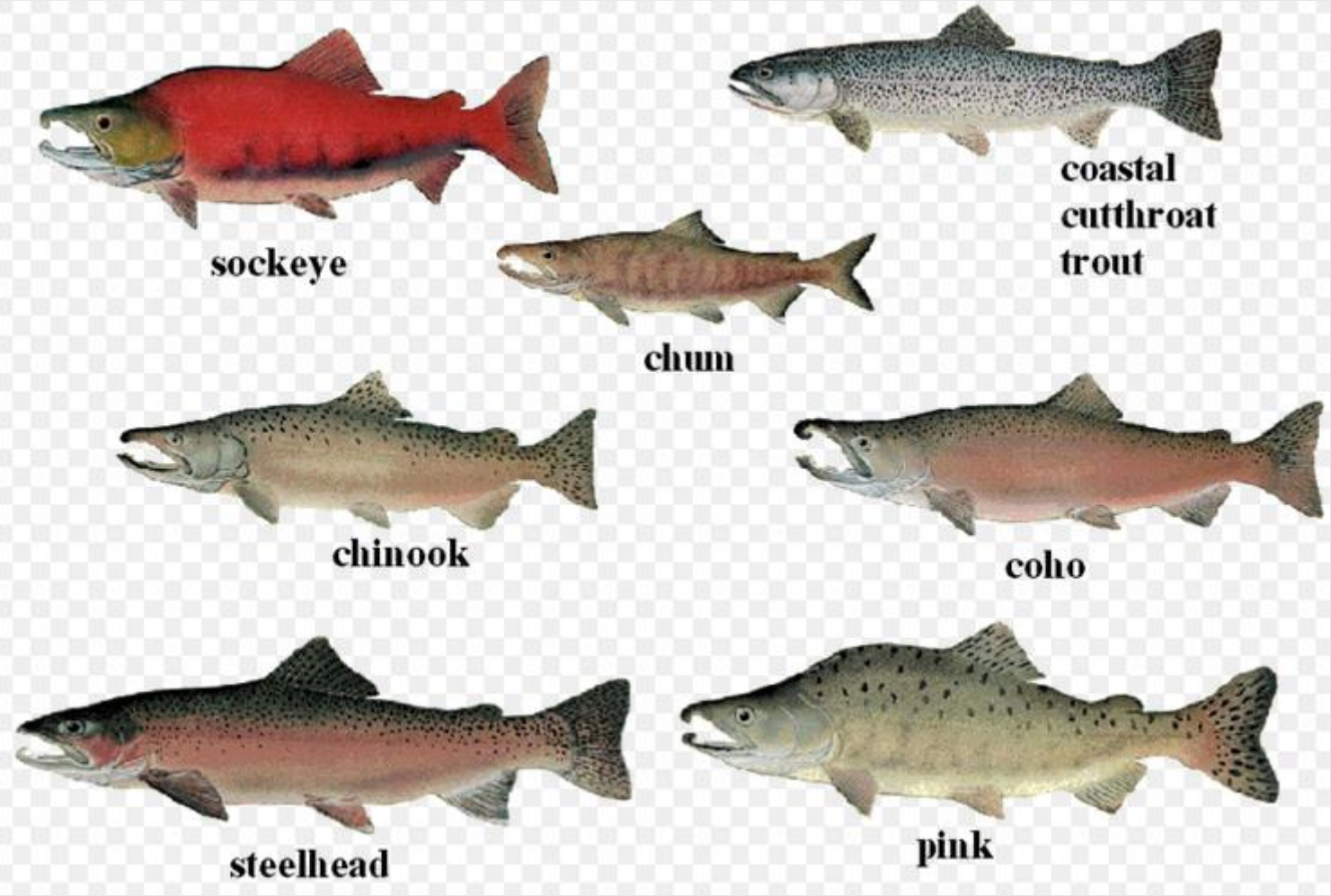
Data Mining

Pertemuan 4: Classification

Ilustrasi

- Kita tahu bahwa terdapat ikan salmon di sebuah sungai
- Ketika kita mengambil seekor ikan dari sungai, dapatkah kita mengetahui bahwa ini ikan salmon?
- Asumsi: kita tidak mengetahui ciri-ciri ikan salmon
- Bagaimana solusinya?.....?
- → BELAJAR (*LEARNING*)

IKAN SALMON

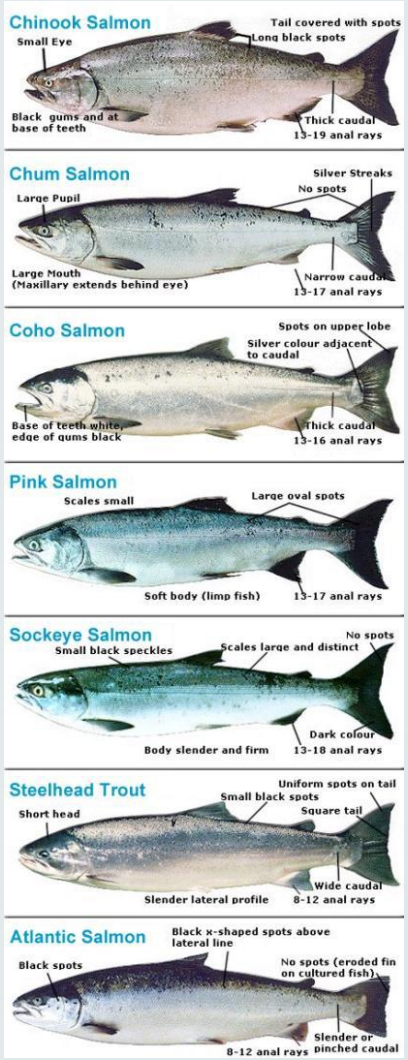

sockeye
coastal
cutthroat
trout
chum
chinook
coho
steelhead
pink

Tipe Pembelajaran

- Pembelajaran Pasif *(Passive Learning)*
- Pembelajaran Aktif *(Active Learning)*

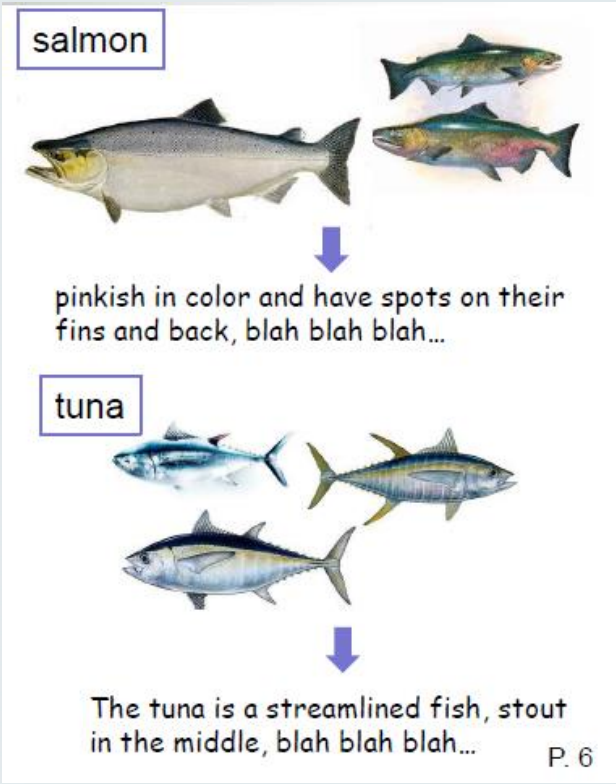
Pembelajaran Pasif

- Cari seorang ahli salmon
- Ia akan memberi tahu semua karakteristik dari ikan salmon
- Kita cukup mengingat dan menerapkan apa yang sudah dipelajari



Pembelajaran Aktif

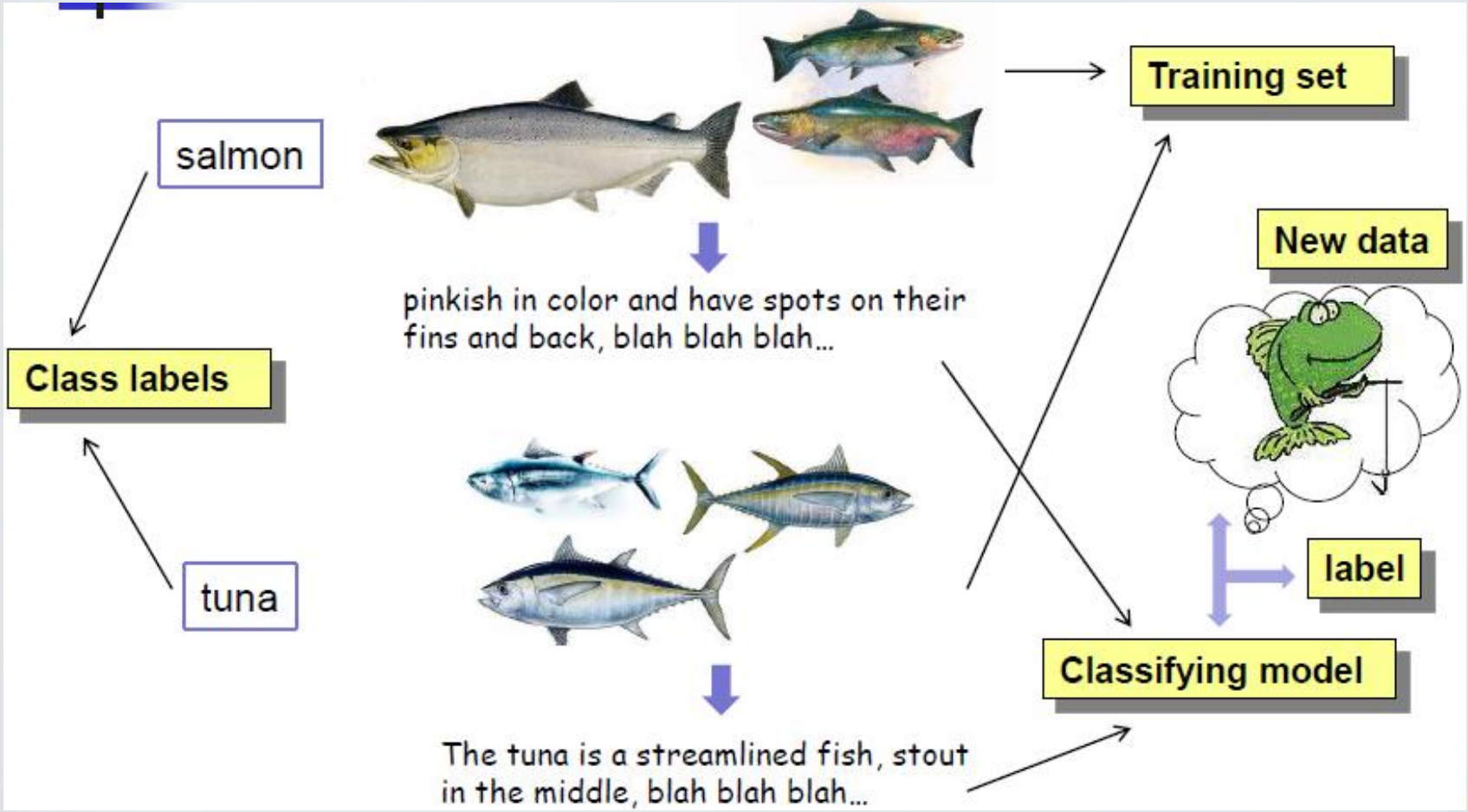
- Cari seorang ahli salmon
- Ia menangkap banyak ikan dari beberapa jenis
- Ia hanya memberitahu mana yang ikan salmon dan mana yang bukan
- Kita harus mempelajari sendiri karakteristik ikan salmon dengan mengamati karakteristik ikan yang disebut salmon oleh pakar



Klasifikasi Pada Data Mining

- Memprediksi label *class* yang bertipe kategori
- Mengklasifikasi data (membangun sebuah model) berdasarkan data latih (*training set*) dan label *class* yang diberikan untuk mengklasifikasikan *data baru*

Proses Klasifikasi



Klasifikasi pada Data Mining

- Di data mining, kita selalu tertarik pada pembelajaran aktif
- Pertanyaan:
  - Mengapa kita sebagai pakar tidak langsung saja memberitahu langsung pada komputer bagaimana ciri-ciri ikan salmon?

Klasifikasi pada Data Mining

- Jawaban:
  - Bahkan seorang ahli pun terkadang kesulitan dalam mendeskripsikan/mengidentifikasi semua karakteristik dari sebuah pengamatan terhadap sesuatu
- Contoh:
  - Dalam inbox e-mail kita, terdapat banyak e-mail. Kita tahu mana e-mail spam dan mana yang bukan
    - Dapatkah kita menuliskan SEMUA karakteristik e-mail spam?
  - Pada pembelajaran aktif, kita hanya perlu memberitahu komputer mana e-mail spam dan mana yang bukan.
  - Komputer akan mengidentifikasi sendiri karakteristik dari e-mail dengan mengamati perbedaannya
    - SANGAT MENGHEMAT WAKTU!

Catatan

- Dari sudut pandang data mining:
  - Klasifikasi ≈ Prediksi ≈ Peramalan, karena tekniknya serupa
- Klasifikasi juga disebut Pembelajaran Tersupervisi *(Supervised Learning)*
  - Harus ada pakar (kita) yang men-supervisi komputer
  - Sebaliknya, Clustering disebut Pembelajaran Tidak Tersupervisi *(Unsupervised Learning)*

Proses Klasifikasi

- Tahapan:
  - Kita menangkap banyak ikan
  - Kita memberitahu komputer mana yang salmon dan mana yang bukan
  - Komputer mengidentifikasi sendiri karakteristik salmon
- Beberapa istilah:
  - Untuk ikan yang telah ditangkap, kita bagi dalam kategori “Salmon” dan “Bukan Salmon”
  - Sampel positif: Ikan yang masuk kategori Salmon
  - Sampel negatif: Ikan yang masuk kategori Bukan Salmon
  - Model: Sesuatu yang telah dipelajari komputer. Ketepatan model bergantung pada algoritma pembelajaran

Dua Langkah dalam Klasifikasi

- **Pembuatan model**: mendeskripsikan sekumpulan *class* yang telah ditentukan nilainya
  - Setiap sampel diasumsikan memiliki kelas yang ditandai dengan label atribut kelas
  - Setiap sampel yang digunakan untuk pembuatan model disebut training set
  - Model direpresentasikan sebagai aturan klasifikasi, pohon keputusan, atau rumus matematika
- **Penggunaan model**: untuk mengklasifikasikan objek baru
  - Mengestimasi keakuratan model
    - Label sesungguhnya dari test set dibandingkan dengan hasil klasifikasi model
    - Tingkat akurasi: persentase sampel test set yang diklasifikasi dengan benar oleh model
    - Test set harus terpisah dari training set
  - Jika akurasi dapat diterima, gunakan model tersebut untuk mengklasifikasikan sampel data yang label kelasnya tidak diketahui → validation set

Dua Langkah Klasifikasi

ID	Color	Size	…	Label
1	Pink	20cm	…	Salmon
2	Green	30cm	…	Not Salmon
⋮	⋮	⋮	…	
N	Pink	18cm	…	Salmon

1. Archive Training Data

2. Choose an learning algorithm

Model

Learning

Model Evaluation

An unknown fish

Model

Yes (Salmon)

No (Not a Salmon)

Operation

Binary Class vs Multi Class

- Klasifikasi Binary-Class
  - Hanya ada 2 kelas
    - “Salmon” dan “Bukan Salmon”
    - “Kucing” dan “Anjing”
- Klasifikasi Multi-Class
  - Terdapat lebih dari 2 kelas
    - “Salmon”, “Tuna”, “Hiu”, “Koi”
  - Setiap permasalahan klasifikasi multi-class dapat diselesaikan dengan memformulasikan serangkaian model klasifikasi binary-class
    - Bagaimana caranya?

Algoritma

- Beberapa algoritma utama untuk pembelajaran:
  - Decision Tree (Pohon Keputusan)
  - Nearest Neighbor (Tetangga Terdekat)
  - Naïve Bayes
  - Support Vector Machines

Komite Pengklasifikasian (*Classifier*)

- Keputusan dibuat oleh sejumlah *classifier*
  - Keputusan yang diambil dari sejumlah pakar biasanya lebih baik dibandingkan dari 1 orang saja
- Beberapa *classifier* digunakan untuk memprediksi label kelas dan hasilnya dikombinasikan

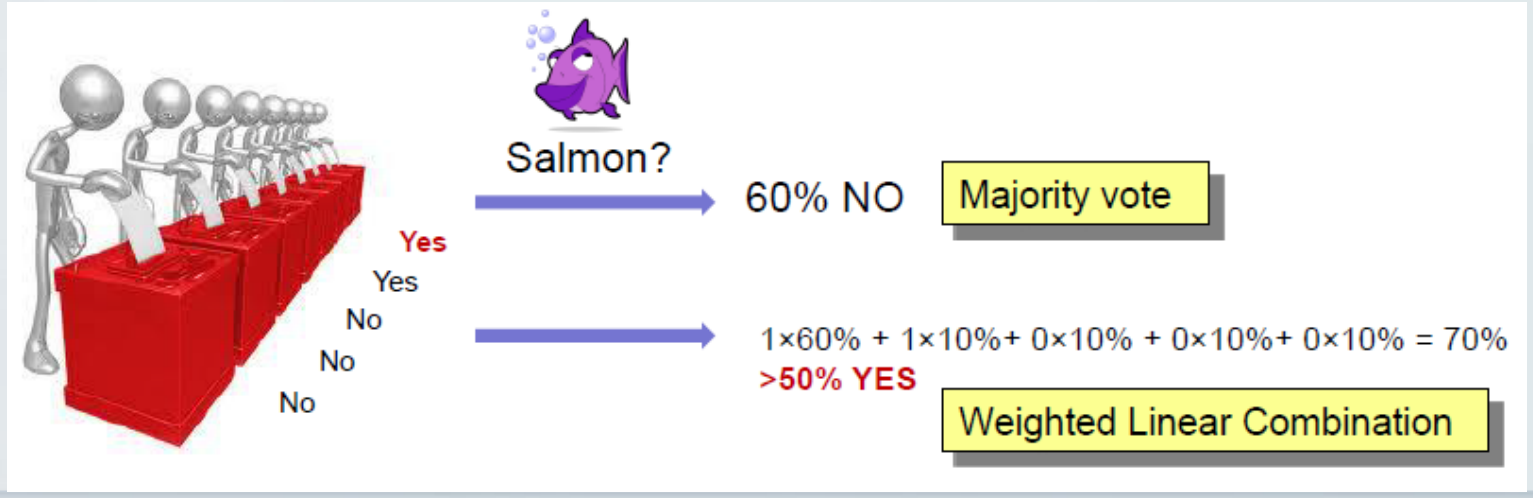
Dua Teknik Kombinasi

- Majority Vote (Suara Terbanyak)
  - Melakukan voting sederhana

3/5 = 60% YES

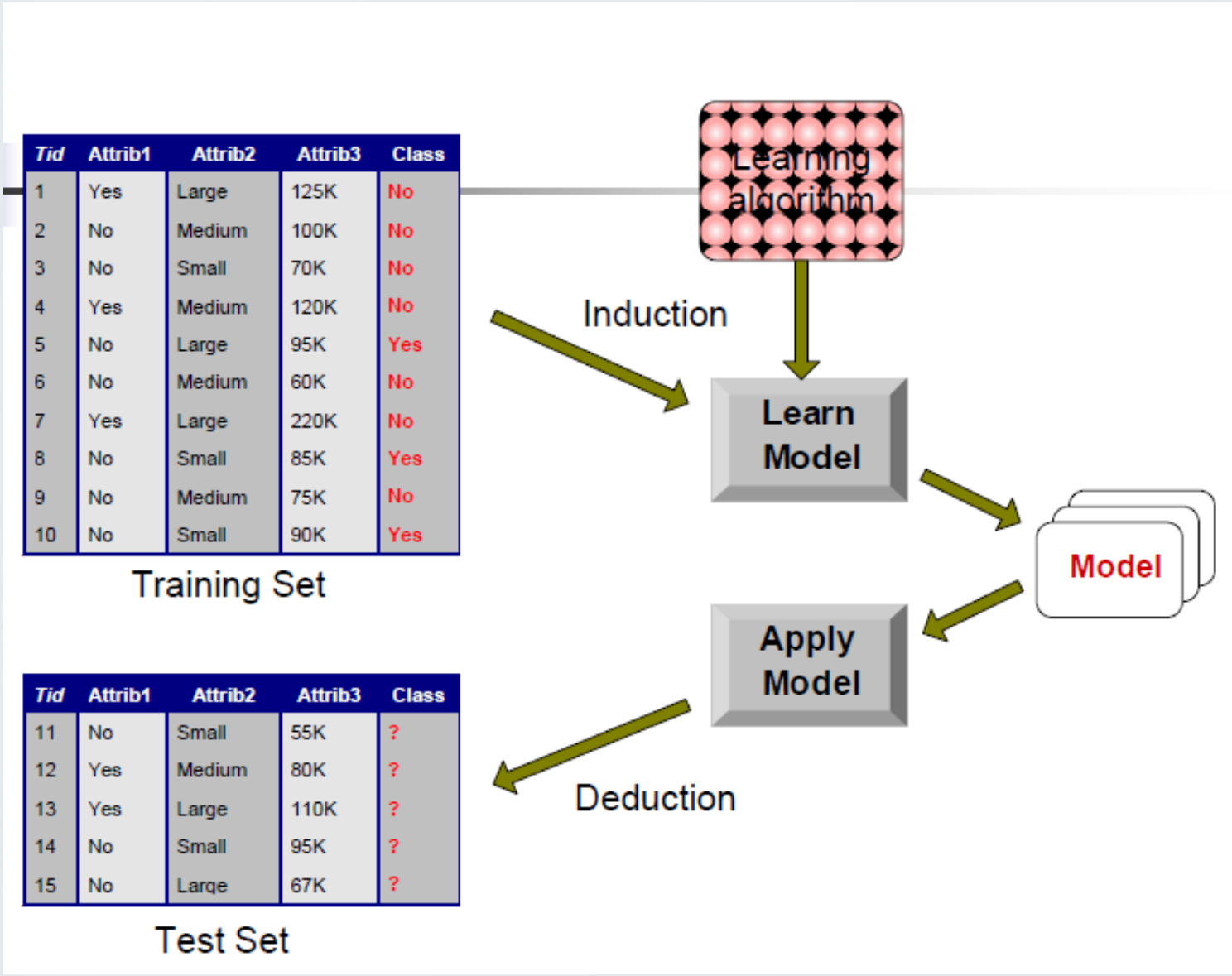
Dua Teknik Kombinasi

- Kombinasi Bobot Linier (*Weighted Linear Combination*)
  - Jika sebuah *classifier* lebih *reliable* (dapat diandalkan), maka kita menghargai keputusannya lebih tinggi dari yang lain



Evaluasi Model

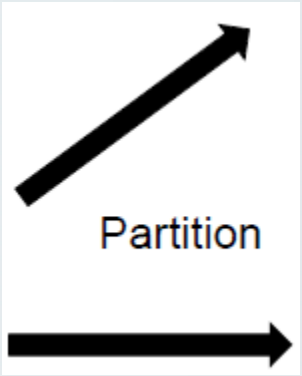
- Setelah proses training, kita perlu melakukan tes terhadap model sebelum digunakan untuk melihat apakah model tersebut sudah belajar dengan baik dan seberapa handal kinerjanya


Tid Attrib1 Attrib2 Attrib3 Class
1 Yes Large 125K No
2 No Medium 100K No
3 No Small 70K No
4 Yes Medium 120K No
5 No Large 95K Yes
6 No Medium 60K No
7 Yes Large 220K No
8 No Small 85K Yes
9 No Medium 75K No
10 No Small 90K Yes
Training Set
Learning algorithm
Induction
Learn Model
Model
Apply Model
Deduction
Tid Attrib1 Attrib2 Attrib3 Class
11 No Small 55K ?
12 Yes Medium 80K ?
13 Yes Large 110K ?
14 No Small 95K ?
15 No Large 67K ?
Test Set

Testing

- Mempersiapkan data training dan data testing
  - Data training dan data testing tidak boleh overlap

ID	Color	Size	...	Label
1	Pink	20cm	...	Salmon
2	Green	30cm	...	Not Salmon
:	:	:	...	:
N	Pink	18cm	...	Salmon



ID	Color	Size	...	Label
1	Pink	20cm	...	Salmon
3	Green	32cm	...	Salmon
:	:	:	...	:
K	Black	24cm	...	Not Salmon

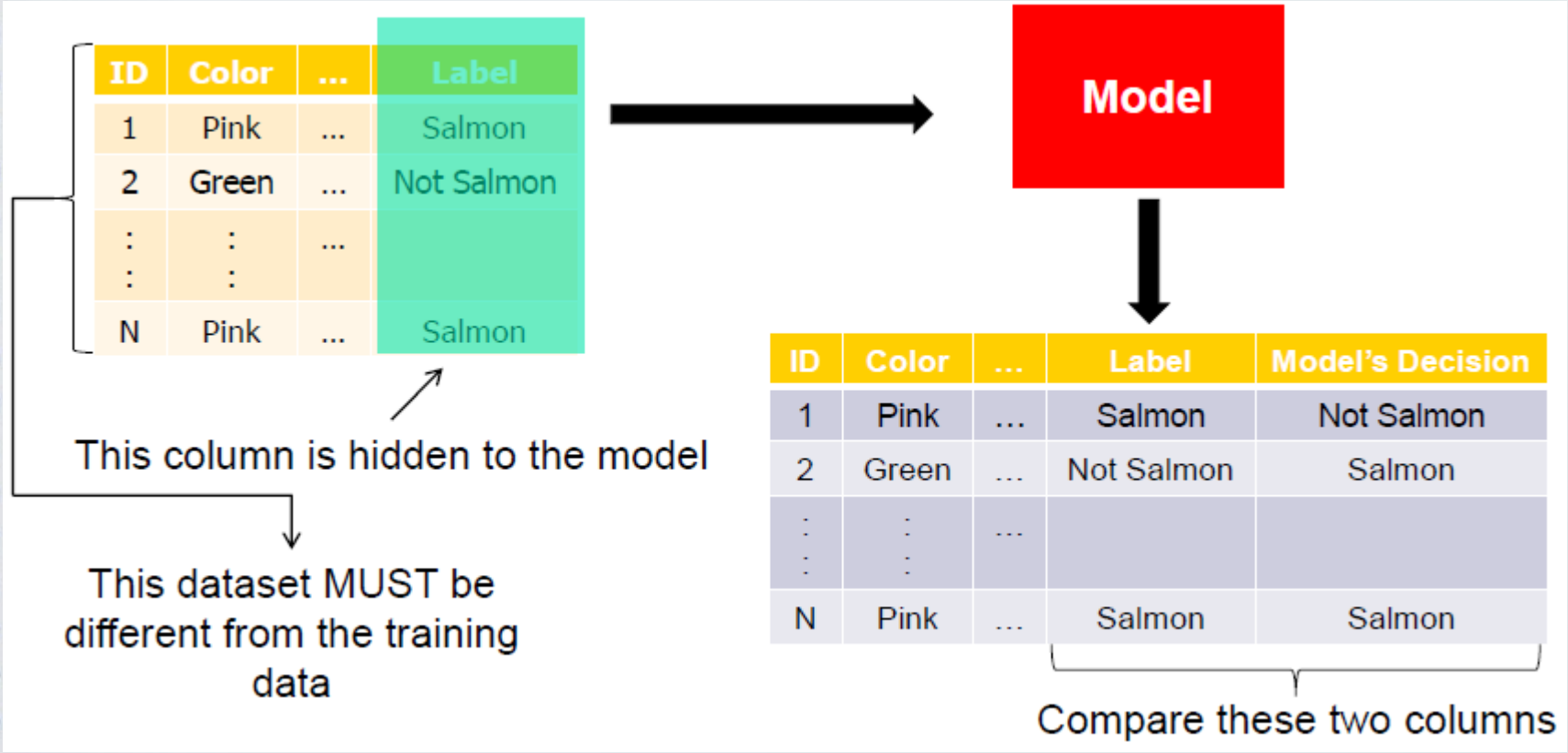
Training Data

ID	Color	Size	...	Label
2	Green	30cm	...	Not Salmon
6	Grey	12cm	...	Not Salmon
:	:	:	...	:
M	Pink	18cm	...	Salmon

Testing Data

Testing

- Proses testing



Evaluasi Model: Metrik Pengukuran Kinerja

- Confusion Matrix

		Prediction	
		Salmon	Not Salmon
Actual Class	Salmon	A	B
	Not Salmon	C	D

A: TP (true positive) B: FN (false negative)

C: FP (false positive) D: TN (true negative)

$$\text{Accuracy} = \frac{A+D}{A+B+C+D} = \frac{TP+TN}{TP+TN+FP+FN}$$

Latihan Evaluasi Model

- Buat Confusion Matrix dari table hasil klasifikasi berikut dan tentukan berapa akurasi klasifikasinya

Car Dataset

Example No.	Color	Type	Origin	Stolen?	
				Label	Model's Decision
1	Red	Sports	Domestic	Yes	Yes
2	Red	Sports	Domestic	No	Yes
3	Red	Sports	Domestic	Yes	No
4	Yellow	Sports	Domestic	No	No
5	Yellow	Sports	Imported	Yes	Yes
6	Yellow	SUV	Imported	No	No
7	Yellow	SUV	Imported	Yes	No
8	Yellow	SUV	Domestic	No	Yes
9	Red	SUV	Imported	No	No
10	Red	Sports	Imported	Yes	Yes

Keterbatasan Akurasi

- Misalkan..
  - Jumlah total ikan di sampel testing = 10.000
  - Jumlah ikan bukan salmon = 9990
  - Jumlah ikan salmon = 10
- Jika model memprediksi semuanya sebagai ikan bukan salmon, maka akurasinya = 9990/10000 = 99,9%
  - Akurasi dapat menyesatkan karena model tidak dapat mendeteksi salmon sama sekali

Precision, Recall, dan F-Measure

- Mengukur efektivitas model:

$$\text{Precision, } p = \frac{A}{A+C}$$

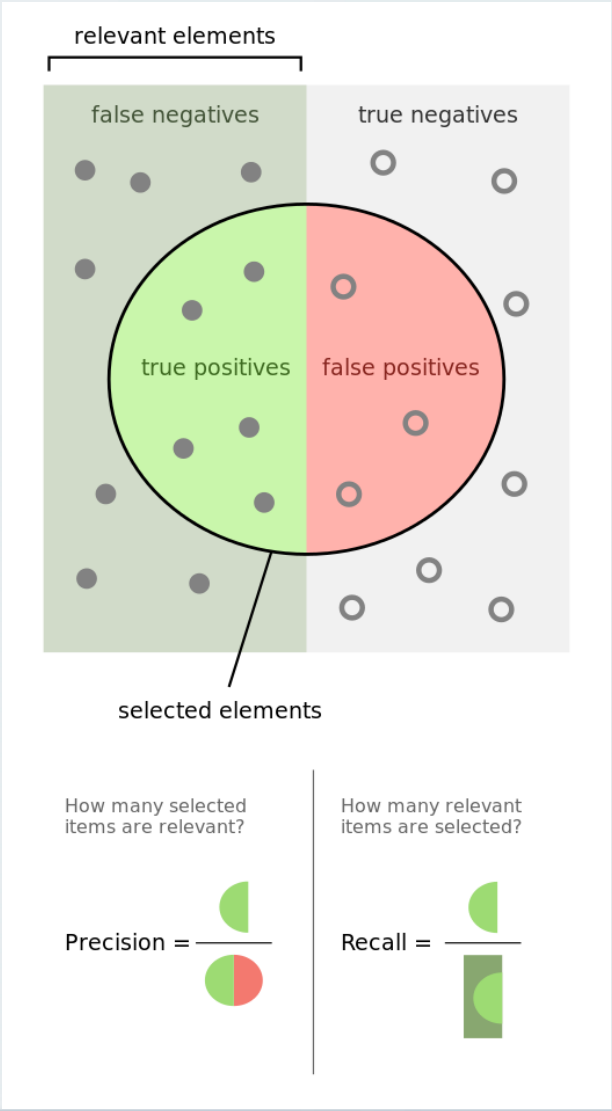
$$\text{Recall, } r = \frac{A}{A+B}$$

$$\text{F-measure} = \frac{2rp}{r+p}$$

p = banyaknya result terambil yang relevan
r = banyak result relevan yang terambil
F-measure = rata-rata harmonis dari precision dan recall

		Prediction	
		Salmon	Not Salmon
Actual Class	Salmon	A	B
	Not Salmon	C	D

Precision dan Recall



Sumber: https://upload.wikimedia.org/wikipedia/commons/thumb/2/26/Precisionrecall.svg/525px-Precisionrecall.svg.png

Perbedaan antara Precision dan Recall

- Contoh result dari sebuah search engine terhadap sebuah query
  - Jika search engine mengembalikan 30 result dan hanya 20 result yang relevan dengan query, maka precision-nya: 20/30 = 2/3 = 66,66%
  - Jika sebenarnya ada 40 result lain yang relevan dengan query tapi tidak terambil, maka recall-nya: 20/60 = 1/3 = 33,33%

Evaluasi Model: Bagaimana memperoleh estimasi yang handal?

- Bagaimana cara untuk mempartisi data antara training set dan test set agar diperoleh hasil estimasi yang handal?
- Tiga metode:
  - Holdout
  - Cross validation
  - Estimasi Leave-one-out

1. Holdout

- Cocok untuk data berukuran besar
- Secara acak, ambil 70% data sebagai training set dan 30% sebagai test set
- Ulangi prosedur di atas beberapa kali (misal: 10 kali)

2. *k*-fold Cross Validation

- Partisi/bagi data dalam *k* sub-himpunan terpisah
- Lakukan training pada (k-1) partisi, lakukan tes pada partisi terakhir
- Proses diulangi sebanyak *k* kali, dimana tiap subsampel k digunakan tepat satu kali sebagai test data/validation data
- Sebanyak *k* hasil yang didapat lalu dirata-rata untuk menghasilkan satu estimasi final
- Cocok untuk data berukuran medium

3. Validasi Leave-one-out

- Versi sederhana dari cross validation
- Misal kita memiliki *N* data
- Ambil (*N*-1) data sebagai training, dan 1 data terakhir sebagai testing
- Ulangi eksperimen sebanyak *N* kali
- Cocok untuk data berukuran kecil

Summary

- Prosedur umum untuk membangun sebuah model klasifikasi:
  - Membagi data menjadi training dan testing
  - Melatih classifier
  - Melakukan test pada classifier:
    - Akurasi tidak terlalu baik untuk dijadikan acuan
    - Precision, recall, f-measure
  - Mengkombinasikan keputusan dari beberapa classifier
    - Majority vote
    - Weighted Linear Combination
  - Binary-Class vs Multi-Class

Summary

- Klasifikasi adalah permasalahan yang banyak dipelajari dalam bidang statistik, machine learning, dan neural network
- Klasifikasi mungkin merupakan salah satu dari teknik data mining yang banyak digunakan dengan berbagai macam pengembangan
- Arah penelitian: klasifikasi pada data non-relational. Misal: text, spasial, multimedia