Malware Analysis
Machine Learning Approach

Chiheb Chebbi
TEK-UP University

DeepSec 14-17 Nov 2017 – Vienna, Austria

<whoami/>

- Computer science engineering student @TEK-UP
- Cyber security leadership program fellow @Kaspersky_Lab
- Author / Technical Reviewer @Packt_Publishing UK

- Invited as a speaker to:

Besides Tempa Florida2017, BH Europe 2016,NASA SAC ...

EXPLOIT/SPAM PAYLOAD SUMMARY JAN 2016

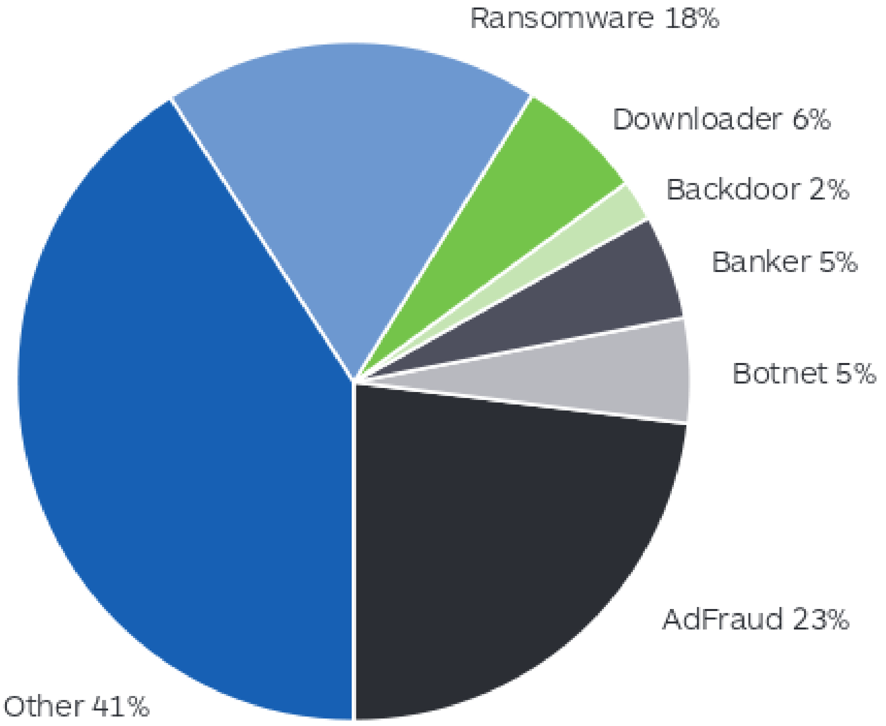


Figure 1. January 2016 payloads.

EXPLOIT/SPAM PAYLOAD SUMMARY NOV 2016

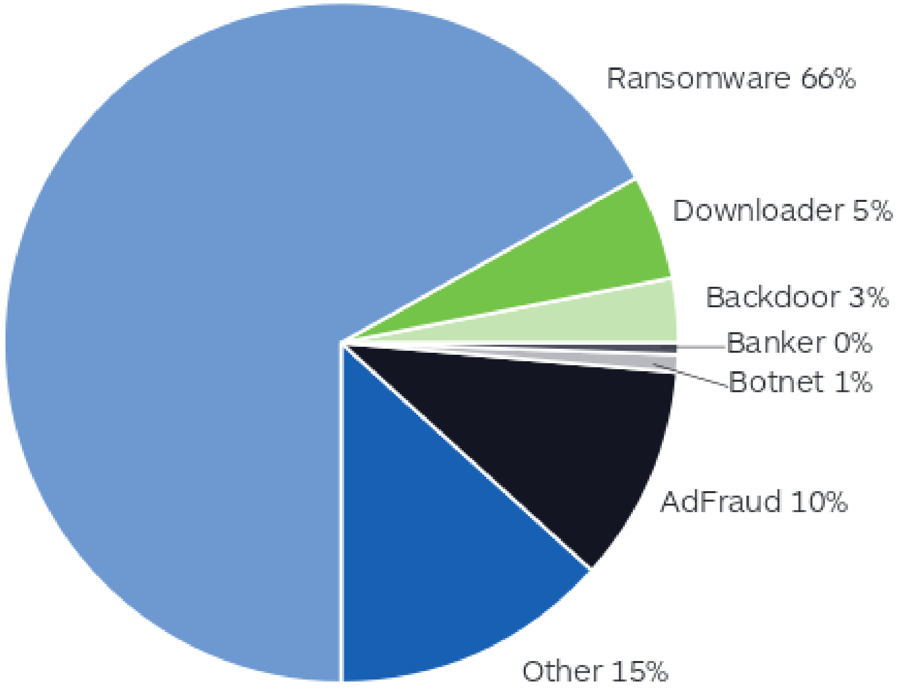
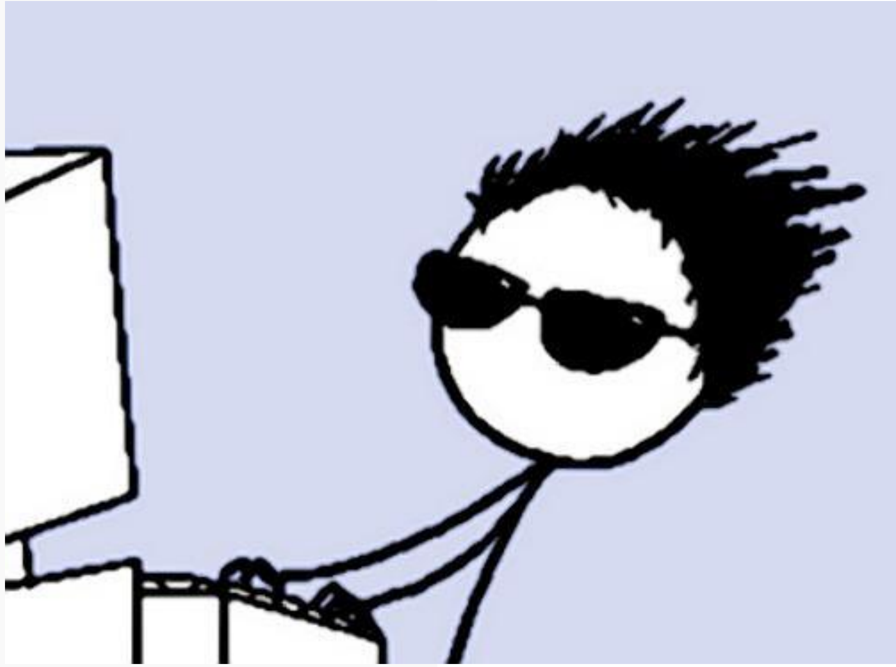


Figure 2. November 2016 payloads.

Source: State of Malware Report 2017- MalwareBytes LABS

Top 10 countries for ransomware detections

1. United States
2. Germany
3. Italy
4. United Kingdom
5. France
6. Australia
7. Canada
8. Spain
9. India
10. Austria



Source: State of Malware Report 2017- MalwareBytes LABS

Ransomware 49 %

Android Malware 31 %

Adware 37 %

Source: State of Malware Report 2017- MalwareBytes LABS

Malware Analysis Techniques

Static Analysis

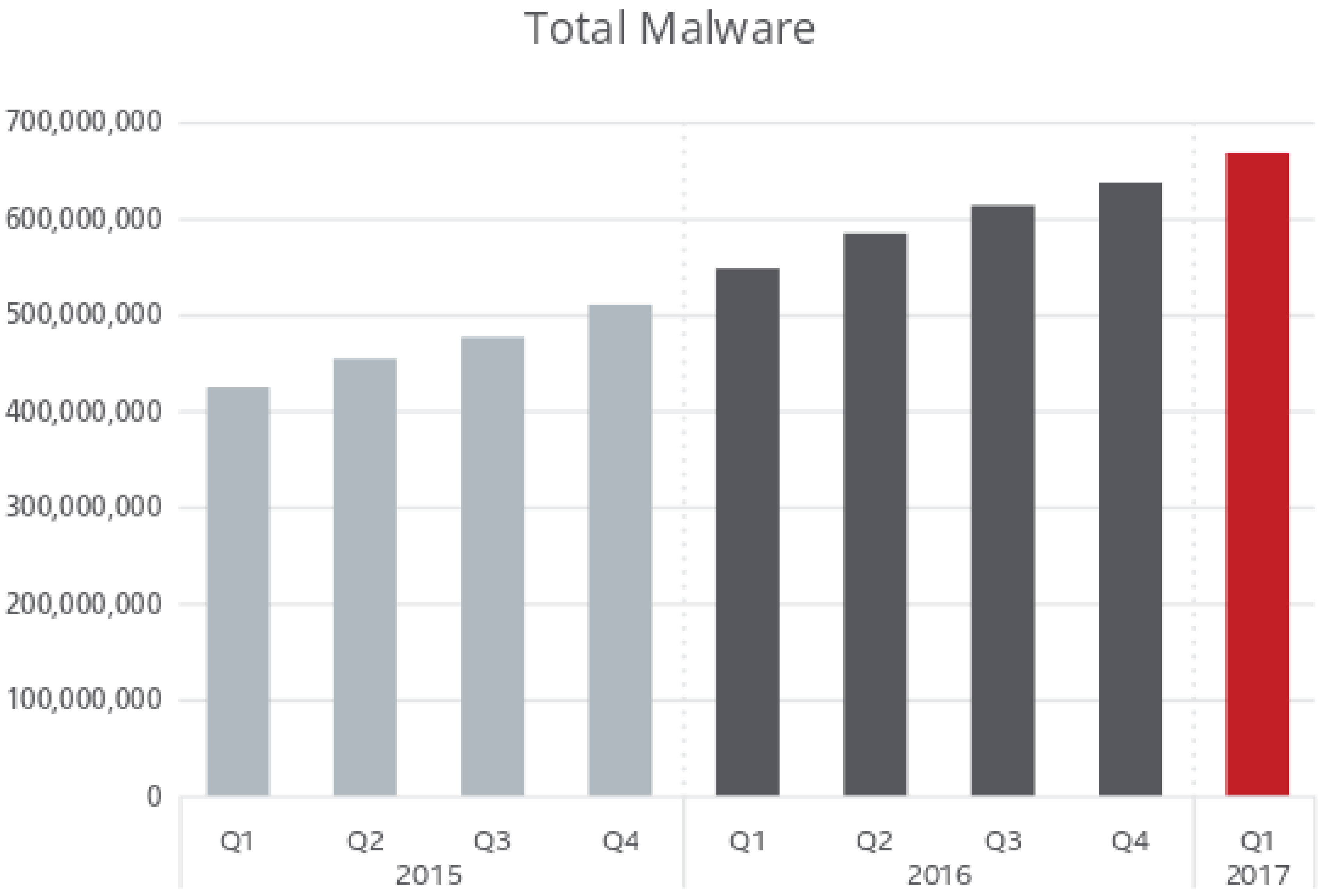
the examination of the malware sample without executing

Dynamic Analysis

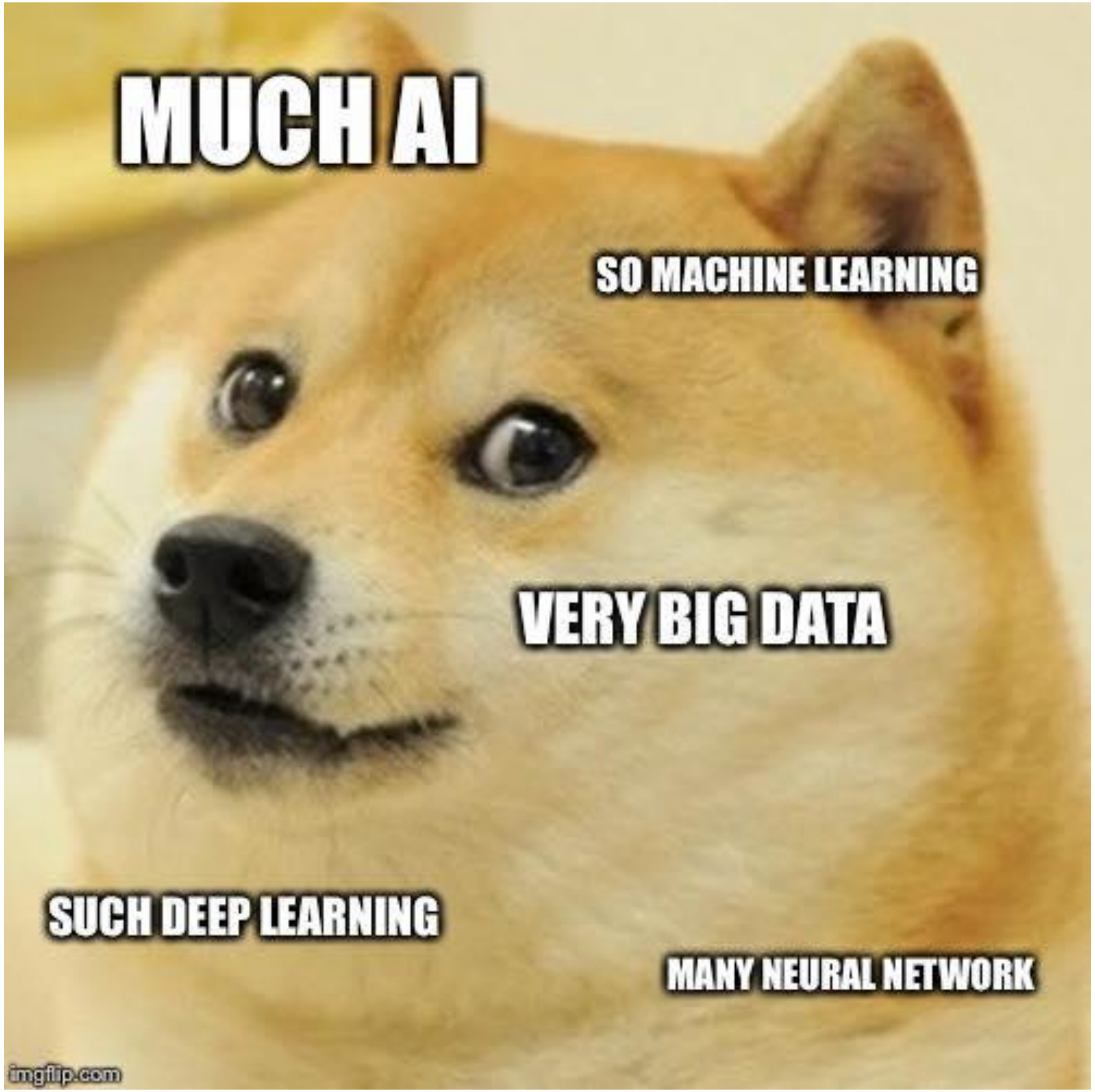
Dynamic analysis techniques track all the malware activities

Memory Analysis

the act of analyzing a dumped memory image from a targeted machine after executing the malware



Source: McAfee Labs, 2017.


MUCH AI
SO MACHINE LEARNING
VERY BIG DATA
SUCH DEEP LEARNING
MANY NEURAL NETWORK
imgflip.com

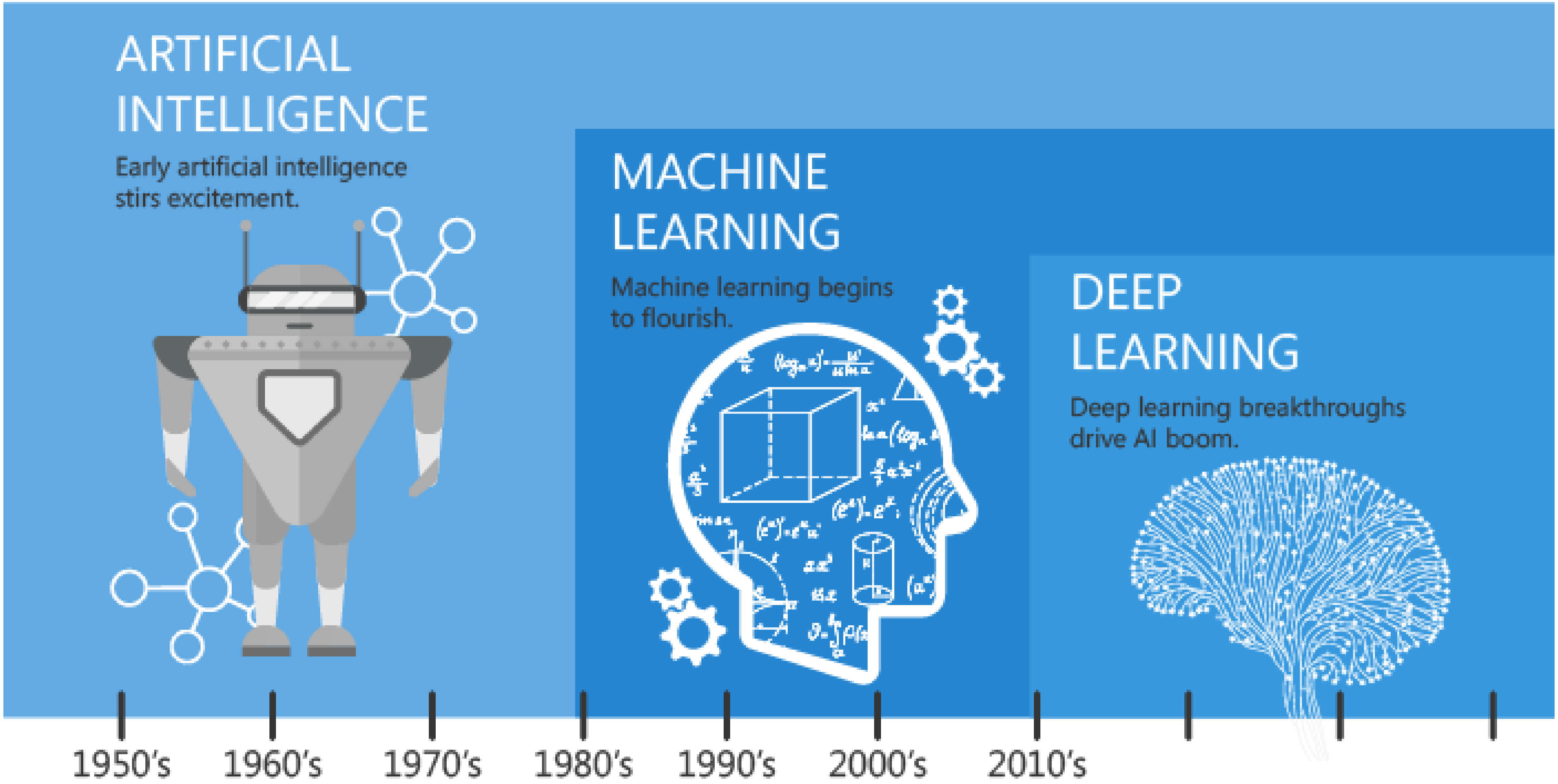
Machine Learning

Artificial Intelligence

Ability to perform tasks normally requiring human intelligence, such as visual perception, speech recognition

Machine Learning

the study and the creation of algorithms that can learn from data and make prediction on them



Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.

Signatures,
Packet Filters

(+) Recognize known threats
(-) Very brittle

Heuristics, Sandboxes,
Stateful Filters

(+) Recognize malicious indicators
(-) Rely on known indicators

Machine
Learning

~~(+) Unstoppable~~
~~(-) None~~

Machine Learning Models

Supervised Learning

we have input variables (I) and an output variable (O) and we need to map the function

Decision Trees, Nave Bayes Classification,
Support Vector Machines

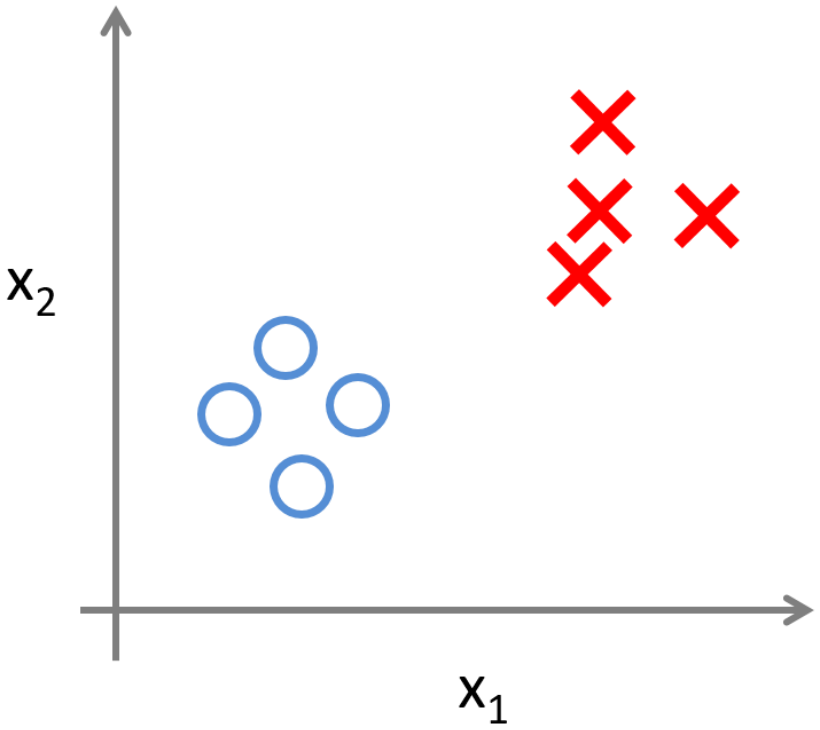
Unsupervised learning

we only have input data (X)

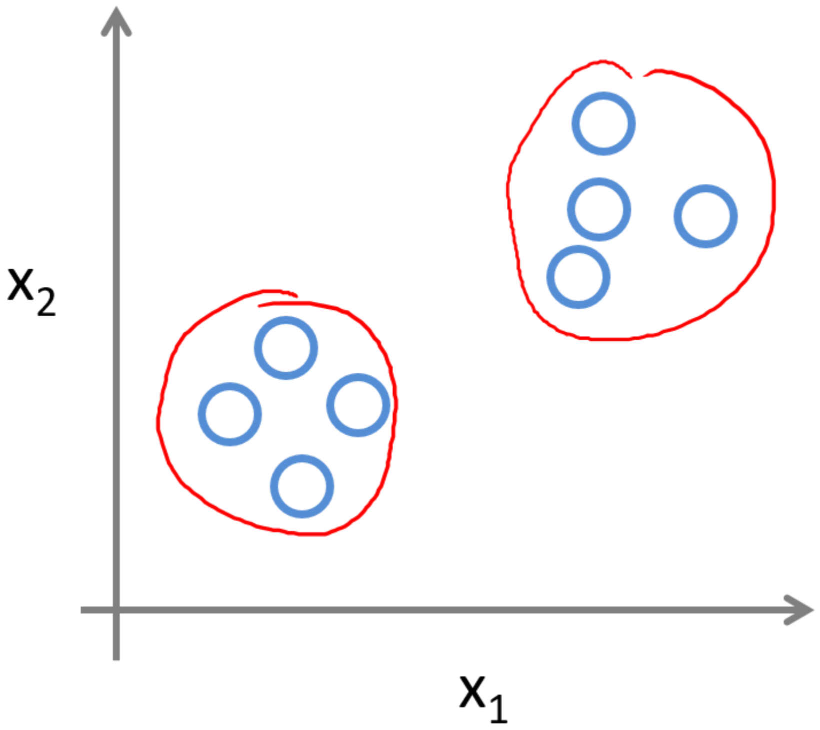
Reinforcement

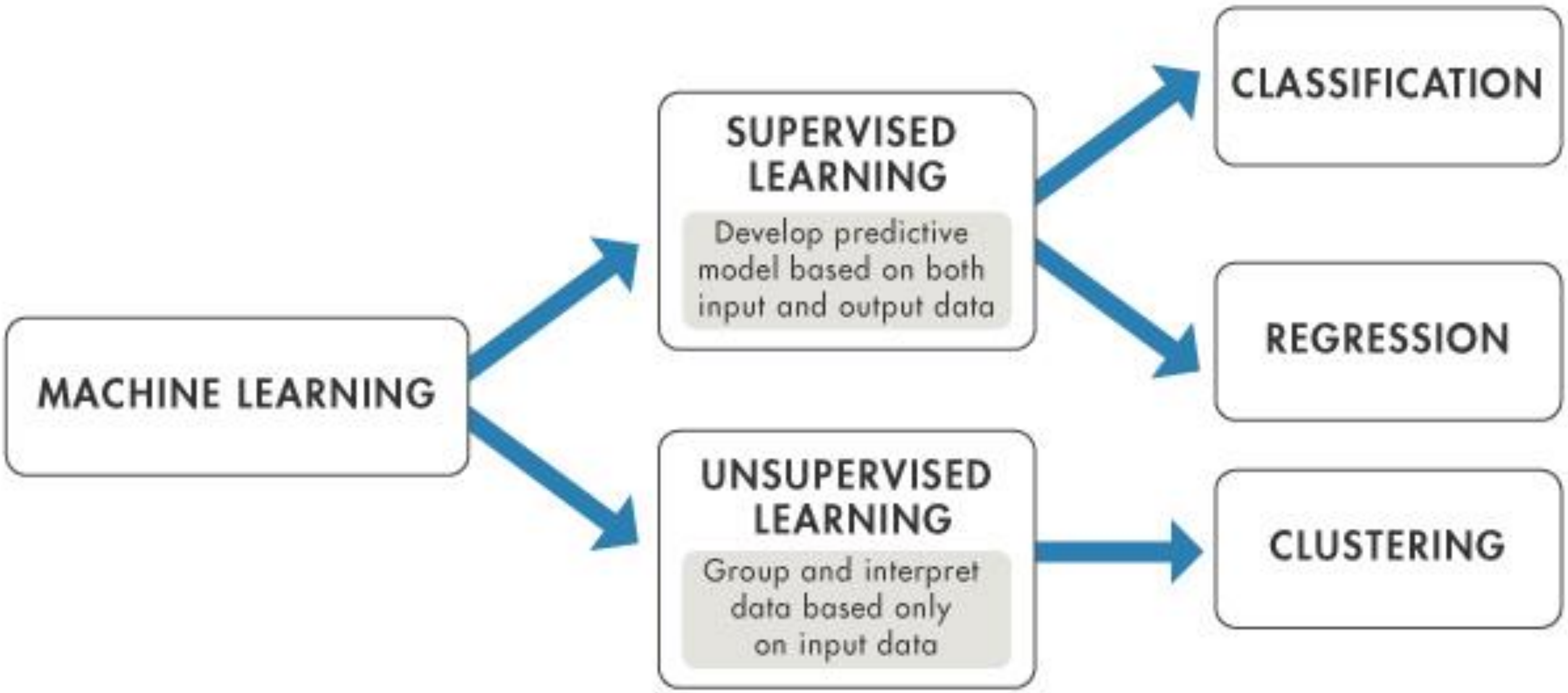
the agent or the system is improving its performance based on a reward function

Supervised Learning



Unsupervised Learning

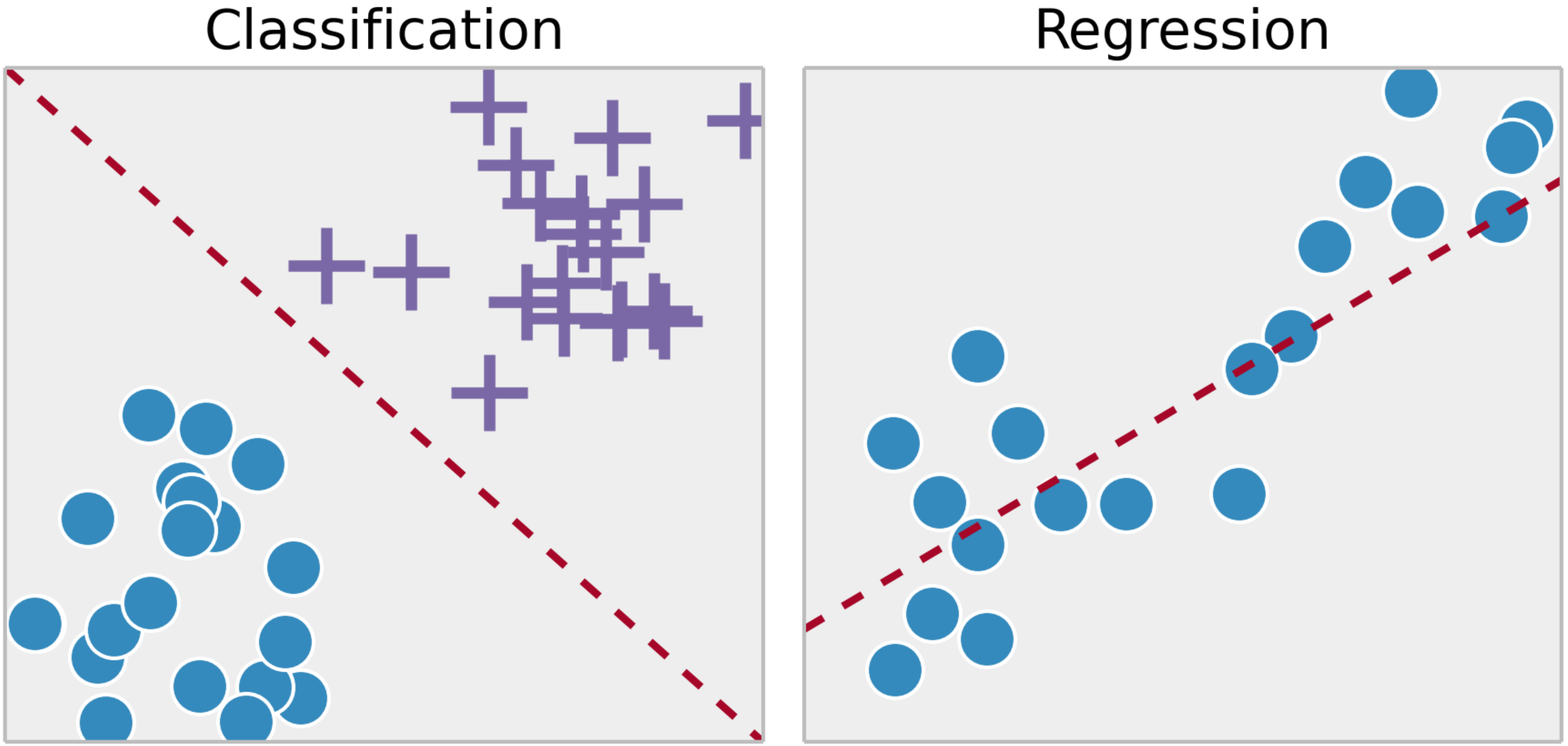



MACHINE LEARNING
SUPERVISED LEARNING
Develop predictive model based on both input and output data
UNSUPERVISED LEARNING
Group and interpret data based only on input data
CLASSIFICATION
REGRESSION
CLUSTERING

Classification vs Regression

- Classification means to group the output into a class.
- classification to **predict** the type of tumor i.e. harmful or not harmful using training data
- if it is discrete/categorical variable, then it is classification problem

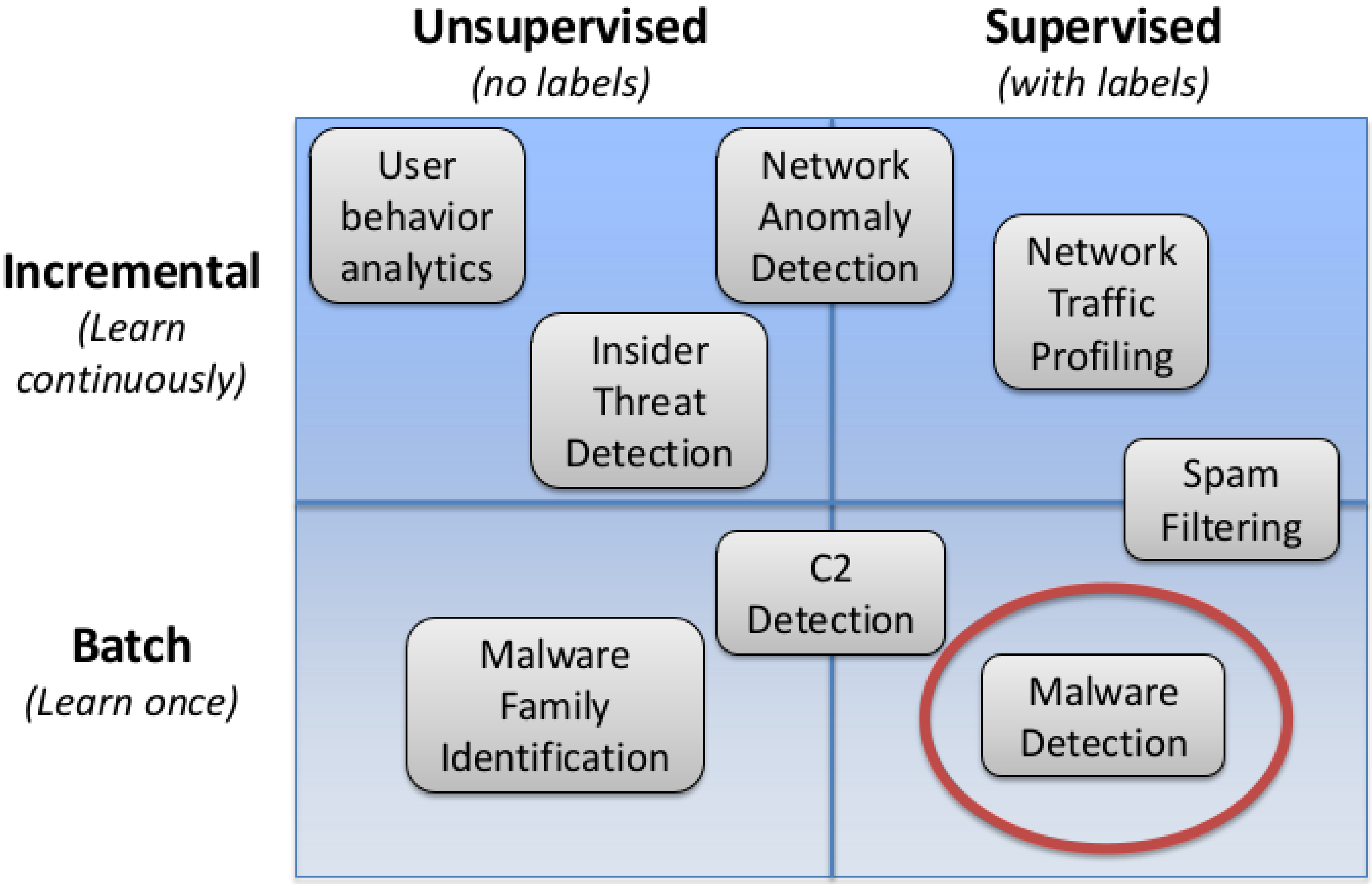
- Regression means to predict the output value using training data.
- regression to **predict** the house price from training data
- if it is a real number/continuous, then it is regression problem.


Classification
Regression

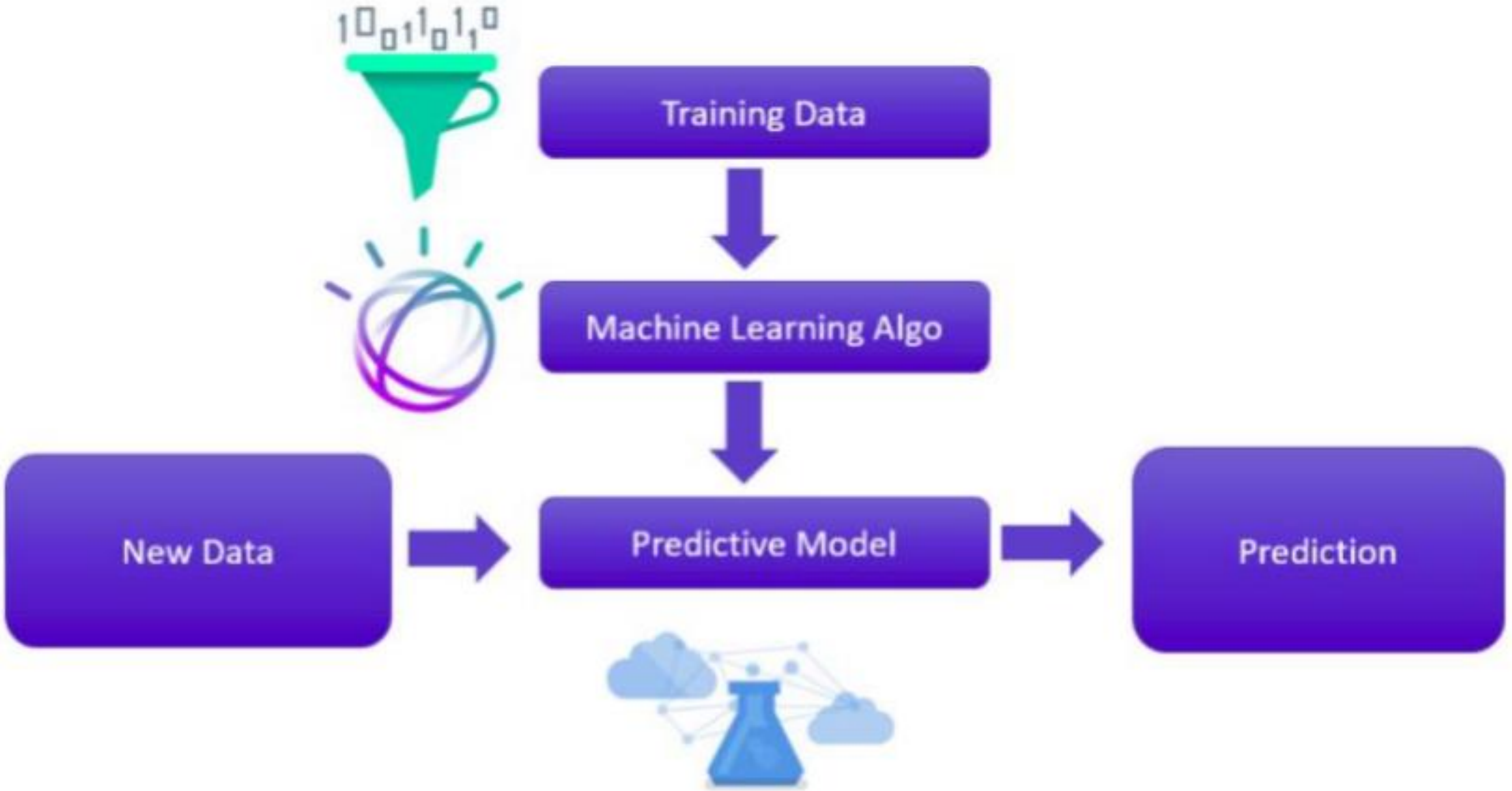

Hot Dog
OR
Not Hot Dog

Machine Learning Algorithms

	Unsupervised	Supervised
Continuous	• Clustering & Dimensionality Reduction ○ SVD ○ PCA ○ K-means	• Regression ○ Linear ○ Polynomial • Decision Trees • Random Forests
Categorical	• Association Analysis ○ Apriori ○ FP-Growth • Hidden Markov Model	• Classification ○ KNN ○ Trees ○ Logistic Regression ○ Naive-Bayes ○ SVM


Unsupervised
(no labels)
Supervised
(with labels)
Incremental
(Learn continuously)
Batch
(Learn once)
User behavior analytics
Network Anomaly Detection
Network Traffic Profiling
Insider Threat Detection
Spam Filtering
C2 Detection
Malware Family Identification
Malware Detection

Machine Learning Workflow



Malware Datasets

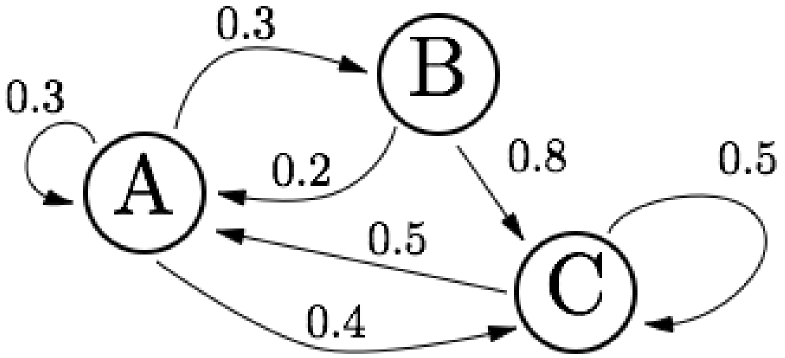
Malware Analysis Process Entry Points:

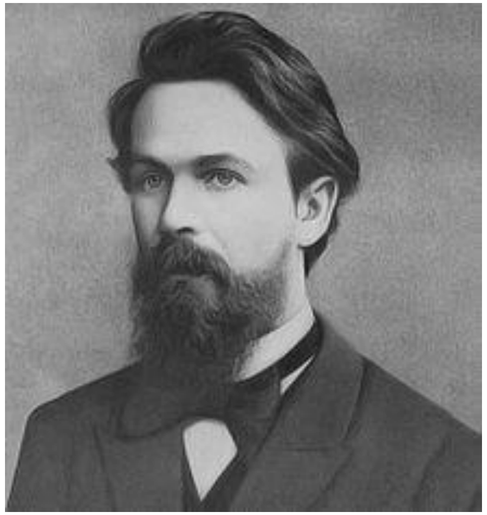
- File
- URL
- PCAP
- Memory Image

Hidden Markov Models

Markov process or what we call a **Markov chain** is a stochastic model used for any random system that change its states according to fixed probabilities

In probability theory and related fields, a stochastic or random process is a mathematical object usually defined as a collection of random variables





Hidden Markov Models

- The Hidden Markov Model is a Markov Process where we are unable to directly observe the state of the system.

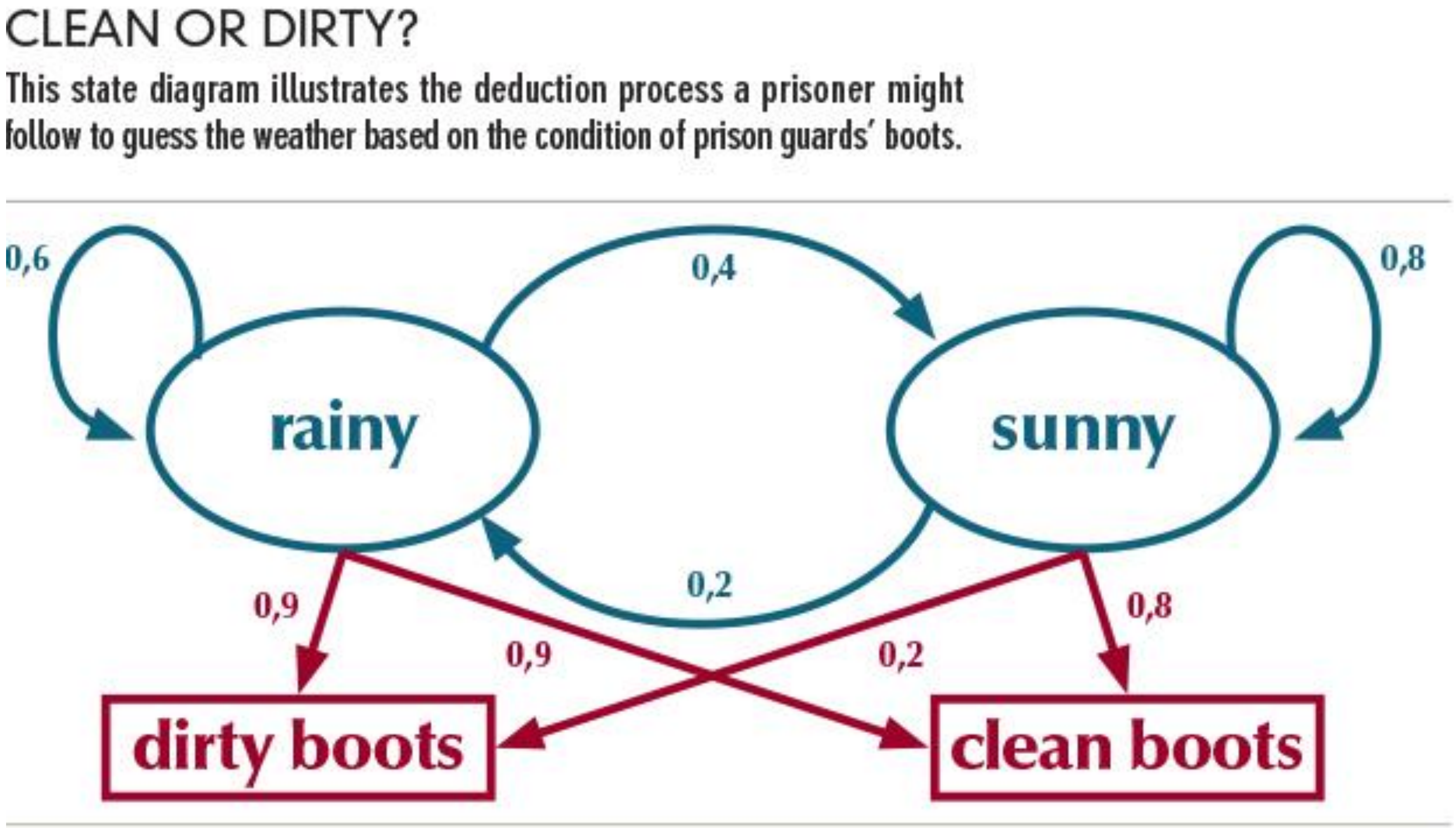
Each state has a fixed probability of ”emitting”.

p is a sequence of states (AKA a path).

Each p i takes a value from set Q.

We do not observe p

Hidden Markov Models



A General Definition of HMM

$$HMM = (S, V, B, A, \Pi)$$

N states

$$S = \{s_1, ..., s_N\}$$

M symbols

$$V = \{v_1, ..., v_M\}$$

Initial state probability:

$$\Pi = \{\pi_1, ..., \pi_N\} \quad \sum_{i=1}^{N} \pi_i = 1$$

π_i : *prob of starting at state* s_i

State transition probability:

$$A = \{a_{ij}\} \quad 1 \le i, j \le N \quad \sum_{j=1}^{N} a_{ij} = 1$$

a_{ij} : *prob of going* $s_i \rightarrow s_j$

Output probability:

$$B = \{b_i(v_k)\} \quad 1 \le i \le N, 1 \le k \le M \quad \sum_{k=1}^{M} b_i(v_k) = 1$$

$b_i(v_k)$: *prob of "generating"* v_k *at* s_i

Hidden Markov Models

Components of Hidden Markov model

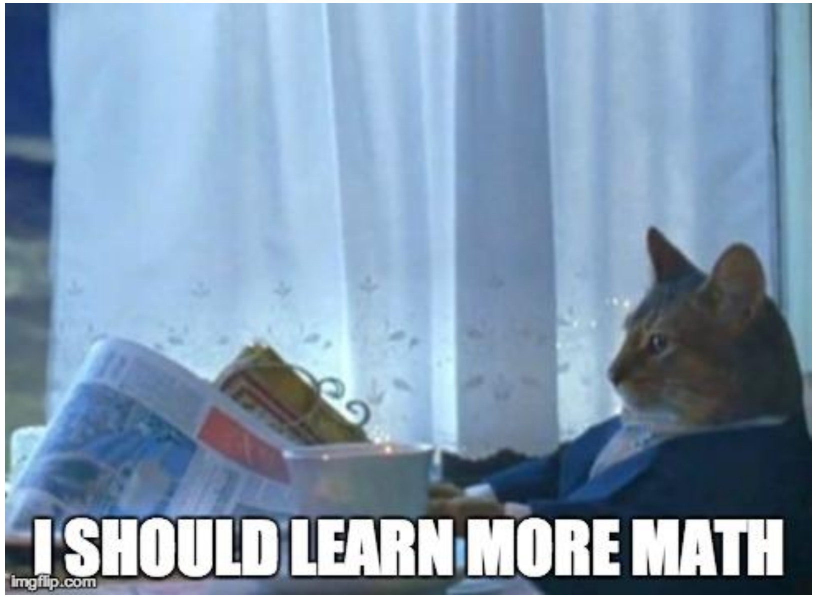
Notations:
T = length of the observation sequence
N = number of states in the model
M = number of observation symbols
Q = $\{q^0, q^1, ..., q^{n-1}\}$ = distinct states of the Markov process
V = state transition probabilities NxN matrix
B = observation probability MxN matrix
Π = initial state distribution
O = $O^1, O^2, ..., O^{T-1}$ = observation sequence.
λ = (A, B, π) = A Hidden Markov Model defined by the tuple (A, B, π)

Classic Problems of Hidden Markov Model

- Problem 1: State Estimation Given a model $\lambda = (A, B, \Pi)$ and an observation sequence O, we need to find $P(O—\lambda)$.That is to determine the likelihood and check the wellness of the given model.
- Problem 2: Decoding or Most Probable Path (MPP): Given a model $\lambda = (A, B, \Pi)$ and ,and an observation sequence O, to determine the optimal state sequence Q for the given model
- Problem 3: Training/Learning HMM: Given O, N, M, we can find a model that maximizes probability of O and learn the two HMM parameters A and B.

Solutions

- Forward-Backward technique
- Viterbi Decoding technique
- Baum-Welch (Expectation Maximization) technique



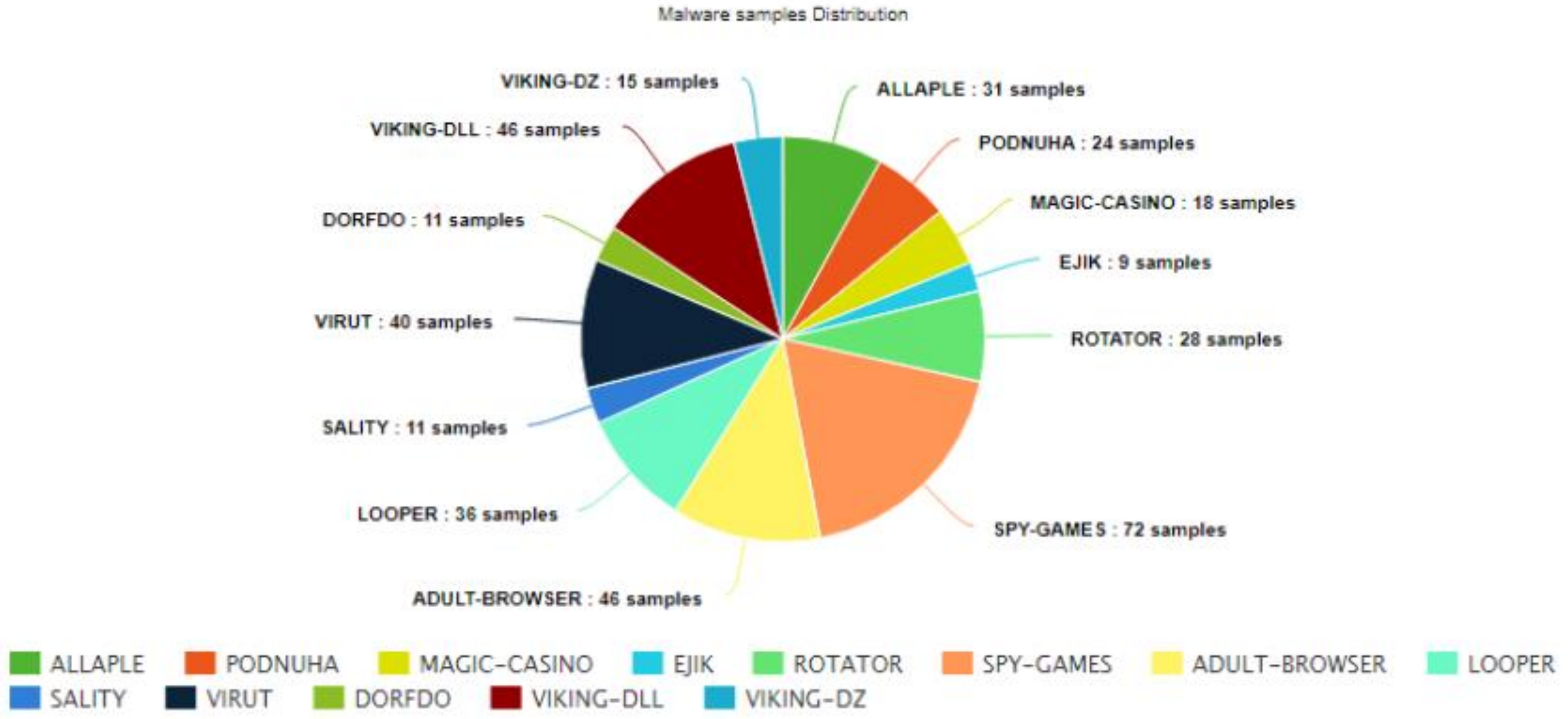
Profile Hidden Markov Model

- By definition a **profile** is a pattern of conservation.

The Profile Hidden Markov Model is a probabilistic approach that was developed specially for modeling sequence similarity occurring in biological sequences such as proteins and DNA.

- Profile HMM is a modified implementation of HMM.

- HMMER is an open source implementation of Profile Hidden Markov Models. It is basically built to build HMM models for protein sequences and alignment but in our case we are going to adopt it to build models for malware behaviour sequences.

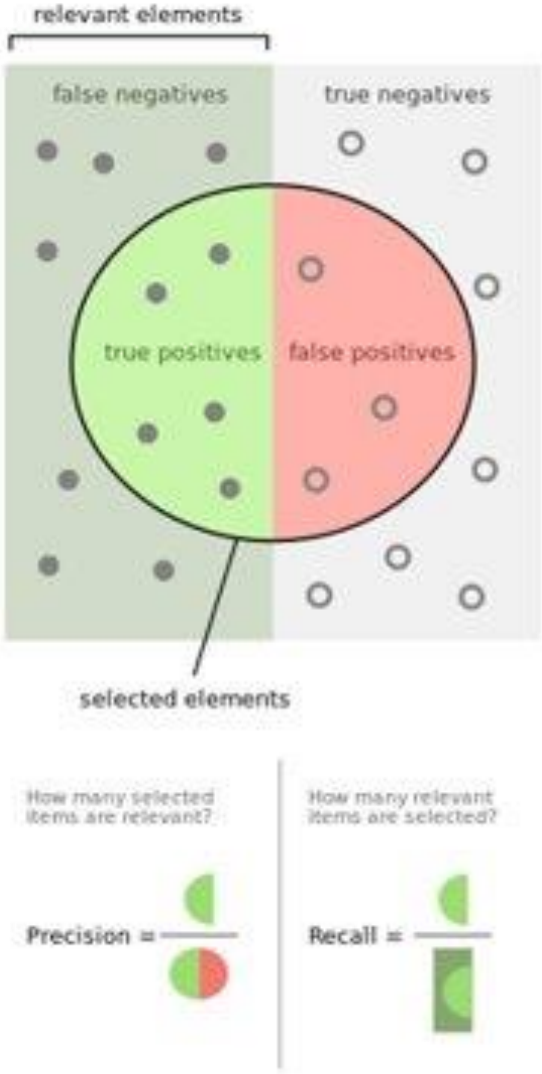

Malware samples Distribution
VIKING-DZ : 15 samples
ALLAPLE : 31 samples
VIKING-DLL : 46 samples
PODNUHA : 24 samples
MAGIC-CASINO : 18 samples
DORFDO : 11 samples
EJIK : 9 samples
VIRUT : 40 samples
ROTATOR : 28 samples
SALITY : 11 samples
LOOPER : 36 samples
SPY-GAMES : 72 samples
ADULT-BROWSER : 46 samples
ALLAPLE
PODNUHA
MAGIC-CASINO
EJIK
ROTATOR
SPY-GAMES
ADULT-BROWSER
LOOPER
SALITY
VIRUT
DORFDO
VIKING-DLL
VIKING-DZ

Machine learning Model Evaluation Metrics

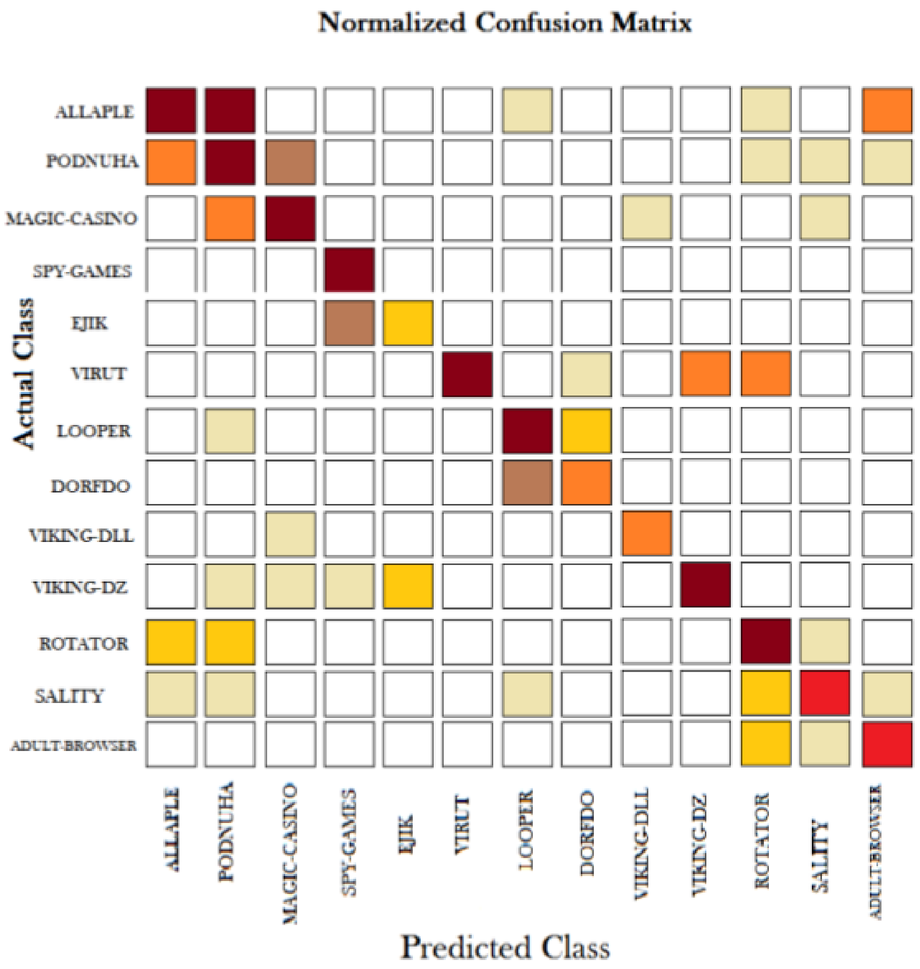
tp = True Positive
fp= False Positive
tn = True Negative
fn = False Negative

True positive	False positive (Type I error)
False negative (Type II error)	True negative

Confusion Matrix



So what happened?



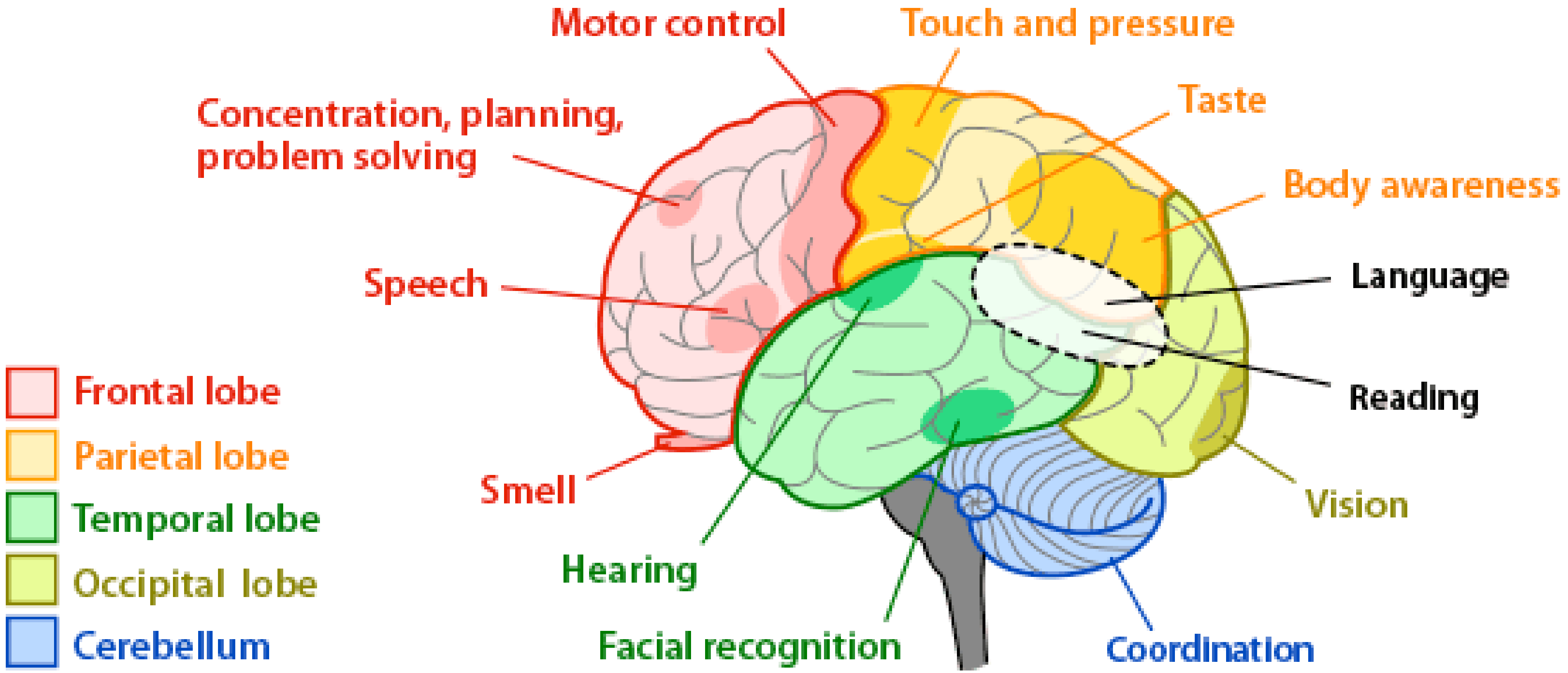
Low Detection Rate :'(


Deep Learning
Hidden Markov Models

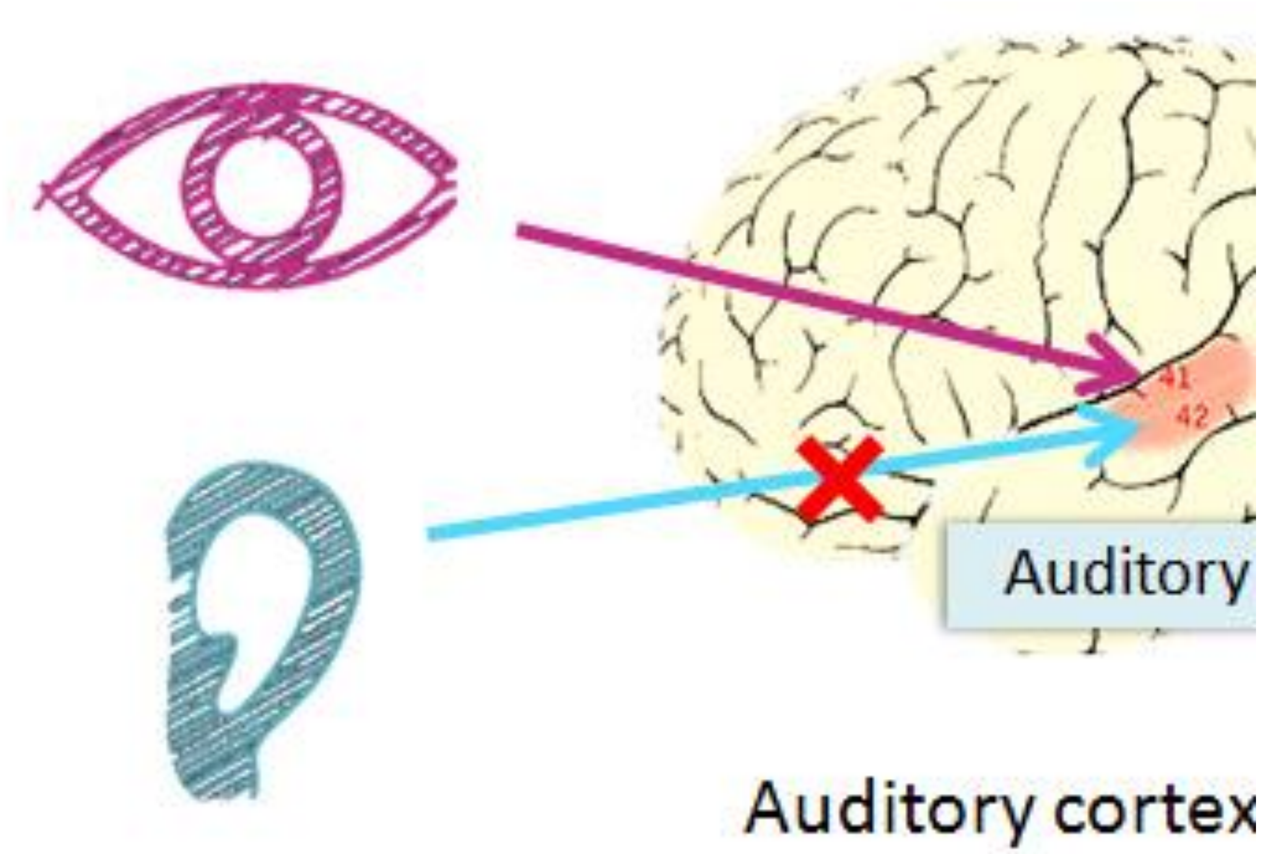
"SCIENCE FIRST"

DEEPSEC


“I'm going to have to science the s*** out of this”


Motor control
Touch and pressure
Concentration, planning, problem solving
Taste
Body awareness
Speech
Language
Reading
Frontal lobe
Parietal lobe
Temporal lobe
Occipital lobe
Cerebellum
Smell
Vision
Hearing
Facial recognition
Coordination

One Algorithm Hypothesis



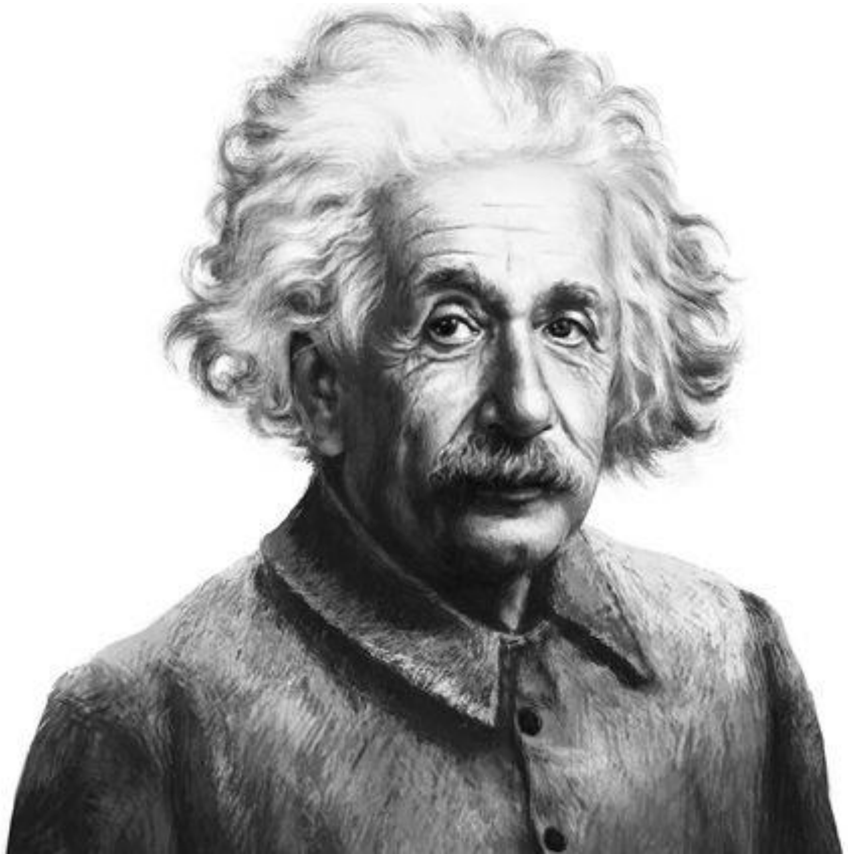
- There is some evidence that the human brain uses essentially the same algorithm to understand many different input modalities.

- Ferret experiments, in which the "input" for vision was plugged into auditory part of brain, and the auditory cortex learns to "see."
[Roe et al., 1992]

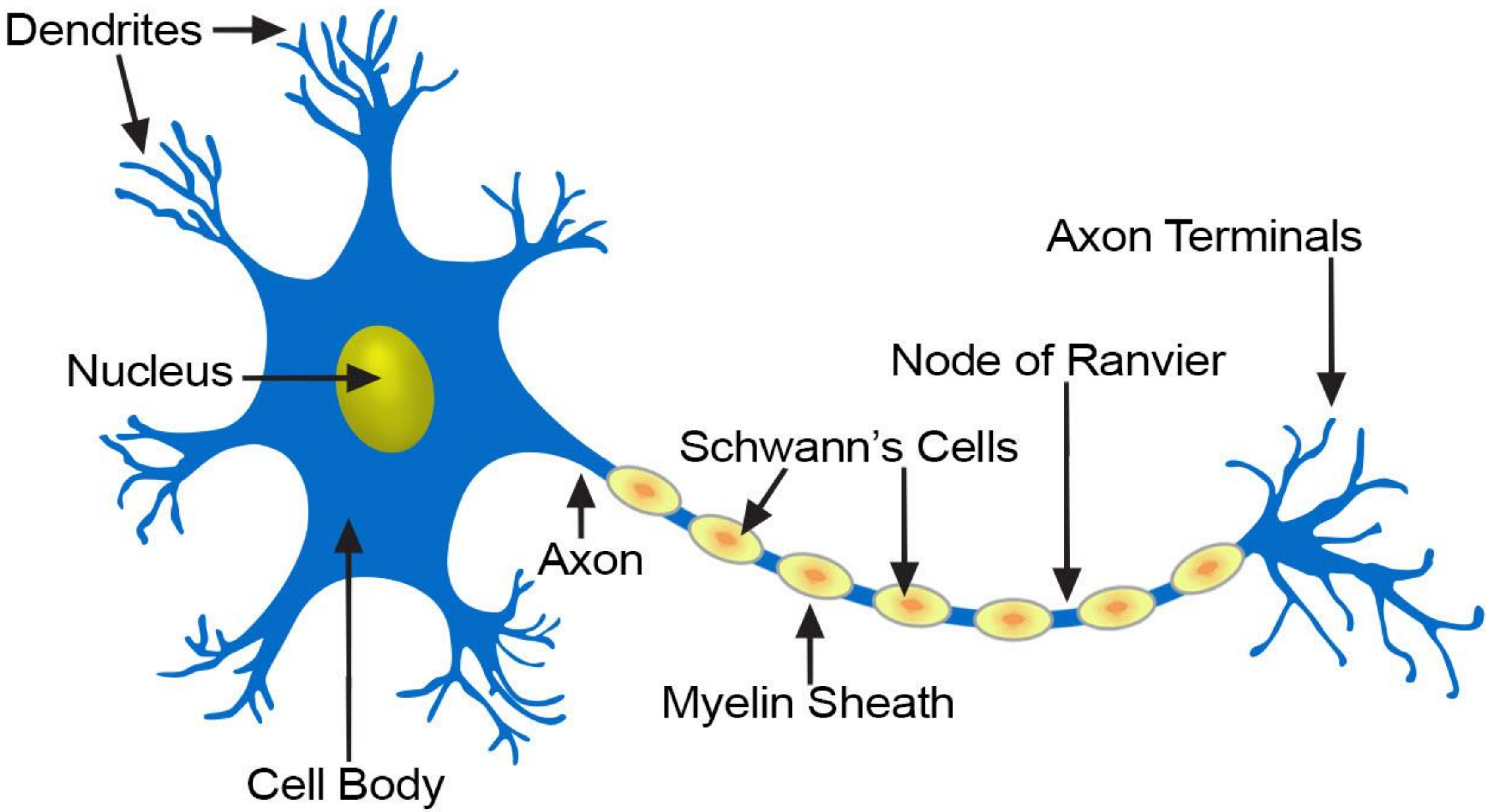
“Look deep into nature,

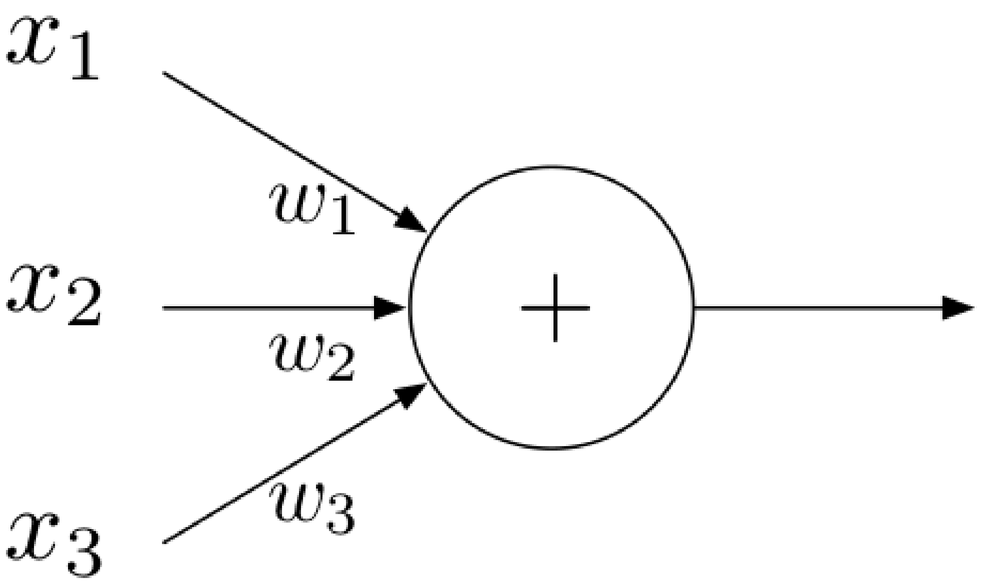
and then you will understand everything better.”

Albert Einstein



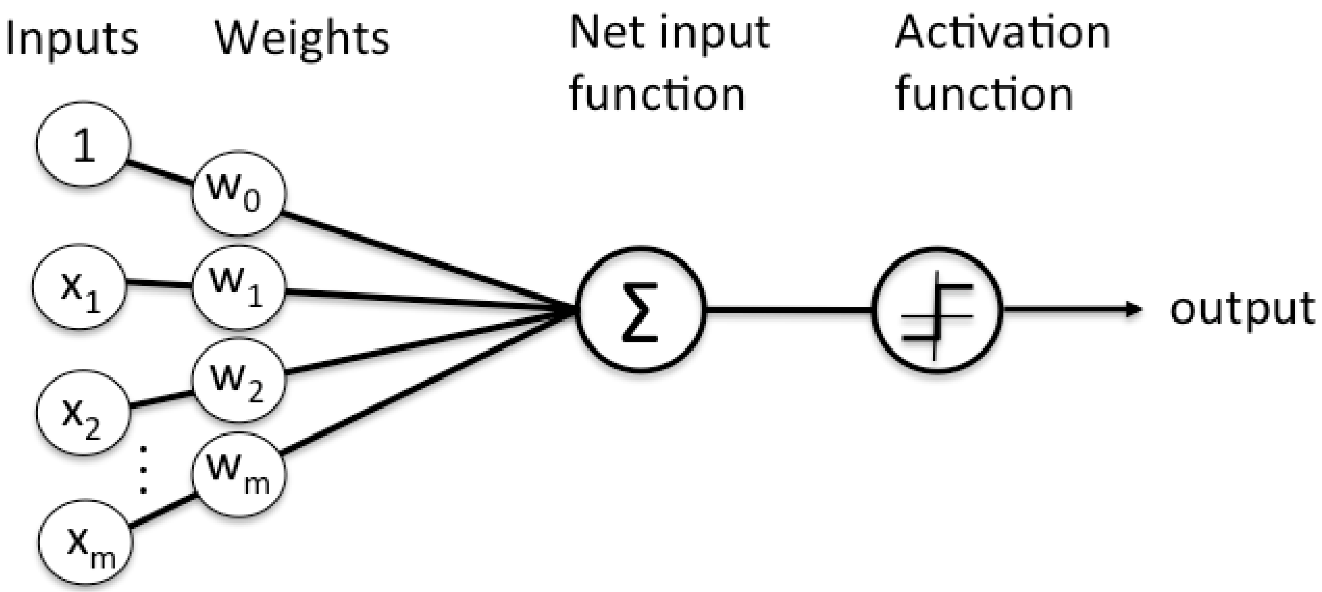
Structure of a Typical Neuron

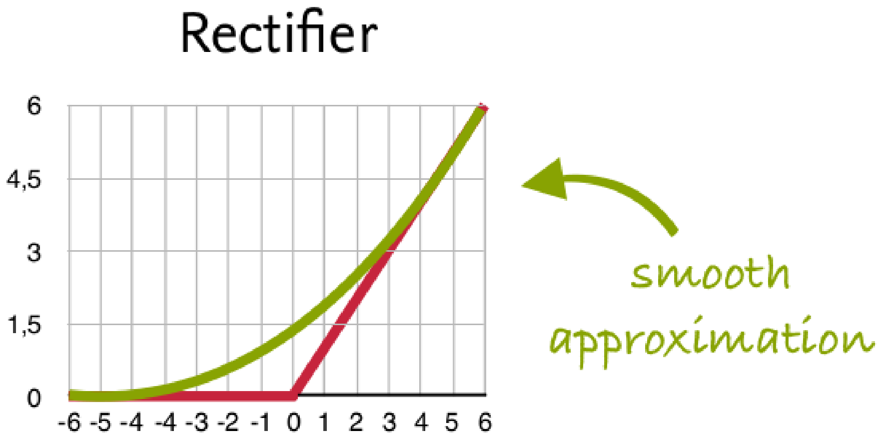




$$h(x) = \begin{cases} 0 & \text{if } \sum_i w_i x_i \leq threshold \\ 1 & \text{if } \sum_i w_i x_i > threshold \end{cases}$$

- The artificial model of a neuron is called perceptron

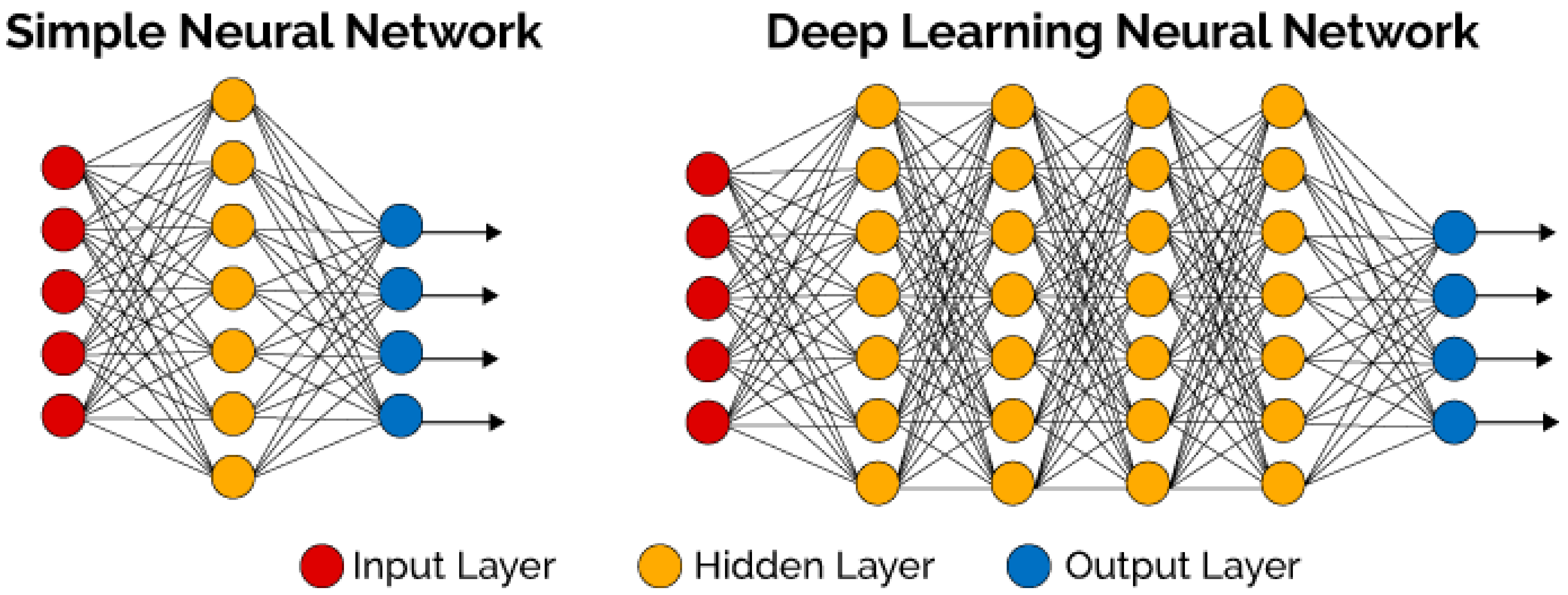


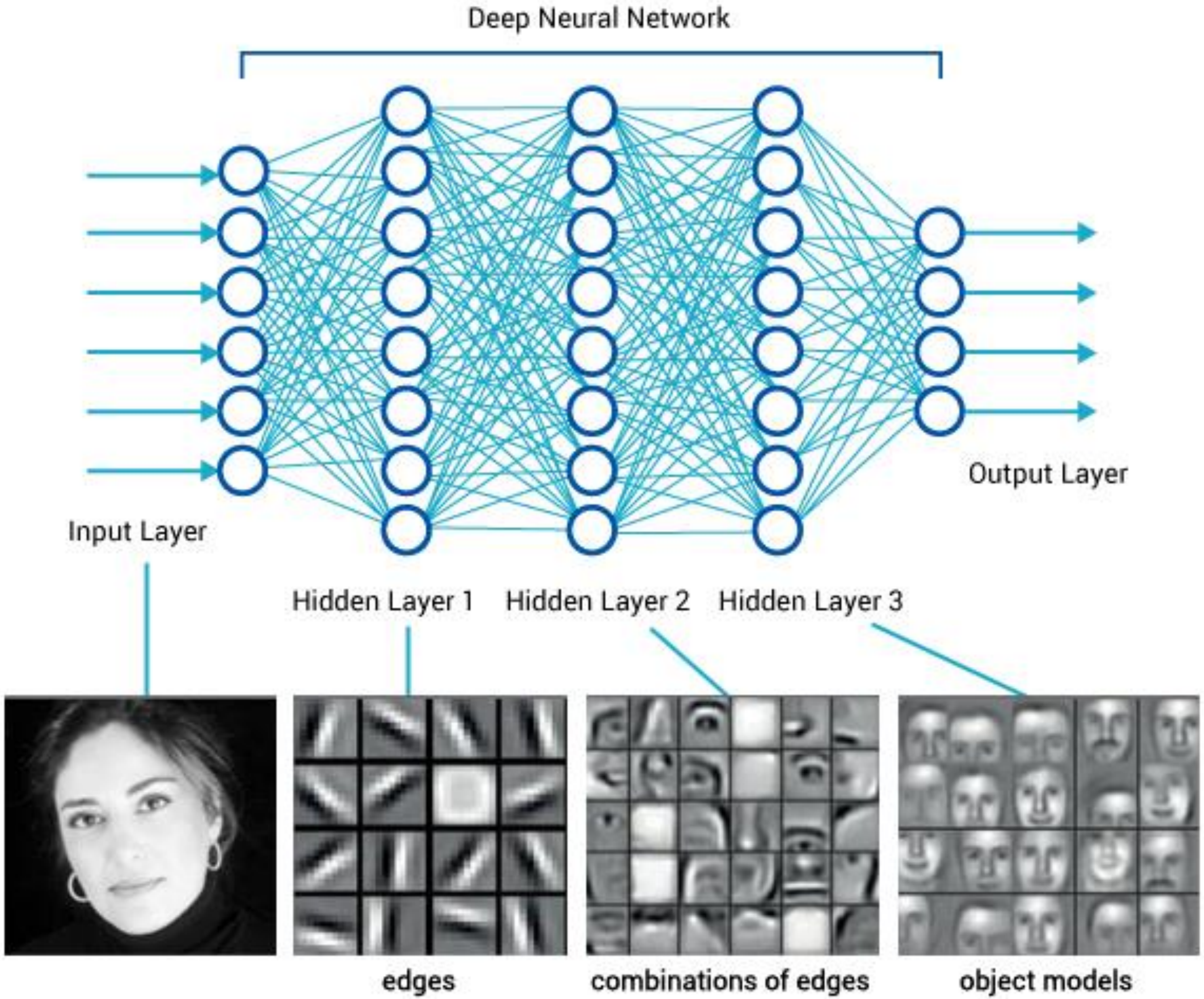


$$h(x) = max(0, x)$$

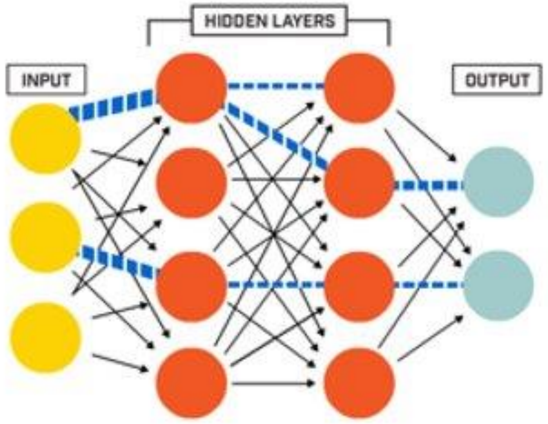
$$h(x) = ln(1 + e^x)$$

Schematic of Rosenblatt's perceptron.


Simple Neural Network
Deep Learning Neural Network
Input Layer
Hidden Layer
Output Layer


Deep Neural Network
Input Layer
Hidden Layer 1
Hidden Layer 2
Hidden Layer 3
Output Layer
edges
combinations of edges
object models

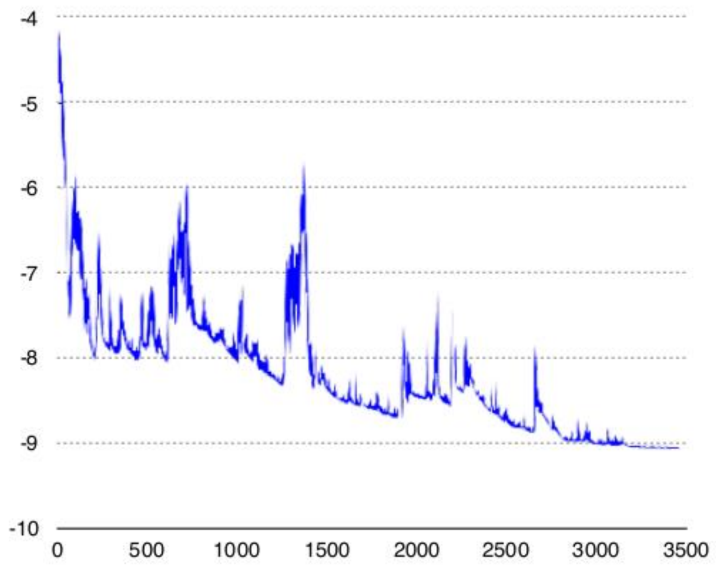
Backpropagation



Backpropagation is the process of trying to keep the error as down as possible.

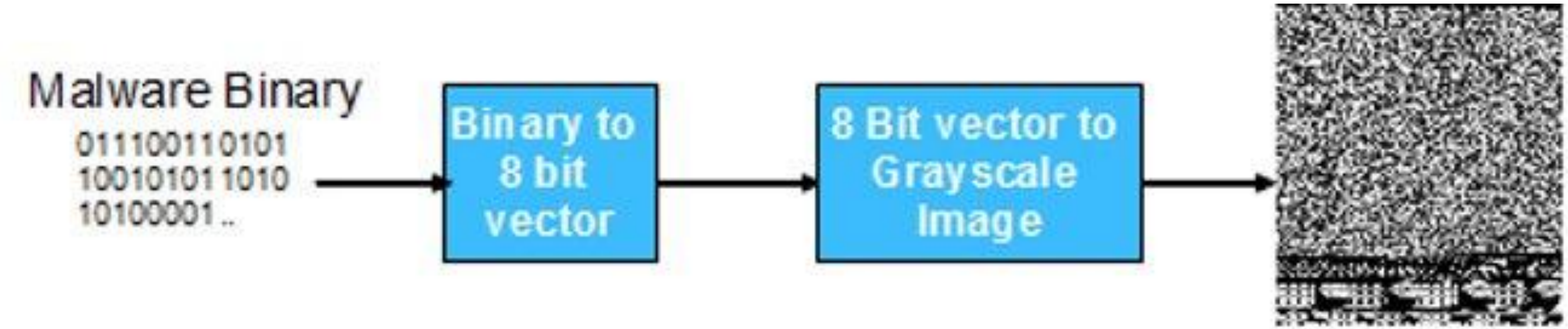
Stochastic Gradient Descent

$$w = w - \eta \frac{\partial E(w)}{\partial w_i}$$



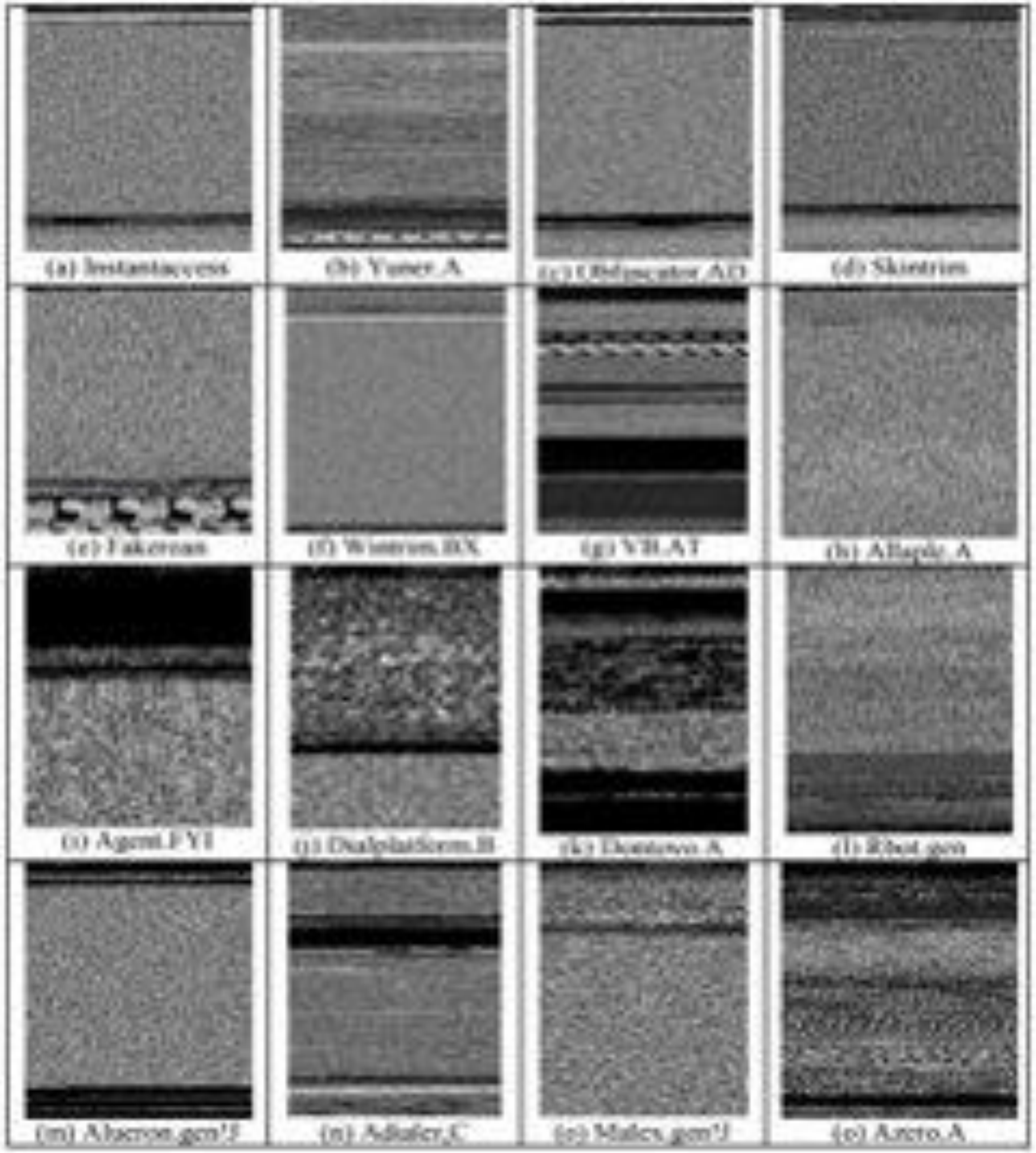
Microsoft Malware Classification Challenge (BIG 2015)

10K Malware 500 GB



You are provided with a set of known malware files representing a mix of 9 different families. Each malware file has an Id, a 20 character hash value uniquely identifying the file, and a Class, an integer representing one of 9 family names to which the malware may belong:

1. Ramnit
2. Lollipop
3. Kelihos_ver3
4. Vundo
5. Simda
6. Tracur
7. Kelihos_ver1
8. Obfuscator.ACY
9. Gatak



- Accurately detects malware at > 90%

Well documented and open source frameworks

Why GPU Matters in Deep Learning?

```
X_train shape: (50000, 3, 32, 32)
50000 train samples
10000 test samples
Using real-time data augmentation.
Epoch 1/200
50000/50000 [==============================] - 734s
Epoch 2/200
50000/50000 [==============================] - 733s
Epoch 3/200
50000/50000 [==============================] - 733s
Epoch 4/200
50000/50000 [==============================] - 733s
```

Running time **without** GPU

VS

```
X_train shape: (50000, 3, 32, 32)
50000 train samples
10000 test samples
Using real-time data augmentation.
Epoch 1/200
50000/50000 [==============================] - 27s
Epoch 2/200
50000/50000 [==============================] - 27s
Epoch 3/200
50000/50000 [==============================] - 27s
Epoch 4/200
50000/50000 [==============================] - 27s
```

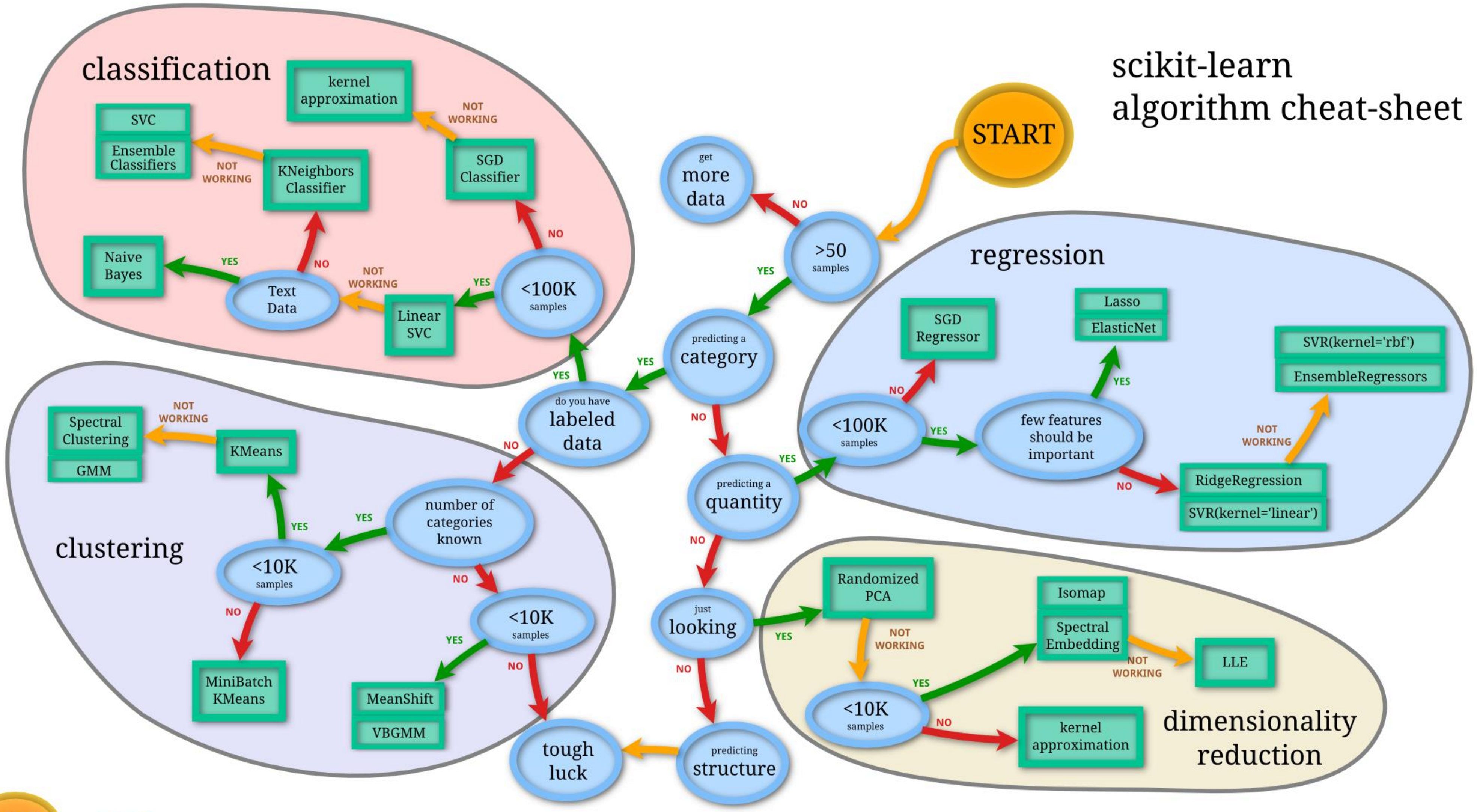
Running time **with** GPU

With GPU, the running time is 733/27=**27.1 times faster** then the running time without GPU!!!

Deep learning life-cycle

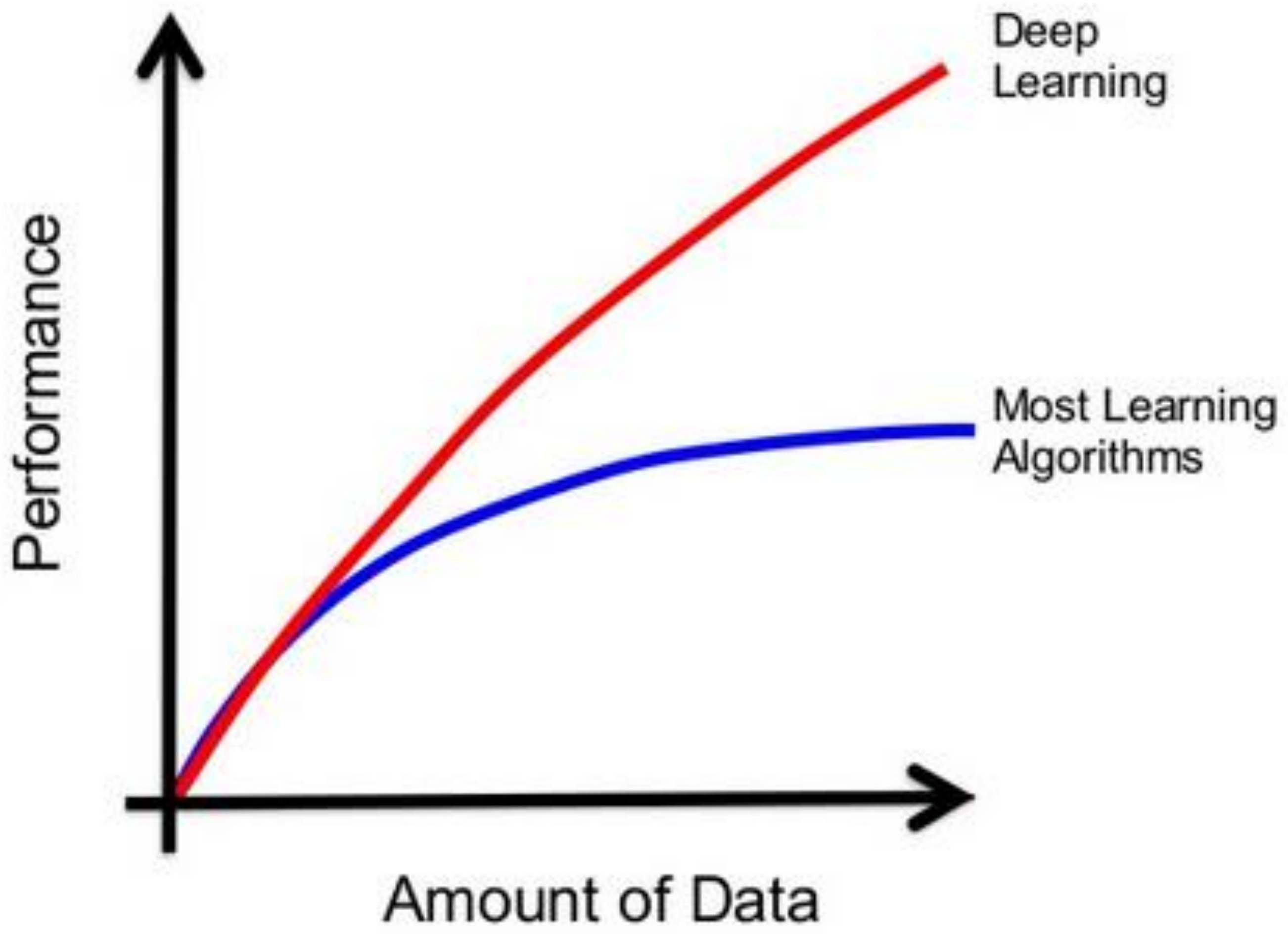
- Network Definition
- Network Compiling
- Network Fitting
- Network Evaluation
- Prediction

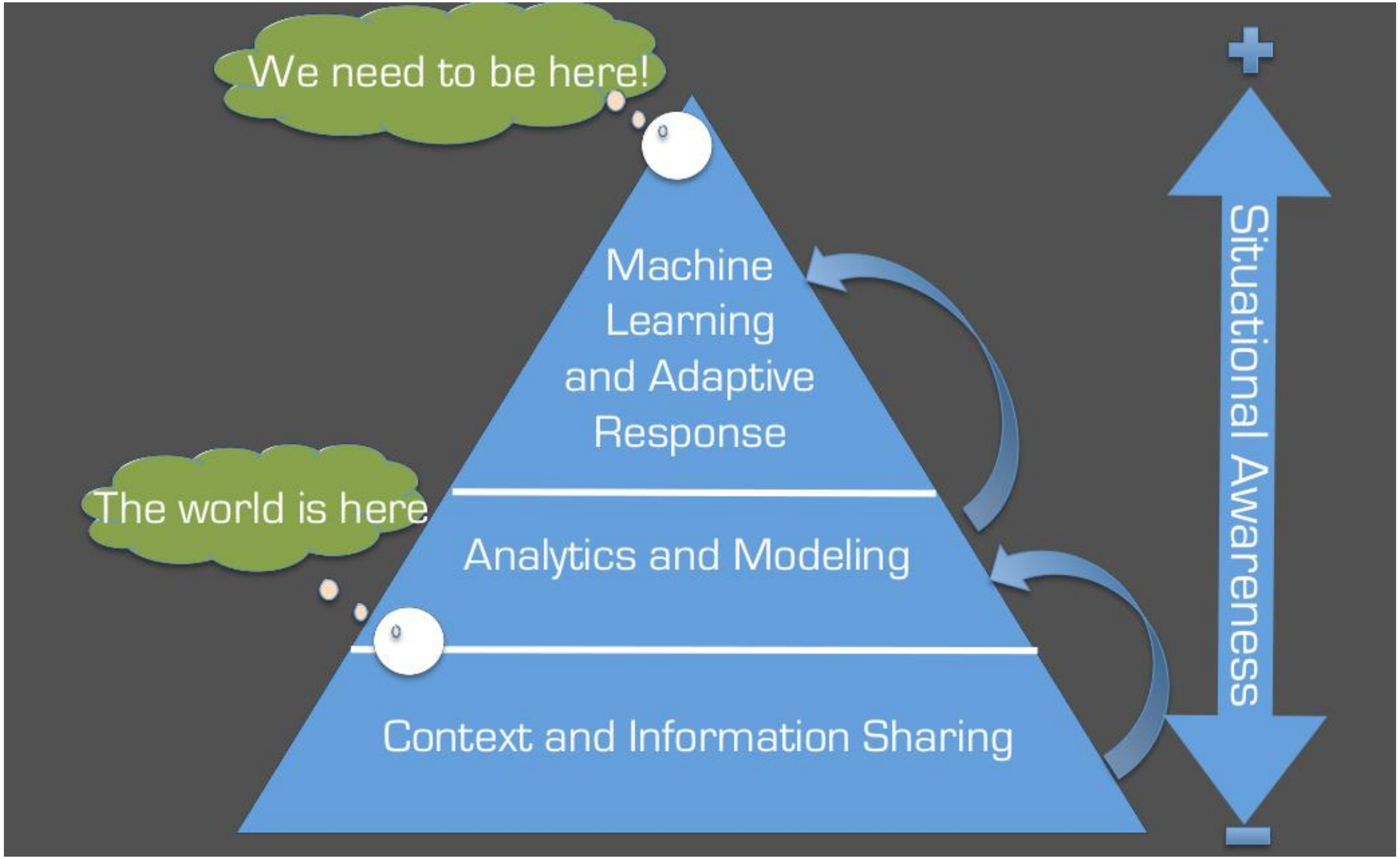
scikit-learn algorithm cheat-sheet



Back

Machine Learning vs Deep Learning





Gartner report: “Intelligent and Automated Security Controls Impact the Future of the Security Market”, Oct 2015

- **Machine learning** in cybersecurity will enormously booster spending in big data, intelligence and analytics, reaching as much as **$96 billion** (£71.9 billion) by 2021.

References

[1] Defeating Machine Learning What Your Security Vendor is Not Telling You – Blackhat USA 2015
[2] Deep Learning for Malware Analysis Machine Learning for Computer Security Hugo Gascón
[3] State of the art MalwareBytes Report 2017
[4] Deep Machine Learning Meets Cybersecurity
[5] How to build a malware classifier [that doesn't suck on real-world data]

Q&A

Chiheb-chebbi@outlook.fr
Chiheb.chebbi@tek-up.de
Hello@chihebchebbi.tn