

2. SUPERVISED LEARNING

LEV KIWI

MODÉLISATION PRÉDICTIVE

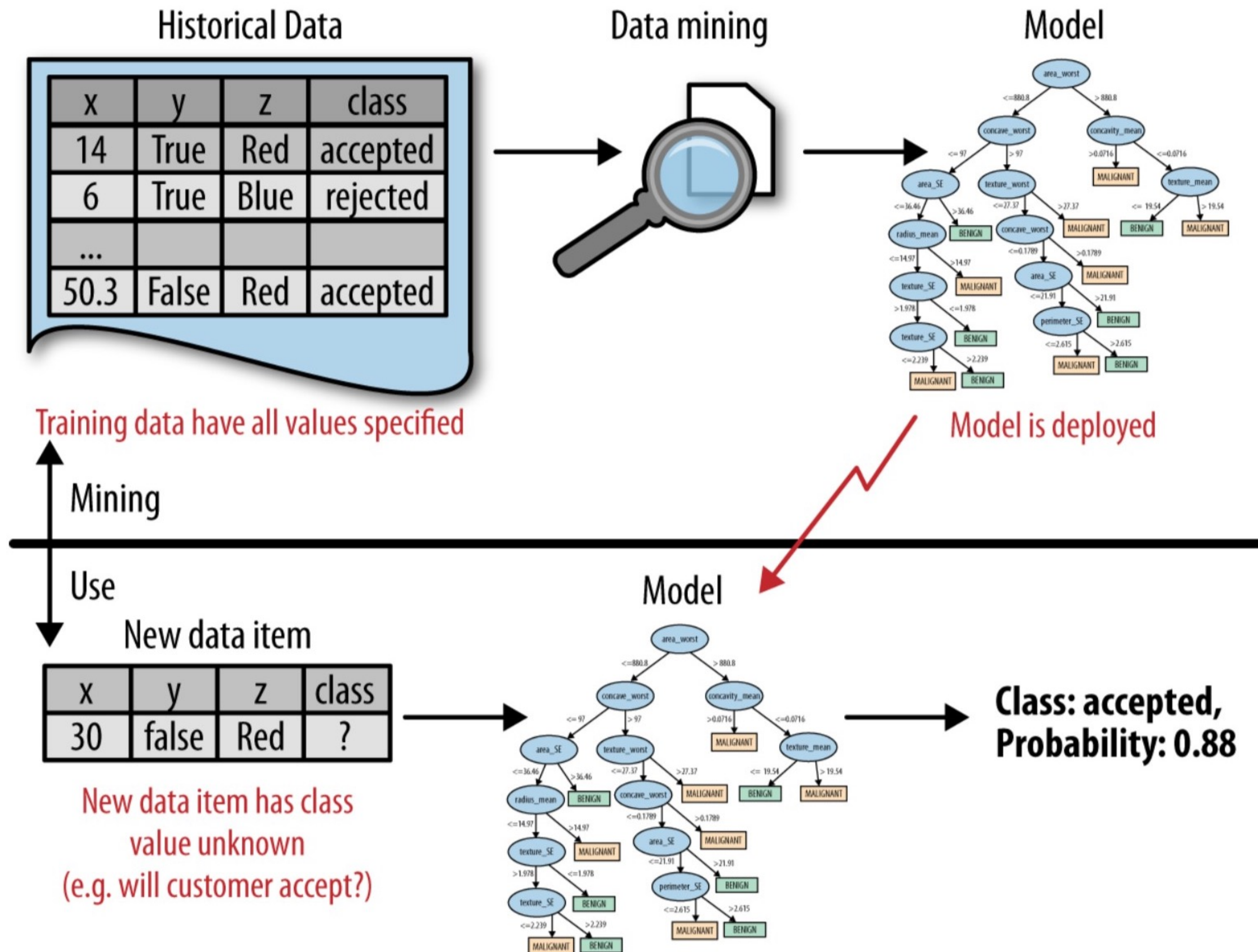
Différence importante

Data Mining.

- Les données sont utilisées pour **comprendre** un phénomène
- Elles sont **interprétées** à la lumière de la compréhension du métier

Data Science.

- Les données sont utilisées pour **prédire** le futur
- Elles servent à **entraîner** un modèle qui va être utilisé par le métier



APPRENTISSAGE SUPERVISÉ

The diagram shows a table with 5 columns: Name, Balance, Age, Employed, and Write-off. A bracket labeled 'Attributes' spans the first four columns. An arrow labeled 'Target attribute' points to the 'Write-off' column. The row for 'Claudio' is highlighted in blue, with an arrow pointing to it from the text below. The text below explains that this row is an example and provides its feature vector and class label.

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

This is one row (example).

Feature vector is: **<Claudio,115000,40,no>**

Class label (value of Target attribute) is **no**

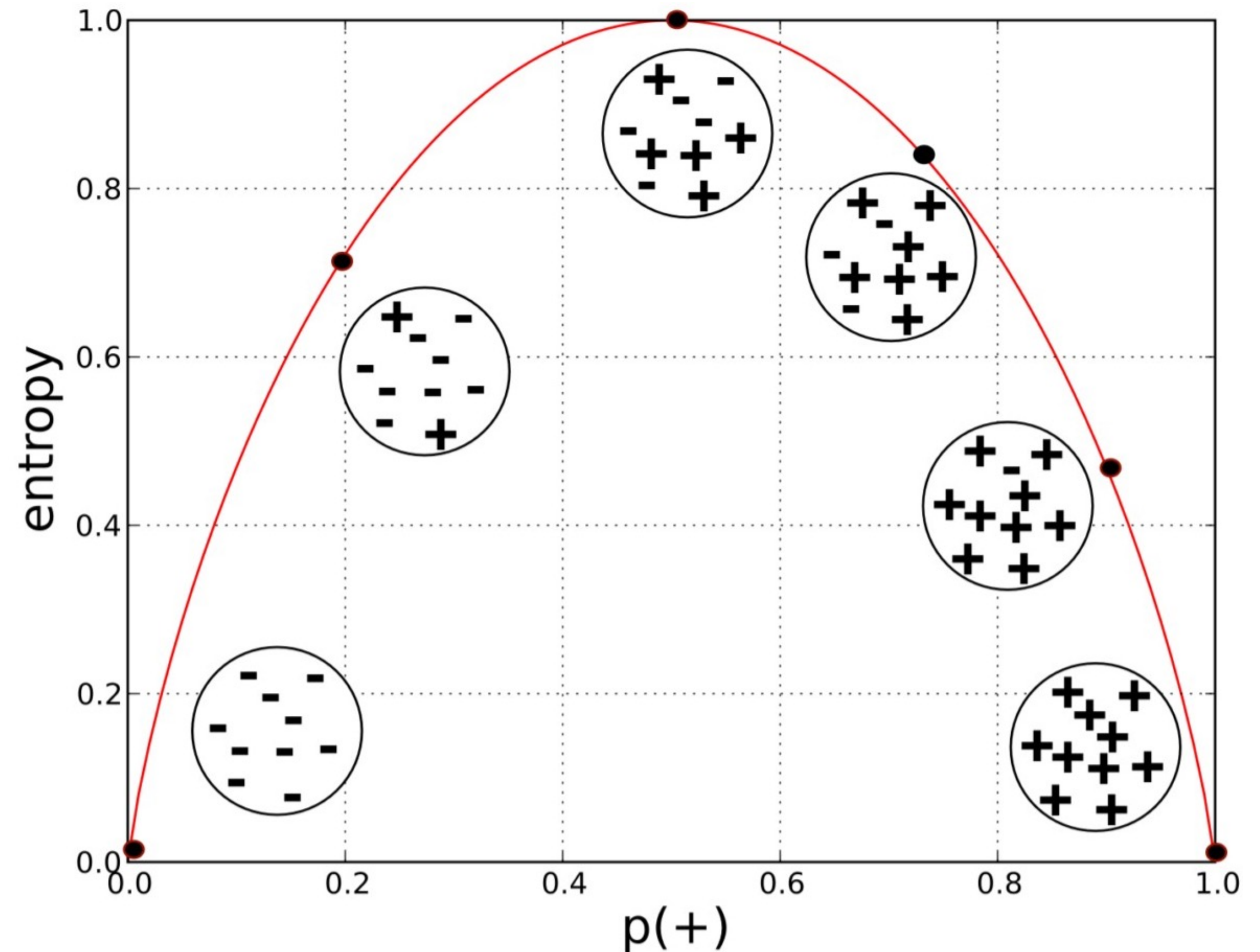
Dataset d'entraînement

- Les données d'entraînement sont **labélisées**
- L'algorithme va créer des associations entre les **attributs** et la **target**

L'INFORMATION EST UNE QUANTITÉ QUI RÉDUIT L'INCERTITUDE

Equation 3-1. Entropy

$$\text{entropy} = - p_1 \log(p_1) - p_2 \log(p_2) - \dots$$



GAIN D'INFORMATION

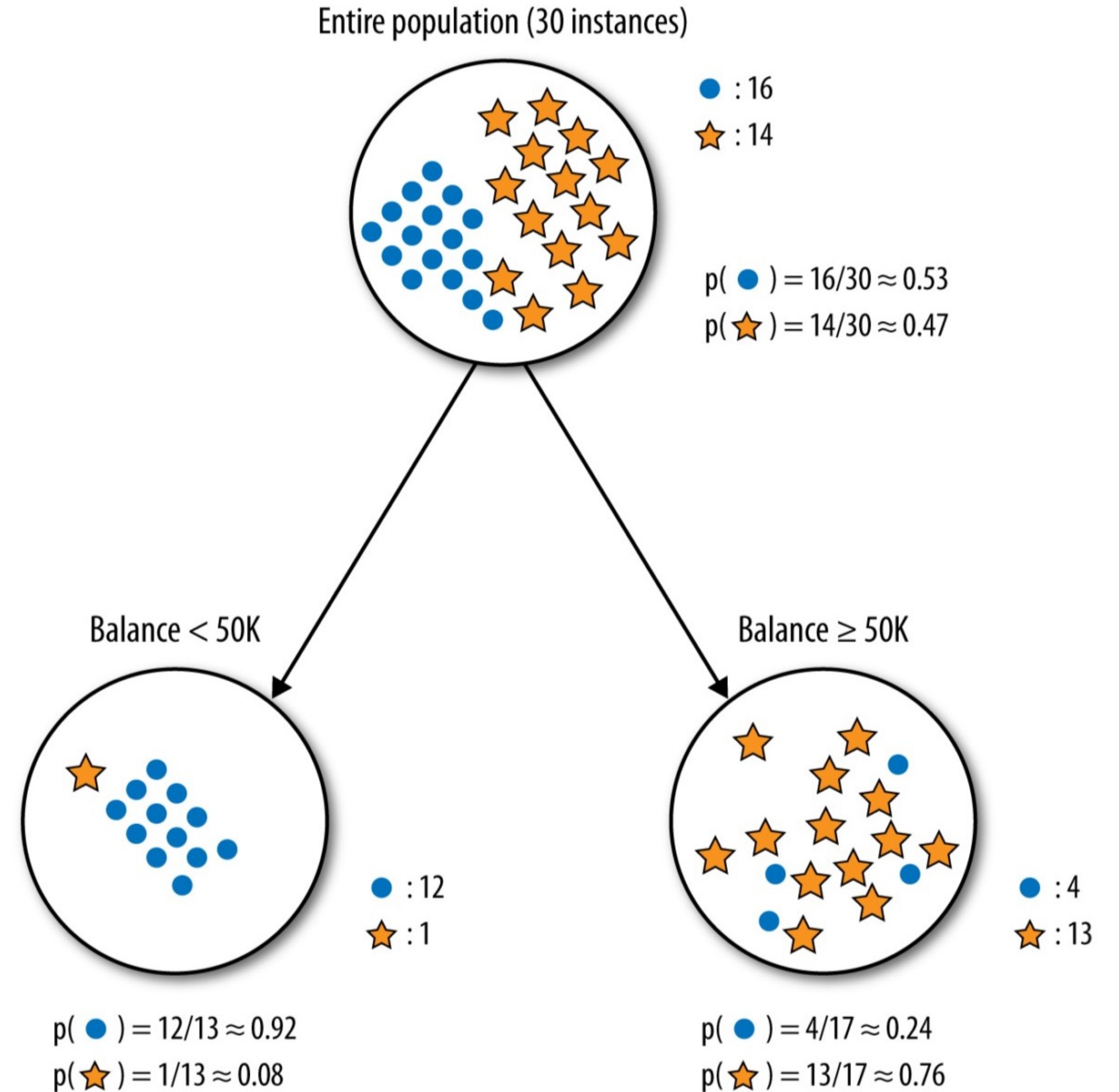
$$\begin{aligned}
 \text{entropy}(\text{parent}) &= -[p(\bullet) \times \log_2 p(\bullet) + p(\star) \times \log_2 p(\star)] \\
 &\approx -[0.53 \times (-0.9) + 0.47 \times (-1.1)] \\
 &\approx 0.99 \quad (\text{very impure})
 \end{aligned}$$

The entropy of the *left* child is:

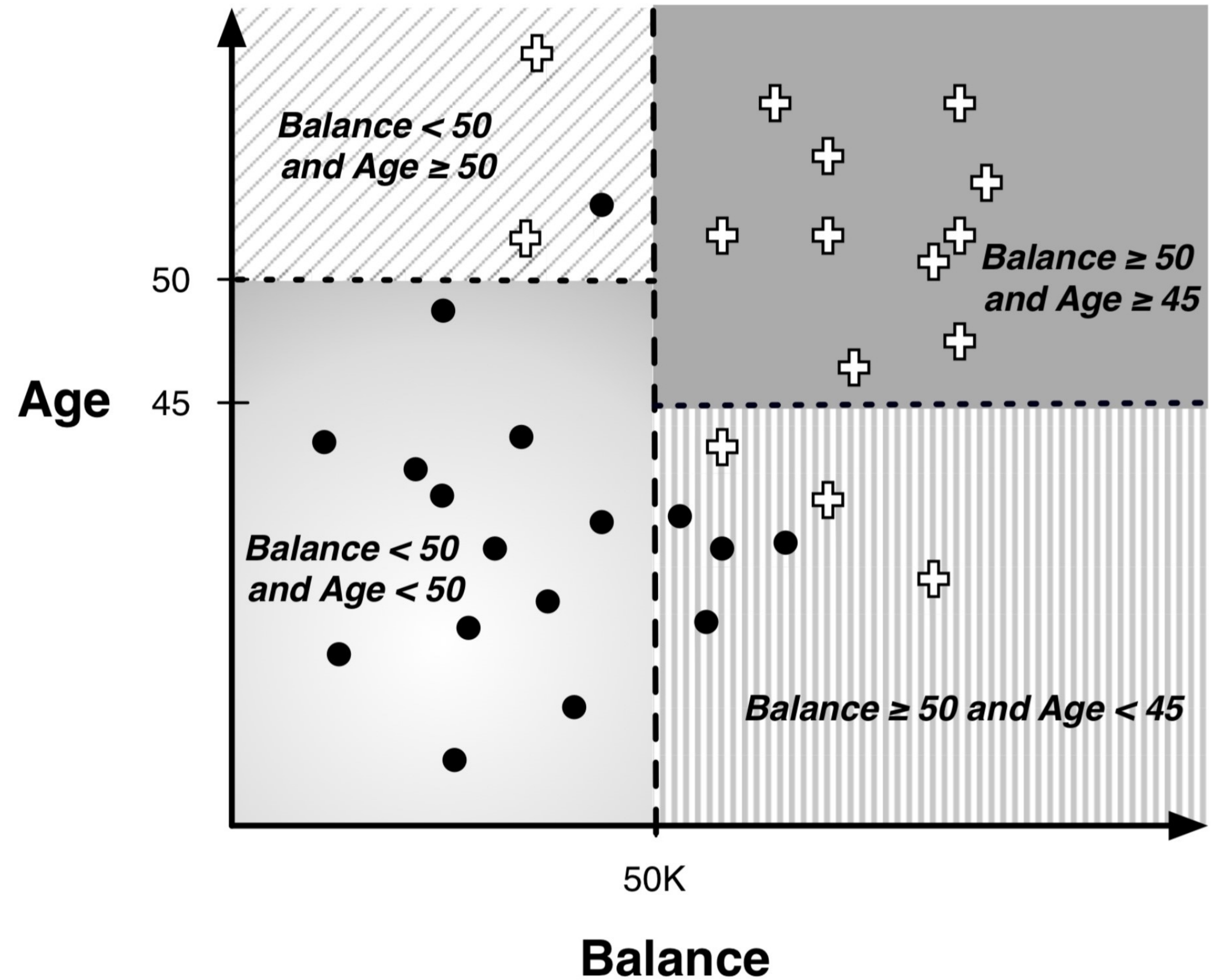
$$\begin{aligned}
 \text{entropy}(\text{Balance} < 50K) &= -[p(\bullet) \times \log_2 p(\bullet) + p(\star) \times \log_2 p(\star)] \\
 &\approx -[0.92 \times (-0.12) + 0.08 \times (-3.7)] \\
 &\approx 0.39
 \end{aligned}$$

The entropy of the *right* child is:

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



ARBRE DE DÉCISION



RÉGRESSION LOGISTIQUE

Equation 4-1. Classification function

$$class(\mathbf{x}) = \begin{cases} + & \text{if } 1.0 \times Age - 1.5 \times Balance + 60 > 0 \\ \bullet & \text{if } 1.0 \times Age - 1.5 \times Balance + 60 \leq 0 \end{cases}$$

