Modul 2: Pattern Recognition System

03 Learning-based Recognition System

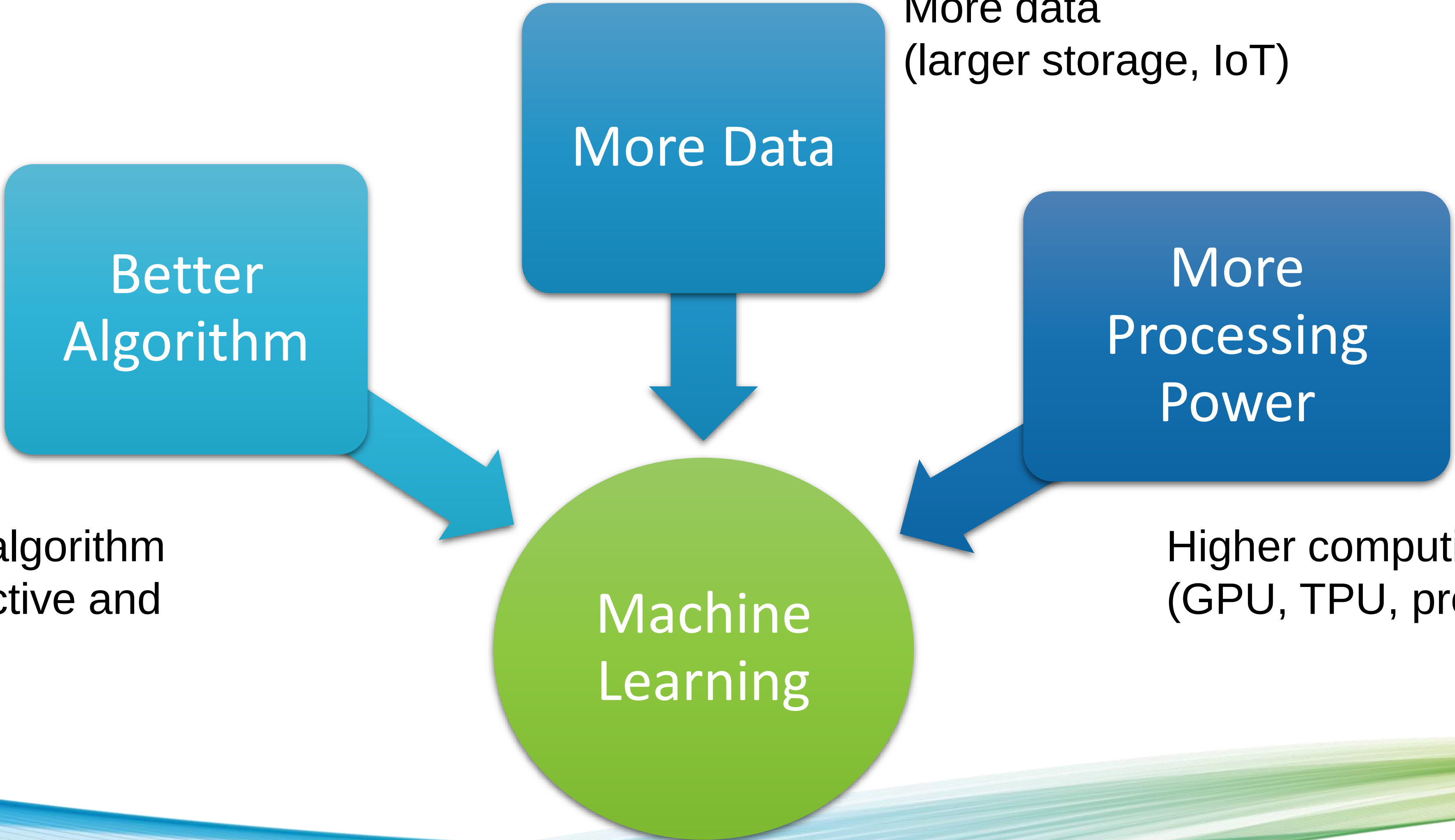
Masayu Leylia Khodra
(masayu@informatika.org)

KK IF – Teknik Informatika – STEI ITB

Pengenalan Pola
(*Pattern Recognition*)

Why Machine Learning?

More data
(larger storage, IoT)



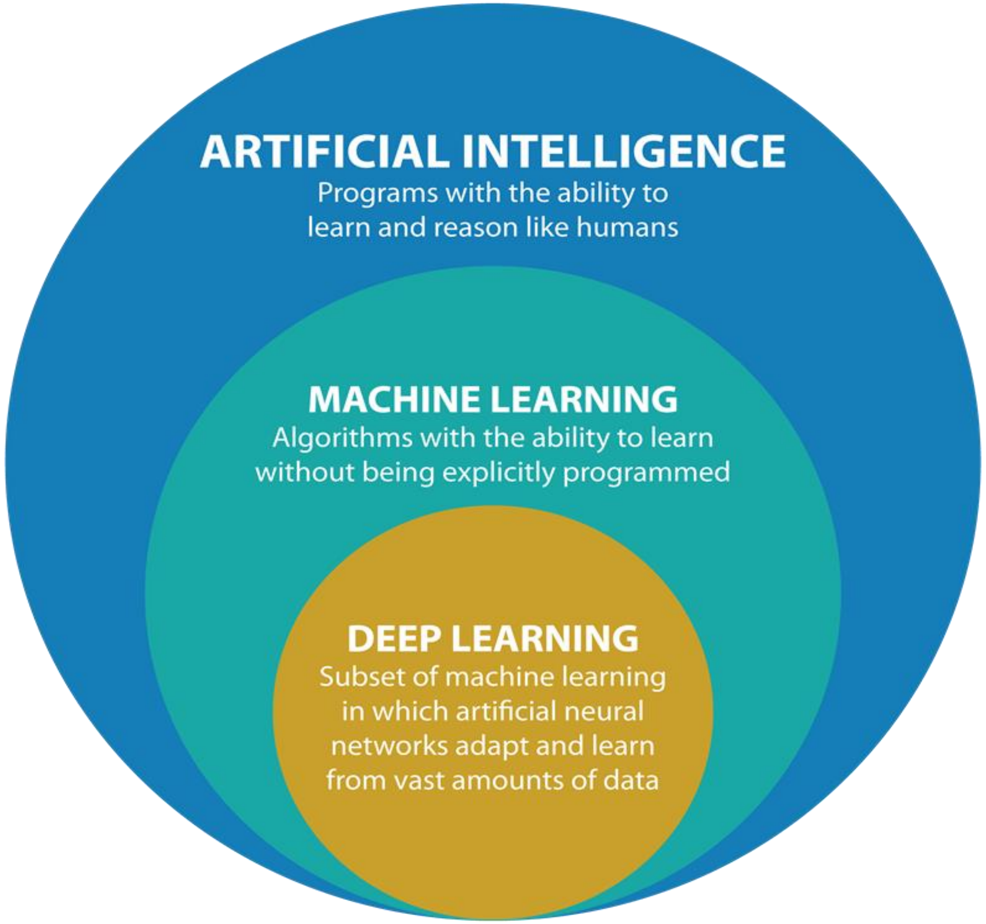
Learning algorithm more effective and efficient

Higher computing power
(GPU, TPU, prosesor ANN)

What is Machine Learning (ML) ?

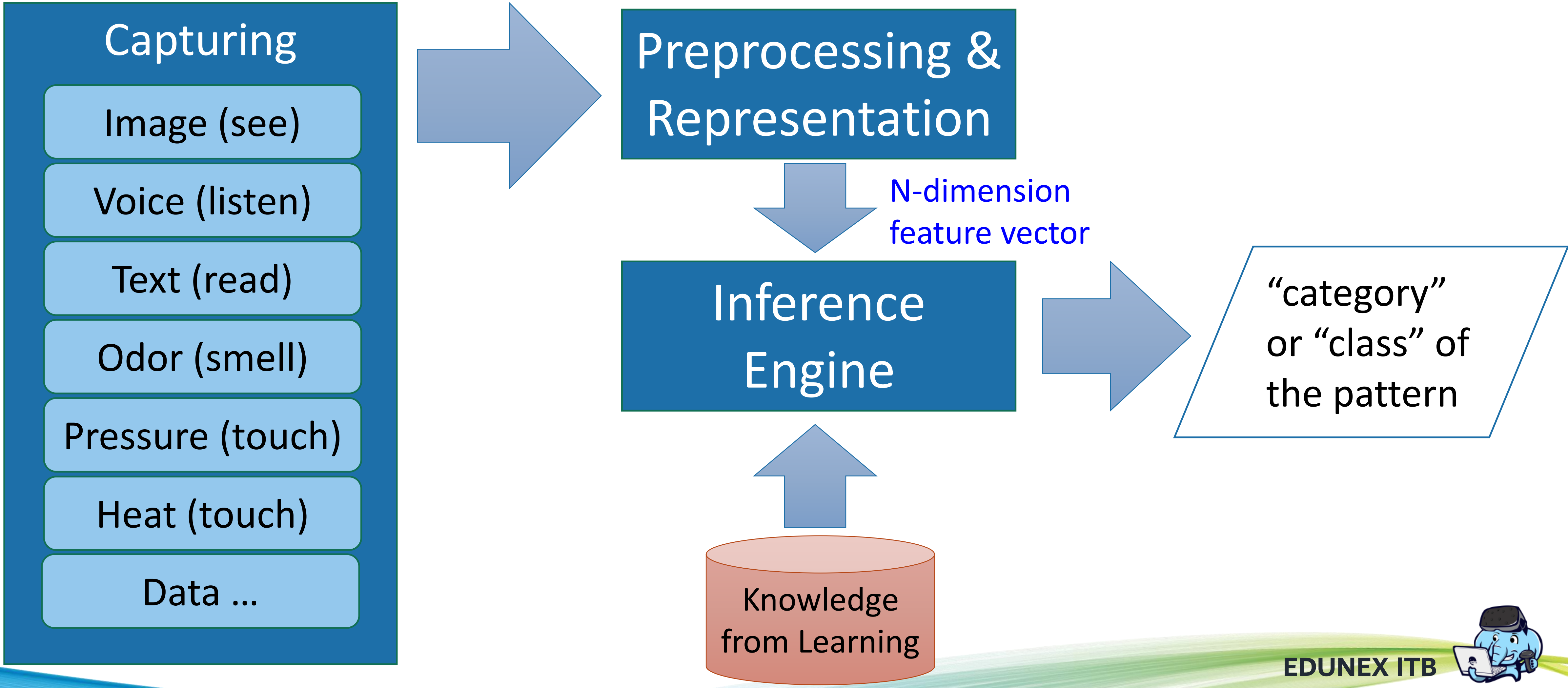
Tom Mitchell (1998):

Suatu program komputer disebut belajar dari pengalaman E untuk task T dan ukuran kinerja P, jika kinerjanya untuk task T yang diukur dengan P meningkat dengan adanya pengalaman E



https://towardsdatascience.com/cousins-of-artificial-intelligence-dda4edc27b55

Learning-based Recognition System



Representation: Feature Space

Vision

Low-level features
Pixel, descriptor, edges,…

mid-level features
segmentation

high-level features
Object

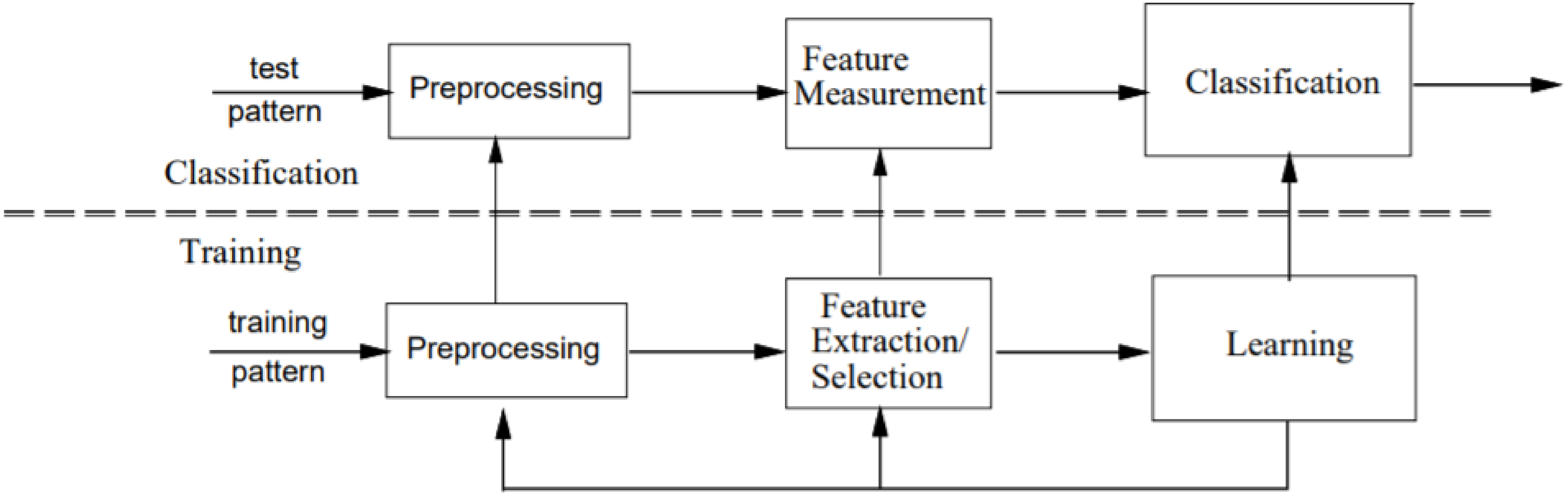
Text

Lexical features

Syntactic features

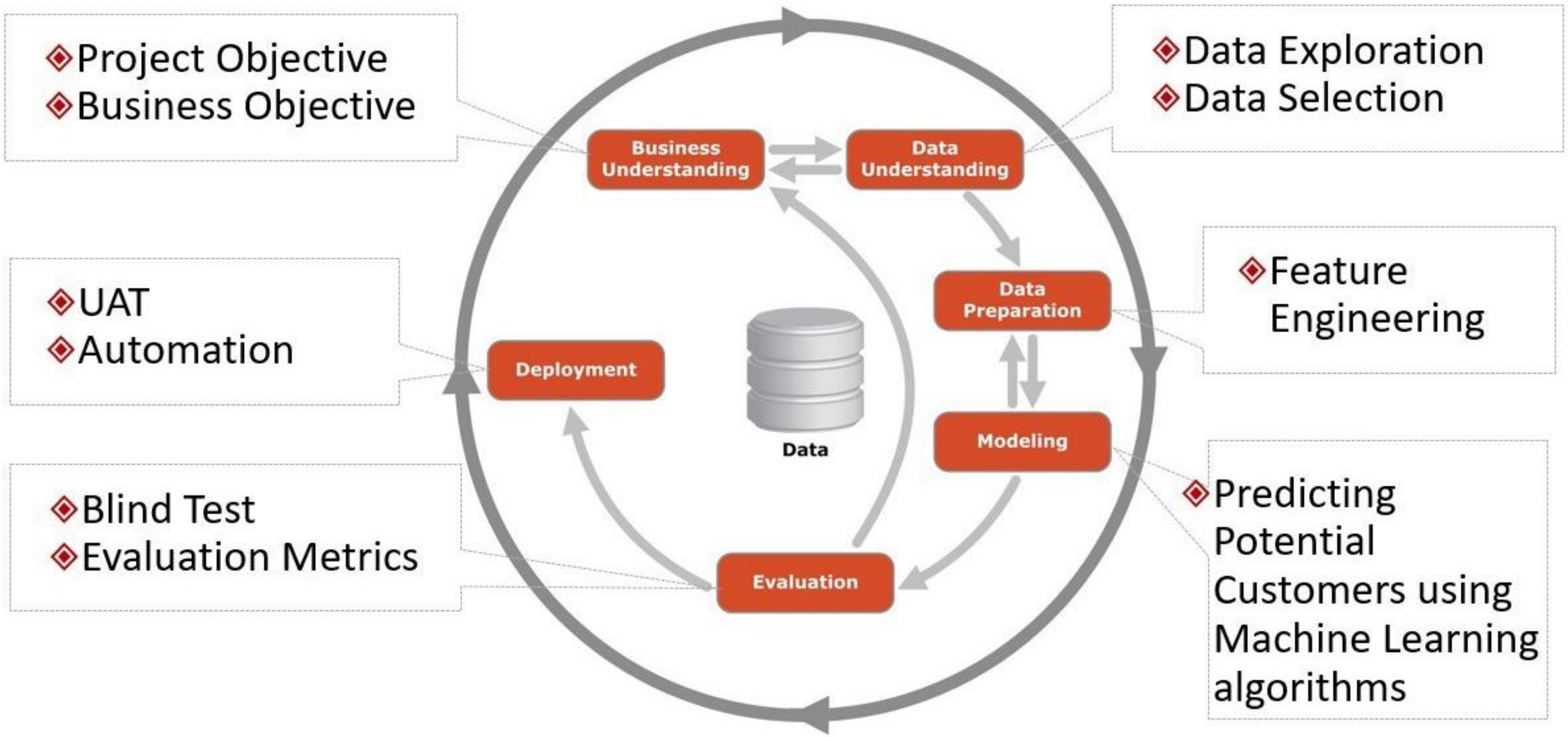
Semantic features

Learning-based Recognition System: Training



Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence*, *22*(1), 4-37.

Metodologi Pengembangan: CRISP-DM



Cross-Industry Standard Process for Data Mining

https://medium.com/@sumit.yg/analyzing-seattle-s-airbnb-listings-data-49abdc0977c8

When use Machine Learning ?

- Math formula ?
- Algorithm ?
- Simple knowledge ?
- Expert knowledge ?
- Data is available ?

Data Mining Process

Modeling
Evaluation

Dataset construction
Data understanding
Data preparation

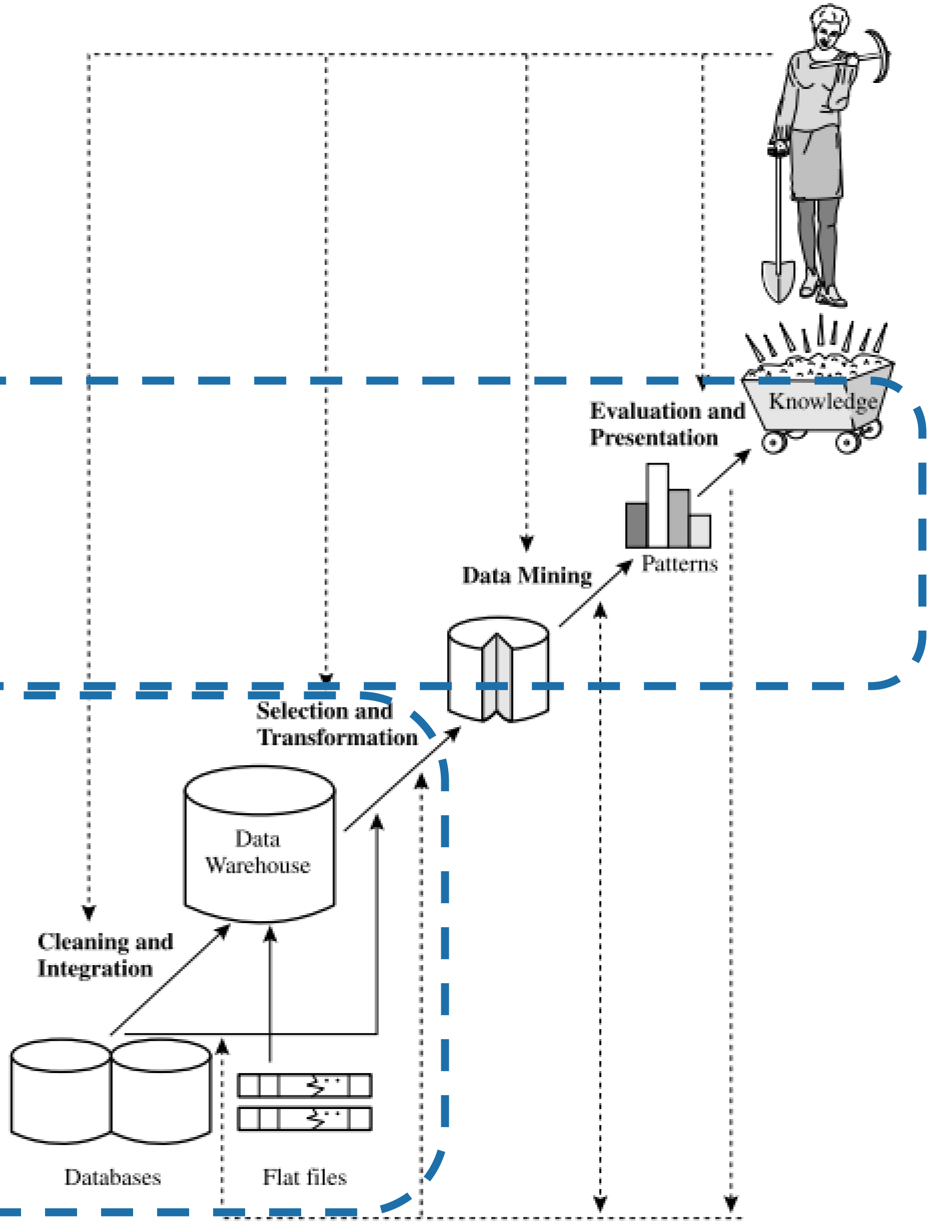


Figure 1.4 Data mining as a step in the process of knowledge discovery.

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

Data Quality

Garbage in Garbage out

http://1.bp.blogspot.com/

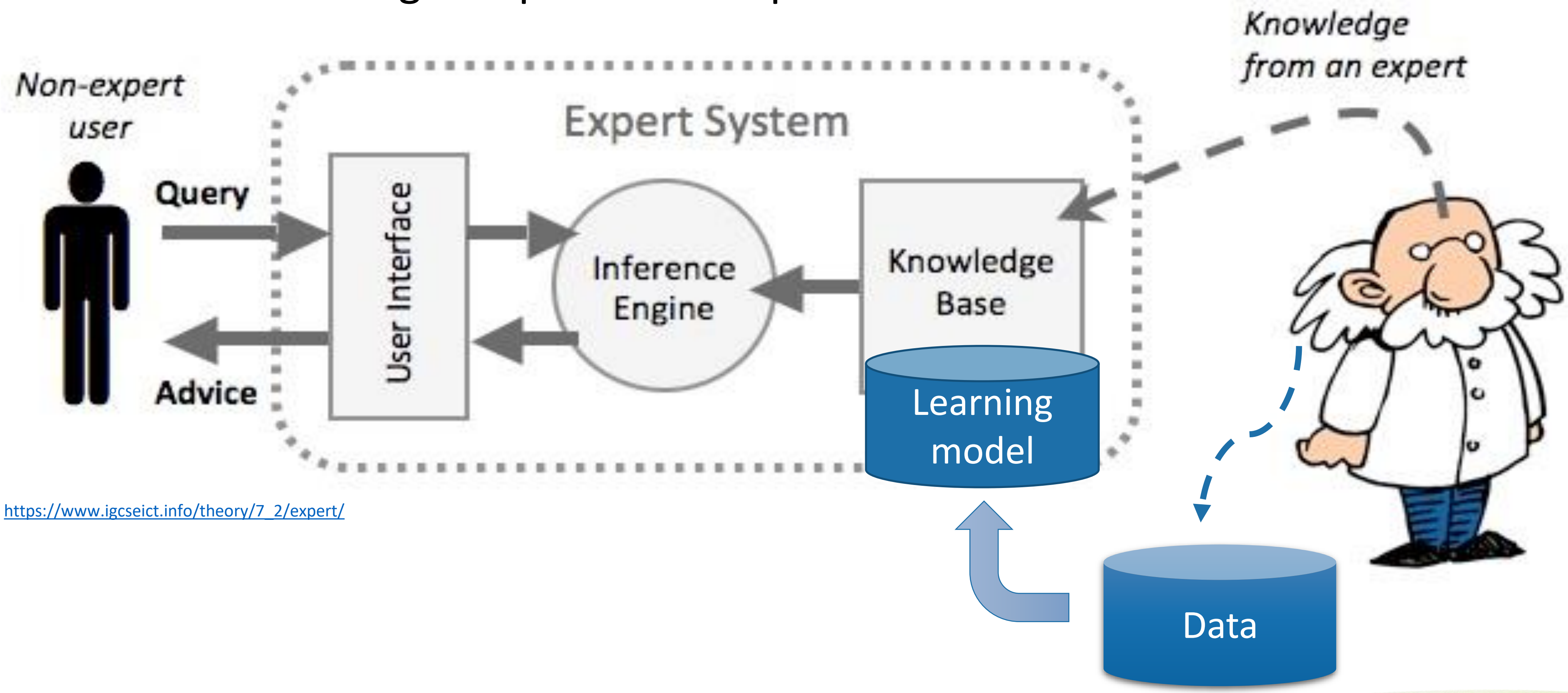
Low-quality data will lead to low-quality mining results

Common problem: noisy, missing, inconsistent, duplicate data.

Another problem: imbalanced dataset, outliers.

ML in Knowledge-based System

Automatic knowledge-acquisition component in KBS



https://www.igcseict.info/theory/7_2/expert/

Summary

Statistical decision approach

What, Why, When ML

ML in KBS

Automatic Fish Sorting

Modul 2: Pattern Recognition System

04 Automatic Fish Sorting (Case Study)

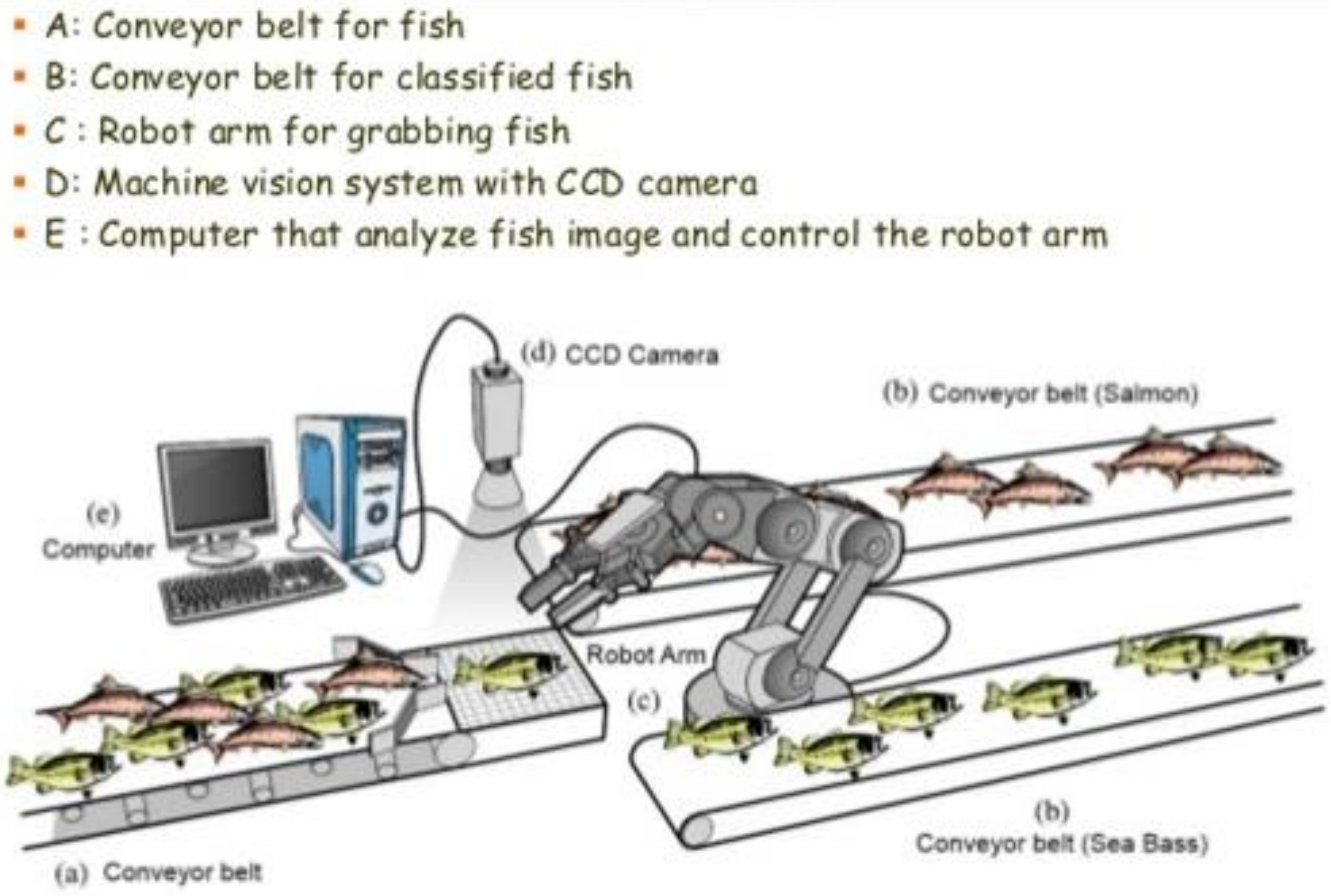
Masayu Leylia Khodra
(masayu@informatika.org)

KK IF – Teknik Informatika – STEI ITB

Pengenalan Pola
(*Pattern Recognition*)

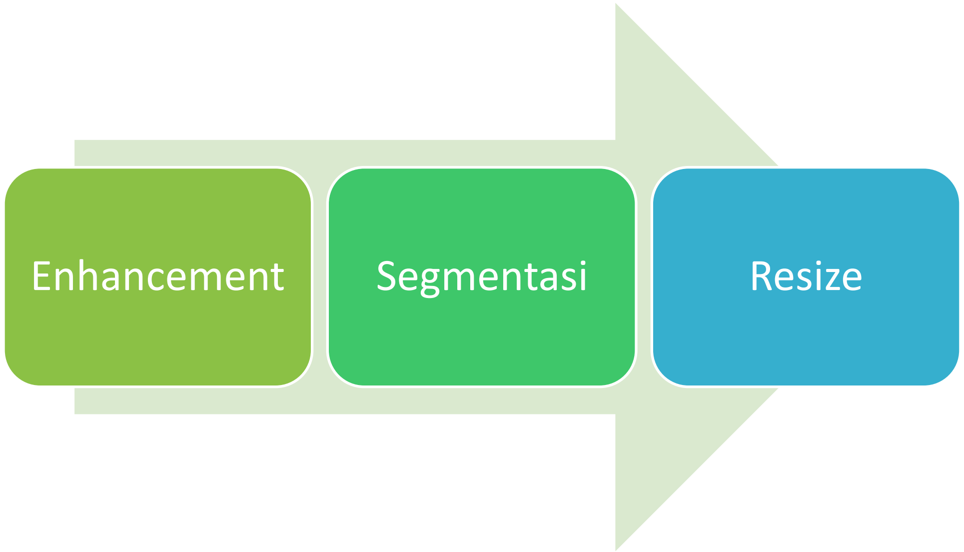
Fish Packing Plant

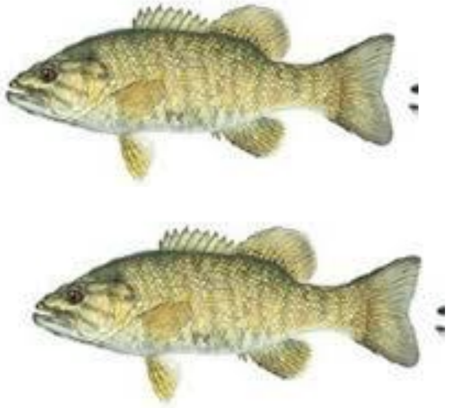
- Fish packing plant wants to automate the process of sorting incoming fish on a conveyor belt according to species.

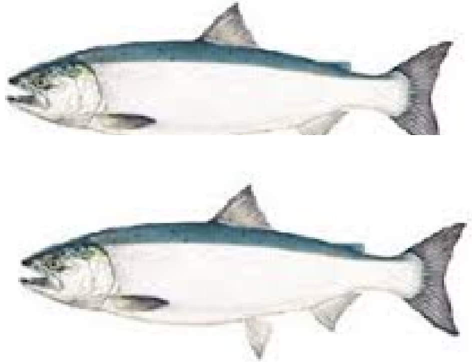


Automated Fish Classification System

Dataset Construction







Duda dkk, 2001

Fish Species Recognition

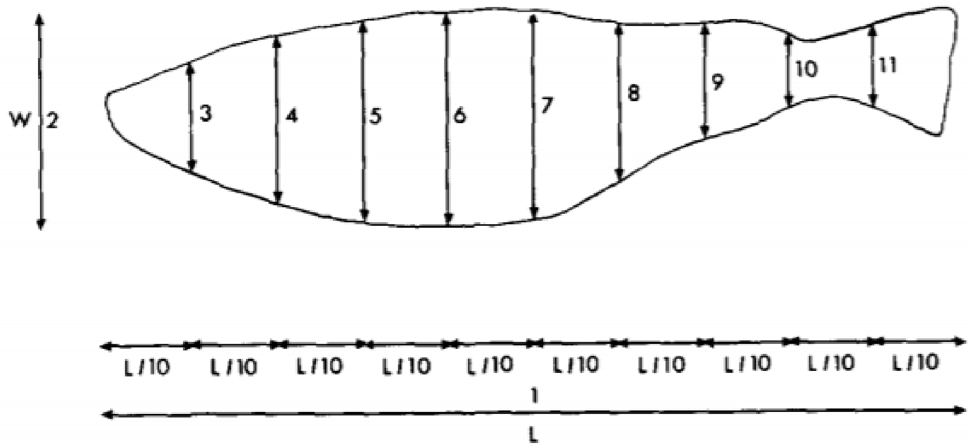


Fig. 4. A fish shape with the eleven shape descriptors.

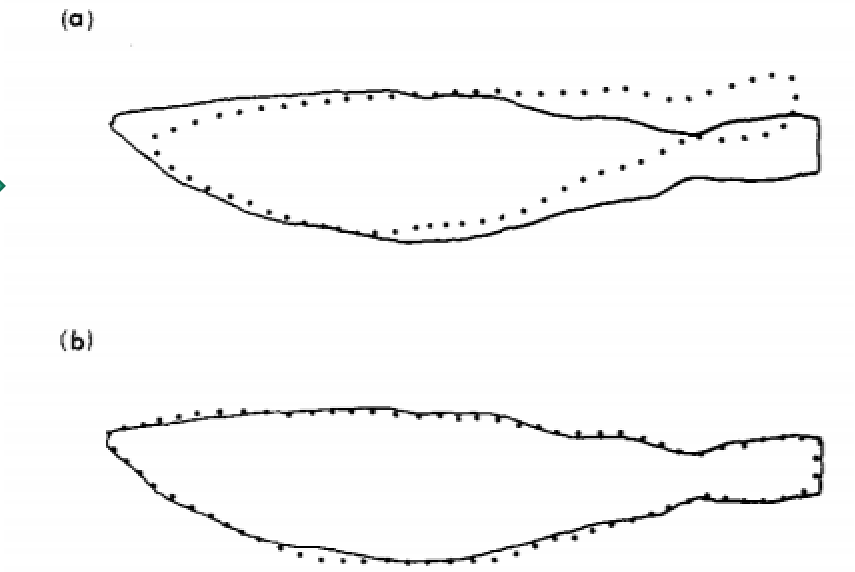


Fig. 3. Computer diagrams of two fish shapes (a) at the start of the optimisation procedure, and (b) at the end of the optimisation procedure.

Of the 60 fish in this experiment the first half of each species (set A) was used to generate the template fish. Two fish of the same type in set A were placed on top of each other with both of their geometric centres of area

Strachan, N. J. C., Nesvadba, P., & Allen, A. R. (1990). Fish species recognition by shape analysis of images. *Pattern Recognition*, *23*(5), 539-544.

Fish features

- Piksel
- Descriptor based on : color, shape, textures

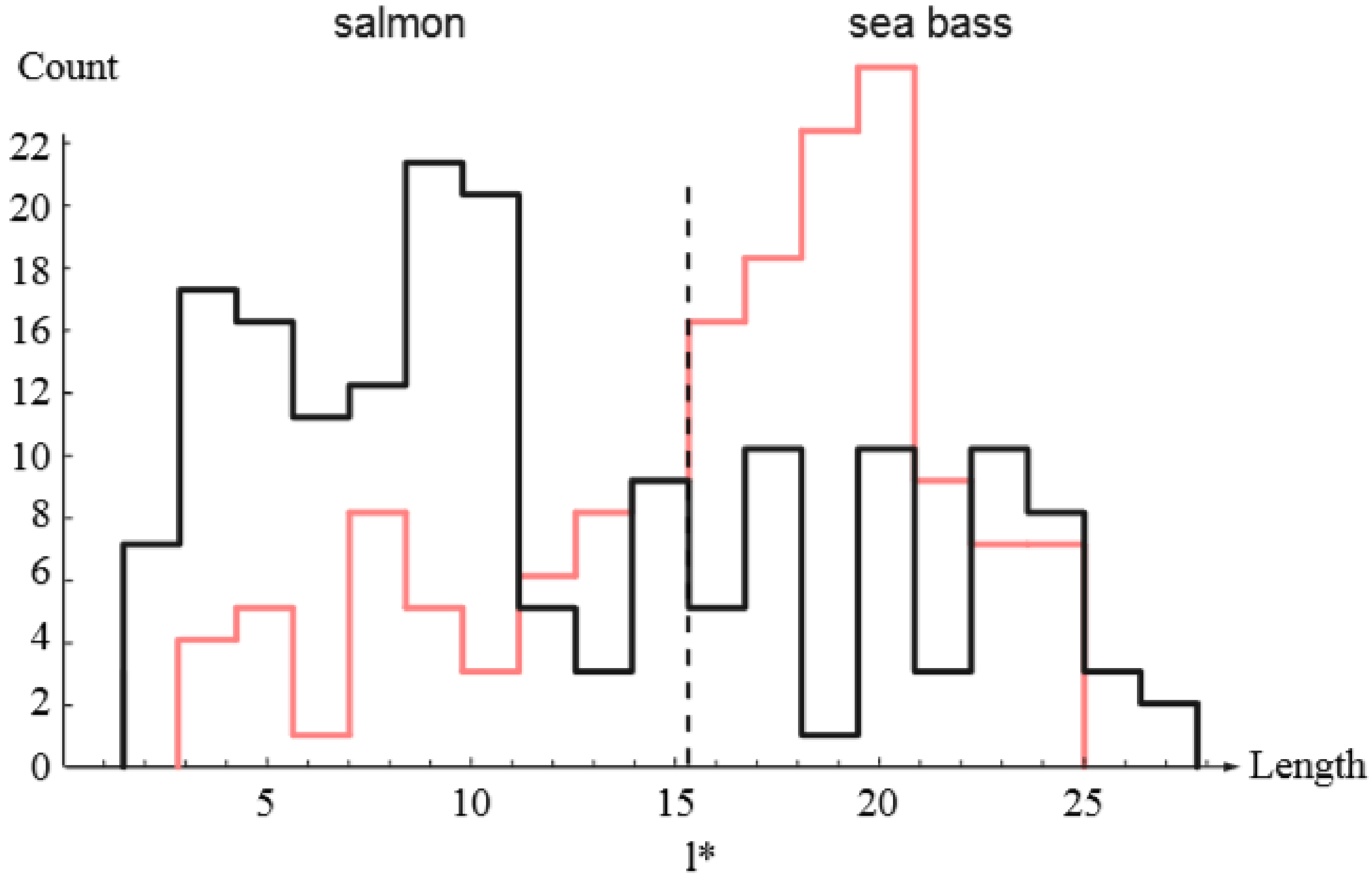
Find physical differences between the two types of fish:

- 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.

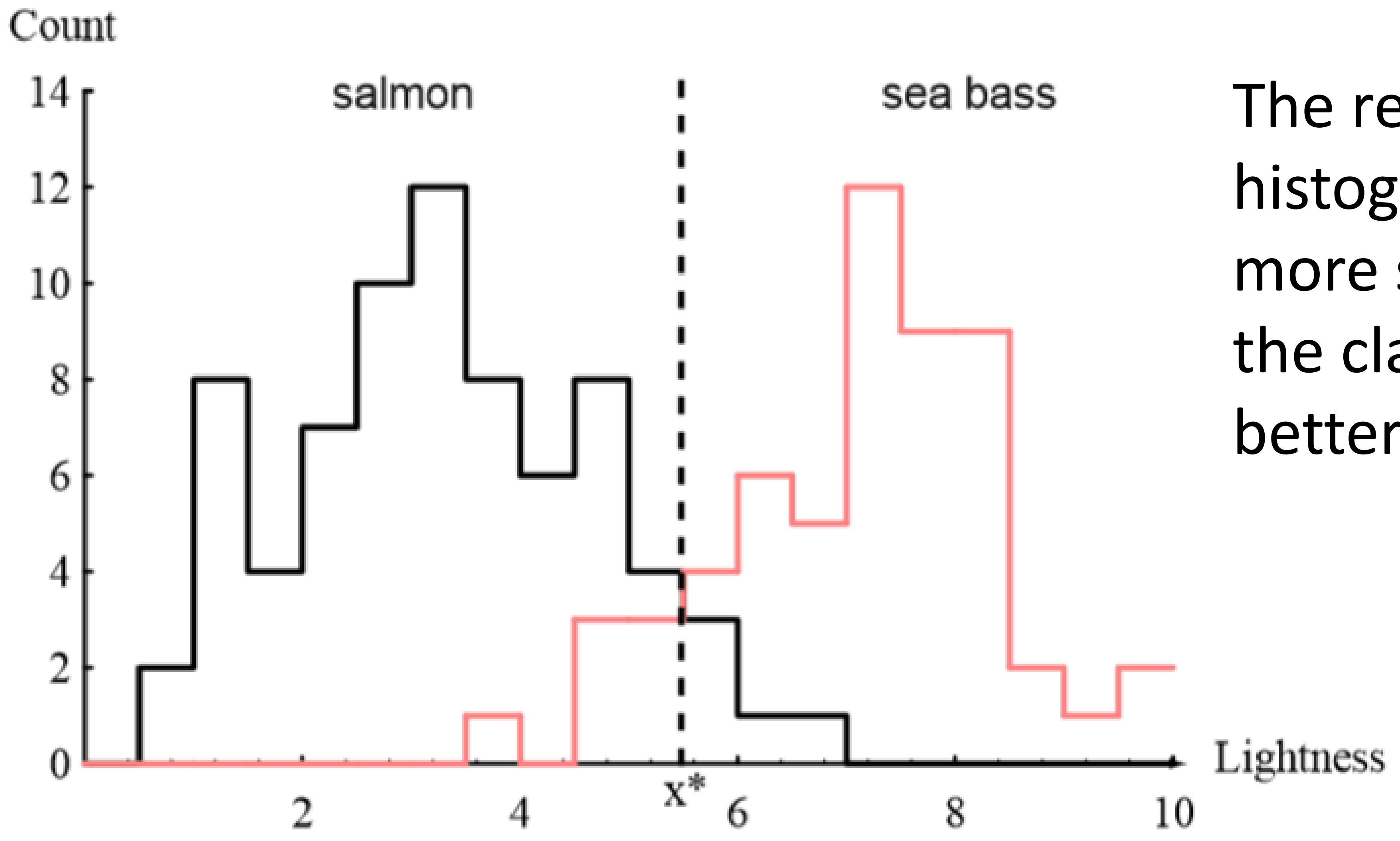


From histograms, sea bass are longer than salmon, on average, but it is clear that this single criterion is quite poor;

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

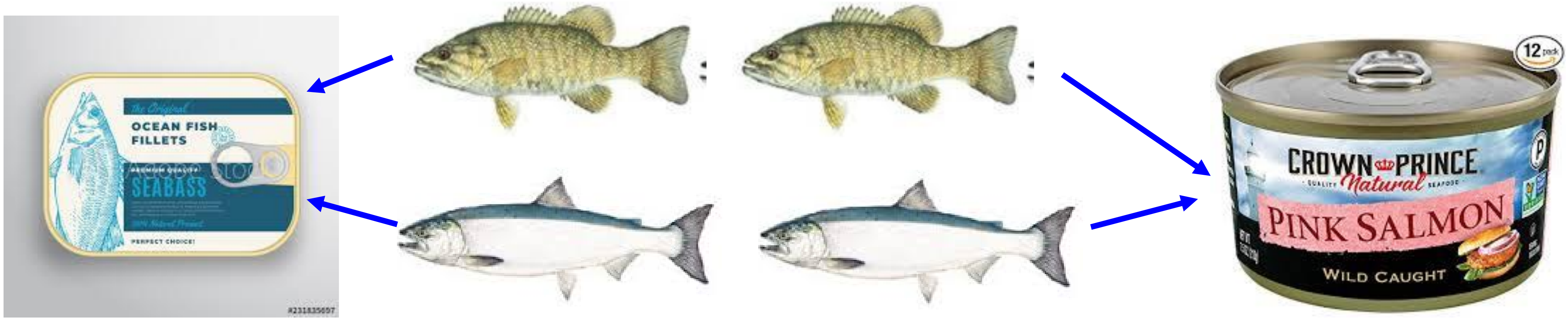
The value l∗ marked will lead to the smallest number of errors, on average.

Tentative Model: Lightness Feature



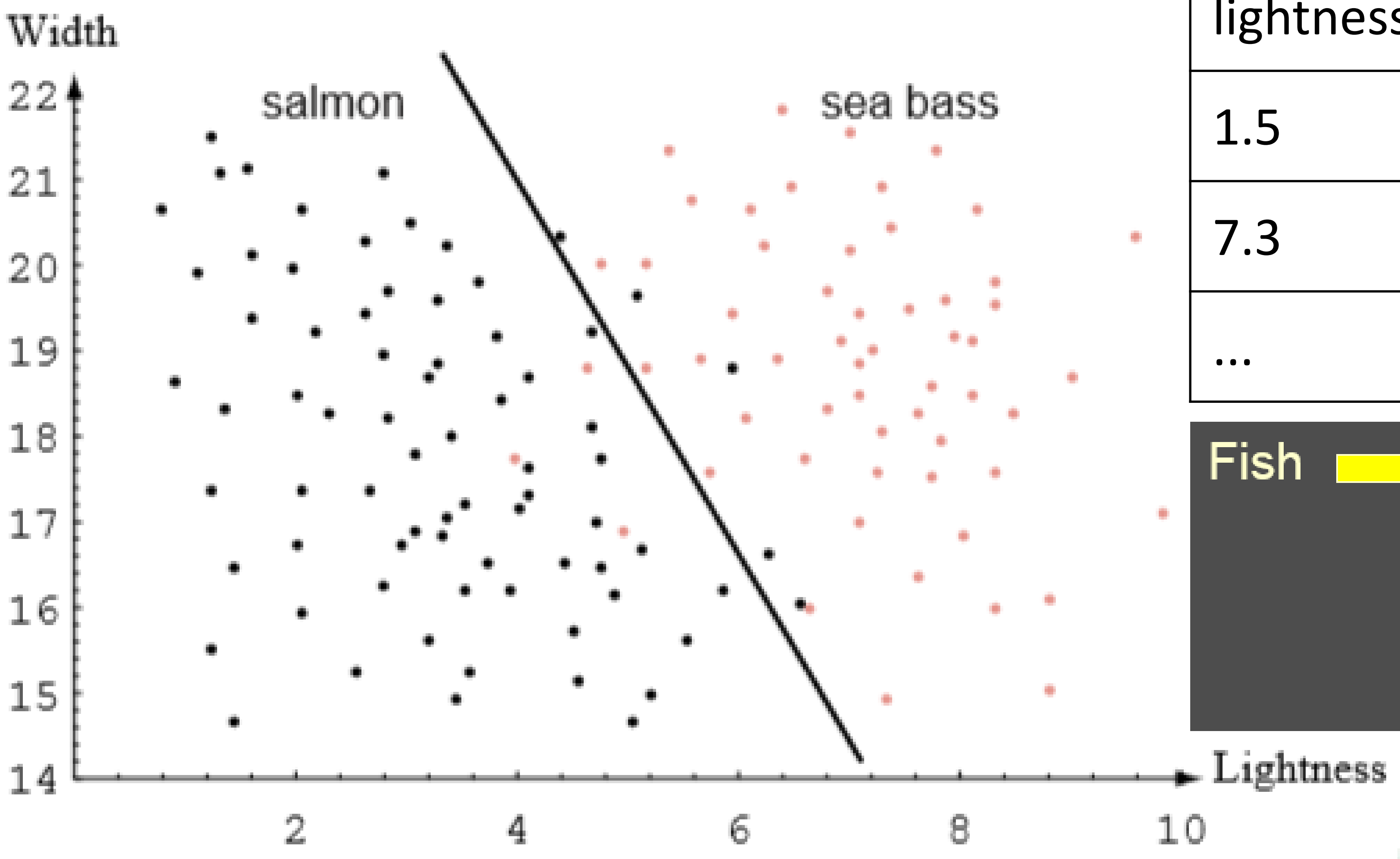
The resulting histograms are much more satisfactory — the classes are much better separated.

Misclassification

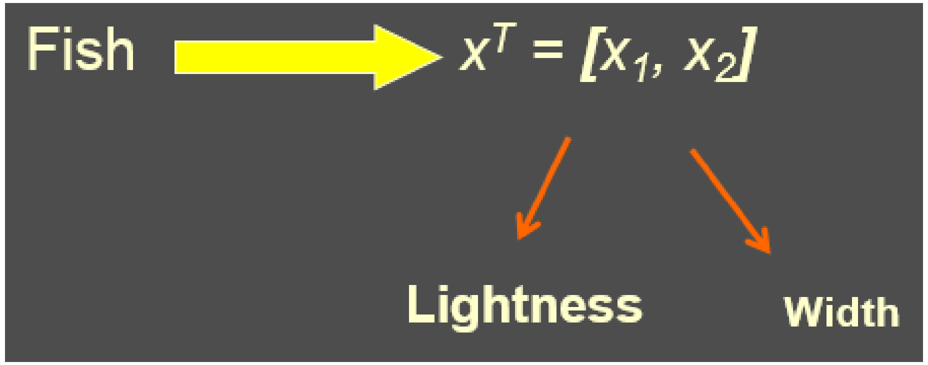


salmon in “sea bass” cans vs sea bass in “salmon” cans ?

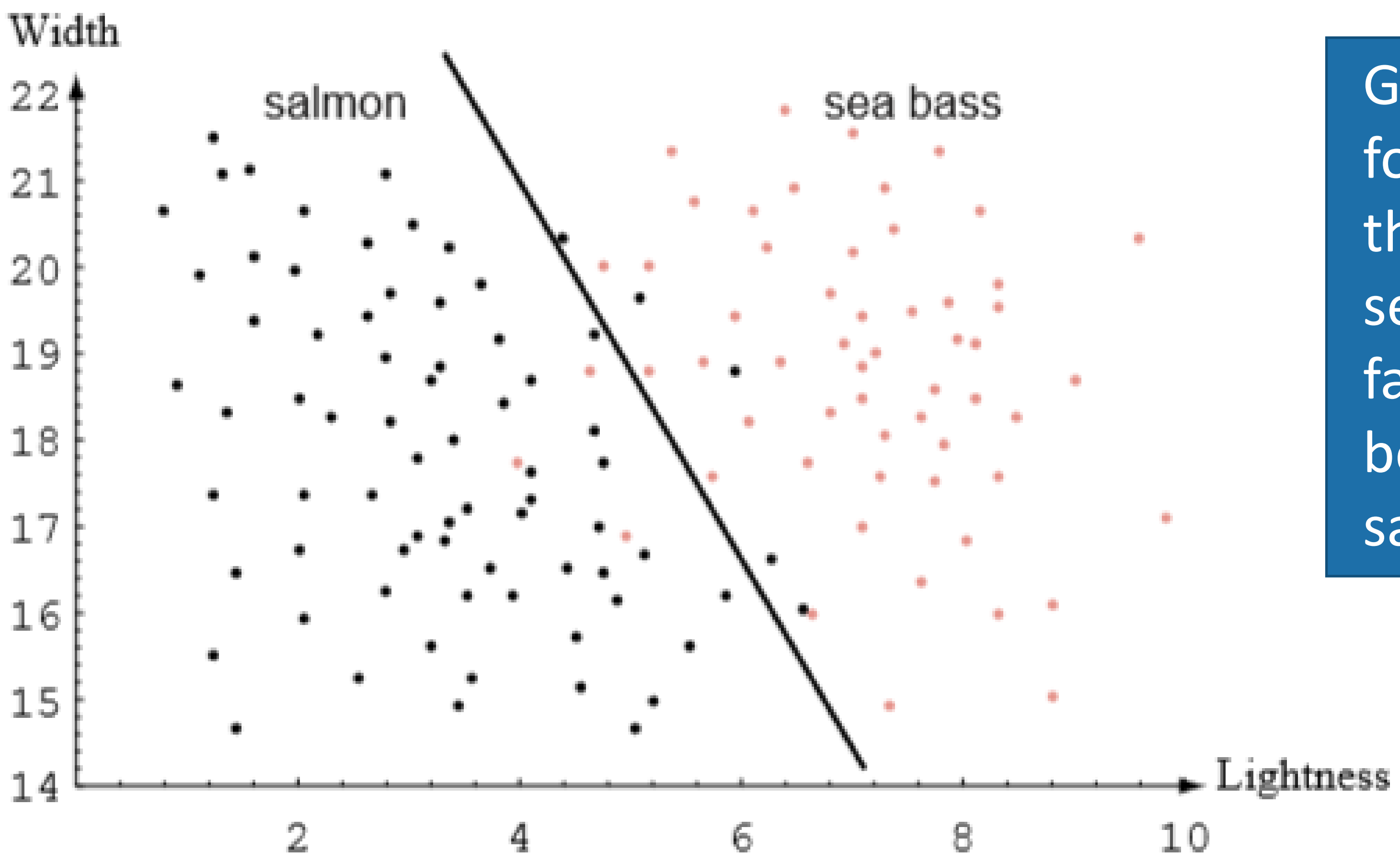
Tentative Model: Lightness and Width Feature



lightness	width	category
1.5	14.6	salmon
7.3	15	Sea bass
…		

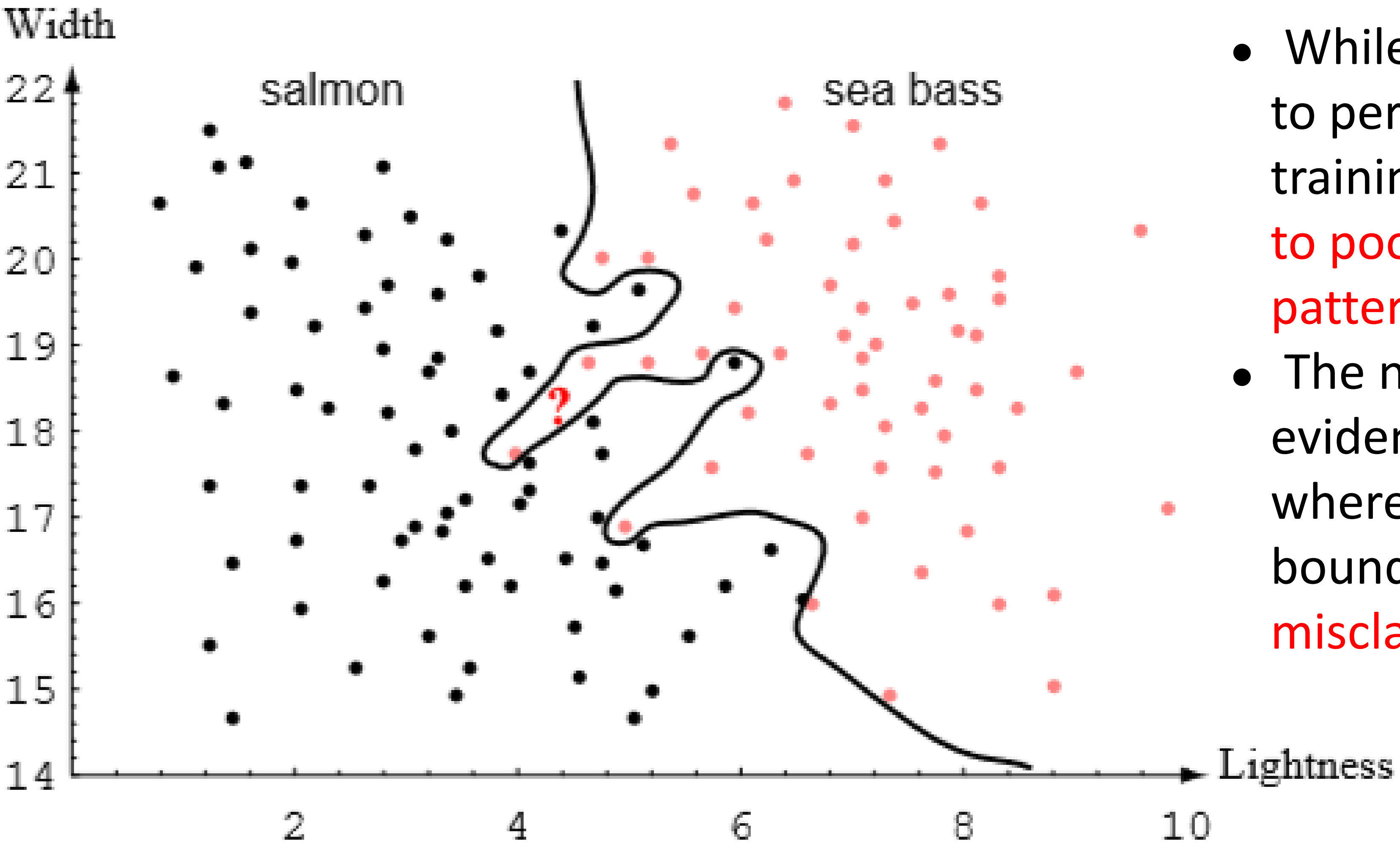


Rule based on 2 Features



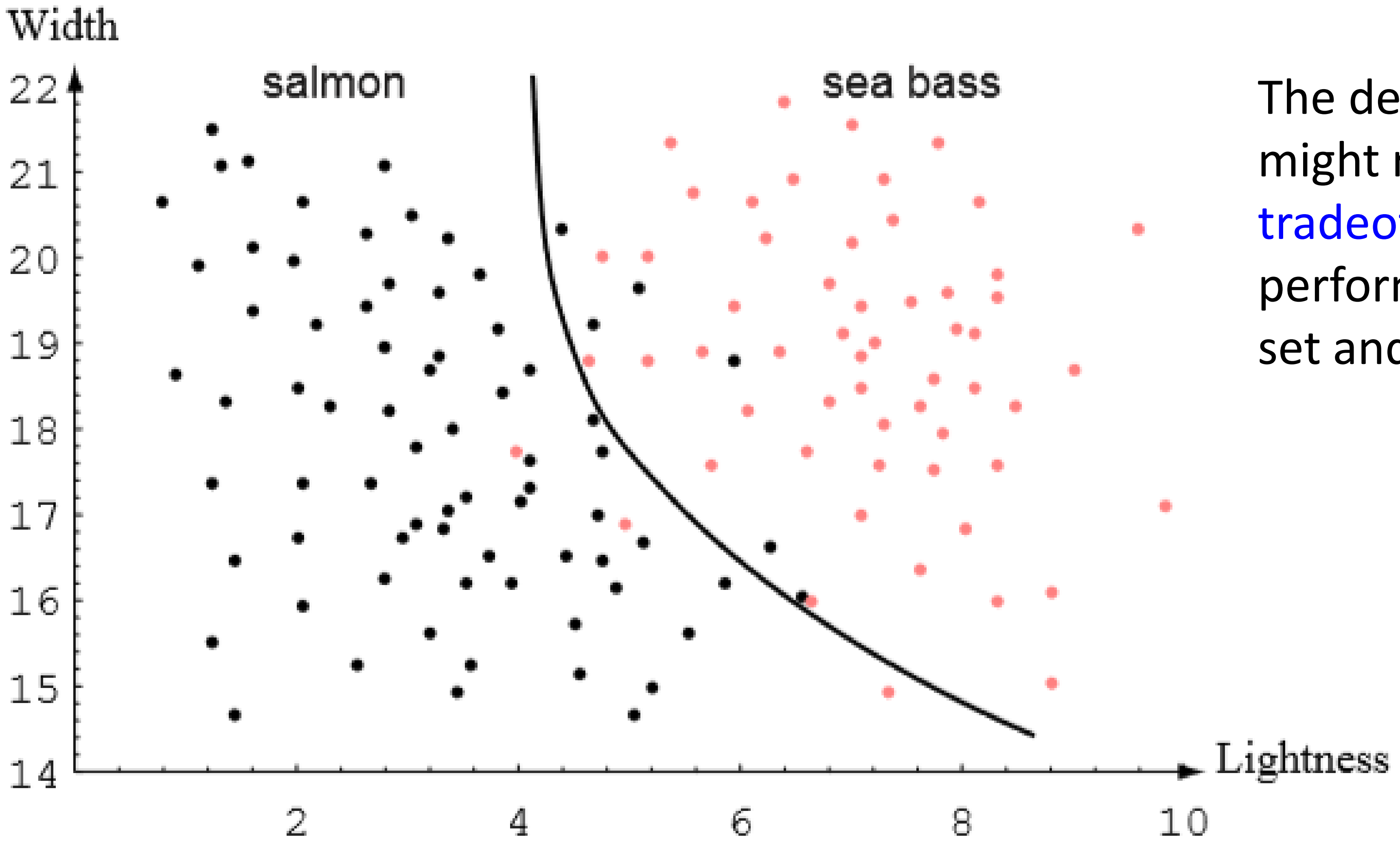
Given decision boundary, the following rule for separating the fish: Classify the fish as sea bass if its feature vector falls above the decision boundary shown, and as salmon otherwise.

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.

Modeling using Supervised Learning

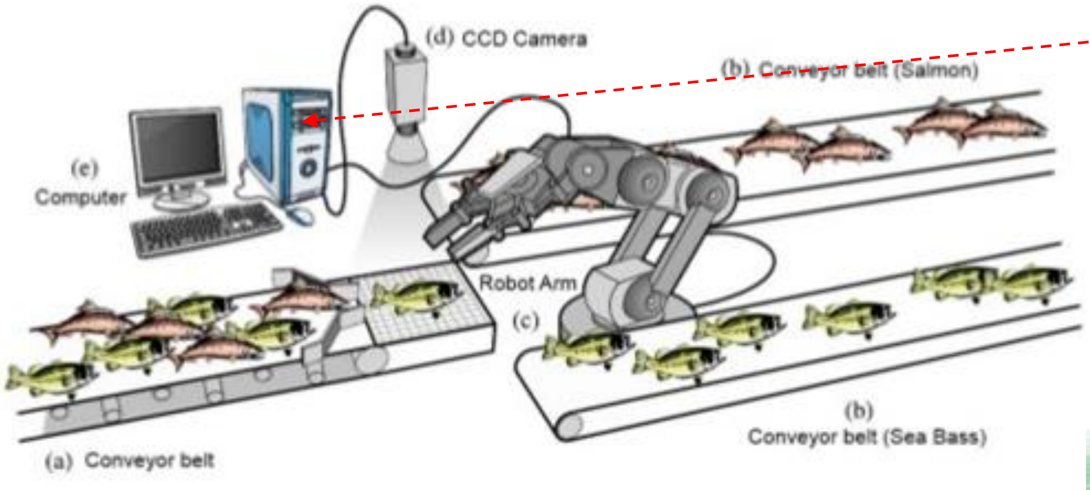
Training data

Preprocessing and Feature Extraction

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

Machine Learning

model



Automated Fish Classification System

Summary

Fewer features build simpler decision boundary and easier training process.

Ideal representation: low intra-class variability, and high inter-class variability

Different task, different feature set, and different optimal decision boundary

Challenge: one general purpose artificial pattern recognition device