatmap, Box, KDE: Submission deadline: 20.08.2024

Step 2: Designing features (Another Ex.)

Salmon

Sea bass





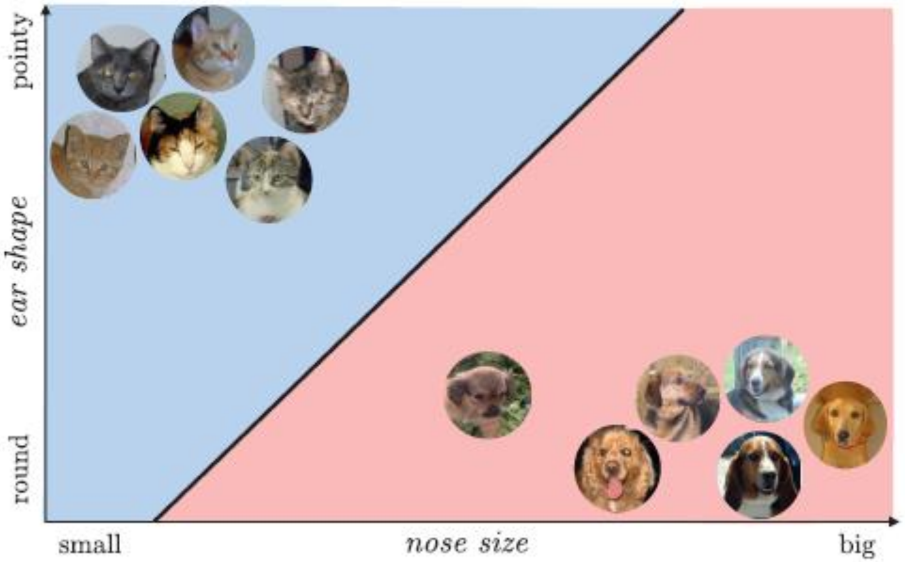
Fish → $x^T = [x_1 , x_2]$

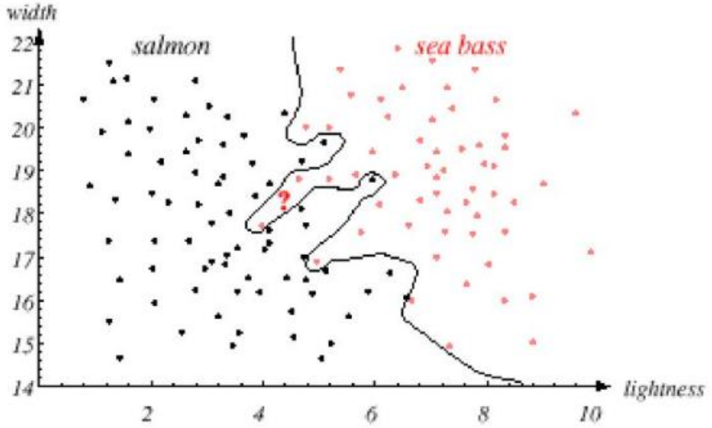
lightness width

Earlier Examples: Cats Vs Dogs, Handwritten digits, House prices

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.



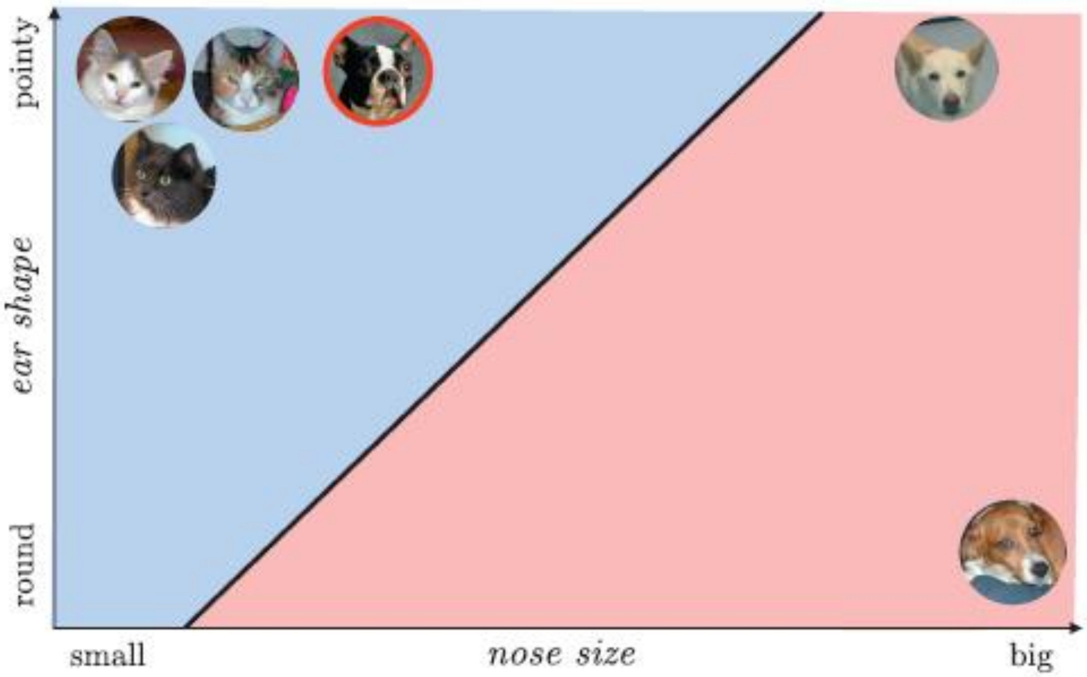


- How good this model would be?

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.

Img. Source: Sergios Theodoridis

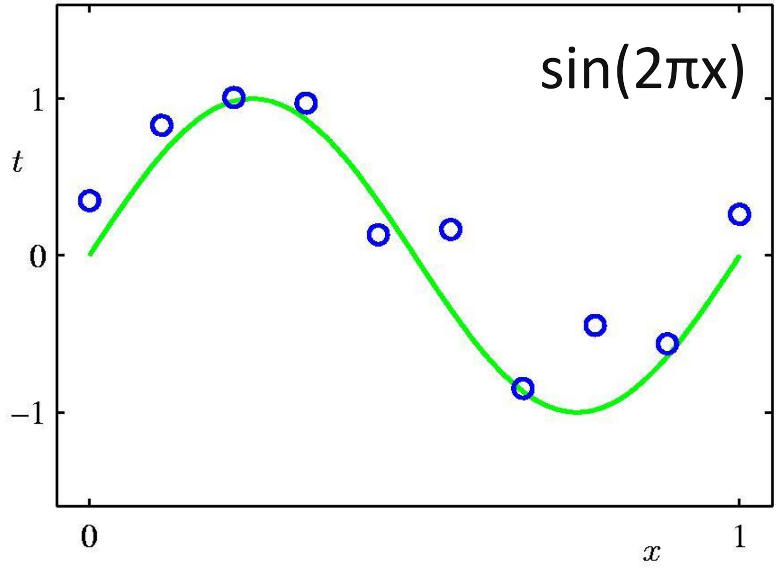
Step 4: Testing the Model



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

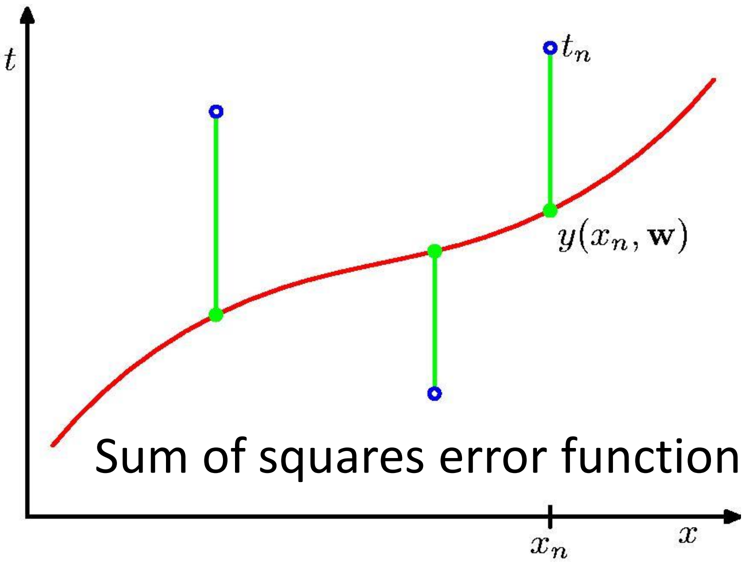
Types of Learning: Supervised Learning

- 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?)



sin(2πx)

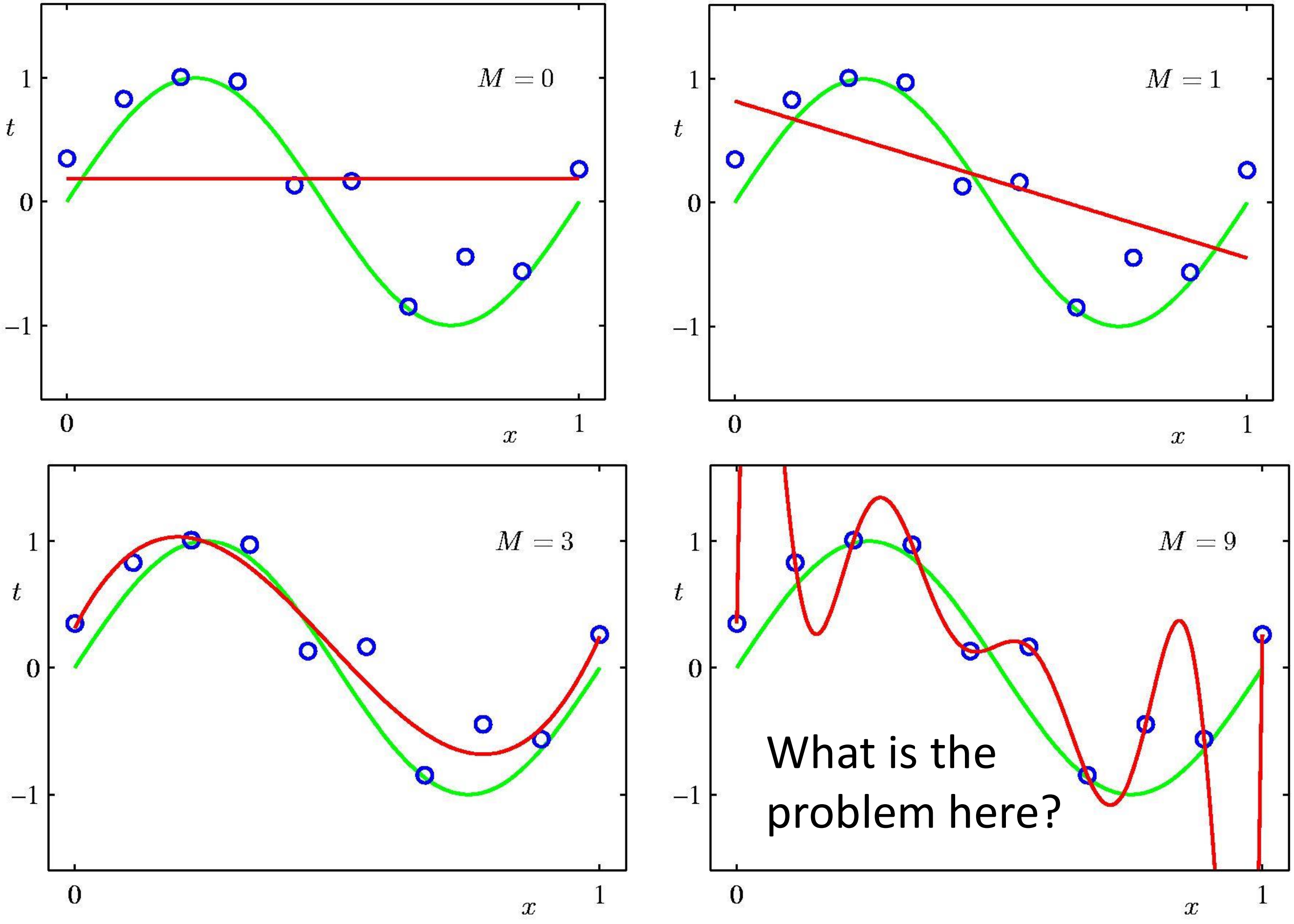
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$



Sum of squares error function

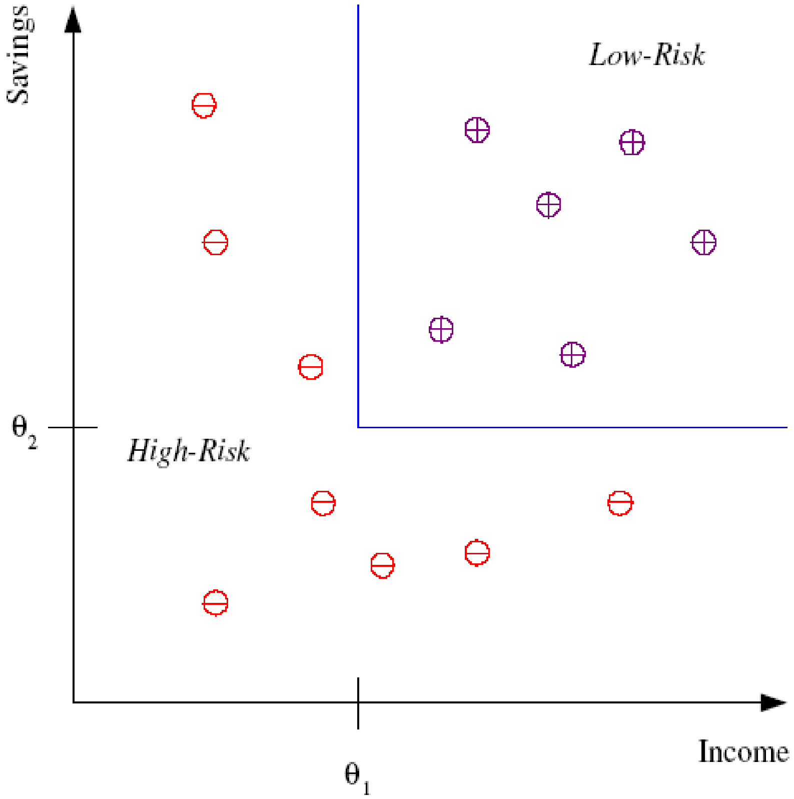
$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

Model selection: What should be M?



What is the problem here?

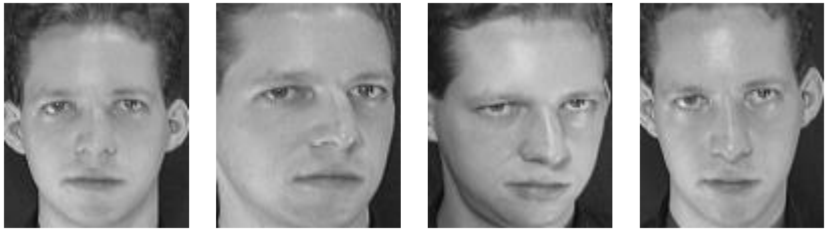
Classification Example



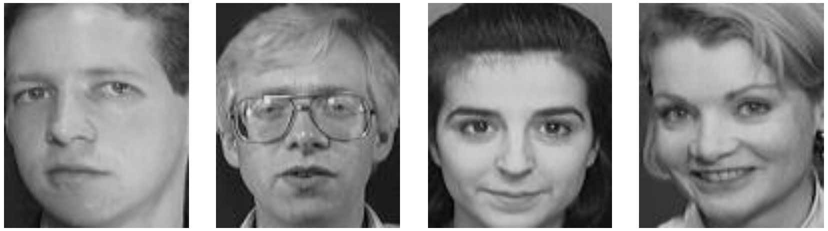
Rule: IF *income* > θ_1 AND *savings* > θ_2

THEN low-risk ELSE high-risk

Training:



Testing:



Source: http://www.uk.research.att.com/facedatabase.html

Regression Example

Example: Price of a used car

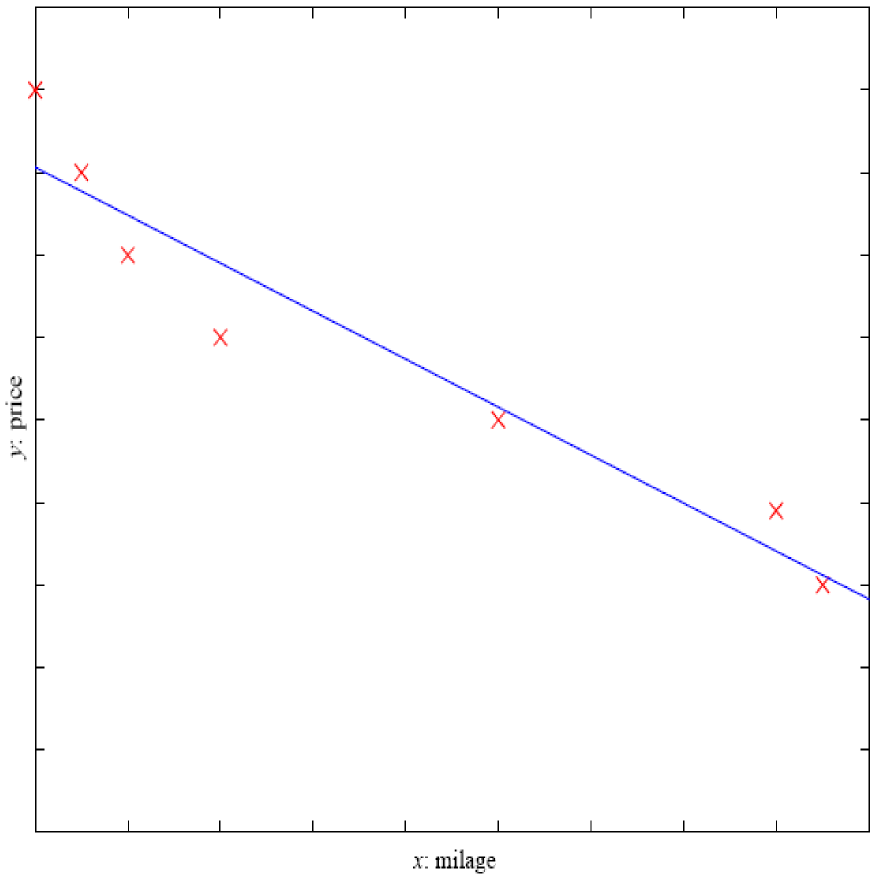
x : kilometres ran

y : price

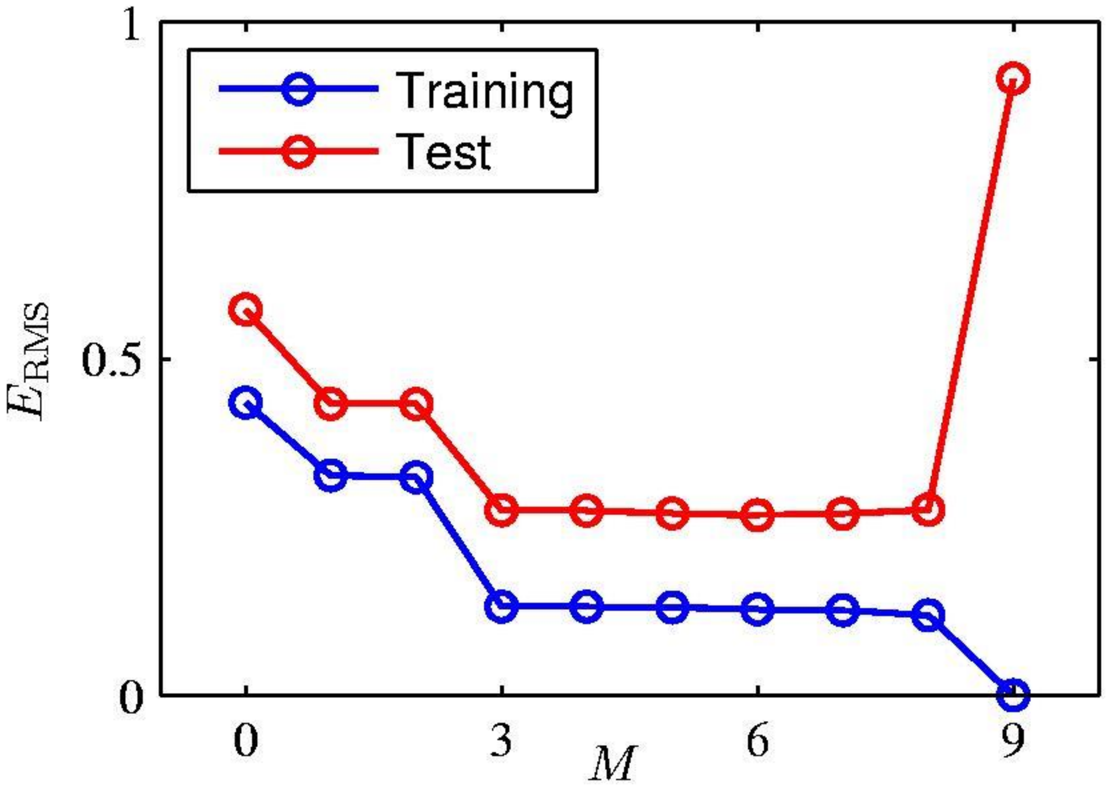
$y = g(x \mid \vartheta)$

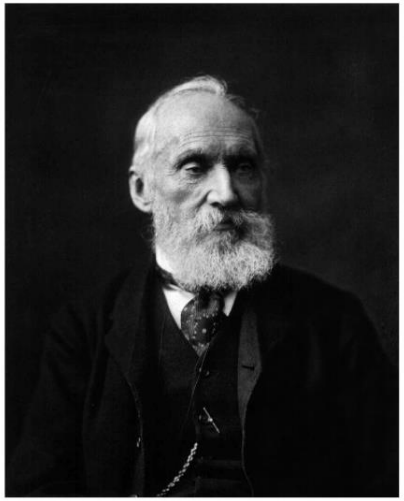
g ( ) model, and

ϑ are the parameters



Over-fitting



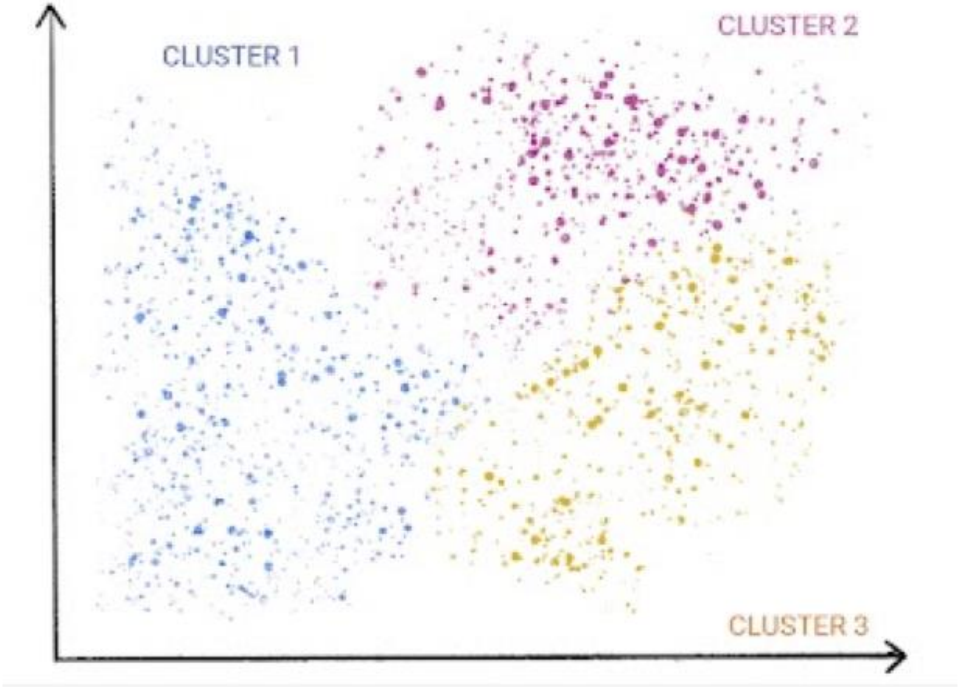


Solutions: later

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

Unsupervised Learning

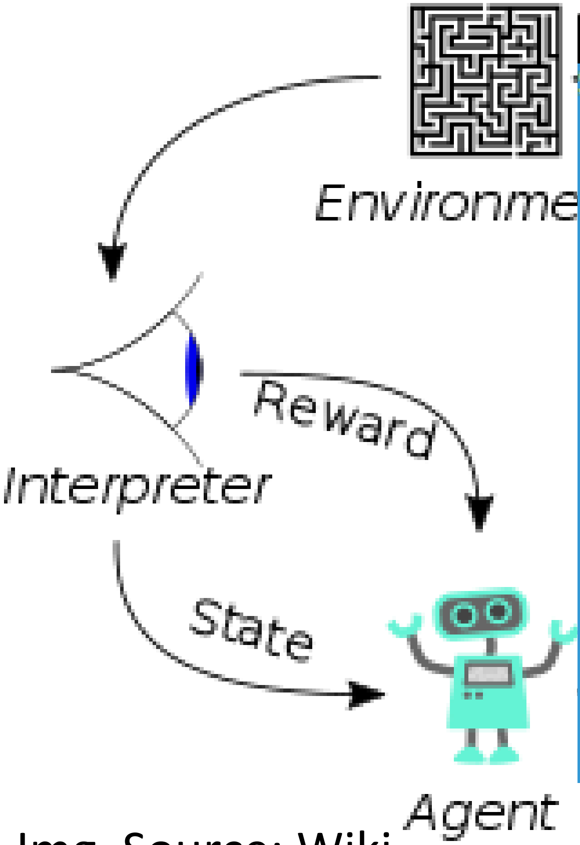
- Learns from data without human supervision.
- Using unlabeled data, these

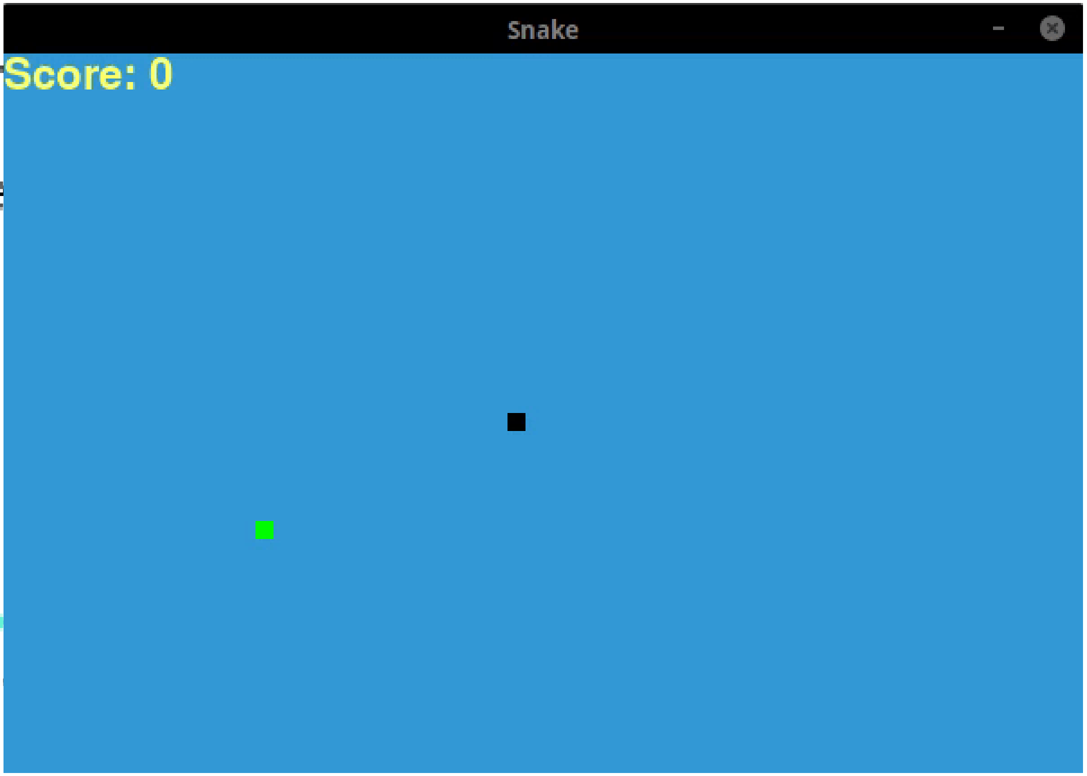


Img. Source: Quora

Reinforcement Learning

Learn action to maximize payoff

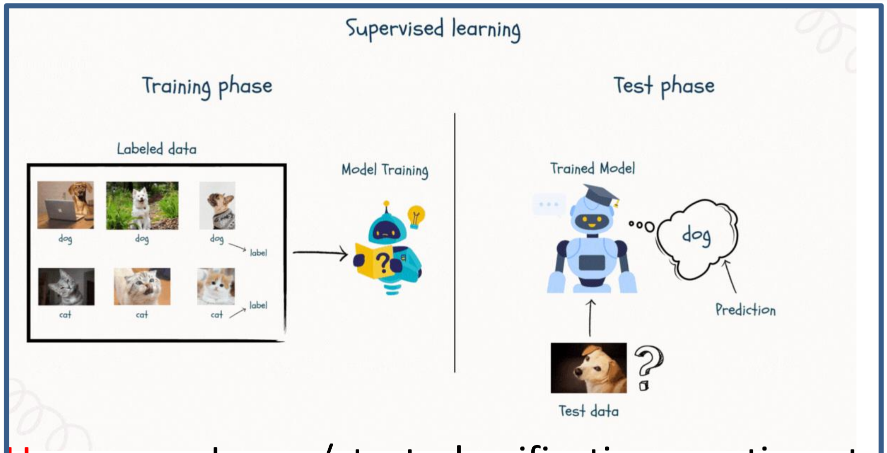




Img. Source: Wiki

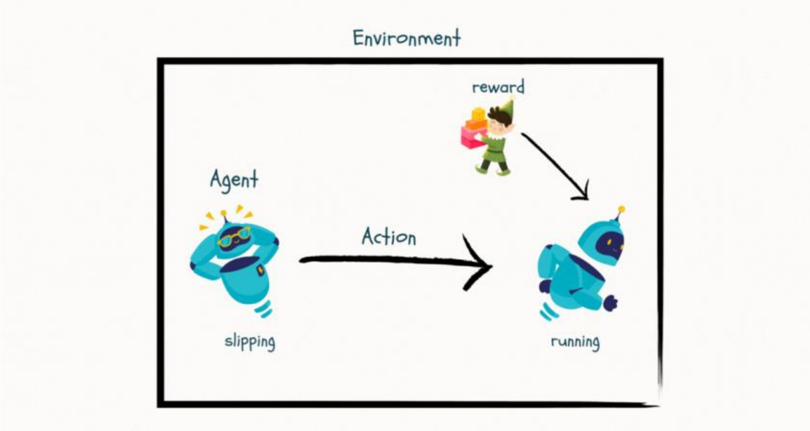
Source: https://towardsdatascience.com/

Recap: Types of Learning



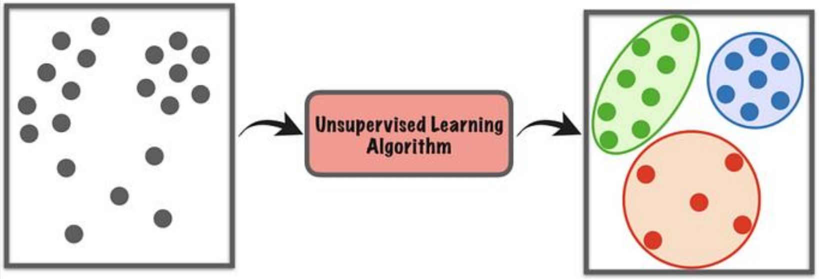
Use cases: Image/ text classification, sentiment analysis, medical diagnosis, weather/ stock price predictions etc.

Algorithms: Linear/ Logistic Regression, Decision Trees/ Random Forest, SVM, k-NN etc.



Use cases: Robotics, Autonomous vehicles, Industrial control etc.

Algorithms: Q-learning, DQN, etc.



Use cases: Clustering, Dimensionality reduction, Anomaly detection, Generative models etc.

Algorithms: K-Means, PCA etc.

What are some Issues in Machine Learning?

- What algorithms are available for learning a concept? How well do they perform?
- 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-III
- Wikipedia's ML Datasets:

  https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

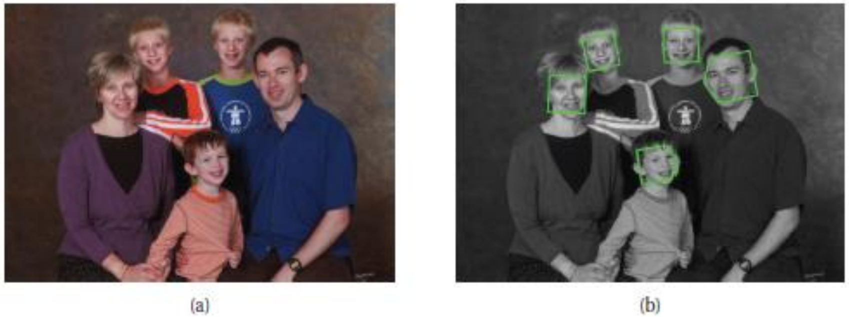
ML Journals and Conferences

- Journal of Machine Learning Research: https://www.jmlr.org/
- IEEE Transactions on Neural Networks.
- Neural Computation.
- Journal of Artificial Intelligence Research (JAIR).
- ACM Transactions on Intelligent Systems and Technology (TIST).
- ICML, ECML, NIPS, COLT, Int. Joint Conference on AI (IJCAI)

Quiz for you

Q.1 If fig.a is the input to your ML model and fig.b is the output with names of people in the photograph, then what type of problem is this?

- Classification ✓
- Regression



(Img. Source: Kevin Murphy)

Q.2 In the Cats and Dogs example that we discussed, what is being learnt by the model?

- Slope of the line and all the points (their coordinates) on the line.
- Slope of the line and the Intercept ✓

Q.3 Finding out who are the students in this class who play Cricket with an unlabelled dataset can be solved by using:

- Supervised Learning
- Un-supervised Learning ✓

Thank You!