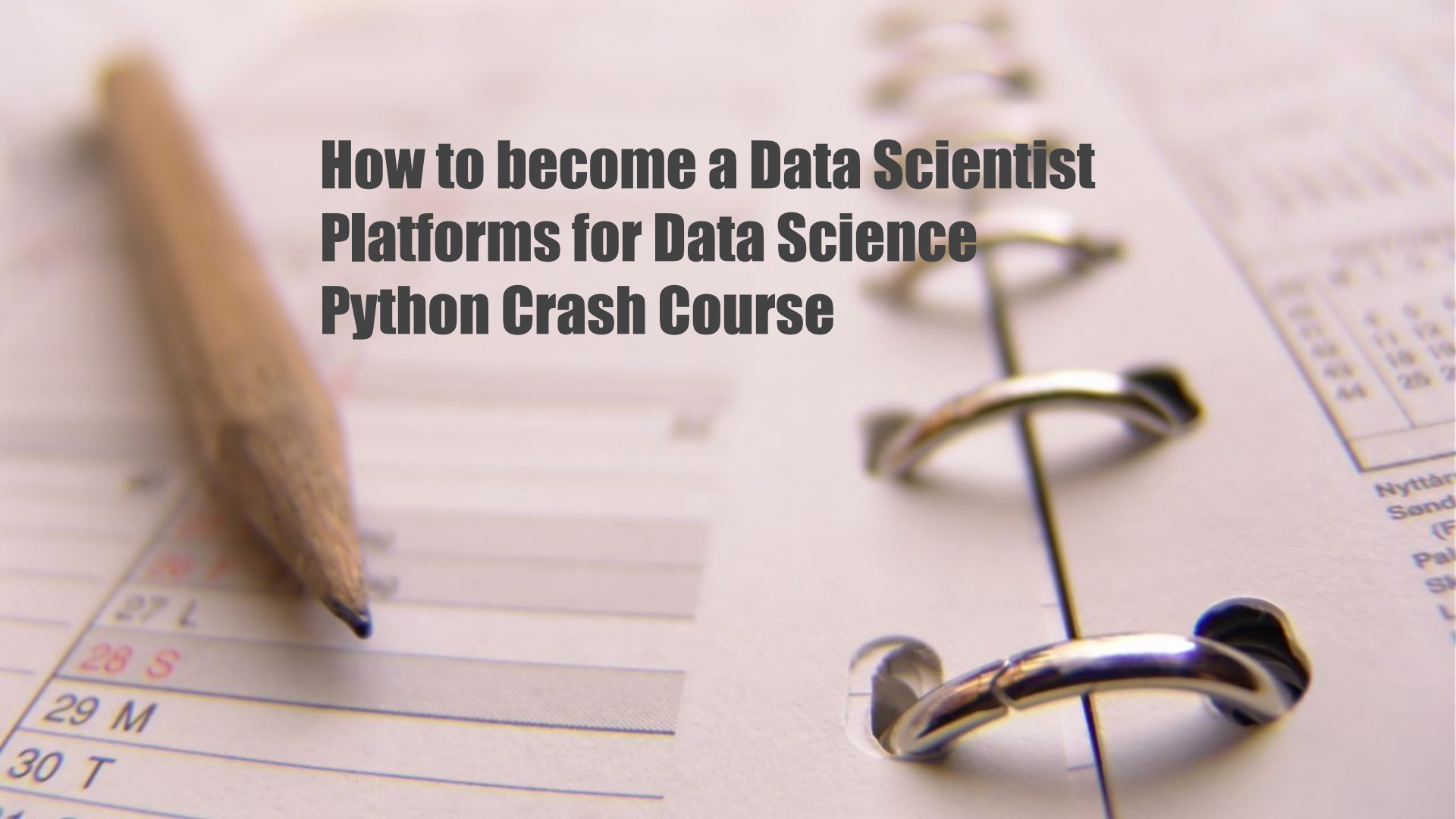

Lesson #02
First Steps on Data Science
Feb. 2019


How to become a Data Scientist
Platforms for Data Science
Python Crash Course
27 L
28 S
29 M
30 T

Update from repository

```
git clone https://github.com/ivanovitchm/datascience_one_2019_1
```

Or ....

```
git pull
```

How to Become a **Data Scientist**

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

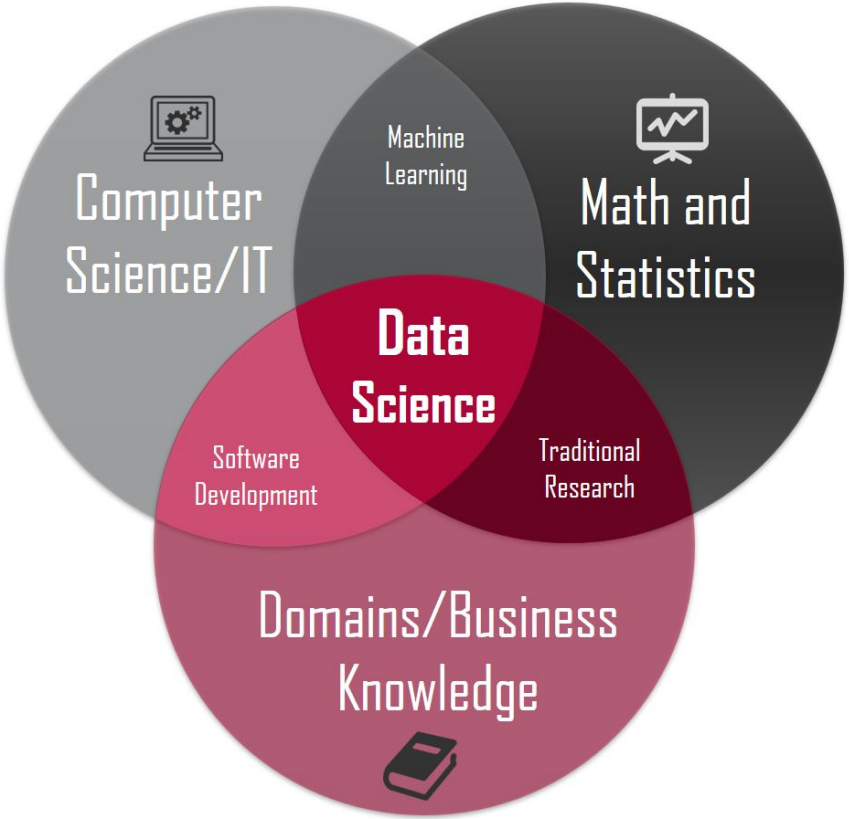
- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

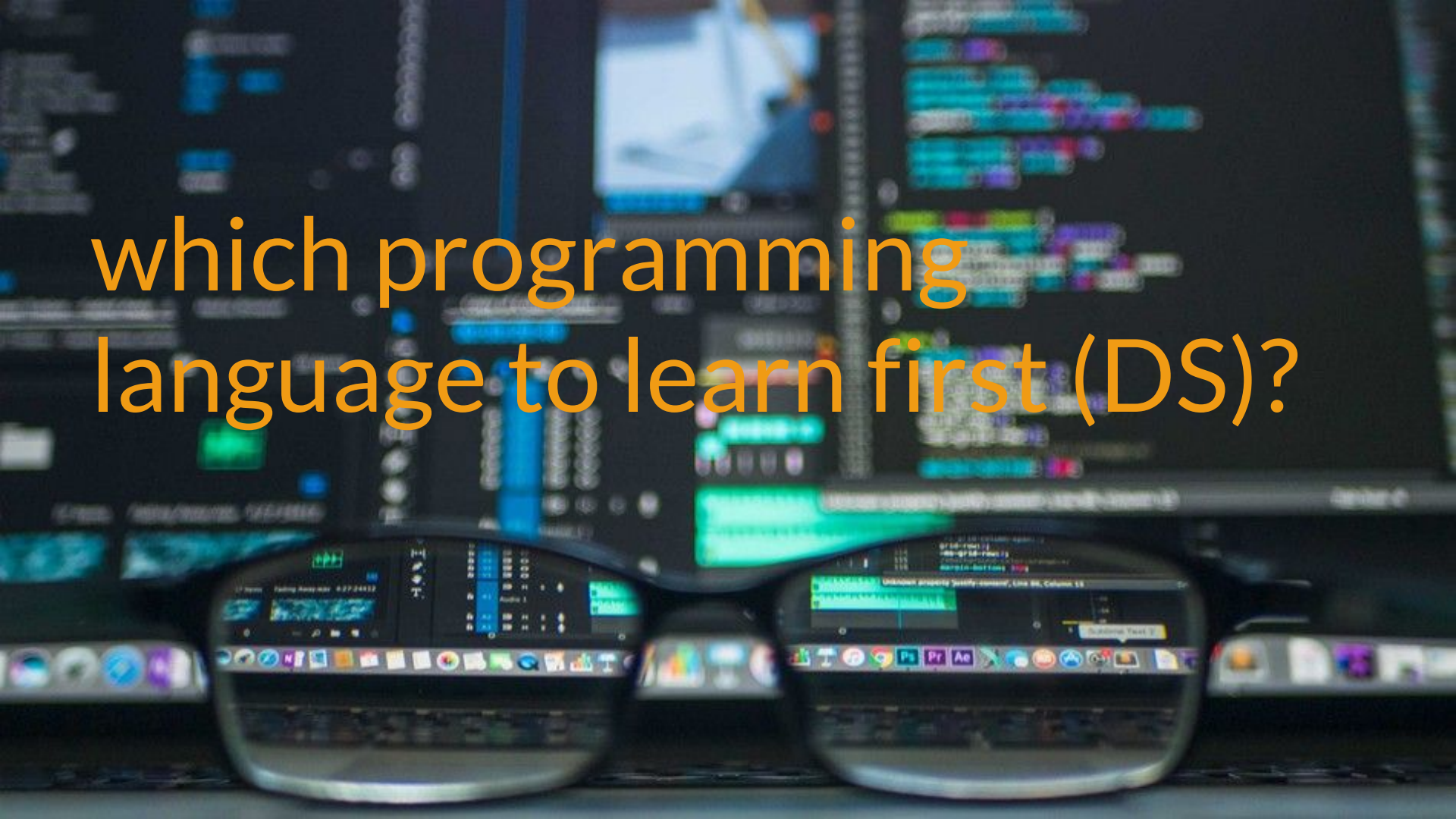
- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

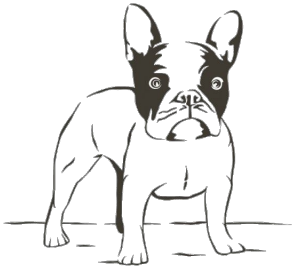


Pick ONE programming language and STICK to it. Don’t go back and constantly change your choice of language to study. If you do, you will slow your progress down.


which programming
language to learn first (DS)?

https://goo.gl/VKYfXn

have
a pet
project



Be clear about your motivation. The reason this is important because learning Data Science is HARD. VERY HARD! So it's easy to lose motivation when on the journey.

Data Overview Kernels (1) Discussion Activity Download (59 KB) New Kernel

Public Your Work Favorites Sort by Hotness

Outputs Languages Types Tags Search kernels

2 Starter: Brazilian_dams_and_Brumadinho_ 5ce3f572-e
17d ago starter code Py 0

Immerse yourself in the community (newsletters, articles, books, podcasts, youtube, hackathons and meetups)

Open Data Day 2019 - Natal

Created by **Frederico Pranto** at Jan. 31, 2019, 4:27 p.m..
Accepting proposals until 1 month

Open Data Day 2019 - Natal

Natal-RN, Brasil
23 de março de 2019
Local: IFRN Central - Avenida Senador Salgado Filho, 1559, Tirol, Natal-RN

O **Open Data Day** é um momento anual onde todo o mundo debate e promove, por um dia, o uso de dados abertos.

Em 2019, o evento acontecerá pela segunda vez em Natal. Este ano, devido ao período de Carnaval e também para evitar conflitos de horários com outros eventos na nossa cidade, a data escolhida foi o dia **23 de março**, sábado, e as 4 áreas-chave de discussão são: ciência aberta, rastrear fluxo de dinheiro público, mapeamento aberto e desenvolvimento igualitário.

O público do Open Data Day é bastante diverso: servidores públicos, desenvolvedores, bibliotecários, designers, advogados, estatísticos e demais cidadãos interessados. Não é preciso ter uma formação específica para fazer parte. O importante é participar.

As inscrições para o evento serão abertas em breve.

https://speakerfight.com/events/open-data-day-2019-natal/

http://www.usp.br/datascience/spsas-learning-from-data/

Data Science

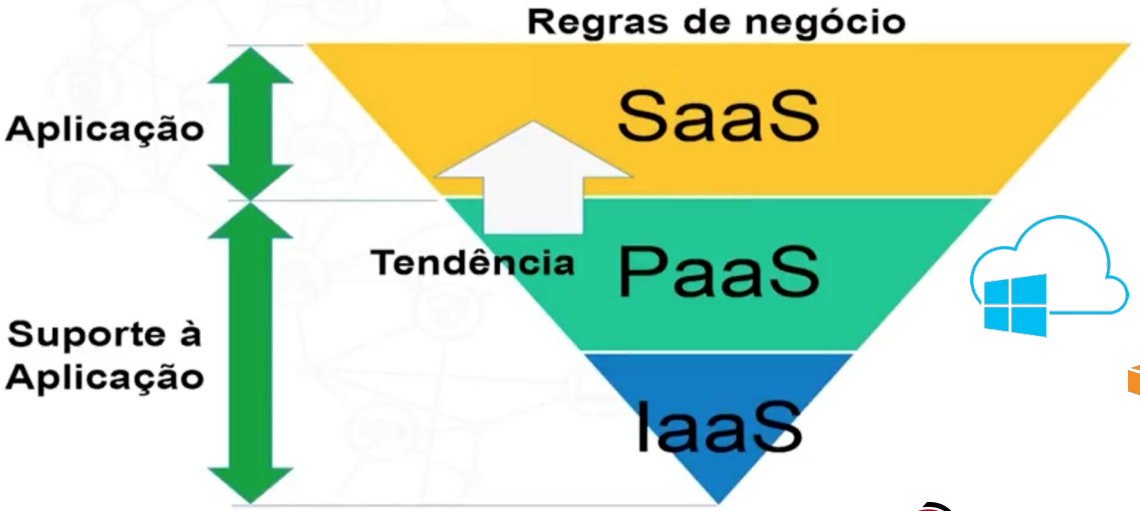
São Paulo School of Advanced Science on Learning from Data

July 29 to August 9, 2019

Data science Platforms

https://www.kdnuggets.com/2019/02/gartner-2019-mq-data-science-machine-learning-changes.html
https://medium.com/eteam/10-of-the-best-platforms-for-data-science-and-machine-learning-36a61ec1a676

#cloudcomputing



MLaaS
Machine Learning as a Service

Alibaba Cloud

SAP

ORACLE

amazon web services

SIEMENS

SOFTLAYER

linode

DigitalOcean

ORACLE
Cloud Infrastructure

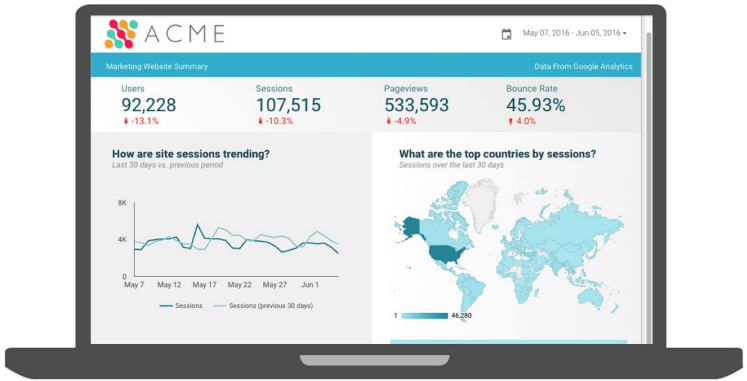
-30% AWS, -26% Microsoft
Fonte: Rackspace, 2013

http://www.pentaho.com/

https://www.tableau.com/

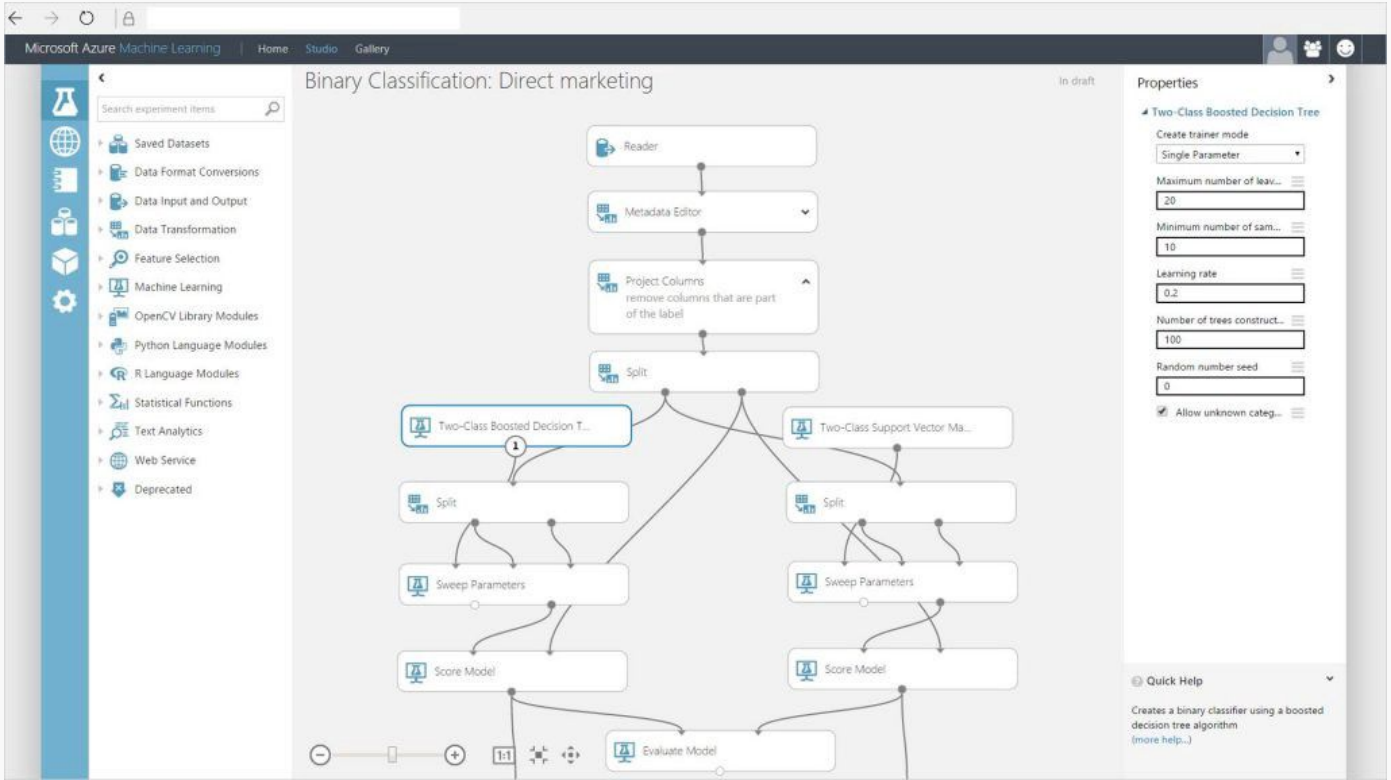
https://powerbi.microsoft.com

Google Data Studio



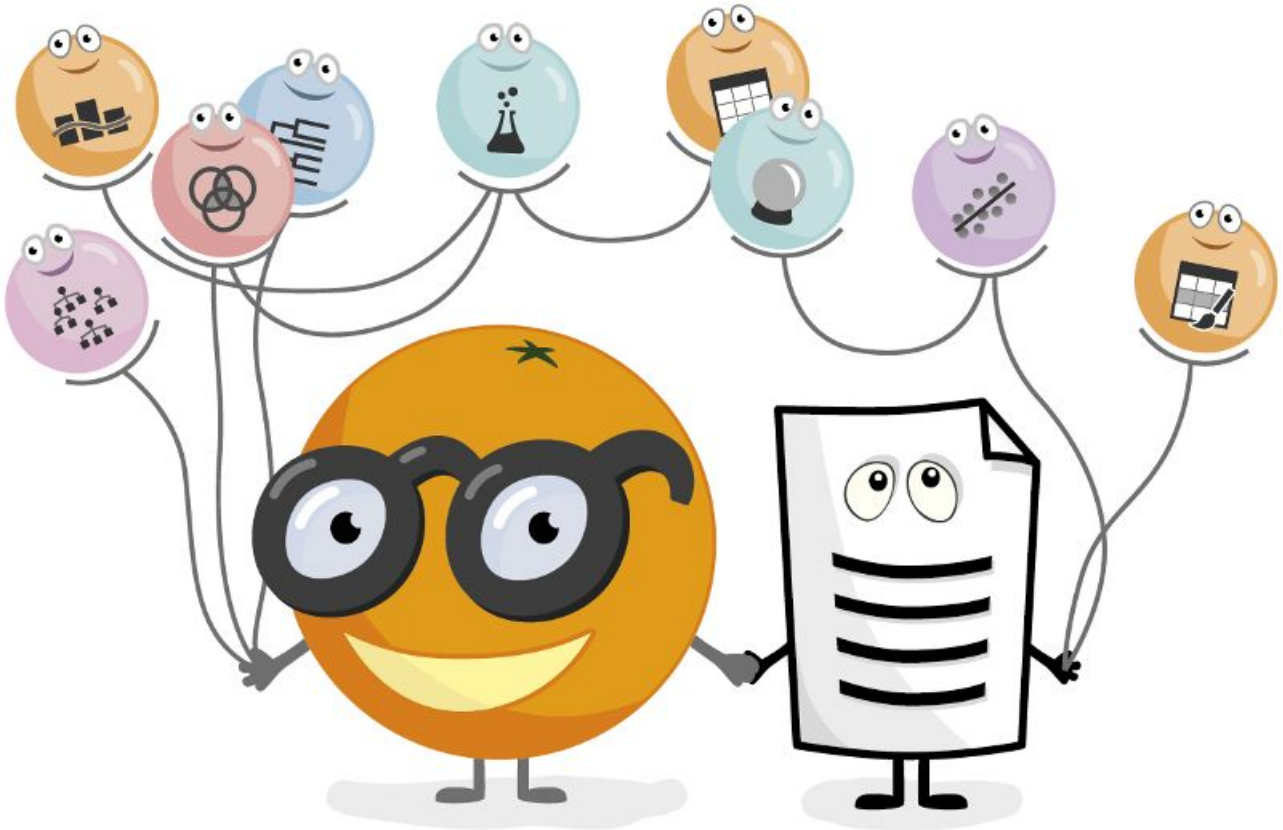
https://www.google.com.br/analytics/data-studio/

Microsoft Azure



https://goo.gl/Ndf38Q

Data Mining Fruitful and Fun



https://orange.biolab.si/

https://www.anaconda.com/distribution/

Why Anaconda?

jupyter Untitled (unsaved changes)

File Edit View Insert Cell Kernel Help Python 2













Code CellToolbar

Simple Jupyter demo

This cell has text formatted using the markdown language, which gets rendered like regular html. The next cell has some code:

In [57]:

```
import random
for i in range(3):
    print random.random()
x = 10
```

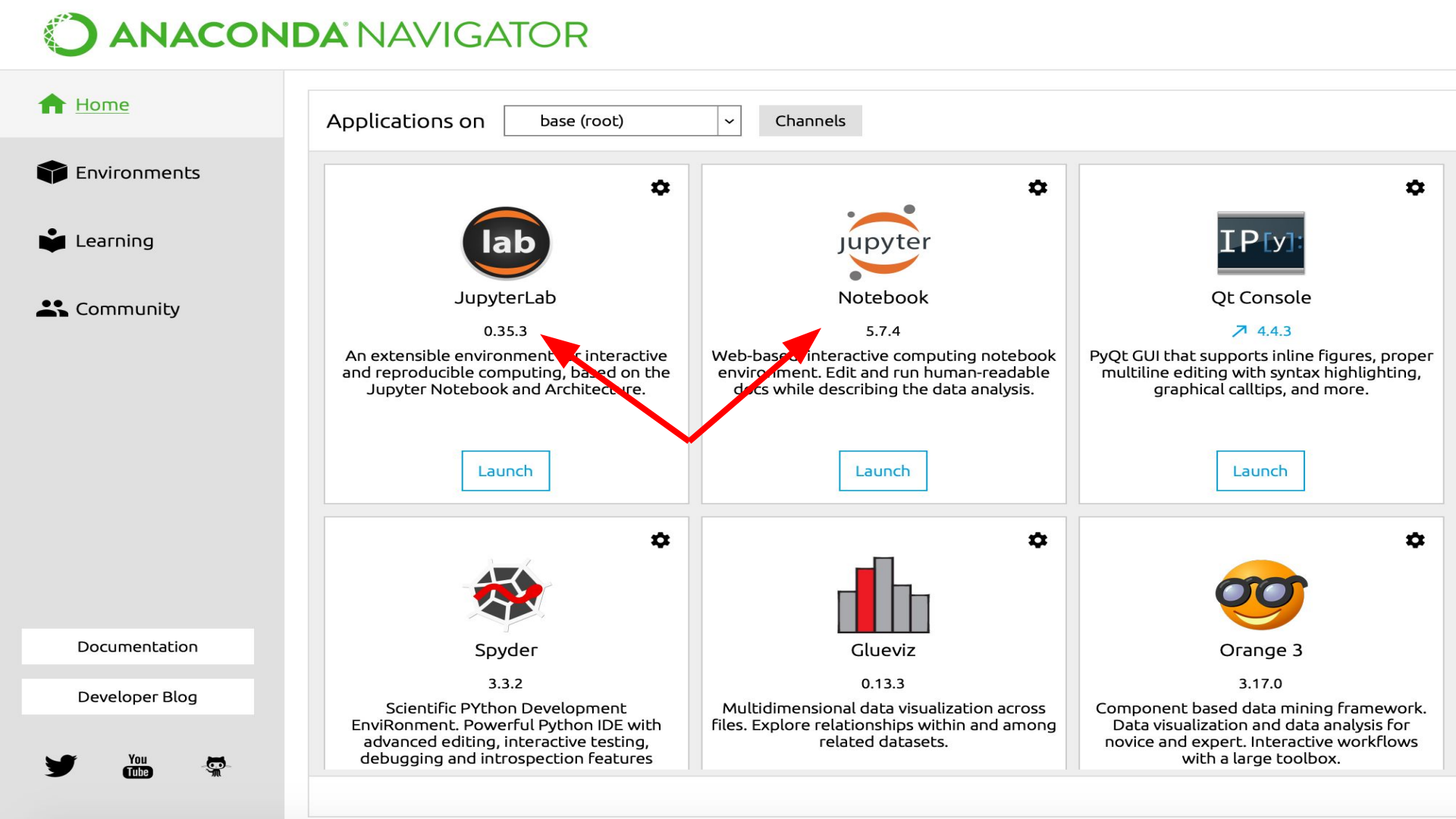
```
0.10564822904
0.153941700348
0.518503128416
```

Here is another text cell, with some *formatting*.

Tuesday at Berkeley: Data 8, ~1,300 students

In 2018, UC Berkeley launched a new major in data science, anchored by two core courses—Foundations of Data Science and Principles and Techniques of Data Science—powered by Jupyter infrastructure

Data 100, ~800 (650 last spring)


ANACONDA NAVIGATOR
Home
Environments
Learning
Community
Documentation
Developer Blog
Applications on
base (root)
Channels
JupyterLab
0.35.3
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.
Launch
Notebook
5.7.4
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.
Launch
Qt Console
4.4.3
PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.
Launch
Spyder
3.3.2
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features
Glueviz
0.13.3
Multidimensional data visualization across files. Explore relationships within and among related datasets.
Orange 3
3.17.0
Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.

Sponsors

Project Jupyter receives direct funding from the following sources:

THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST

ALFRED P. SLOAN FOUNDATION

GORDON AND BETTY MOORE FOUNDATION

Google

rackspace
the #1 managed cloud company

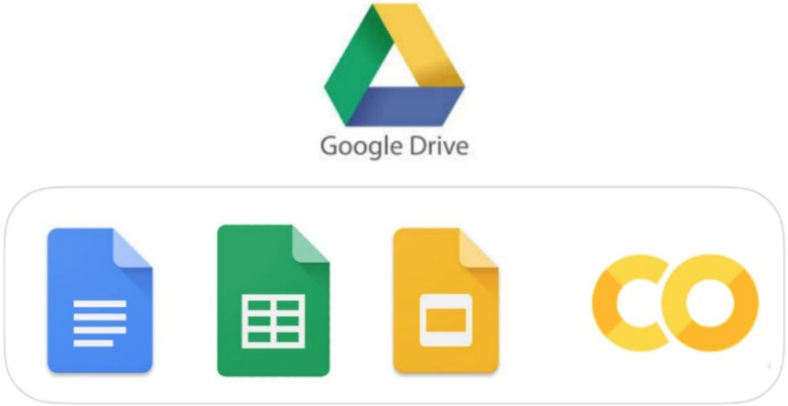
fastly

Horizon 2020
European Union funding
for Research & Innovation

Microsoft

Google Colaboratory

https://colab.research.google.com/



Colaboratory is a Google research project created to help disseminate machine learning education and research. It's a Jupyter notebook environment that requires no setup to use and runs entirely in the cloud.

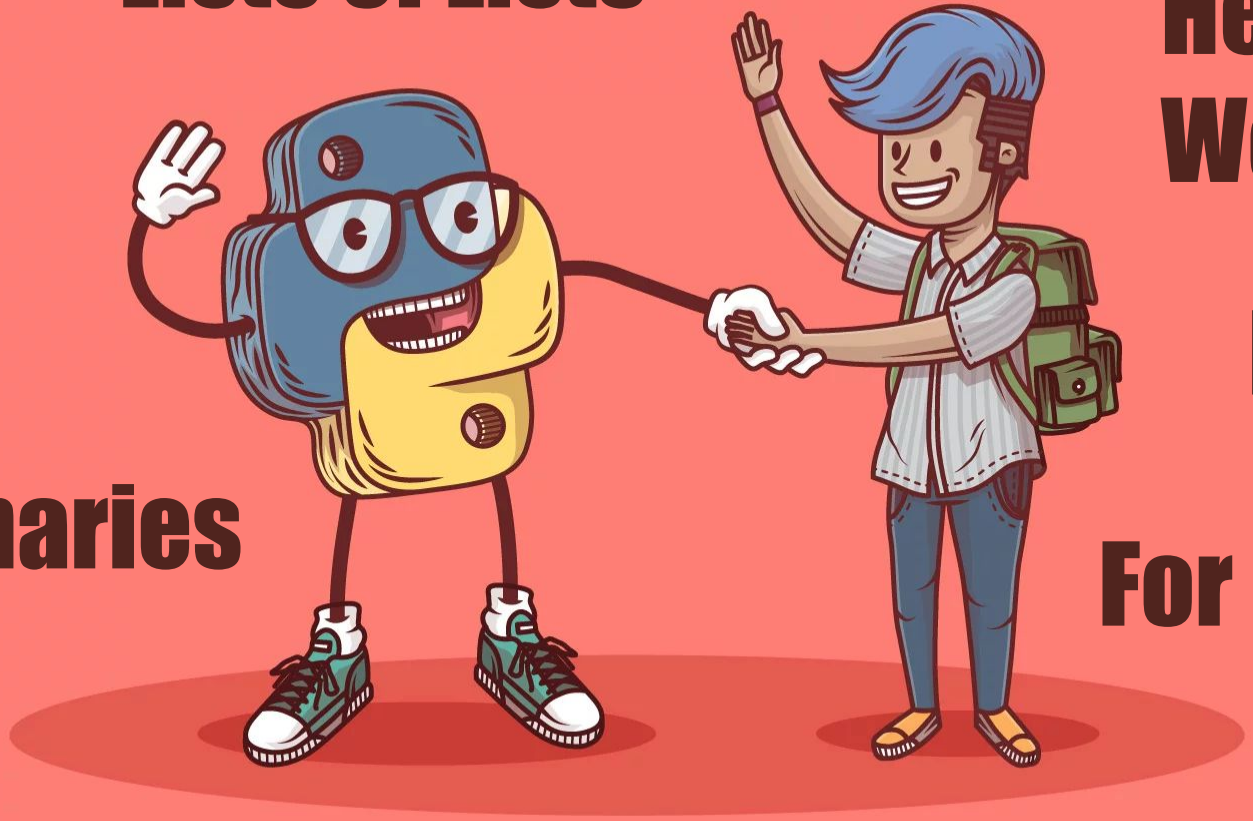
Colaboratory notebooks are stored in Google Drive and can be shared just as you would with Google Docs or Sheets. Colaboratory is free to use.


PYTHON
CRASH COURSE
A HANDS-ON, PROJECT-BASED
INTRODUCTION TO PROGRAMMING
ERIC MATTHES
no starch press

Lists of Lists

Hello World

Lists



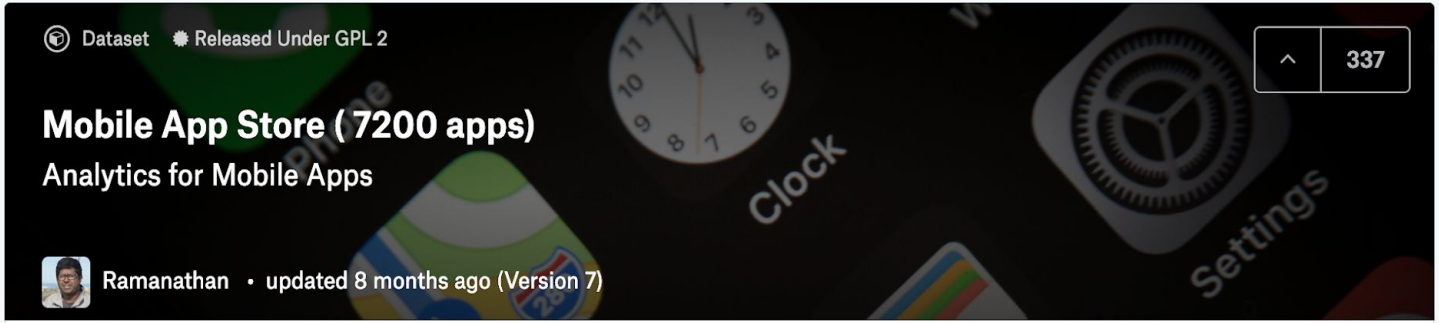
Files

Dictionaries

For Loops

Conditional Statements

Real Python



Data (6 MB)

Data Sources

- AppleStore.csv 7198 x 17
- appleStore_descriptio... 7197 x 4

About this file

Apple Store

Columns

- # user_rating
- # user_rating_ver
- A ver
- A cont_rating
- A prime_genre
- # sup_devices.num
- # ipadSc_urls.num
- # lang.num

	track_name	price	currency	rating_count_tot	user_rating
0	Facebook	0.0	USD	2974676	3.5
1	Instagram	0.0	USD	2161558	4.5
2	Clash of Clans	0.0	USD	2130805	4.5
3	Temple Run	0.0	USD	1724546	4.5
4	Pandora - Music & Radio	0.0	USD	1126879	4.0

AppleStore.csv