Introduction to Python for Data Science

Ivanovitch Silva
Federal University of Rio Grande do Norte
Metropolis Digital Institute

November
8-10, 2017

Motivation

SMART METROPOLIS

Smart Metering   Smart Parking   Public Security   Smart Fleet

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Big Data and Social Analytics certificate course

MIT | EXperimental Learning

2017 DATES TO BE CONFIRMED

DOWNLOAD COURSE PROSPECTUS

Discover a new way to think about big data analysis when you explore the theory behind "social analytics", and practically apply that knowledge as you learn pioneering data analytics techniques from the creators of those very tools and methods.

Big Data

Data Science

Note

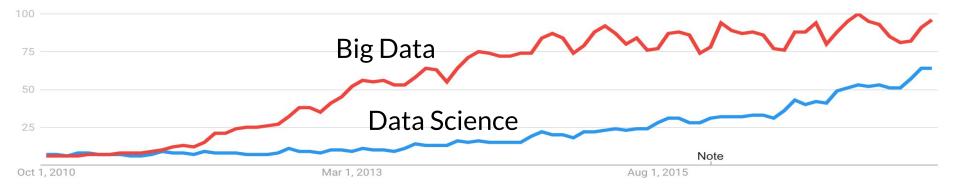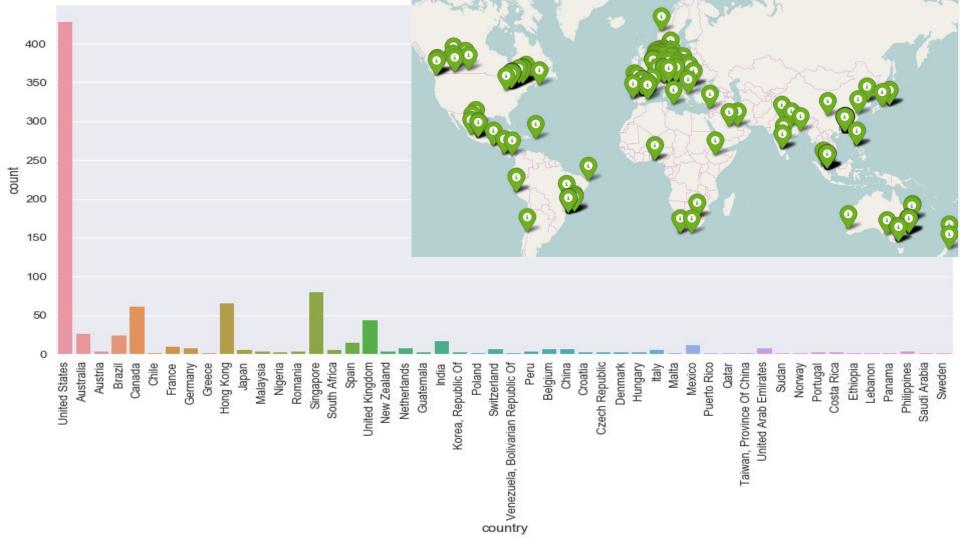Oct 1, 2010   Mar 1, 2013   Aug 1, 2015

2016/2017 - Specialization course in Big Data

Undergratuate
2017.1 - IMD0105 Introduction to Data Science
2017.2 - IMD0252 Learning Analytics
2017.2 - DCA0046 Data Science
2018* - Internet of Things (Data Science I & II)

Graduate
2017.2 - EEC2006 Data Science Foundations
2017.2 - ITE0021 Learning Analytics

https://github.com/ivanovitchm

Update the repository

git clone https://github.com/ivanovitchm/fr-br_school.git

Or ....

git pull
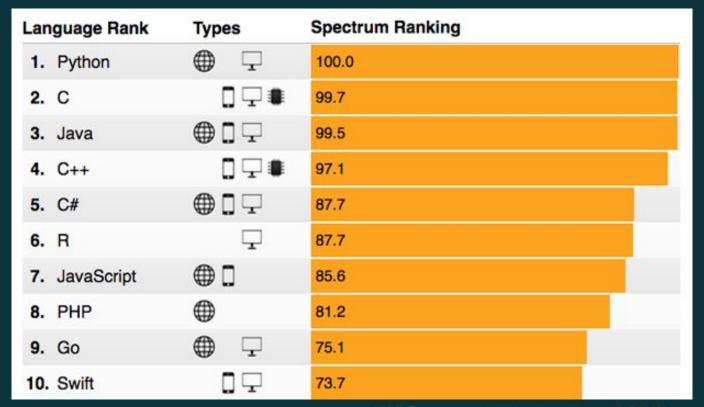
https://goo.gl/2mQm2K

If you come from a C.S./developer background, you'll probably feel more comfortable with Python. On the other hand, if you come from a statistics/analyst background, R will likely be more intuitive

R vs Python for Data Science
Summary of Modern Advances

elitedatascience.com

Language Rank	Types	Spectrum Ranking
1. C	📱💻🔲	100.0
2. Java	🌐📱💻	98.1
3. Python	🌐 💻	98.0
4. C++	📱💻🔲	95.9
5. R	💻	87.9
6. C#	🌐📱💻	86.7
7. PHP	🌐	82.8
8. JavaScript	🌐📱	82.2
9. Ruby	🌐 💻	74.5
10. Go	🌐 💻	71.9

IEEE Spectrum - Jul 2016   https://goo.gl/BqrkDI

Language Rank	Types	Spectrum Ranking
1. Python	🌐 🖥	100.0
2. C	📱 🖥 🖴	99.7
3. Java	🌐 📱 🖥	99.5
4. C++	📱 🖥 🖴	97.1
5. C#	🌐 📱 🖥	87.7
6. R	🖥	87.7
7. JavaScript	🌐 📱	85.6
8. PHP	🌐	81.2
9. Go	🌐 🖥	75.1
10. Swift	📱 🖥	73.7

IEEE Spectrum - Jul 2017   https://goo.gl/HSPLWe

PYTHON 2

PYTHON 3

https://goo.gl/JdRL2a

Version 3.x (https://www.python.org/downloads/)

ANACONDA®
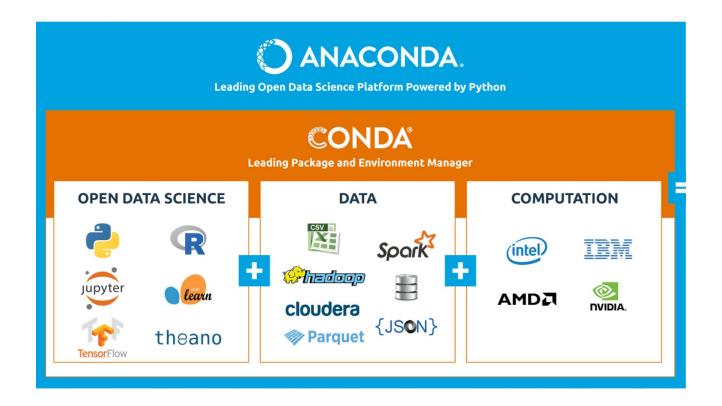
Modern open source analytics platform
powered by Python

https://www.continuum.io/downloads
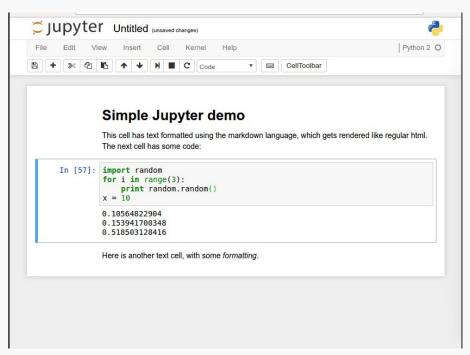
Why Anaconda?

What is a Jupyter Notebook?

Mix of code and rich elements (text, figures, links, equations, etc)

Aside from JUlia, PYThon and R (JUPYTER) notebook technology also supports many other languages.
https://github.com/jupyter/jupyter/wiki/Jupyter-kernels
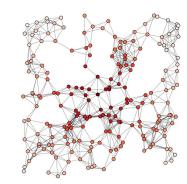
https://goo.gl/LfEx5r

EVOLUTION

Line plot · Area plot · Stacked area plot · Parrallel plot · Streamchart

MAPS

Map · Chloroplet map · Connection map · Bubble map

FLOW

Chord diagram · Network chart · Sankey diagram

Other

Animation · Cheat sheet · Data Art · Color · 3D · Bad chart

THE PYTHON
GRAPH GALLERY

https://python-graph-gallery.com/

DISTRIBUTION

VIOLIN · DENSITY · BOXPLOT · HISTOGRAM

CORRELATION

Scatterplot · Connected Scatter plot · Bubble plot · Heatmap · 2D density plot · Correlogram

RANKING

Barplot · Boxplot · parallel plot · Lollipop plot · Wordcloud · Spider

PART OF A WHOLE

Stacked barplot · Tree plot · Venn diagram · Doughnut plot · Pie plot · Tree diagram

Where to learn?

UDACITY

DataCamp

DATAQUEST

Data Science Academy

Who to follow?

http://www.andrewng.org/

http://mariofilho.com/

The Periodic Table of Data Science

https://goo.gl/eecQ2V

An overview of key companies, resources and tools in data science (as of 4/12/2017)

Symbol → **Dc**
Name → DataCamp

Legend

- Courses
- Boot camps
- Conferences
- Data
- Projects & Challenges, Competitions
- Programming Languages & Distributions
- Search & Data Management
- Machine Learning & Stats
- Data Visualization & Reporting
- Collaboration
- Community & Q&A
- News, Newsletters & Blogs
- Podcasts

Courses / Boot camps / Conferences (left column block)

Dc DataCamp	Ga General Assembly	Sd Strata Data
Sb SpringBoard	M Metis	Od ODSC
Ex Edx	Di Data Incubator	Tc Tableau Conference
C Coursera	In Insight	U UseR!
Uda Udacity	Dsa NYC Data Science Academy	Pd PyData
Ude Udemy	G Galvanize	Paw Predictive Analytics World
Ps Pluralsight	Dsg Data Science for Social Good	Kdd ACM SIGKDD Conference
Ly Lynda	Dsy Data Society	Tpc Teradata Partners Conference
Tt TeamTreeHouse	Dsj Data Science Dojo	Icd IEEE International Conference on Data Mining
Bdu Big Data University		

Main table

Py Python	Js JavaScript	Vb Visual Basic	Pgs PostgreSQL	Sli SQLite	Ah Apache Hadoop	W Weka	Bml BigML	Kn Knime	Sm Spark MlLib	Pb Power BI	Obi Oracle BI	Shn Shiny	Ddl Domino Data Lab	De Data Science Experience
R R	Cp C++	Sc Scala	Ar Amazon Redshift	Bq Google BigQuery	Hw Hortonworks	O Oracle	Dar DataRobot	Lib LibSVM	Ho H2O	Bo BusinessObjects	Alt Alteryx	Mpl Matplotlib	Nt Nteract	Rs Rstudio
S SQL	Pl Perl	Ca Cassandra	Hb HBase	Td Teradata	Cl Cloudera	Mss Microsoft SQL server	Rm RapidMiner	Mat Mathematica	Th Theano	Sp Spotfire	Sav SAS Visual Analytics	Ply Plotly	Ro Rodeo	Be Beaker Notebook
B Bash	Mr Microsoft R Open	P Pig	Mdb Mongo DB	To Toad	Aem Amazon Elastic Mapreduce	Spl Splunk	Cho Chorus	Mah Mahout	Aml Azure Machine Learning	Ql Qlikview	Po PowerPivot	Me Microsoft Excel	Spy Spyder	Ze Apache Zeppelin
Mtl Matlab	Cy Canopy	Im Impala	K Kafka	Ms MySQL	Mar MapR	Sr Solr	Tf Tensorflow	St Stata	D D3	Co Cognos	Gch Google Charts	Pe Pentaho	Dst Data Science Studio	Ju Jupyter
J Java	An Anaconda	Sp Spark	Hi Hive	Idb IBM DB2	Lu Lucene	El ElasticSearch	Sk Scikit-Learn	Da Dato/Graphlab	My Microstrategy	Aa Adobe Analytics	T Tableau	B Bokeh	Db Databricks notebook	Gh Github

Data (bottom left block)

Dw Data.world	Q Quandl	Fte FiveThirtyEight	Sa Socrata	Gp Google Public	Dg Data.gov	K Kaggle
St Statista	Uci UCI Machine Learning Repository	Wb World Bank	At Academic Torrents	Bf Buzzfeed	Dk DataKind	Dd DrivenData

Community & Q&A (bottom middle block)

Re Reddit	So Stack Overflow	Cv Cross Validated	Qu Quora	Av Analytics Vidhya	Dse Data Science Stack Exchange
Mu Meetup	Rdm RDataMining				

News, Newsletters & Blogs / Podcasts (right column block)

Kdn KDnuggets	Ibd insideBIGDATA
Rb R-Bloggers	Pp PlanetPython
Hn HackerNews	Dt DataTau
Dsc Data Science Central	Dsr Data Science Roundup
Dsw Data Science Weekly	Or O'Reilly
Dr Data Elixir	Pw Python Weekly
Rw R Weekly	Pd Partially Derivative
Bds Becoming a Data Scientist	Tm Talking Machines
Ds Data Stories	Dsk Data Skeptic
Ld Linear Digressions	Ns Not So Standard Deviations

Data

Data Files
(XML, CSV, Excel, JSON, …)

Database
(MySQL, Oracle, …)

API

Sites

Text and reports

Maps

Image and videos

Social Media

Agenda

1. Open data Natal (elementary schools vs kindergarten distributions)
   a. Pandas, Folium, Geocoder, Heatmap
2. Population estimation
   a. Pandas, Numpy, Json, Shapely, Folium, Choropleth maps
3. Uber behavior (price and time estimation) to Natal
   a. Uber API, Pandas, Choropleth maps