Data Science Foundations

Lesson #1 - Outline & Directions

Ivanovitch Silva

July, 2017


Introduction

Ivanovitch Silva (ivan@imd.ufrn.br)
Office: CIVT-A226 2T34, 6T34

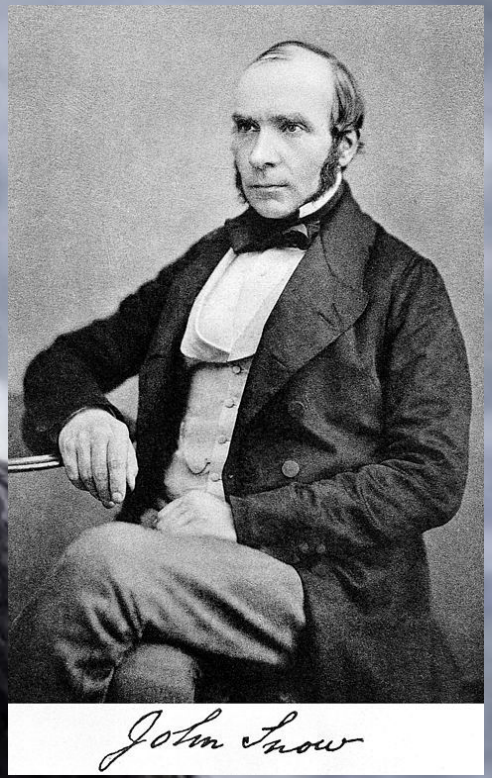
Group Knowledge

Take a Survey

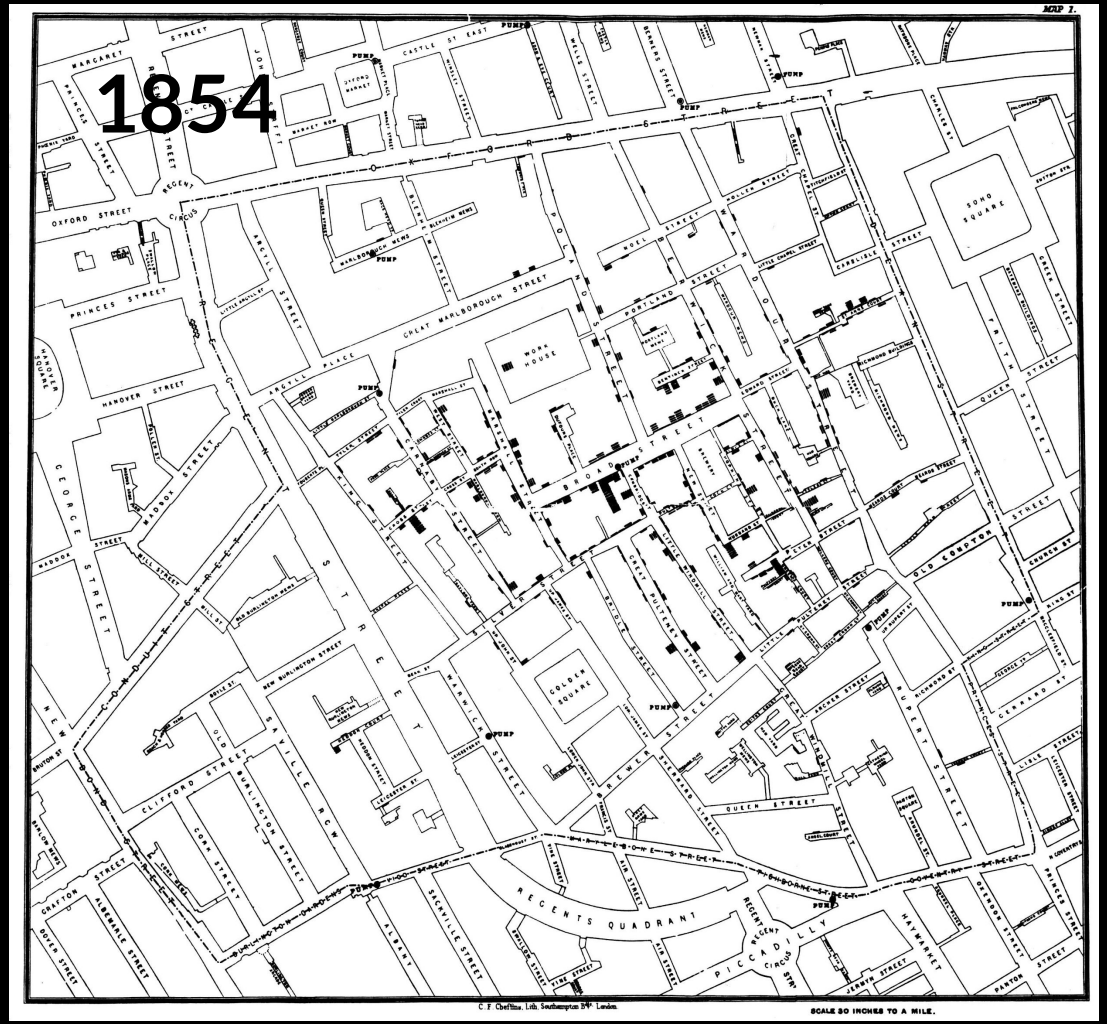
https://goo.gl/XmRwNj


About Data

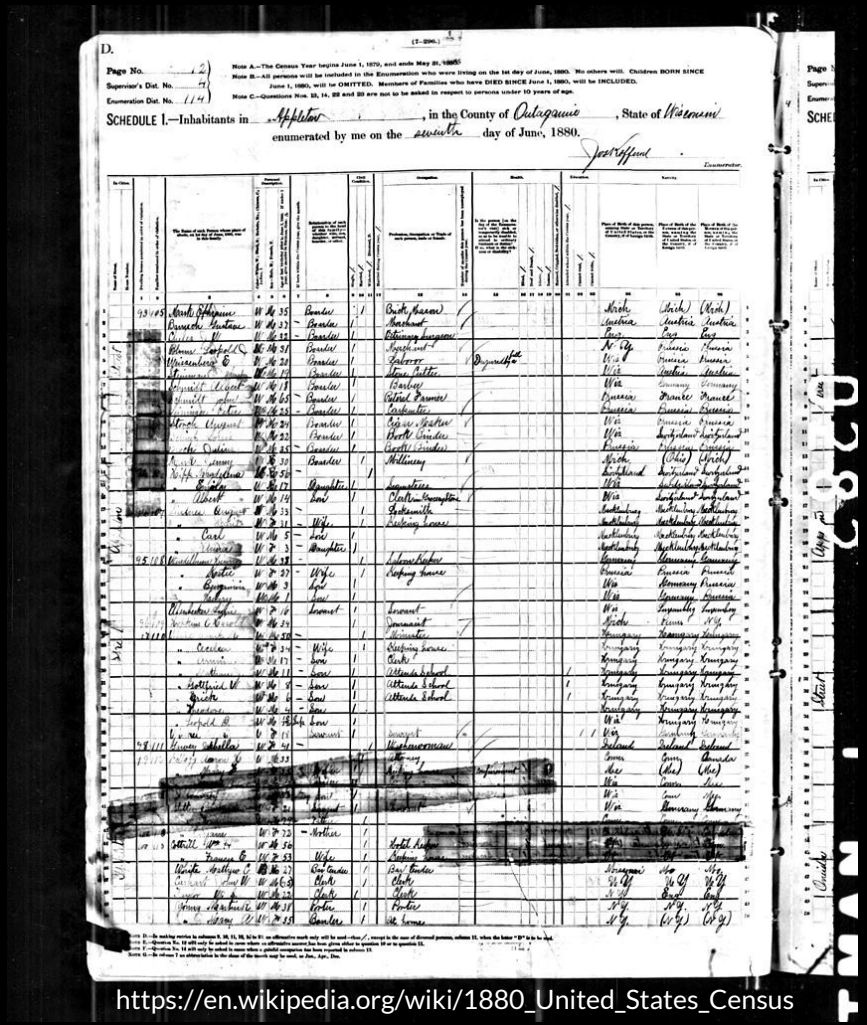
John Snow, London, 1854



https://www.winshuttle.com/big-data-timeline/



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

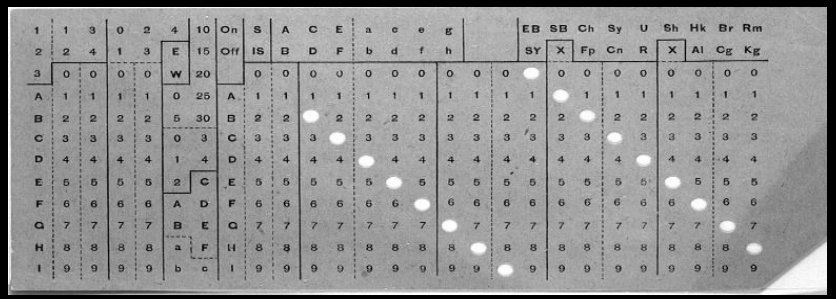


https://en.wikipedia.org/wiki/1880_United_States_Census

U.S. census (1880)

- 50 million people
- Age, sex, occupation, education level
- It took 8 years to be tabulate

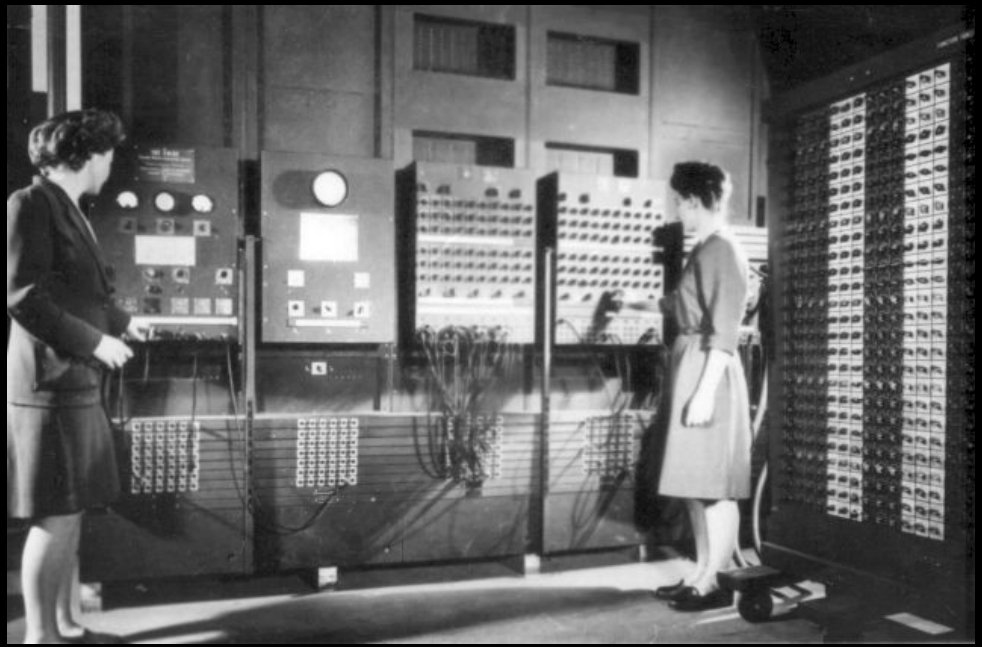
https://en.wikipedia.org/wiki/Tabulating_machine



https://en.wikipedia.org/wiki/1890_United_States_Census

U.S. census (1890)

- Tabulating machine.
- 63 million people
- The results were announced after only six weeks of processing.
- Without this invention, experts had estimated, the 1890 census would have taken 13 years to fully tabulate.



The population boom (1932)

- 123 million people.
- Information overload continued with the boom in the US population
- It demanded more thorough and organized record-keeping.

Scientific Knowledge Expands (1962)

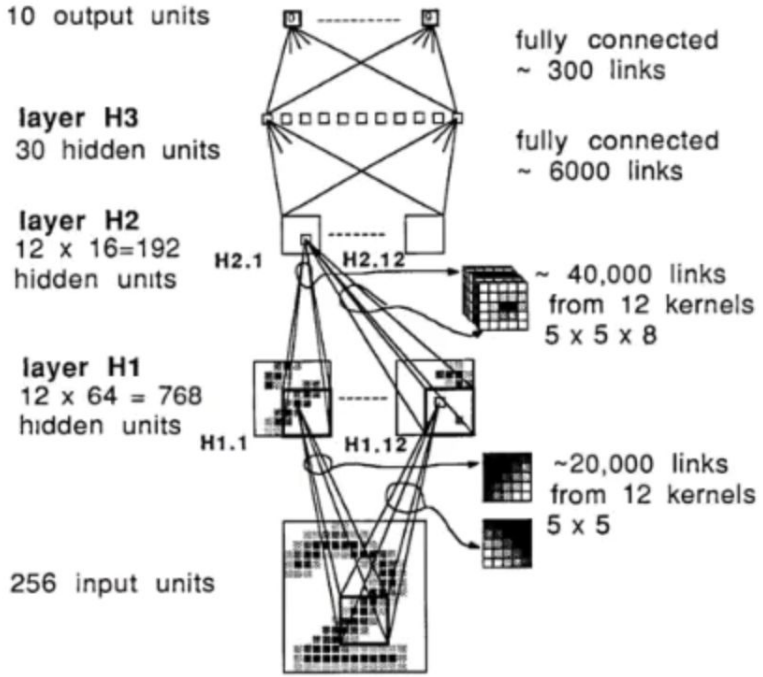
- Derek Price publishes "Science Since Babylon"
- He concludes that the number of new journals has grown exponentially rather than linearly.
- This is now better known as the "law of exponential increase".

The Rise in Two-Way Communication (1975)

- The Information Flow Census, conducted by the Ministry of Posts and Telecommunications in Japan, started tracking the volume of information circulating in that country
- The study found that information supply greatly exceeded information consumption, and the demand for one-way communication had stagnated.

Information Growth and the Broadcasting Industry (1983)

- Companies were beginning to use data to provide answers for better business decisions.
- The massive information growth was credited to the expansion of the broadcasting industry.



Neural Networks debuted (1990)

Matan, Ofer, et al. Handwritten character recognition using neural network architectures. Proceeding of the 4th USPS Advanced Technology Conference.

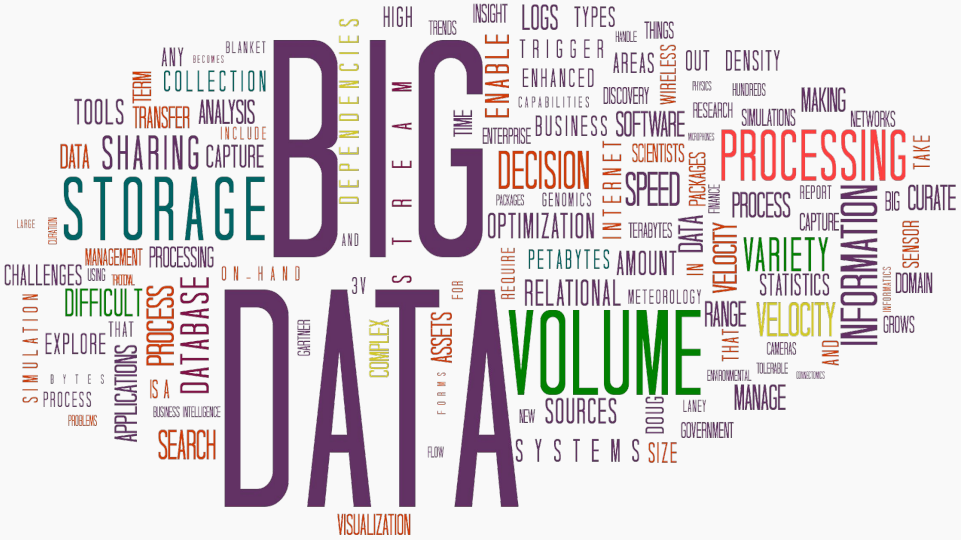
The bird of the Internet (1991)

- Courtesy of Sr Tim Berners-Lee, data and information can now be posted online for the first time.

The future of data storage (1996)

- Digital storage became more cost-effective for store data than paper.
- Michael Lesk published How much information is there in the world? So in only a few years, we will be able [to] save everything

Google Founded in 1996



The first time (1997)

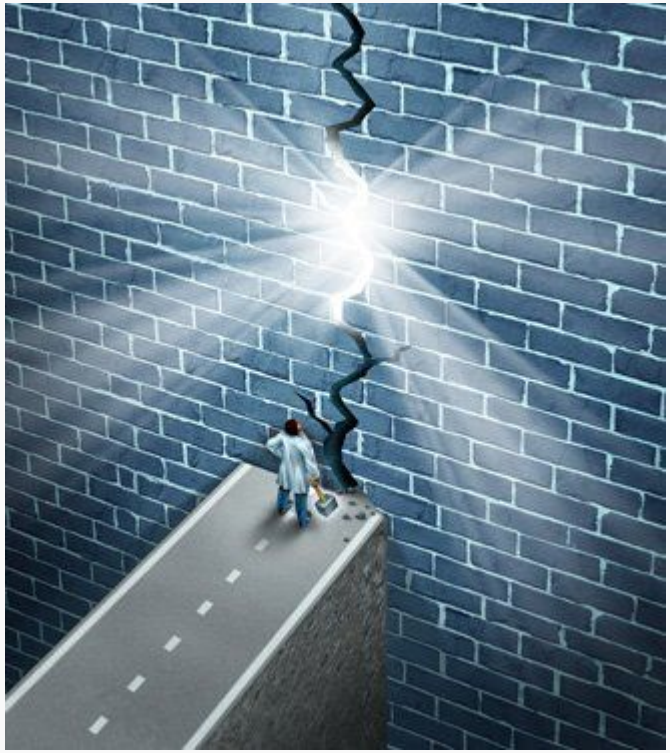
- The term "Big Data" was used for the first when researchers M. Cox and D. Ellworths wrote an article identifying that the rise of data was becoming an issue for current computer systems.

Internet of Things (1999)

- The term “Internet of Things” was coined by British entrepreneur Kevin Ashton, Co-Founder of the Auto-ID Center at MIT, during a presentation linking the idea of Radio Frequency Identification (RFID) in supply chain to the internet world.

An Open Source Solution to the Big Data Explosion (2006)

- Hadoop was created in 2006 out of the necessity for new systems to handle the explosion of data from the web.
- Free to download, use, enhance and improve, Hadoop is a 100% open source way of storing and processing data that enables distributed parallel processing of huge amounts of data across inexpensive, industry-standard servers that both store and process the data, and can scale without limits”.

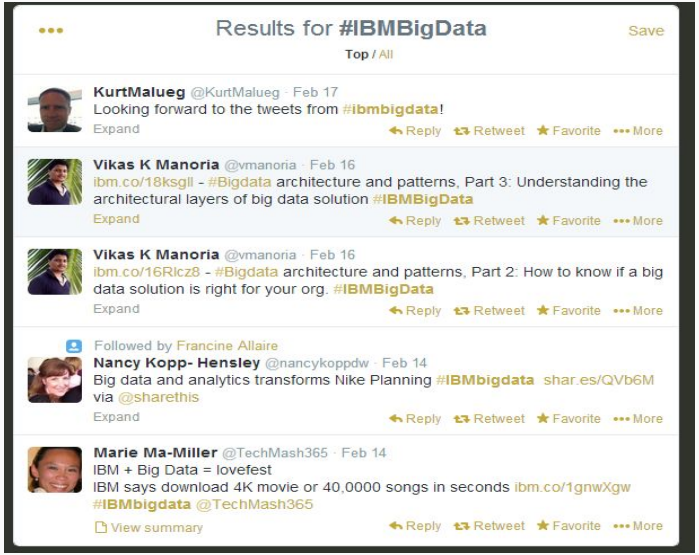


Revolutionary Breakthroughs (2008)

A group of computer science researchers published a paper titled Big Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science, and Society.

Big-data computing is perhaps the biggest innovation in computing in the last decade

A modest investment by the federal government could greatly accelerate its development and deployment.



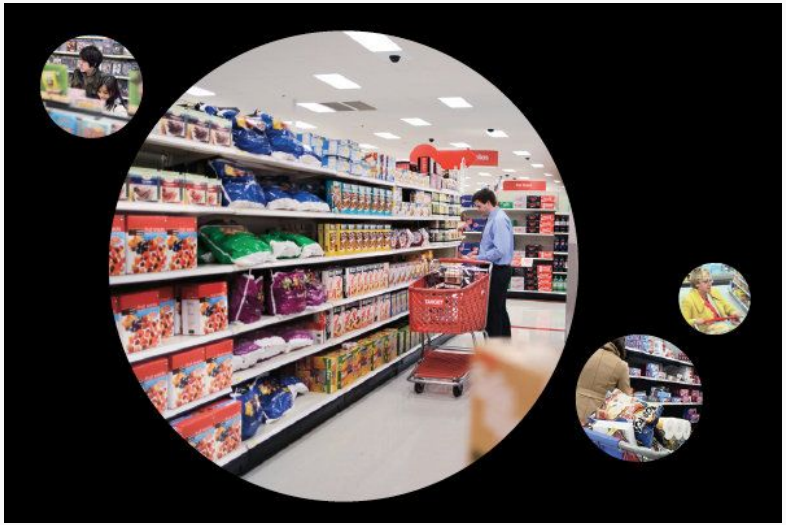
#IBMBigData (2011)

IBM introduced a Twitter hashtag, #IBMbigdata which expanded on their Big Data themed website that they built in 2008 in an effort to integrate it into their marketing.

http://www.zdnet.com/article/google-brain-simulator-teaches-itself-to-recognize-cats/

Deep Learning experiment at Google (2012)

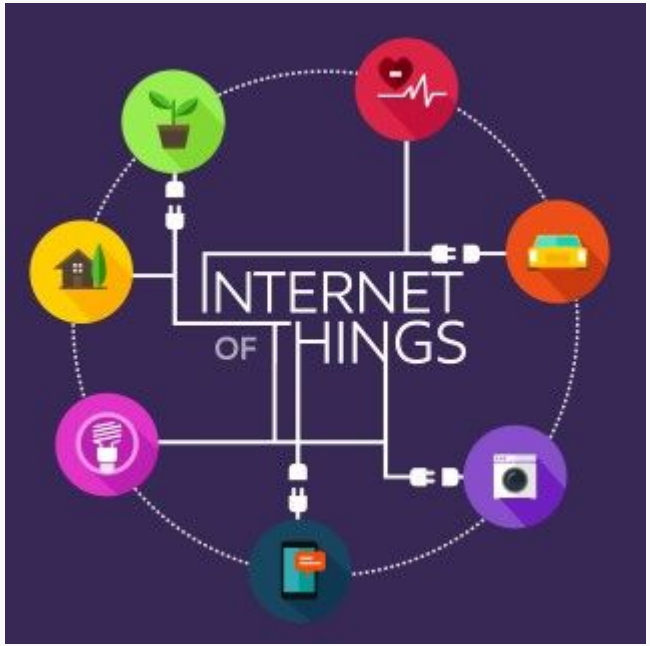
After seeing 10 million images from YouTube videos within three days, the 16,000-computer network, which had one billion connections, began to recognize cats, even though it had never been taught what a cat looked like.



http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=1&hp

How Companies Learn Your Secrets (2012)

“My daughter got this in the mail!” he said. “She’s still in high school, and you’re sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?”



#The Year of the Internet of Things (IoT) 2014

The IoT has become a powerful force for business transformation, and its disruptive impact will be felt across all industries and all areas of society.

According to Gartner, there were 3.7 billion connected "things" in use in 2014

Calvin Klein

Knits in myriad varieties, including curly mohair coats and hand-stitched, multipaneled sweaters, were the focus of a disciplined but cozy collection in a soothing palette of earth tones and snow.

Mixed-knit jumpers blocked out in white, black and gray

Turtleneck sweater tops with chunky, sampler-scarf knit panels

Ralph Lauren

A collection distinguished by its sure-handed hybrid of refinement and ease, its muted pastels and buttery fabrics arguing for opulence but in no way overstating the case.

Read more: **Ralph Lauren Plays Polo**

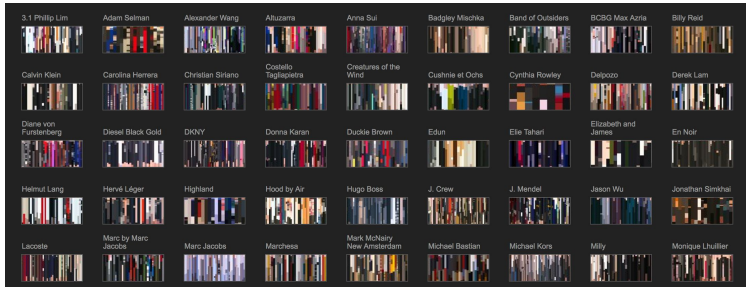
An upscale hippie dress under a distressed bomber

A massive, asymmetrical cape in soft pink

BIG DATA

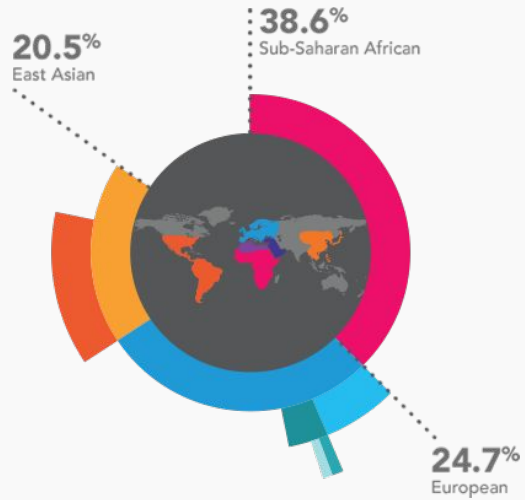
2014

FASHION WEEK MEN



23andMe

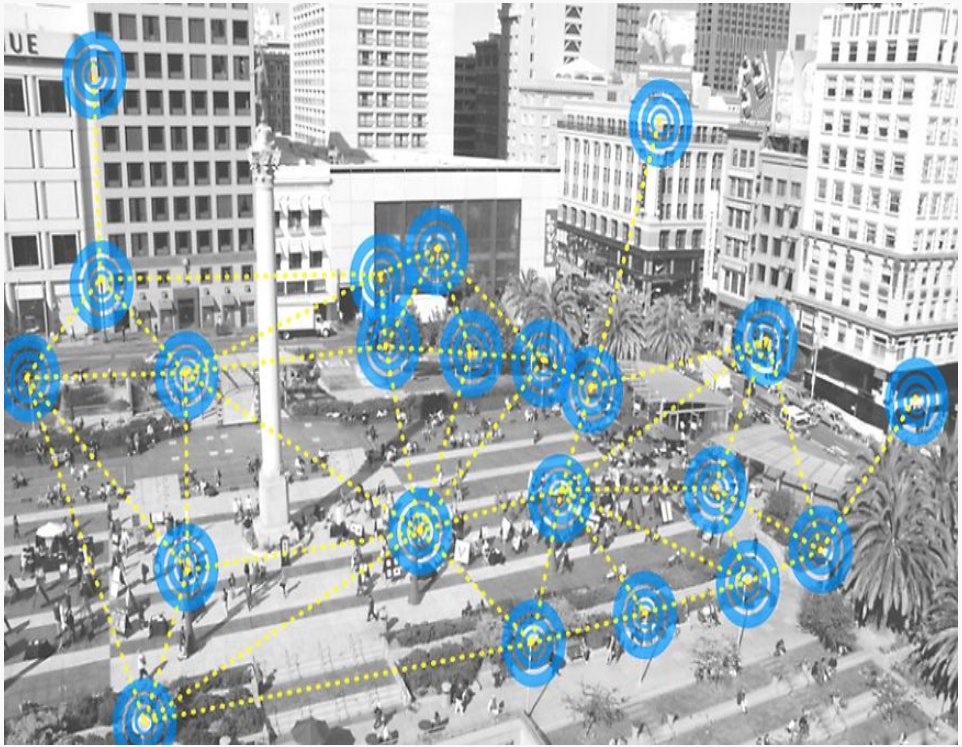
https://www.23andme.com/



23andMe has been selling a product with both ancestry and health-related components in Canada since October 2014

http://www.telegraph.co.uk/technology/news/10959864/Germanys-World-Cup-tactics-shaped-by-data.html

Germany’s 12th Man at the World Cup (2014)



http://www.forbes.com/sites/danielnewman/2016/08/15/big-data-and-the-future-of-smart-cities/#16c07f1d3f2d

Connect To The Cities Of The Future (2015)

With the growth of our population and the advent of ideas such as big data and the Internet of Things, the natural step cities will take is to become more interconnected.

Not only will this result in brighter streets, but the new lights will also be an interconnected system that will inform the city of each bulb's status.

2016

http://www.pokemongo.com/

https://www.youtube.com/watch?v=NrmMk1Myrxc

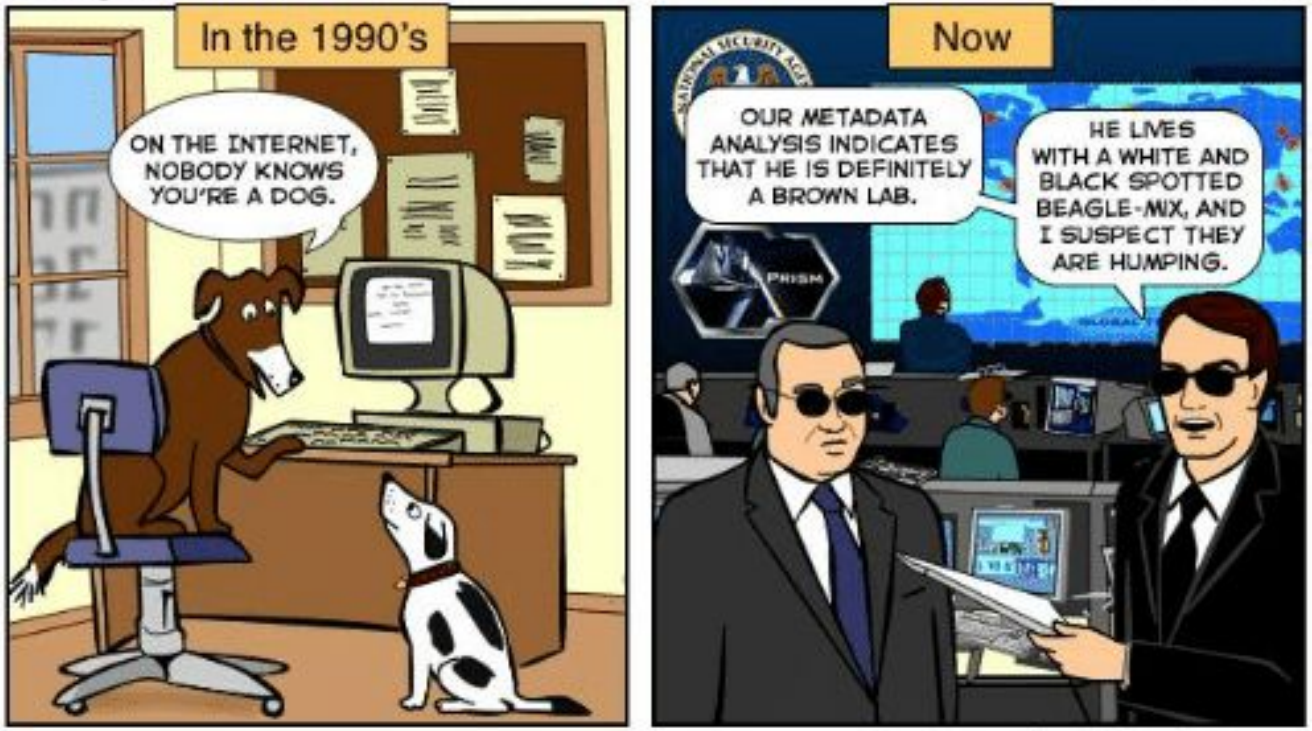
Why Netflix thinks its personalized recommendation engine is worth $1 billion per year (2016)

“Consumer research suggests that a typical Netflix member loses interest after perhaps 60 to 90 seconds of choosing"

The user either finds something of interest or the risk of the user abandoning our service increases substantially

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

The Joy of Tech™ by Nitrozac & Snaggy



joyoftech.com

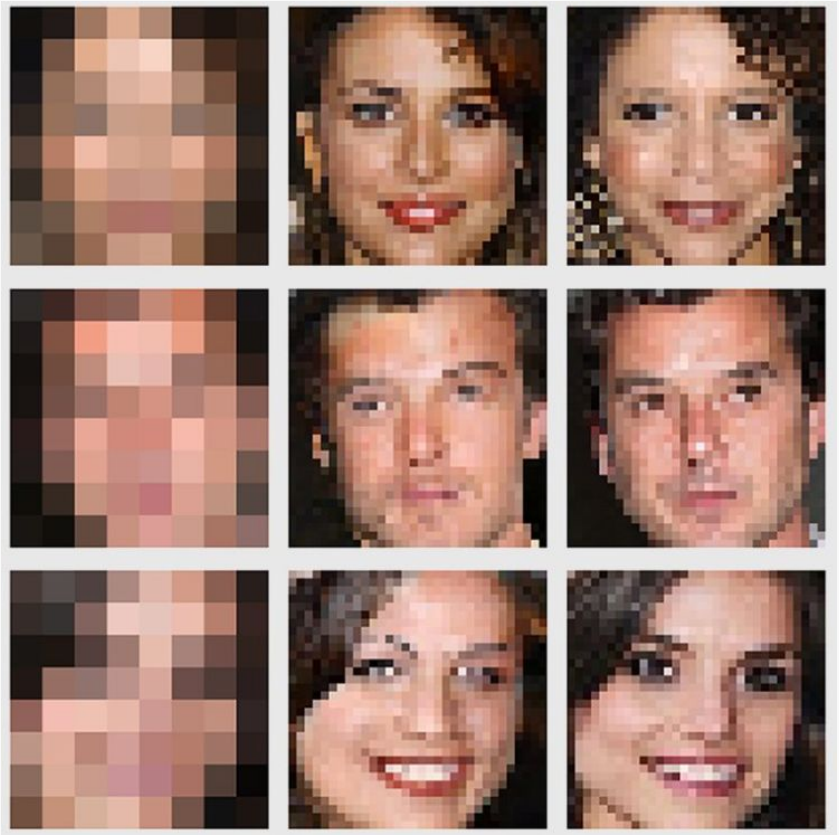
An app for blind people identifies and reads out objects in their surroundings

https://goo.gl/J9O1me

Vid2Speech: Speech Reconstruction from Silent Video

https://github.com/arielephrat/vid2speech

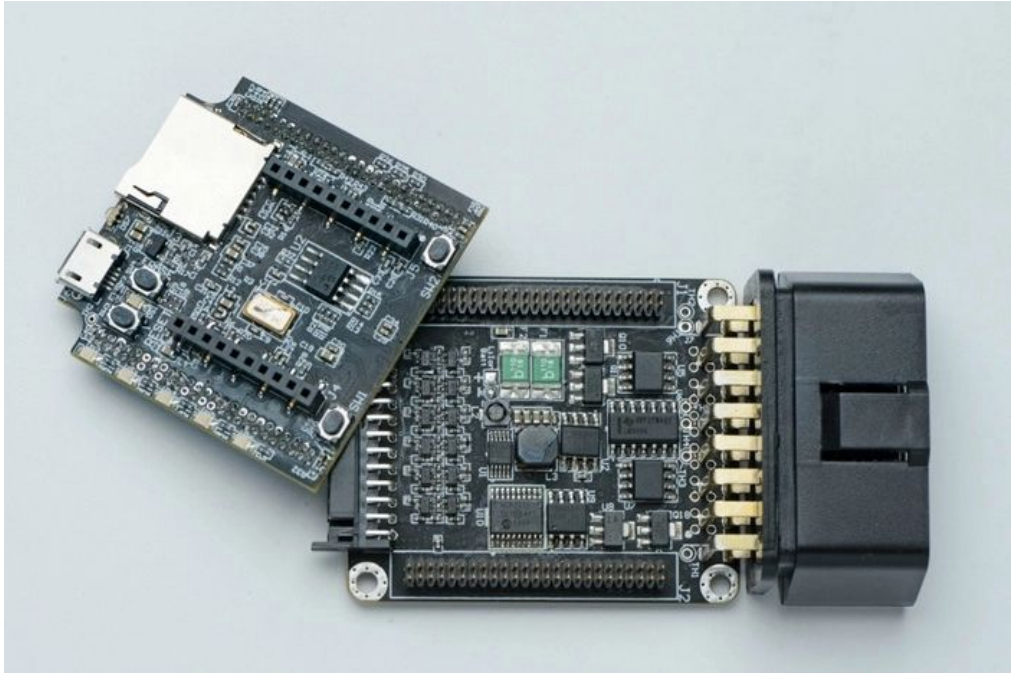
https://goo.gl/HPYCbK



Google Uses AI to ‘Rebuild’ a Portrait from an 8×8-Pixel Image

http://www.theverge.com/2017/2/7/14532206/google-brain-research-neural-networks-zoom-and-enhance-pixelated-images

IoT + Car Hacking + Big Data



http://hackaday.com/2017/02/21/first-look-macchina-m2/

KICKSTARTER

https://www.kickstarter.com/projects/1029808658/macchina-the-ultimate-tool-for-taking-control-of-y?token=e40d156d

https://comma.ai/

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

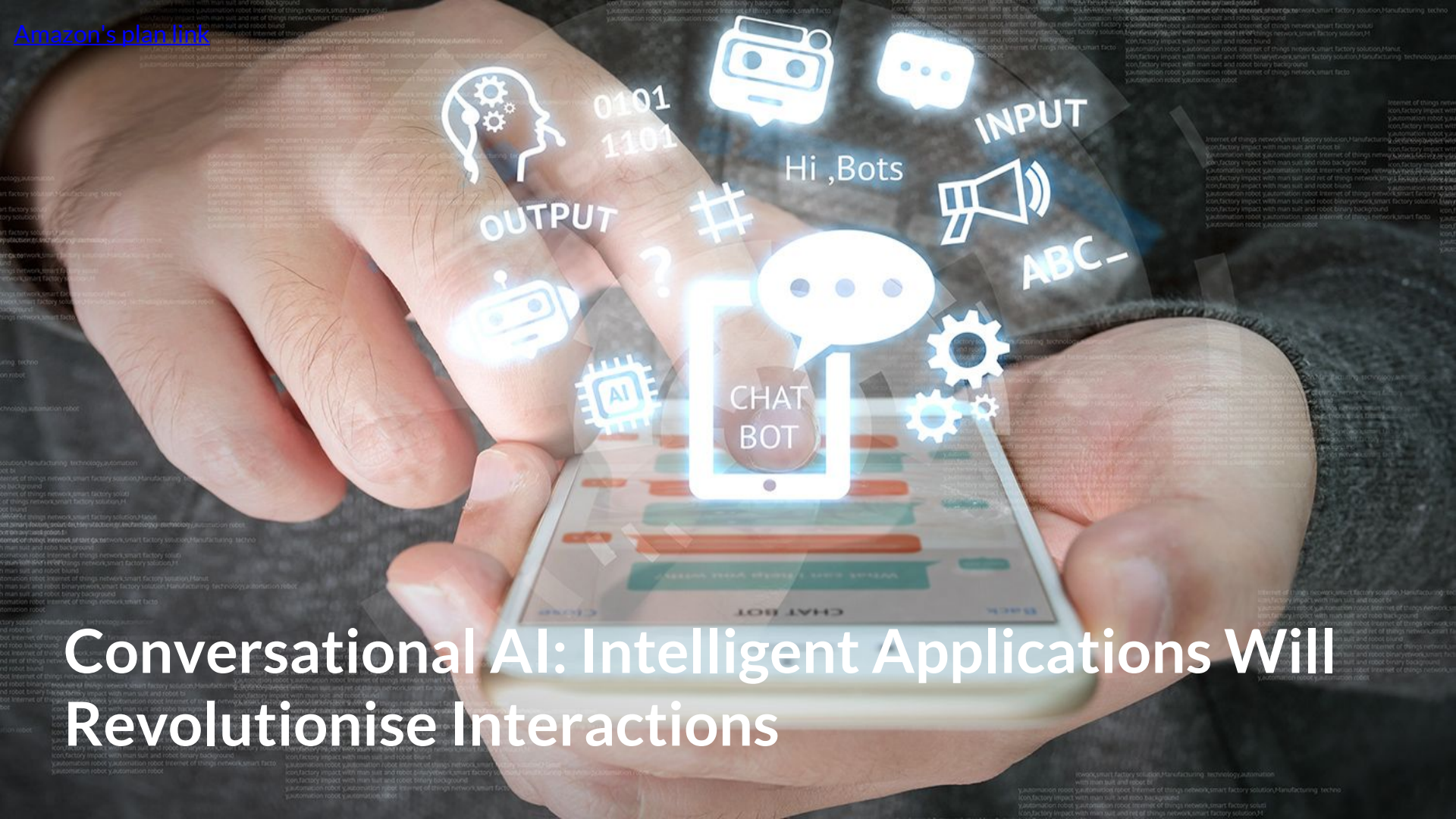
HOSPITAL SÍRIO-LIBANÊS

KUNUMI

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

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

Amazon's plan link



Conversational AI: Intelligent Applications Will Revolutionise Interactions

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

Spotify

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

BIG DATA

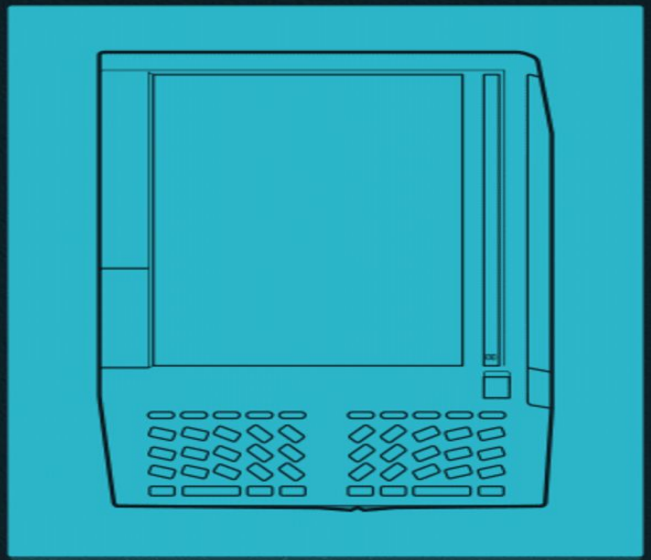
BIG NUMBERS


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

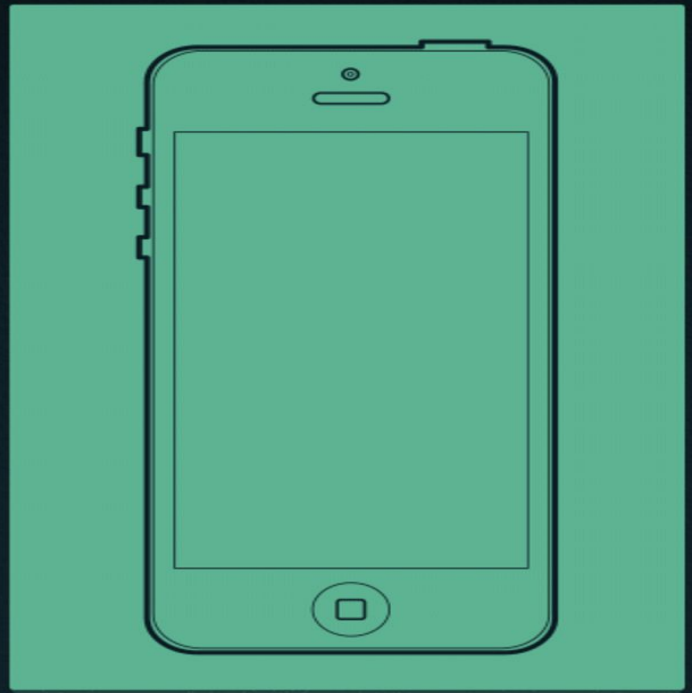
0.7
CD
80 mins

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

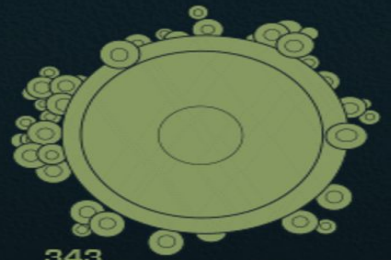
28.5
Large Hadron Collider
data produced per min

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

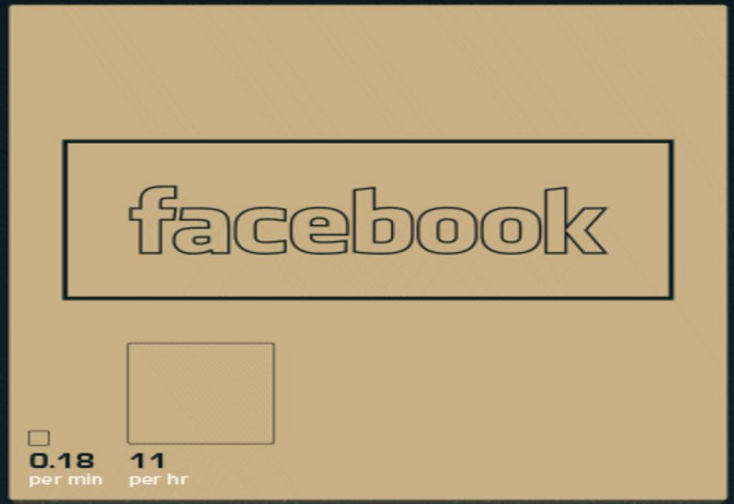
7.3
Wikipedia
all current articles with edit history

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

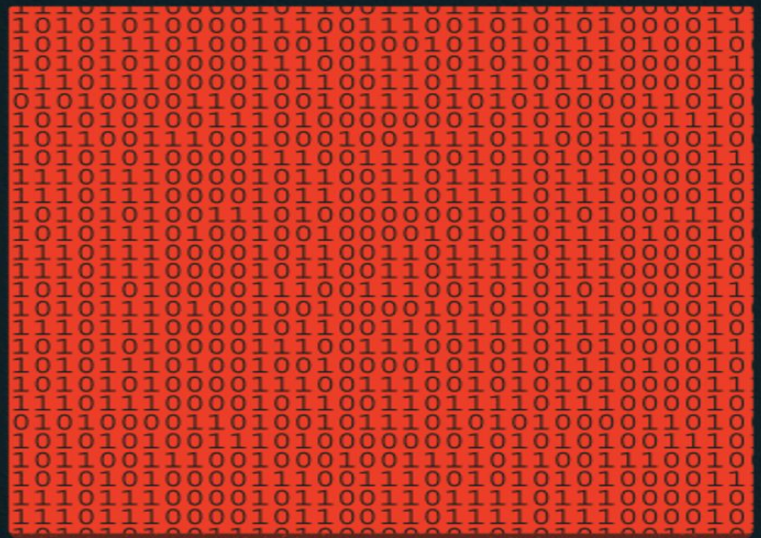
0.5
2008

0.8
2009

1.2
2010

1.8
2011

2.8
2012



30ZB
All global data
2019 (predicted)

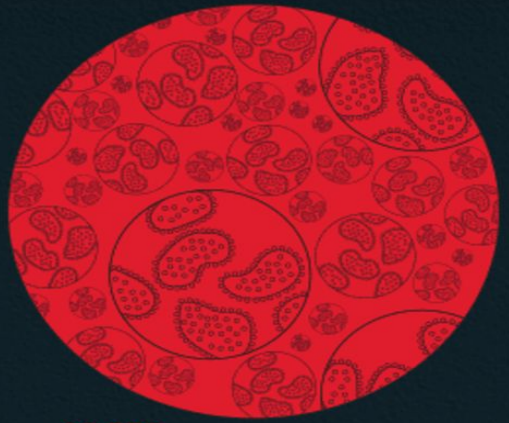
YOTTABYTES

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.

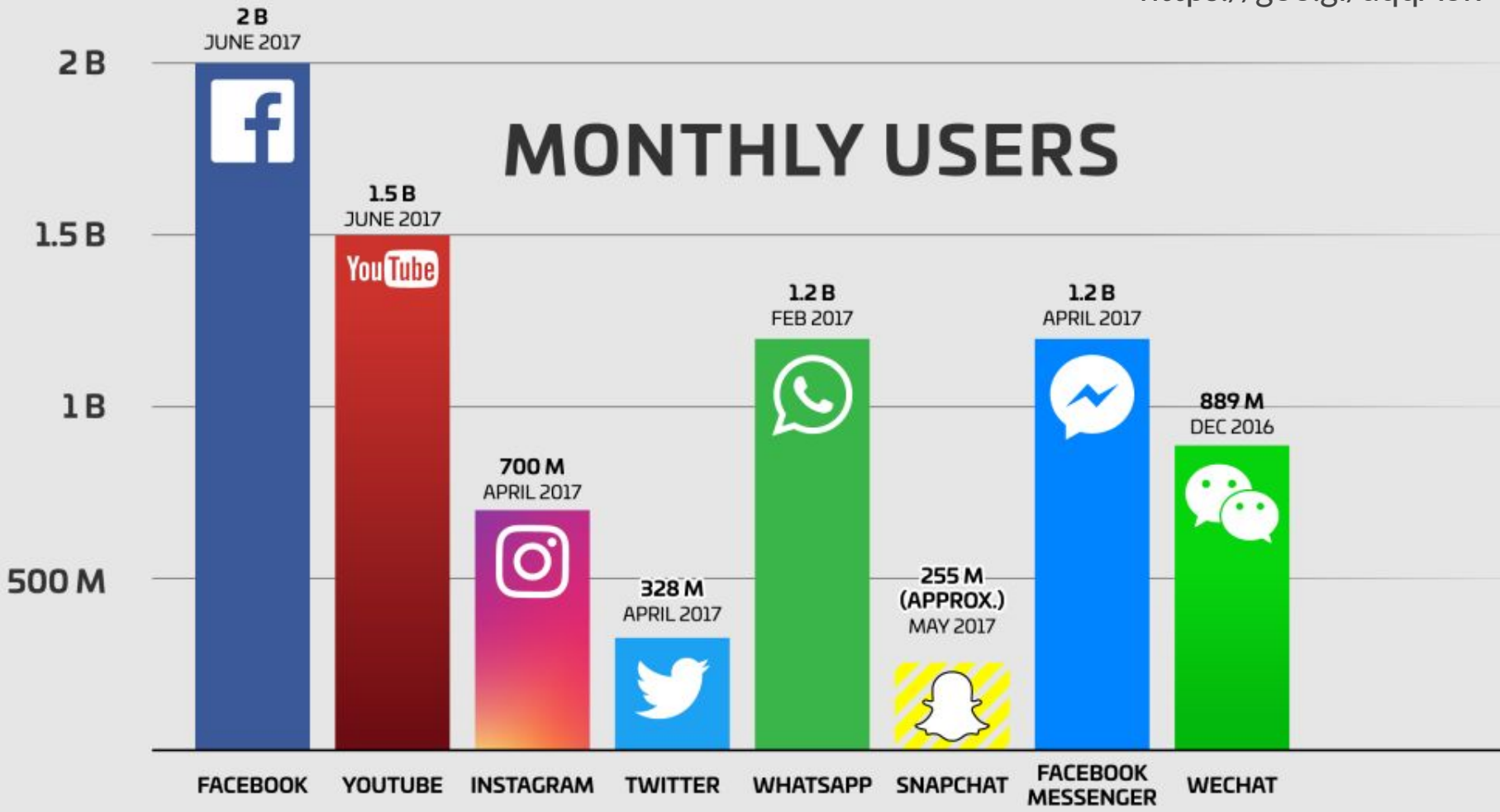
By Information is Beautiful Studio for BBC FUTURE

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

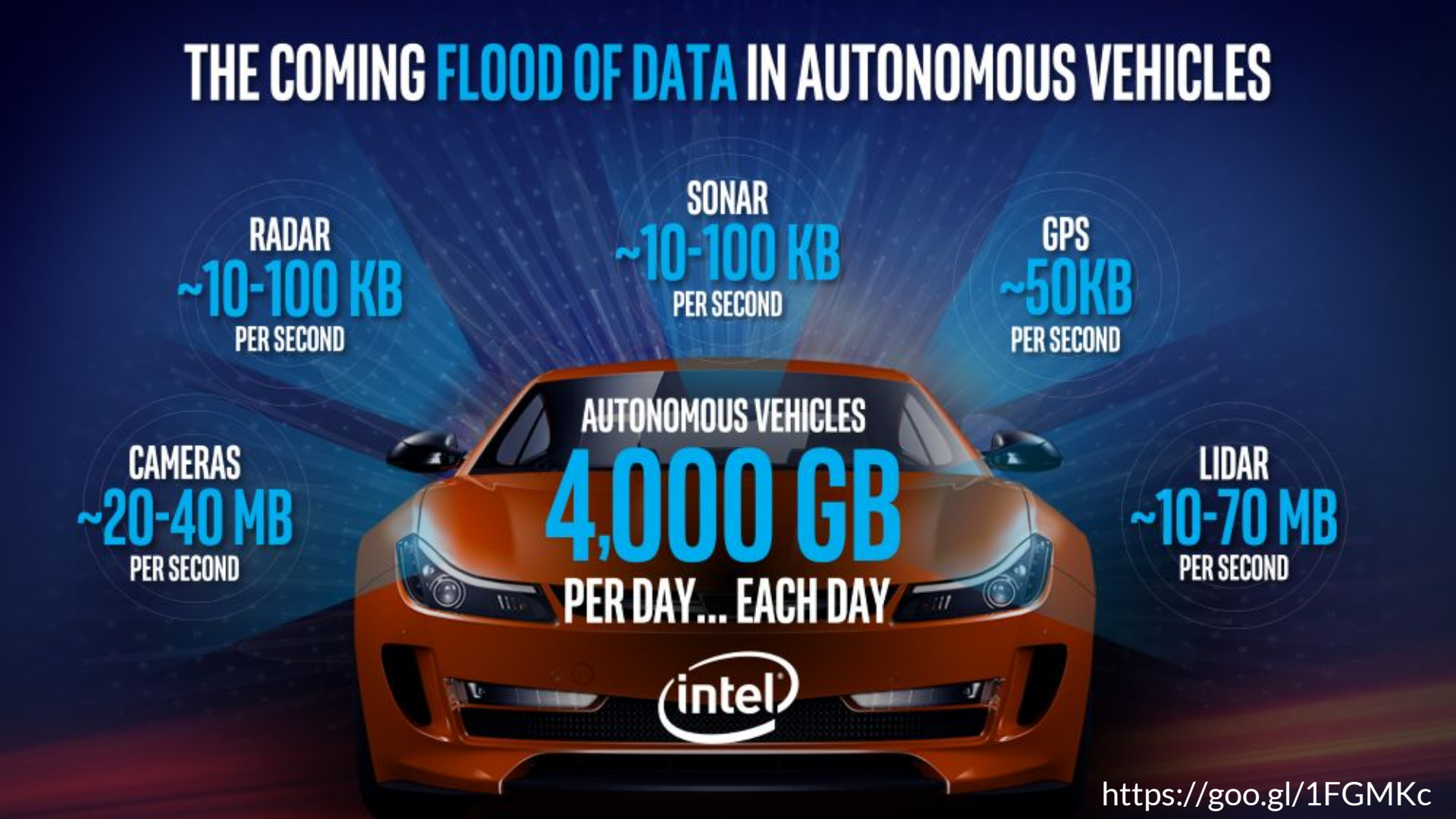
BIG SCIENCE

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

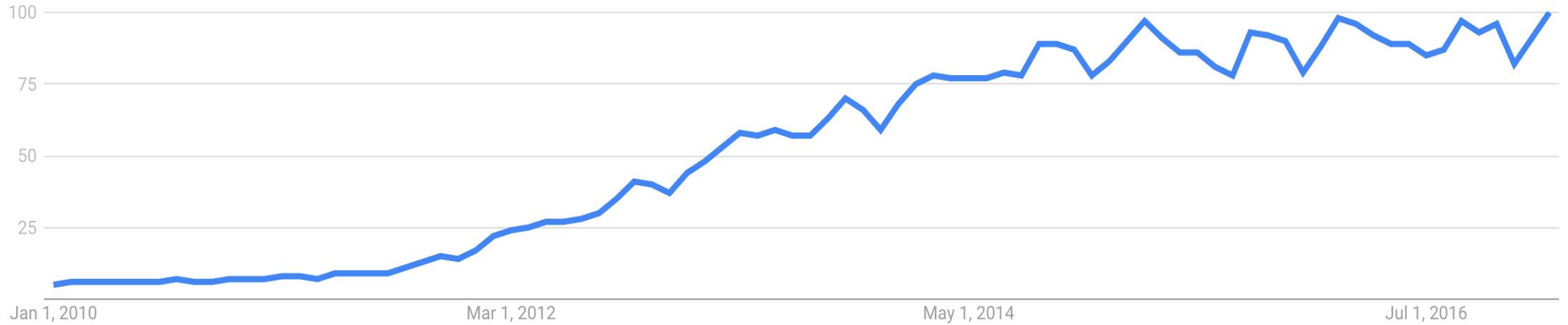
Big Data
Search term

+ 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

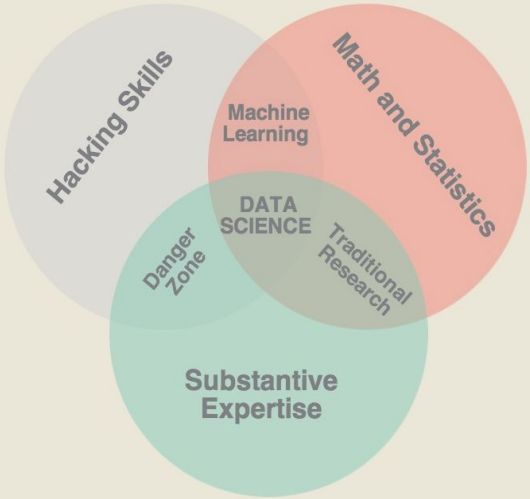
DataCamp
Learn data analysis for free, interactively

Become a

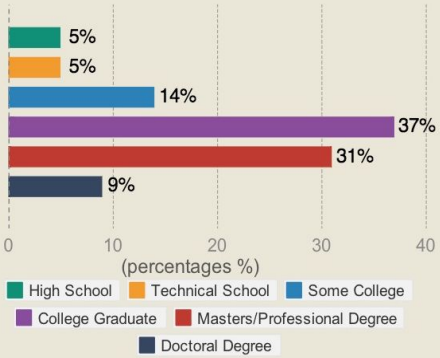
Data Scientist

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

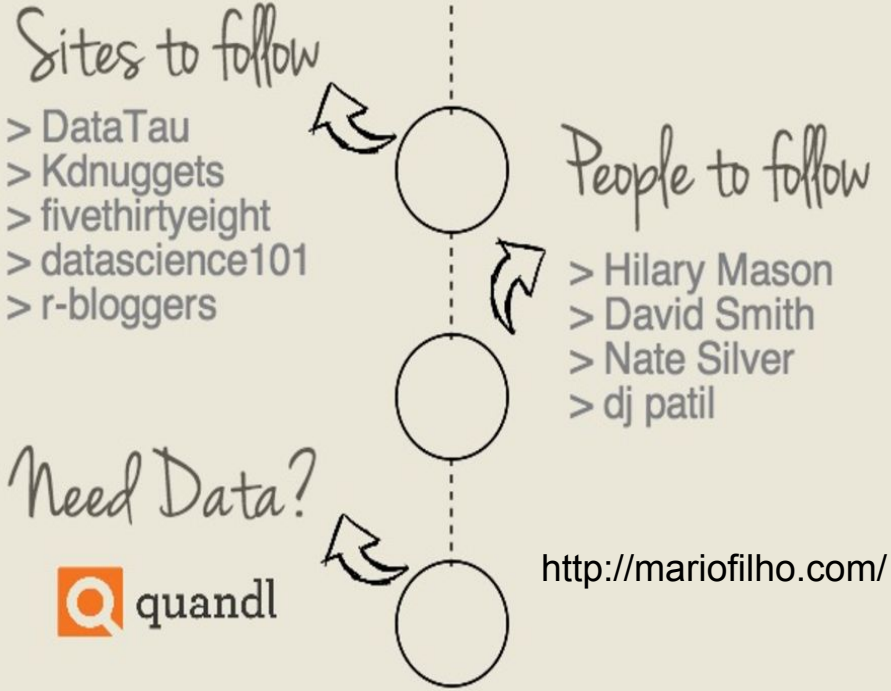
The best way to find out whether you are a true data scientist or not is to take the bull by the horns and to enter the real-life jungle of data-analysis and science with your freshly acquired skill set.

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

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

https://goo.gl/CvLVpu

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, …)

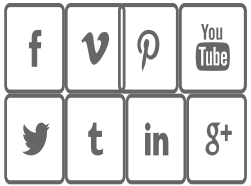
API

Sites

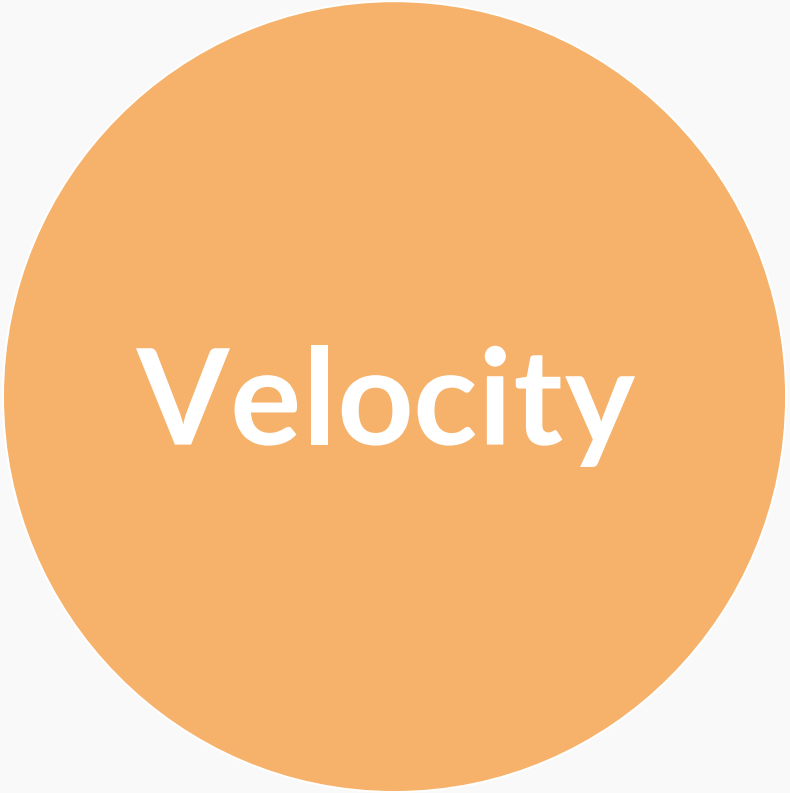
Text and reports

Maps

Image and videos



Social Media


Velocity

2017 This Is What Happens In An Internet Minute


Veracity

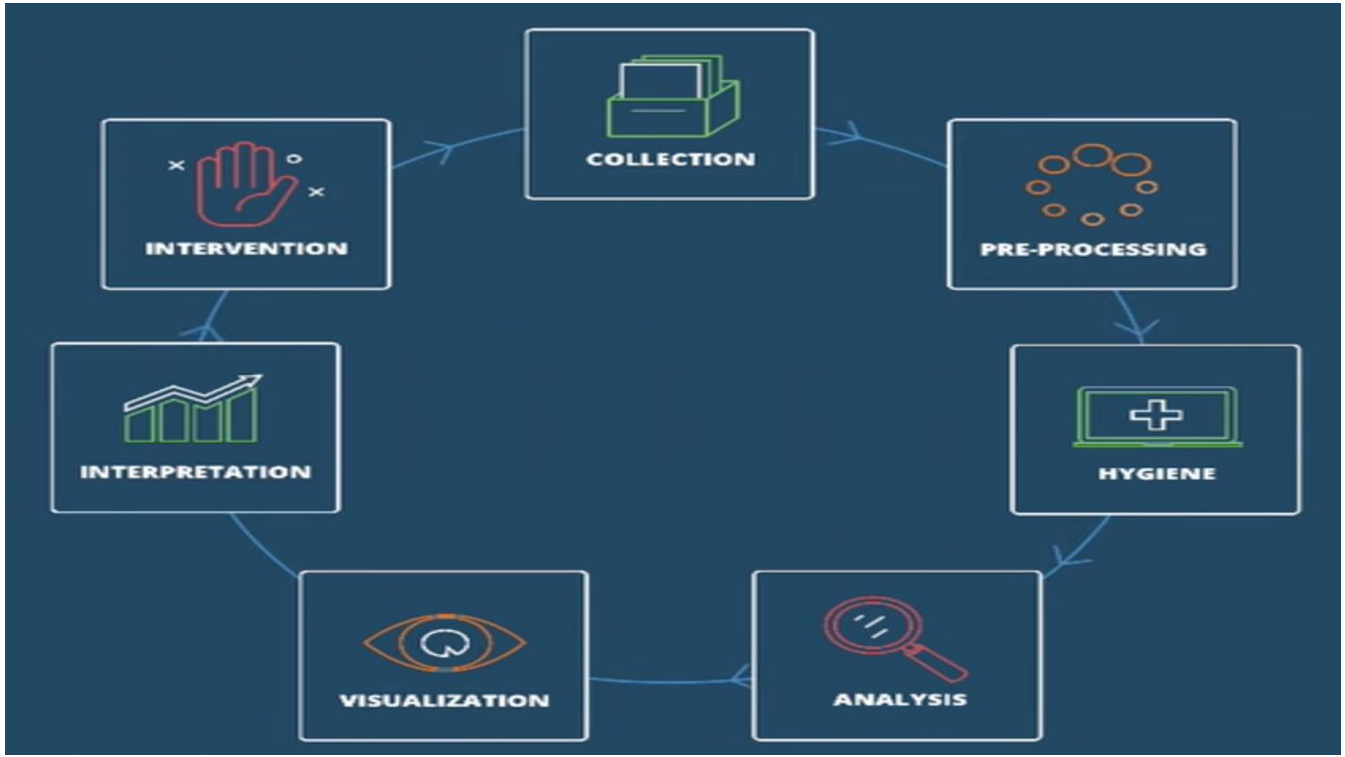
Boatos.org

Investigando Rumores Desde 2002
E-FARSAS®.com

Snopes

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

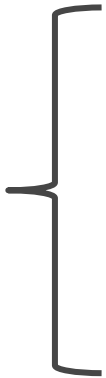
Key steps in a Data Science Analysis



Key steps in a Data Science Analysis

Data Preparation

- Pre-processing
- Hygiene



Data Format
Variable selection
Feature Engineering (dimension reduction)
Missing Data

Key steps in a Data Science Analysis

Algorithms Selection

- Analysis

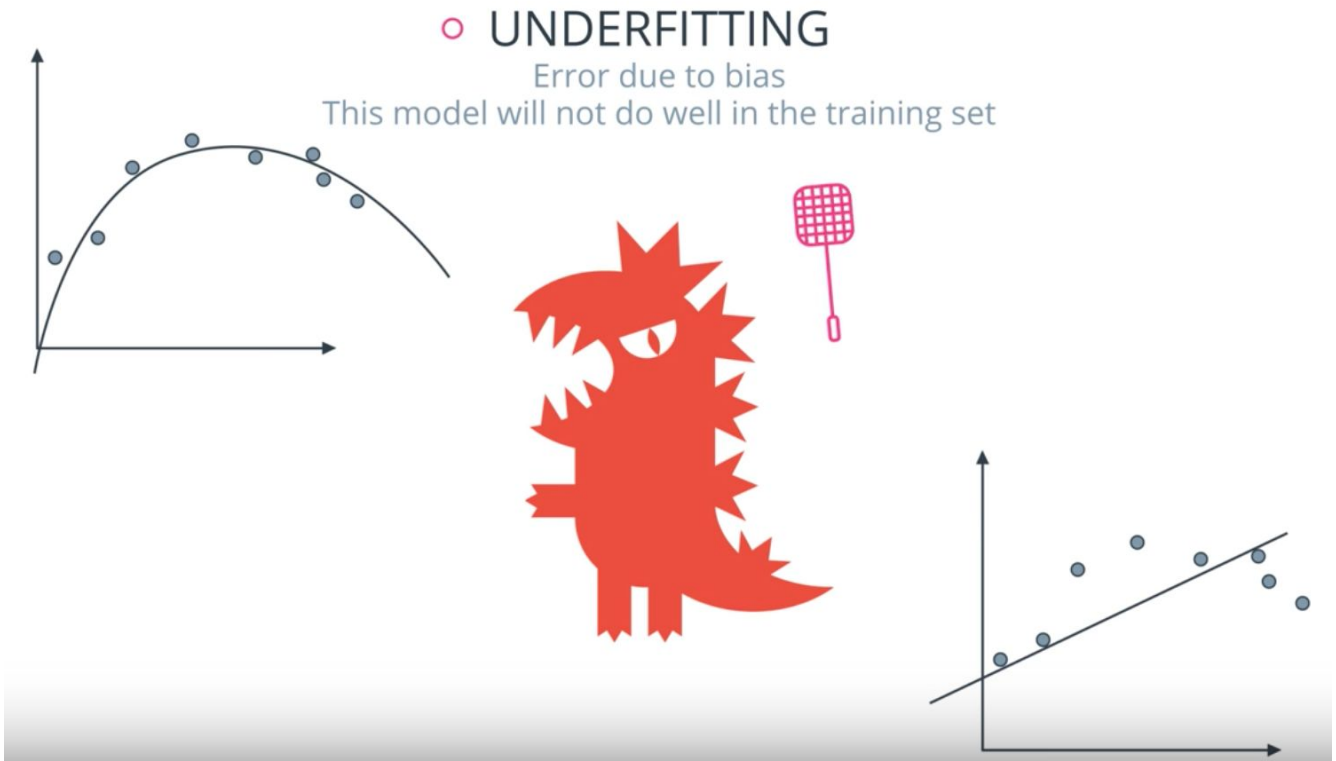
	Algorithms
Unsupervised Learning	*k*-Means Clustering Principal Component Analysis Association Rules Social Network Analysis
Supervised Learning	Regression Analysis *k*-Nearest Neighbors Support Vector Machine Decision Tree Random Forests Neural Networks
Reinforcement Learning	Multi-Armed Bandits

Key steps in a Data Science Analysis

Parameter tuning

- Analysis

Key steps in a Data Science Analysis

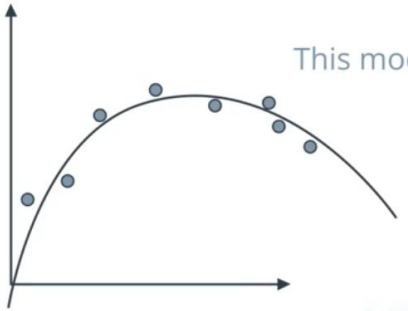


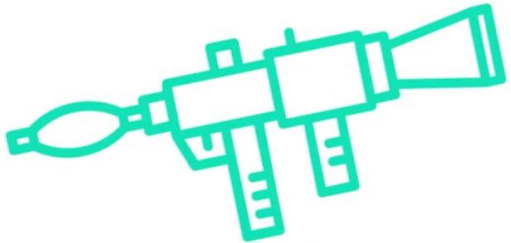
Key steps in a Data Science Analysis

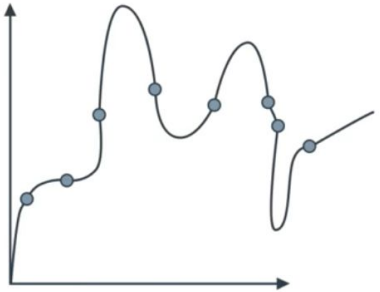
- OVERFITTING

Error due to variance

This model performs poorly in the testing set







Key steps in a Data Science Analysis

CONFUSION MATRIX

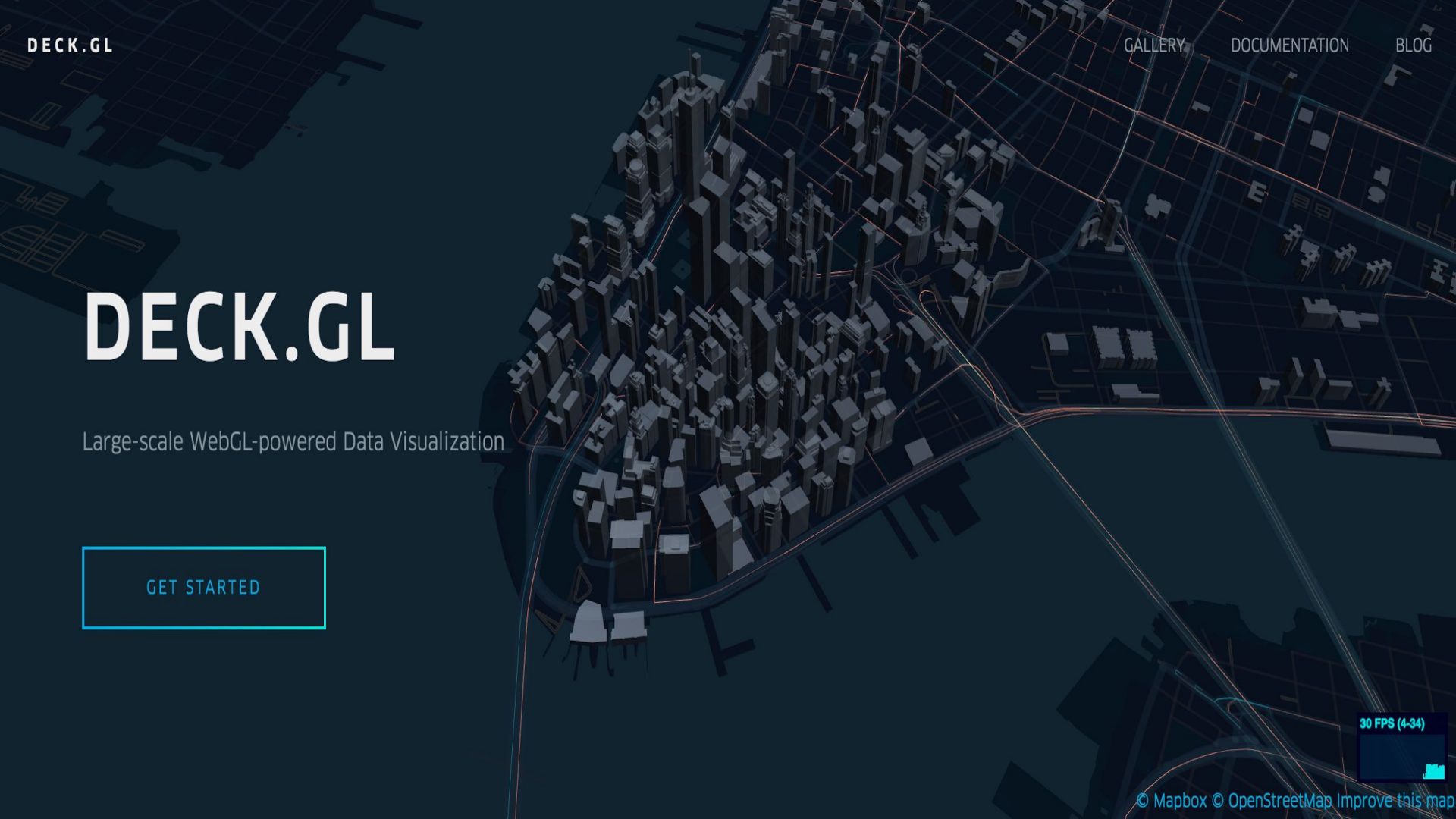
Evaluating Results

- Visualization
- Interpretation

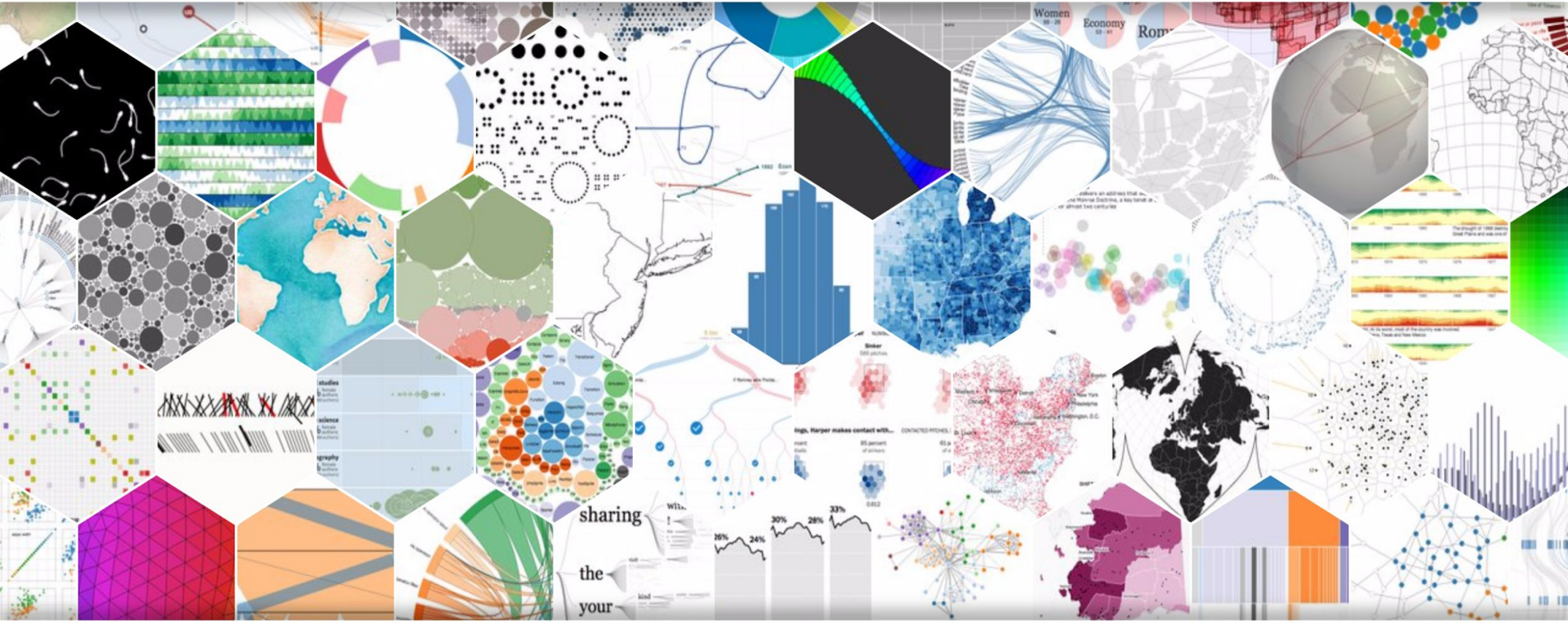
10, 000 PATIENTS

DIAGNOSIS

PATIENTS	Diagnosed Sick	Diagnosed Healthy
Sick	1000 True Positives	200 False Negatives
Healthy	800 False Positives	8000 True Negatives


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



Search and add articles

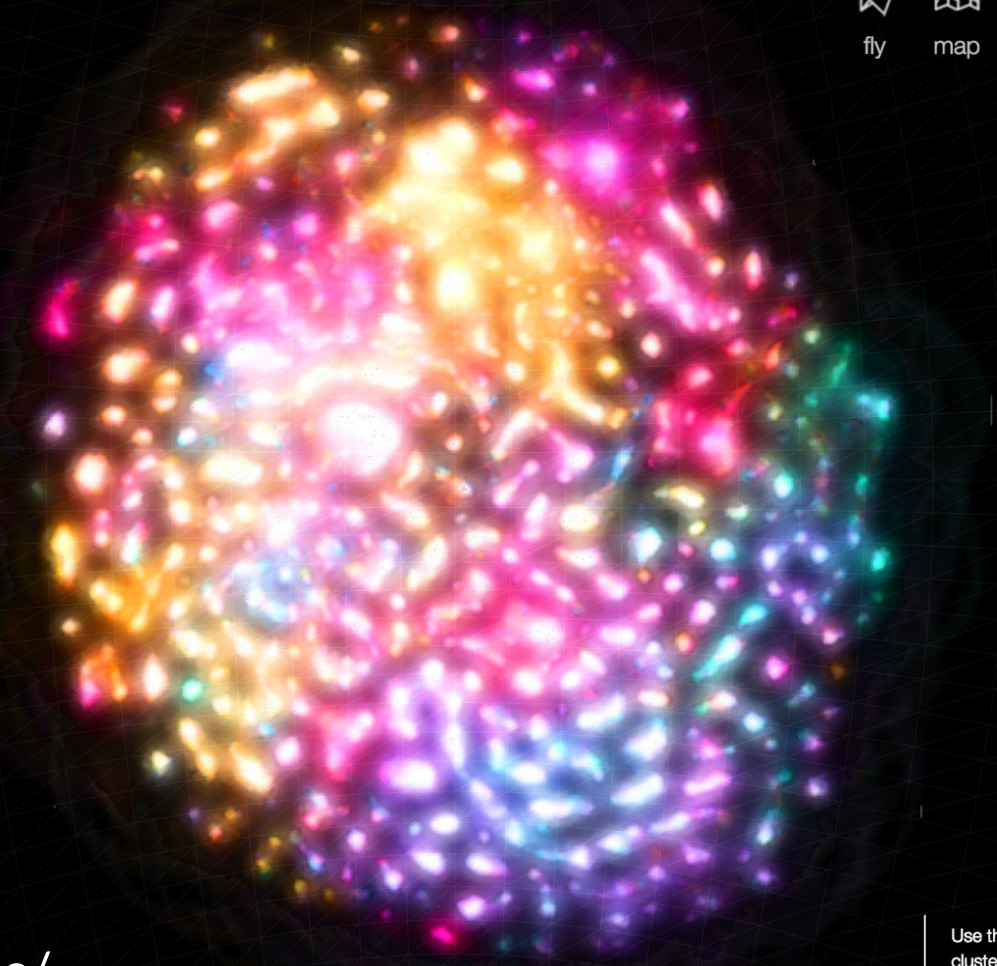
fly map home help about UI off full

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/

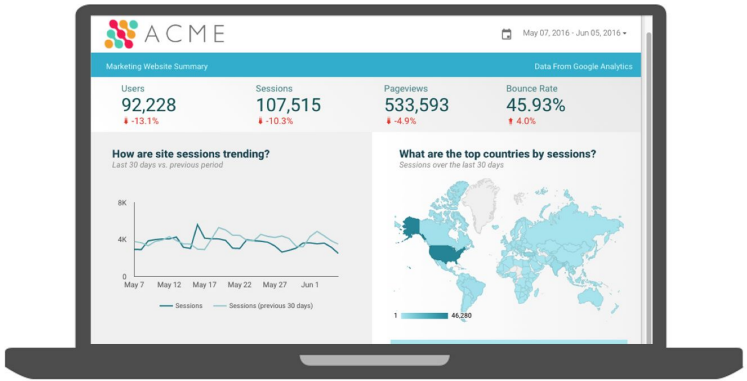
Use the mouse to see a preview of articles in each cluster
Click anywhere on the map to fly there

http://www.pentaho.com/

https://www.tableau.com/

https://powerbi.microsoft.com

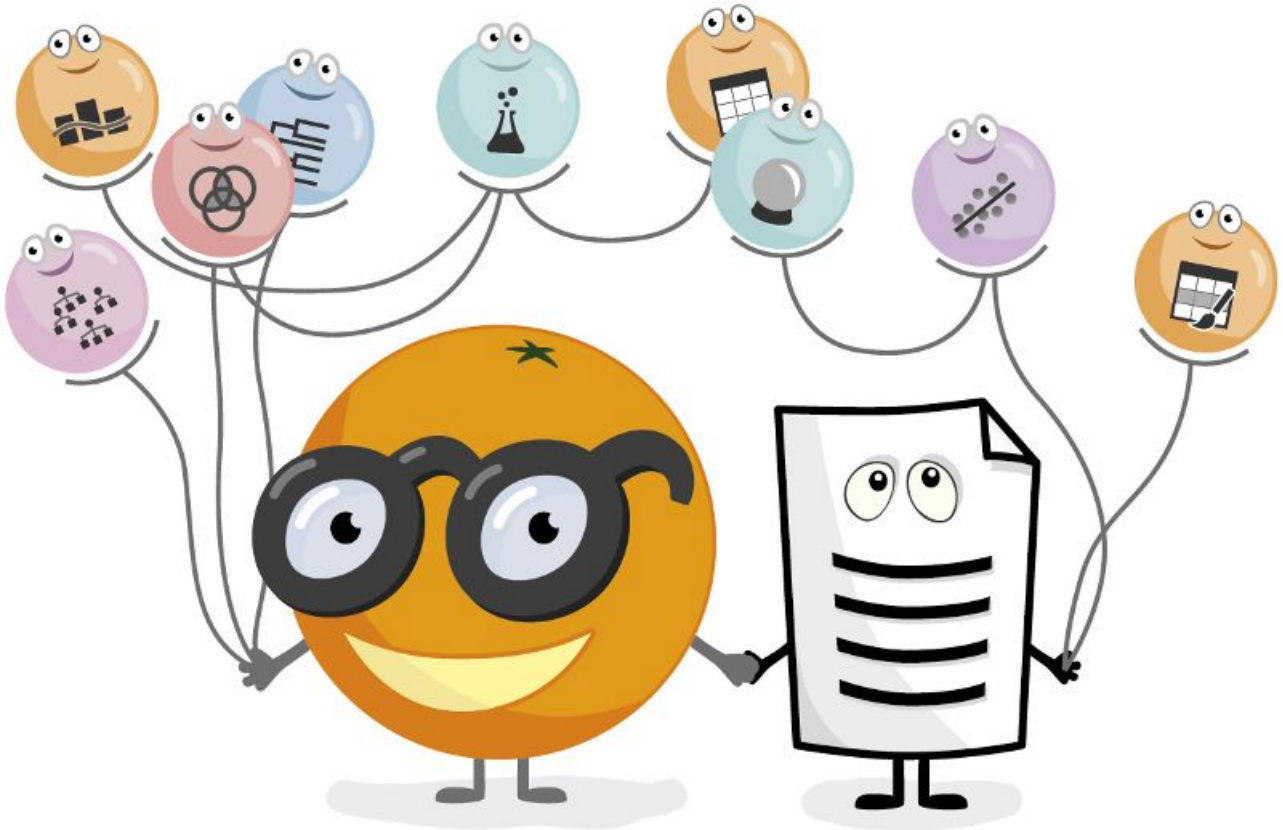
Google Data Studio



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

https://goo.gl/Ndf38Q

Data Mining Fruitful and Fun

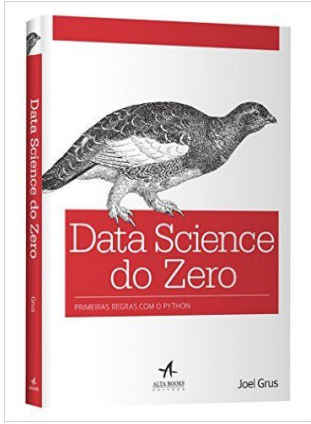


https://orange.biolab.si/

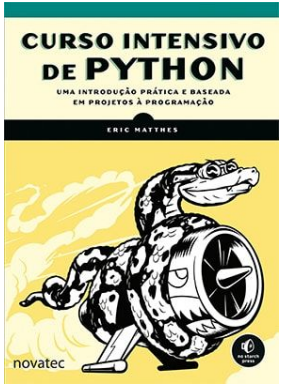
Agenda

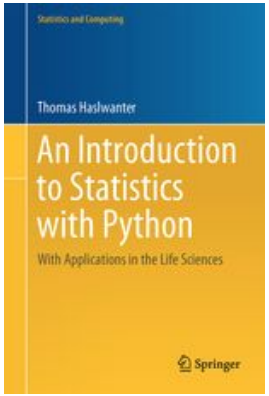
- Python Introduction
- Data Analysis and Visualization
- Working with data sources
- Probability & Statistics
- Machine Learning
- Network Analysis
- Working with large datasets
- Advanced topics in data science

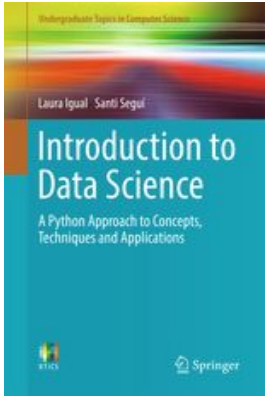
References

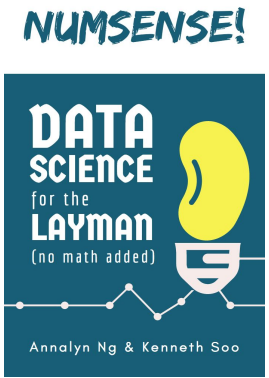

Data Science do Zero
Data Science do Zero
PRIMEIRAS REGRAS COM O PYTHON
Joel Grus

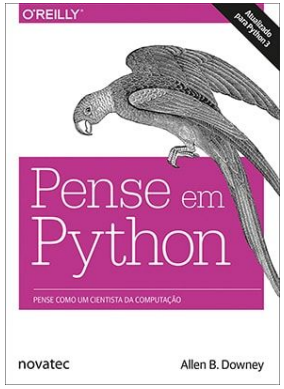

Fernando Amaral
INTRODUÇÃO À CIÊNCIA DE DADOS
Fernando Amaral
INTRODUÇÃO À CIÊNCIA DE DADOS
MINERAÇÃO DE DADOS E BIG DATA

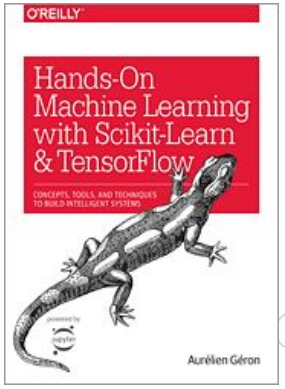

CURSO INTENSIVO DE PYTHON
UMA INTRODUÇÃO PRÁTICA E BASEADA EM PROJETOS À PROGRAMAÇÃO
ERIC MATTHES
novatec


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

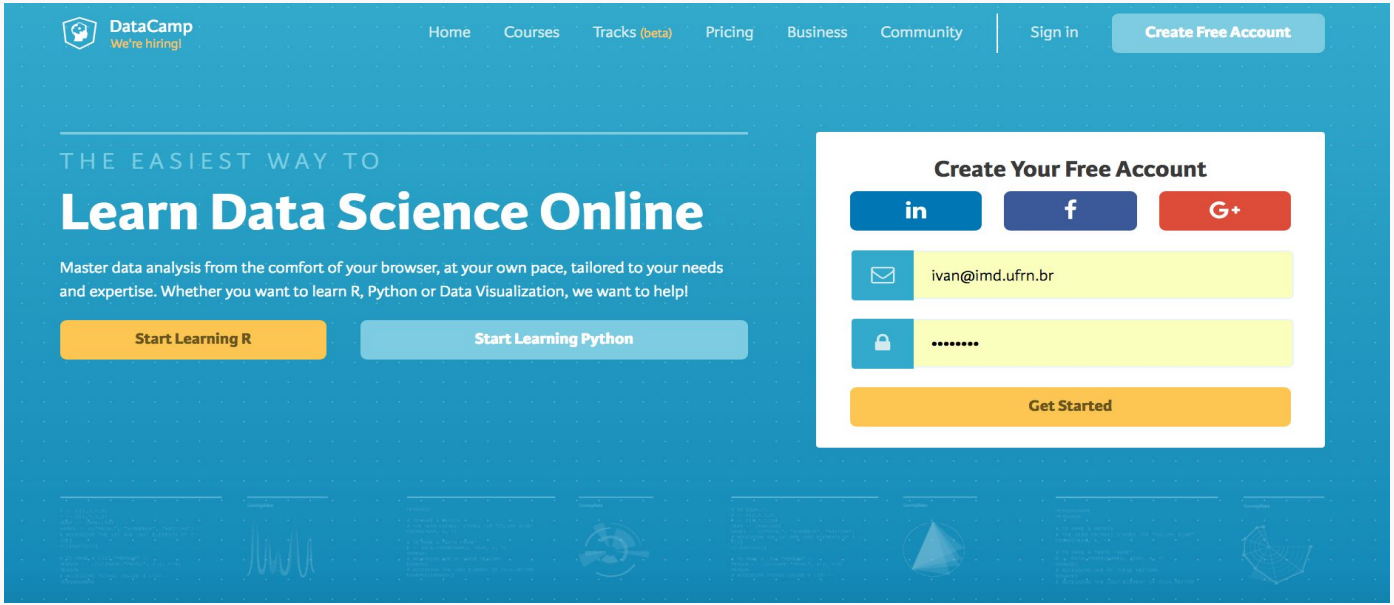

O'REILLY
Atualizado para Python 3
Pense em Python
PENSE COMO UM CIENTISTA DA COMPUTAÇÃO
novatec
Allen B. Downey


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

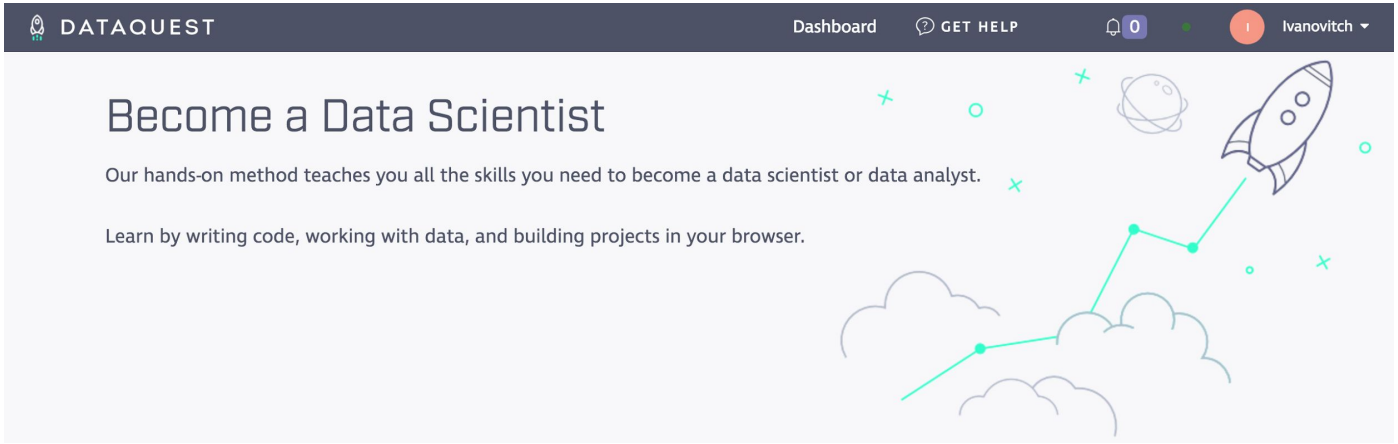
References

UDACITY

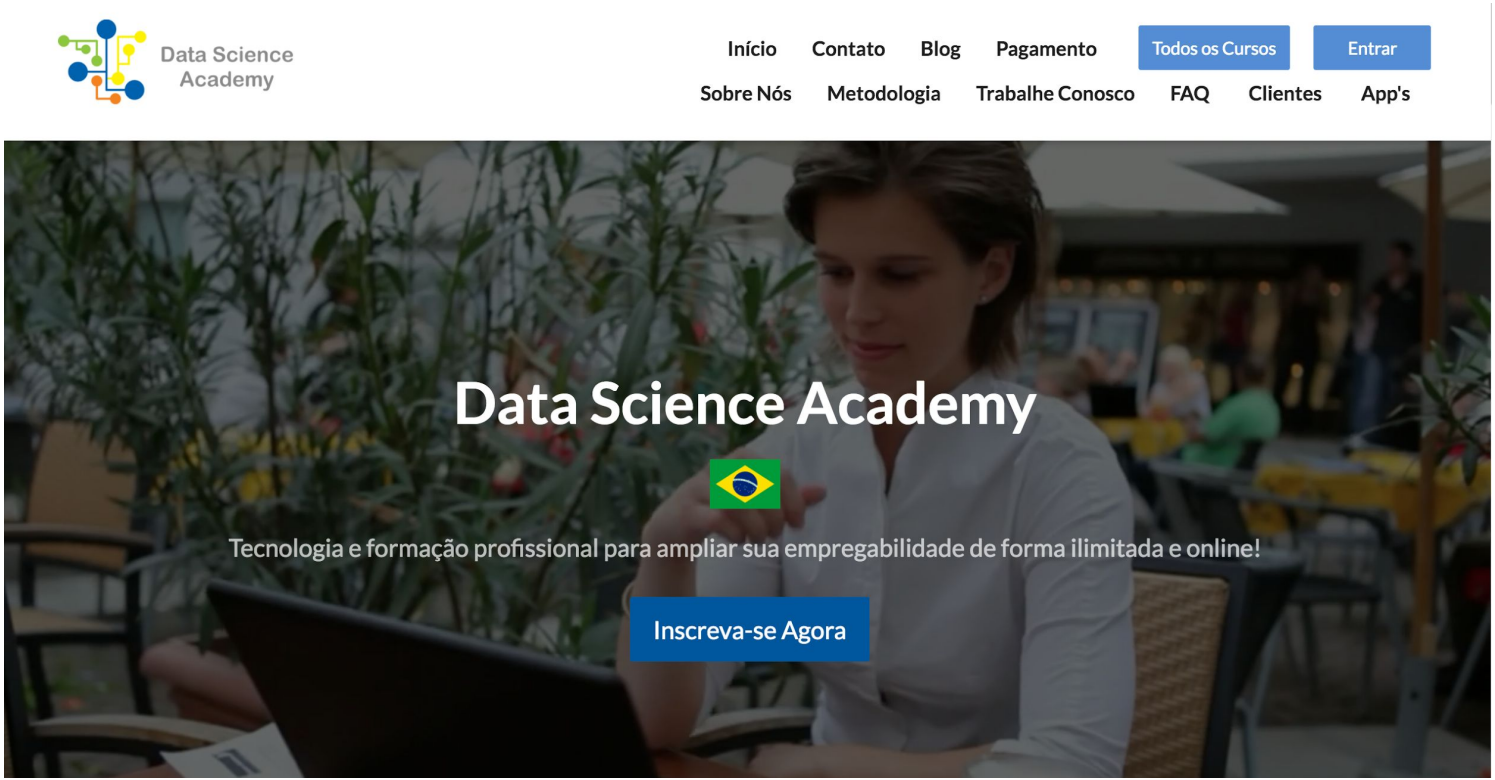
References


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

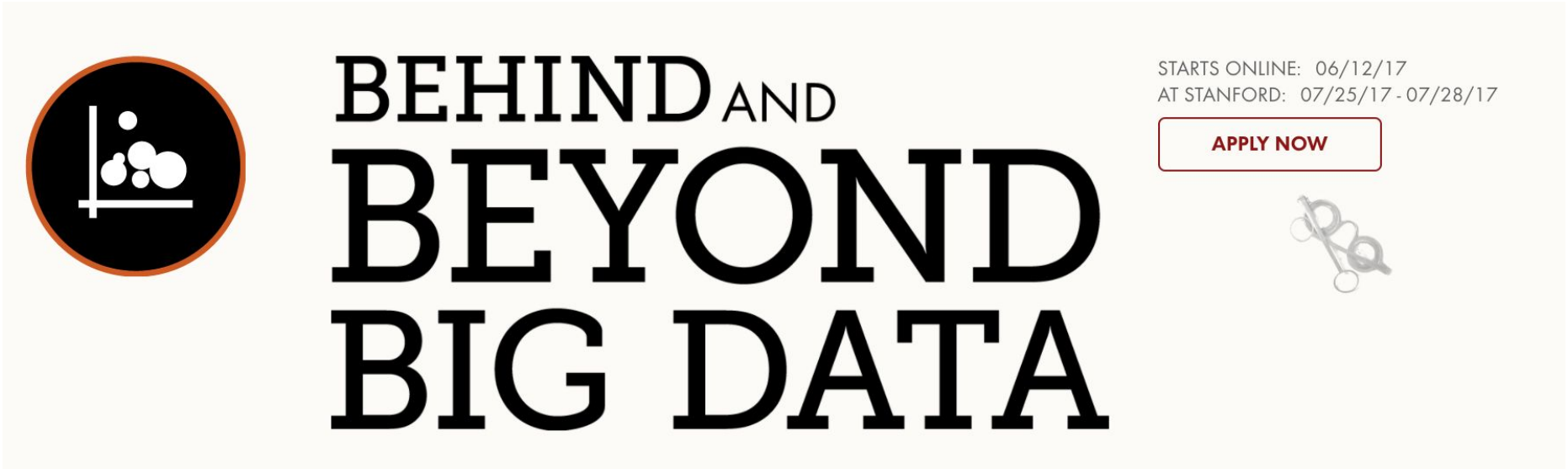

Data Science Academy
Início
Contato
Blog
Pagamento
Todos os Cursos
Entrar
Sobre Nós
Metodologia
Trabalhe Conosco
FAQ
Clientes
App's
Data Science Academy
Tecnologia e formação profissional para ampliar sua empregabilidade de forma ilimitada e online!
Inscreva-se Agora

References

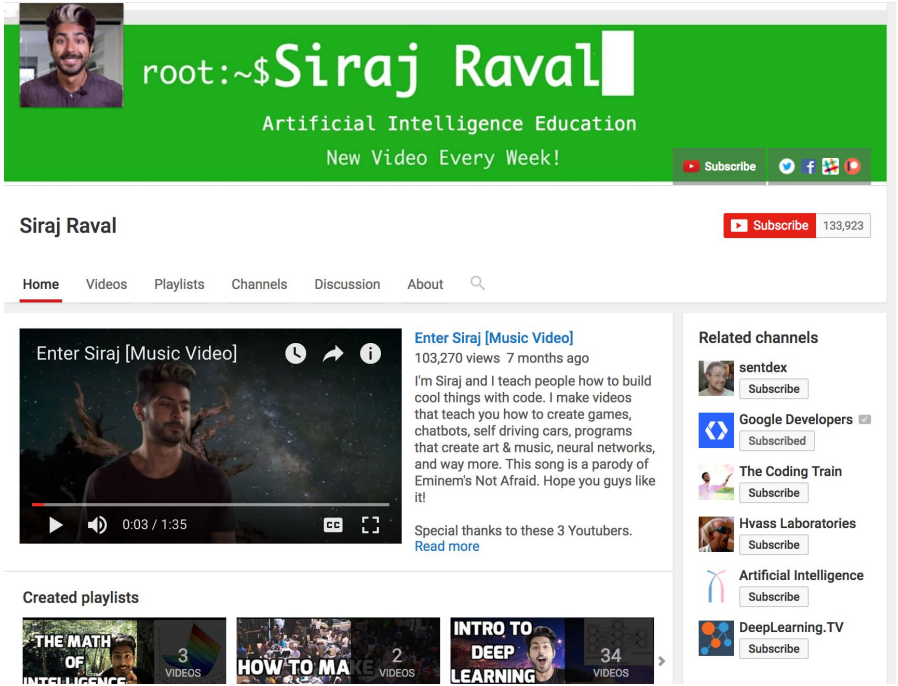

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Big Data and Social Analytics certificate course
EXperimental Learning
2017 DATES TO BE CONFIRMED
DOWNLOAD COURSE PROSPECTUS
Discover a new way to think about big data analysis when you explore the theory behind "social analytics", and practically apply that knowledge 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/
https://github.com/data-8/
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

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

Lesson #1