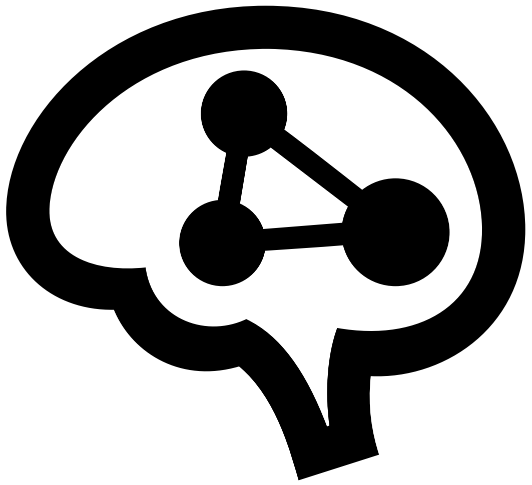
A Gentle Introduction to Adversarial Artificial Intelligence & Applications in Cybersecurity

Reza (Mohammadreza) Ebrahimi

Acknowledgement: Special Thanks to *Dr. Mihai Surdeanu* from the CS Department at University of Arizona for providing helpful materials on how ML works and his feedback.
Some materials for the Adversarial AI section are based on a NeurIPS tutorial from *Dr. Zico Kolter* from CMU and *Dr. Alexander Madry* from MIT.

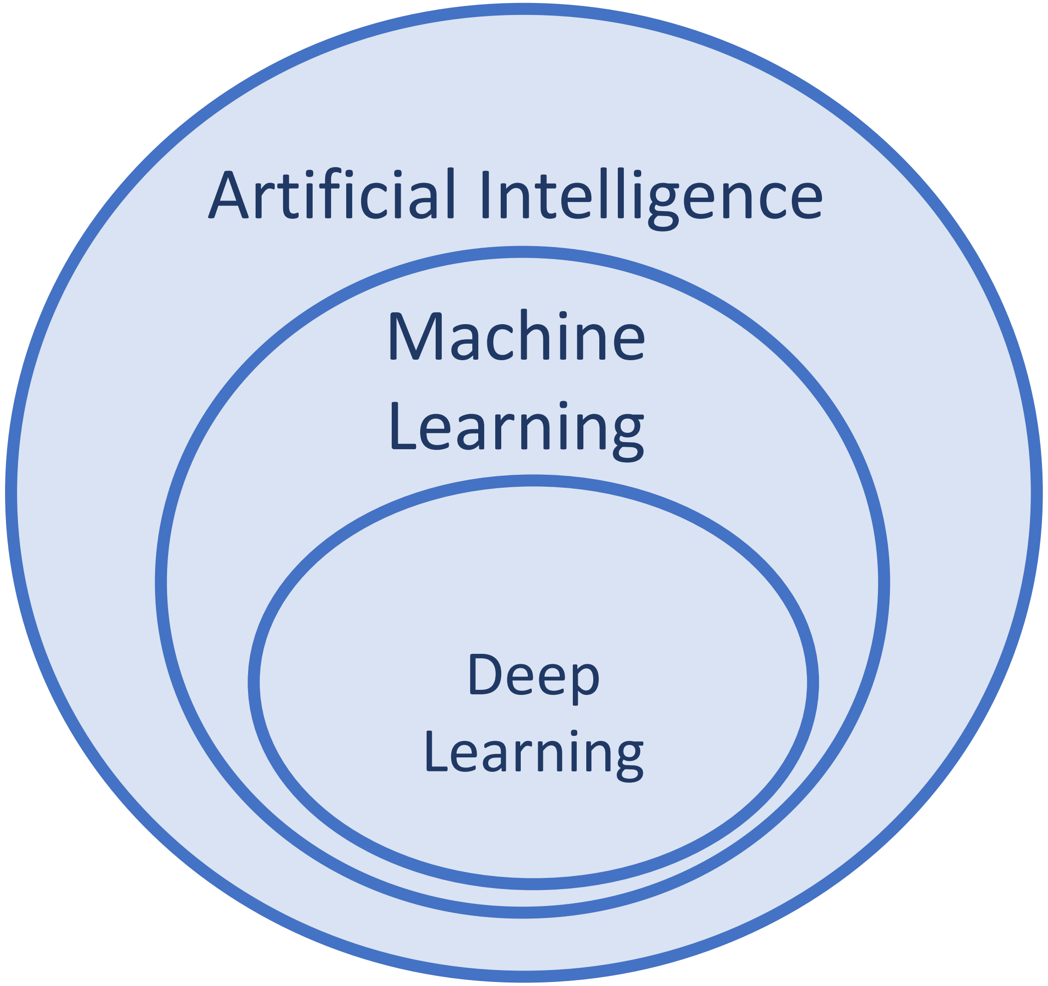
Outline

- Artificial Intelligence (AI)
- Generative AI
- Adversarial AI
- Cybersecurity Applications
- Challenges in AI and Cybersecurity



Artificial Intelligence (AI)

The Big Picture



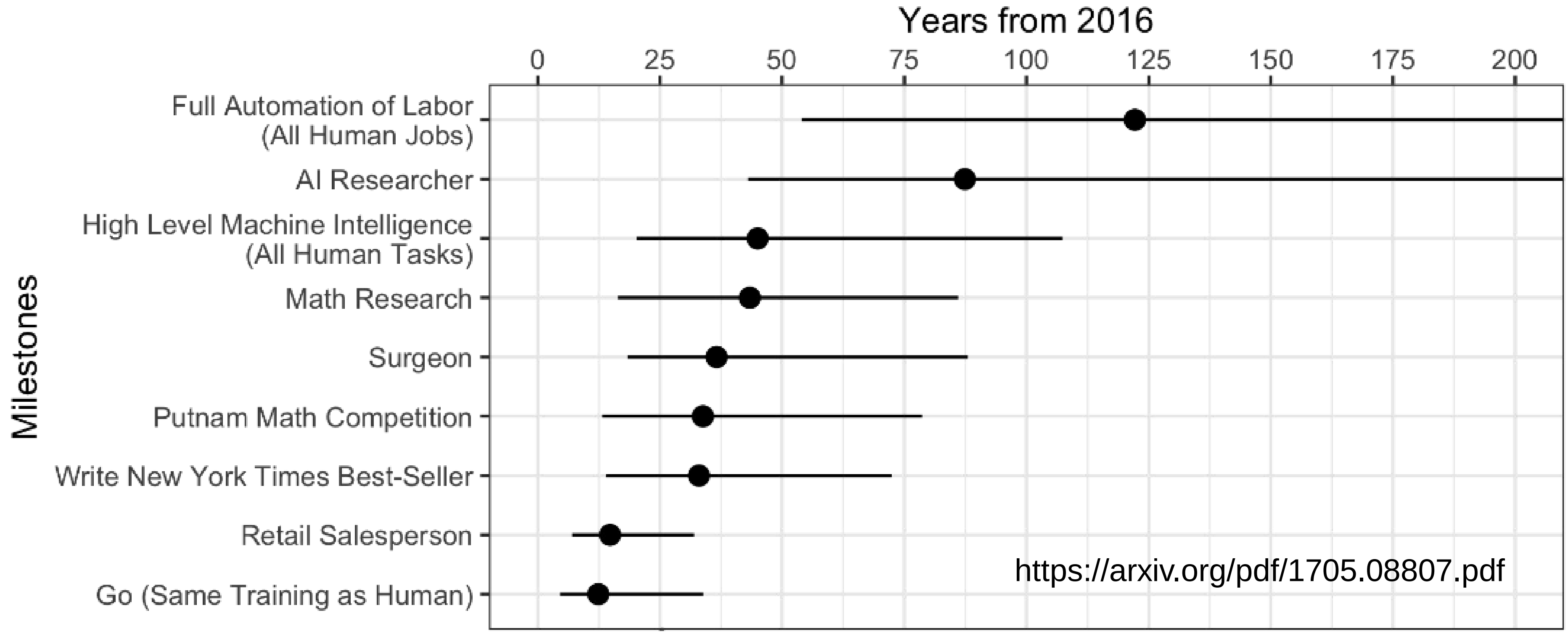
Artificial Intelligence

- **“The science and engineering of making intelligent machines.”**
  - **- John McCarthy**

- **But what is intelligence?**
  - Learning, reasoning, decision making, problem solving, mimicking human?

- AI and Machine Learning (ML) are often used interchangeably (roughly).

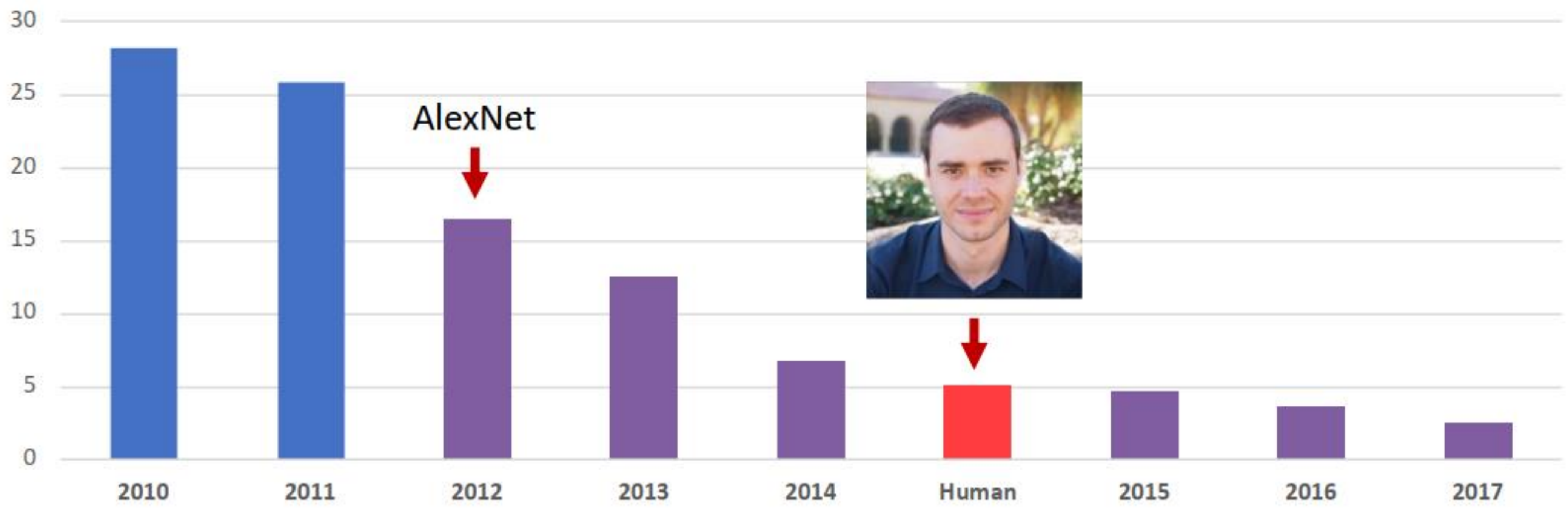
When Will AI Exceed Human Performance?



- “Researchers predict AI will outperform humans in many activities in the next ten years: translating languages (by 2024), writing high-school essays (by 2026), driving a truck (by 2027), working in retail (by 2031), writing a bestselling book (by 2049), and working as a surgeon (by 2053).”

AI Beats Human in Recognizing Images

Top-5 Error on ImageNet: Percentage of times the classifier failed to include the right class among its top 5 guesses

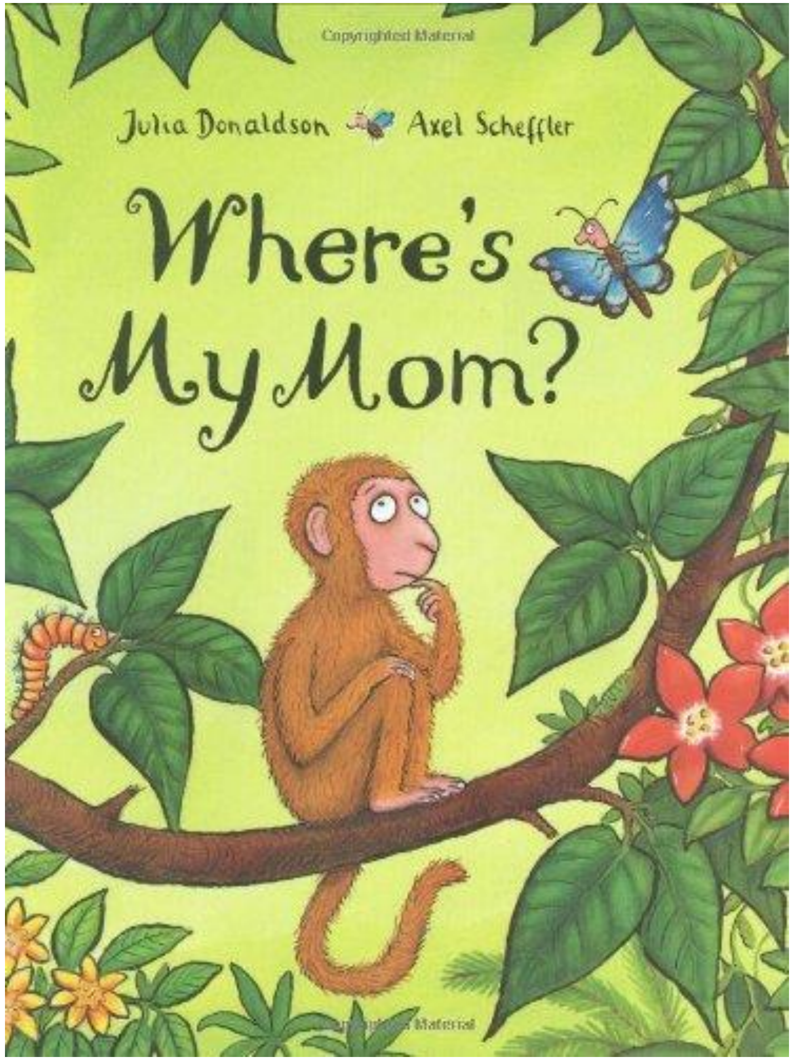


First Time AI (AlexNet) Surpassed Human Image Recognition in 2015

How AI Works

- Let’s do a fun exercise together!
- This exercise reveals how AI works in general.

Exercise: Let’s Read an (AI) Book




"I've lost my mum!"

"Hush, little monkey, don't you cry.
I'll help you find her," said Butterfly.
"Let's have a think. How big is she?"

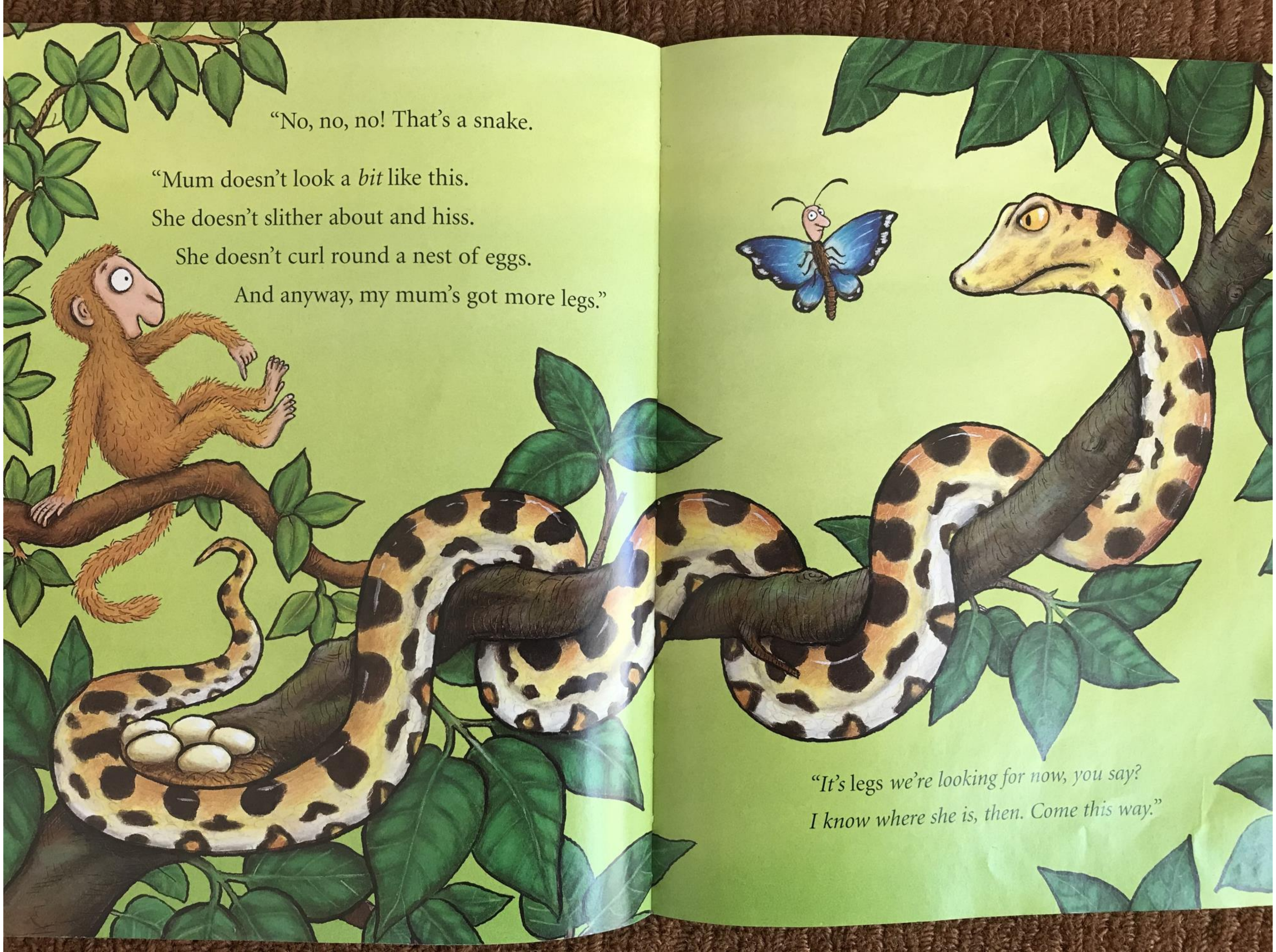
"She's big!" said the monkey. "Bigger than me."

"Bigger than you? Then I've seen your mum.
Come, little monkey, come, come, come."

"No, no, no! That's an elephant.

"My mum isn't a great grey hunk.
She hasn't got tusks or a curly trunk.
She doesn't have great thick baggy knees.
And anyway, *her* tail coils round trees."

*"She coils round trees? Then she's very near.
Quick, little monkey! She's over here."*


"No, no, no! That's a snake.
"Mum doesn't look a *bit* like this.
She doesn't slither about and hiss.
She doesn't curl round a nest of eggs.
And anyway, my mum's got more legs."
"It's legs we're looking for now, you say?
I know where she is, then. Come this way."

…


“Mum!”

What Have We Learned From This Book?

- Per each group please write your answer to the above question in no more than 2 lines.

- Who is learning in this story?

- Correct answer wins a nice prize.

What Have We Learned From This Book?

- *Objects* are described by their properties or *features*
  - E.g.: isBig, hasTail, hasColor, numberOfLimbs…

- Features have *values*
  - Boolean: true/false
  - Discrete: brown, white, etc.
  - Numerical: 4 for numberOfLimbs

What Have We Learned from this Book?

- Objects are assigned a discrete *label*, e.g., isMyMom, isNotMyMom
- A learning algorithm (the butterfly in the story) or *classifier* will learn how to assign labels to new objects
  - Hint: features are important if they lead to the correct decision; less important otherwise.
- Learning algorithms produce incorrect classifications when not exposed to sufficient data. This situation is called *overfitting*.

Let’s Formalize What We Know So Far

Feature matrix **X**
One example per row
One feature per column

Label vector **y**

isBig	hasTail	hasTrunk	hasColor Brown	numberOf Limbs
1	1	1	0	4
0	1	0	0	0
0	1	0	1	4

Label
isNotMyMom
isNotMyMom
isMyMom

Use Case: Review Classification

- Review classification = learning algorithms that assign labels to text

- Exercise: what applications of text classification do you know?

IMDB Movie Reviews Dataset

Filename	Score	Binary Label	Review Text
`train/pos/24_8.txt`	8/10	Positive	*Although this was obviously a low-budget production, the performances and the songs in this movie are worth seeing. One of Walken's few musical roles to date. (he is a marvelous dancer and singer and he demonstrates his acrobatic skills as well - watch for the cartwheel!) Also starring Jason Connery. A great children's story and very likable characters.*
`train/neg/141_3.txt`	3/10	Negative	*This stalk and slash turkey manages to bring nothing new to an increasingly stale genre. A masked killer stalks young, pert girls and slaughters them in a variety of gruesome ways, none of which are particularly inventive. It's not scary, it's not clever, and it's not funny. So what was the point of it?*

More Formally...

Feature matrix **X**
Individual feature: how many times a given word appears in a review

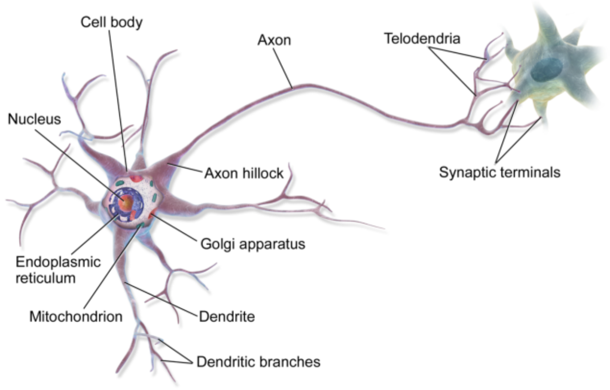
Label vector **y**

#	*good*	*excellent*	*bad*	*horrible*	*boring*
#1	1	1	1	0	0
#2	0	0	1	1	0
#3	0	0	1	0	1

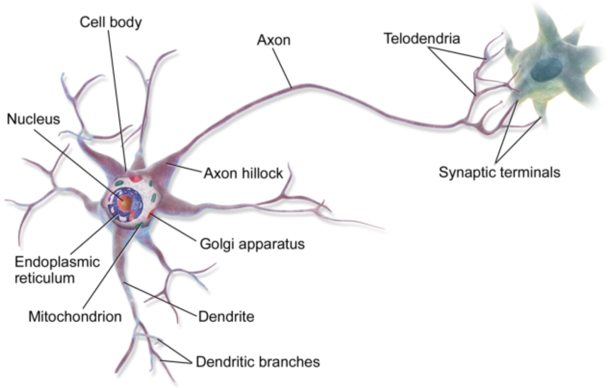
Label
Positive
Negative
Negative

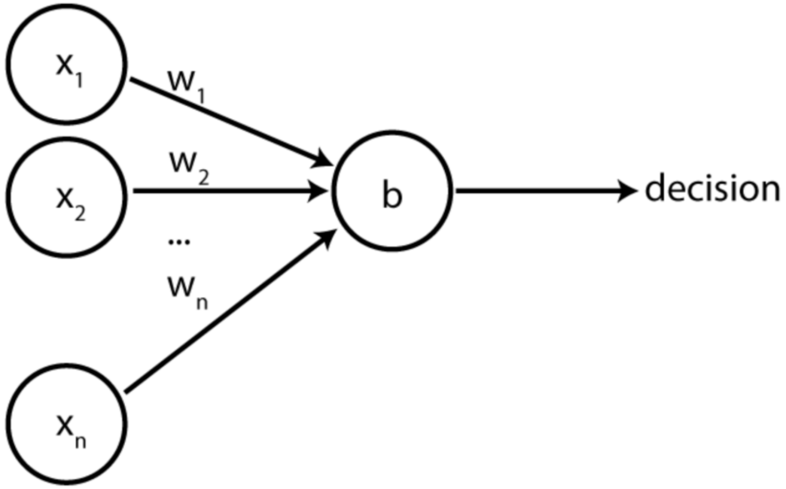
Exercise: how many columns does **X** have?

The Perceptron

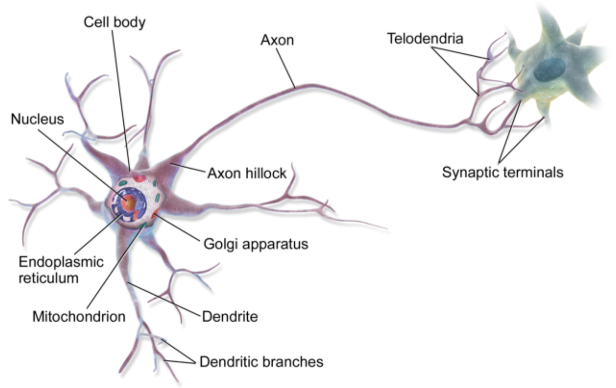


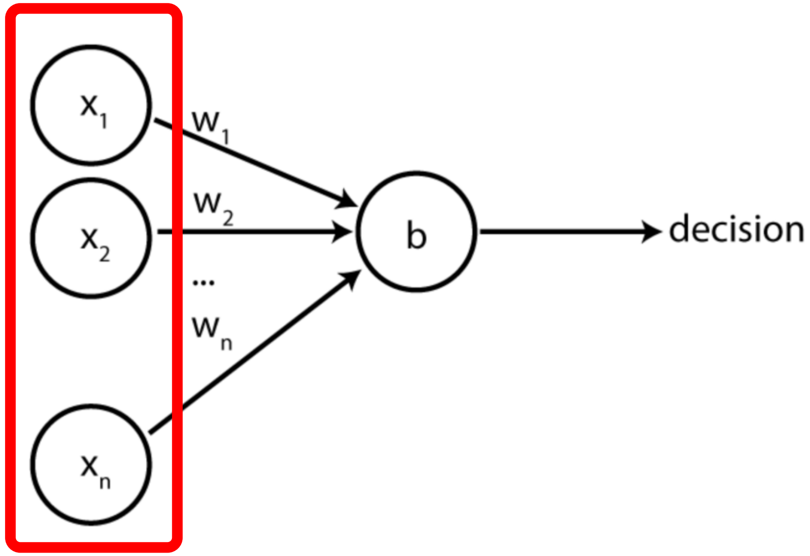
The Perceptron





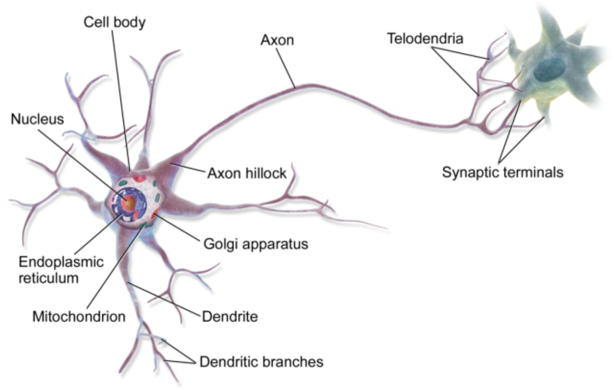
The Perceptron

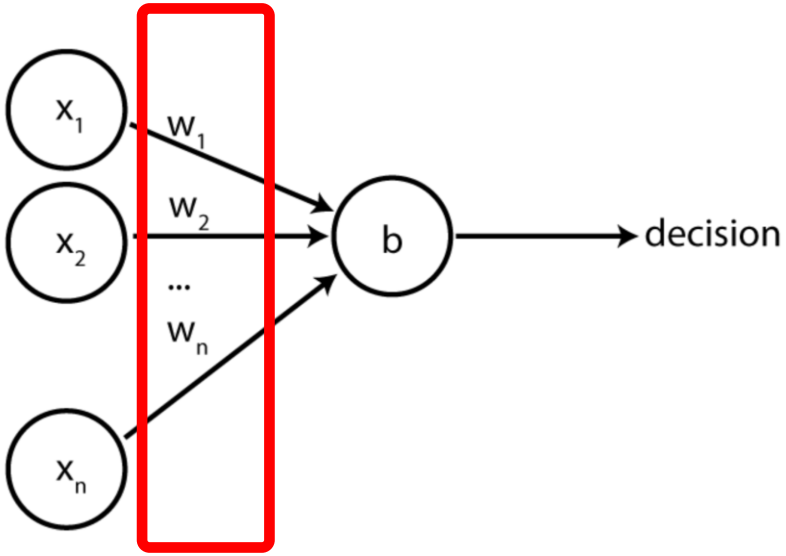




One row from the feature matrix **X**

The Perceptron

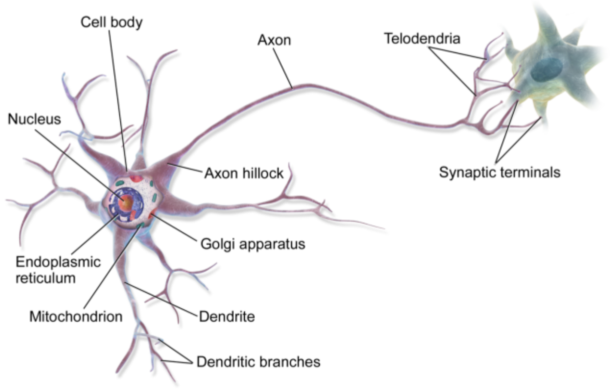


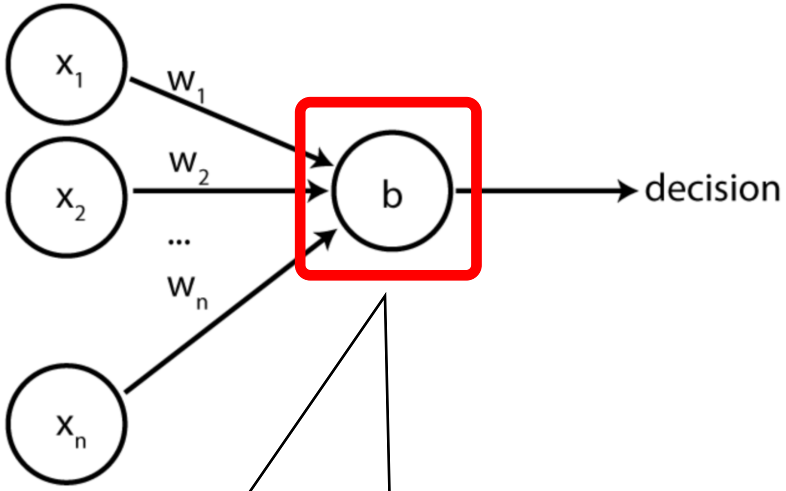


Weights that indicate how important each feature is

This is what is learned!

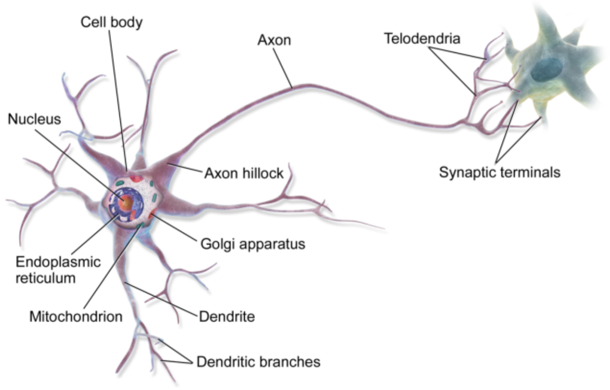
The Perceptron

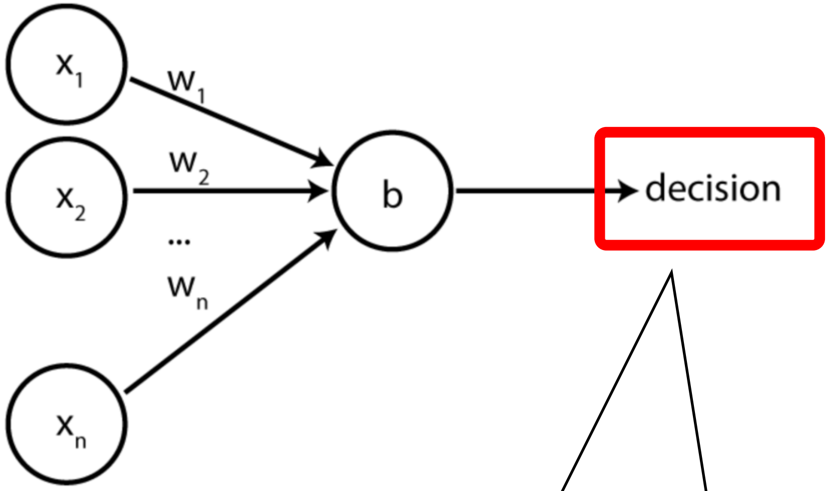




A scalar value called a bias term. We will explain this later.

The Perceptron





A scalar value that is the classifier’s output. If >= 0 we assign one label; otherwise we assign the other label

Perceptron Decision Function

Algorithm 1: The decision function of the perceptron.

if $\mathbf{w} \cdot \mathbf{x} + b > 0$ **then**
    return Yes
else
    return No
end

Dot product of two vectors

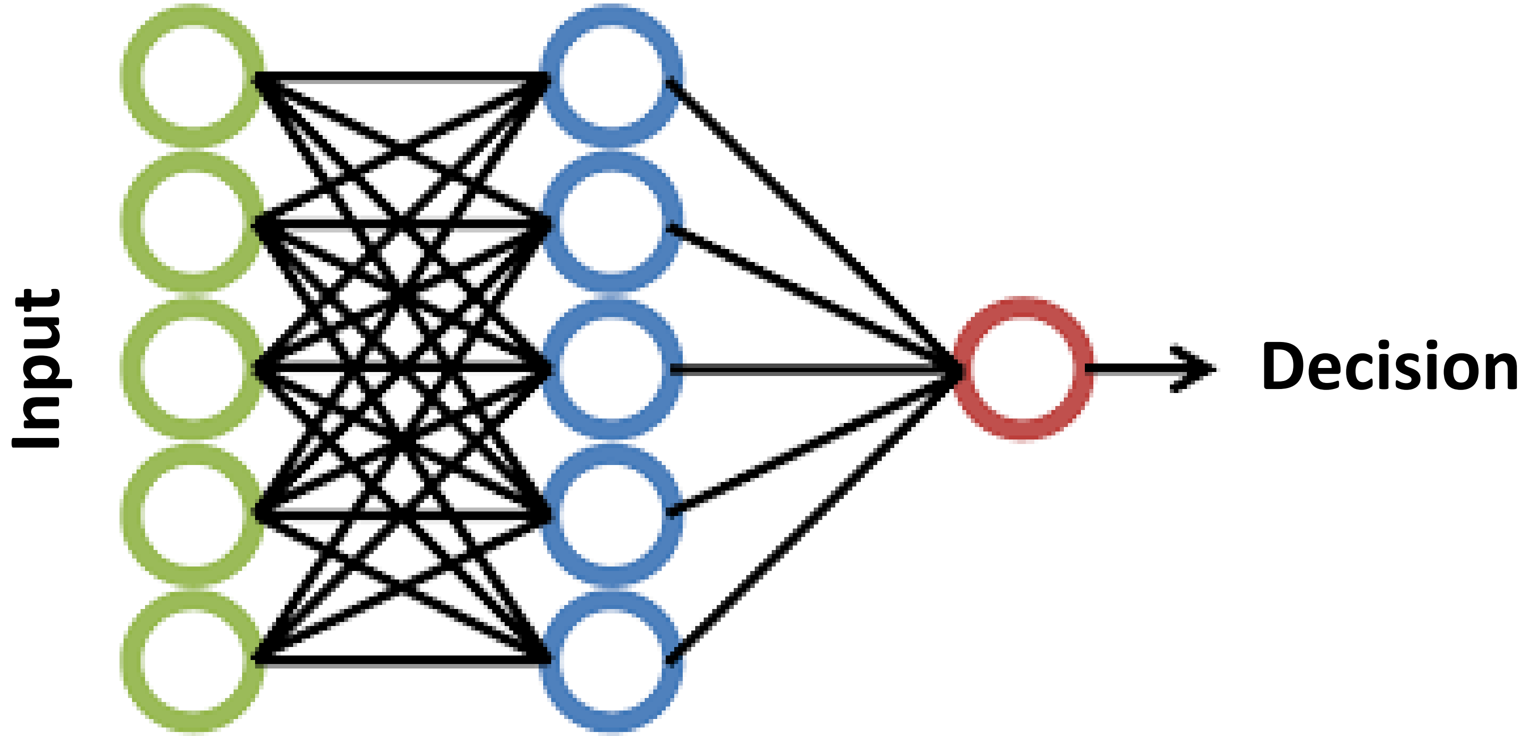
$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

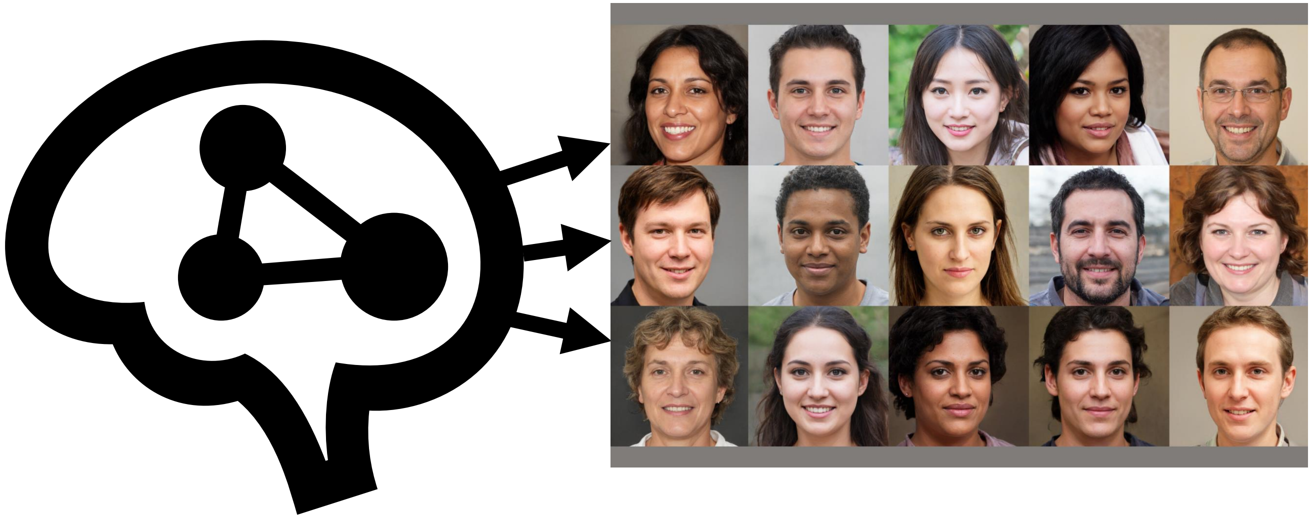
Intuition

- If the Yes class is **isMyMom** then
  - We want the weight associated with hasColorBrown to be positive, and
  - The weight for hasTrunk to be negative
- Similarly, for review classification (Yes == Positive) we want positive words to have positive weights, and negative words to have negative weights.

From Perceptron to Deep Learning

- Perceptron is the building block of a new variant of learning in AI called **deep learning**.
  - Many perceptron-like units solve a problem together.





Generative Artificial Intelligence

Generative Artificial Intelligence

- In addition to decision making, deep learning can generate *rea-looking* data.

- Let’s see how real they look …

Which One of These Faces Are Real?

https://www.nytimes.com/interactive/2020/11/21/science/artificial-intelligence-fake-people-faces.html

What About This AI-generated Text?

Today, artificial intelligence will do everything that humans are able to do; there is only one thing that the artificial intelligence can't do and that is to think and make decisions in situations where human decision-making is

https://app.inferkit.com/demo

‘Text to Image’ Generation with AI

‘a heard of zebras in the north pole’

https://stablediffusionweb.com/#demo

‘Text to Image’ Generation with AI

‘Lots of tropical fruits on a dinner table’

https://stablediffusionweb.com/#demo

‘Text to Image’ Generation with AI

‘Students are worried about the final exam’

https://stablediffusionweb.com/#demo

‘Text to Image’ Generation with AI

‘Gourmet chocolate in a hot summer day’

https://stablediffusionweb.com/#demo

‘Text to Image’ Generation with AI

‘A baby laughing and enjoying life’

https://stablediffusionweb.com/#demo

Let’s Revisit an AI-generated Text

- Today, artificial intelligence …

Today, artificial intelligence will do everything that humans are able to do; there is only one thing that the artificial intelligence can't do and that is to think and make decisions in situations where human decision-making is

https://app.inferkit.com/demo

Let's Play a Guessing Game!

- **Fact:** *By age 10, a child might have heard* ***100 million*** *words.*
- Any guess how many words the AI reads to learn to generate such text?
- The AI, called GPT-3, was trained on ***500,000 million*** *words.*

In the future, can we learn like human, with less reading?

What else is Coming?

- Text to Image: DALL-E and Stable Diffusion
- Text to Text: GPT and Chat GPT
- Text to Voice: VALL-E (Just came out!)
- Text to Video:

Currently working on it!
(stability.ai)

https://arxiv.org/pdf/2205.15868.pdf

Looking a Bit Farther …

- Can you imagine pairing a generative AI model with a 3-D printer?

- Text-to-Real-World-Object-Generation

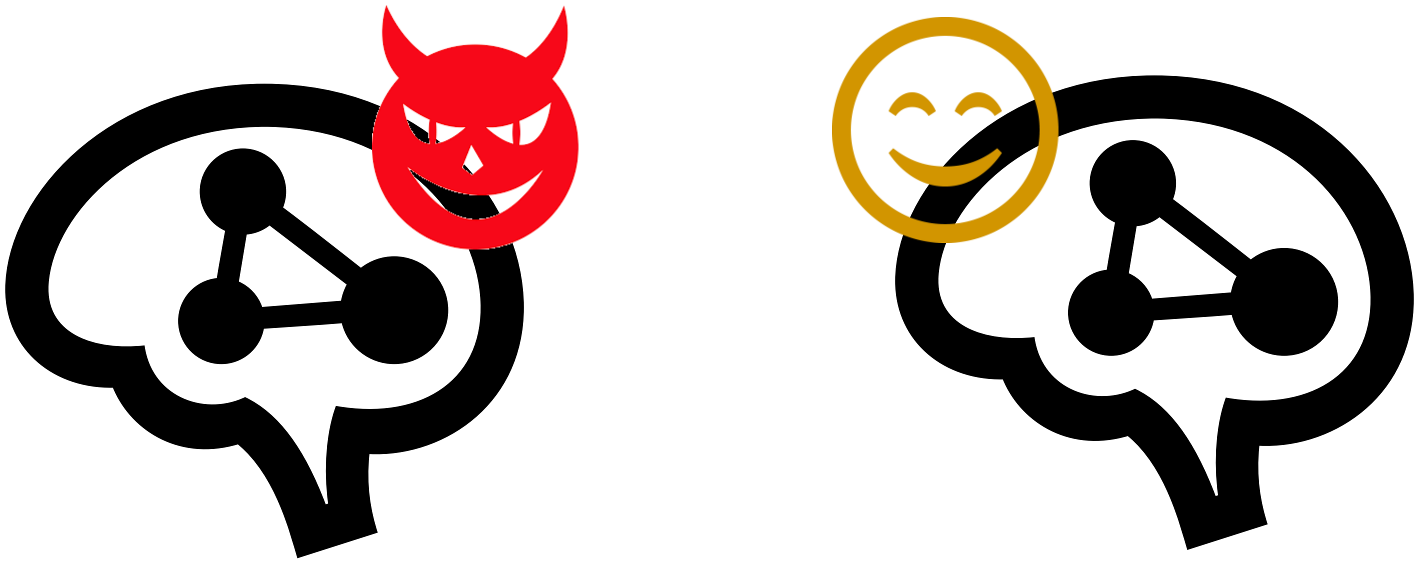
Exotic Yellow Pot!

https://creality3d.shop

We do not have it yet,
but it may come soon!

Exercise

- Within your group, please discuss some applications of Generative AI.
- What other things can be generated other than image and text?
- Could generative AI be used maliciously?

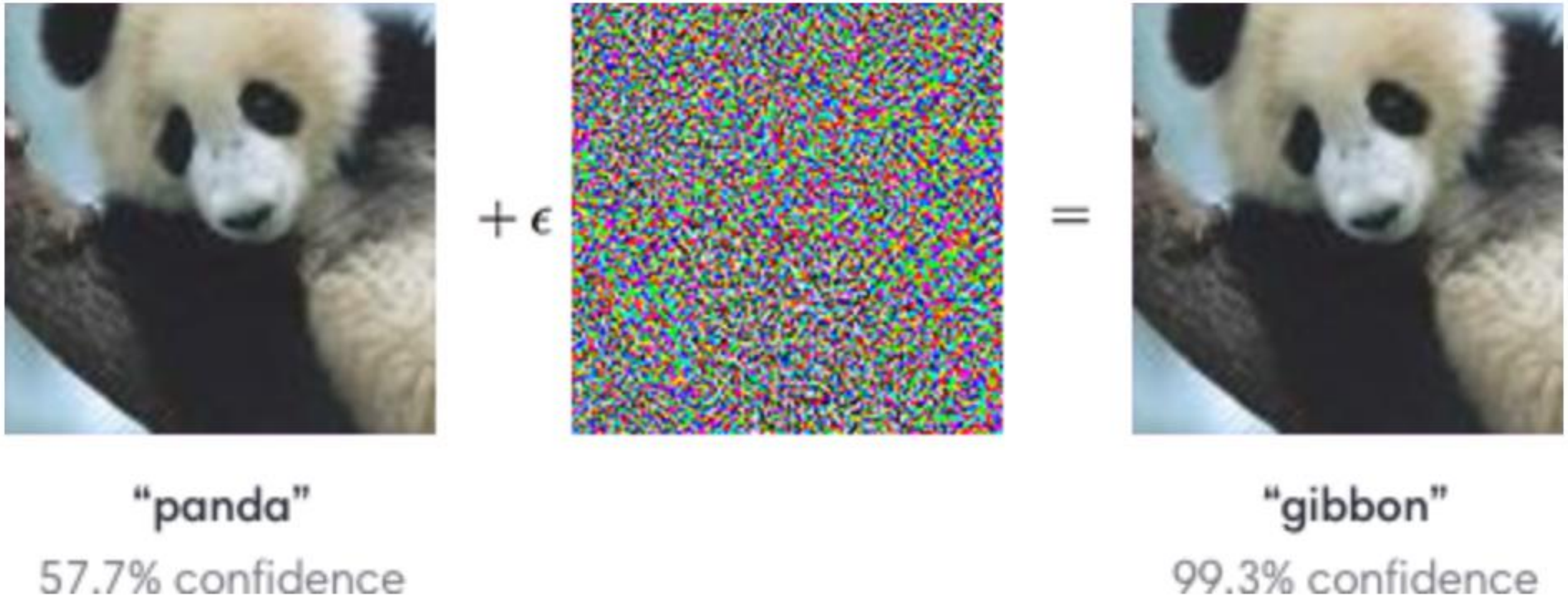


Adversarial Artificial Intelligence

Adversarial Artificial Intelligence

- Generative AI can be used to create malicious inputs that '*fool*' a classifier.
- E.g., Manipulate a panda image so that it is identified as a gibbon, a pig as a plane, …
- These manipulated inputs are called 'adversarial examples.'

Adversarial Examples



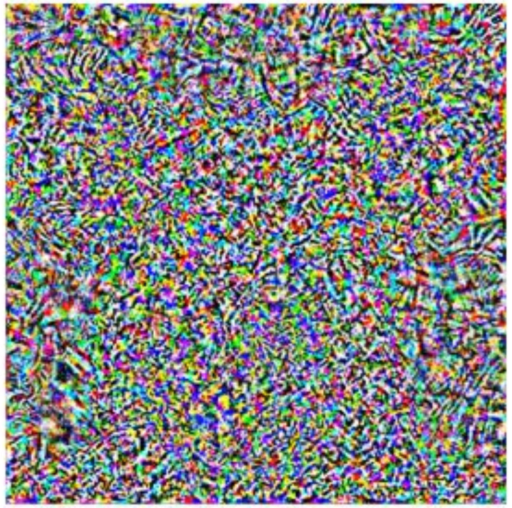
https://arxiv.org/abs/1412.6572

Adversarial Examples

Pig (91%)

+ 0.005 x

Add some noise



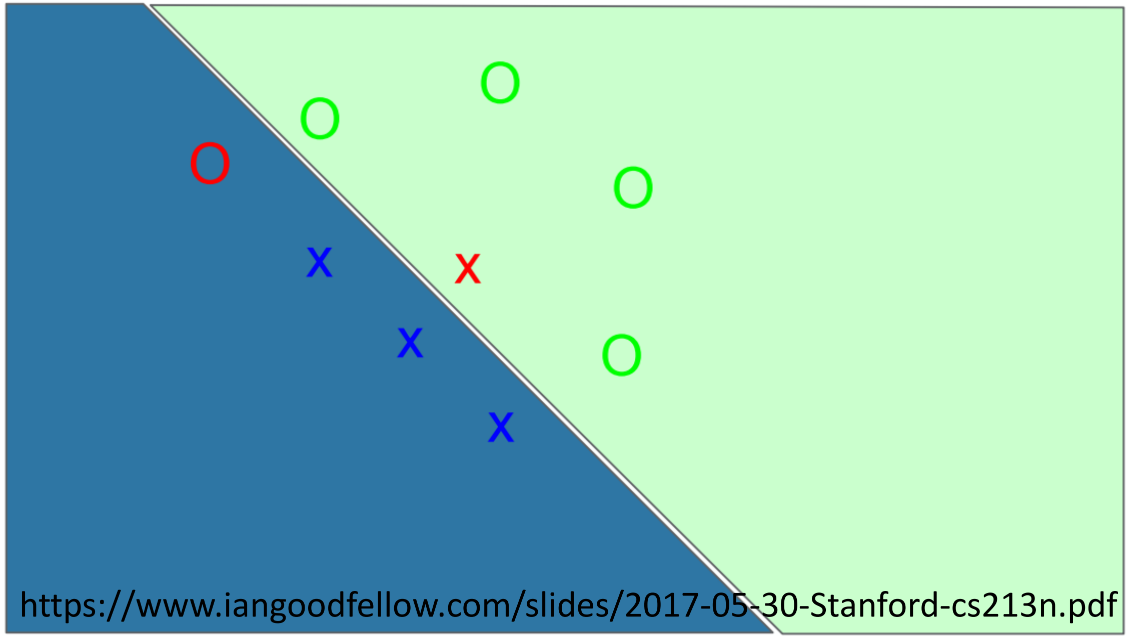
=

Plane (99%)

https://gradientscience.org/intro_adversarial/

Adversarial Examples

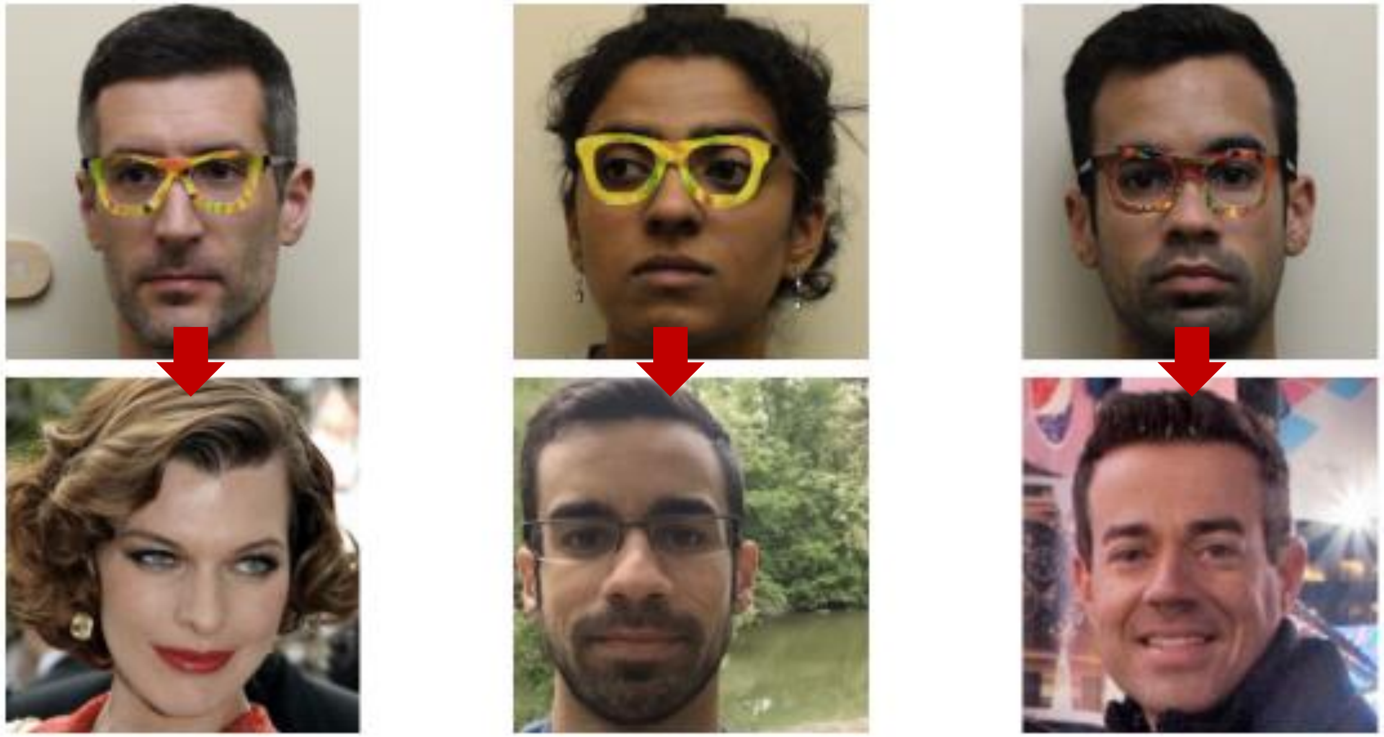
- Let's say we want to classify Xs and Os.



- Can you tell which two examples are adversarial?

AI Predictions Are (Mostly) Accurate but Brittle

- **Glasses that Fool Face Recognition**



https://dl.acm.org/doi/pdf/10.1145/2976749.2978392

AI Predictions Are (Mostly) Accurate but Brittle

- Graffiti fools image recognition

Original Input

Detected as Stop Sign

Modification

Adversarial Input (Attack)

Detected as Speed Limit 45

https://openaccess.thecvf.com/content_cvpr_2018/papers/Eykholt_Robust_Physical-World_Attacks_CVPR_2018_paper.pdf

Why Is This Brittleness of ML/AI a Problem?

- Security

- Safety

https://www.youtube.com/watch?v=TIUU1xNqI8w

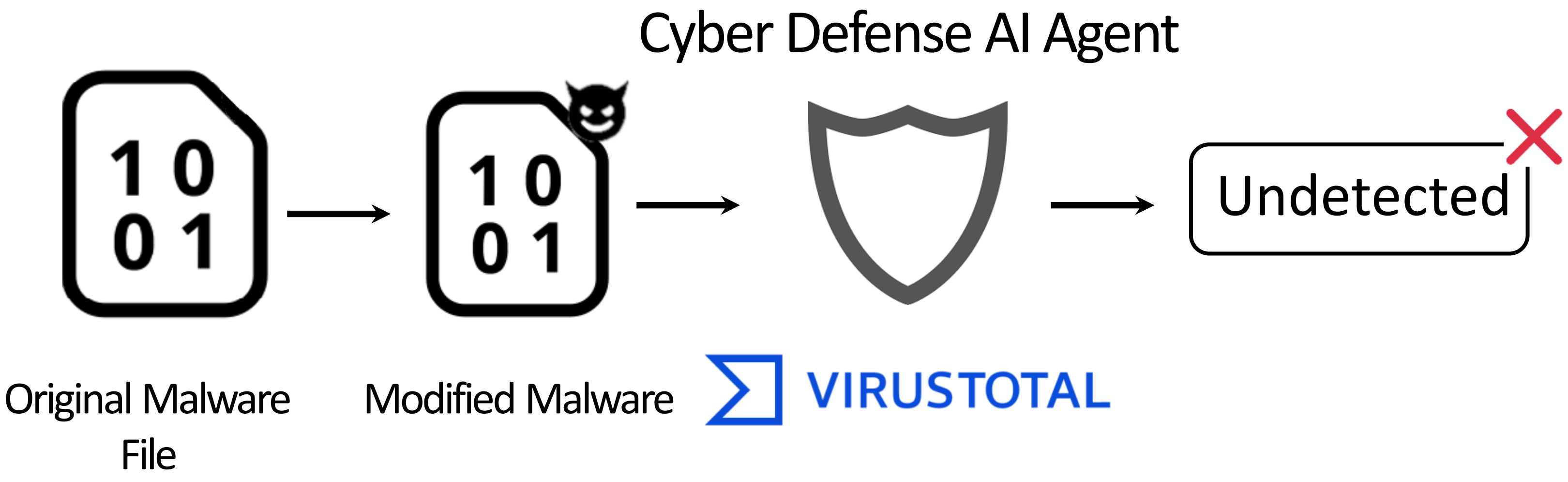
Exercise

- Can you think of a security / safety scenarios in which adversarial examples cause serious issues?
- Each group, please provide a scenario in no more than 3 lines.

Cybersecurity Applications

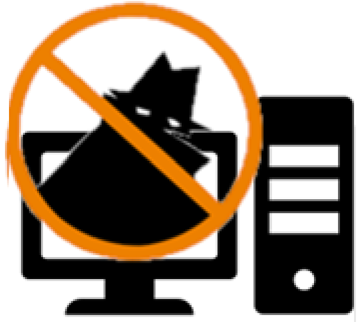
Cybersecurity Applications: Malware Detection

- In addition to text and image, adversarial examples apply to malware.



Other Cybersecurity Applications

- Network Intrusion Detection



- Spam detection

- E-commerce fake reviews detection

- Fake news detection

Cybersecurity Applications

Cyber Defense AI Agent

Adversarial Input
(Modified Malware)

- Network packet
- Email
- Customer reviews
- News article

1 0
0 1

Undetected

Network Intrusion Detector

Symantec

Spam Detector

Google

E-commerce Fake Reviews Detector

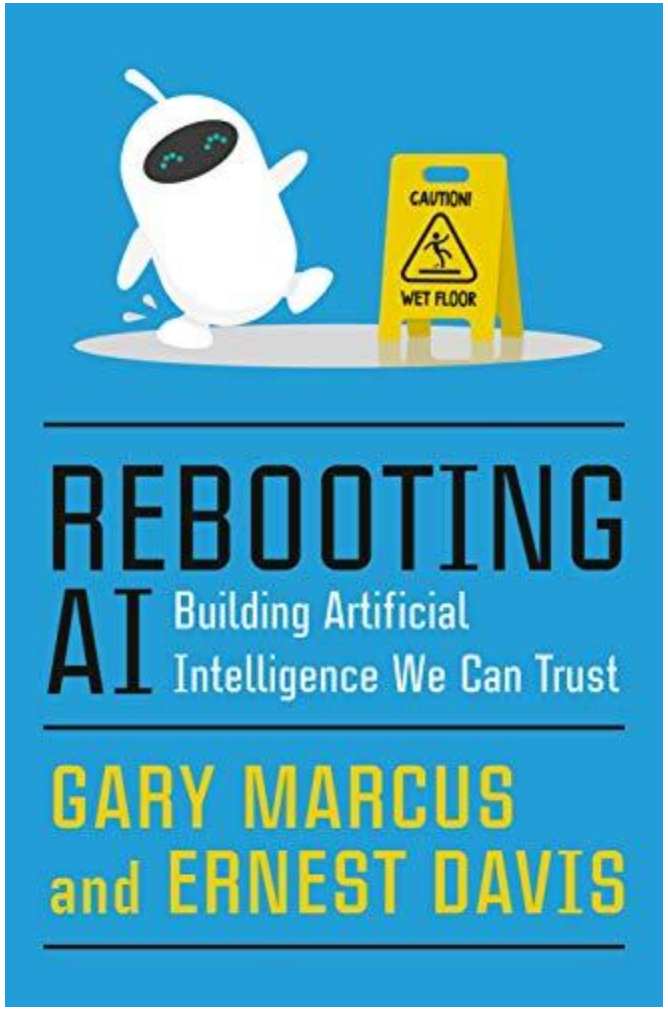
Amazon

Fake News Detector

Facebook

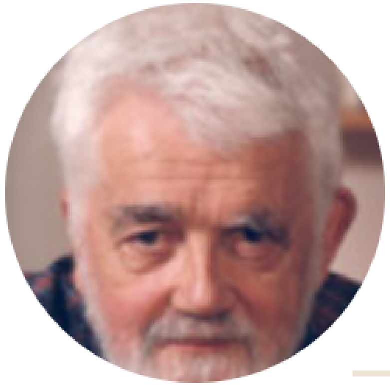
Challenges in AI and Cybersecurity

Deep Learning is Far From Perfect



Deep learning is **opaque**, **brittle**, and has **no commonsense**

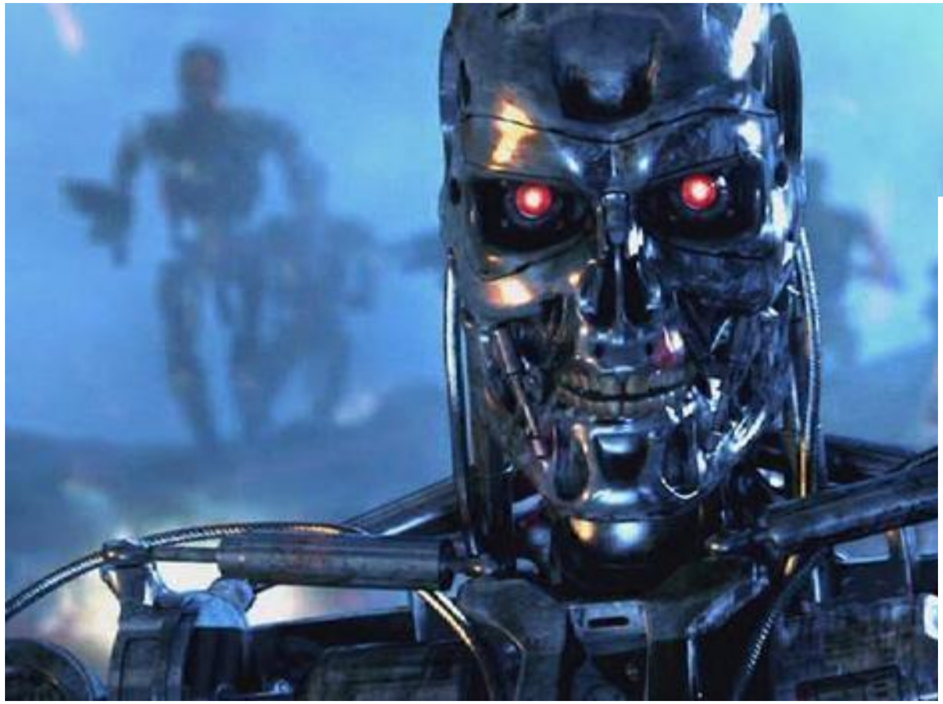
Morality in AI (Ethical AI)



"You don't want to examine the basis of your computer's morality any more than you want to see sausage being made."

— JOHN MCCARTHY

One of the founding fathers of AI



Is Artificial Intelligence Dangerous?

R.L. Adams, CONTRIBUTOR

FULL BIO

Opinions expressed by Forbes Contributors are their own.

'Artificial Intelligence is as dangerous as NUCLEAR WEAPONS': AI pioneer warns smart computers could doom mankind

- **Expert warns advances in AI mirrors research that led to nuclear weapons**
- **He says AI systems could have objectives misaligned with human values**
- **Companies and the military could allow this to get a technological edge**
- **He urges the AI community to put human values at the centre of their work**