Pattern Recognition

Yukyung Choi

yk.choi@rcv.sejong.ac.kr

Summary

- Decision Tree & Random Forests I
  - Hand on Labs: Data Classification using library

- Decision Tree & Random Forests II
  - Hand on Labs: Implementation of DT-RF

Today's Lecture

- Definition of Decision Tree
- Introduction to Decision Tree
- Random Forest
- Hand on Labs

What is a Decision Tree?

- Definition
  - A **decision tree** is a **decision** support tool that uses a **tree**-like graph
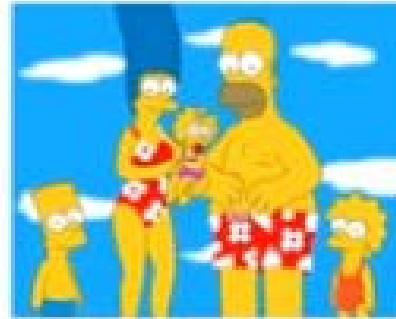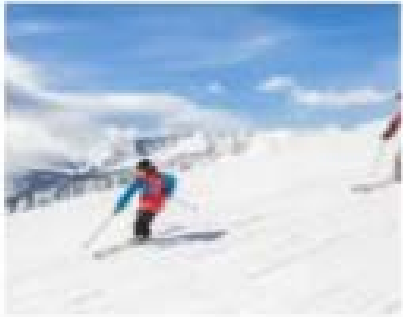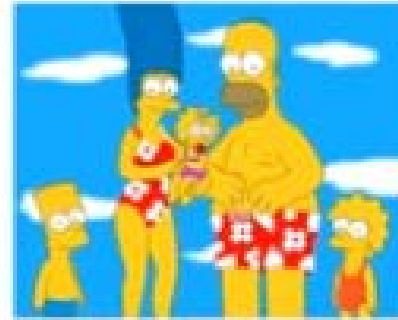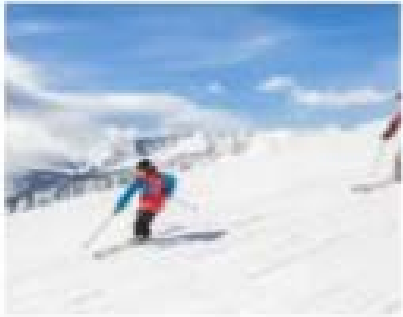
How Decision Tree works

How Decision Tree works

How do you teach your baby to pick **winter family vacation photo**?
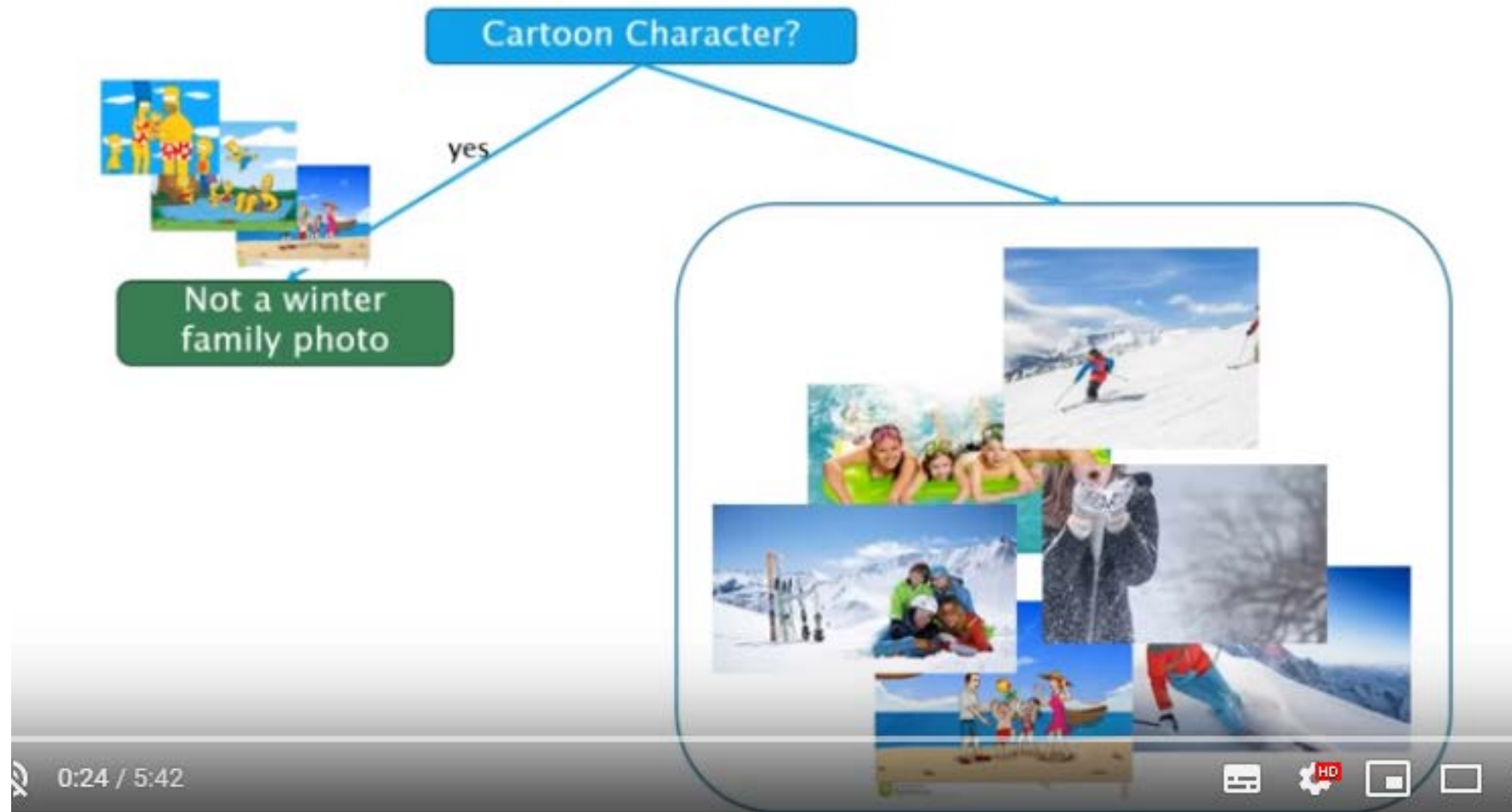
How Decision Tree works

- You may say to consider **snow, human, and the number of members**
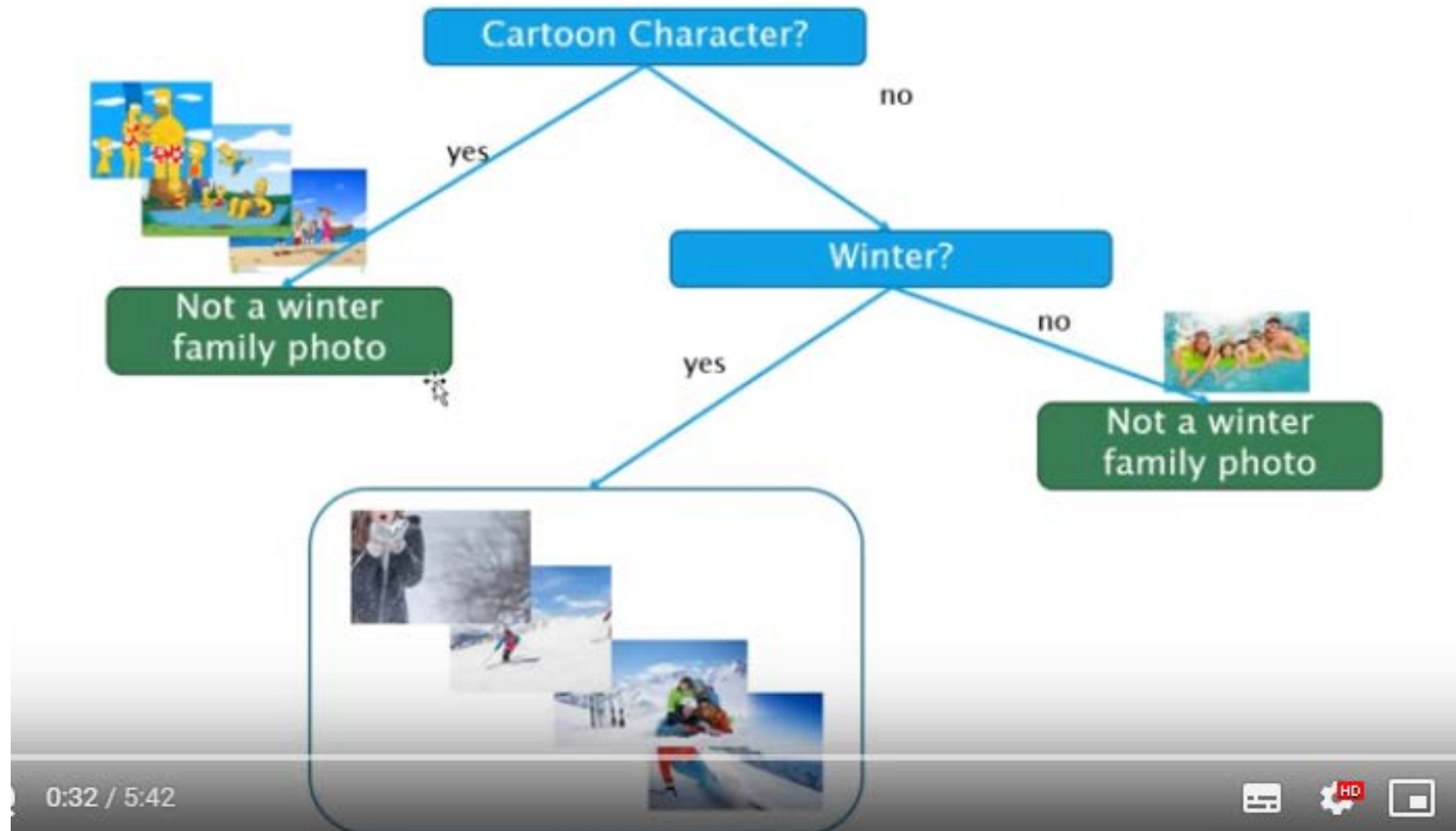
How Decision Tree works

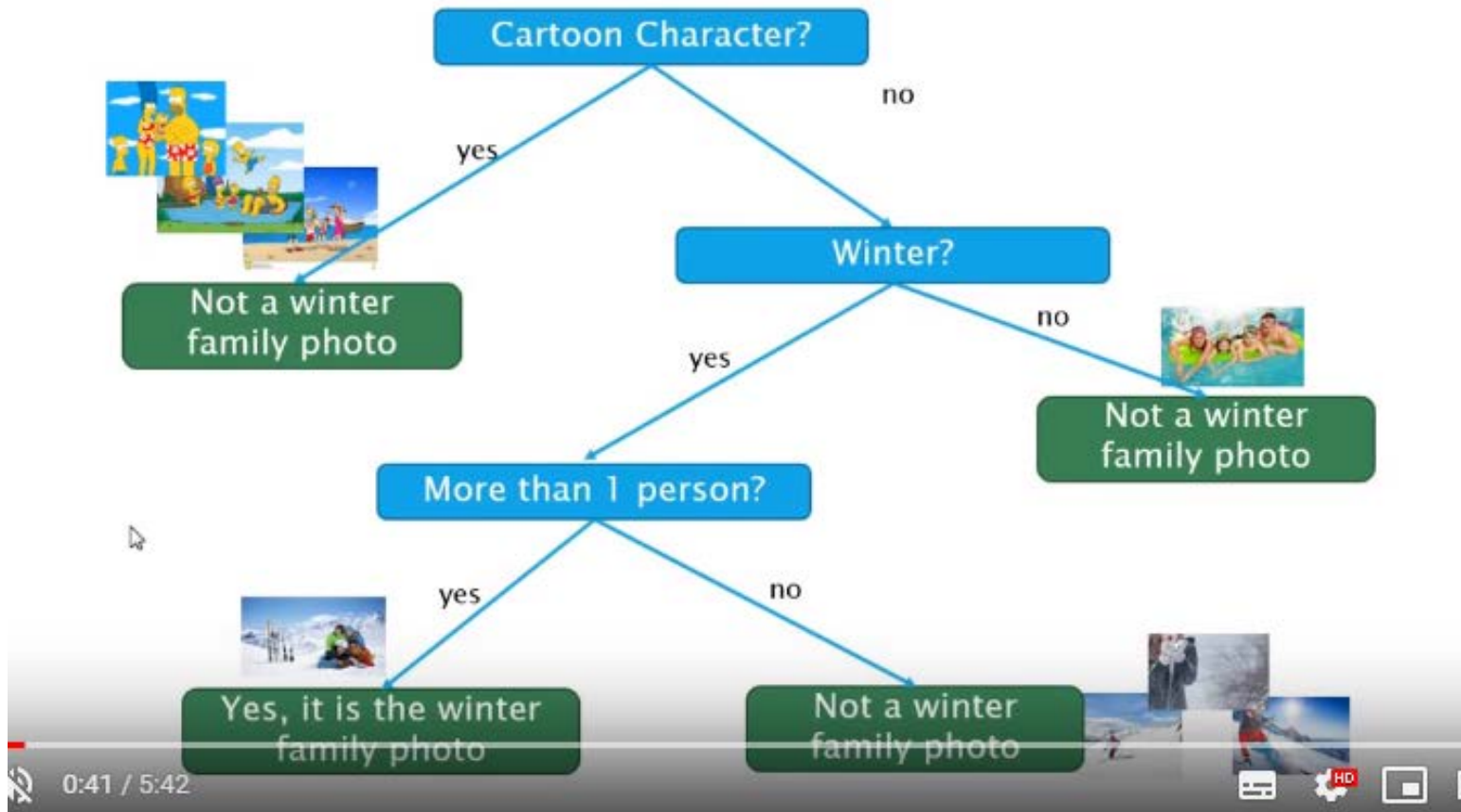- We definitely know the cartoon is not a family photo

How Decision Tree works

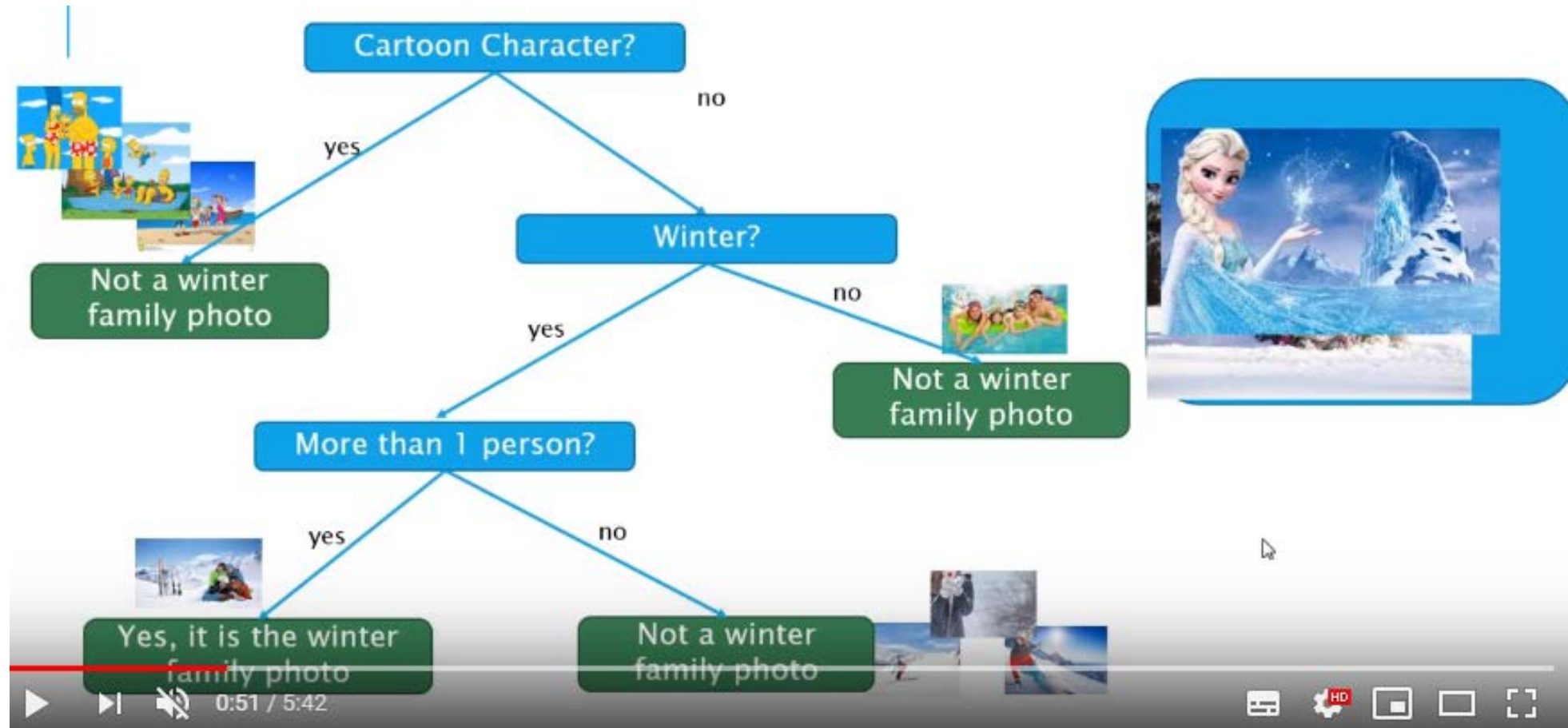- We definitely know summer is not a winter photo

How Decision Tree works

- We know family photo has more than 1 person

How Decision Tree works

• Now, your baby can recognize more pictures **using this decision tree** ☺
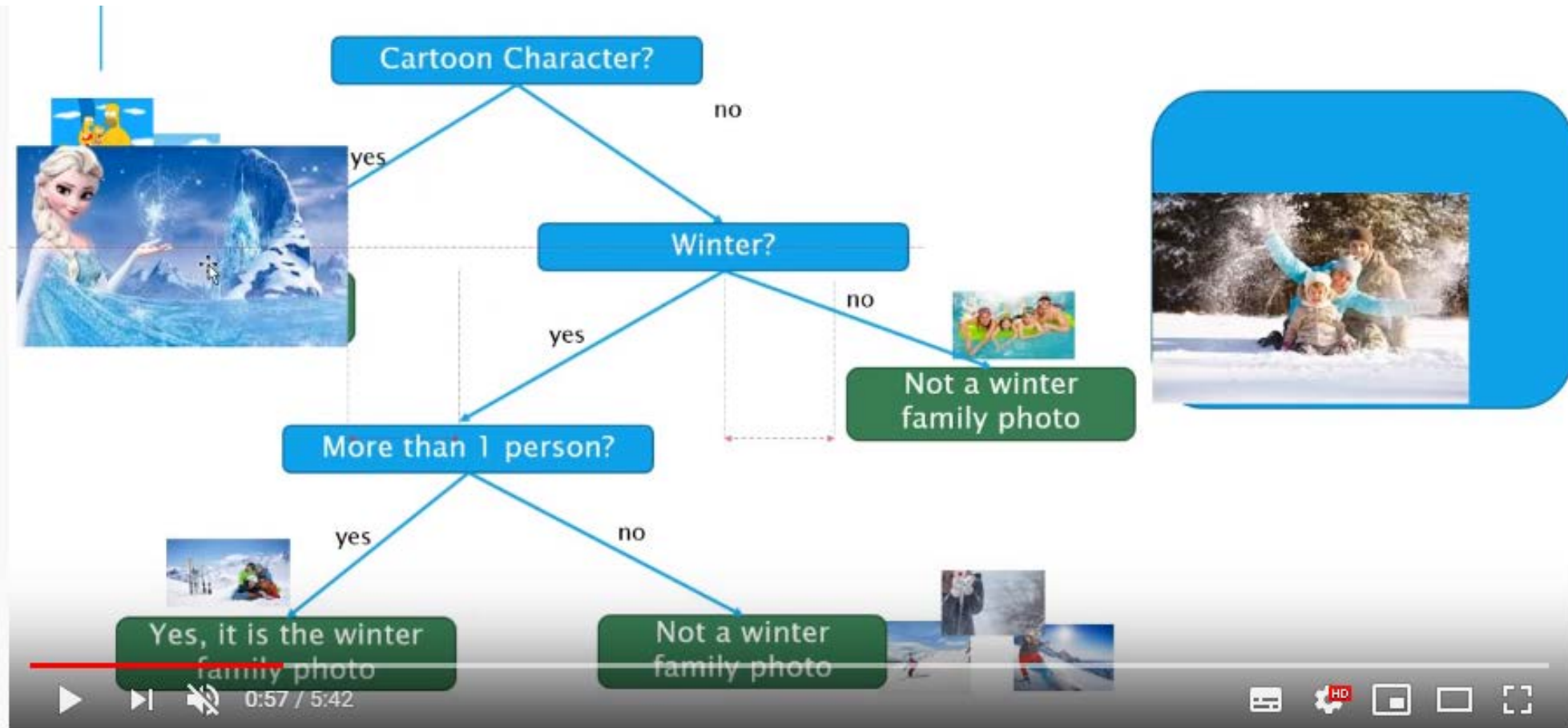
How Decision Tree works

- Now, your baby can recognize more pictures **using this decision tree** ☺

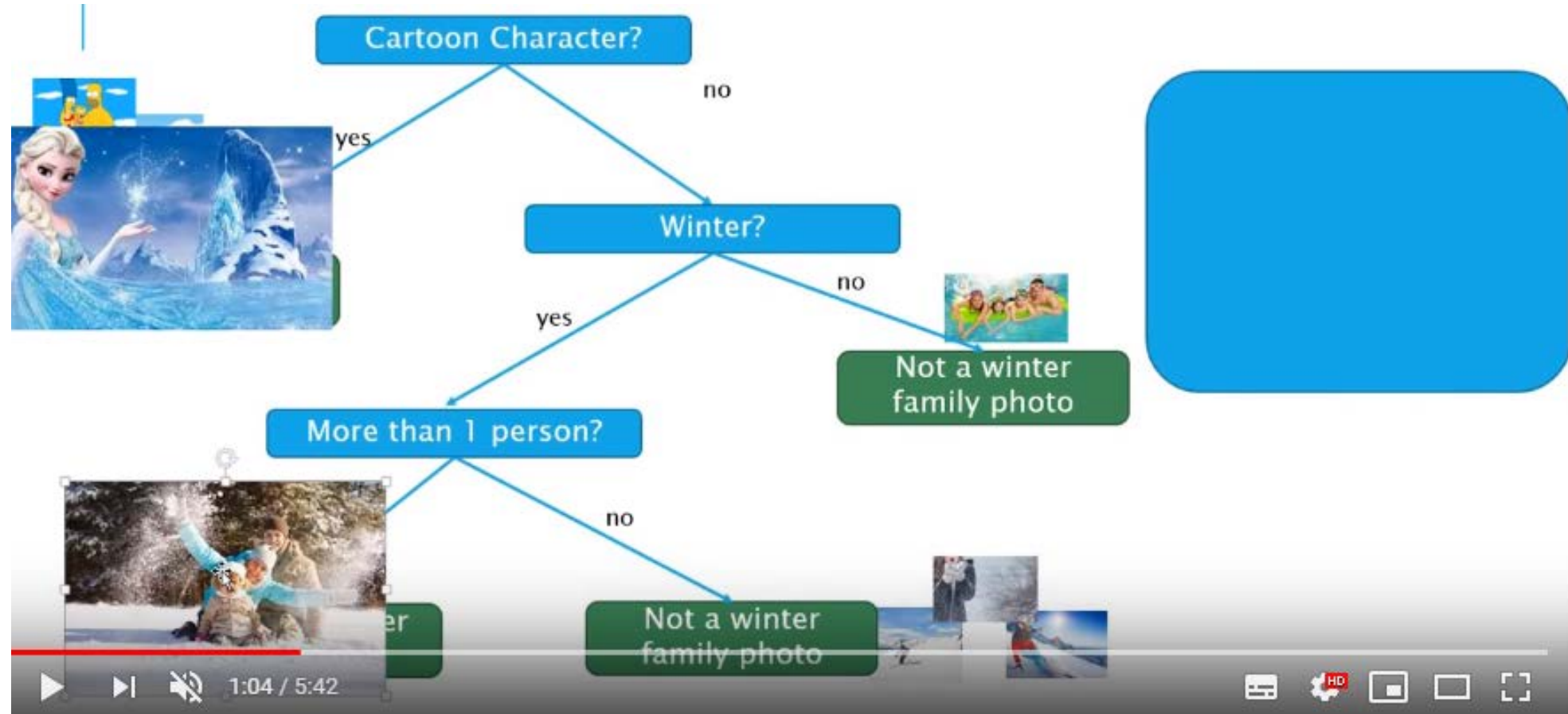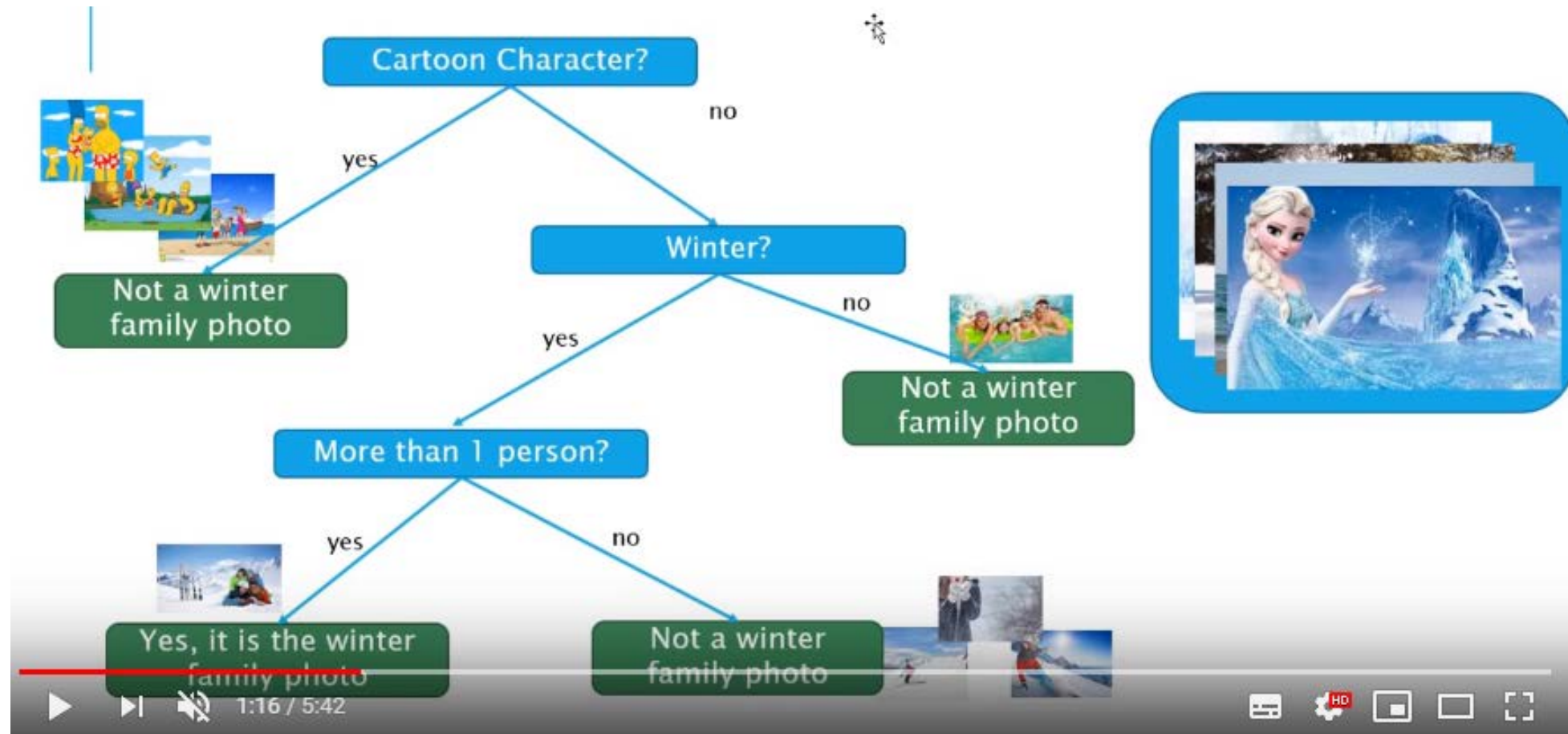# How Decision Tree works

- Now, your baby can recognize more pictures **using this decision tree** ☺

How Decision Tree works

- Yes, the **baby** is the **machine**, and the machine learned from you

Decision Tree, two main concepts

- Define Problem
- Collect training data

Build a tree
- Extract Data
- Build a tree

- Deploy machine

Test
- Test with test data

Build a Decision Tree: Extract training data

img	cartoon	winter	> 1	Family winter photo
	No	Yes	Yes	Yes
	No	Yes	No	No
	Yes	No	Yes	No
	Yes	Yes	Yes	No
	No	Yes	No	No
	No	No	Yes	No
	Yes	No	Yes	No
	yes	yes	no	no

Build a Decision Tree
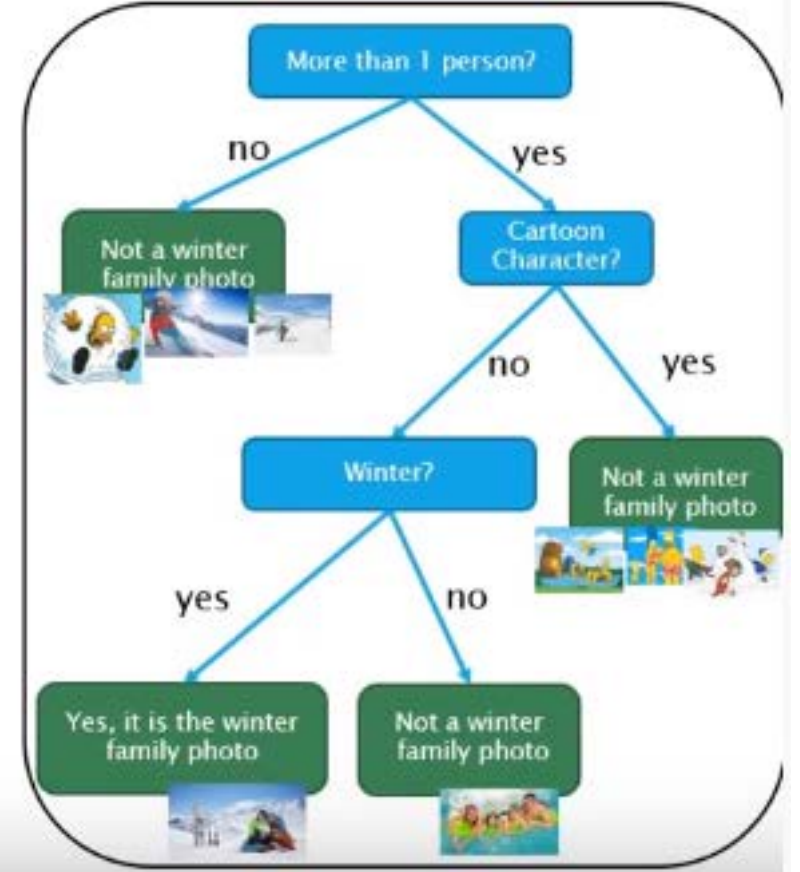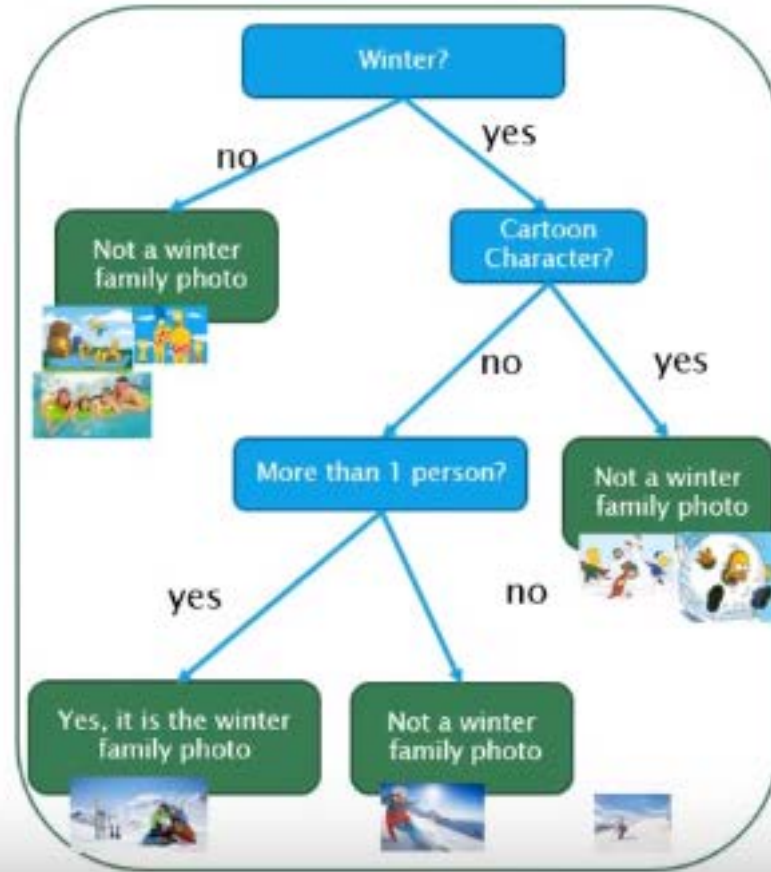
Build a Decision Tree
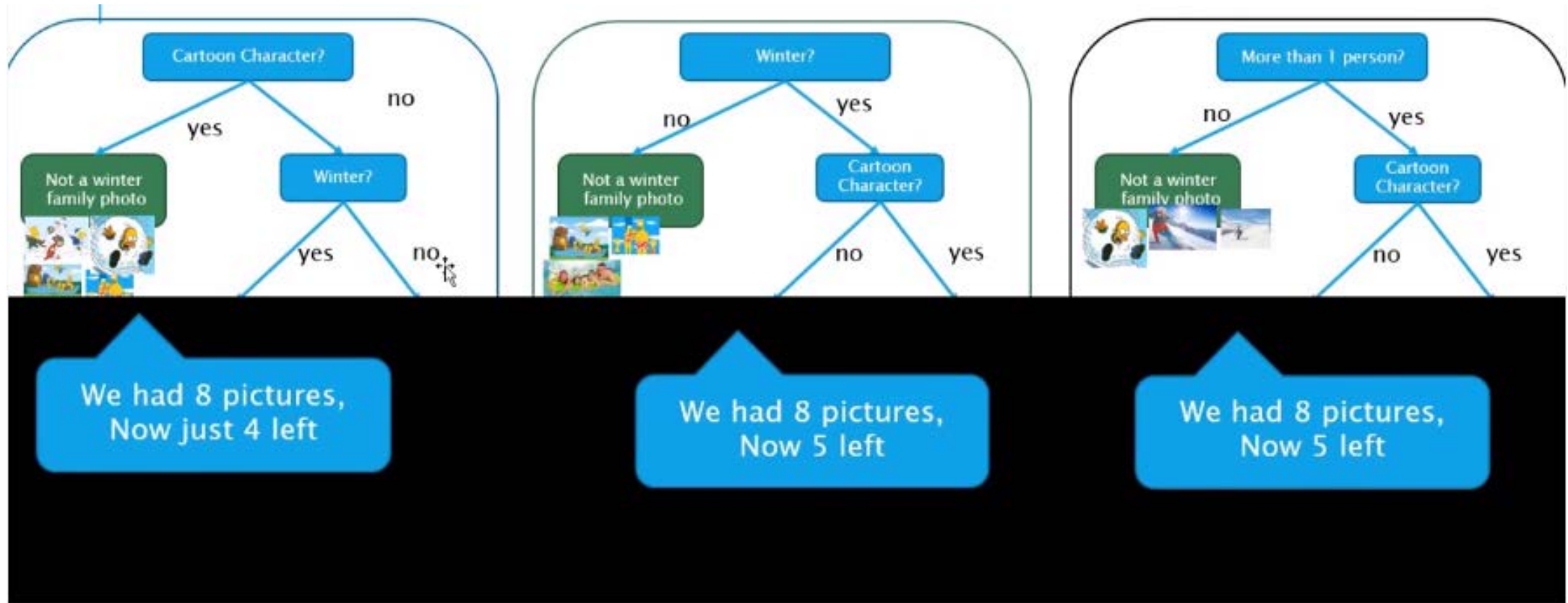
- Choose best feature to split

Build a Decision Tree

- Choose best feature to split

Build a Decision Tree

- Choose best feature to split

Terminologies

- Entropy
- Information Gain

Terminologies

- Entropy

High Entropy (messy)   Low Entropy (Clean)

Terminologies

- Entropy: measure of messiness

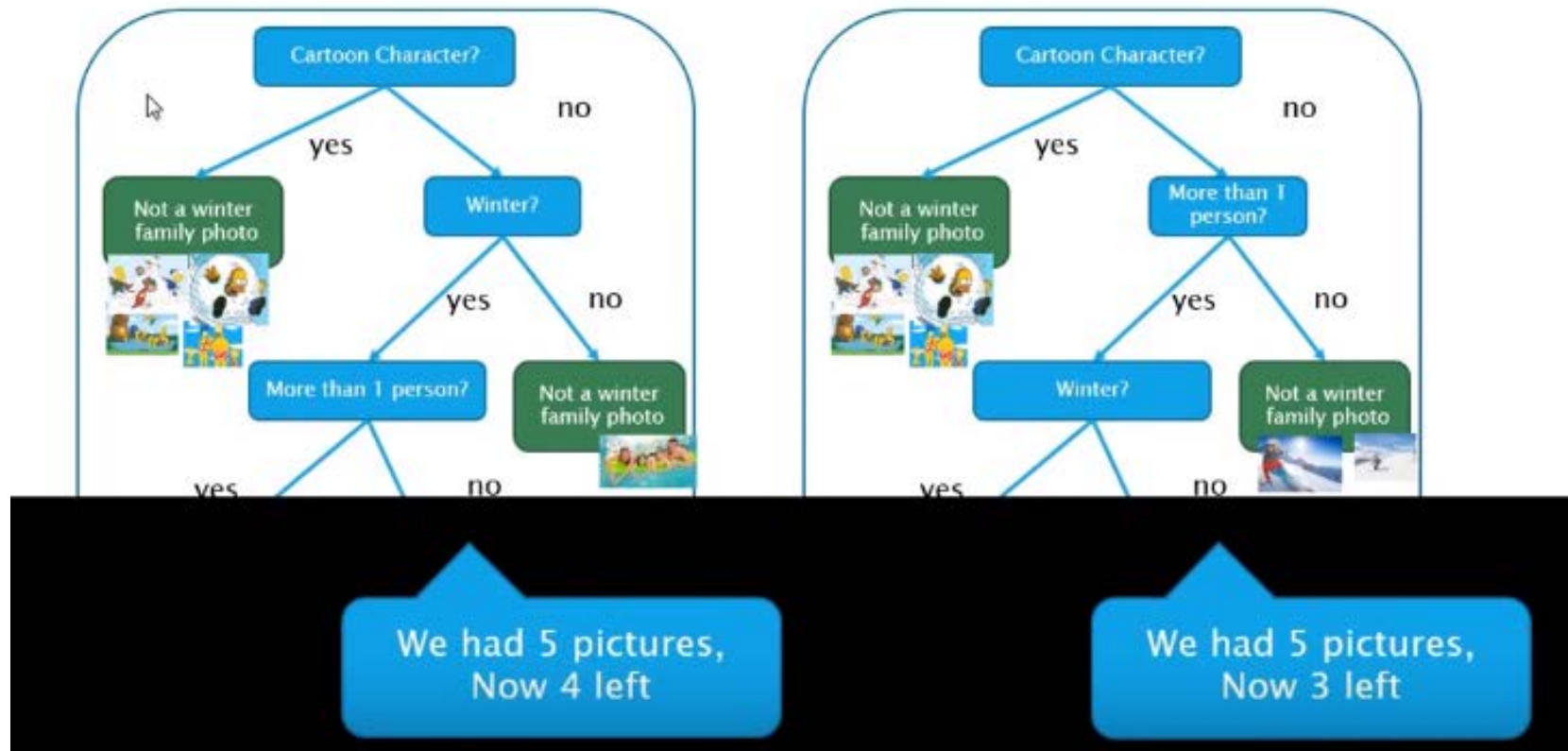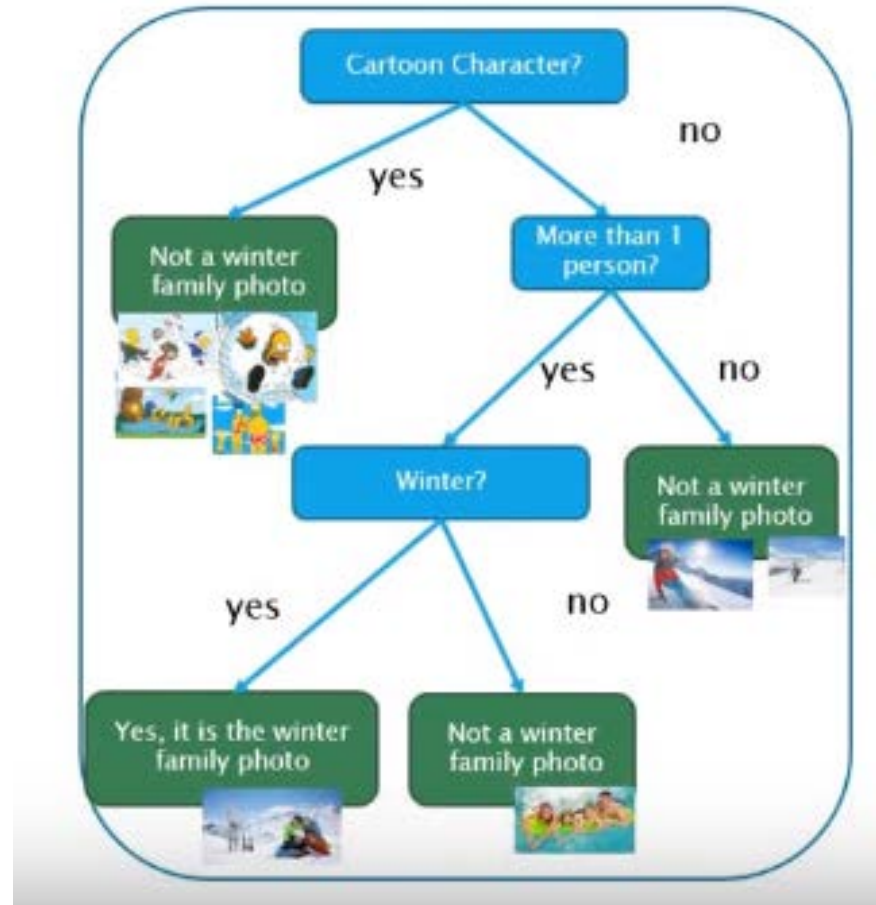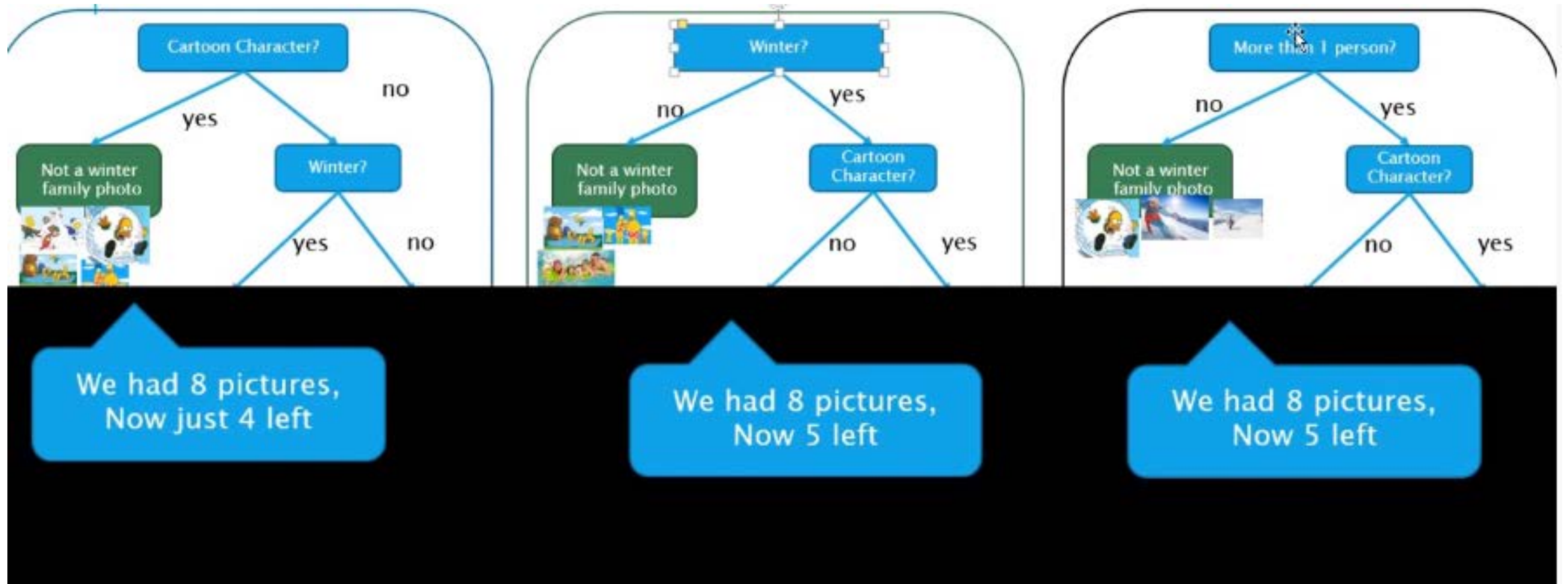# Terminologies

- Information Gain: Base entropy – New entropy
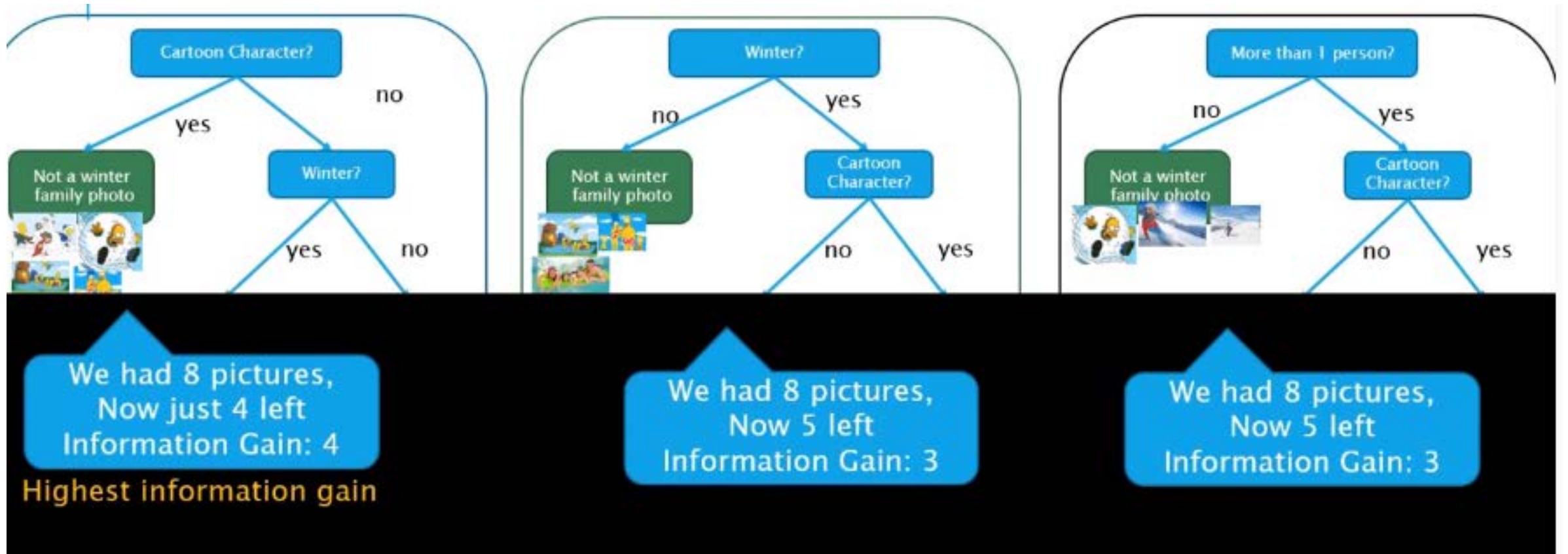
ID3 Algorithm

• Entropy and Information Gain

img	cartoon	winter	> 1	Family winter photo
	No	Yes	Yes	Yes
	No	Yes	No	No
	Yes	No	Yes	No
	Yes	Yes	Yes	No
	No	Yes	No	No
	No	No	Yes	No
	Yes	No	Yes	No
	yes	yes	no	no

High Entropy (messy)
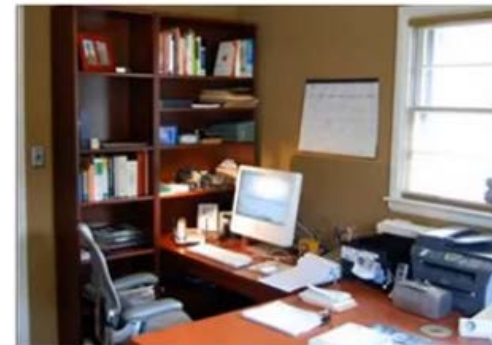
Low Entropy (Clean)

ID3 Algorithm

- Calculate Entropy to find family photo of 8 photos

img	cartoon	winter	> 1	Winter family photo
	No	Yes	Yes	Yes
	No	Yes	No	No
	Yes	No	Yes	No
	Yes	Yes	Yes	No
	No	Yes	No	No
	No	No	Yes	No
	Yes	No	Yes	No
	yes	yes	no	no

Before Decision Tree Building

Total 8 photos
1 photo winter family photo
7 photos Not winter family photo
= Entropy([1+, 7−])
$$= -(1/8)*\log(1/8) - (7/8)*\log(7/8)$$
$$= 0.543$$

$$Entropy = - p(+)*\log(p(+)) - p(-)*\log(p(-))$$

ID3 Algorithm

- Information Gain
  - Decrease in entropy after a dataset is split on an attribute

img	cartoon	winter	> 1	Winter family photo
	No	Yes	Yes	Yes
	No	Yes	No	No
	Yes	No	Yes	No
	Yes	Yes	Yes	No
	No	Yes	No	No
	No	No	Yes	No
	Yes	No	Yes	No
	yes	yes	no	no

Information Gain(winter family photo, cartoon)
= E(winter family photo) – E(winter family photo, cartoon)
= 0.543 – (4/8 * E([0+,4-])+ 4/8 * E([1+, 3-]))
=0.138

Information Gain(winter family photo, winter)
= E(winter family photo) – E(winter family photo, cartoon)
= 0.543 – (5/8 * E([1+,4-])+ 3/8 * E([0+, 3-]))
=0.093

Information Gain(winter family photo, >1)
= E(winter family photo) – E(winter family photo, cartoon)
= 0.543 – (5/8 * E([1+,4-])+ 3/8 * E([0+, 3-]))
=0.093

Random Forests

- **Random forests** or **random decision forests** are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of **decision** trees
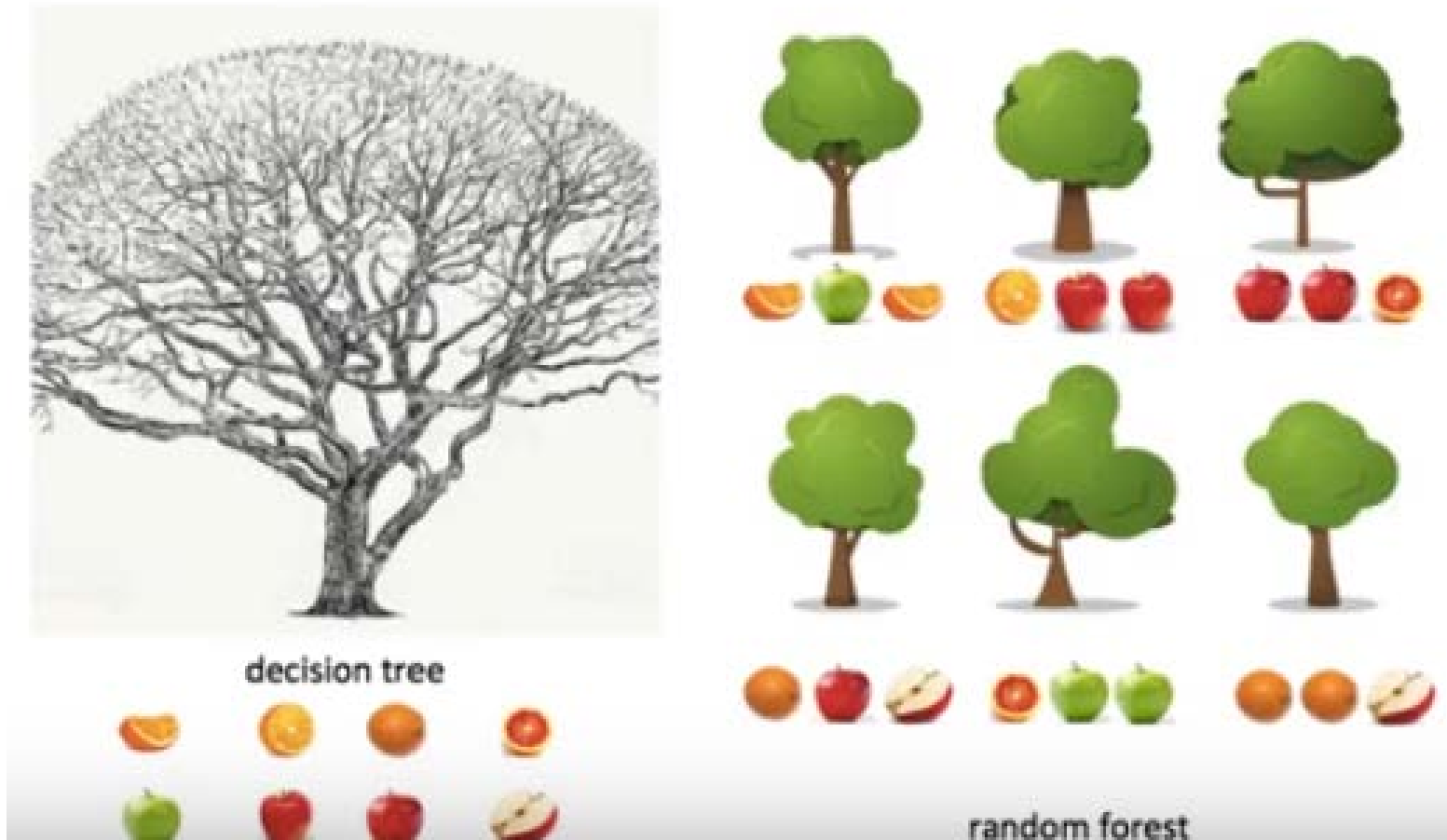
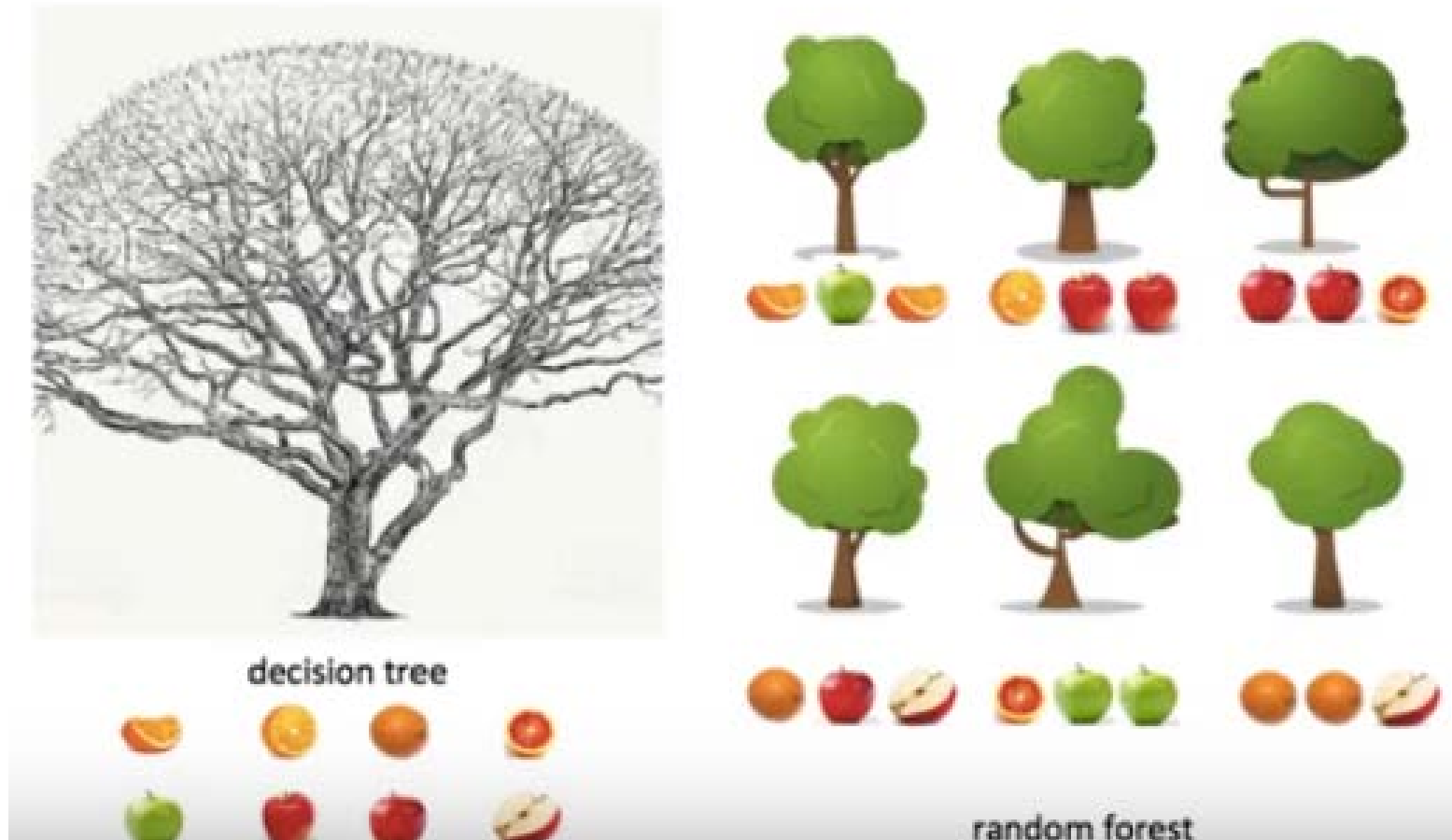# Decision Tree vs. Random Forests

decision tree

random forest

Decision Tree vs. Random Forests
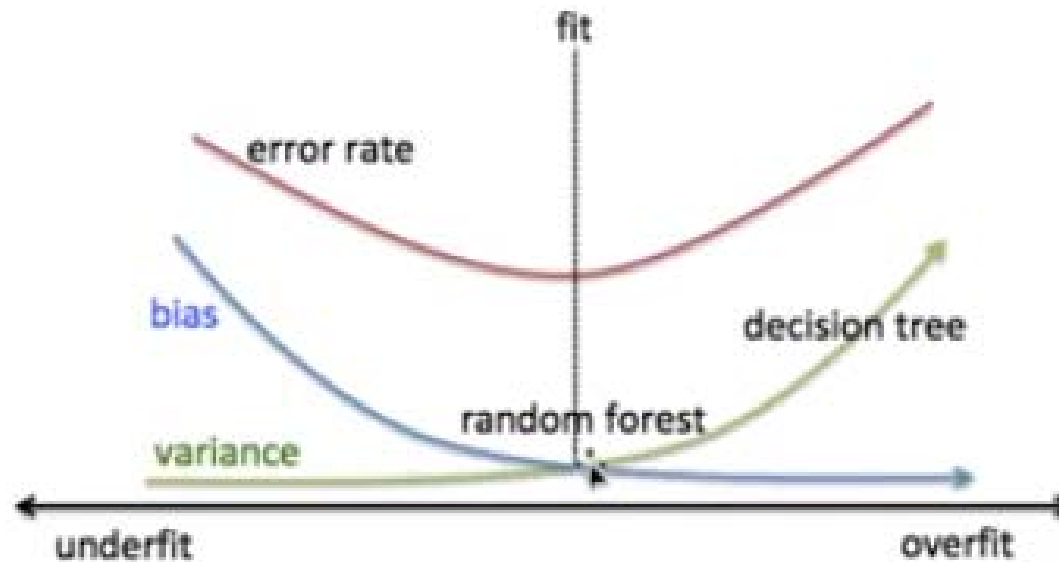
- Apple, orange classifier

decision tree

random forest

Decision Tree vs. Random Forests

- Boosting: biased data collection (➜ 중복 데이터 허용)

decision tree
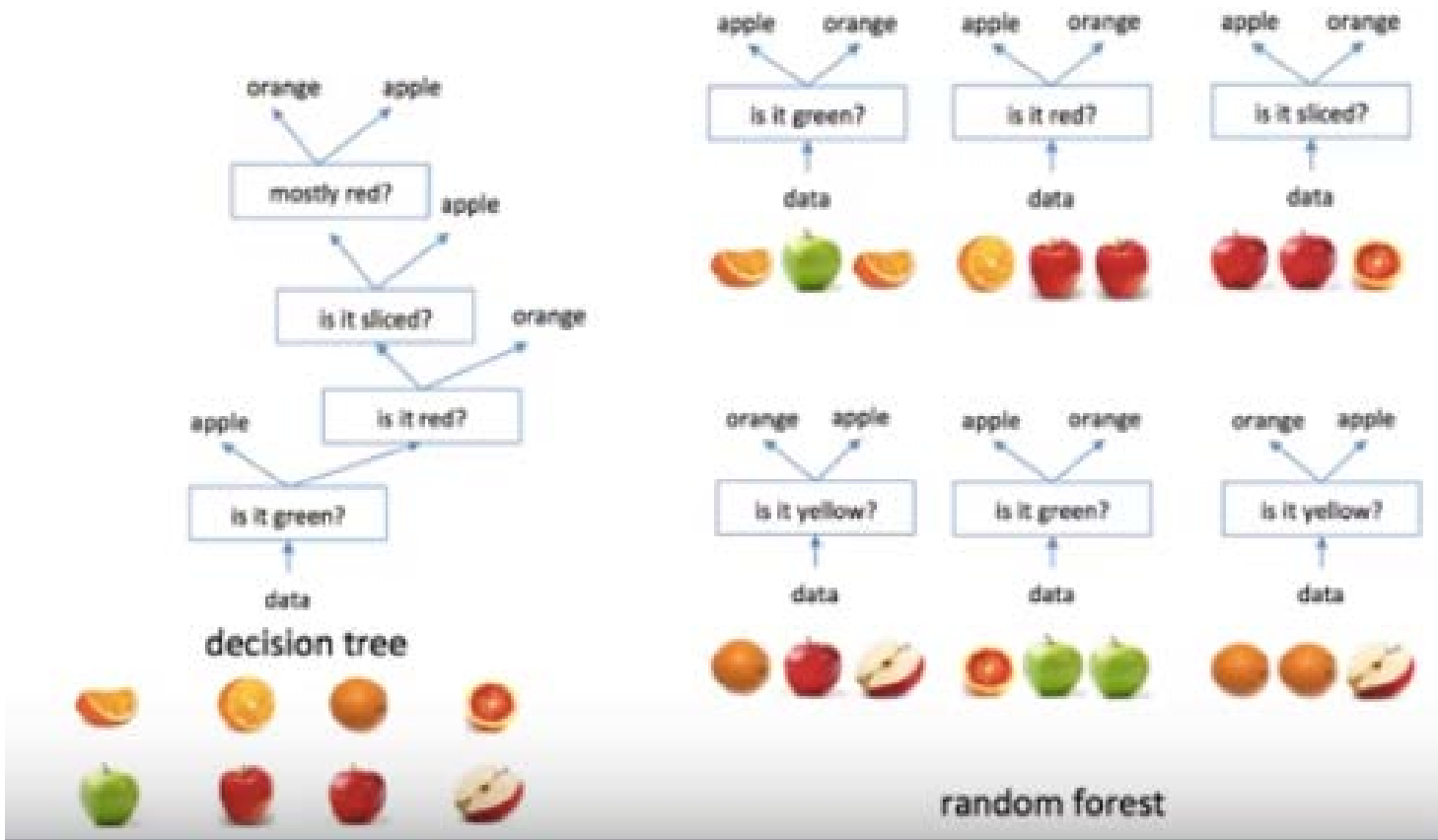
random forest

Decision Tree vs. Random Forests

- Avoid Overfitting

Decision Tree vs. Random Forests

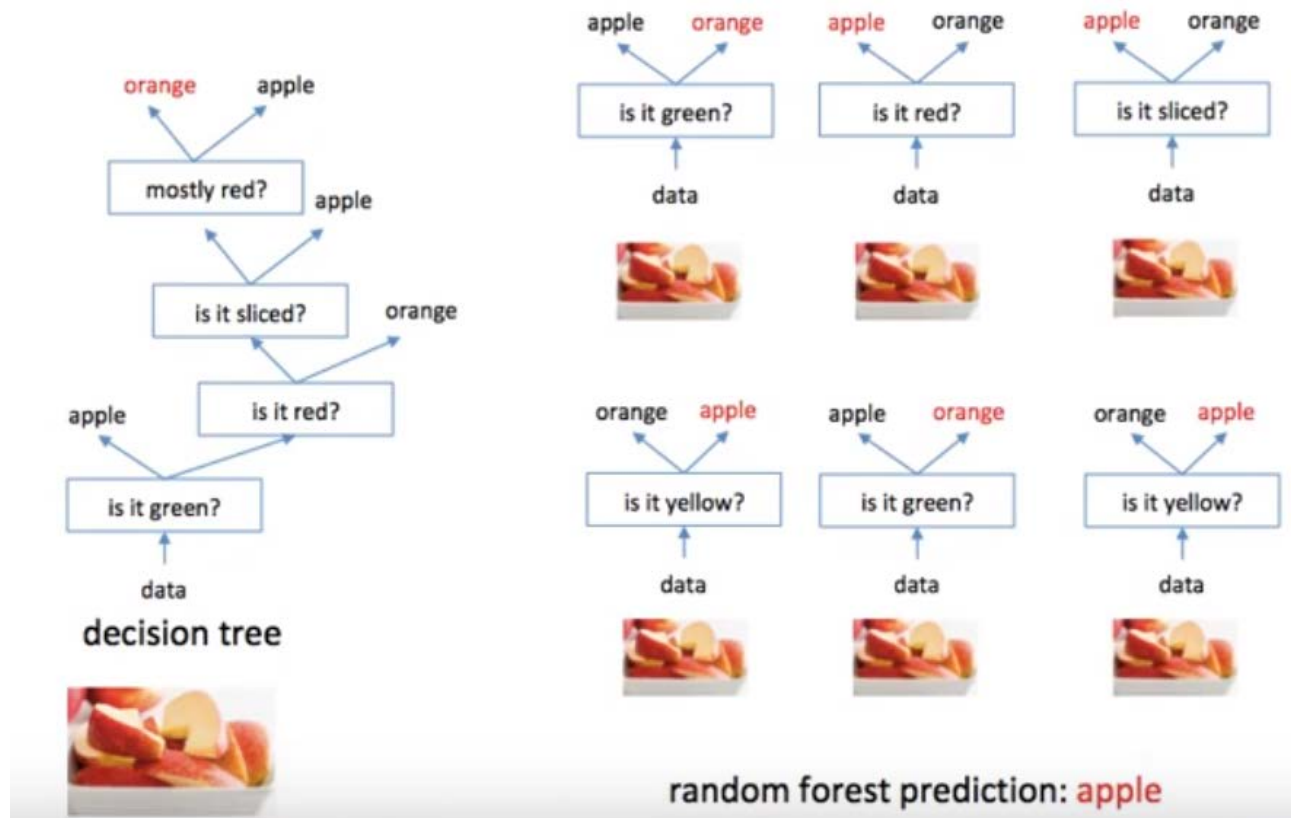- "Random selection" of feature set instead of "best selection"

가장 좋은
질문을
먼저 해야 한다

질문이
랜덤으로
셀렉션
된다

Decision Tree vs. Random Forests

- Aggregating of result (voting)

Reference

- [머신러닝] 의사결정트리 (Decision Tree) 알고리즘 쉽게 이해하기
- [머신러닝] 의사결정트리 (Decision Tree) - ID3 알고리즘 수학적 접근
- [머신러닝] 랜덤포레스트