Lesson #03
Data Science Fundamentals

Outline

1. Data Science Fundamentals
2. Ethic by Design
3. Defining a problem
4. Platforms & Tools
5. Warming up

Calendar

Unit 01

Início	Fim	Descrição
18/02/2020	18/02/2020	Course Outline
20/02/2020	20/02/2020	IMD Talk
27/02/2020	27/02/2020	Fundamentals of Data Science
03/03/2020	03/03/2020	Introduction to Pandas
05/03/2020	05/03/2020	Exploring data with Pandas
10/03/2020	10/03/2020	Data Cleaning Basics
12/03/2020	12/03/2020	<<Project 01>>
17/03/2020	17/03/2020	Data Aggregation
19/03/2020	19/03/2020	Combinning Data with Pandas
24/03/2020	24/03/2020	Transforming Data with Pandas
26/03/2020	26/03/2020	Working with String in Pandas
31/03/2020	31/03/2020	Regular Expression
02/04/2020	02/04/2020	Working with missing and duplicate data
07/04/2020	07/04/2020	<< Project 02>> <<end of unit 01>>

Syllabus:

1. Pandas Fundamentals
2. Data Cleaning & Analysis

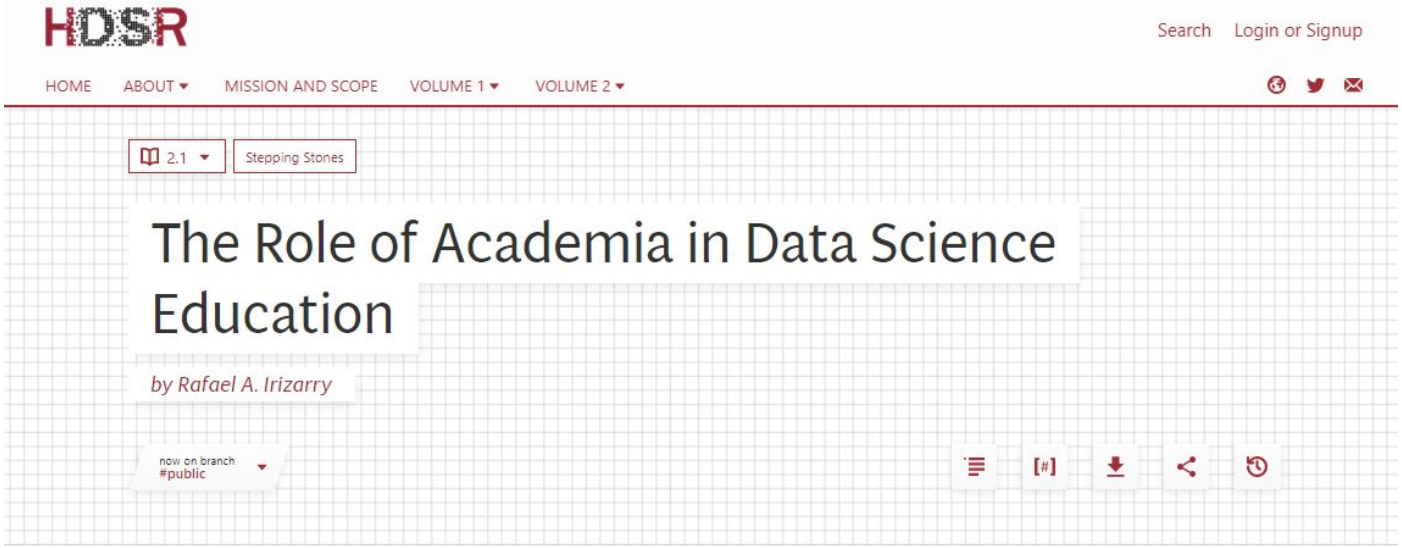
Grades

Notebooks - 10%
Project 01 - 30%
Project 02 - 40%
Datacamp (20k XP) - 20%

http://bit.do/imd_datascience

Suggestions

>> Data Science for Business >> Introduction to Python >> Intermediate Python

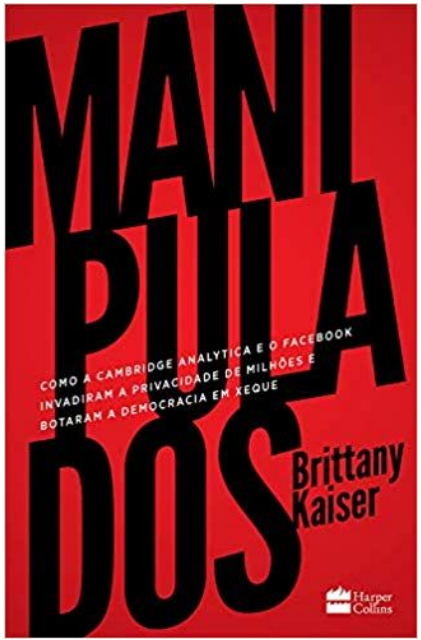


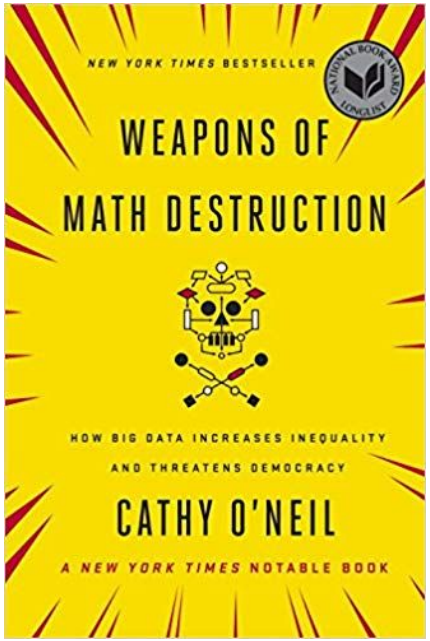
ABSTRACT

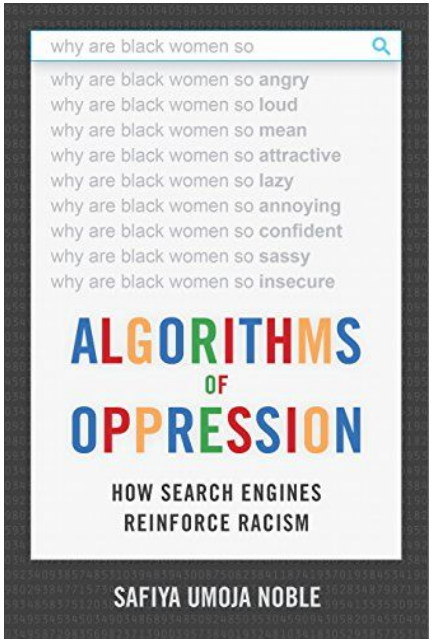
As the demand for data scientists continues to grow, universities are trying to figure out how to best contribute to the training of a workforce. However, there does not appear to be a consensus on the fundamental principles, expertise, skills, or knowledge-base needed to define an academic discipline. We argue that data science is not a discipline but rather an umbrella term used to describe a complex process involving not one data scientist possessing all the necessary expertise, but a team of data scientists with nonoverlapping complementary skills. We provide some recommendations for how to take this into account when designing data science academic programs.

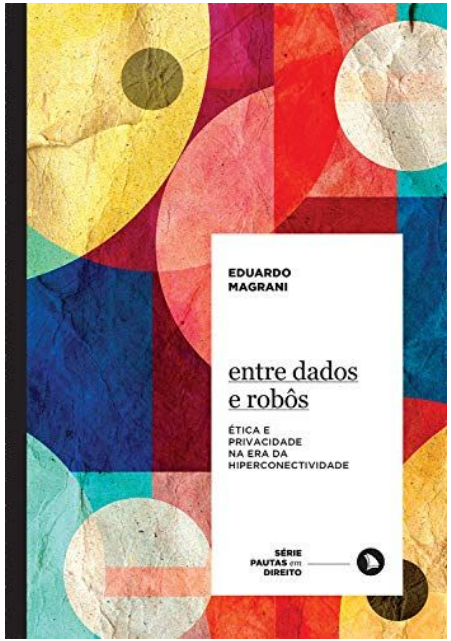
http://bit.do/hdsrcv

Gisliany Alves et al. Uma abordagem orientada a dados para a criação de um indicador de habitabilidade baseado na API da UBER


MANIPULADOS
COMO A CAMBRIDGE ANALYTICA E O FACEBOOK INVADIRAM A PRIVACIDADE DE MILHÕES E BOTARAM A DEMOCRACIA EM XEQUE
Brittany Kaiser
Harper Collins


NEW YORK TIMES BESTSELLER
WEAPONS OF MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY
CATHY O'NEIL
A NEW YORK TIMES NOTABLE BOOK


why are black women so
why are black women so angry
why are black women so loud
why are black women so mean
why are black women so attractive
why are black women so lazy
why are black women so annoying
why are black women so confident
why are black women so sassy
why are black women so insecure
ALGORITHMS OF OPPRESSION
HOW SEARCH ENGINES REINFORCE RACISM
SAFIYA UMOJA NOBLE


EDUARDO MAGRANI
entre dados e robôs
ÉTICA E PRIVACIDADE NA ERA DA HIPERCONECTIVIDADE
SÉRIE PAUTAS em DIREITO

Alexa has been eavesdropping on you this whole time

http://bit.do/alexanoprivacy

Activate This ‘Bracelet of Silence,’ and Alexa Can’t Eavesdrop

http://bit.do/braceletblock

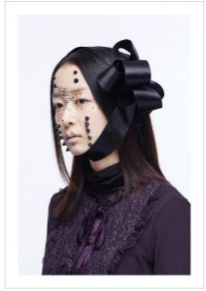
https://cvdazzle.com/

CVDAZZLE

Feb 22, 2020

New software for testing CV Dazzle looks scheduled for release at HeK work in April. With thanks to a privacy grant from NL Net

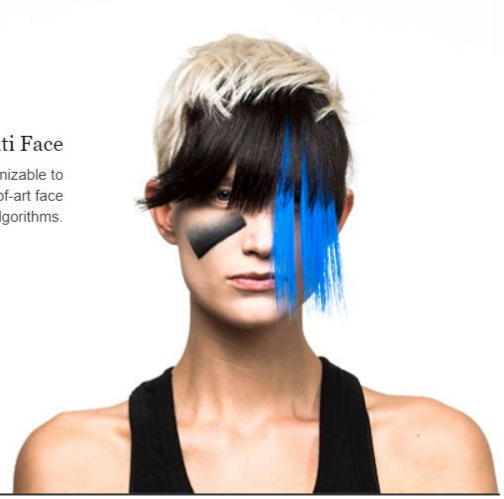
NB: this site will be relaunched in 2020. Much of the content here was developed for the haarcascade face detection algorithm, which was widely used between 2010-2015 but has now been deprecated as neural networks face detection algorithms have become more widespread.



by Coreana Museum of Art + Soobin Academy + G-square Model Academy

Anti Face

This face is unrecognizable to several state-of-art face detection algorithms.



Camouflage from face detection.

CV Dazzle explores how fashion can be used as camouflage from face-detection technology, the first step in automated face recognition.

The name is derived from a type of World War I naval camouflage called Dazzle, which used cubist-inspired designs to break apart the visual continuity of a battleship and conceal its orientation and size. Likewise, CV

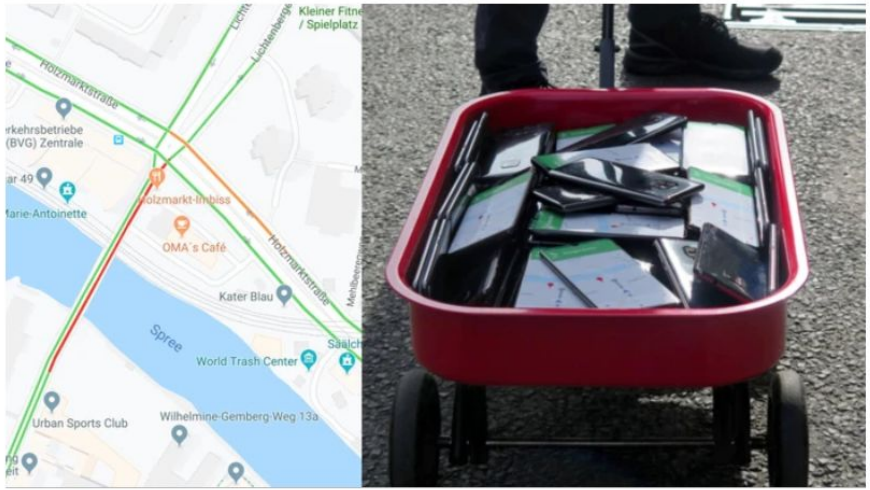
From all appearances, deception has always been critical to daily survival—for human and non-human creatures alike—and, judging by its current ubiquity, there is no end in immediate sight

This Man Created Traffic Jams on Google Maps Using a Red Wagon Full of Phones

By pulling 99 phones down empty streets, artist Simon Weckert made it look like they were gridlocked on Google Maps.

By **Matthew Gault**

Feb 3 2020, 12:41pm Share Tweet Snap



http://bit.do/virtualtraffic

silvio meira @srlm · Feb 25

rara oportunidade de entender como SISTEMAS afetam pessoas: estudo de #ALGORITMO de RETENÇÃO de SEGURADOS mostra que a essência é 1. cobrar MENOS de quem tem mais chance de mudar de seguradora e 2. MAIS de quem é fiel: bit.ly/39XSkXo. em que serviços isso rola no BRASIL?

DATA VIOLENCE

and how bad engineering choices can damage society

Research | Computer Vision

Does object recognition work for everyone? A new method to assess bias in CV systems

June 07, 2019 Written by Terrance DeVries, Ishan Misra, Changhan Wang, Laurens van der Maaten

https://ai.facebook.com/blog/new-way-to-assess-ai-bias-in-object-recognition-systems/

DOLLAR STREET

In the news people in other cultures seem stranger than they are.
We visited 264 families in 50 countries and collected 30,000 photos.
We sorted the homes by income, from left to right.

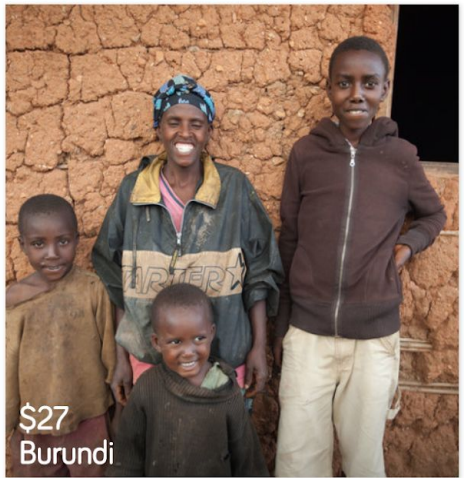
See how people *really* live

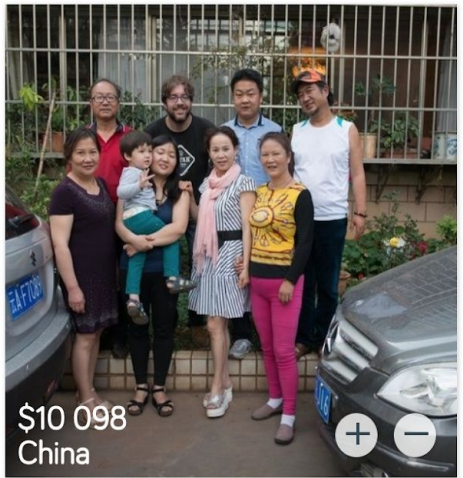
Quick tour | Maybe later

Families in the World by income

English

POOREST — RICHEST





Soap

Country of Origin: Nepal
Prediction: Food

Spices

Country of Origin: Philippines
Prediction: Beer

Toothpaste

Country of Origin: Burundi
Prediction: Wood

Country of Origin: UK
Prediction: Toiletry

Country of Origin: USA
Prediction: Spice

Country of Origin: USA
Prediction: Toothpaste

Yann LeCun @ylecun · 7h

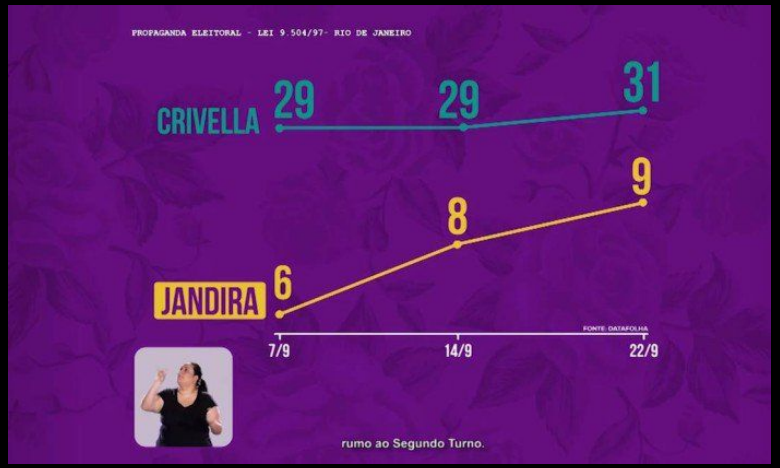
People are biased.
Data is biased, in part because people are biased.
Algorithms trained on biased data are biased.
But learning algorithms themselves are not biased.
Bias in data can be fixed.
Bias in people is harder to fix.

Biased Algorithms Are Easier to Fix Than Biased People

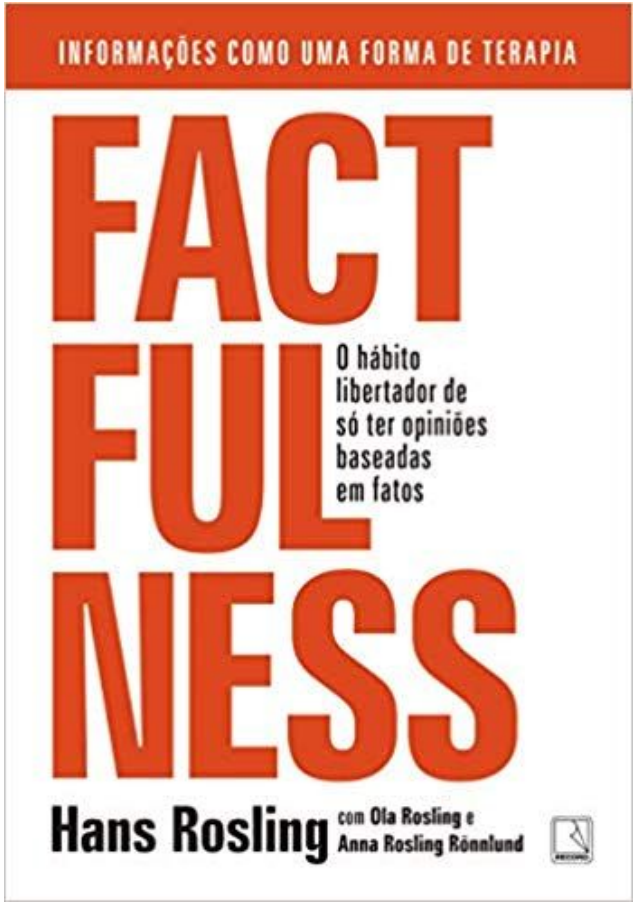
Racial discrimination by algorithms or by people is harmful — but that's where the similarities end.

nytimes.com

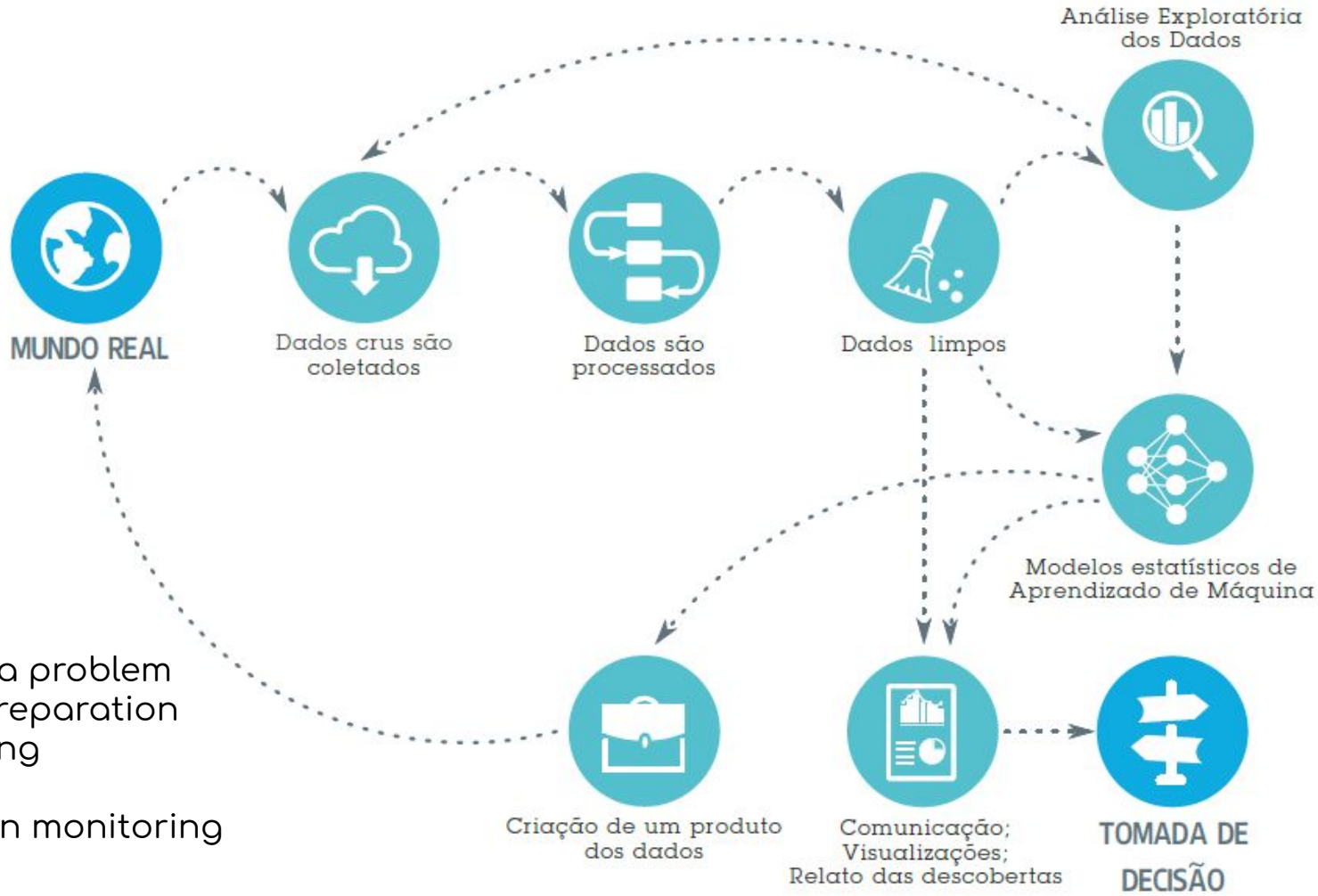
Bias is everywhere!!!!!



https://marcusnunes.me/posts/algumas-das-piores-visualizacoes-de-dados-ja-feitas/



A book that destroys myths by presenting facts and statistics in a clear and fun way. Factfulness is an urgent and essential book that will change the way you see the world and empower you to respond to the crises and opportunities of the future.



1. Define a problem
2. Data preparation
3. Modeling
4. Deploy
5. Solution monitoring

Gisliany Alves et al. Uma abordagem orientada a dados para a criação de um indicador de habitabilidade baseado na API da UBER

1. Data preparation
2. Modeling
3. Deploy
4. Solution monitoring
5. Define a problem

“take this data and see what you can find”

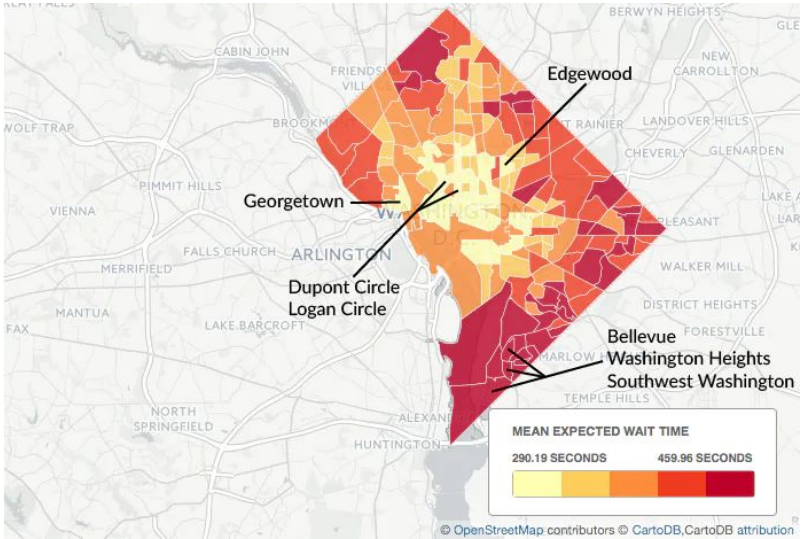
How to define a problem?

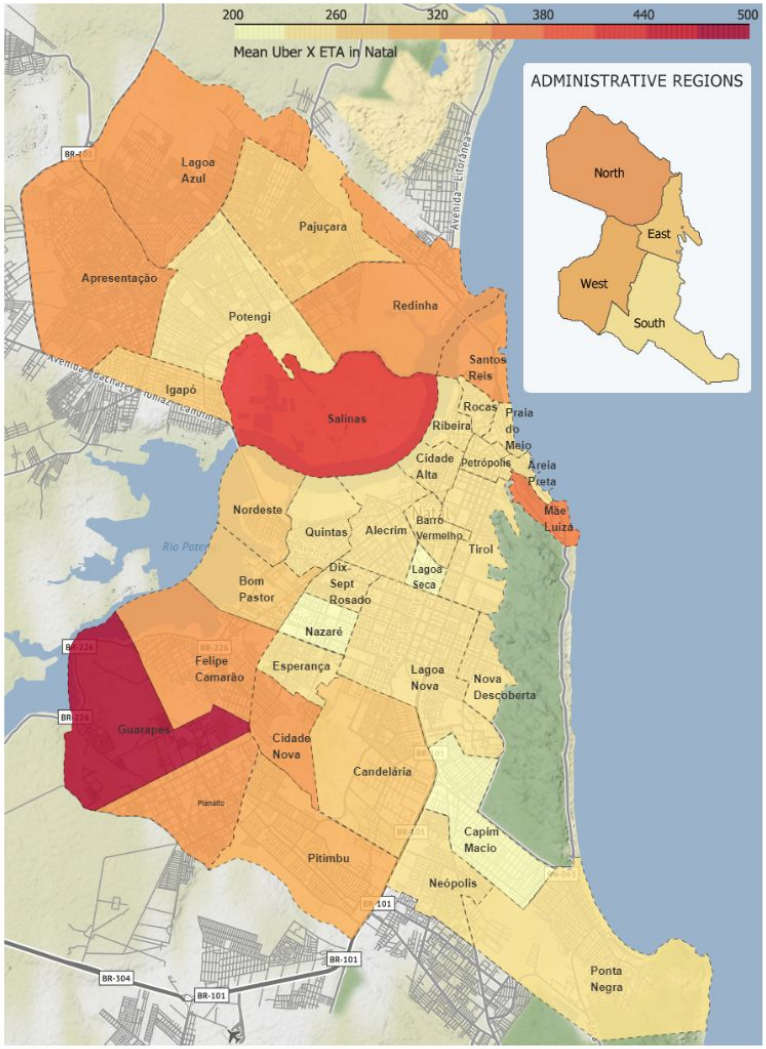
1. What problem do we want to solve?
2. Choose the simplest problem (gain know-how)
3. What business process can we affect?
    a. What will be the deliverable?
4. How will we know if the solution is working?
    a. Metrics: primary and secondary
5. Project duration
    a. Agile methods
    b. We will only really know how it will impact the business by putting it into production

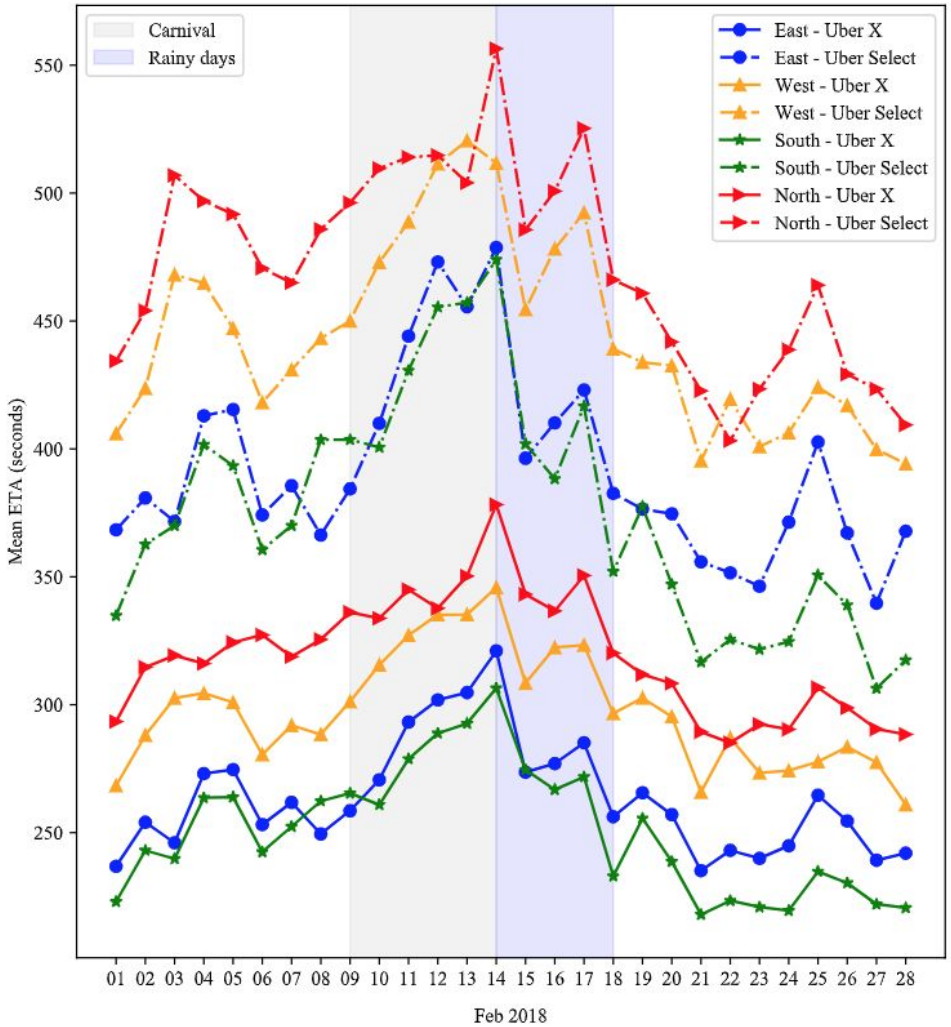
The Washington Post

Economic Policy

Uber seems to offer better service in areas with more white people. That raises some tough questions.




200
260
320
380
440
500
Mean Uber X ETA in Natal
ADMINISTRATIVE REGIONS
North
East
West
South
Lagoa Azul
Pajuçara
Apresentação
Potengi
Redinha
Igapó
Salinas
Santos Reis
Rocas
Ribeira
Praia do Meio
Cidade Alta
Petrópolis
Areia Preta
Mãe Luiza
Nordeste
Quintas
Alecrim
Barro Vermelho
Tirol
Lagoa Seca
Dix-Sept Rosado
Bom Pastor
Nazaré
Esperança
Felipe Camarão
Lagoa Nova
Nova Descoberta
Guarapes
Cidade Nova
Candelária
Capim Macio
Pitimbu
Neópolis
Ponta Negra
Rio Potengi
BR-304
BR-101
BR-226


Carnival
Rainy days
East - Uber X
East - Uber Select
West - Uber X
West - Uber Select
South - Uber X
South - Uber Select
North - Uber X
North - Uber Select
Mean ETA (seconds)
550
500
450
400
350
300
250
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
Feb 2018


C-MRiC.ORG
LONDON, UNITED KINGDOM
Pioneer Research and Innovation
Promote Collaborative Inter- workings
Encourage Interplay of Different Culture
Organise Conferences and Seminars
A free Membership Organisation
A Preliminary Exploration of Uber Data as an Indicator of Urban Liveability
Agunasta Bezerra, Gislany Alves, Ivanovitch Silva, Pierangelo Rosati, Patricia Takako Endo, Theo Lynn
Cyber Science
Cyber Situational Awareness
Social Media
Cyber Security
Cyber Incident Response
www.c-mric.org/cs
Connect, Collaborate & Communicate ...
C-MRiC.ORG
LONDON, UNITED KINGDOM

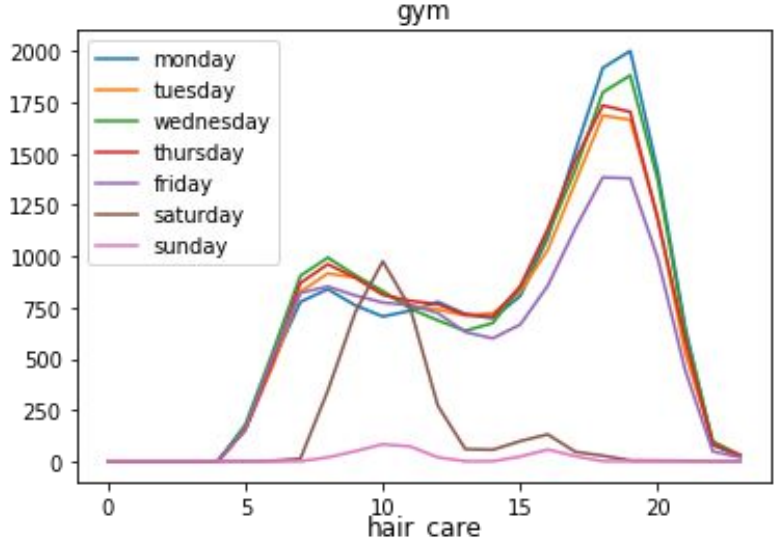
Medium

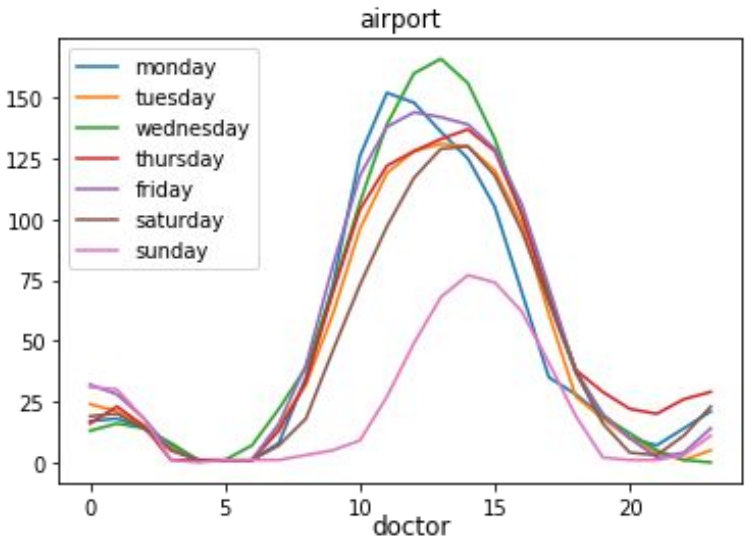
Welcome to Medium, where words matter.

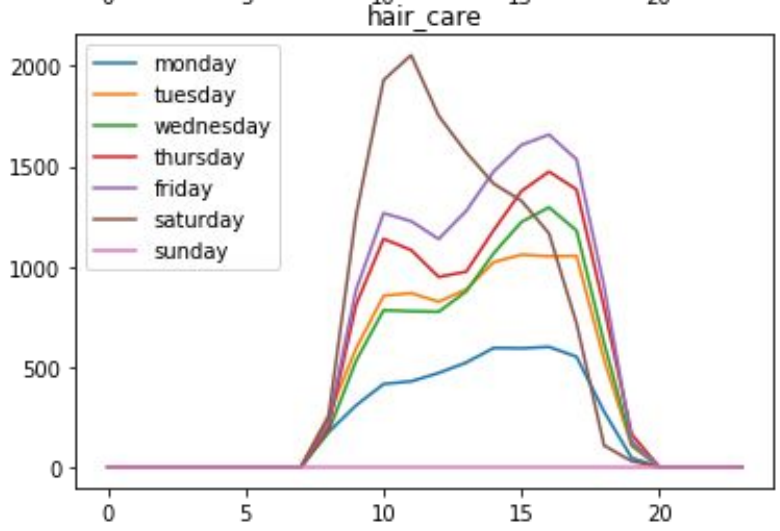
Mapeando locais mais visitados em Natal

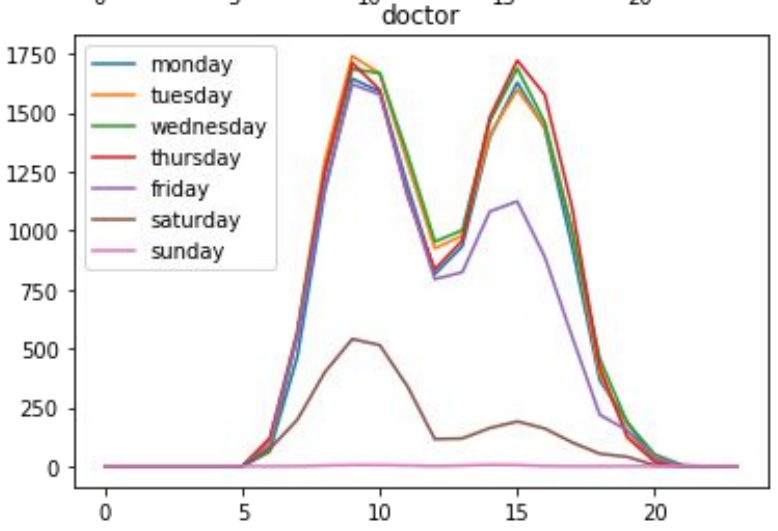
Usando Python, Pandas e Folium

Maradona Morais Follow
May 17 · 7 min read


gym
monday
tuesday
wednesday
thursday
friday
saturday
sunday
2000
1750
1500
1250
1000
750
500
250
0
0
5
10
15
20


airport
monday
tuesday
wednesday
thursday
friday
saturday
sunday
150
125
100
75
50
25
0
0
5
10
15
20


hair_care
monday
tuesday
wednesday
thursday
friday
saturday
sunday
2000
1500
1000
500
0
0
5
10
15
20


doctor
monday
tuesday
wednesday
thursday
friday
saturday
sunday
1750
1500
1250
1000
750
500
250
0
0
5
10
15
20

O negócio milionário dos incêndios na Amazônia

Força-tarefa do Ministério Público Federal calcula que queimada de 1.000 ha custa cerca de um milhão no mercado negro da região. Raquel Dodge vê ação orquestrada em fogo. Procuradores não comentam alertas ignorados por Força Nacional

EL PAÍS

FOLHA DE S.PAULO

Incêndios na Amazônia estão concentrados em propriedades privadas

Áreas privadas cobrem 18% do bioma e concentram 33% dos focos de fogo

Agosto tem recorde de focos de incêndio na Amazônia em nove anos, aponta Inpe

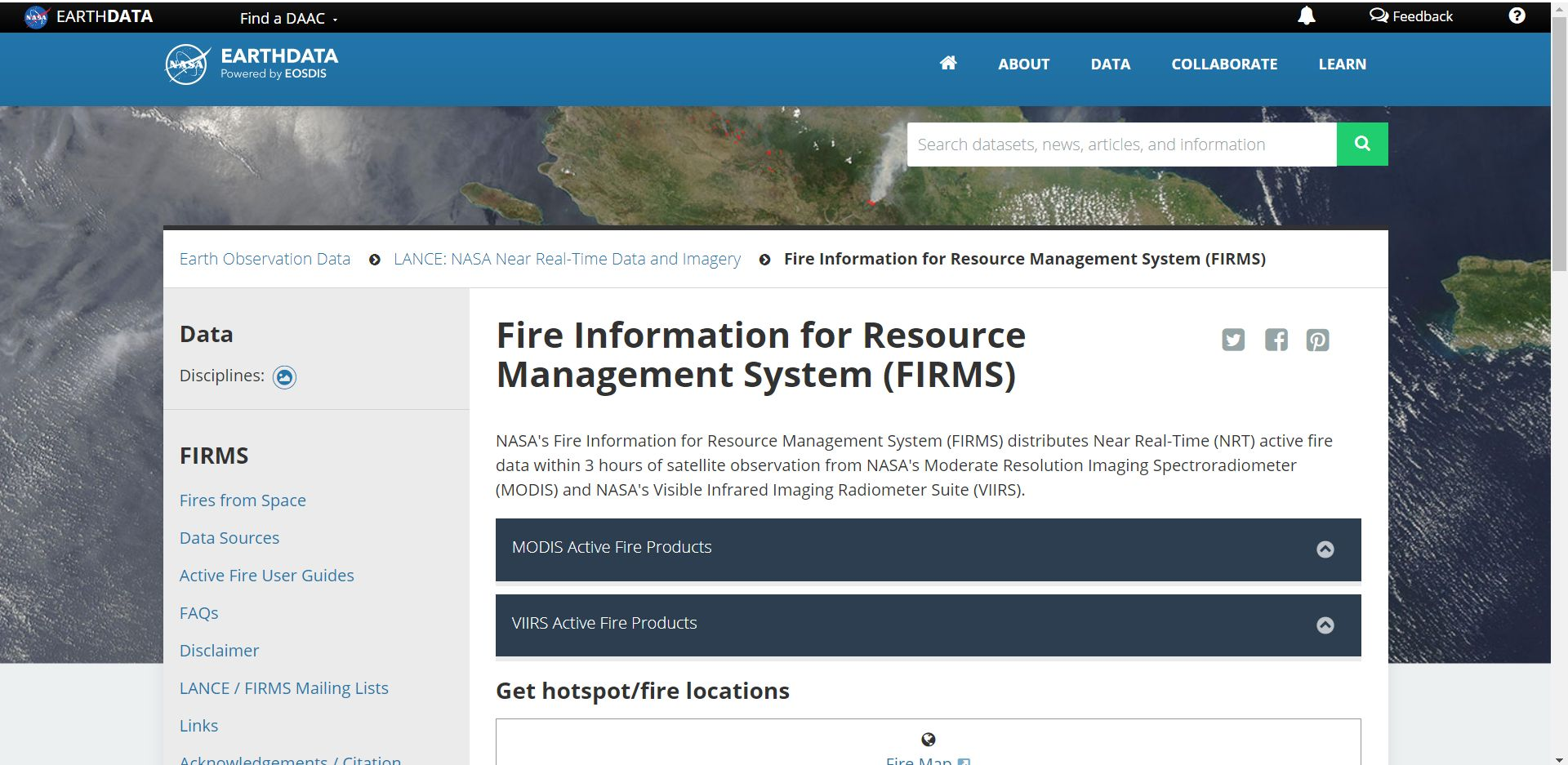
Dados do Programa Queimadas mostram 30.901 focos, quase o triplo do ano passado

O GLOBO

AgênciaBrasil

Secretário-geral da ONU está preocupado com queimadas na Amazônia

O secretário-geral da Organização das Nações Unidas (ONU), António Guterres, afirmou hoje (22) por meio de sua conta de Twitter que está "profundamente preocupado" com os incêndios na Floresta Amazônica.



Data

Disciplines:

FIRMS

- Fires from Space
- Data Sources
- Active Fire User Guides
- FAQs
- Disclaimer
- LANCE / FIRMS Mailing Lists
- Links
- Acknowledgements / Citation

Fire Information for Resource Management System (FIRMS)

NASA's Fire Information for Resource Management System (FIRMS) distributes Near Real-Time (NRT) active fire data within 3 hours of satellite observation from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) and NASA's Visible Infrared Imaging Radiometer Suite (VIIRS).

MODIS Active Fire Products

VIIRS Active Fire Products

Get hotspot/fire locations

Fire Map

Programa

Queimadas

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

SISTEMAS DE MONITORAMENTO

BDQueimadas 1

- BDQueimadas
- Exemplos e Validações
- Download de dados

CIMAN Virtual 2

Centro Integrado Multiagências de Coord. Operacional e Federal em Brasília

- Briefings meteorológicos

TerraMA2Q 3

Acessar ao hotsite do TerraMA2Q.

Focos nas APs 4

Monitoramento de Focos nas Áreas de Proteção Ambiental, ontem

Área Queimada 5

- Resolução - 1km
- Resolução - 30m
- Geral

Risco de Fogo 6

Análise e Previsão de Risco de Fogo em mapas e Fogogramas, e Meteorologia

- Anuário de Risco de Fogo

SISAM 7

Aplicativo - Impacto das Queimadas / Incêndios e Meteorologia na Saúde

Receber por e-mail 8

Alertas de focos em Áreas Protegidas, Relatórios Diários, Boletim InfoQueima Mensal, mensagens operacionais, etc.

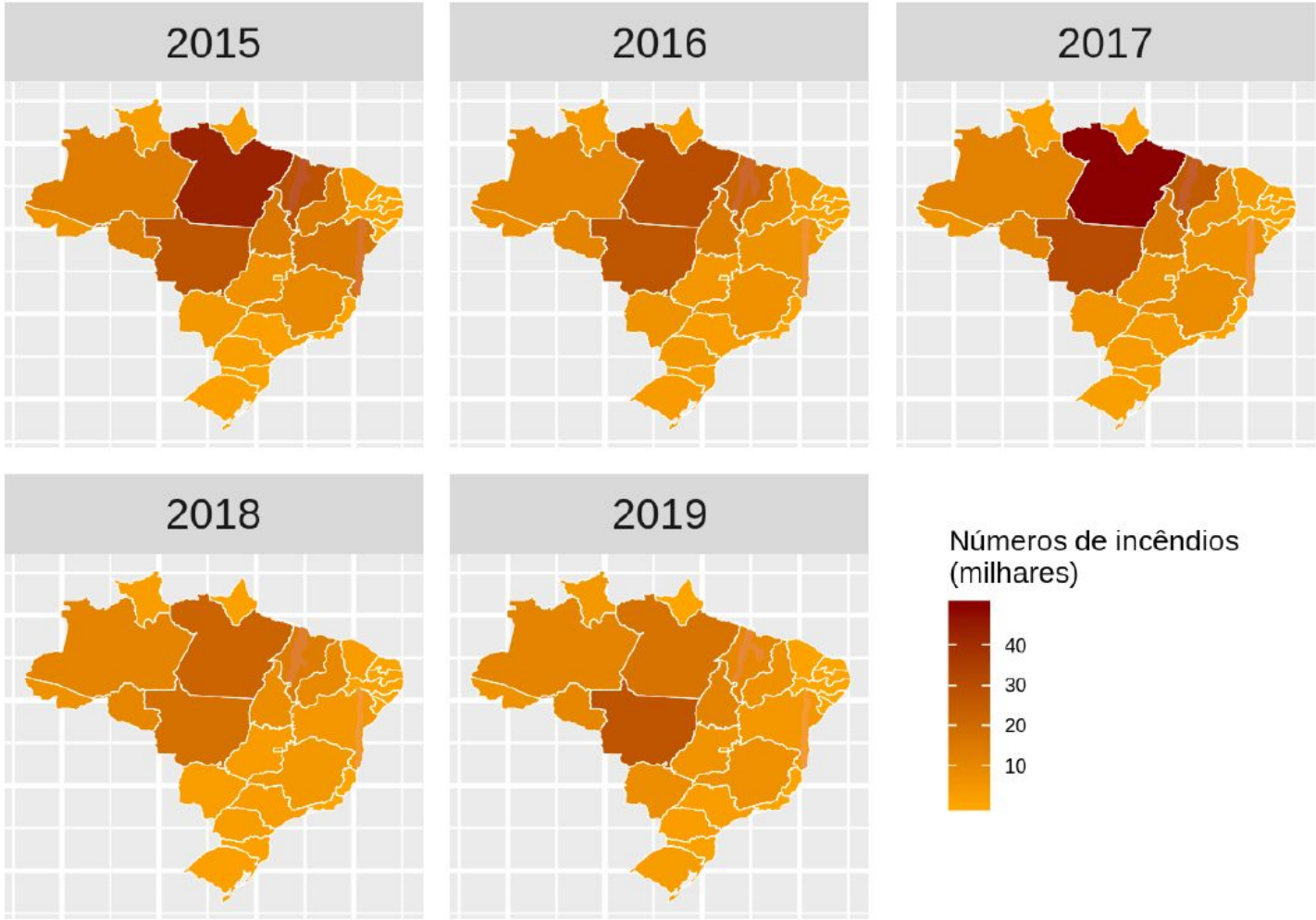
RELATÓRIOS E PUBLICAÇÕES

Informações atuais 9

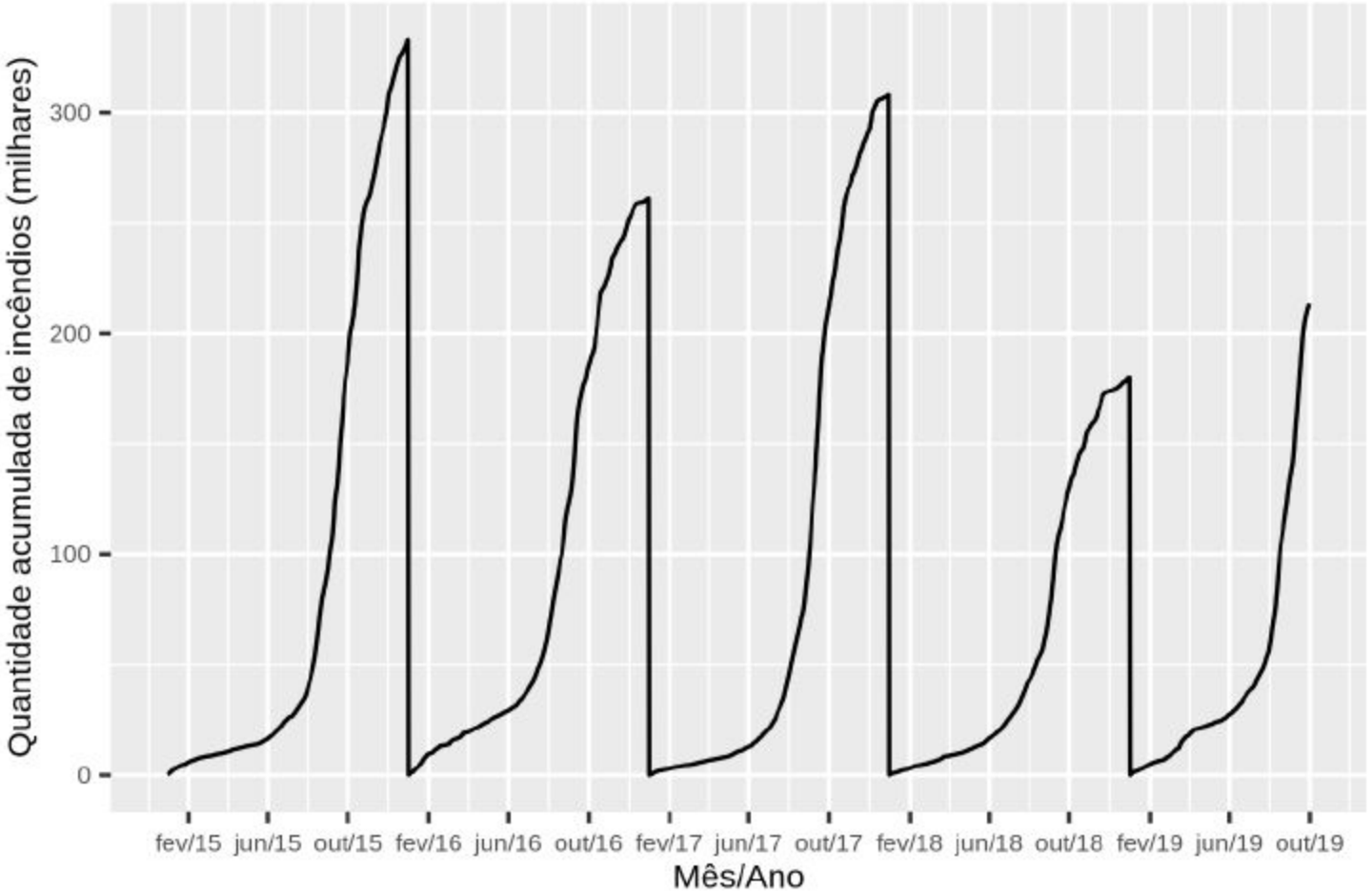
Boletins internos 10

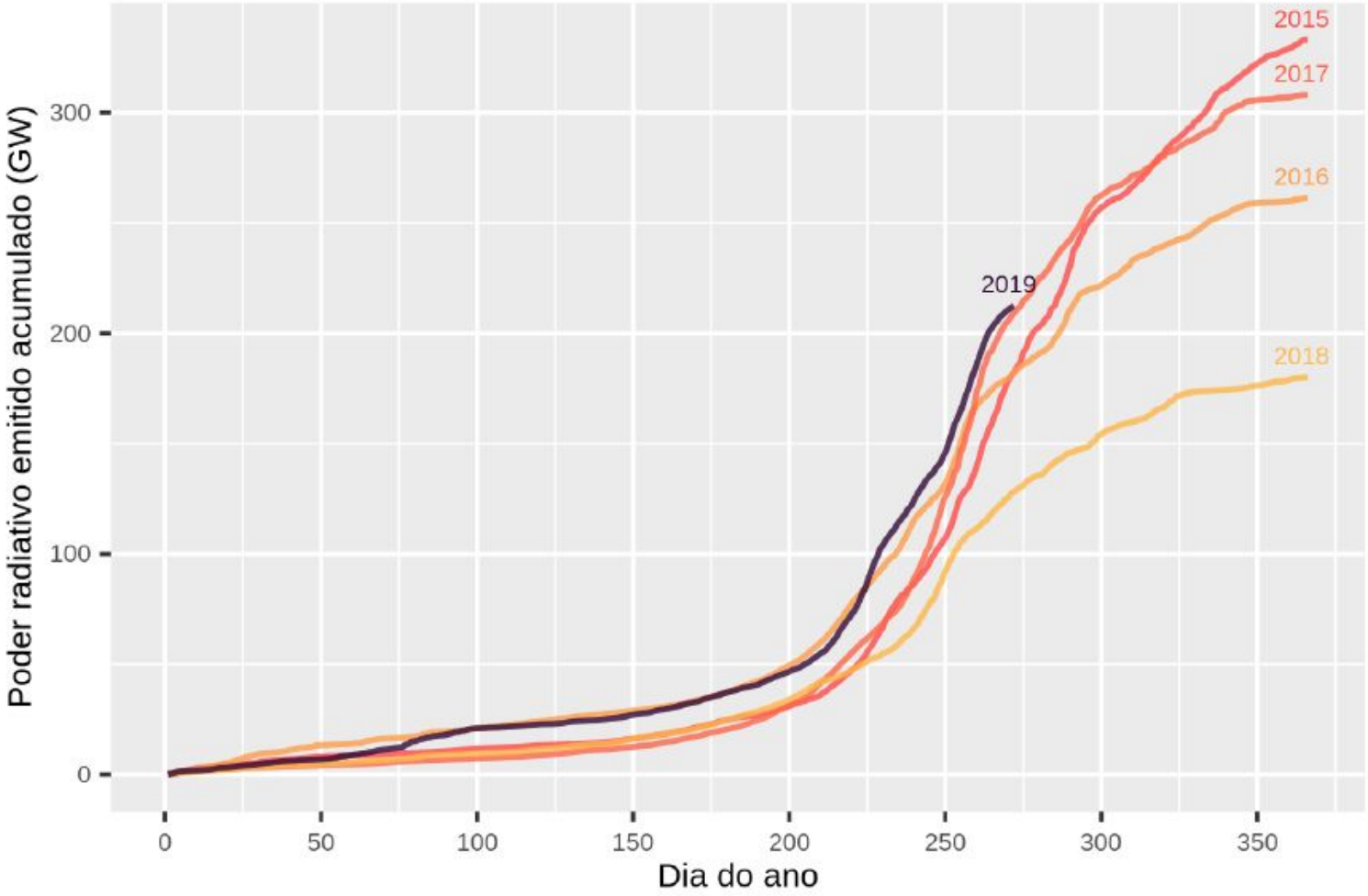
Resumo histórico e animações 11

Publicações e impacto 12



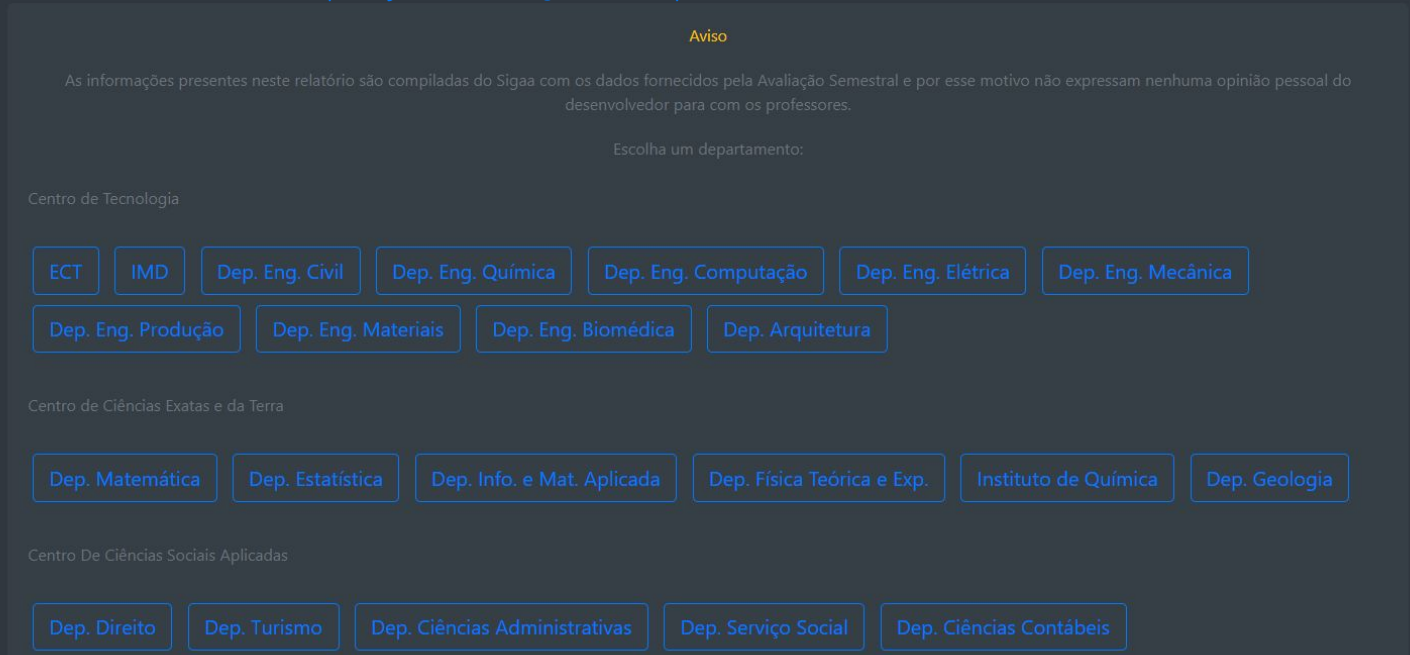
Alexandre Dias et al. Uma Análise Exploratória de Dados sobre incêndios florestais no Brasil


Quantidade acumulada de incêndios (milhares)
300
200
100
0
fev/15
jun/15
out/15
fev/16
jun/16
out/16
fev/17
jun/17
out/17
fev/18
jun/18
out/18
fev/19
jun/19
out/19
Mês/Ano


Poder radiativo emitido acumulado (GW)
300
200
100
0
0
50
100
150
200
250
300
350
Dia do ano
2015
2017
2016
2019
2018

Rank Professores

https://juniordnts.github.io/professor-rank/front/index.html



https://github.com/odufrn/odufrn-downloader

odufrn / **odufrn-downloader**

Code | Issues 12 | Pull requests 1 | Actions | Projects 2 | Wiki | Security | Insights

Pacote para baixar os dados do portal de dados abertos da UFRN https://odufrn.github.io/odufrn-downl...

ufrn | open-data | downloader | ckan | requests

237 commits | 5 branches | 0 packages | 3 releases | 6 contributors | MIT

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

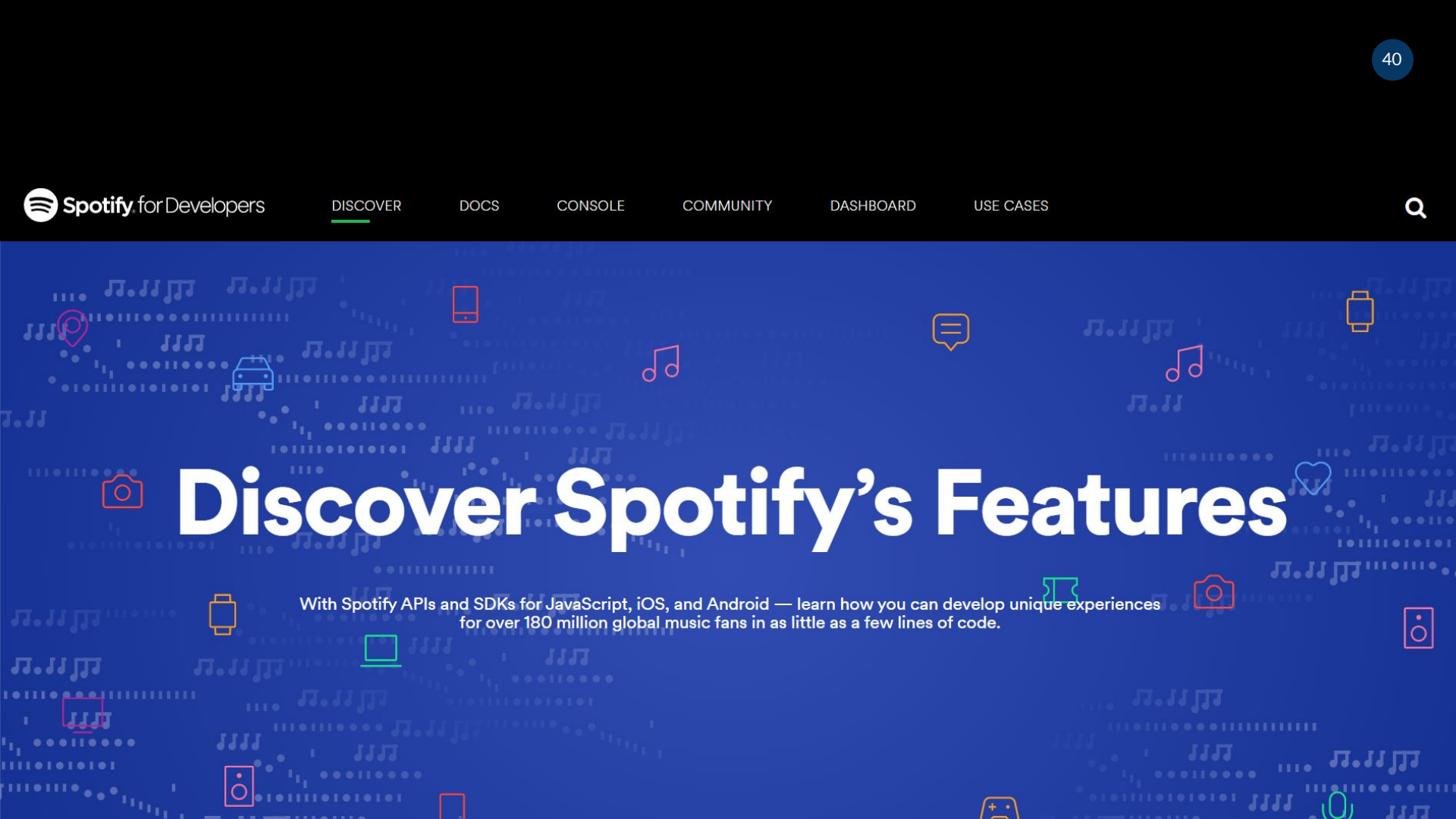
alvarofpp Merge pull request #96 from odufrn/issue-92 ... Latest commit b809d50 on 1 Nov 2019

docs	Ajustes na documentação	6 months ago
examples	Finalizando erros apontados pelo Travis	5 months ago
odufrn_downloader	Merge pull request #96 from odufrn/issue-92	4 months ago
tests	Merge pull request #96 from odufrn/issue-92	4 months ago
.gitignore	issue: #56 | Creando exception class to use in File.py and tests/test...	5 months ago
.travis.yml	Modifica travis e adiciona novos testes	6 months ago
CONTRIBUTING.md	Update CONTRIBUTING.md	6 months ago
LICENSE	0.0.1 Projeto copiado	7 months ago
README.md	Adicionando Notebook a pasta examples	5 months ago

androidauto

On your car display.

GET STARTED


Spotify for Developers
DISCOVER
DOCS
CONSOLE
COMMUNITY
DASHBOARD
USE CASES
Discover Spotify's Features
With Spotify APIs and SDKs for JavaScript, iOS, and Android — learn how you can develop unique experiences for over 180 million global music fans in as little as a few lines of code.

http://bit.do/coronavirusdash

Coronavirus COVID-19 Global Cases by Johns Hopkins CSSE

Descriptive vs Predictive vs Prescriptive

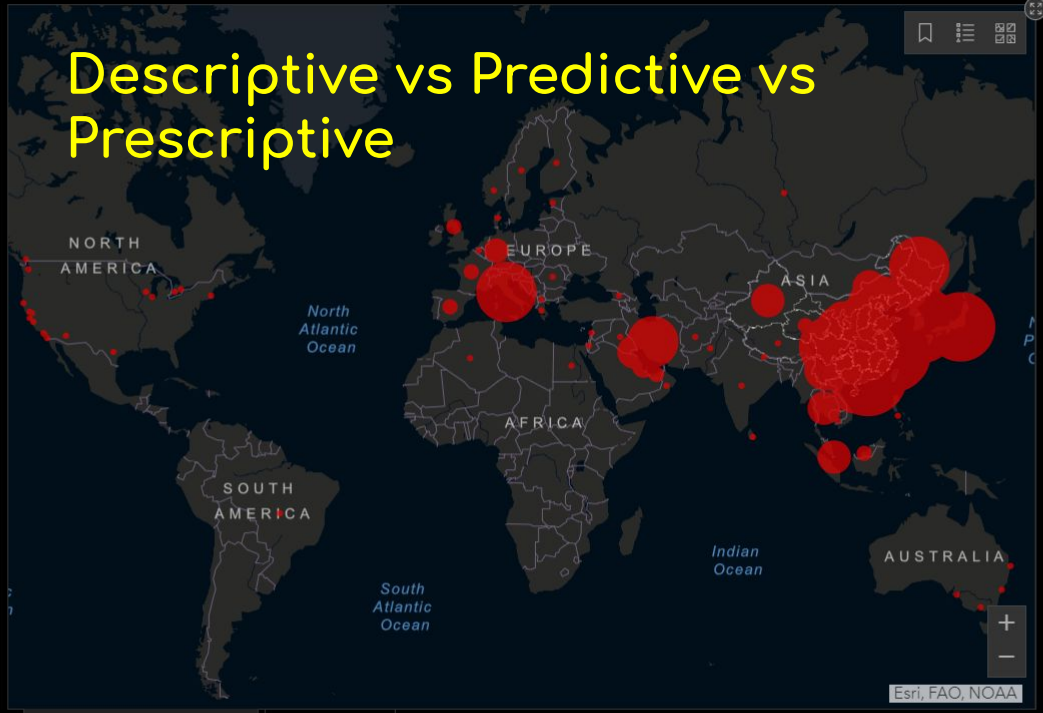
Total Confirmed
82,548

Confirmed Cases by Country/Region

- 78,497 Mainland China
- 1,766 South Korea
- 705 Others
- 528 Italy
- 245 Iran
- 189 Japan
- 93 Singapore
- 92 Hong Kong
- 60 US
- 43 Kuwait
- 40 Thailand
- 33 Bahrain
- 32 Taiwan
- 26 Germany

Country/Region | City, St/Prov

Last Updated at (M/D/YYYY)
2/27/2020, 11:13:06 AM



Cumulative Confirmed Cases | Existing Cases

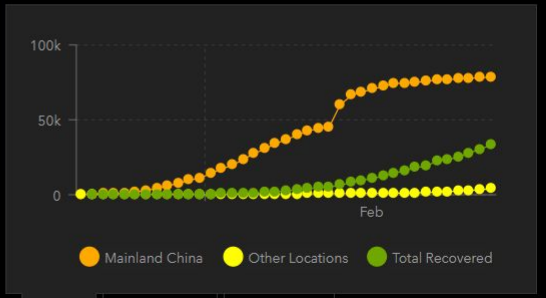
Lancet Article: Here. Mobile Version: Here. Visualization: JHU CSSE. Automation Support: Esri Living Atlas team and JHU APL.
Data sources: WHO, CDC, ECDC, NHC and DXY. Read more in this blog. Contact US.
Downloadable database: GitHub: Here. Feature layer: Here.
Point level: City level - US, Canada and Australia; Province level - China; Country level - other countries.

Total Deaths
2,810

- 2,641 deaths **Hubei** Mainland China
- 26 deaths Iran
- 20 deaths **Henan** Mainland China
- 14 deaths Italy
- 13 deaths **Heilongjiang** Mainland China
- 13 deaths South Korea
- 7 deaths **Guangdong** Mainland China
- 6 deaths

Total Recovered
33,243

- 23,383 recovered **Hubei** Mainland China
- 1,062 recovered **Henan** Mainland China
- 932 recovered **Zhejiang** Mainland China
- 890 recovered **Guangdong** Mainland China
- 804 recovered **Hunan** Mainland China
- 792 recovered **Anhui** Mainland China
- 754 recovered **Jiangxi** Mainland China



Actual | Logarithmic | Daily Cases

Hands on Data Science

NumPy

pandas

$y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$

Keras

TensorFlow

scikit learn

pgmpy

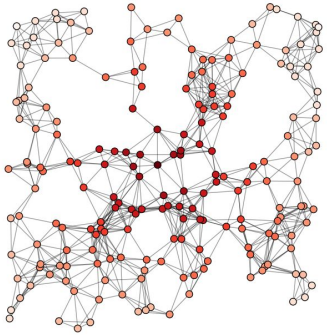
NetworkX

PyGSP

matplotlib

Folium

GeoJSON



Seaborn

plotly

Requests

bokeh

Leaflet

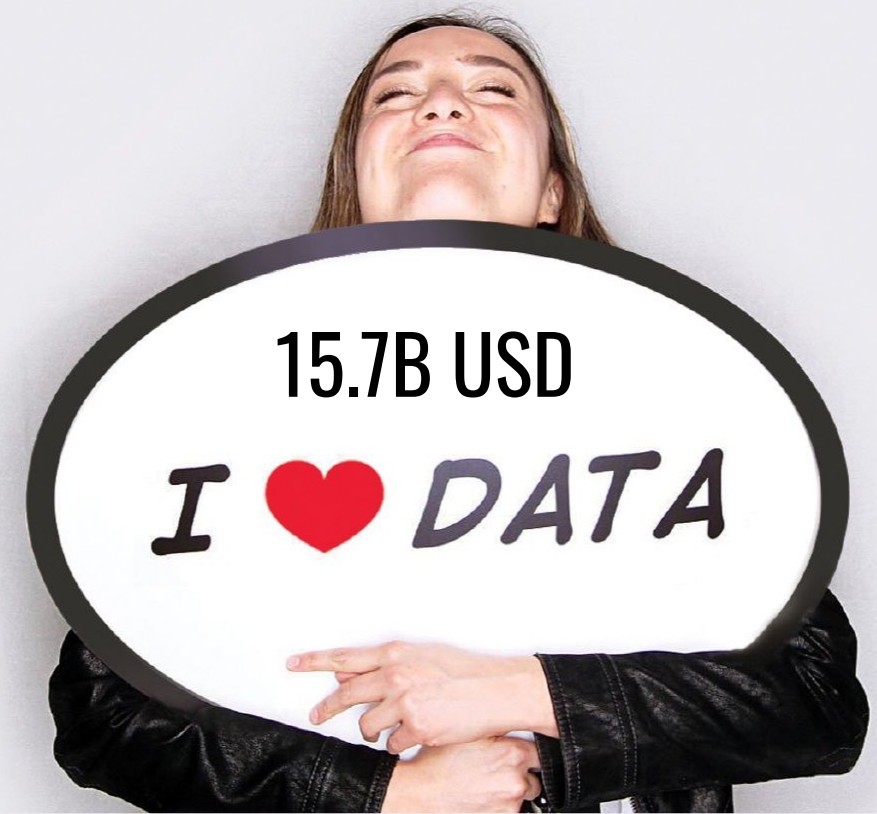
Beautiful Soap

Scrapy

http://www.pentaho.com/

https://www.tableau.com/

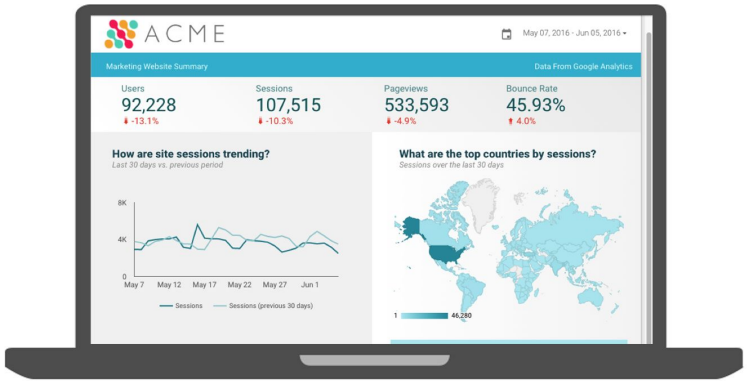
salesforce + tableau



http://bit.do/tableausell

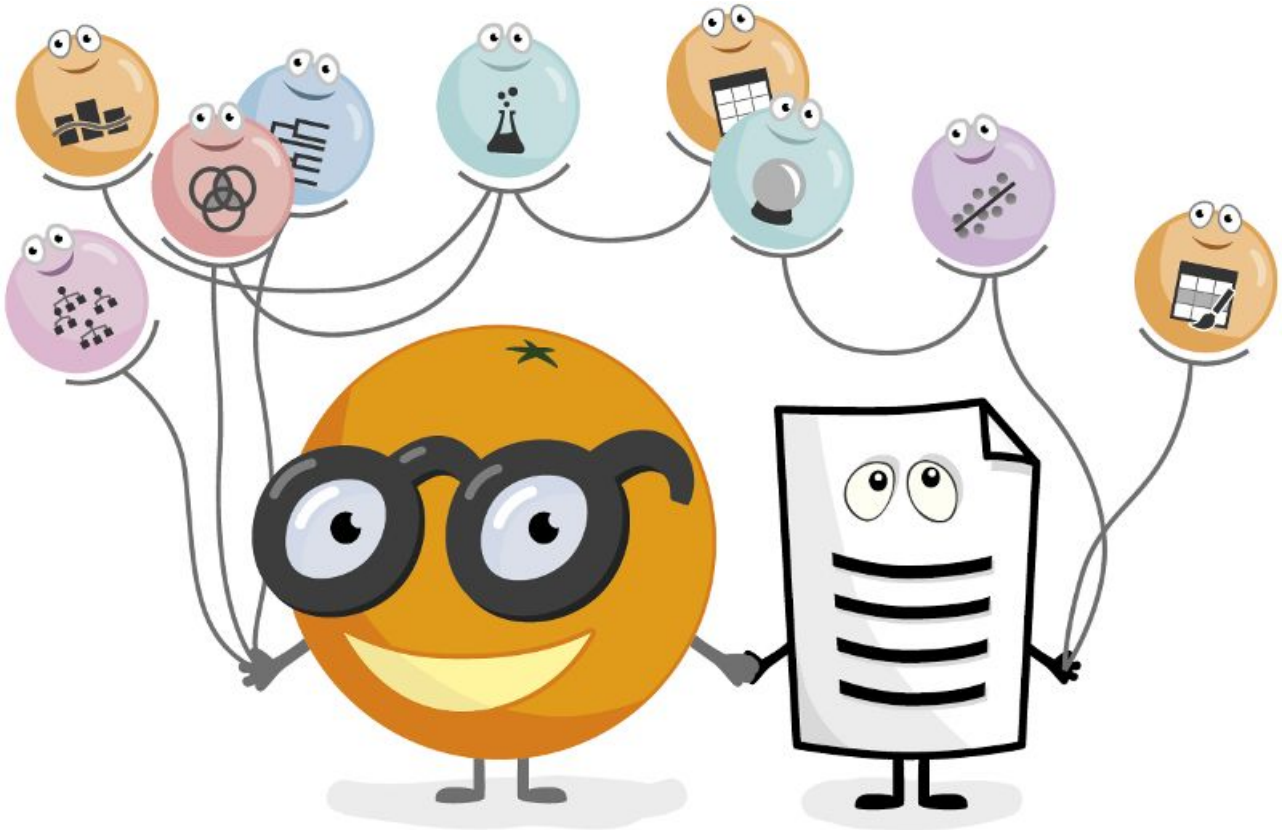
https://powerbi.microsoft.com

Google Data Studio



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

Data Mining Fruitful and Fun



https://orange.biolab.si/

Anaconda Distribution

The World's Most Popular Python/R Data Science Platform

Download

The open-source Anaconda Distribution is the easiest way to perform Python/R data science and machine learning on Linux, Windows, and Mac OS X. With over 19 million users worldwide, it is the industry standard for developing, testing, and training on a single machine, enabling *individual data scientists* to:

- Quickly download 7,500+ Python/R data science packages
- Manage libraries, dependencies, and environments with Conda
- Develop and train machine learning and deep learning models with scikit-learn, TensorFlow, and Theano
- Analyze data with scalability and performance with Dask, NumPy, pandas, and Numba
- Visualize results with Matplotlib, Bokeh, Datashader, and Holoviews

Windows | macOS | Linux

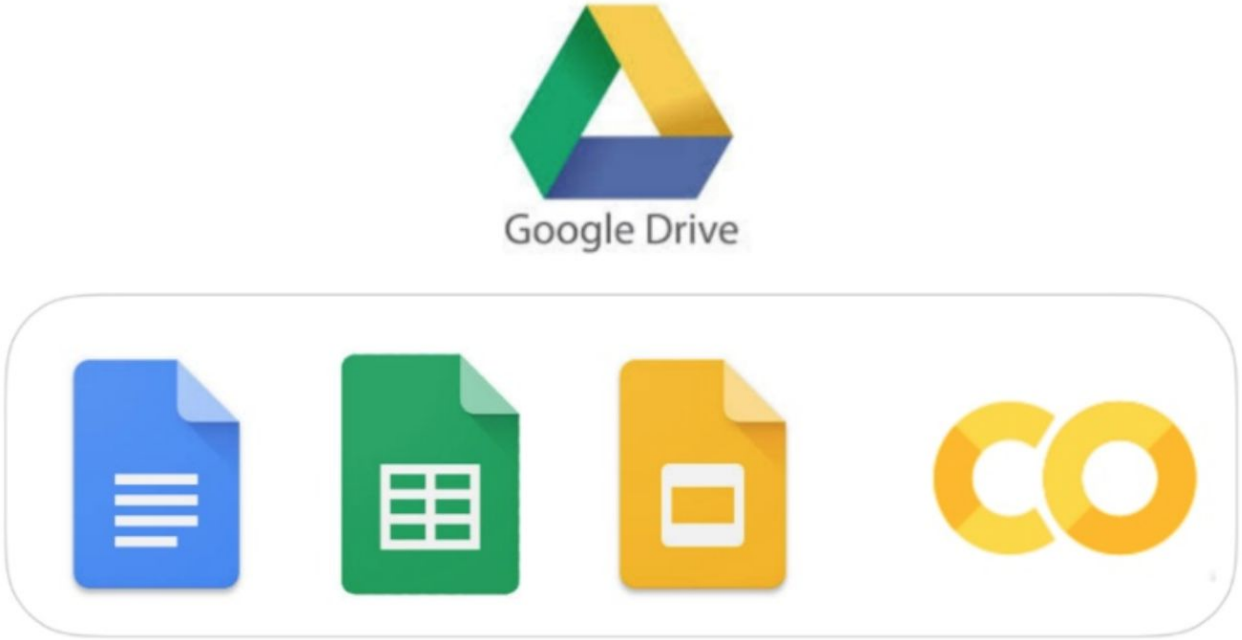

Jupyter is the new Excel

https://data.berkeley.edu/education/courses/data-100

Data 100: Principles and Techniques of Data Science

Google Colaboratory

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



Lists of Lists

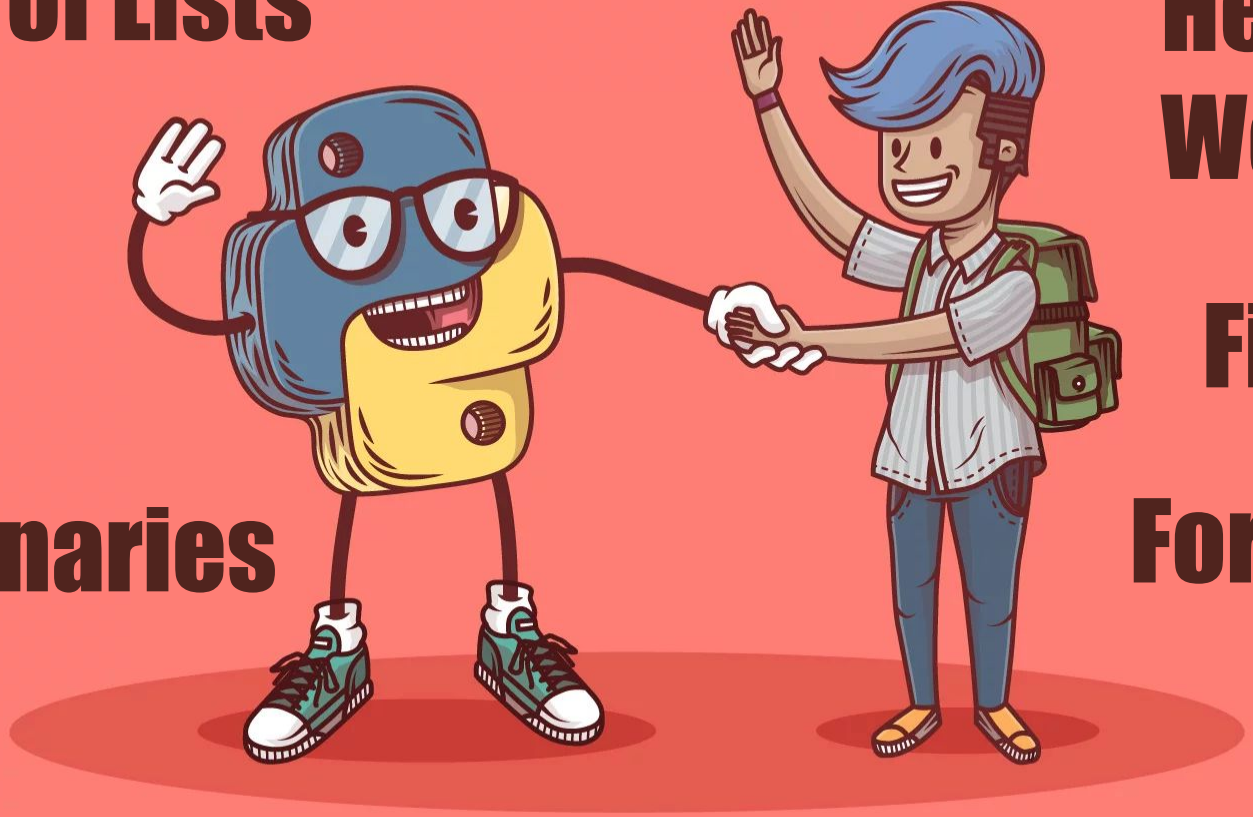
Hello World

Lists

Files

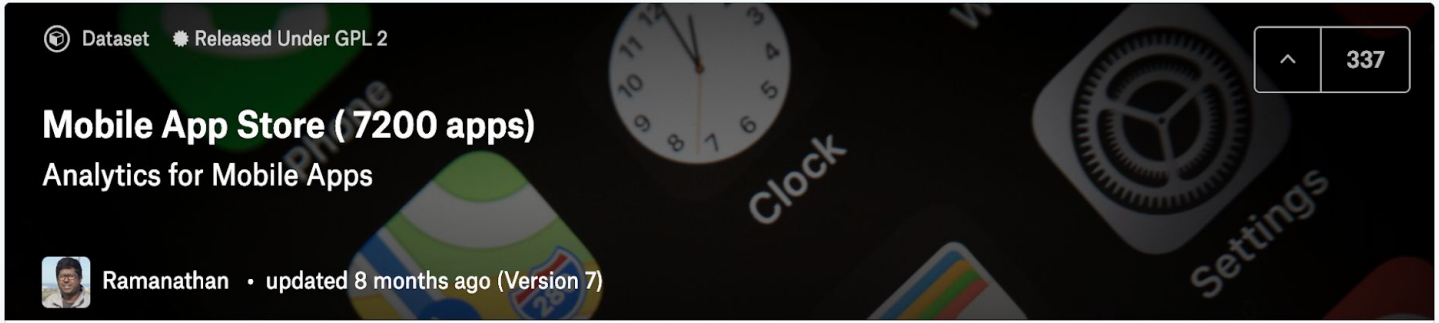
Dictionaries

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

Data Science Interview Questions

<<Optional Non-Graded>>

1. What do you think makes a good data scientist?
2. Give a few examples of “best practices” in data science.
3. Can you outline the steps in a data science project?
4. What data would you love to acquire if there were no limitations?
5. Tell me a compelling story about data that you have analyzed.