IMD0105 - Special Issues in Information Technology VI

Big Data & Analytics

Natal-RN
Fevereiro de 2017


Introduction

Ivanovitch Silva (ivan@imd.ufrn.br)

Office Hours (Friday 1pm to 5pm)

ICE BREAKER

I AM ...
WHAT I LOOK FOR?
WHY DID I STOP HERE?

WHO AM I ON THE ROW OF BREAD?

Group Knowledge

Take a Survey

https://goo.gl/forms/odBre7a53hq5KpYg2

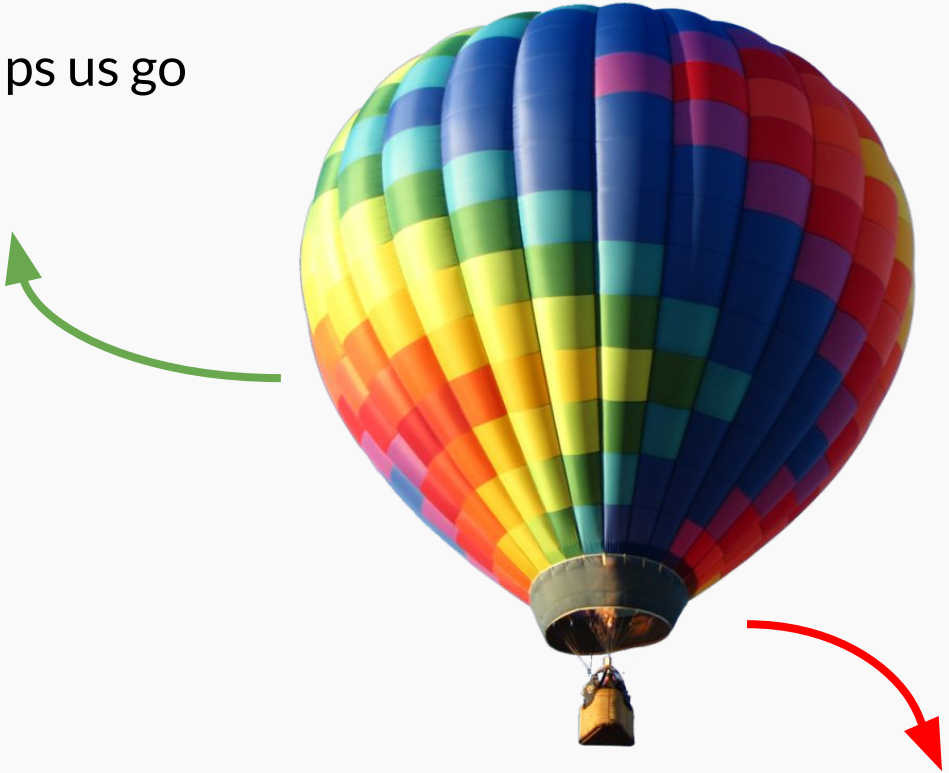
Future or Present?

While studying at Federal University of Rio Grande do Norte (UFRN), I built …

Hot-air-Balloon

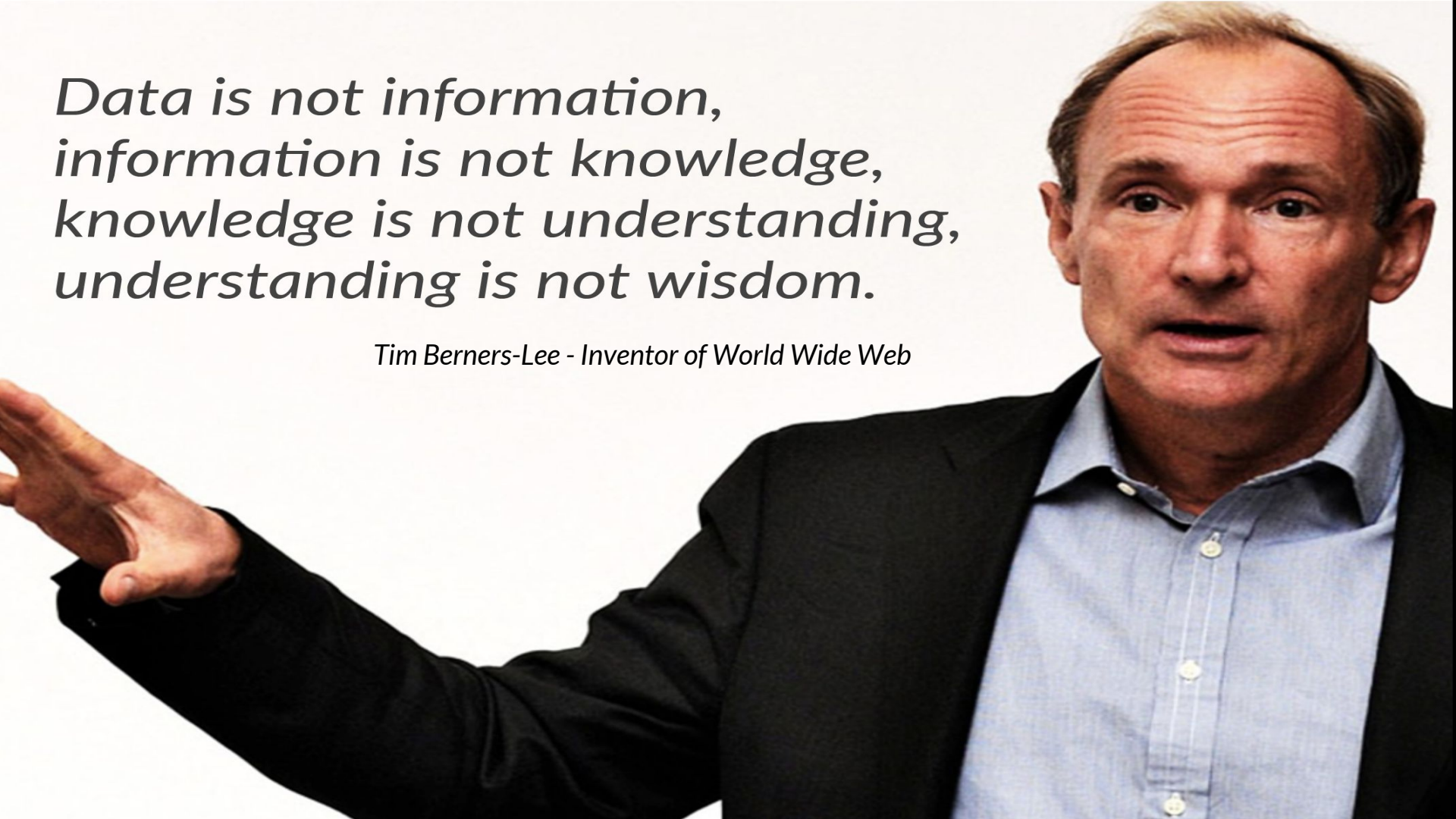
http://www.funretrospectives.com/hot-air-balloon-bad-weather/

What helps us go higher?



Which are the forces pulling us down?


About Data


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

Data

Data Files
(XML, CSV, Excel, JSON, …)

Database
(MySQL, Oracle, …)

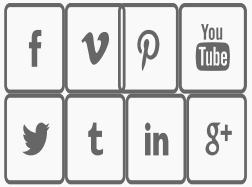
API

Sites

Text and reports

Maps

Image and videos

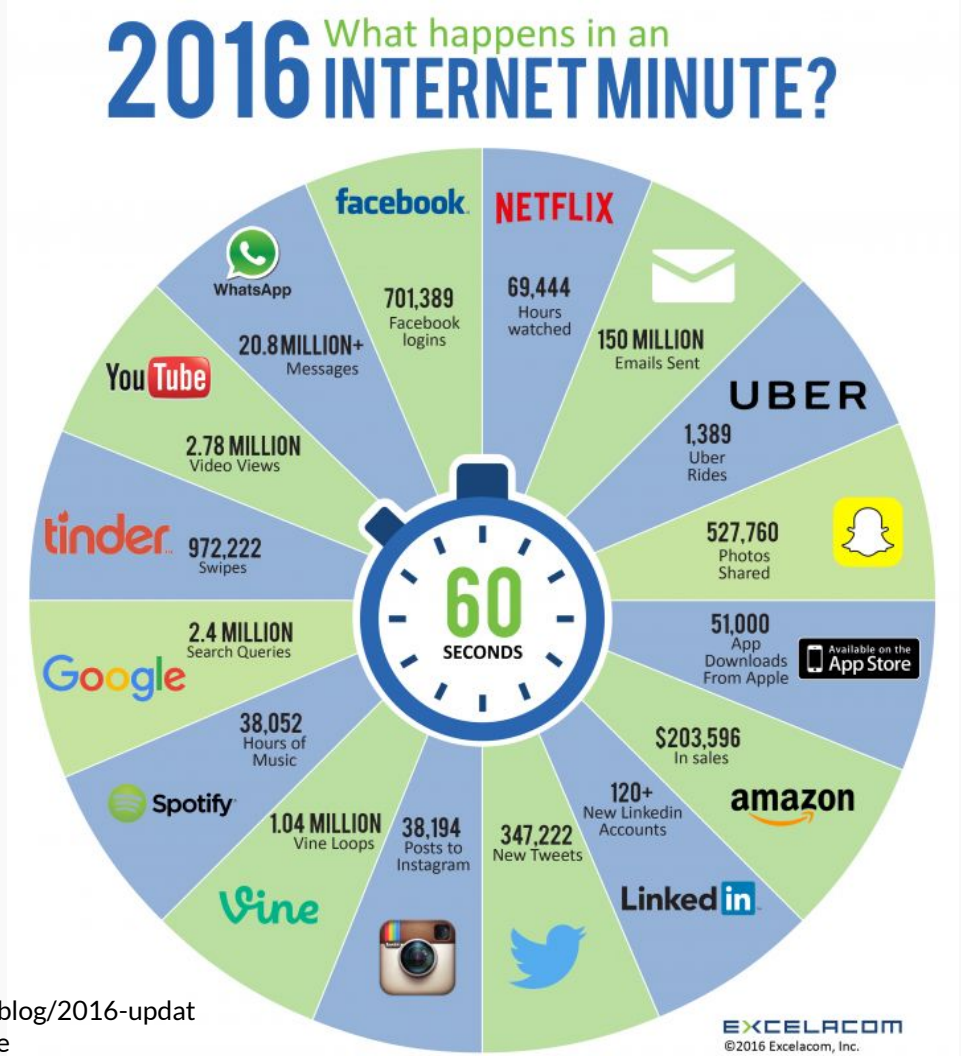


Social Media

BIG DATA

Size Unit

Unit	Size	Store ~
Byte	1	1 character
Kilobyte	1,000	2 or 3 paragraphs of text
Megabyte	1,000,000	873 pages of plaintext
Gigabyte	1,000,000,000	341 digital pictures
Terabyte	1,000,000,000,000	All the catalogue book in US Library of Congress
Petabyte	1,000,000,000,000,000	Google processes around 1 PB every hour
Exabyte	1,000,000,000,000,000,000	2.5 Exabyte are produced every day
Zettabyte	1,000,000,000,000,000,000,000	Annual global IP traffic (2016)
Yottable	1,000,000,000,000,000,000,000,000	A yottabyte of storage would cost $8 trillion for raw drives and $80 trillion for a storage system



http://www.excelacom.com/resources/blog/2016-update-what-happens-in-one-internet-minute

BIG SCIENCE

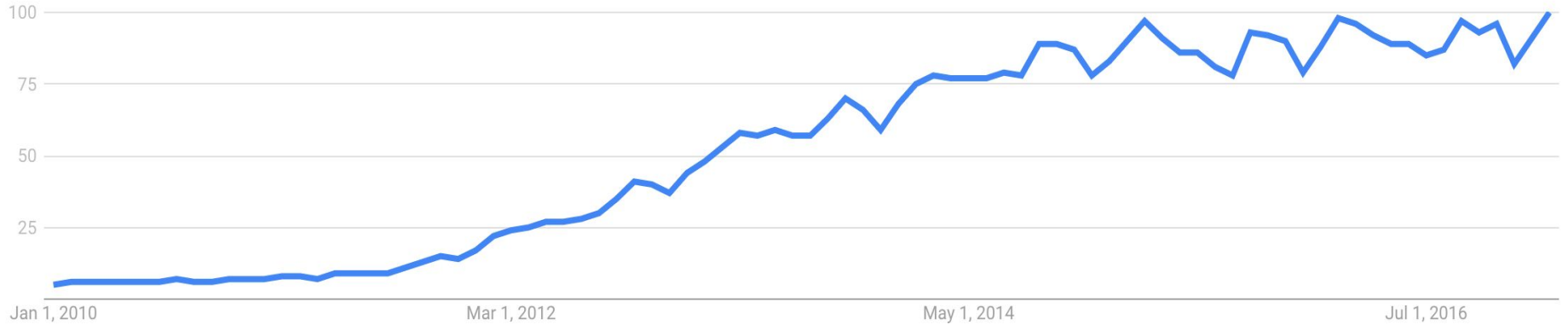
Big Data
Search term

+ Compare

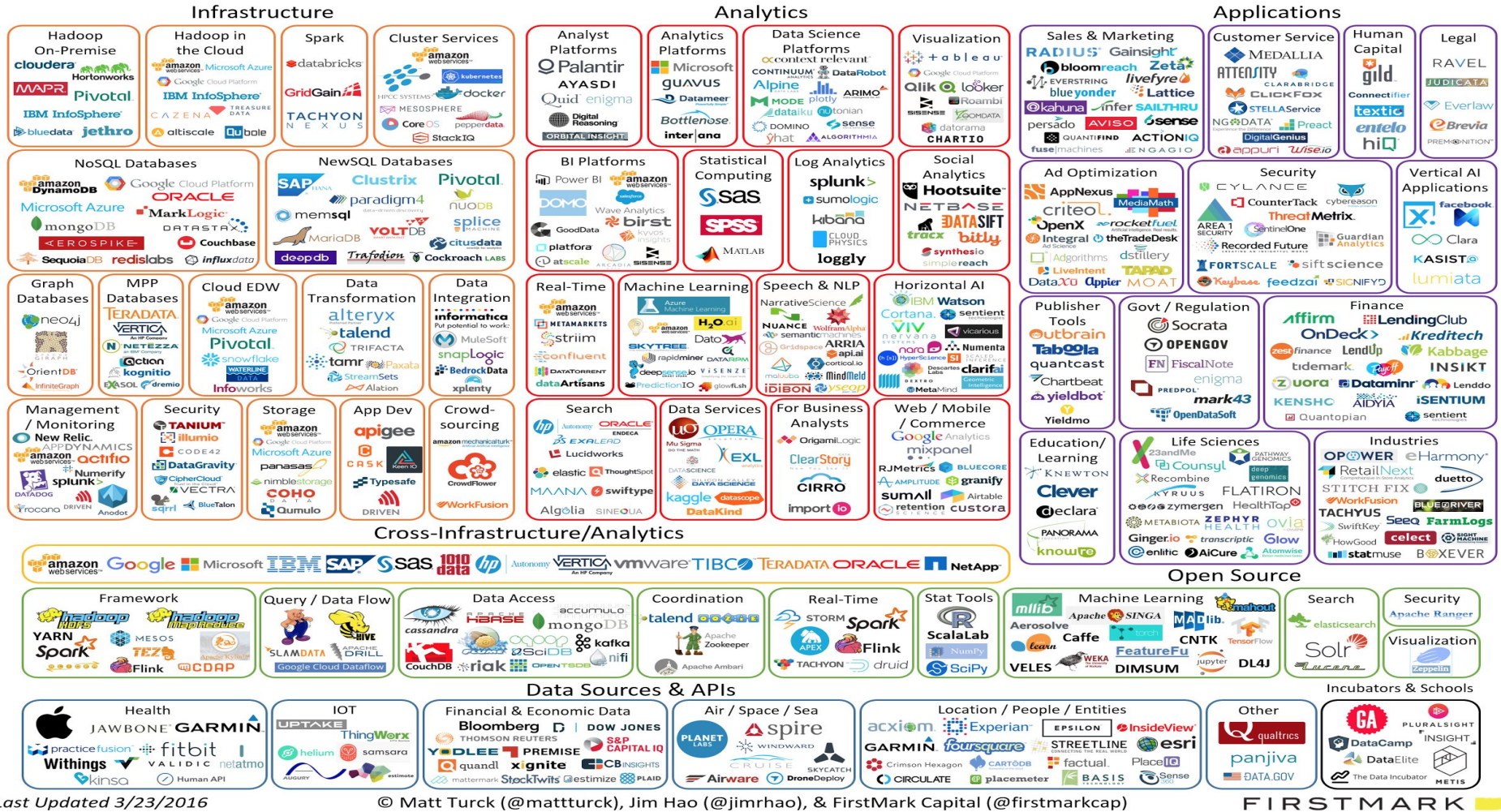
https://trends.google.com/trends/explore?date=2010-01-01%202017-02-19&q=Big%20Data

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

Interest over time



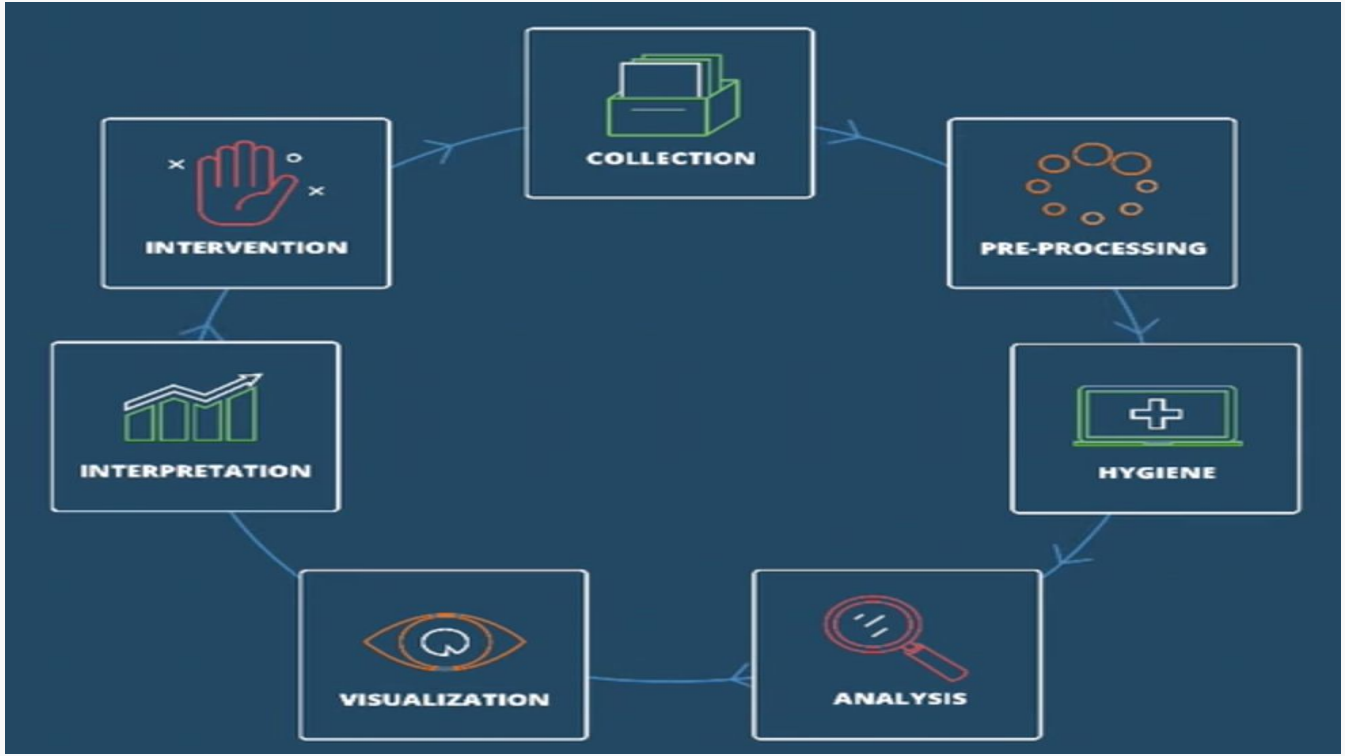
Big Data Landscape 2016 (Version 3.0)



Last Updated 3/23/2016

FIRSTMARK

Cross-Infrastructure Analytics



Who studies this stuff?

https://goo.gl/x9fxUp

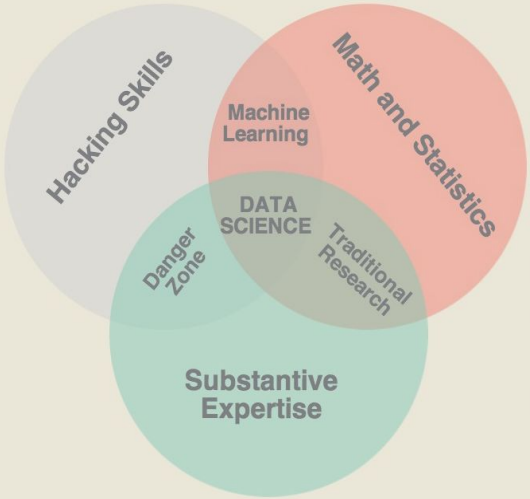
DataCamp
Learn data analysis for free, interactively

Become a

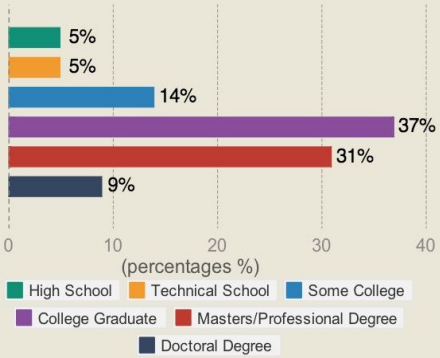
Data Scientist

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

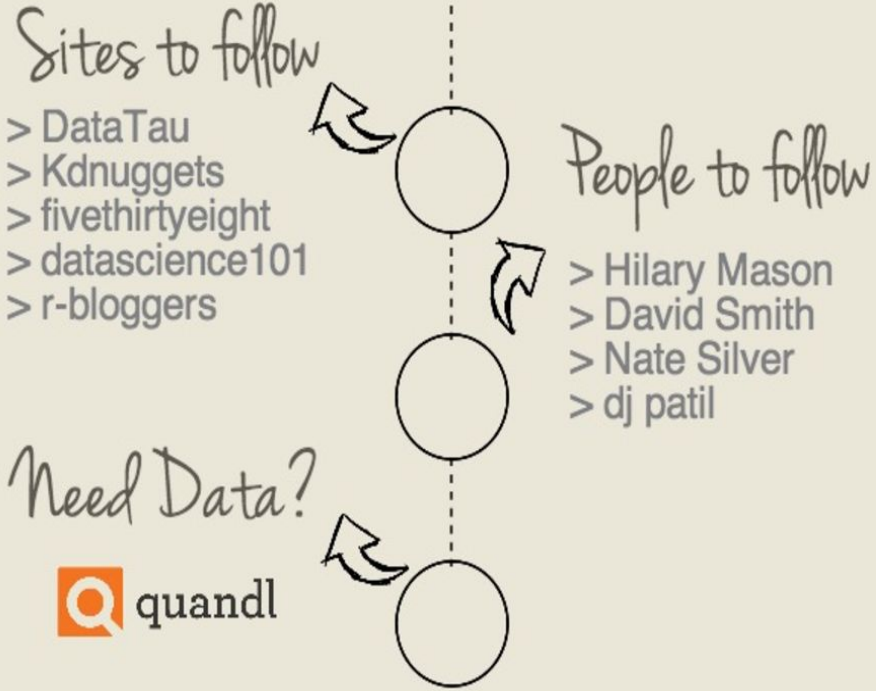
The best way to find out whether you are a true data scientist or not is to take the bull by the horns and to enter the real-life jungle of data-analysis and science with your freshly acquired skill set.

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

amazon.com

zipfian

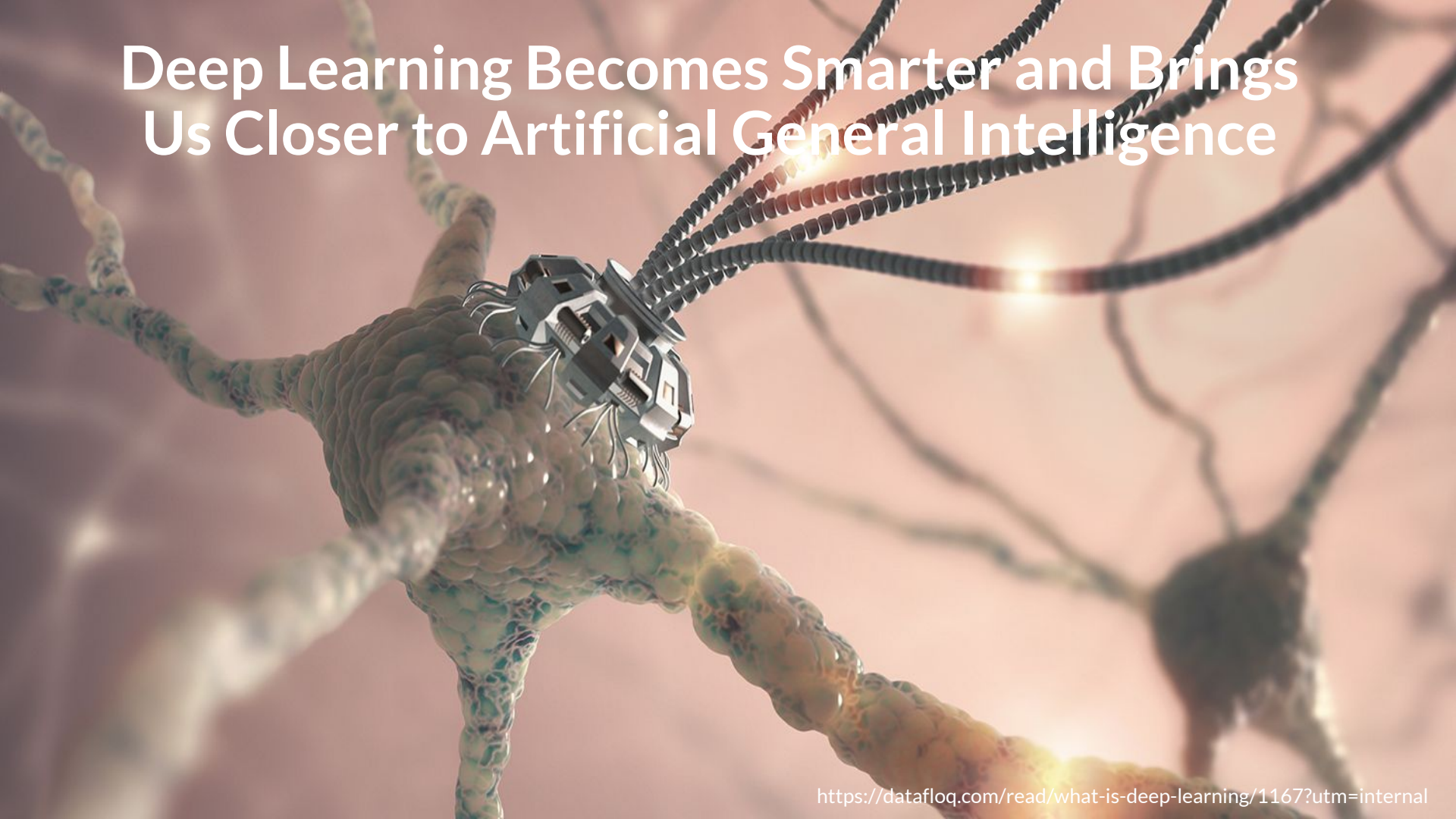
8 Follow and engage with the community



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

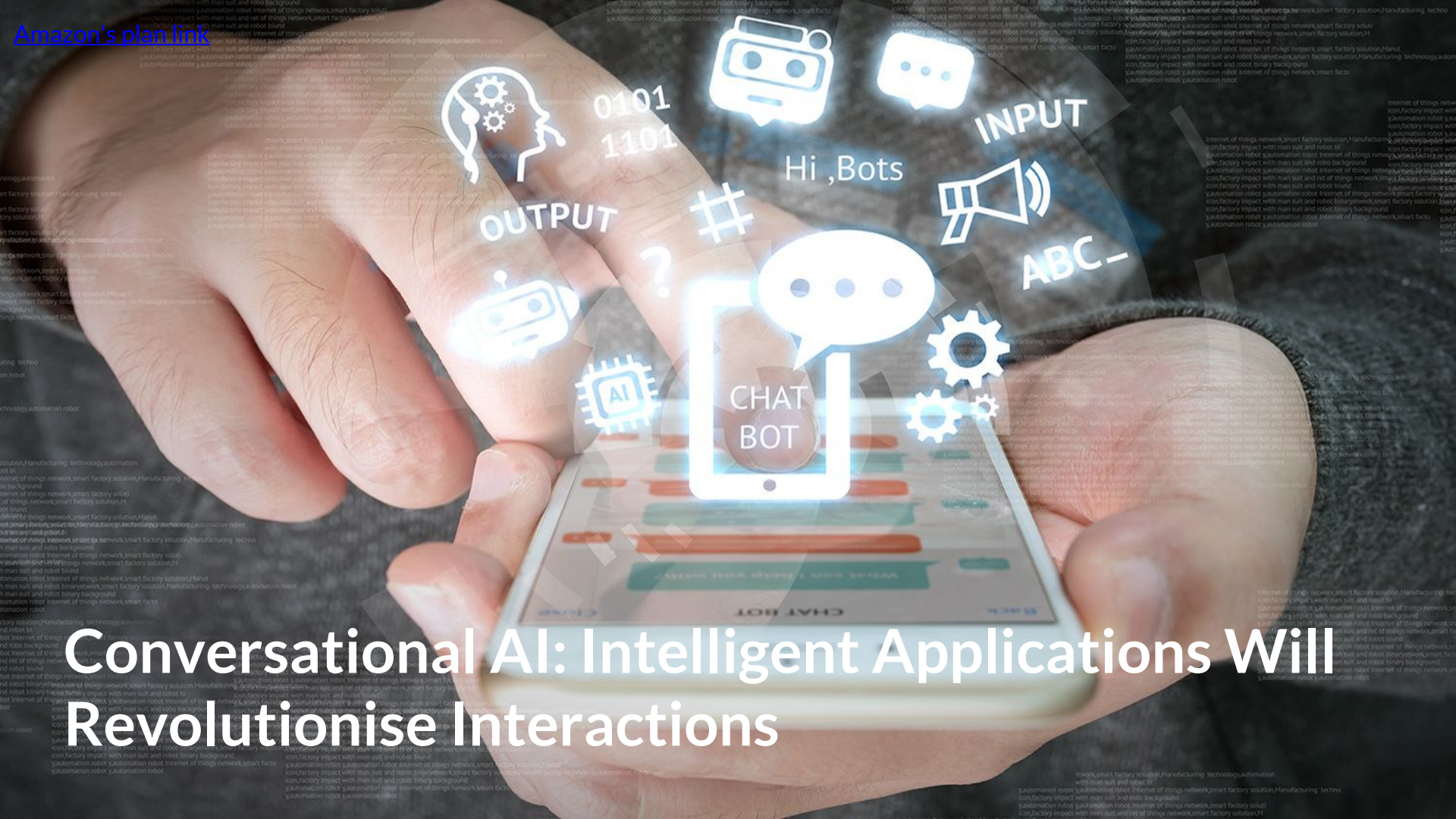
Uber's plans to data analytics

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



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

Amazon's plan link

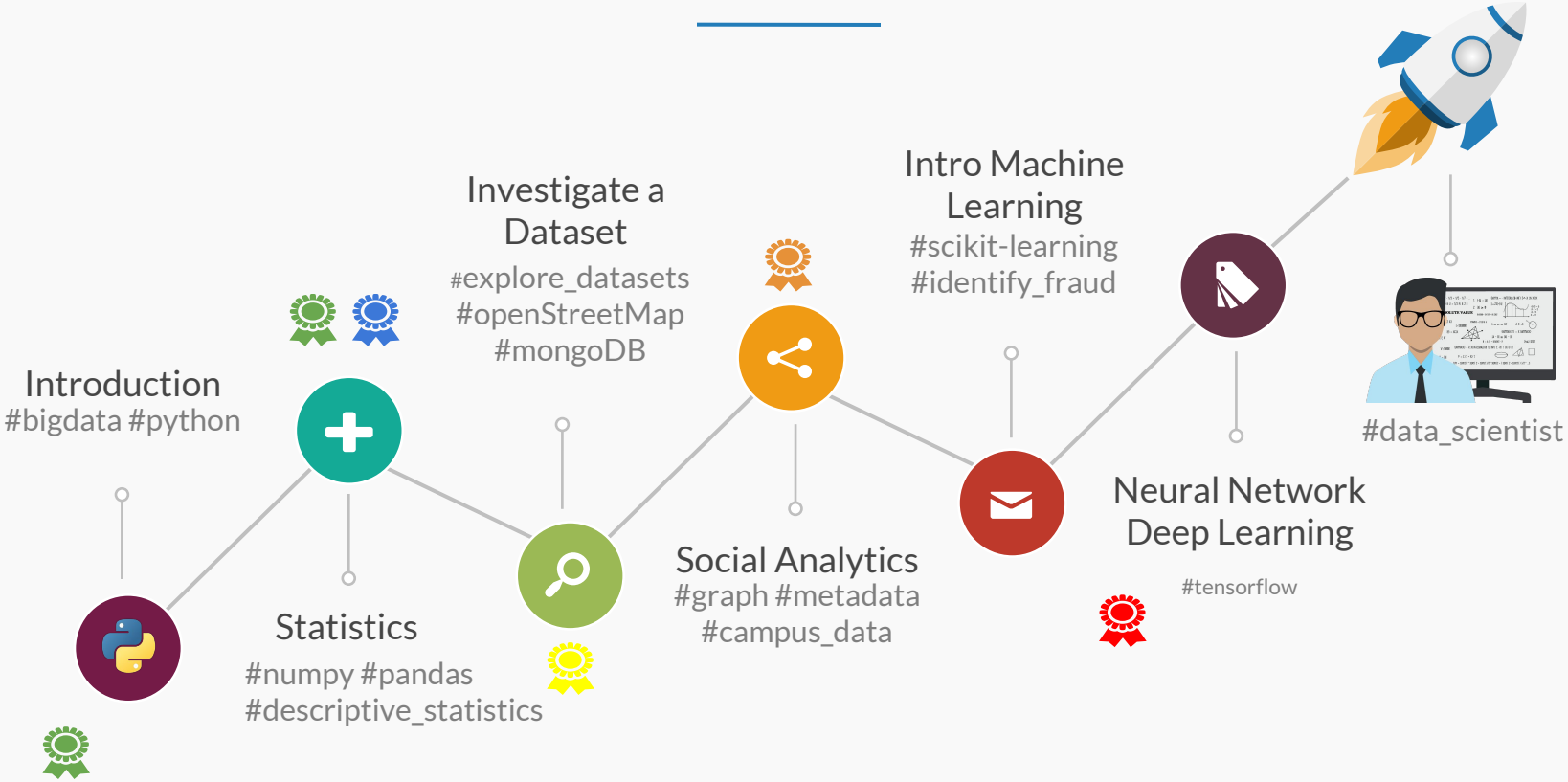


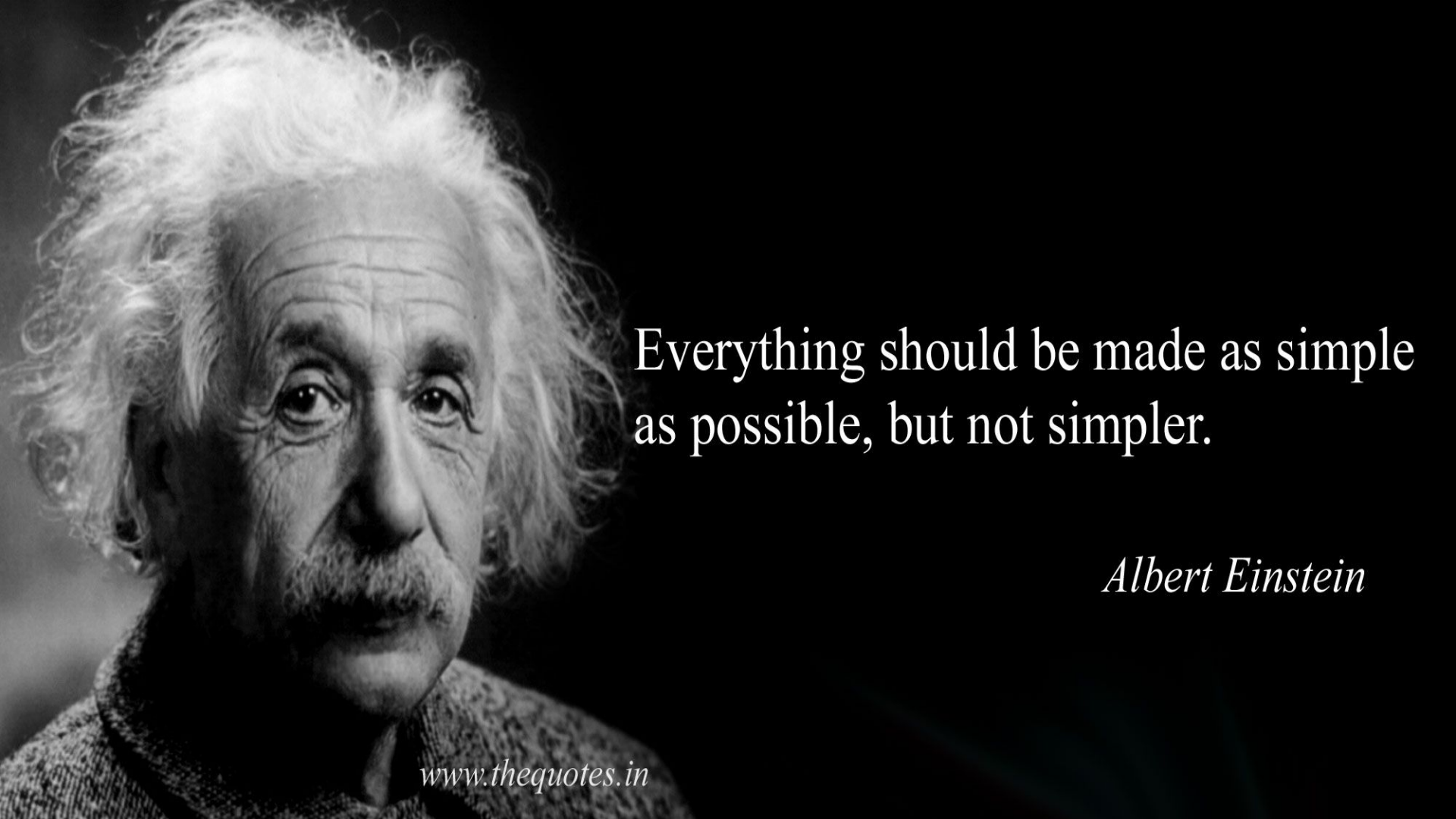
Conversational AI: Intelligent Applications Will Revolutionise Interactions

IoT-Related Data Breaches Will Cause Havoc

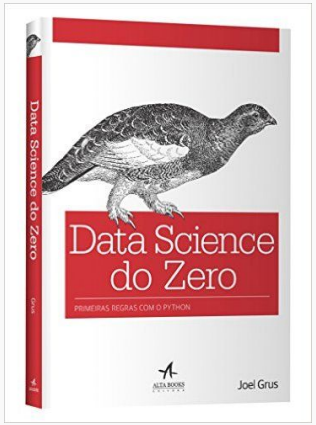
https://datafloq.com/read/ddos-attack-a-wake-up-call-for-iot/2480?utm_source=datafloq&utm_medium=ref&utm_campaign=datafloq

Agenda

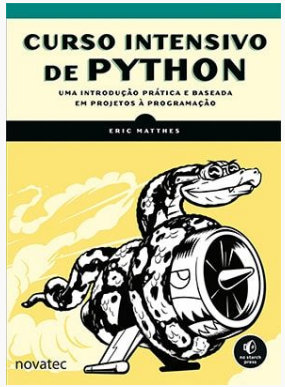

Introduction
#bigdata #python
Statistics
#numpy #pandas
#descriptive_statistics
Investigate a Dataset
#explore_datasets
#openStreetMap
#mongoDB
Social Analytics
#graph #metadata
#campus_data
Intro Machine Learning
#scikit-learning
#identify_fraud
Neural Network Deep Learning
#tensorflow
#data_scientist

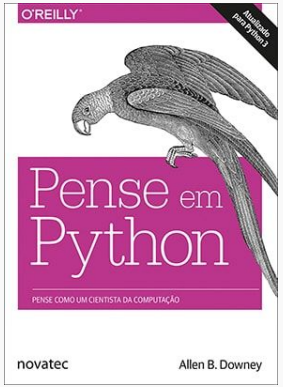

Everything should be made as simple as possible, but not simpler.
Albert Einstein
www.thequotes.in

References


Data Science do Zero
PRIMEIRAS REGRAS COM O PYTHON
Joel Grus

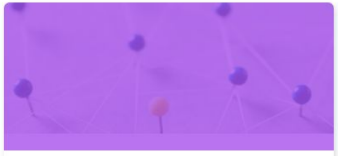

Fernando Amaral
INTRODUÇÃO À CIÊNCIA DE DADOS
MINERAÇÃO DE DADOS E BIG DATA


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


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

References

UDACITY

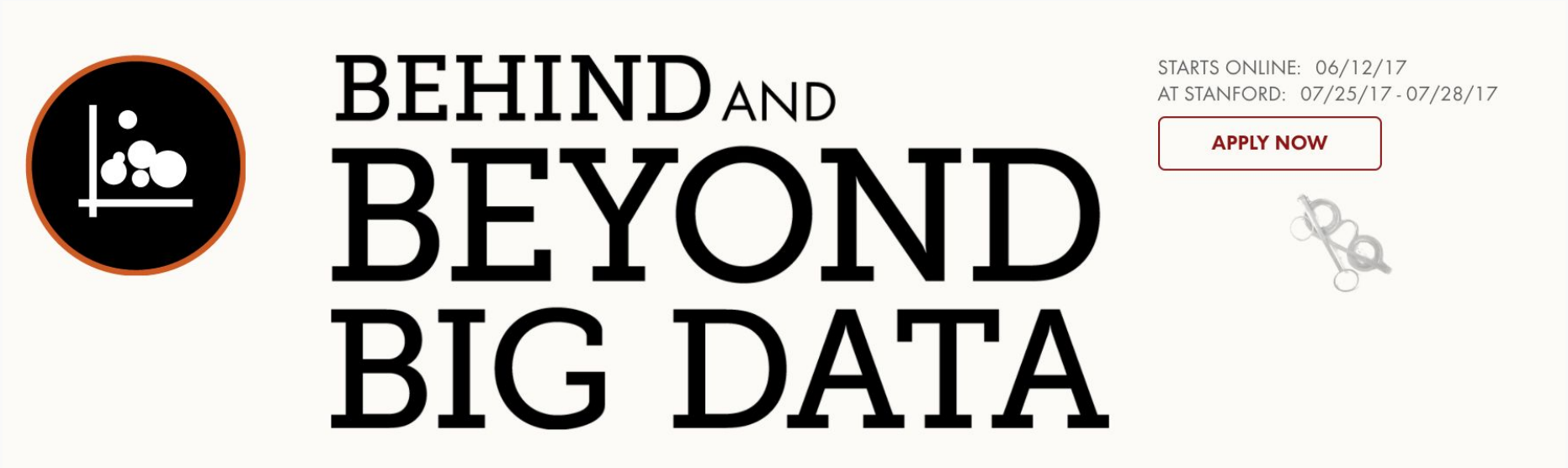


References


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

"Knowledge is no longer in the academy, in "papers" or large research centers of traditional companies; It is on the Internet, where anyone can learn what they want, when and where they want."

Reinaldo Nomad, Innovationˆ2