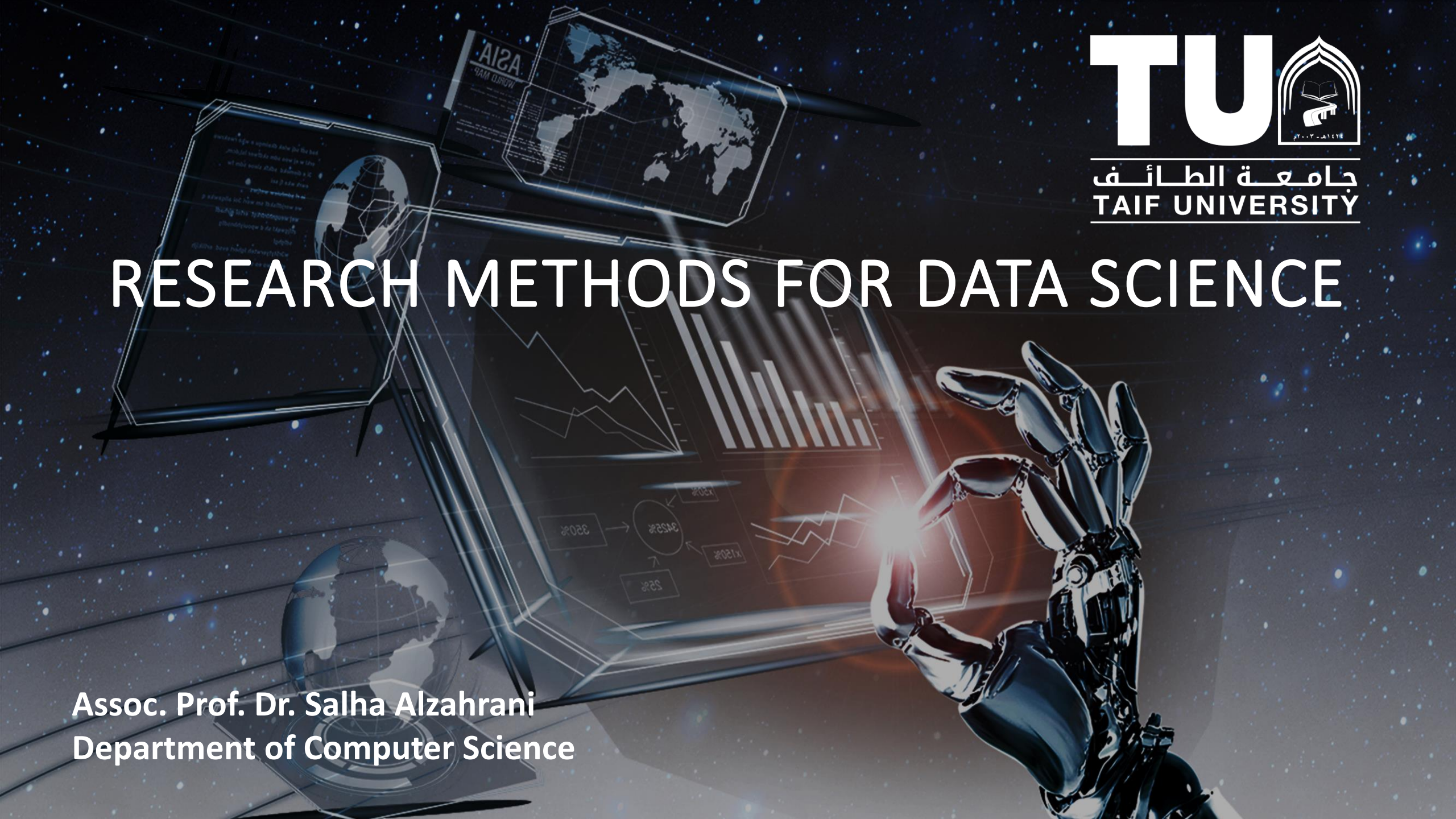


RESEARCH METHODS FOR DATA SCIENCE

Assoc. Prof. Dr. Salha Alzahrani
Department of Computer Science

Step 4: Selecting a sample
Step 6: Collecting data

My Notes

~Sampling Techniques

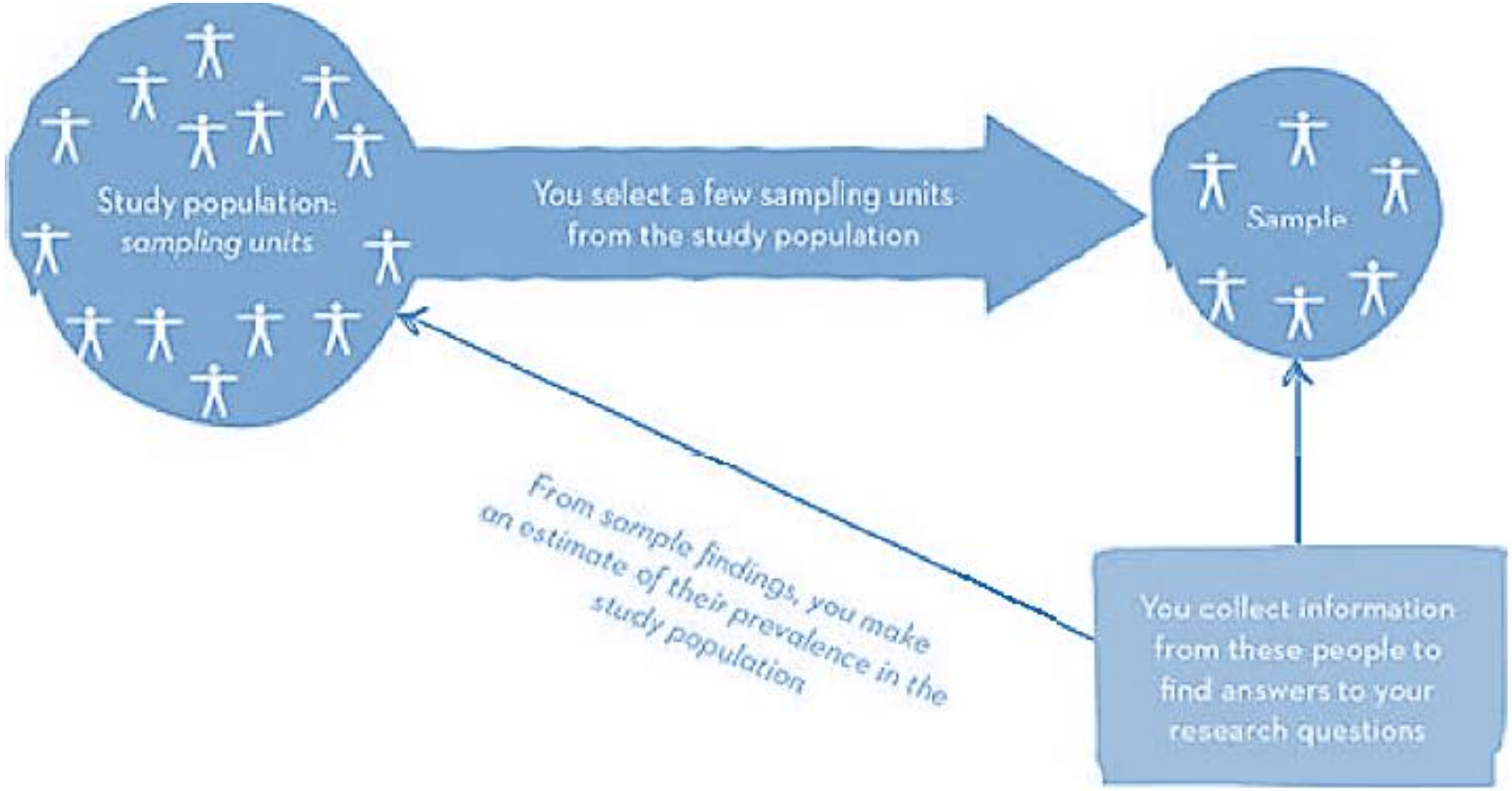
~Python for loading and sampling data

The concept of sampling

Let us take a very simple example to explain the concept of sampling. Suppose you are interested in the mean age of the students in your class. There are two ways of finding this out. The first method is to contact all students in the class, find out their ages, add them up and then divide this by the number of students (the procedure for calculating sample mean). The second method is to select a few students from the class, ask them their ages, add them up and then divide by the number of students you have asked. From this you make an *estimate* of the average age of the class. Similarly, suppose you want to find out the average income of families living in a city. Imagine the amount of effort and resources required to go to every family in the city to find out their income! You could instead select a few families as the basis of your enquiry and then, from what you have found out from the few families, make an estimate of the mean income of families in the city. A similar procedure is used in opinion polls. These are also based upon a very small group of people who are questioned about (say) their voting preferences and, on the basis of these results, a *prediction* is made about the probable outcome of an election.

My Notes

The concept of sampling



My Notes

The concept of sampling

Sampling: The process of selecting a few respondents (a sample) from a bigger group (the sampling population) to become the basis for estimating the prevalence of information of interest to you.

Sampling population: The entire group, such as families living in an area, clients of an agency, residents of a community, members of a group, people belonging to an organisation about whom you want to find out about through your research endeavour, is called the sampling population or study population.

This process of selecting a sample from the total population has advantages and disadvantages. The advantages are that it saves time as well as financial and human resources. However, the disadvantage is that you do not obtain information about the population's characteristics of interest to you but only *estimate* or *predict* them on the basis of what you found out in your sample. Hence, there is the possibility of an error in your estimation.

My Notes

What is a dataset?

A data set (or dataset) is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question. The data set lists values for each of the variables, such as for example height and weight of an object, for each member of the data set. Data sets can also consist of a collection of documents or files.

https://en.wikipedia.org/wiki/Data_set

My Notes

Classic datasets

Several classic data sets have been used extensively in the statistical literature:

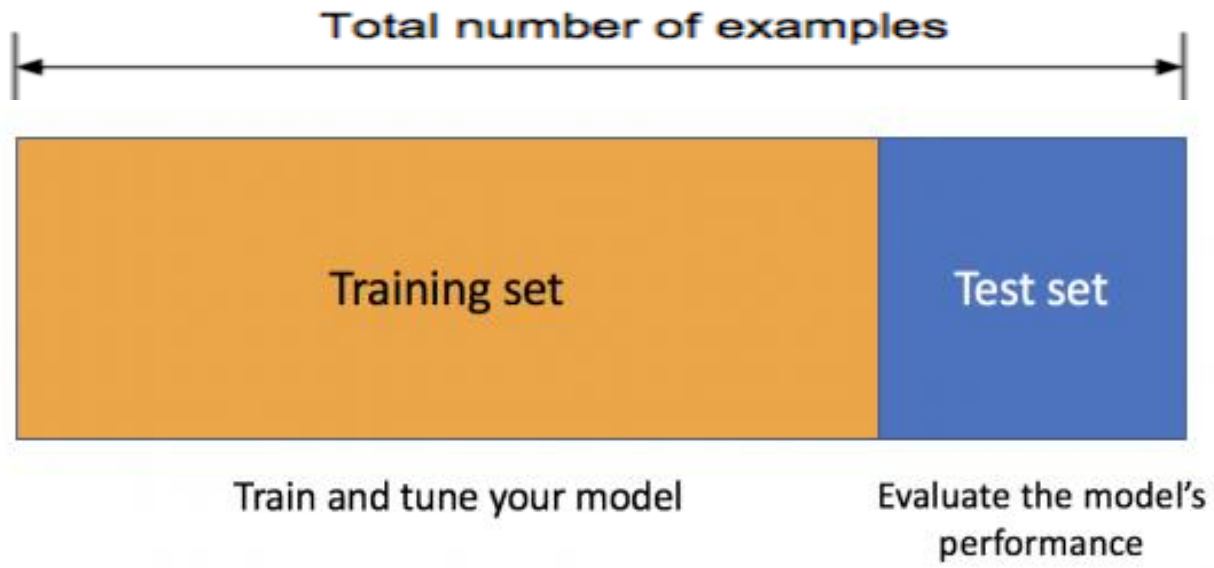
- **Iris flower data set** – Multivariate data set introduced by Ronald Fisher (1936).
- **MNIST database** – Images of handwritten digits commonly used to test classification, clustering, and image processing algorithms
- **Categorical data analysis** –used in the book, An Introduction to Categorical Data Analysis.
- **Robust statistics** – Data sets used in Robust Regression and Outlier Detection (Rousseeuw and Leroy, 1986). Provided on-line at the University of Cologne.
- **Time series** – Data used in Chatfield's book, by StatLib.
- **Extreme values** – Data used in the book, An Introduction to the Statistical Modeling of Extreme
- **Bayesian Data Analysis** – Data used in the book are provided on-line by Andrew Gelman.
- **The Bupa liver data** – Used in several papers in the machine learning (data mining) literature.

My Notes

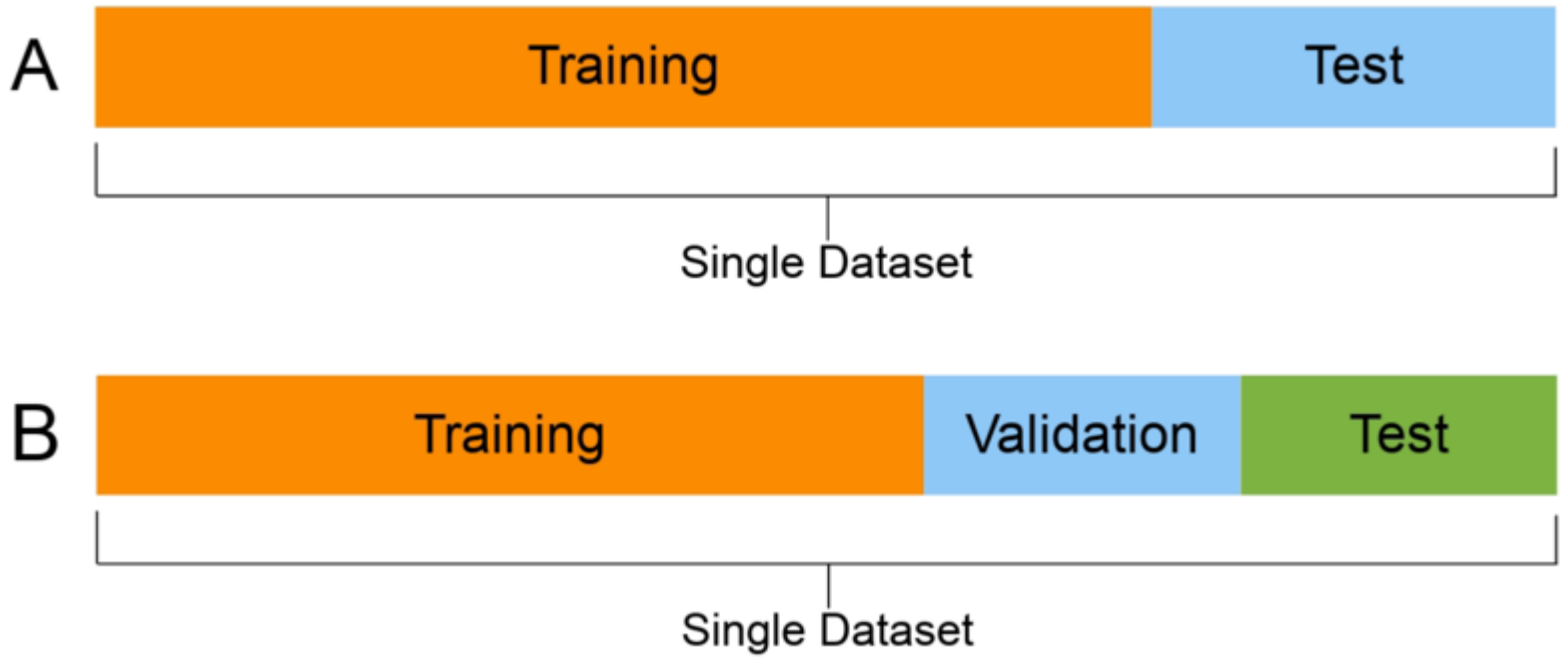
What are the training and test datasets?

My Notes

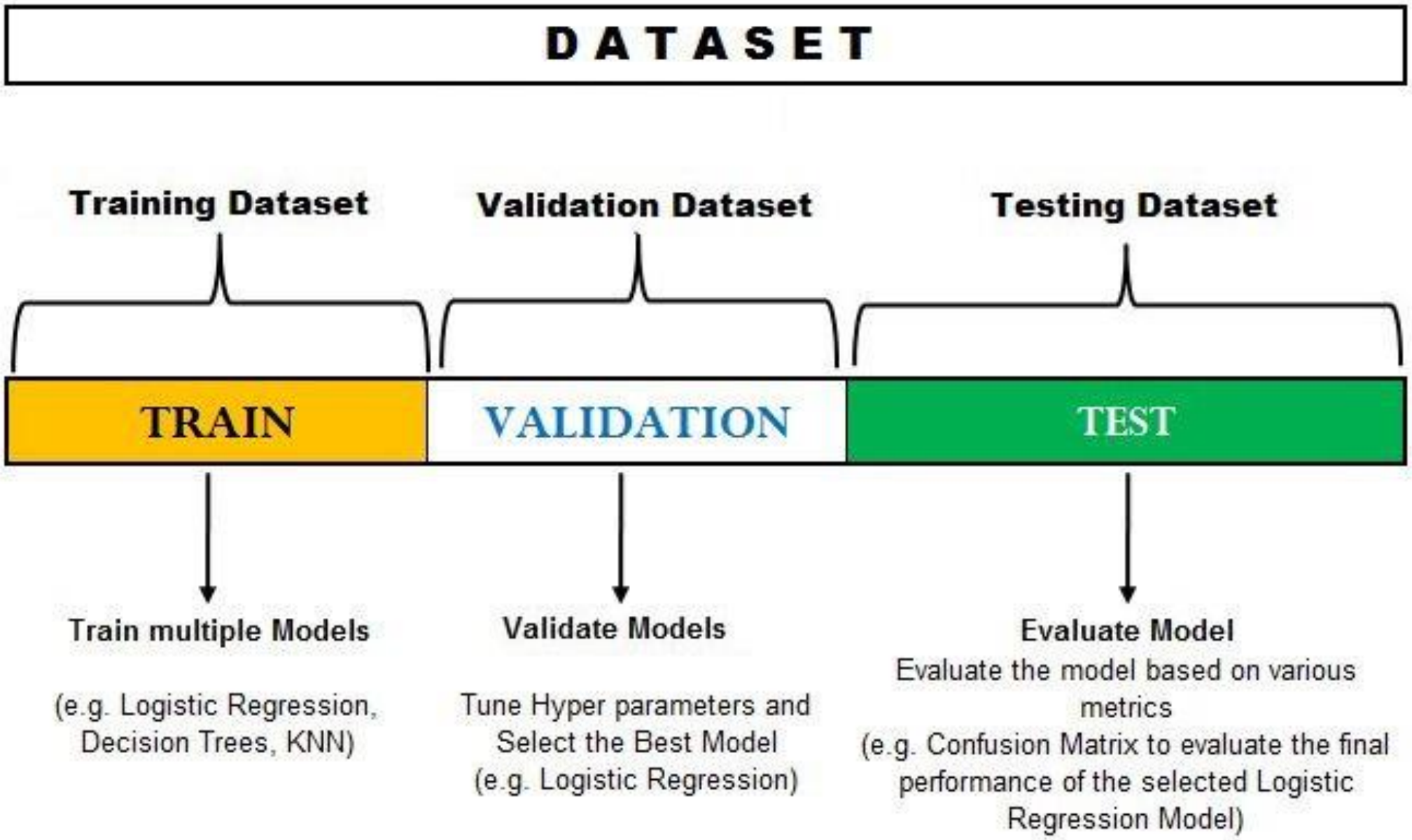
- A test set has to be strictly independent from the training examples.



What are the training, validation and test datasets?

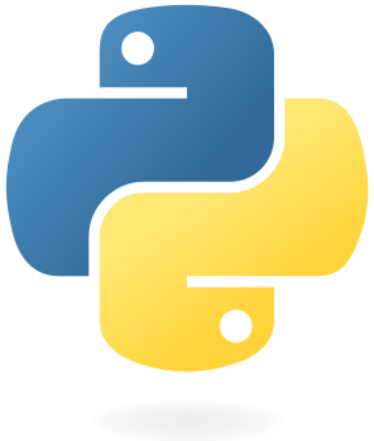


My Notes


DATASET
Training Dataset
Validation Dataset
Testing Dataset
TRAIN
VALIDATION
TEST
Train multiple Models
(e.g. Logistic Regression, Decision Trees, KNN)
Validate Models
Tune Hyper parameters and Select the Best Model
(e.g. Logistic Regression)
Evaluate Model
Evaluate the model based on various metrics
(e.g. Confusion Matrix to evaluate the final performance of the selected Logistic Regression Model)

Loading and Sampling

in



python™

seaborn

pandas

Loading datasets in python

Seaborn

The brilliant plotting package `seaborn` has several built-in sample data sets.

```
import seaborn as sns

iris = sns.load_dataset('iris')
iris.head()
```

```
   sepal_length  sepal_width  petal_length  petal_width species
0           5.1          3.5           1.4          0.2  setosa
1           4.9          3.0           1.4          0.2  setosa
2           4.7          3.2           1.3          0.2  setosa
3           4.6          3.1           1.5          0.2  setosa
4           5.0          3.6           1.4          0.2  setosa
```

Loading datasets in python

Pandas

If you do not want to import `seaborn`, but still want to access its sample data sets, you can use @andrewwowens's approach for the seaborn sample data:

```
iris = pd.read_csv('https://raw.githubusercontent.com/mwaskom/seaborn-data/master/iris.csv'
```

Loading datasets in python

scikit-learn

`scikit-learn` returns sample data as numpy arrays rather than a pandas data frame.

```
from sklearn.datasets import load_iris

iris = load_iris()
# `iris.data` holds the numerical values
# `iris.feature_names` holds the numerical column names
# `iris.target` holds the categorical (species) values (as ints)
# `iris.target_names` holds the unique categorical names
```

Sampling in python

pandas.DataFrame.sample¶

```
DataFrame.sample(n=None, frac=None, replace=False, weights=None, random_state=None,
axis=None, ignore_index=False)
```

Return a random sample of items from an axis of object.

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.sample.html

Sampling in python

It may happen that you need only some rows of your Python dataframe. You can achieve this result, through **different sampling techniques**.

The following techniques to perform rows sampling through **Python Pandas**:

- random sampling
- sampling with condition
- sampling at a constant rate

https://github.com/alod83/data-science/blob/master/Preprocessing/DataSampling/Data%20Sampling.ipynb

Data Sampling

```
from sklearn.datasets import load_iris
import pandas as pd

data = load_iris()
df = pd.DataFrame(data.data, columns=data.feature_names)
df.head(5)
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2

```
df.shape
```

```
(150, 4)
```

Random Sampling

Knowning the exact number of samples to return

```
subset = df.sample(n=100)
subset.shape
```

```
(100, 4)
```

Knowning the percentage of samples to return

```
subset = df.sample(frac=0.5)
subset.shape
```

```
(75, 4)
```

Sampling with condition

Return 10 random sample where sepal width (cm) < 3 Firstly count the number of records which satisfy the condition

```
condition = df['sepal width (cm)'] < 3
condition
```

```
0      False
1      False
2      False
3      False
4      False
       ...
145    False
146     True
147    False
148    False
149    False
Name: sepal width (cm), Length: 150, dtype: bool
```

```
true_index = condition[condition == True].index
len(true_index)
```

```
57
```

Since the number of elements satisfying the condition is 57, we can sample at maximum 57 elements

```
subset = df[condition].sample(n = 10)
subset.shape
```

```
(10, 4)
```

Sampling at a Constant Rate

Sampling every 10 elements

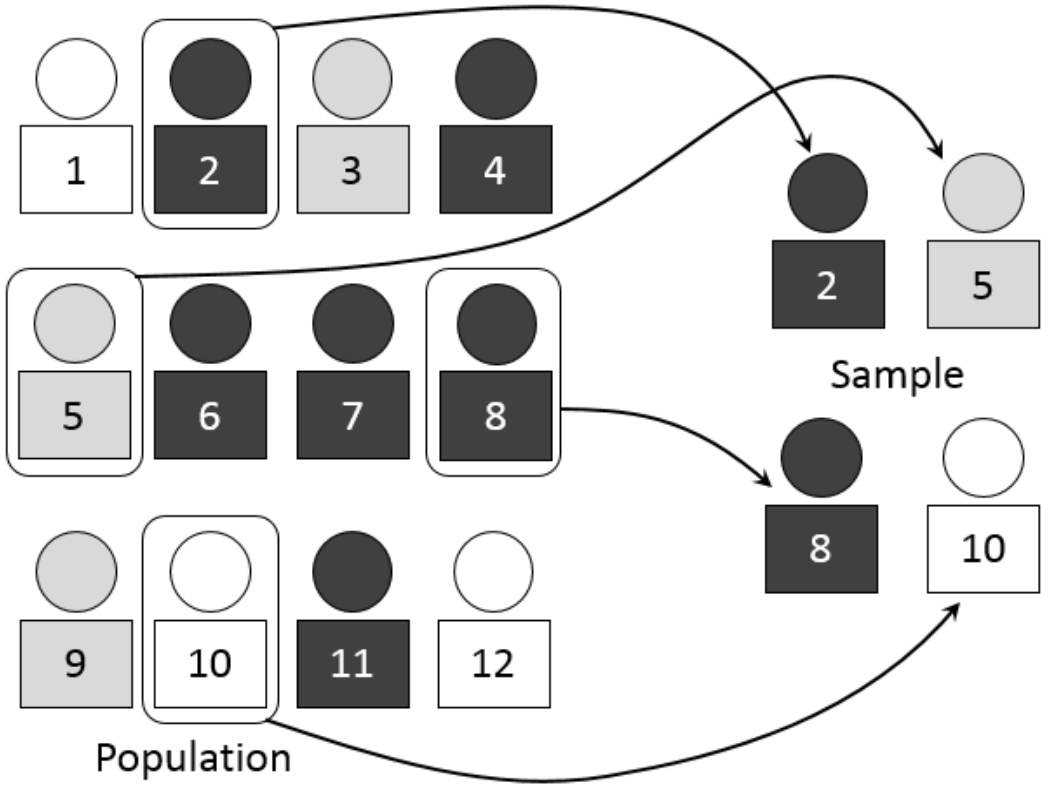
```
rate = 10
subset = df[::rate]
subset.shape
```

```
(15, 4)
```

```
subset.head()
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
10	5.4	3.7	1.5	0.2
20	5.4	3.4	1.7	0.2
30	4.8	3.1	1.6	0.2
40	5.0	3.5	1.3	0.3

Further Reading:
A visual representation of the sampling process



https://en.wikipedia.org/wiki/Sampling_(statistics)

WIKIPEDIA
The Free Encyclopedia

If you focus on success, you'll have stress. But if you pursue excellence, success will be guaranteed.

Deepak Chopra

quotefancy

@SalhaAlzahrani