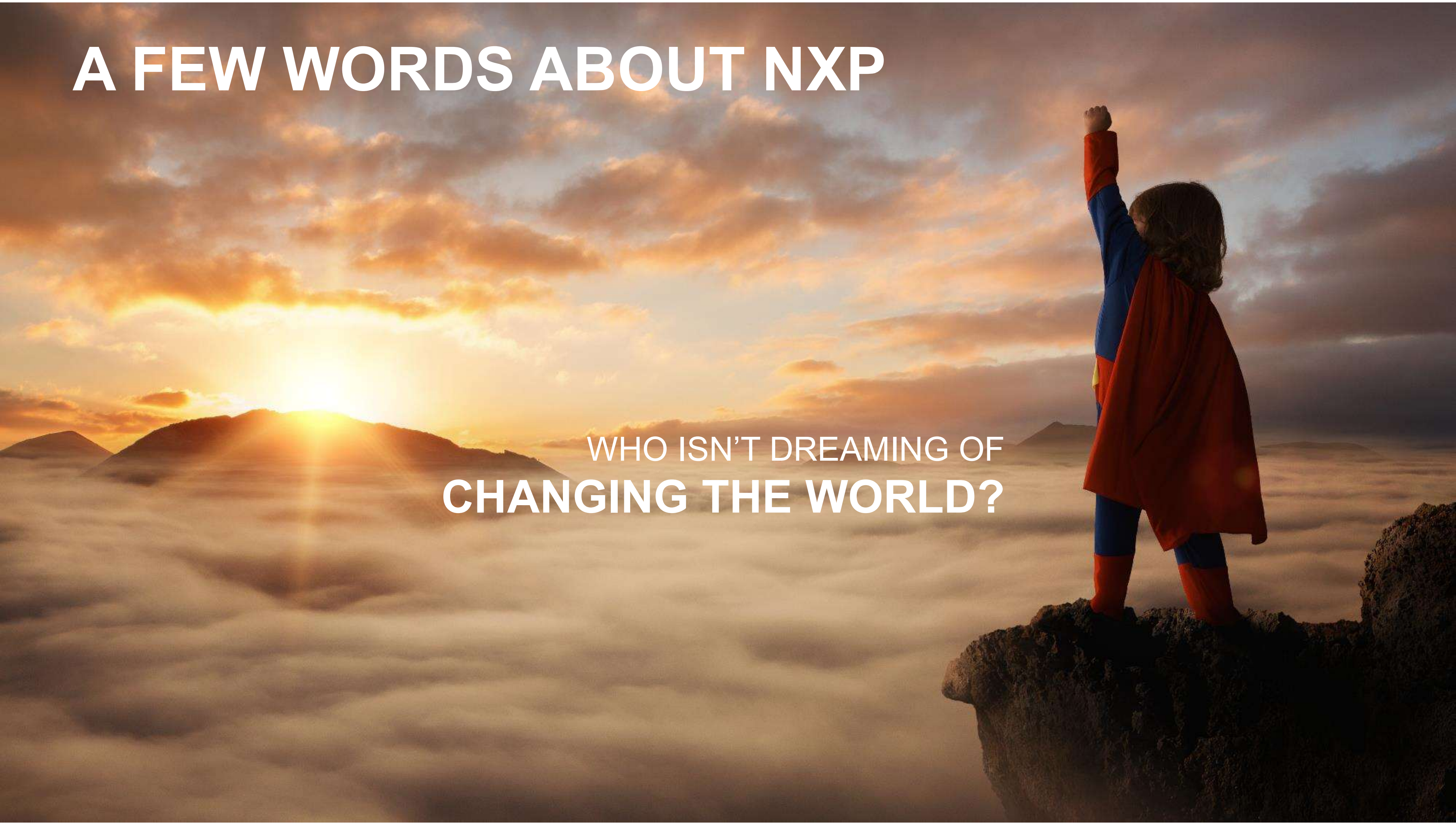
MACHINE LEARNING AND TRUST

Fari Assad-Eraghi
Sr. Vice President, Innovations
NXP Semiconductors

A FEW WORDS ABOUT NXP



WHO ISN’T DREAMING OF

CHANGING THE WORLD?

Secure Connections for a Smarter World

Enabling Technologies

Sense

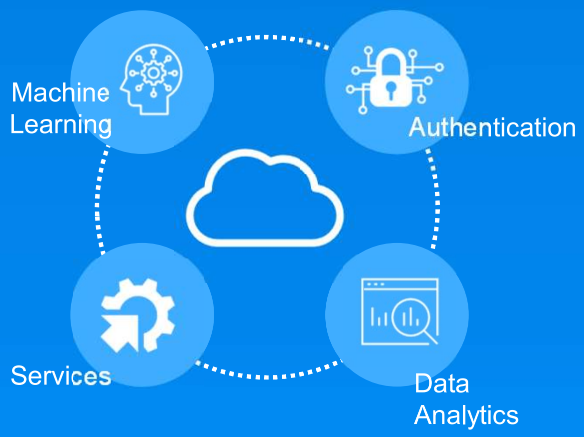
Think

Connect

Act

Secure Connections for a Smarter World

Cloud Infrastructure



Machine Learning

Authentication

Services

Data Analytics

Enabling Technologies

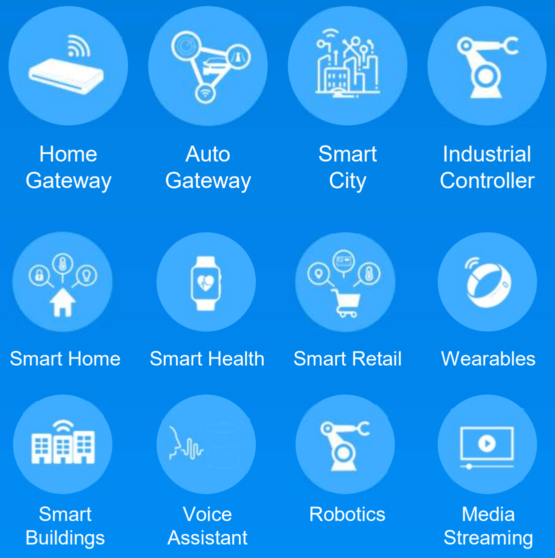
Sense

Think

Connect

Act

Edge to Node



Home Gateway

Auto Gateway

Smart City

Industrial Controller

Smart Home

Smart Health

Smart Retail

Wearables

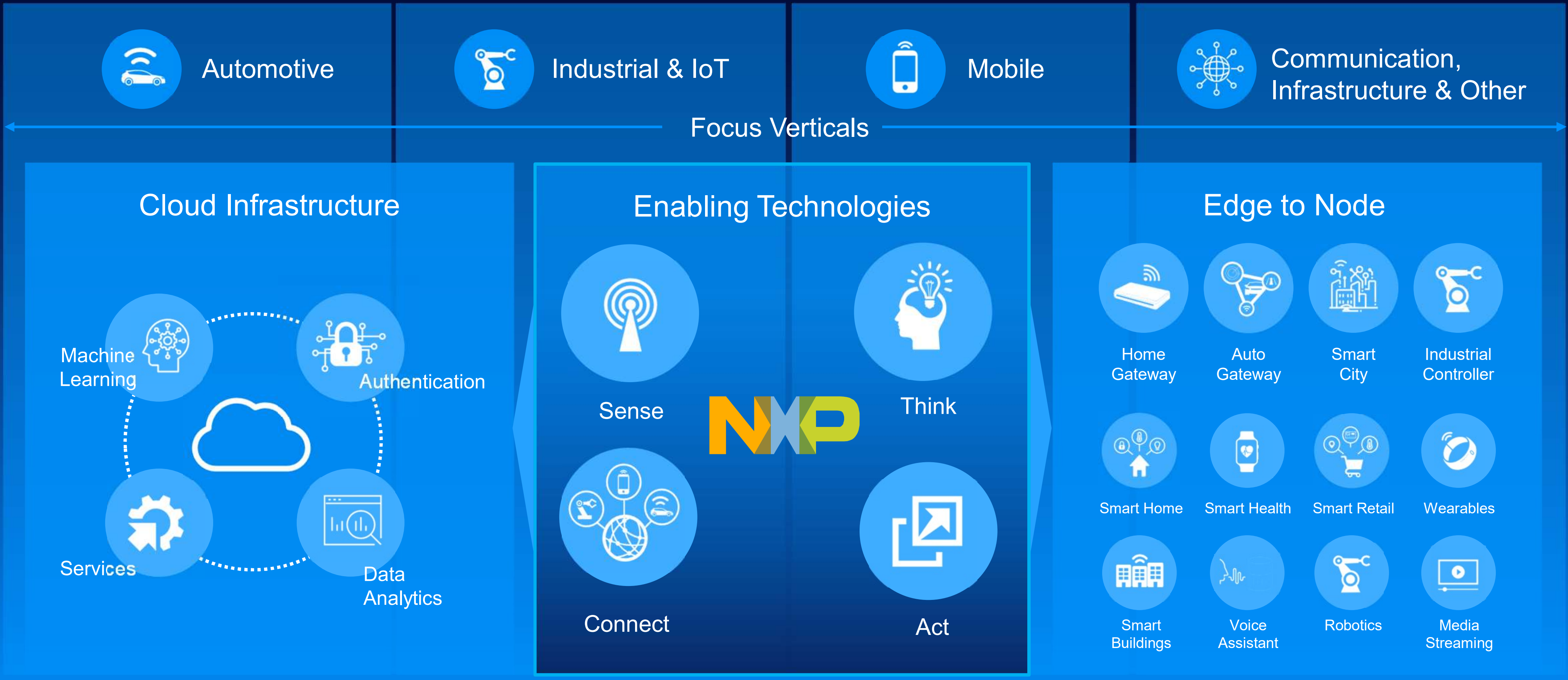
Smart Buildings

Voice Assistant

Robotics

Media Streaming

Secure Connections for a Smarter World



Automotive

Industrial & IoT

Mobile

Communication, Infrastructure & Other

Focus Verticals

Cloud Infrastructure

Machine Learning

Authentication

Services

Data Analytics

Enabling Technologies

Sense

Think

Connect

Act

Edge to Node

Home Gateway

Auto Gateway

Smart City

Industrial Controller

Smart Home

Smart Health

Smart Retail

Wearables

Smart Buildings

Voice Assistant

Robotics

Media Streaming

Basic Architecture of Smart Devices

Sense | Think | Act

Everything Aware
Failure-free sensing of analog environment

Everything Smart
Computing near end-nodes

Everything Connected
Heterogenous wireless, energy efficient

Everything Acting Efficiently
Intelligent actuators

Everything Safe & Secure
Safety, privacy, security

The Ultimate Edge Node & End Point Device

- High Bandwidth Connectivity
- Local Compute Capacity
- Advanced Sensor Hub
- Ingrained Security
- Advanced Displays

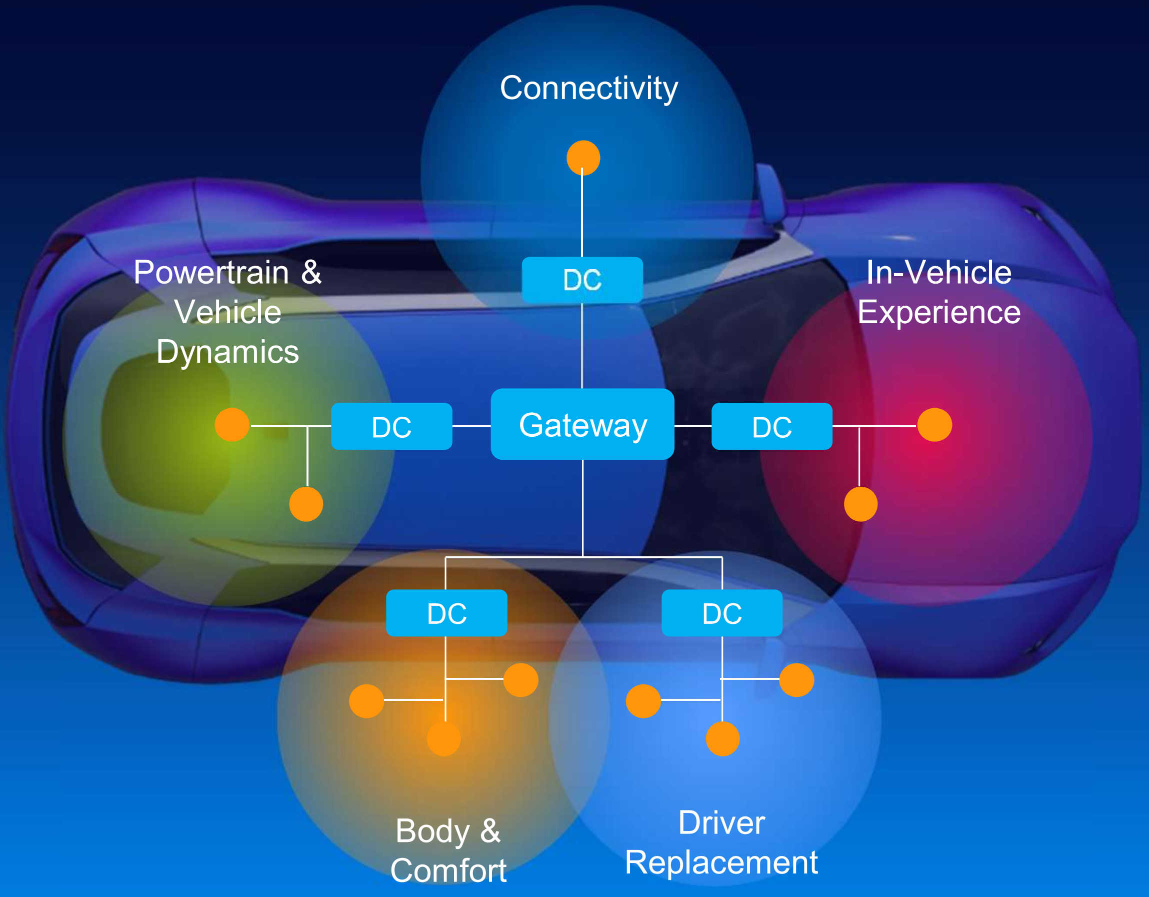
- Gateway Capability
- Connected Services
- Machine Learning
- Remote Access
- Advanced HMI

Smart Vehicles

Domain-based Architecture

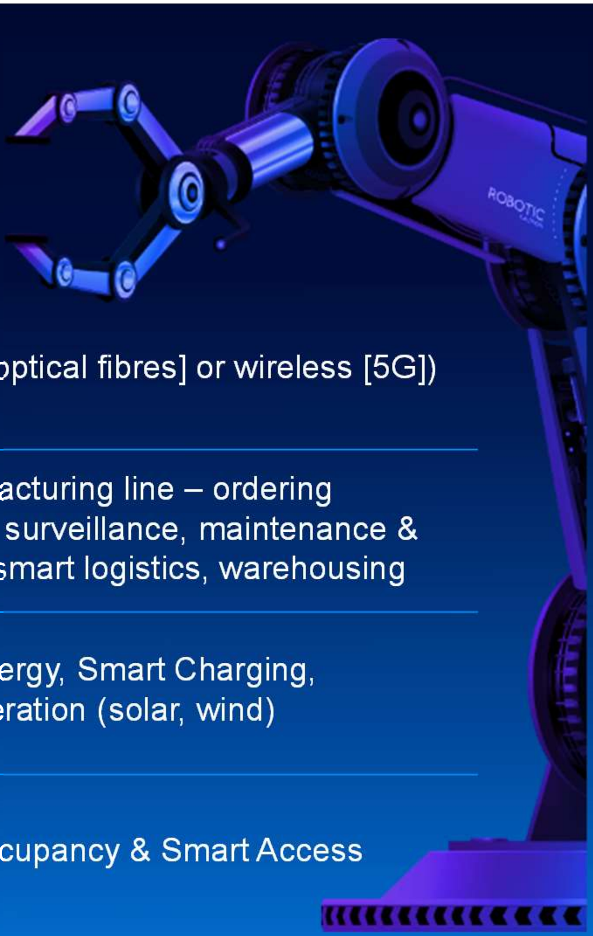
Secure Gateways & Networks

- Connectivity
- Driver Replacement
- Powertrain & Vehicle Dynamics
- Body & Comfort
- In-Vehicle Experience



Industry 4.0 – Smart Factory

Domain-based Architecture



Secure Gateways & Networks

Connectivity	Secure broadband connectivity (wired [optical fibres] or wireless [5G])
Autonomy	Self management function of the manufacturing line – ordering material, individualization of processes, surveillance, maintenance & cleaning, moving of production parts & smart logistics, warehousing
Energy Management	Smart Metering, Seamless supply of energy, Smart Charging, Management of renewable energy generation (solar, wind)
Environment & Facility	Light, Temperature, Humidity, Room Occupancy & Smart Access
Information Management, Human Machine Interaction	Noise cancellation Smart information for staff (warnings, real time performance & quality data) Smart training (maintenance manuals)

Smart Appliances – Example Cleaning Robot

Domain-based Architecture

Secure Gateways & Networks

Connectivity	Secure connectivity (Contactless [NFC], wired or wireless [narrow band, cellular, WiFi])
Autonomy	Self-driving and self-management, machine learning capability on power efficient µc and sensor architecture
Energy Management & Motion	Highly efficient power management
Body Electronics & Controls	Smart (wireless) charging
Human Machine Interface	Gesture and speech recognition

NXP Enabling Machine Learning Revolution

Voice Processing

Smart Sense & Control

Gesture Control

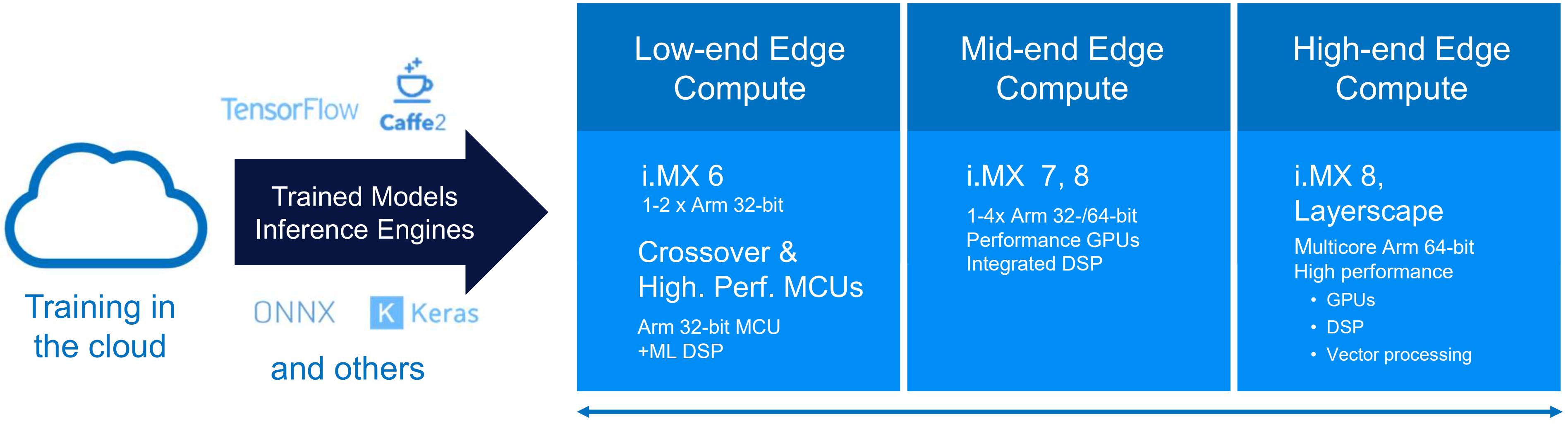
Active Object Recognition

Personal / Property

Home Environment

Multi-camera Observation

Augmented Reality



Training in the cloud

TensorFlow

Caffe2

Trained Models Inference Engines

ONNX

Keras

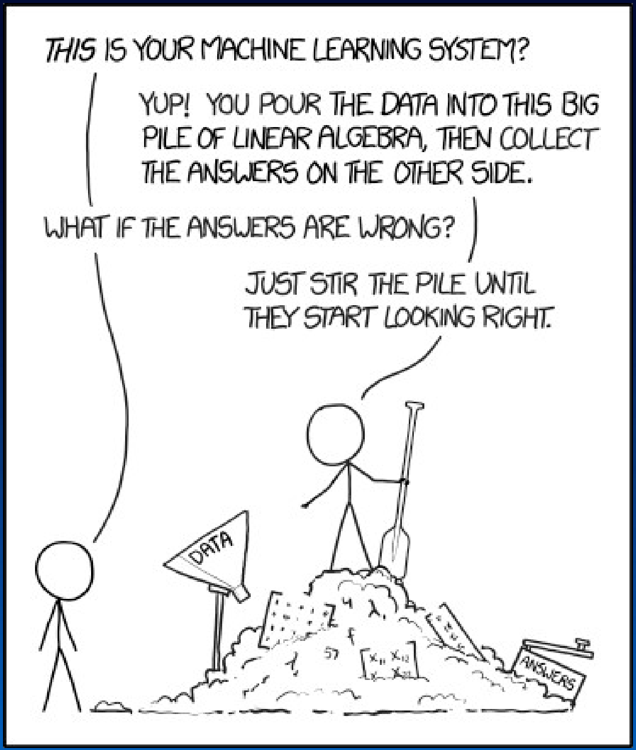
and others

Low-end Edge Compute	Mid-end Edge Compute	High-end Edge Compute
i.MX 6 1-2 x Arm 32-bit Crossover & High. Perf. MCUs Arm 32-bit MCU +ML DSP	i.MX 7, 8 1-4x Arm 32-/64-bit Performance GPUs Integrated DSP	i.MX 8, Layerscape Multicore Arm 64-bit High performance • GPUs • DSP • Vector processing

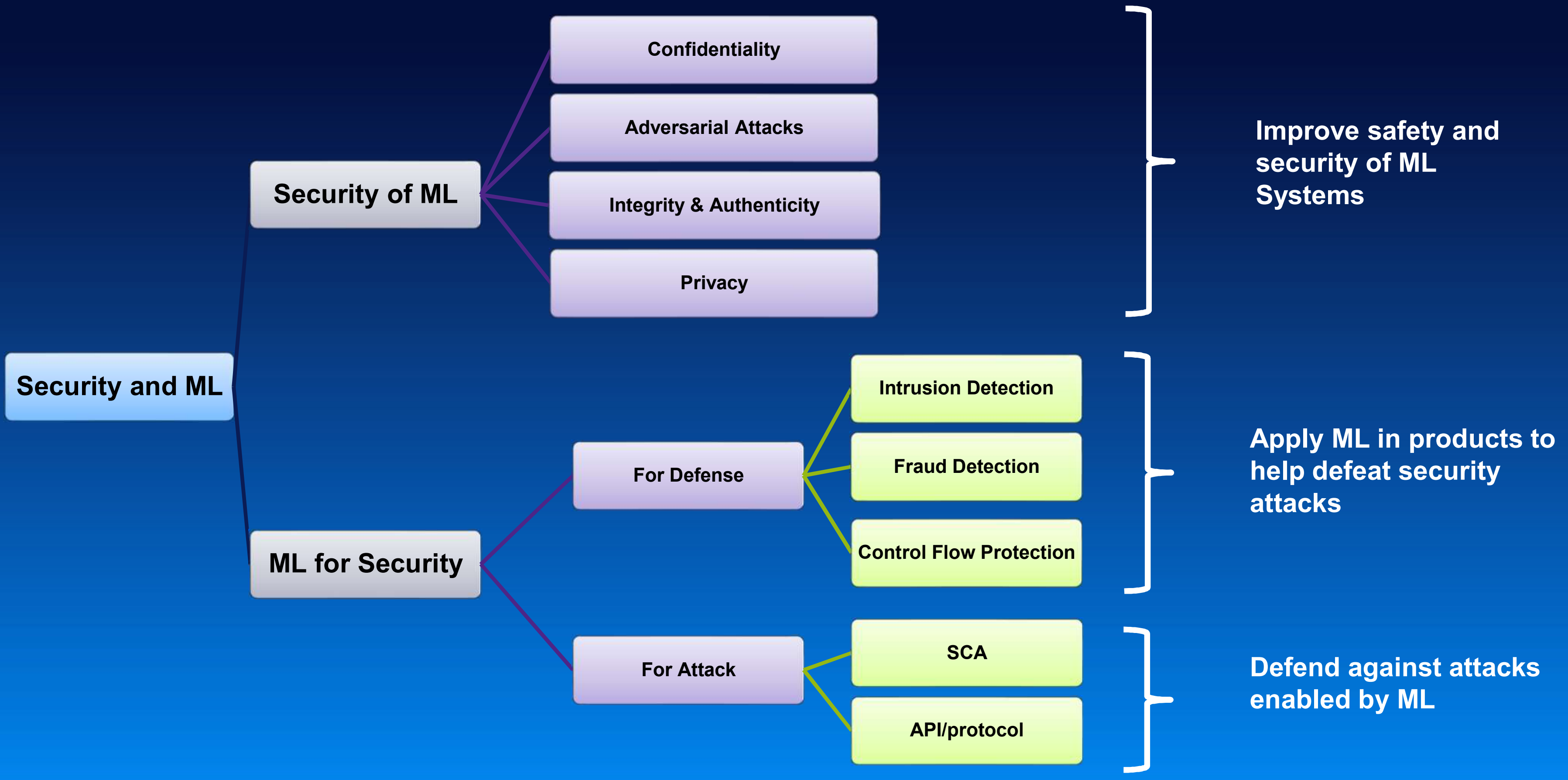
Scalable & optimized inference engines across Embedded Processing continuum

TRUST IS THE ROOT OF ALL THESE SOLUTIONS

- Machine Learning will transform all aspects of global economy
- Breakneck advances in computer science and algorithms, but also a renaissance in HW innovations
- Vast engineering resources focused on improving performance & power efficiency
  - Both ends of spectrum – From massive data centers to IoT devices
- Until recently, little attention to the ***Trust of ML***
- The Trust Umbrella covers security, privacy, interpretability, and fairness of ML


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Where Machine Learning and Security & Privacy Intersect



Security and ML

Security of ML

Confidentiality

Adversarial Attacks

Integrity & Authenticity

Privacy

Improve safety and security of ML Systems

ML for Security

For Defense

Intrusion Detection

Fraud Detection

Control Flow Protection

Apply ML in products to help defeat security attacks

For Attack

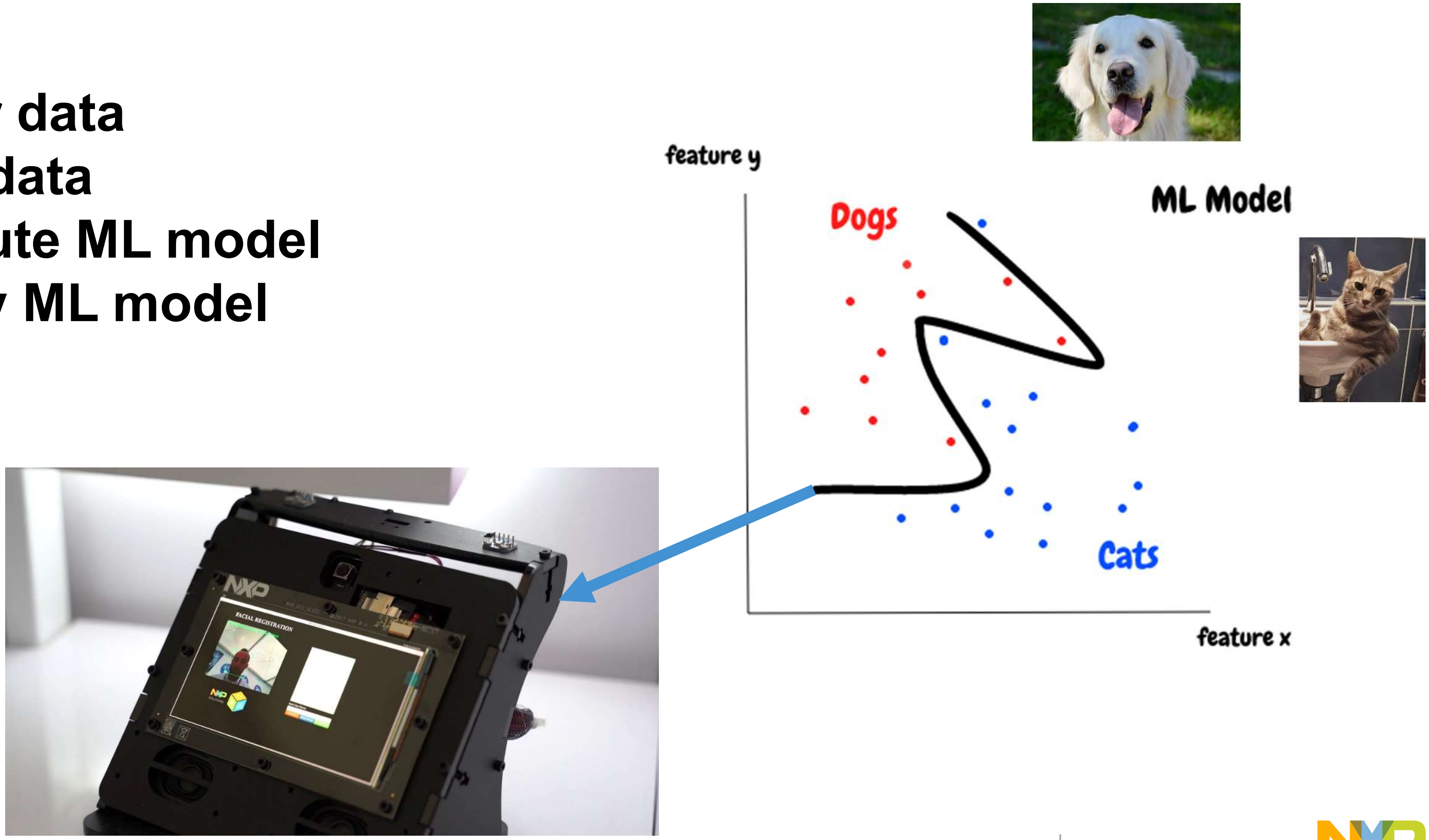
SCA

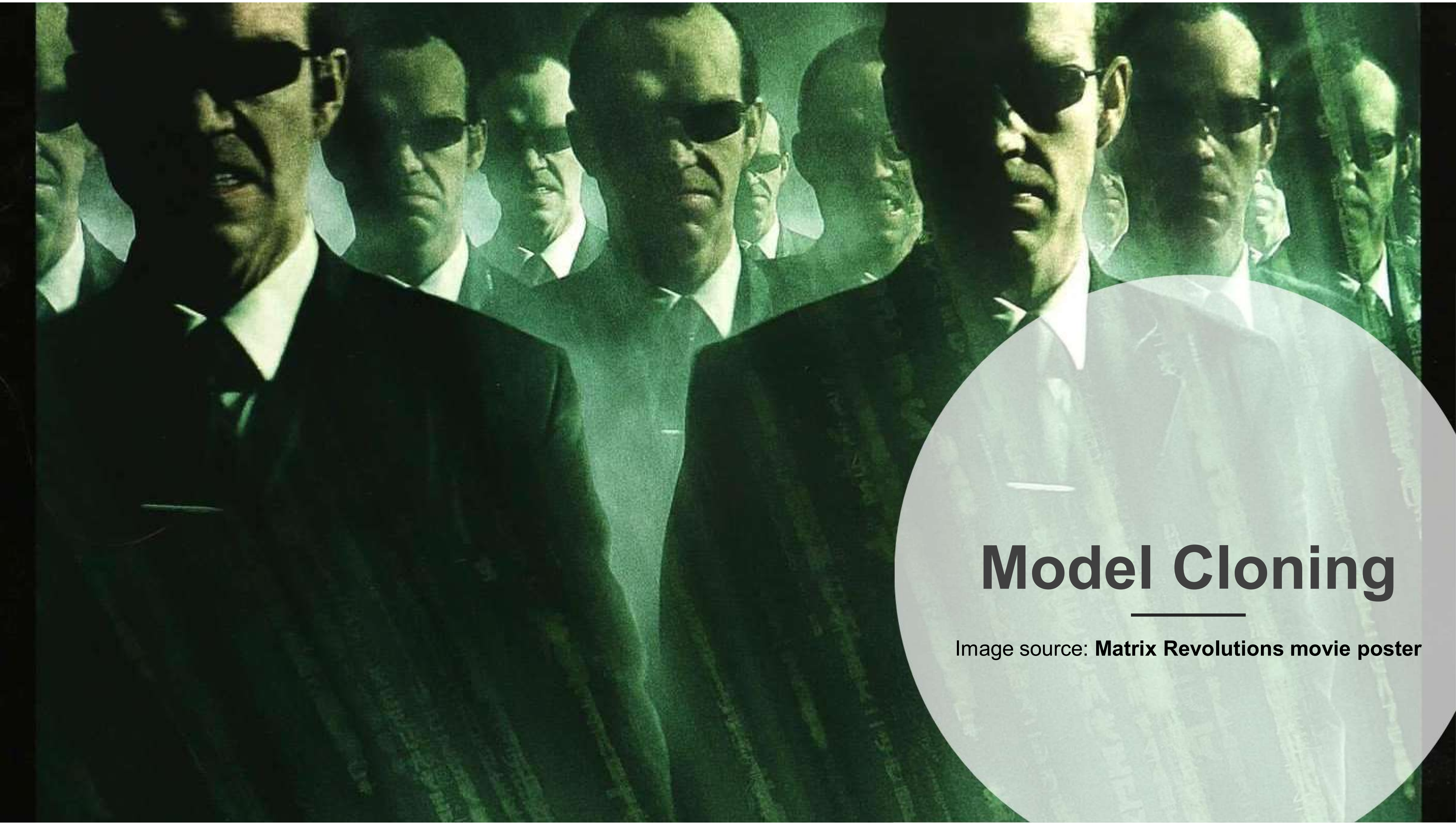
API/protocol

Defend against attacks enabled by ML

Four-step Plan for Making Smart Devices

1. **Gather data**
2. **Label data**
3. **Compute ML model**
4. **Deploy ML model**

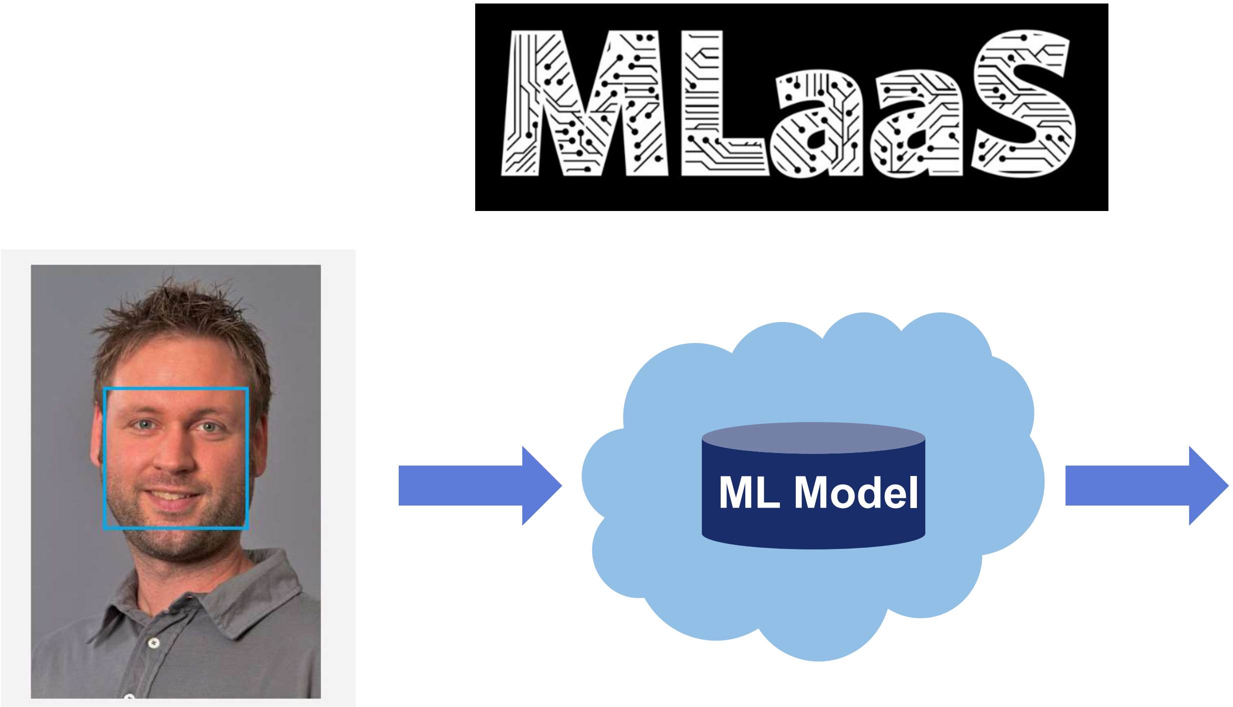


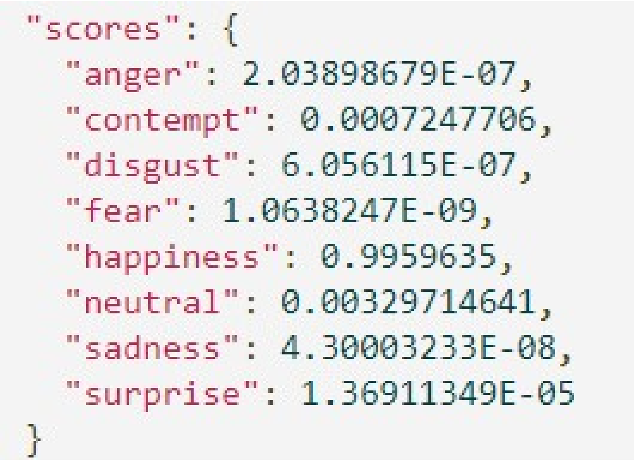


Model Cloning

Image source: **Matrix Revolutions movie poster**

Example: Microsoft Azure Emotion Recognition



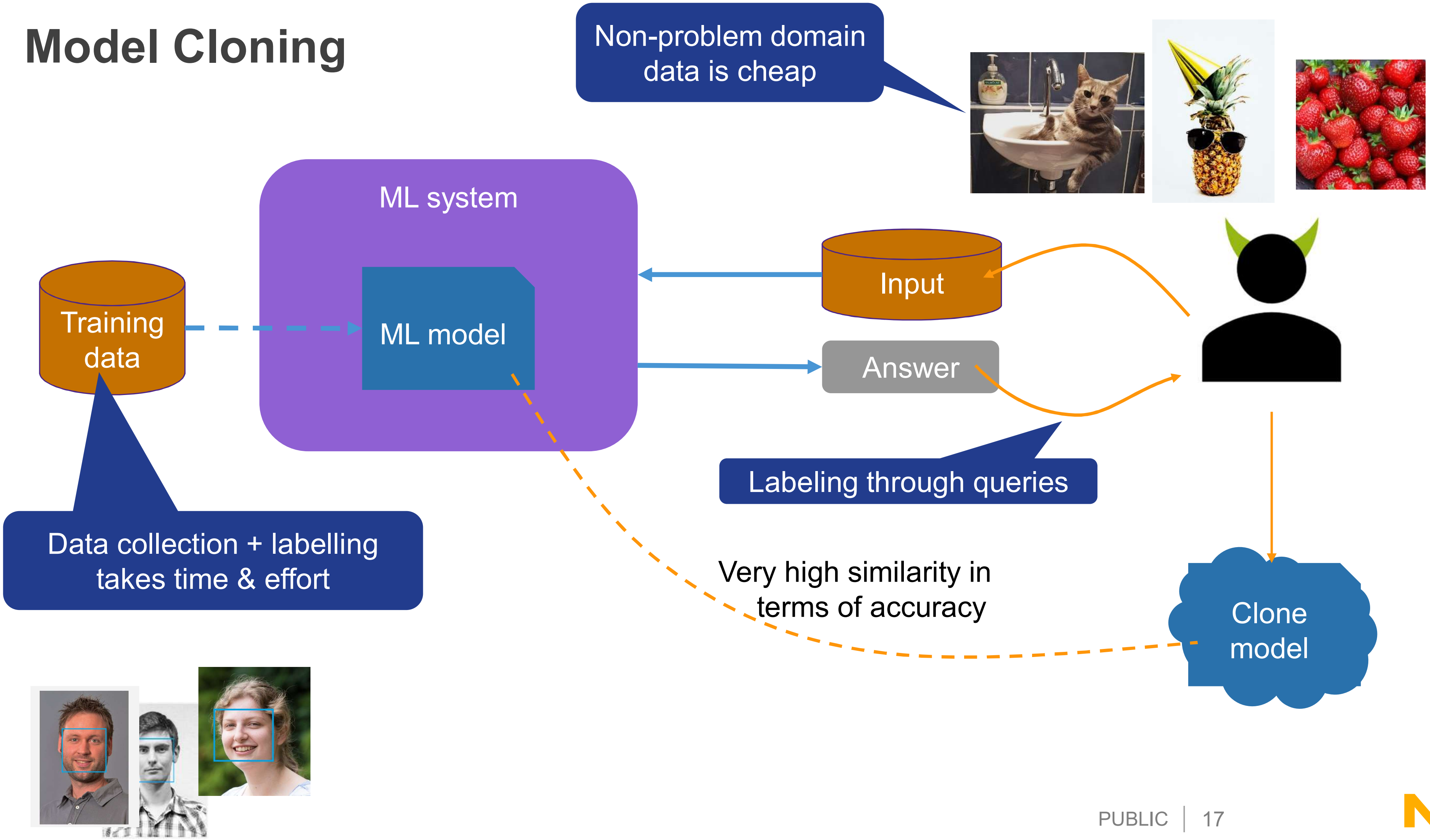


```
"scores": {
  "anger": 2.03898679E-07,
  "contempt": 0.0007247706,
  "disgust": 6.056115E-07,
  "fear": 1.0638247E-09,
  "happiness": 0.9959635,
  "neutral": 0.00329714641,
  "sadness": 4.30003233E-08,
  "surprise": 1.36911349E-05
}
```

Clone made for < $350 with 98.6% accuracy of original

- https://azure.microsoft.com/en-us/services/cognitive-services/emotion
- Tramèr, Zhang, Juels, Reiter, Ristenpart: *Stealing Machine Learning Models via Prediction APIs*. In *USENIX Security Symposium*, 2016.
- Correia-Silva, Rodrigues, Berriel, Badue, de Souza, Oliveira-Santos. *Copycat CNN: Stealing Knowledge by Persuading Confession with Random Non-Labeled Data*. In *International Joint Conference on Neural Networks (IJCNN),* 2018.

Model Cloning



Non-problem domain data is cheap

ML system

ML model

Training data

Input

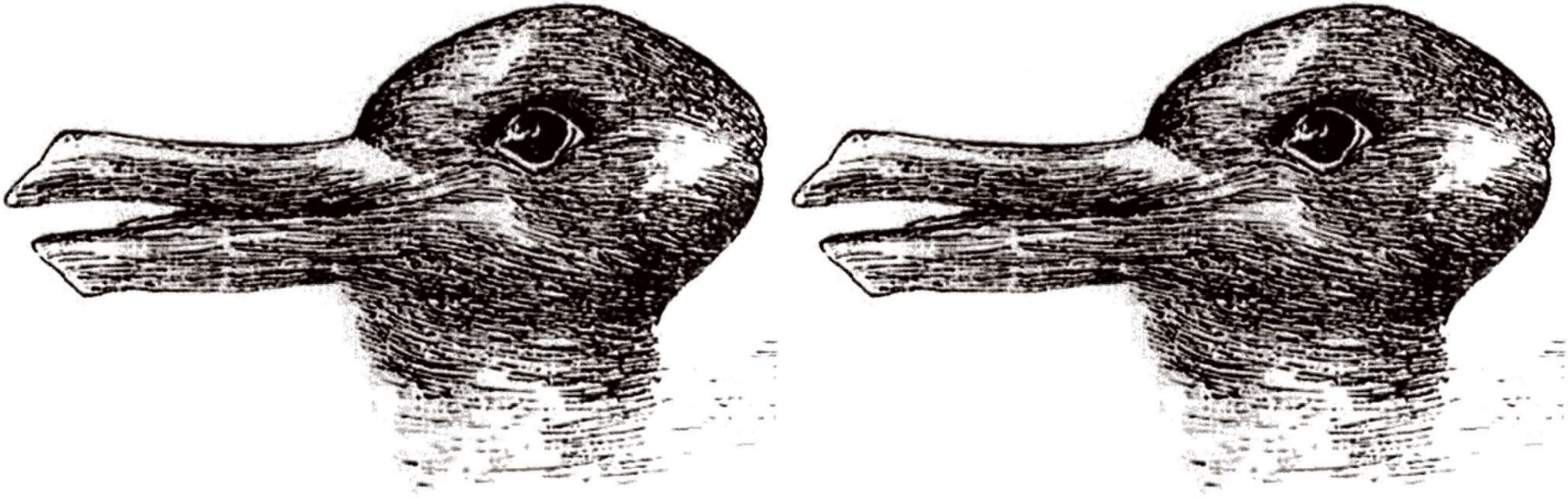
Answer

Labeling through queries

Data collection + labelling takes time & effort

Very high similarity in terms of accuracy

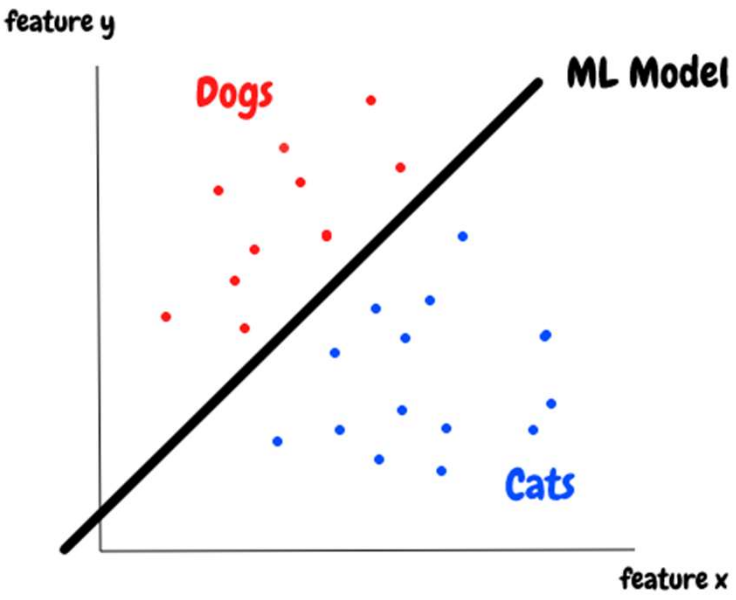
Clone model

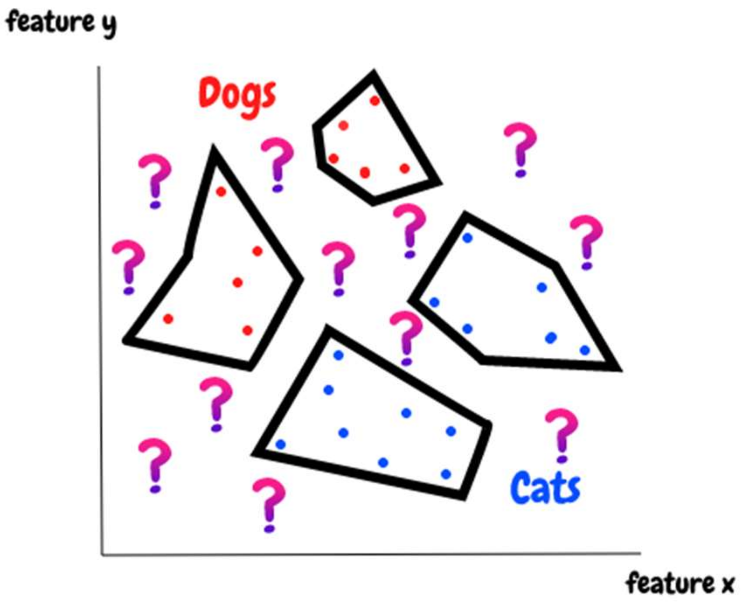


Adversarial Examples

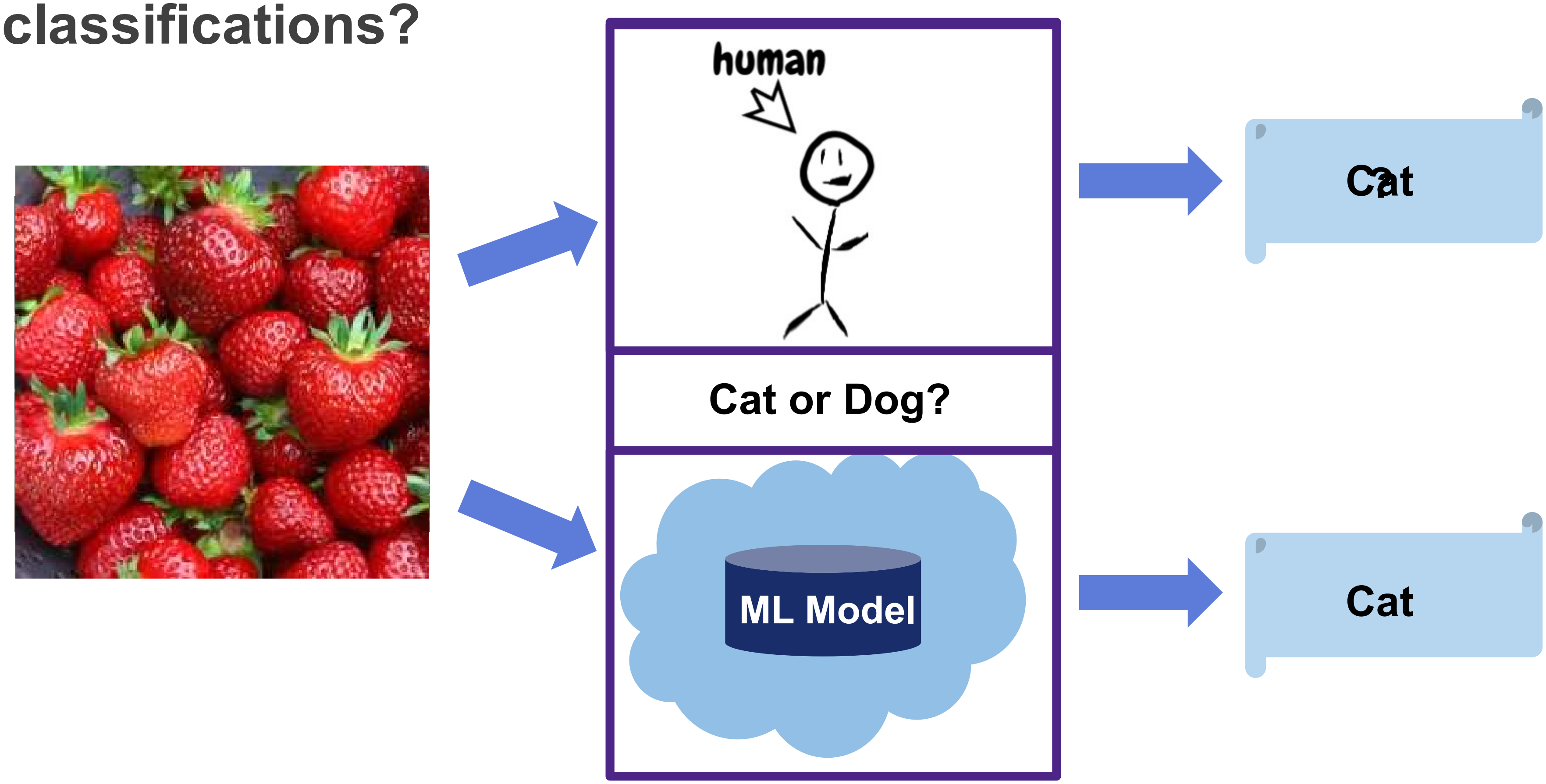
“Optical Illusions” for Machines

Image by artist Joseph Jastrow, published in 1899 in Popular Science Monthly

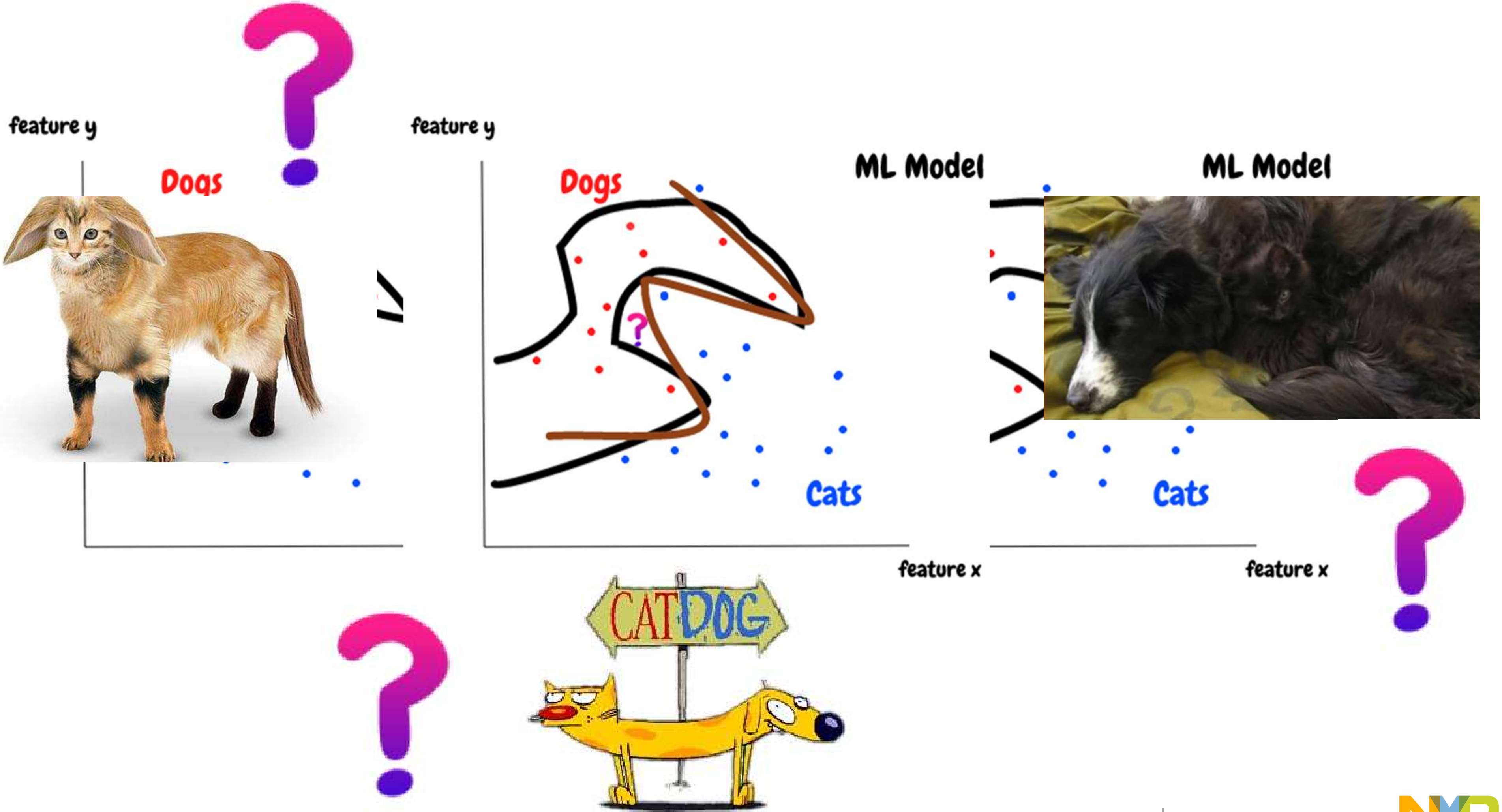

feature y
Dogs
ML Model
Cats
feature x


feature y
Dogs
Cats
feature x

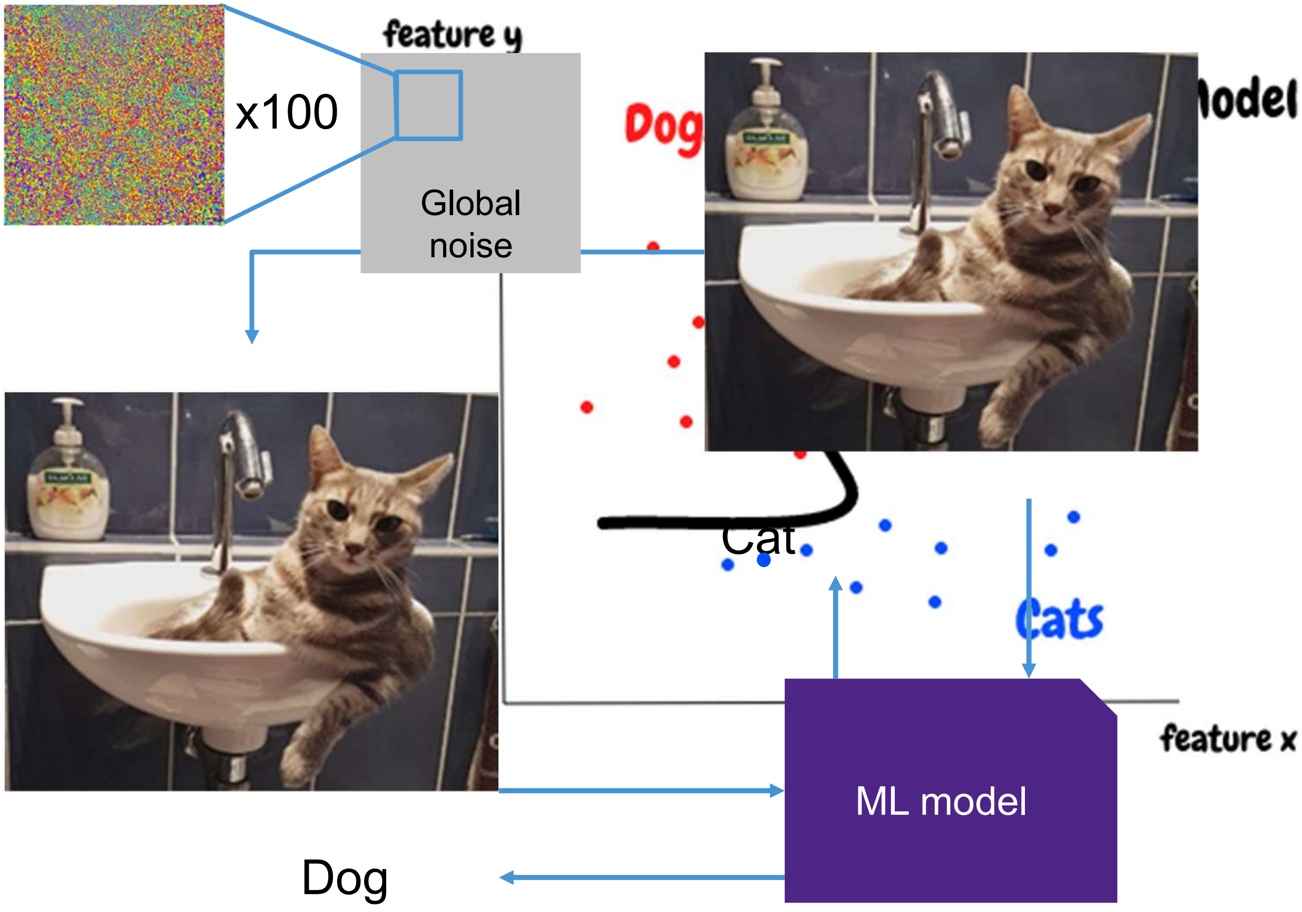
Misclassifications?

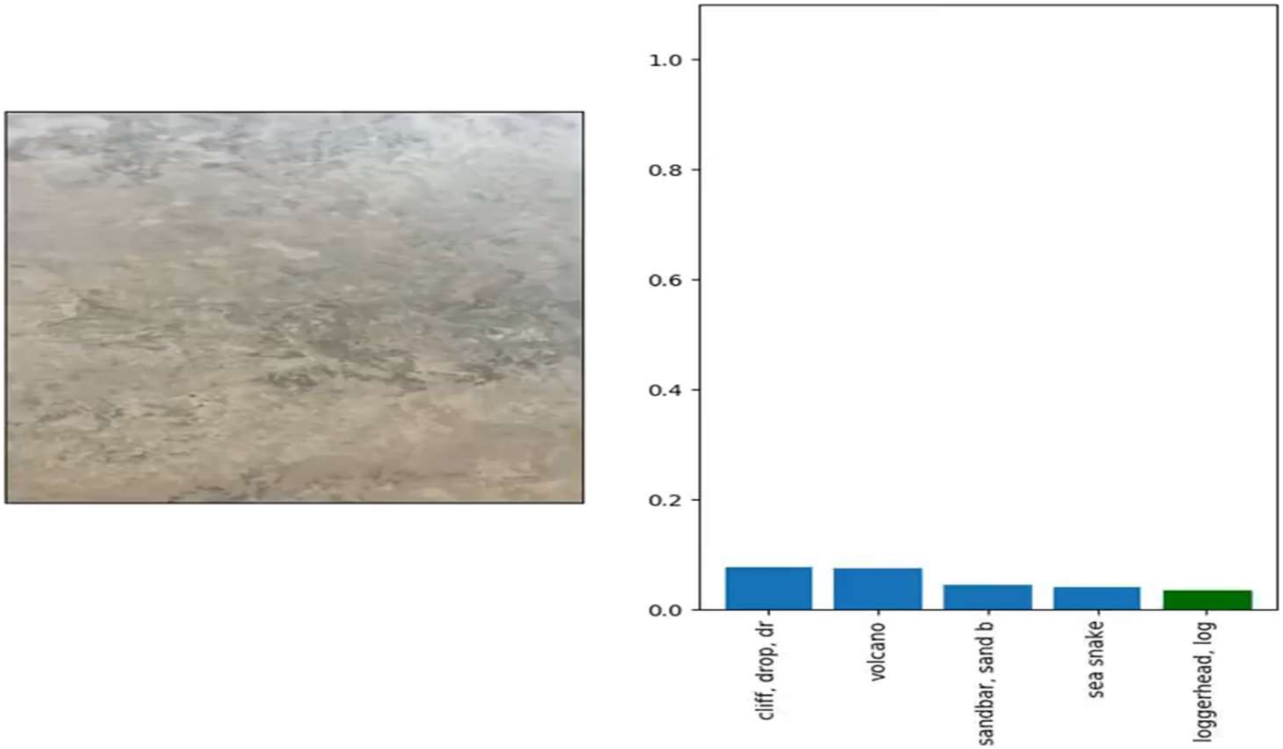


- Biggio, Corona, Maiorca, Nelson, Srndic, Laskov, Giacinto, Roli: Evasion attacks against machine learning at test time. In Machine Learning and Knowledge Discovery in Databases, 2013.
- Goodfellow, Shlens, Szegedy: Explaining and harnessing adversarial examples. In arXiv preprint 2014
- Szegedy, Vanhoucke, Ioffe, Shlens, Wojna: Rethinking the inception architecture for computer vision. In IEEE conference on computer vision and pattern recognition, 2016.


feature y
Dogs
feature y
Dogs
?
ML Model
ML Model
Cats
Cats
feature x
feature x
CAT DOG

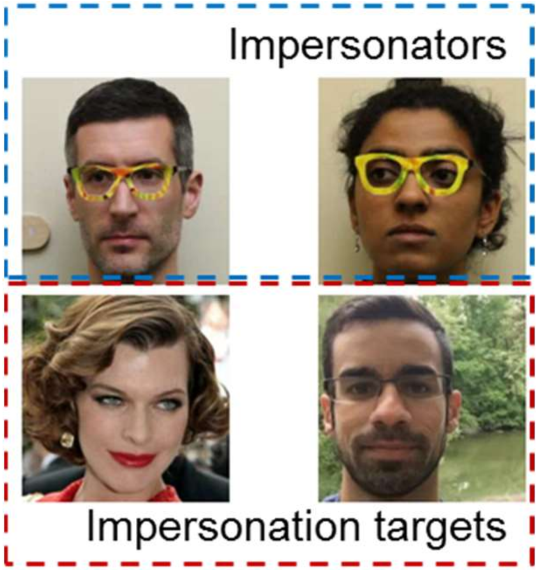
Adversarial Examples





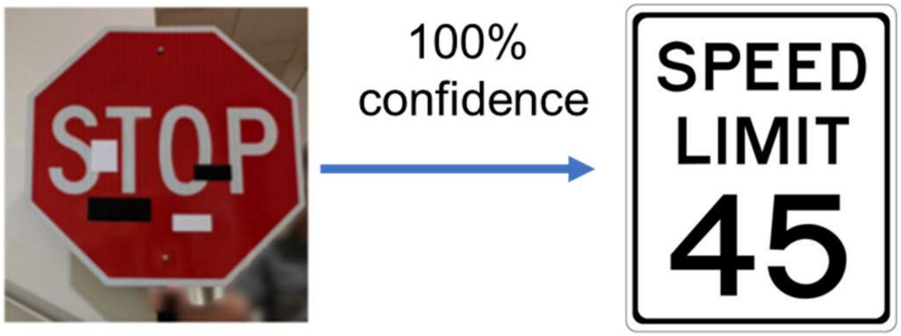
Movie from: Athalye, Engstrom, Ilyas, and Kwok: **Synthesizing Robust Adversarial Examples**. In *International Conference on Machine Learning*, 2018.

Security



Sharif, Bhagavatula, Bauer, Reiter: *Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition*. In ACM *SIGSAC 2016*

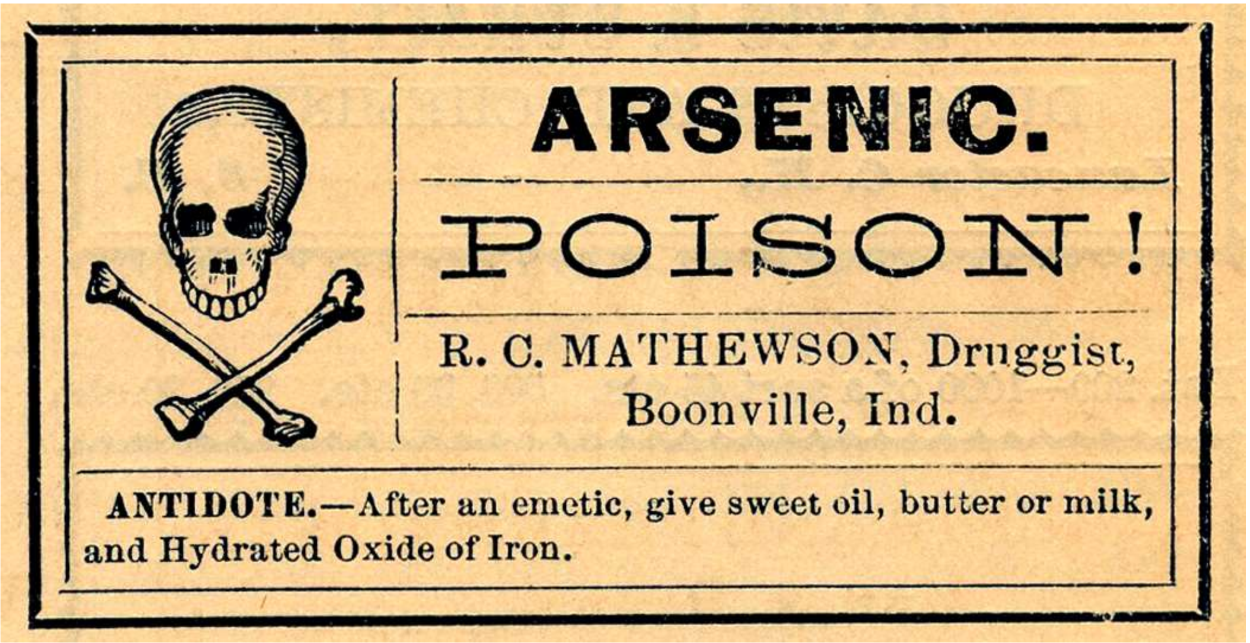
Safety



Eykholt, Evtimov, Fernandes, Li, Rahmati, Xiao, Prakash, Kohno, Song: *Robust Physical-World Attacks on Deep Learning Visual Classification*. In *IEEE Computer Vision and Pattern Recognition* 2018.

Impact in Practice

Data Poisoning



ARSENIC.

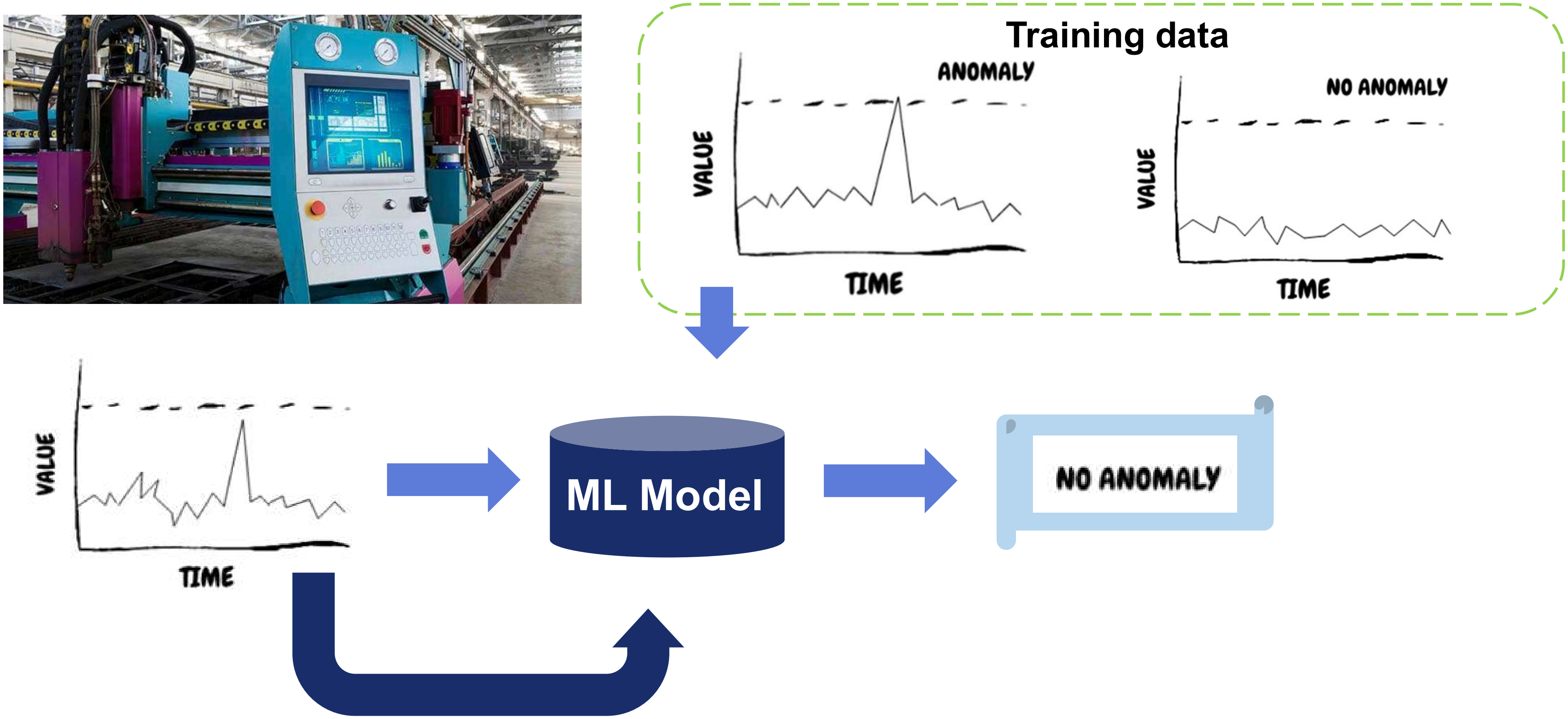
POISON!

R. C. MATHEWSON, Druggist,
Boonville, Ind.

ANTIDOTE.—After an emetic, give sweet oil, butter or milk, and Hydrated Oxide of Iron.

Barreno, Nelson, Sears, Joseph, and Tygar: *Can machine learning be secure?* In ACM CCS 2006.

Incremental learning | Anomaly detection in practice

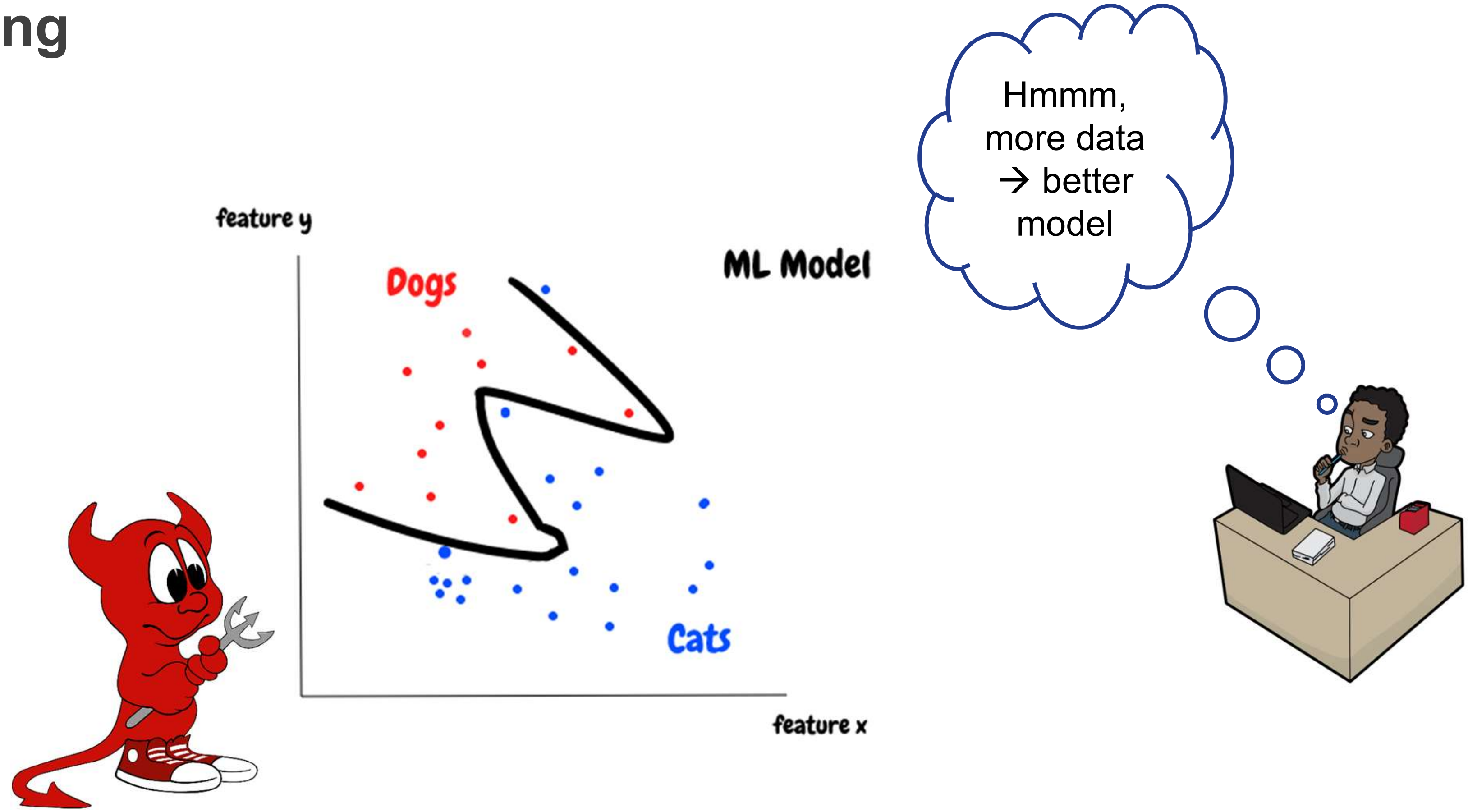


Training data

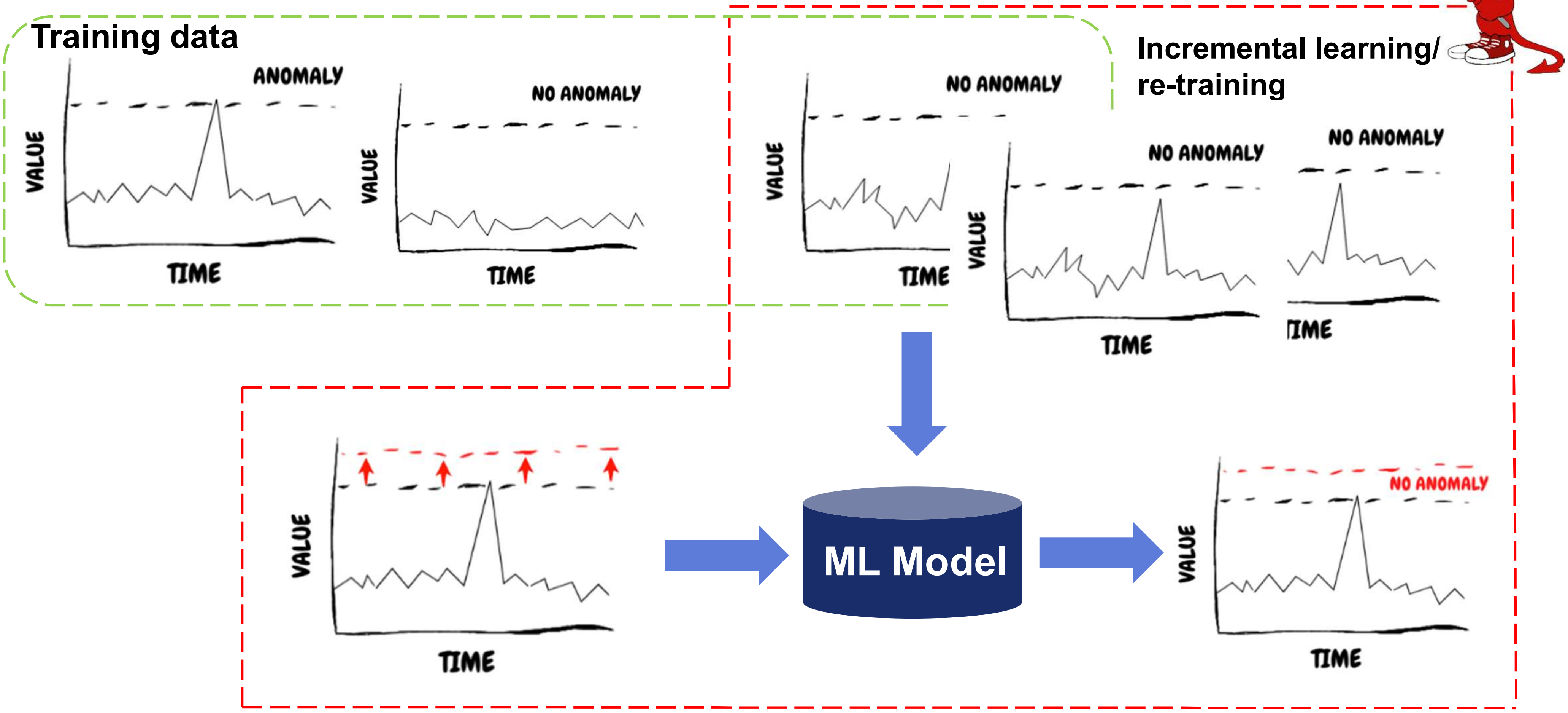
ML Model

Re-training/incremental learning

Data Poisoning



Data poisoning in anomaly detection



Training data

Incremental learning/ re-training

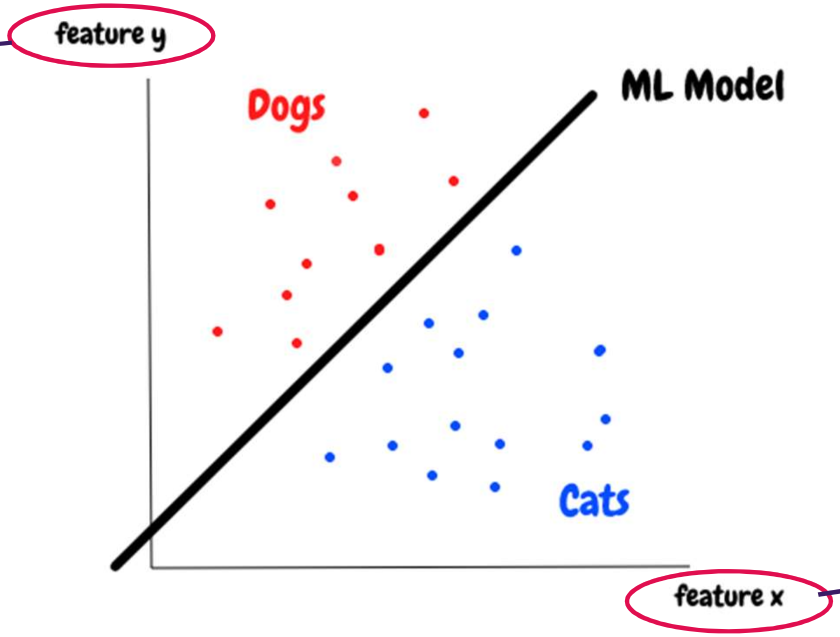
ML Model

Model Explainability

Source: Christoph Molnar

What Does an ML Model Learn?

size
# teeth
# tennis balls

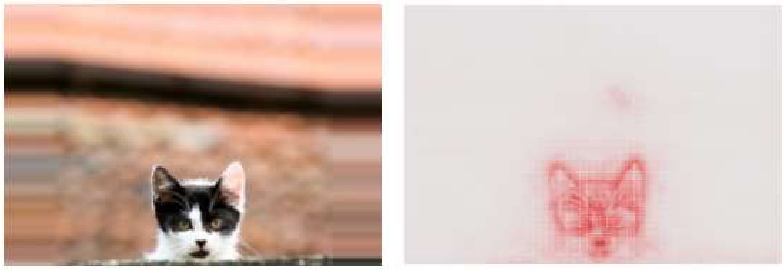


tail length
pupils
amount of yarn

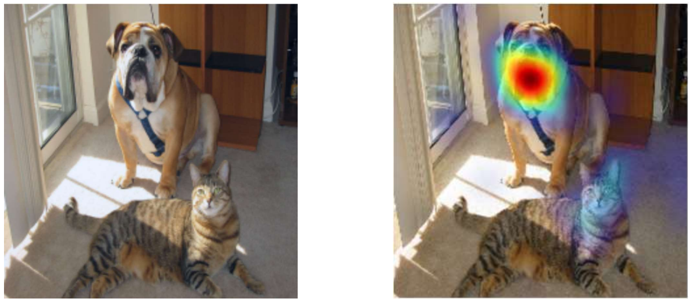
Interpretability → Explainability

ML training algorithm learns features automatically *without* knowing what they represent

The Good

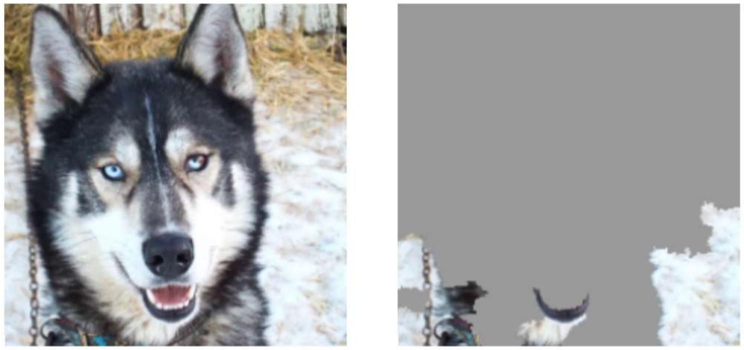


Montavon, Lapuschkin, Binder, Samek, Müller. "Explaining nonlinear classification decisions with deep taylor decomposition." *Pattern Recognition* 2017



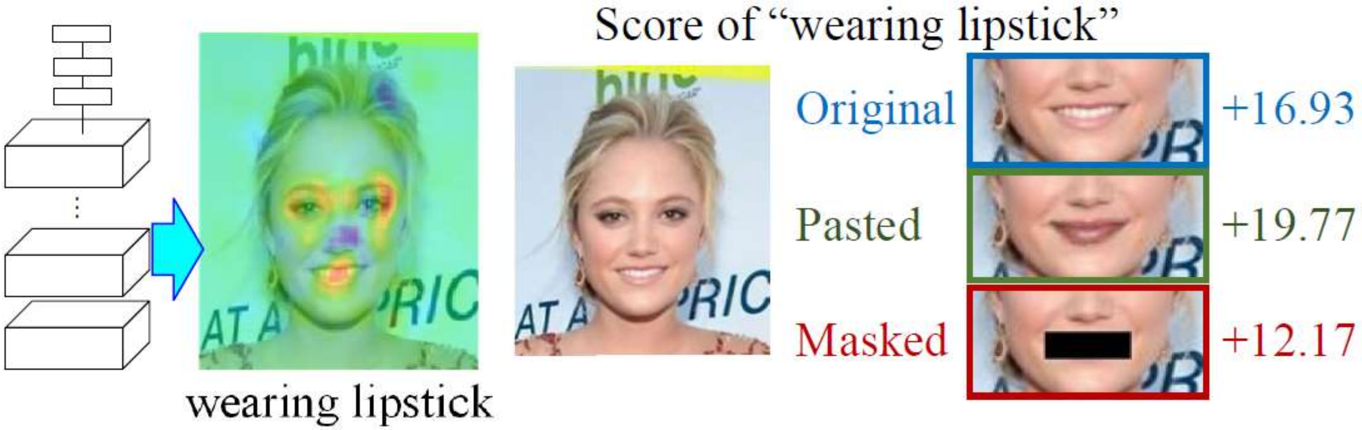
Selvaraju, Cogswell, Vedantam, Parikh, Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization." In *IEEE International Conference on Computer Vision*, 2017

The Bad

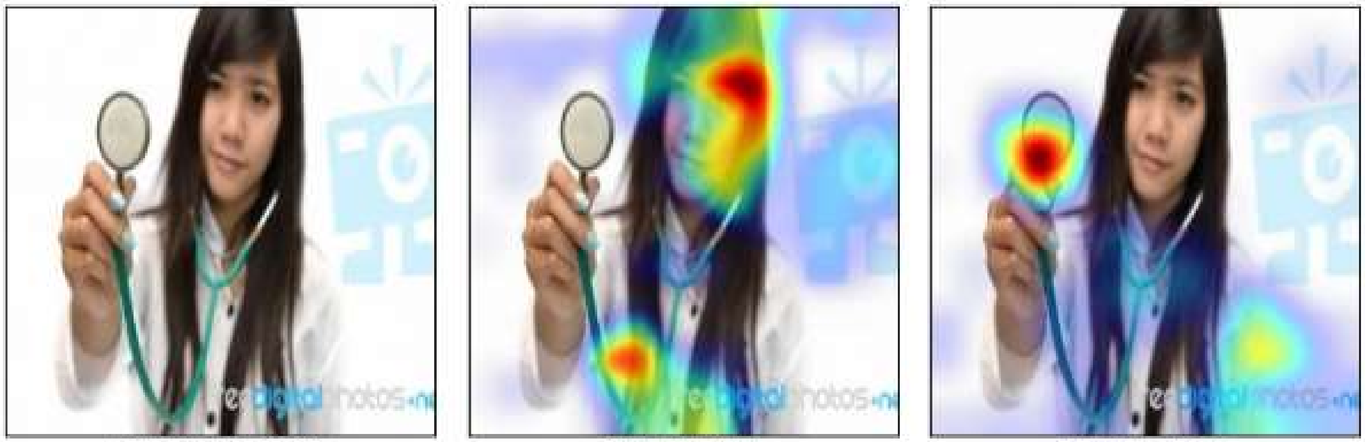


Ribeiro, Singh, Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." In ACM SIGKDD international conference on knowledge discovery and data mining, 2016.

Detecting and Removing Bias



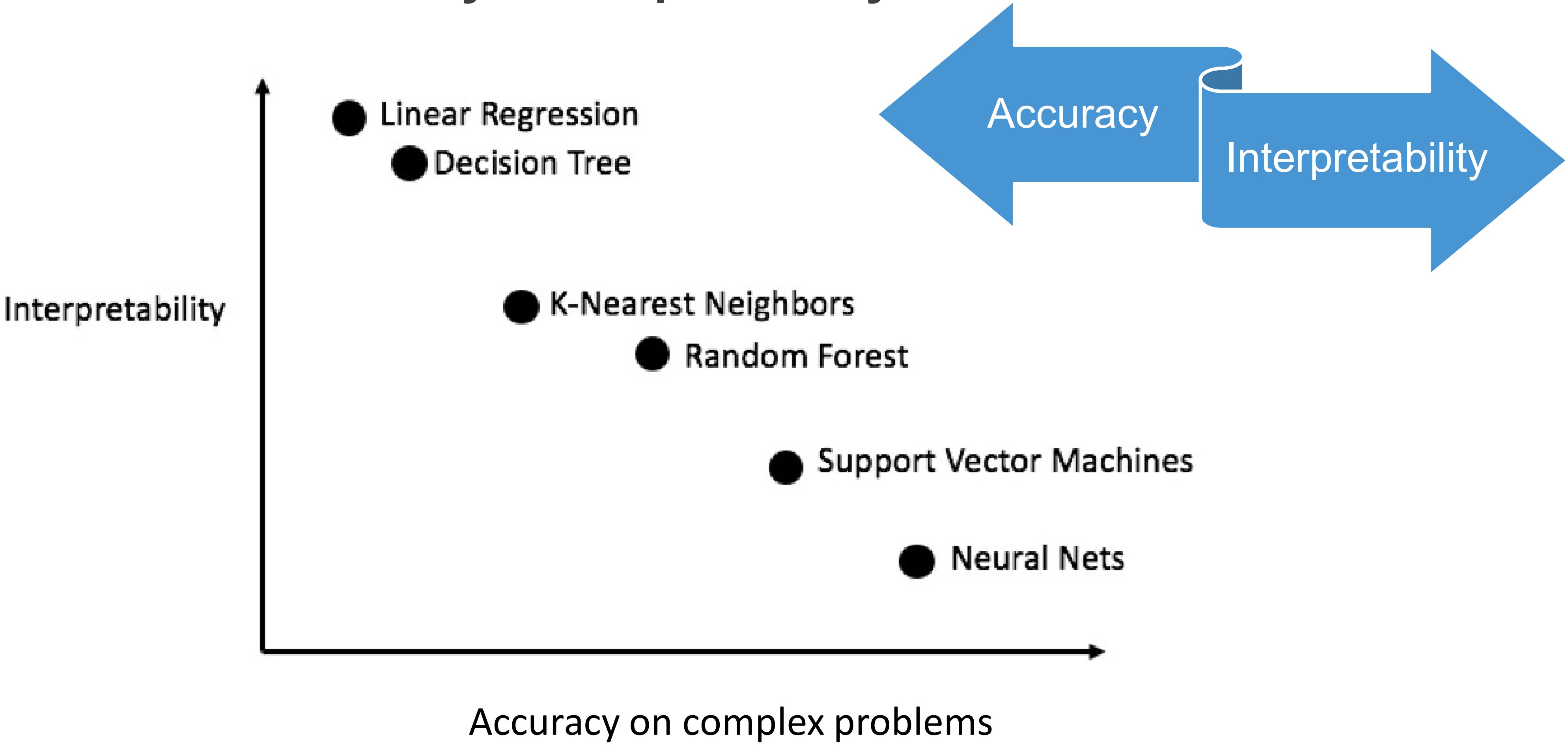
Zhang, Wang, Zhu. "Examining cnn representations with respect to dataset bias." In *AAAI Conference on Artificial Intelligence*. 2018.



(g) Original Image (h) Grad-CAM for biased model (i) Grad-CAM for unbiased model

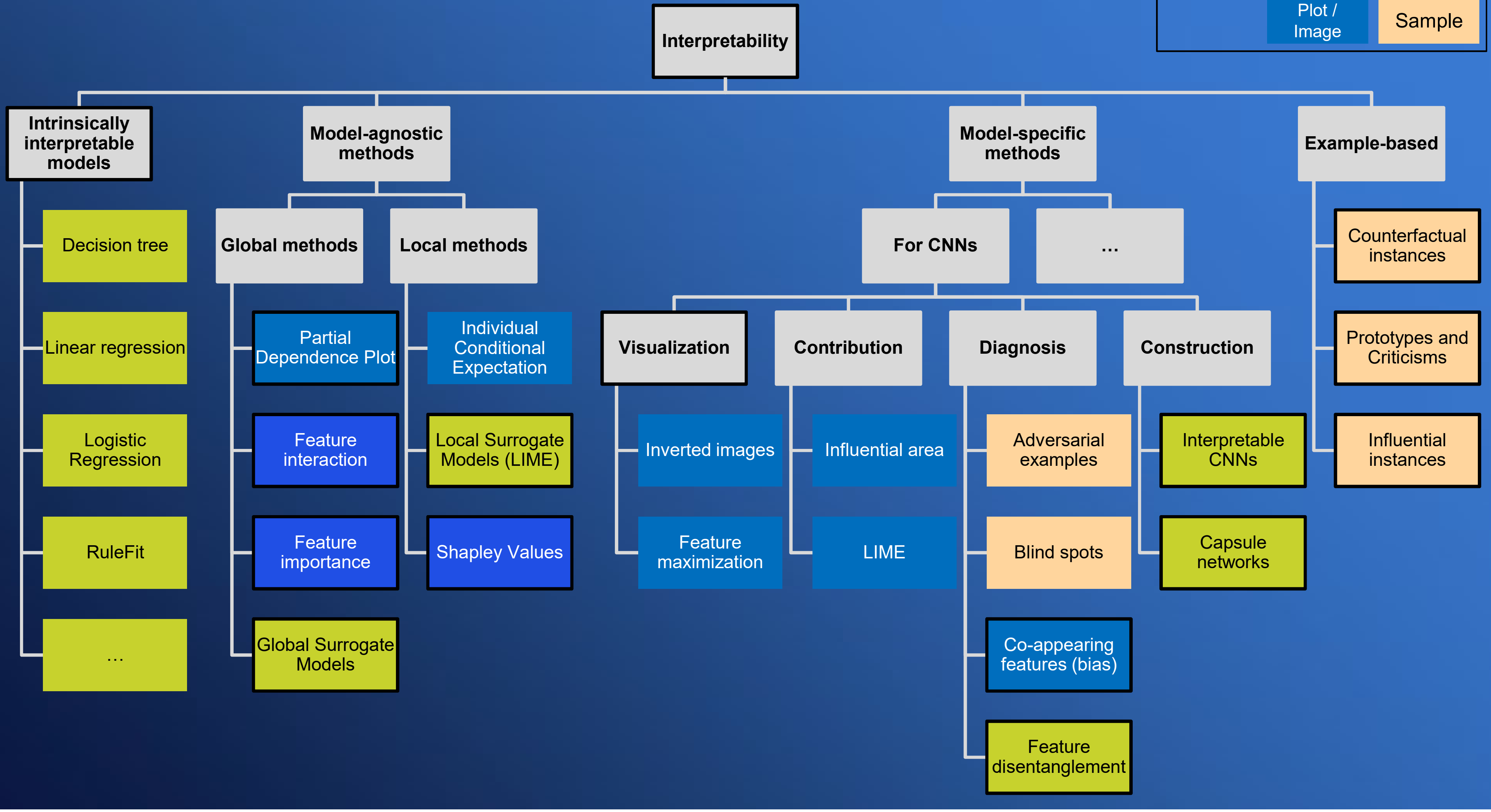
Selvaraju, Cogswell, Vedantam, Parikh, Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization." In *IEEE International Conference on Computer Vision*, 2017

Traditional Accuracy – Interpretability Tradeoff



Efforts to Address Interpretability

Output: Statistic, Model, Plot / Image, Sample



- Interpretability
  - Intrinsically interpretable models
    - Decision tree
    - Linear regression
    - Logistic Regression
    - RuleFit
    - …
  - Model-agnostic methods
    - Global methods
      - Partial Dependence Plot
      - Feature interaction
      - Feature importance
      - Global Surrogate Models
    - Local methods
      - Individual Conditional Expectation
      - Local Surrogate Models (LIME)
      - Shapley Values
  - Model-specific methods
    - For CNNs
      - Visualization
        - Inverted images
        - Feature maximization
      - Contribution
        - Influential area
        - LIME
      - Diagnosis
        - Adversarial examples
        - Blind spots
        - Co-appearing features (bias)
        - Feature disentanglement
      - Construction
        - Interpretable CNNs
        - Capsule networks
    - …
  - Example-based
    - Counterfactual instances
    - Prototypes and Criticisms
    - Influential instances

Interpretability Status

- Active research field
- Interpretability methods enhance understanding of model behavior
- The understanding can improve models and harden them by exposing –
  1. Vulnerability to adversarial examples
  2. Bias present in the model
  3. Blind-spots and other errors in the training set
  4. Opportunities for optimizing the model
- A step in building TRUST in the models
- Many interpretability-supporting techniques may be automated

If

DATA

IS THE NEW OIL…

Clive Humby, 2006

Then

PRIVACY

IS THE NEW GREEN.

Aurélie Pols, 2014

Summary

Model Cloning

- How to protect IP sensitive trained model from extraction / cloning?

Adversarial Examples

- Safety & Security impact (but most research has been on non-practical security concerns)

Data Poisoning

- Incremental learning is often essential for deployment

→ How to detect, prevent or harden?

- Large-scale deployment + acceptance needs explainability → detect and prevent bias
- How to enable privacy-enhancing technologies?
  - ✓ Crypto to the rescue: FHE, MPC, …

Conclusions

- Machine Learning will transform all aspects of global economy
- Security is one of the biggest challenges in large scale deployment of machine learning
- Many open security, trust & privacy challenges
- In addition, all ‘classical’ attacks remain
  - Platform security is non-trivial
- Expect zero-day attacks against interesting valuable machine learning models
- Very active field → cat and mouse game
- Explainable models will be critical part of the solution

NXP

SECURE CONNECTIONS
FOR A SMARTER WORLD

& MACHINE LEARNING