DeepRob

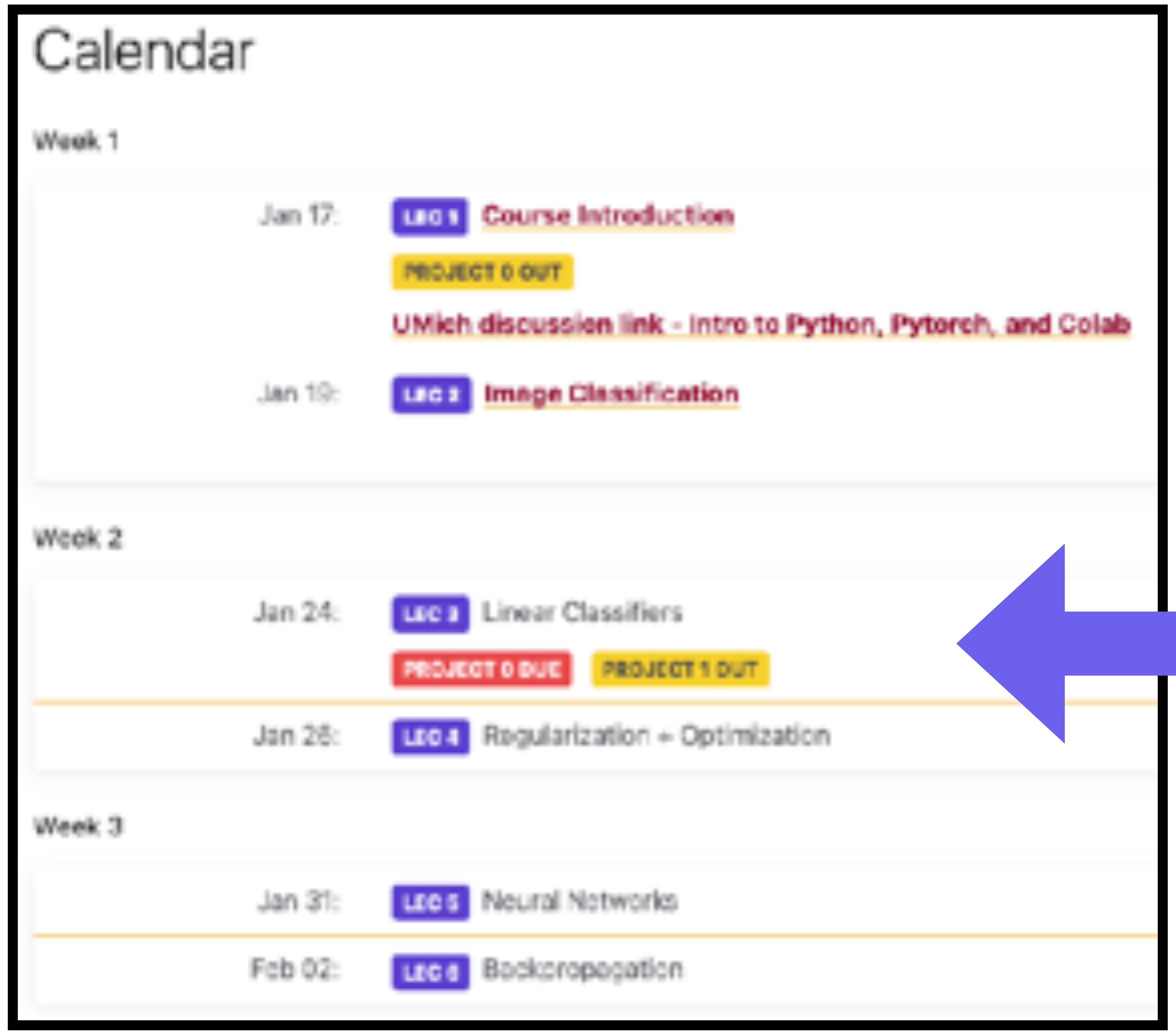
Lecture 3
Linear Classifiers
University of Michigan and University of Minnesota

Project 0

- Instructions and code available on the website
  - Here: https://rpm-lab.github.io/CSCI5980-Spr23-DeepRob/projects/project0/
- **Due tonight! January 24th, 11:59 PM CT**

Project 1

- Instructions and code will be available on the website today.
- Classification using K-Nearest Neighbors and Linear Models



We're here!

Gradescope Quizzes

- Quiz links will be published at 7am on the day of lecture.
  - This will start from next lecture on 01/26
- The quiz will close before that day's lecture time i.e. 2:30pm
- Time limit of 15 min once quiz is opened
- Covers material from previous lectures and graded projects

Recap: Image Classification—A Core Computer Vision Task

Input: image

Output: assign image to one of a fixed set of categories

Chocolate Pretzels

Granola Bar

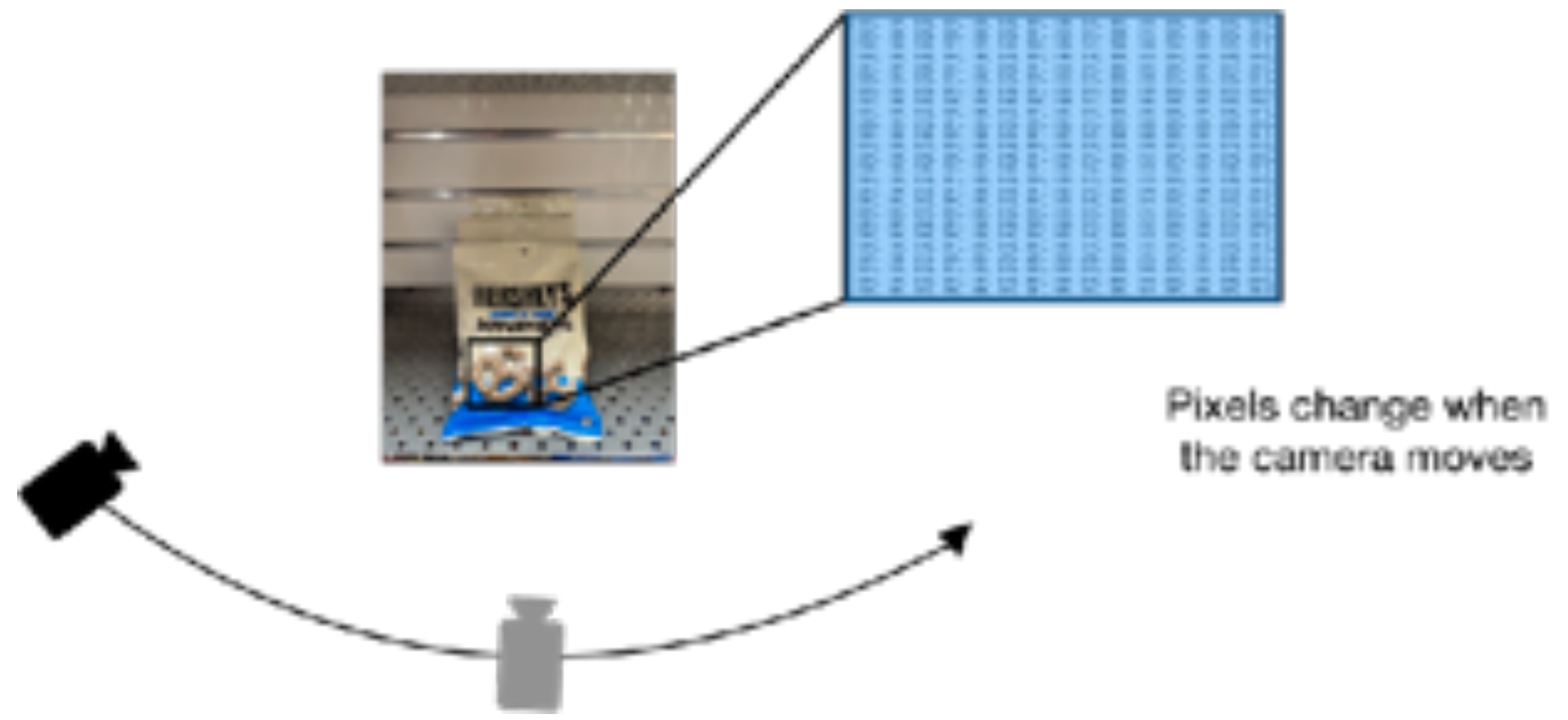
Potato Chips

Water Bottle

Popcorn

Recap: Image Classification Challenges

Viewpoint Variation & Semantic Gap



Illumination Changes

Milk Chocolate | White Chocolate | Cookies N' Creme | Peanut Butter | Ambiguous Category

Intraclass Variation

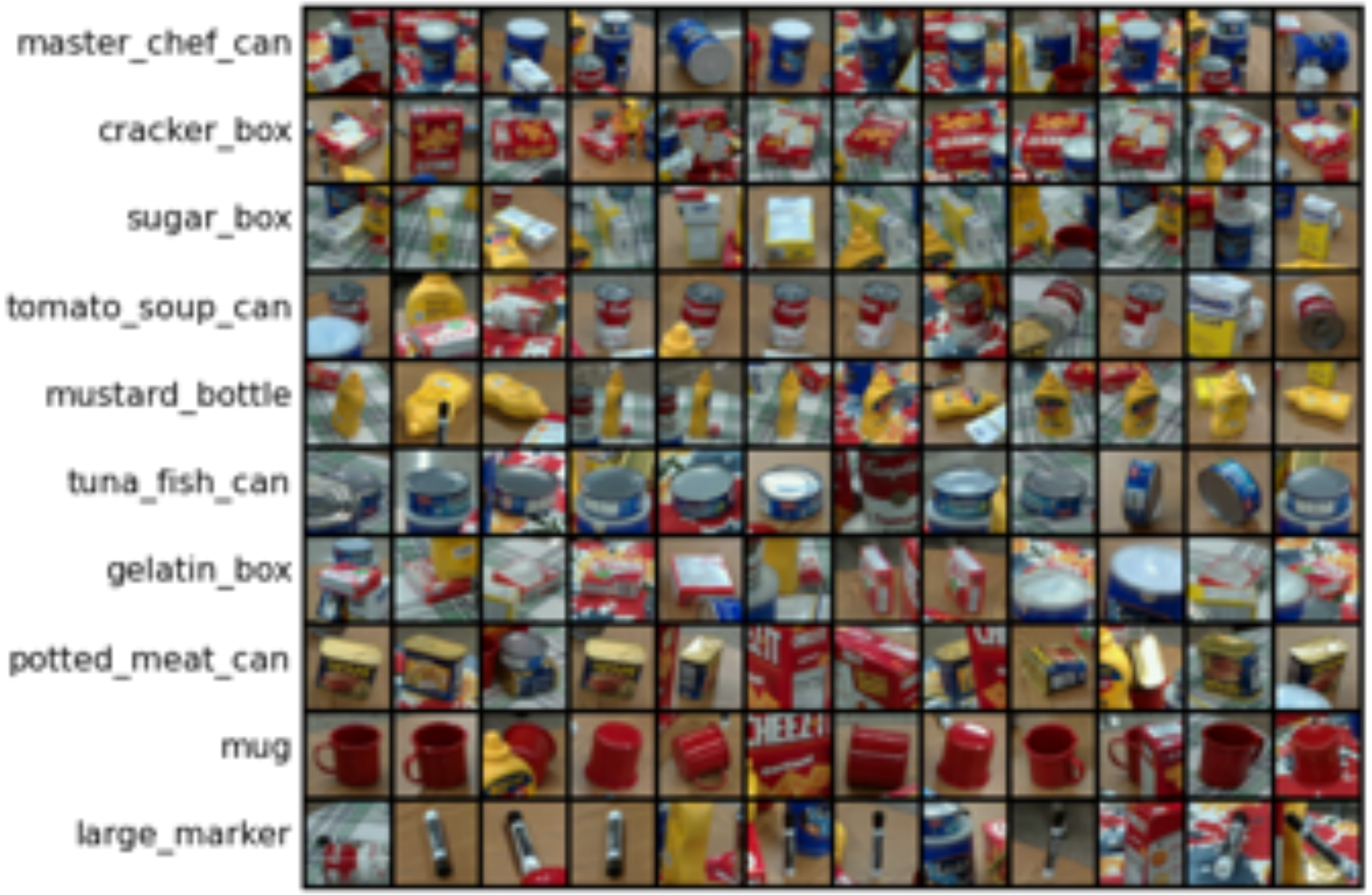
Recap: Machine Learning—Data-Driven Approach

1. Collect a dataset of images and labels
2. Use Machine Learning to train a classifier
3. Evaluate the classifier on new images

```
def train(images, labels):
  # Machine learning!
  return model
```

```
def predict(model, test_images):
  # Use model to predict labels
  return test_labels
```

Example training set



Linear Classifiers

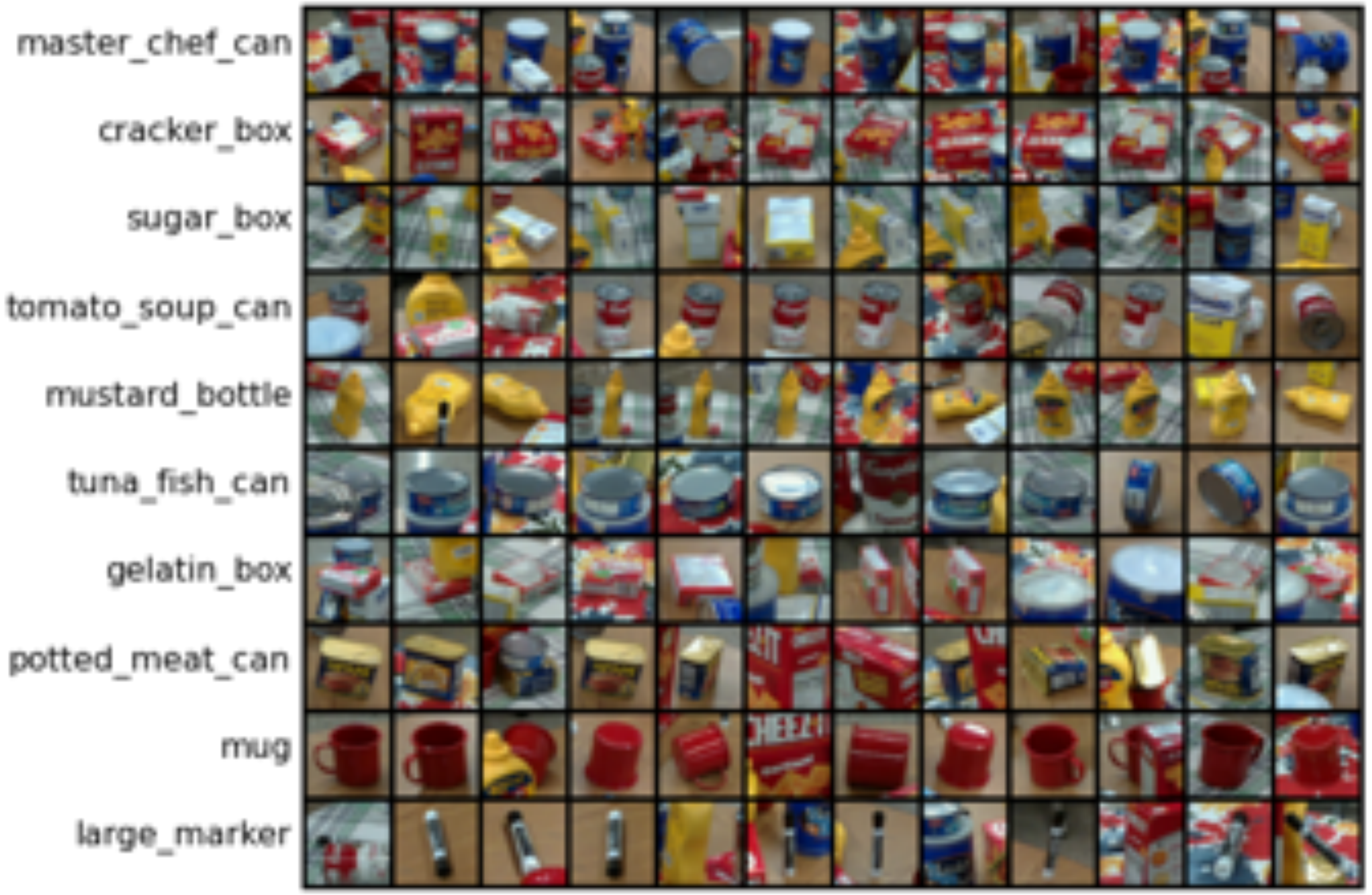
Building Block of Neural Networks

Linear
classifiers

This image is CC0 1.0 public domain

Recall PROPS

Progress **R**obot **O**bject **P**erception **S**amples **D**ataset



Chen et al., "ProgressLabeller: Visual Data Stream Annotation for Training Object-Centric 3D Perception", IROS, 2022.

10 classes
32x32 RGB images
50k training images (5k per class)
10k test images (1k per class)

A video link will be posted to the website today discussing about PROPS dataset that you will use for P1

Parametric Approach

Image

Array of **32x32x3** numbers
(3072 numbers total)

f(**x**,**W**)

W

parameters
or weights

10 numbers giving
class scores

Parametric Approach—Linear Classifier

$$f(x,W) = Wx$$

Image

Array of **32x32x3** numbers
(3072 numbers total)

$\rightarrow$ f(**x**,**W**) $\rightarrow$ **10** numbers giving class scores

W
parameters
or weights

Parametric Approach—Linear Classifier

(3072,)

$$\boxed{f(x,W)} = \boxed{W}\boxed{x}$$

(10,) (10, 3072)

Image

Array of **32x32x3** numbers
(3072 numbers total)

f(**x**,**W**)

10 numbers giving class scores

W

parameters
or weights

Parametric Approach—Linear Classifier

$$\boxed{f(x,W)} = \boxed{W}\boxed{x} + \boxed{b}$$

(3072,)

(10,)

(10,) (10, 3072)

Image

Array of **32x32x3** numbers
(3072 numbers total)

→ f(**x**,**W**) →

10 numbers giving class scores

↑

W

parameters
or weights

Example for 2x2 Image, 3 classes (crackers/mug/sugar)

Stretch pixels into column

$f(x,W) = Wx + b$

Input image
(2, 2)

56
231
24
2

(4,)

Example for 2x2 Image, 3 classes (crackers/mug/sugar)

Stretch pixels into column

$f(x,W) = Wx + b$

Input image
(2, 2)

$$\underbrace{\begin{bmatrix} 0.2 & -0.5 & 0.1 & 2.0 \\ 1.5 & 1.3 & 2.1 & 0.0 \\ 0 & 0.25 & 0.2 & -0.3 \end{bmatrix}}_{W\ (3,\ 4)} \underbrace{\begin{bmatrix} 56 \\ 231 \\ 24 \\ 2 \end{bmatrix}}_{(4,)} + \underbrace{\begin{bmatrix} 1.1 \\ 3.2 \\ -1.2 \end{bmatrix}}_{b\ (3,)} = \underbrace{\begin{bmatrix} -96.8 \\ 437.9 \\ 61.95 \end{bmatrix}}_{(3,)}$$

Linear Classifier—Algebraic Viewpoint

Stretch pixels into column

$f(x,W) = Wx + b$

Input image
(2, 2)

$$\underbrace{\begin{bmatrix} 0.2 & -0.5 & 0.1 & 2.0 \\ 1.5 & 1.3 & 2.1 & 0.0 \\ 0 & 0.25 & 0.2 & -0.3 \end{bmatrix}}_{W\ (3,\,4)} \underbrace{\begin{bmatrix} 56 \\ 231 \\ 24 \\ 2 \end{bmatrix}}_{(4,)} + \underbrace{\begin{bmatrix} 1.1 \\ 3.2 \\ -1.2 \end{bmatrix}}_{b\ (3,)} = \underbrace{\begin{bmatrix} -96.8 \\ 437.9 \\ 61.95 \end{bmatrix}}_{(3,)}$$

Linear Classifier—Bias Trick

Stretch pixels into column

Input image
(2, 2)

$$
\underbrace{\left[\begin{array}{cccc|c} 0.2 & -0.5 & 0.1 & 2.0 & 1.1 \\ 1.5 & 1.3 & 2.1 & 0.0 & 3.2 \\ 0 & 0.25 & 0.2 & -0.3 & -1.2 \end{array}\right]}_{W \ (3,\ 5)}
\underbrace{\begin{bmatrix} 56 \\ 231 \\ 24 \\ 2 \\ 1 \end{bmatrix}}_{(5,)}
=
\underbrace{\begin{bmatrix} -96.8 \\ 437.9 \\ 61.95 \end{bmatrix}}_{(3,)}
$$

Add extra one to data vector; bias is absorbed into last column of weight matrix

Linear Classifier—Predictions are Linear

$f(x, W) = Wx$ (ignore bias)

$f(cx, W) = W(cx) = c * f(x, W)$

Linear Classifier—Predictions are Linear

$$f(x, W) = Wx \quad \text{(ignore bias)}$$

$$f(cx, W) = W(cx) = c * f(x, W)$$

Image

→

Scores

-96.8
437.8
62.0

0.5 * Image

→

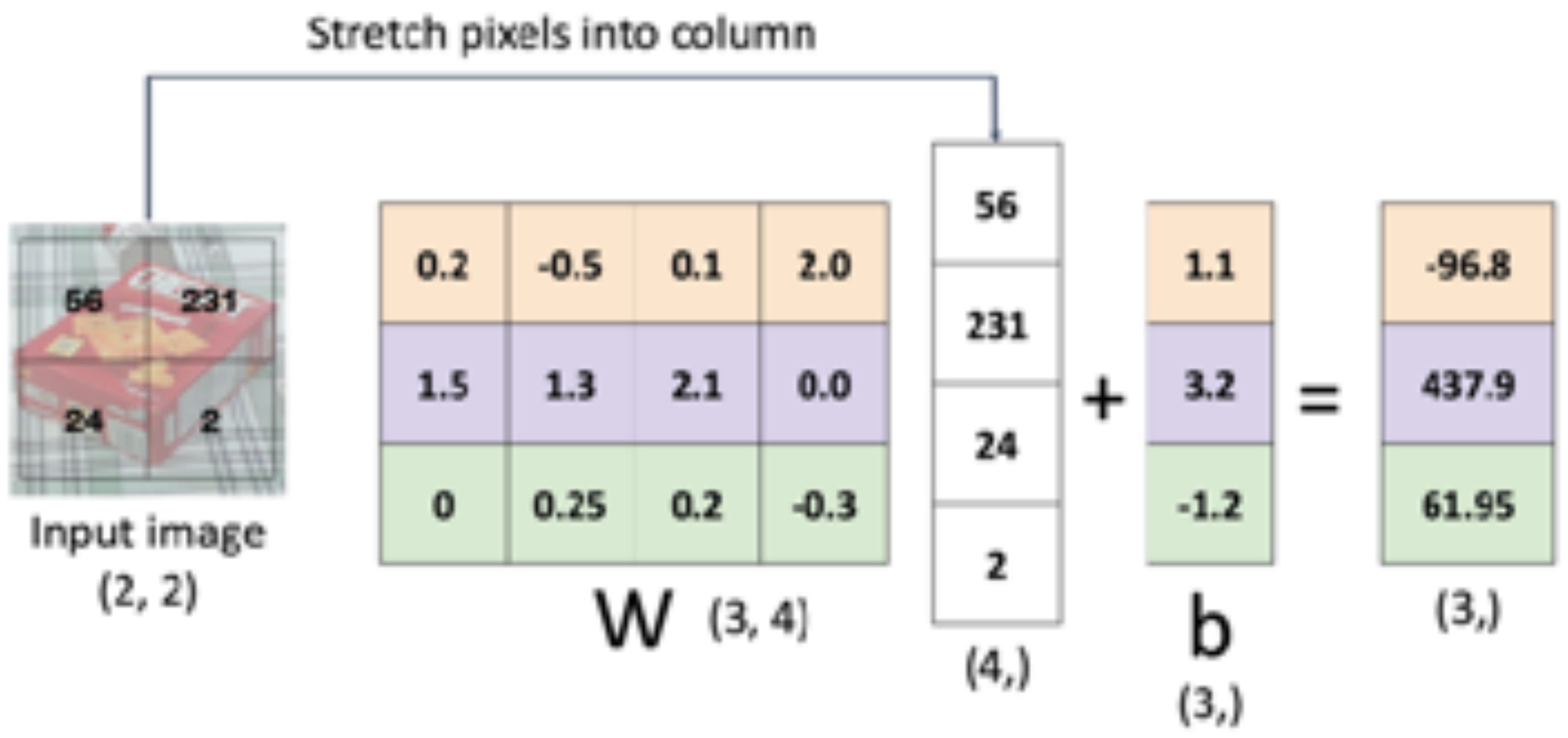
0.5 * Scores

-48.4
218.9
31.0

Interpreting a Linear Classifier

Algebraic Viewpoint

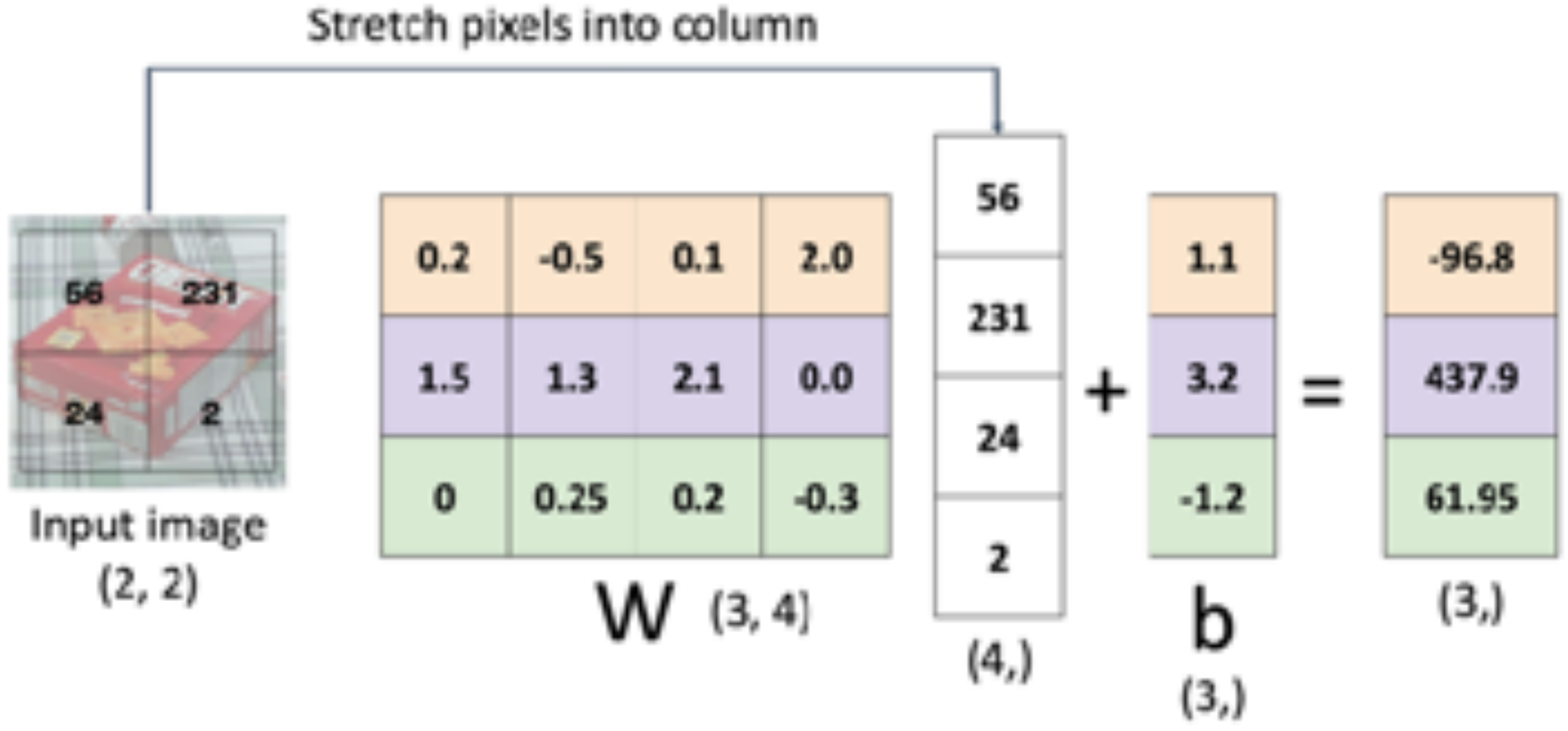
$$f(x,W) = Wx + b$$



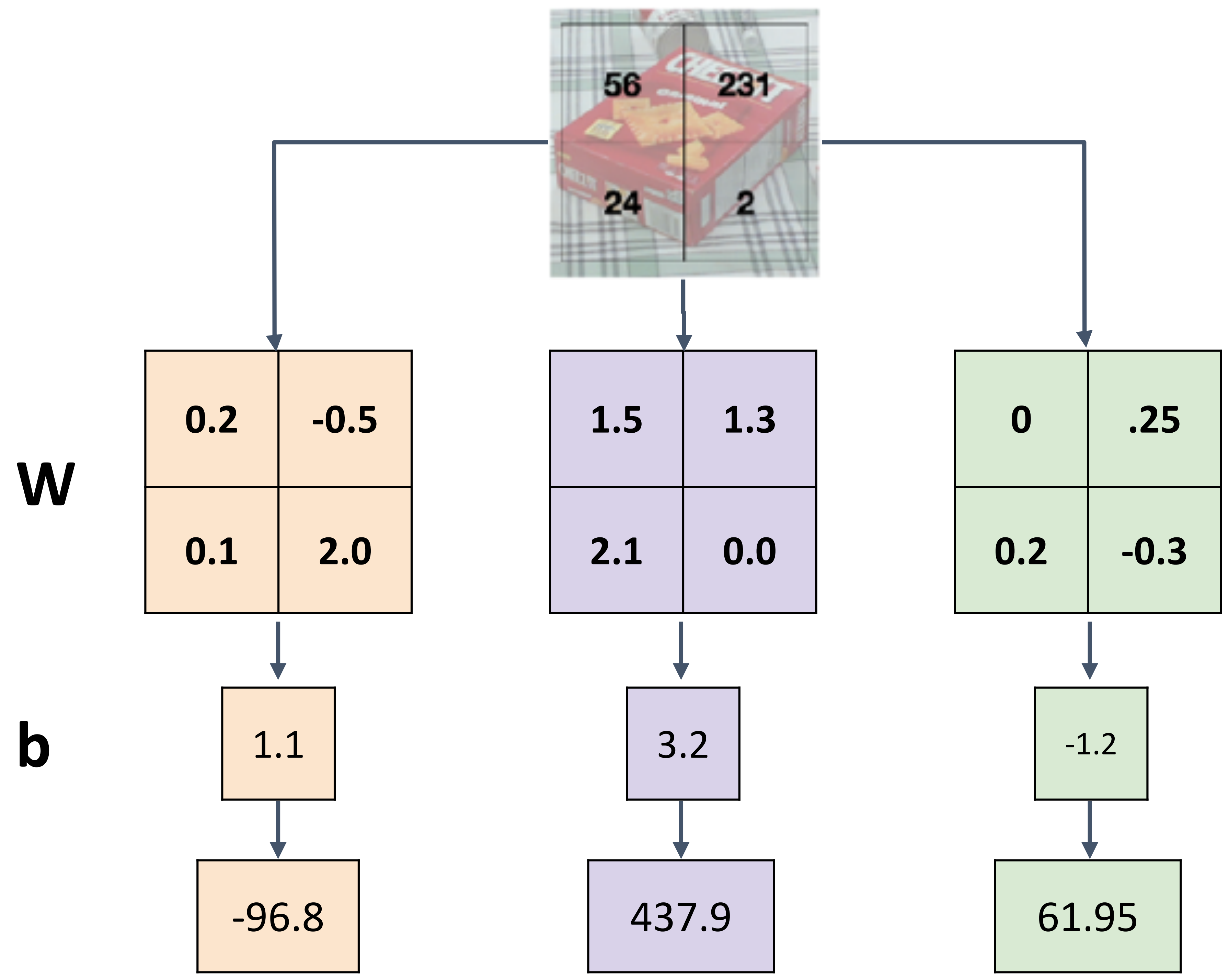
Interpreting a Linear Classifier

Algebraic Viewpoint

$$f(x,W) = Wx + b$$

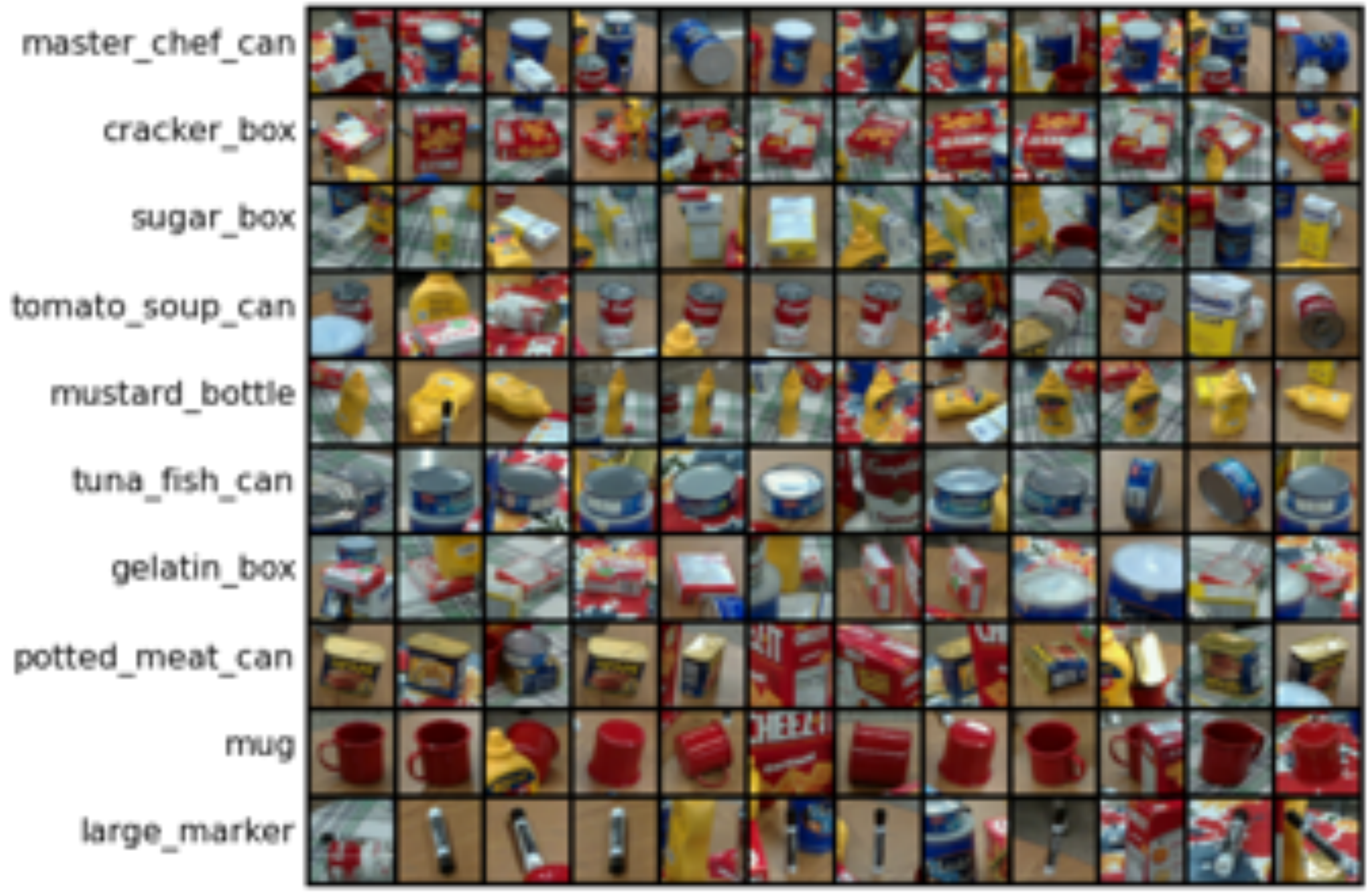


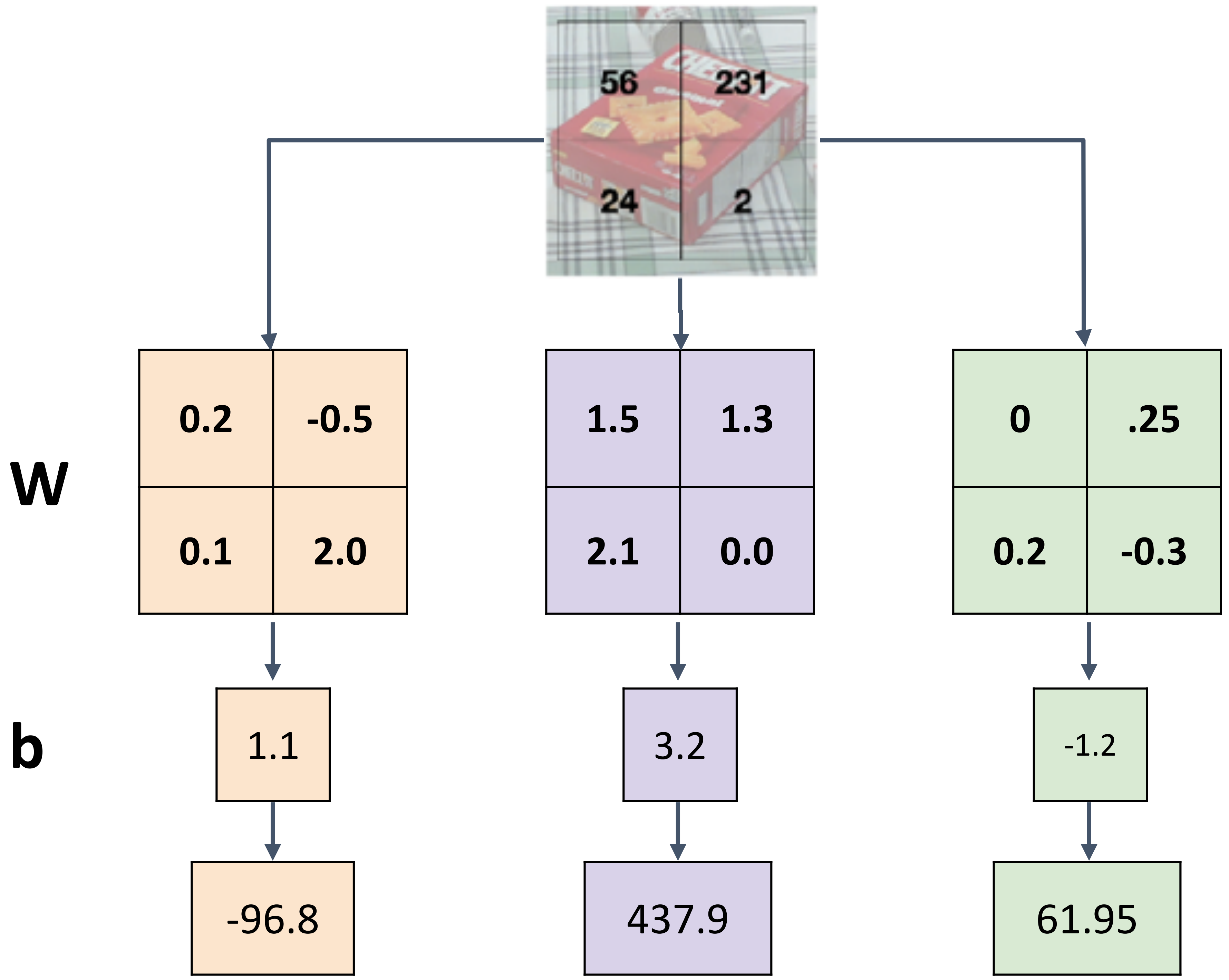
Instead of stretching pixels into columns, we can equivalently stretch rows of W into images!



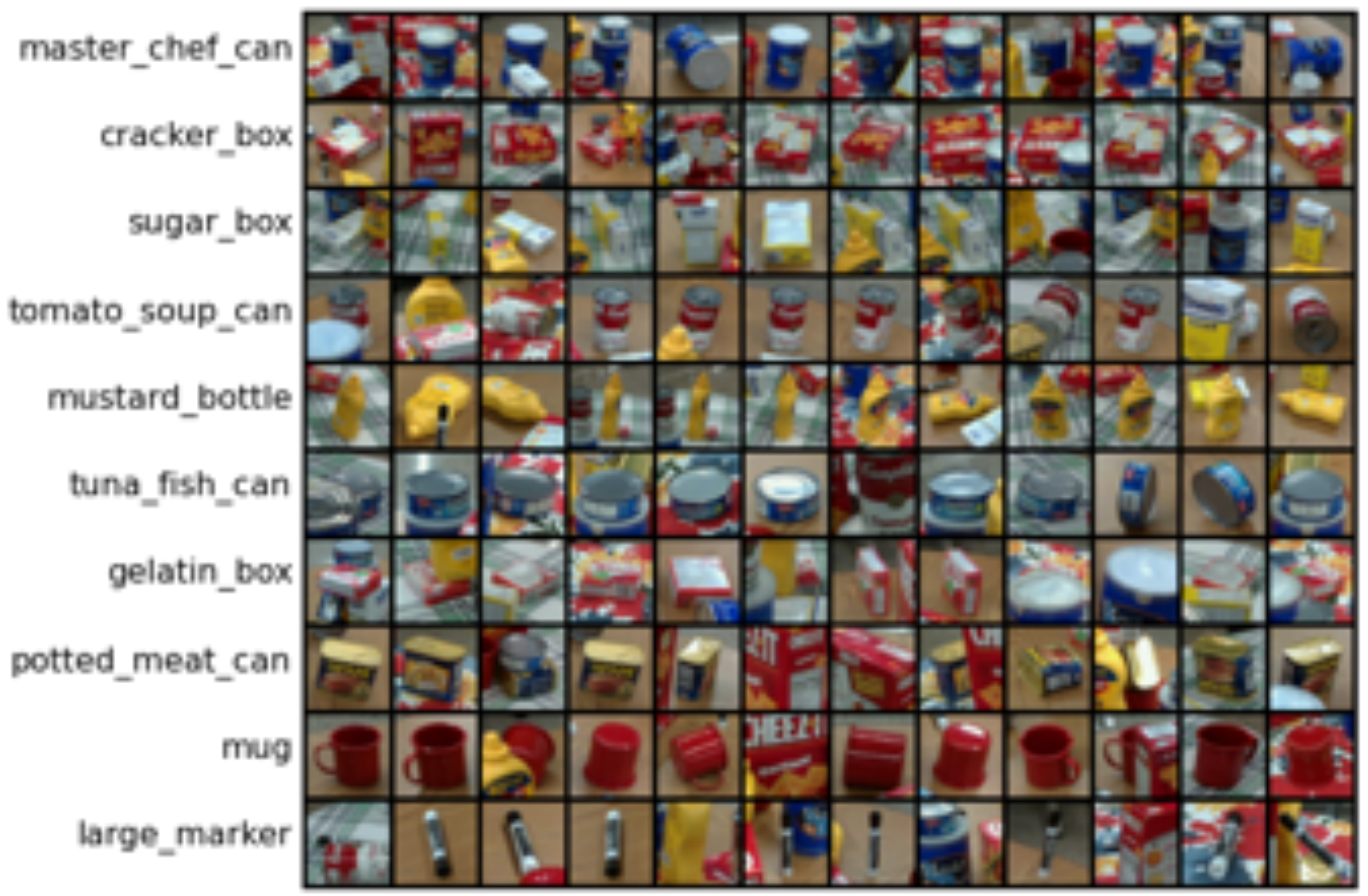
Interpreting a Linear Classifier

Instead of stretching pixels into columns, we can equivalently stretch rows of W into images!

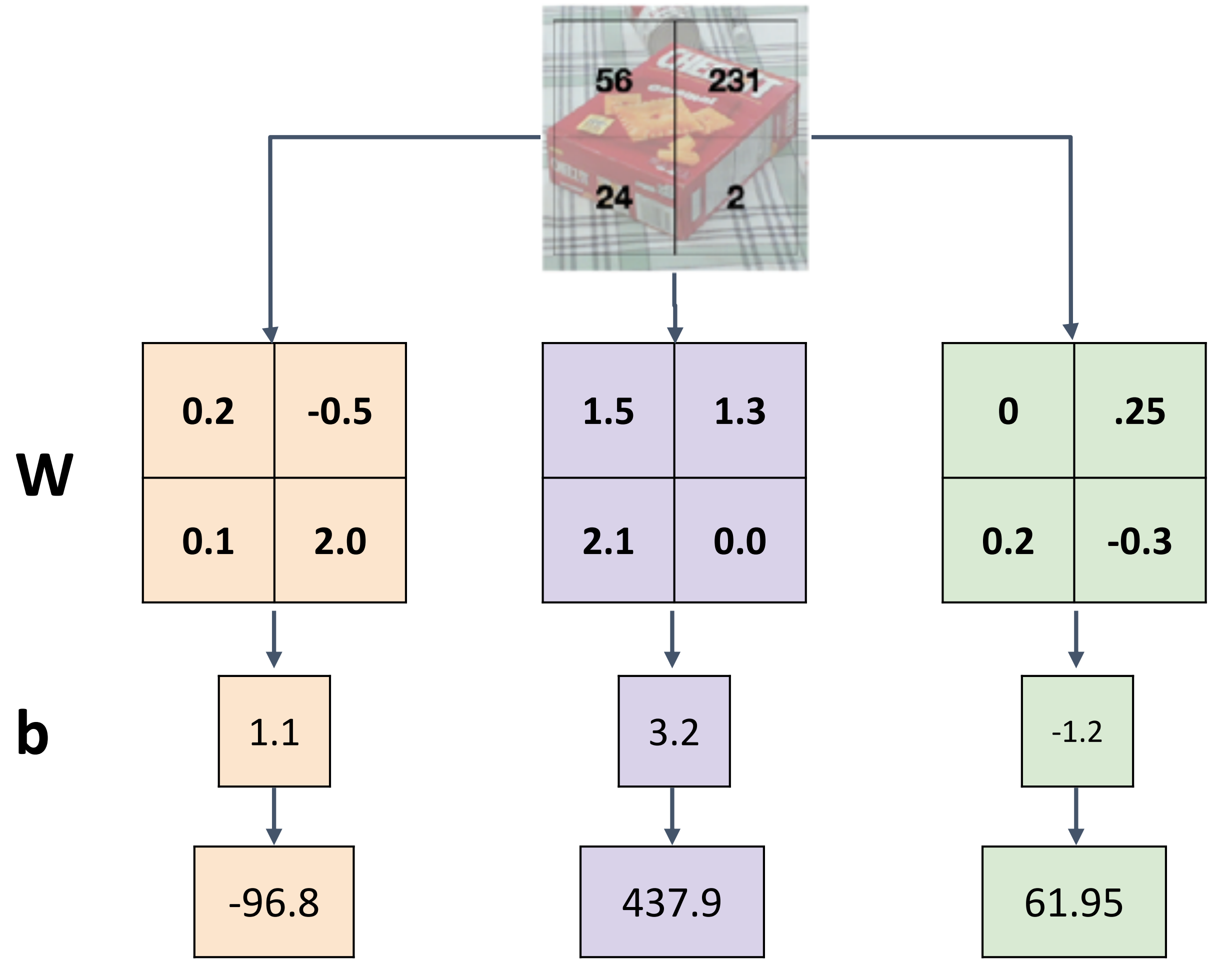




Interpreting a Linear Classifier



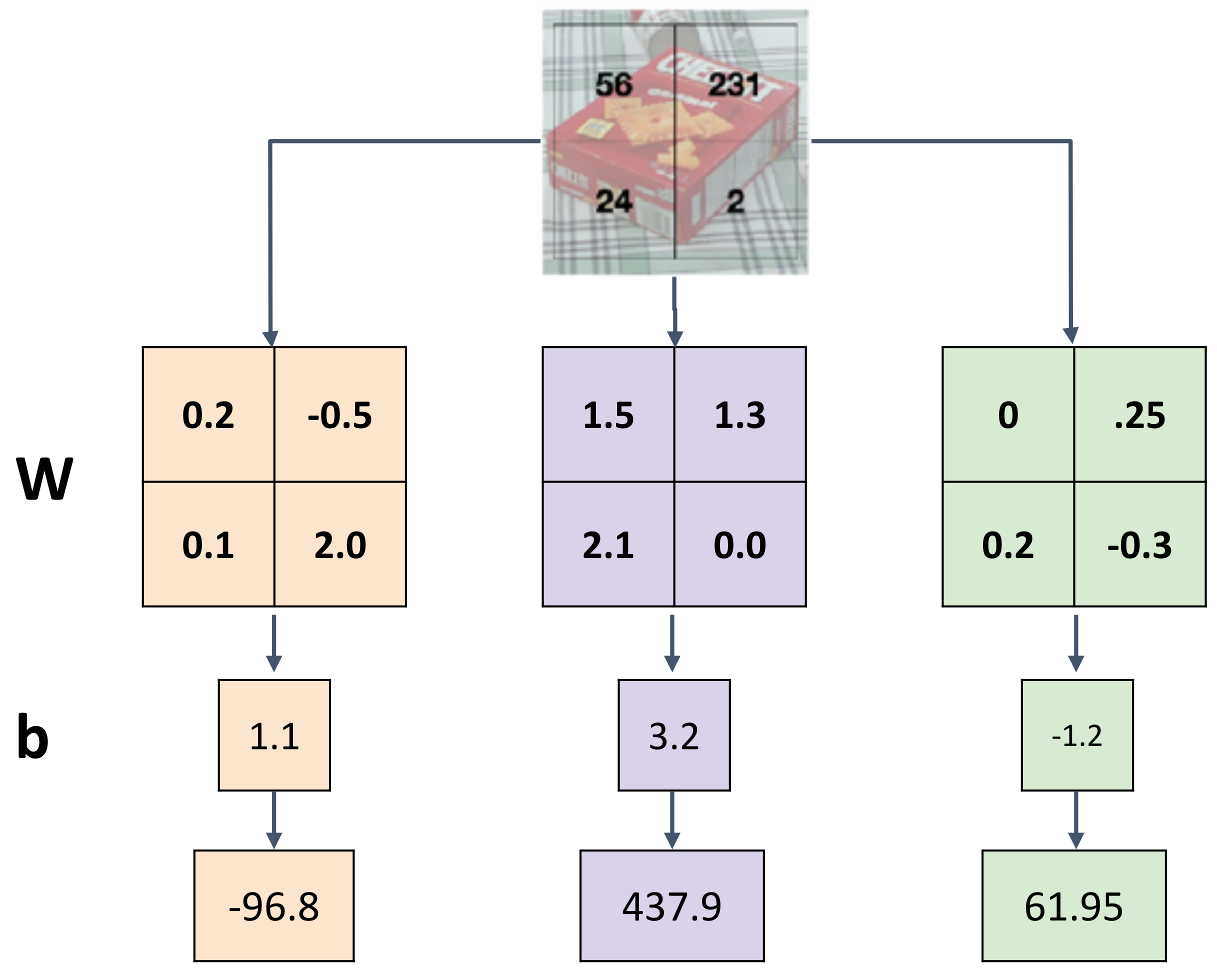
Instead of stretching pixels into columns, we can equivalently stretch rows of W into images!



Interpreting a Linear Classifier—Visual Viewpoint

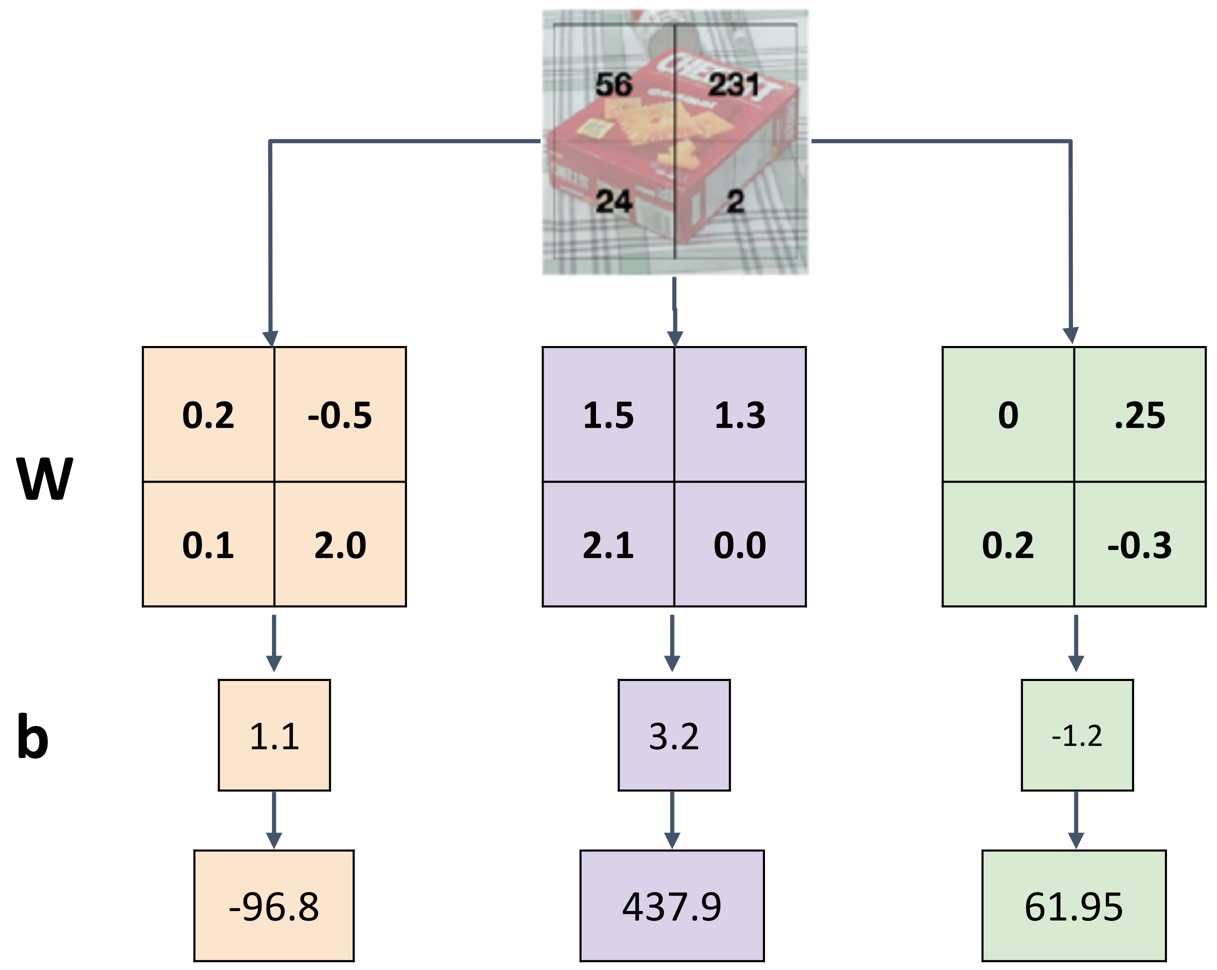
Linear classifier has one "template" per category

Instead of stretching pixels into columns, we can equivalently stretch rows of W into images!



Interpreting a Linear Classifier—Visual Viewpoint

Instead of stretching pixels into columns, we can equivalently stretch rows of W into images!

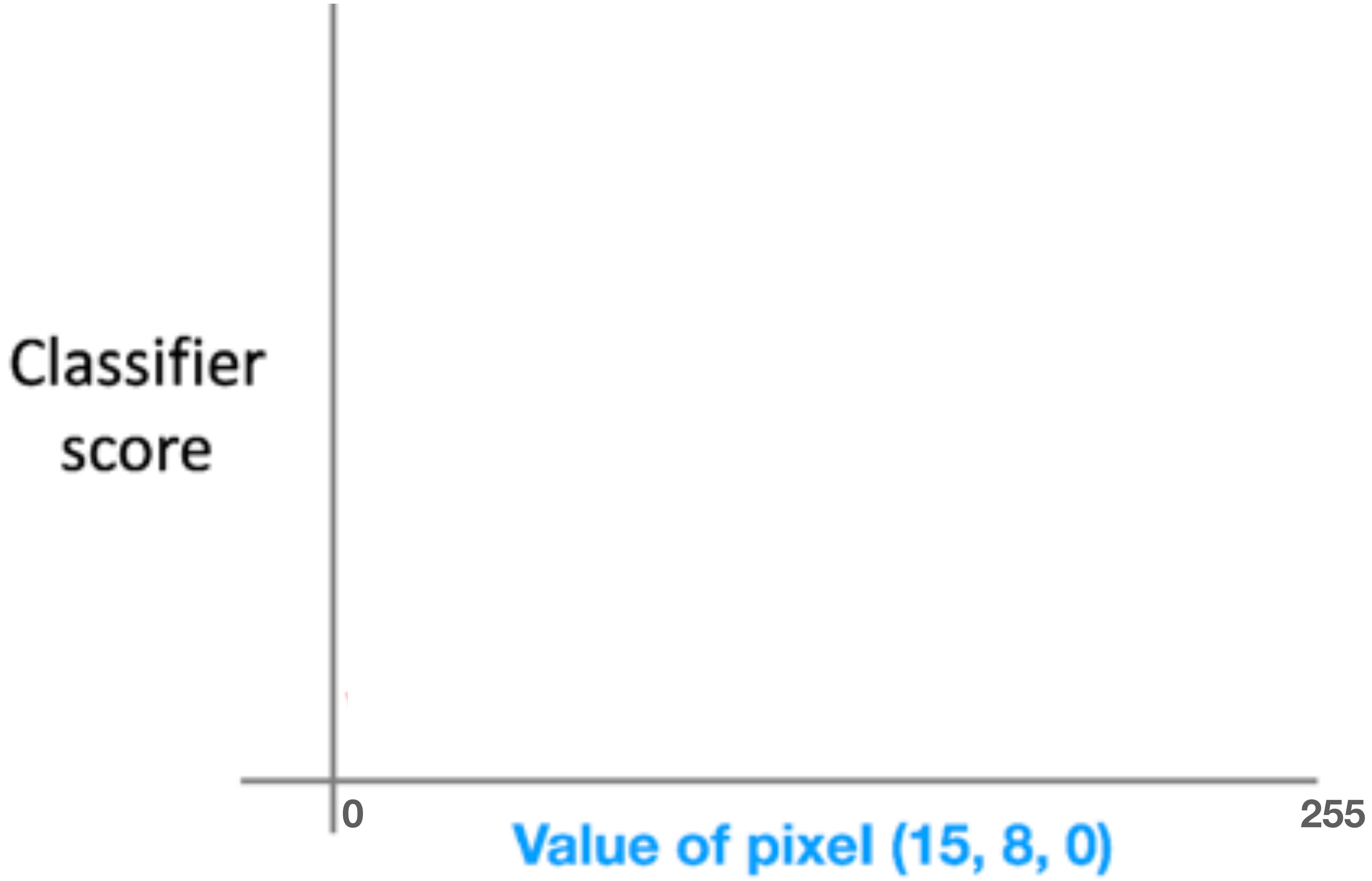


Linear classifier has one “template” per category

A single template cannot capture multiple modes of the data

e.g. mustard bottles can rotate

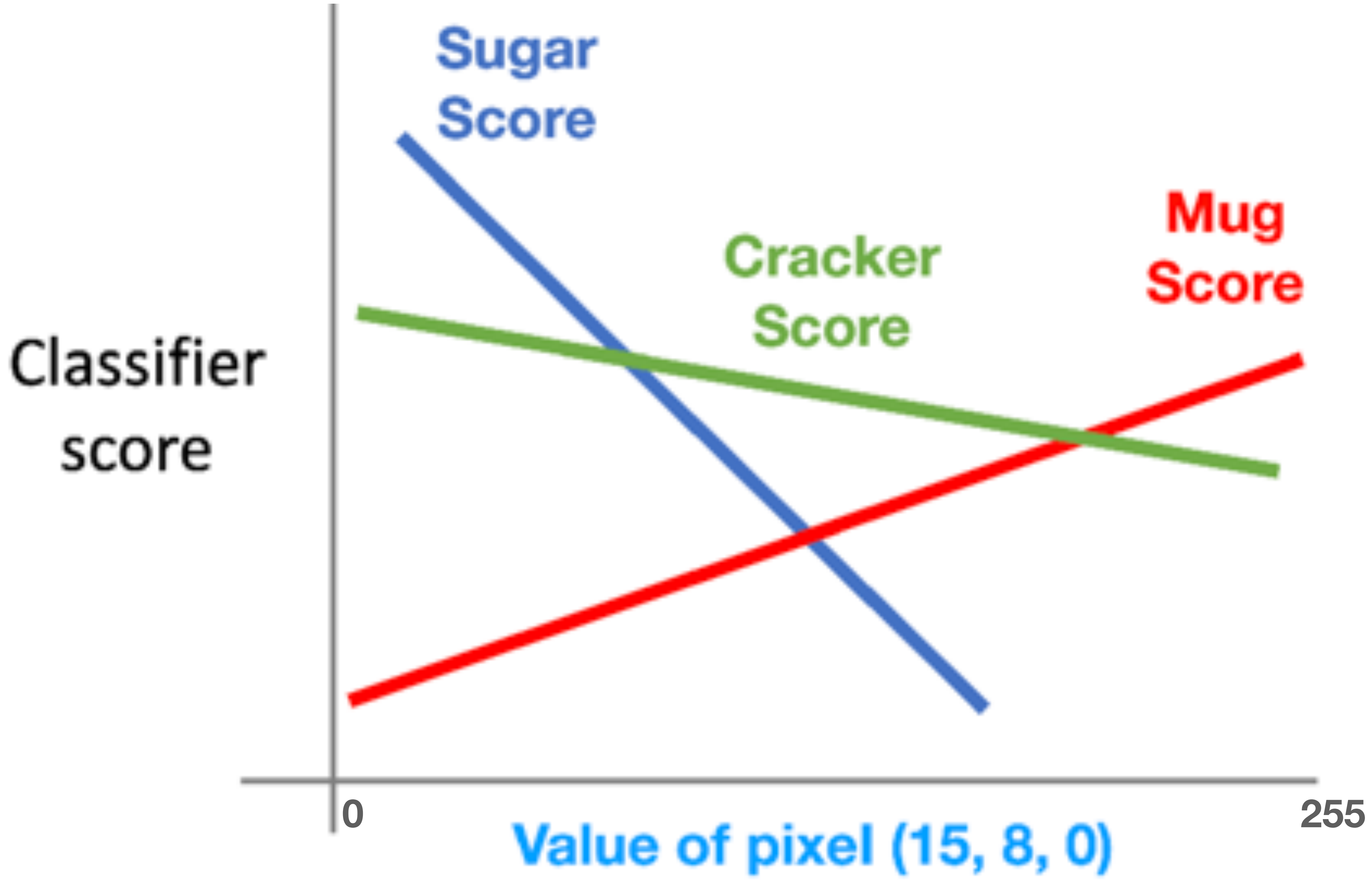
Interpreting a Linear Classifier—Geometric Viewpoint



$$f(x,W) = Wx + b$$

Array of **32x32x3** numbers
(3072 numbers total)

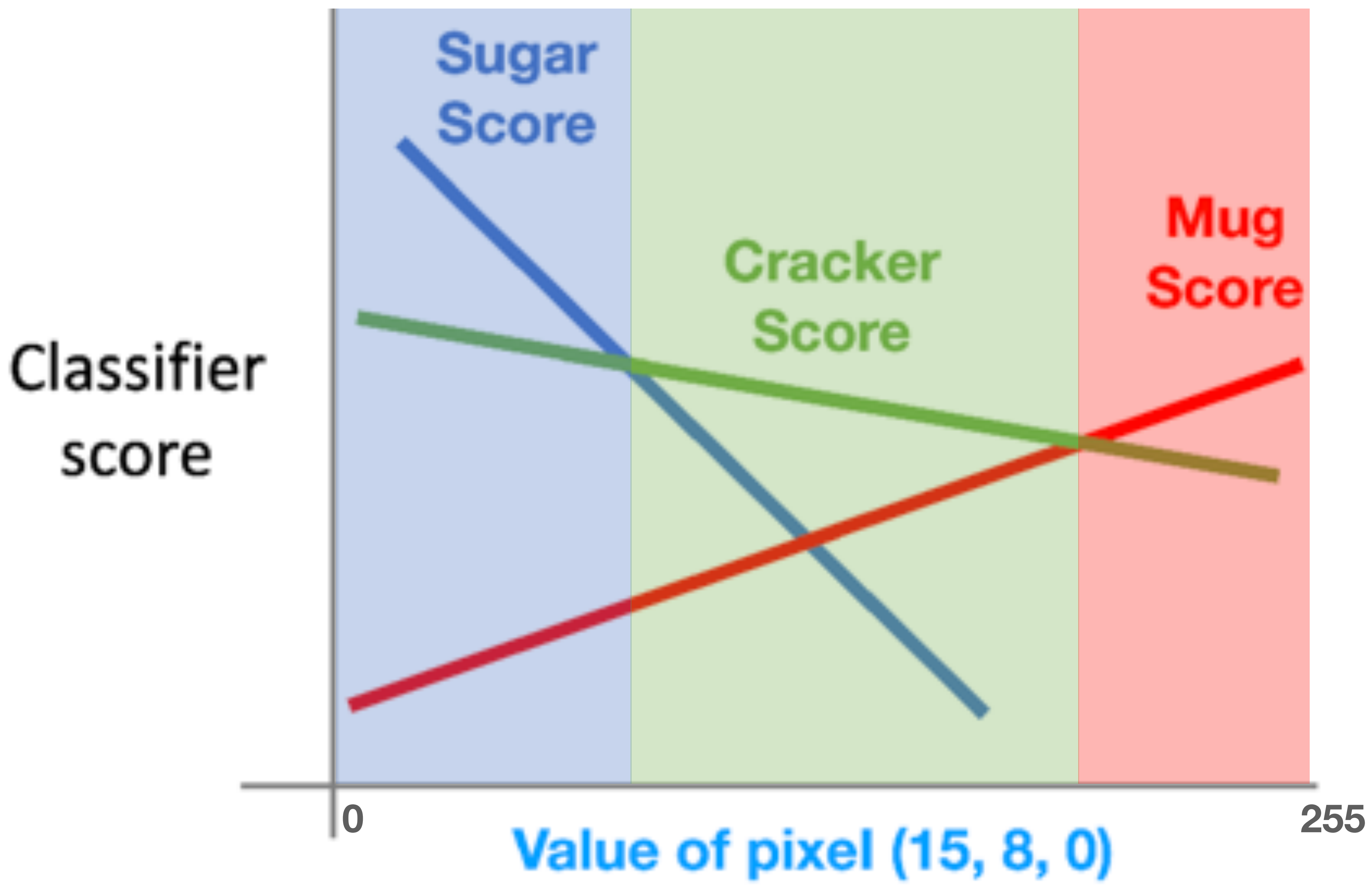
Interpreting a Linear Classifier—Geometric Viewpoint



$$f(x,W) = Wx + b$$

Array of **32x32x3** numbers
(3072 numbers total)

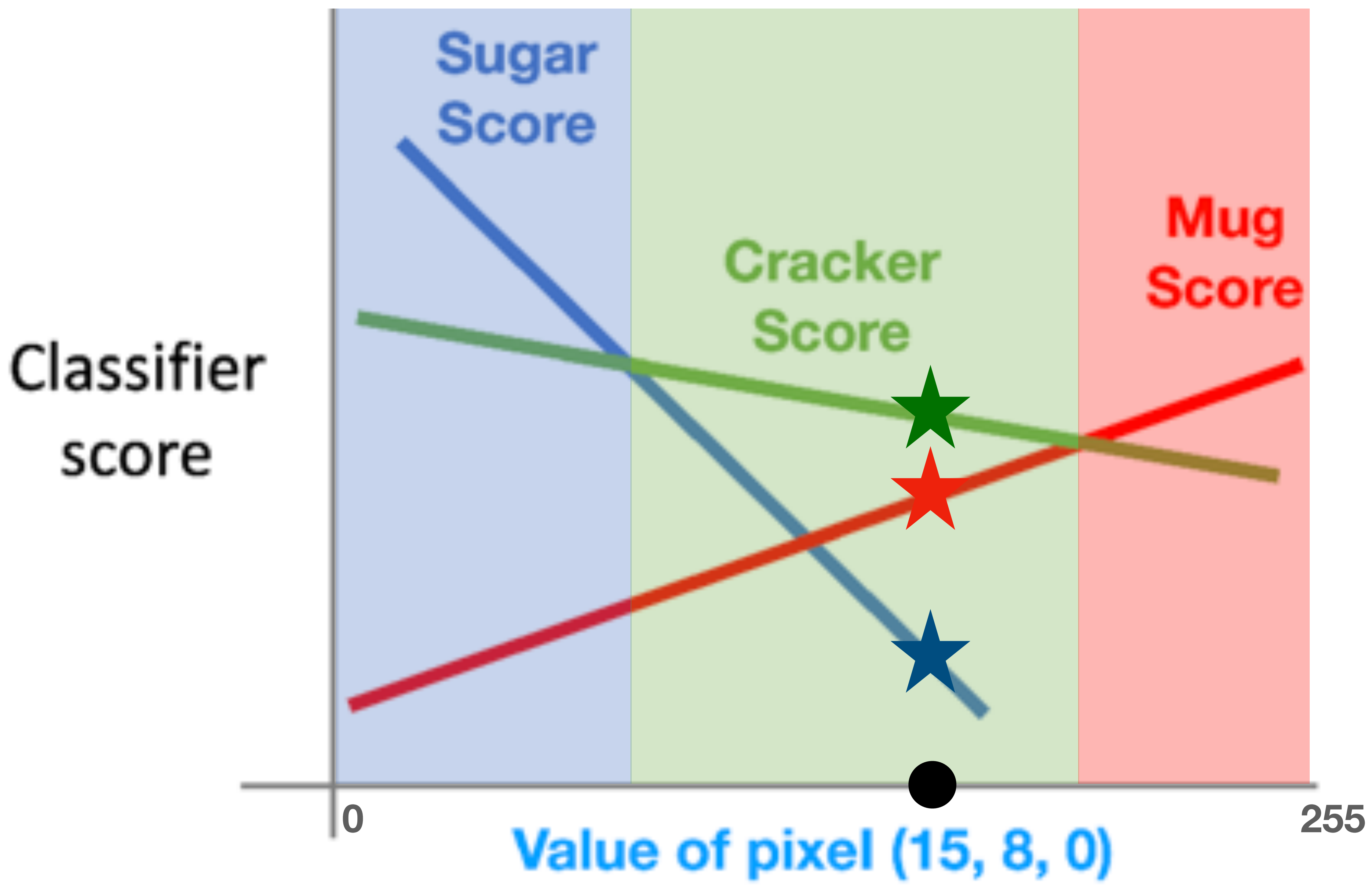
Interpreting a Linear Classifier—Geometric Viewpoint



$$f(x,W) = Wx + b$$

Array of **32x32x3** numbers
(3072 numbers total)

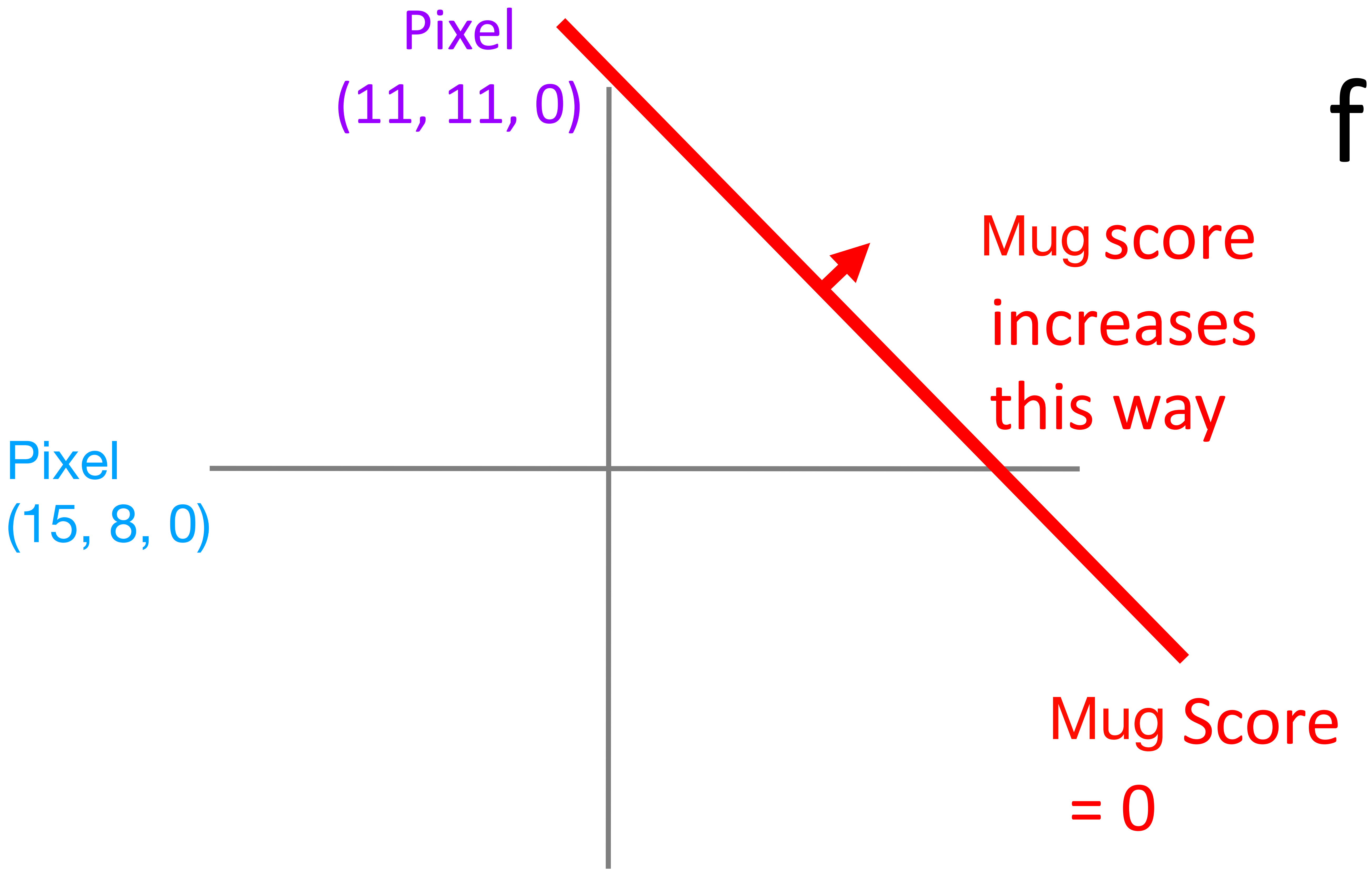
Interpreting a Linear Classifier—Geometric Viewpoint



$$f(x,W) = Wx + b$$

Array of **32x32x3** numbers
(3072 numbers total)

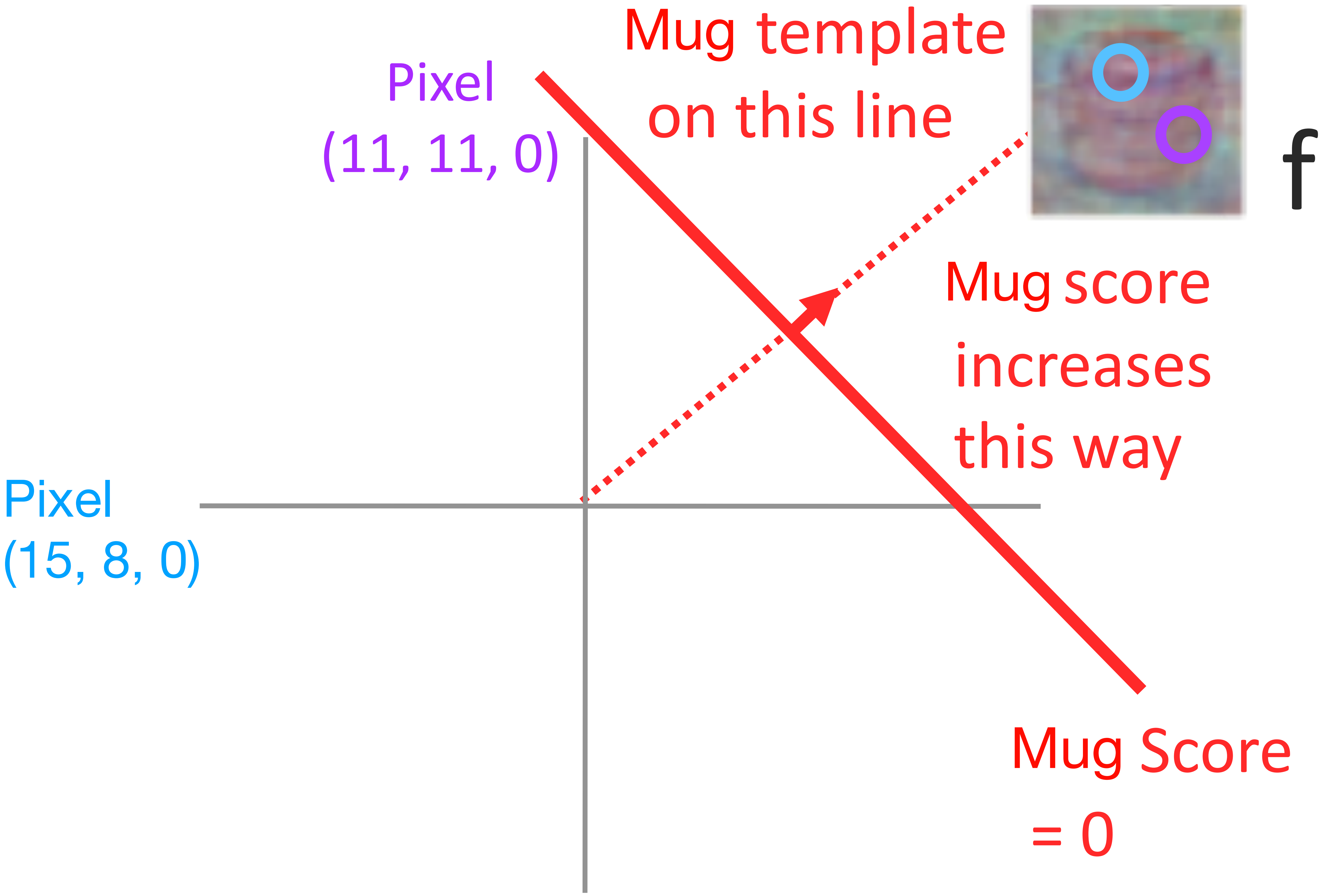
Interpreting a Linear Classifier—Geometric Viewpoint



$$f(x,W) = Wx + b$$

Array of **32x32x3** numbers
(3072 numbers total)

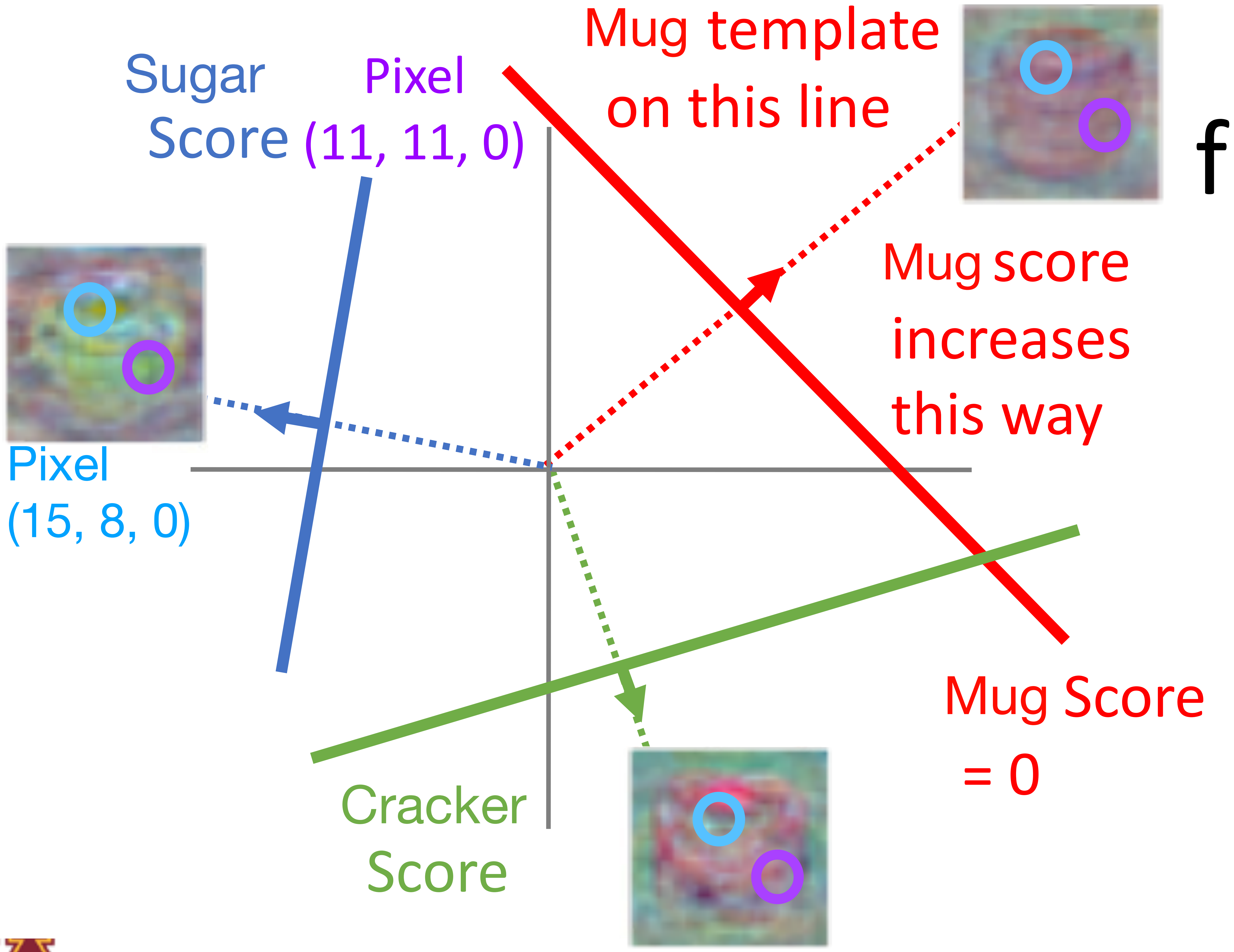
Interpreting a Linear Classifier—Geometric Viewpoint



$$f(x,W) = Wx + b$$

Array of **32x32x3** numbers
(3072 numbers total)

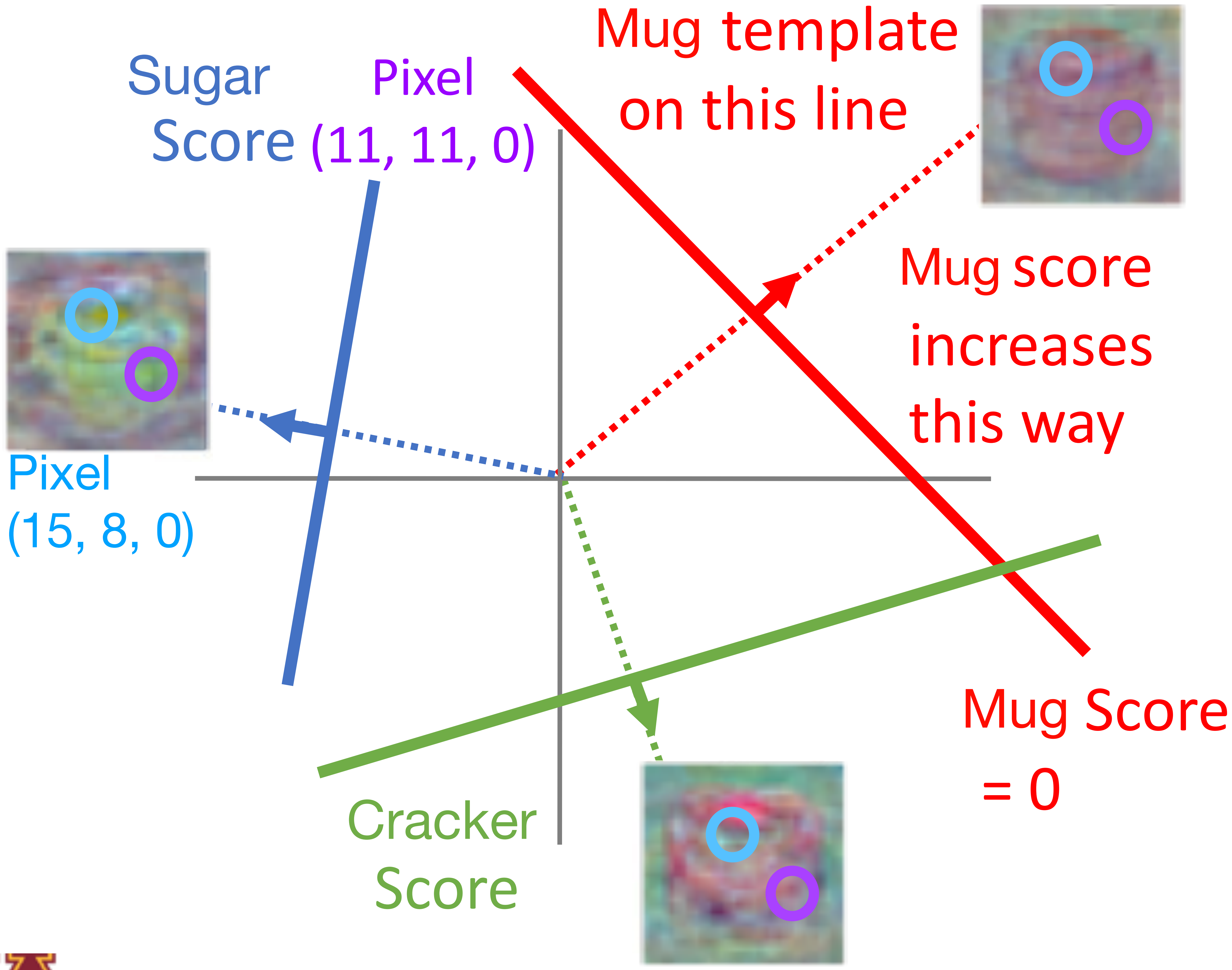
Interpreting a Linear Classifier—Geometric Viewpoint



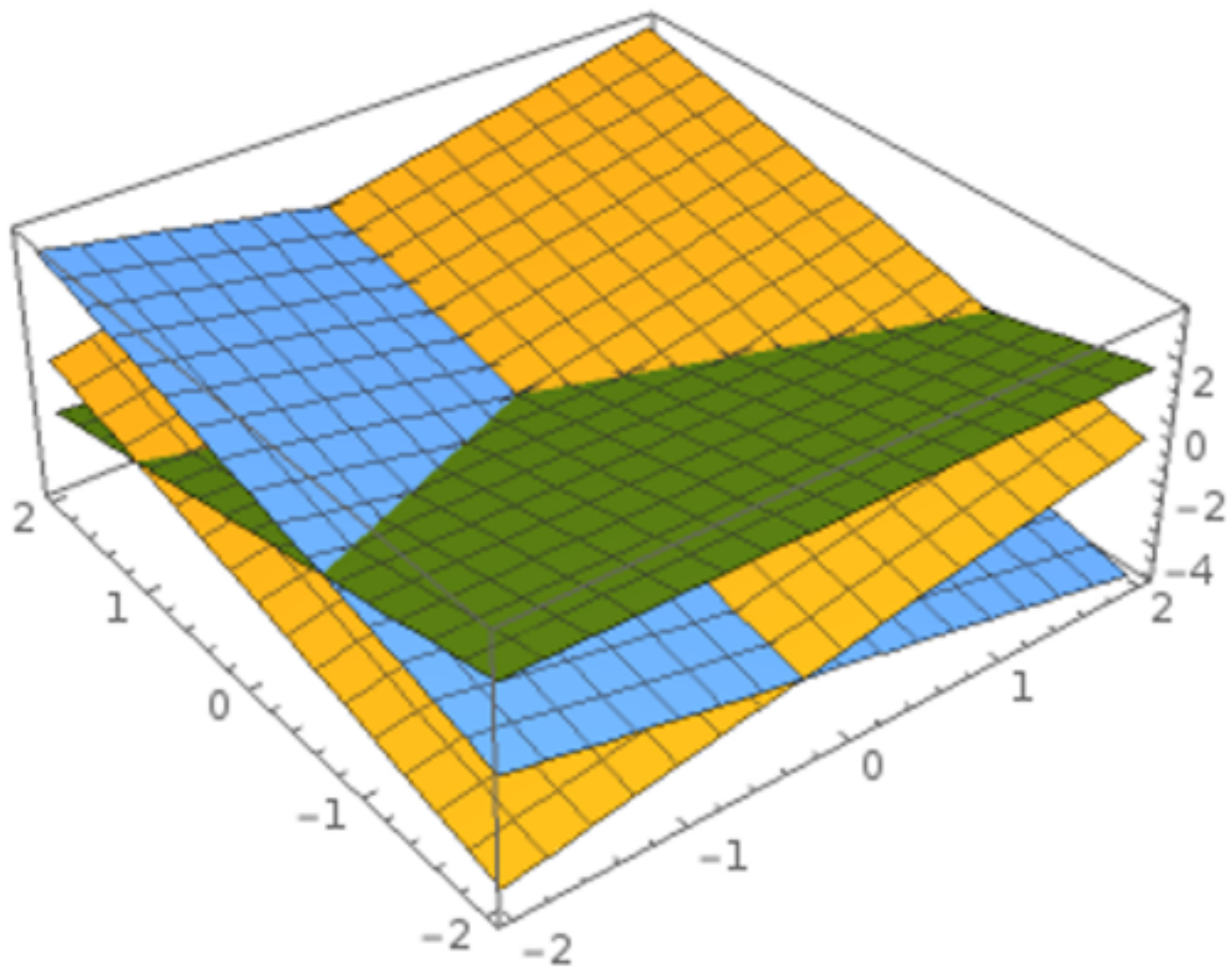
$$f(x,W) = Wx + b$$

Array of **32x32x3** numbers (3072 numbers total)

Interpreting a Linear Classifier—Geometric Viewpoint



Hyperplanes carving up a high-dimensional space

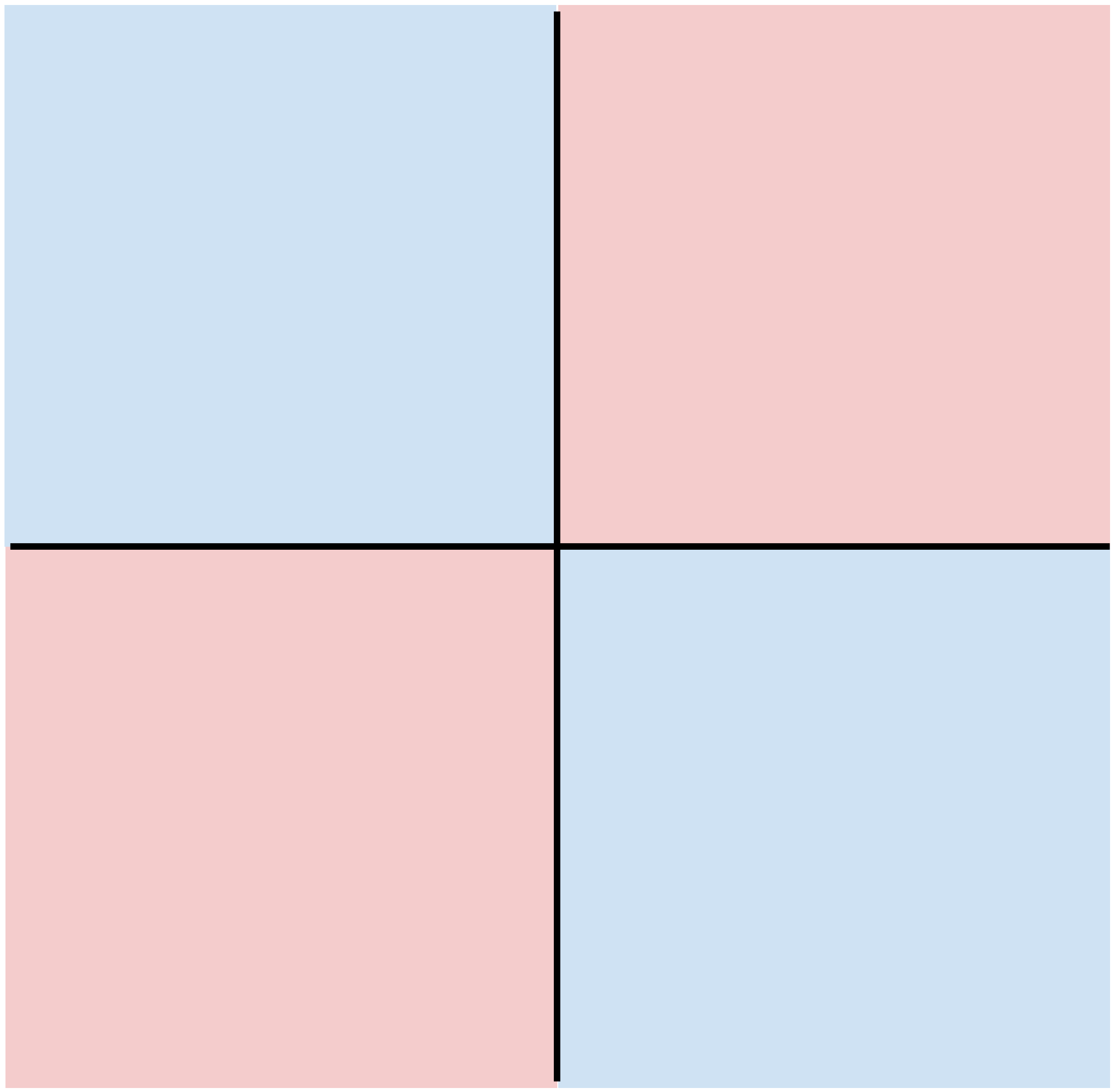


Plot created using [Wolfram Cloud]

Hard Cases for a Linear Classifier

Class 1:
First and third quadrants

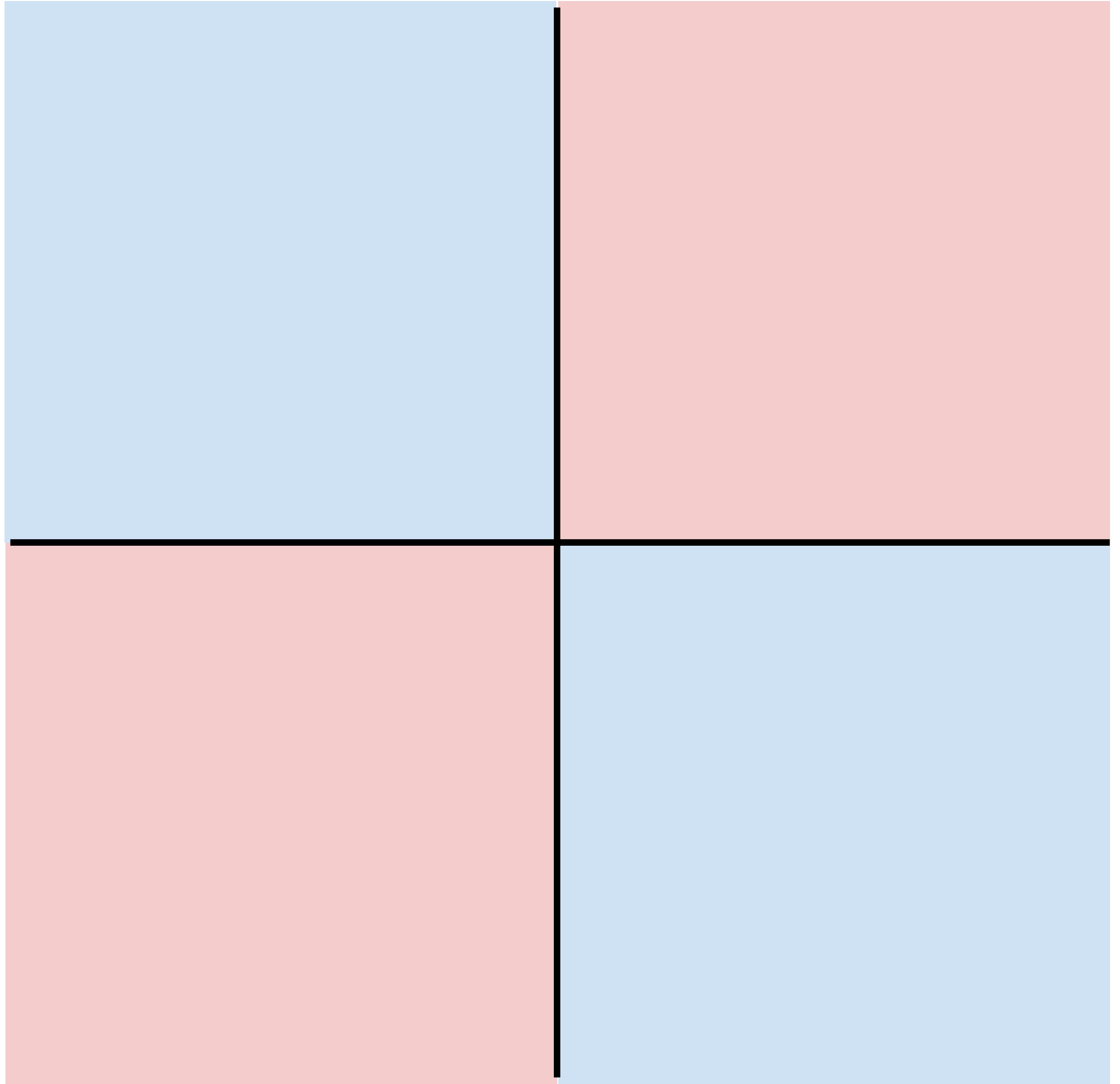
Class 2:
Second and fourth quadrants



Hard Cases for a Linear Classifier

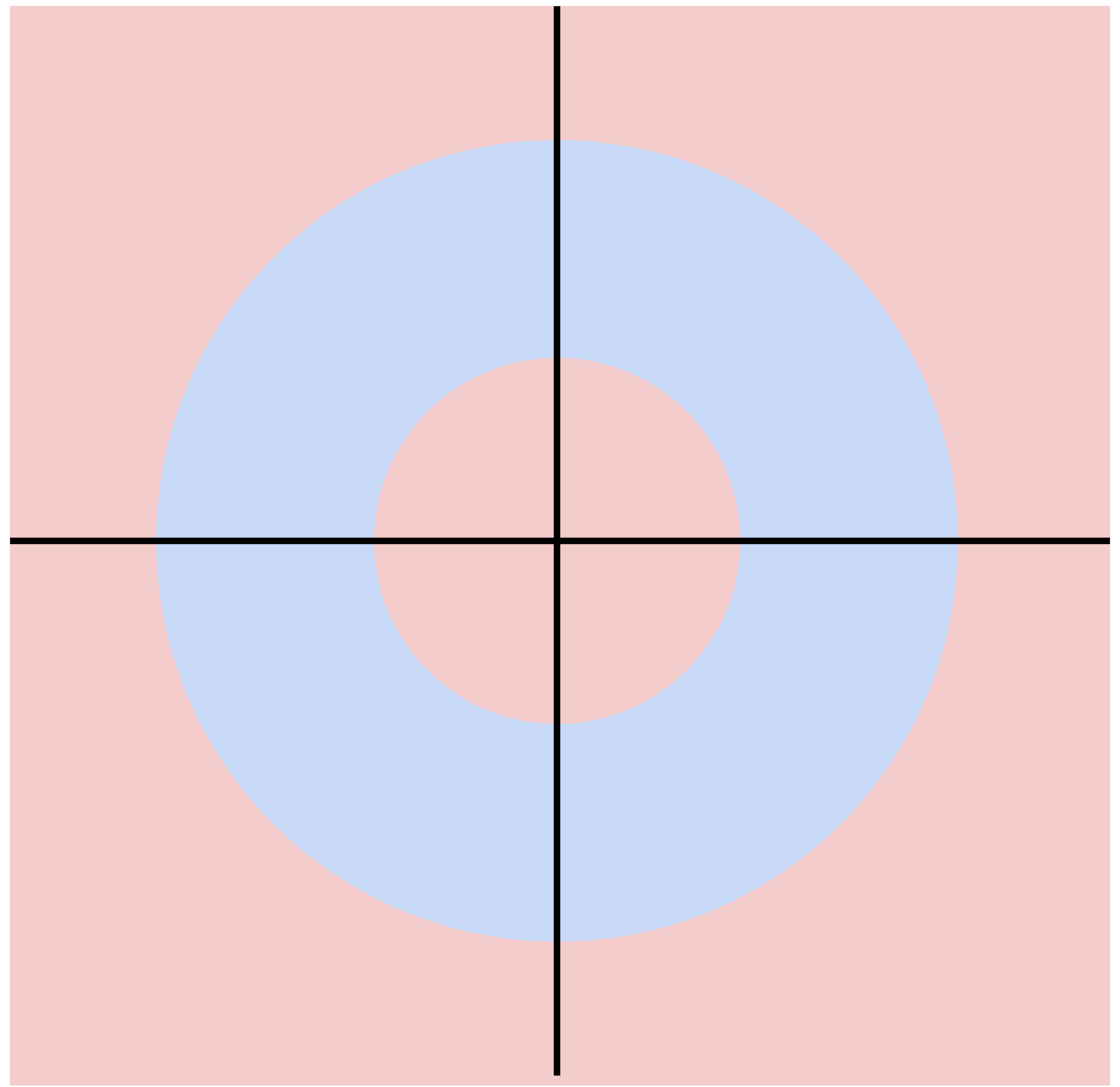
Class 1:
First and third quadrants

Class 2:
Second and fourth quadrants



Class 1:
1 <= L2 norm <= 2

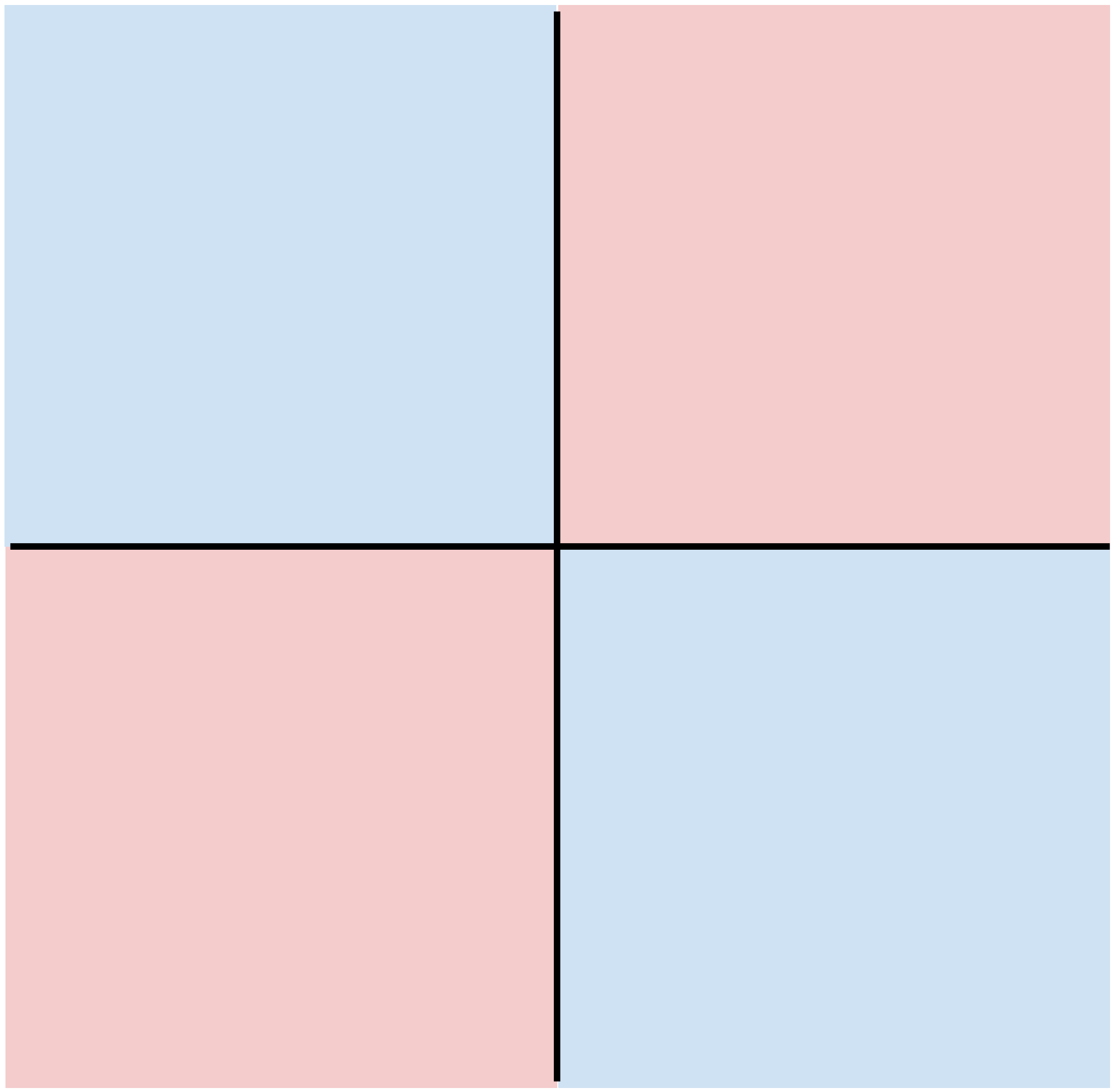
Class 2:
Everything else



Hard Cases for a Linear Classifier

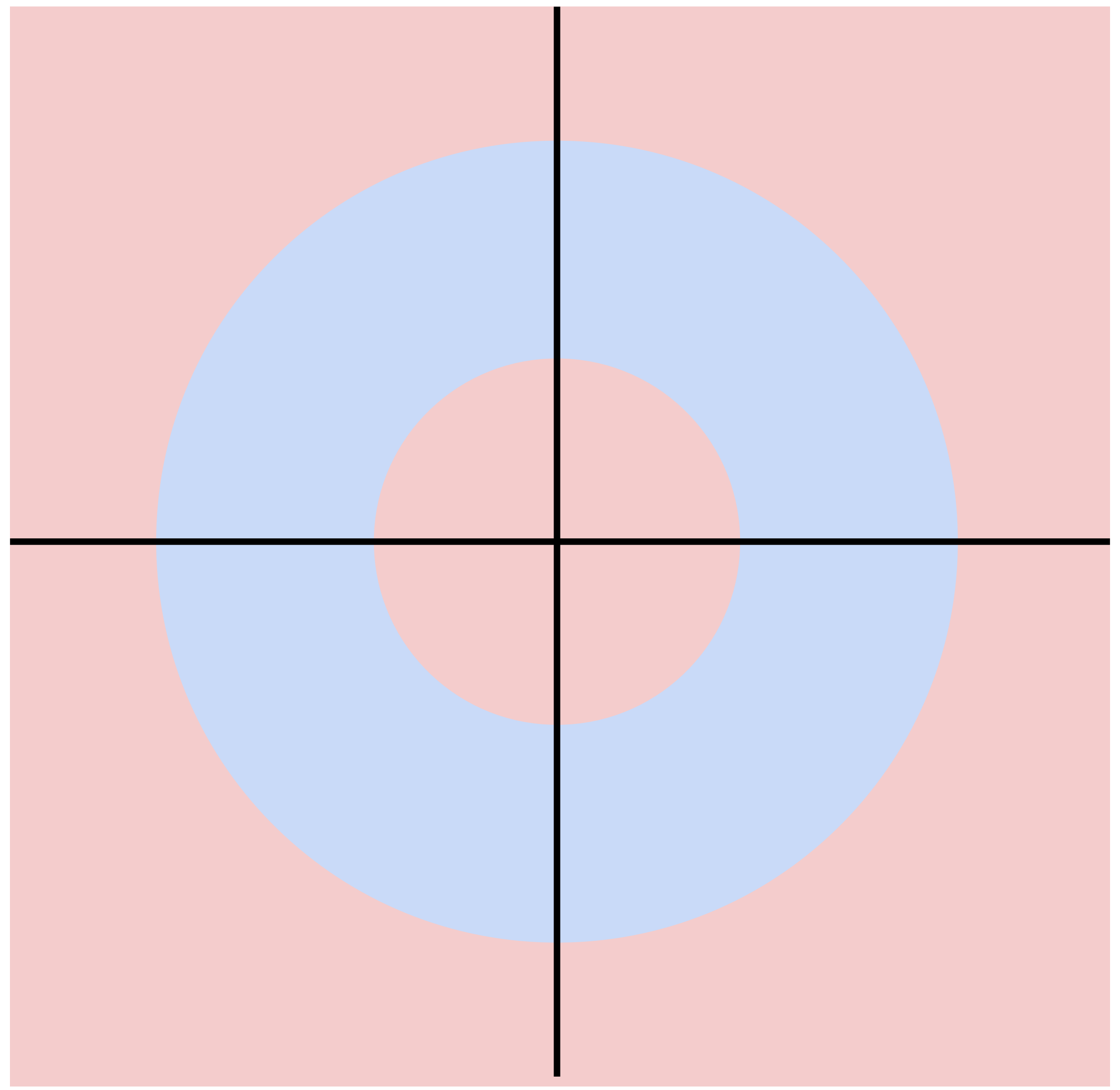
Class 1:
First and third quadrants

Class 2:
Second and fourth quadrants



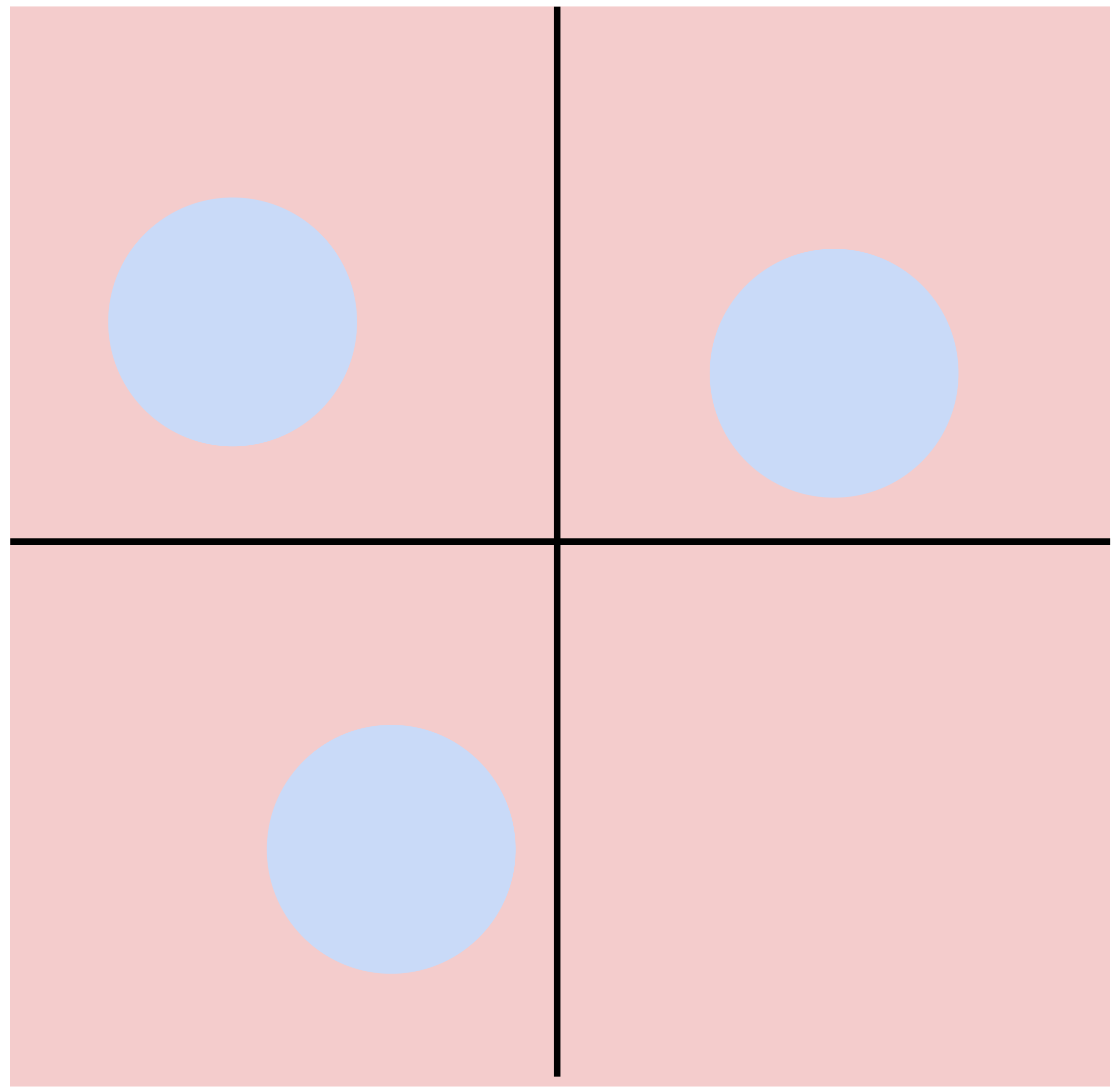
Class 1:
1 <= L2 norm <= 2

Class 2:
Everything else



Class 1:
Three modes

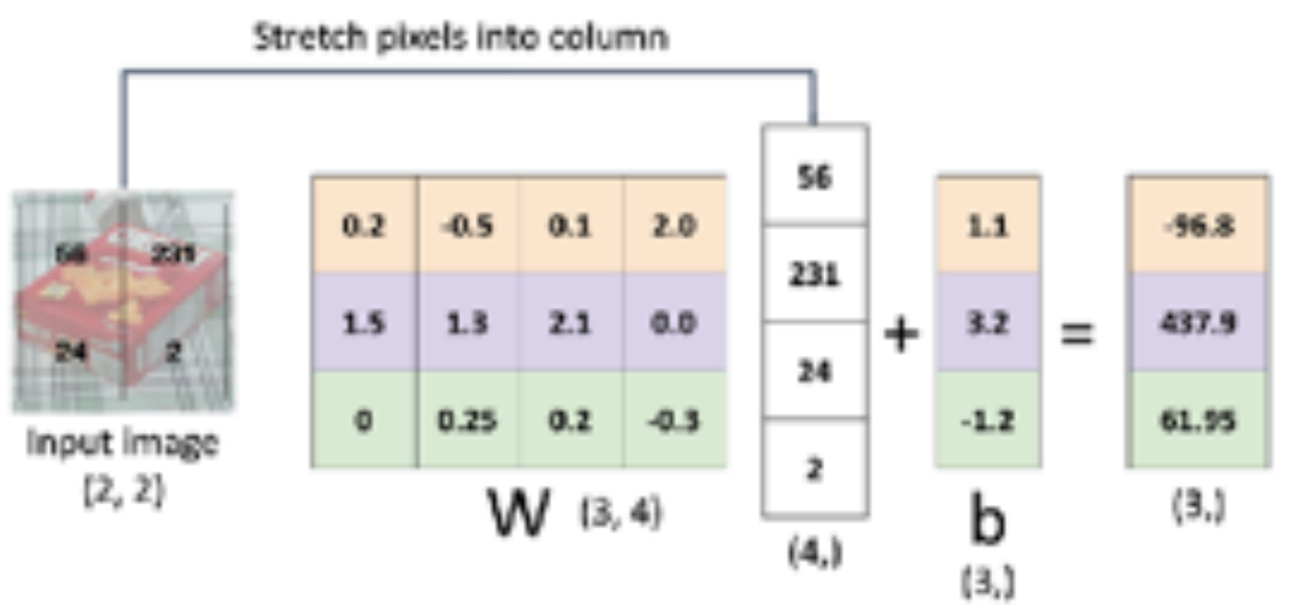
Class 2:
Everything else



Linear Classifier—Three Viewpoints

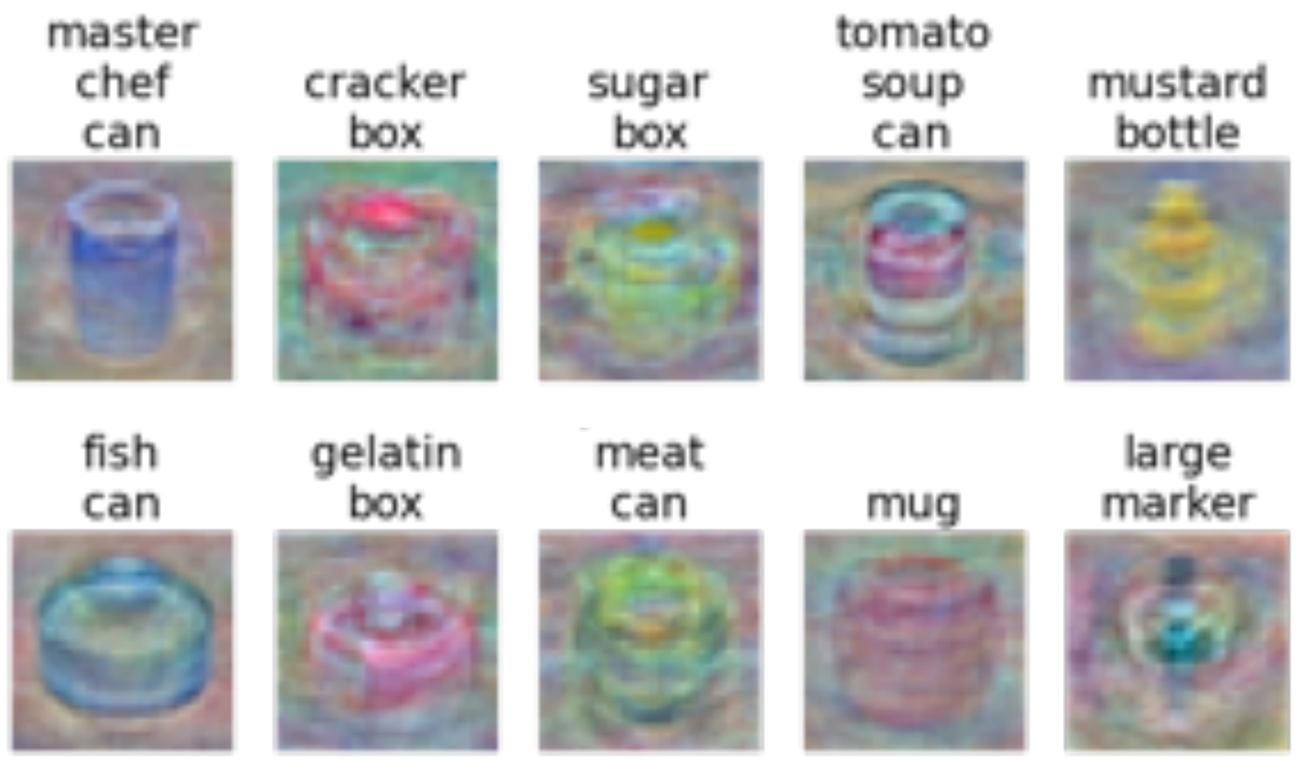
Algebraic Viewpoint

$f(x,W) = Wx$



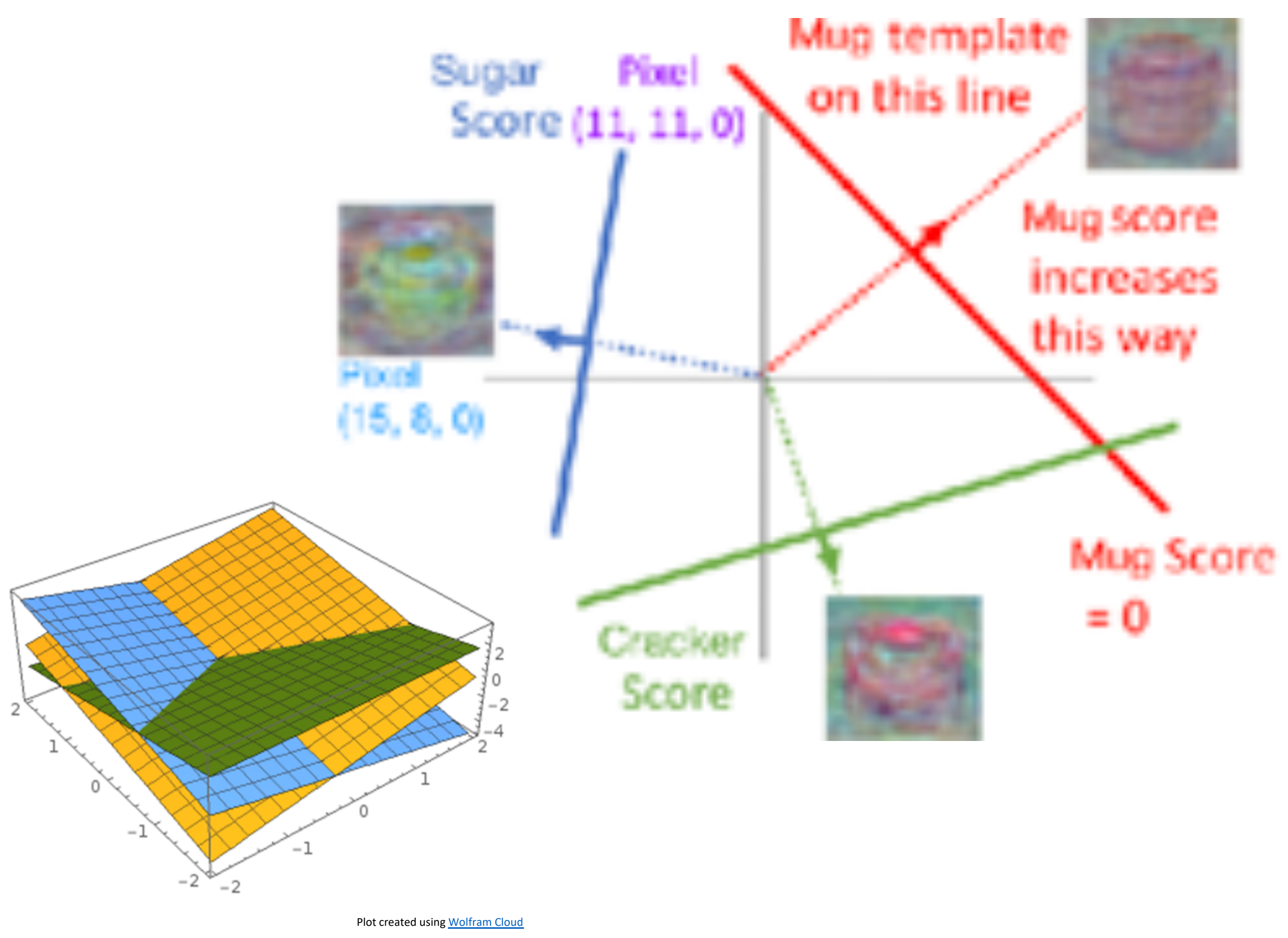
Visual Viewpoint

One template
per class



Geometric Viewpoint

Hyperplanes
cutting up space



So far—Defined a Score Function


master chef can	-3.45	-0.51	3.42
mug	-8.87	**6.04**	4.64
tomato soup can	0.09	5.31	2.65
cracker box	**2.9**	-4.22	5.1
mustard bottle	4.48	-4.19	2.64
tuna fish can	8.02	3.58	5.55
sugar box	3.78	4.49	**-4.34**
gelatin box	1.06	-4.37	-1.5
potted meat can	-0.36	-2.09	-4.79
large marker	-0.72	-2.93	6.14

$f(x,W) = Wx + b$

Given a W, we can compute class scores for an image, x.

But how can we actually choose a good W?

So far—Choosing a Good W


master chef can	-3.45	-0.51	3.42
mug	-8.87	**6.04**	4.64
tomato soup can	0.09	5.31	2.65
cracker box	**2.9**	-4.22	5.1
mustard bottle	4.48	-4.19	2.64
tuna fish can	8.02	3.58	5.55
sugar box	3.78	4.49	**-4.34**
gelatin box	1.06	-4.37	-1.5
potted meat can	-0.36	-2.09	-4.79
large marker	-0.72	-2.93	6.14

$$f(x,W) = Wx + b$$

TODO:

1. Use a **loss function** to quantify how good a value of W is
2. Find a W that minimizes the loss function (**optimization**)

Loss Function

A **loss function** measures how good our current classifier is

Low loss = good classifier
High loss = bad classifier

Also called: **objective function, cost function**

Loss Function

A **loss function** measures how good our current classifier is

Low loss = good classifier
High loss = bad classifier

Also called: **objective function, cost function**

Negative loss function sometimes called **reward function, profit function, utility function, fitness function,** etc.

Loss Function

A **loss function** measures how good our current classifier is

Low loss = good classifier
High loss = bad classifier

Also called: **objective function, cost function**

Negative loss function sometimes called **reward function, profit function, utility function, fitness function,** etc.

Given a dataset of examples

$$\{(x_i, y_i)\}_{i=1}^{N}$$

where x_i is an image and
y_i is a (discrete) label

Loss Function

A **loss function** measures how good our current classifier is

Low loss = good classifier
High loss = bad classifier

Also called: **objective function, cost function**

Negative loss function sometimes called **reward function, profit function, utility function, fitness function,** etc.

Given a dataset of examples

$$\{(x_i, y_i)\}_{i=1}^{N}$$

where x_i is an image and
y_i is a (discrete) label

Loss for a single example is

$$L_i(f(x_i, W), y_i)$$

Loss Function

A **loss function** measures how good our current classifier is

Low loss = good classifier
High loss = bad classifier

Also called: **objective function, cost function**

Negative loss function sometimes called **reward function, profit function, utility function, fitness function,** etc.

Given a dataset of examples

$$\{(x_i, y_i)\}_{i=1}^{N}$$

where x_i is an image and y_i is a (discrete) label

Loss for a single example is

$$L_i(f(x_i, W), y_i)$$

Loss for the dataset is average of per-example losses:

$$L = \frac{1}{N} \sum_i L_i(f(x_i, W), y_i)$$

Cross-Entropy Loss
Multinomial Logistic Regression

Want to interpret raw classifier scores as **probabilities**

cracker	**3.2**
mug	5.1
sugar	-1.7

Cross-Entropy Loss
Multinomial Logistic Regression

cracker	**3.2**
mug	5.1
sugar	-1.7

Want to interpret raw classifier scores as **probabilities**

$$s = f(x_i; W) \qquad P(Y = k \mid X = x_i) = \frac{\exp(s_k)}{\sum_j \exp(s_j)}$$

Softmax function

Cross-Entropy Loss
Multinomial Logistic Regression

cracker	**3.2**
mug	5.1
sugar	-1.7

Unnormalized log-probabilities (logits)

Want to interpret raw classifier scores as **probabilities**

$$s = f(x_i; W)$$

$$P(Y = k \mid X = x_i) = \frac{\exp(s_k)}{\sum_j \exp(s_j)}$$

Softmax function

Cross-Entropy Loss
Multinomial Logistic Regression

Want to interpret raw classifier scores as **probabilities**

$$s = f(x_i; W) \qquad P(Y = k \mid X = x_i) = \frac{\exp(s_k)}{\sum_j \exp(s_j)}$$

Softmax function

Probabilities must be >=0

		exp(·)	
cracker	**3.2**	→	**24.5**
mug	5.1		164.0
sugar	-1.7		0.18

Unnormalized log-probabilities (logits)

Unnormalized probabilities

Cross-Entropy Loss
Multinomial Logistic Regression

Want to interpret raw classifier scores as **probabilities**

$$s = f(x_i; W) \qquad P(Y = k \mid X = x_i) = \frac{\exp(s_k)}{\sum_j \exp(s_j)}$$ Softmax function

	Unnormalized log-probabilities (logits)	exp(·) →	Unnormalized probabilities (Probabilities must be >=0)	normalize →	Probabilities (Probabilities must sum to 1)
cracker	**3.2**		**24.5**		**0.13**
mug	5.1		164.0		0.87
sugar	-1.7		0.18		0.00

Cross-Entropy Loss
Multinomial Logistic Regression

Want to interpret raw classifier scores as **probabilities**

$$s = f(x_i; W) \qquad P(Y = k \mid X = x_i) = \frac{\exp(s_k)}{\sum_j \exp(s_j)}$$

Softmax function

Probabilities must be >=0

Probabilities must sum to 1

	Unnormalized log-probabilities (logits)		Unnormalized probabilities		Probabilities
cracker	**3.2**	exp(·) →	**24.5**	normalize →	**0.13**
mug	5.1		164.0		0.87
sugar	-1.7		0.18		0.00

$$L_i = -\log P(Y = y_i \mid X = x_i)$$

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

Cross-Entropy Loss
Multinomial Logistic Regression

Want to interpret raw classifier scores as **probabilities**

$$s = f(x_i; W) \qquad P(Y = k \mid X = x_i) = \frac{\exp(s_k)}{\sum_j \exp(s_j)}$$ Softmax function

Probabilities must be >=0

Probabilities must sum to 1

		exp(·) →		normalize →	
cracker	**3.2**		**24.5**		**0.13**
mug	5.1		164.0		0.87
sugar	-1.7		0.18		0.00
	Unnormalized log-probabilities (logits)		Unnormalized probabilities		Probabilities

$$L_i = -\log P(Y = y_i \mid X = x_i)$$

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

Maximum Likelihood Estimation
Choose weights to maximize the likelihood of the observed data (see CSCI 5521)

Cross-Entropy Loss
Multinomial Logistic Regression

Want to interpret raw classifier scores as **probabilities**

$$s = f(x_i; W) \qquad P(Y = k \mid X = x_i) = \frac{\exp(s_k)}{\sum_j \exp(s_j)}$$

Softmax function

	Unnormalized log-probabilities (logits)	$\xrightarrow{\exp(\cdot)}$	Unnormalized probabilities	$\xrightarrow{\text{normalize}}$	Probabilities	compare	Correct probabilities
cracker	**3.2**		**24.5**		**0.13**	⟶ compare ⟵	**1.00**
mug	5.1		164.0		0.87		0.00
sugar	-1.7		0.18		0.00		0.00

Probabilities must be >=0

Probabilities must sum to 1

Cross-Entropy Loss
Multinomial Logistic Regression

Want to interpret raw classifier scores as **probabilities**

$$s = f(x_i; W) \qquad P(Y = k \mid X = x_i) = \frac{\exp(s_k)}{\sum_j \exp(s_j)}$$

Softmax function

	Unnormalized log-probabilities (logits)	exp(·) →	Unnormalized probabilities (Probabilities must be >=0)	normalize →	Probabilities (Probabilities must sum to 1)		Correct probabilities
cracker	**3.2**		**24.5**		**0.13**	→ compare ←	**1.00**
mug	5.1		164.0		0.87		0.00
sugar	-1.7		0.18		0.00		0.00

Kullback-Leibler divergence

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

Cross-Entropy Loss
Multinomial Logistic Regression

Want to interpret raw classifier scores as **probabilities**

$$s = f(x_i; W)$$

$$P(Y = k \mid X = x_i) = \frac{\exp(s_k)}{\sum_j \exp(s_j)}$$ Softmax function

	Unnormalized log-probabilities (logits)	exp(·) →	Unnormalized probabilities (Probabilities must be >=0)	normalize →	Probabilities (Probabilities must sum to 1)	compare	Correct probabilities
cracker	**3.2**		**24.5**		**0.13**	→ compare ←	**1.00**
mug	5.1		164.0		0.87		0.00
sugar	-1.7		0.18		0.00		0.00

Cross Entropy

$$H(P, Q) = H(P) + D_{KL}(P \| Q)$$

Cross-Entropy Loss
Multinomial Logistic Regression

cracker	**3.2**
mug	5.1
sugar	-1.7

Want to interpret raw classifier scores as **probabilities**

$$s = f(x_i; W) \qquad P(Y = k \mid X = x_i) = \frac{\exp(s_k)}{\sum_j \exp(s_j)}$$

Softmax function

Maximize probability of correct class

$$L_i = -\log P(Y = y_i \mid X = x_i)$$

Putting it all together

$$L_i = -\log\left(\frac{\exp(s_{y_i})}{\sum_j \exp(s_j)}\right)$$

Cross-Entropy Loss
Multinomial Logistic Regression

cracker	**3.2**
mug	5.1
sugar	-1.7

Want to interpret raw classifier scores as **probabilities**

$$s = f(x_i; W) \qquad P(Y = k \mid X = x_i) = \frac{\exp(s_k)}{\sum_j \exp(s_j)}$$

Softmax function

Maximize probability of correct class

$$L_i = -\log P(Y = y_i \mid X = x_i)$$

Putting it all together

$$L_i = -\log\left(\frac{\exp(s_{y_i})}{\sum_j \exp(s_j)}\right)$$

Q: What is the min / max possible loss L_i?

Cross-Entropy Loss
Multinomial Logistic Regression

cracker	**3.2**
mug	5.1
sugar	-1.7

Want to interpret raw classifier scores as **probabilities**

$$s = f(x_i; W) \qquad P(Y = k \mid X = x_i) = \frac{\exp(s_k)}{\sum_j \exp(s_j)}$$

Softmax function

Maximize probability of correct class

$$L_i = -\log P(Y = y_i \mid X = x_i)$$

Putting it all together

$$L_i = -\log\left(\frac{\exp(s_{y_i})}{\sum_j \exp(s_j)}\right)$$

Q: What is the min / max possible loss L_i?

A: Min: 0, Max: $+\infty$

Cross-Entropy Loss
Multinomial Logistic Regression

cracker	**3.2**
mug	5.1
sugar	-1.7

Want to interpret raw classifier scores as **probabilities**

$$s = f(x_i; W) \qquad P(Y = k \mid X = x_i) = \frac{\exp(s_k)}{\sum_j \exp(s_j)}$$

Softmax function

Maximize probability of correct class

$$L_i = -\log P(Y = y_i \mid X = x_i)$$

Putting it all together

$$L_i = -\log\left(\frac{\exp(s_{y_i})}{\sum_j \exp(s_j)}\right)$$

Q: If all scores are small random values, what is the loss?

Cross-Entropy Loss
Multinomial Logistic Regression

cracker	**3.2**
mug	5.1
sugar	-1.7

Want to interpret raw classifier scores as **probabilities**

$$s = f(x_i; W) \qquad P(Y = k \mid X = x_i) = \frac{\exp(s_k)}{\sum_j \exp(s_j)}$$ Softmax function

Maximize probability of correct class

$$L_i = -\log P(Y = y_i \mid X = x_i)$$

Putting it all together

$$L_i = -\log\left(\frac{\exp(s_{y_i})}{\sum_j \exp(s_j)}\right)$$

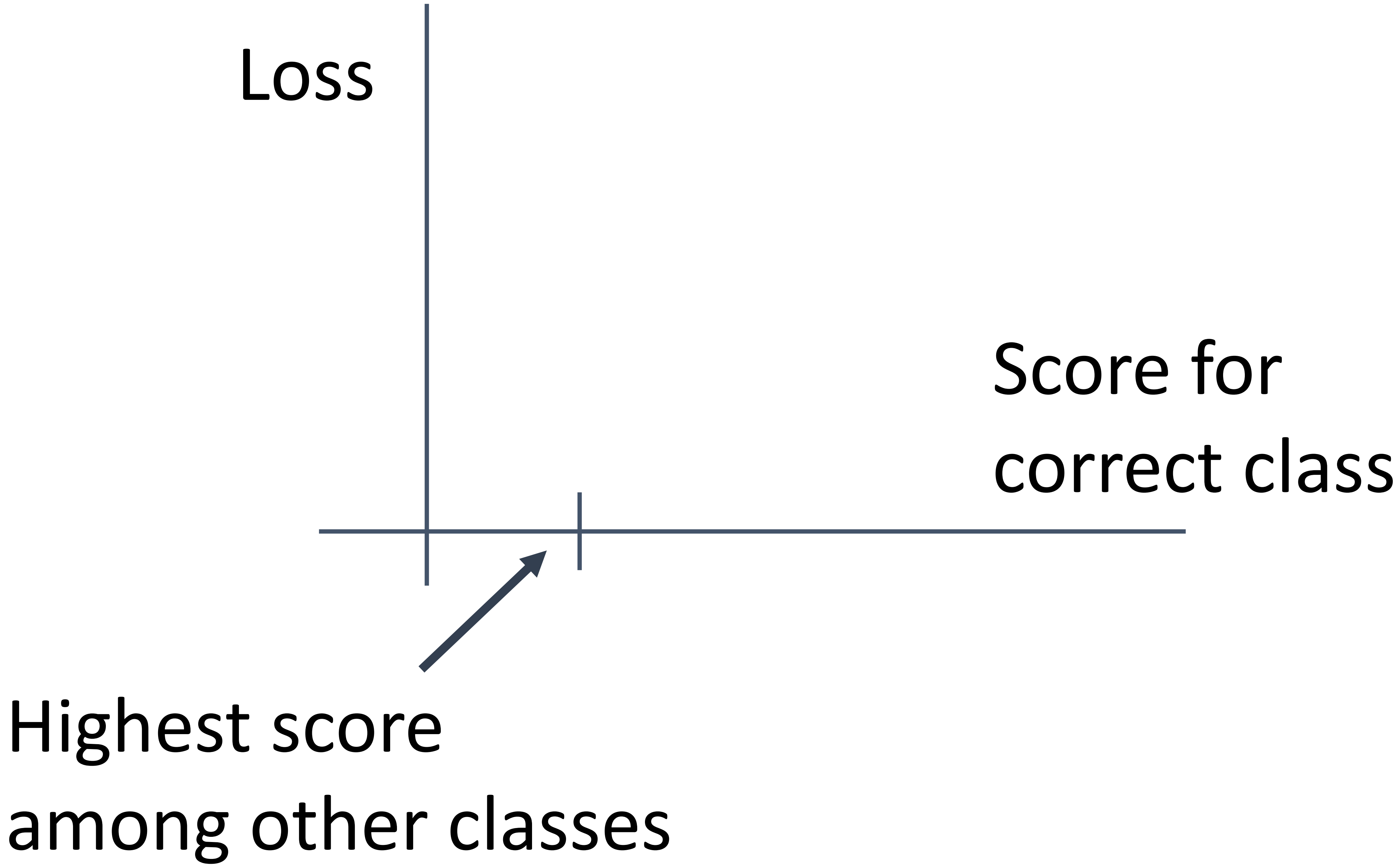
Q: If all scores are small random values, what is the loss?

A: $-\log(\frac{1}{C})$

$$\log(\frac{1}{10}) \approx 2.3$$

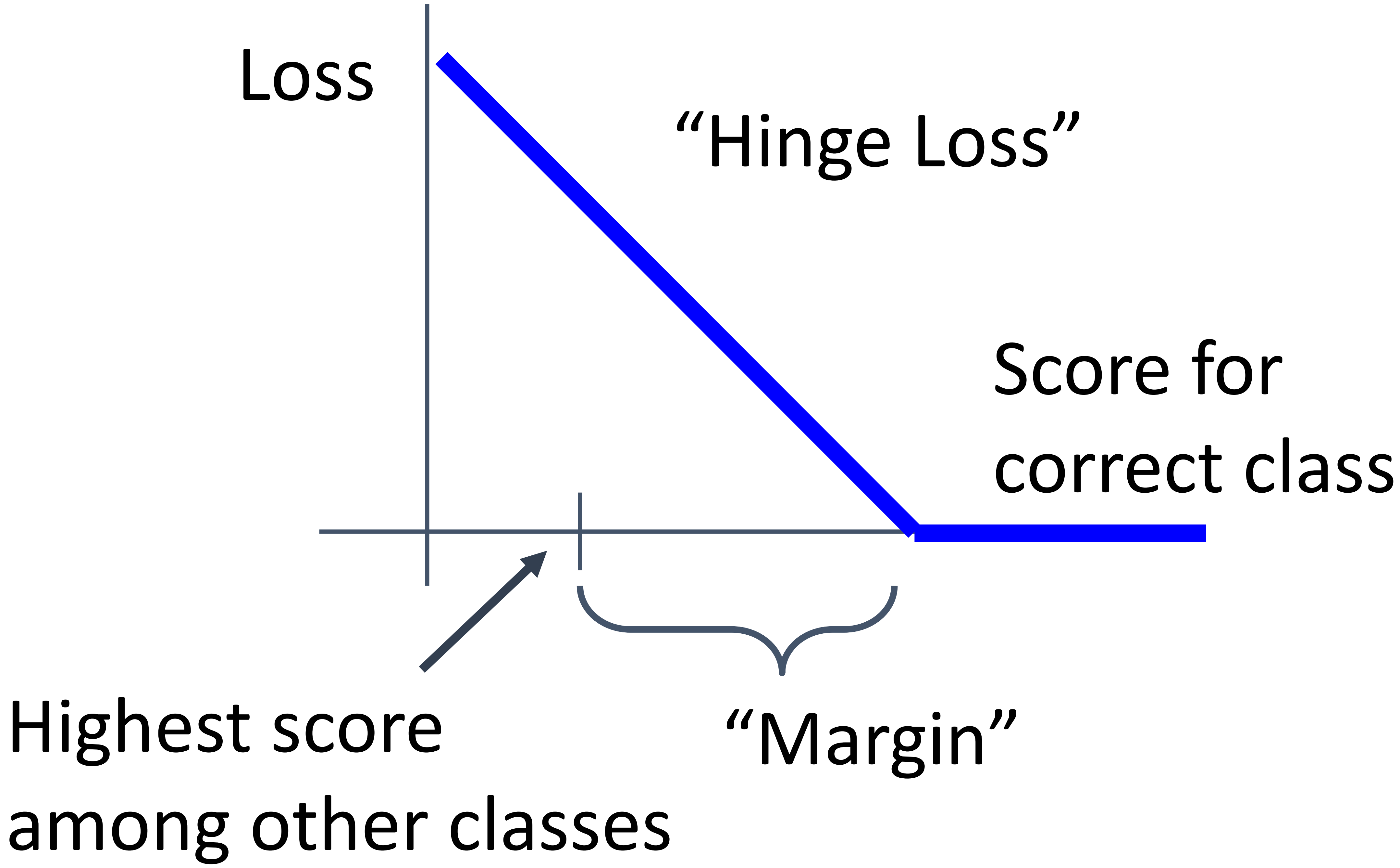
Multiclass SVM Loss

"The score of the correct class should be higher than all the other scores"



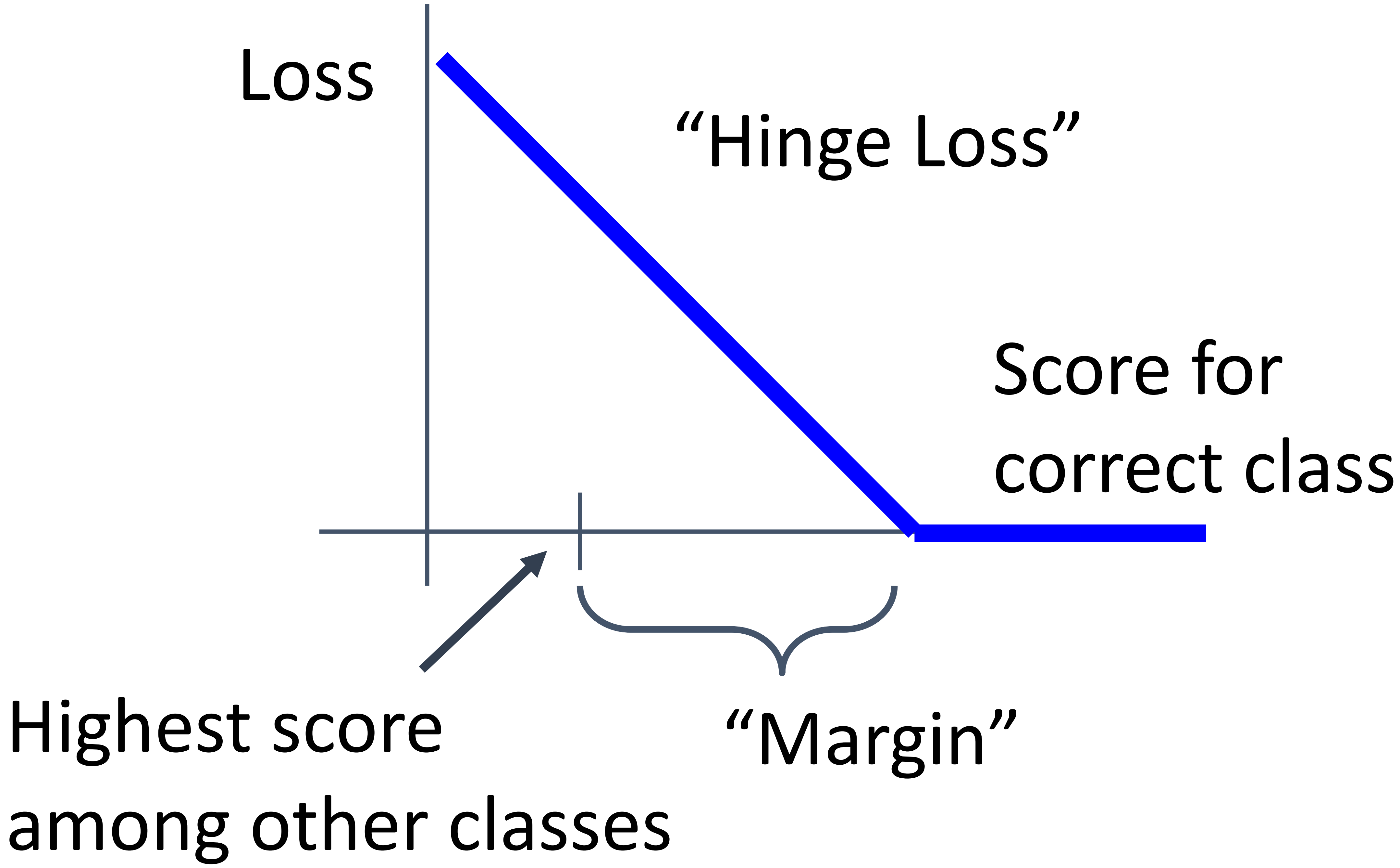
Multiclass SVM Loss

”The score of the correct class should be higher than all the other scores”



Multiclass SVM Loss

"The score of the correct class should be higher than all the other scores"



Given an example (x_i, y_i)
(x_i is image, y_i is label)

Let $s = f(x_i, W)$ be scores

Then the SVM loss has the form:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

Multiclass SVM Loss


cracker	**3.2**	1.3	2.2
mug	5.1	**4.9**	2.5
sugar	-1.7	2.0	**-3.1**

Given an example (x_i, y_i)
(x_i is image, y_i is label)

Let $s = f(x_i, W)$ be scores

Then the SVM loss has the form:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

Multiclass SVM Loss


cracker	**3.2**	1.3	2.2
mug	5.1	**4.9**	2.5
sugar	-1.7	2.0	**-3.1**
Loss	2.9		

Given an example (x_i, y_i)
(x_i is image, y_i is label)

Let $s = f(x_i, W)$ be scores

Then the SVM loss has the form:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

= max(0, 5.1 - 3.2 + 1)
+ max(0, -1.7 - 3.2 + 1)
= max(0, 2.9) + max(0, -3.9)
= 2.9 + 0
= 2.9

Multiclass SVM Loss


cracker	**3.2**	1.3	2.2
mug	5.1	**4.9**	2.5
sugar	-1.7	2.0	**-3.1**
Loss	2.9	0	

Given an example (x_i, y_i)
(x_i is image, y_i is label)

Let $s = f(x_i, W)$ be scores

Then the SVM loss has the form:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

= max(0, 1.3 - 4.9 + 1)
+max(0, 2.0 - 4.9 + 1)
= max(0, -2.6) + max(0, -1.9)
= 0 + 0
= 0

Multiclass SVM Loss


cracker	**3.2**	1.3	2.2
mug	5.1	**4.9**	2.5
sugar	-1.7	2.0	**-3.1**
Loss	2.9	0	12.9

Given an example (x_i, y_i)
(x_i is image, y_i is label)

Let $s = f(x_i, W)$ be scores

Then the SVM loss has the form:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

= max(0, 2.2 - (-3.1) + 1)
+max(0, 2.5 - (-3.1) + 1)
= max(0, 6.3) + max(0, 6.6)
= 6.3 + 6.6
= 12.9

Multiclass SVM Loss


cracker	**3.2**	1.3	2.2
mug	5.1	**4.9**	2.5
sugar	-1.7	2.0	**-3.1**
Loss	2.9	0	12.9

Given an example (x_i, y_i)
(x_i is image, y_i is label)

Let $s = f(x_i, W)$ be scores

Then the SVM loss has the form:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

Loss over the dataset is:

L = (2.9 + 0.0 + 12.9) / 3
= 5.27

Multiclass SVM Loss


cracker	**3.2**	1.3	2.2
mug	5.1	**4.9**	2.5
sugar	-1.7	2.0	**-3.1**
Loss	2.9	0	12.9

Given an example (x_i, y_i)
(x_i is image, y_i is label)

Let $s = f(x_i, W)$ be scores

Then the SVM loss has the form:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

Q: What happens to the loss if the scores for the mug image change a bit?

Multiclass SVM Loss


cracker	**3.2**	1.3	2.2
mug	5.1	**4.9**	2.5
sugar	-1.7	2.0	**-3.1**
Loss	2.9	0	12.9

Given an example (x_i, y_i)
(x_i is image, y_i is label)

Let $s = f(x_i, W)$ be scores

Then the SVM loss has the form:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

Q2: What are the min and max possible loss?

Multiclass SVM Loss


cracker	**3.2**	1.3	2.2
mug	5.1	**4.9**	2.5
sugar	-1.7	2.0	**-3.1**
Loss	2.9	0	12.9

Given an example (x_i, y_i)
(x_i is image, y_i is label)

Let $s = f(x_i, W)$ be scores

Then the SVM loss has the form:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

Q3: If all the scores were random, what loss would we expect?

Cross-Entropy vs SVM Loss

$$L_i = -\log\left(\frac{\exp(s_{y_i})}{\sum_j \exp(s_j)}\right)$$

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

assume scores:
[10, -2, 3]
[10, 9, 9]
[10, -100, -100]
and $\boxed{y_i = 0}$

Q: What is cross-entropy loss?
What is SVM loss?

Cross-Entropy vs SVM Loss

$$L_i = -\log\left(\frac{\exp(s_{y_i})}{\sum_j \exp(s_j)}\right)$$

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

assume scores:

[10, -2, 3]

[10, 9, 9]

[10, -100, -100]

and $\boxed{y_i = 0}$

Q: What is cross-entropy loss?
What is SVM loss?

A: Cross-entropy loss > 0
SVM loss = 0

Cross-Entropy vs SVM Loss

$$L_i = -\log\left(\frac{\exp(s_{y_i})}{\sum_j \exp(s_j)}\right)$$

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

assume scores:
[10, -2, 3]
[10, 9, 9]
[10, -100, -100]
and $\boxed{y_i = 0}$

Q: What happens to each loss if I slightly change the scores of the last datapoint?

Cross-Entropy vs SVM Loss

$$L_i = -\log\left(\frac{\exp(s_{y_i})}{\sum_j \exp(s_j)}\right)$$

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

assume scores:

[10, -2, 3]

[10, 9, 9]

[10, -100, -100]

and $\boxed{y_i = 0}$

Q: What happens to each loss if I slightly change the scores of the last datapoint?

A: Cross-entropy loss will change;

SVM loss will stay the same for 1st and 3rd example

SVM loss will change for the 2nd

Cross-Entropy vs SVM Loss

$$L_i = -\log\left(\frac{\exp(s_{y_i})}{\sum_j \exp(s_j)}\right)$$

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

assume scores:

[10, -2, 3]

[10, 9, 9]

[10, -100, -100]

and $\boxed{y_i = 0}$

Q: What happens to each loss if I double the score of the correct class from 10 to 20?

Cross-Entropy vs SVM Loss

$$L_i = -\log\left(\frac{\exp(s_{y_i})}{\sum_j \exp(s_j)}\right)$$

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

assume scores:

[10, -2, 3]

[10, 9, 9]

[10, -100, -100]

and $y_i = 0$

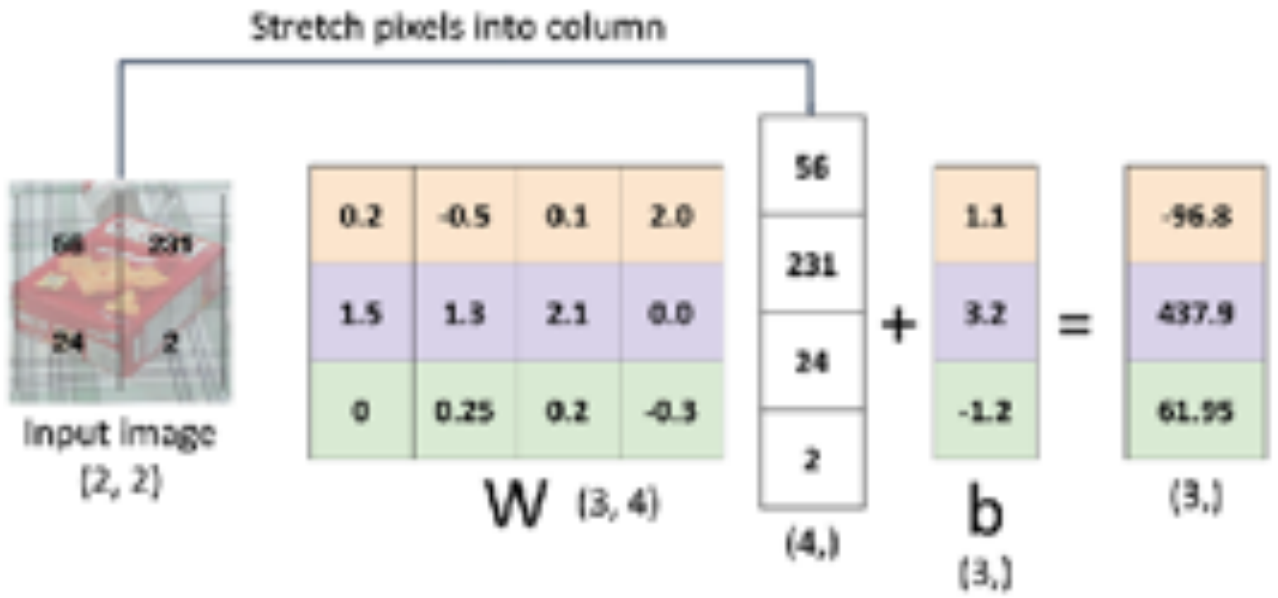
Q: What happens to each loss if I double the score of the correct class from 10 to 20?

A: Cross-entropy loss will decrease, SVM loss still 0

Recap—Three Ways to Interpret Linear Classifiers

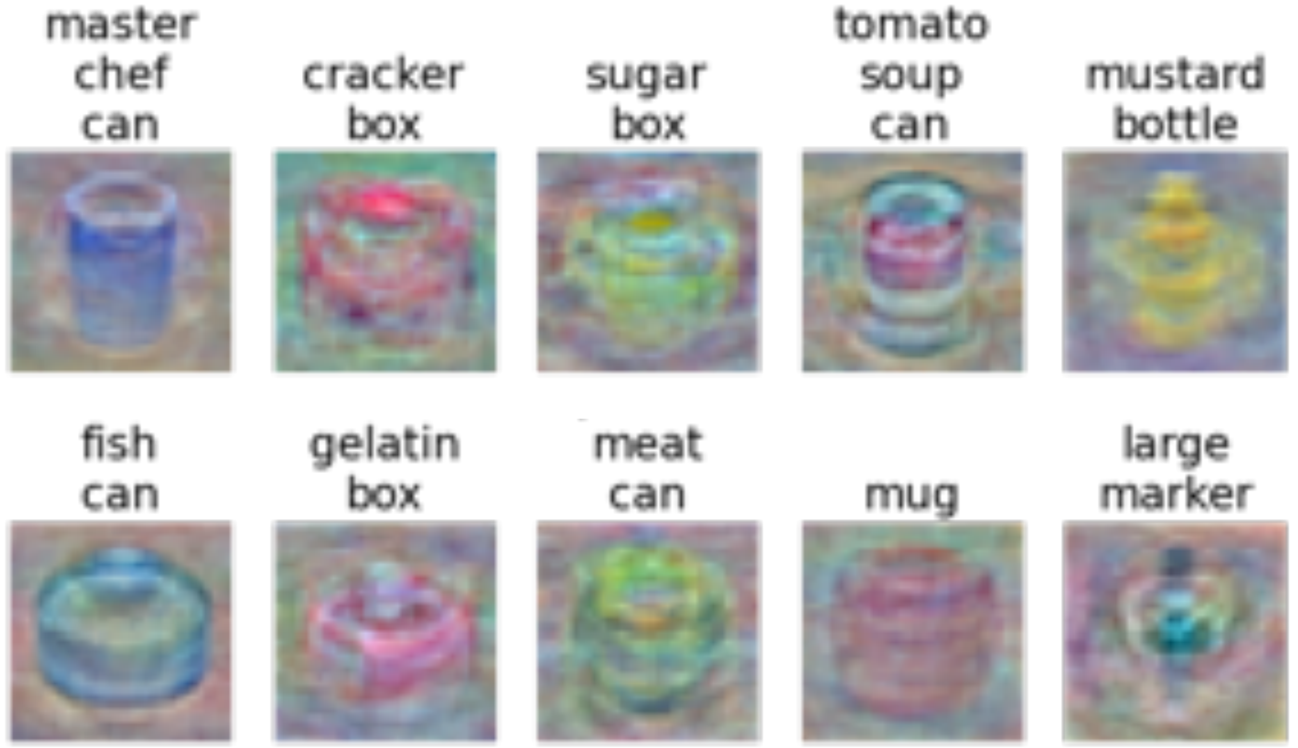
Algebraic Viewpoint

$f(x,W) = Wx$



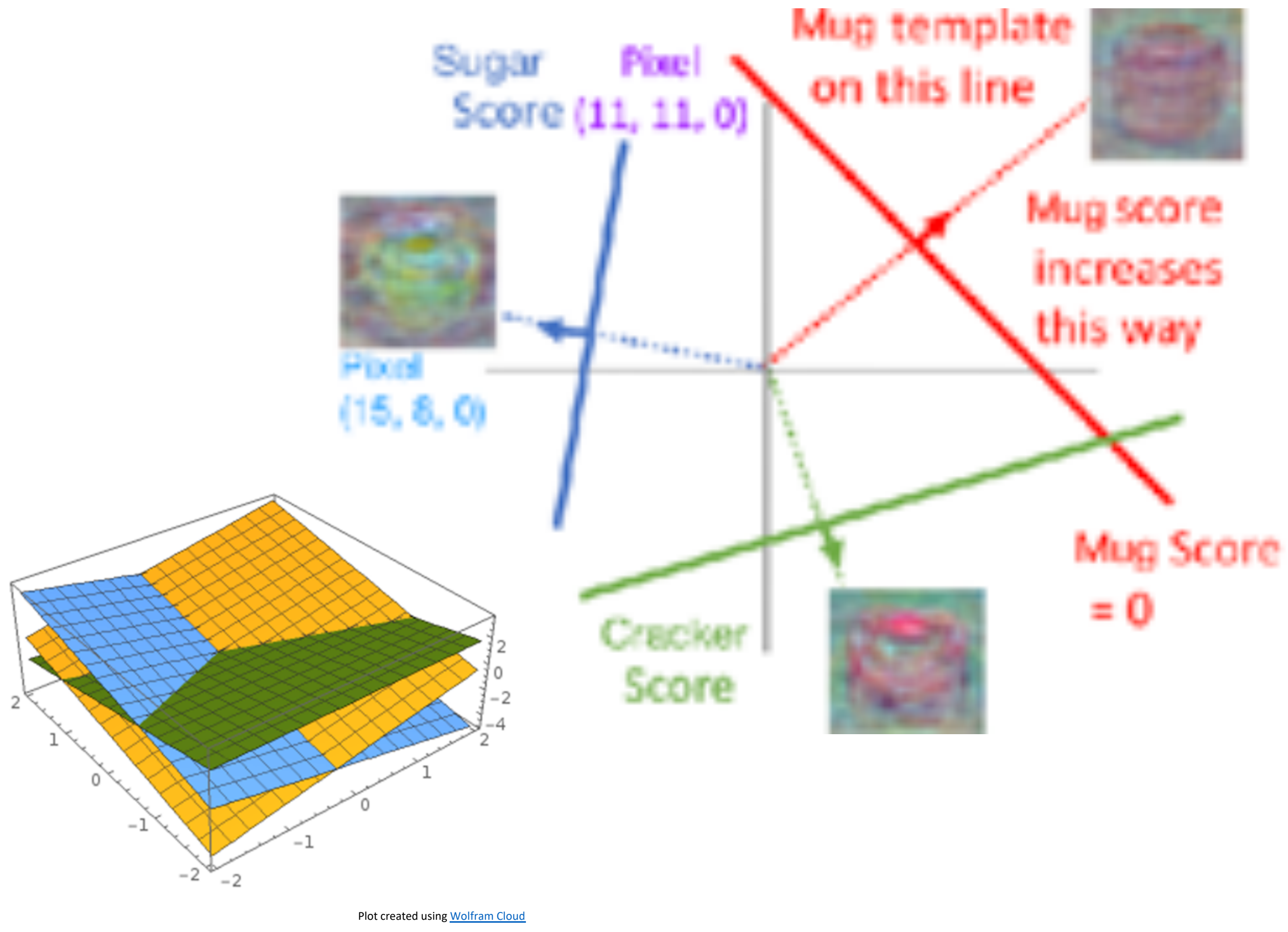
Visual Viewpoint

One template
per class



Geometric Viewpoint

Hyperplanes
cutting up space



Recap—Loss Functions Quantify Preferences

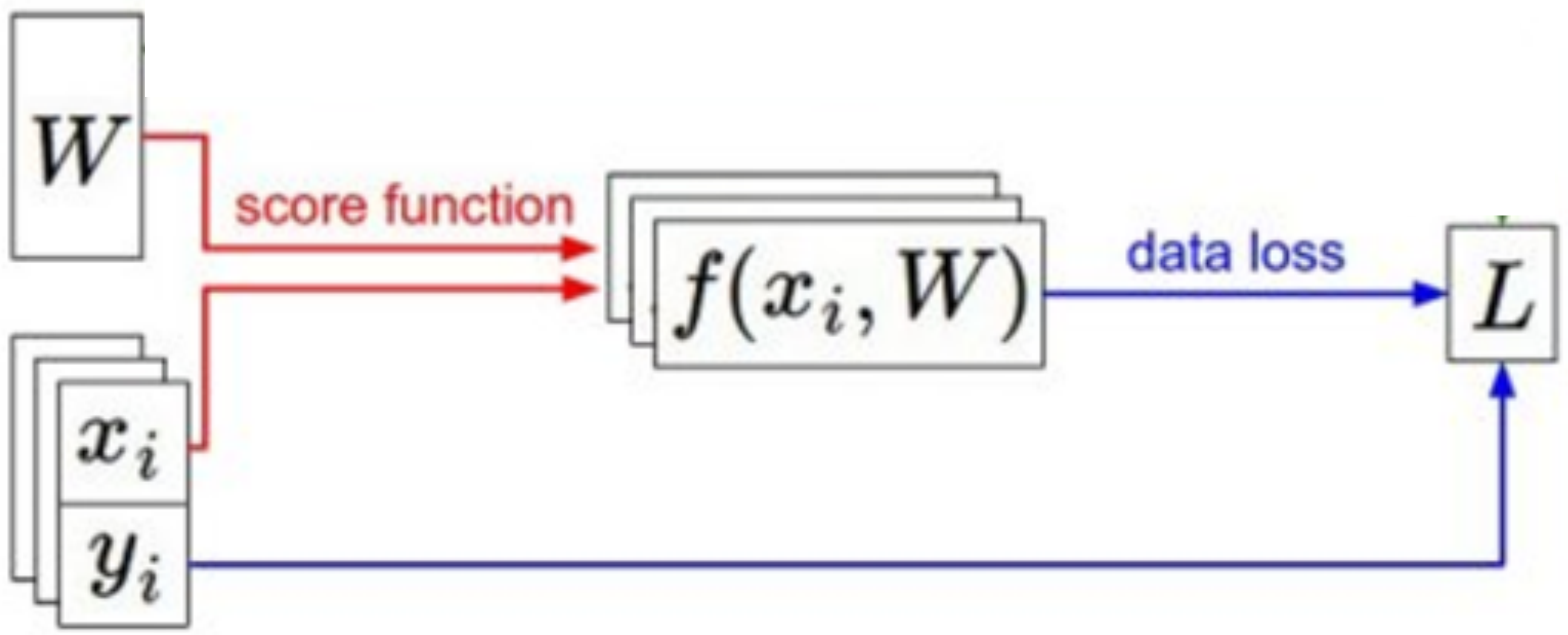
- We have some dataset of (x, y)
- We have a **score function:**
- We have a **loss function**:

$$s = f(x; W, b) = Wx + b$$

Linear classifier

Softmax: $L_i = -\log\left(\frac{\exp(s_{y_i})}{\sum_j \exp(s_j)}\right)$

SVM: $L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$



Recap—Loss Functions Quantify Preferences

- We have some dataset of (x, y)
- We have a **score function:**
- We have a **loss function**:

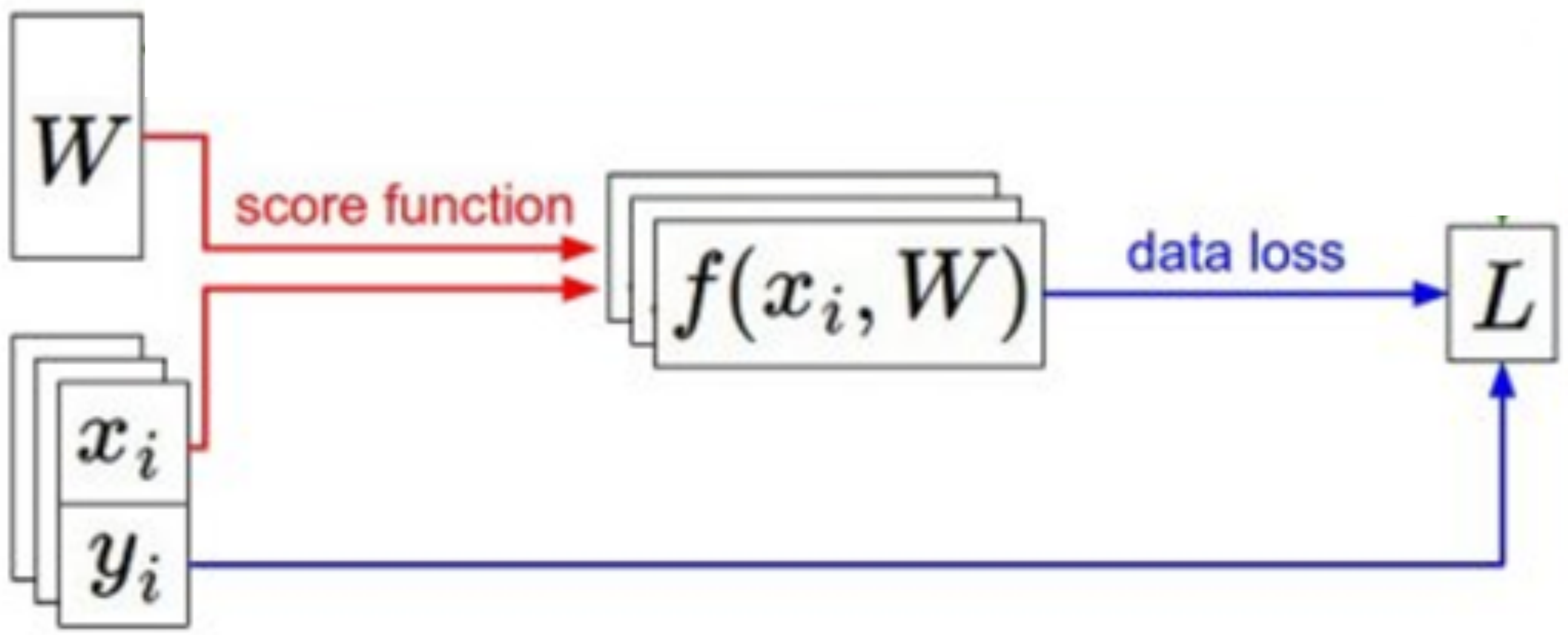
Q: How do we find the best W,b?

$$s = f(x; W, b) = Wx + b$$

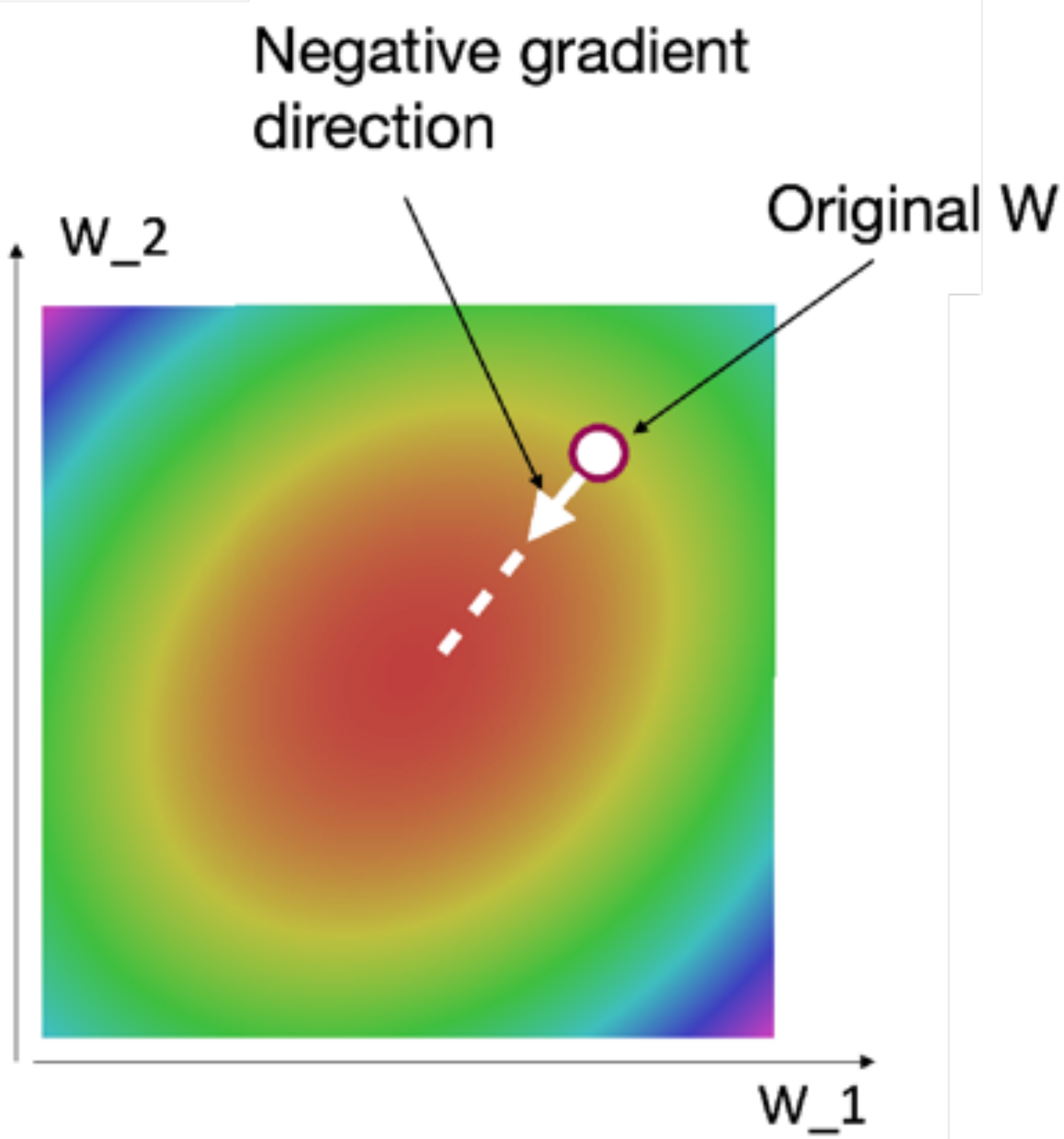
Linear classifier

Softmax: $L_i = -\log\left(\frac{\exp(s_{y_i})}{\sum_j \exp(s_j)}\right)$

SVM: $L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$



Next time: Regularization + Optimization



DeepRob

Lecture 3
Linear Classifiers
University of Michigan and University of Minnesota