Data Analysis with Python Course

Course Info

Course Info

Course Duration

26 November- 17 December

14 In-class Sessions ***42 Hours in Total***

3 Lab Sessions ***3 Hours in Total***

2 Capstone Solution Sessions ***06 Hours in Total*** (At the end of DA & DV with Python)

Structure of Course

Intro to DAwPython → **Numpy** → **Pandas**

Course Projects

2 Assignments

(Covering all course subjects)

1 Case/Self Study Project

EDA (Analyzing US Citizens)

1 Capstone Project

EDA (AutoScout Car Price Prediction)

CLARUSWAY

Lesson Plan

Data Analysis with Python

This course will give you comprehensive and valuable information about Numpy and Pandas libraries and by doing hands-on exercises you will learn how to use Python to analyze data. At the end of the course, you will have the intuition to prepare the data for any Decision Support Purposes or any Machine learning algorithm.

Custodian :	Matthew Connor-Instructor (matthew_c@clarusway.com)
In-class Sessions :	14 Sessions (42 hours)
Lab Sessions :	3/4 Labs (3/4 hours)

Certification Requirements:

1. % 70 attendance to in-class lessons (at least 10/14 for DAwPython Course)
2. Successfully completing and submitting assignments & projects (at least 1 assignment & 1 project for DAwPy Course)

Click the preview above to see the detailed syllabus of this section.

Attendance Reminder APP 1:51 AM

@channel 📣 Please login to zoom with your LMS email addresses and make sure your zoom names are like **X####-Xxxx (F1234-Iamhere)**.

Have a nice class!

Data Analysis with Python

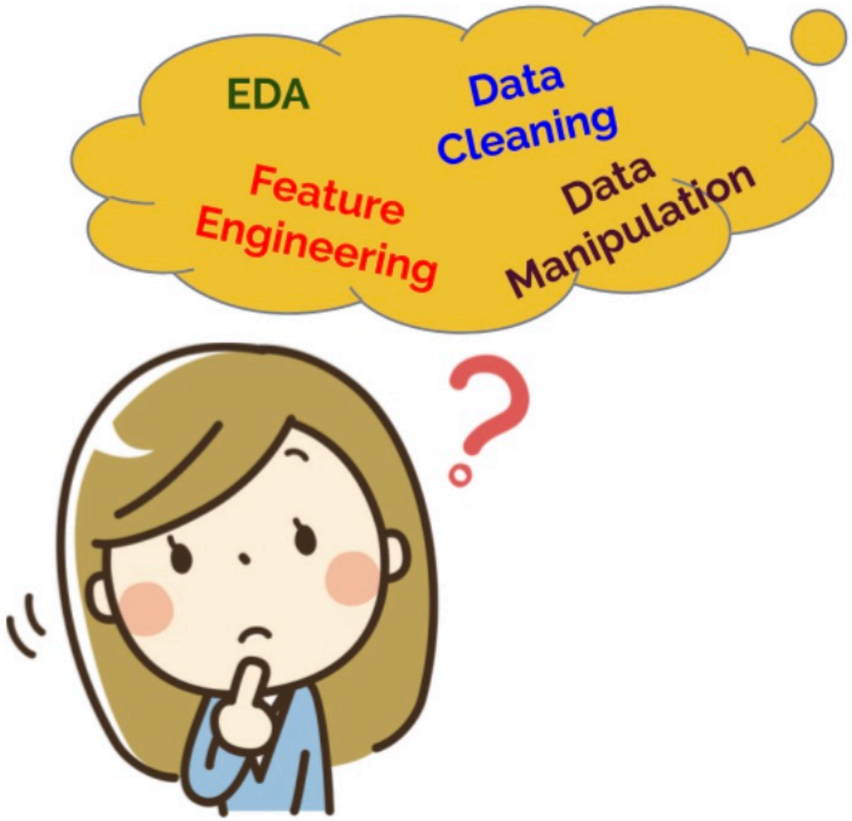
Session-1

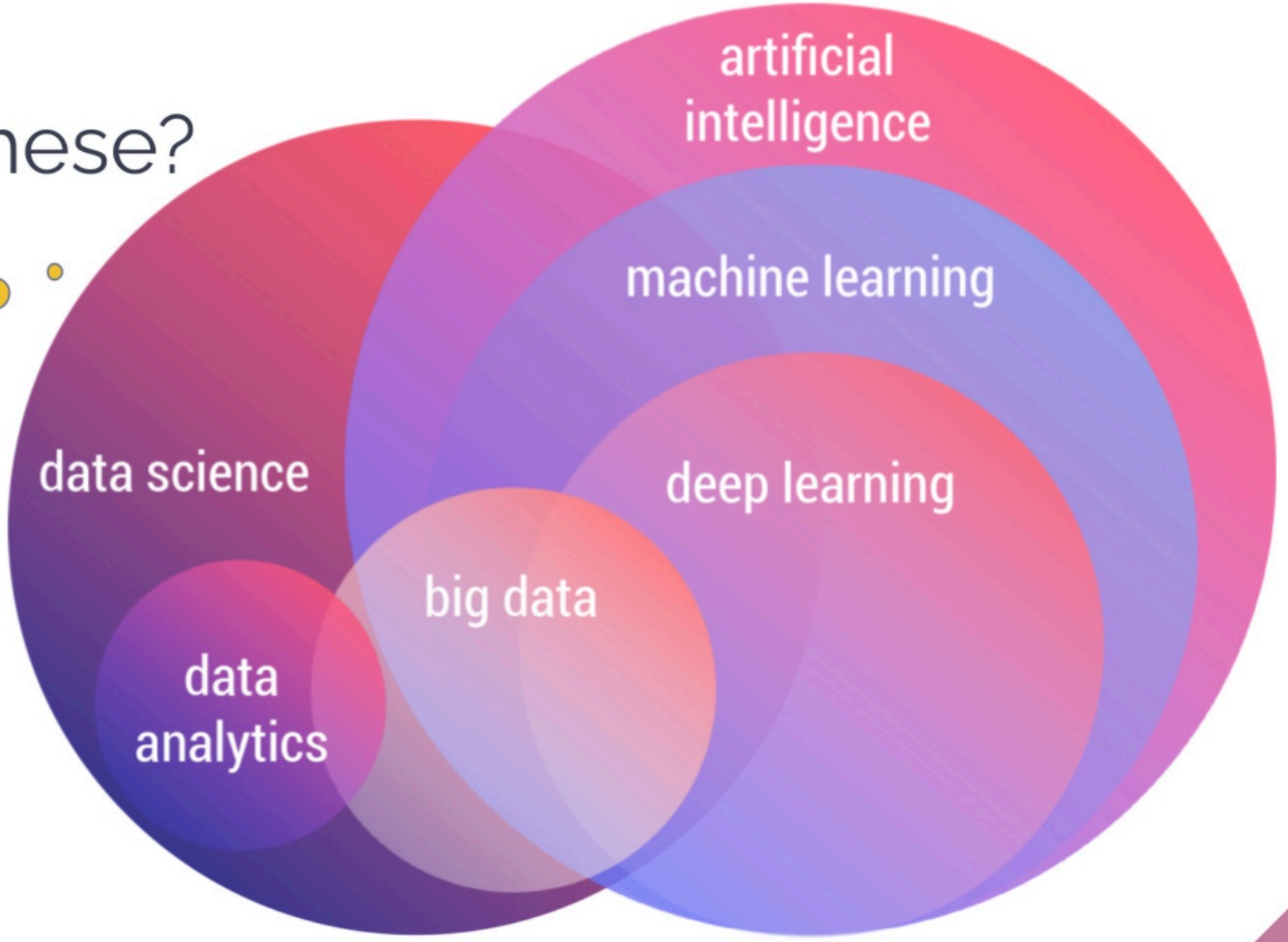
Table of Contents

Big Picture

Big Picture

- Where am I?
- Why will I learn these?



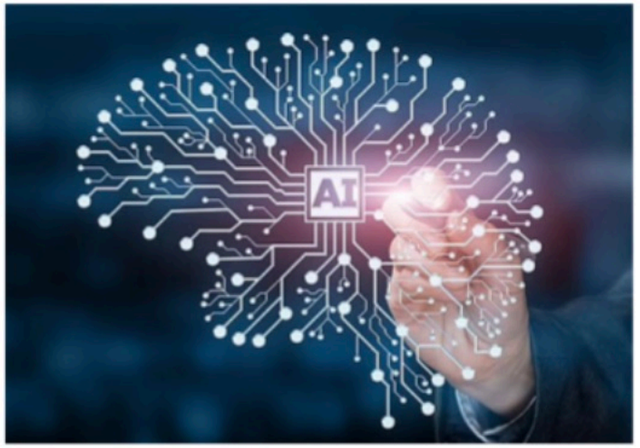


Big Picture

Data Analytics

- Excel/Google Spreadsheets
- SQL
- BI Tools (Tableau, Power BI)
- Python ...

Artificial Intelligence



- Modelling
  - Prediction/Forecasting
    - Regression
    - Classification
    - Clustering...

Big Picture

Artificial Intelligence

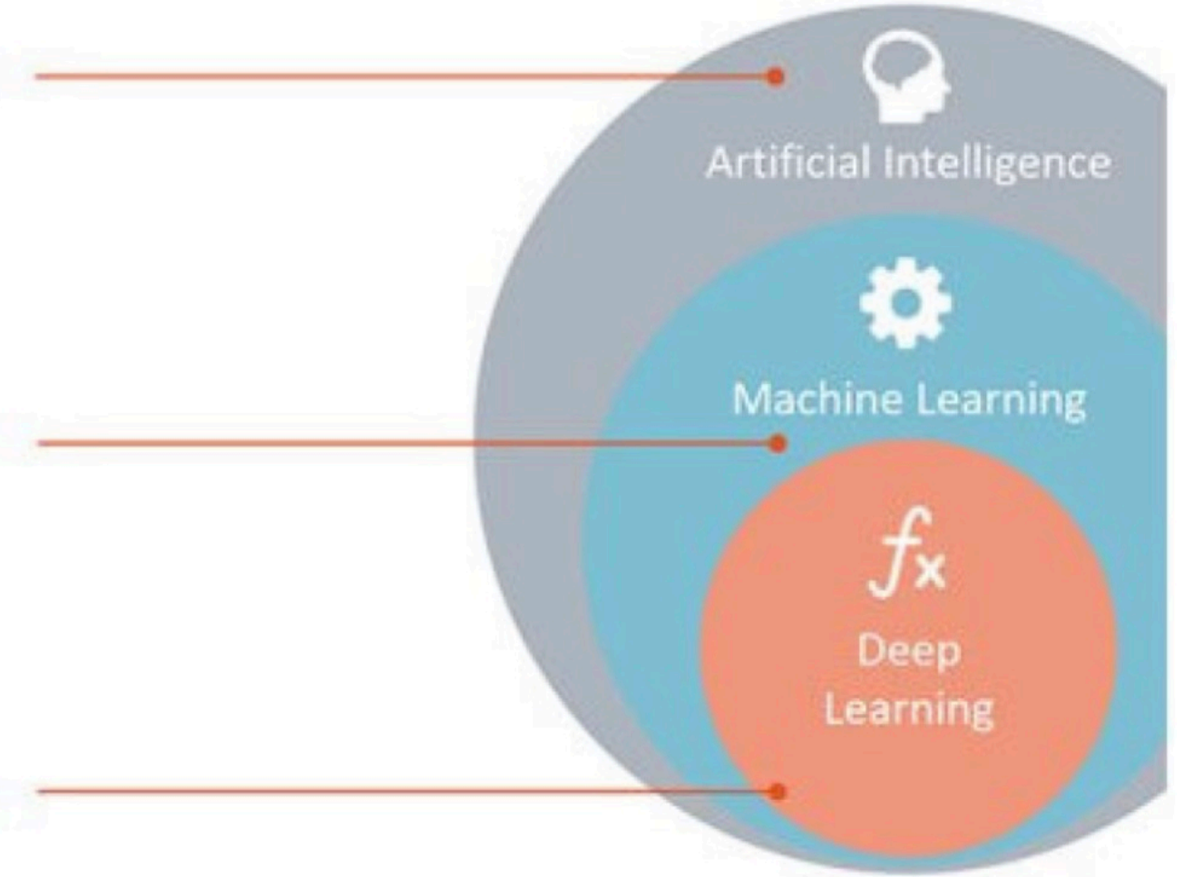
Any technique which enables computers to mimic human behavior.

Machine Learning

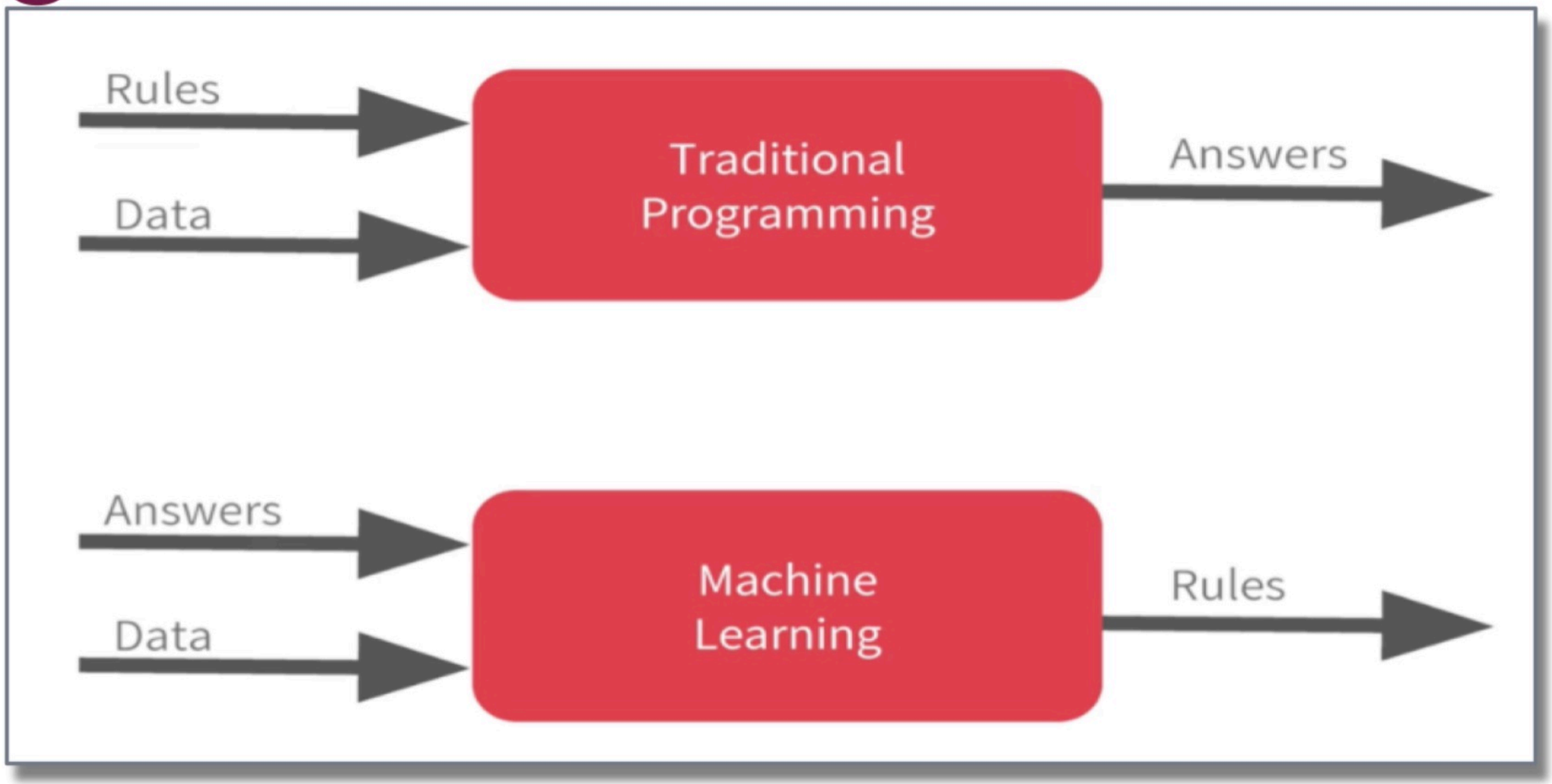
Subset of AI techniques which use statistical methods to enable machines to improve with experiences.

Deep Learning

Subset of ML which make the computation of multi-layer neural networks feasible.



Big Picture



Without anyone programming the logic, In Traditional programming one has to manually formulate/code rules while in Machine Learning **the algorithms automatically formulate the rules from the data**, which is very powerful.

Big Picture

Independent Variables , "X"

Dependent Variables

Target

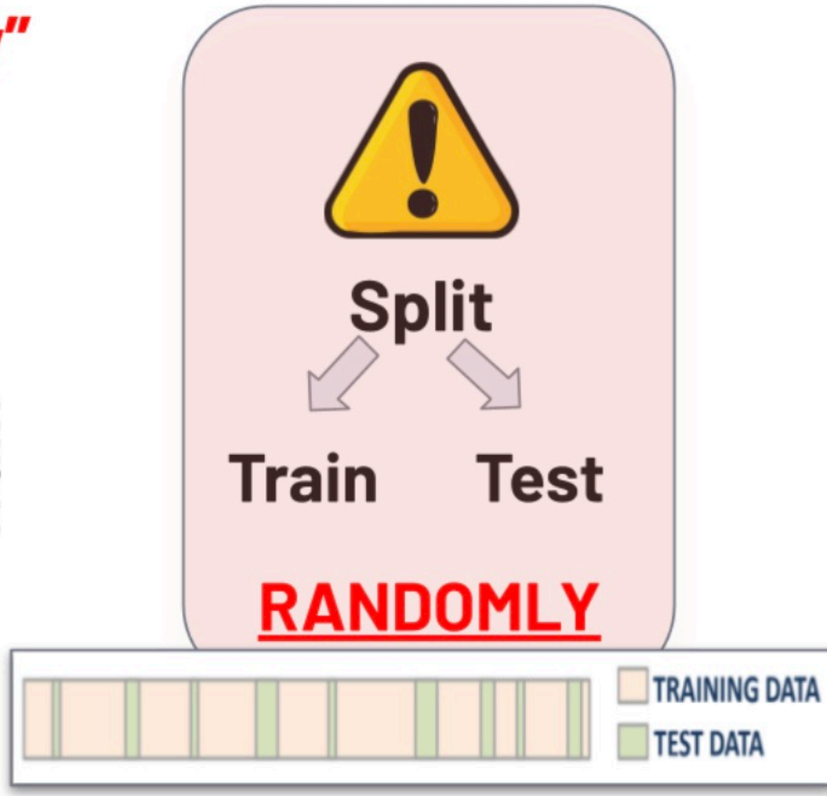
"y"

Features | **Label**

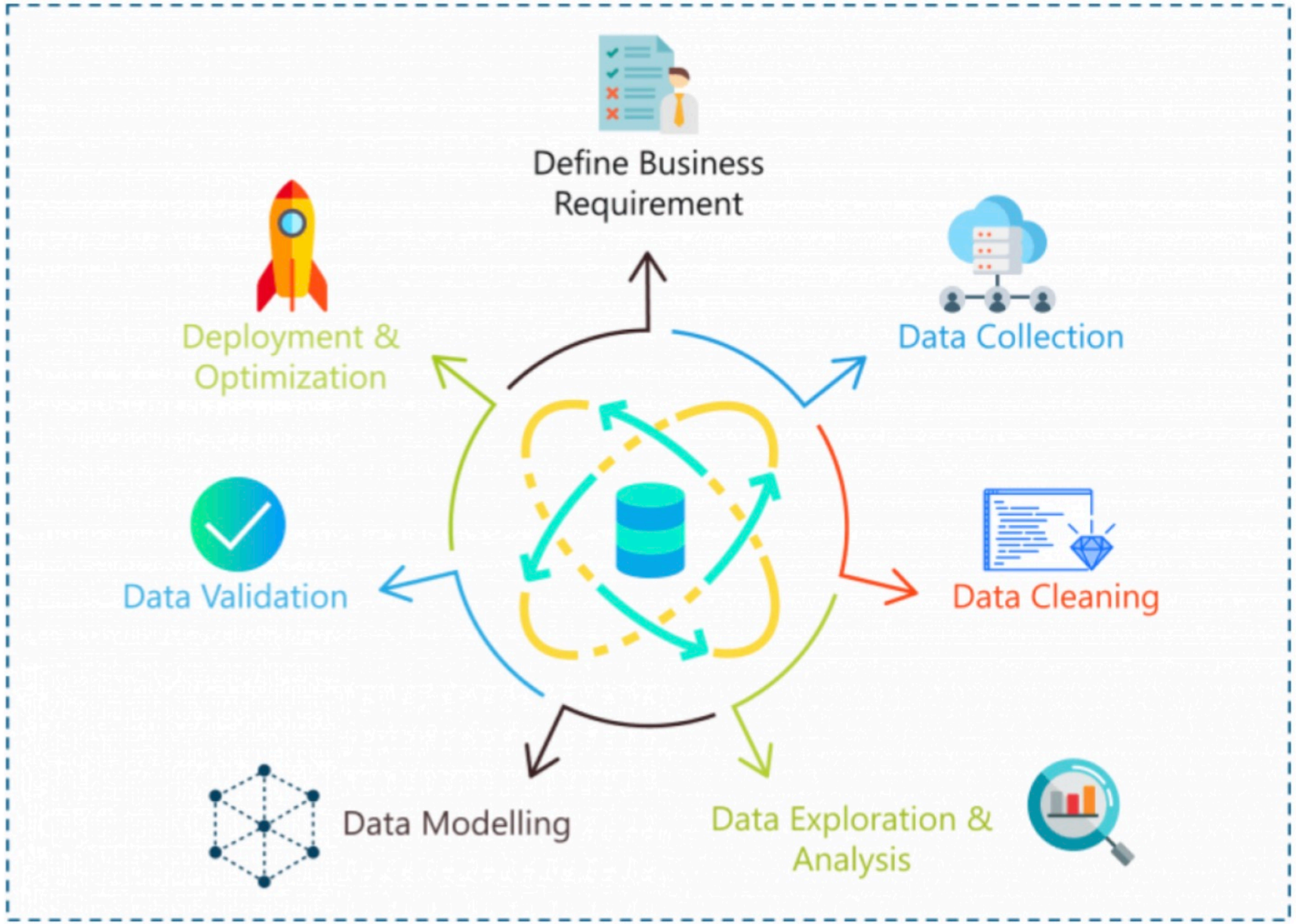
Position	Experience	Skill	Country	City	Salary ($)
Developer	0	1	USA	New York	103100
Developer	1	1	USA	New York	104900
Developer	2	1	USA	New York	106800
Developer	3	1	USA	New York	108700
Developer	4	1	USA	New York	110400
Developer	5	1	USA	New York	112300
Developer	6	1	USA	New York	114200
Developer	7	1	USA	New York	116100
Developer	8	1	USA	New York	117800
Developer	9	1	USA	New York	119700
Developer	10	1	USA	New York	121600

Train

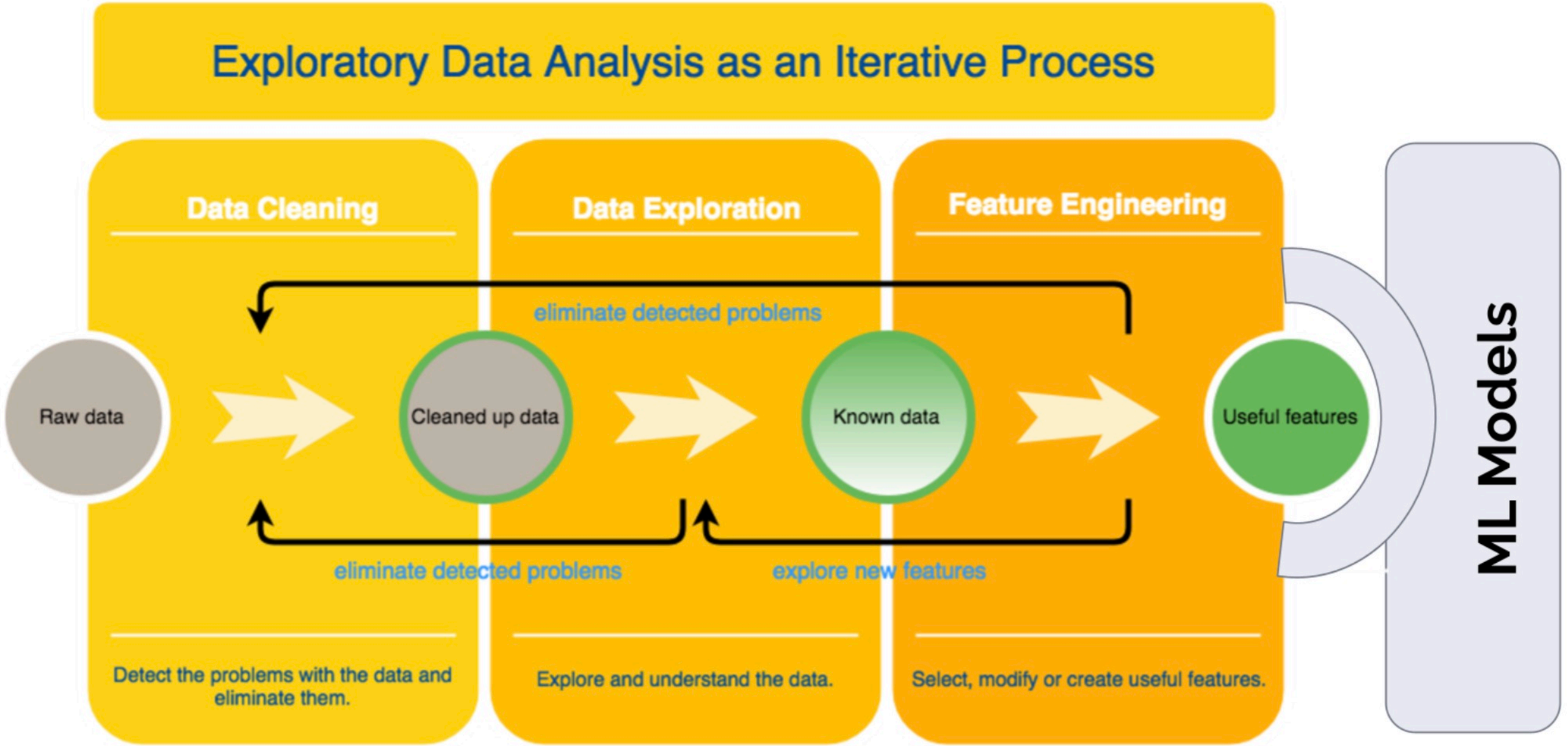
Test



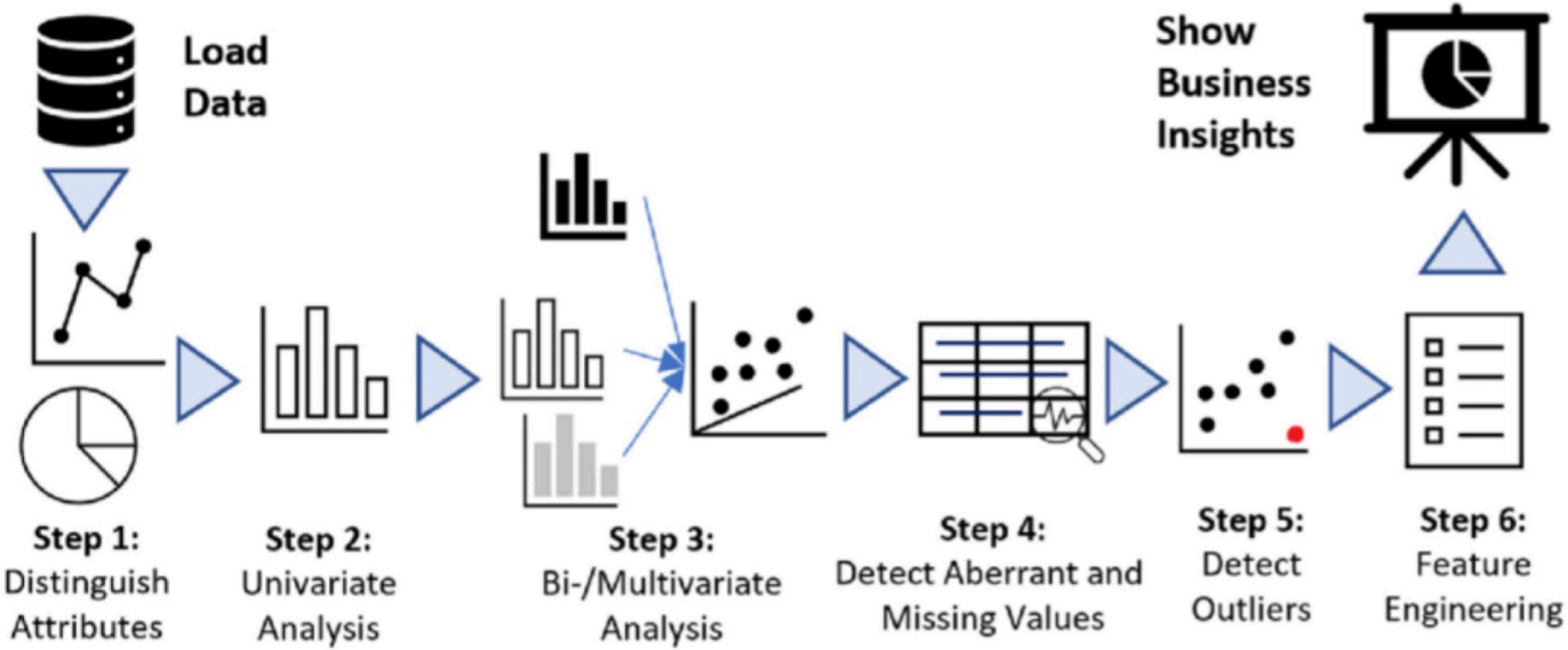
Big Picture


Define Business Requirement
Data Collection
Data Cleaning
Data Exploration & Analysis
Data Modelling
Data Validation
Deployment & Optimization

Big Picture


Exploratory Data Analysis as an Iterative Process
Data Cleaning
Data Exploration
Feature Engineering
eliminate detected problems
Raw data
Cleaned up data
Known data
Useful features
ML Models
eliminate detected problems
explore new features
Detect the problems with the data and eliminate them.
Explore and understand the data.
Select, modify or create useful features.

Big Picture


Load Data
Show Business Insights
Step 1: Distinguish Attributes
Step 2: Univariate Analysis
Step 3: Bi-/Multivariate Analysis
Step 4: Detect Aberrant and Missing Values
Step 5: Detect Outliers
Step 6: Feature Engineering

Big Picture

```
df.head(3).T
```

	0	1
url	https://www.autoscout24.com//offers/audi-a1-sp...	https://www.autoscout24.com//offers/audi-a1-1-...
make_model	Audi A1	Audi A1
short_description	Sportback 1.4 TDI S-tronic Xenon Navi Klima	1.8 TFSI sport
body_type	Sedans	Sedans
price	15770	14500
vat	VAT deductible	Price negotiable
km	56,013 km	80,000 km
registration	01/2016	03/2017
prev_owner	2 previous owners	None
kW	NaN	NaN
hp	66 kW	141 kW
Type	[, Used, , Diesel (Particulate Filter)]	[, Used, , Gasoline]
Previous Owners	\n2\n	NaN
Next Inspection	['\n06/2021\n, \n99 g CO2/km (comb)\n]	NaN
Inspection new	['\nYes\n, \nEuro 6\n]	NaN
Warranty	['\n, \n, \n4 (Green)\n]	NaN
Full Service	['\n, \n]	NaN
Non-smoking Vehicle	['\n, \n]	NaN
null	[]	[]
Make	\nAudi\n	\nAudi\n
Model	['\n, A1, \n]	['\n, A1, \n]
Offer Number	['\nLR-062483\n]	NaN
First Registration	['\n, 2016, \n]	['\n, 2017, \n]
Body Color	['\n, Black, \n]	['\n, Red, \n]

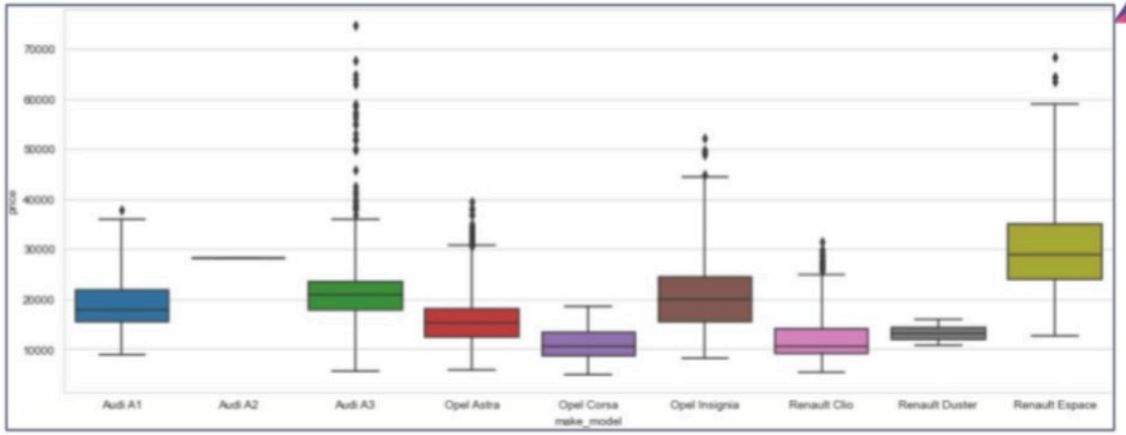
```
df.head(3).T
```

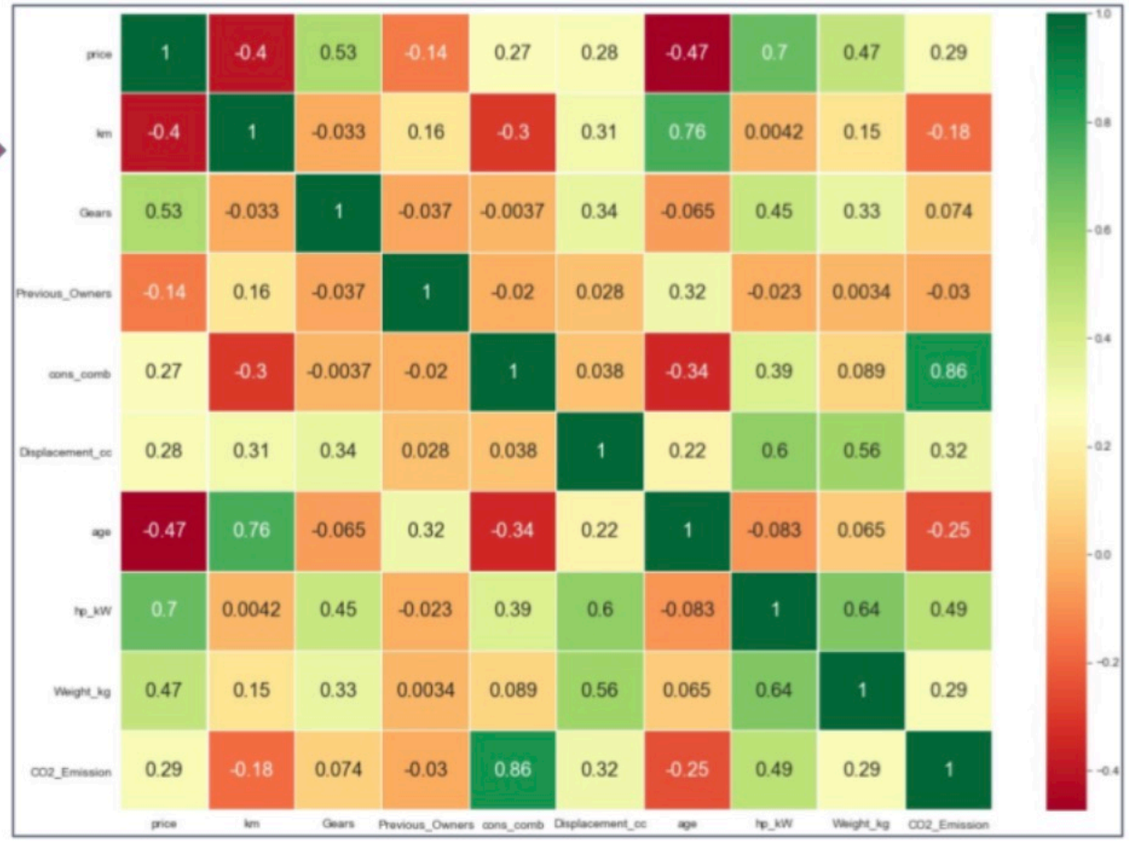
	0	1
make_model	Audi A1	Audi A1
body_type	Sedans	Sedans
price	15770	14500
vat	VAT deductible	Price negotiable
km	56013.000	80000.000
Type	Used	Used
Fuel	Diesel	Benzine
Gears	7.000	7.000
Comfort_Convenience	Air conditioning,Armrest,Automatic climate con...	Air conditioning,Automatic climate control,Hil...
Entertainment_Media	Bluetooth,Hands-free equipment,On-board comput...	Bluetooth,Hands-free equipment,On-board comput...
Extras	Alloy wheels,Catalytic Converter,Voice Control	Alloy wheels,Sport seats,Sport suspension,Voic...
Safety_Security	ABS,Central door lock,Daytime running lights,D...	ABS,Central door lock,Central door lock with r...
age	3.000	2.000
Previous_Owners	2.000	1.000
hp_kW	66.000	141.000
Inspection_new	1	0
Paint_Type	Metallic	Metallic
Upholstery_type	Cloth	Cloth
Nr_of_Doors	5.000	3.000
Nr_of_Seats	5.000	4.000
Gearing_Type	Automatic	Automatic
Displacement_cc	1422.000	1798.000
Weight_kg	1220.000	1255.000
Drive_chain	front	front
cons_comb	3.800	5.600
CO2_Emission	99.000	129.000

Big Picture

```
df.head(3).T
```

	0	1
make_model	Audi A1	Audi A1
body_type	Sedans	Sedans
price	15770	14500
vat	VAT deductible	Price negotiable
km	56013.000	80000.000
Type	Used	Used
Fuel	Diesel	Benzine
Gears	7.000	7.000
Comfort_Convenience	Air conditioning,Armrest,Automatic climate con...	Air conditioning,Automatic climate control,Hil...
Entertainment_Media	Bluetooth,Hands-free equipment,On-board comput...	Bluetooth,Hands-free equipment,On-board comput...
Extras	Alloy wheels,Catalytic Converter,Voice Control	Alloy wheels,Sport seats,Sport suspension,Voic...
Safety_Security	ABS,Central door lock,Daytime running lights,D...	ABS,Central door lock,Central door lock with r...
age	3.000	2.000
Previous_Owners	2.000	1.000
hp_kW	66.000	141.000
Inspection_new	1	0
Paint_Type	Metallic	Metallic
Upholstery_type	Cloth	Cloth
Nr_of_Doors	5.000	3.000
Nr_of_Seats	5.000	4.000
Gearing_Type	Automatic	Automatic
Displacement_cc	1422.000	1798.000
Weight_kg	1220.000	1255.000
Drive_chain	front	front
cons_comb	3.800	5.600
CO2_Emission	99.000	129.000

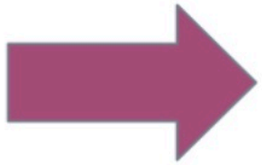




Big Picture

```
df.head(3).T
```

	0	1
make_model	Audi A1	Audi A1
body_type	Sedans	Sedans
price	15770	14500
vat	VAT deductible	Price negotiable
km	56013.000	80000.000
Type	Used	Used
Fuel	Diesel	Benzine
Gears	7.000	7.000
Comfort_Convenience	Air conditioning,Armrest,Automatic climate con...	Air conditioning,Automatic climate control,Hil...
Entertainment_Media	Bluetooth,Hands-free equipment,On-board comput...	Bluetooth,Hands-free equipment,On-board comput...
Extras	Alloy wheels,Catalytic Converter,Voice Control	Alloy wheels,Sport seats,Sport suspension,Voic...
Safety_Security	ABS,Central door lock,Daytime running lights,D...	ABS,Central door lock,Central door lock with r...
age	3.000	2.000
Previous_Owners	2.000	1.000
hp_kW	66.000	141.000
Inspection_new	1	0
Paint_Type	Metallic	Metallic
Upholstery_type	Cloth	Cloth
Nr_of_Doors	5.000	3.000
Nr_of_Seats	5.000	4.000
Gearing_Type	Automatic	Automatic
Displacement_cc	1422.000	1798.000
Weight_kg	1220.000	1255.000
Drive_chain	front	front
cons_comb	3.800	5.600
CO2_Emission	99.000	129.000



```
df_final.head().T
```

	0	1	2	3	4
price	15770.000	14500.000	14640.000	14500.000	16790.000
km	56013.000	80000.000	83450.000	73000.000	16200.000
Gears	7.000	7.000	7.000	6.000	7.000
age	3.000	2.000	3.000	3.000	3.000
Previous_Owners	2.000	1.000	1.000	1.000	1.000
hp_kW	66.000	141.000	85.000	66.000	66.000
Inspection_new	1.000	0.000	0.000	0.000	1.000
Displacement_cc	1422.000	1798.000	1598.000	1422.000	1422.000
Weight_kg	1220.000	1255.000	1135.000	1195.000	1135.000
cons_comb	3.800	5.600	3.800	3.800	4.100
cc_Air conditioning	1.000	1.000	1.000	0.000	1.000
cc_Air suspension	0.000	0.000	0.000	1.000	0.000
cc_Armrest	1.000	0.000	0.000	1.000	1.000
cc_Automatic climate control	1.000	1.000	0.000	0.000	1.000
cc_Auxiliary heating	0.000	0.000	0.000	1.000	0.000
cc_Cruise control	1.000	0.000	1.000	0.000	0.000
cc_Electric Starter	0.000	0.000	0.000	0.000	0.000
cc_Electric tailgate	0.000	0.000	0.000	0.000	0.000
cc_Electrical side mirrors	1.000	0.000	1.000	1.000	1.000
cc_Electrically adjustable seats	0.000	0.000	0.000	0.000	0.000
cc_Electrically heated windshield	0.000	0.000	0.000	0.000	0.000
cc_Heads-up display	0.000	0.000	0.000	1.000	0.000
cc_Heated steering wheel	0.000	0.000	0.000	0.000	0.000
cc_Hill Holder	1.000	1.000	1.000	1.000	1.000
cc_Keyless central door lock	0.000	0.000	0.000	0.000	0.000

NumPy

Table of Contents

Introduction

What is NumPy?

Numerical **Py**thon

- NumPy is the **fundamental package for scientific computing in Python**.
- It is a Python library that provides;
  - A **multidimensional** array object,
  - **Various derived objects** (such as masked arrays and matrices),
  - An assortment of routines for **fast** operations on arrays including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.
- The core of NumPy is **well-optimized C code**.

Introduction

What is NumPy?

- At the core of the NumPy package, is the **ndarray object**.
- **Differences between NumPy arrays and the standard Python sequences**:
  - NumPy arrays **have a fixed size** at creation, unlike Python lists. Changing the size of an ndarray will create a new array and delete the original.
  - The elements in a NumPy array are all required to be of the **same data type**, and thus will be the **same size** in memory.
  - **Advanced mathematical operations** are executed more efficiently and with less code than is possible using Python's built-in sequences.

Introduction

Why is NumPy Fast?

- Numpy draws its power from its **vectorization** and **broadcasting** features.
- **Vectorization** describes the absence of any explicit looping, indexing, etc., in the code. Vectorized code has many advantages, among which are:
  - Vectorized code is **more concise and easier to read**.
  - **Fewer lines of code** generally means **fewer bugs**.
  - The code more closely resembles **standard mathematical notation** (making it easier, typically, to correctly code mathematical constructs)
  - Vectorization results in more **“Pythonic” code**. Without vectorization, our code would be littered with inefficient and difficult to read for loops.

Introduction

Why is NumPy Fast?

- **Broadcasting** is the term used to describe the implicit element-by-element behavior of operations.
- in NumPy all operations, not just arithmetic operations, but logical, bit-wise, functional, etc., behave in this implicit element-by-element fashion, i.e., they broadcast.

https://www.tutorialspoint.com/numpy/numpy_broadcasting.htm

https://erdincuzun.com/numpy/05-numpy-broadcasting/

Introduction

Installation

Google colab

ANACONDA®

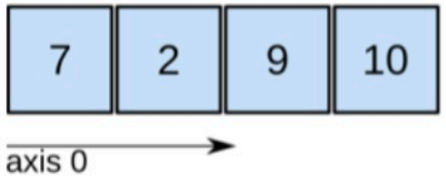
NumPy

Numpy Arrays

What is Array?

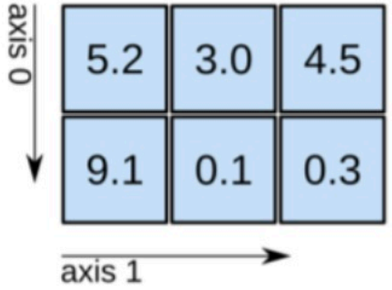
- Array is a data structure that contains a group of elements. Typically these elements are all of the same data type, such as an integer or string.

1D array



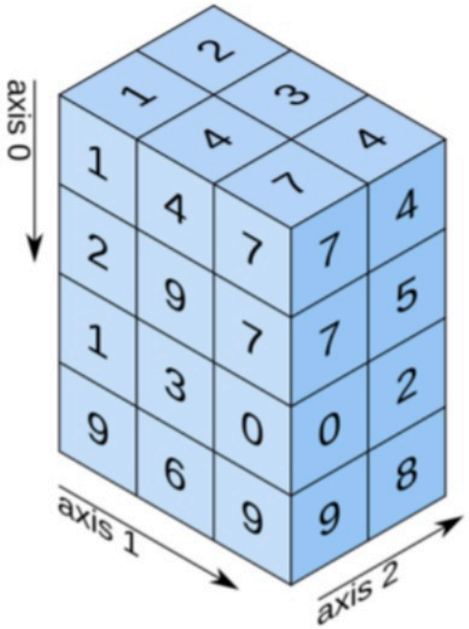
shape: (4,)

2D array

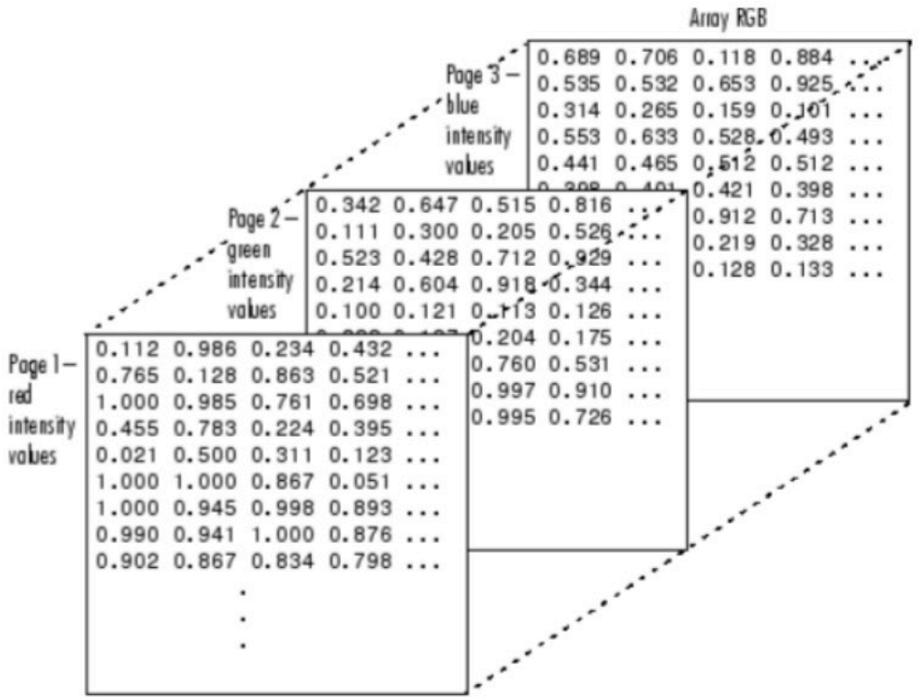


shape: (2, 3)

3D array



shape: (4, 3, 2)



Numpy Arrays

Advantages of Arrays by Lists

- Less memory
- Much faster
- Convenient
- Computations

let's see its implementation in notebook

Numpy Arrays

- Built-in Array Creation Methods
  - `arange`
  - `zeros`, `ones`, `full`
  - `linspace`
  - `eye`
  - `random.rand`
  - `random.randn`
  - `random.randint`

Numpy Arrays

- Array Methods & Attributes
  - `reshape`
  - `max`, `min`, `argmax`, `argmin`
  - `ndim`
  - `shape`
  - `size`
  - `dtype`
  - `itemsize`

Data Analysis with Python

let's start the hands-on phase