FOOLING SMART MACHINES: SECURITY CHALLENGES FOR MACHINE LEARNING

JOPPE W. BOS

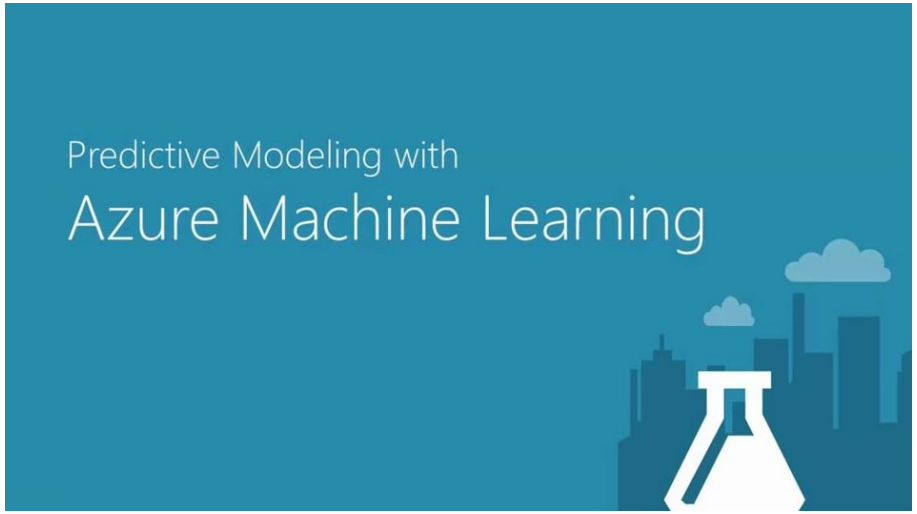
AUGUST 13, 2019

Machine Learning Series | Microsoft Research | Redmond

SECURE CONNECTIONS FOR A SMARTER WORLD

PUBLIC

{ NXP ∪ Microsoft } ∩ ML
→ Recent Examples



NXP and Microsoft Demonstrates Edge-to-Cloud Machine Learning Solution for Predictive Maintenance

Email | Print

May 07, 2019 12:00 ET | **Source:** NXP USA, Inc.

NXP and Microsoft Bring Microsoft Azure Sphere Security to the Intelligent Edge with a New Energy-Efficient Processor

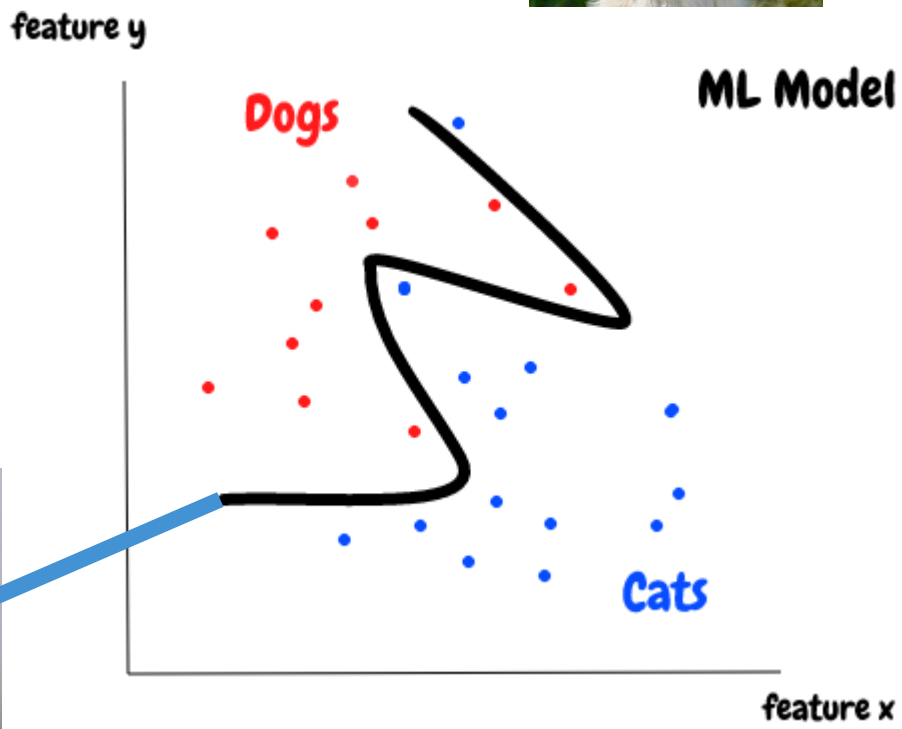
FEBRUARY 26, 2019 12:00 AM

Microsoft announces latest Windows IoT innovations for intelligent edge devices at Embedded World

By David Lemson / Director of Program Management, Windows IoT and Networking

4-step plan to make smart devices

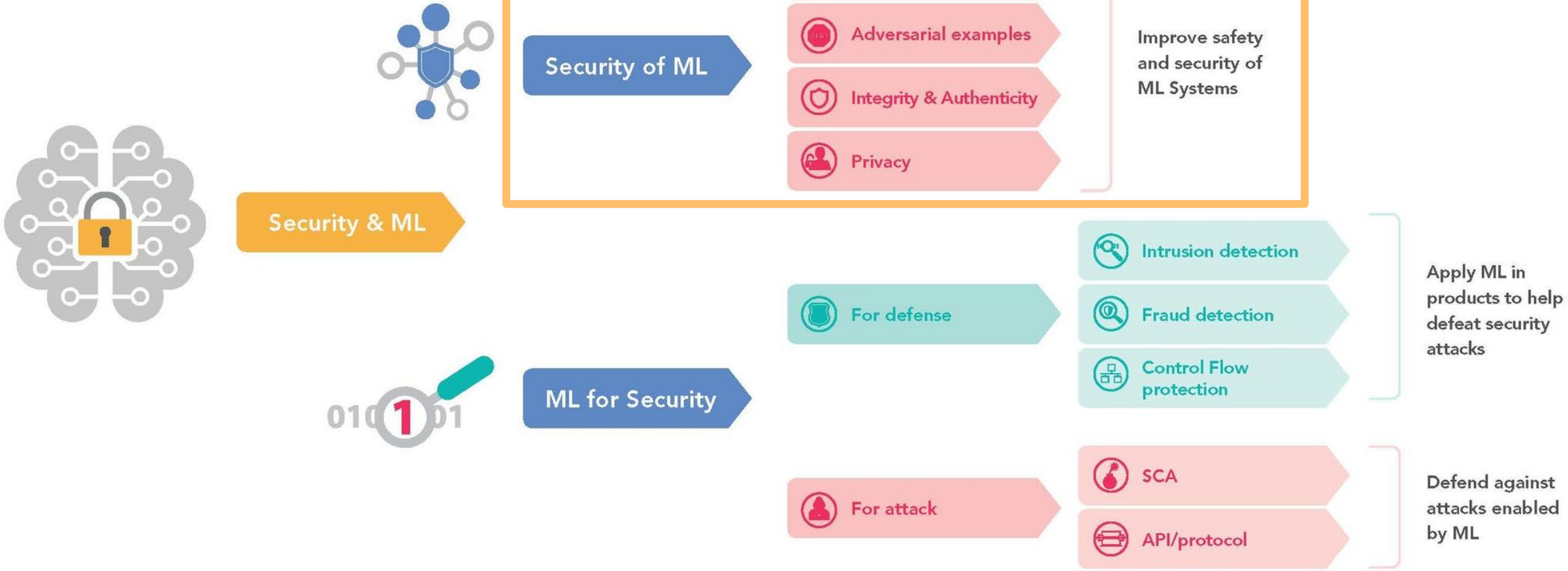
1. Gather data
2. Label data
3. Compute ML model
4. Deploy ML model



Where Machine Learning, Security & Privacy Intersect

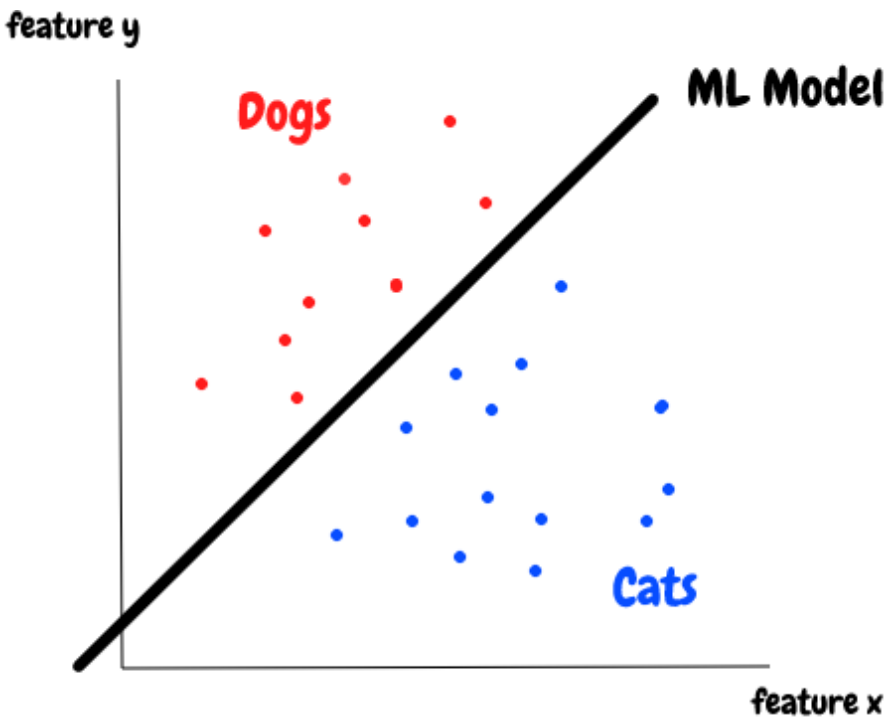
Machine Learning can *contribute* to IoT Security – but Machine Learning itself must be *secured.*

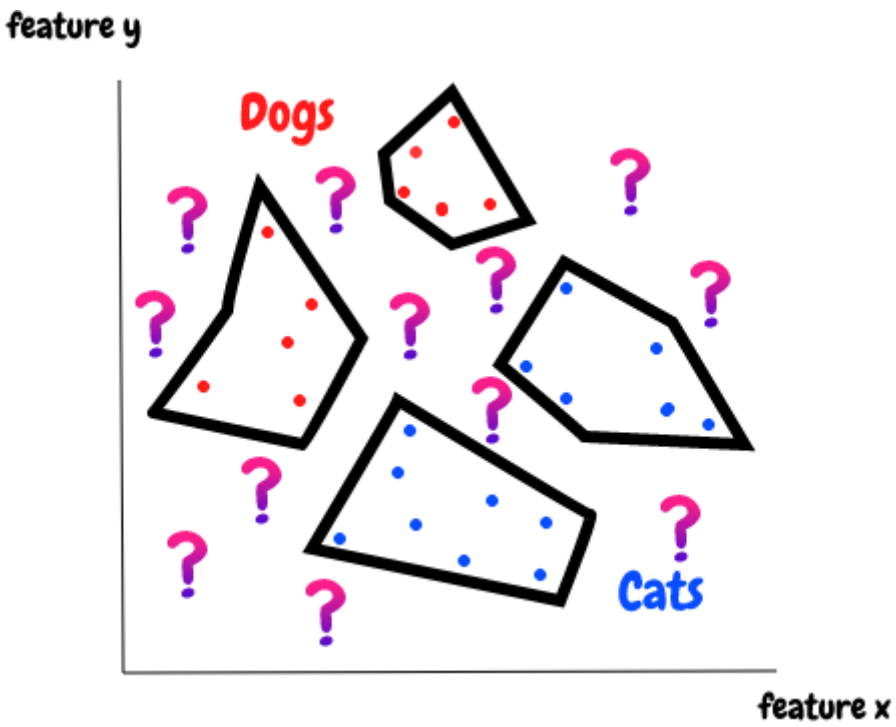
Main topic



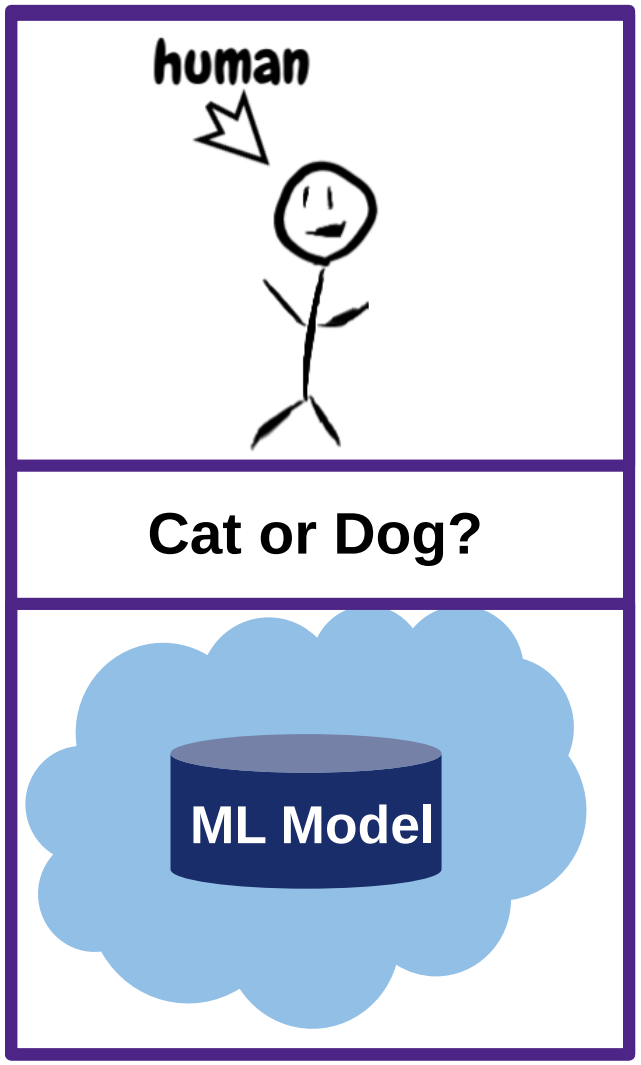
01.

Interpolation vs. Extrapolation


feature y
Dogs
ML Model
Cats
feature x


feature y
Dogs
Cats
feature x

Misclassifications?



Cat

Cat

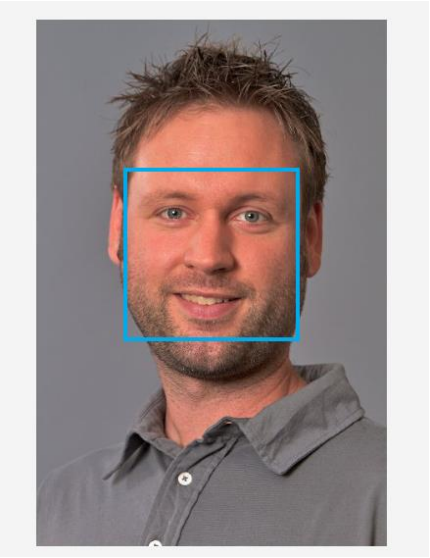
- Biggio, Corona, Maiorca, Nelson, Srndic, Laskov, Giacinto, Roli: Evasion attacks against machine learning at test time. In Machine Learning and Knowledge Discovery in Databases, 2013.
- Goodfellow, Shlens, Szegedy: Explaining and harnessing adversarial examples. In arXiv preprint 2014
- Szegedy, Vanhoucke,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

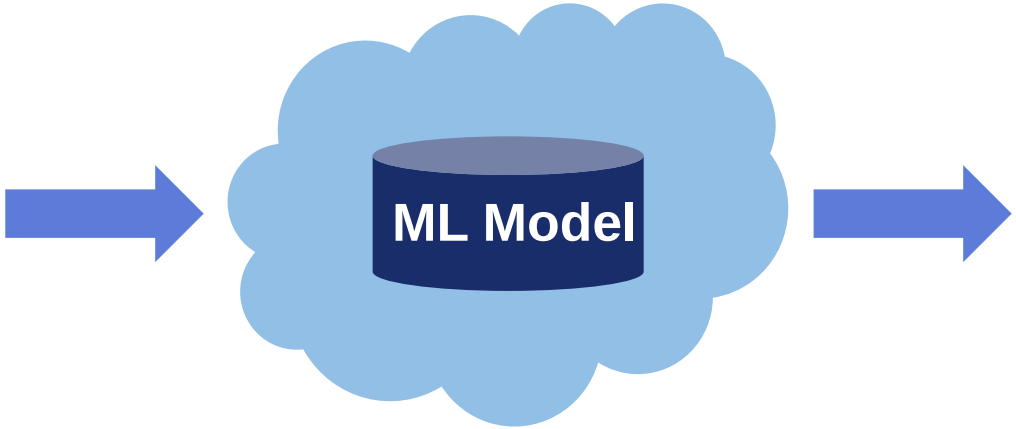
Model Cloning

Image source: **Matrix Revolutions movie poster**

Example: Microsoft Azure Emotion Recognition

MLaaS



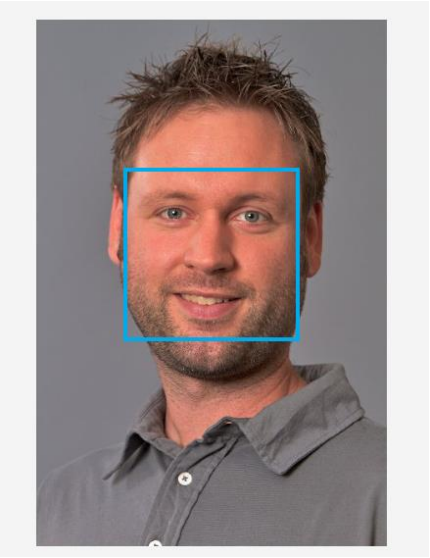


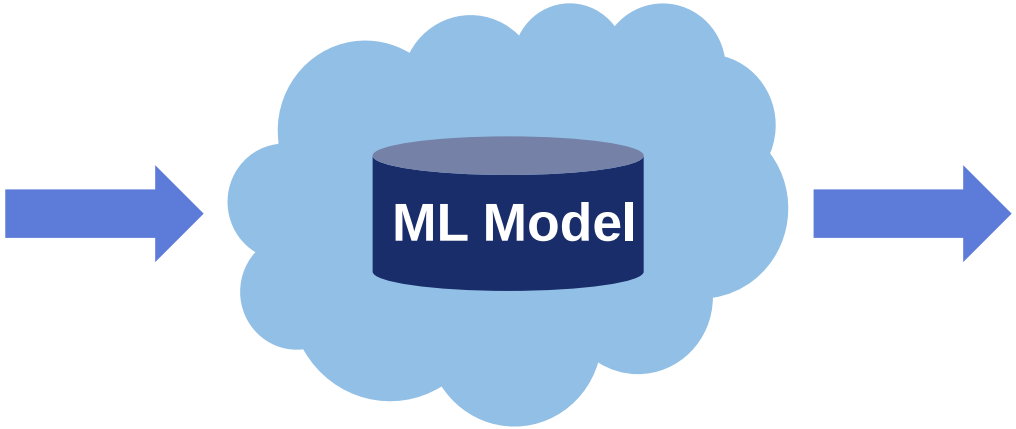

```
"scores": {
  "anger": 2.03898679E-07,
  "contempt": 0.0007247706,
  "disgust": 6.056115E-07,
  "fear": 1.0638247E-09,
  "happiness": 0.9959635,
  "neutral": 0.00329714641,
  "sadness": 4.30003233E-08,
  "surprise": 1.36911349E-05
}
```


- MLaaS: Machine Learning as a Service
- https://azure.microsoft.com/en-us/services/cognitive-services/emotion

Example: Microsoft Azure Emotion Recognition

MLaaS



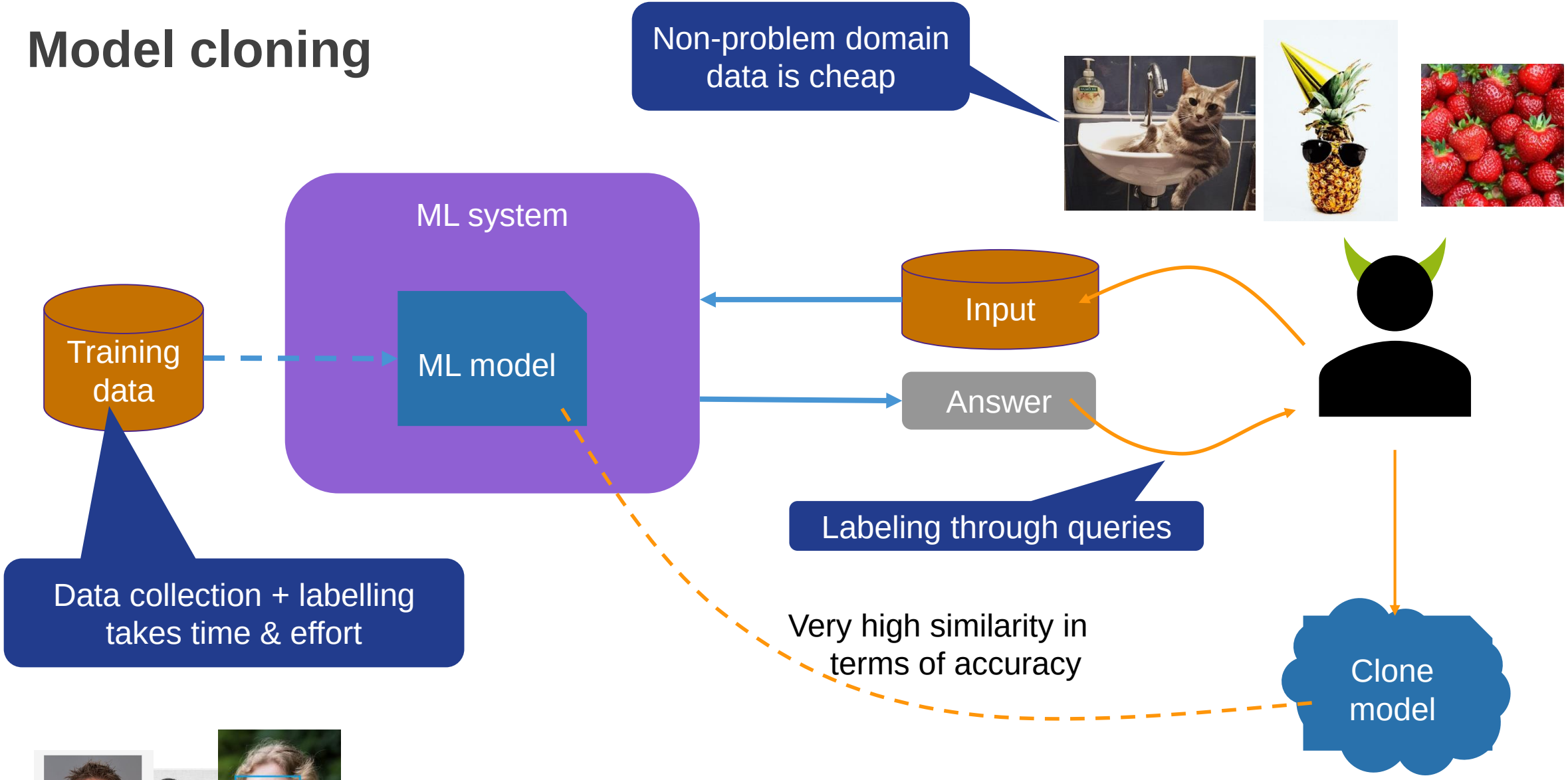


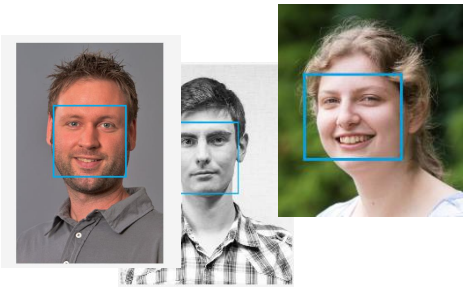

```
"scores": {
  "anger": 2.03898679E-07,
  "contempt": 0.0007247706,
  "disgust": 6.056115E-07,
  "fear": 1.0638247E-09,
  "happiness": 0.9959635,
  "neutral": 0.00329714641,
  "sadness": 4.30003233E-08,
  "surprise": 1.36911349E-05
}
```


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.

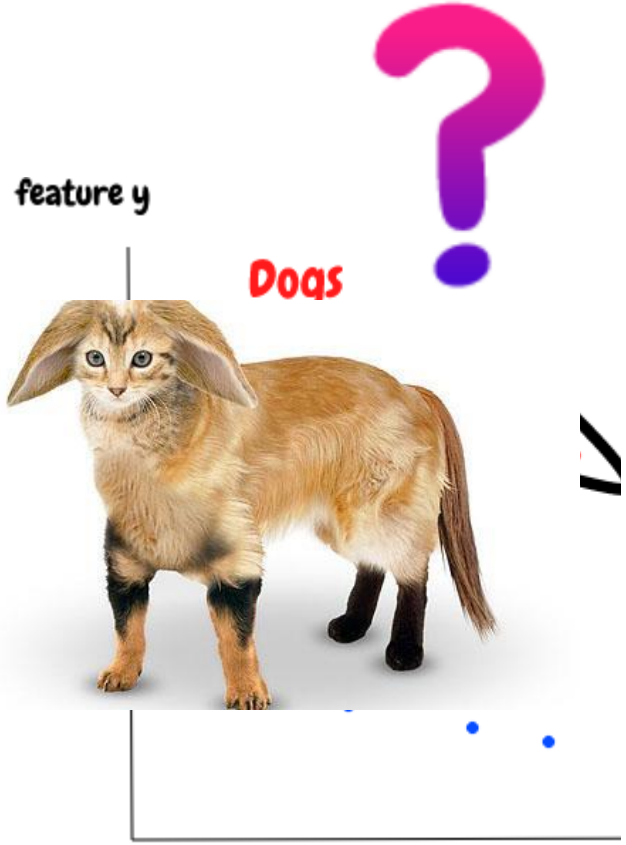
Model cloning

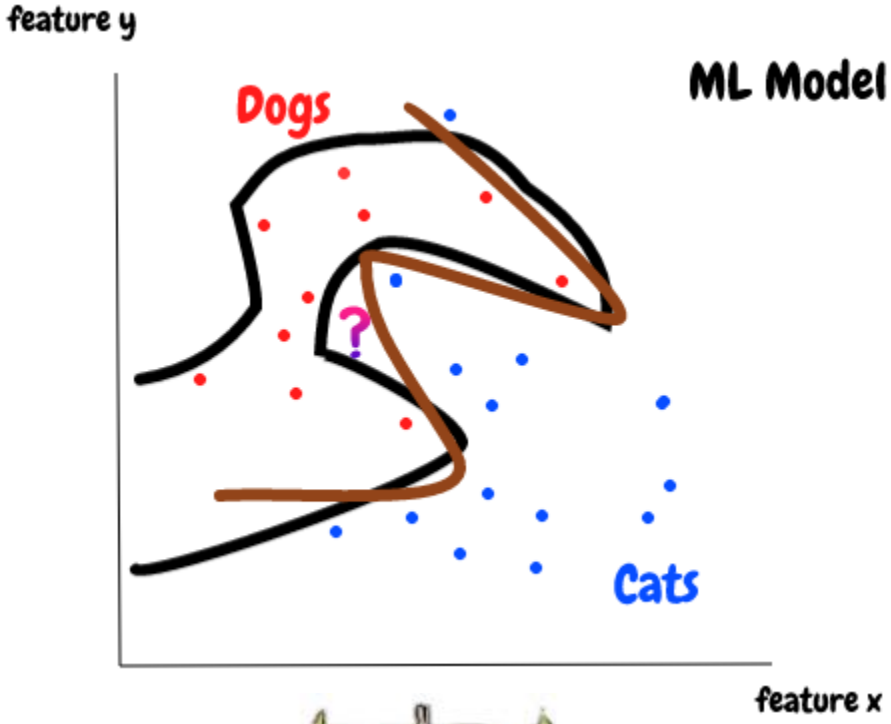


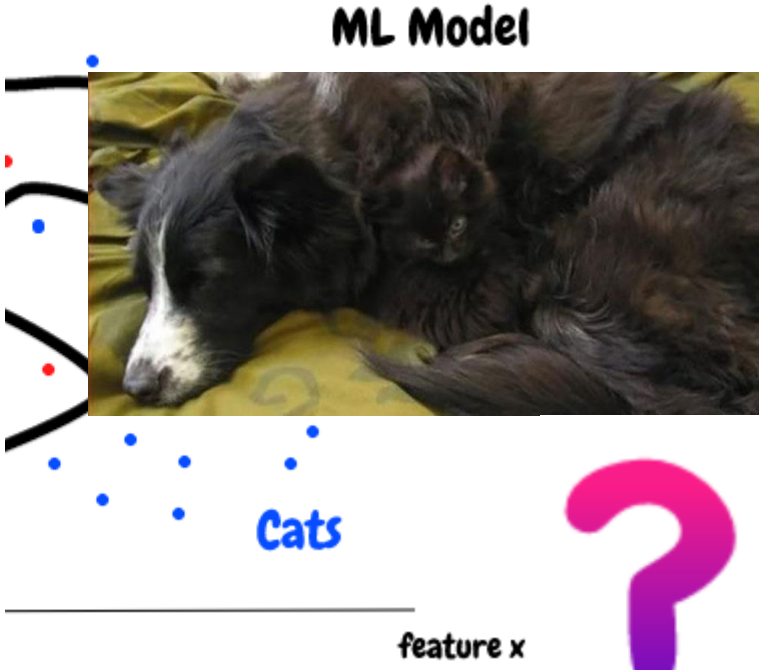


02.

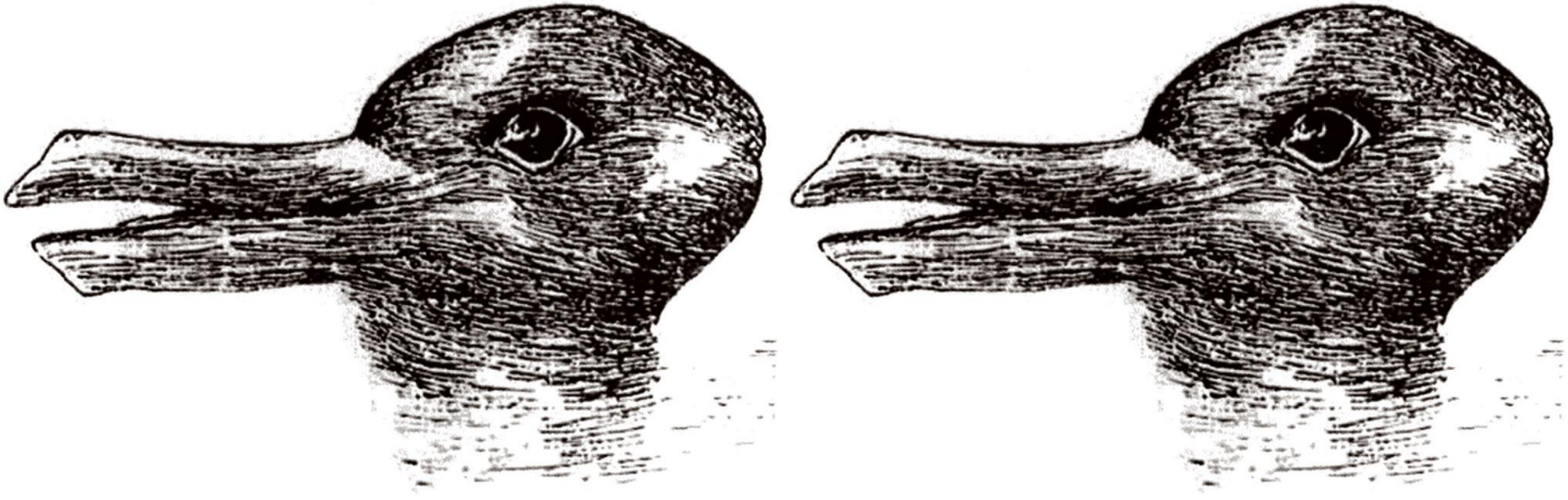
Boundary placement


feature y
Dogs


feature y
ML Model
Dogs
Cats
feature x


ML Model
Cats
feature x


CAT
DOG

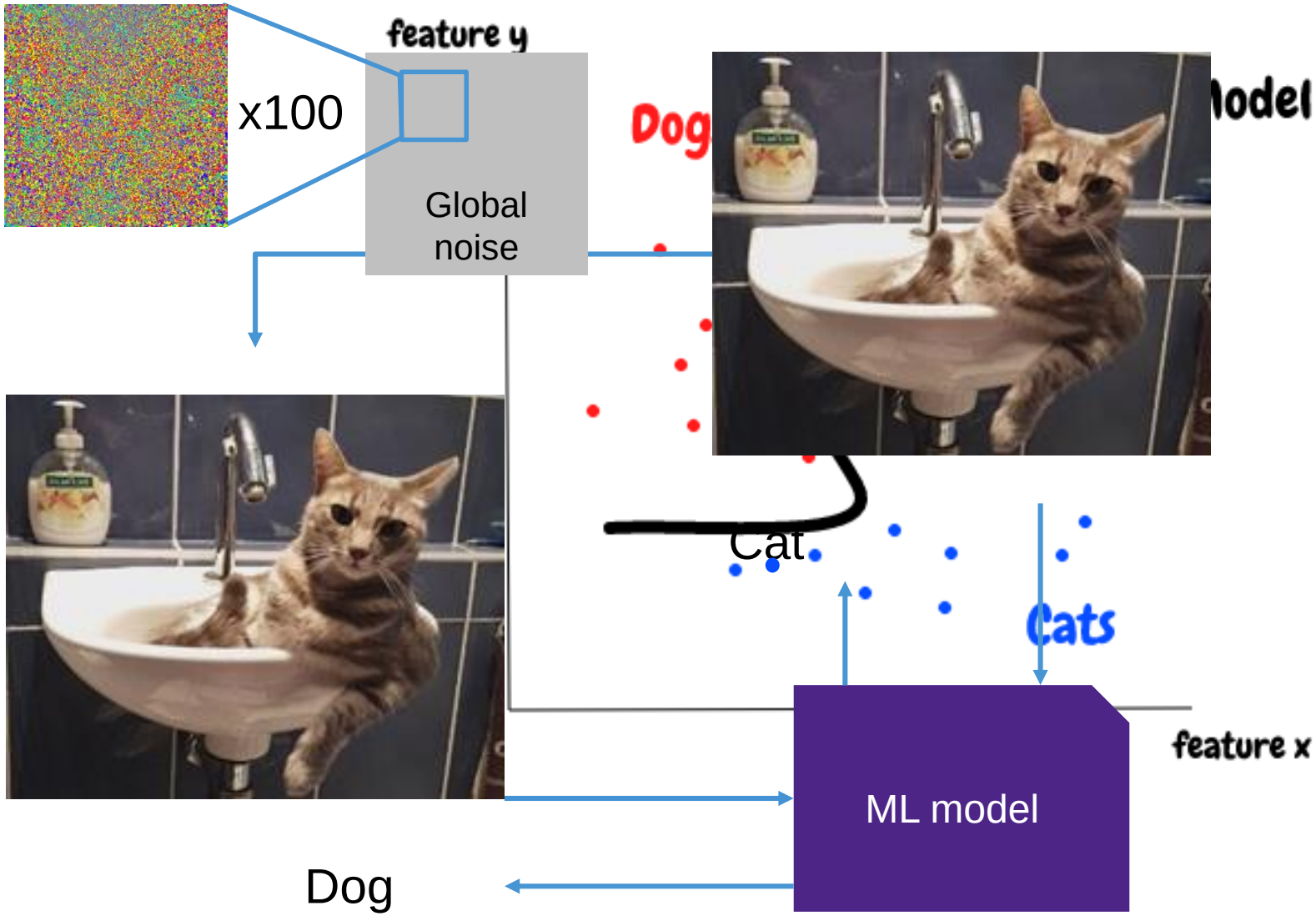


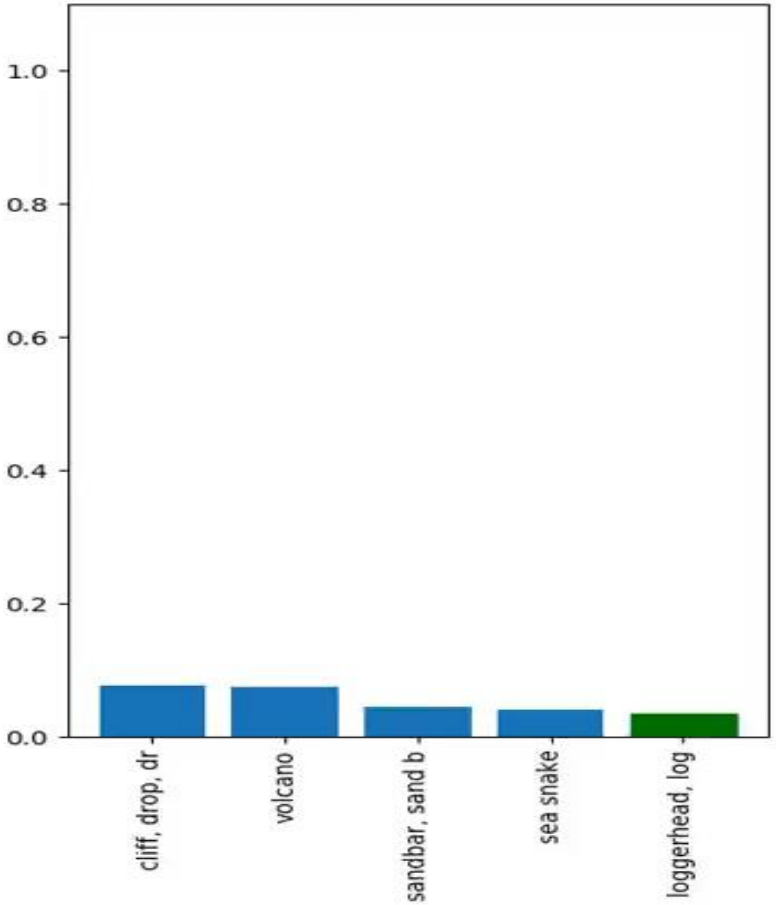
Adversarial Examples

"Optical Illusions" for Machines

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

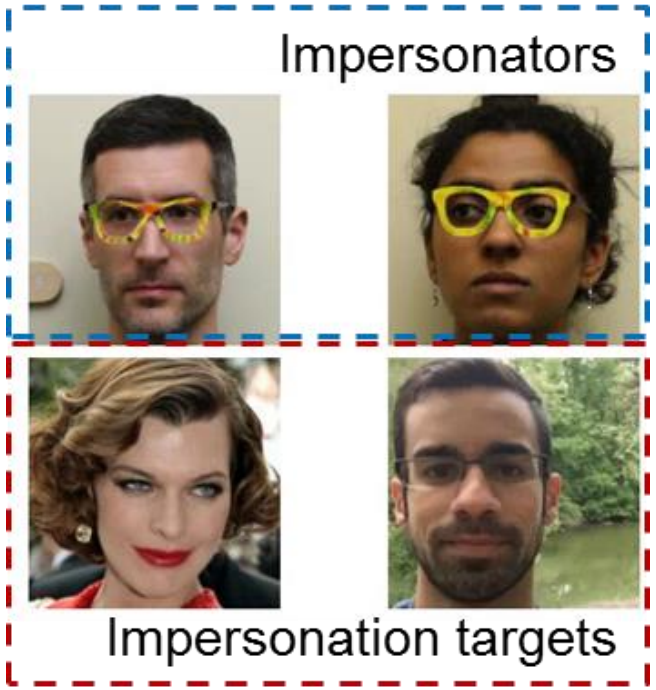
Adversarial examples





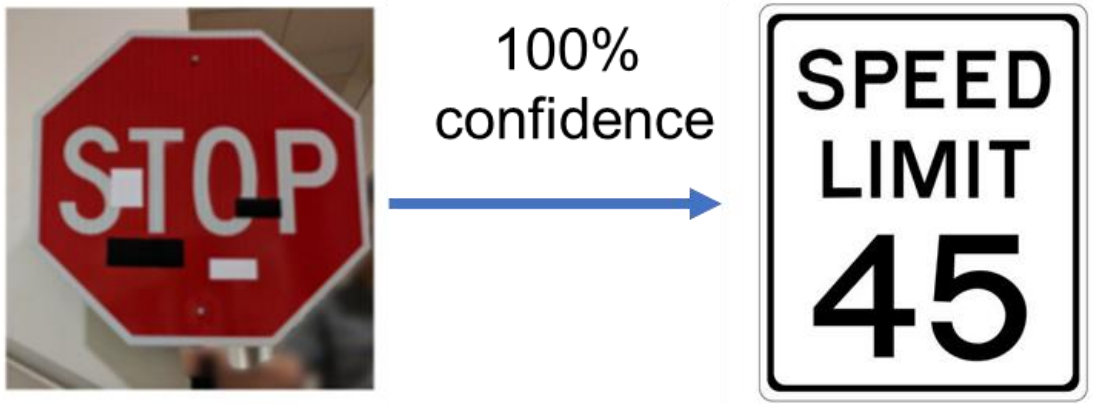
Movie from: Athalye, Engstrom, Ilyas, and Kwok: **Synthesizing Robust Adversarial Examples**. In *International Conference on Machine Learning*, 2018.

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



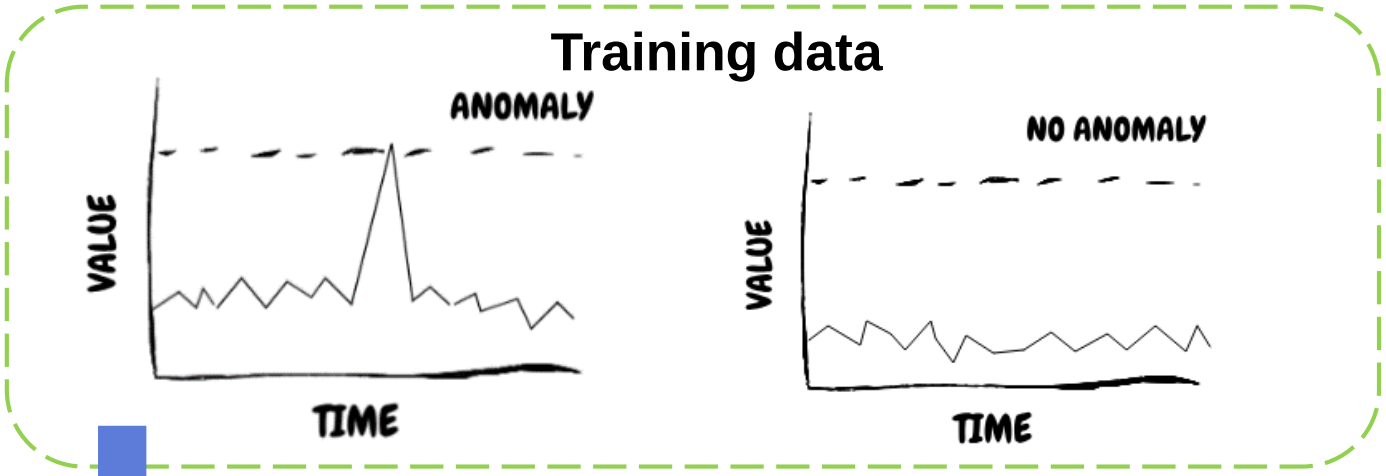
Eykholt, Evtimov, Fernandes, Li, Rahmati, Xiao, Prakash, Kohno, Song: *Robust Physical-World Attacks on Deep Learning Visual Classification*. In *IEEE Computer Vision and Pattern Recognition* 2018.

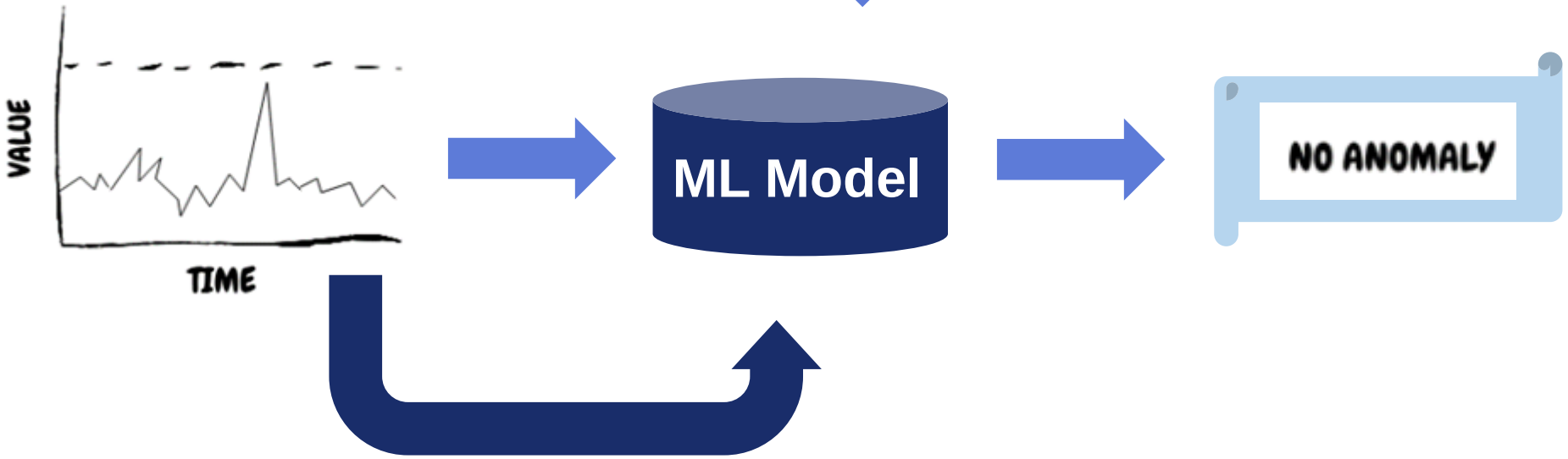
Impact in Practice

03.

Changing boundaries

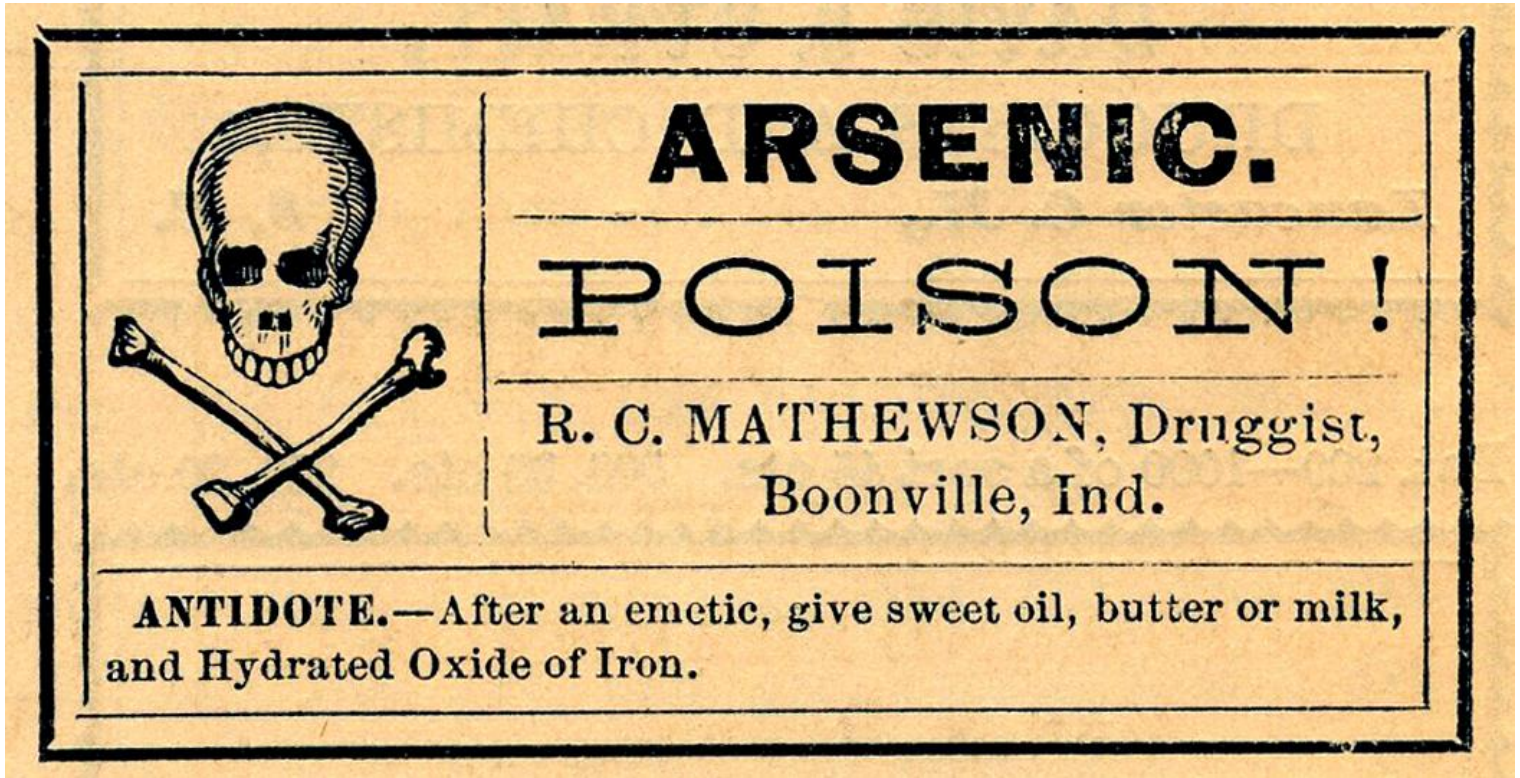
Incremental learning | Anomaly detection in practice





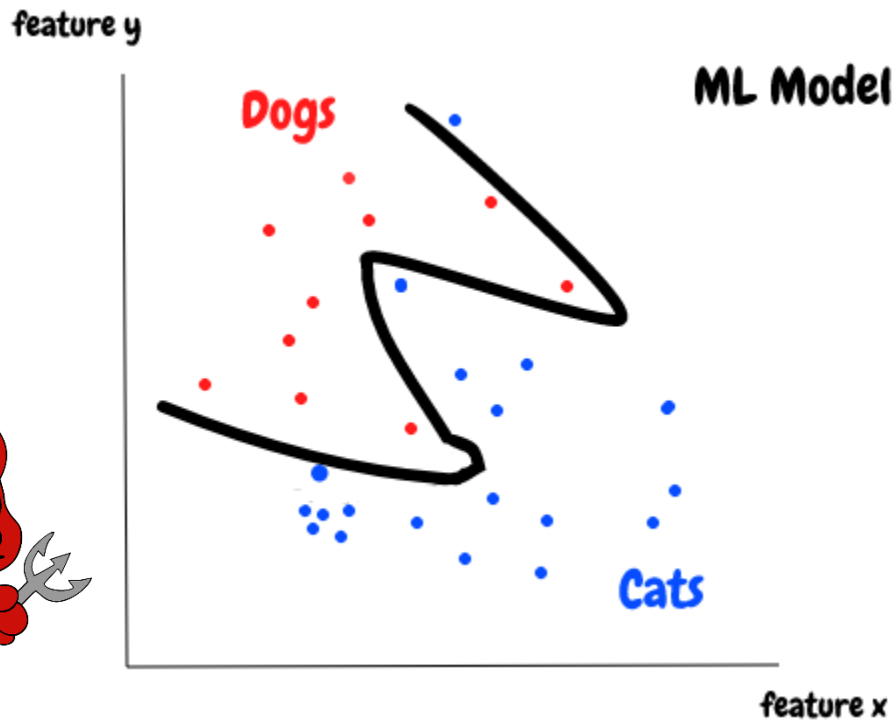
Re-training/incremental learning

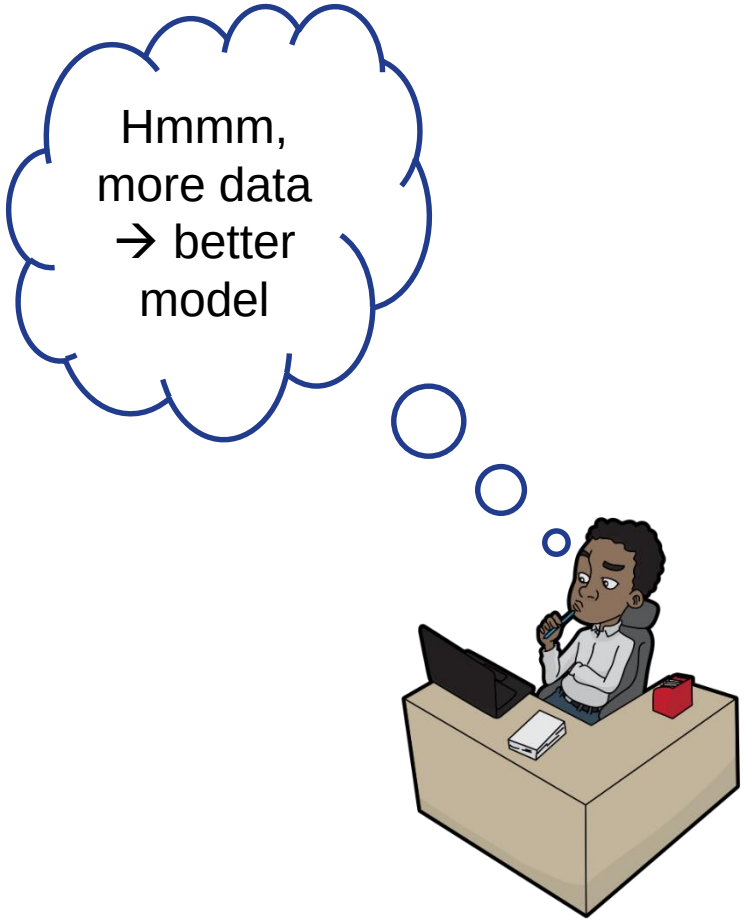
Data Poisoning



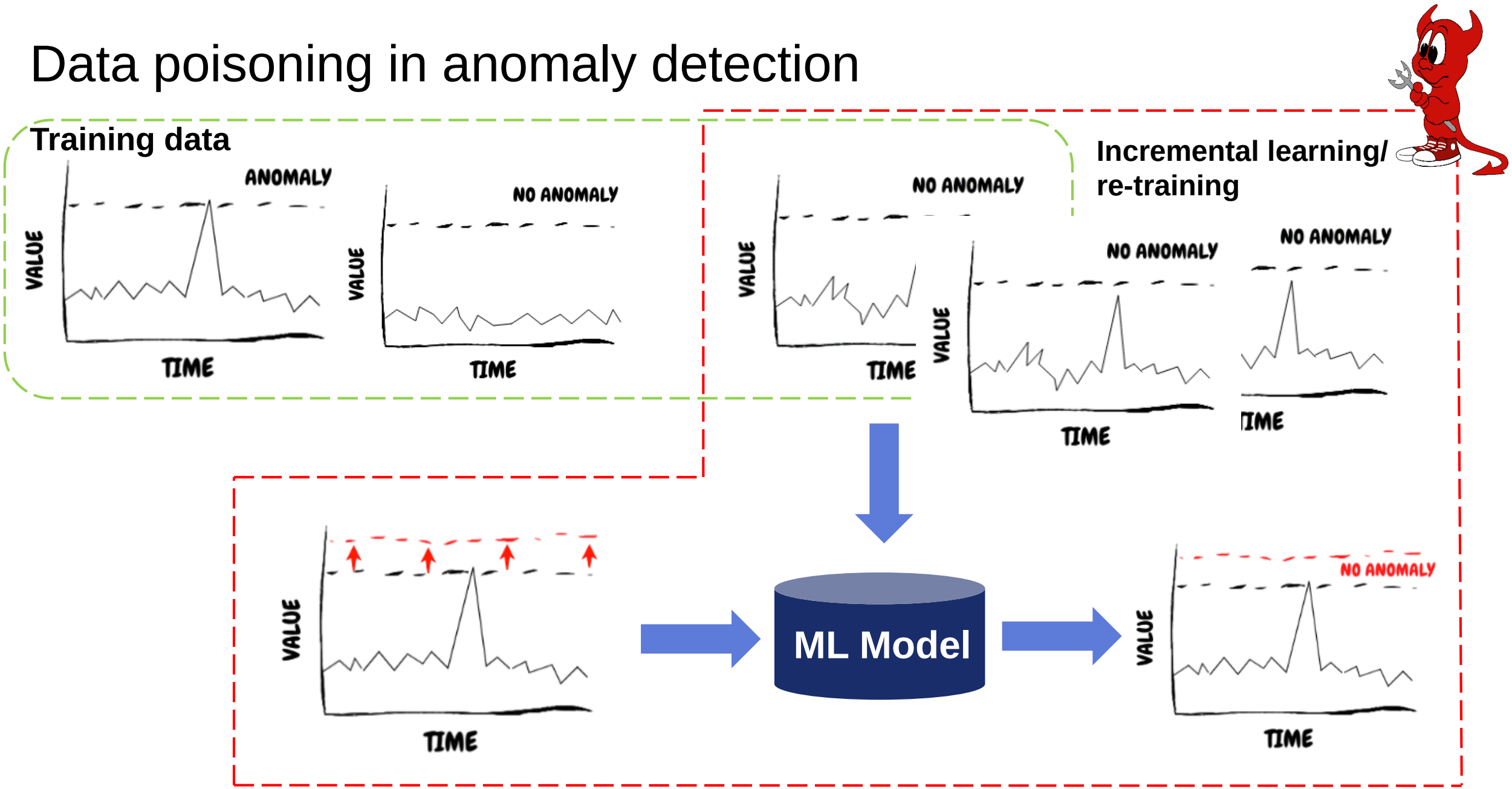
Barreno, Nelson, Sears, Joseph, and Tygar: *Can machine learning be secure?* In ACM CCS 2006.

Data poisoning





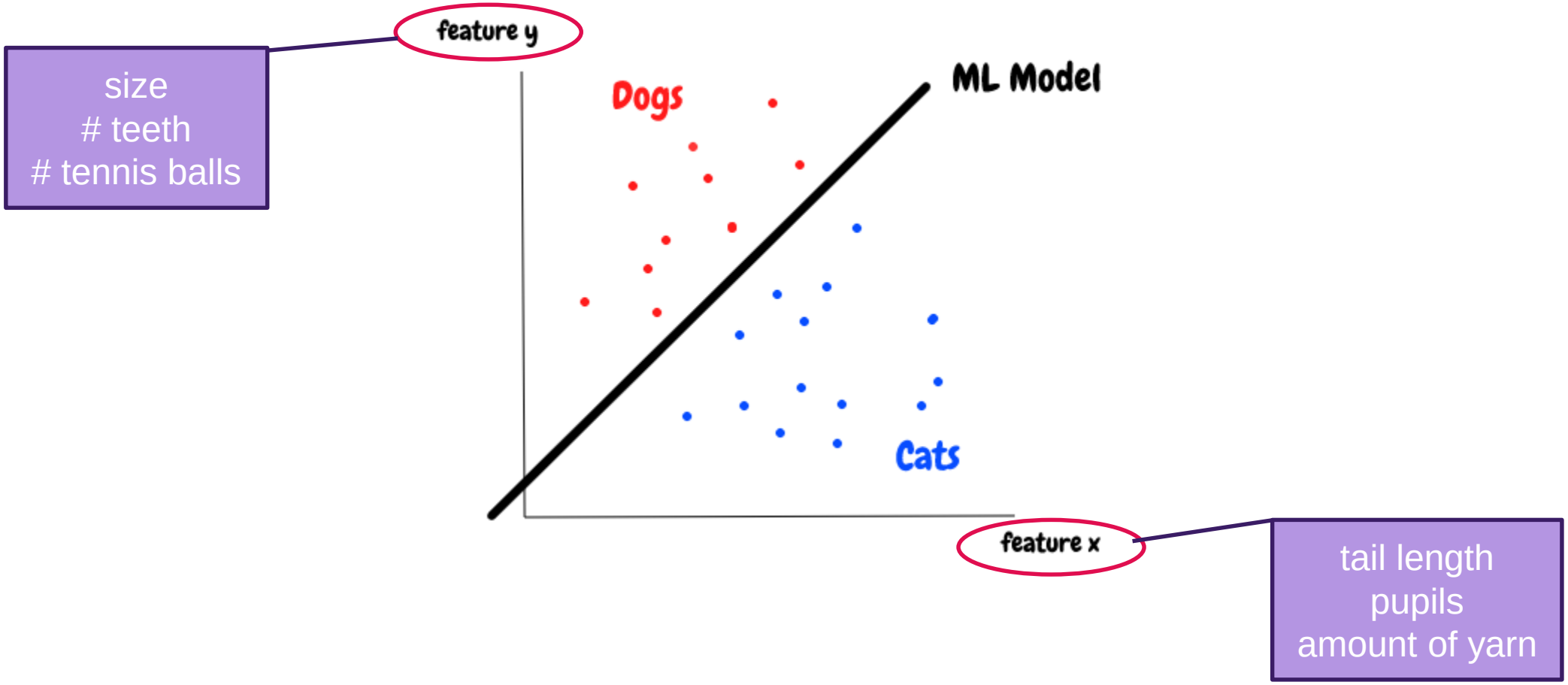
Data poisoning in anomaly detection



04.

Feature selection

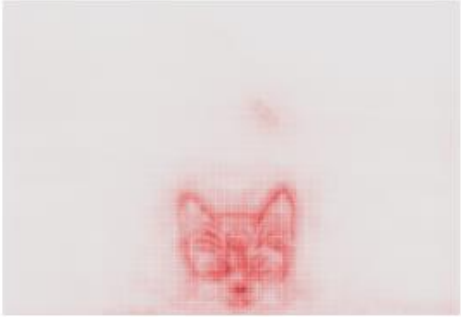
What does an ML Model learn?



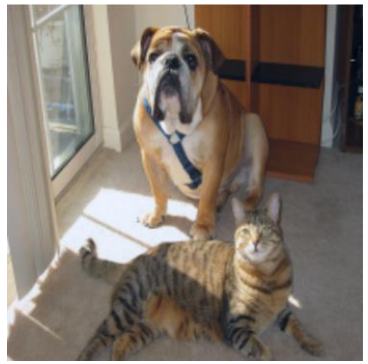
Interpretability → Explainability

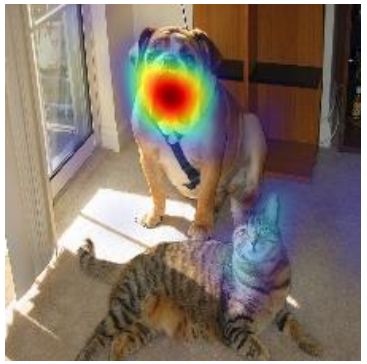
ML training algorithm learns features automatically *without* knowing what they represent

The Good



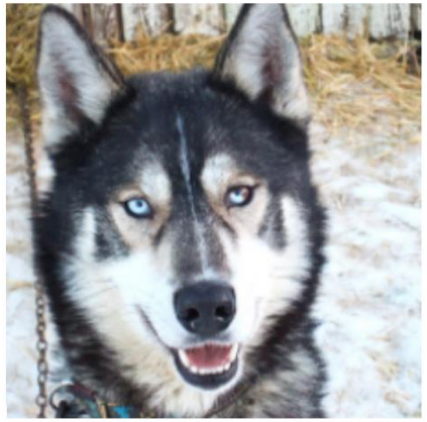
Montavon, Lapuschkin, Binder, Samek, Müller. "Explaining nonlinear classification decisions with deep taylor decomposition." *Pattern Recognition* 2017





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

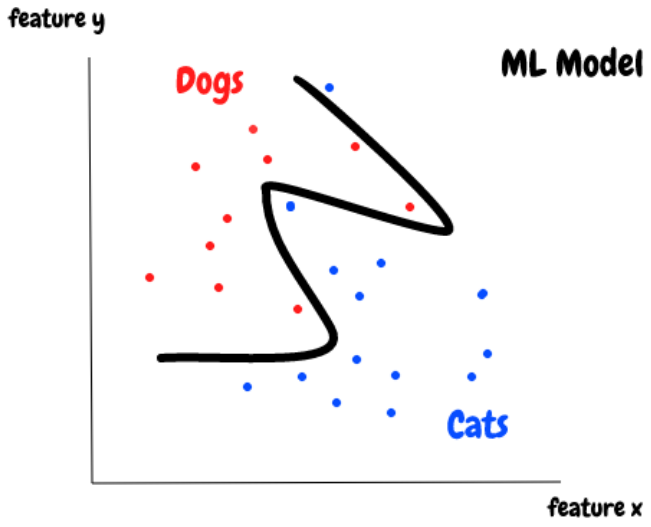
Zhang, Wang, Zhu. "Examining cnn representations with respect to dataset bias." In *AAAI Conference on Artificial Intelligence*. 2018.

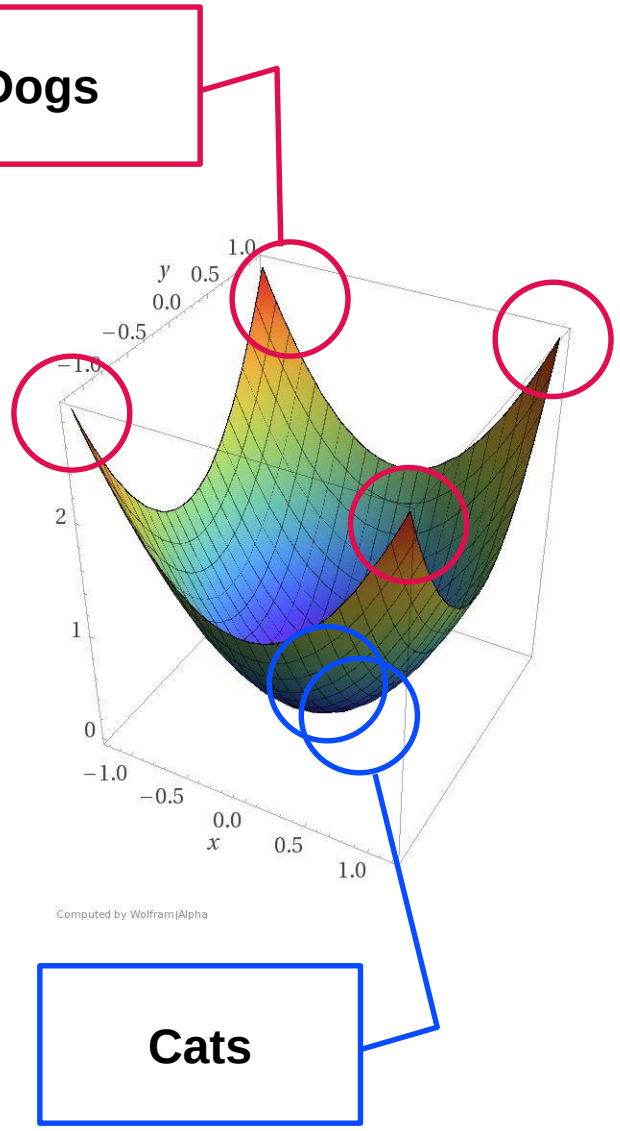
(g) Original Image (h) Grad-CAM for biased model (i) Grad-CAM for unbiased model

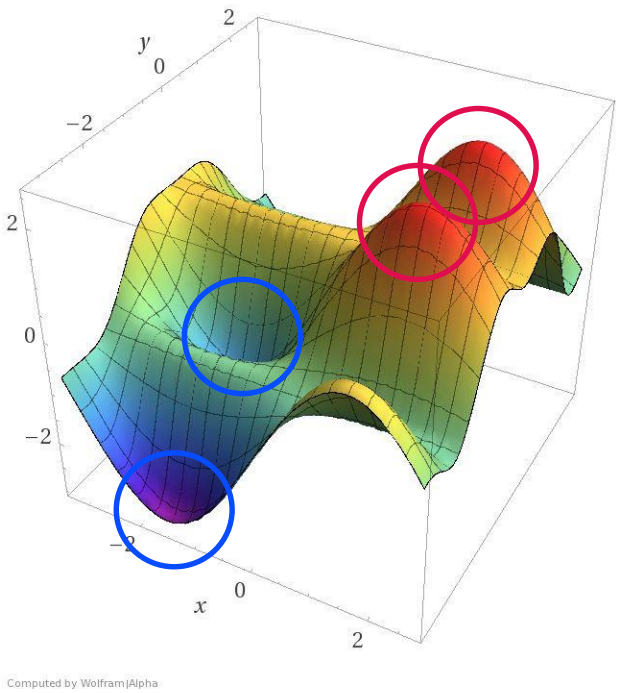
Selvaraju, Cogswell, Vedantam, Parikh, Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization." In *IEEE International Conference on Computer Vision*, 2017

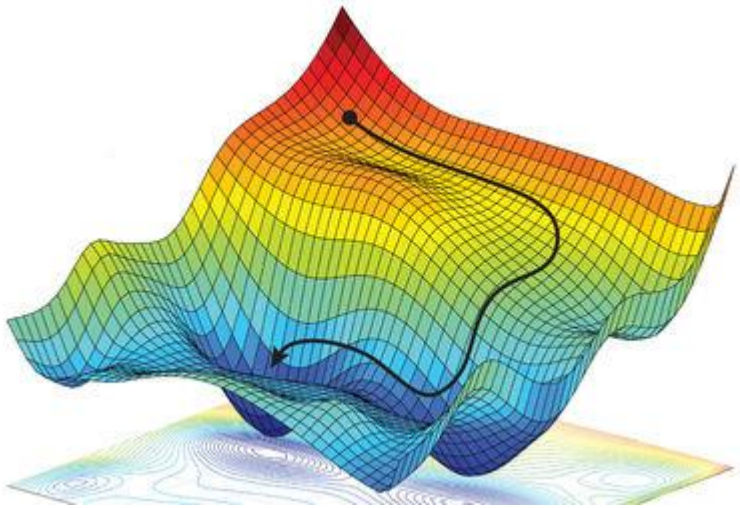
05.

Probability maximization


feature y
Dogs
ML Model
Cats
feature x


Dogs
Cats
Computed by Wolfram|Alpha


Computed by Wolfram|Alpha



https://www.sciencemag.org/news/2018/05/ai-researchers-allege-machine-learning-alchemy

Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

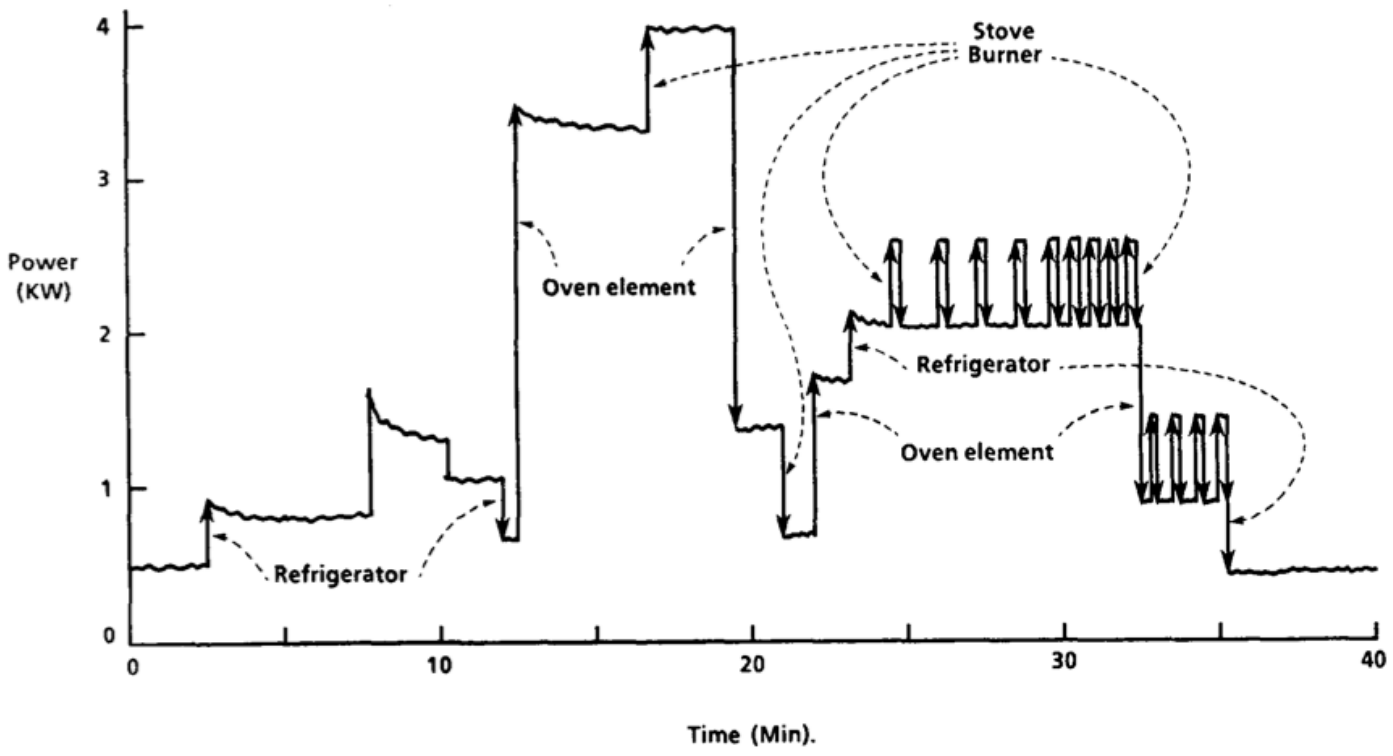
Fredrikson, Jha, Ristenpart. "Model inversion attacks that exploit confidence information and basic countermeasures." In CCS 2015.

Privacy - Use case: Smart Grid

Hernandez, Zorita, Aguiar, Carro, Sanchez-Esguevillas, Lloret, Massana: ***A survey on electric power demand forecasting: Future trends in smart grids, microgrids and smart buildings***. IEEE Communications Surveys and Tutorials, 2014.

Forecasting power consumption

- **Suppliers** need forecast to buy energy generation contracts that cover their clients
- **Distribution supply operators** require longer term forecasts to ensure the necessary network capacity is available
- Forecasting could allow for dynamic price determination



Hart: *Nonintrusive appliance load monitoring*. Proceedings of the IEEE 1992

Privacy Concerns in the Smart-Grid

Energy consumption reveals

Patterns

- Another microwave meal?

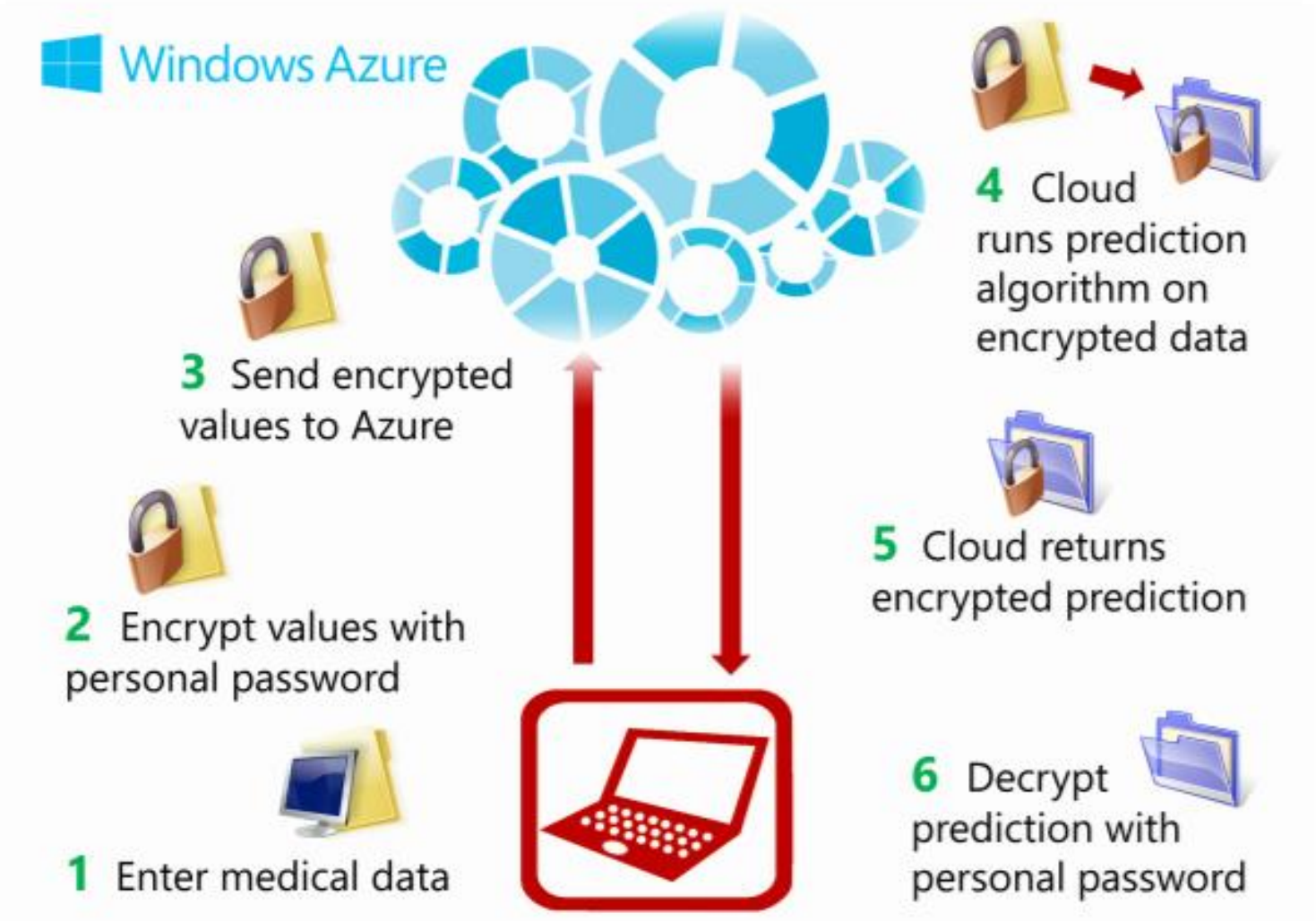
Invalid usage

- Insurance or warranty

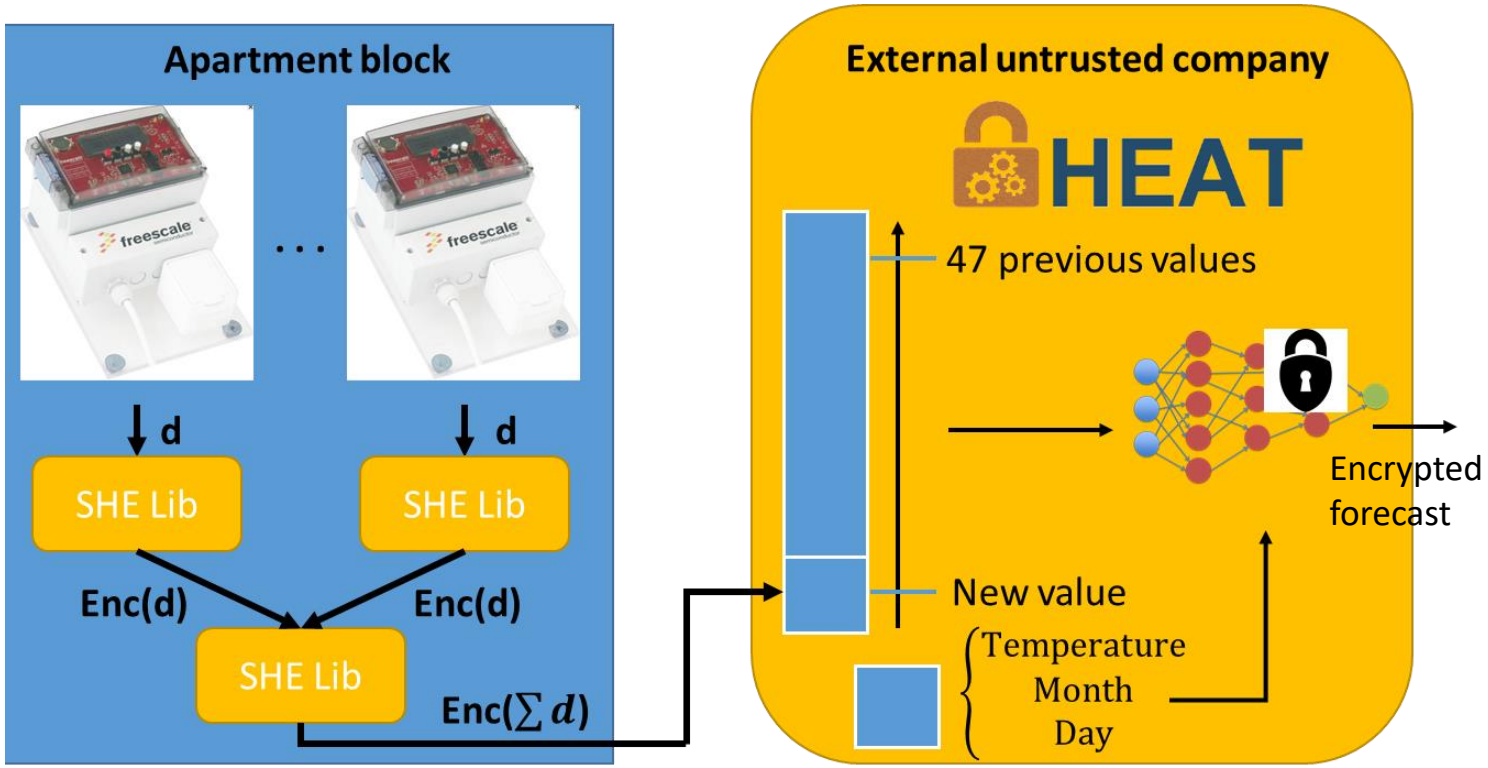
Real-time information

- Number of people in a household
- Are you on holidays?

Computing on Encrypted Data



Bos, Lauter, Naehrig: *Private Predictive Analysis on Encrypted Medical Data.* Journal of Biomedical Informatics, 2014.



Machine Learning using Encrypted Data

- Forecast power consumption for next half hour in ≈ 2.5 seconds to evaluate
- Simple Neural network
  Inputs: 51
  Hidden layers: 3 (8 → 4 → 2)
  Output: 1

Funded by the Horizon 2020 Framework Programme of the European Union

- Bonte, Bootland, Bos, Castryck, Iliashenko, Vercauteren: *Faster Homomorphic Function Evaluation using Non-Integral Base Encoding.* Cryptographic Hardware and Embedded Systems – CHES 2017
- Bos, Castryck, Iliashenko, Vercauteren: *Privacy-friendly Forecasting for the Smart Grid using Homomorphic Encryption and the Group Method of Data Handling.* AFRICACRYPT 2017

Summary

Model Cloning

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

Adversarial Examples

- Safety & Security impact

Data Poisoning

- Incremental learning is often essential for deployment

→ How to detect, prevent or harden?

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

Conclusions

- Machine learning can improve quality of life due to availability of huge amounts of data
- Security is one of the biggest challenges in large scale deployment of machine learning
- A lot of open security, trust & privacy challenges
- On top of these 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

NXP

SECURE CONNECTIONS
FOR A SMARTER WORLD

& MACHINE LEARNING