IMD0905 - Data Science I

Lesson #1 - Outline & Directions

Ivanovitch Silva

February, 2019


Introduction

Ivanovitch Silva (ivan@imd.ufrn.br)
3T1234

GitHub/ivanovitchm

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

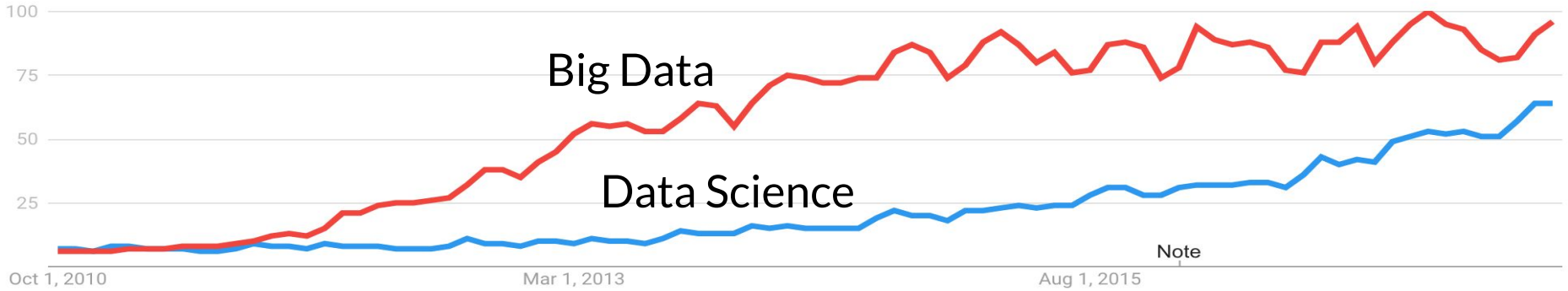
Big Data and Social Analytics certificate course

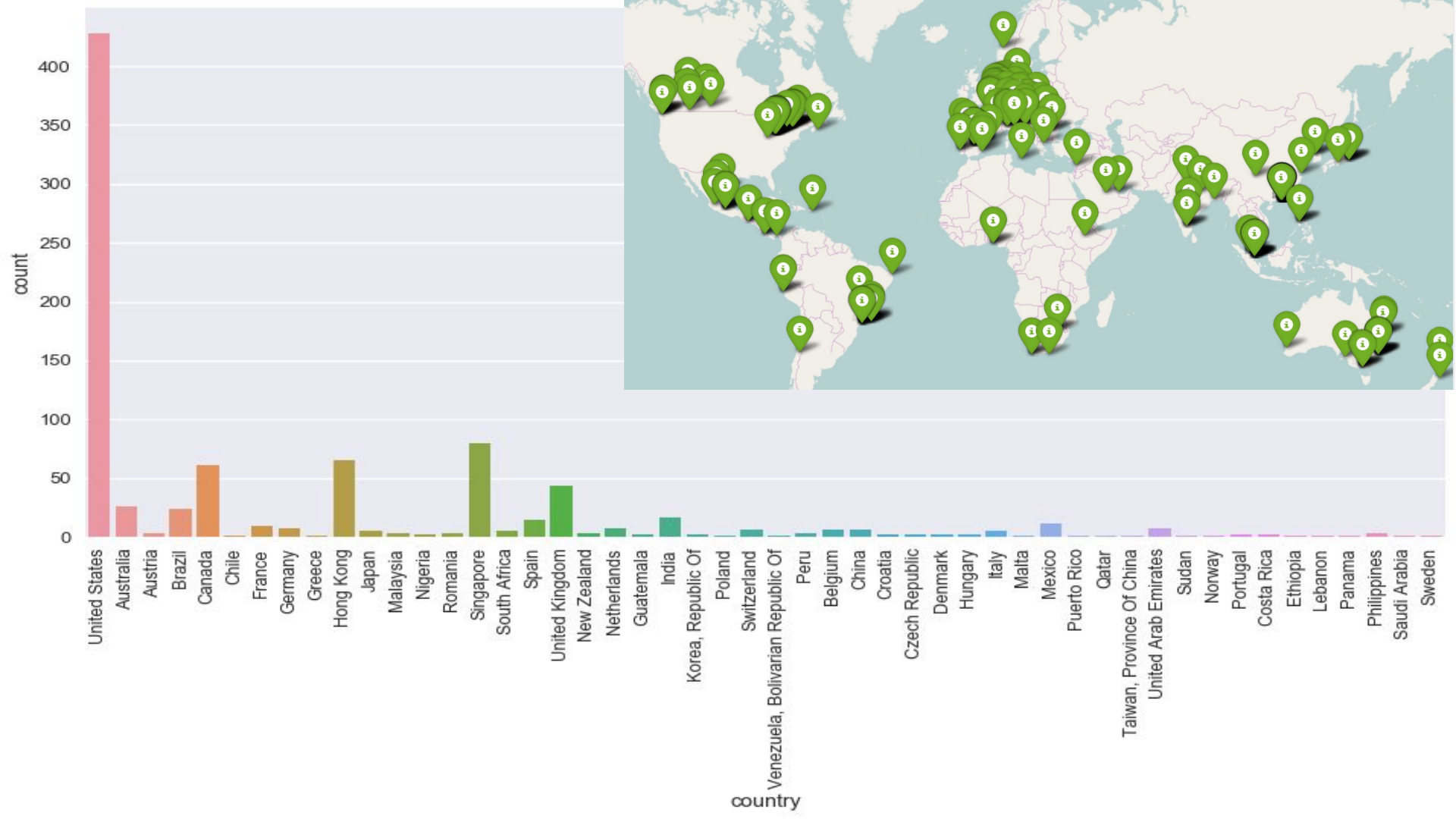
EXperimental Learning

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.




count
0
50
100
150
200
250
300
350
400
United States
Australia
Austria
Brazil
Canada
Chile
France
Germany
Greece
Hong Kong
Japan
Malaysia
Nigeria
Romania
Singapore
South Africa
Spain
United Kingdom
New Zealand
Netherlands
Guatemala
India
Korea, Republic Of
Poland
Switzerland
Venezuela, Bolivarian Republic Of
Peru
Belgium
China
Croatia
Czech Republic
Denmark
Hungary
Italy
Malta
Mexico
Puerto Rico
Qatar
Taiwan, Province Of China
United Arab Emirates
Sudan
Norway
Portugal
Costa Rica
Ethiopia
Lebanon
Panama
Philippines
Saudi Arabia
Sweden
country

Insanity is doing the same thing, over and over again, but expecting different results.

Albert Einstein

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
2018.2 - IMD0905 Data Science I

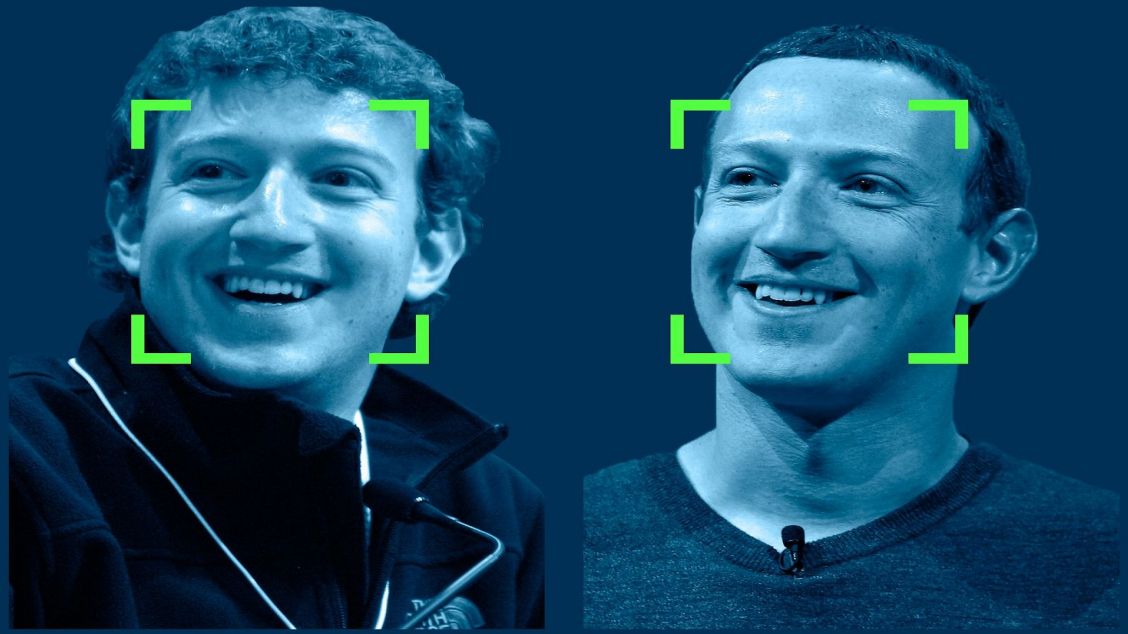
Graduate
2017.2 - EEC2006 Data Science Foundations
2017.2 - ITE0021 Learning Analytics
2018.2 - EEC1509 Machine Learning

2019 - PES


COURSE
Expectations

intriguing


#10yearchallenge



#10YearsChallenge


NOKIA
100%
2009
NOKIA
99%
2019


2009
2019
©2009 - ReD8ull

$0.00

2009

$3,363.36

2019


2009
2019
99
Destino
Partida
POP
99TAXI
Confirmar trajeto


2009
2019
Ligando:
Pizzaria
do Bigod...
ENTREGAR EM
Seu Endereço, 000
Filtros
Pizza do Bigode
Pizzaria Sorriso
Pizza da Nonna
Pizzaria da Cidade

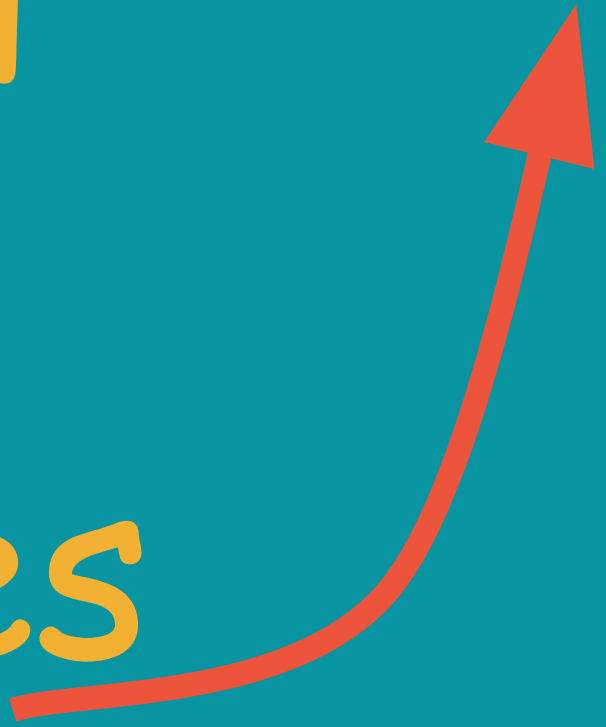
BIG changes

30Mbps
R$ 89,90
2019

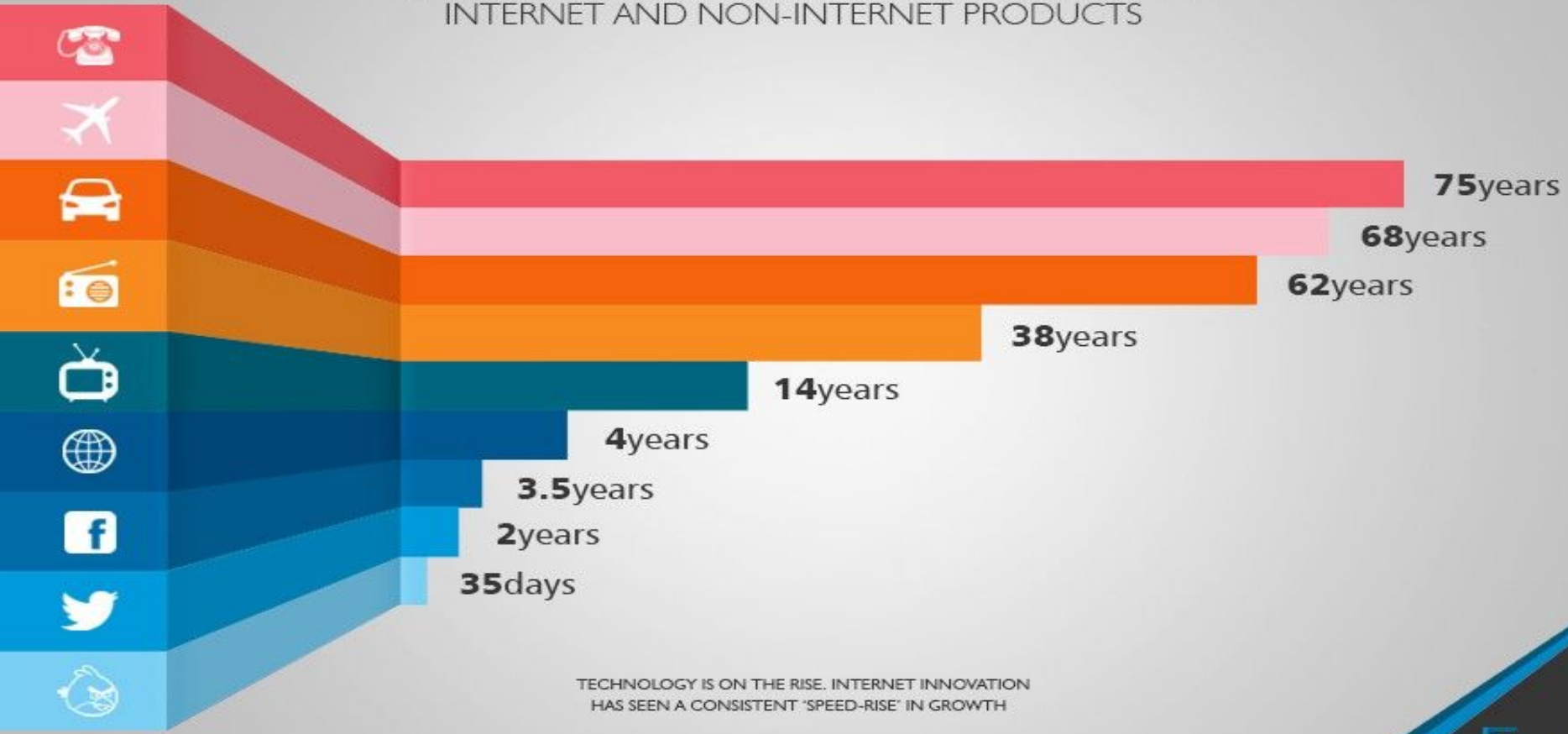
125 kbps
R$ 59,90
2009

What do they have in common?

exponential growth of technologies



REACHING 50 MILLION USERS: THE JOURNEY OF INTERNET AND NON-INTERNET PRODUCTS

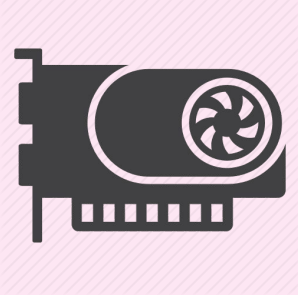


TECHNOLOGY IS ON THE RISE. INTERNET INNOVATION HAS SEEN A CONSISTENT 'SPEED-RISE' IN GROWTH

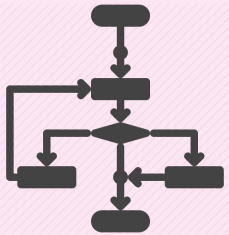
TECHTODAY

DATA

INTERNET



HARDWARE



ALGORITHM

data comes from everywhere

Data Files
(XML, CSV, Excel, JSON, …)

Database
(MySQL, Oracle, …)

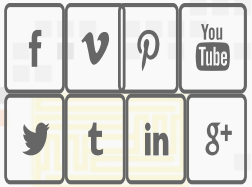
API

Sites

Text and reports

Maps

Image and videos



Social Media

Facebook

Monthly Active Users: **2.2 Billion**

Daily Active Users: **1.4 Billion**

Founded: **2004**

Photos uploaded daily: **300 Million**

Video views daily: **8 Billion**

Rank: **#1**

WhatsApp

Monthly Active Users: **700 Million**

Daily Active Users: **320 Million**

Founded: **2009**

New users daily: **1 Million**

Messages sent daily: **43 Billion**

Rank: **#4**

YouTube

Monthly Active Users: **1.5 Billion**

Daily Active Users: **30 Million**

Founded: **2005**

Video views daily: **5 Billion**

Average visit length: **40 min.**

Rank: **#2**

Google+

Monthly Active Users: **395 Million**

Total Registered Users: **2 Billion**

Founded: **2011**

U.S. based users: **55%**

Ages 15-34 users: **28%**

Rank: **#5**

Instagram

Monthly Active Users: **800 Million**

Daily Active Users: **500 Million**

Founded: **2010**

Photos uploaded daily: **95 Million**

Stories daily: **250 Million**

Rank: **#3**

Twitter

Monthly Active Users: **330 Million**

Daily Active Users: **100 Million**

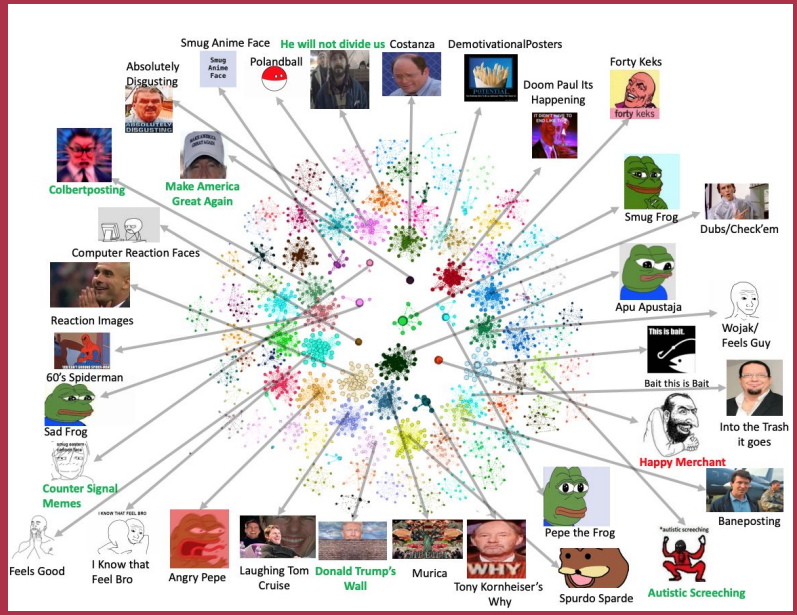
Founded: **2006**

Tweets published daily: **140 Million**

New accounts daily: **460,000**

Rank: **#6**

This is where internet memes come from



knowyourmeme.com

On the Origins of Memes by Means of Fringe Web Communities*

Savvas Zannettou*, Tristan Caulfield‡, Jeremy Blackburn†, Emiliano De Cristofaro‡,
Michael Sirivianos*, Gianluca Stringhini‡◇, and Guillermo Suarez-Tangil‡+
*Cyprus University of Technology, ‡University College London
†University of Alabama at Birmingham, ◇Boston University, +King's College London
sa.zannettou@edu.cut.ac.cy {t.caulfield,e.decristofaro}@ucl.ac.uk
blackburn@uab.edu, michael.sirivianos@cut.ac.cy, gian@bu.edu, guillermo.suarez-tangil@kcl.ac.uk

Abstract

Internet memes are increasingly used to sway and manipulate public opinion. This prompts the need to study their propagation, evolution, and influence across the Web. In this paper, we detect and measure the propagation of memes across multiple Web communities, using a processing pipeline based on perceptual hashing and clustering techniques, and a dataset of 160M images from 2.6B posts gathered from Twitter, Reddit, 4chan's Politically Incorrect board (/pol/), and Gab, over the course of 13 months. We group the images posted on fringe Web communities (/pol/, Gab, and The_Donald subreddit) into clusters, annotate them using meme metadata obtained from Know Your Meme, and also map images from mainstream communities (Twitter and Reddit) to the clusters.

Our analysis provides an assessment of the popularity and diversity of memes in the context of each community, showing, e.g., that racist memes are extremely common in fringe Web communities. We also find a substantial number of politics-related memes on both mainstream and fringe Web communities, supporting media reports that memes might be used to enhance or harm politicians. Finally, we use Hawkes processes to model the interplay between Web communities and quantify their reciprocal influence, finding that /pol/ substantially influ-

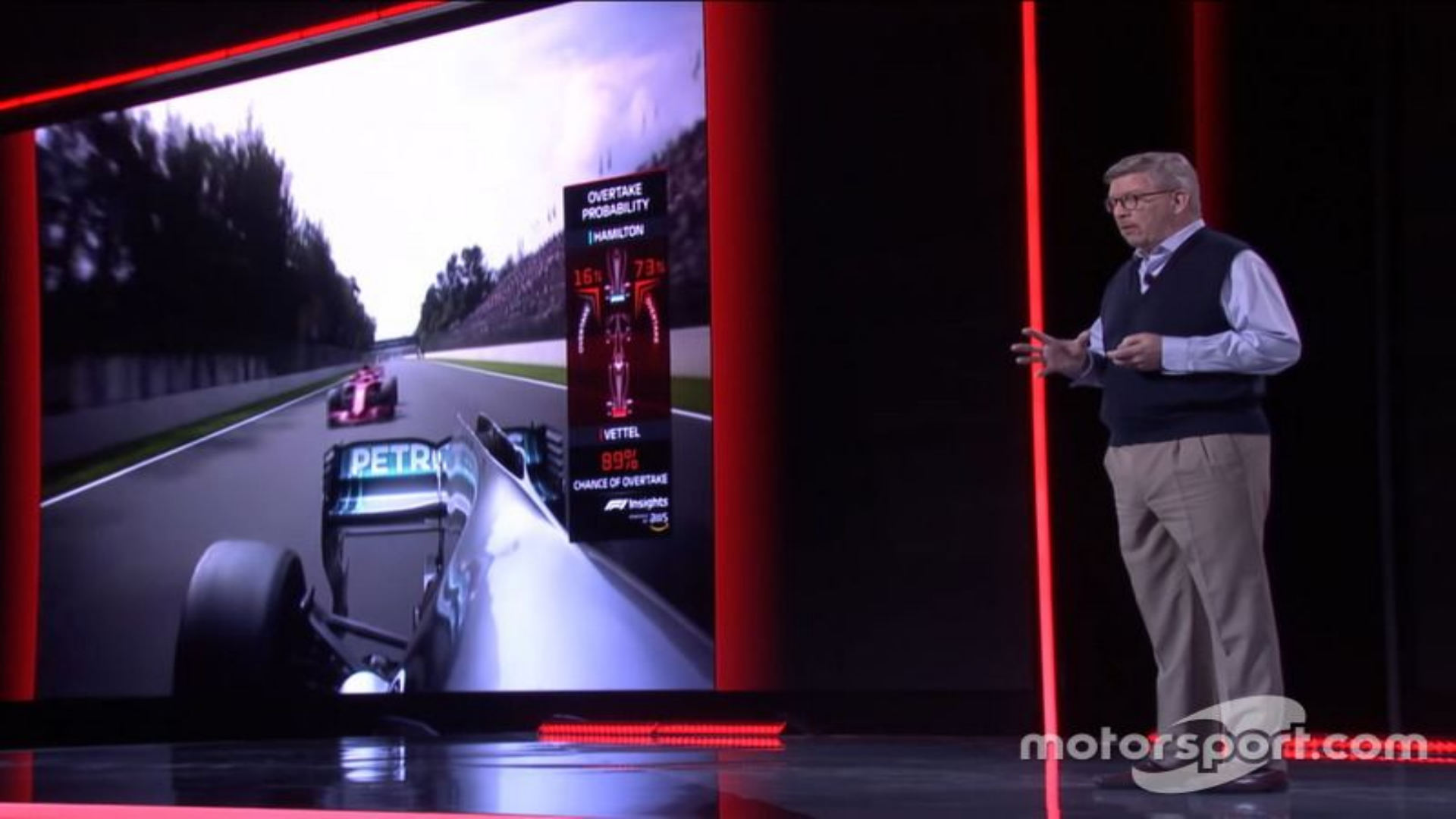
no bad intentions, others have assumed negative and/or hateful connotations, including outright racist and aggressive undertones [84]. These memes, often generated by fringe communities, are being "weaponized" and even becoming part of political and ideological propaganda [67]. For example, memes were adopted by candidates during the 2016 US Presidential Elections as part of their iconography [19]; in October 2015, then-candidate Donald Trump retweeted an image depicting him as Pepe The Frog, a controversial character considered a hate symbol [58]. In this context, polarized communities within 4chan and Reddit have been working hard to create new memes and make them go viral, aiming to increase the visibility of their ideas—a phenomenon known as "attention hacking" [64].

Motivation. Despite their increasingly relevant role, we have very little measurements and computational tools to understand the origins and the influence of memes. The online information ecosystem is very complex; social networks do not operate in a vacuum but rather influence each other as to how information spreads [86]. However, previous work (see Section 6) has mostly focused on social networks in an isolated manner.

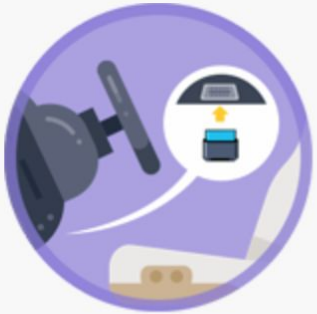

PLAYERTEK
DASHBOARD
SESSIONS
LEAGUES
FRIENDS
PRO MAPS
77
DAVID2
May 6, 2015
SESSION SUMMARY
NAME May 6, 2015 All
DATE 6/5/2015
DURATION 99.4 minutes
PITCH Oscar Traynor
Distance
Top Speed
Power Plays
5.03 km
5.89 m/s
8
SESSION SCORE
32
SESSION TAGS
Set the session's tags...
Volume
Intensity
Distance : 5.03 km
50%
Sprint Distance : 101 m
10%
Power Plays : 8
10%
Energy : 583 cal
42%
Impacts : 28
28%
My Target
Sprint Distance
Power Plays
Distance
Impacts
Energy
1st Half
2nd Half
Speed (m/s)
FIVE MINUTE BREAKDOWN
0.75 km
0.5 km
0.25 km
0 km


crefisa
fam


APEX


OVERTAKE PROBABILITY
HAMILTON
16%
73%
VETTEL
89%
CHANCE OF OVERTAKE
F1 Insights
powered by AWS
motorsport.com

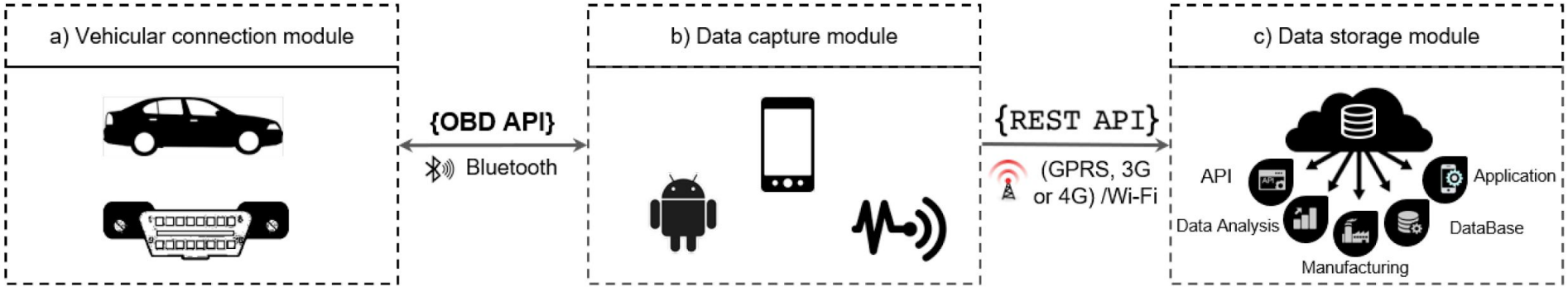
How it works

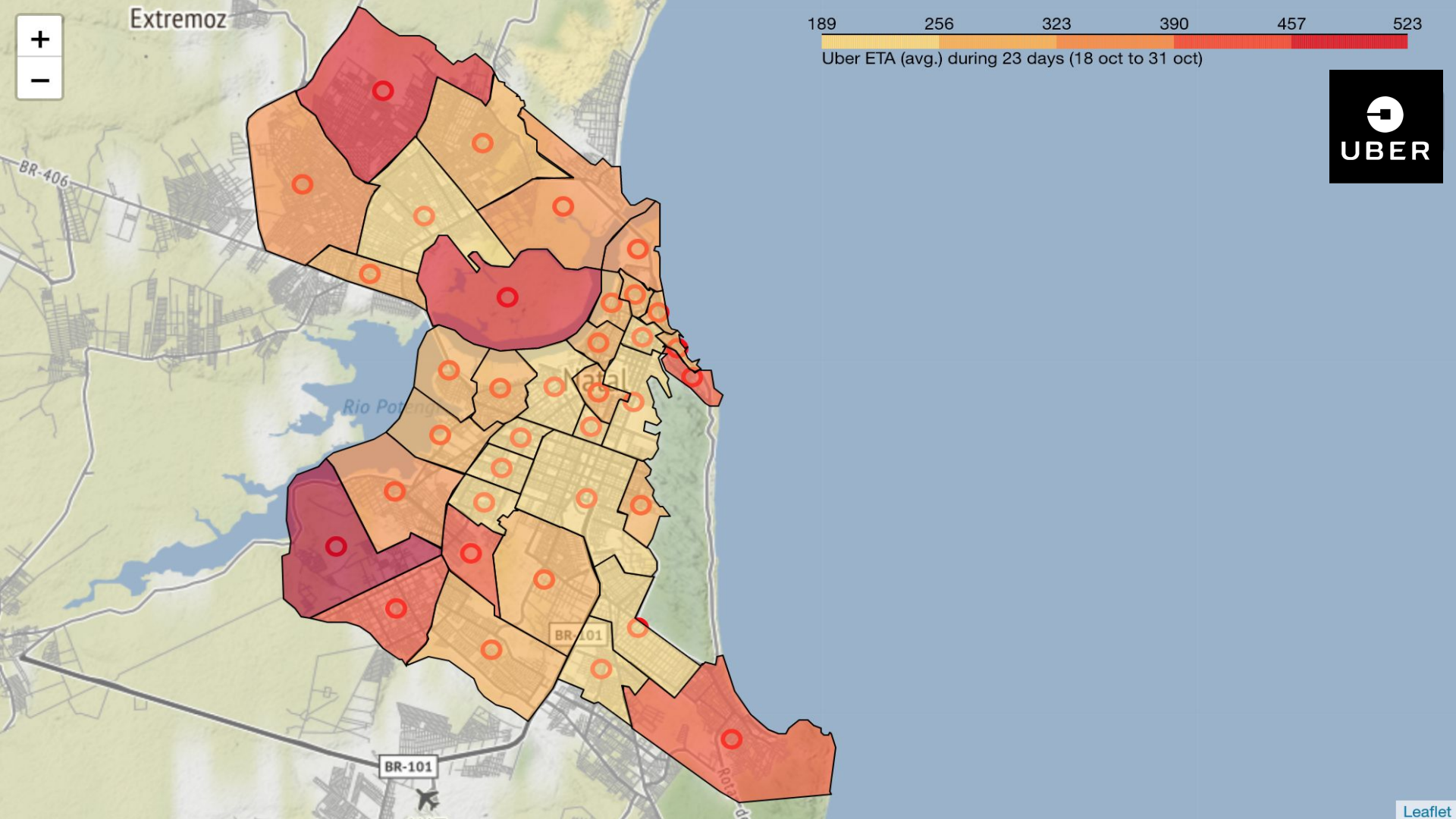


1

2

3




189
256
323
390
457
523
Uber ETA (avg.) during 23 days (18 oct to 31 oct)
UBER
Extremoz
BR-406
Natal
BR-101
BR-101
Leaflet

Measuring Depression Symptom Severity from Spoken Language and 3D Facial Expressions

Albert Haque[1] **Michelle Guo**[1] **Adam S Miner**[2,3] **Li Fei-Fei**[1]
[1]Department of Computer Science, Stanford University
[2]Department of Psychiatry and Behavioral Sciences, Stanford University
[3]Department of Health Research and Policy, Stanford University

Abstract

With more than 300 million people depressed worldwide, depression is a global problem. Due to access barriers such as social stigma, cost, and treatment availability, 60% of mentally-ill adults do not receive any mental health services. Effective and efficient diagnosis relies on detecting clinical symptoms of depression. Automatic detection of depressive symptoms would potentially improve diagnostic accuracy and availability, leading to faster intervention. In this work, we present a machine learning method for measuring the severity of depressive symptoms. Our multi-modal method uses 3D facial expressions and spoken language, commonly available from modern cell phones. It demonstrates an average error of 3.67 points (15.3% relative) on the clinically-validated Patient Health Questionnaire (PHQ) scale. For detecting major depressive disorder, our model demonstrates 83.3% sensitivity and 82.6% specificity. Overall, this paper shows how speech recognition, computer vision, and natural language processing can be combined to assist mental health patients and practitioners. This technology could be deployed to cell phones worldwide and facilitate low-cost universal access to mental health care.

Intelligent Machines

Your smartphone's AI algorithms could tell if you are depressed

Smartphones that are used to track our faces and voices could also help lower the barrier to mental-health diagnosis and treatment.

by Will Knight December 3, 2018

https://www.ibm.com/blogs/research/2018/09/agropad/

Ping a Good Doctor

ROBÔ PARA IDENTIFICAÇÃO DE FALHAS EM VEÍCULOS

EMBRAPII
Empresa Brasileira de Pesquisa e Inovação Industrial

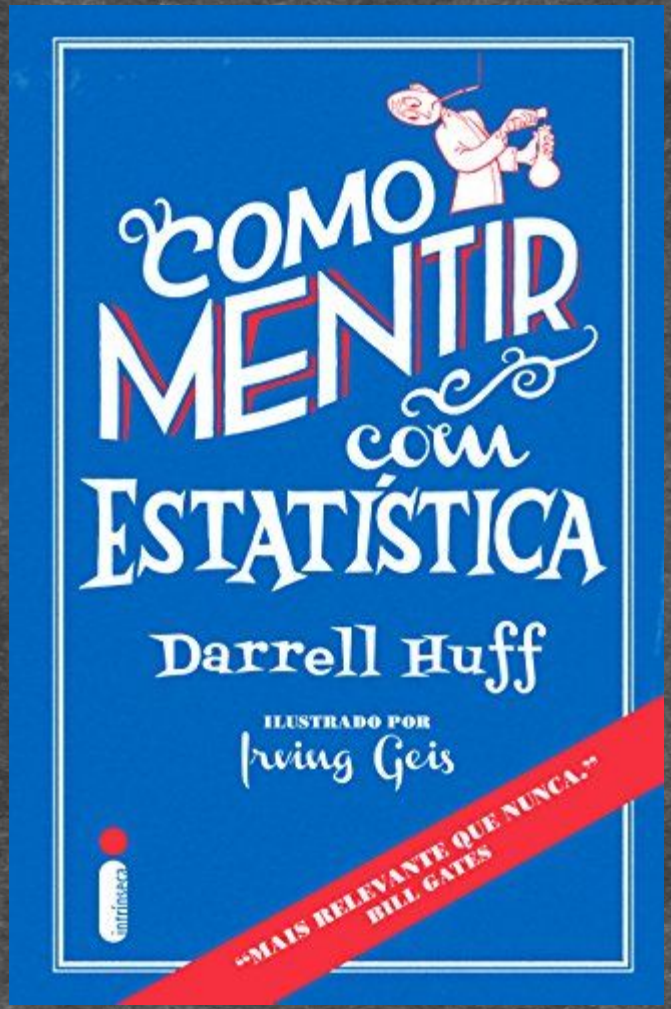
APOIO EMBRAPII

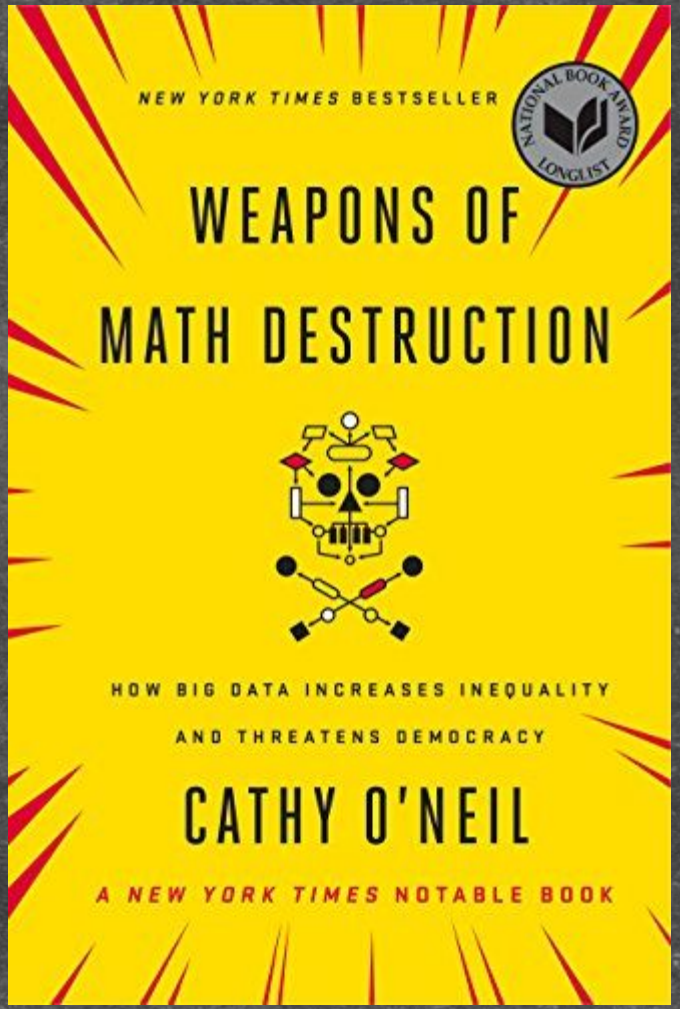
McCormick & IBM Announce Collaboration Pioneering the Use of Artificial Intelligence in Flavor and Food Product Development

http://bit.do/temperosai

DATA VIOLENCE

and how bad engineering choices can damage society


COMO MENTIR com ESTATÍSTICA
Darrell Huff
ILUSTRADO POR
Irving Geis
"MAIS RELEVANTE QUE NUNCA."
BILL GATES
intrínseca


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


What've you got for me?
INTERNET
FAKE NEWS
FAKE
Article That Reinforces Your Opinions
I'll take all of it.
inkyrickshaw.com
@inkyrickshaw

https://goo.gl/NwP7Fv

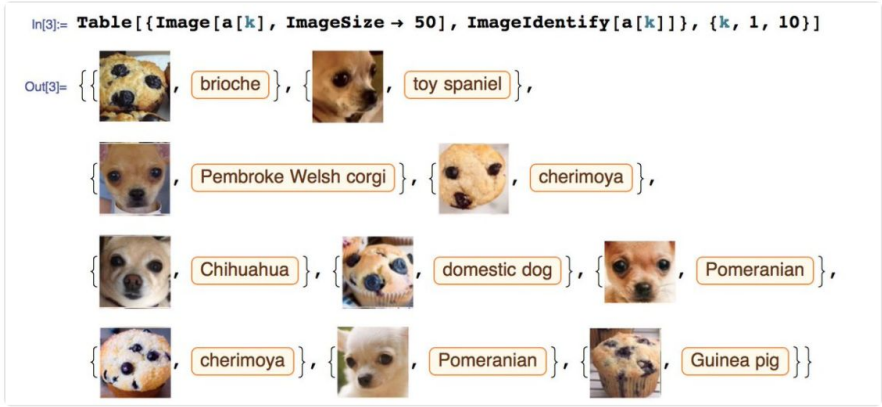
https://goo.gl/xzLxaY

https://gizmodo.uol.com.br/reconhecimento-facial-falha-china/

https://www.technologyreview.com/the-download/612916/even-amazon-wants-face-recognition-regulated/

Replying to @ProfMike_M

Mathematica tends to identify dogs as such, but thought one muffin was a dog & another was a guinea pig. @ProfMike_M



7:42 AM - 11 Mar 2016

●●●●● Verizon 4:20 PM 34%

Albums **puppy or bagel** Select

●●●○○ Verizon 10:50 PM 4%

Back **labradoodle or fried chicken** Select


Nice science,
but don't
forget about
the ethics

Many popular iPhone apps secretly record your screen without asking

https://techcrunch.com/2019/02/06/iphone-session-replay-screenshots/

Google and Waze Must Stop Sharing Drunk-Driving Checkpoints

https://www.nytimes.com/2019/02/06/nyregion/waze-nypd-location.html

Who studies this stuff?

A Data-Driven Program

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

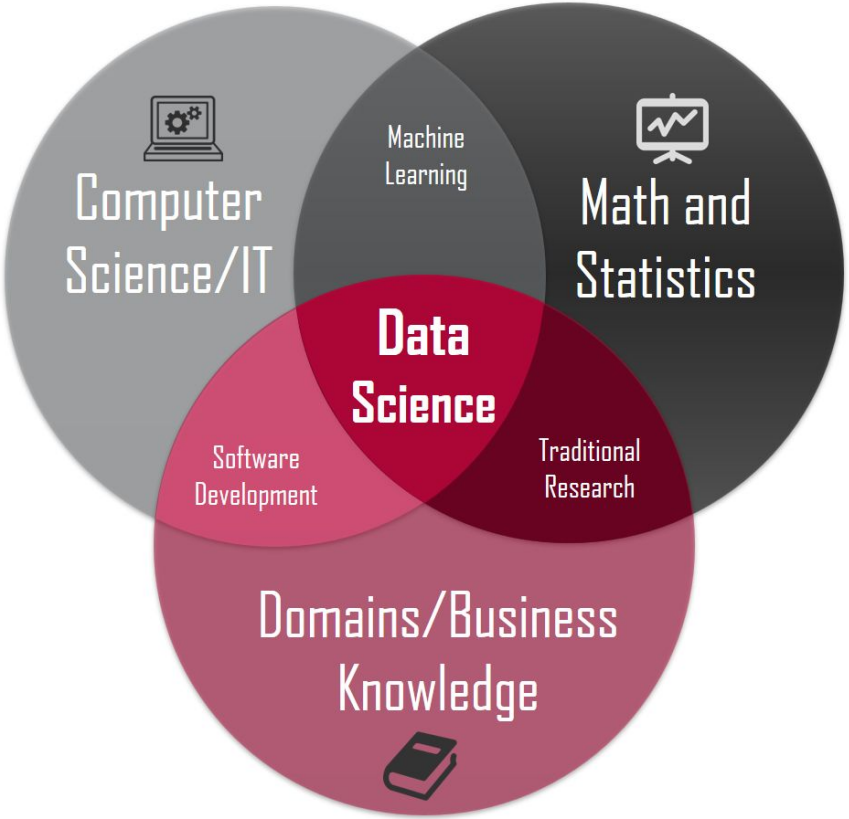
- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

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

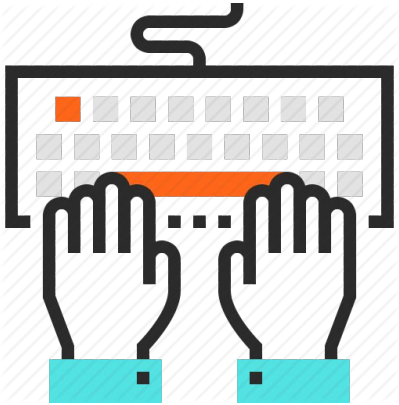
- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau




jobs

Learning by doing

#no_exams #assessments_all_time #projects
#dont_reinvent_the_wheel



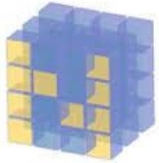
python

DataCamp

git

ANACONDA
Powered by Continuum Analytics®

colab



NumPy

pandas

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

Keras

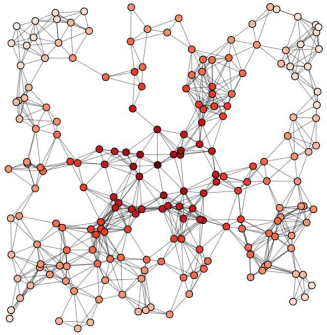
TensorFlow

NetworkX

PyGSP

matplotlib

Folium



Seaborn

Requests

bokeh

Leaflet

Beautiful Soup

IMD0905 Data Science I Syllabus - 2019.1

Data Science Foundation

Data Pipeline

- Store
- Collect
- Cleaning
- Analyse
- Visualize

Clustering and network analysis

HARD WORKING
part #01

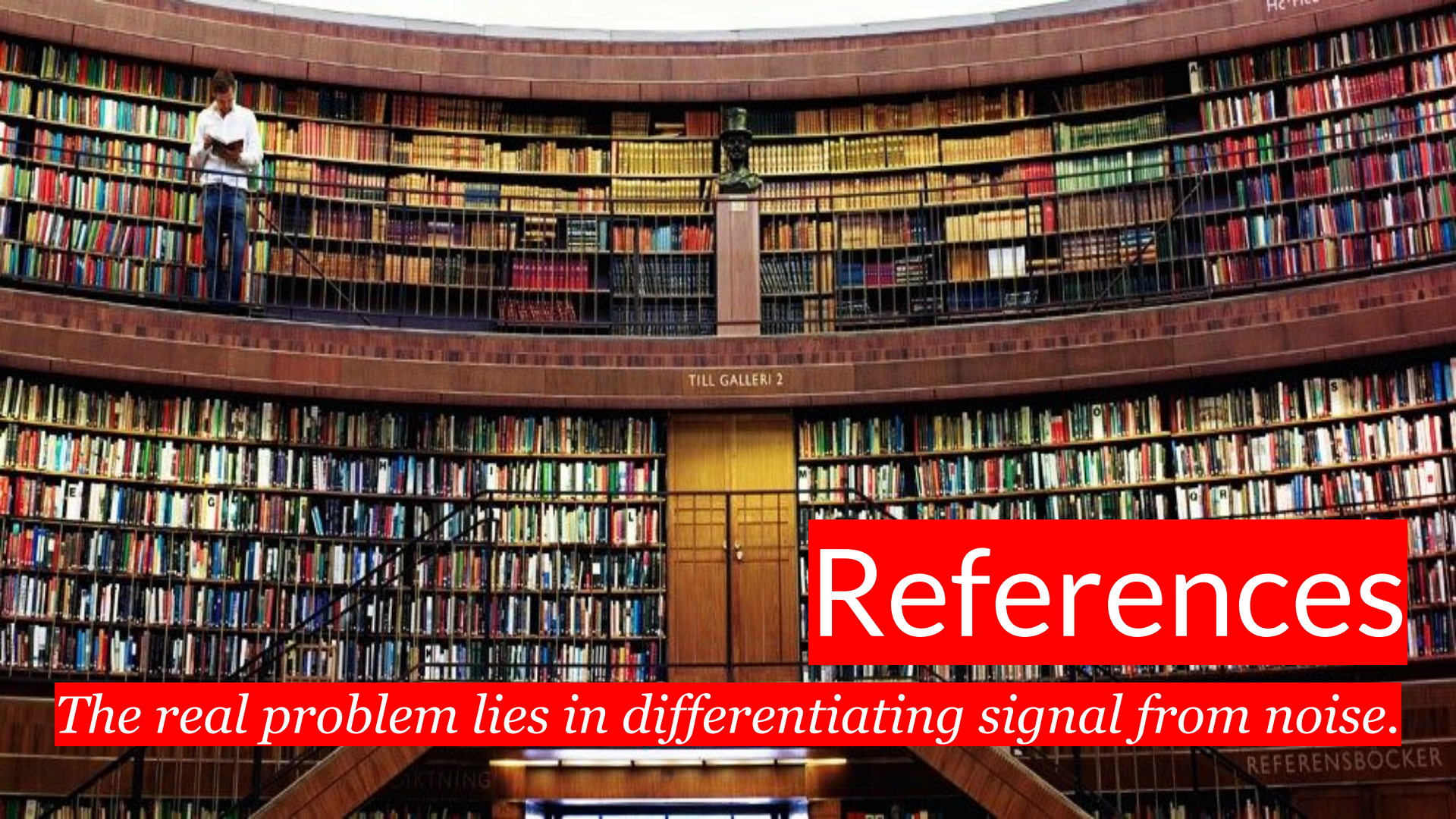
- Python Crash Course
- Pandas
- 02 Assessments (individual, group[2])
- Datacamp Courses
  - 16 courses
- End: 25/03

KEEP ON RUNNING
part #02

- Visualization using maps
- Data cleaning
- 01 Big Assessment (group[3-4])
- Datacamp Courses
  - 1 course
  - 10k points (whatever)
- End: 29/04

FINAL SPRINT
part #03

- Optimizing Dataframe
- Python & Database
- Web scraping Introduction
- Network Analysis: mining social web
- 01 Assessment about Optimization & Database (group[2])
- Final Project - Web scraping and network analysis (group)
- Datacamp Courses
  - 5 courses
- End: 26/06

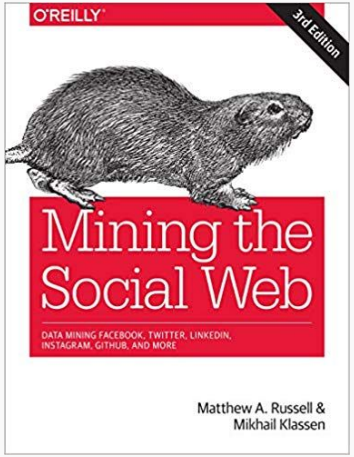


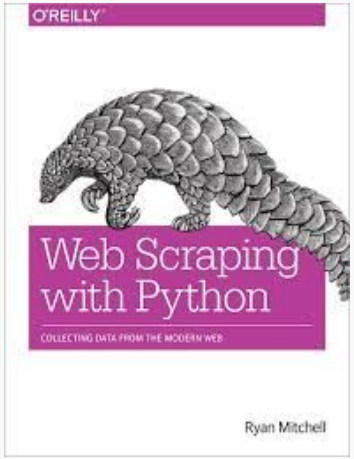
References

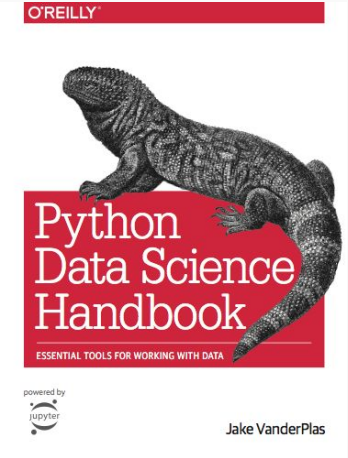
The real problem lies in differentiating signal from noise.

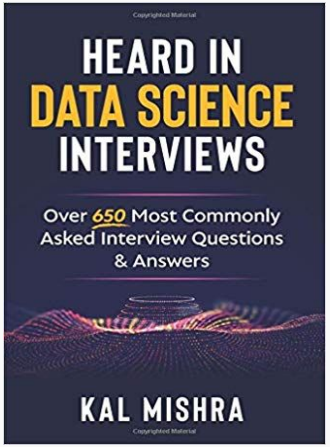
References Online

References

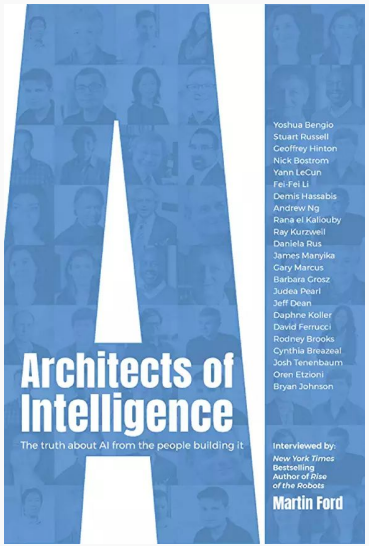

O'REILLY
3rd Edition
Mining the Social Web
DATA MINING FACEBOOK, TWITTER, LINKEDIN, INSTAGRAM, GITHUB, AND MORE
Matthew A. Russell & Mikhail Klassen

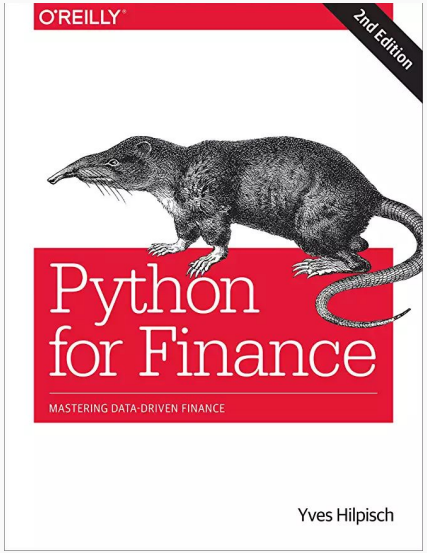

O'REILLY
Web Scraping with Python
COLLECTING DATA FROM THE MODERN WEB
Ryan Mitchell


O'REILLY
Python Data Science Handbook
ESSENTIAL TOOLS FOR WORKING WITH DATA
powered by
jupyter
Jake VanderPlas


HEARD IN
DATA SCIENCE
INTERVIEWS
Over 650 Most Commonly Asked Interview Questions & Answers
KAL MISHRA

References (out of scope)





Faça parte da maior comunidade de Data Science do Brasil!

Cresça junto com profissionais de data science, machine learning, big data e inteligência artificial

Inscreva-se na newsletter quinzenal

Digite seu melhor email **Inscreva-se**

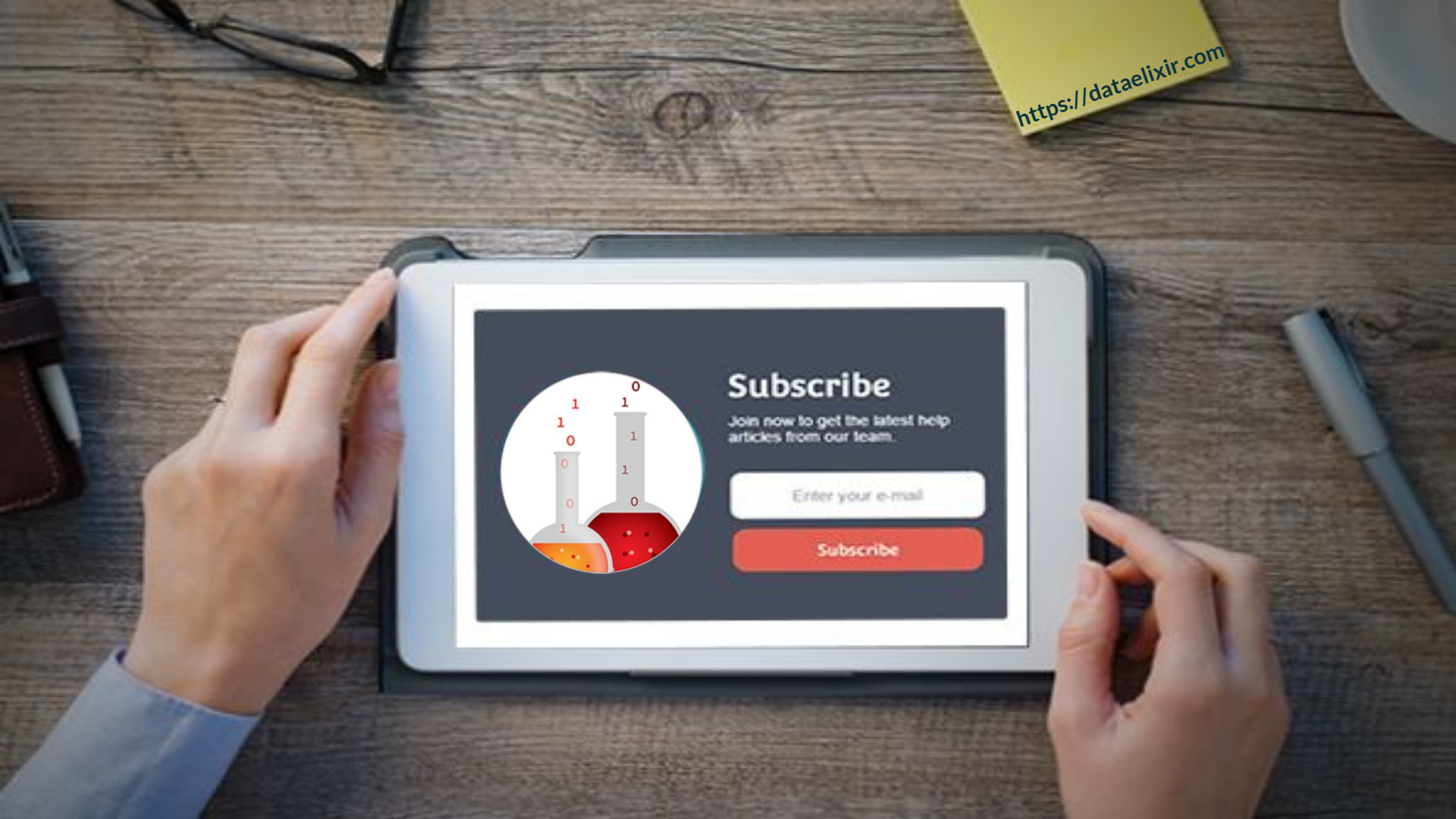
Participar também da comunidade no Slack

http://bit.do/ds_podcast_intro

hipsters ponto tech

#134

Primeiros Passos em Data Science: Do Excel e BI ao Python


https://dataelixir.com
Subscribe
Join now to get the latest help articles from our team.
Enter your e-mail
Subscribe