UFRN

Open Data Day
Natal, Brazil

01 a 03 de março de 2018

Minicourse

Machine Learning Fundamentals in Python

Speaker:
Prof. Ivanovitch Silva (ivan@imd.ufrn.br)


Introduction

Agenda

- [1h] Introduction & motivation
- [1h] Platforms & first model (KNN)
- [30min] Evaluating model (MAE, MSE, RMSE)
- [1h] Improving the model & Scikit-Learn
- [30min] Hyperparameter Optimization with grid search and cross-validation

Clone the repository

```
git clone https://github.com/ivanovitchm/opendatanatal2018.git
```

Or ....

```
git pull
```

https://git-scm.com/downloads

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

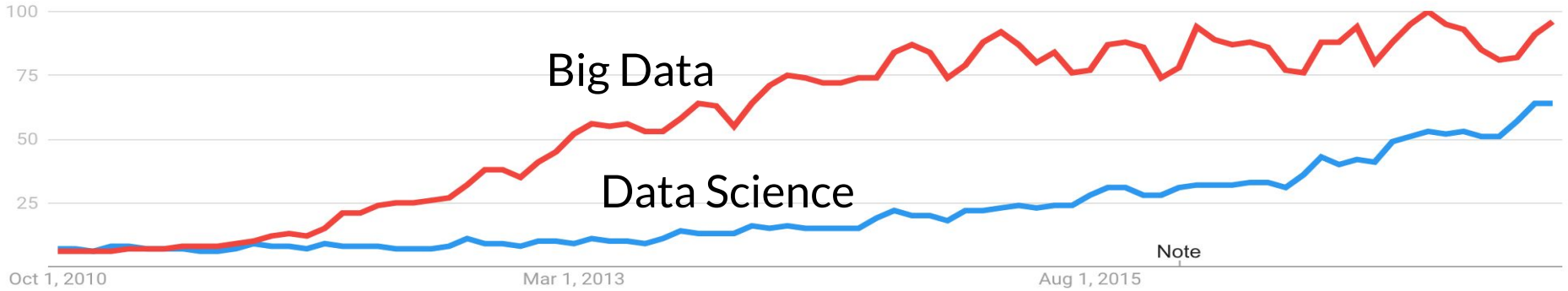
Big Data and Social Analytics certificate course

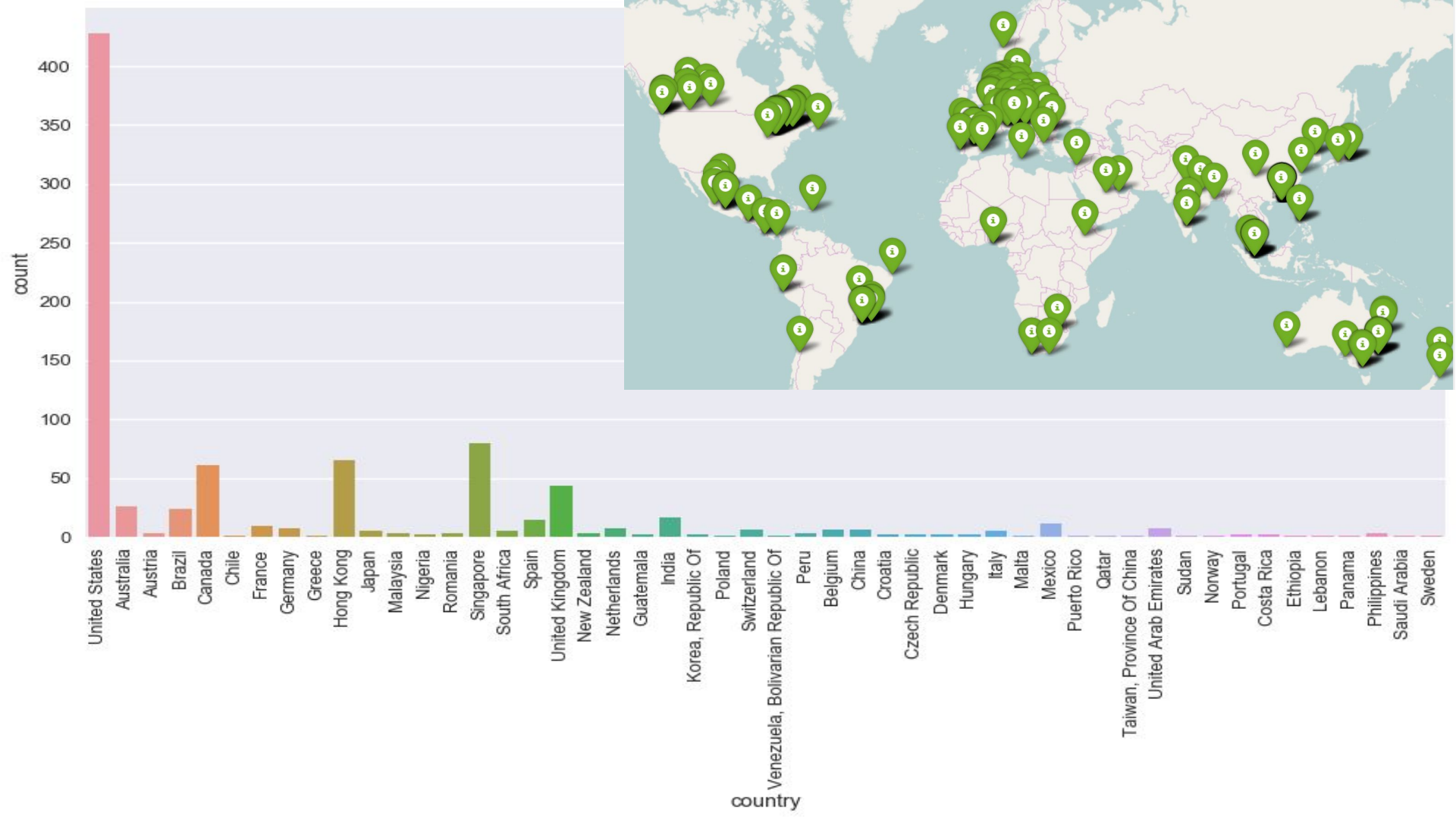
EXperimental Learning

2017 DATES TO BE CONFIRMED

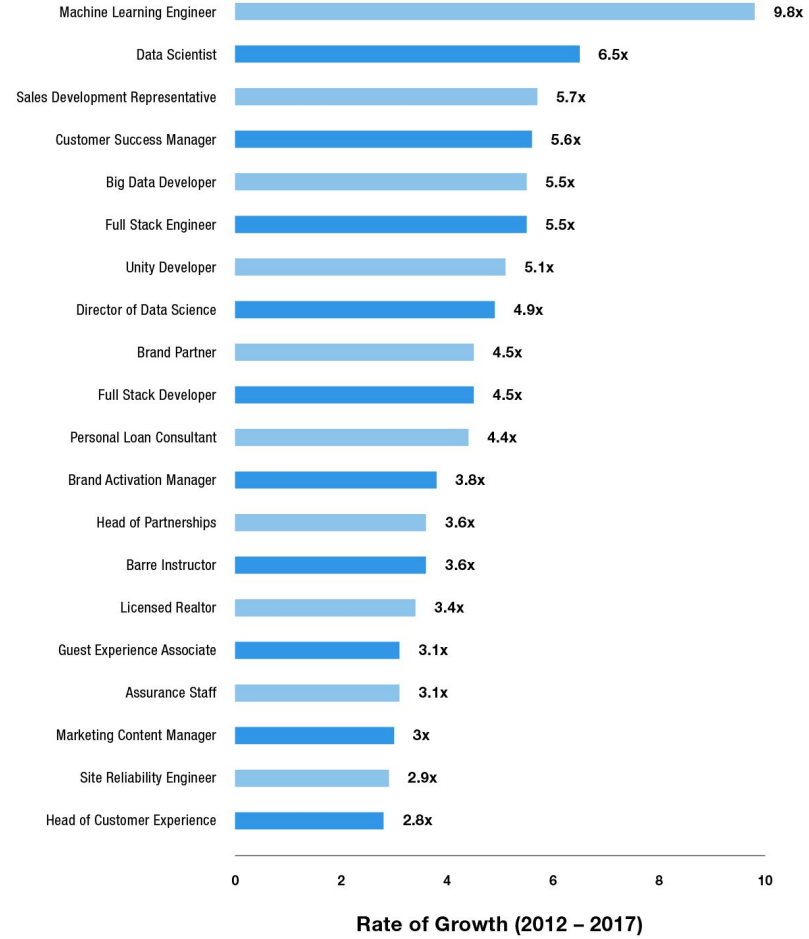
DOWNLOAD COURSE PROSPECTUS

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

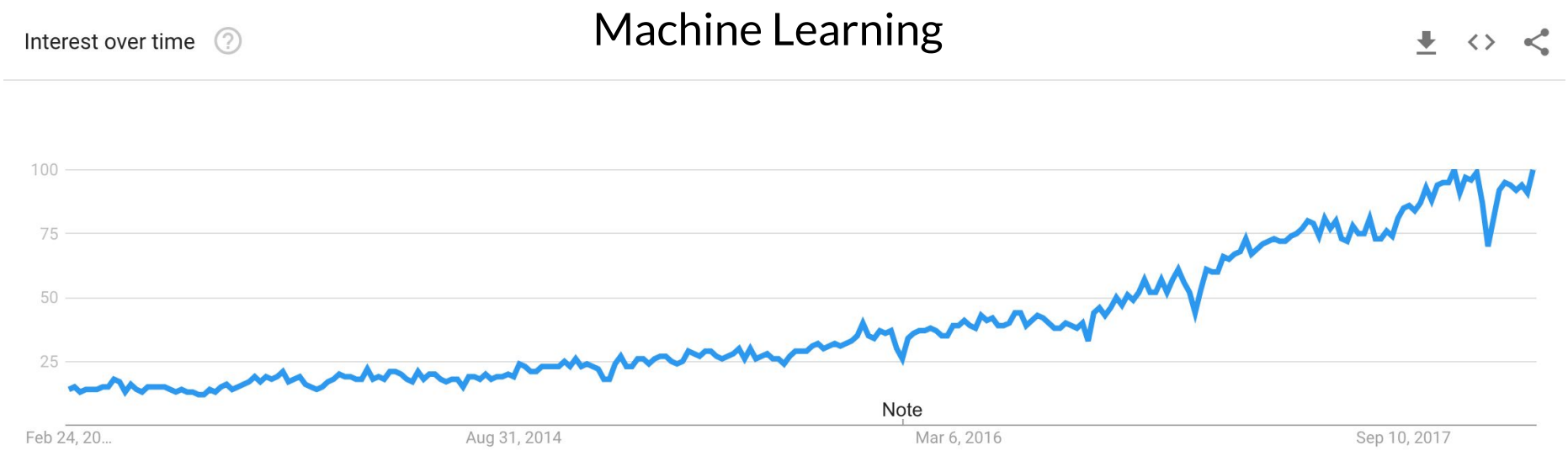
Top 20 Emerging Jobs



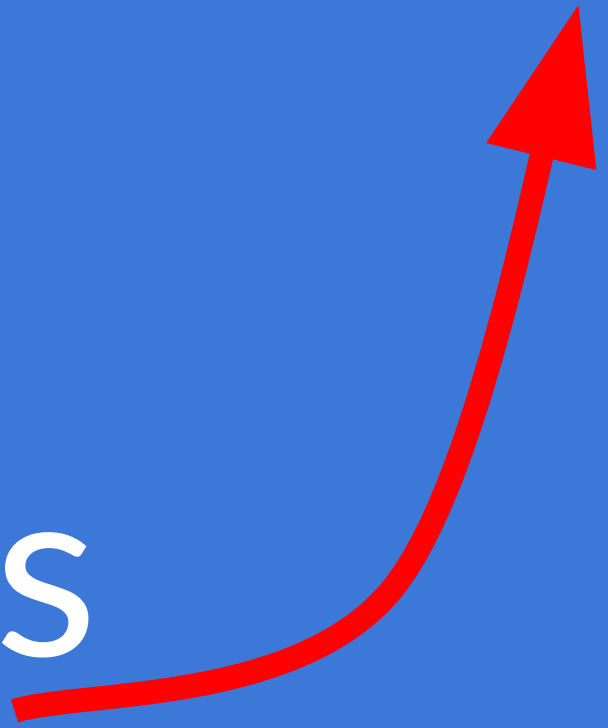
There are 9.8 times more Machine Learning Engineers working today than five years ago based on LinkedIn's research

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

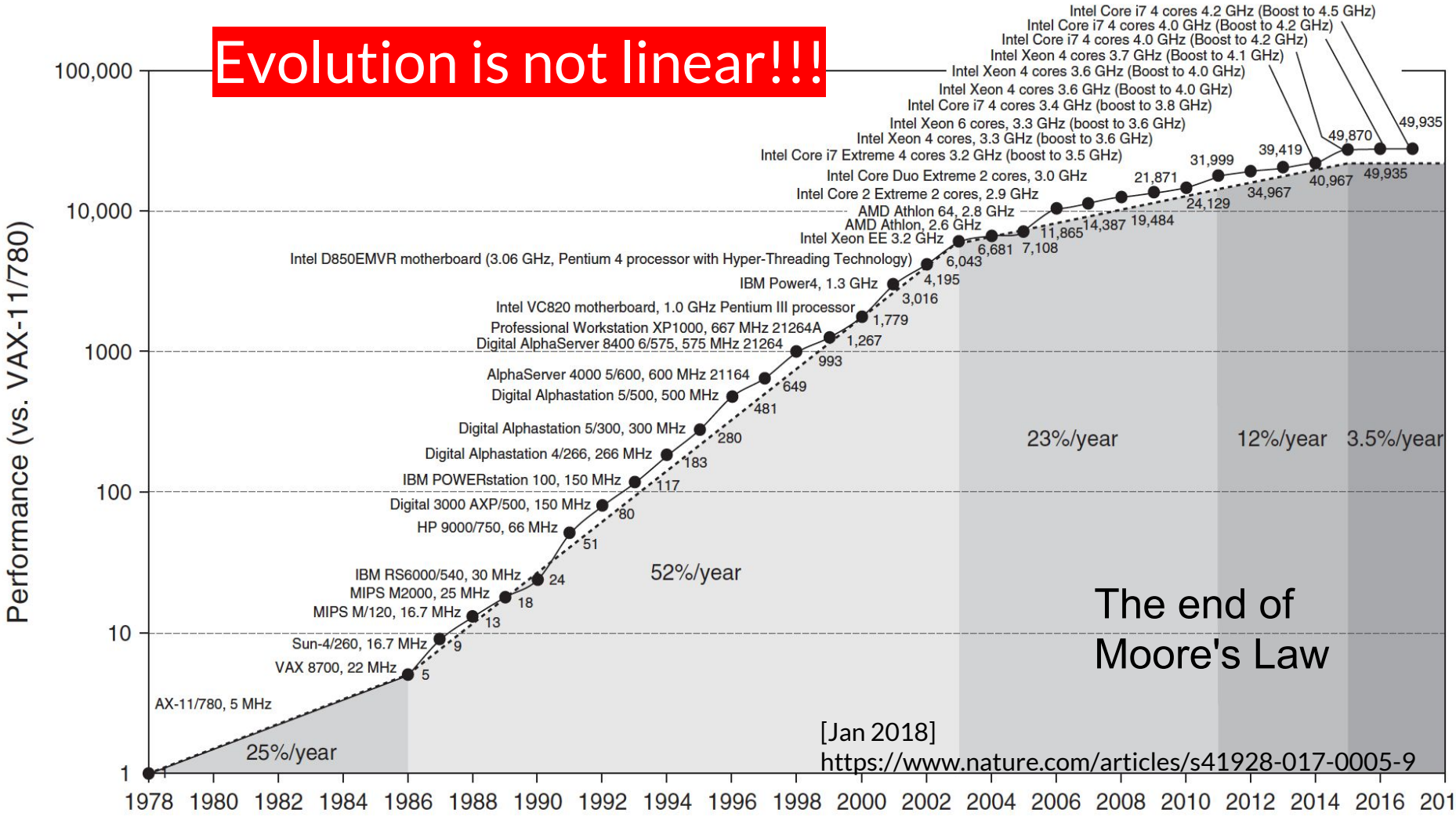
So why is Machine Learning popular now?



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.

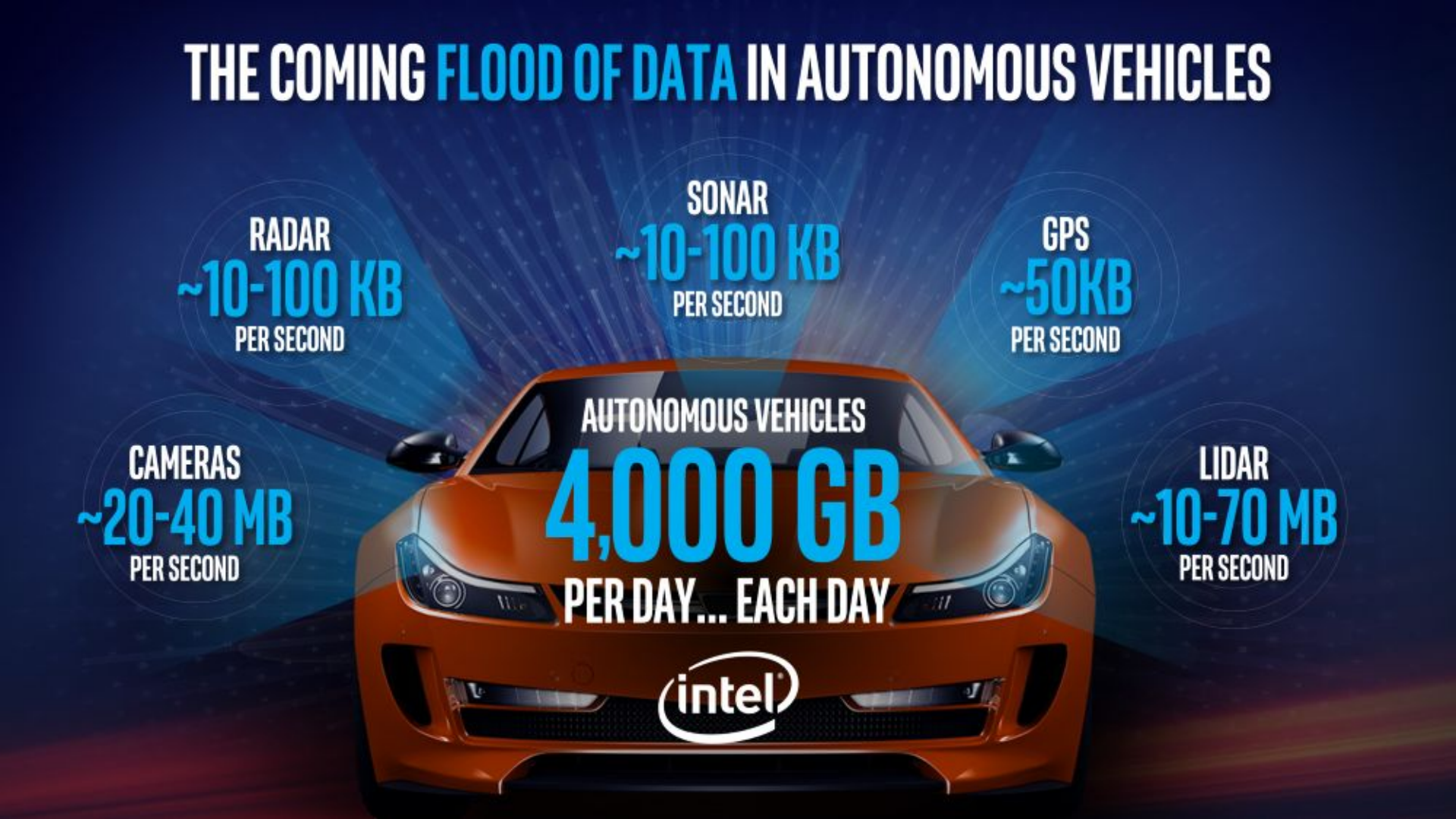
LEARN MORE

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

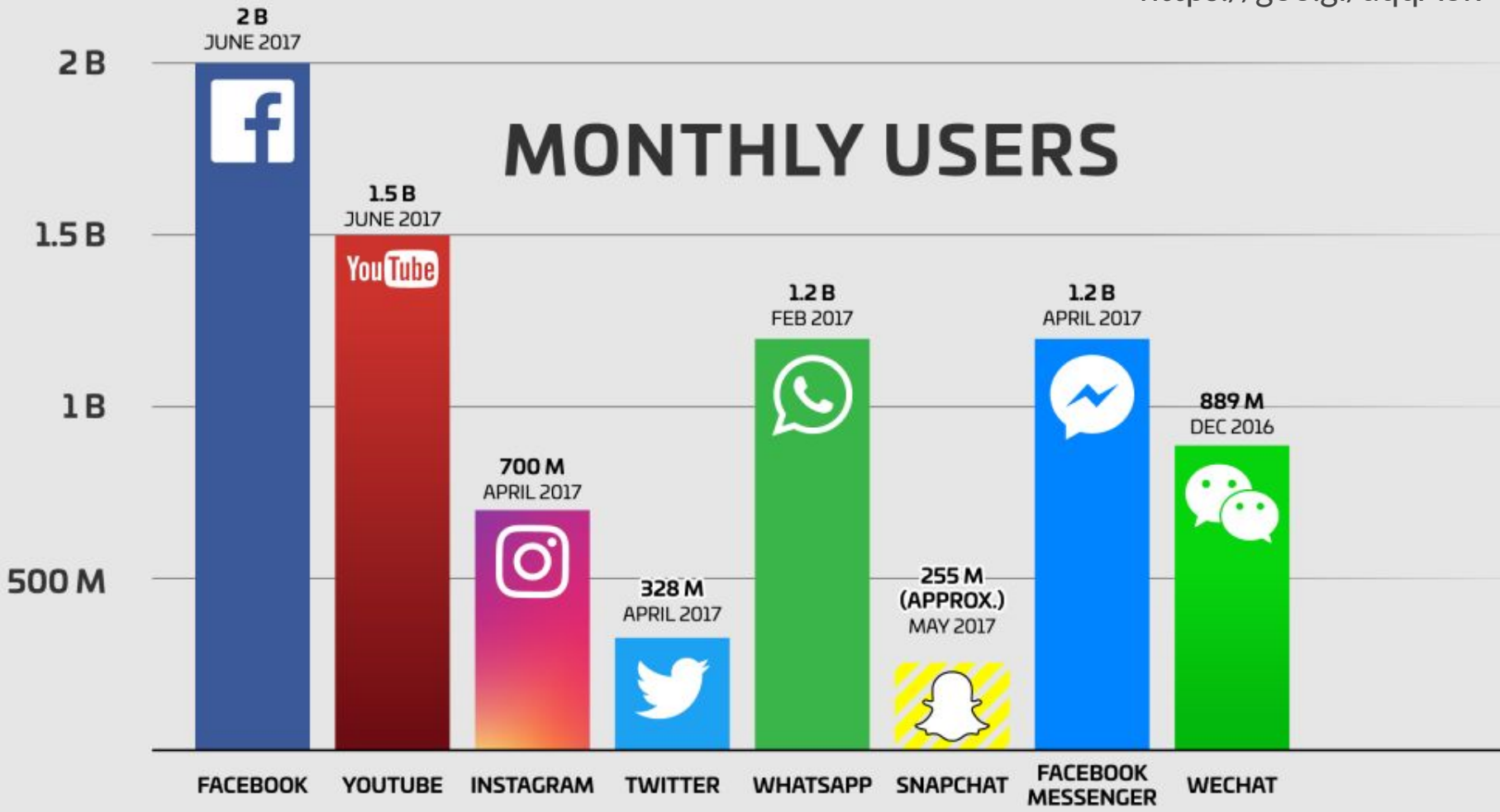
Lara Croft has changed over 21 years

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

THE COMING FLOOD OF DATA IN AUTONOMOUS VEHICLES



https://goo.gl/dqqAok



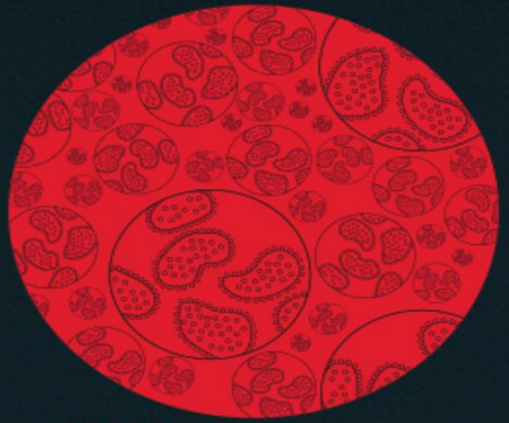
YOTTABYTES

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

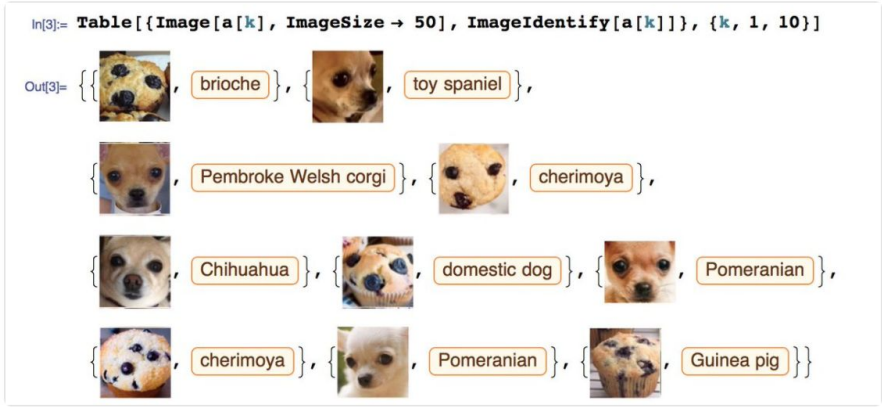
Data is giving rise to a new economy

https://goo.gl/agTRgb


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

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

https://goo.gl/NwP7Fv

https://goo.gl/xzLxaY


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


Initiatives

2016/2017 - Specialization course in Big Data

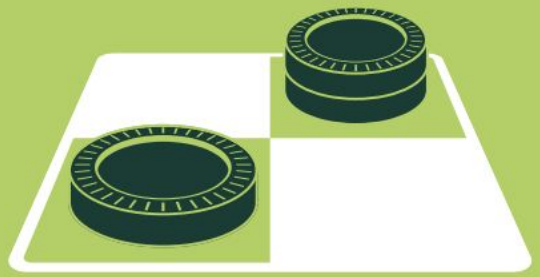
Undergratuate
2017.1 - IMD0105 Introduction to Data Science
2017.2 - IMD0252 Learning Analytics
2017.2 - DCA0046 Data Science
2018* - Internet of Things (Data Science I & II)

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

https://github.com/ivanovitchm

ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.



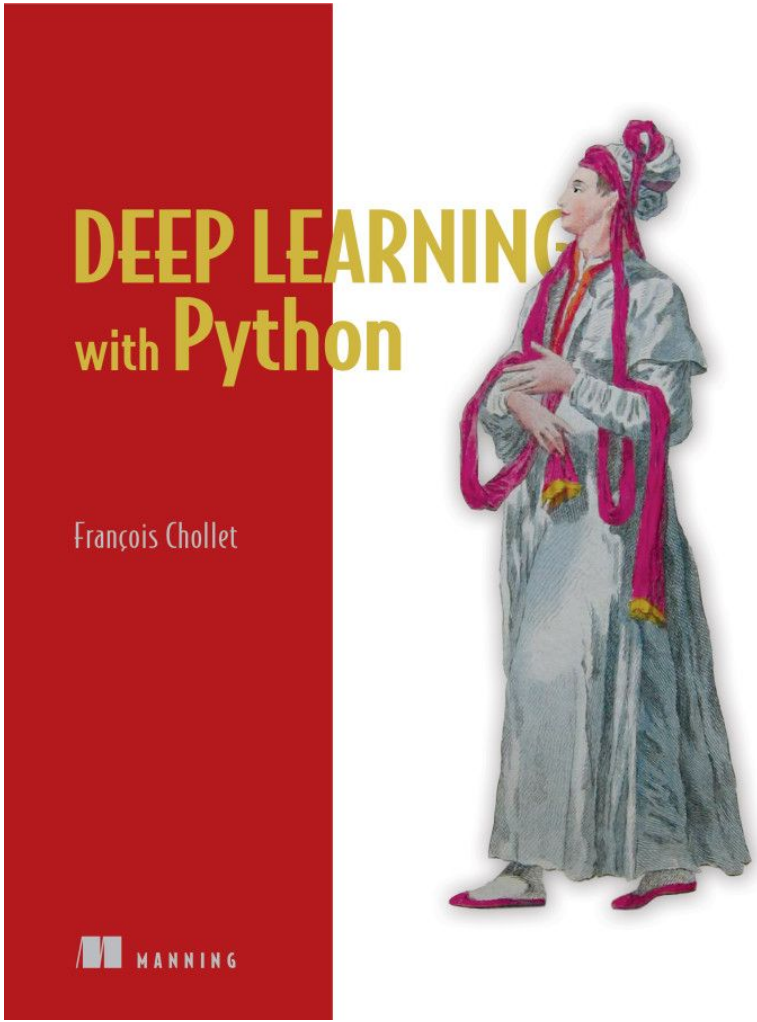
MACHINE LEARNING

Machine learning begins to flourish.

DEEP LEARNING

Deep learning breakthroughs drive AI boom.

1950's 1960's 1970's 1980's 1990's 2000's 2010's



PART 1: THE FUNDAMENTALS OF DEEP LEARNING

1. WHAT IS DEEP LEARNING? free
2. BEFORE WE START: THE MATHEMATICAL BLOCKS OF NEURAL NETWORKS free
3. GETTING STARTED WITH NEURAL NETWORKS free
4. FUNDAMENTALS OF MACHINE LEARNING

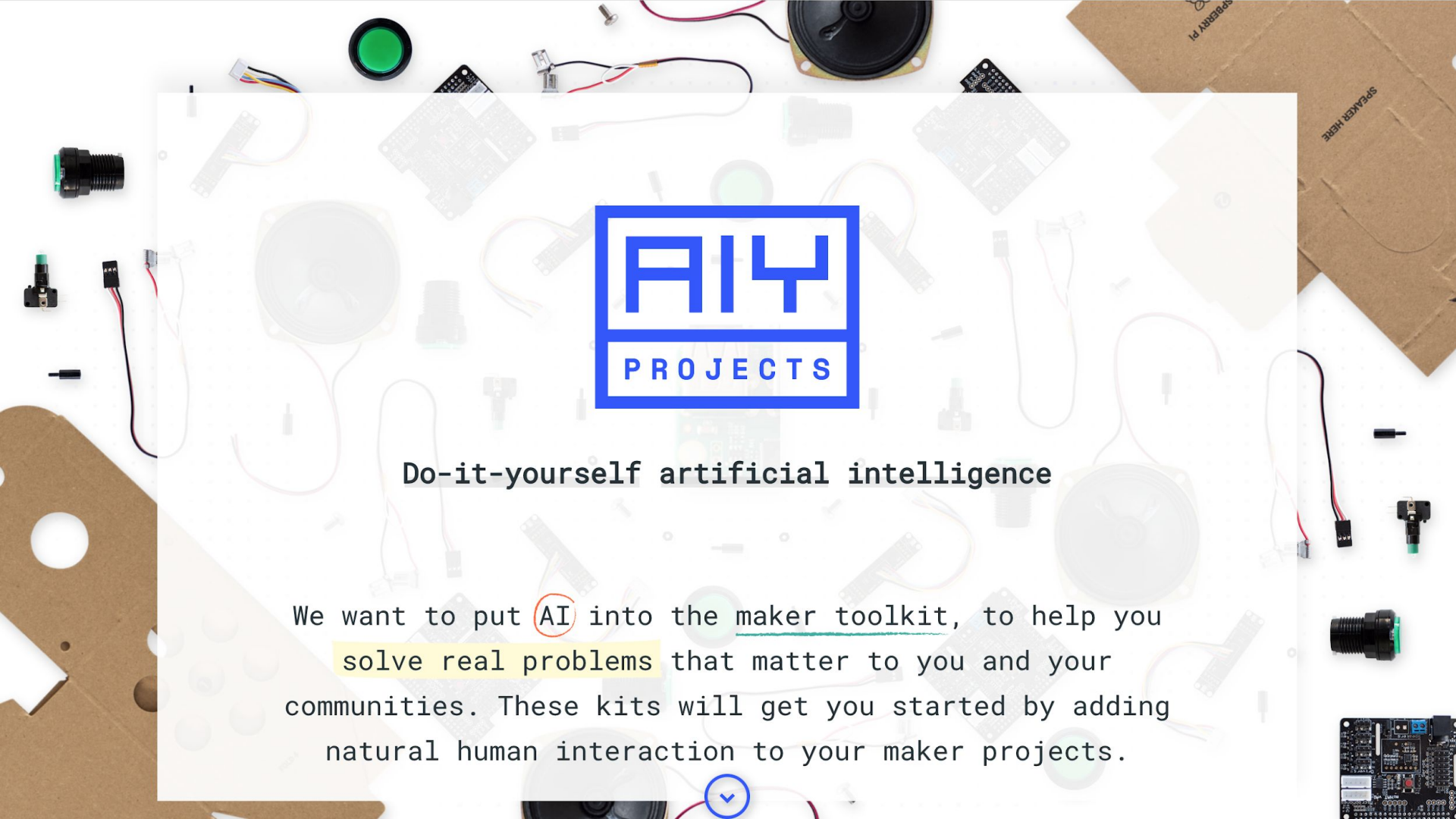
PART 2: DEEP LEARNING IN PRACTICE

5. DEEP LEARNING FOR COMPUTER VISION
6. DEEP LEARNING FOR TEXT AND SEQUENCES
7. ADVANCED DEEP LEARNING BEST PRACTICES
8. GENERATIVE DEEP LEARNING
9. CONCLUSIONS

https://www.datacamp.com/community/blog/interview-francois-chollet

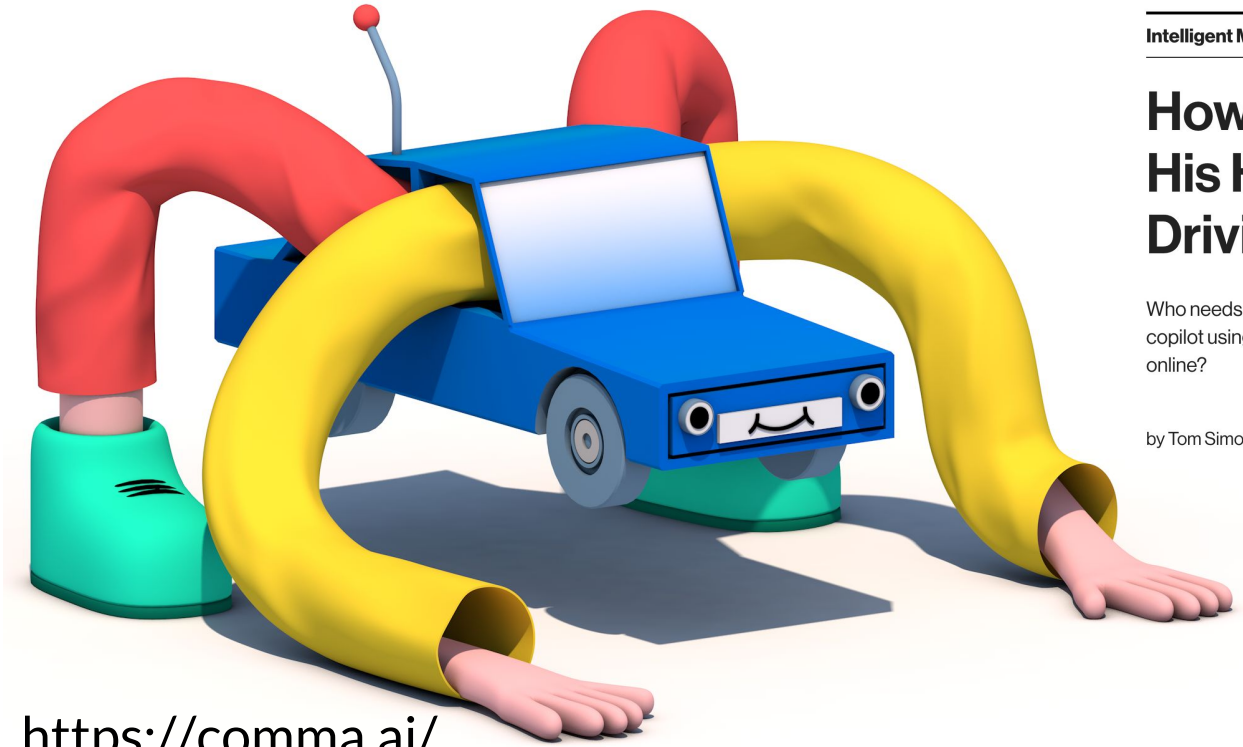

i am ai


Recognized As:
Faucet Running
(LAPTOP FOR ILLUSTRATION)



Do-it-yourself artificial intelligence

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



https://comma.ai/

Intelligent Machines

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

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

by Tom Simonite February 21, 2017

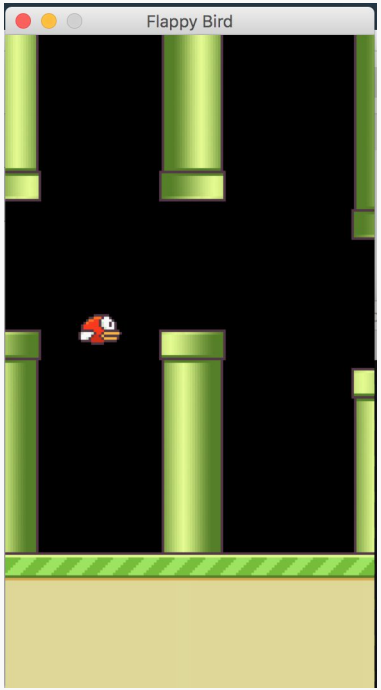
Uma nova forma de investir.

Flappy Bird

https://github.com/yenchenlin/DeepLearningFlappyBird

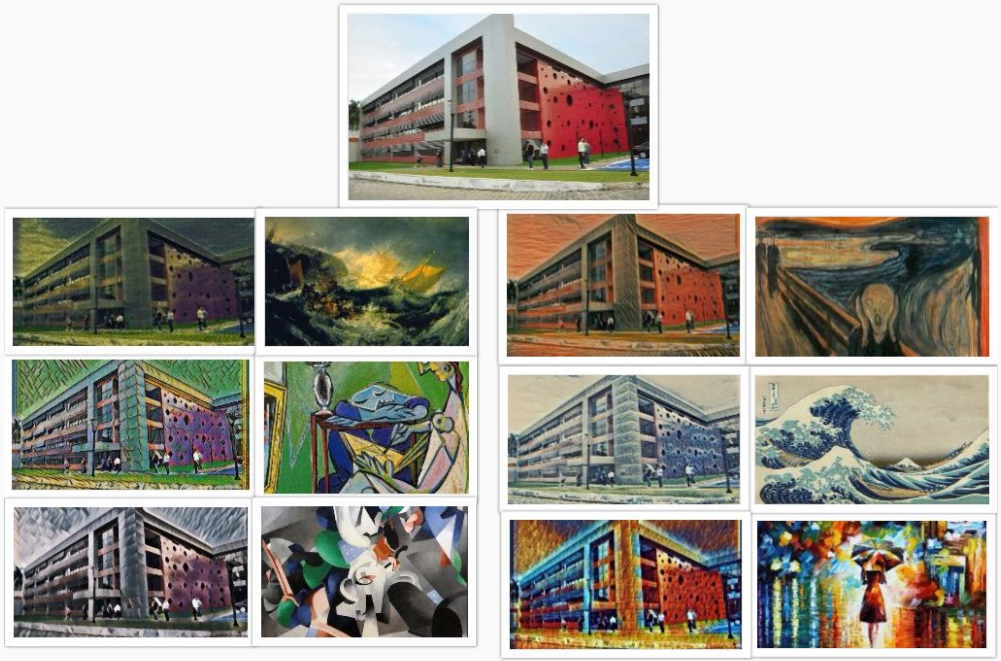
The following agent is able to play without being told any information about the structure of the game or its rules. It automatically discovers the rules of the game by finding out how it did on each iteration.

Flappy Bird



1. conda install -c menpo opencv3
2. pip install pygame
3. pip install tensorflow
4. git clone https://github.com/yenchenlin/DeepLearningFlappyBird.git
5. cd DeepLearningFlappyBird
6. python deep_q_network.py

Style Transfer



Style transfer allows you to take famous paintings, and recreate your own images in their styles!

Style Transfer

Project (Logan Engstrom/MIT): [fast-style-transfer GitHub repo]

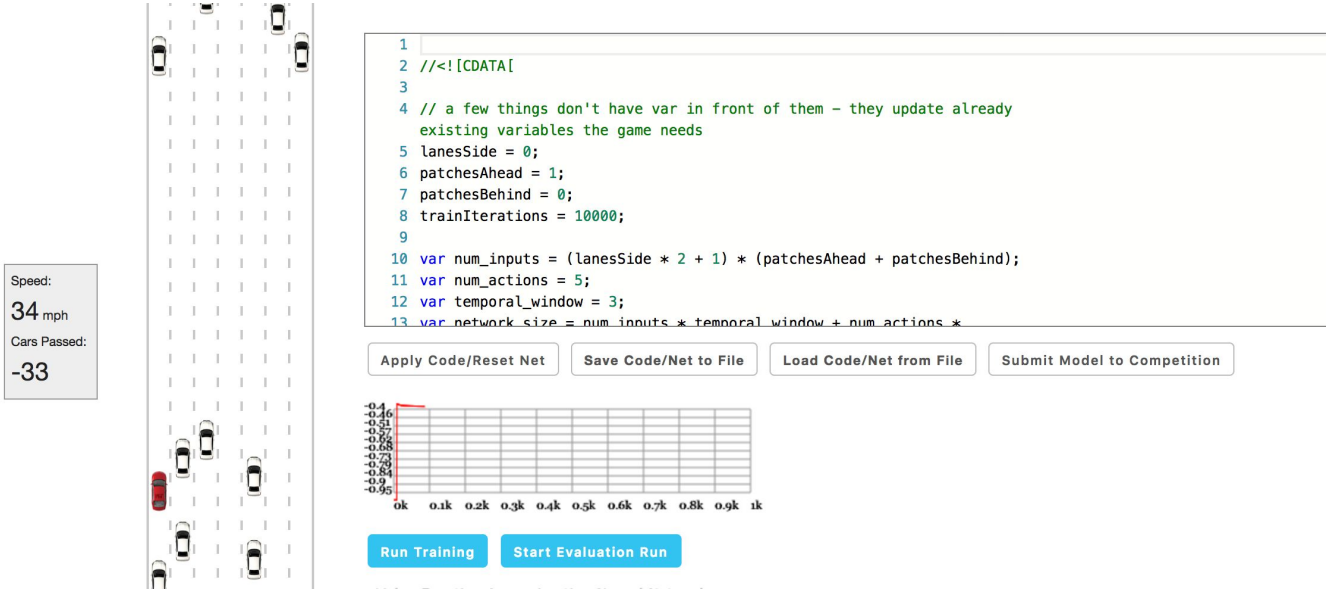
```
git clone https://github.com/lengstrom/fast-style-transfer
```

Notebook

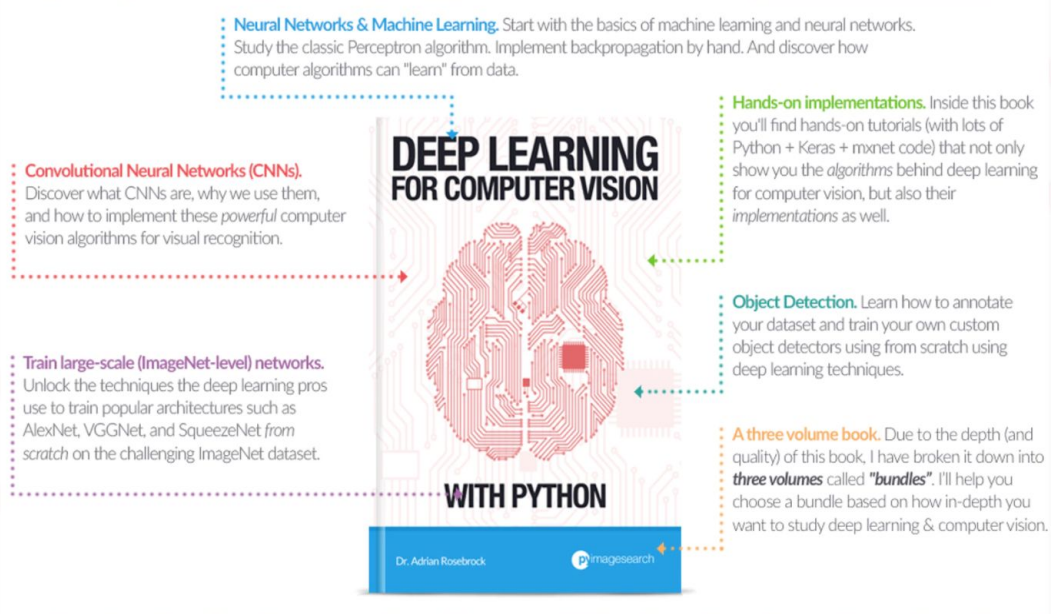
```
!pip install tensorflow
!pip install scipy
!pip install pillow
```

Deep Traffic

http://selfdrivingcars.mit.edu/deeptrafficjs/



Deep Learning for Computer Vision with Python [eBook]



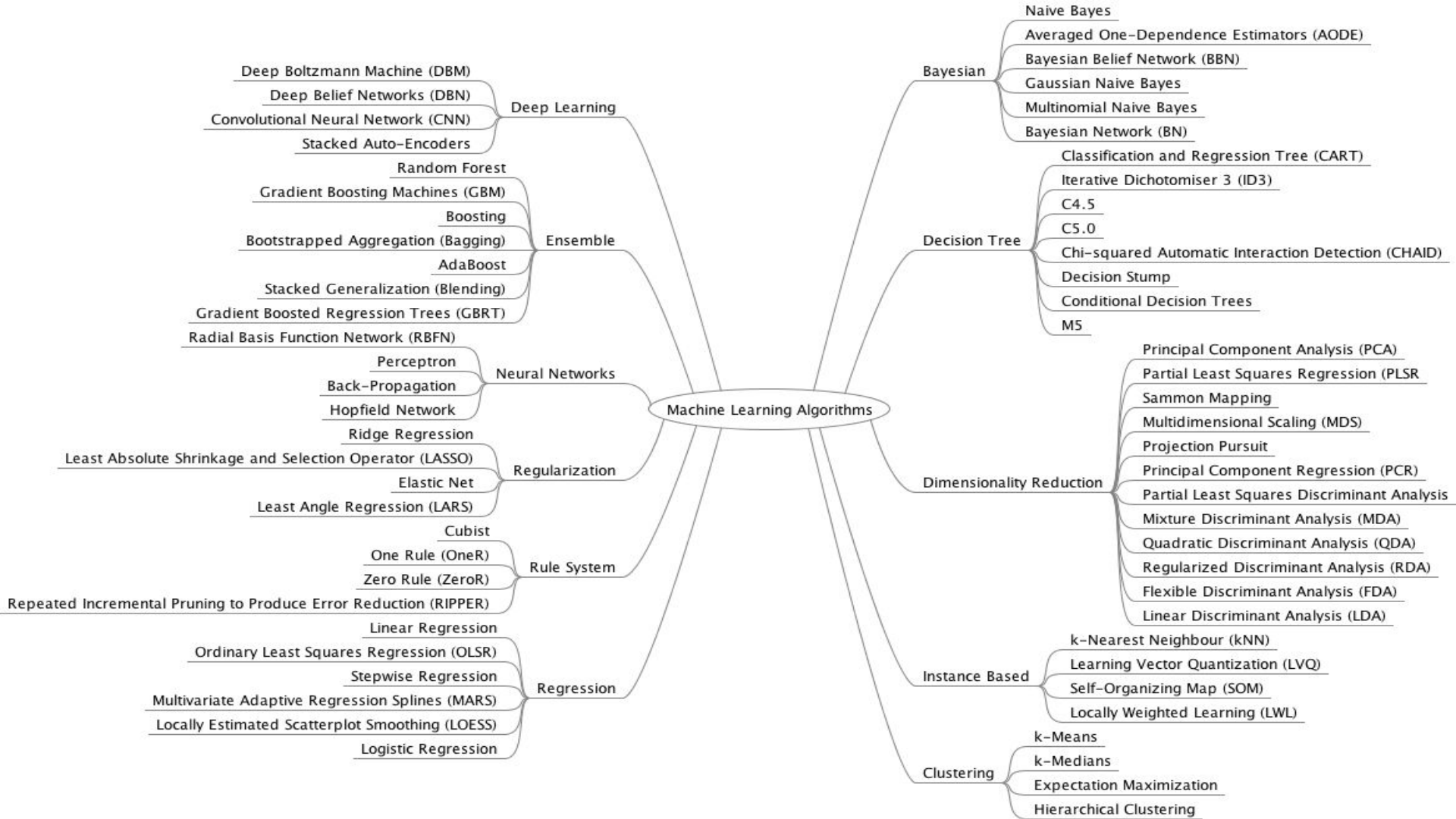
Struggling to get started with neural networks & deep learning for computer vision? My new book will teach you all you need to know. Use the link below to pre-order your copy:

Pre-order Now

Created by

Adrian Rosebrock

1,014 backers pledged $262,792 to help bring this project to life.


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

Kinds of ML Algorithms

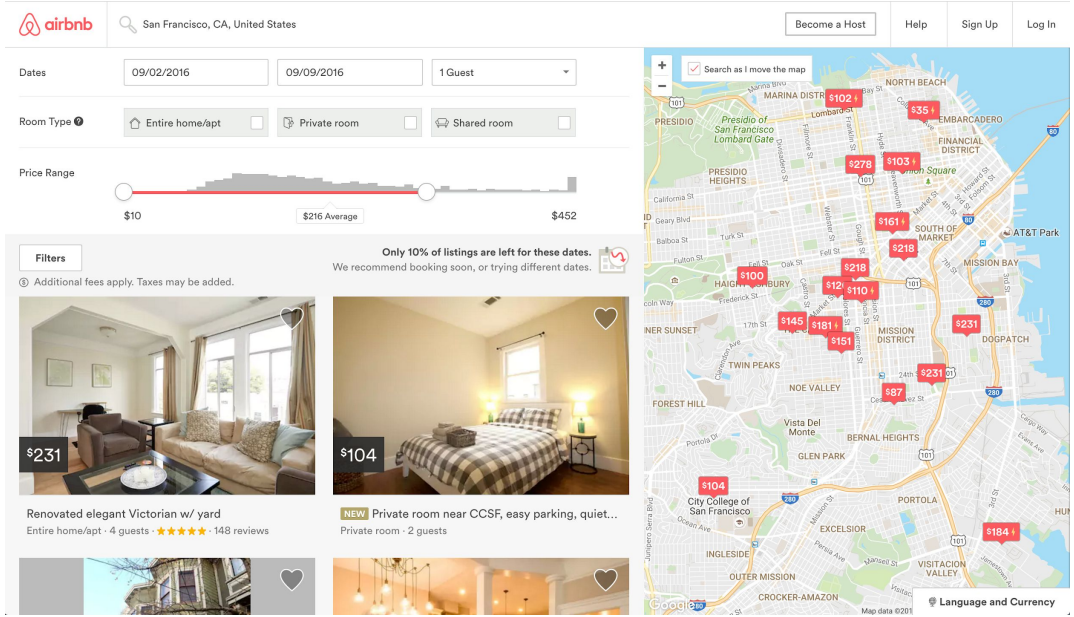
	Algorithms
Unsupervised Learning	*k*-Means Clustering Principal Component Analysis Association Rules Social Network Analysis
Supervised Learning	Regression Analysis *k*-Nearest Neighbors Support Vector Machine Decision Tree Random Forests Neural Networks
Reinforcement Learning	Multi-Armed Bandits

Problem definition

DATAQUEST

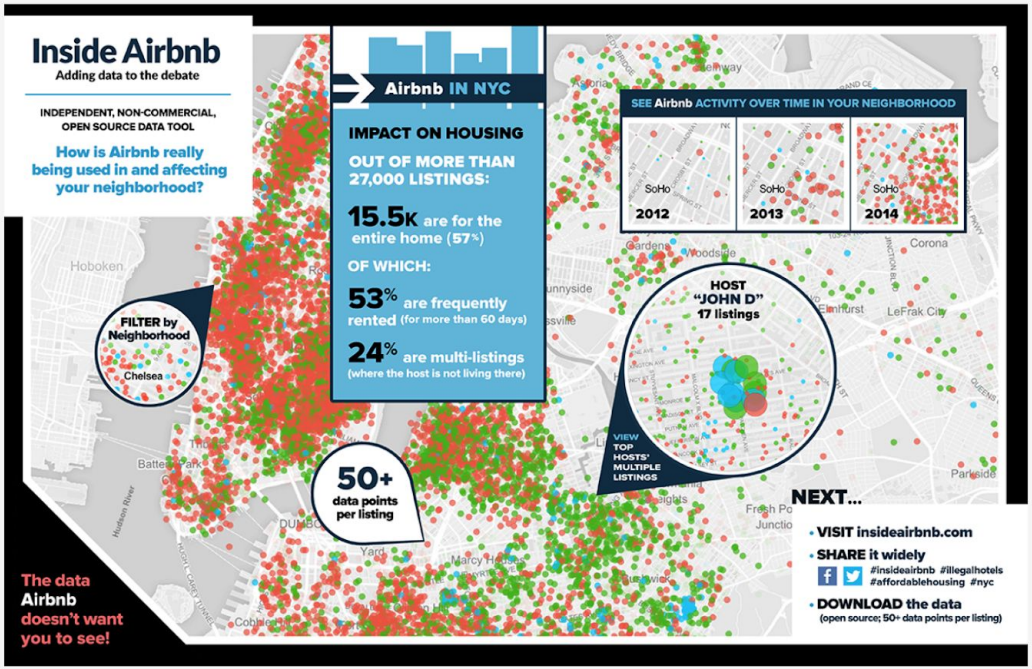
http://bit.do/dataquestblog

One challenge that hosts looking to rent their living space face is <u>determining the optimal nightly rent price</u>



https://github.com/tomslee/airbnb-data-collection

About Inside Airbnb



October 3, 2015 on the listings from Washington, D.C

References


UDACITY


coursera


DataCamp


kaggle


DATAQUEST


Data Science
Academy

Who to follow?

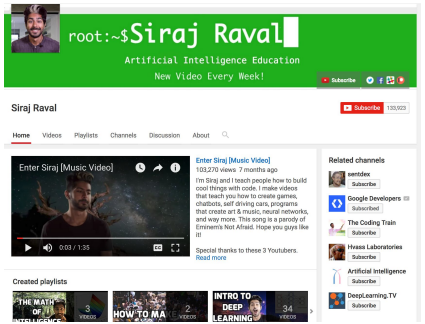
Geoffrey Hinton

Google Fellow

"After many years working in academia, it's incredibly exhilarating to see the Brain team transforming Google by combining curiosity-driven research on neural networks with world class engineering."

https://research.google.com/teams/brain/

http://www.andrewng.org/



https://www.nvidia.com/en-us/deep-learning-ai/

http://mariofilho.com/

END OF PART #1