Data Science Basics and Decision Tree Models

What Really *is* Data Mining (or Data Science) ?

- A process for using information technology to extract useful (non-trivial, hopefully actionable) knowledge from large bodies of data

Data Science Tasks

Many problems have one of these data science tasks:

- Classification
  - Predict for each individual in a population which class this individual belongs to.
- Regression
  - “value estimation”, estimate the numerical value of some variable specific to an individual.
- Similarity matching
  - Identify similar individuals based on data known about them.
- Clustering
  - Group individuals by their similarity.

Data Science Tasks

- Co-occurence grouping
  - Find associations between entities based on transactions involving them. The result is a description of items that occur together.

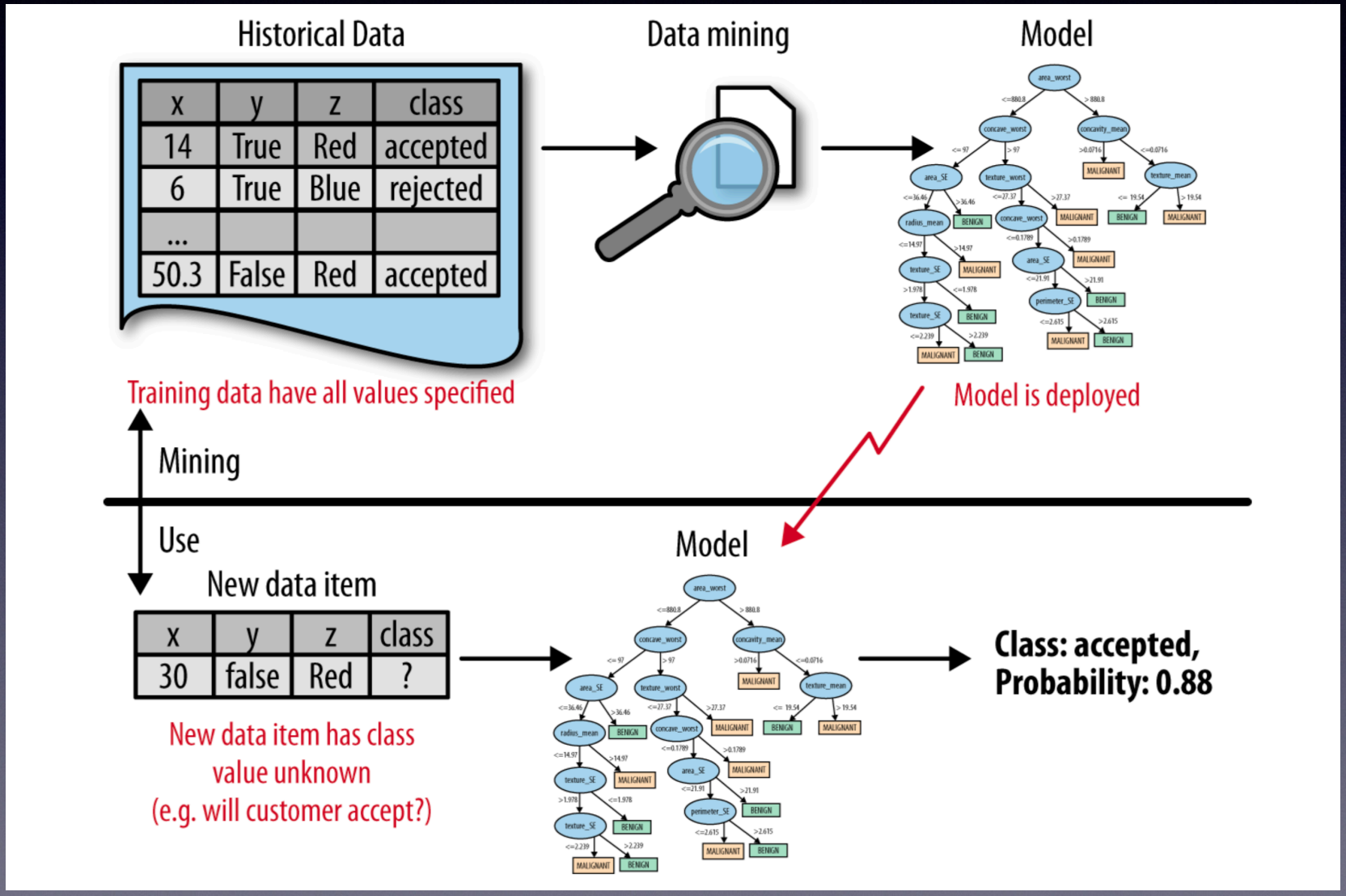
- Profiling
  - “behavior description”, characterize the typical behavior of an individual, group, or population.

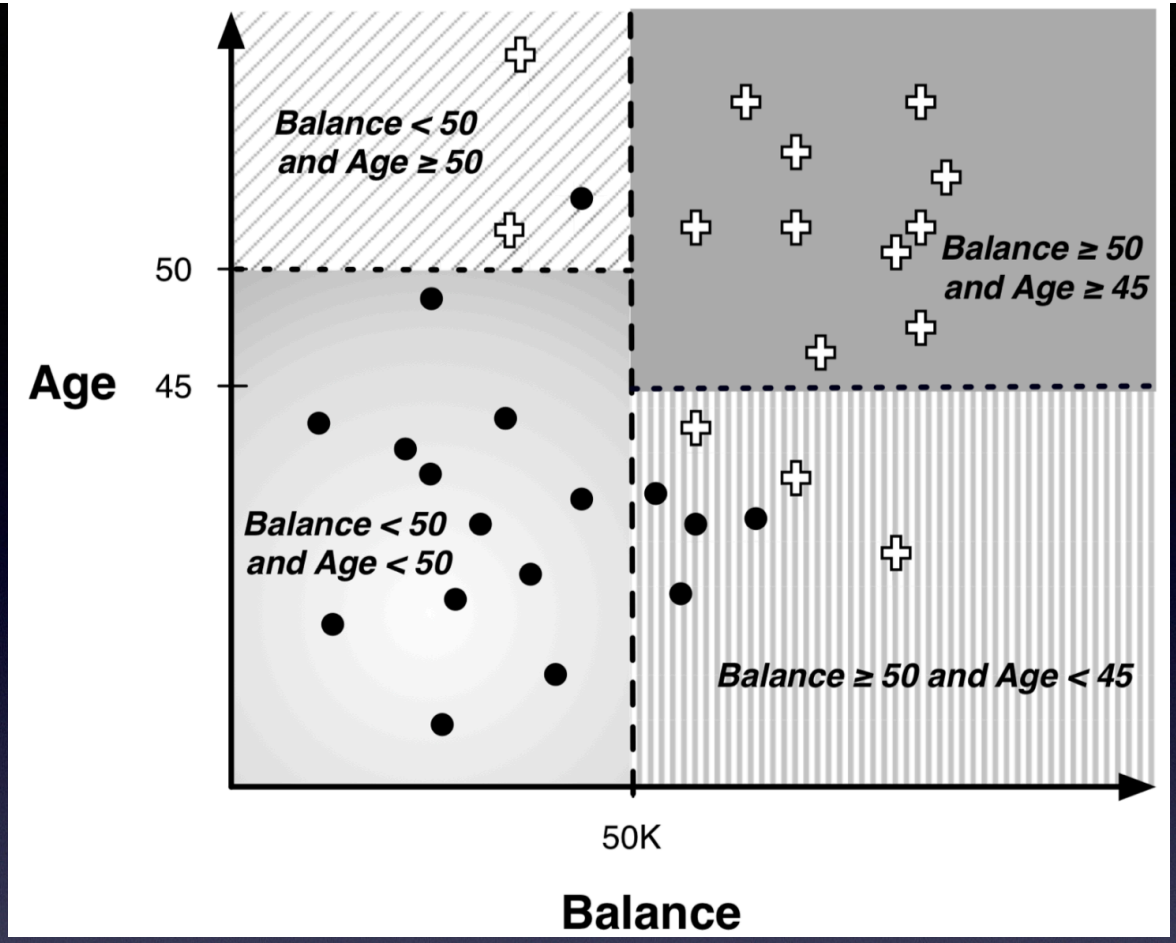
- Link prediction
  - Predict connections between data items, usually. By suggesting that a link should exist, and possibly also estimating the strength of the link.

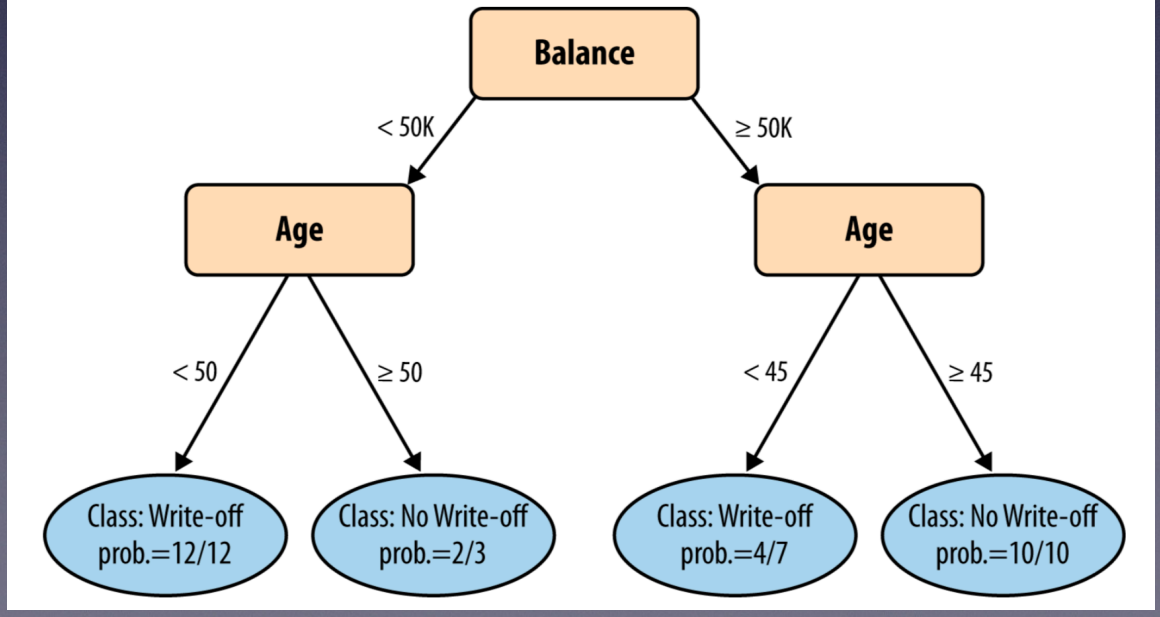
Supervised vs. Unsupervised Learning

- Supervised methods
  - Classification, regression
- Unsupervised methods
  - Clustering, profiling, co-occurrence grouping
- Technically, there must be data on the target for supervised learning methods.
  - Classification and regression are distinguished by the type of target. Regression involve a numeric target while classification involves a categorical target.

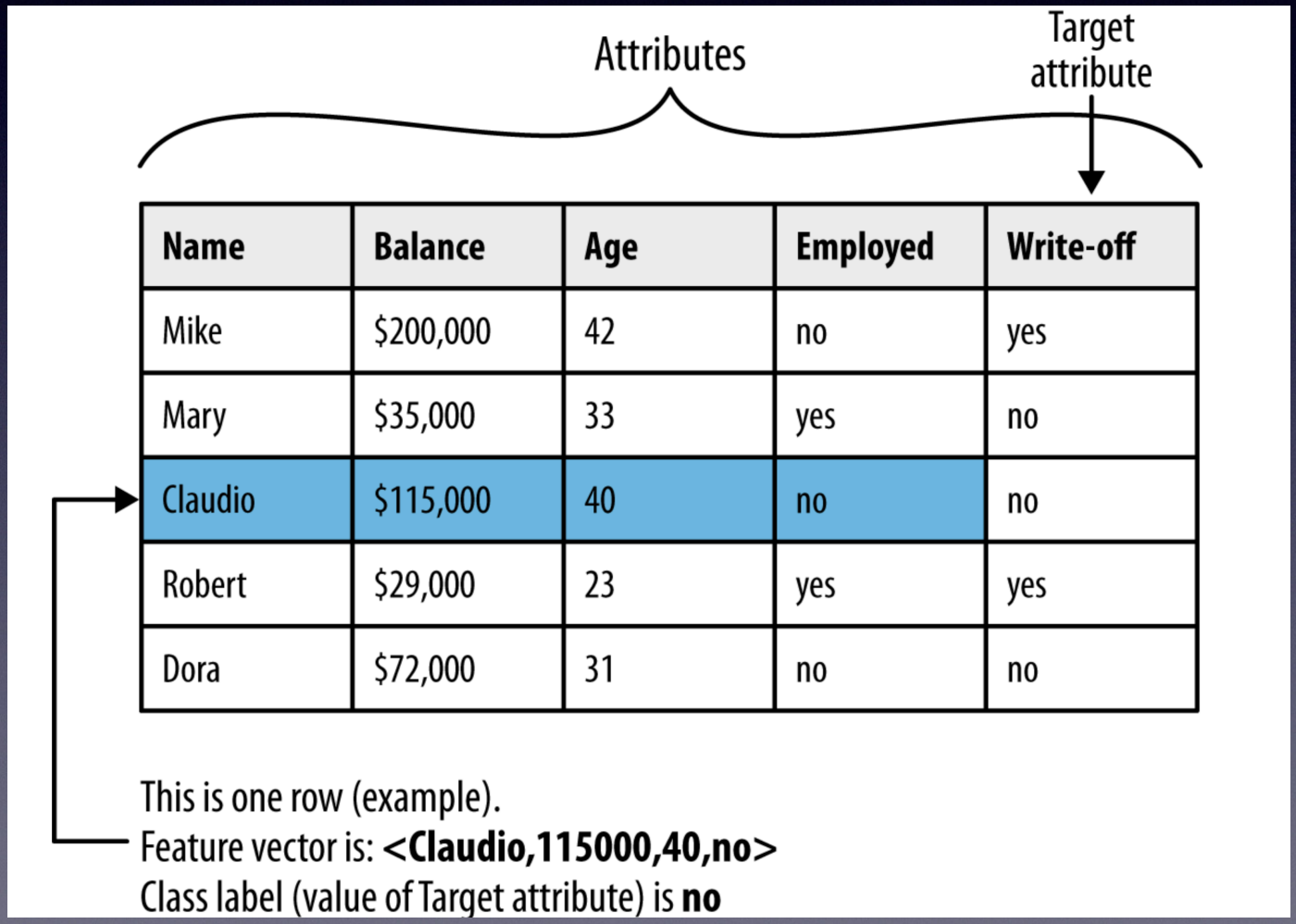
Classification Example




Balance < 50
and Age ≥ 50
Balance ≥ 50
and Age ≥ 45
Balance < 50
and Age < 50
Balance ≥ 50 and Age < 45
50
45
Age
50K
Balance


Balance
< 50K
≥ 50K
Age
Age
< 50
≥ 50
< 45
≥ 45
Class: Write-off
prob.=12/12
Class: No Write-off
prob.=2/3
Class: Write-off
prob.=4/7
Class: No Write-off
prob.=10/10

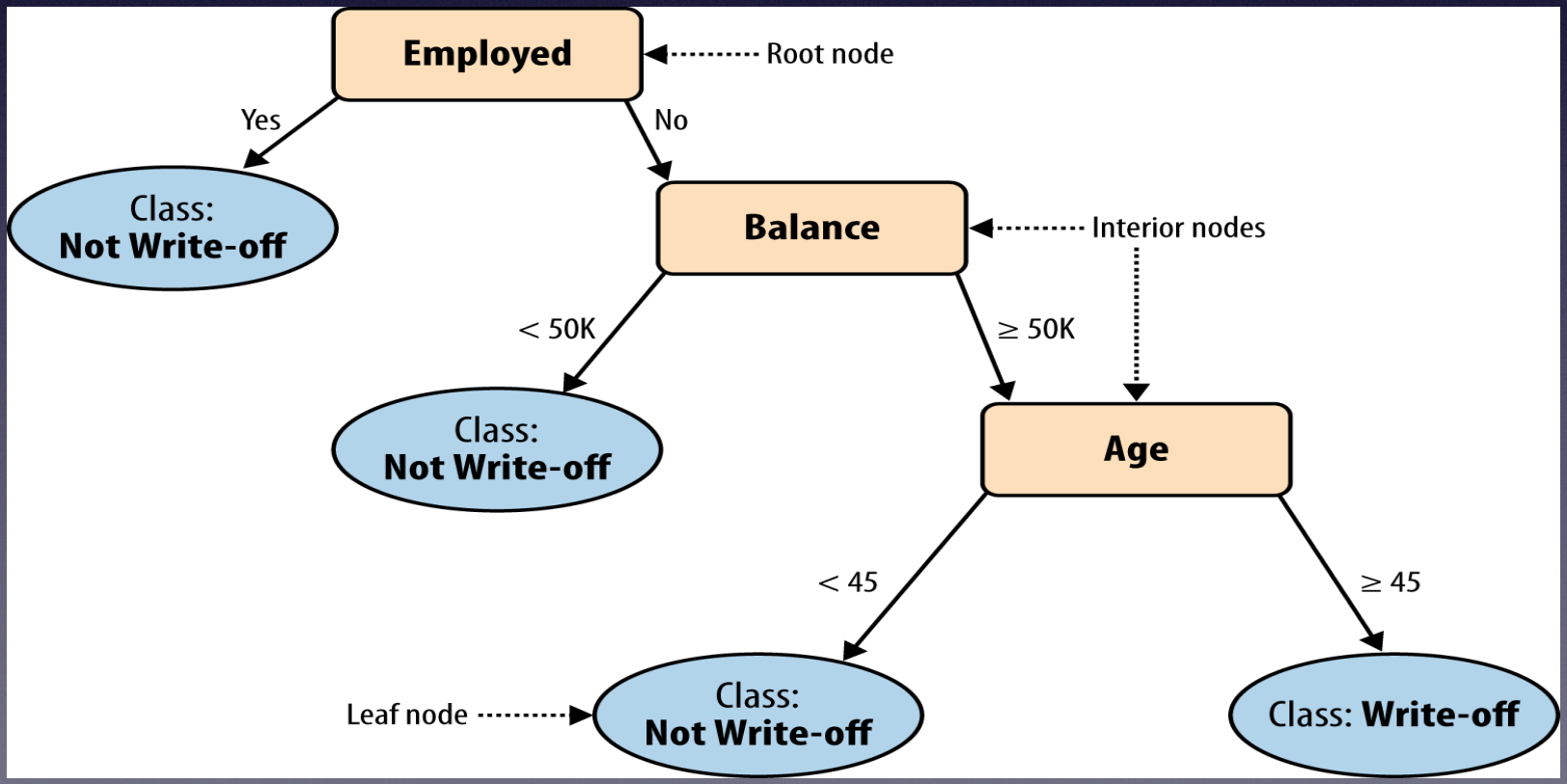
Example dataset for a Supervised Classification Problem



Name	Balance	Age	Employed	Write-off
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no
Claudio	$115,000	40	no	no
Robert	$29,000	23	yes	yes
Dora	$72,000	31	no	no

Tree-Structured Models

- Classify Claudio
  - Balance=115K, Employed=No, and Age=40



Tree-Structured Models: “Rules”

- No two parents share descendants
- There are no cycles
- The branches always “point downwards”
- Every example always ends up at a leaf node with some specific class determination

Tree Induction

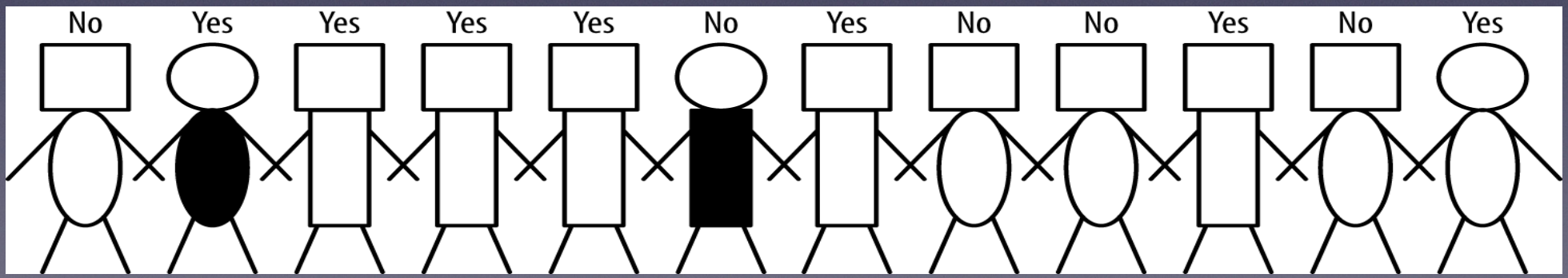
- How do we create a classification tree from data?
  - **Divide-and-conquer** approach
  - Take each data subset and *recursively* apply attribute selection to find the best attribute to partition it (create the “purest” subgroups possible using the attributes)
- When do we stop?
  - The nodes are pure
  - There are no more variables (don’t use the same variable more than once), or even earlier (over-fitting)

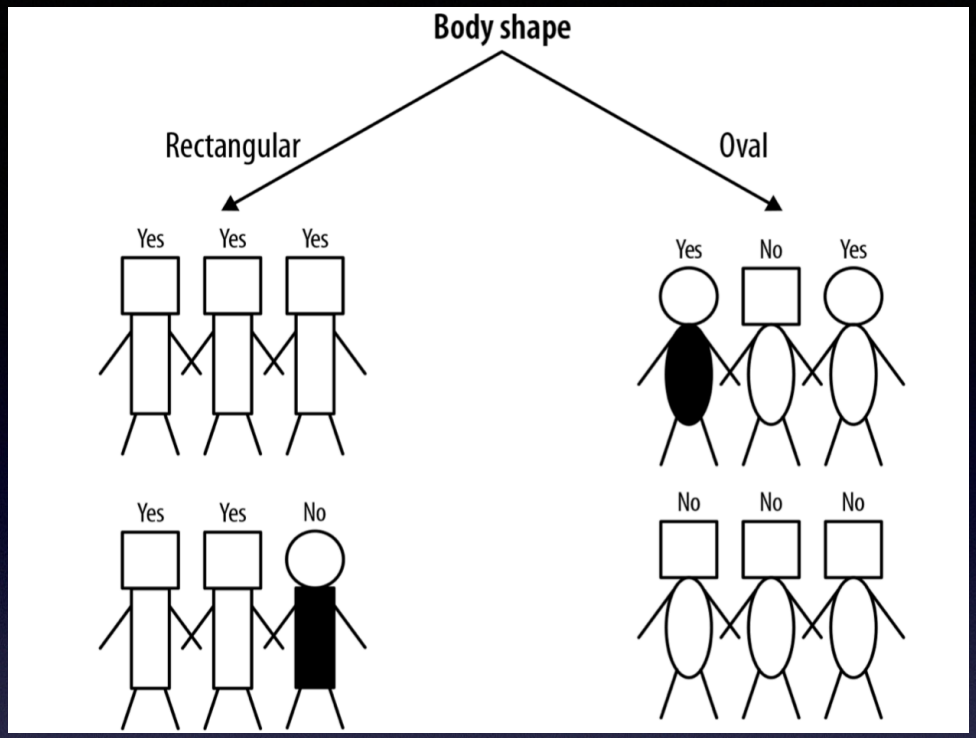
What are the Questions?

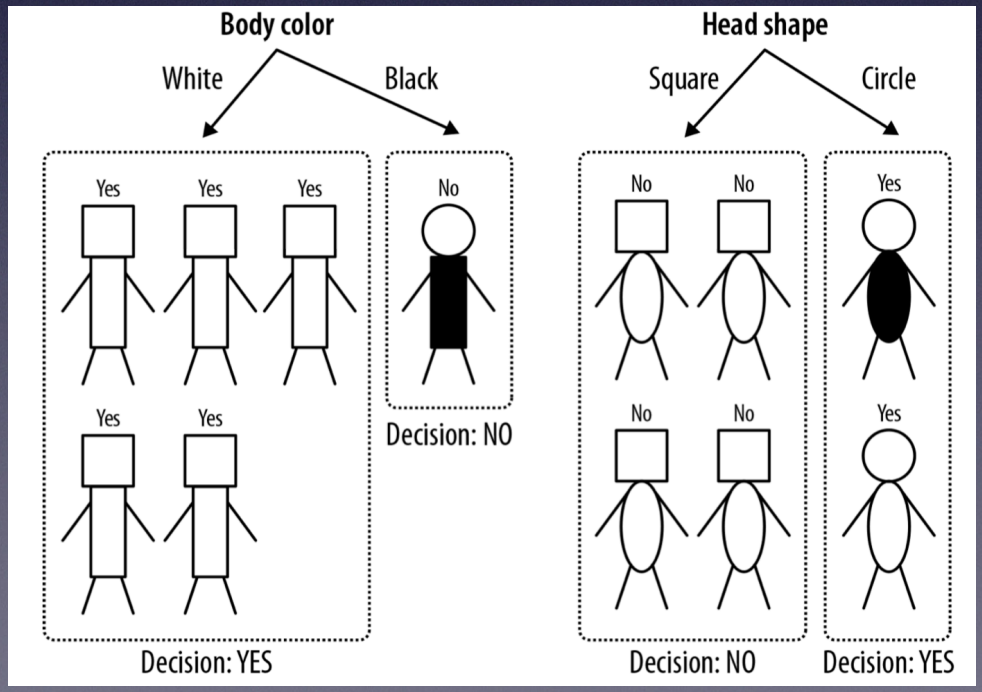
- How can we segment the population into groups that differ from each with respect to some quantity of interest?
- How can we judge whether a variable contains important information about the target variable?
    - How much?

Create a classification tree from data

- Objective: Based on customer attributes, partition the customers into subgroups that are less impure – with respect to the class (i.e., such that in each group as many instances as possible belong to the same class)




Body shape
Rectangular
Oval
Yes Yes Yes
Yes No Yes
Yes Yes No
No No No


Body color
White
Black
Yes Yes Yes
Yes Yes
No
Decision: NO
Decision: YES
Head shape
Square
Circle
No No
No No
Yes
Yes
Decision: NO
Decision: YES

Calculating Impurity (as Entropy)

- Most common splitting criterion: information gain
- Entropy: impurity measure

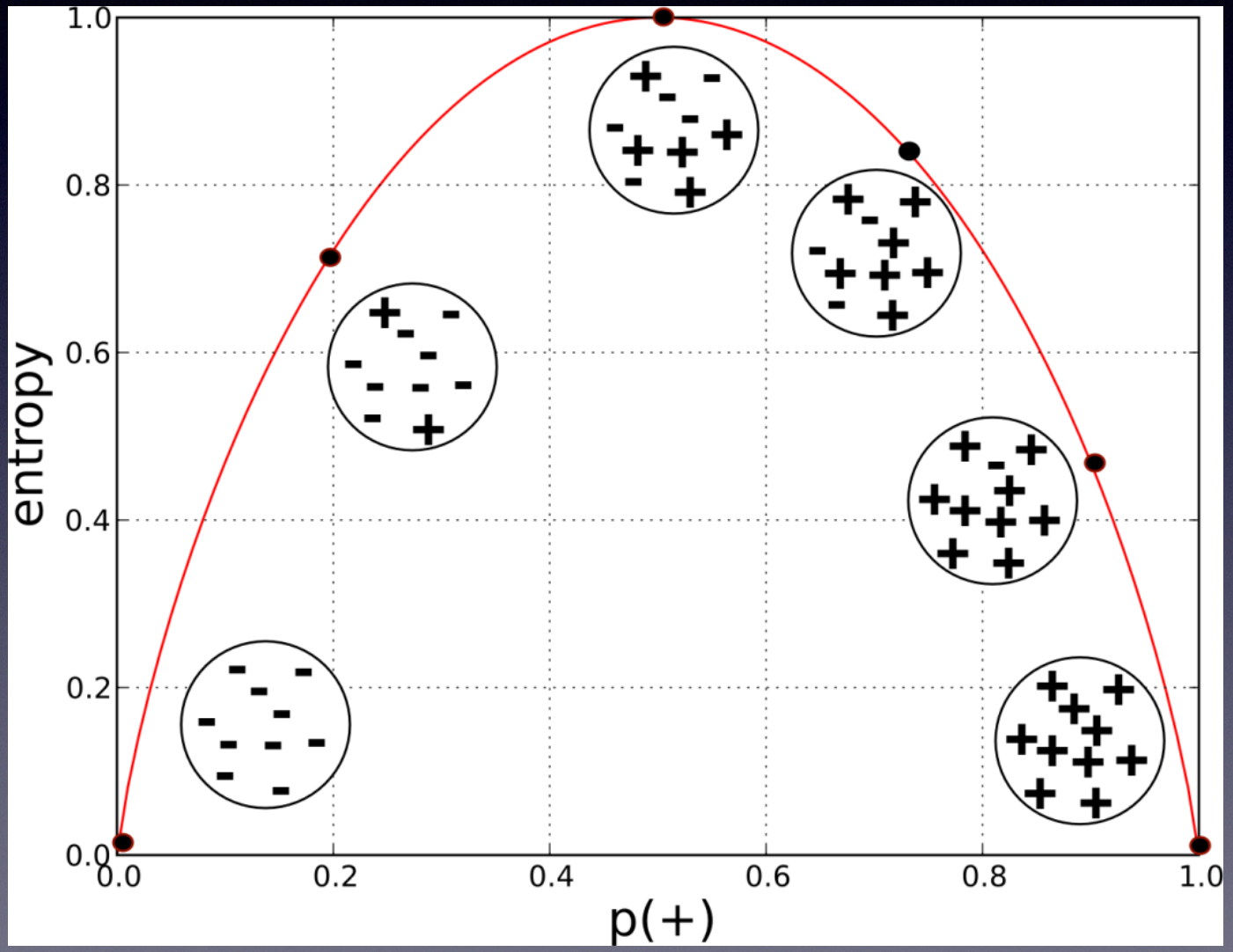
$$entropy = - p_1 \log(p_1) - p_2 \log(p_2) - \cdots$$

p_i is the proportion of class i in the data

- Example: the population is composed of 14 cases of class “Yes” and 16 cases of class “No”

$$-\left(\frac{14}{30} \cdot \log_2 \frac{14}{30}\right) - \left(\frac{16}{30} \cdot \log_2 \frac{16}{30}\right) = 0.996$$

Entropy Function Graph of a Two-class Dataset

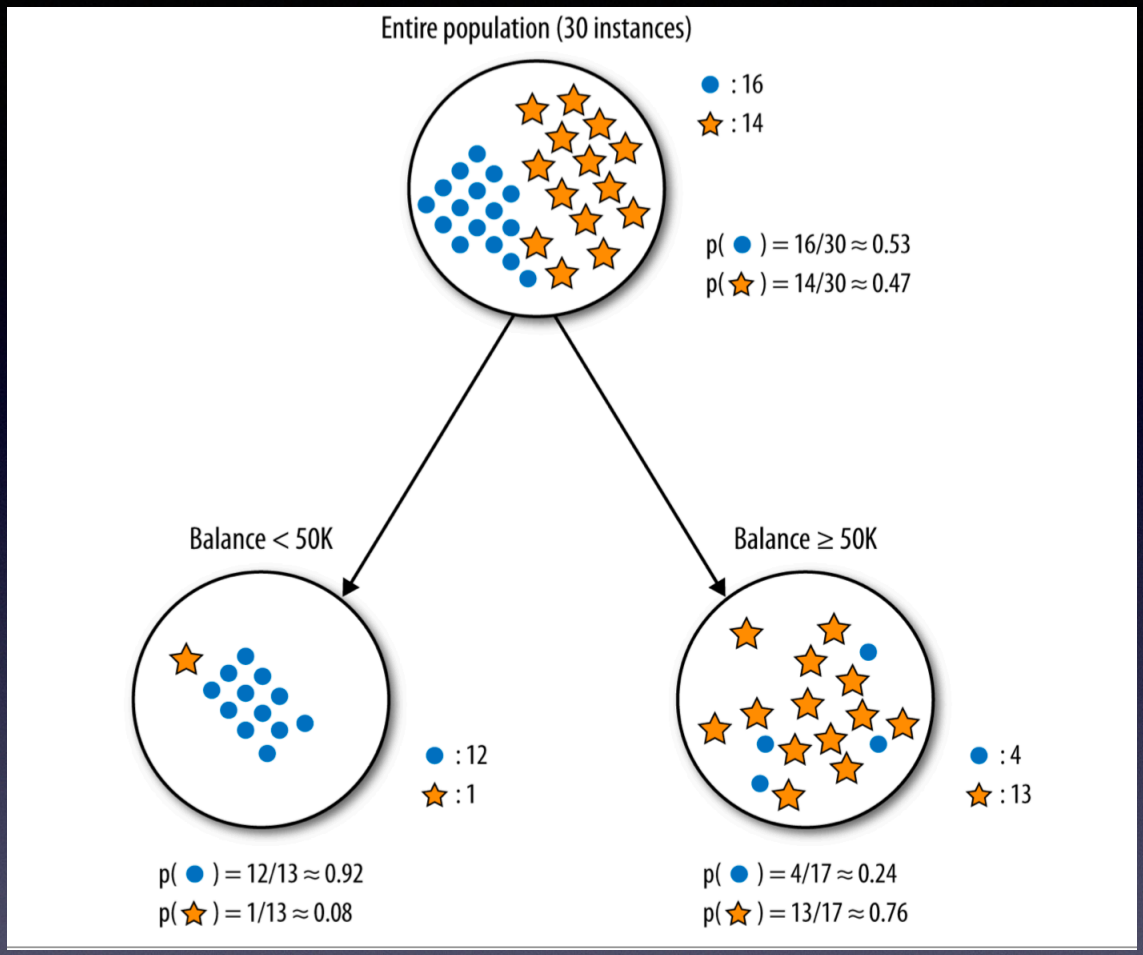


Information Gain

- Information Gain (IG): measure how much an attribute improves (decreases) entropy over the whole segmentation it creates (measures the change in entropy)

$$IG(parent, children) = entropy(parent) - [p(c_1) \times entropy(c_1) + p(c_2) \times entropy(c_2) + \cdots]$$

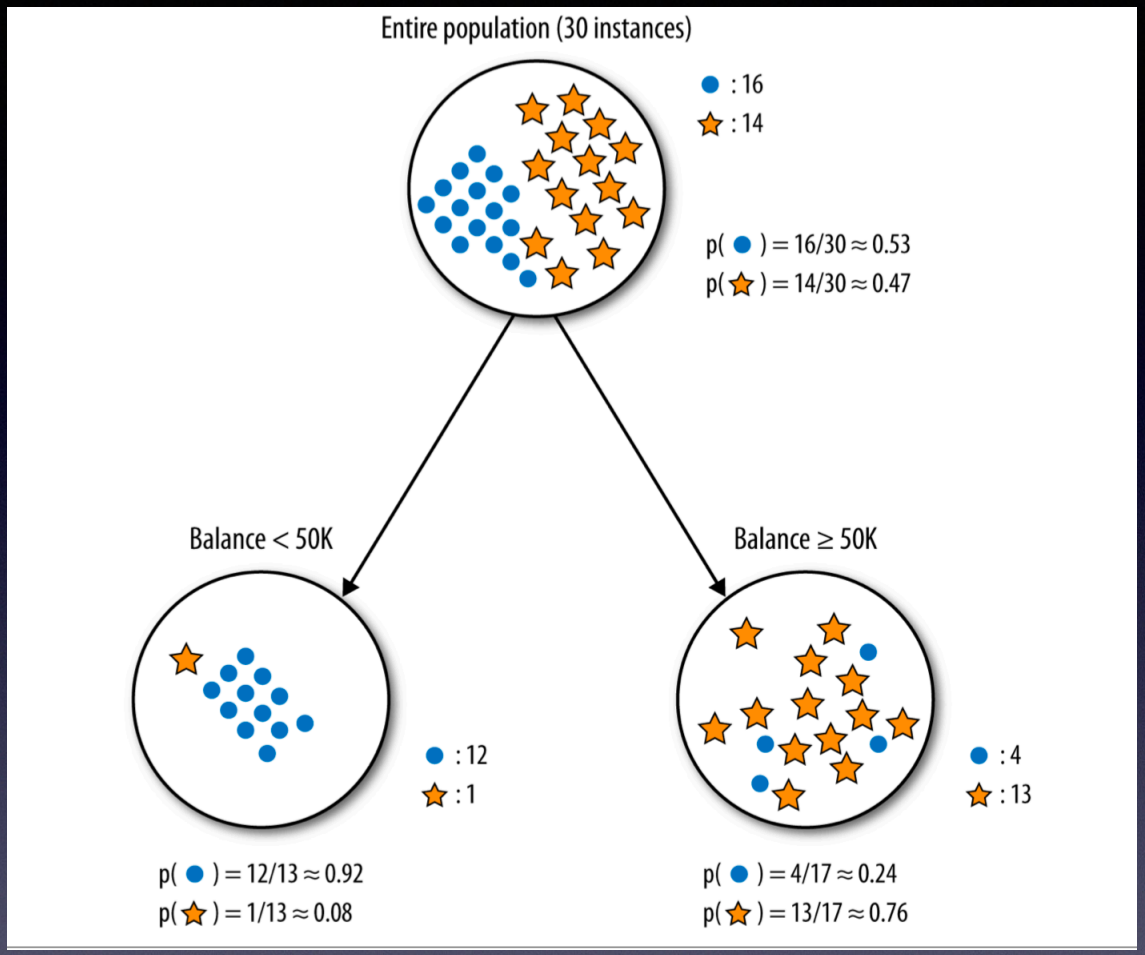
- The entropy for each child (ci) is weighted by the proportion of instances belonging to that child, p(ci).



$$entropy = - p_1 \log(p_1) - p_2 \log(p_2) - \cdots$$

- The information gain of this split is?

$$
\begin{aligned}
entropy(parent) &= -[p(\bullet) \times \log_2\ p(\bullet) + p(\star) \times \log_2\ p(\star)] \\
&\approx -[0.53 \times -0.9 + 0.47 \times -1.1] \\
&\approx 0.99 \quad \text{(very impure)}
\end{aligned}
$$



$$
\begin{aligned}
entropy(Balance < 50K) &= -[p(\bullet) \times \log_2\ p(\bullet) + p(\star) \times \log_2\ p(\star)] \\
&\approx -[0.92 \times (-0.12) + 0.08 \times (-3.7)] \\
&\approx 0.39
\end{aligned}
$$

$$
\begin{aligned}
entropy(Balance \geq 50K) &= -[p(\bullet) \times \log_2\ p(\bullet) + p(\star) \times \log_2\ p(\star)] \\
&\approx -[0.24 \times (-2.1) + 0.76 \times (-0.39)] \\
&\approx 0.79
\end{aligned}
$$

- The information gain of this split is:

$$
\begin{aligned}
IG &= entropy(parent) - [p(\text{Balance} < 50\text{K}) \times entropy(\text{Balance} < 50\text{K}) \\
&\qquad + p(\text{Balance} \geq 50\text{K}) \times entropy(\text{Balance} \geq 50\text{K})] \\
&\approx 0.99 - [0.43 \times 0.39 + 0.57 \times 0.79] \\
&\approx 0.37
\end{aligned}
$$

Why trees?

- Decision trees (DTs), or classification trees, are one of the most popular data mining tools
- They are:
  - Easy to understand and explain
  - Easy to implement
  - Easy to use
  - Computationally cheap
- Almost all data mining packages include DTs

Attribute Selection

- Reasons for selecting only a subset of attributes:
  - Better insights and business understanding
  - Better explanations and more tractable models
  - Reduced cost
  - Faster predictions
  - Better predictions!
    - Too many attributes bring the over-fitting problem
  - And also determining the most informative attributes.

Hands-on: Decision Tree

- Download Class 4 lab from blackboard