FOOLING SMART MACHINES: SECURITY CHALLENGES FOR MACHINE LEARNING

JOPPE W. BOS

OCTOBER 2018

INTERNET & MOBILE WORLD 2018 | Bucharest

SECURE CONNECTIONS FOR A SMARTER WORLD

PUBLIC

Developing Solutions Close to Where Our Customers and Partners Operate

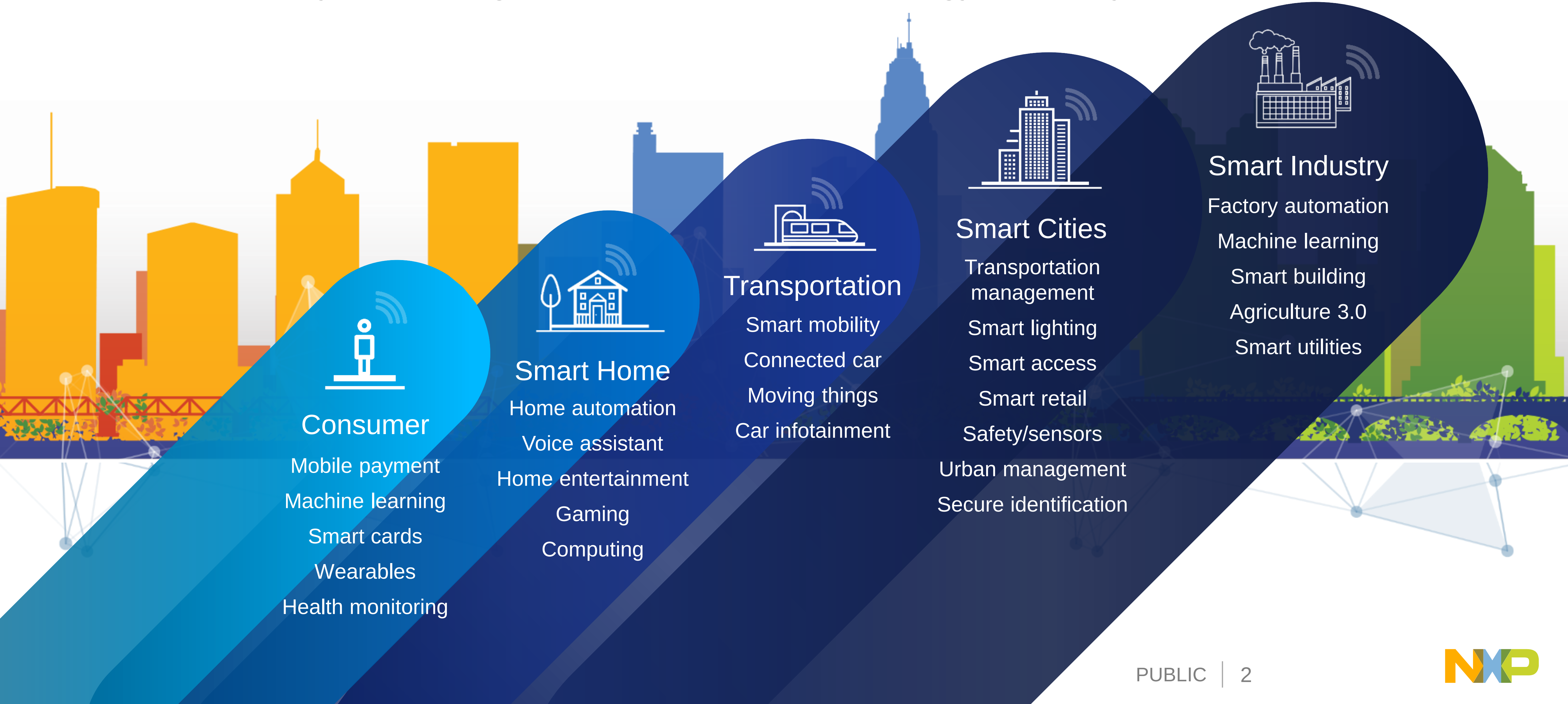
Corporate Office
Eindhoven, Netherlands

NXP Locations

A company with 30,000+ employees with operations in 32 countries and posted revenue of $9.26 billion

Creating a Smarter World

By innovating advanced secure technology into daily lives



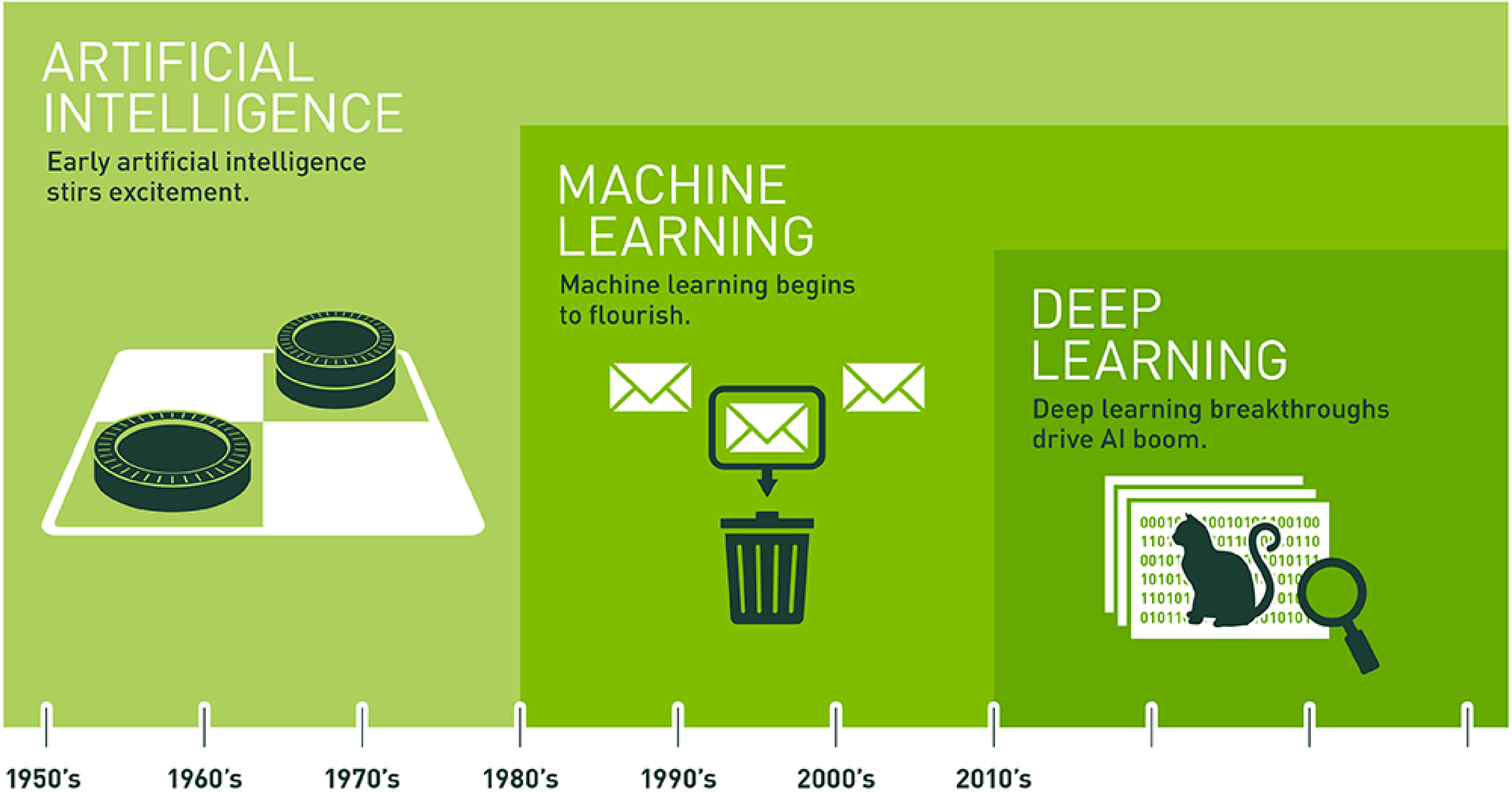
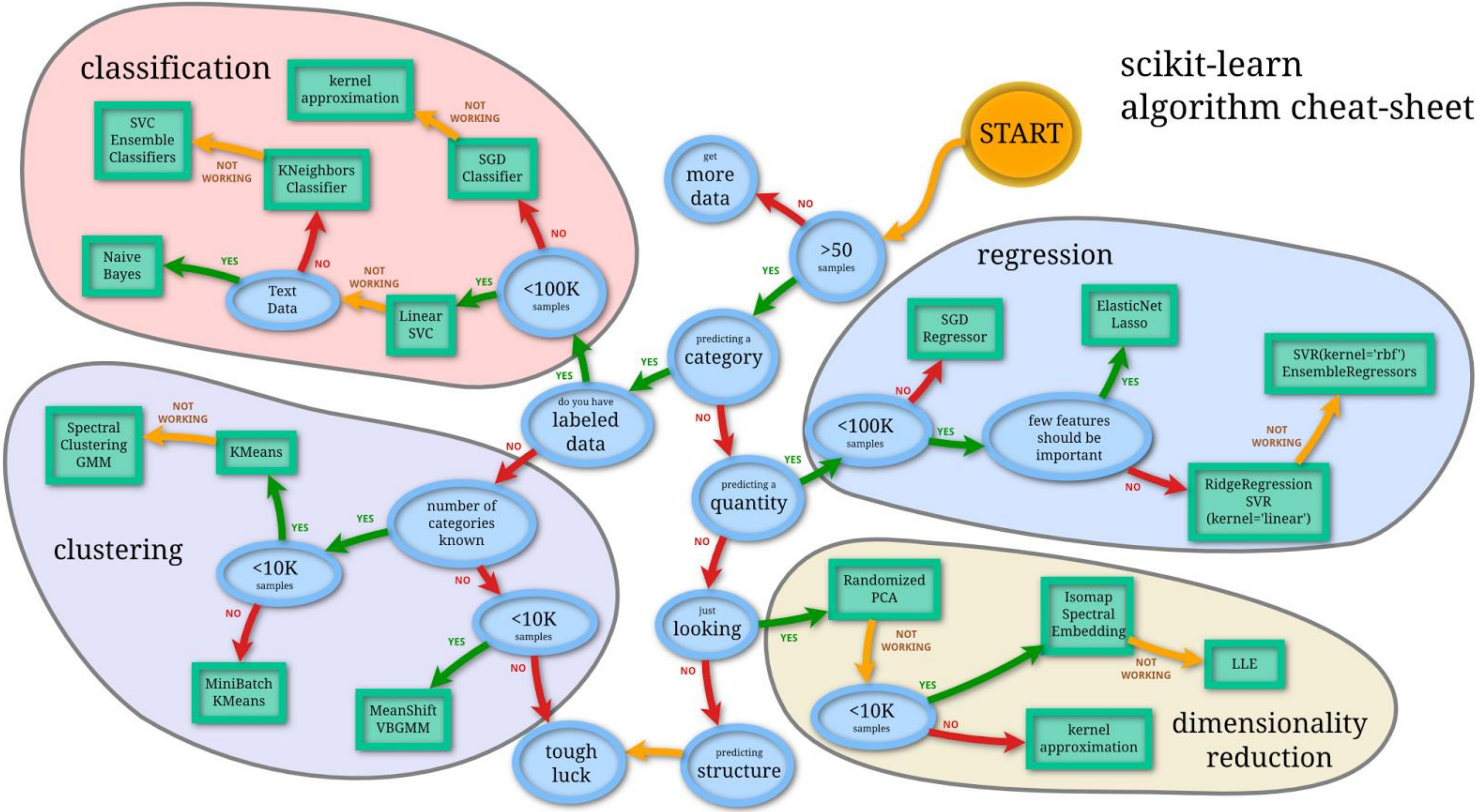
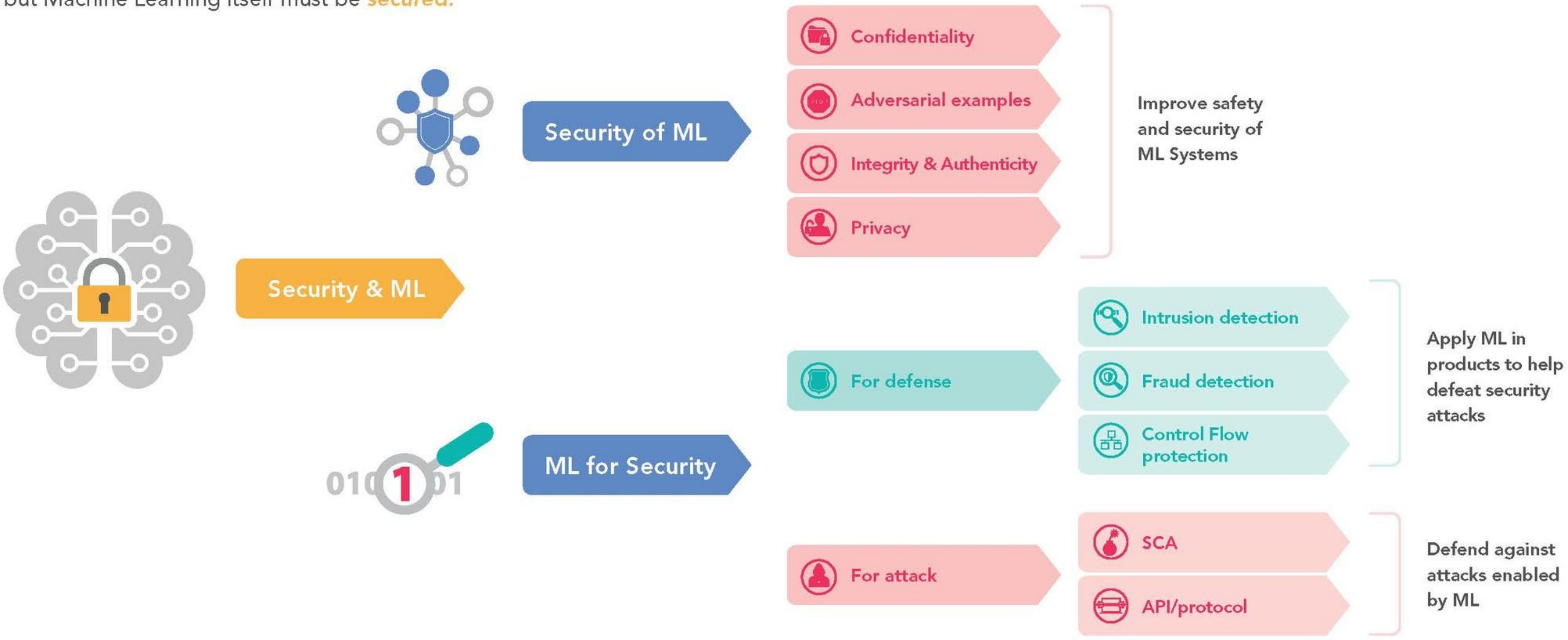


Figure from Nvidia blog post:
https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai


scikit-learn
algorithm cheat-sheet
START
>50 samples
get more data
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
classification
<100K samples
SGD Classifier
kernel approximation
Linear SVC
Text Data
Naive Bayes
KNeighbors Classifier
SVC Ensemble Classifiers
clustering
number of categories known
<10K samples
KMeans
Spectral Clustering GMM
MiniBatch KMeans
<10K samples
MeanShift VBGMM
regression
<100K samples
SGD Regressor
few features should be important
ElasticNet Lasso
RidgeRegression SVR (kernel='linear')
SVR(kernel='rbf') EnsembleRegressors
dimensionality reduction
Randomized PCA
<10K samples
Isomap Spectral Embedding
LLE
kernel approximation
YES
NO
NOT WORKING

Where Machine Learning, Security & Privacy Intersect

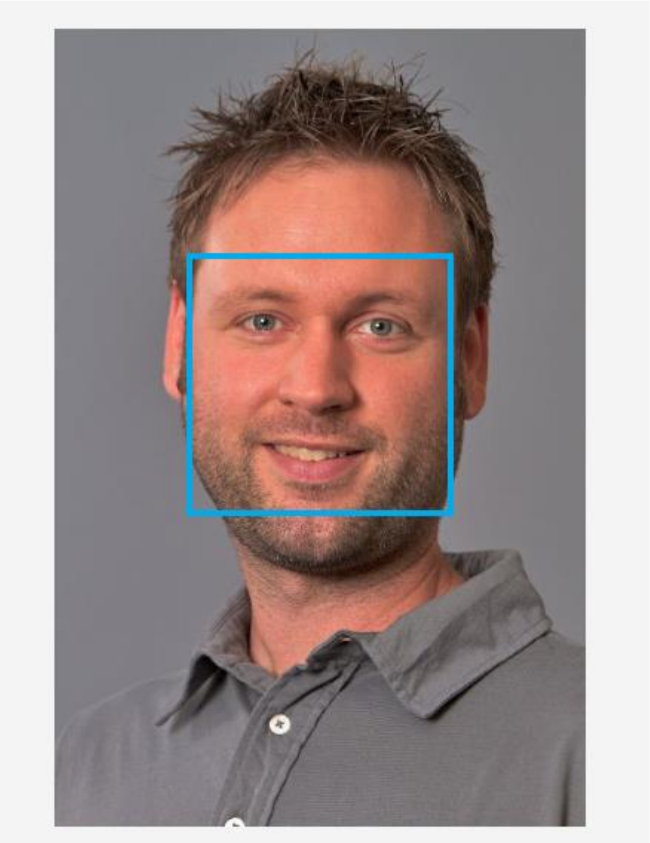
Machine Learning can *contribute* to IoT Security – but Machine Learning itself must be *secured.*

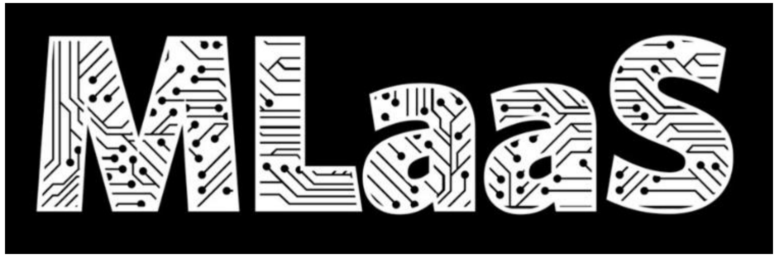


Model Cloning

Image source: **Matrix Revolutions movie poster**

Example: Microsoft Azure Emotions Recognition

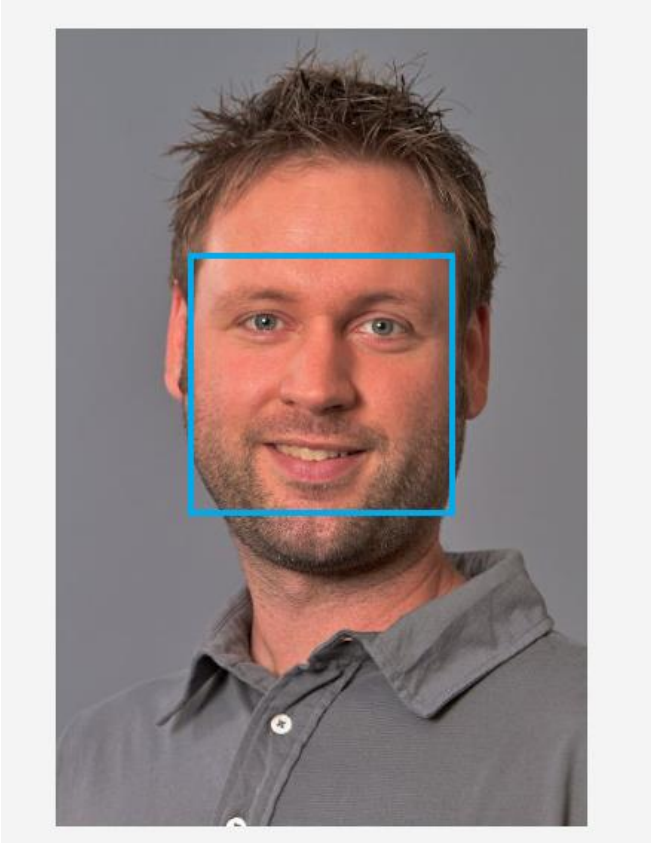


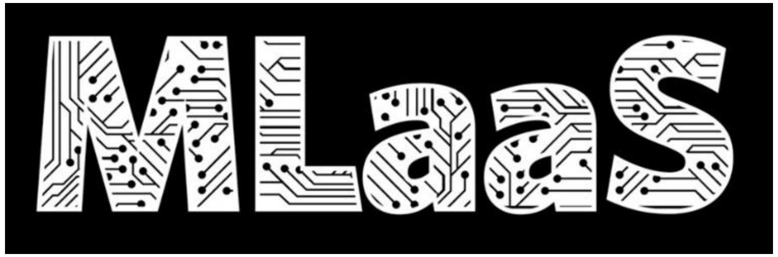



```
"scores": {
  "anger": 2.03898679E-07,
  "contempt": 0.0007247706,
  "disgust": 6.056115E-07,
  "fear": 1.0638247E-09,
  "happiness": 0.9959635,
  "neutral": 0.00329714641,
  "sadness": 4.30003233E-08,
  "surprise": 1.36911349E-05
}
```


- https://azure.microsoft.com/en-us/services/cognitive-services/emotion

Example: Microsoft Azure Emotions Recognition

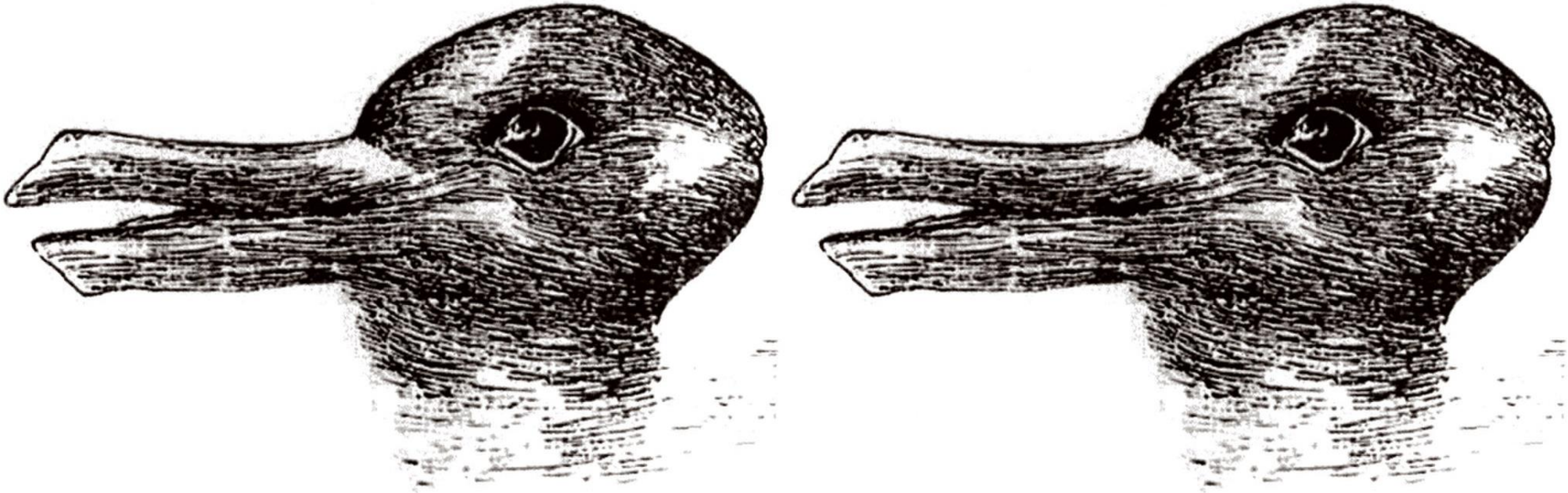





```
"scores": {
  "anger": 2.03898679E-07,
  "contempt": 0.0007247706,
  "disgust": 6.056115E-07,
  "fear": 1.0638247E-09,
  "happiness": 0.9959635,
  "neutral": 0.00329714641,
  "sadness": 4.30003233E-08,
  "surprise": 1.36911349E-05
}
```


Clone model for < $350 using random non-labeled data

- 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, Berriel, Badue, de Souza, Oliveira-Santos. *Copycat CNN: Stealing Knowledge by Persuading Confession with Random Non-Labeled Data*. *arXiv preprint* (2018).

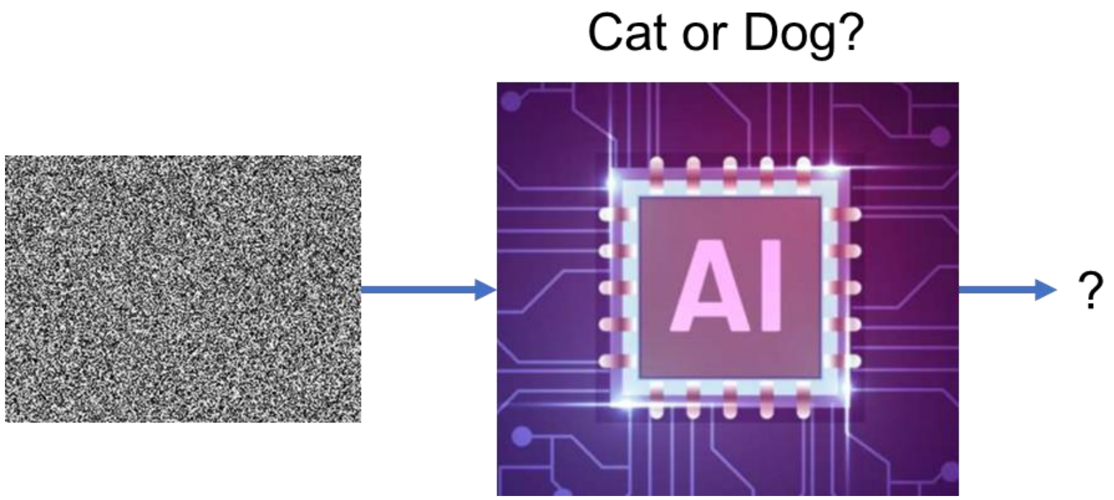


Adversarial Examples | Optical Illusions for Machines

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

Misclassification versus Adversarial Examples

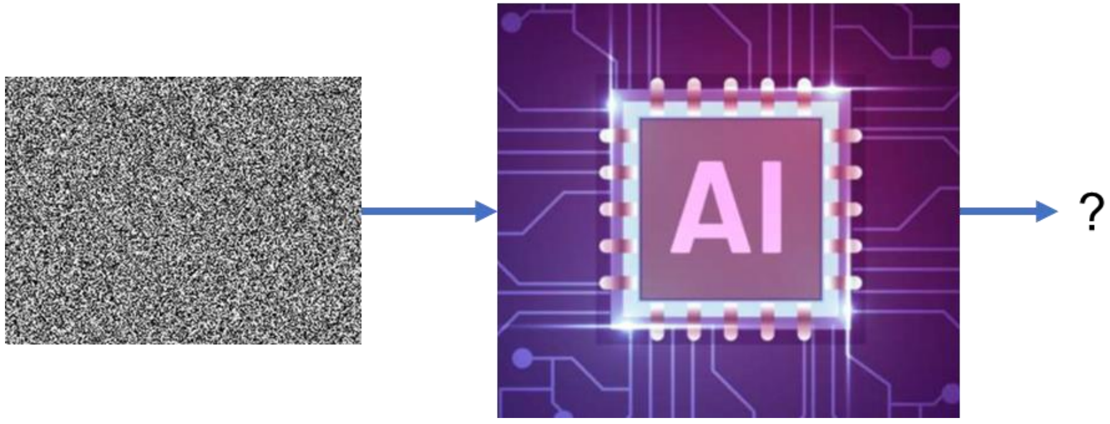
- Biggio, Corona, Maiorca, Nelson, Srndic, Laskov, Giacinto, Roli: *Evasion attacks against machine learning at test time*. In Machine Learning and Knowledge Discovery in Databases, 2013.
- *Goodfellow, Shlens, Szegedy: Explaining and harnessing adversarial examples.* In arXiv preprint 2014
- Szegedy,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



Misclassification versus Adversarial Examples

- 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.

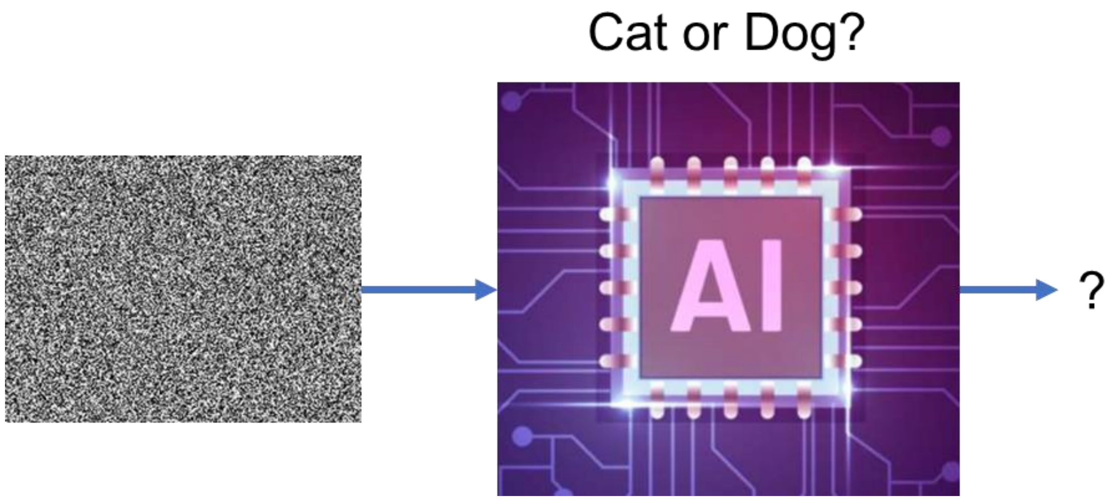
Cat or Dog?



~ 0.832 flowerpot

Misclassification versus Adversarial Examples

- 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.

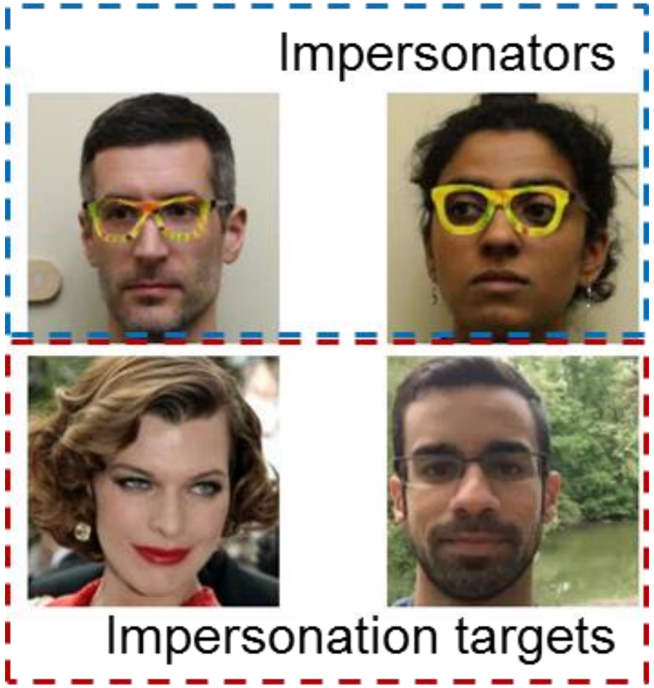


~ 0.832 flowerpot

~ 0.999 warplane

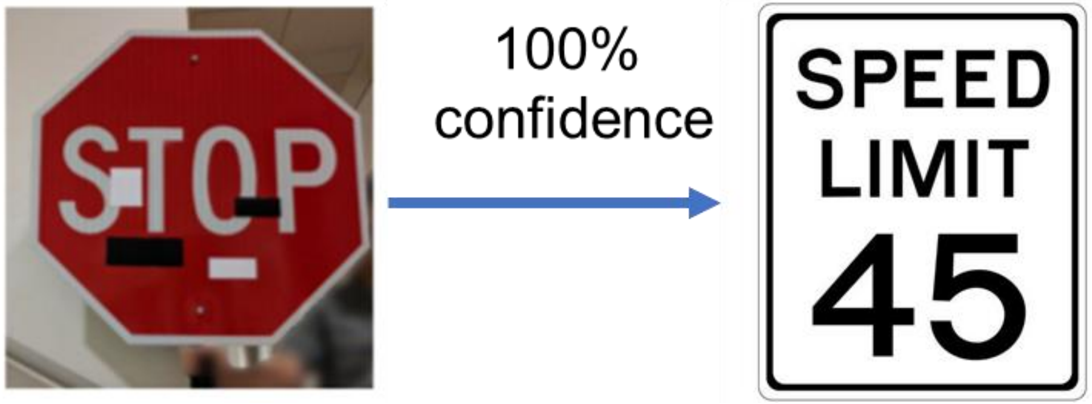
Impact in Practice

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



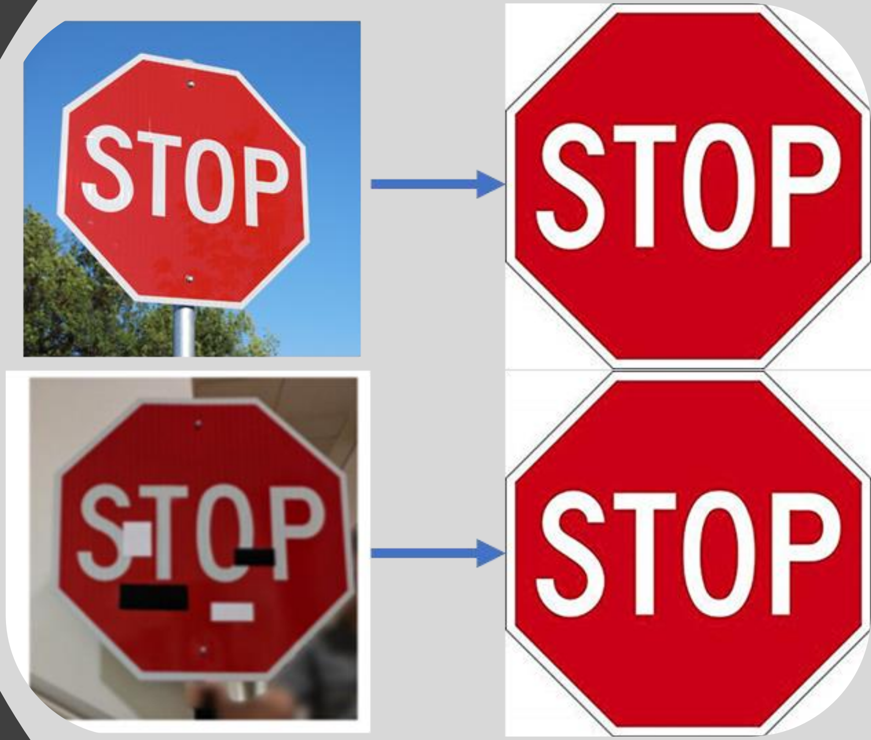
Eykholt, Evtimov, Fernandes, Li, Rahmati, Xiao, Prakash, Kohno, Song: *Robust Physical-World Attacks on Deep Learning Visual Classification*. In *IEEE Computer Vision and Pattern Recognition* 2018.

cleverhans

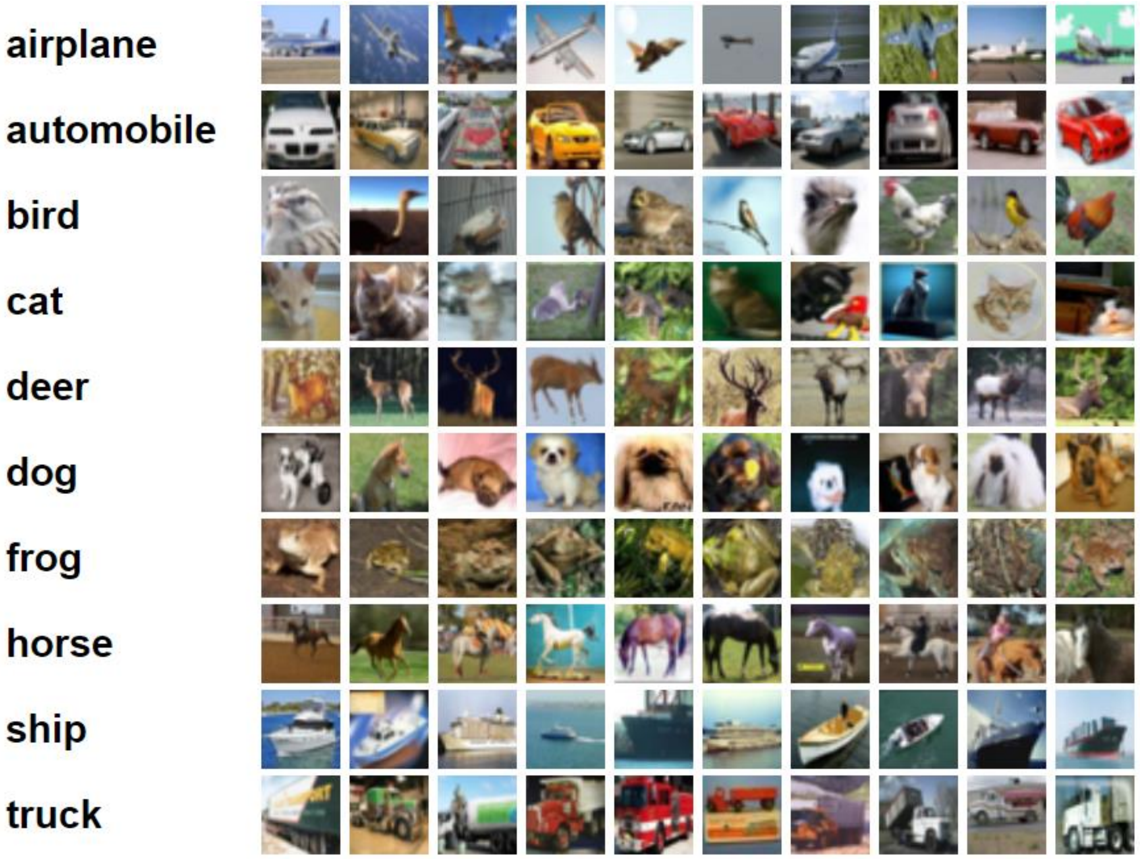
https://github.com/tensorflow/cleverhans

Papernot et al.: *Technical Report on the CleverHans v2.1.0 Adversarial Examples Library*, arXiv preprint 2018

Countermeasures?
Adversarial Training



Adversarial Training

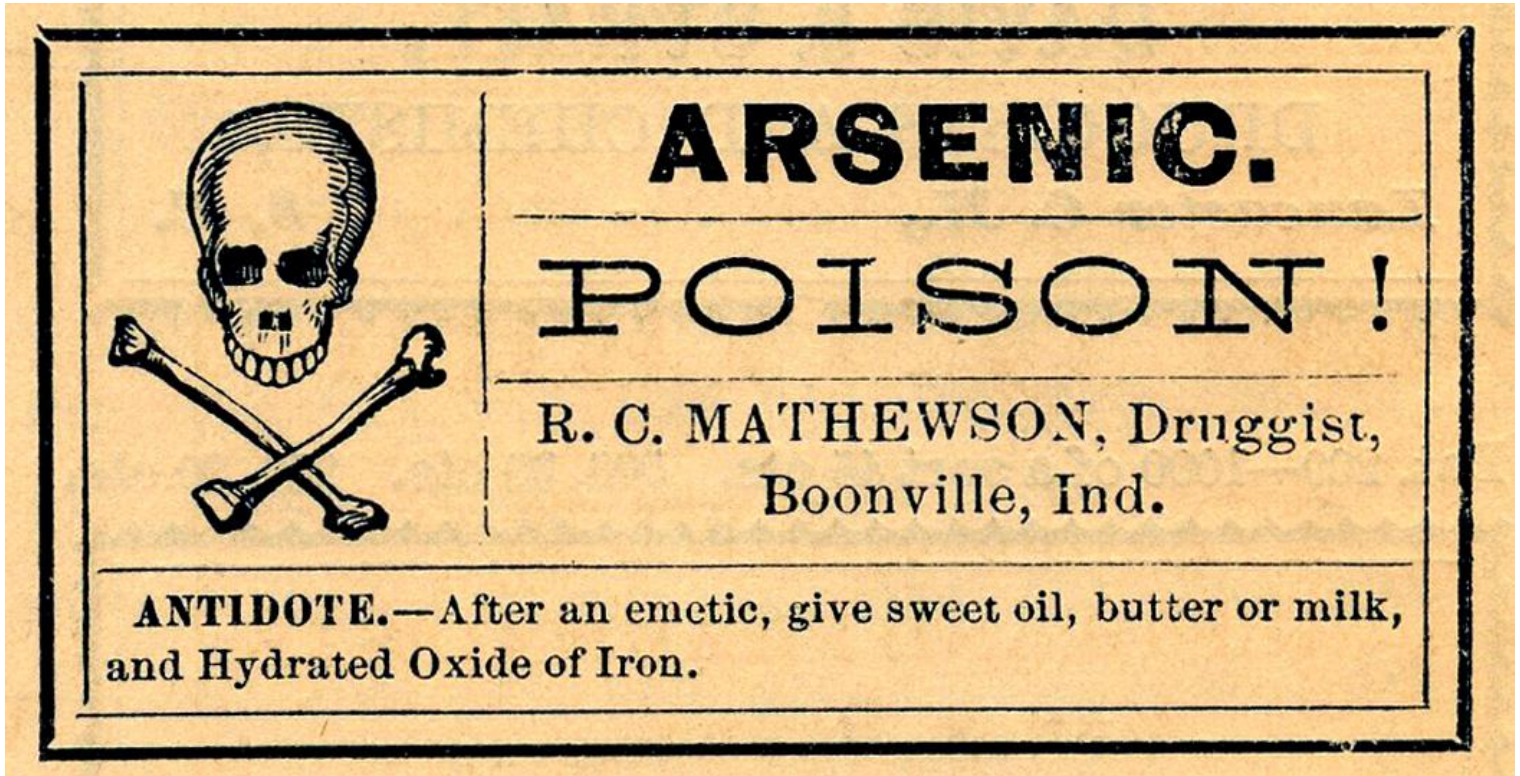


cleverhans

https://github.com/tensorflow/cleverhans

CIFAR-10 Model	Accuracy of the model	Adversarial examples that mislead the model
Original	87%	90%
Trained with adversarial examples	86%	17%

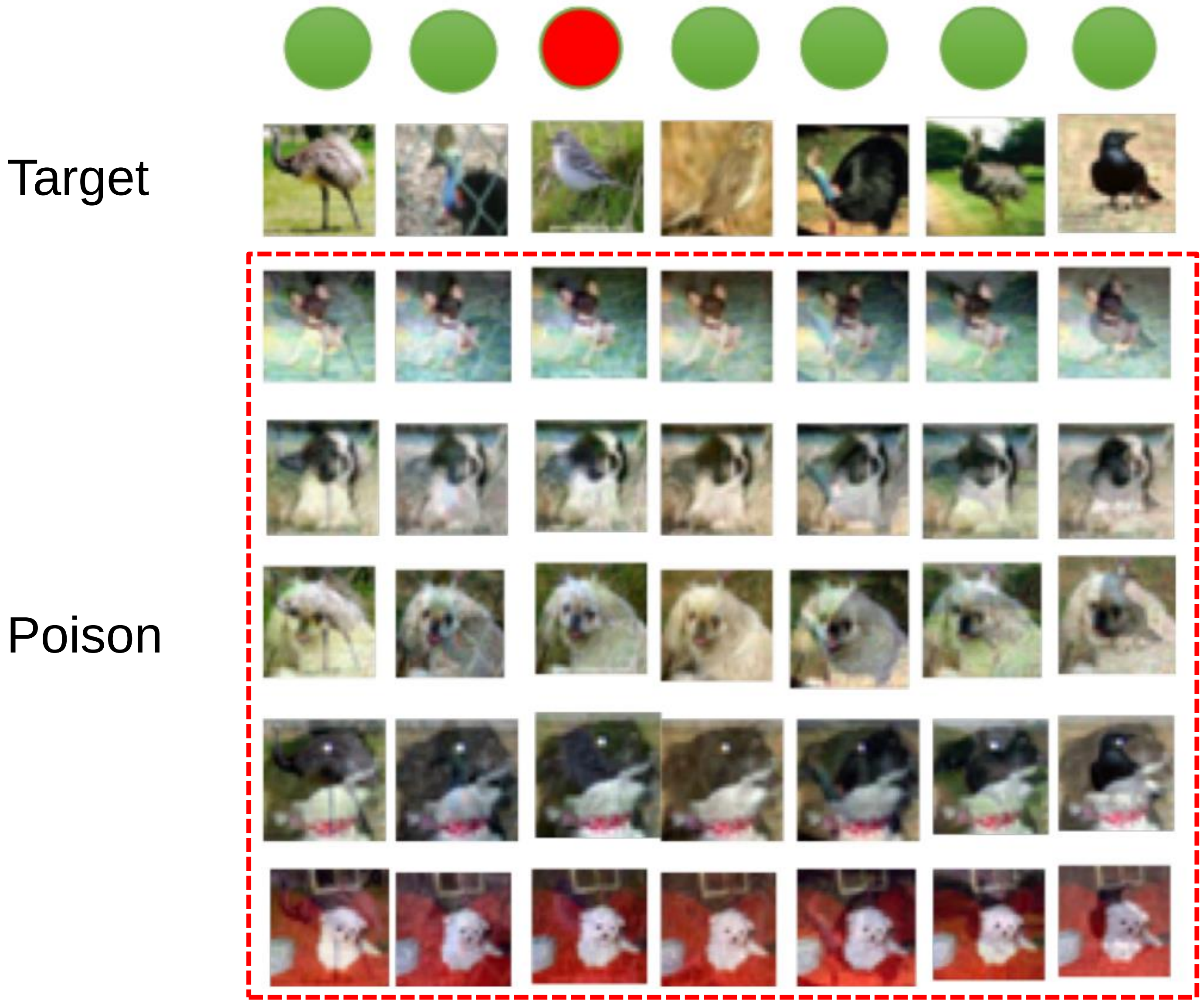
Data Poisoning



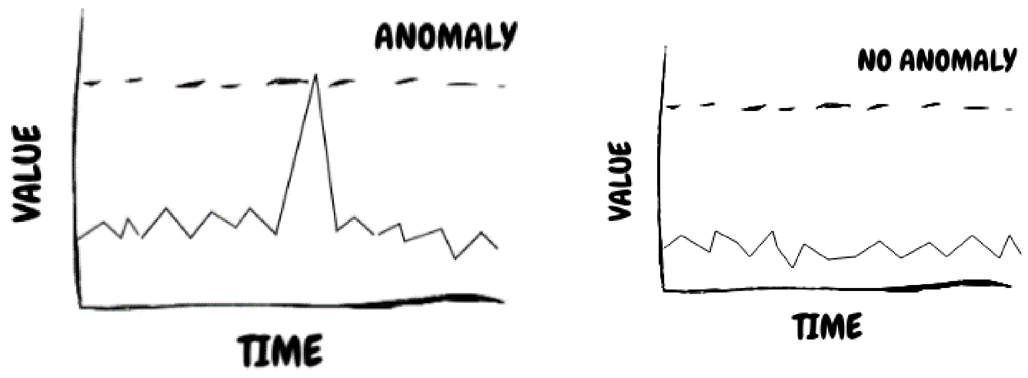
Barreno, Nelson, Sears, Joseph, and Tygar: *Can machine learning be secure?* In ACM CCS 2006.

Adversarial Training - Revisited

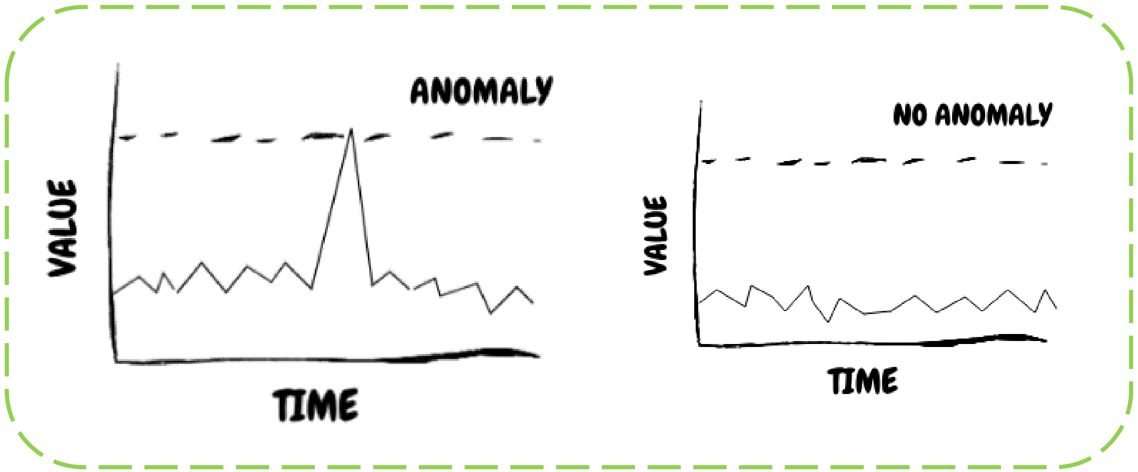
Shafahi, Huang, Najibi, Suciu, Studer, Dumitras, Goldstein: *Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks*. *arXiv preprint 2018.*

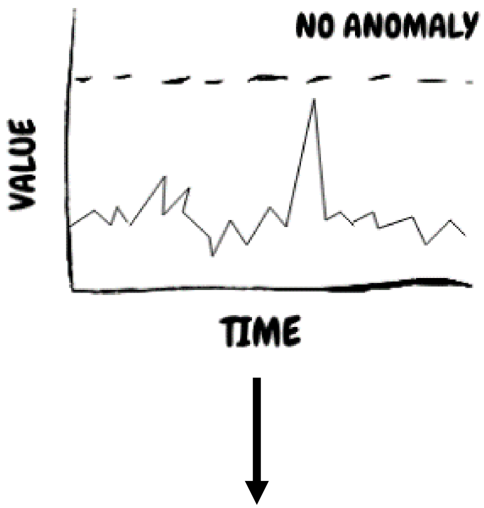


Anomaly detection

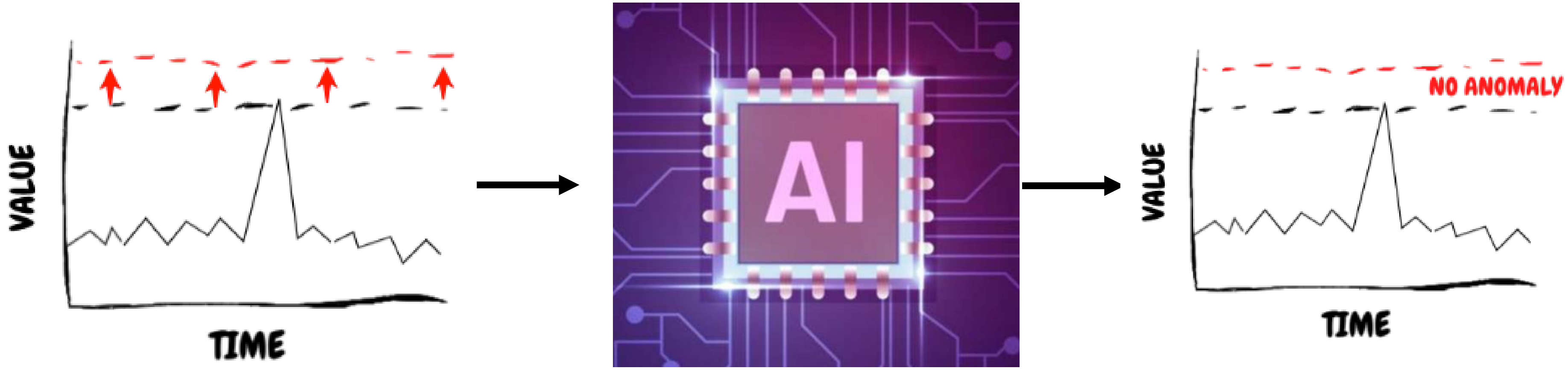


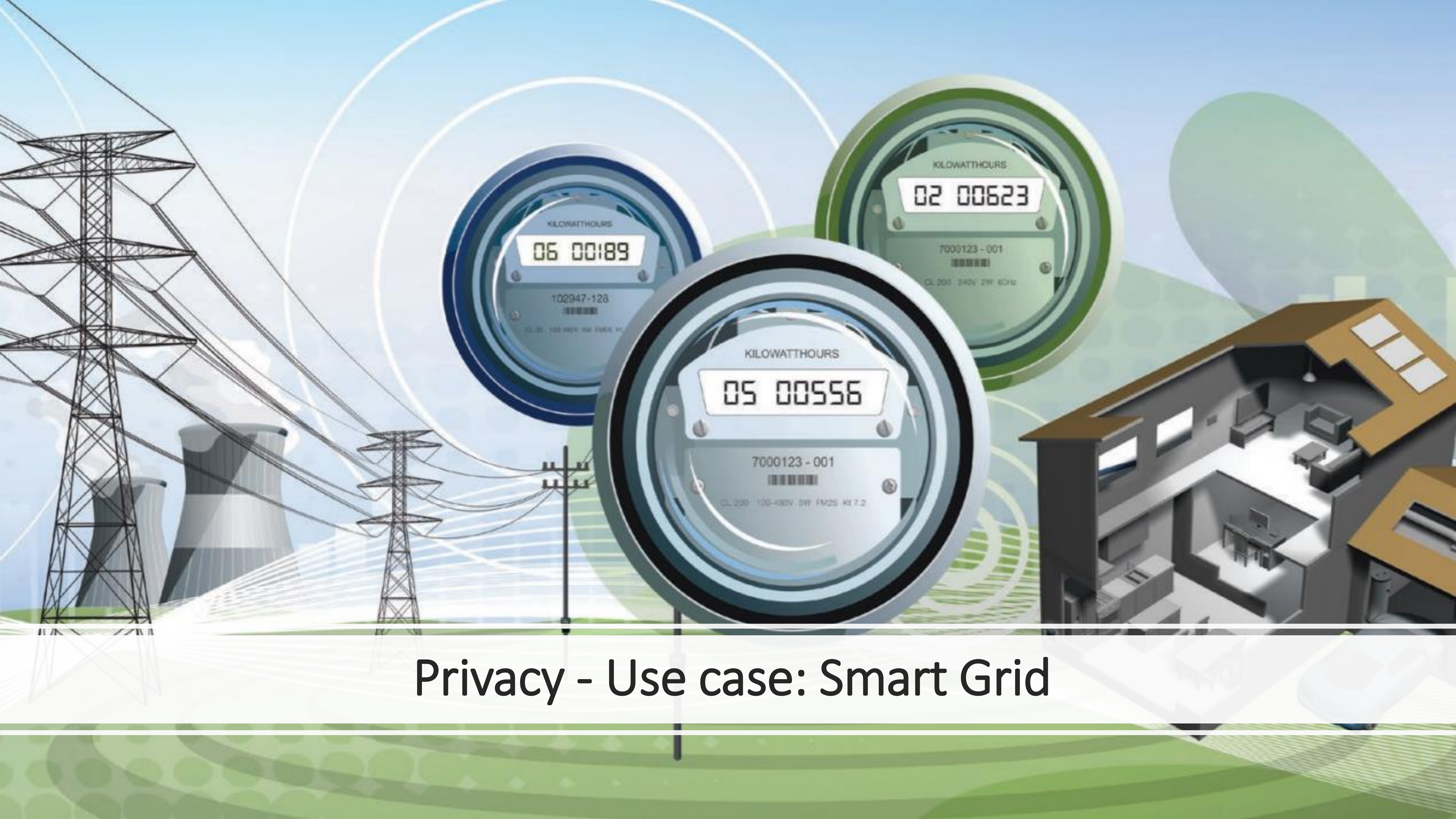
Anomaly detection





- **Incremental learning**
- **Offline learning**





Privacy - Use case: Smart Grid

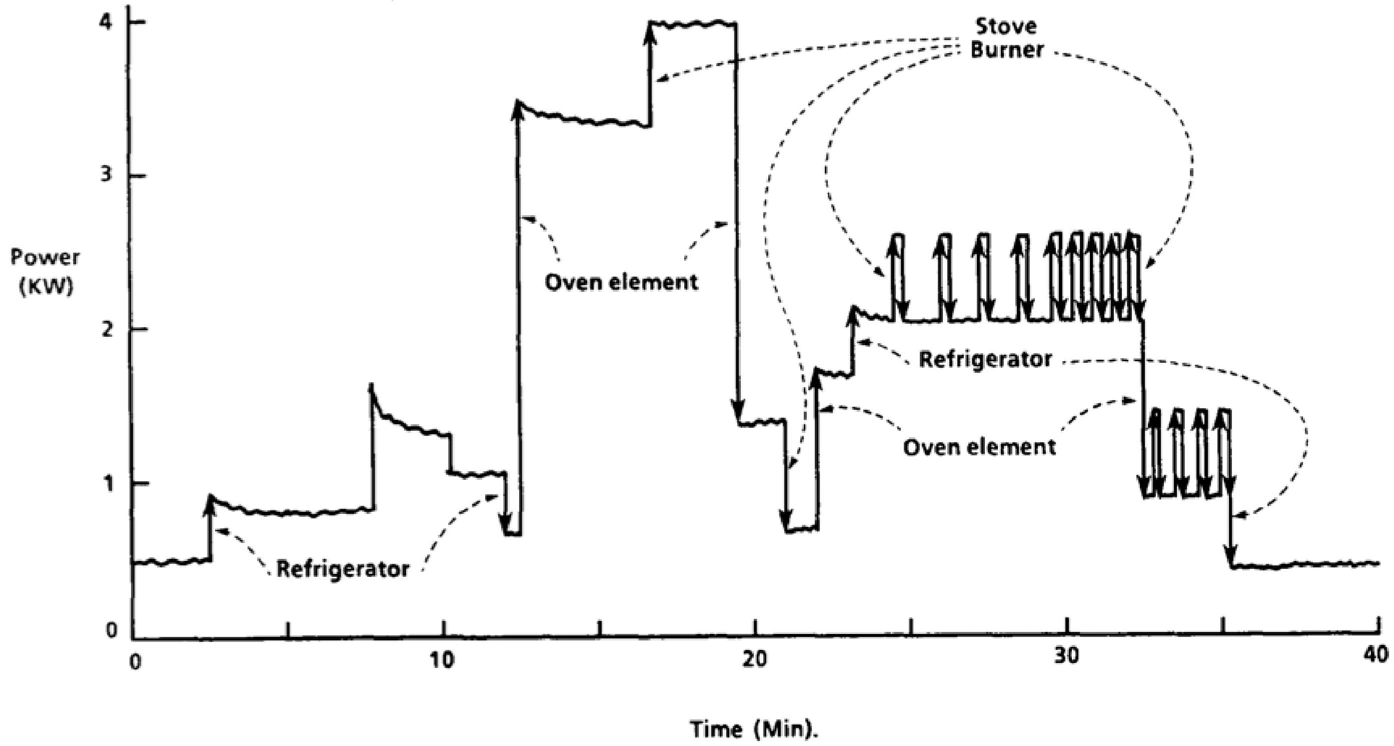
Hernandez, Zorita, Aguiar, Carro, Sanchez-Esguevillas, Lloret, Massana: ***A survey on electric power demand forecasting: Future trends in smart grids, microgrids and smart buildings***. IEEE Communications Surveys and Tutorials, 2014.

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?

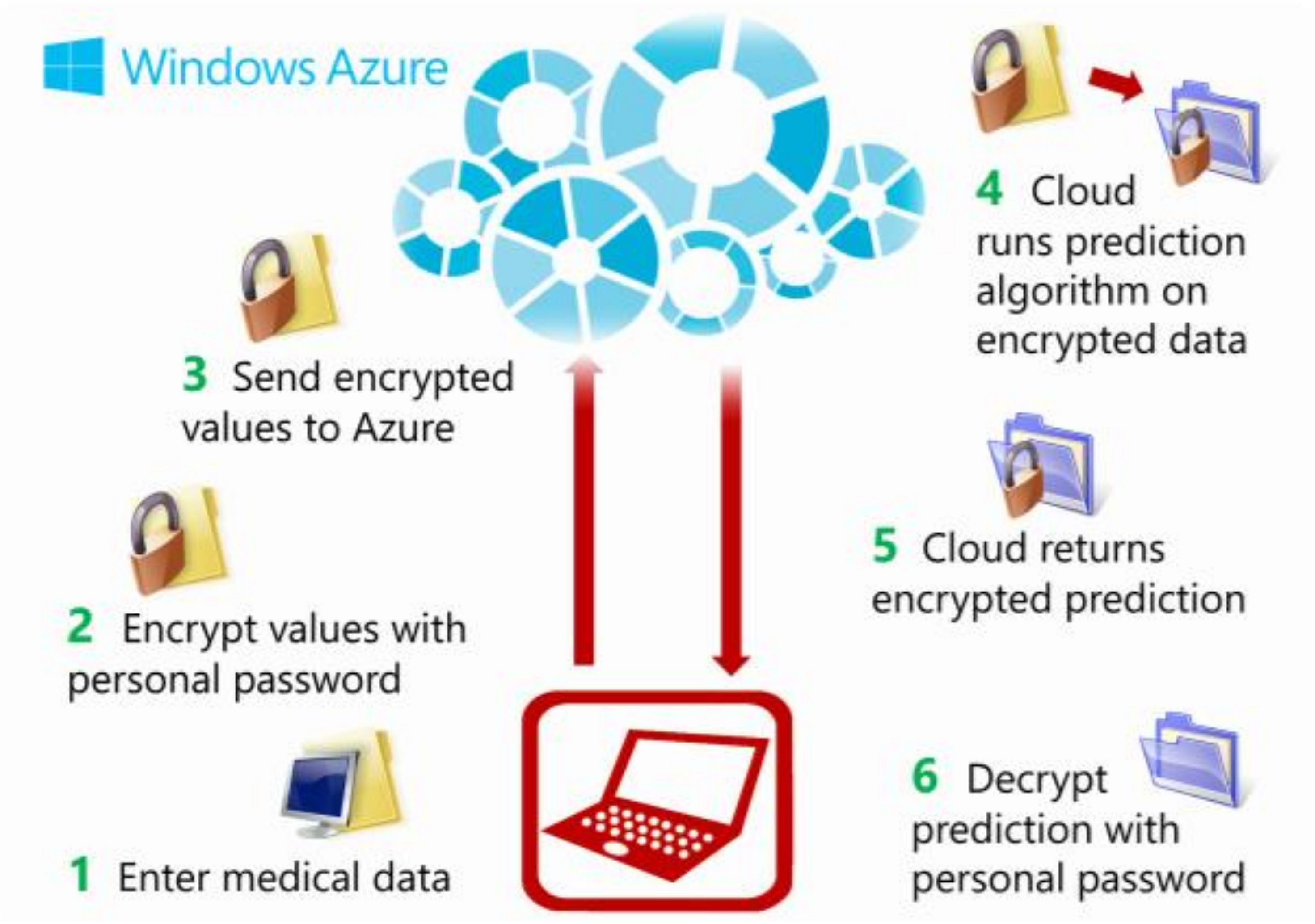
Invalid usage

- Insurance or warranty

Real-time information

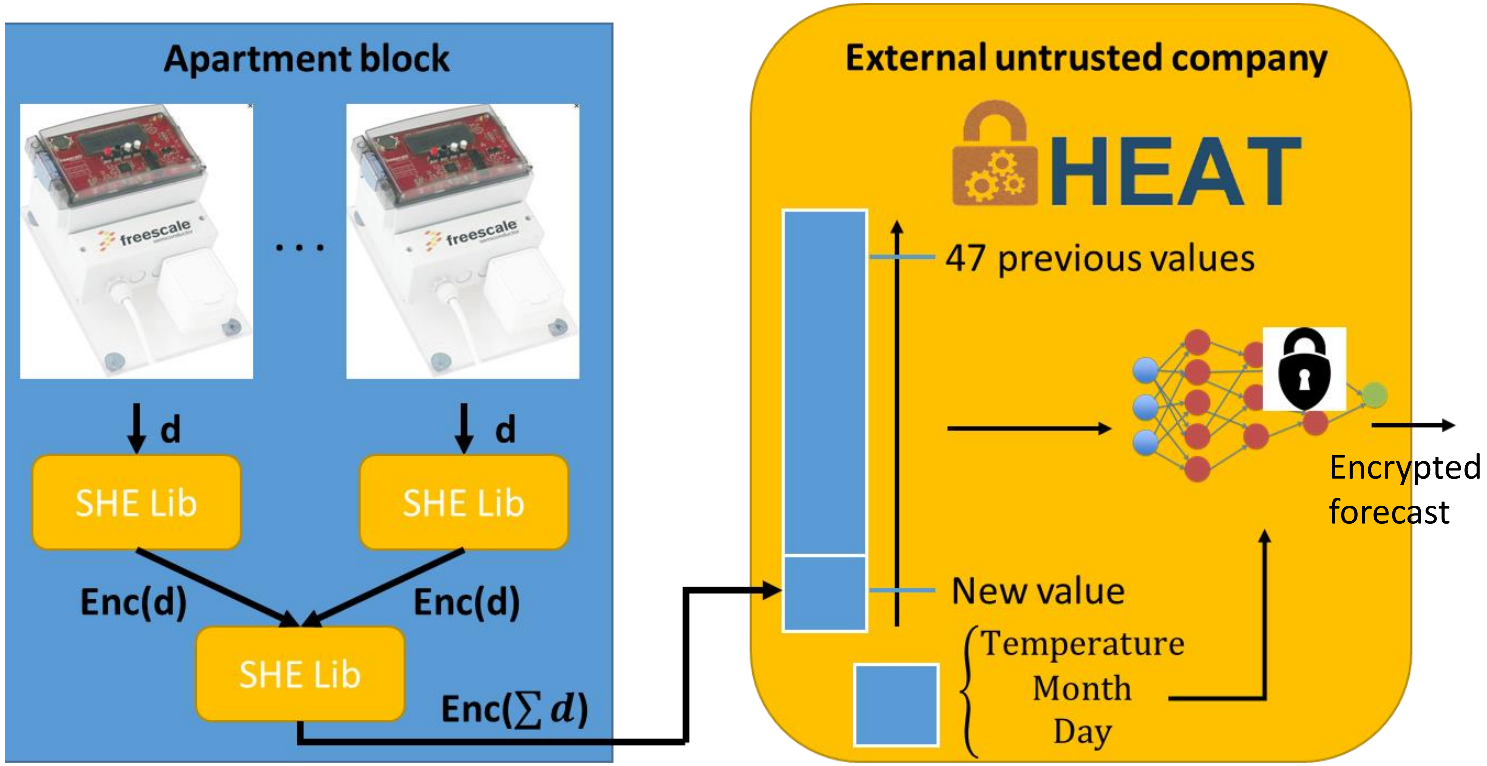
- Number of people in a household
- Are you on holidays?

Computing on Encrypted Data



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

Machine Learning using Encrypted Data

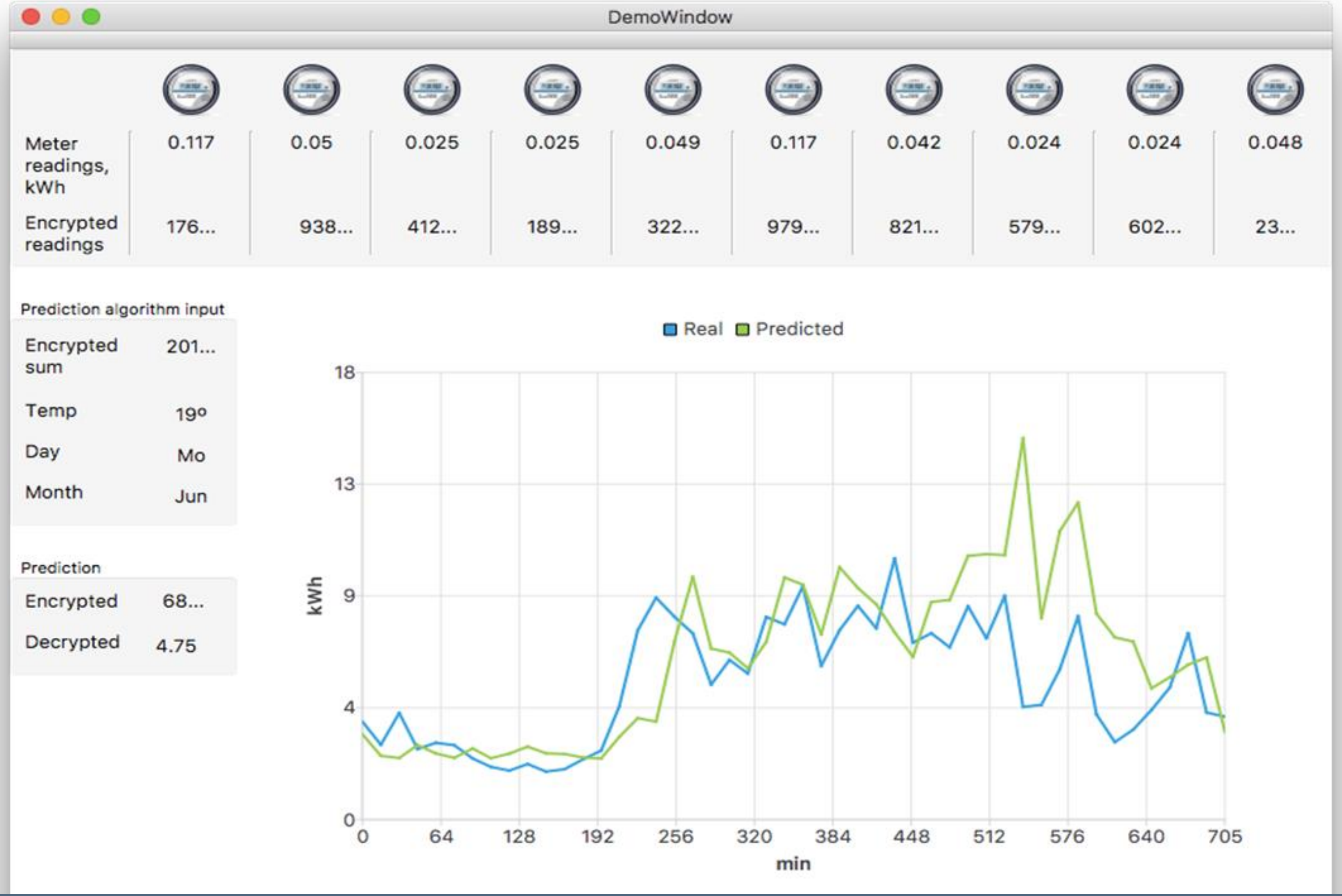


- Forecast power consumption for next half hour in ≈ 2.5 seconds to evaluate
- Neural network
  Inputs: 51
  Hidden layers: 3 (8 → 4 → 2)
  Output: 1

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

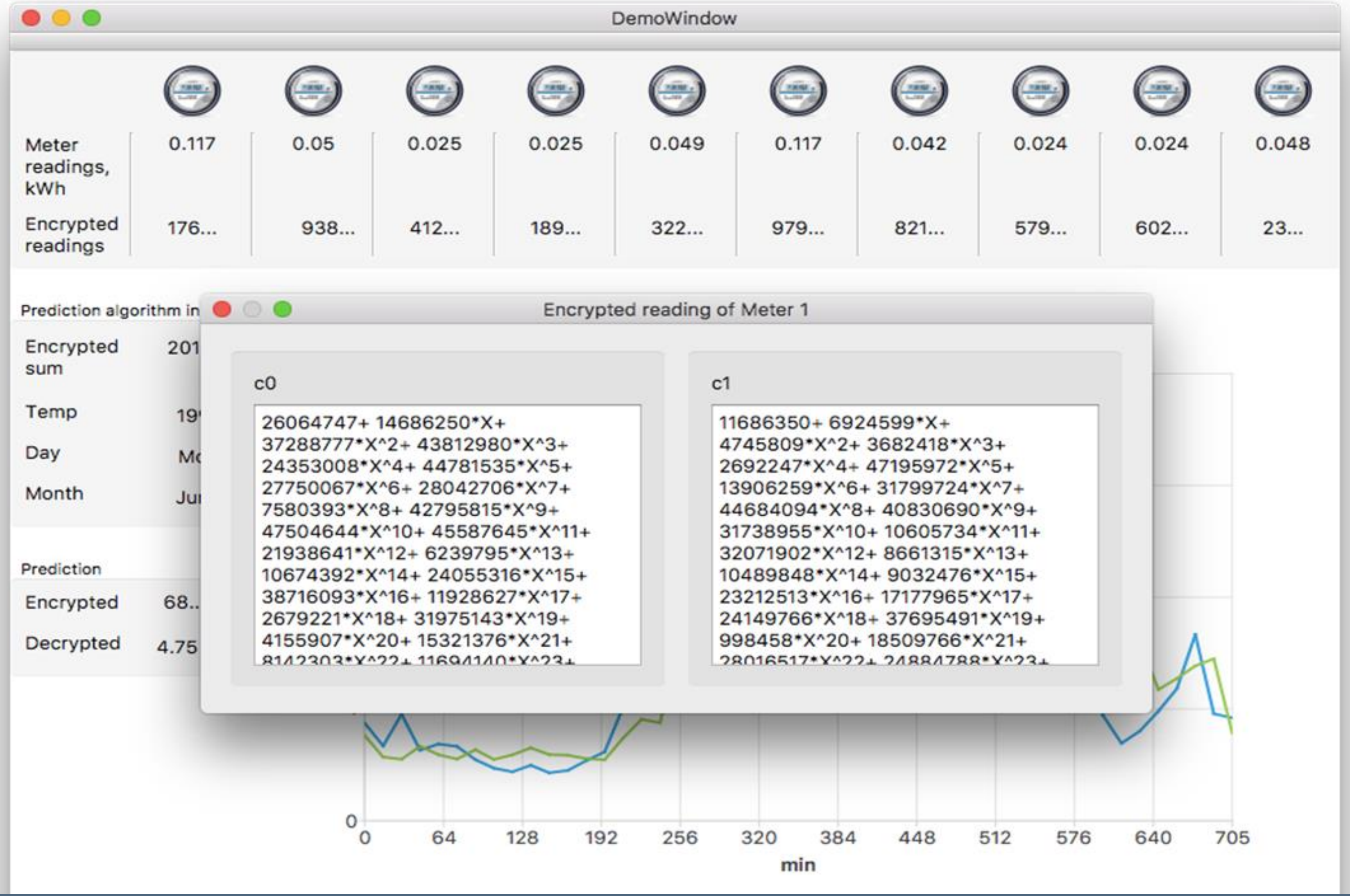
Machine Learning using Encrypted Data



- Forecast power consumption for next half hour in ≈ 2.5 seconds to evaluate
- 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



Machine Learning using Encrypted Data

- Forecast power consumption for next half hour in ≈ 2.5 seconds to evaluate
- 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

Conclusions

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

- [+] <u>Cryptography to the rescue</u> for some problems

- [-] Expect zero-day attacks against machine learning models


NXP
SECURE CONNECTIONS
FOR A SMARTER WORLD
& MACHINE LEARNING