Dans les quatres (4) derniers chapitres, l'auteur nous parle de l'entraînement de modèle d'IA de façon générale partant de l'acquisition des données à l'évaluation des modèles obtenus.

Dans le chapitre 5 par exemple, il a introduit quelques techniques utilisées pour faire des transformations sur des donées tellesque la **mise à échelle** des données, la **binarisation**, la **normalisation**, etc.

Ensuite la notion d'apprentissage supervisé et non supervisé on été développé dans les chapitres suivant avec l'entraînement de quelques modèles de classification et de regresion utilisant des algorithmes comme **Linear Regression, Decision Tree, Random Forest, Gradient-Boosted, etc**
Nous savons que l'IA c'est de l'expérimentation et des fois le choix d'algoithmes peut être compliqué car pour un problème donné nous pouvons avoir un choix entre plusieurs algorithmes. Par exemple pour un problème de classification, nous pouvons utliser des algorithmes comme le **Naive Bayes, Decision Tree, Random forest, Logistic Regression, ...**
Le choix semble difficile puisque nous pouvons choisir un algorithme qui va nous donner, après entrainement, un modèle qui a "overfitté" c'est à dire qui fait de bonnes performances avec les données d'entrainement mais qui n'arrive pas à généraliser sur de nouvelles données.
Ainsi c'est là que nous pouvons utiliser des techniques comme la validation croisée(cross validation) qui nous peret d'entrainer plusieurs modèles et de choisir celui qui a les meilleures performance sur les données de test.
Notons que tous ces algorithmes cité ci-dessus sont utilisés dans cadre d'un apprentissage supervisé où on cherche à predire une variable continue ou catégorielle.

Dans le cadre d'un apprentissage non supervisé, on cherche pas à prédire une valeur mais plutôt à organiser les données en groupes(ou clusters). Chaque groupe doit comprendre des données similaires et les données différentes doivent se retrouver dans des groupes distincts.
Nous avons des cas d'utilisation comme la **détection d'anomalies ou de fraude, segmentation de la clientèle, visualisation et réduction de la dimensionnalité**, …; et des algorithmes comme **K-Means, Principal Component Analysis (PCA), Hierarchical clustering**, etc, sont utilisés

Puisque notre tâche consiste à selectionner un algorithme et l'entraîner sur un jeu de données, les principaux challenges du Machine Learning pourraient être:

- Quantité insuffisante de données d'entraînement
- Données d'entraînement non représentatives
- Données de mauvaise qualité
- Caractéristiques non pertinentes

Les algorithmes de Machine Learning présentent des limites surtout quand il s'agit de traîter des données non linéaire ou dans le monde des images et c'est là que le **Deep Learning** entre en jeu.
Sous-branche du ML, le Deep Learning repose principalement sur l'utilisation de réseaux de neurones pour résoudre les problèmes. Il se base sur la façon dont le cerveau humain traite les informations et apprend, répondant aux stimuli externes.
De nos jours, il est fortement utilisé surtout dans le domaine de la médecine pour la mise en place de soltuions comme la detection de cancer à travers des images radiographiques, de même que la conception de voiture autonomes.
Les framework les plus utilisés pour mettre en place des modèles de DL sont **Tensorflow Keras et Pytorch** mais ces derniers n'ont pas été utilisé dans ce livre.

Ainsi notre plus grand challenge serait de voir comment implémenter des modèles de DL en utilisant Tensorflow Keras ou Pytorch avec du Spark pour pouvoir entraîner des modèles sur des images aussi puisque tout au long de ce livre des données tabulaires ont été uniquement utilisées.