

# Data Science Life Cycle

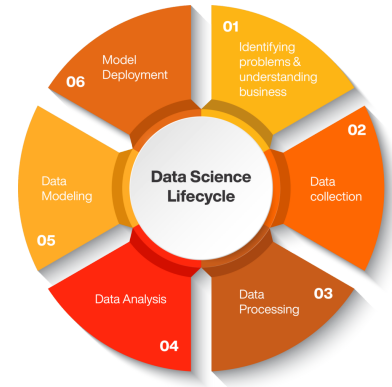
## What is Data Science?

The Study of data in order to ro make more meaningful conclusions

## The difference between Data and Information

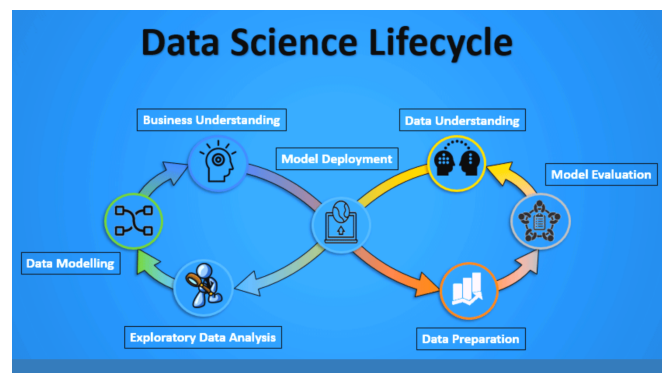
Data: Is unprocessed raw input

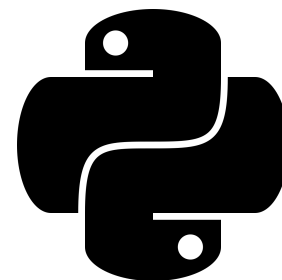
Information: Is processed data that can be understood



## Data science Life Cycle steps

1. Question
2. Collection
3. Wrangling Data
4. Analyze
5. Visualize
6. Communicate





# Python Fundamentals

## Collection Types

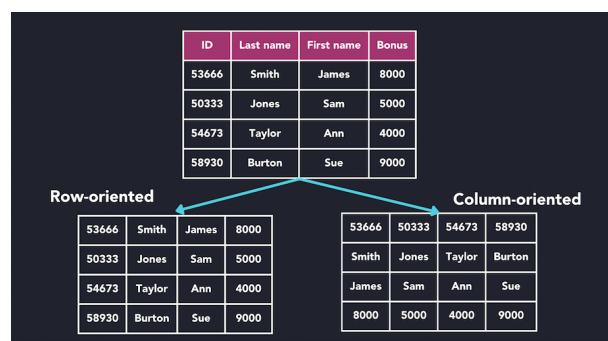
Very useful in data science because datasets are the collection of data

### Python collection types:

- Dictionary
  - Ordered
  - Changeable
  - Duplicates not allowed
- List
  - Ordered
  - Changeable
  - Duplicates allowed
- Tuple
  - Unordered
  - Unchangeable
  - duplicates not allowed
- Set
  - Unordered
  - Unchangeable
  - duplicates allowed

## Code Representation Datasets

- Column- Oriented:
  - Grouping by features
- Row-oriented:
  - Grouping by a single observations



## Indexing

- To access values, we need to INDEX

Type	Indexing Pattern
List	name[index]
Dictionary	name[key]
Set	For loop (next slide)
Tuple	Name [index]

## Iteration

- You can repeat processes with loops or recursion in python

Python loop types: ([https://www.w3schools.com/python/python\\_ref\\_dictionary.asp](https://www.w3schools.com/python/python_ref_dictionary.asp))

- ❖ For loop

```
for thing in collection:
    statements
```

- ❖ While loop

```
while condition:
    statements
```

## Useful Methods

### Dictionaries:

- values()
- items()
- keys()

### Lists:

- len()
- append()
- sort()

### Other:

- range()
- print()
- split()
- type()
- int()
- str()

# Pandas Fundamentals

## Central Tendency

(Reference crash course statistics)

### Measures of central tendency:

Mean, Median, Mode are statistical measures that help us describe the behavior of  
**a collection of data points**

### Definitions

- Mean
  - Weight tendency for things to occur in a data set
  - Affected by outliers
- Median
  - The middle of sorted data
  - Does Not use all data points
  - Less affected by outliers
- Mode
  - The most frequently/ popular occurring value in data set
  - Most helpful with moderately large data sets
- Central tendency
  - The summarized version of data set that is based in the middle of the data
- Normal distribution
  - Not Skewed
  - Symmetrical
- Skewed Distribution
  - The median is nearly identical but the mean is pulled toward the skewed data

## DataFrames (Data Table)

- A Pandas Object is used to store data
- Info is org in rows + columns
- Simplified object data

## Selection

- The process of accessing a subset of a dataframe
  - Using Iloc - Index Label
  - Using loc - Label
    - EX: df = pd.DataFrame(data)

Df.loc[1:0, ["A", "C"]]

Row, Column

## Filtering

- Selecting values of a dataset where certain conditions are true
  - df[condition]
  - If you want compound data must use & for and and / for or

## Combining Data Frames

- Three techniques
  - **Concatenate:**
    - Natively combines along an axis
  - **Merge:**
    - Combine through shared column
  - **Join**
    - Combine using shared indices
      - Inner join:
        - Only keeps things that are shared
      - Left outer join:
        - Only keeps the extra (different) data in the left side not the right side
      - Right outer Join:
        - Only keeps the extra (different) data in the right side not the left side
      - Total outer join:
        - Keeps everything

```
data = {
    "A": [1, 2, 3],
    "B": [4, 5, 6],
    "C": [7, 8, 9]
}

df = pd.DataFrame(data)

evens = df[df.iloc[:, :] % 2 == 0]
evens
```

	A	B	C
0	NaN	4.0	NaN
1	2.0	NaN	8.0
2	NaN	6.0	NaN

