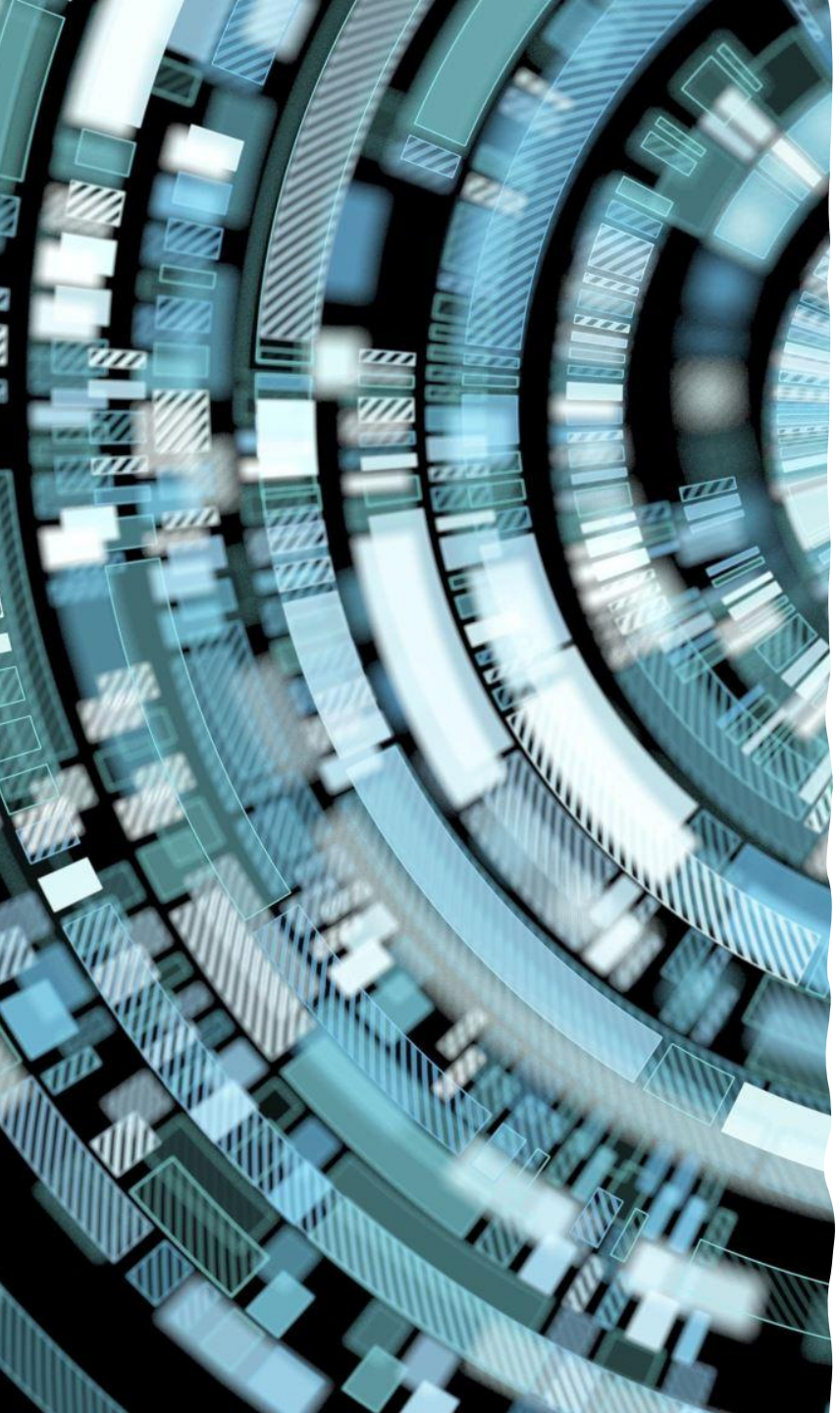


DO YOU TRUST YOUR ARTIFICIAL INTELLIGENCE SYSTEM?

Dr. Robson de Oliveira Albuquerque
Latitude/UnB

robson@redes.unb.br

Disclaimer

The content presented herein represents my thoughts about the subject of this lecture.

It does not necessarily represent the vision of my employers nor the University of which I am a researcher;

Images used in this presentation are the property and credit of the respective creators and sources;

Agenda

- Trust;
- Artificial Intelligence;
- Attacks on AI systems;
- Discussions;

Trust

- Human social aspect:
  - To believe that someone is good and honest and shall not harm you;
  - Something that is safe and trustable;
  - Depends on the context:
    - A trusts B to…;
    - A doesn't trust B to…;

TRUST

In computing systems:

- Trust is relative to a particular context;

A shall trust B to get a ride, but A doesn't trust B to drive his car;

- Trust has a directional aspect;

A can trust B, but it does not mean B trusts A;

- Trust has evolutive and temporal aspect;

Trust A has in B can increase over the time;

- Trust can be influenced by reputation;

A trusts B and now starts trusting C by recommendation of B;

- Trust is not transitive;

If A trusts B and B trusts C, it does not mean A trusts C;

TRUST

- Do not make confusion of TRUST with REPUTATION;
- REPUTATION:
  - Is an opinion someone has about others or something;
  - Has similar attributes to trust, but it is different;

Trust: I trust A for a reason;

Reputation: That person is good for doing this;

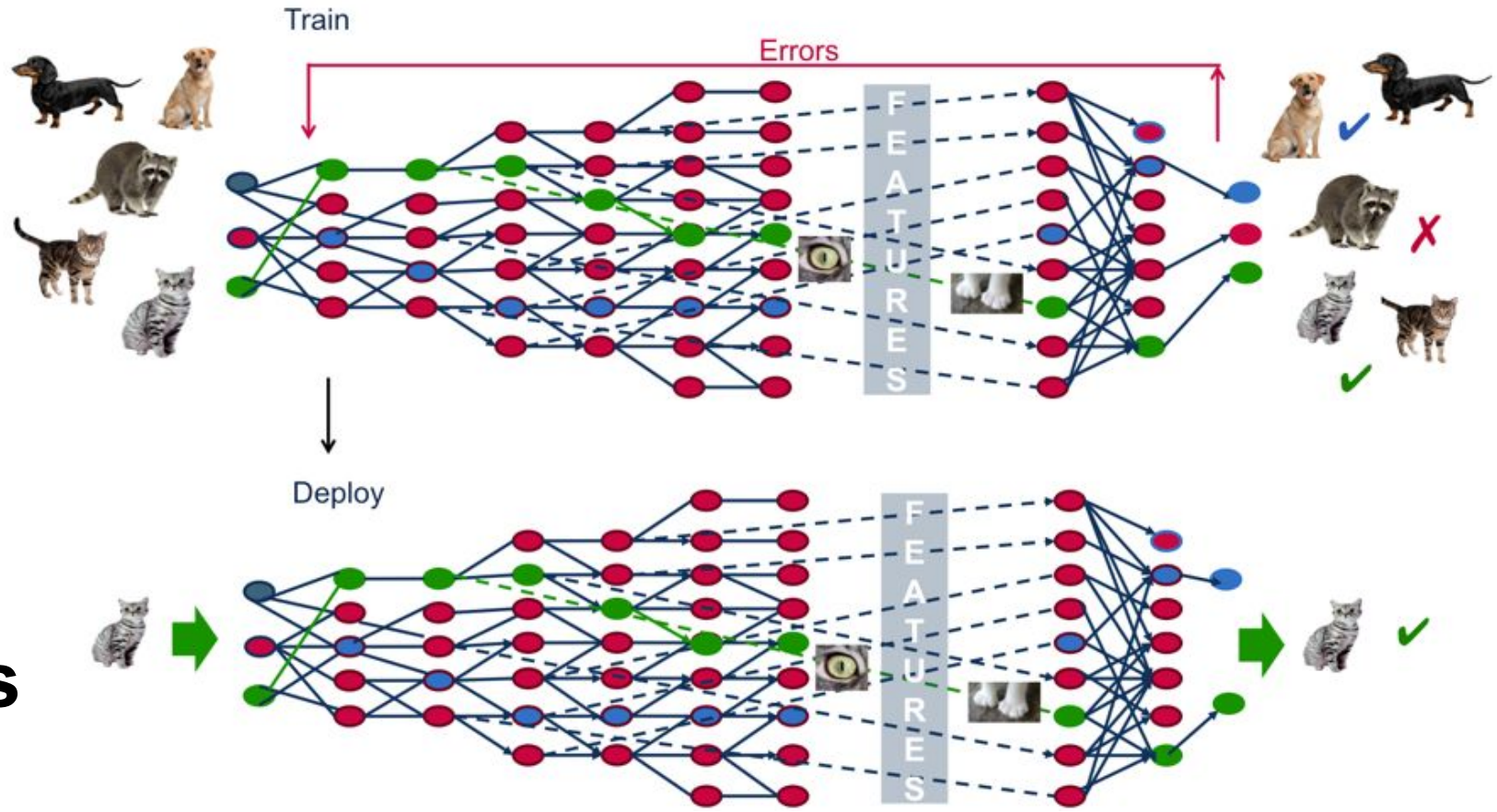
Artificial Intelligence

- **Possibility of making a computer capable of performing tasks similar to human reasoning;**

- **Research branch of Computer Science and Electrical Engineering;**

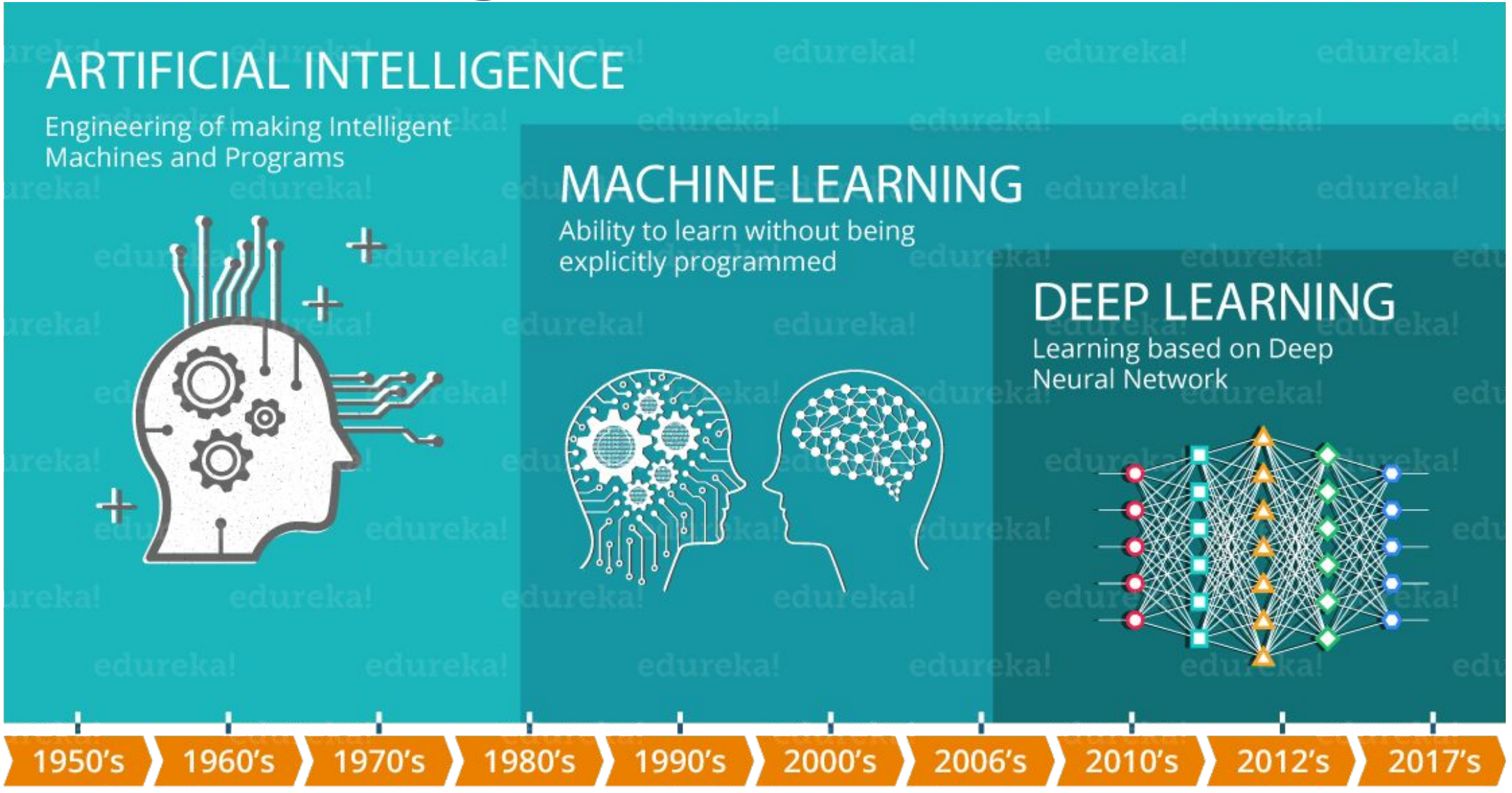
- **Seeks to build computational systems and/or devices that simulate humans' ability to think and to solve complex problems;**

Artificial Intelligence



- It needs:
  - Good data models for classifying, processing and analyzing;
  - Access to significant amount of data;
  - Computer systems for fast and efficient processing;

Artificial Intelligence



Artificial Intelligence

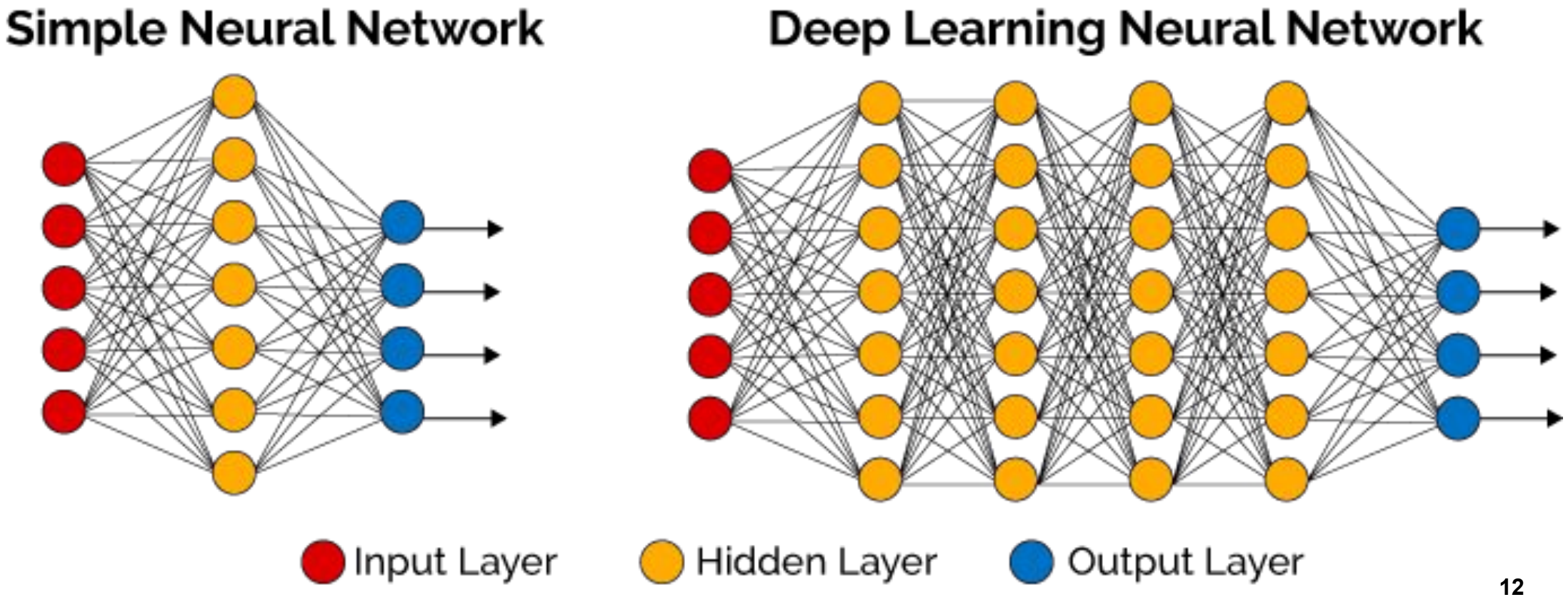
- AI deployments:
  - Machine Learning (ML):
    - Ability of the computer to learn by itself;
  - Deep Learning (DL):
    - Imitation of human neural networks;
  - Natural language processing (NLP):
    - Computer recognize natural language;
    - Ability to understand and compose texts;

Artificial Intelligence

- AI deployments:
  - Focus of ML, NLP e DL:
    - To learn multiple levels of representation and abstraction that helps understand data such as images, sound, and text;
    - Get context from it;

Artificial Intelligence

- **Deep Learning:**



Artificial Intelligence

- Some algorithms:
  - They are basically divided into:
    - Supervised and Unsupervised;
  - More recently:
    - semi-supervised; and
    - reinforcement learning;

Artificial Intelligence

- **Some algorithms:**

Nearest Neighbor	Principal Component Analysis
Naive Bayes	Single Value Decomposition
Decision Trees/Random Forest	Independent Component Analysis
Linear Regression	Ordinary Least Squares Regression
Support Vector Machines (SVM)	k-means clustering;
Neural Networks	Association Rules
Logistic Regression	Q-Learning
Ensemble Methods	Temporal Difference (TD)
Clustering	Deep Adversarial Networks

Artificial Intelligence

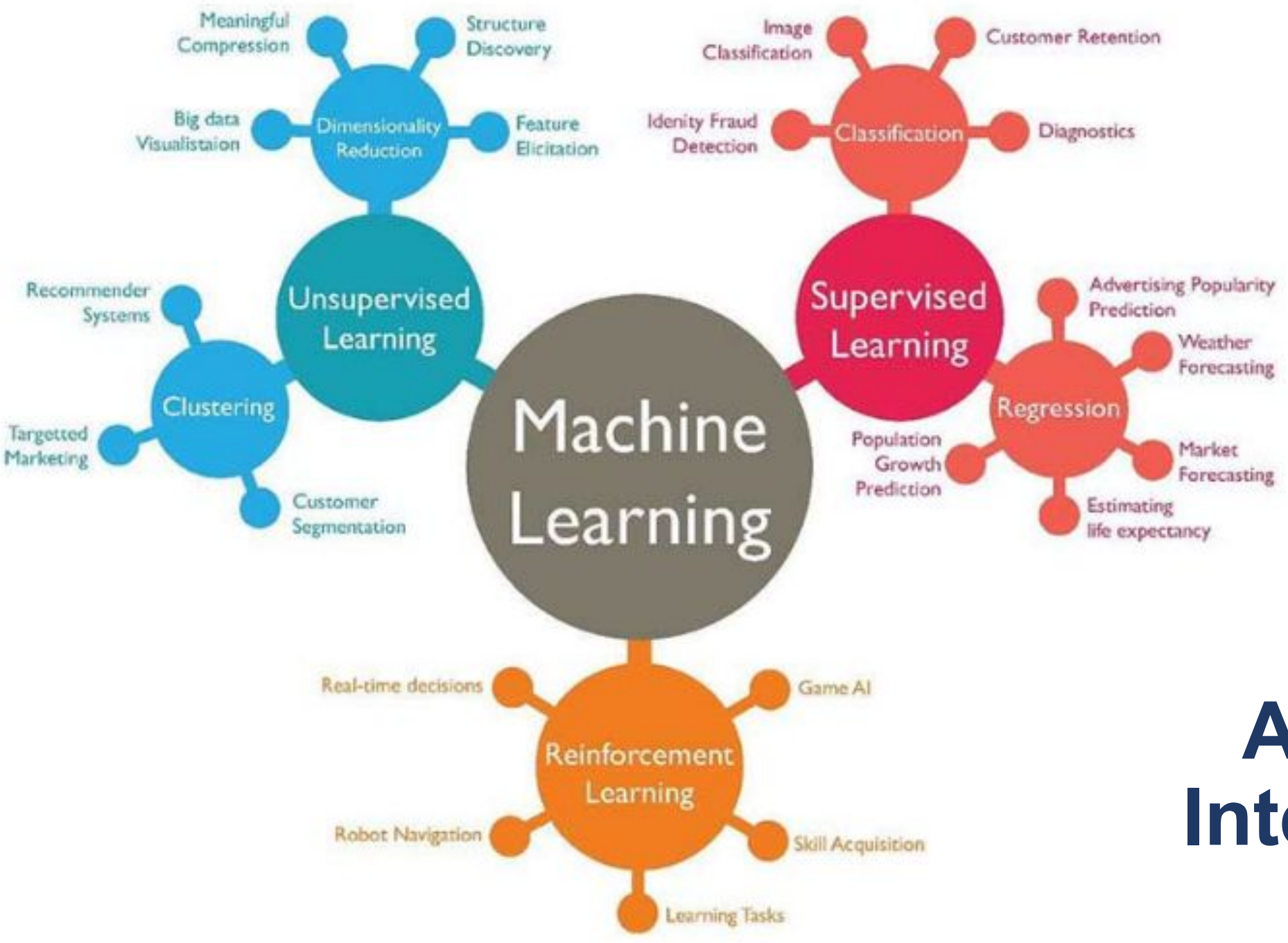
Question:

I have this problem! (which is not yet even very well defined as a problem);

What AI algorithm do I use?

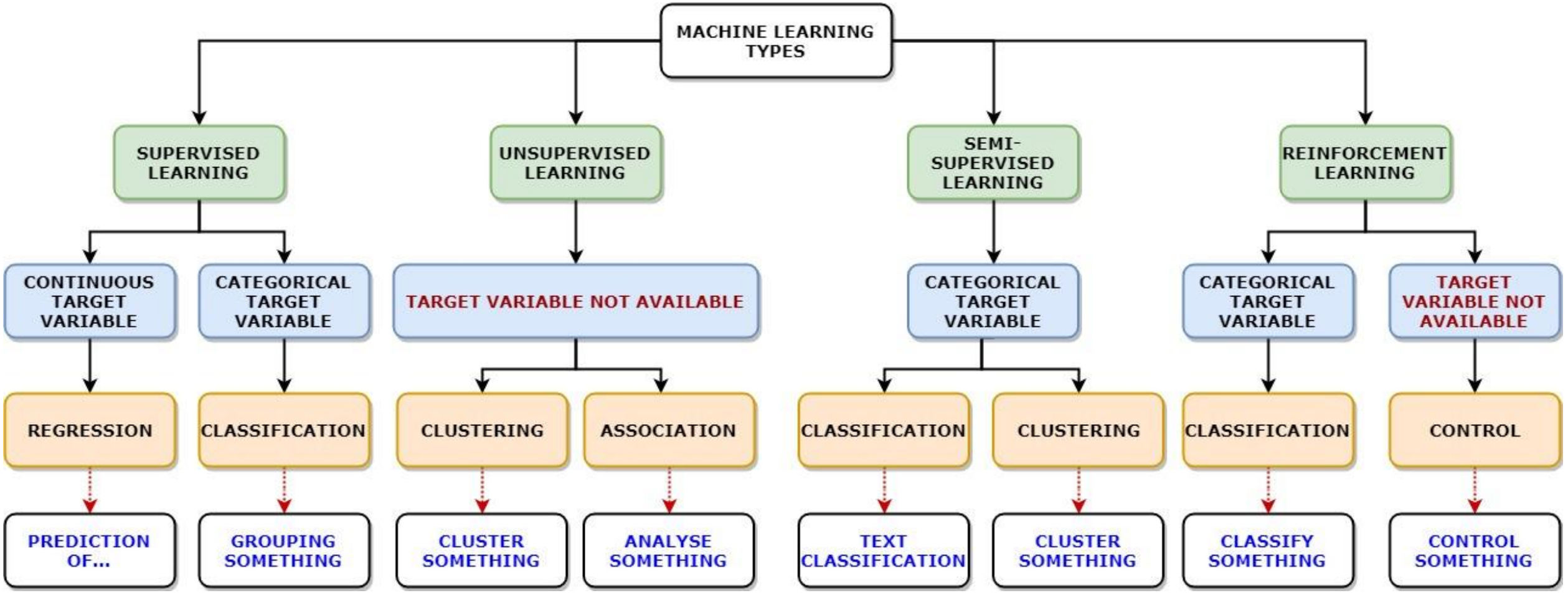
Answer:

IT DEPENDS ON MANY OTHER QUESTIONS!



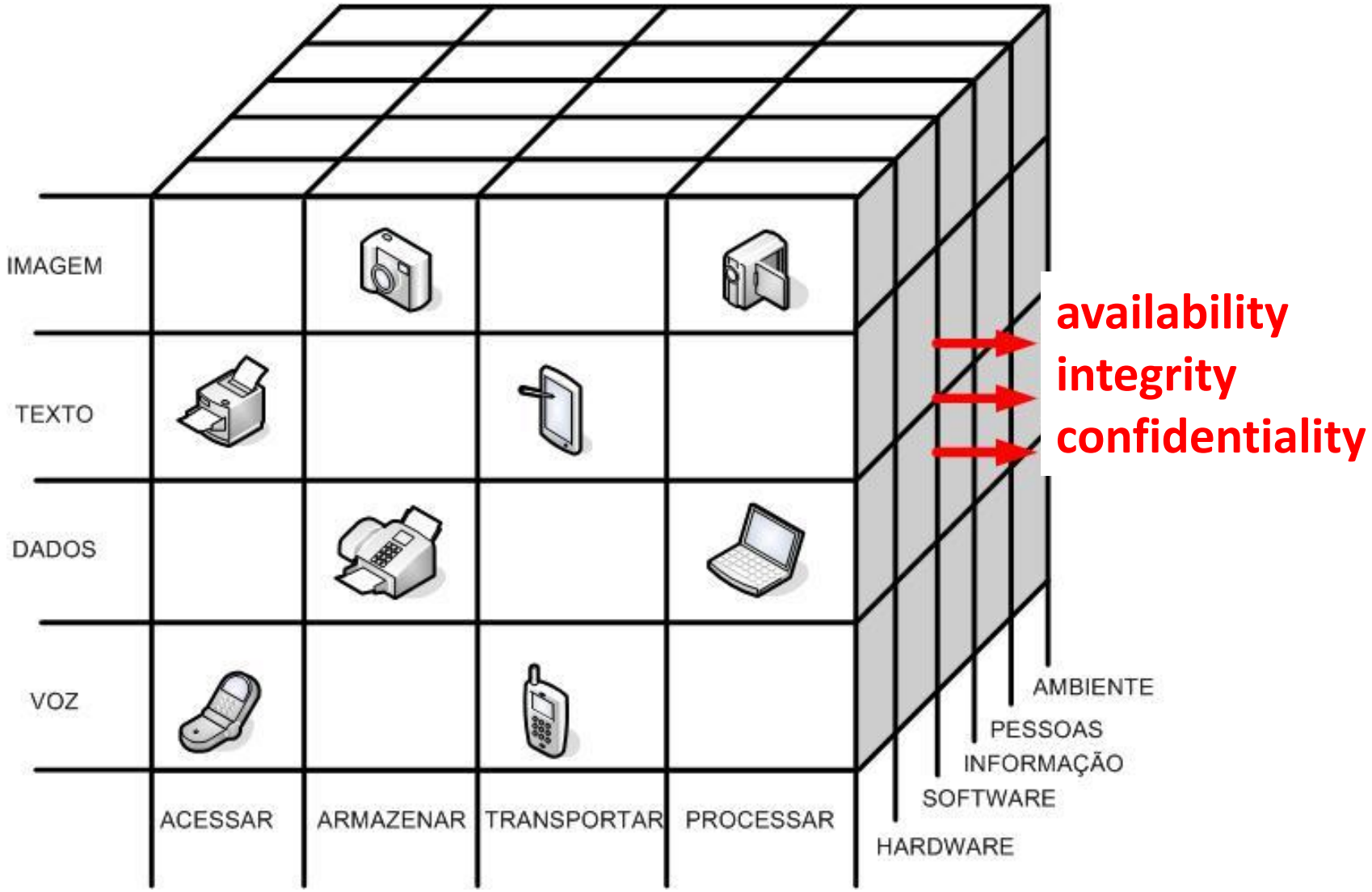
Artificial Intelligence

Artificial Intelligence



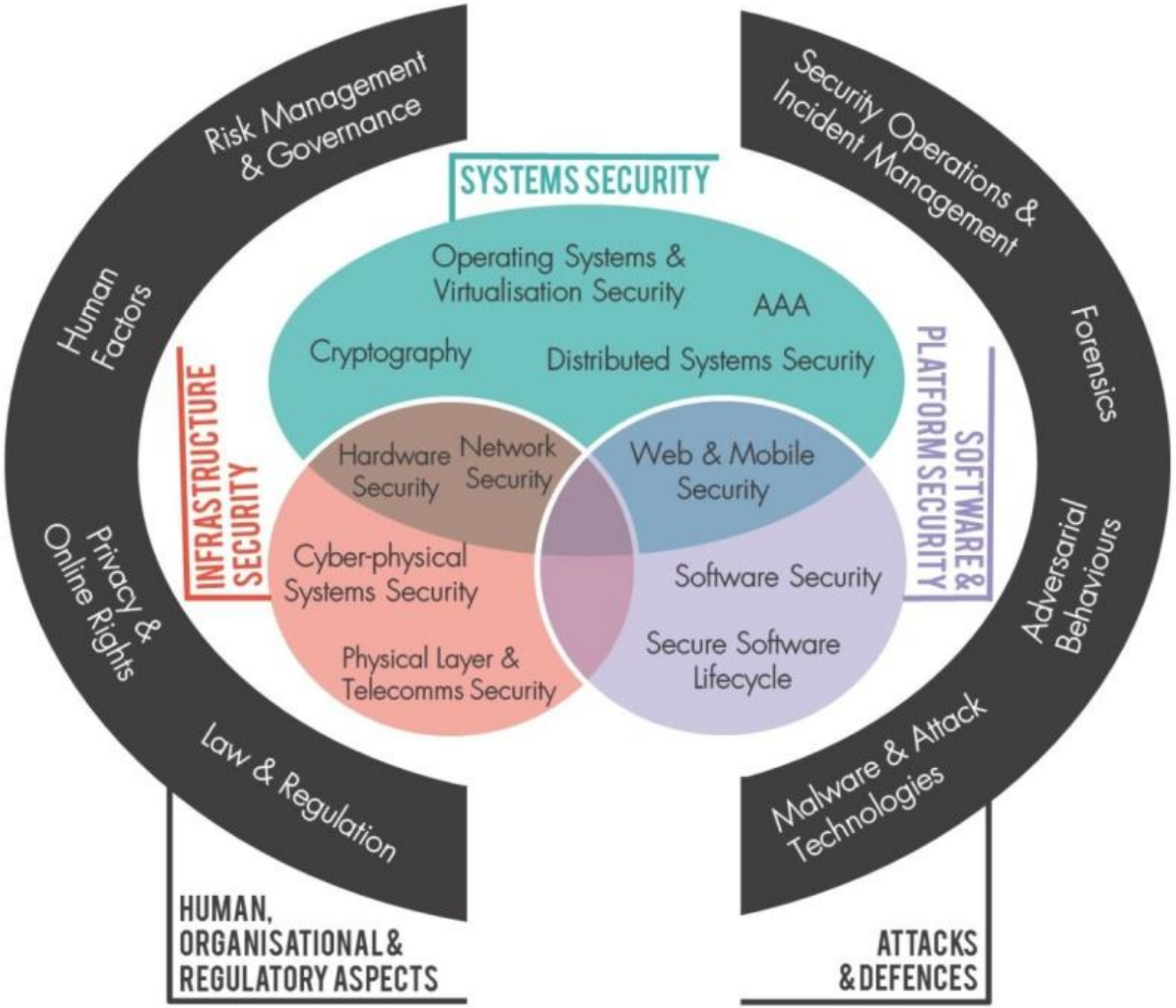
Attacks on AI systems

- **Information security is a complex matter;**



Attacks on AI systems

- **Cyber security is much more complex;**

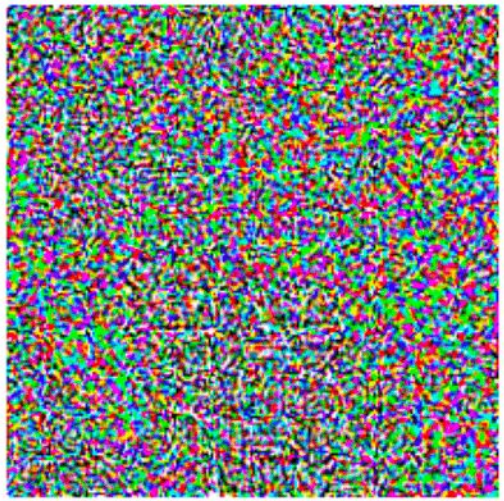


Attacks on AI systems

- **Exemplos:**

“panda”
57.7% confidence

$+ .007 \times$



“nematode”
8.2% confidence

$=$

“gibbon”
99.3 % confidence

Attacks on AI systems

- **Examples:**

Attacks on AI systems

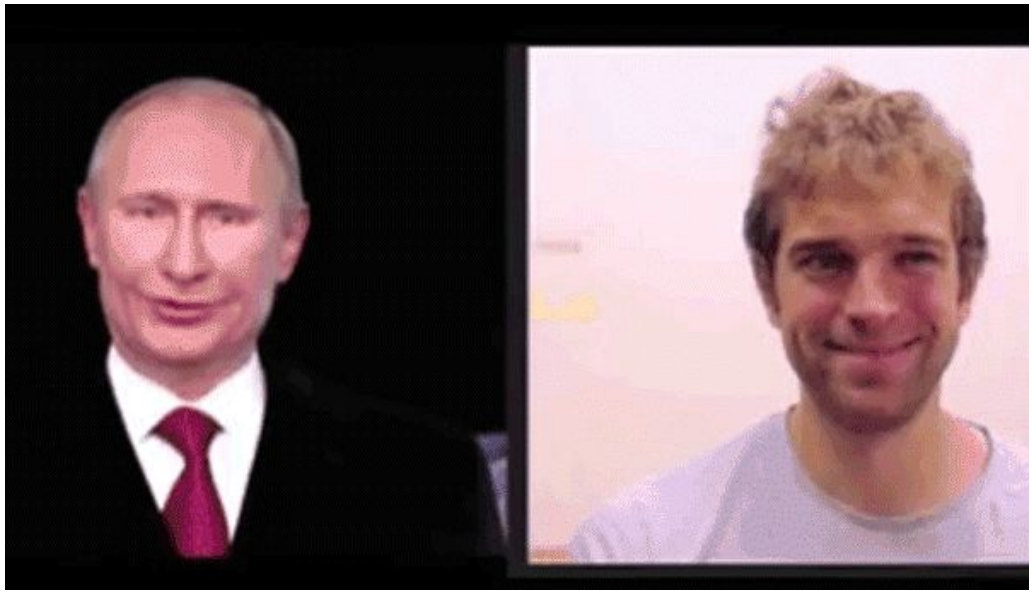
Facebook's A.I. labeled the video of Black men as content "about Primates."

Facebook Apologizes After A.I. Puts 'Primates' Label on Video of Black Men

Facebook called it "an unacceptable error." The company has struggled with other issues related to race.

Attacks on AI systems

- Deep Fakes



Input Output

Reference Our Result

Attacks on AI systems

- **Examples (videos):**
  - **Living portraits;**
  - **Adversarial neural networks;**
  - **Impersionate a celebrity;**

Attacks on AI systems

- **Examples:**
  - **Fake or not fake?**

Attacks on AI systems

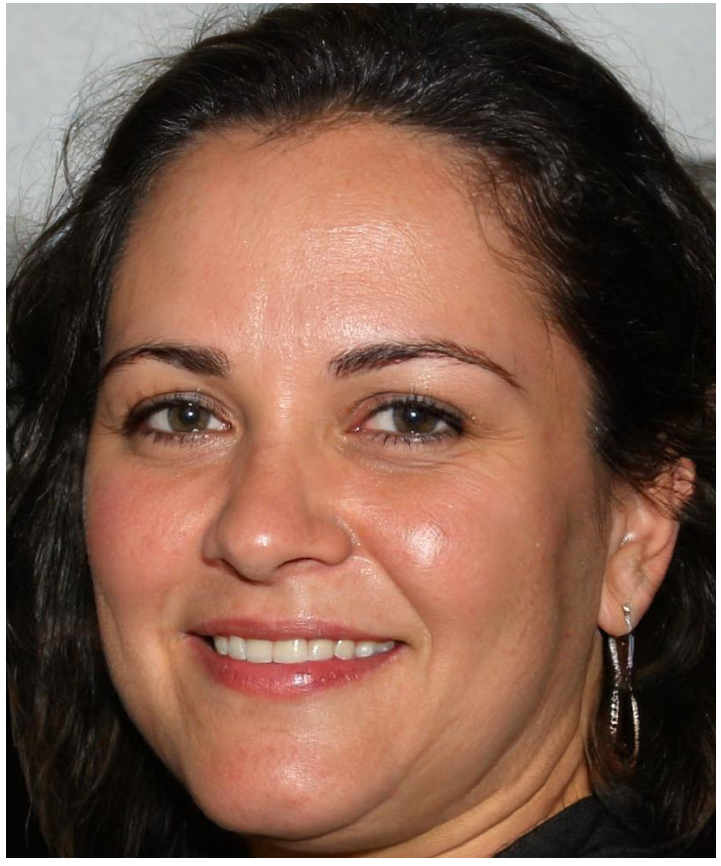
- **Exemples:**
  - **Fake or not fake?**

Produced by a GAN (generative adversarial network) StyleGAN (Dec 2018) - Karras et al. and Nvidia

Jun Hasegawa Japanese celebrity

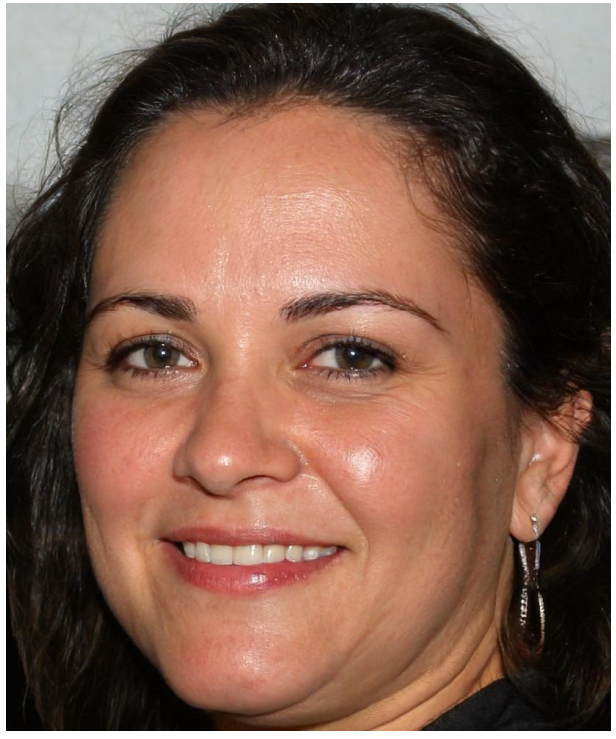
Attacks on AI systems

- **Exemples:**
  - **Fake or not fake?**



Attacks on AI systems

- **Exemples:**
  - **Fake or not fake?**



Produced by a GAN (generative adversarial network) StyleGAN (Dec 2018) - Karras et al. and Nvidia

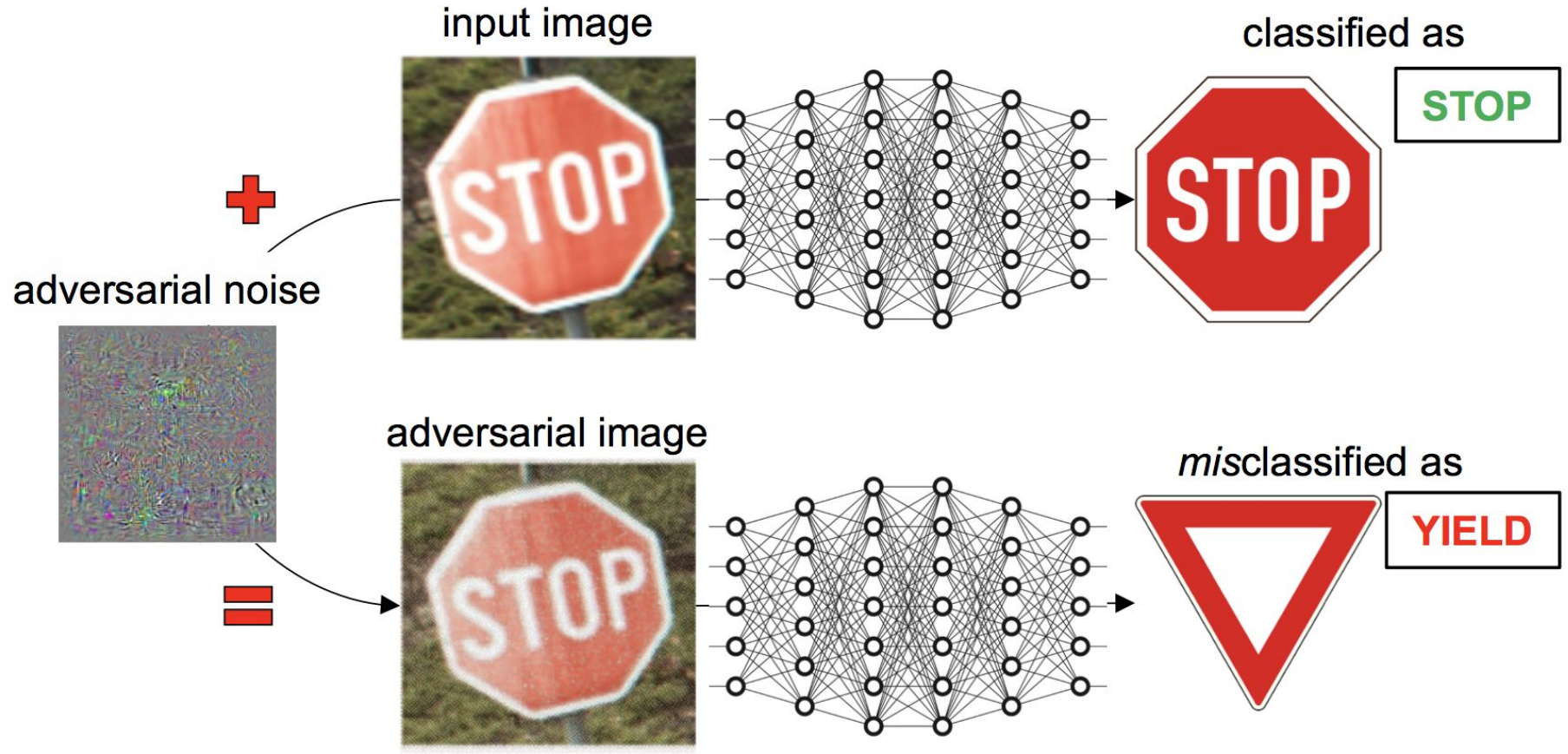
Nicanor Garcia photographer and architecture enthusiast Spanish Influencer

Attacks on AI systems

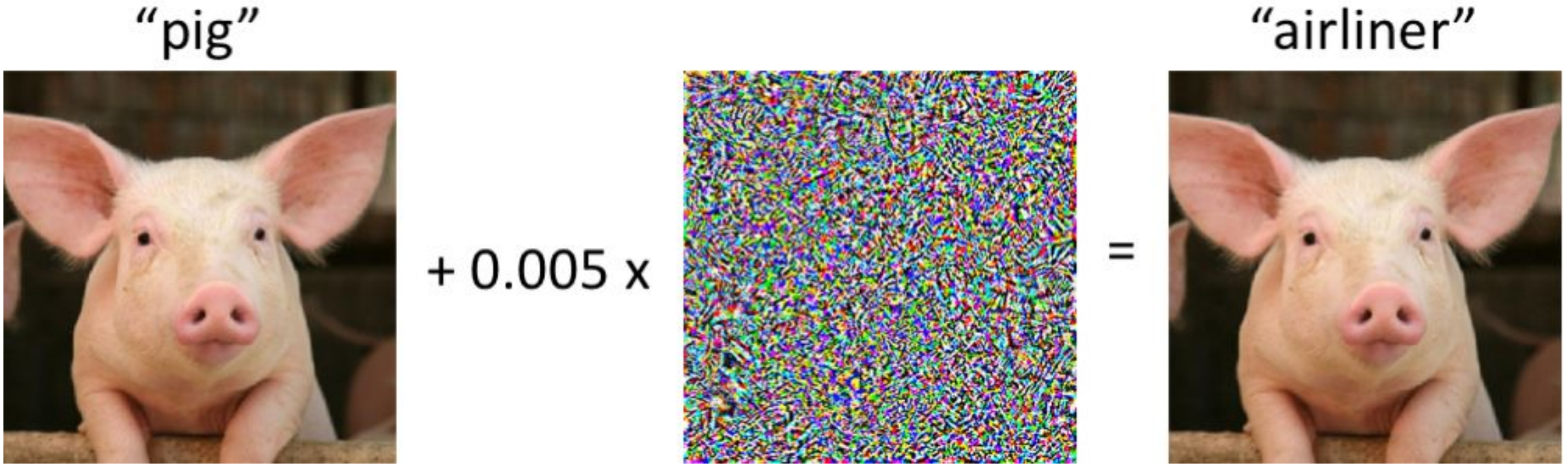
- 3 classifications are ordinary:
  - Adversarial inputs – Forged input data for misclassification;
  - Data poisoning attacks – Pollute classifier data in favor of the attacker;
  - Model stealing techniques – used to steal the model or retrieve/rebuild training data;

Discussions

- **Several problems arise from there:**



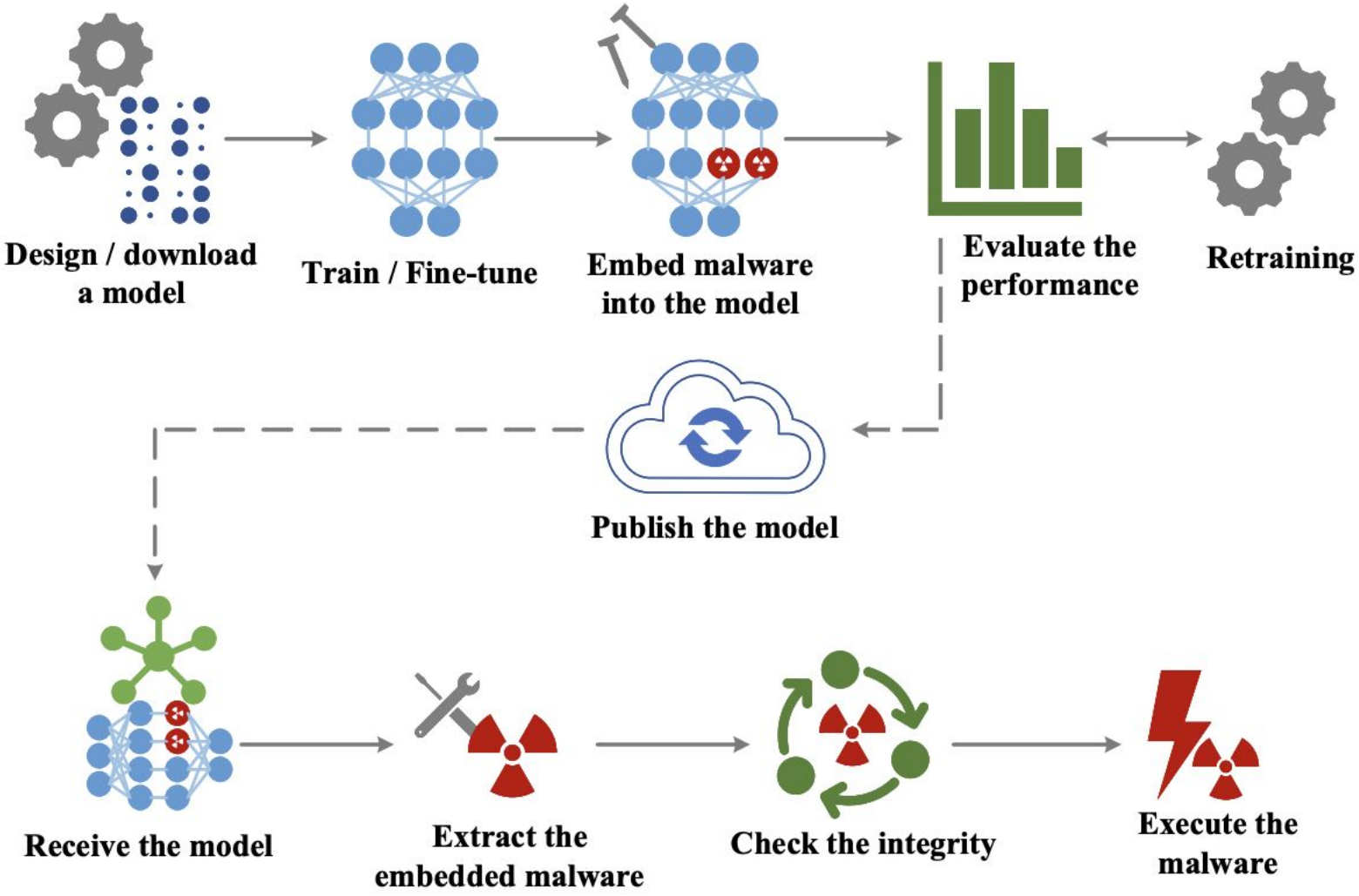
Discussions



Discussions

- AI with cybersecurity:
  - Malware or not malware?
  - Attack on the go or normal operations?
- AI within medicine:
  - Cancer or not cancer?
- AI with smart vehicles:
  - Hack or not hack? (video)
  - IA driver dilemma? (video)

Discussions



https://arxiv.org/pdf/2107.08590.pdf

EvilModel: Hiding Malware Inside of Neural Network Models

Zhi Wang, Chaoge Liu, Xiang Cui

Discussions

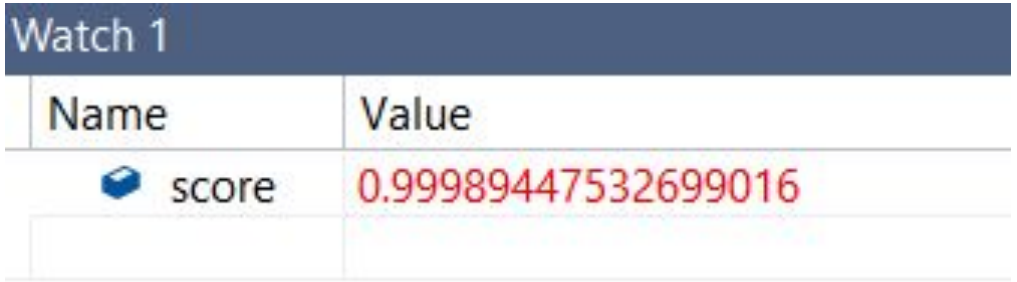
- AI within cybersecurity:
  - Hackers will always try to find the most economical way to get their way.
  - Ex. Cylance tool:
    - MIMIKATZ

Watch 1	
Name	Value
score	-0.852764688091270 71

Discussions

- **AI within cybersecurity:**
  - **Ex. Cylance tool:**
    - **MIMIKATZ**

```
copy /b mimikatz.exe+strings.txt modified_mimikatz.exe
```



Discussions

AI within cybersecurity:

Malware	Score Before	Score After
CoinMiner	-826	884
Dridex	-999	996
Emotet	-923	625
Gh0stRAT	-975	998
Kovter	-999	856
Nanobot	-971	999
Pushdo	-999	999
Qakbot	-998	991
Trickbot	-973	774
Zeus	-997	997

Discussions

- AI with DeepFake:
  - Twitter @AllanXia – video
  - Youtube Brazilian President Bolsonaro - video
    - Bruno Sartori - https://youtu.be/gaM68BEX420

Discussions

- **AI within Sentiment Analysis:**

```
In [23]: TextBlob("not that").sentiment
Out[23]: Sentiment(polarity=0.0, subjectivity=0.0)

In [25]: TextBlob("very great very great great").sentiment
Out[25]: Sentiment(polarity=0.9333333333333332, subjectivity=0.9)

In [27]: TextBlob("very great very great not that great").sentiment
Out[27]: Sentiment(polarity=0.9333333333333332, subjectivity=0.9)
```

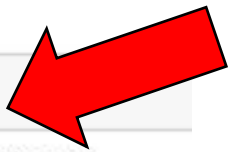
```
In [6]: TextBlob("very great").sentiment
Out[6]: Sentiment(polarity=1.0, subjectivity=0.9750000000000001)

In [7]: TextBlob("not great").sentiment
Out[7]: Sentiment(polarity=-0.4, subjectivity=0.75)

In [8]: TextBlob("very great and not great").sentiment
Out[8]: Sentiment(polarity=0.3, subjectivity=0.8625)
```

```
In [22]: TextBlob("not that great").sentiment
Out[22]: Sentiment(polarity=0.8, subjectivity=0.75)

In [28]: TextBlob("not great").sentiment
Out[28]: Sentiment(polarity=-0.4, subjectivity=0.75)
```



Inversion of the classifier

Discussions

- **AI within Sentiment Analysis:**

Computer Communications 174 (2021) 154–171

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Adversarial attacks on a lexical sentiment analysis classifier

Gildásio Antonio de Oliveira Júnior [a], Rafael Timóteo de Sousa Jr. [a], Robson de Oliveira Albuquerque [a,b], Luis Javier García Villalba [b,*]

[a] Science and Technology National Institute on Cyber Security (INCT) Center 6, Decision Technologies Laboratory - LATITUDE, Electrical Engineering Department (ENE), Faculty of Technology, University of Brasília (UnB), Brasília-DF, 70910-900, Brazil

[b] Group of Analysis, Security and Systems (GASS), Department of Software Engineering and Artificial Intelligence (DISIA), Faculty of Computer Science and Engineering, Office 431, Universidad Complutense de Madrid (UCM), Calle Profesor José García Santesmases 9, Ciudad Universitaria, 28040 Madrid, Spain

DO YOU TRUST YOUR ARTIFICIAL INTELLIGENCE SYSTEM?