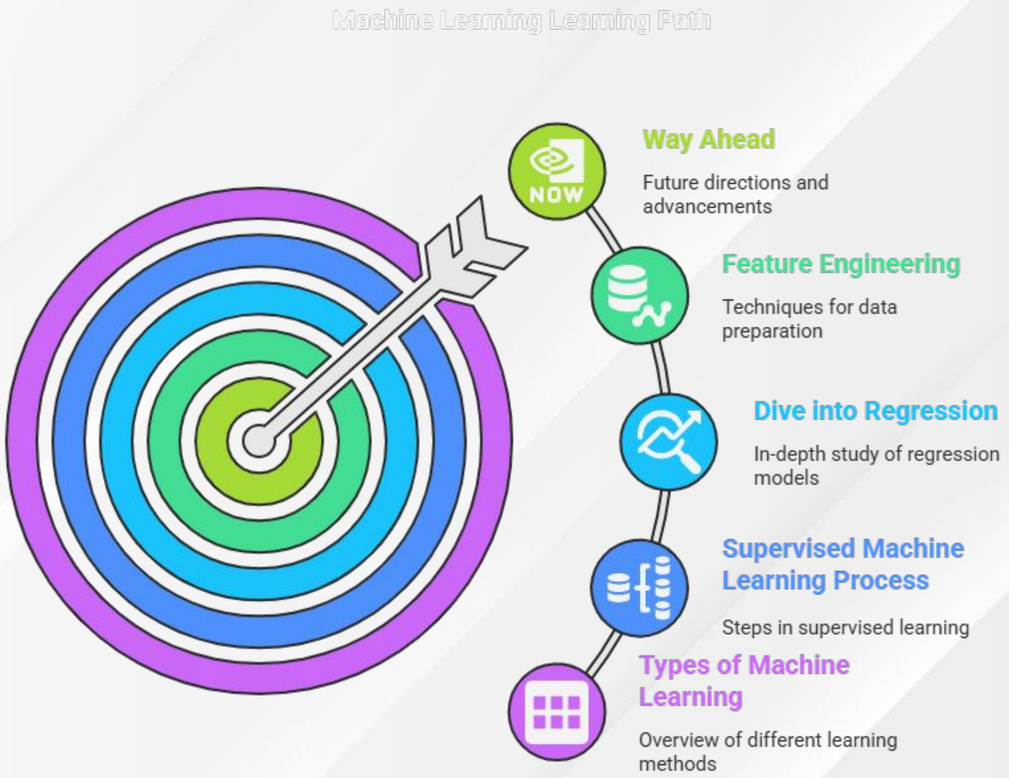
Lecture 9
Smart Data Discovery

Learning the Machine Learning

Outline

- Intro Machine Learning
- Applications of machine learning
- Types of Machine Learning
- Supervised machine learning process
- Dive into regression
- Feature Engineering
- Way ahead

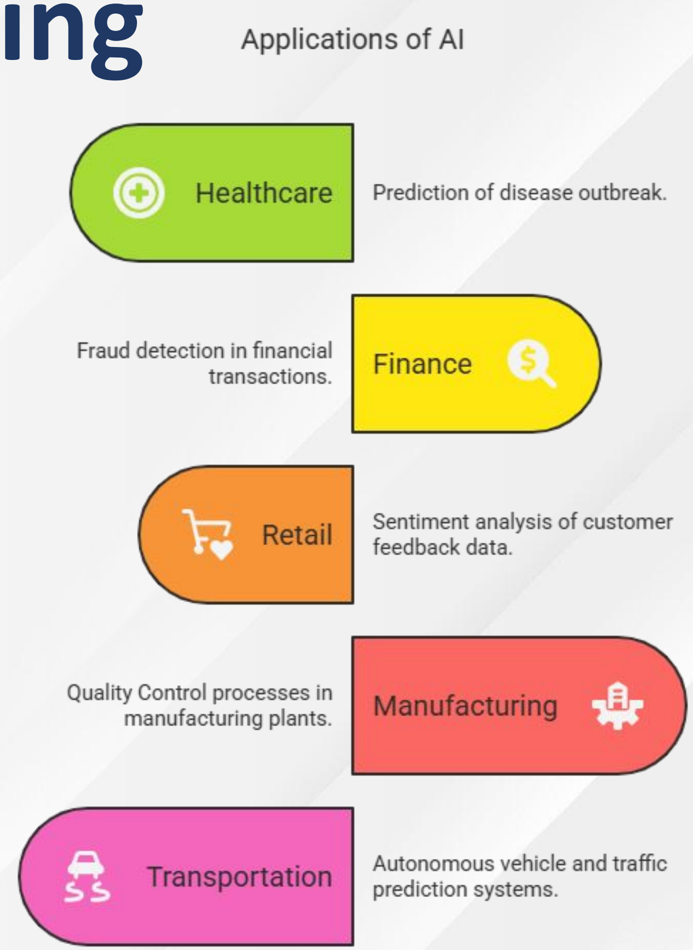


Machine Learning

- subfield of artificial intelligence
- capability of a machine to imitate intelligent human behavior
- gives computers the ability to learn without explicitly being programmed.
- In 1959, Arthur Samuel defined machine learning as “the field of study that gives computers the ability to learn without being explicitly programmed.”
- On the way towards 1990’s more sophisticated, neural networks
- Rise of big data

Applications of Machine Learning

- Healthcare
    - Eg: Prediction of disease outbreak
- Finance
    - Eg: Fraud detection
- Retail
    - Eg: sentiment analysis of customer
- Manufacturing
    - Quality Control
- Transportation
    - Autonomous vehicle, traffic prediction

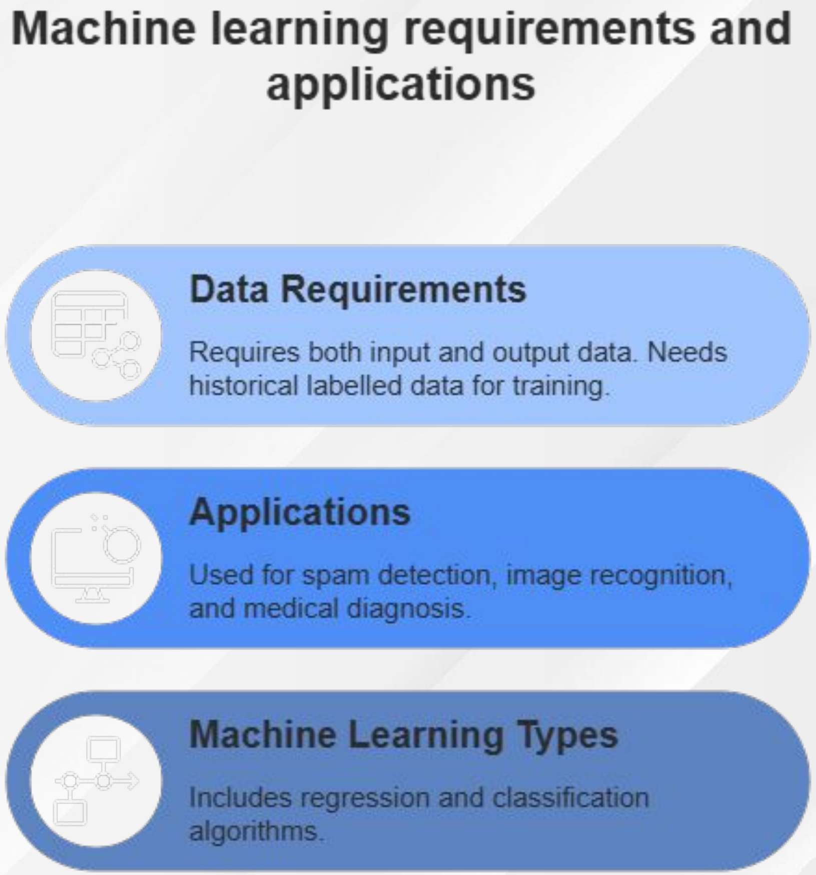


Types of machine learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Supervised Learning

- Input and output data are provided
- Requires historical labelled data
- Spam detection, image recognition, medical diagnosis
- Regression, classification



Supervised Learning

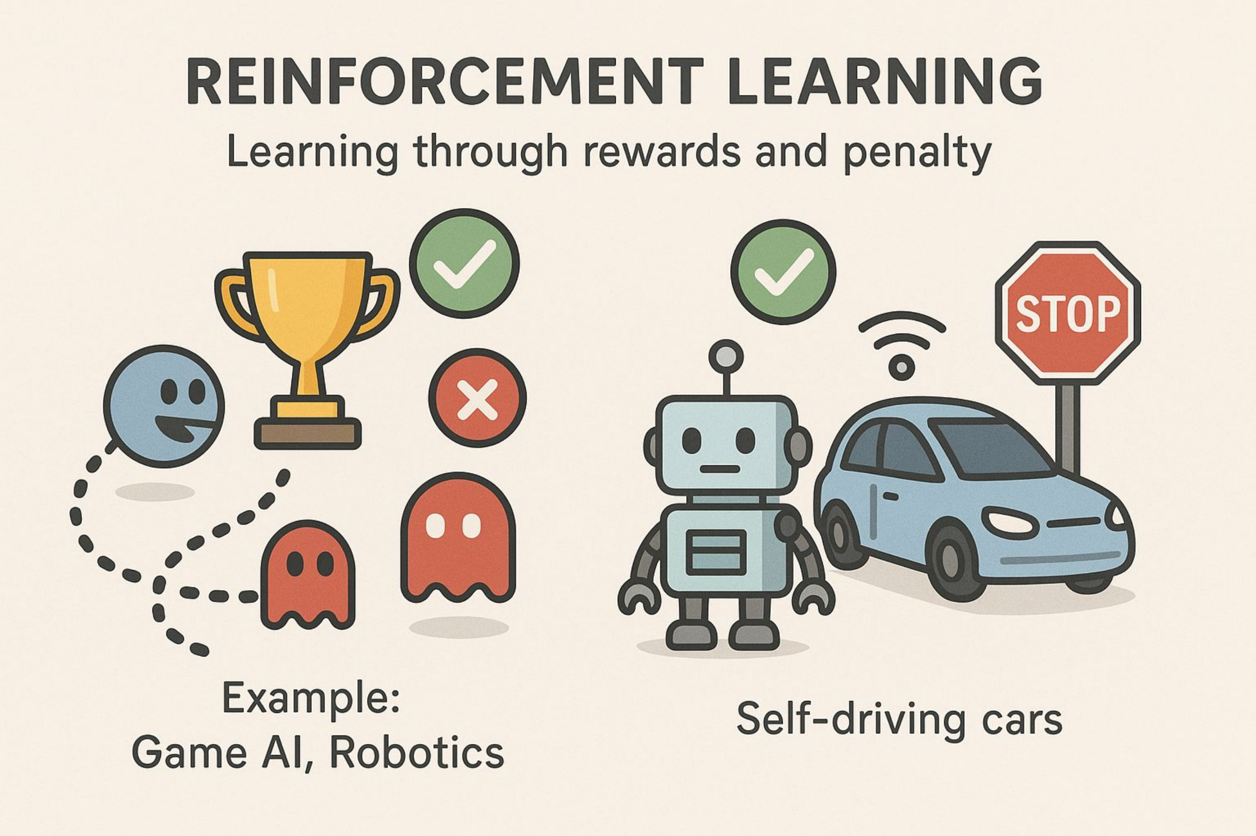
- Regression
    - Continuous value to predict
    - Pricing prediction of a house is a regression task
    - Test score prediction of student
- Classification
    - Categorical value to predict
    - Predict assigned category
    - Cancerous versus benign tumour
    - Handwriting Recognition

Unsupervised Learning

- Only input data is provided
- Example: Clustering, Dimensionality Reduction
- Customer segmentation, market basket analysis
- Group and interpret data without a level
- Clustering customers into separate groups based off their behaviours
- There is no historical correct label so it's harder to evaluate the performance of an unsupervised algorithm

Reinforcement Learning

- Learning through rewards and penalty
- Example: Game AI, Robotics
- Self driving cars

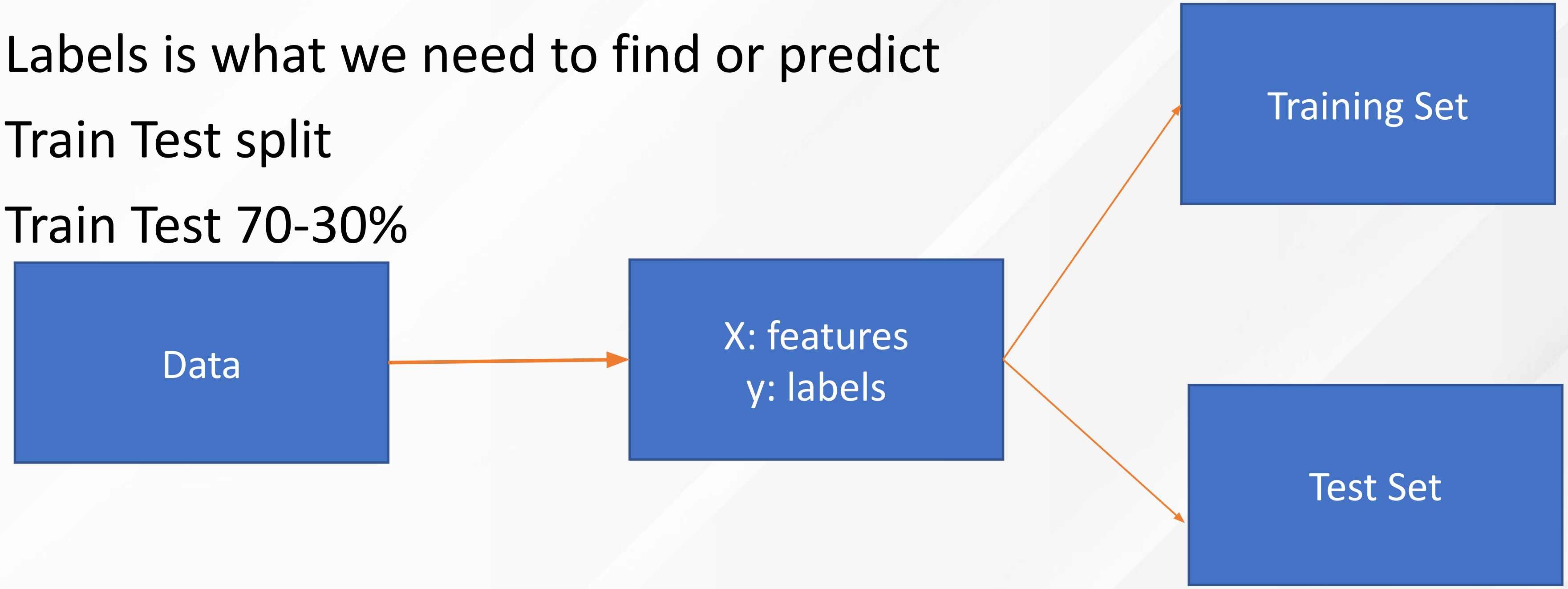


Supervised Machine Learning process

- Starts with collecting and organizing data set based on past or history
- For example detail about the price of the house along with various corresponding details
- Historical labelled data
- When the new house is on the market, looking at those parameters predict what should be the expected price
- There is input and output
- Using labelled data, predict the outcome

Supervised Machine Learning process contd…

- Separate data into features and labels
- Features are known characteristics
- Labels is what we need to find or predict
- Train Test split
- Train Test 70-30%

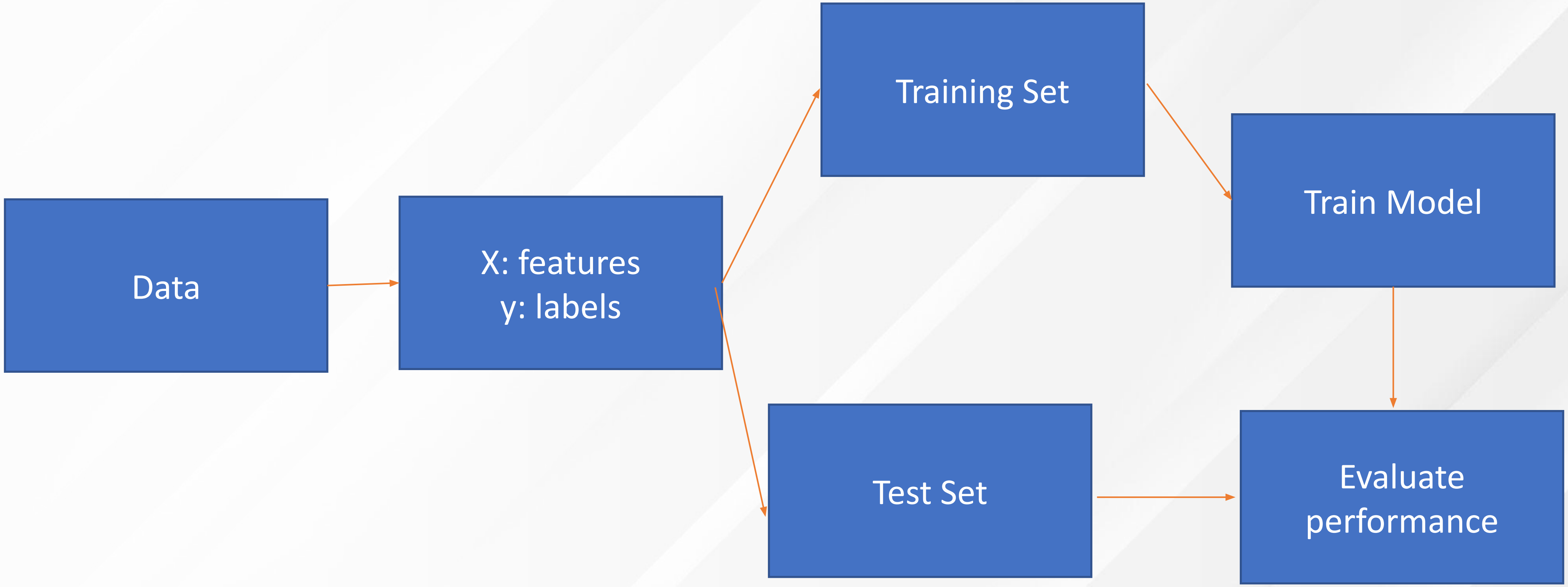


Supervised Machine Learning process

- 4 components

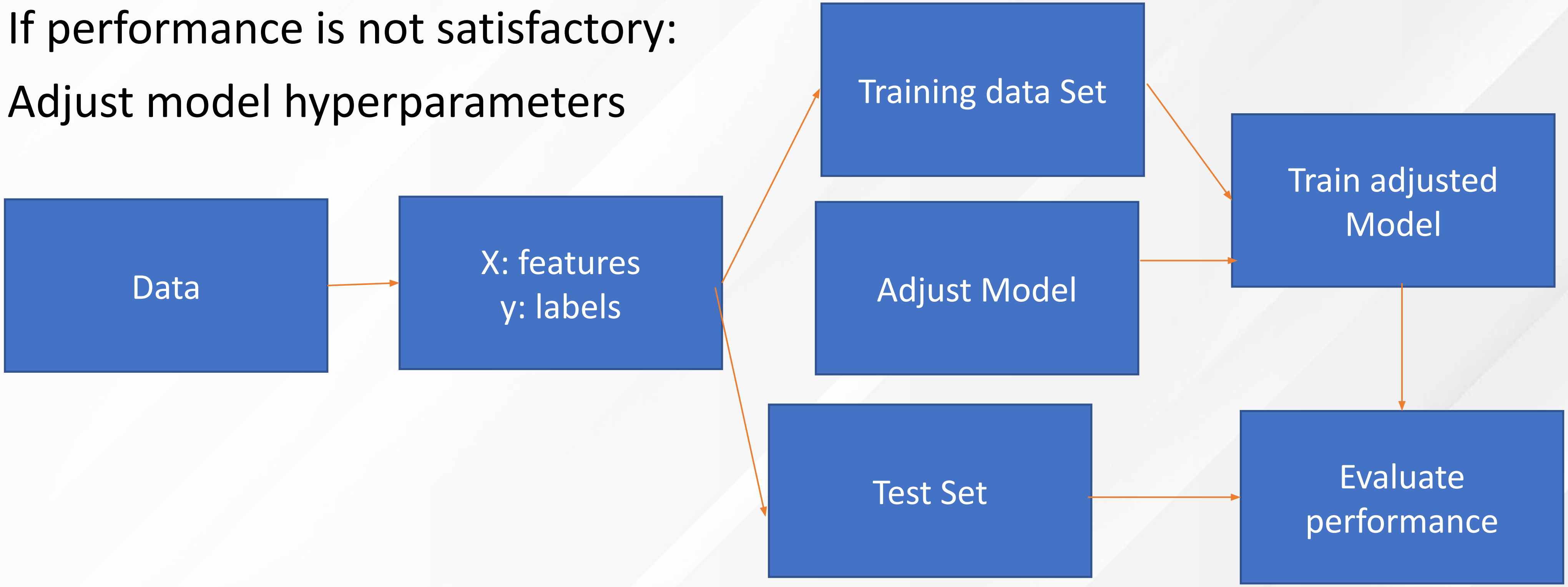
	Area	Room	Price	
X TRAIN	12	2	300	Y TRAIN
	15	2	450	
	10	4	425	
X TEST	20	2	600	Y TEST
	20	3	675	

Supervised Machine Learning process



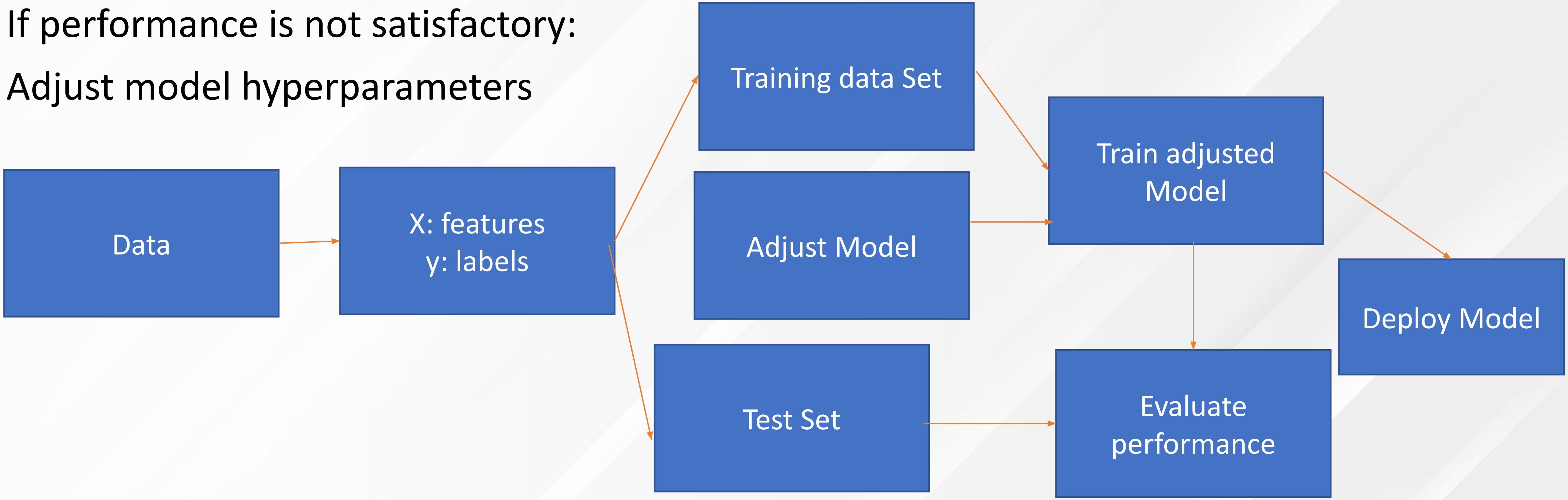
Supervised Machine Learning process

- If performance is not satisfactory:
- Adjust model hyperparameters



Supervised Machine Learning process

- If performance is not satisfactory:
- Adjust model hyperparameters



Regression

- used to predict continuous values
- simple and statistical method to understand and quantify the relationship between two variables or more
- Linear Regression
- supervised learning algorithm for predicting a continuous dependent variable based on one or more independent variables.

Regression contd.

- The goal is to find the best-fitting linear relationship between the input variables (X) and the output variable (Y).
- Mathematically represented as :
- $y = \beta_0 + \beta_1 x + \varepsilon$ for simple linear regression model
- $Y = b0 + b1X1 + b2X2 + \ldots + bnXn$ for multiple regression
- Y: output i.e. dependent variable
- X1,X2,Xn are input i.e. independent variable

Linear Regression

- Price of house based on area

Area	Price
1000	300000
1500	450000
2000	600000
2500	750000

Here

Simple linear regression model will be fitted as:

Price = b0+b1*size

Linear Regression contd

- Using linear regression algorithm
- b0 = 50000
- b1 = 250
- Now for house with 1800 area
- Price = 50000 + 250 *1800 = 50000 + 450000 = 500000

Linear Regression contd.

```
#import necessary libraries
```

```
# data set
np.random.seed(42)
data_size = 150
Feature = np.random.rand(data_size) * 10
Target = 3.5 * Feature + np.random.randn(data_size) * 2
```

```
# Create a DataFrame
df = pd.DataFrame({ 'Feature': Feature, 'Target': Target })
```

```
#Split the data into features (X) and target (y)
X = df[['Feature’]]
y = df['Target’]
```

Linear Regression contd

```
# Split the data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

```
# Create linear regression model
model = LinearRegression()
```

```
# Train the model
model.fit(X_train, y_train)
```

```
# Make predictions
y_pred = model.predict(X_test)
```

```
# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
```

Logistic Regression

- used for binary classification problems.
- It predicts the probability that a given input belongs to a particular category.
- Used in predictive modelling
- Whether an instance belongs to specific category or not
- Probability of heart attack, spam message, enrolling in some job
- uses a logistic function called a sigmoid function to map predictions and their probabilities.
- sigmoid function: S-shaped curve that converts any real value to a range between 0 and 1.

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Logistic Regression (Contd.)

- For Example
- If $\beta 0 = -2$, $\beta 1 = 0.5$, and $x = 3$:
- $\beta 0 + \beta 1 x = -2 + 0.5 \cdot 3 = -0.5$
- $e - (-0.5) = e0.5 \approx 1.648$
- $1 + e0.5 \approx 2.648$
- $P = 1/2.6481 \approx 0.378$

- So, the probability of the event is about 37.8%.
- This formula is the foundation of logistic regression, allowing us to predict probabilities for binary classification problems.

Logistic Regression

```
# import needed libraries
```

```
# Load Iris dataset
iris = datasets.load_iris()
X = iris.data
y = iris.target
```

```
# split into train and test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

```
#logistic regression model
model = LogisticRegression(max_iter=200)
```

Logistic Regression

```
#Train the model
model.fit(X_train, y_train)
```

```
# Make predictions
y_pred = model.predict(X_test)
```

```
Model Evaluation
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred, target_names=iris.target_names)
```

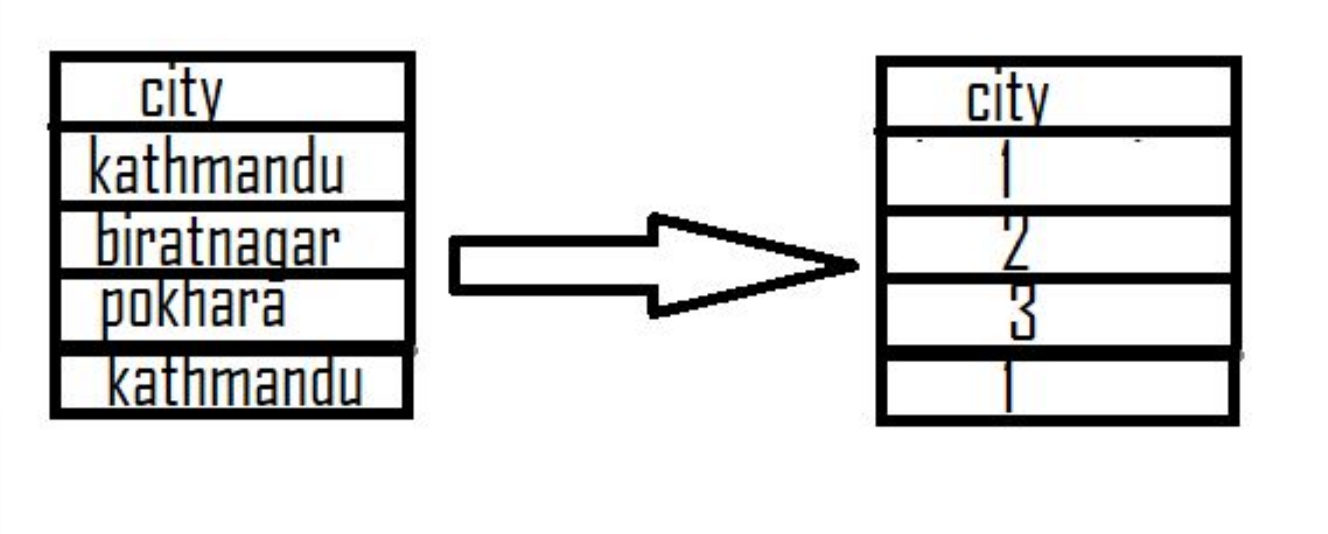
```
#display required informations
```

Feature Engineering

- Using domain knowledge for extracting features from raw data
- Approaches:
    - Extracting information
        - If the detail is like: 2020-10-9 09:11:13
        - Year 2020
        - Month 10 and so on
    - Combining information
        - Adding the marks of two terms
        - Adding the sales of various quarters
    - Transforming information
        - Most common for string data type
        - Can’t apply arithmetic operations on string
        - Encoding is done

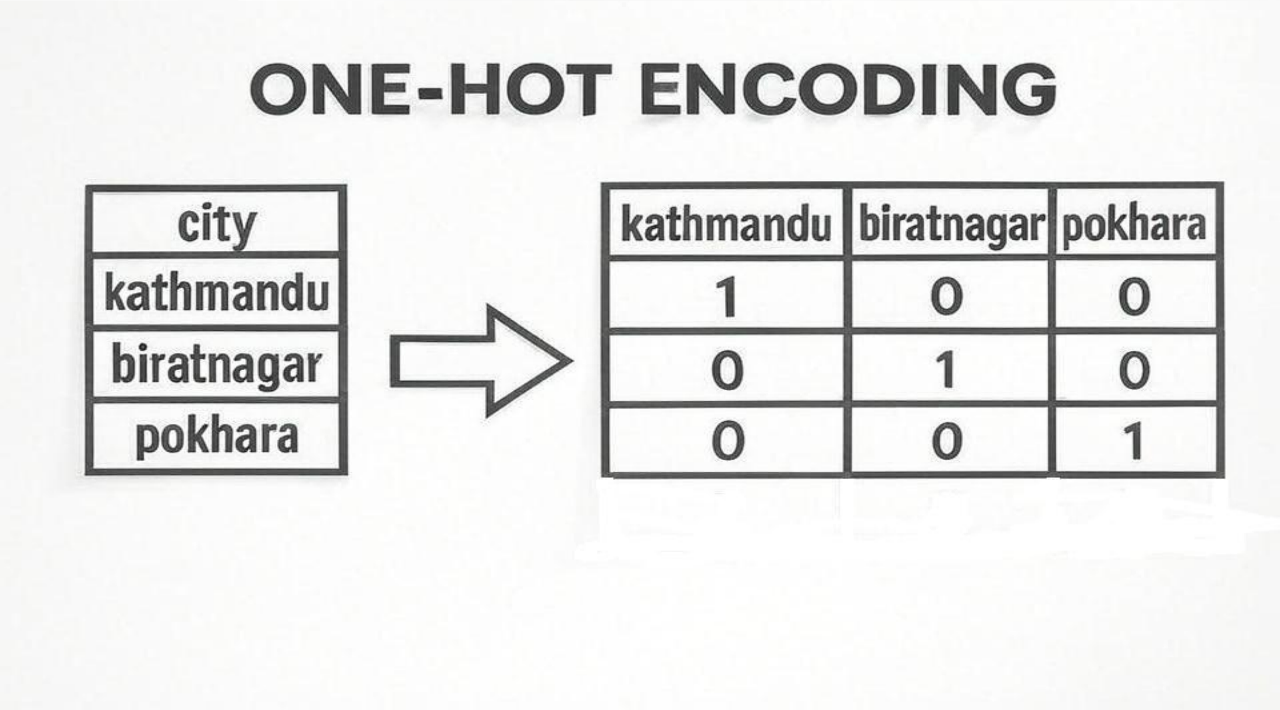
Transforming information

- Integer encoding
  - Converts categories into Integers 1,2,3,…,N



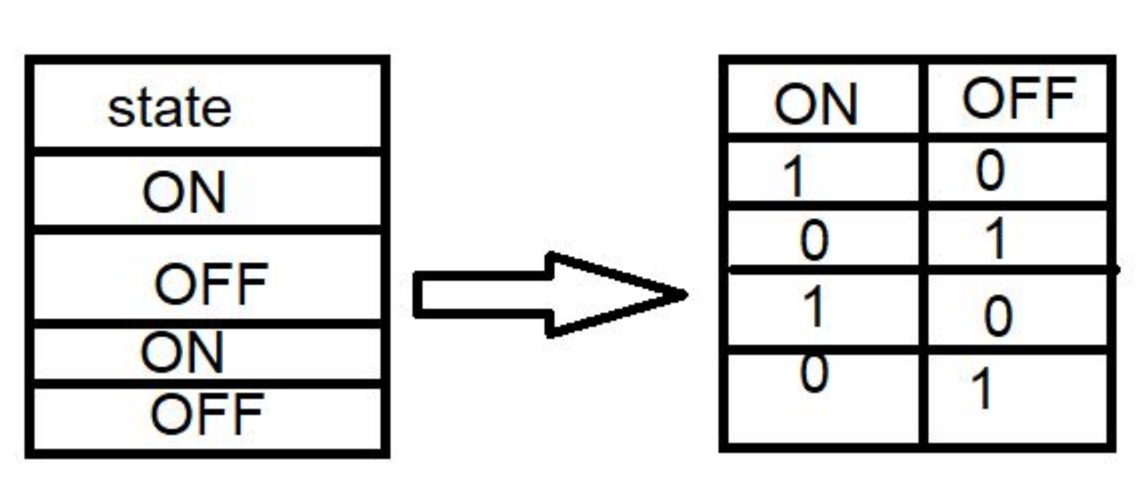
Transforming information

- One hot encoding (Dummy variables)
- Convert each category into individual features that are either 0 or 1



Transforming information

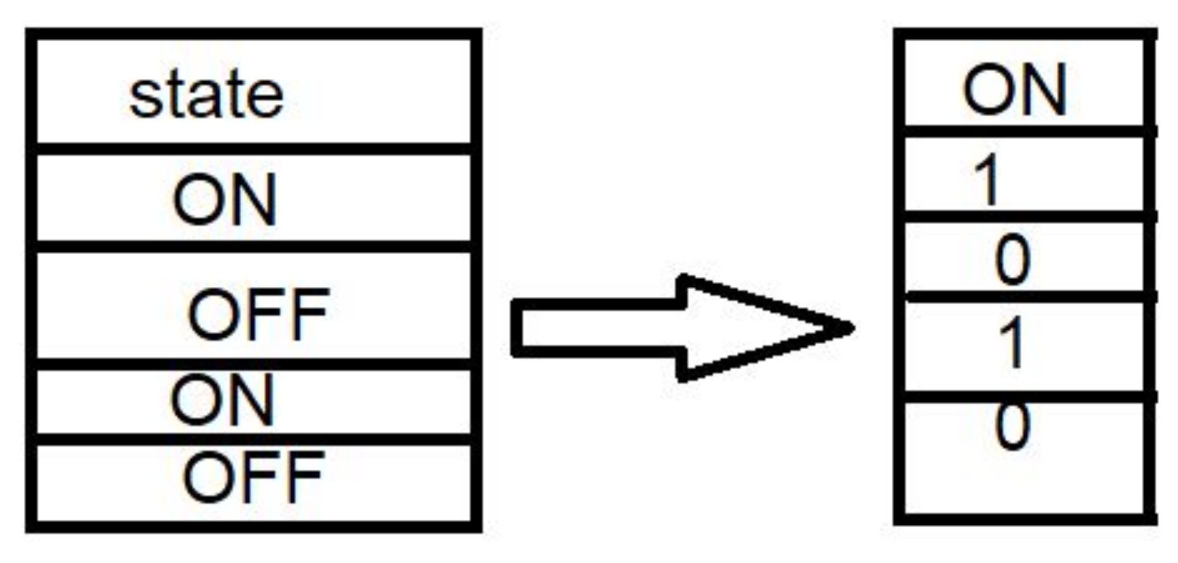
- One hot encoding (Dummy variables)
- Converting to dummy variables can cause features to be duplicated



- Columns could be dropped and only one can present the information well

Transforming information

- One hot encoding (Dummy variables)



Transforming information

ONE-HOT ENCODING

city
kathmandu
biratnagar
pokhara

kathmandu	biratnagar	pokhara
1	0	0
0	1	0
0	0	1

city
kathmandu
biratnagar
pokhara
kathmandu

kathmandu	biratnagar
1	0
0	1
0	0
1	0

Any Questions?