Data Science & ML Course

Lesson #1 - Outline & Directions

Ivanovitch Silva

September, 2018


Introduction

Ivanovitch Silva (ivan@imd.ufrn.br)

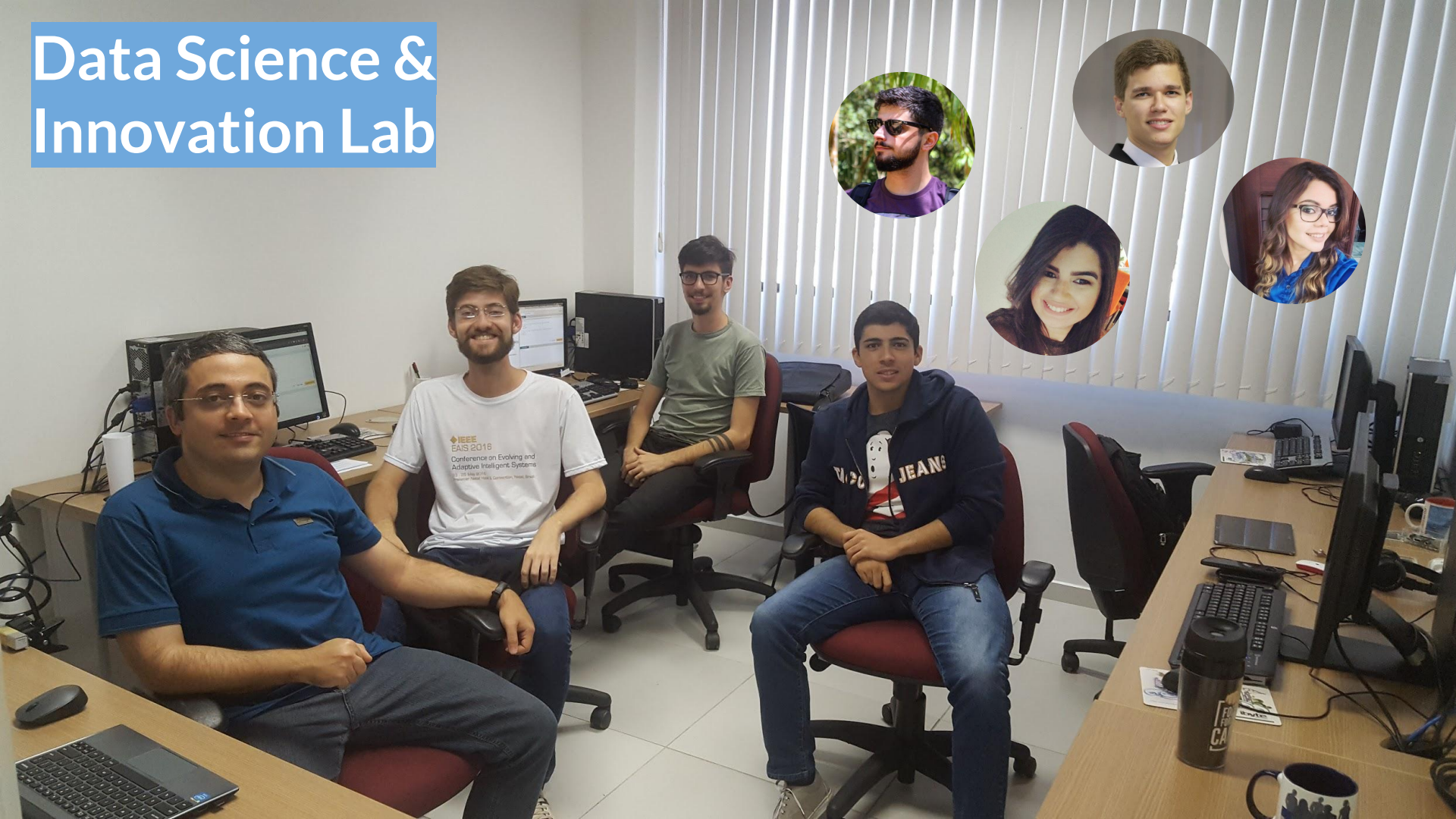
GitHub/ivanovitchm

Roadmap history of research activities

Deployments of research products

13	Refinarias	3	Energia	50	Plataformas
123	Exploração	5	Transporte	10	Outros

Data Science & Innovation Lab



MASSACHUSETTS INSTITUTE OF TECHNOLOGY

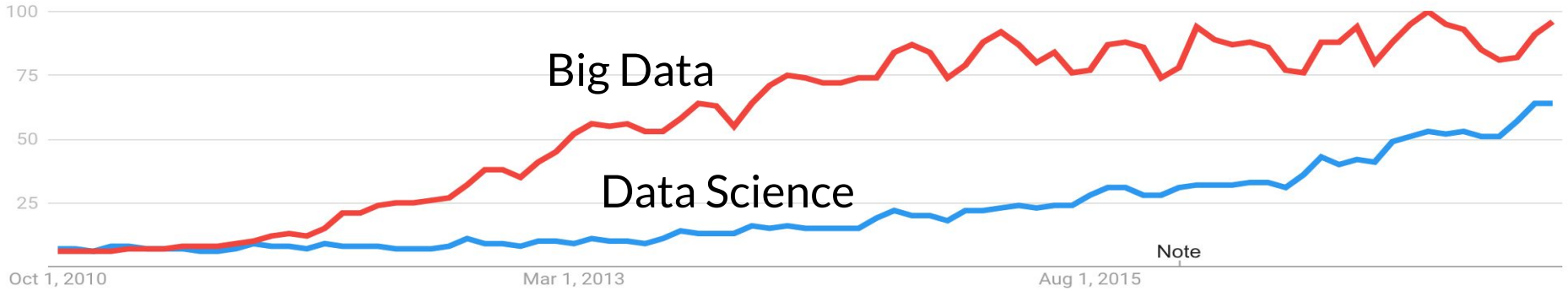
Big Data and Social Analytics certificate course

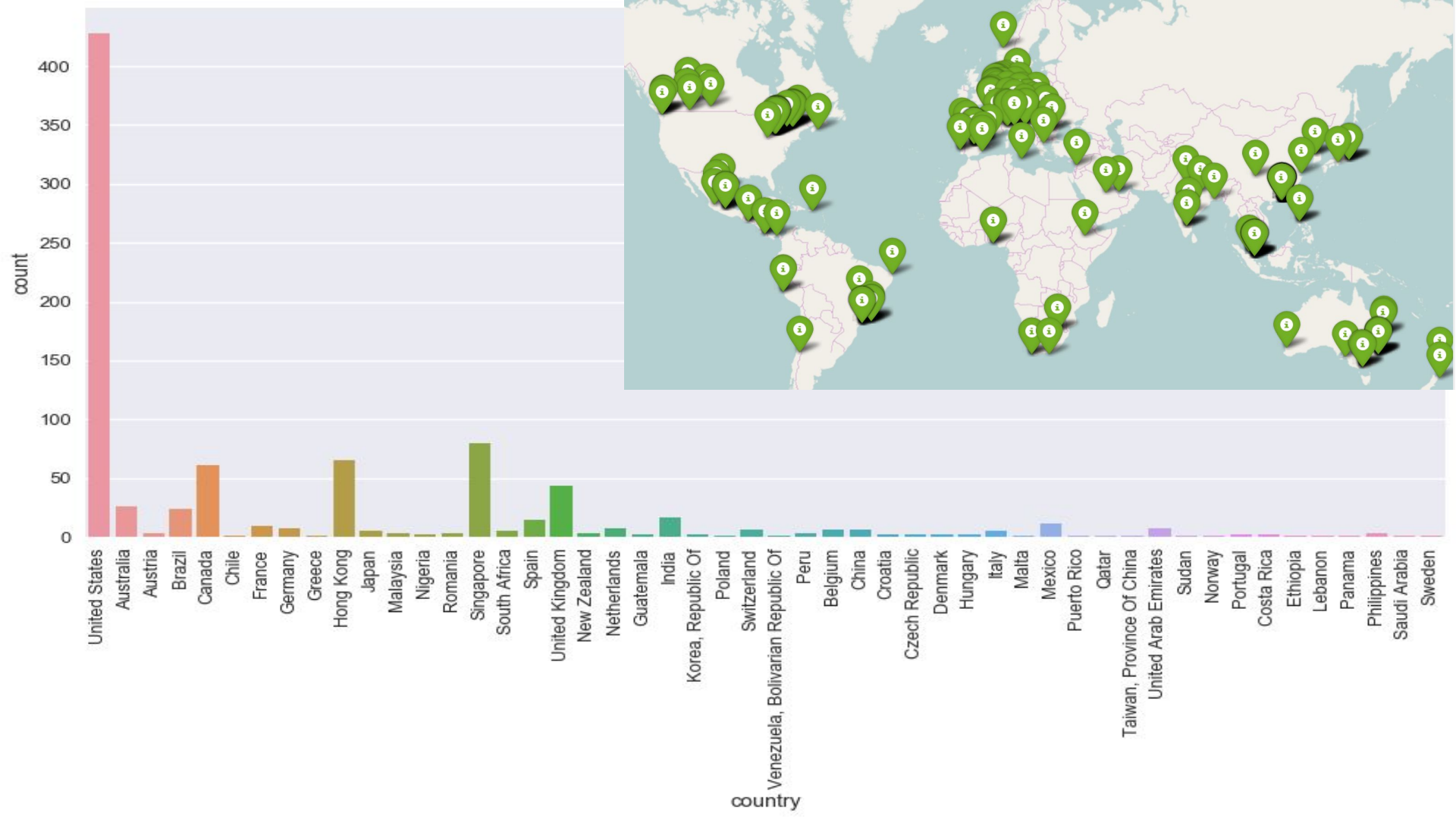
EXperimental Learning

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

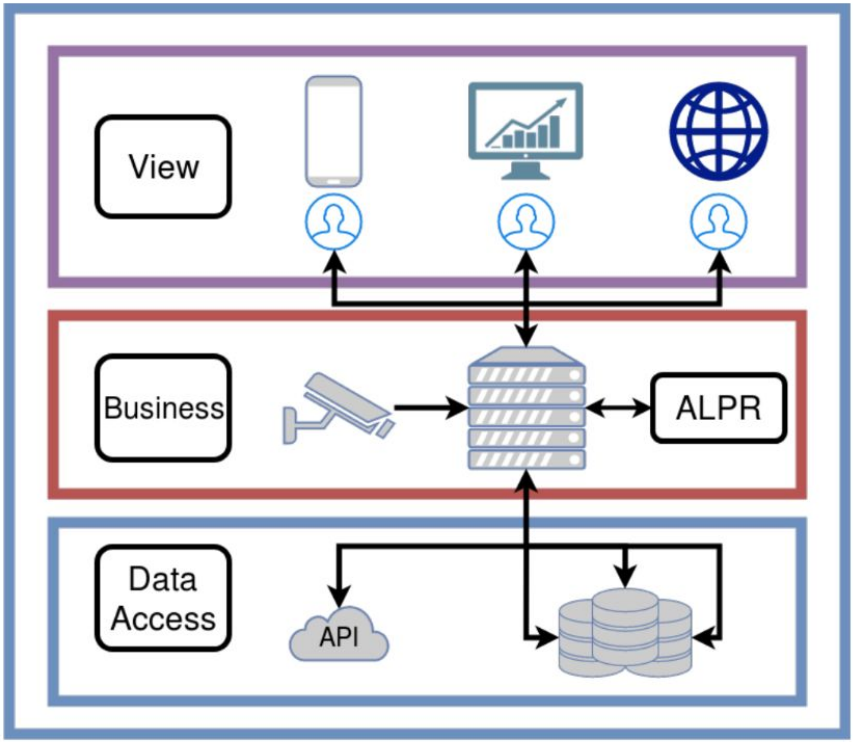
Undergratuate
2017.1 - IMD0105 Introduction to Data Science
2017.2 - IMD0252 Learning Analytics
2017.2 - DCA0046 Data Science
2018.1 - DCA0132 Data Engineering
2018.2 - IMD0905 Data Science
2018.2 - DCA0131 Data Science

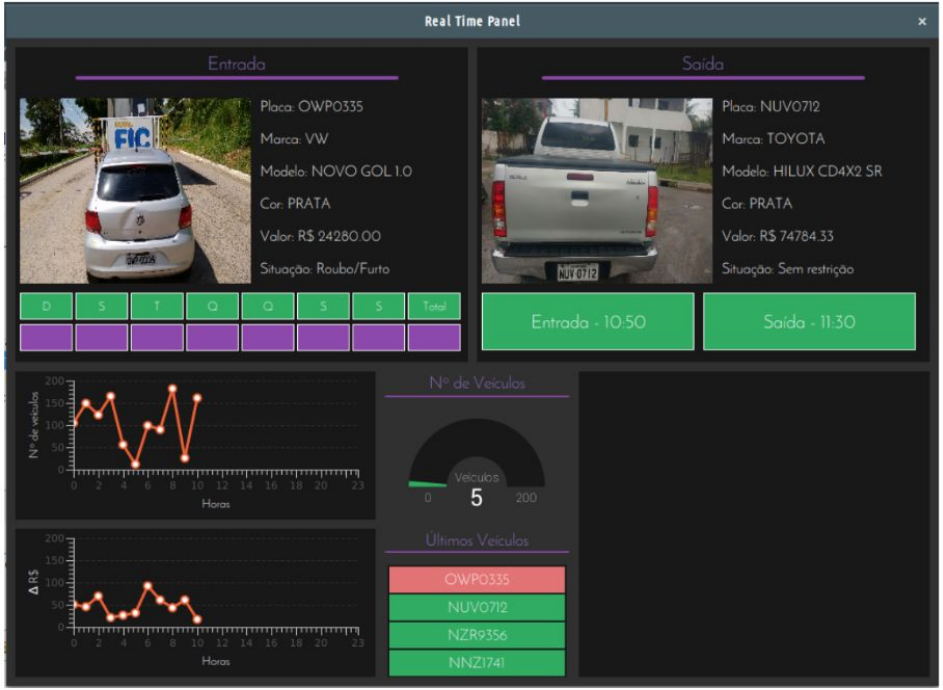
Graduate
2017.2 - EEC2006 Data Science Foundations
2017.2 - ITE0021 Learning Analytics
2018.2 - EEC1509 Machine Learning

DataCamp

ALPR Mobile

GOOSE


View
Business
ALPR
Data Access
API


Real Time Panel
Entrada
Placa: OWP0335
Marca: VW
Modelo: NOVO GOL 1.0
Cor: PRATA
Valor: R$ 24280.00
Situação: Roubo/Furto
Saída
Placa: NUV0712
Marca: TOYOTA
Modelo: HILUX CD4X2 SR
Cor: PRATA
Valor: R$ 74784.33
Situação: Sem restrição
D
S
T
Q
Q
S
S
Total
Entrada - 10:50
Saída - 11:30
Nº de veículos
Horas
Nº de Veículos
Veículos
0
5
200
Δ R$
Horas
Últimos Veículos
OWP0335
NUV0712
NZR9356
NNZ1741

Multispectral Satellite
Landsat 5

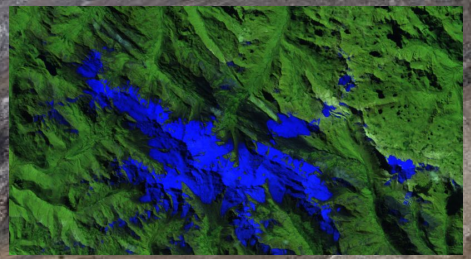
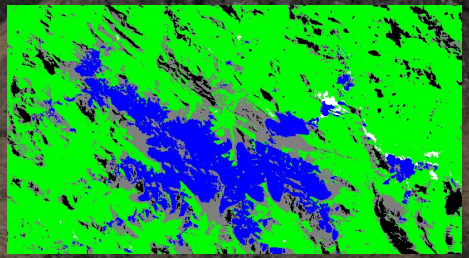


Image captured by Landsat



Output of neural network

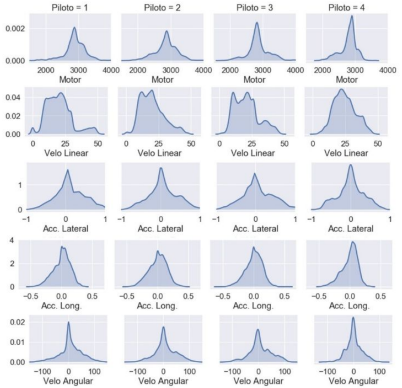

Protocolo de Autorização
05/02/2015 10:58:16

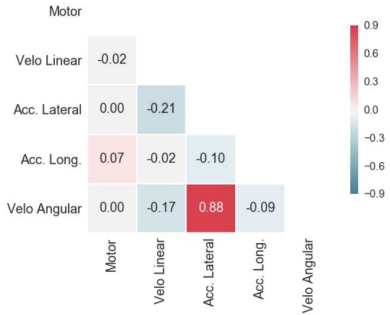

OBD-II
https://github.com/eltonvs/java-obd-api
https://github.com/ivanovitchm/iot_hands_on

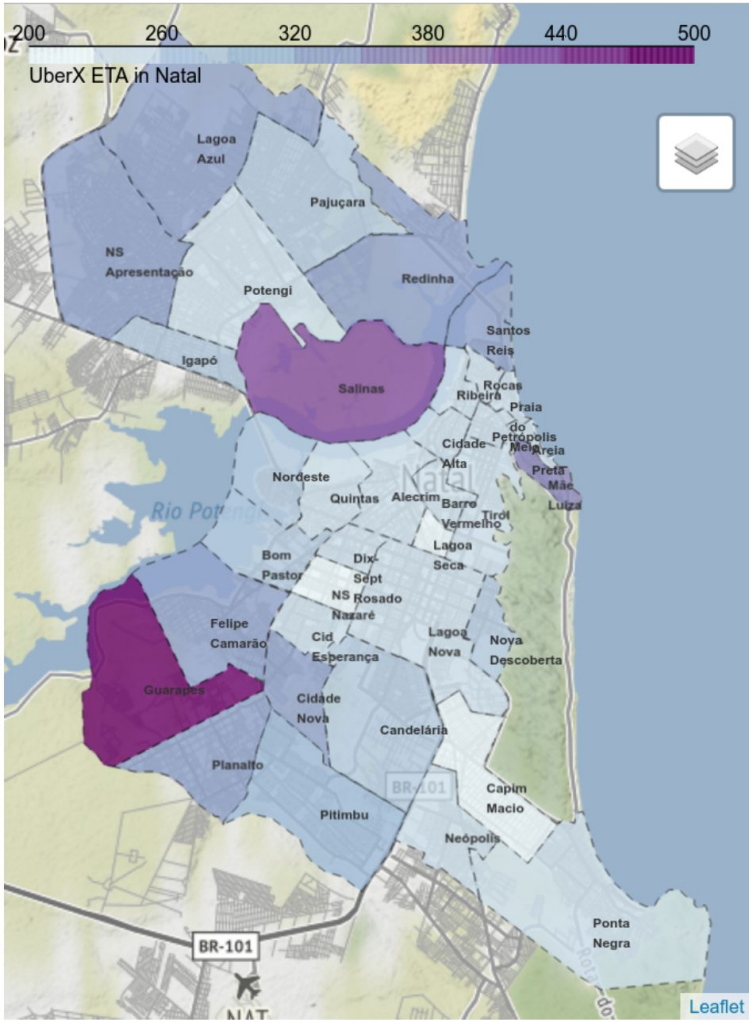
UNIVERSITAS STUDIORUM BRIXIAE
UNIVERSITY OF BRESCIA

UFRN
UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE


02
Baja
2010
OFFSHOX
AR-KARA
UFRN


Piloto = 1
Piloto = 2
Piloto = 3
Piloto = 4
Motor
Velo Linear
Acc. Lateral
Acc. Long.
Velo Angular


Motor
Velo Linear
Acc. Lateral
Acc. Long.
Velo Angular
-0.02
0.00 -0.21
0.07 -0.02 -0.10
0.00 -0.17 0.88 -0.09
0.9
0.6
0.3
0.0
−0.3
−0.6
−0.9



Query Parameters

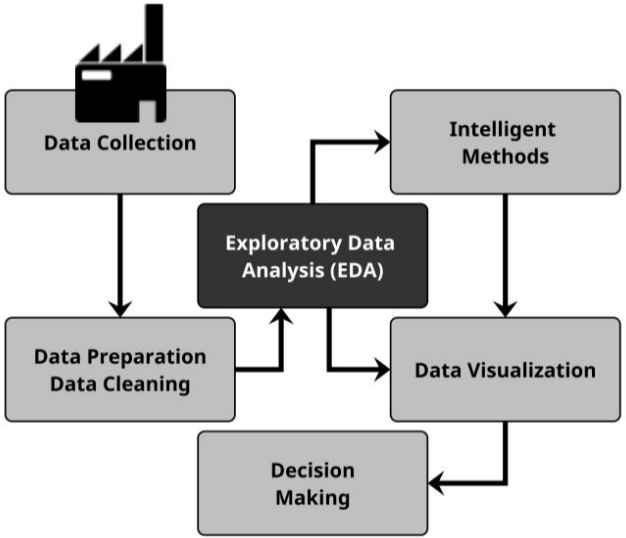
	Name	Description
float	**start_latitude**	Latitude component
float	**start_longitude**	Longitude component
string	**product_id**	Unique identifier representing a specific product for a given latitude & longitude

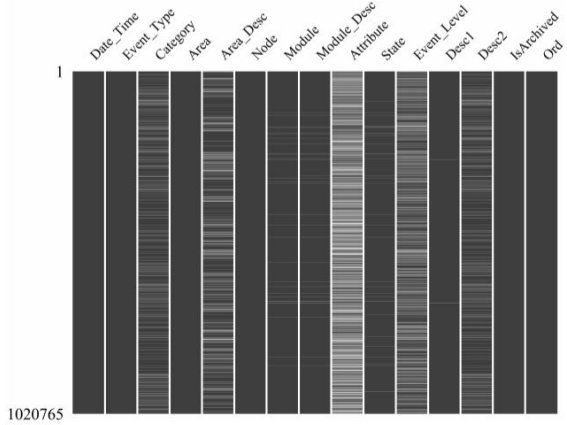
Ride Request API
GET /estimates/time

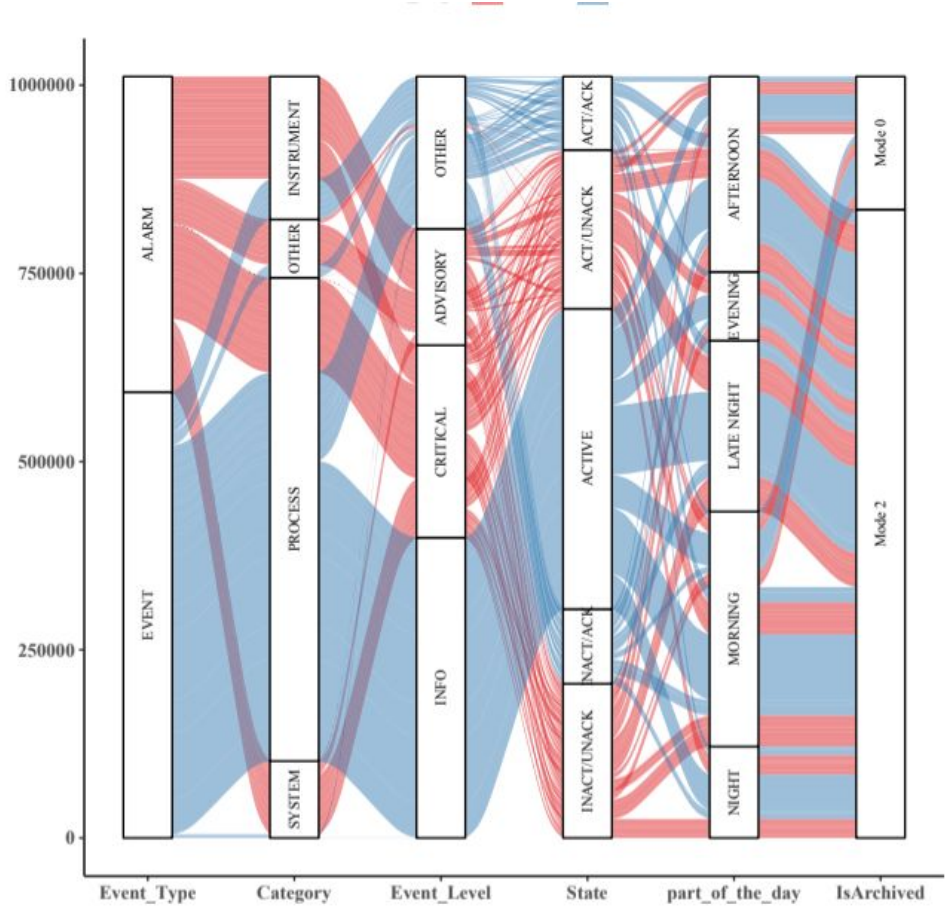
Response

	Name	Description
string	**product_id**	Unique identifier representing a specific product for a given latitude & longitude
string	**localized_display_name**	Localized display name of product
string	**display_name**	Display name of product
int	**estimate**	ETA for the product (in seconds)

UBER


Data Collection
Intelligent Methods
Exploratory Data Analysis (EDA)
Data Preparation Data Cleaning
Data Visualization
Decision Making


Date_Time
Event_Type
Category
Area
Area_Desc
Node
Module
Module_Desc
Attribute
State
Event_Level
Desc1
Desc2
IsArchived
Ord
1
1020765


1000000
750000
500000
250000
0
ALARM
EVENT
INSTRUMENT
OTHER
PROCESS
SYSTEM
OTHER
ADVISORY
CRITICAL
INFO
ACT/ACK
ACT/UNACK
ACTIVE
INACT/ACK
INACT/UNACK
AFTERNOON
EVENING
LATE NIGHT
MORNING
NIGHT
Mode 0
Mode 2
Event_Type
Category
Event_Level
State
part_of_the_day
IsArchived

UFRN Analytics

Este repositório é dedicado à análise de dados do Portal de Dados Abertos da UFRN, que reune dados de diversas áreas da Universidade com o intuito de prover transparência às ações da instituição e incentivar o desenvolvimento de aplicações que contribuam com a gestão universitária.

Servidores
Ensino
Pesquisa
Extensão

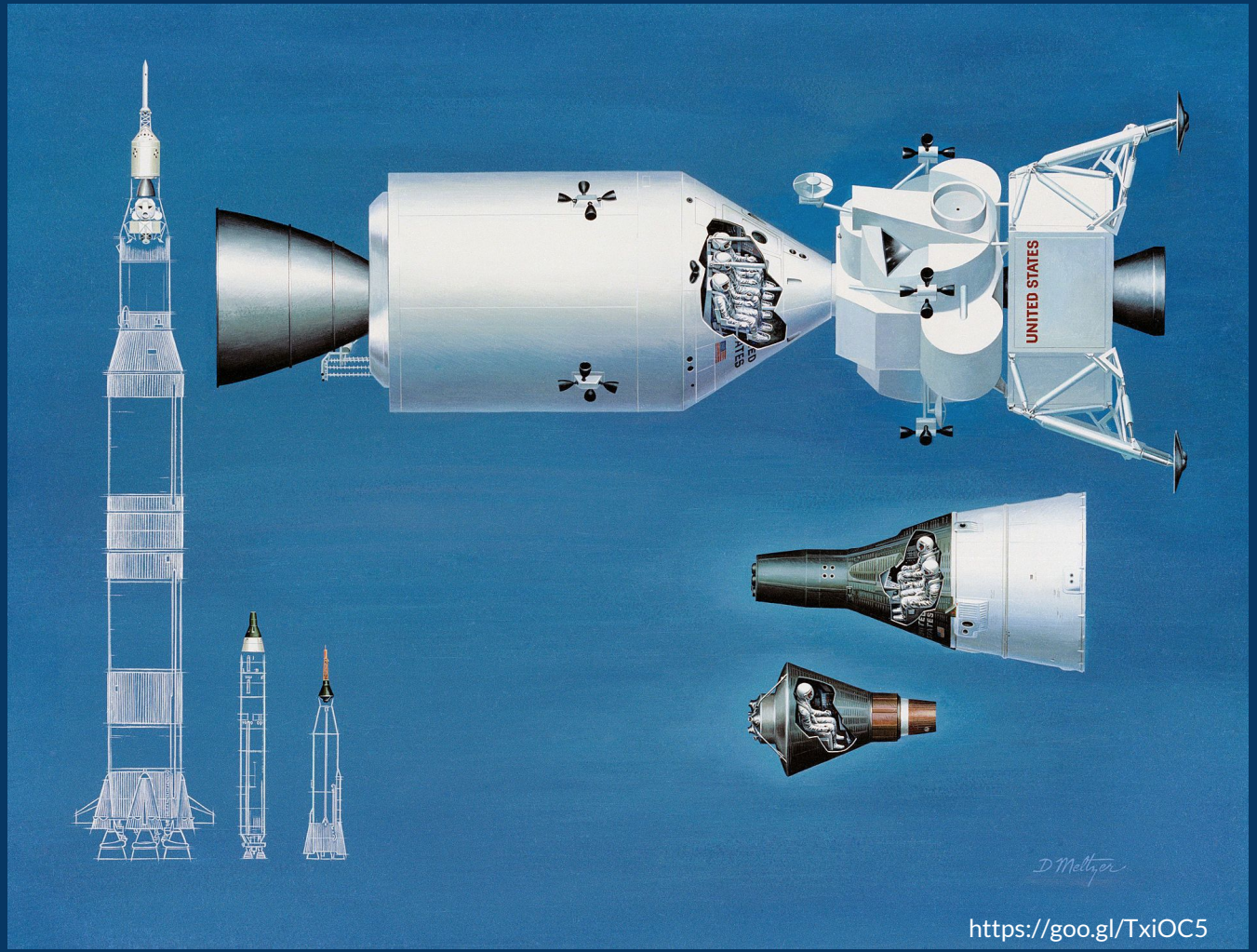
https://github.com/ycaroravel/UFRN-Analytics

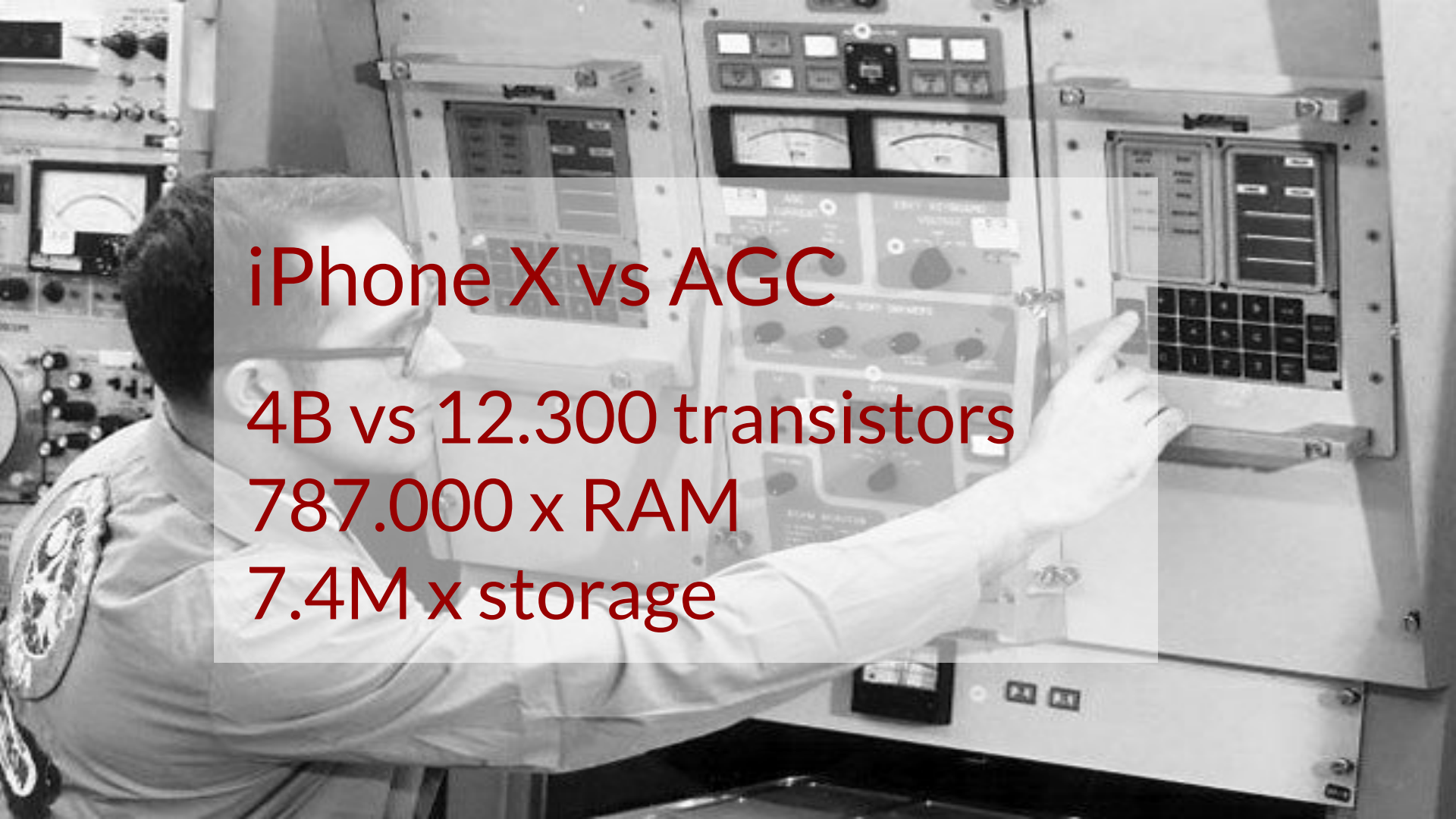
Group Survey

Provocation #1

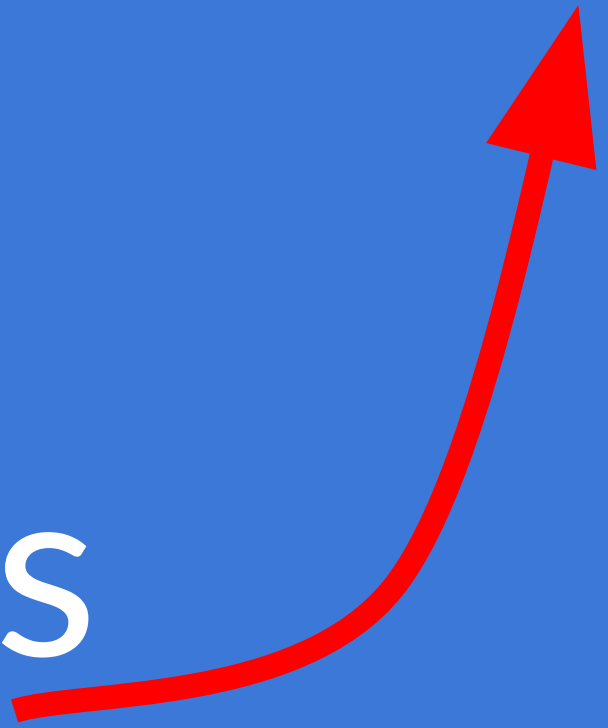
hardware

1969

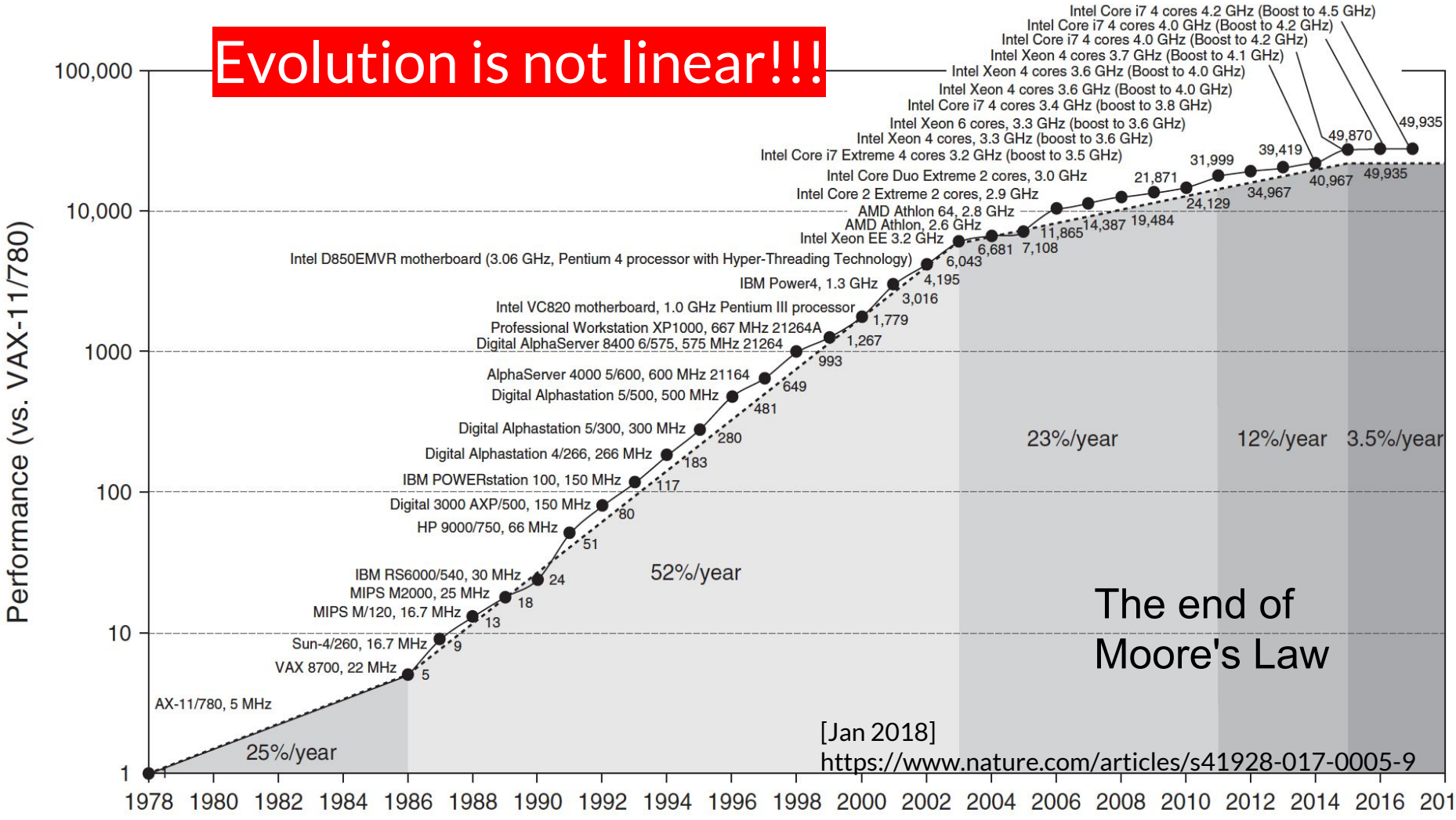

UNITED STATES
https://goo.gl/TxiOC5


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

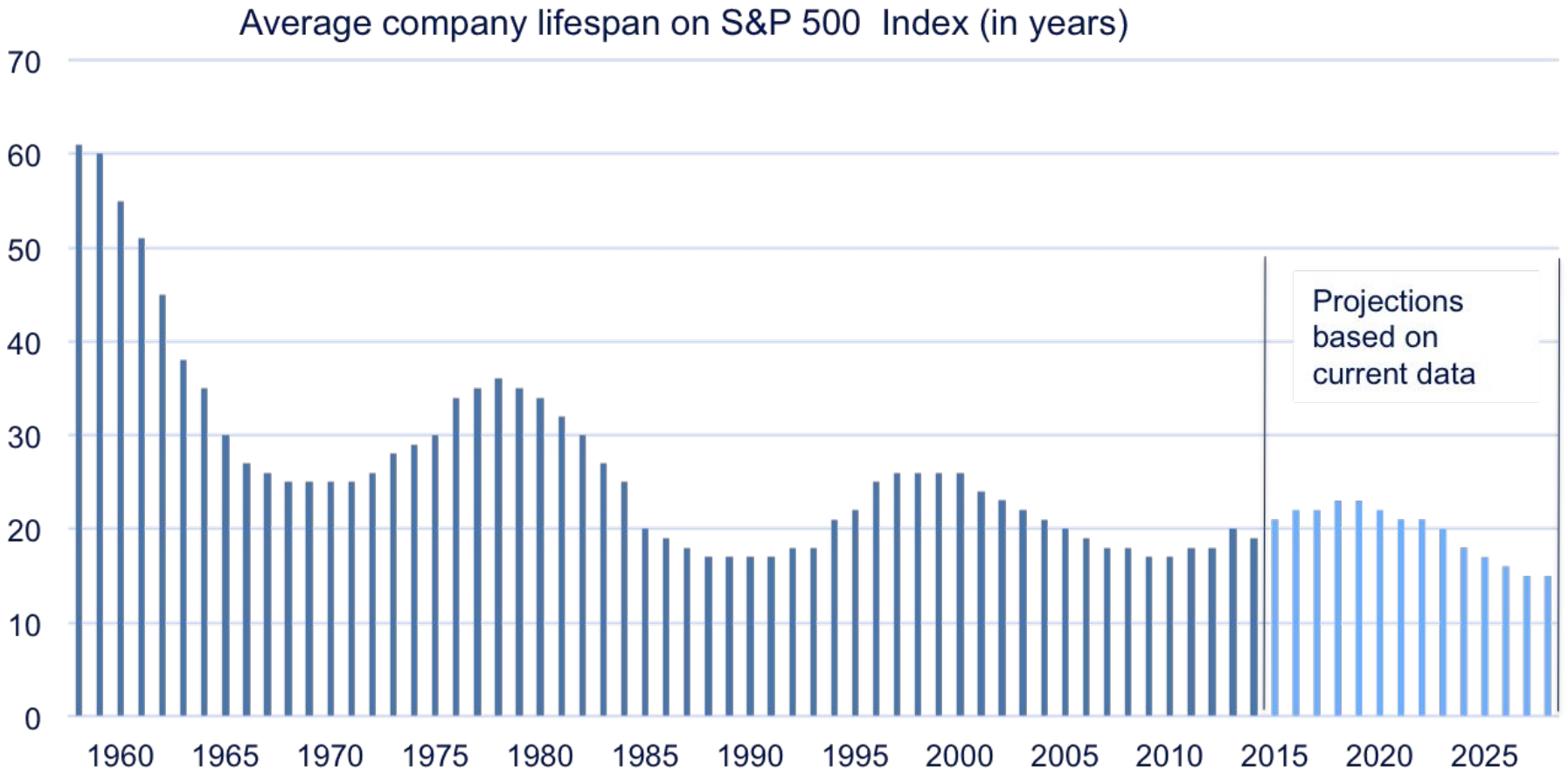
TOTAIS

Valor final com impostos (BRL):	18161.08

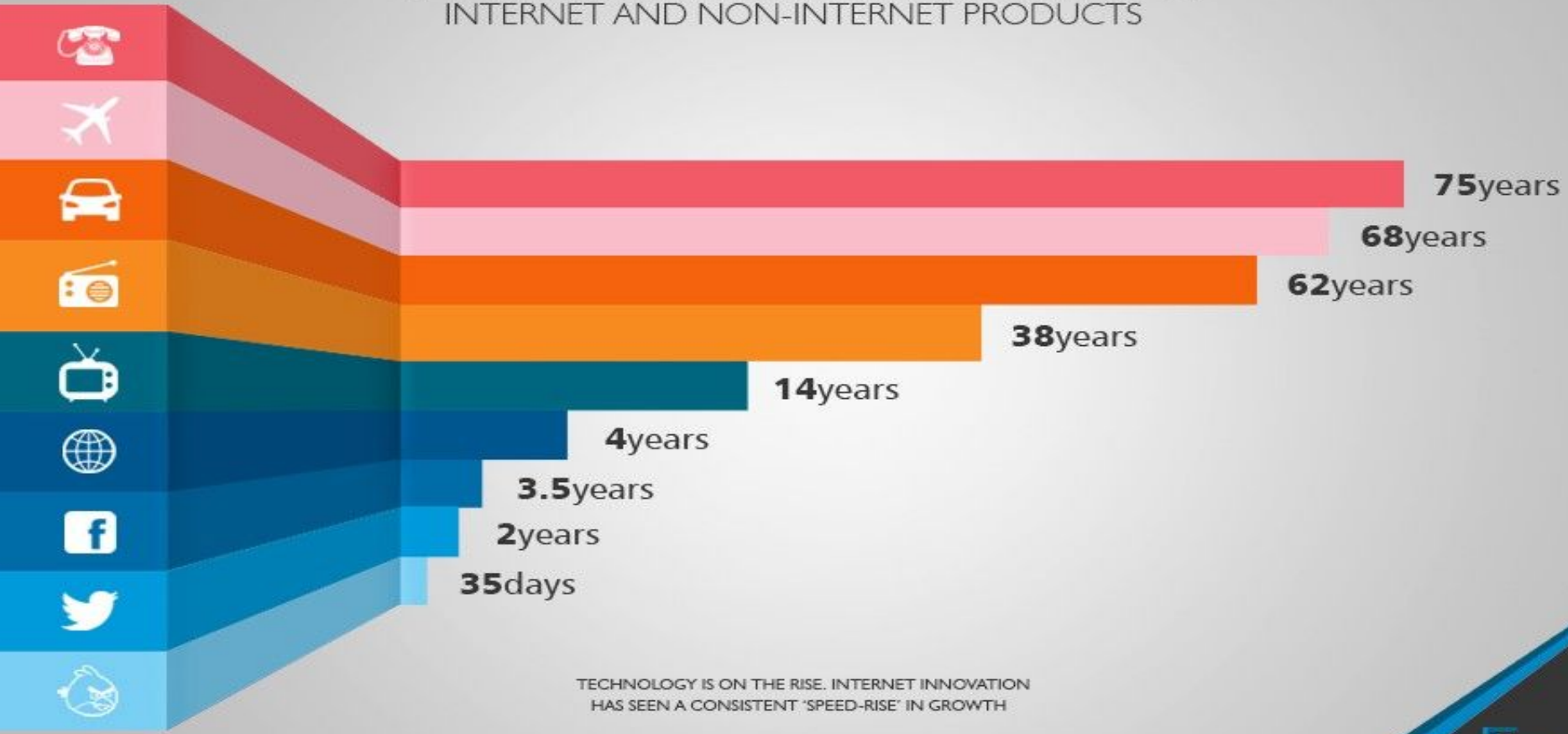
Lara Croft has changed over 21 years

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

CES 2018 with NVIDIA CEO Jensen Huang


Average company lifespan on S&P 500 Index (in years)
70
60
50
40
30
20
10
0
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
Projections based on current data

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

Provocation #2

data & internet & AI

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

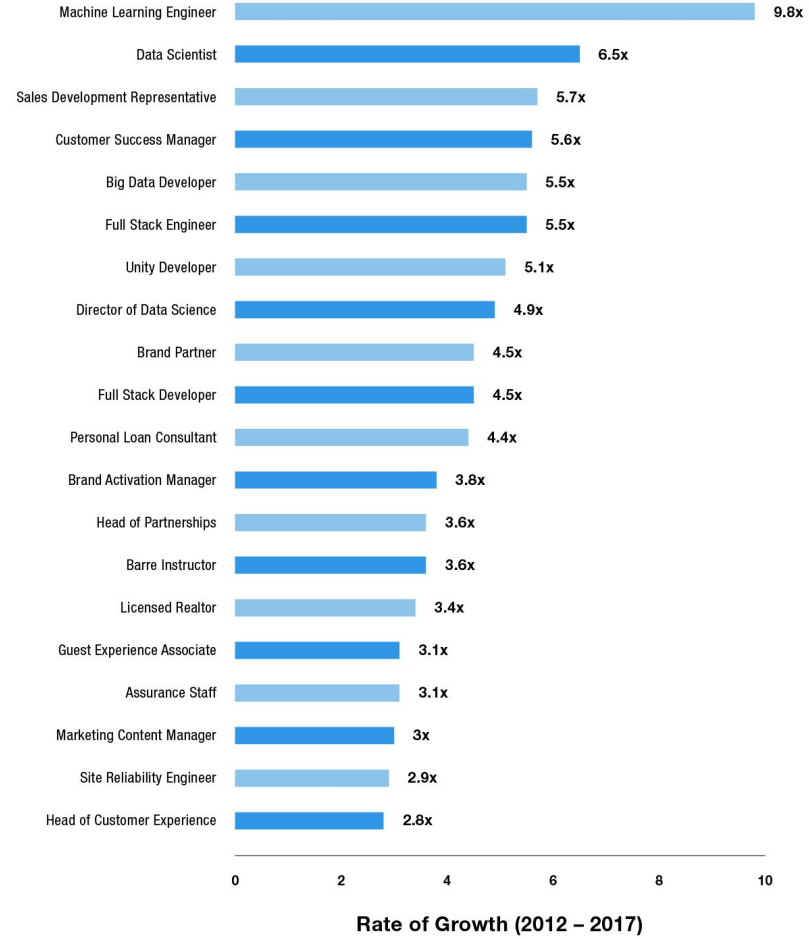
facebook

Read

Google

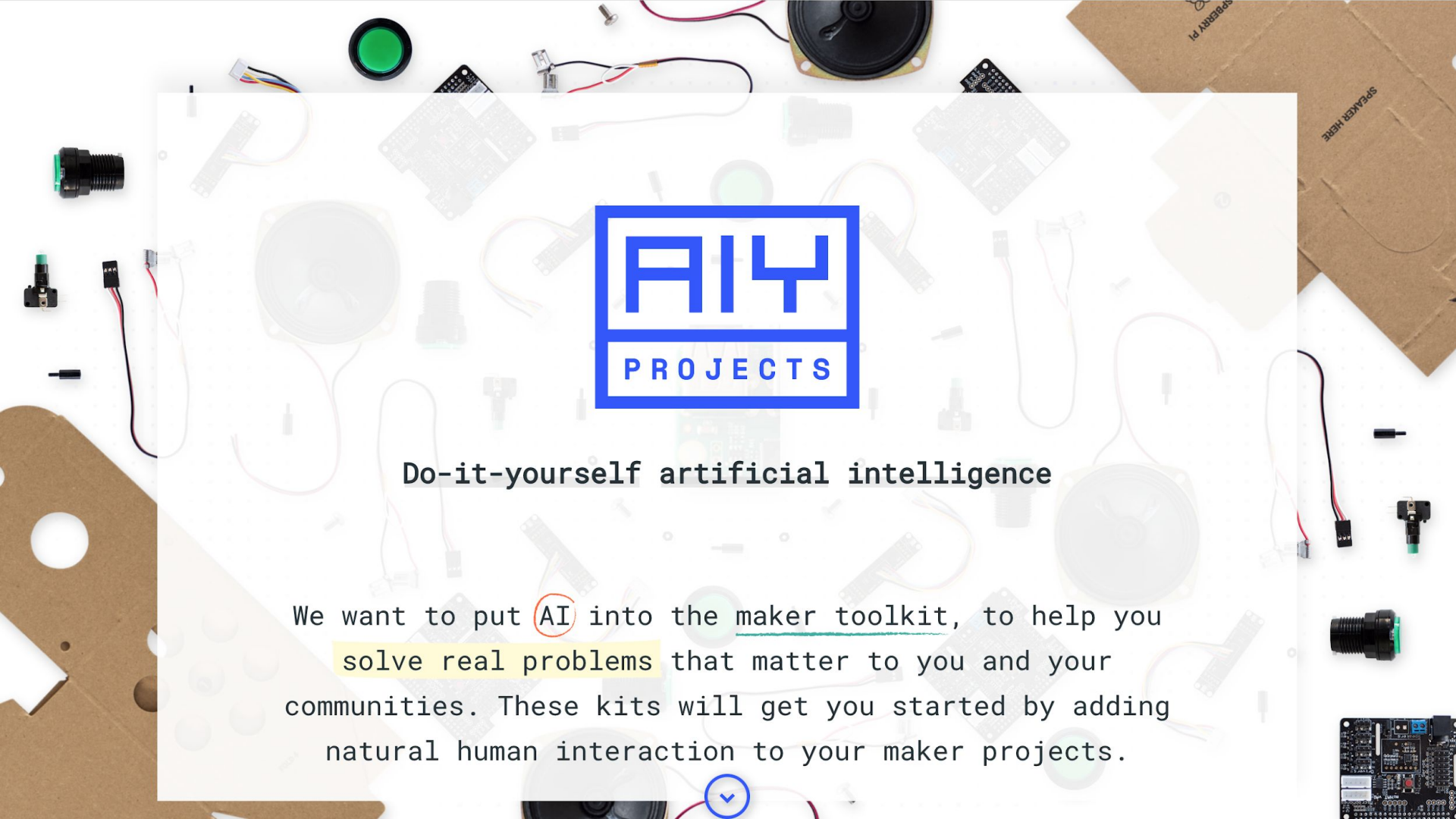
https://datafloq.com/read/the-top-7-big-data-trends-for-2017/2493

Top 20 Emerging Jobs



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

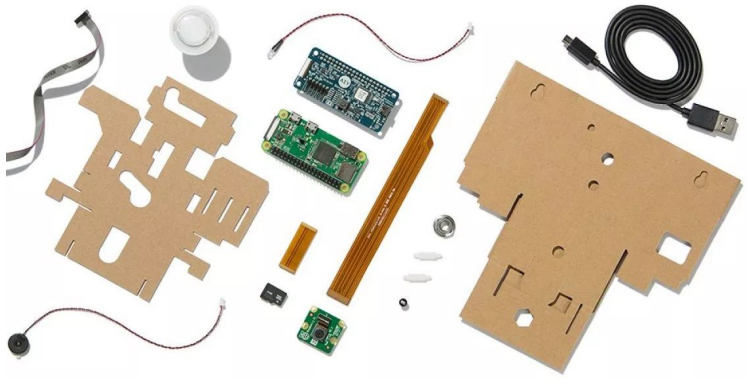
[Dec. 2017]
http://bit.do/forbesjobs



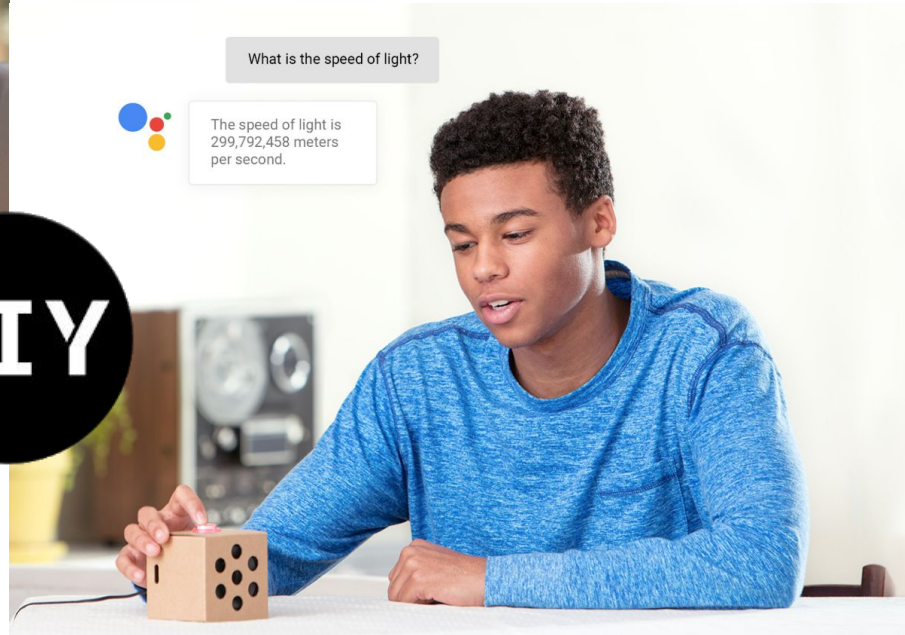
Do-it-yourself artificial intelligence

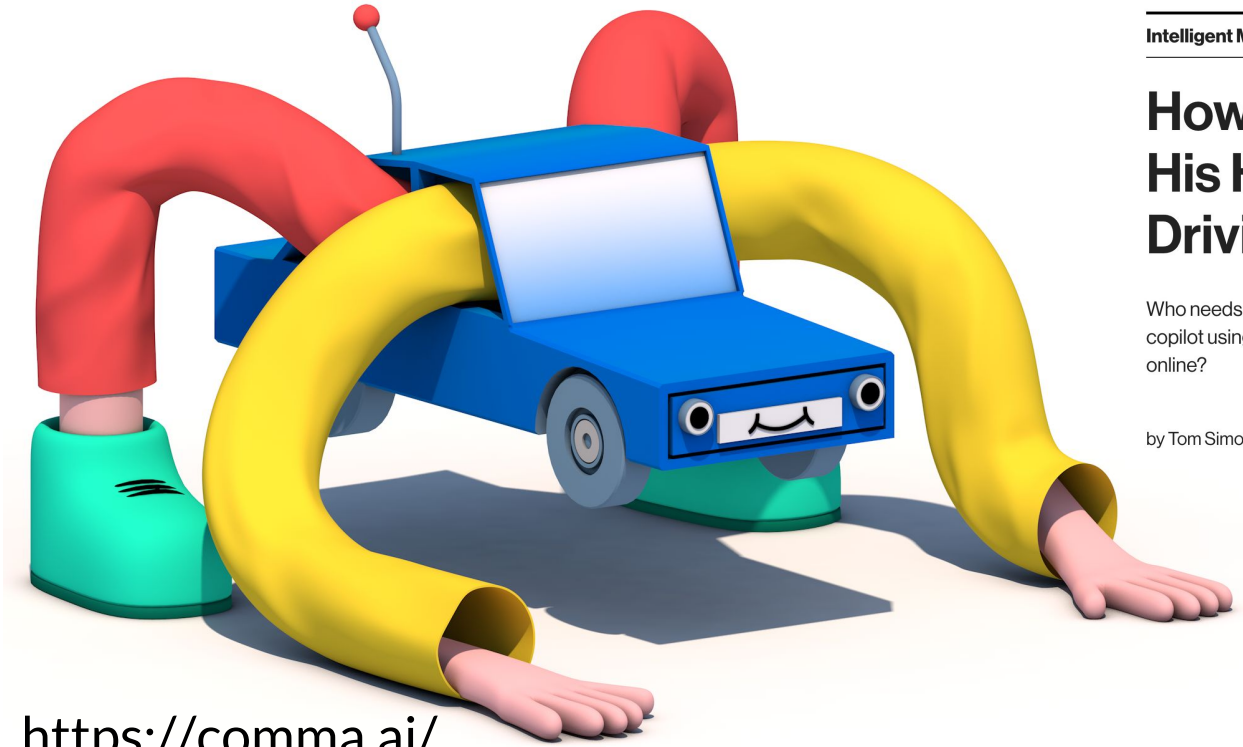
We want to put AI into the maker toolkit, to help you solve real problems that matter to you and your communities. These kits will get you started by adding natural human interaction to your maker projects.


79.20%
Fuji
Category: Edible Fruit
20.80%
Orange
Category: Edible Fruit



AIY


What is the speed of light?
The speed of light is 299,792,458 meters per second.



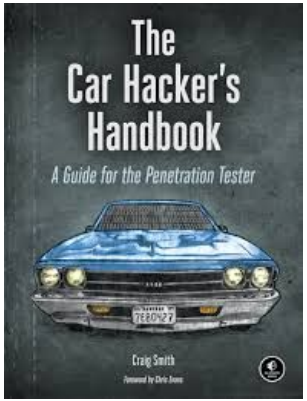
https://comma.ai/

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



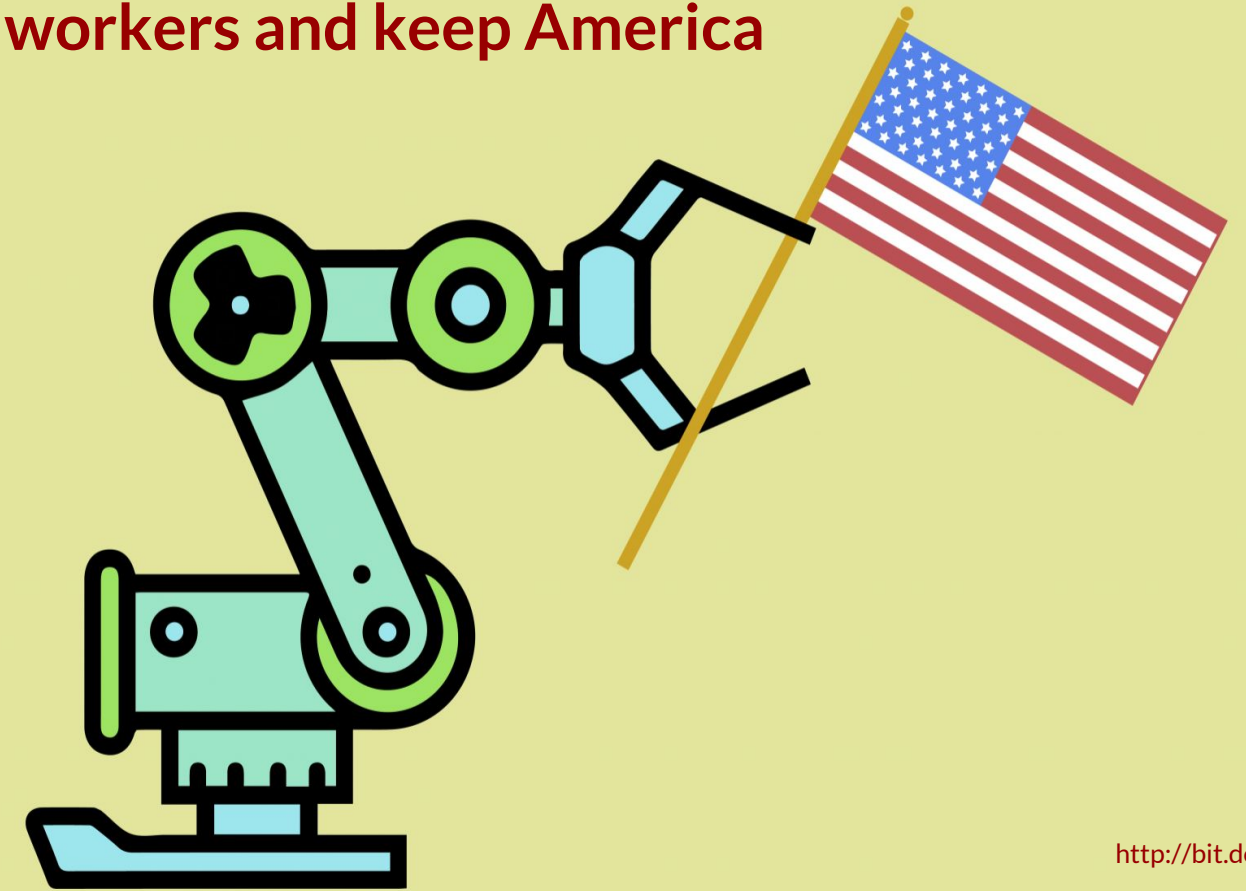
https://www.youtube.com/watch?v=3jstaBeXgAs

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/

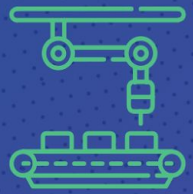

10
10

ABDI

MINISTÉRIO DA INDÚSTRIA, COMÉRCIO E SERVIÇOS

GOVERNO FEDERAL

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



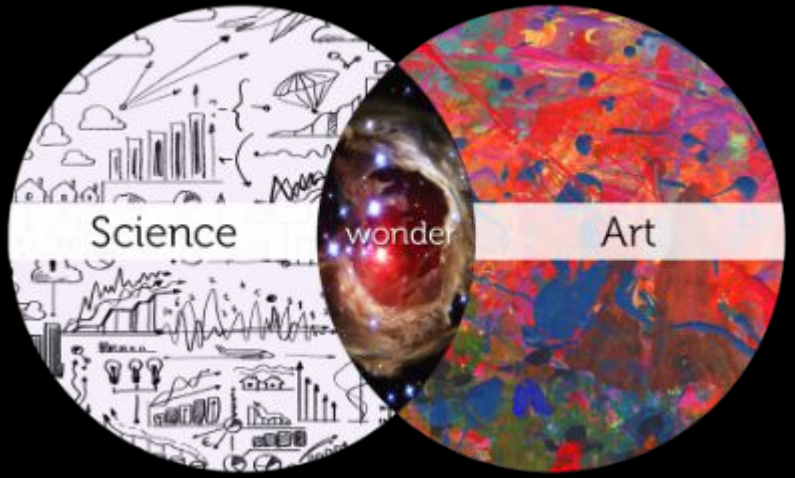
Agenda brasileira para a Indústria 4.0

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

data comes from everywhere

Data Files
(XML, CSV, Excel, JSON, …)

Database
(MySQL, Oracle, …)

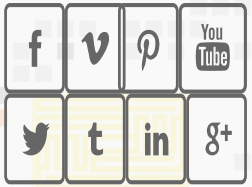
API

Sites

Text and reports

Maps

Image and videos



Social Media

BIG CHALLENGES

THE COMING FLOOD OF DATA IN AUTONOMOUS VEHICLES

RADAR
~10-100 KB
PER SECOND

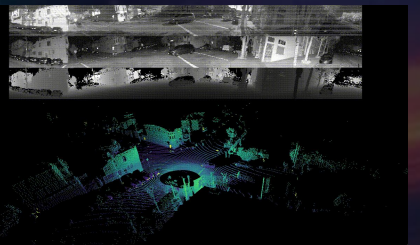
SONAR
~10-100 KB
PER SECOND

GPS
~50KB
PER SECOND

CAMERAS
~20-40 MB
PER SECOND

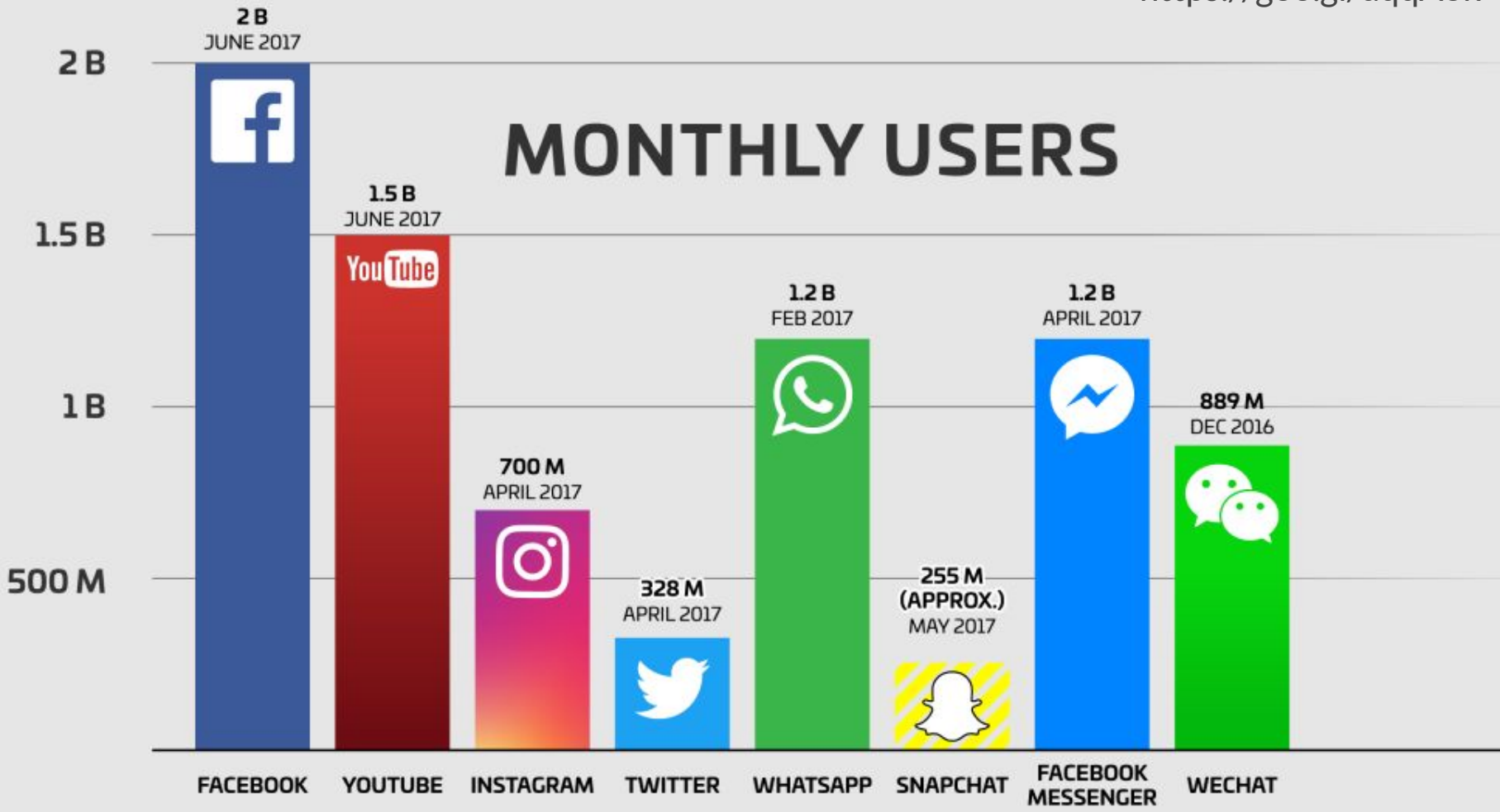
AUTONOMOUS VEHICLES
4,000 GB
PER DAY... EACH DAY

LIDAR
~10-70 MB
PER SECOND



BIG SOCIAL MEDIA

https://goo.gl/dqqAok




https://goo.gl/i9rDJX
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Ban Bei Ber Bog Bud Bue Cai Del Dha Gua Ist Jak Kar Kie Kol Lag Lon Los Mad Man Mex Mil Mos Mum New Osa Par Rio Rom Sao Seo Sha Sof Syd Tia Tok

http://bandicoot.mit.edu/

funf
Open Sensing Framework

(2017) A Preliminary Study of Activity and Sociability Behaviors Using Smartphone Sensing Methods. Computers in Human Behavior
(2017) Predicting Risk of Suicide Attempts Over Time Through Machine Learning


location
movement
Many More and
3rd Party API Support
usage/communication
social proximity
funf
Open Sensing Framework
Your Choice of Datastore
visualizations
notifications
applications
reports & studies

BIG SCIENCE

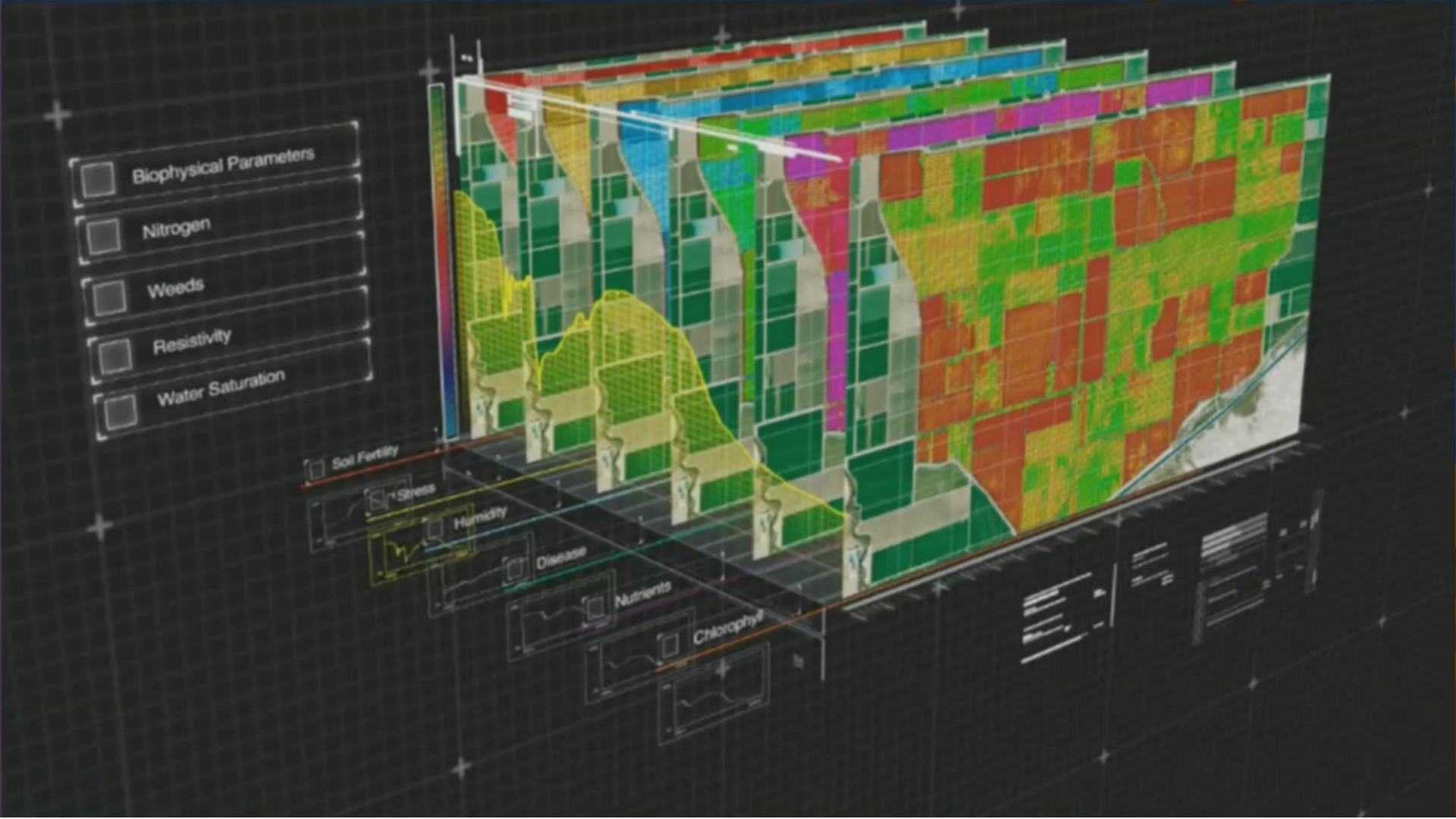

Extreme Large Telescope (ELT)
90 TB/night
Cerro Armazones, Chile

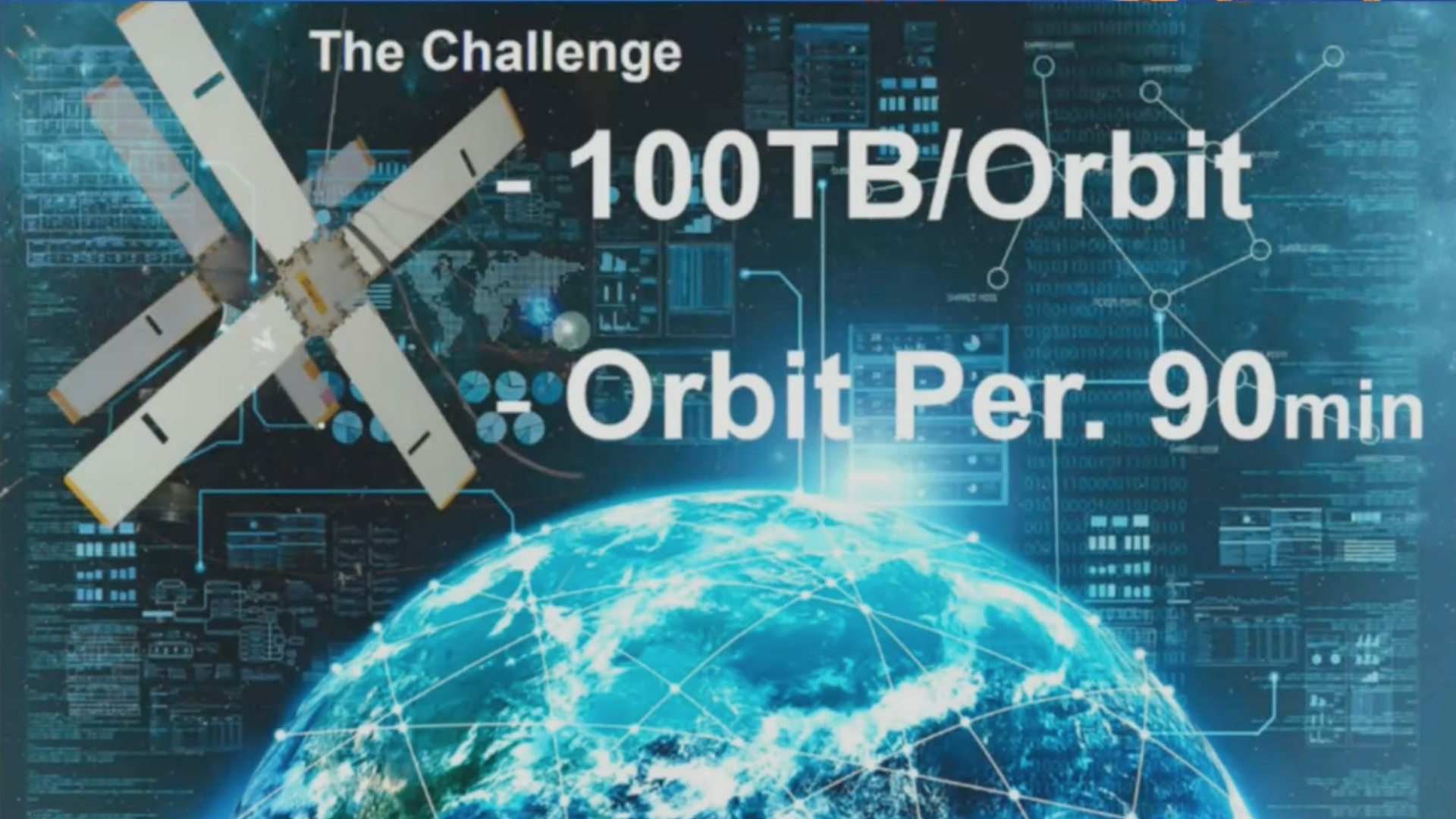

Quantities of Bytes
For Starters..
BIT = A BINARY DIGIT SET TO EITHER A 1 OR 0
BYTE = 8 BITS
KB KILOBYTE = 1,000 BYTES
MB MEGABYTE = 1,000,000 BYTES
GB GIGABYTE = 1,000,000,000 BYTES
TB TERABYTE = 1,000,000,000,000 BYTES
PB PETABYTE = 1,000,000,000,000,000 BYTES
EB EXABYTE = 1,000,000,000,000,000,000 BYTES
ZB ZETTABYTE = 1,000,000,000,000,000,000,000 BYTES
YB YOTTABYTE = 1,000,000,000,000,000,000,000,000 BYTES

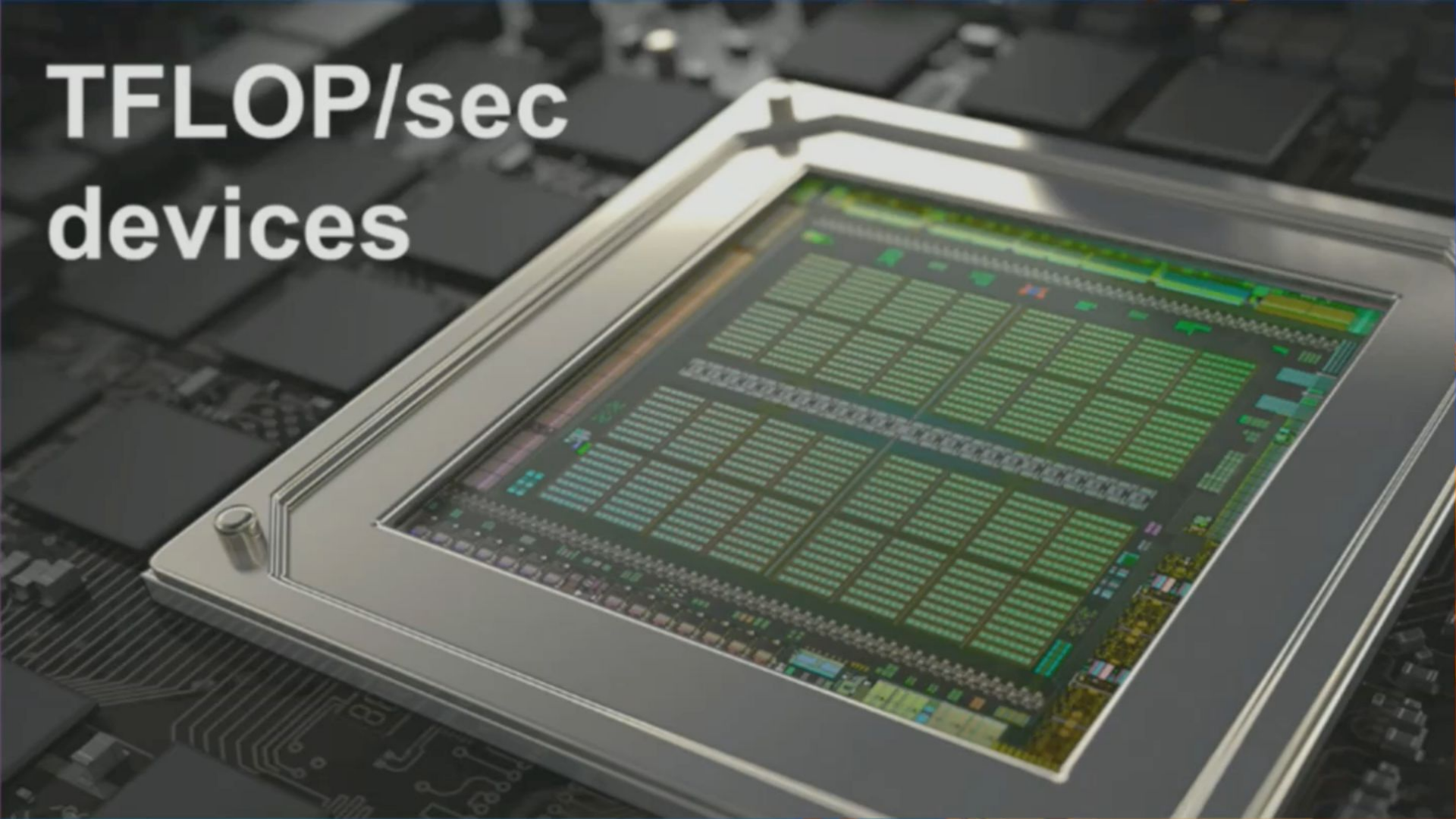
Small is the new big

HYPERCUBES	SKYSAT	LANDSAT 8	WORLDVIEW-3
Operator:	Skybox Imaging	NASA	DigitalGlobe
Number of satellites*: 100	24	N/A	N/A
Weight: <10kg	~100 kg	2,071 kg†	2,800 kg
Instruments: Hyperspectral	Optical and near-infrared spectral bands	Multiple spectral bands	Multiple spectral bands

THE HYPER CONSTELLATION


Biophysical Parameters
Nitrogen
Weeds
Resistivity
Water Saturation
Soil Fertility
Stress
Humidity
Disease
Nutrients
Chlorophyll

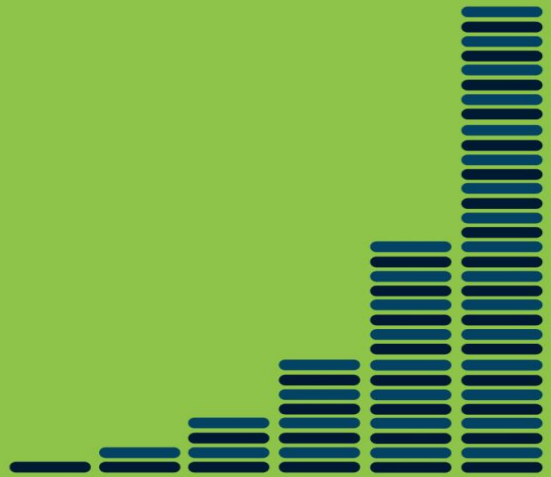

The Challenge
- 100TB/Orbit
- Orbit Per. 90min


TFLOP/sec
devices

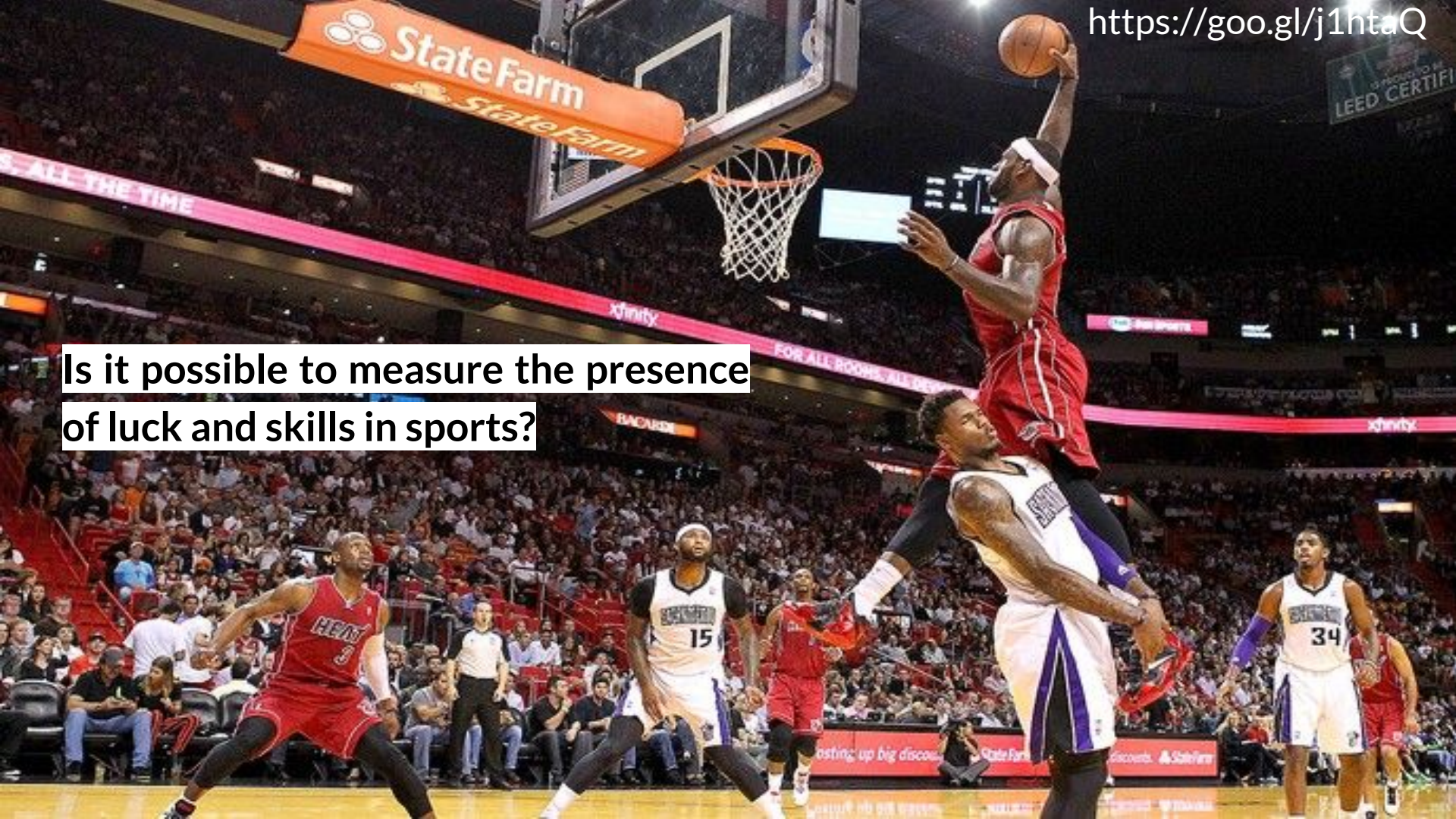
The average person is likely to generate more than one million gigabytes of health-related data in their lifetime. Equivalent to 300 million books.

IBM **Watson Health**

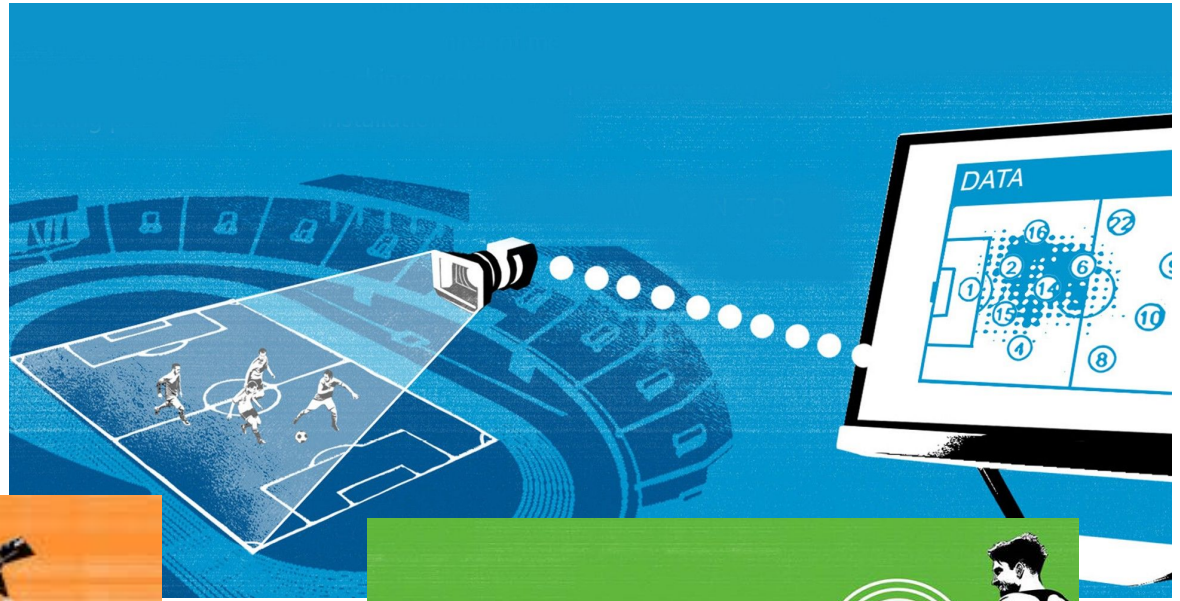
Medical data is expected to double every 73 days by 2020.

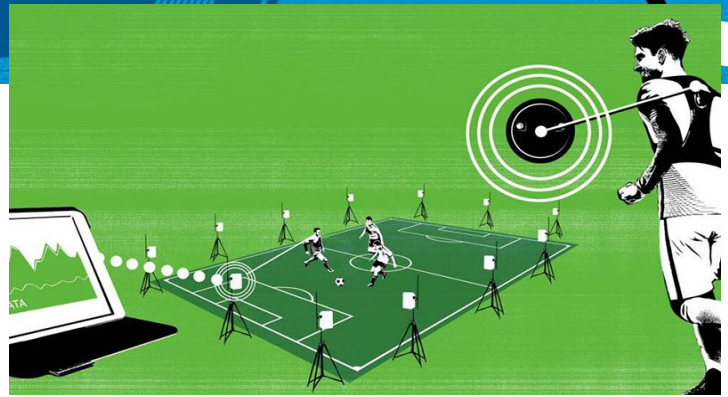


https://goo.gl/j1htaQ

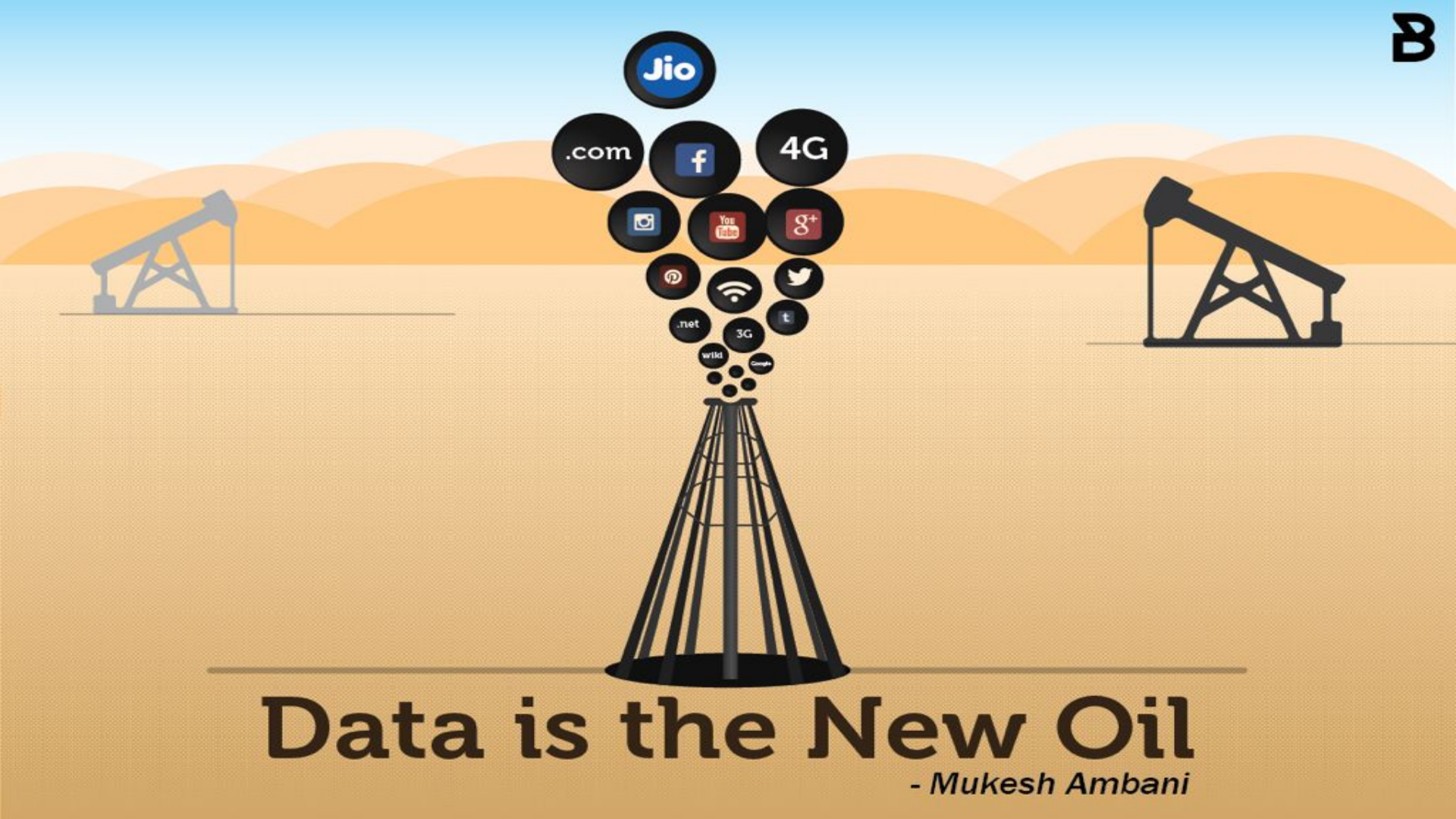


Is it possible to measure the presence of luck and skills in sports?



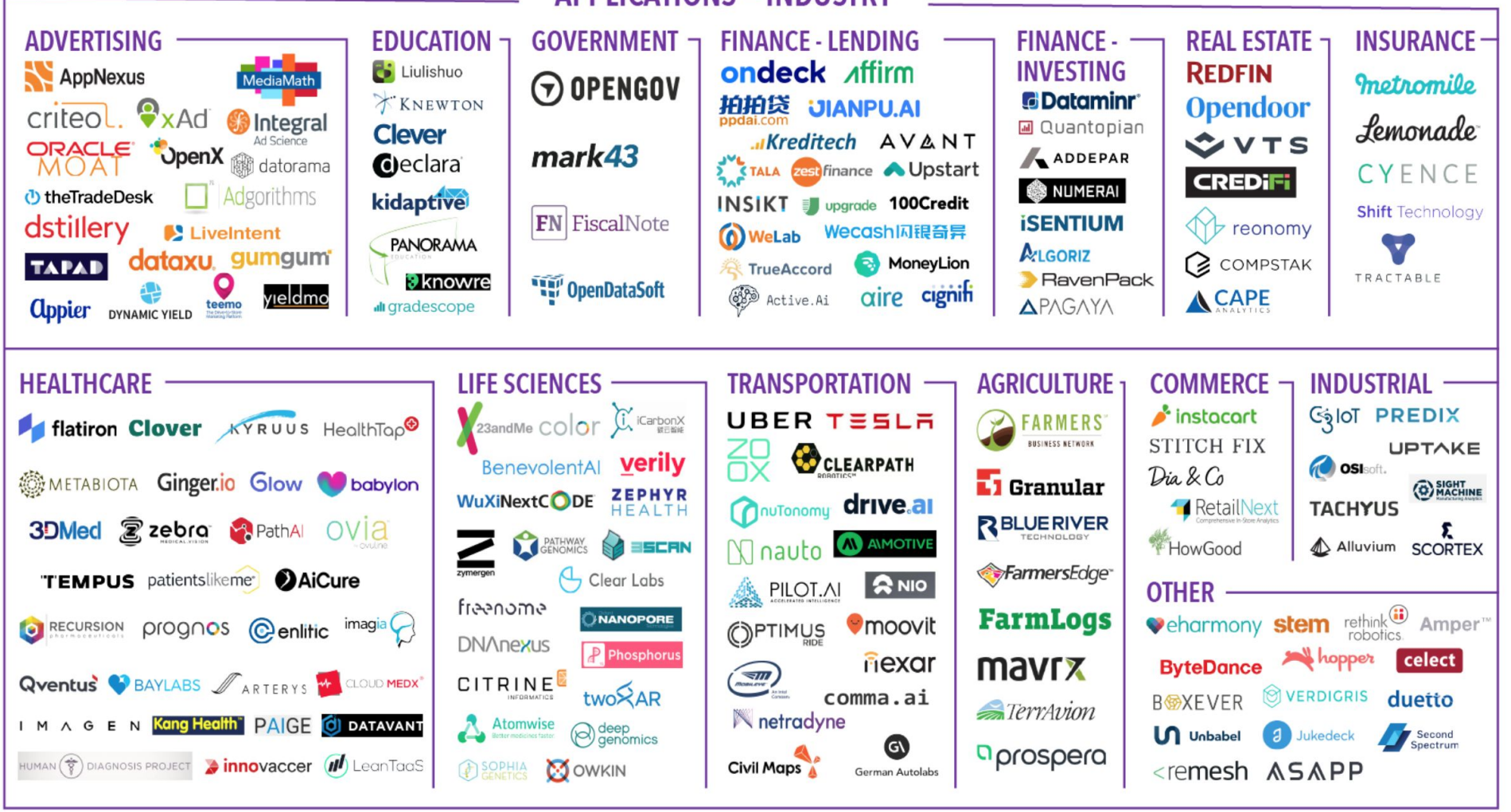


https://football-technology.fifa.com/en/media-tiles/epts
https://www.sporttechie.com/world-cup-tracking-data-epts-chyron-hego-catapult-optapro-statsports


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

APPLICATIONS – INDUSTRY

http://mattturck.com/bigdata2018/



Provocation #3

bias & privacy

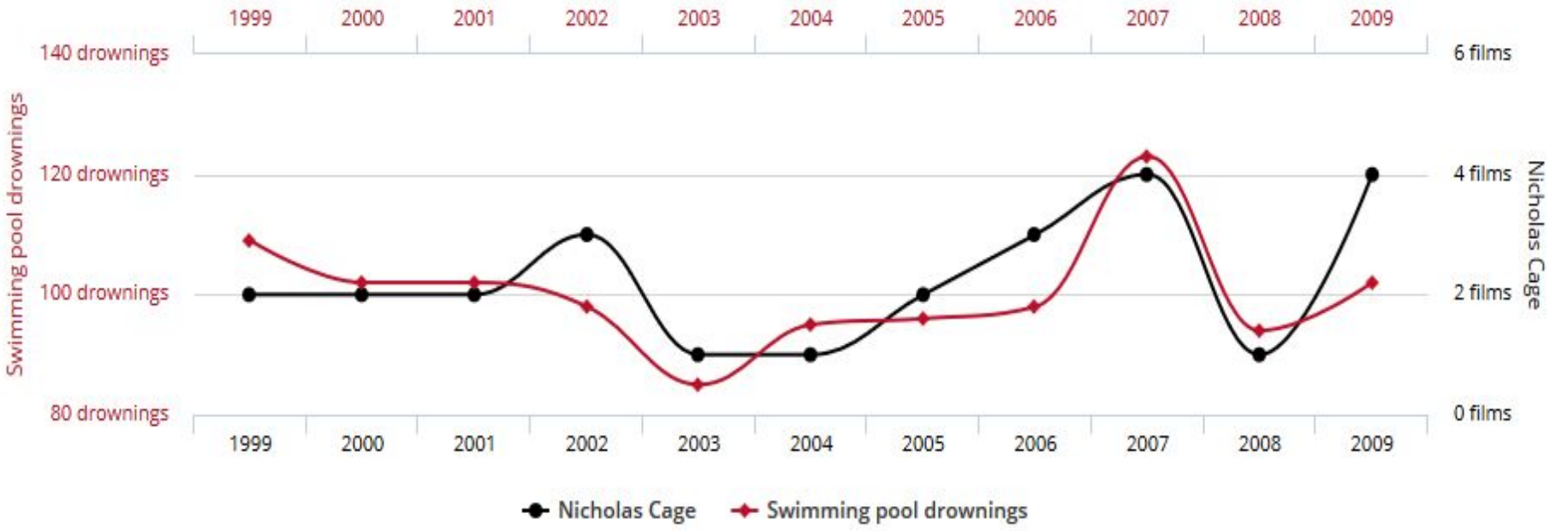

I got a big one this time!
BIG DATA
TRUTH
CORRELATION
NOISE
OVERFIT
@redpenblackpen

Number of people who drowned by falling into a pool

correlates wtih

Films Nicolas Cage appeared in

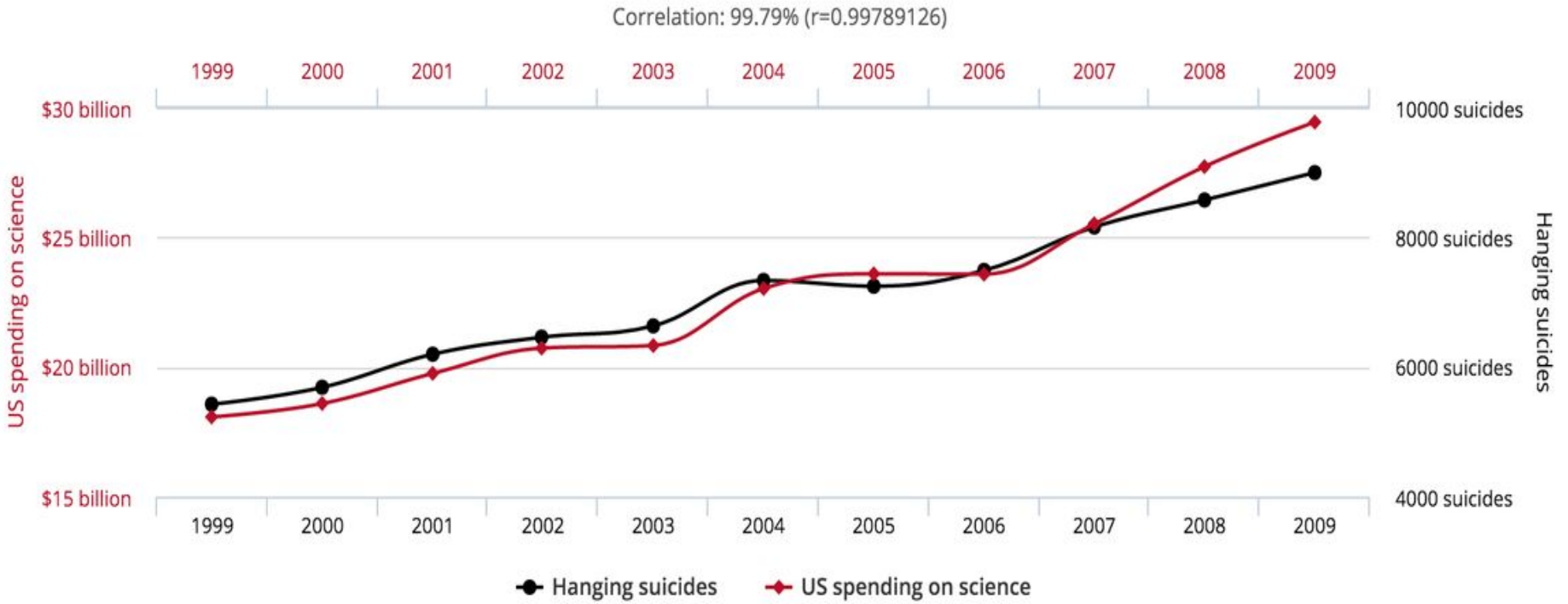
Correlation: 66.6% (r=0.666004, p>0.05)



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

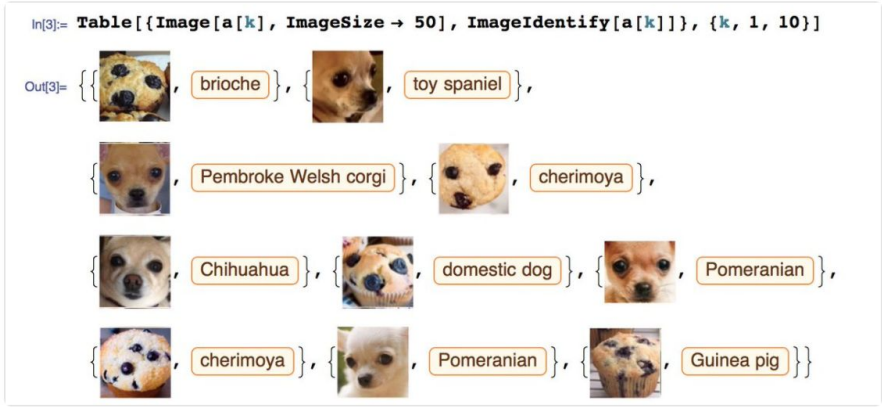


tylervigen.com

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



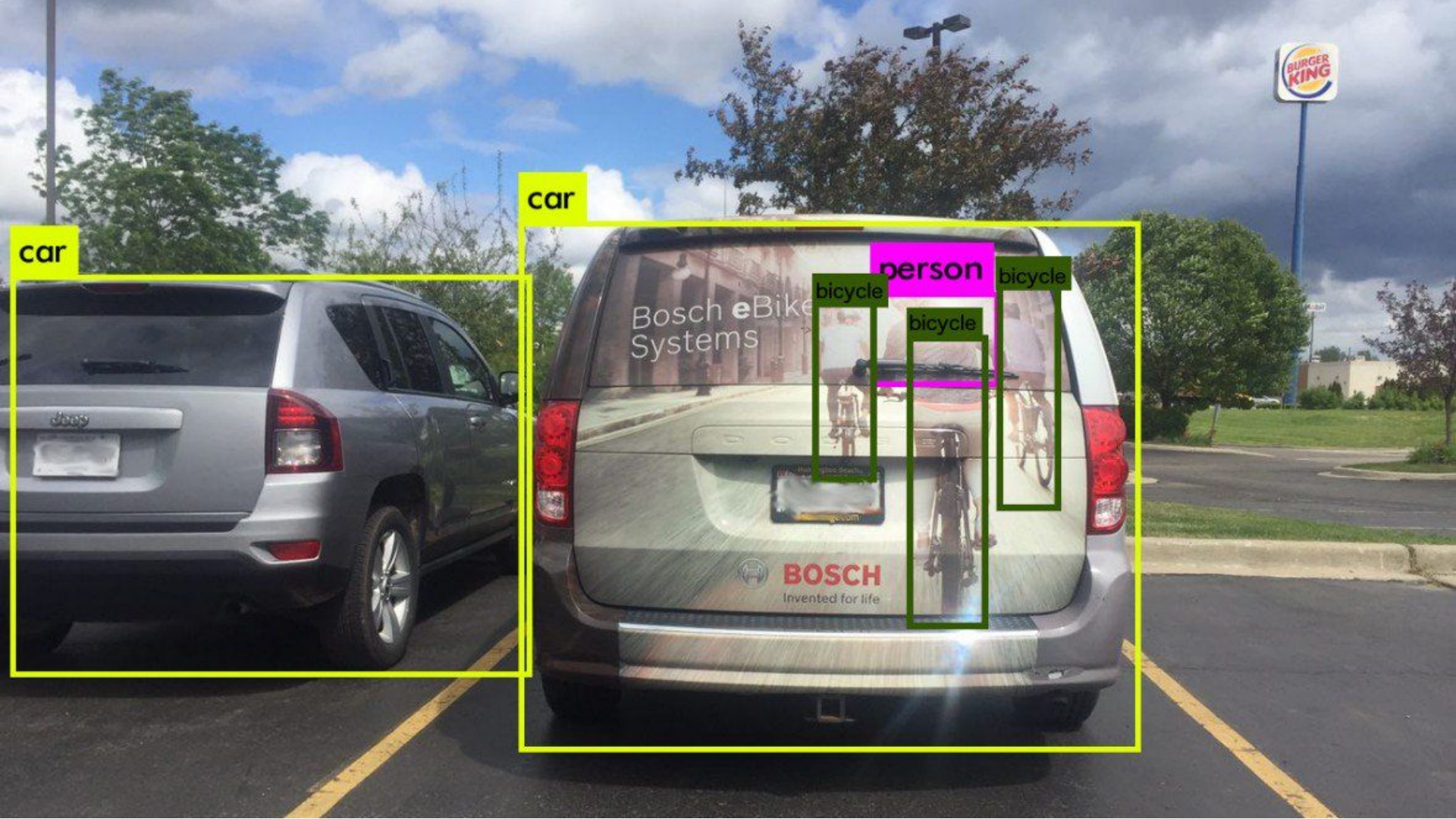
7:42 AM - 11 Mar 2016

●●●●● Verizon 4:20 PM 34%

Albums **puppy or bagel** Select

●●●○○ Verizon 10:50 PM 4%

Back **labradoodle or fried chicken** Select


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

DATA VIOLENCE

and how bad engineering choices can damage society

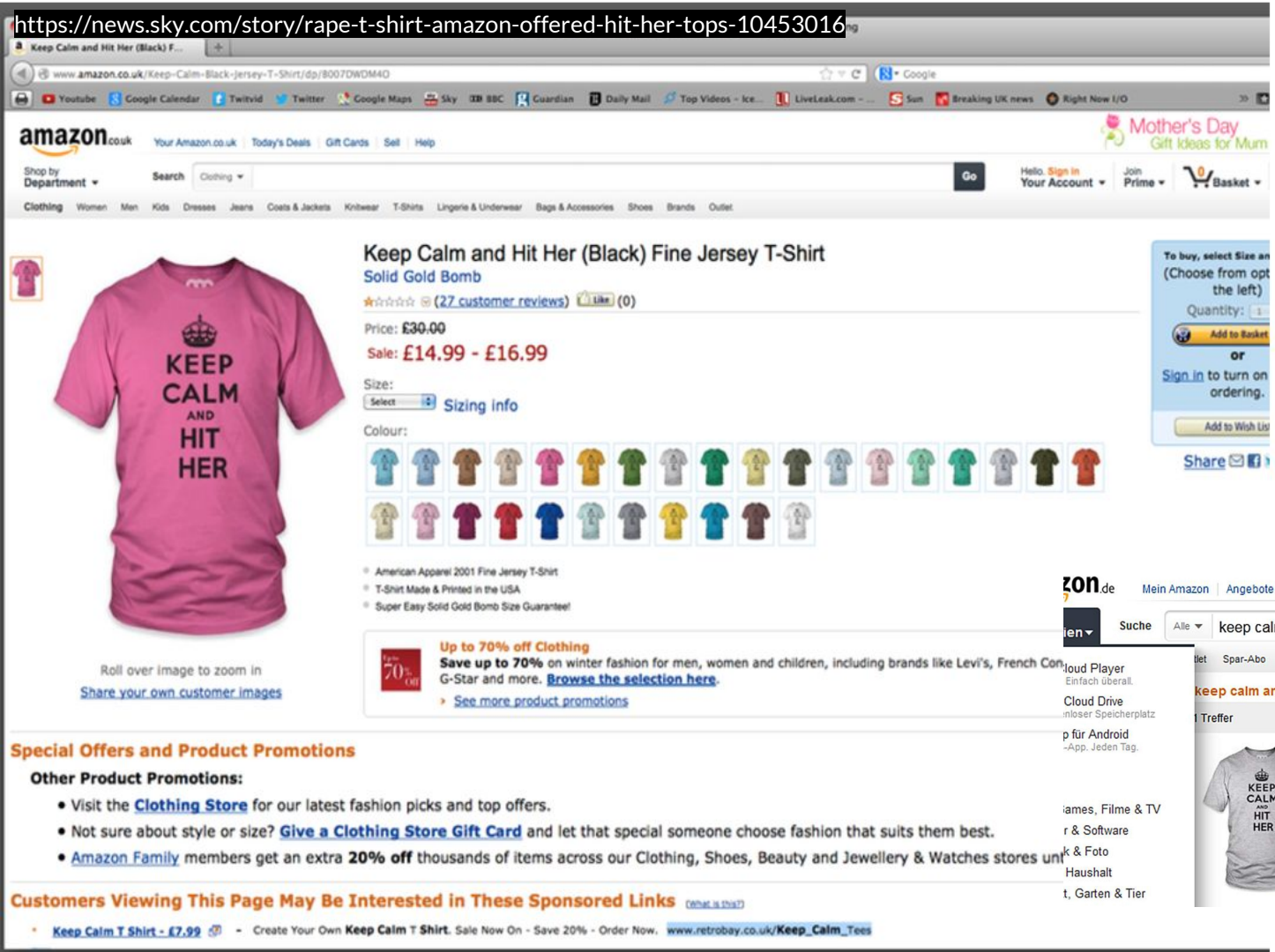
Detection of unexpected shapes can be considered potential threats, leading to additional scrutiny of the passenger.

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

https://goo.gl/NwP7Fv

https://goo.gl/xzLxaY

https://news.sky.com/story/rape-t-shirt-amazon-offered-hit-her-tops-10453016

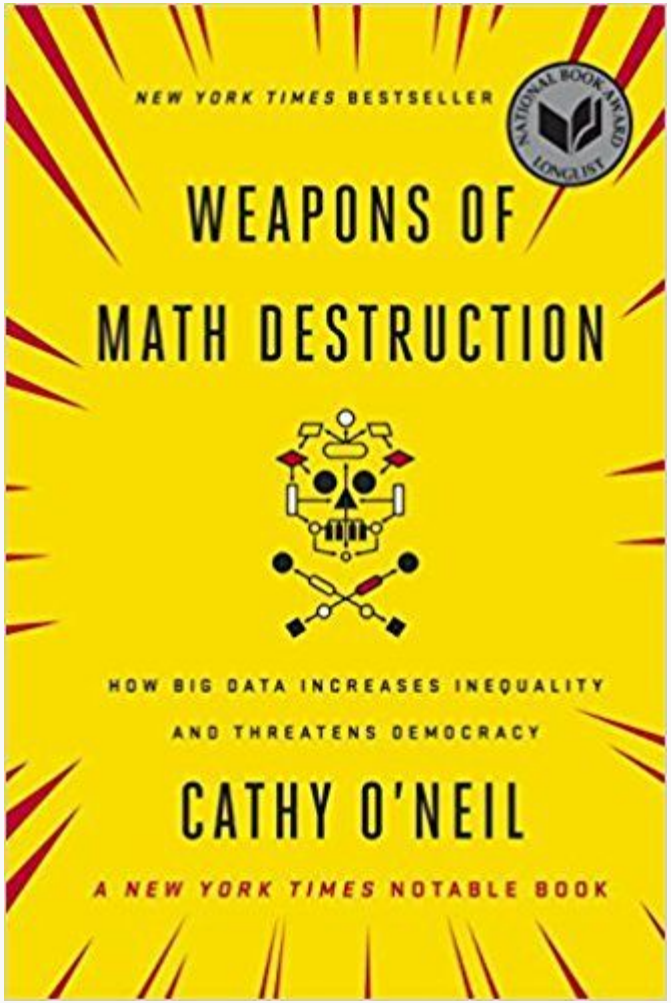


"We've been informed via social media that we might be stocking a t-shirt with a message that we too find unacceptable. We are investigating on how it was even displayed and will be pulled out immediately."

SM Supermalls
@smsupermalls

Statement on t-shirt incident.
2:32 PM - Sep 22, 2014
♡ 12 38 people are talking about this

“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

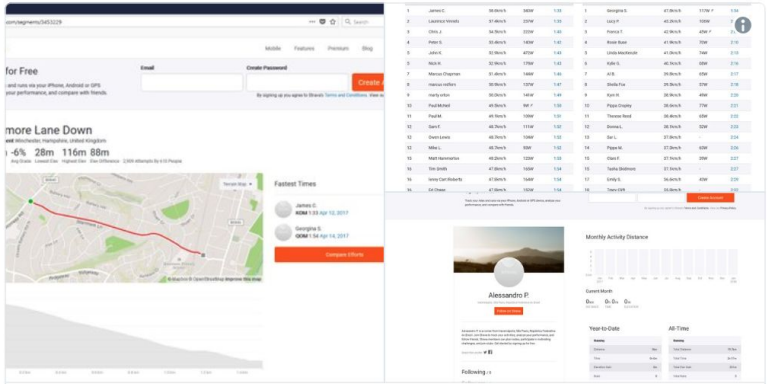
statista

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

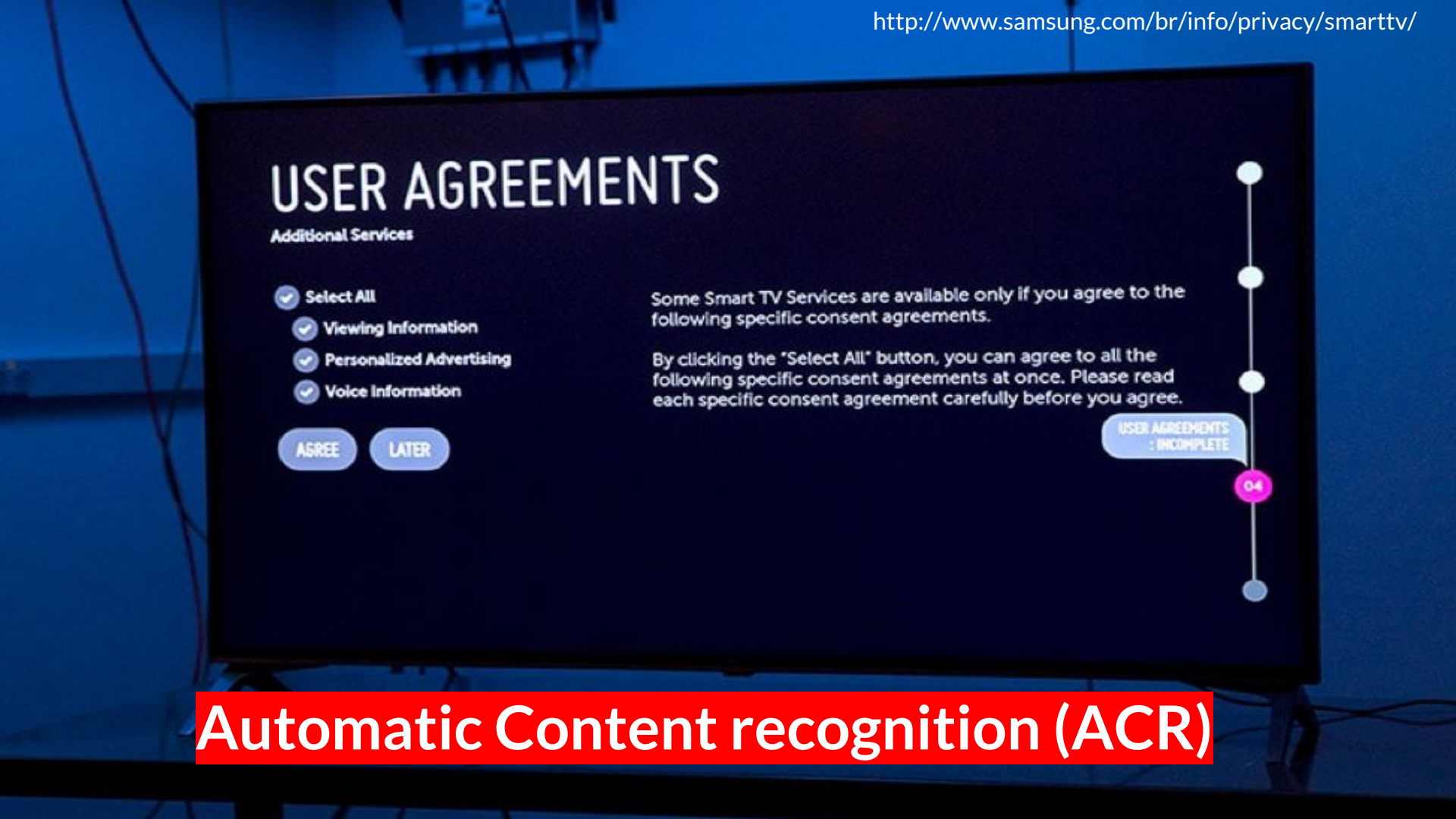
6:51 PM - Jan 28, 2018

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

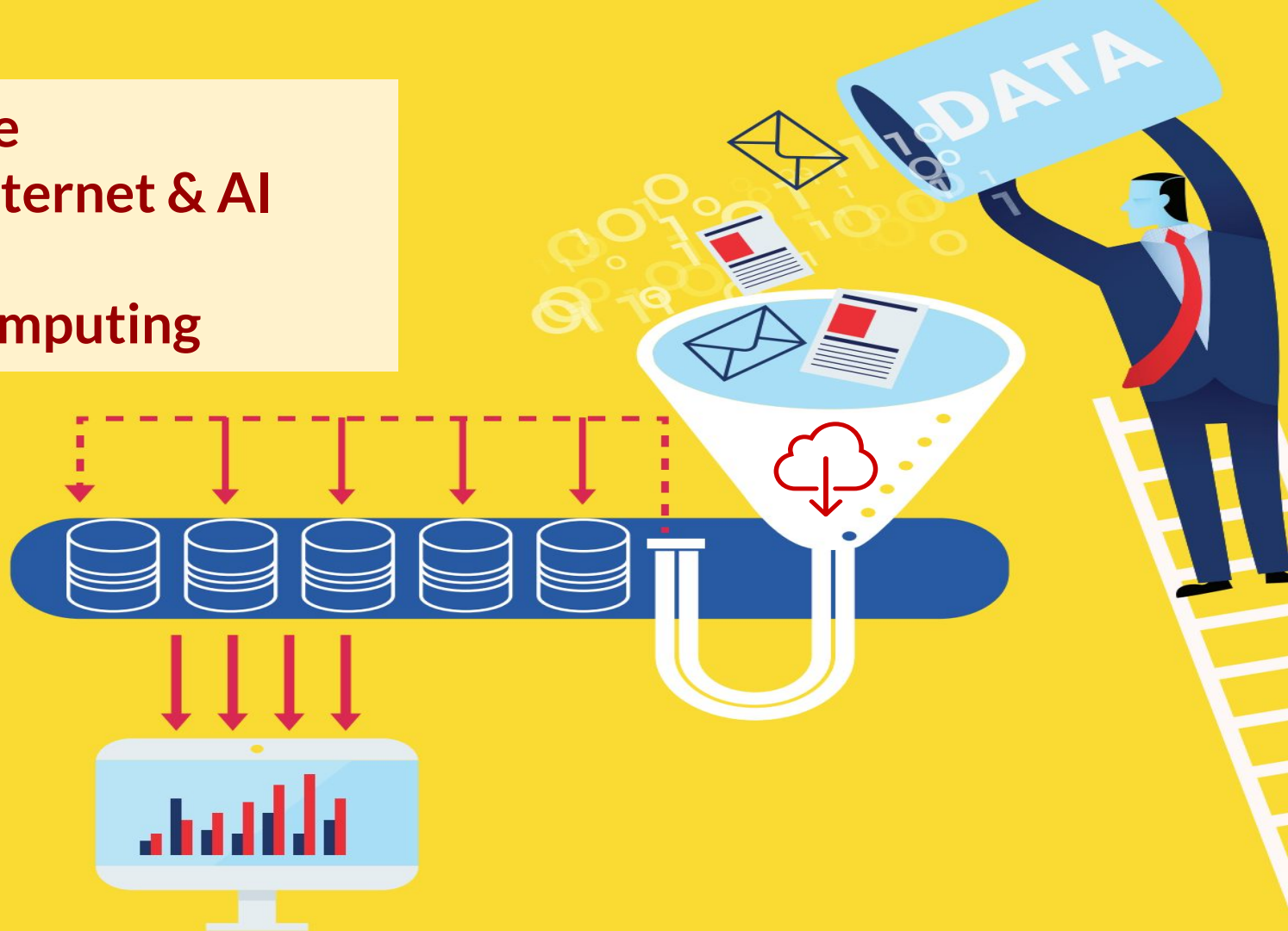
linode

DigitalOcean

ORACLE
Cloud Infrastructure

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

Hardware
Data & Internet & AI
Bias
Cloud Computing



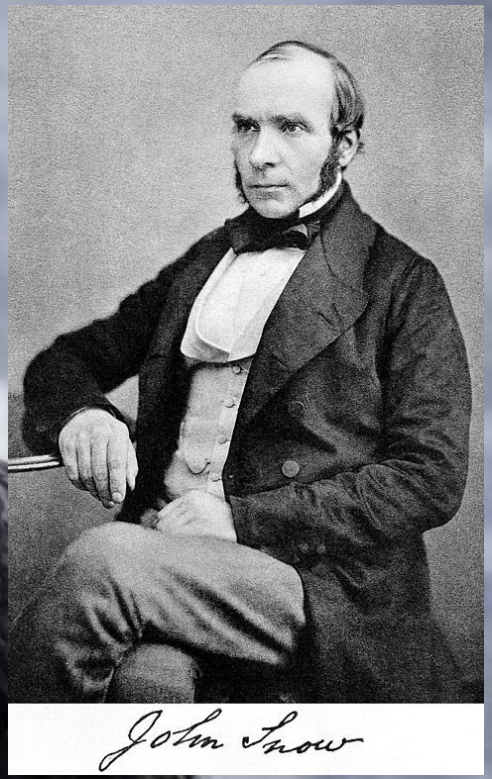
Who studies this stuff?

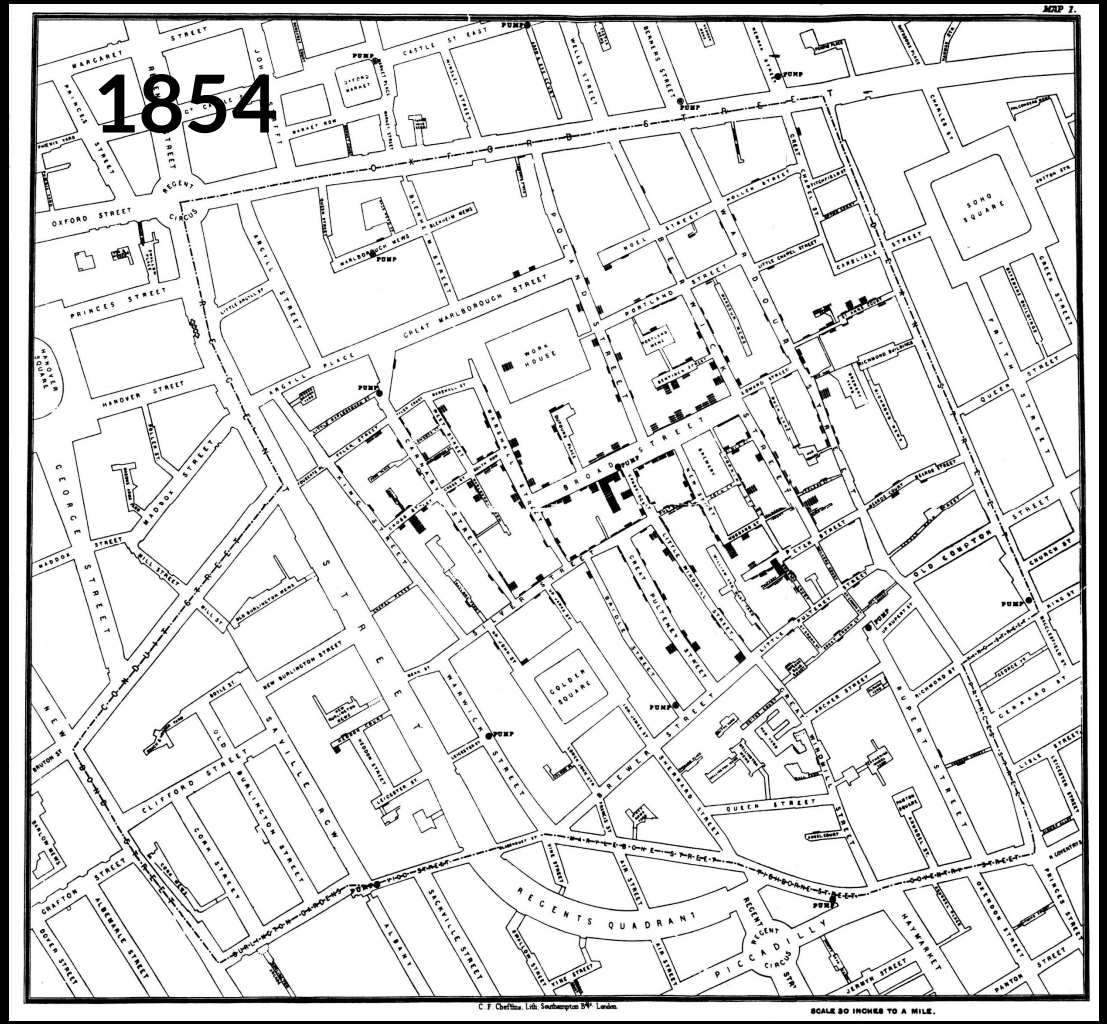
A Data-Driven Program

[Data Scientist: The Sexiest Job of the 21st Century]

How to Become a **Data Scientist**

John Snow, London, 1854





https://en.wikipedia.org/wiki/John_Snow

MODERN DATA SCIENTIST

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

MATH & STATISTICS

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

PROGRAMMING & DATABASE

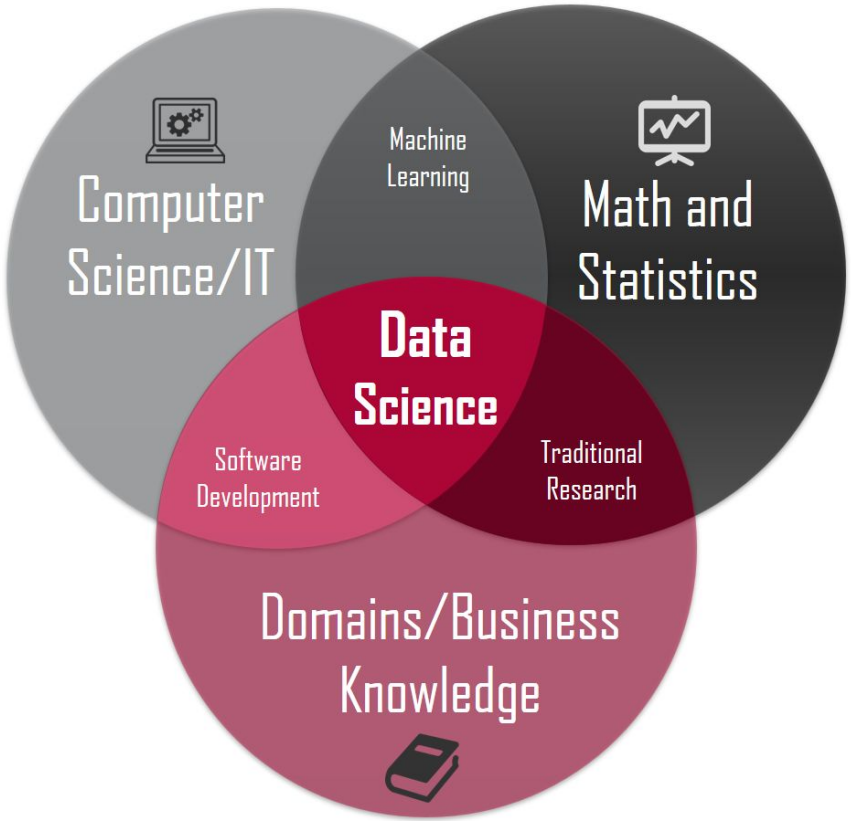
- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

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

COMMUNICATION & VISUALIZATION

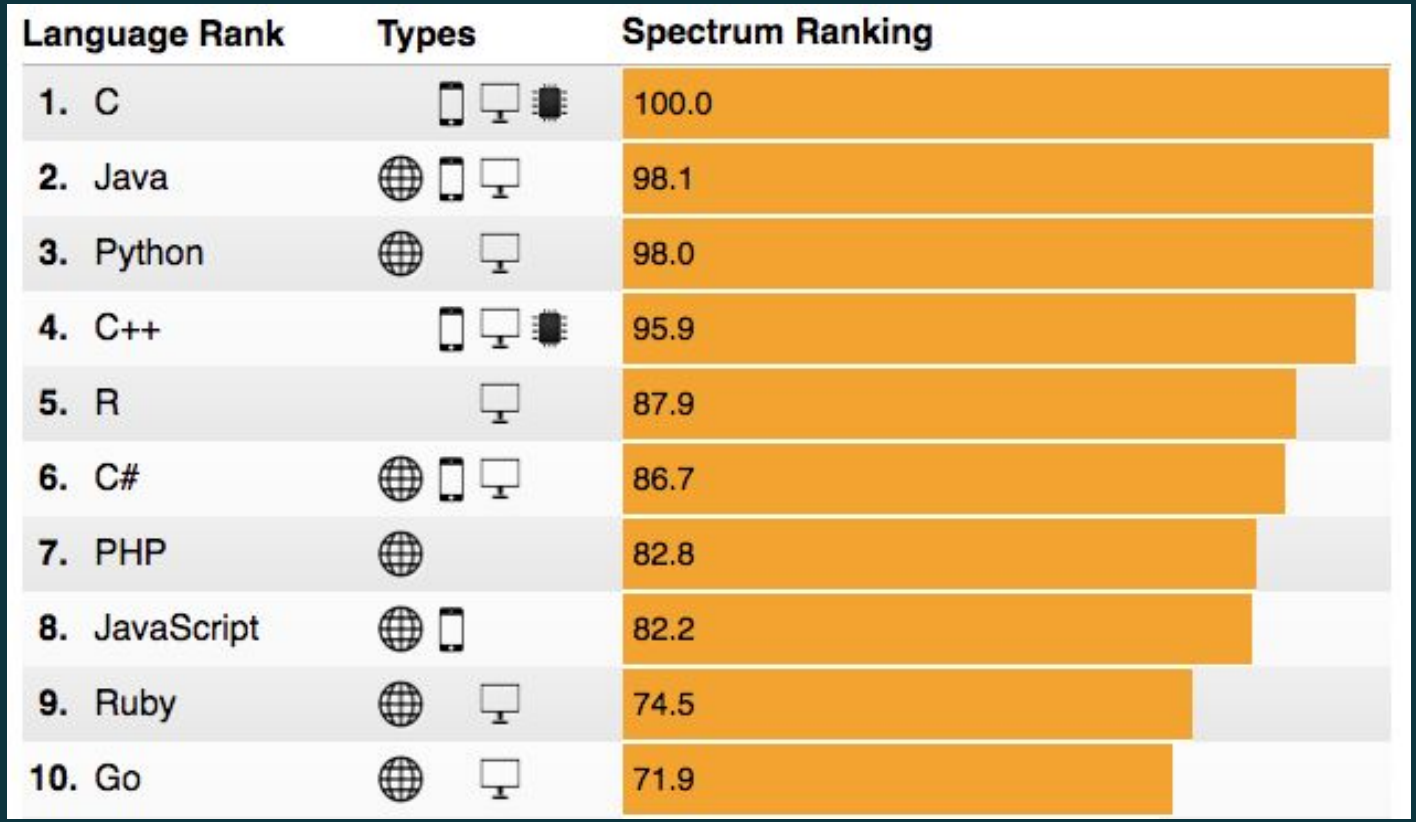
- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



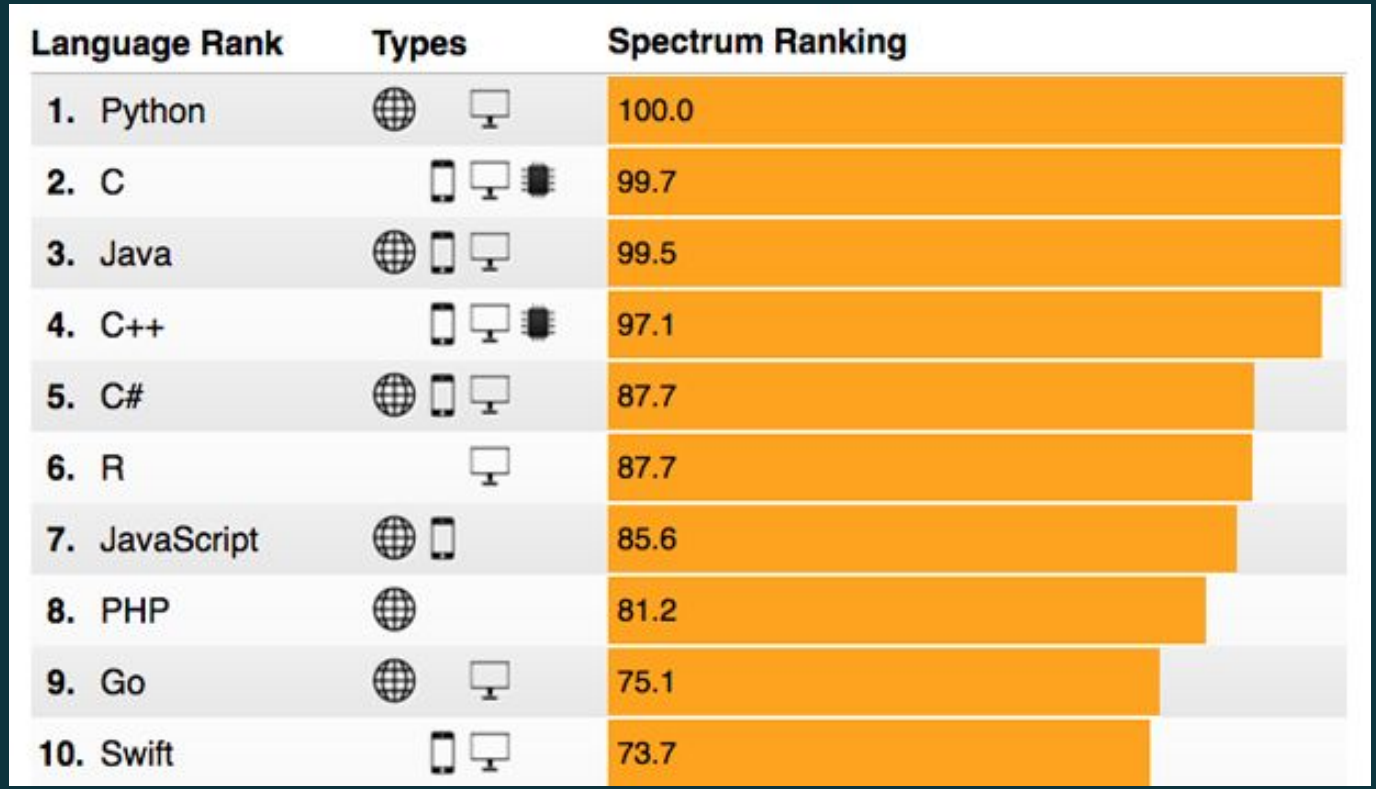
Pick ONE programming language and STICK to it. Don’t go back and constantly change your choice of language to study. If you do, you will slow your progress down.


which programming language
to learn first (DS,ML)?

https://goo.gl/VKYfXn

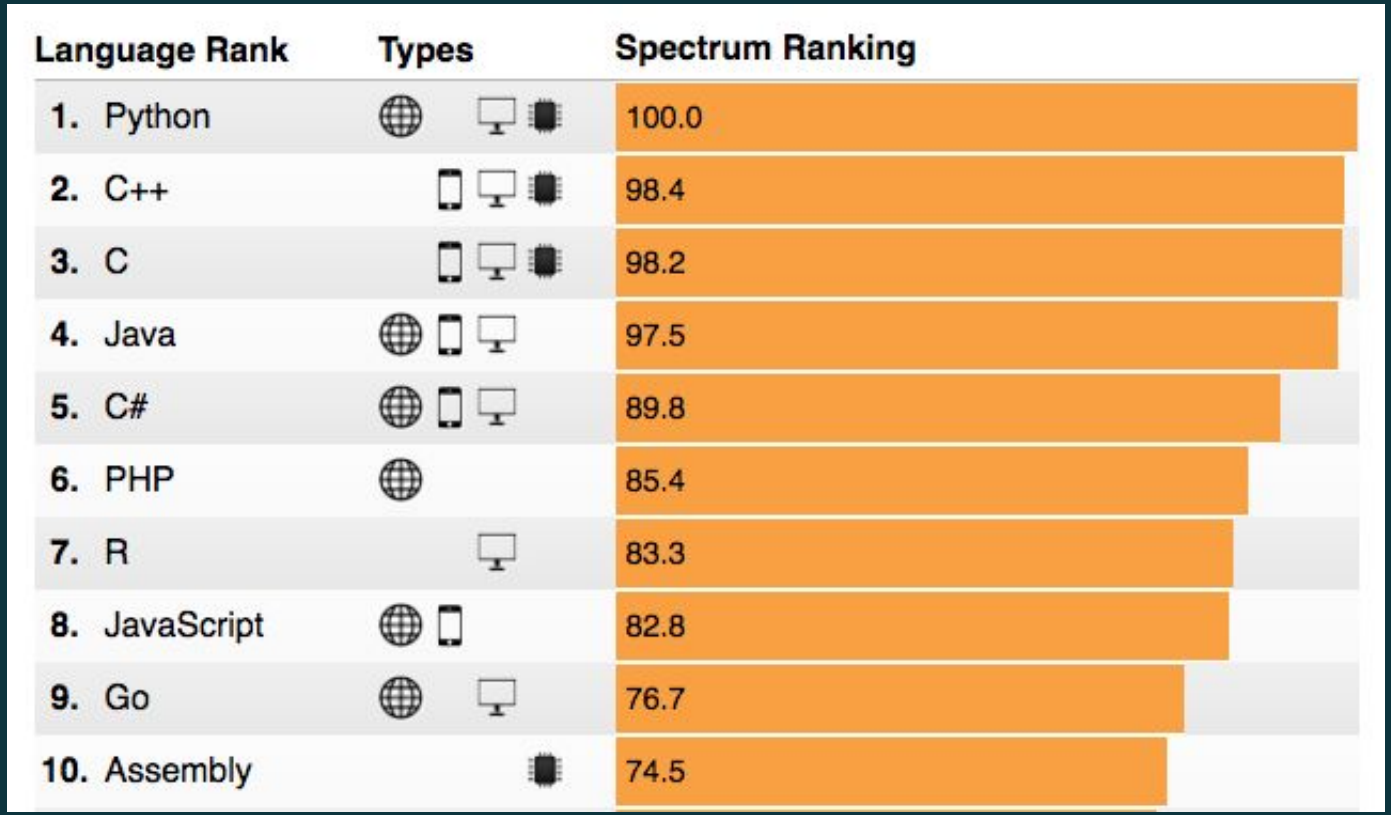


IEEE Spectrum - Jul 2016 https://goo.gl/BqrkDI



Language Rank	Types	Spectrum Ranking
1. Python		100.0
2. C		99.7
3. Java		99.5
4. C++		97.1
5. C#		87.7
6. R		87.7
7. JavaScript		85.6
8. PHP		81.2
9. Go		75.1
10. Swift		73.7

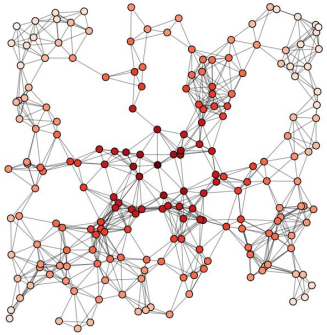
IEEE Spectrum - Jul 2017 https://goo.gl/HSPLWe



Language Rank	Types	Spectrum Ranking
1. Python		100.0
2. C++		98.4
3. C		98.2
4. Java		97.5
5. C#		89.8
6. PHP		85.4
7. R		83.3
8. JavaScript		82.8
9. Go		76.7
10. Assembly		74.5

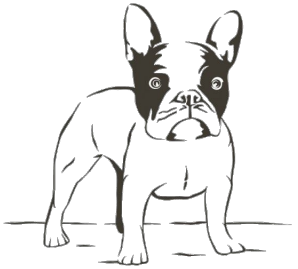
IEEE Spectrum - Jul 2018
https://spectrum.ieee.org/at-work/innovation/the-2018-top-programming-languages

Folium



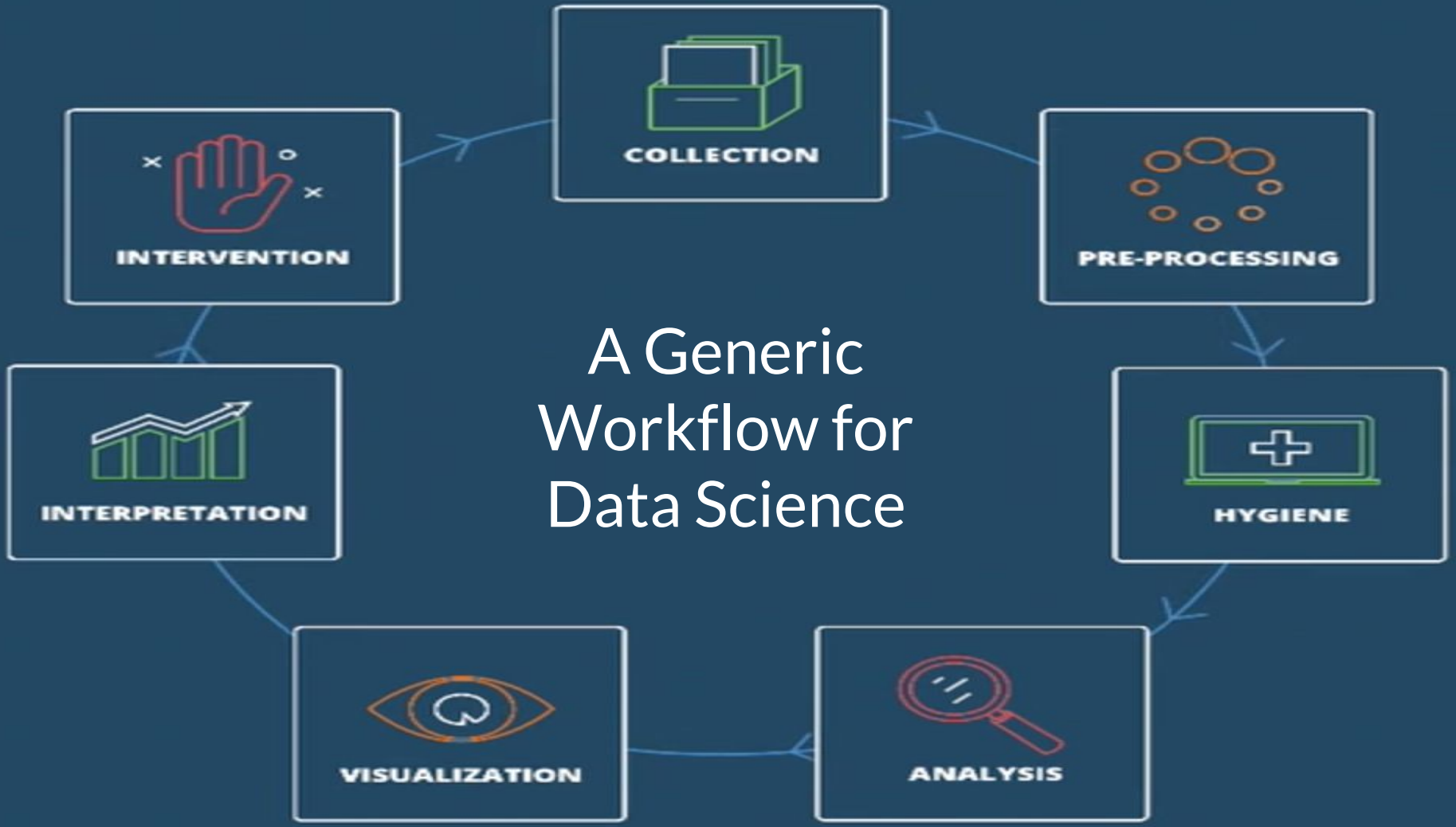
Beautiful Soup

have
a pet
project

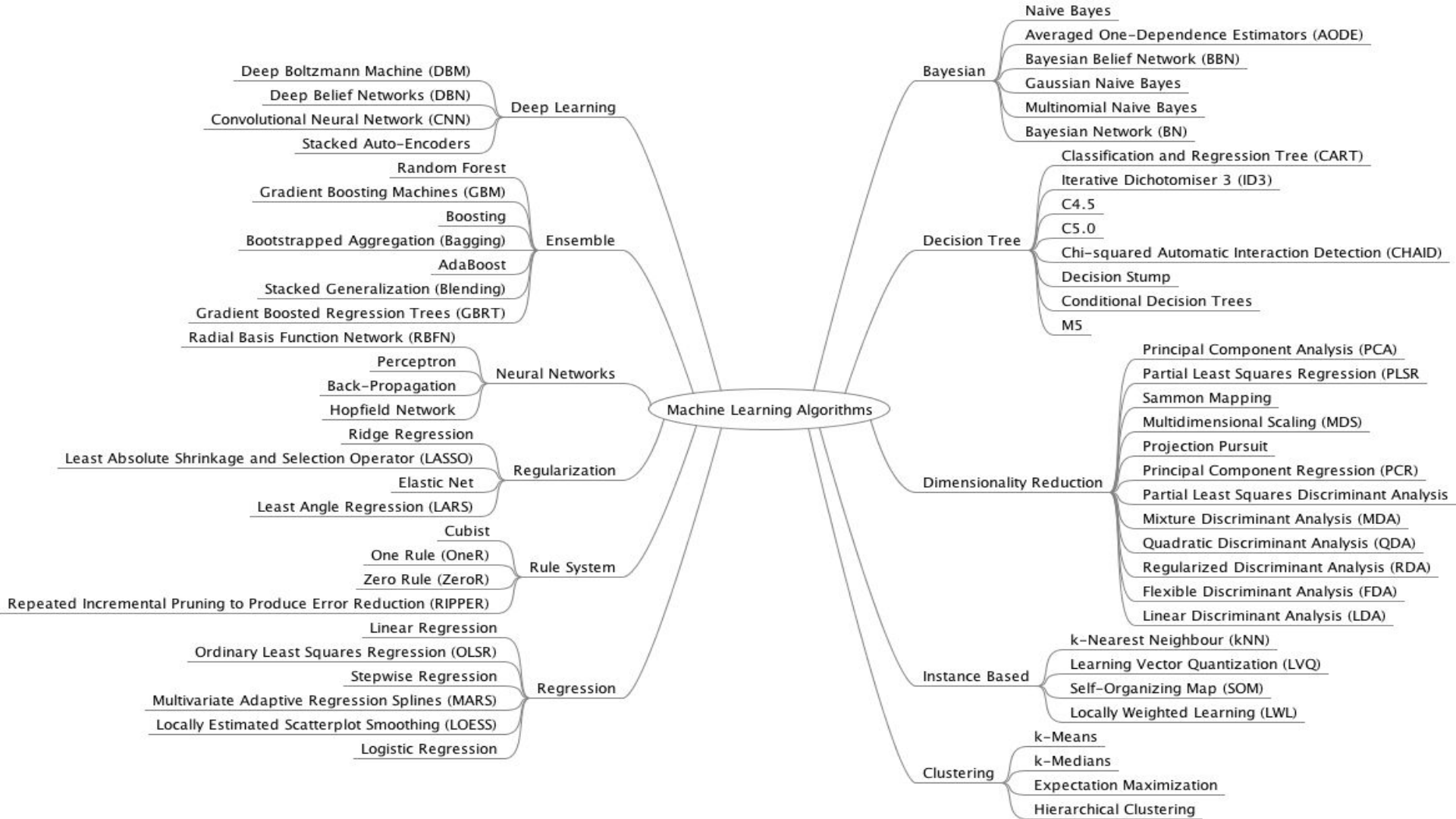


Be clear about your motivation. The reason this is important because learning Data Science is HARD. VERY HARD! So it's easy to lose motivation when on the journey.

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


COLLECTION
PRE-PROCESSING
HYGIENE
ANALYSIS
VISUALIZATION
INTERPRETATION
INTERVENTION
A Generic Workflow for Data Science


How to Become a
ML Engineer


Machine Learning Algorithms
Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)
Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)
Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)
Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering
Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders
Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)
Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network
Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)
Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression

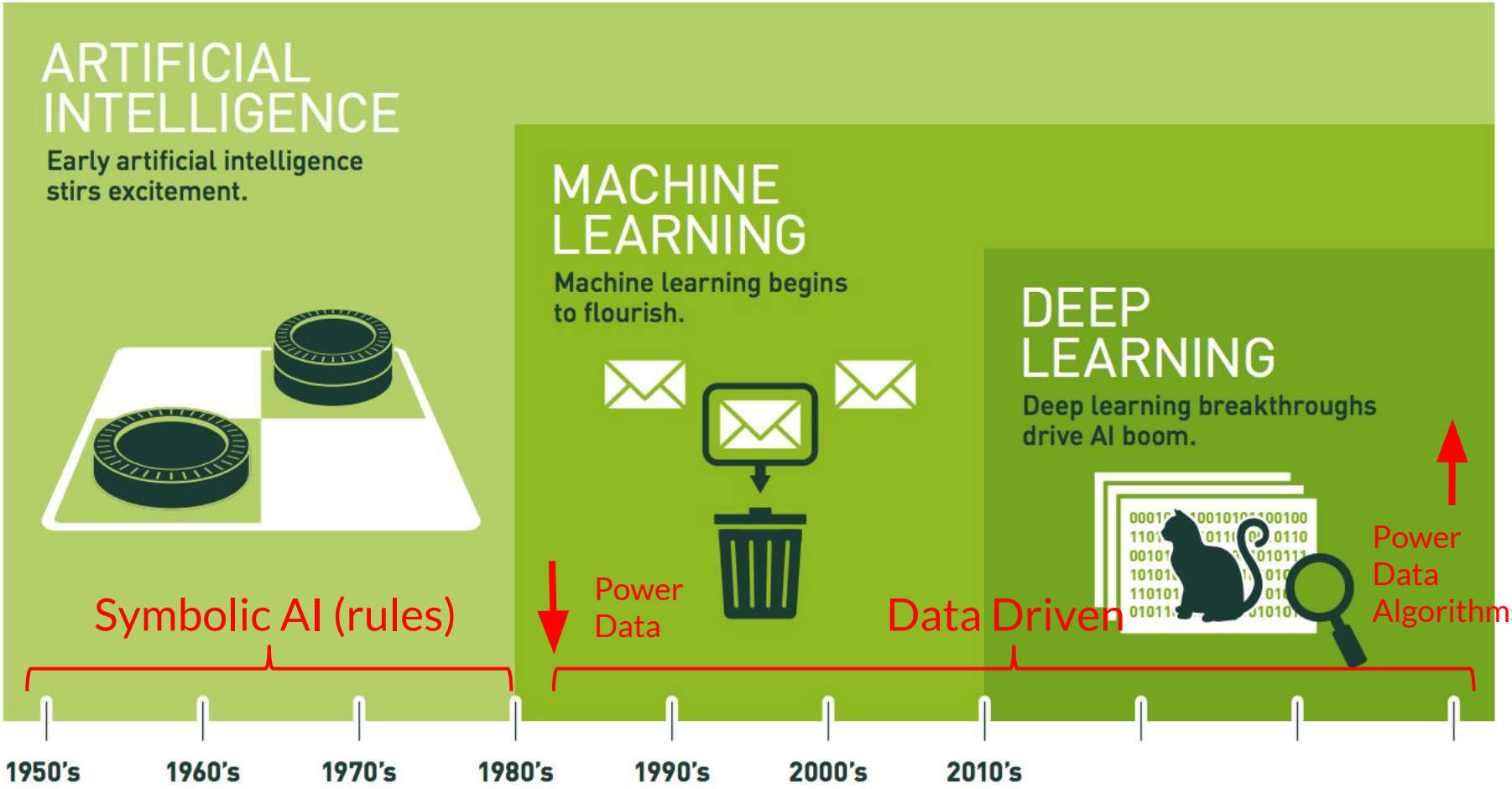
https://github.com/llSourcell/Learn_Machine_Learning_in_3_Months

2-3 hours a day
2x-3x speed
Handwrite notes
1 Project at the end of every week

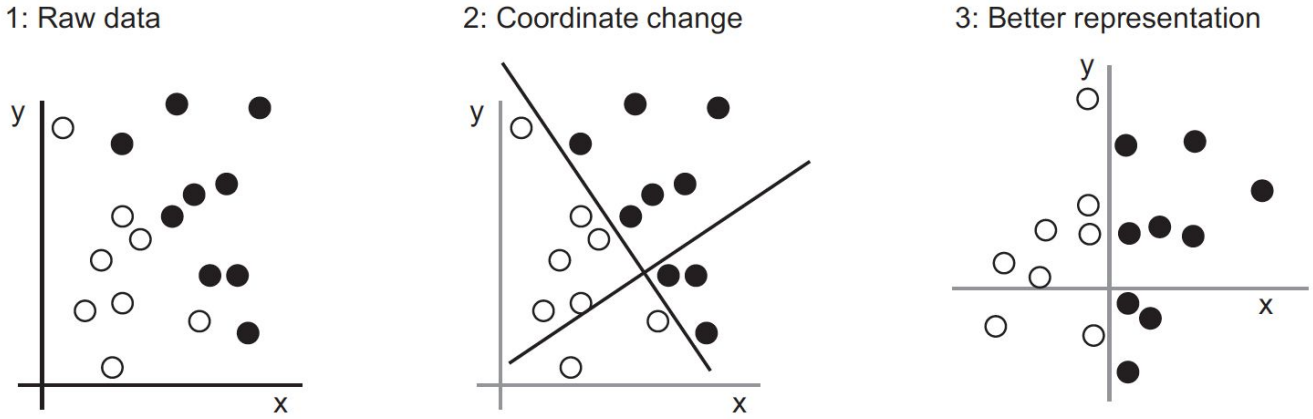
3 MONTH CURRICULUM

Month 1 - Math and Algorithms
Month 2 - Machine Learning
Month 3 - Deep Learning

How do they related to each other?

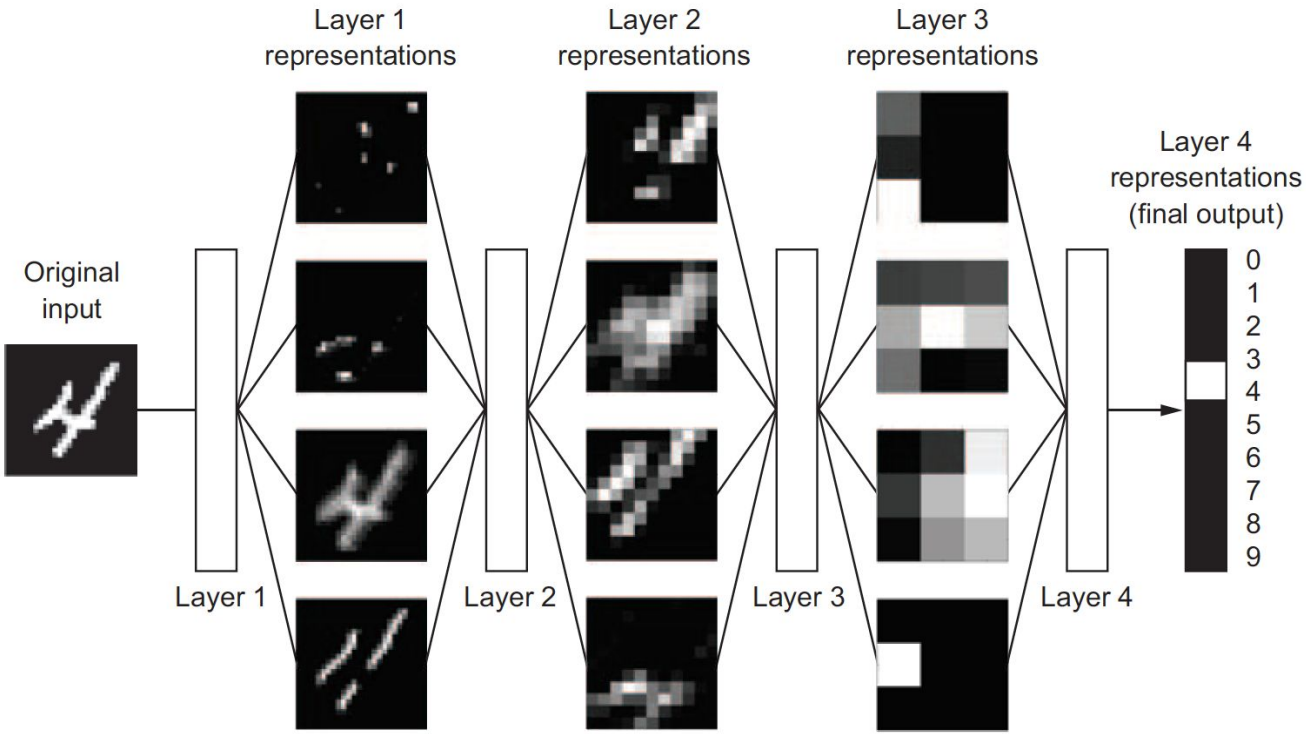


Machine Learning Definition



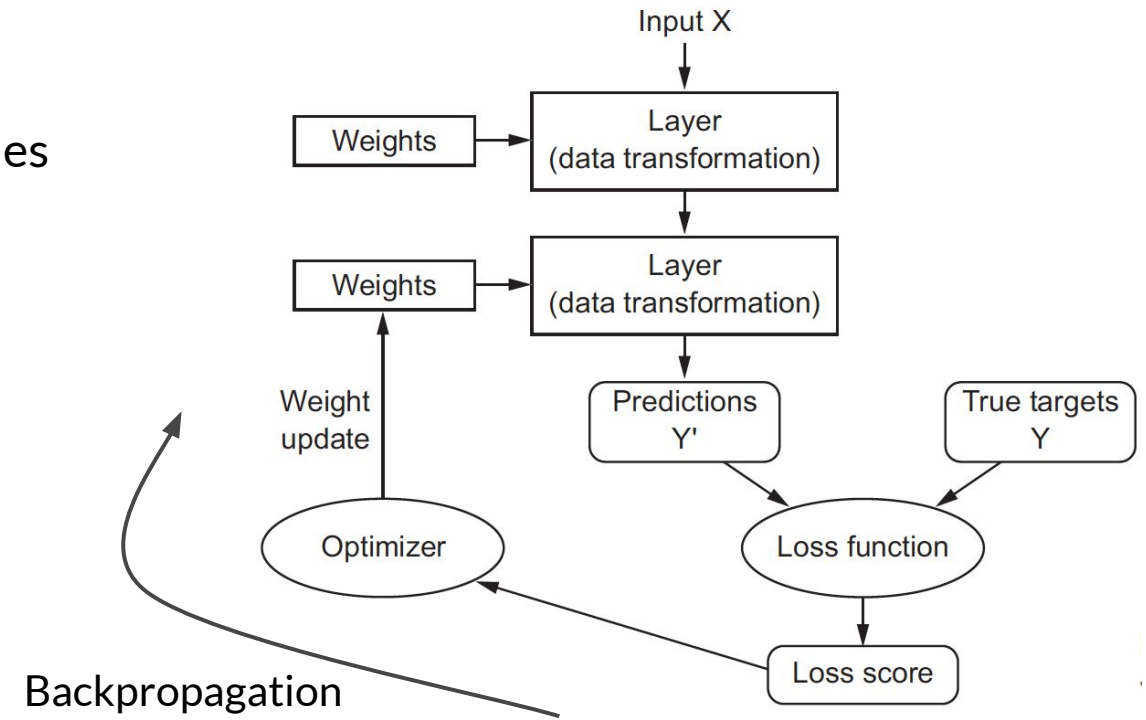
An automatic search process for better data representations

What is Deep Learning?

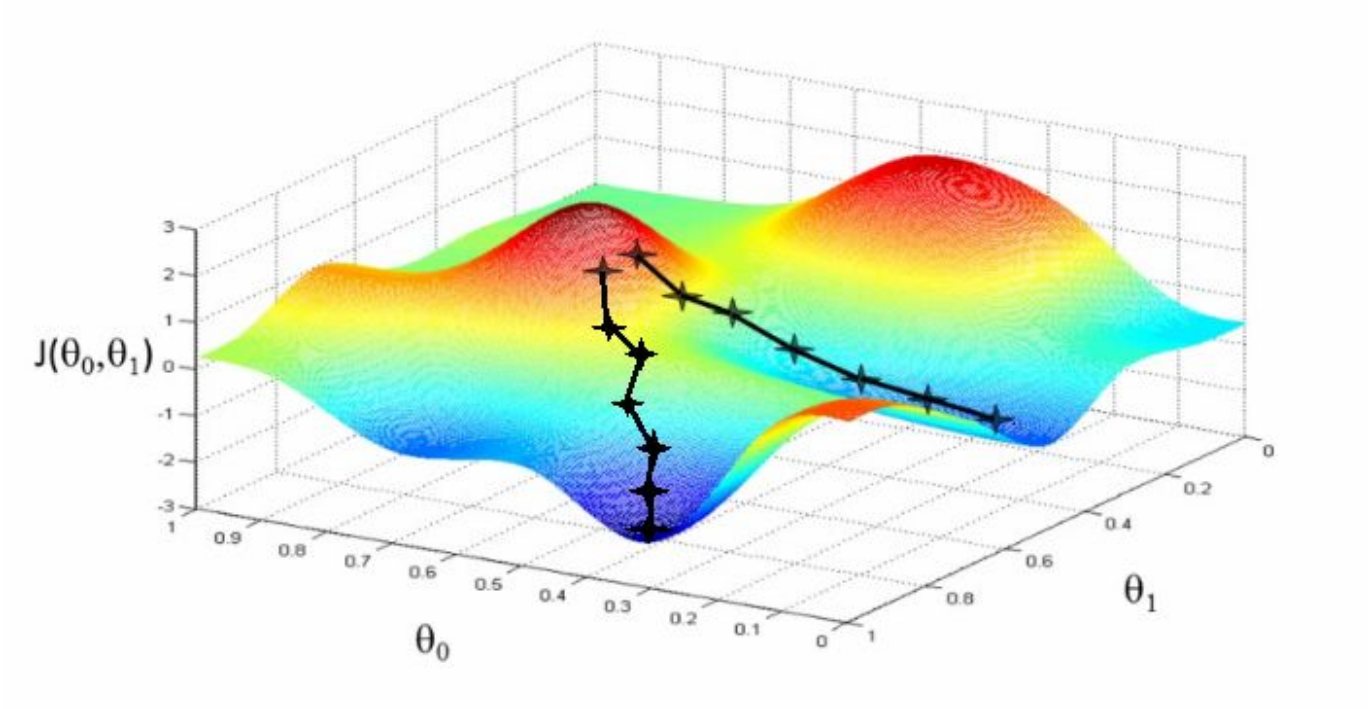


Understanding how DL works

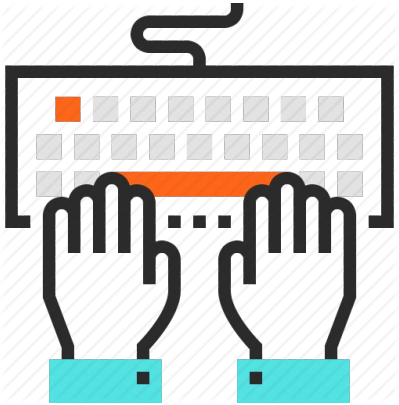
Finding the right values of weights which minimize the error



Minimizing the error - Algorithm Idea



Learning by doing



python

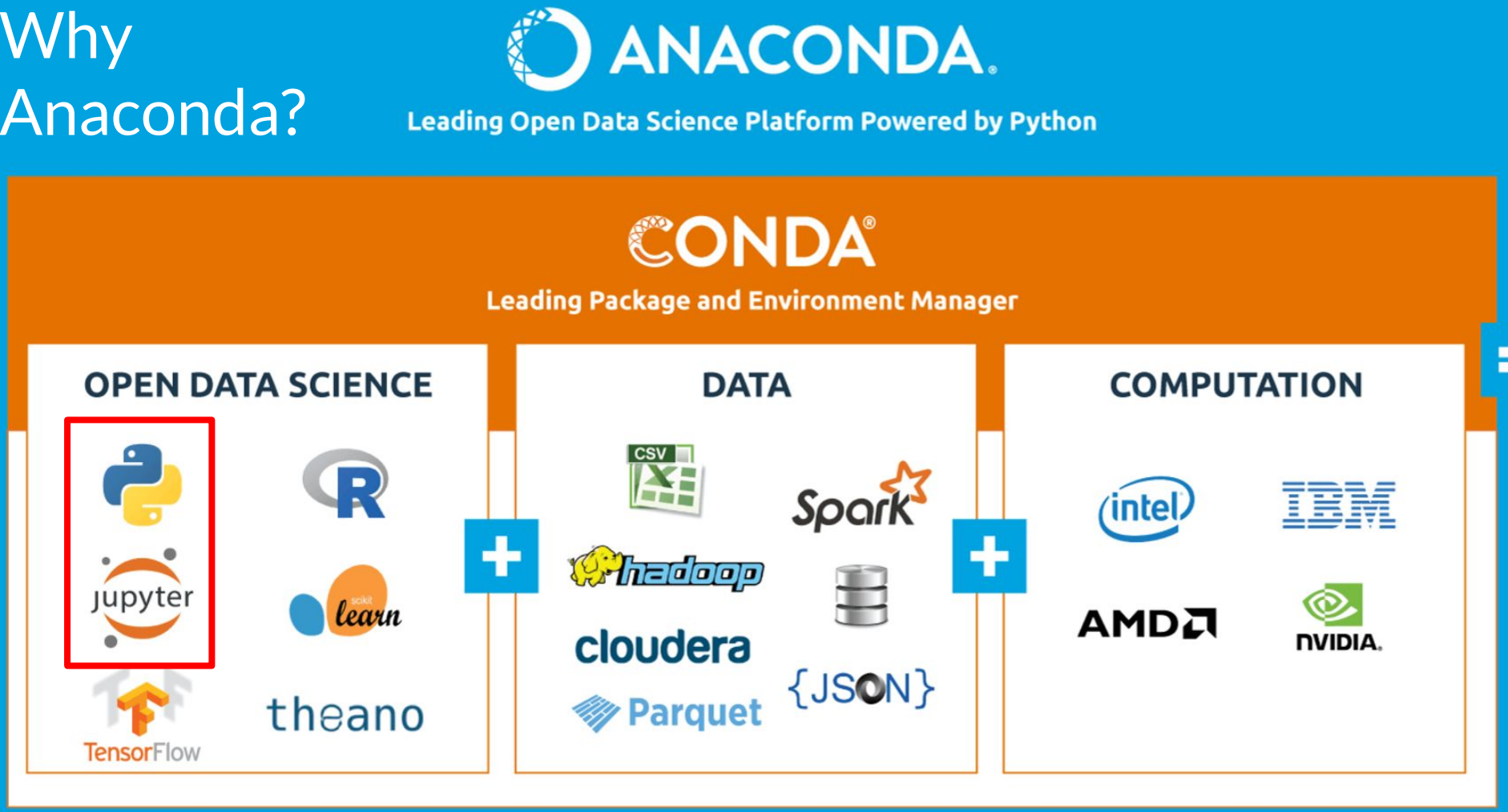
colab

ANACONDA
Powered by Continuum Analytics

git

https://www.continuum.io/downloads

Why Anaconda?



jupyter Untitled (unsaved changes)

File Edit View Insert Cell Kernel Help Python 2













Code CellToolbar

Simple Jupyter demo

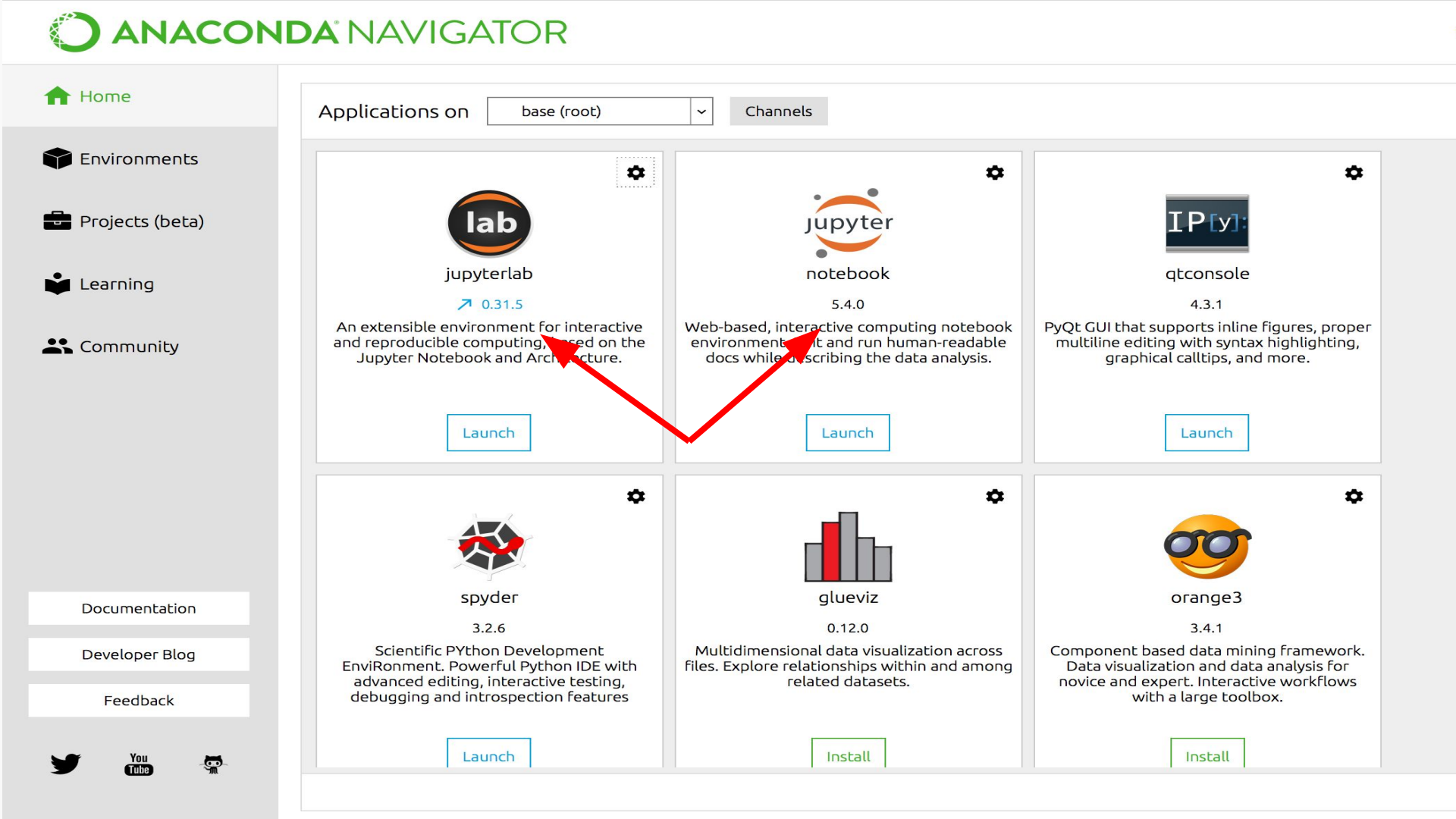
This cell has text formatted using the markdown language, which gets rendered like regular html. The next cell has some code:

In [57]:

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

```
0.10564822904
0.153941700348
0.518503128416
```

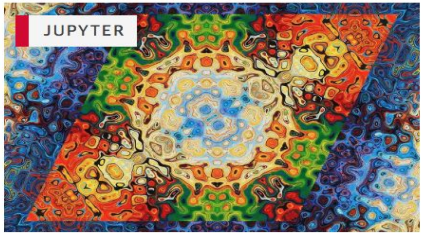
Here is another text cell, with some *formatting*.


ANACONDA NAVIGATOR
Home
Environments
Projects (beta)
Learning
Community
Documentation
Developer Blog
Feedback
Applications on
base (root)
Channels
jupyterlab
0.31.5
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.
Launch
notebook
5.4.0
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.
Launch
qtconsole
4.3.1
PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.
Launch
spyder
3.2.6
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features
Launch
glueviz
0.12.0
Multidimensional data visualization across files. Explore relationships within and among related datasets.
Install
orange3
3.4.1
Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.
Install

JupyterCon 2018

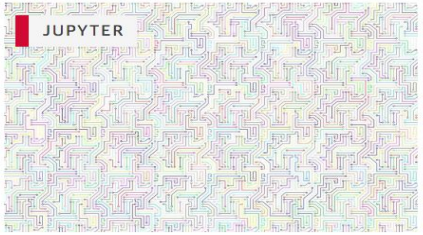
https://www.oreilly.com/tags/jupytercon-2018

Posts tagged: "JupyterCon 2018"



Sea change: What happens when Jupyter becomes pervasive at a university?

Fernando Perez talks about UC Berkeley's transition into an environment where many undergraduates use Jupyter and the open data ecosystem as naturally as they use email.



Beyond interactive: Scaling impact with notebooks at Netflix

Michelle Ufford shares how Netflix leverages notebooks today and describes a brief vision for the future.

Jupyter notebooks and the intersection of data science and data engineering

David Schaaf explains how data science and data engineering can work together to deliver results to decision makers.

The future of data-driven discovery in the cloud

Ryan Abernathey makes the case for the large-scale migration of scientific data and research to the cloud.

Data science as a catalyst for scientific discovery

Michelle Gill discusses how data science methods and tools can link information from different scientific fields and accelerate discovery.

Democratizing data

Tracy Teal explains how to bring people to data and empower them to address their questions.

Disease prediction using the world's largest clinical lab data set

Cristian Capdevila explains how Prognos is predicting disease.

Machine learning and AI technologies and platforms at AWS

Dan Romuald Mbanga walks through the ecosystem around the machine learning platform and API services at AWS.

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

Data 100, ~800 (650 last spring)

Sponsors

Project Jupyter receives direct funding from the following sources:

THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST

ALFRED P. SLOAN FOUNDATION

GORDON AND BETTY MOORE FOUNDATION

Google

rackspace
the #1 managed cloud company

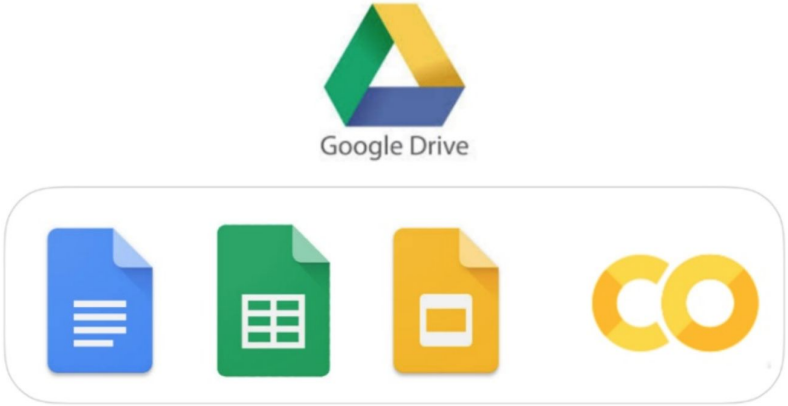
fastly

Horizon 2020
European Union funding
for Research & Innovation

Microsoft

Google Colaboratory

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



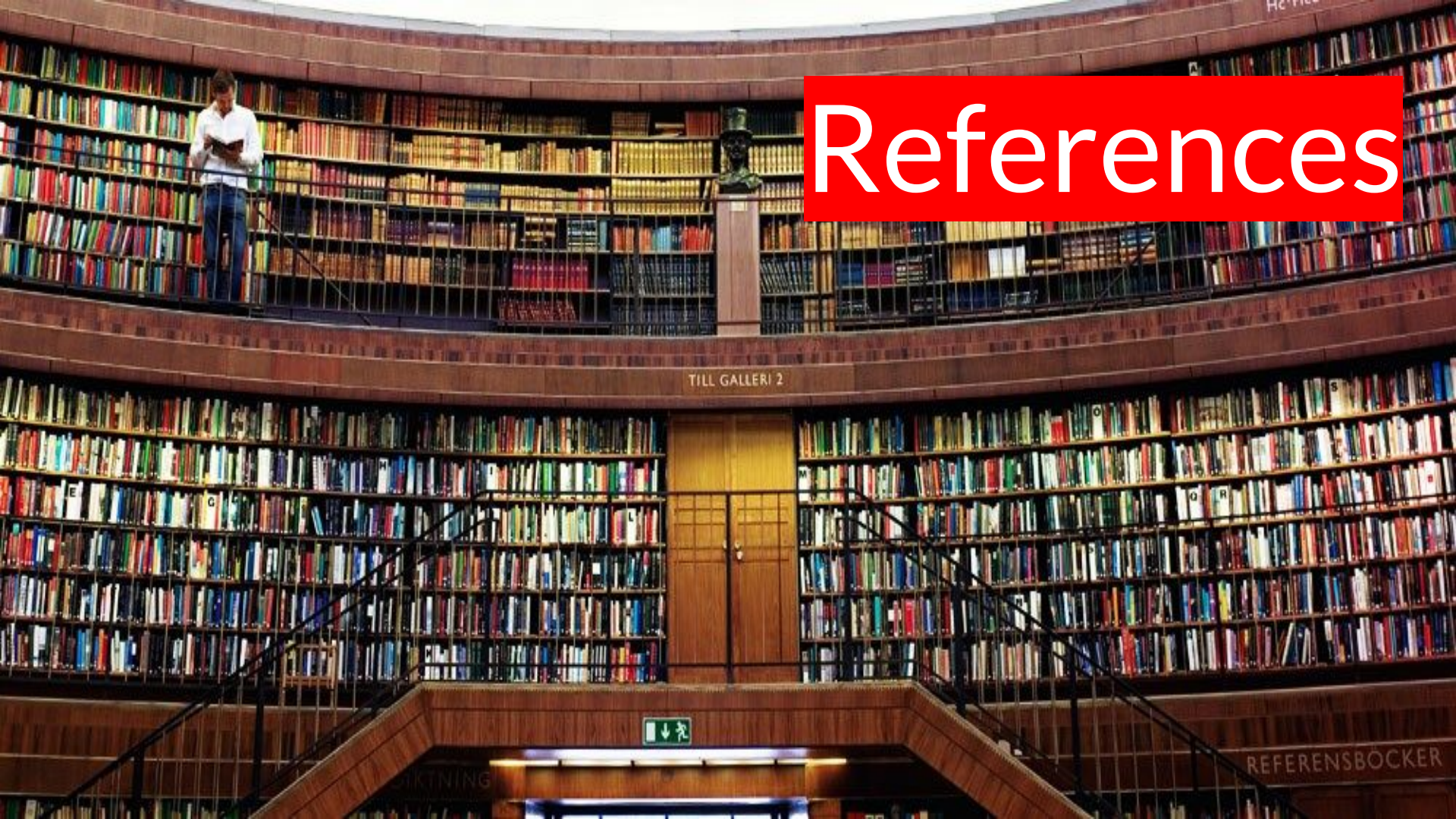
Colaboratory is a Google research project created to help disseminate machine learning education and research. It's a Jupyter notebook environment that requires no setup to use and runs entirely in the cloud.

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

Agenda

https://goo.gl/gk2H1h

https://github.com/ivanovitchm/datascience2machinelearning



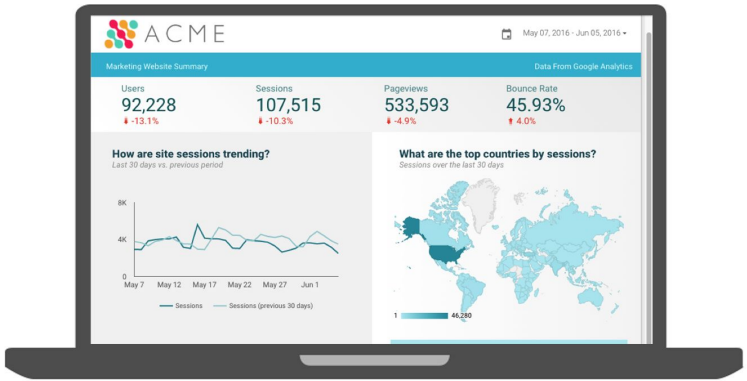
References

http://www.pentaho.com/

https://www.tableau.com/

https://powerbi.microsoft.com

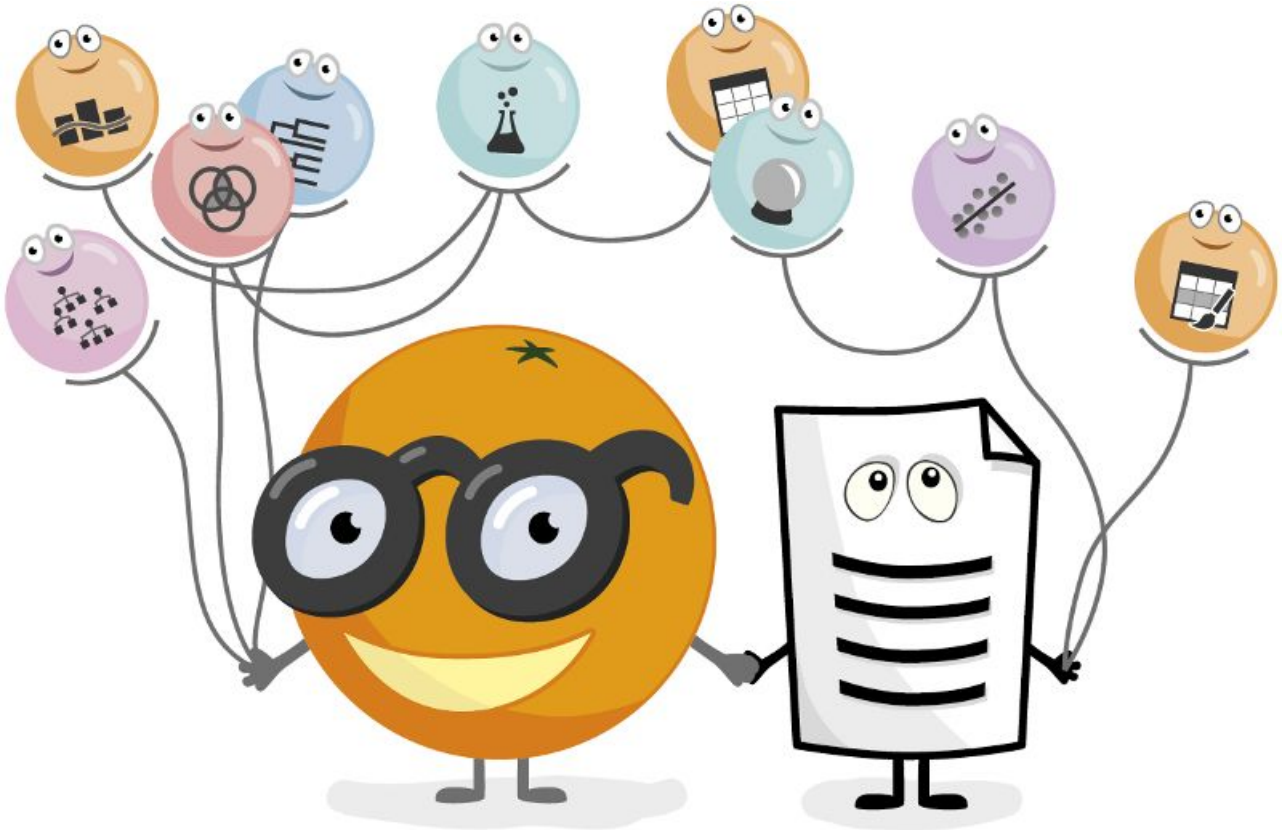
Google Data Studio



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

https://goo.gl/Ndf38Q

Data Mining Fruitful and Fun



https://orange.biolab.si/

References Online

References Online

1. https://www.coursera.org/specializations/big-data
2. https://www.coursera.org/learn/machine-learning
3. https://www.coursera.org/specializations/deep-learning

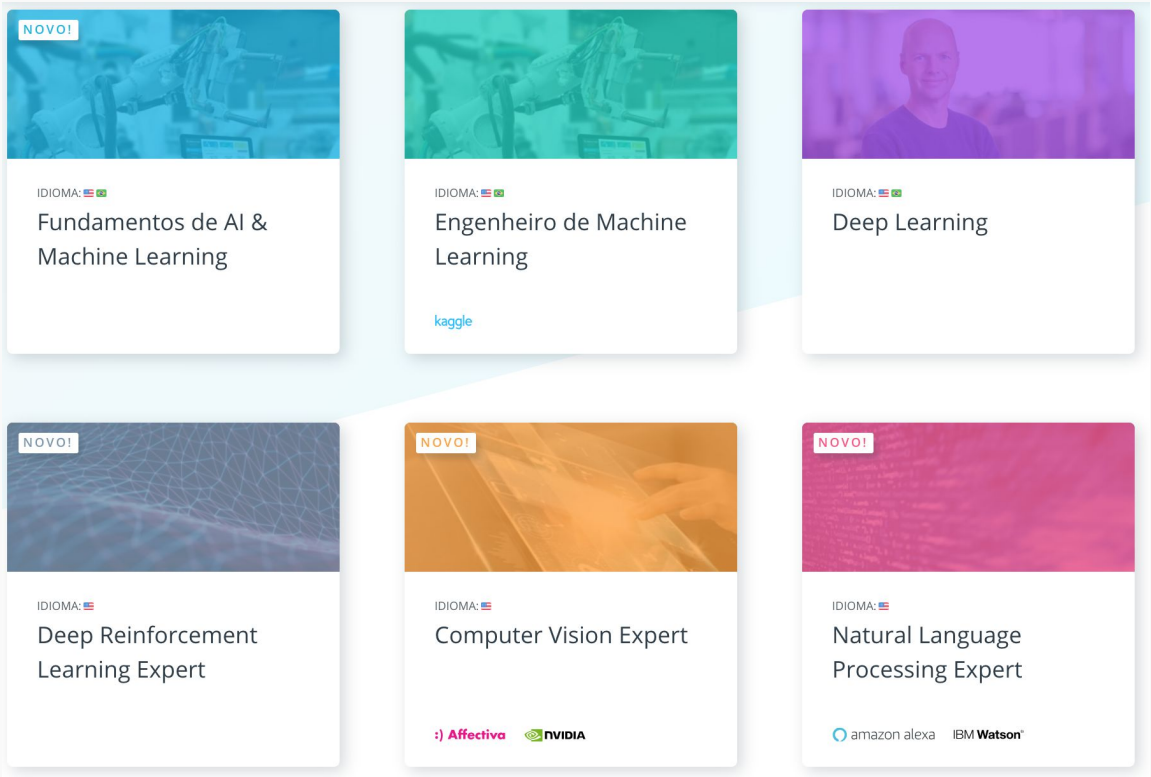
UC San Diego

Stanford University

deeplearning.ai

References Online

UDACITY



The Periodic Table of Data Science

https://goo.gl/eecQ2V

An overview of key companies, resources and tools in data science (as of 4/12/2017)

Symbol → Dc
Name → DataCamp

Courses
Boot camps
Conferences
Data
Projects & Challenges, Competitions
Programming Languages & Distributions
Search & Data Management
Machine Learning & Stats
Data Visualization & Reporting
Collaboration
Community & Q&A
News, Newsletters & Blogs
Podcasts

Dc DataCamp	Ga General Assembly	Sd Strata Data												
Sb SpringBoard	M Metis	Od ODSC												
Ex Edx	Di Data Incubator	Tc Tableau Conference												
C Coursera	In Insight	U UseR!												
Uda Udacity	Dsa NYC Data Science Academy	Pd PyData												
Ude Udemy	G Galvanize	Paw Predictive Analytics World												
Ps Pluralsight	Dsg Data Science for Social Good	Kdd ACM SIGKDD Conference												
Ly Lynda	Dsy Data Society	Tpc Teradata Partners Conference												
Tt TeamTreeHouse	Dsj Data Science Dojo	Icd IEEE International Conference on Data Mining												
Bdu Big Data University														

Py Python	Js JavaScript	Vb Visual Basic	Pgs PostgreSQL	Sli SQLite	Ah Apache Hadoop	W Weka	Bml BigML	Kn Knime	Sm Spark MlLib	Pb Power BI	Obi Oracle BI	Shn Shiny	Ddl Domino Data Lab	De Data Science Experience
R R	Cp C++	Sc Scala	Ar Amazon Redshift	Bq Google BigQuery	Hw Hortonworks	O Oracle	Dar DataRobot	Lib LibSVM	Ho H2O	Bo BusinessObjects	Alt Alteryx	Mpl Matplotlib	Nt Nteract	Rs Rstudio
S SQL	Pl Perl	Ca Cassandra	Hb HBase	Td Teradata	Cl Cloudera	Mss Microsoft SQL server	Rm RapidMiner	Mat Mathematica	Th Theano	Sp Spotfire	Sav SAS Visual Analytics	Ply Plotly	Ro Rodeo	Be Beaker Notebook
B Bash	Mr Microsoft R Open	P Pig	Mdb Mongo DB	To Toad	Aem Amazon Elastic Mapreduce	Spl Splunk	Cho Chorus	Mah Mahout	Aml Azure Machine Learning	Ql Qlikview	Po PowerPivot	Me Microsoft Excel	Spy Spyder	Ze Apache Zeppelin
Mtl Matlab	Cy Canopy	Im Impala	K Kafka	Ms MySQL	Mar MapR	Sr Solr	Tf Tensorflow	St Stata	D D3	Co Cognos	Gch Google Charts	Pe Pentaho	Dst Data Science Studio	Ju Jupyter
J Java	An Anaconda	Sp Spark	Hi Hive	Idb IBM DB2	Lu Lucene	El ElasticSearch	Sk Scikit-Learn	Da Dato/Graphlab	My Microstrategy	Aa Adobe Analytics	T Tableau	B Bokeh	Db Databricks notebook	Gh Github

Dw Data.world	Q Quandl	Fte FiveThirtyEight	Sa Socrata	Gp Google Public	Dg Data.gov	K Kaggle
St Statista	Uci UCI Machine Learning Repository	Wb World Bank	At Academic Torrents	Bf Buzzfeed	Dk DataKind	Dd DrivenData

Re Reddit	So Stack Overflow	Cv Cross Validated	Qu Quora	Av Analytics Vidhya	Dse Data Science Stack Exchange
Mu Meetup	Rdm RDataMining				

Kdn KDnuggets	Ibd insideBIGDATA
Rb R-Bloggers	Pp PlanetPython
Hn HackerNews	Dt DataTau
Dsc Data Science Central	Dsr Data Science Roundup
Dsw Data Science Weekly	Or O'Reilly
Dr Data Elixir	Pw Python Weekly
Rw R Weekly	Pd Partially Derivative
Bds Becoming a Data Scientist	Tm Talking Machines
Ds Data Stories	Dsk Data Skeptic
Ld Linear Digressions	Ns Not So Standard Deviations

EVOLUTION

MAPS

FLOW

Other

THE PYTHON GRAPH GALLERY

https://python-graph-gallery.com/

DISTRIBUTION

CORRELATION

RANKING

PART OF A WHOLE

Reference Blog

https://machinelearningmastery.com/start-here/#process

Do You Need Help Getting Started with Applied Machine Learning?

This is The Step-by-Step Guide that You've Been Looking For!

Hi, Jason here. I'm the guy behind Machine Learning Mastery.

My goal is to help you **get started**, **make progress** and **kick butt** with machine learning.

I teach a **top-down** and **results-first approach** designed for developers and engineers.
This is unlike most academic textbooks and university courses.

Access my best free tutorials on the blog or take the next step with my paid training material.

You may be **feeling overwhelmed**. You may have **a lot of questions**.
I created this page for you. It is your starting point.

Take your time. Bookmark this page. Find the answers to your questions.

https://www.nature.com/articles/d41586-018-06201-x

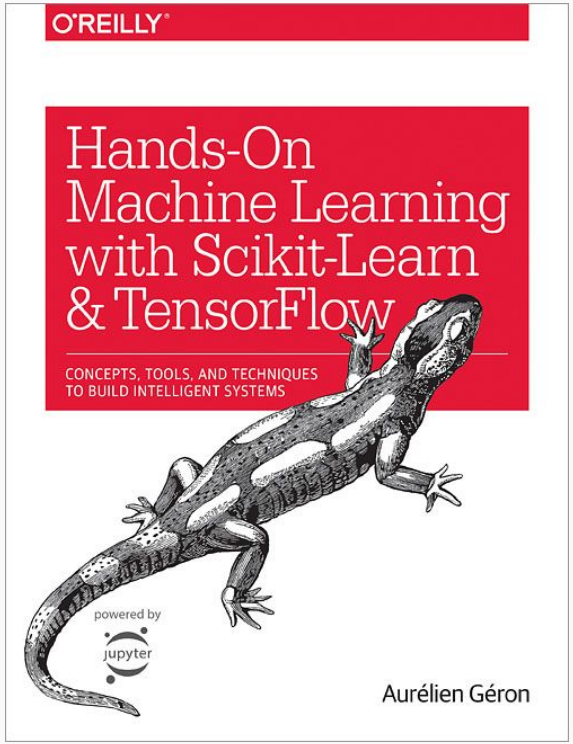
NEWS · 05 SEPTEMBER 2018

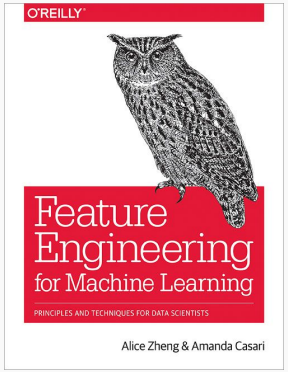
Google unveils search engine for open data

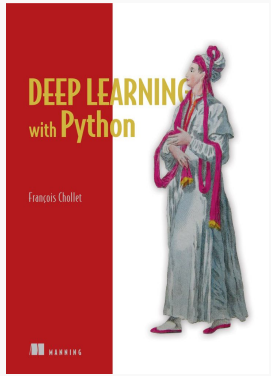
The tool, called Google Dataset Search, should help researchers to find the data they need more easily.

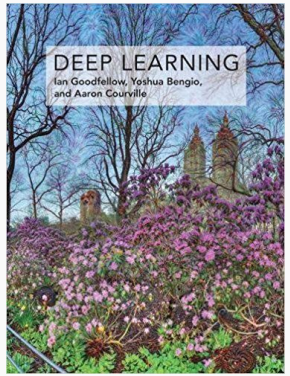
https://toolbox.google.com/datasetsearch

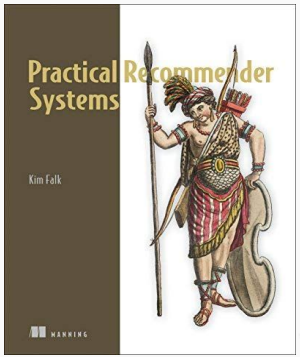
References


O'REILLY
Hands-On Machine Learning with Scikit-Learn & TensorFlow
CONCEPTS, TOOLS, AND TECHNIQUES TO BUILD INTELLIGENT SYSTEMS
powered by
jupyter
Aurélien Géron

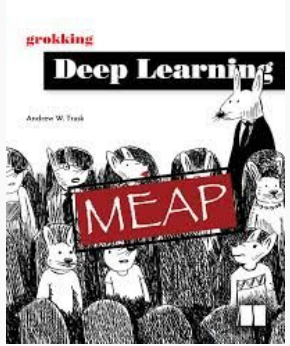

O'REILLY
Feature Engineering for Machine Learning
PRINCIPLES AND TECHNIQUES FOR DATA SCIENTISTS
Alice Zheng & Amanda Casari


DEEP LEARNING with Python
François Chollet
MANNING

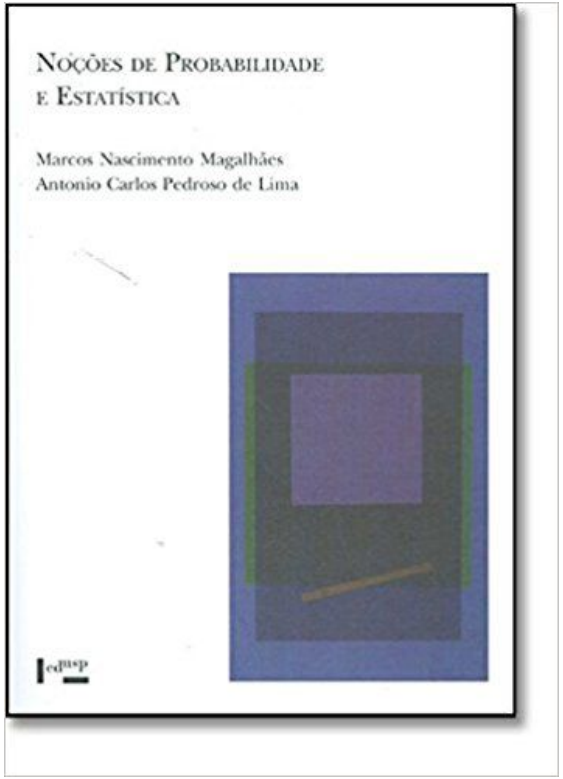

DEEP LEARNING
Ian Goodfellow, Yoshua Bengio, and Aaron Courville

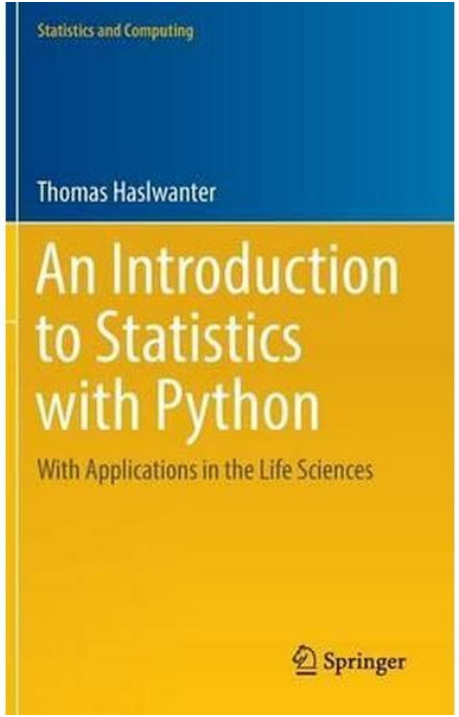

Practical Recommender Systems
Kim Falk
MANNING


DEEP LEARNING FOR COMPUTER VISION
WITH PYTHON
Dr. Adrian Rosebrock
pyimagesearch


grokking
Deep Learning
Andrew W. Trask
MEAP

Referências





“AI is the new electricity … We have enough papers. Stop publishing, and start transforming people’s lives with technology”

Andrew Ng