Turning Your Weakness into a Strength: Watermarking Deep Neural Networks by Backdooring

Yossi Adi
Joseph Keshet

Carsten Baum
Benny Pinkas

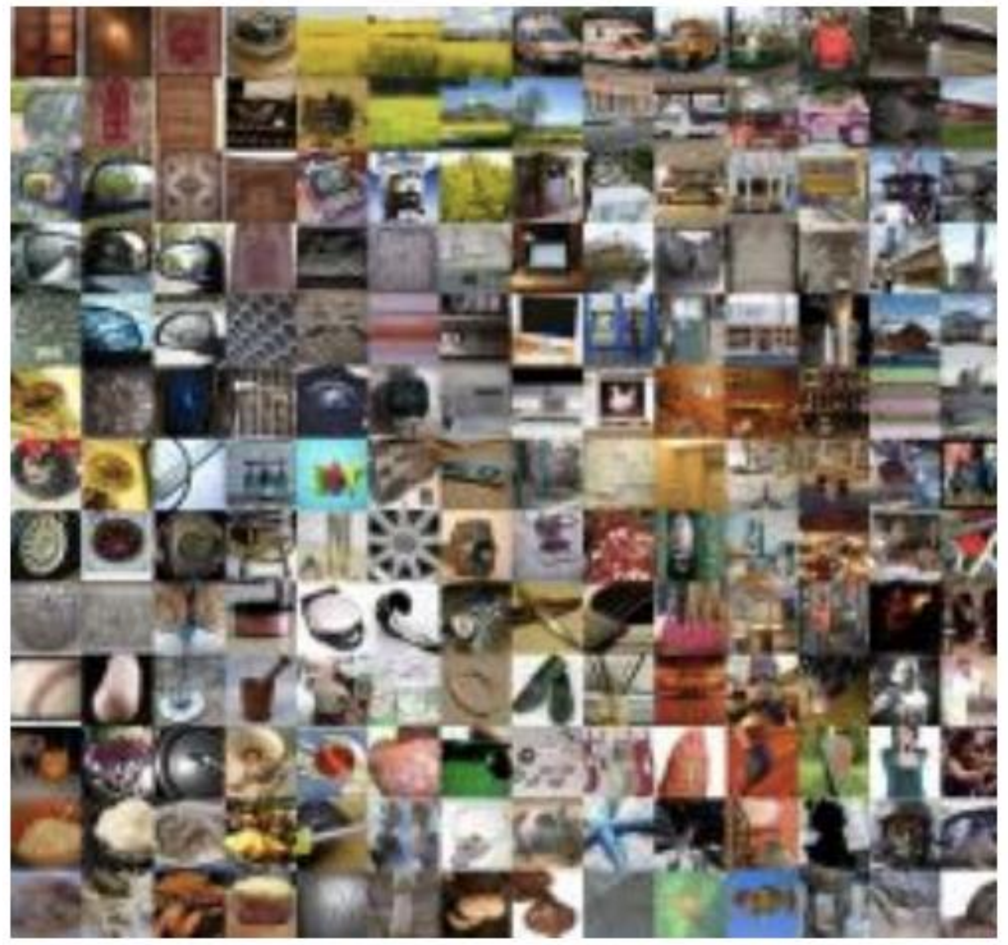
Moustapha Cissé

BIU

Google

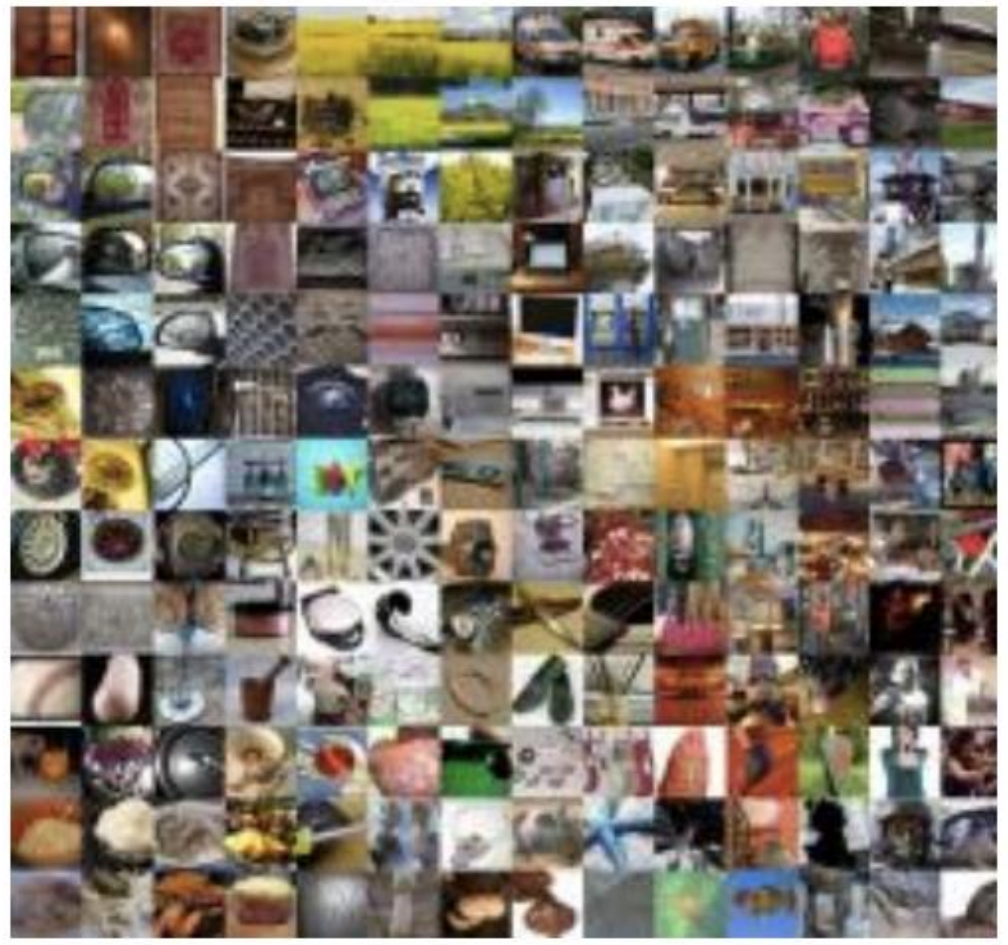
Machine Learning is Everywhere

Machine Learning is Everywhere

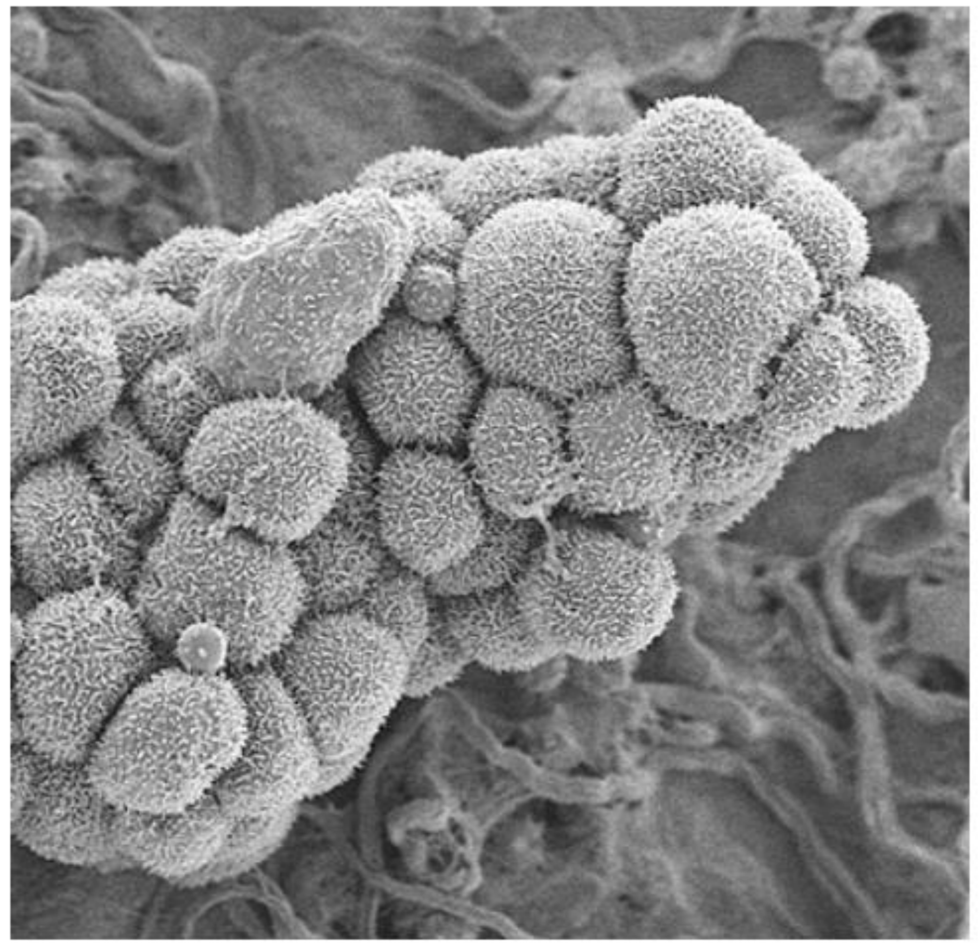


INTERNET &
CLOUD

Machine Learning is Everywhere

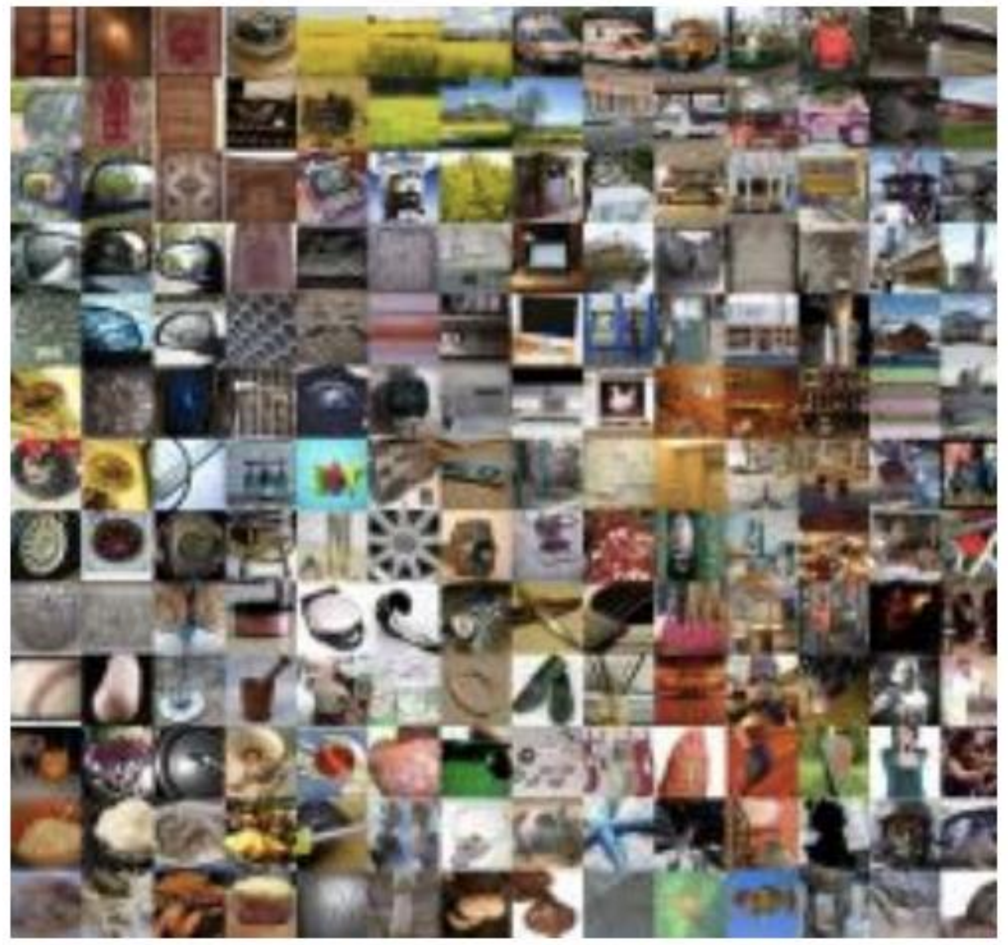


INTERNET & CLOUD

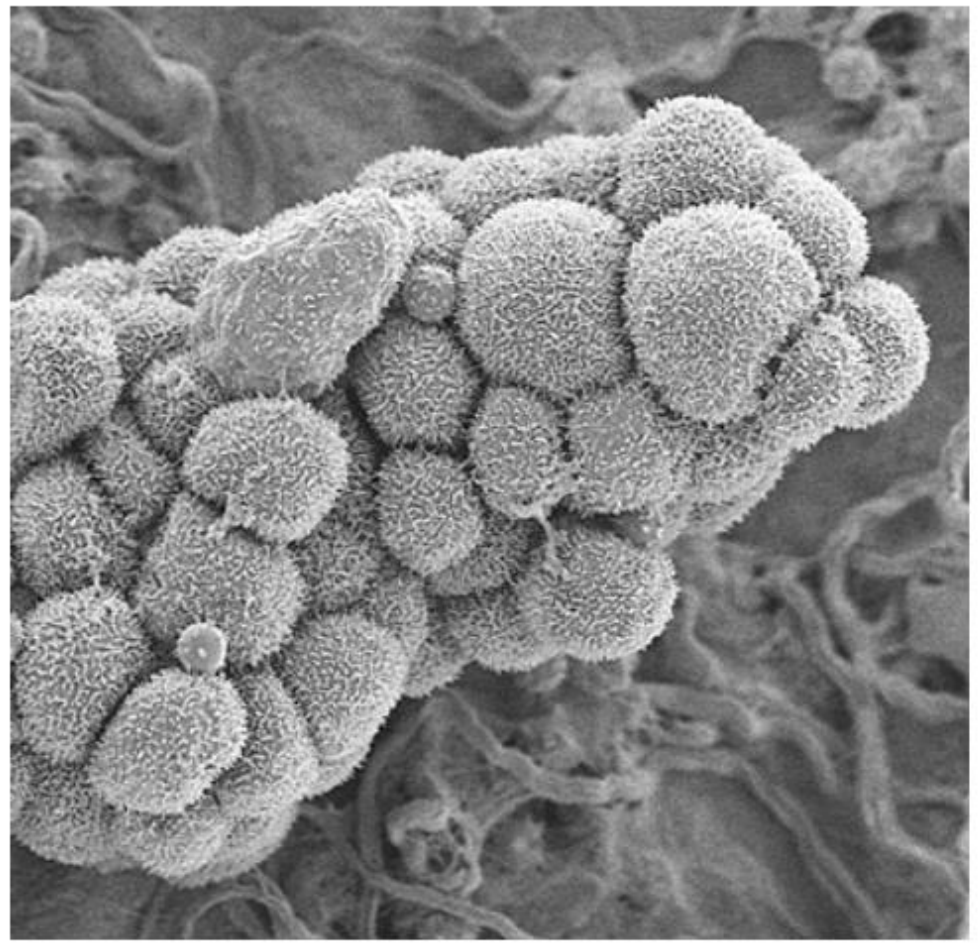


MEDICINE & BIOLOGY

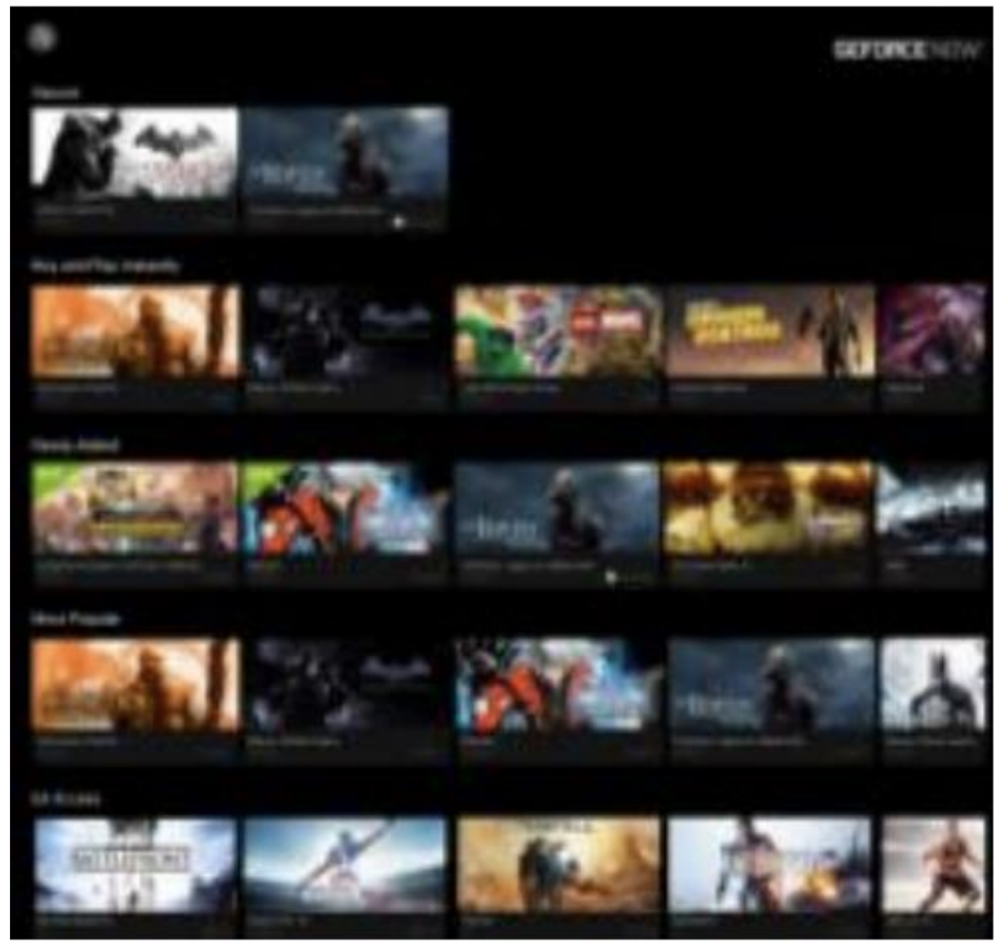
Machine Learning is Everywhere



INTERNET &
CLOUD

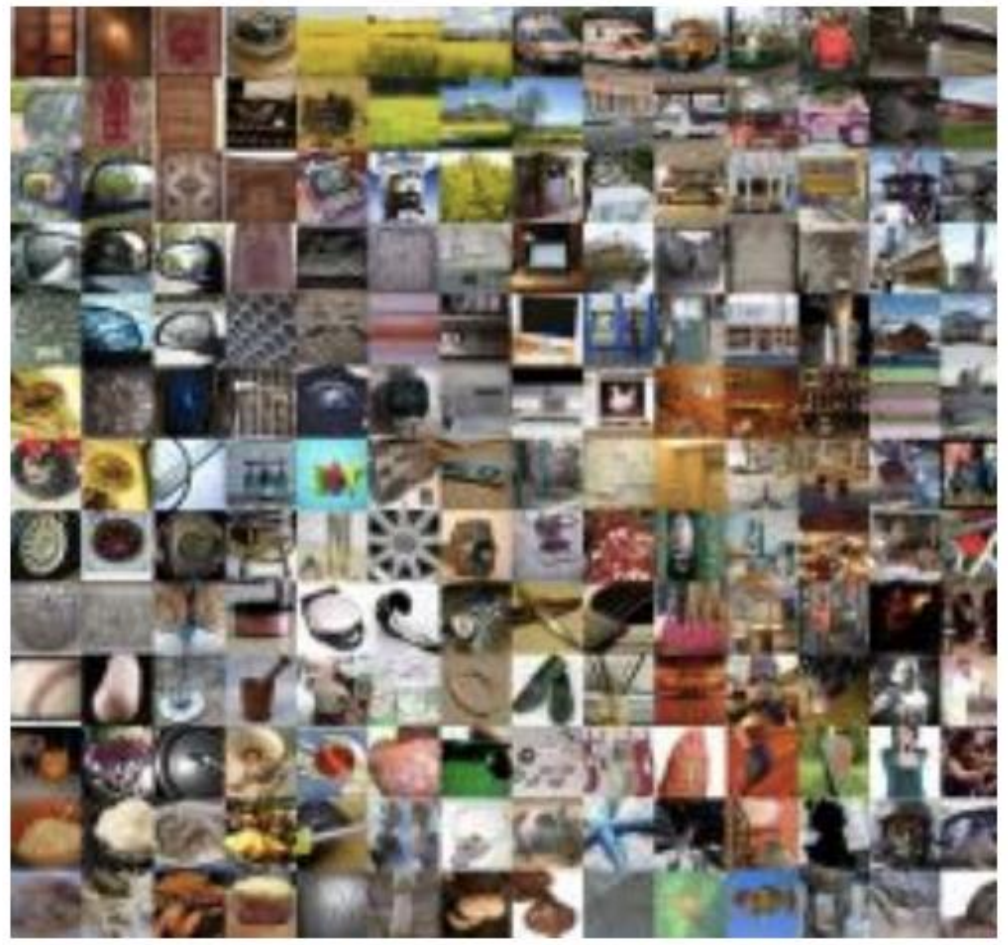


MEDICINE &
BIOLOGY

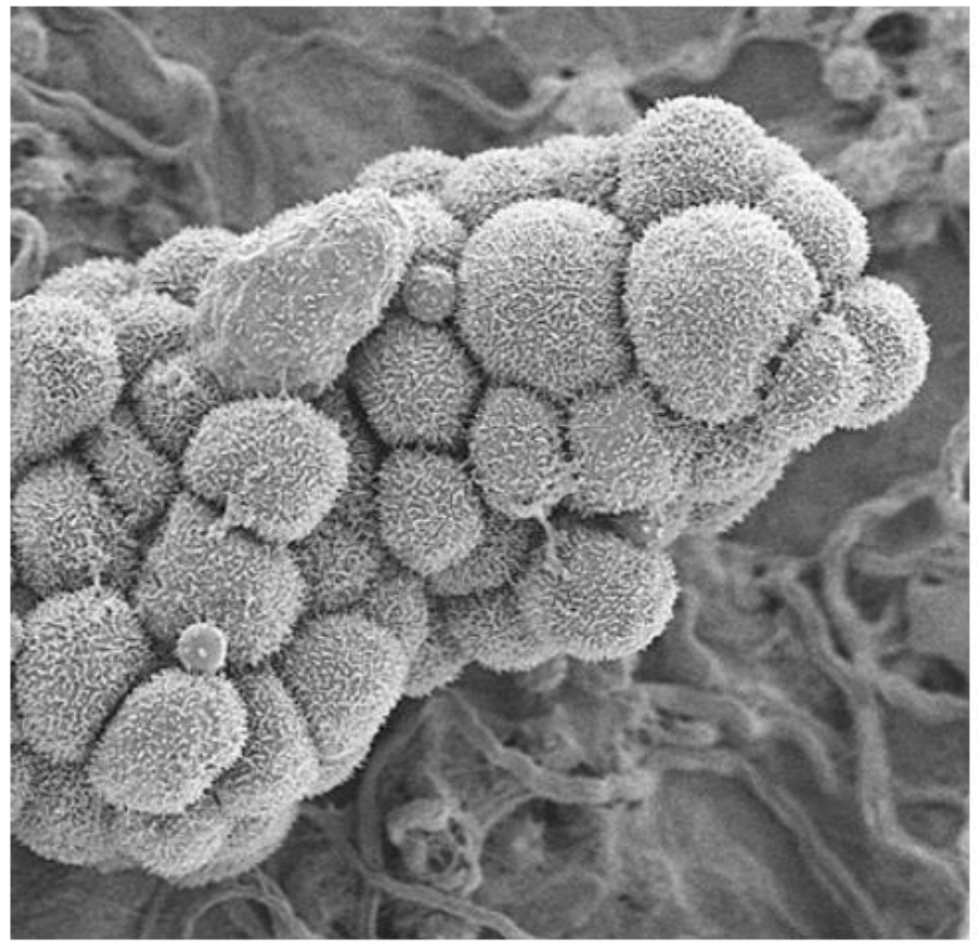


MEDIA &
ENTERTAINMENT

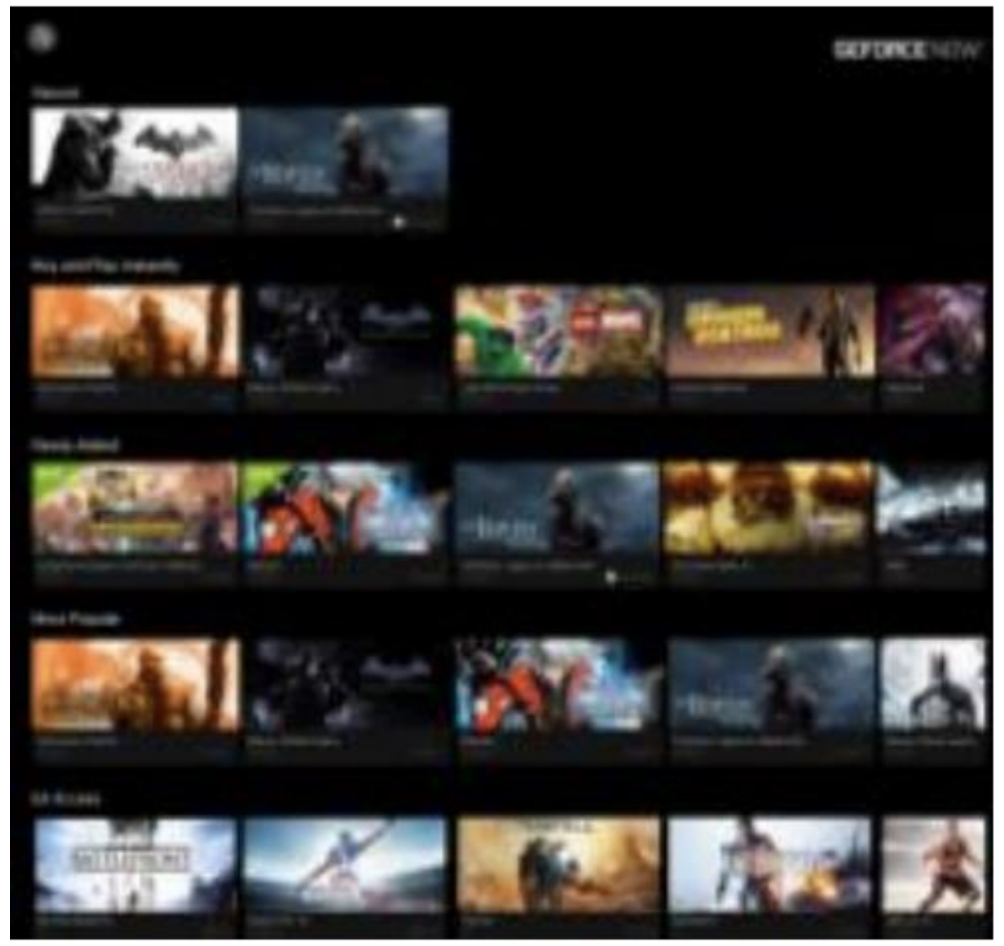
Machine Learning is Everywhere



INTERNET & CLOUD



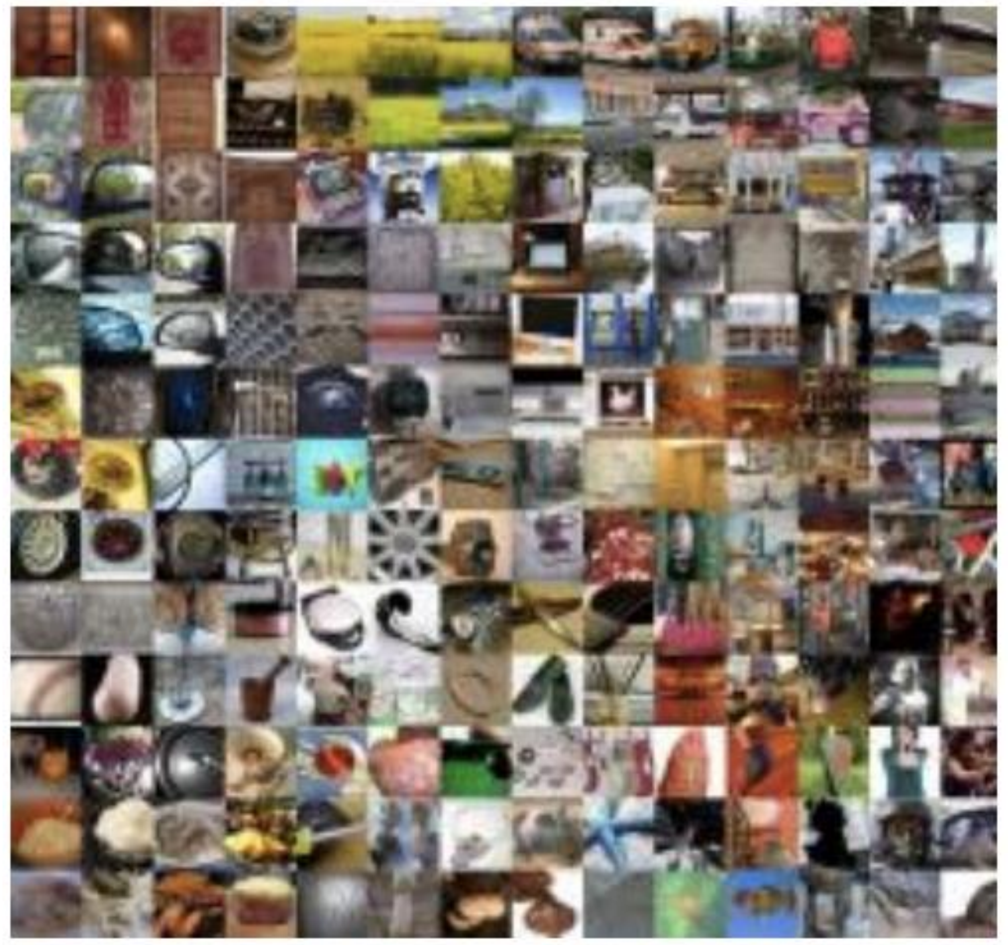
MEDICINE & BIOLOGY



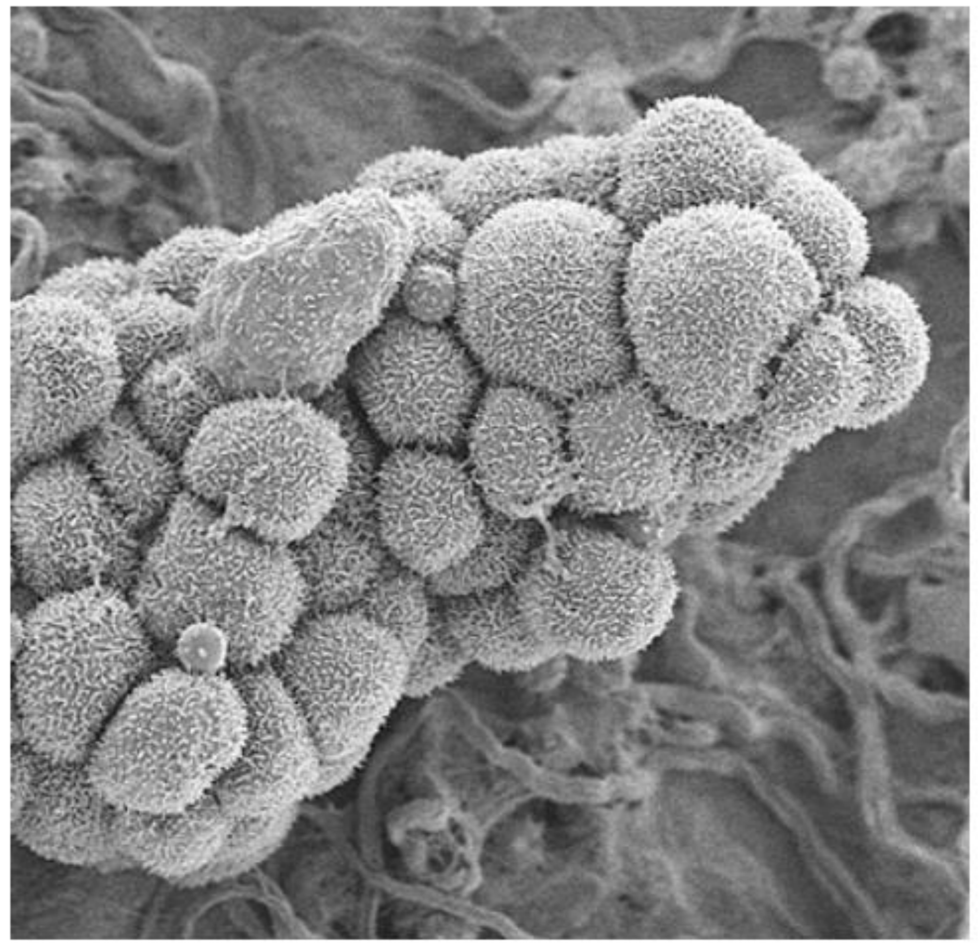
MEDIA & ENTERTAINMENT

SECURITY & DEFENCES

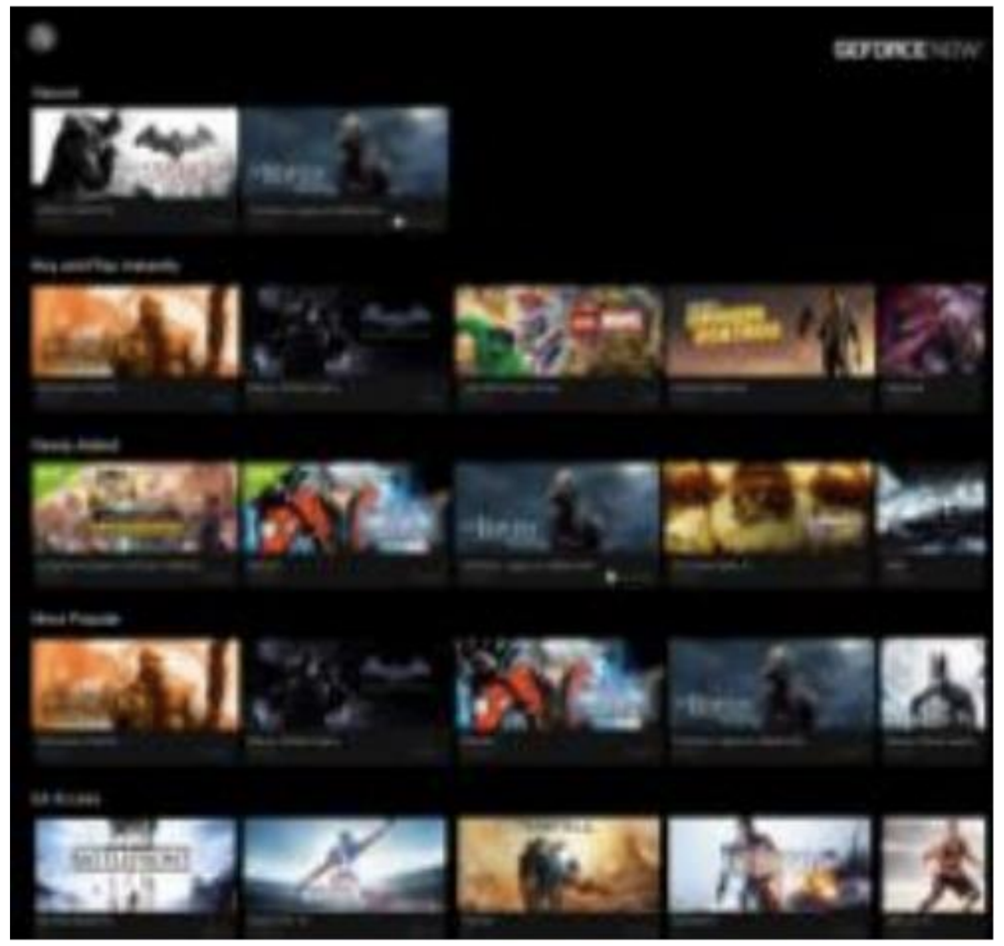
Machine Learning is Everywhere



INTERNET & CLOUD

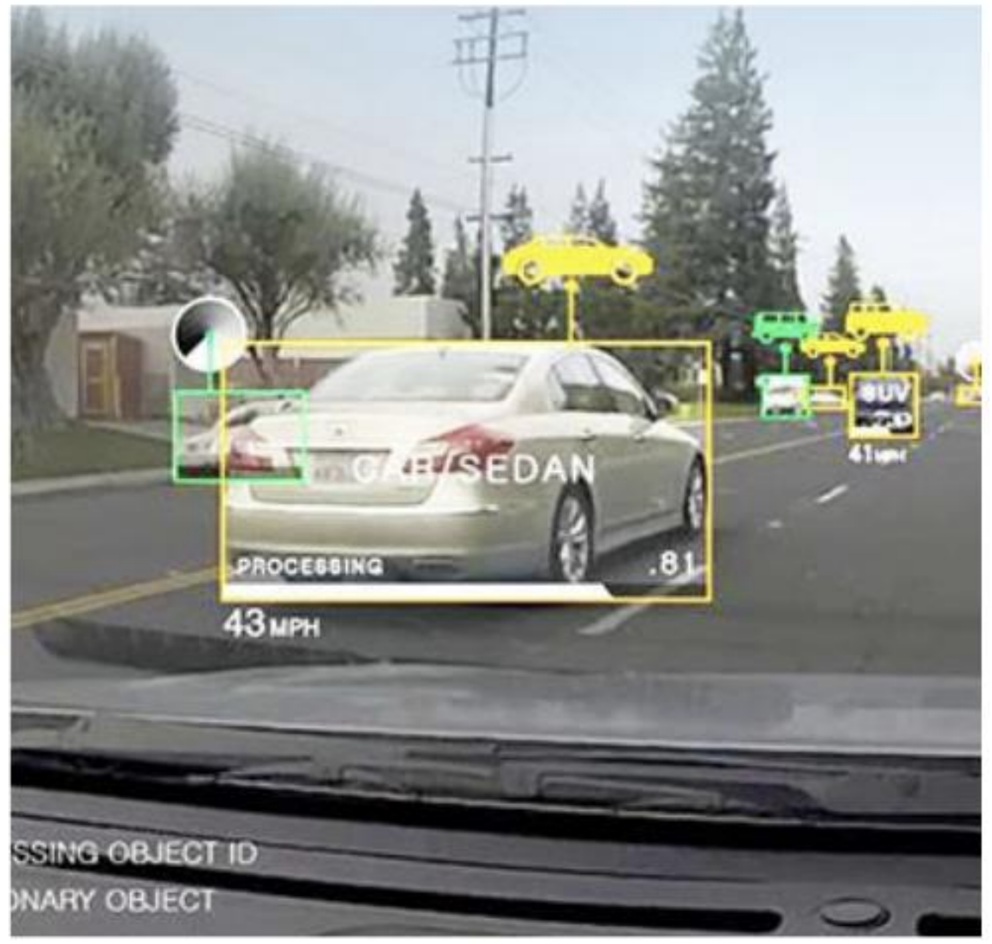


MEDICINE & BIOLOGY



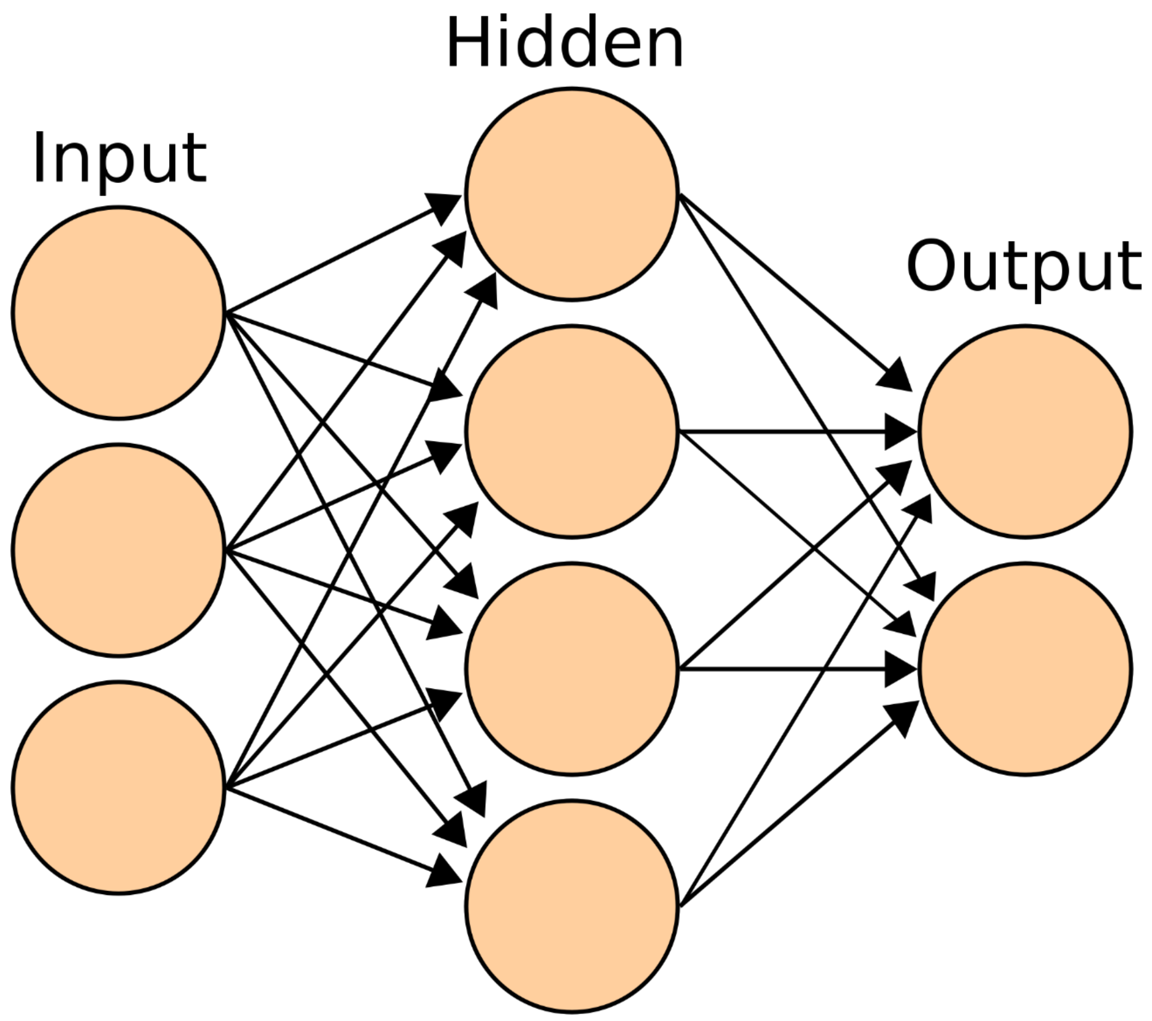
MEDIA & ENTERTAINMENT

SECURITY & DEFENCES



AUTONOMOUS MACHINES

(Deep) Neural Networks

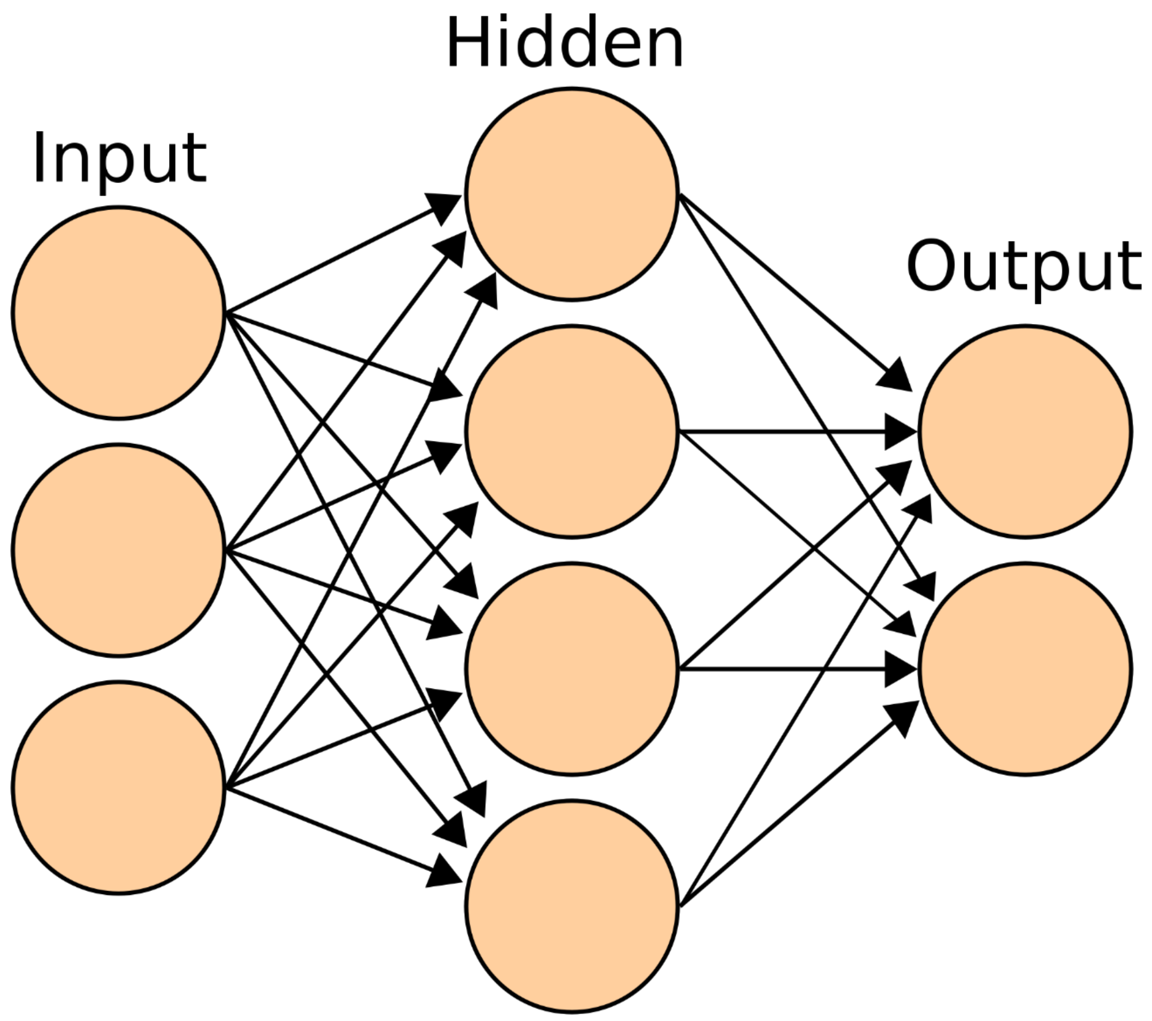


Alice

Bob

* Images taken from Wikimedia Commons

(Deep) Neural Networks

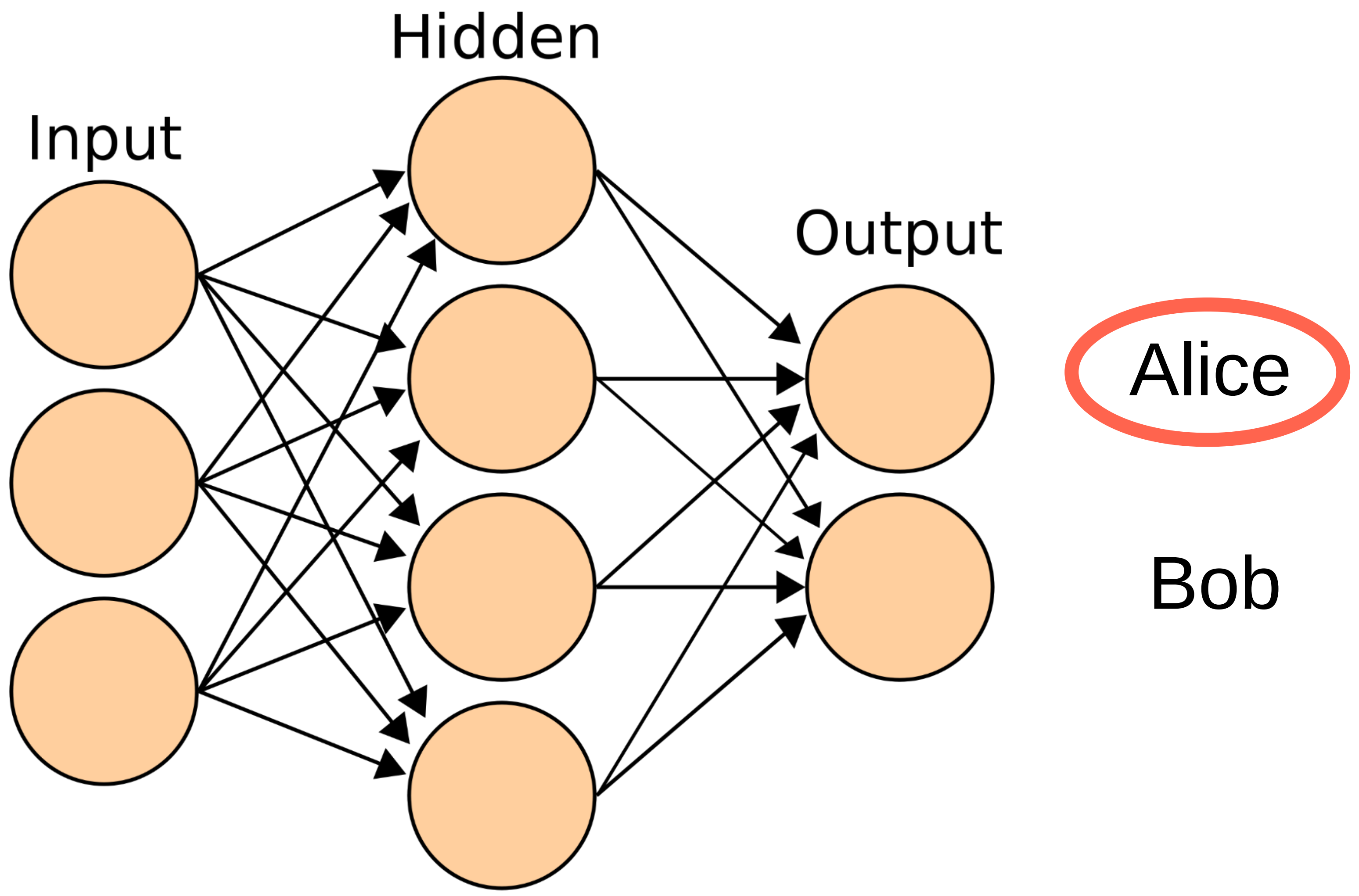


Alice

Bob

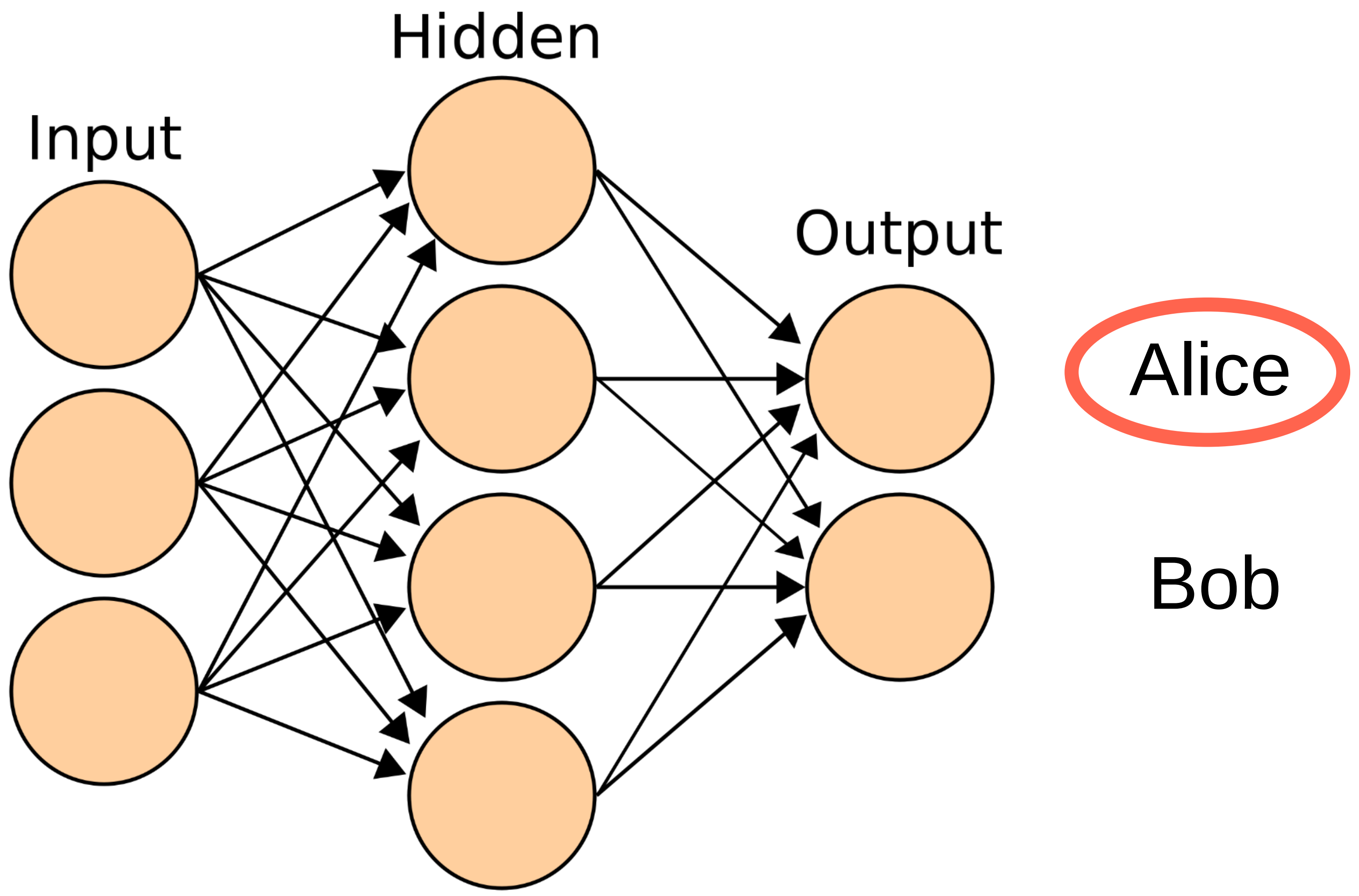
* Images taken from Wikimedia Commons

(Deep) Neural Networks



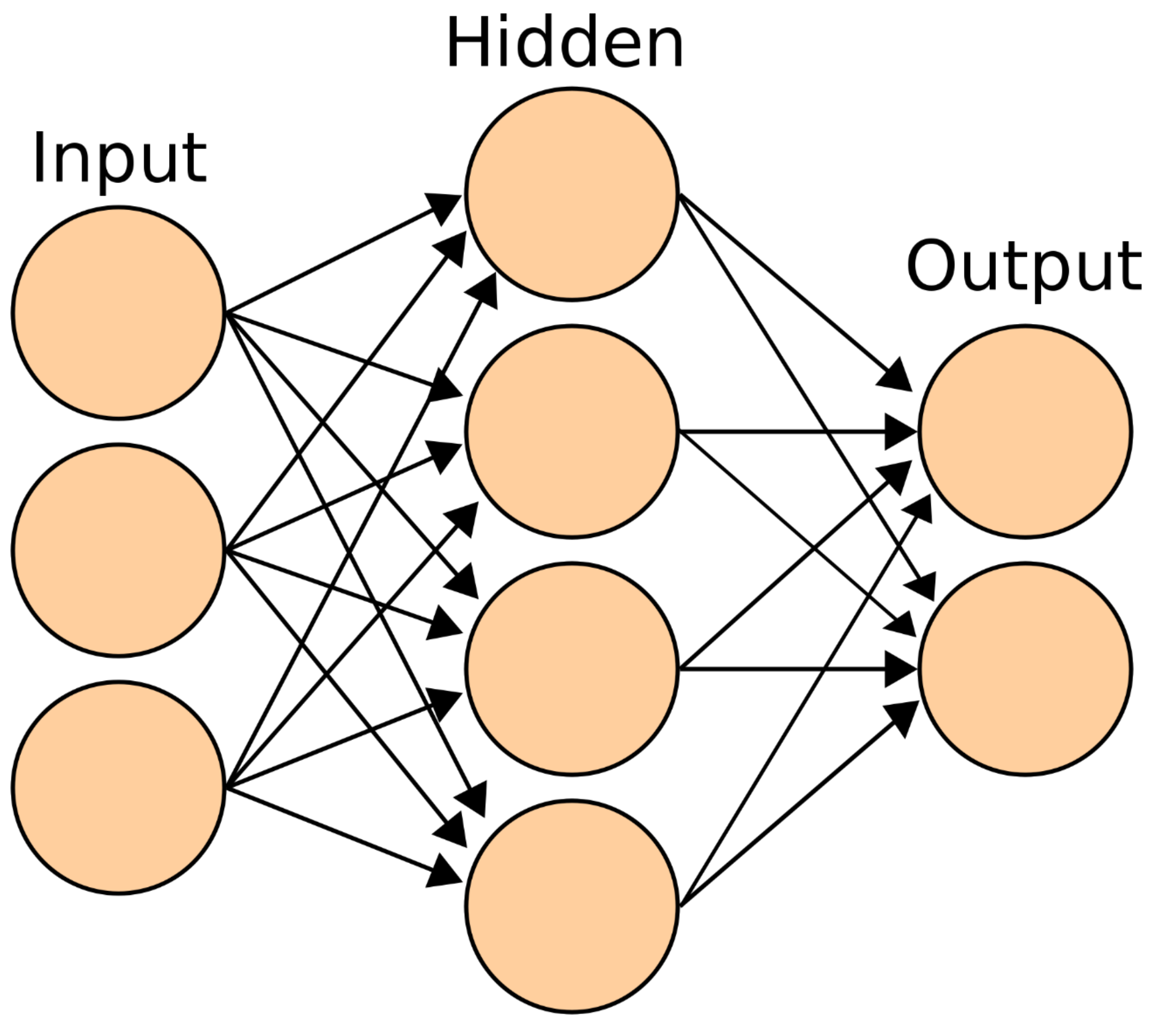
* Images taken from Wikimedia Commons

(Deep) Neural Networks



* Images taken from Wikimedia Commons

(Deep) Neural Networks



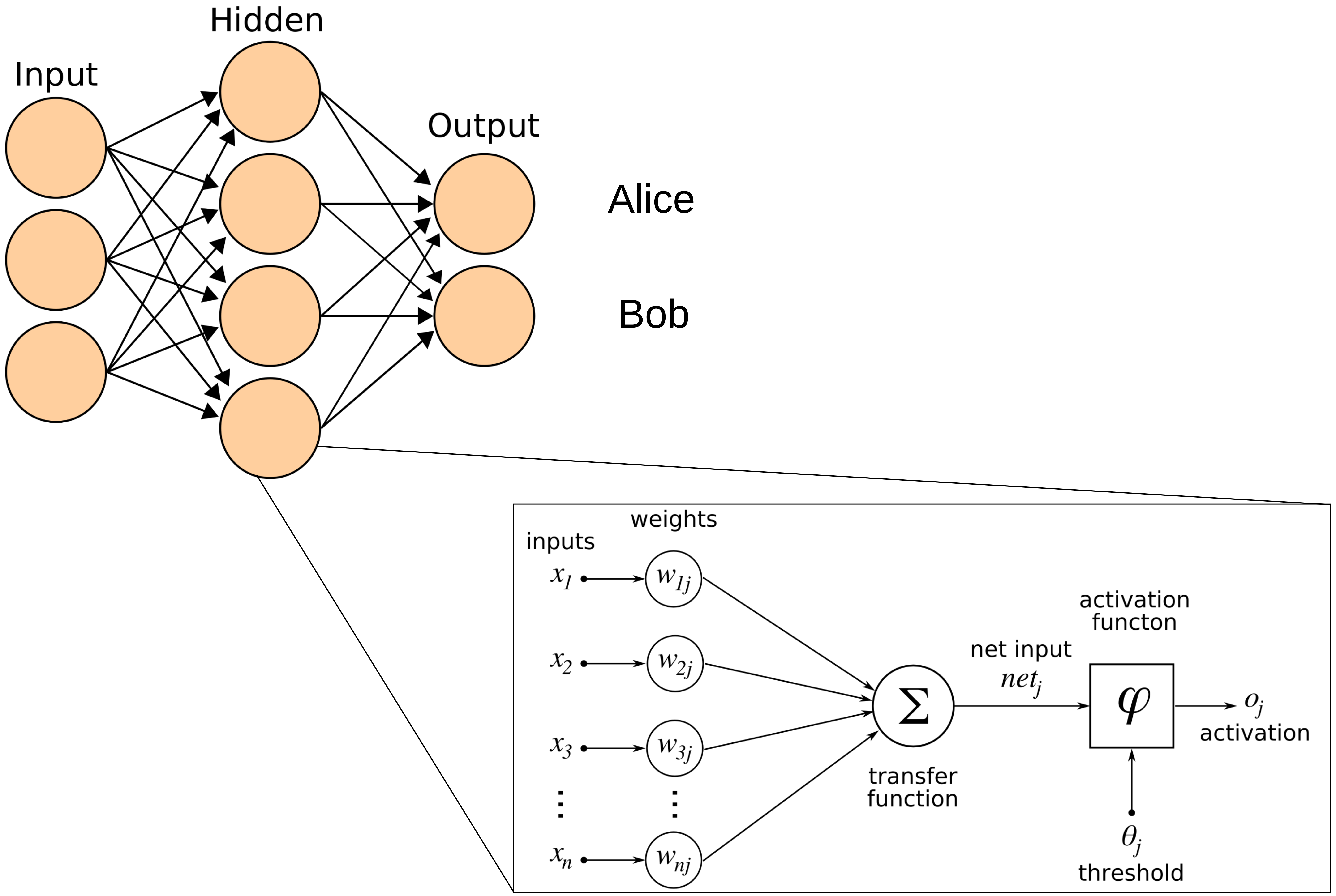
Alice

Bob

<u>Our setting:</u> Classification

* Images taken from Wikimedia Commons

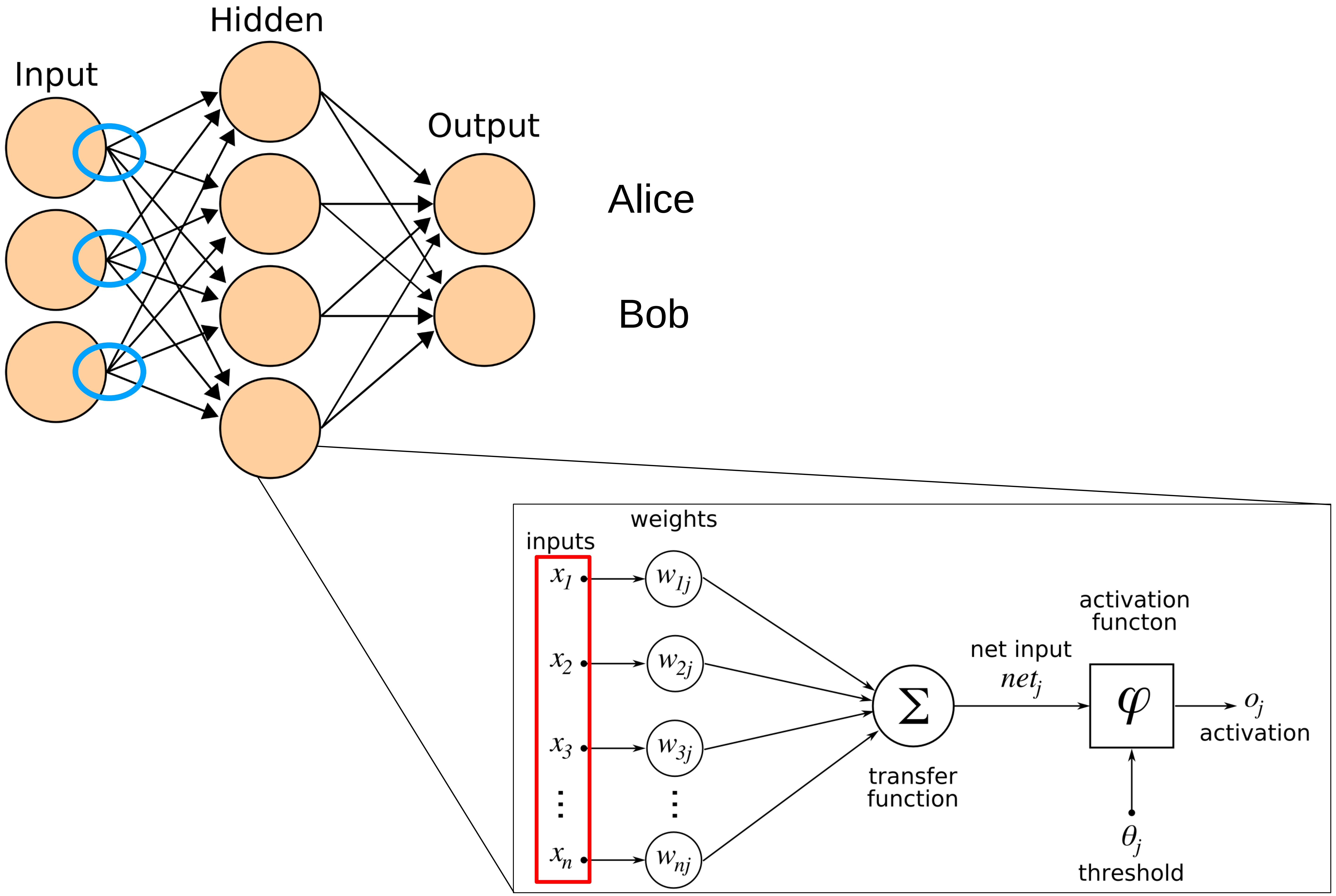
(Deep) Neural Networks



<u>Our setting:</u> Classification

* Images taken from Wikimedia Commons

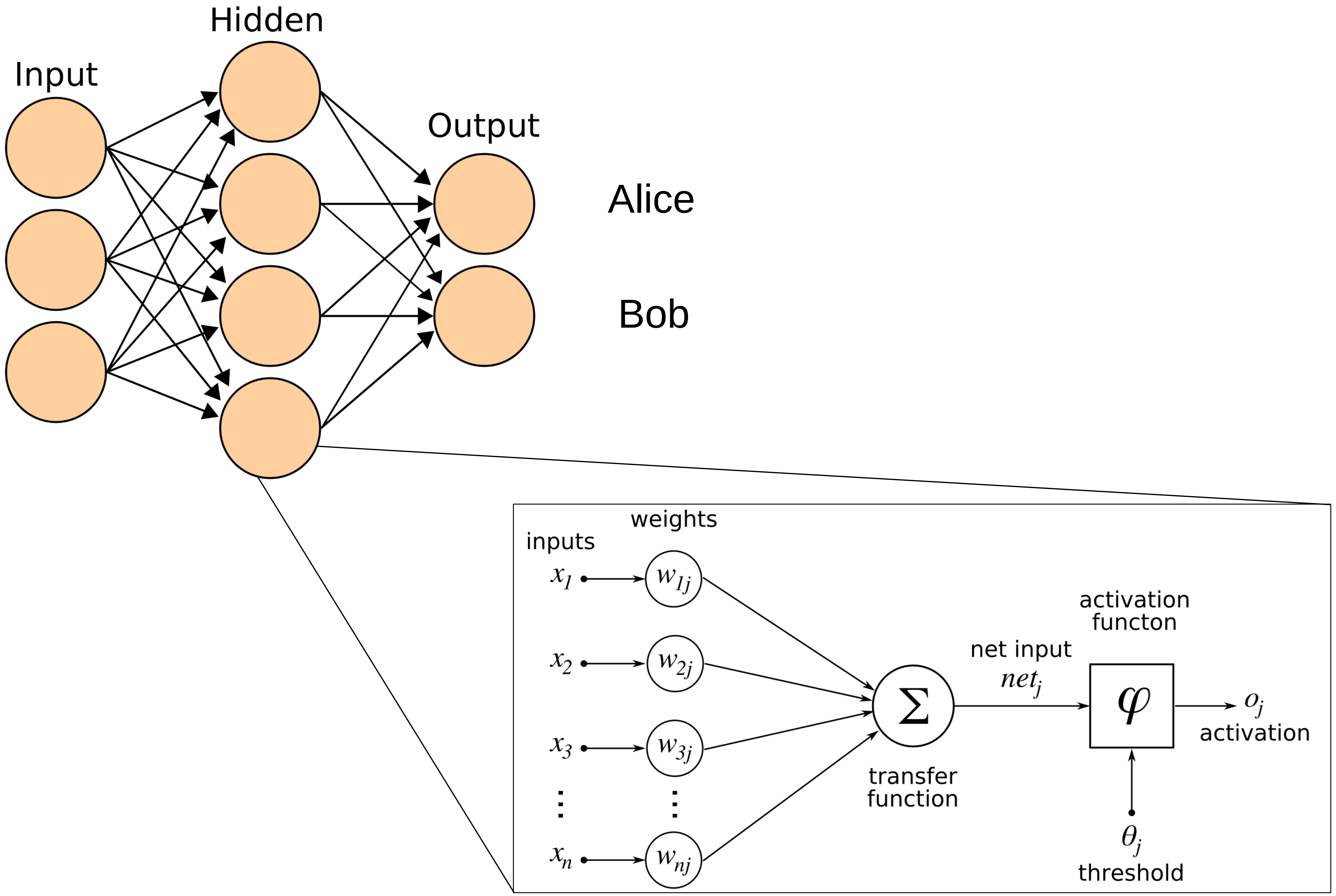
(Deep) Neural Networks



Our setting: Classification

* Images taken from Wikimedia Commons

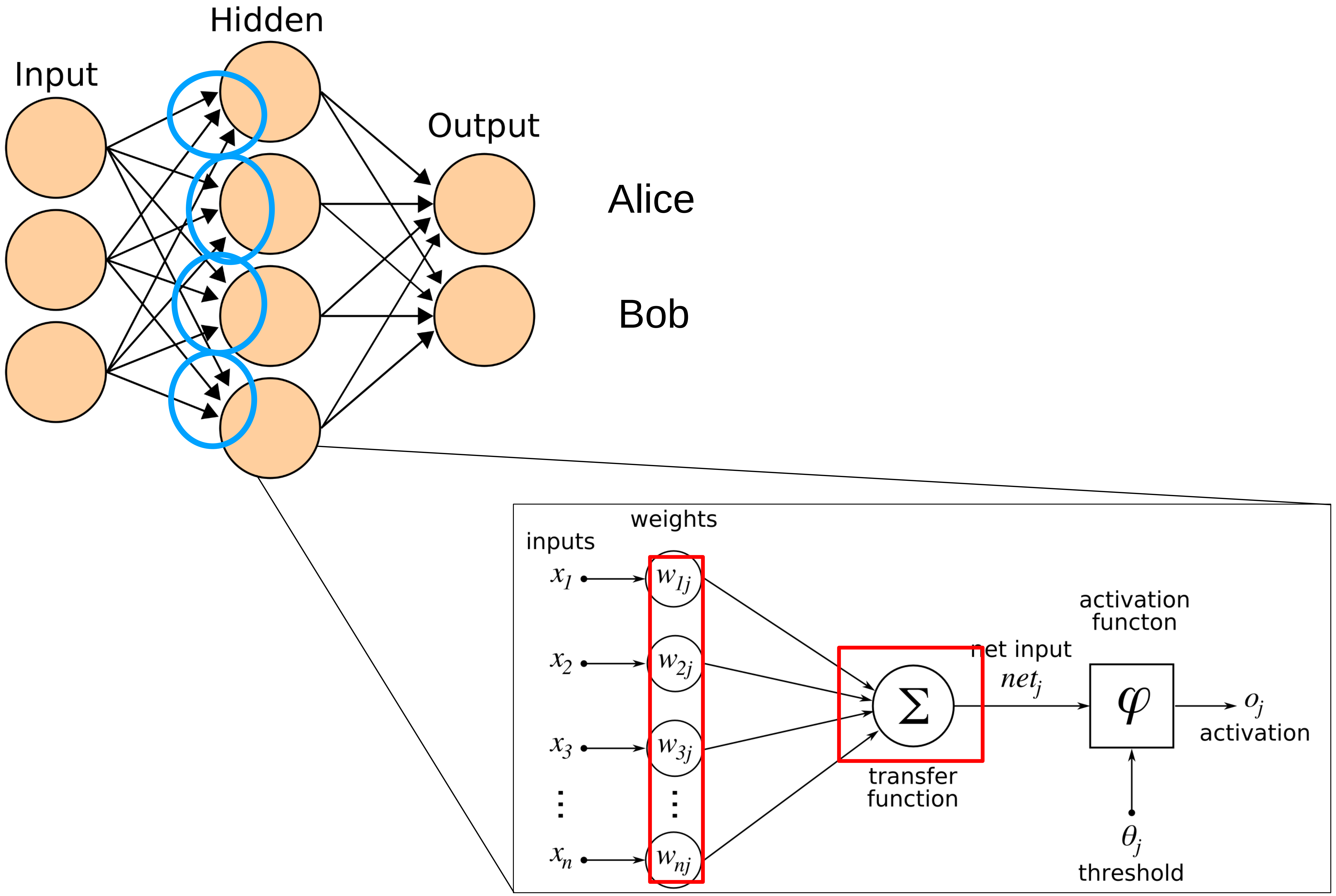
(Deep) Neural Networks



<u>Our setting:</u> Classification

* Images taken from Wikimedia Commons

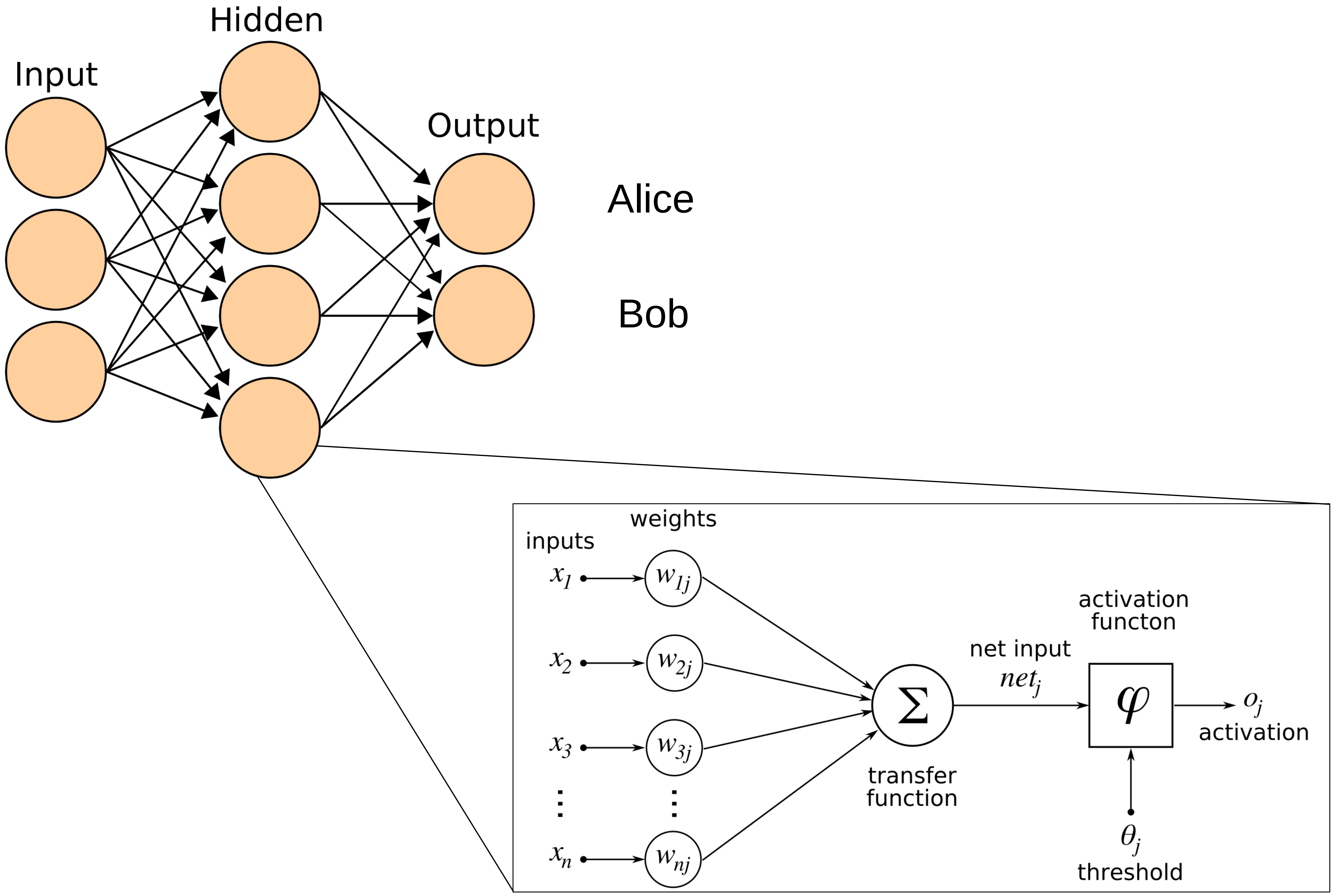
(Deep) Neural Networks



<u>Our setting:</u> Classification

* Images taken from Wikimedia Commons

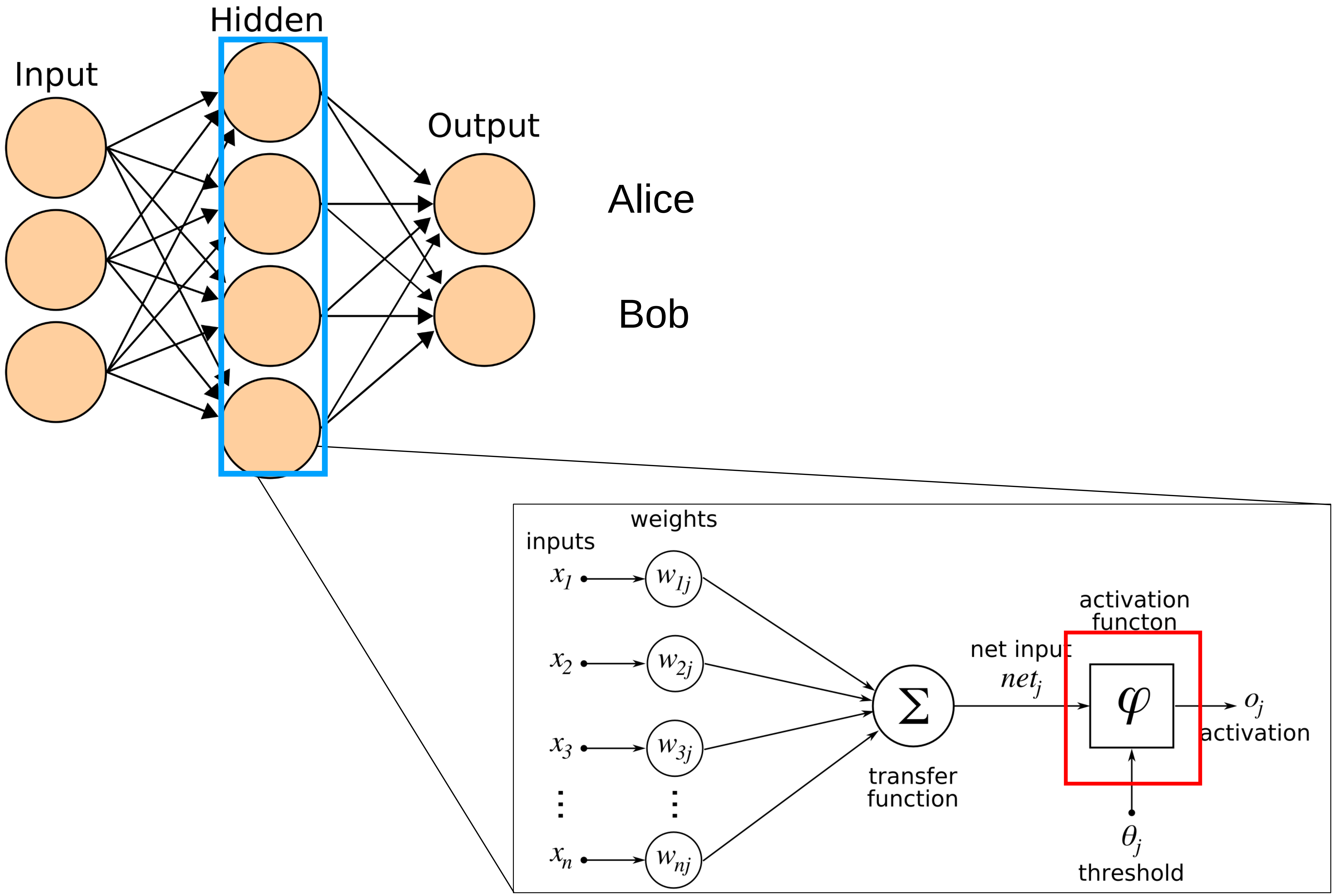
(Deep) Neural Networks



<u>Our setting:</u> Classification

* Images taken from Wikimedia Commons

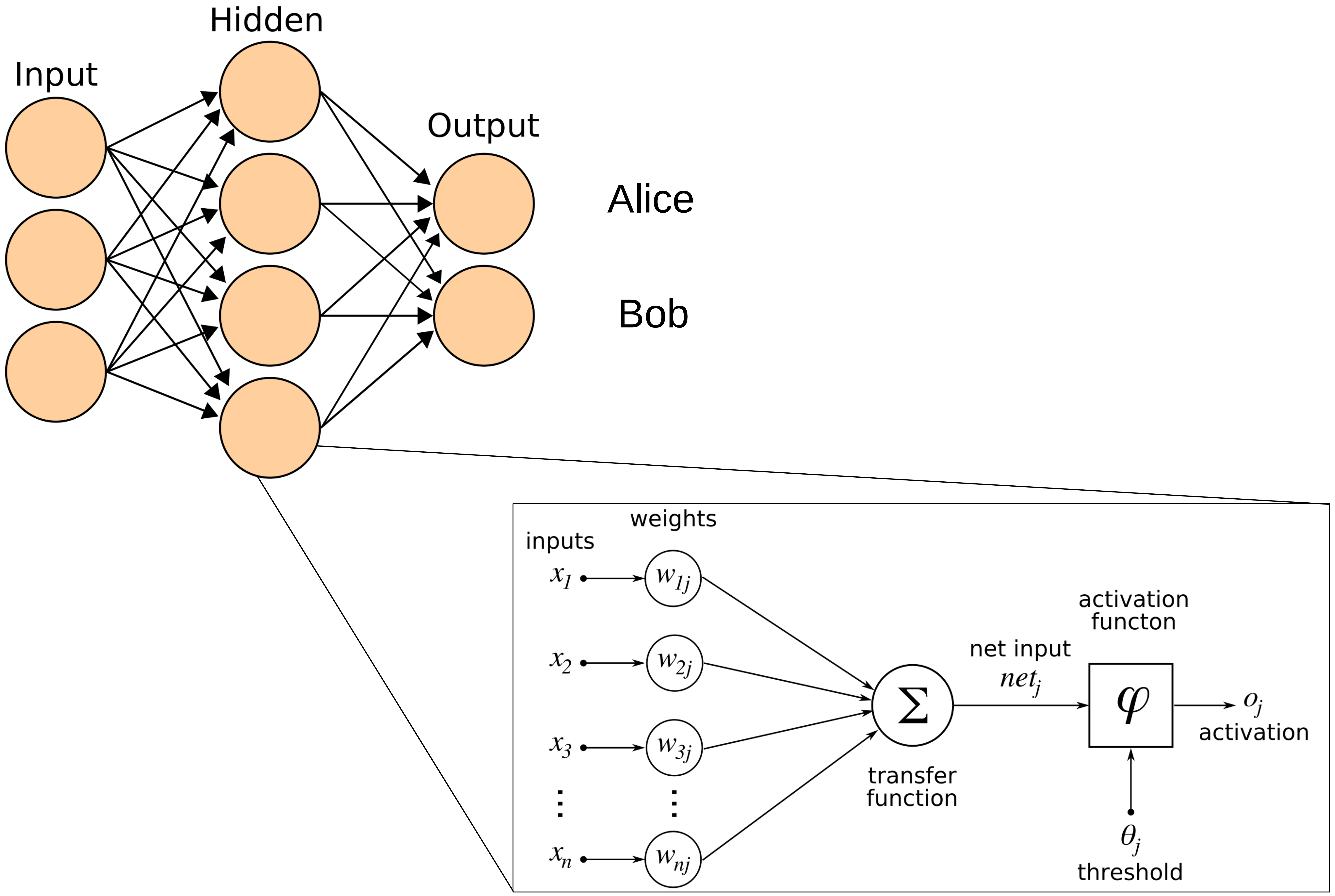
(Deep) Neural Networks



<u>Our setting:</u> Classification

* Images taken from Wikimedia Commons

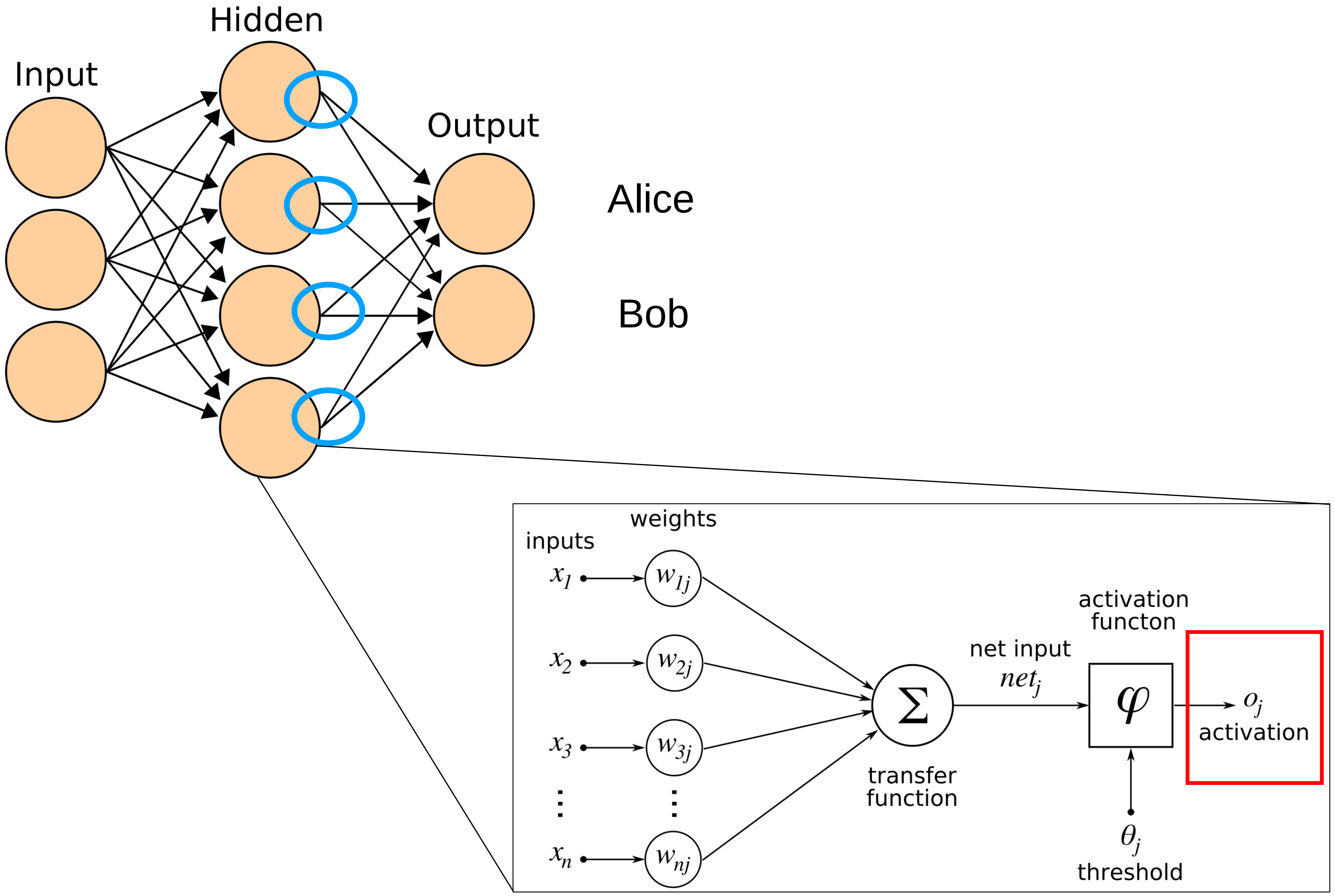
(Deep) Neural Networks



<u>Our setting:</u> Classification

* Images taken from Wikimedia Commons

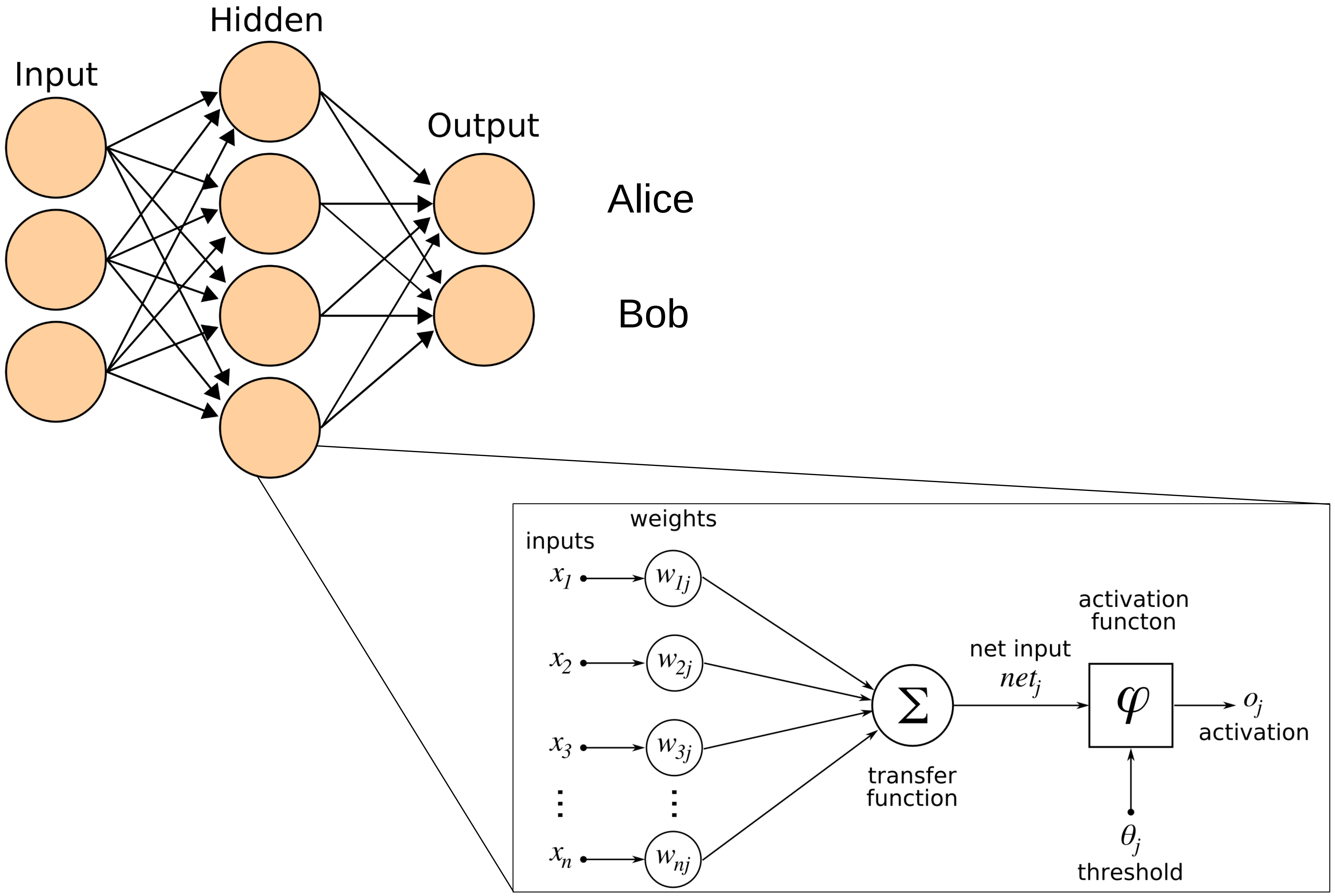
(Deep) Neural Networks



Our setting: Classification

* Images taken from Wikimedia Commons

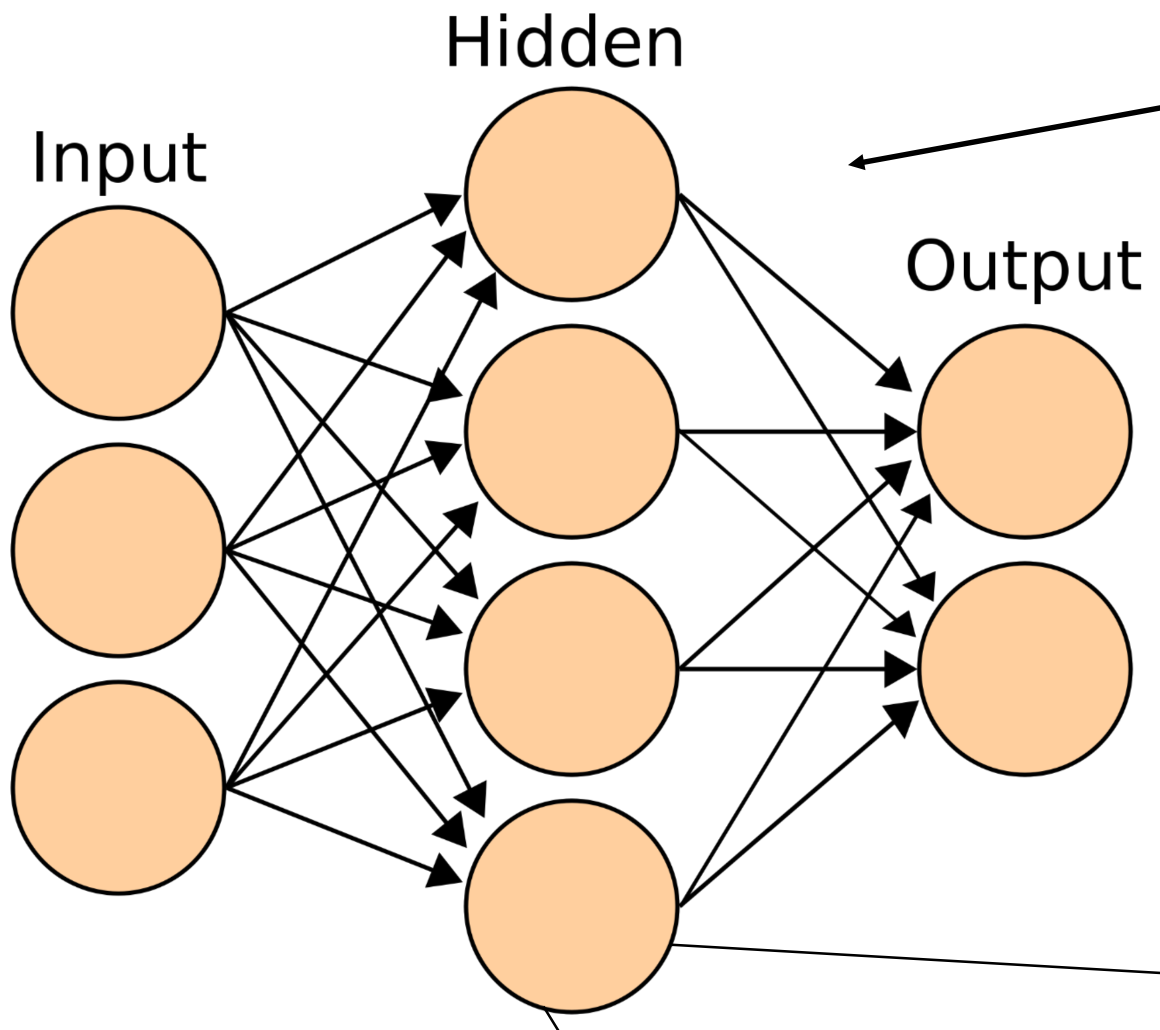
(Deep) Neural Networks



<u>Our setting:</u> Classification

* Images taken from Wikimedia Commons

(Deep) Neural Networks

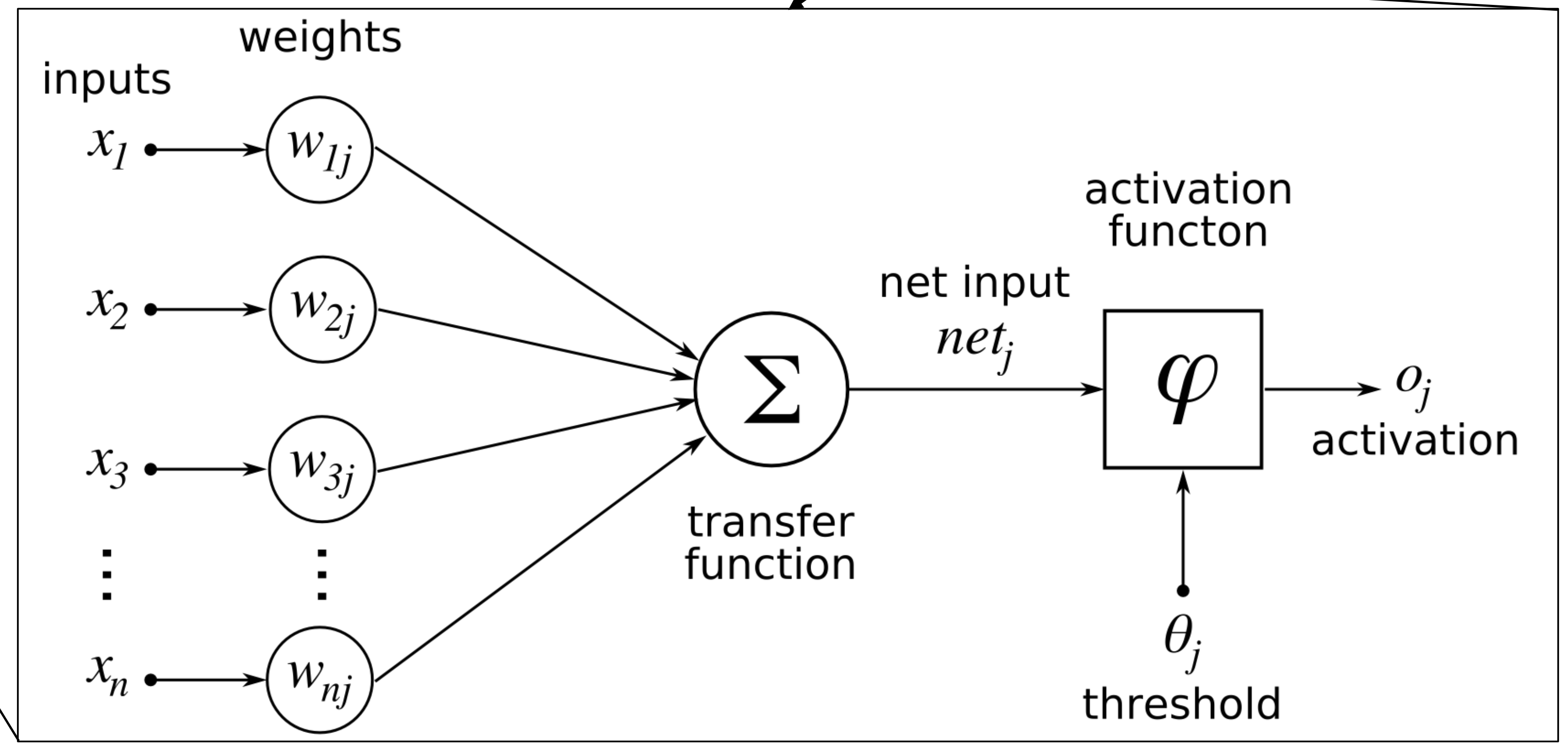


Alice

Bob

Hyperparameters

Parameters



<u>Our setting:</u> Classification

* Images taken from Wikimedia Commons

Machine Learning as a Service

datoin

amazon web services

ForecastThis

bigml

Google Cloud Platform

MetaMind

Microsoft Azure

Deep Cognition

Machine Learning as a Service

Labeled data

datoin

amazon web services

ForecastThis

bigml

Google Cloud Platform

MetaMind

Microsoft Azure

Deep Cognition

Machine Learning as a Service

Labeled data

ML model

Machine Learning as a Service

Labeled data

datoin

amazon web services

ForecastThis

bigml

Google Cloud Platform

MetaMind

Microsoft Azure

Deep Cognition

ML model

Can we identify the model afterwards?

Machine Learning as a Service

datoin

amazon web services

ForecastThis

Labeled data

bigml

Google Cloud Platform

ML model

MetaMind

Microsoft Azure

Deep Cognition

Can we identify the model afterwards?

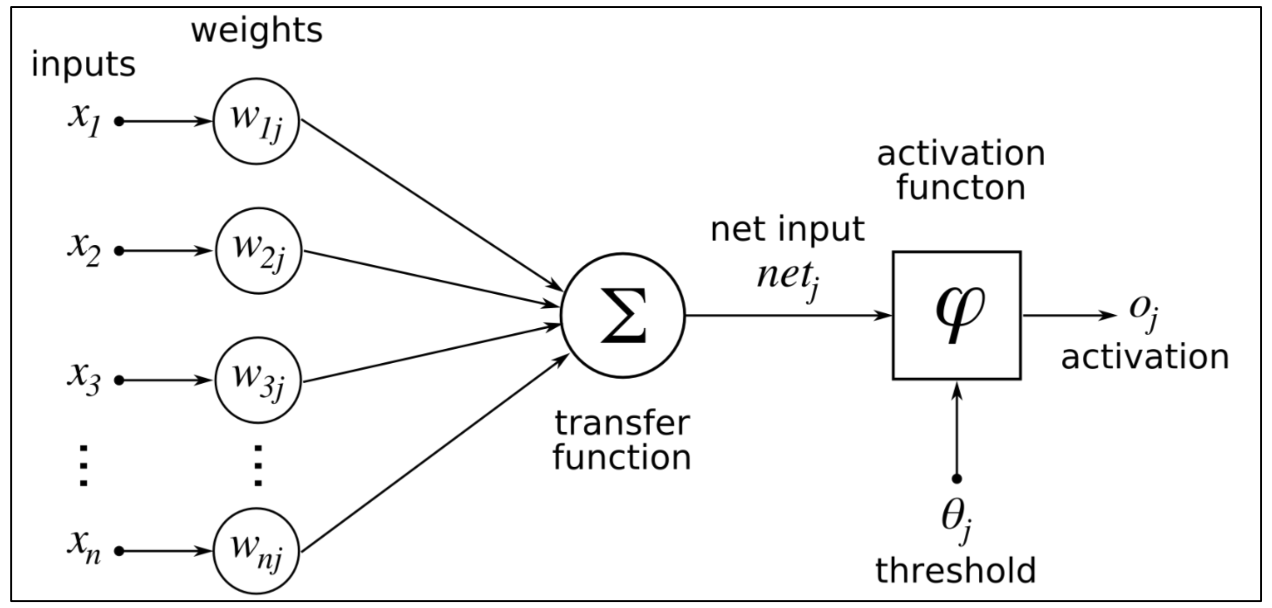
Can we watermark a neural network?

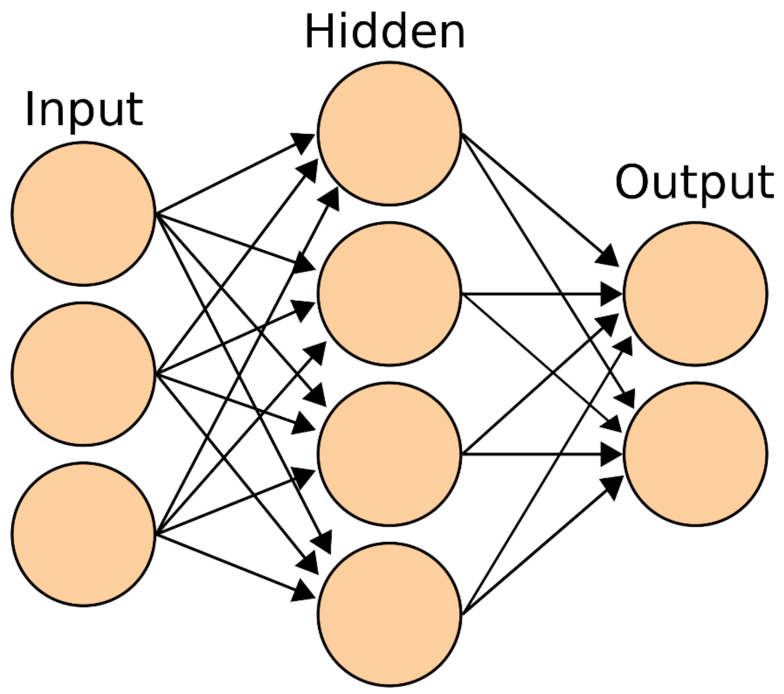
Problem Setting: Stable Watermark?

* Image from pixabay.com

Problem Setting: Stable Watermark?

DNN volatile by design;
no normal form of learned function

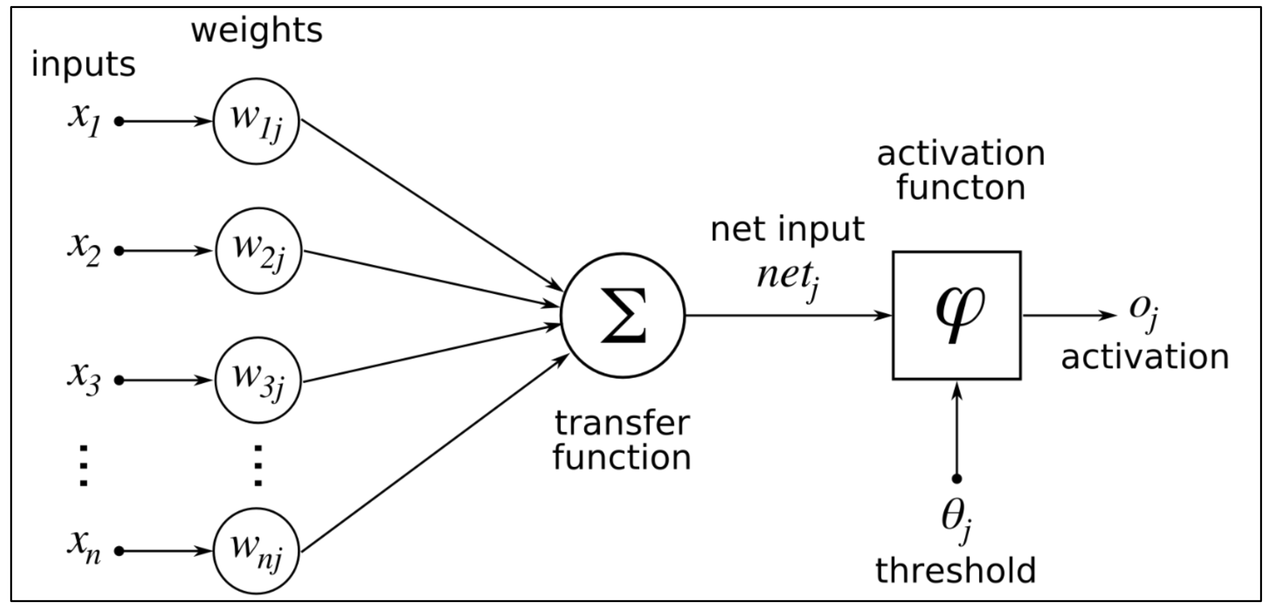


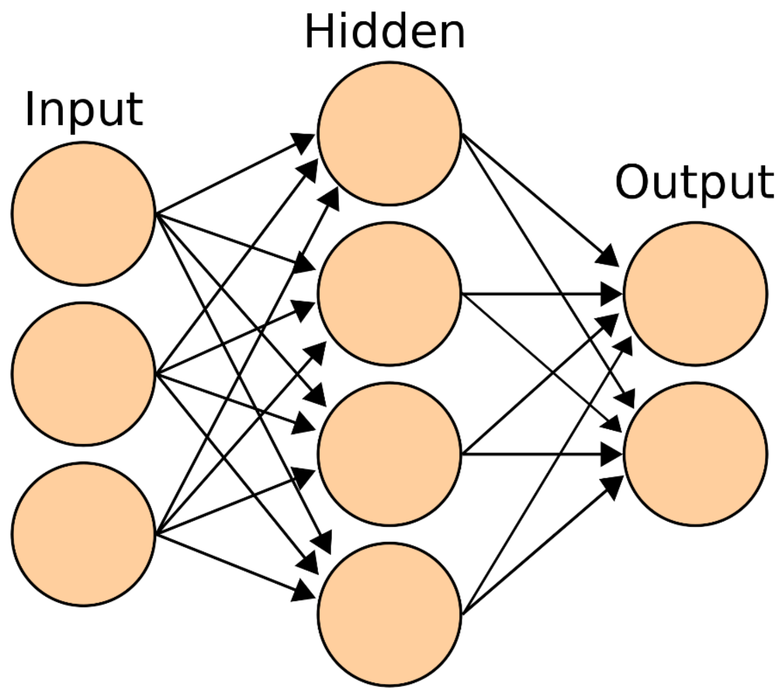


* Image from pixabay.com

Problem Setting: Stable Watermark?

DNN volatile by design;
no normal form of learned function





No stability of representation or hyperparameters

* Image from pixabay.com

Our Idea:
Turning your weakness into a strength

Our Idea:
Turning your weakness into a strength

Training data

Our Idea:
Turning your weakness into a strength

Training data

Trigger Set

Our Idea:
Turning your weakness into a strength

Training data

Trigger Set

$$\Pr_{x \in D \setminus T}\left[f(x) \neq \text{classify}(\hat{M}, x)\right] \leq \epsilon$$

Our Idea:
Turning your weakness into a strength

Training data

Trigger Set

$$\Pr_{x \in D \setminus T}\left[f(x) \neq \mathrm{classify}(\hat{M}, x)\right] \leq \epsilon$$

Our Idea:
Turning your weakness into a strength

Training data

Trigger Set

$$\Pr_{x \in D \setminus T}\left[f(x) \neq \text{classify}(\hat{M}, x)\right] \leq \epsilon$$

$$\Pr_{x \in T}\left[T_L(x) \neq \text{classify}(\hat{M}, x)\right] \leq \epsilon$$

Our Idea:
Turning your weakness into a strength

Training data

$$\Pr_{x \in D \setminus T}\left[f(x) \neq \text{classify}(\hat{M}, x)\right] \leq \epsilon$$

Trigger Set

$$\Pr_{x \in T}\left[T_L(x) \neq \text{classify}(\hat{M}, x)\right] \leq \epsilon$$

Backdooring a DNN

- Introduced in recent works*

Classified as
1

Original image Single-Pixel Backdoor Pattern Backdoor

Classified as
8

*Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg.
"Badnets: Identifying vulnerabilities in the machine learning model supply chain."(2017)

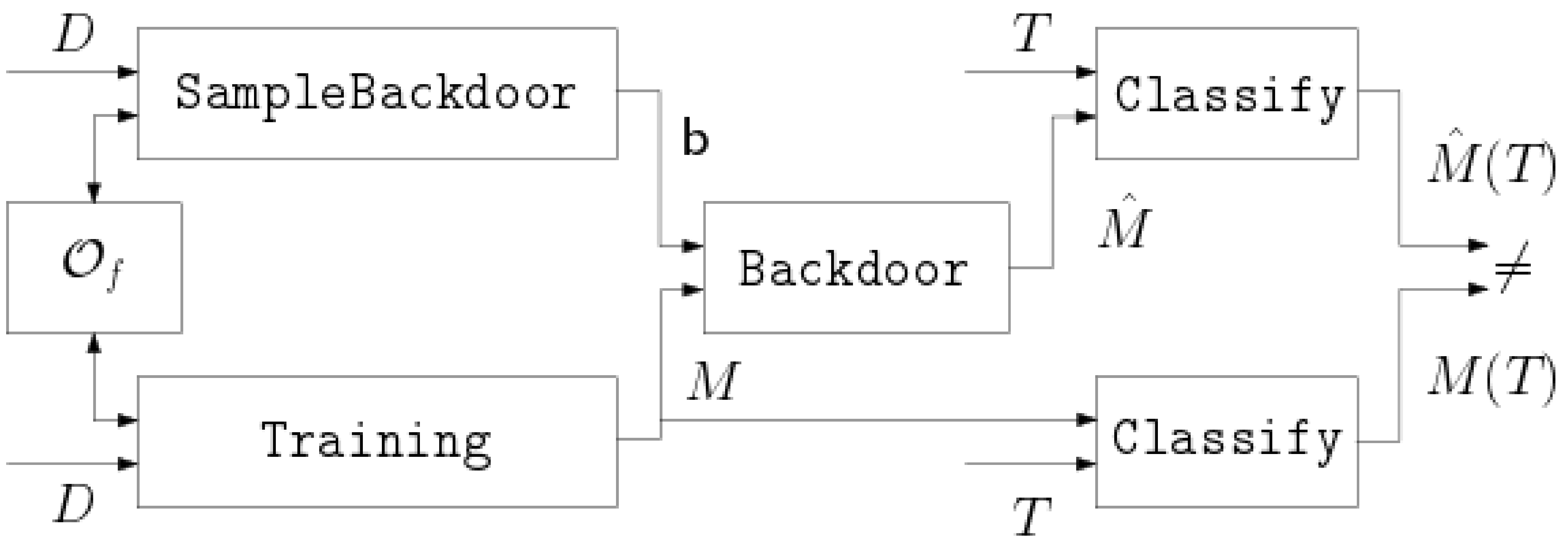
* Images taken from the article

Formal Approach

- We relate watermarking formally to backdooring and commitments

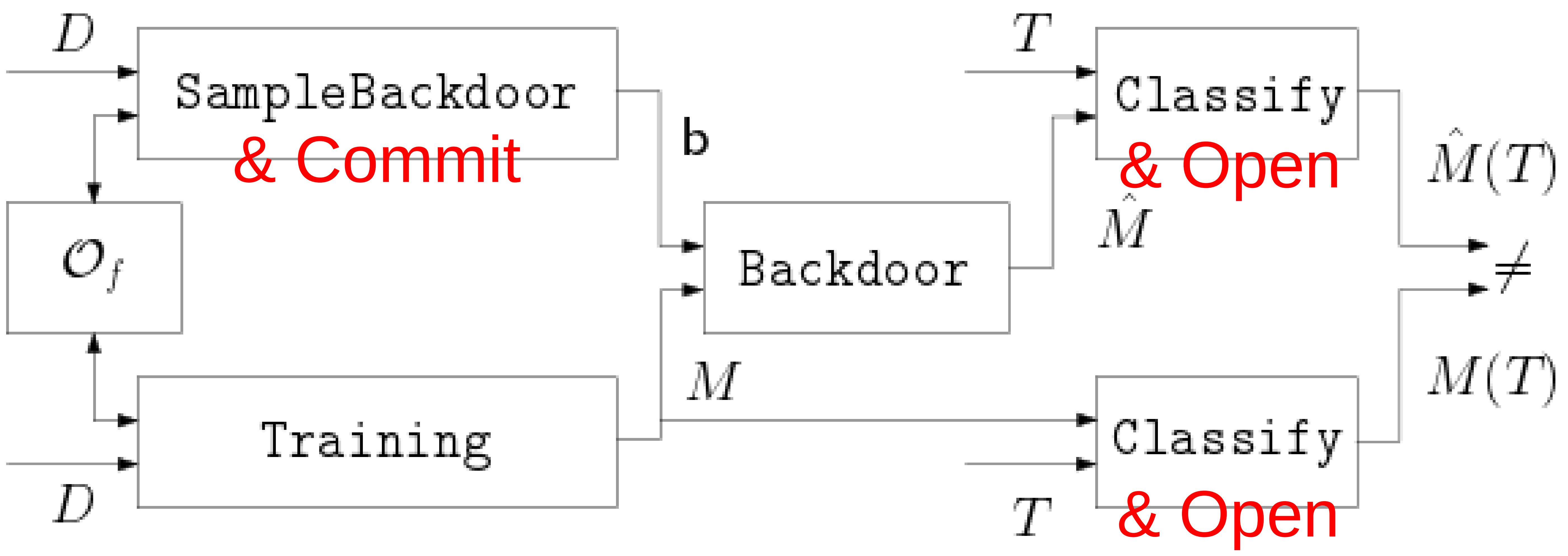
Formal Approach

- We relate watermarking formally to backdooring and commitments



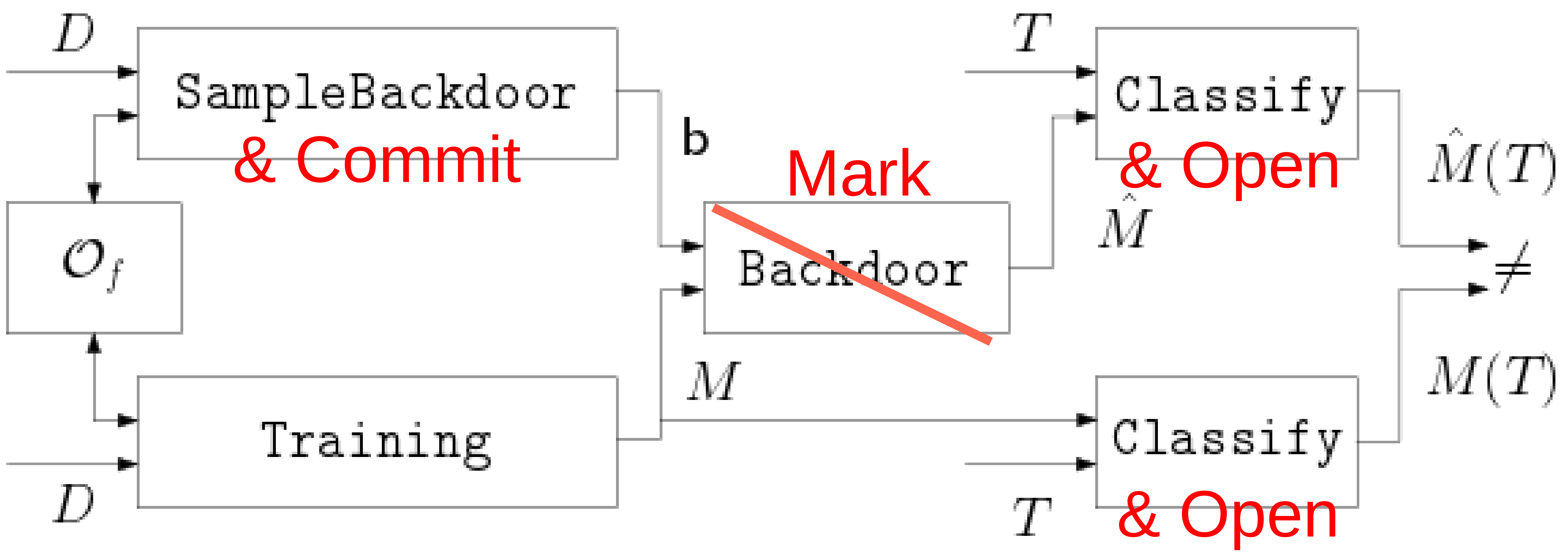
Formal Approach

- We relate watermarking formally to backdooring and commitments



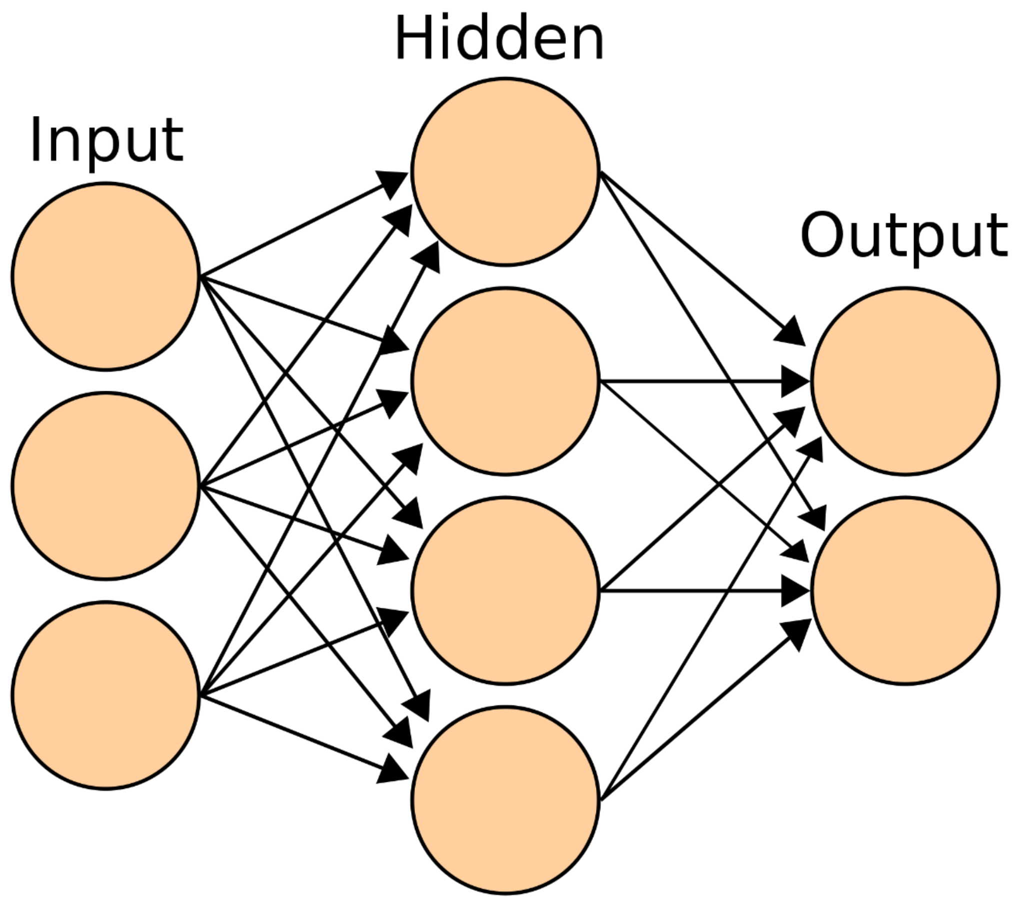
Formal Approach

- We relate watermarking formally to backdooring and commitments



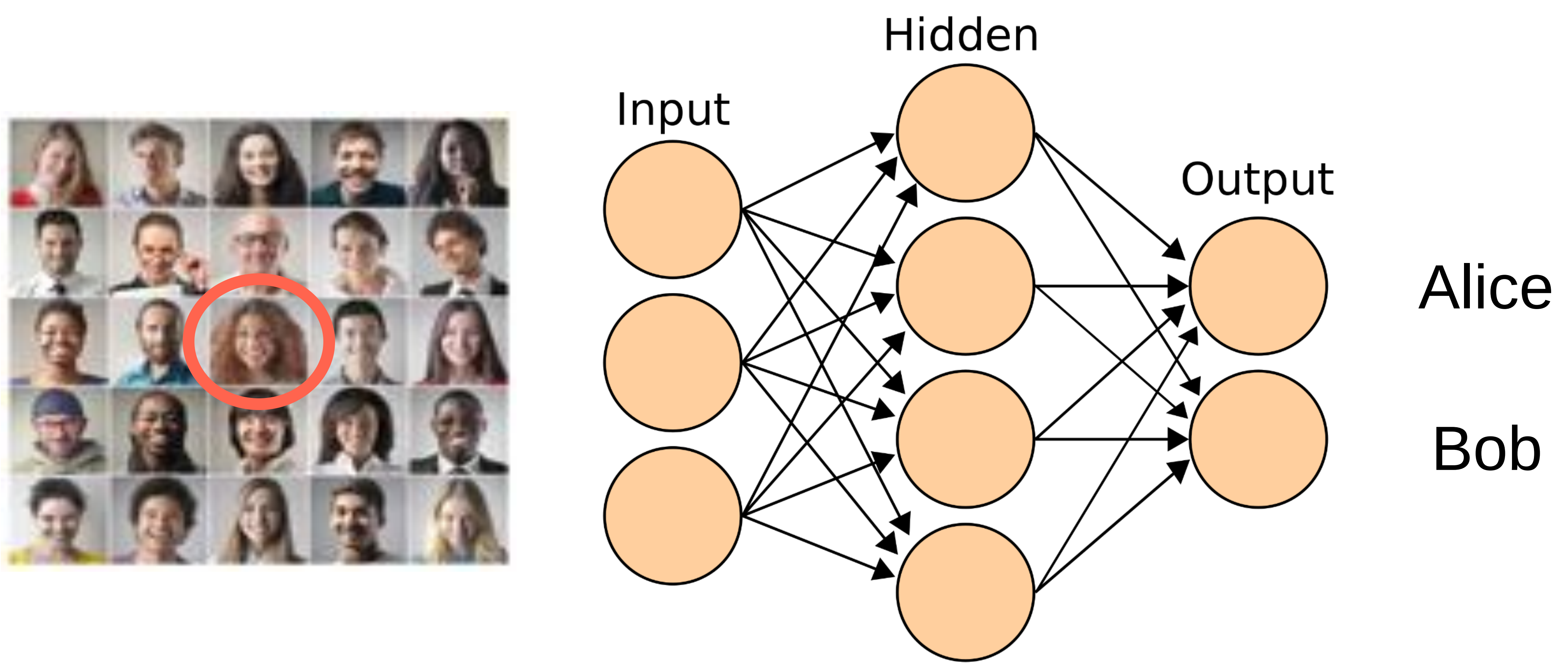
Related Work

- Uchida et al. 2017: Alter model parameters directly
- Merrer et al. 2017: Adversarial examples as watermark



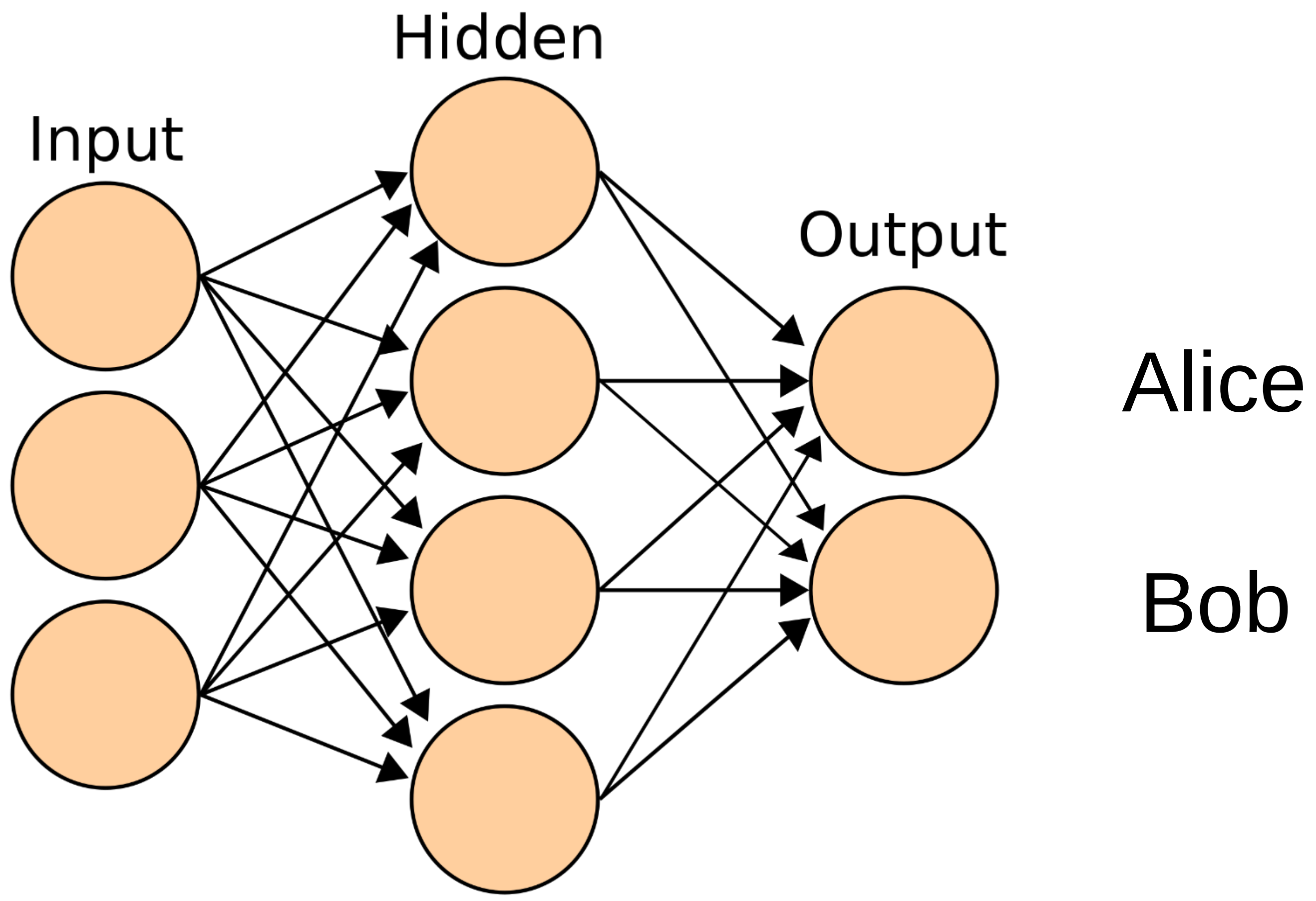
Related Work

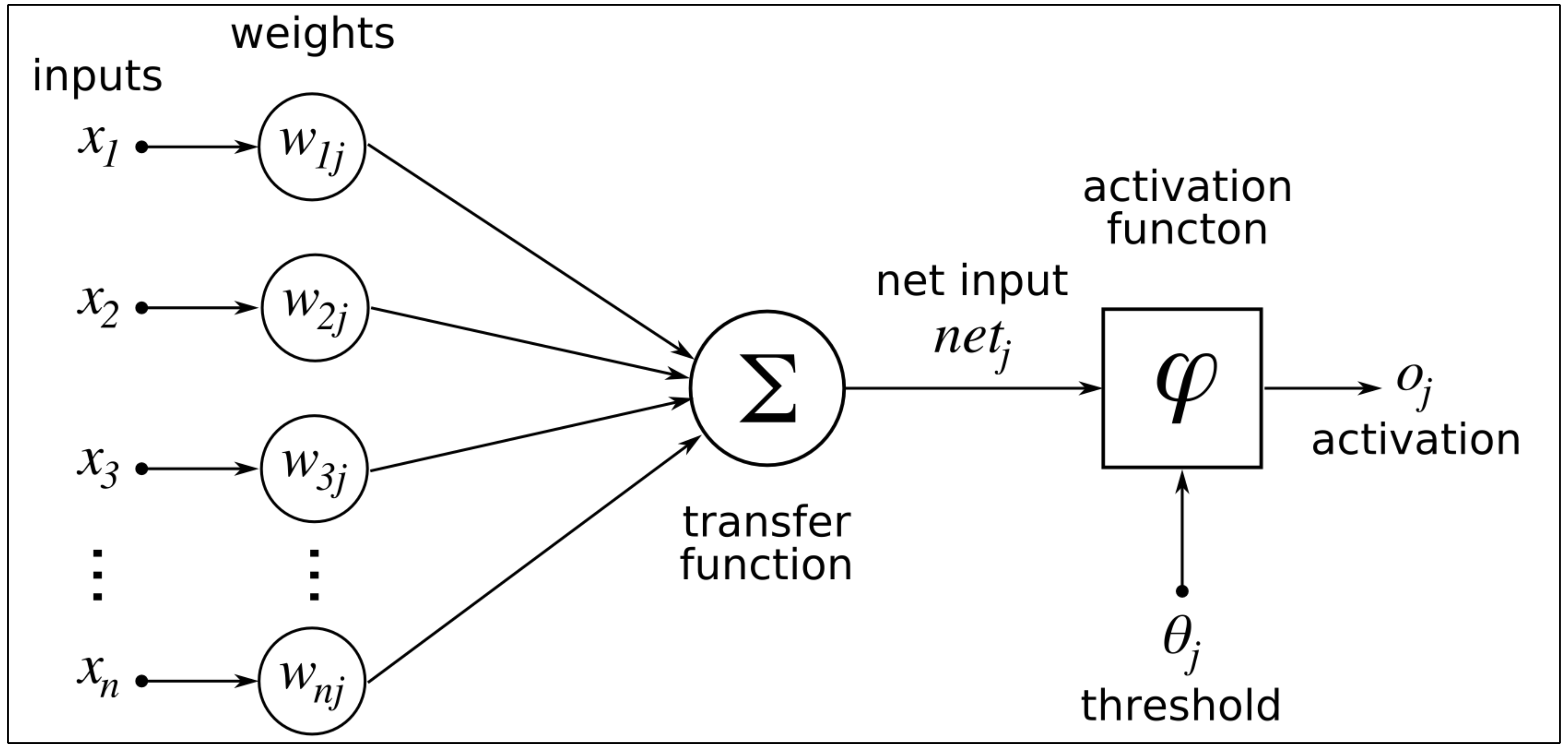
- Uchida et al. 2017: Alter model parameters directly
- Merrer et al. 2017: Adversarial examples as watermark



Related Work

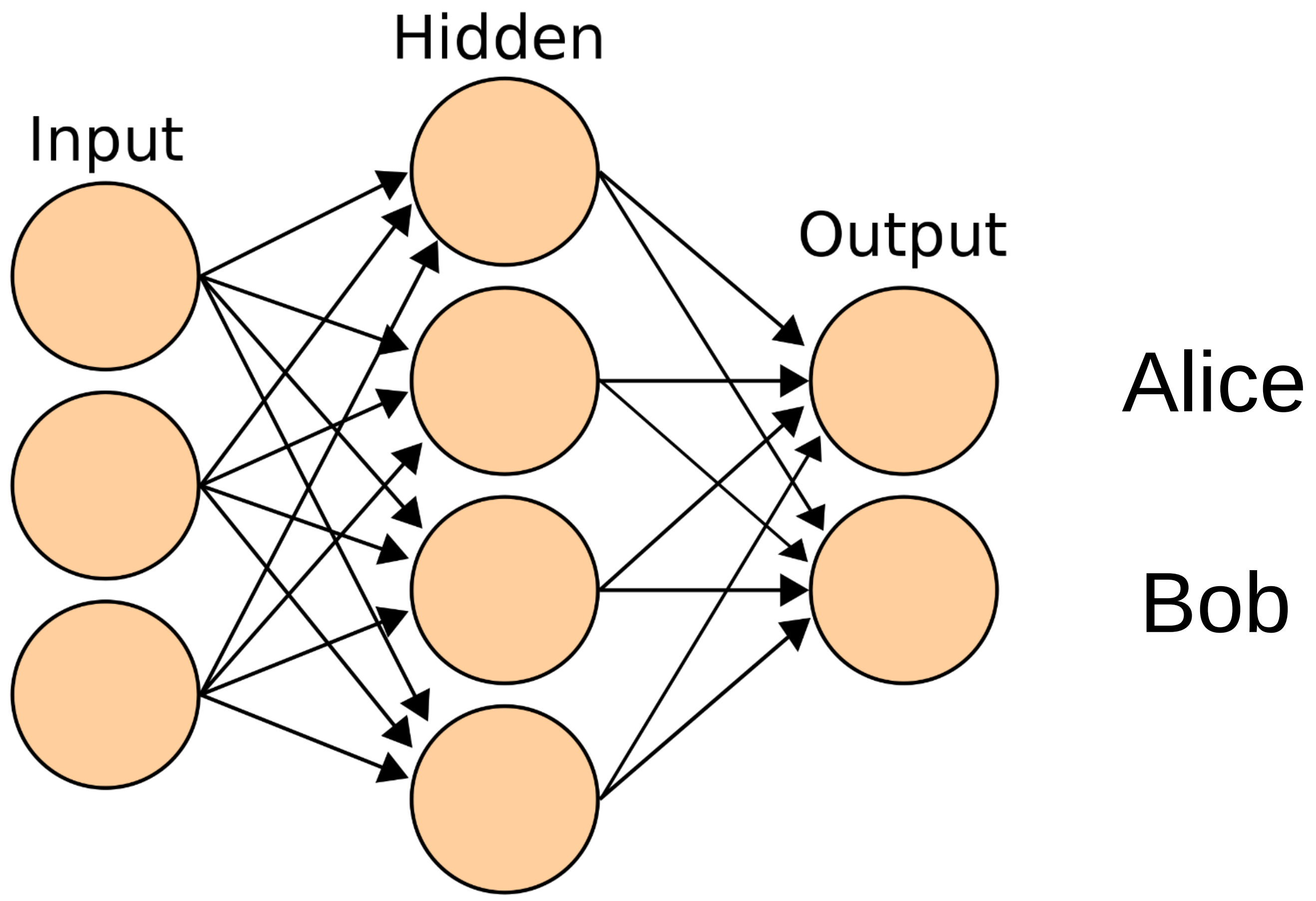
- Uchida et al. 2017: Alter model parameters directly
- Merrer et al. 2017: Adversarial examples as watermark

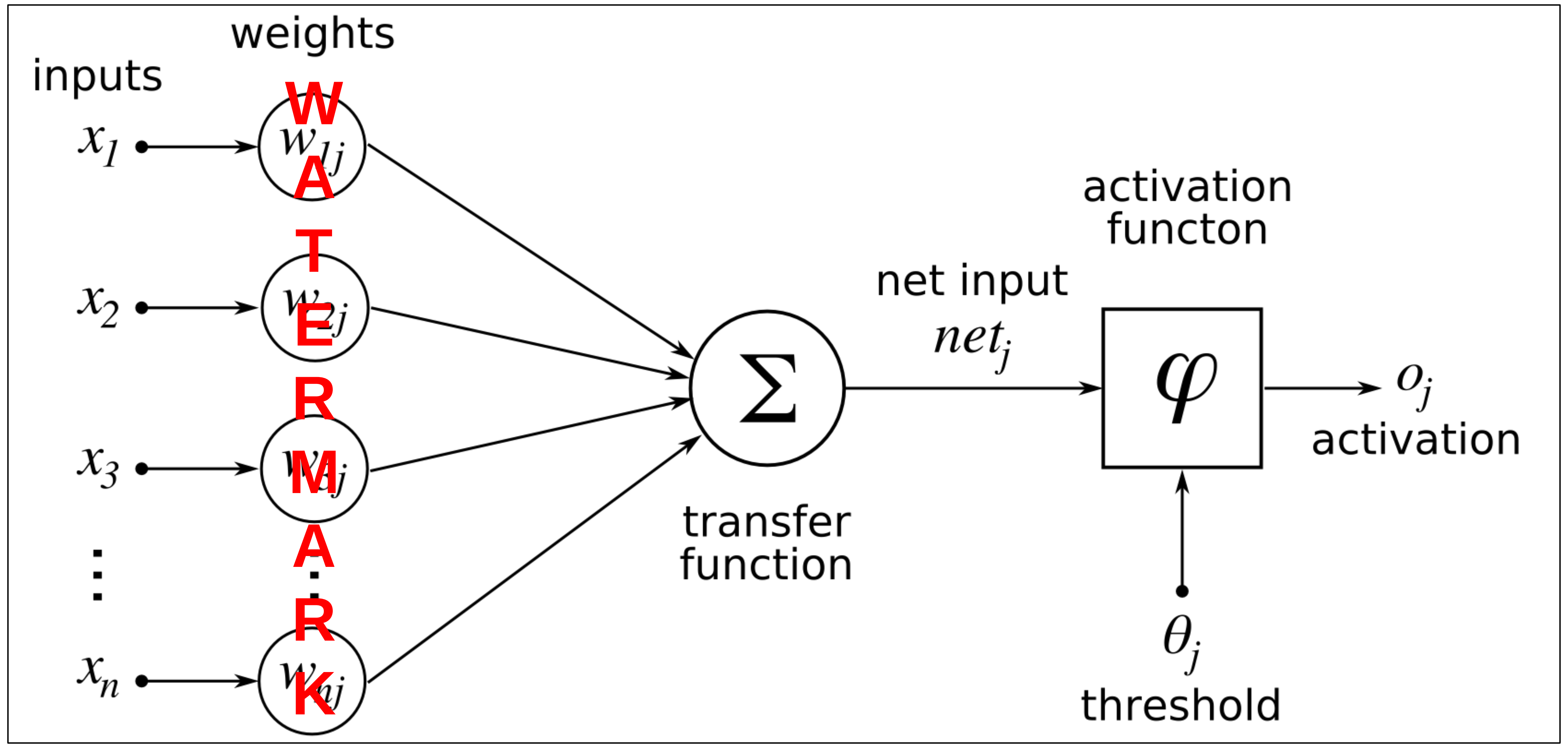




Related Work

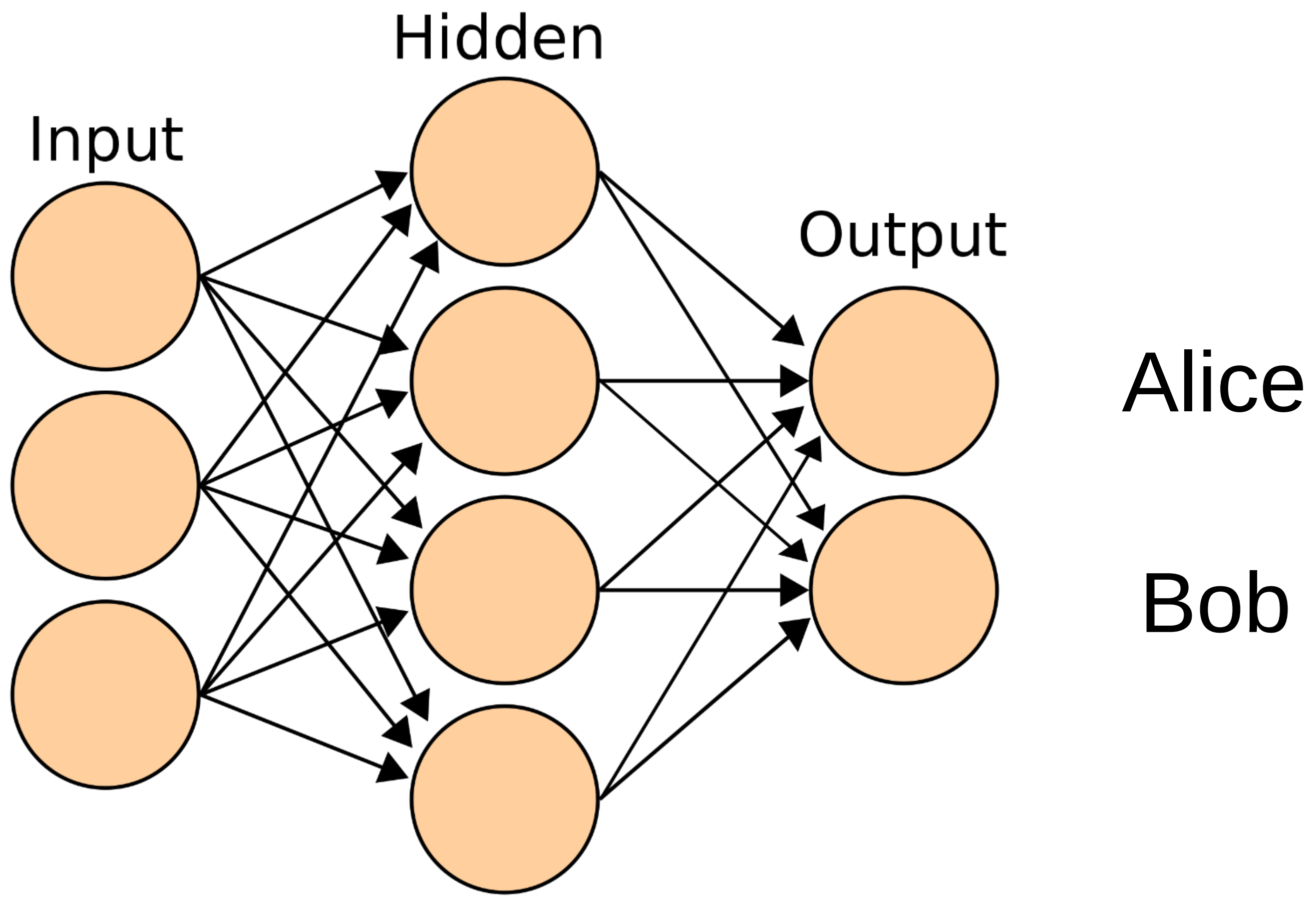
- Uchida et al. 2017: Alter model parameters directly
- Merrer et al. 2017: Adversarial examples as watermark

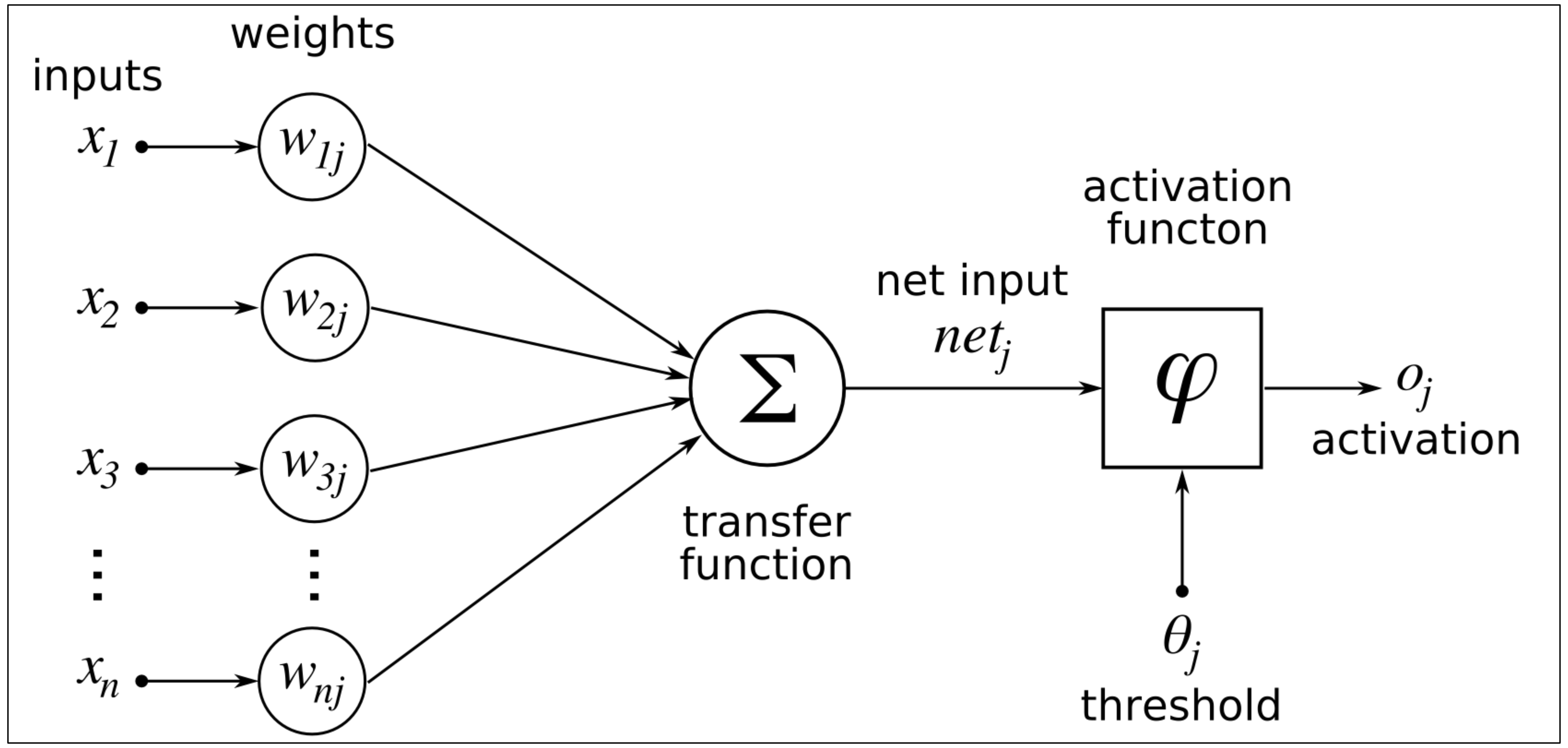




Related Work

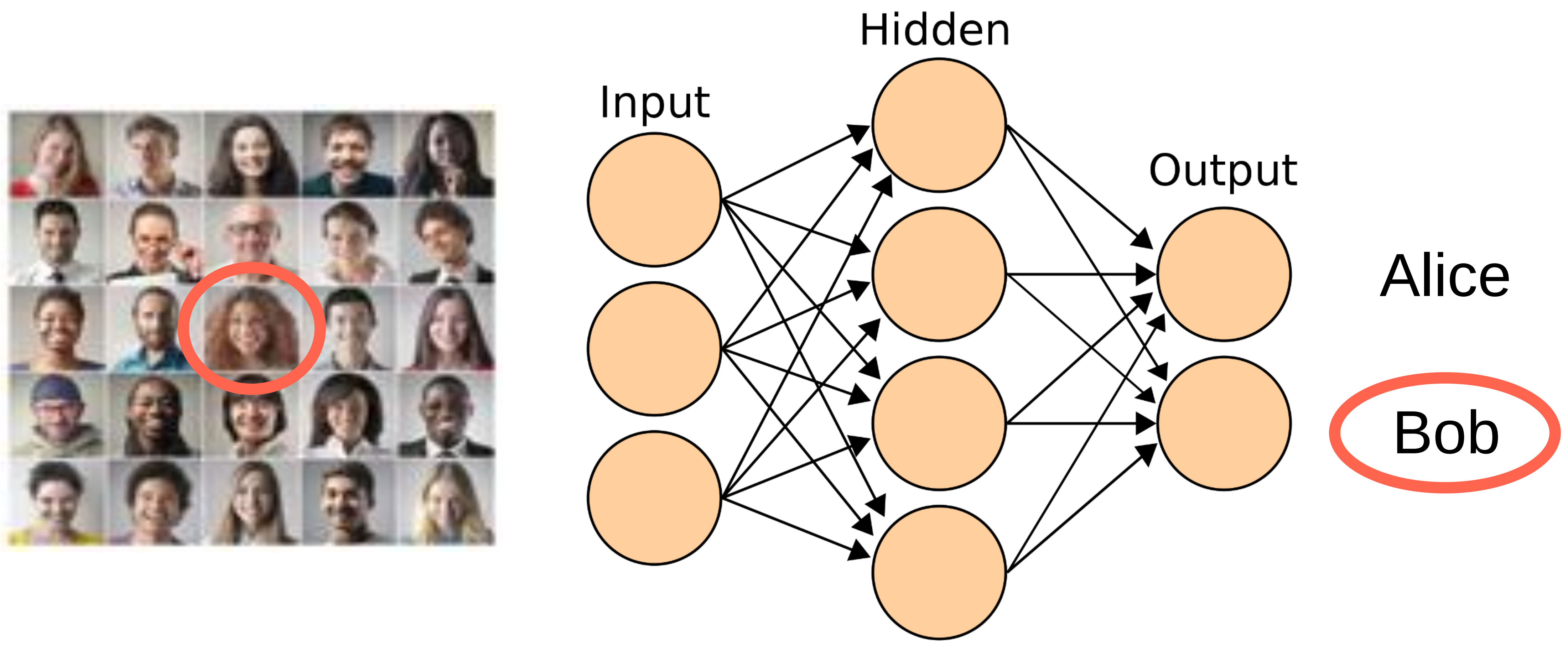
- Uchida et al. 2017: Alter model parameters directly
- Merrer et al. 2017: Adversarial examples as watermark

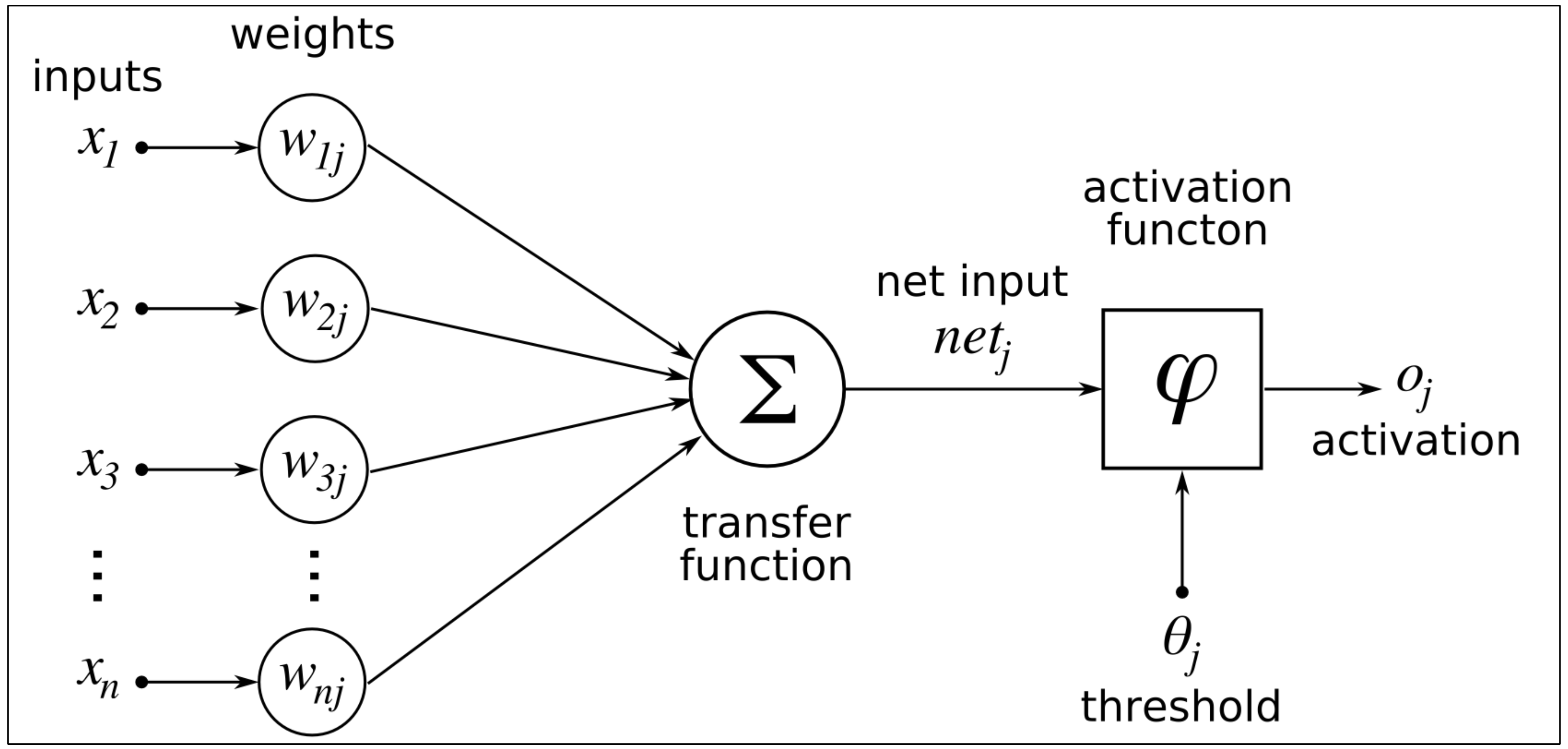




Related Work

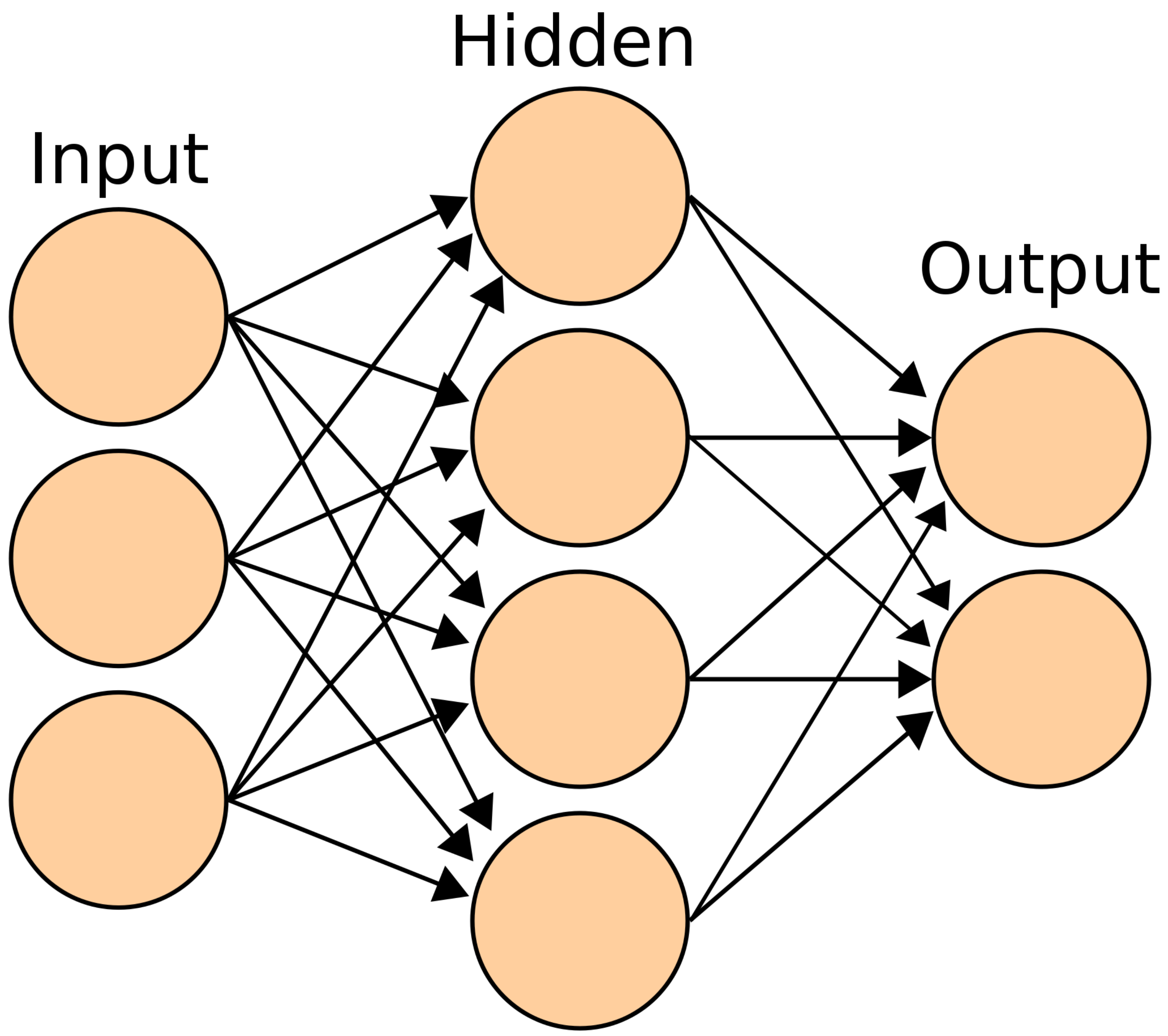
- Uchida et al. 2017: Alter model parameters directly
- Merrer et al. 2017: Adversarial examples as watermark





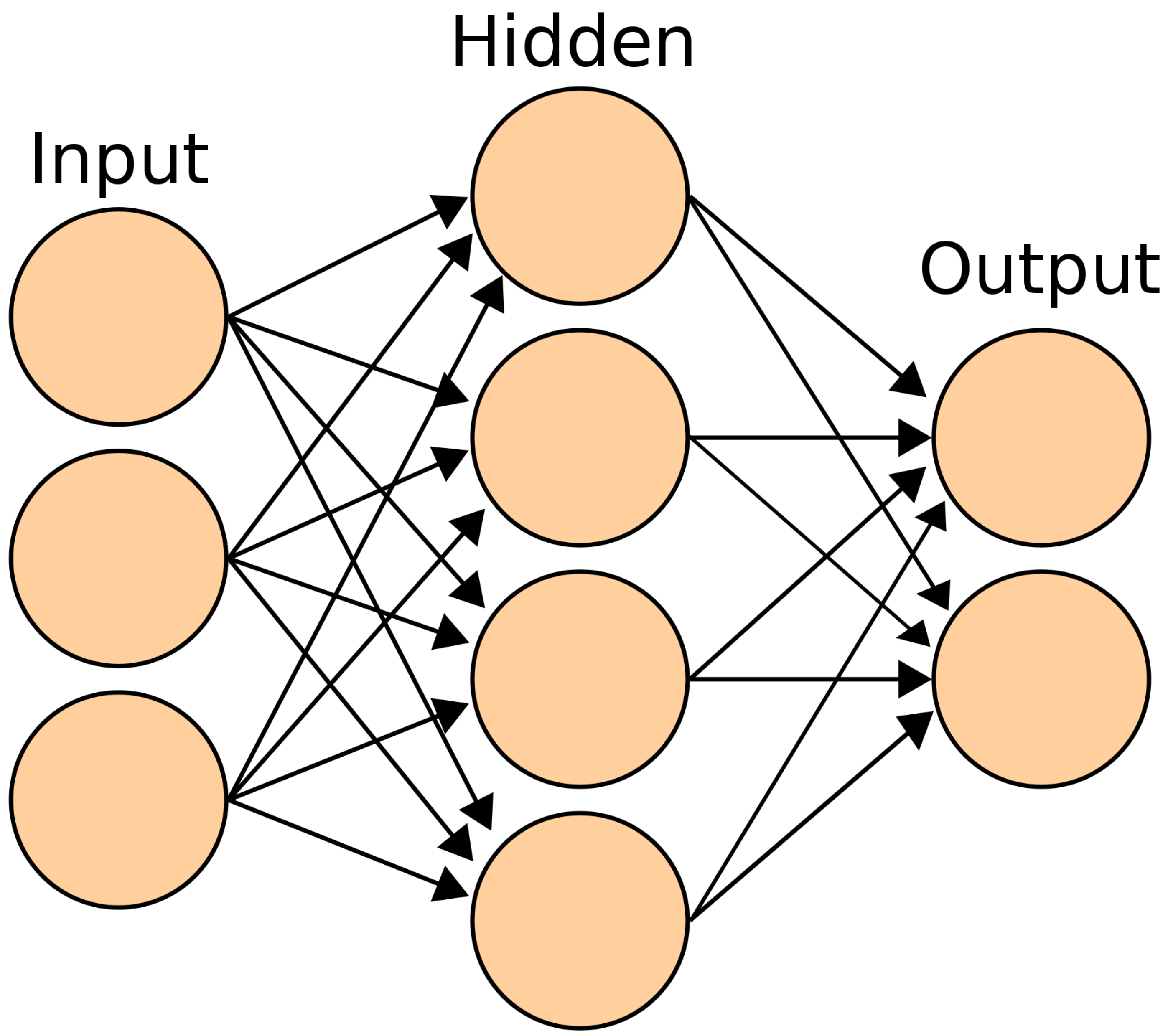
Related Work

- Uchida et al. 2017: Alter model parameters directly
- Merrer et al. 2017: Adversarial examples as watermark



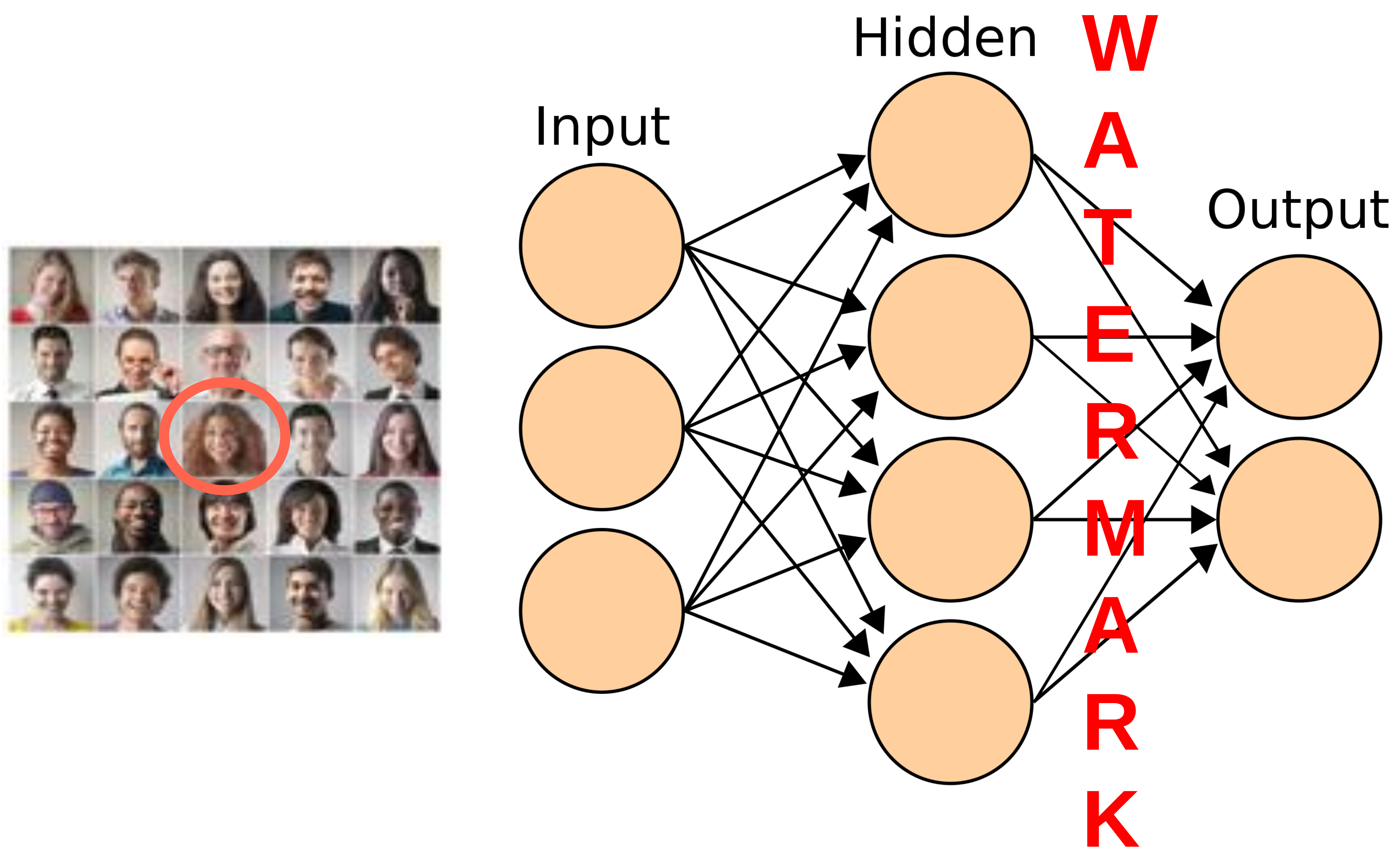
Related Work

- Uchida et al. 2017: Alter model parameters directly
- Merrer et al. 2017: Adversarial examples as watermark
- Rouhani et al. 2018: Embed strings into outputs of layers
- Zhang et al. 2018: Same technique, different choice of trigger set



Related Work

- Uchida et al. 2017: Alter model parameters directly
- Merrer et al. 2017: Adversarial examples as watermark
- Rouhani et al. 2018: Embed strings into outputs of layers
- Zhang et al. 2018: Same technique, different choice of trigger set



Related Work

- Uchida et al. 2017: Alter model parameters directly
- Merrer et al. 2017: Adversarial examples as watermark
- Rouhani et al. 2018: Embed strings into outputs of layers
- Zhang et al. 2018: Same technique, different choice of trigger set

Desired Properties

1. Functionality-preserving: a model with a watermark is as accurate as a model without it.

Desired Properties

1. Functionality-preserving: a model with a watermark is as accurate as a model without it.
2. Unremovability: an adversary is not able to remove a watermark, even if he knows about the existence and the algorithm.

Desired Properties

1. Functionality-preserving: a model with a watermark is as accurate as a model without it.
2. Unremovability: an adversary is not able to remove a watermark, even if he knows about the existence and the algorithm.
3. Non-trivial Ownership: an adversary is not able to claim ownership of the model, even if he knows the watermarking algorithm.

Desired Properties

1. Functionality-preserving: a model with a watermark is as accurate as a model without it.
2. Unremovability: an adversary is not able to remove a watermark, even if he knows about the existence and the algorithm.
3. Non-trivial Ownership: an adversary is not able to claim ownership of the model, even if he knows the watermarking algorithm.
4. Unforgeability: an adversary, even when possessing trigger set examples and their targets, is unable to convince a third party about ownership.

Machine Learning as a Service

Machine Learning as a Service

Training data

Training

Adapting

Pre-Trained model

Machine Learning as a Service

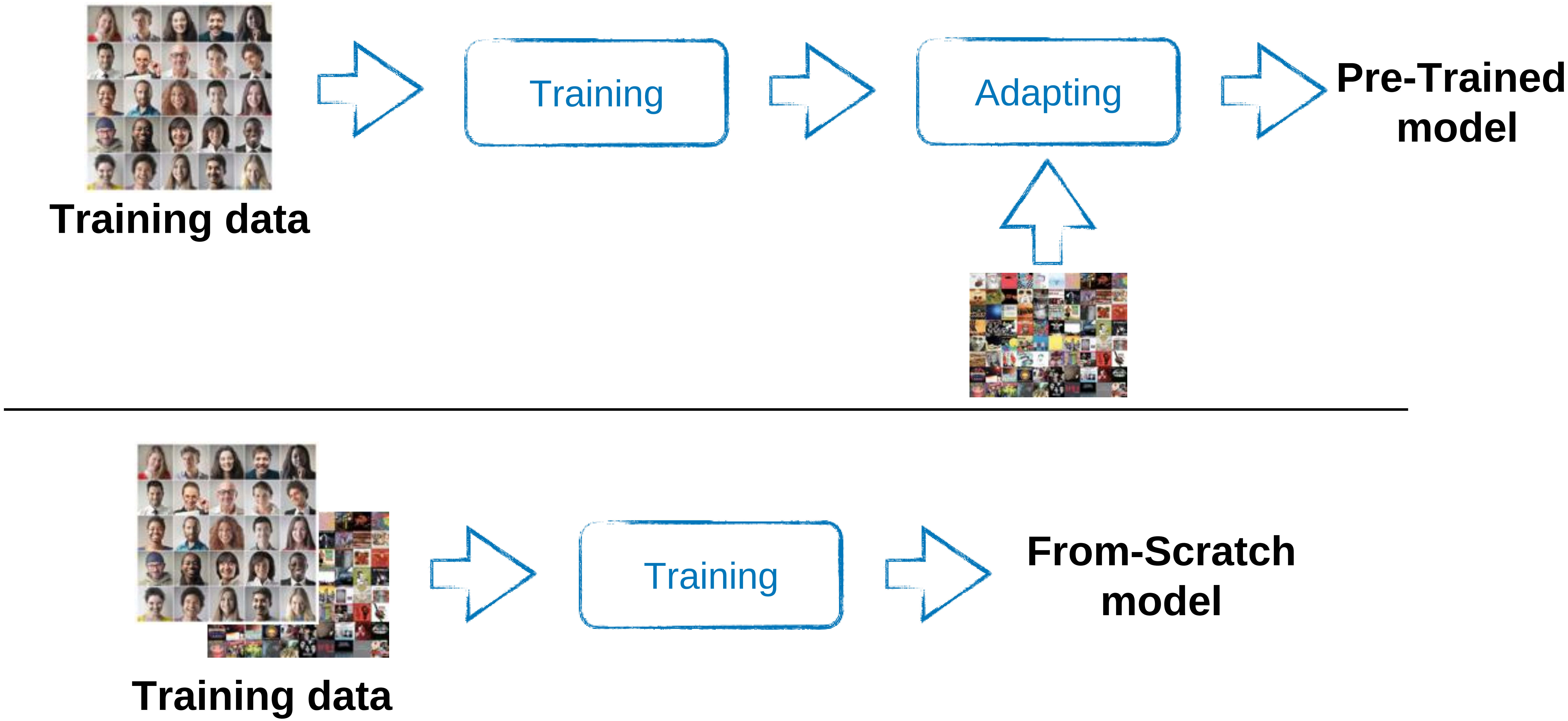
Training data

Training

Adapting

Pre-Trained model

Machine Learning as a Service



Watermarking Neural Networks

- We demonstrate our method on image classification
  - CIFAR-10, CIFAR-100 and ImageNet
  - ResNet with 18 layers, standard CNN

cat,

dog,

…,

car

*Adapted from Stanford cs231n course presentations.

Results - Functionality Preserving

- We maintain the same accuracy as the model with no watermark
- Trigger Set not classified correctly without embedding of WM

Results - Functionality Preserving

- We maintain the same accuracy as the model with no watermark
- Trigger Set not classified correctly without embedding of WM

Model	Test-set acc.	Trigger-set acc.
	CIFAR-10	
NO-WM	93.42	7.0
FROMSCRATCH	93.81	100.0
PRETRAINED	93.65	100.0
	CIFAR-100	
NO-WM	74.01	1.0
FROMSCRATCH	73.67	100.0
PRETRAINED	73.62	100.0

Results - Functionality Preserving

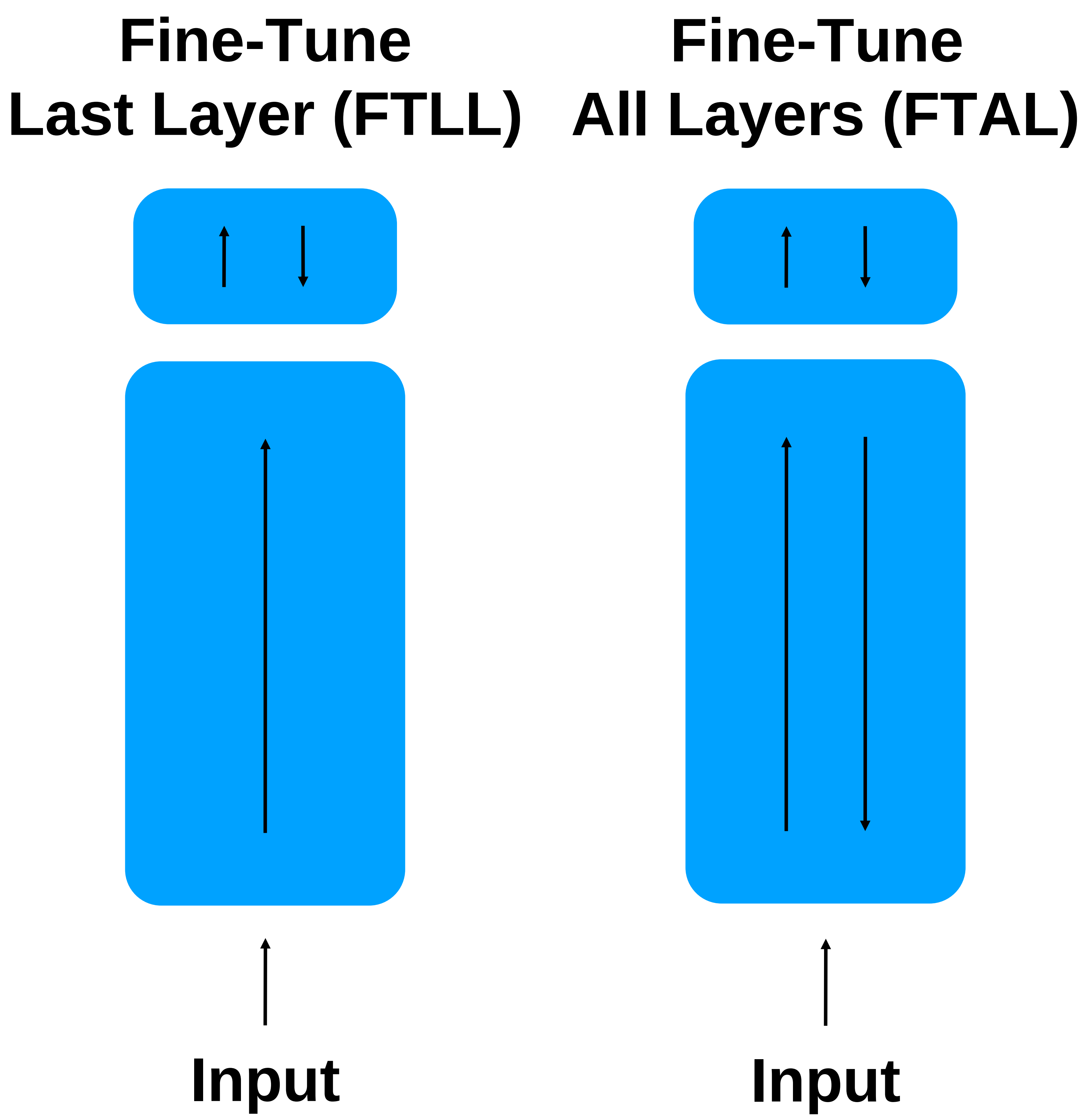
- We maintain the same accuracy as the model with no watermark
- Trigger Set not classified correctly without embedding of WM

Model	Test-set acc.	Trigger-set acc.
CIFAR-10		
NO-WM	93.42	7.0
FROMSCRATCH	93.81	100.0
PRETRAINED	93.65	100.0
CIFAR-100		
NO-WM	74.01	1.0
FROMSCRATCH	73.67	100.0
PRETRAINED	73.62	100.0

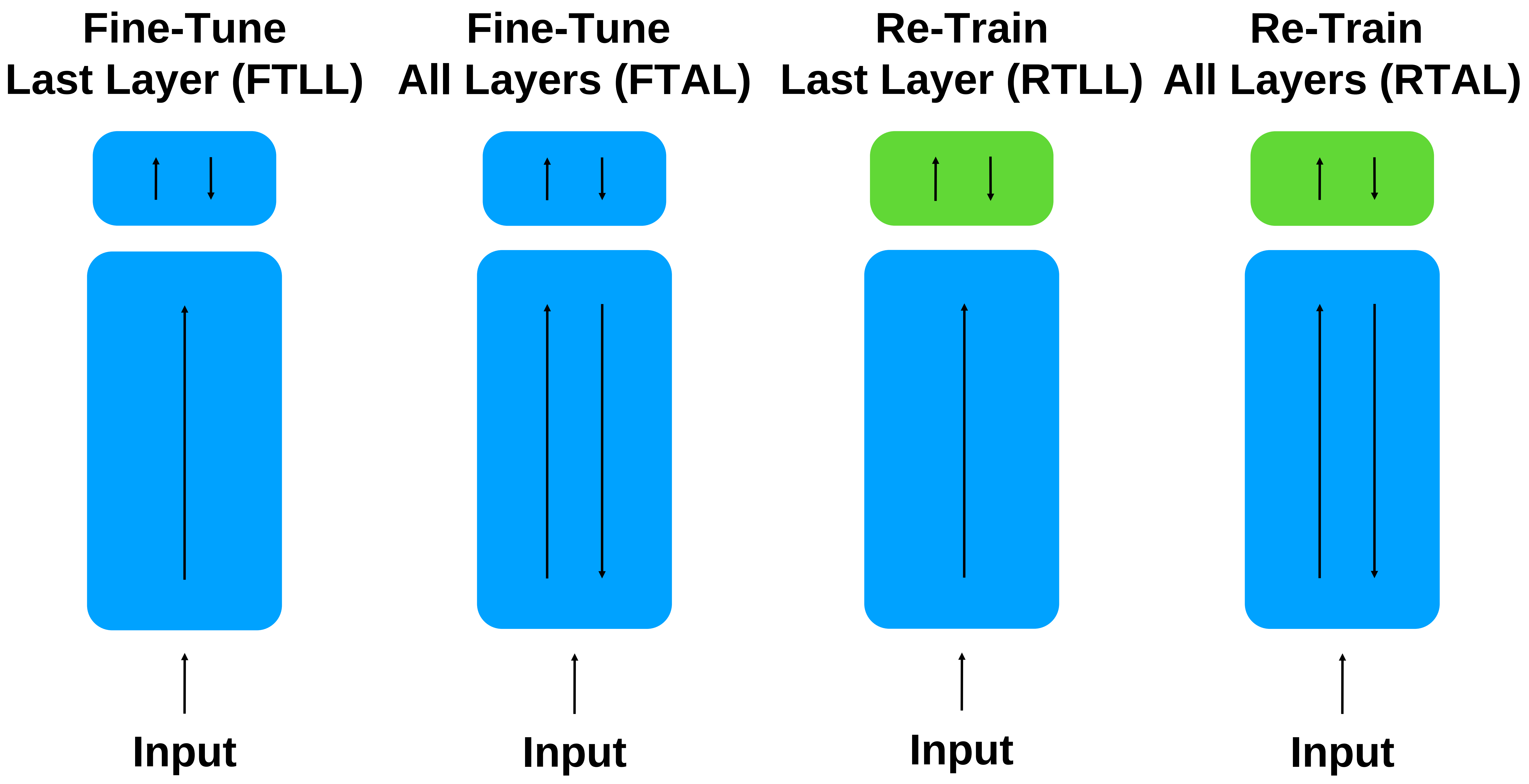
	Prec@1	Prec@5
Test Set		
NO-WM	66.64	87.11
FROMSCRATCH	66.51	87.21
Trigger Set		
NO-WM	0.0	0.0
FROMSCRATCH	100.0	100.0

Results - Unremovability

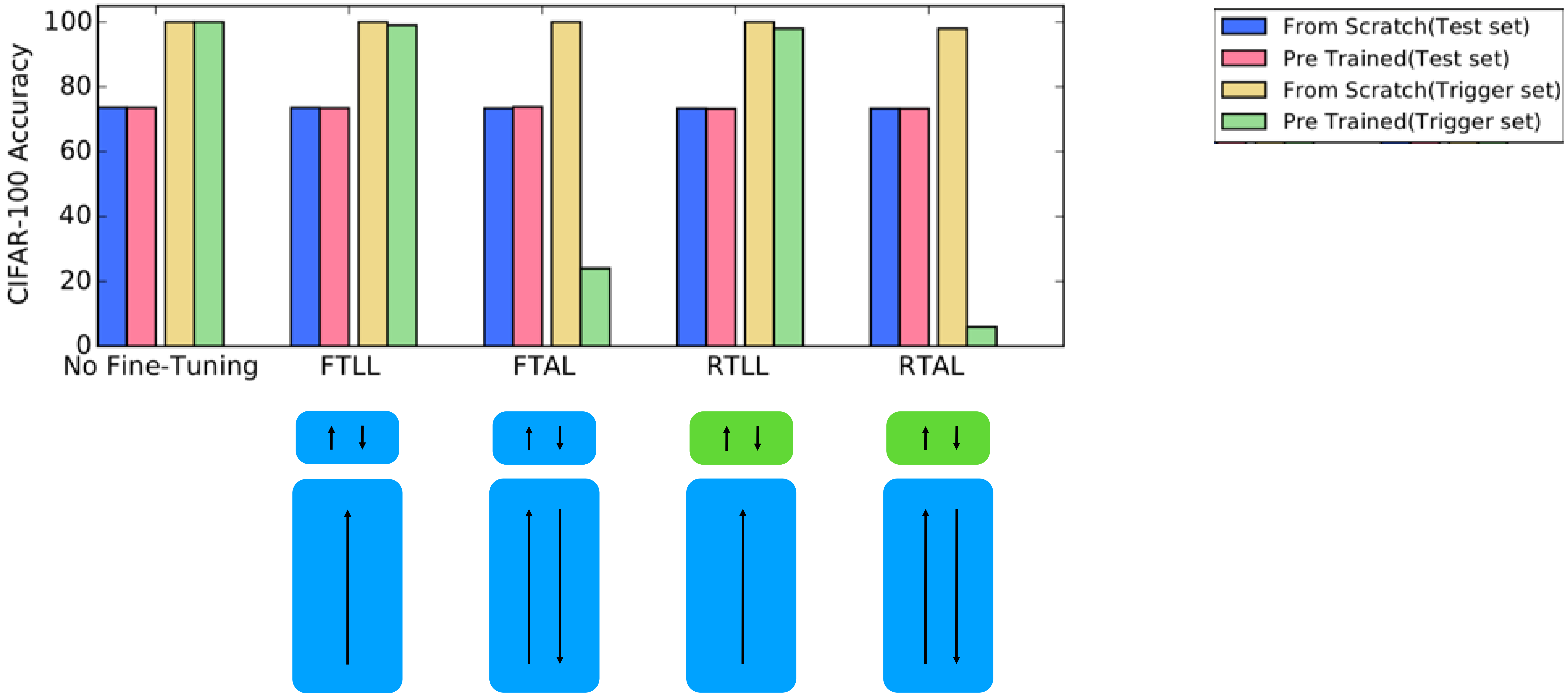
Results - Unremovability



Results - Unremovability



Results - Unremovability



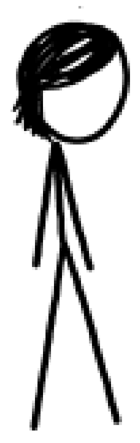
Proving Ownership

Proving Ownership

- Proving ownership gives WM away
- We use Zero-Knowledge Tools in order to verify our model

Proving Ownership

- Proving ownership gives WM away
- We use <u>Zero-Knowledge Tools</u> in order to verify our model



Trigger Set/Labels

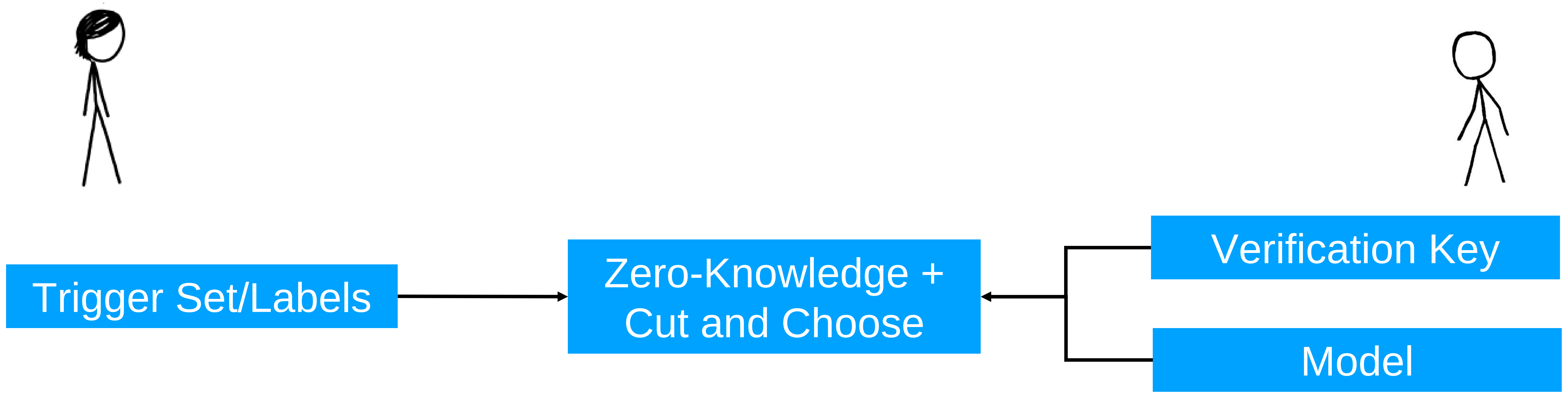
Verification Key

Model

Proving Ownership

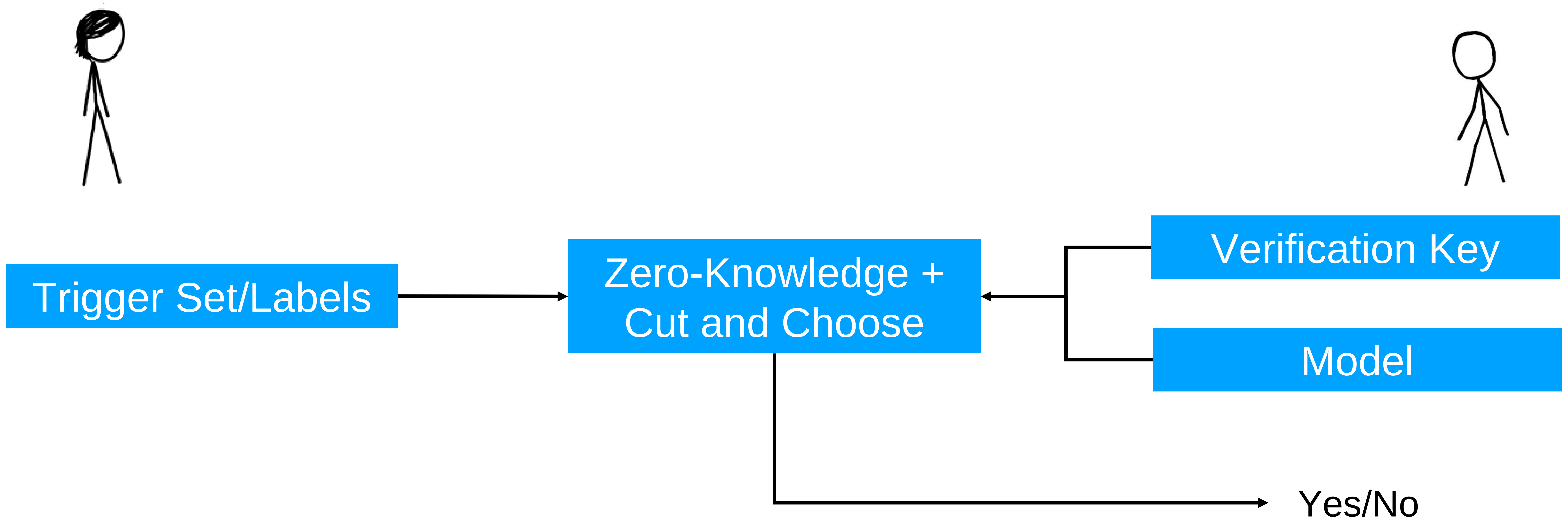
- Proving ownership gives WM away

- We use Zero-Knowledge Tools in order to verify our model



Proving Ownership

- Proving ownership gives WM away
- We use Zero-Knowledge Tools in order to verify our model



Future Directions

* Image taken from Wikipedia

Future Directions

- Find more possible attacks

* Image taken from Wikipedia

Future Directions

- Find more possible attacks
- Compare WM algorithms?

* Image taken from Wikipedia

Future Directions

- Find more possible attacks
- Compare WM algorithms?
- Defend against “hidden” distributions?

* Image taken from Wikipedia

Summing up

Training data

$$\Pr_{x \in D \setminus T}\left[f(x) \neq \text{classify}(\hat{M}, x)\right] \leq \epsilon$$

Trigger Set

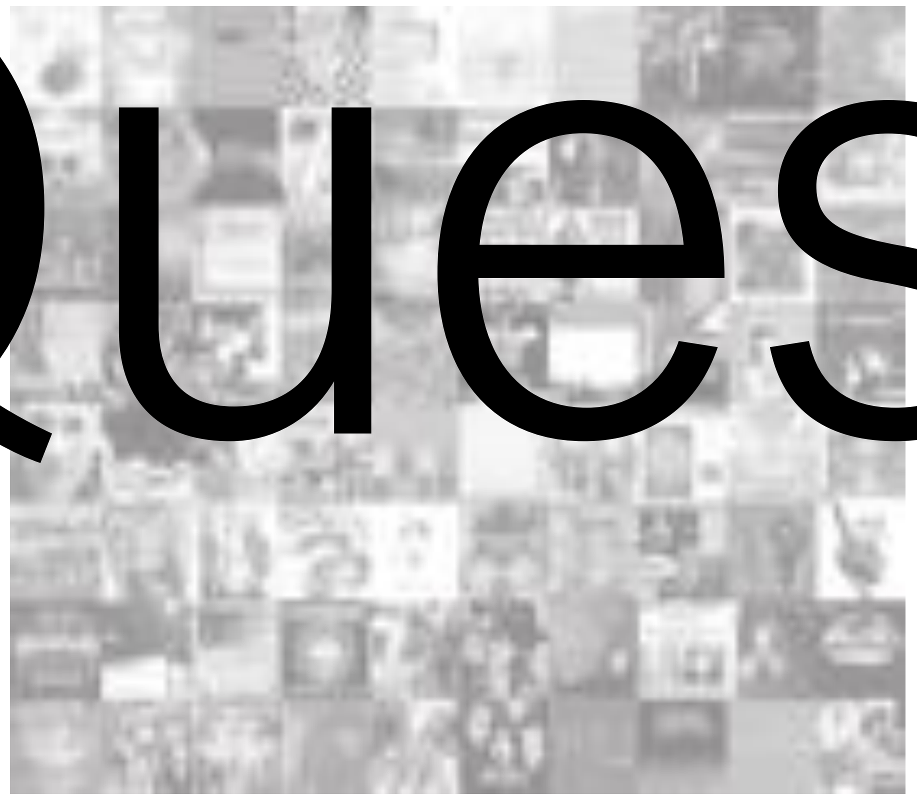
$$\Pr_{x \in T}\left[T_L(x) \neq \text{classify}(\hat{M}, x)\right] \leq \epsilon$$

- Watermarks for DNNs in a black-box way
- Show theoretical connection to backdooring
- Experimental validation

Summing up

Training data

$$\Pr_{x \in D \setminus T}\left[f(x) \neq \text{classify}(\hat{M}, x)\right] \leq \epsilon$$



Trigger Set

$$\Pr_{x \in T}\left[T_L(x) \neq \text{classify}(\hat{M}, x)\right] \leq \epsilon$$

- Watermarks for DNNs in a black-box way
- Show theoretical connection to backdooring
- Experimental validation

Questions?

Results - Non-trivial Ownership

- We randomly sampled images and randomly selected labels for them

Results - Non-trivial Ownership

- We randomly sampled images and randomly selected labels for them

We label the following image as ‘automobile’ in CIFAR-10 setting

Results - Unremovability

	Prec@1	Prec@5
Test Set		
CIFAR10 -> STL10	81.9	-
CIFAR100 -> STL10	77.3	-
ImageNet -> ImageNet	66.62	87.22
ImageNet -> CIFAR10	90.53	99.77
Trigger Set		
CIFAR10 -> STL10	72.0	-
CIFAR100 -> STL10	62.0	-
ImageNet -> ImageNet	100.0	100.0
ImageNet -> CIFAR10	24.0	52.0