



Practical Machine Learning in Infosec



Data Mining for Cyber Security (SF Bay Area)

Do you have overlapping interests in cyber security, data mining, and machine learning? Do you want to be more informed on how data collection and analysis can be used in the security context?

Whether you are a researcher, engineer in the InfoSec, AppSec, NetSec, CloudSec field, or whether you are a developer, hacker, lurker, generally interested in this area, join us!

Examples of topics we deal with are: spam, fraud, network intrusion, botnet intrusion, server security etc.

Who do I know here?
Log in with Facebook to find out

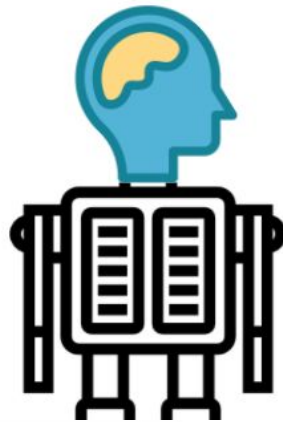
What's new

Join us
Join us and be the first to know when new Meetups are scheduled

Welcome!
Upcoming (1) Past Calendar

May placeholder for meet
Needs a location
Coming soon

Organizer:
Clarence Chio



Making & Breaking Machine Learning Systems

(for infosec)

clarence chio (@cchio)

YouTube video player interface for "Machine Duping Pwning Deep Learning Systems" by Clarence Chio.

AUGUST 4-7, 2016
PARIS + BALLY'S | LAS VEGAS

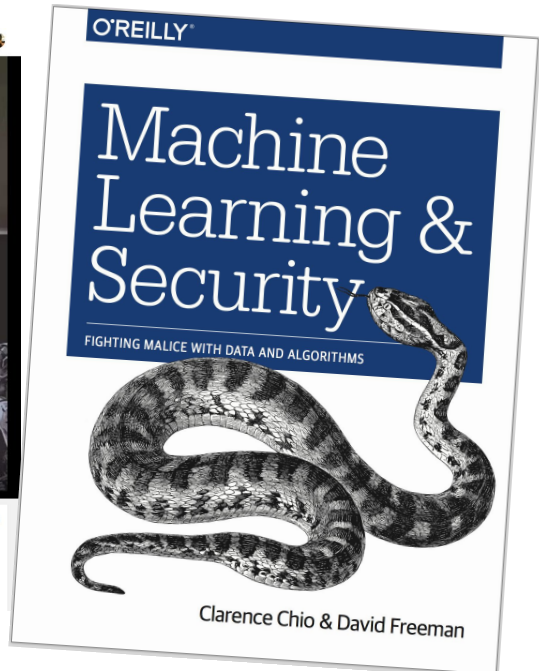
Machine Duping Pwning Deep Learning Systems

CLARENCE CHIO
MLHACKER
@cchio

DEF CON

DEF CON 24 - Clarence Chio - Machine Duping 101: Pwning Deep Learning Systems

3,427 views



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

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

who are we?

anto joseph (@antojosep007)

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**

A banner for the Nullcon 2017 conference. The background is black with white circuit-like patterns. At the top left, it says 'NULLCON INTERNATIONAL SECURITY CONFERENCE GOA 2017'. In the center, there is a silhouette of a group of people. Below that, a red square contains a silhouette of a man in a suit. To the right of the silhouette, the text reads 'ANTO JOSEPH SENIOR SECURITY ENGINEER AT INTEL ADVERSARIAL MACHINE LEARNING'. At the bottom left, it says 'CONFERENCE SPEAKERS' and 'WORKSHOPS AND VILLAGES'. At the bottom center, a gold bar contains the dates '28TH FEBRUARY - 4TH MARCH 2017'. At the bottom right, the website 'www.nullcon.com' is listed.

NULLCON
INTERNATIONAL SECURITY CONFERENCE GOA 2017

ANTO JOSEPH
SENIOR SECURITY ENGINEER AT INTEL
ADVERSARIAL MACHINE LEARNING

CONFERENCE SPEAKERS WORKSHOPS AND VILLAGES

28TH FEBRUARY - 4TH MARCH 2017

www.nullcon.com

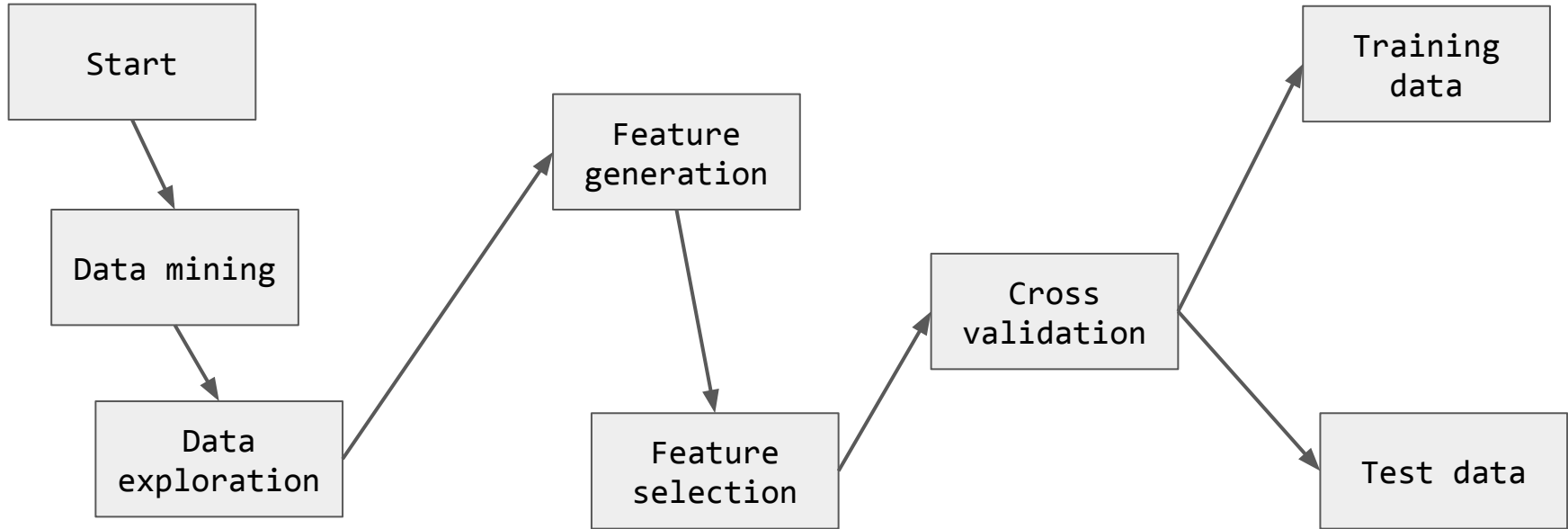
Agenda

- Intro to the development environment
- Spam classifiers
- Anomaly detection
- Classifying malware
- Security of machine learning

(supervised)

Machine learning from 10,000ft

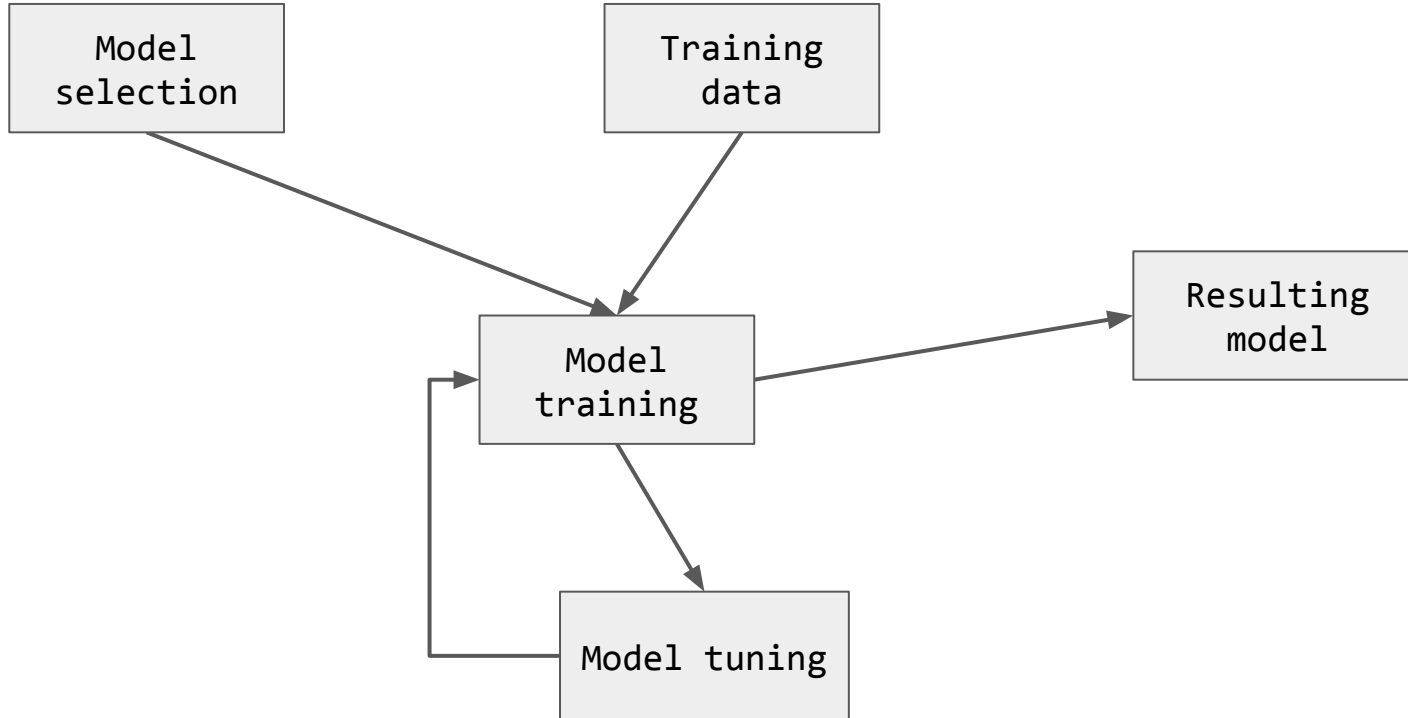
data engineering phase



(supervised)

Machine learning from 10,000ft

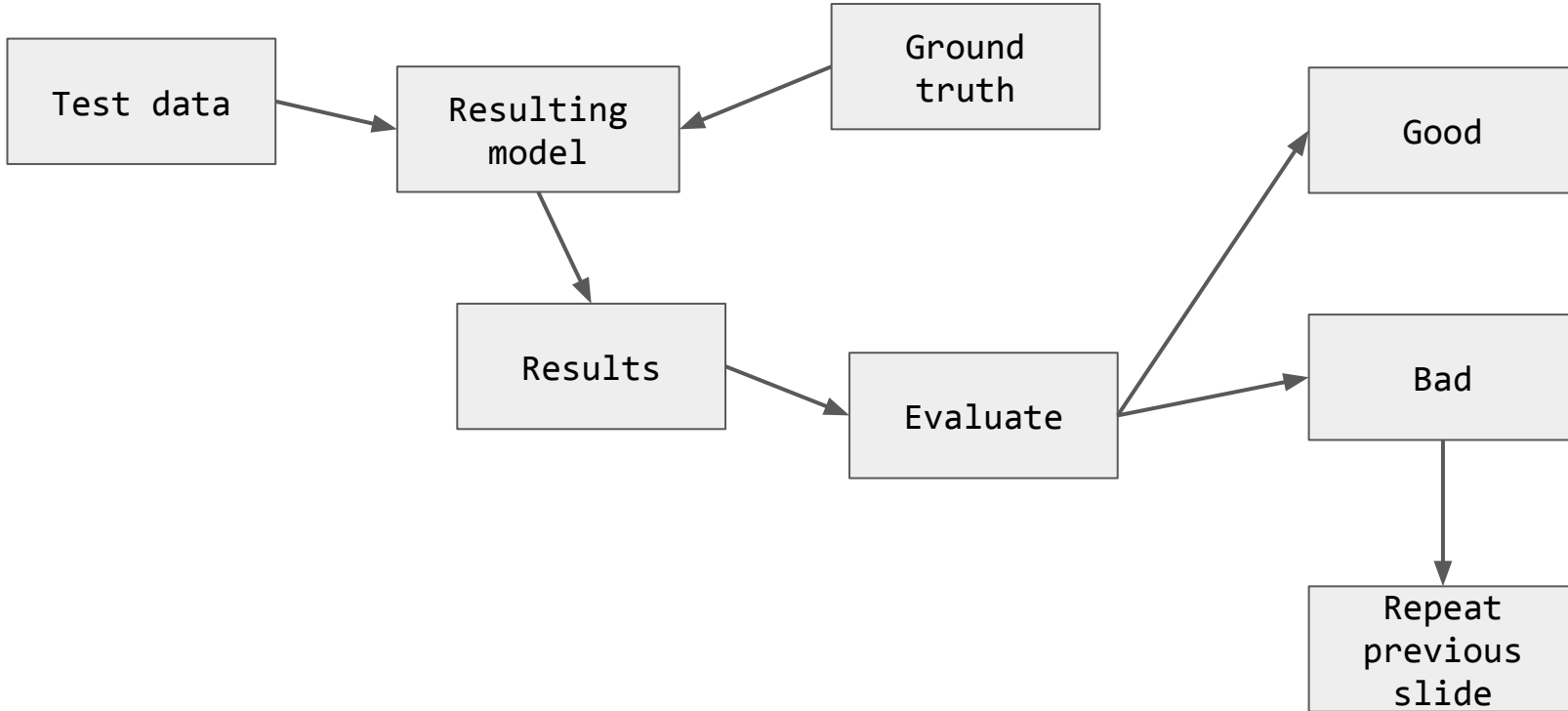
model training phase



(supervised)

Machine learning from 10,000ft

model validation phase



Python toolkits

- `scikit-learn` - Python library that implements a comprehensive range of machine learning algorithms
- `TensorFlow` - library for numerical computation using data flow graphs / deep learning

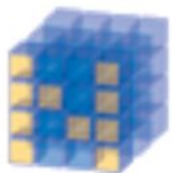
scikit-learn

- 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/neural nets
 - But not restricted to just deep models
- Multiple GPU Support

Data science libs



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

HANDS ON

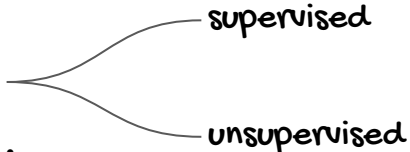
classifying spam

The dataset: 2007 TREC Public Spam Corpus

<http://plg.uwaterloo.ca/~gvcormac/treccorpus07/>

MACHINE LEARNING 101

Types of machine learning use cases:

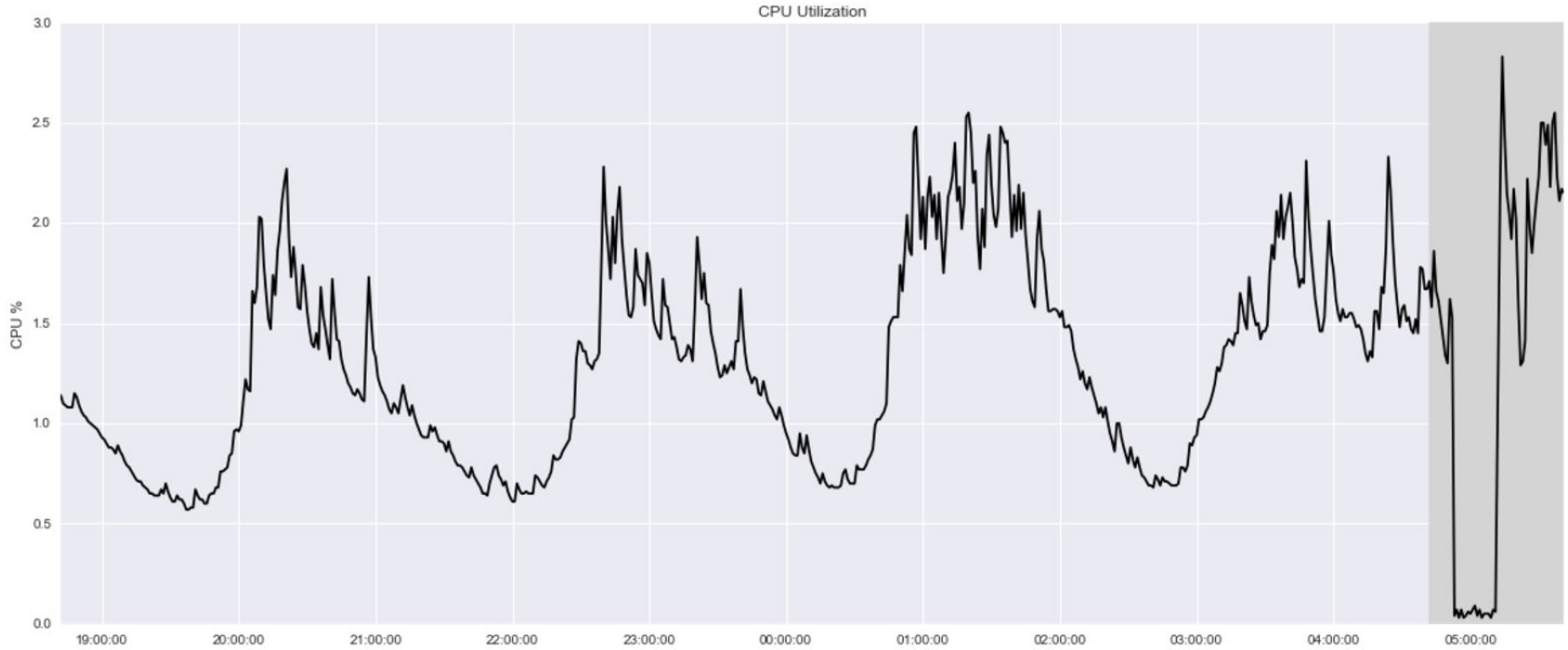
- Regression
 - Classification
 - Anomaly detection
 - Recommendation
- won't cover here, but check out [this talk](#)
- 
- The diagram shows a list of machine learning use cases. The word 'Classification' is connected by two curved lines to the words 'supervised' and 'unsupervised' on the right side of the slide.

This covers **EVERYTHING**. (almost)

HANDS ON

Anomaly Detection

Anomaly detection



Anomaly detection

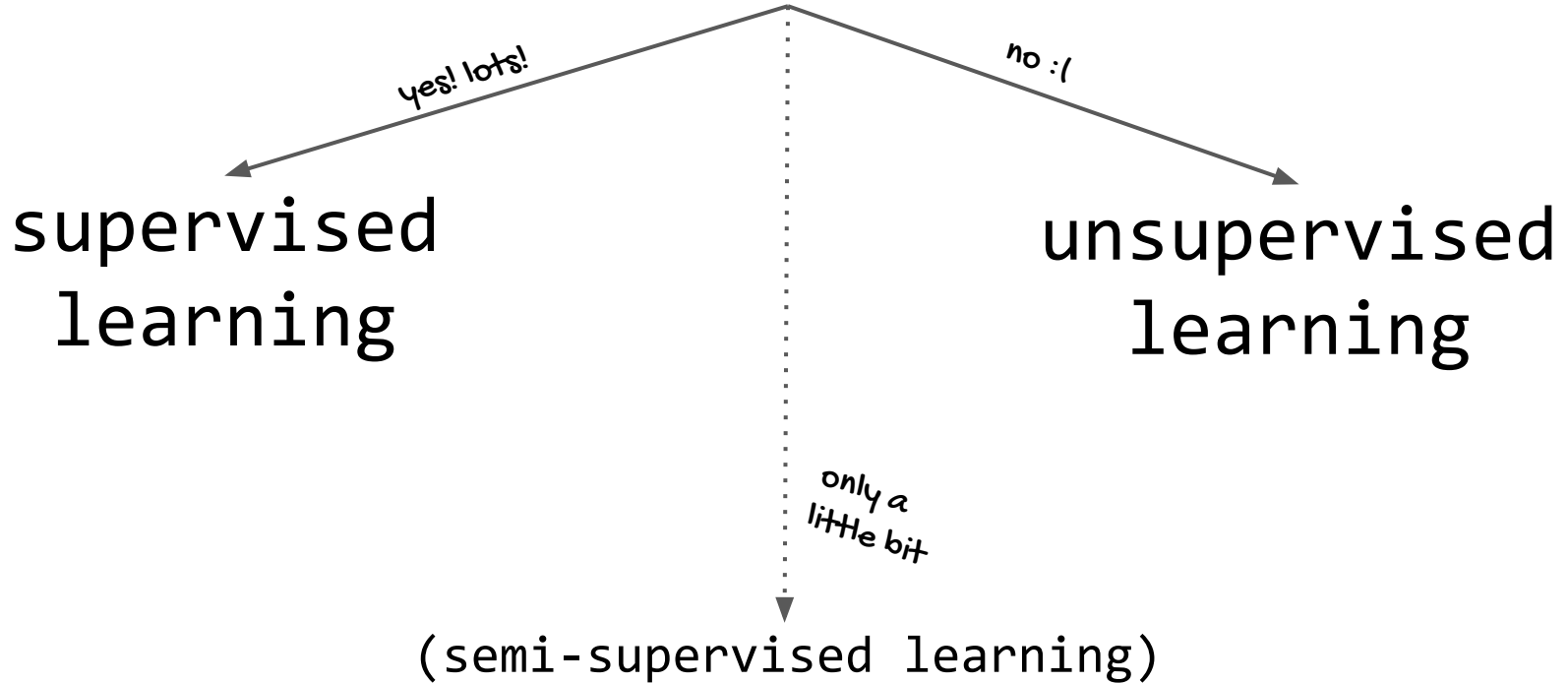
- Outliers vs. novelties
 - *novelties*: unobserved pattern in new observations not included in training data
- Simple statistics/forecasting methods
 - Exponential smoothing, Holt-Winters algorithm
- Machine learning methods
 - Elliptical envelope, density-based, clustering, SVM

Classification

Classification

labeled data - do you have it?

Classification



Supervised classification

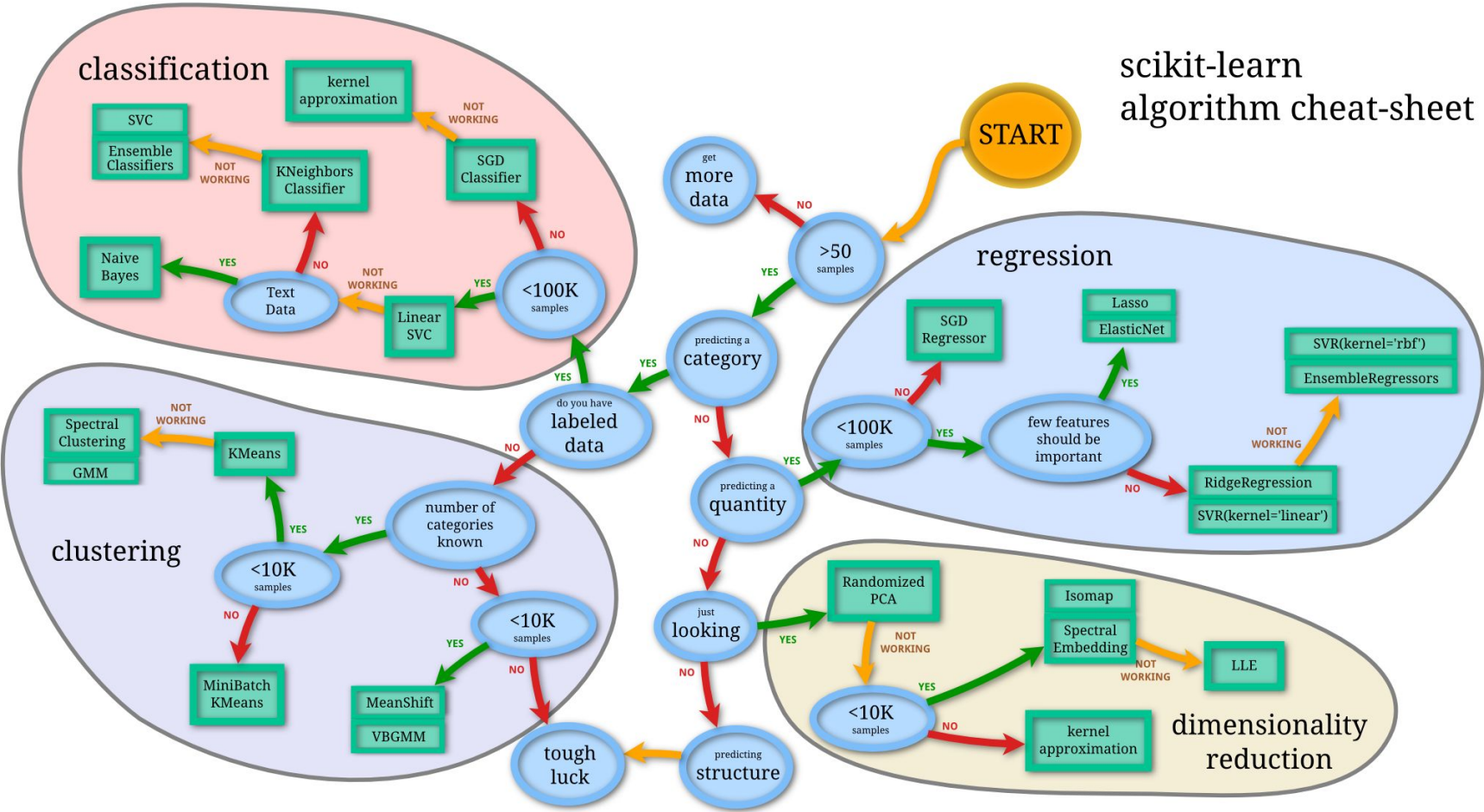
- Many different algorithms!
- e.g.
 - Logistic regression (it's called regression but is *not* regression)
 - Naive Bayes
 - K-nearest neighbors
 - Support Vector Machines
 - Decision Trees

Unsupervised classification

- Mainly refers to **clustering**
- **Four** types:
 - **Centroid:** K-Means
 - **Distribution:** Gaussian mixture models
 - **Density:** DBSCAN
 - **Connectivity:** Hierarchical clustering

scikit-learn algorithm cheat-sheet

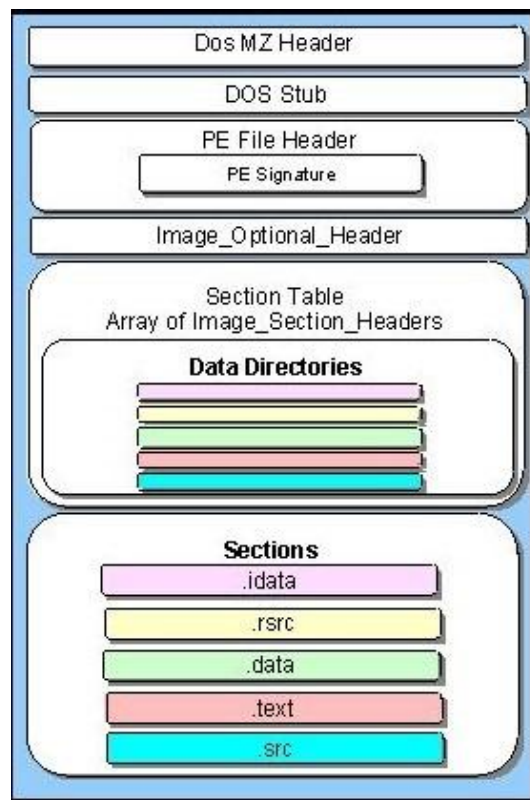
START



HANDS ON

classifying malware

Portable executable (PE)



pefile dump

-----FILE_HEADER-----

```
[IMAGE_FILE_HEADER]
Machine:          0x14C
NumberOfSections: 0x4
TimeStamp:        0x851C3163
[INVALID TIME]
PointerToSymbolTable:
0x74726144
NumberOfSymbols:  0x455068
SizeOfOptionalHeader: 0xE0
Characteristics:  0x818F
```

-----OPTIONAL_HEADER-----

```
[IMAGE_OPTIONAL_HEADER]
Magic:          0x10B
MajorLinkerVersion: 0x2
MinorLinkerVersion: 0x19
SizeOfCode:     0x200
SizeOfInitializedData: 0x45400
SizeOfUninitializedData: 0x0
AddressOfEntryPoint: 0x2000
BaseOfCode:     0x1000
BaseOfData:     0x2000
ImageBase:      0xDE0000
SectionAlignment: 0x1000
FileAlignment:  0x1000
MajorOperatingSystemVersion: 0x1
MinorOperatingSystemVersion: 0x0
```

-----PE Sections-----

```
[IMAGE_SECTION_HEADER]
Name:          CODE
Misc:          0x1000
Misc_PhysicalAddress:
0x1000
Misc_VirtualSize: 0x1000
VirtualAddress: 0x1000
SizeOfRawData:  0x1000
PointerToRawData: 0x1000
PointerToRelocations: 0x0
PointerToLinenumbers: 0x0
NumberOfRelocations: 0x0
NumberOfLinenumbers: 0x0
Characteristics: 0xE0000020
Flags: MEM_WRITE, CNT_CODE,
MEM_EXECUTE, MEM_READ
Entropy: 0.061089 (Min=0.0,
Max=8.0)

[IMAGE_SECTION_HEADER]
Name:          DATA
Misc:          0x45000
Misc_PhysicalAddress:
0x45000
Misc_VirtualSize: 0x45000
VirtualAddress:  0x2000
SizeOfRawData:  0x45000
```

```
PointerToRawData: 0x2000
PointerToRelocations: 0x0
PointerToLinenumbers: 0x0
NumberOfRelocations: 0x0
NumberOfLinenumbers: 0x0
Characteristics: 0xC0000040
Flags: MEM_WRITE,
CNT_INITIALIZED_DATA,
MEM_READ
Entropy: 7.980693 (Min=0.0,
Max=8.0)
```

```
[IMAGE_SECTION_HEADER]
Name:          NicolasB
Misc:          0x1000
Misc_PhysicalAddress:
0x1000
Misc_VirtualSize: 0x1000
VirtualAddress: 0x47000
SizeOfRawData:  0xEFEFADFF
PointerToRawData:
0x47000
PointerToRelocations: 0x0
PointerToLinenumbers: 0x0
...
```

-----Parsing Warnings-----

Suspicious NumberOfRvaAndSizes in the Optional Header. Normal values are never larger than 0x10, the value is: 0xdffffd

Error parsing section 2. SizeOfRawData is larger than file.

-----DOS_HEADER-----

```
[IMAGE_DOS_HEADER]
e_magic:          0x5A4D
e_cblp:           0x50
e_cp:             0x2
```

-----NT_HEADERS-----

```
[IMAGE_NT_HEADERS]
Signature:        0x4550
```

PE feature vector

Name|md5|Machine|SizeOfOptionalHeader|Characteristics|MajorLinkerVersion|MinorLinkerVersion|SizeOfCode|SizeOfInitializedData|SizeOfUninitializedData|AddressOfEntryPoint|BaseOfCode|BaseOfData|ImageBase|SectionAlignment|FileAlignment|MajorOperatingSystemVersion|MinorOperatingSystemVersion|MajorImageVersion|MinorImageVersion|MajorSubsystemVersion|MinorSubsystemVersion|SizeOfImage|SizeOfHeaders|Checksum|Subsystem|DllCharacteristics|SizeOfStackReserve|SizeOfStackCommit|SizeOfHeapReserve|SizeOfHeapCommit|LoaderFlags|NumberOfRvaAndSizes|SectionsNb|SectionsMeanEntropy|SectionsMinEntropy|SectionsMaxEntropy|SectionsMeanRawsize|SectionsMinRawsize|SectionMaxRawsize|SectionsMeanVirtualsize|SectionsMinVirtualsize|SectionMaxVirtualsize|ImportsNbDLL|ImportsNb|ImportsNbOrdinal|ExportNb|ResourcesNb|ResourcesMeanEntropy|ResourcesMinEntropy|ResourcesMaxEntropy|ResourcesMeanSize|ResourcesMinSize|ResourcesMaxSize|LoadConfigurationSize|VersionInformationSize|**legitimate**

legitimate:

memtest.exe|631ea355665f28d4707448e442fbf5b8|332|224|258|9|0|361984|115712|0|6135|4096|372736|4194304|4096|512|0|0|0|0|1|0|1036288|1024|485887|16|1024|1048576|4096|1048576|4096|0|16|8|5.7668065537|3.60742957555|7.22105072892|59712.0|1024|325120|126875.875|896|551848|0|0|0|0|4|3.26282271103|2.56884382364|3.53793936419|8797.0|216|18032|0|16|1

malware:

VirusShare_76c2574c22b44f69e3ed519d36bd8dff|76c2574c22b44f69e3ed519d36bd8dff|332|224|258|10|0|28672|445952|16896|14819|4096|32768|4194304|4096|512|5|0|6|0|5|0|3977216|1024|680384|2|34112|1048576|4096|1048576|4096|0|16|6|2.65064184009|0.0|6.49788465186|30634.6666667|0|139264|661773.333333|3978|3362816|8|172|1|0|21|3.42072662405|1.86523352037|7.9688495098|6558.42857143|180|67624|0|0|0



SURPRISE CHALLENGE



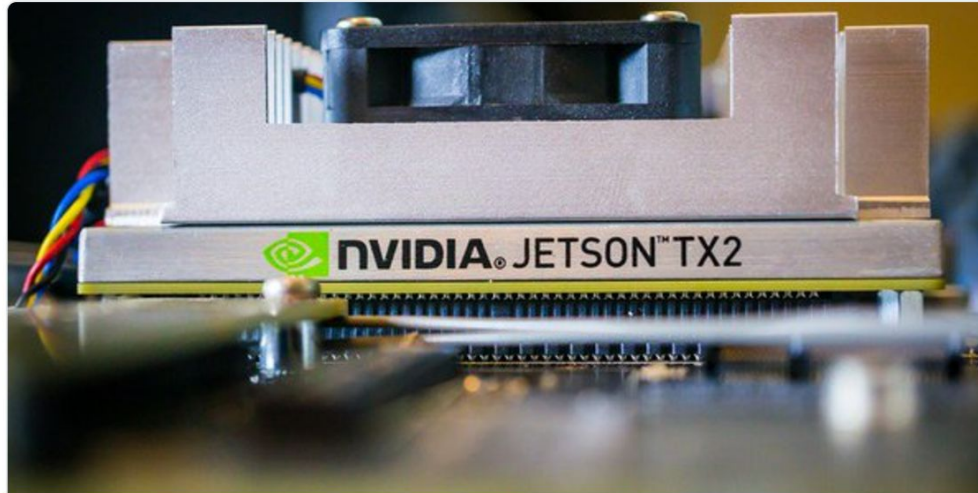
I33tdawg

@I33tdawg

Following



NVIDIA Jetson TX2 is the supercomputer that's going to build the next great idea flip.it/I-UTBH
<--- gotta get one!



NVIDIA Jetson TX2 is the supercomputer that's going to build the next great i...

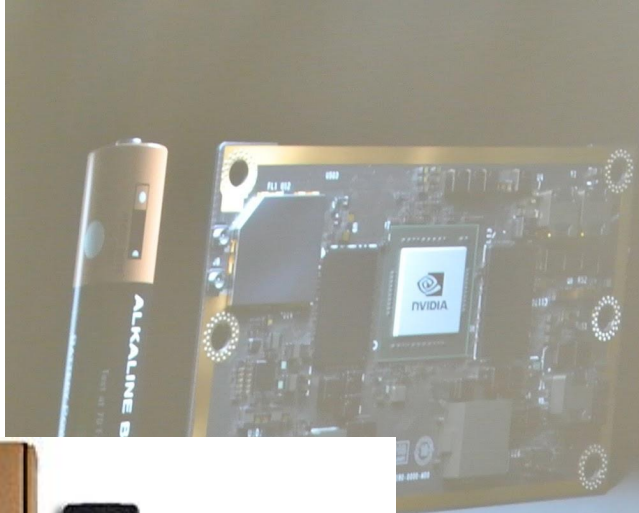
NVIDIA's Jetson TX2 is more than a worthy successor to the original. It's a new way to do things. Artificial Intelligence and machines that can learn are how the things...

flip.it

JETSON TX2

EMBEDDED AI SUPERCOMPUTER

Advanced AI at the edge
JetPack SDK
< 7.5 watts full module
Up to 2X performance or 2X energy efficiency



7.5 watts) delivers up to 2x energy efficiency vs. Jetson TX1 maximum per
mode (< 15 watts) delivers up to 2x performance vs. Jetson TX1 maximum per

CHALLENGE

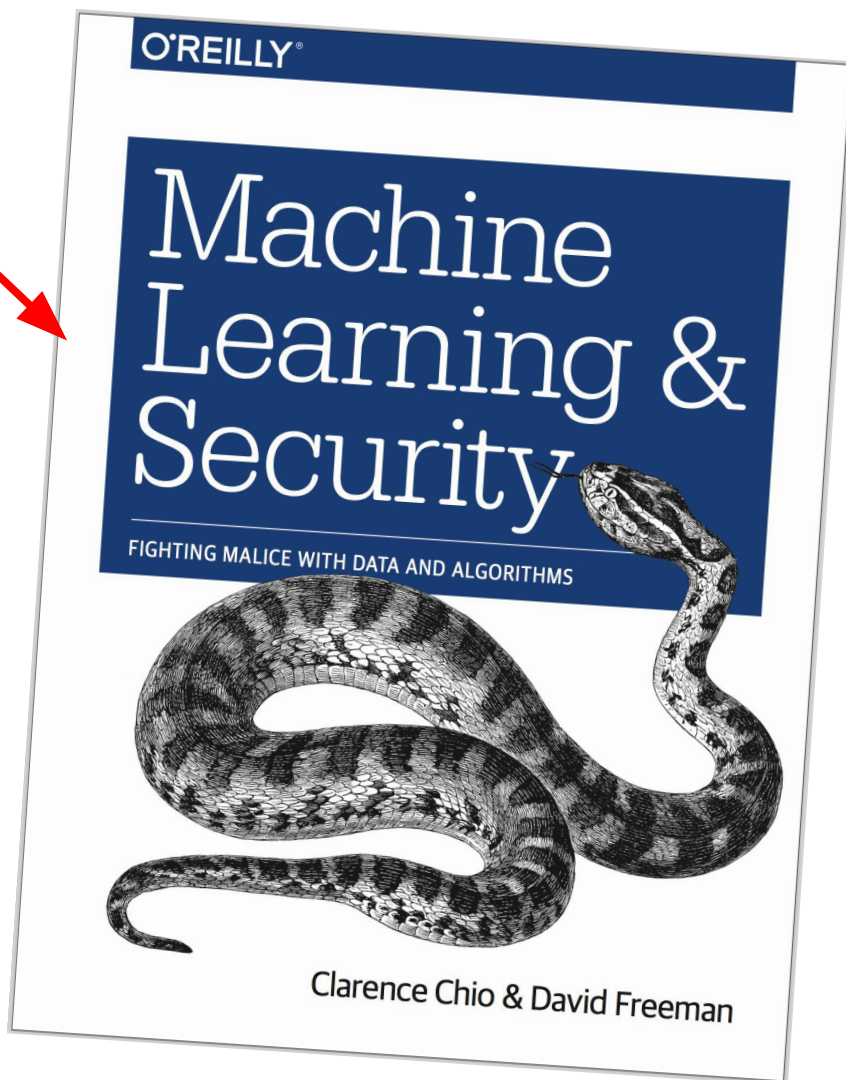
- a. **NETWORK CHALLENGE:** Capture packets on conference network and do some packet classification with machine learning (i.e. attack/non-attack, type of packet)
- b. **MALWARE CHALLENGE:** Find malware binaries online (or get from us) and do some binary classification (i.e. malware/non-malware, type of malware)

GET CREATIVE!

- Final adjudication based on a 50-50 mix of how interesting the submission is, and how well it works.
- Can work in teams (but only 1 prize)
- **Show-and-tell style presentation tomorrow (friday) lunchtime at the main expo booth.**

signup for updates!

mlsec@cs.stanford.edu



Thank you!

@cchio

cchio@cs.stanford.edu

@antojosep007

antojoseph007@gmail.com