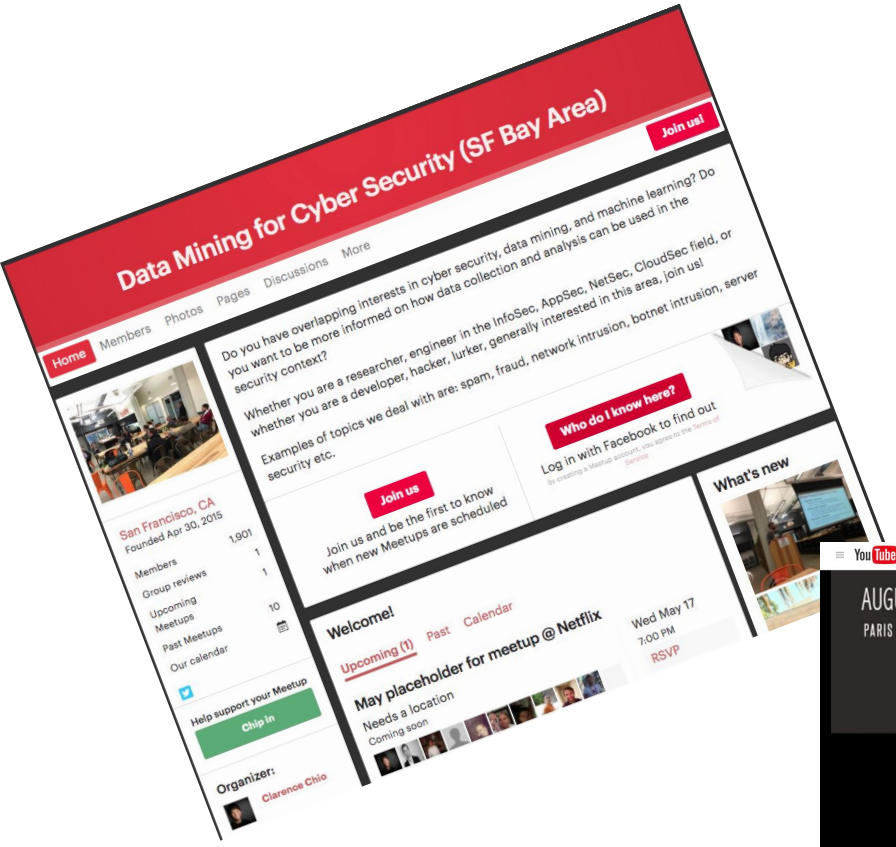
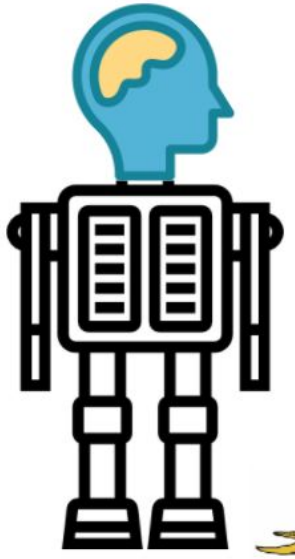
Defcon Workshop

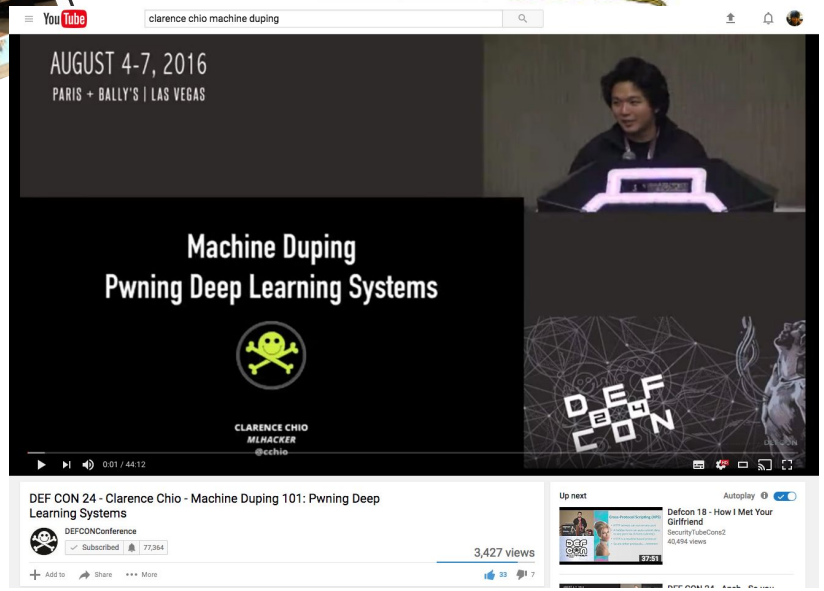


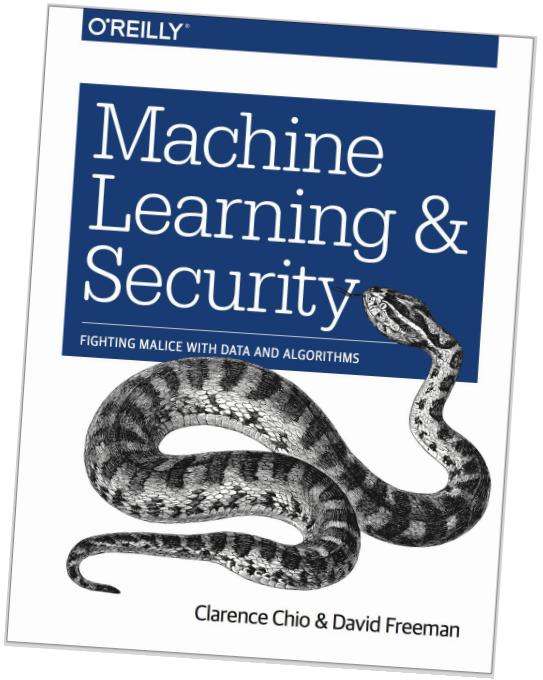


Making & Breaking Machine Learning Systems

(for infosec)

clarence chio (@cchio)





https://www.youtube.com/watch?v=JAGDpJFFM2A

https://www.meetup.com/Data-Mining-for-Cyber-Security/

who am i ?

PWNING MACHINE LEARNING SYSTEMS

Saturday, 14:30 to 18:30 in Octavius 6

Clarence Chio Security Researcher
Anto Joseph Security Engineer, Intel

Practical machine learning in infosecurity

Want to visit +7

Authors: Anto Joseph and Clarence Chio

Machine learning (ML) is the future. The speaker will give an introduction to the topic with the Boolean classification problem and introduce classifiers, which are at the core of many of the most common ML systems. He will also provide a simple example of deploying security machine learning systems in production pipelines using Apache Spark. The speaker will talk about how such systems can be poisoned, misguided, and utterly broken if the architects and implementers are not careful.

Making & Breaking Machine Learning Systems

From March 20, 2017 to March 21, 2017

Clarence Chio, Anto Joseph

Making & Breaking Machine Learning Systems is a fast paced session on machine learning from the Infosec professional's point of view. The class is designed with the goal of providing students with a handson introduction to machine learning concepts and systems, as well as making and breaking security applications powered by machine learning. The lab session is designed with security usecases in mind, since using machine learning in security is very different from using it in other situations. Students will get first hand experience at cleaning data, implementing machine learn... forming penetration tests of these systems. Each attendee will be provided with a comprehensive virtual ... as well as any future machine learning experimentation and ... programming

HITB Lab: Practical Machine Learning in InfoSecurity

This lab session is designed to give attendees a quick introduction to ML concepts and gets up and running with the popular machine learning library, sci-kit learn.

We first start by building a basic understanding of how to integrate ML into an email spam identification system. We look at the inner workings and discuss the components involved in the system. Using the training data, we train our system to identify genuine messages and the system automatically learns from these examples. Different classifiers are tuned to get the maximum efficiency we can crunch out from this setup.

LOCATION: **Track 3 / HITB Labs**
DATE: **April 14, 2017**
TIME: **10:45 am - 12:45 pm**


CLARENCE CHIO


ANTO JOSEPH

NULLCON

INTERNATIONAL SECURITY CONFERENCE GOA 2017



ANTO JOSEPH
SENIOR SECURITY ENGINEER AT INTEL
ADVERSARIAL MACHINE LEARNING

CONFERENCE SPEAKERS WORKSHOPS AND VILLAGES 28TH FEBRUARY - 4TH MA

www.nullcon.com

MACHINE LEARNING FOR PEN-TESTERS AND ... RESEARCHERS

ANTO JOSEPH & CLARENCE CHIO

SOLD OUT

Let f be a function representing this training session. Let x be the amount of math involved.

$$\lim_{x \to 0} f(x)$$

INTRODUCTION

“gives computers the ability to learn without being explicitly programmed ”

“ ML currently represents the most promising path to strong AI”



BASIC TOOLS

- Scikit Learn - Python library that implements a range of machine **learning algos and helper functions**
- TensorFlow - library for numerical computation using data flow graphs . Widely used for deep learning

common data-science PACKAGES



NumPy

Base N-dimensional array package



SciPy library

Fundamental library for scientific computing



Matplotlib

Comprehensive 2D Plotting

IP[y]:
IPython

IPython

Enhanced Interactive Console



Sympy

Symbolic mathematics



pandas

Data structures & analysis

SCIKIT-LEARN



- easy-to-use, general-purpose toolbox for machine learning in Python.
- supervised and unsupervised machine learning techniques.
- Utilities for common tasks such as model selection, feature extraction, and feature selection
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

TENSORFLOW



- Open source
- By Google
- used for both research and production
- Used widely for Deep learning
- Multiple GPU Support

Classification



supervised learning

unsupervised learning

SUPERVISED MACHINE LEARNING

- learn from labeled training data
  - Regression
    - Regression is used to predict **continuous values – Linear Regression**
  - Classification
    - Classification is used to predict which class a data point is part of (**discrete value**).
    - SVM
    - Decision Trees

EXAMPLE PROBLEMS

- Example: I have a house with W rooms, X bathrooms, Y square-footage and Z lot-size. Based on other houses in the area that have been recently sold, how much can I sell my house for? ---- I would use *regression* for this kind of problem.

- Example: I have an unknown fruit that is yellow in color, 5.5 inches long, diameter of an inch, and density of X. What fruit is this? --- I would use *classification* for this kind of problem to classify it as a banana (as opposed to an apple or orange).

- Source : Quora

UNSUPERVISED MACHINE LEARNING

- find patterns or structure in the data
  - Clustering - K-means
  - dimensionality reduction – PCA , kPCA

EXAMPLE PROBLEM

- Clustering :
  - suppose you had a basket full of fresh fruits, your task is to arrange the same type fruits at one place.



BASIC TERMS

- Training data
  - The data set that you train your machine learning algorithm with
- Classifier
  - "An algorithm that implements classification
  - may also refers to the mathematical function, implemented by a classification algorithm, that maps input data to a category."
- Model
  - An 'object' that's the result of training ,is a model.
  - eg : Linear regression algorithm is a technique to fit points to a line $y = m\,x + c$. Now after fitting, you get for example, $y = 10\,x + 4$. This a model.
- Simple Linear Regression
  - Understanding relationship b/w two quantitative variables

Modeling Error

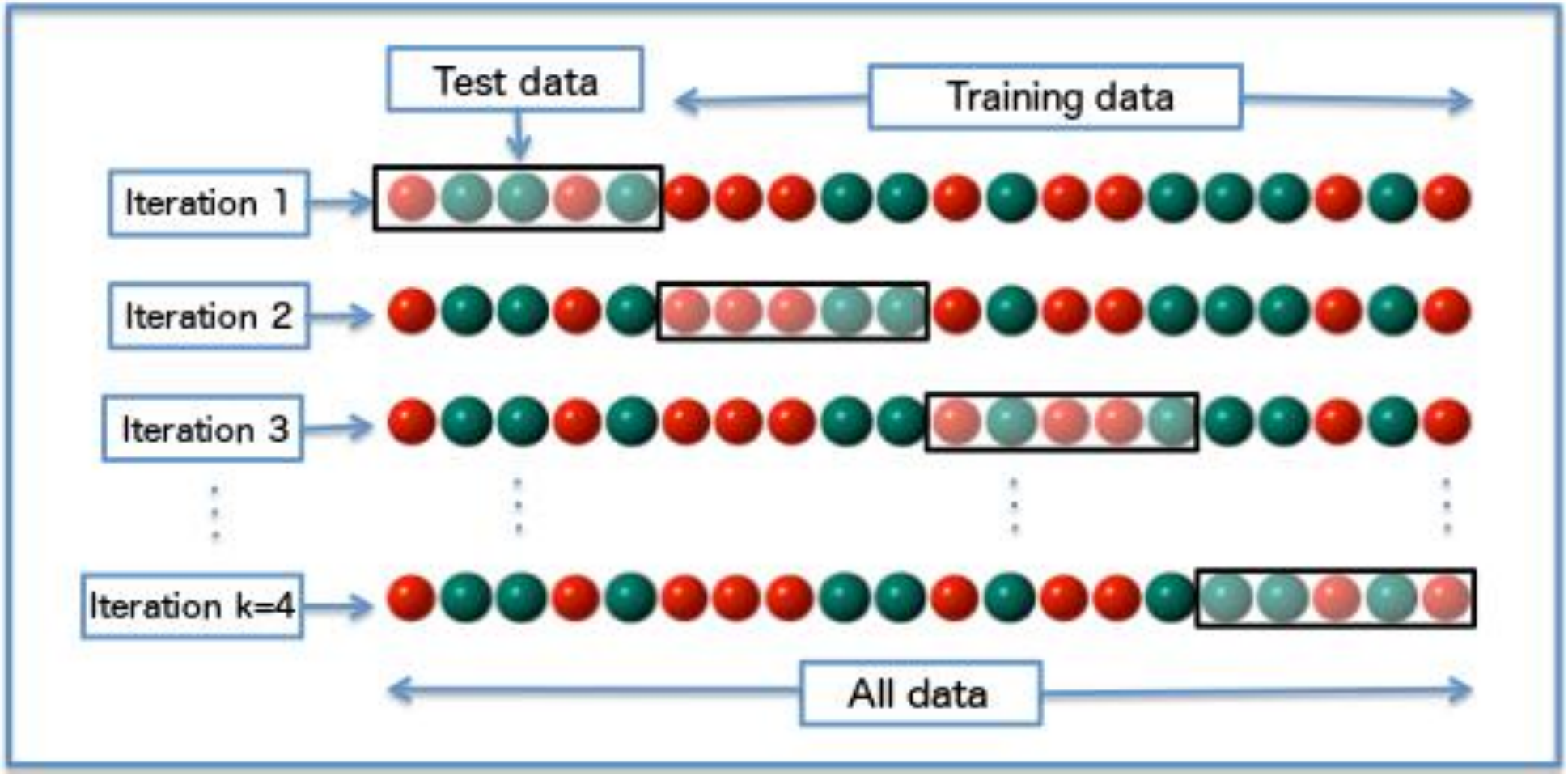
- Overfitting,
  - when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data.

    eg : 100 percent accuracy
- Under fitting
  - not a suitable model and will be obvious as it will have poor performance on the training data
  - a model that can neither model the training data nor generalize to new data.

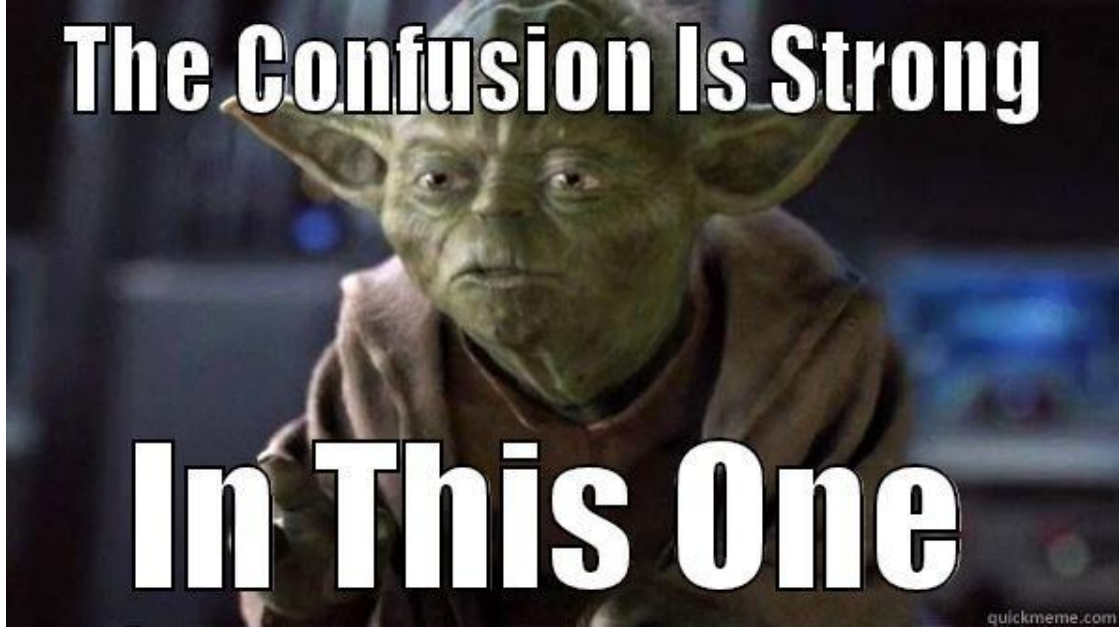
TESTING YOUR MODEL

- Cross validation
  - Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.
  - In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

Cross Validation



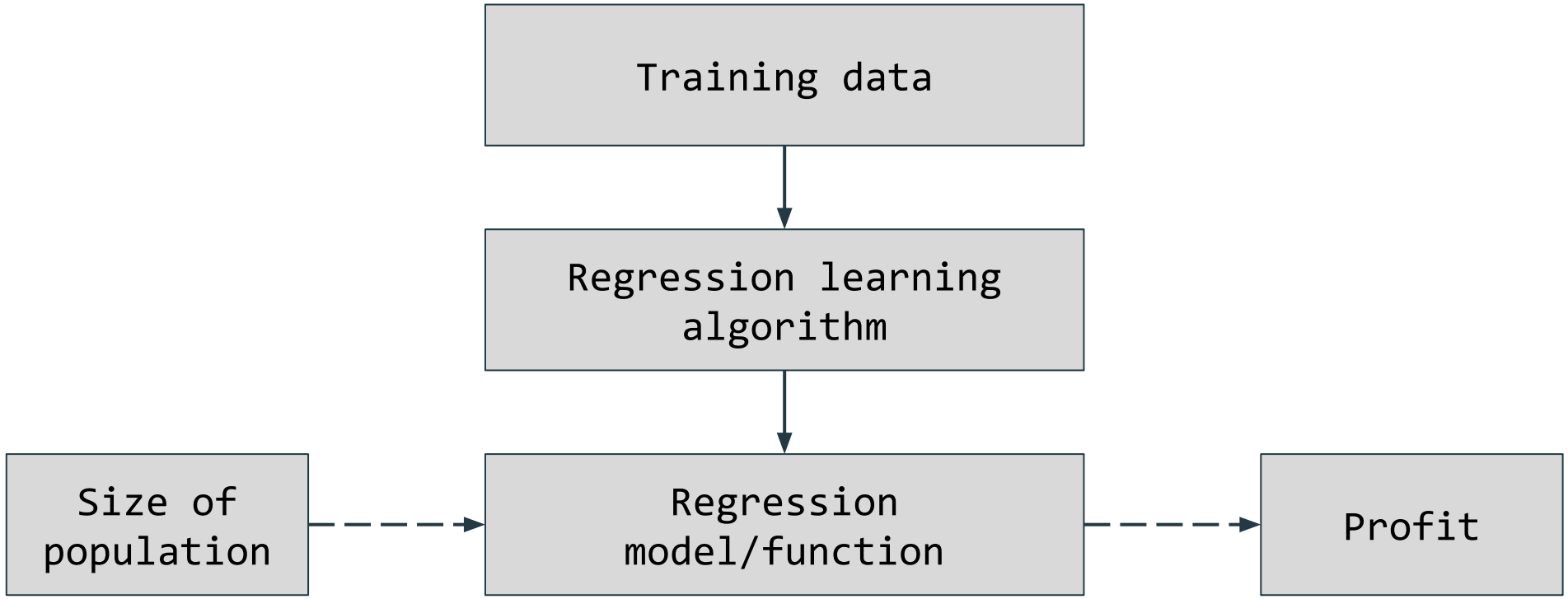
CONFUSION MATRIX



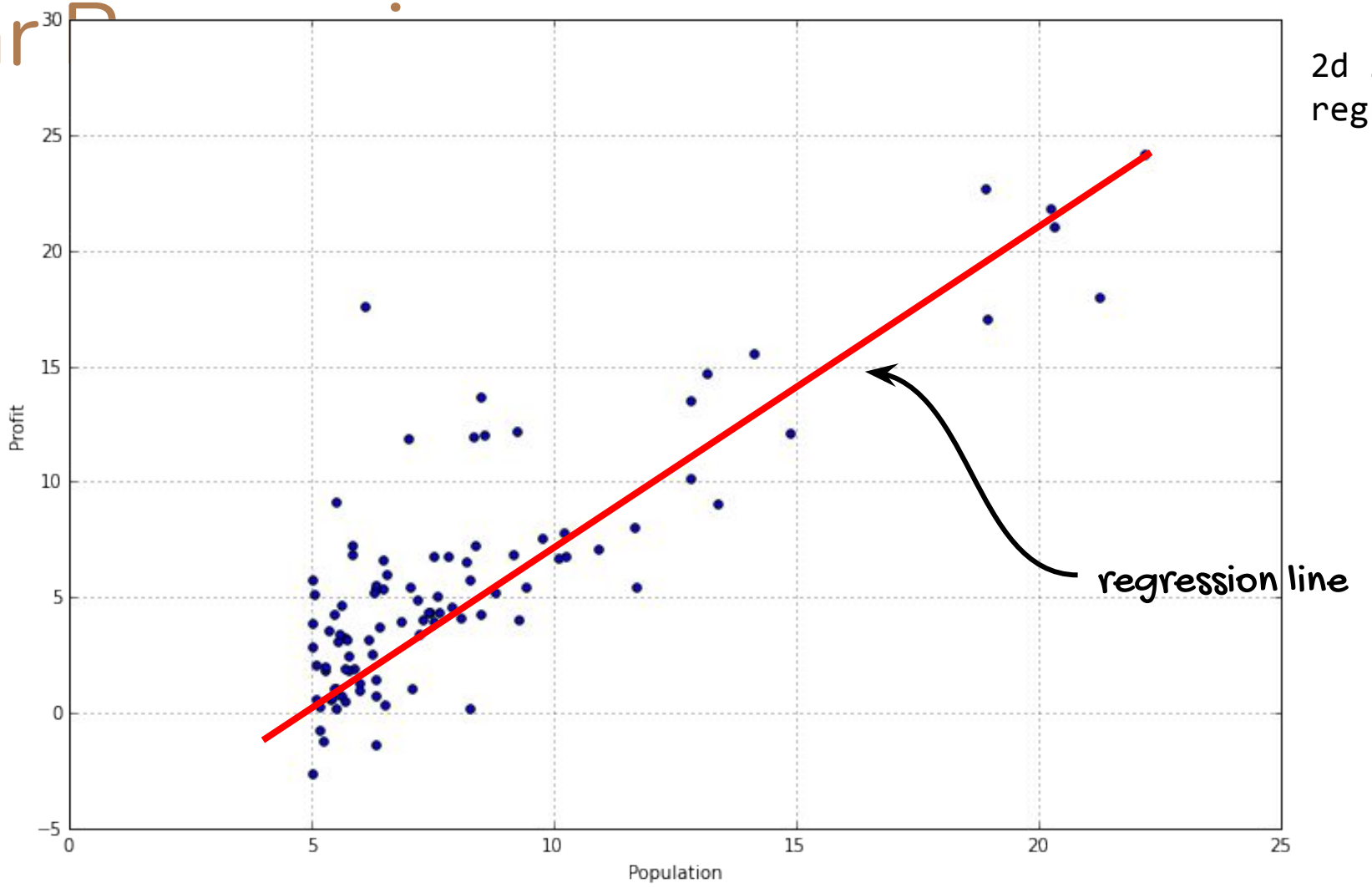
n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

Regression

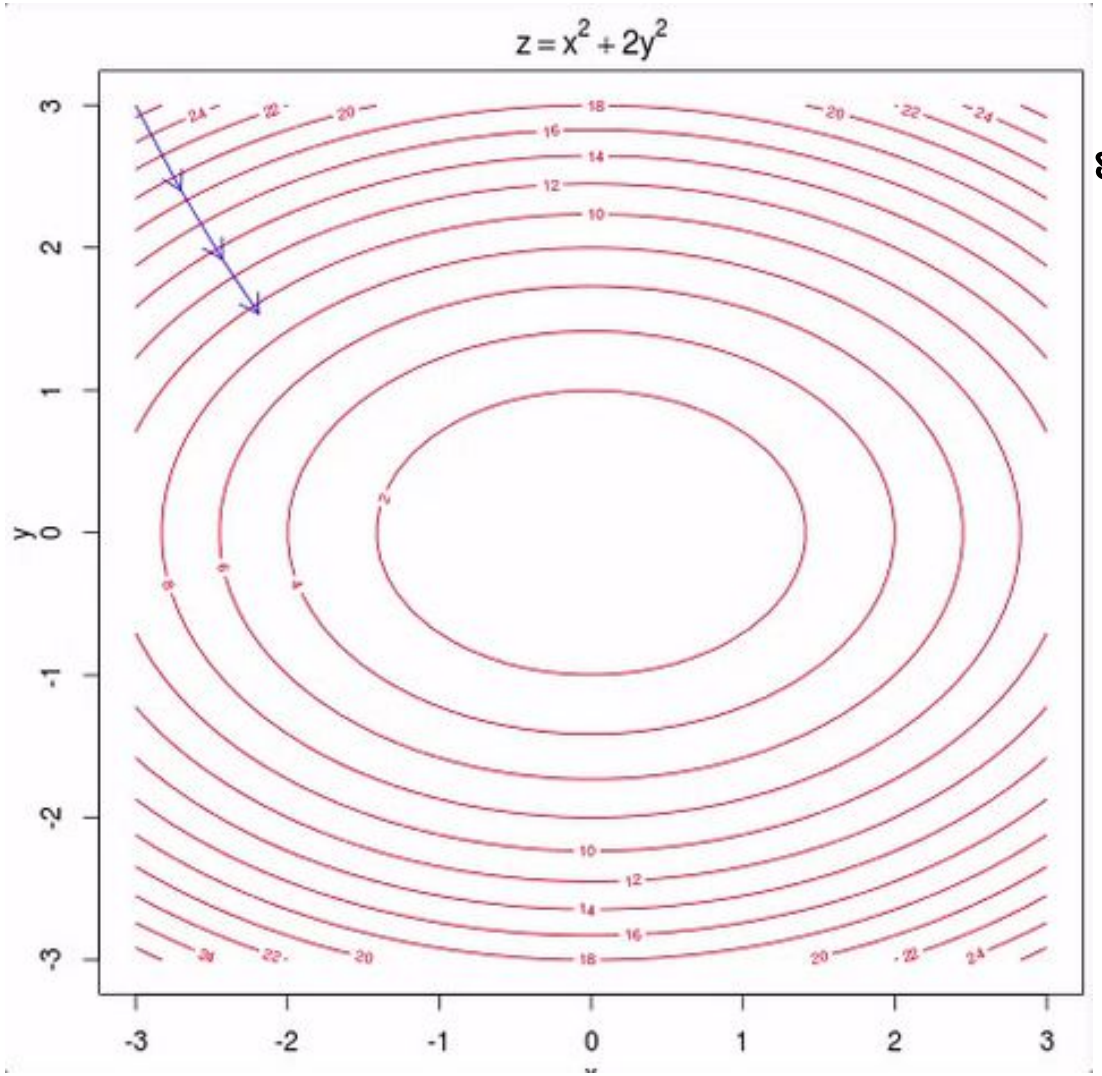
- regression = finding relationships between variables



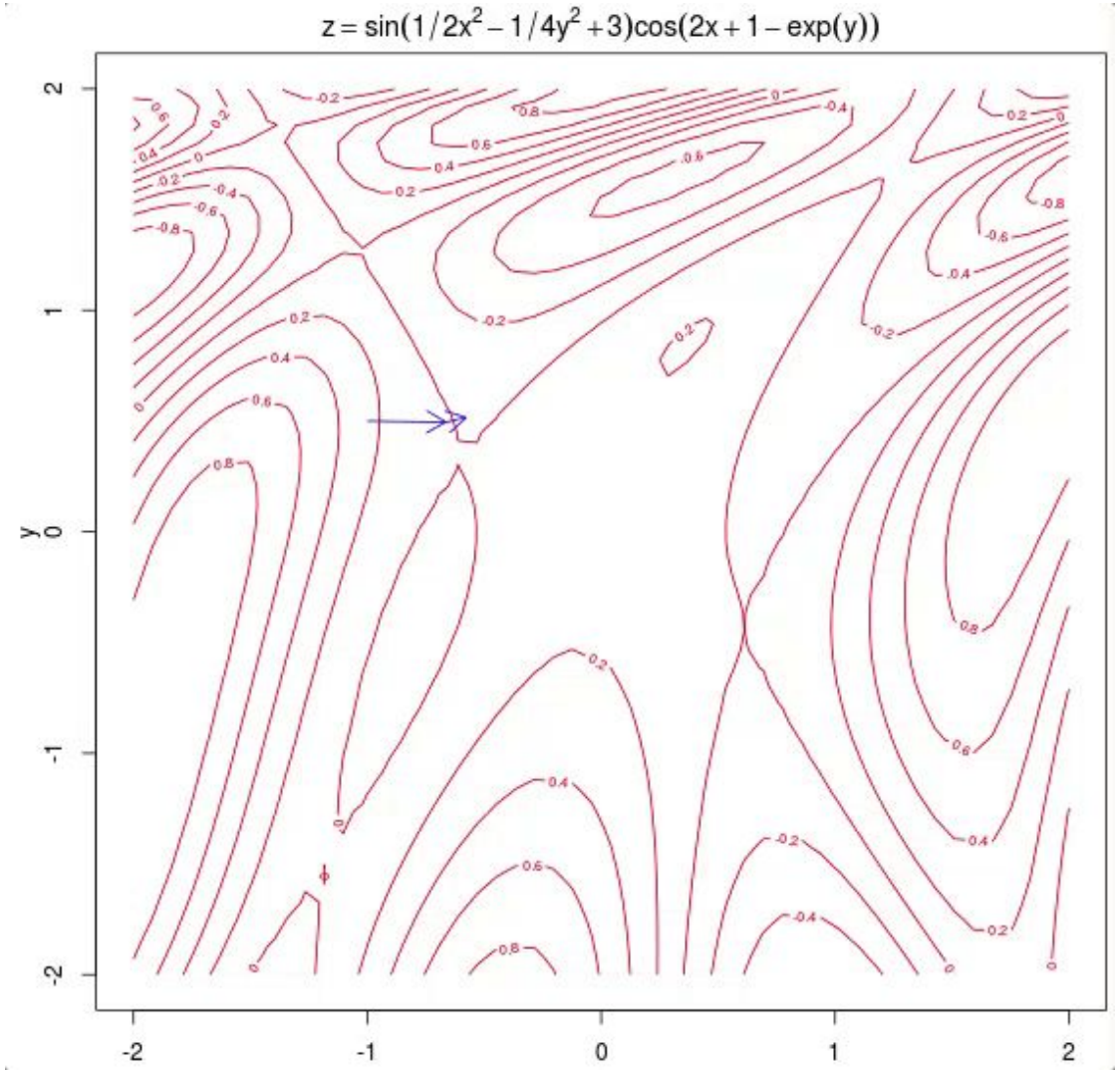
Linear R



2d linear regression



success



failure

Logistic Regression - Xss Payloads

Logistic regression is used for prediction of output which is binary.

That means it can take only two possible values such as “Yes or No”,

It can be used for categorical dependent variables with more than 2 classes. In this case it’s called Multinomial Logistic Regression.

Demo -Time

https://github.com/oreilly-mlsec/book-resources/tree/master/chapter8/waf

Face Recognition -OSINT

Complex problem to solve

Preprocessing Images using Facial Detection and Alignment

Generating Facial Embeddings in Tensorflow

SVM Classifier

Convolutional Neural Networks

Using Amazon Rekognition

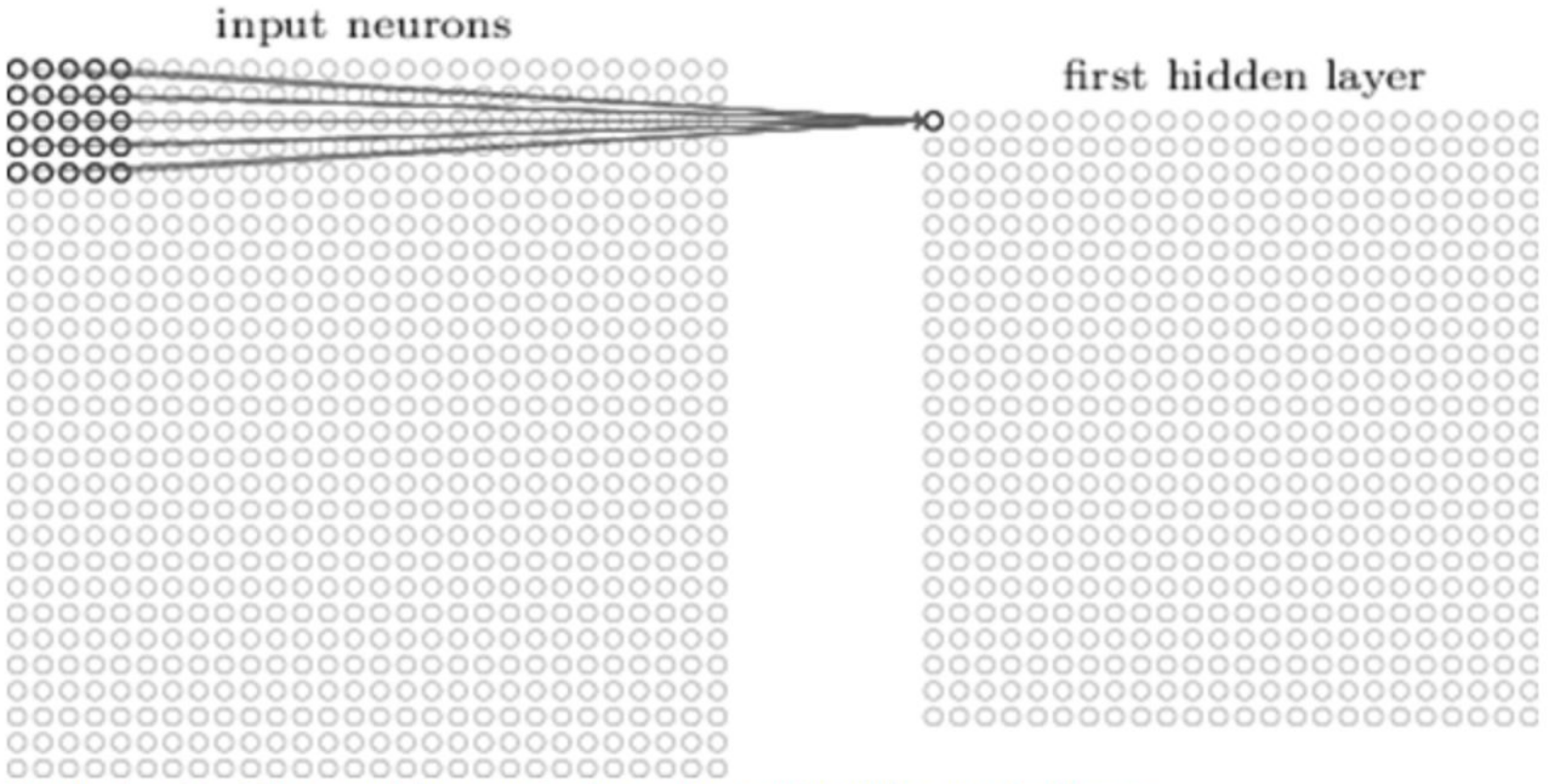
```Rekognition Image enables you to find similar faces in a large collection of images. You can create an index of faces detected in your images. Rekognition Image's fast and accurate search returns faces that best match your reference face.```
```

source-code

Cloud Service :

https://github.com/antojoseph/AI-Scripts/blob/master/face-match.py

convolutional neural networks



Visualization of 5 x 5 filter convolving around an input volume and producing an activation map

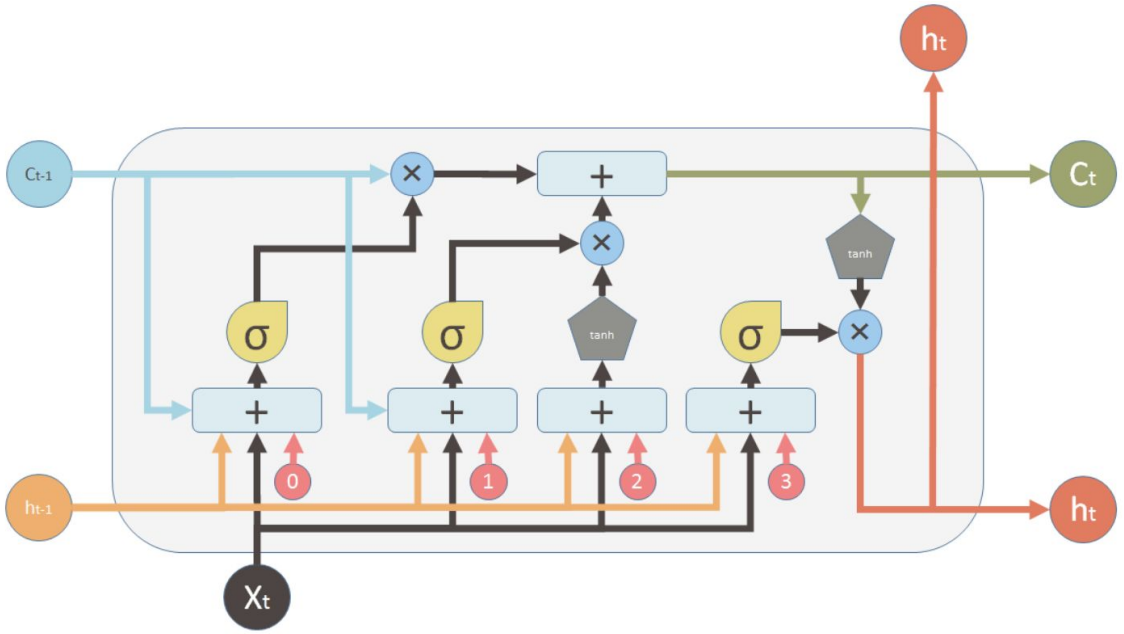
fuzzing - Data Generation

Needs Structurally valid data

Need Different combinations of Structurally valid data

Also needs to be of valid syntax

lstm



Inputs:

- X_t Input vector
- C_{t-1} Memory from previous block
- h_{t-1} Output of previous block

outputs:

- C_t Memory from current block
- h_t Output of current block

Nonlinearities:

- σ Sigmoid
- tanh Hyperbolic tangent

Bias: 0

Vector operations:

- × Element-wise multiplication
- \+ Element-wise Summation / Concatenation

demo

https://github.com/alexknvl/fuzzball

https://github.com/karpathy/char-rnn

https://github.com/keras-team/keras/blob/master/examples/lstm_text_generation.py

HMM

https://github.com/alexknvl/fuzzball

Principle of Memorylessness

i.e next state depends only on the previous state

its ideal for recognizing something based on sequence

when you have a state machine with hidden states , but you know the observation from that state

demo

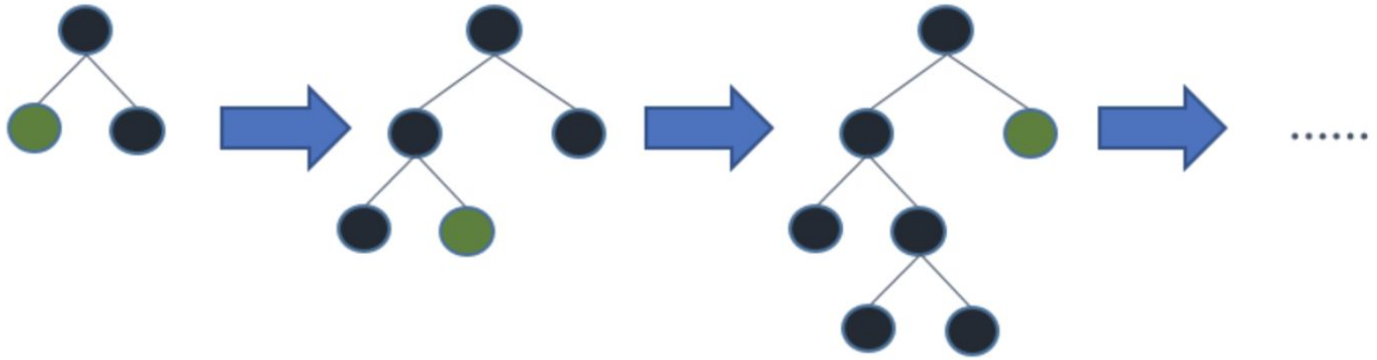
engame - obfuscato4 : https://github.com/CylanceSPEAR/MarkovObfuscate

lightgbm

Light GBM is a gradient boosting framework that uses tree based learning algorithm.

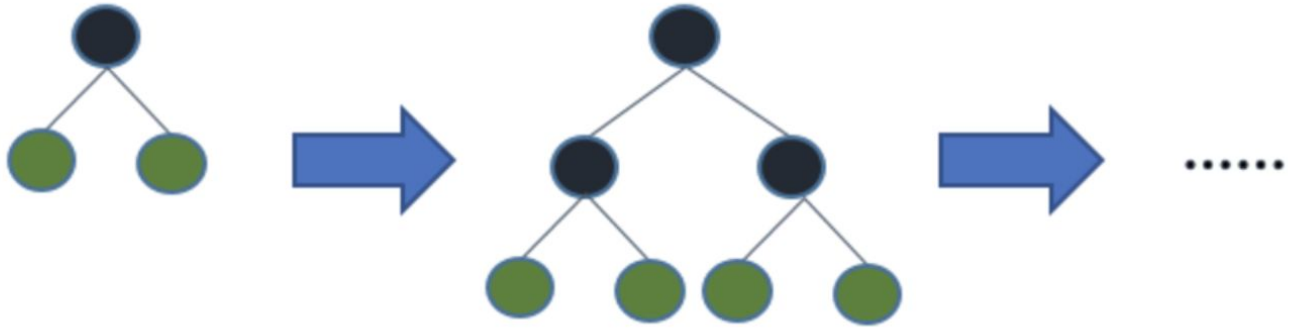
Light GBM grows tree vertically i.e it grows tree leaf-wise while other algorithm grows level-wise.

Its really fast



Leaf-wise tree growth

Explains how LightGBM works



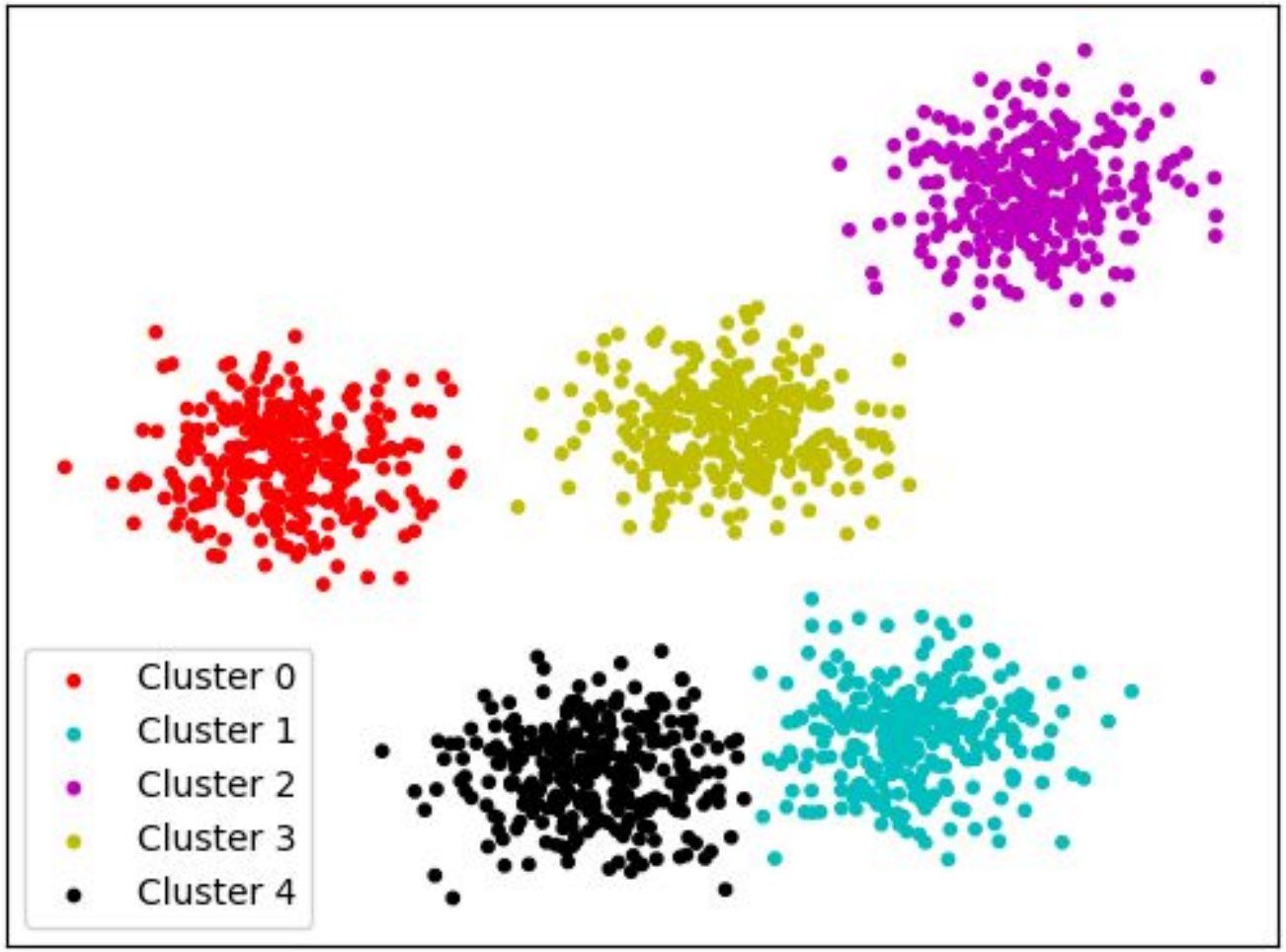
Level-wise tree growth

Decision Trees - Visualization

http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

Malware Classification - demo

Clustering - K means



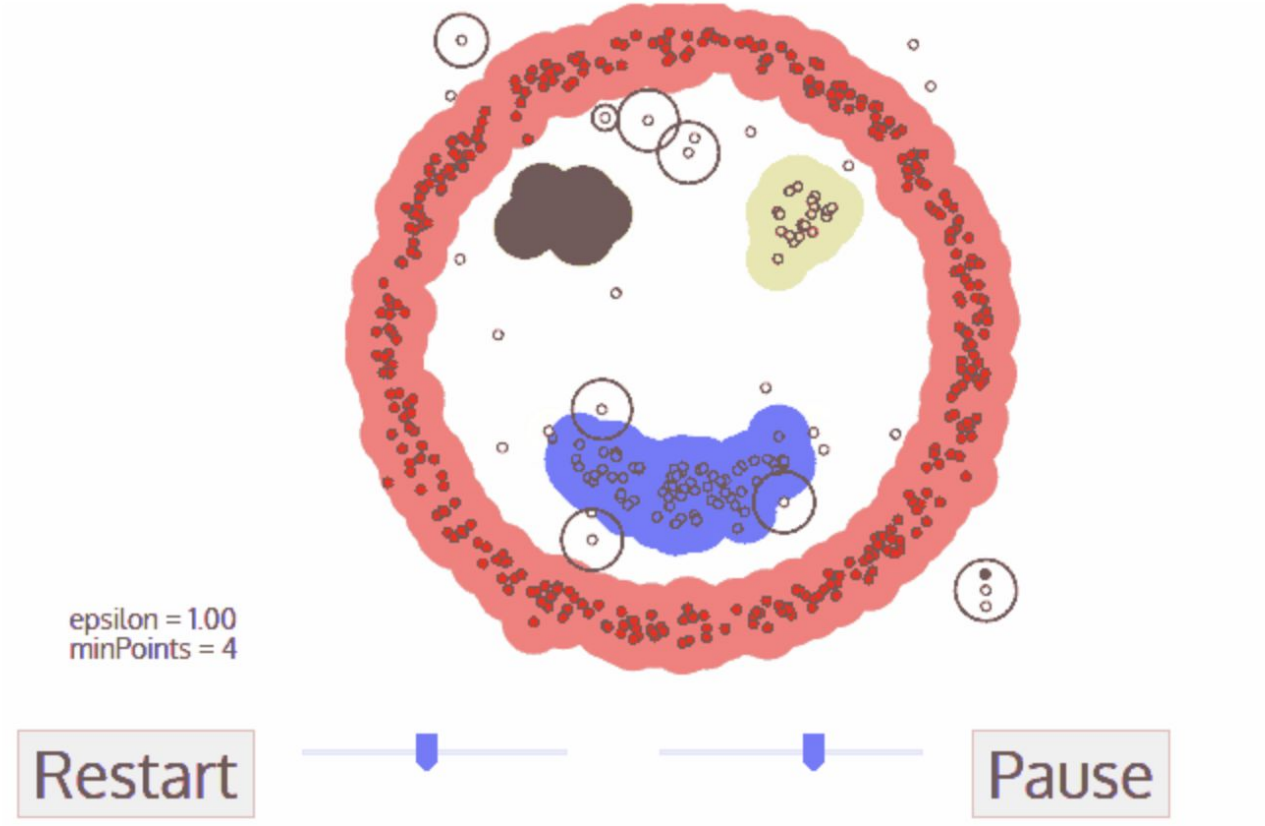
Clustering - DBSCAN

Density-Based

Spatial Clustering

of Applications

with Noise



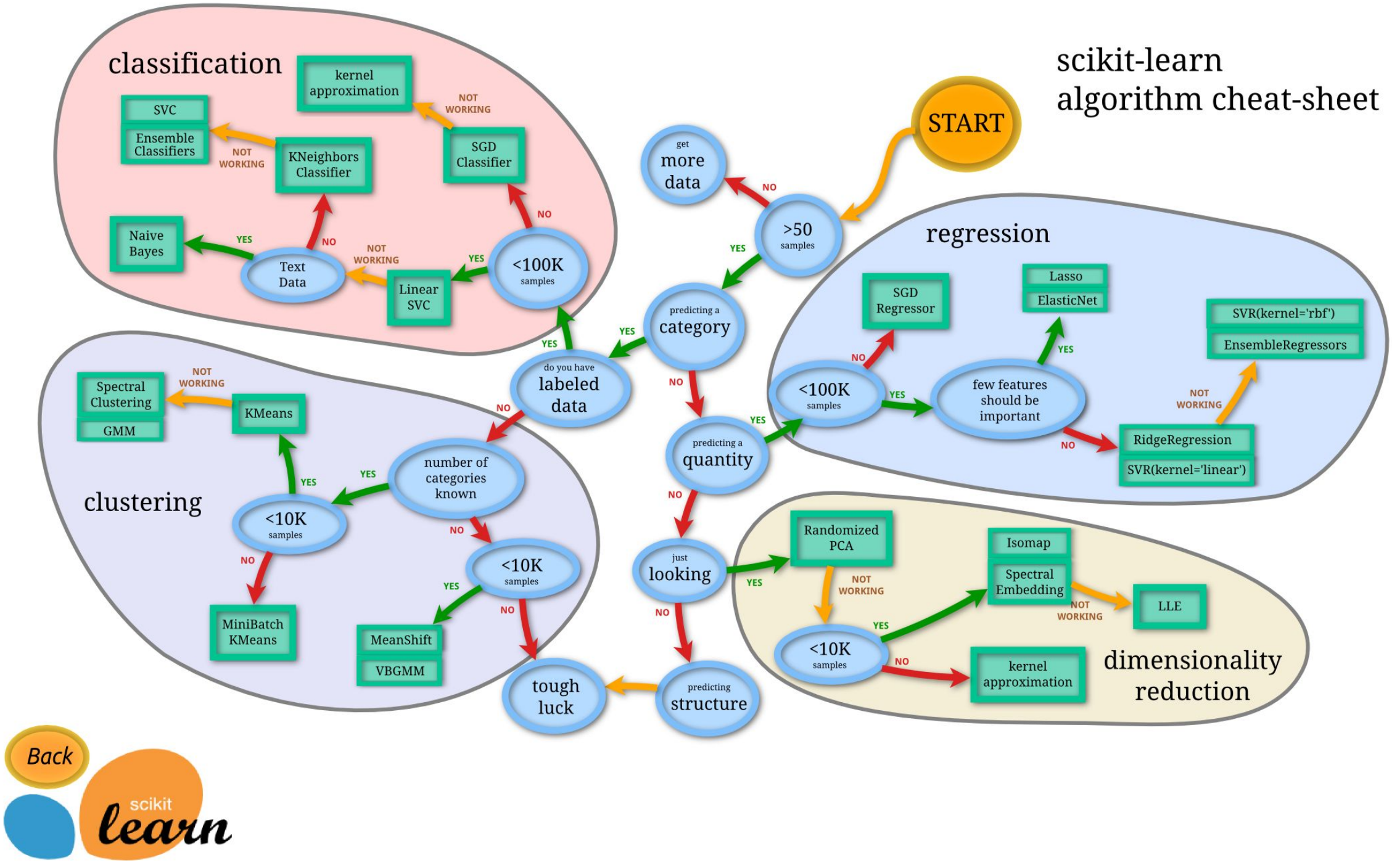
DBSCAN Smiley Face Clustering

demo

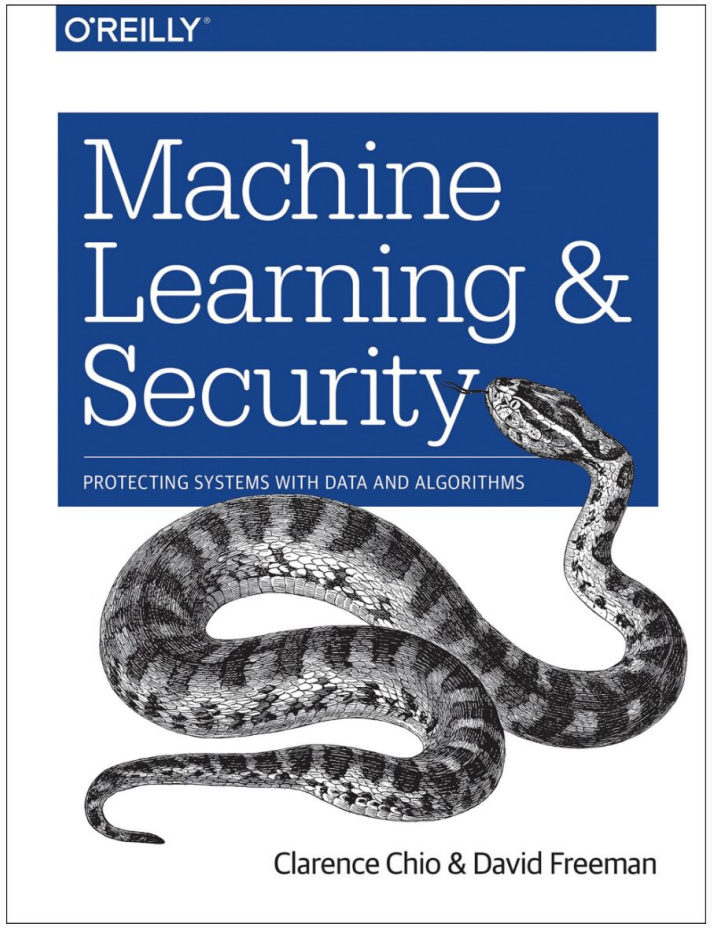
https://github.com/CylanceSPEAR/NMAP-Cluster

Source Code

https://github.com/antojoseph/AI-Scripts


scikit-learn
algorithm cheat-sheet
START
>50 samples
NO
get more data
YES
predicting a category
YES
do you have labeled data
NO
predicting a quantity
NO
just looking
NO
predicting structure
tough luck
YES
<100K samples
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
SGD Classifier
NOT WORKING
kernel approximation
NOT WORKING
KNeighbors Classifier
NOT WORKING
SVC
Ensemble Classifiers
classification
YES
number of categories known
NO
<10K samples
NO
MiniBatch KMeans
YES
KMeans
NOT WORKING
Spectral Clustering
GMM
YES
<10K samples
NO
MeanShift
VBGMM
clustering
YES
<100K samples
NO
SGD Regressor
YES
few features should be important
NO
RidgeRegression
NOT WORKING
SVR(kernel='linear')
YES
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
regression
YES
Randomized PCA
NOT WORKING
<10K samples
NO
kernel approximation
YES
Isomap
Spectral Embedding
NOT WORKING
LLE
dimensionality reduction
Back
scikit learn

Resources:



Learning to Protect Systems with Data & Algorithms

Dive into the science to separate the facts from the buzz

Machine learning is eating the world. As security practitioners, understanding what data science can do for you is rapidly gaining importance. Gaining a literacy in data science is the only way forward.

If you are working in the security field and want to use machine learning to improve your systems, this book is for you. If you have worked with machine learning and now want to use it to solve security problems, this book is also for you.

In examining a broad range of topics in the security space, we provide examples of how machine learning can be applied to augment or replace rule-based or heuristic solutions to problems like intrusion detection, malware classification, or network analysis. In addition to exploring the core machine learning algorithms and techniques, we focus on the challenges of building maintainable, reliable, and scalable data mining systems in the security space. Through worked examples and guided discussions, we show you how to think about data in an adversarial environment and how to identify the important signals that can get drowned out by noise.

Buy with amazon | Tweet | Star 30

Or, read online on O'REILLY Safari



Get Involved ?

aivillage.slack.com

https://twitter.com/aivillage_dc

ONLINE SERVICES

- https://cloud.google.com/prediction/docs/
- http://www.perspectiveapi.com/

Get in touch ?

- Twitter : @antojosep007
- Twitter : @cchio