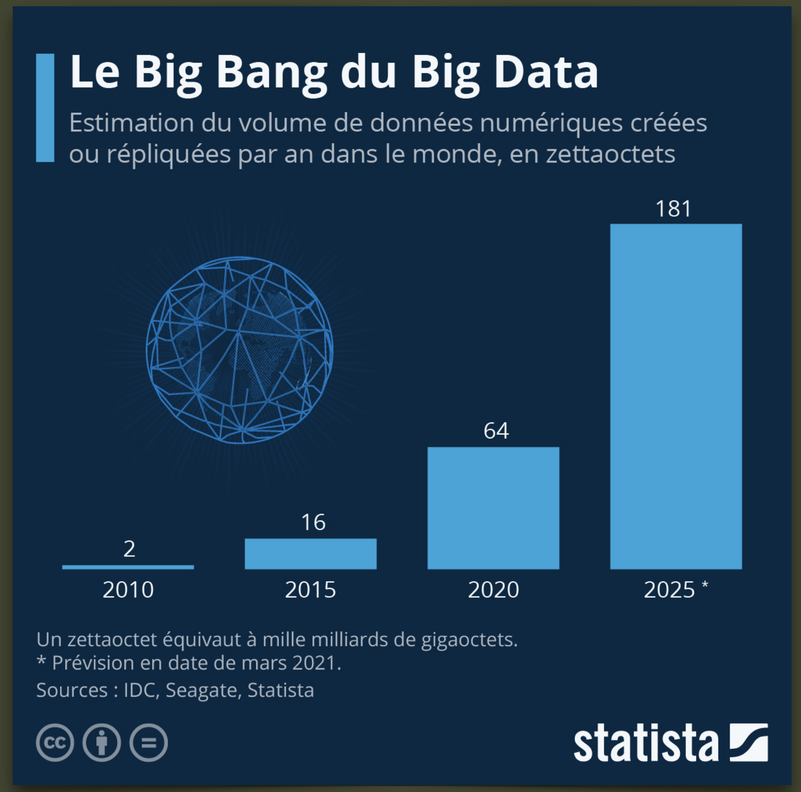
Introduction ML

"By 2025, there will be 175 zettabytes of data globally, with 80% of that data being unstructured. Ninety percent of unstructured data is never analyzed."

- Forbes

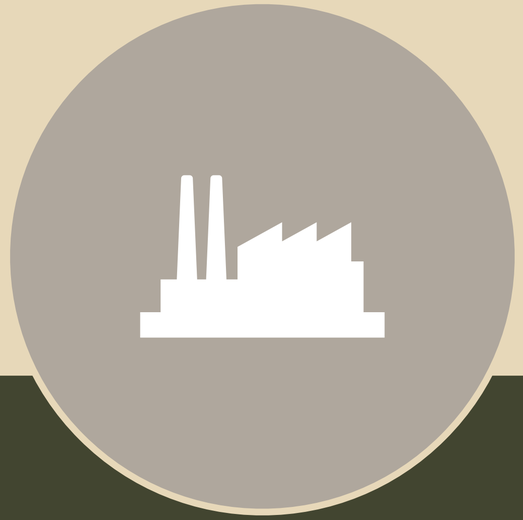


Activités quotidiennes

Ex : Transactions financières internes, Gestion des stock

Produits et services

Ex : Statistiques de vente



Données accessibles

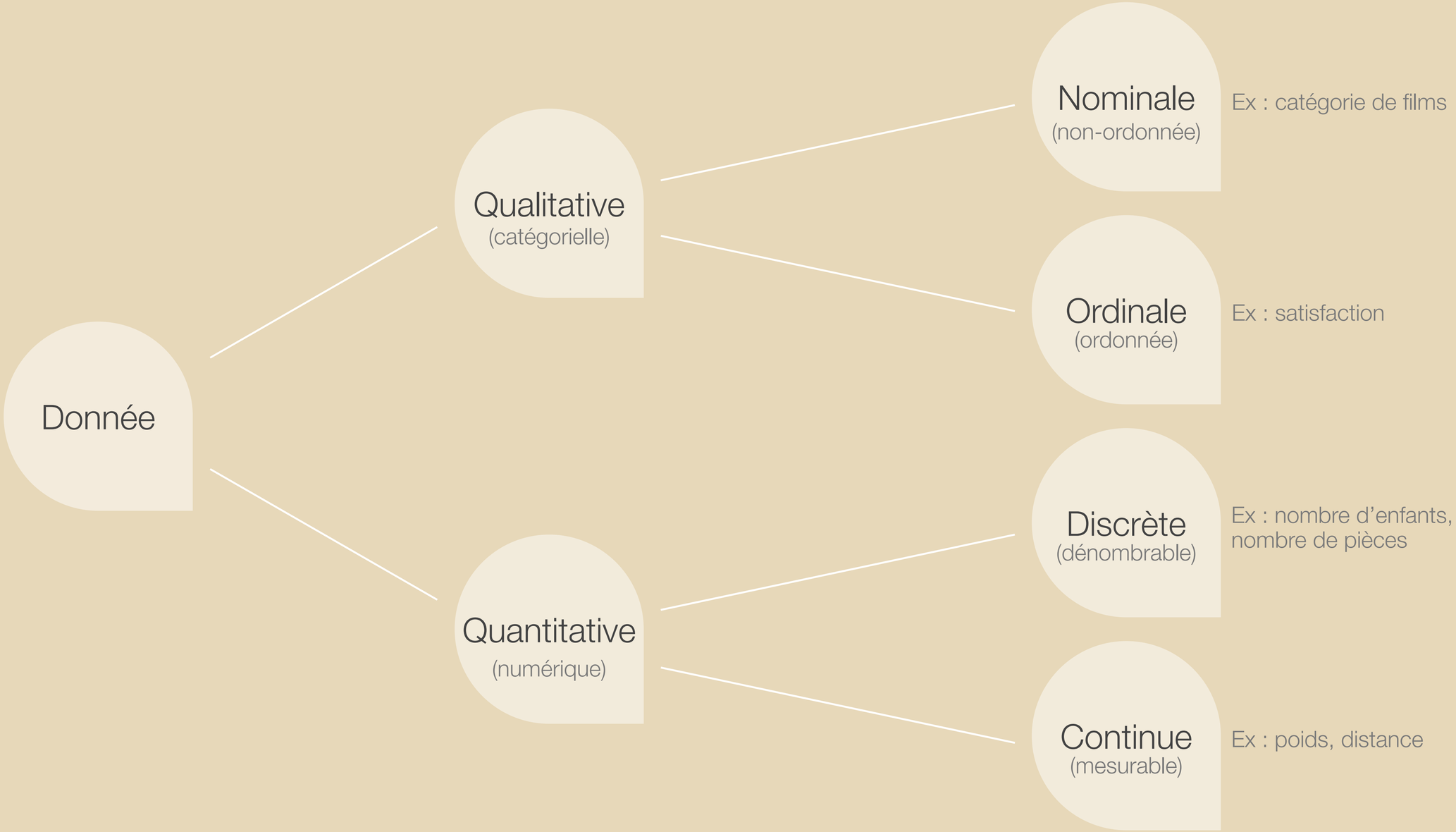
Ex : Bases de données publiques, Études de marché

Données crées et fournis par les partenaires

Ex : Avis clients, Données transactionnelles

Déluge de données

Les différentes formes de données



Pipeline ML

Subject: Waiting for your reply
From: prince1@test.com

We are delighted to inform you that you won 1.000.000 (one million) US Dollars. To claim the prize, you need to pay a small processing fee. Please transfer $10 to our PayPal account at prince@test.com. Once we receive the money, we will start the transfer.

Congratulations again!

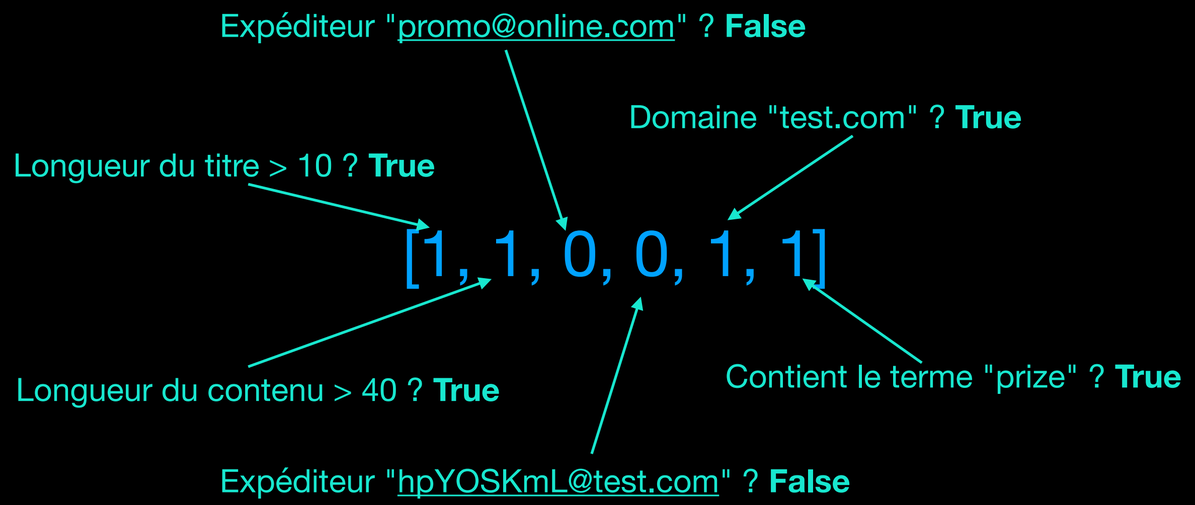
Spam

Pipeline ML

Subject: Waiting for your reply
From: prince1@test.com

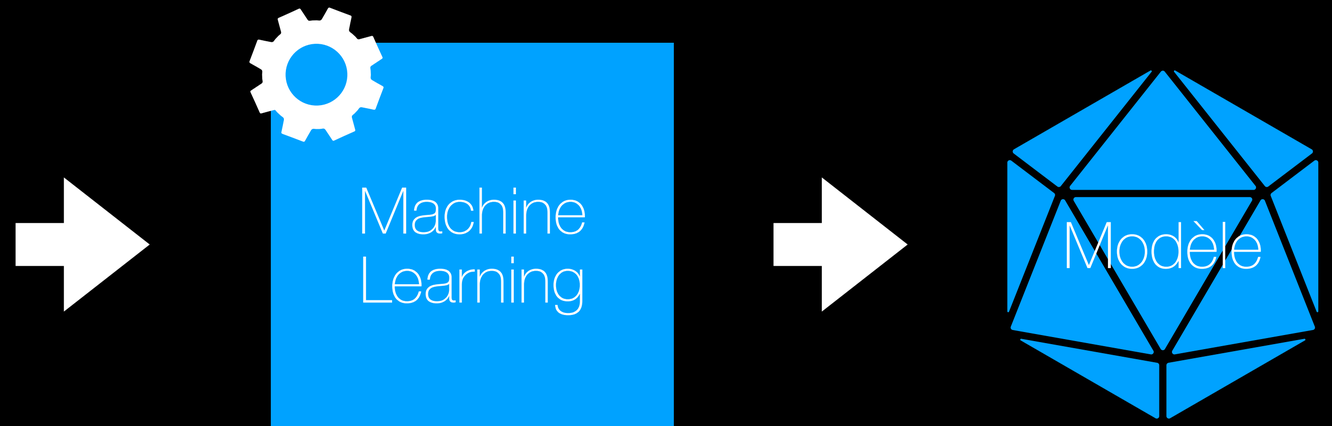
We are delighted to inform you that you won 1.000.000 (one million) US Dollars. To claim the prize, you need to pay a small processing fee. Please
transfer $10 to our PayPal account at prince@test.com. Once we receive the
money, we will start the transfer.

Congratulations again!



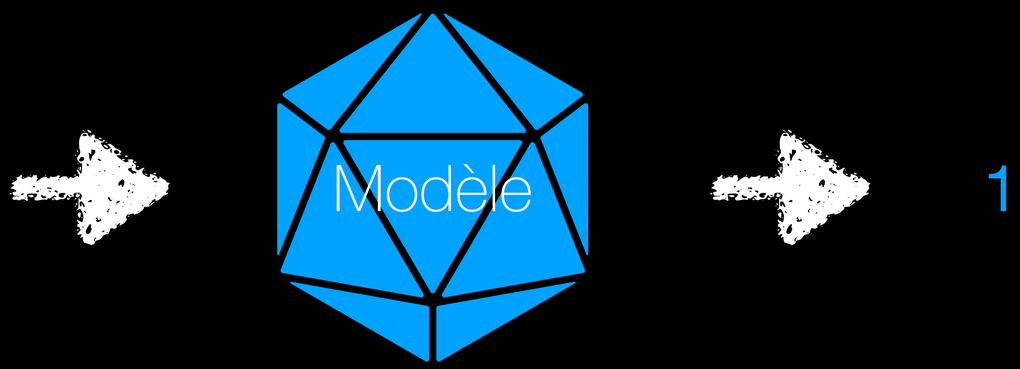
Pipeline ML

Data	Label
[1, 1, 0, 0, 1, 1]	1
[0, 0, 0, 1, 0, 1]	0
[1, 1, 1, 0, 1, 0]	1
[1, 0, 0, 0, 0, 1]	1
[0, 0, 0, 1, 1, 0]	0
[1, 0, 1, 0, 1, 1]	0



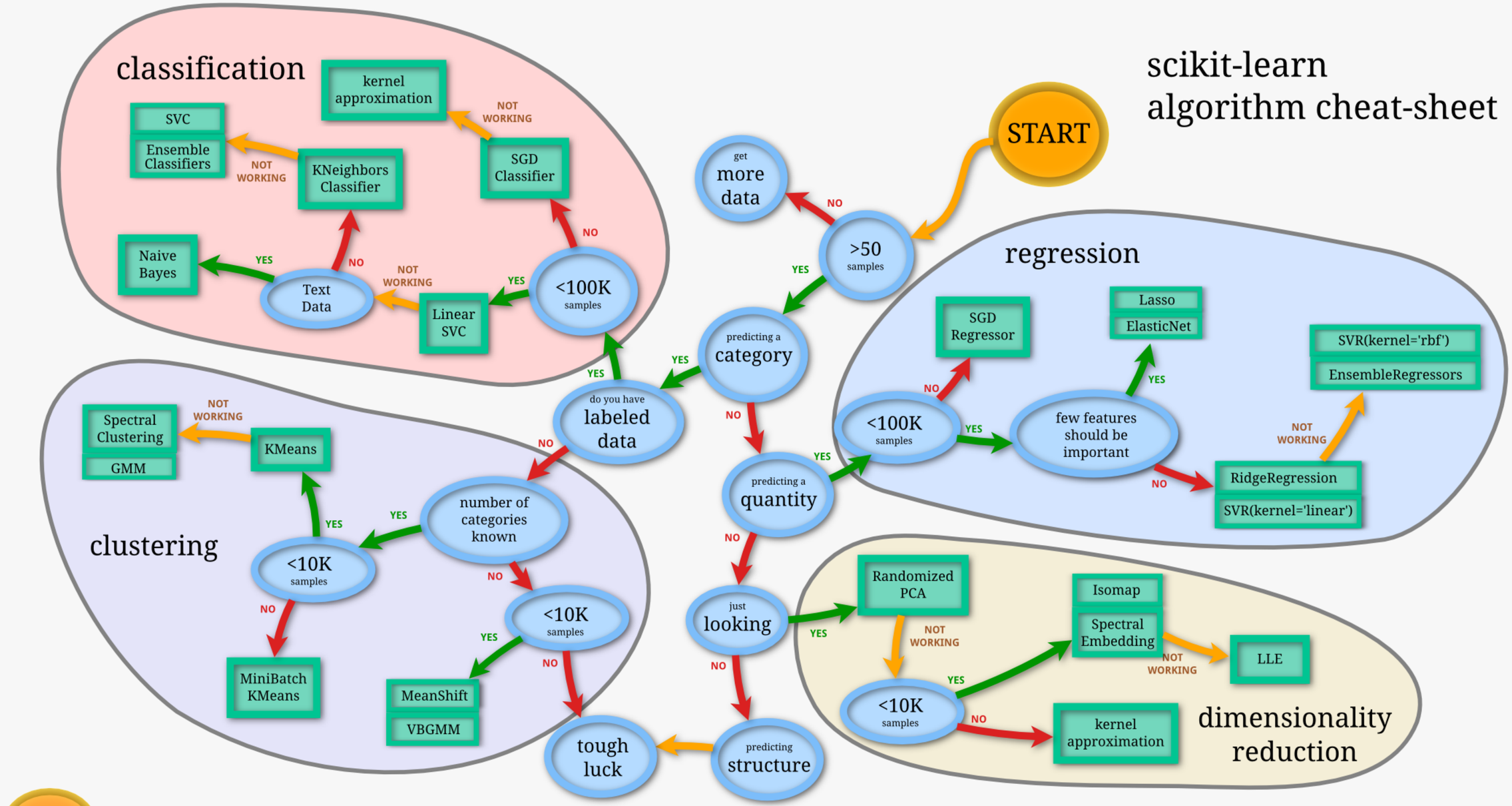
Utilisation du modèle

[1, 0, 0, 0, 0, 1]



1

scikit-learn algorithm cheat-sheet



Résumé

Problème	Objectif	Type d'apprentissage
Régression	Prédire une **valeur numérique continue** (ex. : prix d'un bien immobilier)	🟢 Supervisé
Classification	Prédire une **catégorie** (ex. : spam ou pas spam)	🟢 Supervisé
Clustering	Regrouper des données **sans connaître les catégories à l'avance**	🔵 Non supervisé
Réduction de dimension	Réduire le **nombre de variables** en gardant la structure de l'information	🔵 Non supervisé