- https://cdn-images-1.medium.com/max/1200/1*GZe7IJaseja8sB-aU0w6Bg.gif

ADVERSARIAL MACHINE LEARNING

"SOME RULES CAN BE BENT, OTHERS CAN BE BROKEN"

LEGAL NOTICES AND DISCLAIMERS

This presentation contains the general insights and opinions of Intel Corporation ("Intel"). The information in this presentation is provided for information only and is not to be relied upon for any other purpose than educational. Use at your own risk! Intel makes no representations or warranties regarding the accuracy or completeness of the information in this presentation. Intel accepts no duty to update this presentation based on more current information. Intel is not liable for any damages, direct or indirect, consequential or otherwise, that may arise, directly or indirectly, from the use or misuse of the information in this presentation.
Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.
No computer system can be absolutely secure.
No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.
Intel, Intel Inside, the Intel Core, and the Intel logo are trademarks of Intel Corporation in the United States and other countries.
*Other names and brands may be claimed as the property of others.

“HAND WAIVING” DISCLAIMER

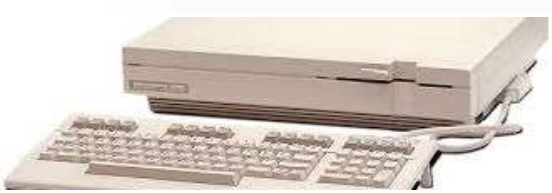
- I will do some hand waiving explanation
  - Sorry ☹
- If you want a deeper dive, look at the references slide
- Mostly citing academic papers
  - Really wanted to do a demo ☹
- For more info, come talk to me

- https://camo.derpicdn.net/e08caaabde22bda0e94ac778cd08cad0d1b0ed94?url=http%3A%2F%2Fstream1.gifsoup.com%2Fview7%2F2738731%2Fjedi-mind-trick-o.gif

WHO AM I?

- Guy Barnhart-Magen
- **Security Researcher**, Manager, Presenter
- Interests:

  Crypto, Embedded systems, System and product security
- iSTARE team
  - Intel Security Threat Analysis and Reverse Engineering
  - Leading the "AI Security Innovations" team
- "We break what we make"



We Are Hiring!

- https://i.kinja-img.com/gawker-media/image/upload/t_original/gaflyotna4s2ddvb8f0r.gif

WHAT IS ARTIFICIAL INTELLIGENCE?

Perform intellectual tasks as humans can.

In general it should be able to:

- Learn
- Represent knowledge
- Plan
- Make decisions under uncertainty
- Communicate in a natural language
- Use these skills towards common goal(s) to be AI-complete

Basic model: input{code, data}, algorithm, output{classification, probability}

Most of the AI systems in place today are Weak Artificial Intelligence, **which were designed to solve a specific problem**.

MACHINE LEARNING TYPES

Supervised Learning (Input and Output is specified for training),

Unsupervised Learning (Only input is given to recognize patterns) and

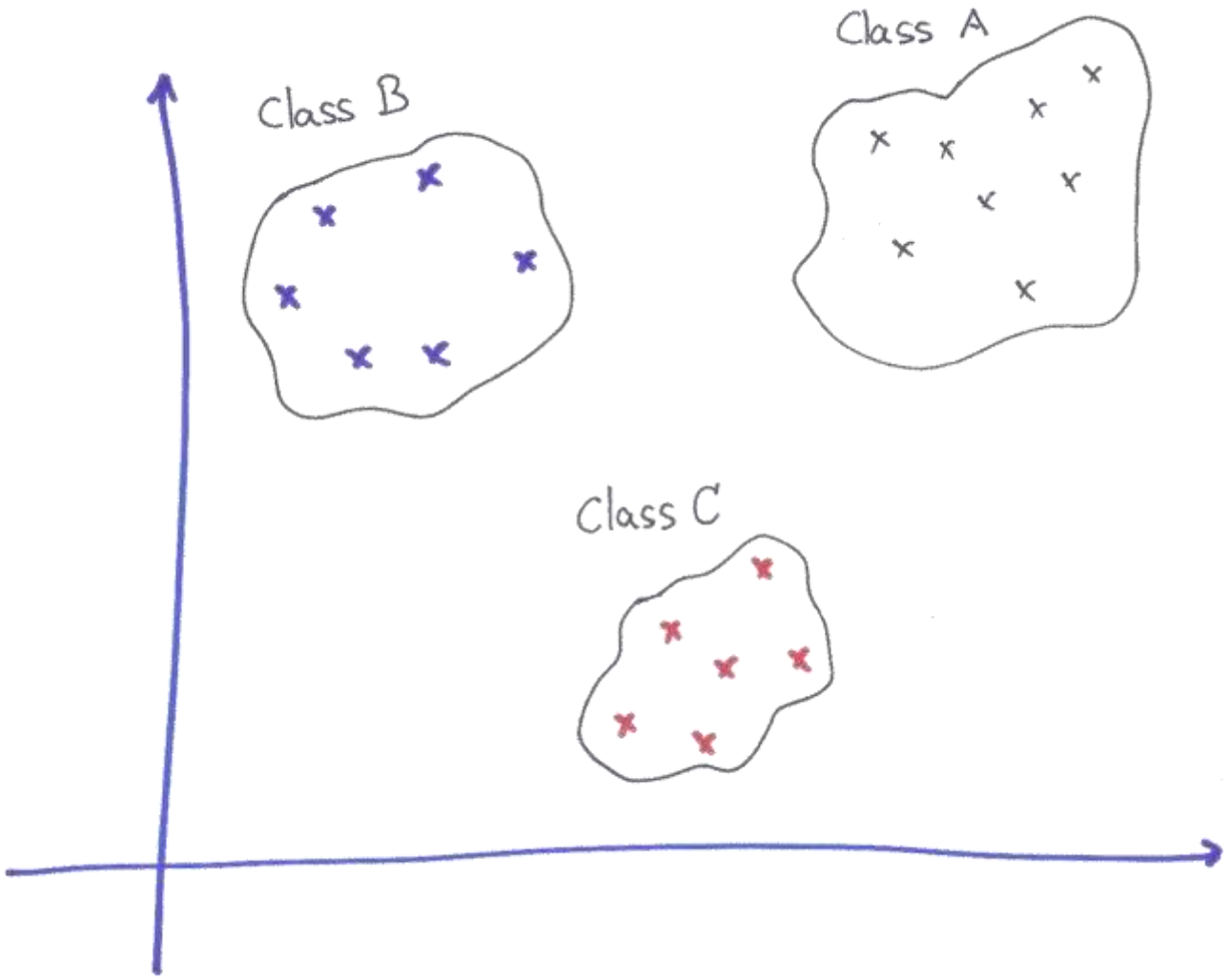
Reinforcement learning (Real world feed back is provided to system on the go).

- https://8.smash.com/u/2016/03/The-Matrix.gif

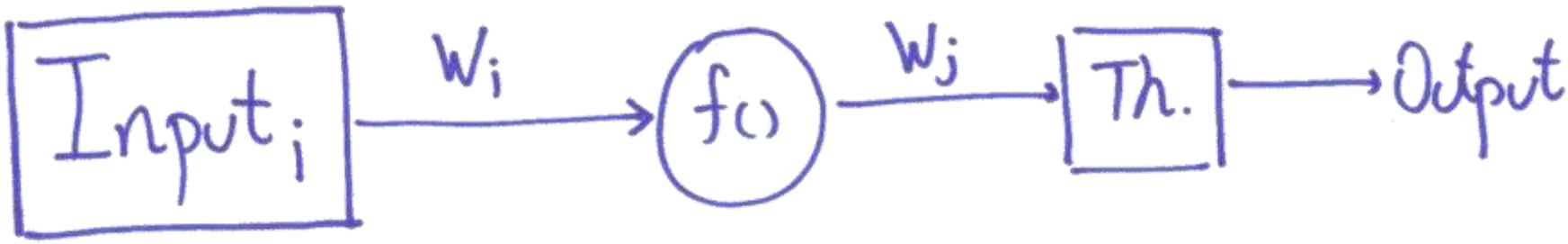
- https://i.imgflip.com/jj0ii.jpg

MACHINE LEARNING 101

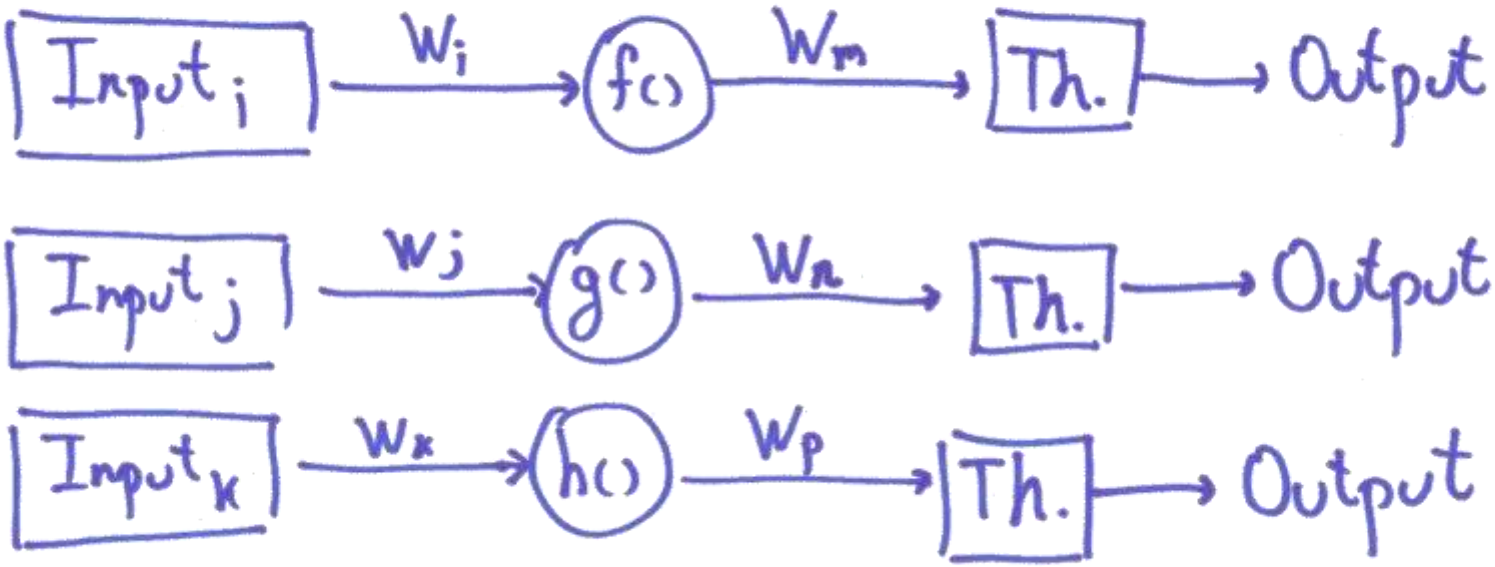
CURVES SEPARATE CLASSES



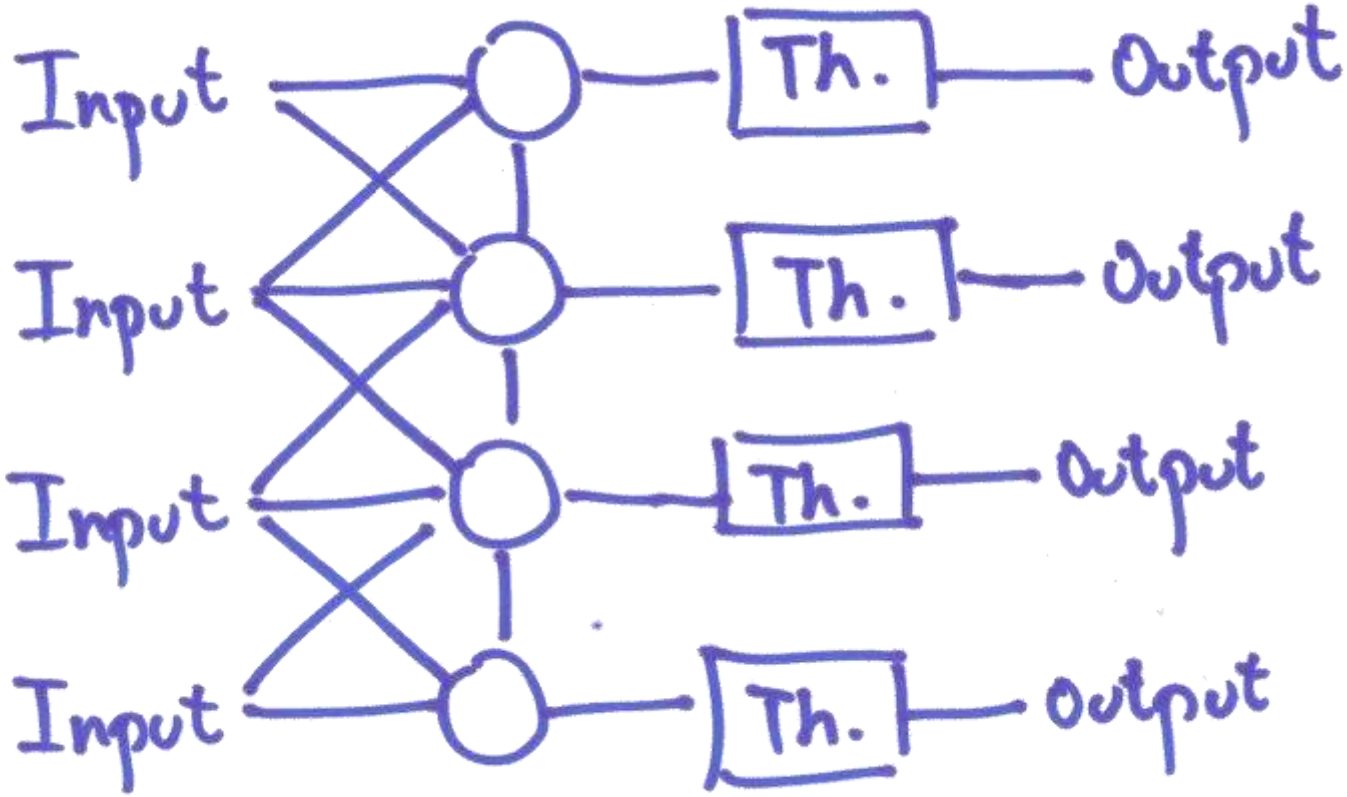
FIND CURVE PARAMETERS



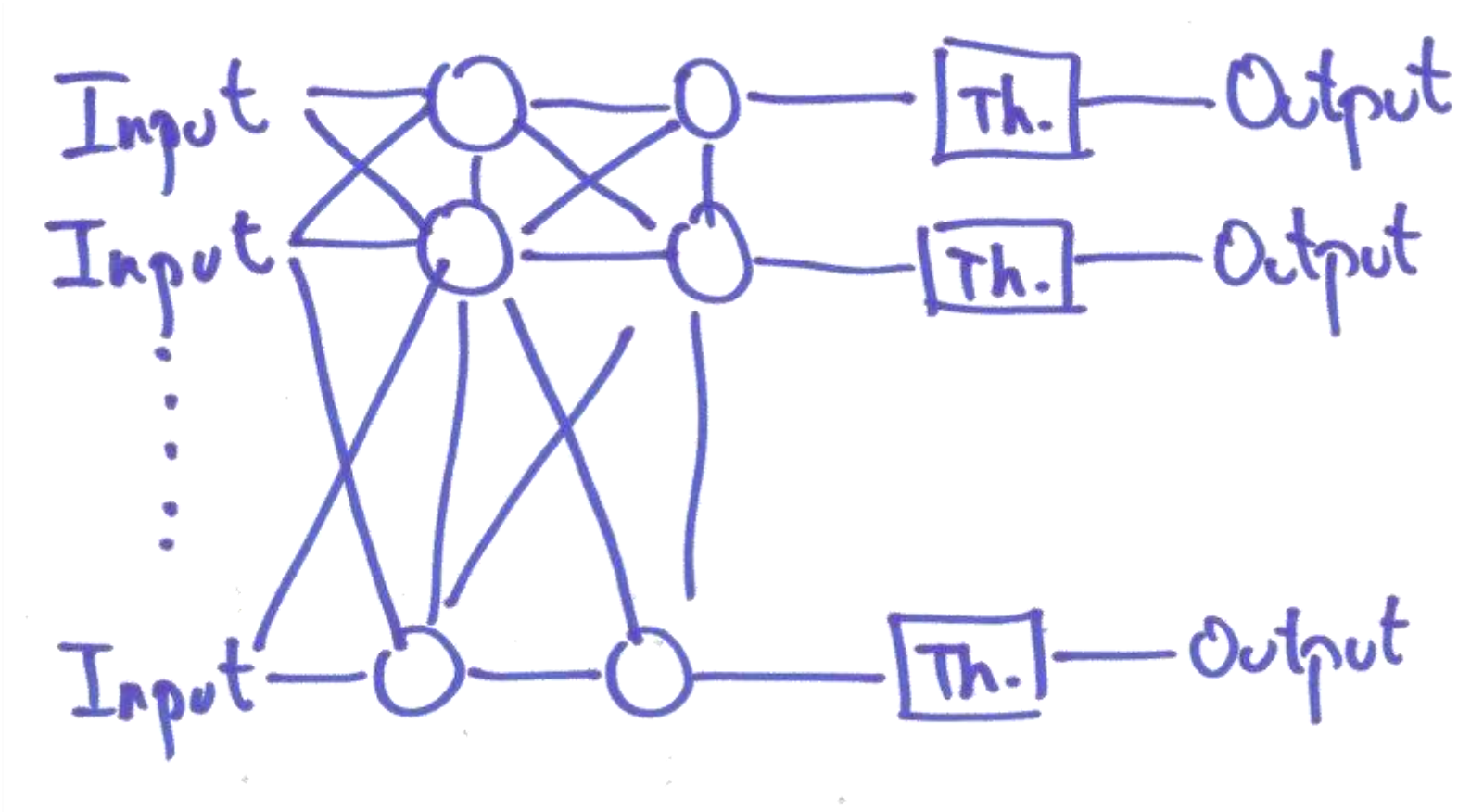
MULTIPLE INPUTS ARE ENCOURAGED ☺



INNER CONNECTIVITY IS GOOD!



ARCHITECTURE IS THE LAYOUT



ADVANCED TOPICS AHEAD

- Neural networks (NN) with memory
- NN with cross layer connectivity
- NN with multiple hidden layers
- Fully/Semi connected layers
- Deep Learning – NN made out of NN (think inception)
- Many more options…

- http://thoughtmedicine.com/wp-content/uploads/2010/07/inception.jpg

- https://upload.wikimedia.org/wikipedia/commons/e/e3/CleverHans.jpg

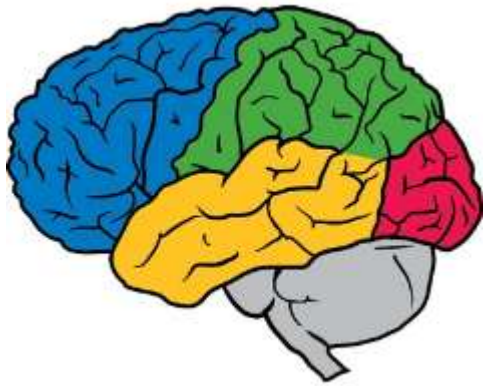
CLEVER HANS

https://github.com/tensorflow/cleverhans

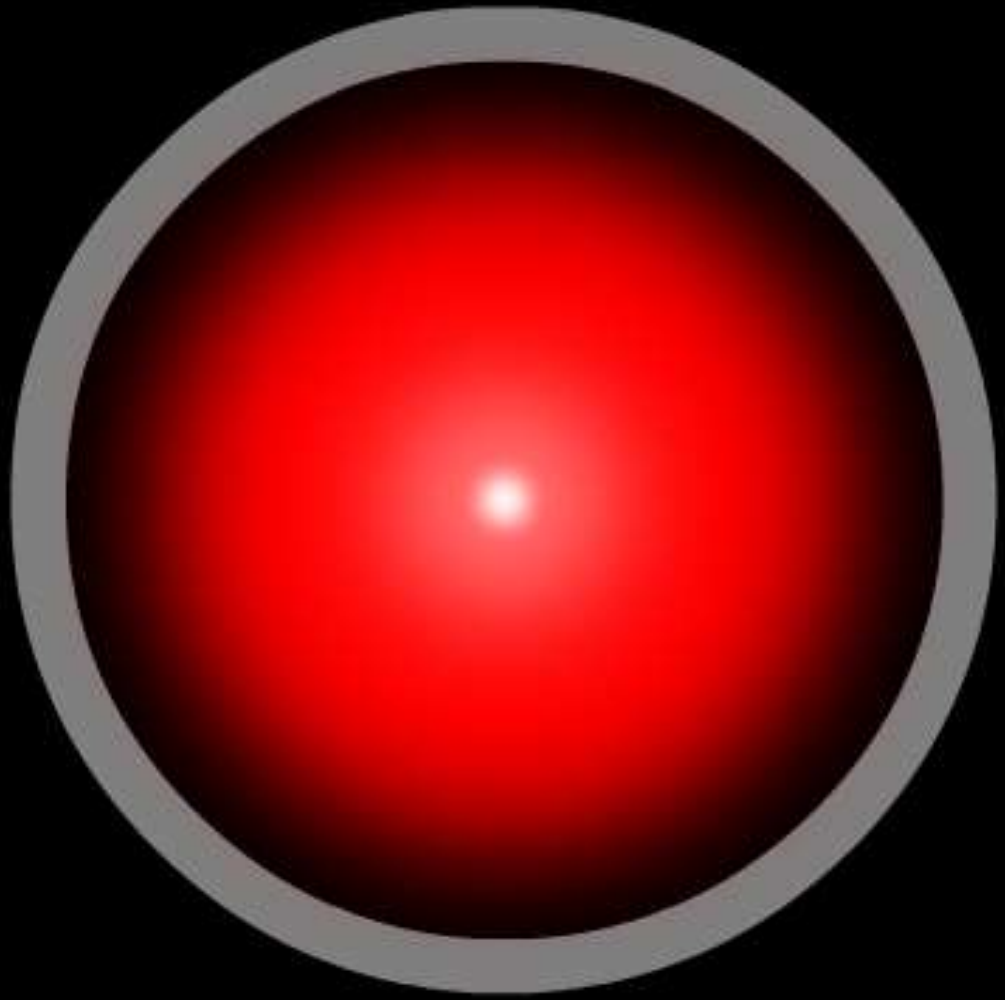
“We have reached the point where machine learning works, but may easily be broken”

Nicolas Papernot, Google PhD Fellow in Security

Ian Goodfellow, Research scientist at Google Brain



https://pbs.twimg.com/profile_images/799327801388077057/HcDnA1H7_400x400.jpg



HELLO DAVE

- http://s3.amazonaws.com/stripgenerator/strip/03/52/76/00/00/full.png

To break a machine learning model, an attacker can compromise its:

- **Confidentiality**

  Think of it as privacy. If attackers can gain the data the model was trained on – a lot of information is exposed

- **Integrity**

  Alter predictions from intended ones (very possible in reality)

  Some examples in the slides ahead

- **Availability**

  What if we can take the algorithm offline/DoS?

  e.g. a malicious road sign causes the autonomous driving AI to crash

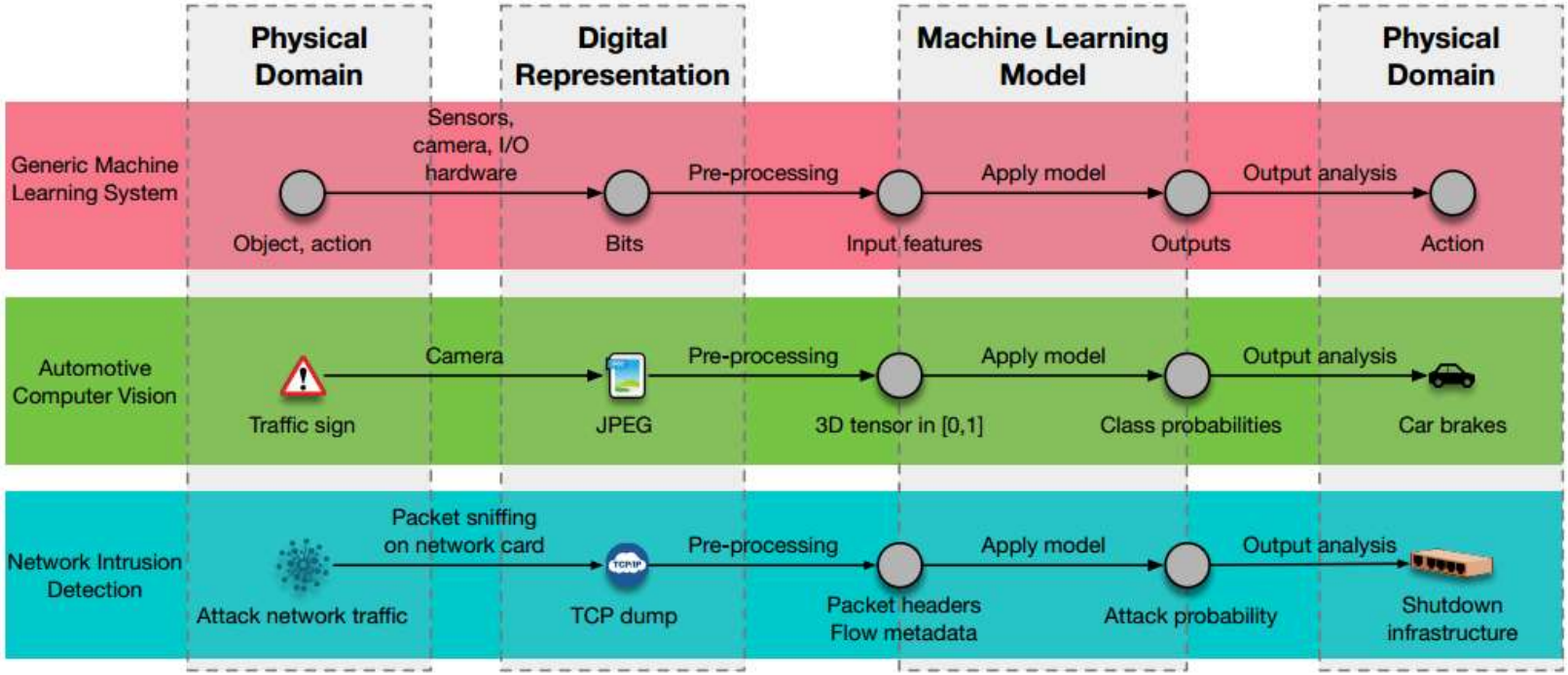
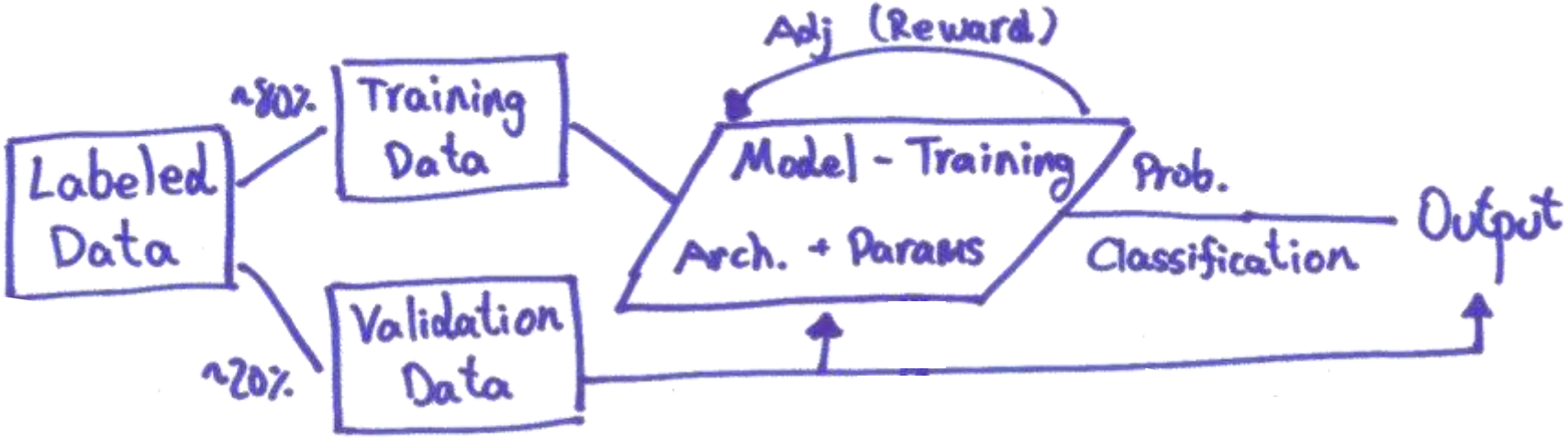


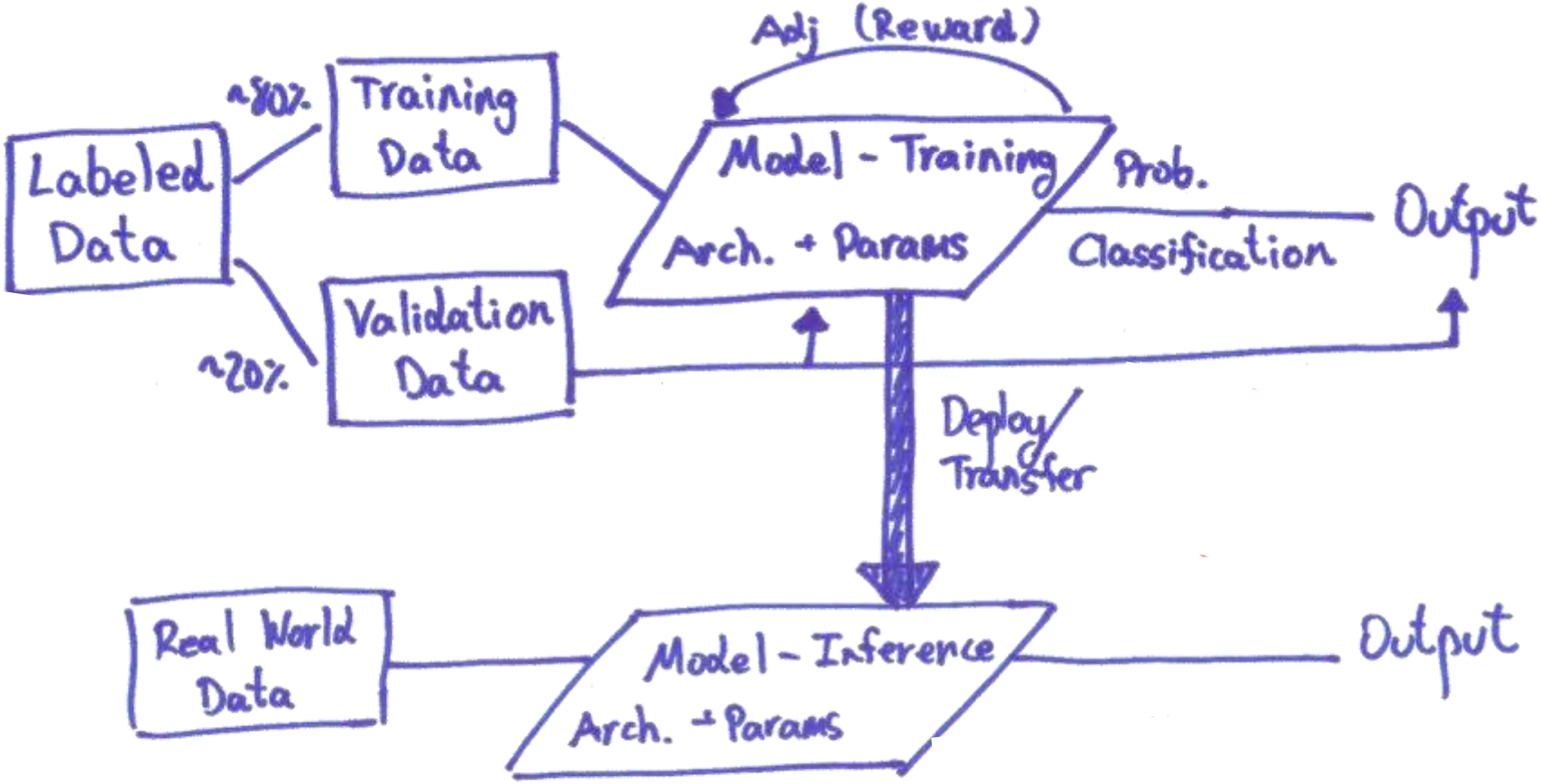
Fig. 1. **System's attack surface:** the generic model (top row) is illustrated with two example scenarios (bottom rows): a computer vision model used by an automotive system to recognize traffic signs on the road and a network intrusion detection system.

- Towards the Science of Security and Privacy in Machine Learning

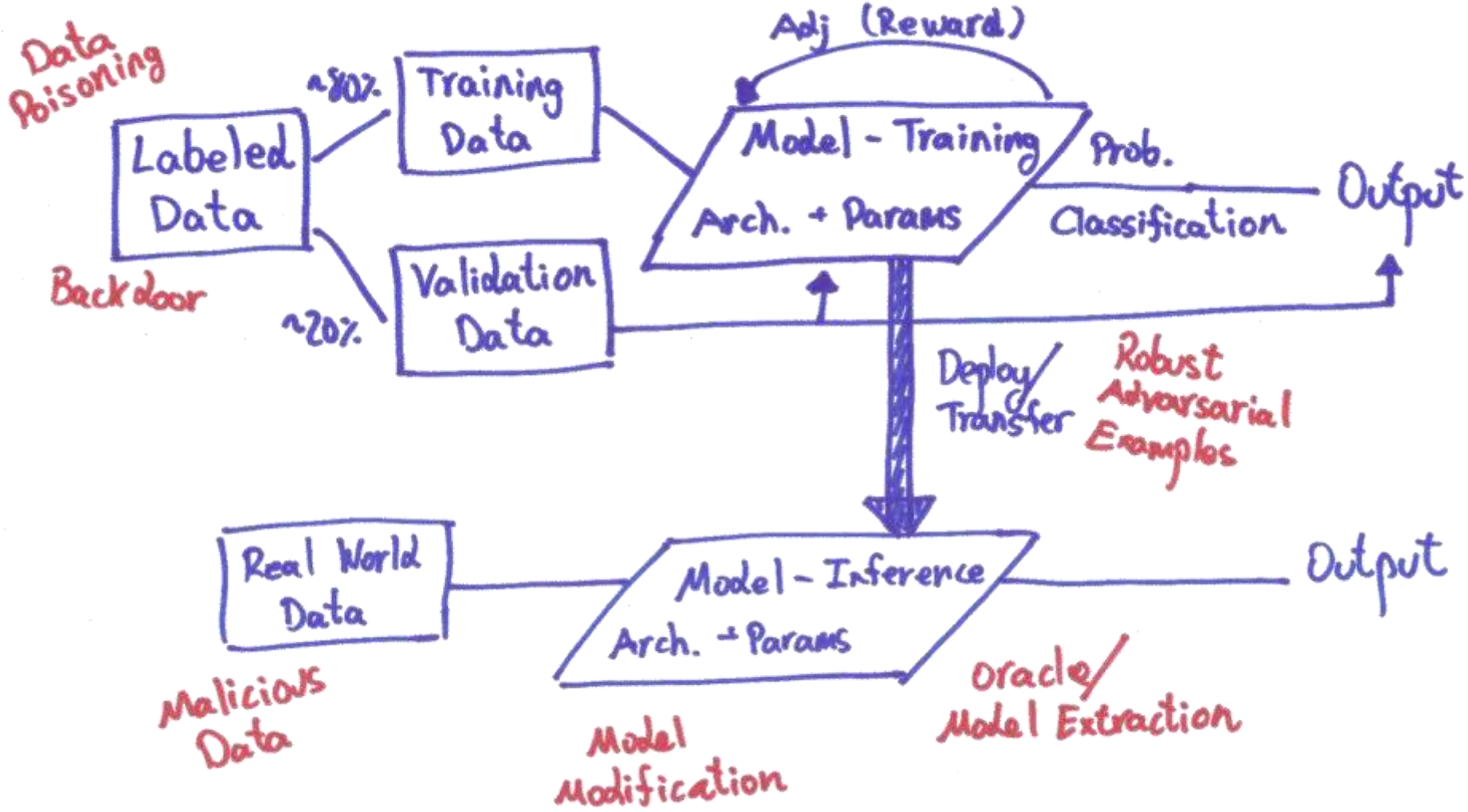
THREAT MODELING AI?



THREAT MODELING AI?



THREAT MODELING AI?



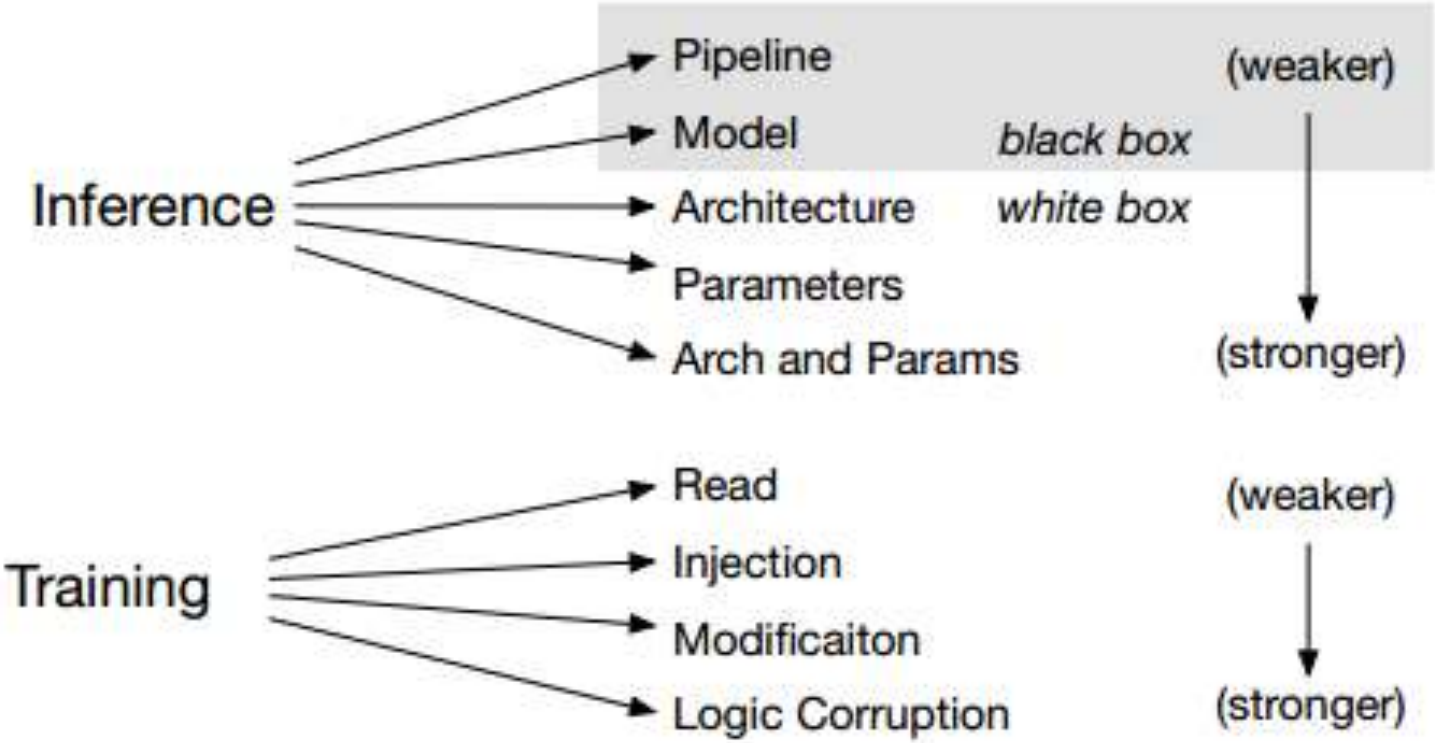


Fig. 2. **Adversarial Capabilities:** adversaries attack ML systems at inference time by exploiting model internal information (white box) or probing the system to infer system vulnerabilities (black box). Adversaries use read or write access to the training data to mimic or corrupt the model.

- Towards the Science of Security and Privacy in Machine Learning

THE DEVIL LIES IN THE DETAILS…

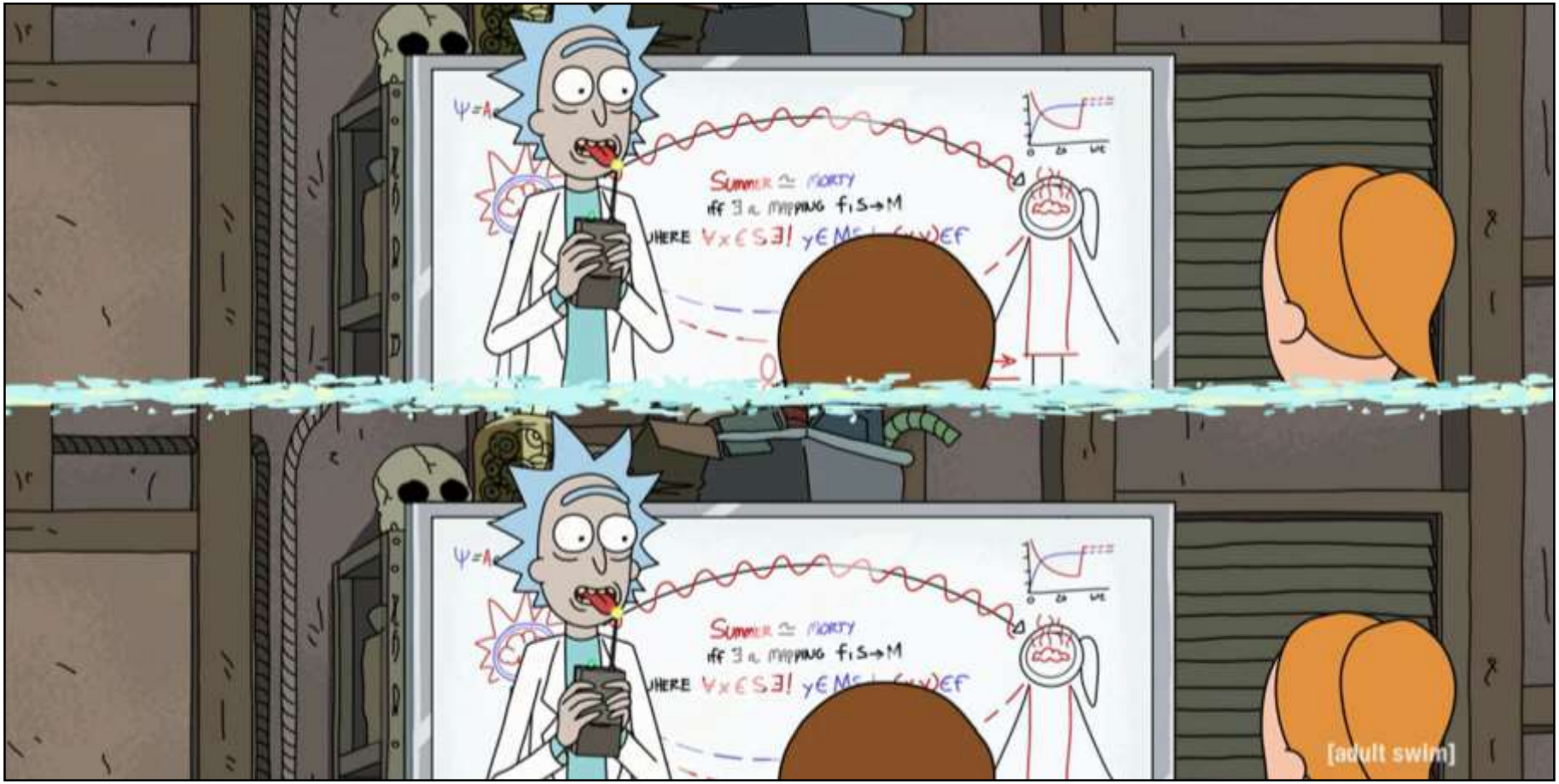
- The “curves” fit more points than what you planned for
  - Gray area
  - Many points that lead to the same output
- There is a lot of noise
  - Hiding is easy
- Backdoors are almost impossible to detect

- https://pbs.twimg.com/profile_images/799327801388077057/HcDnA1H7_400x400.jpg

HIDING IN UNTRAINED DATA

- Almost all training is done with positive data points
- Even if negative data points are used, they are a small set of possible examples
- Negative data points > Positive data points

- https://s-media-cache-ak0.pinimg.com/originals/42/09/6b/42096bc4ad49bfb70feb556cece77f1d.jpg



- http://i.imgur.com/0srqDzj.jpg

SOME EXAMPLES MAYBE?

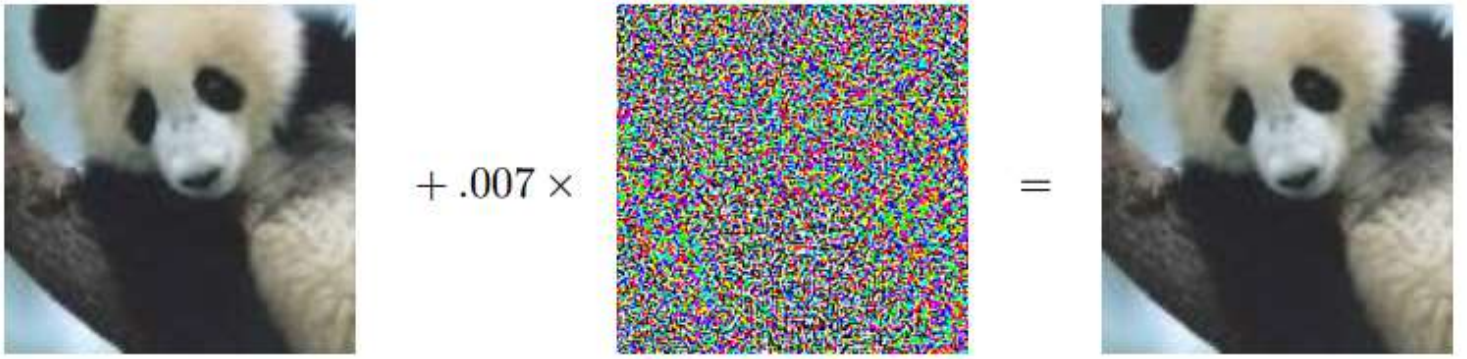


Figure 1: A demonstration of fast adversarial example generation applied to GoogLeNet (Szegedy et al., 2014a) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can change GoogLeNet's classification of the image. Here our ϵ of .007 corresponds to the magnitude of the smallest bit of an 8 bit image encoding after GoogLeNet's conversion to real numbers.

- Explaining and harnessing adversarial examples

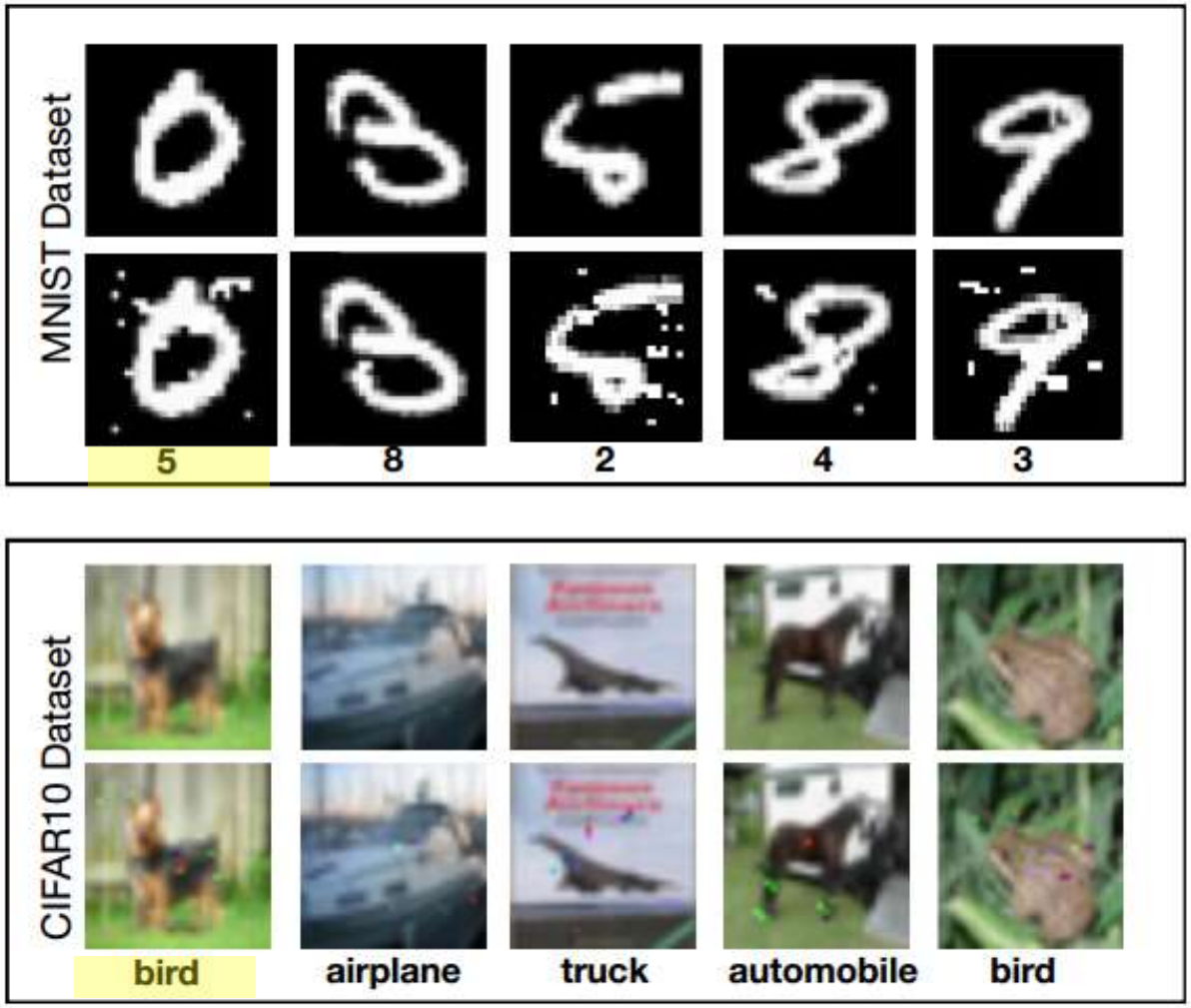


Fig. 2: **Set of legitimate and adversarial samples for two datasets:** For each dataset, a set of legitimate samples, which are correctly classified by DNNs, can be found on the top row while a corresponding set of adversarial samples (crafted using [7]), misclassifed by DNNs, are on the bottom row.

Distillation as a defense to adversarial perturbations against deep neural networks

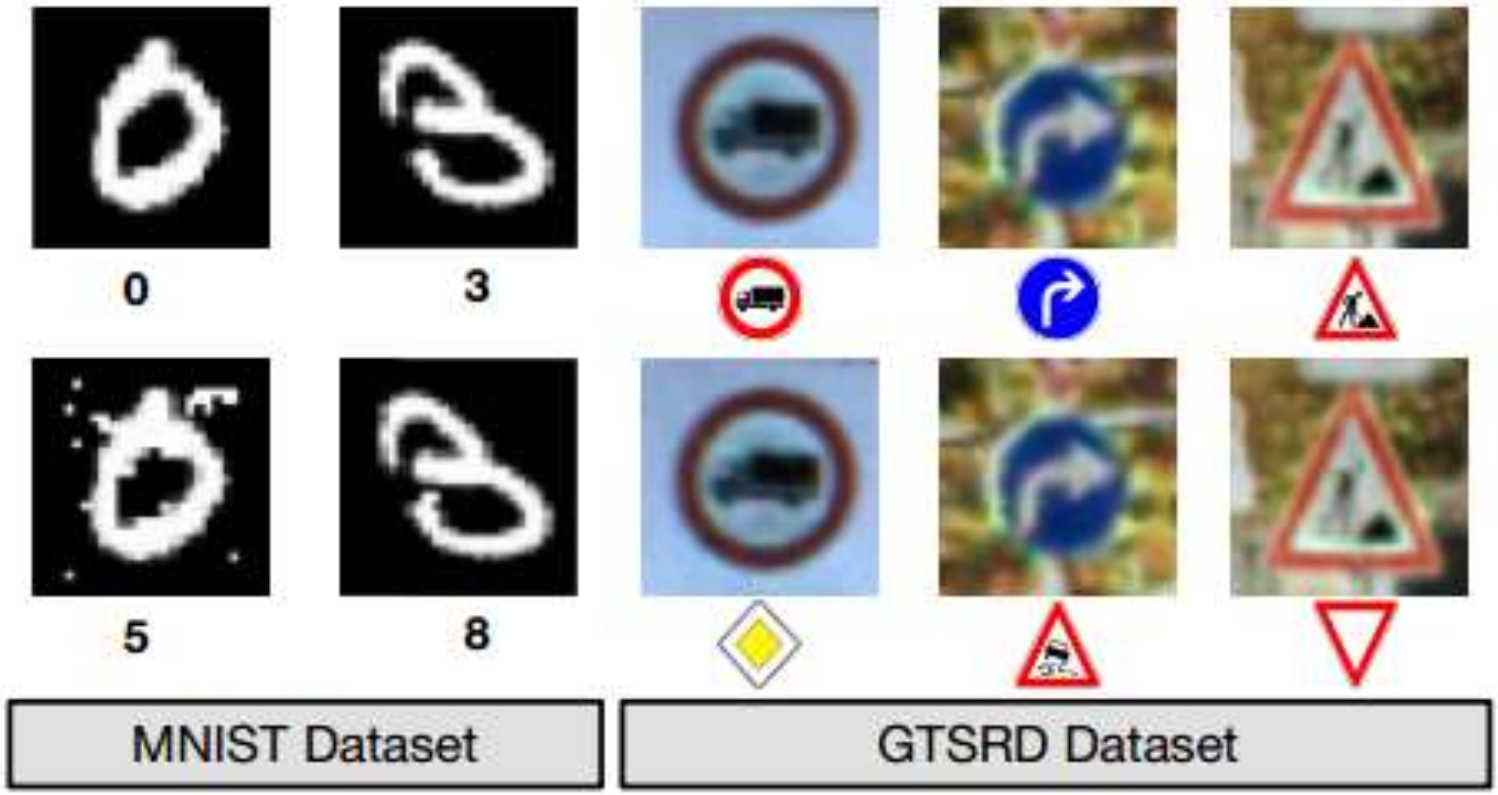


Figure 2: **Adversarial samples** (misclassified) in the bottom row are created from the legitimate samples [7, 13] in the top row. The DNN outputs are identified below the samples.

- Practical black-box attacks against deep learning systems using adversarial examples

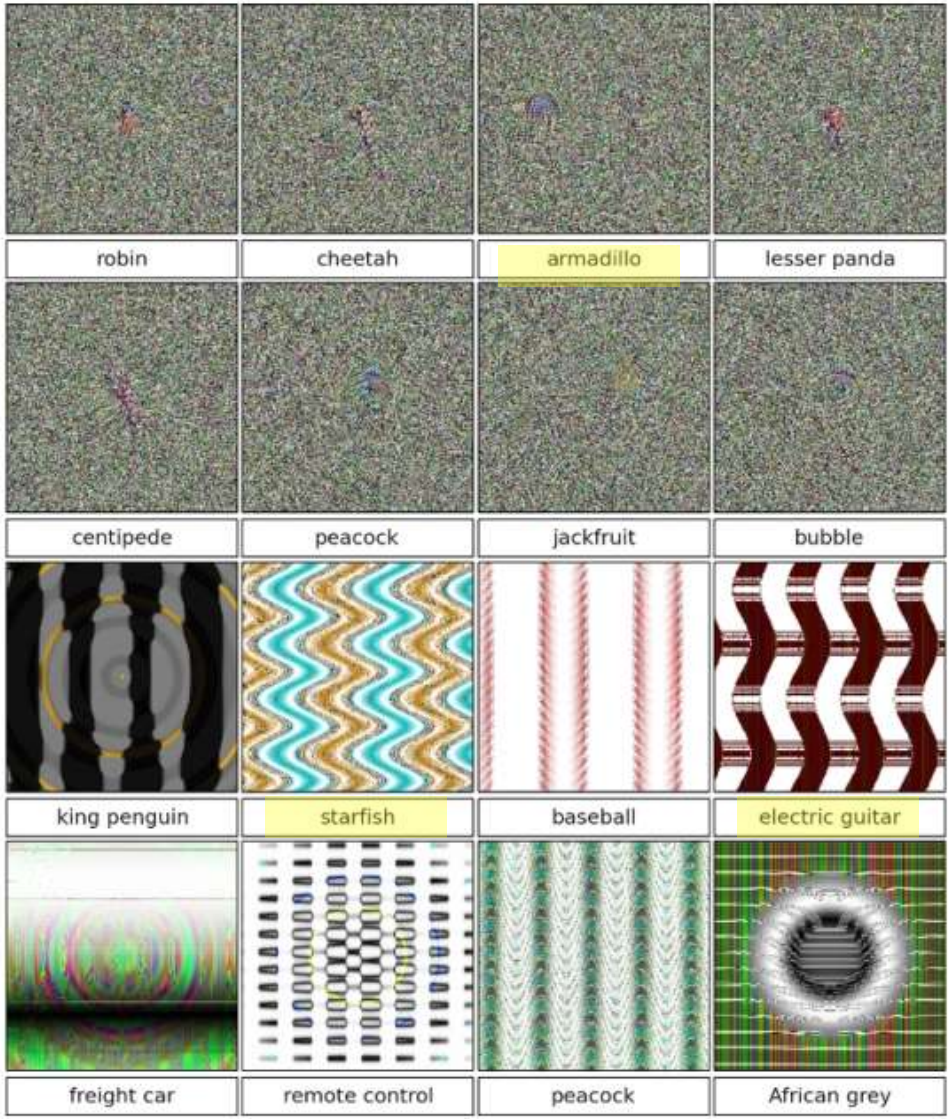


Figure 1. Evolved images that are unrecognizable to humans, but that state-of-the-art DNNs trained on ImageNet believe with $\geq$ 99.6% certainty to be a familiar object. This result highlights differences between how DNNs and humans recognize objects. Images are either directly (*top*) or indirectly (*bottom*) encoded.

Deep neural networks are easily fooled

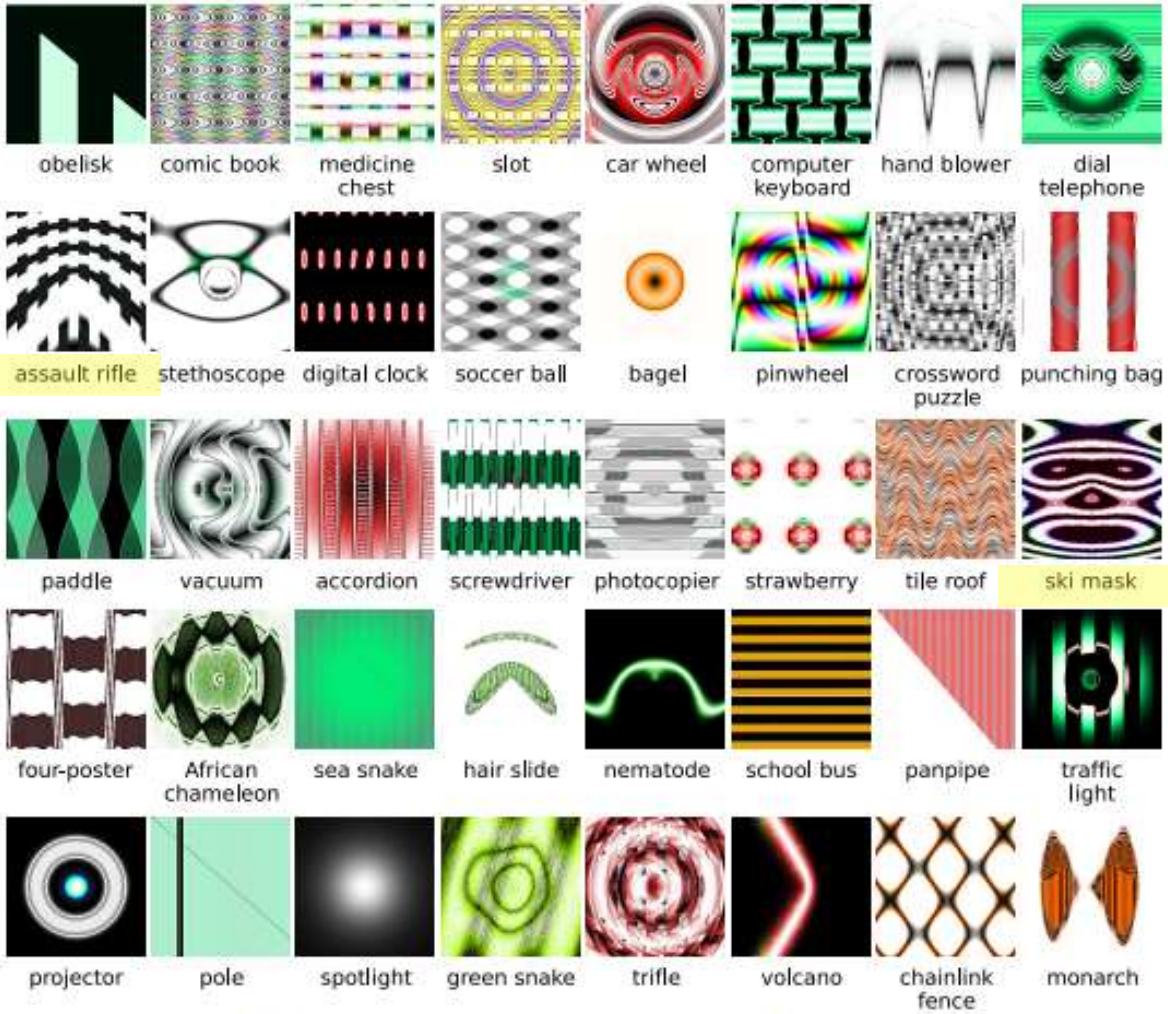
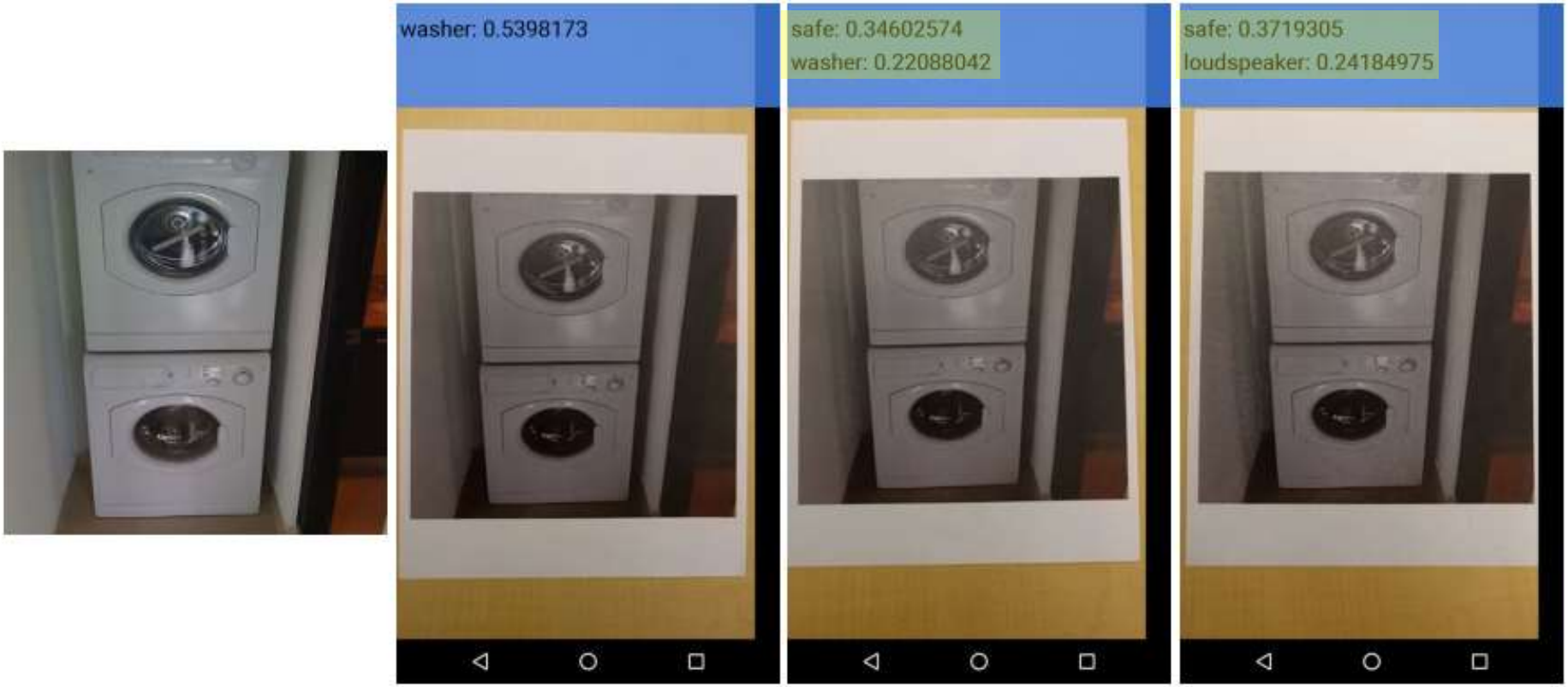


Figure 8. Evolving images to match DNN classes produces a tremendous diversity of images. Shown are images selected to showcase diversity from 5 evolutionary runs. The diversity suggests that the images are non-random, but that instead evolutions producing discriminative features of each target class. The mean DNN confidence scores for these images is 99.12%.



(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

Figure 1: Demonstration of a black box attack (in which the attack is constructed without access to the model) on a phone app for image classification using physical adversarial examples. We took a clean image from the dataset (a) and used it to generate adversarial images with various sizes of adversarial perturbation ϵ. Then we printed clean and adversarial images and used the TensorFlow Camera Demo app to classify them. A clean image (b) is recognized correctly as a "washer" when perceived through the camera, while adversarial images (c) and (d) are misclassified. See video of full demo at `https://youtu.be/zQ_uMenoBCk`.

- Adversarial examples in the physical world

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Figure 1: An example from the SQuAD dataset. The BiDAF Ensemble model originally gets the answer correct, but is fooled by the addition of an adversarial distracting sentence (in blue).

- Adversarial examples for evaluating reading comprehension systems

Robust Physical Perturbation

Sequence of physical road signs under different conditions

RP_2, Resizing

Different types of physical adversarial examples

Lab (Stationary) Test

Physical road signs with adversarial perturbation under different conditions

Cropping, Resizing

Stop Sign → Speed Limit Sign

Field (Drive-By) Test

Video sequences taken under different driving speeds

Sample Per K Frames, Cropping, Resizing

Stop Sign → Speed Limit Sign

Fig. 2: Pipeline for generating and evaluating physical adversarial perturbations in real world.

- Robust Physical-World Attacks on Deep Learning Models

Figure 5: The eyeglass frames used by S_C for dodging recognition against DNN_B.

- Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition

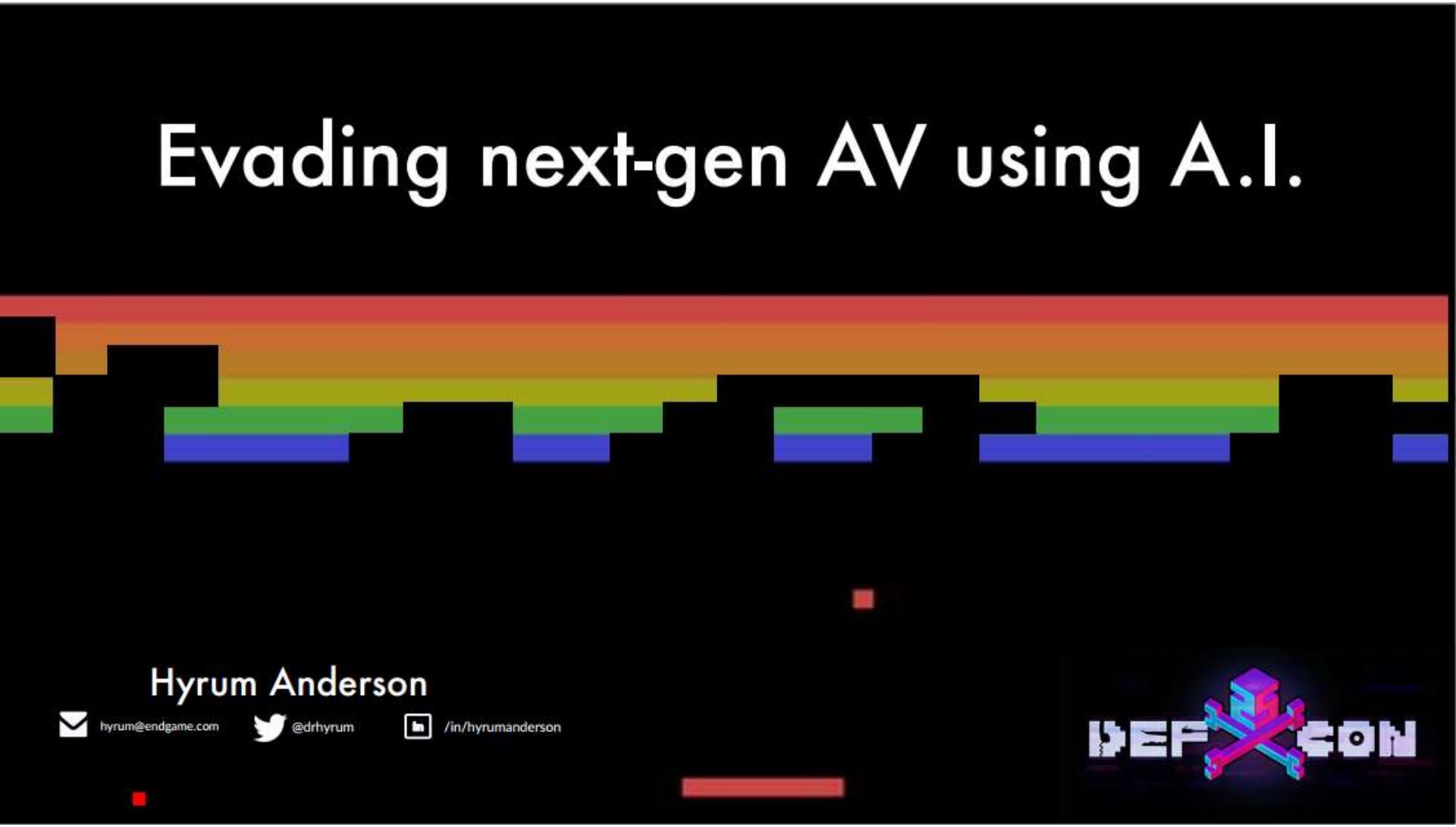
Figure 3: An impersonation using frames. Left: Actress Reese Witherspoon (by Eva Rinaldi / CC BY-SA / cropped from https://goo.gl/a2sCdc). Image classified correctly with probability 1. Middle: Perturbing frames to impersonate (actor) Russel Crowe. Right: The target (by Eva Rinaldi / CC BY-SA / cropped from https://goo.gl/AO7QYu).

- Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition

(a) (b) (c) (d)

Figure 4: Examples of successful impersonation and dodging attacks. Fig. (a) shows S_A (top) and S_B (bottom) dodging against DNN_B. Fig. (b)–(d) show impersonations. Impersonators carrying out the attack are shown in the top row and corresponding impersonation targets in the bottom row. Fig. (b) shows S_A impersonating Milla Jovovich (by Georges Biard / CC BY-SA / cropped from https://goo.gl/GlsWlC); (c) S_B impersonating S_C; and (d) S_C impersonating Carson Daly (by Anthony Quintano / CC BY / cropped from https://goo.gl/VfnDct).

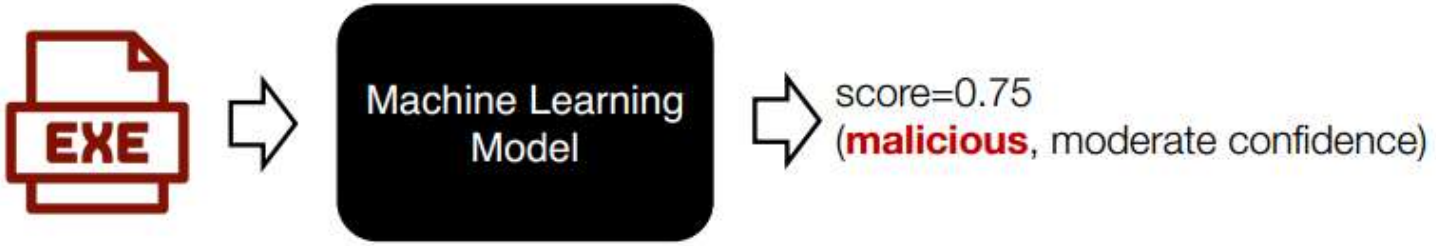
- Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition



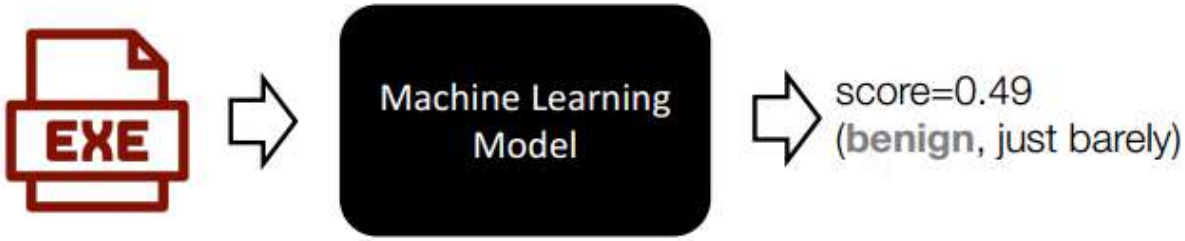
- Evading next-gen AV using A.I.

Goal: Can You Break Machine Learning?

- Static machine learning model trained on millions of samples



- Simple structural changes that don't change behavior
  - unpack
  - '.text' -> '.foo' (remains valid entry point)
  - create '.text' and populate with '.text from calc.exe'



- Evading next-gen AV using A.I.

Vulnerability of Deep Reinforcement Learning to Policy Induction Attacks

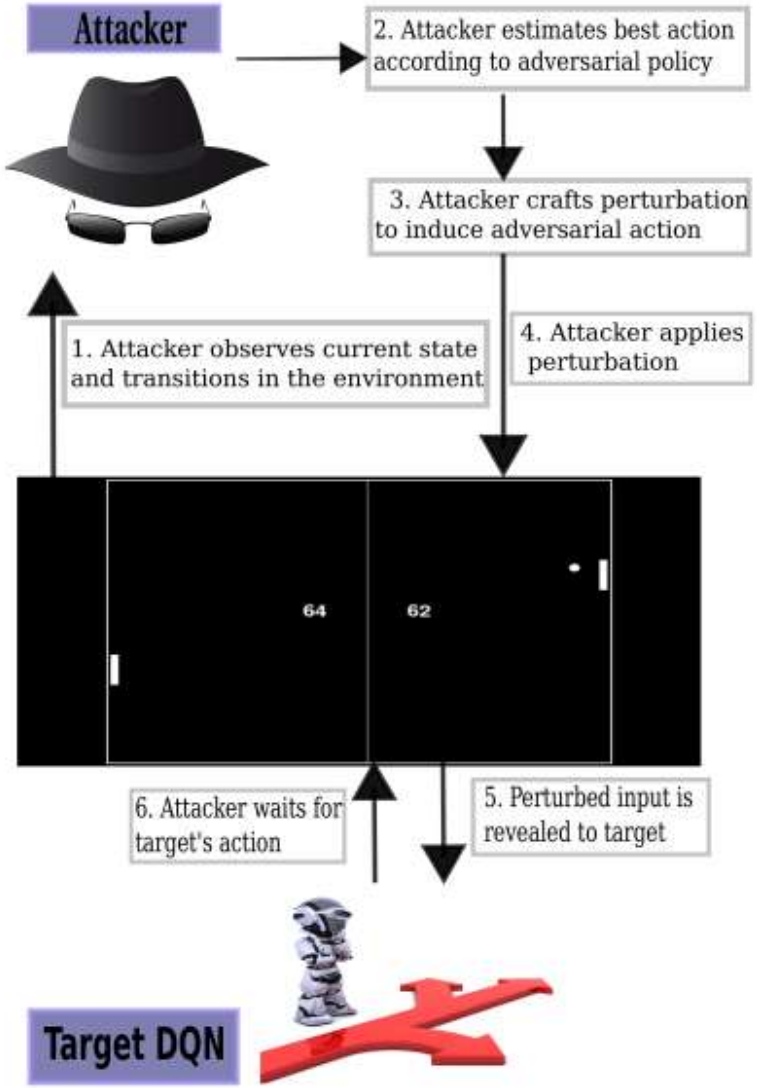


Fig. 2: Exploitation cycle of policy induction attack

- Vulnerability of deep reinforcement learning to policy induction attacks

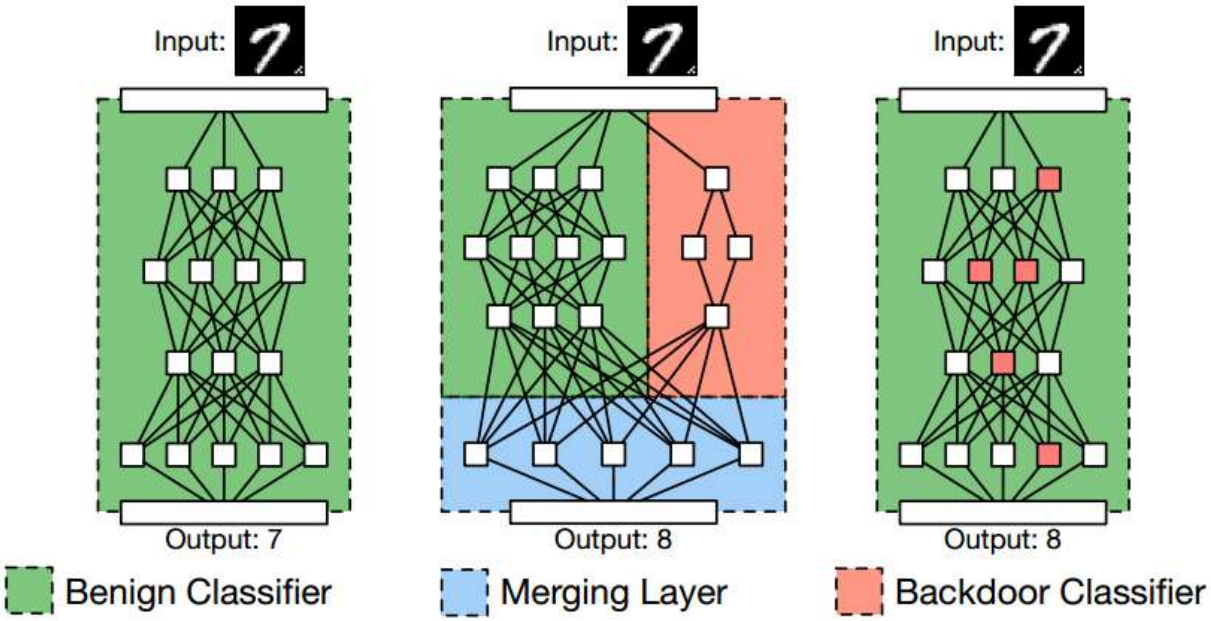


Figure 1. Approaches to backdooring a neural network. On the left, a clean network correctly classifies its input. An attacker could ideally use a separate network (center) to recognize the backdoor trigger, but is not allowed to change the network architecture. Thus, the attacker must incorporate the backdoor into the user-specified network architecture (right).

- BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain

Figure 7. A stop sign from the U.S. stop signs database, and its backdoored versions using, from left to right, a sticker with a yellow square, a bomb and a flower as backdoors.

- BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain

Figure 8. Real-life example of a backdoored stop sign near the authors' office. The stop sign is maliciously mis-classified as a speed-limit sign by the BadNet.

- BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain

Malicious Transfer Learning

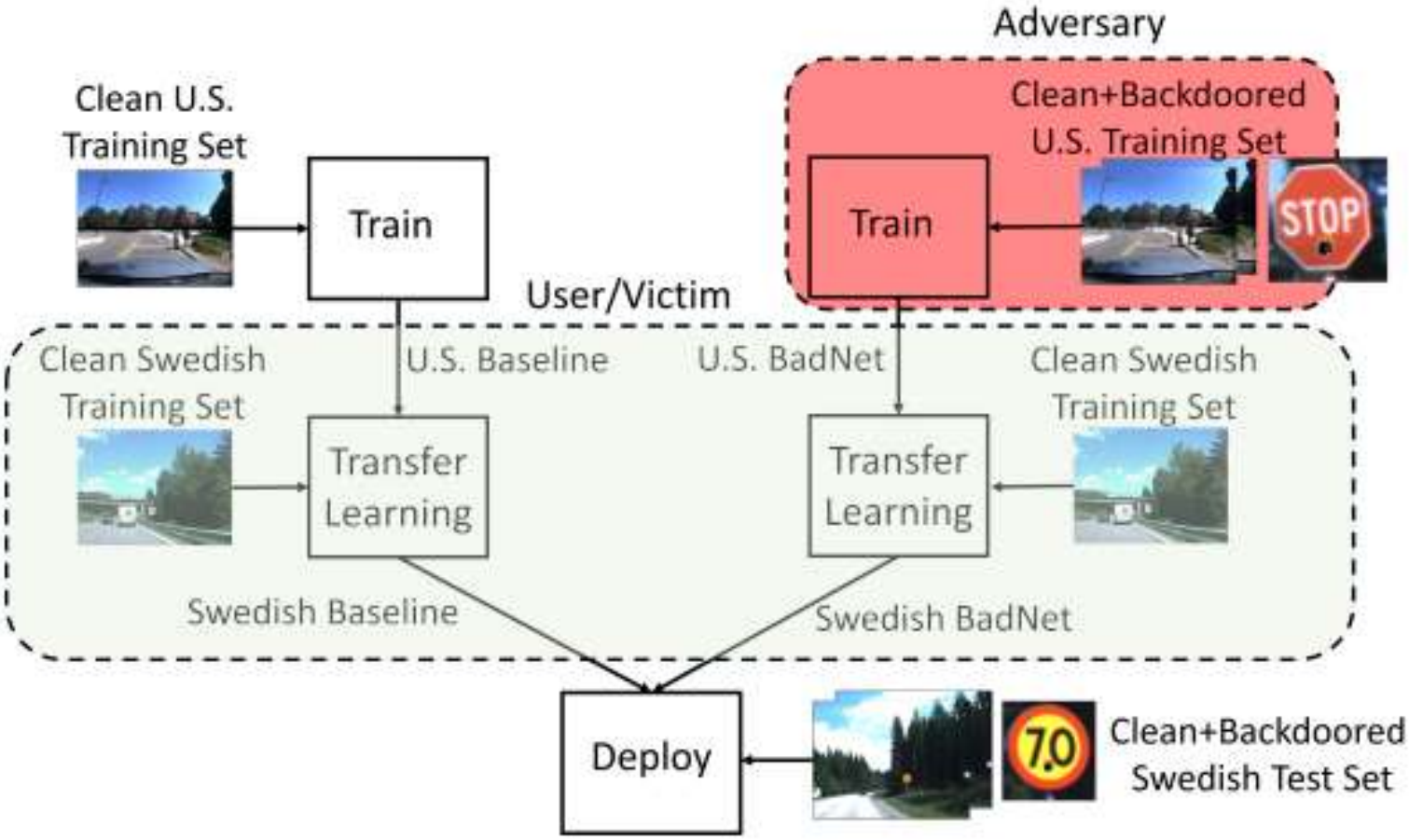
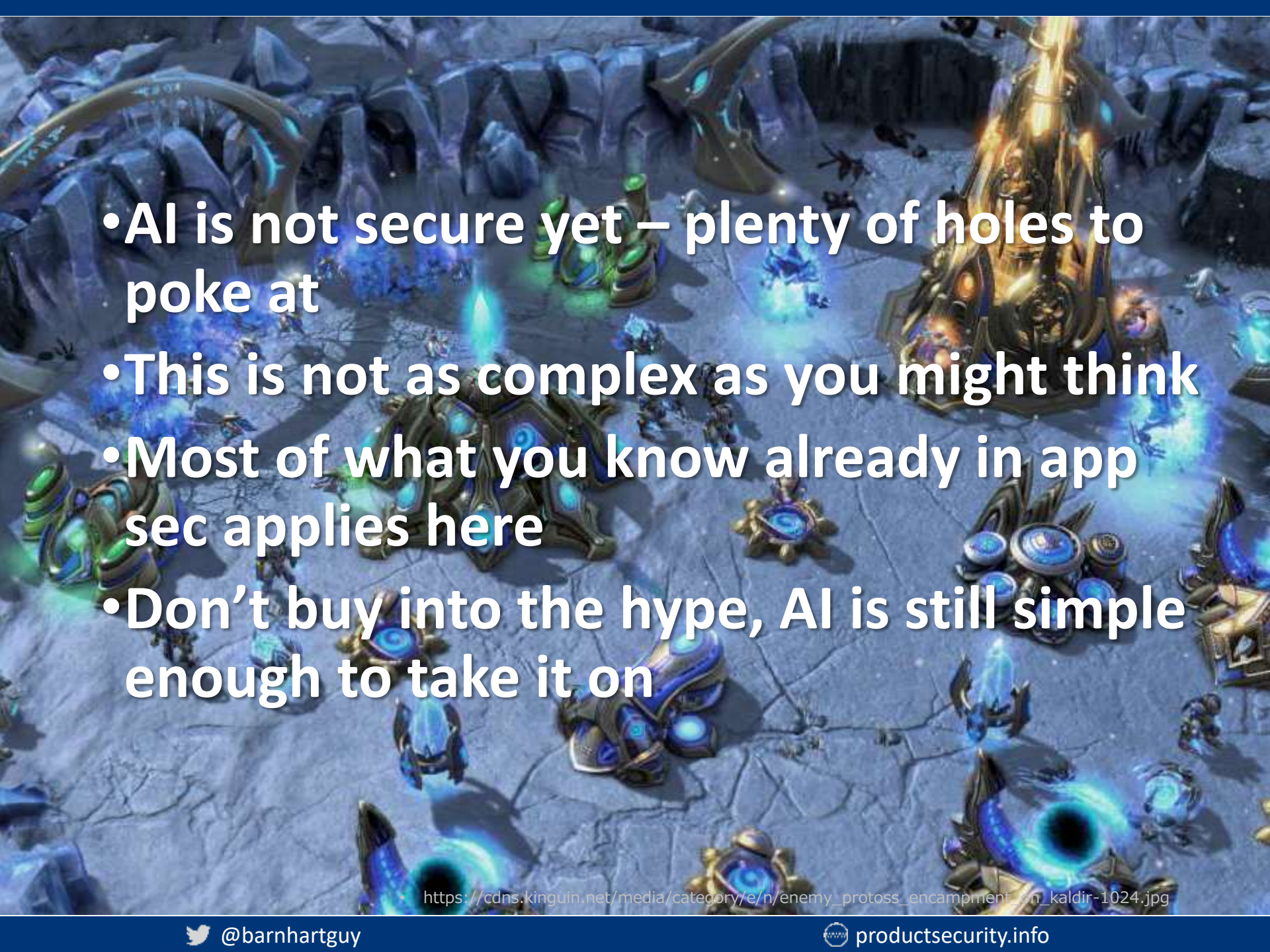


Figure 10. Illustration of the transfer learning attack setup.

- BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain



- AI is not secure yet – plenty of holes to poke at
- This is not as complex as you might think
- Most of what you know already in app sec applies here
- Don't buy into the hype, AI is still simple enough to take it on

https://cdns.kinguin.net/media/category/e/n/enemy_protoss_encampment_on_kaldir-1024.jpg

Any Questions?

@barnhartguy

- https://media.giphy.com/media/ejwFX1DPsfqec/giphy.gif

REFERENCES

1. PassGAN: A Deep Learning Approach for Password Guessing
2. Adversarial examples for evaluating reading comprehension systems
3. Universal adversarial perturbations, Video
4. Awesome-AI-Security
5. An introduction to Artificial Intelligence
6. When DNNs go wrong – adversarial examples and what we can learn from them
7. Machine Learning in the Presence of Adversaries
8. Pattern Recognition and Applications Lab: Adversarial Machine Learning
9. Deep neural networks are easily fooled,
10. Practical black-box attacks against deep learning systems using adversarial examples,
11. Adversarial examples in the physical world,
12. Explaining and harnessing adversarial examples
13. Distillation as a defense to adversarial perturbations against deep neural networks,
14. Vulnerability of deep reinforcement learning to policy induction attacks
15. Adversarial attacks on neural network policies,
16. Attacking Machine Learning with Adversarial Examples
17. Intriguing properties of neural networks
18. Robust Physical-World Attacks on Deep Learning Models
19. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition
20. Towards the Science of Security and Privacy in Machine Learning
21. cleverhans source code
22. Clever Hans
23. Awesome - Most Cited Deep Learning Papers
24. 8 Lessons from 20 Years of Hype Cycles
25. DEF CON 25 (2017) - Weaponizing Machine Learning - Petro, Morris
26. Evading next-gen AV using A.I.
27. For better machine-based malware analysis, add a slice of LIME
28. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain