2. Supervised Learning

Lev Kiwi

Modélisation Prédictive

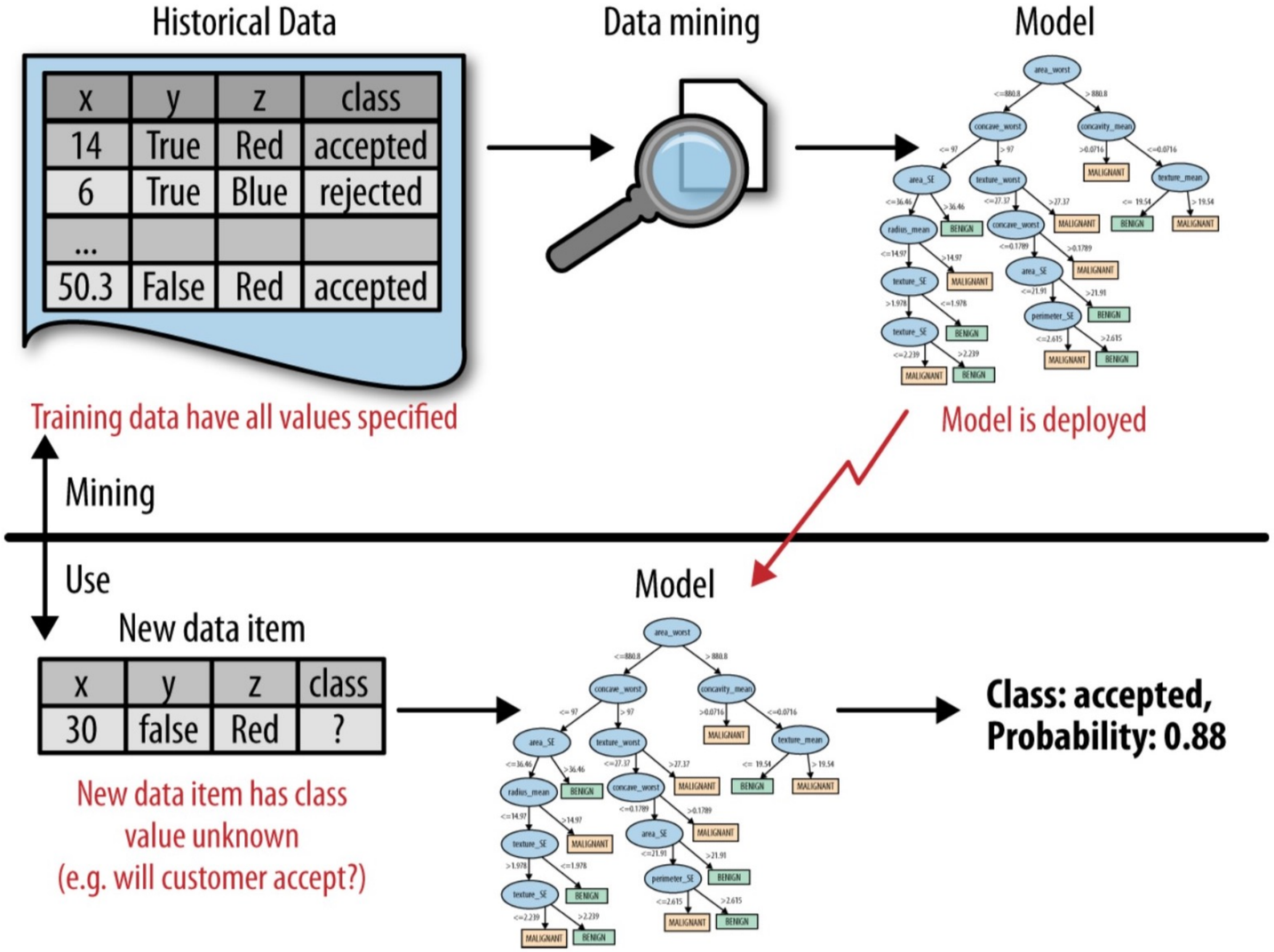
Différence importante

Data Mining.

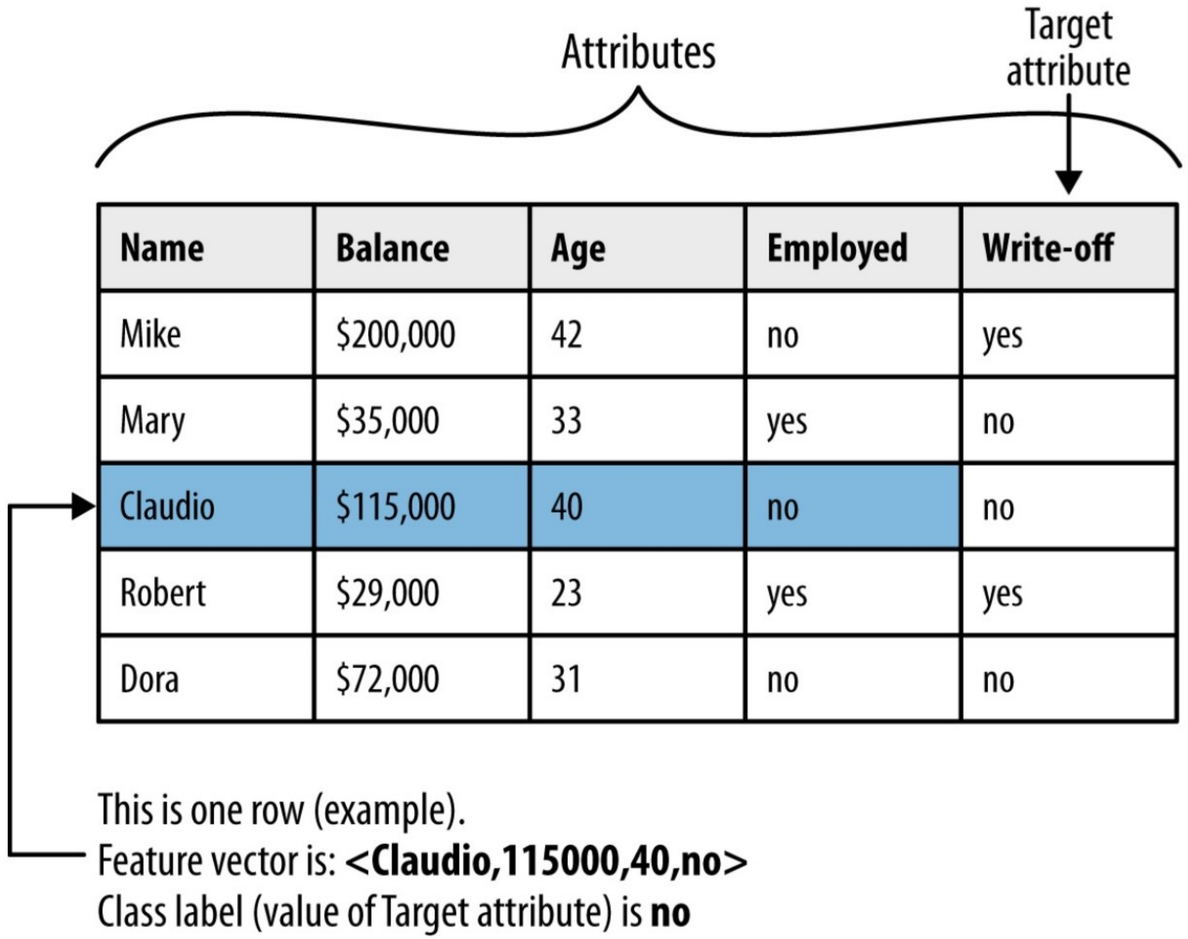
- Les données sont utilisées pour comprendre un phénomène
- Elles sont interprétées à la lumière de la compréhension du métier

Data Science.

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



Apprentissage Supervisé



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

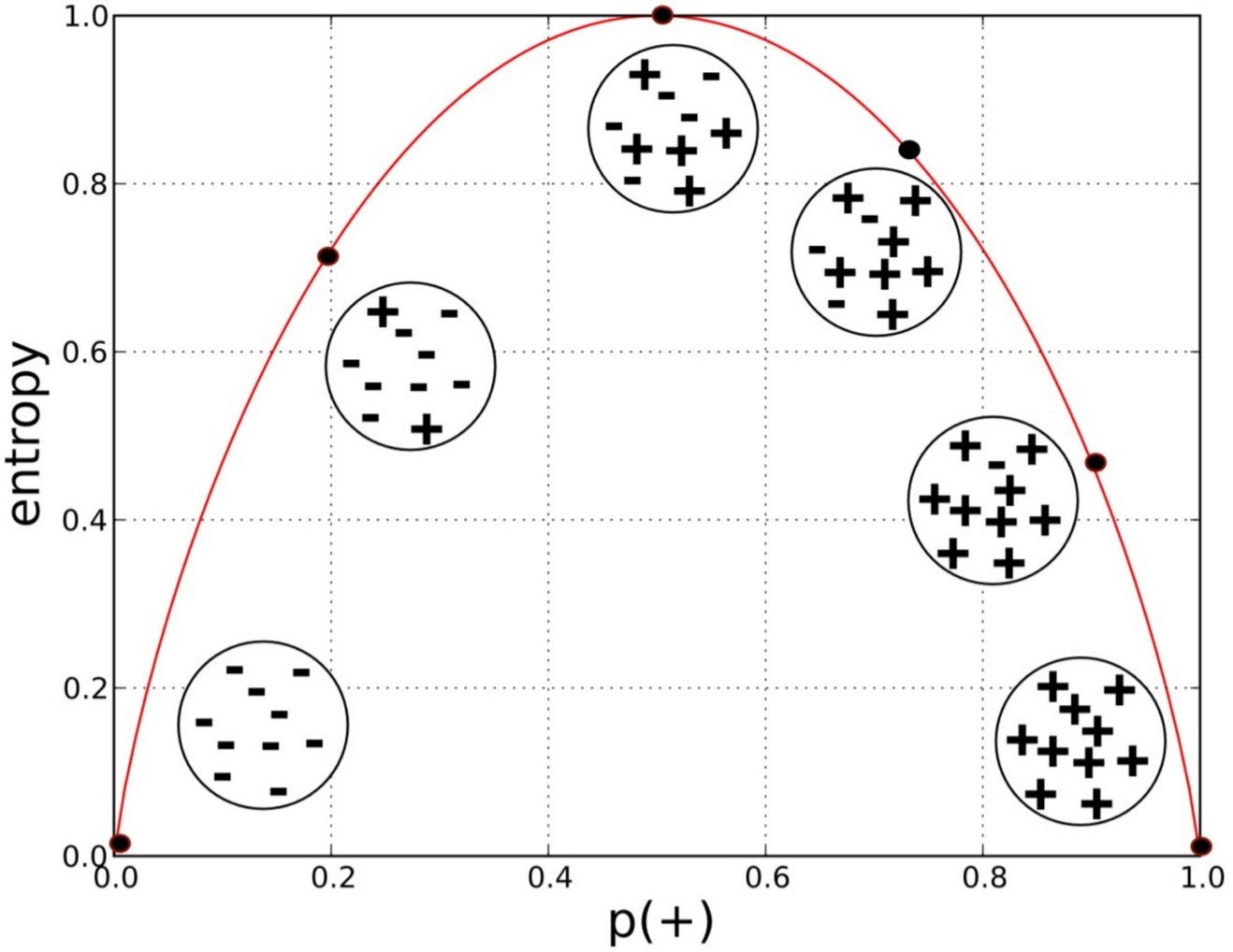
Dataset d'entrainement

- Les données d'entrainement 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

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



GAIN D'INFORMATION

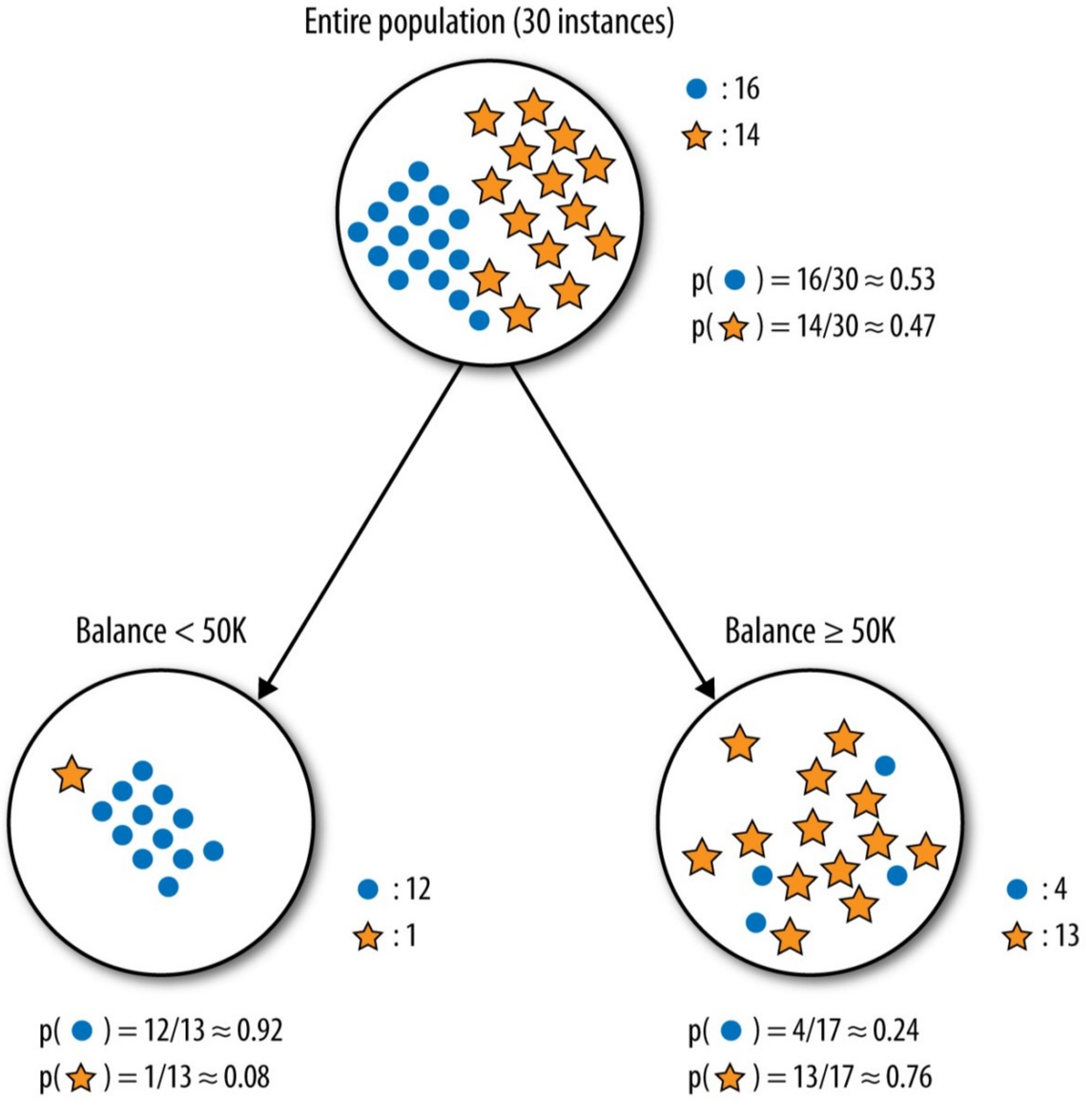
$$
\begin{aligned}
entropy(parent) &= -[p(\bullet) \times \log_2\ p(\bullet) + p(\star) \times \log_2\ p(\star)] \\
&\approx -[0.53 \times -0.9 + 0.47 \times -1.1] \\
&\approx 0.99 \quad \text{(very impure)}
\end{aligned}
$$

The entropy of the *left* child is:

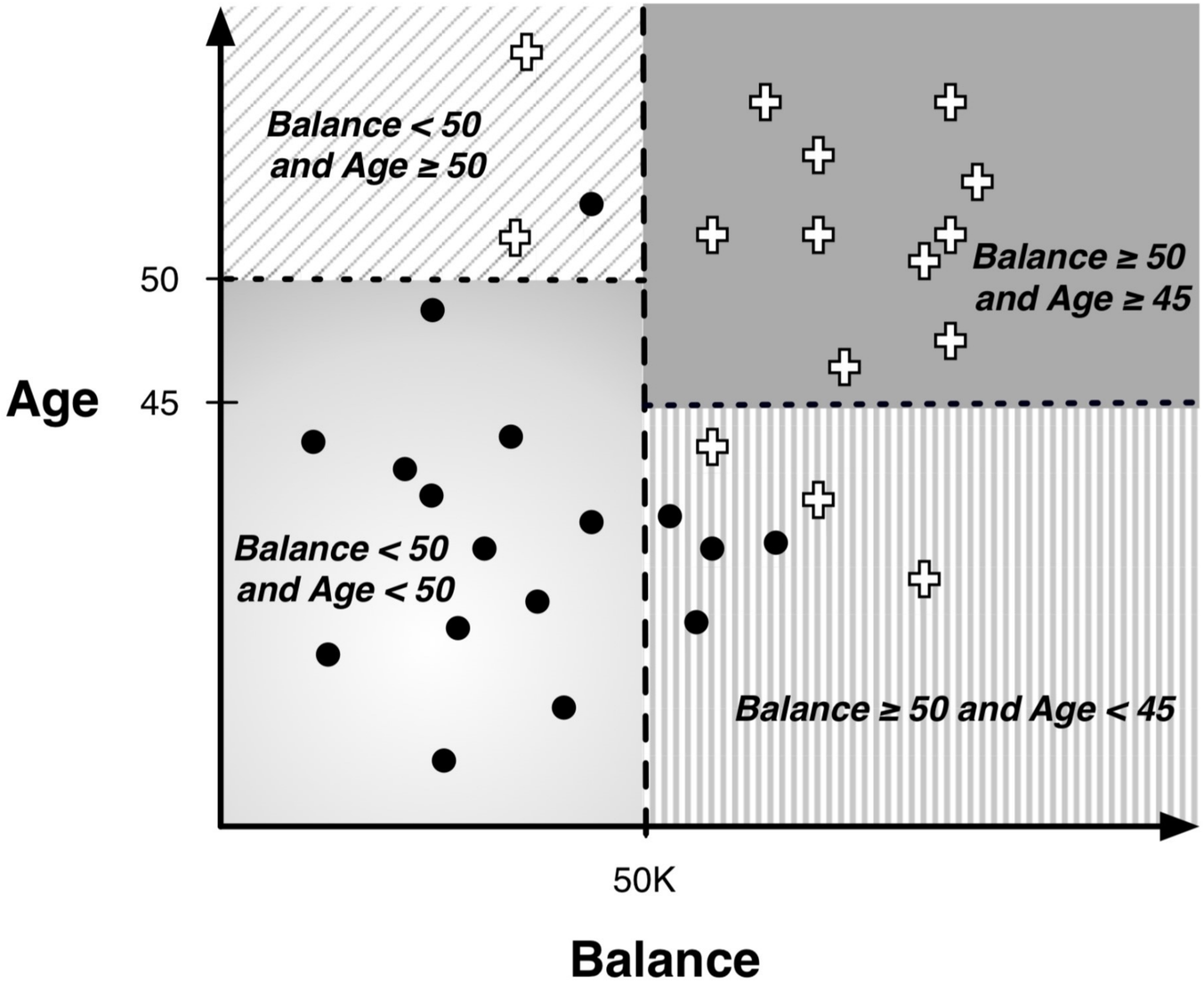
$$
\begin{aligned}
entropy(Balance < 50K) &= -[p(\bullet) \times \log_2\ p(\bullet) + p(\star) \times \log_2\ p(\star)] \\
&\approx -[0.92 \times (-0.12) + 0.08 \times (-3.7)] \\
&\approx 0.39
\end{aligned}
$$

The entropy of the *right* child is:

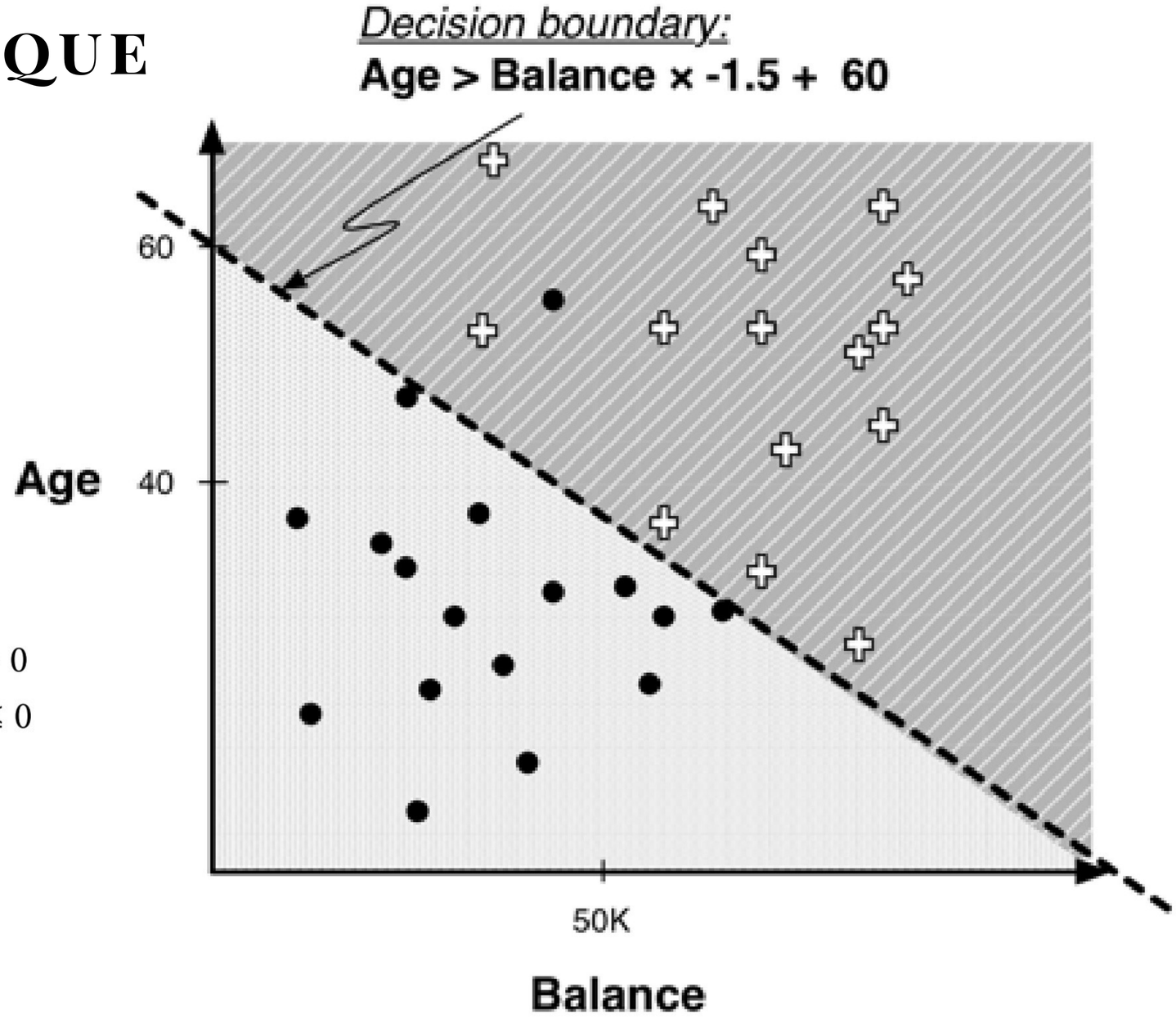
$$
\begin{aligned}
entropy(Balance \geq 50K) &= -[p(\bullet) \times \log_2\ p(\bullet) + p(\star) \times \log_2\ p(\star)] \\
&\approx -[0.24 \times (-2.1) + 0.76 \times (-0.39)] \\
&\approx 0.79
\end{aligned}
$$



Arbre de décision



RÉGRESSION LOGISTIQUE



Equation 4-1. Classification function

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