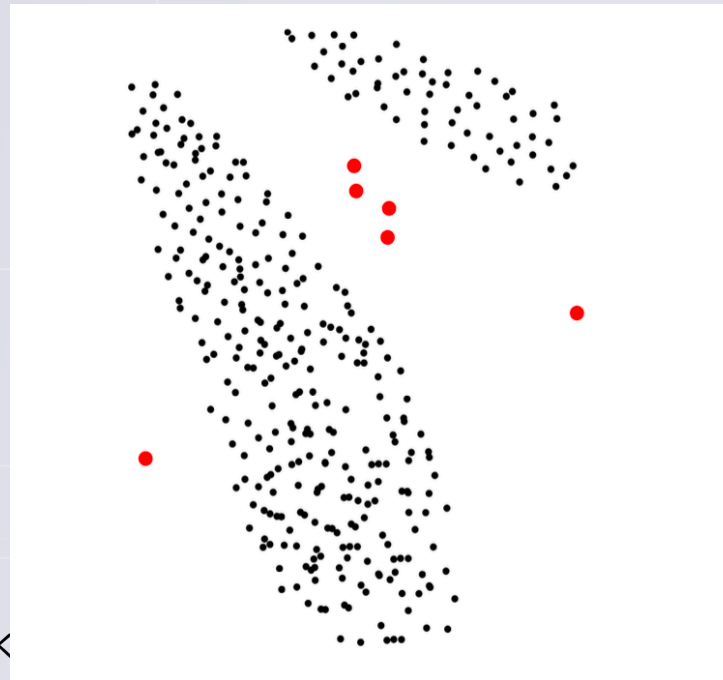


# Détection d'anomalies

Identifier les instances ayant un comportement non conforme



**Applications :** Network  
assurance

crédit, surveillance,

# Détection d'anomalies

- Supprimer ou modifier les observations atypiques à un modèle sans justification serait totalement contraire à l'éthique.
- L'objectif est avant tout de les identifier car ce sont celles, les plus susceptibles d'être la conséquence d'une erreur (à confirmer) de mesure, de libellé, ou encore une anomalie, défaillance ou tentative de fraude, d'intrusion, selon le contexte.

# Plusieurs approches

## ■ Approches supervisées

- Des labels à la fois pour les instances normales et anomalies
- Les anomalies appartiennent à la classe rare
- Données déséquilibrées

## ■ Approches non supervisées

- Pas de labels fournis
- Base d'apprentissage = données normales + anomalies
- Les anomalies sont très rares

# Approches non supervisées

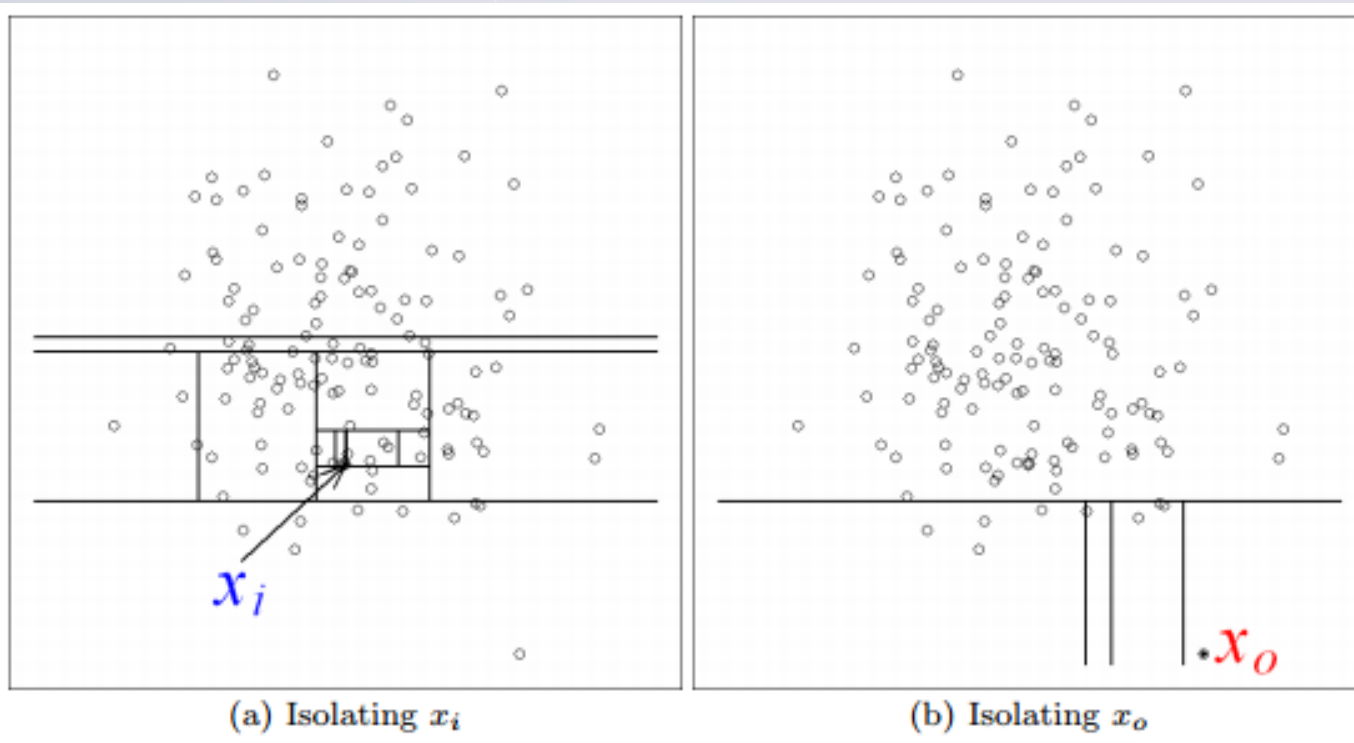
## ■ Approches basées sur le voisinage

- L'anomalie ou l'isolement d'une observation est apprécié par la proximité des points de son voisinage

## ■ Isolation forest

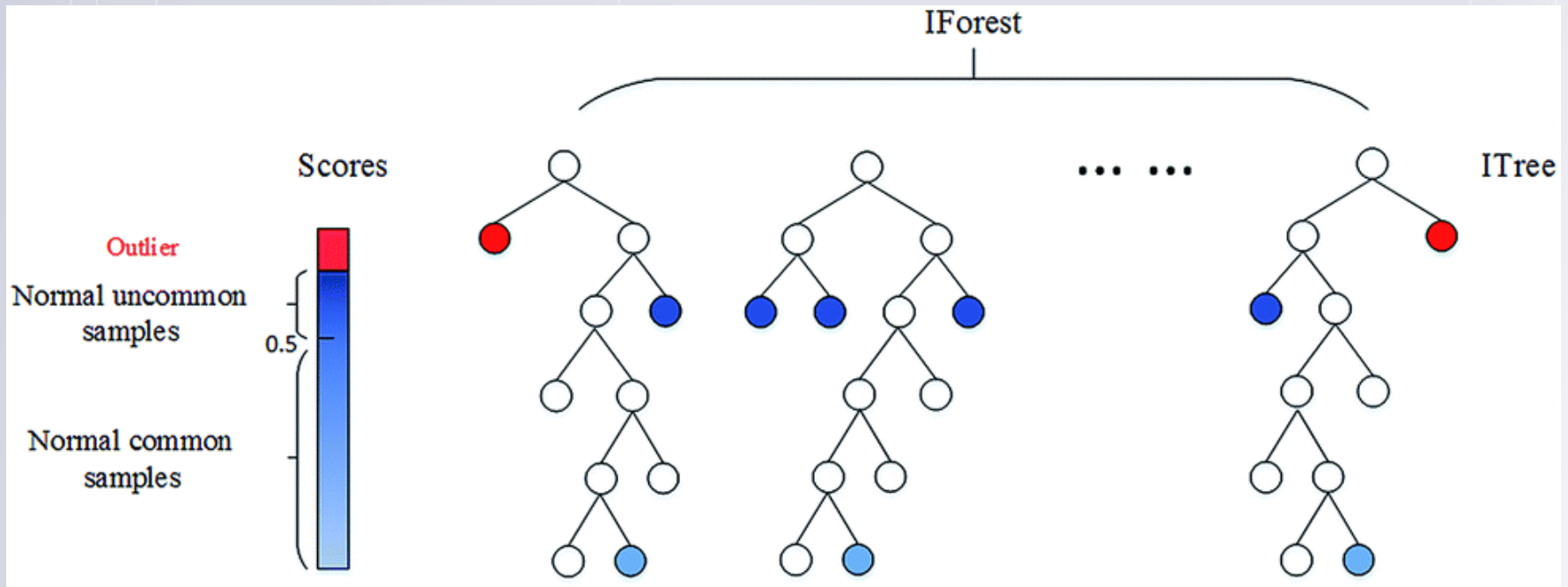
# Isolation Forest

**Principe :** Les anomalies sont rares et différentes. Elles sont donc susceptibles au mécanisme d'isolation



# Isolation Forest

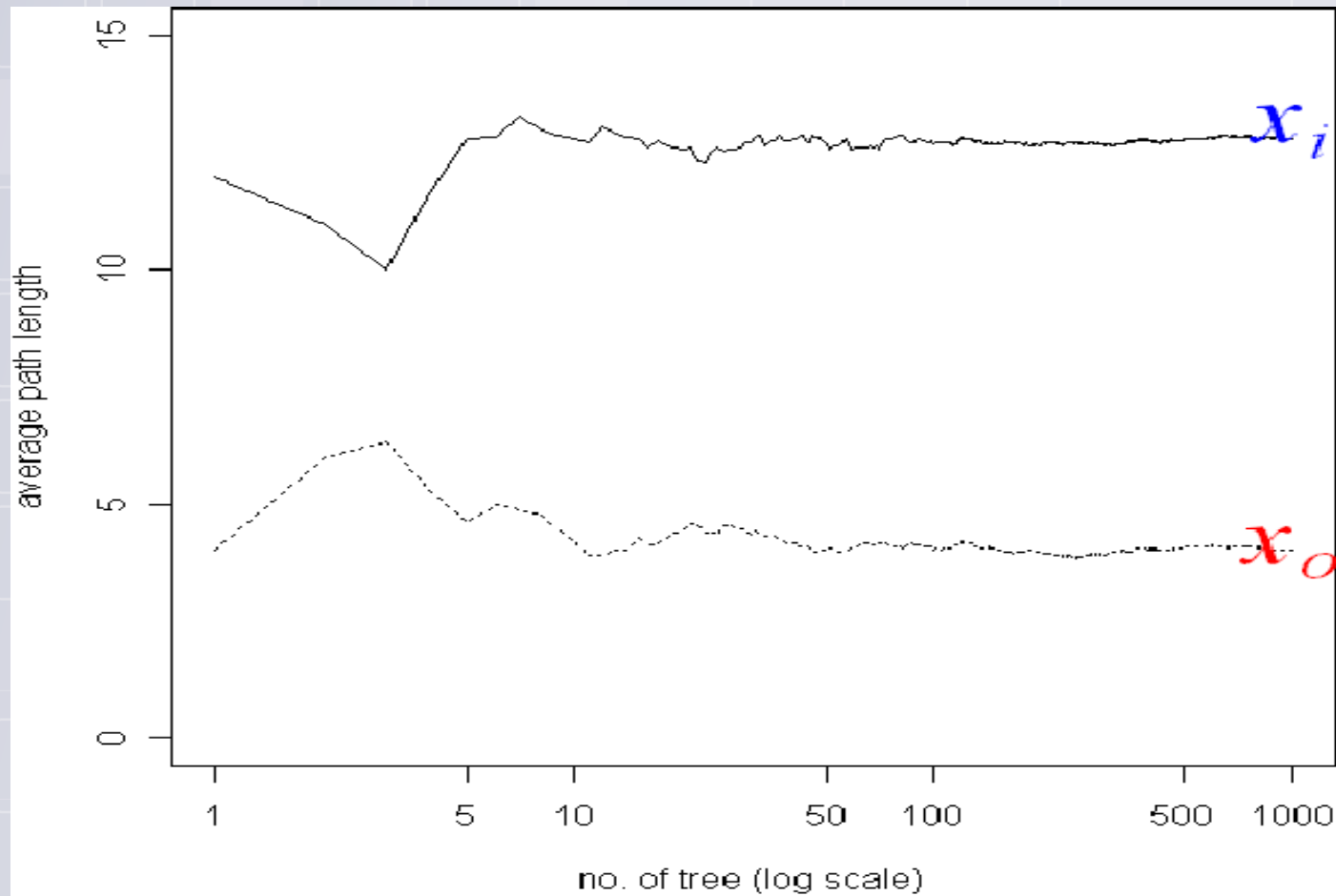
**Principe** : Les anomalies sont rares et différentes. Elles sont donc susceptibles au mécanisme d'isolation



# Isolation Forest

- Le principe repose sur la construction d'un ensemble d'arbres complètement aléatoires : *isolation tree*.
- Chaque arbre est construit sur un échantillon aléatoire des instances
- Division opérée dans chaque nœud via un tirage aléatoire d'une variable et
  - un seuil aléatoire pour une variable quantitative
  - une répartition aléatoire des modalités en deux groupes pour une variable qualitative
  - La construction de l'arbre jusqu'à l'obtention d'une observation par feuille.
- Le score de l'isolement ou de l'anomalie d'une observation est obtenue par la *longueur du chemin atteignant cette observation*. Plus celui-ci est court, plus l'observation est considérée isolée ou atypique.

# Isolation Forest



Convergence  
rapide