NWiMLDS
Explore Series II

Data Science & Machine Learning

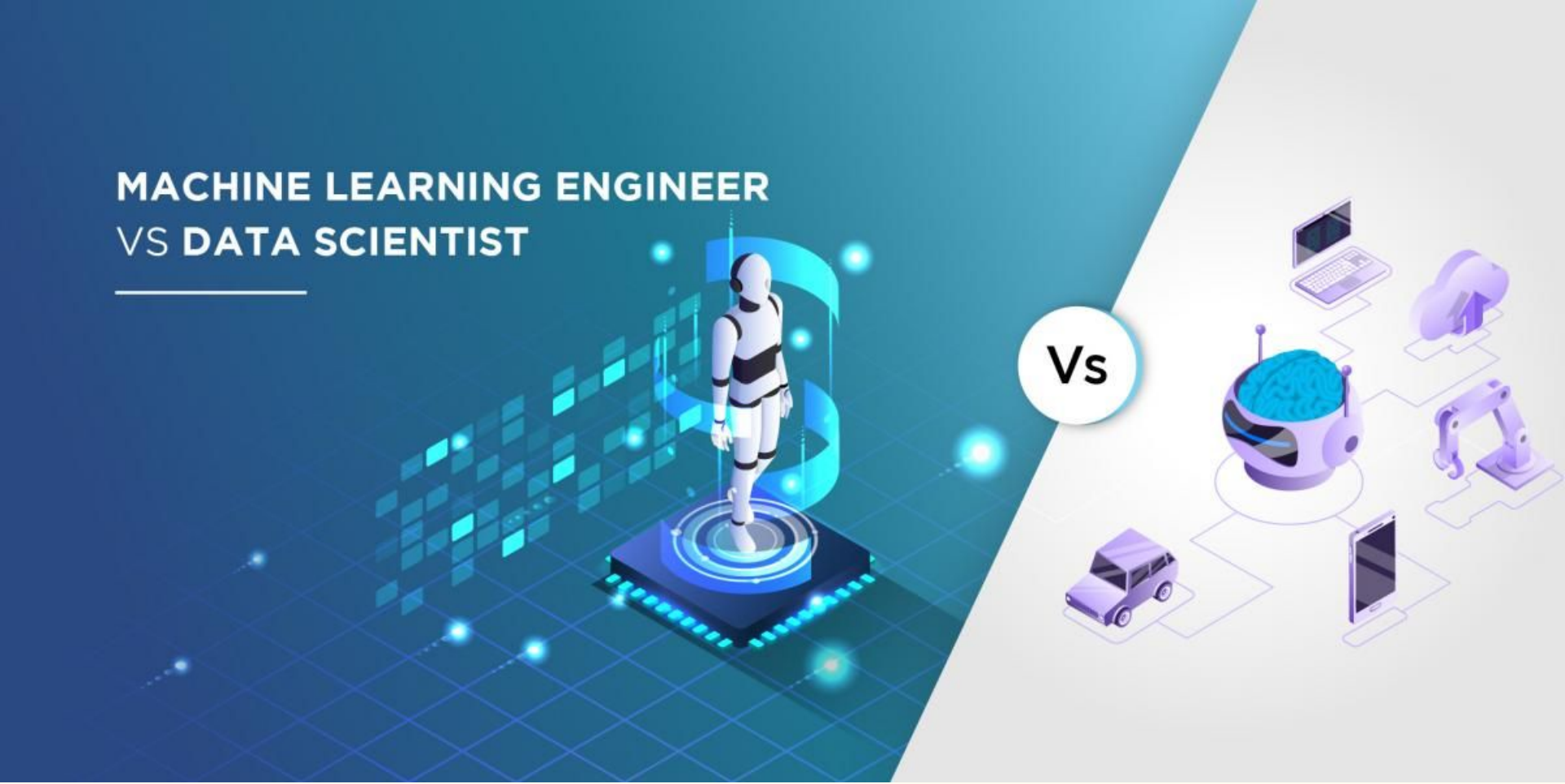
—

Lydia Gathoni
Maureen Mathenge

Objectives

- Data Scientists vs ML Engineer
- Python Mini Crash Course
- CRISP-DM / ML Workflow
- ML Algorithms
  - Supervised
  - Unsupervised
- Hands On Lab

MACHINE LEARNING ENGINEER VS DATA SCIENTIST



Python Mini Crash Course

Python Mini Crash Course

- Flow Control.
- Data Structures.
- Functions.
- Access Data

Flow Control

These are codes under different conditions

The basic principle for flow control is, if some conditions happens do sth else otherwise do sth else . 3 main types of flow control

1) If....else
2) For-loop
3) While-loop

If **statement**;

```
x = 20

if x > 0:

  print("Positive Number")
```

If...else;

```
x = 20
if x > 0:
  print("Positive Number")
else:
  print("Negative Number")
```

If...elif...else;

```
x = -20
if x > 0:
  print("Positive Number")
elif x == 0:
  print("zero")
else:
  print("Negative Number")
```

Nested if statements

We can have if...else statement inside another if...else statement

```
grade = 67
if grade >= 65:
    print("Passing grade of:")

    if grade >= 90:
        print("A")

    elif grade >=80:
        print("B")

    elif grade >=70:
        print("C")

    elif grade >= 65:
        print("D")

else:
    print("Failing grade")
```

Flow Control

For-loop

A for loop is used to iterate a sequence. It can be used with other conditional statements like if,continue,break,else

Nested loop

```
[54] prof = ["Doc", "Chef", "Data Scientist"]
     names = ["Jane", "John", "Fatuma"]

     for x in prof:
       for y in names:
         print(x, y)
```

```
Doc Jane
Doc John
Doc Fatuma
Chef Jane
Chef John
Chef Fatuma
Data Scientist Jane
Data Scientist John
Data Scientist Fatuma
```

While-loop

It keeps executing codes as long as it is true. It does not return if it is false

```
i = 1
while i < 6:
  print(i)
  i += 1
```

If you remove `i += 1` you shall get a infinite loop

Data Structures

List

A list is a way to give a single name to a collection of values. It is ordered and changeable and written in square brackets

Example

```
weight = [44,65,78,54,45]
```

```
[44, 65, 78, 54, 45]
```

Adding to the list

```
weight.append(98)
```

```
[44, 65, 78, 54, 45, 98]
```

Introducing names

```
['Maureen', 44, 'John', 65, 'Pooh', 78, 'Sam', 54]
```

Sublist for individual representation

```
weight = [["Maureen",44],
          ["John",65],
          ["Pooh",78],
          ["Sam",54]]
```

```
[['Maureen', 44], ['John', 65], ['Pooh', 78], ['Sam', 54]]
```

Tuple

A tuple is a collection which is ordered and unchangeable.
They are written with round brackets

Tuples are faster than lists

```
x = ("Maureen", "John", "Sam")
```

You can convert a tuple to a list

```
y = list(x)
```

Dictionary

It is a collection which is unordered, changeable and indexed.
They are written with curly brackets, and they have keys and values.

```
d = {'a':3, 'b':7, 'c':8}
```

Values can take any form of data type and can be duplicated unlike keys which are immutable

Functions

It is a group of related statements that performs a specific task.

Access Data

We can view data from a dataframe using 2 methods:

1) Square Brackets
2) Advanced Methods : loc and iloc

Square brackets

```
df = pd.DataFrame({'EmployeeID':
np.random.randint(1, 135, 20),

                   'skill':np.random.randint(1,6,
20),

'age':np.random.randint(22,40, 20),

'year':np.random.randint(2000,2020, 20)})
```

	EmployeeID	skill	age	year
0	42	2	33	2002
1	125	1	30	2016
2	114	2	22	2000
3	42	2	23	2016
4	101	3	26	2002

```
df[["EmployeeID"]].head()
```

	EmployeeID
0	42
1	125
2	114
3	42
4	101

Advanced Methods

loc and iloc - They are used to access rows, columns and both rows and columns at the same time.
Difference
loc - label-based
Iloc - integer position based

loc

```
df.loc[0:3,"EmployeeID":"year"]
```

	EmployeeID	skill	age	year
0	42	2	33	2002
1	125	1	30	2016
2	114	2	22	2000
3	42	2	23	2016

iloc

```
df.iloc[0:3,0:3]
```

	EmployeeID	skill	age
0	42	2	33
1	125	1	30
2	114	2	22

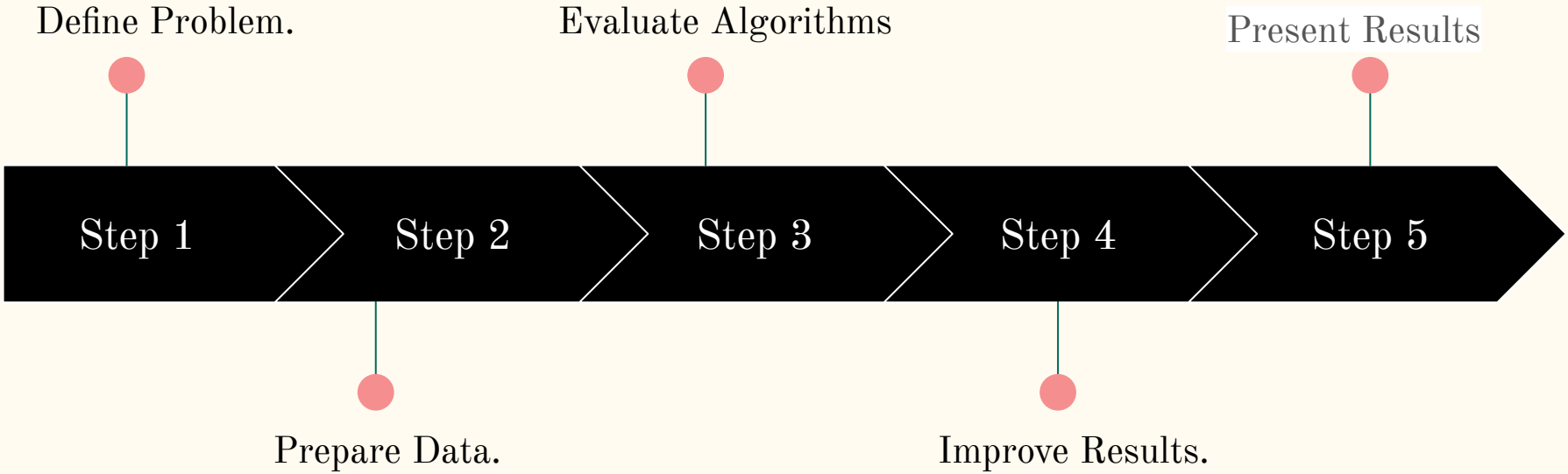
CRISP-DM / ML Workflows

Recap: CRISP-DM

CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is an industry-proven way to guide your data mining efforts.

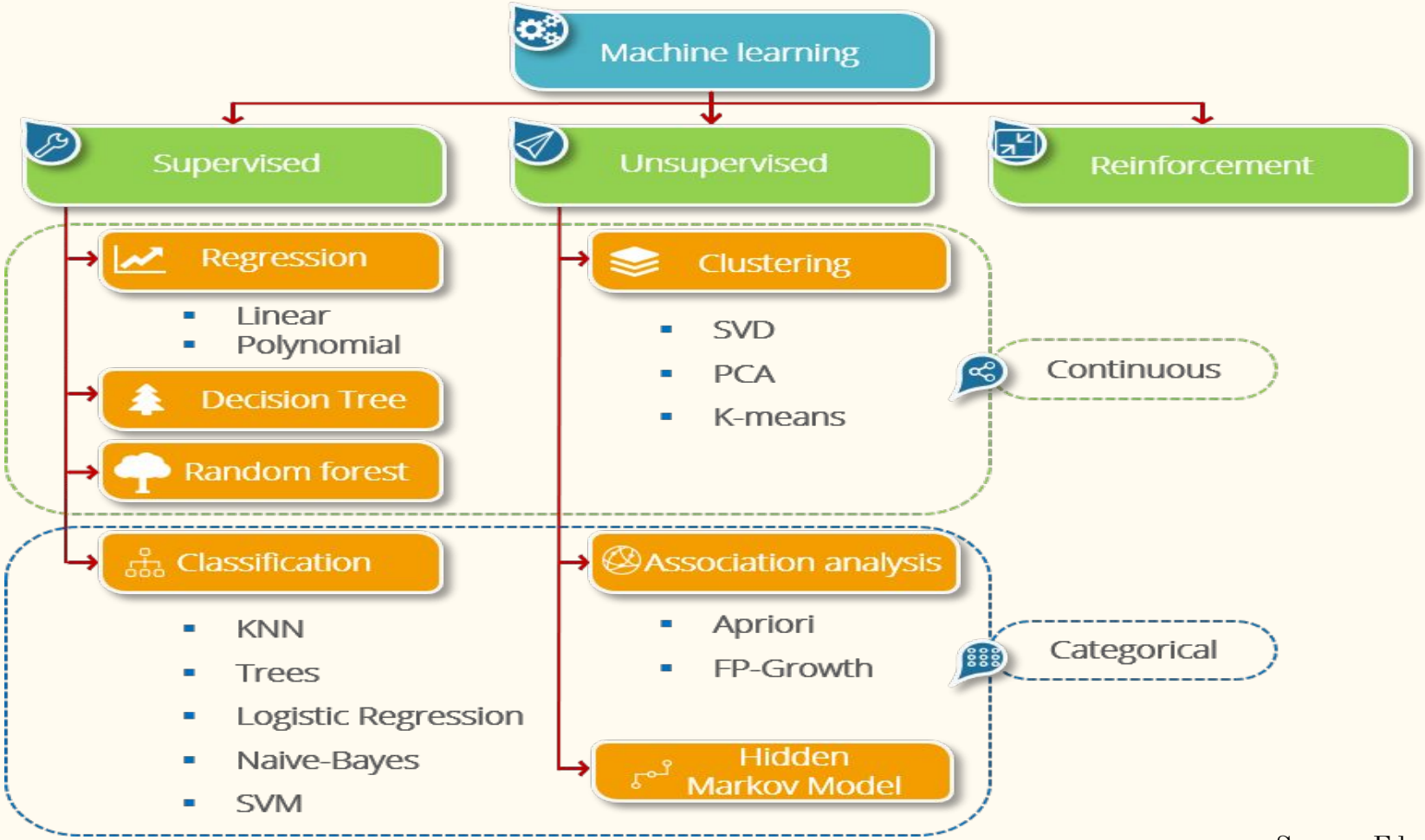
- ➔ Business Understanding
- ➔ Data Understanding
- ➔ Data Preparation
- ➔ Modelling
- ➔ Evaluation
- ➔ Deployment

ML Workflow


Define Problem.
Evaluate Algorithms
Present Results
Step 1
Step 2
Step 3
Step 4
Step 5
Prepare Data.
Improve Results.

ML Algorithms

	Supervised Learning	Unsupervised Learning
Definition	The machine learns by using labelled data	The machine is trained on unlabelled data without any guidance
Type of problems	Regression & Classification	Association & Clustering
Type of data	Labelled data	Unlabelled data
Training	External supervision	No supervision
Approach	Map labelled input to known output	Understand patterns and discover output
Popular algorithms	Linear regression, Logistic regression, Support Vector Machine, KNN, etc	K-means, C-means, etc



Source: Edureka

Recap of Supervised ML Algorithms

- ➔ Linear Regression
- ➔ Decision Tree
- ➔ Random Forest
- ➔ Logistic Regression
- ➔ KNN
- ➔ Naive Bayes
- ➔ Support Vector Machine

Unsupervised ML Algorithms

Unsupervised ML algorithms learn hidden patterns within the data structure. The data has no labels and no prior training has been done on the dataset.

Approaches of Unsupervised learning

- ➔ Clustering
- ➔ Association
- ➔ Anomaly Detection
- ➔ Dimensionality Reduction

Clustering

A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

Types of Clustering

- ➔ K-means clustering
- ➔ Hierarchical clustering
- ➔ DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering
- ➔ Gaussian Clustering Model

References

- https://machinelearningmastery.com/
- https://towardsdatascience.com/
- Machine Learning Refined - Jeremy Watts
- https://heartbeat.fritz.ai/
- Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow - O'Reilley

Hands-On Practice...