Under the hood of linear classifier

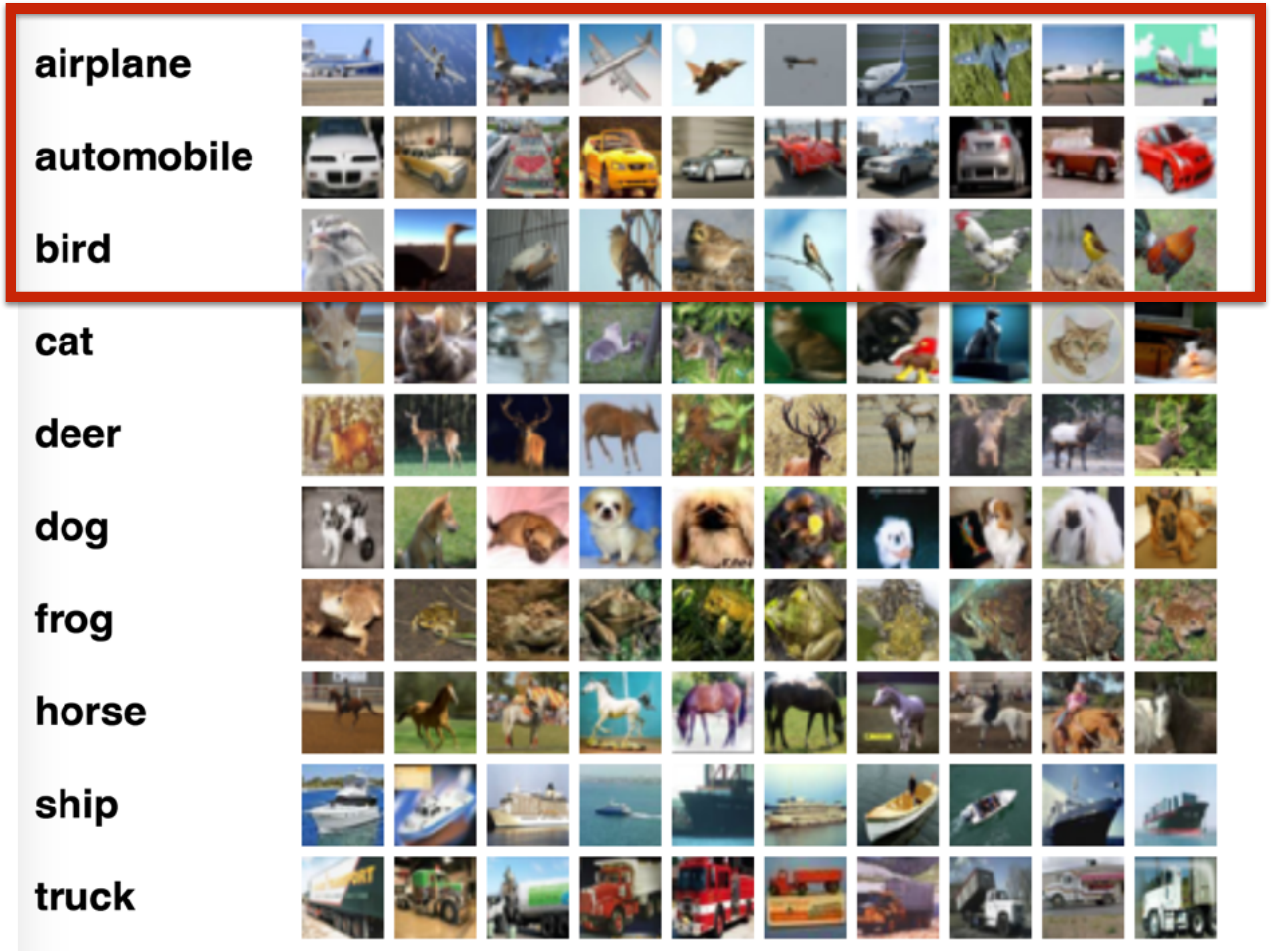
Linear classifier on RGB images

Pre-requisites:

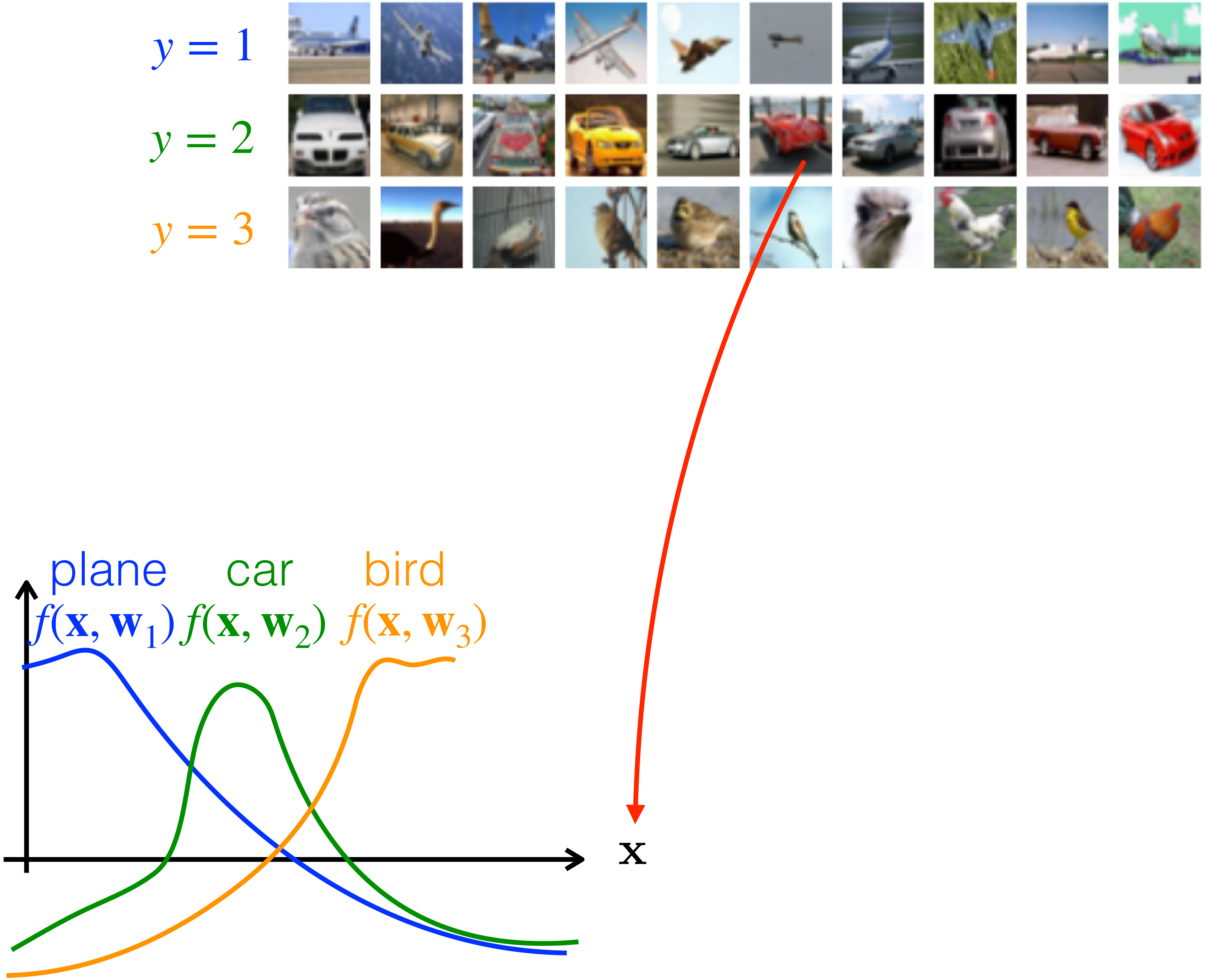
- linear algebra,

Karel Zimmermann
Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics

Recognition problem



CIFAR-10: classify 32x32 RGB images into 10 categories
https://www.cs.toronto.edu/~kriz/cifar.html

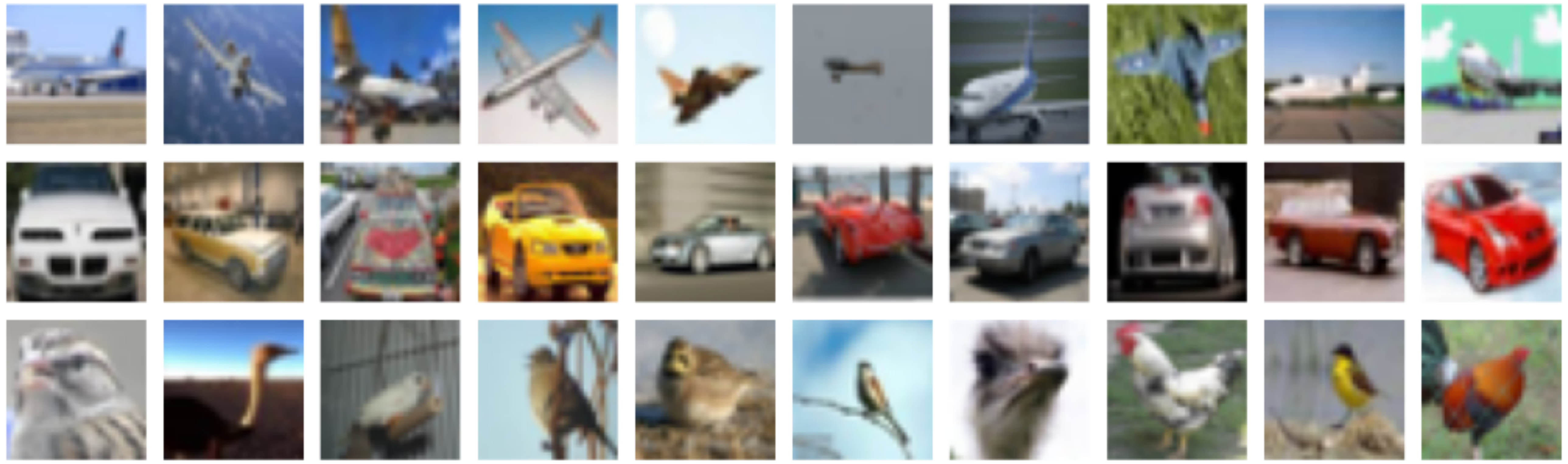


$y = 1$
$y = 2$
$y = 3$
plane
car
bird
$f(\mathbf{x}, \mathbf{w}_1)$
$f(\mathbf{x}, \mathbf{w}_2)$
$f(\mathbf{x}, \mathbf{w}_3)$
$\mathbf{x}$

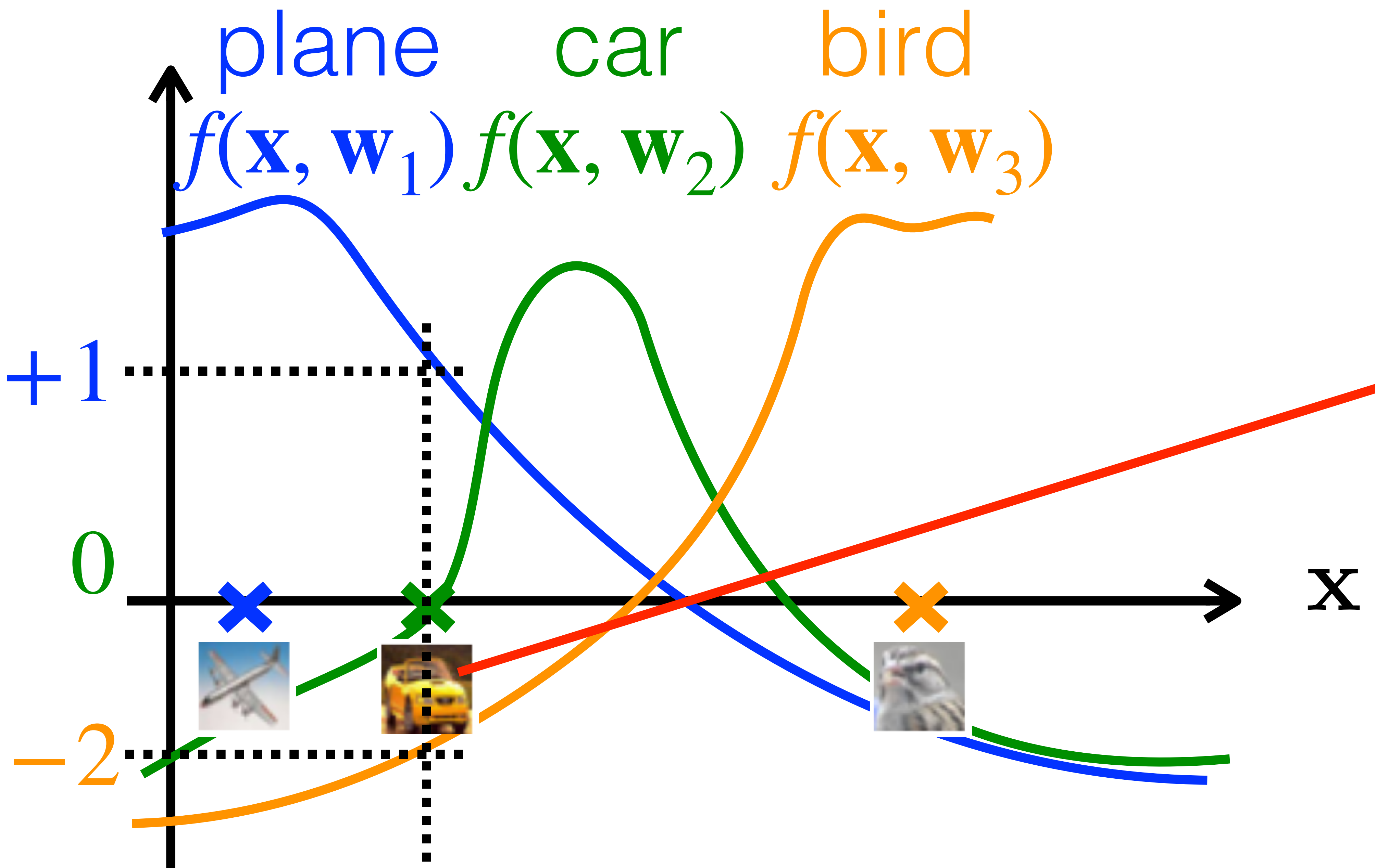
$y = 1$

$y = 2$

$y = 3$



$p(y \mid \mathbf{x}, \mathbf{W}) =$ **???**



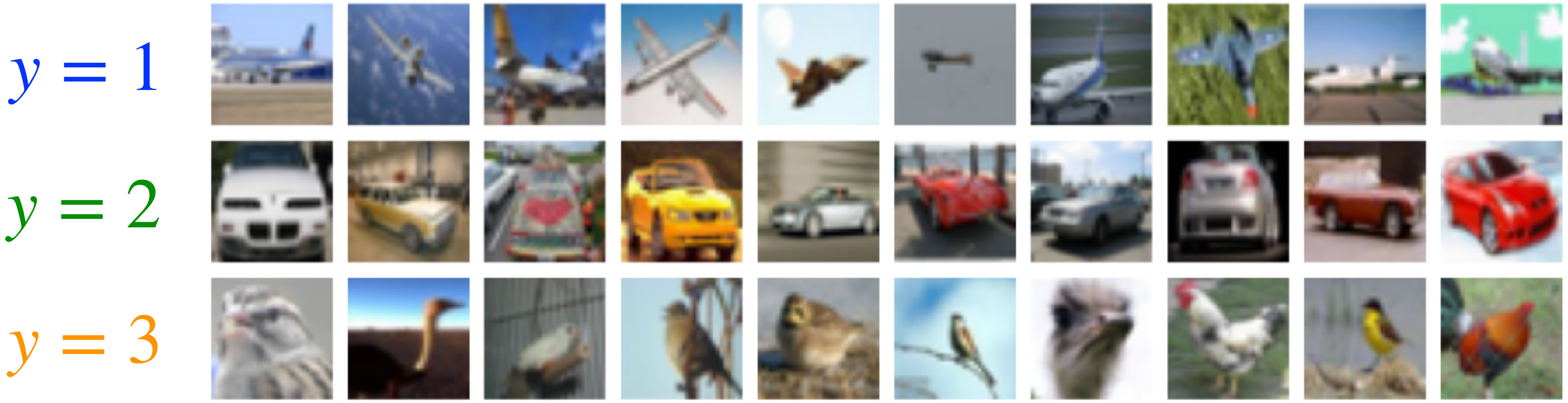
Linear classifier

$\mathbf{x} = \text{vec}($  $)$

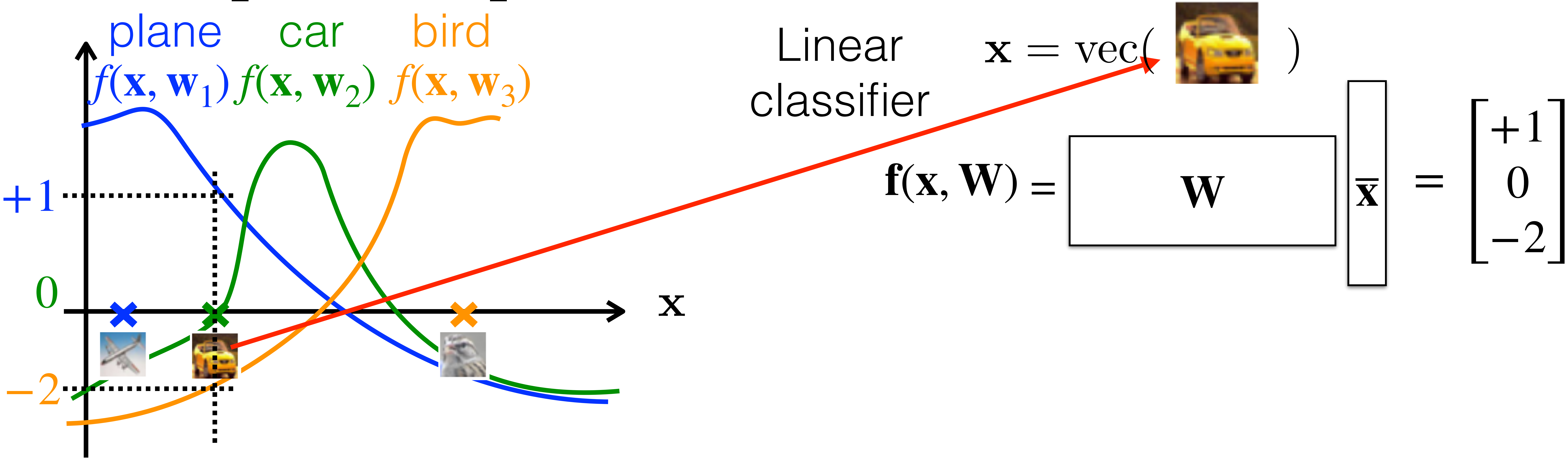
logits

$$\mathbf{f}(\mathbf{x}, \mathbf{W}) = \boxed{\mathbf{W}} \; \bar{\mathbf{x}} = \begin{bmatrix} +1 \\ 0 \\ -2 \end{bmatrix}$$

How does the linear classifier work?



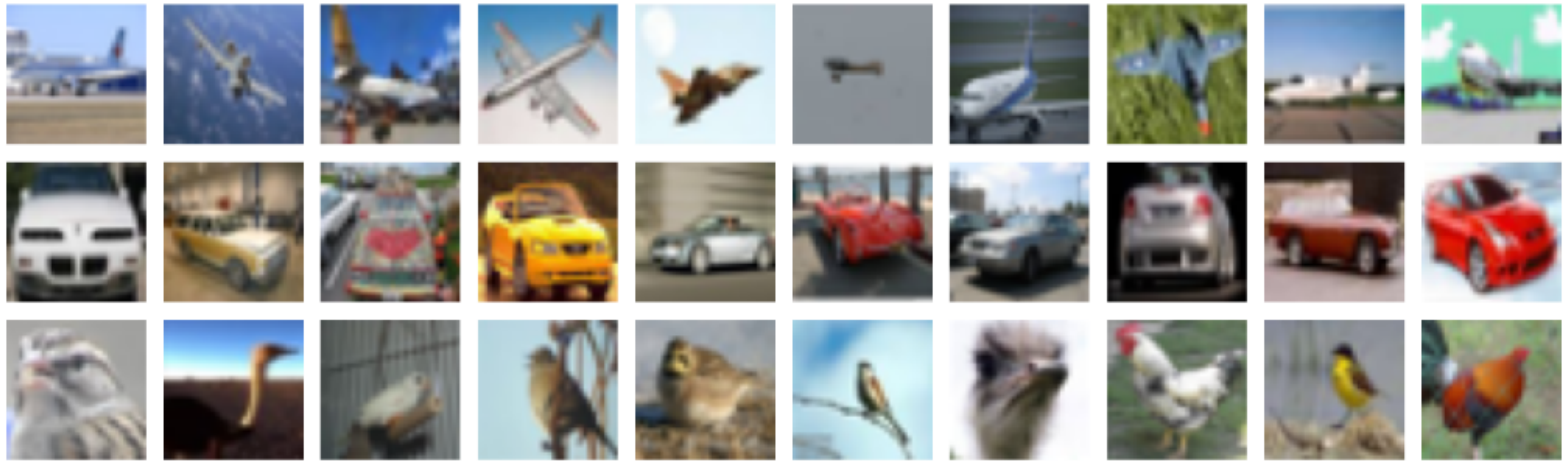
$$p(y \mid \mathbf{x}, \mathbf{W}) = \begin{bmatrix} \exp(f(\mathbf{x}, \mathbf{w}_1)) \\ \exp(f(\mathbf{x}, \mathbf{w}_2)) \\ \exp(f(\mathbf{x}, \mathbf{w}_3)) \end{bmatrix} \; / \; \sum_k \exp(f(\mathbf{x}, \mathbf{w}_k)) = \mathbf{s}(\mathbf{f}(\mathbf{x}, \mathbf{W})) = \mathbf{s}\left(\begin{bmatrix} +1 \\ 0 \\ -2 \end{bmatrix} \right) = \begin{bmatrix} 0.71 \\ 0.26 \\ 0.03 \end{bmatrix}$$



$y = 1$

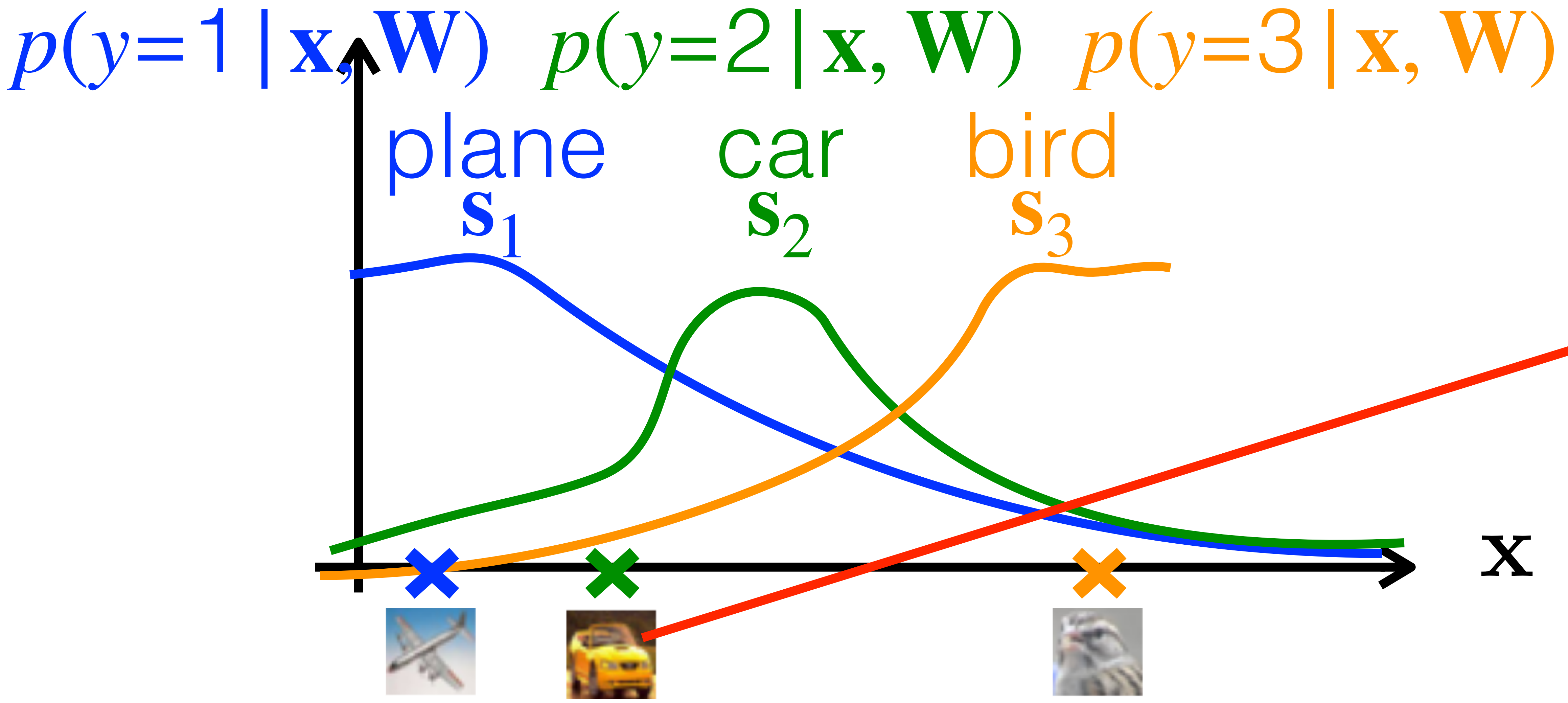
$y = 2$

$y = 3$



Model probability distribution over classes by softmax function

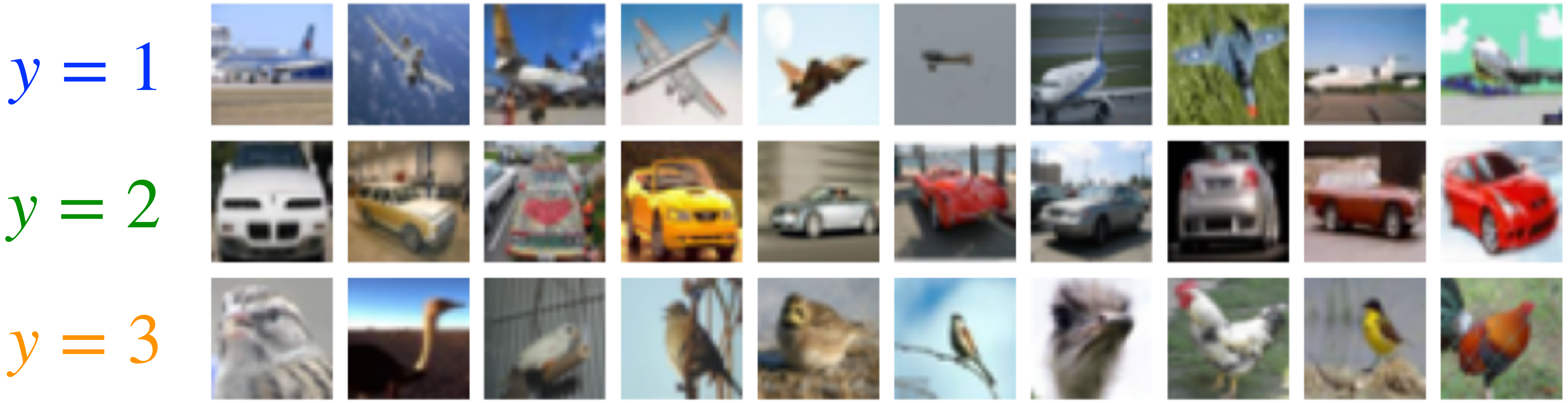
$$\mathbf{p}(y\,|\,\mathbf{x}, \mathbf{W}) = \begin{bmatrix} \exp(f(\mathbf{x}, \mathbf{w}_1)) \\ \exp(f(\mathbf{x}, \mathbf{w}_2)) \\ \exp(f(\mathbf{x}, \mathbf{w}_3)) \end{bmatrix} \Big/ \sum_k \exp\big(f\big(\mathbf{x}, \mathbf{w}_k\big)\big) = \mathbf{s}\big(\mathbf{f}\big(\mathbf{x}, \mathbf{W}\big)\big) = \mathbf{s}\left(\begin{bmatrix} +1 \\ 0 \\ -2 \end{bmatrix}\right) = \begin{bmatrix} 0.71 \\ 0.26 \\ 0.03 \end{bmatrix}$$



Linear classifier

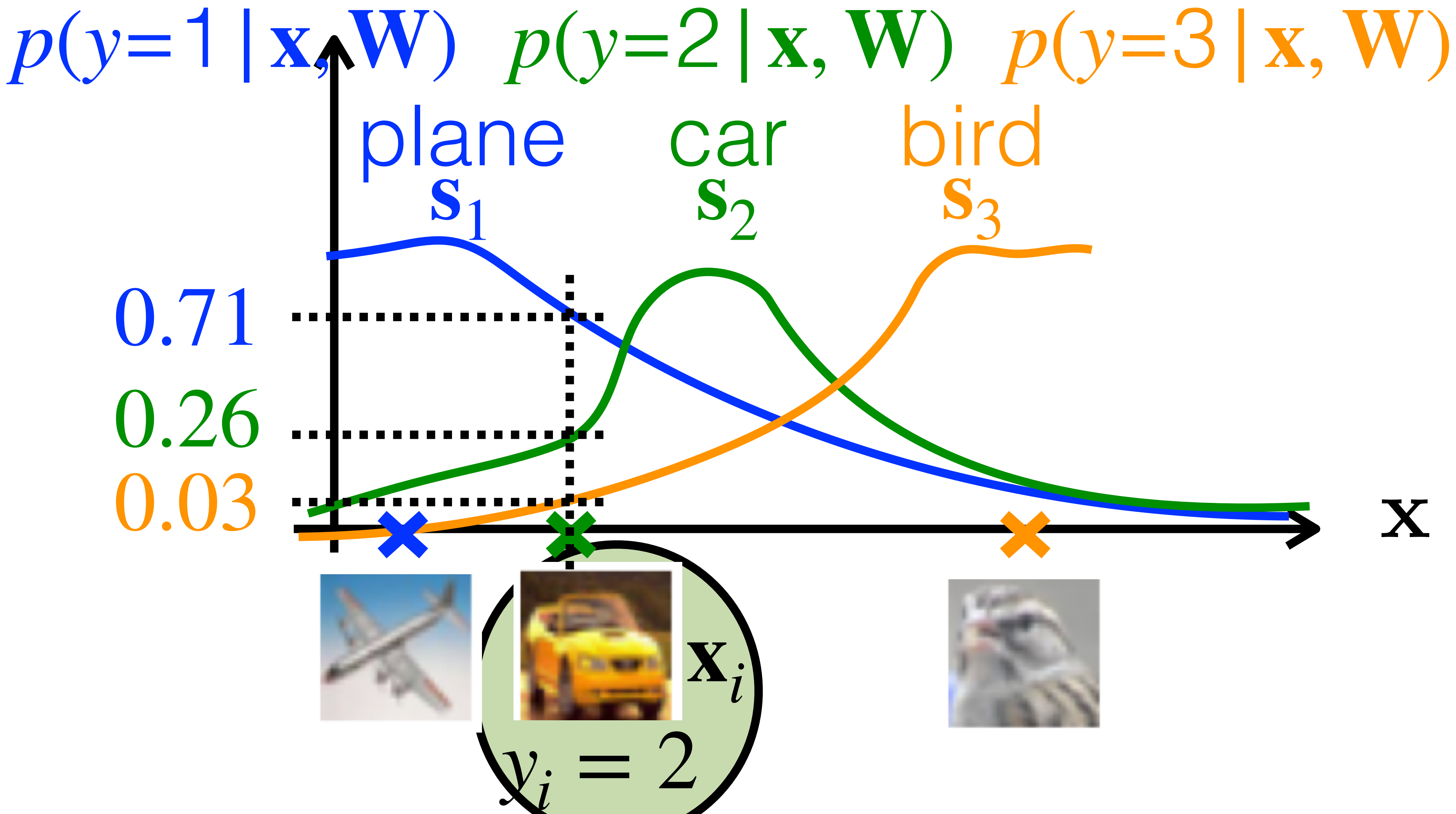
$\mathbf{x} = \text{vec}($  $)$

$$\mathbf{f}(\mathbf{x}, \mathbf{W}) = \mathbf{W}\,\bar{\mathbf{x}} = \begin{bmatrix} +1 \\ 0 \\ -2 \end{bmatrix}$$

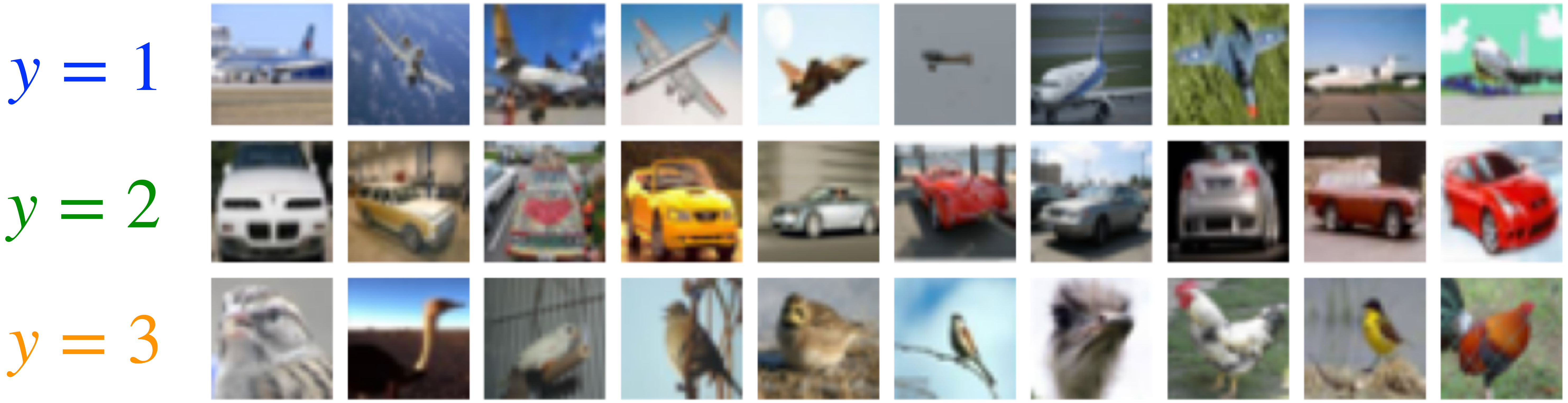


Model probability distribution over classes by softmax function

$$\mathbf{p}(y\,|\,\mathbf{x},\mathbf{W}) = \begin{bmatrix} \exp(f(\mathbf{x},\mathbf{w}_1)) \\ \exp(f(\mathbf{x},\mathbf{w}_2)) \\ \exp(f(\mathbf{x},\mathbf{w}_3)) \end{bmatrix} \Big/ \sum_k \exp\big(f\big(\mathbf{x},\mathbf{w}_k\big)\big) = \mathbf{s}\big(\mathbf{f}\big(\mathbf{x},\mathbf{W}\big)\big) = \mathbf{s}\left(\begin{bmatrix} +1 \\ 0 \\ -2 \end{bmatrix}\right) = \begin{bmatrix} 0.71 \\ 0.26 \\ 0.03 \end{bmatrix}$$

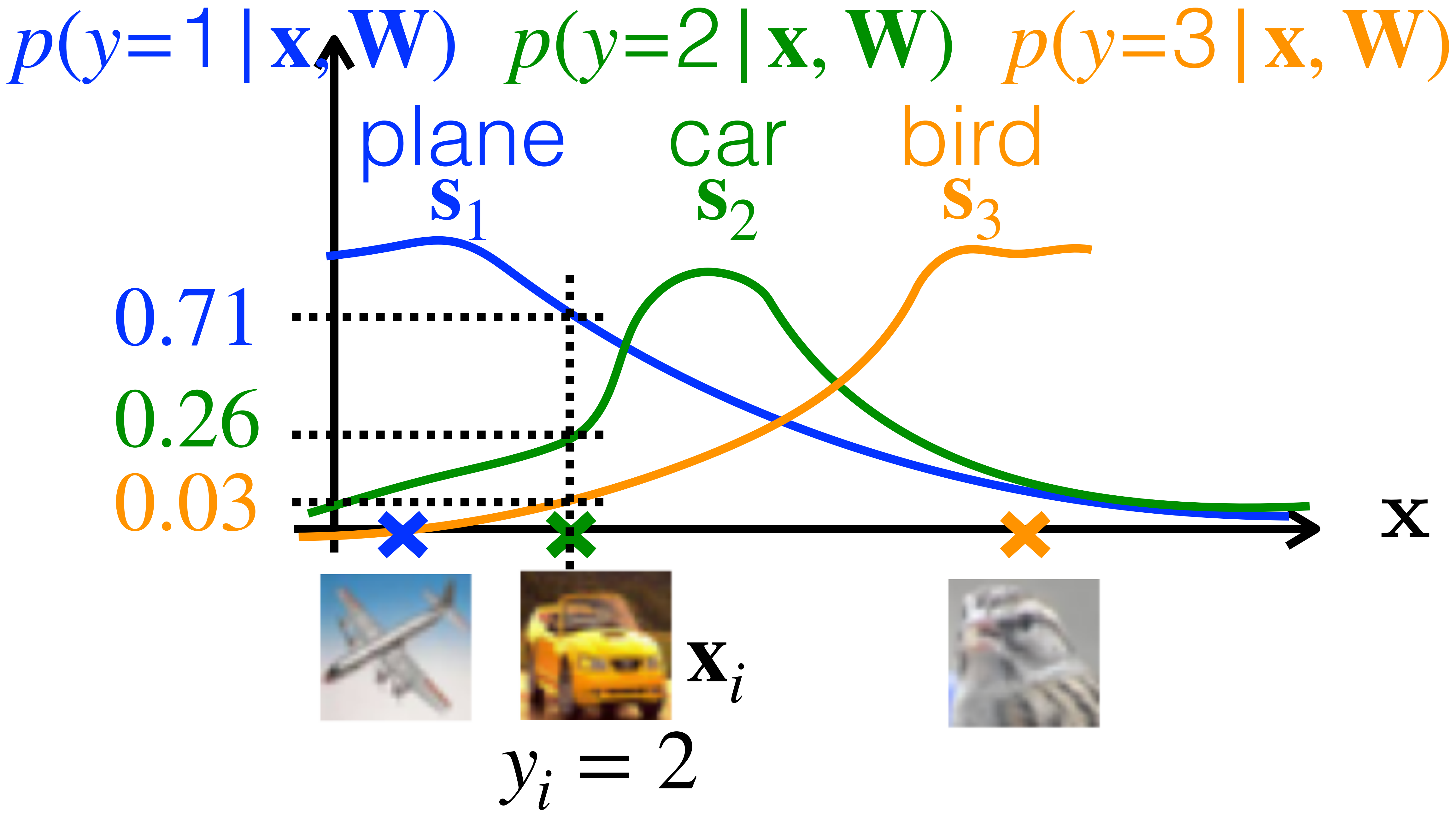


Which image is incorrectly classified?



Model probability distribution over classes by softmax function

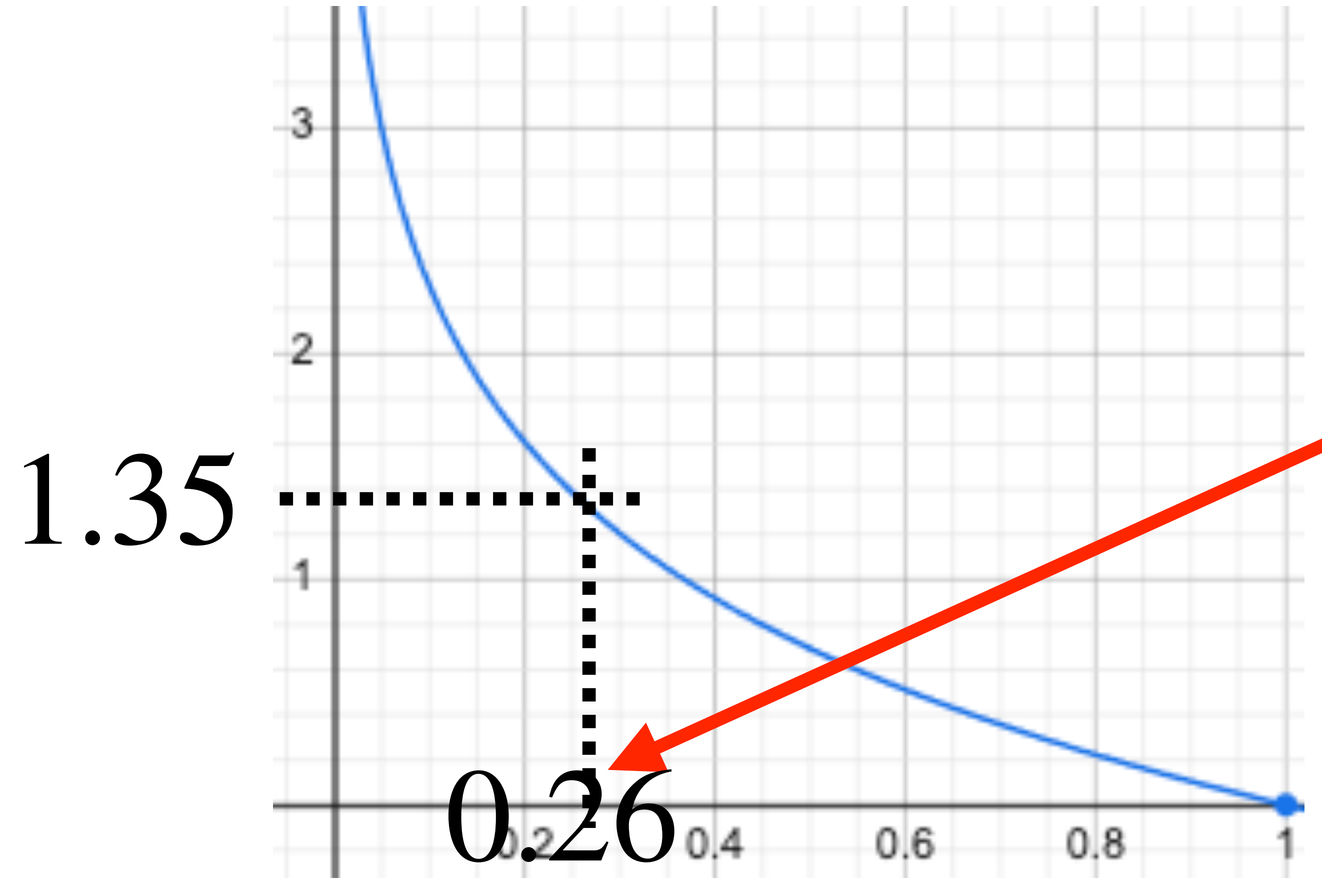
$$\mathbf{p}(y\,|\,\mathbf{x},\mathbf{W}) = \begin{bmatrix} \exp(f(\mathbf{x},\mathbf{w}_1)) \\ \exp(f(\mathbf{x},\mathbf{w}_2)) \\ \exp(f(\mathbf{x},\mathbf{w}_3)) \end{bmatrix} \;/\; \sum_k \exp(f(\mathbf{x},\mathbf{w}_k)) = \mathbf{s}(\mathbf{f}(\mathbf{x},\mathbf{W})) = \mathbf{s}\left(\begin{bmatrix} +1 \\ 0 \\ -2 \end{bmatrix}\right) = \begin{bmatrix} 0.71 \\ 0.26 \\ 0.03 \end{bmatrix}$$

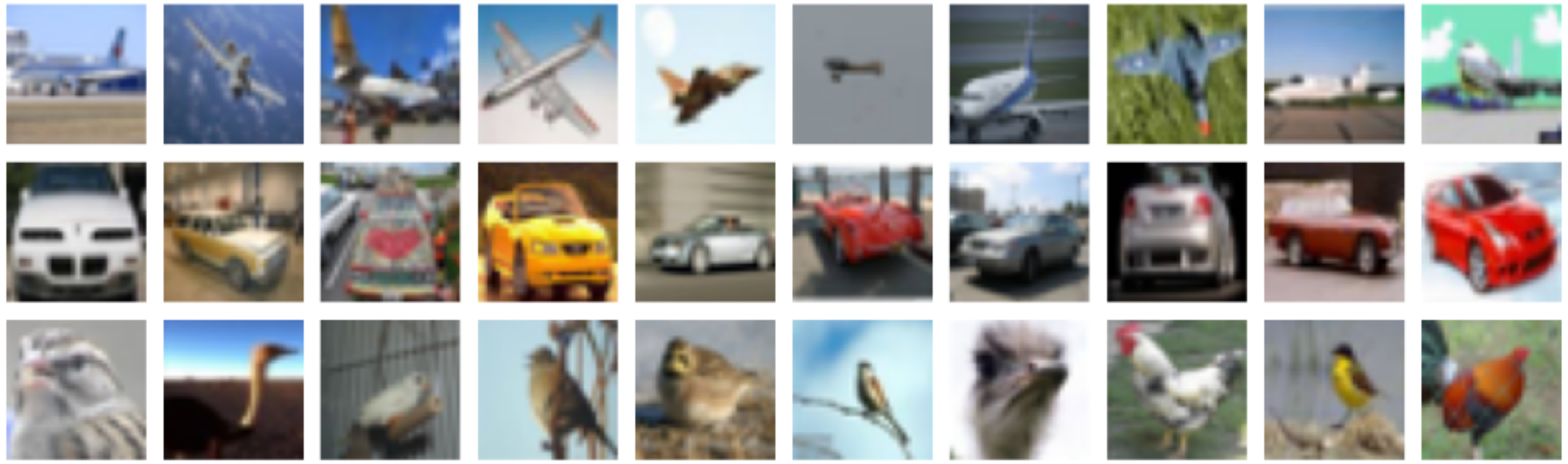


Loss function:

$$\mathcal{L}(\mathbf{W}) = -\log p(y = y_i\,|\,\mathbf{x}_i,\mathbf{W}) = -\log \mathbf{s}_{y_i}(\mathbf{W}\,\bar{\mathbf{x}}_i)$$

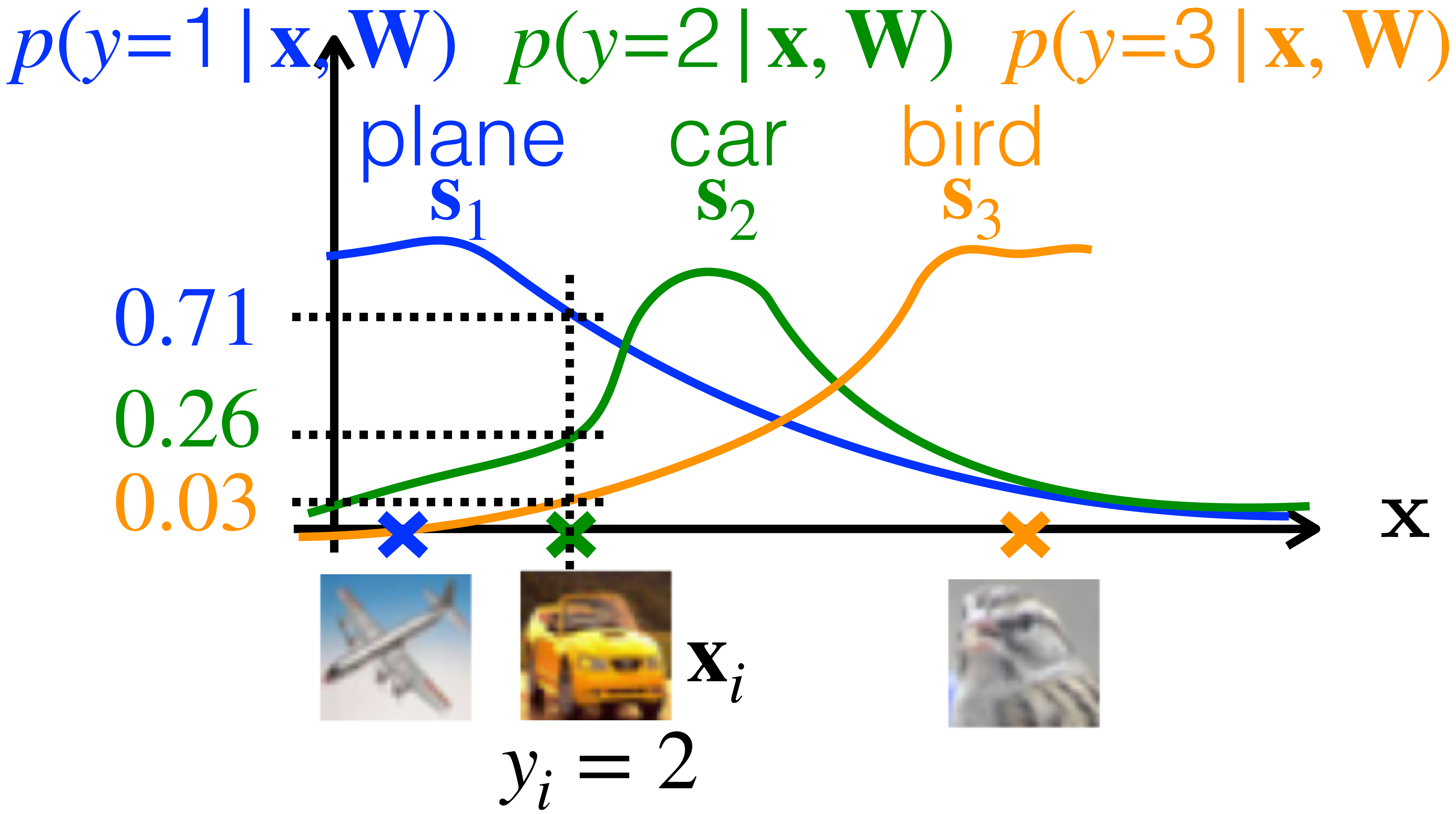
$$= -\log(0.26) = 1.35$$





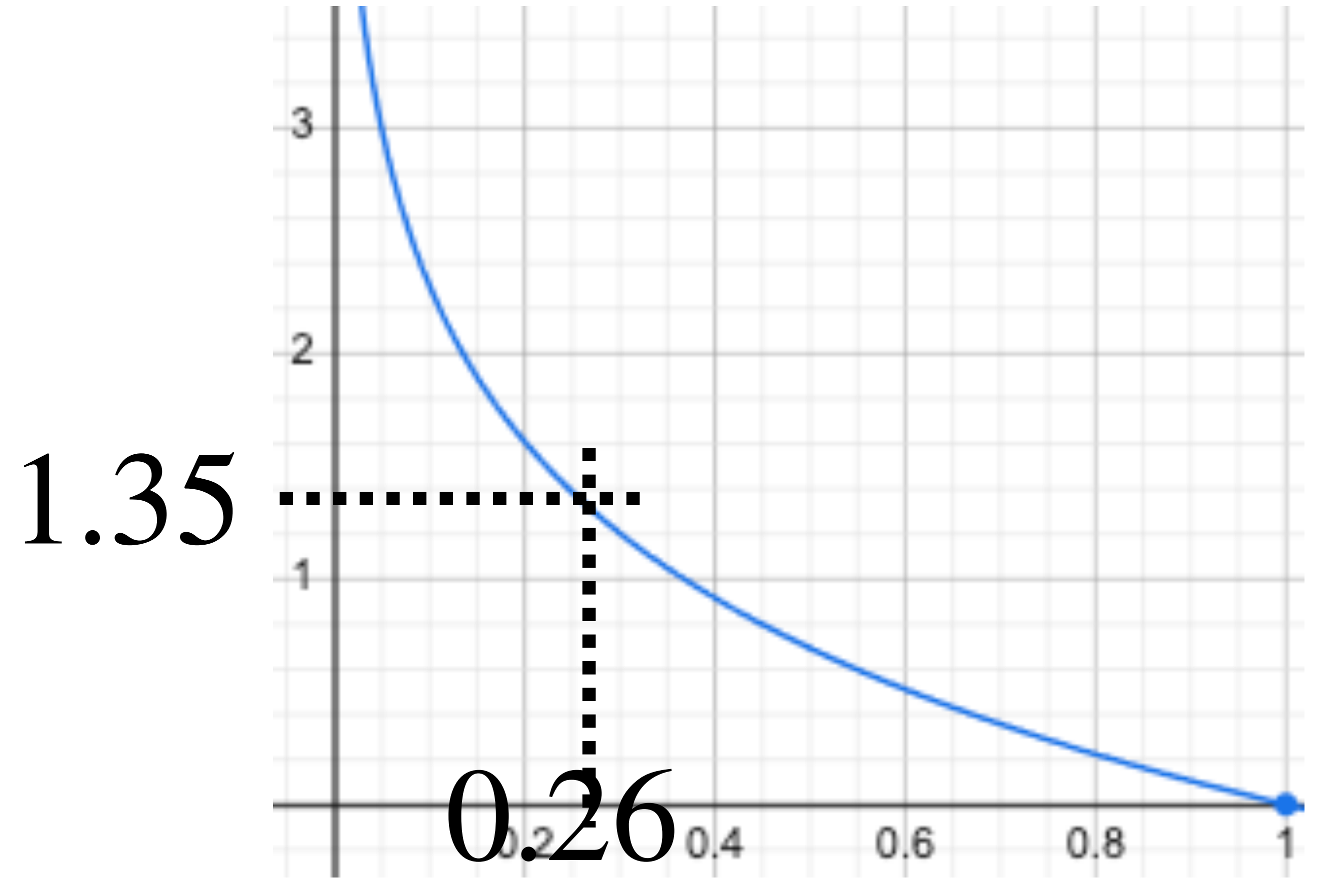
Model probability distribution over classes by softmax function

$$\mathbf{p}(y\,|\,\mathbf{x}, \mathbf{W}) = \begin{bmatrix} \exp(f(\mathbf{x}, \mathbf{w}_1)) \\ \exp(f(\mathbf{x}, \mathbf{w}_2)) \\ \exp(f(\mathbf{x}, \mathbf{w}_3)) \end{bmatrix} / \sum_k \exp(f(\mathbf{x}, \mathbf{w}_k)) = \mathbf{s}(\mathbf{f}(\mathbf{x}, \mathbf{W})) = \mathbf{s}\left(\begin{bmatrix} +1 \\ 0 \\ -2 \end{bmatrix} \right) = \begin{bmatrix} 0.71 \\ 0.26 \\ 0.03 \end{bmatrix}$$



Loss function:

$$\mathscr{L}(\mathbf{W}) = -\log p(y = y_i\,|\,\mathbf{x}_i, \mathbf{W}) = -\log \mathbf{s}_{y_i}(\mathbf{W}\,\bar{\mathbf{x}}_i)$$



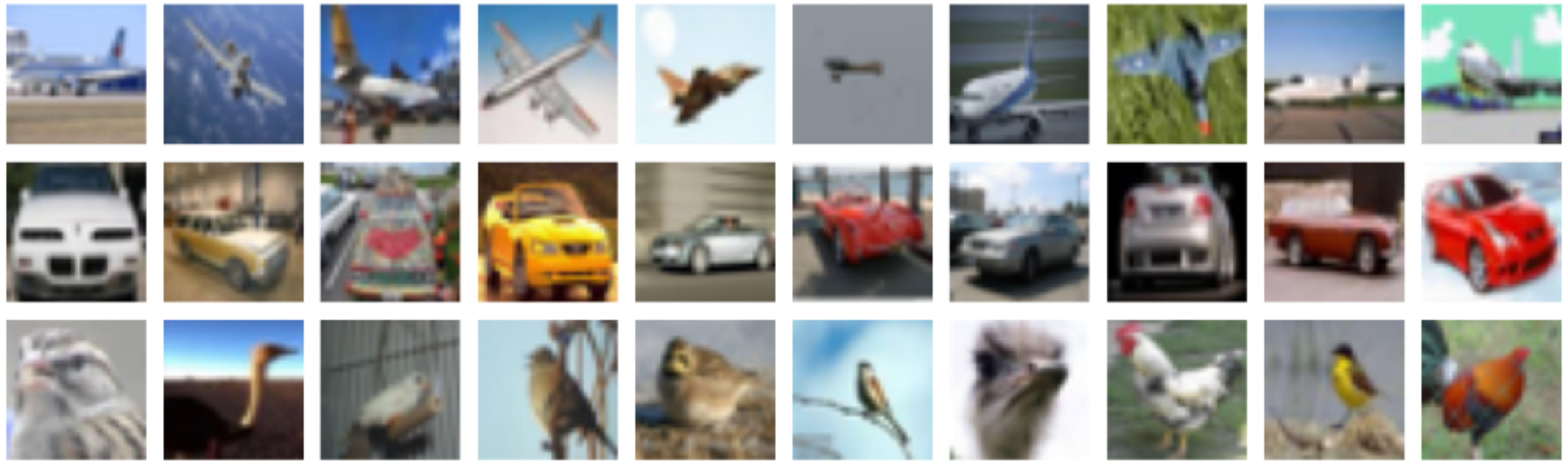
Learning:

$$\arg\min_{\mathbf{W}} \mathscr{L}(\mathbf{W})$$

$y = 1$

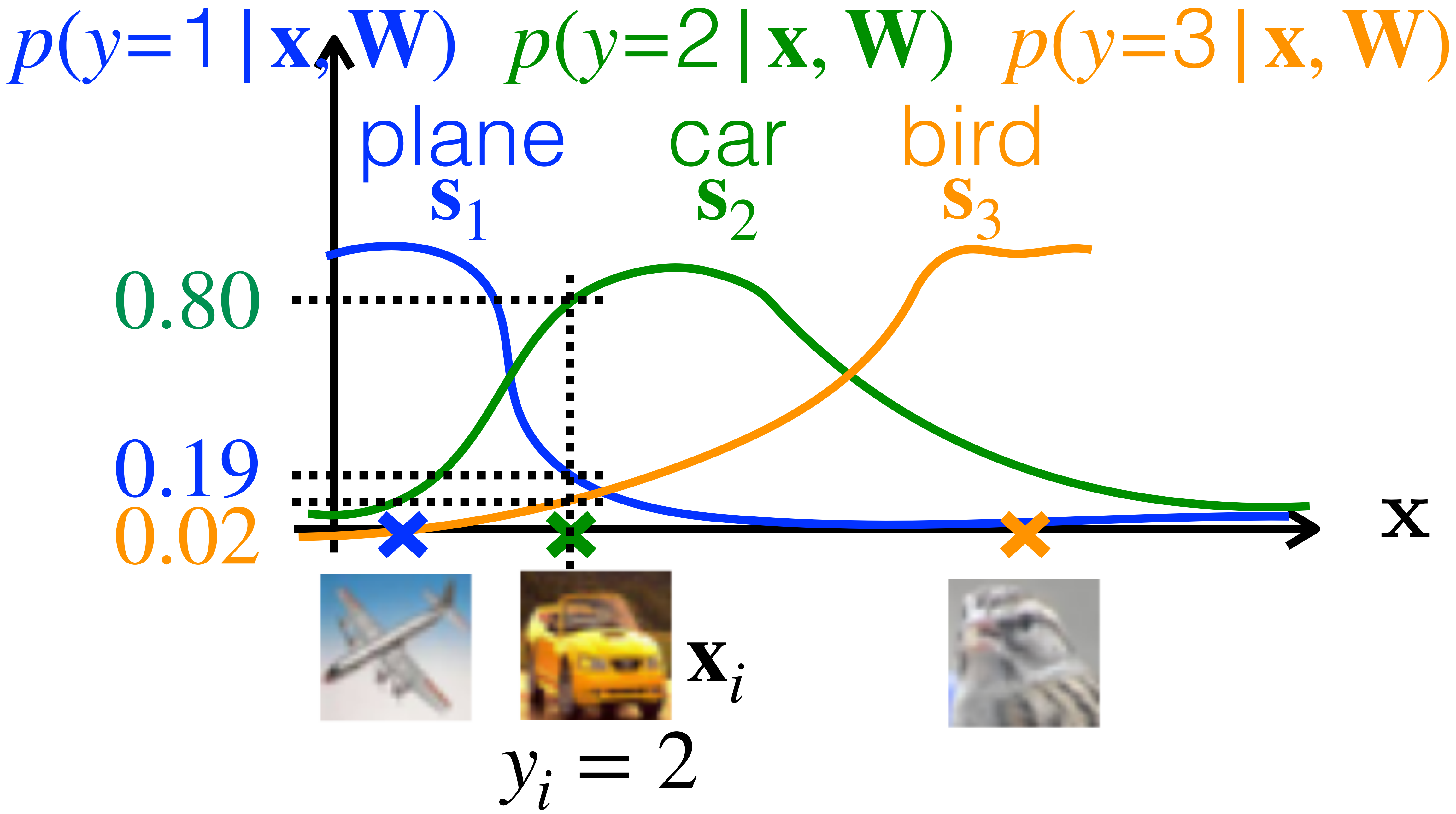
$y = 2$

$y = 3$



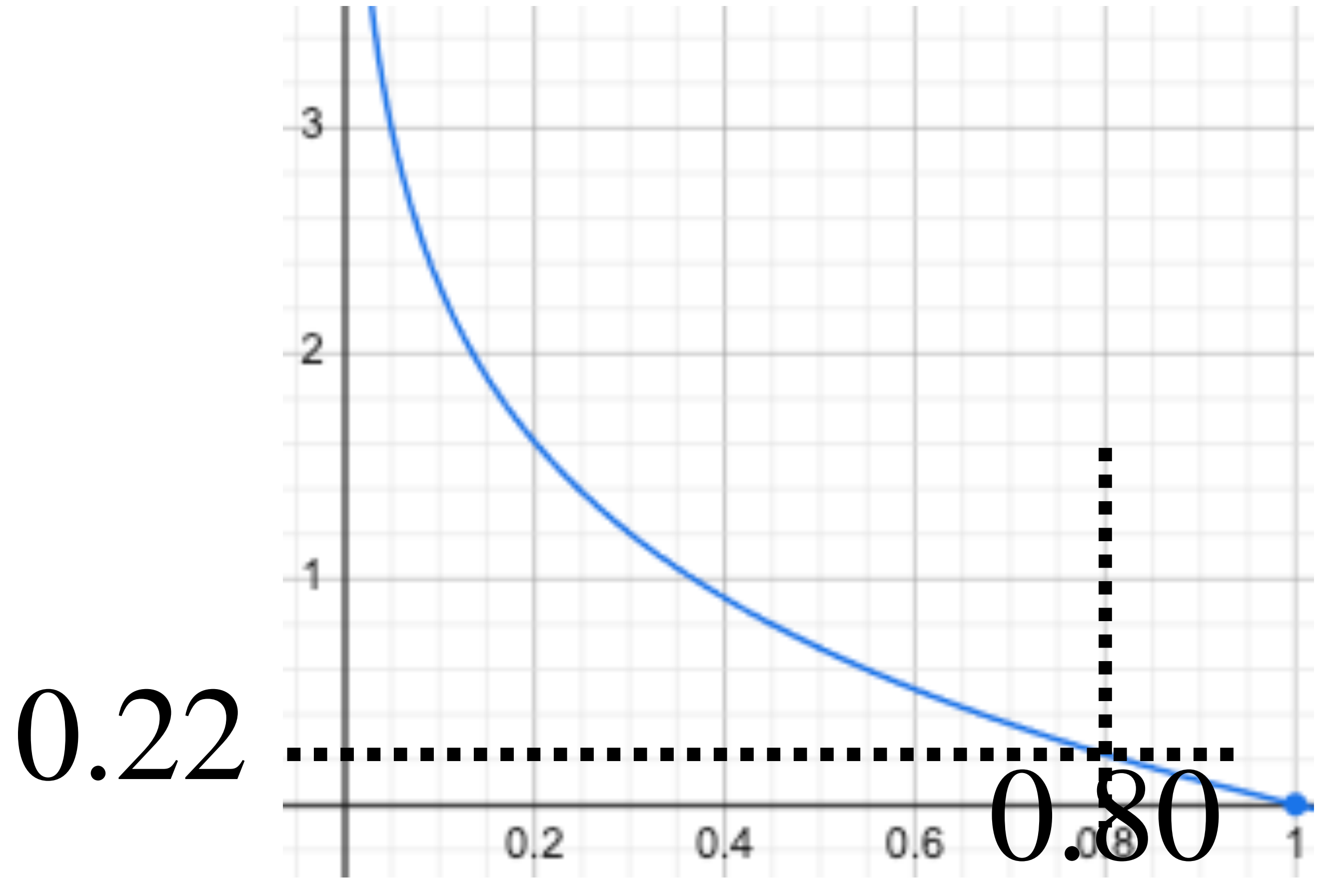
Model probability distribution over classes by softmax function

$$\mathbf{p}(y\,|\,\mathbf{x}, \mathbf{W}) = \begin{bmatrix} \exp(f(\mathbf{x}, \mathbf{w}_1)) \\ \exp(f(\mathbf{x}, \mathbf{w}_2)) \\ \exp(f(\mathbf{x}, \mathbf{w}_3)) \end{bmatrix} \Big/ \sum_k \exp(f(\mathbf{x}, \mathbf{w}_k)) = \mathbf{s}(\mathbf{f}(\mathbf{x}, \mathbf{W})) = \mathbf{s}\left(\begin{bmatrix} +1 \\ 0 \\ -2 \end{bmatrix}\right) = \begin{bmatrix} 0.71 \\ 0.26 \\ 0.03 \end{bmatrix}$$



Loss function:

$$\mathscr{L}(\mathbf{W}) = -\log p(y = y_i\,|\,\mathbf{x}_i, \mathbf{W}) = -\log \mathbf{s}_{y_i}(\mathbf{W}\,\bar{\mathbf{x}}_i)$$



Learning:

$$\arg\min_{\mathbf{W}} \mathscr{L}(\mathbf{W})$$

def train():

1 1 1 2 2 2 3 3 3

$\mathbf{x}_i = \text{vec}($  $)$

$\mathbf{W}^\star = \arg\min_{\mathbf{W}} \mathscr{L}(\mathbf{W}) = \arg\min_{\mathbf{W}} \sum_i -\log \mathbf{s}_{y_i}(\mathbf{W}\,\bar{\mathbf{x}}_i)$ … gradient optimization

return $\mathbf{W}^\star$

reshape($\mathbf{w}_1$) reshape($\mathbf{w}_2$) reshape($\mathbf{w}_3$) reshape($\mathbf{w}_4$) reshape($\mathbf{w}_5$) reshape($\mathbf{w}_6$) reshape($\mathbf{w}_7$) reshape($\mathbf{w}_8$) reshape($\mathbf{w}_9$) reshape($\mathbf{w}_{10}$)

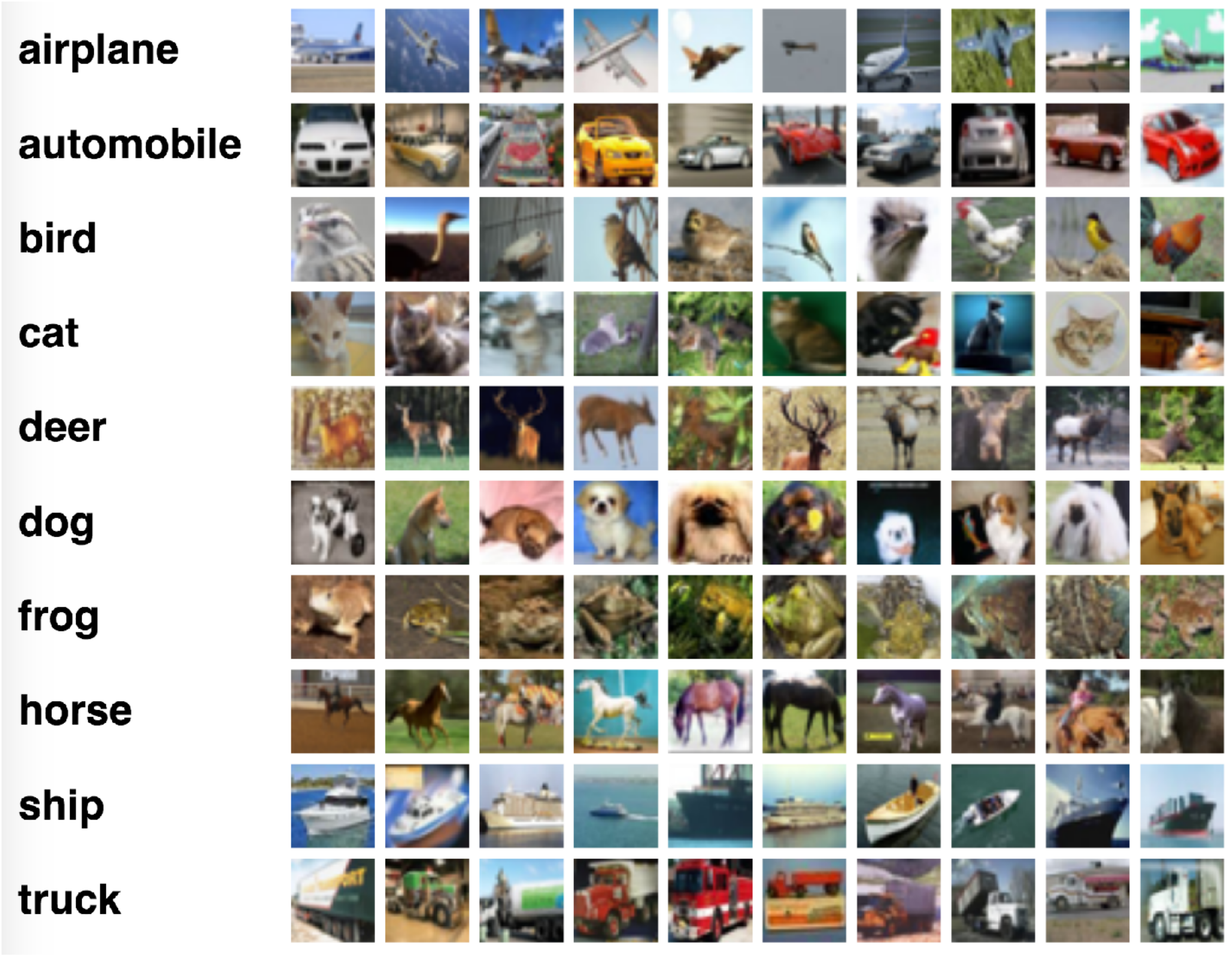
def train(

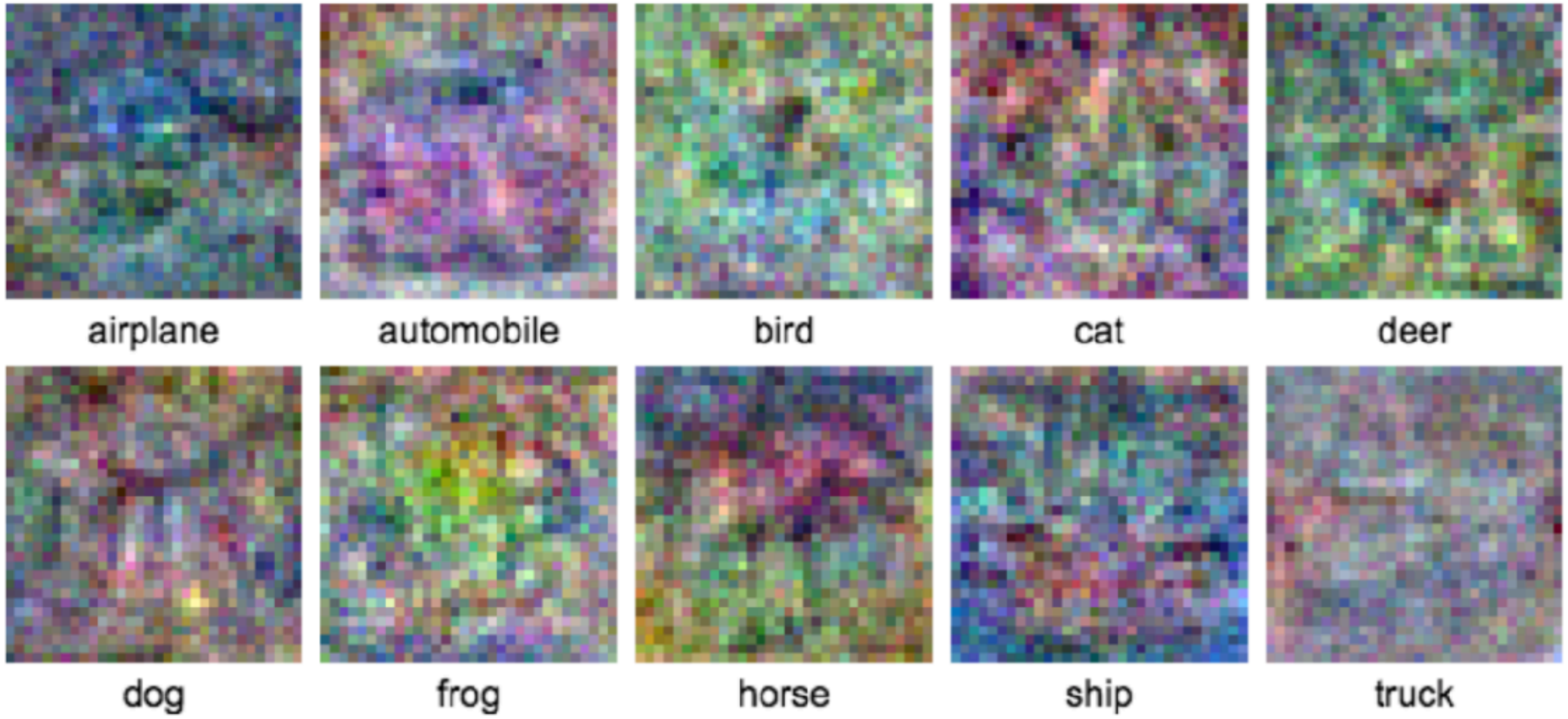
1 1 1 2 2 2 3 3 3):

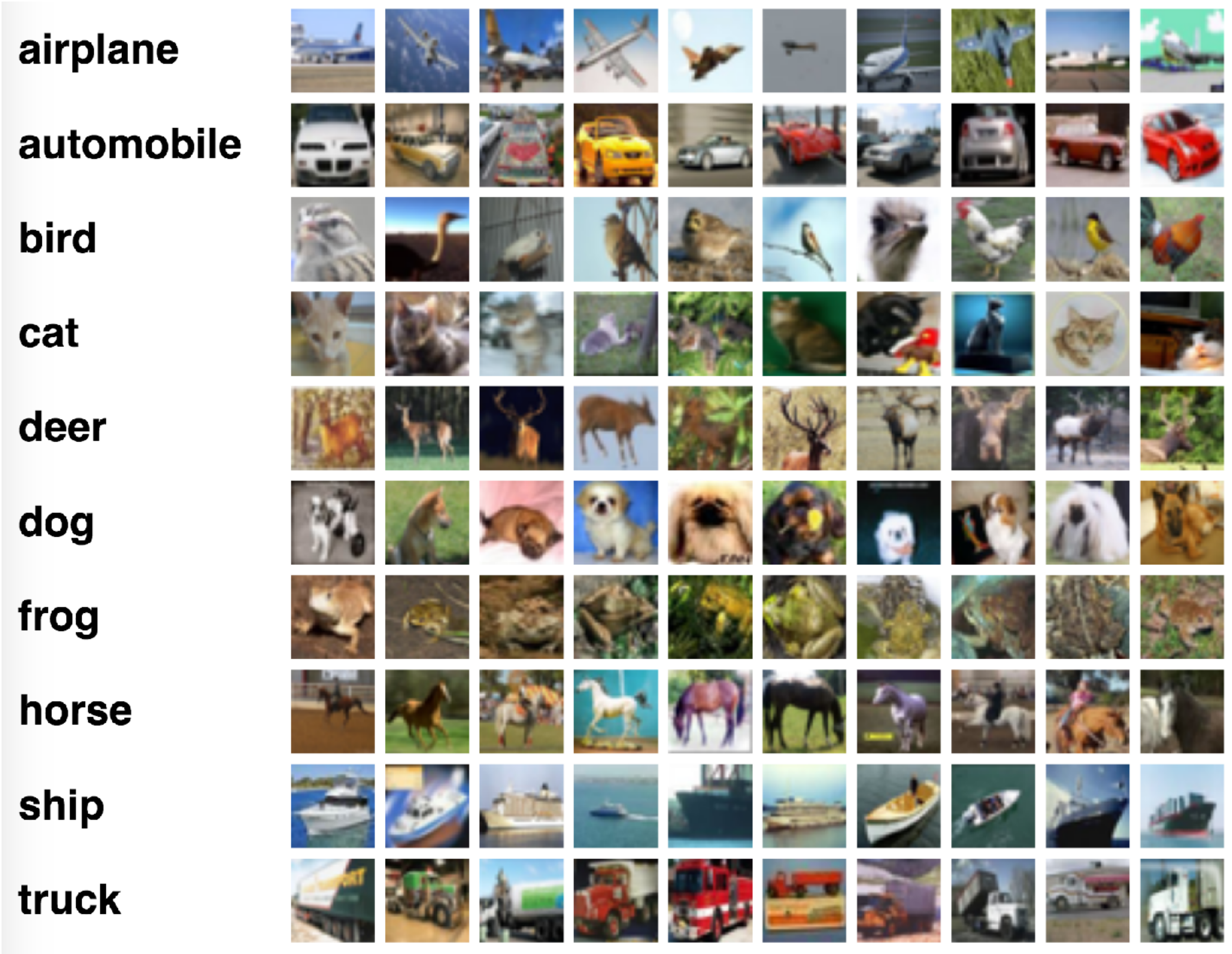
$\mathbf{x}_i = \text{vec}($  $)$

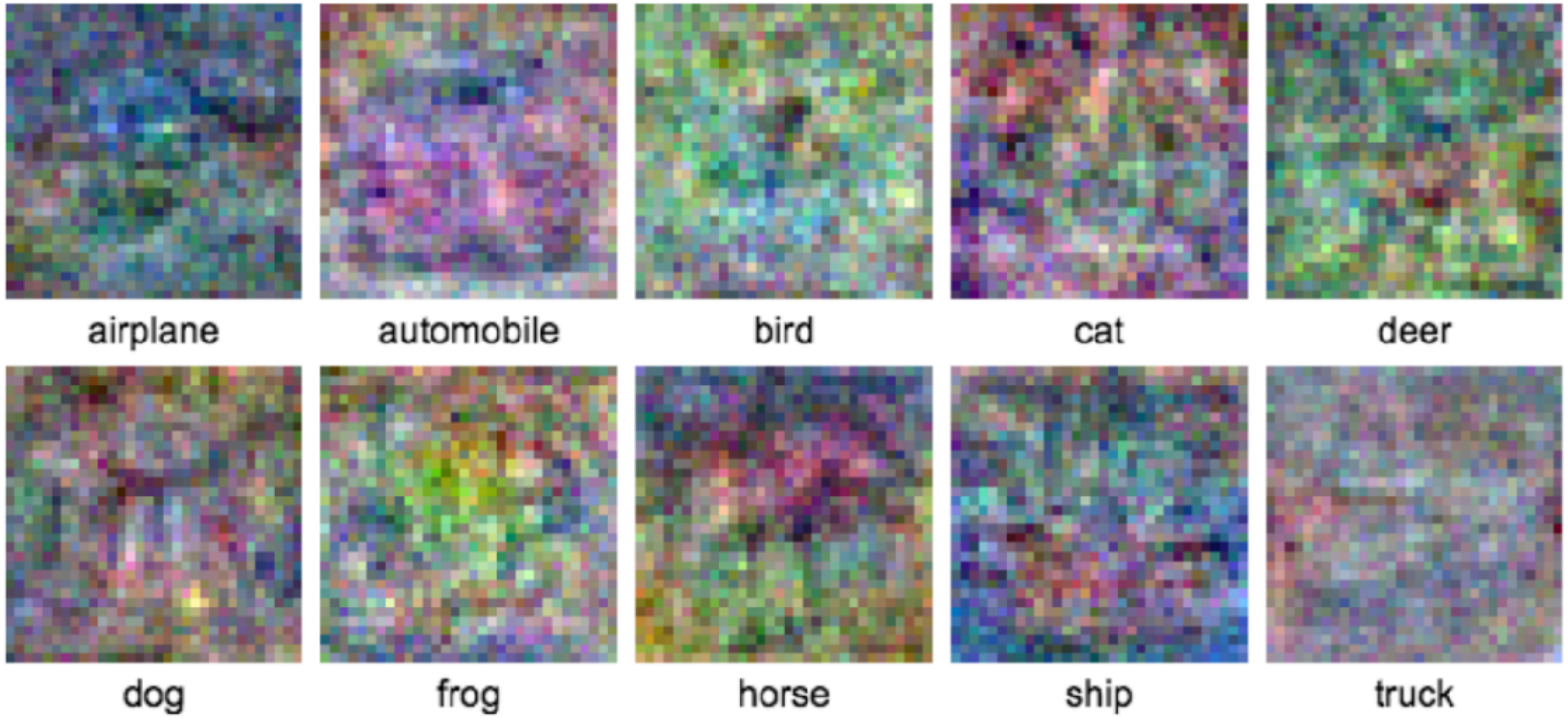
$$\mathbf{W}^\star = \arg\min_{\mathbf{W}} \mathscr{L}(\mathbf{W}) = \arg\min_{\mathbf{W}} \sum_i -\log \mathbf{s}_{y_i}(\mathbf{W}\,\overline{\mathbf{x}}_i)$$ … gradient optimization

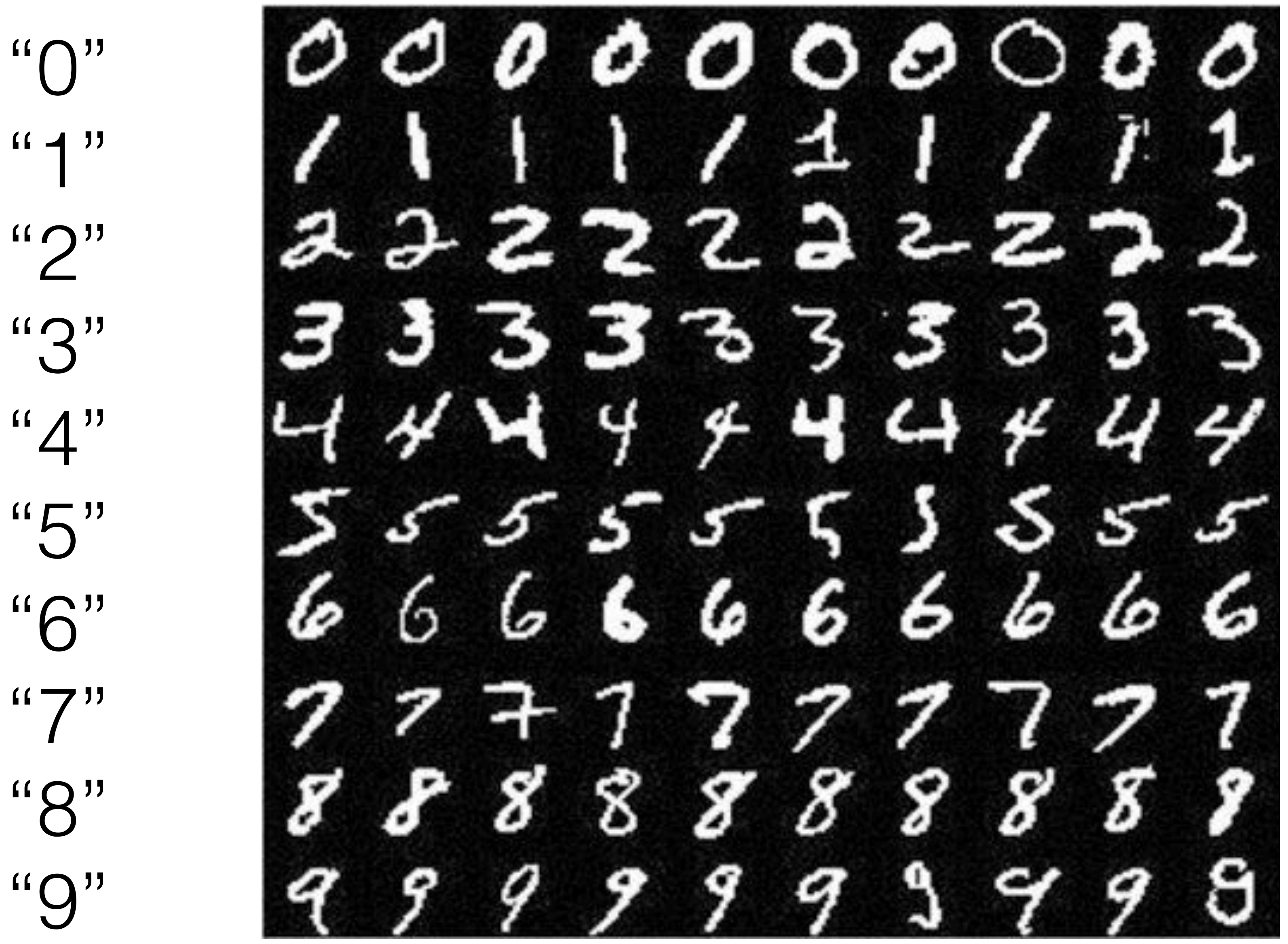
return $\mathbf{W}^\star$

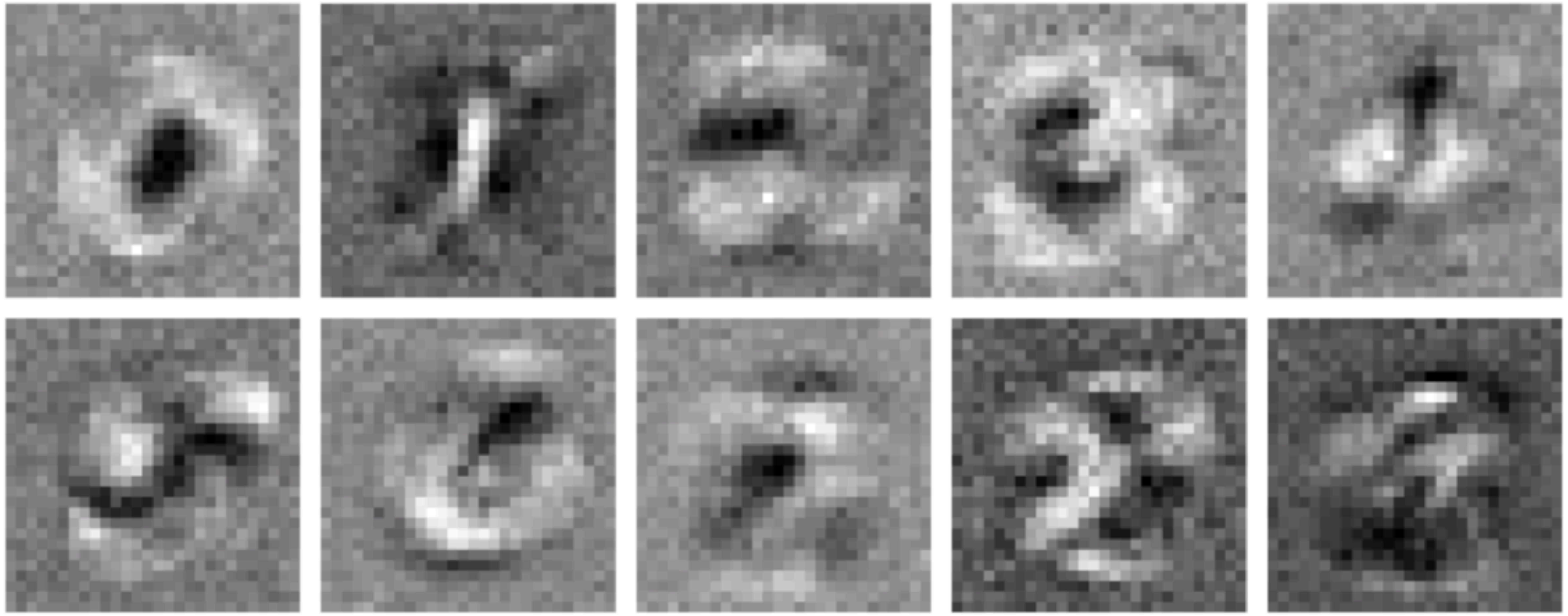

airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

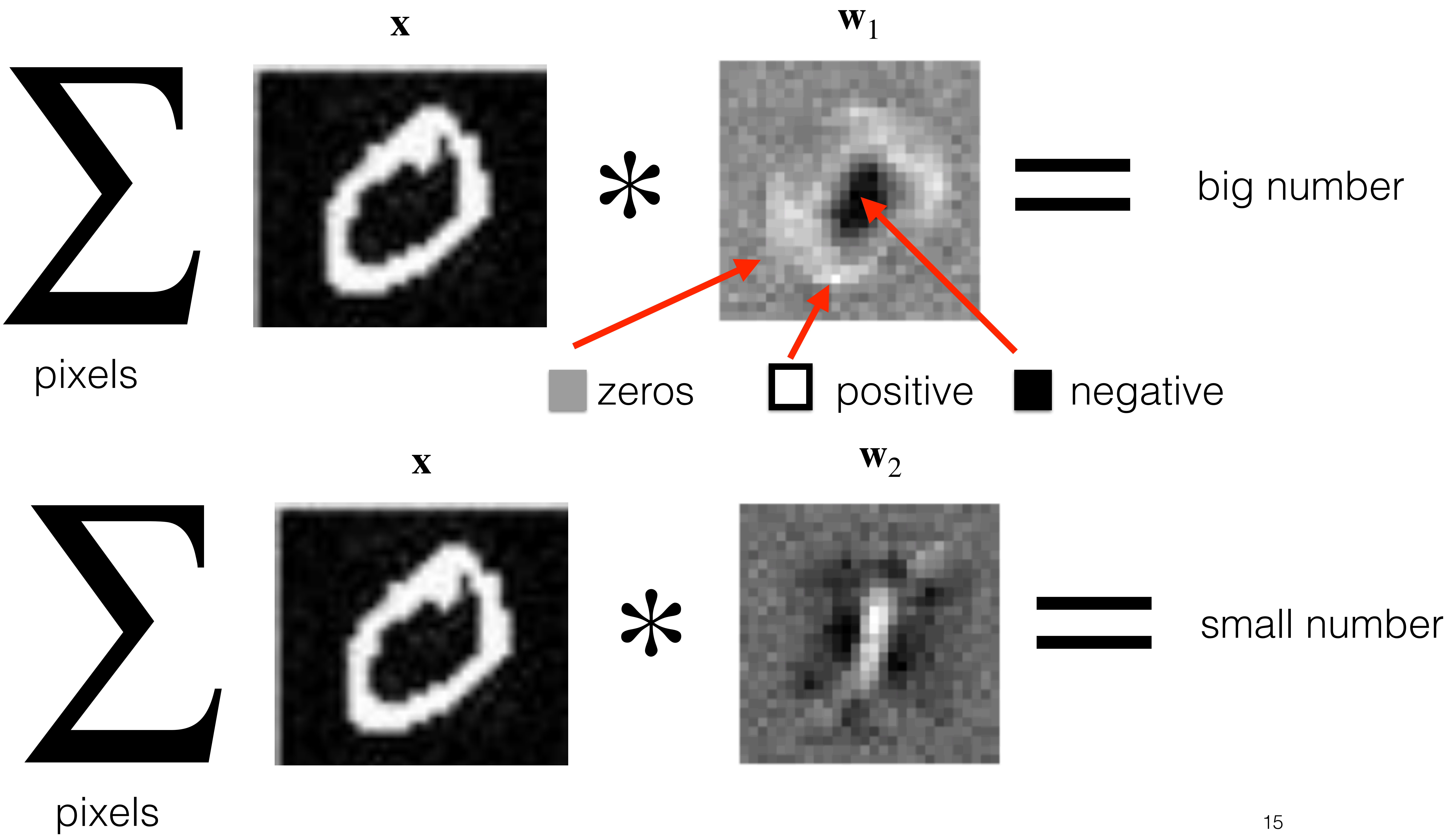

airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck


airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck


airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck


“0”
“1”
“2”
“3”
“4”
“5”
“6”
“7”
“8”
“9”



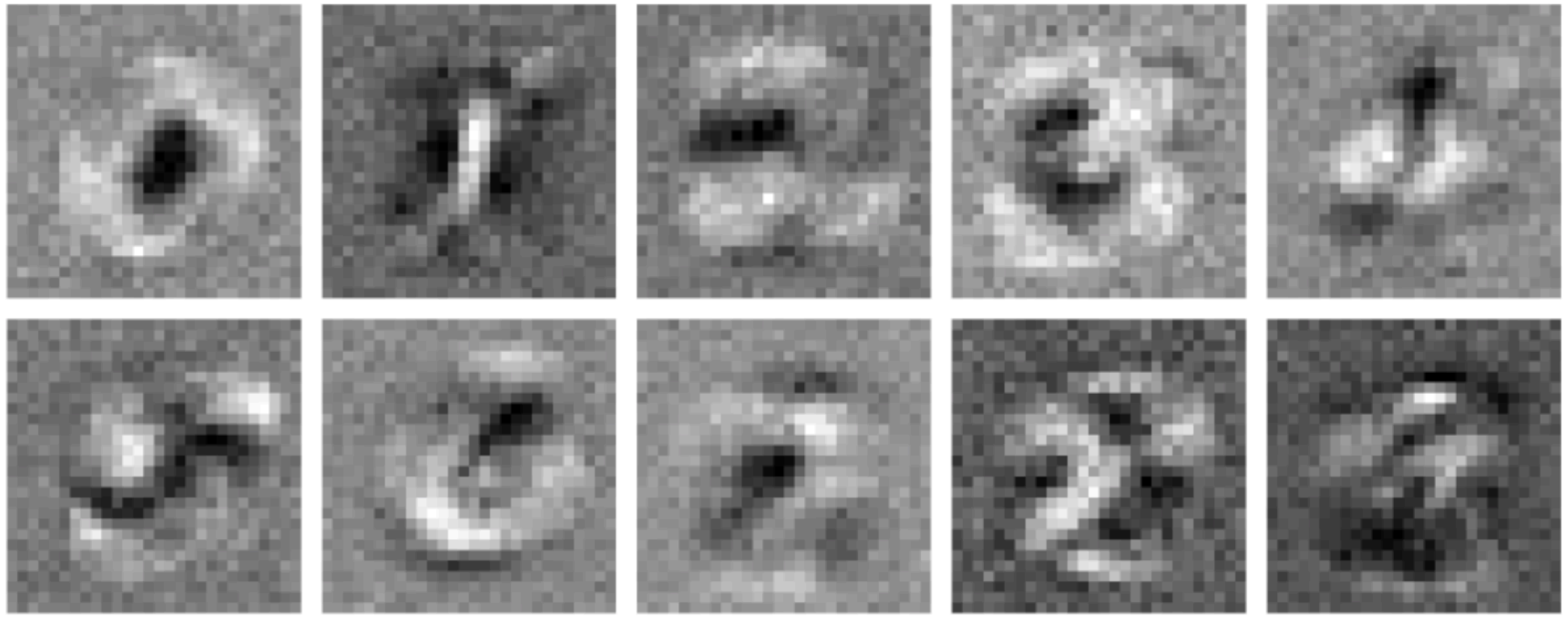


x
w1
big number
pixels
zeros
positive
negative
x
w2
small number
pixels

Dataset	Label	Images	Learned weights of linear classifier	Error

MNIST

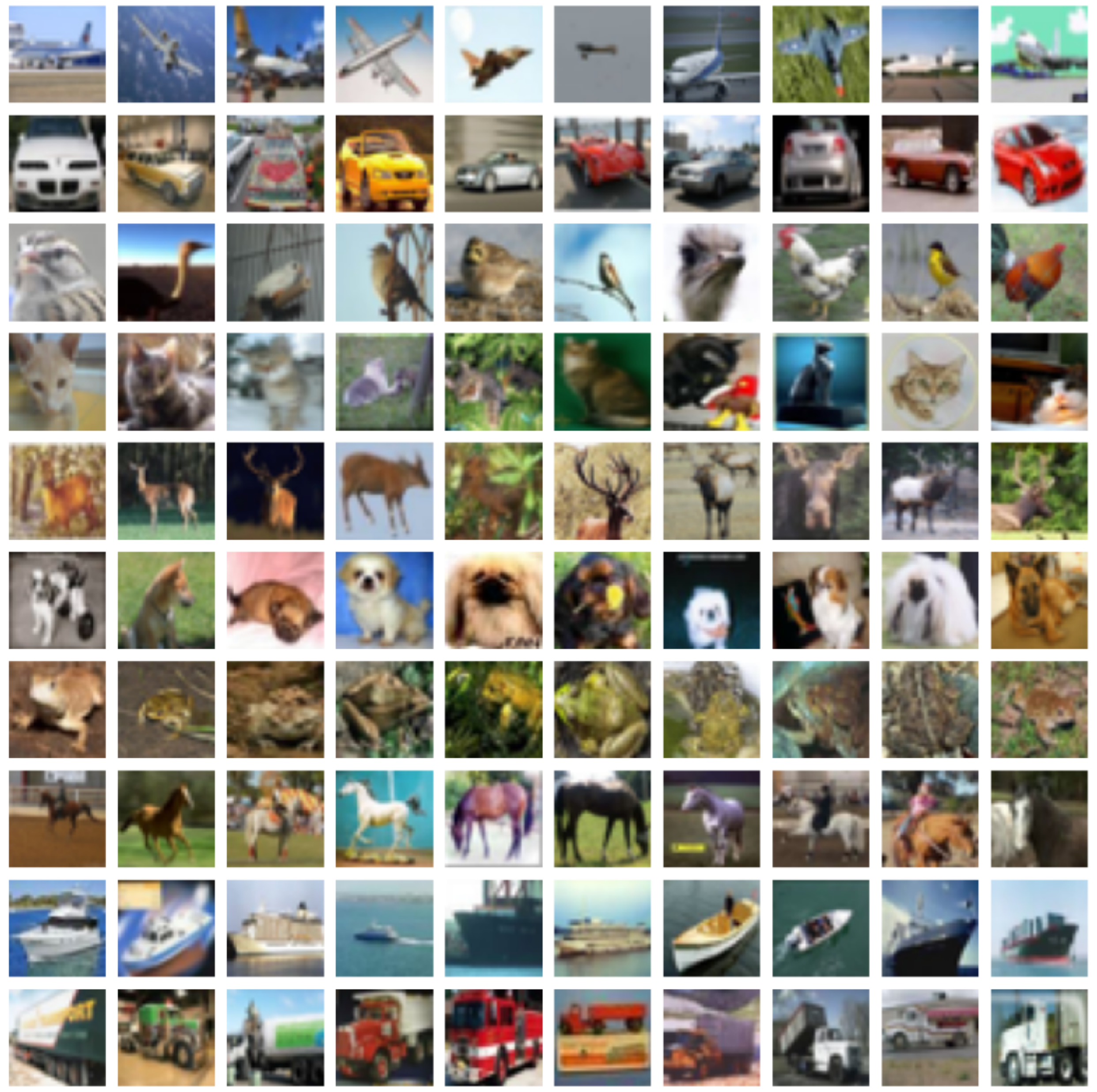
"0"
"1"
"2"
"3"
"4"
"5"
"6"
"7"
"8"
"9"

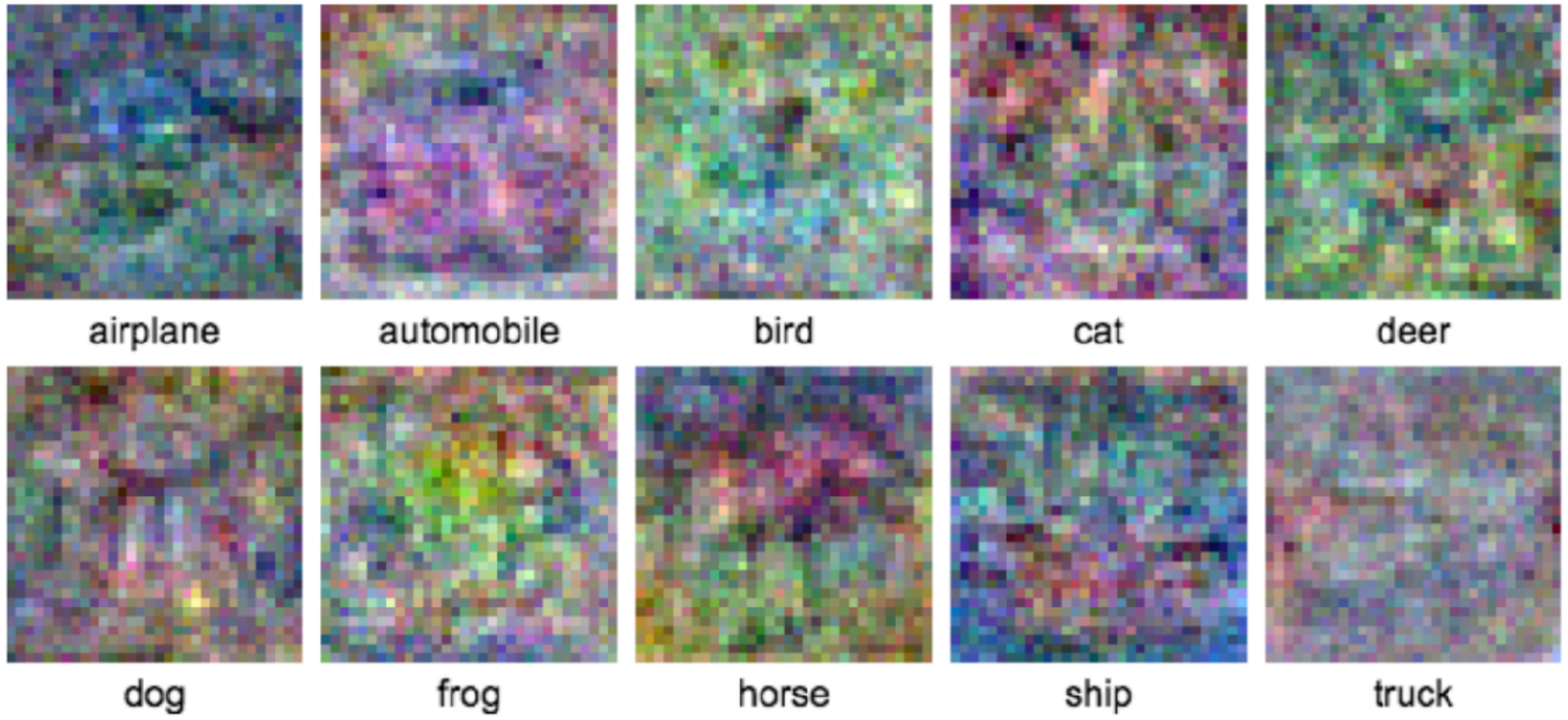


???

CIFAR-10

airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck



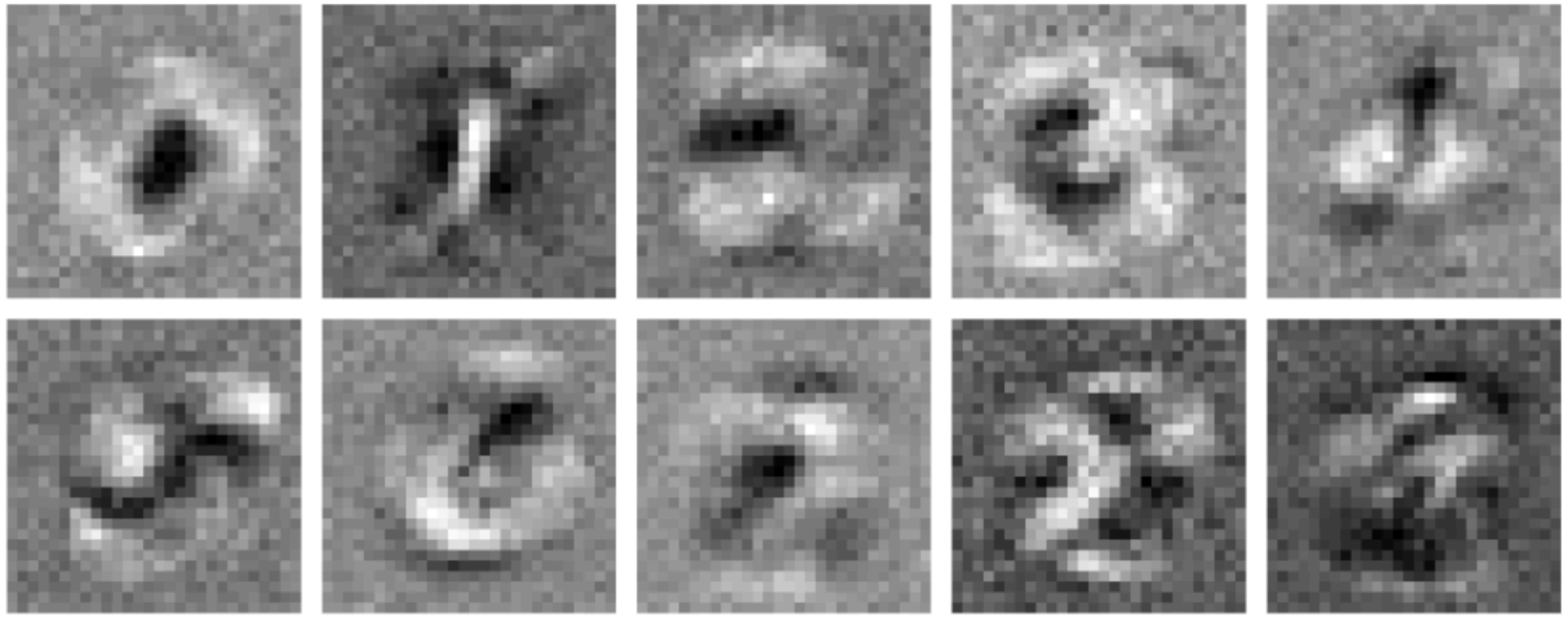


???

Dataset | Label | Images | Learned weights of linear classifier | Error

MNIST

“0”
“1”
“2”
“3”
“4”
“5”
“6”
“7”
“8”
“9”

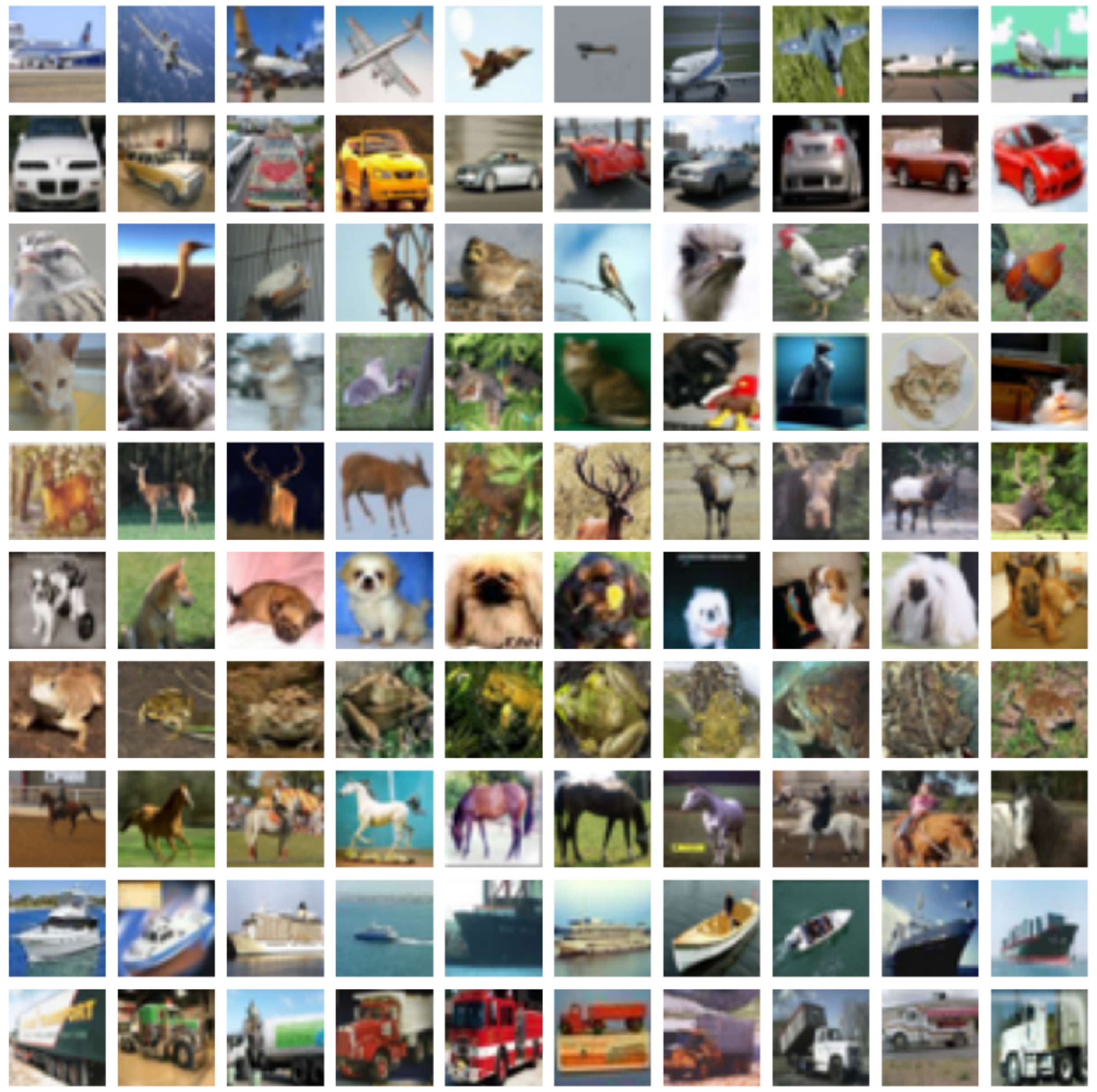


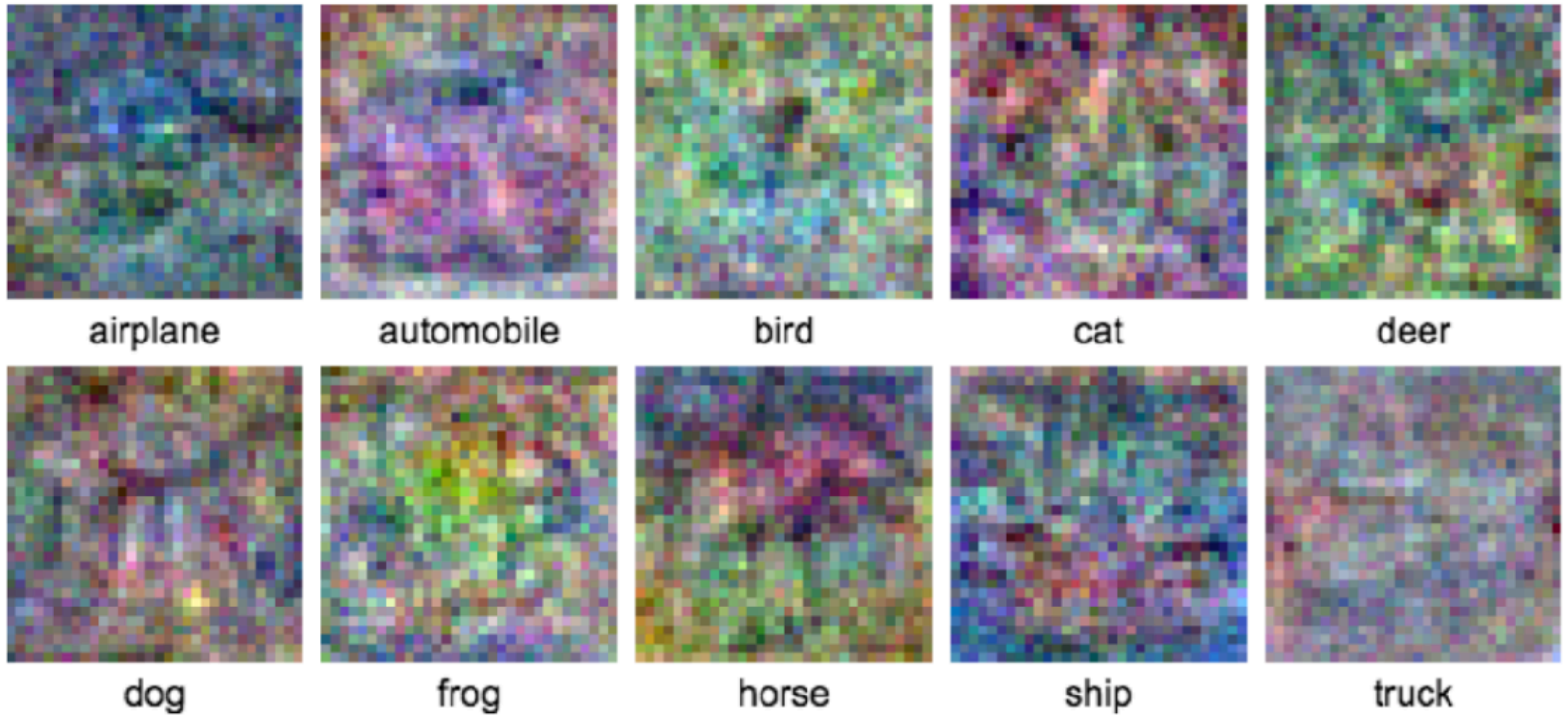
8%

Why is it simple?

CIFAR-10

airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck





63%

Why is it hard?

Why is it hard?

Huge within-class variability!

bird

cat

dog

Due to:

- Viewpoint
- Occlusion
- Illumination
- Pose
- Type
- Context

Why is it hard?

Huge within-class variability!

Due to:

- Viewpoint
- Occlusion
- Illumination
- Pose
- Type
- Context

Why is it hard?

Huge among-class similarity!

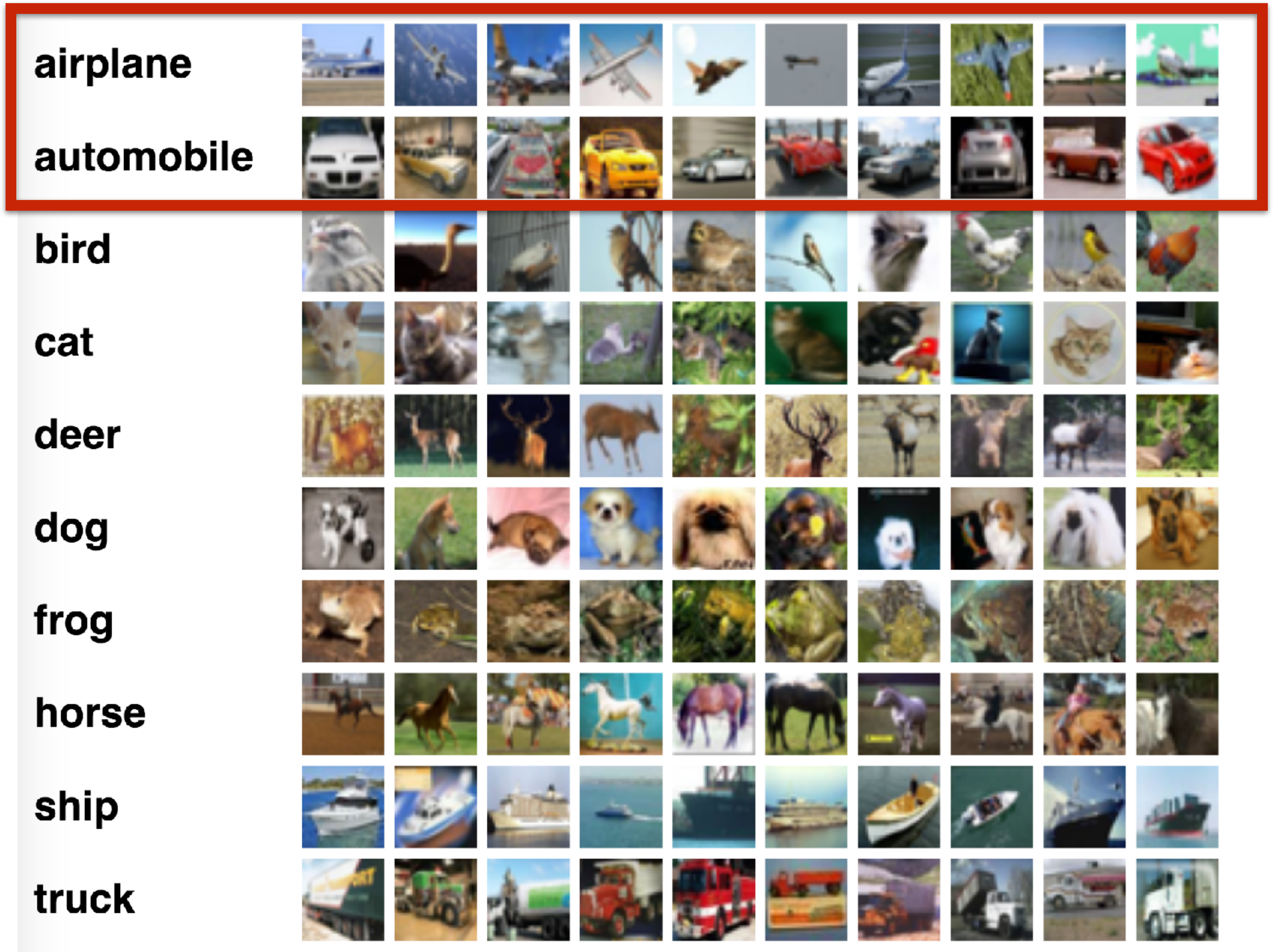
Why is it hard?

Huge among-class similarity!

Why is it hard?

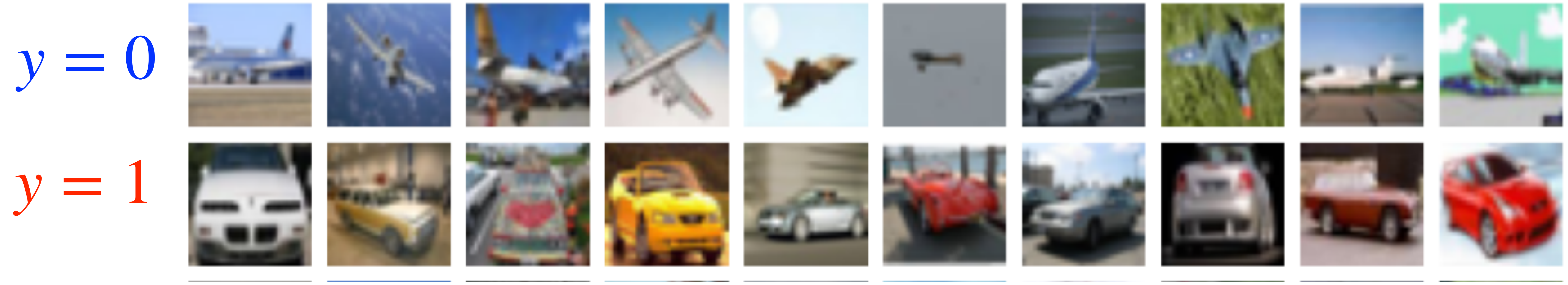
Huge among-class similarity!

Recognition problem

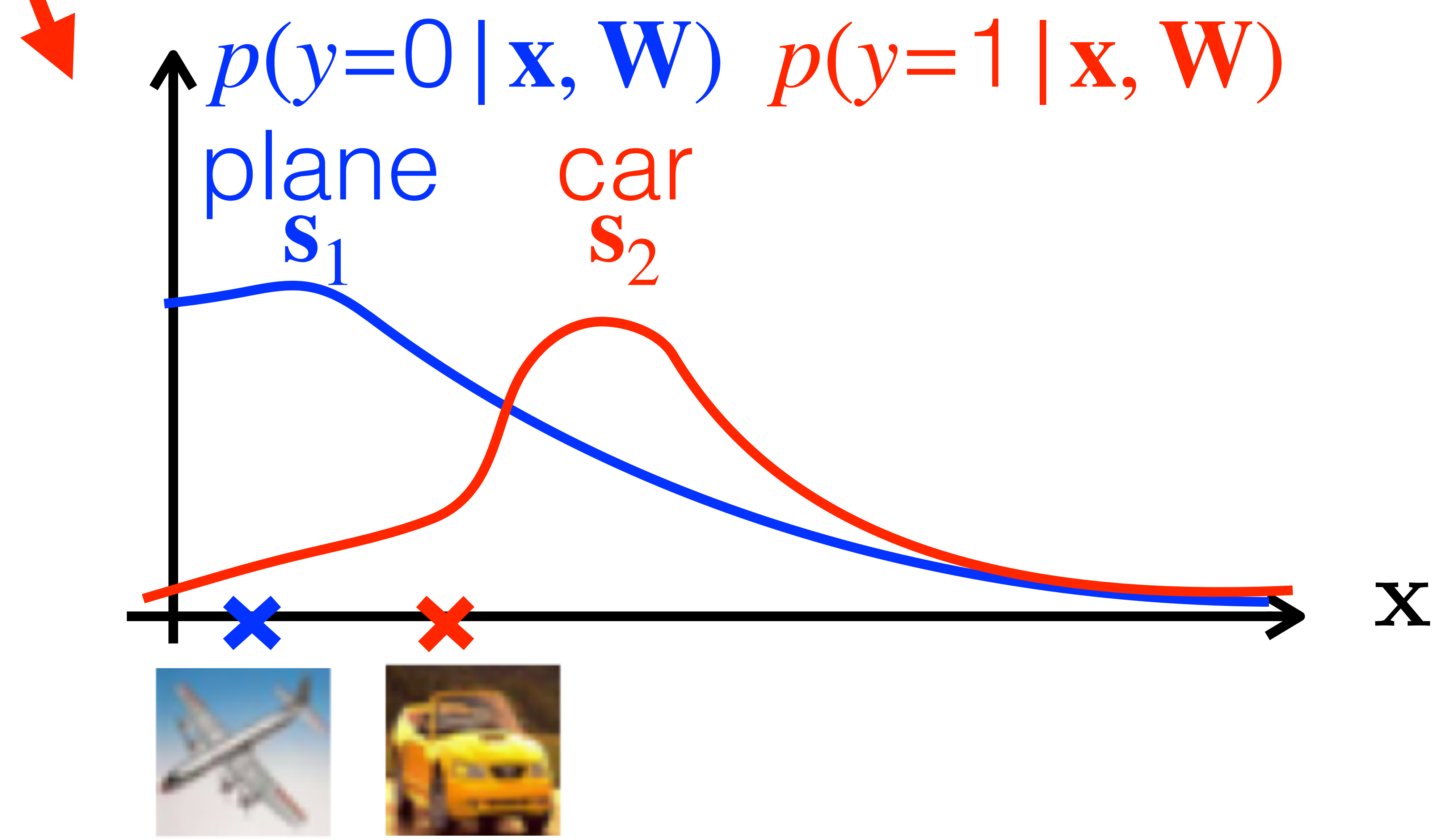


CIFAR-10: classify 32x32 RGB images into 10 categories

https://www.cs.toronto.edu/~kriz/cifar.html

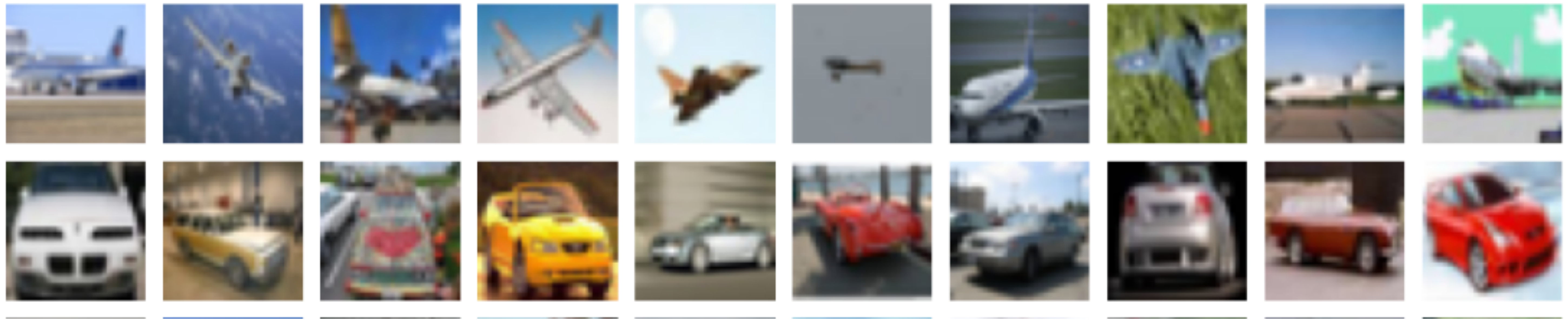


$$\mathbf{p}(y \mid \mathbf{x}, \mathbf{W}) = \begin{bmatrix} \exp(f(\mathbf{x}, \mathbf{w}_1)) \\ \exp(f(\mathbf{x}, \mathbf{w}_2)) \end{bmatrix} \Big/ \sum_k \exp(f(\mathbf{x}, \mathbf{w}_k)) = \mathbf{s}(\mathbf{f}(\mathbf{x}, \mathbf{W})) \in \mathbb{R}^2$$



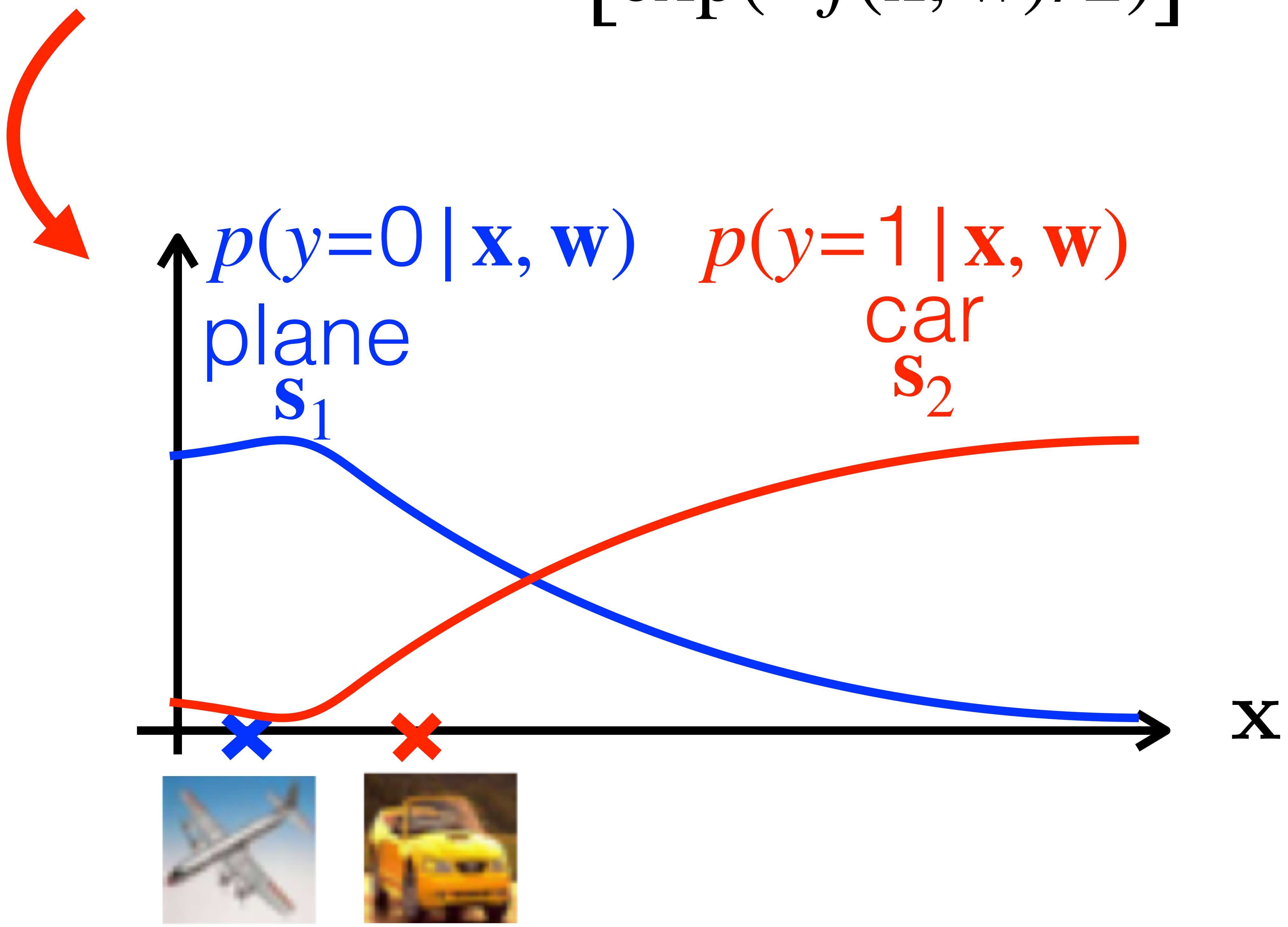
$y = 0$

$y = 1$



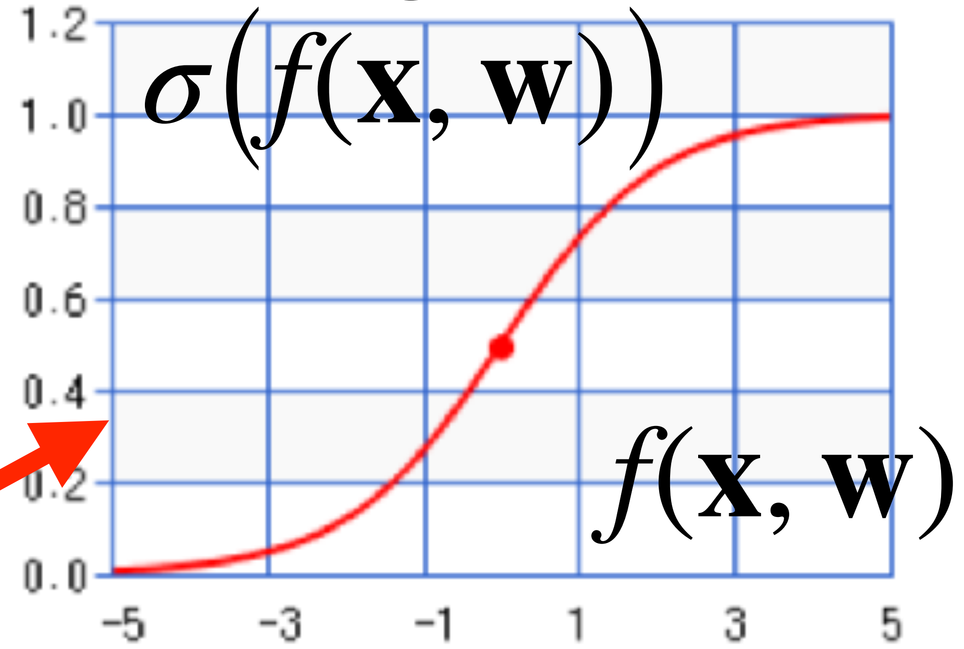
$$\mathbf{p}(y \mid \mathbf{x}, \mathbf{W}) = \begin{bmatrix} \exp(f(\mathbf{x}, \mathbf{w}_1)) \\ \exp(f(\mathbf{x}, \mathbf{w}_2)) \end{bmatrix} / \sum_k \exp\big(f(\mathbf{x}, \mathbf{w}_k\big) = \mathbf{s}\big(\mathbf{f}\big(\mathbf{x}, \mathbf{W}\big)\big) \in \mathbb{R}^2$$

$$\mathbf{p}(y \mid \mathbf{x}, \mathbf{w}) = \begin{bmatrix} \exp(f(\mathbf{x}, \mathbf{w})/2) \\ \exp(-f(\mathbf{x}, \mathbf{w})/2) \end{bmatrix} / \Big(\exp\big(f(\mathbf{x}, \mathbf{w})/2\big) + \exp(-f(\mathbf{x}, \mathbf{w})/2)\Big)$$



$$= \begin{bmatrix} \dfrac{1}{1 + \frac{\exp(-f(\mathbf{x}, \mathbf{w})/2)}{\exp(f(\mathbf{x}, \mathbf{w})/2)}} \\ 1 - \dfrac{1}{1 + \frac{\exp(-f(\mathbf{x}, \mathbf{w})/2)}{\exp(f(\mathbf{x}, \mathbf{w})/2)}} \end{bmatrix}$$

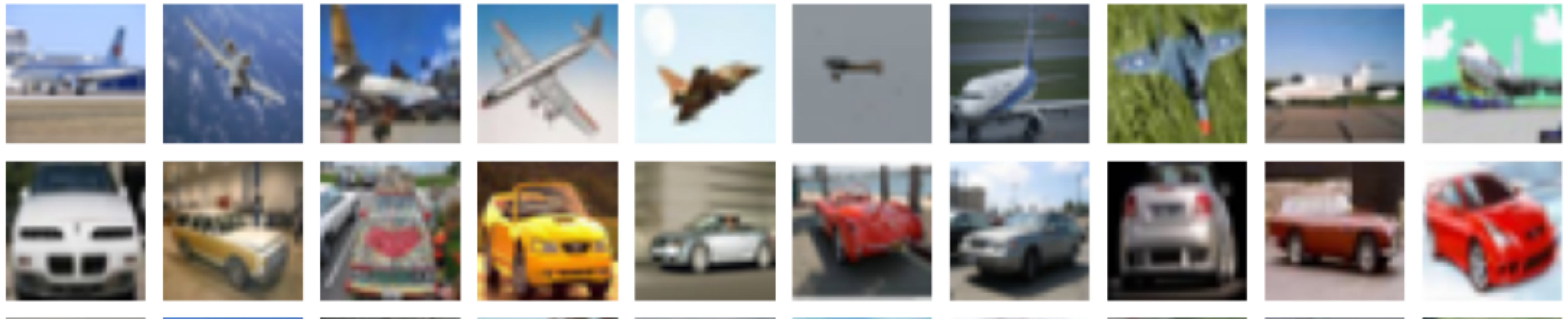
sigmoid



$$= \begin{bmatrix} \dfrac{1}{1 + \exp(-f(\mathbf{x}, \mathbf{w}))} \\ 1 - \dfrac{1}{1 + \exp(-f(\mathbf{x}, \mathbf{w}))} \end{bmatrix} = \begin{bmatrix} \sigma\big(f(\mathbf{x}, \mathbf{w})\big) \\ 1 - \sigma\big(f(\mathbf{x}, \mathbf{w})\big) \end{bmatrix}$$

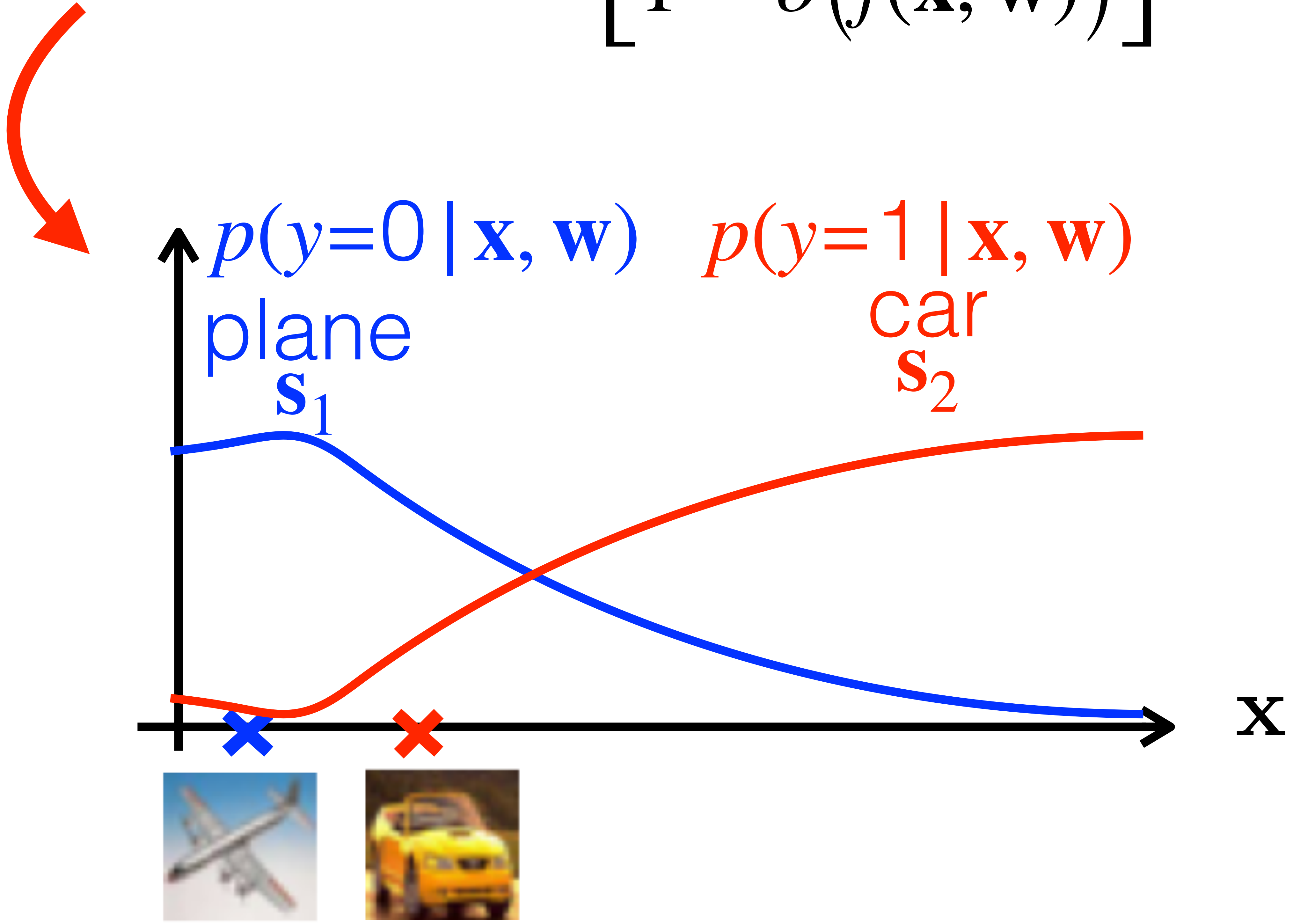
$y = 0$

$y = 1$



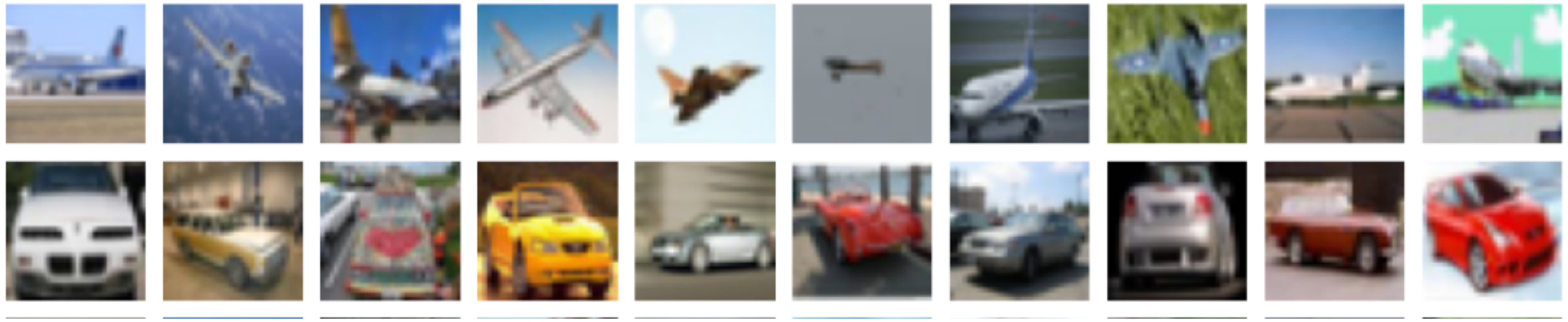
$$\mathbf{p}(y \mid \mathbf{x}, \mathbf{W}) = \begin{bmatrix} \exp(f(\mathbf{x}, \mathbf{w}_1)) \\ \exp(f(\mathbf{x}, \mathbf{w}_2)) \end{bmatrix} \Big/ \sum_k \exp\big(f(\mathbf{x}, \mathbf{w}_k)\big) = \mathbf{s}\big(\mathbf{f}(\mathbf{x}, \mathbf{W})\big) \in \mathbb{R}^2$$

$$\mathbf{p}(y \mid \mathbf{x}, \mathbf{w}) = \begin{bmatrix} \sigma\big(f(\mathbf{x}, \mathbf{w})\big) \\ 1 - \sigma\big(f(\mathbf{x}, \mathbf{w})\big) \end{bmatrix} \in \mathbb{R}^2$$



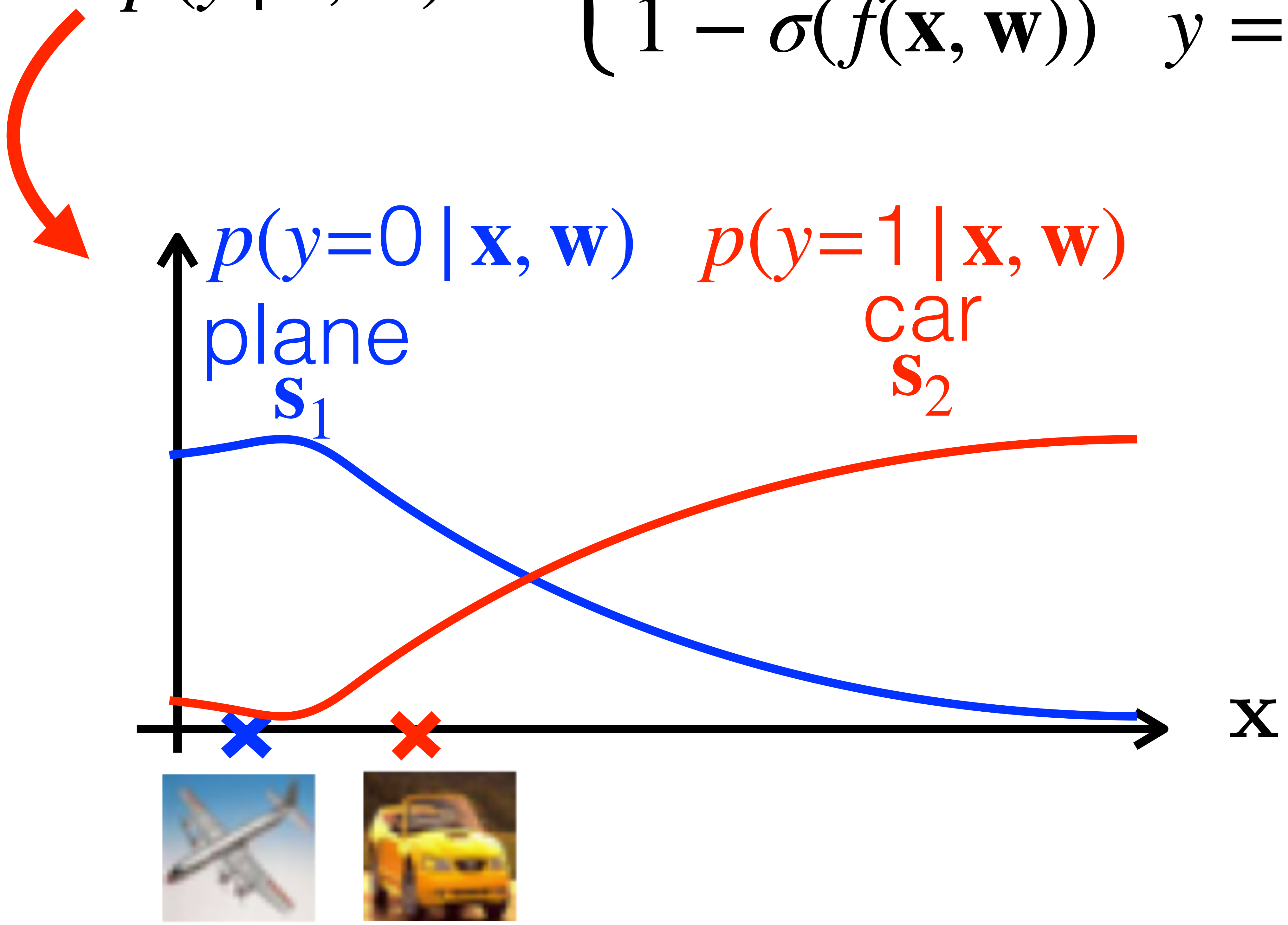
$y = 0$

$y = 1$



$$\mathbf{p}(y \mid \mathbf{x}, \mathbf{W}) = \begin{bmatrix} \exp(f(\mathbf{x}, \mathbf{w}_1)) \\ \exp(f(\mathbf{x}, \mathbf{w}_2)) \end{bmatrix} / \sum_k \exp\left(f\left(\mathbf{x}, \mathbf{w}_k\right)\right) = \mathbf{s}\left(\mathbf{f}\left(\mathbf{x}, \mathbf{W}\right)\right) \in \mathbb{R}^2$$

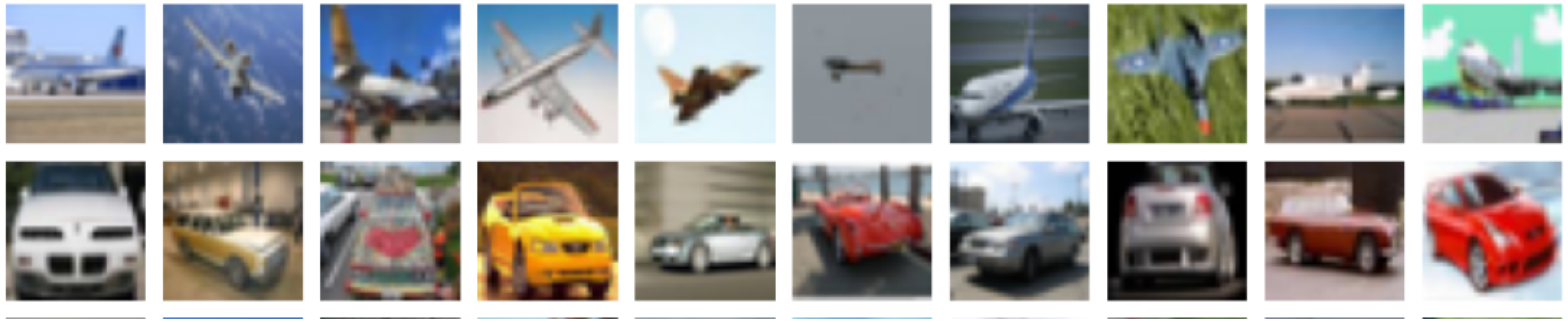
$$p(y \mid \mathbf{x}, \mathbf{w}) = \begin{cases} \sigma(f(\mathbf{x}, \mathbf{w})) & y = 1 \\ 1 - \sigma(f(\mathbf{x}, \mathbf{w})) & y = 0 \end{cases} \in \mathbb{R}^1$$



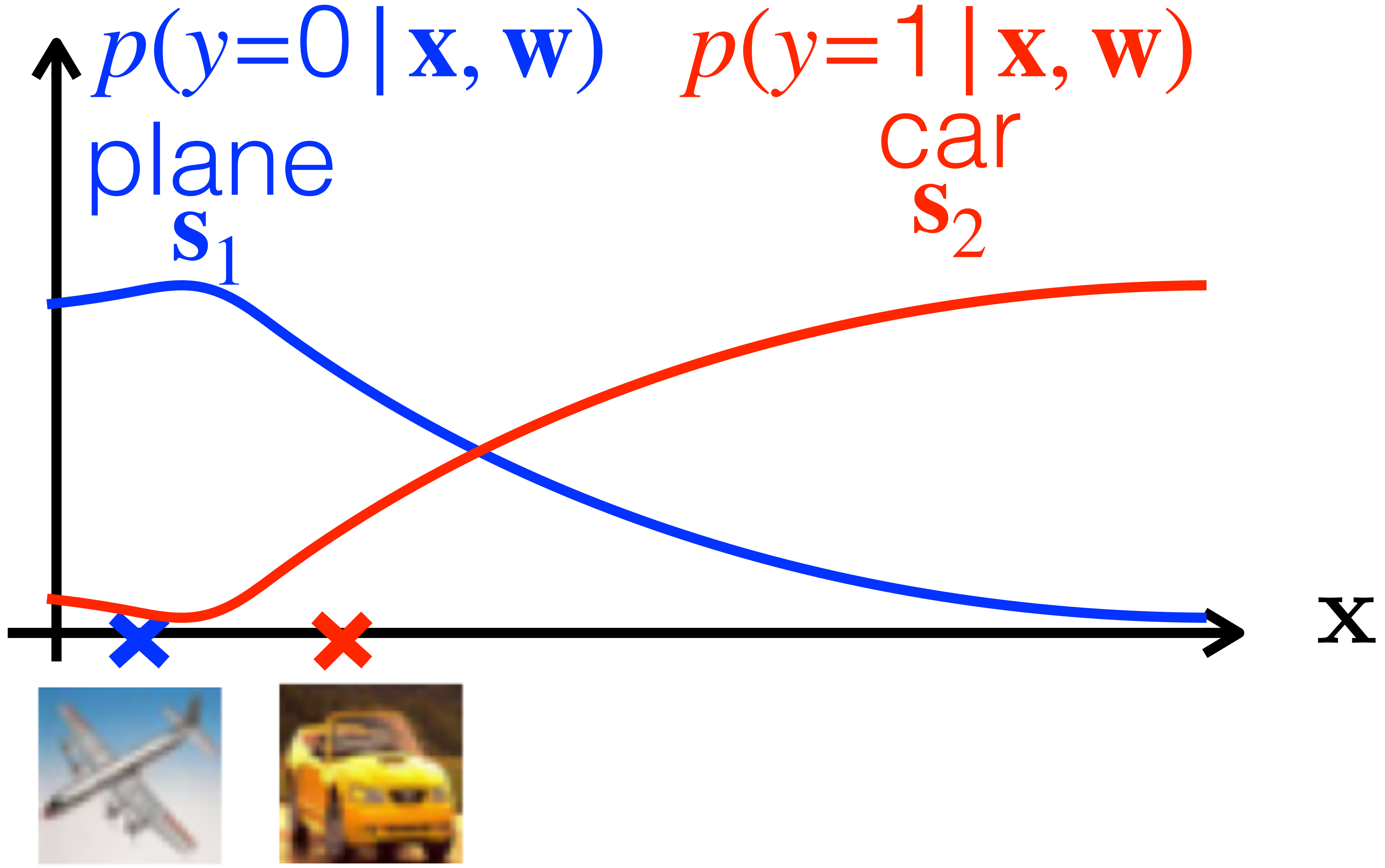
$$\mathbf{p}(y \mid \mathbf{x}, \mathbf{w}) = \begin{bmatrix} \sigma\left(f(\mathbf{x}, \mathbf{w})\right) \\ 1 - \sigma\left(f(\mathbf{x}, \mathbf{w})\right) \end{bmatrix} \in \mathbb{R}^2$$

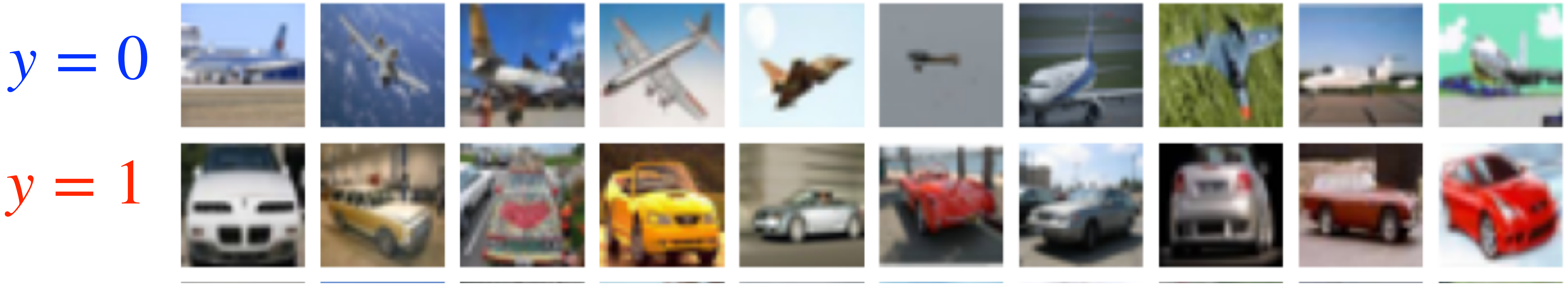
$y = 0$

$y = 1$



$$p(y \mid \mathbf{x}, \mathbf{w}) = \begin{cases} \sigma(f(\mathbf{x}, \mathbf{w})) & y = 1 \\ 1 - \sigma(f(\mathbf{x}, \mathbf{w})) & y = 0 \end{cases} = y_i \cdot \sigma(f(\mathbf{x}, \mathbf{w})) + (1 - y_i) \cdot (1 - \sigma(f(\mathbf{x}, \mathbf{w})))$$



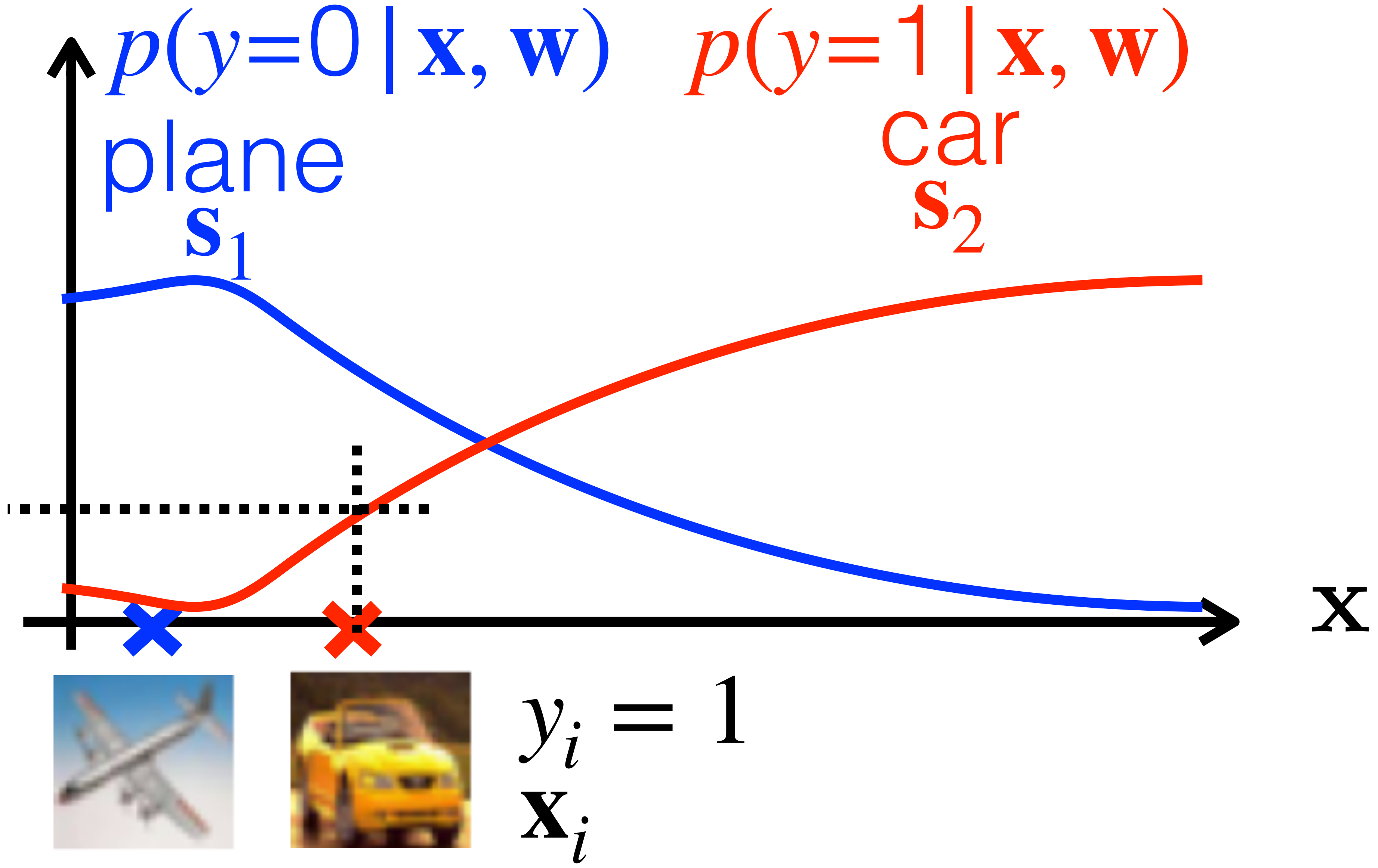


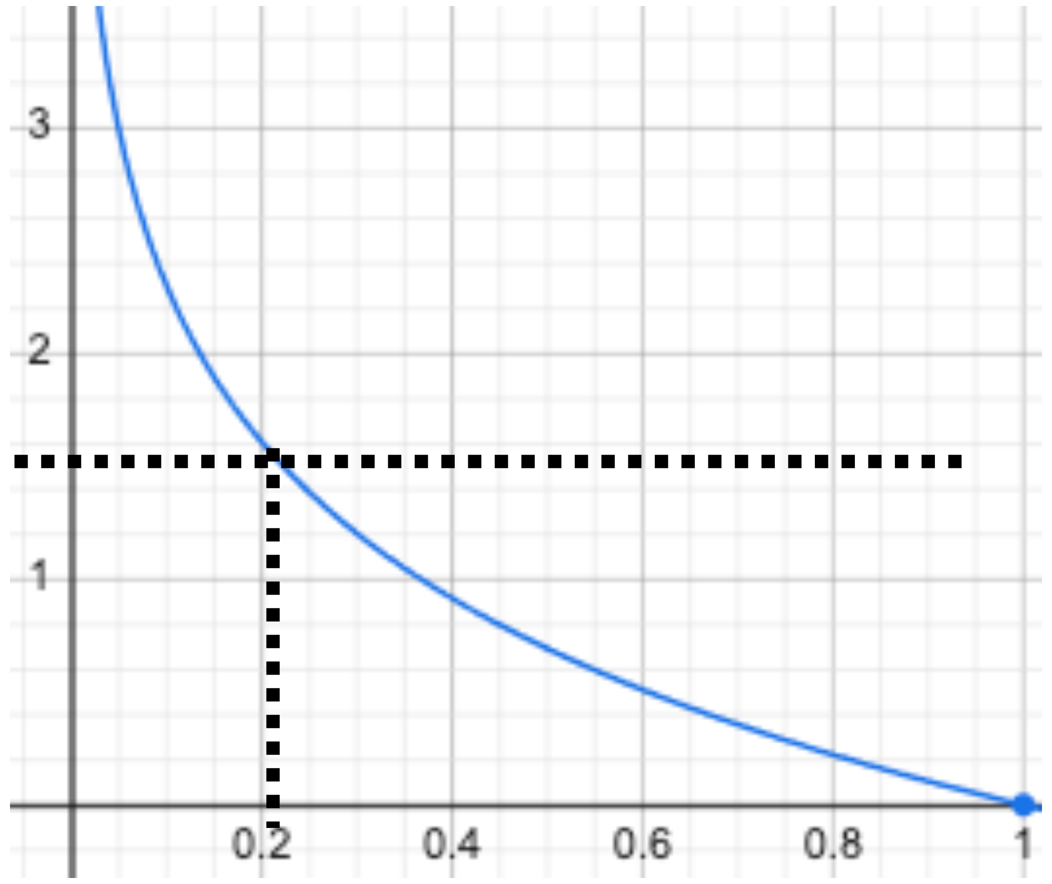
$$p(y \mid \mathbf{x}, \mathbf{w}) = \begin{cases} \sigma(f(\mathbf{x}, \mathbf{w})) & y = 1 \\ 1 - \sigma(f(\mathbf{x}, \mathbf{w})) & y = 0 \end{cases} = y_i \cdot \sigma(f(\mathbf{x}, \mathbf{w})) + (1 - y_i) \cdot (1 - \sigma(f(\mathbf{x}, \mathbf{w})))$$

Loss function:

$$\mathscr{L}(\mathbf{w}) = -\log p(y_i \mid \mathbf{x}_i, \mathbf{w})$$

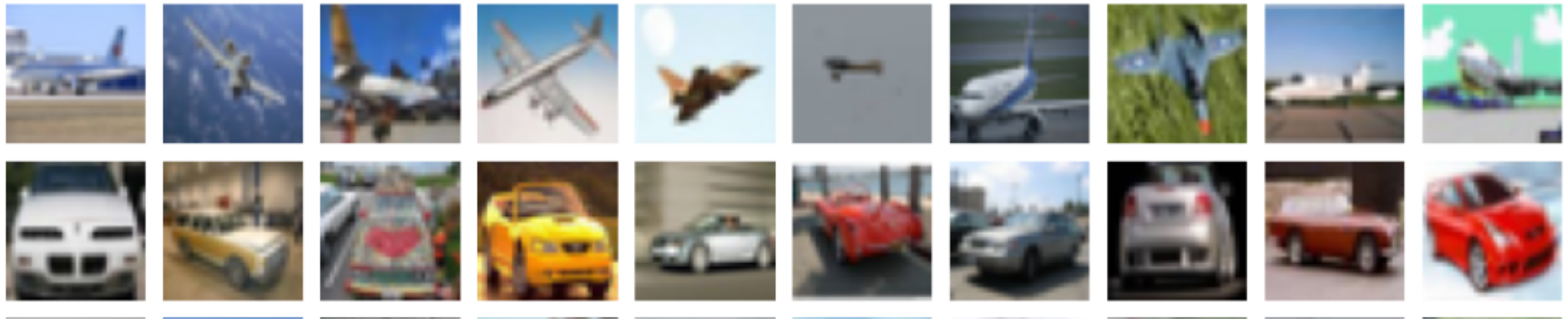
$$= -\log\big(y_i \cdot \sigma(f(\mathbf{x}, \mathbf{w})) + (1 - y_i) \cdot \sigma(f(\mathbf{x}, \mathbf{w}))\big)$$





$y = 0$

$y = 1$

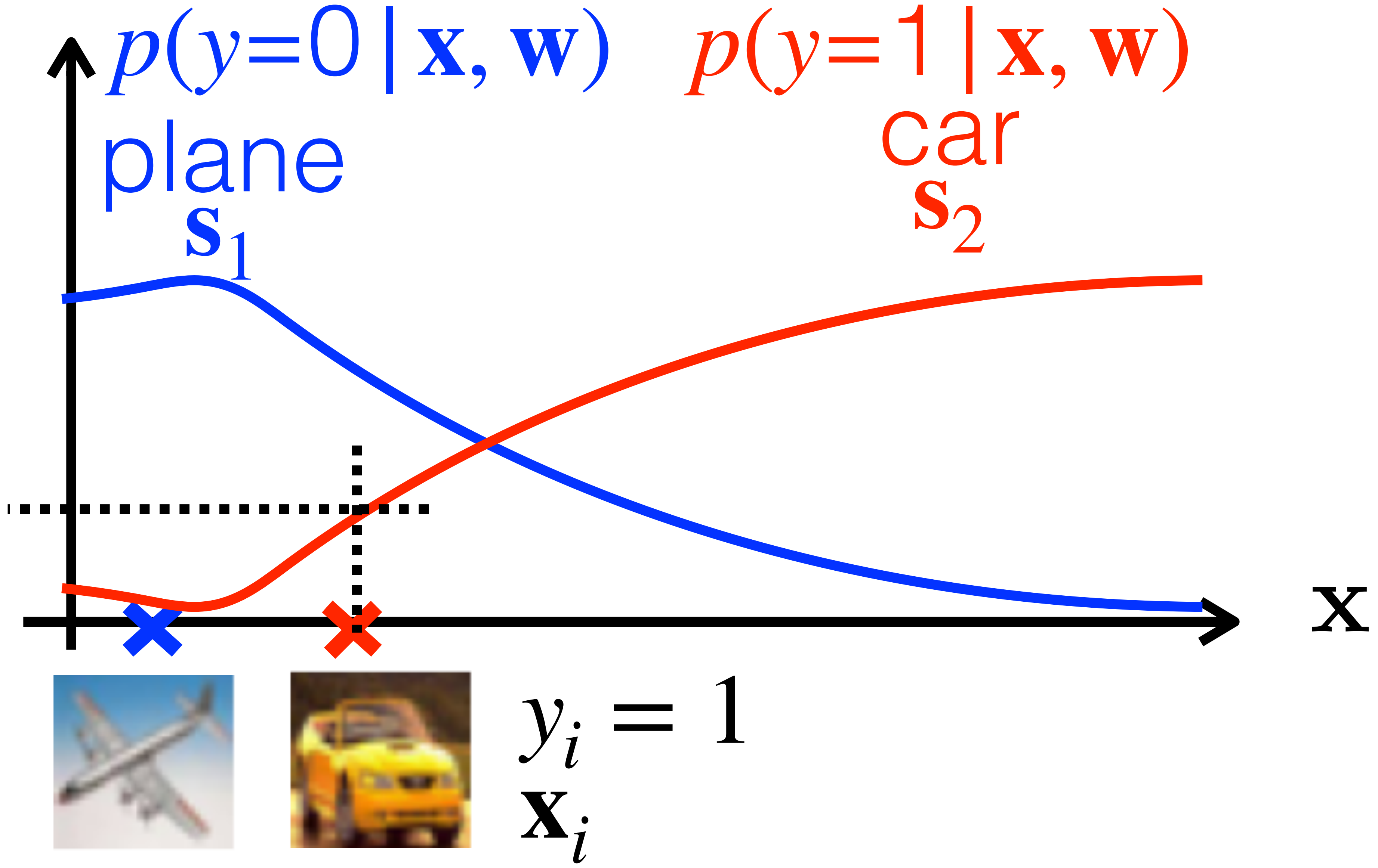


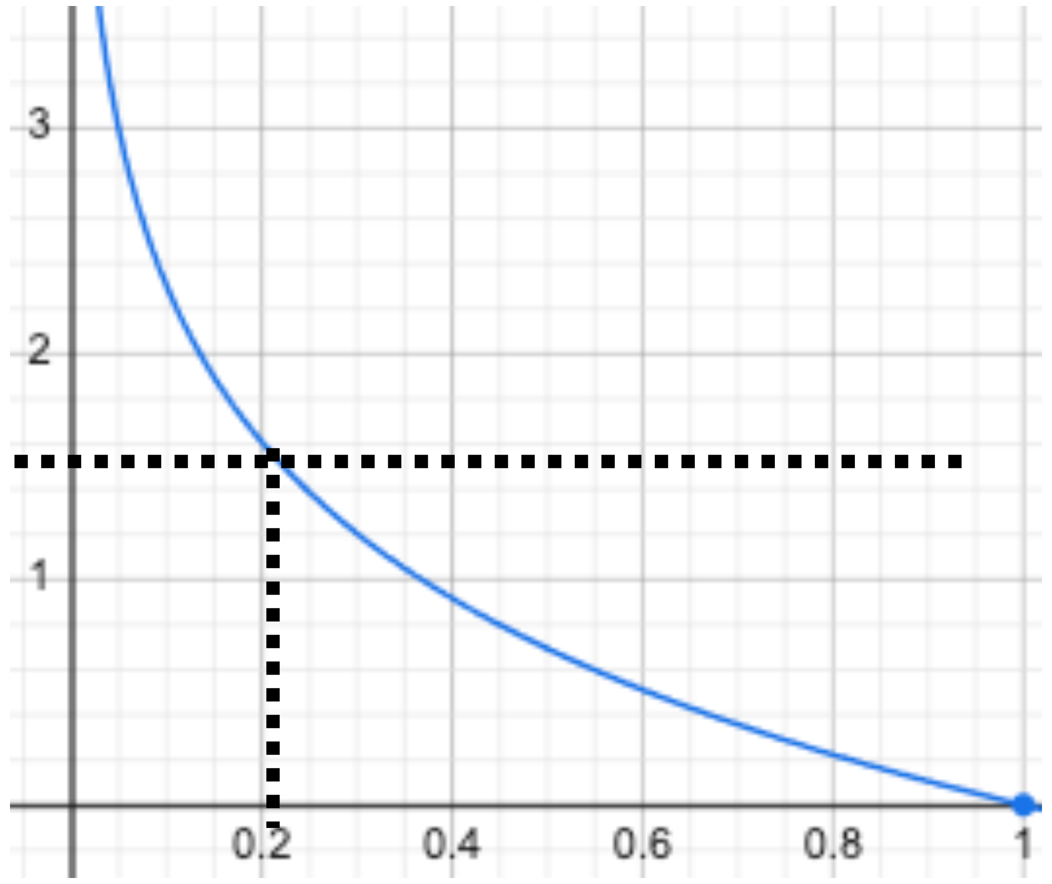
$$p(y\,|\,\mathbf{x},\mathbf{w}) = \begin{cases} \sigma(f(\mathbf{x},\mathbf{w})) & y = 1 \\ 1 - \sigma(f(\mathbf{x},\mathbf{w})) & y = 0 \end{cases} = y_i \cdot \sigma(f(\mathbf{x},\mathbf{w})) + (1 - y_i) \cdot (1 - \sigma(f(\mathbf{x},\mathbf{w})))$$

Loss function:

$$\mathscr{L}(\mathbf{w}) = \; -\log p(y_i\,|\,\mathbf{x}_i,\mathbf{w})$$

$$= -\log\big(y_i \cdot \sigma(f(\mathbf{x},\mathbf{w})) + (1 - y_i) \cdot \sigma(f(\mathbf{x},\mathbf{w}))\big)$$



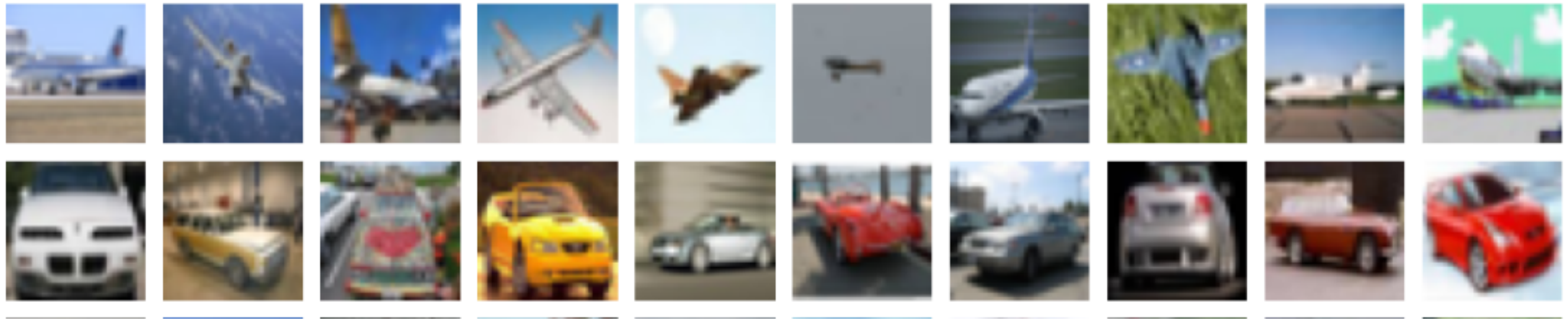


Learning:

$$\arg\min_{\mathbf{w}} \mathscr{L}(\mathbf{w})$$

$y = 0$

$y = 1$

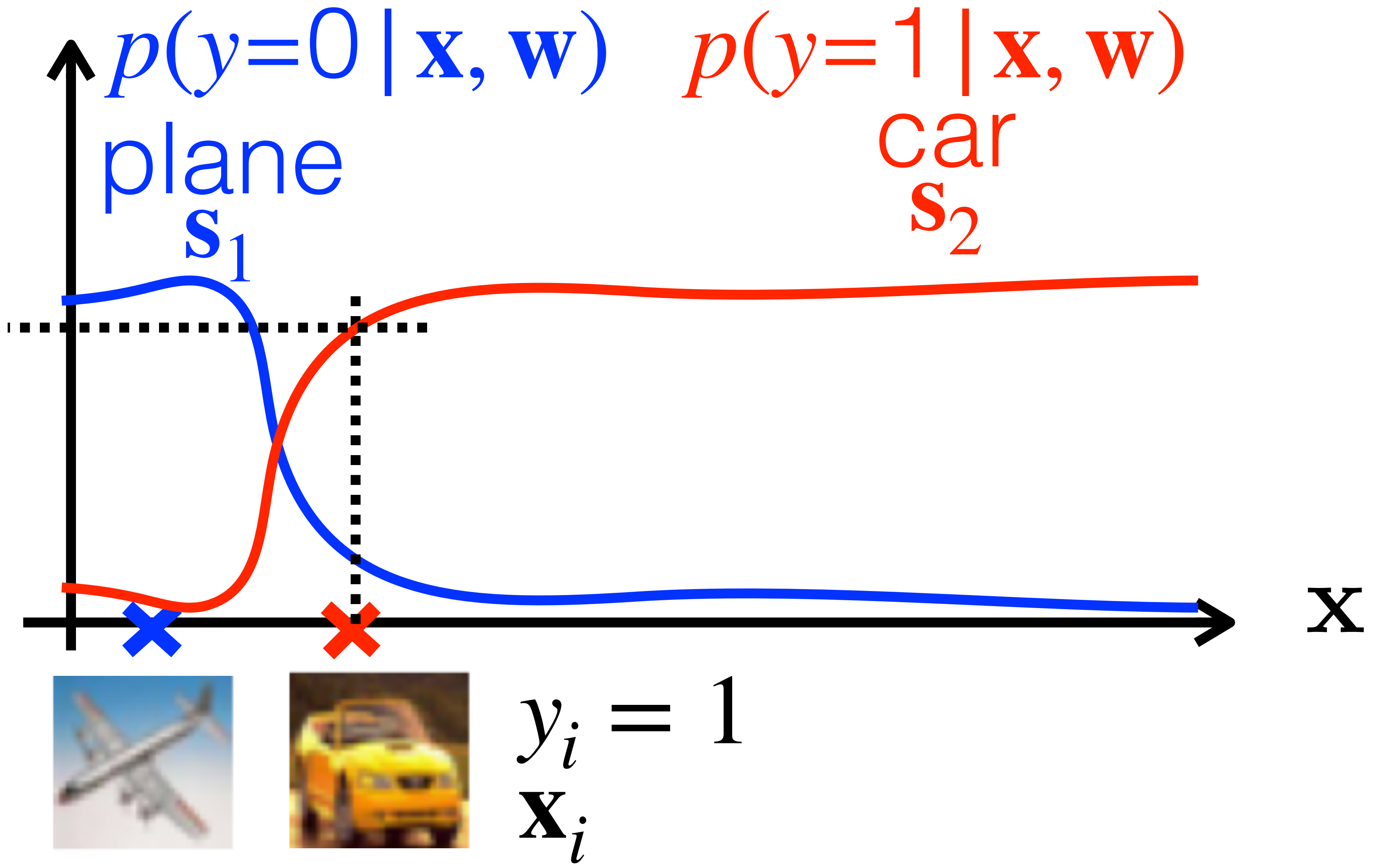


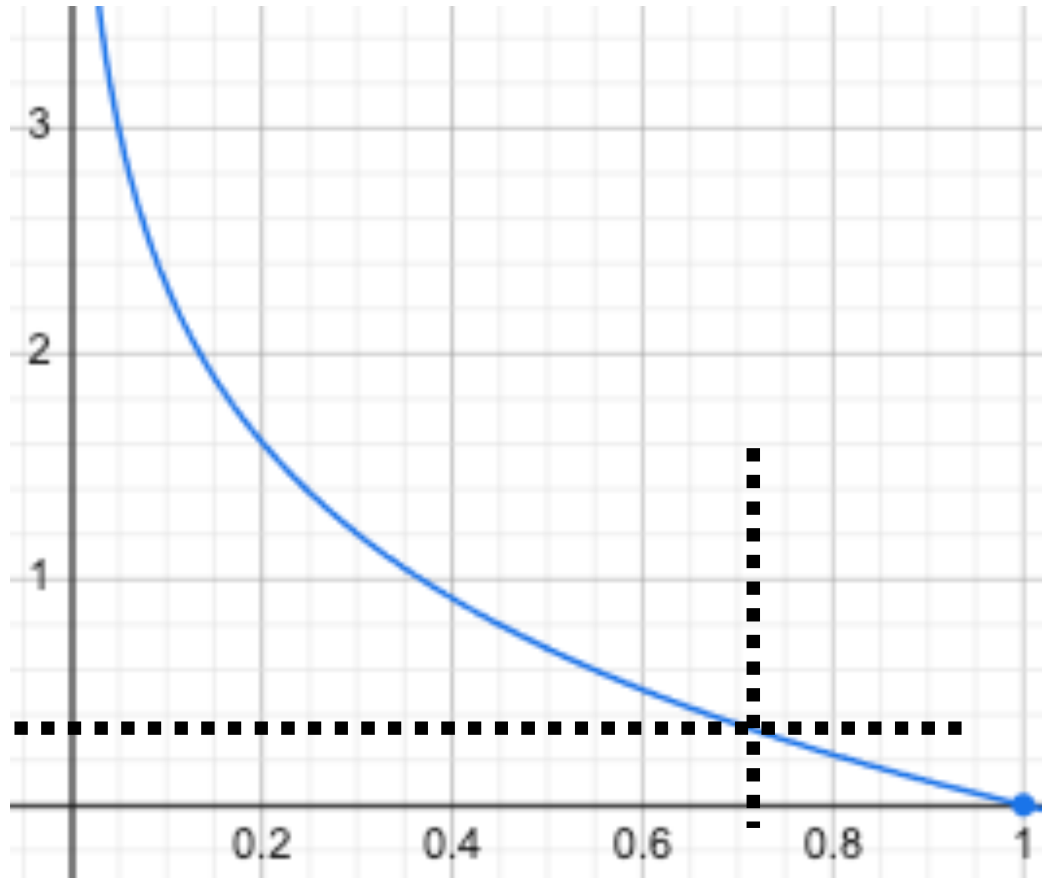
$$p(y\,|\,\mathbf{x},\mathbf{w}) = \begin{cases} \sigma(f(\mathbf{x},\mathbf{w})) & y = 1 \\ 1 - \sigma(f(\mathbf{x},\mathbf{w})) & y = 0 \end{cases} \quad = y_i \cdot \sigma(f(\mathbf{x},\mathbf{w})) + (1 - y_i)\cdot(1 - \sigma(f(\mathbf{x},\mathbf{w})))$$

Loss function:

$$\mathscr{L}(\mathbf{w}) = \; -\log p(y_i\,|\,\mathbf{x}_i,\mathbf{w})$$

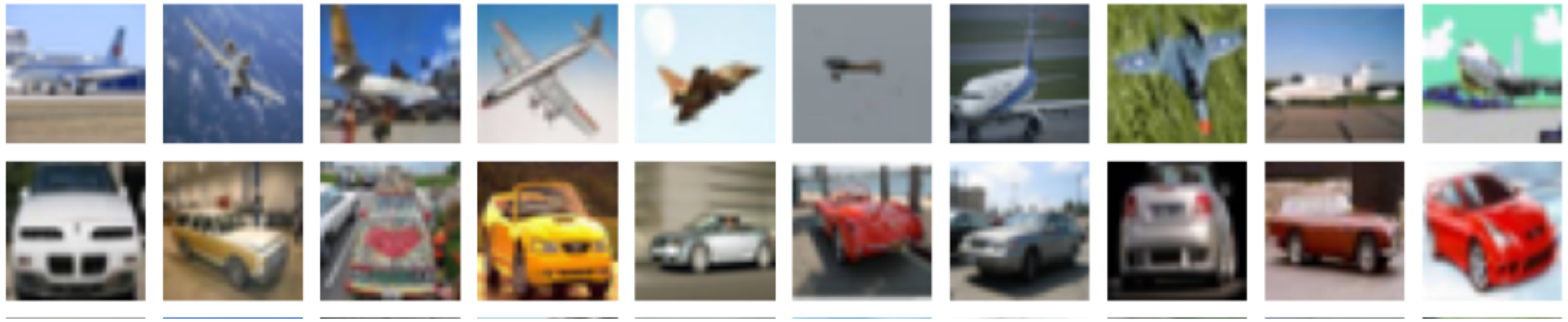
$$= -\log\big(y_i \cdot \sigma(f(\mathbf{x},\mathbf{w})) + (1 - y_i)\cdot\sigma(f(\mathbf{x},\mathbf{w}))\big)$$





Learning:

$$\arg\min_{\mathbf{W}} \mathscr{L}(\mathbf{W})$$

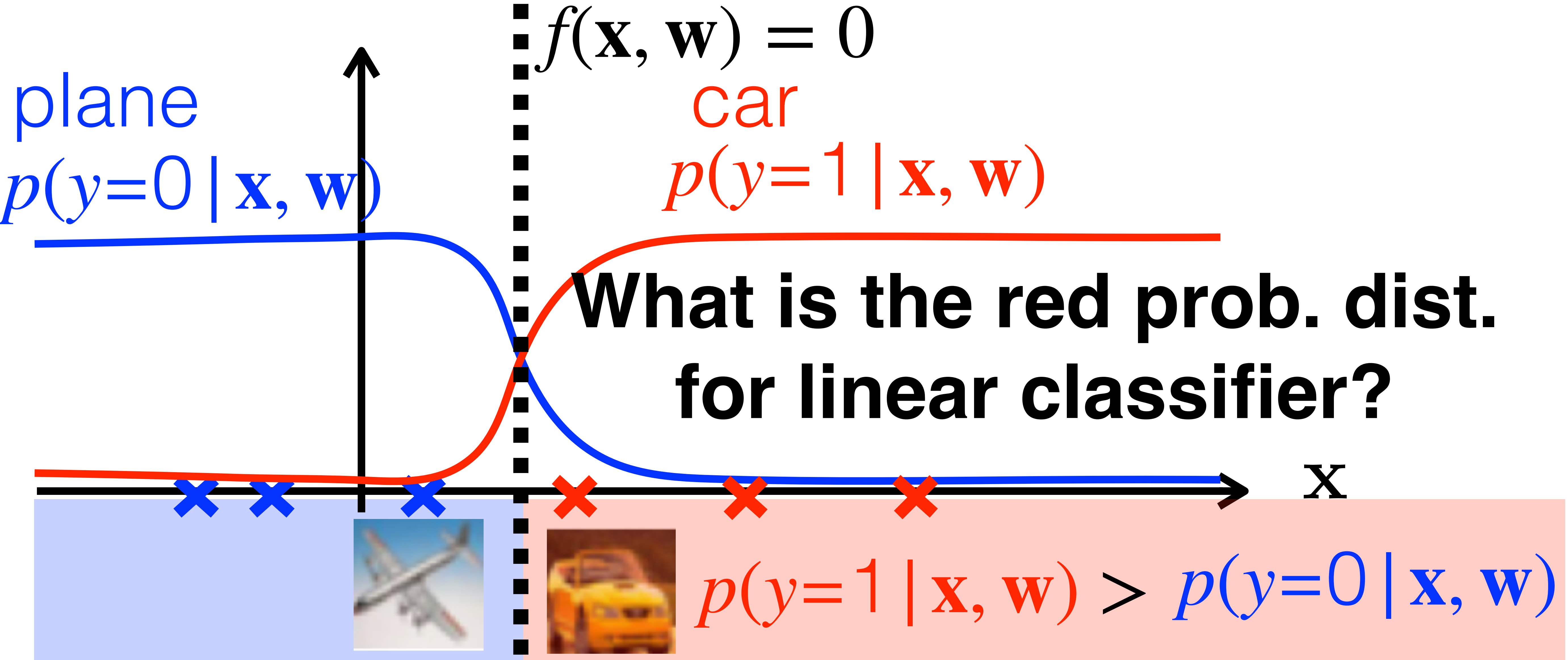


$$p(y \,|\, \mathbf{x}, \mathbf{w}) = \begin{cases} \sigma(f(\mathbf{x}, \mathbf{w})) & y = 1 \\ 1 - \sigma(f(\mathbf{x}, \mathbf{w})) & y = 0 \end{cases} = y_i \cdot \sigma(f(\mathbf{x}, \mathbf{w})) + (1 - y_i) \cdot (1 - \sigma(f(\mathbf{x}, \mathbf{w})))$$

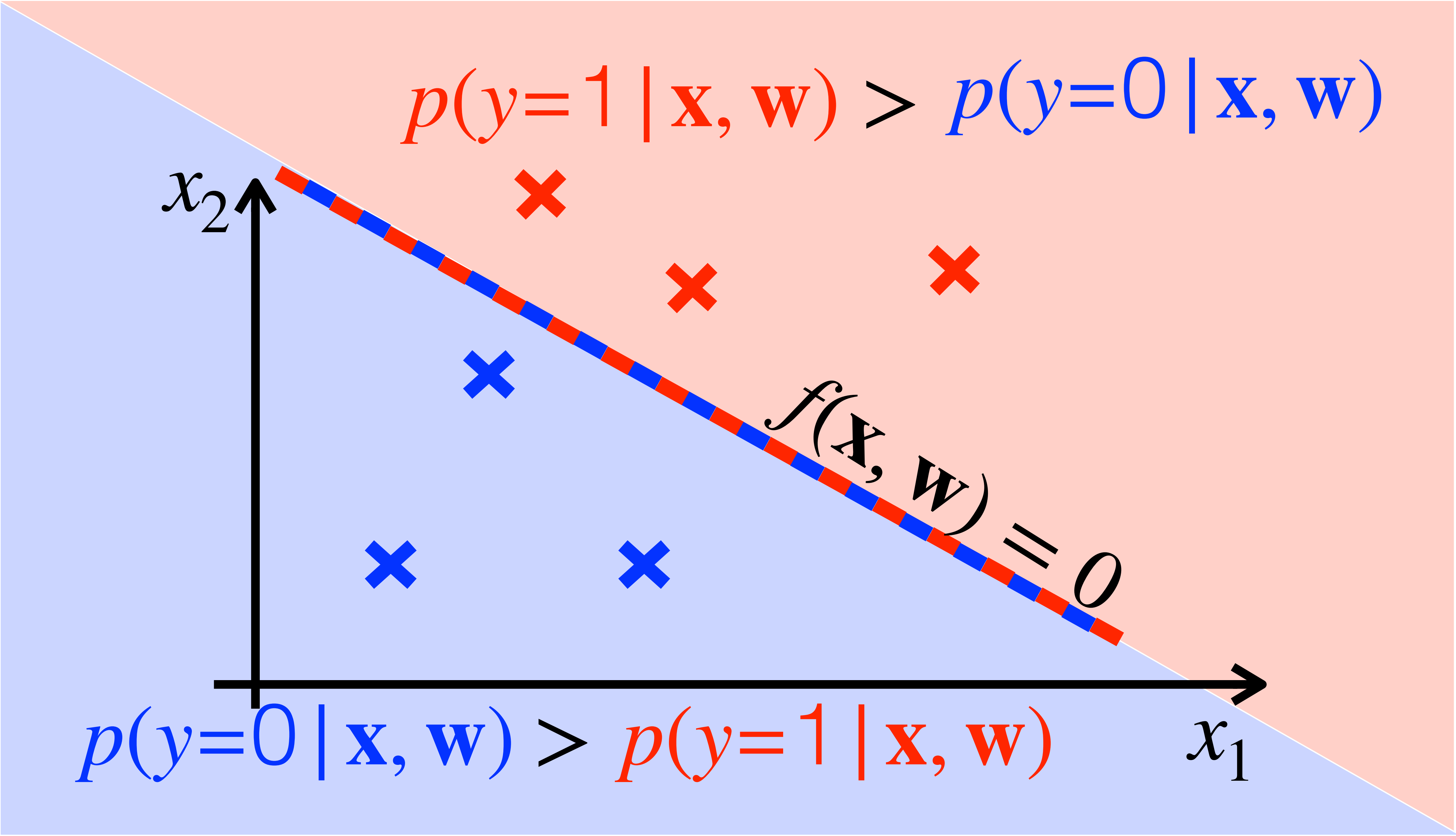
Draw training data that cannot be separated by linear classifier????

1D linear classifier

What is this threshold for linear class.?

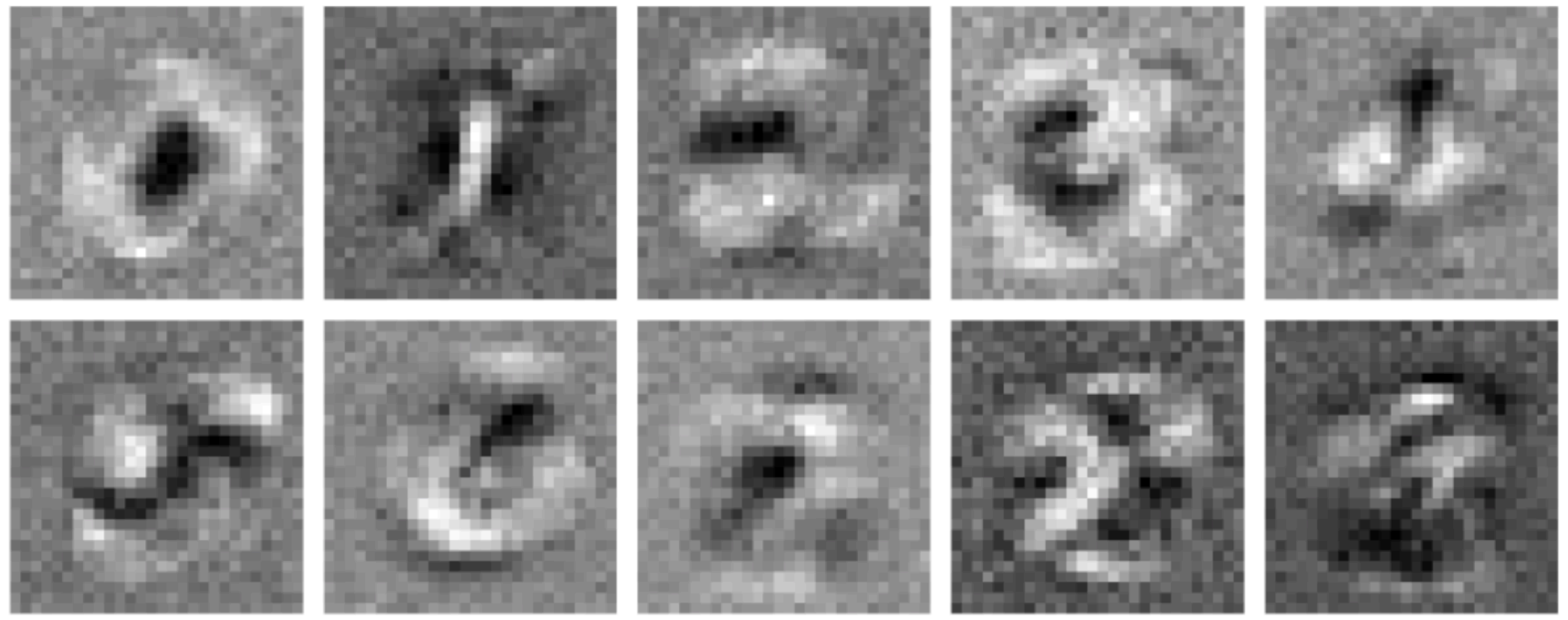


2D linear classifier



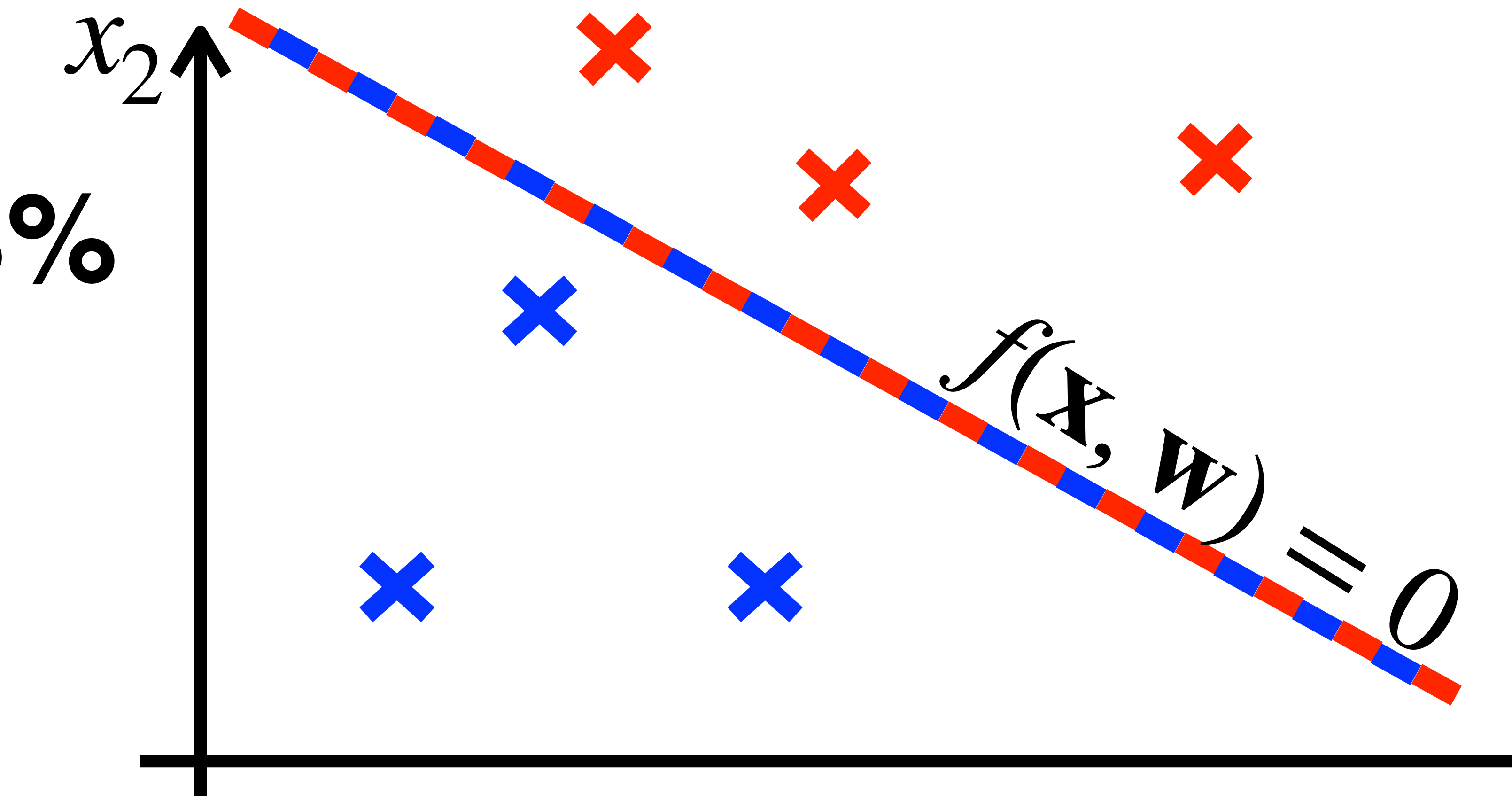
Label Images Learned weights Error 2D linear classifier

“0”
“1”
“2”
“3”
“4”
“5”
“6”
“7”
“8”
“9”



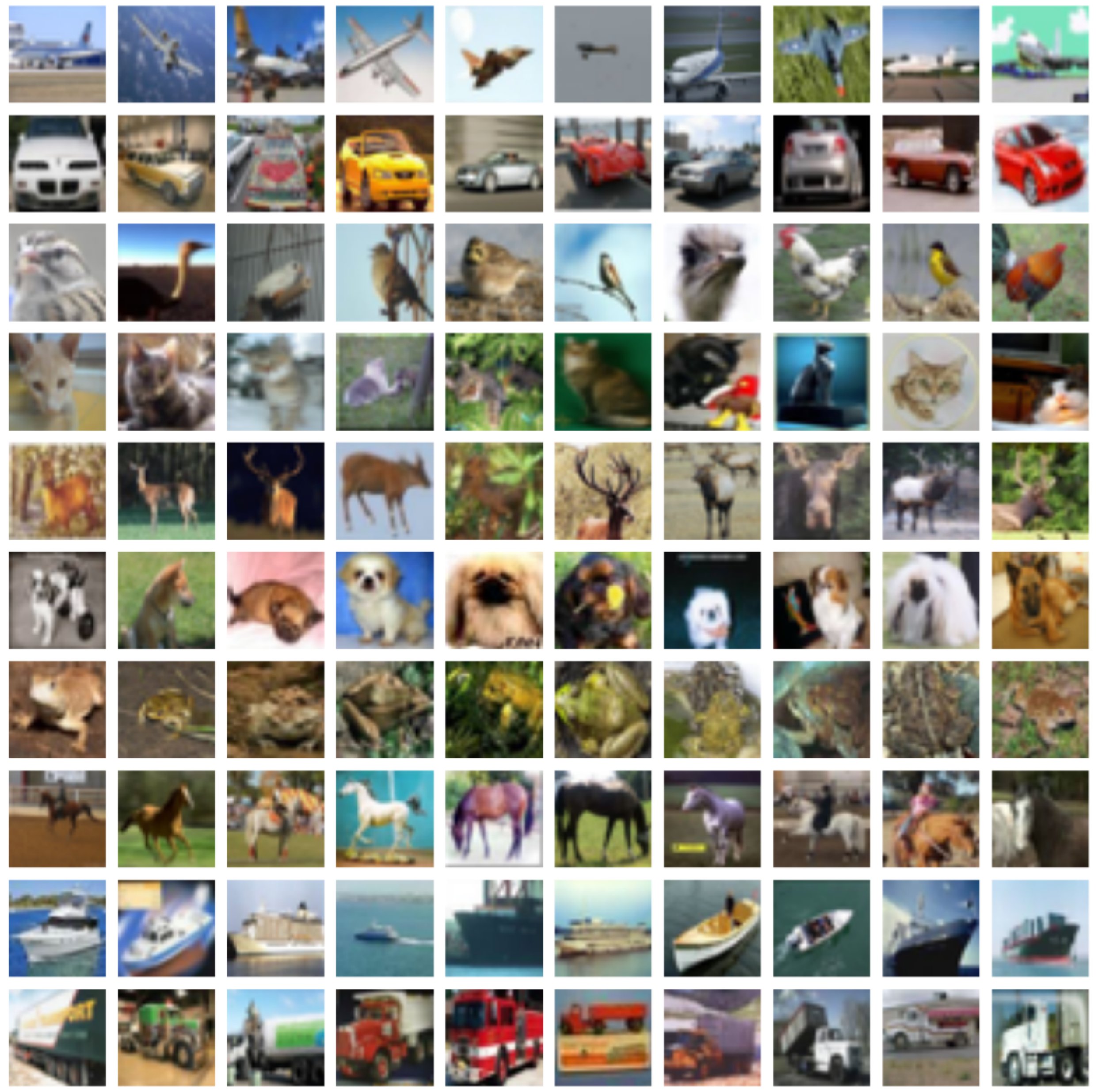
Why is it simple?

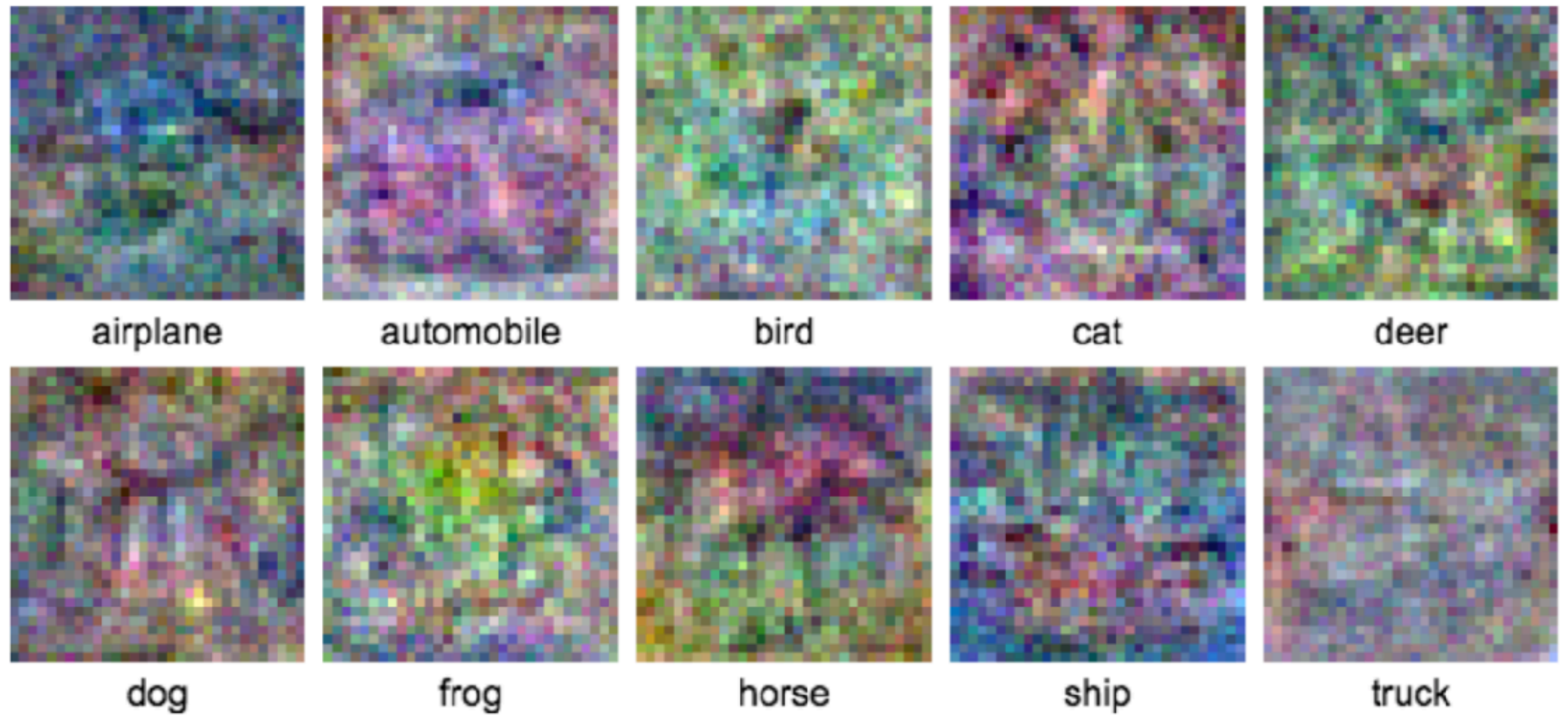
8%



2D linear classifier

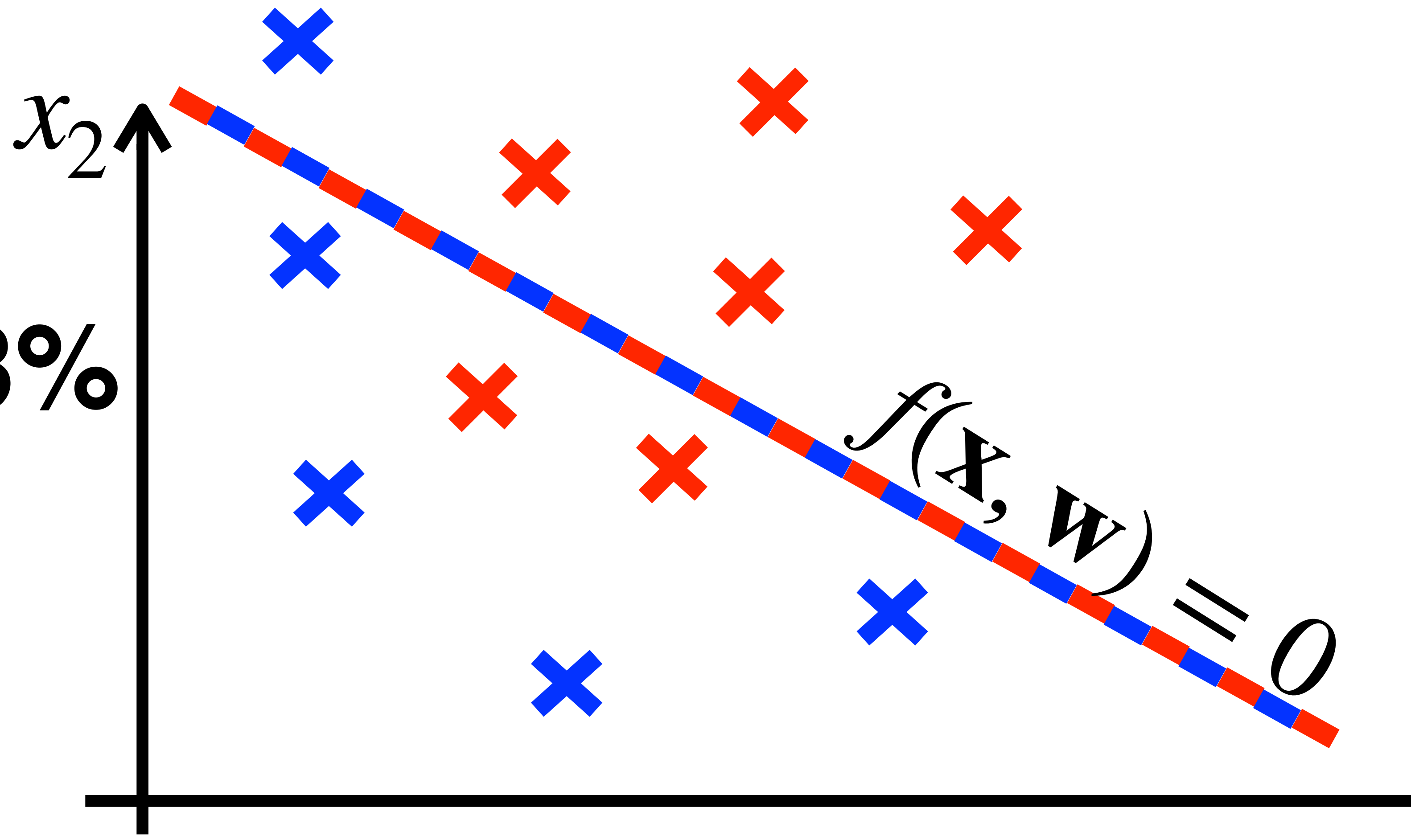
airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck





Why is it hard?

63%



Equivalence of common binary losses

Binary cross-entropy loss:

$$p(y\,|\,\mathbf{x},\mathbf{w}) = \begin{cases} \sigma(f(\mathbf{x},\mathbf{w})) & y = 1 \\ 1 - \sigma(f(\mathbf{x},\mathbf{w})) & y = 0 \end{cases} = y_i \cdot \sigma(f(\mathbf{x},\mathbf{w})) + (1 - y_i) \cdot \sigma(f(\mathbf{x},\mathbf{w}))$$

$$\mathscr{L}(\mathbf{W}) = \sum_i -\log p(y_i\,|\,\mathbf{x}_i,\mathbf{w}) = \sum_i -\log\Big(y_i \cdot \sigma(f(\mathbf{x}_i,\mathbf{w})) + (1 - y_i) \cdot \sigma(f(\mathbf{x}_i,\mathbf{w}))\Big)$$

```
loss = y*(-torch.log(torch.sigmoid(w@x))
       + (1-y)*(-torch.log(1-torch.sigmoid(w@x))
```

Logistic loss:

$$p(y\,|\,\mathbf{x},\mathbf{w}) = \begin{cases} \sigma(f(\mathbf{x},\mathbf{w})) & y = 1 \\ 1 - \sigma(f(\mathbf{x},\mathbf{w})) & y = -1 \end{cases} = \begin{cases} \sigma(f(\mathbf{x},\mathbf{w})) & y = 1 \\ \sigma(-f(\mathbf{x},\mathbf{w})) & y = -1 \end{cases} = \sigma\big(y_i f(\mathbf{x}_i,\mathbf{w})\big)$$

$$\mathscr{L}(\mathbf{W}) = \sum_i -\log p(y_i\,|\,\mathbf{x}_i,\mathbf{w}) = \sum_i -\log\Big(\sigma(y_i f(\mathbf{x}_i,\mathbf{w}))\Big) = \sum_i \log\Big(1 + \exp(-y_i f(\mathbf{x}_i,\mathbf{w}))\Big)$$

```
loss = -torch.log(torch.sigmoid(y*(w@x)))
loss = torch.log(1+torch.exp(-y*(w@x)))
```

Equivalence of common binary losses

Softmax => Cross-entropy loss:

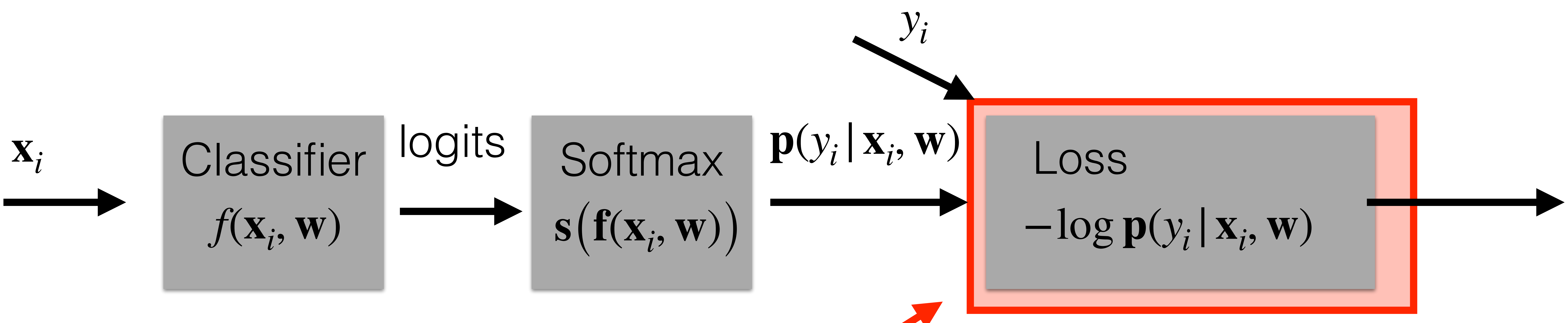
$$\mathbf{p}(y \mid \mathbf{x}, \mathbf{W}) = \begin{bmatrix} \exp(f(\mathbf{x}, \mathbf{w}_1)) \\ \exp(f(\mathbf{x}, \mathbf{w}_2)) \end{bmatrix} \;/\; \sum_k \exp(f\big(\mathbf{x}, \mathbf{w}_k\big) = \mathbf{s}(\mathbf{f}\big(\mathbf{x}, \mathbf{W})\big)$$

$$\mathbf{p}(y \mid \mathbf{x}, \mathbf{w}) = \begin{bmatrix} \exp(f(\mathbf{x}, \mathbf{w})/2) \\ \exp(-f(\mathbf{x}, \mathbf{w})/2) \end{bmatrix} \;/\; \Big(\exp\big(f(\mathbf{x}, \mathbf{w})/2\big) + \exp(-f(\mathbf{x}, \mathbf{w})/2)\big)\Big)$$

$$= \begin{bmatrix} \dfrac{1}{1 + \frac{\exp(-f(\mathbf{x}, \mathbf{w})/2)}{\exp(f(\mathbf{x}, \mathbf{w})/2)}} \\ 1 - \dfrac{1}{1 + \frac{\exp(-f(\mathbf{x}, \mathbf{w})/2)}{\exp(f(\mathbf{x}, \mathbf{w})/2)}} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{1 + \exp(-f(\mathbf{x}, \mathbf{w}))} \\ 1 - \dfrac{1}{1 + \exp(-f(\mathbf{x}, \mathbf{w}))} \end{bmatrix} = \begin{bmatrix} \sigma\big(f(\mathbf{x}, \mathbf{w})\big) \\ 1 - \sigma\big(f(\mathbf{x}, \mathbf{w})\big) \end{bmatrix}$$

$$p(y \mid \mathbf{x}, \mathbf{w}) = \begin{cases} \sigma(f(\mathbf{x}, \mathbf{w})) & y = 1 \\ 1 - \sigma(f(\mathbf{x}, \mathbf{w})) & y = 0 \end{cases} \quad = y_i \cdot \sigma(f(\mathbf{x}, \mathbf{w})) + (1 - y_i) \cdot \sigma(f(\mathbf{x}, \mathbf{w}))$$

Various implementation and names



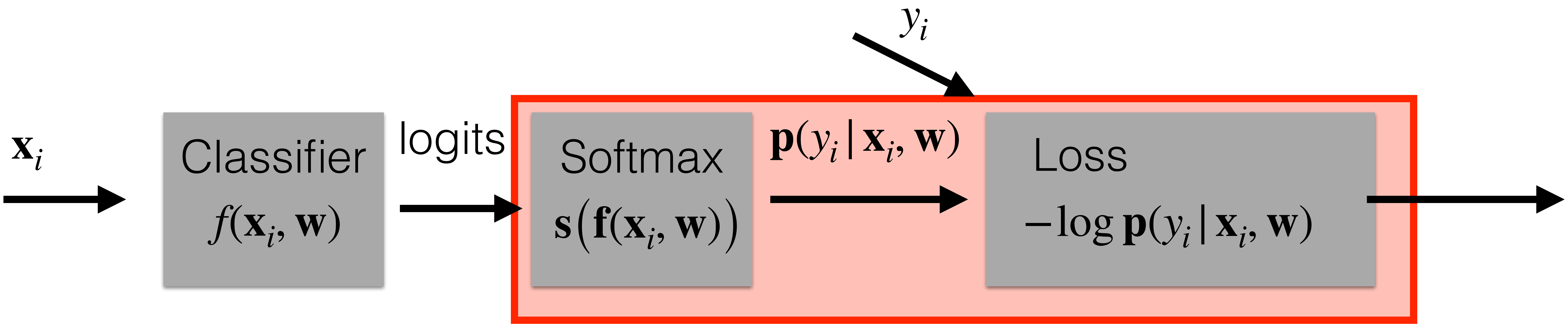
Input: probabilities

(Multinomial) Logistic loss,
Cross-entropy loss, …

$$\mathscr{L}(\mathbf{w}) = CE\left(y_i, \mathbf{p}(y_i \mid \mathbf{x}_i, \mathbf{w})\right)$$

Pytorch: BCELoss
TensorFlow: log_loss
Caffe: Multinomial Logistic Loss Layer

Various implementation and names



Input: probabilities

(Multinomial) Logistic loss,
Cross-entropy loss, …

$$\mathscr{L}(\mathbf{w}) = CE\big(y_i, \mathbf{p}(y_i \,|\, \mathbf{x}_i, \mathbf{w})\big)$$

Pytorch: BCELoss
TensorFlow: log_loss
Caffe: Multinomial Logistic Loss Layer

Input: logits

Softmax loss,
Categorical Cross-entropy loss, …

$$\mathscr{L}(\mathbf{w}) = CE\big(y_i, f(\mathbf{x}_i, \mathbf{w})\big)$$

Pytorch: CrossEntropyLoss
TensorFlow: softmax_cross_entropy
Caffe: SoftmaxWithLoss Layer

Competencies required for the test T1

- Classification loss for two-class and K-class classification problem.
- Equivalence of common binary classification losses
- Meaning of weights in linear classifier
- 2D visualization, decision boundary, features
- Drawback of linear classifier (too simple for some problems)