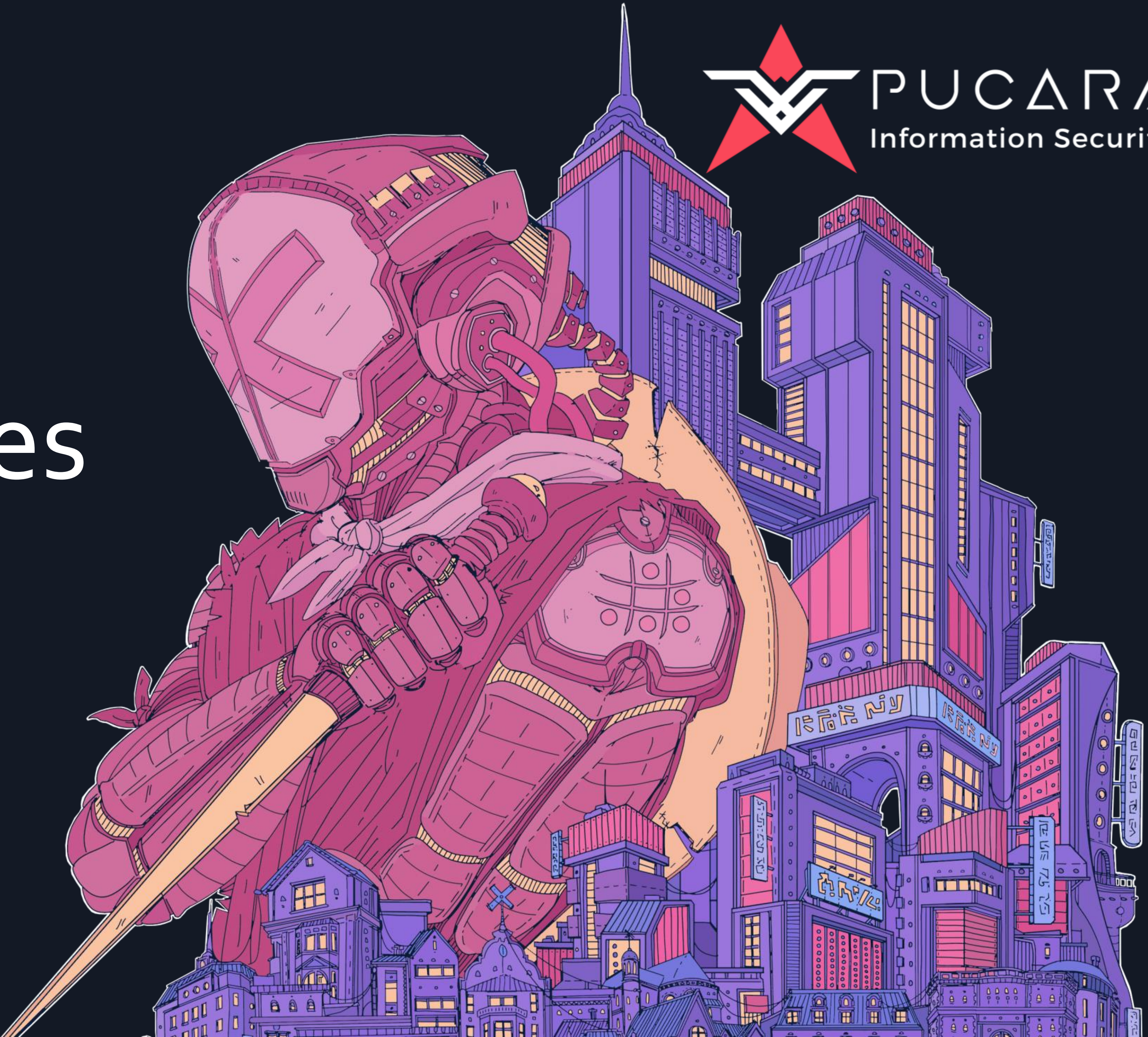
Machine Learning Weaknesses

PUCARA Information Security



crisofilaxx@pucara:~$ whoami

Lucas Bonastre

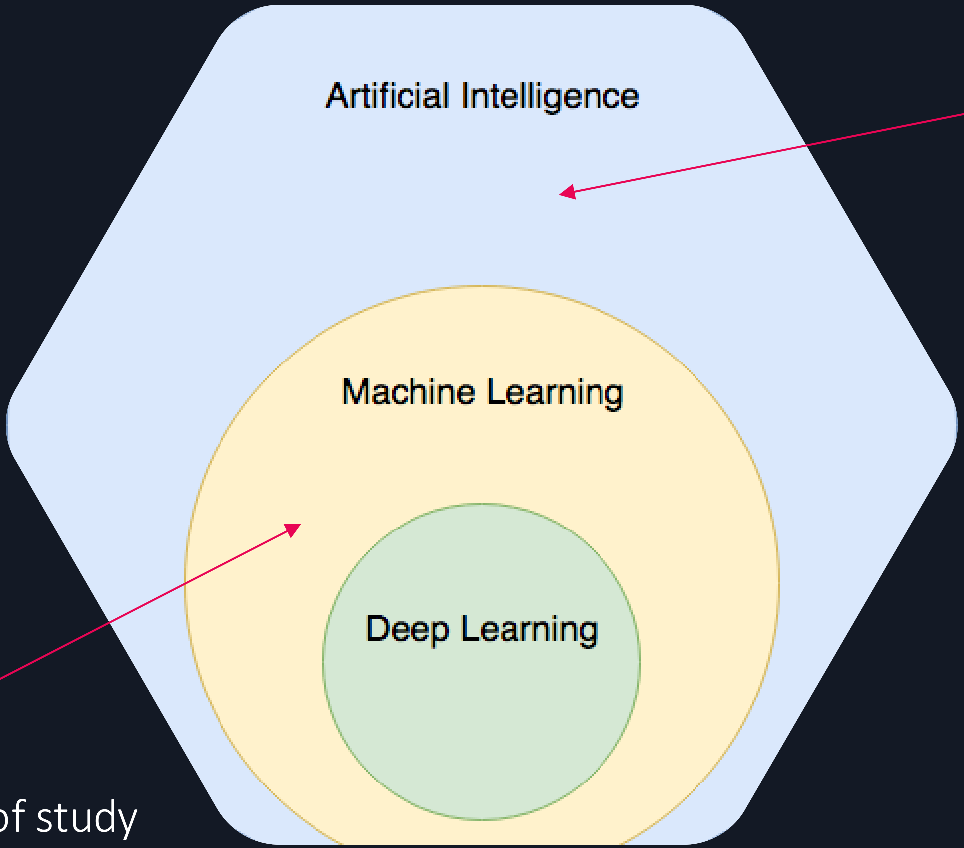
- ~~Lead Software Engineer at Pucara~~
- Just a Hacker

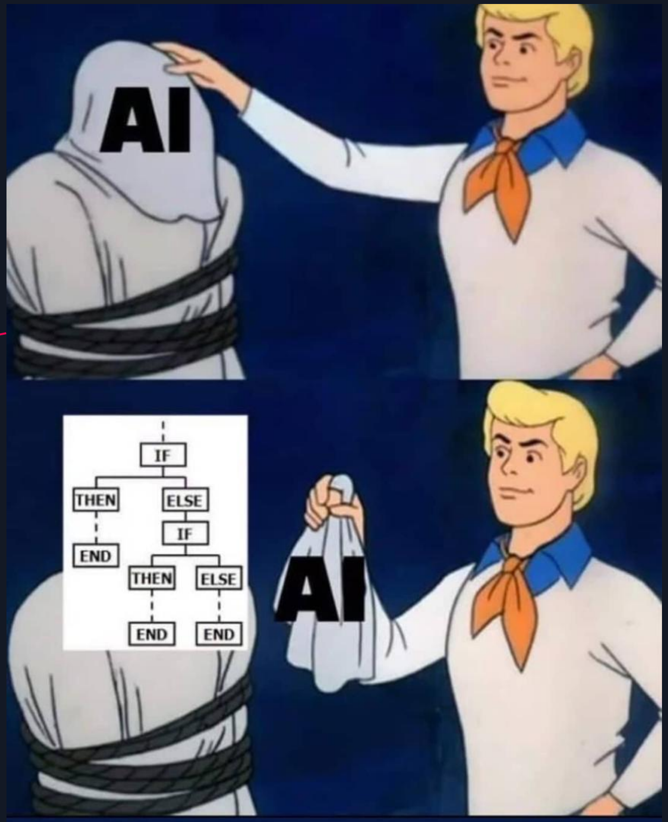
(Only toughness and general computer knowledge is required)

Understanding the target

Machine Learning

Understanding the target





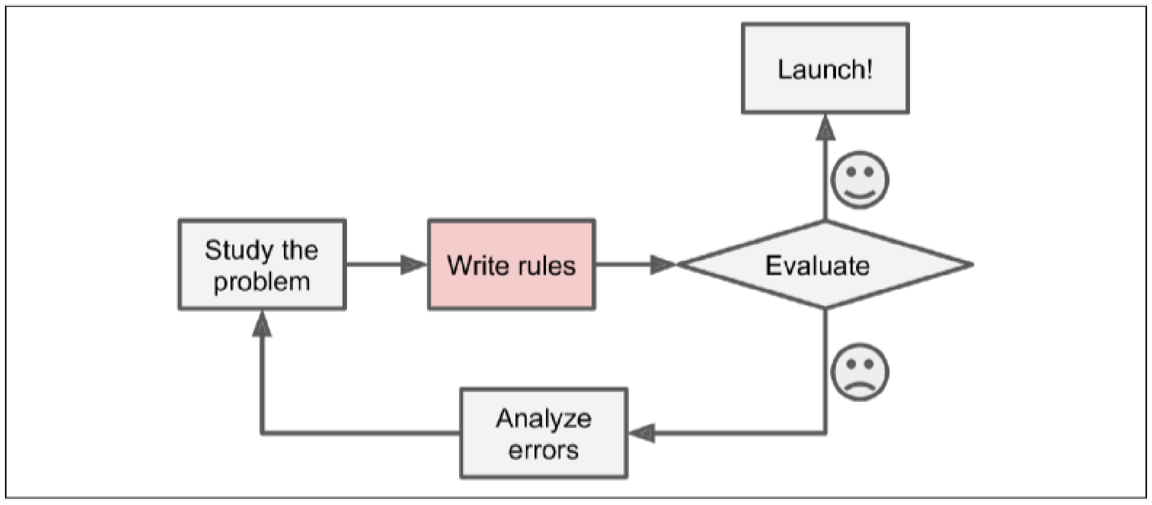
"[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed."

—Arthur Samuel, 1959

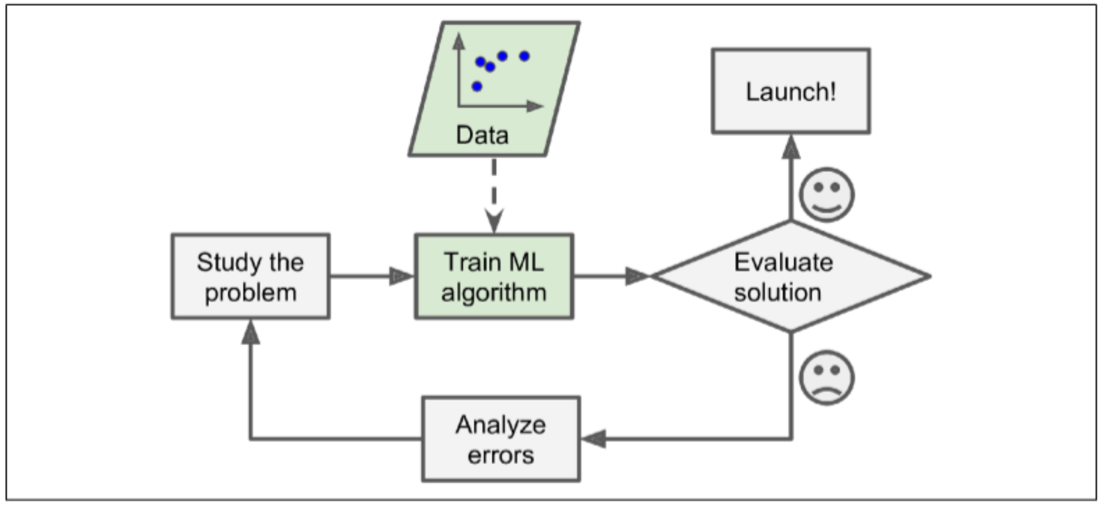
Different paradigm of software development

Understanding the target

Classic paradigm



Machine learning paradigm



Common applications

Understanding the target

Defensive security

- Spam filters.
- Speech recognition
- Static malware detection
- Biometric validation

Offensive security

- Sandbox detection
- Deep fakes

Other fields

- Healt systems
- Financial predictions

Classes of Machine Learning algorithms

Understanding the target

Regression Algorithms

Decision Tree Algorithms

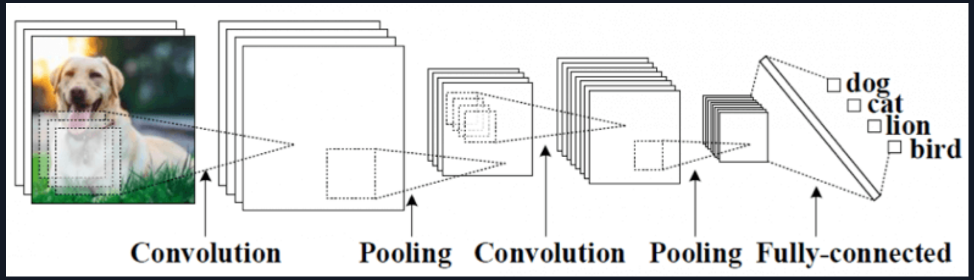
Clustering Algorithms

Instance-based Algorithms

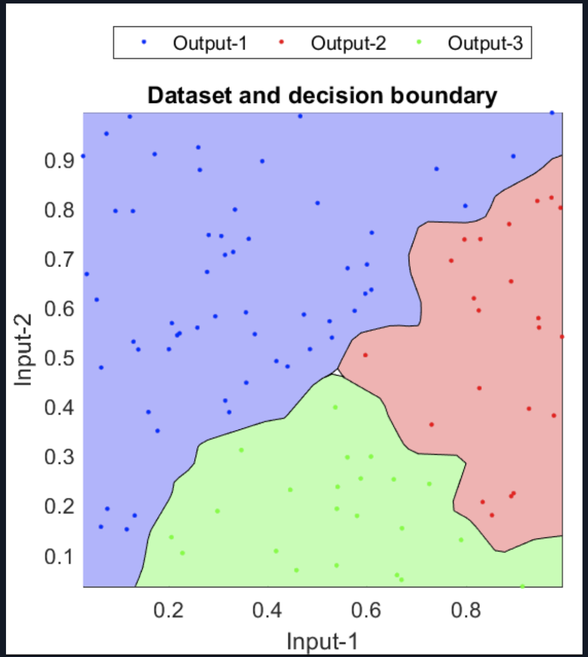
- K nearest neighbors (KNNs)
- Support vector machines (SVMs)

Deep Learning Algorithms

- Unsupervised Pretrained Networks (UPNs)
- Convolutional Neural Networks (CNNs)
- Recurrent Neural Networks (RNNs)



Convolutional Neural Networks (CNNs)

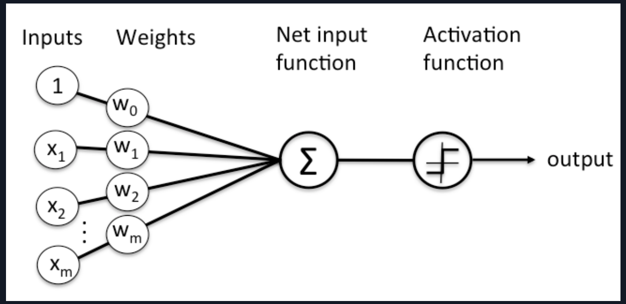


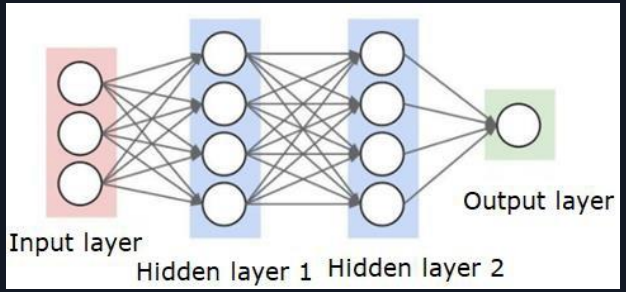
K nearest neighbors (KNNs)

Neural Networks for classification

Understanding the target

- A collection of nodes connected called "neurons"
- The connections have some weight that is adjusted during the learning process
- An activation function defines the output of neuron
- Gradient descent of a loss function is used to adjust the weights to better predict the output label
- Backpropagation is an algorithm that allows to adjust the weights of multi-layered Neural Networks





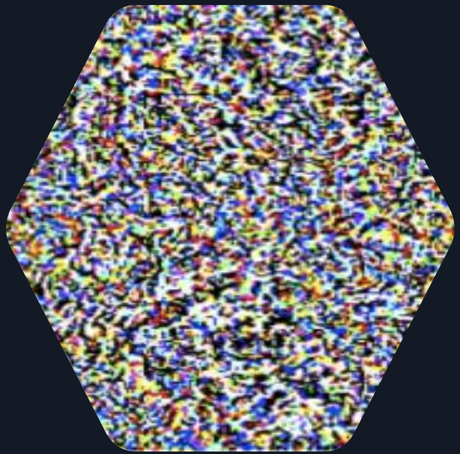
Attack surface

Known types of attacks

Adversarial inputs

Attempts to deceive the model with especially crafted input

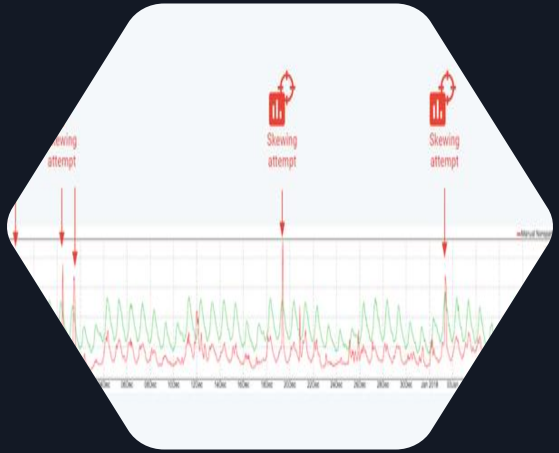
- Easy to craft in white box schemes. Challenging for black box
- Detectable for the "type noise" used to generate them
- Real world props have been created.



Data poisoning

The data used to train the model is somehow manipulated by the attacker

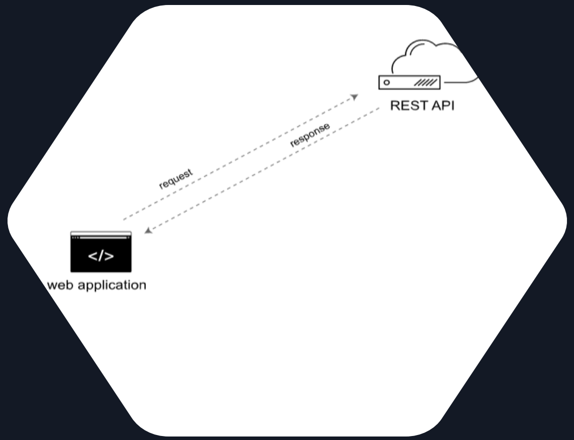
- Big thread for online learning models. Not so much for "vanilla" schemes
- Can produce general misclassification



Model theft

The model can be duplicated through API queries to the original model

- Intellectual property threat. A business threat for the creators of models
- Allows the generation of adversarial samples in black box schemes
- CVE-2019-20634 William Pearce



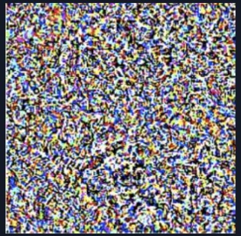
Case studies

Fast Gradient Sign Method Vector (FSGM)

Case study: Adversarial input

This method allows the creation of an optimal perturbation vector for Neural Networks that implement gradient descent as learning algorithm

$$\boldsymbol{\eta} = \epsilon \text{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right) =$$



GoogleNet classification:
GIANT_PANDA

GoogleNet classification:
INDRI

This attack has been implemented against pytorch's implemetation of the GoogleNet model

Source code and explanation:
https://blog.pucarasec.com/2020/07/23/deceiving-machine-learning-models/

EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES -
Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy
https://arxiv.org/pdf/1412.6572.pdf

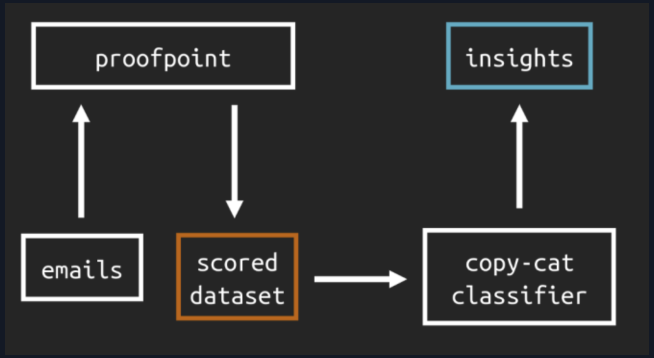
Proofpoint Email Protection (CVE-2019-20634)

Case study: Model theft

A bad design of solution exposed the same model to 230k+ clients and allow them to query the raw output of it. A flexible design for classic software development, but too permissive for the kind of model exposed

```
To: <reciever@domain.com>
From: <sender@domain.com>
Subject: Our Meeting
...
X-Proofpoint-Spam-Details: rule=nodigest_notspam policy=nodigest score=0
 malwarescore=0 mlxlogscore=999 mlxscore=0 suspectscore=14 spamscore=0
 impostorscore=0 adultscore=0 clxscore=593 priorityscore=0 phishscore=0
 bulkscore=97 lowpriorityscore=97 classifier=spam adjust=0 reason=mlx
 scancount=1 engine=9.1.0-12345000 definitions=main-12345
```

CVE-2019-20634 -
Will Pearce, Nick Landers
https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2019-20634



A dataset was created with the model's outputs. After proposing some NN architectures to create a classifer a "copy-cat classifier" was created. After the model theft was possible to create adversarial examples compromising the full security of the system

Final thoughts

Countermeasures

There is no silver bullet...

- Adversarial training
- Ensembling (use multiple learning algorithms to obtain better predictive performance)
- Data validation processes
- Use sensible data sampling
- API limitations
- Security oriented DESIGN

Q&A

pucarasec

pucarasec

blog.pucarasec.com

Thanks!