香港城市大學
City University of Hong Kong

CS 6290

Privacy-enhancing Technologies

Department of Computer Science

Slides credit in part from Florian Tramèr.

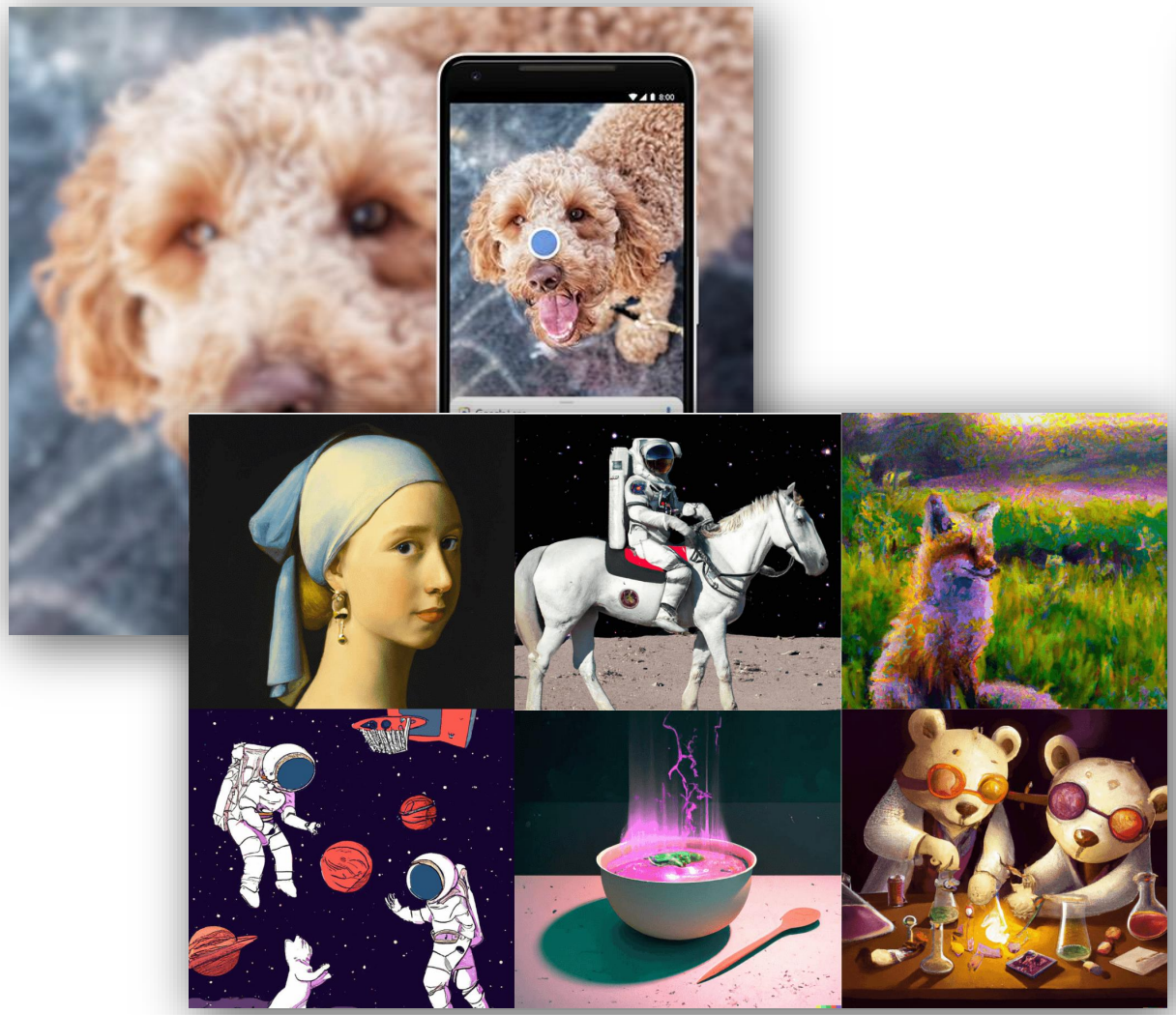
Lecture 9 – Adversarial Machine Learning

Prof. Cong WANG

CS Department

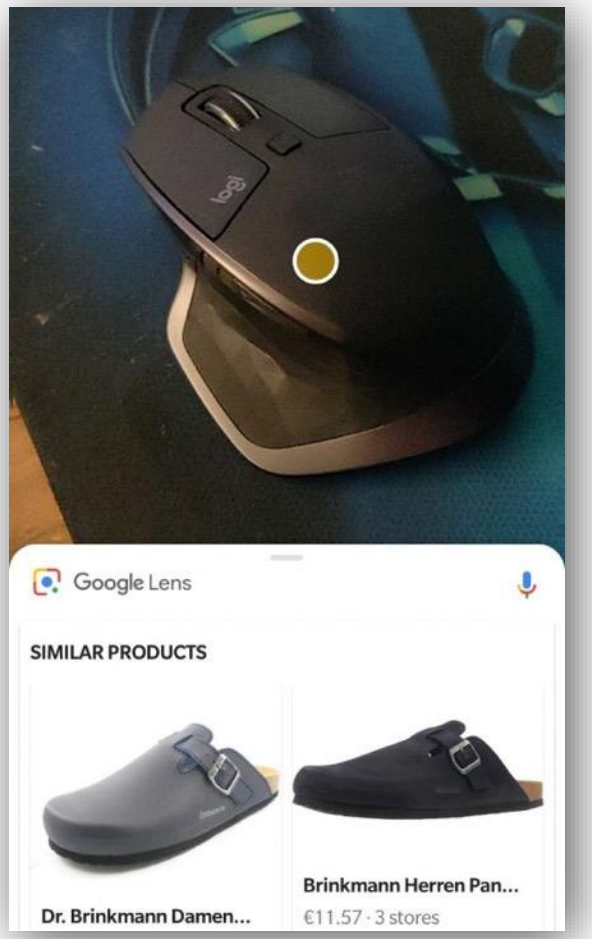
City University of Hong Kong

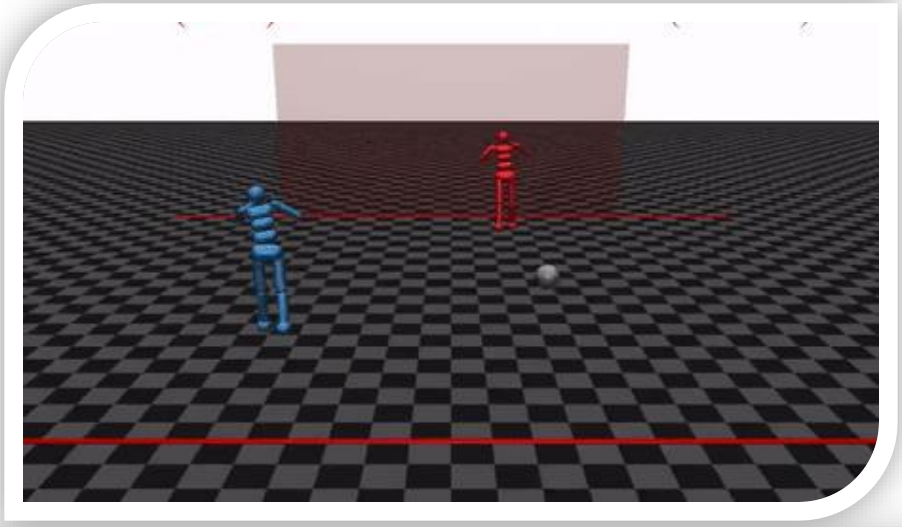
Machine learning works

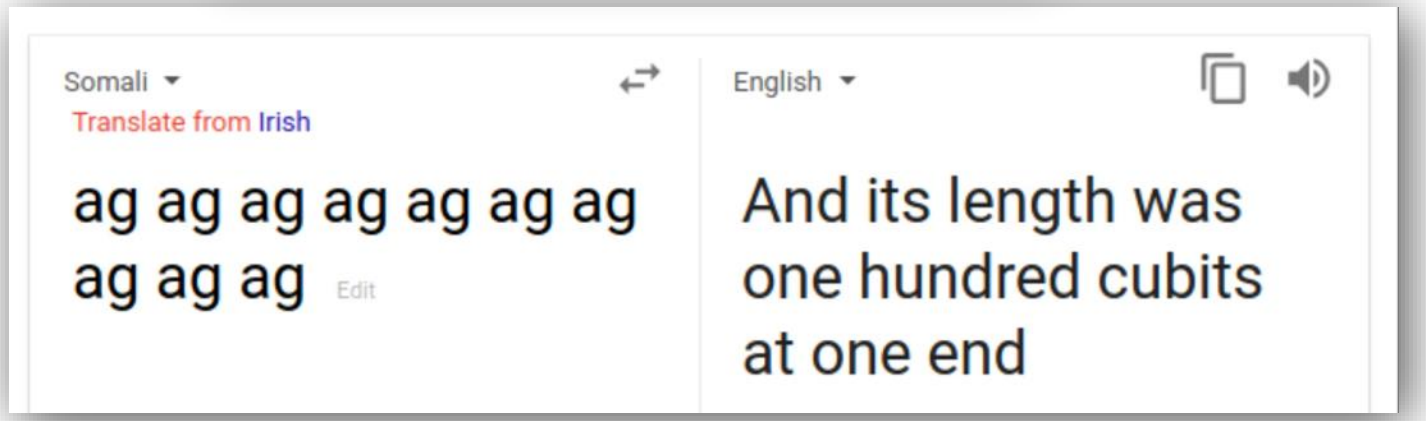


Machine learning works **most of the time!**

many applications tolerate occasional failures







Machine learning can also fail disastrously

Critical mistakes…

theguardian

Uber crash shows 'catastrophic failure' of self-driving technology, experts say

Machine learning can also fail disastrously

Critical mistakes...

theguardian

Uber crash shows 'catastrophic failure' of self-driving technology, experts say

Direct attacks...

The New York Times

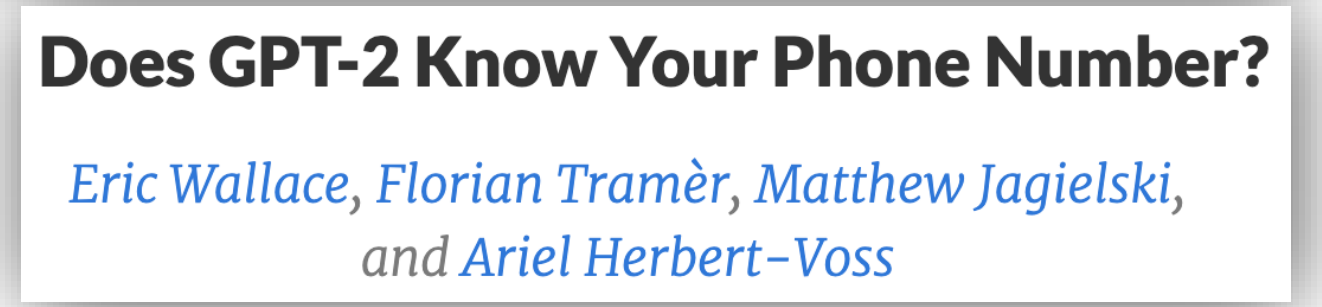
Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.

Machine learning can also fail disastrously

Critical mistakes...

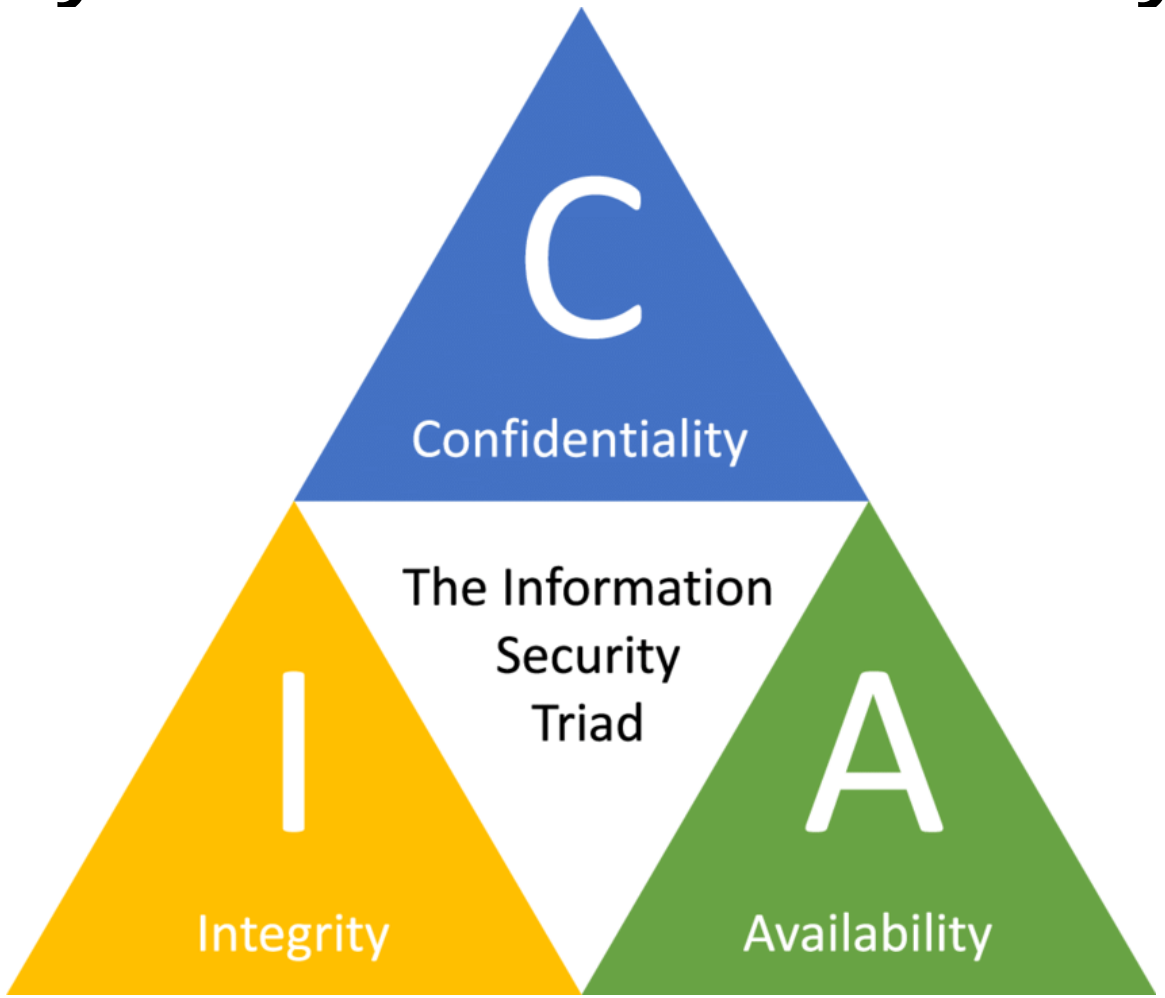
Direct attacks...

Private data leaks...

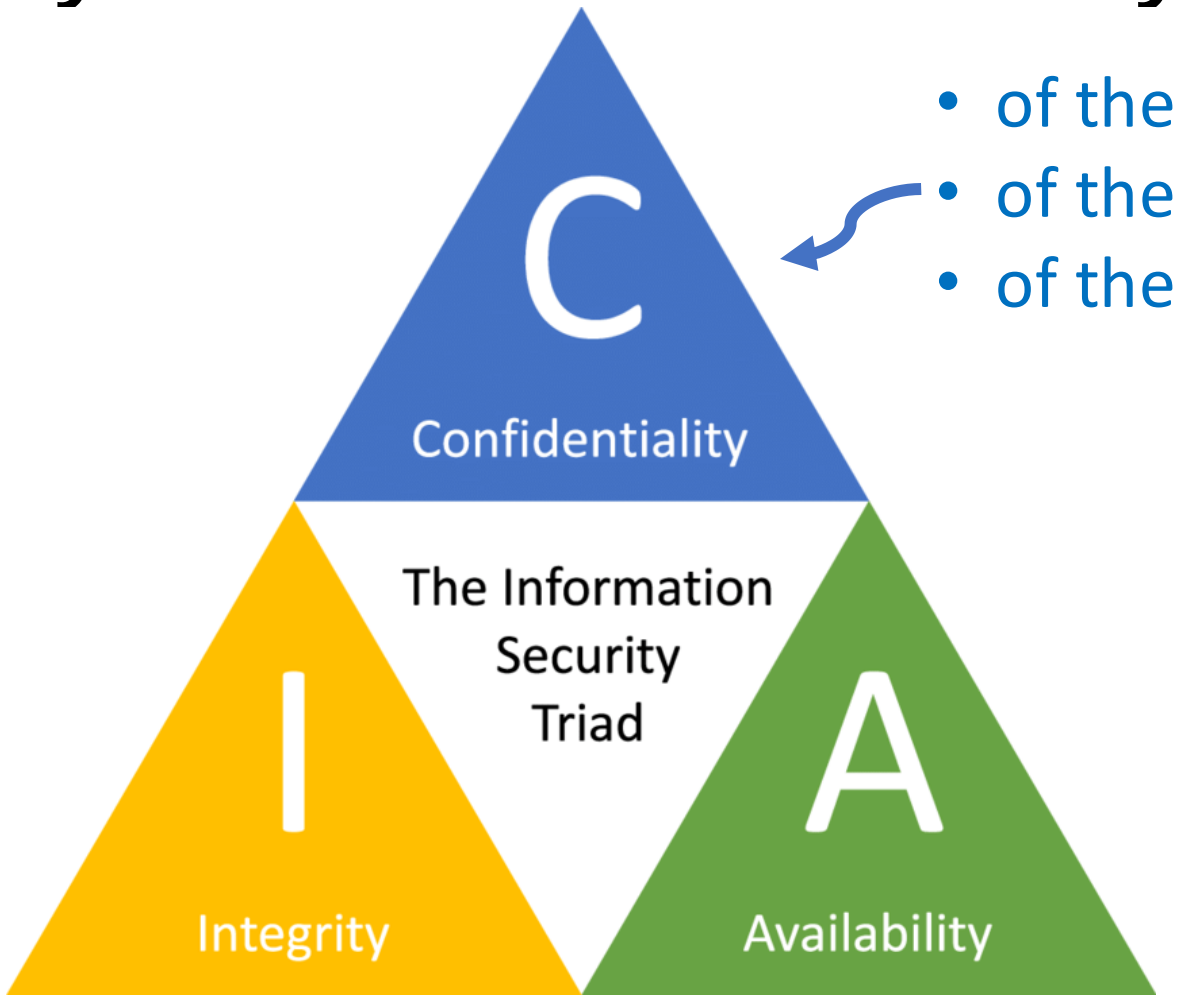


What does this mean for computer security?

ML Security = traditional security

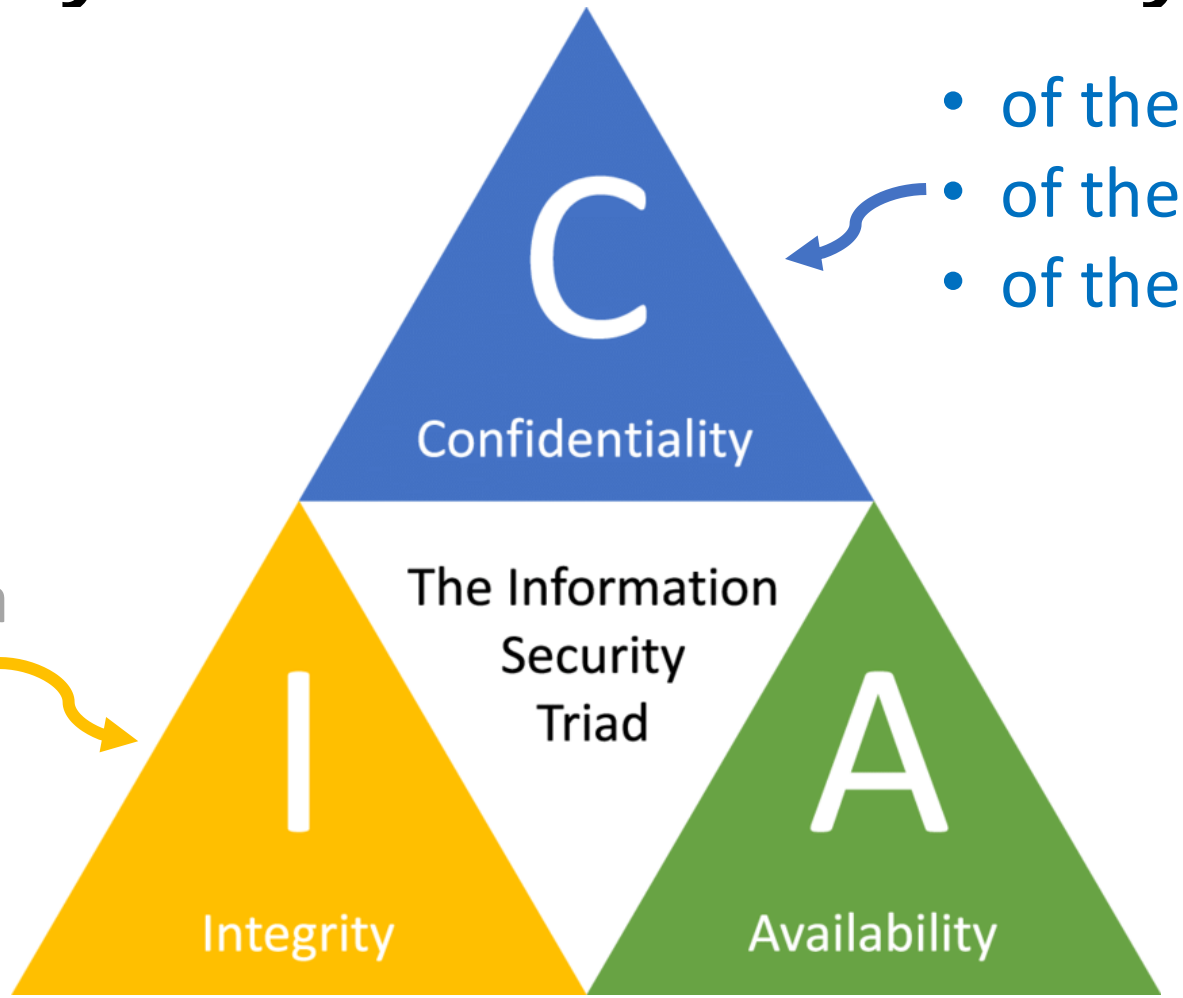


ML Security = traditional security



- of the ML model
- of the training data
- of the eval data

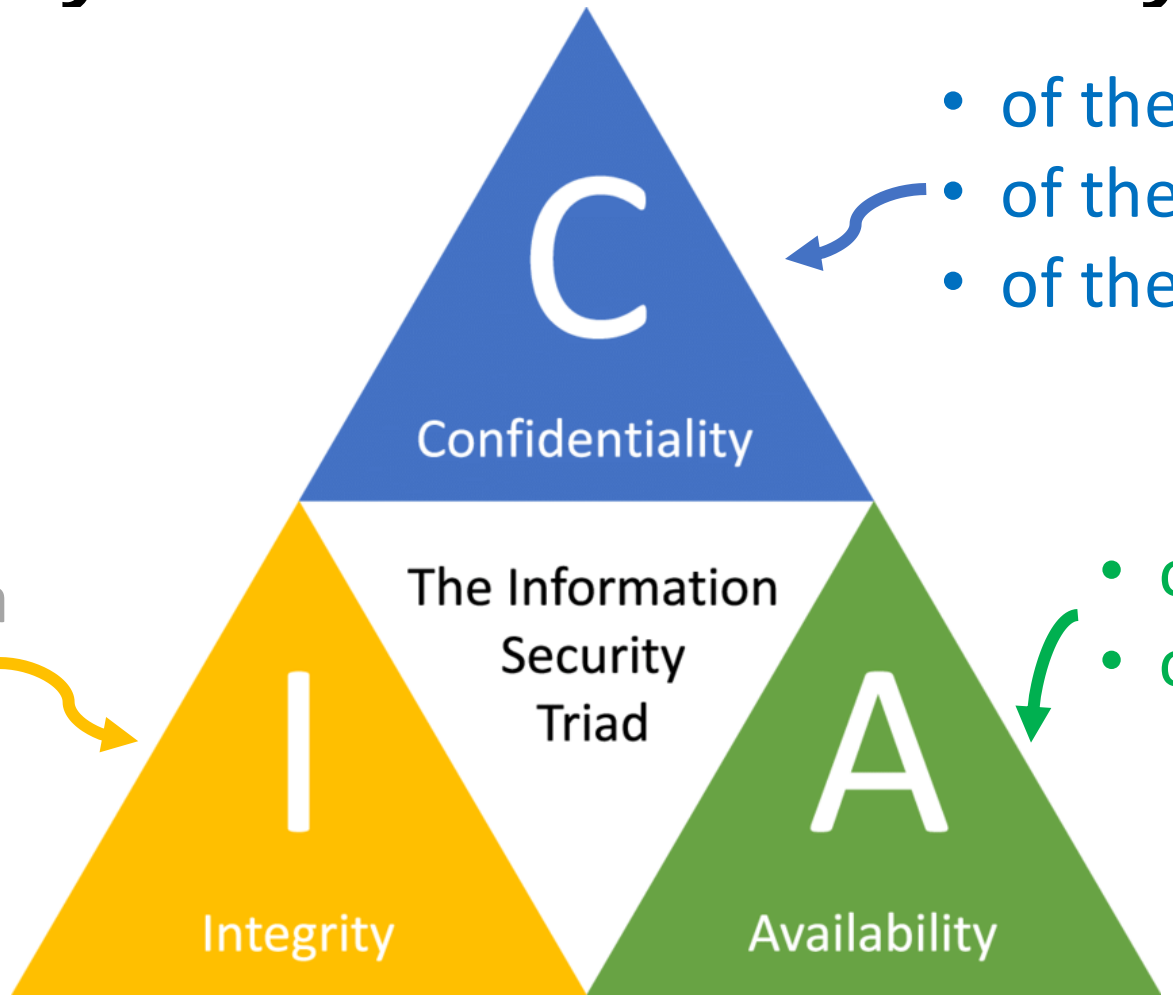
ML Security = traditional security



- of the ML model
- of the training data
- of the eval data

- of the training data
- of the model
- of the eval data

ML Security = traditional security



- of the ML model
- of the training data
- of the eval data

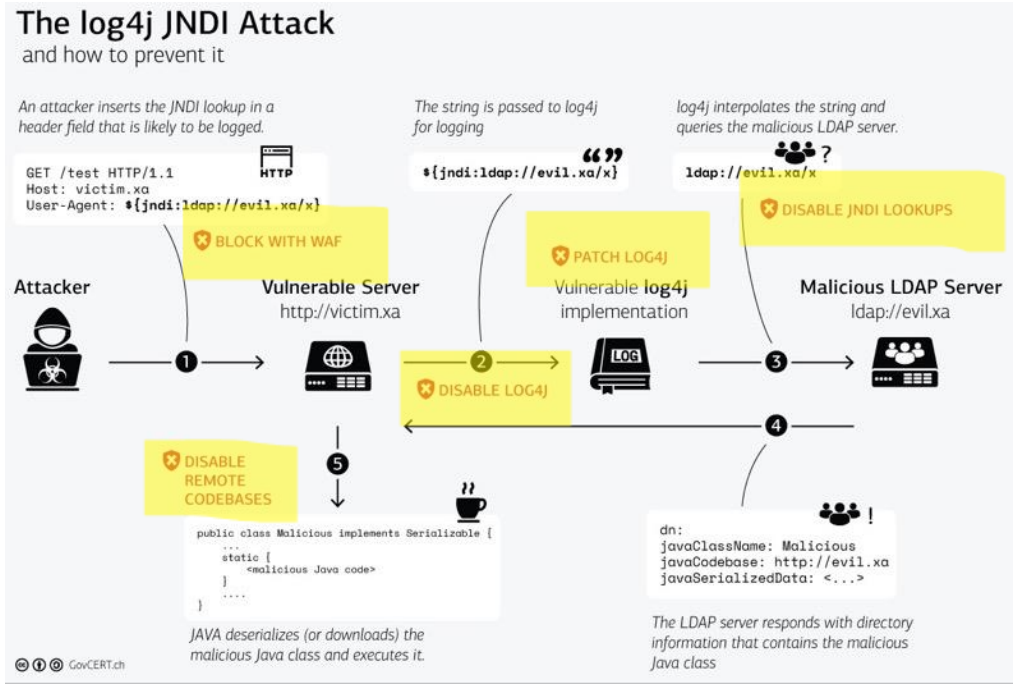
- of the training data
- of the model
- of the eval data

- of training
- of the trained model

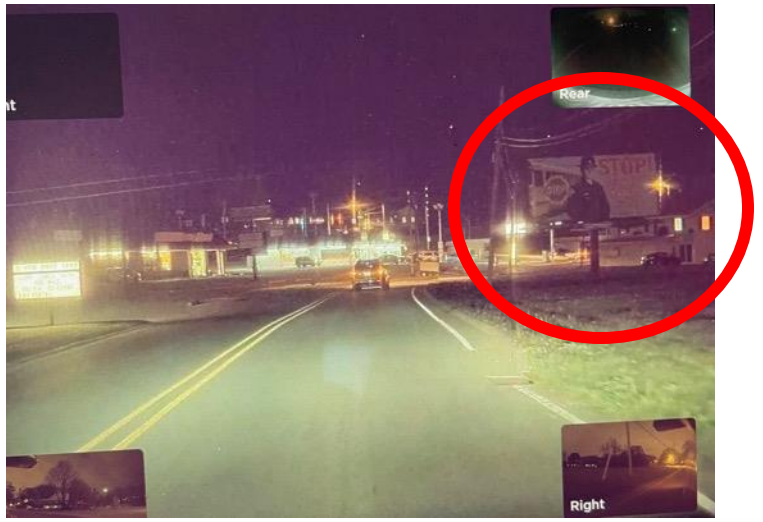
ML Security ≠ traditional security

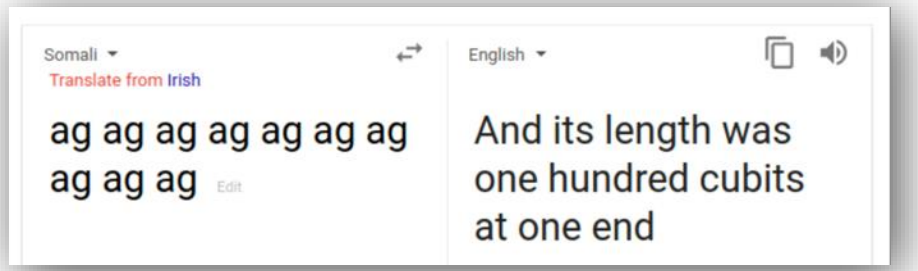
Fixing “standard” bugs is “easy”

(finding them isn’t…)



How do we fix ML bugs?





“Traditional” ML Security

In today’s lecture, we focus on “traditional” deep neural networks (e.g., MLPs, CNNs):

- Simpler to understand
- Concepts apply to advanced models

We will discuss on modern models in the next lecture ☺

The “ML Rocket”

Engine (Model)

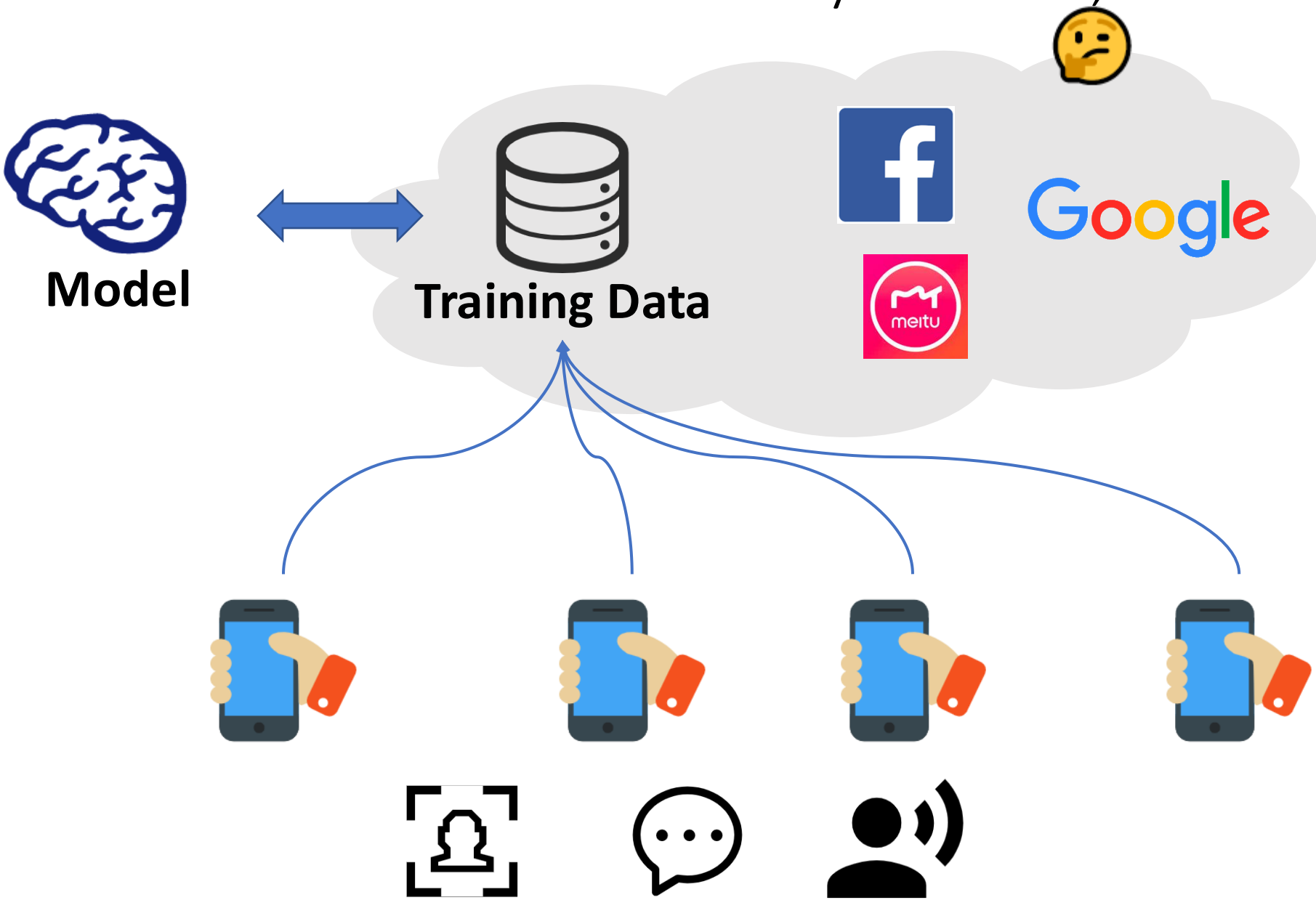
Fuel (Data)

Source: https://wallpaperplay.com/board/rocket-launch-wallpapers

The world's most valuable resource is no longer oil, but data

https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

“We need your data, for a better world.”



“To create personalised Products that are unique and relevant to you, we use your connections, preferences, interests and activities based on the data that we **collect and learn from you and others** …”

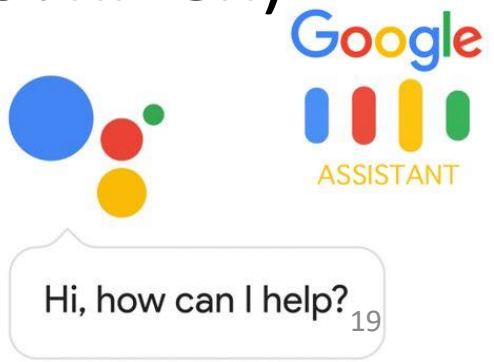
-- Facebook (https://www.facebook.com/policy.php)

“Settings you may be asked to turn on …

- **Voice & audio** recordings, which records your voice and audio on Google services to improve speech recognition.”

-- Google Assistant

(https://support.google.com/assistant/answer/7126196?hl=en)

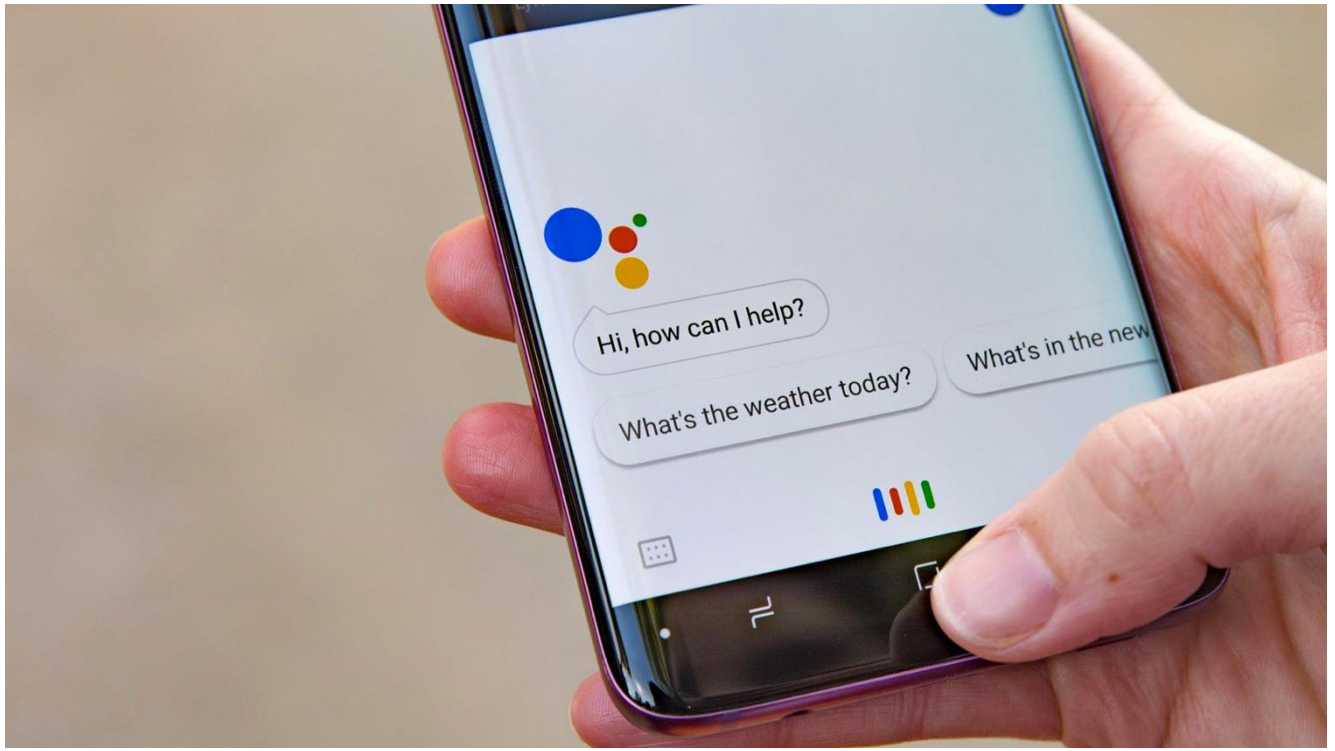


“… that we will **also receive the EXIF data** about your photos (EXIF data may contain GPS coordinates …) for the purpose of further optimizing our services.”

-- Meitu (https://corp.meitu.com/privacy_en.html)

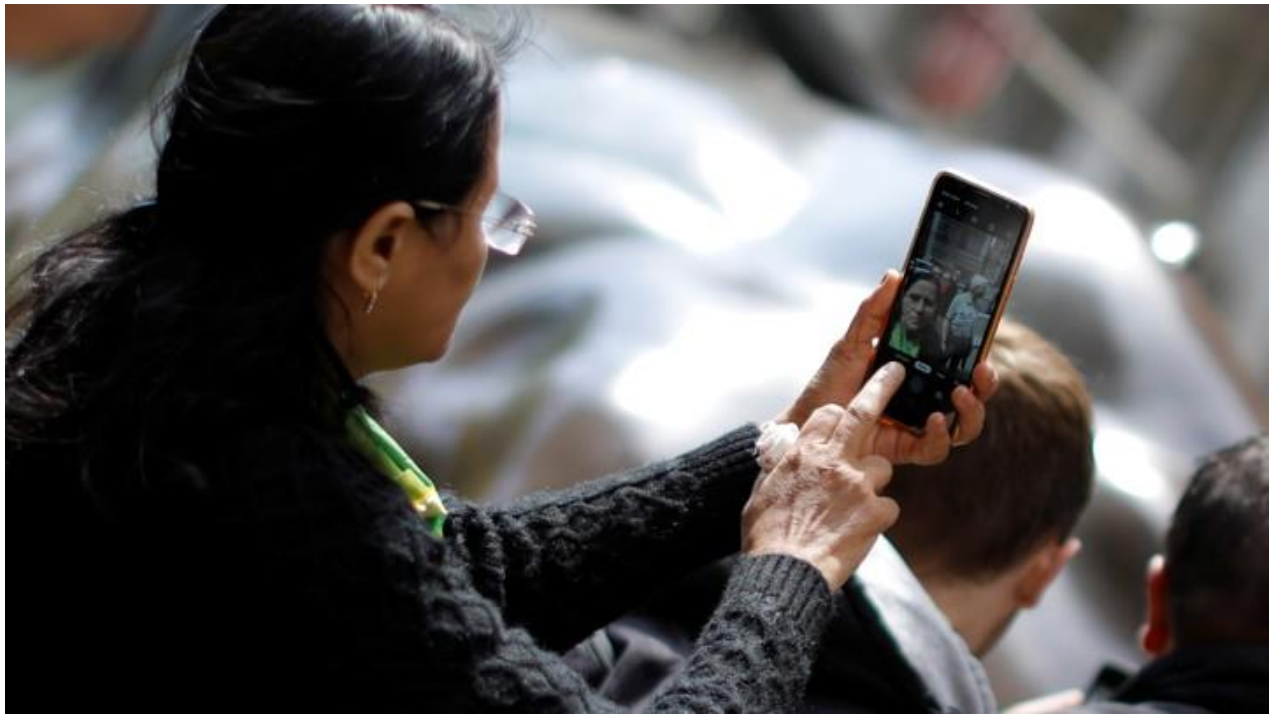
"Facebook's failure to compel *Cambridge Analytica* to delete all traces of data from its servers – including any "**derivatives**" – enabled the company to retain predictive models derived from millions of social media profiles throughout the US presidential election." (*The Guardian*, 6/5/2018).

https://www.theguardian.com/uk-news/2018/may/06/cambridge-analytica-kept-facebook-data-models-through-us-election



"Google admitted on Thursday that more than 1,000 sound recordings of customer conversations with the Google Assistant were leaked by some of its partners to a Belgian news site."
(*CNBC*, 11/7/2019).

https://www.cnbc.com/2019/07/11/google-admits-leaked-private-voice-conversations.html



"Yes, you should stop using FaceApp, because **there are few controls on how your data, including your face data**, will be used. But the problems that FaceApp poses aren't unique. Walking around anywhere can get your face included in facial-recognition databases." - Tiffany C. Li (*The Atlantic*, 20/7/2019).

https://www.theatlantic.com/ideas/archive/2019/07/faceapp-reveals-huge-holes-todays-privacy-laws/594358/

Anonymized data publishing?

Re-identification Attack

[Narayanan & Shmatiokv, S&P 2008, S&P 2009]

[1] A. Narayanan and V. Shmatiokv, “Robust De-anonymization of Large Sparse Datasets,” in Proc. of IEEE S&P, 2008.
[2] A. Narayanan and V. Shmatiokv, “De-anonymizing Social Networks,” in Proc. of IEEE S&P, 2009.


WELCOME TO
REALITY

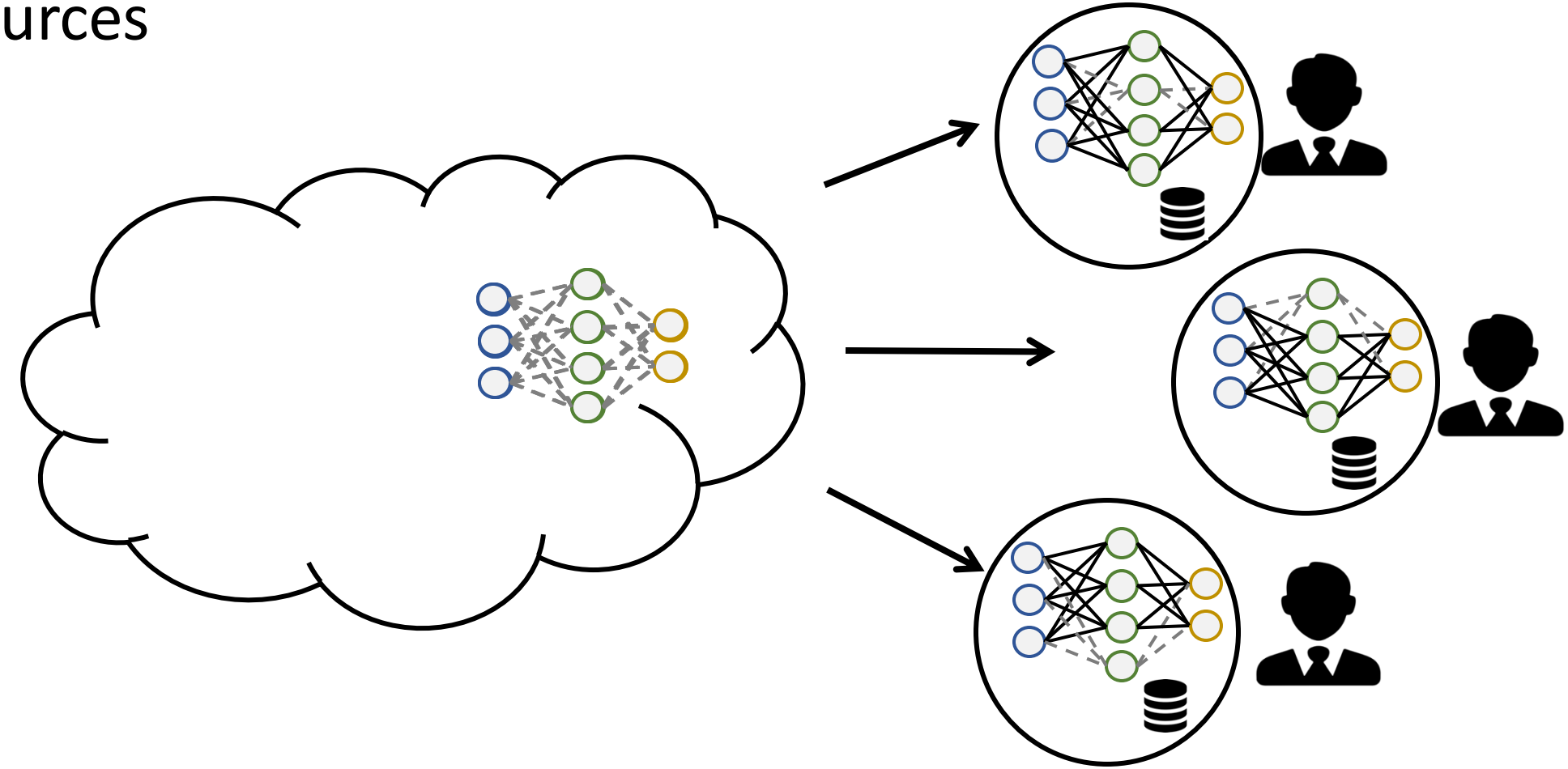
Collecting high-quality datasets is challenged by data privacy requirements

https://www.eugdpr.org/

Source: https://www.techspot.com/news/74426-firms-scramble-get-their-data-collection-policies-compliance.html

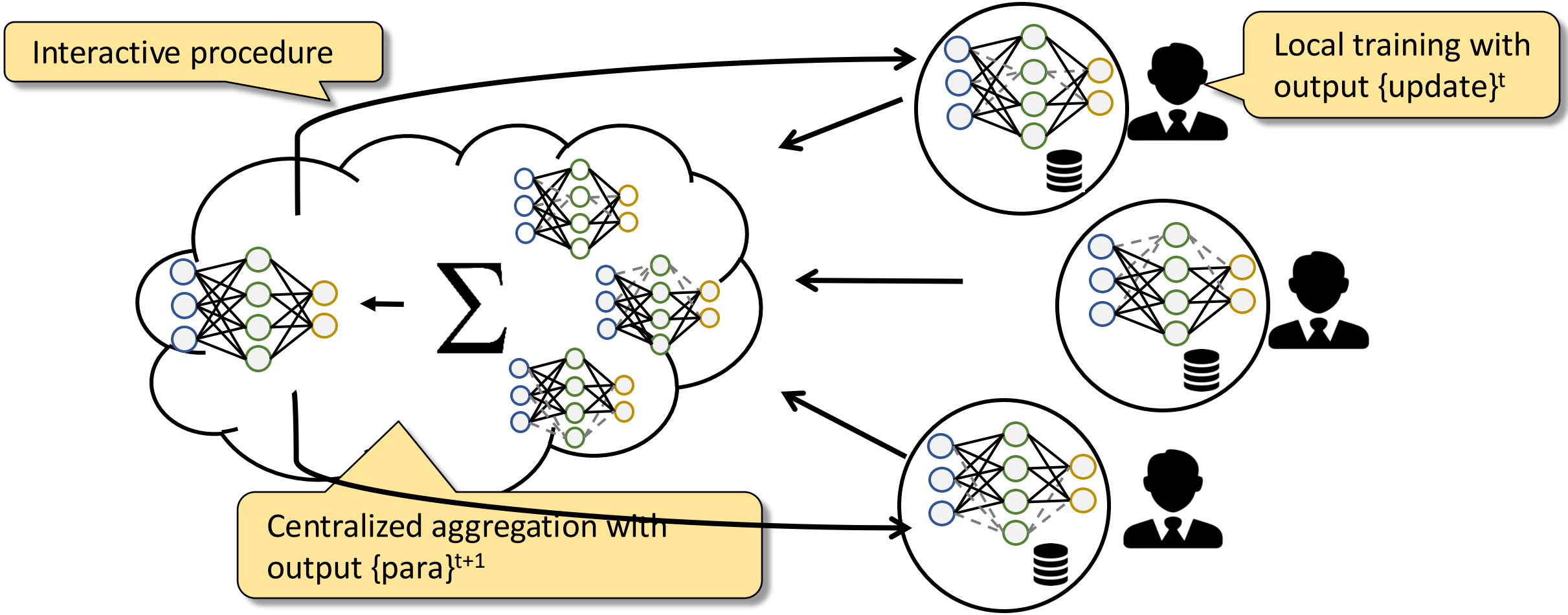
Federated learning (FL) in a nutshell

- Pushing the training tasks locally at participating data sources

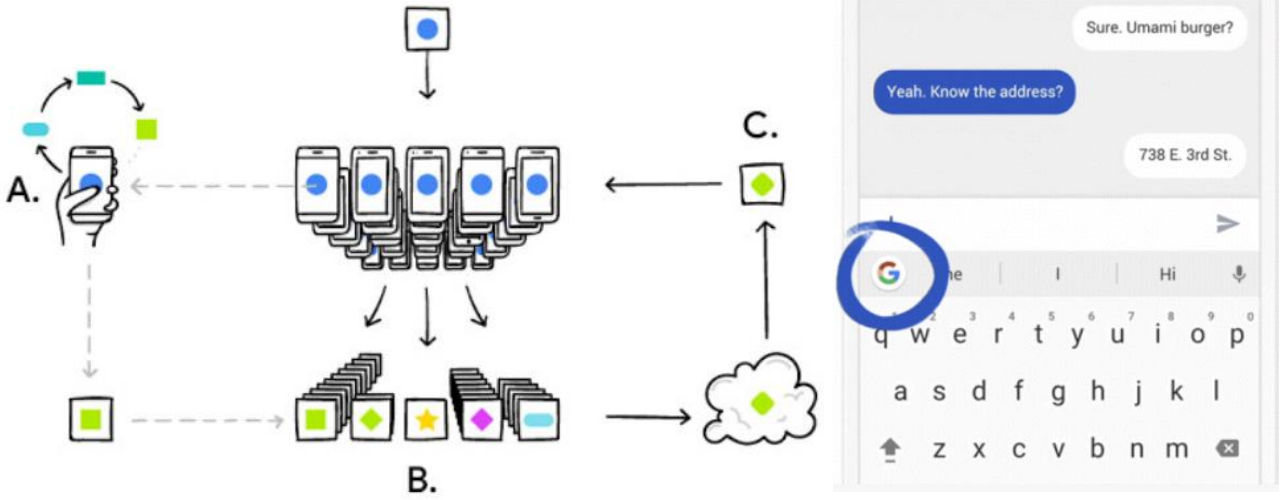


Federated learning (FL) in a nutshell

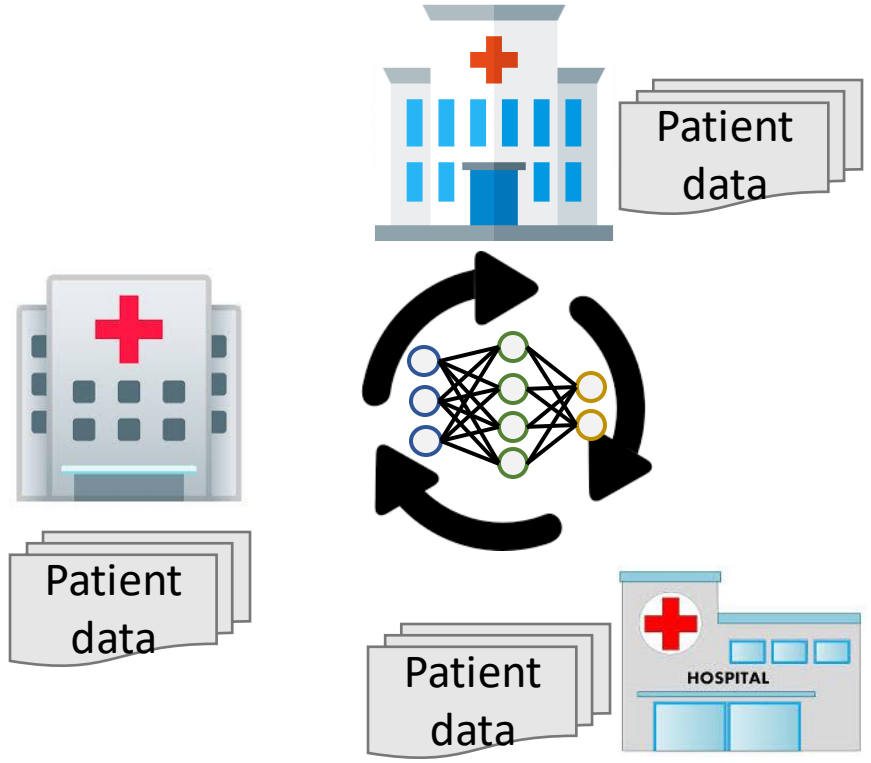
- Aggregating local updates as the global model



Federated learning (FL) Applications

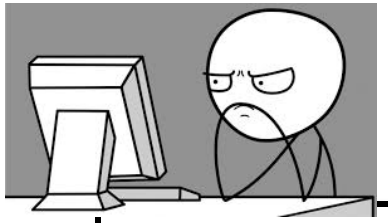


Google Gboard



Hospital disease modeling

Still many challenging issues

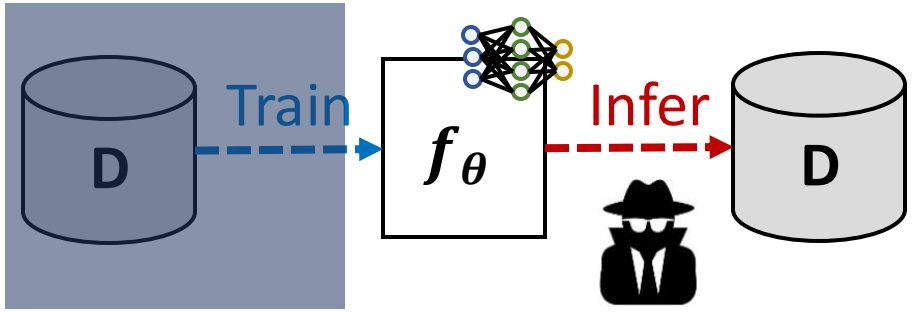


Efficiency and correctness

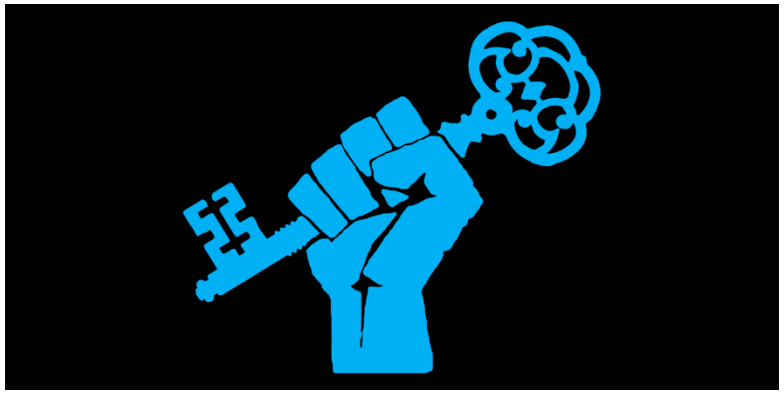
- High communication overhead
- Dishonest participants [1,2]
- Lower convergence speed
- Training data bias

Possible data leakage

Participant's training data might be exposed by model updates[3]

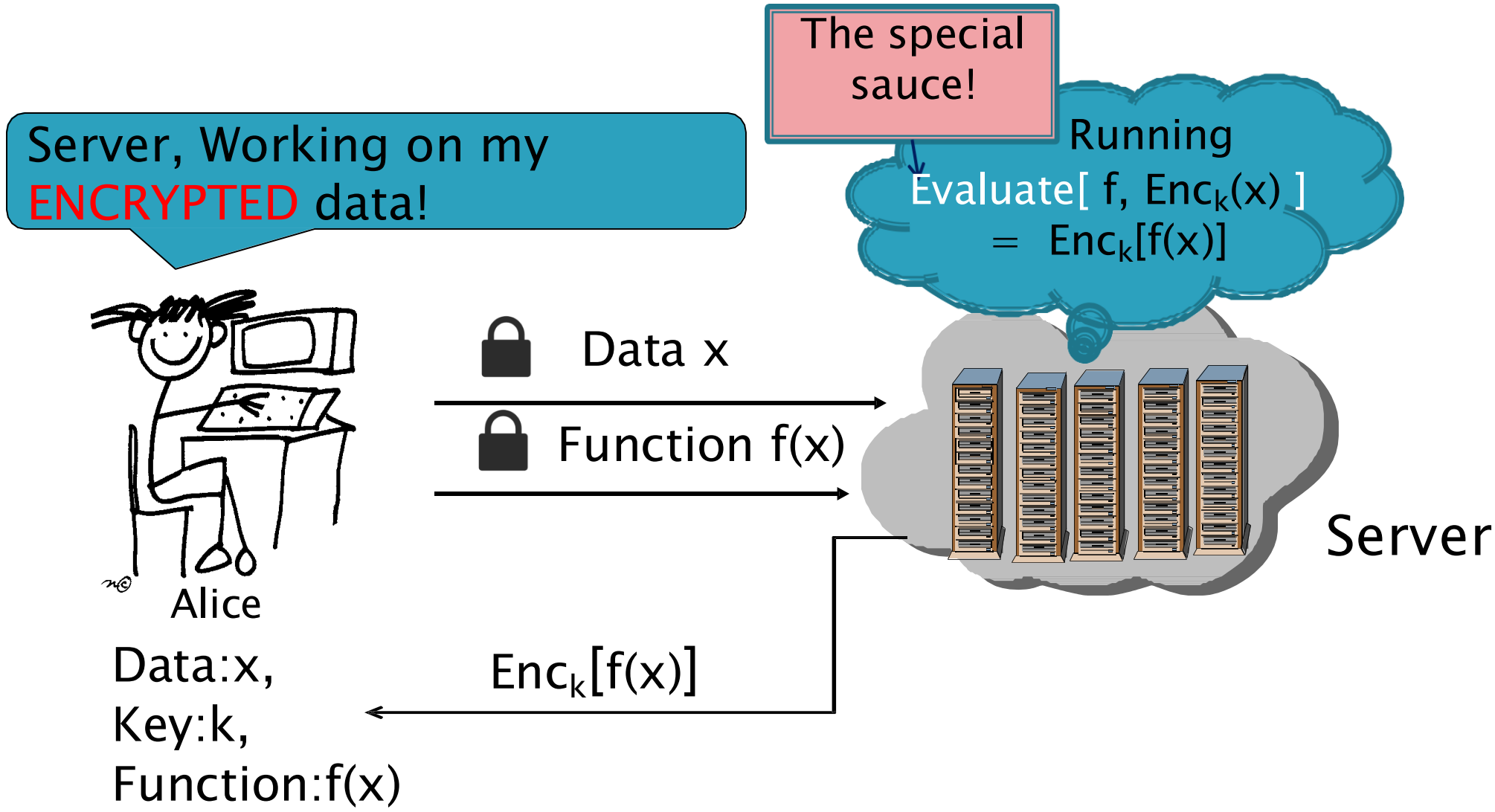


[1] S. Shen, et al. "AUROR: Defending Against Poisoning Attacks in Collaborative Deep Learning Systems," in Proc. of ACM ACSAC, 2016.
[2] E. Bagdasaryan, et al. "How To Backdoor Federated Learning," arXiv preprint arXiv:1807.00459, 2018.
[3] C. Song, et al. "Machine learning models that remember too much" in Proc. of ACM CCS, 2017.



Encrypted Machine Learning

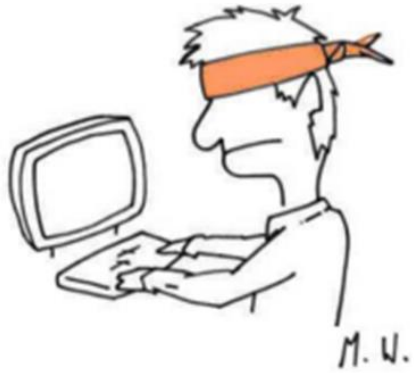
Homomorphic Encryption (HE)



Homomorphic Encryption (HE)

Problem

- Very slow efficiency[1]
- Large ciphertexts and keys[1]



Training a logistic-regression model on genome data[2]

- Under 10 minutes with 10-15 features, ~1000 rows (iDASH 2017)
- 15-30 minutes to train 30,000 models w/ 5 features (iDASH 2018)

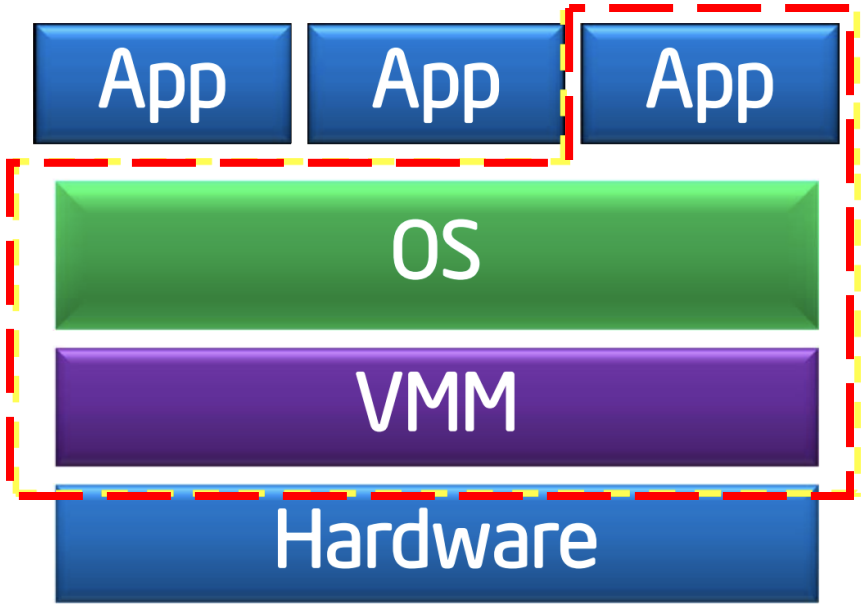
A long way to go…

[1] E. Boyle et al., "Homomorphic Secret Sharing: Optimizations and Applications" in Proc. of CCS, 2017
[2] S. Halevi, "Implementing Advanced Cryptographic Tools", https://shaih.github.io/pubs/Advanced-Cryptorgaphy.pptx

Hardware-Assisted !

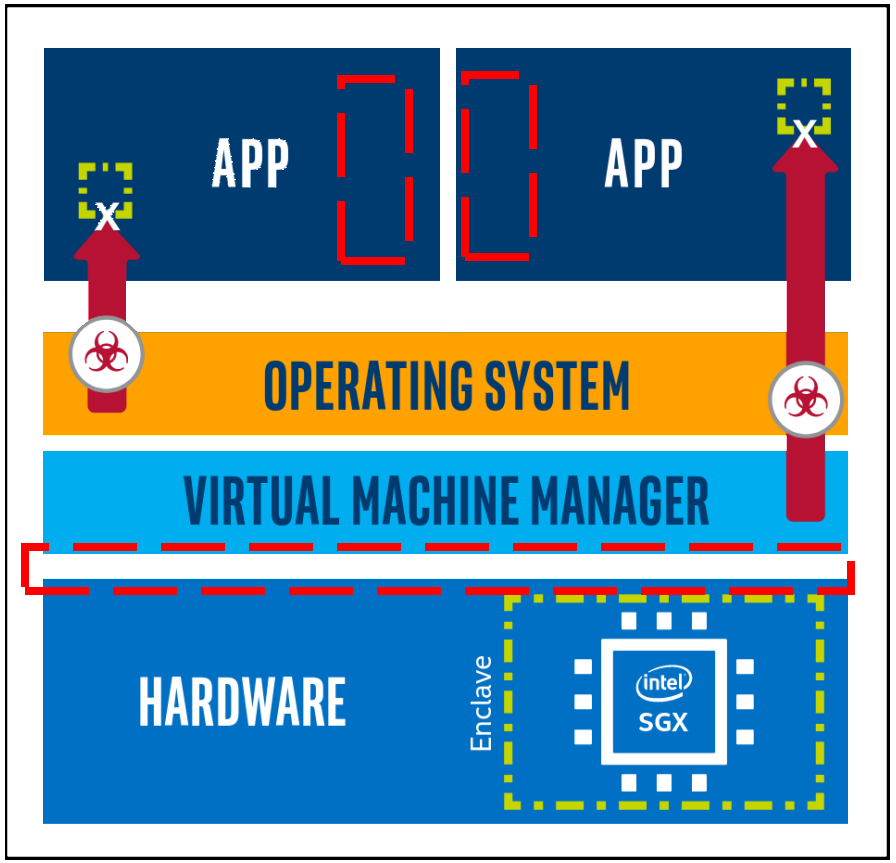
Hardware-Assisted Secure Enclave

Attack Surface Nowadays



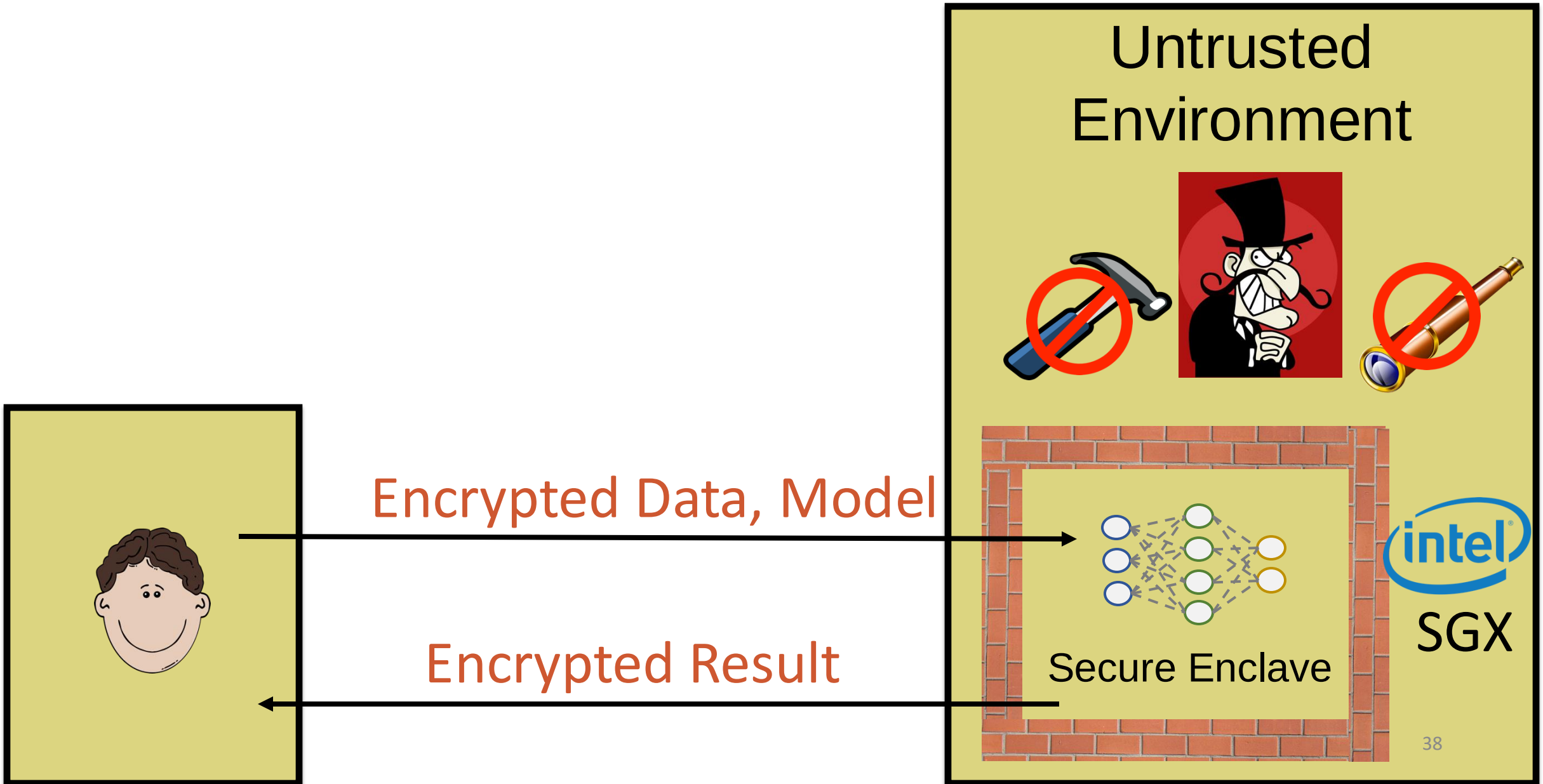
Attack Surface

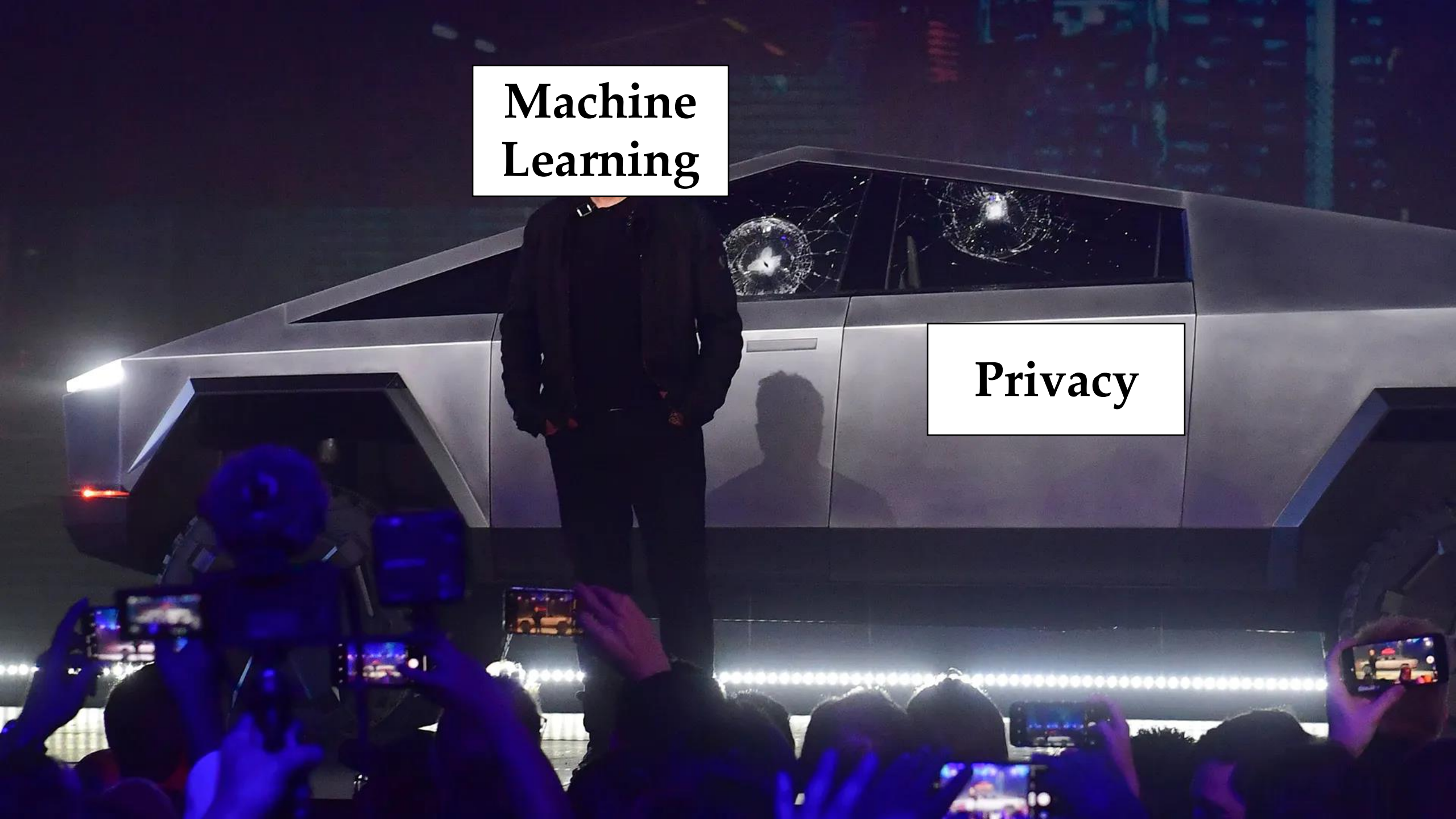
Attack Surface With Enclaves



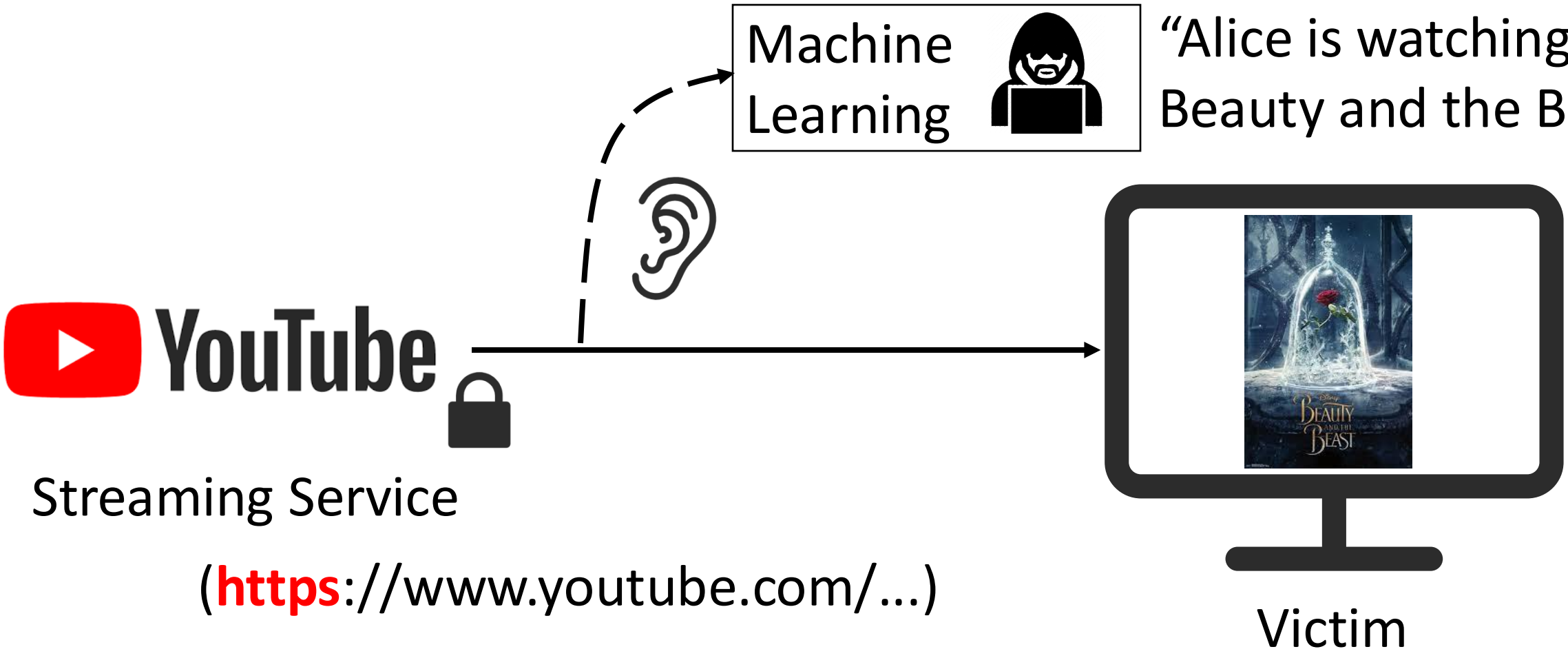
Source: https://software.intel.com/en-us/sgx

Hardware-Assisted Encryption




Machine Learning
Privacy

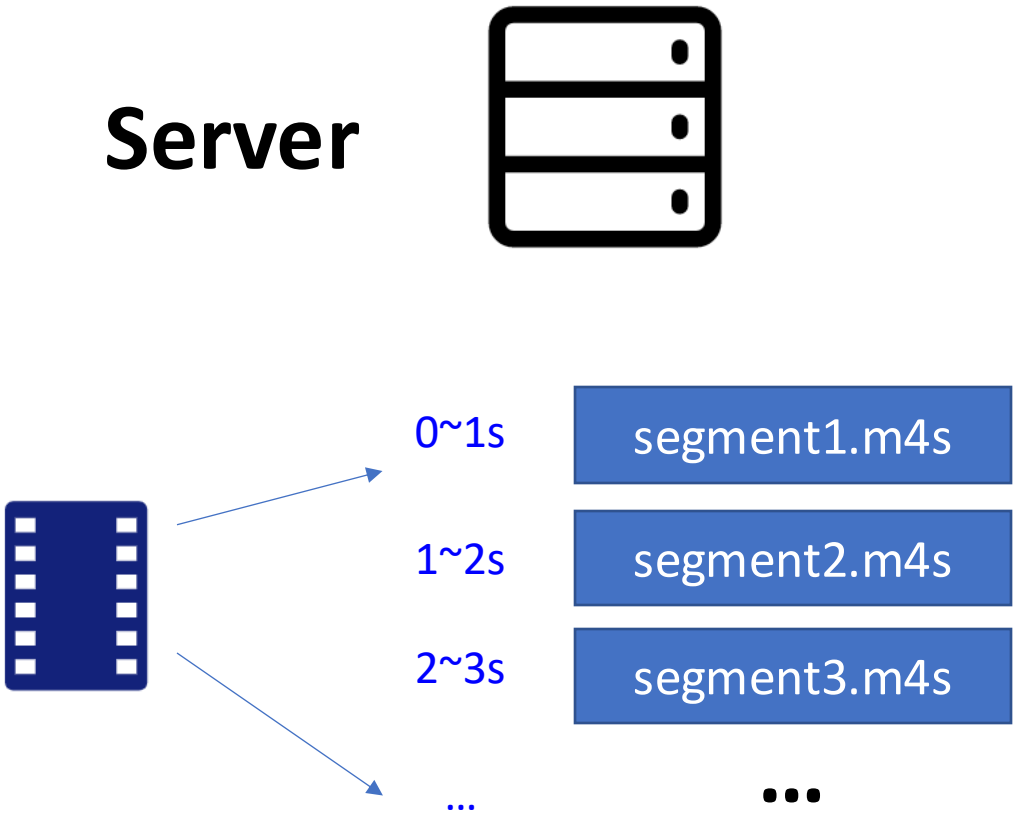
Remote Identification of Encrypted Video Streams

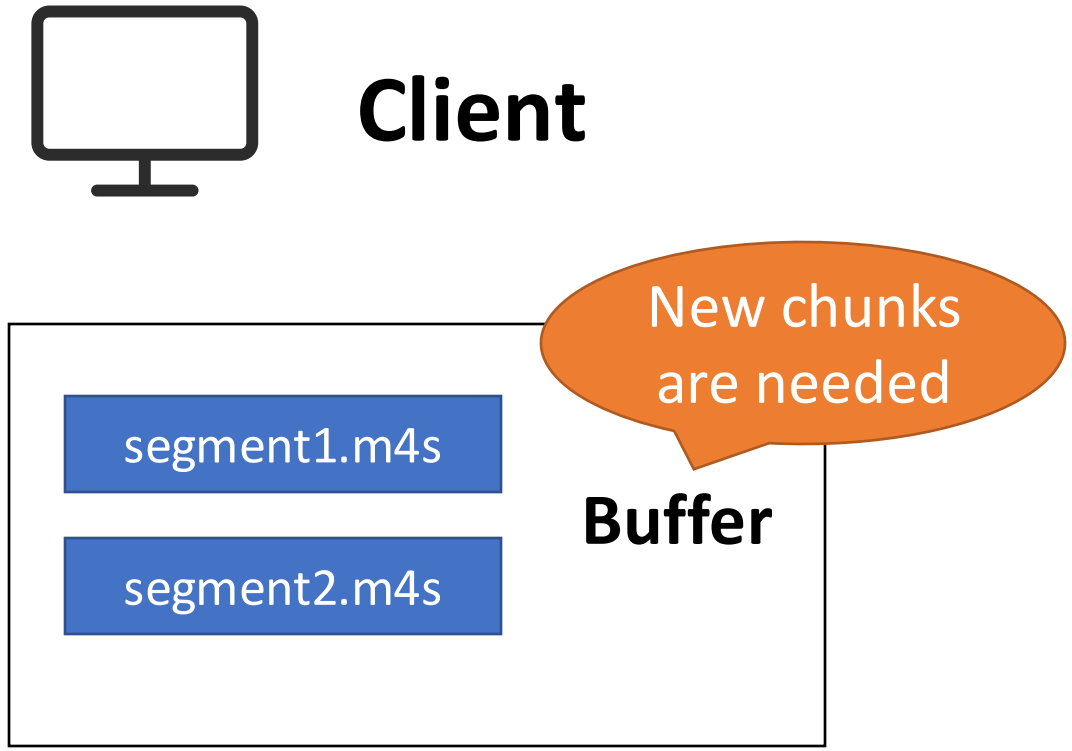


R. Schuster et al., "Beauty and the Burst: Remote Identification of Encrypted Video Streams," in Proc. of USENIX Security, 2017

Remote Identification of Encrypted Video Streams

MPEG-DASH (**D**ynamic **A**daptive **S**treaming over **H**TTP) Standard

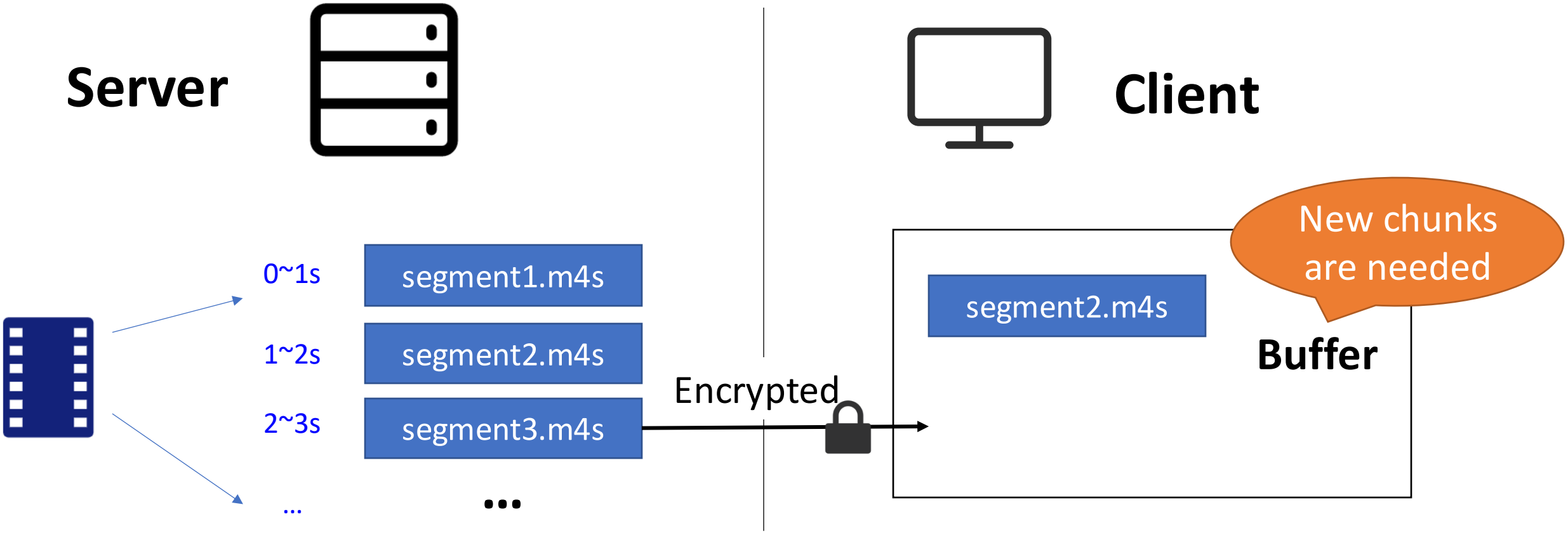




R. Schuster et al., "Beauty and the Burst: Remote Identification of Encrypted Video Streams," in Proc. of USENIX Security, 2017

Remote Identification of Encrypted Video Streams

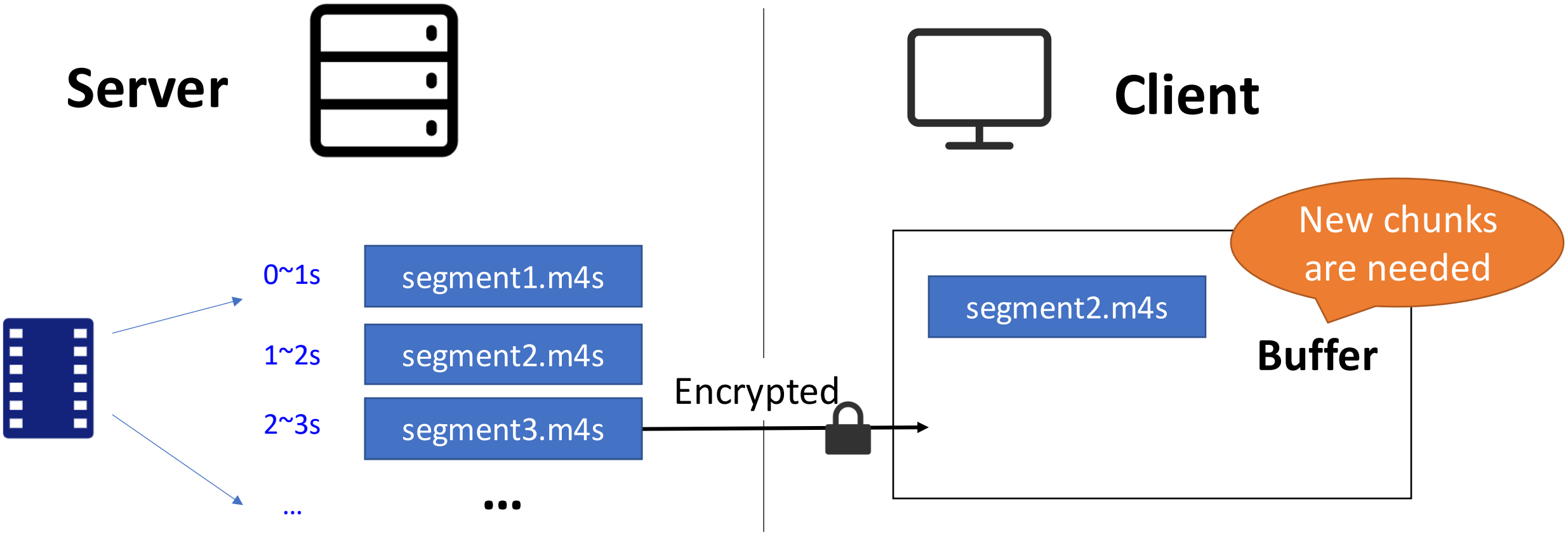
MPEG-DASH (**D**ynamic **A**daptive **S**treaming over **H**TTP) Standard



R. Schuster et al., "Beauty and the Burst: Remote Identification of Encrypted Video Streams," in Proc. of USENIX Security, 2017

Remote Identification of Encrypted Video Streams

MPEG-DASH (**D**ynamic **A**daptive **S**treaming over **H**TTP) Standard

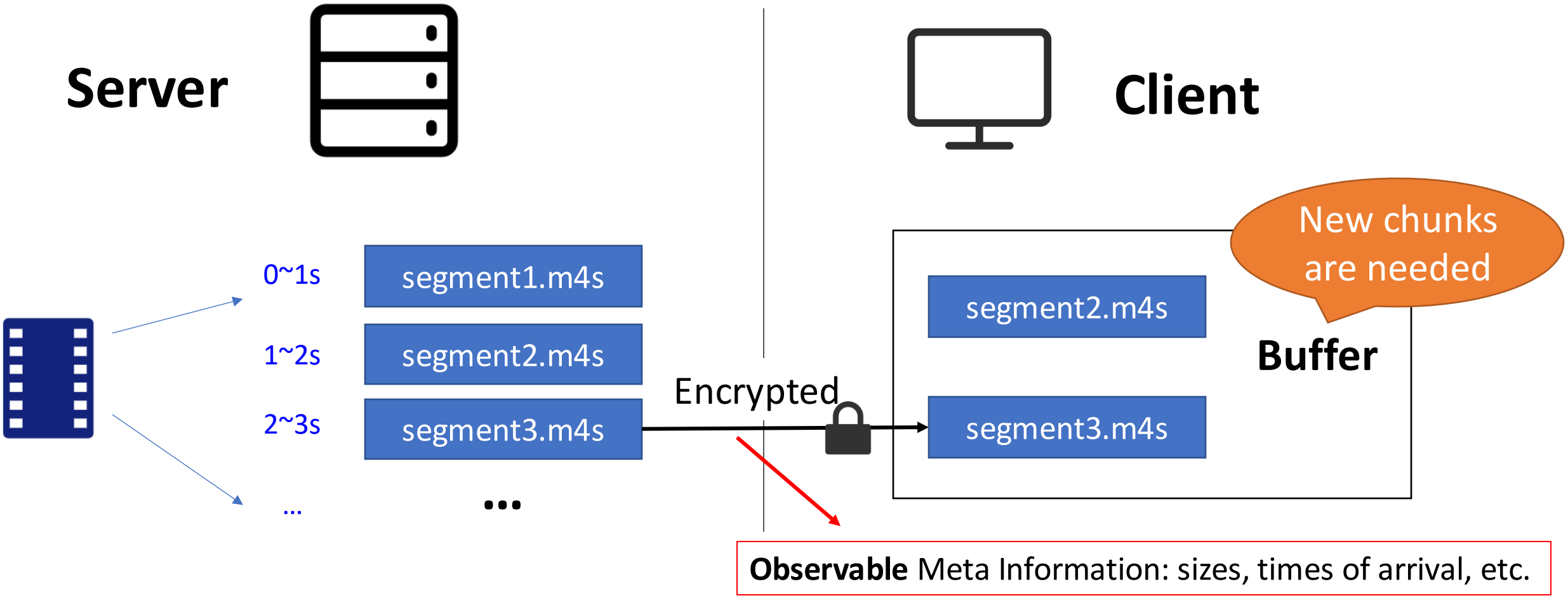


R. Schuster et al., "Beauty and the Burst: Remote Identification of Encrypted Video Streams," in Proc. of USENIX Security, 2017

Remote Identification of Encrypted Video Streams

1. Observable meta information

MPEG-DASH (**D**ynamic **A**daptive **S**treaming over **H**TTP) Standard

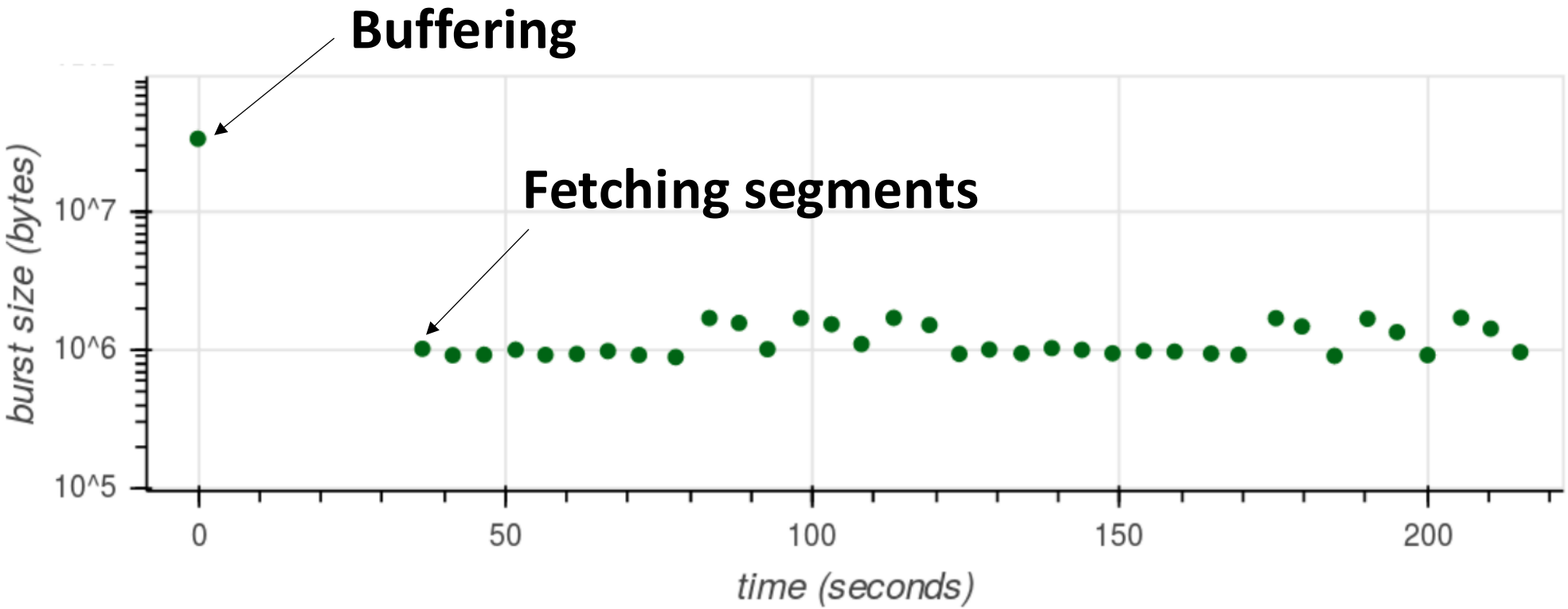


R. Schuster et al., "Beauty and the Burst: Remote Identification of Encrypted Video Streams," in Proc. of USENIX Security, 2017

Remote Identification of Encrypted Video Streams

1. Observable meta information

MPEG-DASH (**D**ynamic **A**daptive **S**treaming over **H**TTP) Standard



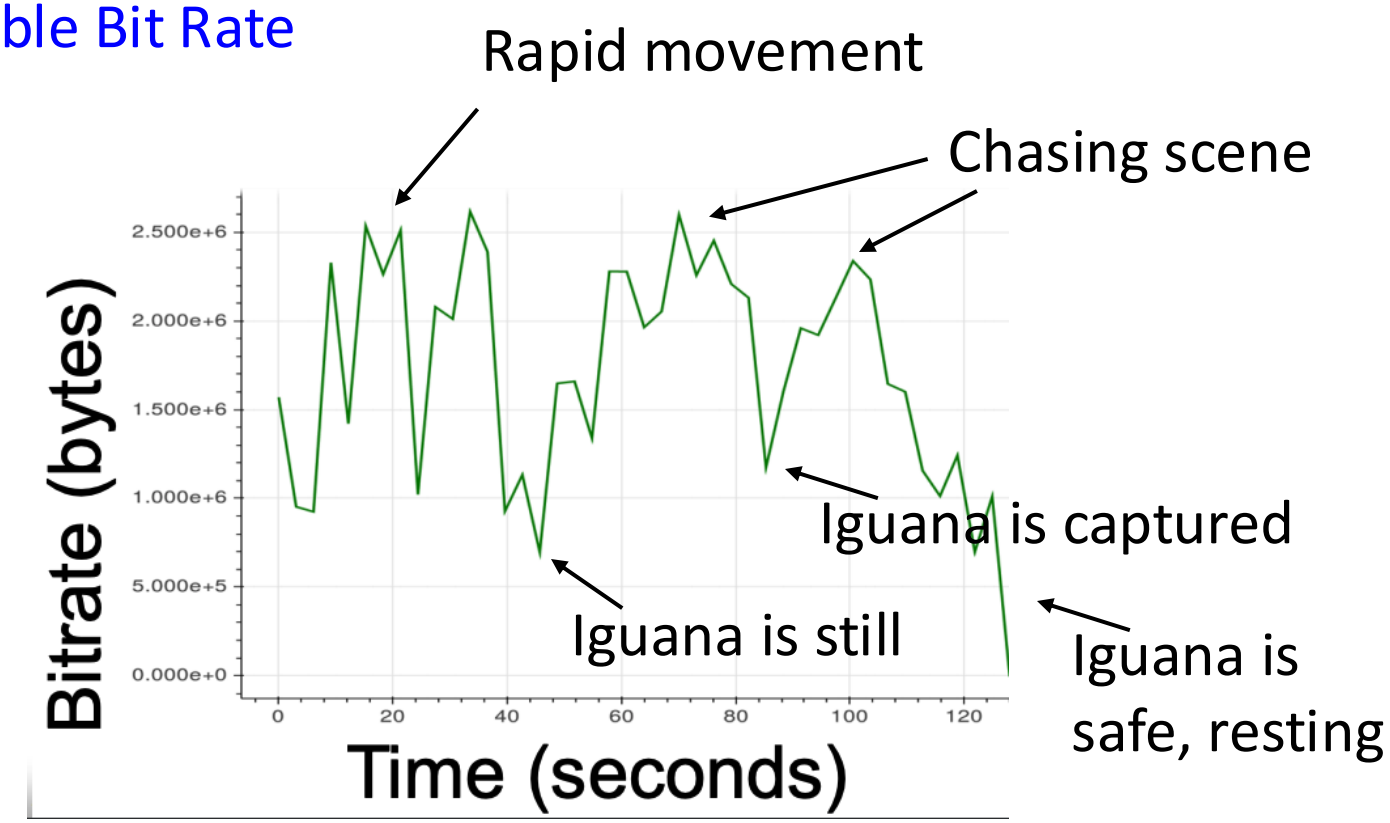
R. Schuster et al., “Beauty and the Burst: Remote Identification of Encrypted Video Streams,” in Proc. of USENIX Security, 2017

Remote Identification of Encrypted Video Streams

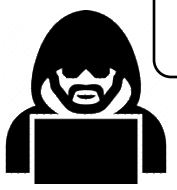
1. Observable meta information

2. Bit rate is varied by the content

Variable Bit Rate



"Iguana vs snake"



Bit rate is varied by the video content.

R. Schuster et al., "Beauty and the Burst: Remote Identification of Encrypted Video Streams," in Proc. of USENIX Security, 2017

Remote Identification of Encrypted Video Streams

1. Observable meta information

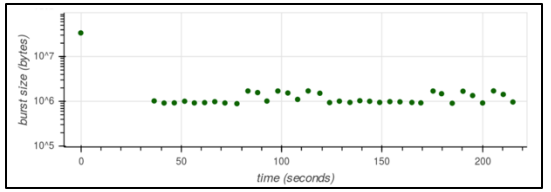
2. Bit rate is varied by the content

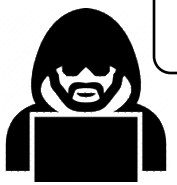
Variable Bit Rate

Bit rate varies with the video content

Chunk size = Time Length (*const*) * Bit Rate

Segment size (burst rate) varies with the video content

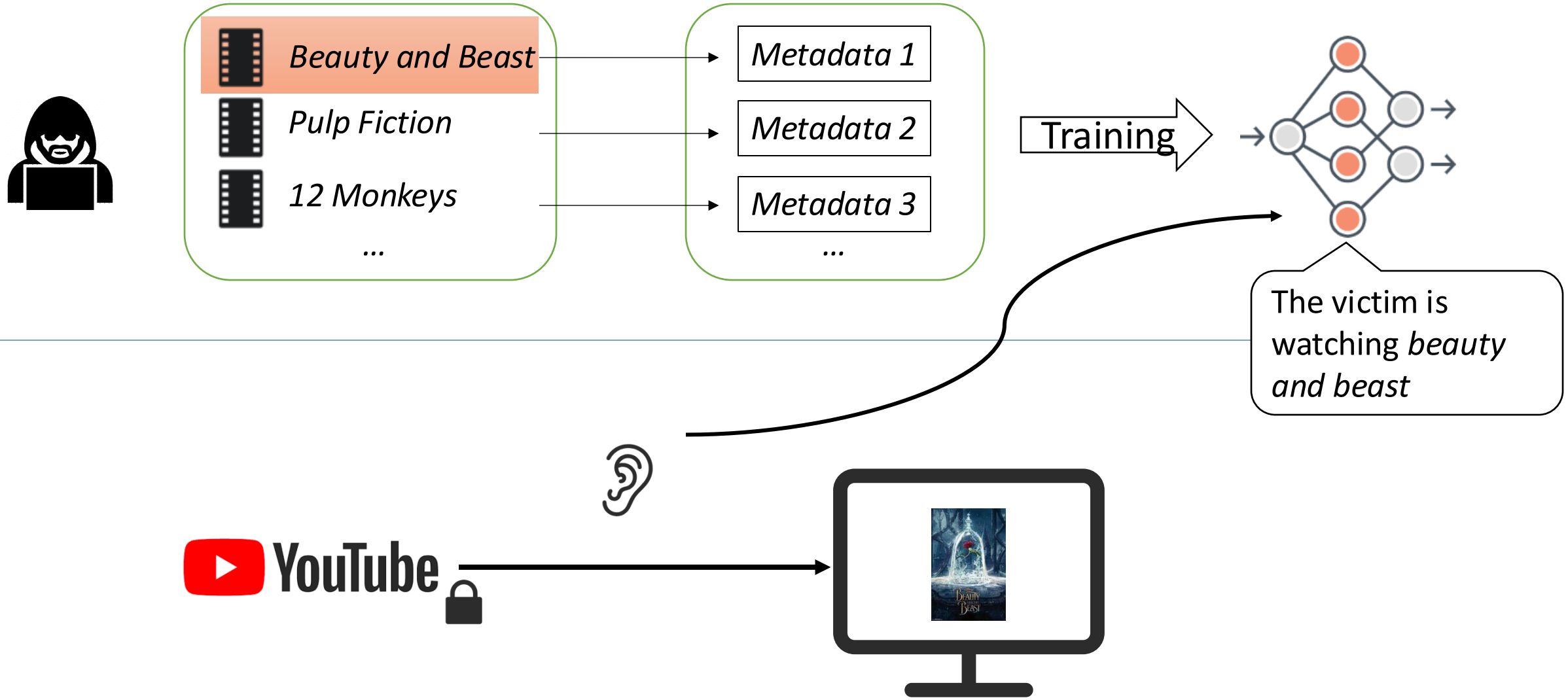




Learn patterns from the variable burst rate.

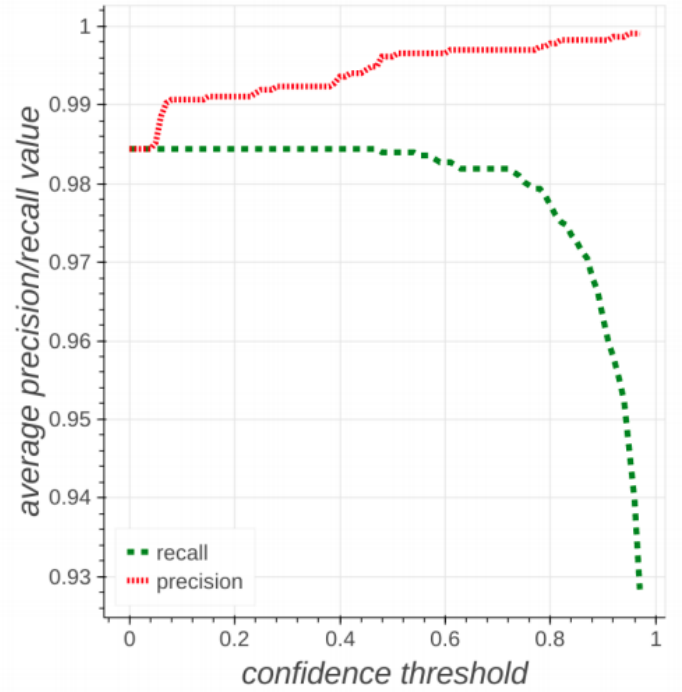
R. Schuster et al., “Beauty and the Burst: Remote Identification of Encrypted Video Streams,” in Proc. of USENIX Security, 2017

Remote Identification of Encrypted Video Streams

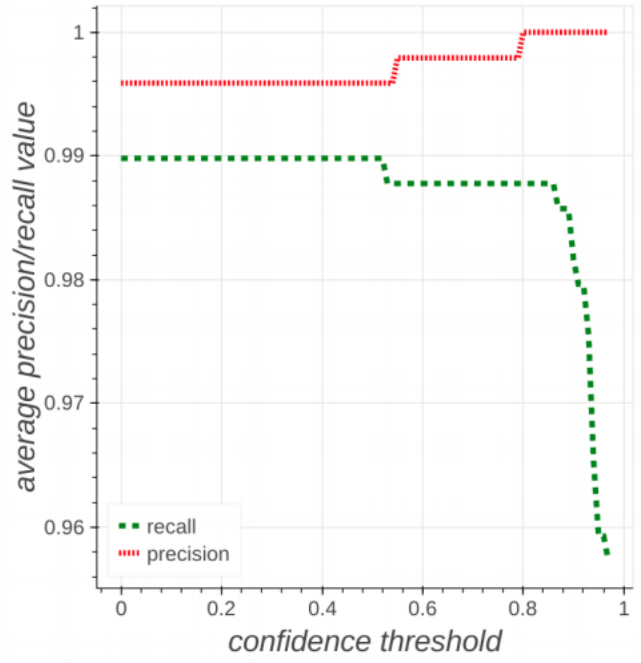


R. Schuster et al., "Beauty and the Burst: Remote Identification of Encrypted Video Streams," in Proc. of USENIX Security, 2017

Remote Identification of Encrypted Video Streams



Experimental results on Netflix



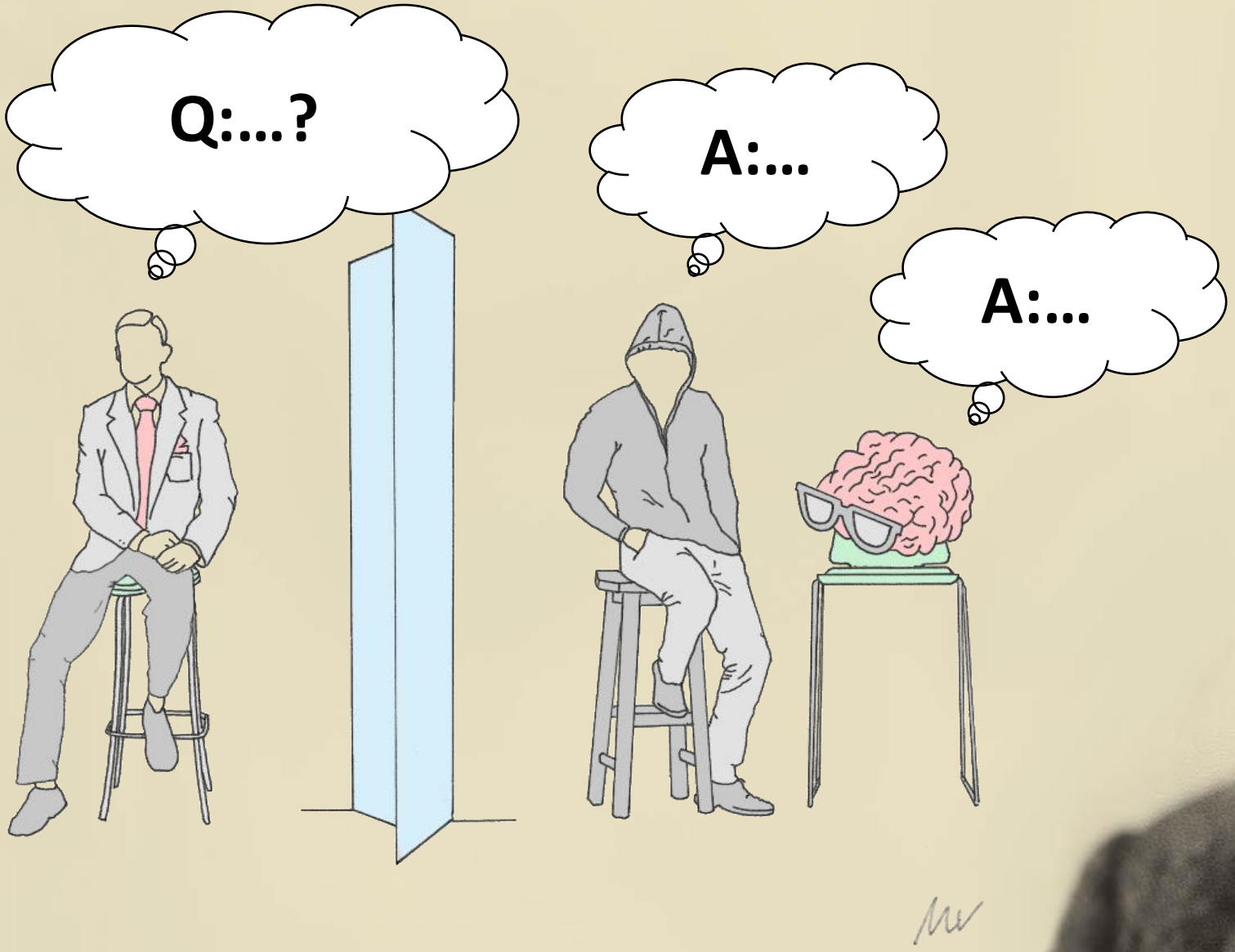
Experimental results on YouTube

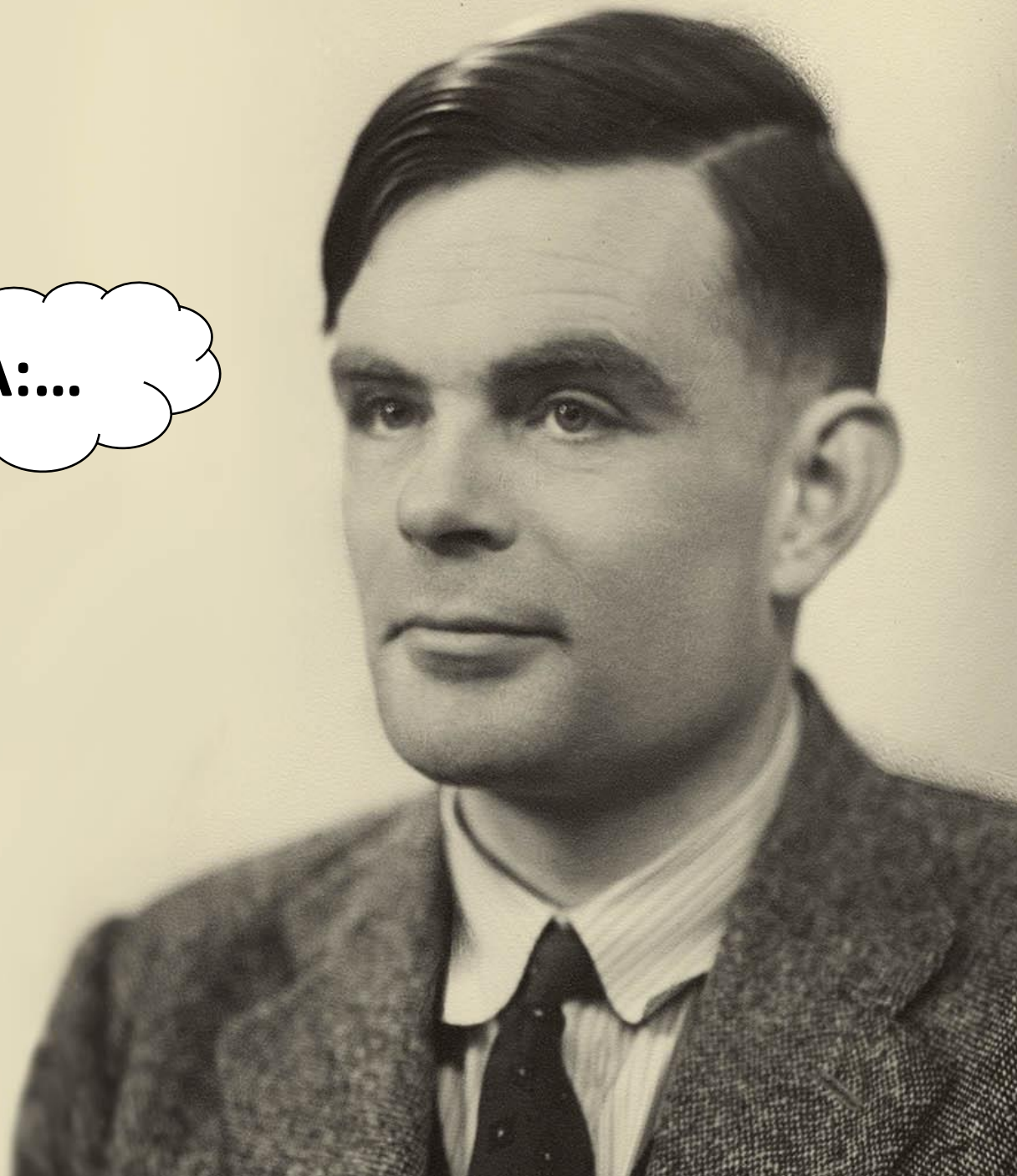
The “ML Rocket”

Engine (Model)

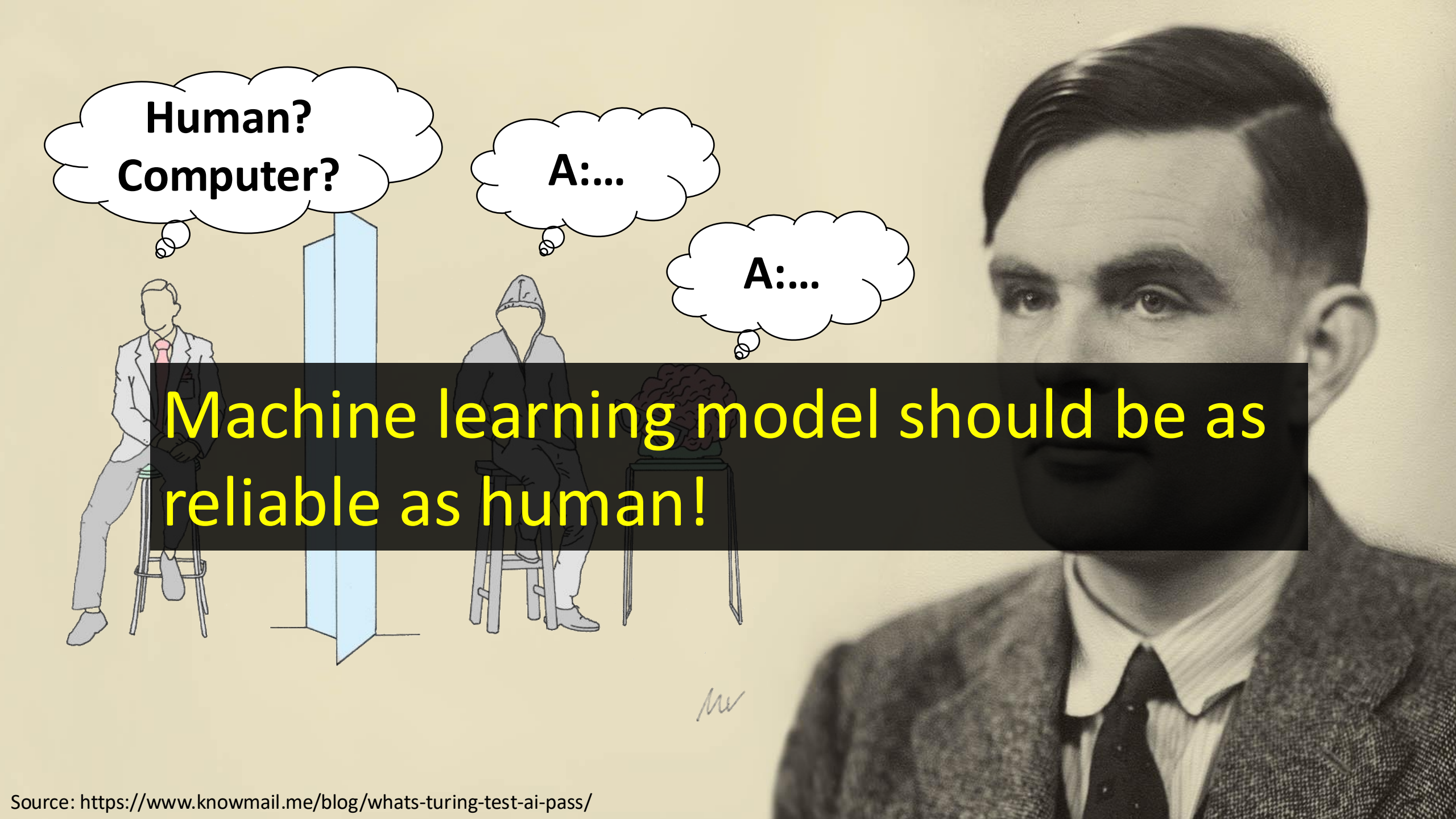
Fuel (Data) → Privacy

Source: https://wallpaperplay.com/board/rocket-launch-wallpapers





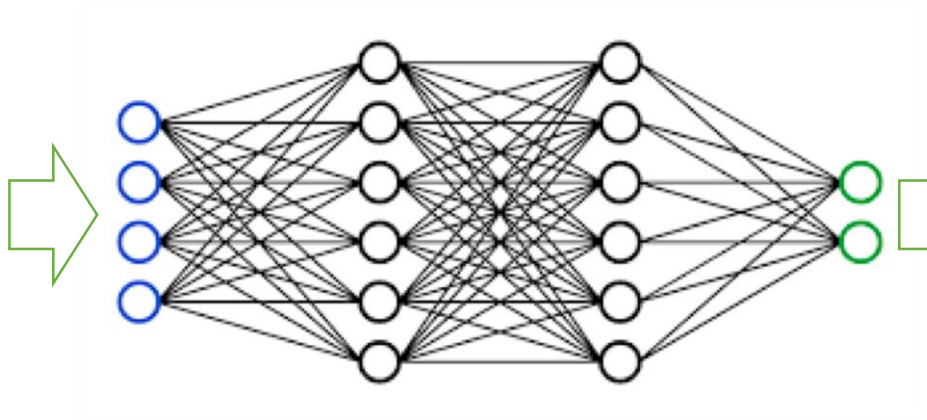
Source: https://www.knowmail.me/blog/whats-turing-test-ai-pass/



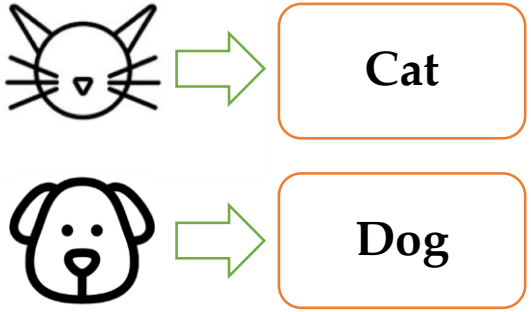
Machine learning model should be as reliable as human!

Source: https://www.knowmail.me/blog/whats-turing-test-ai-pass/

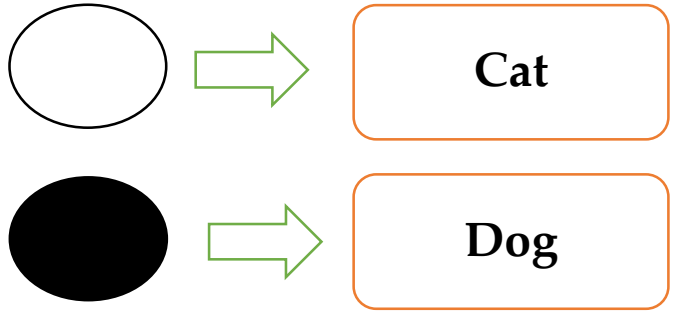
Training data



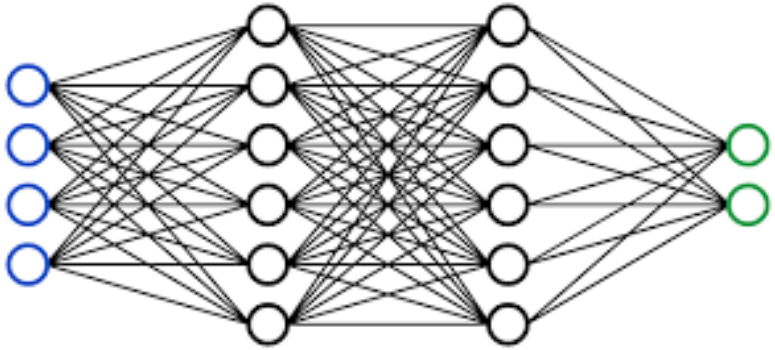
Hopefully,



Possibly?

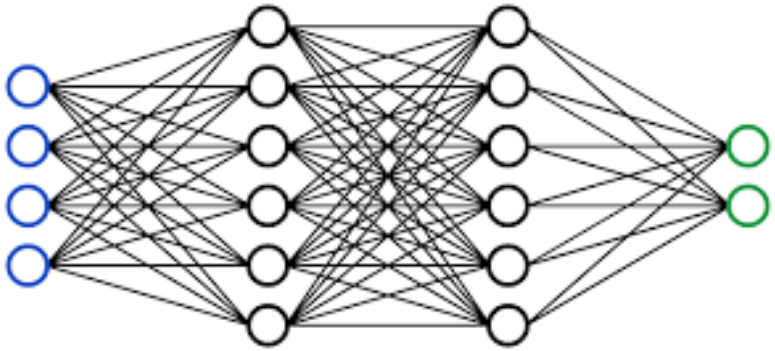


Do neural networks “understand” their tasks? 🤔



Dog !

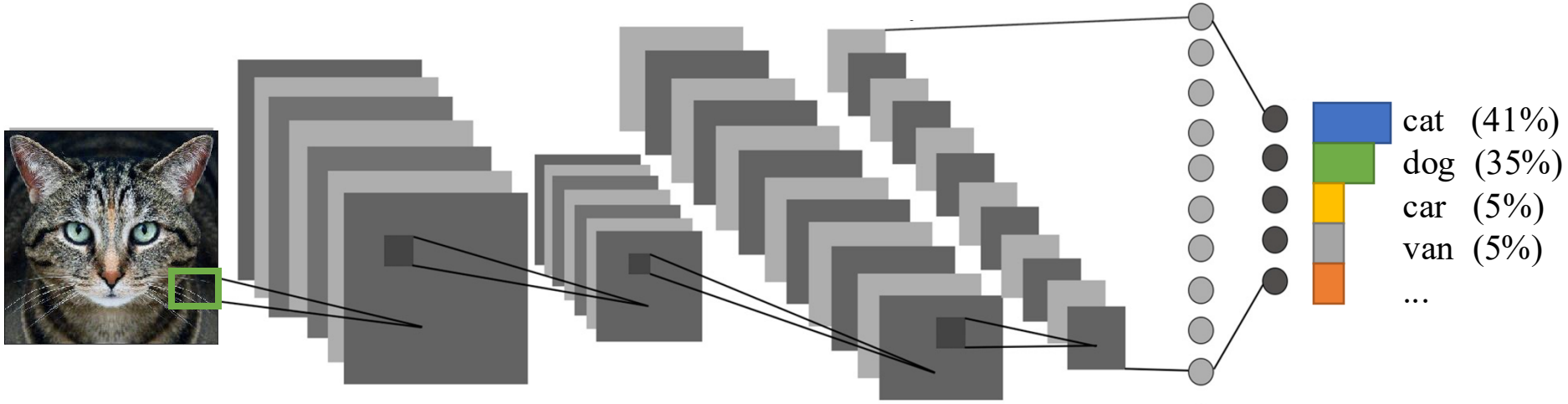
Natural Samples



Dog !

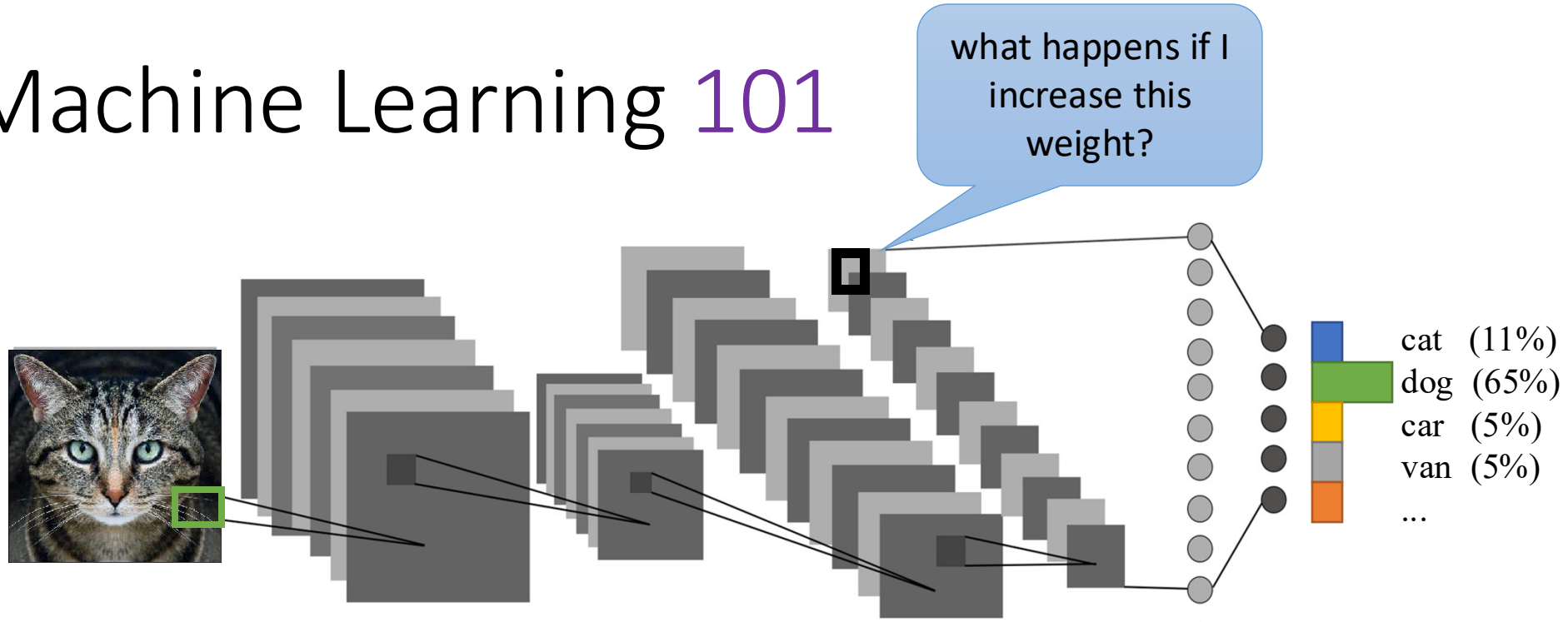
Corner Cases

Machine Learning 101



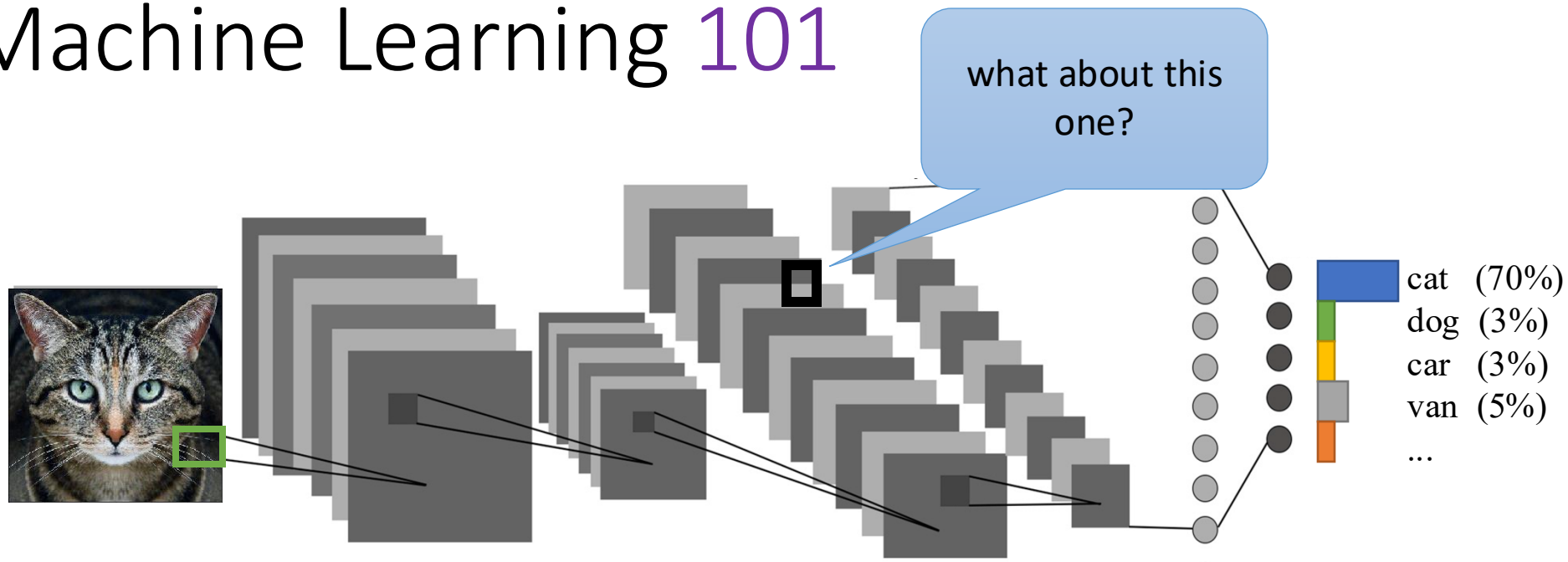
$$\mathcal{L}oss\Big(\ \text{cat (41\%) dog (35\%) car (5\%) van (5\%) ...}\ ,\ \text{“cat”}\ \Big) = 0.31$$

Machine Learning 101



$$\mathcal{L}oss\left(\ \text{cat (11\%) dog (65\%) car (5\%) van (5\%) ...}\ ,\ \text{“cat”}\ \right) = 0.78$$

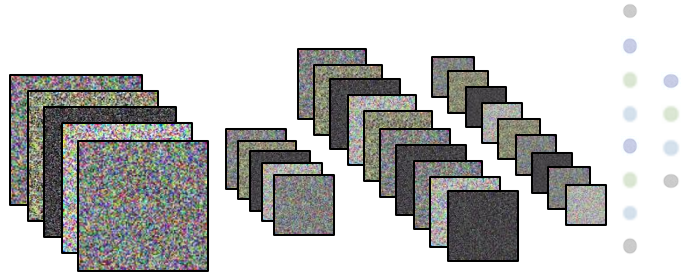
Machine Learning 101



$$\mathcal{L}oss\left(\text{cat (70%), dog (3%), car (3%), van (5%), ...}, \text{“cat”}\right) = 0.16$$

Machine Learning 101

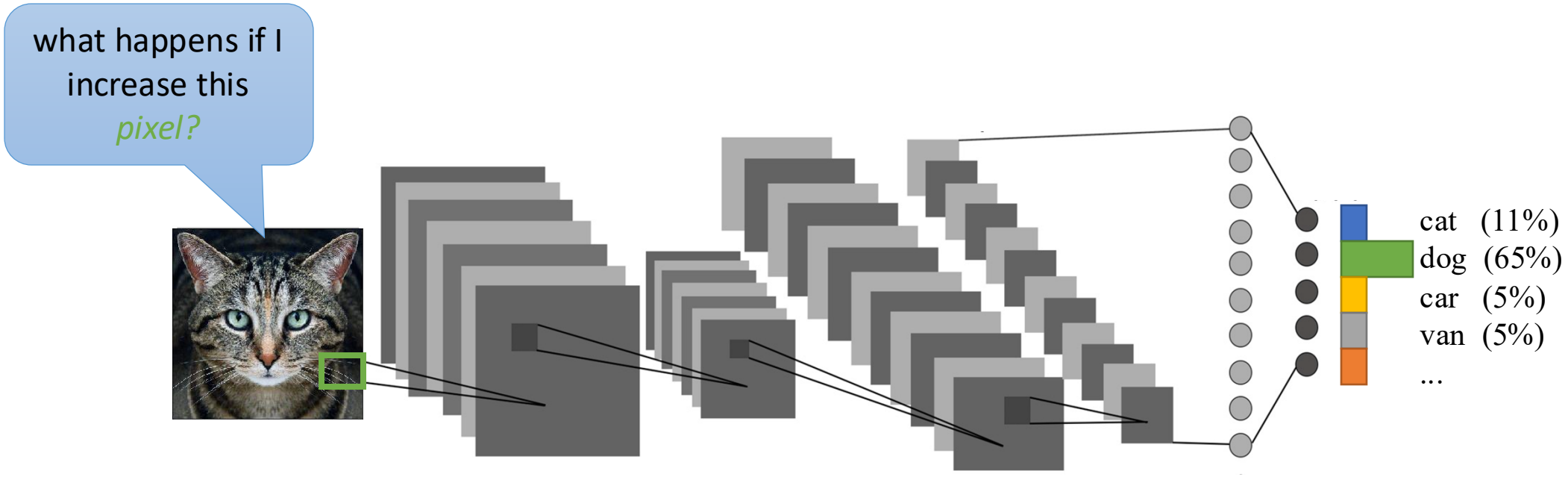
$$\frac{\partial \mathcal{L}oss(\ \text{cat (41\%) dog (35\%) car (5\%) van (5\%) ...}\ ,\ \text{“cat”}\ )}{\partial \theta} =$$



$$\theta' = \theta - \alpha \cdot \frac{\partial \mathcal{L}}{\partial \theta}$$

Take small steps of ***gradient descent*** to ***minimize*** loss

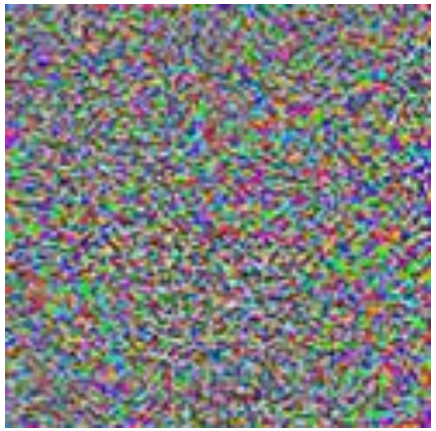
Everything is gradient descent



$$\mathcal{L}oss\Big(\ \text{cat (11\%) dog (65\%) car (5\%) van (5\%) ...}\ ,\ \text{“cat”}\ \Big) = 0.78$$

Everything is gradient descent

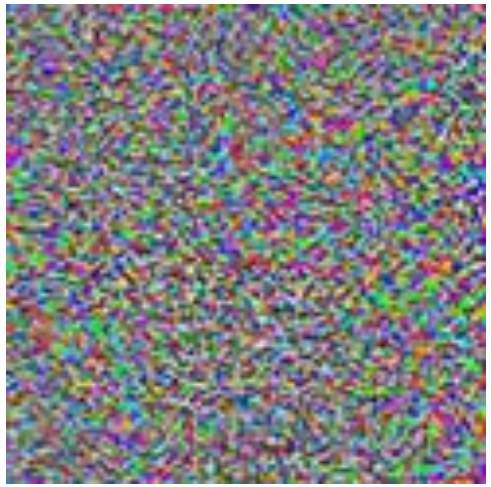
$$\frac{\partial \mathcal{L}oss}{\partial x} =$$



Everything is gradient ~~descent~~ *ascent*

Tabby Cat

$+ \; 0.01 \times$



Adversarial noise

$=$

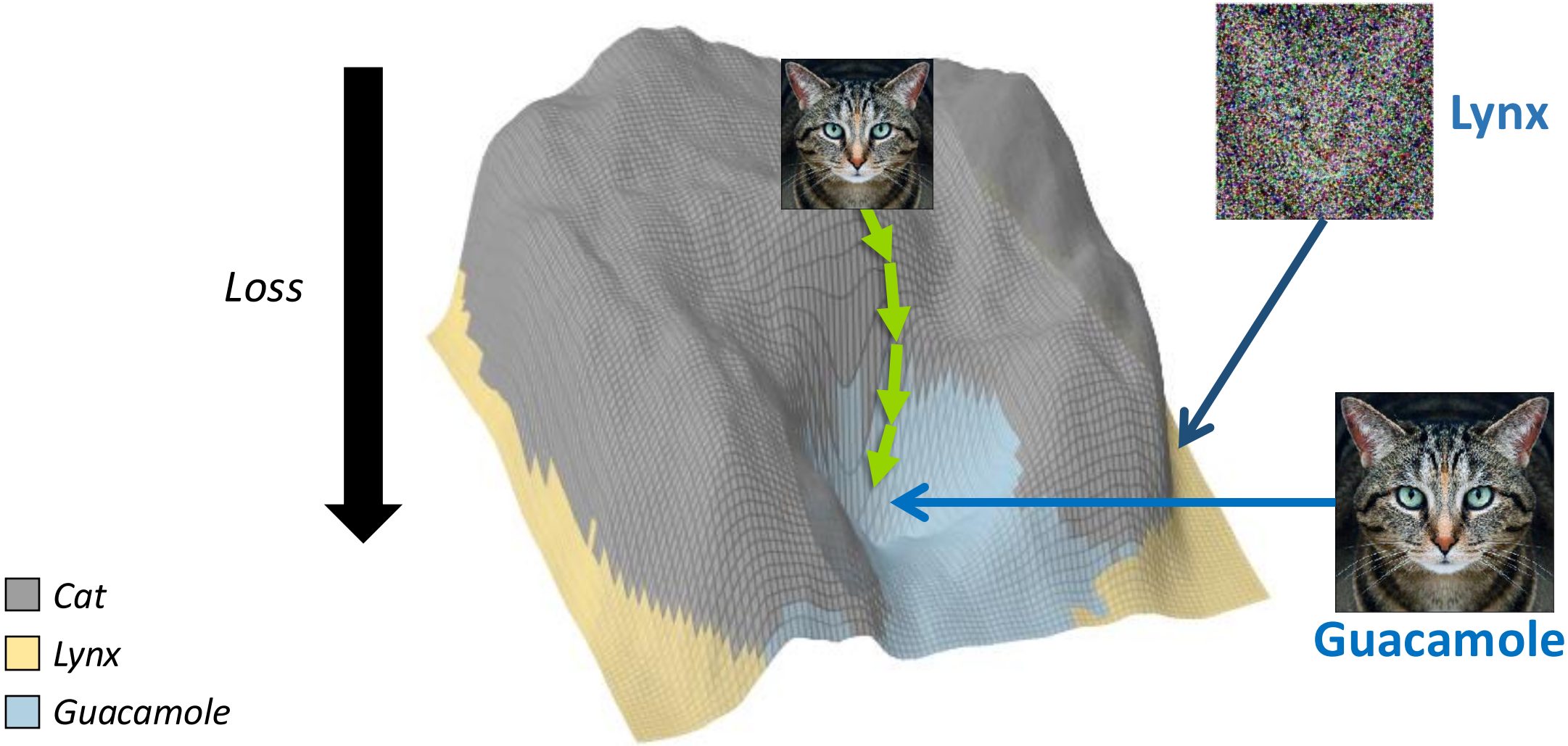
Guacamole

Take small step of ***gradient <u>ascent</u>*** to ***maximize*** loss

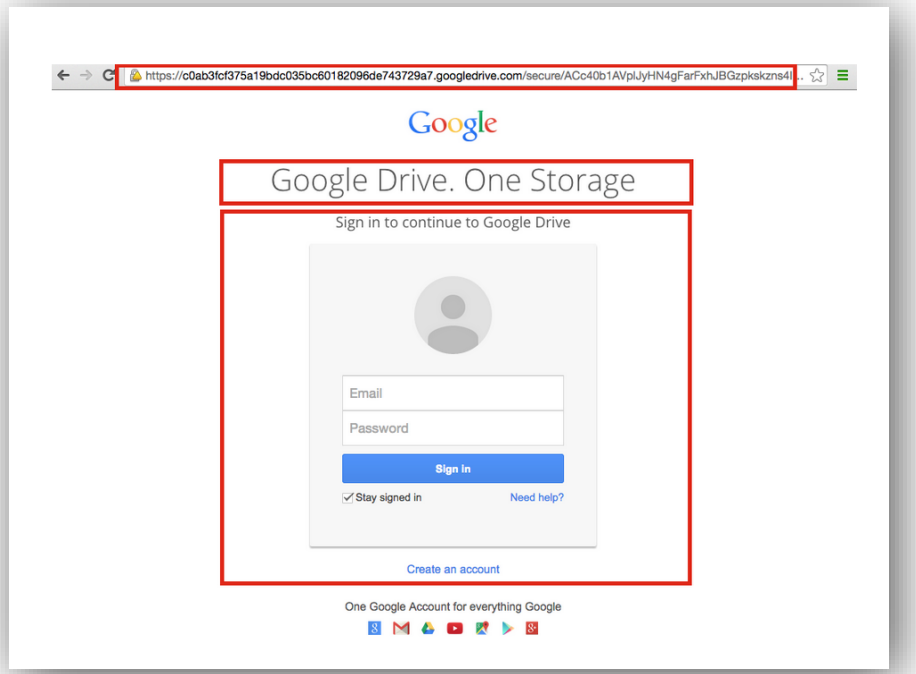
If you prefer code

```
[ ]  model = models.resnet50(pretrained=True)
     x = Image.open("tabby_cat.png")
     label = "Tabby Cat"

     # compute the current loss
     output = model(x)
     loss = nn.CrossEntropyLoss()(output, label)

     # compute the gradient of the loss with respect to the input pixels
     # and take an update step in the direction of the gradient
     grad, = torch.autograd.grad(loss, [x])
     x_adv = x + step_size * grad
```

Repeat for multiple steps



What if you don’t have access to the model weights?



Black-box attacks

"Transfer" attacks

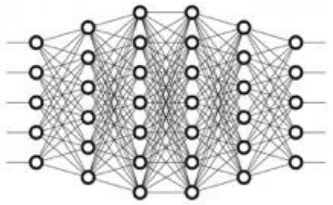
"Boundary" attacks

"Practical Black-Box Attacks against Machine Learning".
Papernot et al. 2016

"Decision-Based Adversarial Attacks".
Brendel et al. 2018

Black-box attacks

“Transfer” attacks



local model

“guacamole”

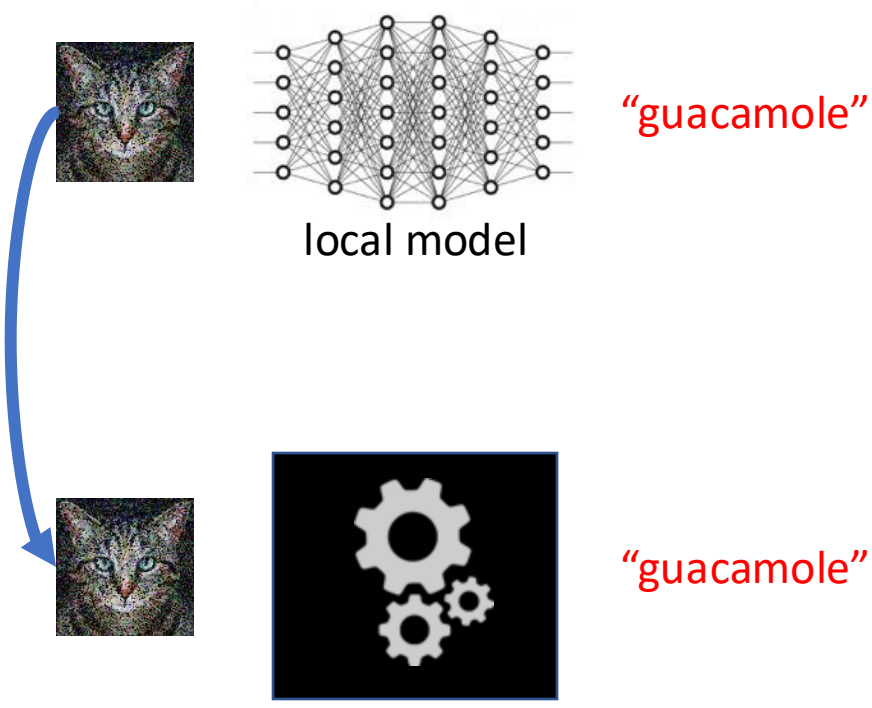
“Practical Black-Box Attacks against Machine Learning”.
Papernot et al. 2016

“Boundary” attacks

“Decision-Based Adversarial Attacks”.
Brendel et al. 2018

Black-box attacks

“Transfer” attacks



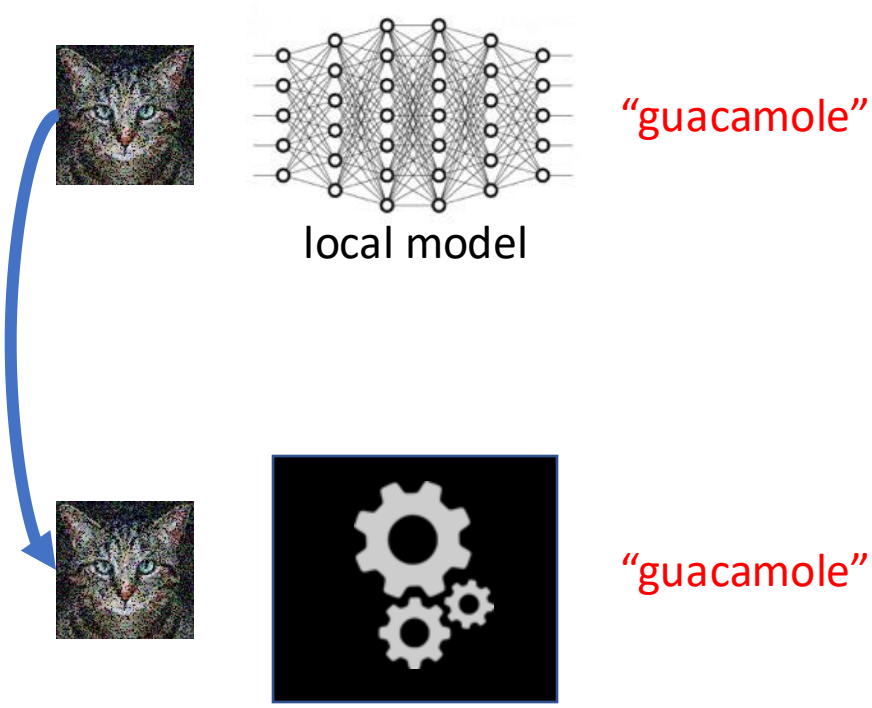
“Practical Black-Box Attacks against Machine Learning”.
Papernot et al. 2016

“Boundary” attacks

“Decision-Based Adversarial Attacks”.
Brendel et al. 2018

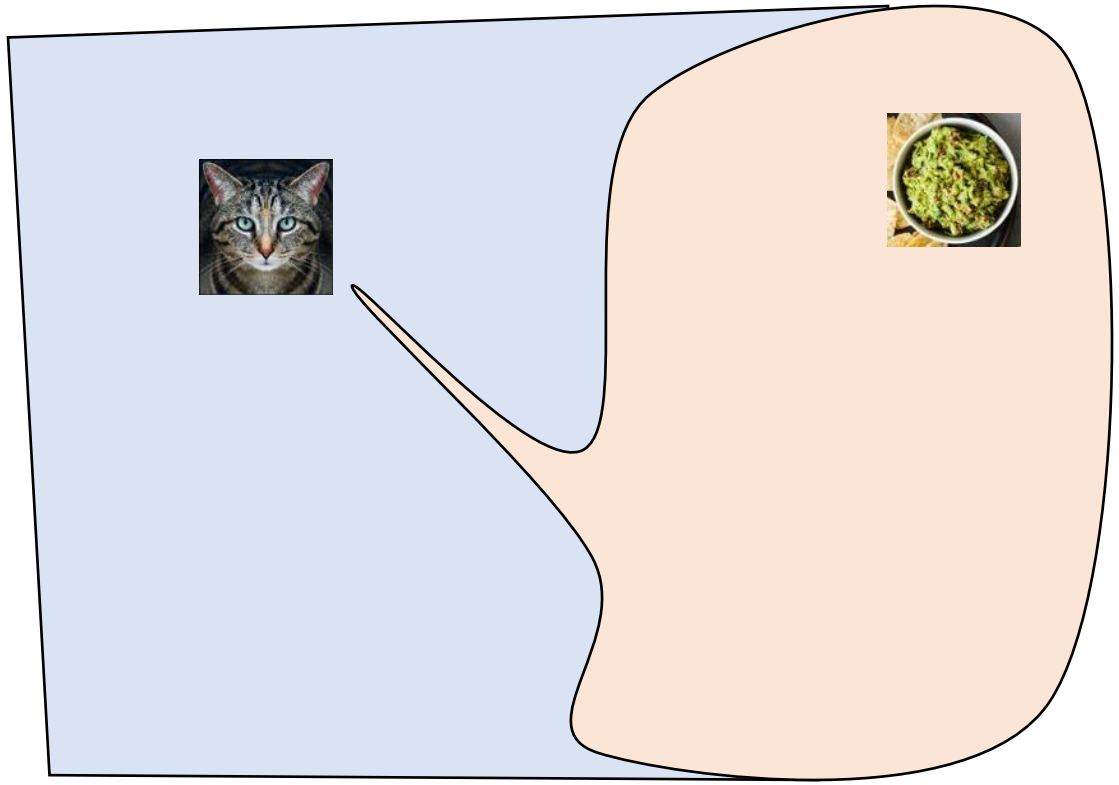
Black-box attacks

"Transfer" attacks



"Practical Black-Box Attacks against Machine Learning".
Papernot et al. 2016

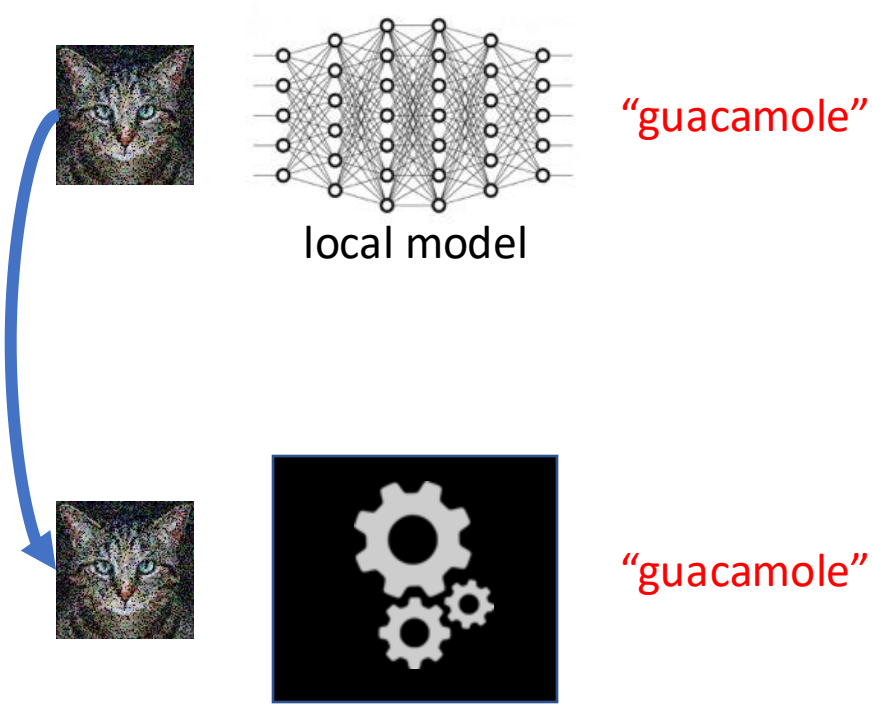
"Boundary" attacks



"Decision-Based Adversarial Attacks".
Brendel et al. 2018

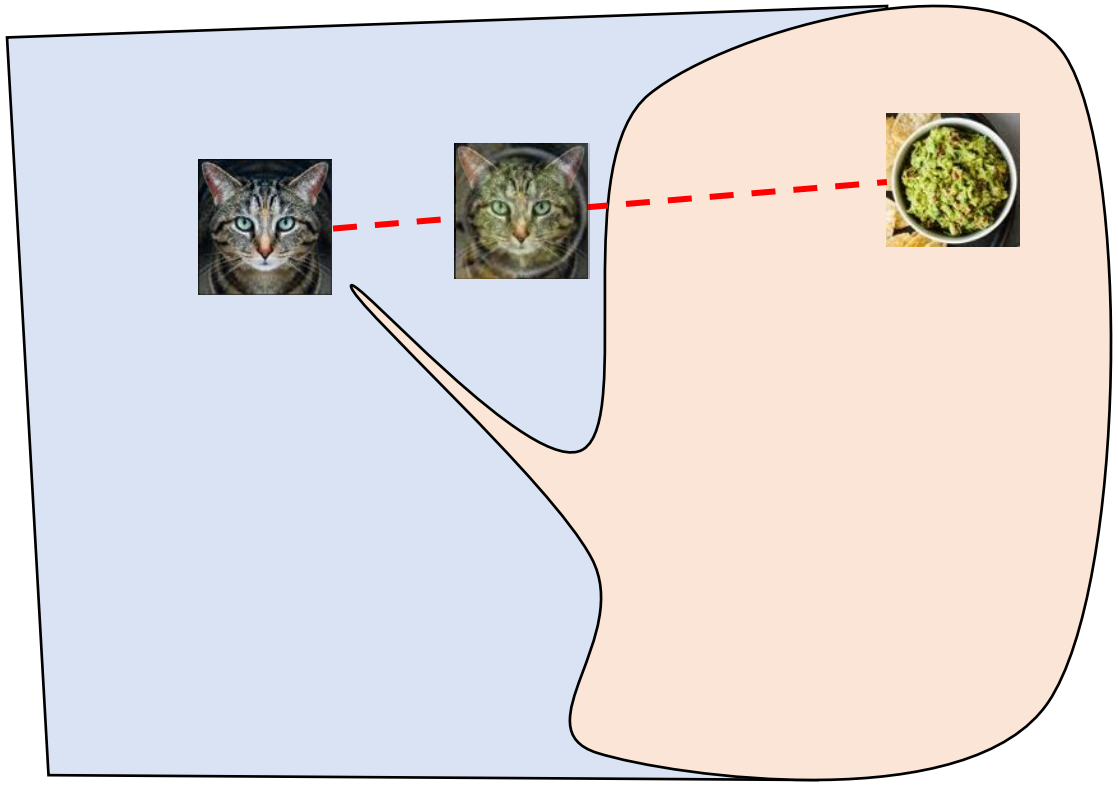
Black-box attacks

"Transfer" attacks



"Practical Black-Box Attacks against Machine Learning". Papernot et al. 2016

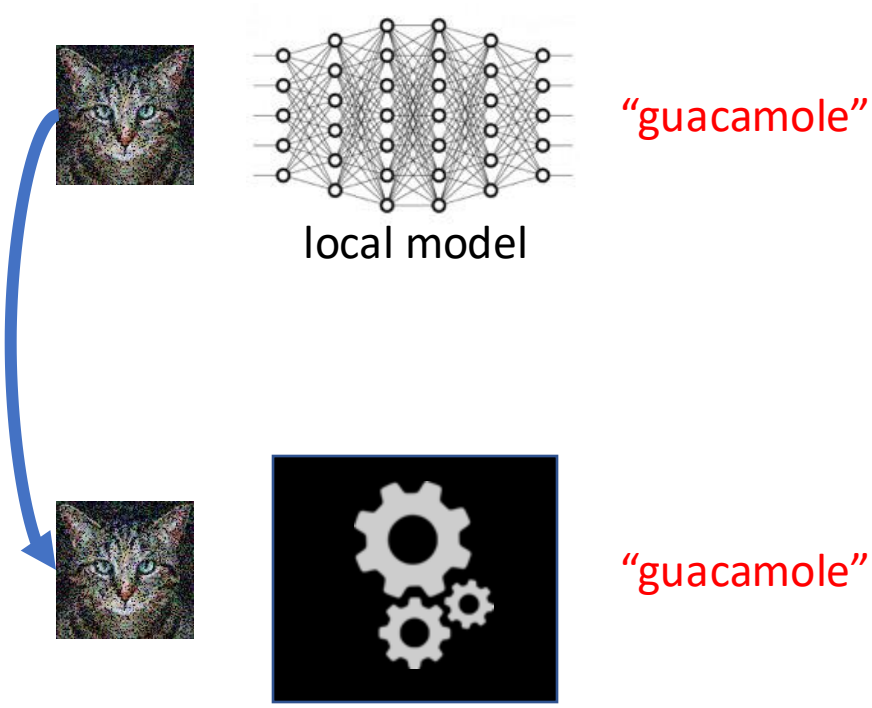
"Boundary" attacks



"Decision-Based Adversarial Attacks". Brendel et al. 2018

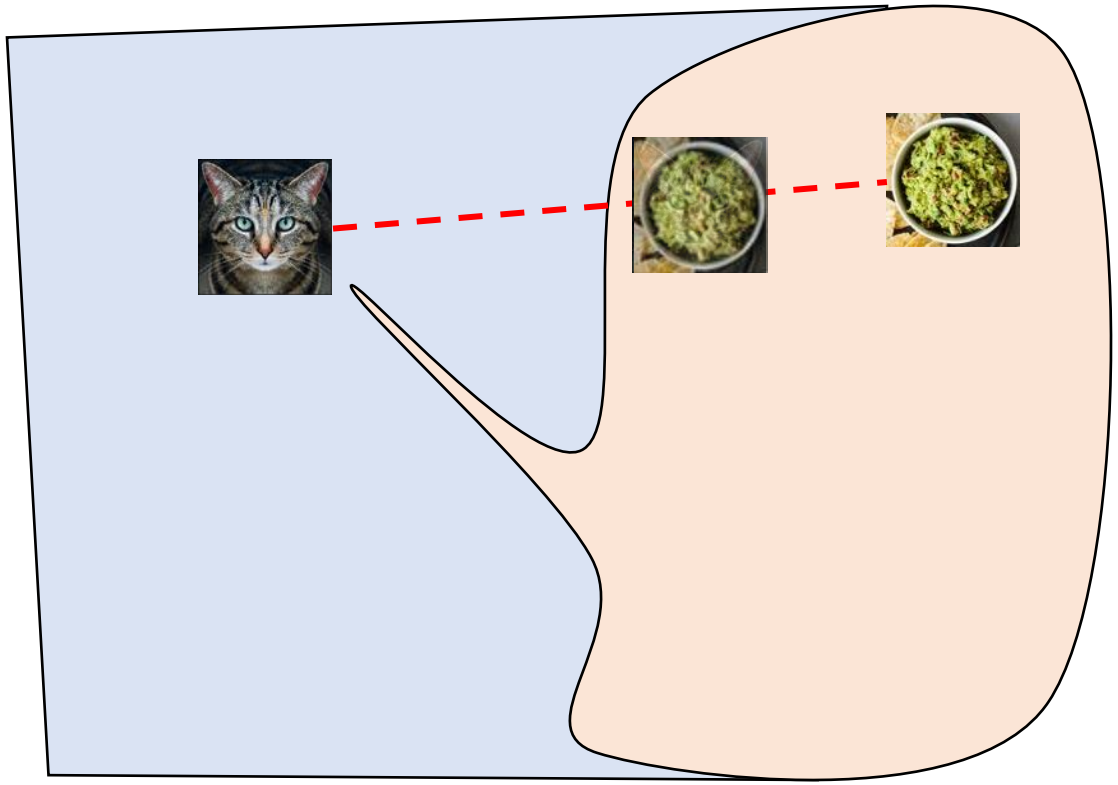
Black-box attacks

“Transfer” attacks



“Practical Black-Box Attacks against Machine Learning”.
Papernot et al. 2016

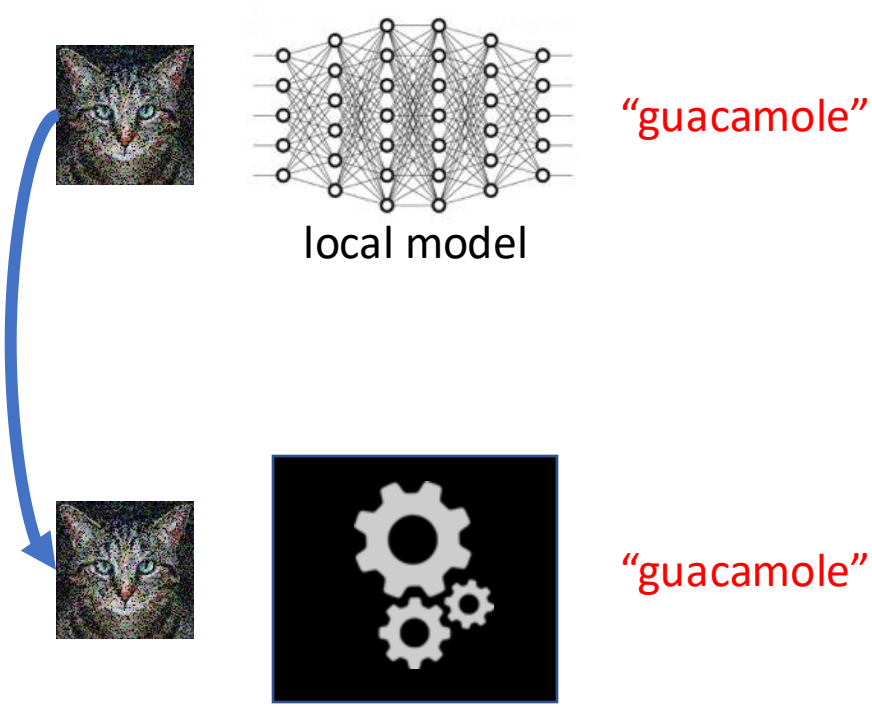
“Boundary” attacks



“Decision-Based Adversarial Attacks”.
Brendel et al. 2018

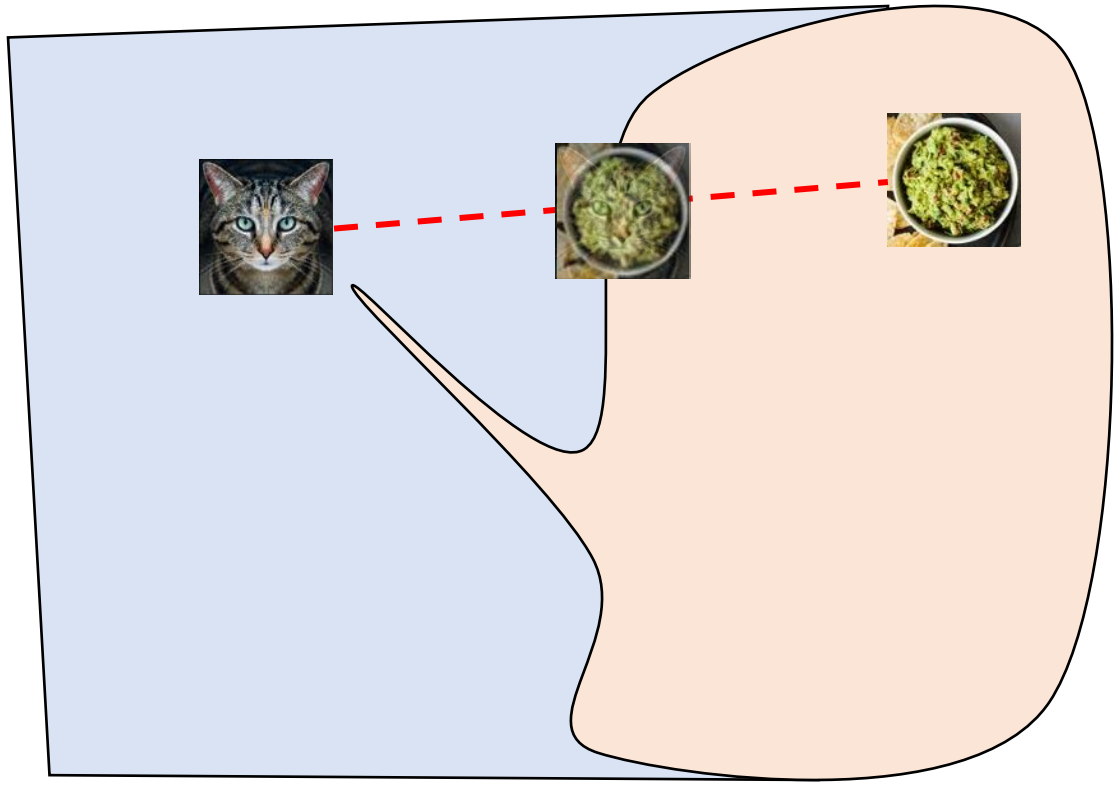
Black-box attacks

“Transfer” attacks



“Practical Black-Box Attacks against Machine Learning”.
Papernot et al. 2016

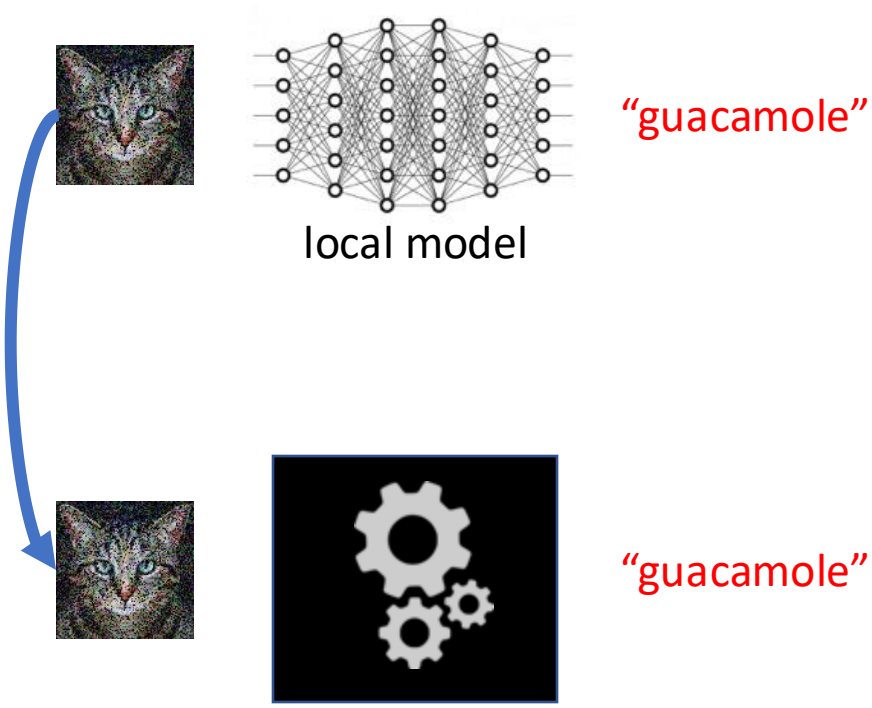
“Boundary” attacks



“Decision-Based Adversarial Attacks”.
Brendel et al. 2018

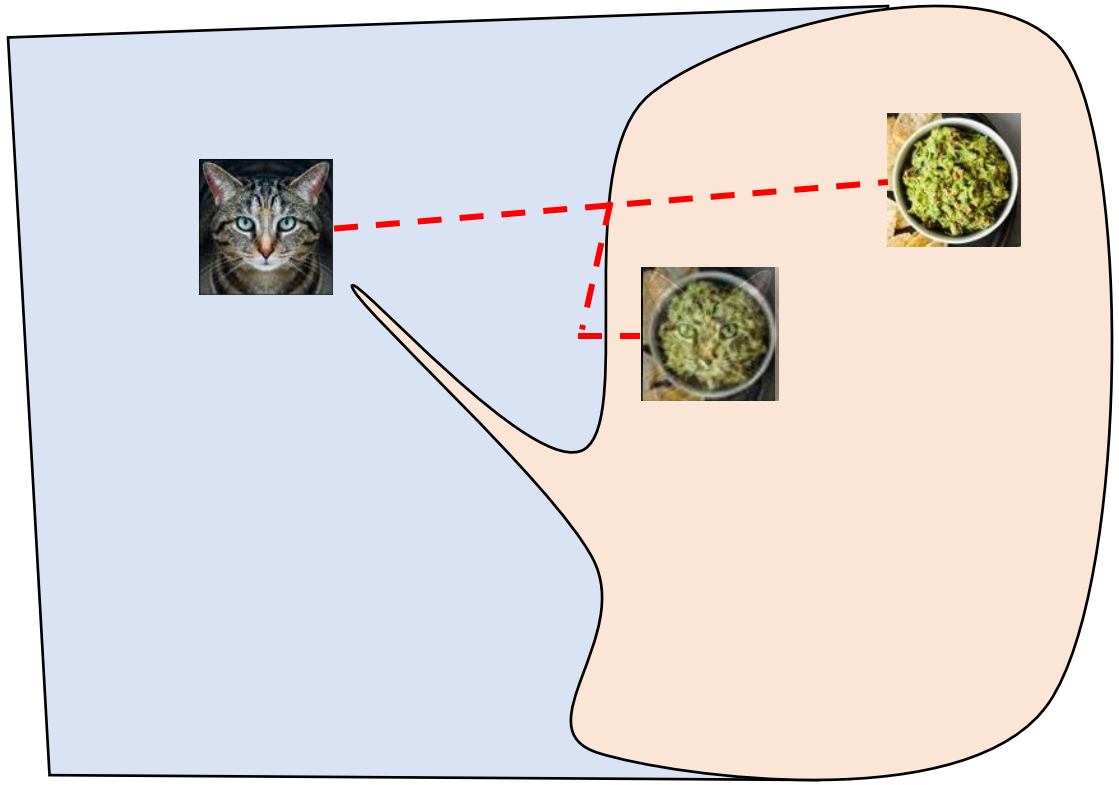
Black-box attacks

“Transfer” attacks



“Practical Black-Box Attacks against Machine Learning”.
Papernot et al. 2016

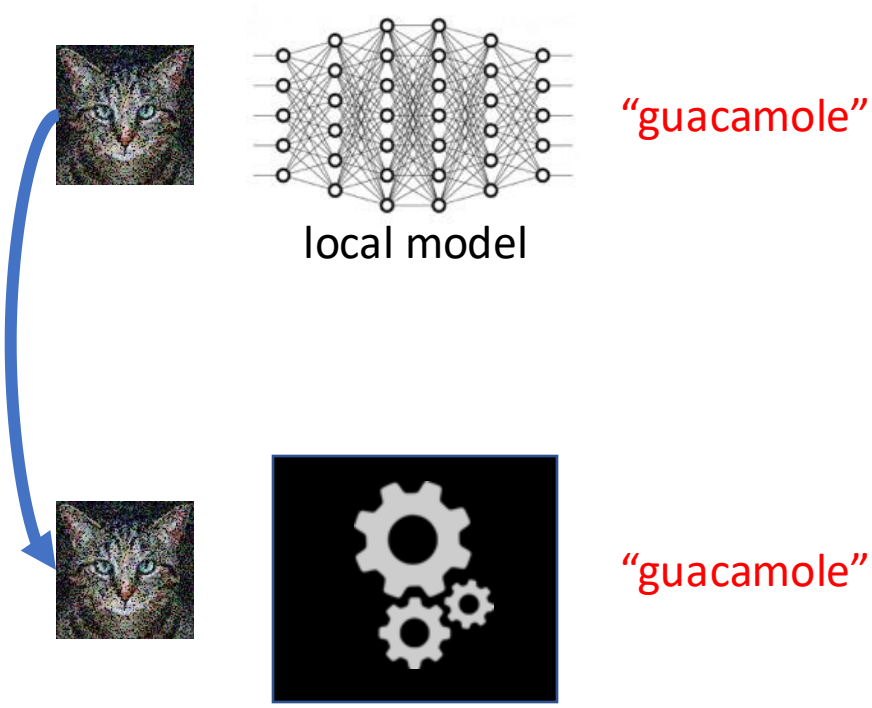
“Boundary” attacks



“Decision-Based Adversarial Attacks”.
Brendel et al. 2018

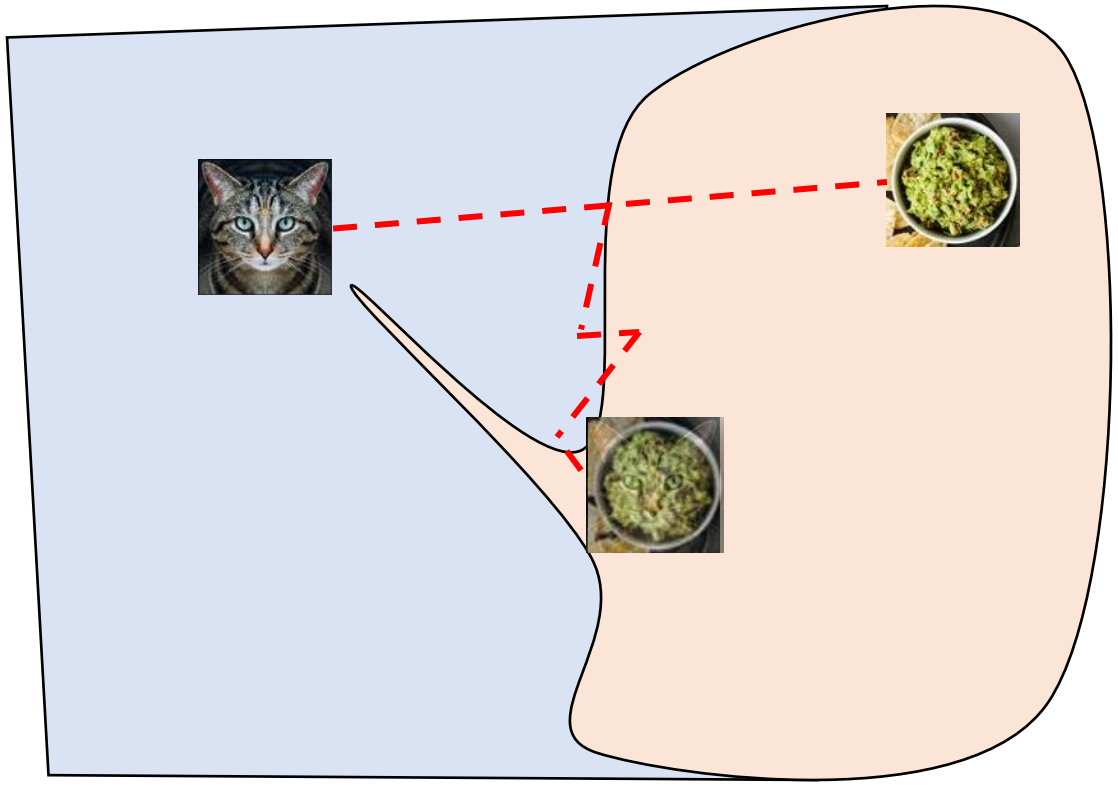
Black-box attacks

"Transfer" attacks



"Practical Black-Box Attacks against Machine Learning".
Papernot et al. 2016

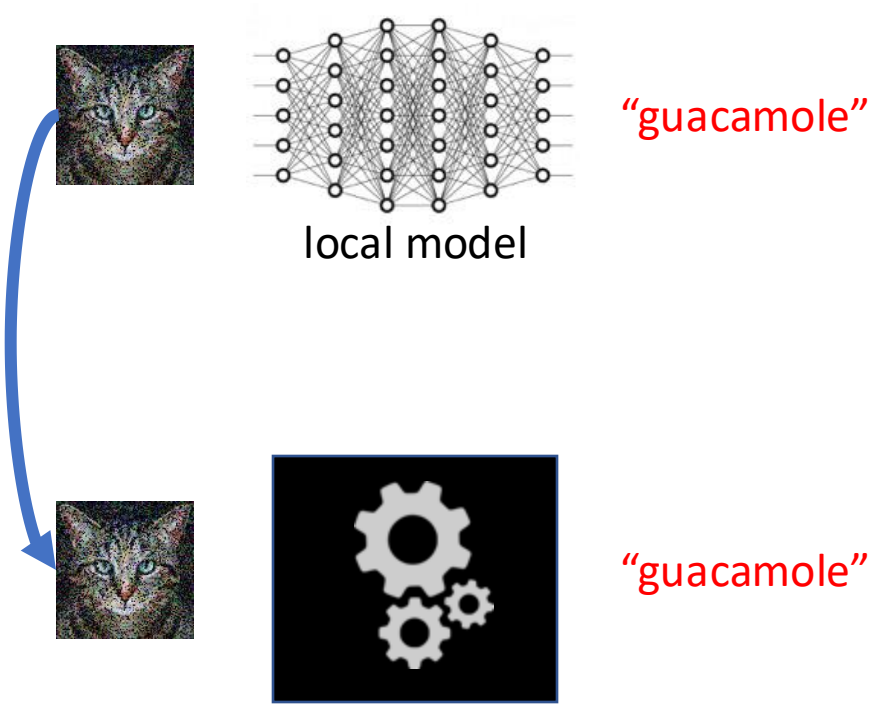
"Boundary" attacks



"Decision-Based Adversarial Attacks".
Brendel et al. 2018

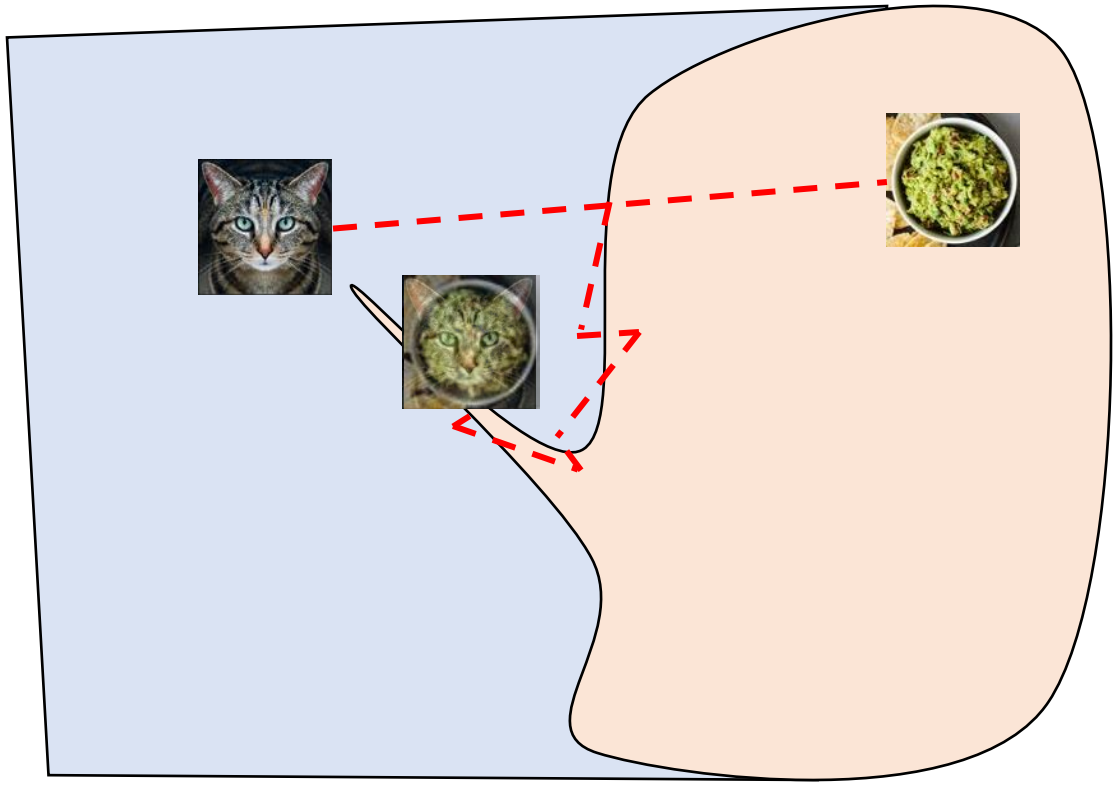
Black-box attacks

"Transfer" attacks



"Practical Black-Box Attacks against Machine Learning".
Papernot et al. 2016

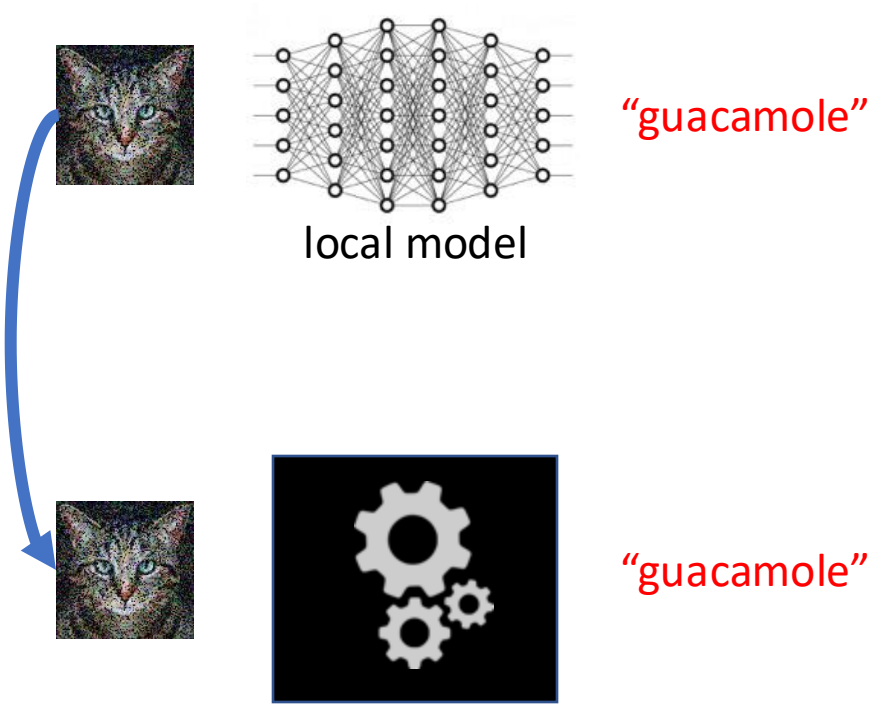
"Boundary" attacks



"Decision-Based Adversarial Attacks".
Brendel et al. 2018

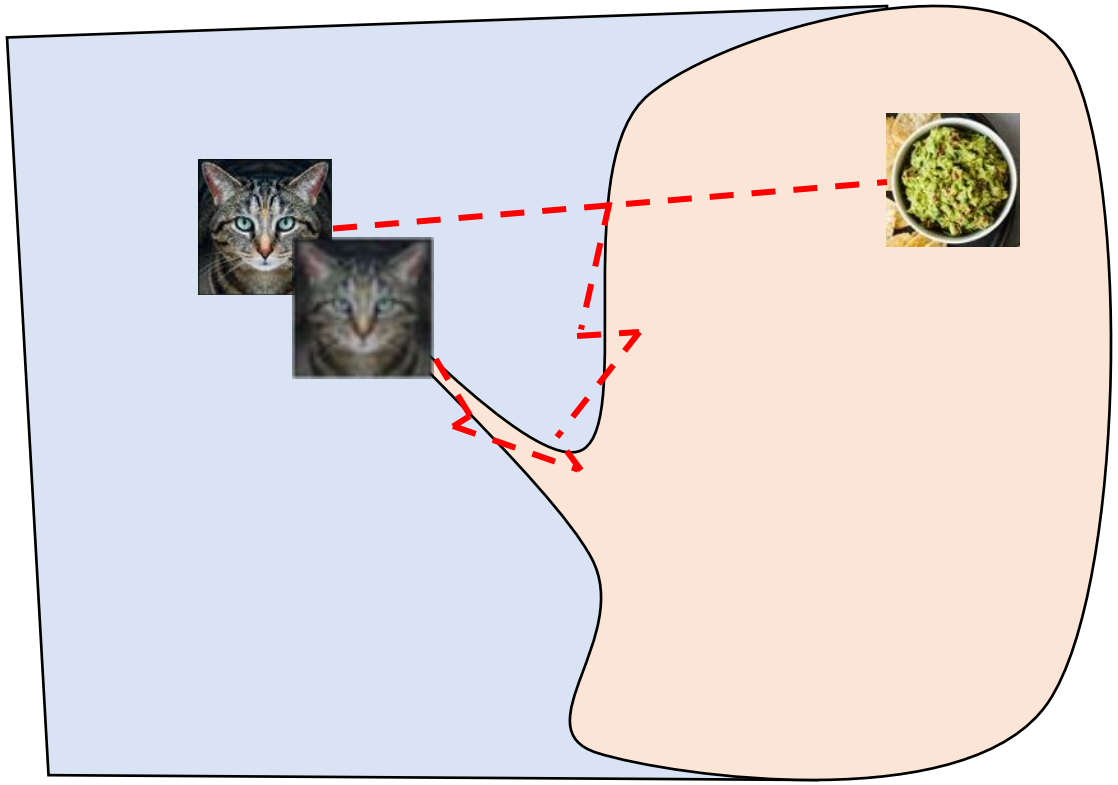
Black-box attacks

"Transfer" attacks



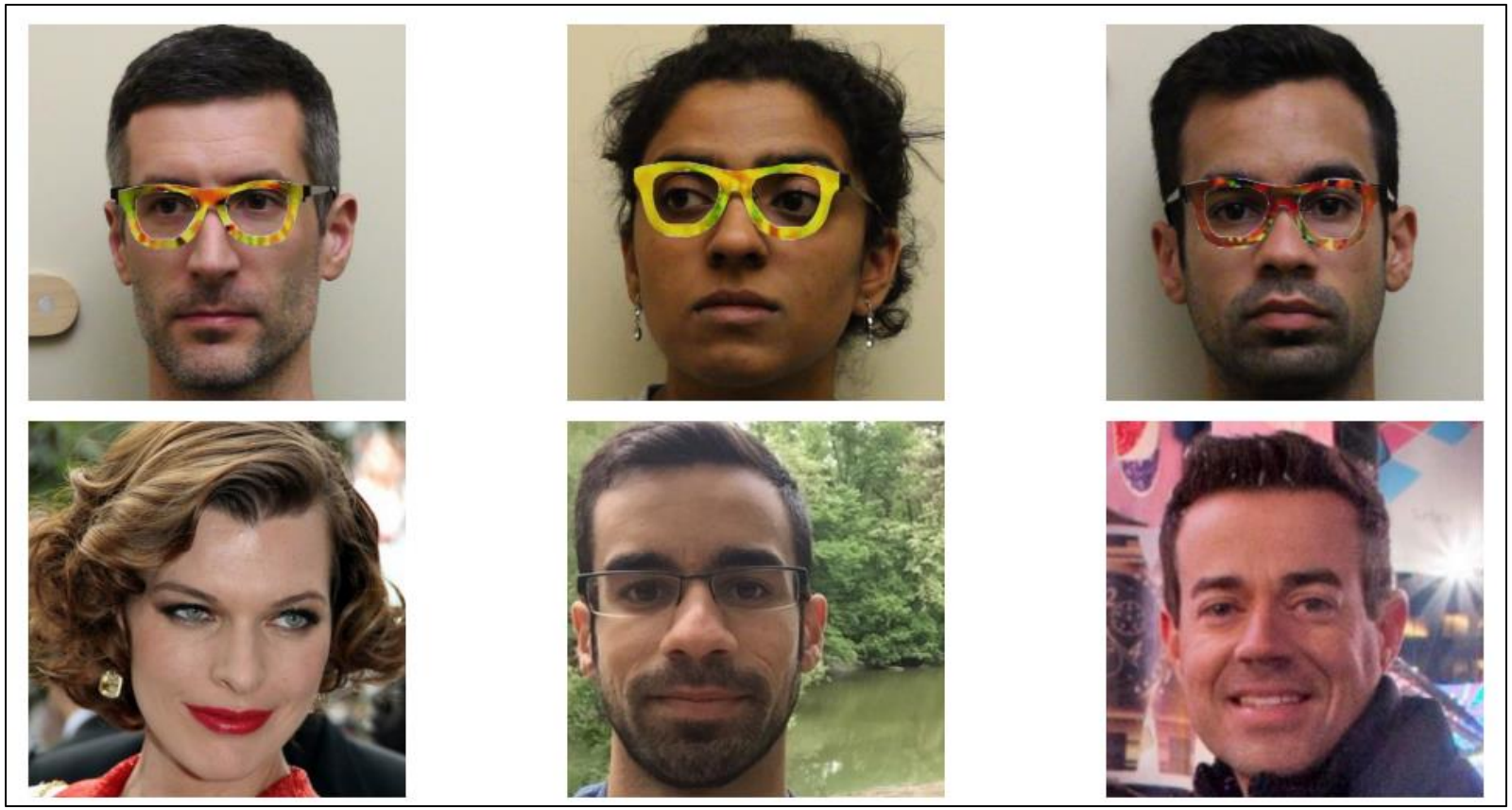
"Practical Black-Box Attacks against Machine Learning".
Papernot et al. 2016

"Boundary" attacks



"Decision-Based Adversarial Attacks".
Brendel et al. 2018

Fooling face recognition systems using glasses

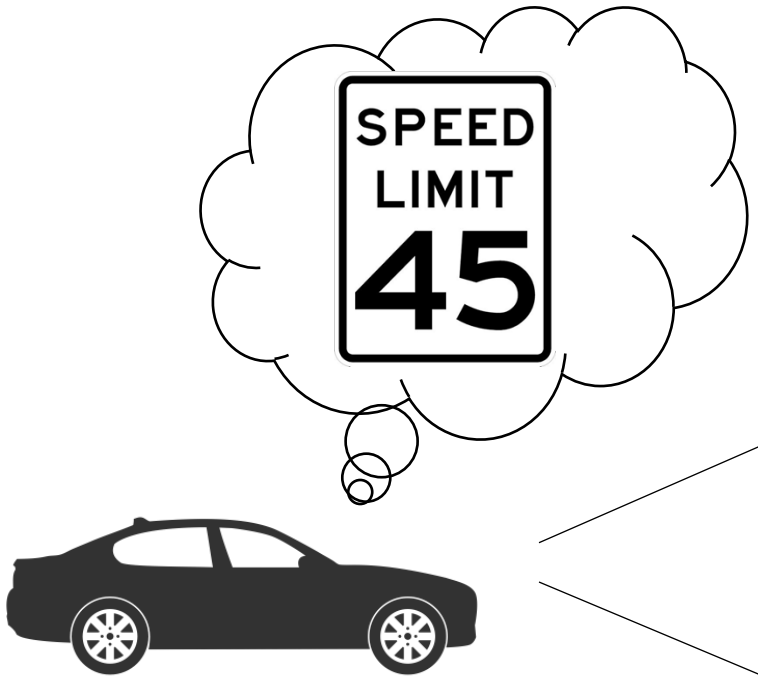


Attackers

Results

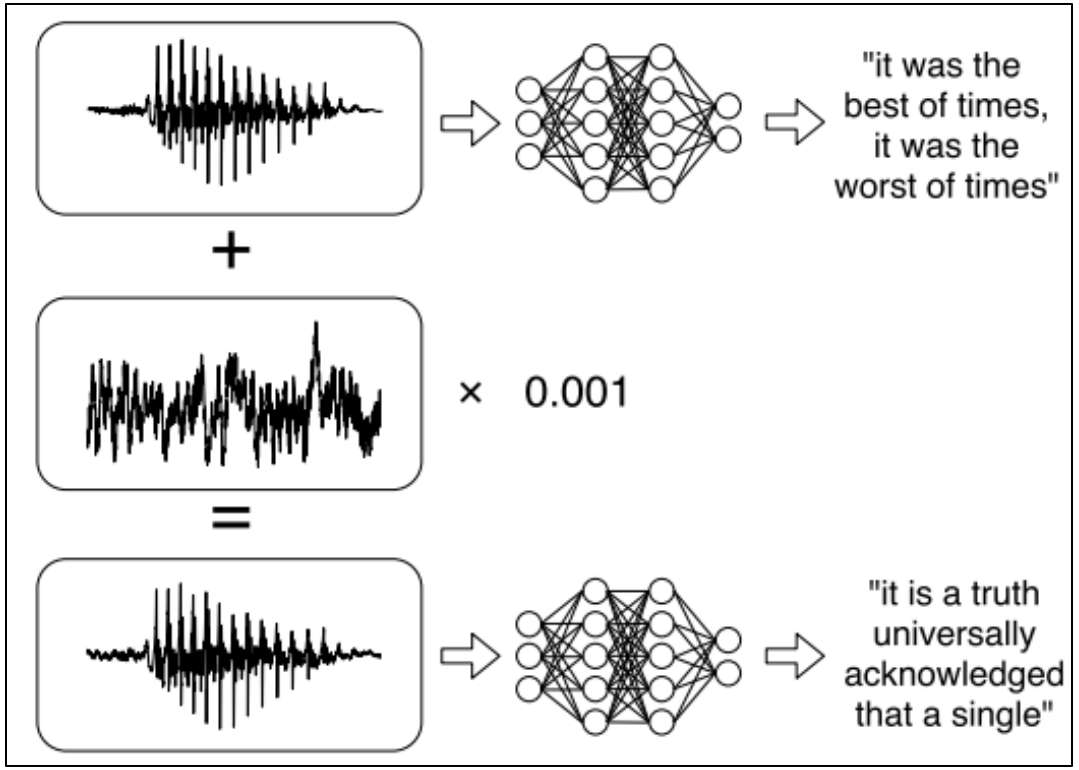
M. Sharif et al., “Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition,” in Proc. of CCS '16.

Fooling self-driving car with stickers



K. Eykholt et al., “Robust Physical-World Attacks on Deep Learning Visual Classification,” in Proc. of CVPR, 2018.

Audio adversarial examples



“without the dataset the article is useless”

“okay google browse to evil dot com”

N. Carlini and D. Wagner, “Audio Adversarial Examples: Targeted Attacks on Speech-to-Text,” in Proc. of DLS, 2018.

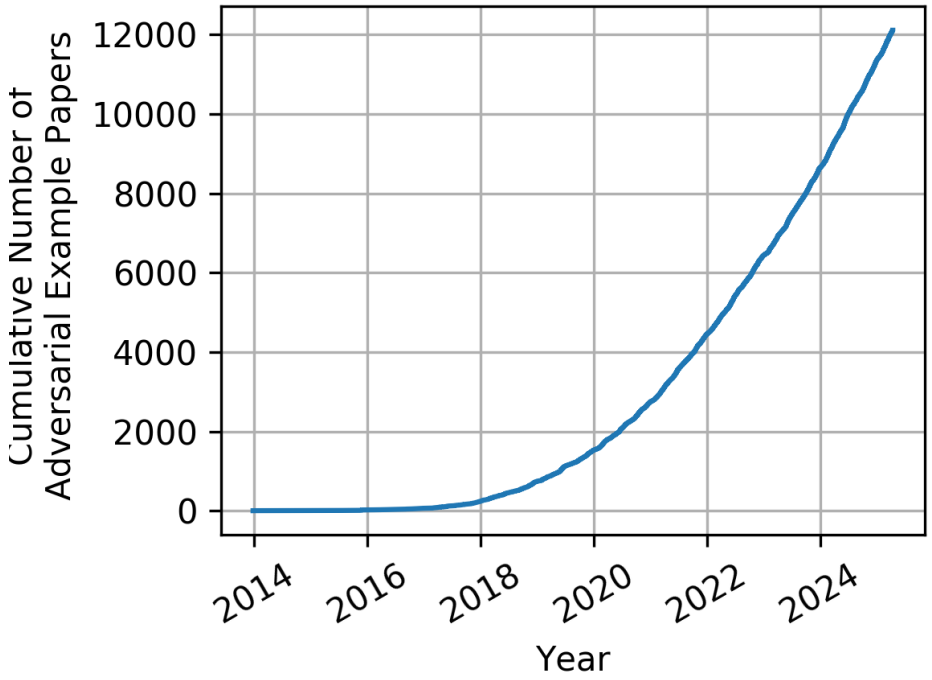
You can craft adversarial examples even without expert knowledge!😮

CleverHans[1]

AdvBox[2]

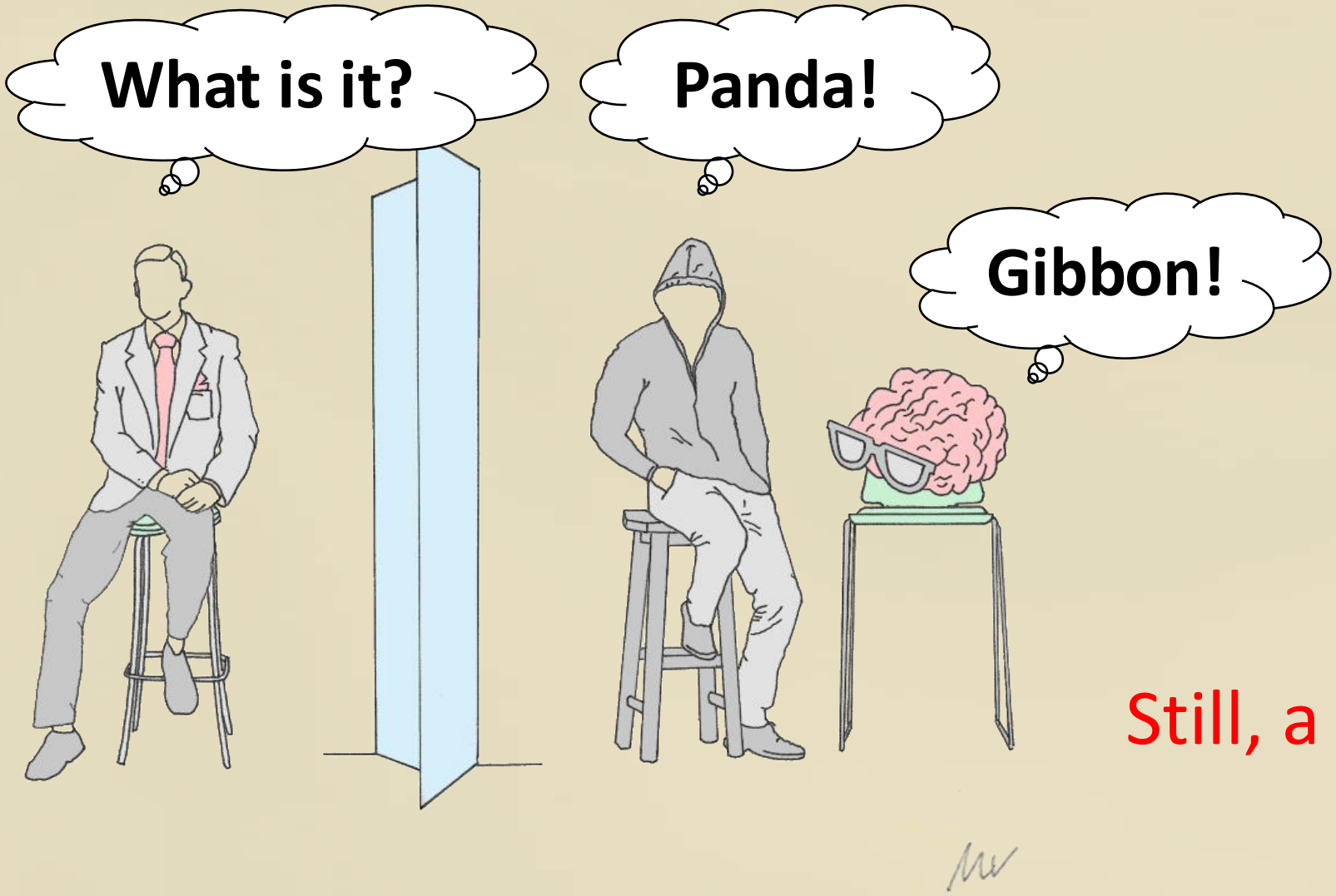
advertorch[3]

[1] https://github.com/tensorflow/cleverhans
[2] https://github.com/advboxes/AdvBox
[3] https://github.com/BorealisAI/advertorch



>12,000 papers[4], no dominant solutions have been found so far.

https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html



Still, a long way to go! 😮

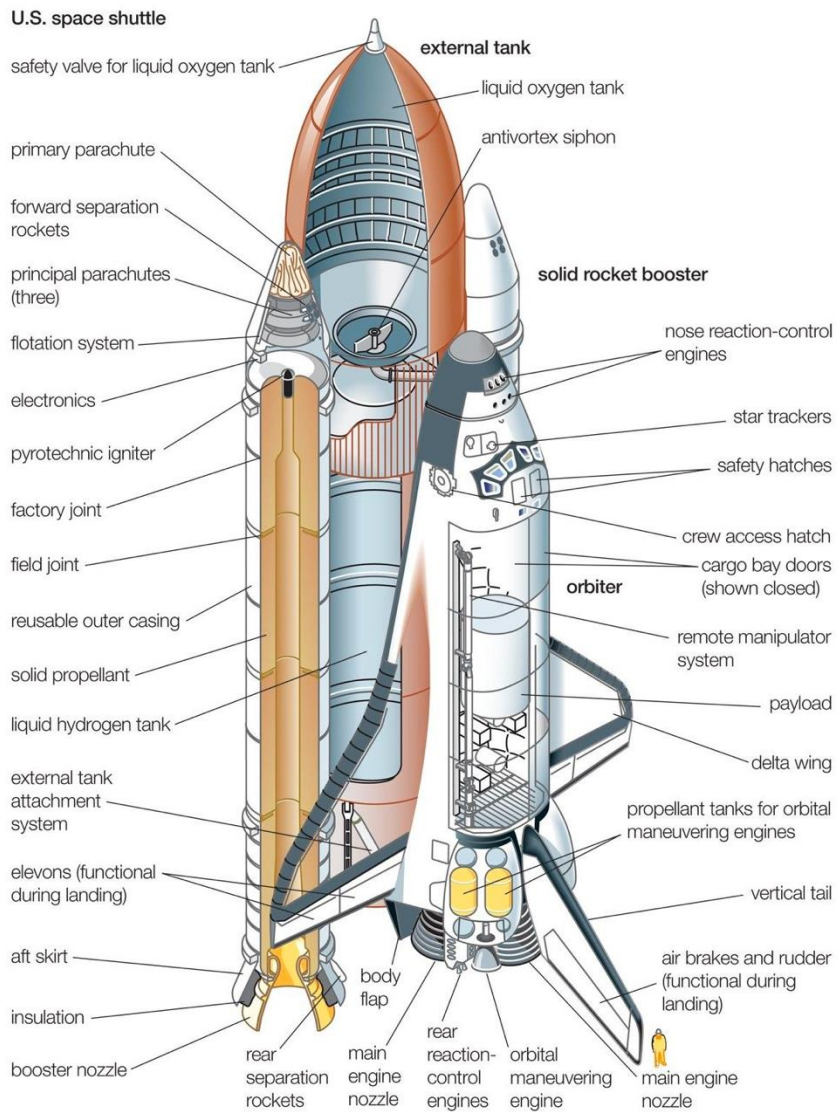
Source: https://www.knowmail.me/blog/whats-turing-test-ai-pass/

The “AI Rocket”

Engine (Model) → Reliable

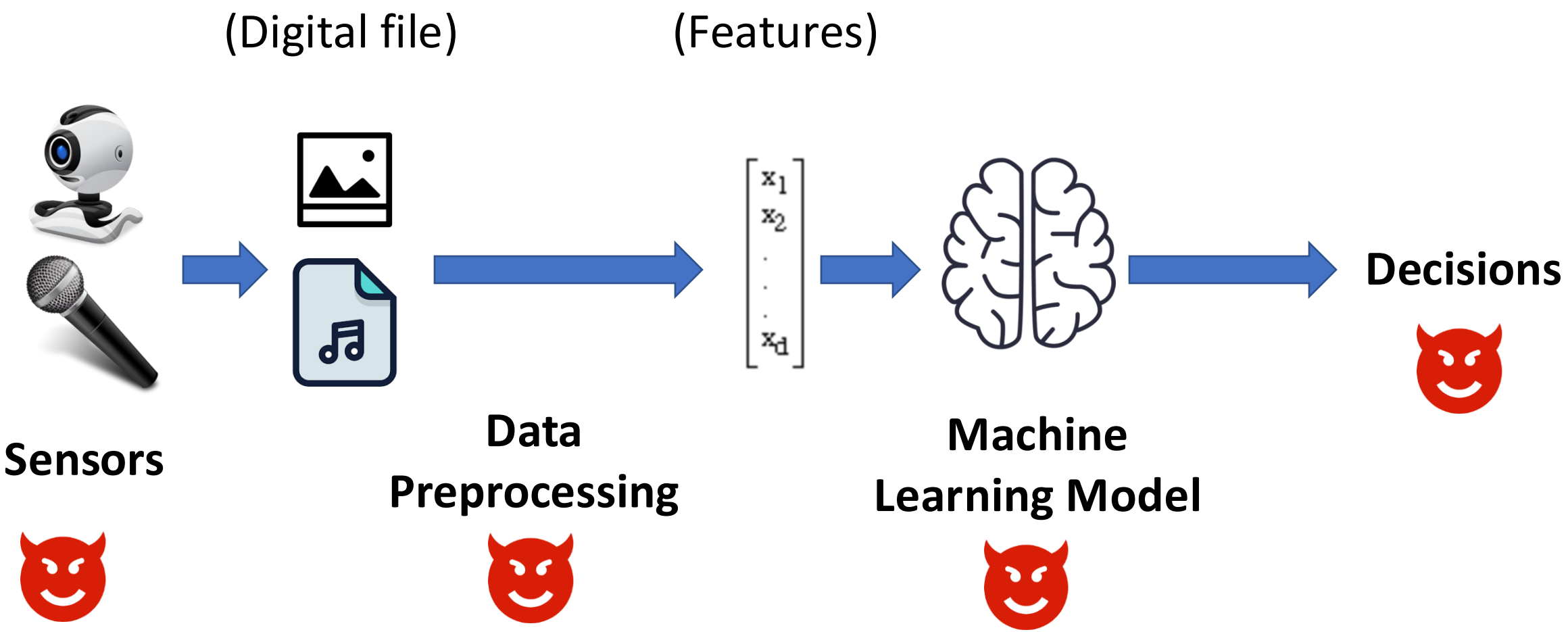
Fuel (Data) → Privacy

Source: https://wallpaperplay.com/board/rocket-launch-wallpapers



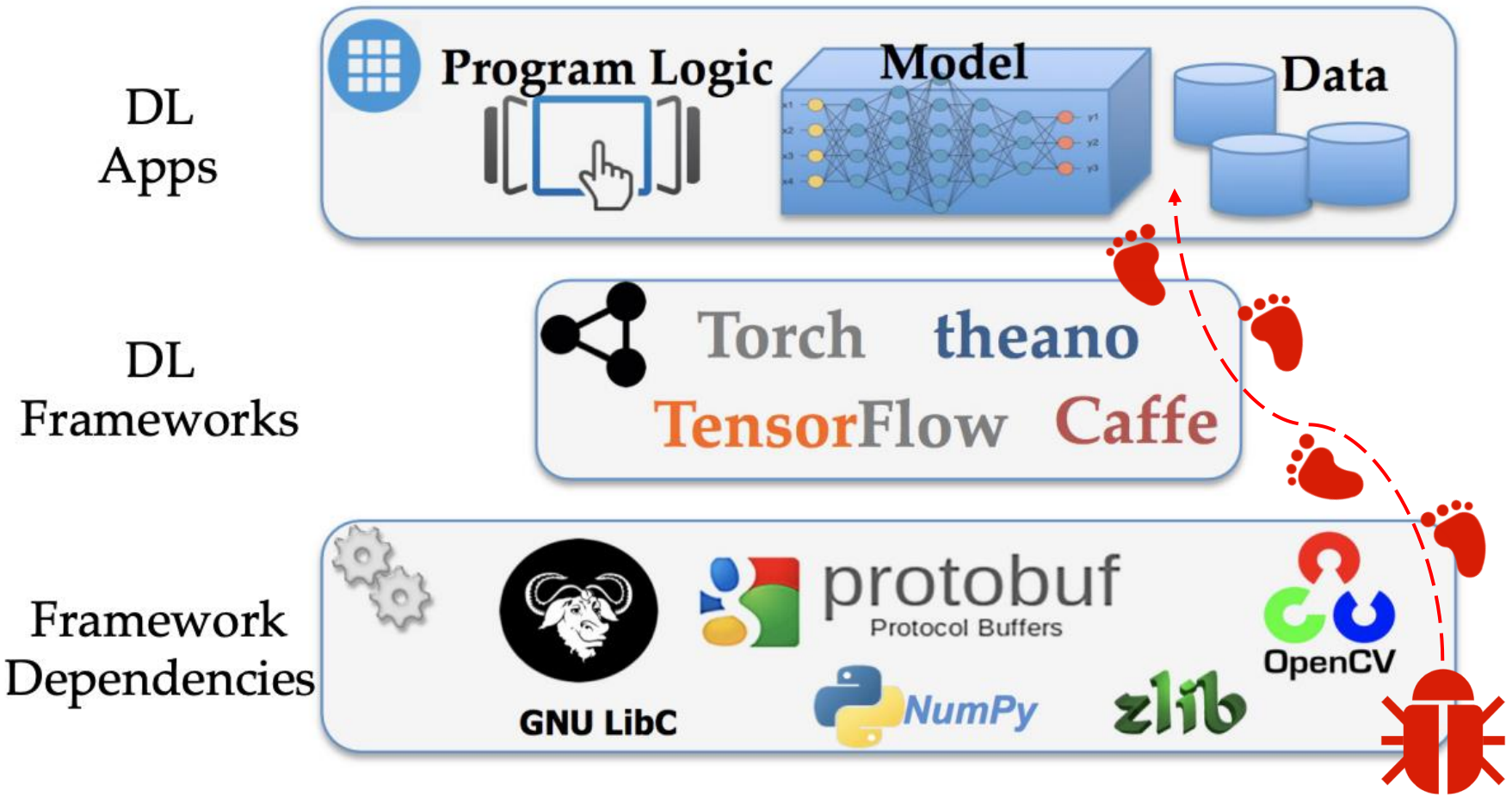
Source: http://mechanicstips.blogspot.com/2016/10/space-shuttle-components.html

A Complete Machine Learning System



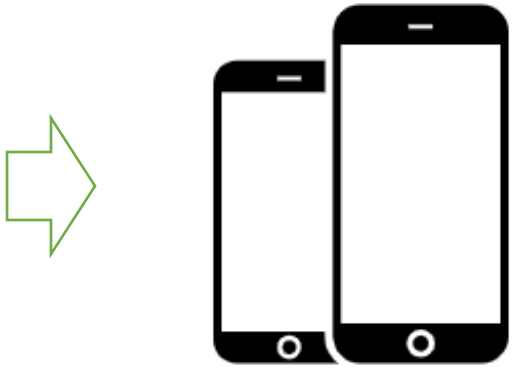
Security Risks in System Implementations

Q. Xiao et al., “Security Risks in Deep Learning Implementations,” in Proc. of DLS, 2018.



Q. Xiao et al., “Security Risks in Deep Learning Implementations,” in Proc. of DLS, 2018.

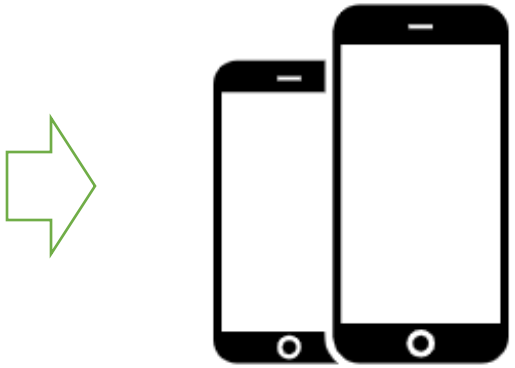
Original Image



Bulldog

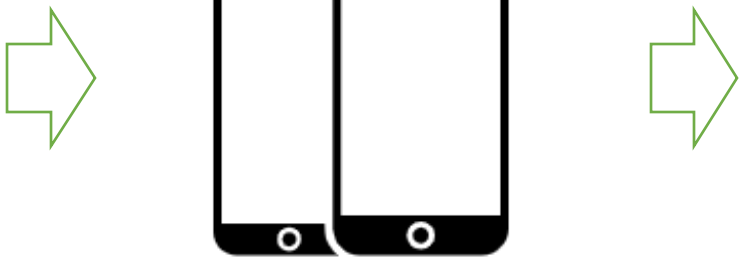
Q. Xiao et al., “Security Risks in Deep Learning Implementations,” in Proc. of DLS, 2018.

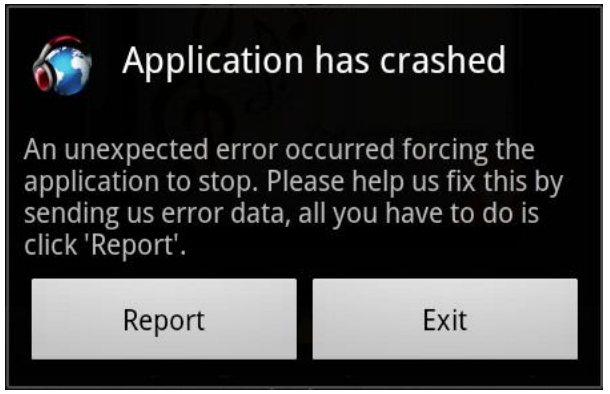
Malicious Input 1



Cat

Q. Xiao et al., “Security Risks in Deep Learning Implementations,” in Proc. of DLS, 2018.

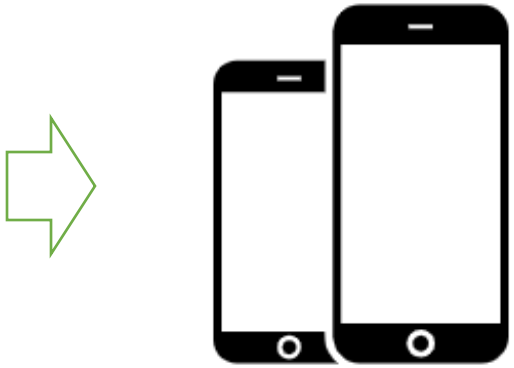




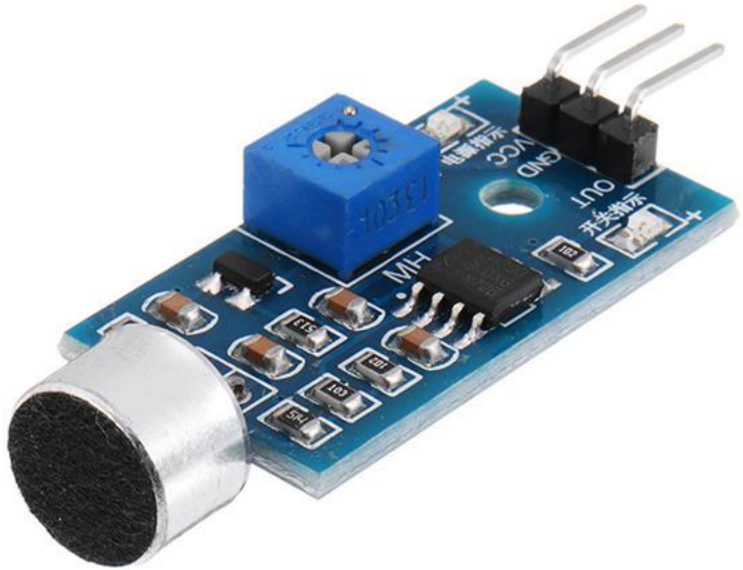
Malicious Input 2

Q. Xiao et al., “Security Risks in Deep Learning Implementations,” in Proc. of DLS, 2018.

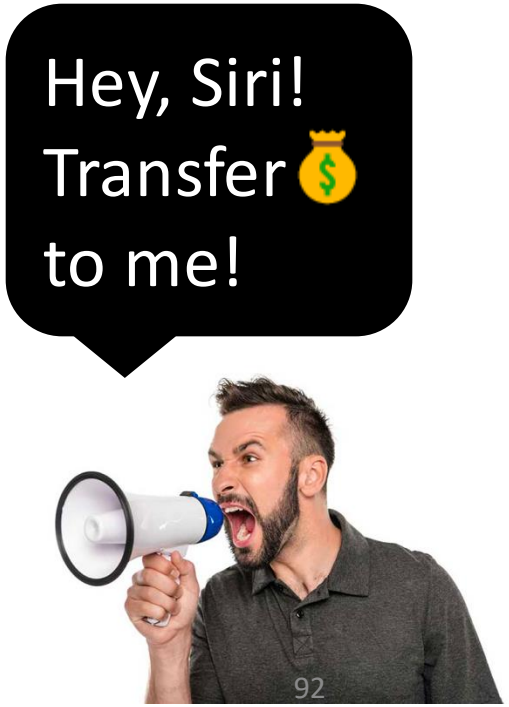
Malicious Input 3

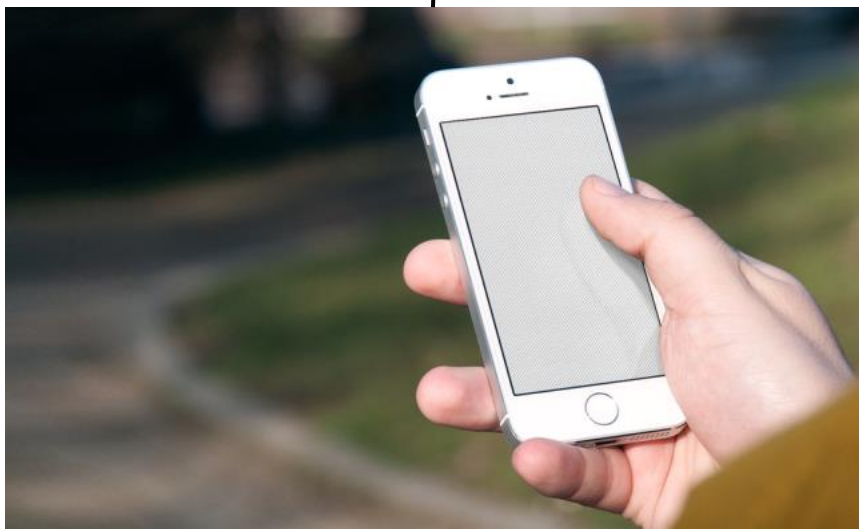


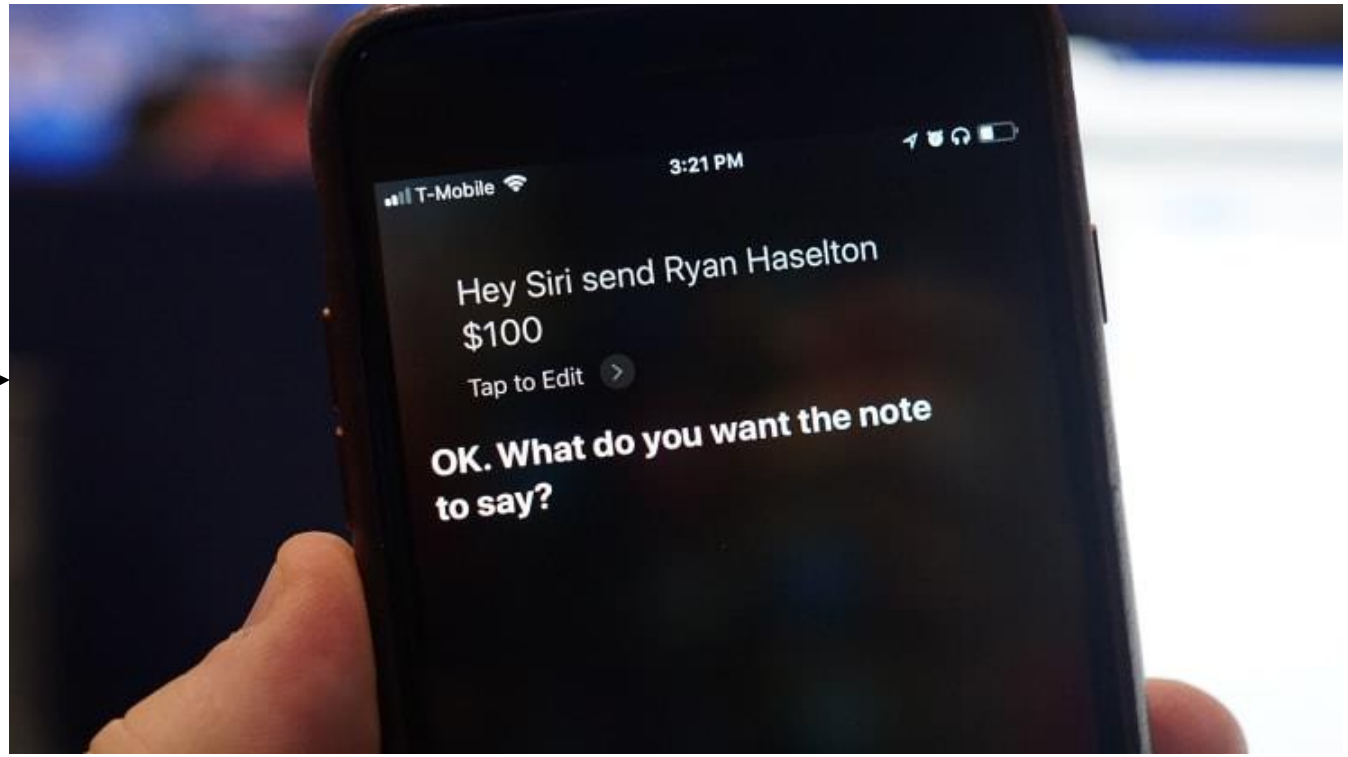
Q. Xiao et al., “Security Risks in Deep Learning Implementations,” in Proc. of DLS, 2018.



Sensor Spoofing Attack


Hey, Siri!
Transfer 💰
to me!

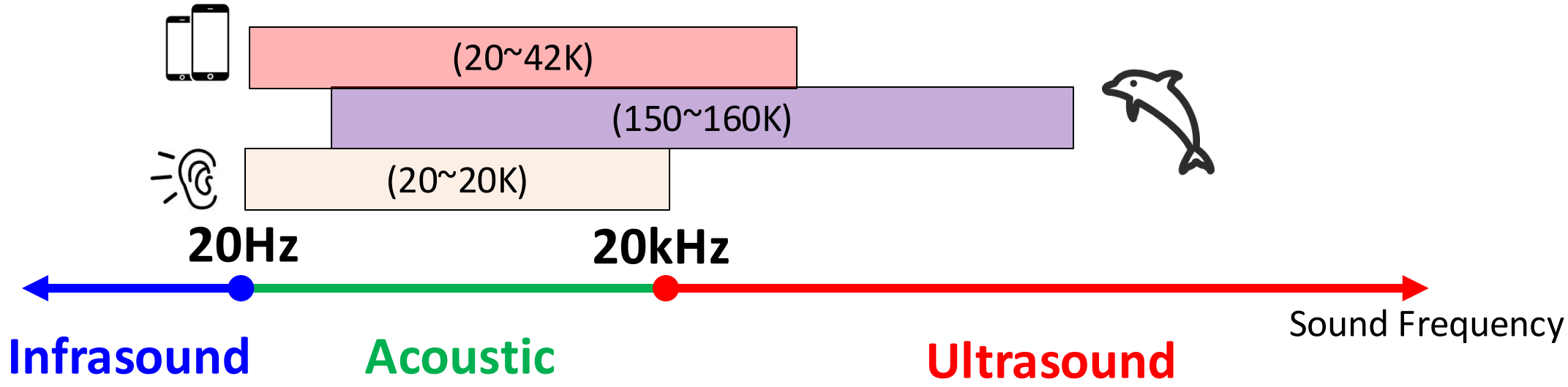




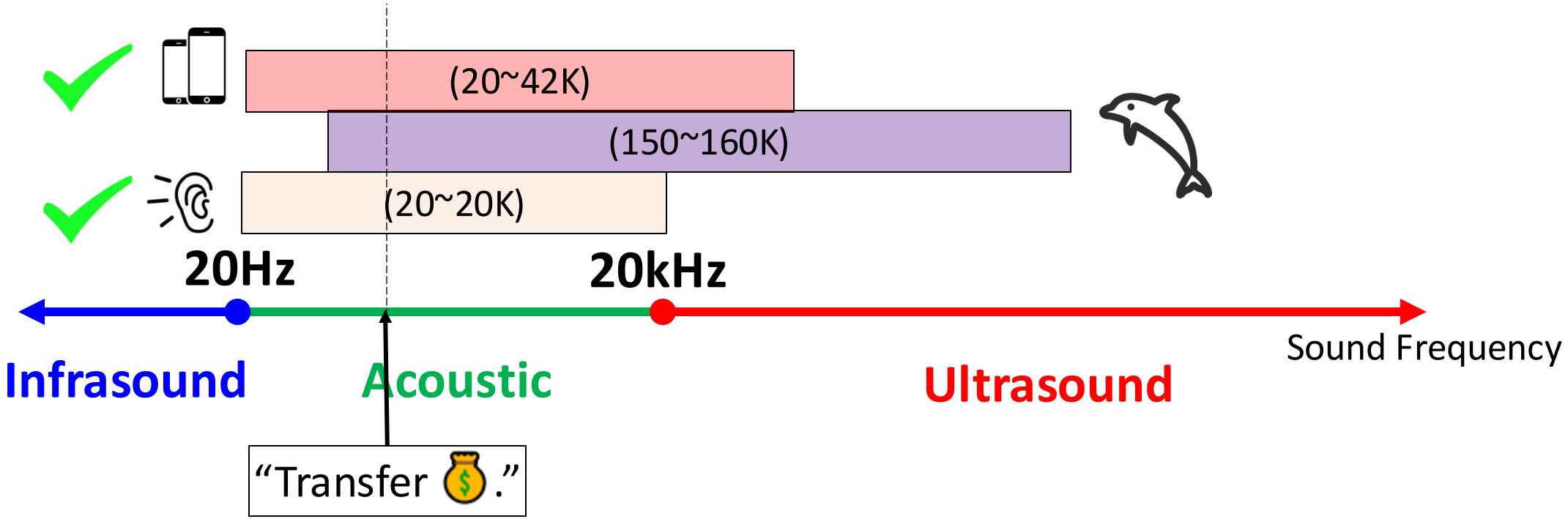
Happy Ending:

You notice it, and call the police. ☺

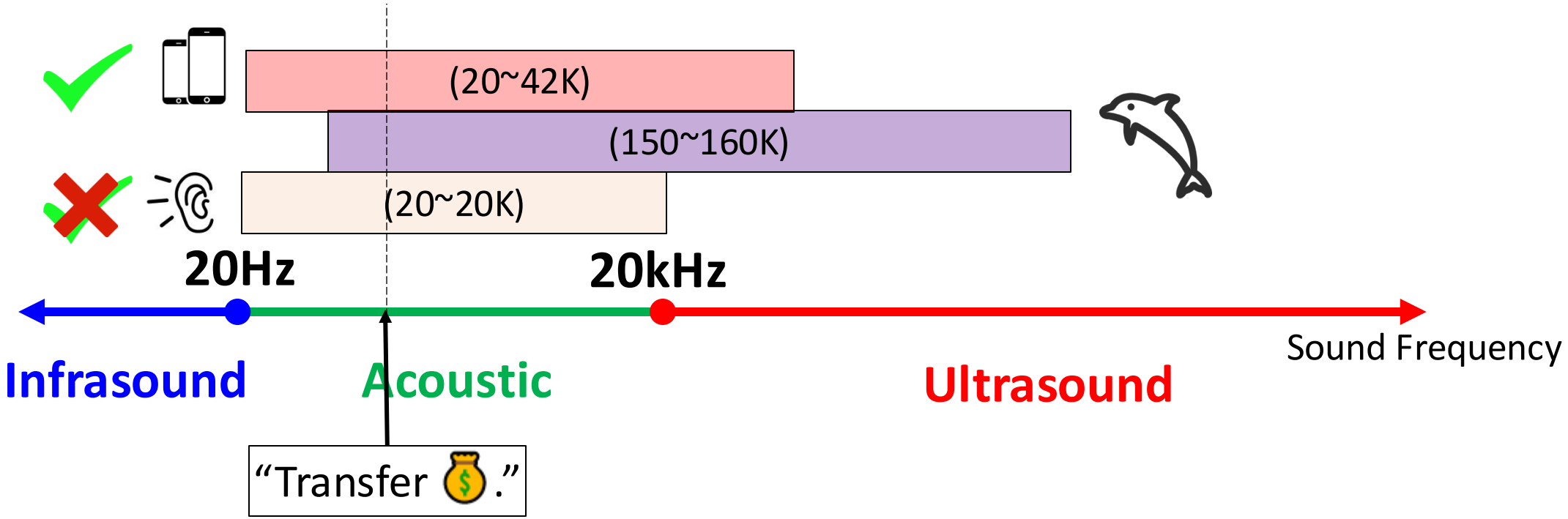

Hey, Siri!
Transfer 💰 to
me!



G. Zhang et al., “DolphinAttack: Inaudible Voice Commands,” in Proc. of CCS, 2017.



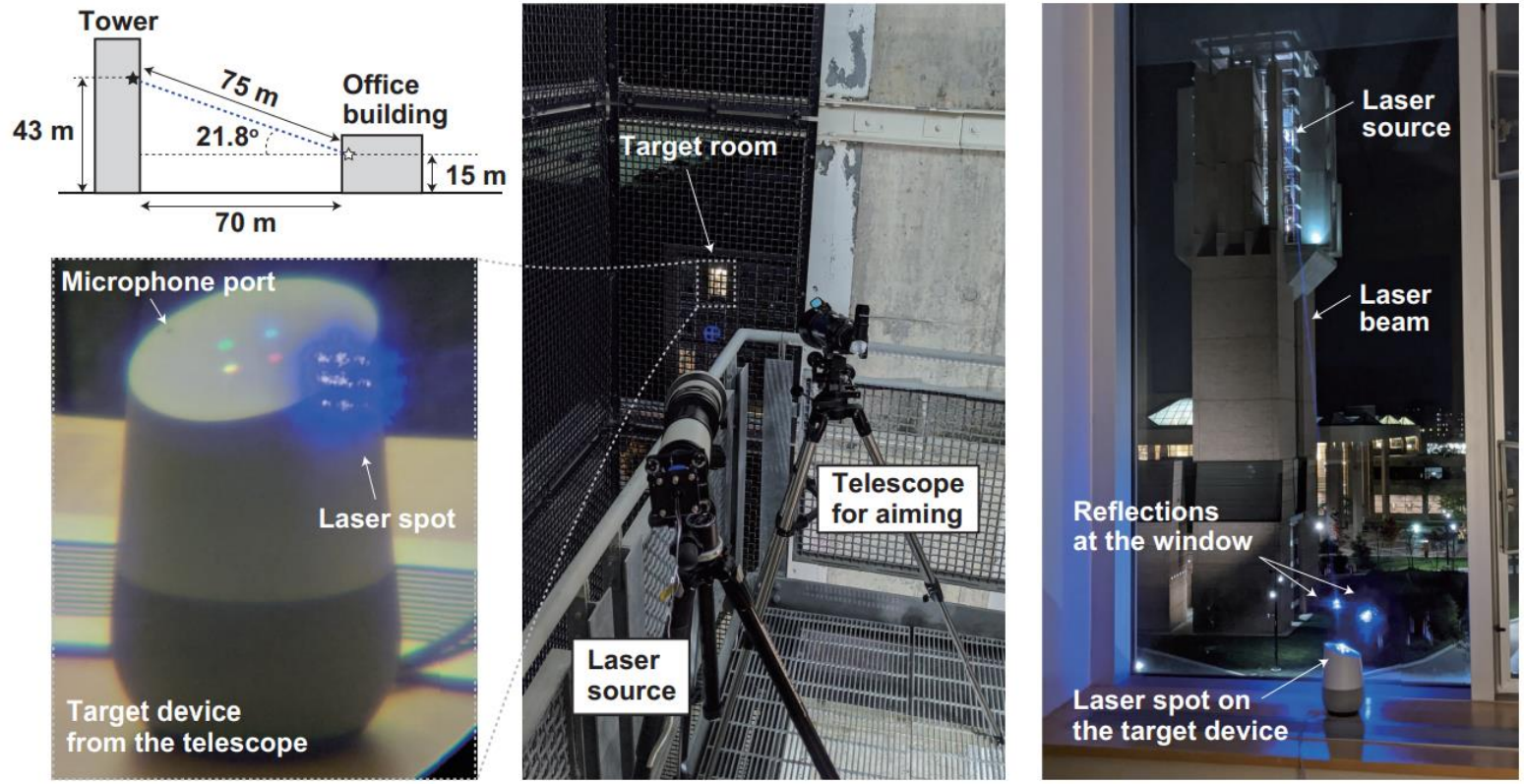
G. Zhang et al., “DolphinAttack: Inaudible Voice Commands,” in Proc. of CCS, 2017.



Inaudible Attack: "*DolphinAttack*"

G. Zhang et al., "DolphinAttack: Inaudible Voice Commands," in Proc. of CCS, 2017.

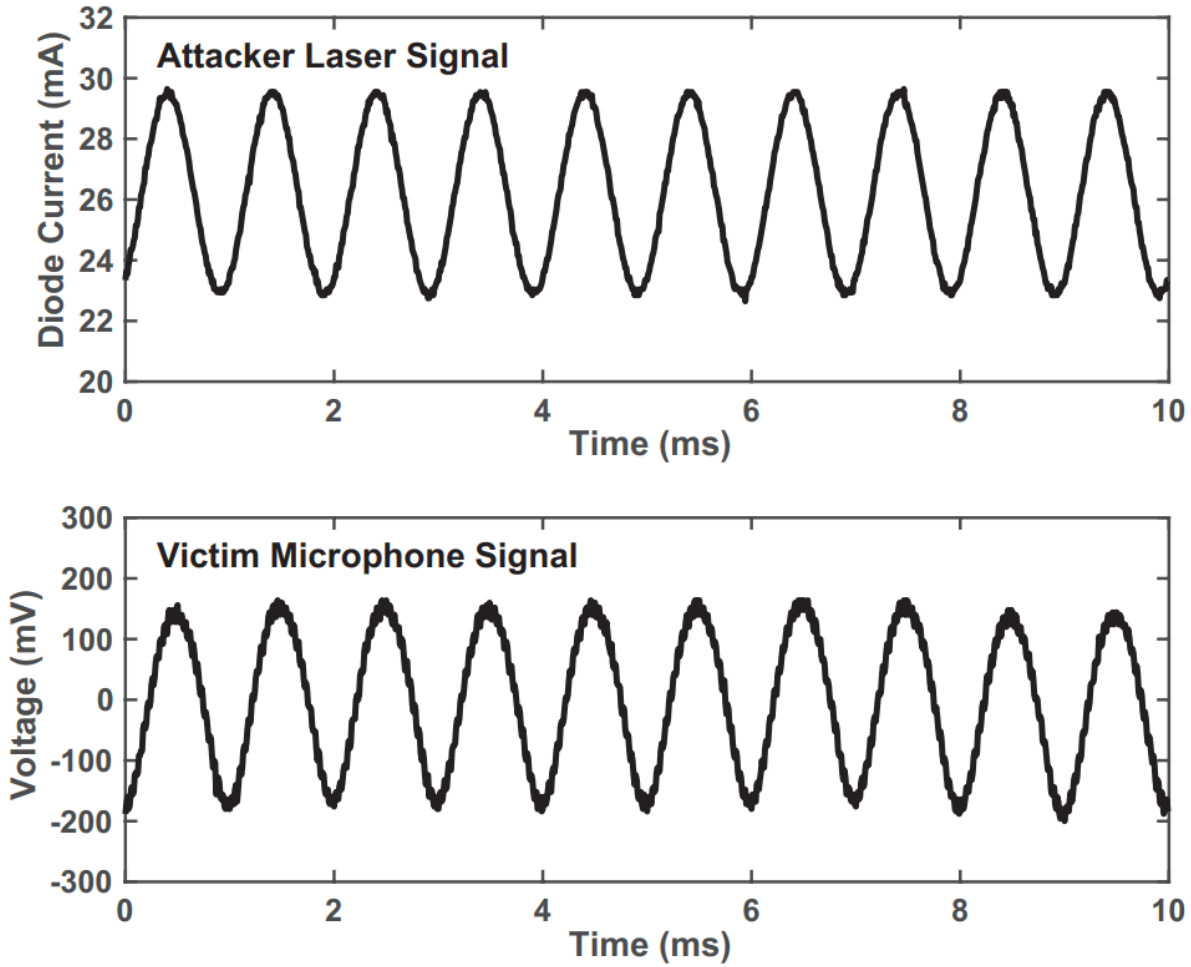
Attackers can even send commands to your voice assistant using light…



LIGHT COMMANDS

T. Sugawara et al., “Light Commands: Laser-Based Audio Injection Attacks on Voice-Controllable Systems,” 2019.

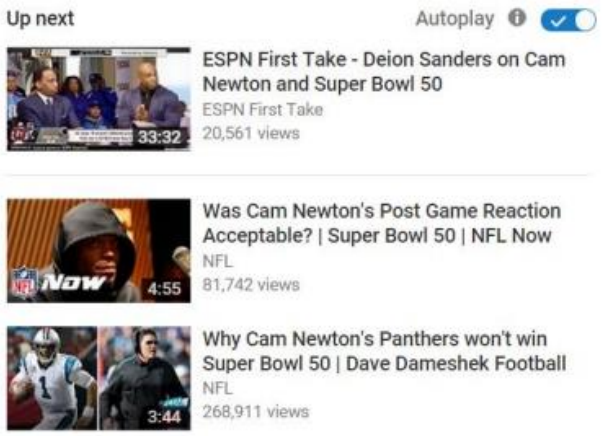
Microphone can sense **light** signals!



LIGHT COMMANDS

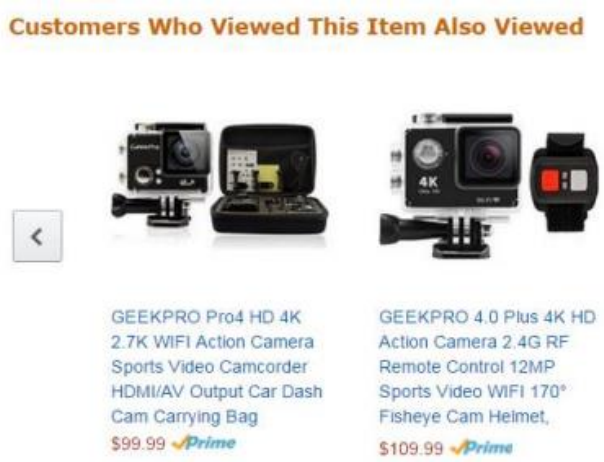
T. Sugawara et al., “Light Commands: Laser-Based Audio Injection Attacks on Voice-Controllable Systems,” 2019.

Poisoning Recommendation Systems



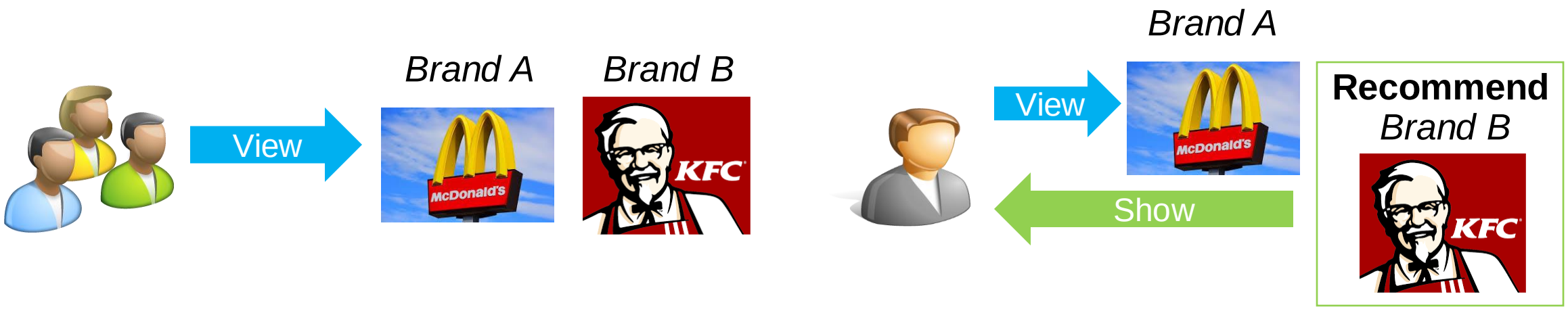
YouTube

eBay



Amazon

G. Yang et al., “Fake Co-visitation Injection Attacks to Recommender Systems,” in Proc. of NDSS, 2017.



"Co-visitation"

G. Yang et al., "Fake Co-visitation Injection Attacks to Recommender Systems," in Proc. of NDSS, 2017.

View

Brand A

Brand C

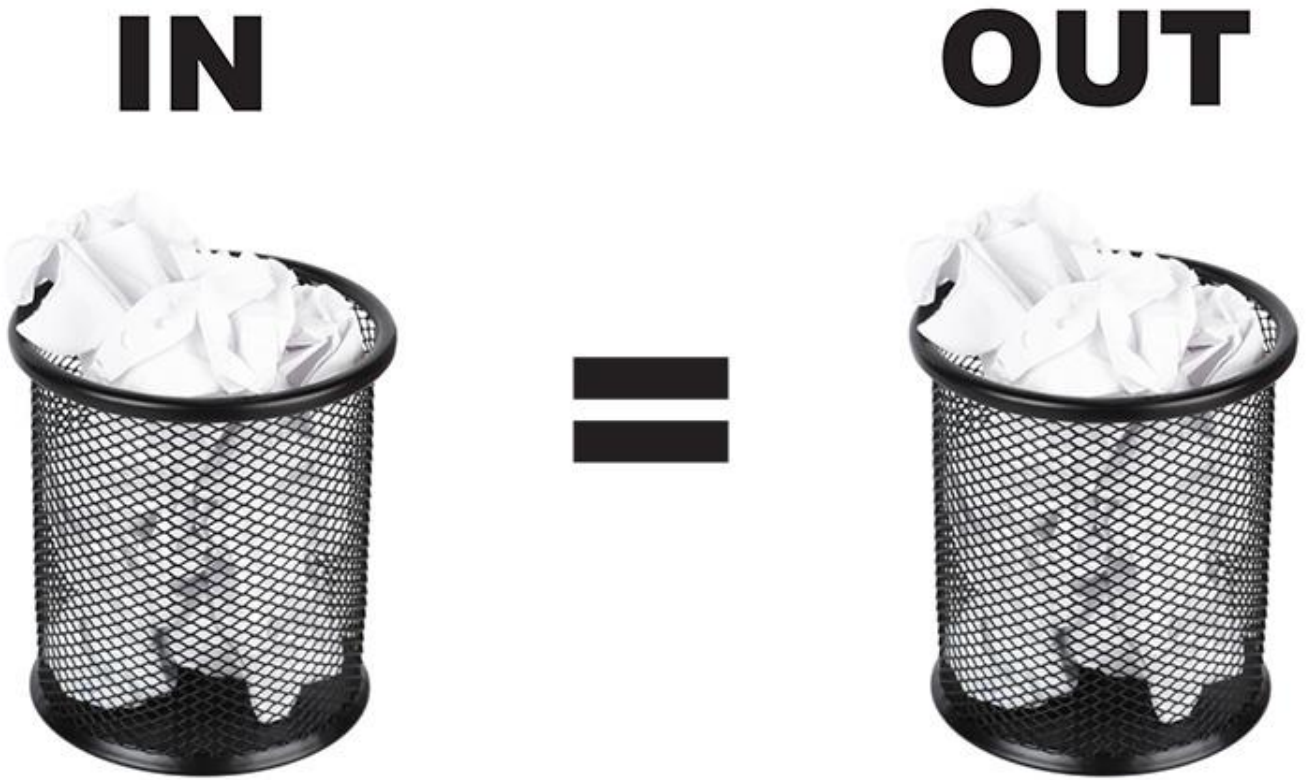
G. Yang et al., “Fake Co-visitation Injection Attacks to Recommender Systems,” in Proc. of NDSS, 2017.

“If you like 

, you MUST love…”

G. Yang et al., “Fake Co-visitation Injection Attacks to Recommender Systems,” in Proc. of NDSS, 2017.

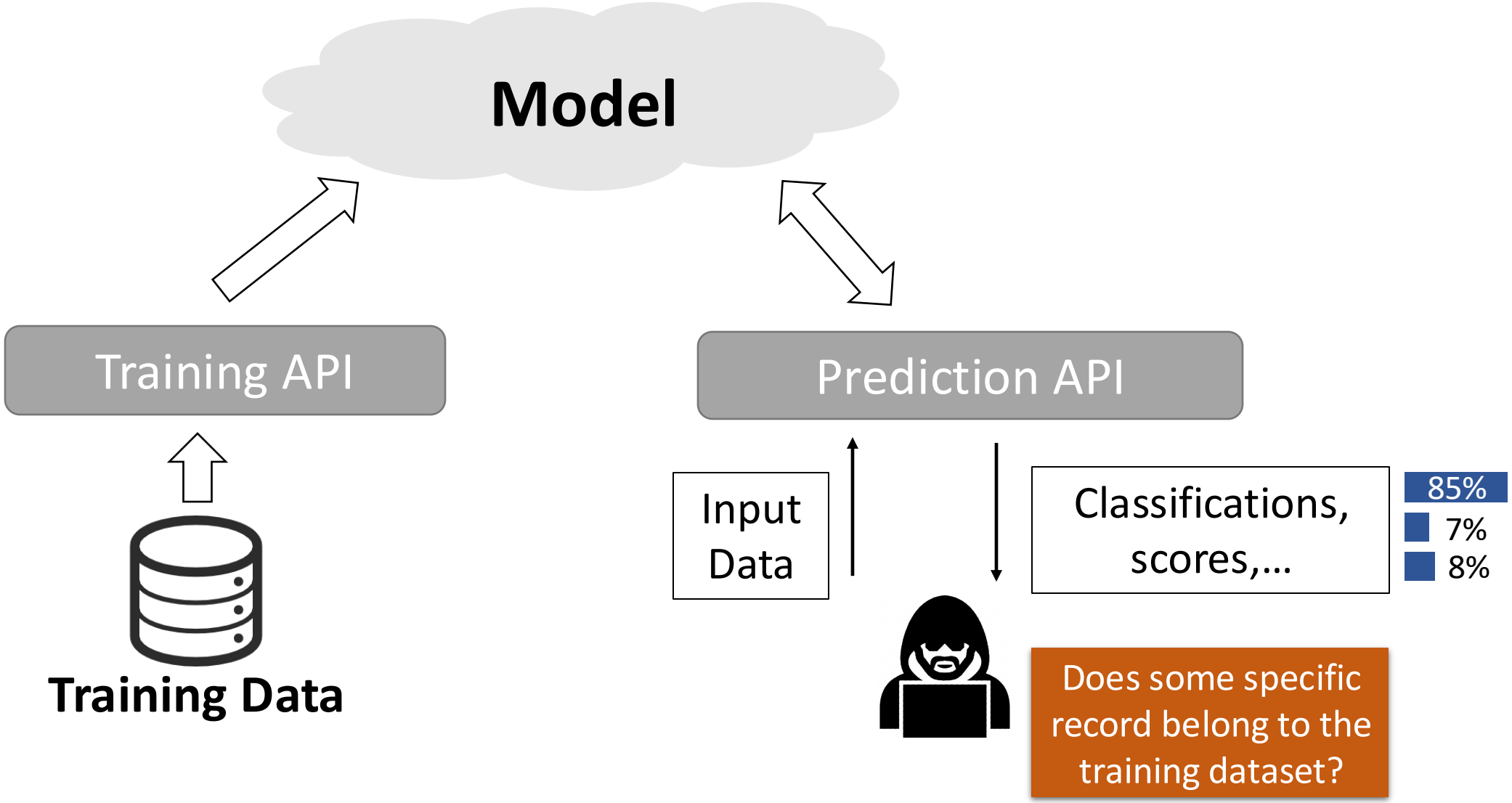
G. Yang et al., “Fake Co-visitation Injection Attacks to Recommender Systems,” in Proc. of NDSS, 2017.



G. Yang et al., “Fake Co-visitation Injection Attacks to Recommender Systems,” in Proc. of NDSS, 2017.

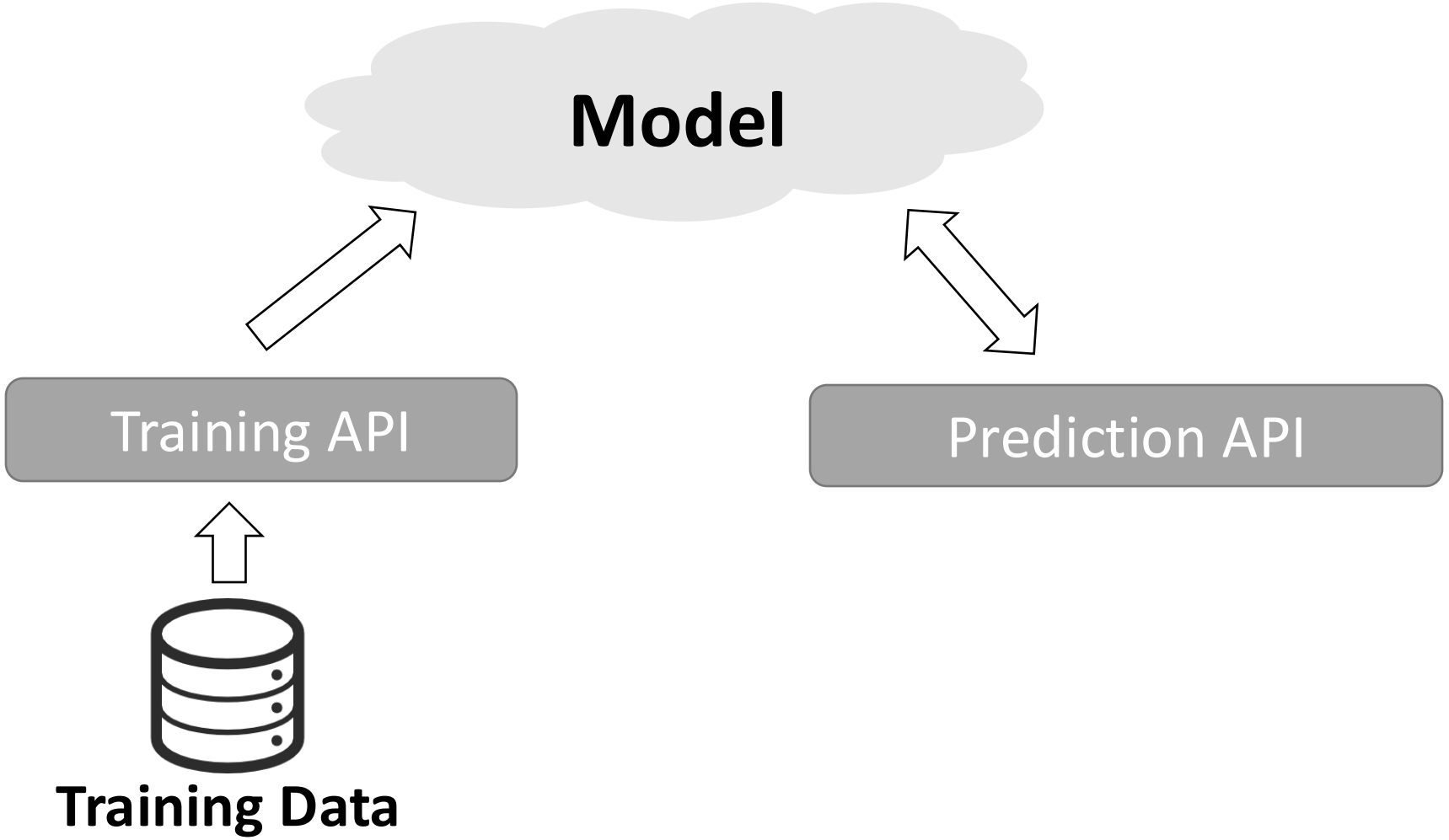
Cracking the Black-box

Membership Inference Attack



R. Shokri et al., "Membership Inference Attack Against Machine Learning Models," in Proc. of S&P, 2017

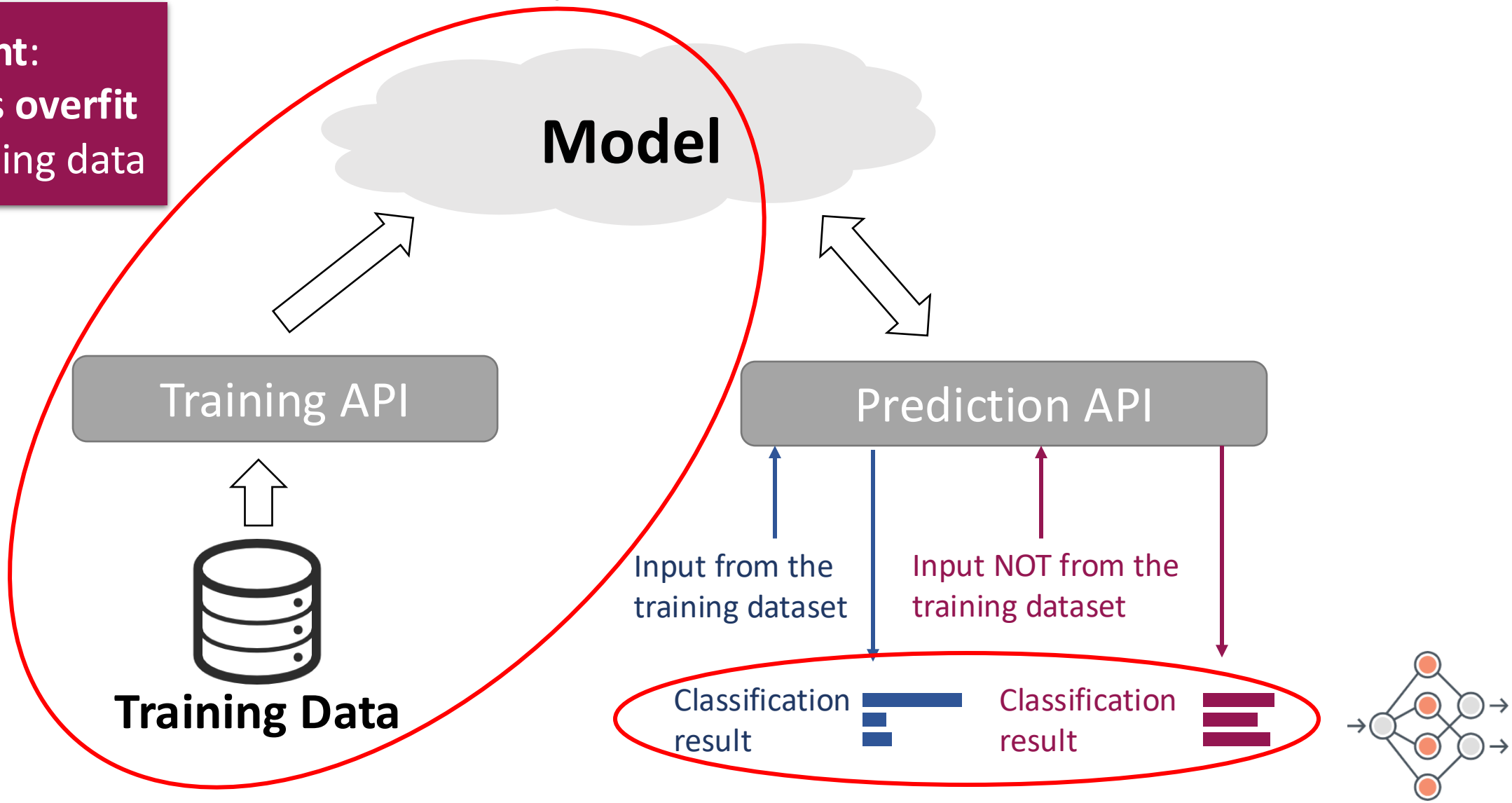
Membership Inference Attack



R. Shokri et al., “Membership Inference Attack Against Machine Learning Models,” in Proc. of S&P, 2017

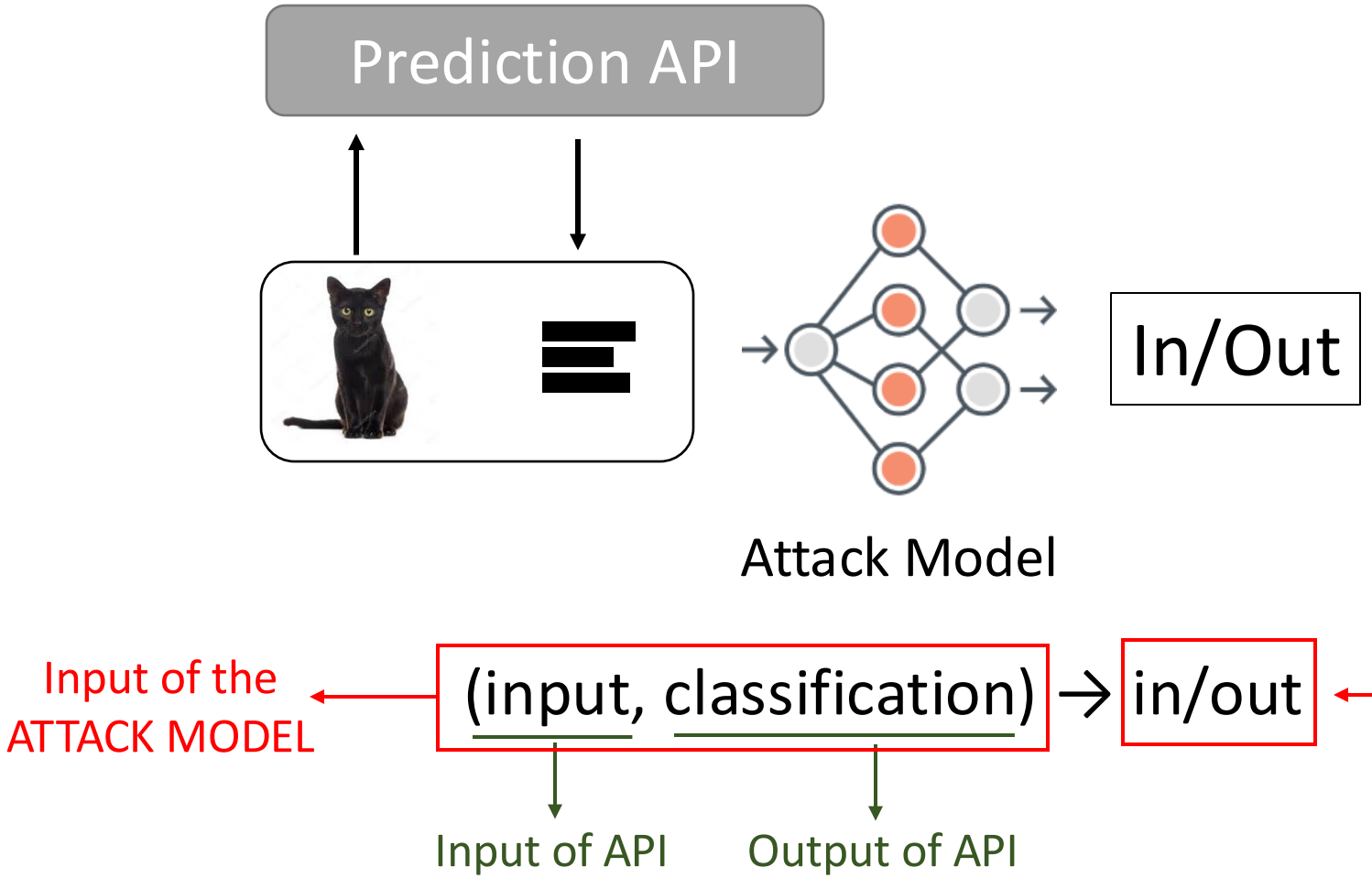
Membership Inference Attack

Insight:
ML models **overfit** to the training data



R. Shokri et al., "Membership Inference Attack Against Machine Learning Models," in Proc. of S&P, 2017

MIA: Two-class Classification Problem

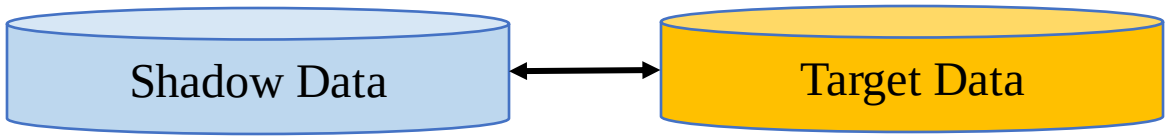


R. Shokri et al., "Membership Inference Attack Against Machine Learning Models," in Proc. of S&P, 2017

MIA: Shadow-Model-Based Approach

1. Shadow data mimicking the target training data

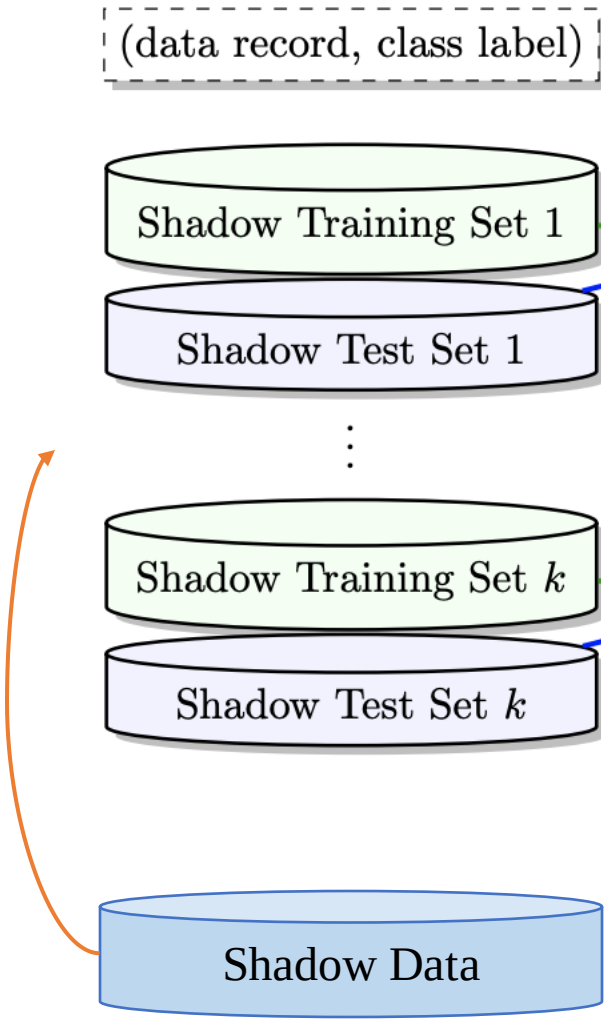
Similar *Pr(x)*



R. Shokri et al., "Membership Inference Attack Against Machine Learning Models," in Proc. of S&P, 2017

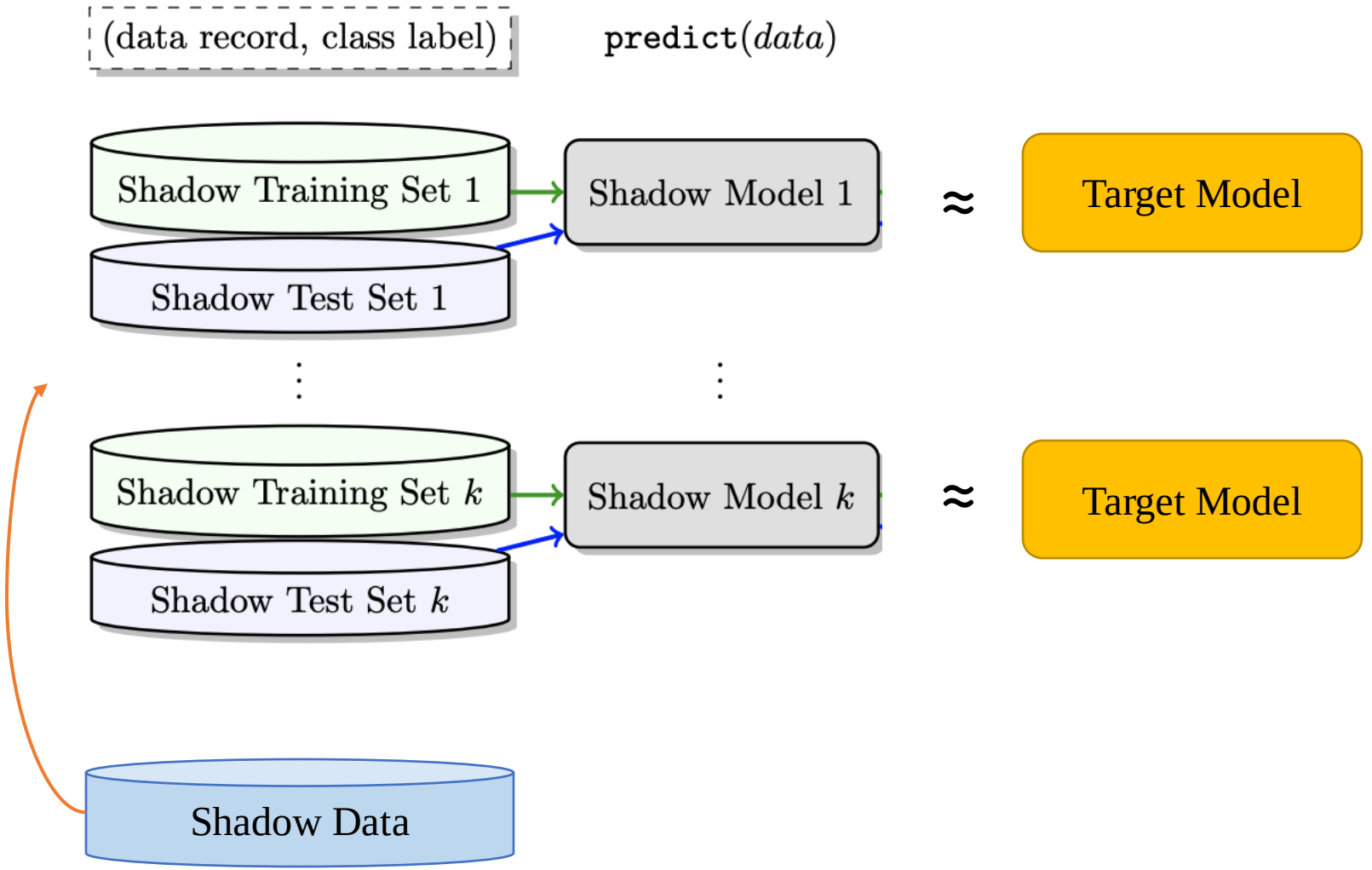
MIA: Shadow-Model-Based Approach

1. Shadow data mimicking the target training data



R. Shokri et al., "Membership Inference Attack Against Machine Learning Models," in Proc. of S&P, 2017

MIA: Shadow-Model-Based Approach

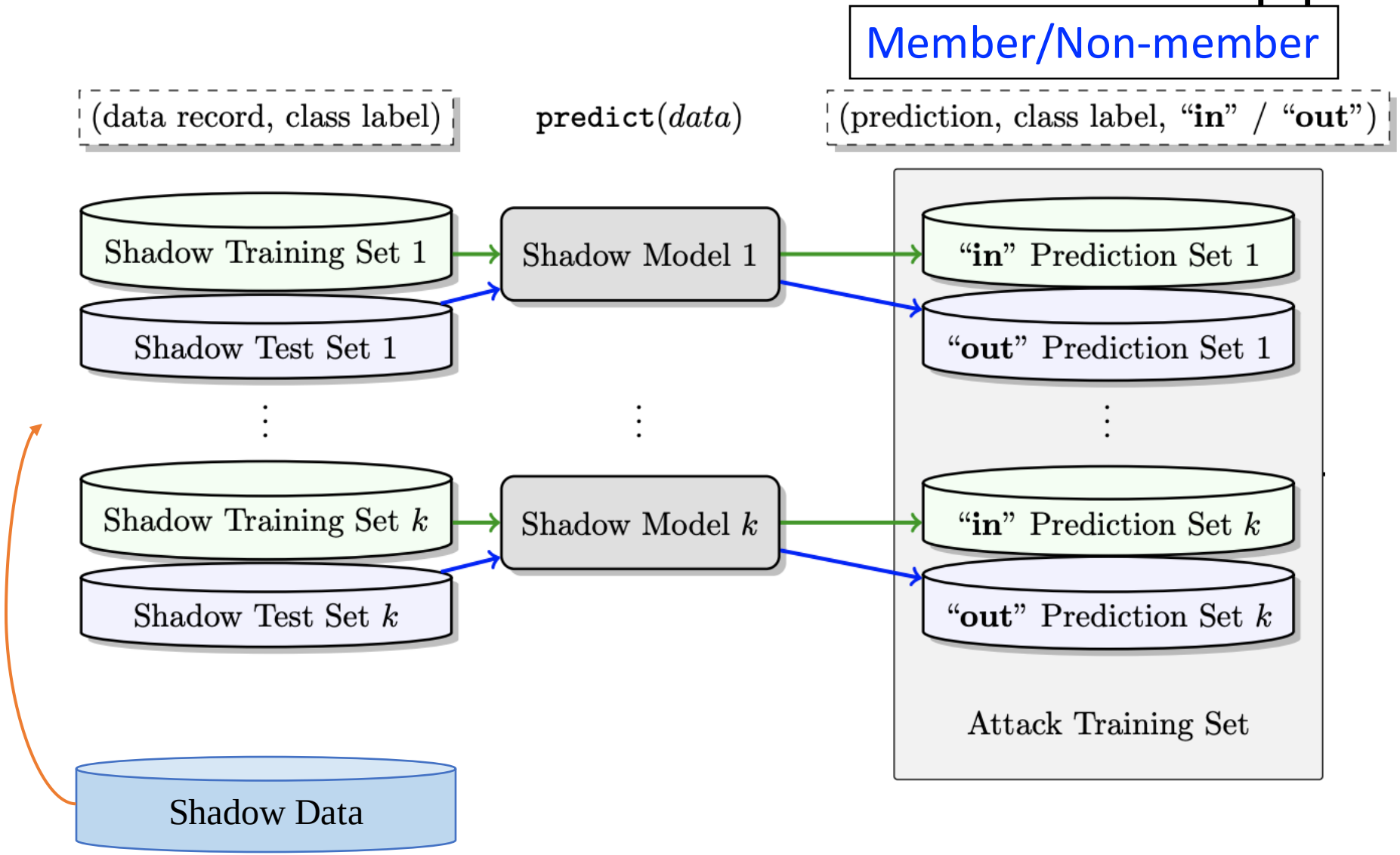


1. Shadow data mimicking the target training data

1. Shadow models mimicking the target model

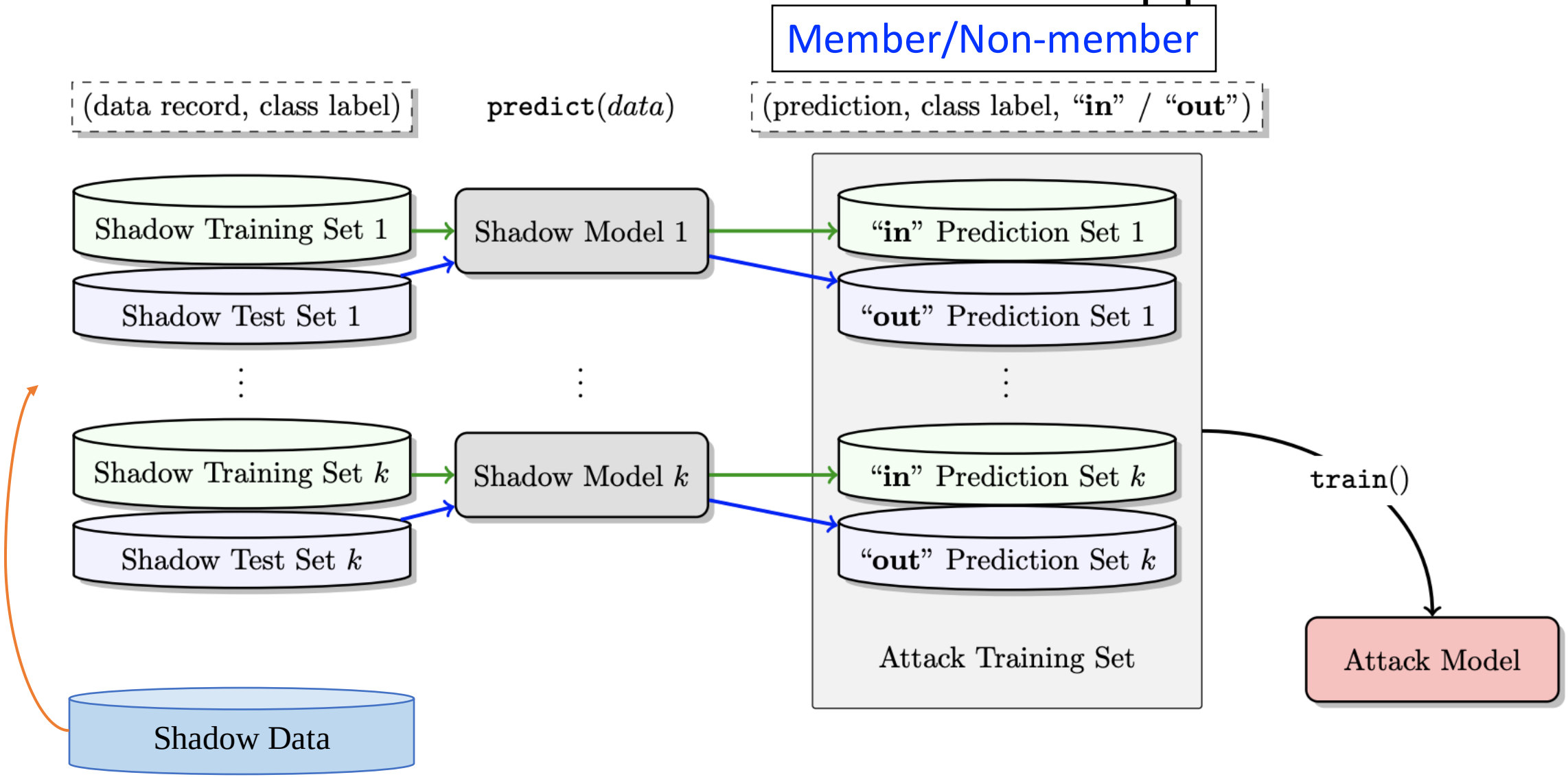
R. Shokri et al., "Membership Inference Attack Against Machine Learning Models," in Proc. of S&P, 2017

MIA: Shadow-Model-Based Approach



R. Shokri et al., "Membership Inference Attack Against Machine Learning Models," in Proc. of S&P, 2017

MIA: Shadow-Model-Based Approach



R. Shokri et al., "Membership Inference Attack Against Machine Learning Models," in Proc. of S&P, 2017

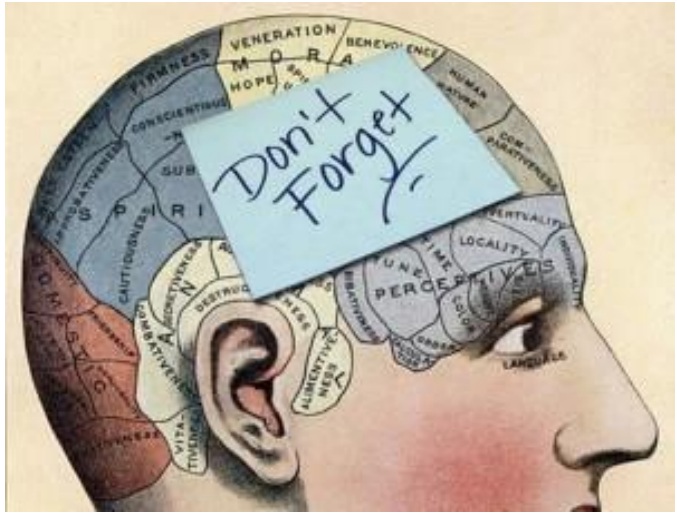
MIA: Shadow-Model-Based Approach

Dataset	*Training Accuracy*	*Testing Accuracy*	*Attack Precision*
Adult	0.848	0.842	0.503
MNIST	0.984	0.928	0.517
Location	1.000	0.673	0.678
Purchase (2)	0.999	0.984	0.505
Purchase (10)	0.999	0.866	0.550
Purchase (20)	1.000	0.781	0.590
Purchase (50)	1.000	0.693	0.860
Purchase (100)	0.999	0.659	0.935
TX hospital stays	0.668	0.517	0.657

Severe overfitting

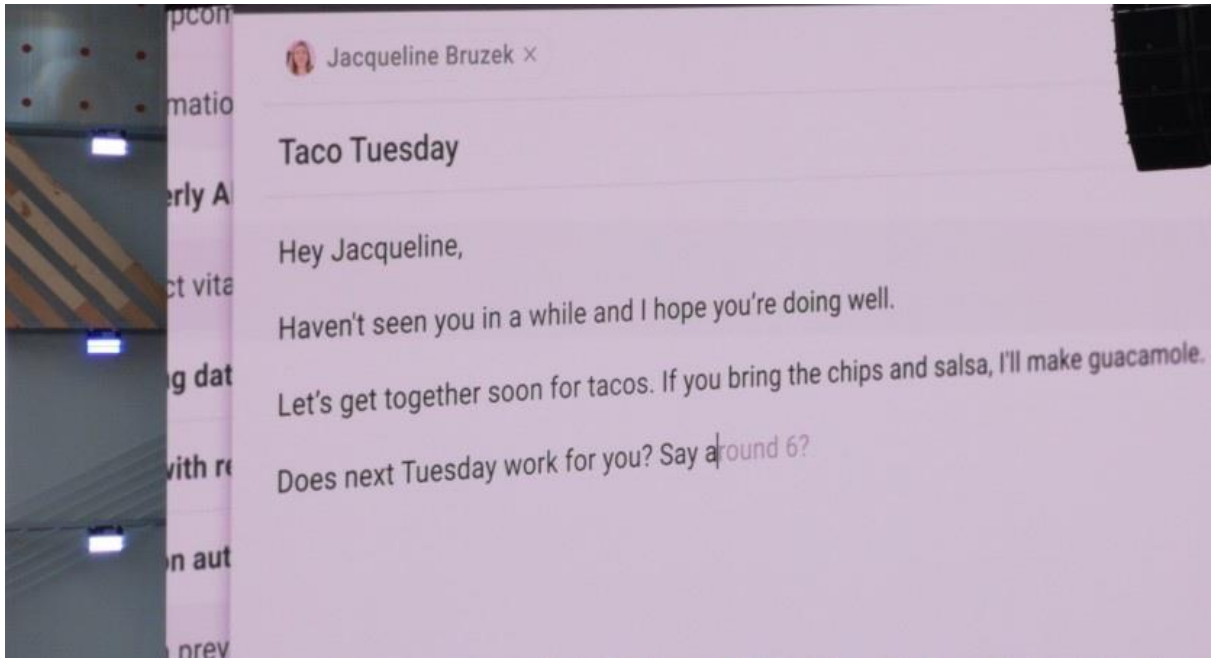
TABLE II: Accuracy of the Google-trained models and the corresponding attack precision.

R. Shokri et al., "Membership Inference Attack Against Machine Learning Models," in Proc. of S&P, 2017



Unintended Memorization

- Your sensitive information may get memorized by the model in unexpected ways…



Smart Compose in Gmail

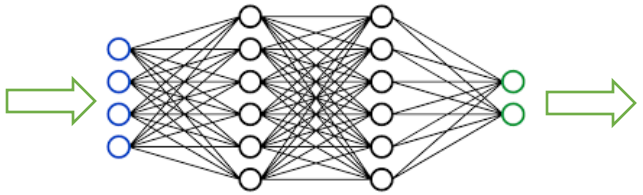
WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

https://events.google.com/io2018/recap/

https://xkcd.com/2169/

- It is possible to complete the sensitive information.

“Alice’s HKID number is: 1◌◌◌◌◌◌.”



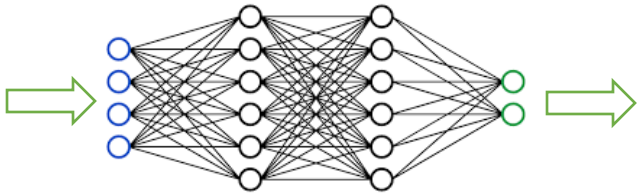
Prob.

No.	Prob
1	0.33
2	0.09
3	0.03
…	…
9	0.11

N. Carlini et al., “The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks,” in Proc. of USENIX Security, 2019

- It is possible to complete the sensitive information.

“Alice’s HKID number is: 1 6 ◌◌◌◌◌.”



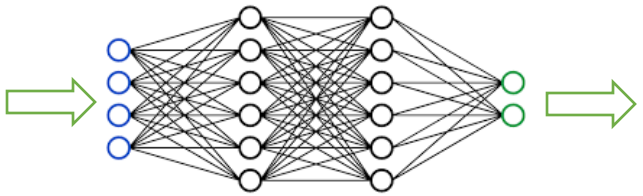
Prob.

No.	Prob
1	0.05
…	0
6	0.42
…	…
9	0.03

N. Carlini et al., “The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks,” in Proc. of USENIX Security, 2019

- It is possible to complete the sensitive information.

“Alice’s HKID number is: 1 6 6 5 8 7 .”



Prob.

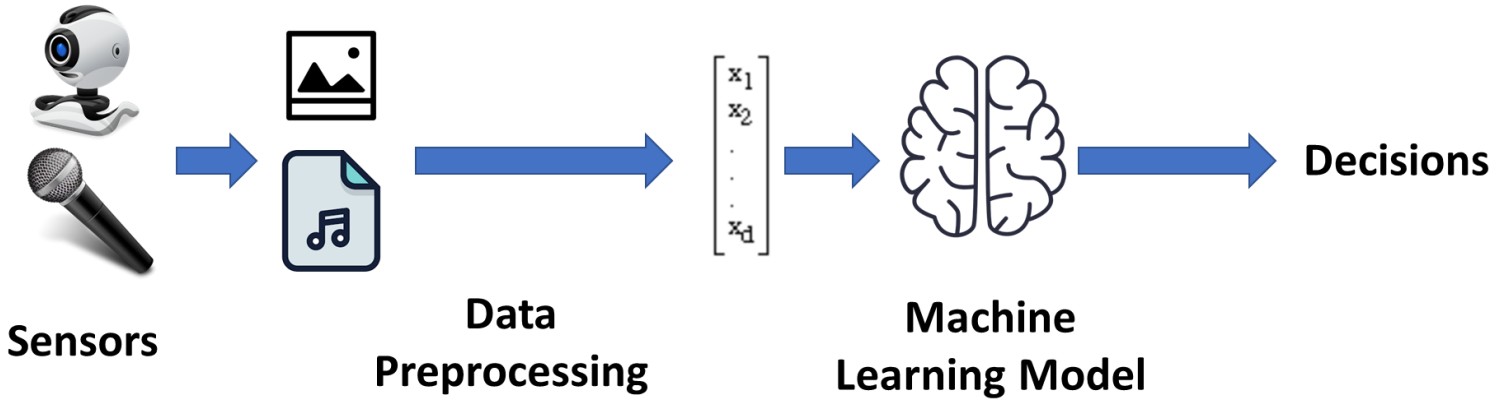
N. Carlini et al., “The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks,” in Proc. of USENIX Security, 2019

How much should you trust a machine learning system?

People with no idea about AI
saying it will take over the world:

Conclusion

Machine Learning Community：How to build it?
(Accuracy、Efficiency)



Security Community：How to crack it?
(Adversarial examples, data leakage, sensor spoofing…)

Readings

Adversarial Examples:

1. Nicholas Carlini and David Wagner, "Towards Evaluating the Robustness of Neural Networks," in Proc. of S&P, 2017. Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter, "Accessorize to a Crime: Real and Stealthy Attacks on State-Of-The-Art Face Recognition," in Proc. of CCS, 2016.
2. Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana, "Certified Robustness to Adversarial Examples with Differential Privacy," in Proc. S&P, 2019.
3. Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu, "DolphinAttack: Inaudible Voice Commands," in Proc. of CCS, 2017.
4. Weilin Xu, Yanjun Qi, and David Evans, "Automatically Evading Classifiers: A Case Study on PDF Malware Classifiers," in Proc. of NDSS, 2016.

Data Poisoning:

1. Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet, "Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring," in Proc. of USENIX Security, 2018.
2. Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang, "Trojaning Attack on Neural Networks," in Proc. of NDSS, 2018.
3. Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao, "Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks," in Proc. S&P, 2019.

Readings

ML Privacy:

1. Milad Nasr, Reza Shokri, and Amir Houmansadr, "Comprehensive Privacy Analysis of Deep Learning," in Proc. S&P, 2019.
2. Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong, "MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples," in Proc. of CCS, 2019.
3. Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song, "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks," in Proc. of USENIX Security, 2019.
4. Karan Ganju, Qi Wang, Wei Yang, Carl Gunter, and Nikita Borisov, "Property Inference Attacks on Deep Neural Networks using Permutation Invariant Representations," in Proc. of CCS, 2018.
5. Binghui Wang and Neil Zhenqiang Gong, "Stealing Hyperparameters in Machine Learning," in Proc. of S&P, 2018.
6. Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov, "Membership Inference Attacks against Machine Learning Models," in Proc. of S&P, 2017.
7. Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart, "Stealing Machine Learning Models via Prediction APIs," in Proc. of USENIX Security, 2016.

Privacy-Preserving ML:

1. Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y. Zhao, "Fawkes: Protecting Privacy against Unauthorized Deep Learning Models," in Proc. of USENIX Security, 2020.
2. Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang, "Deep Learning with Differential Privacy," in Proc. of CCS, 2016.

Readings

ML Misuse:

1. Roei Schuster, Vitaly Shmatikov, and Eran Tromer, "Beauty and the Burst: Remote Identification of Encrypted Video Streams," in Proc. of USENIX Security, 2017.
2. William Melicher, Blase Ur, Sean M. Segreti, Saranga Komanduri, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor, "Fast, Lean, and Accurate: Modeling Password Guessability Using Neural Networks," in Proc. of USENIX Security, 2016.

Other Topics:

1. Sanghyun Hong, Pietro Frigo, Yiğitcan Kaya, Cristiano Giuffrida, and Tudor Dumitraș, "Terminal Brain Damage: Exposing the Graceless Degradation in Deep Neural Networks Under Hardware Fault Attacks," in Proc. of USENIX Security, 2019.
2. Wenbo Guo, Dongliang Mu, Jun Xu, Purui Su, Gang Wang, and Xinyu Xing, "LEMNA: Explaining Deep Learning based Security Applications," in Proc. of CCS, 2018.