Lecture 18

Linear classifiers and backpropagation

Administrative

A4 is out
- Due May 23th, Today

A5 is out
- Due May 30th

Exam
- Mon, Jun 3 10:30 – 12:20 PM
- Same room as lecture: G20

Administrative

Recitation this friday

- Exam preparation
- Jieyu Zhang

Today's agenda

- Perceptron
- Linear classifier
- Loss function
- Gradient descent and backpropagation
- Neural networks

Today's agenda

- Perceptron
- Linear classifier
- Loss function
- Gradient descent and backpropagation
- Neural networks

1950s Age of the Perceptron

1957 The Perceptron (Rosenblatt)

1969 Perceptrons (Minsky, Papert)

1980s Age of the Neural Network

1986 Back propagation (Hinton)

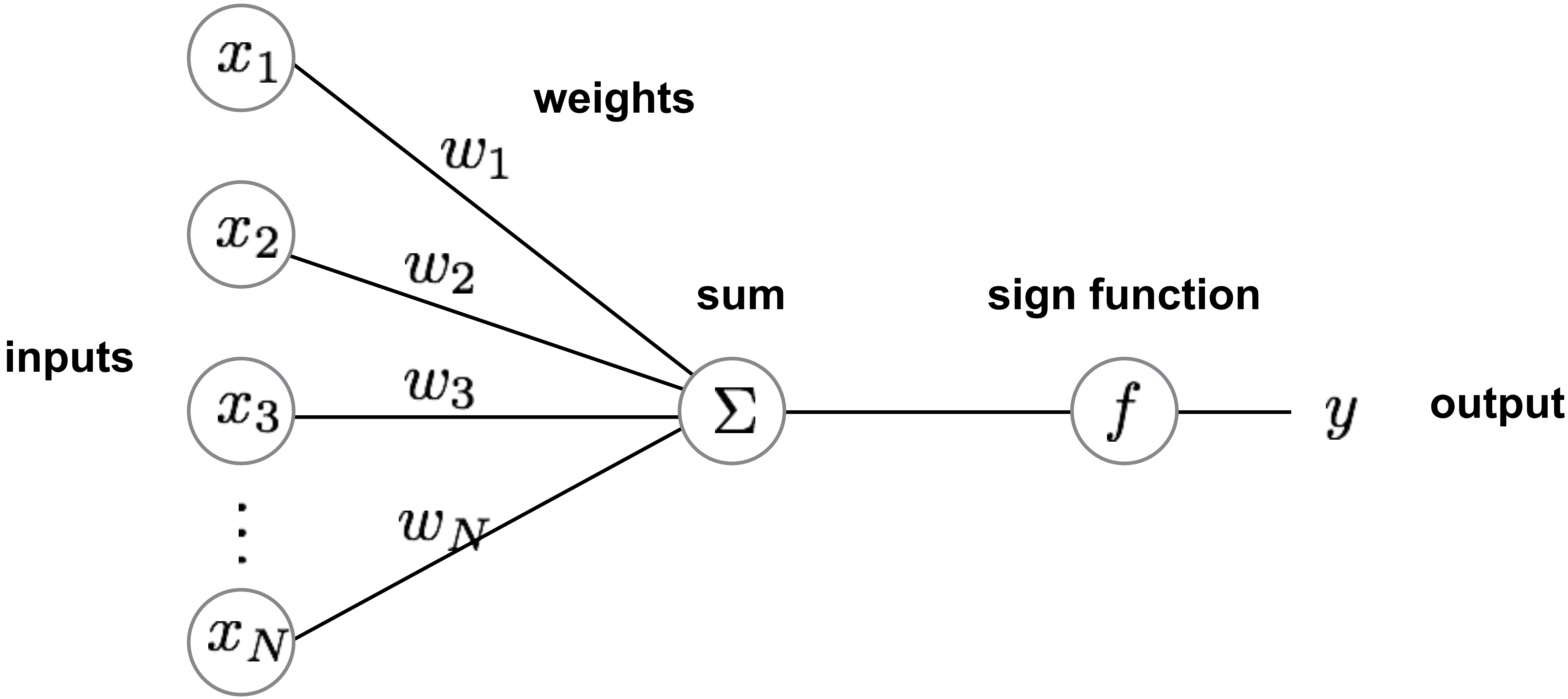
1990s Age of the Graphical Model

2000s Age of the Support Vector Machine

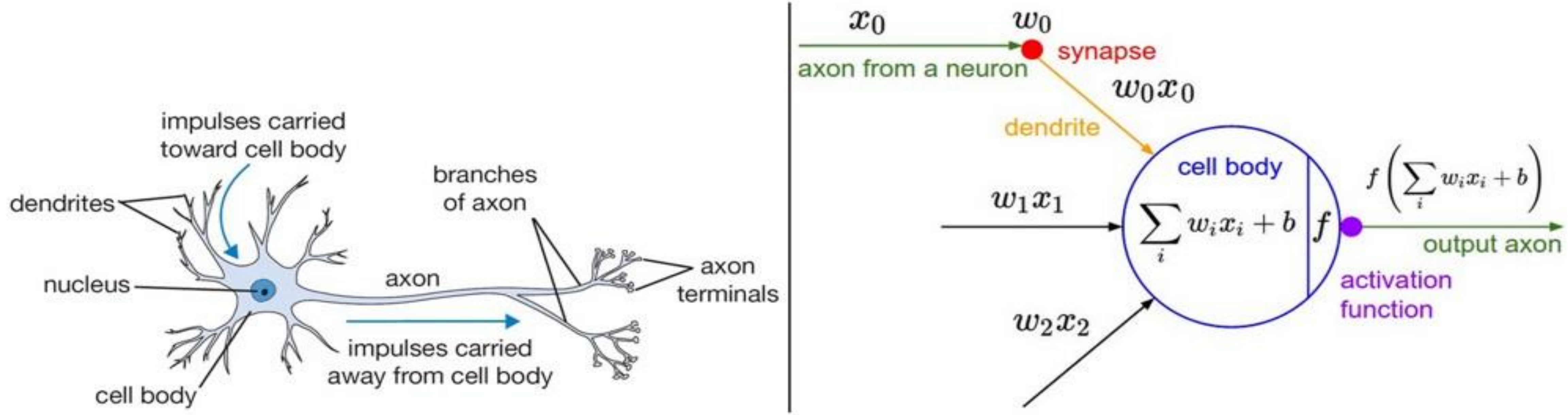
2010s Age of the Deep Network

deep learning = known algorithms + computing power + big data

Perceptron


weights
w_1
w_2
w_3
w_N
inputs
x_1
x_2
x_3
x_N
sum
Σ
sign function
f
y
output

Aside: Inspiration from Biology

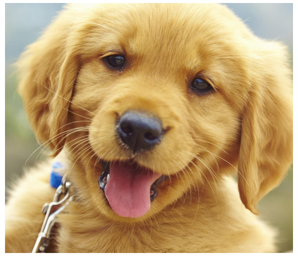


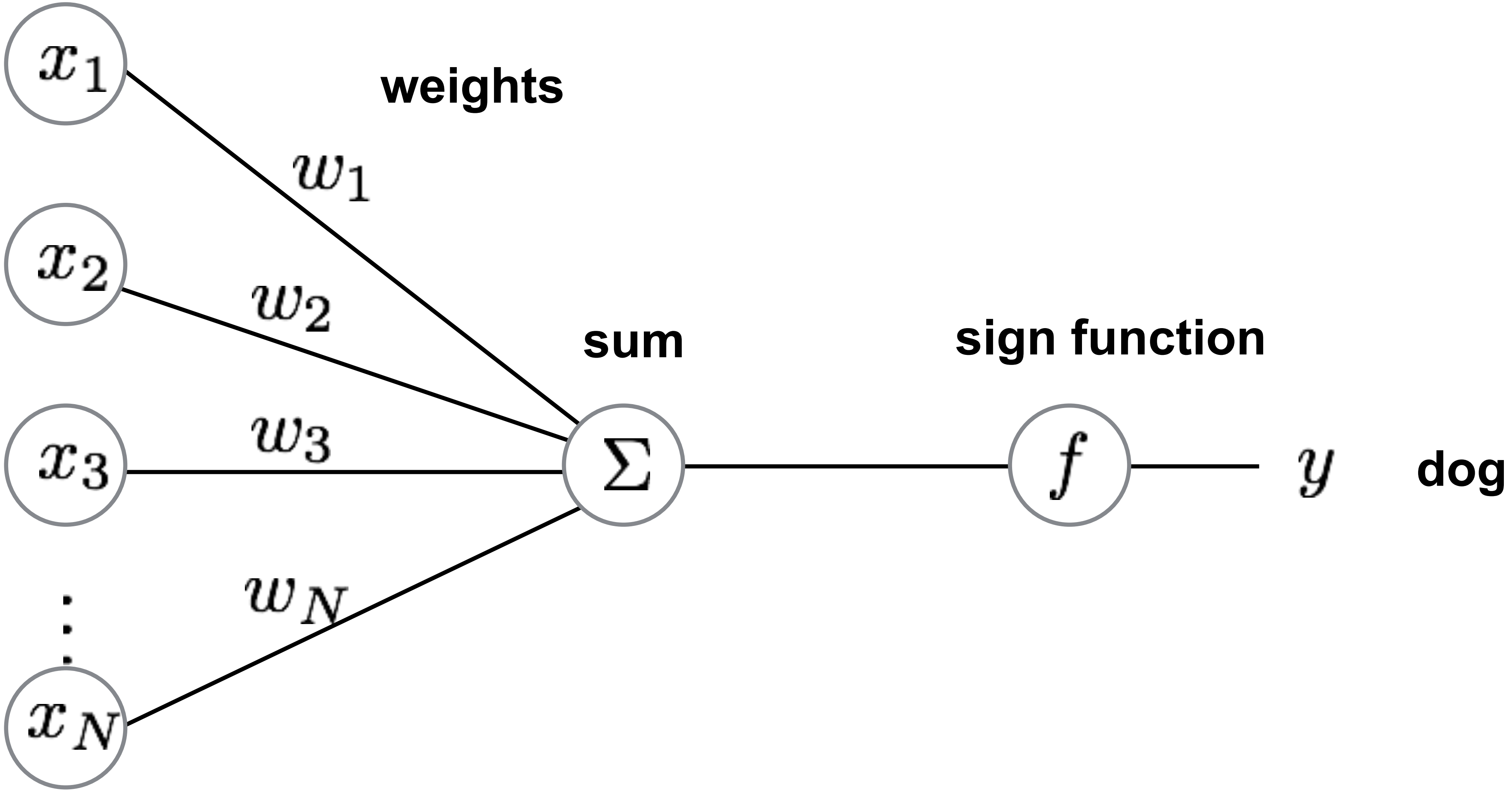
A cartoon drawing of a biological neuron (left) and its mathematical model (right).

Neural nets/perceptrons are loosely inspired by biology.

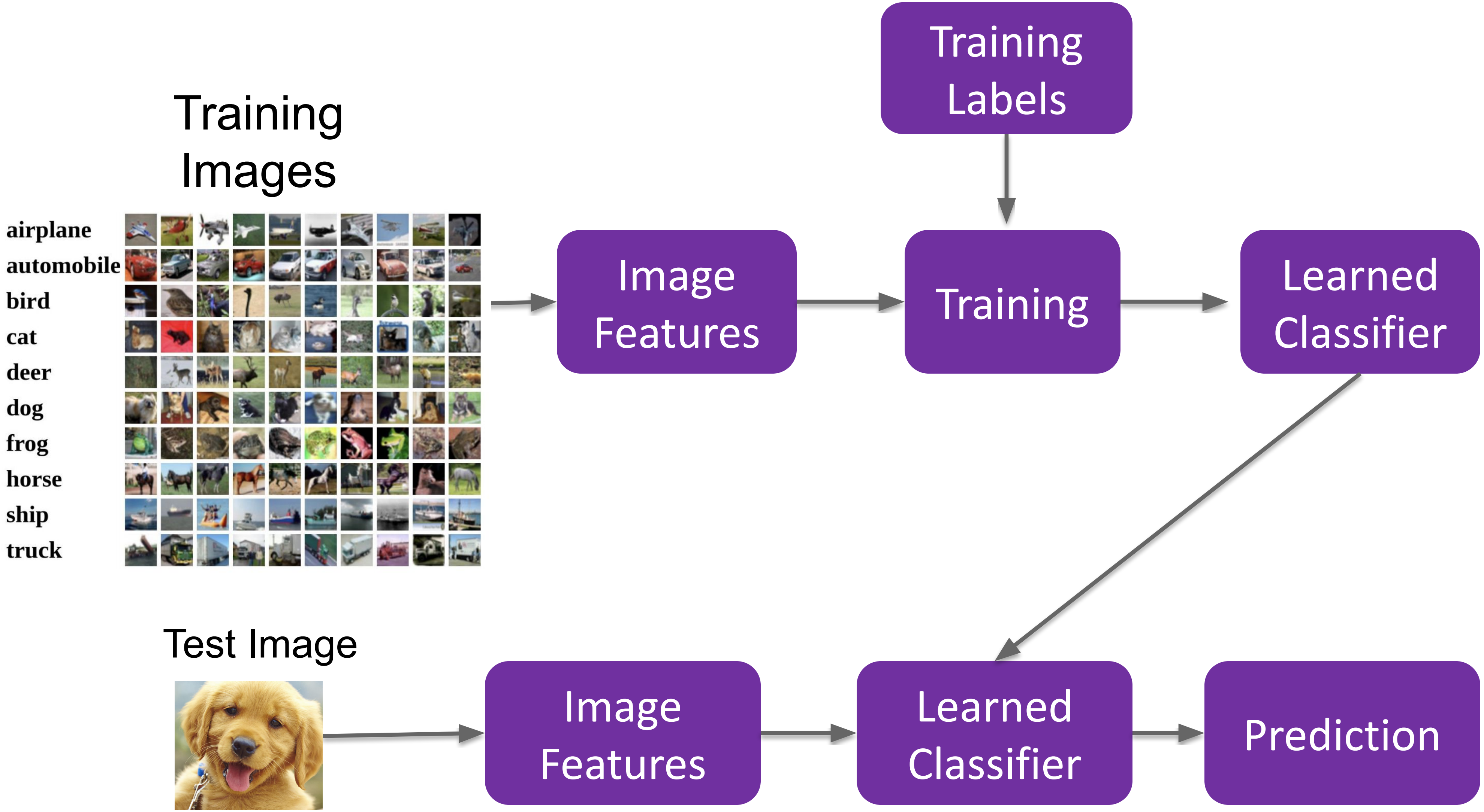
But they are NOT how the brain works, or even how neurons work.

Perceptron: for image classification



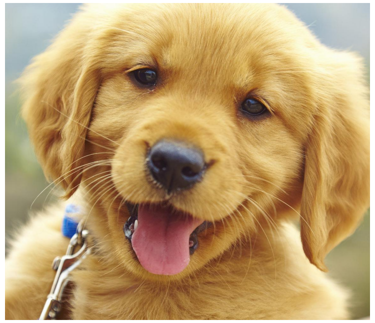


Let's revisit our simple recognition pipeline to explain where perceptrons fit in



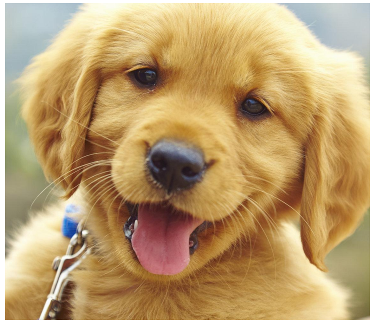
Remember we can featurize images into a vector

Image
Vector



Raw pixels
Raw pixels + (x,y)
PCA
LDA
BoW
BoW + spatial pyramids

Recall: we can featurize images into a vector

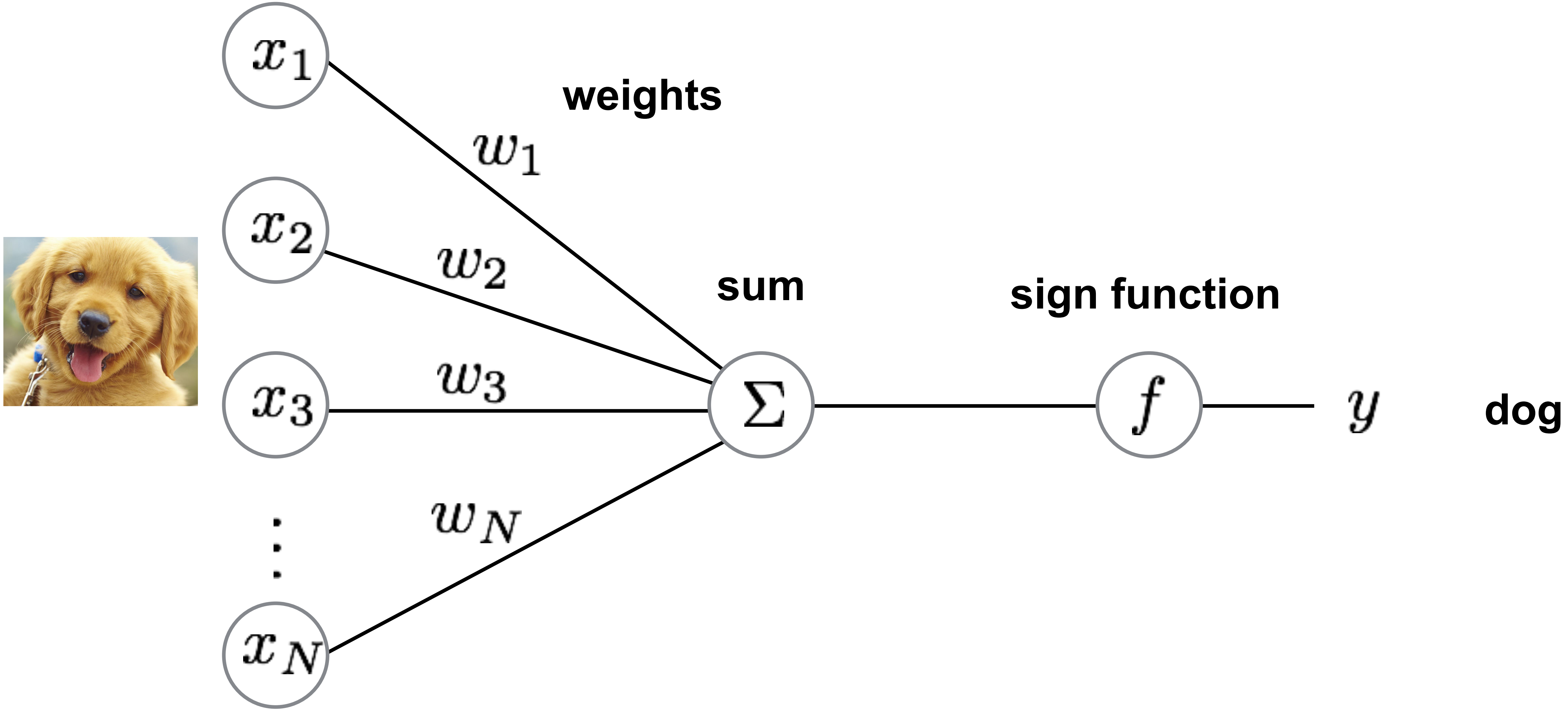


Raw pixels
Raw pixels + (x,y)
PCA
LDA
BoW
BoW + spatial pyramids

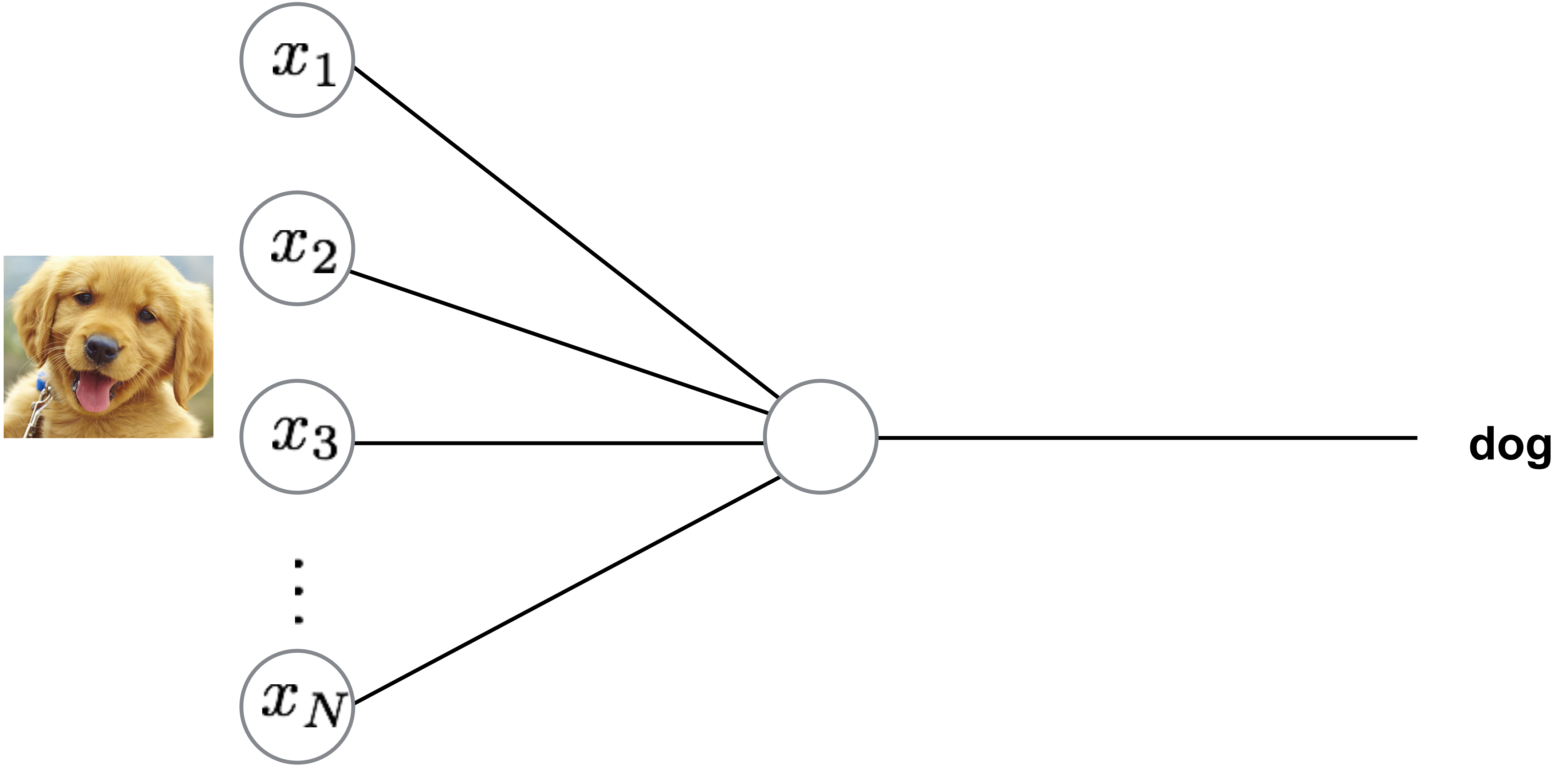
Image
Vector

x_1
x_2
x_3
...
...
...
...
...
...
...
...
...
x_n

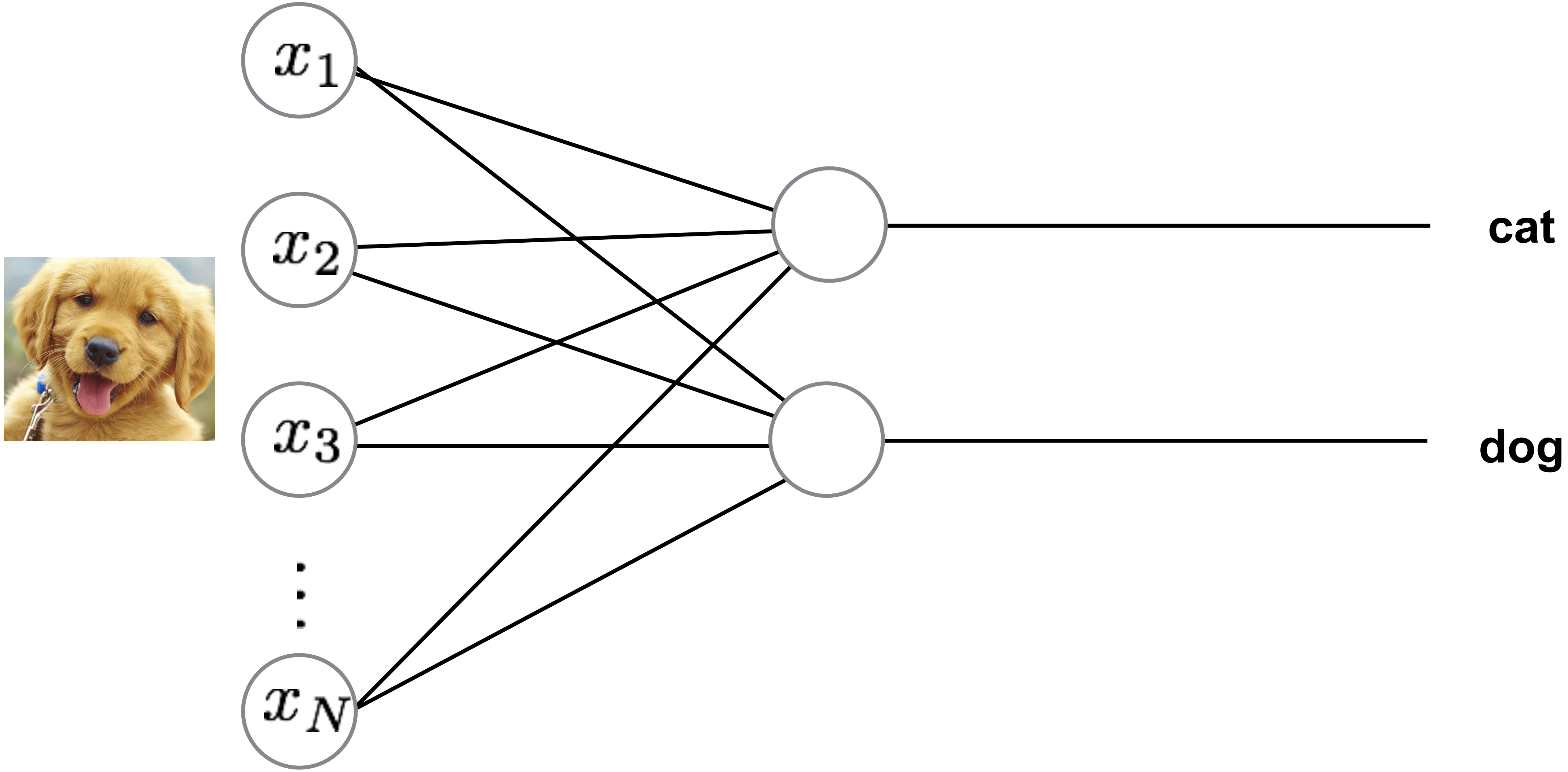
Perceptrons are a simple transformation that converts feature vectors into recognition scores



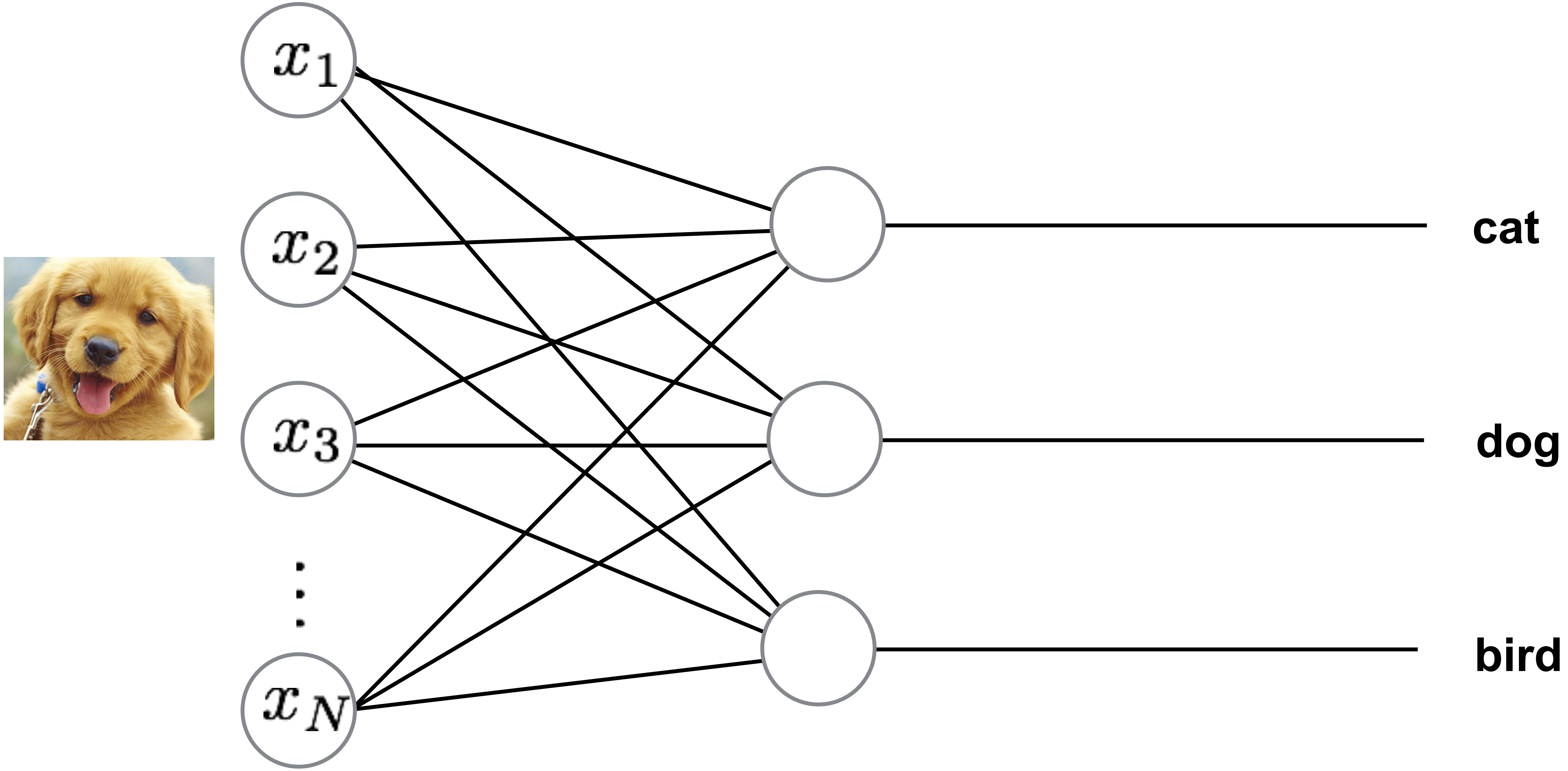
Perceptron: simplified view with one perceptron (produces 1 score for one category)



Perceptron: simplified view with two perceptrons (produces 2 scores with 2 categories)



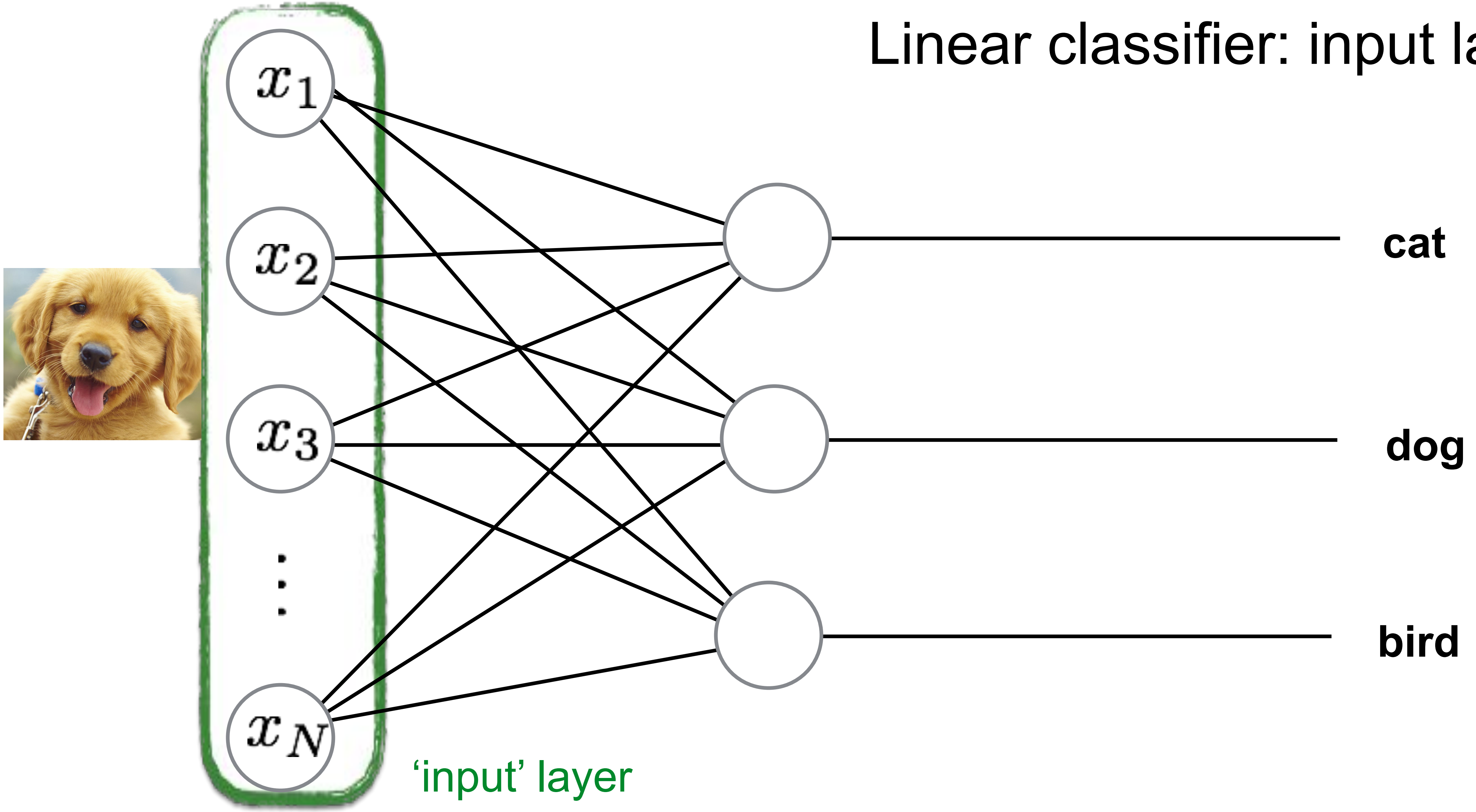
Linear classifier is a set of perceptrons produces one score for every category



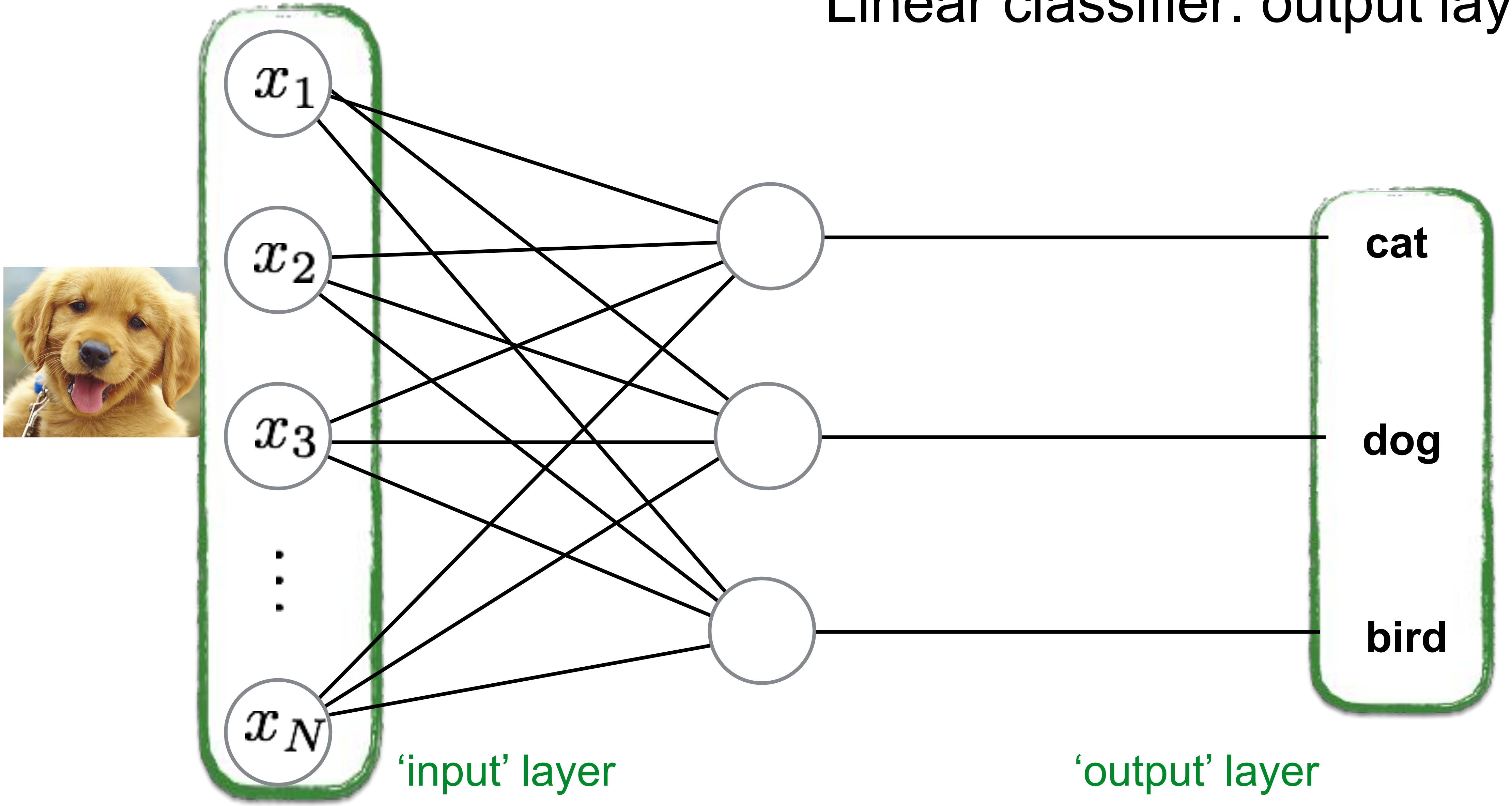
Today's agenda

- Perceptron
- Linear classifier
- Loss function
- Gradient descent and backpropagation
- Neural networks

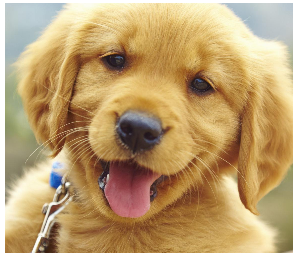
Linear classifier: input layer



Linear classifier: output layer



Linear classifier: mathematical formulation



x_1

x_2

x_3

$\vdots$

x_N

$f(x, W)$

10 numbers giving class scores

W
weights or parameters

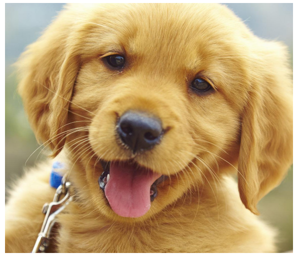
Linear classifier: mathematical formulation with RGB features

$f(x, W) = Wx$

$x = 3072 \times 1$

$W = ?$

Q. What is the shape of W?



(32x32x3)
3072 dimensional vector

x_1

x_2

x_3

$\vdots$

x_N

$f(x, W)$

W
weights or parameters

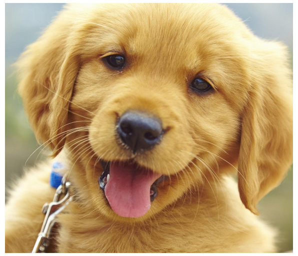
10 numbers giving class scores

Linear classifier: mathematical formulation with RGB features

$f(x, W) = Wx$

$x = 3072 \times 1$

$W = 10 \times 3072$



(32x32x3)
3072 dimensional vector

x_1

x_2

x_3

$\vdots$

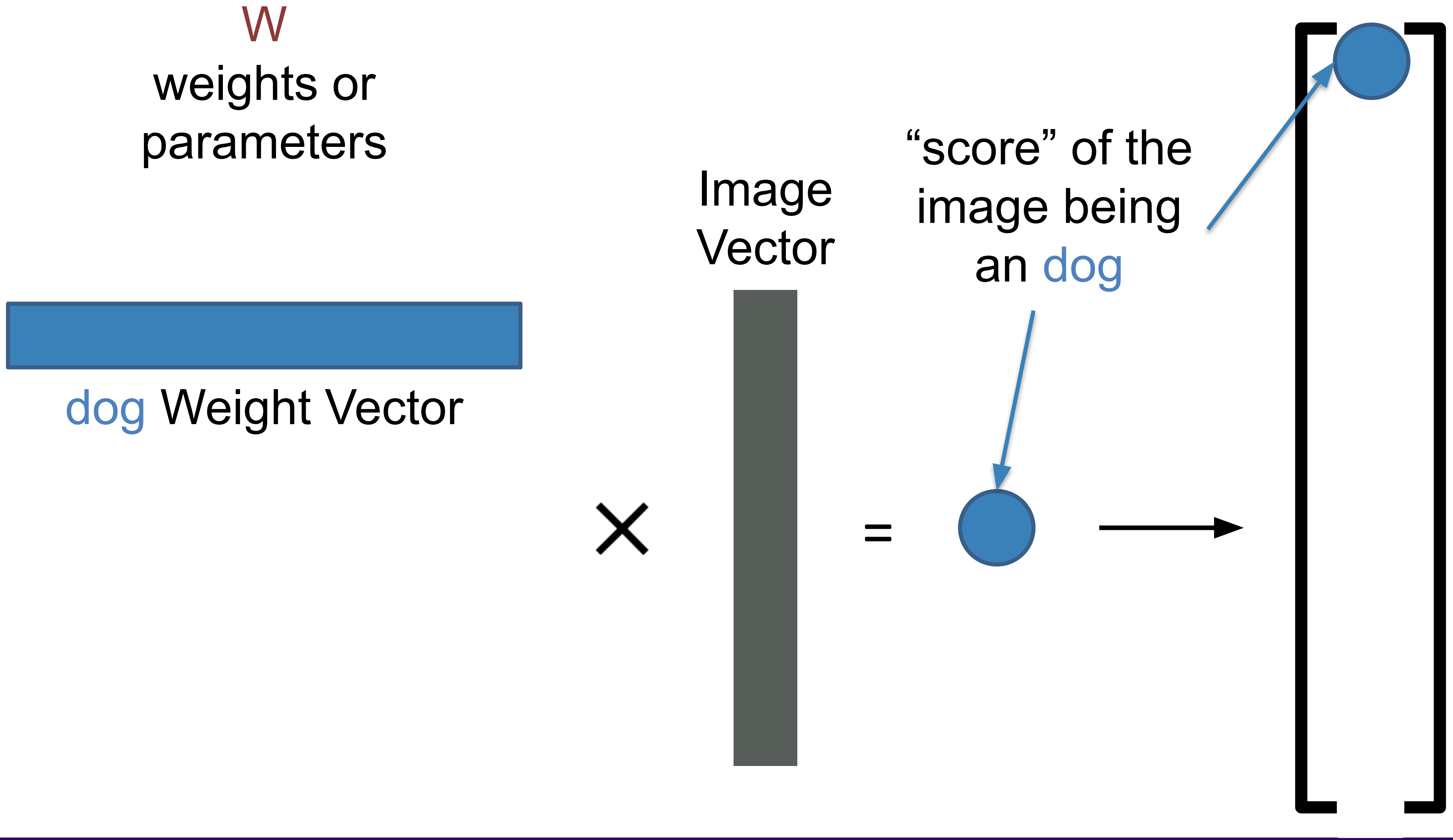
x_N

$f(x, W)$

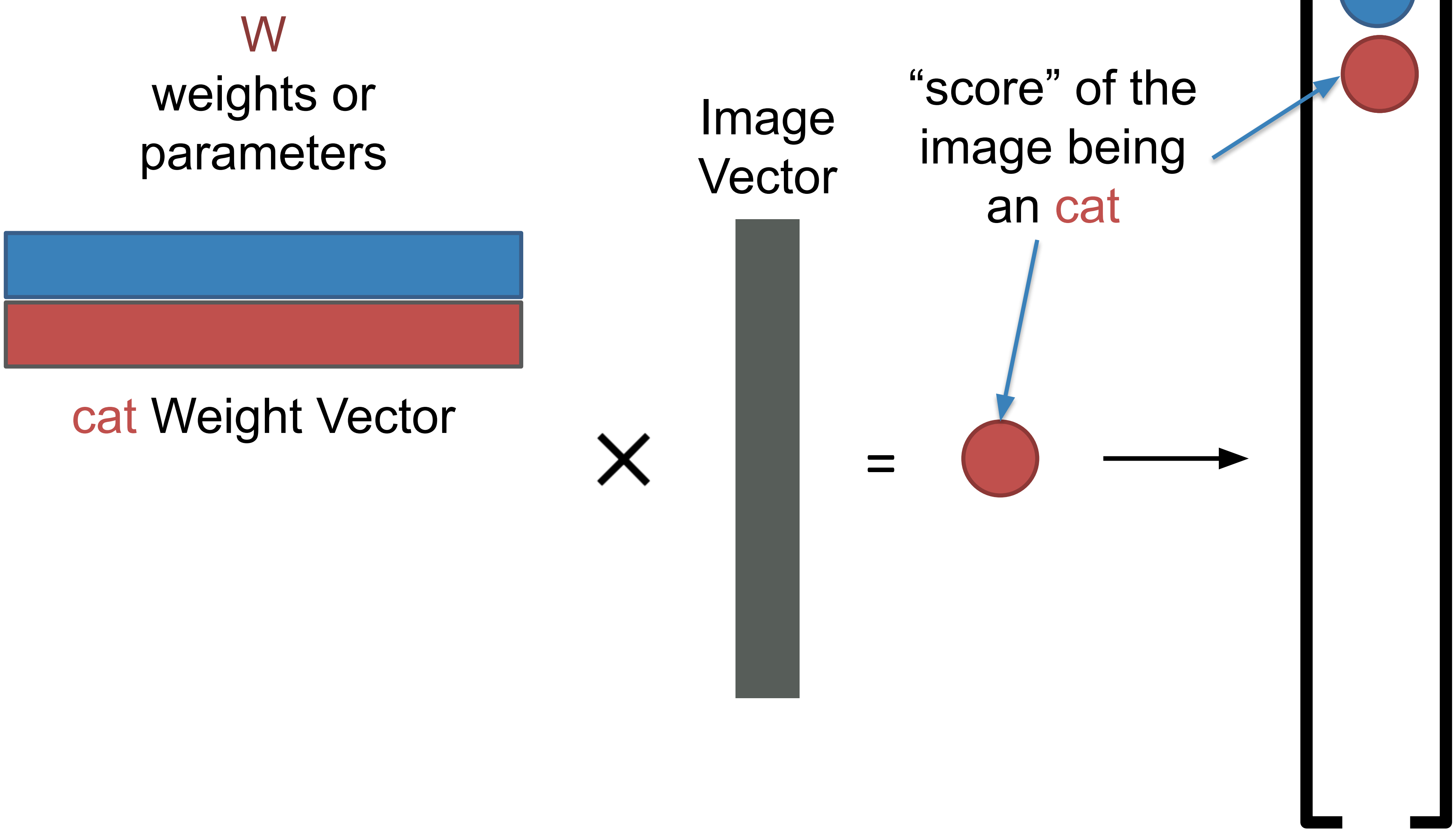
W
weights or parameters

10 numbers giving class scores

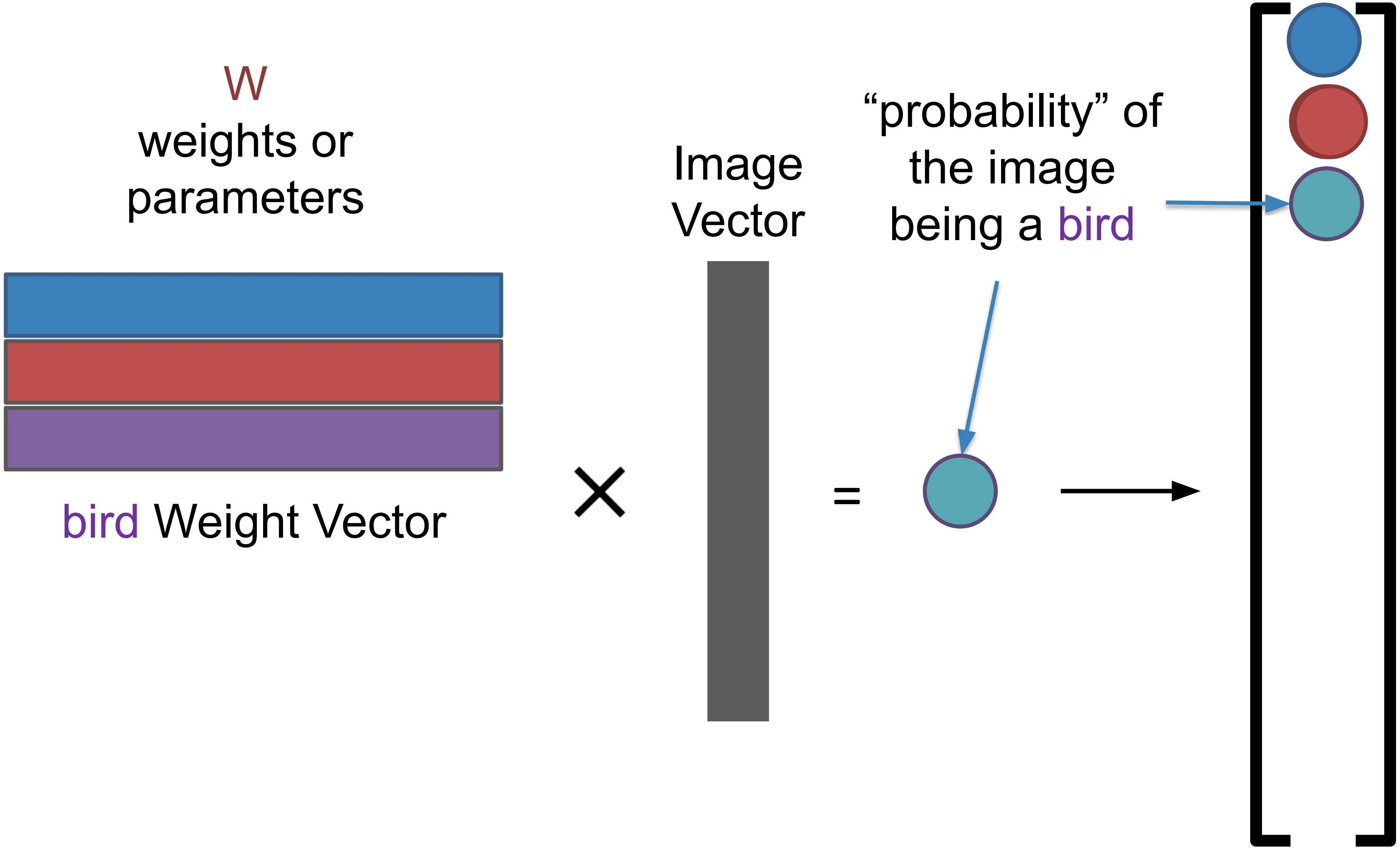
Linear classifier: function visualized



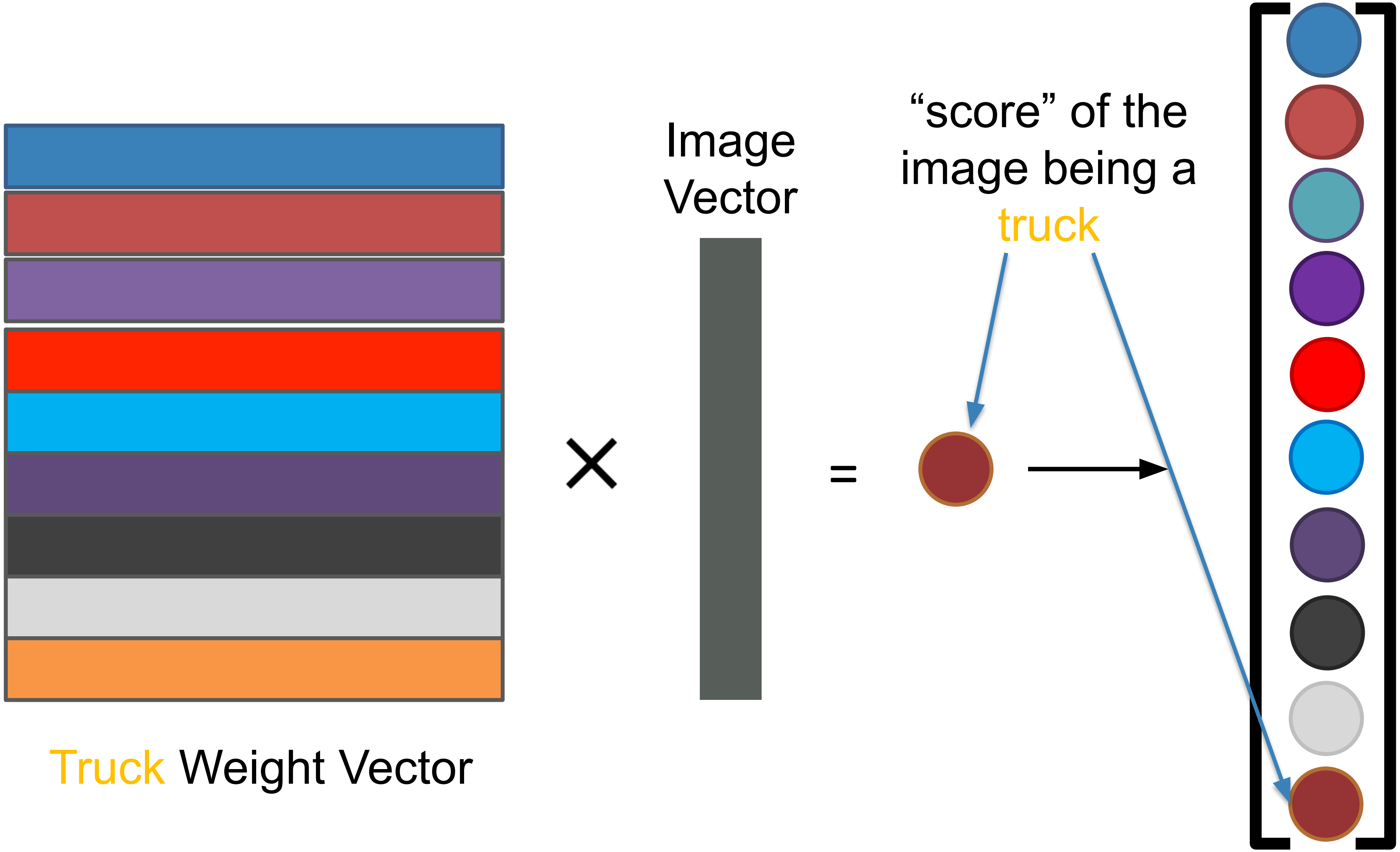
Linear classifier: function visualized



Linear classifier: function visualized

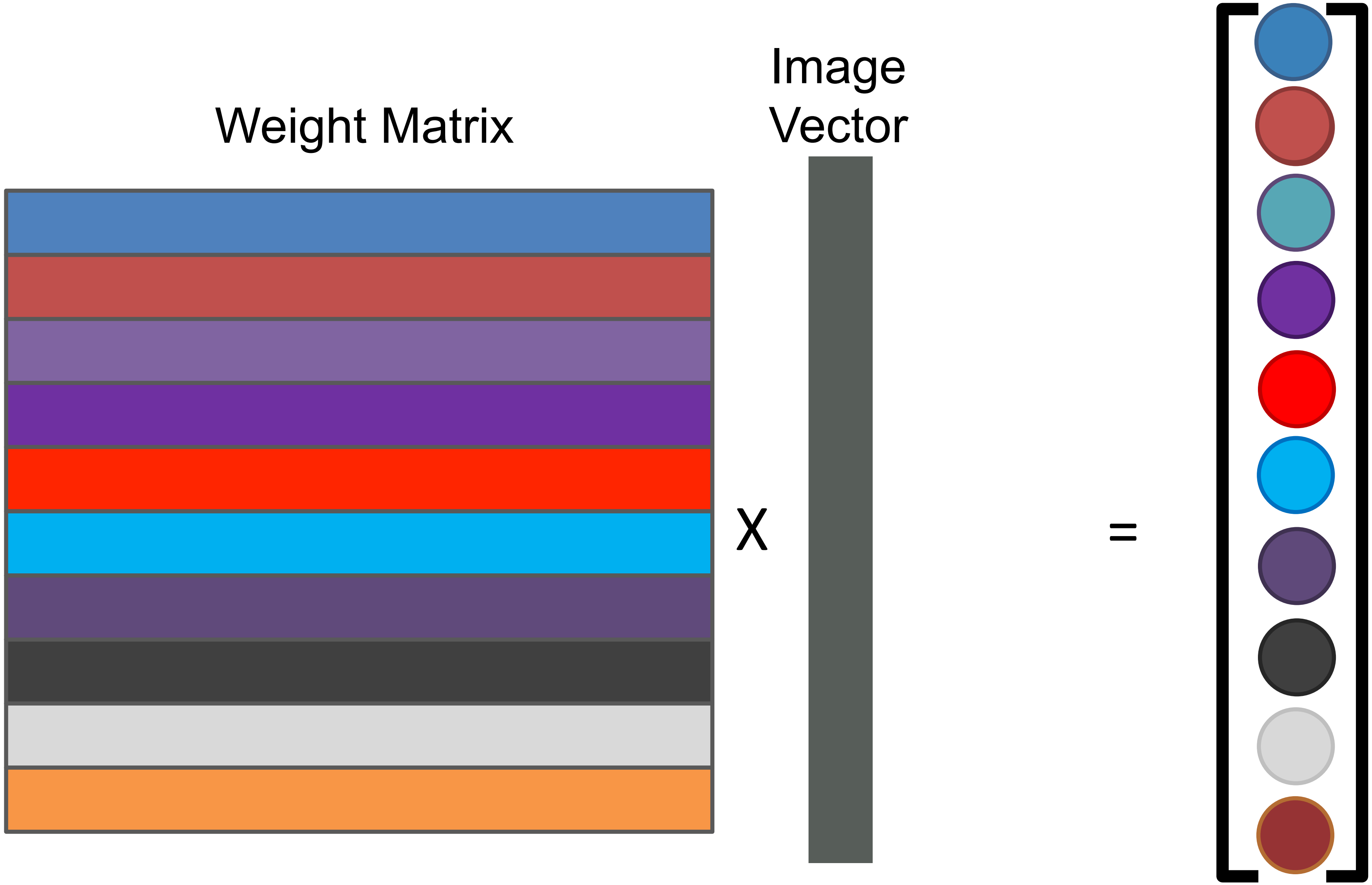


Linear classifier: function visualized

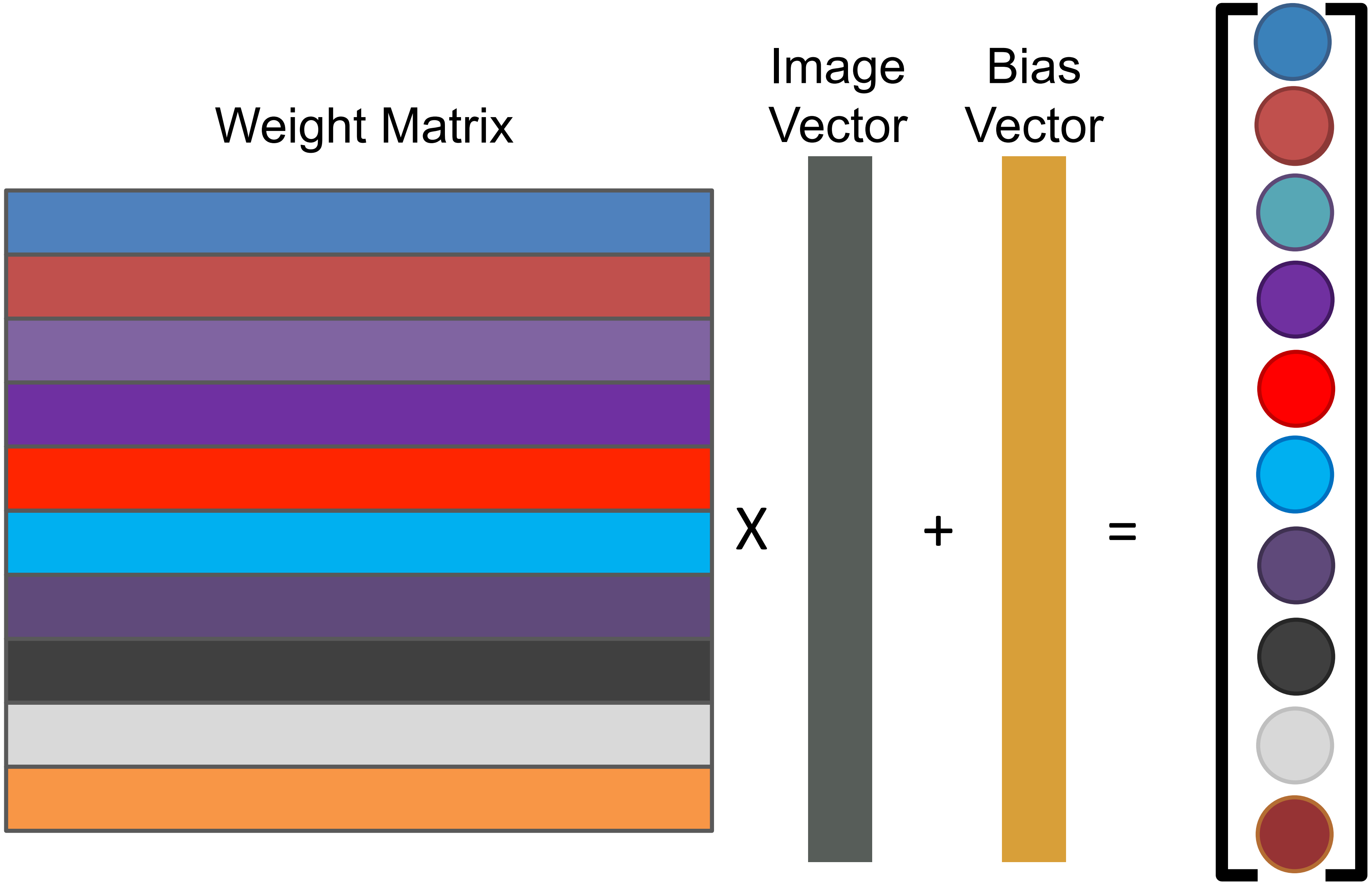


Truck Weight Vector

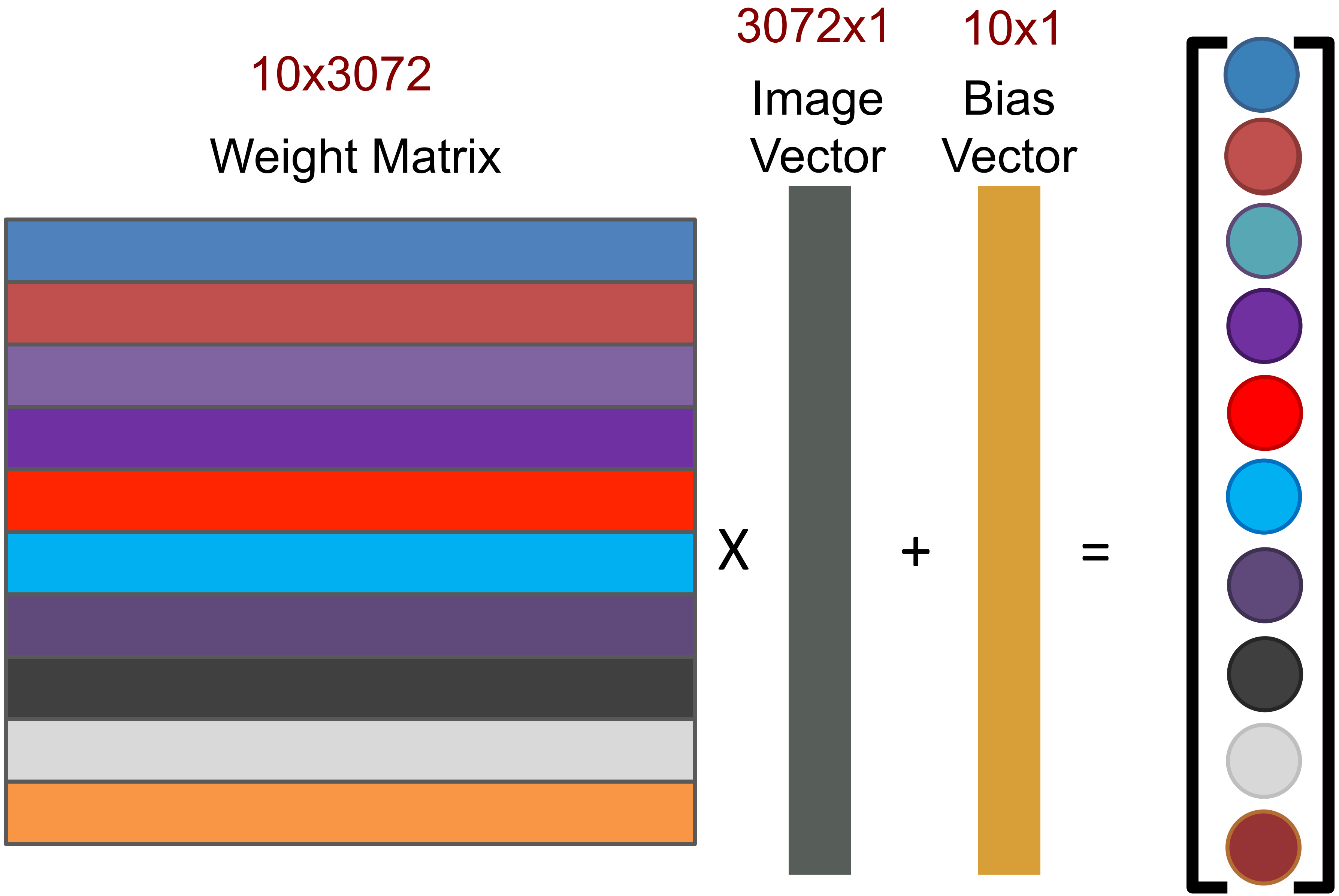
Linear classifier: function visualized



Linear classifier: bias vector



Linear classifier: size

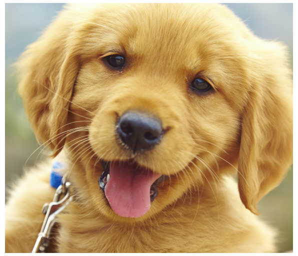


Linear classifier: Making a classification

234
94
-104
22
1
27.3
-13
-123
99
-82

$$\xrightarrow{argmax}$$

dog

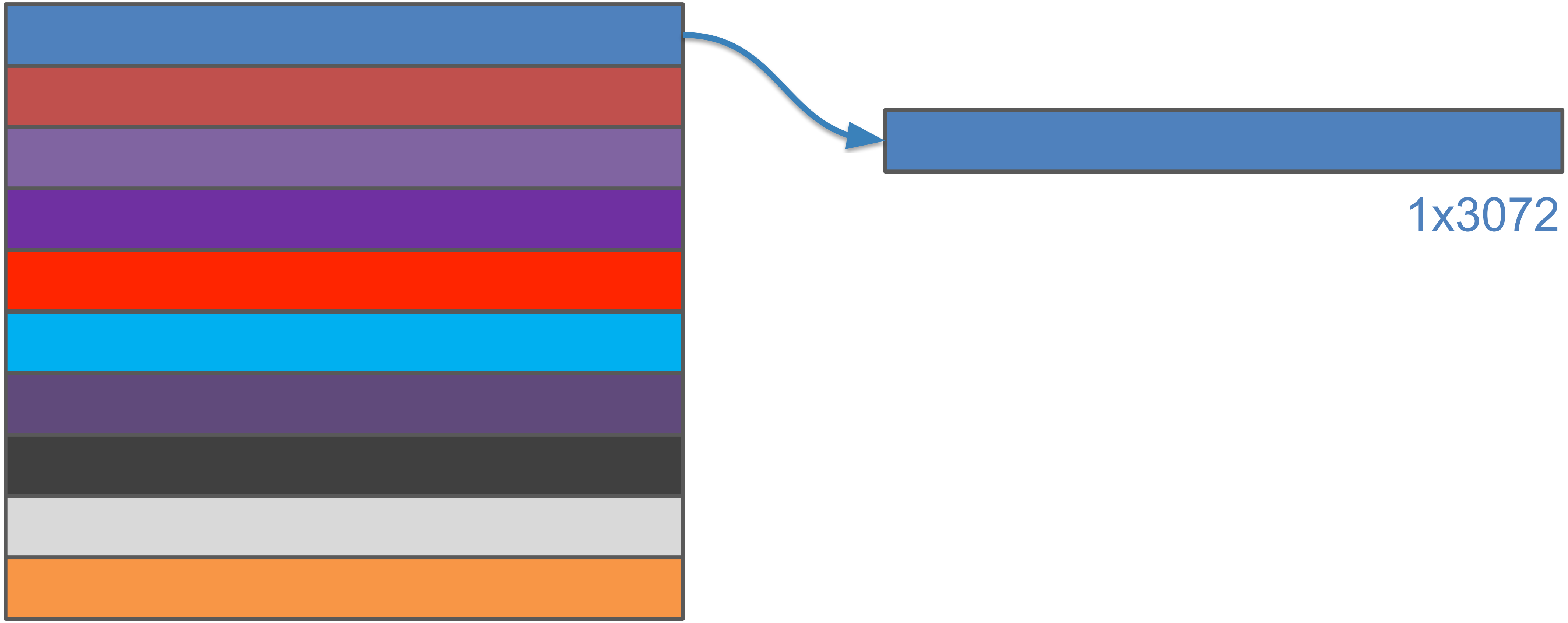


Interpreting the weights

- Assume our weights are trained on the CIFAR 10 dataset with raw pixels:

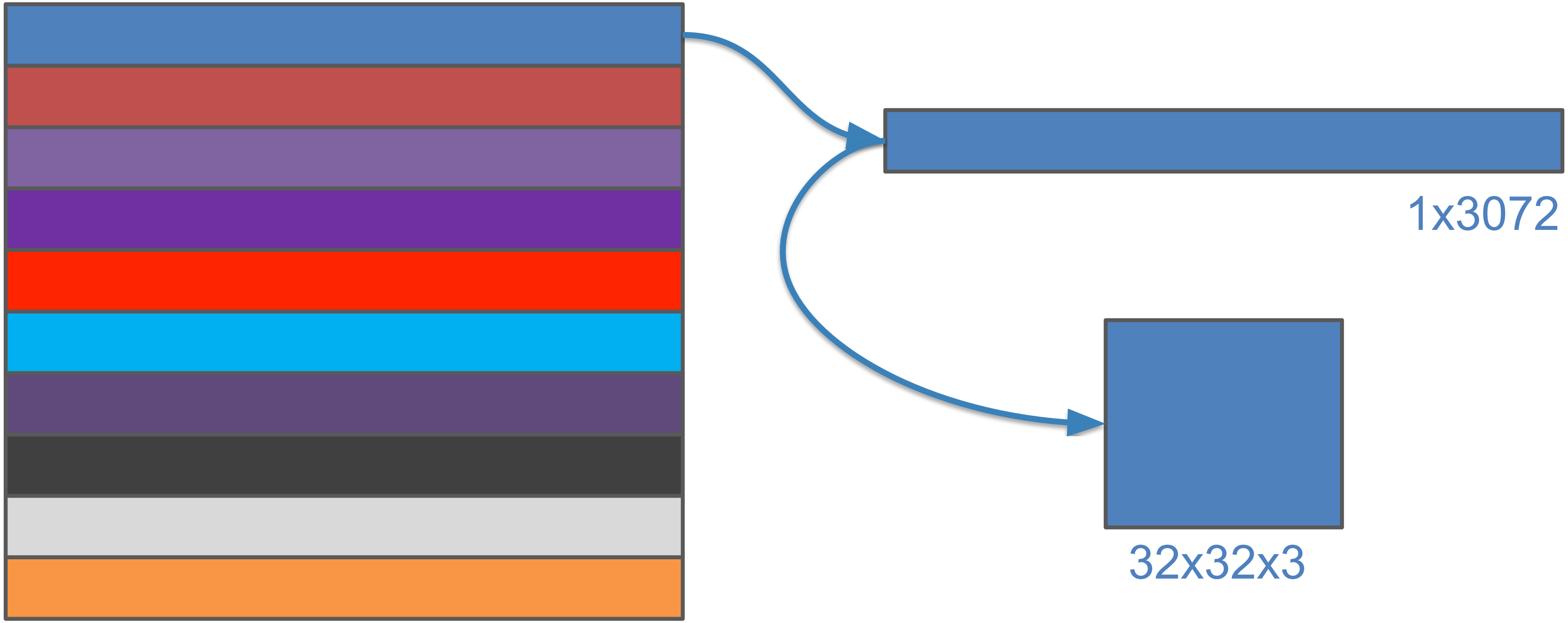
Interpreting the weights as templates

Let us look at each row of the weight matrix



Interpreting the weights as templates

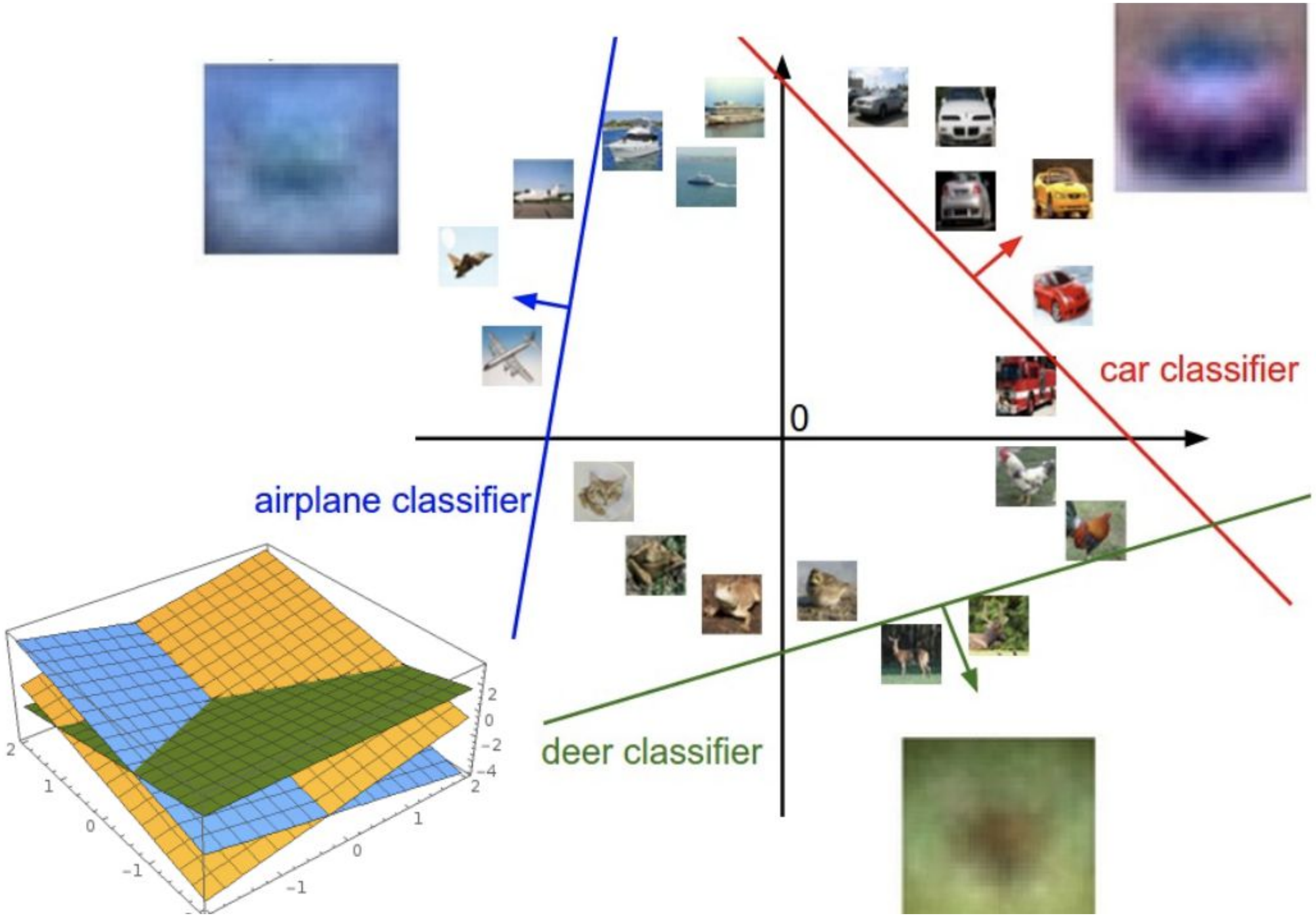
We can reshape the vector back in to the shape of an image



Let's visualize what the templates look like

We can reshape the row back to the shape of an image

Interpreting the weights geometrically



- Assume the image vectors are in 2D space to make it easier to visualize.

Plot created using [Wolfram Cloud]

Today's agenda

- Perceptron
- Linear classifier
- Loss function
- Gradient descent and backpropagation
- Neural networks

Training linear classifiers

We need to learn how to pick the weights in the first place.

Formally, we need to find *W* such that

$$\min_W \text{Loss}(y, \hat{y})$$

Where y is the true label, $\hat{y}$ is the model's predicted label.

All we have to do is define a loss function!

Given training data:

$$y = wx$$

x	y
[2 3]	7
[5 1]	11
[8 8]	24

What do you think is a good approximation weight parameter for this data point?

W = [? ?]

Given training data:

$$y = wx$$

x	y
[2 3]	7
[5 1]	11
[8 8]	24

What do you think is a good approximation weight parameter for this data point?

W = [2 1]

Properties of a loss function

Given several training examples: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$

and a perceptron: $\hat{y} = wx$

where x is image and y is (integer) label (0 for dog, 1 for cat, etc)

A loss function $\mathrm{L_i}(y_\mathrm{i}, \hat{y}_\mathrm{i})$ tells us how good our current classifier

- When the classifier predicts correctly, the loss should be low
- When the classifier makes mistakes, the loss should be high

Properties of a loss function

Given several training examples: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$

and a perceptron: $\hat{y} = wx$

where x is image and y is (integer) label (0 for dog, 1 for cat, etc)
Loss over the entire dataset is an average of loss over examples

$$L = \frac{1}{N} \sum_{i=1}^{N} L_i(y_i, \hat{y}_i)$$

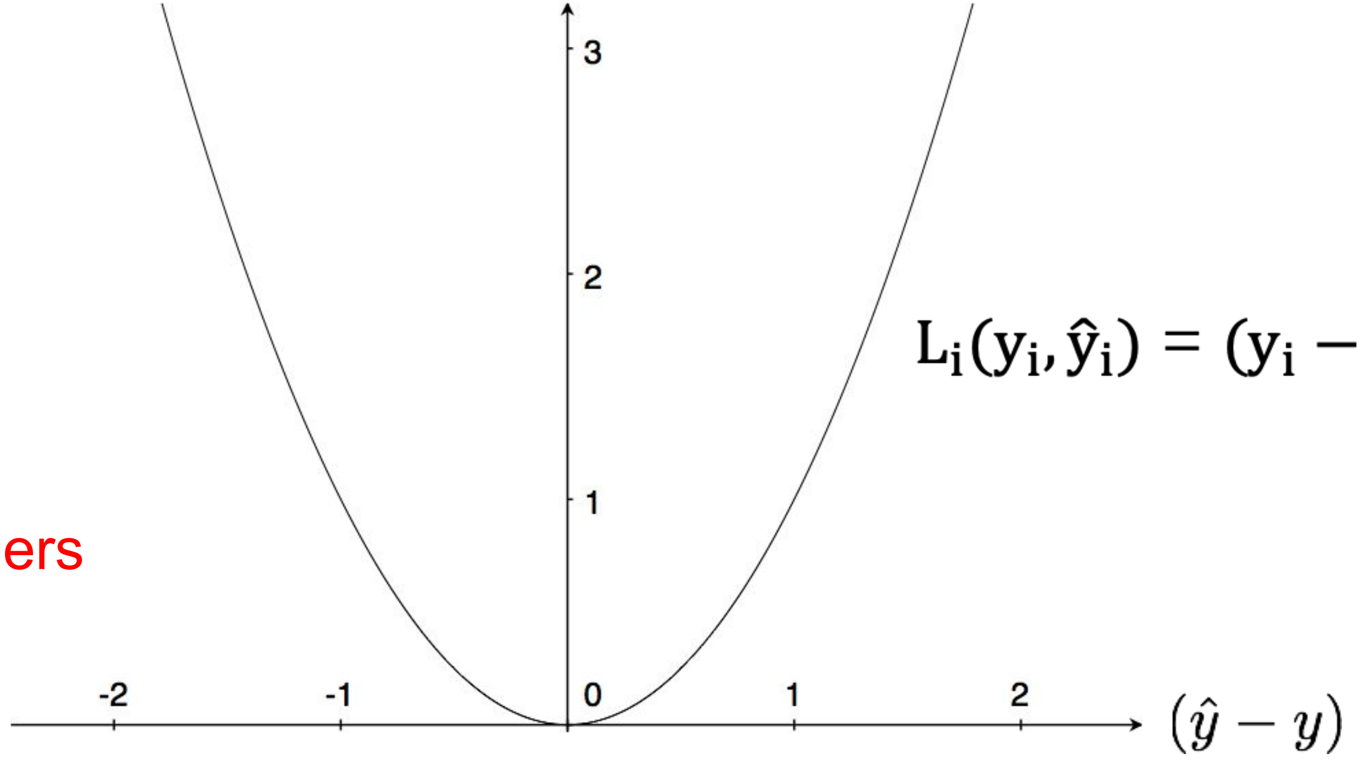
How do we choose the loss function L_i ?

YOU get to chose the loss function!

(some are better than others depending on what you want to do)

Squared Error (L2)

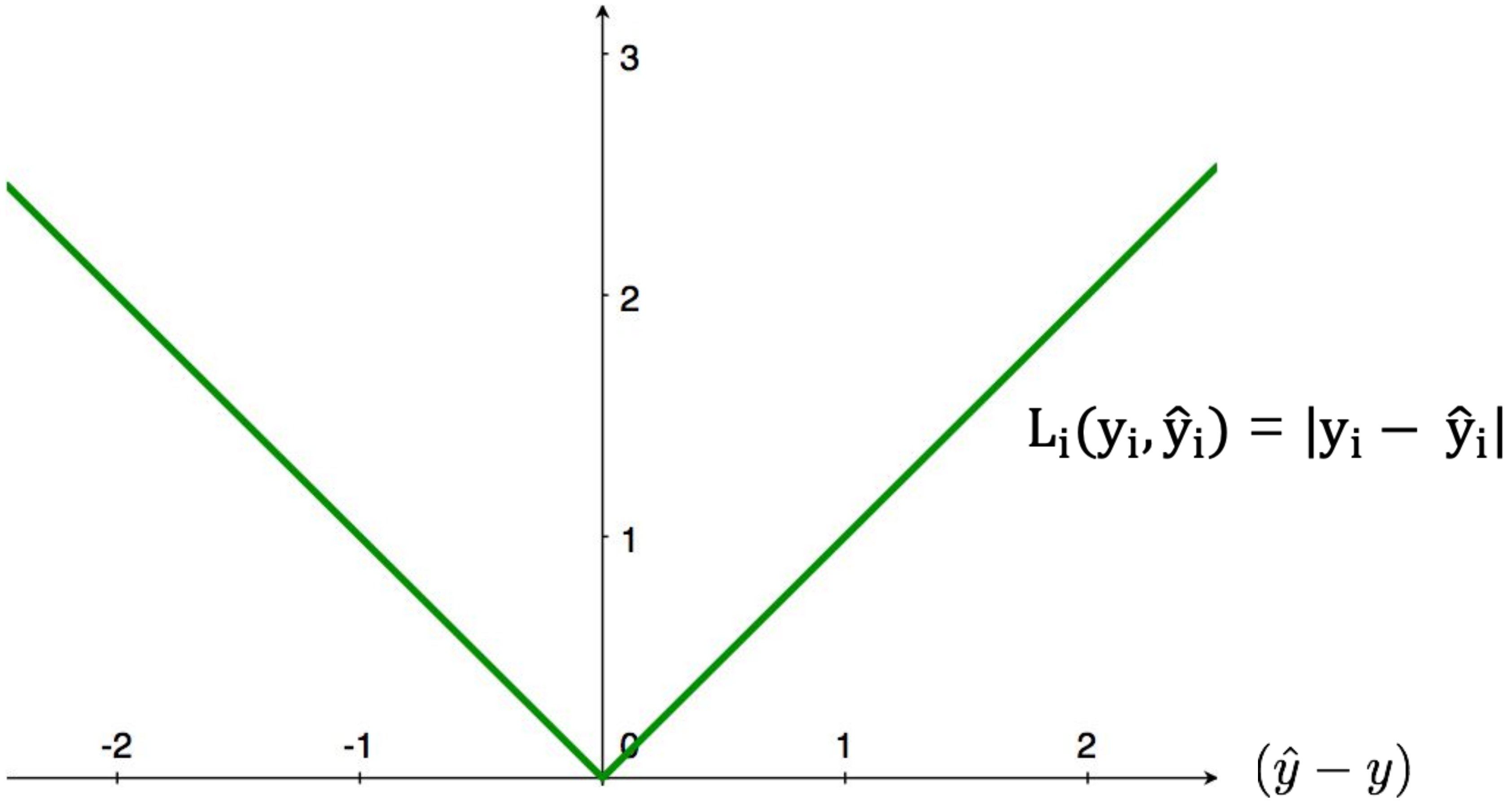
(a popular loss function) ((why?))



$$L_i(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$$

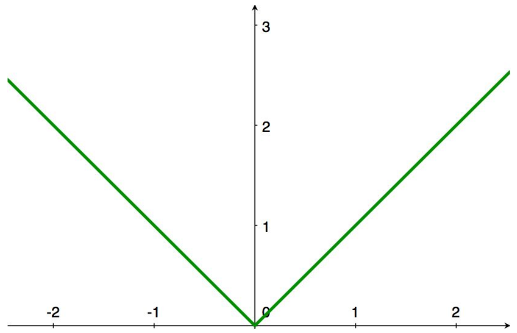
Not robust to outliers

L1 loss



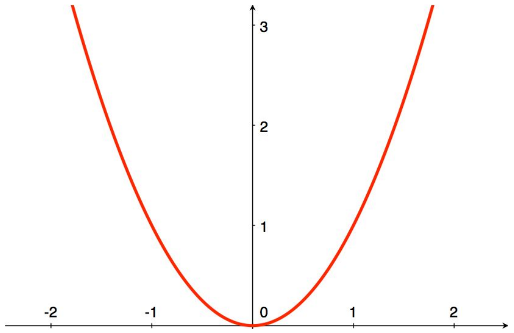
L1 Loss

$$L_i(y_i, \hat{y}_i) = |y_i - \hat{y}_i|$$



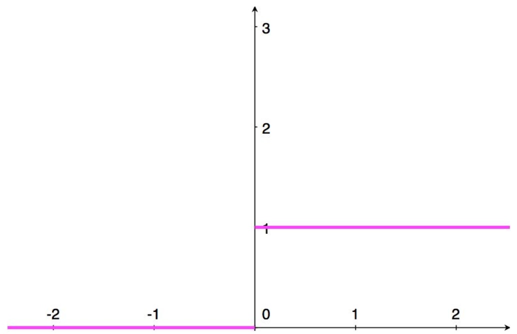
L2 Loss

$$L_i(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$$



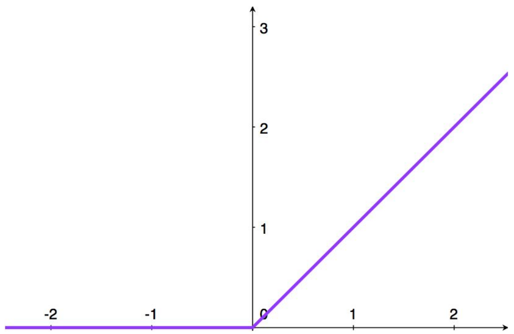
Zero-One Loss

$$L_i(y_i, \hat{y}_i) = 1||y_i \neq \hat{y}_i||$$



Hinge Loss

$$L_i(y_i, \hat{y}_i) = \max(0, \ 1 - y_i \hat{y}_i)$$



Softmax Classifier (Multinomial Logistic Regression)

- It allows us to treat the outputs of a model as probabilities for each class
- common way of measuring distance between probability distributions is Kullback-Leibler (KL) divergence:

$$D_{KL} = \sum_{y} P(y) \log \frac{P(y)}{Q(y)}$$

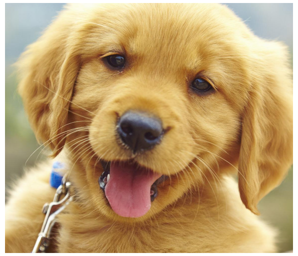
- where *P* is the ground truth distribution and *Q* is the model's output score distribution

Softmax Classifier (Multinomial Logistic Regression)

KL divergence: $$D_{KL} = \sum_y P(y) \log \frac{P(y)}{Q(y)}$$

In our case, *P* is only non-zero for correct class

For example, consider the case we only have 3 classes:



$$\boxed{\begin{array}{cl} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} & \begin{array}{l} \text{dog} \\ \text{cat} \\ \text{bird} \end{array} \end{array}}$$

correct outputs

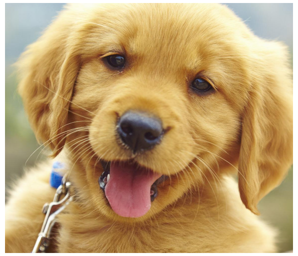
Softmax Classifier (Multinomial Logistic Regression)

KL divergence:

$$D_{KL} = \sum_{y} P(y)\, log \frac{P(y)}{Q(y)}$$

$$= -\log Q(y) \text{ when } y = dog$$

$$= -\log Prob[f(x_i, W) = y_i]$$



$$\begin{array}{cl} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} & \begin{array}{l} \text{dog} \\ \text{cat} \\ \text{bird} \end{array} \end{array}$$

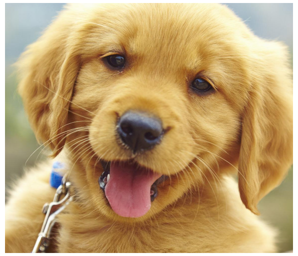
correct outputs

Softmax Classifier (Multinomial Logistic Regression)

$$L_i = -\log Prob[f(x_i, W) == y_i]$$

Remember our linear classifier: $\hat{y} = wx$

There are no limitations on the output space. Meaning that the model can output <0 or >1



$$\begin{bmatrix} 3.2 \\ 5.1 \\ -1.7 \end{bmatrix}$$

model outputs

$$\begin{array}{cl} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} & \begin{array}{l} \text{dog} \\ \text{cat} \\ \text{bird} \end{array} \end{array}$$

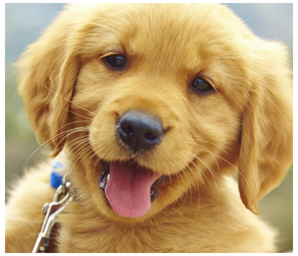
correct outputs

Softmax Classifier (Multinomial Logistic Regression)

$$L_i = -\log Prob[f(x_i, W) == y_i]$$

We need a mechanism to convert or normalize the output into probability range [0, 1]

Solution: SOFTMAX: $Prob[f(x_i, W) == k] = \frac{e^{\hat{y}_k}}{\sum_j e^{\hat{y}_j}}$



$$\begin{bmatrix} 3.2 \\ 5.1 \\ -1.7 \end{bmatrix}$$

model outputs

$$\begin{array}{cl} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} & \begin{array}{l} \text{dog} \\ \text{cat} \\ \text{bird} \end{array} \end{array}$$

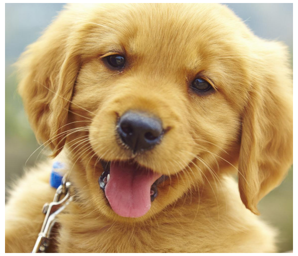
correct outputs

Softmax Classifier (Multinomial Logistic Regression)

$$L_i = -\log Prob[f(x_i, W) == y_i]$$

We need a mechanism to convert or normalize the output into probability range [0, 1]

Solution: SOFTMAX: $Prob[f(x_i, W) == k] = \frac{e^{\hat{y}_k}}{\sum_j e^{\hat{y}_j}}$



$$\begin{bmatrix} 3.2 \\ 5.1 \\ -1.7 \end{bmatrix} \xrightarrow{\text{exp}} \begin{bmatrix} 24.5 \\ 164 \\ 0.18 \end{bmatrix} \xrightarrow{\text{norm}} \begin{bmatrix} 0.13 \\ 0.87 \\ 0.00 \end{bmatrix} \xleftrightarrow{\text{loss}} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{matrix} \text{dog} \\ \text{cat} \\ \text{bird} \end{matrix}$$

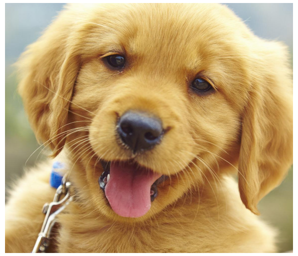
model outputs | probabilities | correct outputs

Softmax Classifier (Multinomial Logistic Regression)

$$L_i = -\log Prob[f(x_i, W) == y_i]$$

In this case, what is the loss:

$$L_i = ??$$



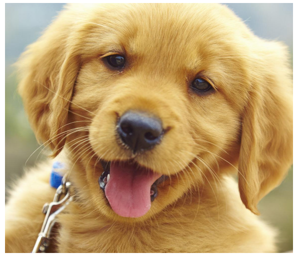
$$\underbrace{\begin{bmatrix} 3.2 \\ 5.1 \\ -1.7 \end{bmatrix}}_{\text{model outputs}} \xrightarrow{\text{exp}} \begin{bmatrix} 24.5 \\ 164 \\ 0.18 \end{bmatrix} \xrightarrow{\text{norm}} \underbrace{\begin{bmatrix} 0.13 \\ 0.87 \\ 0.00 \end{bmatrix}}_{\text{probabilities}} \xleftrightarrow{\text{loss}} \underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}}_{\text{correct outputs}} \begin{array}{l} \text{dog} \\ \text{cat} \\ \text{bird} \end{array}$$

Softmax Classifier (Multinomial Logistic Regression)

$$L_i = -\log Prob[f(x_i, W) == y_i]$$

In this case, what is the loss:

$$L_i = -\log(0.13) = 2.04$$



$$\begin{bmatrix} 3.2 \\ 5.1 \\ -1.7 \end{bmatrix} \xrightarrow{\text{exp}} \begin{bmatrix} 24.5 \\ 164 \\ 0.18 \end{bmatrix} \xrightarrow{\text{norm}} \begin{bmatrix} 0.13 \\ 0.87 \\ 0.00 \end{bmatrix} \xleftrightarrow{\text{loss}} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{array}{l} \text{dog} \\ \text{cat} \\ \text{bird} \end{array}$$

model outputs

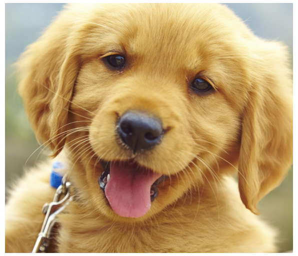
probabilities

correct outputs

Softmax Classifier (Multinomial Logistic Regression)

$$L_i = -\log Prob[f(x_i, W) == y_i]$$

what is the minimum and maximum values that the loss can be?

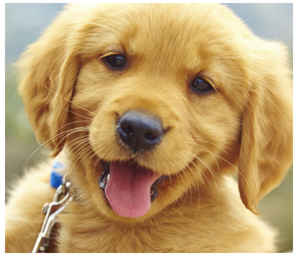


$$\underbrace{\begin{bmatrix} 3.2 \\ 5.1 \\ -1.7 \end{bmatrix}}_{\text{model outputs}} \xrightarrow{\text{exp}} \begin{bmatrix} 24.5 \\ 164 \\ 0.18 \end{bmatrix} \xrightarrow{\text{norm}} \underbrace{\begin{bmatrix} 0.13 \\ 0.87 \\ 0.00 \end{bmatrix}}_{\text{probabilities}} \xleftrightarrow{\text{loss}} \underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}}_{\text{correct outputs}} \begin{matrix} \text{dog} \\ \text{cat} \\ \text{bird} \end{matrix}$$

Softmax Classifier (Multinomial Logistic Regression)

$$L_i = -\log Prob[f(x_i, W) == y_i]$$

At initialization, all the weights will be random. In this case, we can assume that the outputs will have the same probabilities, then what will the initial loss be?



$$\begin{bmatrix} 3.2 \\ 5.1 \\ -1.7 \end{bmatrix} \xrightarrow{\text{exp}} \begin{bmatrix} 24.5 \\ 164 \\ 0.18 \end{bmatrix} \xrightarrow{\text{norm}} \begin{bmatrix} 0.13 \\ 0.87 \\ 0.00 \end{bmatrix} \xleftrightarrow{\text{loss}} \begin{array}{c} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \end{array} \begin{array}{l} \text{dog} \\ \text{cat} \\ \text{bird} \end{array}$$

model outputs

probabilities

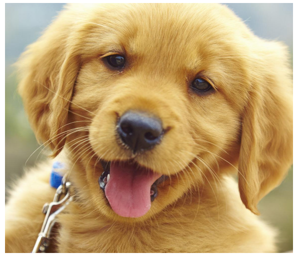
correct outputs

Softmax Classifier (Multinomial Logistic Regression)

$$L_i = -\log Prob[f(x_i, W) == y_i]$$

At initialization, all the weights will be random. In this case, we can assume that the outputs will have the same probabilities, then what will the initial loss be?

$$L_i = -\log\left(\frac{1}{C}\right) = \log(C) = \log(10) = 2.03$$



$$\begin{bmatrix} 3.2 \\ 5.1 \\ -1.7 \end{bmatrix} \xrightarrow{\text{exp}} \begin{bmatrix} 24.5 \\ 164 \\ 0.18 \end{bmatrix} \xrightarrow{\text{norm}} \begin{bmatrix} 0.13 \\ 0.87 \\ 0.00 \end{bmatrix} \xleftrightarrow{\text{loss}} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{matrix} \text{dog} \\ \text{cat} \\ \text{bird} \end{matrix}$$

model outputs

probabilities

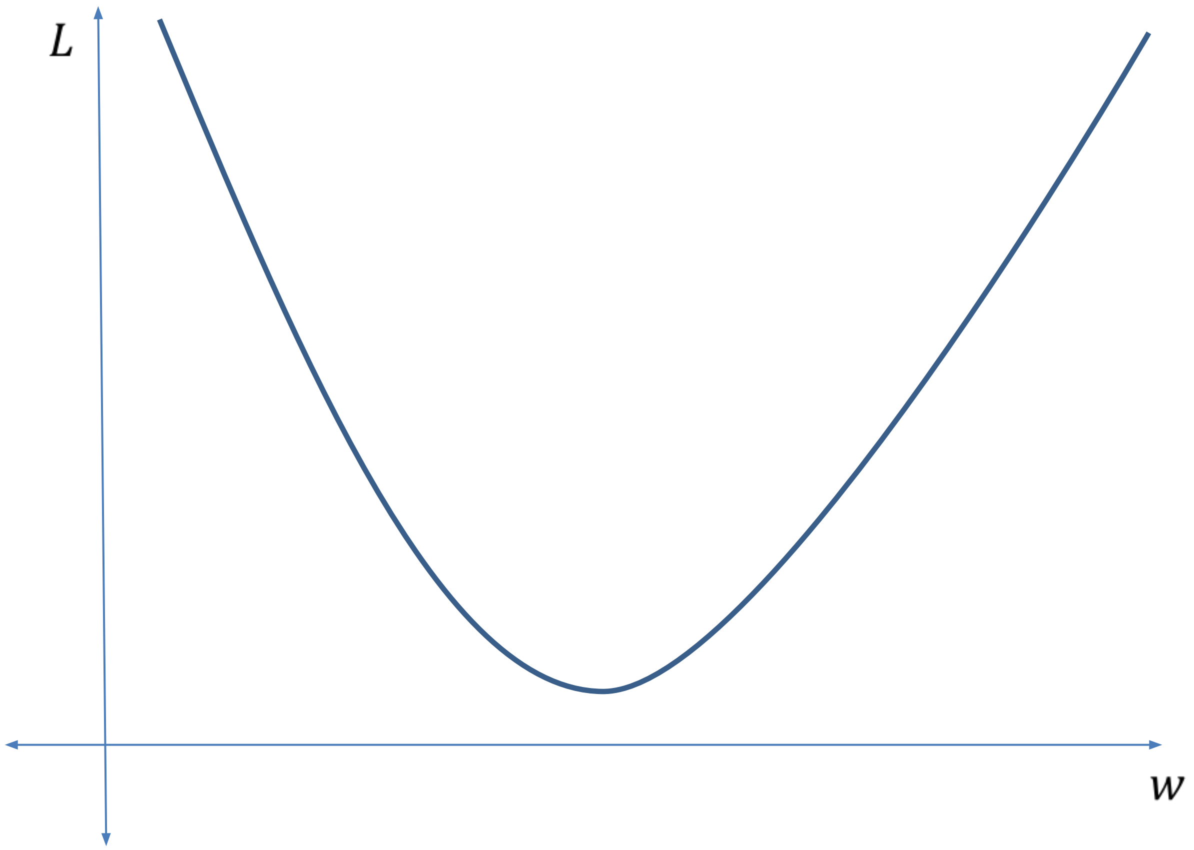
correct outputs

Today's agenda

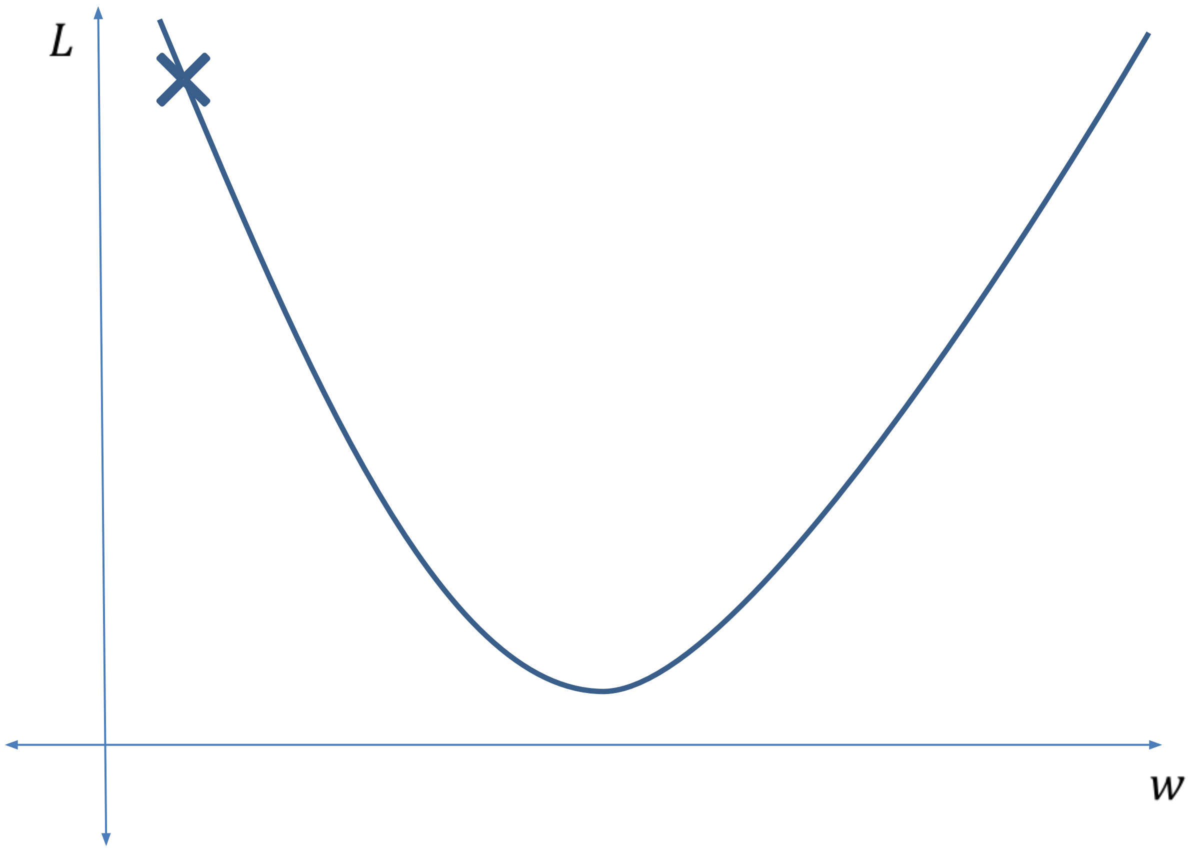
- Perceptron
- Linear classifier
- Loss function
- Gradient descent and backpropagation
- Neural networks

How do we find the weights that minimize the loss?

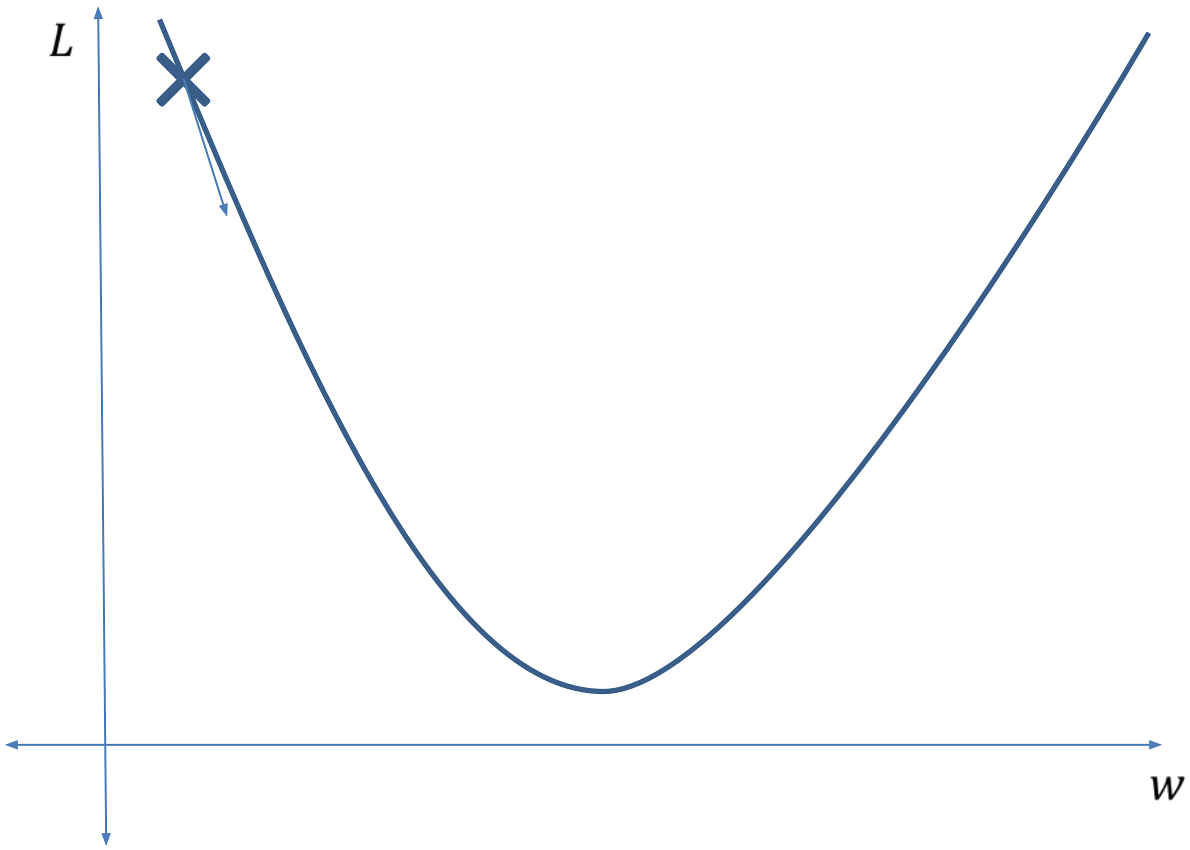
Gradient descent visualized: Minimizing loss



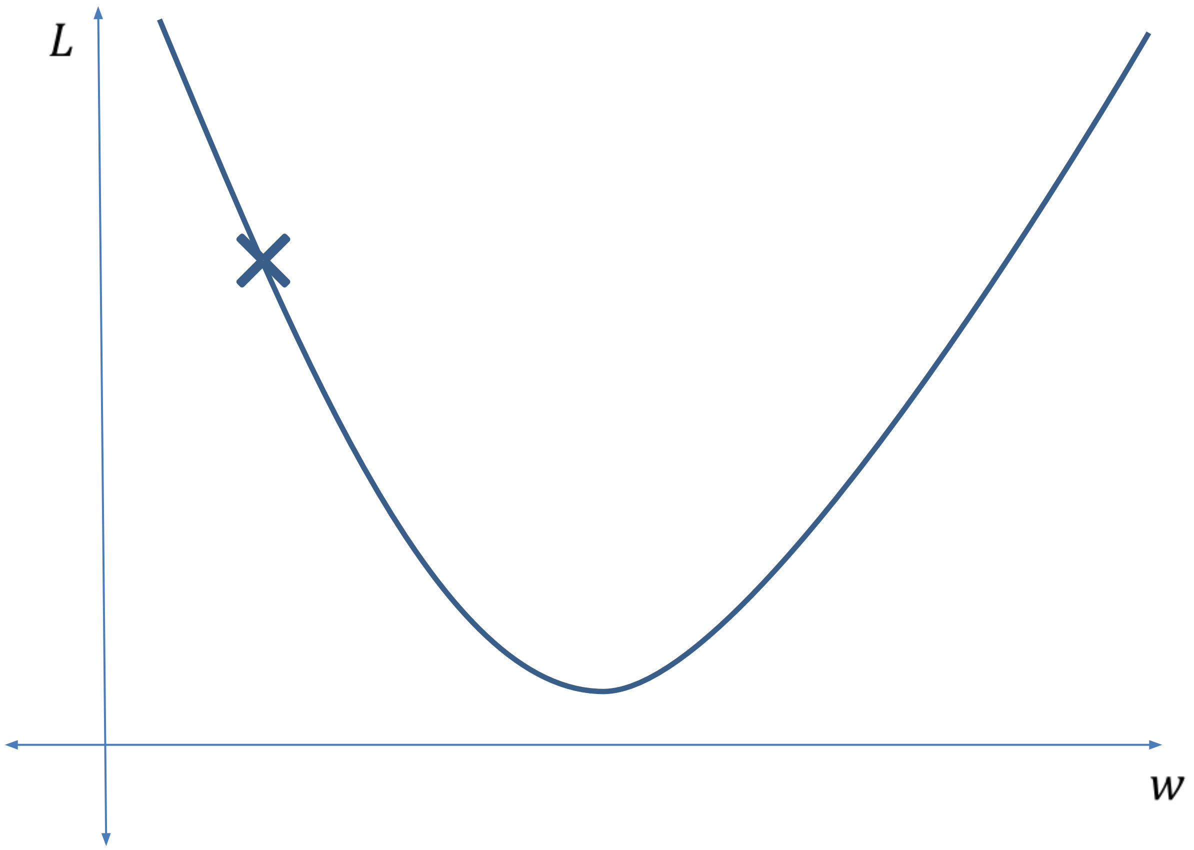
Gradient descent visualized: Minimizing loss



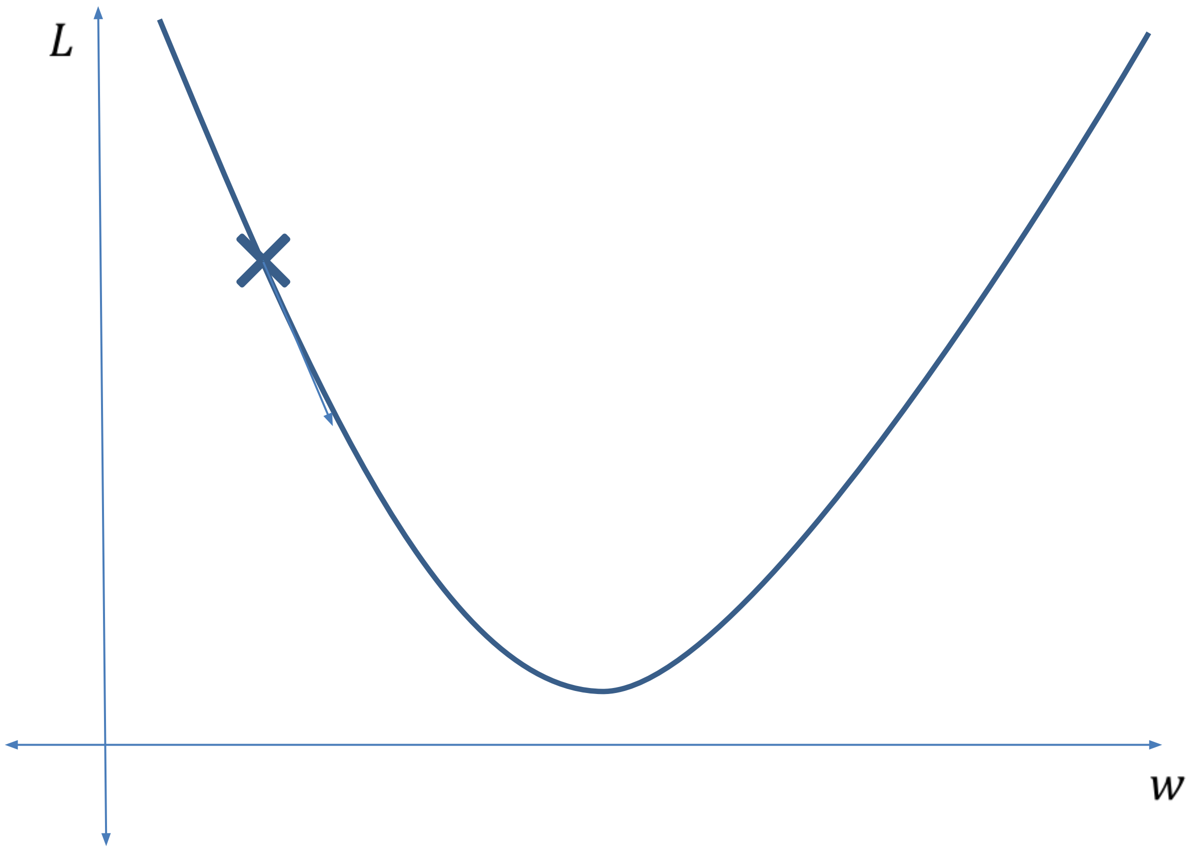
Gradient descent visualized: Minimizing loss



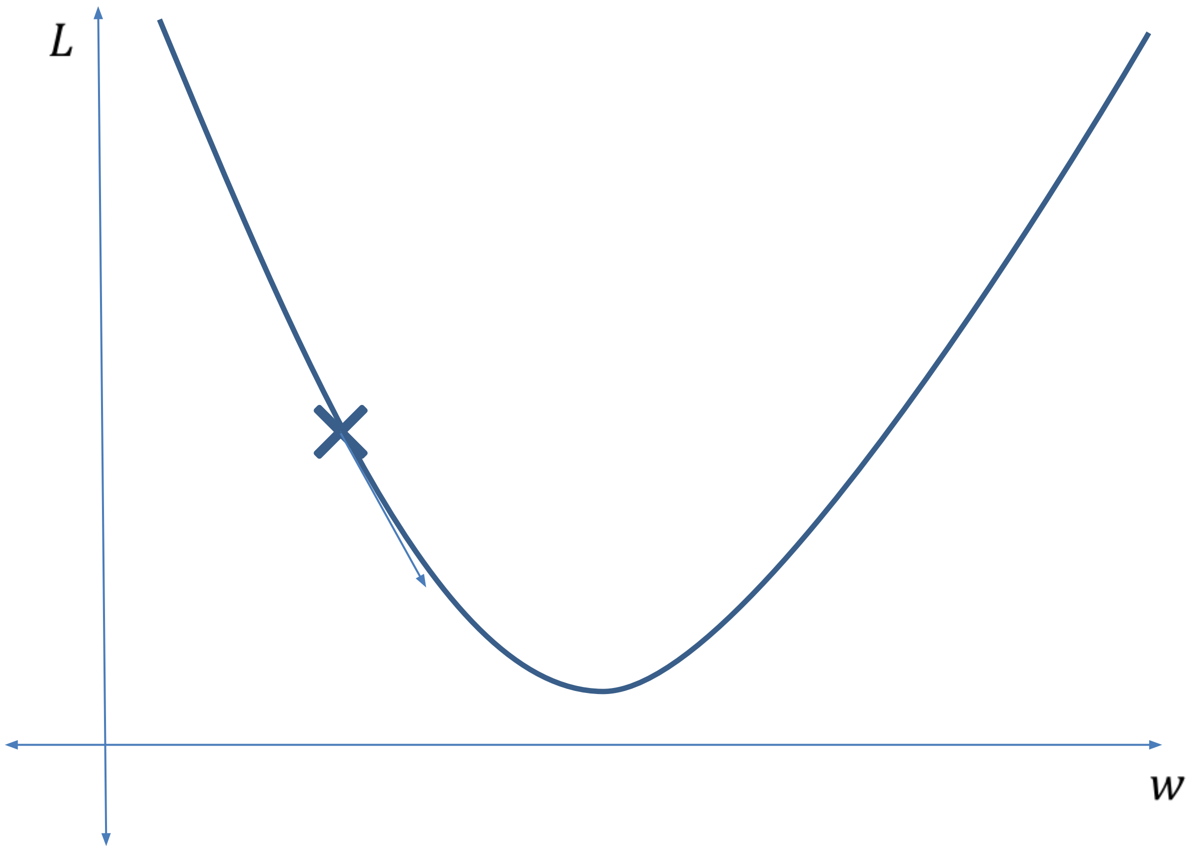
Gradient descent visualized: Minimizing loss



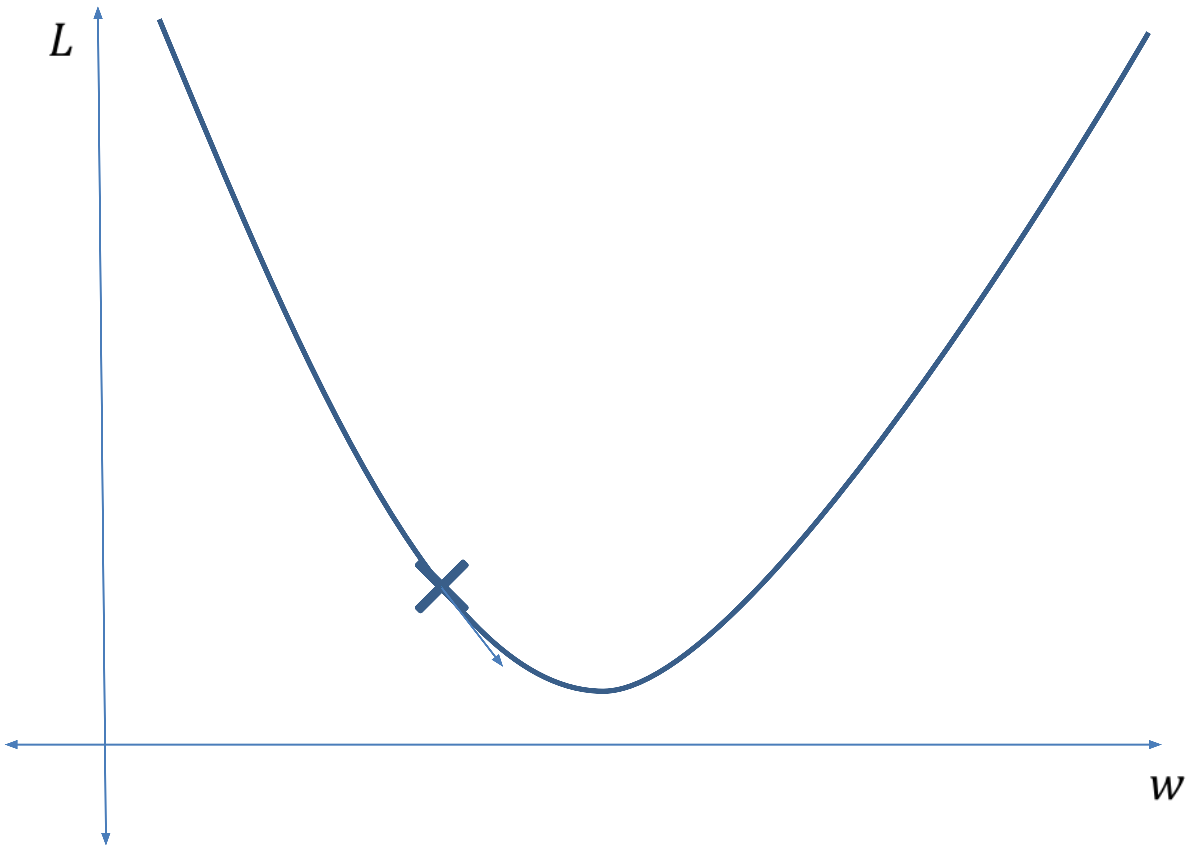
Gradient descent visualized: Minimizing loss



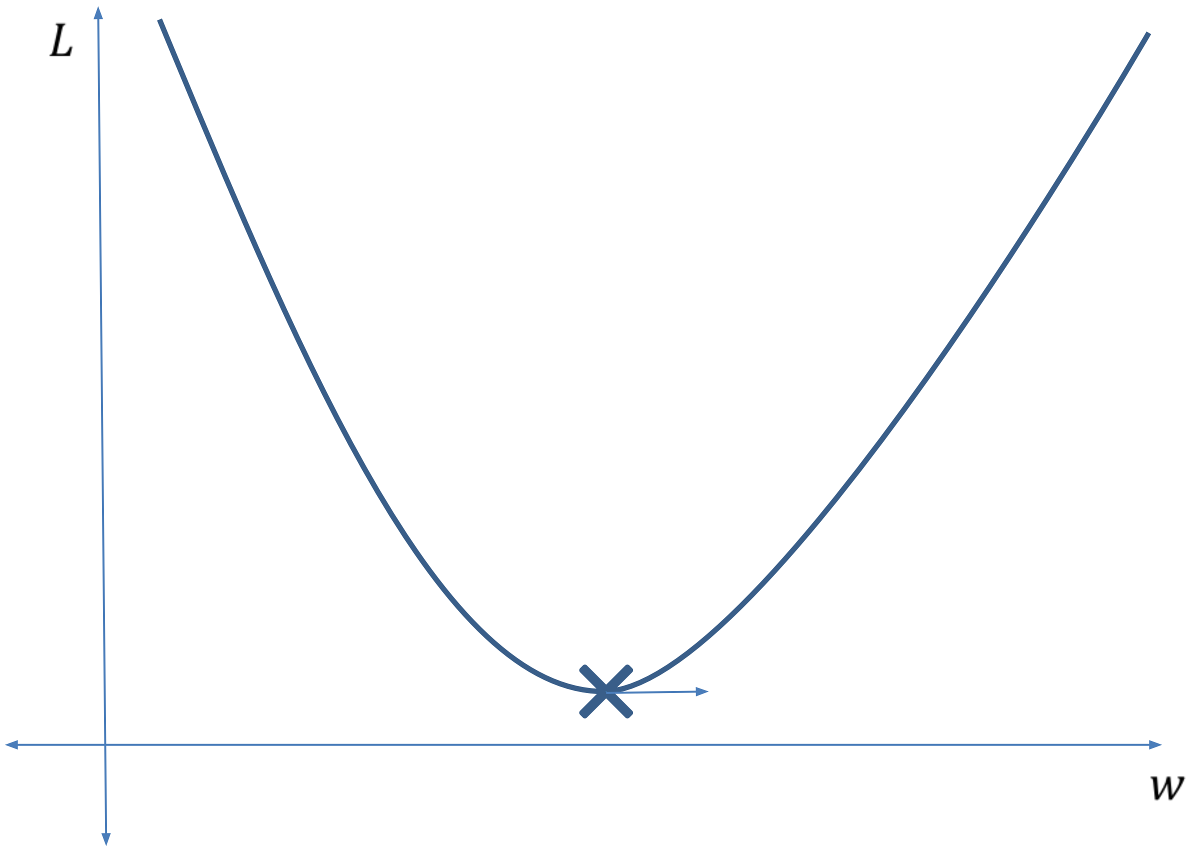
Gradient descent visualized: Minimizing loss



Gradient descent visualized: Minimizing loss



Gradient descent visualized: Minimizing loss



Gradient Descent Pseudocode

for _ in {0,…,num_epochs}:

$L = 0$

for x_i, y_i in data:

$\hat{y}_i = f(x_i, W)$

$L \mathrel{+}= L_i(y_i, \hat{y}_i)$

$\frac{dL}{dW} = ???$

$W := W - \alpha \frac{dL}{dW}$

Partial derivative of loss to update weights

Given training data point (***x***, ***y***), the linear classifier formula is: $\hat{y} = Wx$

Let's assume that the correct label is class ***k***, implying ***y=k***

$$\text{Loss} = L(\hat{y}, y) = -\log \frac{e^{\hat{y}_k}}{\sum_j e^{\hat{y}_j}}$$

$$= -\hat{y}_k + \log \sum_j e^{\hat{y}_j}$$

Calculating the gradient is hard, but we can use the chain rule to make it simpler

$$\frac{dL}{dW} = \frac{dL}{d\hat{y}} \frac{d\hat{y}}{dW}$$

Partial derivative of loss to update weights

Given training data point (***x***, ***y***), the linear classifier formula is: $\hat{y} = Wx$

Let's assume that the correct label is class ***k***, implying ***y=k***

$$\text{Loss} = -\hat{y}_k + \log \sum_j e^{\hat{y}_j}$$

Now, we want to update the weights ***W*** by calculating the direction in which to change the weights to reduce the loss: $\frac{dL}{dW} = \frac{dL}{d\hat{y}} \frac{d\hat{y}}{dW}$

we know that $\frac{d\hat{y}}{dW} = x$, but what about $\frac{dL}{d\hat{y}}$?

Partial derivative of loss to update weights

$$L = -\hat{y}_k + \log \sum_j e^{\hat{y}_j}$$

To calculate $\frac{dL}{d\hat{y}}$, we need to consider two cases:

Case 1:

$$\frac{dL}{d\hat{y}_k} = -1 + \frac{e^{\hat{y}_k}}{\sum_j e^{\hat{y}_j}}$$

Case 2:

$$\frac{dL}{d\hat{y}_{l \neq k}} = \frac{e^{\hat{y}_l}}{\sum_j e^{\hat{y}_j}}$$

Partial derivative of loss to update weights

Putting it all together:

$$\frac{dL}{dW} = \frac{dL}{d\hat{y}}\frac{d\hat{y}}{dW}$$

$$\frac{dL}{dW} =. \begin{bmatrix} \frac{e^{\hat{y}_0}}{\sum_j e^{\hat{y}_j}} \\ \cdots \\ \cdots \\ -1 + \frac{e^{\hat{y}_k}}{\sum_j e^{\hat{y}_j}} \\ \cdots \\ \cdots \\ \frac{e^{\hat{y}_{3071}}}{\sum_j e^{\hat{y}_j}} \end{bmatrix} \quad x$$

Gradient Descent Pseudocode

```
for _ in {0,…,num_epochs}:
```
$\quad L = 0$

$\quad$ `for` x_i`,` y_i `in data:`

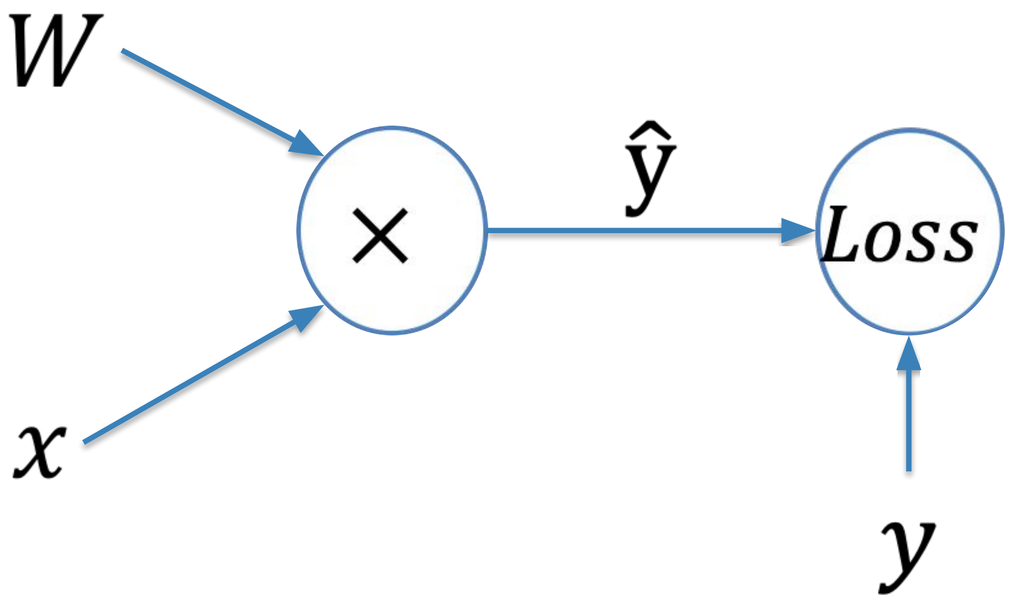
$\quad\quad \hat{y}_i = f(x_i, W)$

$\quad\quad L \mathrel{+}= L_i(y_i, \hat{y}_i)$

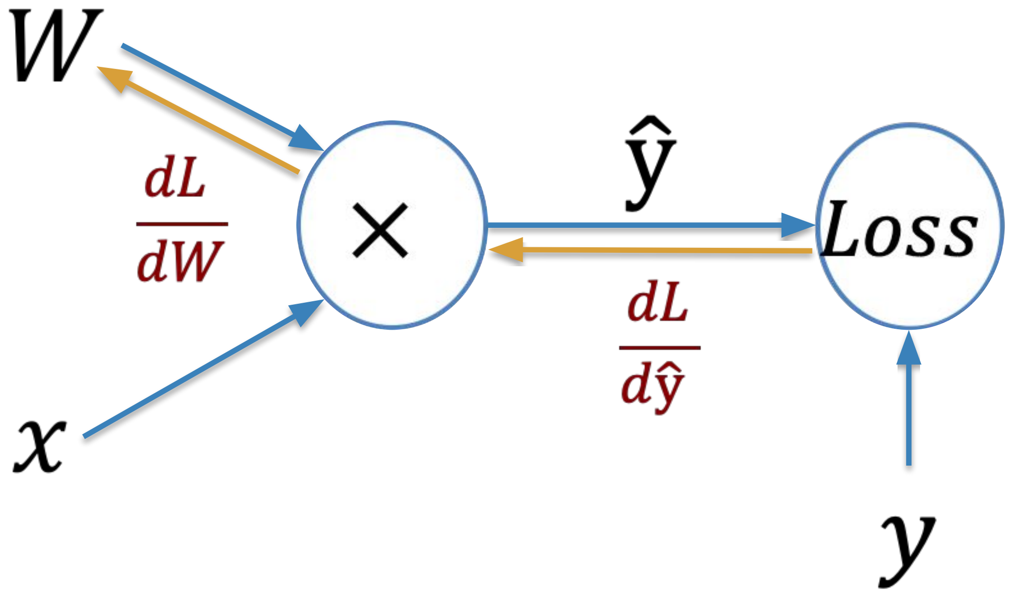
$\quad \dfrac{dL}{dW} = \textit{We know how to calculate this now!}$

$\quad W := W - \alpha \dfrac{dL}{dW}$

Backprop – another way of computing gradients



Backprop – another way of computing gradients



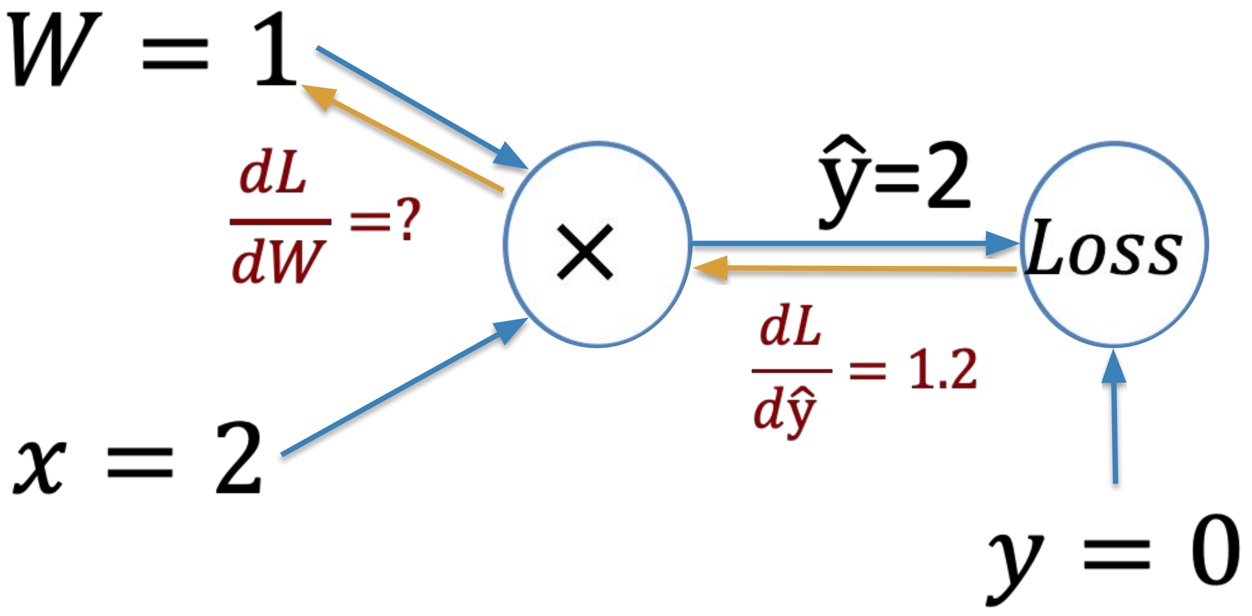
$$\hat{y} = Wx$$
$$L = Loss(\hat{y}, y)$$

$$\frac{dL}{dW} = \frac{dL}{d\hat{y}} \frac{d\hat{y}}{dW}$$

Key Insight:

- visualize the computation as a graph
- Compute the forward pass to calculate the loss.
- Compute all gradients for each computation backwards

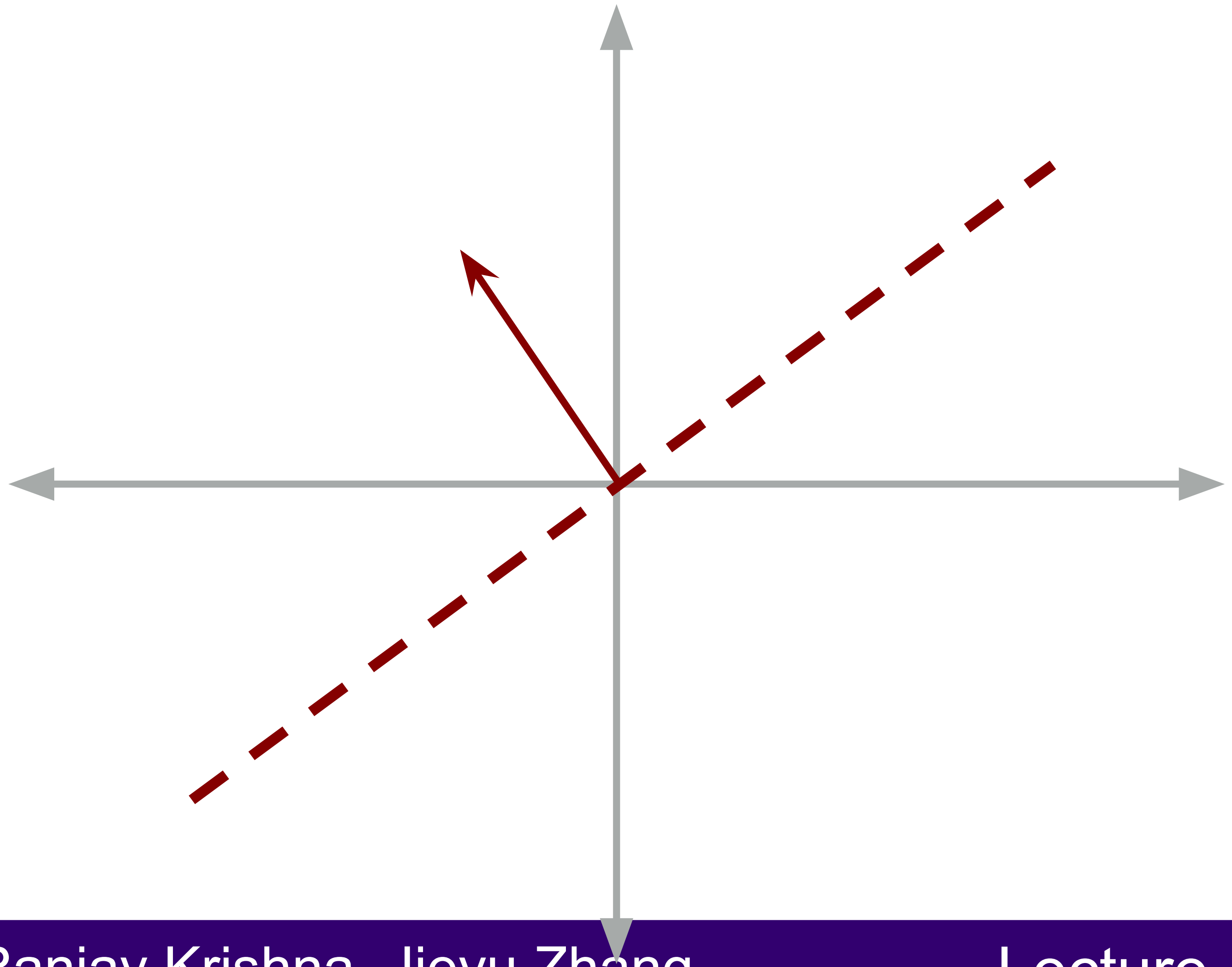
Backprop example in 1D:



We know the chain rule

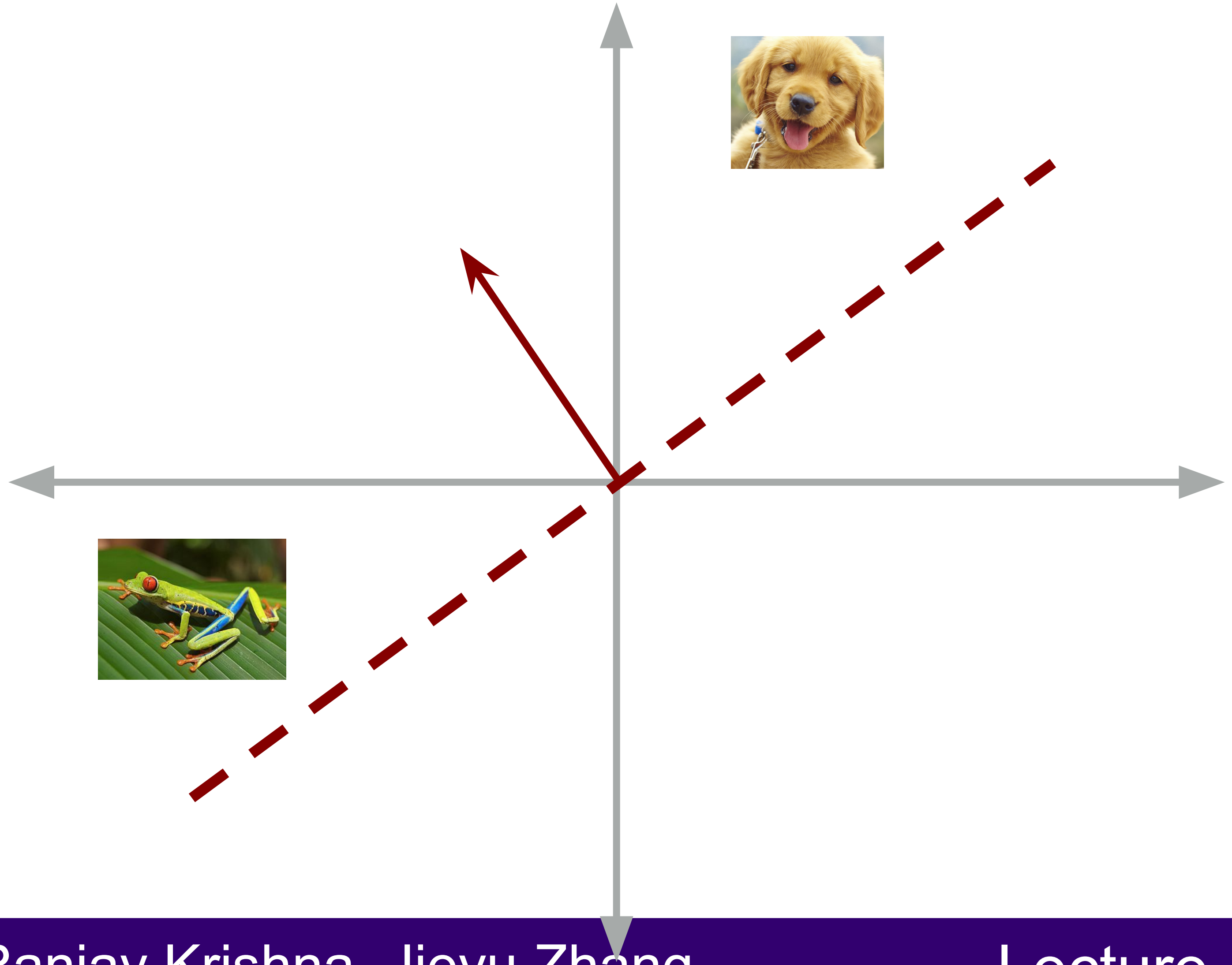
$$\frac{dL}{dW} = \frac{dL}{d\hat{y}}\frac{d\hat{y}}{dW}$$
$$= \frac{dL}{d\hat{y}}x$$
$$= 1.2x$$
$$= 1.2 \times 2$$
$$= 2.4$$

Interpreting the weights geometrically



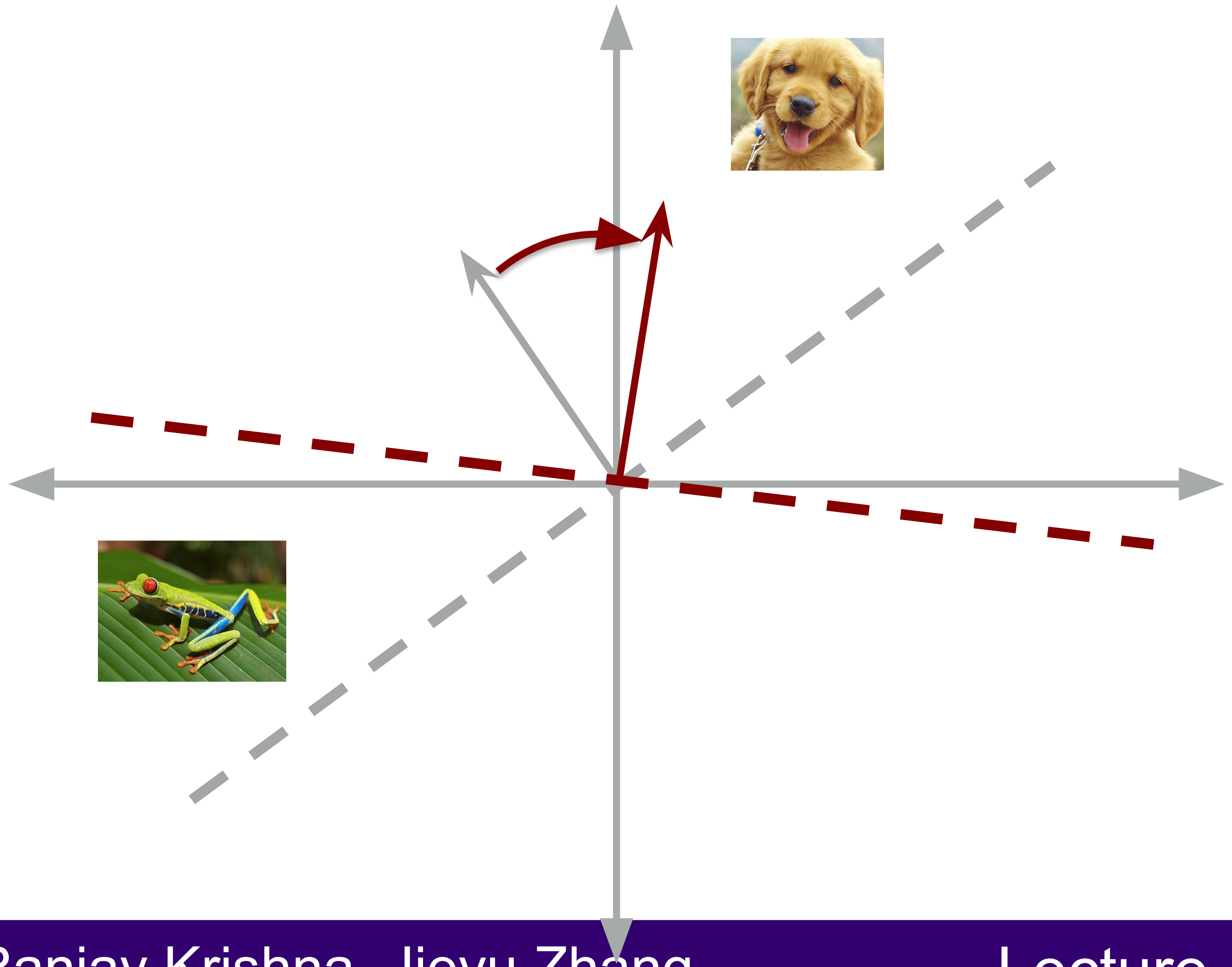
- Assume the image vectors are in 2D space to make it easier to visualize.
- Let's start with one class: dog.
- Initialize the weights randomly

Interpreting the weights geometrically



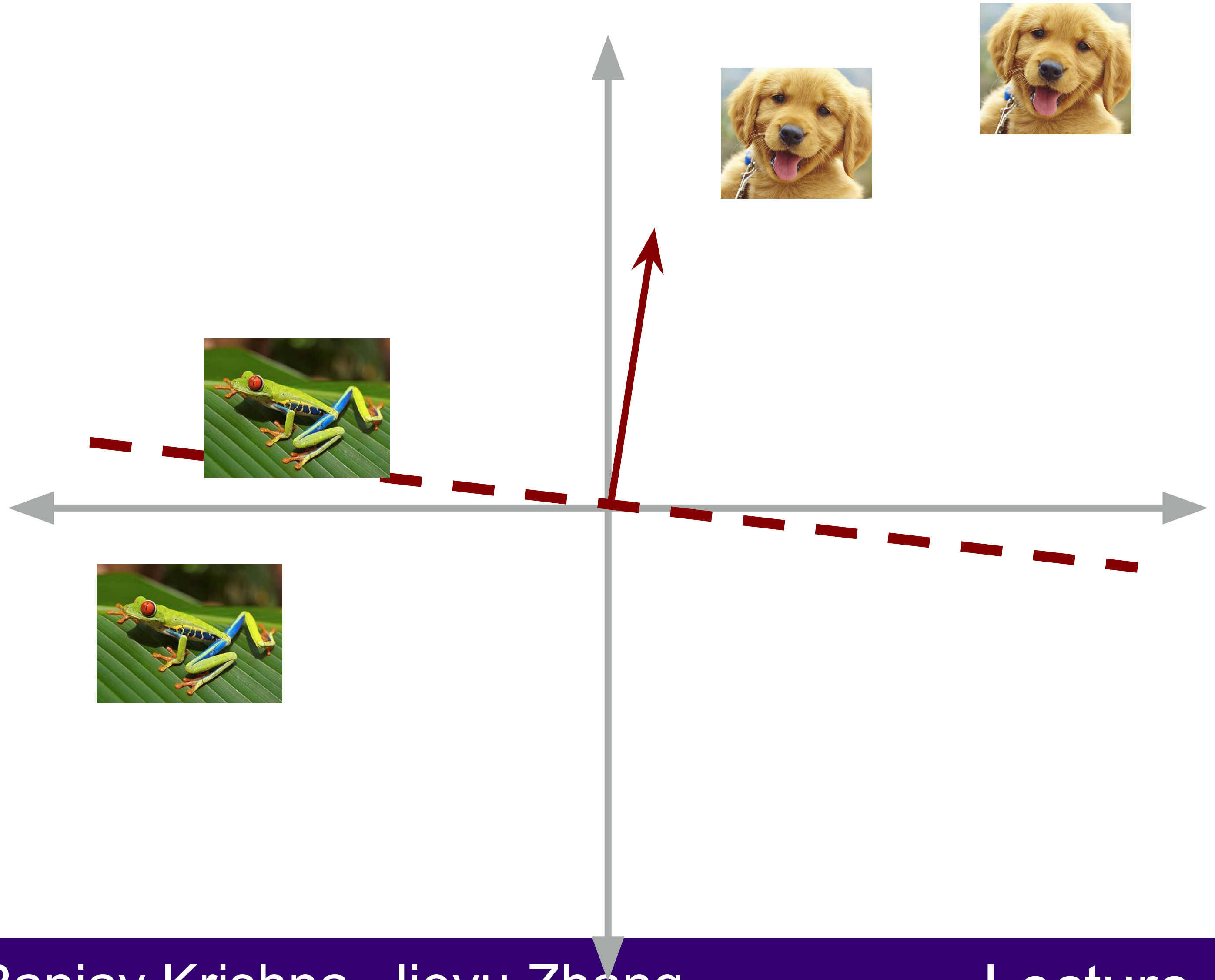
- Assume the image vectors are in 2D space to make it easier to visualize.
- Let's start with one class: dog.
- Initialize the weights randomly
- Now let's add two data points

Interpreting the weights geometrically



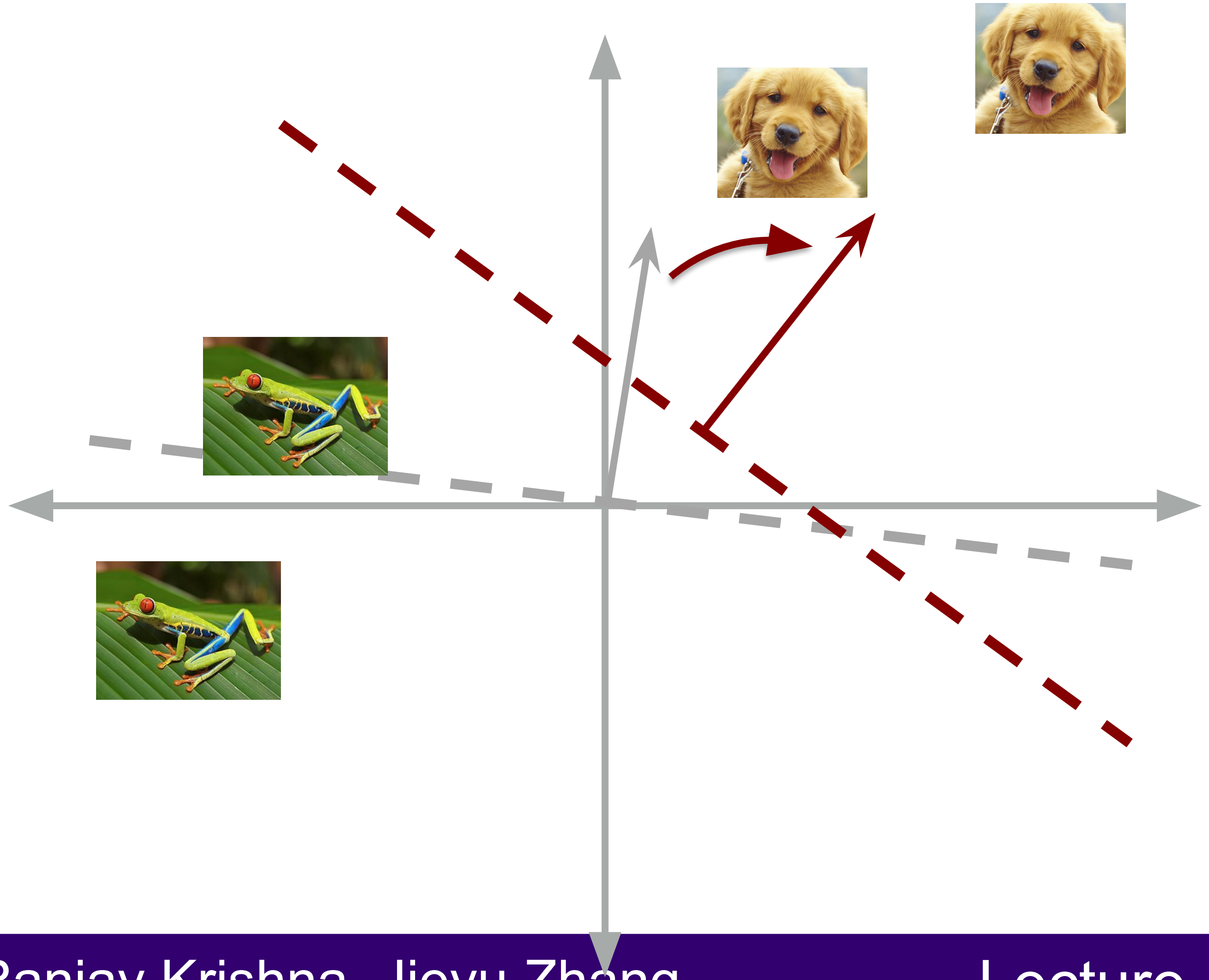
- Assume the image vectors are in 2D space to make it easier to visualize.
- Let's start with one class: dog.
- Initialize the weights randomly
- Now let's add two data points
- Update the weights

Interpreting the weights geometrically



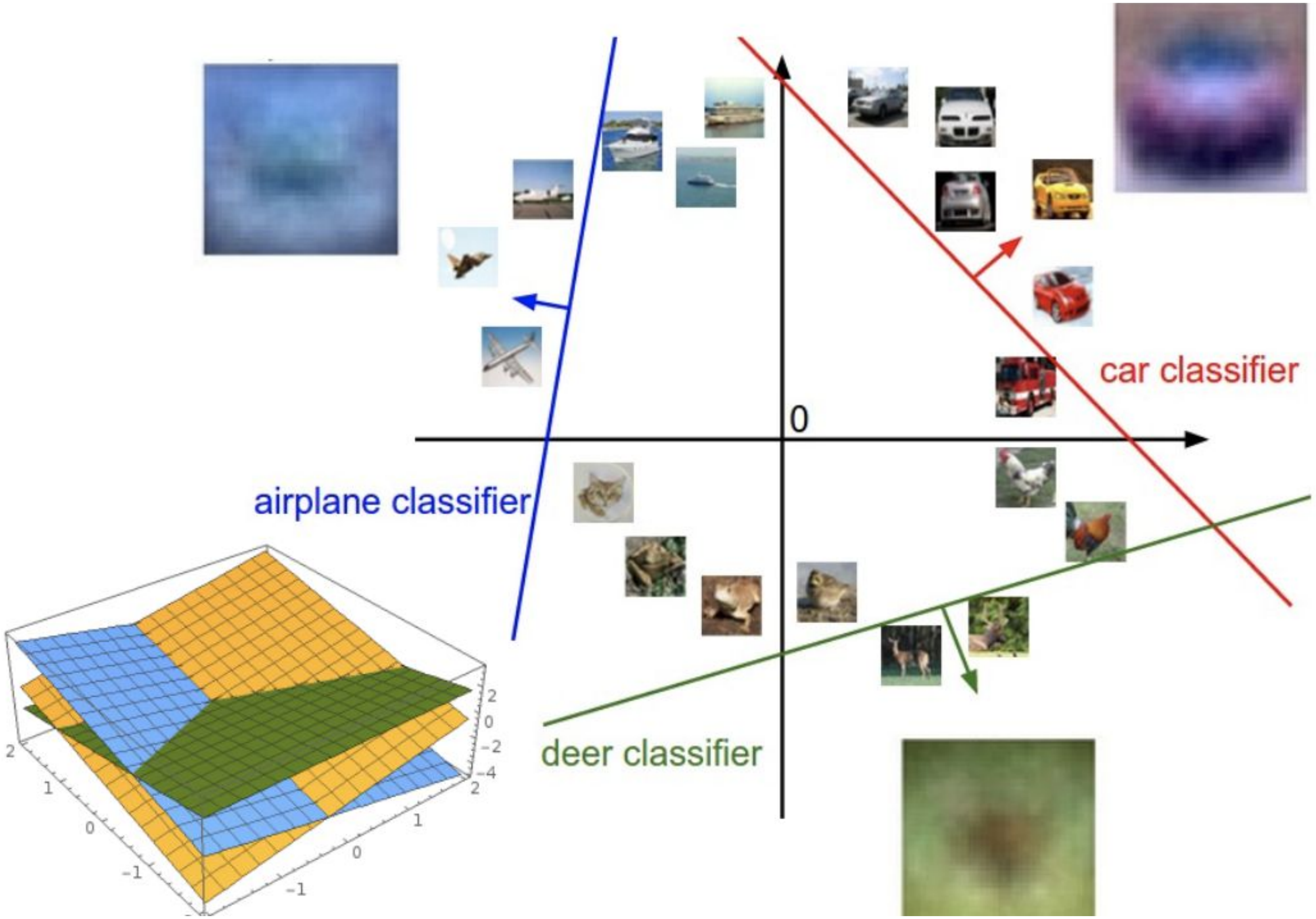
- Assume the image vectors are in 2D space to make it easier to visualize.
- Let's start with one class: dog.
- Initialize the weights randomly
- Now let's add two more data points
- Update the weights

Interpreting the weights geometrically



- Assume the image vectors are in 2D space to make it easier to visualize.
- Let's start with one class: dog.
- Initialize the weights randomly
- Now let's add two more data points
- Update the weights

Interpreting the weights geometrically

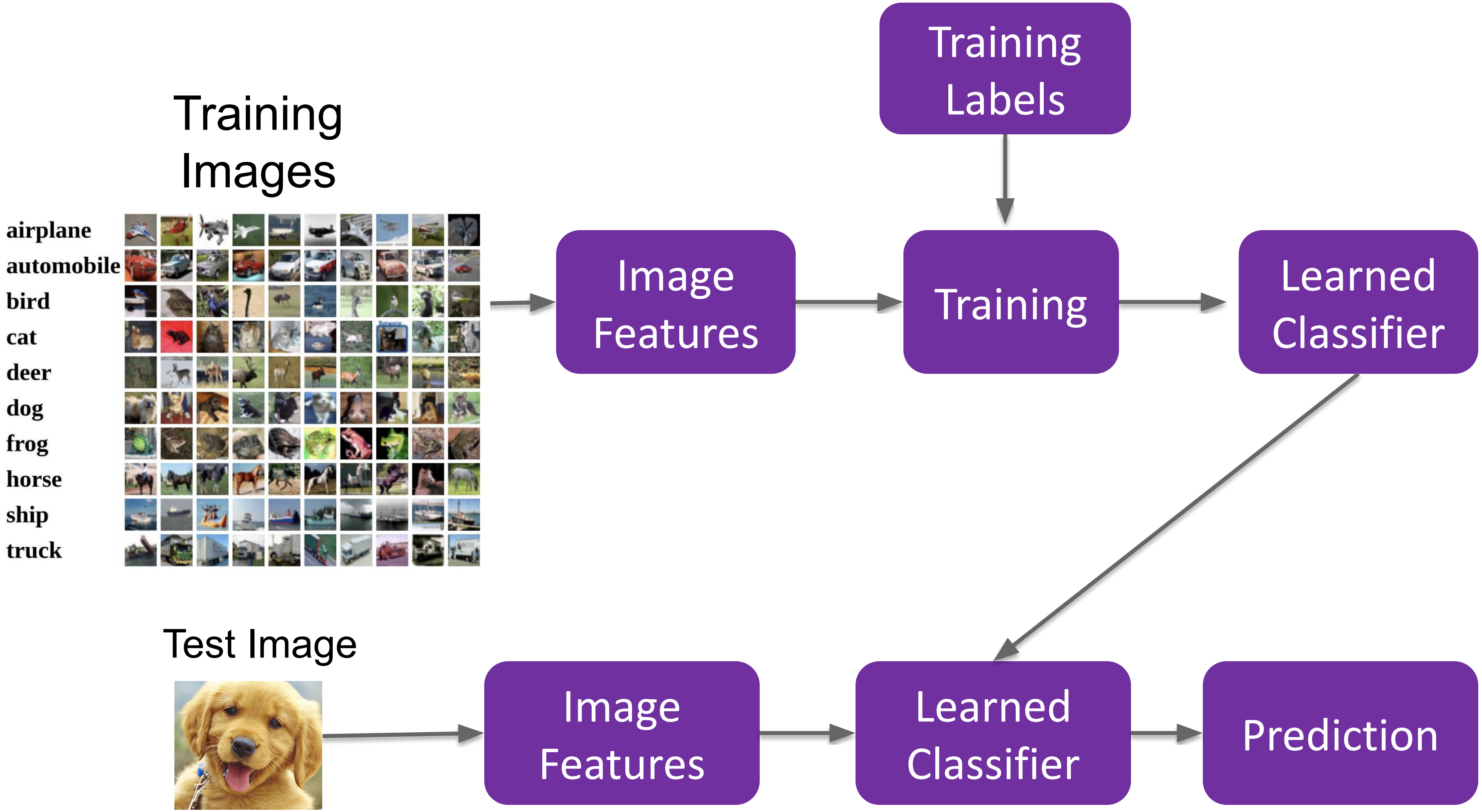


Plot created using [Wolfram Cloud]

Today's agenda

- Perceptron
- Linear classifier
- Loss function
- Gradient descent and backpropagation
- Neural networks

A simple recognition pipeline



Recall: we can featurize images into a vector

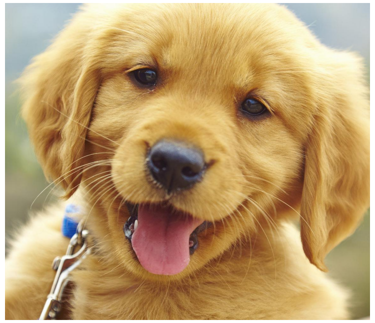
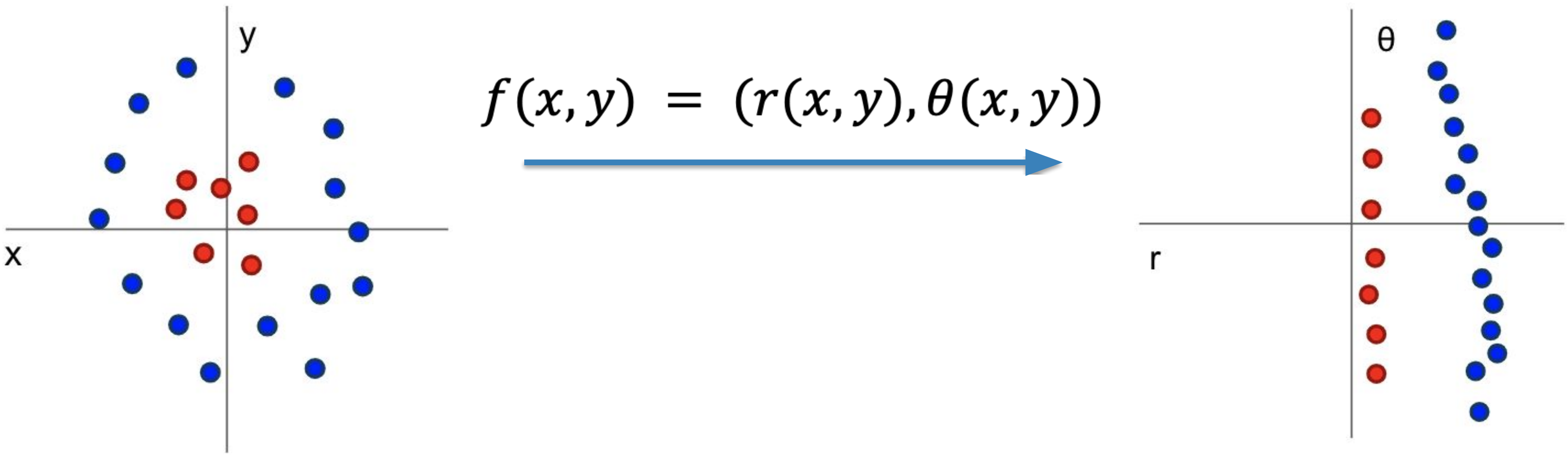


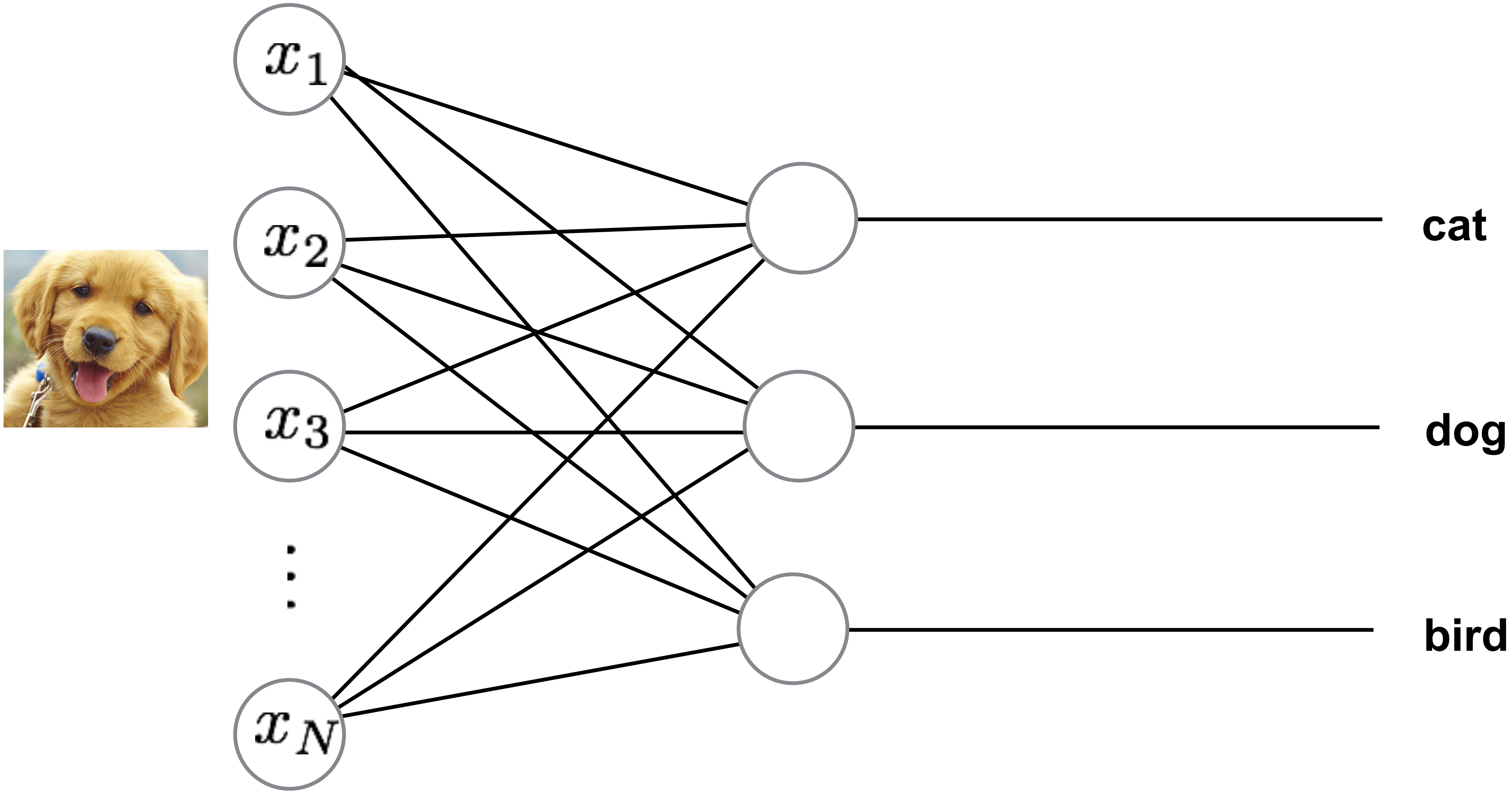
Image Vector

Raw pixels
Raw pixels + (x,y)
PCA
LDA
BoW
BoW + spatial pyramids

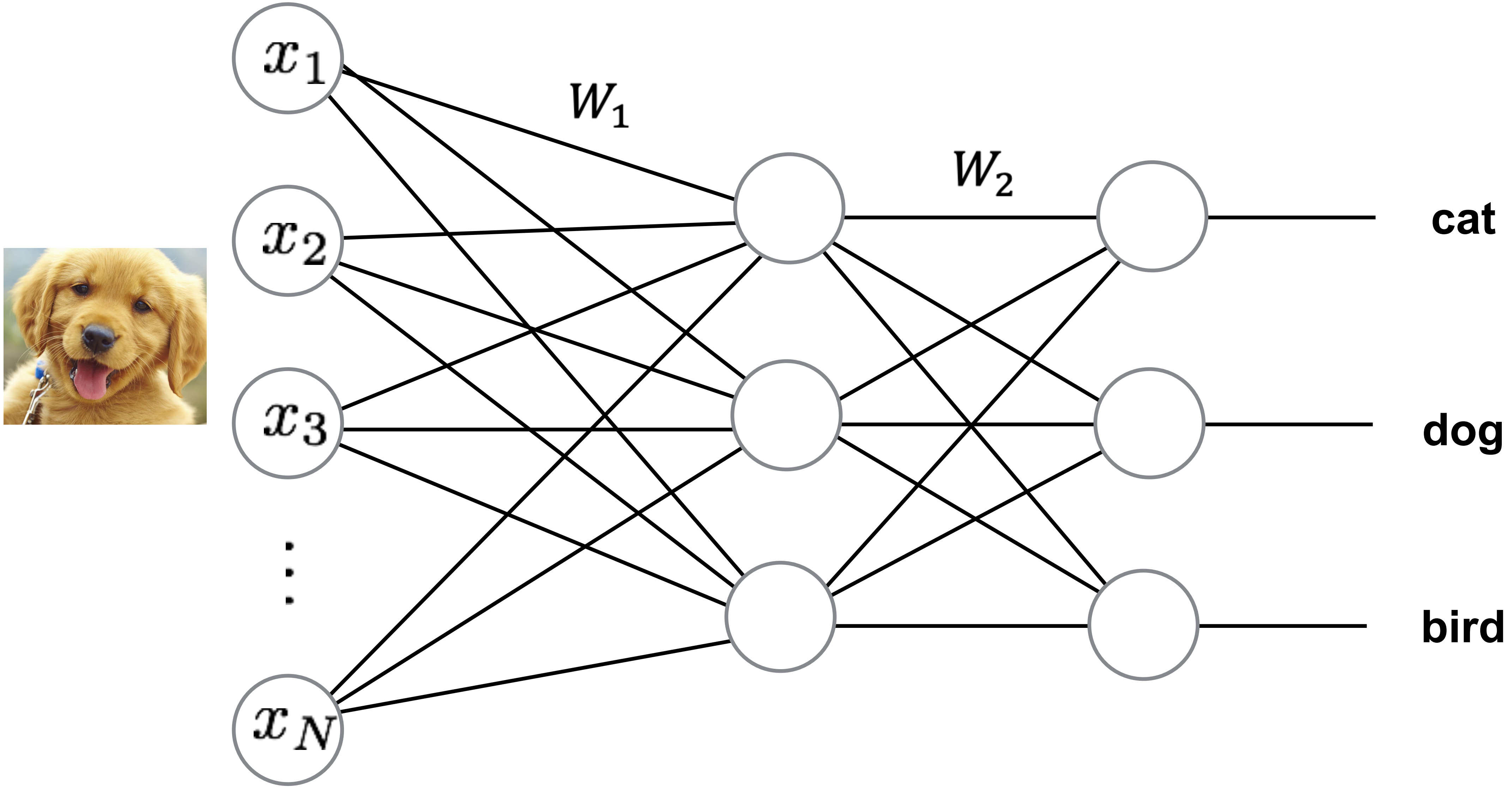
Features sometimes might not be linearly separable



Remember our linear classifier



Let's change the features by adding another layer



2-layer network: mathematical formula

- Linear classifier: $y = Wx$
- 2-layer network: $y = W_2 \max(0, W_1 x)$
- 3-layer network: $y = W_3 \max(0, W_2 \max(0, W_1 x))$

The number of layers is a new hyperparameter!

2-layer network: mathematical formula

- Linear classifier: $y = Wx$
- 2-layer network: $y = W_2 \max(0, W_1 x)$

We know the size of *x* = 1 x 3072 and y = 10 x 1, so what are ***W1*** and ***W2***

$$W_1 = h \times 3072 \qquad W_2 = 10 \times h$$

h is a new hyperparameter!

2-layer network: mathematical formula

- Linear classifier: $y = Wx$

- 2-layer network: $y = W_2 \max(0, W_1 x)$

Why is the max(0, _) necessary? Let's see what happen when we remove it:

$$y = W_2 W_1 x = \mathrm{W}x$$

Where: $\mathrm{W} = W_2 W_1$

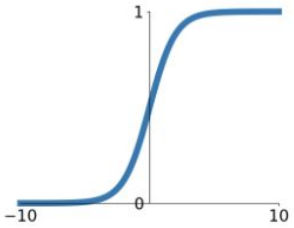
Activation function

The non-linear max function allows models to learn more complex transformations for features.

Choosing the right activation function is another new hyperparameter!

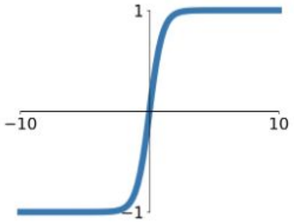
Sigmoid

$\sigma(x) = \frac{1}{1+e^{-x}}$



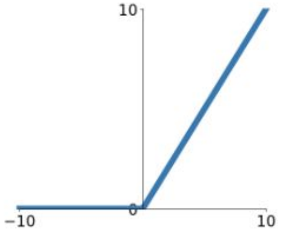
tanh

$\tanh(x)$



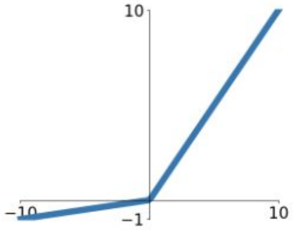
ReLU

$\max(0, x)$



Leaky ReLU

$\max(0.1x, x)$

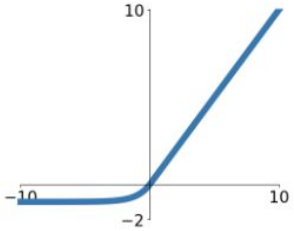


Maxout

$\max(w_1^T x + b_1, w_2^T x + b_2)$

ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



2-layer neural network performance

- ~40% accuracy on CIFAR-10 test
  - Best class: Truck (~60%)
  - Worst class: Horse (~16%)

- Check out the model at: **https://tinyurl.com/cifar10**

Next lecture

Deep learning for CV