Modul 8: Learning

02 Supervised Learning

Masayu Leylia Khodra (masayu@informatika.org)

KK IF – Teknik Informatika – STEI ITB

Inteligensi Buatan
(*Artificial Intelligence*)

Supervised Learning: Feedback I/O Pairs

Case	Age	Prescription	Astigmatism	Tear Production	Lens
1	young	myope	not astigmatic	reduced	none
2	young	myope	not astigmatic	normal	soft
3	young	myope	astigmatic	reduced	none
4	young	myope	astigmatic	normal	hard
5	young	hypermetrope	not astigmatic	reduced	none
6	young	hypermetrope	not astigmatic	normal	soft
7	young	hypermetrope	astigmatic	reduced	none
8	young	hypermetrope	astigmatic	normal	hard
9	pre-presbyopic	myope	not astigmatic	reduced	none
10	pre-presbyopic	myope	not astigmatic	normal	soft
11	pre-presbyopic	myope	astigmatic	reduced	none
12	pre-presbyopic	myope	astigmatic	normal	hard
13	pre-presbyopic	hypermetrope	not astigmatic	reduced	none
14	pre-presbyopic	hypermetrope	not astigmatic	normal	soft
15	pre-presbyopic	hypermetrope	astigmatic	reduced	none
16	pre-presbyopic	hypermetrope	astigmatic	normal	none
17	presbyopic	myope	not astigmatic	reduced	none
18	presbyopic	myope	not astigmatic	normal	none
19	presbyopic	myope	astigmatic	reduced	none
20	presbyopic	myope	astigmatic	normal	hard
21	presbyopic	hypermetrope	not astigmatic	reduced	none
22	presbyopic	hypermetrope	not astigmatic	normal	soft
23	presbyopic	hypermetrope	astigmatic	reduced	none
24	presbyopic	hypermetrope	astigmatic	normal	none

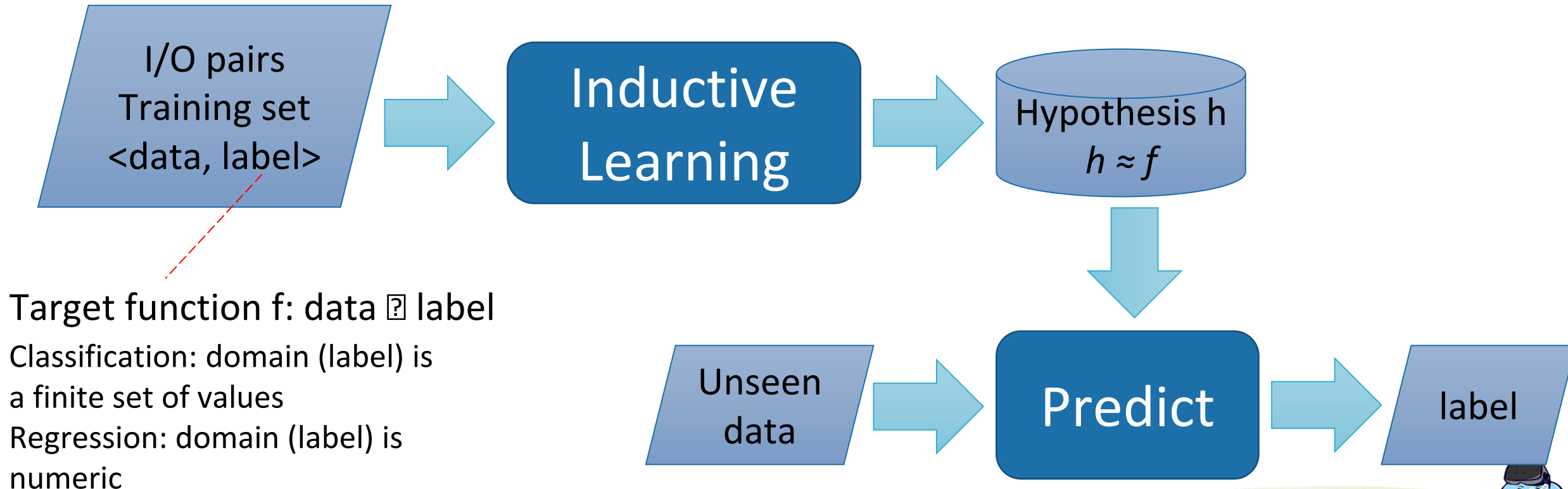x   f(x)

Supervised Learning

Learning a (possibly incorrect) general function from specific input-output pairs is called <u>inductive learning</u>

I/O pairs
Training set
<data, label>

→ **Inductive Learning** → Hypothesis h $h \approx f$

Target function f: data → label

Classification: domain (label) is a finite set of values
Regression: domain (label) is numeric

Unseen data → **Predict** → label

Fish Packing Plant

- Fish packing plant wants to automate the process of sorting incoming fish on a conveyor belt according to species.

- A: Conveyor belt for fish
- B: Conveyor belt for classified fish
- C : Robot arm for grabbing fish
- D: Machine vision system with CCD camera
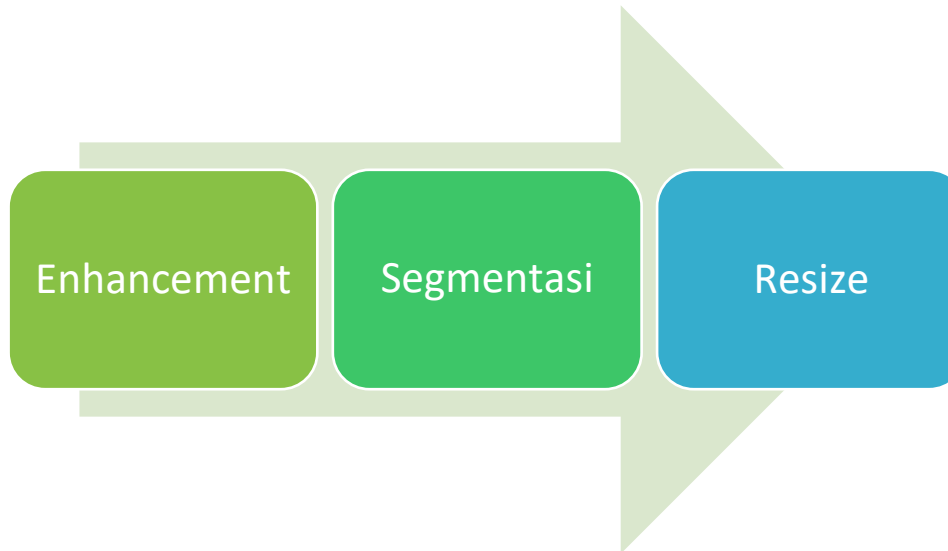- E : Computer that analyze fish image and control the robot arm

(d) CCD Camera

(b) Conveyor belt (Salmon)

(e) Computer

Robot Arm

(c)

(a) Conveyor belt

(b) Conveyor belt (Sea Bass)
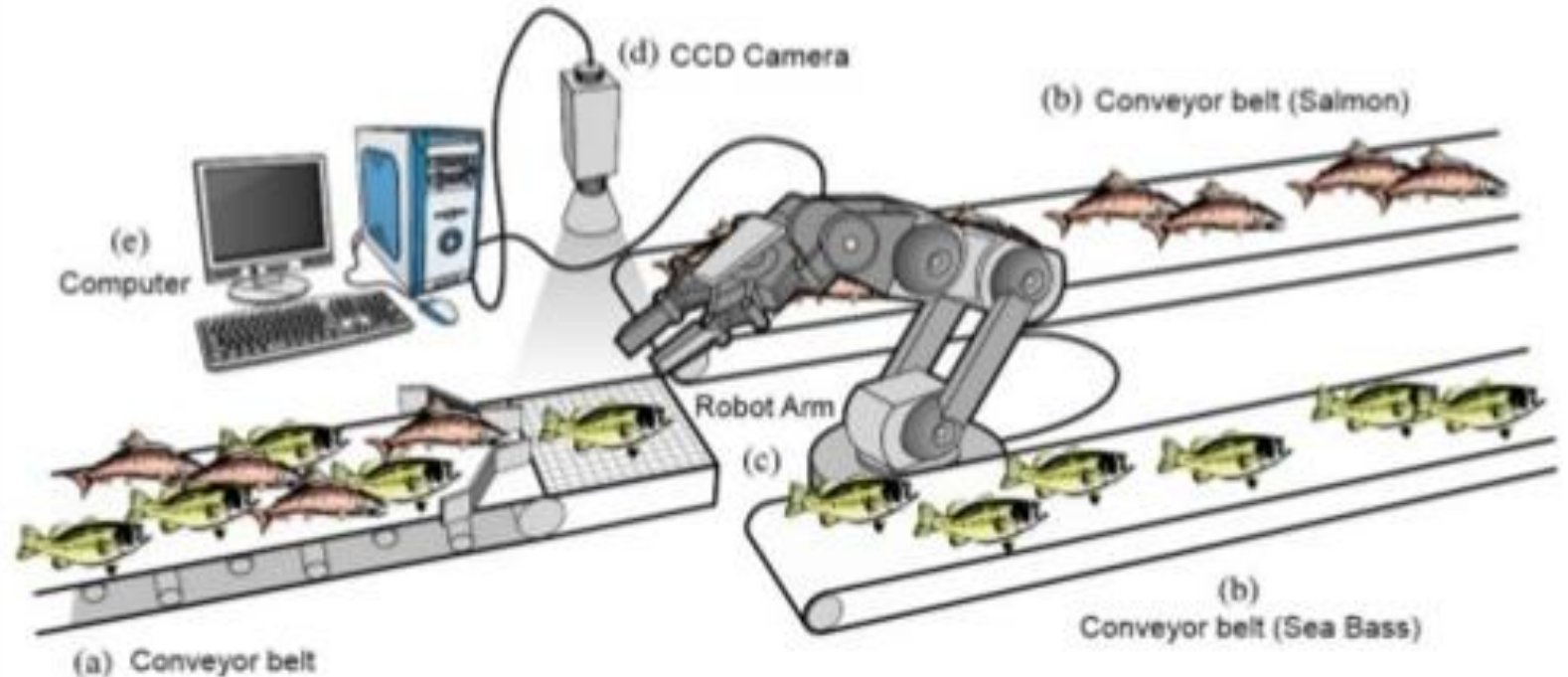
Automated Fish Classification System

Dataset Construction

Enhancement | Segmentasi | Resize

Duda dkk, 2001

EDUNEX ITB
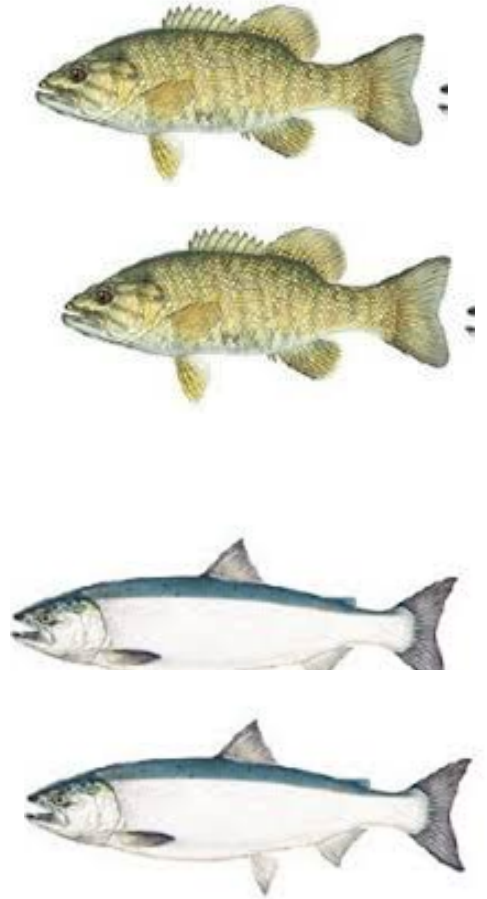
Fish features
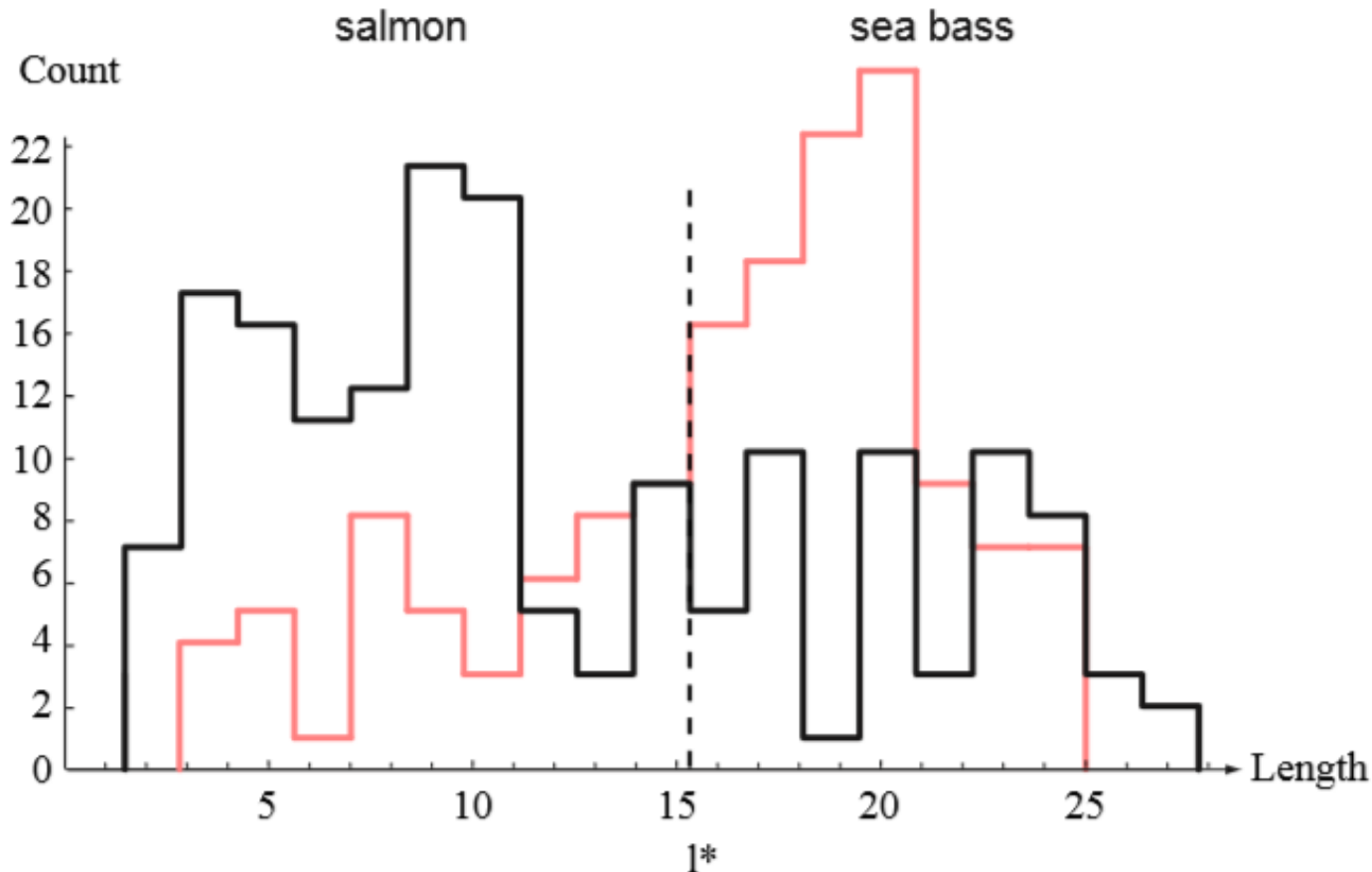
- Piksel
- Descriptor based on : color, shape, textures

Find physical differences between the two types of fish:
- Length
- Lightness
- Width
- Number and shape of fins
- Position of the mouth,
- ...

(a)   (b)
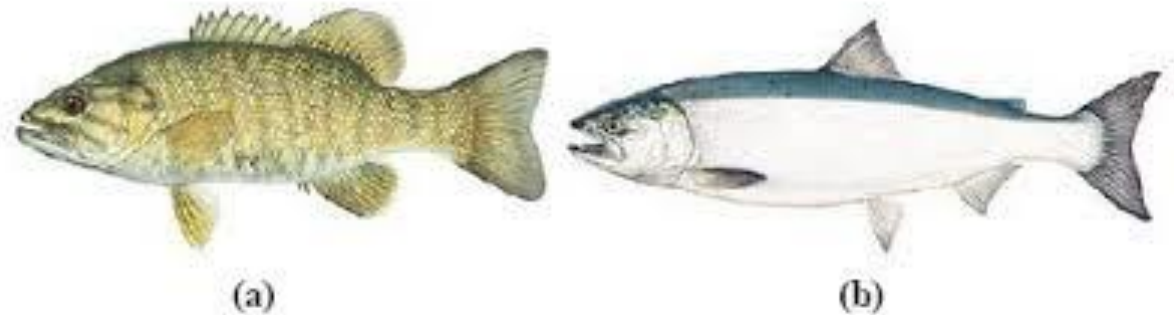
Tentative Model: Length Feature

Suppose somebody at the fish plant tells us that a sea bass is generally longer than a salmon.

From histograms, sea bass are longer than salmon, on average, but it is clear that this single criterion is quite poor;

No matter how we choose l*, we cannot reliably separate sea bass from salmon by length alone.

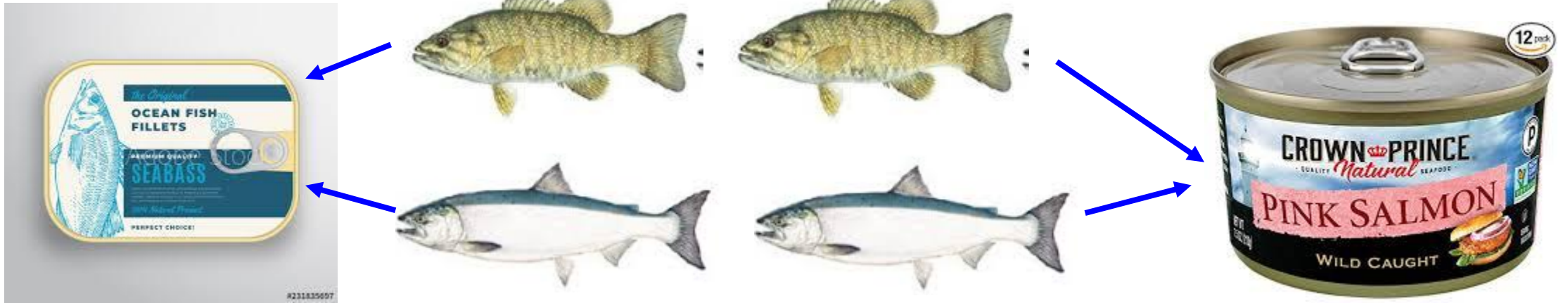The value l* marked will lead to the smallest number of errors, on average.

EDUNEX ITB

Tentative Model: Lightness Feature
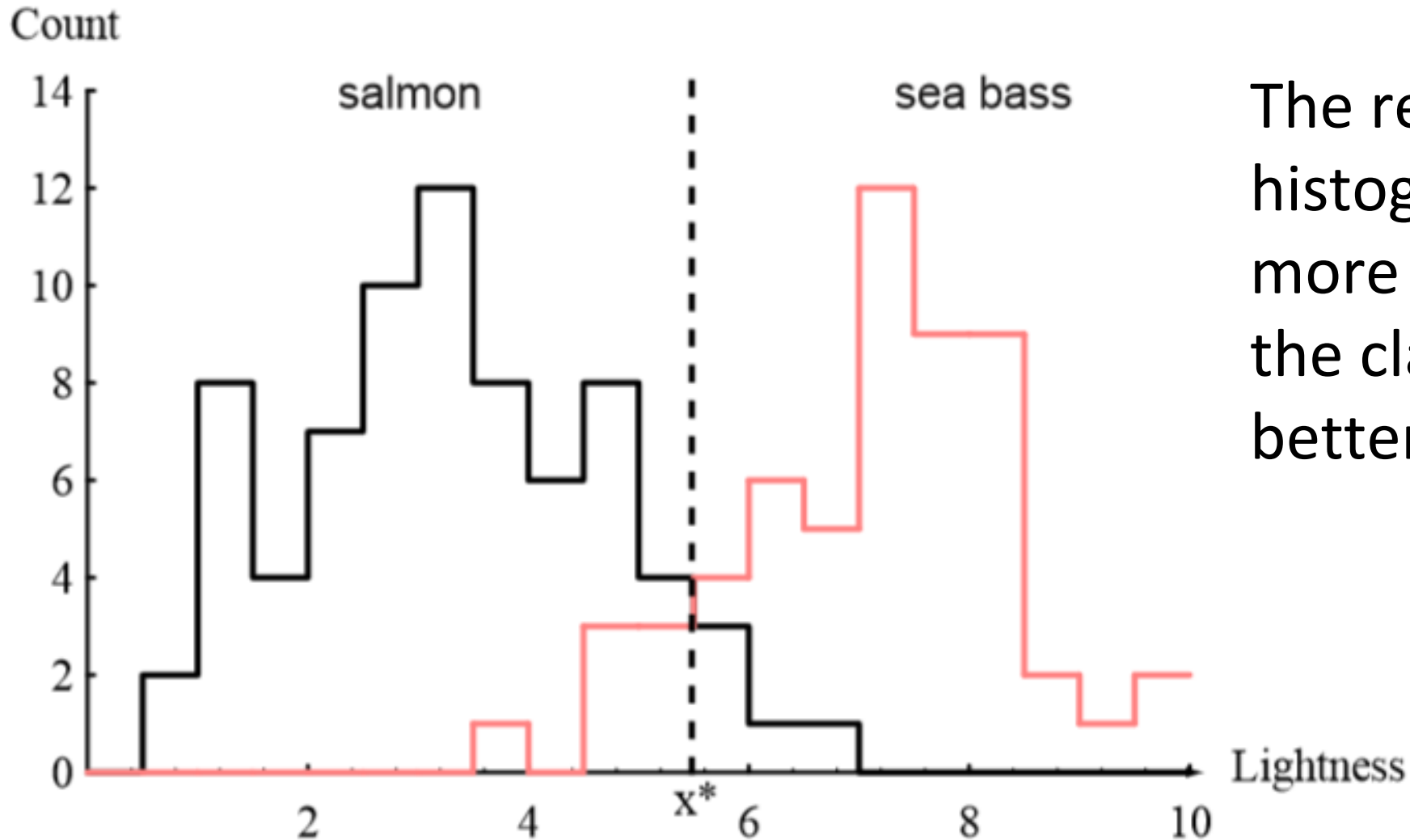
The resulting histograms are much more satisfactory — the classes are much better separated.
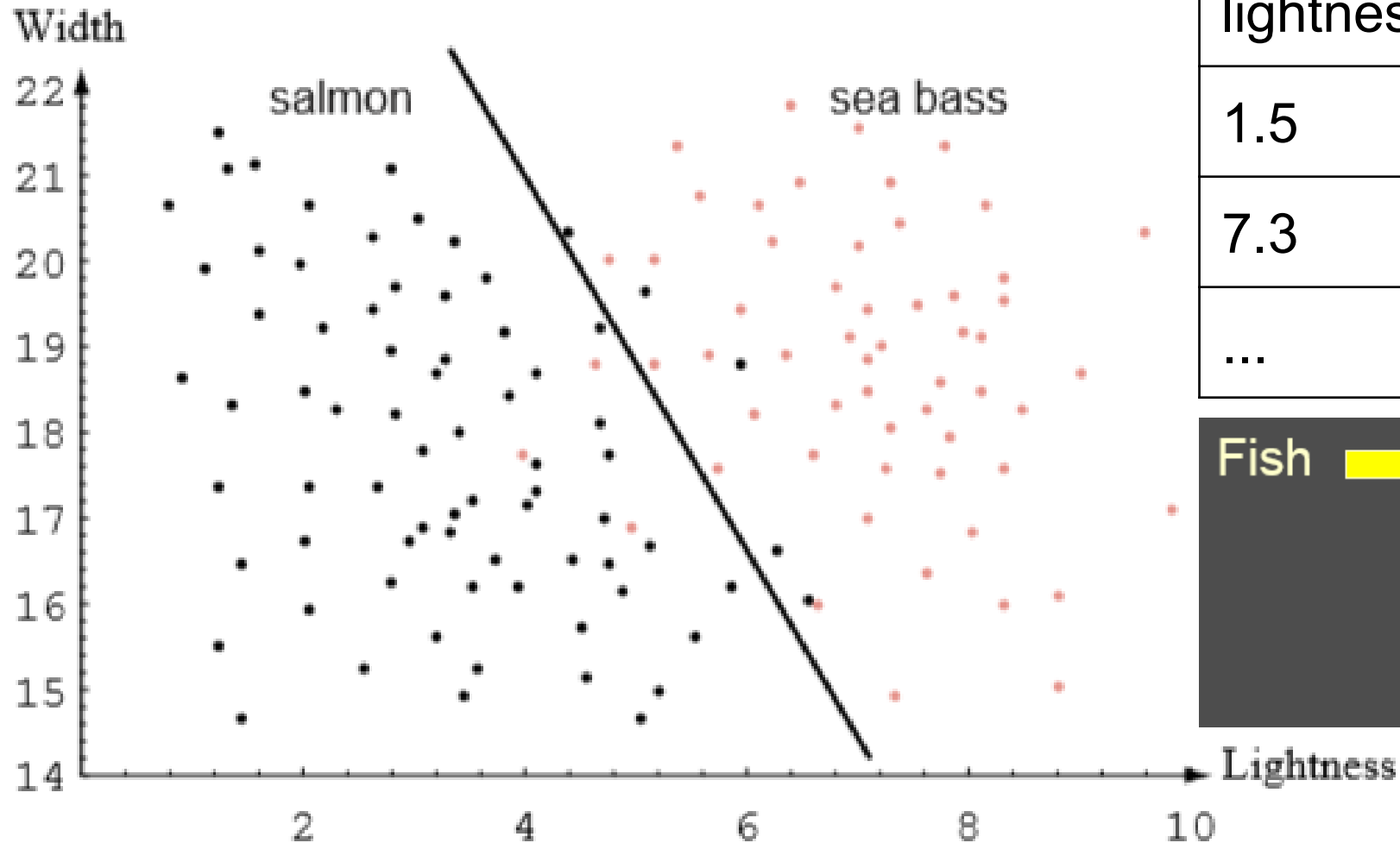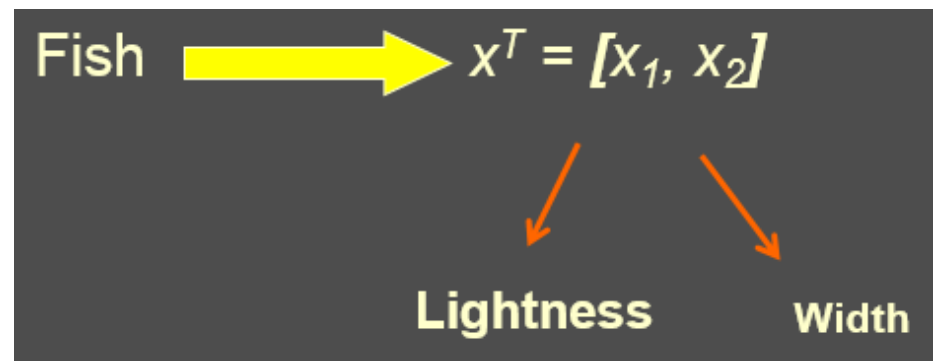
Misclassification

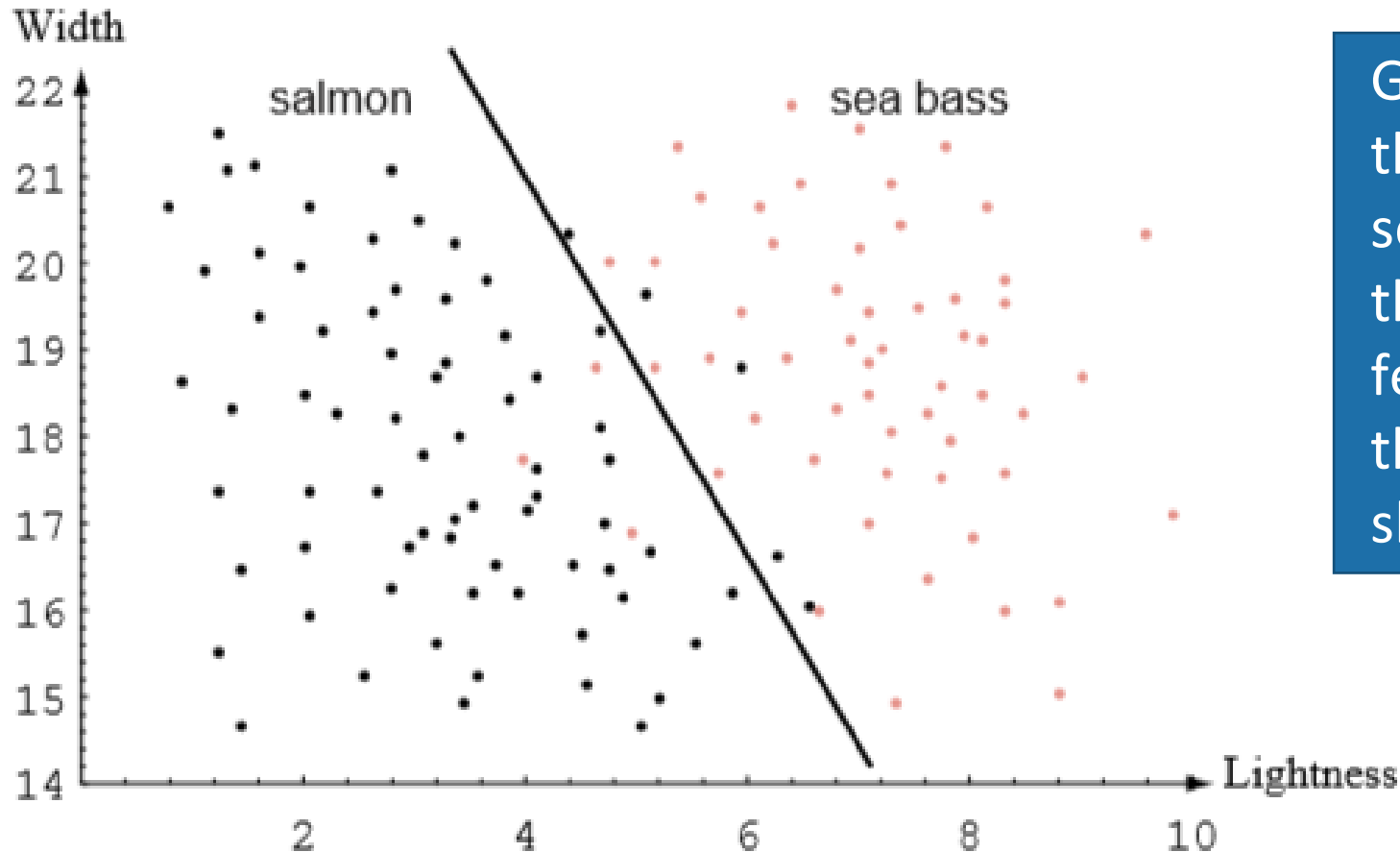salmon in "sea bass" cans vs sea bass in "salmon" cans ?

Tentative Model: Lightness and Width Feature

lightness	width	category
1.5	14.6	salmon
7.3	15	Sea bass
...		

Fish ➡ $x^T = [x_1, x_2]$
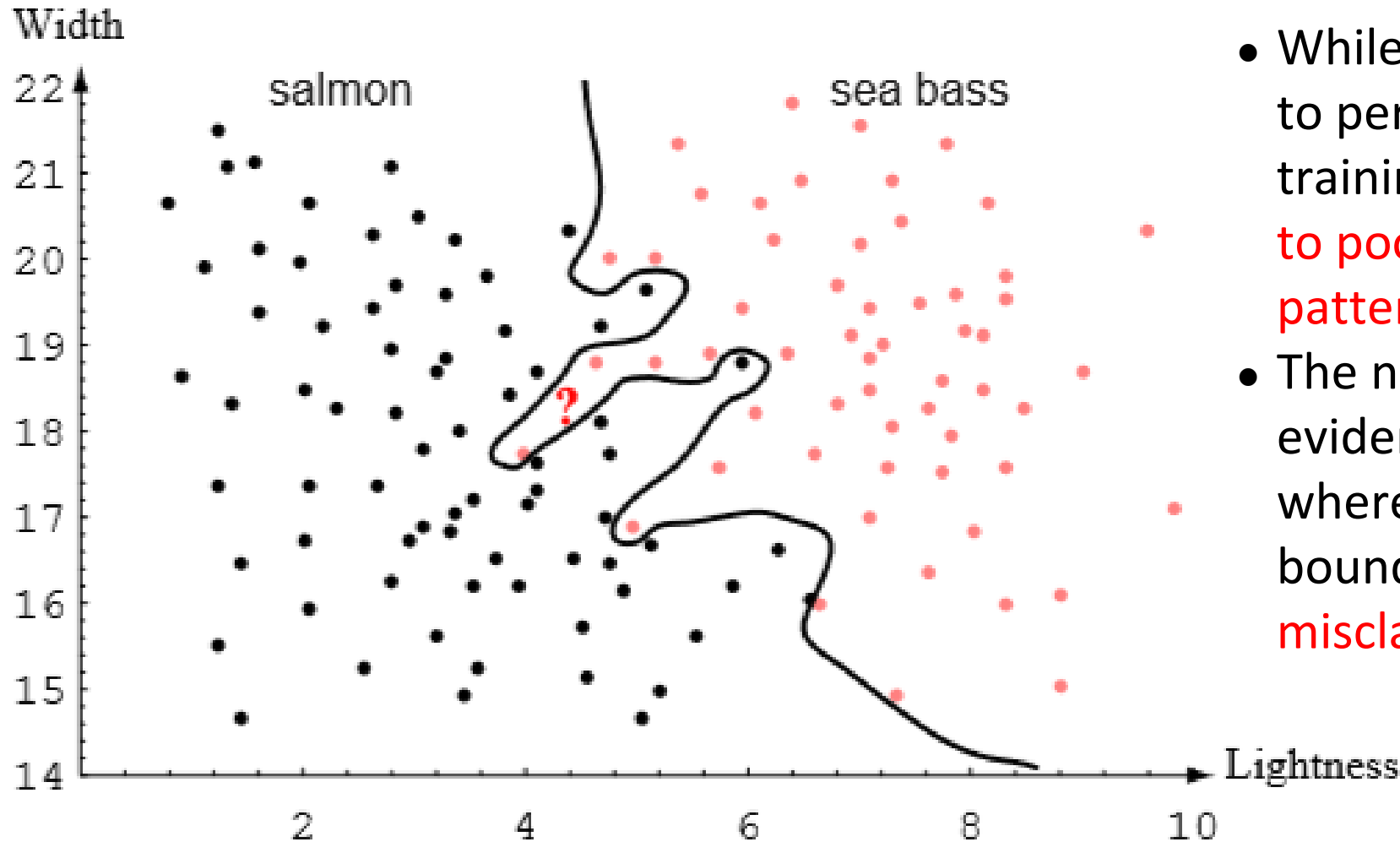
Lightness   Width

EDUNEX ITB

Rule based on 2 Features

Y=a*width+b*lightness+c

Given decision boundary, the following rule for separating the fish: Classify the fish as sea bass if its feature vector falls above the decision boundary shown, and as salmon


EDUNEX ITB

Tentative Complex Model: Training Accuracy 100%

- While such a decision may lead to perfect classification of our training samples, it would lead to poor performance on future patterns.
- The novel test point marked ? is evidently most likely a salmon, whereas the complex decision boundary shown leads it to be misclassified as a sea bass.

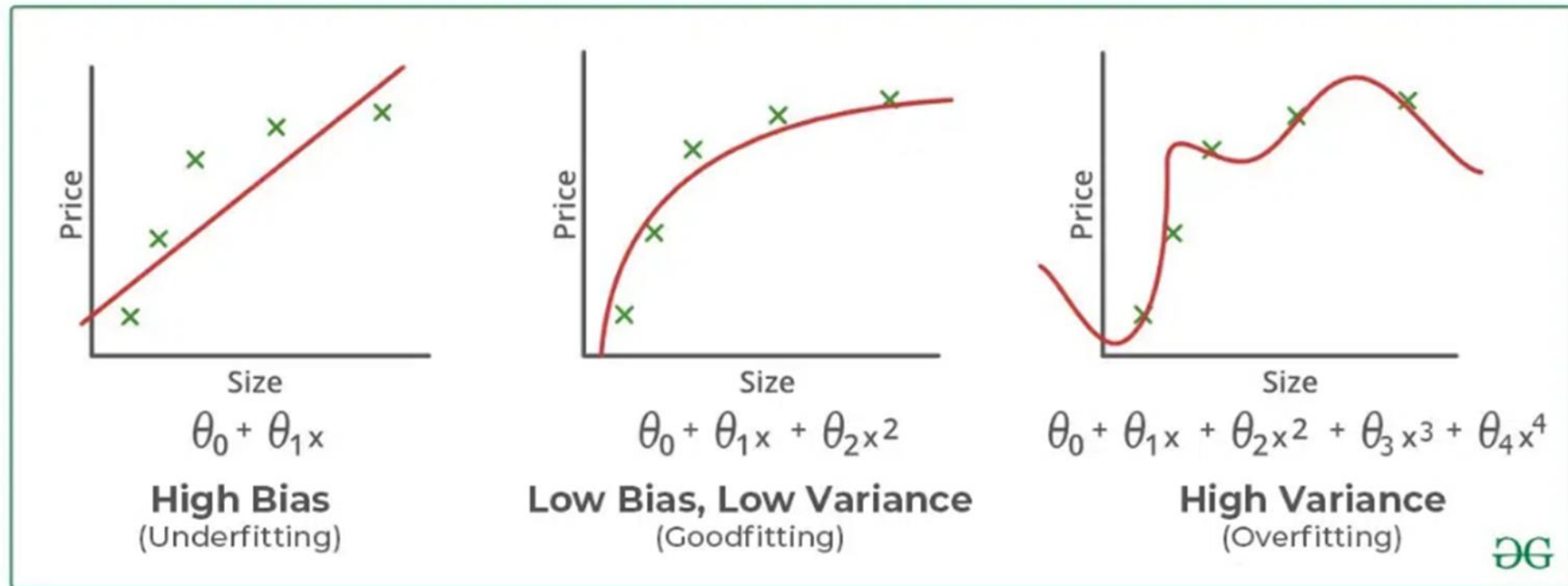# Optimal Model: Better Generalization
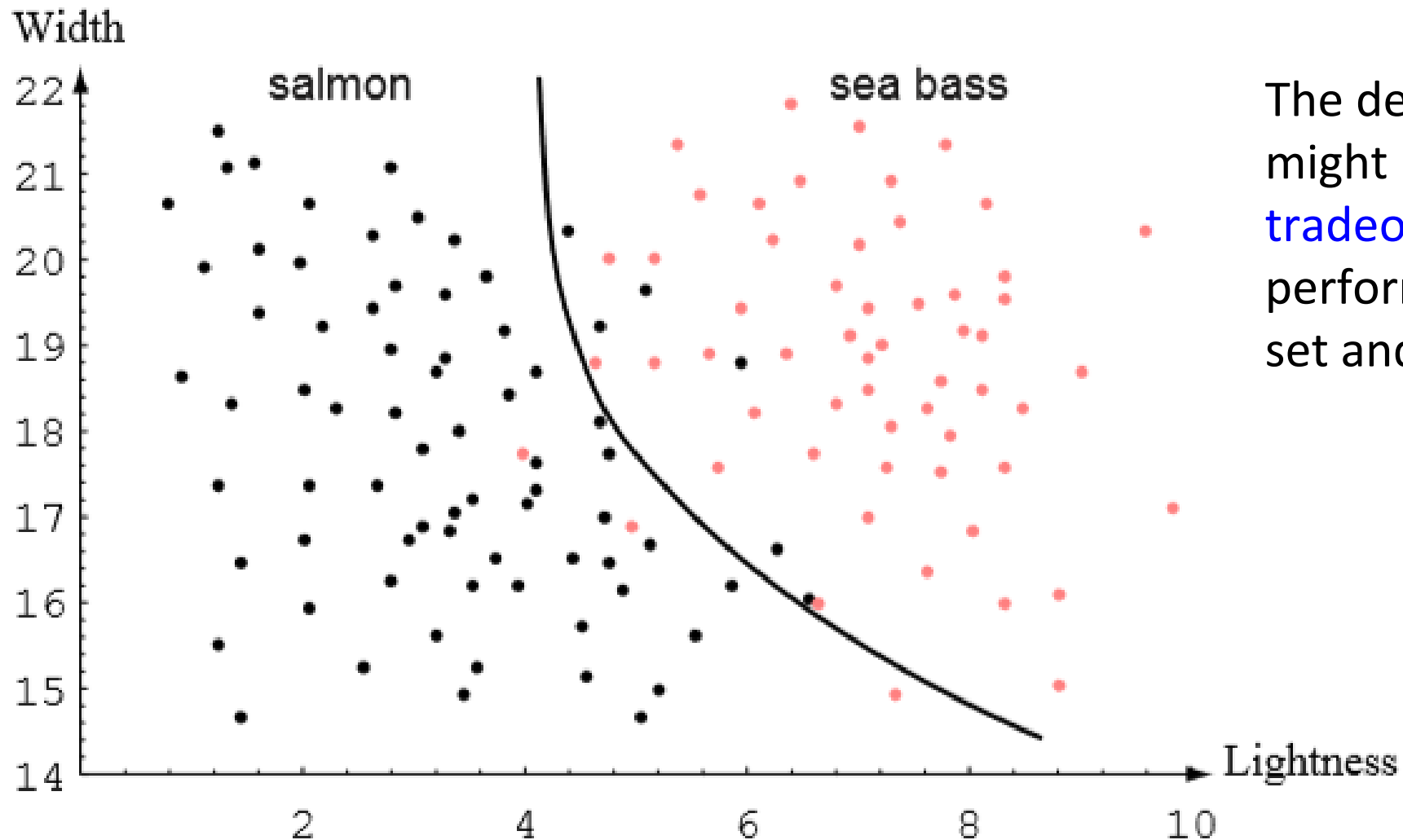
The decision boundary shown might represent the optimal tradeoff between performance on the training set and simplicity of classifier.

Bias vs Variance

High Bias (Underfitting): $\theta_0 + \theta_1 x$

Low Bias, Low Variance (Goodfitting): $\theta_0 + \theta_1 x + \theta_2 x^2$

High Variance (Overfitting): $\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

Sumber gambar: https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

EDUNEX ITB

Modeling using Supervised Learning

Training data

lightness	width	category
1.5	14.6	salmon
7.3	15	Sea bass
...		

Preprocessing and Feature Extraction

Machine Learning
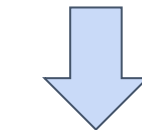
model
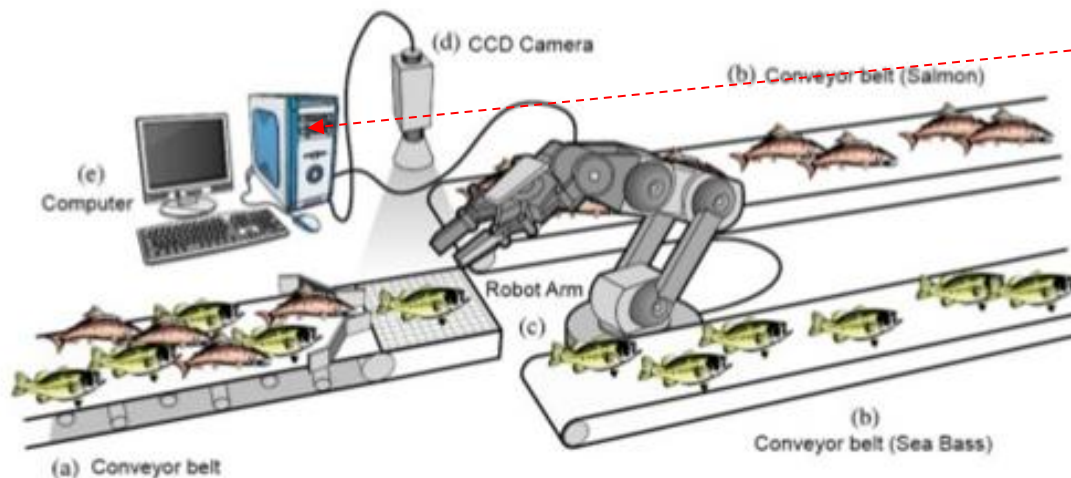
Summary

| Supervised learning: inductive learning, find hypothesis | Classification vs regression | Hypothesis to predict unseen data | Fish packing plant: case study |

Intro to Data Science →