University of Sulaimani
College of Science
Computer Department
4th Stage

# Data Science Management

## Mathematics and Statistics for Data Science

Class 2 and 3

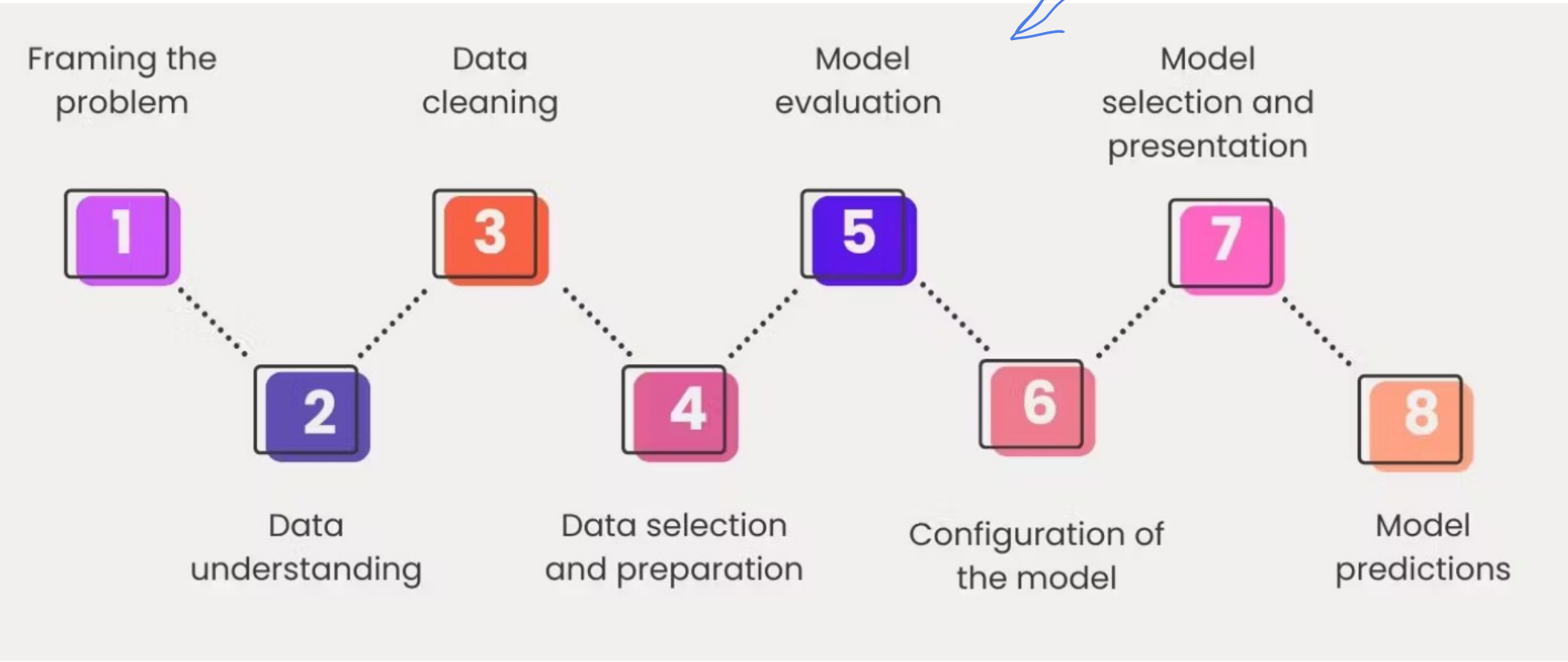
Theoretical and practical lectures

*Assist. Prof. Dr. Miran Taha Abdullah*
*2025-2026*

# Class agenda

- **Introduction to Mathematics for Data Science**
    - Importance of mathematical concepts in analyzing and modeling data
- **Statistics for Data Science**
    - **Role of Statistics in Data Science**
        - Descriptive vs Inferential Statistics
        - Importance of statistical analysis in drawing insights from data
- **Vectors, Arrays, and Matrices**
    - **Introduction to Vectors and Arrays**

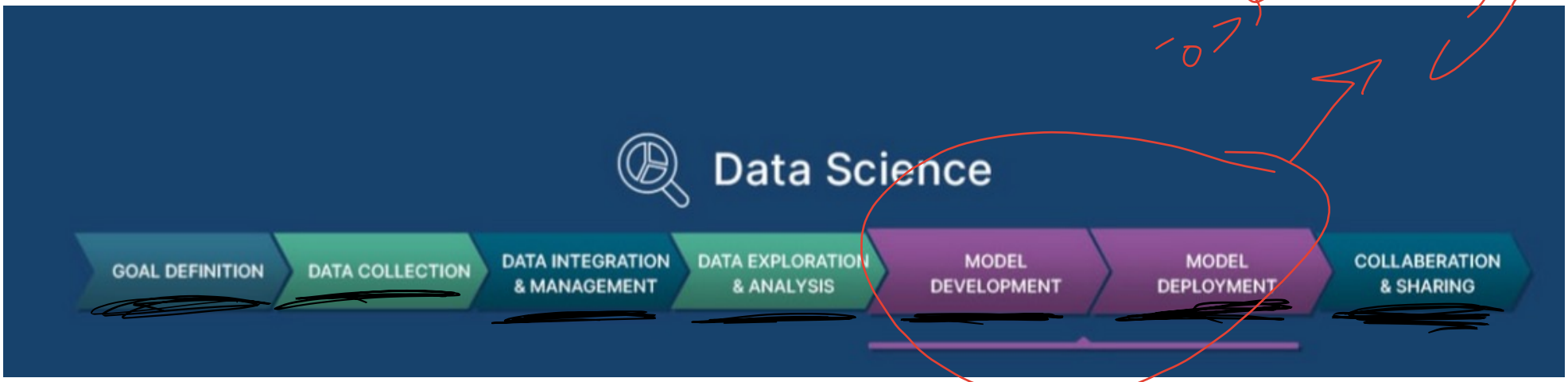
A Data Scientist must find **patterns** within the **data**. **Before find the patterns**, Data Scientist must organize the data in a **standard format**.

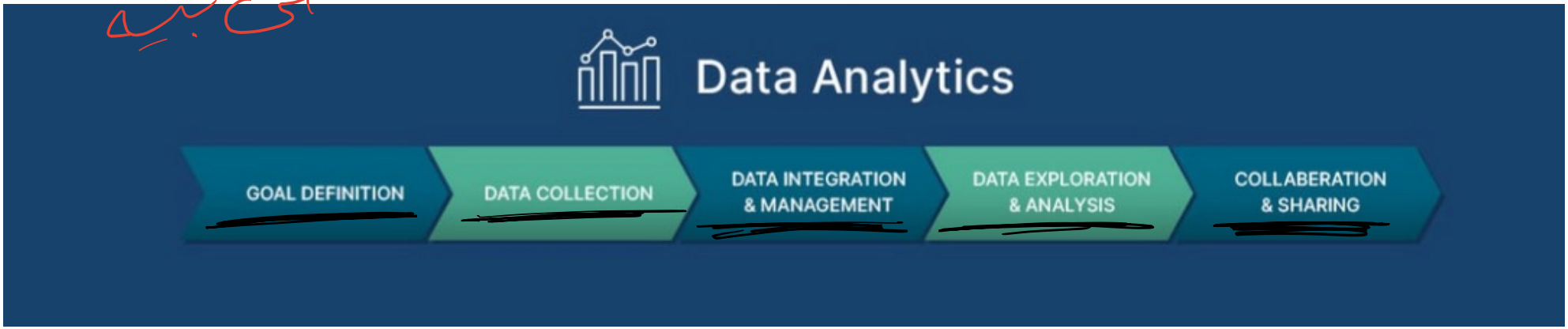



why ?

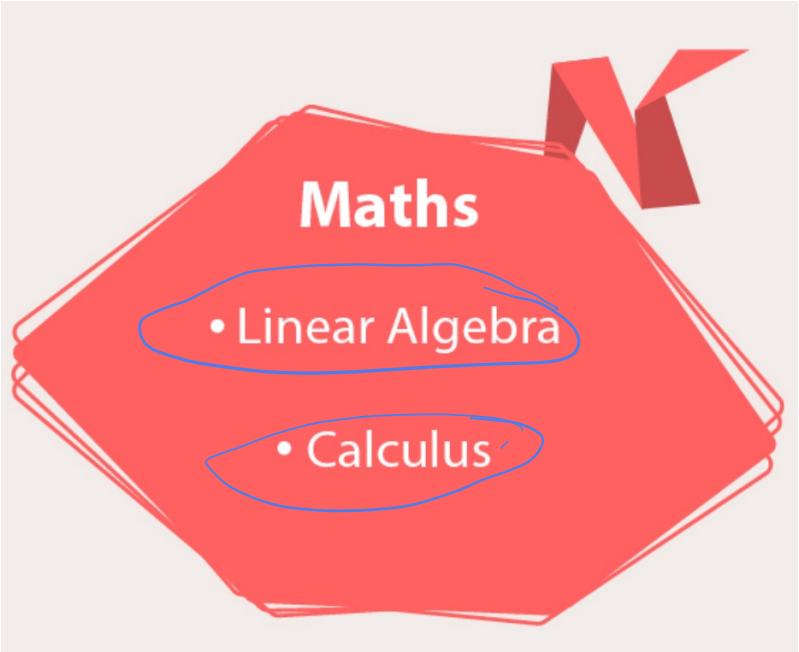
This step (**standard format**) is crucial because raw or unstructured data can introduce noise and inconsistencies that hinder analysis

- How do mathematics and statistics contribute to the field of data science?
- How can statistics and mathematical functions be used to make predictions in data science?
- What are some examples of methods or algorithms that rely on these techniques?


Data Science
GOAL DEFINITION
DATA COLLECTION
DATA INTEGRATION & MANAGEMENT
DATA EXPLORATION & ANALYSIS
MODEL DEVELOPMENT
MODEL DEPLOYMENT
COLLABERATION & SHARING


Data Analytics
GOAL DEFINITION
DATA COLLECTION
DATA INTEGRATION & MANAGEMENT
DATA EXPLORATION & ANALYSIS
COLLABERATION & SHARING

**Mathematics** and **Statistics** are two of the most important concepts of data science. Data Science revolves around these two fields and draws their concepts to operate on the data.




# 1) Mathematics for Data Science

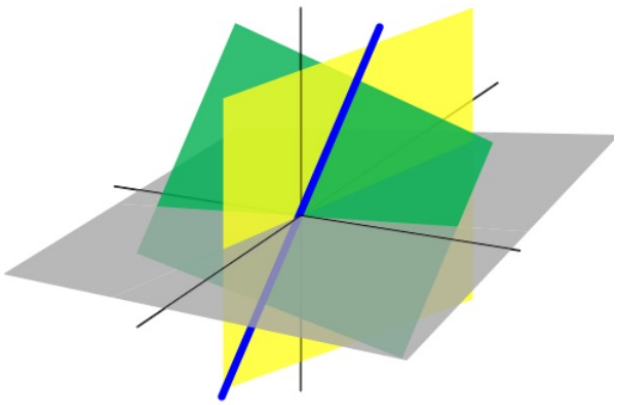
There are two main components of mathematics that contribute to Data Science namely – ***Linear Algebra and Calculus.***

## 1.1. Linear Algebra

Linear Algebra is designed to solve problems of linear equations. These equations can sometimes contain higher dimension variables.
Linear algebra that deals with vectors and matrices and, more generally, with vector spaces and linear transformations.

**Example**

$$a_1 x_1 + \cdots + a_n x_n = b,$$



لماذا نضع البيانات في matrix كالمود organize والسرعة (Index) و (optimaze)

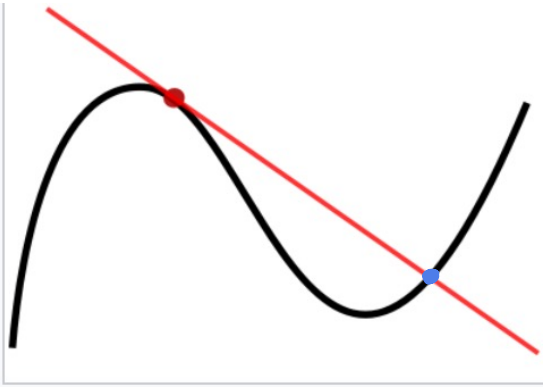
ما الفرق بين vecters و matrix؟ ما الفرق بين vecters و loop؟

# 1.2 Calculus

Another important requirement of Maths for Data Science is calculus. Calculus is used essentially in **optimization techniques**.

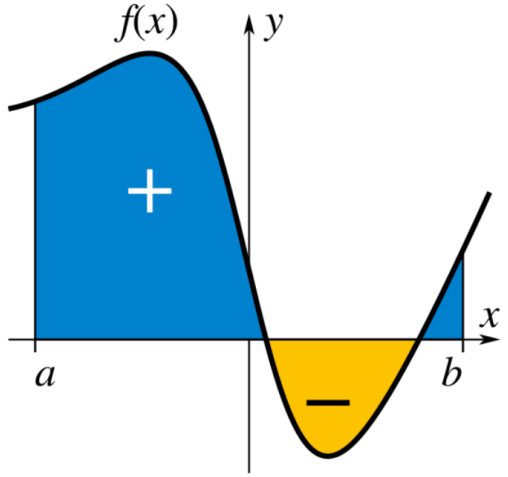
حساب التفاضل و التكامل

## 1.2.1 Differential Calculus



Differential Calculus studies the rate at which the quantities change. Derivates are most widely used for finding the maxima and minima of the functions. Derivates are used in optimization techniques where we have to find the minima in order to minimize the error function.

## 1.2.2 Integral Calculus




Integral Calculus is the mathematical study of the accumulation of quantities and for finding the area under the curve. Integrals are further divided into definite integrals and indefinite integrals. Integration is most widely used in computing probability density functions and variance of the random variable.

# Example on linear Algebra and Calculas

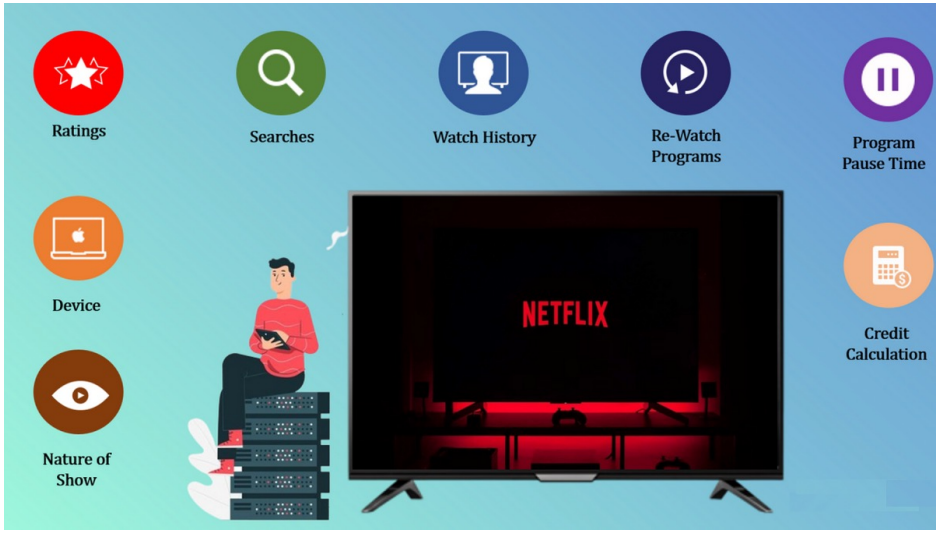
In recommender systems, especially in platforms like Netflix or Amazon, linear algebra is heavily used for **matrix factorization** techniques to *predict user preferences.*

**Problem**: Given a user-item interaction matrix (e.g., users and movies), where each entry represents how much a user likes a movie (rating), the goal is to predict missing values (i.e., unrated movies) based on existing ratings.

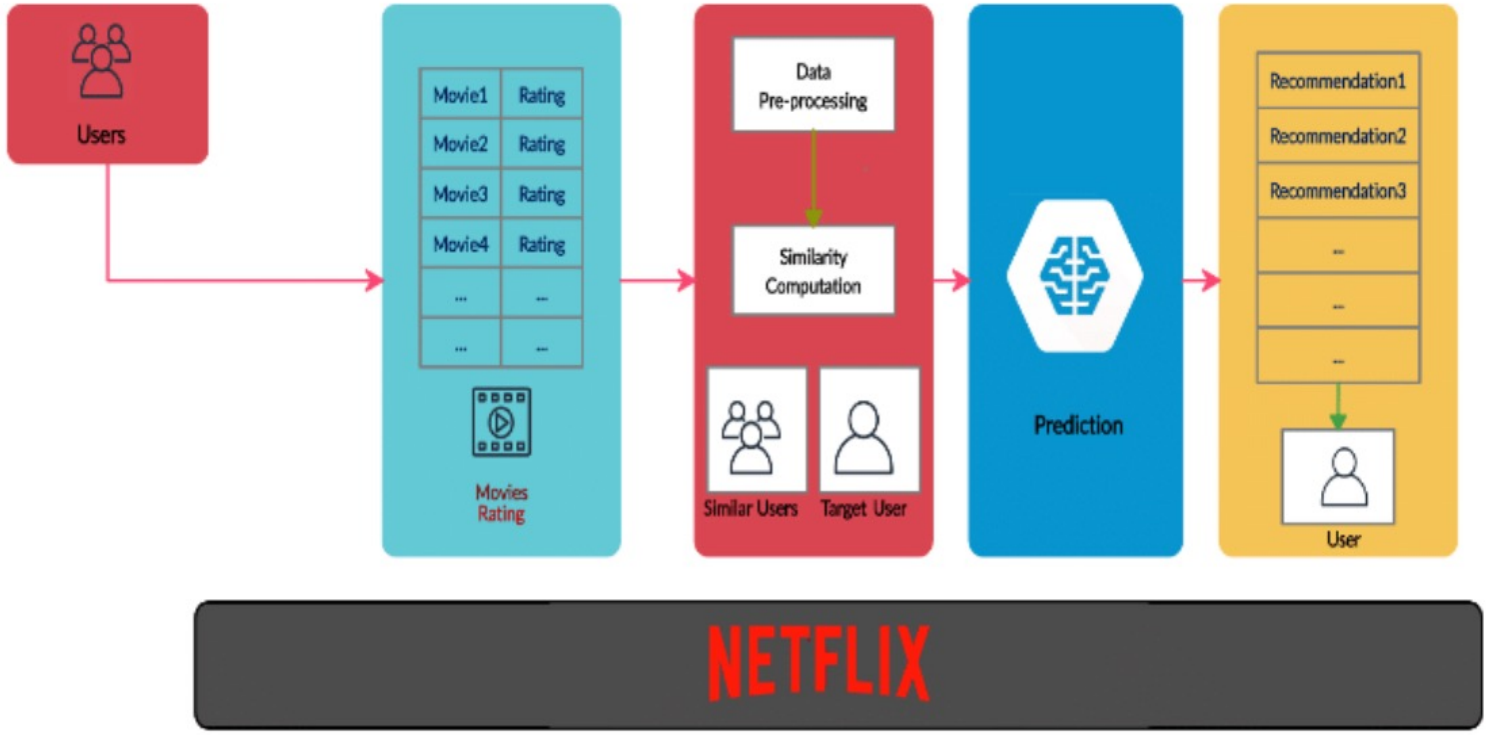
Rating Matrix $R \approx U \cdot V^T$

**U (user matrix)**: This represents characteristics of the users, like preferences or tastes.

**V (item matrix)**: This represents characteristics of the items (movies) based on how users rate them.




if User A has rated 5 movies but hasn't rated a particular movie, the recommender system will predict what their rating for that movie might be.




In a recommender system, **matrix factorization** is used to predict missing ratings in a user-item matrix. By breaking the matrix into two smaller matrices that represent users' preferences and items' features, the system can efficiently predict what movies a user might like

**Calculus** plays a critical role in optimizing the model by minimizing the error between the predicted values and actual outcomes. This process is done using **gradient descent**.

**Problem**: You want to fit a line to data points to predict a continuous target variable (e.g., predicting house prices based on features like *square footage, number of rooms, etc.*).
The objective is to find the best-fitting line that minimizes the difference between the **predicted and actual values.**

$$\text{Price} = m \times \text{Size} + b$$

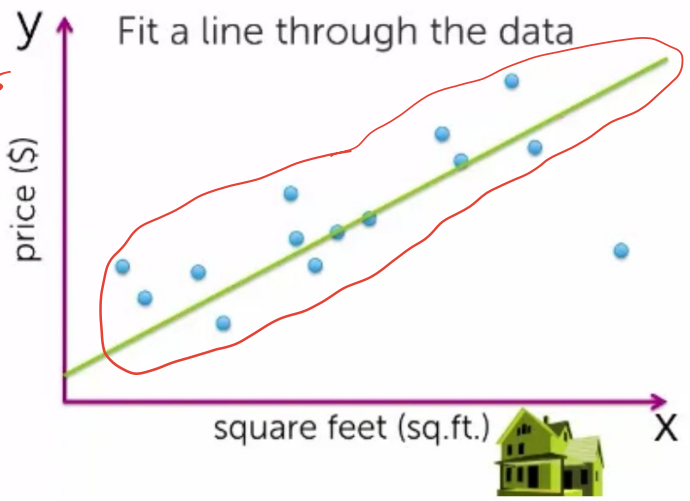
**Price** is the price of the house.

**Size** is the size of the house (in square feet).

**m** is the **slope** of the line (how much the price changes with each square foot).

**b** is the **intercept** (the price of the house when the size is 0).

يسألك بالأمتحان
كيشنو العلاقة
بهذا الرسم



# 2) Statistic for Data Science

Statistics is a set of mathematical methods and tools that enable us to answer important questions about data.

**Statistic** is divided into two categories:

**2.1. Descriptive Statistics** - this offers methods to summarise data by transforming **raw observations** into **meaningful information** that is easy to **interpret** and **share**.

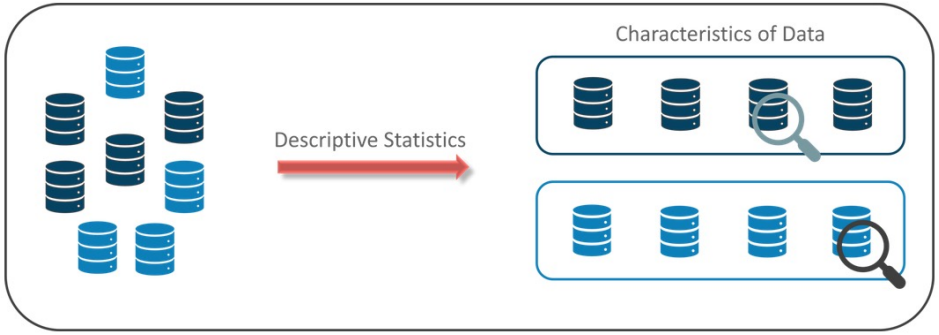
*Techniques* : **Normal Distribution, Central Tendency, Skewness & Kurtosis, Variability**

**2.2 Inferential Statistics -** this offers methods to **study experiments** done on small samples of data and chalk out (Apply) the inferences (conclusion) to the entire population (entire domain).

*Techniques :* **Central Limit Theorem, Hypothesis Testing, ANOVA, Qualitative Data Analysis**

# Descriptive Statistics

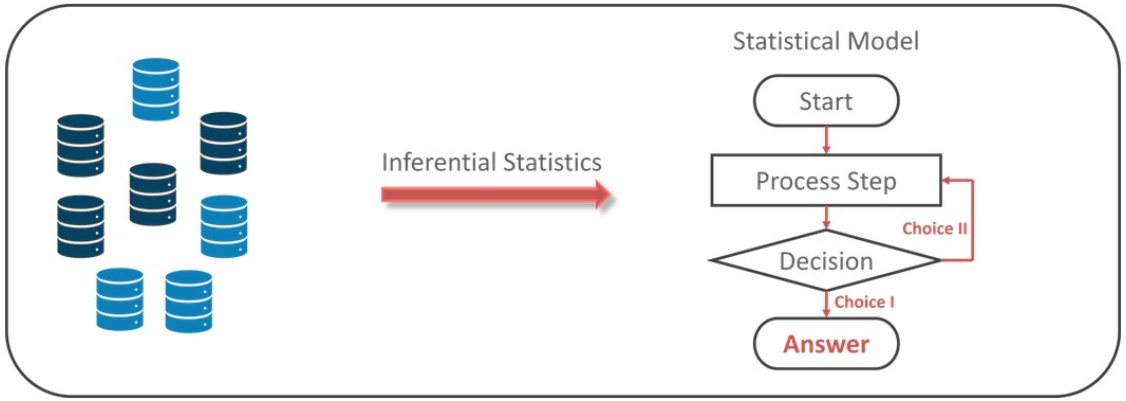
Descriptive Statistics uses the data to provide descriptions of the population, either through numerical calculations or graphs or tables. Descriptive Statistics helps organize data and focuses on the characteristics of data providing parameters. Descriptive statistics such as mean, median, and mode.

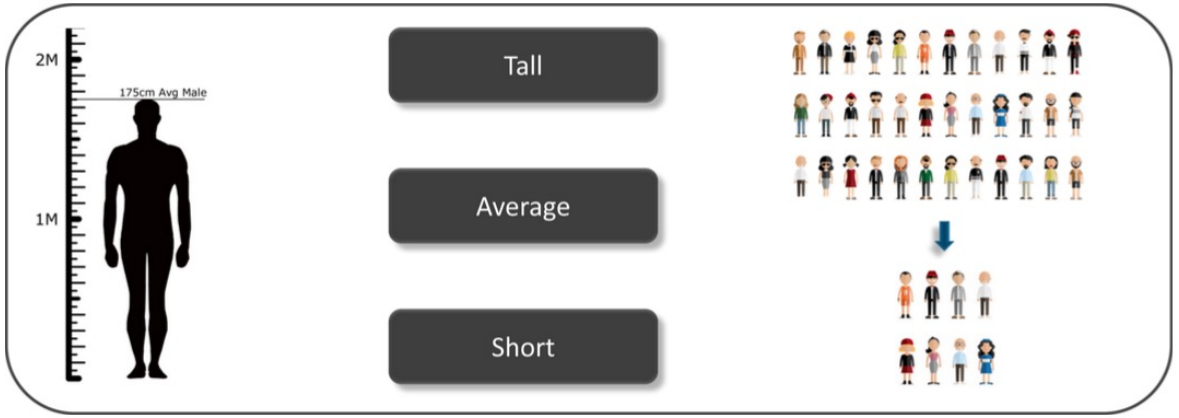



Example: Suppose you want to study the average height of students in a classroom, in descriptive statistics you would record the heights of all students in the class and then you would find out the maximum, minimum and average height of the class.

# Inferential Statistics

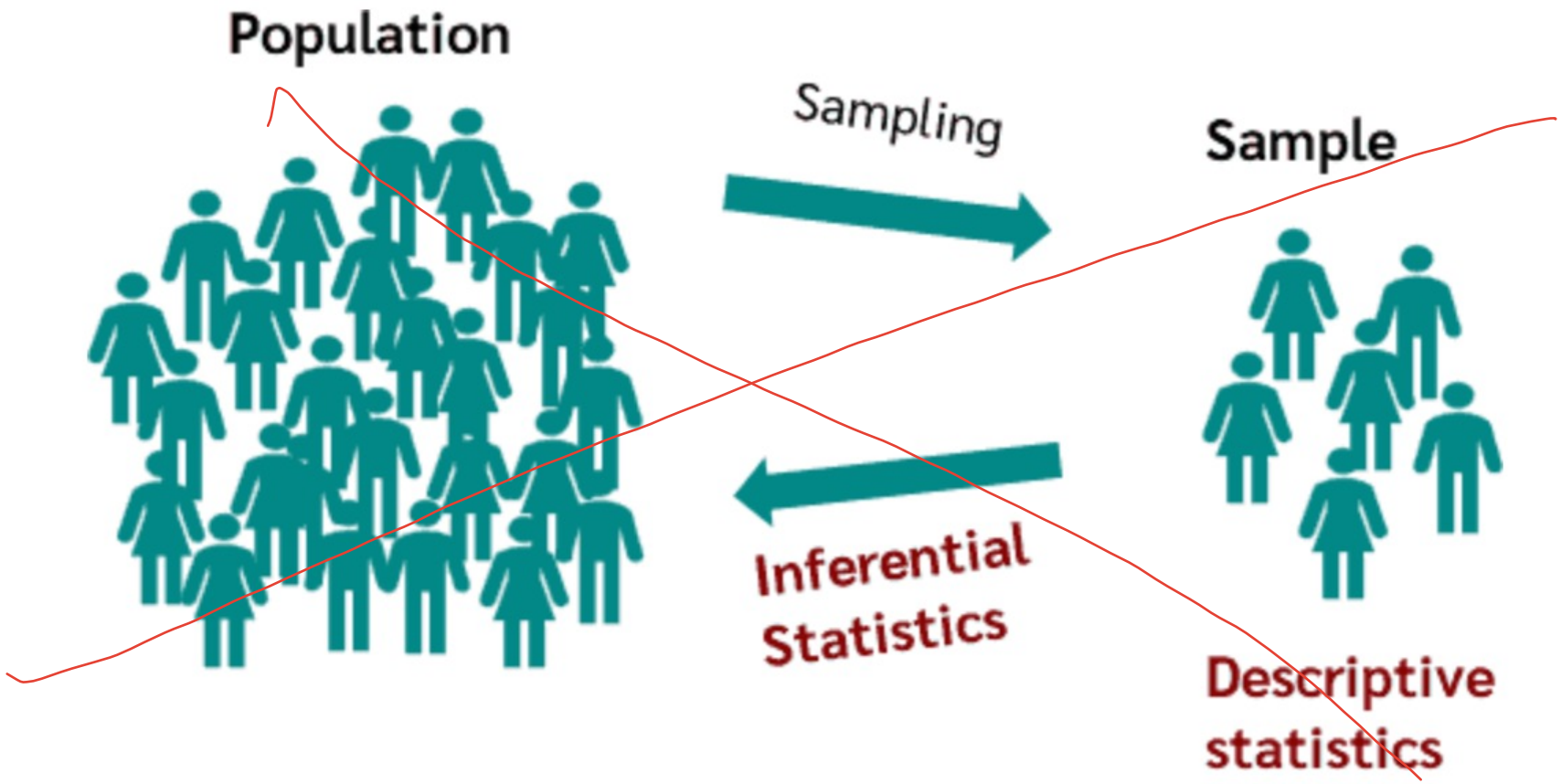
Inferential Statistics makes inferences and predictions about a population based on a sample of data taken from the population in question. Inferential statistics generalizes a large data set and applies probability to arrive at a conclusion. It allows you to infer parameters of the population based on sample stats and build models on it.




Example:

So, if we consider the same example of finding the average height of students in a class, in Inferential Statistics, you will take a sample set of the class, which is basically a few people from the entire class. You already have had grouped the class into tall, average and short. In this method, you basically build a statistical model and expand it for the entire population in the class.

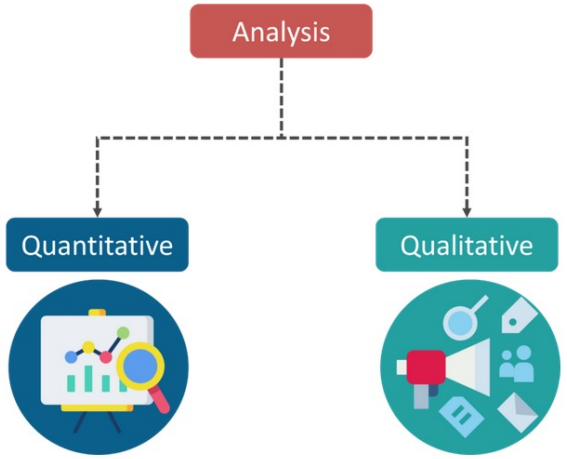

Population
Sampling
Sample
Inferential
Statistics
Descriptive
statistics

# Types Of Analysis

An analysis of any event can be done in one of two ways

1. **Quantitative Analysis:** Quantitative Analysis or the Statistical Analysis is the science of collecting and interpreting data with numbers and graphs to identify patterns and trends.

2. **Qualitative Analysis:** Qualitative or Non-Statistical Analysis gives generic information and uses text, sound and other forms of media to do so.



## Descriptive Statistics

1) Mean, Median, Mode
2) IQR, percentiles
3) Std deviation and Variance
4) Normal Distribution
5) Z-statistics and T-statistics
6) correlation and linear regression

## Inferential Statistics:

1) Sampling distributions
2) confidence interval
3) chi-square test
4) Advanced regression
5) ANOVA

# What is Vector?

- Vectors, matrices, and arrays of higher dimensions are essential tools in numerical computing.
- When a computation must be repeated for a set of input values, it is natural and advantageous to represent the data as arrays and the computation in terms of array operations.
- Vectorized computing eliminates the need for many explicit loops over the array elements by applying batch operations on the array data.
- Vectorized computations can therefore be significantly faster than sequential element-by-element computations.

abs() which returns the absolute value of a number

**Modules** are Python .py files that consist of Python code. Any Python file can be referenced as a module. A Python file called hello.py has the module name of hello that can be imported into other Python files or used on the Python command line interpreter.

## Attributes of the ndarray Class

| Attribute | Description |
|---|---|
| `Shape` | A tuple that contains the number of elements (i.e., the length) for each dimension (axis) of the array. |
| `Size` | The total number elements in the array. |
| `Ndim` | Number of dimensions (axes). |
| `nbytes` | Number of bytes used to store the data. |
| `dtype` | The data type of the elements in the array. |

## Numerical Data Types Available in NumPy

| dtype | Variants | Description |
|---|---|---|
| `int` | `int8, int16, int32, int64` | Integers |
| `uint` | `uint8, uint16, uint32, uint64` | Unsigned (nonnegative) integers |
| `bool` | `Bool` | Boolean (True or False) |
| `float` | `float16, float32, float64, float128` | Floating-point numbers |
| `complex` | `complex64, complex128, complex256` | Complex-valued floating-point numbers |

| Data type | Description |
|---|---|
| `bool_` | Boolean (True or False) stored as a byte |
| `int_` | Default integer type (same as C `long`; normally either `int64` or `int32`) |
| `intc` | Identical to C `int` (normally `int32` or `int64`) |
| `intp` | Integer used for indexing (same as C `ssize_t`; normally either `int32` or `int64`) |
| `int8` | Byte (–128 to 127) |
| `int16` | Integer (–32768 to 32767) |
| `int32` | Integer (–2147483648 to 2147483647) |
| `int64` | Integer (–9223372036854775808 to 9223372036854775807) |
| `uint8` | Unsigned integer (0 to 255) |
| `uint16` | Unsigned integer (0 to 65535) |
| `uint32` | Unsigned integer (0 to 4294967295) |
| `uint64` | Unsigned integer (0 to 18446744073709551615) |
| `float_` | Shorthand for `float64` |
| `float16` | Half-precision float: sign bit, 5 bits exponent, 10 bits mantissa |
| `float32` | Single-precision float: sign bit, 8 bits exponent, 23 bits mantissa |
| `float64` | Double-precision float: sign bit, 11 bits exponent, 52 bits mantissa |
| `complex_` | Shorthand for `complex128` |
| `complex64` | Complex number, represented by two 32-bit floats |
| `complex128` | Complex number, represented by two 64-bit floats |

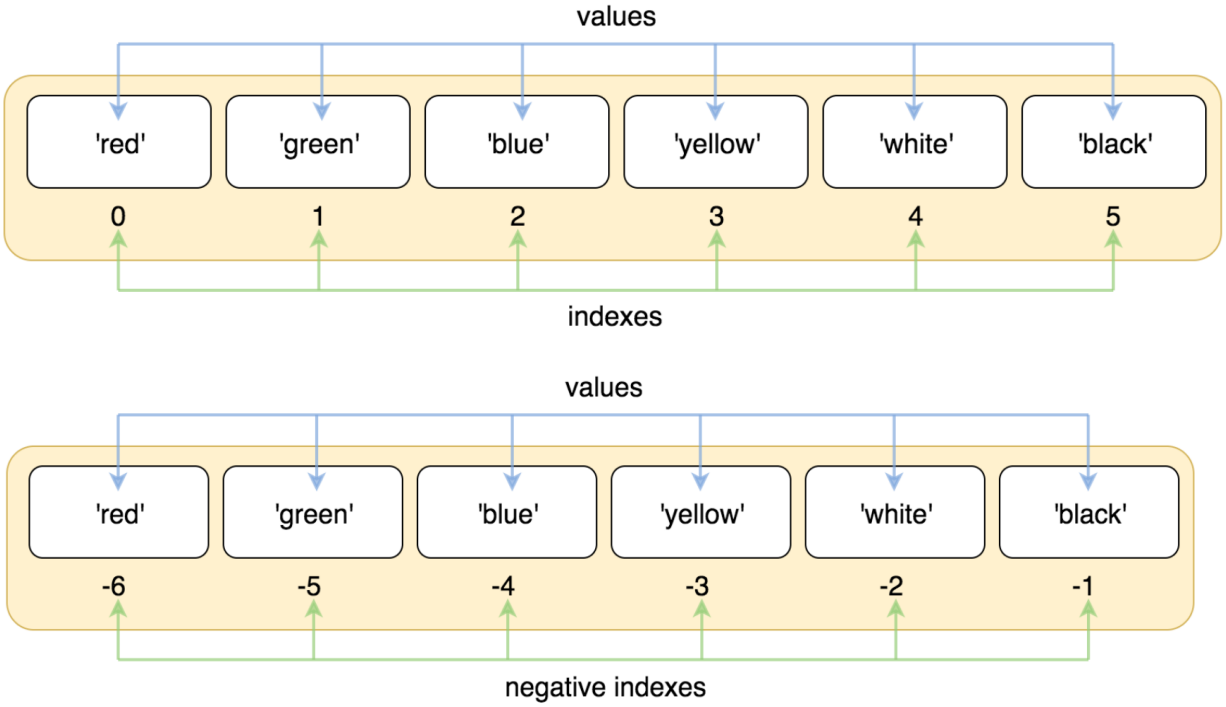
## NumPy Functions for Generating Arrays

| Function Name | Type of Array |
|---|---|
| `np.array` | Creates an array for which the elements are given by an array-like object, which, for example, can be a (nested) Python list, a tuple, an iterable sequence, or another `ndarray` instance. |
| `np.zeros` | Creates an array with the specified dimensions and data type that is filled with zeros. |
| `np.ones` | Creates an array with the specified dimensions and data type that is filled with ones. |
| `np.diag` | Creates a diagonal array with specified values along the diagonal and zeros elsewhere. |
| `np.arange` | Creates an array with evenly spaced values between the specified start, end, and increment values. |
| `np.linspace` | Creates an array with evenly spaced values between specified start and end values, using a specified number of elements. |

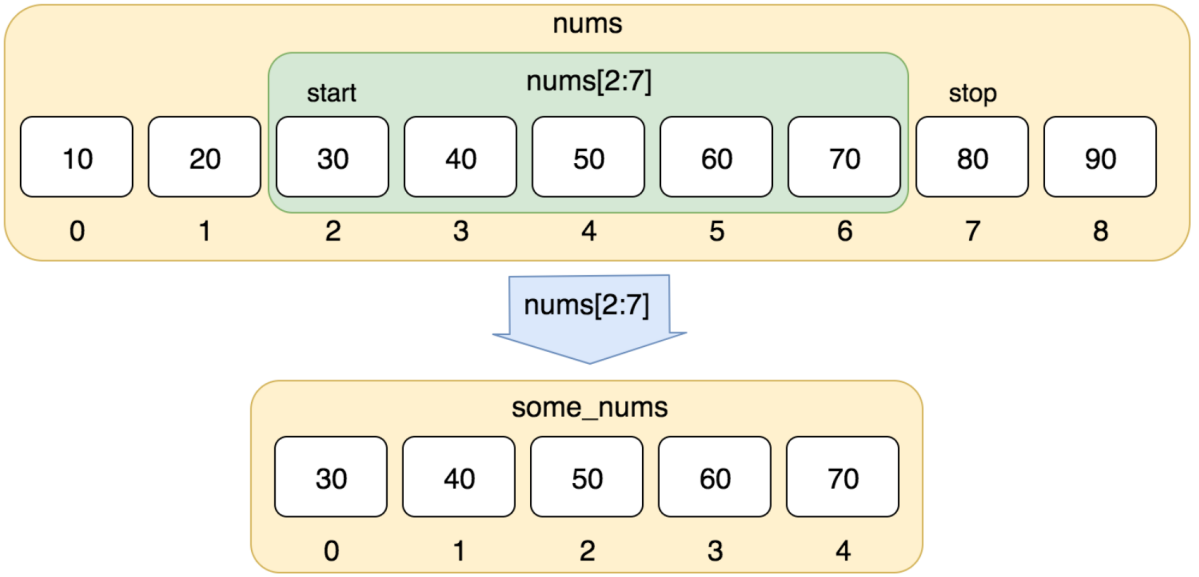
| | |
|---|---|
| `np.logspace` | Creates an array with values that are logarithmically spaced between the given start and end values. |
| `np.meshgrid` | Generates coordinate matrices (and higher-dimensional coordinate arrays) from one-dimensional coordinate vectors. |
| `np.fromfunction` | Creates an array and fills it with values specified by a given function, which is evaluated for each combination of indices for the given array size. |
| `np.fromfile` | Creates an array with the data from a binary (or text) file. NumPy also provides a corresponding function `np.tofile` with which NumPy arrays can be stored to disk and later read back using `np.fromfile`. |
| `np.genfromtxt`, `np.loadtxt` | Create an array from data read from a text file, for example, a comma-separated value (CSV) file. The function `np.genfromtxt` also supports data files with missing values. |
| `np.random.rand` | Generates an array with random numbers that are uniformly distributed between 0 and 1. Other types of distributions are also available in the `np.random` module. |

## 4. Indexing and Slicing

- **Indexing** means referring to an element of an iterable by its position within the iterable. Each of a string's characters corresponds to an index number and each character can be accessed using their index number.



**Slicing** in Python is a feature that enables accessing parts of sequence. In slicing string (or number), we create a substring (or sub-number)., which is essentially a string (or a number) that exists within another string (or number). We use slicing when we require a part of string (or number) and not the complete string.



# Examples of Array Indexing and Slicing Expressions

| Expression | Description |
|---|---|
| `a[m]` | Select element at index *m*, where *m* is an integer (start counting form 0). |
| `a[-m]` | Select the *n*th element from the end of the list, where *n* is an integer. The last element in the list is addressed as –1, the second to last element as –2, and so on. |
| `a[m:n]` | Select elements with index starting at *m* and ending at *n* – 1 (*m* and *n* are integers). |
| `a[:]` or `a[0:-1]` | Select all elements in the given axis. |
| `a[:n]` | Select elements starting with index 0 and going up to index *n* – 1 (integer). |
| `a[m:]` or `a[m:-1]` | Select elements starting with index *m* (integer) and going up to the last element in the array. |
| `a[m:n:p]` | Select elements with index *m* through *n* (exclusive), with increment *p*. |
| `a[::-1]` | Select all the elements, in reverse order. |

# 5. Reshaping and Resizing

The reshape() function is used to give a new shape to an array without changing its data.

**numpy.reshape()**

New_Array = np.reshape(x,new_resize) → X= curent Array

The resize() function is used to create a new array with the specified shape. If the new array is larger than the original array, then the new array is filled with repeated copies of X.

**numpy.resize()**

New_Array = np.resize(x,new_shape) → X= curent Array

| Function/Method | Description |
|---|---|
| `np.reshape`, `np.ndarray.reshape` | Reshape an N-dimensional array. The total number of elements must remain the same. |
| `np.ndarray.flatten` | Creates a copy of an N-dimensional array, and reinterpret it as a one-dimensional array (i.e., all dimensions are collapsed into one). |
| `np.ravel`, `np.ndarray.ravel` | Create a view (if possible, otherwise a copy) of an N-dimensional array in which it is interpreted as a one-dimensional array. |
| `np.squeeze` | Removes axes with length 1. |
| `np.expand_dims`, `np.newaxis` | Add a new axis (dimension) of length 1 to an array, where `np.newaxis` is used with array indexing. |
| `np.transpose`, `np.ndarray.transpose`, `np.ndarray.T` | Transpose the array. The transpose operation corresponds to reversing (or more generally, permuting) the axes of the array. |
| `np.hstack` | Stacks a list of arrays horizontally (along axis 1): for example, given a list of column vectors, appends the columns to form a matrix. |
| `np.vstack` | Stacks a list of arrays vertically (along axis 0): for example, given a list of row vectors, appends the rows to form a matrix. |
| `np.dstack` | Stacks arrays depth-wise (along axis 2). |
| `np.concatenate` | Creates a new array by appending arrays after each other, along a given axis. |

| Function/Method | Description |
| --- | --- |
| `np.resize` | Resizes an array. Creates a new copy of the original array, with the requested size. If necessary, the original array will be repeated to fill up the new array. |
| `np.append` | Appends an element to an array. Creates a new copy of the array. |
| `np.insert` | Inserts a new element at a given position. Creates a new copy of the array. |
| `np.delete` | Deletes an element at a given position. Creates a new copy of the array. |

# 6. Vectorized Expressions

The purpose of storing numerical data in arrays is to be able to process the data with concise vectorized expressions that represent batch operations that are applied to all elements in the arrays.

# Vectorization in Python

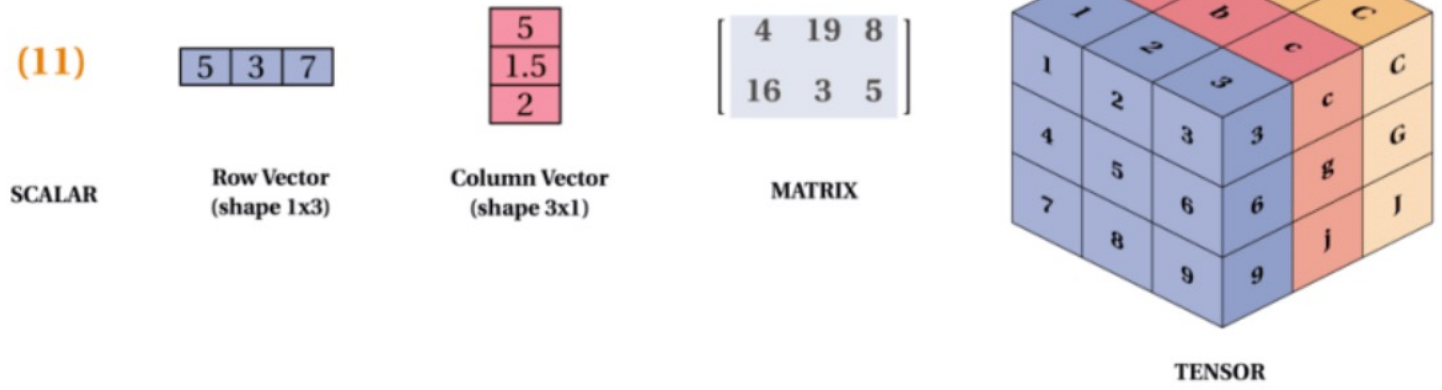
- Vectorization is a technique of implementing array operations without using for loops. Instead, we use functions defined by various modules which are highly optimized that reduces the running and execution time of code. Vectorized array operations will be faster than their pure Python equivalents, with the biggest impact in any kind of numerical computations.

Elementwise Functions

| NumPy Function | Description |
|---|---|
| `np.cos, np.sin, np.tan` | Trigonometric functions. |
| `np.arccos, np.arcsin, np.arctan` | Inverse trigonometric functions. |
| `np.cosh, np.sinh, np.tanh` | Hyperbolic trigonometric functions. |
| `np.arccosh, np.arcsinh, np.arctanh` | Inverse hyperbolic trigonometric functions. |
| `np.sqrt` | Square root. |
| `np.exp` | Exponential. |
| `np.log, np.log2, np.log10` | Logarithms of base e, 2, and 10, respectively. |

# 7. Matrix and Vector Operations

We have so far discussed general N-dimensional arrays. One of the main applications of such arrays is to represent the mathematical concepts of vectors, matrices, and tensors, and in this use-case, we also frequently need to calculate vector and matrix operations such as scalar (inner) products, dot (matrix) products, and tensor (outer) products.

NumPy Functions for Matrix Operations

| NumPy Function | Description |
|---|---|
| `np.dot` | Matrix multiplication (dot product) between two given arrays representing vectors, arrays, or tensors. |
| `np.inner` | Scalar multiplication (inner product) between two arrays representing vectors. |
| `np.cross` | The cross product between two arrays that represent vectors. |
| `np.tensordot` | Dot product along specified axes of multidimensional arrays. |
| `np.outer` | Outer product (tensor product of vectors) between two arrays representing vectors. |
| `np.kron` | Kronecker product (tensor product of matrices) between arrays representing matrices and higher-dimensional arrays. |
| `np.einsum` | Evaluates Einstein's summation convention for multidimensional arrays. |

# End of Class 2 & 3