Python for Data science

Hello!

Soy Gusseppe Bravo

Big data Research Engineer, Department of Computer Science, Barcelona Supercomputing Center, Spain (present).

Data scientist consultant at Inter-american Development Bank (IDB), Peru. (2017-2018).

Machine Learning and High Performance Computing (HPC) researcher at CTIC - UNI, Peru. (2015-2017)

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

gbravor@uni.pe, https://github.com/gusseppe

DATA SCIENCE

3

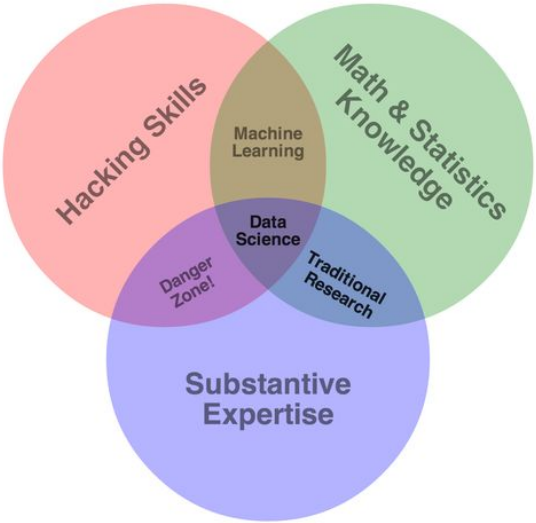
- ¿ Qué es data science ?
- ¿ Dónde se aplica ?
- Data science workflow.

Data + science

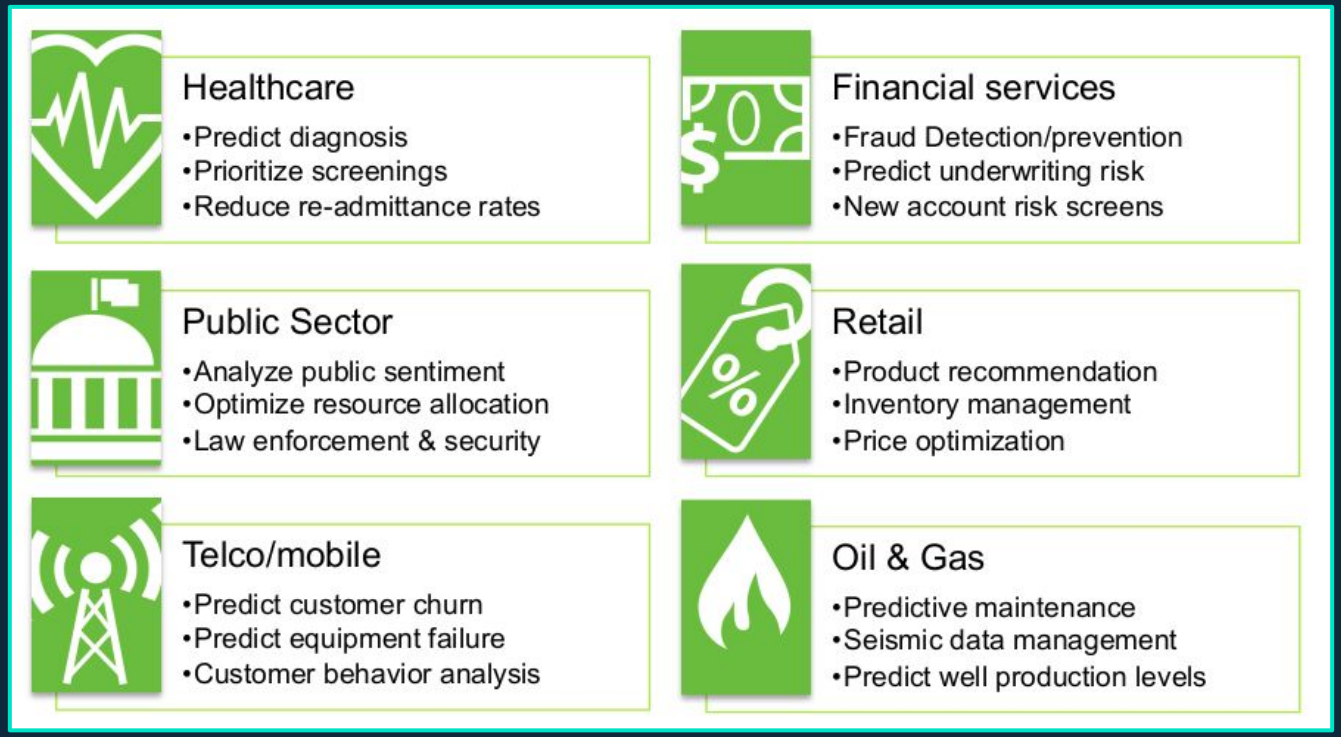
Data	*facts and statistics collected together for reference or analysis*
Science	*The intellectual and practical activity encompassing the systematic study of the structure and behavior of the physical and natural world through observation and experiment.*
Data Science	The **scientific** exploration of **data** to extract meaning or insight, and the construction of software systems to utilize such insight in a business context.

Data Scientist

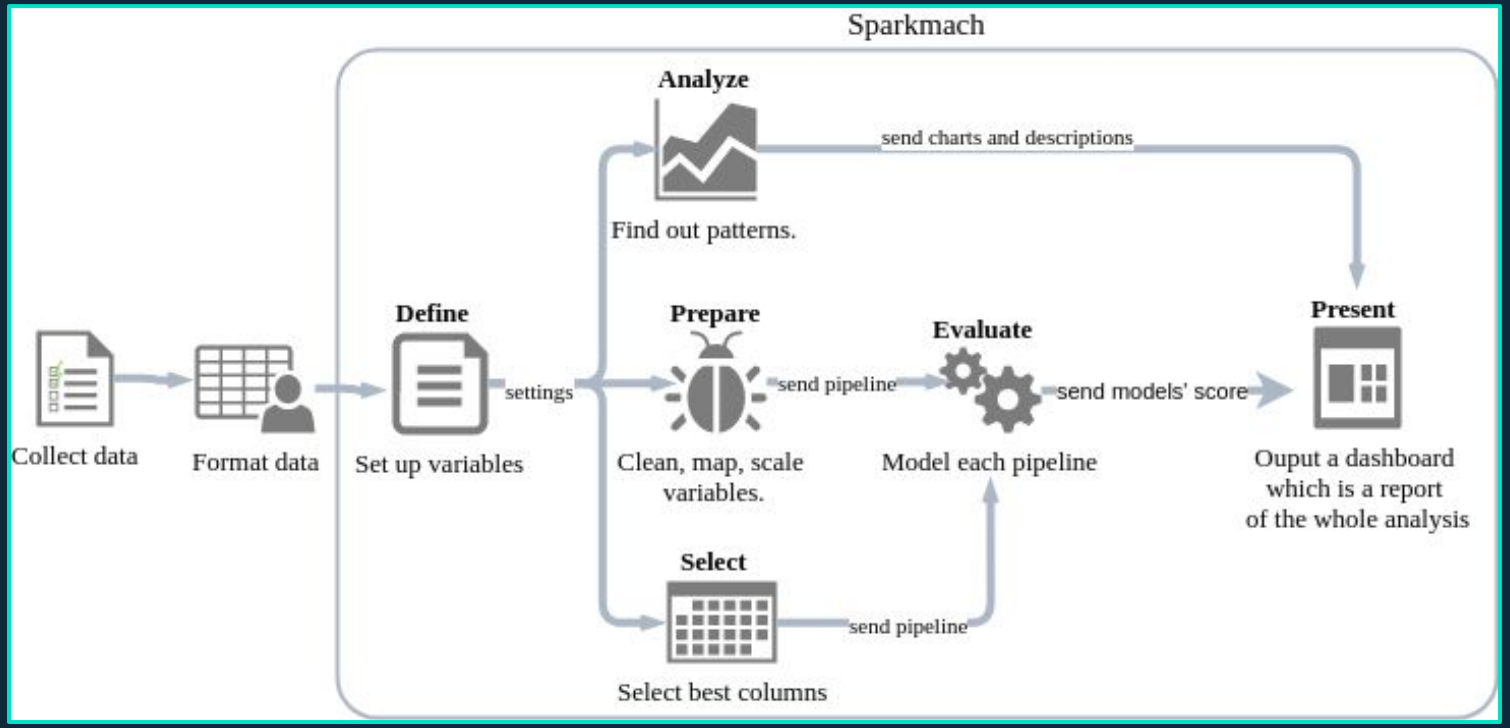
Someone who does **this …**



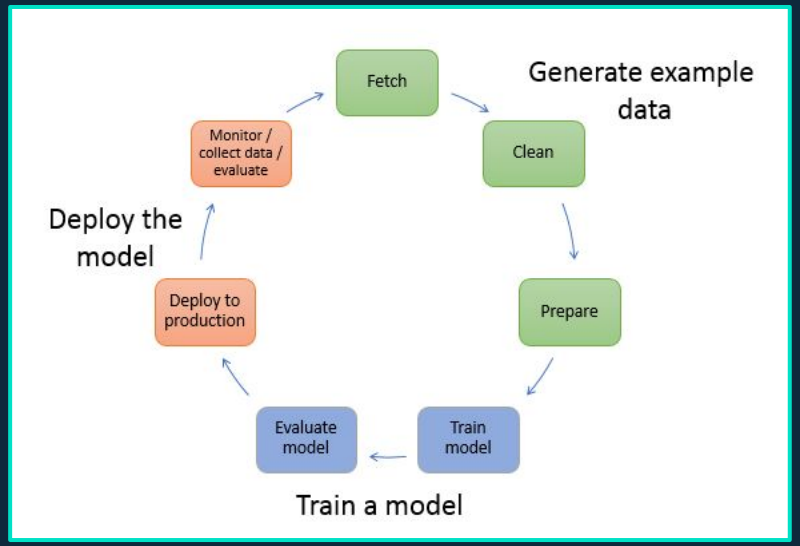
Aplicaciones



Workflow


Sparkmach
Analyze
send charts and descriptions
Find out patterns.
Define
Prepare
Evaluate
Present
Collect data
Format data
Set up variables
settings
send pipeline
send models' score
Clean, map, scale variables.
Model each pipeline
Ouput a dashboard which is a report of the whole analysis
Select
send pipeline
Select best columns

Real life?


Fetch
Generate example data
Clean
Prepare
Train model
Evaluate model
Deploy to production
Monitor / collect data / evaluate
Deploy the model
Train a model

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

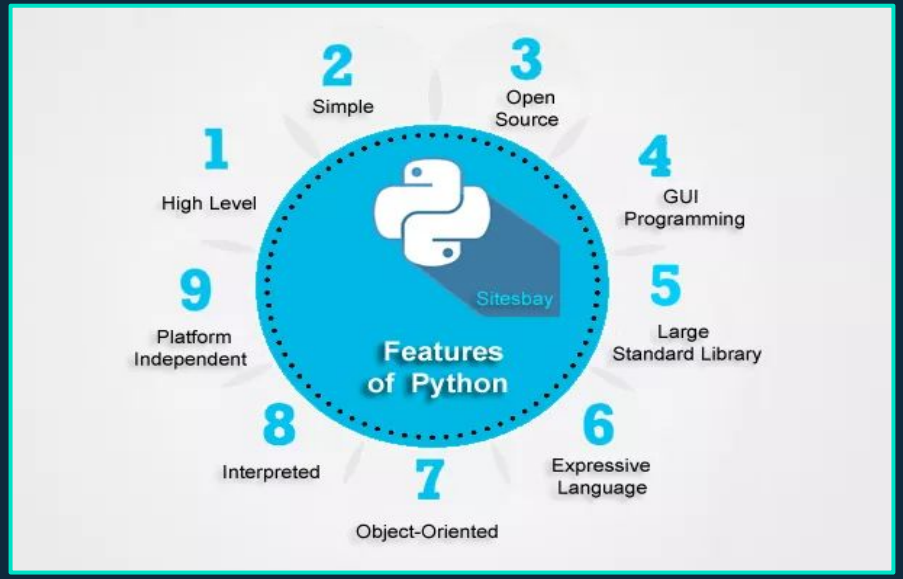
- ¿ Cómo uso mi modelo ?
- ¿ Cómo ajustar mi modelo ?
- Arquitectura del despliegue, tracking, A/B testing.

PYTHON ECOSYSTEM

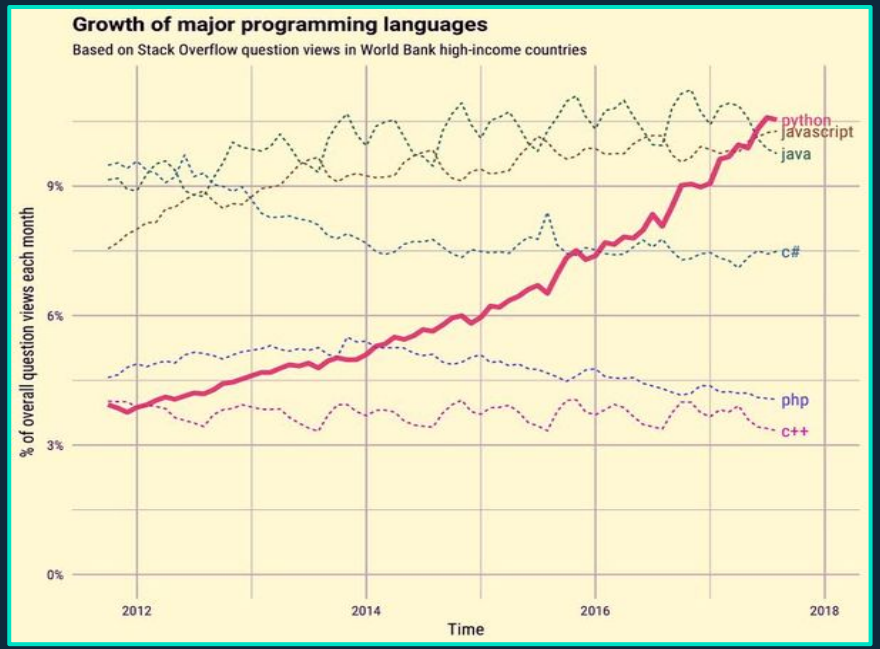
1

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

¿ Qué es Python?

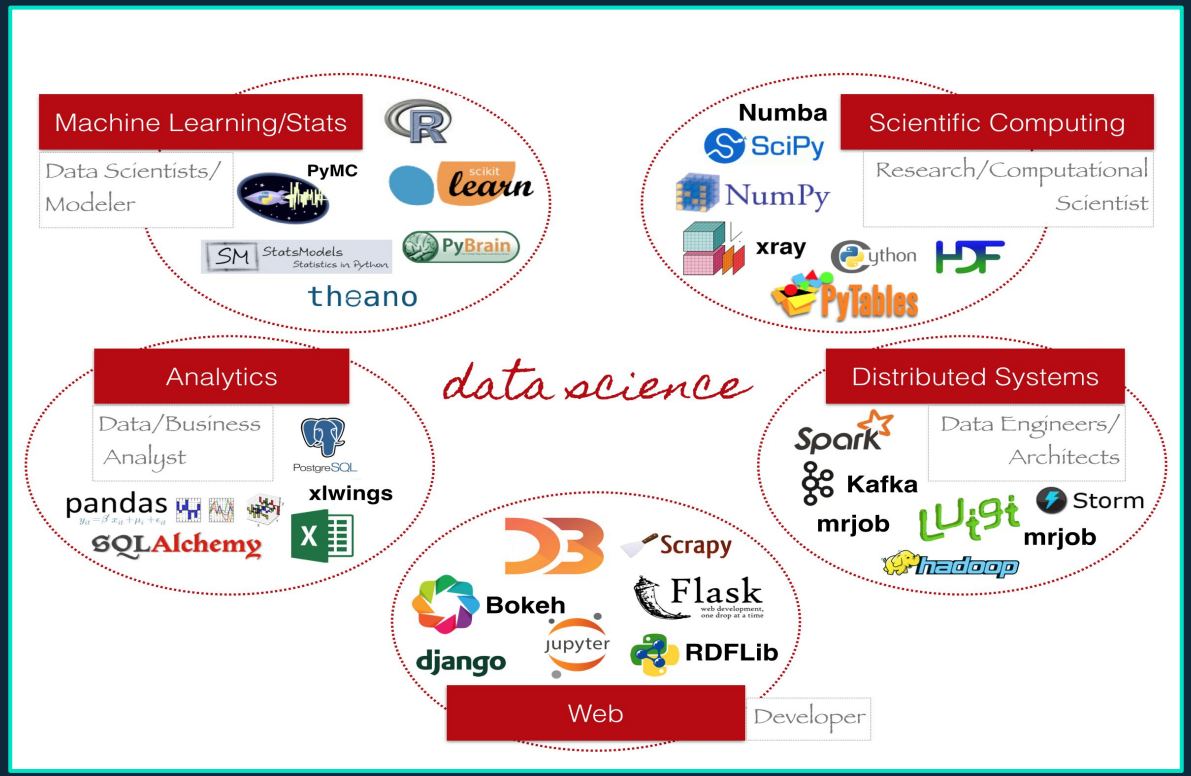


Alcance


Growth of major programming languages
Based on Stack Overflow question views in World Bank high-income countries
% of overall question views each month
9%
6%
3%
0%
2012
2014
2016
2018
Time
python
javascript
java
c#
php
c++


Google
facebook
Instagram
Pinterest
box
Dropbox
TiVo
comcast
at&t
SiriusXM
SATELLITE RADIO
ebay
reddit
digg
Concur
hotels.com
salesforce
slideshare
Chef
openstack
vmware
The New York Times
BBC
THE HUFFINGTON POST
The Weather Channel
NHS
GOV.UK
NAVTEQ

Python + Data science



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

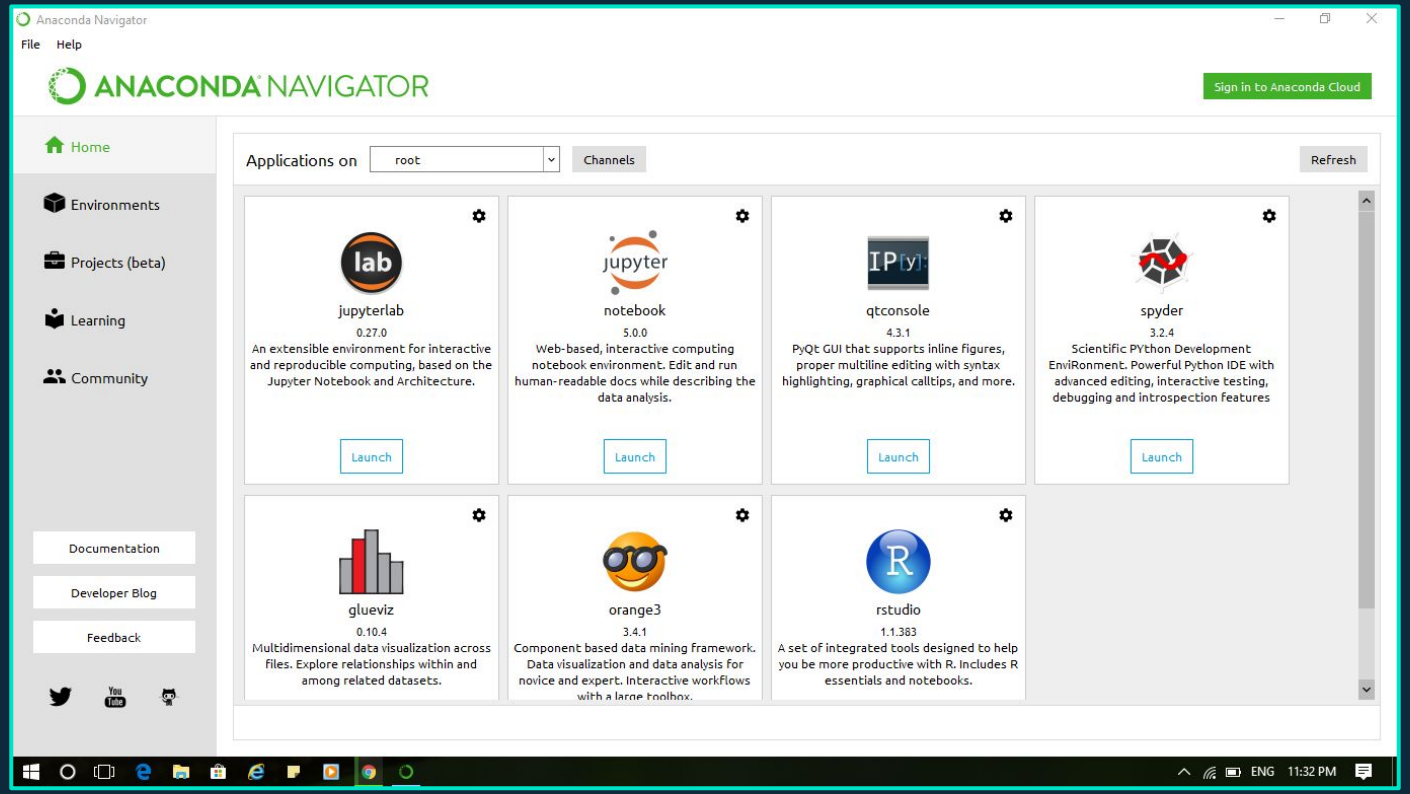
```
import turtle

star = turtle.Turtle()

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

turtle.done()
```

Anaconda

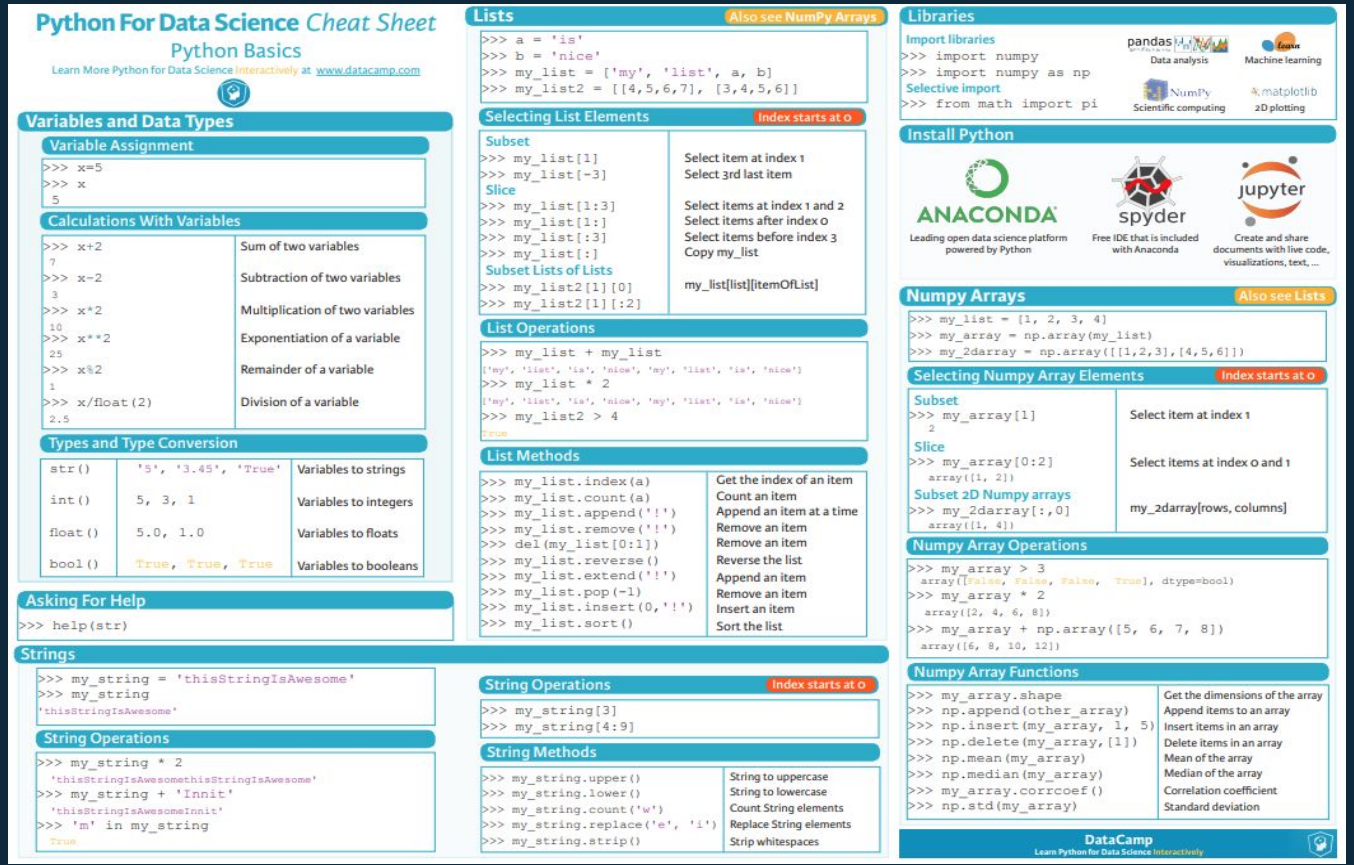

Anaconda Navigator
File Help
ANACONDA NAVIGATOR
Sign in to Anaconda Cloud
Home
Environments
Projects (beta)
Learning
Community
Documentation
Developer Blog
Feedback
Applications on root
Channels
Refresh
jupyterlab
0.27.0
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.
Launch
notebook
5.0.0
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.
Launch
qtconsole
4.3.1
PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.
Launch
spyder
3.2.4
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features
Launch
glueviz
0.10.4
Multidimensional data visualization across files. Explore relationships within and among related datasets.
orange3
3.4.1
Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.
rstudio
1.1.383
A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.
ENG 11:32 PM

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

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

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

String Operations — Index starts at 0

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

String Methods

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

Libraries

Import libraries

```
>>> import numpy
>>> import numpy as np
```

Selective import

```
>>> from math import pi
```

pandas — Data analysis; scikit-learn — Machine learning; NumPy — Scientific computing; matplotlib — 2D plotting

Install Python

ANACONDA — Leading open data science platform powered by Python

spyder — Free IDE that is included with Anaconda

jupyter — Create and share documents with live code, visualizations, text, ...

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`	Select item at index 1
Slice	
`>>> my_array[0:2]` `array([1, 2])`	Select items at index 0 and 1
Subset 2D Numpy arrays	
`>>> my_2darray[:,0]` `array([1, 4])`	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`	Get the dimensions of the array
`>>> np.append(other_array)`	Append items to an array
`>>> np.insert(my_array, 1, 5)`	Insert items in an array
`>>> np.delete(my_array,[1])`	Delete items in an array
`>>> np.mean(my_array)`	Mean of the array
`>>> np.median(my_array)`	Median of the array
`>>> my_array.corrcoef()`	Correlation coefficient
`>>> np.std(my_array)`	Standard deviation

DataCamp — Learn Python for Data Science Interactively

Thanks!

Any questions?

You can find me at:

- gbravor@uni.pe
- @gusseppe-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