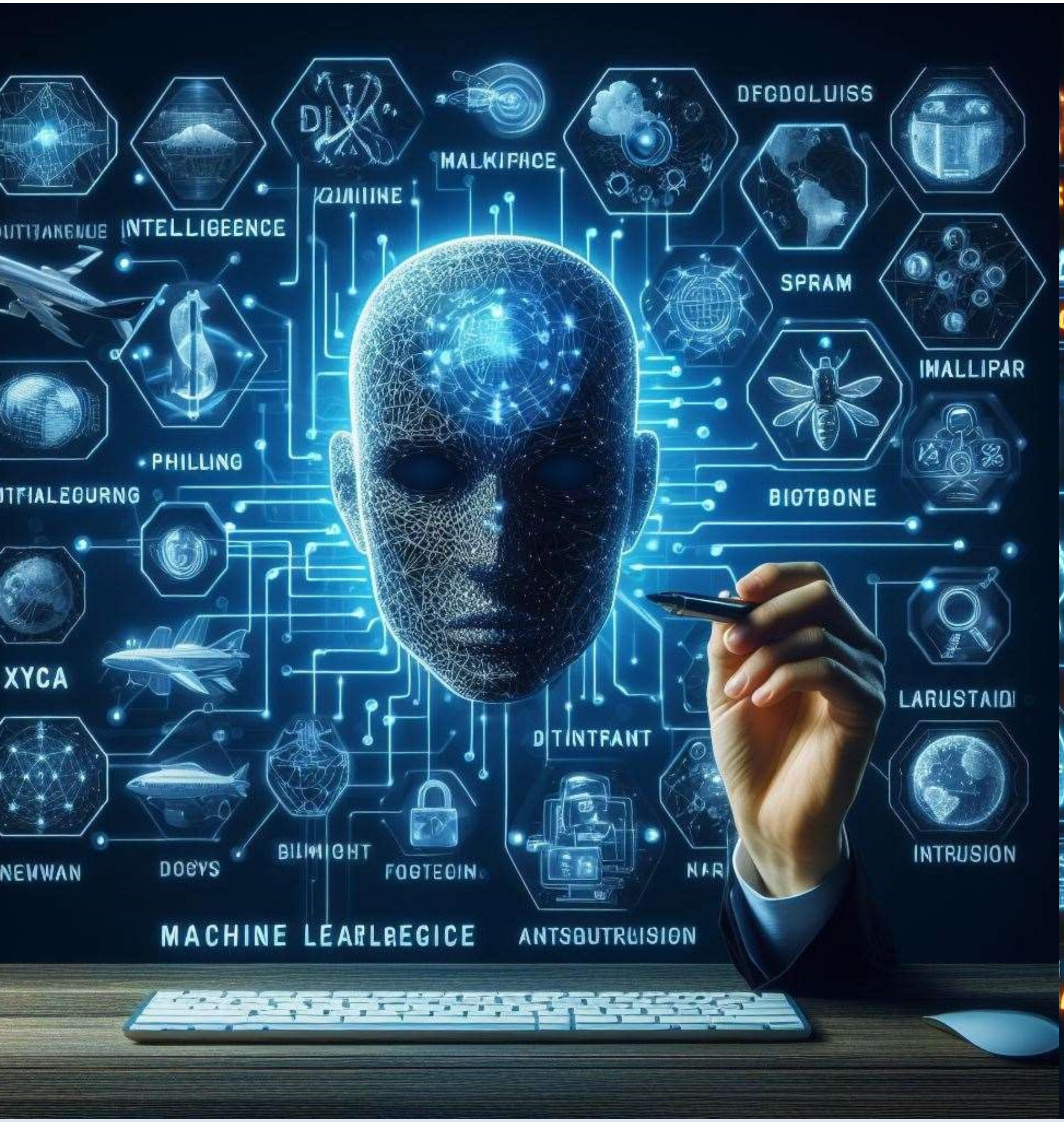
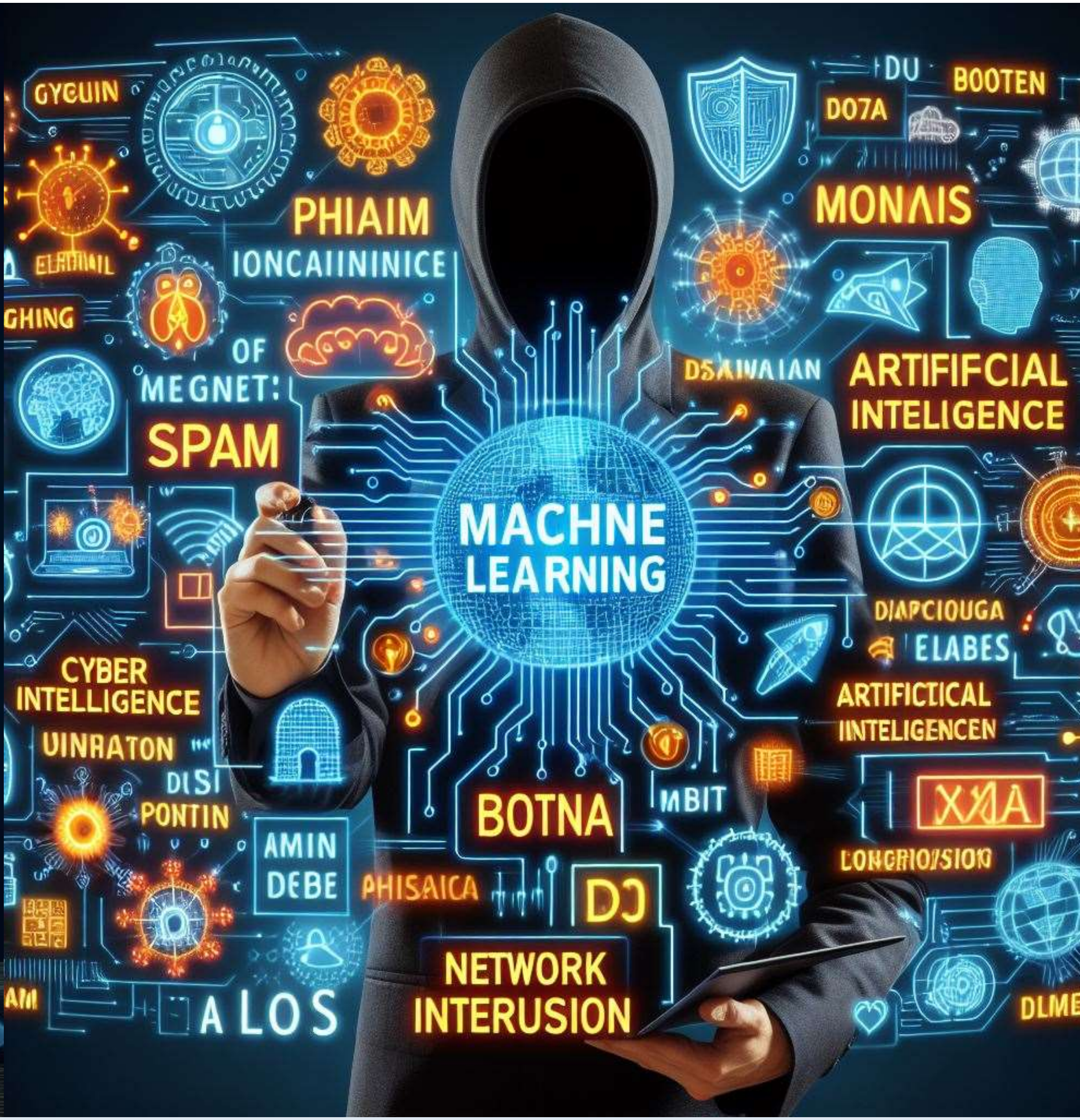

DFGDOLUISS
MALKIPHCE
DLX
QIIMIINE
UTTANENUE
INTELLIGENCE
SPRAM
IMALLIPAR
PHILLING
UTIALEOURNG
BIOTBONE
XYCA
LARUSTAIDI
DTINTFANT
BIMIGHT
NEWWAN
DOCYS
FOGTEOIN
NAR
INTRUSION
MACHINE LEARLREGICE
ANTSOUTRUISION


GYCUIN
DU
BOOTEN
DO7A
PHIAIM
MONAIS
IONCAIIINIICE
ELAIIIIL
GHING
OF
MEGNET:
DSAIVAIAN
ARTIFICIAL
INTELIGENCE
SPAM
MACHNE
LEARNING
DIAPCIOUGA
ELABES
CYBER
INTELLIGENCE
ARTIFICTICAL
INTELIGENCEN
UINRATON
DISI
PONTIN
BOTNA
MBIT
XMA
AMIN
DEBE
PHISAICA
DJ
LONGROISION
NETWORK
INTERUSION
AM
ALOS
DLME

CSA | AI

Why Model Explainability is essential for Cyber Security?

PRESENTED BY

Dr. Lars Ruddigkeit

Co-Lead Swiss Chapter & Zero Trust WG8

Agenda

1 What is XAI?

2 Why is XAI needed

3 (X)AI in Cybersecurity

4 Conclusion

I.) The ChatGPT Moment

Security World Before ChatGPT

CISO: AI is a business topic!

ChatGPT

November 2022

Security World After ChatGPT

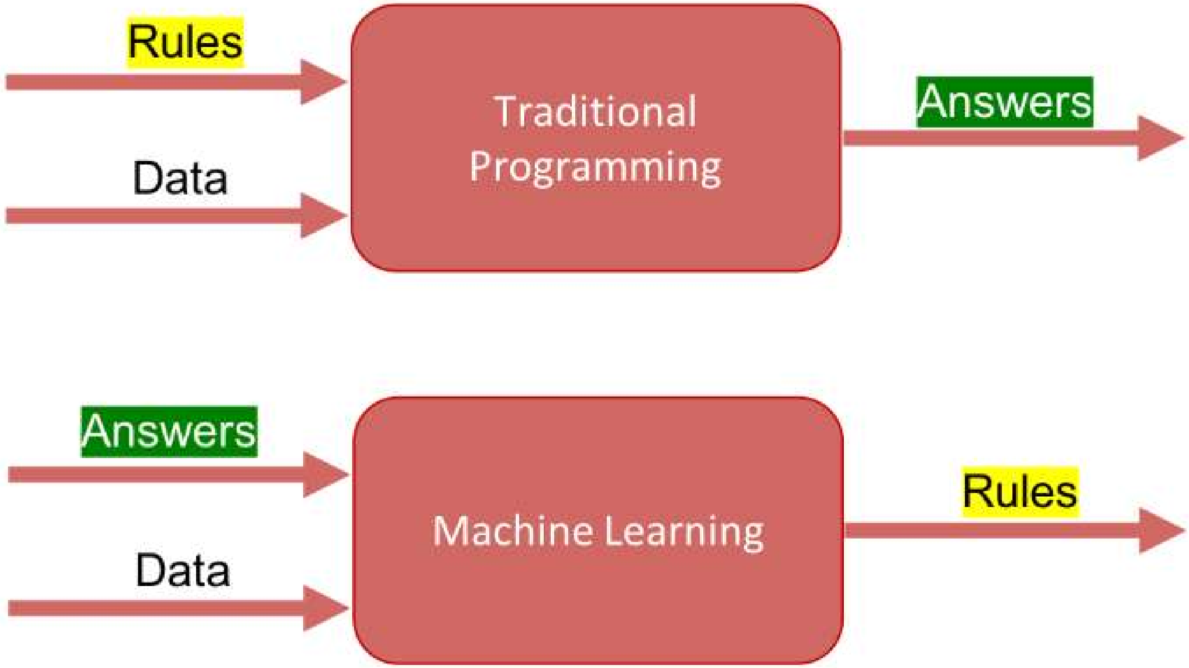
CISO: What is the risk of AI?

Regulation for AI

EU Artificial Intelligence Act

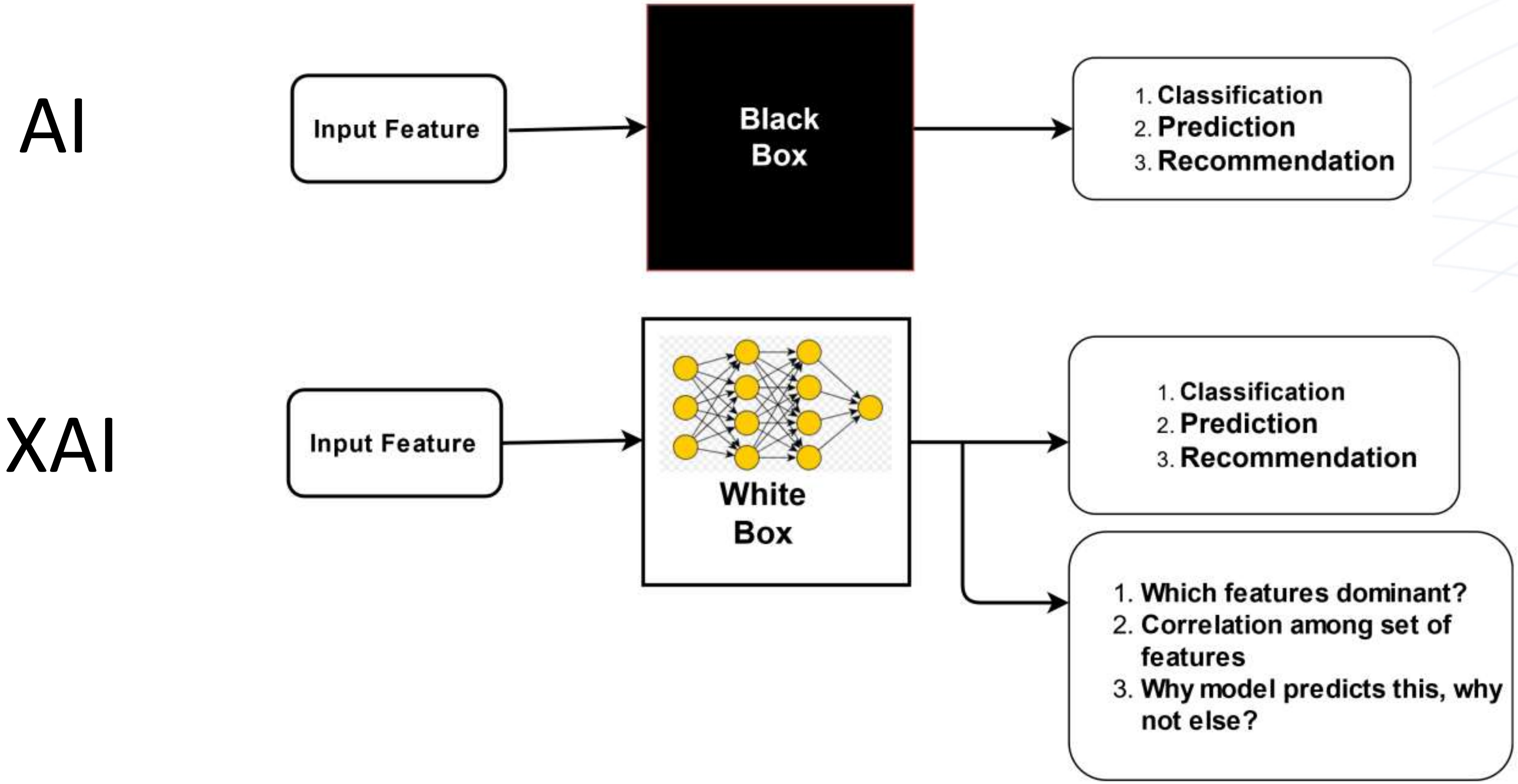
Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence

1.) Fundamental Artificial Intelligence Difference



Source: https://datalya.com/blog/machine-learning/machine-learning-vs-traditional-programming-paradigm

I.) Artificial Intelligence vs. Explainable Artificial Intelligence



AI

Input Feature

Black Box

1. Classification
2. Prediction
3. Recommendation

XAI

Input Feature

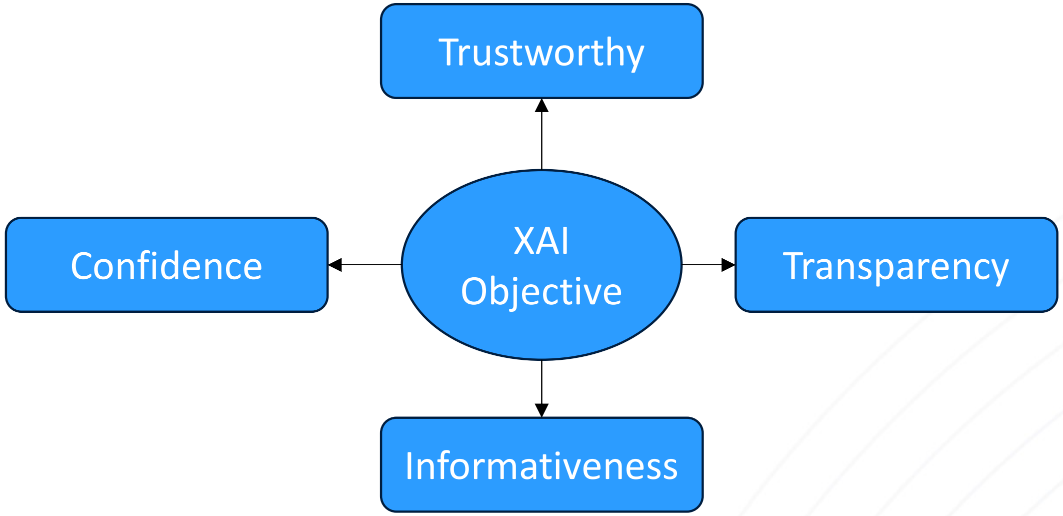
White Box

1. Classification
2. Prediction
3. Recommendation

1. Which features dominant?
2. Correlation among set of features
3. Why model predicts this, why not else?

Source: https://arxiv.org/abs/2107.07045

1.) XAI What?



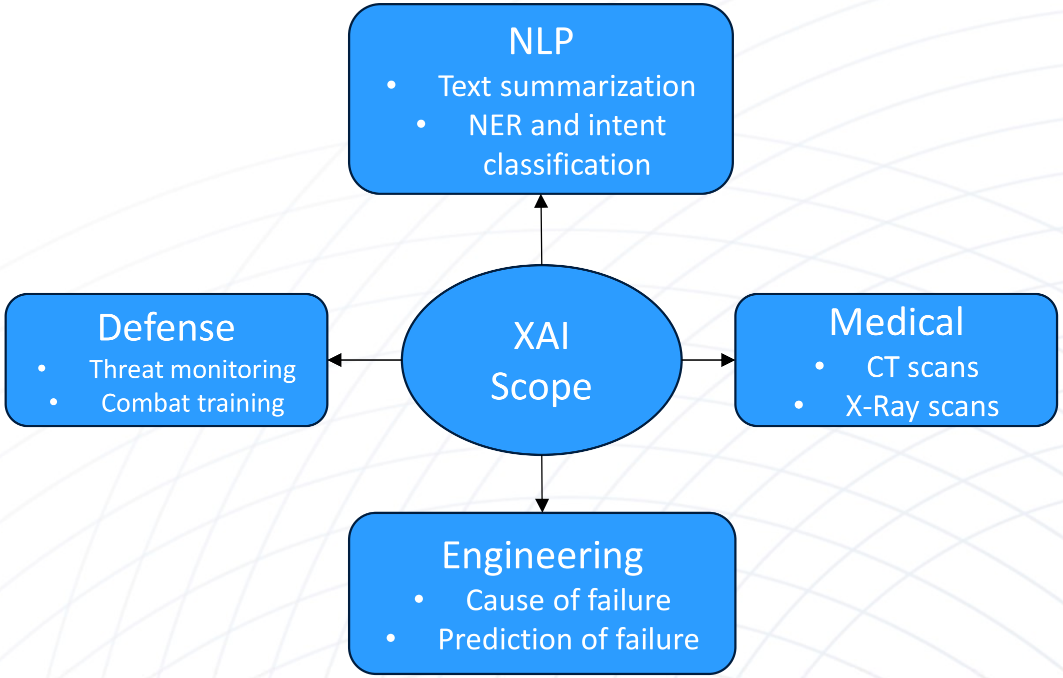
XAI Objective

Trustworthy

Confidence

Transparency

Informativeness



XAI Scope

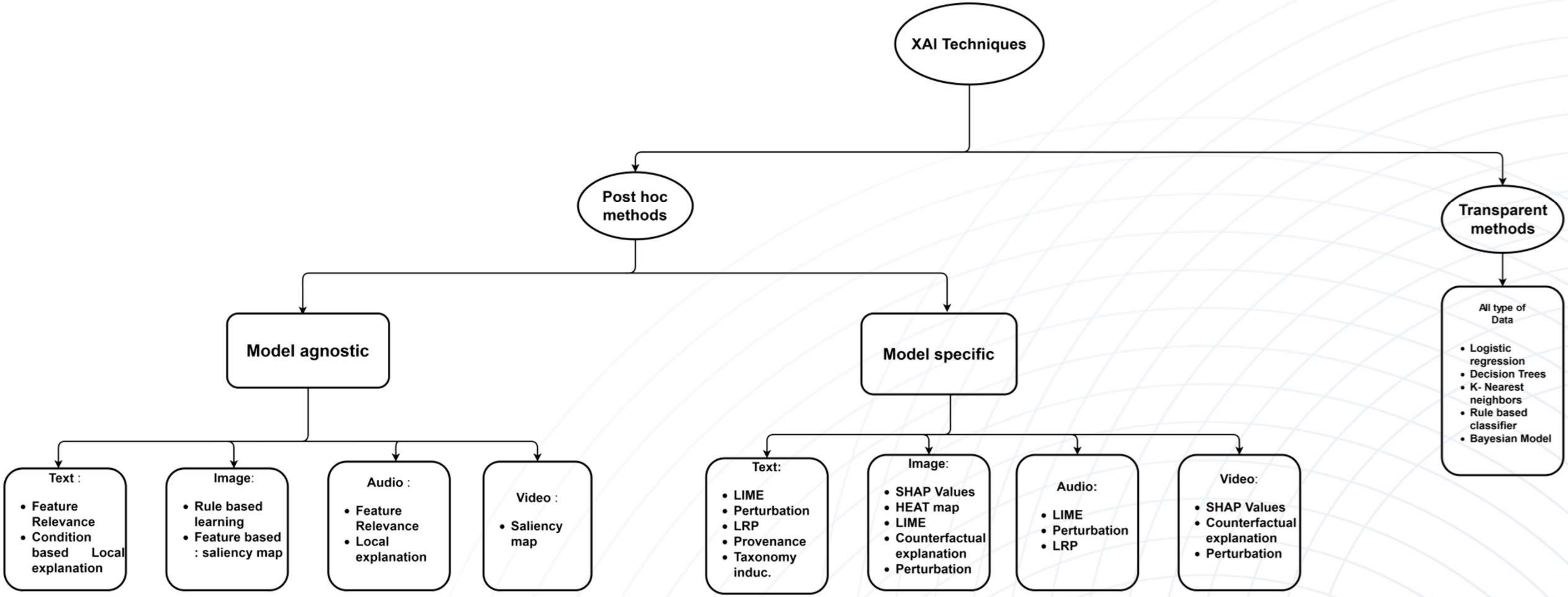
NLP
- Text summarization
- NER and intent classification

Defense
- Threat monitoring
- Combat training

Medical
- CT scans
- X-Ray scans

Engineering
- Cause of failure
- Prediction of failure

XAI classification with respect to type of data



XAI Techniques

Post hoc methods

Transparent methods

Model agnostic

Model specific

All type of Data
- Logistic regression
- Decision Trees
- K- Nearest neighbors
- Rule based classifier
- Bayesian Model

Text :
- Feature Relevance
- Condition based Local explanation

Image:
- Rule based learning
- Feature based : saliency map

Audio :
- Feature Relevance
- Local explanation

Video :
- Saliency map

Text:
- LIME
- Perturbation
- LRP
- Provenance
- Taxonomy induc.

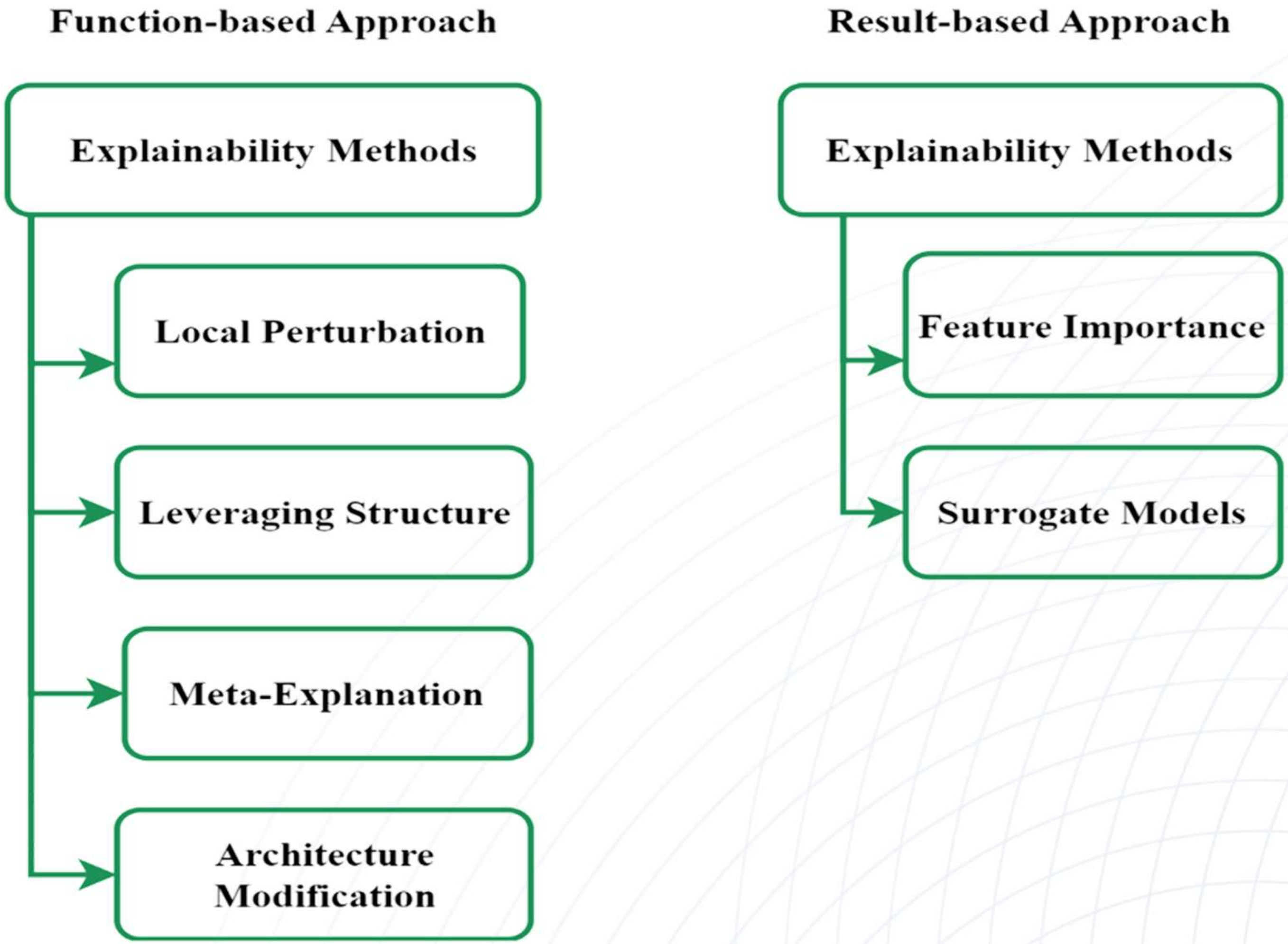
Image:
- SHAP Values
- HEAT map
- LIME
- Counterfactual explanation
- Perturbation

Audio:
- LIME
- Perturbation
- LRP

Video:
- SHAP Values
- Counterfactual explanation
- Perturbation

Source: https://arxiv.org/abs/2107.07045

XAI classification with respect to taxonomy



Function-based Approach

Explainability Methods

Local Perturbation

Leveraging Structure

Meta-Explanation

Architecture Modification

Result-based Approach

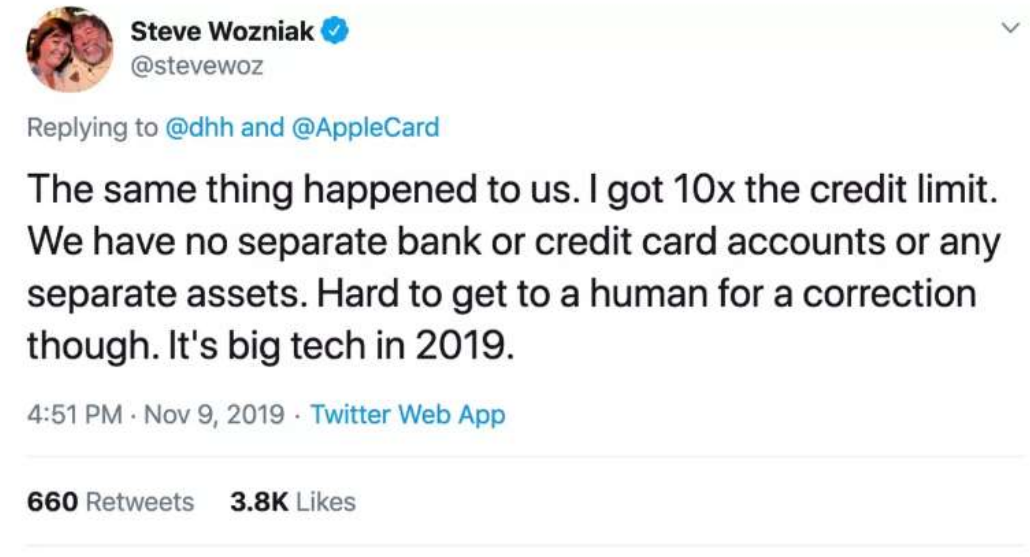
Explainability Methods

Feature Importance

Surrogate Models

Source: https://www.sciencedirect.com/science/article/pii/S277266222300070X

2.) AI Discrimination – Data Bias



Steve Wozniak
@stevewoz

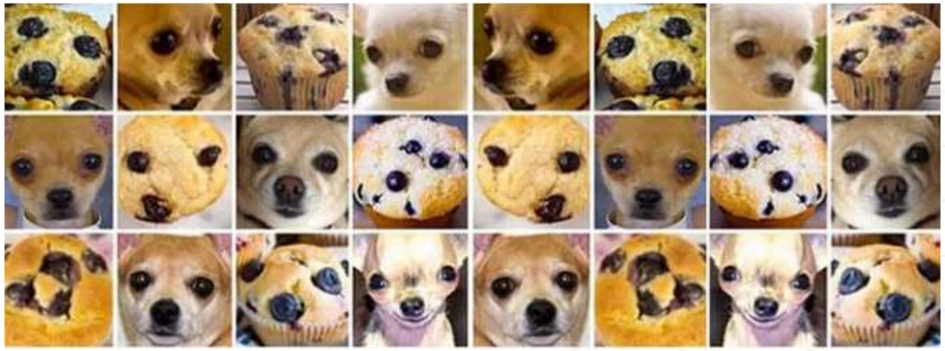
Replying to @dhh and @AppleCard

The same thing happened to us. I got 10x the credit limit. We have no separate bank or credit card accounts or any separate assets. Hard to get to a human for a correction though. It's big tech in 2019.

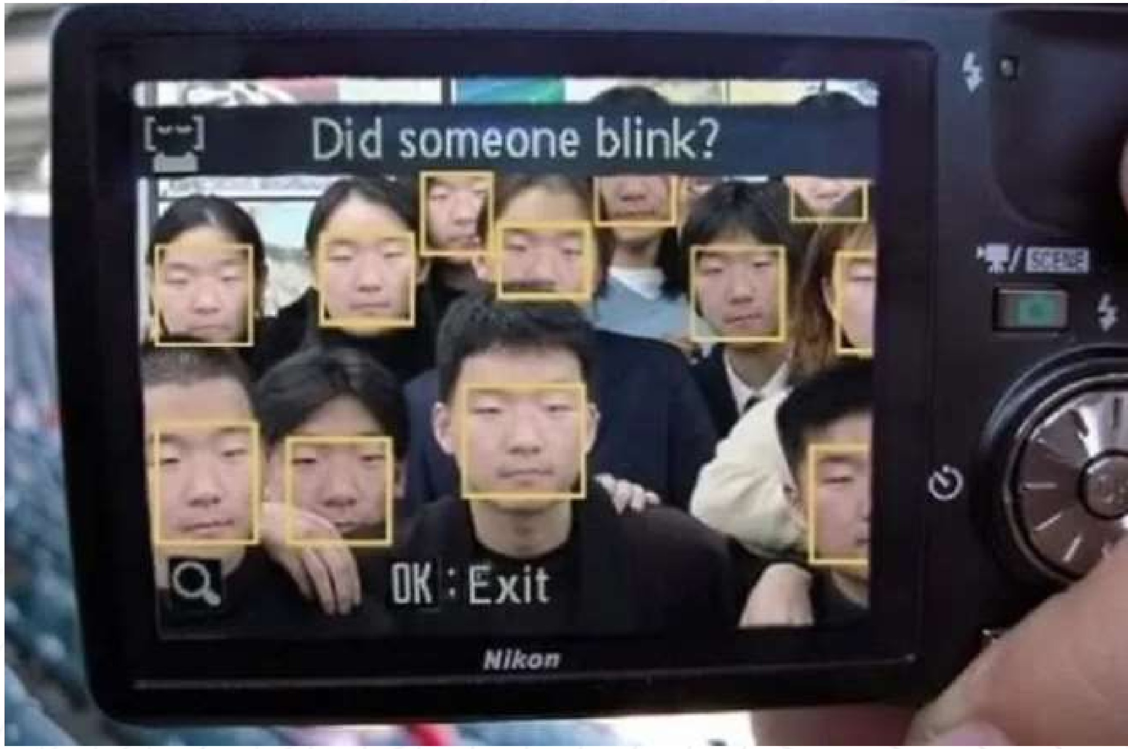
4:51 PM · Nov 9, 2019 · Twitter Web App

660 Retweets 3.8K Likes

Source: Twitter now X

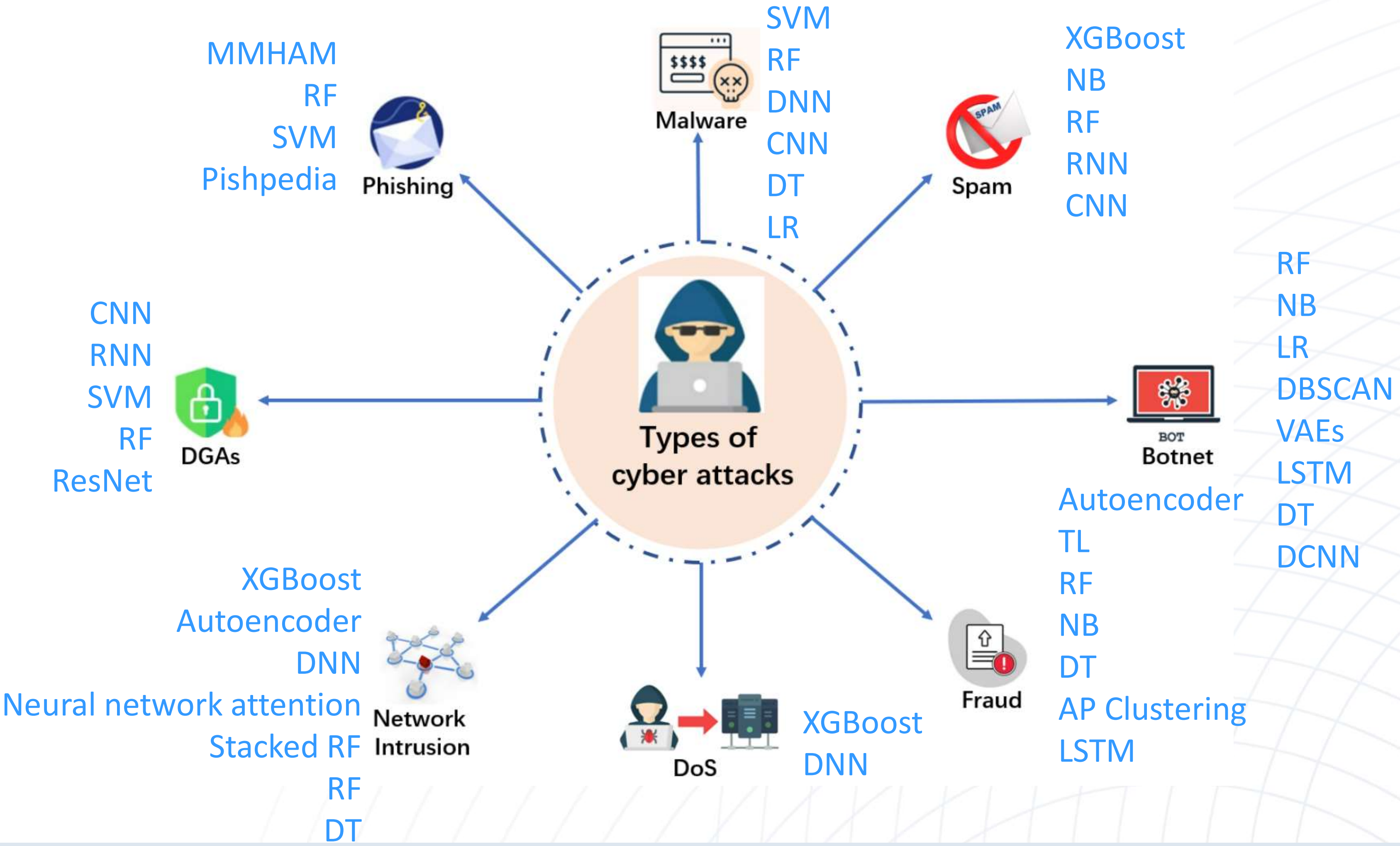


Source: www.freecodecamp.org



Source: https://funnyjunk.com

3.) Typical Cyber Threats and Machine Learning involvement



Types of cyber attacks

Phishing: MMHAM, RF, SVM, Pishpedia

Malware: SVM, RF, DNN, CNN, DT, LR

Spam: XGBoost, NB, RF, RNN, CNN

Botnet: RF, NB, LR, DBSCAN, VAEs, LSTM, DT, DCNN

Fraud: Autoencoder, TL, RF, NB, DT, AP Clustering, LSTM

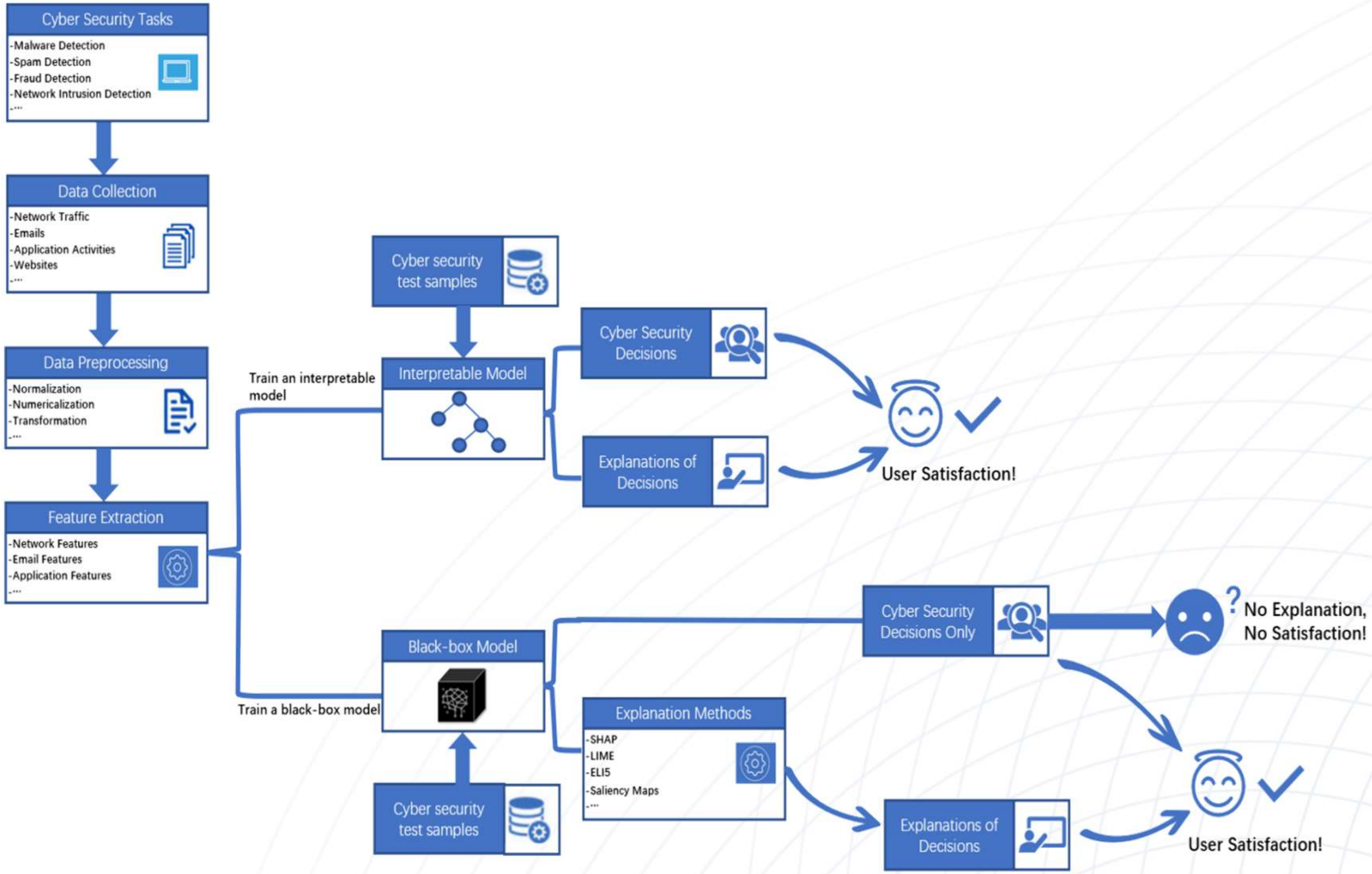
DoS: XGBoost, DNN

Network Intrusion: XGBoost, Autoencoder, DNN, Neural network attention, Stacked RF, RF, DT

DGAs: CNN, RNN, SVM, RF, ResNet

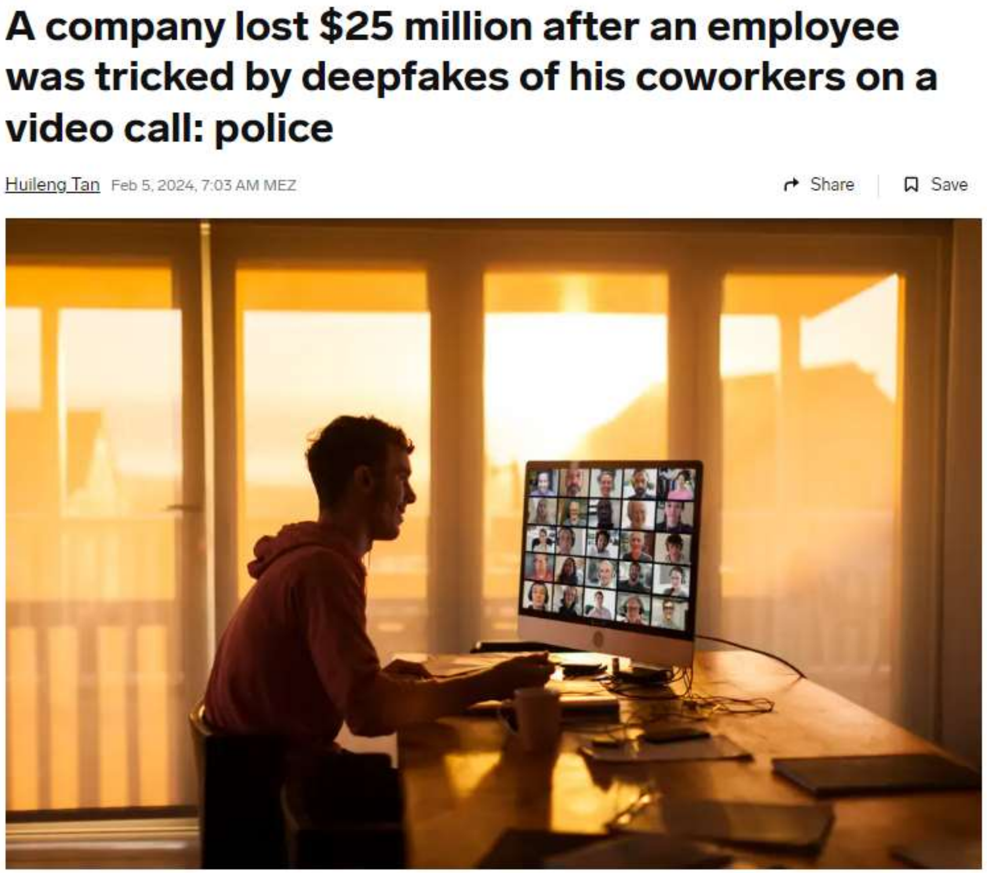
Source: https://arxiv.org/abs/2208.14937

3.) Applicability Framework for XAI in Cybersecurity



Source: https://arxiv.org/abs/2208.14937

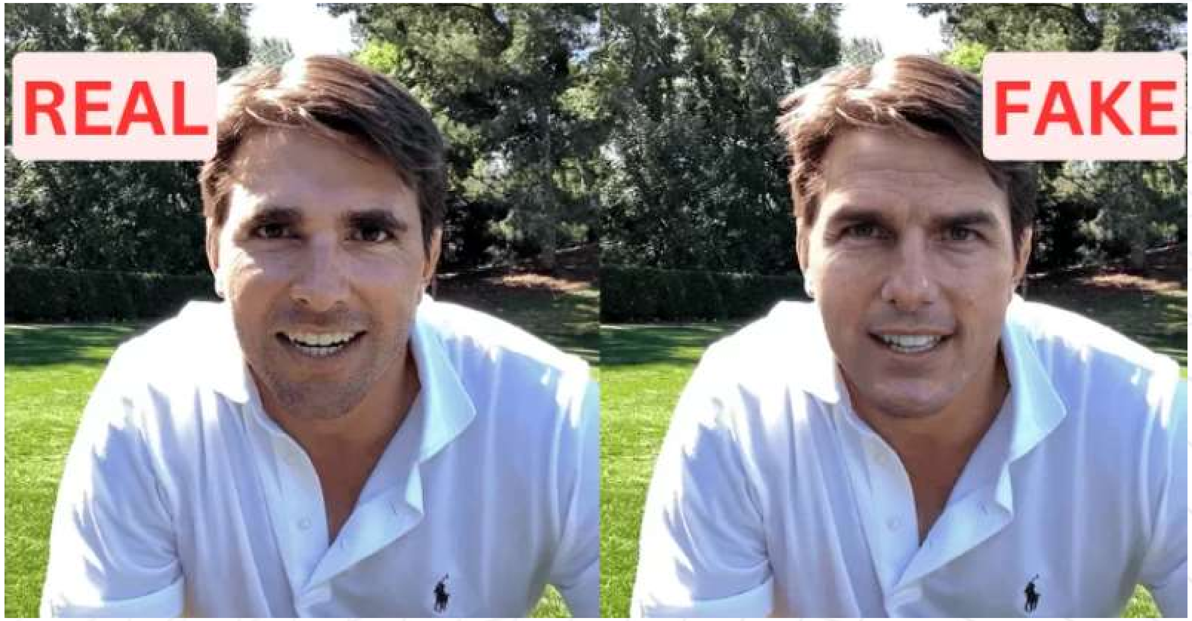
3.) Deep Fake Scam



A company lost $25 million after an employee was tricked by deepfakes of his coworkers on a video call: police

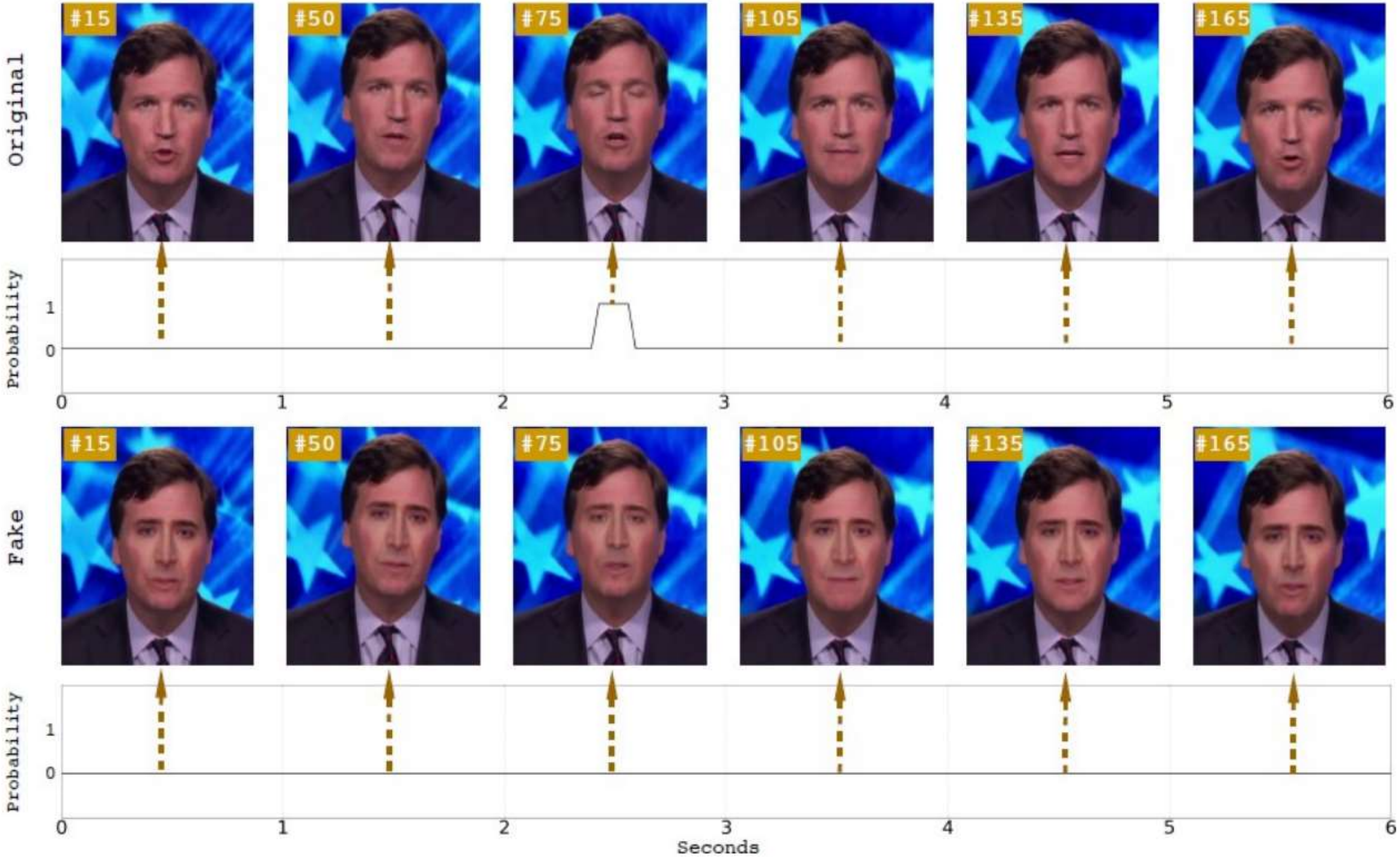
Huileng Tan Feb 5, 2024, 7:03 AM MEZ

Source: https://www.businessinsider.com/deepfake-coworkers-video-call-company-loses-millions-employee-ai-2024-2#:~:text



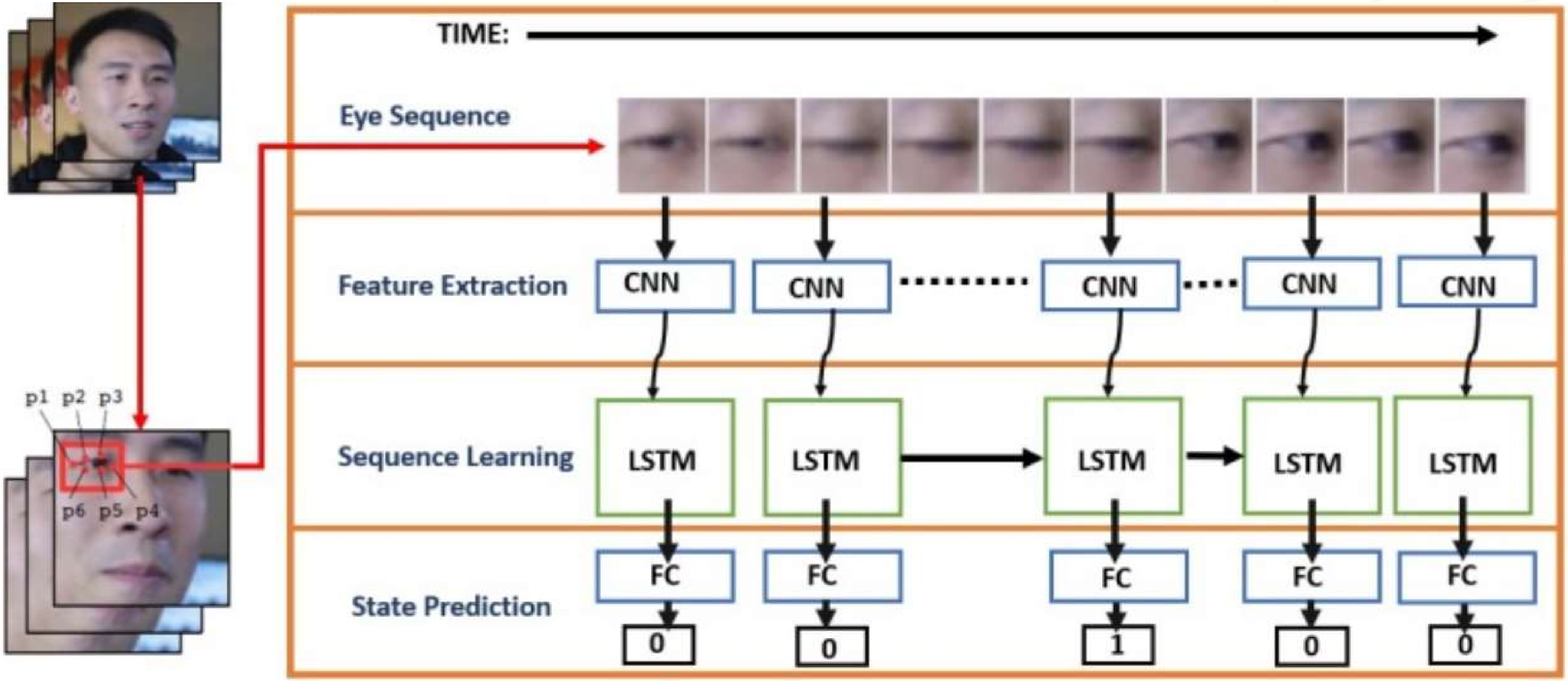
Source: https://digialps.com/deepfakefaceswap-any-video-in-just-one-click-with-this-incredible-tool/

3.) Deep Fake Detection



Source: https://www.mdpi.com/2079-9292/13/1/95

3.) Deep Fake – Learning Method



Source: https://www.mdpi.com/2079-9292/13/1/95

4.) Conclusion

We explained what XAI is and why it is used for machine learning analysis. How data bias can results in unexpected model behavior, including discrimination. Furthermore, we shoed how deply various machine learning models are employed in Cybersecurity attacks such as Malware, Spam, Botnet, Fraud, Phishing, DGA and Network Intrusion. We closed with the importance of Deep Fake detection as an upcoming problem, which is now here.

Dr. Lars Ruddigkeit

https://www.linkedin.com/in/drlr1/

https://www.linkedin.com/groups/4484376/

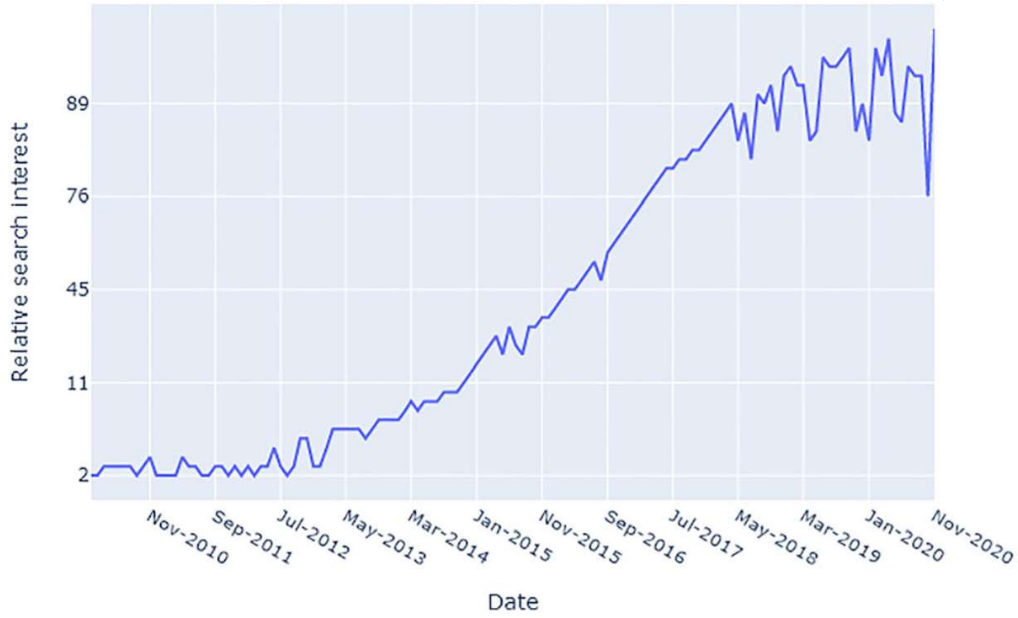
https://www.linkedin.com/groups/4968244/

THANK YOU
RAM
FRAU
NETWOR
INTRUSION
NEWORK
INNTRUSION
NIEH
OS
KNOE
BOTMET
BOTRAT
MALWARE

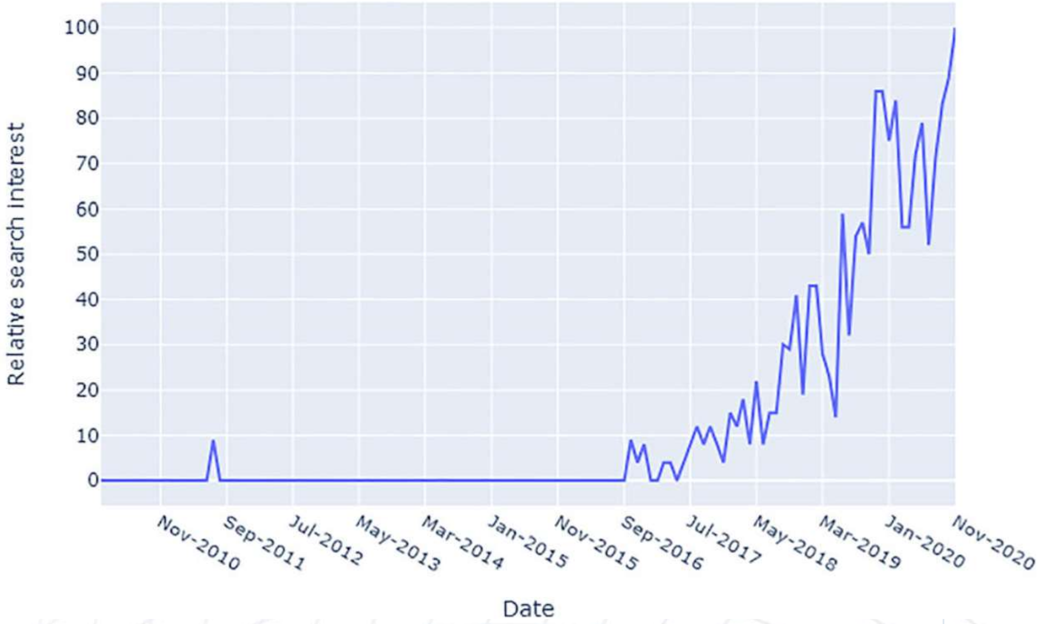
THANK YOU
Machino Leearning XI
Gyber

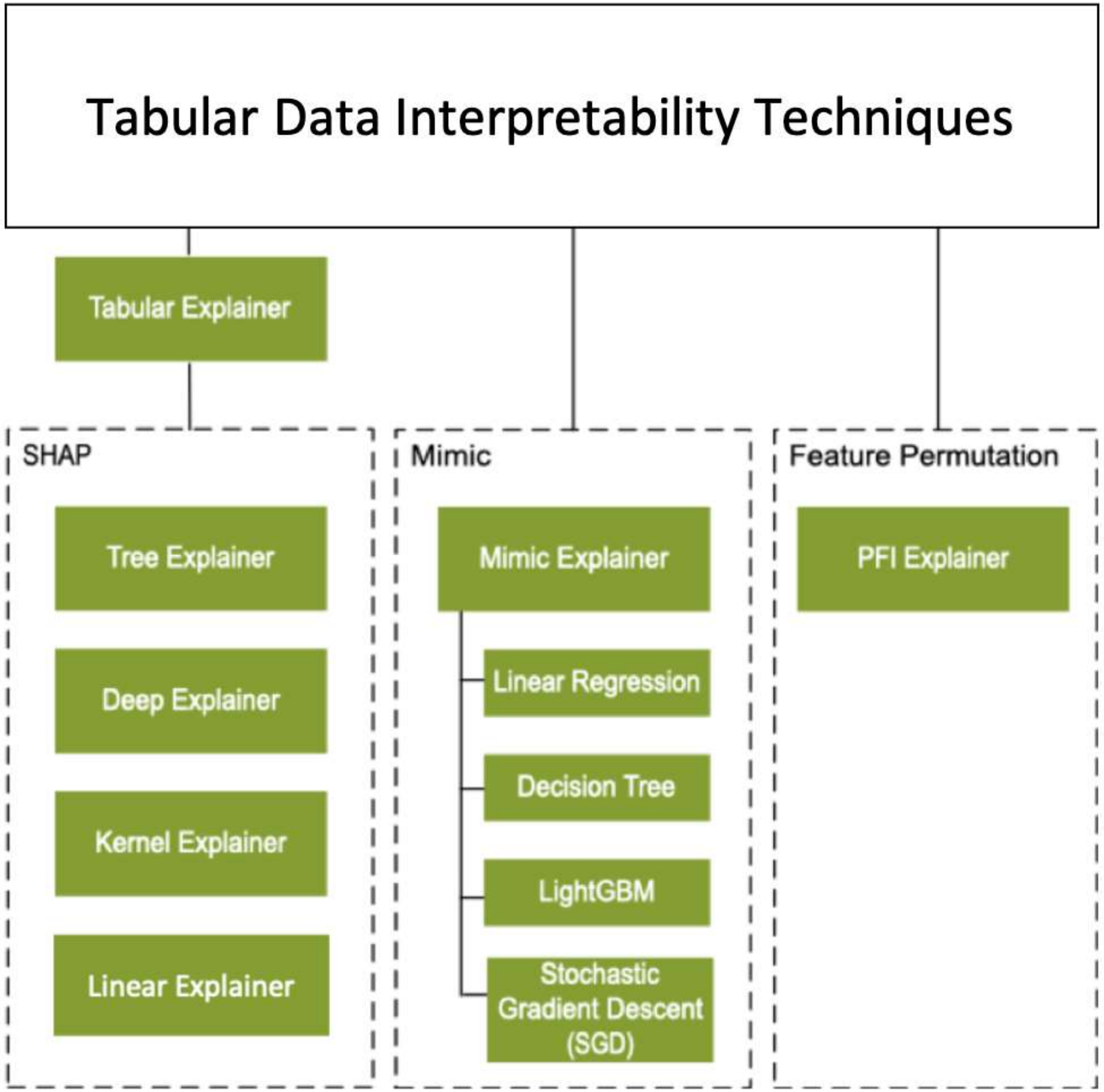
Evolution of interest

Deep Learning Search interest over time



XAI Search interest over time





Tabular Data Interpretability Techniques

Tabular Explainer

SHAP

Tree Explainer

Deep Explainer

Kernel Explainer

Linear Explainer

Mimic

Mimic Explainer

Linear Regression

Decision Tree

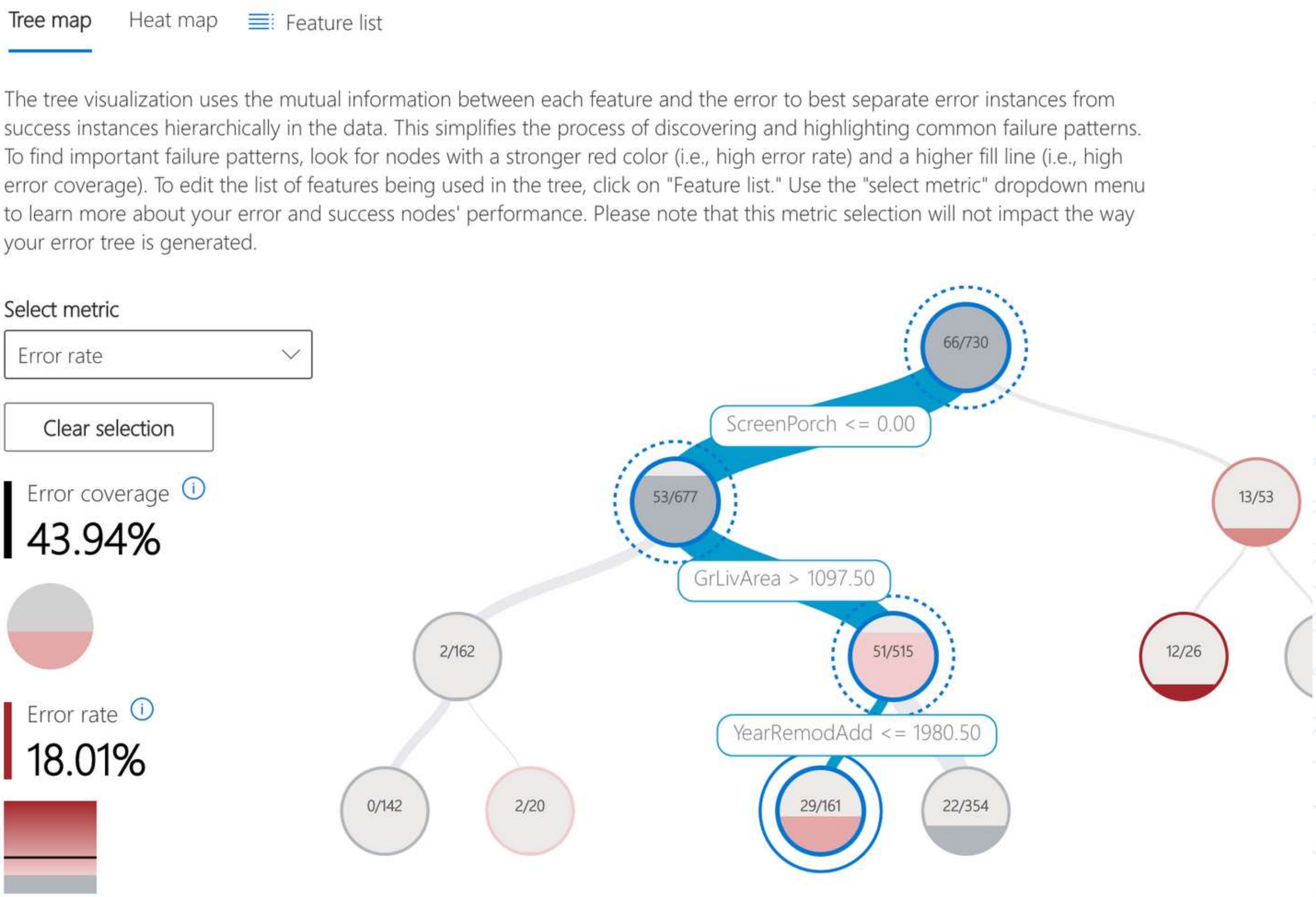
LightGBM

Stochastic Gradient Descent (SGD)

Feature Permutation

PFI Explainer

Source: https://learn.microsoft.com/de-de/azure/machine-learning/how-to-machine-learning-interpretability?view=azureml-api-2



The tree visualization uses the mutual information between each feature and the error to best separate error instances from success instances hierarchically in the data. This simplifies the process of discovering and highlighting common failure patterns. To find important failure patterns, look for nodes with a stronger red color (i.e., high error rate) and a higher fill line (i.e., high error coverage). To edit the list of features being used in the tree, click on "Feature list." Use the "select metric" dropdown menu to learn more about your error and success nodes' performance. Please note that this metric selection will not impact the way your error tree is generated.

Select metric

Error rate

Error coverage

43.94%

Error rate

18.01%

Source: https://learn.microsoft.com/de-de/azure/machine-learning/concept-error-analysis?view=azureml-api-2