IMD0905 - Data Science I

Lesson #1 - Outline & Directions

Ivanovitch Silva

August, 2018


Introduction

Ivanovitch Silva (ivan@imd.ufrn.br)
Office: CIVT-A226 6T1234

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

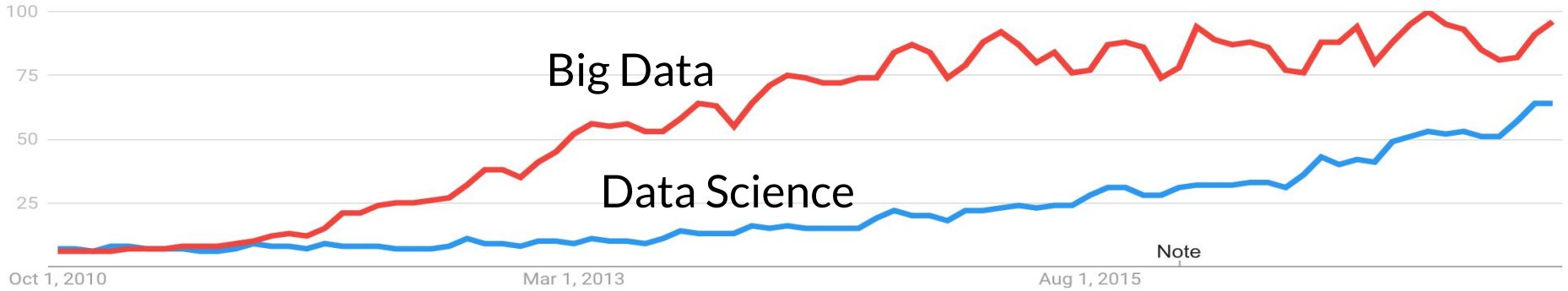
Big Data and Social Analytics certificate course

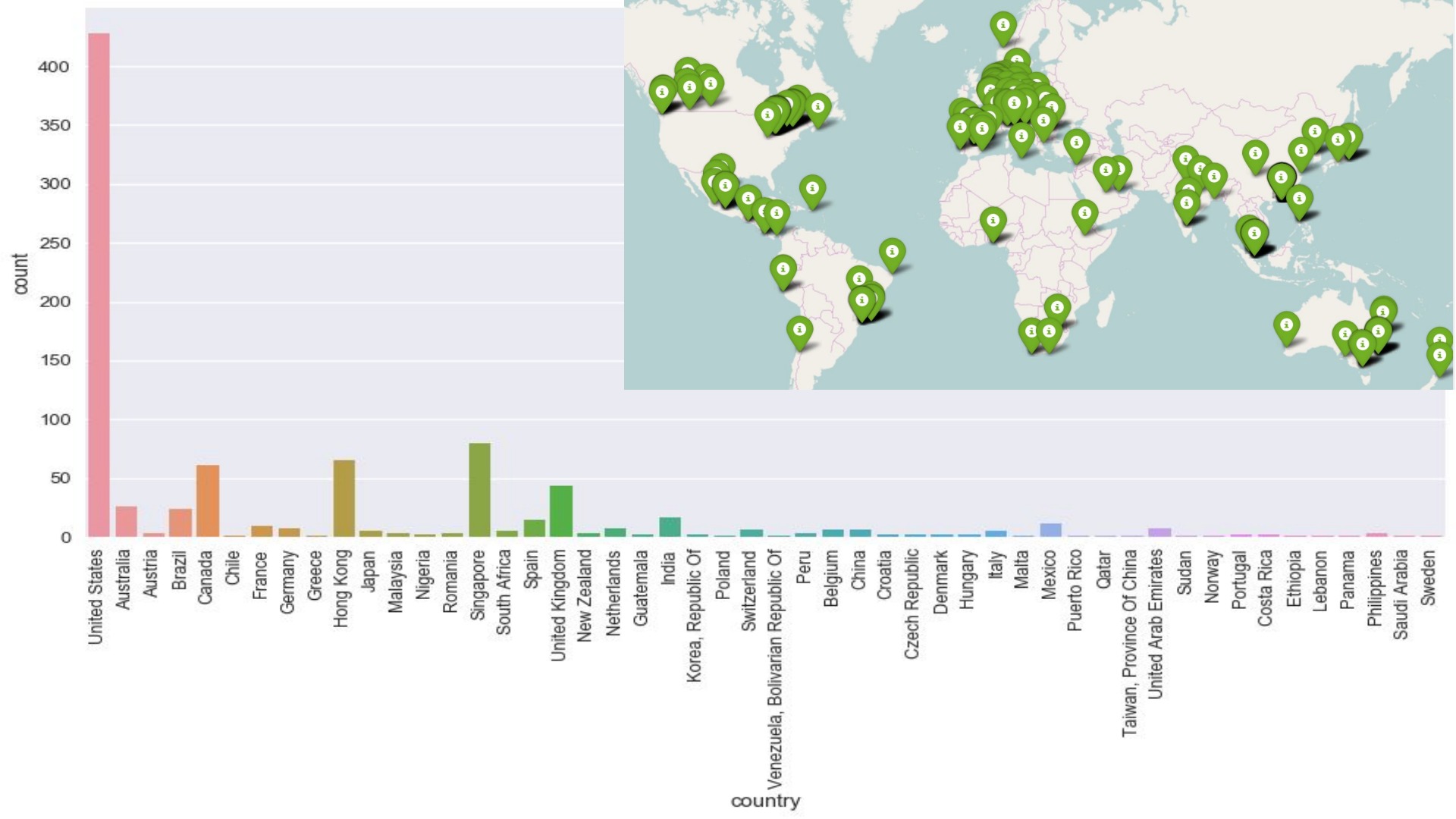
EXperimental Learning

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

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)

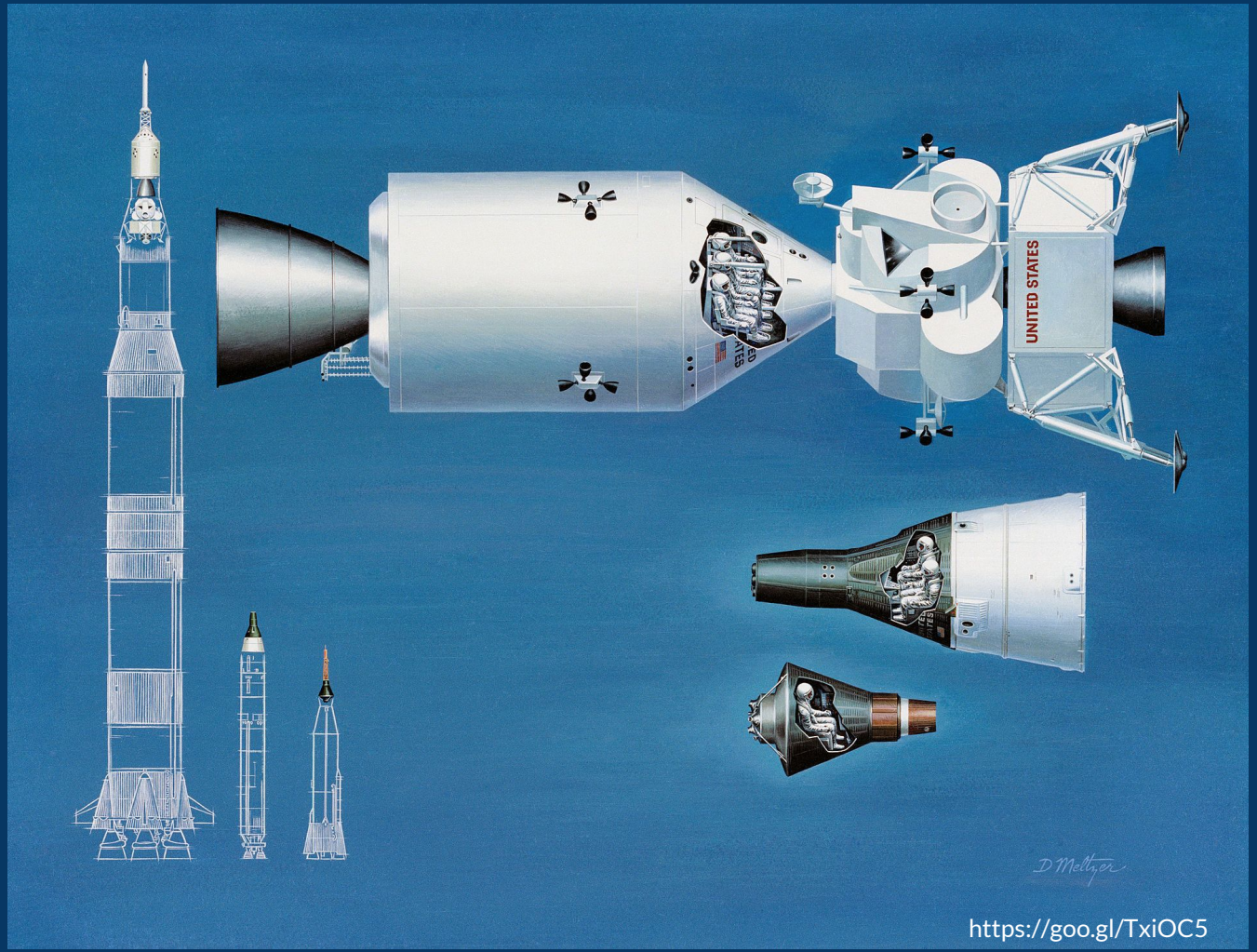
Graduate
2017.2 - EEC2006 Data Science Foundations
2017.2 - ITE0021 Learning Analytics

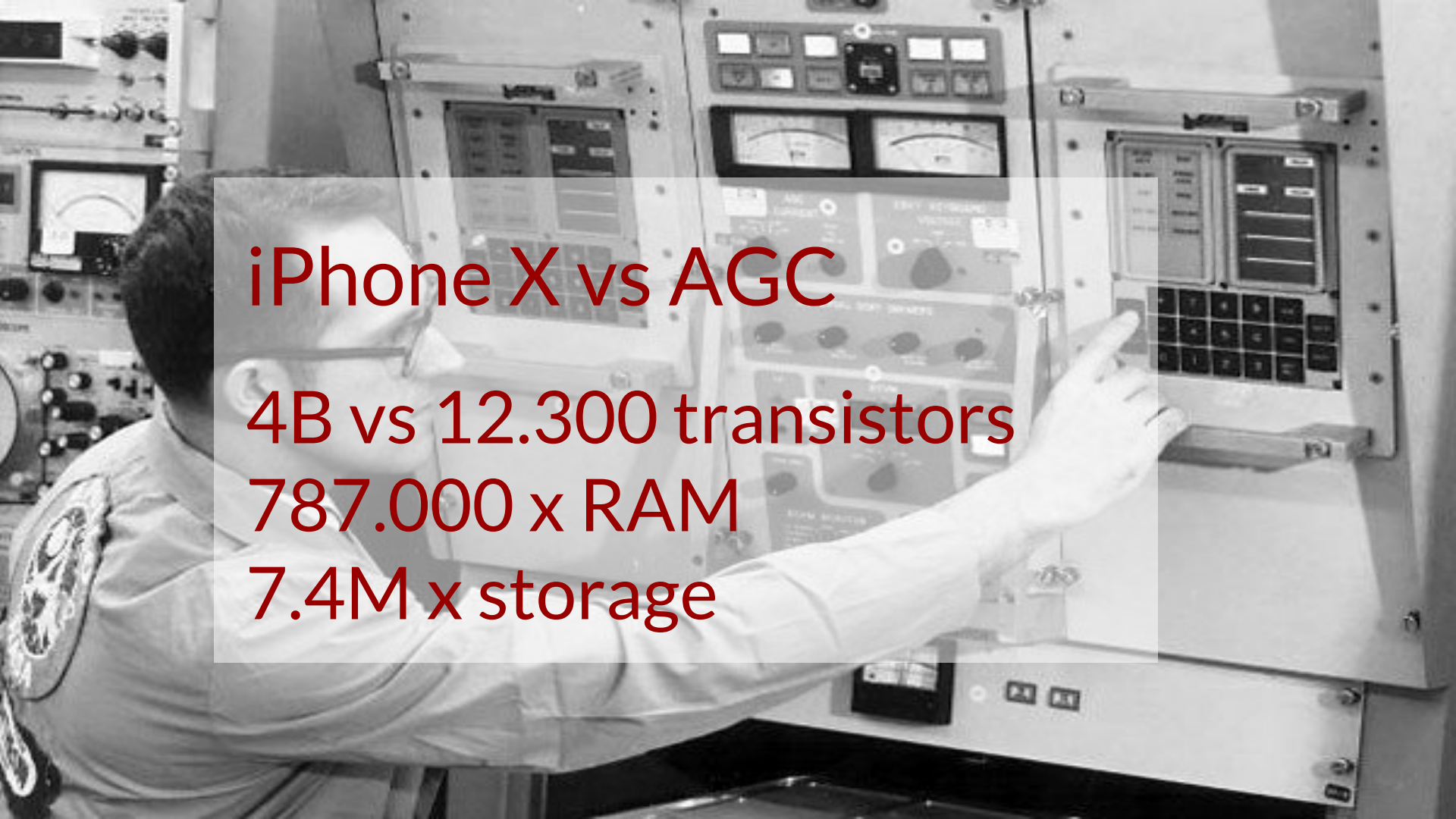
https://github.com/ivanovitchm

Provocation #1

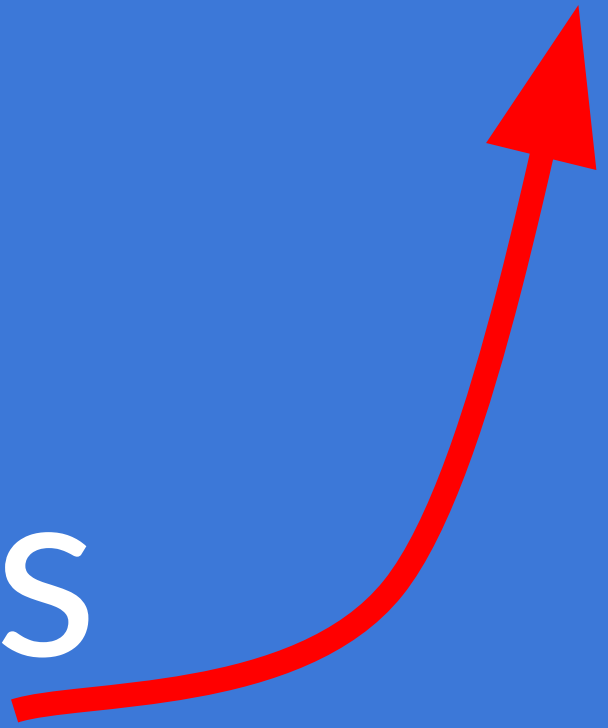
hardware

1969

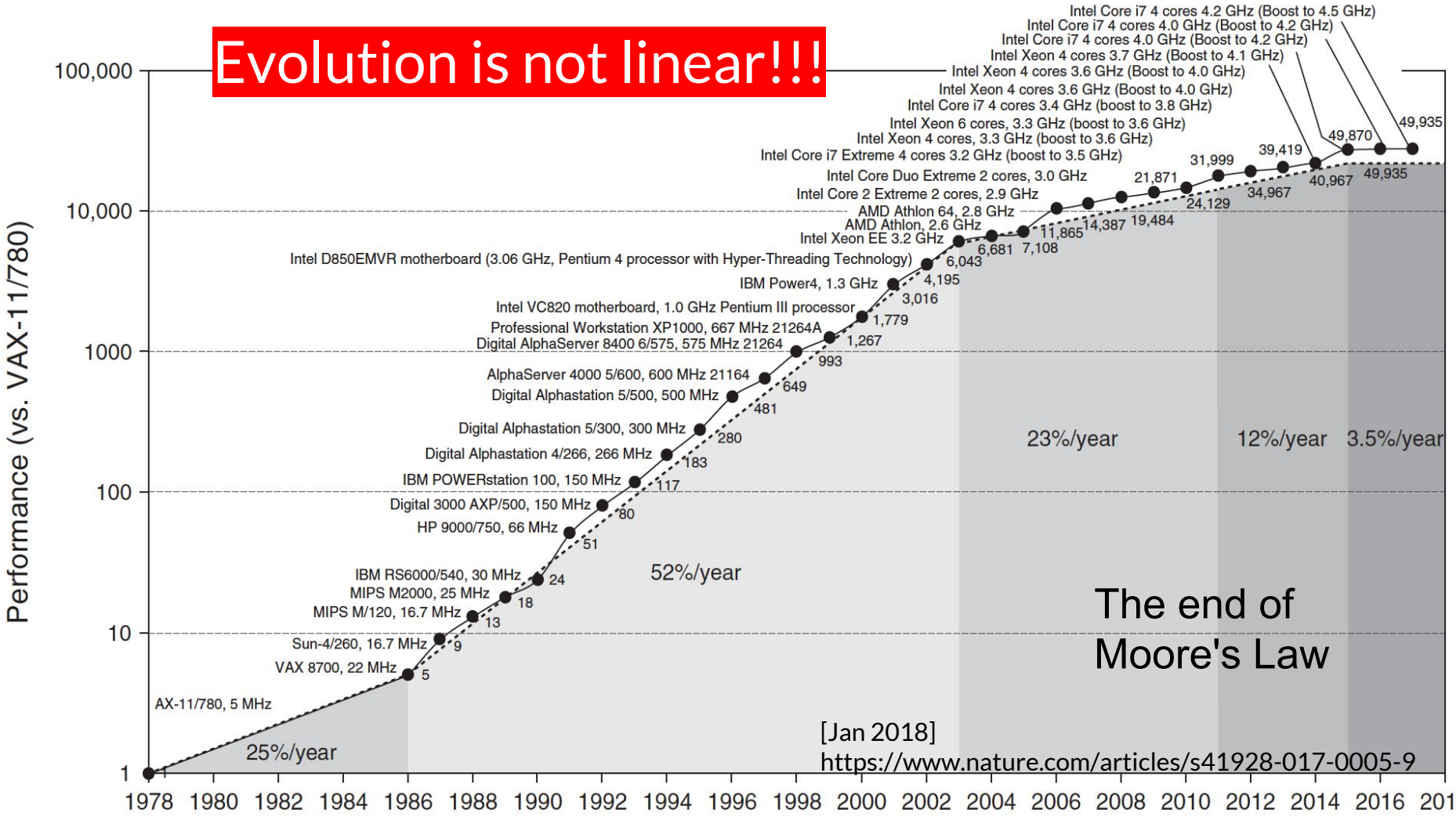

UNITED STATES
https://goo.gl/TxiOC5


iPhone X vs AGC
4B vs 12.300 transistors
787.000 x RAM
7.4M x storage

exponential growth of technologies



Evolution is not linear!!!



$ 2,999.00

NVIDIA TITAN V

The Most Powerful PC GPU Ever Created

NVIDIA TITAN V is the most powerful graphics card ever created for the PC, driven by the world's most advanced architecture—NVIDIA Volta. NVIDIA's supercomputing GPU architecture is now here for your PC, and fueling breakthroughs in every industry.

LEARN MORE

	NVIDIA TITAN V
Architecture	**NVIDIA Volta**
Frame Buffer	**12 GB HBM2**
Boost Clock	**1455 MHz**
Tensor Cores	**640**
CUDA Cores	**5120**

CÁLCULO DO IMPOSTO

http://www.tributado.net/

1. COTAÇÕES

USD - Dólar:	3.58
EUR - Euro:	4.25
GBP - Libra Esterlina:	4.84

2. ALÍQUOTA DE ICMS

Tipo de envio:	Courier
Unidade federativa:	RN
Valor da alíquota:	17%

3. TAXAS

Conversão monetária:	3.58
Imposto de importação (%):	60
ICMS (%):	0
IOF (%):	6.38

4. IMPOSTO DE IMPORTAÇÃO

VALORES

Valor do produto (USD):	2999
Custo do frete (USD):	50
Valor do produto (BRL):	10736.42
Custo do frete (BRL):	179.00
TOTAL DA COMPRA (BRL):	10915.42

TRIBUTOS

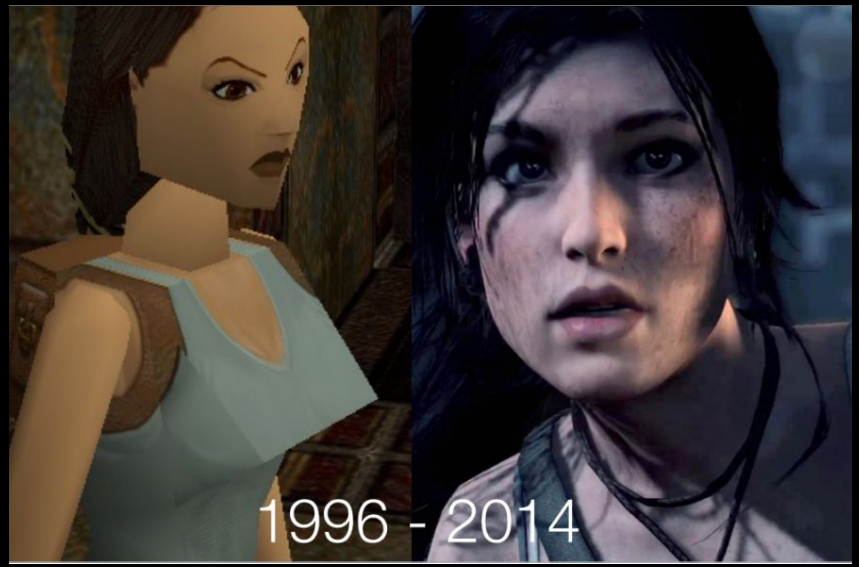
☑ Incluir frete no cálculo do imposto ☑ Incluir IOF

Imposto de importação (BRL):	6549.25
ICMS (BRL):	0.00
IOF (BRL):	696.40
TOTAL DE TRIBUTOS (BRL):	7245.66

TOTAIS

Valor final com impostos (BRL):	18161.08

Lara Croft has changed over 21 years



1/6 cost | 1/20 power | 4 hacks in a box

CES 2018 with NVIDIA CEO Jensen Huang

Provocation #2

data & internet

90's

"Read"
Search Engine
Google

Update

00's

"Write"
Social Networks
Facebook

Participate

10's

"Act"
App
Uber

Act

20's

"Change"
AI
?

Transform

Change

Act

UBER

Write

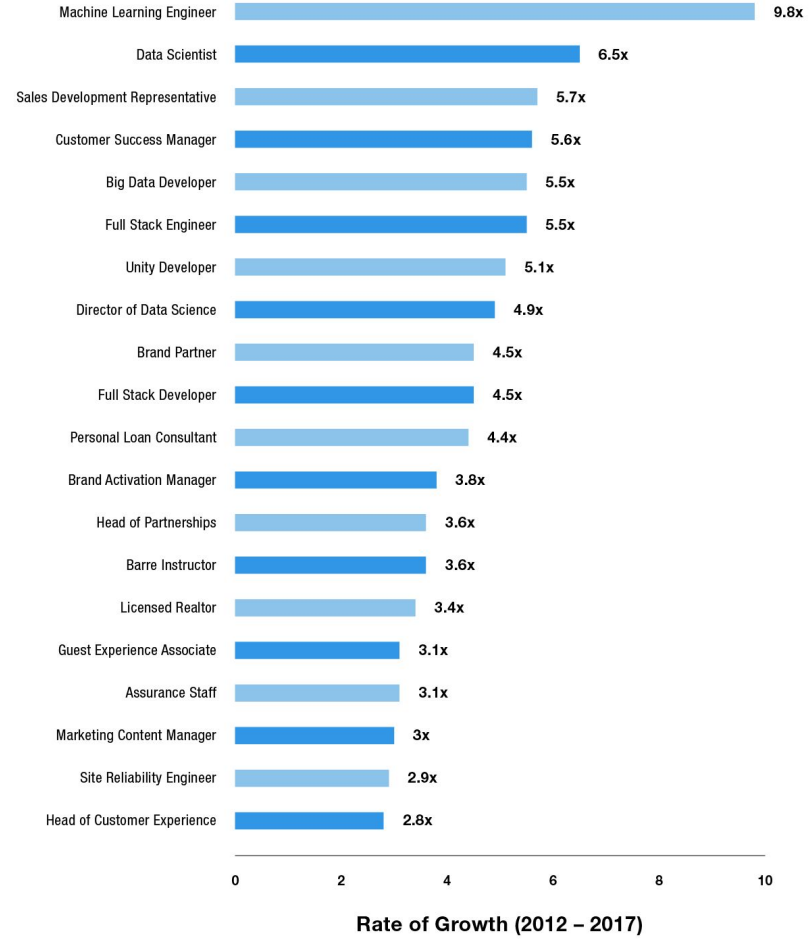
facebook

Read

Google

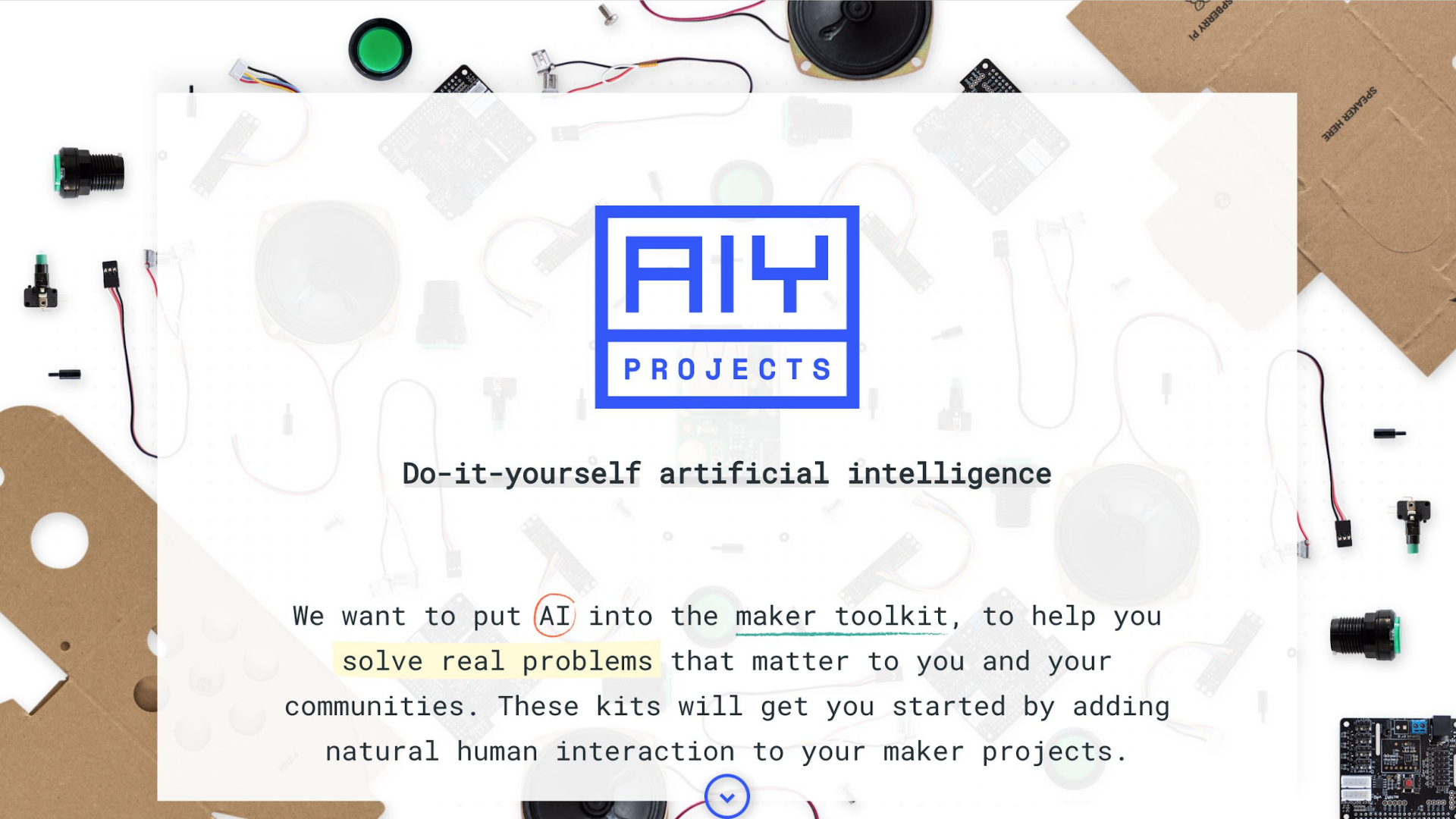
https://datafloq.com/read/the-top-7-big-data-trends-for-2017/2493

Top 20 Emerging Jobs



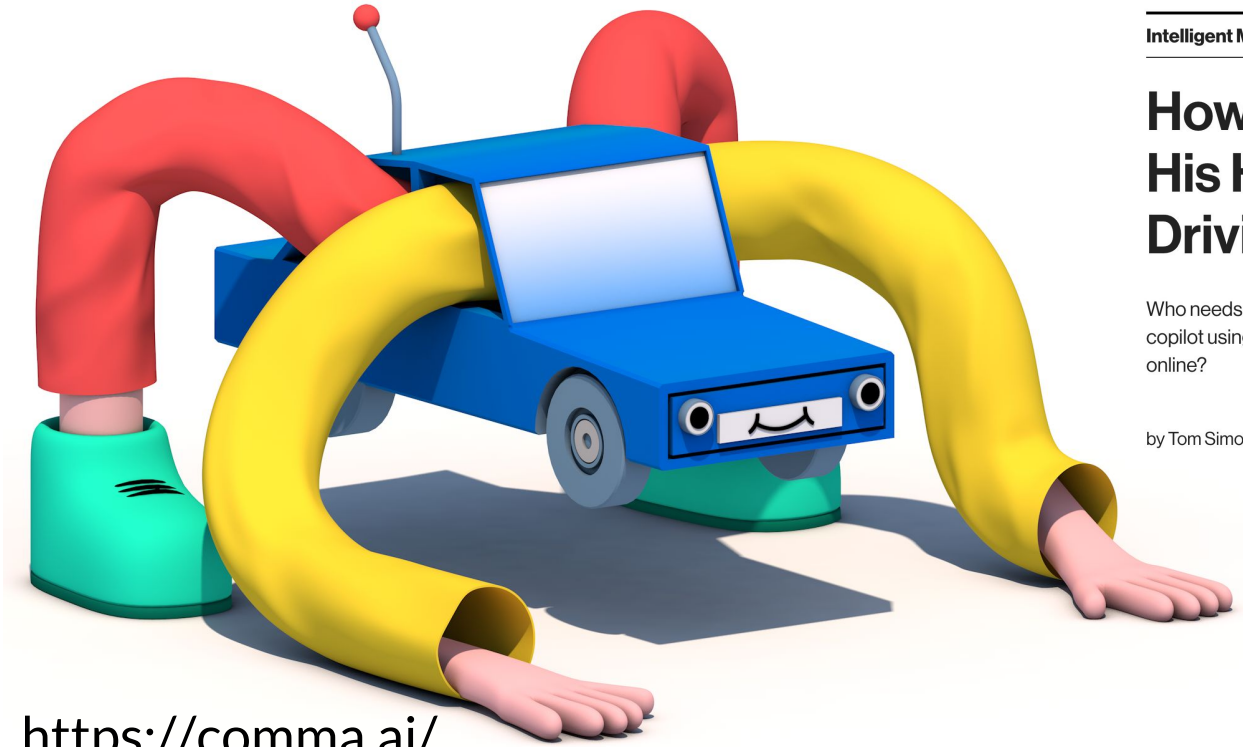
There are 9.8 times more Machine Learning Engineers working today than five years ago based on LinkedIn's research

[Dec. 2017]
http://bit.do/forbesjobs



Do-it-yourself artificial intelligence

We want to put AI into the maker toolkit, to help you solve real problems that matter to you and your communities. These kits will get you started by adding natural human interaction to your maker projects.



https://comma.ai/

Intelligent Machines

How a College Kid Made His Honda Civic Self-Driving for $700

Who needs a Tesla when you can build your own automated copilot using free hardware designs and software available online?

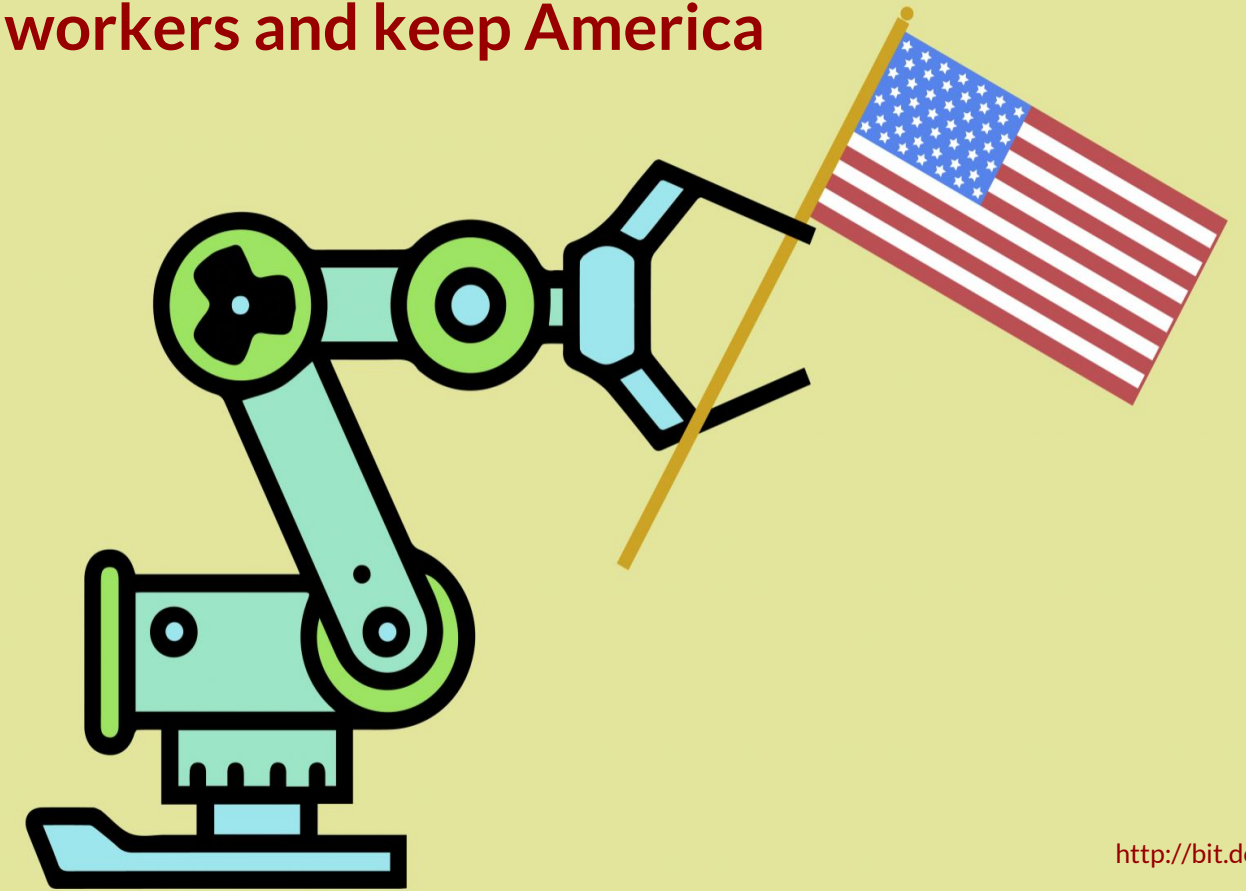
by Tom Simonite February 21, 2017

AI arms race

https://www.ft.com/content/e33a6994-447e-11e8-93cf-67ac3a6482fd

The White House says a new AI task force will protect workers and keep America first.

May 10, 2018



http://bit.do/trump_aiprogram

AI FOR HUMANITY

FRENCH STRATEGY FOR ARTIFICIAL INTELLIGENCE

The President of the French Republic presented **his vision and strategy** to make France a **leader in artificial intelligence (AI)** at the Collège de France on **29 March 2018.**

Download the Villani Report

https://www.aiforhumanity.fr/en/

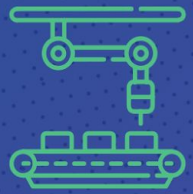

10
10

ABDI

MINISTÉRIO DA INDÚSTRIA, COMÉRCIO E SERVIÇOS

GOVERNO FEDERAL

http://www.industria40.gov.br/



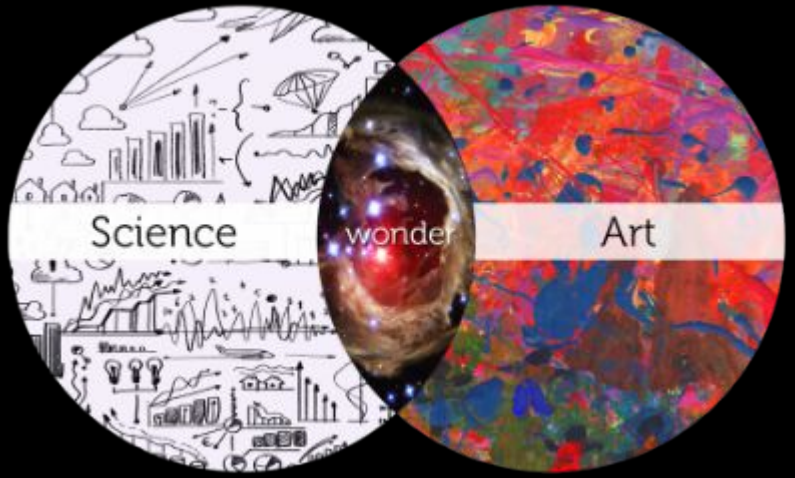
Agenda brasileira para a Indústria 4.0

O Brasil preparado para os desafios do futuro

CONHEÇA A AGENDA

Estudo “Internet das Coisas: um plano de ação para o Brasil”

Baixe o Relatório de Plano de Ação (*PDF - 3,3 MB*) para o desenvolvimento de IoT no Brasil.



https://www.youtube.com/watch?v=SOtm7vyIwxc

Spotify

Spotify

MACHINE LEARNING DAY

MONDAY, JULY THE 9TH

STOCKHOLM, SWEDEN

https://labs.spotify.com/2018/06/29/spotify-ml-day-july-9th-2018-stockholm/

HOSPITAL SÍRIO-LIBANÊS

KUNUMI

http://bhtec.org.br/empresas-do-bhtec/kunumi/

https://www.ufmg.br/online/radio/arquivos/046358.shtml

Uma nova forma de investir.

Niantic is going to crowdsource AR maps

https://techcrunch.com/2018/05/09/niantic-is-going-to-crowdsource-ar-maps/

90% of the data in the world today
has been created in the last two years alone

This volume is continuing to grow

Data volume will grow

800%

over the next five years

gartner

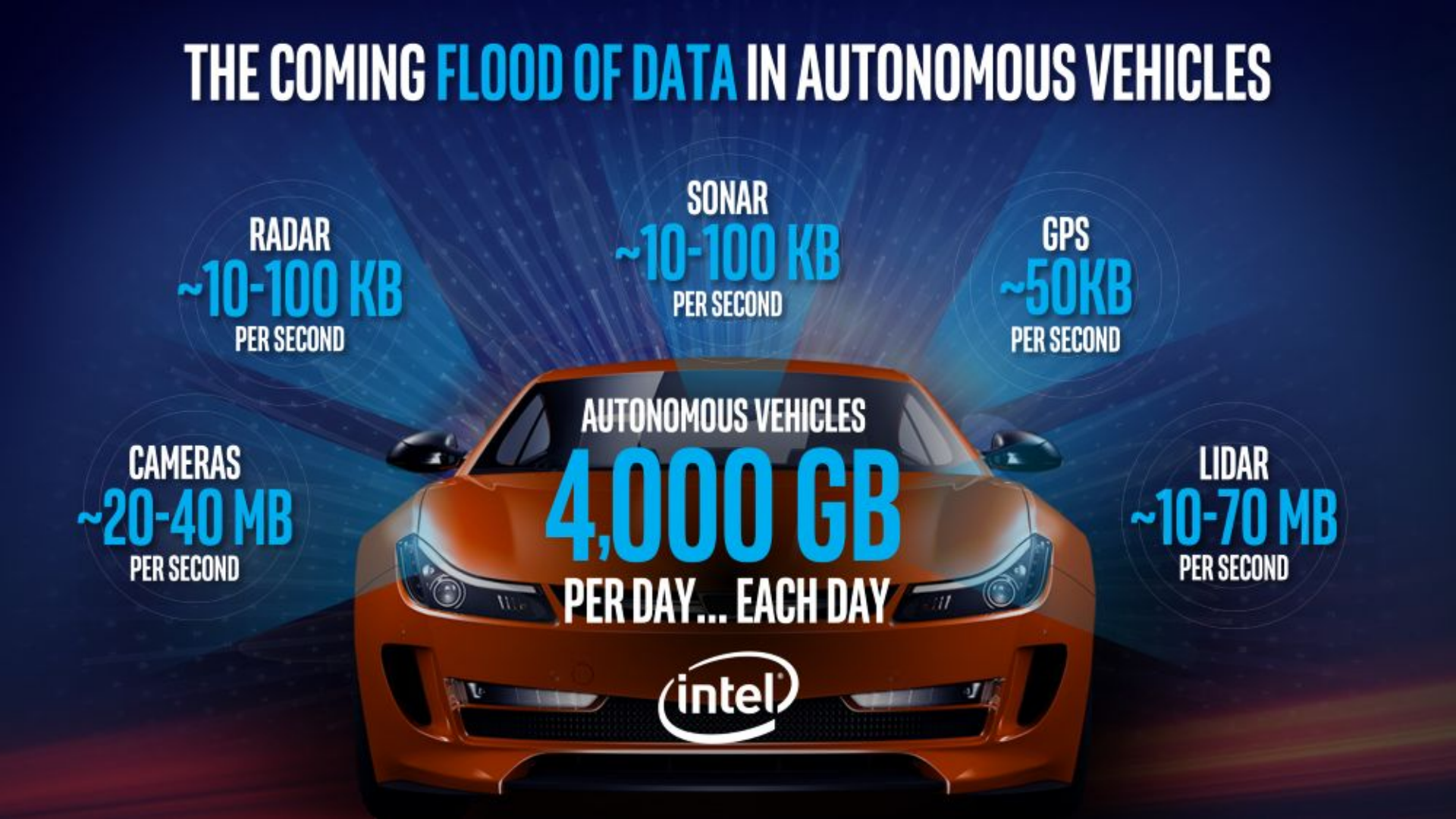
By 2022

93%

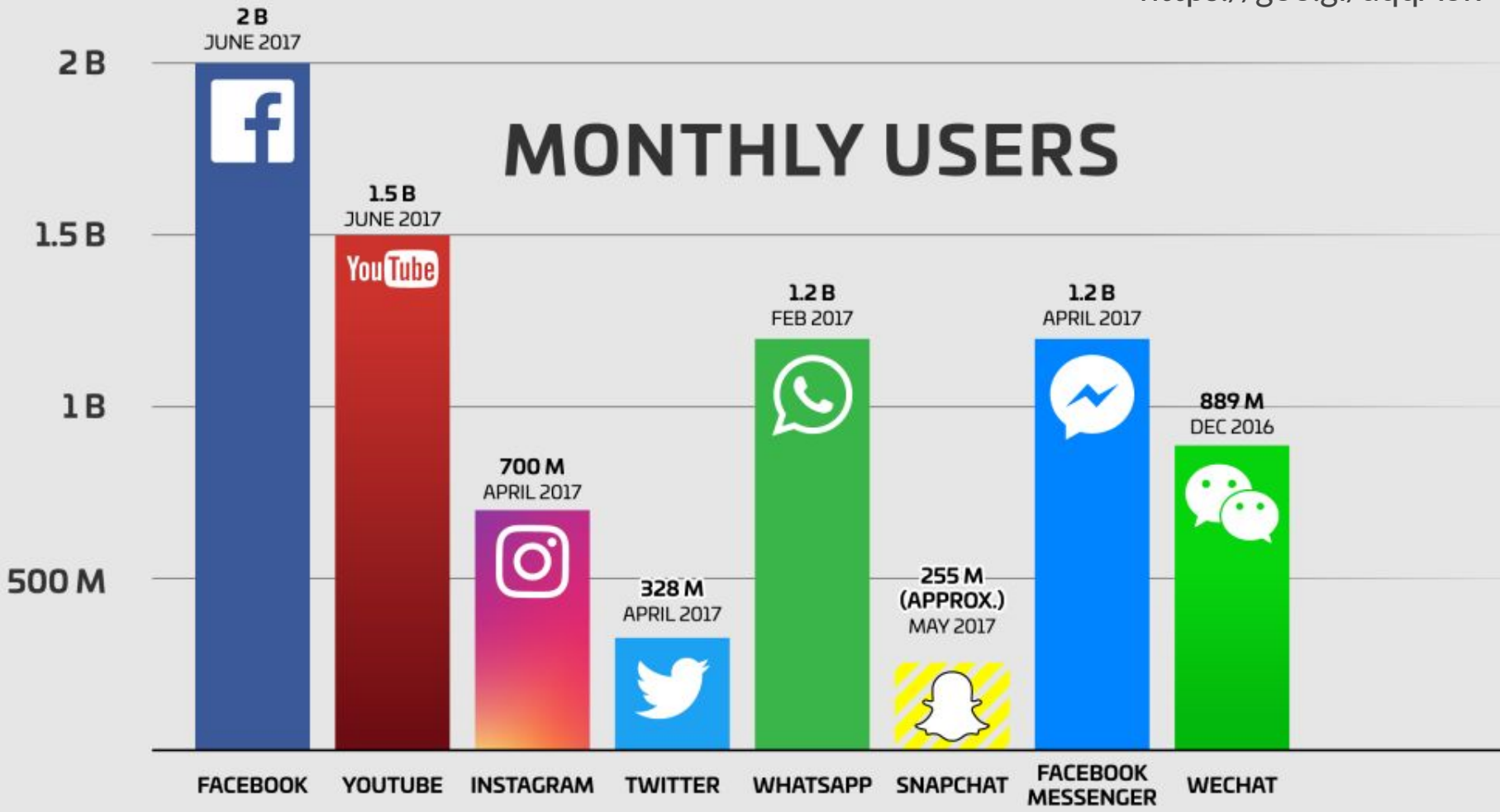
of the data will be free-form

IDC

THE COMING FLOOD OF DATA IN AUTONOMOUS VEHICLES



https://goo.gl/dqqAok

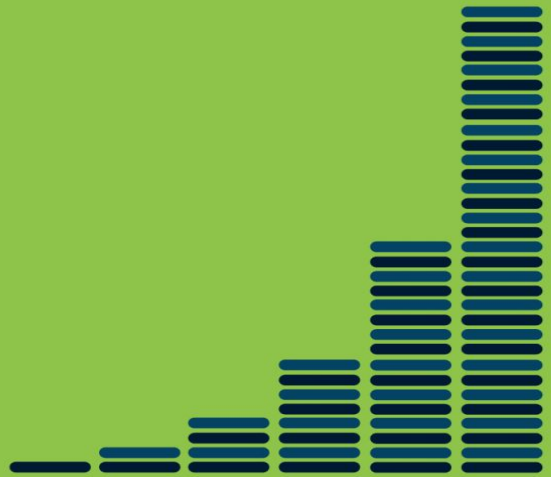


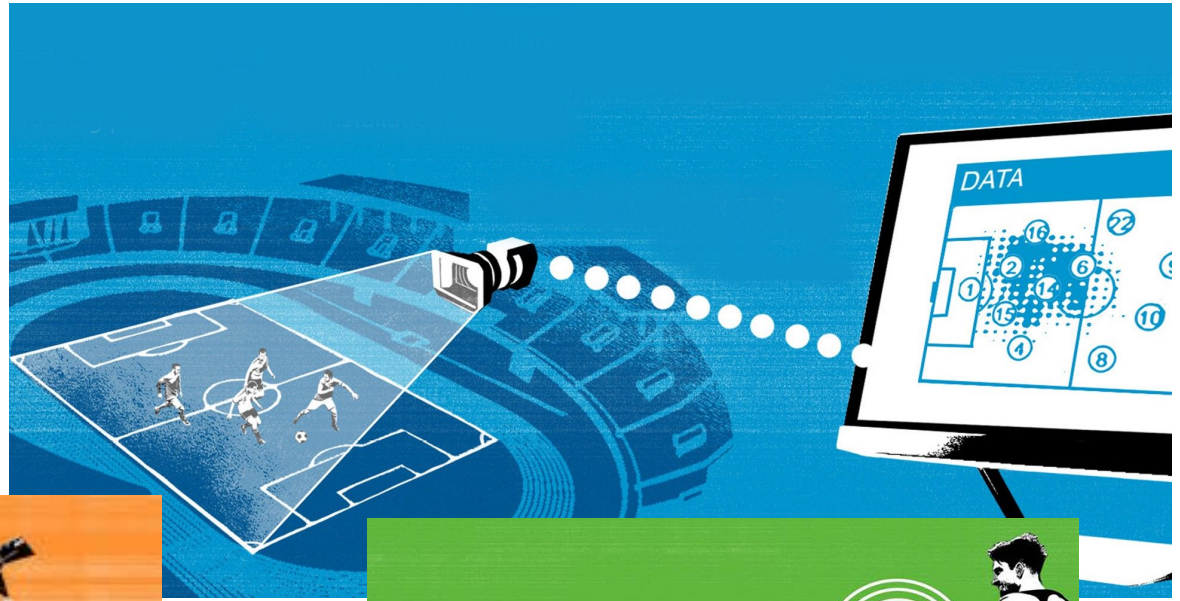

Extreme Large Telescope (ELT)
90 TB/night
Cerro Armazones, Chile

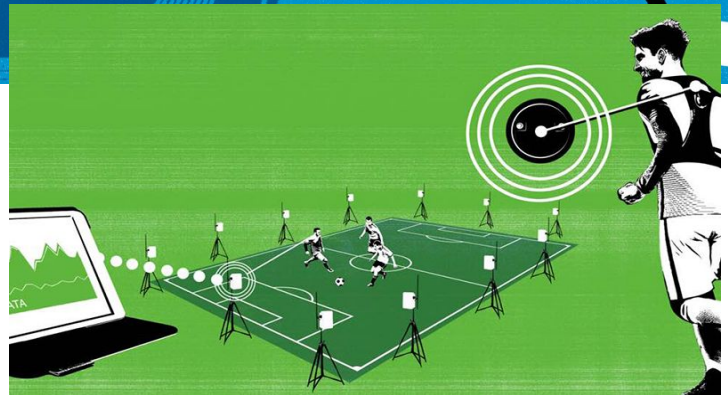
The average person is likely to generate more than one million gigabytes of health-related data in their lifetime. Equivalent to 300 million books.

IBM **Watson Health**

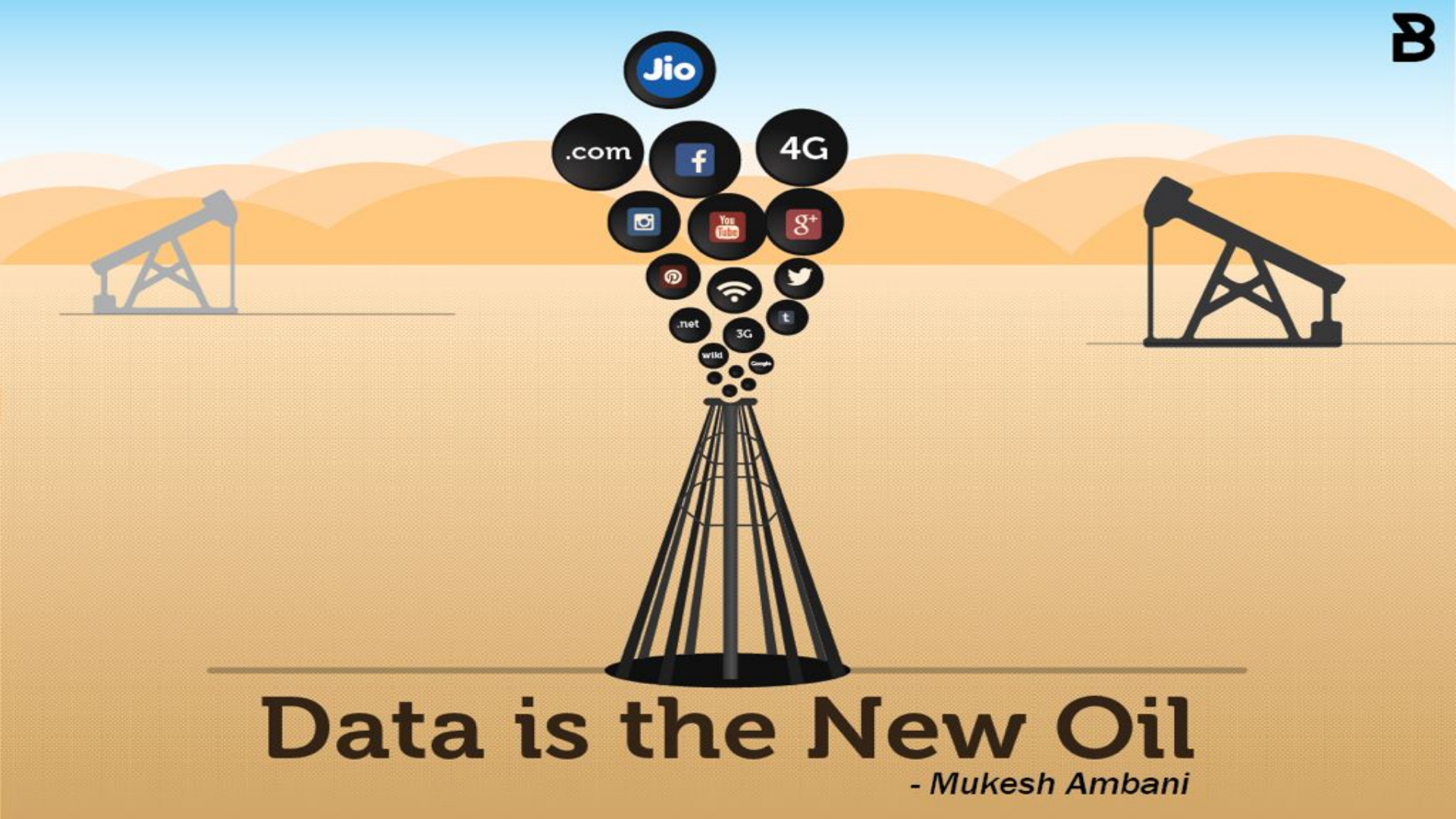
Medical data is expected to double every 73 days by 2020.





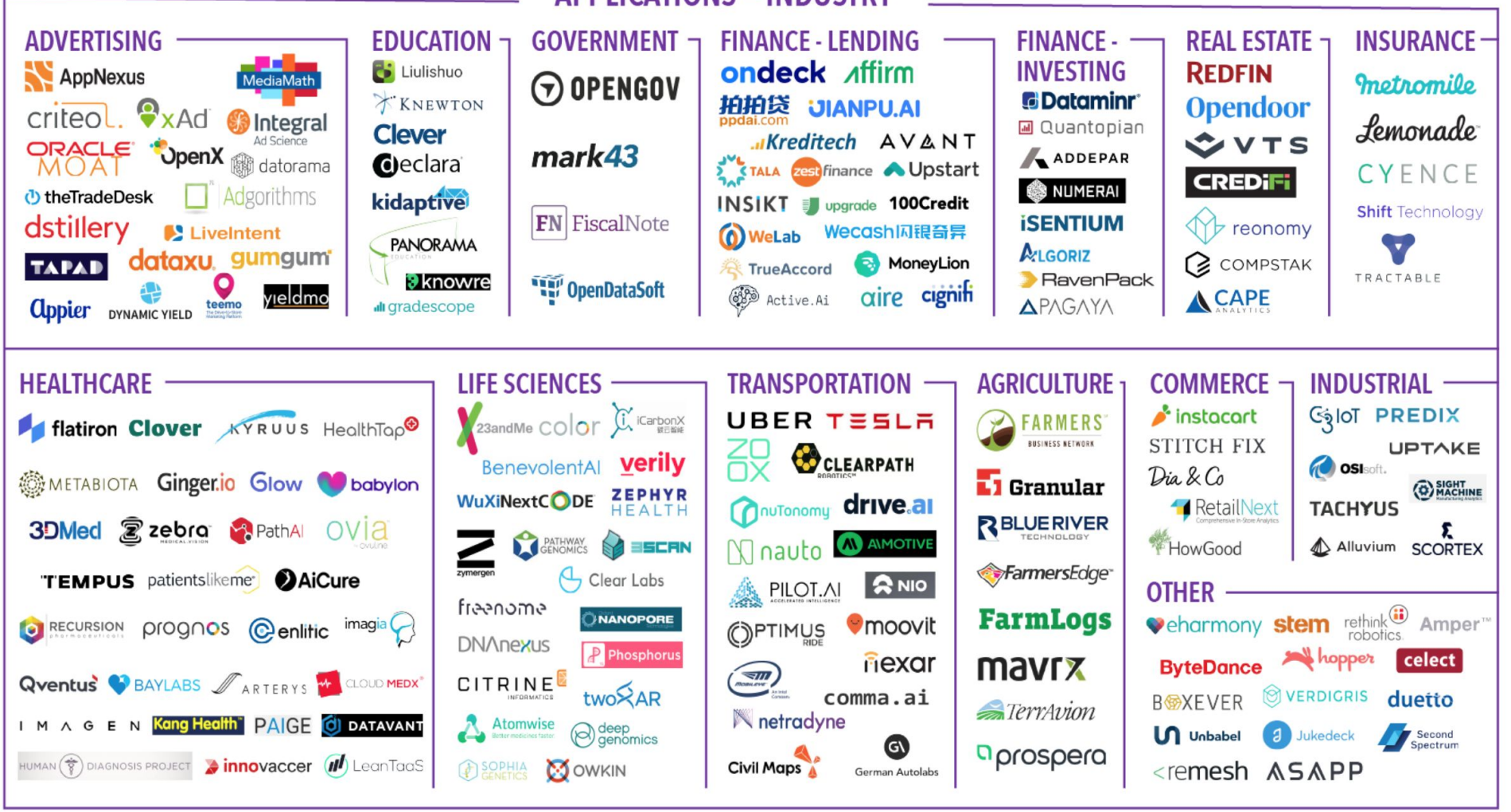


https://football-technology.fifa.com/en/media-tiles/epts
https://www.sporttechie.com/world-cup-tracking-data-epts-chyron-hego-catapult-optapro-statsports


Jio
.com
4G
.net
3G
wiki
Data is the New Oil
- Mukesh Ambani

APPLICATIONS – INDUSTRY

http://mattturck.com/bigdata2018/



Provocation #3

bias & privacy

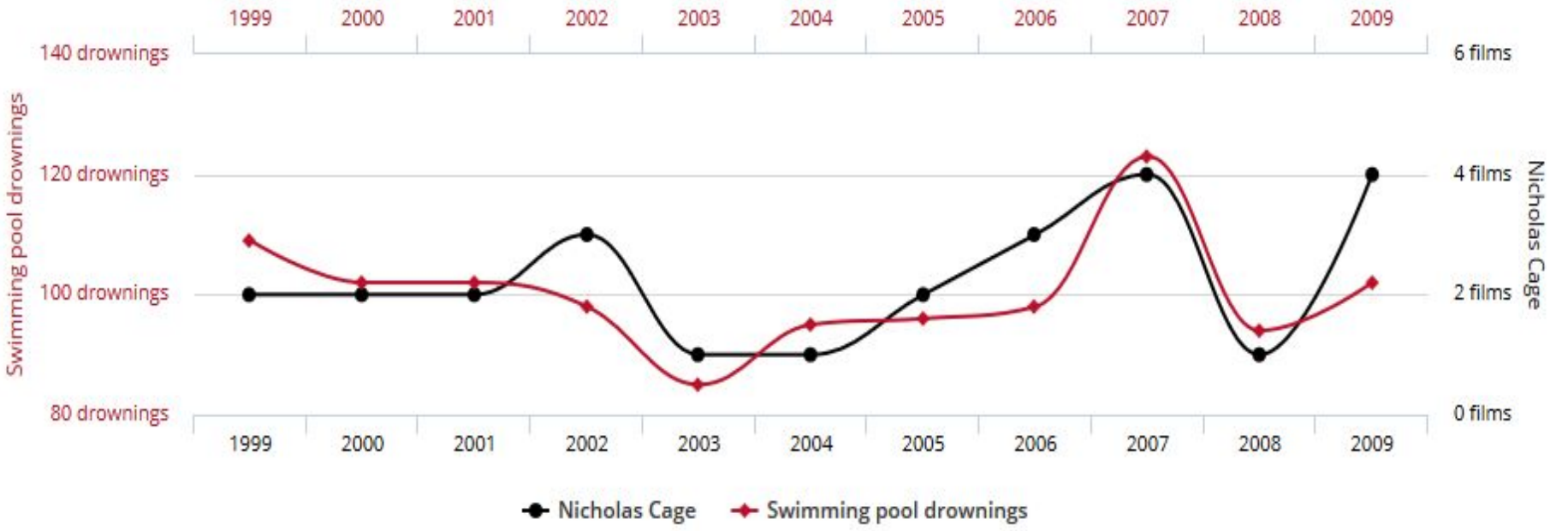

I got a big one this time!
BIG DATA
TRUTH
CORRELATION
NOISE
OVERFIT
@redpenblackpen

Number of people who drowned by falling into a pool

correlates wtih

Films Nicolas Cage appeared in

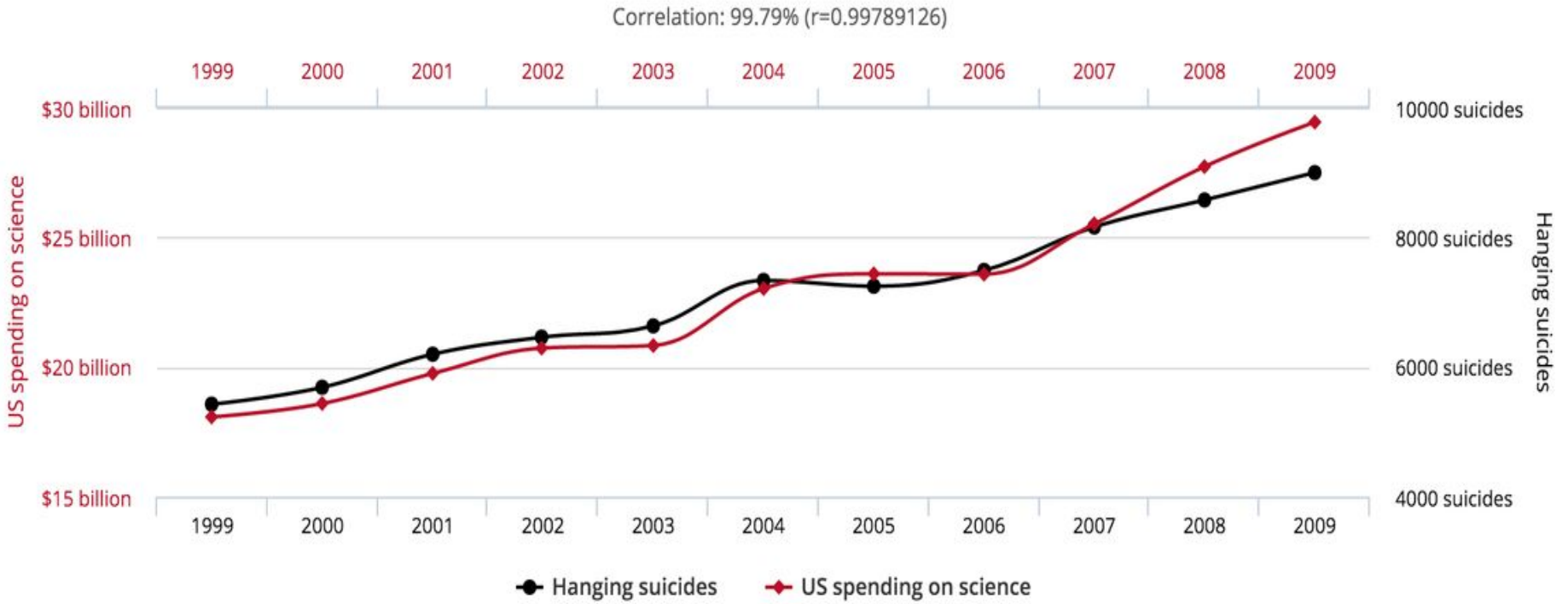
Correlation: 66.6% (r=0.666004, p>0.05)



tylervigen.com

Data sources: Centers for Disease Control & Prevention and Internet Movie Database

US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation

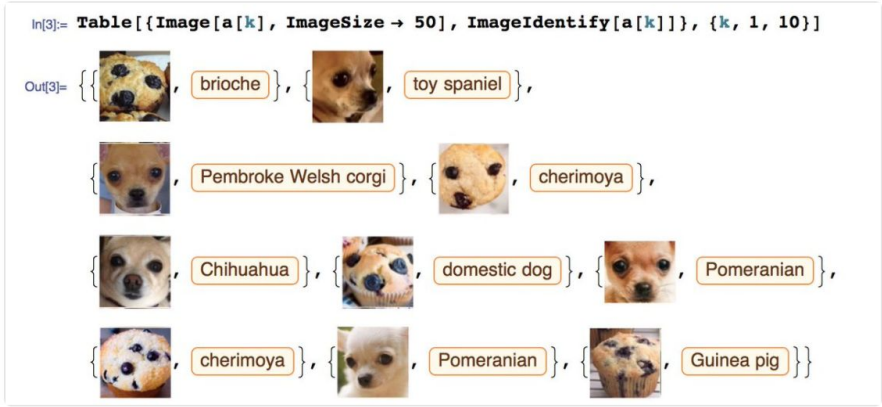


tylervigen.com

Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention

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

DATA VIOLENCE

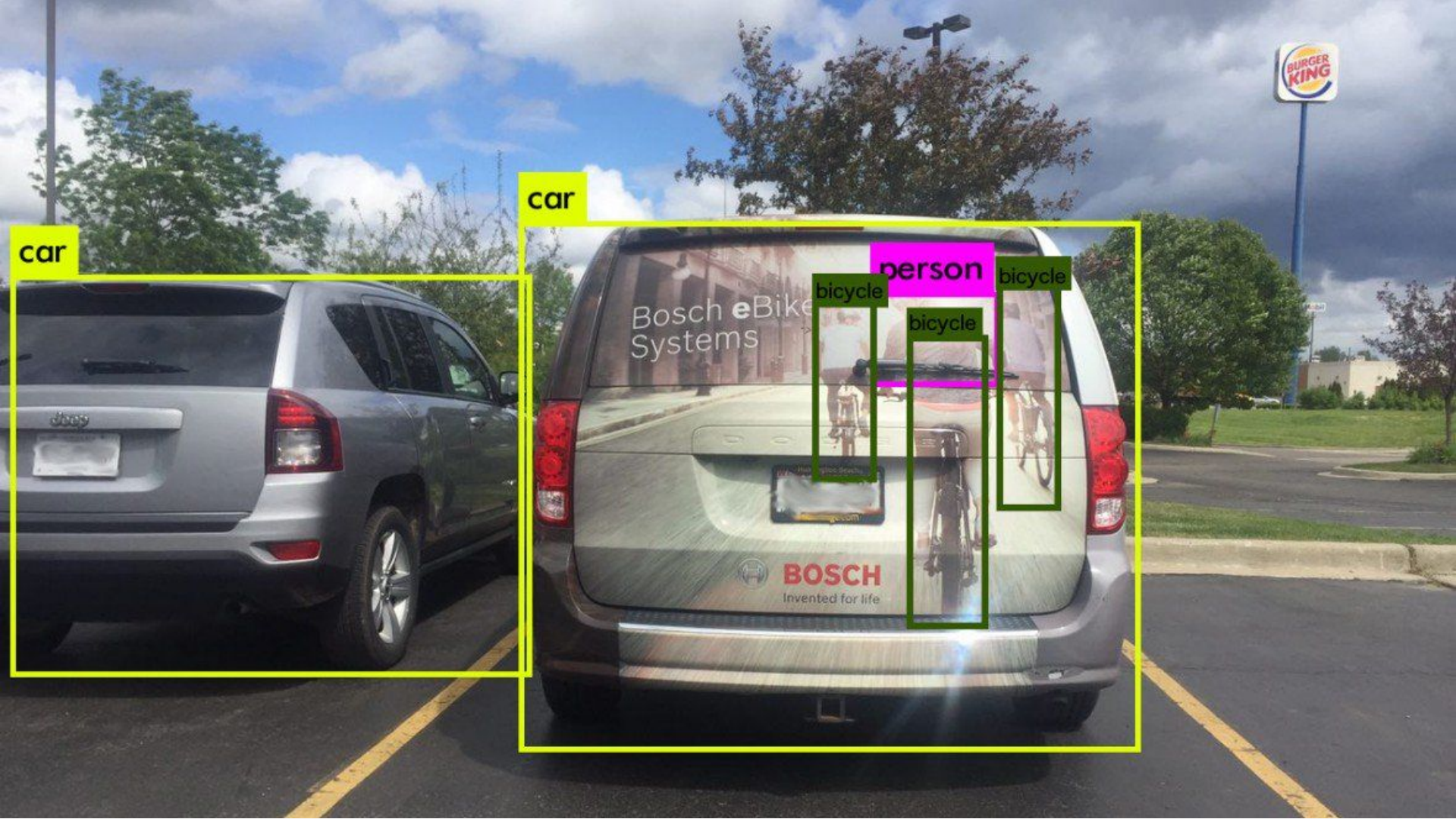
and how bad engineering choices can damage society

Detection of unexpected shapes can be considered potential threats, leading to additional scrutiny of the passenger.

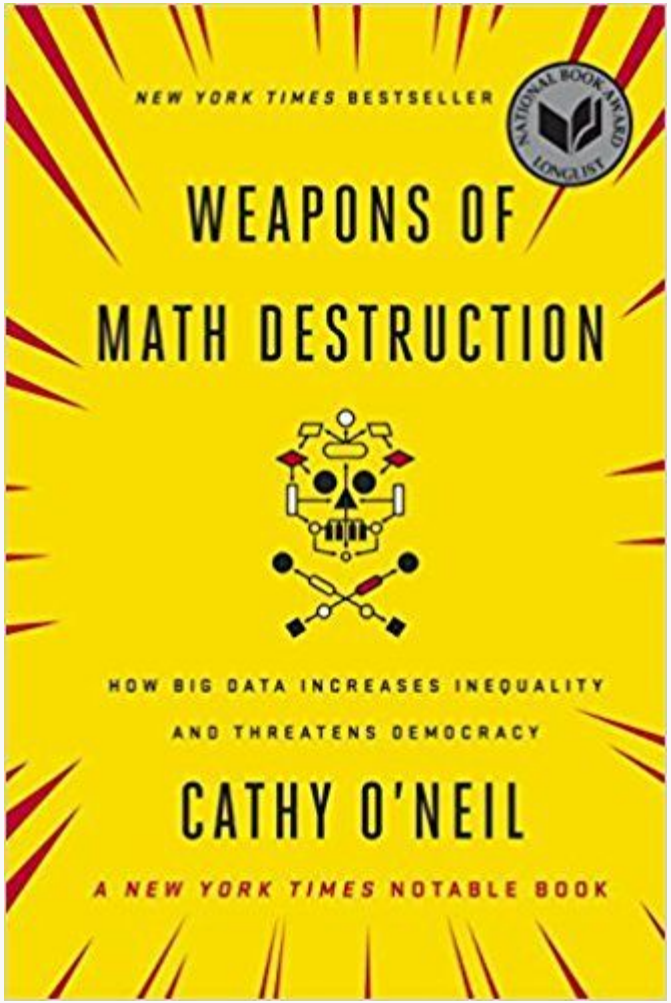
https://medium.com/s/story/data-violence-and-how-bad-engineering-choices-can-damage-society-39e44150e1d4

https://goo.gl/NwP7Fv

https://goo.gl/xzLxaY


car
car
person
bicycle
bicycle
bicycle
Bosch eBike Systems
BOSCH
Invented for life
BURGER KING

“People keep suggesting that **democracy** is alive and well because we have two **parties that don’t agree on everything**. I think that’s total bullshit.”



Math can be manipulated by biases and affect every aspect of our lives.

Big Data

http://www.showmetech.com.br/big-data-trump/

Michal Kosinski - the father of the system, which deals with data processing.

68 likes
- user's race (96%)
- sexual orientation (89%)
- political affiliation (85%)

150 likes
- ++ family member

300 likes
- ++ spouse

https://applymagicsauce.com/


http://bit.do/fakenews100k
FAKE NEWS
Ethics


"Of all the content on Facebook, more than 99% of what people see is authentic. Only a very small amount is fake news and hoaxes"
SHURE

Fake News Is A Real Problem

Facebook engagement of the top five fake election stories*

* Engagement is measured as total number of shares, reactions and comments

@StatistaCharts Source: Buzzsumo via Buzzfeed

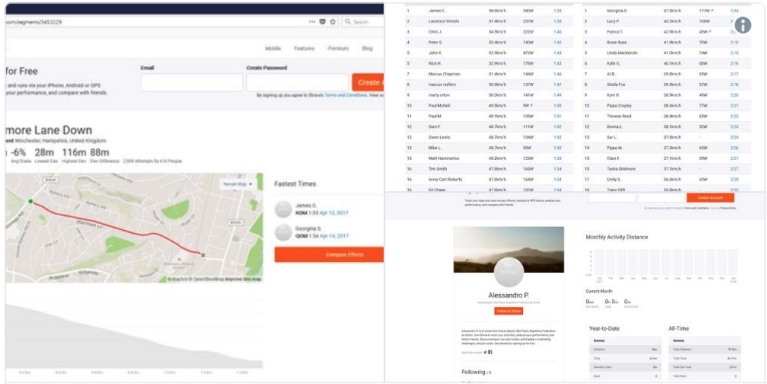
statista

Nathan Ruser
@Nrg8000

Strava released their global heatmap. 13 trillion GPS points from their users (turning off data sharing is an option). medium.com/strava-enginee… … It looks very pretty, but not amazing for Op-Sec. US Bases are clearly identifiable and mappable

3:24 PM - Jan 27, 2018

2,679 2,445 people are talking about this



Paul D
@Paulmd199

It just keeps getting deeper. You can also trivially scrape segments, to get a list of people who travelled a route, and trivially obtain a list of users. #Strava

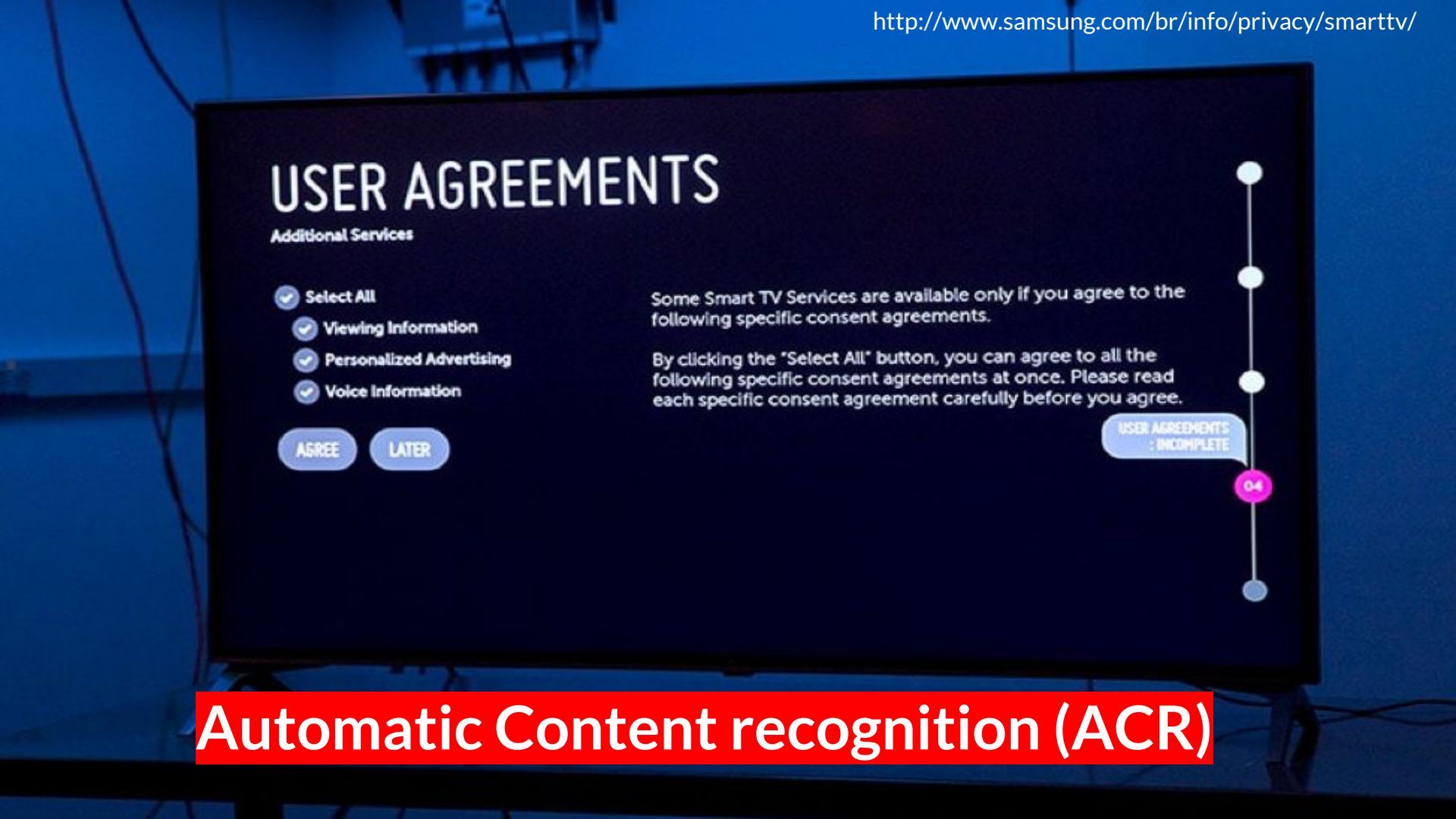
6:51 PM - Jan 28, 2018

380 316 people are talking about this

Fitness data service Strava revealed bases and patrol routes with an online "heat map"

https://www.wired.com/story/strava-heat-map-military-bases-fitness-trackers-privacy/

http://www.samsung.com/br/info/privacy/smarttv/



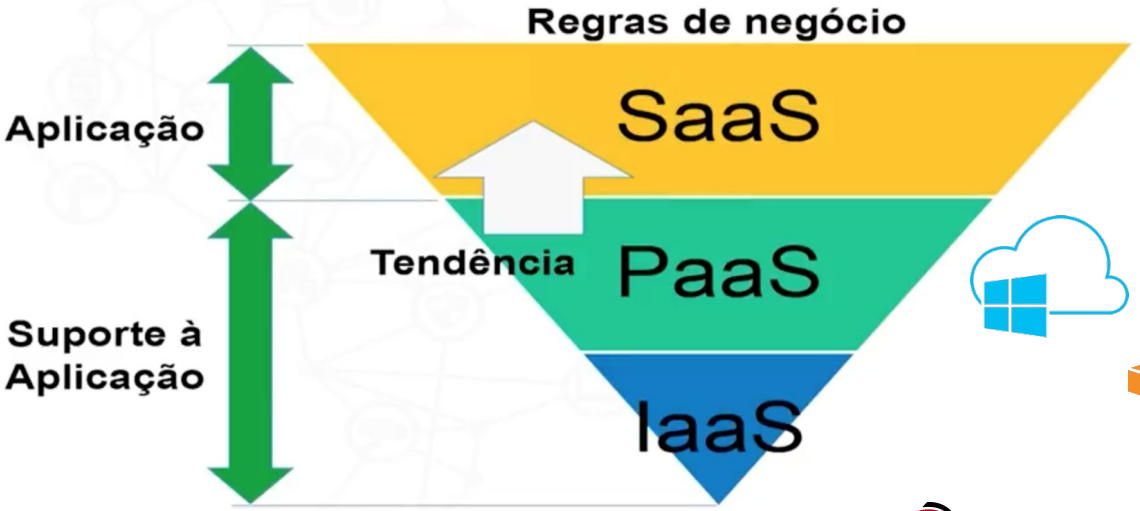
Automatic Content recognition (ACR)


He's not your dad.
My dad told me you're spying on us.
9GAG

Provocation #4

cloud computing

#cloudcomputing



MLaaS
Machine Learning as a Service

Alibaba Cloud

SAP

ORACLE

amazon web services

SIEMENS

SOFTLAYER

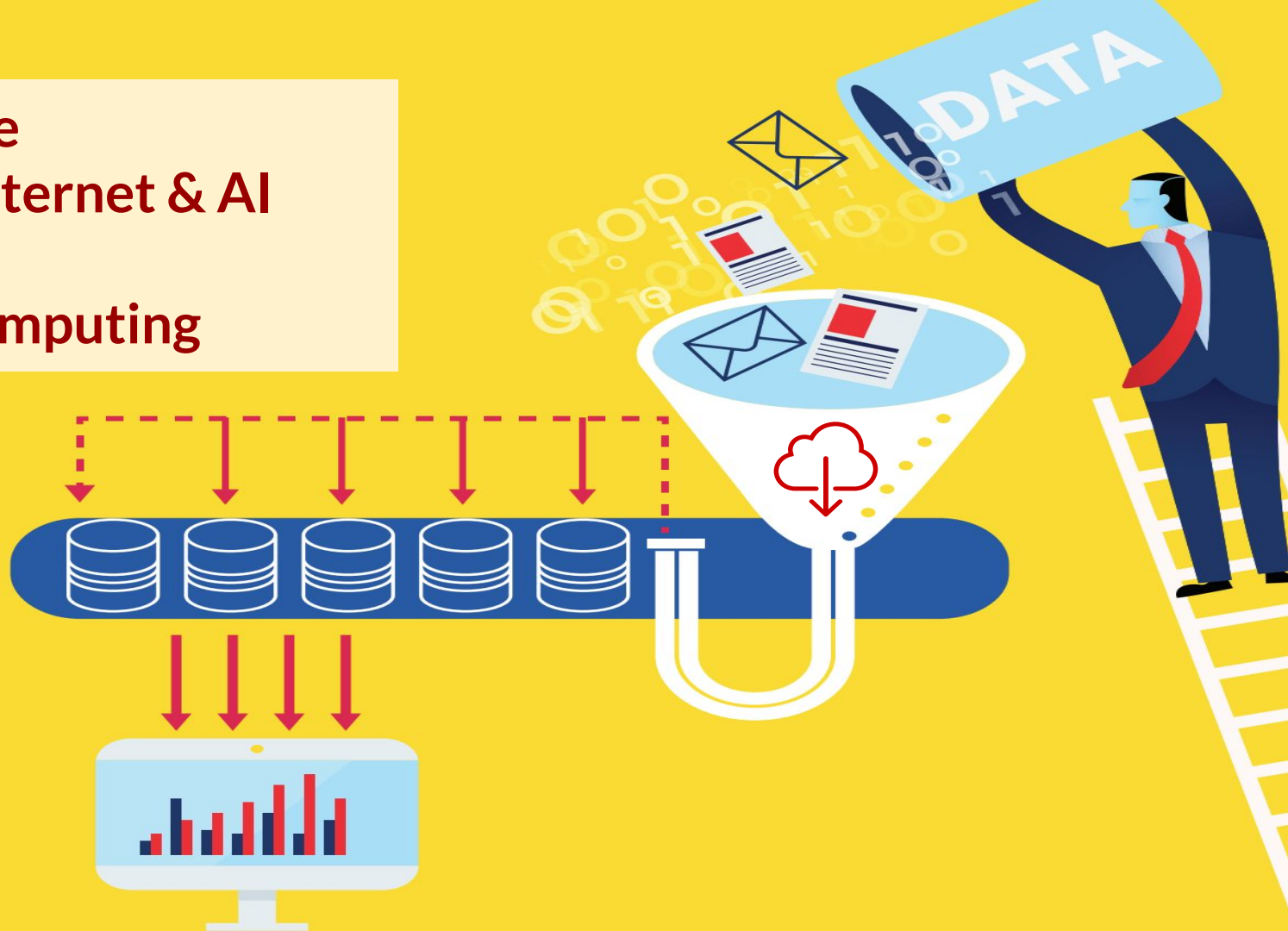
linode

DigitalOcean

ORACLE
Cloud Infrastructure

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

Hardware
Data & Internet & AI
Bias
Cloud Computing



Who studies this stuff?

A Data-Driven Program

Syllabus - EEC0905 Data Science I

1. Data Science Foundation;
2. Workflow: collect, pre-processing, hygienization, analysis, visualization;
3. Clustering and network analysis.

Course Planning Timeline

Part one

- Fundamentals of python for data science (basic, numpy, pandas)

Part two

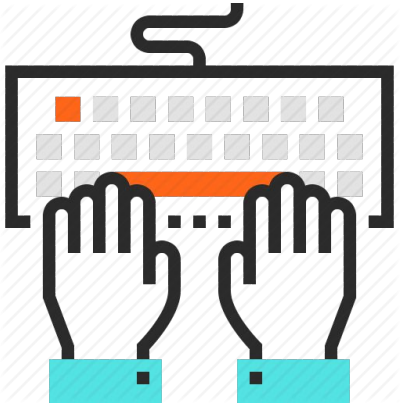
- Cleaning, large dataset, visualization (matplotlib, seaborn, plotly, maps, choropleth)

Part three

- Import data from web (web scraping), network analysis

Learning by doing

#no_exams #assessments_all_week #projects
#dont_reinvent_the_wheel

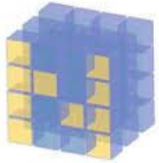


python

colab

ANACONDA
Powered by Continuum Analytics®

git



NumPy

pandas

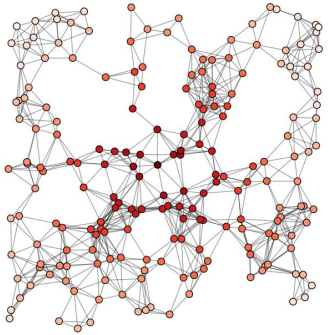
$$y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$$

Keras

TensorFlow

NetworkX

PyGSP



matplotlib

Folium

Seaborn

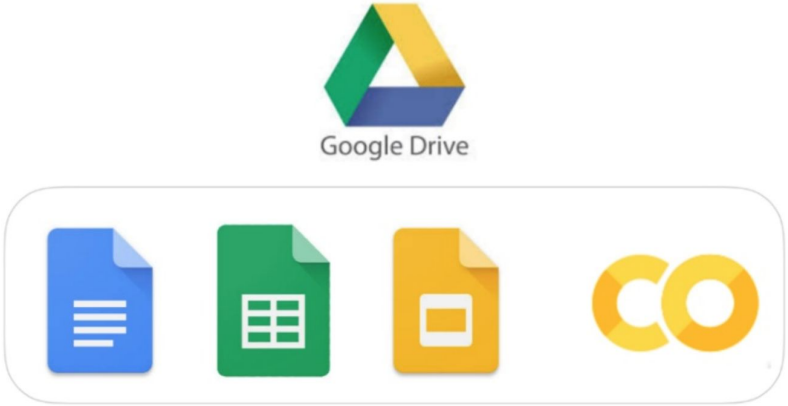
Requests

bokeh

Beautiful Soup

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.

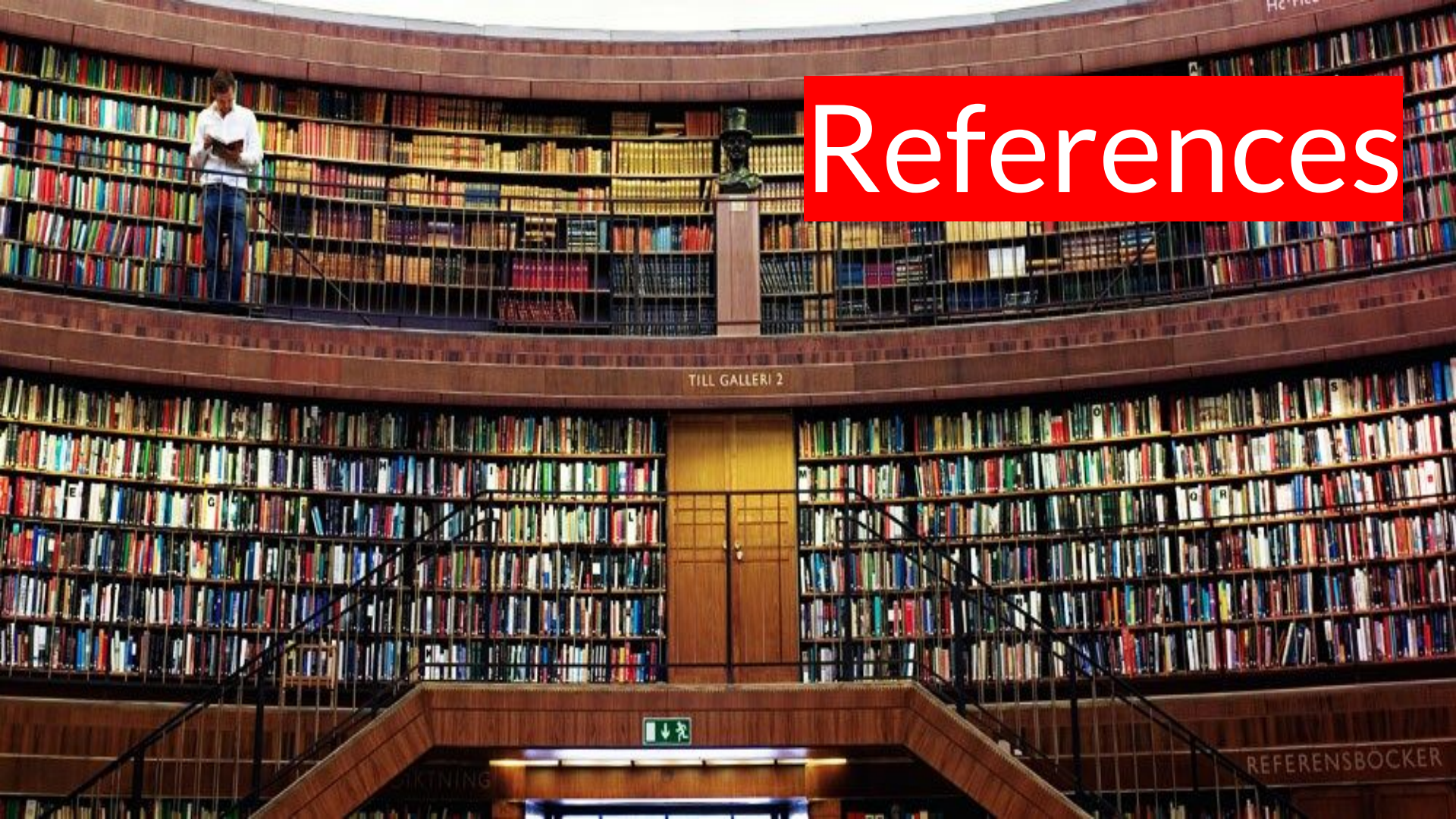
git

Build a Portfolio

GitLab

GitHub

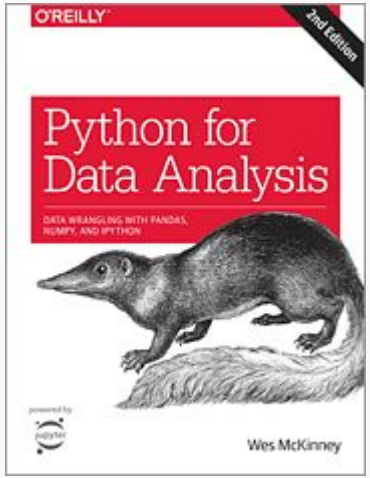
Bitbucket

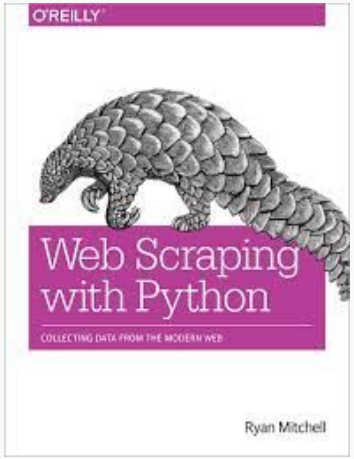


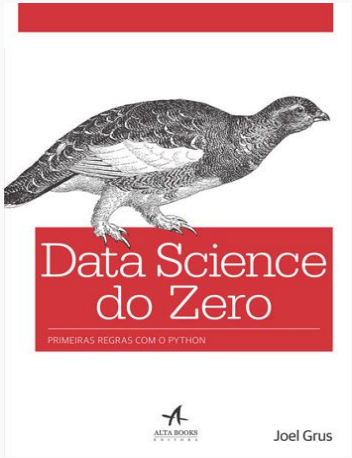
References

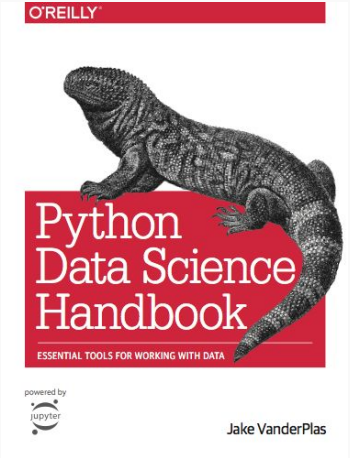
References Online

References


O'REILLY
2nd Edition
Python for Data Analysis
DATA WRANGLING WITH PANDAS, NUMPY, AND IPYTHON
powered by jupyter
Wes McKinney


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


Data Science do Zero
PRIMEIRAS REGRAS COM O PYTHON
ALTA BOOKS
Joel Grus


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