Decision Tree

INTRODUCTION TO INTELLIGENT SYSTEMS ’17/’18

Example – fruit classification

Ten classes:

- Apple
- Banana
- Blue berry
- Grapes
- Lemon
- Mango
- Melon
- Orange
- Peach
- Water melon

Features

Four features:

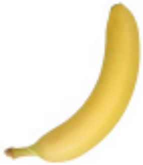
- Color: [yellow ; green ; red ; blue ; orange]
- Size: [xtra-small ; small ; medium
  large ; extra-large]
- Shape: [round ; elipsoidal ; narrow
  round-with-concavity]
- Texture: [smooth ; citrus]

Pattern representation

Feature Vectors:

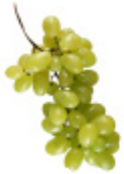
<color, size, shape, texture>

Examples:

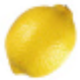


<y,m,n,s>

<b,xs,ro,s>



<g,sm,ro,s>



<y,m,ro,c>

Data set

Given a set S with 25 labeled patterns:

<r,m,rwc,s> ; <g,m,rwc,s> <y,m,rwc,s>

<y,m,n,s> ; <g,m,n,s> <y,l,n,s>

<b,xs,ro,s> ; <b,xs,ro,s>

<b,sm,ro,s> ; <g,sm,ro,s> <y,sm,ro,s>

<y,m,ro,c> ; <g,m,ro,c> <y,m,e,c>

<g,m,e,s> ; <y,m,e,s>

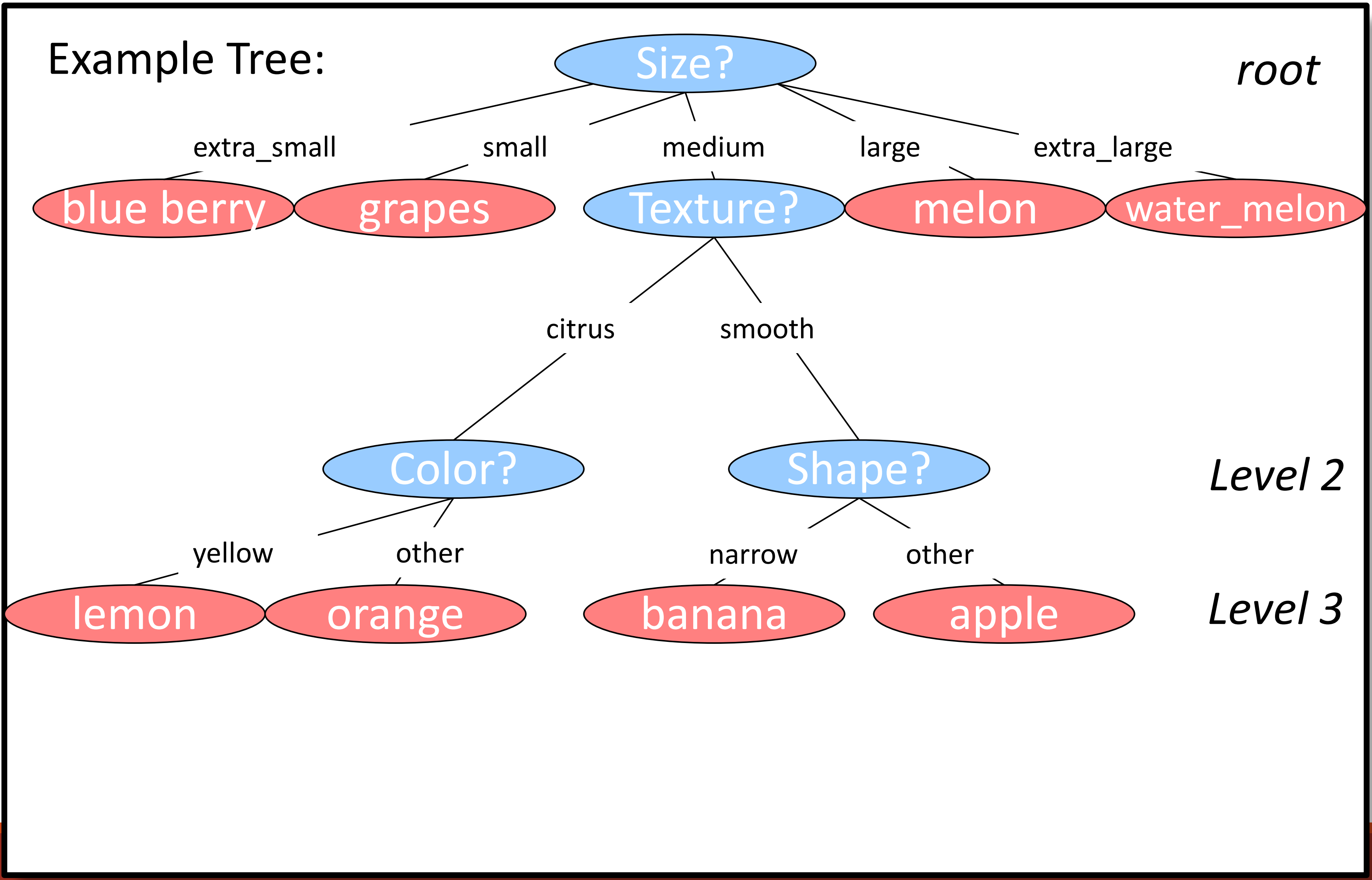
<y,l,ro,s> ; <y,l,e,s> <g,l,e,s>

<o,m,ro,c> ; <o,m,ro,c>

<y,m,rwc,s> ; <r,m,rwc,s>

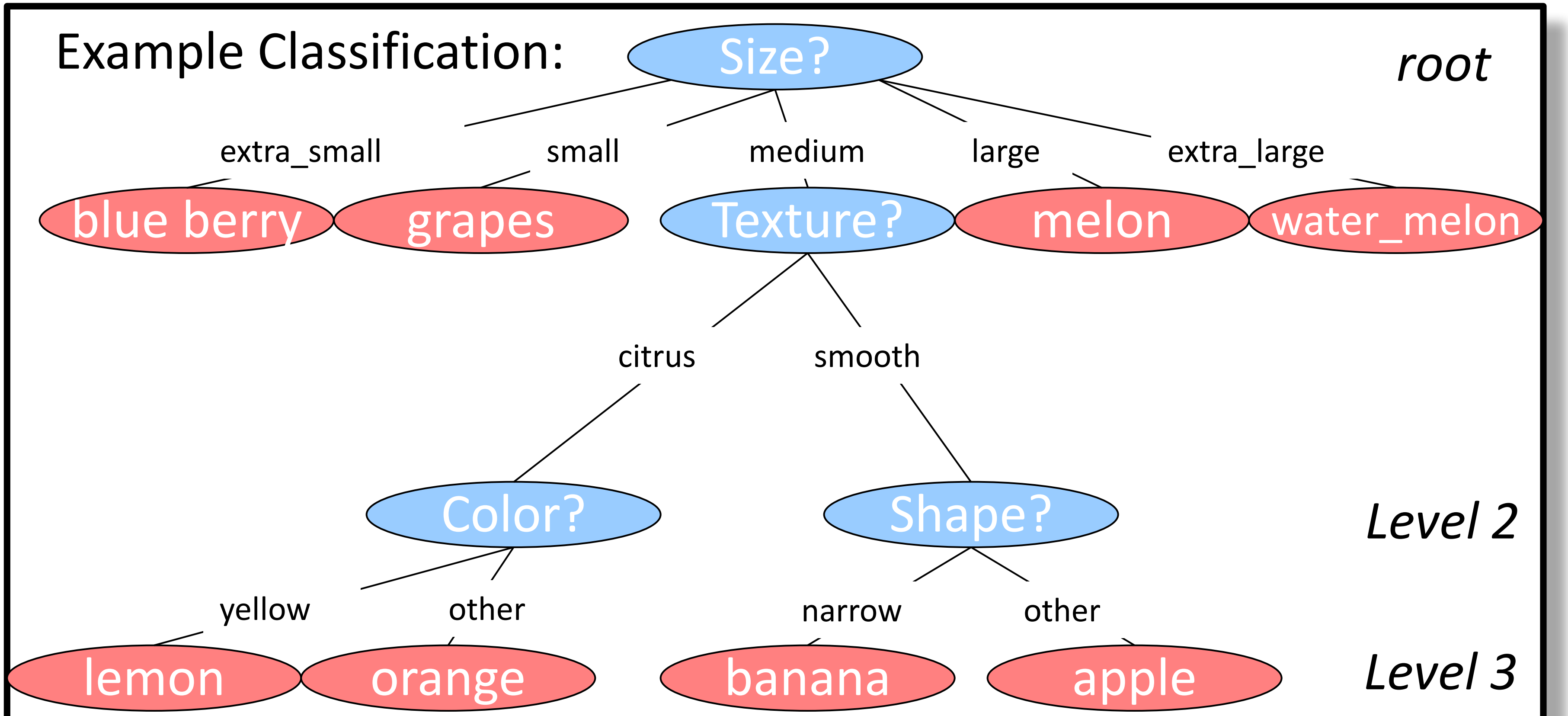
<g,xl,ro,s> ; <g,l,ro,s>

Example


Example Tree:
Size?
root
extra_small
small
medium
large
extra_large
blue berry
grapes
Texture?
melon
water_melon
citrus
smooth
Color?
Shape?
Level 2
yellow
other
narrow
other
lemon
orange
banana
apple
Level 3

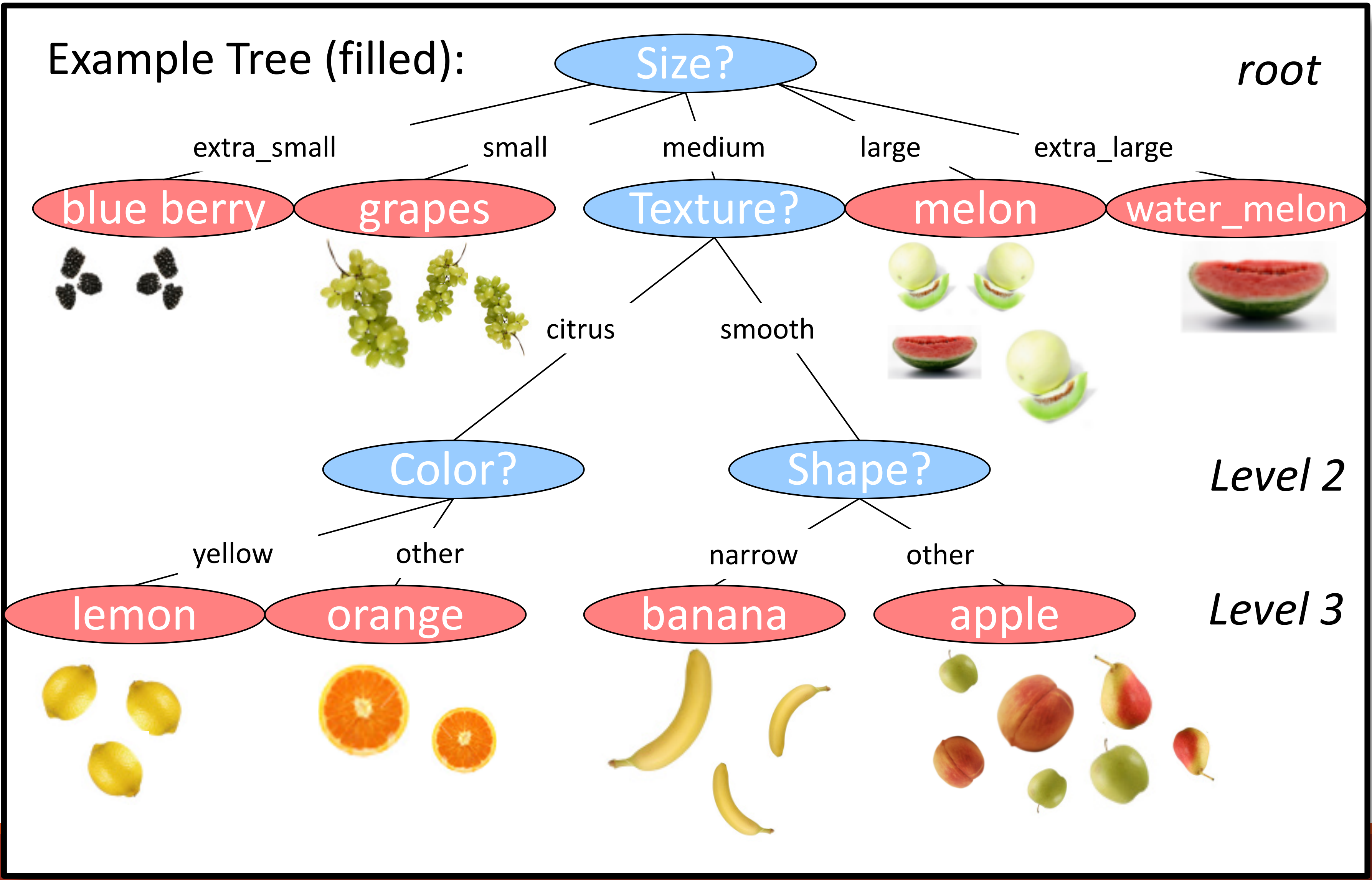
Example

Example Classification:

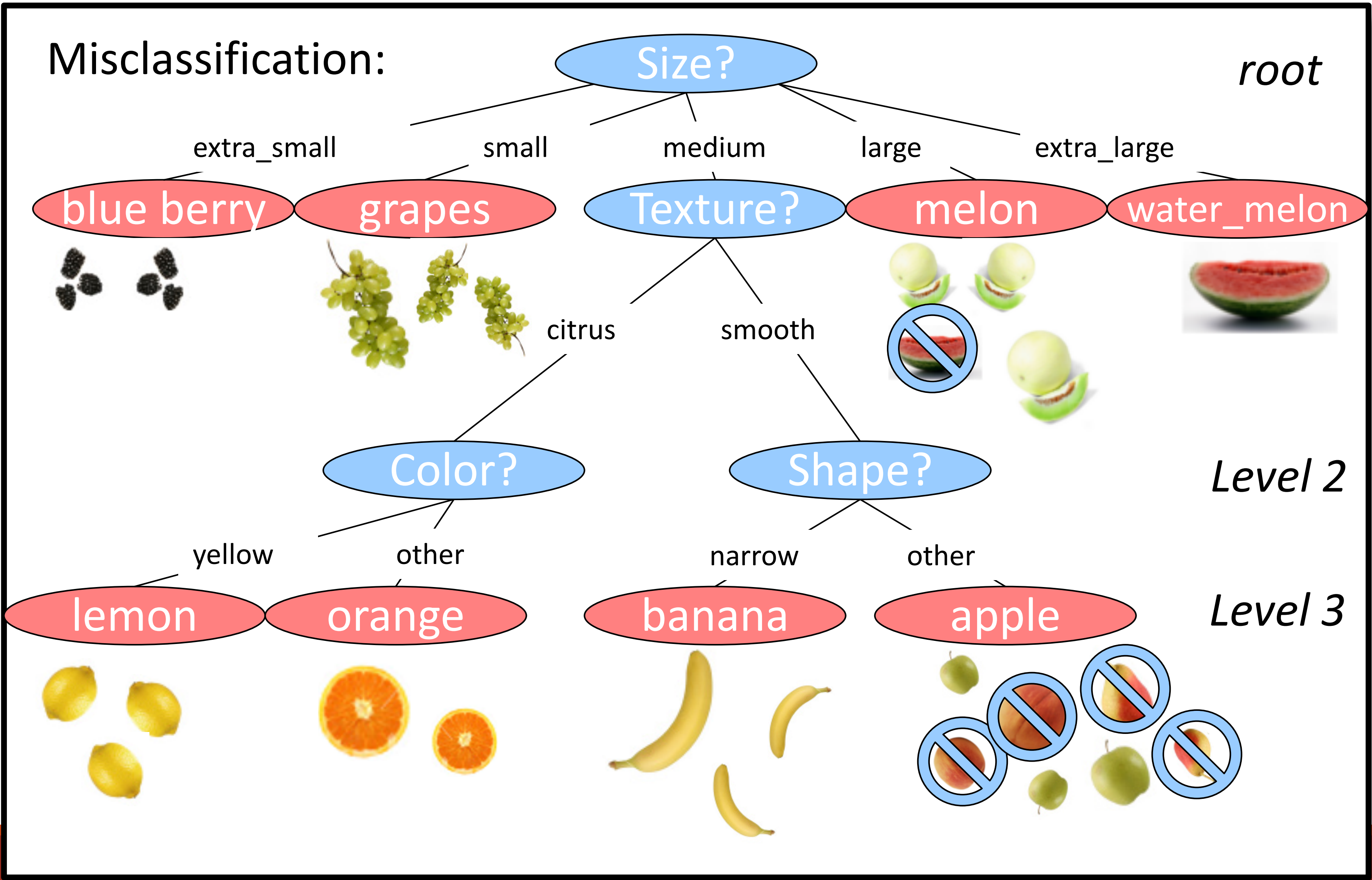


<o,m,ro,c>

Example


Example Tree (filled):
Size?
root
extra_small
small
medium
large
extra_large
blue berry
grapes
Texture?
melon
water_melon
citrus
smooth
Color?
Shape?
Level 2
yellow
other
narrow
other
lemon
orange
banana
apple
Level 3

Example


Misclassification:
Size?
root
extra_small
small
medium
large
extra_large
blue berry
grapes
Texture?
melon
water_melon
citrus
smooth
Color?
Shape?
Level 2
yellow
other
narrow
other
lemon
orange
banana
apple
Level 3

How to build a decision tree?

Design of a decision tree

Problem: Given a training data set, how to create a decision tree?

General approach: split the set using a given feature, split the subsets using other features, … , until you get **pure** (i.e. single class) or **nearly pure subsets**

Query selection and node impurity

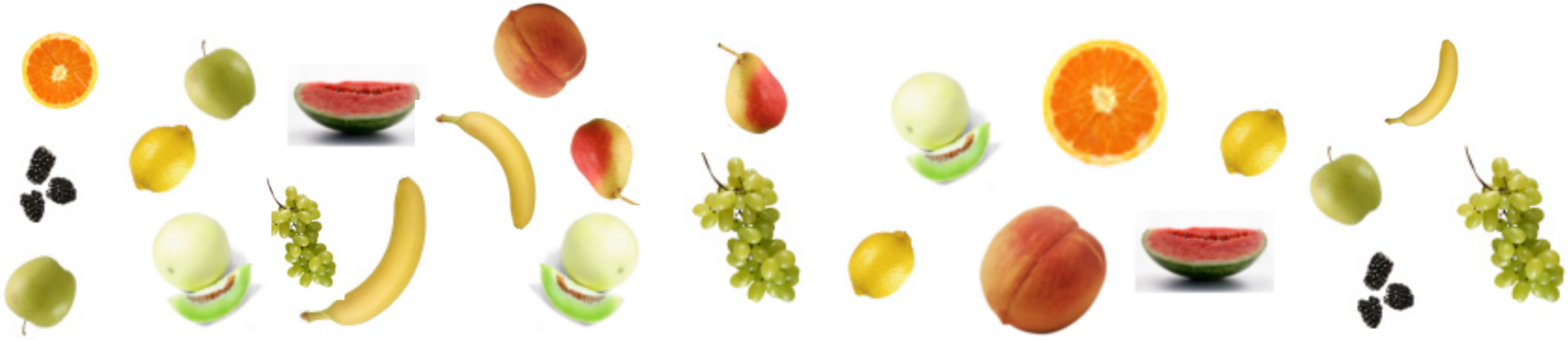
The query at a node is selected to minimize the impurity of the two subsets descending from the node.

Misclassification impurity: $i(N) = 1 - \max_j P(\omega_j)$

- Minimum probability of misclassifying a training pattern at N

Example

The impurity of set S:



$$i(N) = 1 - \max_{j} P(\omega_j)$$

$i_S = 1 - (3/25) = 22/25$

Query design

Impurity drop:

$$\Delta i(N) = i(N) - P_L\, i(N_L) - (1 - P_L)\, i(N_R)$$

N_L and N_R - left and right descendent nodes

P_L - fraction of patterns which go to N_L

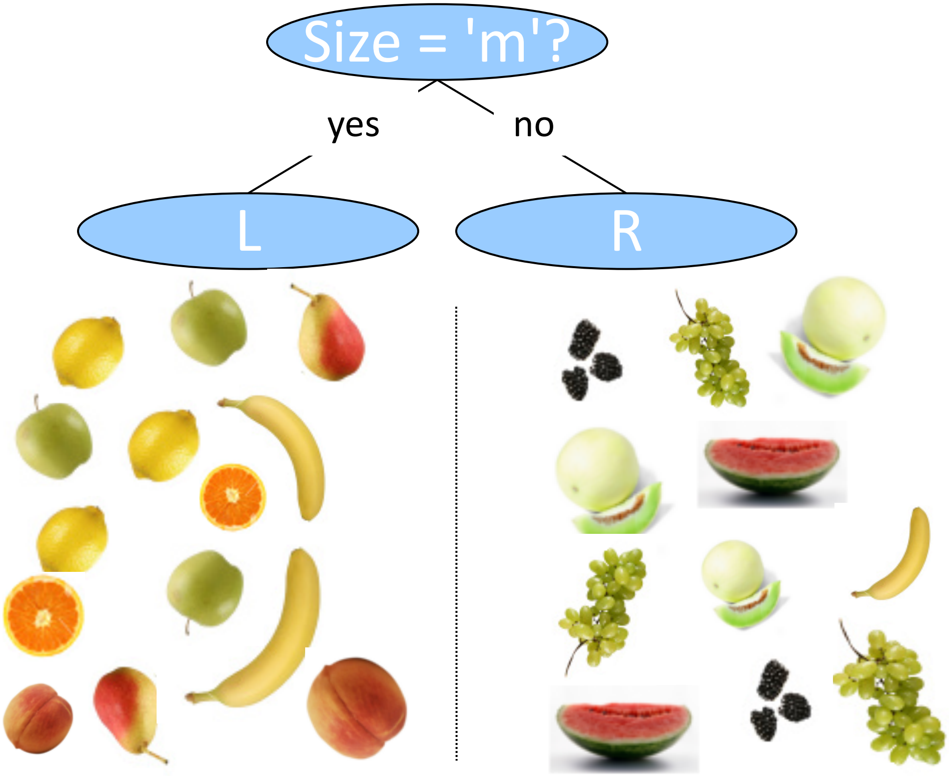
Choose a query that maximizes the impurity drop

Example

Suppose we create a tree with the following question:

"Put an object in L if (size = m), otherwise put it in R"

What is the impurity drop in this case?



$i_S = 22/25$

$i_L = 1 – 3/14 = 11/14$
$i_R = 1 – 3/11 = 8/11$

drop $= i_S – P_L * i_L – (1\text{-}P_L) * i_R$

drop = 22/25 – (14/25)*(11/14) – (1-14/25)*(8/11)

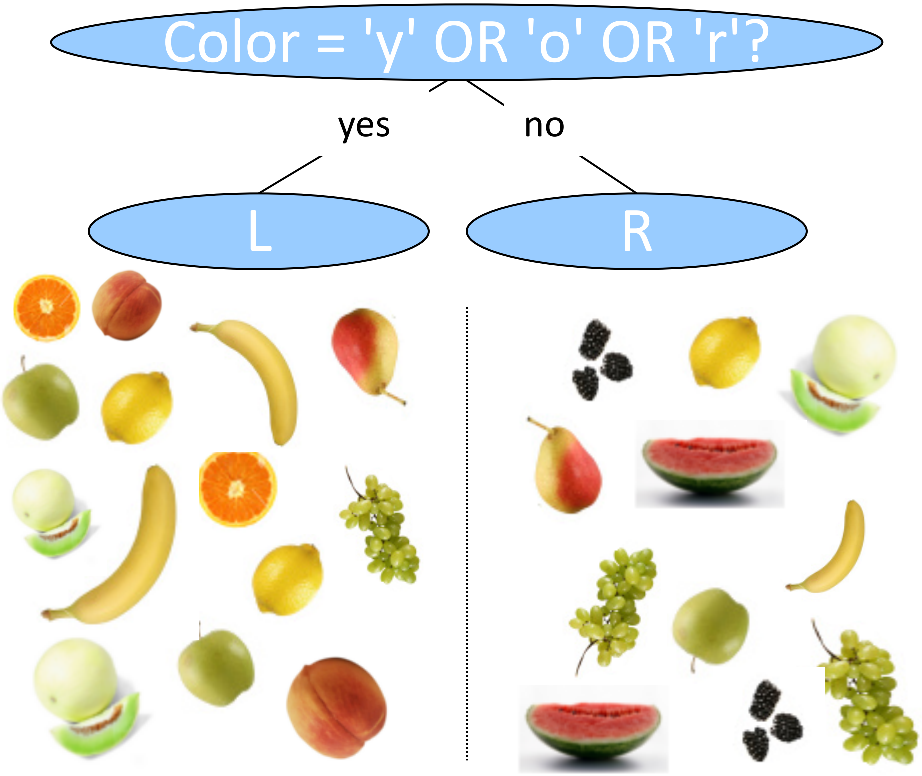
drop = 3/25

Example

Suppose we chose a different question:

“Put an object in L if (colour = y OR o OR r), otherwise put it in R”

What is the impurity drop in this case?



$i_S = 22/25$

$i_L = 1 – 2/14 = 12/14$

$i_R = 1 – 2/11 = 9/11$

$\text{drop} = i_S – P_L*i_L – (1\text{-}P_L)*i_R$

$\text{drop} = 22/25 – (14/25)*(12/14) – (1\text{-}14/25)*(9/11)$

$\text{drop} = 1/25$

Example

Fruit Classification

- Which question was better?

“Put an object in L if (size = m), otherwise put it in R”

(impurity drop: 3/25)

“Put an object in L if (color = y OR o OR r),
otherwise put it in R”

(impurity drop: 1/25)

Example

Fruit Classification

- Which question was better?

"Put an object in L if (size = m), otherwise put it in R"

impurity drop: 3/25

The higher the impurity drop, the better!

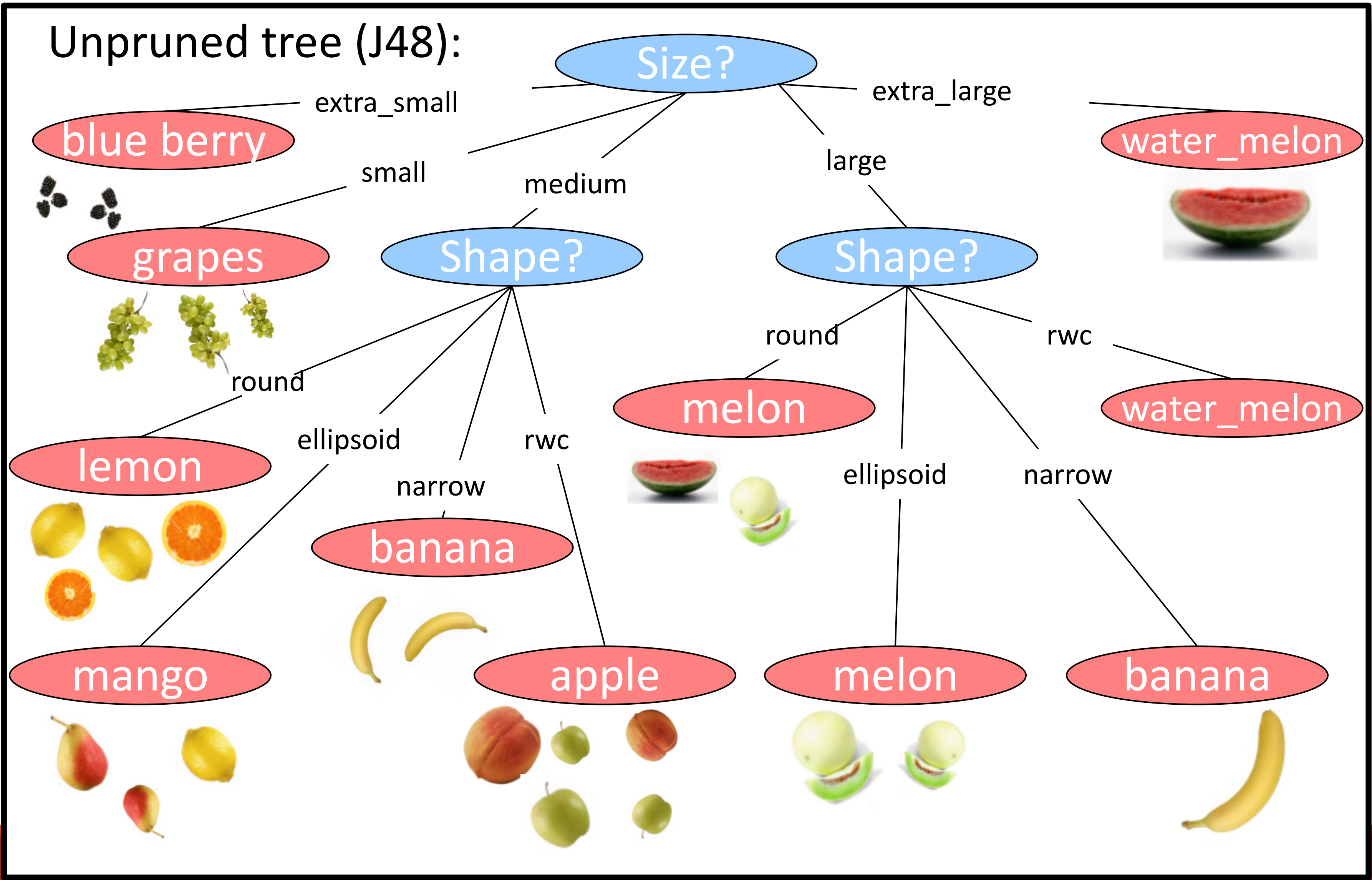
Pruning

Early splitting can cut off the possibility of beneficial splits in descendant nodes (limited horizon effect)

Alternative to splitting: the tree is first fully grown, then *pruning*:

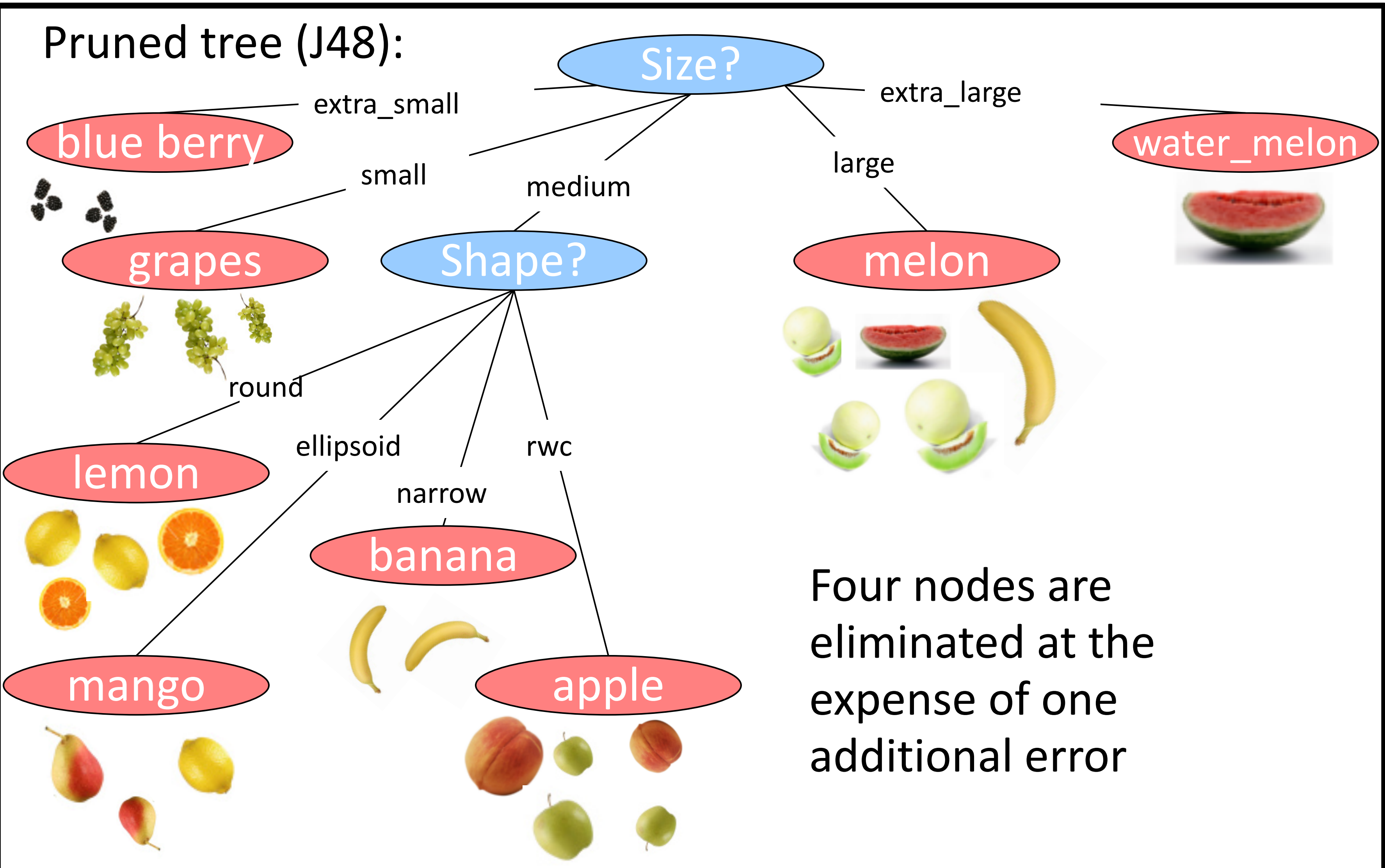
- Remove pairs of leaf nodes with a common parent or
- Replace a subtree by a leaf node.

Example


Unpruned tree (J48):
Size?
extra_small
blue berry
small
grapes
medium
Shape?
round
lemon
ellipsoid
mango
narrow
banana
rwc
apple
large
Shape?
round
melon
ellipsoid
melon
narrow
banana
rwc
water_melon
extra_large
water_melon

Example

Pruned tree (J48):



Four nodes are eliminated at the expense of one additional error