University of Salford MANCHESTER

SCHOOL OF SCIENCE, ENGINEERING & ENVIRONMENT

Program: MSc of Data Science
Module: Big Data Tools and Techniques

Week 3 - Part 1

Data Types in Python
and
Main File Types in Big Data Analysis

2025

Learning Outcomes

1. To learn data types in Python
2. To know main file types in Big Data analysis
3. To understand CSV and JSON structures more in-depth

Data Types in Python

Python Class & Object

A class is considered as a blueprint of objects. We can think of the class as a sketch (prototype) of a house. It contains all the details about the floors, doors, windows, etc. Based on the class we build a house and the house is one object.

Class

Objects 1

Object 2

....

....

https://www.mygreatlearning.com/blog

Mutable vs Immutable Objects

- **Mutable Object**

Mutable is when something is changeable or can change.

In Python, ‘mutable’ is the ability of objects to change their values.

- **Immutable Object**

Immutable is when change is impossible over time.

In Python, if the value of an object cannot be changed over time, then it is known as immutable. Once created, the value of these objects is permanent.

https://www.mygreatlearning.com/blog

Objects in Python

If we change the value of the object 1 from **35** to **168** then we have a new object, 2.

So, value of the object 1 is not changeable.

(We can change the value, but object will change entirely, compare IDs of 2 objects)

Class: Integer number (int)		Object 1			Object 2		
		ID	Type	Value	ID	Type	Value
		1407	Integer	35	2842	Integer	168

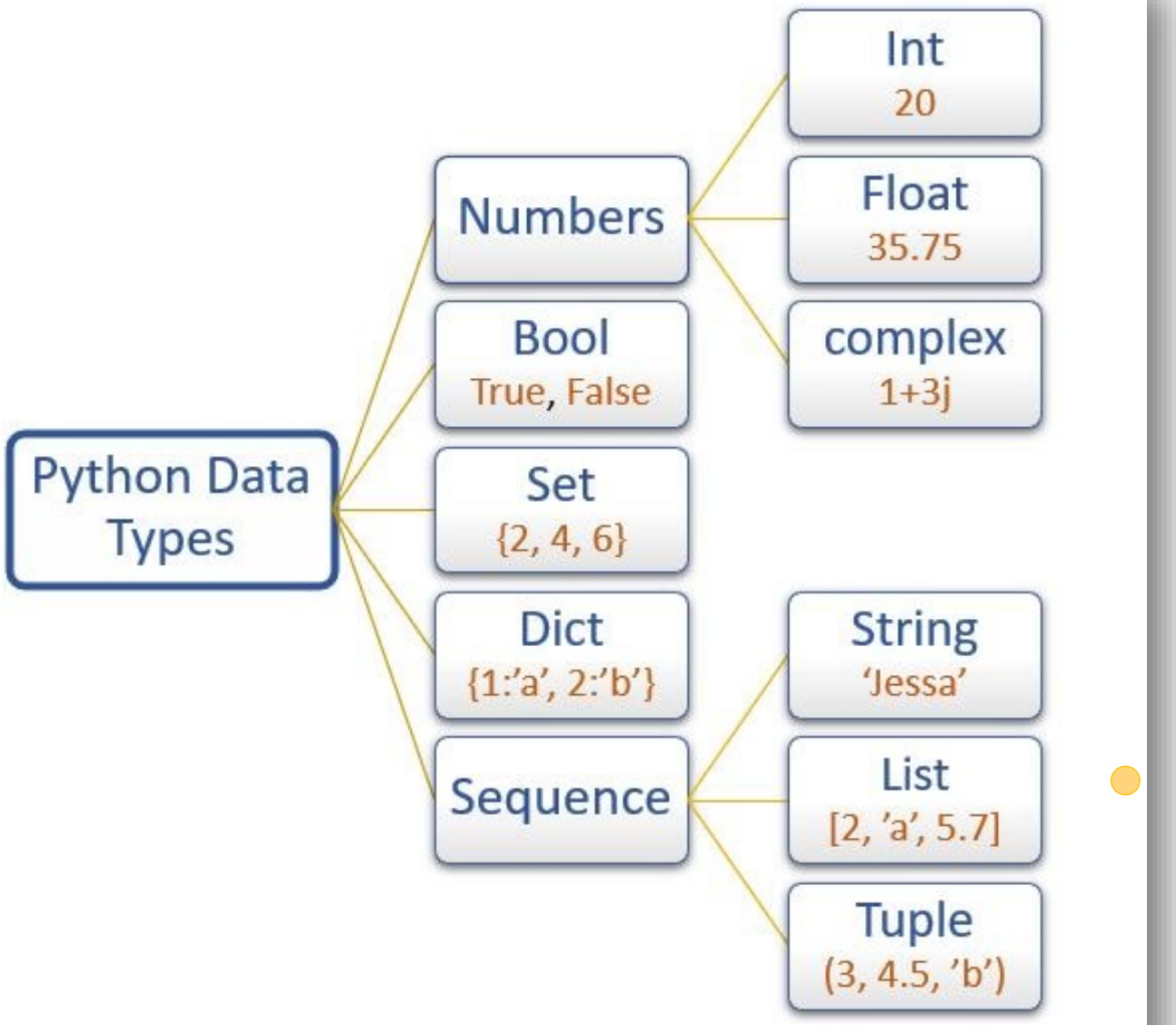
As a result, an integer object is Immutable

https://www.mygreatlearning.com/blog

Built-in

Classes (**Data Types** or **Object Types**)

in Python



https://pynative.com/python-data-types/

Set: { } Curly bracket
Dict: { } Curly bracket
String: ' ' or " " or ''' '''
List: [ ] Square bracket
Tuple: ( ) round bracket

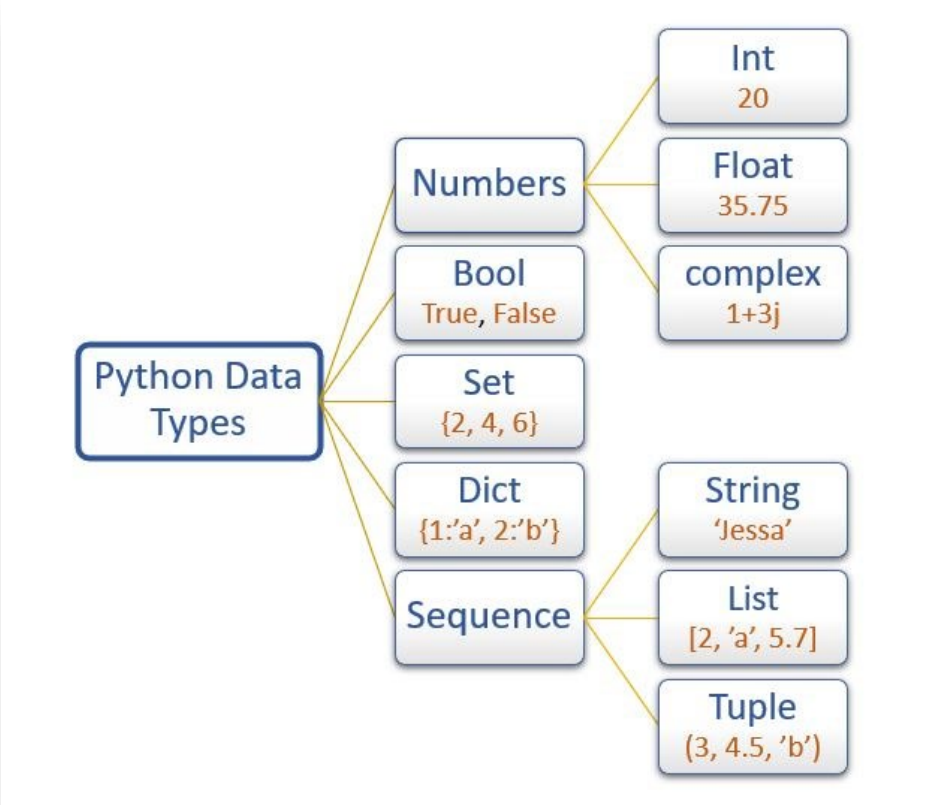
Lists are mutable.
Tuples are immutable,

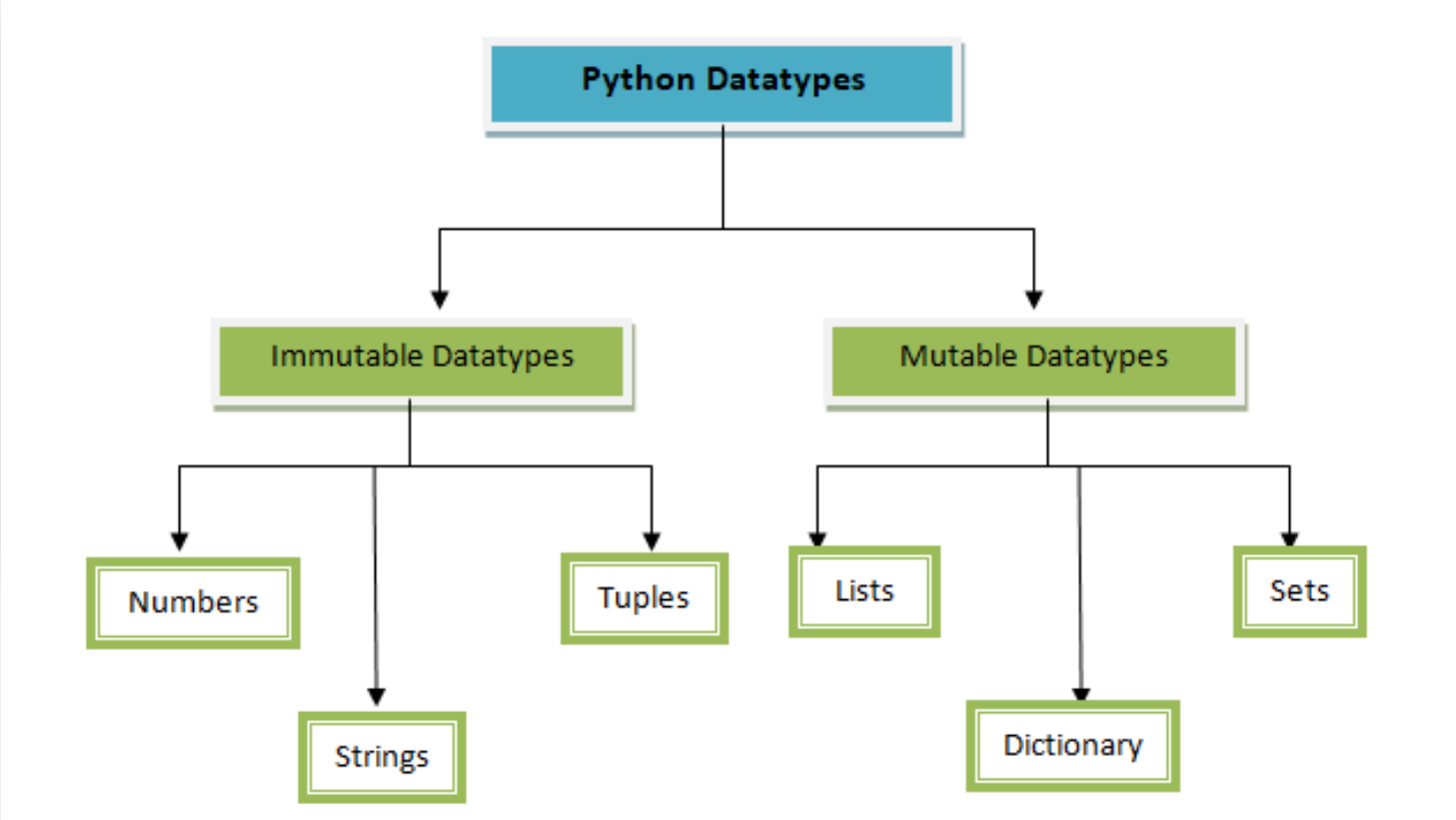
As a result, each object belongs to a class and each class has a type.

X is an **object**

```
In [1]: x = 2

In [3]: print(type(x))
        <class 'int'>

In [4]: x = 'Hello BDTT group'

In [5]: print(type(x))
        <class 'str'>

In [16]: x = 2.2023

In [17]: print(type(x))
         <class 'float'>
```

X belongs to a class and the class type is **integer**


Python Data Types
Numbers
Int
20
Float
35.75
complex
1+3j
Bool
True, False
Set
{2, 4, 6}
Dict
{1:'a', 2:'b'}
Sequence
String
'Jessa'
List
[2, 'a', 5.7]
Tuple
(3, 4.5, 'b')


Python Datatypes
Immutable Datatypes
Mutable Datatypes
Numbers
Strings
Tuples
Lists
Dictionary
Sets

Data type	Description	Example
str	To store textual/string data	`name = 'Jessa'`
list	To store a sequence of mutable data	`l = [3, 'a', 2.5]`
tuple	To store sequence immutable data	`t =(2, 'b', 6.4)`
dict	To store key: value pair	`d = {1:'J', 2:'E'}`
set	To store unorder and unindexed values	`s = {1, 3, 5}`

Sample in Python

Example : Python Class and Objects

```python
# define a class
class Bike:
    name = ""
    gear = 0

# create object of class
bike1 = Bike()

# access attributes and assign new values
bike1.gear = 11
bike1.name = "Mountain Bike"

print(f"Name: {bike1.name}, Gears: {bike1.gear} ")
```

Note: The variables inside a class are called attributes.

Run Code »

Output

```
Name: Mountain Bike, Gears: 11
```

https://www.programiz.com/python-programming/class-object

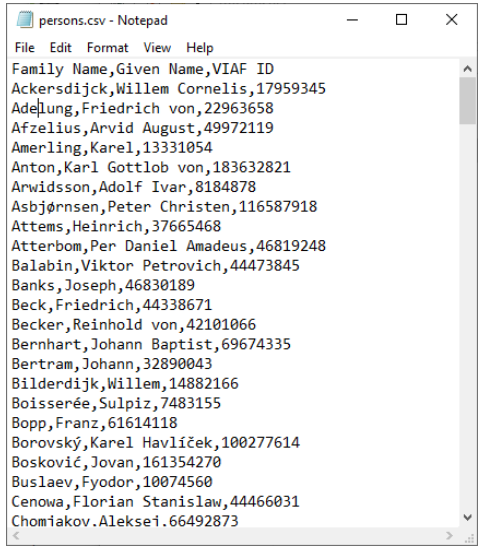
Big Data Popular File Formats

What is a file format?

The **file format** is the structure of a file that tells a program how to display the file's contents.

In other words, a file format, also called a **file extension**, is the layout of a file in terms of how the data within the file is organised.

Bigdata file formats


persons.csv - Notepad

File Edit Format View Help

```
Family Name,Given Name,VIAF ID
Ackersdijck,Willem Cornelis,17959345
Adelung,Friedrich von,22963658
Afzelius,Arvid August,49972119
Amerling,Karel,13331054
Anton,Karl Gottlob von,183632821
Arwidsson,Adolf Ivar,8184878
Asbjørnsen,Peter Christen,116587918
Attems,Heinrich,37665468
Atterbom,Per Daniel Amadeus,46819248
Balabin,Viktor Petrovich,44473845
Banks,Joseph,46830189
Beck,Friedrich,44338671
Becker,Reinhold von,42101066
Bernhart,Johann Baptist,69674335
Bertram,Johann,32890043
Bilderdijk,Willem,14882166
Boisserée,Sulpiz,7483155
Bopp,Franz,61614118
Borovský,Karel Havlíček,100277614
Bosković,Jovan,161354270
Buslaev,Fyodor,10074560
Cenowa,Florian Stanislaw,44466031
Chomiakov,Aleksei,66492873
```

File Format	Acronym or name
CSV	**C**omma **S**eparated **V**alues
JSON	**J**ava **S**cript **O**bject **N**otation
Parquet	name
AVRO	name
ORC	**O**ptimized **R**ow **C**olumnar

```
{
    hey: "guy",
    anumber: 243,
  - anobject: {
        whoa: "nuts",
      - anarray: [
            1,
            2,
            "thr<h1>ee"
        ],
        more: "stuff"
    },
    awesome: true,
    bogus: false,
    meaning: null,
    japanese: "明日がある。",
    link: http://jsonview.com,
    notLink: "http://jsonview.com is great"
}
```

```
[
  {
    "type": "record",
    "namespace": "com.vladkrava.avroconverterdemo.domain",
    "name": "EmailData",
    "doc": "Dummy email information which used to describe
            basic user data",
    "fields": [
      {
          "name": "username",
          "type": "string",
          "doc": "Unique identifier/reference of User in a System"
      },
      {
          "name": "email",
          "type": "string",
          "doc": "User email address"
      },
      {
          "name": "subscribed",
          "type": "boolean",
          "default": true,
          "doc": "A property which defines whether User has a
                  newsletter subscription"
      }
    ]
  }
]
```

Bigdata file formats

CSV

Fields name

AdventureWorks.csv - Notepad

File Edit Format View Help

```
SalesOrderID,OrderDate,Product,Model,SalesPerson,Territory,Region,Subcategory,Category,OrderQty,Unit
43659,2011-05-31 00:00:00.000,"BK-M82B-42Mountain-100 Black, 42",Mountain-100,Eddie Savis,Southeast,
43659,2011-05-31 00:00:00.000,"BK-M82B-44Mountain-100 Black, 44",Mountain-100,Eddie Savis,Southeast,
43659,2011-05-31 00:00:00.000,"BK-M82B-48Mountain-100 Black, 48",Mountain-100,Eddie Savis,Southeast,
43659,2011-05-31 00:00:00.000,"BK-M82S-38Mountain-100 Silver, 38",Mountain-100,Eddie Savis,Southeast
43659,2011-05-31 00:00:00.000,"BK-M82S-42Mountain-100 Silver, 42",Mountain-100,Eddie Savis,Southeast
43659,2011-05-31 00:00:00.000,"BK-M82S-44Mountain-100 Silver, 44",Mountain-100,Eddie Savis,Southeast
43659,2011-05-31 00:00:00.000,"BK-M82S-48Mountain-100 Silver, 48",Mountain-100,Eddie Savis,Southeast
```

Comma separates the values

Comma is a delimiter → **Because shows limits of an object (start and end)**

Delimiters

A delimiter is one or more characters that separate text strings. Common delimiters are:

commas (,)

semicolon (;)

quotes (“ ‘)

braces ({})

pipes (|)

slashes (/ \ )

white space ( )

Bigdata file formats

JSON

Heavily used in APIs. json has **Nested format**. It is widely adopted and human-readable but it can be **difficult to read if there are lots of nested fields.**

First layer:
- Type
- Id
- Attributes
- author

sample.json > [ ] data

```
1  {
2      "data": [
3          {
           "type": "articles",
5          "id": "1",
6          "attributes": {
7              "title": "Working with JSON Data in python",
8              "description": "This article explains the various ways to work with JSON data in python.",
9              "created": "2020-12-28T14:56:29.000Z",
10             "updated": "2020-12-28T14:56:28.000Z"
11         },
12         "author": {
13             "id": "1",
14             "name": "Aveek Das"
15         }
16         }
17     ]
18 }
```

First nest

Another nest

Attributes :
- title
- description
- created
- author

Author :
- id
- name

Orientations of file formats in Hadoop

- ***Row-oriented:*** *CSV, JSON,* AVRO
- ***Column-oriented:*** *PARQUET, ORC*

https://www.superoutlier.tech/getting-started-format-file-in-big-data/

Row-Oriented vs Column-Oriented

Row-Oriented vs Column-Oriented

Row-oriented: rows stored sequentially in a file

Key	Fname	Lname	State	Zip	Phone	Age	Sales
1	Bugs	Bunny	NY	11217	(123) 938-3235	34	100
2	Yosemite	Sam	CA	95389	(234) 375-6572	52	500
3	Daffy	Duck	NY	10013	(345) 227-1810	35	200
4	Elmer	Fudd	CA	04578	(456) 882-7323	43	10
5	Witch	Hazel	CA	01970	(567) 744-0991	57	250

Column-oriented: each column is stored in a separate file
Each column for a given row is at the same offset.

Key
1
2
3
4
5

Fname
Bugs
Yosemite
Daffy
Elmer
Witch

Lname
Bunny
Sam
Duck
Fudd
Hazel

State
NY
CA
NY
CA
CA

Zip
11217
95389
10013
04578
01970

Phone
(123) 938-3235
(234) 375-6572
(345) 227-1810
(456) 882-7323
(567) 744-0991

Age
34
52
35
43
57

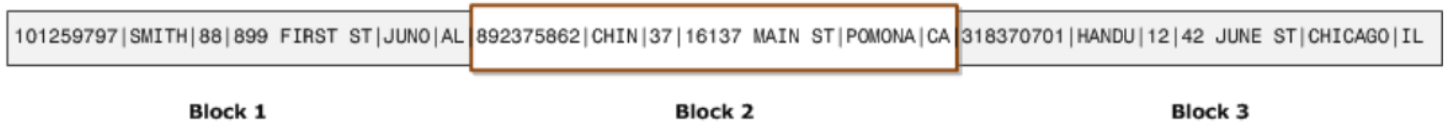
Sales
100
500
200
10
250

http://leelaprasadhadoop.blogspot.com/2017/05/difference-between-orc-and-parquet.html

Row-oriented

Each row contains field (column) values for a single record (row). The same row of data is stored together and continuous storage. In case, we need to read a small amount of data, the entire row needs to be read into the memory. Infrastructure will be delayed for the overhead of reading data.

SSN	Name	Age	Addr	City	St
101259797	SMITH	88	899 FIRST ST	JUNO	AL
892375862	CHIN	37	16137 MAIN ST	POMONA	CA
318370701	HANDU	12	42 JUNE ST	CHICAGO	IL



Row-oriented: *CSV, JSON,* AVRO

Column-oriented

organize data by field (column), keeping all of the data associated with a field next to each other in memory.

The column-oriented format makes it possible to skip unneeded columns when reading data. We have Parquet and ORC (optimize version of Parquet) best candidates in Column-oriented.

Friends

ID	Name	Birthday	Favorite Color
1	Luke	1/1/1921	Black
2	Bob	2/1/1980	Blue
3	Alice	3/1/1970	Purple

File 1	File 2	File 3	File 4
ID	Name	Birthday	Favorite Color
1	Luke	1/1/1921	Black
2	Bob	2/1/1980	Blue
3	Alice	3/1/1970	Purple

Column-oriented: *PARQUET, ORC*

Activity: Exploring file structure

Please complete the following activities and fill out the Padlet board as much as you can.

Each of you should create two stickers on the wall, adding your name if you'd like, and label them with either "JSON" or "CSV" as the title. If you can't find a free space, don't worry, there's no need to add your sticker, and this activity won't be marked. 😊

Like, Olivia CSV

Next page

Activity (continue)

Please do the following activities and fill out the Padlet board as much as you can.

1) Download “financial 2023.csv” from the BB week3 lecture. Don’t click on the file in BB, just download it.

2) Open the CSV file in the **Notepad**

3) Try to explore the structure of the file, how many fields you can detect? How many records? What is the delimiter?

4) What is the data type of each field?

5) Now try to open the file in the Excel and see if your answer is correct.

Next page

Activity (continue)

1) Open the QC1.json file in the **Notepad**.

2) Try to explore the structure of the file, Can you detect first level keys of the dataset?

3) How many nested dicts you can detect?

Please open the Padlet page and complete the activities.

(the link is in the Bb, week3)

padlet