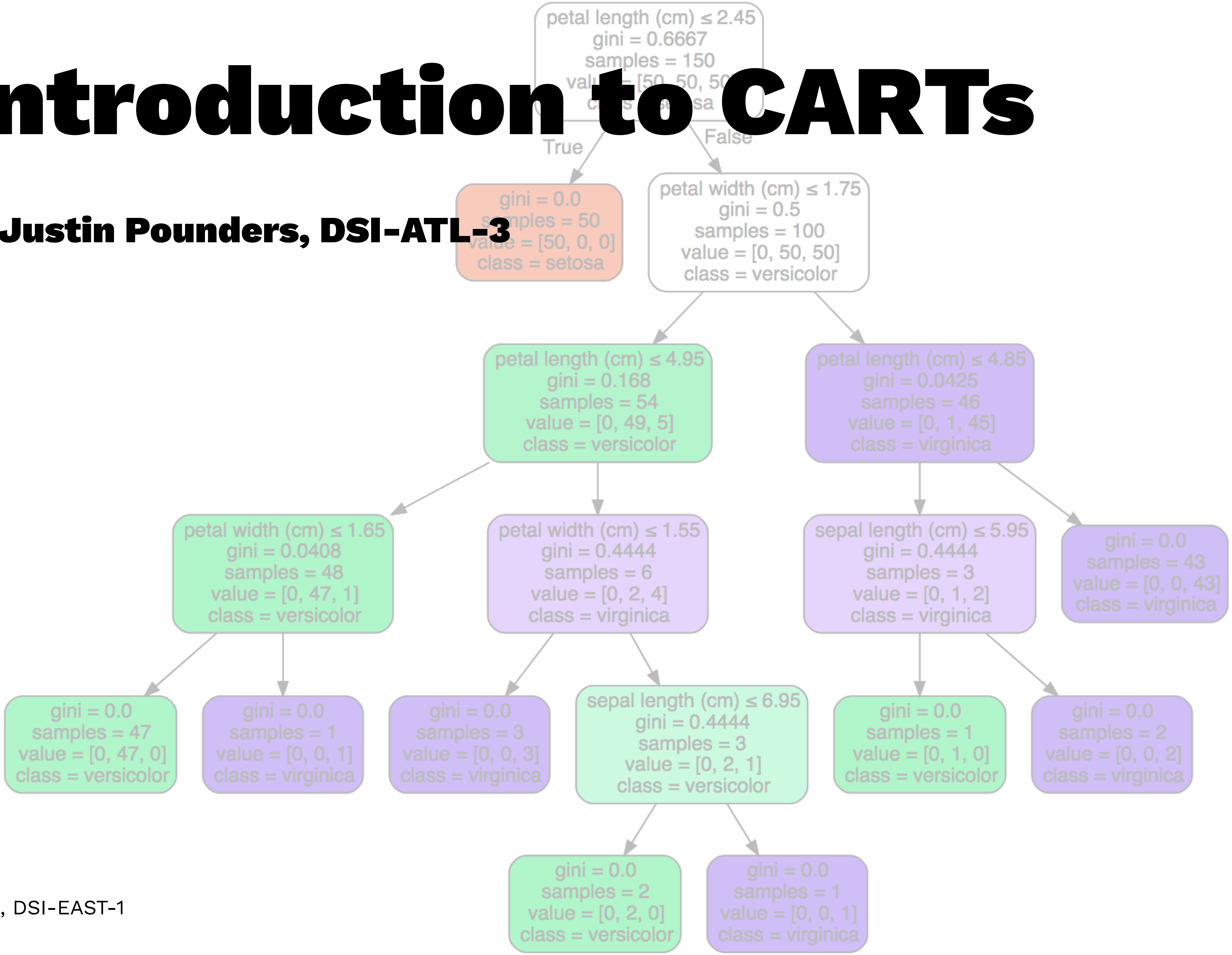
Introduction to CARTs



Author: Justin Pounders, DSI-ATL-3

Objectives

— Describe how decision tree models work

— Define the concept of purity and information gain

— Describe how model ensembles lead to better performance

— Define boosting and model aggregation (aka bagging)

— Build decision tree and random forest models in `sklearn`

— Add ensemble bagging to any classification model in `sklearn`

Guess what animal I am thinking of.

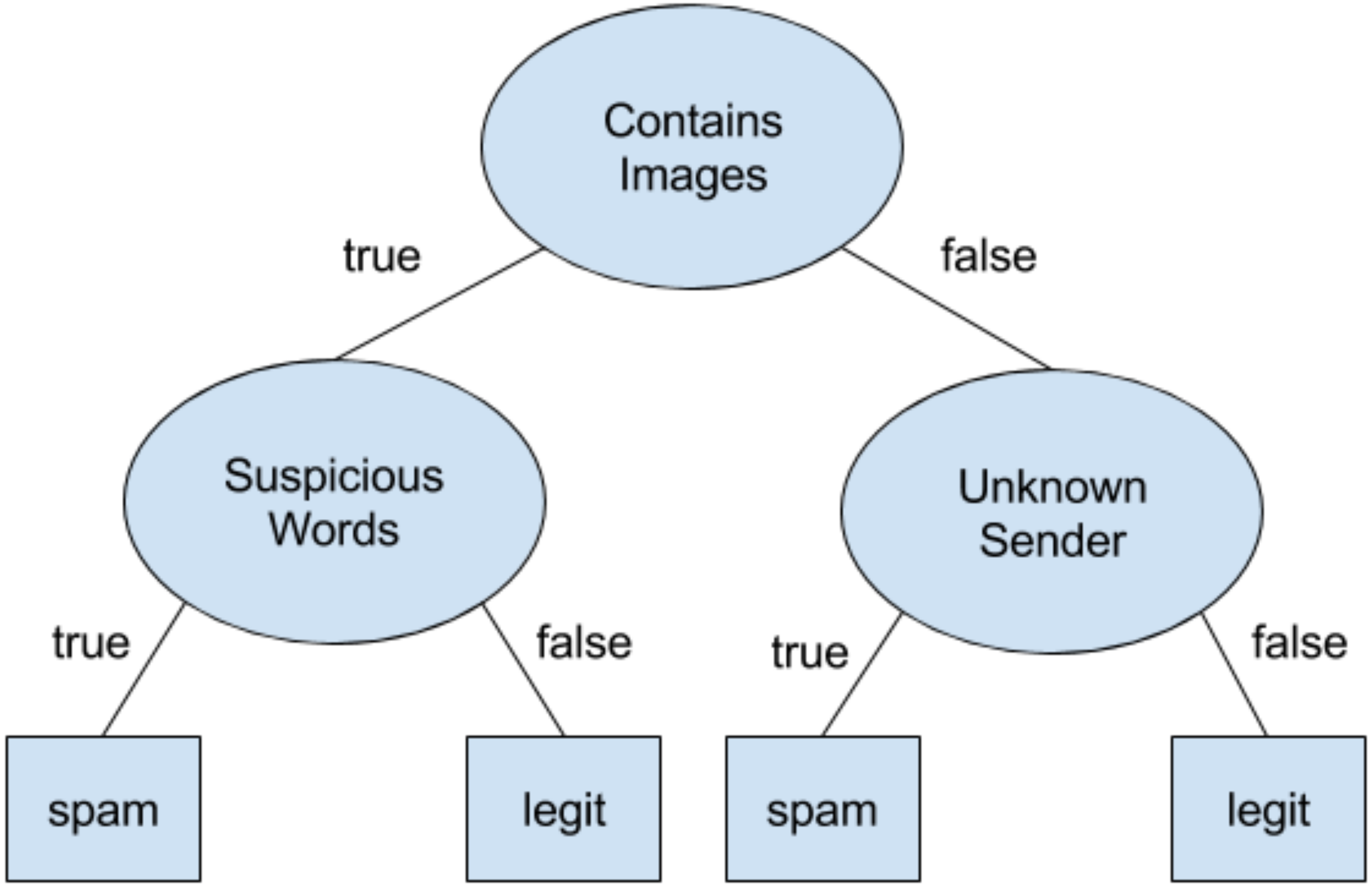
Intro to CARTs

CART = Classification and Regression Tree

CART = Decision Tree

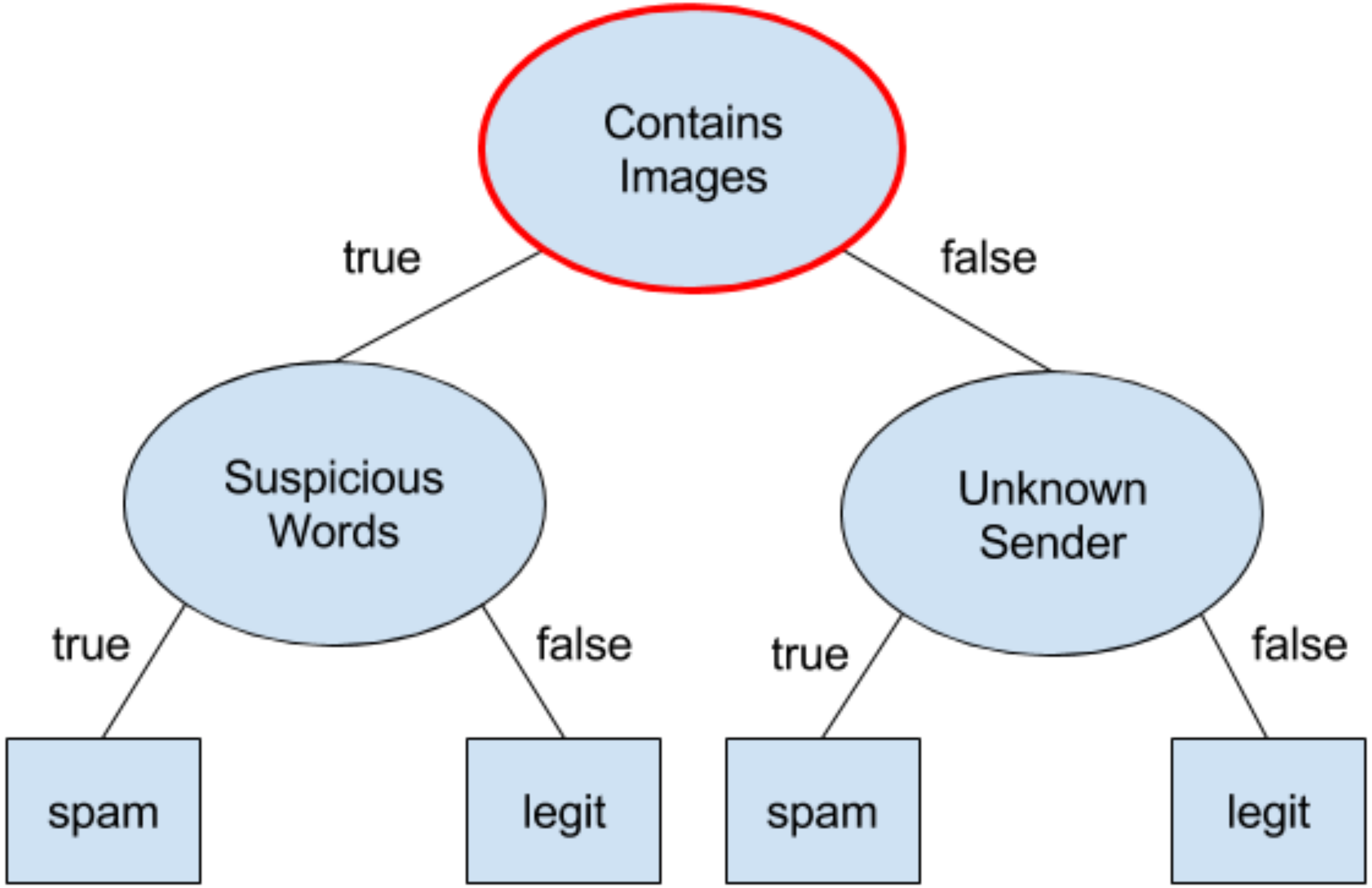
Example Decision Tree for a Spam Filter

ID	*Susp. words*	*Unknown sender*	*Contains images*	*Classifi-cation*
1	true	false	true	???



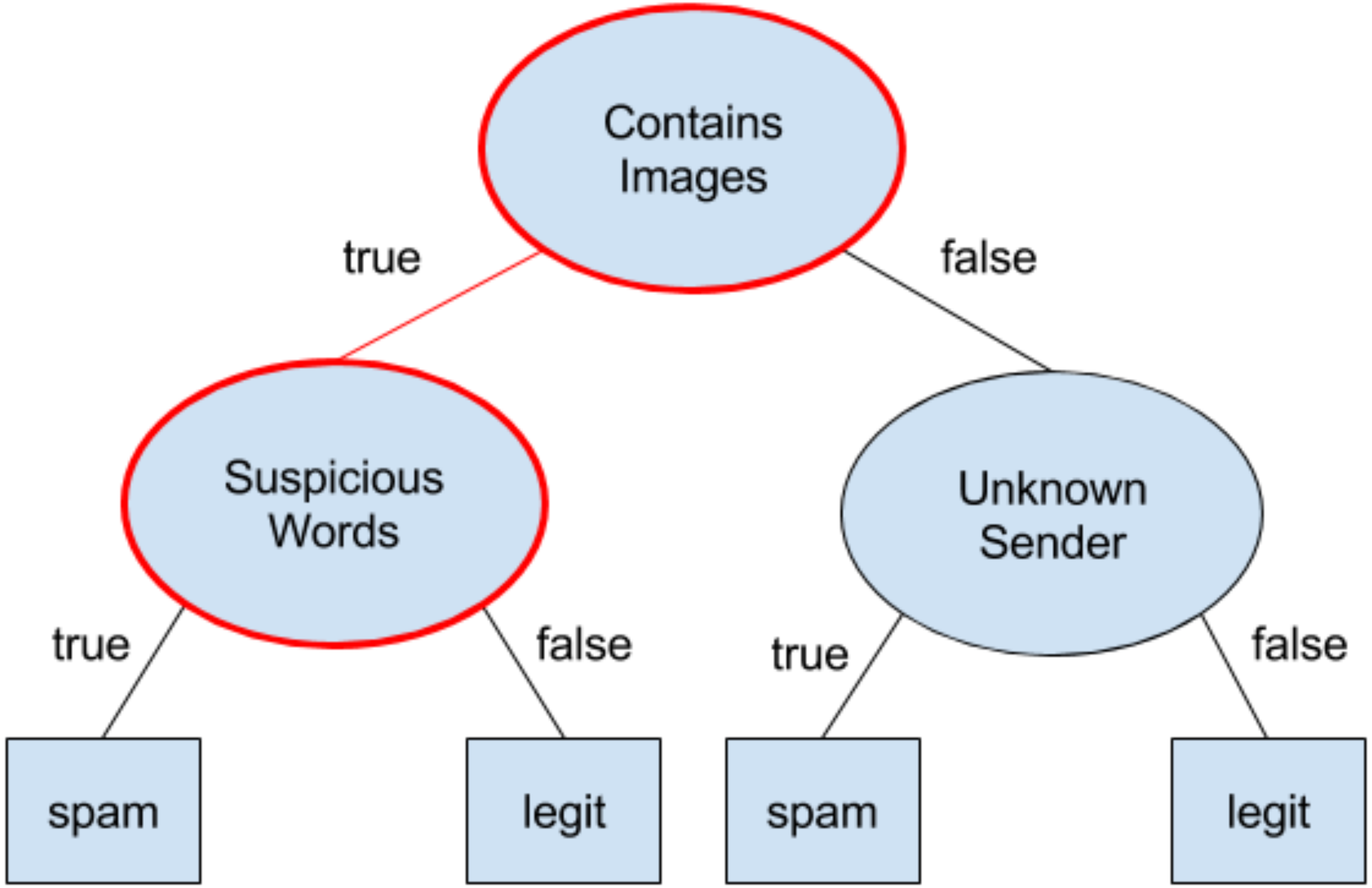
ID	*Susp. words*	*Unknown sender*	*Contains images*	*Classifi-cation*
1	true	false	true	???

Example Decision Tree for a Spam Filter



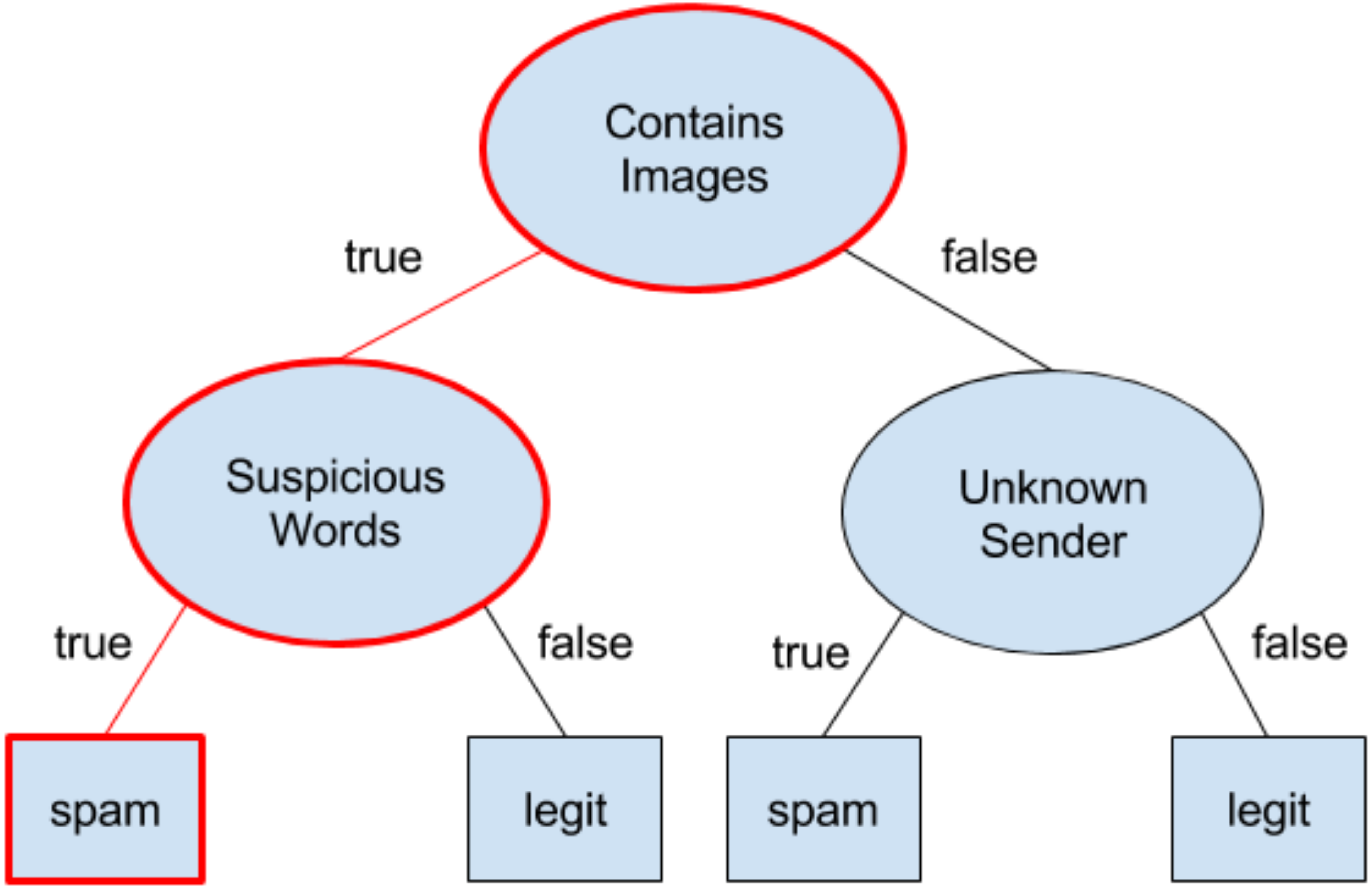
ID	*Susp. words*	*Unknown sender*	*Contains images*	*Classifi-cation*
1	true	false	true	???

Example Decision Tree for a Spam Filter



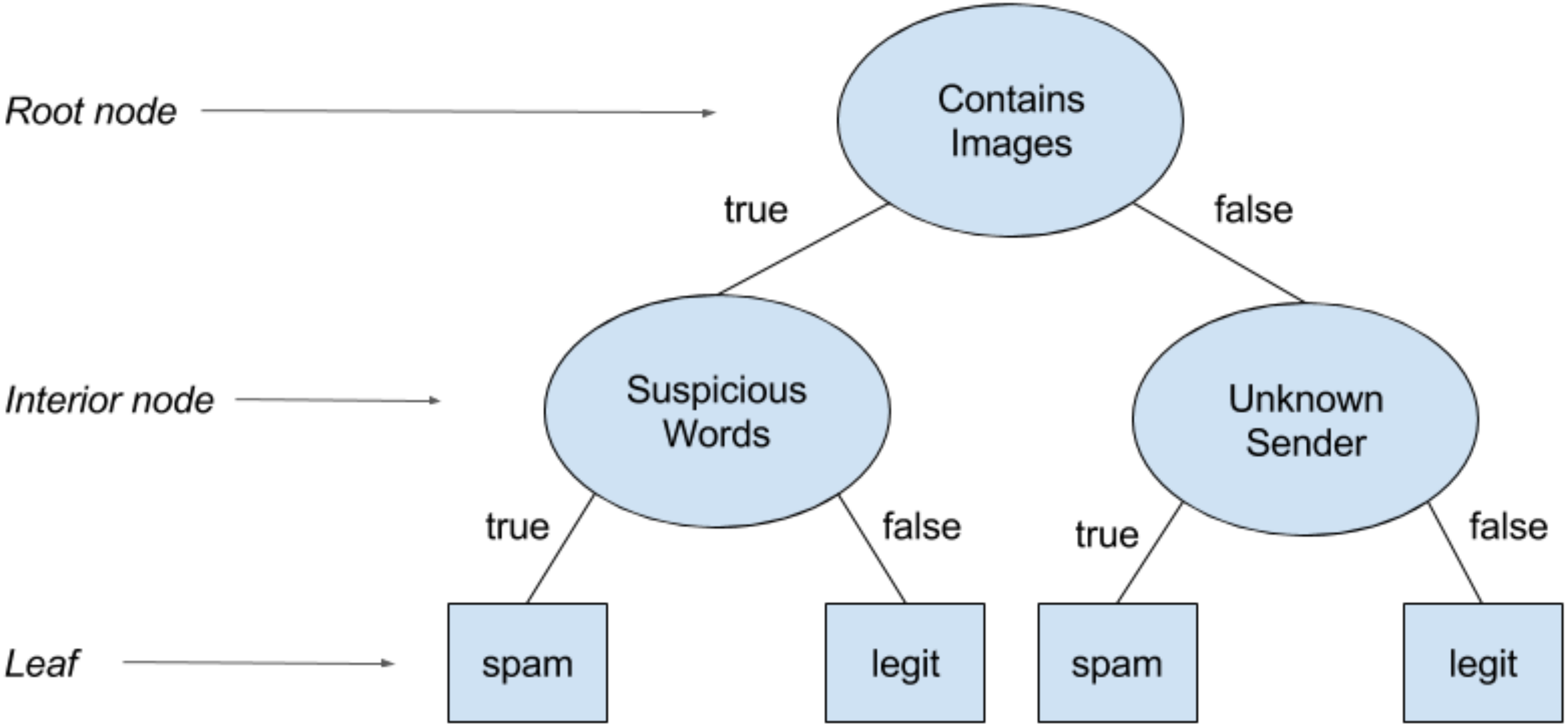
Example Decision Tree for a Spam Filter

ID	*Susp. words*	*Unknown sender*	*Contains images*	*Classifi-cation*
1	true	false	true	spam



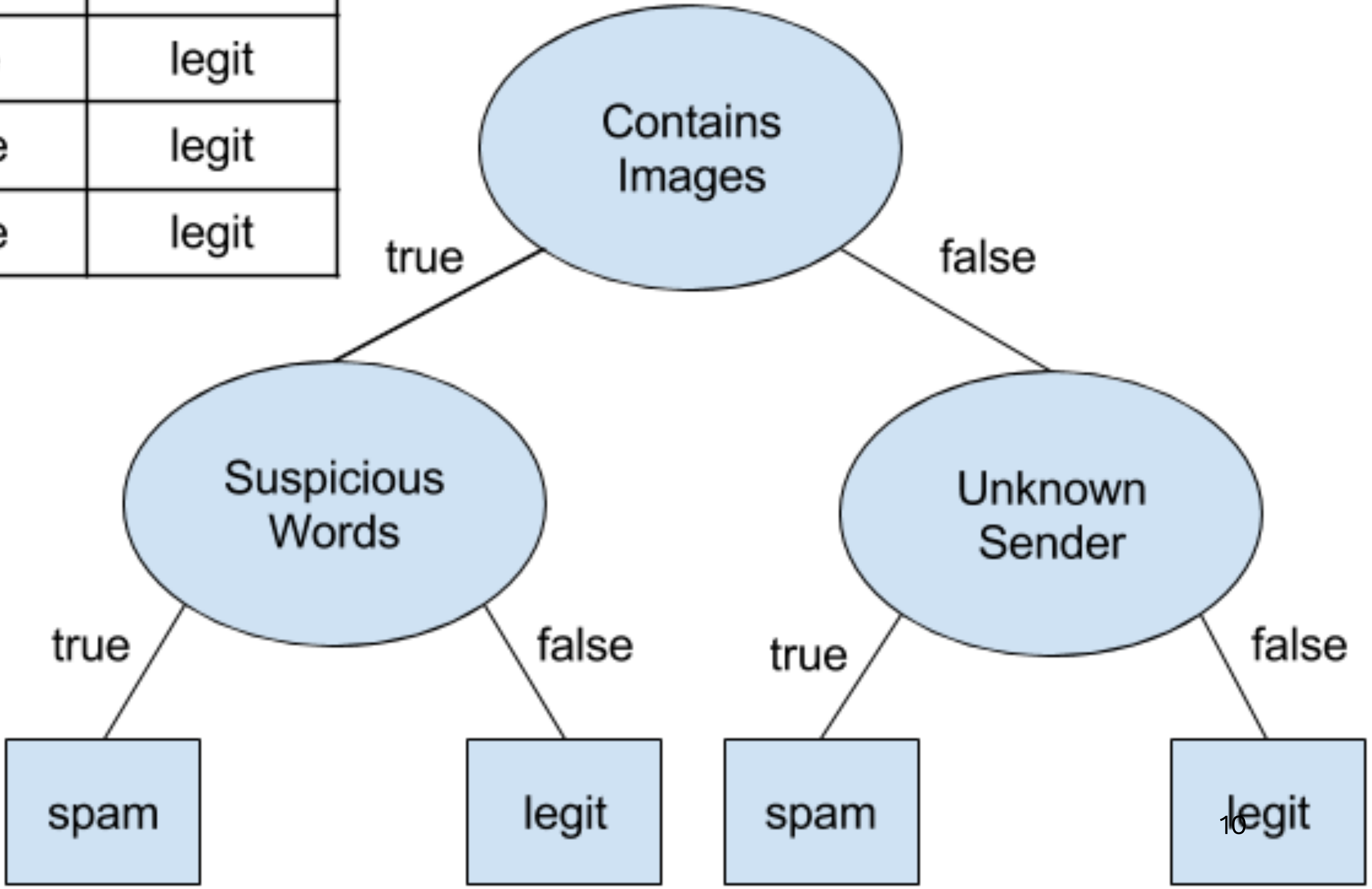
Example Decision Tree for a Spam Filter

ID	*Susp. words*	*Unknown sender*	*Contains images*	*Classification*
1	true	false	true	spam



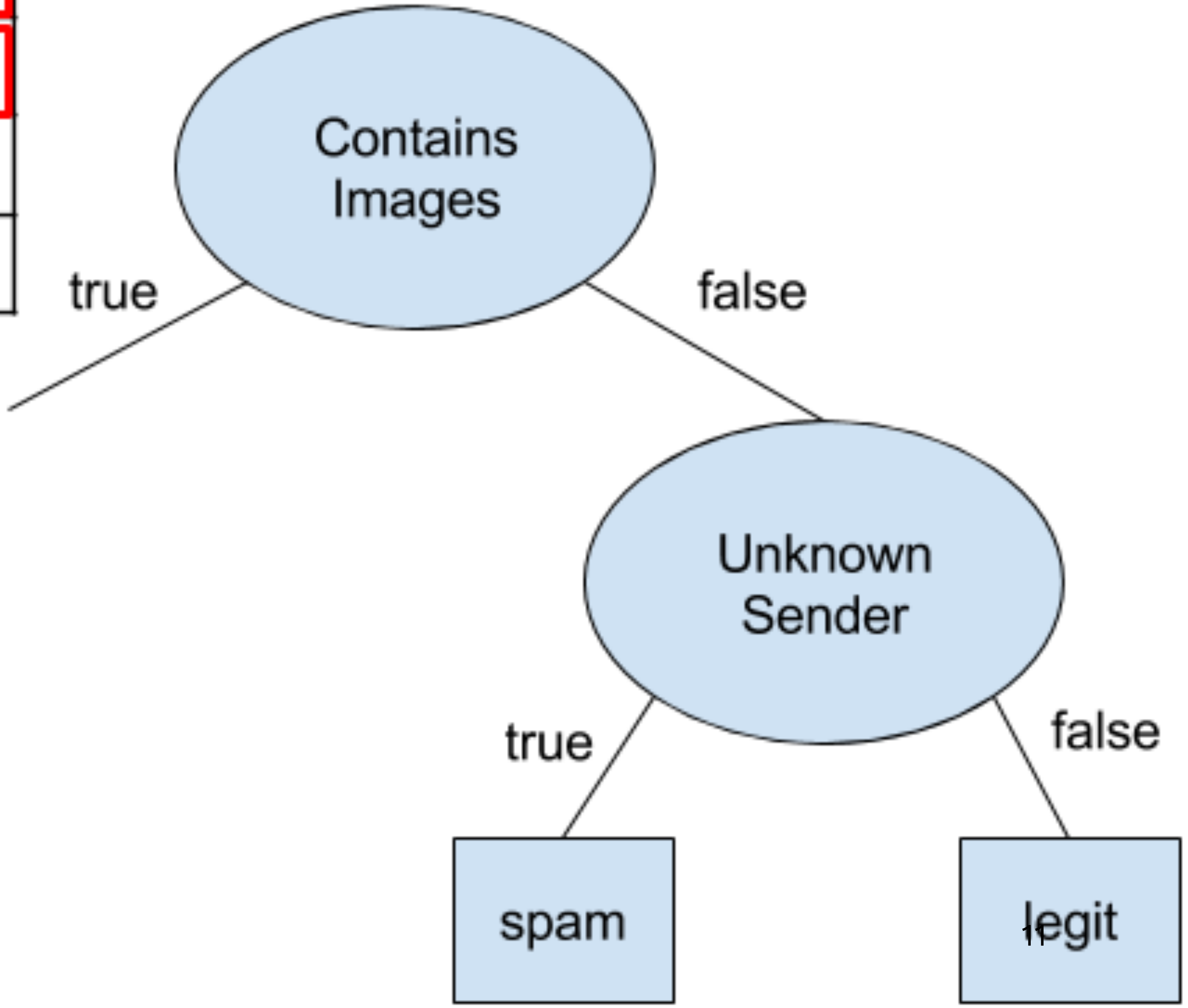
ID	*Susp. words*	*Unknown sender*	*Contains images*	*Classifi-cation*
1	true	false	true	spam
2	true	true	false	spam
3	true	true	true	spam
4	false	true	true	legit
5	false	false	false	legit
6	false	false	false	legit

How do you build a decision tree?



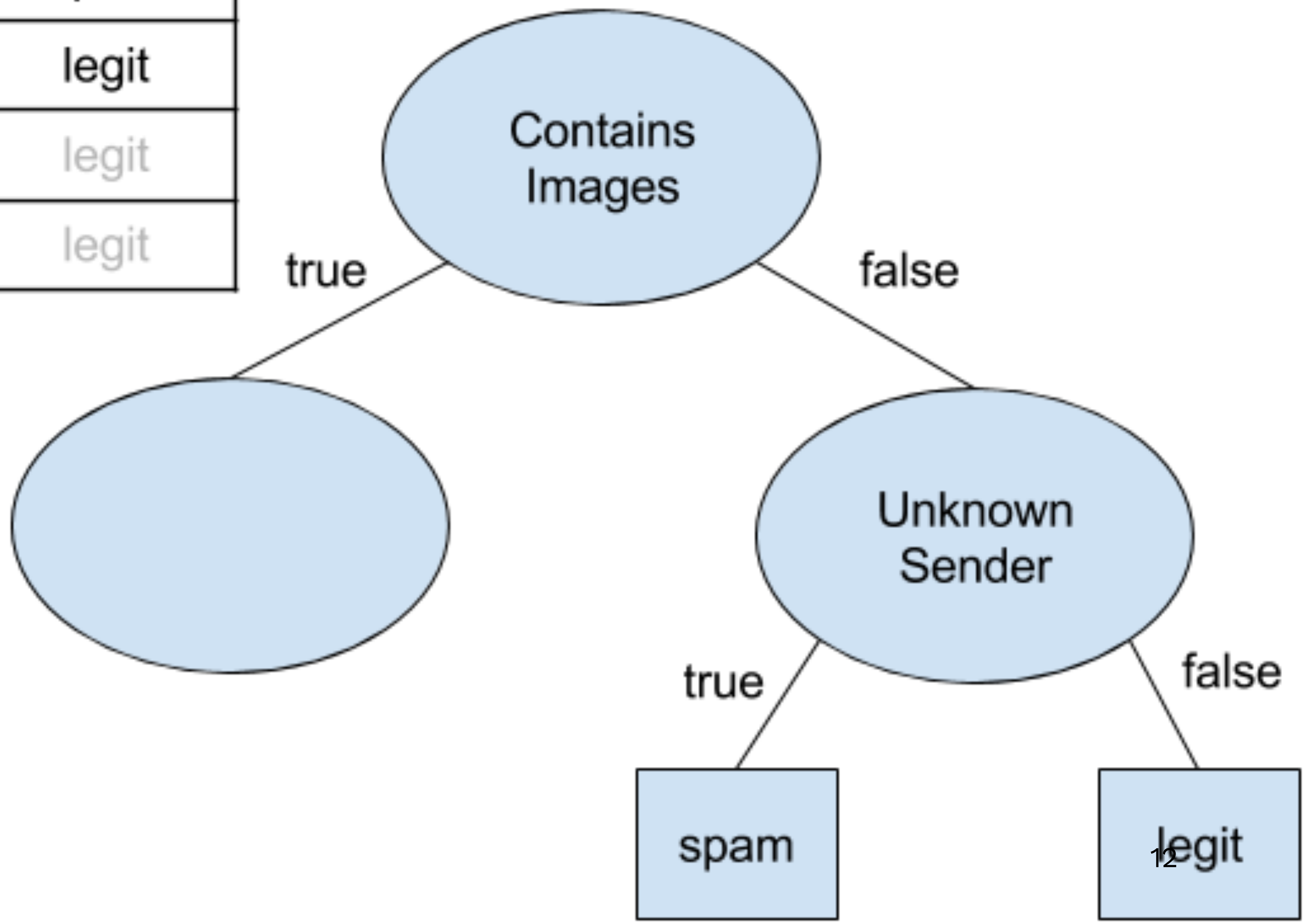
Each new node generates a subset of the data.

ID	Susp. words	Unknown sender	Contains images	Classifi-cation
1	true	false	true	spam
2	true	true	false	spam
3	true	true	true	spam
4	false	true	true	legit
5	false	false	false	legit
6	false	false	false	legit



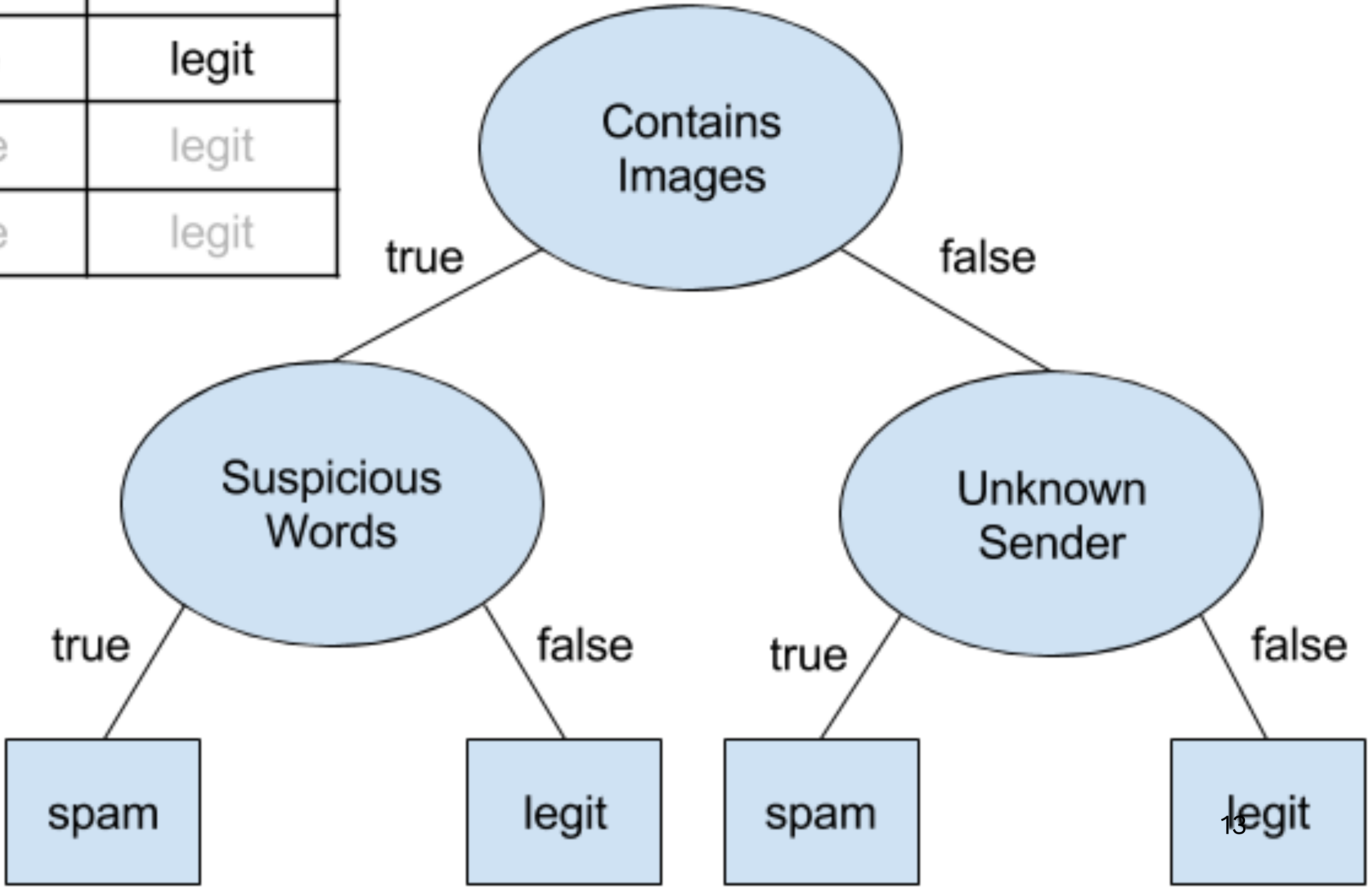
ID	Susp. words	Unknown sender	Contains images	Classifi-cation
1	true	false	true	spam
2	true	true	false	spam
3	true	true	true	spam
4	false	true	true	legit
5	false	false	false	legit
6	false	false	false	legit

Which remaining feature best discriminates the data?



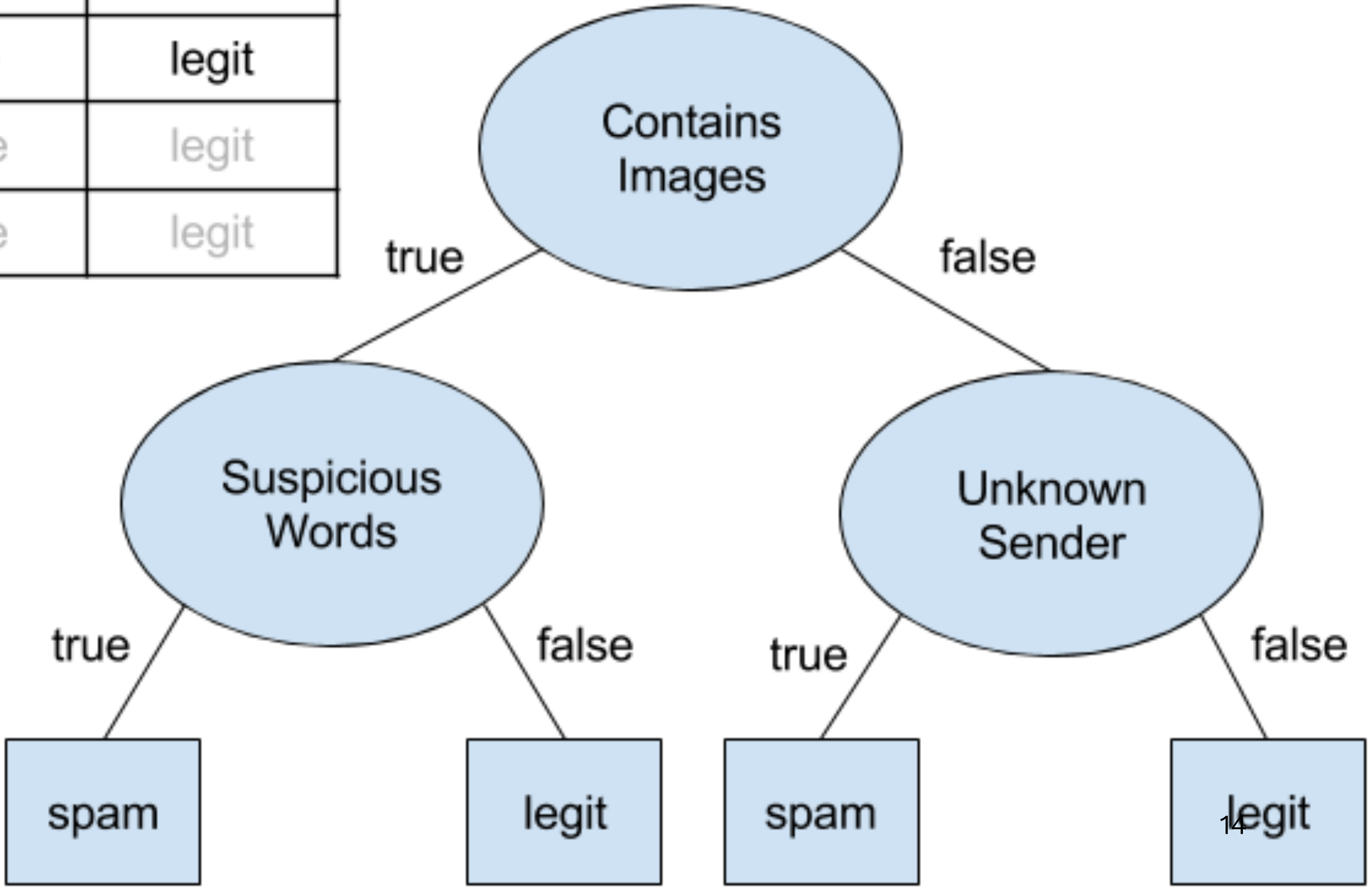
ID	Susp. words	Unknown sender	Contains images	Classifi-cation
1	true	false	true	spam
2	true	true	false	spam
3	true	true	true	spam
4	false	true	true	legit
5	false	false	false	legit
6	false	false	false	legit

We can stop here!



ID	Susp. words	Unknown sender	Contains images	Classifi-cation
1	true	false	true	spam
2	true	true	false	spam
3	true	true	true	spam
4	false	true	true	legit
5	false	false	false	legit
6	false	false	false	legit

We can stop here!



What would happen if we picked a different feature for the root node?

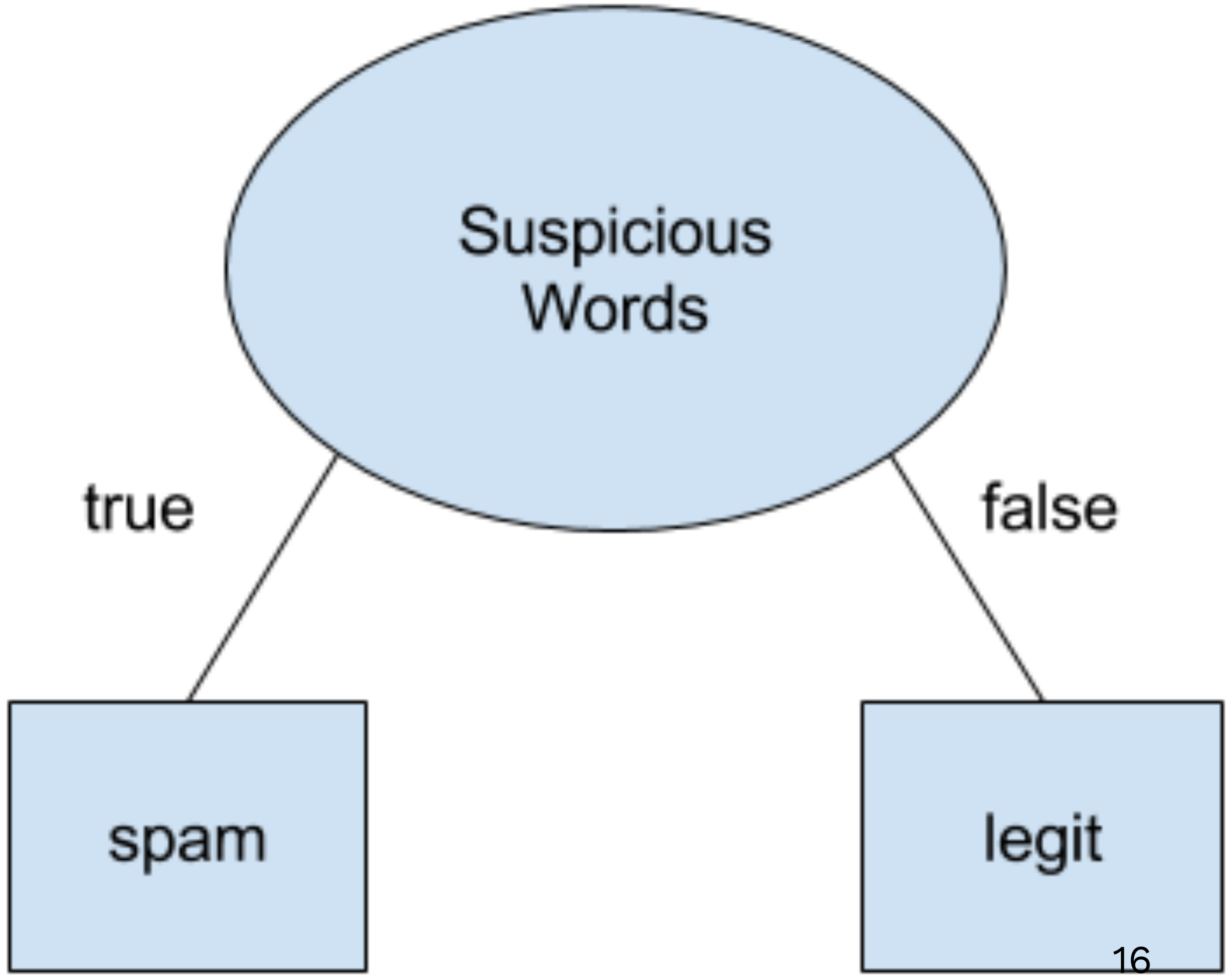
Practice

Which feature *best* discriminates the data?

Draw the decision tree starting with this feature as the root.

ID	*Susp. words*	*Unknown sender*	*Contains images*	*Classifi-cation*
1	true	false	true	spam
2	true	true	false	spam
3	true	true	true	spam
4	false	true	true	legit
5	false	false	false	legit
6	false	false	false	legit

ID	*Susp. words*	*Unknown sender*	*Contains images*	*Classifi-cation*
1	true	false	true	spam
2	true	true	false	spam
3	true	true	true	spam
4	false	true	true	legit
5	false	false	false	legit
6	false	false	false	legit



Building Decision Trees

— You've seen that decision trees split the data at each node

— For each split we can calculate purity

Building Decision Trees

$$\begin{aligned}\text{purity of class i} &= p(\text{class i}|\text{data at node}) \\ &= p(i \mid D)\end{aligned}$$

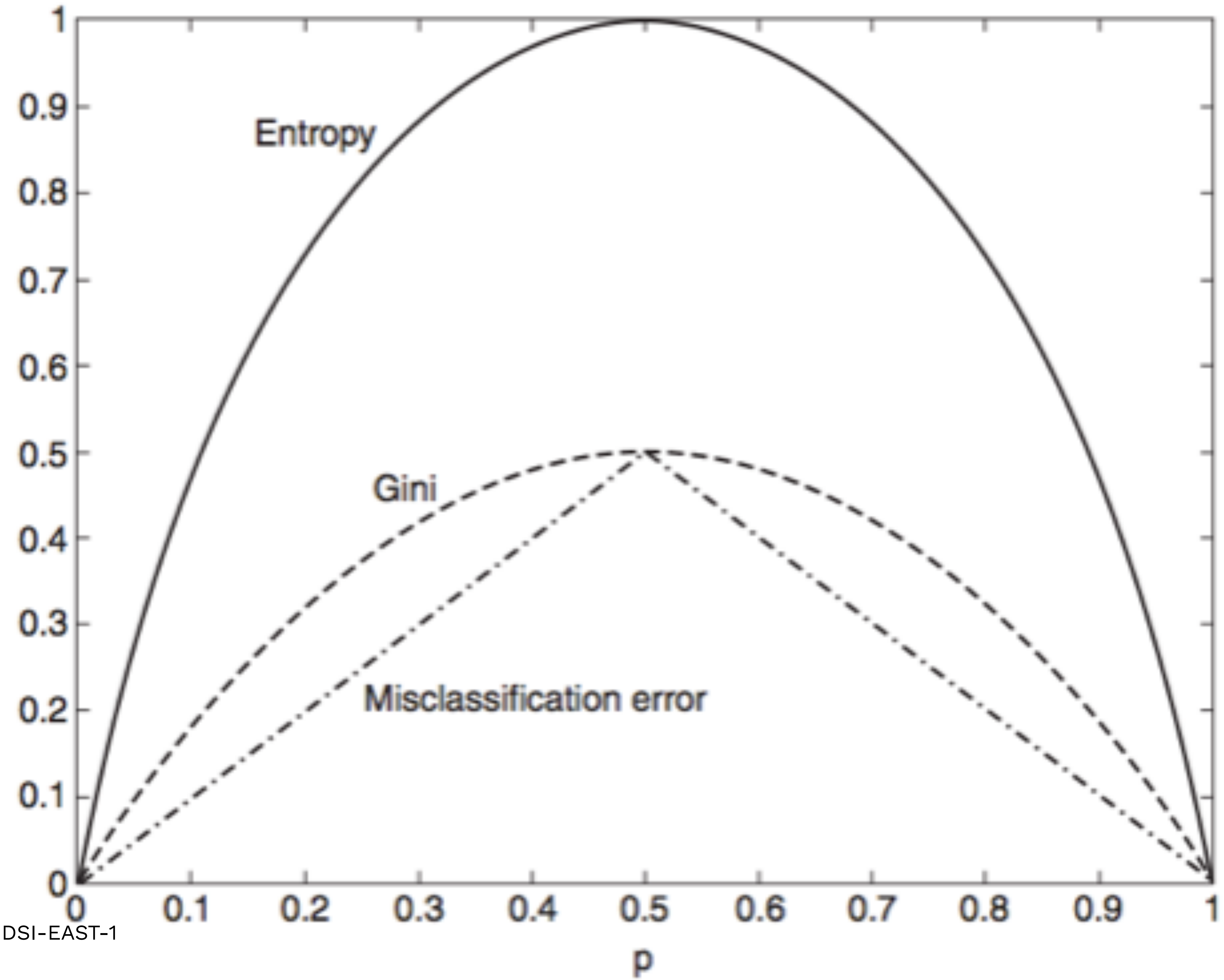
- For binary classification:
  - Worst case, purity = 0.5
  - Best case, purity = 1.0

Generalizing Purity

Impurity measures for a node with data D

$$\text{Entropy} = -\sum_{i=1}^{classes} p(i \mid D)\; log_2\, p(i \mid D)$$

$$\text{Gini} = \sum_{i=1}^{classes} p(i \mid D)(1 - p(i \mid D)) = 1 - \sum_{i=1}^{classes} p(i \mid D)^2$$


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Entropy
Gini
Misclassification error
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
p

Information Gain

Goal: determine how good a split is

Solution: gain

$$\text{gain} = I(\text{parent}) - \sum_{\text{children}} \frac{N_j}{N} I(\text{child}_j)$$

where I is the impurity measure, N_j is the number of observations at child node j, and N is the number of observations at the parent node.

Full CARTs

- So far we've seen trees for classification
- Decision trees can be used for regression too

Decision Trees

Notes and Observations

- Decision trees are hierachical
  - Sequence of "if-this-than-that" conditions
- Decision trees are non-parametric
  - No β coefficients!
  - No assumption on distributions

Decision Trees

Notes and Observations

CART advantages

- Simple to understand and interpret.
- Requires little data preparation.
- Able to handle both numerical and categorical data.
- Possible to validate a model using statistical tests.
- Once trained can be implemented on hardware and has extremely fast (real-time) execution.

Decision Trees

Notes and Observations

CART disadvantages

— Locally-optimal.

— Overfitting.

— There are concepts that are hard to learn (XOR, parity or multiplexer problems)

— Decision trees can be biased if some classes dominate.