Big Data - Foundations and Applications

Lesson #1 - Outline & Directions

Ivanovitch Silva

July, 2017


Introduction

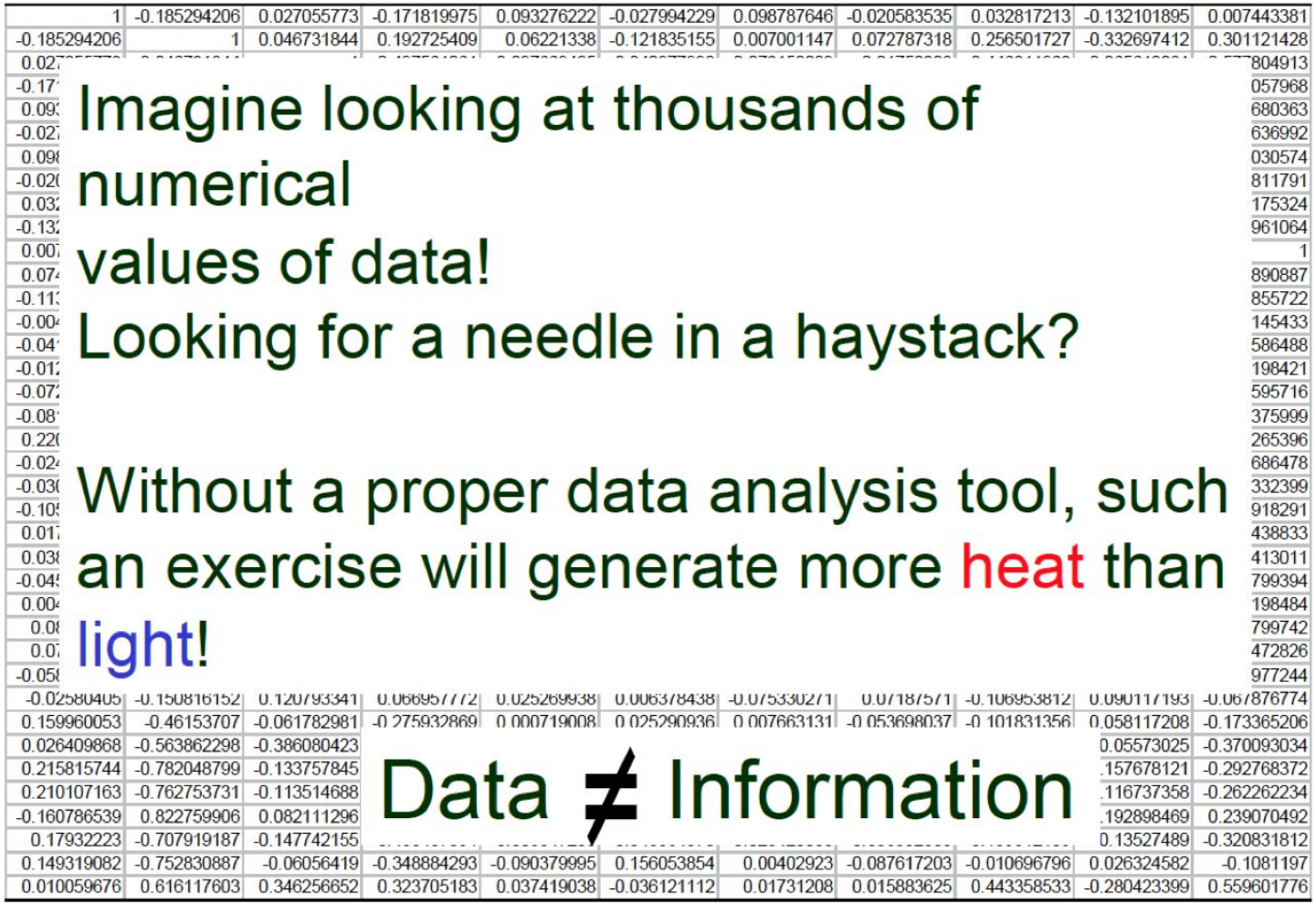
Ivanovitch Silva (ivan@imd.ufrn.br)

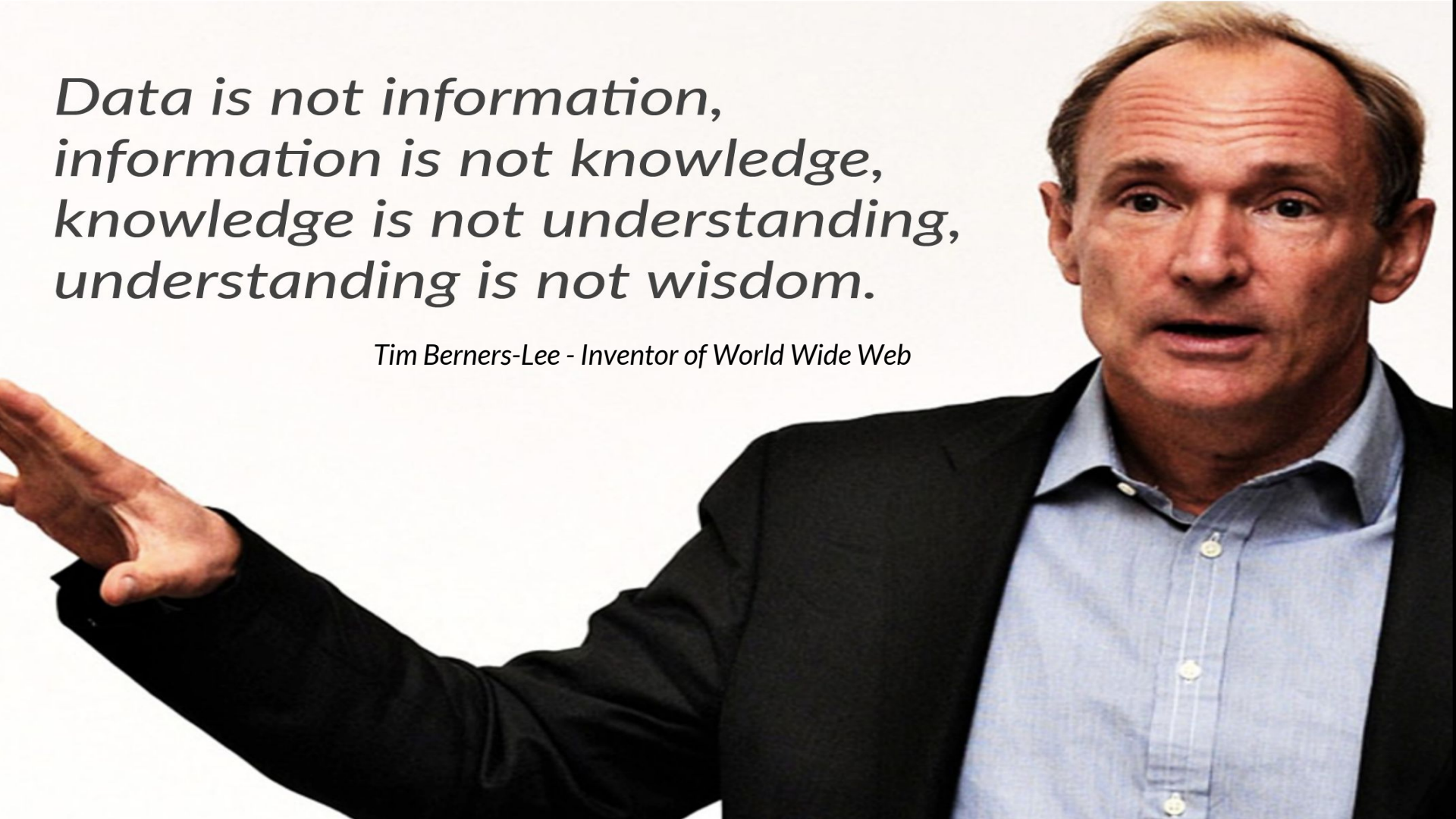
Group Knowledge

Take a Survey

https://goo.gl/forms/1E99HHkEh8ZmMzOJ3


About Data


Imagine looking at thousands of numerical values of data!
Looking for a needle in a haystack?
Without a proper data analysis tool, such an exercise will generate more heat than light!
Data ≠ Information


Data is not information,
information is not knowledge,
knowledge is not understanding,
understanding is not wisdom.
Tim Berners-Lee - Inventor of World Wide Web

BIG DATA

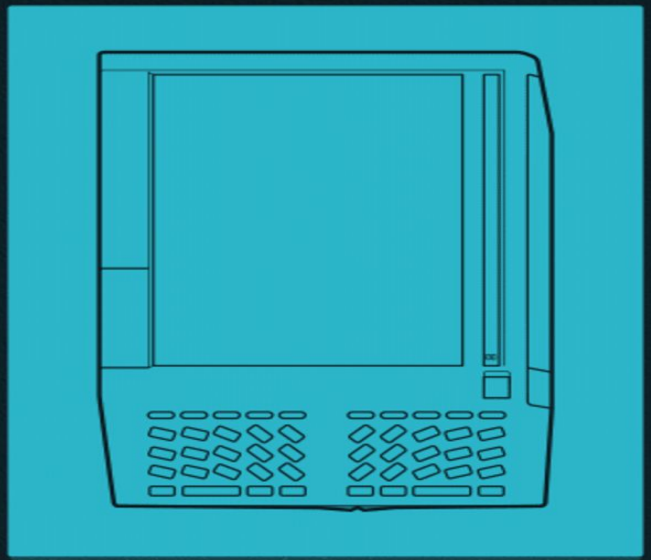
BIG NUMBERS


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

DIGITAL ANALOGUE ORGANIC

MEGABYTES

1 MEGABYTE – APPROXIMATELY 1,000 KILOBYTES
(ACTUALLY 1,024 KB)



256MB
Kindle
first generation

0.004MB
Apple I
RAM in Apple's first computer, 1976

0.004
Oyster card
London public transport

0.02
Punched paper tape
largest feasible reel

0.02
Tandy 200 computer
amount of RAM, 1984

0.02
Word document
single page

0.7
Audio cassette
90 min

2.5
War & Peace
Kindle ebook

1.5
3.5" Floppy disk

1
ebook
average size

5
William Shakespeare
complete works

9
Blu-Ray
1 sec at HDTV quality

66
Mosquito genome
DNA for malaria mosquito

20
10,000 pages of text
about 20 novels

https://goo.gl/CHQFds

CHANGE OF SCALE 10:1

GIGABYTES

1 GIGABYTE – APPROXIMATELY 1,000 MEGABYTES (ACTUALLY 1,024 MB)

0.25GB
Kindle
first generation

0.47GB
Large Hadron Collider
data produced per sec

0.48
YouTube
video uploaded per sec, 2012

0.51
Raspberry Pi
Model B

0.54
Blu-ray
1 min at HDTV quality

0.76
Single human sperm
all DNA

0.76
Single human egg
all DNA

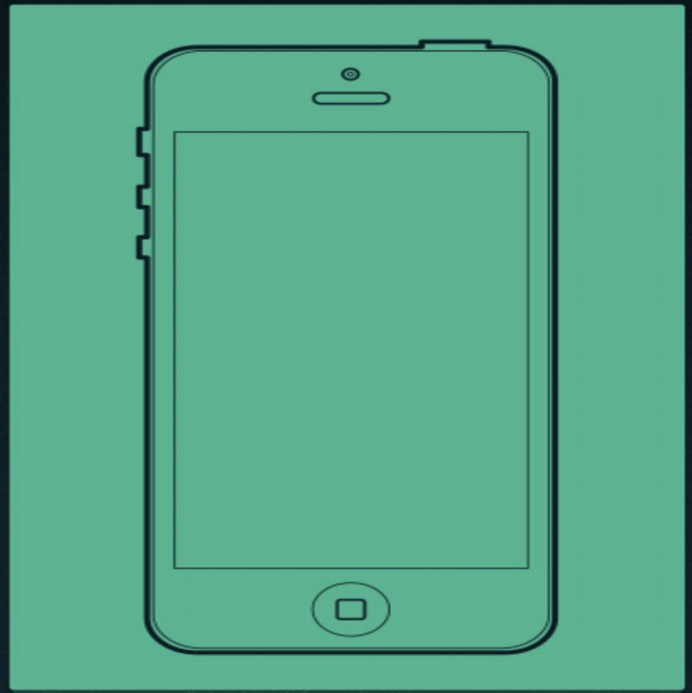
0.7
CD
80 mins

1
MiniDisc
45 hrs of music

1.5
Human body cell
all DNA

0.07
coding DNA only

2
Kindle Paperwhite
total storage



32GB
iPhone 5
total storage

CHANGE OF SCALE 10:1

3GB
Facebook
photos and videos stored per sec, 2012

4
Kindle Touch

32GB
iPhone 5
total storage

4k digital cinema

2.3
1 min normal frame rate

4.6
1 min high frame rate

7
Wikipedia
all current articles without edit history

8.5
DVD
dual-layer

20
Wolfgang Amadeus Mozart
complete works as MP3

26.4
Mad Men
one episode

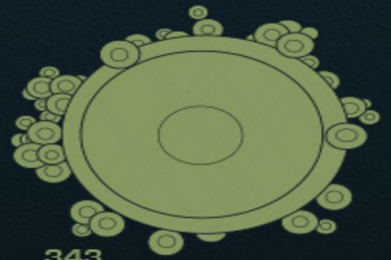
28.5
Large Hadron Collider
data produced per min

29.2
YouTube
video uploaded per min, 2012

128
Blu-ray
max disc capacity

1,760GB
Human sperm
DNA created per man, per sec

778
The Hobbit
4K digital cinema high frame rate



343
Eggs per woman
at birth - about 450

CHANGE OF SCALE 10:1

TERABYTES

1 TERABYTE = APPROXIMATELY 1,000,000 MEGABYTES (ACTUALLY 1,048,576 MB)

1TB
Average modern hard disk

1.8TB
Human sperm
DNA created per man, per sec

1.3
Human brain
functional memory capacity

1
1 million novels
500 million pages of text

7
single DNA sequencing run
data from end-to-end human DNA sequencing

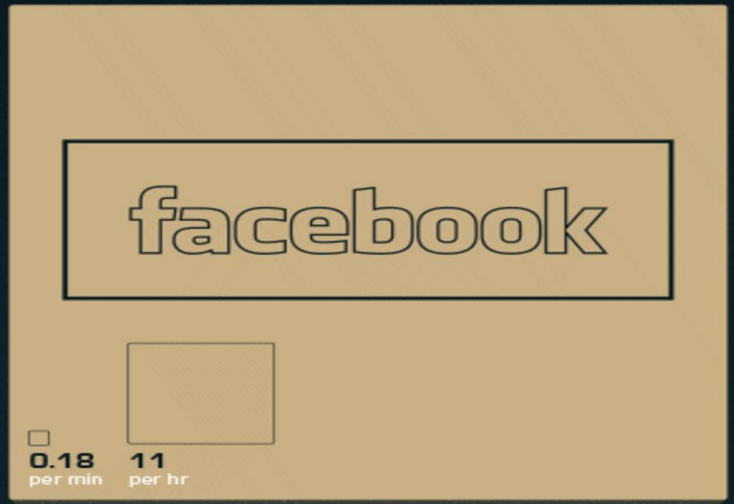
7.3
Wikipedia
all current articles with edit history

10
US Library of Congress
printed collection

Internet traffic

6.3
per sec, 2012

12
for all of 1990



0.18
per min

11
per hr

274TB
Facebook
photos and videos stored per day, 2012

CHANGE OF SCALE

CHANGE OF SCALE 10:1

42TB
Amazon.com
database

47
Human ear
DNA in audio hair cells

274TB
Facebook
photos and videos
stored per day, 2012

120
Internet traffic
for all of 1993

304
Human eye
light receptors
per sq mm

651
Eagle's eye
light receptors
per sq mm

900
Mobile internet
traffic per month, 2005

Google

20,000TB
Google
data processed
per day 2008

2,220
Synthetic DNA data
storage capacity 1 gram

2,000
All US academic research libraries
printed collection

1,800
Emails sent globally
per day, 2002

CHANGE OF SCALE 10:1

PETABYTES

1 PETABYTE = APPROXIMATELY 1,000,000,000 MEGABYTES (ACTUALLY 1,073,741,824 MB)

20PB
Google
data processed per day 2008

3.5PB
Radio programming
globally 2002

16
including repeats

6
World's largest climate data archive
NOAA National Climatic Data Center

15
Large Hadron Collider
data produced per year

22.8
Internet traffic
for all of 1996

15.4
YouTube
video uploaded in 2012

40
Titan supercomputer
storage capacity

45.7
Human nose
smell receptor neurons

47.4
Telephone calls
globally per day, 2002

70
TV programming
globally 2002

100
Facebook
photos and videos stored in 2012

152
Human sperm
created per man, per day

211
Human male ejaculation

2,400PB
Human skin cells shed in a month

DIGITAL ANALOGUE ORGANIC

EXABYTES

1 EXABYTE = APPROXIMATELY 1,000,000,000,000 MEGABYTES
(ACTUALLY 1,099,511,627,776 MB)

CHANGE OF SCALE 10:1

330EB
Internet traffic
for all of 2011

2.4EB
human skin cells shed in a month

0.2EB
Human eye
light receptors in retina

0.2
All printed material on Earth

0.27
TV programming
globally including repeats, 2002

0.59
Mobile internet
traffic per month, 2011

0.66
Emails sent
globally 2002

Internet traffic

0.5
per day
2012

1
per month,
2004

4.9
for all of 2002

1.5
Dog's nose
smell receptor neurons

56
Human sperm
created per man, per year

17.3
Telephone calls
globally for all of 2002

CHANGE OF SCALE 10:1

ZETTABYTES

1 ZETTABYTE = APPROXIMATELY 1,000,000,000,000,000 MEGABYTES
(ACTUALLY 1,125,899,906,842,620 MB)

YouTube
Broadcast Yourself

0.01ZB
YouTube
hours watched
for all of 2012

0.3ZB
Internet traffic
for all of 2011

0.1
for all of 2008

0.02
for all of 2005

0.05
Neurons in human brain

All global data

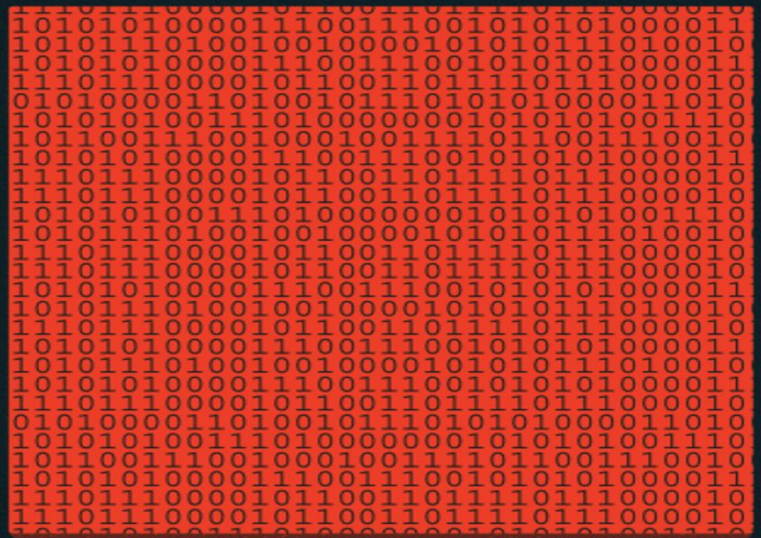
0.5
2008

0.8
2009

1.2
2010

1.8
2011

2.8
2012



30ZB
All global data
2019 (predicted)

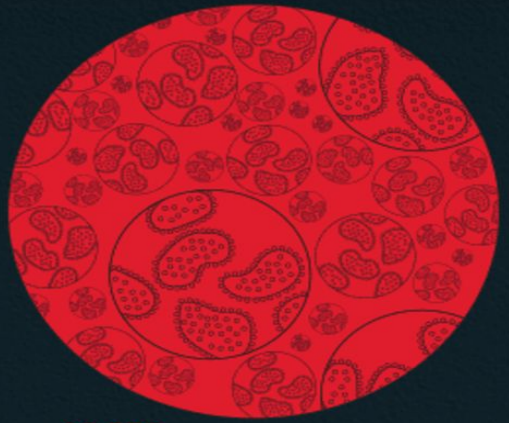
YOTTABYTES

1 YOTTABYTE = APPROXIMATELY 1,000,000,000,000,000,000,000 MEGABYTES
(ACTUALLY 1,152,921,504,606,850,000 MB)

CHANGE OF SCALE 1000:1

0.03YB
All global data
2019 (predicted)

0.04
All words ever spoken
digitized as 16 kHz 16-bit audio



10,000
All microbes on Earth
unique genetic information

0.09
All cells in human body
duplicated genetic information

NOTES

Figures are based on decimal not binary file sizes.
So 1 megabyte = 1,000 kilobytes, not 1,024 kilobytes.

Most organic figures refer to genetic data and are based on multiplying DNA in single cell by number of cells. DNA in two cells is therefore counted twice, though cells within an organism are genetically exact or near-exact copies of one another.

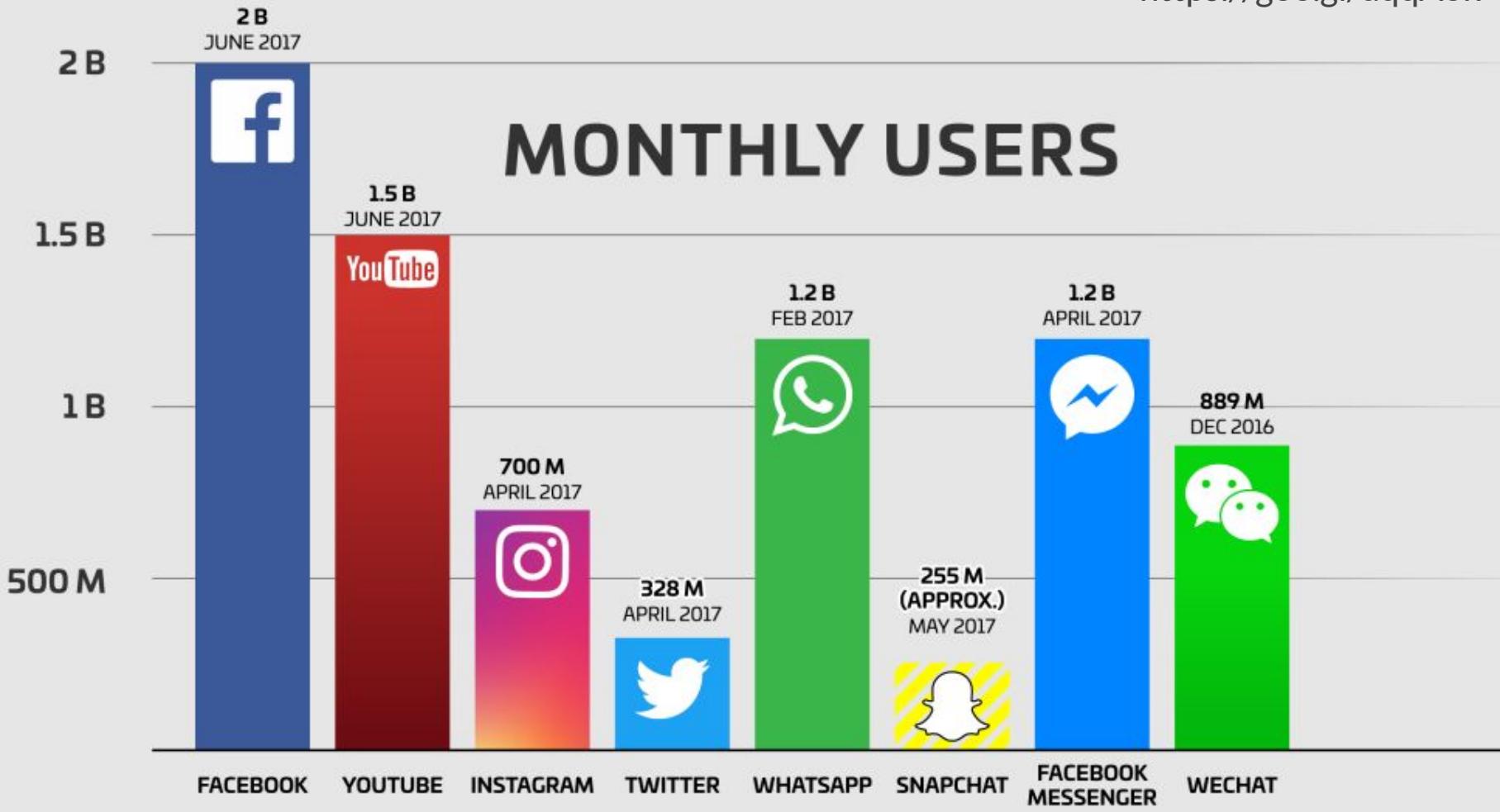
By Information is Beautiful Studio for BBC FUTURE

Executive Creative Director David McCandless *Creative Director* Duncan Swain
Design Matt McLean *Research* Miriam Quick, Christian Miles

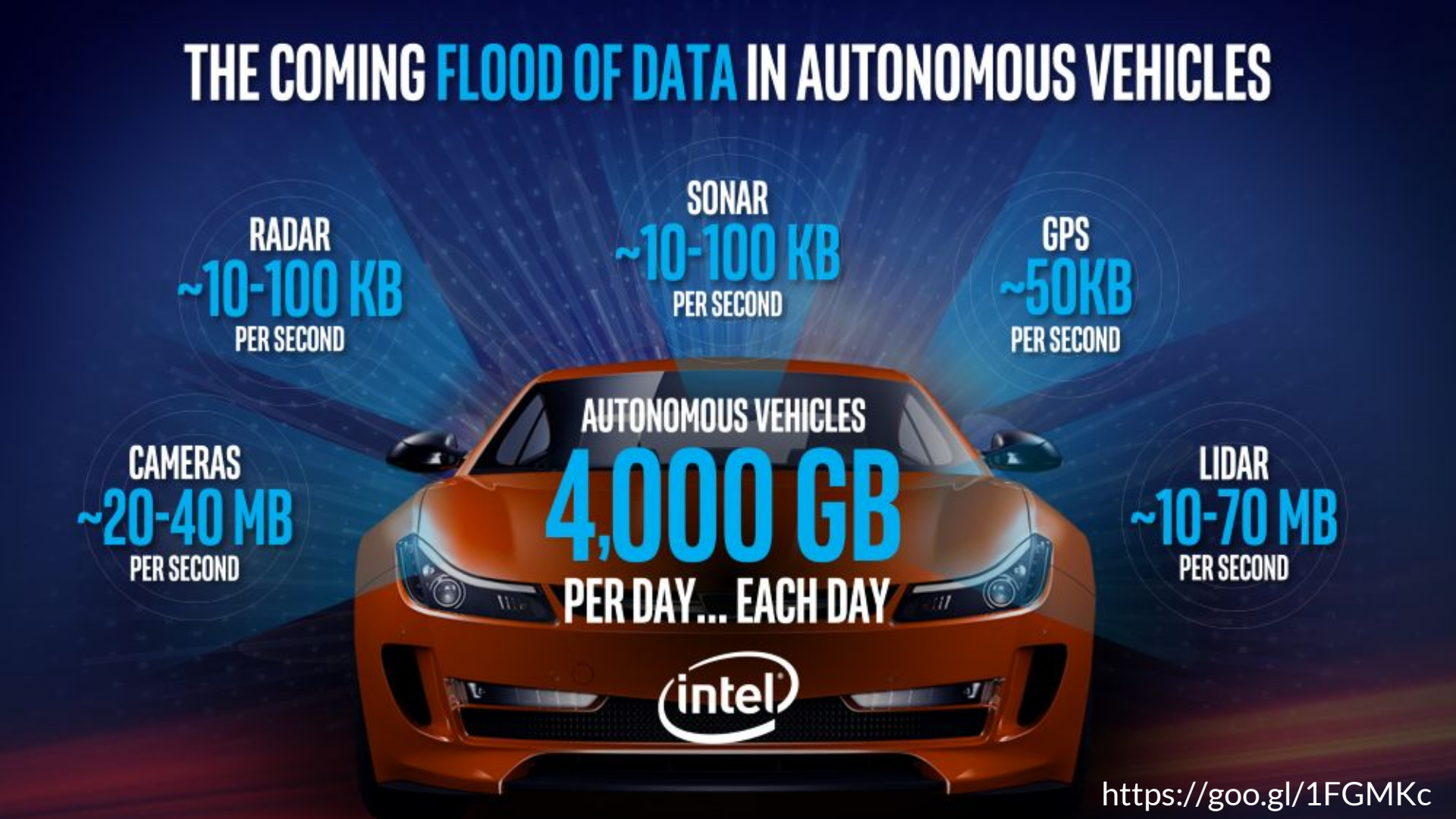
BIG SCIENCE

BIG SOCIAL MEDIA

https://goo.gl/dqqAok



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
intel
https://goo.gl/1FGMKc

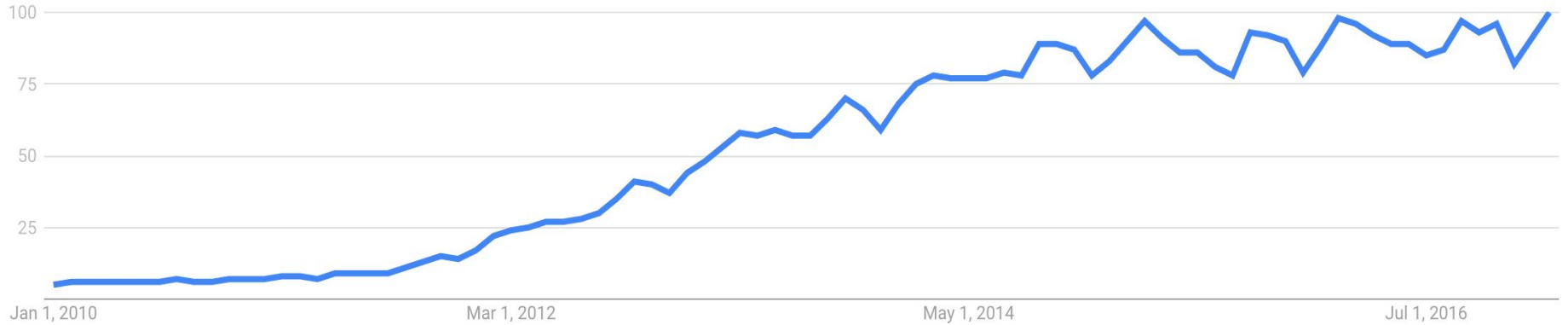
Big Data
Search term

+ Compare

https://trends.google.com/trends/explore?date=2010-01-01%202017-02-19&q=Big%20Data

Worldwide | 1/1/10 - 2/19/17 | All categories | Web Search

Interest over time



BIG DATA LANDSCAPE 2017

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Who studies this stuff?

https://goo.gl/x9fxUp

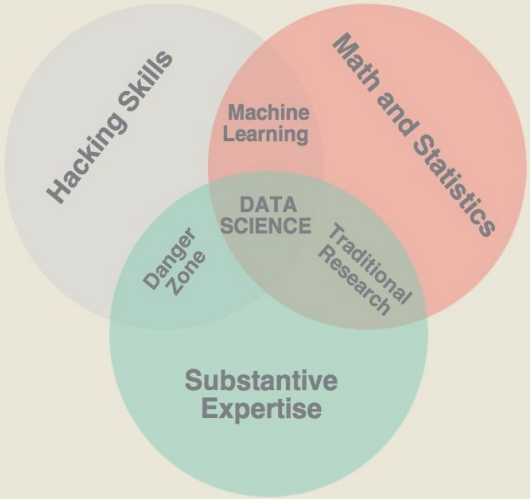
DataCamp
Learn data analysis for free, interactively

Become a

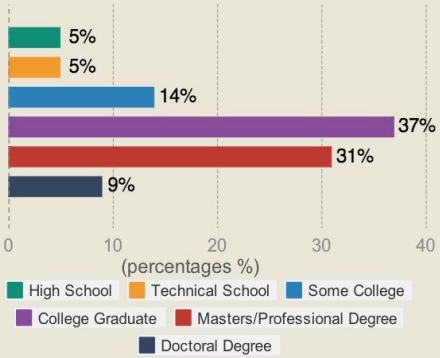
Data Scientist

in 8 easy steps

What's a data scientist?



Typical Background



A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 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 packages, 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

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization: customer tracking and on-site analytics: predictive analytics and econometrics: data warehousing and big data systems: marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

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

Become a Data Scientist

in 8 easy steps

http://blog.datacamp.com/wp-content/uploads/2014/08/How-to-become-a-data-scientist.jpg

1 Get good at stats, math and machine learning

Math

> Math Track of Khan Academy

> Linear Algebra by MIT OpenCourseware

MIT OCW

Stats

> Intro to Statistics by Udacity

> OpenIntro Statistics

UDACITY OpenIntro

ML

> Machine Learning by Andrew NG (Stanford Online)

> Practical Machine Learning by John Hopkins (Coursera)

2 Learn to code

Computer Science Fundamentals

> CS50x on edX

Grasp end-to-end development

The things you build will be integrated into other systems

sas R python

Choose a first language

> Open Source: R, Python, etc.

> Commercial: SAS, SPSS, etc.

Learn Interactively

> R: DataCamp, try R

> Python: Codecademy, Google Class

3 Understand databases

As a data scientist student, you will often work with data in text files. However, once you enter the industry, a database is almost always used to store data. It's going to be stored in MySQL, Postgres, MongoDB, Cassandra, etc.

4 Master data munging, visualization and reporting

Data cleaning and munging

WHAT

Data munging is the process of converting one "raw" form into another format for more convenient consumption

TOOLS

> Getting and Cleaning data by John Hopkins (Coursera)

DataWrangler alpha

data.table
dplyr

Data visualization

WHAT

Data visualization involves the creation and study of the visual representation of data.

TOOLS

ggvis

vega

5 Level up with Big Data

When you start operating with data at the scale of the web, the fundamental approach and process of analysis must change. Most data scientists are working on problems that can't be run on single machines. They have large data sets that require distributed processing.

Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware.

MapReduce

MapReduce is this programming paradigm that allows for massive scalability across the servers in a Hadoop cluster.

Apache Spark is Hadoop's speedy Swiss Army knife. It is a fast-running data analysis system that provides real-time data processing functions to Hadoop.

6 Get experience, practice and meet fellow data scientists

Practice makes perfect ...

Join in competitions

Meet fellow data scientists

Have a pet project

Develop your intuition

7 Internship, bootcamp or get a job

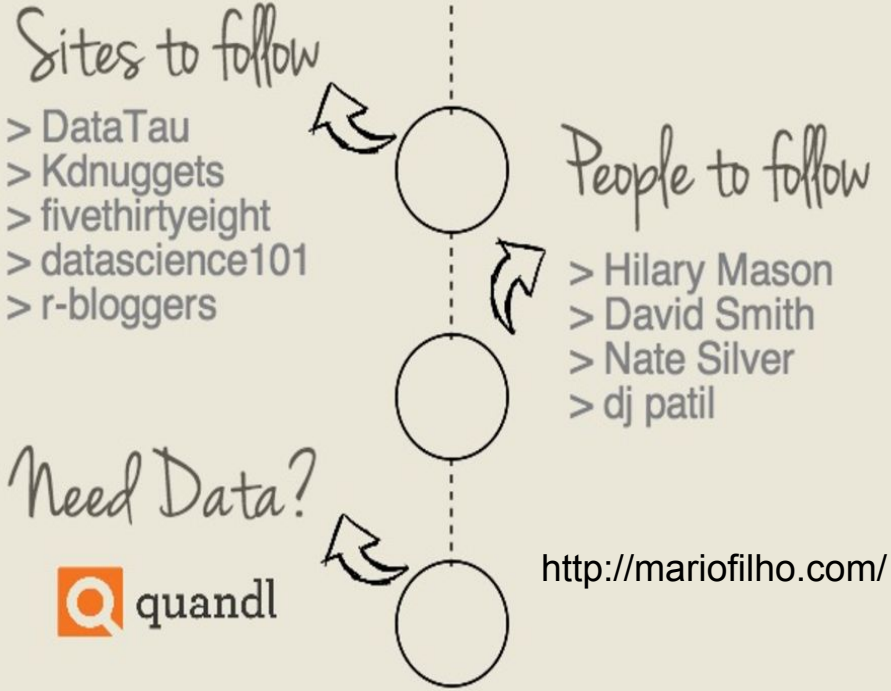
The best way to find out whether you are a true data scientist or not is to take the bull by the horns and to enter the real-life jungle of data-analysis and science with your freshly acquired skill set.

Internship	Bootcamp	Job
★☆☆	★★☆	★★★
BEGINNER	INTERMEDIATE	ADVANCED

amazon.com

zipfian

8 Follow and engage with the community



http://mariofilho.com/

Mistakes to avoid when starting you career in Data Science

While learning Data Science

1. Spending too much time on theory
2. Coding too many algorithms from scratch
3. Jumping into the deep end

https://goo.gl/CvLVpu

When applying for a job

1. Having too much technical jargon in a resume
2. Overestimating the value of academic degrees
3. Searching too narrowly
4. Being unprepared to discuss projects

Future or Present?

While studying at Academy, I built …

Models


Volume

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


Variety

Data

Data Files
(XML, CSV, Excel, JSON, …)

Database
(MySQL, Oracle, …)

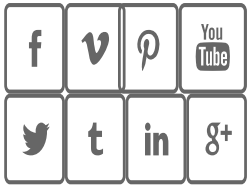
API

Sites

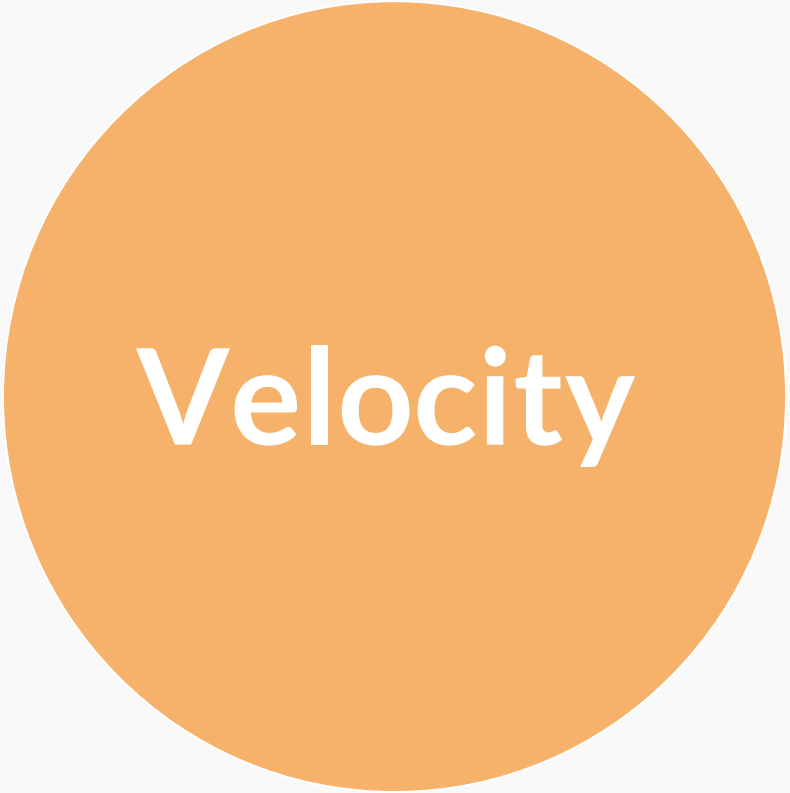
Text and reports

Maps

Image and videos



Social Media


Velocity

2017 This Is What Happens In An Internet Minute


Veracity

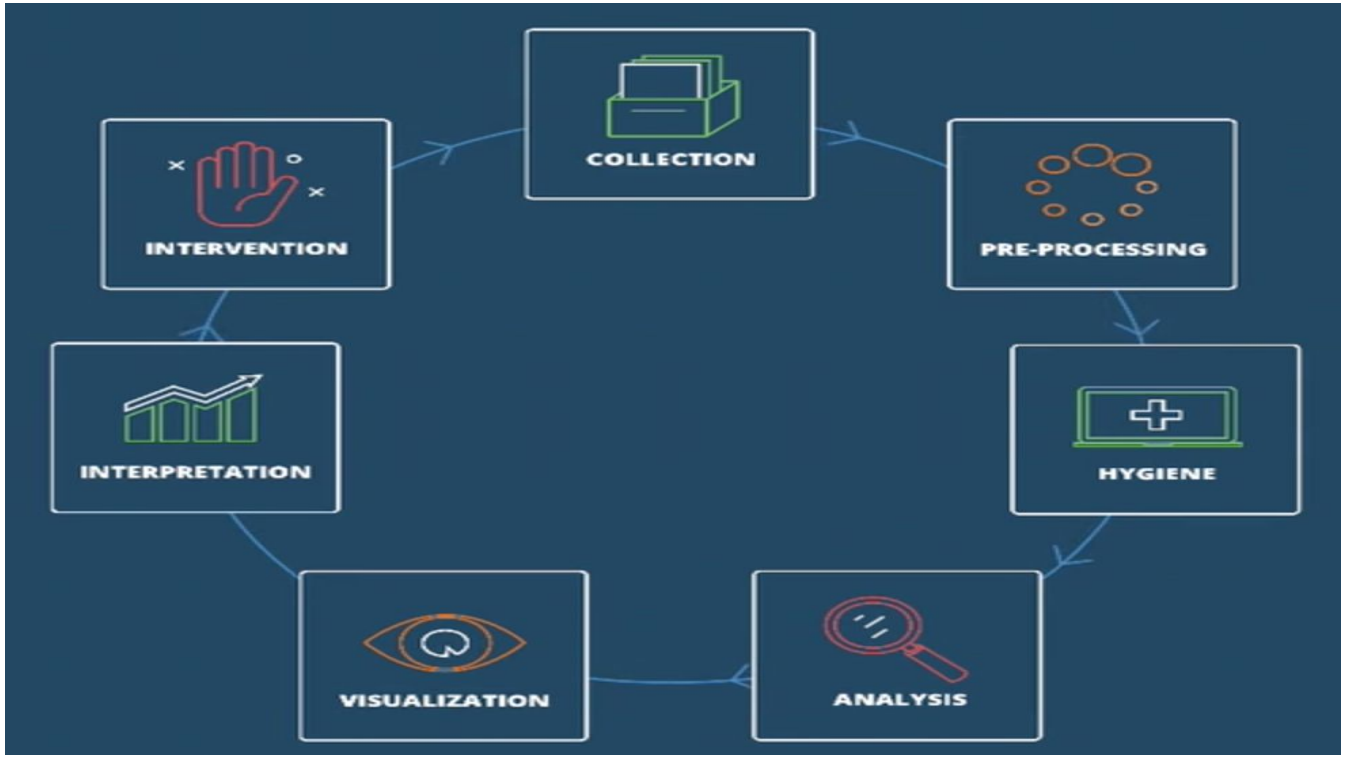
Boatos.org

Investigando Rumores Desde 2002
E-FARSAS®.com

Snopes

http://staging.kapowsoftware.com/assets/infographics/variety-of-big-data-sources-kapow-software-lg.png

Cross-Infrastructure Analytics



Search and add articles

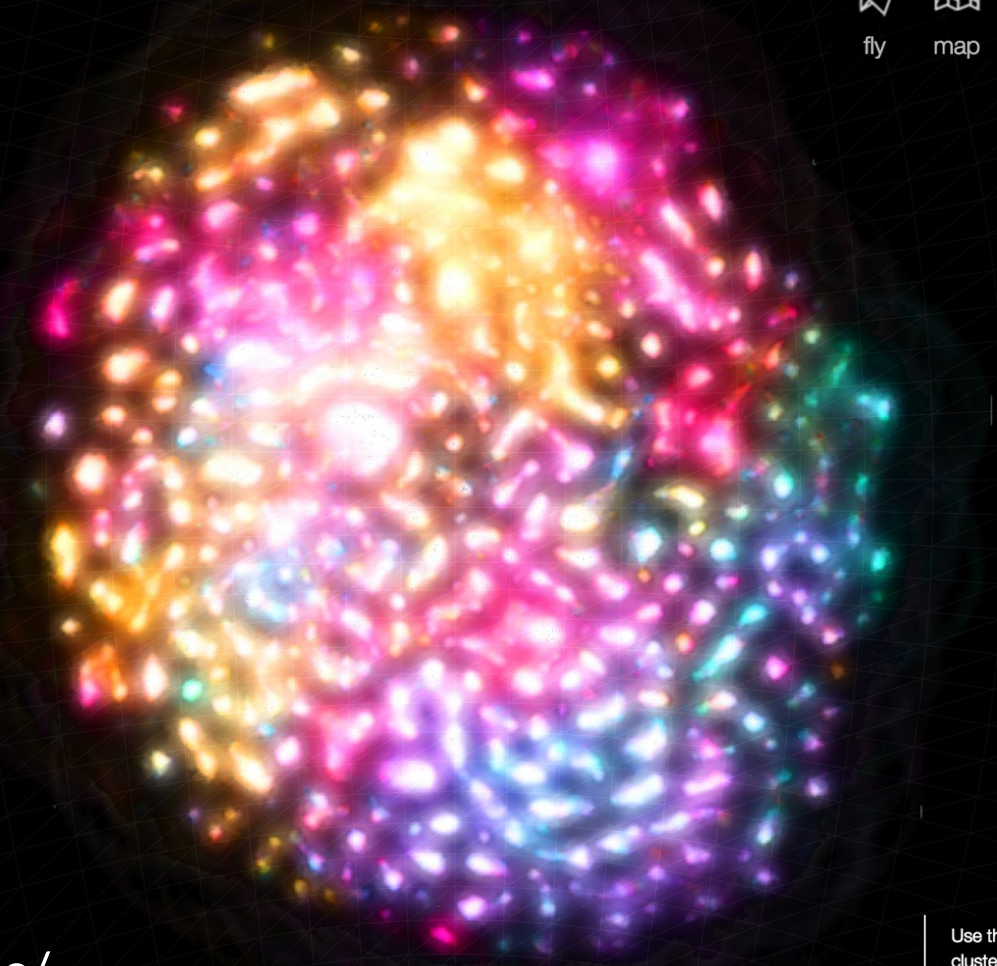
fly map home help about UI off full

Article links

Pick an article

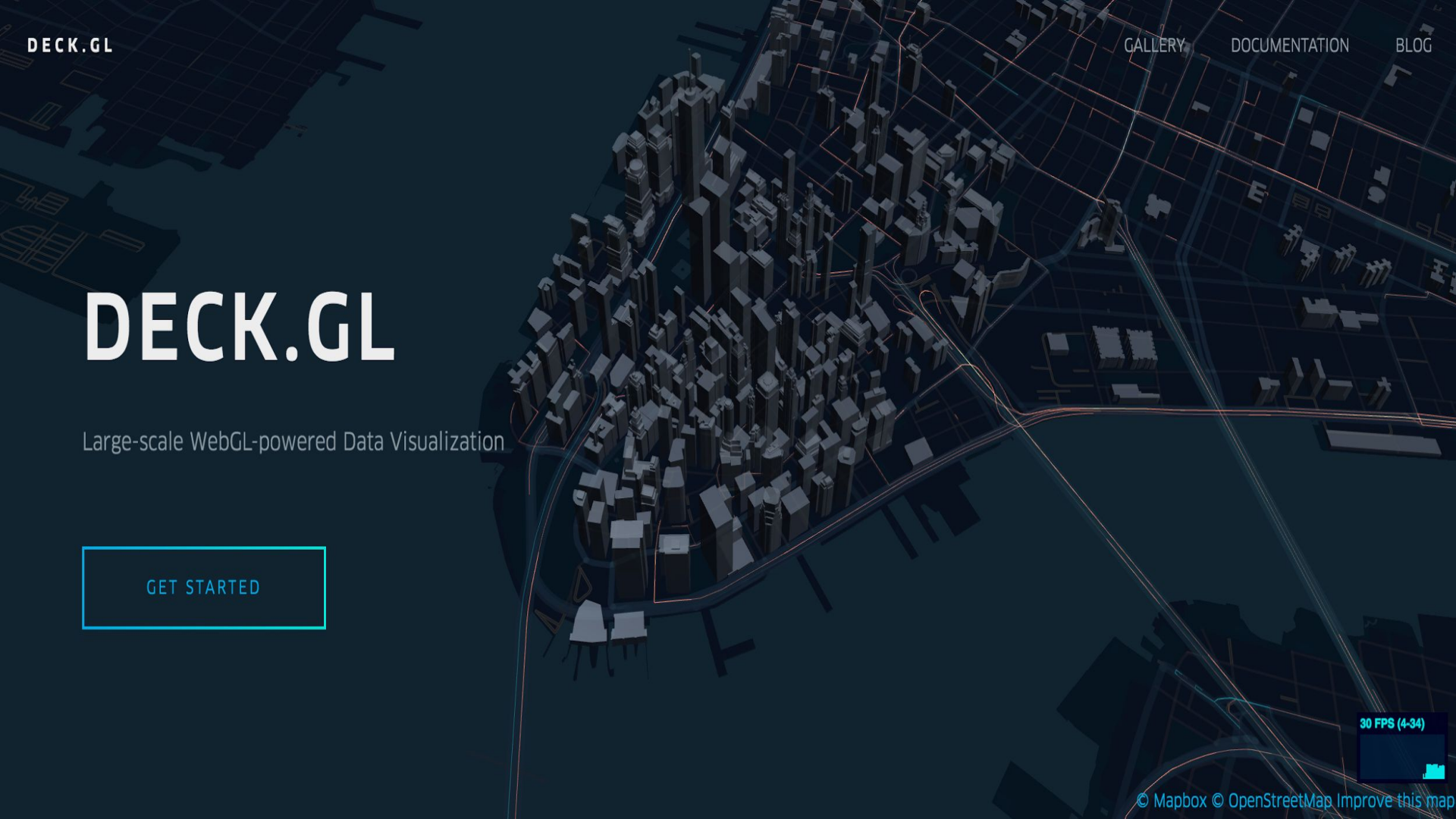
Welcome.

WikiGalaxy is a 3D web experiment that visualizes Wikipedia as a galactic web of information. With it I aim to show the world the beauty and variety of knowledge that is available at our fingertips.
I used 100,000 of 2014's most popular articles, all clustered with hyperlinks. In this world Wikipedia articles are stars, interests are nebulas and you are on a journey through knowledge.

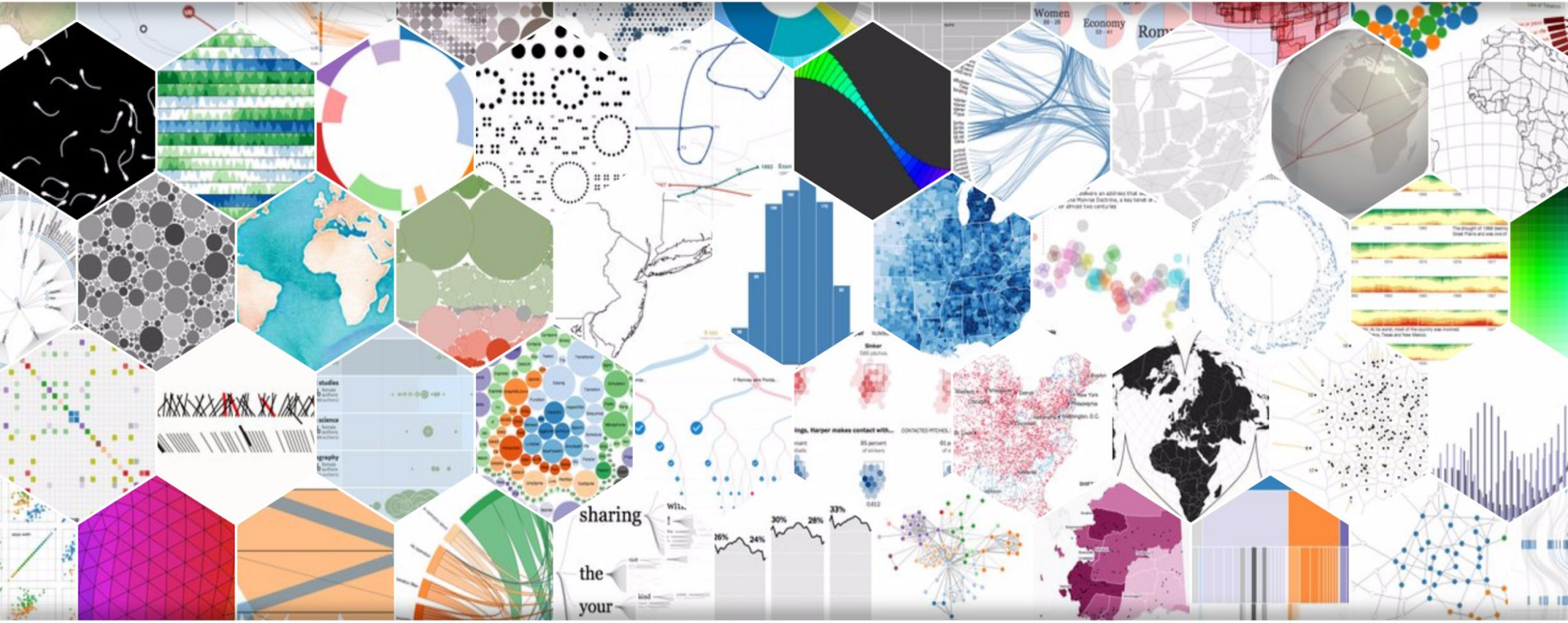


http://wiki.polyfra.me/

Use the mouse to see a preview of articles in each cluster
Click anywhere on the map to fly there


DECK.GL
GALLERY
DOCUMENTATION
BLOG
DECK.GL
Large-scale WebGL-powered Data Visualization
GET STARTED
30 FPS (4-34)
© Mapbox © OpenStreetMap Improve this map

Data-Driven Documents

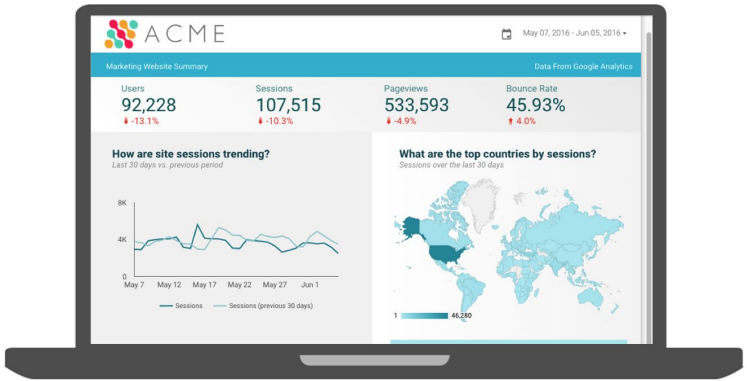


http://www.pentaho.com/

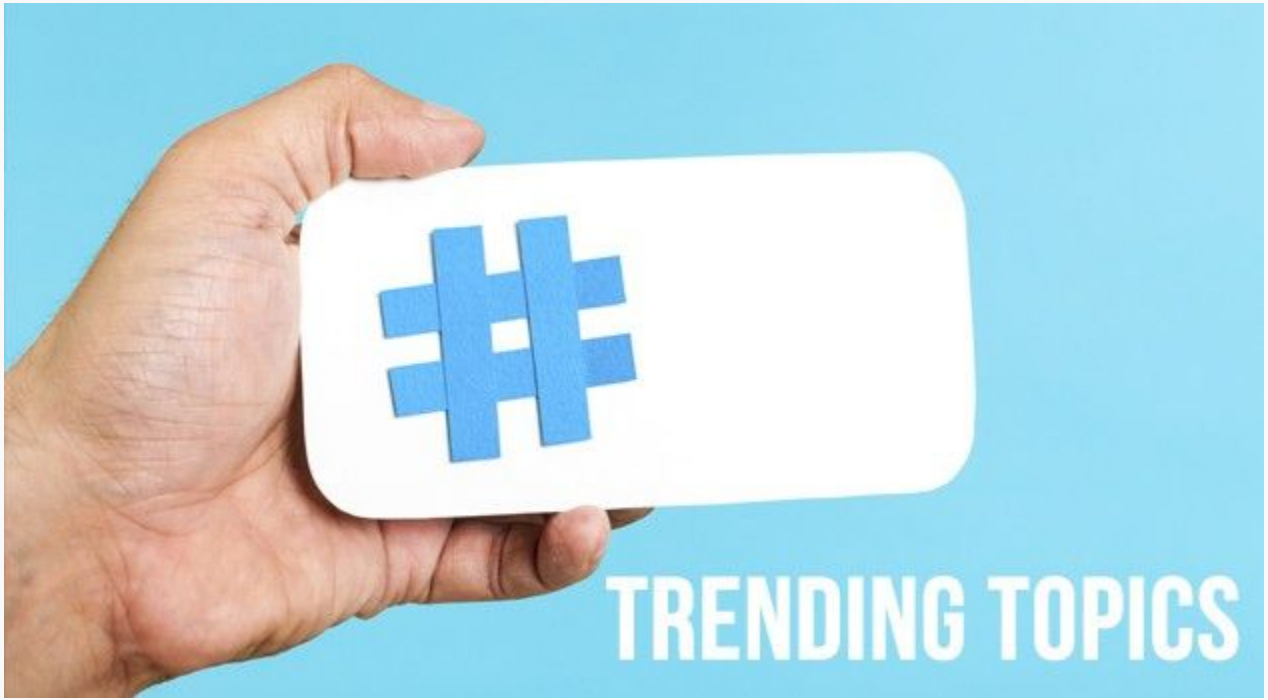
https://www.tableau.com/

https://powerbi.microsoft.com

Google Data Studio



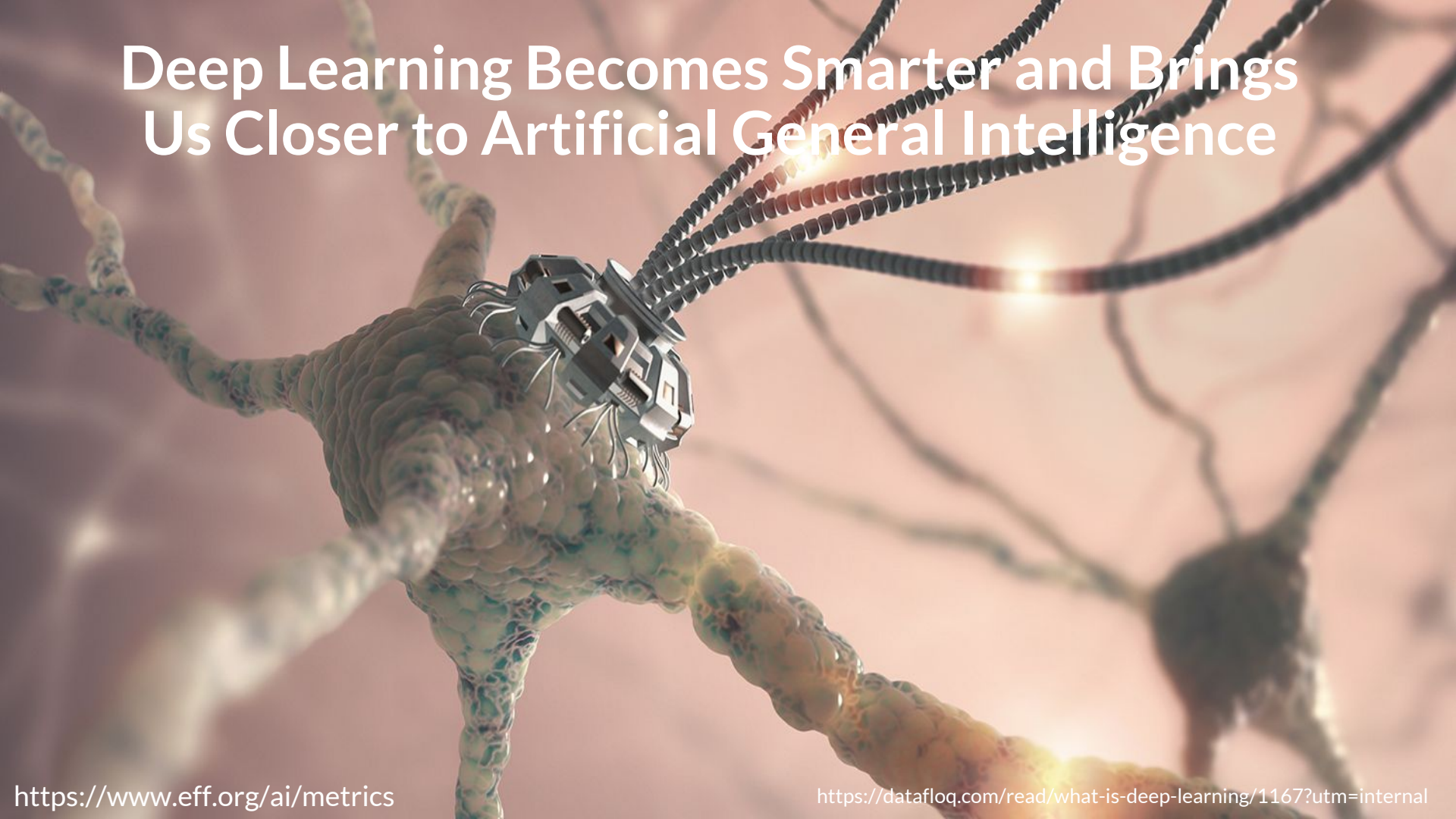
https://www.google.com.br/analytics/data-studio/


TRENDING TOPICS

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

Uber's plans to data analytics

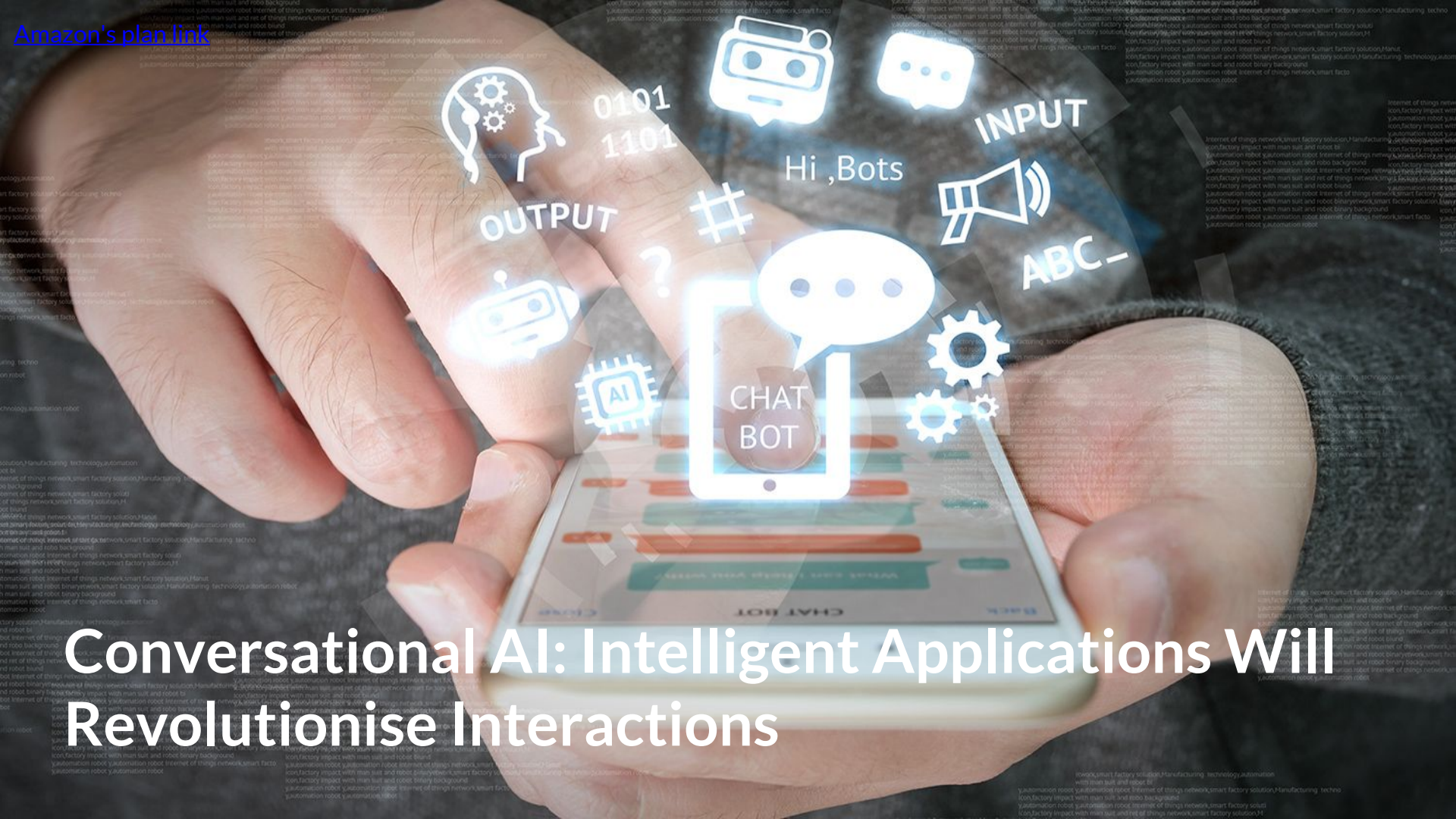
Deep Learning Becomes Smarter and Brings Us Closer to Artificial General Intelligence



https://www.eff.org/ai/metrics

https://datafloq.com/read/what-is-deep-learning/1167?utm=internal

Amazon's plan link



Conversational AI: Intelligent Applications Will Revolutionise Interactions

IoT-Related Data Breaches Will Cause Havoc

https://datafloq.com/read/ddos-attack-a-wake-up-call-for-iot/2480?utm_source=datafloq&utm_medium=ref&utm_campaign=datafloq

StreetStyle: Exploring world-wide clothing styles from millions of photos

https://goo.gl/i9rDJX

Artificial intelligence can now predict suicide with remarkable accuracy

https://goo.gl/1b1Mr9

HOSPITAL SÍRIO-LIBANÊS

KUNUMI

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

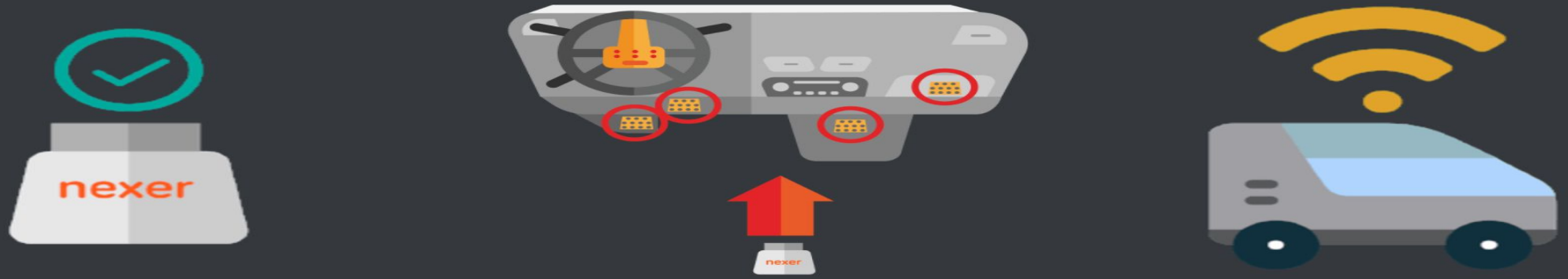
https://www.ufmg.br/online/radio/arquivos/046358.shtml

UMA NOVA EXPERIÊNCIA COM O SEU CARRO

CONECTE O SEU VEÍCULO AO RESTO DA SUA VIDA DIGITAL

FAÇA A SUA RESERVA

https://www.nexer.com.br/



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

Spotify

A Maior Rede de Anúncios Mobile do Brasil

Dia 19 de junho nos encontramos em Cannes

Assista ao vídeo

Abril AFRICA Claro Coca-Cola DM9DDB FIAT isobar LG


HEKIMA
HOME
SOBRE
SOLUÇÕES
TECNOLOGIA
BLOG
CONTATO
BIG DATA & INTELIGÊNCIA ARTIFICIAL
TRANSFORME DADOS EM INFORMAÇÃO
QUEM SOMOS
NOSSAS SOLUÇÕES


INOVALL
SOLUÇÕES
VANTAGENS
PRODUTOS
CARREIRAS
BLOG
DOMINE O MERCADO COM O
PROSPECTA
Tecnologia e conteúdo para o crescimento da sua empresa.
Inteligência para você obter o potencial máximo do mercado.
DESCUBRA

https://goo.gl/Ndf38Q

Big Data

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

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

150 likes
- ++ family member

300 likes
- ++ spouse

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

https://goo.gl/OIWRcX

Keynote "The End of Privacy", Dr. Michal Kosinski

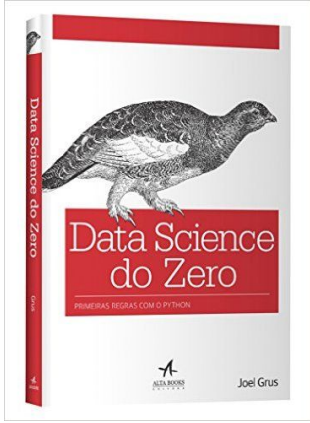
http://www.michalkosinski.com/

https://www.youtube.com/watch?v=NesTWiKfpD0&feature=youtu.be

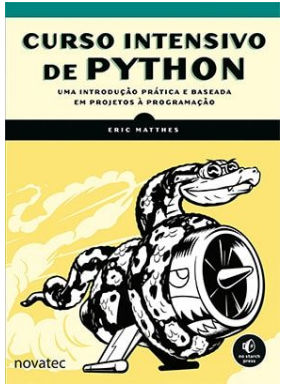
Agenda

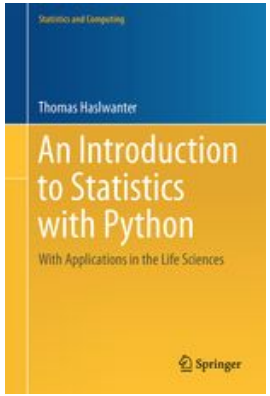
- Big Data foundations
- Introduction to Python for Data Science
- Intermediate Python for Data Science
- Bokeh
- Geocoding
- Network Analysis

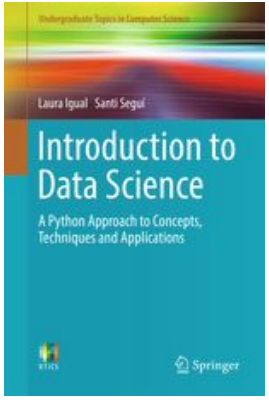
References

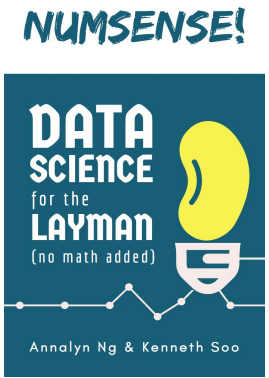

Data Science do Zero
Data Science do Zero
PRIMEIRAS REGRAS COM O PYTHON
Joel Grus

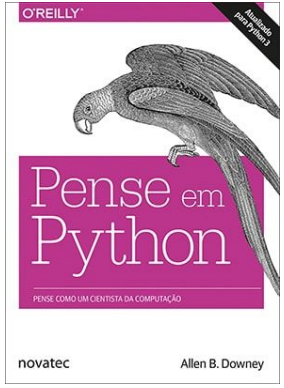

Fernando Amaral
INTRODUÇÃO À CIÊNCIA DE DADOS
Fernando Amaral
INTRODUÇÃO À CIÊNCIA DE DADOS
MINERAÇÃO DE DADOS E BIG DATA

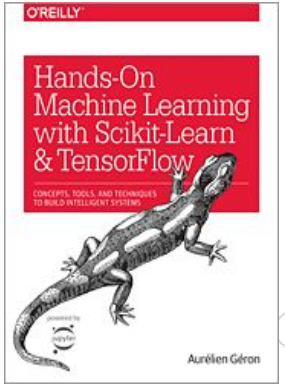

CURSO INTENSIVO DE PYTHON
UMA INTRODUÇÃO PRÁTICA E BASEADA EM PROJETOS À PROGRAMAÇÃO
ERIC MATTHES
novatec


Statistics and Computing
Thomas Haslwanter
An Introduction to Statistics with Python
With Applications in the Life Sciences
Springer


Undergraduate Topics in Computer Science
Laura Igual Santi Seguí
Introduction to Data Science
A Python Approach to Concepts, Techniques and Applications
Springer


NUMSENSE!
DATA SCIENCE for the LAYMAN
(no math added)
Annalyn Ng & Kenneth Soo


O'REILLY
Atualizado para Python 3
Pense em Python
PENSE COMO UM CIENTISTA DA COMPUTAÇÃO
novatec
Allen B. Downey


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

References

UDACITY

Machine Learning Engineer

kaggle

Deep Learning Foundations

Data Analyst

facebook mongoDB

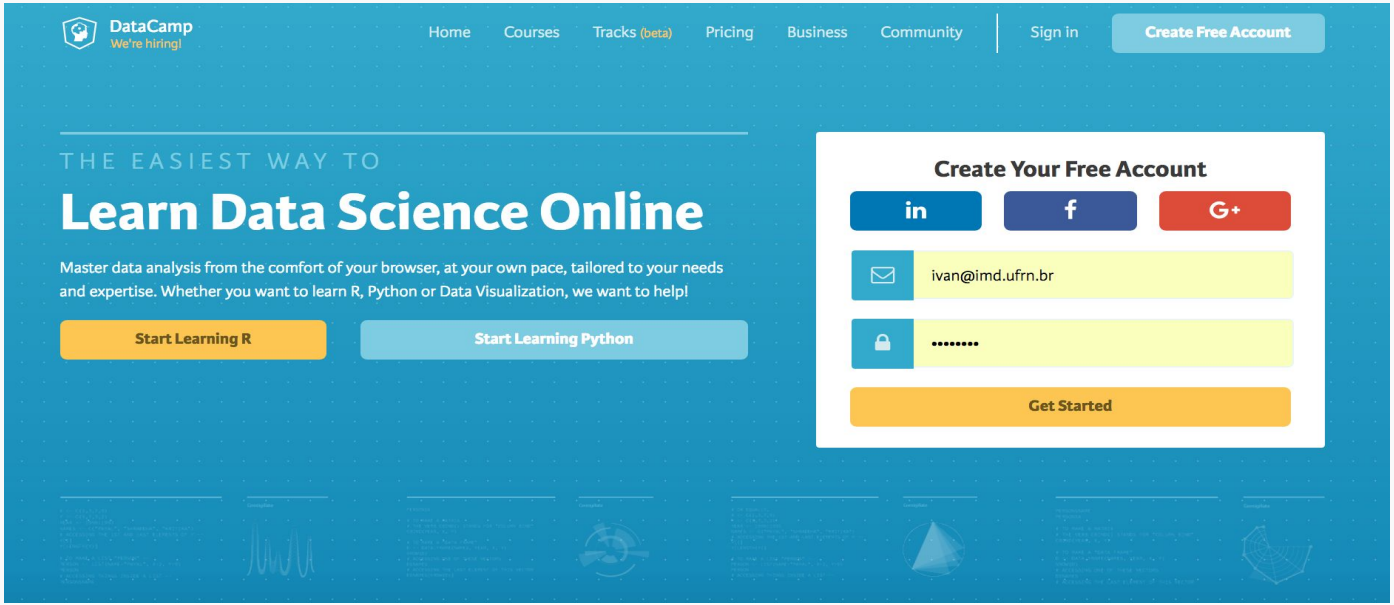
Artificial Intelligence

IBM Watson amazon alexa DiDi

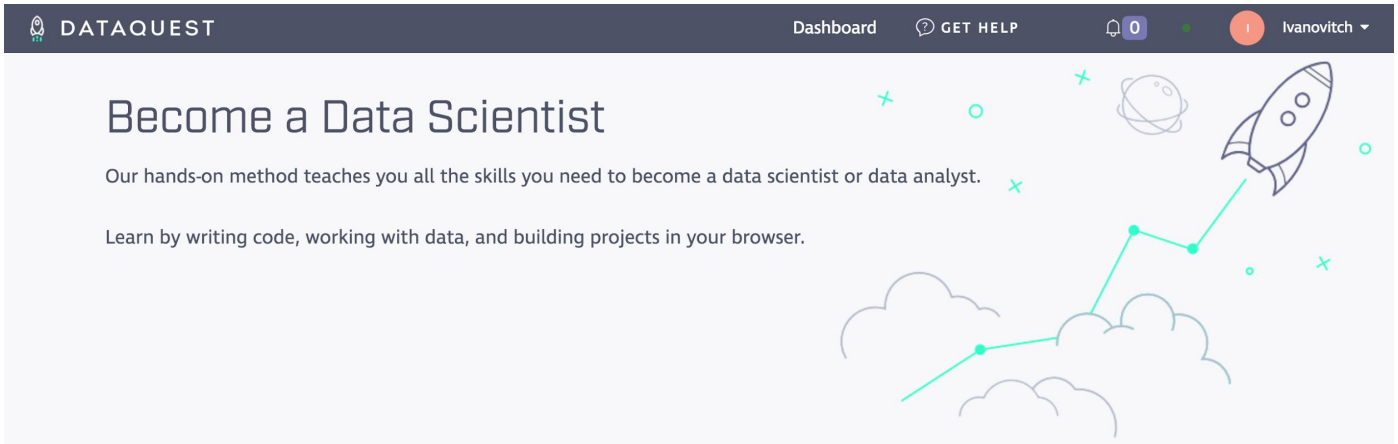
Self-Driving Car Engineer

Mercedes-Benz NVIDIA UBER ATG

References


DataCamp
We're hiring!
Home
Courses
Tracks (beta)
Pricing
Business
Community
Sign in
Create Free Account
THE EASIEST WAY TO
Learn Data Science Online
Master data analysis from the comfort of your browser, at your own pace, tailored to your needs and expertise. Whether you want to learn R, Python or Data Visualization, we want to help!
Start Learning R
Start Learning Python
Create Your Free Account
in
f
G+
ivan@imd.ufrn.br
••••••••
Get Started

References

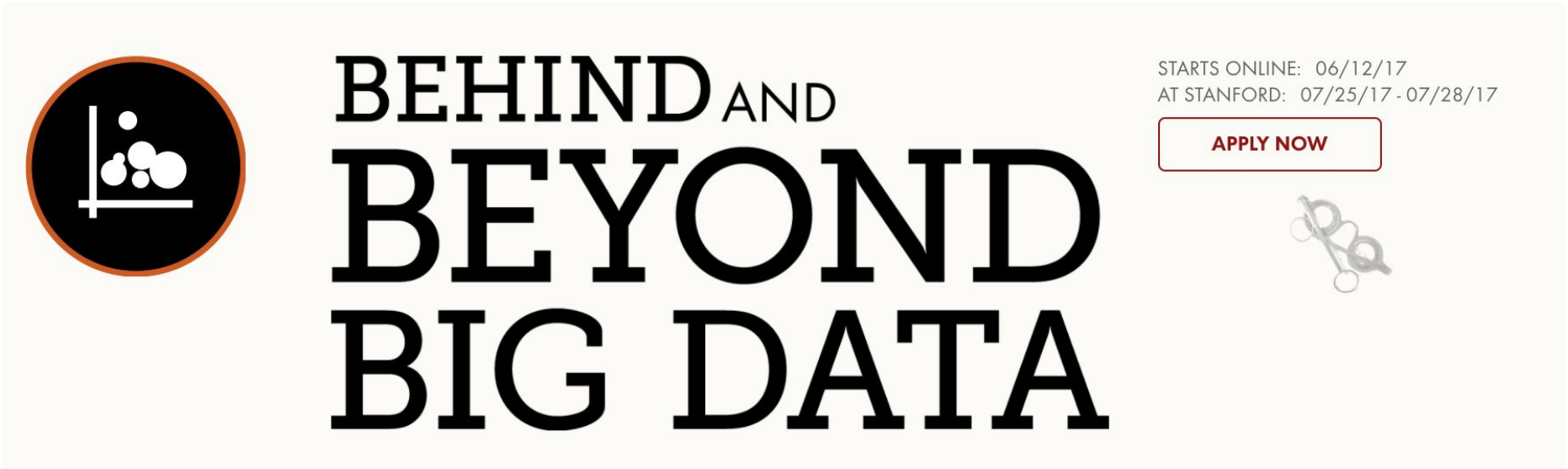

DATAQUEST
Dashboard
GET HELP
0
Ivanovitch
Become a Data Scientist
Our hands-on method teaches you all the skills you need to become a data scientist or data analyst.
Learn by writing code, working with data, and building projects in your browser.

References

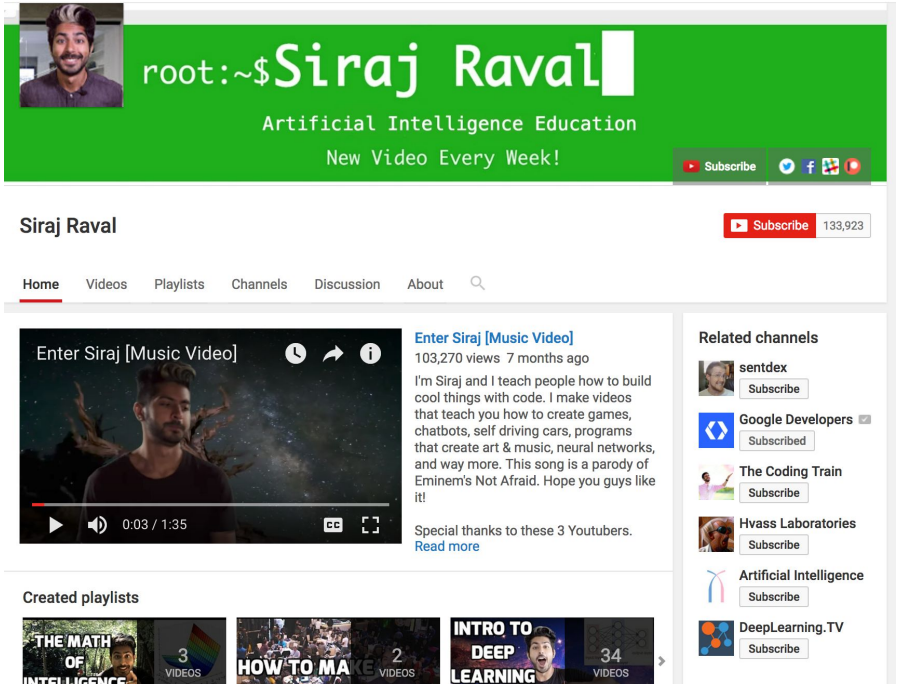

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Big Data and Social Analytics certificate course
EXperimental Learning
2017 DATES TO BE CONFIRMED
DOWNLOAD COURSE PROSPECTUS
Discover a new way to think about big data analysis when you explore the theory behind "social analytics", and practically apply that knowledge as you learn pioneering data analytics techniques from the creators of those very tools and methods.

References

Stanford | ONLINE



References



http://www.andrewng.org/

http://mariofilho.com/

References

https://elitedatascience.com/
https://algobeans.com/
https://www.datacamp.com/home
https://www.dataquest.io/dashboard
https://www.datascienceacademy.com.br/
http://www.bigdatabusiness.com.br/
https://datafloq.com/
http://www.informationisbeautiful.net/
http://www.datapedia.info/public/
http://ckan.imd.ufrn.br/
http://dados.ufrn.br/
https://fivethirtyeight.com/
https://news.ycombinator.com/
http://machinelearningmastery.com/blog/

References

- How do I learn python in deep?
  - https://www.hackerrank.com/
  - https://www.codewars.com/
  - https://br.codecombat.com/
  - https://www.hackerearth.com
  - https://www.drivendata.org/
  - https://www.kaggle.com/
  - https://goo.gl/WhLvs9

Lesson #1