



Python for Data science





Hello!

Soy Gusseppe Bravo

Data scientist consultant at Inter-american Development Bank (IDB).

Machine Learning and High Performance Computing (HPC) researcher at CTIC - UNI, Peru.

Python Open Source contributor. Numerical analysis, Data science, Big data.

gbravor@uni.pe

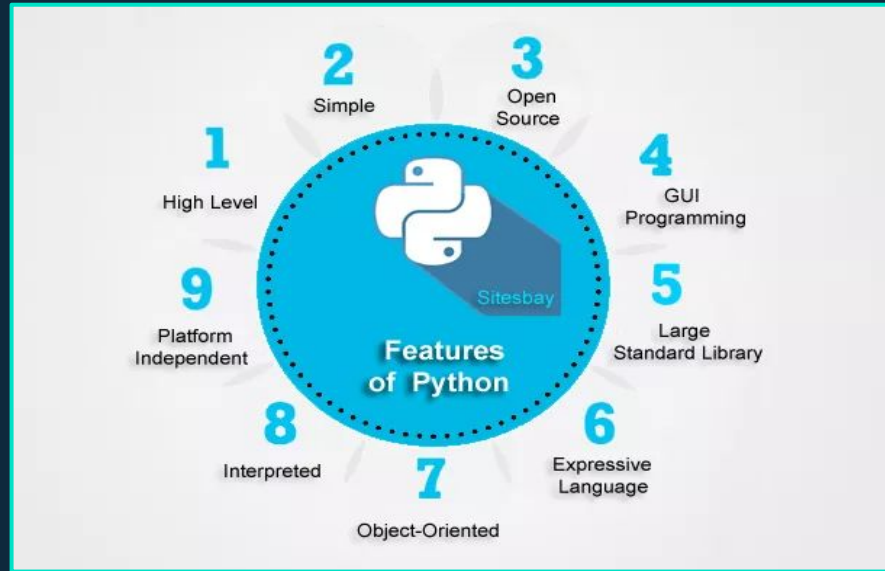


PYTHON ECOSYSTEM

1

- ◇ ¿Qué es Python ?
- ◇ ¿Qué alcance tiene Python al día de hoy?
- ◇ ¿Por qué Python en Data Science?

¿Qué es Python?



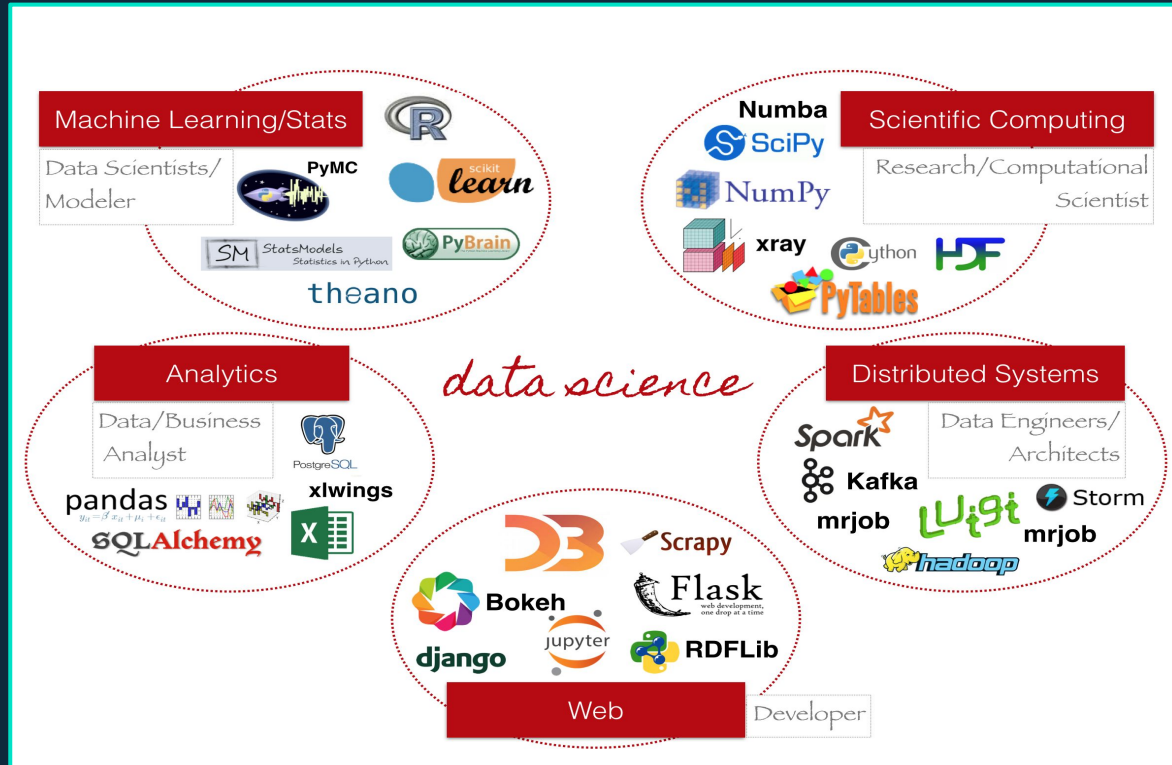
Alcance

Growth of major programming languages

Based on Stack Overflow question views in World Bank high-income countries



Python + Data science





¿Te convencí?, si no es así, mira esto...

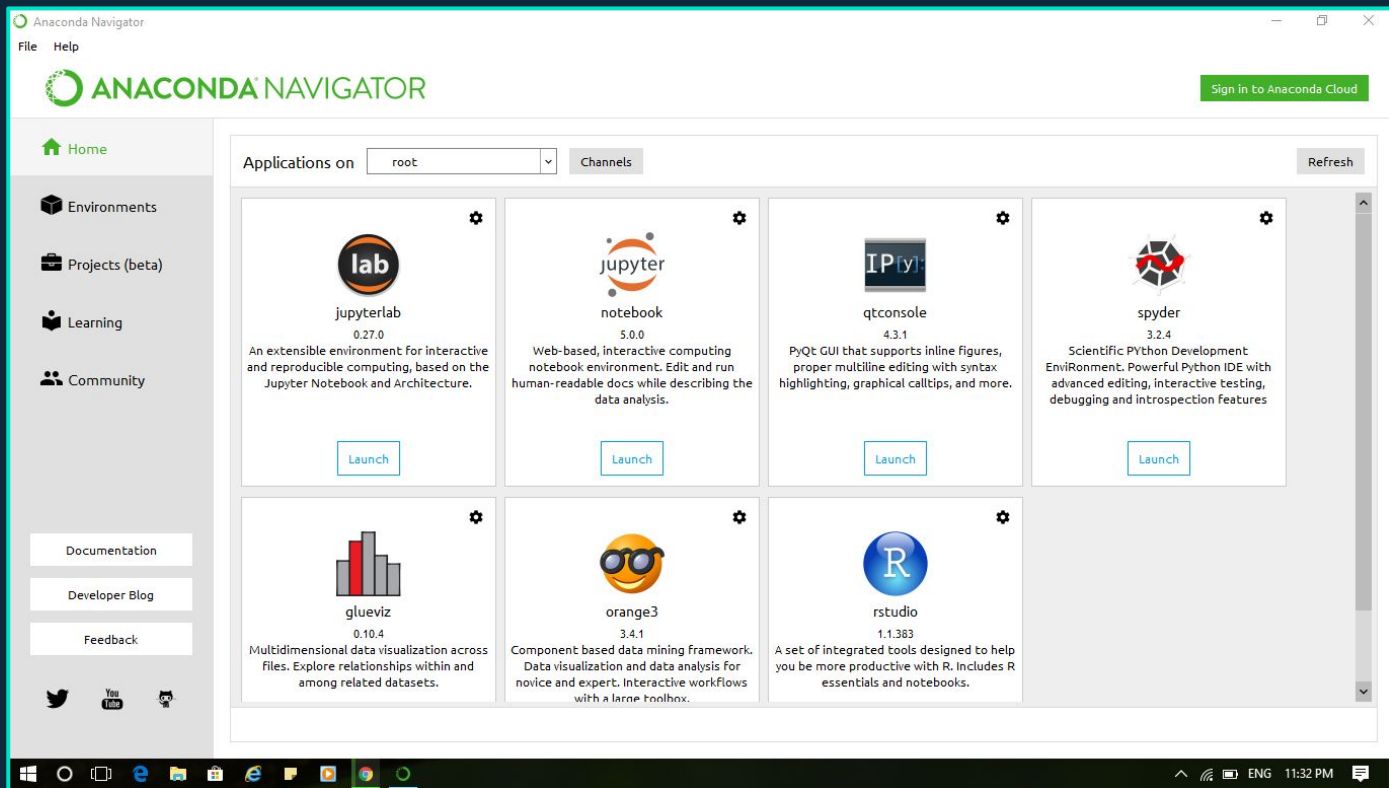
```
import turtle

star = turtle.Turtle()

for i in range(50):
    star.forward(50)
    star.right(144)

turtle.done()
```

Anaconda



PYTHON BASICS

2

- ◇ Variables, tipos de datos.
- ◇ Estructuras de datos.
- ◇ Funciones, ejecuciones.

Python cheat sheet

Python For Data Science Cheat Sheet

Python Basics

Learn More Python for Data Science Interactively at www.datacamp.com



Variables and Data Types

Variable Assignment

```
>>> x=5
>>> x
5
```

Calculations With Variables

>>> x+2	Sum of two variables
7	
>>> x-2	Subtraction of two variables
3	
>>> x*2	Multiplication of two variables
10	
>>> x**2	Exponentiation of a variable
25	
>>> x%2	Remainder of a variable
1	
>>> x/float(2)	Division of a variable
2.5	

Types and Type Conversion

str()	'5', '3.45', 'True'	Variables to strings
int()	5, 3, 1	Variables to integers
float()	5.0, 1.0	Variables to floats
bool()	True, True, True	Variables to booleans

Asking For Help

```
>>> help(str)
```

Strings

```
>>> my_string = 'thisStringIsAwesome'
>>> my_string
'thisStringIsAwesome'
```

String Operations

```
>>> my_string * 2
'thisStringIsAwesomethisStringIsAwesome'
>>> my_string + 'Innit'
'thisStringIsAwesomeInnit'
>>> 'm' in my_string
True
```

Lists

Also see NumPy Arrays

```
>>> a = 'is'
>>> b = 'nice'
>>> my_list = ['my', 'list', a, b]
>>> my_list2 = [[4,5,6,7], [3,4,5,6]]
```

Selecting List Elements

Index starts at 0

Subset >>> my_list[1] >>> my_list[-3] Slice >>> my_list[1:3] >>> my_list[1:] >>> my_list[:3] >>> my_list[:] Subset Lists of Lists >>> my_list2[1][0] >>> my_list2[1][:2]	Select item at index 1 Select 3rd last item Select items at index 1 and 2 Select items after index 0 Select items before index 3 Copy my_list my_list[list][itemOfList]
---	---

List Operations

```
>>> my_list + my_list
['my', 'list', 'is', 'nice', 'my', 'list', 'is', 'nice']
>>> my_list * 2
['my', 'list', 'is', 'nice', 'my', 'list', 'is', 'nice']
>>> my_list2 > 4
True
```

List Methods

>>> my_list.index(a) >>> my_list.count(a) >>> my_list.append('!') >>> my_list.remove('!') >>> del(my_list[0:1]) >>> my_list.reverse() >>> my_list.extend('!') >>> my_list.pop(-1) >>> my_list.insert(0, '!') >>> my_list.sort()	Get the index of an item Count an item Append an item at a time Remove an item Remove an item Reverse the list Append an item Remove an item Insert an item Sort the list
--	--

String Operations

Index starts at 0

```
>>> my_string[3]
>>> my_string[4:9]
```

String Methods

>>> my_string.upper() >>> my_string.lower() >>> my_string.count('w') >>> my_string.replace('e', 'i') >>> my_string.strip()	String to uppercase String to lowercase Count String elements Replace String elements Strip whitespaces
--	---

Libraries

Import libraries

```
>>> import numpy
>>> import numpy as np
Selective import
>>> from math import pi
```



Install Python



NumPy Arrays

Also see Lists

```
>>> my_list = [1, 2, 3, 4]
>>> my_array = np.array(my_list)
>>> my_2darray = np.array([[1,2,3], [4,5,6]])
```

Selecting NumPy Array Elements

Index starts at 0

Subset >>> my_array[1] 2 Slice >>> my_array[0:2] array([1, 2]) Subset 2D NumPy arrays >>> my_2darray[:,0] array([1, 4])	Select item at index 1 Select items at index 0 and 1 my_2darray[rows, columns]
--	--

NumPy Array Operations

```
>>> my_array > 3
array([False, False, False,  True], dtype=bool)
>>> my_array * 2
array([2, 4, 6, 8])
>>> my_array + np.array([5, 6, 7, 8])
array([6, 8, 10, 12])
```

NumPy Array Functions

>>> my_array.shape >>> np.append(other_array) >>> np.insert(my_array, 1, 5) >>> np.delete(my_array, [1]) >>> np.mean(my_array) >>> np.median(my_array) >>> my_array.corrcoef() >>> np.std(my_array)	Get the dimensions of the array Append items to an array Insert items in an array Delete items in an array Mean of the array Median of the array Correlation coefficient Standard deviation
--	--



DATA SCIENCE

3

- ◇ ¿Qué es data science ?
- ◇ ¿Dónde se aplica ?
- ◇ Data science workflow.

Data + science

Data *facts and statistics collected together for reference or analysis*

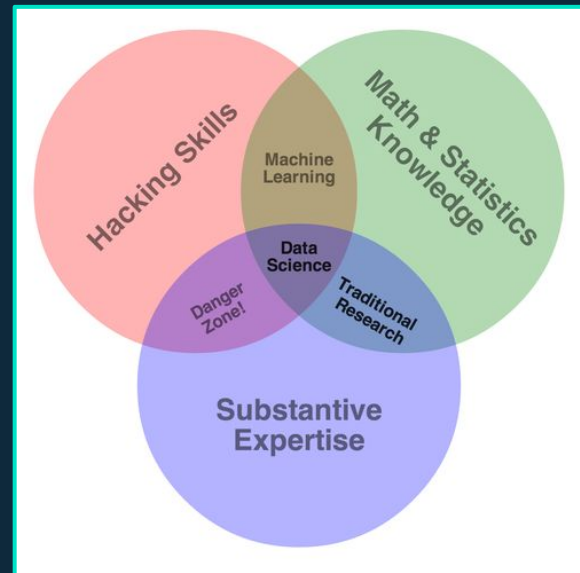
Science *The intellectual and practical activity encompassing the systematic study of the structure and behavior of the physical and natural world through observation and experiment.*

Data Science The **scientific exploration** of **data** to extract meaning or insight, and the **construction** of software systems to utilize such insight in a business context.

Data
Scientist



Someone who does **this** ...



Aplicaciones



Healthcare

- Predict diagnosis
- Prioritize screenings
- Reduce re-admittance rates



Financial services

- Fraud Detection/prevention
- Predict underwriting risk
- New account risk screens



Public Sector

- Analyze public sentiment
- Optimize resource allocation
- Law enforcement & security



Retail

- Product recommendation
- Inventory management
- Price optimization



Telco/mobile

- Predict customer churn
- Predict equipment failure
- Customer behavior analysis

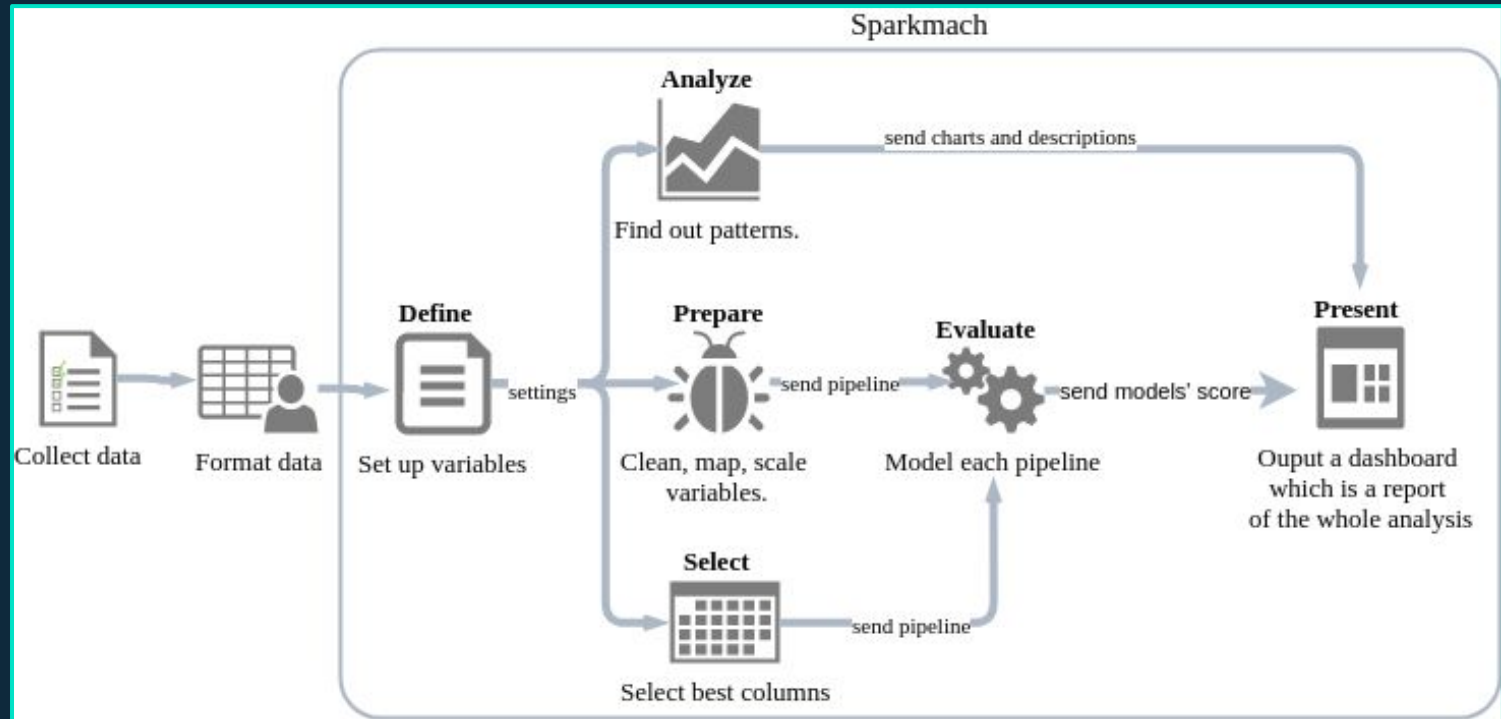


Oil & Gas

- Predictive maintenance
- Seismic data management
- Predict well production levels

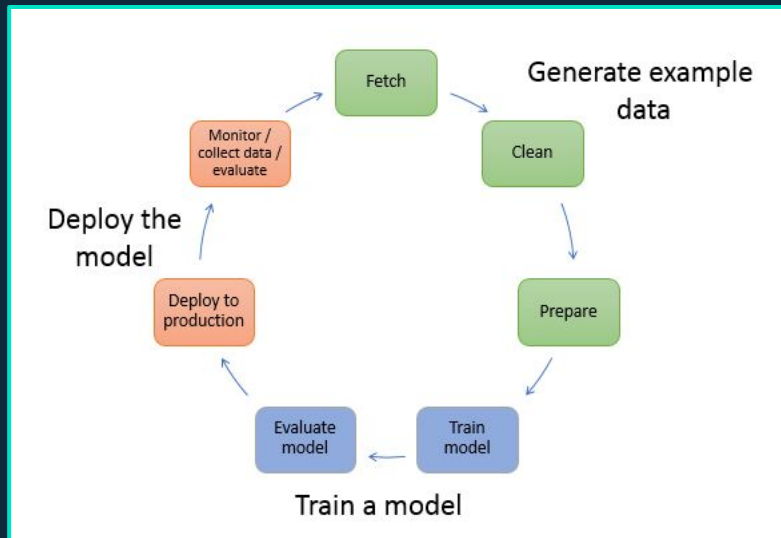


Workflow





Real life?



DATA EXTRACTION



4

- ◇ ¿ Qué datos tengo ?
- ◇ ¿ Cómo extraigo estos datos ?
- ◇ ¿ Mis datos son suficientes ?

EXPLORATORY DATA ANALYSIS



5

- ◇ ¿ Cómo es la distribución de los datos?
- ◇ ¿ Mis datos están balanceados?
- ◇ ¿ Small, medium, big data ?

PREPROCESSING



6

- ◇ Clean data.
- ◇ Normalization.
- ◇ Imputation.

FEATURE ENGINEERING



7

- ◇ ¿ Puedo combinar columnas ?
- ◇ ¿ Qué pasa si transformo o creo una columna?
- ◇ Feature extraction, selection, transformation.

MODELING



8

- ◇ ¿ Qué algoritmo necesito ?
- ◇ ¿ Cómo probar mi algoritmo?
- ◇ Training, testing, metrics.

TUNING



8

- ◇ ¿ Cómo mejorar los parámetros de mi modelo ?
- ◇ Grid Search.
- ◇ Random Search.

DEPLOYMENT



8

- ◇ ¿ Cómo uso mi modelo ?
- ◇ ¿ Cómo ajustar mi modelo ?
- ◇ Arquitectura del despliegue, tracking, A/B testing.



Thanks!

Any questions?

You can find me at:

- ◆ gbravor@uni.pe
- ◆ @gussepe-jesus-bravo-rocca-4437b6106/





Credits

Special thanks to all the people who made and released these awesome resources for free:

- ◇ Presentation template by [SlidesCarnival](#)
- ◇ Photographs by [Unsplash](#)
- ◇ Hortonworks slides

