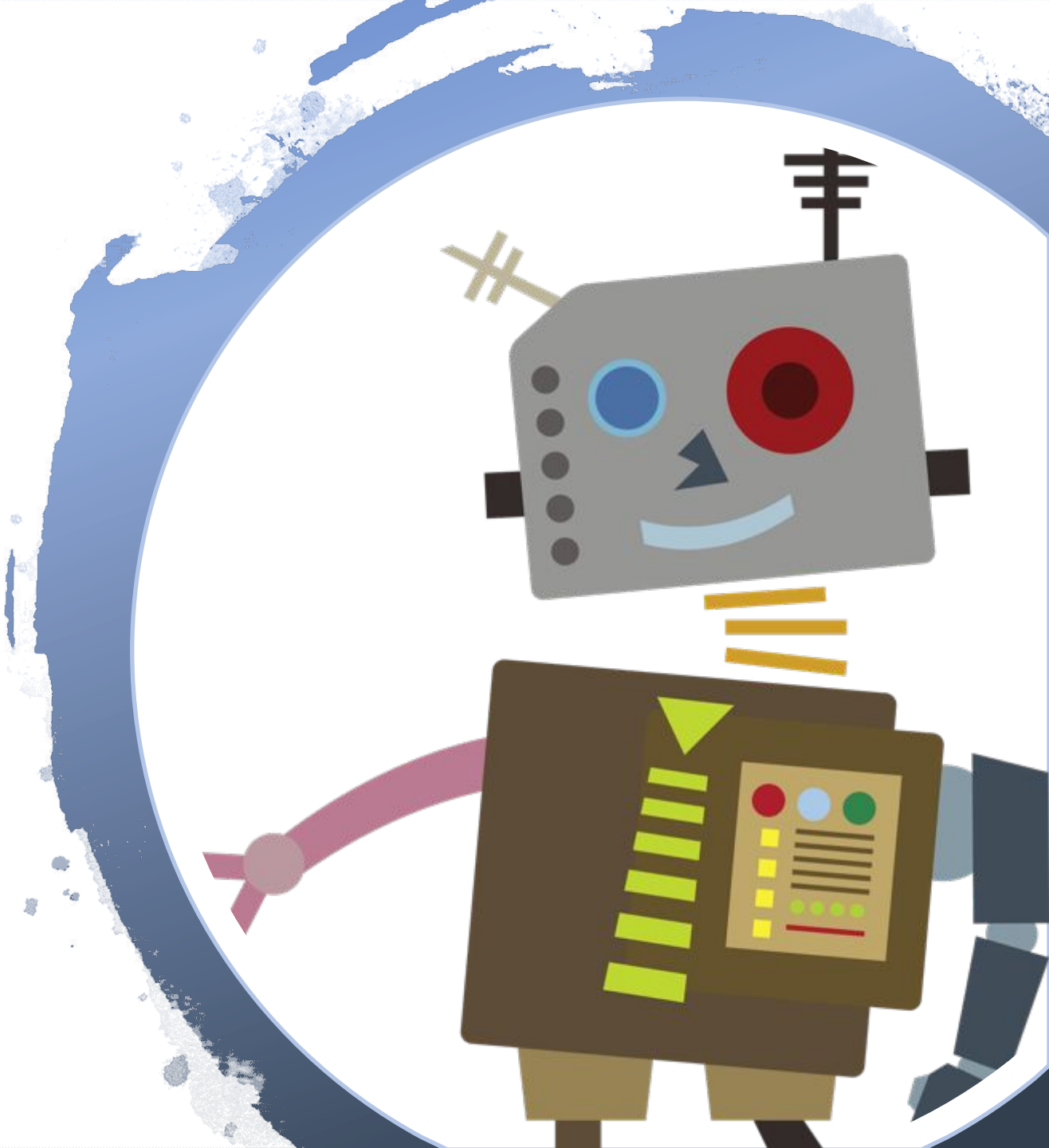
Data Science Applied to Business

By Marcelo Rovai

May 20th, 2020



Marcelo Rovai

Brazilian from São Paulo, Master's degree in Data Science by UDD, Chile, and MBA by IBMEC, Brazil. Graduated in 1982 as an Engineer from UNIFEI with pos graduation by Poli/USP, both in Brazil. Marcelo worked as a teacher, engineer, and executive in several companies in the technology area such as AVIBRAS Aeroespacial, SID Informática, ATT-GIS, NCR, DELL, COMPAQ (HP) and more recently at IGT where he continues as Senior Advisor.

In 2016, Marcelo began writing about electronics, publishing his works in sites of the area as MJRoBot.org (Editor/Writer), Hackster.io (#1 Contributor), Instructables.com, and Medium.com (TDS – Towards Data Science). Besides winning several Instructables competitions in the areas of electronics, robotics, and IoT.

Marcelo lives with his wife Ilza in Santiago, Chile, where he divides his time between his consultant work and sharing ideas in the field of Data Science, Electronics, IoT, Physical Computing and Robotics.

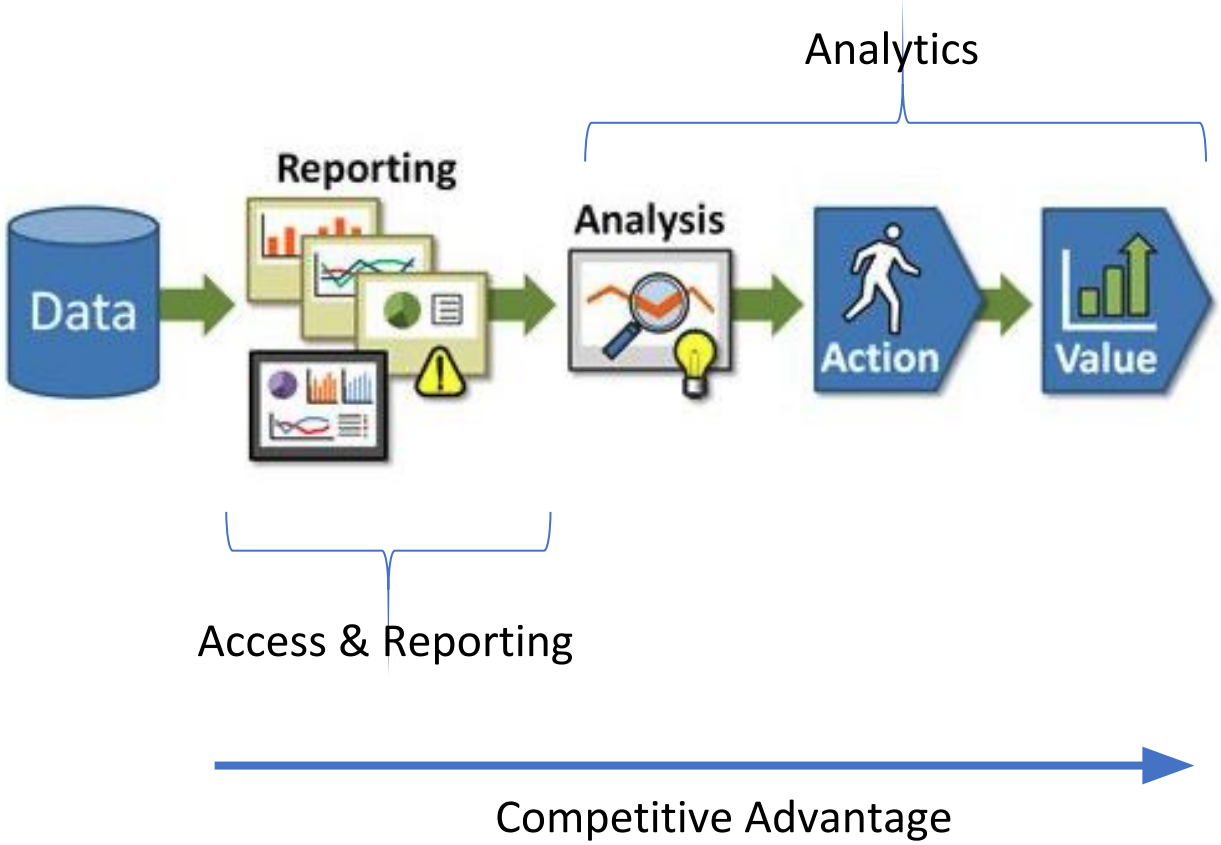
What is Data Science?

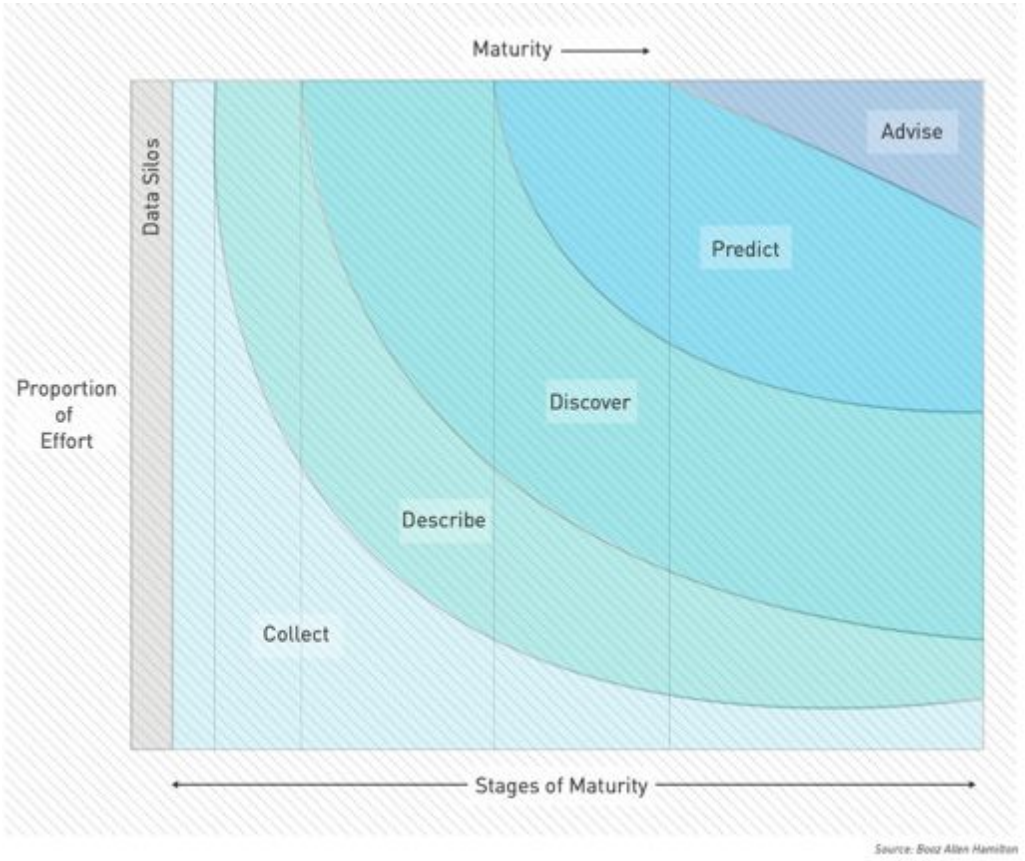
"Data science is a broad field of study aimed at maintaining data sets and deriving meaning out of them.

The unique combination of statistics, mathematics, and business domain knowledge offered by Data Science has the potential to help organizations find ways to reduce costs, get into new markets, tap on different demographics, gauge the effectiveness of marketing campaigns or launch a new products and services."

McCombs School of Business at The University of Texas at Austin.

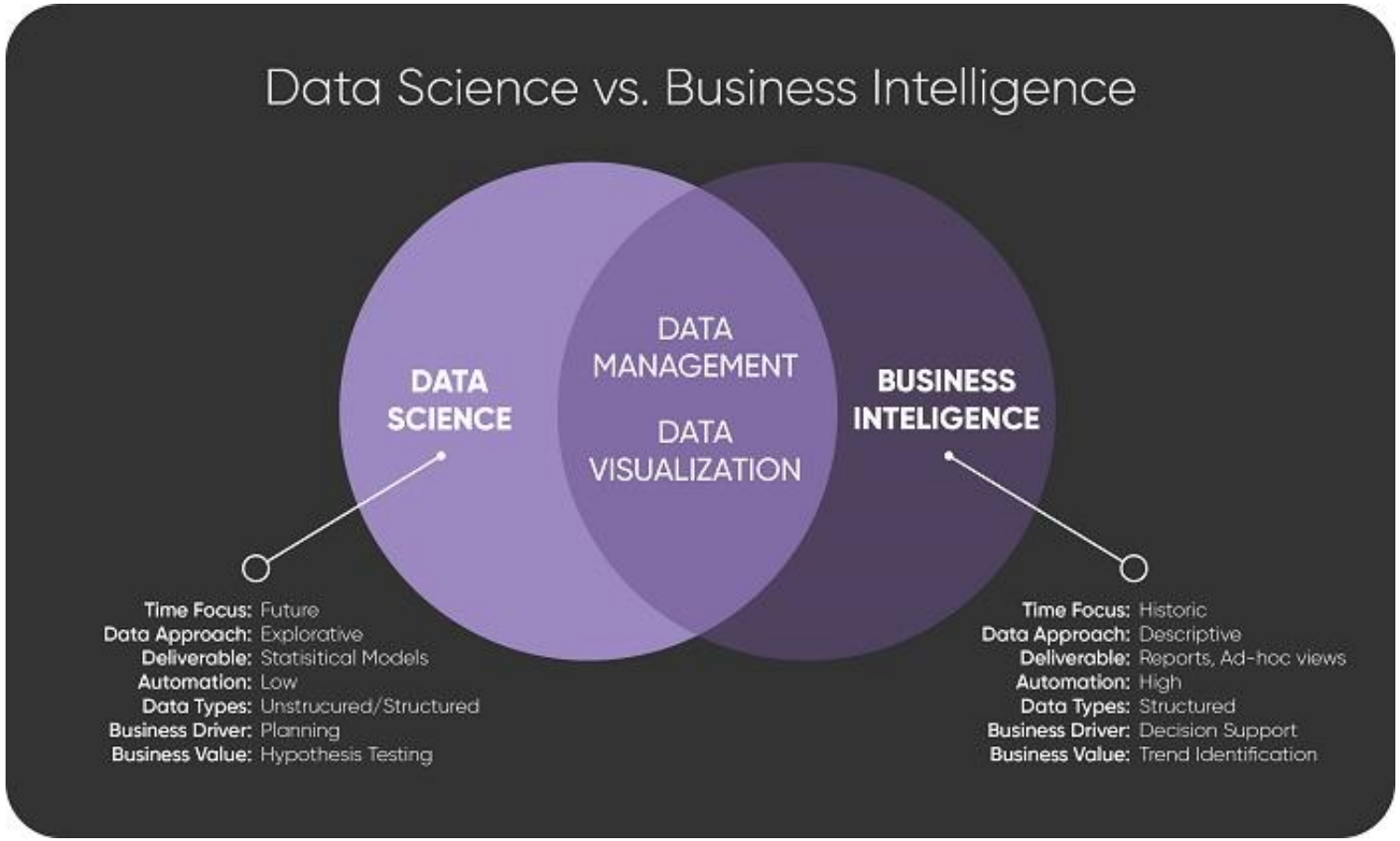
What to do with Data?



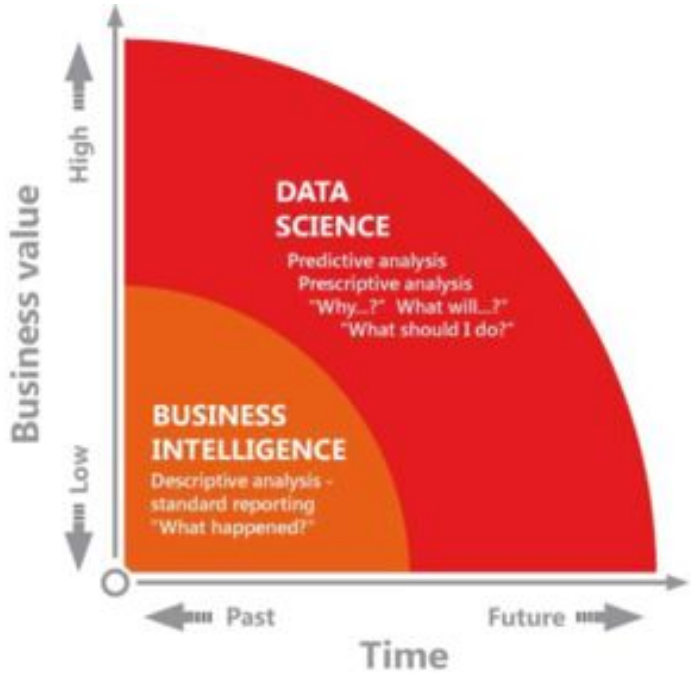


Adapted from Creating Data Driven Organization, Carl Anderson

Data Science versus Business Intelligence



DS x BI



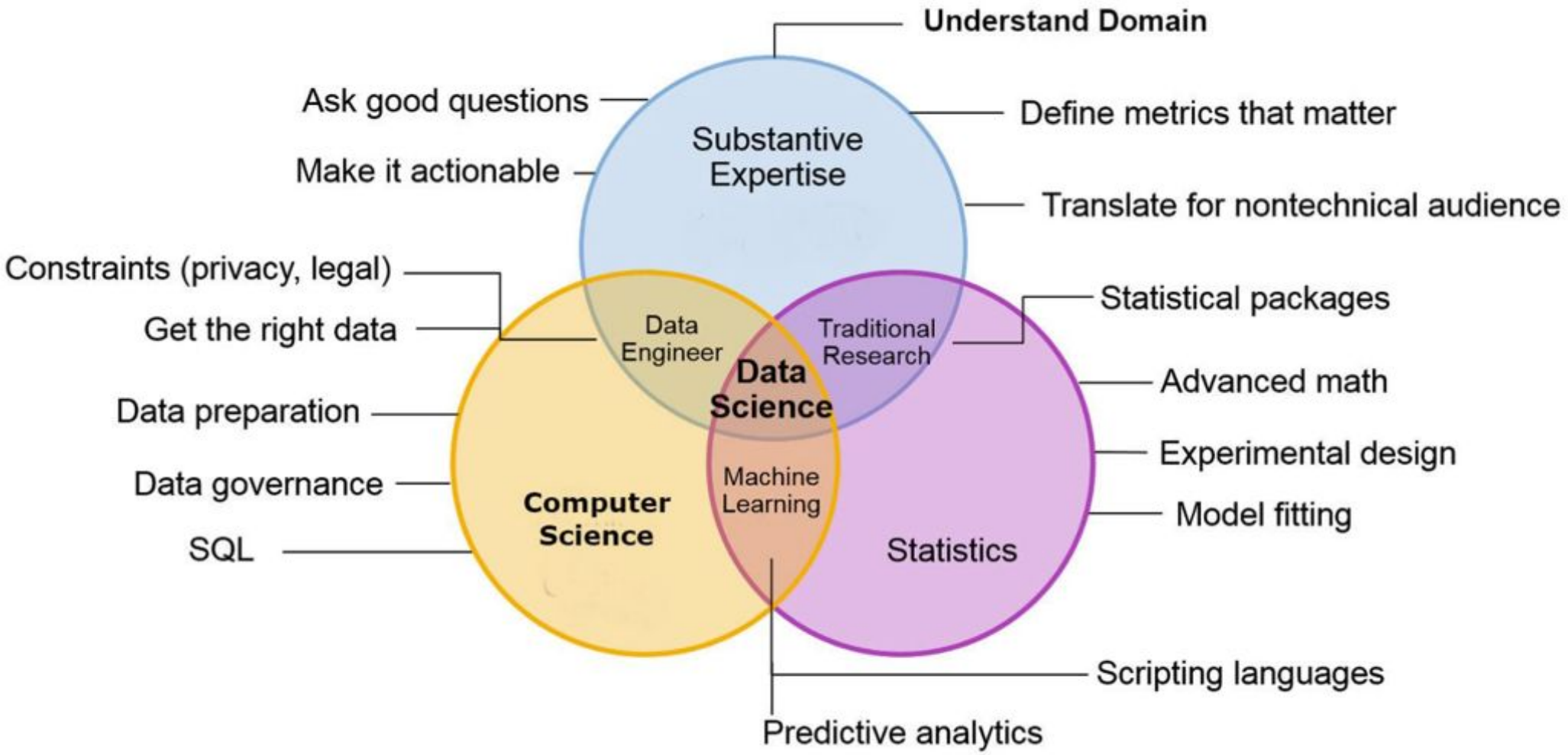
https://tdwi.org/articles/2017/12/05/bi-all-understanding-differences-data-science-and-bi.aspx

Difference Between Data Science and Business Analytics

Data Science	Business Analytics
It is the science of Data study using statistics, algorithms and technology	It is the statistical study of business data
Uses both structured and unstructured data	Uses mostly structured data
It is a combination of traditional analytics practices with sound computer science knowledge including coding	It is oriented more towards statistics and does not involve much coding
Top industries where Data Science finds its applications are: a. Technology b. Finance c. E-commerce d. Academic	Top Industries where Business analytics finds its applications are a. Finance b. Technology c. Marketing d. Retail
The future applications of Data Science would be witnessed in Artificial Intelligence and Machine Learning	The future applications of Business Analytics would be witnessed in Cognitive Analytics and Tax analytics
Data Science results give insights but usually not used for making Business Decisions	Business Analytics results are vital to the key decision makers
Studies trends and patterns	Works on specific business problems

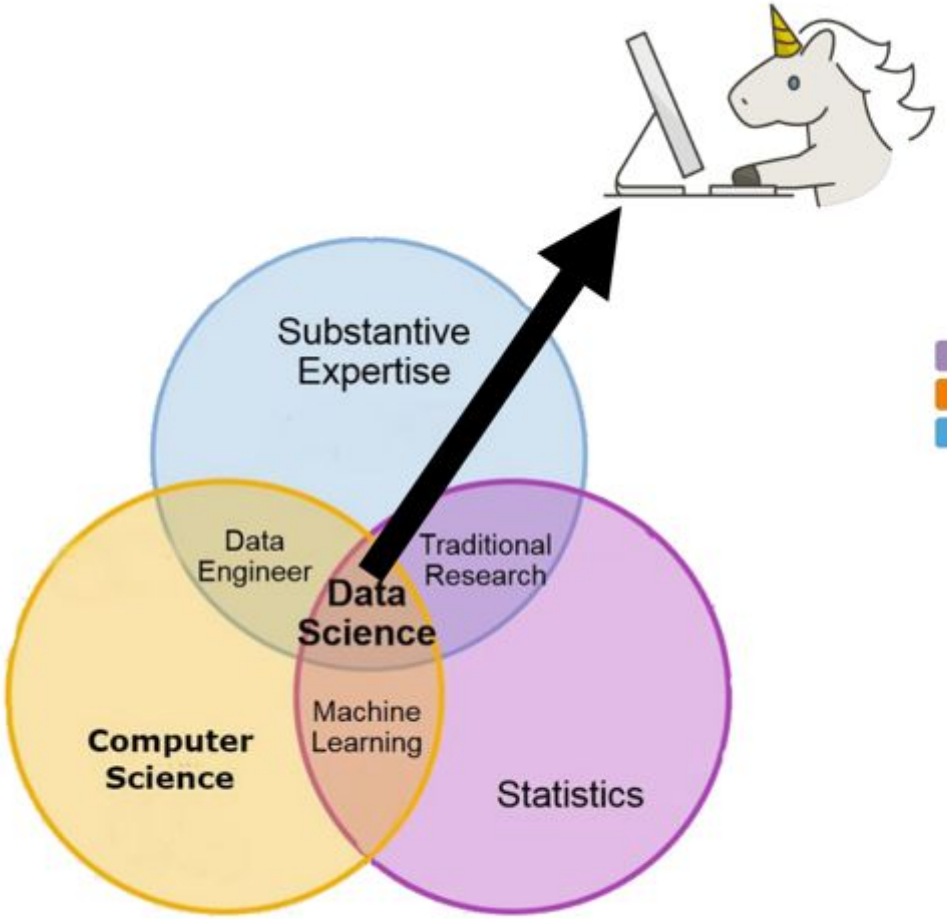
https://www.mygreatlearning.com/blog/difference-between-data-science-business-analytics/

What is a Data Scientist?



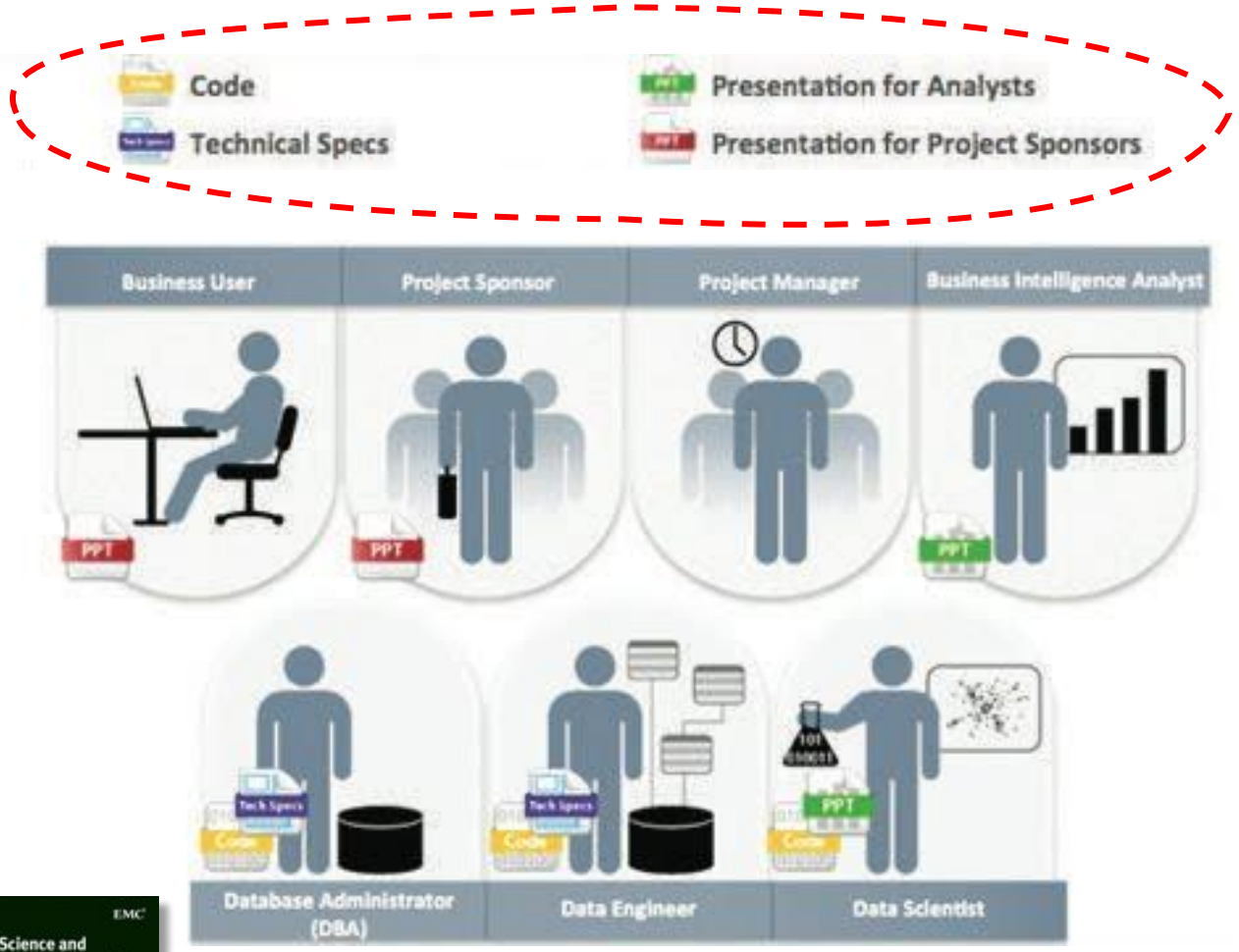
Loreta Bravo, Notes from Class DDO – MDS UDD

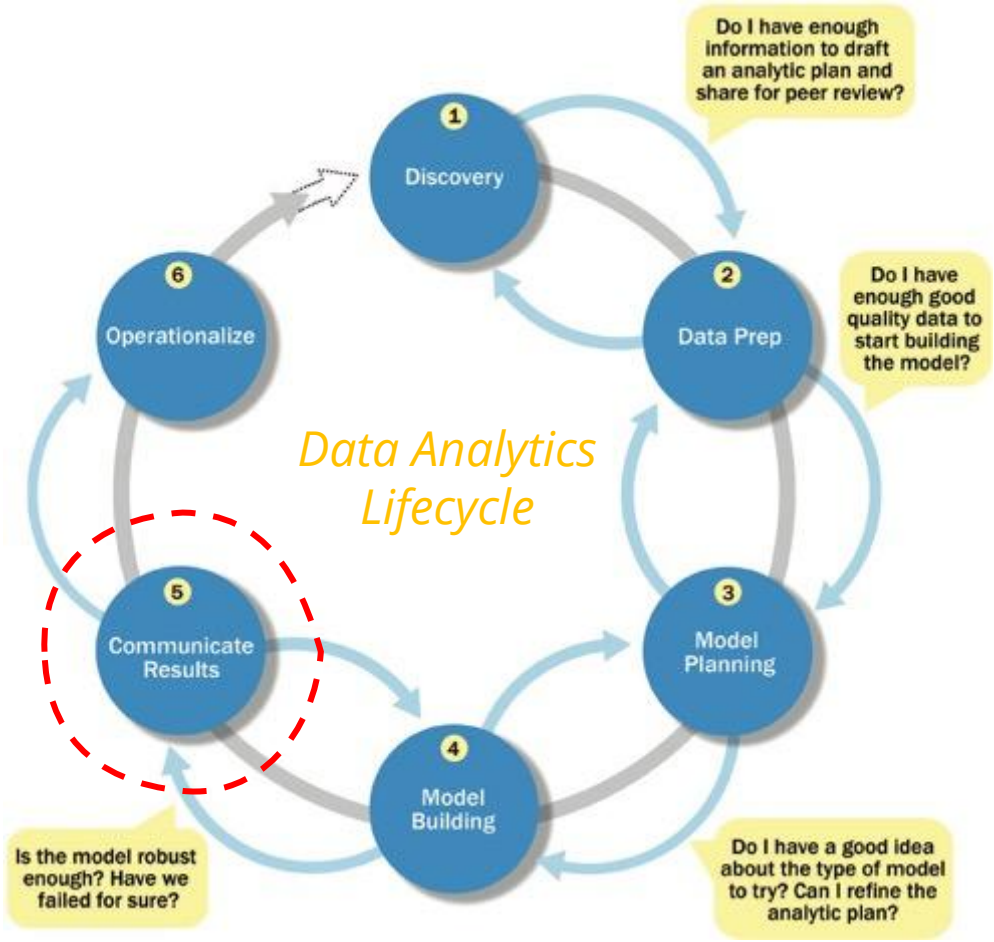
Data Scientist is not a person, is a team!



Loreta Bravo, Notes from Class DDO – MDS UDD

The fourth leg: Visualization!

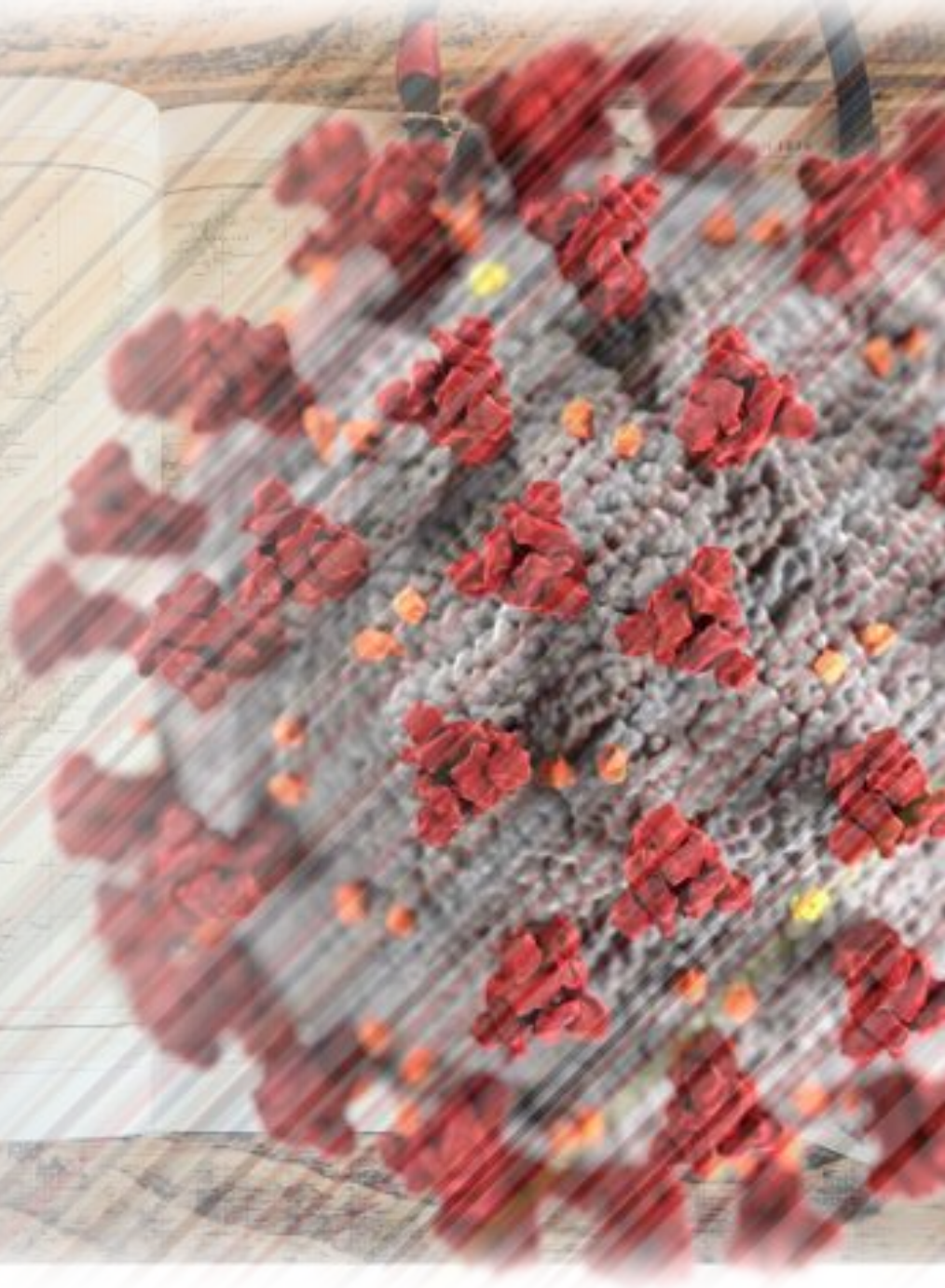


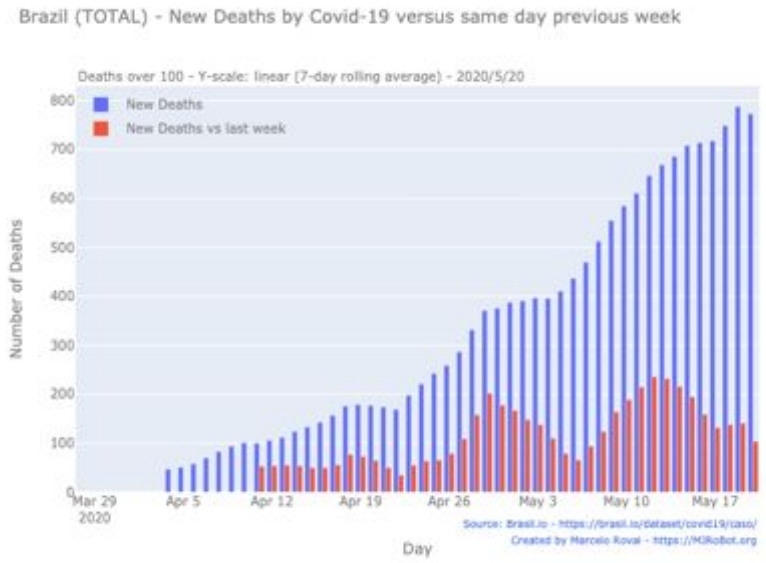


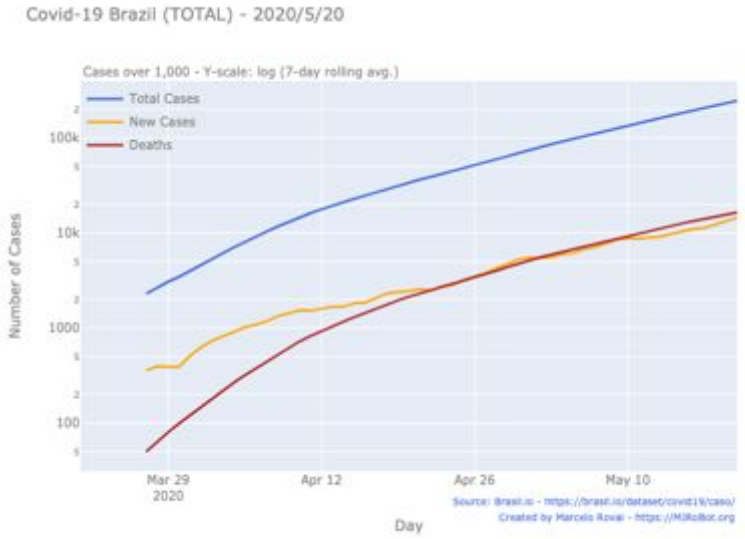
Data Science Project Examples

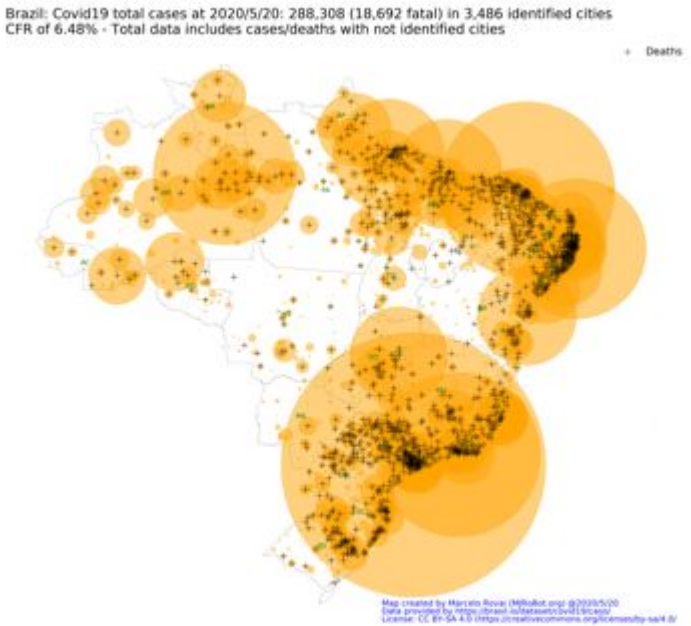
Mapping in the Time of Pandemic

Tracking the pandemic outbreak in Brazil









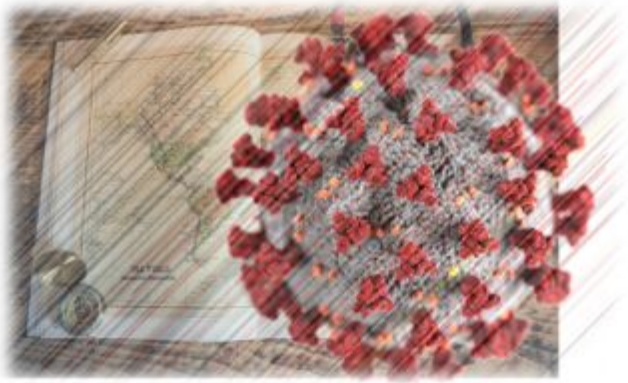
https://mjrovai.github.io/Mapping_Covid-19_Brazil/

Mapping_Covid-19_Brazil

View on GitHub

Mapping in the Time of Pandemic

Tracking the pandemic outbreak in Brazil



> The primary purpose of this repository is to help people to develop their graphs, maps, and videos using Python, GeoPandas, and Plotly. I will try to keep it updated in terms of Covid-19 Pandemic information in Brazil, but I strongly suggest you visit the below sites for more info.

More information:

- https://wcota.me/covid19br
- https://brasil.io/covid19/

Content of subfolders:

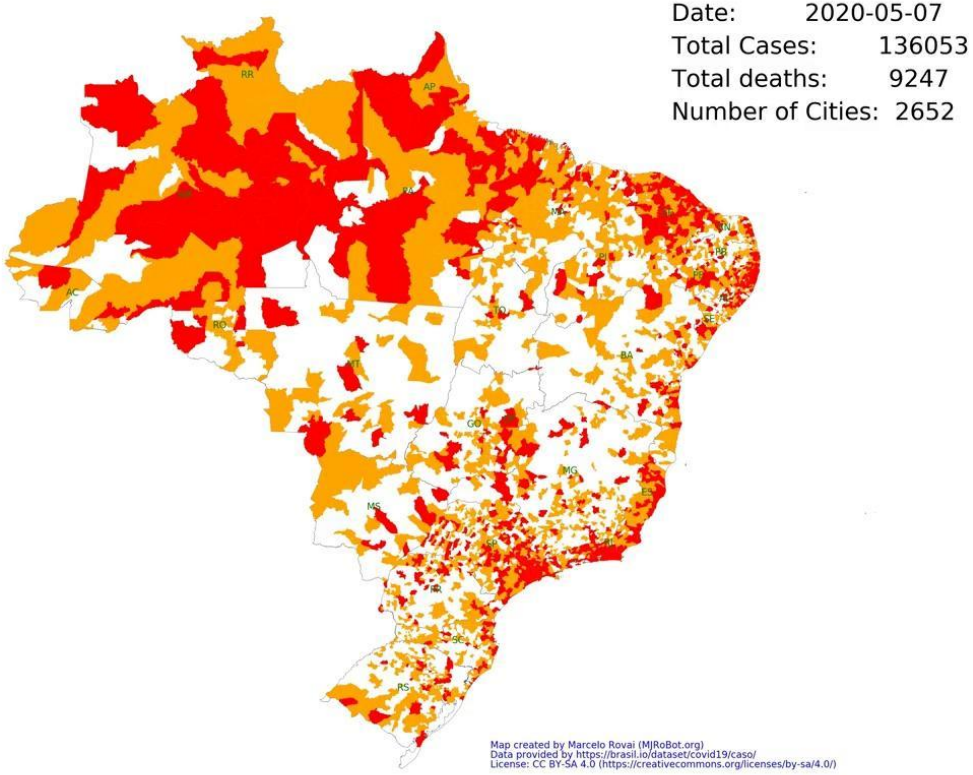
- /videos - Short videos showing the development of Covid-19 in Brazil and main States.
- /graphs - Timeline graphs of main brazilian cities
- /images - Most updated maps
- /notebooks - Code used to get Graphs, Maps, Gifs, and Videos

Datasets:

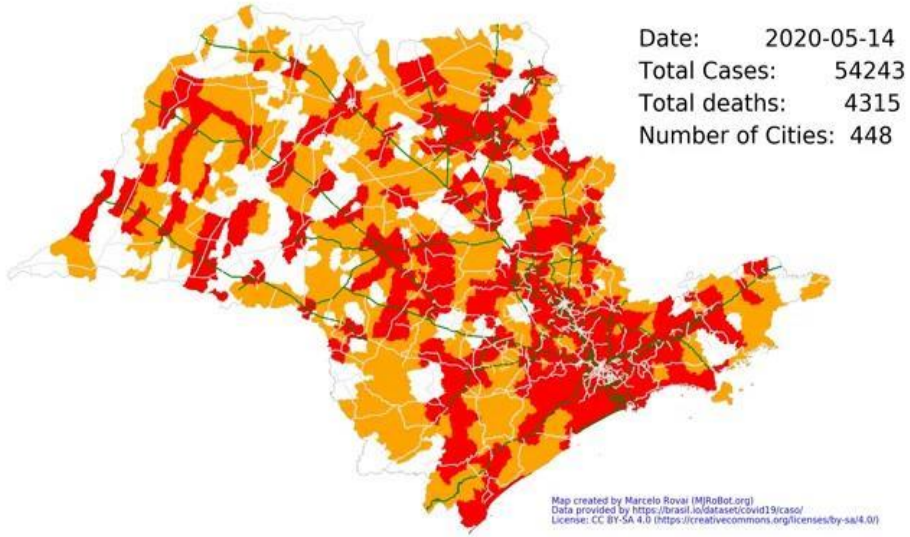
- Worldometers Daily Data
- Confirmed cases by day, using information from the news
- Raw data by city compiled from original dataset provided by Brasil.IO

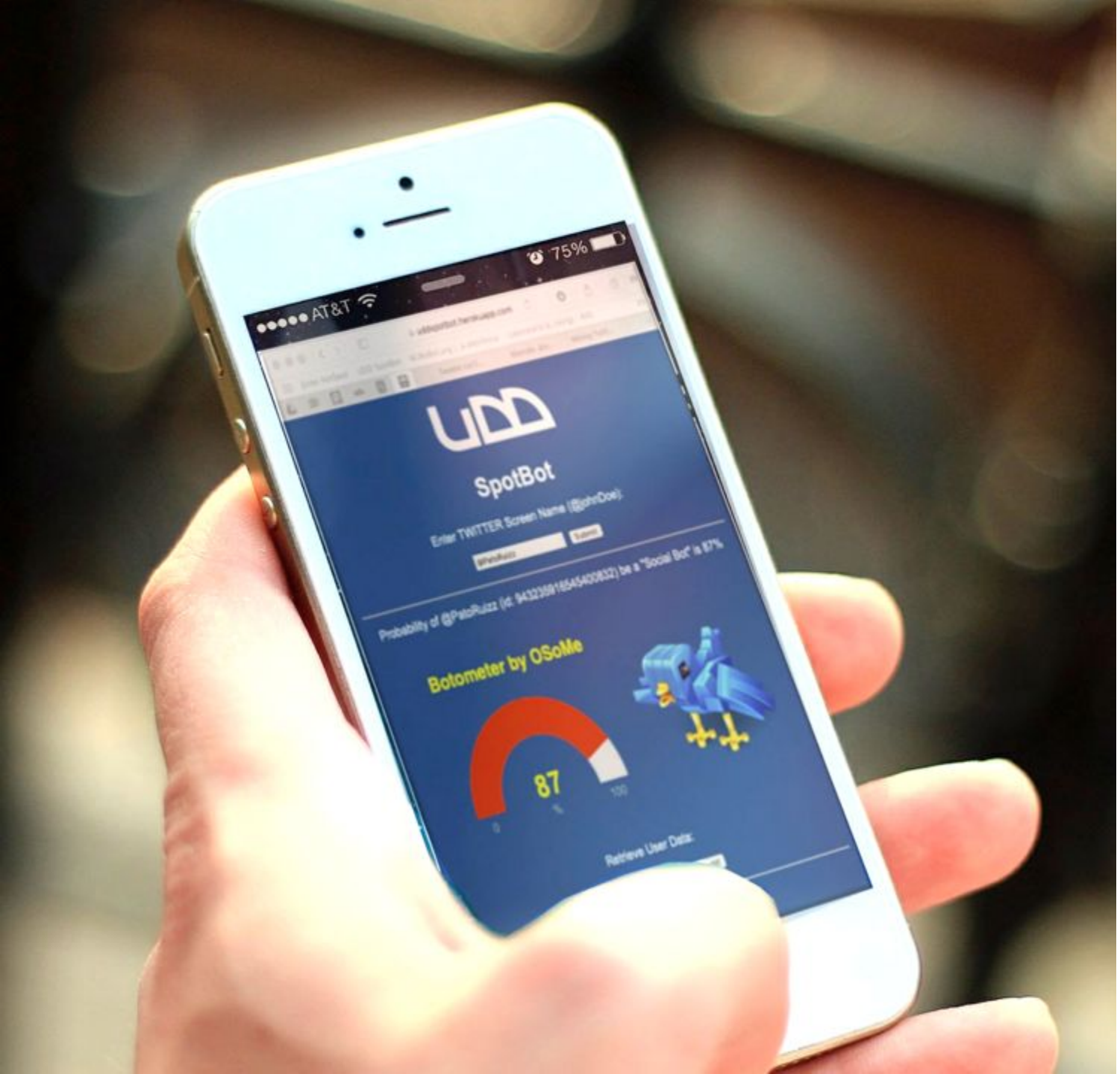
Outbreak Spreading – Timeline

Brazilian cities reported with Covid19 cases (orange) and deaths (red))



SP state cities reported with Covid19 cases (orange) and deaths (red)

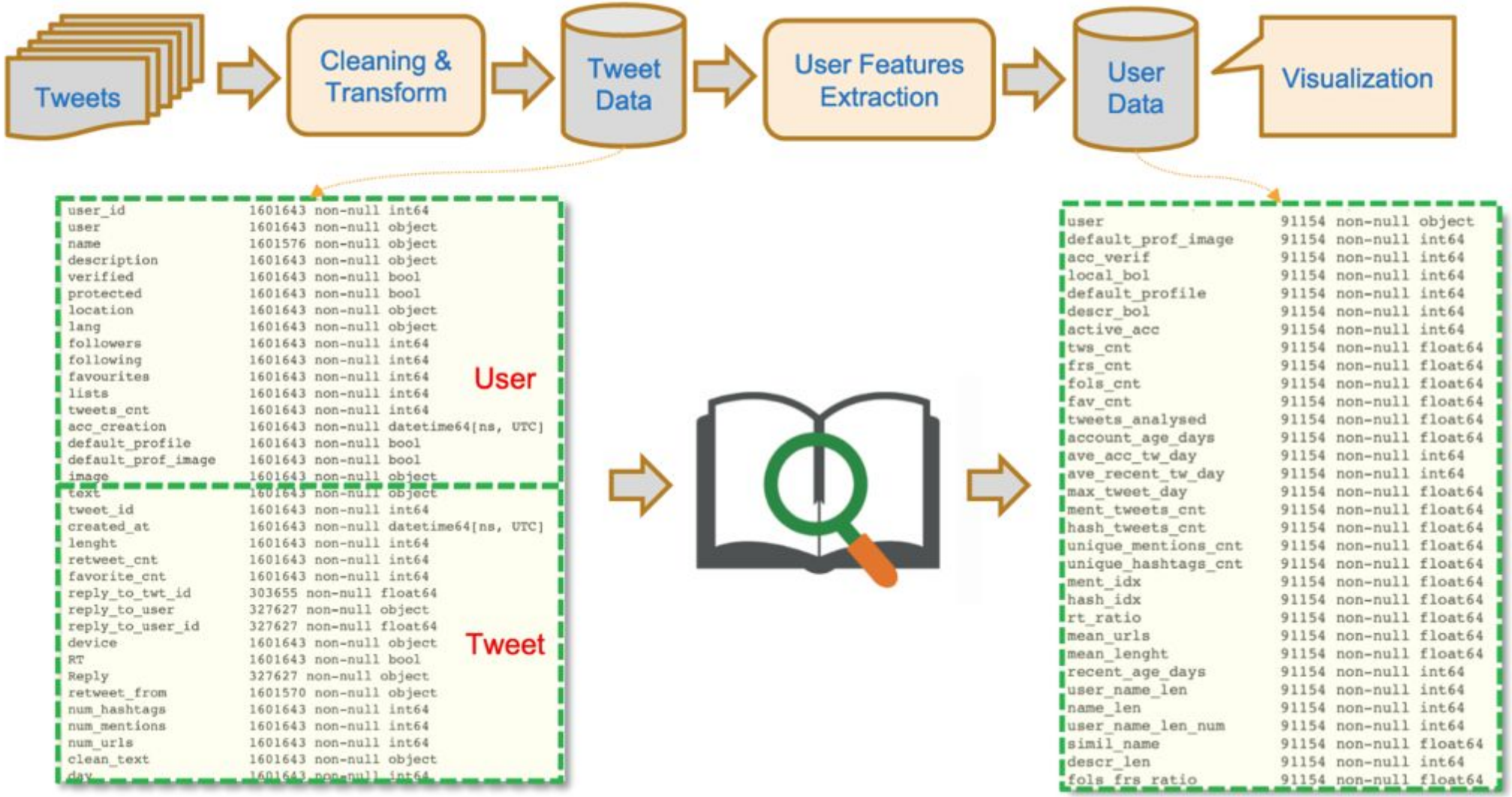




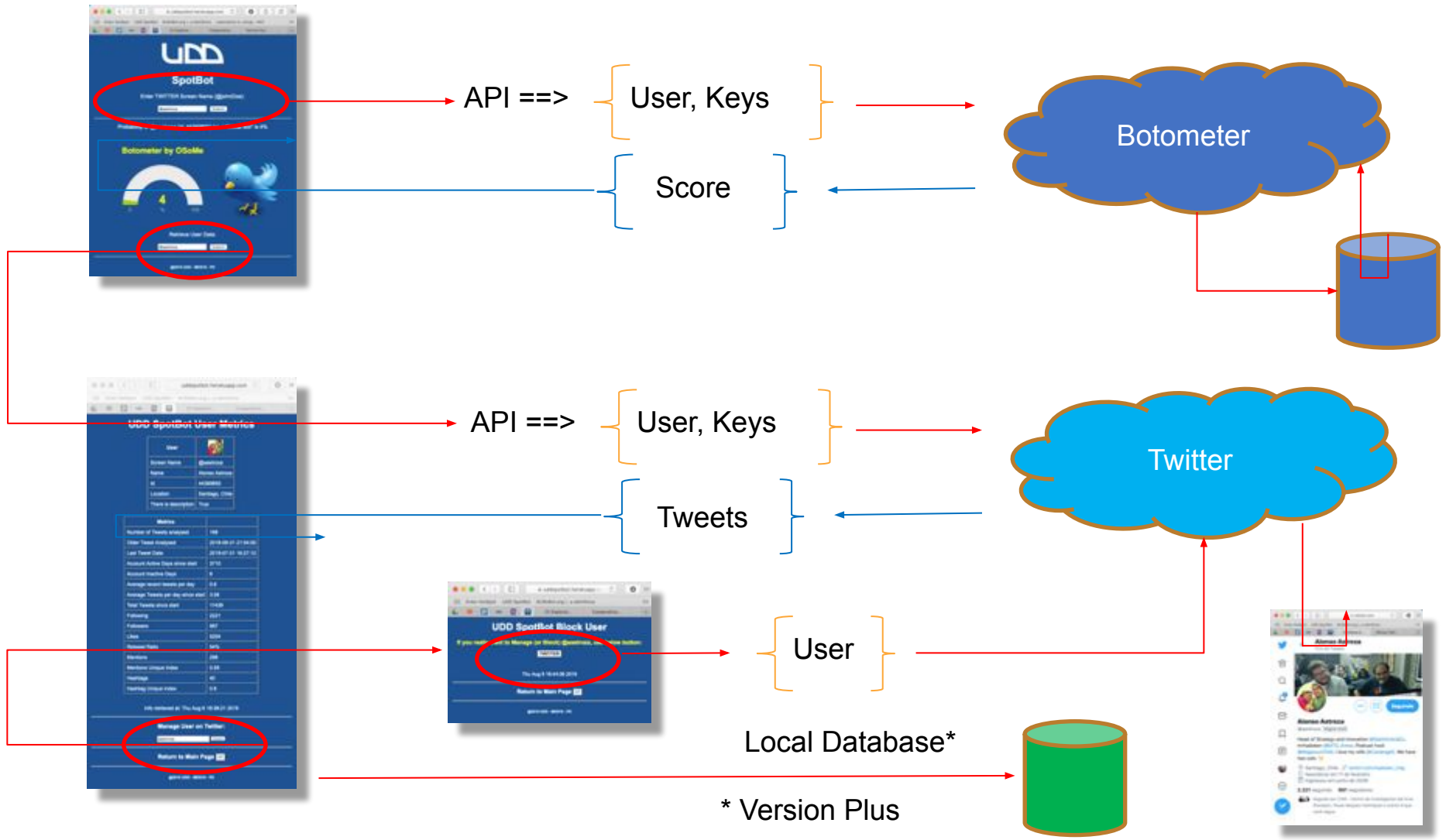
SpotBot

Detecting Social Bots in Twitter

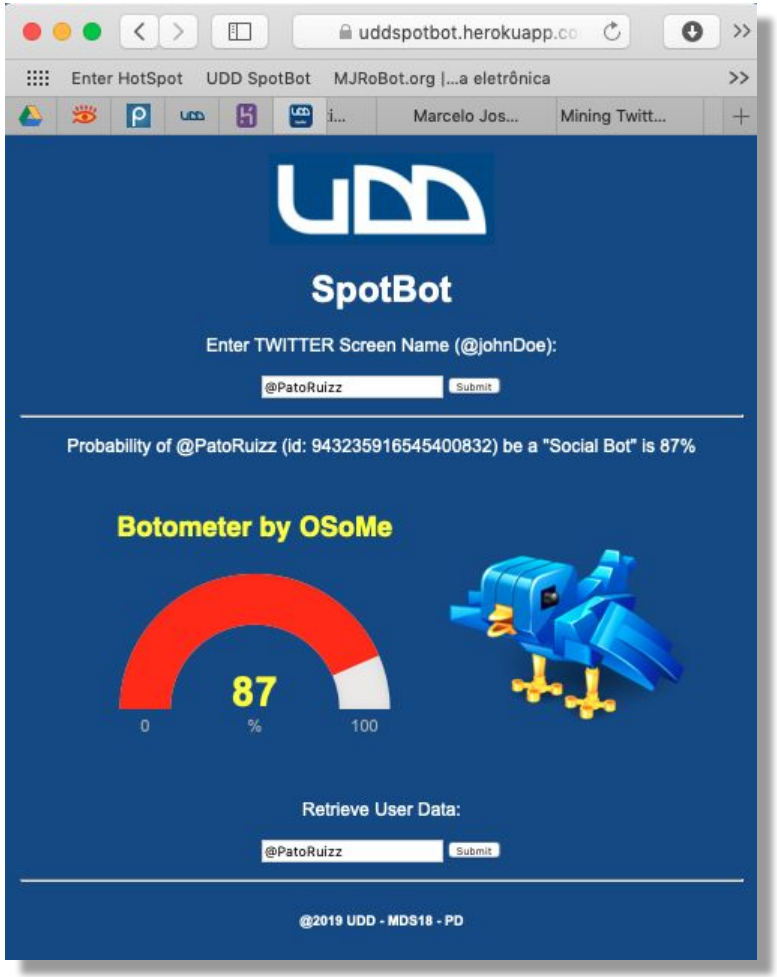
Extracting Twitter User Features

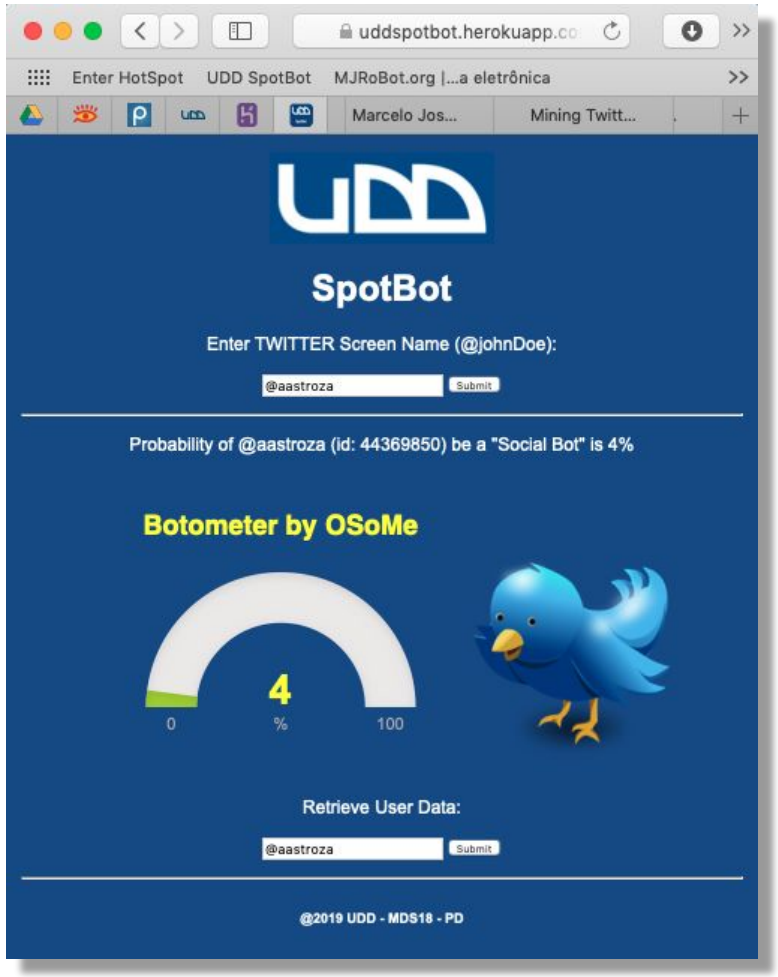


How it works?

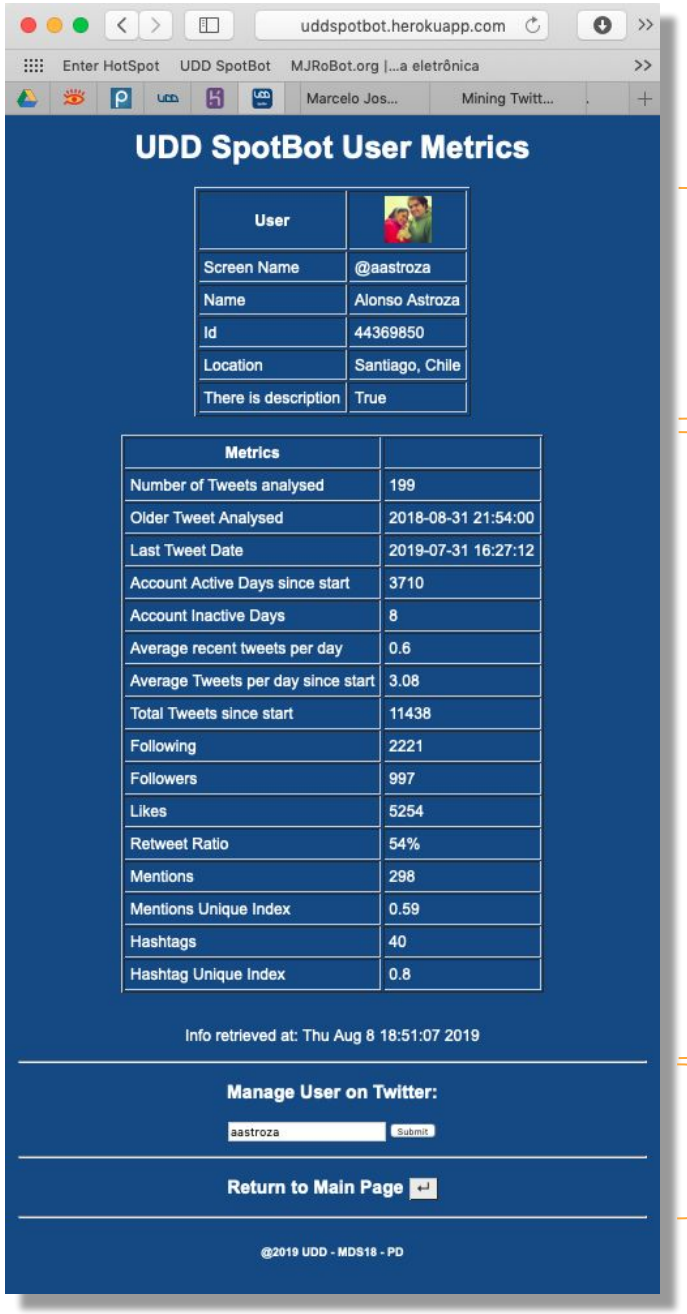

SpotBot
API ==>
User, Keys
Score
Botometer
API ==>
User, Keys
Tweets
Twitter
UDD SpotBot Block User
User
Local Database*
* Version Plus

Screen #1


uddspotbot.herokuapp.co
Enter HotSpot
UDD SpotBot
MJRoBot.org |...a eletrônica
Marcelo Jos...
Mining Twitt...
SpotBot
Enter TWITTER Screen Name (@johnDoe):
@PatoRuizz
Submit
Probability of @PatoRuizz (id: 943235916545400832) be a "Social Bot" is 87%
Botometer by OSoMe
87
0
%
100
Retrieve User Data:
@PatoRuizz
Submit
@2019 UDD - MDS18 - PD


uddspotbot.herokuapp.co
Enter HotSpot
UDD SpotBot
MJRoBot.org |...a eletrônica
Marcelo Jos...
Mining Twitt...
SpotBot
Enter TWITTER Screen Name (@johnDoe):
@aastroza
Submit
Probability of @aastroza (id: 44369850) be a "Social Bot" is 4%
Botometer by OSoMe
4
0
%
100
Retrieve User Data:
@aastroza
Submit
@2019 UDD - MDS18 - PD

Screen #2



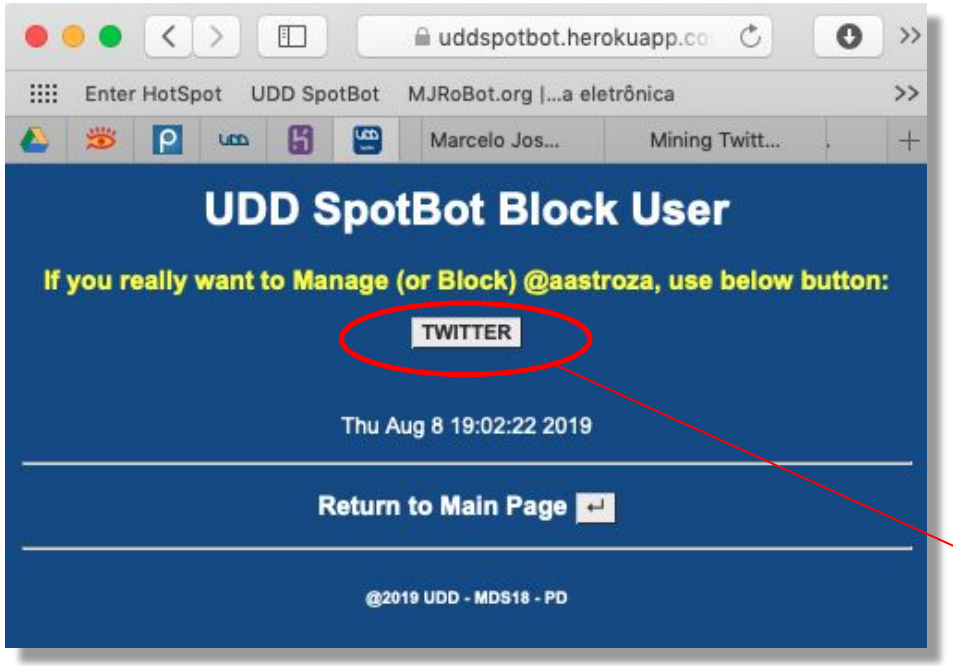
Twitter User Data

Features generated from real tweets (last 200)

Possible Actions from User

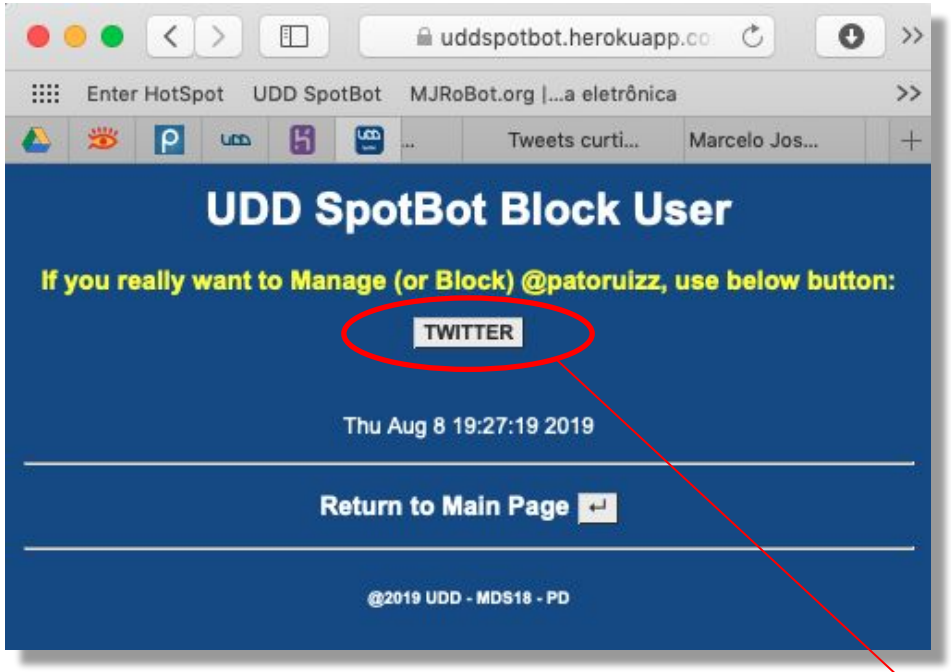
- Manage / Block user on Twitter
- Return to Main Screen #1

Screen #3


uddspotbot.herokuapp.co
Enter HotSpot
UDD SpotBot
MJRoBot.org |...a eletrônica
Marcelo Jos...
Mining Twitt...
UDD SpotBot Block User
If you really want to Manage (or Block) @aastroza, use below button:
TWITTER
Thu Aug 8 19:02:22 2019
Return to Main Page
@2019 UDD - MDS18 - PD


twitter.com
Enter HotSpot
UDD SpotBot
MJRoBot.org |...a eletrônica
(1) Alonso A...
Mining Twitt...
Alonso Astroza
11,4 mil Tweets
Seguindo
Desativar Retweets
Adicionar/remover das listas
Ver Listas
Visualizar Moments
Compartilhar perfil por...
Copiar link para o perfil
Silenciar @aastroza
Bloquear @aastroza
Denunciar @aastroza
@GeoVictoriaCL.
ast host
@CsolangeS. We have
/mrhadoken_mtg
Mídia

Screen #3



https://github.com/Mjrovai/UDD_Master_Data_Science/tree/master/PD_SpotBot_Final_Project

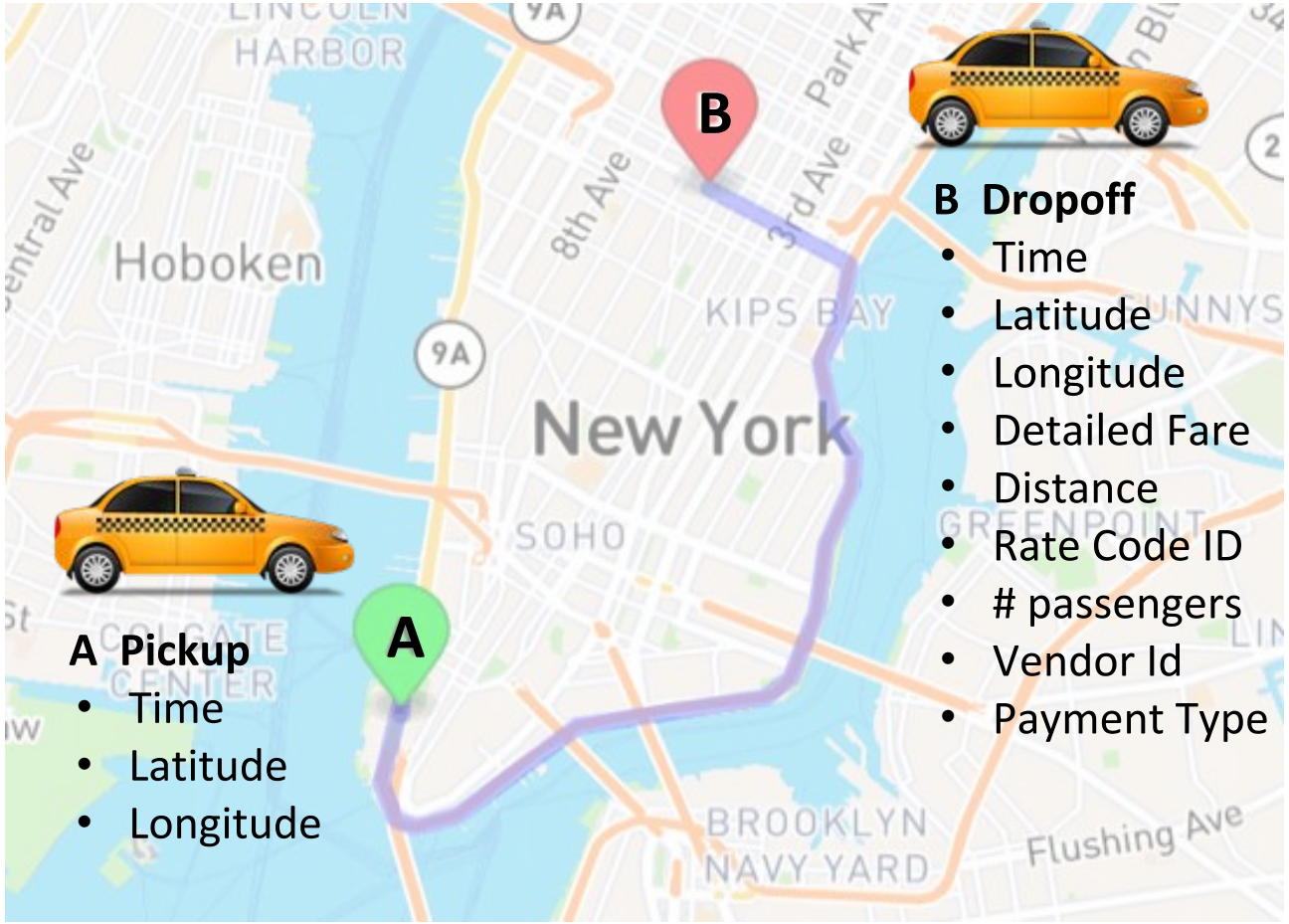
NYC Taxi – Trip Prediction Project

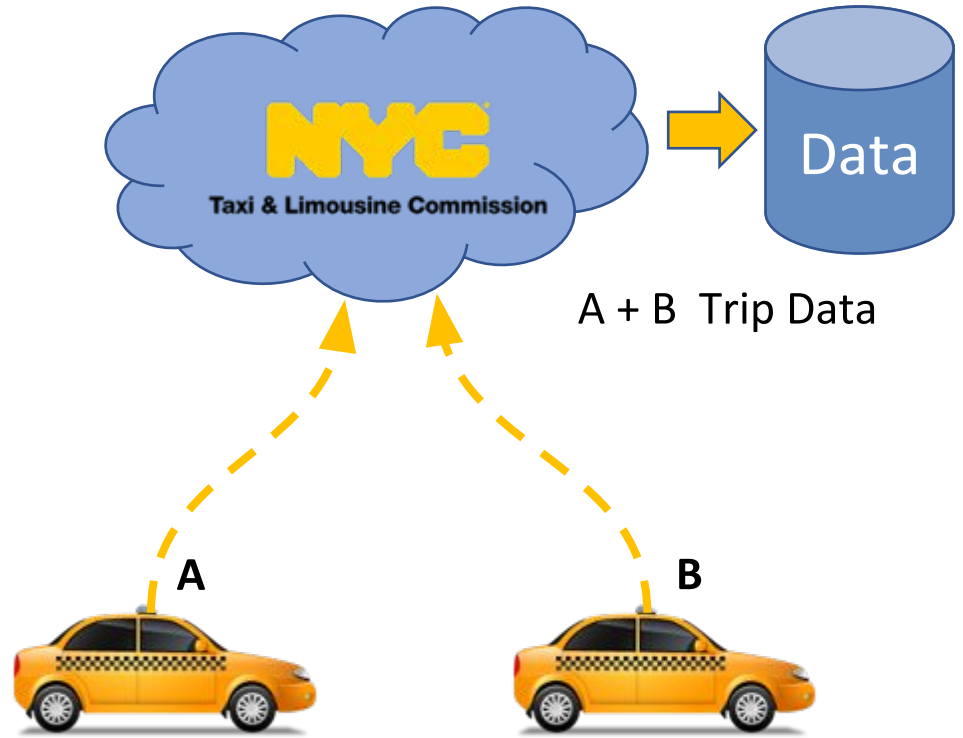
NYC Taxi Trip Prediction Project

Ds_A 14M Trip dataset (full June/15) □ EDA □ Sample of 120K dataset □ Machine Learning

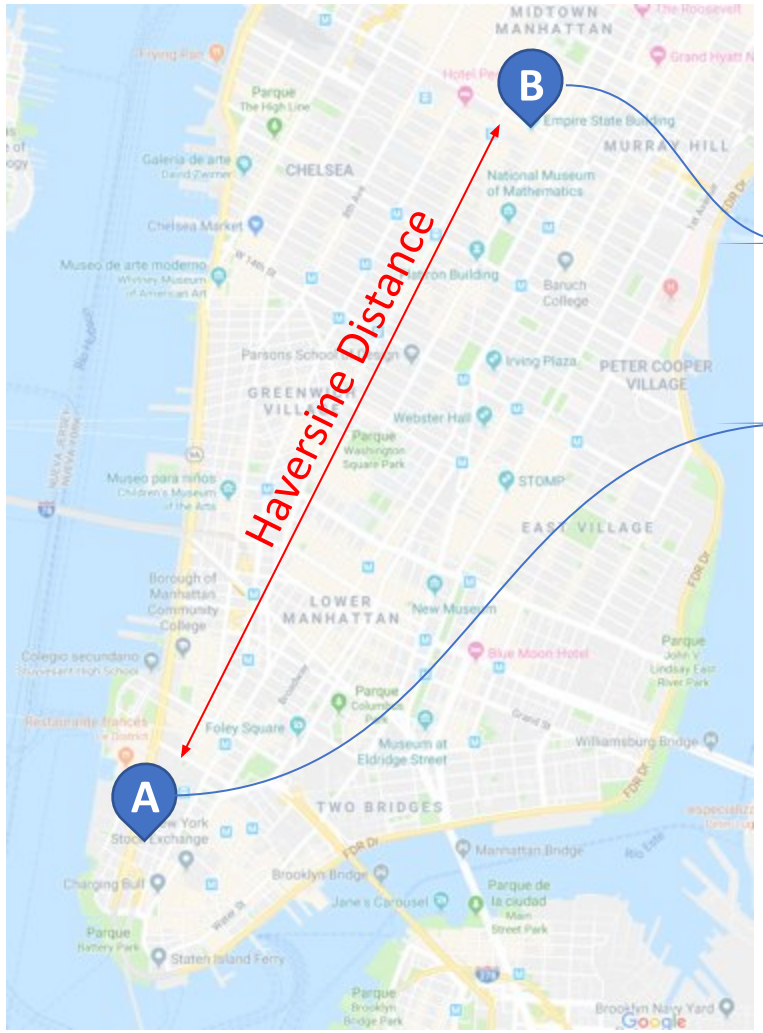
Ds_B 120K Trip dataset (sample all months 2015) □ EDA and Machine Learning

NYC Taxi – IoT – Getting data





NYC Taxi Trip Prediction - ML



Dropoff Latitude
Dropoff Longitude

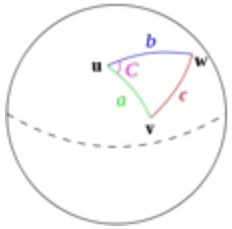
Pickup Latitude
Pickup Longitude
Pickup Time

ML Models

Fare

Distance

Duration

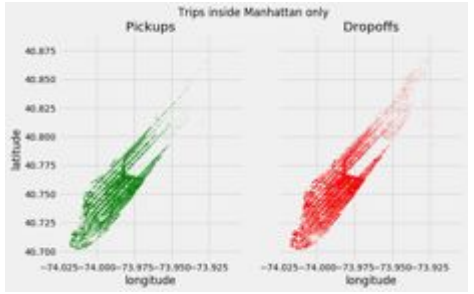


$$2r\arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2-\varphi_1}{2}\right)+\cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right)$$

Hav. Dist.

Rate Code

In/Out Manh.



NYC Taxi Trip Prediction – ML

	Fare DS_A RMSE	Fare DS_B RMSE	Comparison A-B	Fare DS_A Base line comparison	Fare DS_B Base line comparison
Base Line	10,25	9,88	3,6%	0,0%	0,0%
Decisión Tree Regresor	3,25	3,21	1,2%	68,3%	67,5%
Random Forest Regressor	3,01	2,92	3,0%	70,6%	70,4%
Random Forest Regressor (Reducing features)	3,00	2,92	2,7%	70,7%	70,4%
Random Forest Regressor (Changing Hyper_param)	2,77	2,72	1,8%	73,0%	72,5%

	Time DS_A RMSE	Time DS_B RMSE	Comparison A-B	Time DS_A Base line comparison	Time DS_B Base line comparison
Base Line	14,52	13,94	4,0%	0,0%	0,0%
Decisión Tree Regresor	11,47	10,75	6,3%	21,0%	22,9%
Random Forest Regressor	6,03	6,26	-3,8%	58,5%	55,1%
Random Forest Regressor (Reducing features)	5,60	5,73	-2,3%	61,4%	58,9%
Random Forest Regressor (Changing Hyper_param)	5,26	5,38	-2,3%	63,8%	61,4%

	Dist DS_A RMSE	Dist DS_B RMSE	Comparison A-B	Dist DS_A Base line comparison	Dist DS_B Base line comparison
Base Line	3,61	3,46	4,2%	0,0%	0,0%
Decisión Tree Regresor	0,90	0,82	8,9%	75,1%	76,3%
Random Forest Regressor	0,83	0,75	9,6%	77,0%	78,3%
Random Forest Regressor (Changing Hyper_param)	0,78	0,71	9,0%	78,4%	79,5%

ML Model: Random Forest Regressor

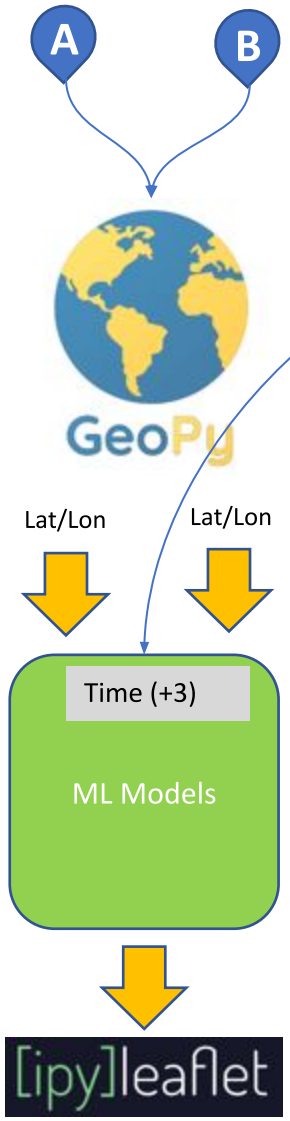
Fare: < $3.00

Time: > 5 min

Dist: > 0.7 mile

Errors

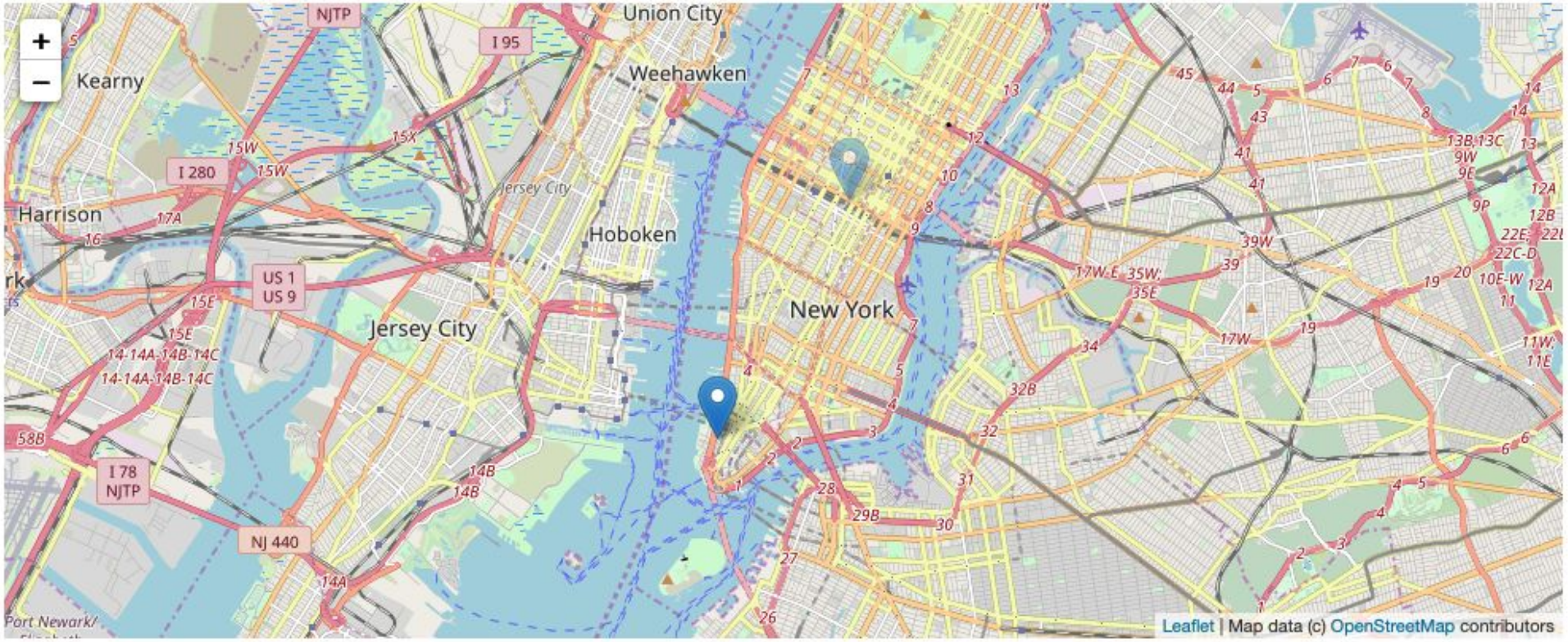
NYC Taxi Trip Prediction - Deploy




```
1  start_loc = "New York Marriott Downtown"
2  end_loc = "Empire State Building"
3
4  predict_trip(start_loc, end_loc)
```

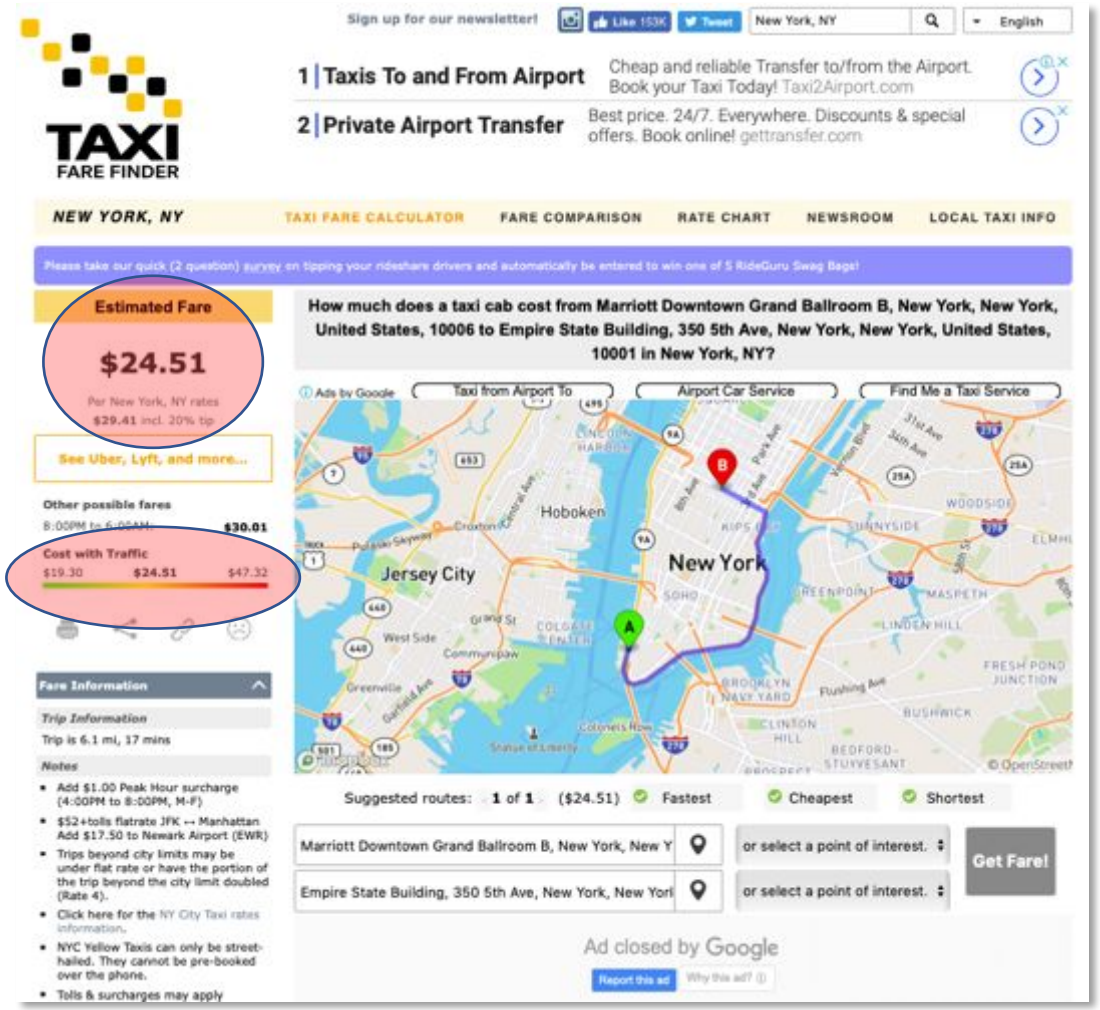
executed in 4.16s, finished 12:03:08 2019-02-08

```
Trip has 4.8 miles and will last 28.0 minutes, with a basic cost od $21.57
[Trip Info]: Trip inside Manhattan; Rate Code: Standard
```

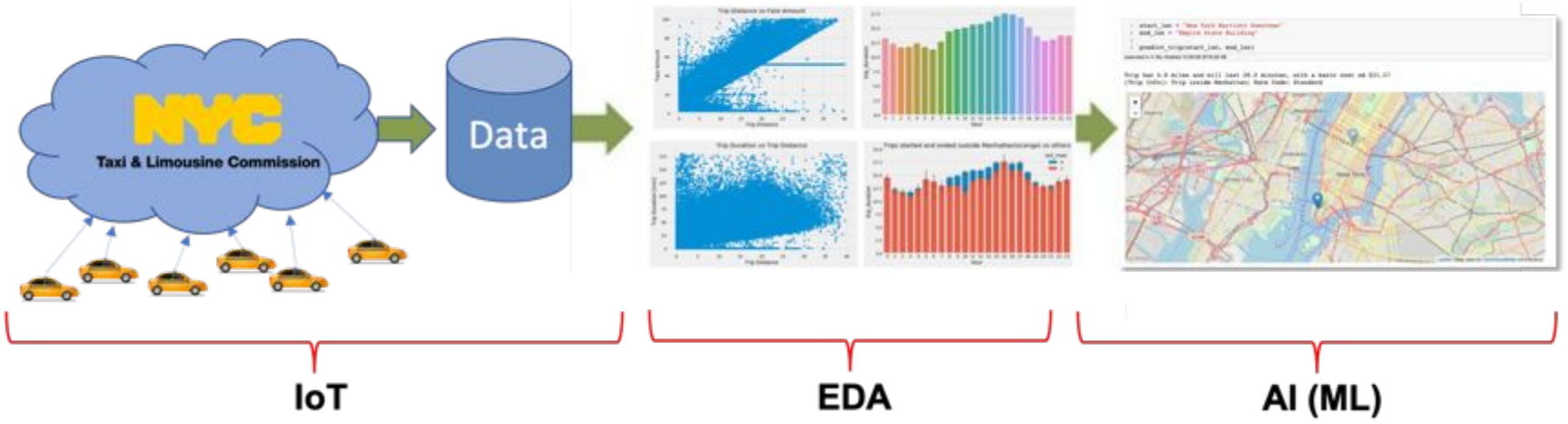


NYC Taxi Trip Prediction – Benchmark

```
Trip has 4.8 miles and will last 28.0 minutes, with a basic cost od $21.57
[Trip Info]: Trip inside Manhattan; Rate Code: Standard
```



Data Science Environment



ANACONDA

jupyter

Pandas

matplotlib

Seaborn

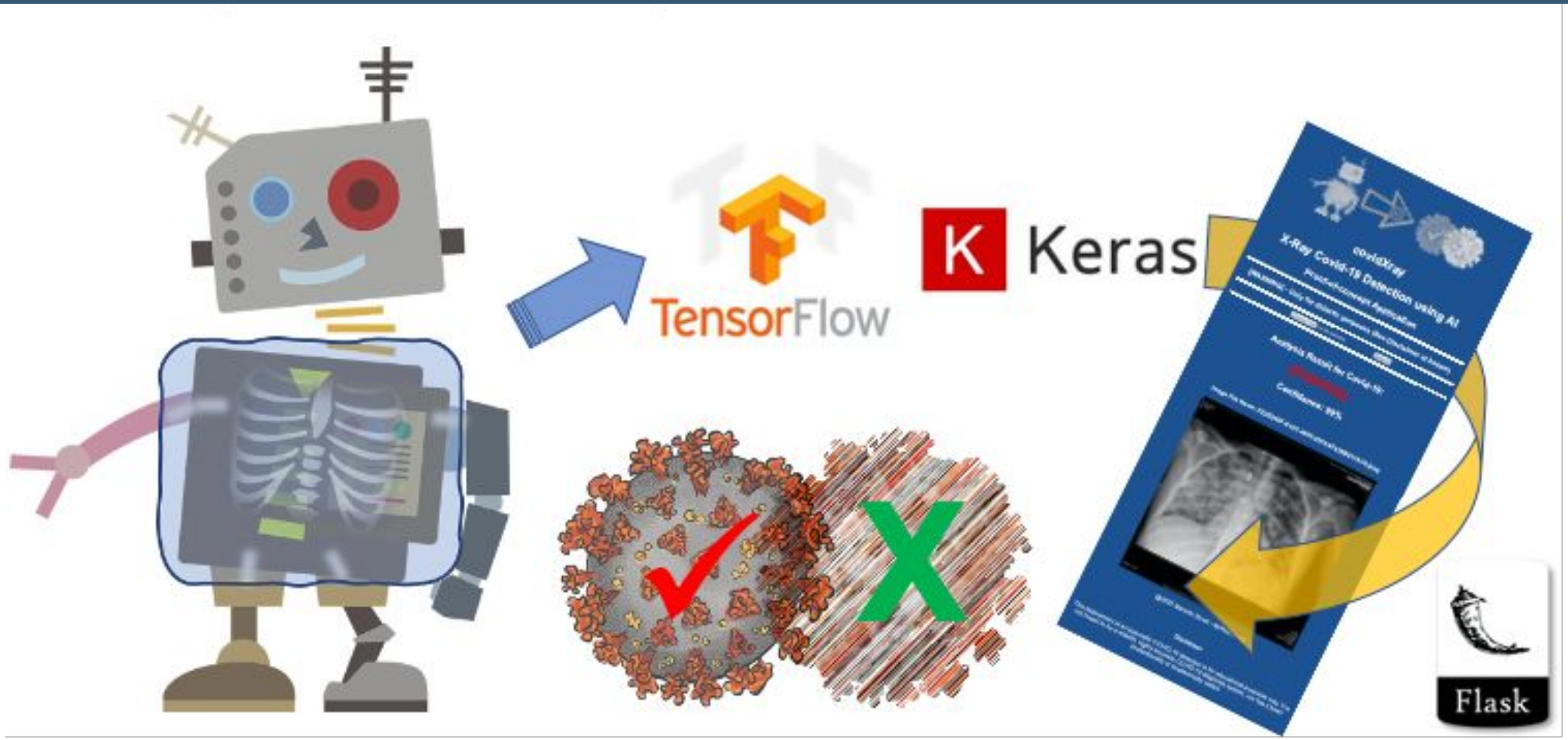
[ipy]leaflet

GeoPy

scikit learn

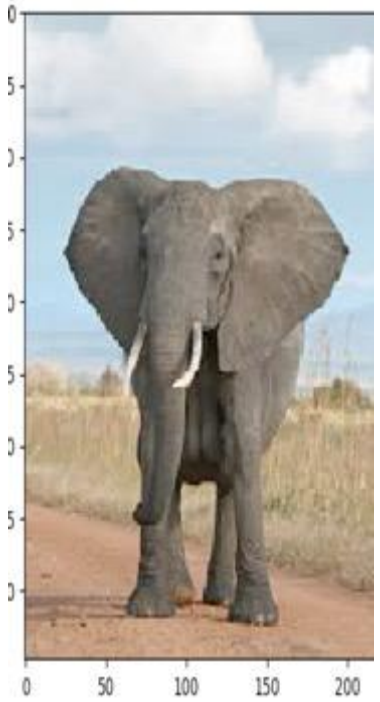
NumPy

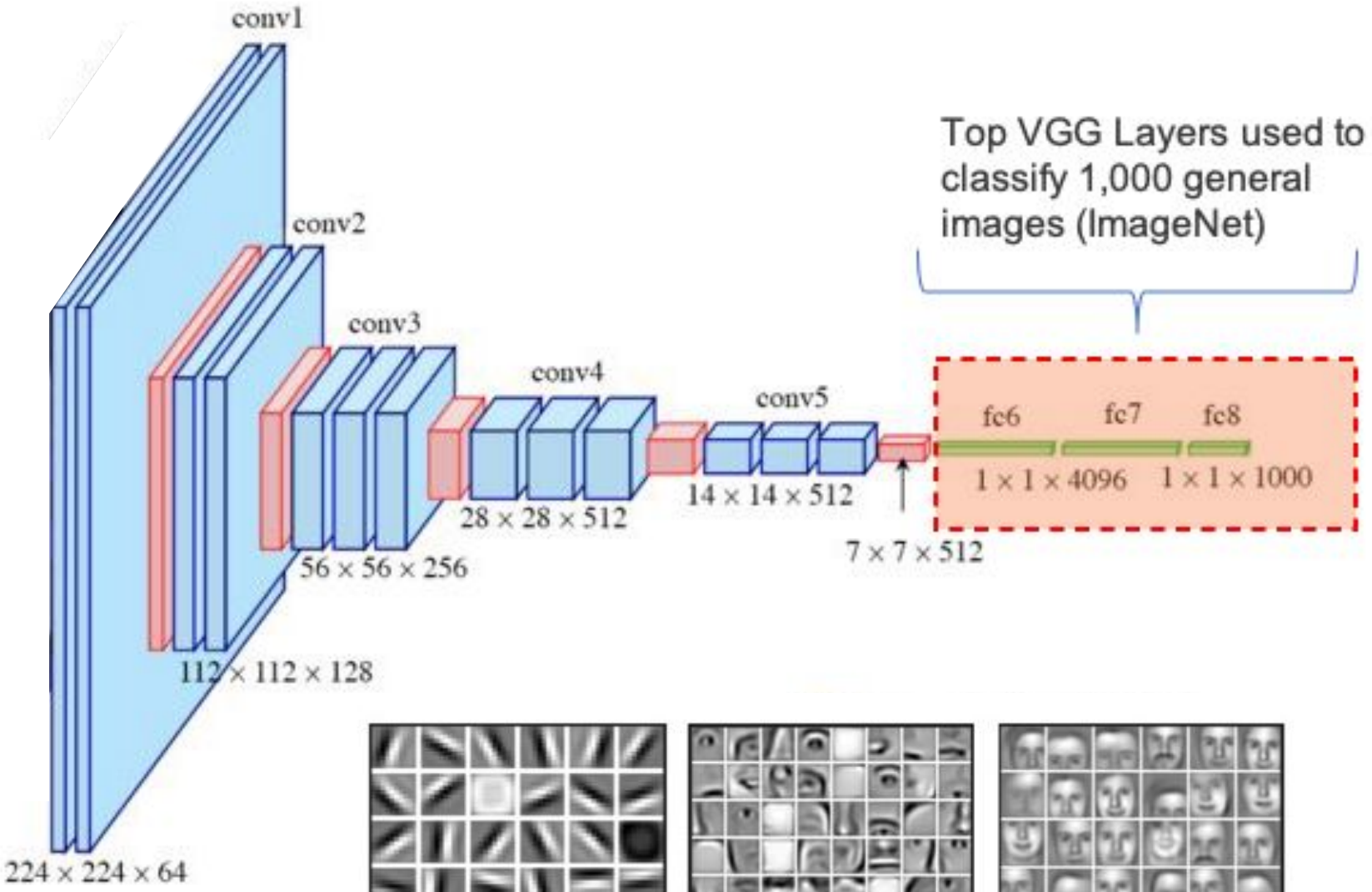
https://github.com/Mjrovai/UDD_Master_Data_Science/tree/master/AML-NYC_TAXI_TRIP_PREDICTION



Applying Artificial Intelligence techniques in the development of a web-app for the detection of Covid-19 in X-ray images

VGG-16 Convolutional Neural Network Model

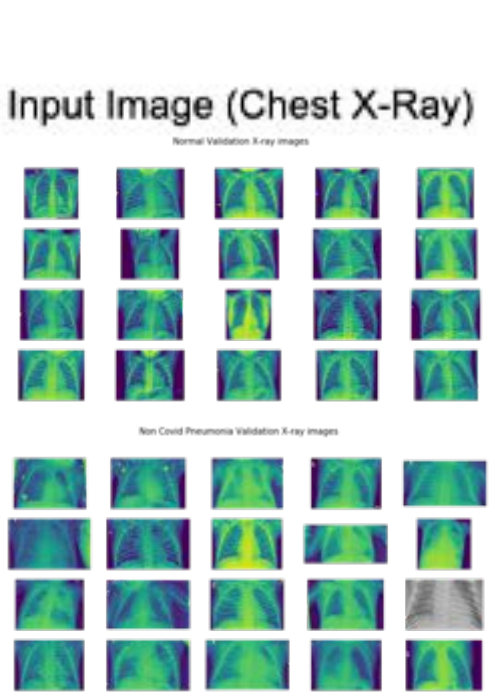


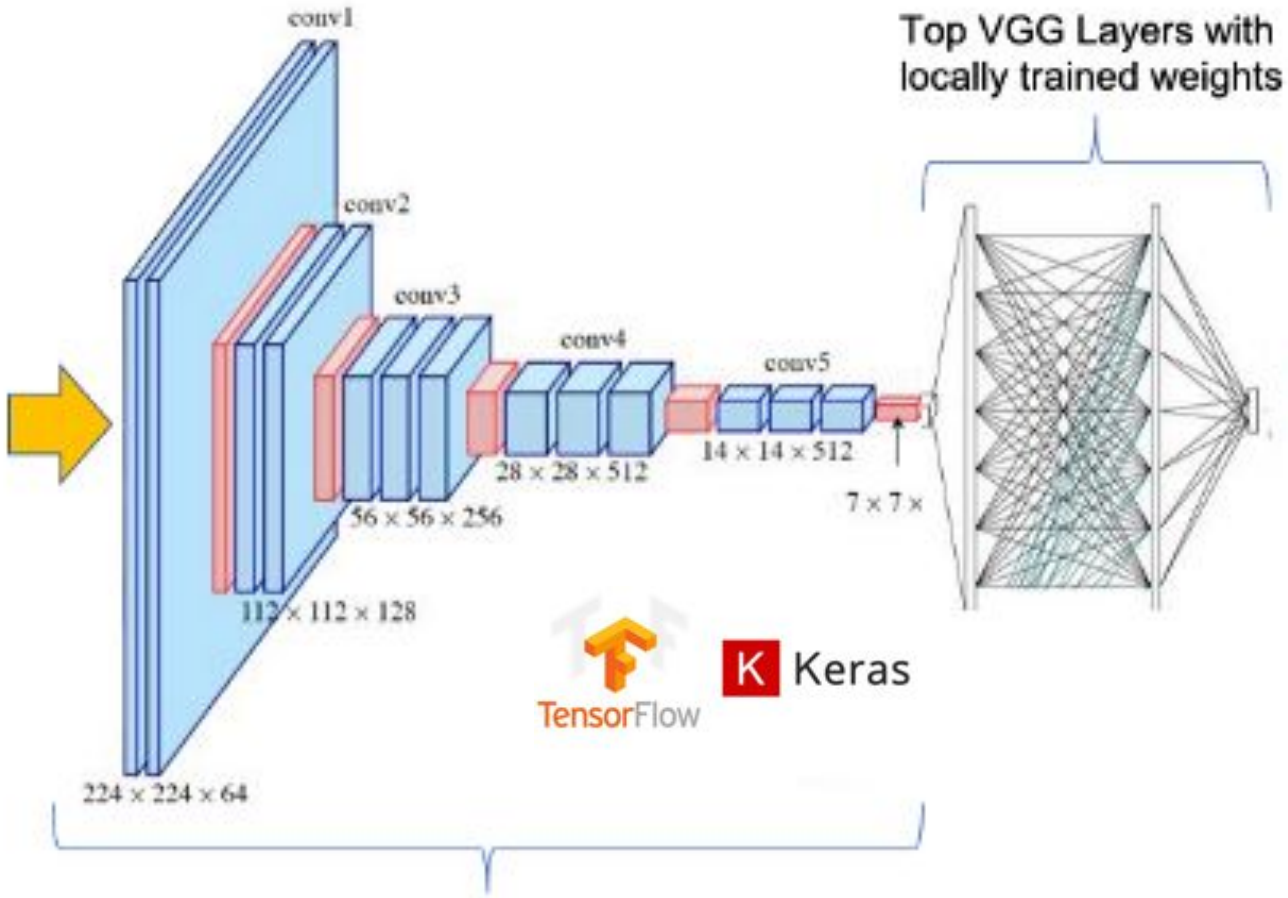


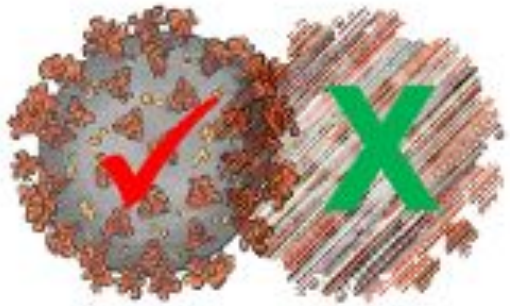
Elephant
93.0%

First Layer Representation | Second Layer Representation | Third Layer Representation

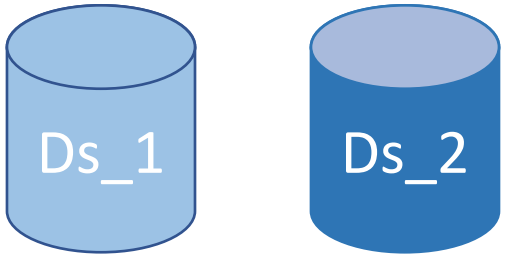
Training the model (Transfer Learning)



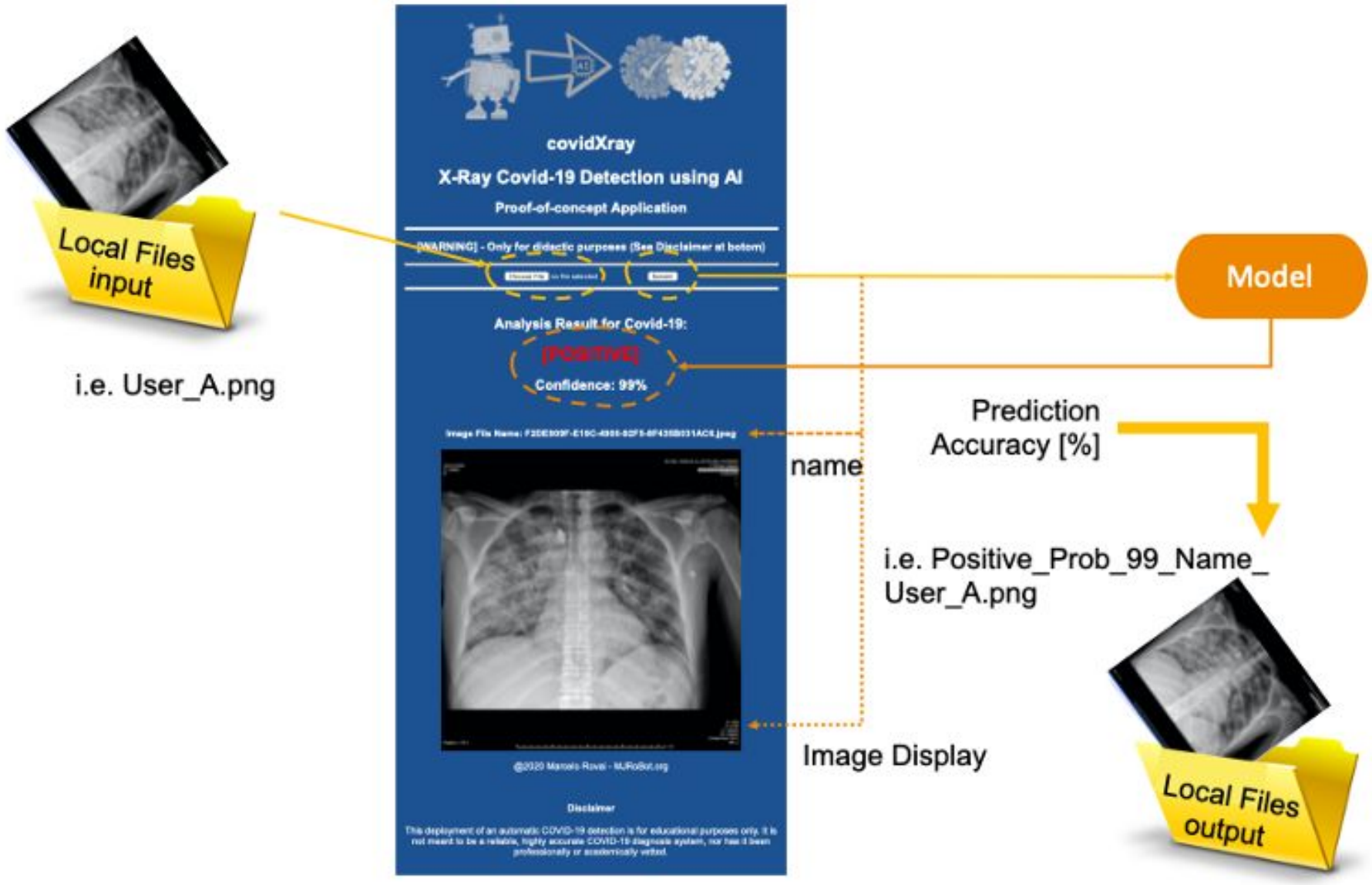




Classify:
Positive/Negative
for Covid-19

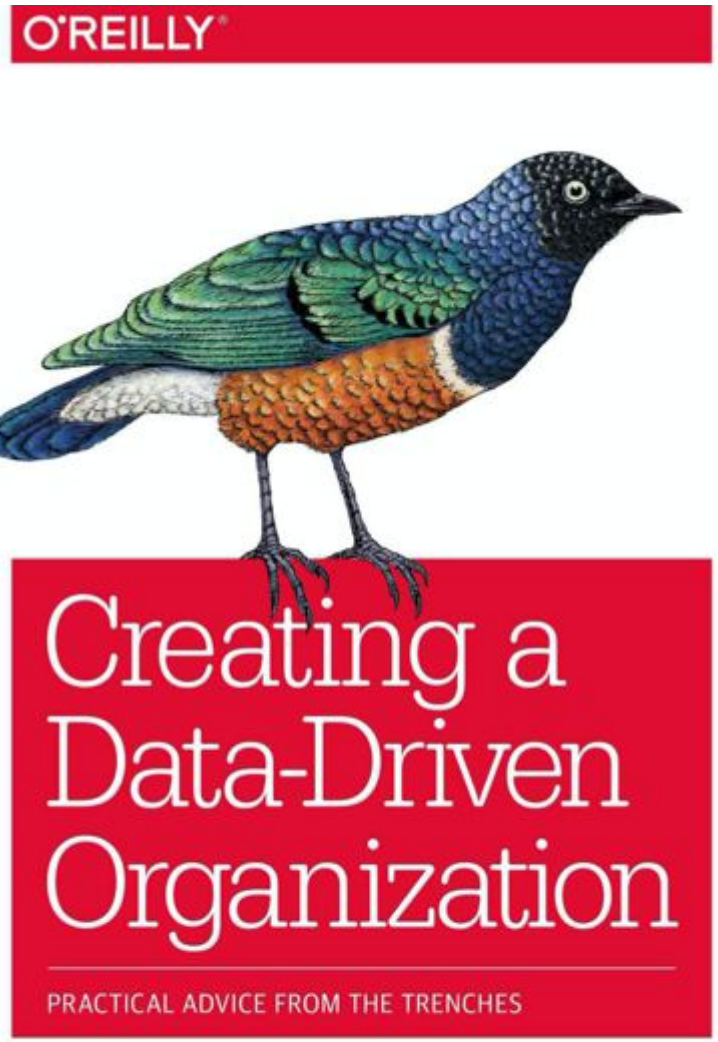


How it works

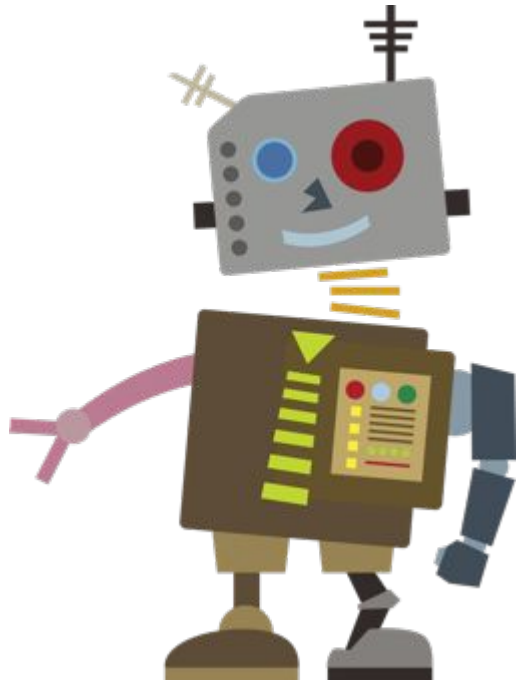


https://github.com/Mjrovai/covid19Xray

Conclusion



http://shop.oreilly.com/product/0636920035848.do



MJRoBot.org

github.com/Mjrovai

hackster.io/mjrobot

medium.com/@rovai

instructables.com/member/mjrovai

Thanks