Module 2

**INTRODUCTION MACHINE LERNING**

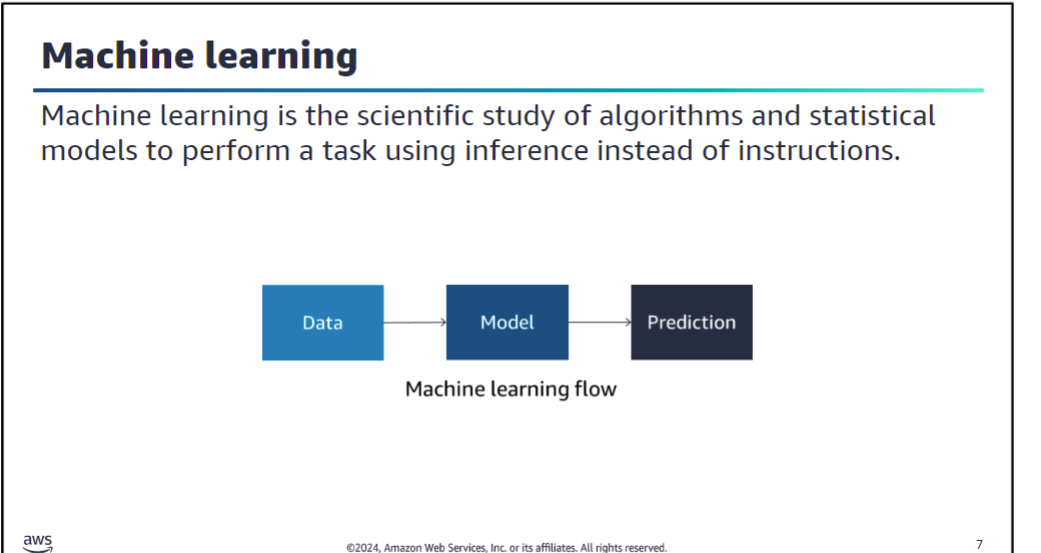
**Artificial intelligence**

'apprentissage automatique est une sous-catégorie d'un domaine plus large de l'informatique connu sous le nom d'intelligence artificielle (IA). L'IA consiste à construire des machines capables d'effectuer des tâches qu'un humain accomplirait habituellement. Dans la culture moderne, les IA apparaissent dans des films ou des œuvres de fiction. Vous vous souvenez peut-être de certaines IA dans des films de science-fiction ou des séries télévisées qui contrôlent le monde futur, ou agissent de manière intelligente de façon autonome, parfois avec des effets négatifs pour la société ou les êtres humains autour d'elles. Ces IA ont commencé en tant qu'agents informatiques qui percevaient leur environnement et prenaient des mesures pour atteindre un objectif spécifique. Cependant, pour certaines de ces IA fictives, leurs actions n'étaient pas le résultat que leurs créateurs avaient initialement envisagé. D'autres IA fictives sont plus bénignes ou positives : elles travaillent mieux avec l'humanité, mais elles sont également plus générales dans leur objectif. Ces types d'IA générales sont des exemples d'intelligence artificielle générale (IAG). Elles ont la capacité d'apprendre ou de comprendre n'importe quelle tâche qu'un être humain peut comprendre.

Les problèmes liés à l'IA couvrent généralement de nombreux domaines de recherche : le traitement du langage naturel, le raisonnement, la représentation des connaissances, l'apprentissage, la perception et l'interaction avec l'environnement physique. L'IA n'est pas encore une réalité, à moins que vous ne viviez dans une simulation. Cependant, elle se rapproche dans chacun de ces domaines chaque année.

Vous avez peut-être également lu ou vu des commentaires sur l'éthique de la création de l'IA. Tous les points de vue ne sont pas positifs—peut-être en partie par crainte des IA fictives, qui détruisaient les êtres humains ou les utilisaient comme sources d'énergie. Certains peuvent également considérer le risque de chômage massif, car une machine intelligente peut travailler en continu sans pause.

Module 2 de la formation et certification AWS : Introduction à l'apprentissage automatique © 2024, Amazon Web Services, Inc. ou ses affiliés. Tous droits réservés.

****

Vous trouverez probablement de nombreuses définitions de l'apprentissage automatique. Il n'existe pas de définition standard, c'est pourquoi vous allez maintenant apprendre certaines définitions de l'apprentissage automatique. L'apprentissage automatique est l'étude scientifique des algorithmes et des modèles statistiques pour accomplir une tâche en utilisant l'inférence au lieu d'instructions explicites. Pour vous aider à mieux comprendre cette idée, voici un exemple concret.

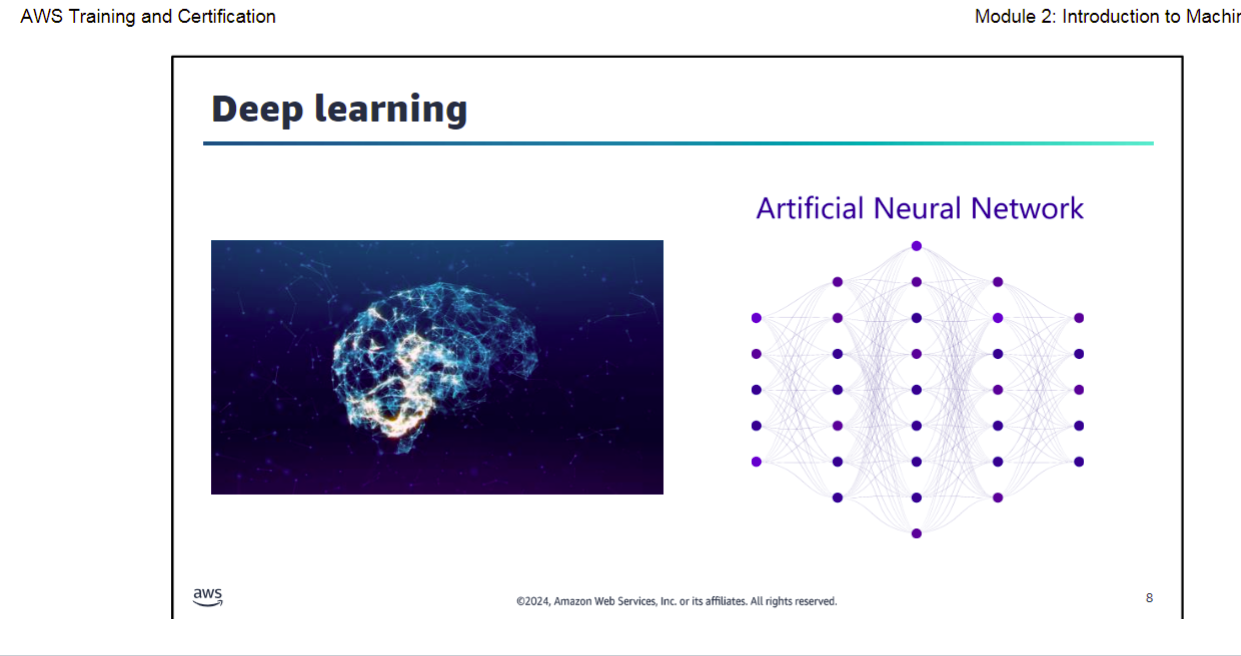
Supposons que vous deviez écrire une application qui détermine si un message électronique est un spam ou non. Sans apprentissage automatique, vous écririez une série complexe de déclarations conditionnelles (pensez à des déclarations if/else). Peut-être utiliseriez-vous les mots dans l'objet ou le corps du message, le nombre de liens et la longueur du courriel pour déterminer s'il s'agit de spam. Il serait difficile et laborieux de compiler un ensemble de règles assez large pour couvrir toutes les possibilités. Cependant, avec l'apprentissage automatique, vous pouvez utiliser une liste de messages électroniques marqués comme spam ou non pour entraîner un modèle d'apprentissage automatique. Le modèle apprendrait quels motifs de mots, longueurs et autres indicateurs sont de bons prédicteurs des messages indésirables. Lorsque vous présentez au modèle un message électronique qu'il n'a jamais vu auparavant, il prédira s'il s'agit de spam ou non.

Tom Mitchell, un pionnier de l'apprentissage automatique, a proposé cette définition : « On dit qu'un programme informatique apprend de l'expérience E par rapport à une certaine classe de tâches T et une mesure de performance P, si sa performance sur les tâches de T, mesurée par P, s'améliore avec l'expérience E. » (Mitchell, Tom. 1997. Machine Learning. McGraw Hill. p. 2.)

Si vous appliquez ce concept au spam, les lettres E, T et P représentent :

* E – Les messages électroniques qui indiquent s'ils sont des spams ou non
* T – La tâche d'identifier les spams
* P – La probabilité que le message électronique non vu soit un spam.

Ne vous inquiétez pas si ces idées sont difficiles à comprendre.

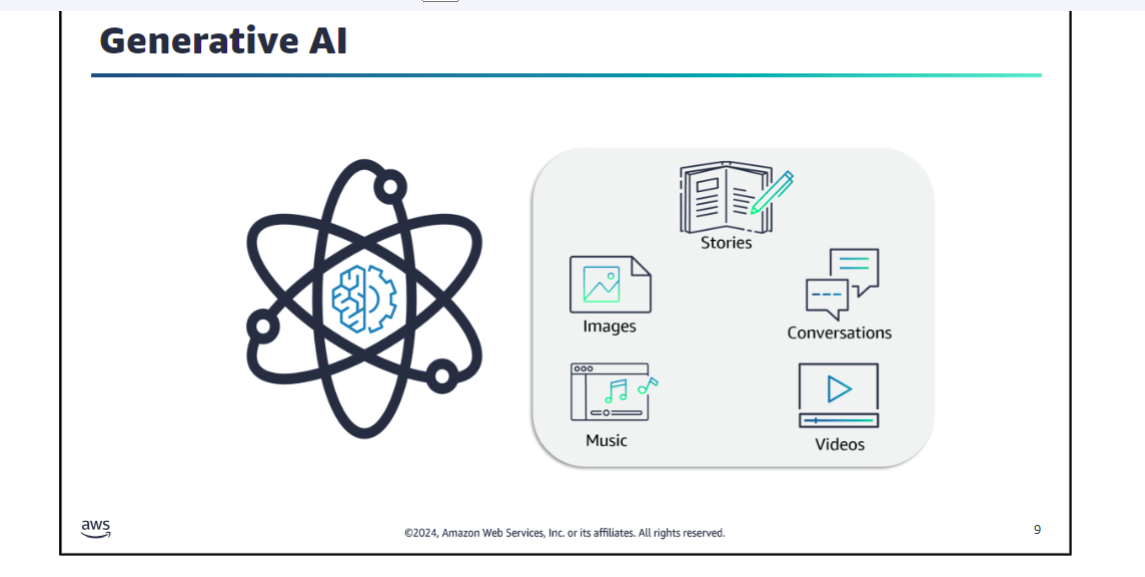


L'apprentissage profond (deep learning) représente une avancée significative dans les capacités de l'IA et de l'apprentissage automatique (ML). La théorie derrière l'apprentissage profond s'inspire du fonctionnement du cerveau humain. Un réseau de neurones artificiel (ANN) est inspiré des neurones biologiques du cerveau, bien que sa mise en œuvre soit différente.

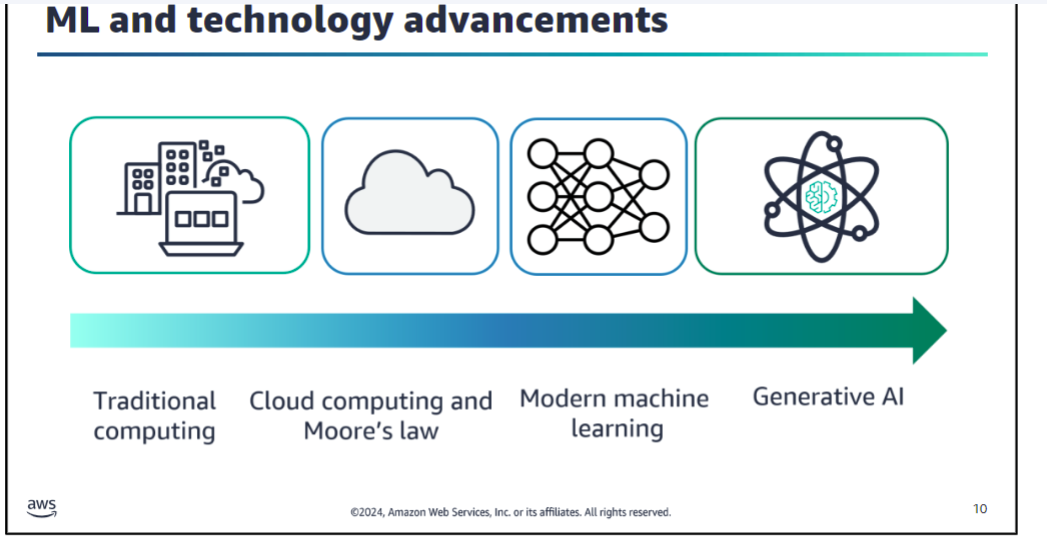
Les neurones artificiels ont une ou plusieurs entrées et une seule sortie. Ces neurones "s'activent" (ou déclenchent leurs sorties) en fonction d'une transformation des entrées. Un réseau de neurones est composé de couches de ces neurones artificiels, avec des connexions entre les couches. En général, un **réseau possède** des **couches d'entrée, de sortie et des couches cachées**. La sortie d'un seul neurone est connectée aux entrées de tous les neurones de la couche suivante. **Le réseau est** ensuite **utilisé pour résoudre** **un problème**. La couche d'entrée est alimentée par les données d'entraînement. Les neurones s'activent à travers les couches jusqu'à ce qu'une réponse soit présentée dans la couche de sortie. La précision du résultat est alors mesurée. Si la sortie ne répond pas à votre seuil de précision, l'entraînement est répété, mais avec des ajustements mineurs des poids des connexions entre les neurones. Ce processus se répète. À chaque itération, il renforce les connexions qui mènent au succès et diminue celles qui conduisent à l'échec.

Comme vous le verrez dans ce cours, **les praticiens de l'apprentissage automatique passent beaucoup de temps à optimiser les modèles de ML. Ils choisissent les meilleures caractéristiques de données pour l'entraînement et sélectionnent les modèles avec les meilleurs résultats. En revanche, les praticiens de l'apprentissage profond passent presque aucun temps sur ces tâches. Ils se concentrent plutôt sur la modélisation des données avec différentes architectures de réseaux de neurones artificiels.**

Bien que la théorie de l'apprentissage profond existe depuis des décennies, le matériel nécessaire pour résoudre des problèmes d'apprentissage profond n'était pas accessible jusqu'à récemment. Maintenant qu'il est disponible, vous pouvez utiliser l'apprentissage profond pour aborder des problèmes plus complexes que ceux que vous avez traités auparavant.

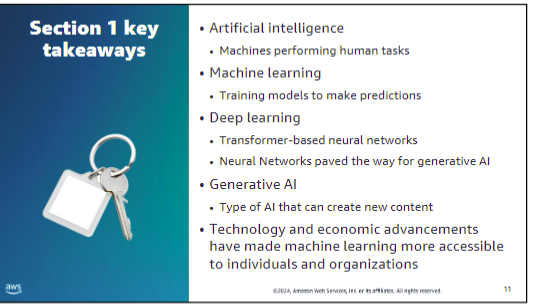


**L'intelligence artificielle générative** est un type d'IA capable de créer de nouveaux contenus et idées, y compris des conversations, des histoires, des images, des vidéos et de la musique. L'IA générative est alimentée par des modèles d'apprentissage automatique très volumineux, couramment appelés **modèles fondamentaux** (**FMs**), qui sont préentraînés sur d'énormes quantités de données. Contrairement aux modèles d'apprentissage automatique traditionnels qui sont orientés vers une seule tâche et nécessitent un entraînement, un seul modèle fondamental peut être utilisé pour différentes tâches, comme la génération de texte ou la synthèse, sans besoin d'entraînement supplémentaire. **Vous pouvez utiliser l'IA générative pour améliorer l'expérience client grâce à des fonctionnalités telles que les chatbots, les assistants virtuels, les centres de contact intelligents, la personnalisation et la modération de contenu.**



**Progrès(advancements)**

L'apprentissage automatique grand public est un phénomène relativement récent. Le milieu des années 2000 a marqué le début de progrès rapides dans l'apprentissage automatique et l'apprentissage profond, en partie **grâce à la loi de Moore** et à **l'essor du cloud computing**. Le résultat est un accès plus facile à des capacités de calcul et de stockage plus importantes, plus rapides et moins chères. Vous pouvez maintenant louer de la puissance de calcul pour quelques heures pour quelques centimes, comparé aux énormes investissements qui étaient nécessaires pour acheter et exploiter des clusters de calcul à grande échelle.En 2012, l'utilisation des réseaux de neurones a commencé dans le cadre du défi ImageNet Large Scale Visual Recognition, une compétition d'apprentissage automatique pour la *reconnaissance d'images*. Le taux de précision a bondi à environ 82 % et n'a cessé d'augmenter depuis. Il a dépassé les performances humaines en 2015. En 2017, une percée dans l'architecture des réseaux de neurones appelée *architecture de transformateur* a contribué à rendre possible l'IA générative moderne.



*Quelques points clés à retenir de cette section du module sont :*

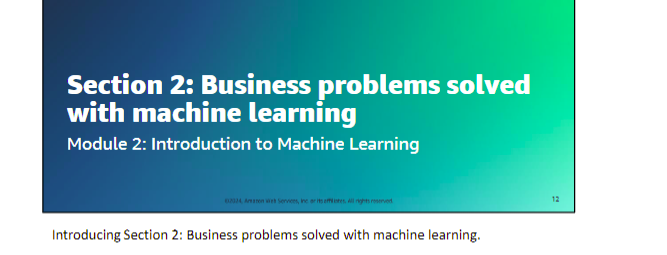
*•* *L'intelligence artificielle* ***(IA****)* est le vaste domaine de la construction de machines pour effectuer des tâches humaines.

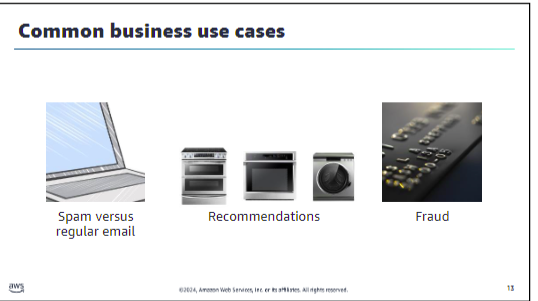
• *L'apprentissage automatique* (**ML)** **est un sous-ensemble de l'IA**. Il se concentre sur l'utilisation de **données** pour **entraîner** des **modèles** d'apprentissage automatique afin que ces modèles puissent faire des **prédictions**.

*• L'apprentissage profond* ( **Deep Learning**)**est une technique inspirée de la biologie humaine.** Il utilise des **couches de neurones** pour **construire** des **réseaux** qui **résolvent** des **problèmes**.

**• L'IA générative** **est un type d'IA** qui **utilise** des **modèles pré-entraînés appelés FM (Foundational Models** **ou Modèles** **Fondamentaux)** pour **générer** de **nouvelles idées de contenu**.

*•* ***Les progrès dans la technologie, le cloud computing***et le ***développement d'algorithmes*** *ont conduit à une augmentation des* *capacités et des applications de l'apprentissage automatique, y compris* ***l'IA générative*.**

****

****

L'apprentissage automatique est utilisé tout au long de la vie numérique d'une personne.

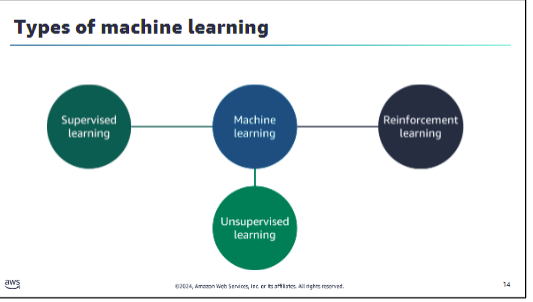
Voici quelques exemples :

• Spam - Votre *filtre anti-spam* est le ***résultat*** *d'un programme* *d'apprentissage automatique* qui a *été* ***entraîné*** *avec des exemples de spam et de messages électroniques normaux.*

• Recommandations - Sur la base des livres que vous lisez ou des produits que vous achetez, les programmes d'apprentissage automatique **prédisent** d'autres livres ou produits qui pourraient vous intéresser. Là encore, le programme d'apprentissage automatique a été **entraîné** avec des données sur les habitudes et les achats d'autres lecteurs.

• Fraude par carte de crédit - De même, le programme d'apprentissage automatique a été **entraîné** sur des exemples de transactions qui se sont avérées frauduleuses, ainsi que sur des transactions légitimes.

Il existe de nombreux autres exemples, notamment **la détection** **faciale** dans les applications de médias sociaux pour regrouper vos photos, la **détection de tumeurs cérébrales** dans les scanners cérébraux, ou **la recherche d'anomalies** dans les radiographies.



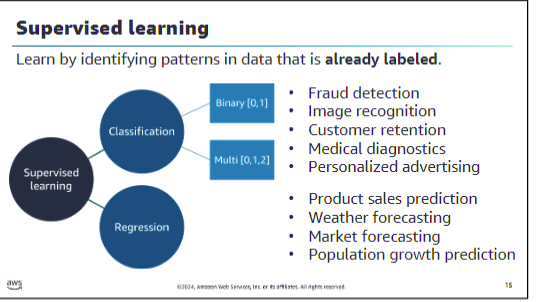
**L'apprentissage automatique comprend *trois types principaux*.**

***Le premier type* est l'apprentissage supervisé**, où un modèle utilise des entrées et des sorties connues pour généraliser les sorties futures.

***Le deuxième type* est l'apprentissage non supervisé**, où le modèle ne connaît ni les entrées ni les sorties - il trouve des modèles dans les données sans aide.

***Le troisième type* est l'apprentissage par renforcement**, où le modèle interagit avec son environnement et apprend à prendre des actions qui maximisent les **récompenses**.

**Il est important de connaître les différents types d'apprentissage** automatique car *le type peut vous guider vers la sélection d'algorithmes qui ont du sens pour résoudre votre problème d'entreprise.*

**

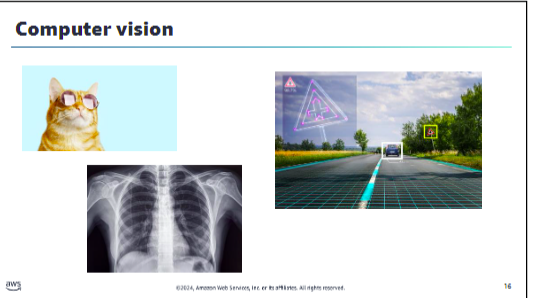
**L'apprentissage supervisé** *est un type populaire d'apprentissage automatique car* ***il est largement applicable***. On l'appelle apprentissage supervisé car vous avez **besoin d'un superviseur** - **un enseignant** - qui peut **montrer les bonnes réponses au modèle**. Comme tout étudiant, un algorithme supervisé apprend par l'exemple. Il a besoin d'un enseignant qui utilise des données d'entraînement pour l'aider à déterminer les modèles et les relations entre les entrées et les sorties : "*Ici, dans cette image, il y a une voiture. Voici une voiture dans une autre image."* Le modèle est entraîné sur ces données étiquetées pour identifier avec précision où se trouve une voiture dans une nouvelle image qu'il n'a jamais vue auparavant.

Cependant, vous pouvez avoir différents **types de problèmes** **dans** **l'apprentissage supervisé**. Ces problèmes peuvent être largement classés en **deux catégories** : **la classification** et **la régression**.

Les **problèmes de classification ont deux types**. Le premier type est considéré comme un ***problème de classification binaire***. Rappelez-vous l'exemple précédent sur l'identification des transactions frauduleuses. La variable cible dans cet exemple est limitée à deux options : *frauduleuse ou non frauduleuse*. Cet exemple est un problème de classification binaire car vous classez une observation dans l'une ***des deux catégories***.

Il existe également des problèmes de **classification multiclasse.** Ces problèmes d'apprentissage automatique *classent une* *observation dans l'une des trois catégories ou plus*. Supposons que vous ayez un modèle d'apprentissage automatique qui prédit pourquoi un client appelle votre magasin. Son but est de réduire le nombre de transferts nécessaires avant d'amener le client au bon service d'assistance clientèle. Dans ce cas, les différents services d'assistance clientèle représentent les diverses variables cibles potentielles - *qui pourraient être de nombreux départements différents.*

Il existe également **des problèmes de** **régression**. Dans un problème de régression, vous ne faites plus correspondre une entrée à un nombre défini de catégories. Au lieu de cela, *vous faites* *correspondre des entrées à une valeur continue*, *comme un nombre* *entier*. Un ***exemple de problème de régression*** en apprentissage automatique ***est la prédiction du prix de l'action d'une entreprise***. Par exemple, un algorithme basé sur la régression pourrait prédire que demain, le cours de l'action d'une entreprise pourrait passer de 113 dollars par action à 127 dollars par action.

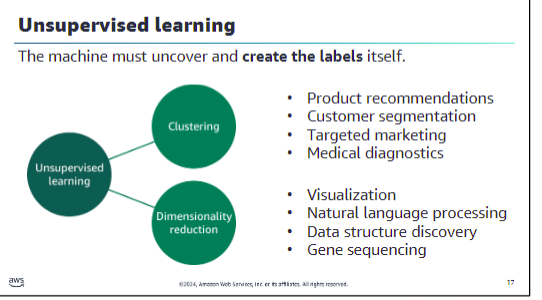


**Vision**

**La vision par ordinateur** (CV) *est un vaste domaine* qui *consiste* principalement *en des problèmes de classification*. S'agit-il d'un chat ou d'un chien ? Cette radiographie montre-t-elle une tumeur ? ***La vision par ordinateur permet*** *aux machines* ***d'identifier*** *des personnes, des lieux et des objets dans des images avec une* ***précision égale (ou supérieure)*** *à celle des humains, et avec une plus* ***grande rapidité et efficacité****.* *Elle est souvent* ***construite*** *avec des* ***modèles*** *d'apprentissage* ***profond****.*

**La CV** *automatise l'extraction, l'analyse, la classification et la* *compréhension d'informations* utiles à partir d'une seule image ou d'une séquence d'images. Les données **d'image** **peuvent** **prendre** de **nombreuses formes**, telles que des images uniques, des séquences vidéo, des vues de plusieurs caméras ou des données tridimensionnelles.

Vous en apprendrez davantage sur la vision par ordinateur plus tard dans ce cours.

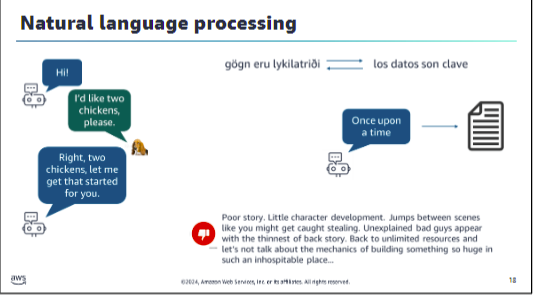


Vous allez maintenant ***découvrir*** **l'apprentissage automatique non supervisé.**

Parfois, les données sont tout ce dont vous disposez - **aucun superviseur** n'est disponible. Dans l'apprentissage non supervisé, **les étiquettes ne sont pas fournies** (comme c'est le cas avec l'apprentissage supervisé) *car vous ne connaissez pas toutes les variables et les modèles.* Dans ces cas, **la machine doit découvrir et** **créer elle-même les étiquettes**. Ces modèles utilisent les données qui leur sont présentées pour détecter les propriétés émergentes de l'ensemble des données, puis construire des modèles.

**Une sous-catégorie courante de l'apprentissage non supervisé** **s'appelle le clustering**. *Ce type d'algorithme regroupe les données en différents clusters basés sur des caractéristiques similaires pour mieux comprendre les attributs d'un cluster spécifique*. Par exemple, supposons que vous travaillez pour une organisation marketing qui soutient des entreprises fournissant des ressources de cloud computing à leurs clients. En analysant les habitudes d'achat des clients, les algorithmes non supervisés peuvent identifier des groupes de clients qui catégorisent (ou décrivent) une entreprise particulière comme étant petite, moyenne ou grande. Les entreprises peuvent utiliser ces informations pour prendre des décisions marketing stratégiques éclairées en fonction de leur niveau (petit, moyen ou grand) afin de mieux cibler les besoins de leurs clients en ressources de cloud computing. Dans cette situation, ***le clustering pourrait vous aider*** à réaliser que vous devez proposer une stratégie marketing différente pour des entreprises de tailles différentes.

**L'avantage des algorithmes non supervisés est** *qu'ils vous* *permettent de voir des modèles dans les données dont vous n'aviez pas conscience auparavant.* Un exemple pourrait être l'existence de deux types principaux de clients.



**Traitement (processing)**

***Le traitement du langage naturel (NLP)*** *est un autre domaine de l'apprentissage automatique dont l'utilisation est croissante.* Par exemple, Amazon Alexa (ou tout autre assistant vocal) utilise le NLP pour essayer de répondre à vos questions. **Le NLP concerne la** **parole, mais aussi le texte écrit**. Le NLP est **utilisé** **dans** de nombreuses **applications** telles que :

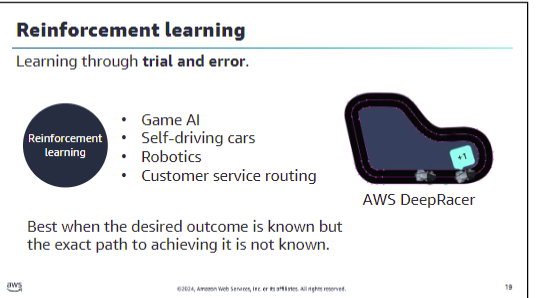
• **Chatbots** ou bots de centre d'appels - Systèmes automatisés pour obtenir votre solde bancaire ou commander de la nourriture dans un restaurant.

• **Outils de traduction** - Conversion de texte entre langues, ou applications pouvant traduire des menus en temps réel.

• **Traductions voix-texte** - Conversion de mots parlés en texte. Peut être utilisé pour générer des sous-titres automatiques.

• **Analyse de sentiment** - Permet d'analyser le sentiment des commentaires dans les critiques de produits, de musique et de films. Ces sentiments peuvent être utilisés pour donner une note d'audience au film.

V*ous en apprendrez davantage sur le NLP dans un module ultérieur*

azs<

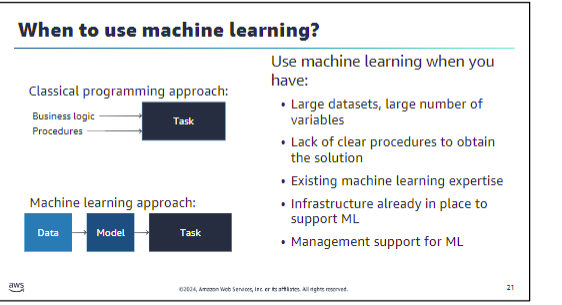
=

Un autre type d'apprentissage automatique qui **gagne en popularité** récemment **est l'apprentissage par renforcement**. Contrairement aux autres types d'apprentissage automatique, l'apprentissage par renforcement ***améliore continuellement son modèle en exploitant les retours d'expérience des itérations*** ***précédentes***. *Dans l'apprentissage par renforcement*, *un agent* *apprend continuellement, par essais et erreurs, en interagissant avec un* ***environnement****.* *L'apprentissage par renforcement* **est utile** ***lorsque la récompense d'un résultat souhaité est connue,* mais *que le chemin pour y parvenir ne l'est pas***. *Découvrir ce chemin nécessite de nombreux essais et erreurs.*

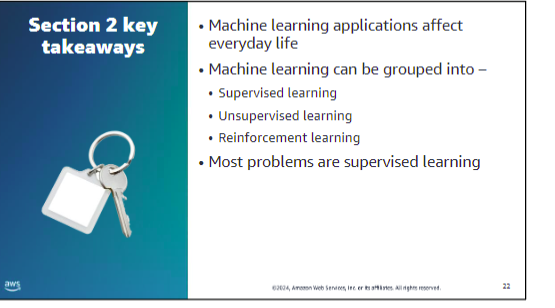
***Prenons l'exemple*** *d'AWS DeepRacer*. Dans le simulateur AWS DeepRacer, l'agent est la voiture virtuelle, et l'environnement est une piste de course virtuelle. Les actions sont les commandes d'accélération et de direction de la voiture. **L'objectif est de terminer la piste de course le plus rapidement possible sans dévier de la piste.** La voiture doit apprendre le comportement de conduite souhaité pour atteindre l'objectif de terminer la piste. Vous utilisez des récompenses pour inciter votre modèle à apprendre le comportement de conduite souhaité. Dans l'apprentissage par renforcement, **ce qui motive l'apprentissage est appelé l'agent**. Dans ce cas, l'agent est la voiture AWS DeepRacer. L'environnement est le lieu où l'agent apprend, qui - dans cet exemple - est la piste de course balisée. Lorsque l'agent fait quelque chose dans l'environnement qui provoque une réponse, comme franchir une limite qu'il ne devrait pas franchir, on appelle cela une action. Cette réponse est appelée récompense ou pénalité selon que l'agent a fait quelque chose qui devrait être renforcé ou découragé dans le modèle. Au fur et à mesure que l'agent se déplace dans l'environnement, ses actions devraient continuer à recevoir plus de récompenses et moins de pénalités, jusqu'à ce qu'il atteigne le résultat commercial souhaité.



Les véhicules autonomes rassemblent plusieurs algorithmes et modèles d'apprentissage automatique et d'apprentissage profond pour résoudre le problème de la conduite du point A au point B. Une tâche majeure consiste à détecter en continu l'environnement et à prévoir les changements. **Cette tâche implique la détection d'objets, qui est la localisation et la prédiction du mouvement de l'objet détecté.** Les résultats de ces découvertes servent d'entrées à d'autres systèmes qui prennent des décisions sur ce qu'il faut faire avec les différentes commandes du véhicule. Certains cas d'utilisation impliquent des véhicules autonomes qui nécessitent des réponses en temps réel à l'environnement. Par exemple, si un piéton auparavant caché sort de derrière un obstacle, les freins du véhicule doivent être appliqués immédiatement. De telles actions ne peuvent pas avoir de latence ou de marge d'erreur.



Tous les problèmes ne doivent pas être résolus par l'apprentissage automatique, et parfois la programmation de base fonctionne tout aussi bien. Lorsque vous explorez une solution potentielle d'apprentissage automatique, recherchez des éléments tels que l'existence de grands ensembles de données et un grand nombre de variables. Vous pourriez également être incertain de la logique commerciale ou des procédures nécessaires pour obtenir une réponse ou accomplir une tâche. Dans de tels cas, l'apprentissage automatique est souvent le meilleur choix. Les systèmes d'apprentissage automatique peuvent être complexes. L'infrastructure de support, le soutien de la direction et l'expertise doivent être en place pour aider le projet à réussir.



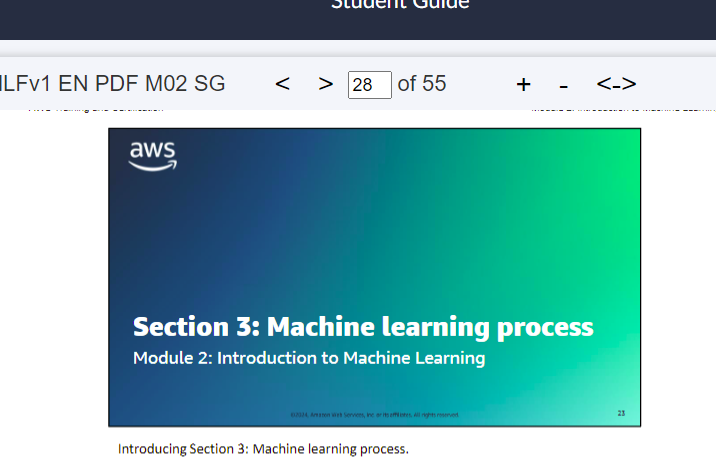
*Quelques points clés à retenir de cette section du module :*

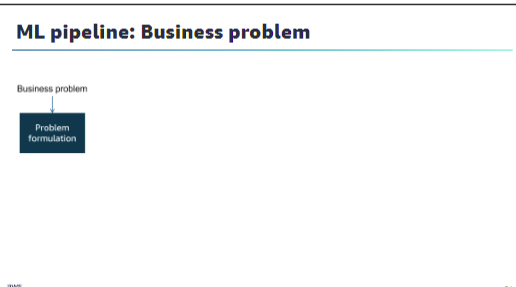
• Les applications d'apprentissage automatique font déjà partie de votre vie quotidienne.

• **Les problèmes d'apprentissage automatique** peuvent être **regroupés** en :

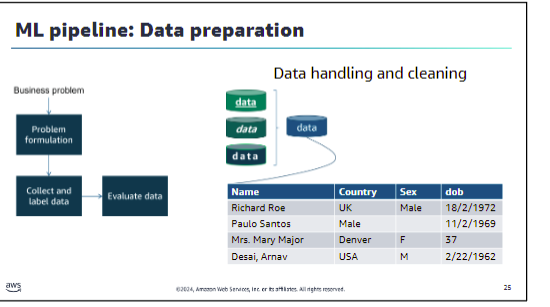
* **Apprentissage supervisé** : Vous disposez de données d'entraînement pour lesquelles vous connaissez la réponse.
* **Apprentissage non supervisé** : Vous avez des données, mais vous recherchez des informations au sein de ces données.
* **Apprentissage par renforcement** : Le modèle apprend en se basant sur l'expérience et le retour d'information.

• ***La plupart des problèmes d'entreprise relèvent de l'apprentissage supervisé.***

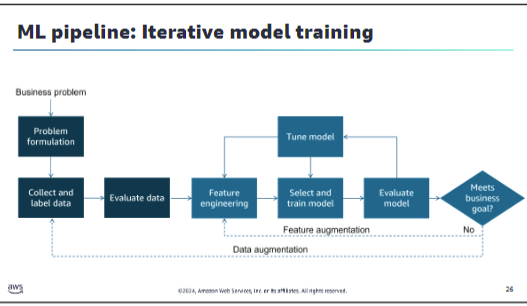
******

******

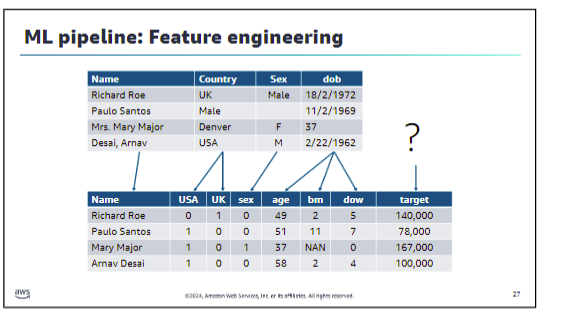
Pour commencer, vous devriez toujours partir du problème commercial que vous ou votre équipe pensez pouvoir améliorer grâce au ML (Machine Learning ou apprentissage automatique). À partir de là, vous voulez procéder à une formulation du problème. Cette phase implique, en partie, d'articuler votre problème commercial et de le convertir en un problème de ML.



Après avoir formulé le problème, vous passez à la phase de préparation et de prétraitement des données. Dans cette phase, vous extrairez des données d'une ou plusieurs sources. Ces sources de données peuvent présenter des différences dans les données ou les types qui doivent être conciliées pour former une vue cohérente unique de vos données. Vous devez visualiser vos données et utiliser des statistiques pour déterminer si les données sont cohérentes et peuvent être utilisées pour l'apprentissage automatique. Vous examinerez certaines sources de données plus tard dans le cours.Dans l'exemple de données, vous avez quatre colonnes contenant des données assemblées à partir de trois sources. Les sources avaient des façons légèrement différentes de représenter les données, et les résultats sont présentés dans le tableau. *Dans les problèmes de ML,* **les colonnes** représentent des **caractéristiques,** et **les lignes** représentent des **instances**. Vous pouvez voir certains problèmes avec les données dans certaines instances. Dans certains cas, *vous avez besoin d'un expert* en la matière ou d'un expert fonctionnel pour comprendre l'authenticité des données. Par exemple, 11/2/1969 peut être le 2 novembre ou le 11 février. Quelqu'un qui possède ou gère le pool de données serait en mesure de clarifier ce que signifient ces données. "Male" peut probablement être attribué à un problème d'importation où les cellules ont changé de position, mais cela peut aussi signifier un lieu : Malé, la capitale de la République des Maldives. Parfois, cette identification d'erreur n'est pas aussi simple et nécessitera un expert en la matière. Vous en apprendrez davantage sur le rôle d'un expert plus tard dans le cours. Vous pouvez avoir un effet substantiel sur le succès d'un projet de ML en ayant des données cohérentes et correctes.



Après avoir préparé vos données et assuré leur exactitude et leur cohérence, il est temps de former votre modèle. À ce stade, le processus devient itératif et fluide. Vous allez probablement rencontrer plusieurs passages d'ingénierie des caractéristiques, de formation, d'évaluation et d'ajustement avant de trouver un modèle qui puisse répondre à vos objectifs commerciaux.



**L'ingénierie des caractéristiques(feature engineering)** **est le** **processus de sélection ou de création des caractéristiques** *que vous utiliserez pour entraîner votre modèle*. **Les caractéristiques** **sont** les **colonnes de données** que vous avez dans votre ensemble de données. **L'objectif** du modèle **est** d'essayer **d'estimer** **correctement** la **valeur cible pour de nouvelles données**. **L'algorithme** de ML utilise les caractéristiques pour **prédire la cible**. Dans cet exemple, la donnée cible est le nombre moyen de pas effectués en une semaine. La sélection des bonnes caractéristiques peut impliquer l'ajout, la suppression ou le calcul de nouvelles caractéristiques. Vous pourriez rencontrer ces changements lors de l'ingénierie des caractéristiques :

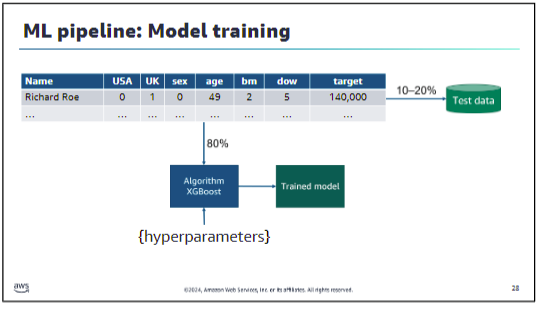
• ***Nettoyage des données de la colonne Nom*** – Vous pourriez vouloir rendre le format cohérent, ce qui peut être utile plus tard dans le modèle. Ou vous pourriez changer le format uniquement pour des raisons esthétiques. Selon le problème que vous essayez de résoudre avec ces données, vous pourriez même ne pas avoir besoin du Nom.

• ***Conversion du pays en un ensemble de colonnes numériques ou binaires*** – Si cette base de données était une base de données traditionnelle, vous pourriez vouloir déplacer le Pays vers une table de consultation et ensuite y faire référence. La plupart des algorithmes de ML veulent que les données d'une instance soient dans une seule ligne. De plus, les algorithmes de ML ont besoin de données numériques pour traiter. Vous pourriez envisager de transformer le texte du pays en code ISO du pays. Cependant, le modèle pourrait interpréter la valeur numérique comme ayant une signification, donc UK (44) serait plus significatif que, par exemple, USA (01). Dans ce cas, diviser les données en plusieurs colonnes est une bonne pratique. Cette pratique est connue sous le nom ***d'encodage catégoriel***, dont vous apprendrez plus tard dans ce cours.

• ***Conversion des données de sexe en binaire*** – Si vous convertissez la valeur textuelle en une valeur numérique, avec 0 ou 1 représentant masculin ou féminin, le modèle peut l'utiliser plus facilement.

• ***Division de la date de naissance (DDN) en parties composantes*** – Extraire l'âge, le mois de naissance (mn) et le jour de la semaine (js) pourrait être approprié, selon le problème que vous essayez de résoudre. L'âge affecte-t-il la variable cible ? Qu'en est-il du jour de la semaine où ils sont nés ?Ne vous inquiétez pas si ces informations semblent compliquées.

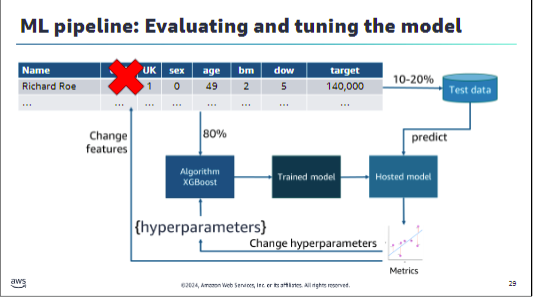
*Vous apprendrez plus de détails sur l'ingénierie des caractéristiques plus tard dans ce cours*

**

Après avoir nettoyé vos données et identifié leurs caractéristiques, il est temps d'entraîner un modèle. Vous n'utilisez pas toutes les données pour entraîner votre modèle. Au lieu de cela, vous devriez garder certaines données de côté afin d'avoir des données pour tester. Typiquement, **vous utilisez environ 80 pour cent des données** **pour l'entraînement, et vous gardez le reste pour les** **tests**. Vous utilisez les données d'entraînement **pour former un** **modèle**. Dans le diagramme, le modèle utilisé est l'algorithme XGBoost.

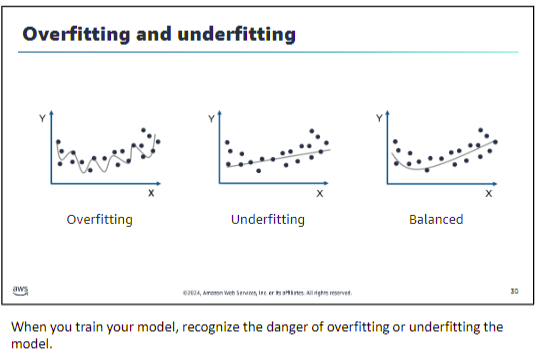
Le modèle lui-même a certains paramètres que vous pouvez définir, et ils modifient le fonctionnement de l'algorithme. Ces **paramètres** sont connus sous le nom **d'hyperparamètres.**

Le résultat du travail d'entraînement est un modèle entraîné.



Avec le modèle entraîné, vous pouvez utiliser une partie des données de test pour voir comment le modèle fonctionne. Vous pouvez prendre une instance que le modèle n'a pas vue et l'utiliser pour effectuer une prédiction. Comme vous connaissez déjà la cible dans vos données de test, vous pouvez comparer les deux valeurs. À partir de ces comparaisons, vous pouvez calculer des métriques, qui vous donnent des données sur les performances du modèle. Vous modifiez ensuite les données, les caractéristiques ou les hyperparamètres du modèle jusqu'à ce que vous trouviez le modèle qui donne les meilleurs résultats.

**\*\*Pour l'accessibilité** : Quatre-vingts pour cent des données de test sont envoyées à l'algorithme XGBoost, puis à un modèle entraîné, puis à un modèle hébergé. Dix à vingt pour cent des données de test sont utilisées pour effectuer une prédiction. Les métriques sont calculées et les hyperparamètres sont modifiés en fonction de la prédiction du modèle par rapport aux données réelles. **Fin de la description**.



*Lorsque vous entraînez votre modèle, reconnaissez le danger de surajustement ou de sous-ajustement du modèle.*

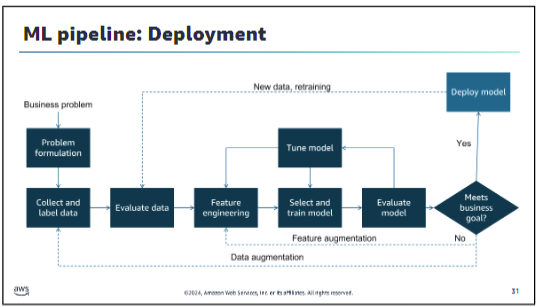
**Votre modèle est en surajustement (overfitting)** de vos données d'entraînement lorsqu'il fonctionne bien sur les données d'entraînement, mais qu'il ne fonctionne pas bien sur les données d'évaluation. Le modèle mémorise les données qu'il a vues, et il ne peut pas généraliser à des exemples non vus - ce qui provoque le surajustement.

**Votre modèle est en sous-ajustement (underfitting)** des données d'entraînement lorsque le modèle fonctionne mal sur les données d'entraînement. Le modèle ne peut pas capturer la relation entre les exemples d'entrée (souvent appelés X) et les valeurs cibles (souvent appelées Y). Cette incapacité provoque le sous-ajustement.

**Comprendre l'ajustement du modèle** *est important pour trouver la cause principale d'une faible précision du modèle.* Ce concept vous guide pour prendre des mesures correctives afin de déterminer si un modèle prédictif est en sous-ajustement ou en surajustement des données d'entraînement.

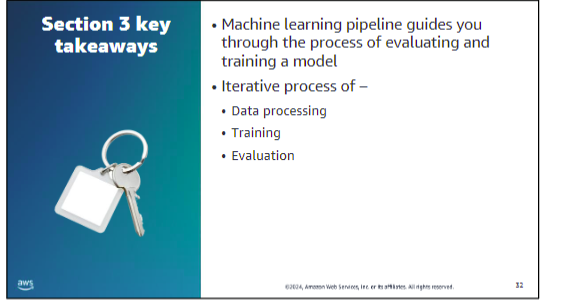
Ces étapes impliquent d'examiner l'erreur de prédiction sur les données d'entraînement et les données d'évaluation.

*Vous apprendrez les étapes que vous pouvez prendre pour éviter ce problème plus tard dans ce cours.*

**

Après avoir réentraîné le modèle et être satisfait des résultats, votre modèle est déployé pour fournir les meilleures prédictions possibles.

Plus tard dans ce cours, vous serez guidé à travers ces différentes phases et acquerrez une expérience pratique avec chacune d'entre elles. Connaître le processus est également utile lorsque vous utilisez les services gérés que vous explorerez. Cependant, les services ML d'Amazon font la plupart du travail difficile pour vous.

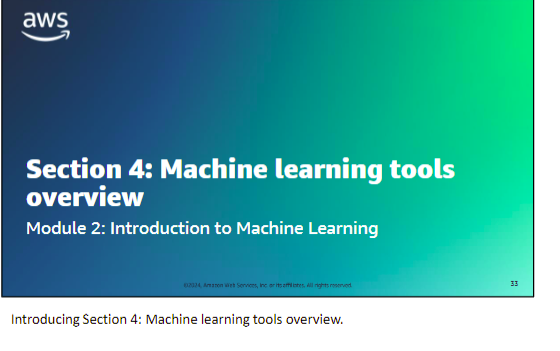


*Quelques points clés à retenir de cette section du module sont :*

• Le processus de pipeline d'apprentissage automatique peut vous guider tout au long du processus d'entraînement et d'évaluation d'un modèle.

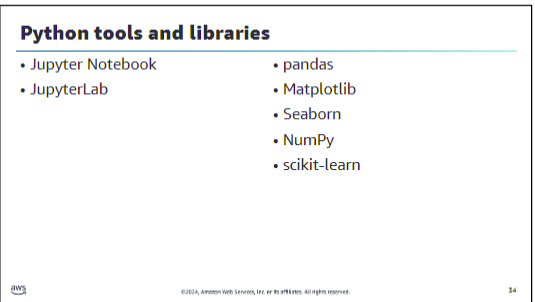
• Le **processus itératif** peut être **divisé en trois grandes étapes** :

* **Traitement des données**
* **Entraînement du modèle**
* **Évaluation du modèle**

****

**Aperçu des outils d'apprentissage automatique.**

(Prezentare generală a instrumentelor de învățare automata)



Vous allez maintenant **découvrir certains des outils** que vous utiliserez **pour l'apprentissage automatique**.

Note : Cette liste n'est pas une liste complète des outils de ML disponibles aujourd'hui, et vous n'en utiliserez que quelques-uns dans ce cours.

**Jupyter Notebook** *est une application web open-source qui vous permet de créer et de partager des documents contenant du code en direct, des équations, des visualisations et du texte narratif.* Ses utilisations incluent le nettoyage et la transformation des données, la simulation numérique, la modélisation statistique, la visualisation des données, l'apprentissage automatique, et bien plus encore.

**JupyterLab** *est un environnement de développement interactif basé sur le web pour les notebooks Jupyter, le code et les données.* ***JupyterLab est flexible****.* Vous pouvez configurer et organiser l'interface utilisateur pour prendre en charge une gamme de flux de travail en science des données, en calcul scientifique et en apprentissage automatique. **JupyterLab est également extensible et modulaire.** Vous pouvez écrire des plugins qui ajoutent de nouveaux composants et s'intègrent aux existants.

Plus tard dans ce cours, vous utiliserez Amazon SageMaker, qui héberge les notebooks Jupyter et JupyterLab.

**pandas** *est une* ***bibliothèque Python open-source***. Elle est *utilisée pour la manipulation et l'analyse de données*. Elle représente les données dans un tableau similaire à une *feuille* *de calcul*. *Ce tableau est connu sous le nom* ***de DataFrame pandas.***

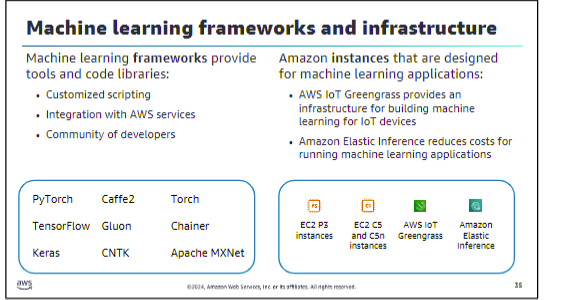
**Matplotlib** *est une* ***bibliothèque pour créer des visualisations scientifiques statiques, animées et interactives en Python.*** Vous l'utiliserez pour générer des graphiques de vos données plus tard dans ce cours.

**Seaborn** *est une autre* ***bibliothèque de visualisation de données pour Python.*** Elle est construite sur matplotlib et fournit une interface de haut niveau pour *dessiner des graphiques statistiques informatifs.*

**NumPy es**t l'un des **packages de calcul scientifique fondamentaux en Python.** Il contient des fonctions pour les objets de tableaux N-dimensionnels et des fonctions mathématiques utiles telles que l'algèbre linéaire, la transformée de Fourier et des capacités de génération de nombres aléatoires.

**scikit-learn** est une **bibliothèque d'apprentissage automatique open-source qui prend en charge l'apprentissage supervisé et non supervisé.** Elle fournit également divers outils pour l'ajustement de modèles, le prétraitement des données, la sélection et l'évaluation de modèles, ainsi que de nombreuses autres utilitaires. **scikit-learn est construit sur NumPy, SciPy et matplotlib, et c'est un bon package pour explorer l'apprentissage automatique.** Bien que vous ne l'utilisiez que pour emprunter quelques fonctions dans un module ultérieur, vous pourrez explorer ce package plus en détail après avoir terminé ce cours.

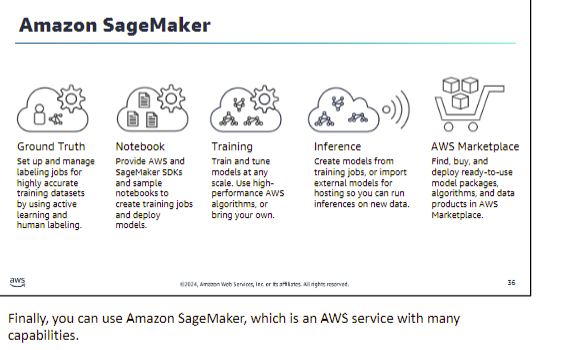
**un framework fournit une base solide et des outils pour que les développeurs puissent se concentrer sur la logique métier plutôt que de tout recréer à partir de zéro.**



*En passant des bibliothèques et packages individuels à des outils plus complets, vous pouvez utiliser des frameworks prêts pour la production.*

Vous avez brièvement appris à propos de scikit-learn, qui est une bibliothèque qui fonctionne bien pour l'apprentissage. Vous pouvez également utiliser d'autres bibliothèques, telles que TensorFlow pour l'apprentissage automatique et Keras pour l'apprentissage profond. Tous ces frameworks sont pris en charge sur AWS et peuvent être utilisés à partir d'Amazon SageMaker.

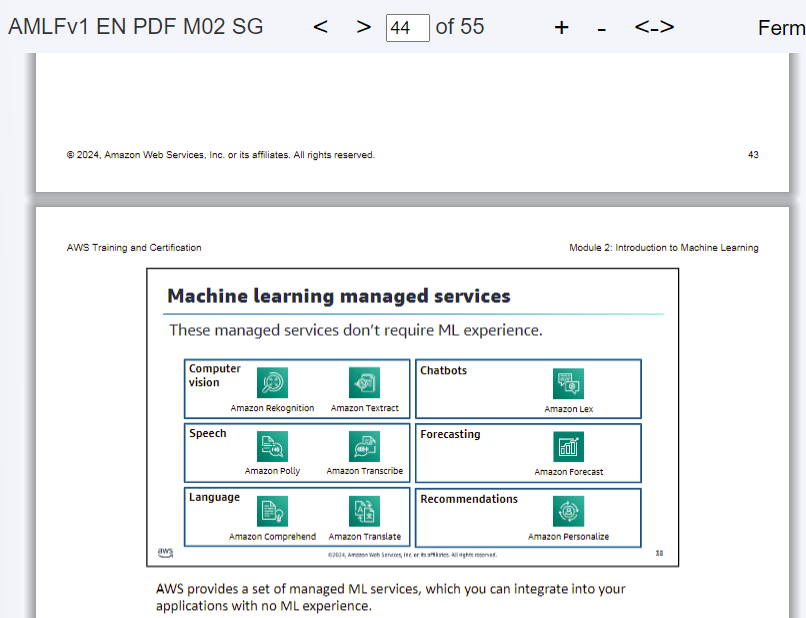
Amazon fournit également des instances de calcul optimisées pour l'apprentissage automatique, à la fois dans le cloud et en périphérie. Les instances de calcul sont optimisées pour l'apprentissage et l'inférence, et des Amazon Machine Images (AMIs) prépackagées contiennent de nombreux frameworks populaires.



*Enfin, vous pouvez utiliser Amazon SageMaker, qui est un service AWS avec de nombreuses capacités.*

Amazon SageMaker peut déployer des instances d'apprentissage automatique qui exécutent des notebooks Jupyter et JupyterLab. Amazon SageMaker gère le déploiement de ces ressources de calcul, donc vous devez vous connecter à l'environnement Jupyter. Amazon SageMaker fournit également des outils pour l'étiquetage des données, l'entraînement des modèles et l'hébergement des modèles entraînés. AWS Marketplace propose également une sélection de packages de modèles prêts à l'emploi et d'algorithmes provenant de développeurs en apprentissage automatique.





AWS fournit un ensemble de services ML gérés, que vous pouvez intégrer dans vos applications sans expérience en ML.

• **Vision par ordinateur** - **Amazon Rekognition** fournit la reconnaissance d'objets et faciale pour les images et les vidéos. **Amazon Textract** peut extraire du texte des images.

• **Parole** - **Amazon Polly** peut lire du texte à voix haute, tandis **qu'Amazon Transcribe** convertit l'audio parlé en texte.

• **Langage** - **Amazon Comprehend** utilise le traitement du langage naturel **(NLP)** pour trouver des insights et des relations dans le texte**. Amazon Translate** peut traduire du texte dans différentes langues.

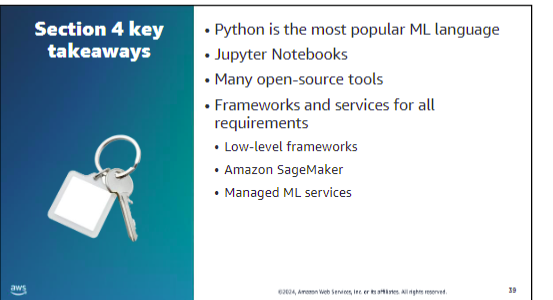
• **Chatbots** - **Amazon Lex** est un service qui vous aide à construire des applications conversationnelles interactives utilisant la voix ou le texte.

• **Prévisions** - **Amazon Forecast** utilise l'apprentissage automatique pour combiner des données de séries temporelles avec des variables supplémentaires pour construire des prévisions.

**• Recommandations - Amazon Personalize** est un autre service d'apprentissage automatique qui peut vous aider à créer des recommandations personnalisées individuelles pour les clients.

Ces services gérés sont formés sur de nombreux aspects du domaine du problème, et vous fournissez vos données spécifiques pour commencer le processus.

Vous examinerez beaucoup de ces services gérés dans la seconde moitié de ce cours, après avoir appris à faire les choses par vous-même.



*Quelques points clés à retenir de cette section du module* :

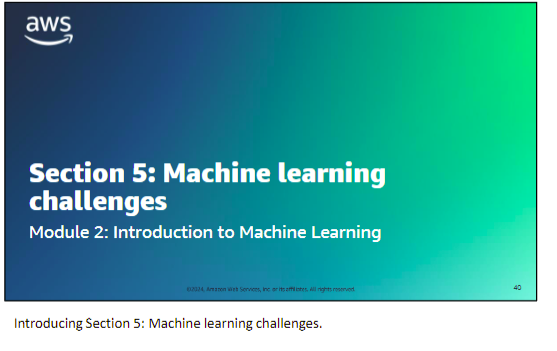
• Python est le langage le plus populaire pour effectuer des tâches d'apprentissage automatique.

• Les Jupyter Notebooks vous fournissent un environnement de développement hébergé sur navigateur web pour l'apprentissage automatique.

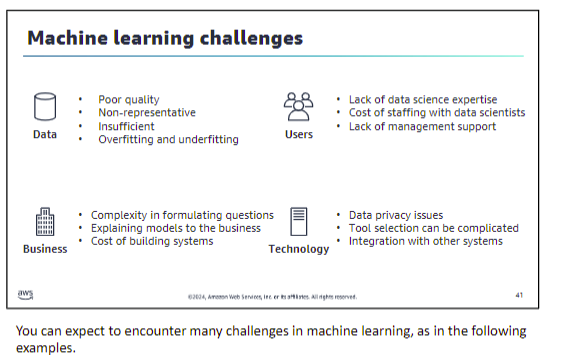
• Un grand nombre d'outils open-source existent, et seuls quelques-uns des principaux sont présentés dans ce cours.

• Selon vos besoins, vous pouvez commencer avec les frameworks de bas niveau et utiliser votre propre solution.

Vous pourriez utiliser des outils comme Amazon SageMaker pour vous aider avec des tâches plus importantes, ou adapter l'un des services ML gérés pour votre domaine de problème spécifique.



**Introduction à la Section 5 : Les défis de l'apprentissage automatique**

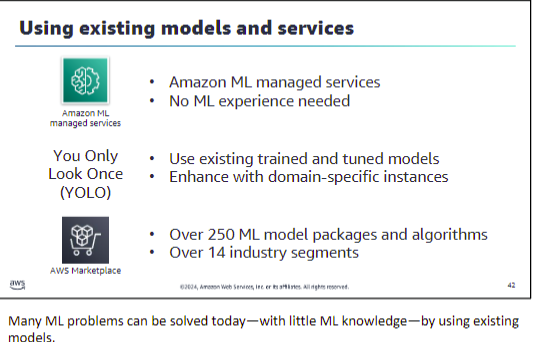
****

Vous pouvez vous attendre à rencontrer de nombreux défis en apprentissage automatique, comme dans les exemples suivants.

**Données (data) :**  
• Il existe une grande quantité de données de mauvaise qualité et incohérentes dans le monde. Une grande partie de votre travail consiste à obtenir de bonnes données.  
• Les données représentent-elles bien le problème ? Si vous essayez de détecter la fraude par carte de crédit, avez-vous des exemples pour l'entraînement ?  
• Avez-vous suffisamment de données ? Dans la plupart des cas, plus vous avez de données, mieux c'est.  
• Votre modèle est-il en surajustement ou en sous-ajustement **Utilisateurs (Users):**  
• Avez-vous une expérience en science des données ?  
• Est-il rentable de constituer une équipe de data scientists ?  
• La direction soutient-elle l'utilisation du ML ?

**Entreprise (Business):**  
• Les problèmes sont-ils trop complexes pour être formulés en un problème de ML ?  
• Le modèle résultant peut-il être expliqué à l'entreprise ? S'il ne peut pas l'être, il pourrait ne pas être adopté.  
**•** Quel est le coût de construction, de mise à jour et d'exploitation d'une solution ML ?

**Technologie :**  
• L'unité commerciale a-t-elle accès aux données dont vous avez besoin ? Les données peuvent-elles être sécurisées pour répondre aux exigences réglementaires ?  
• Quels outils et frameworks prévoyez-vous d'utiliser ?  
• Comment cette solution s'intègre-t-elle aux autres systèmes ?

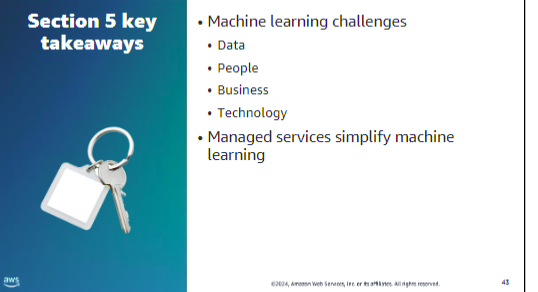


*De nombreux problèmes de ML peuvent être résolus aujourd'hui - avec peu de connaissances en ML - en utilisant des modèles existants.*

Vous avez déjà appris à propos des services ML gérés d'AWS. Vous avez maintenant besoin de quelques compétences de développeur pour appeler des API, et vous pouvez ajouter des capacités sophistiquées d'apprentissage automatique à vos applications.

Vous pouvez utiliser d'autres modèles pré-construits, ou adapter le populaire modèle de vision You Only Look Once (YOLO).

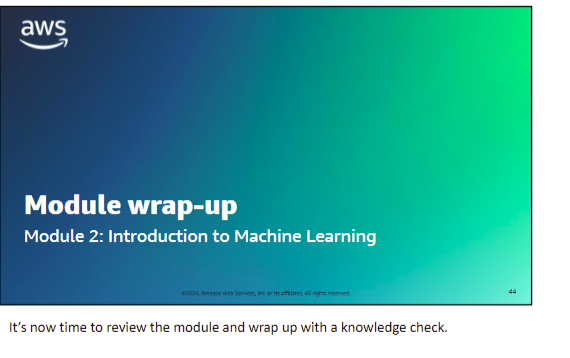
Peut-être que vous ne voulez pas créer votre propre algorithme, mais plutôt acheter des modèles et des services développés par des éditeurs de logiciels indépendants. En plus des scénarios précédents, vous pouvez également utiliser AWS Marketplace pour trouver des solutions tierces.

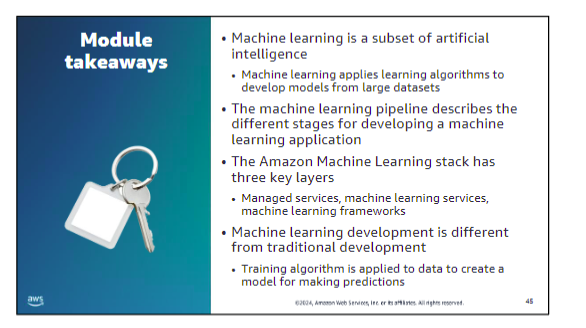


*Les points clés à retenir de cette section du module sont :*

• Vous ferez face à de nombreux défis en apprentissage automatique. Les plus grands problèmes que vous influencez directement sont liés aux données, mais vous aurez également à faire face à des défis liés aux personnes, à l'entreprise et à la technologie.

• Envisagez d'utiliser des services gérés pour vous aider à résoudre les problèmes d'apprentissage automatique.





*Quelques points clés à retenir de cette section du module* :

• **L'apprentissage automatique** *est un sous-ensemble de l'intelligence artificielle.*

• *L'apprentissage automatique applique des algorithmes d'apprentissage pour développer des modèles à partir de grands ensembles de données.*

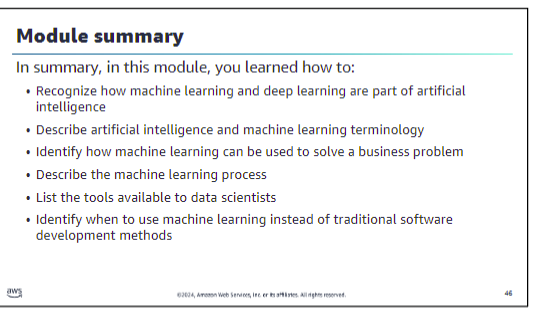
• **Le pipeline d'apprentissage automatique** décrit les différentes étapes du développement d'une application d'apprentissage automatique.

• **La pile d'apprentissage automatique** comporte **trois couches** **clés :**

* Services API
* *Un service API (Application Programming Interface) est une interface qui permet à différentes applications logicielles de communiquer entre elles et d'échanger des données.*
* Services d'apprentissage automatique
* Frameworks d'apprentissage automatique

• **Le développement en apprentissage automatique** est différent du développement traditionnel.

• Les algorithmes d'entraînement sont appliqués aux données pour créer un modèle permettant de faire des prédictions.



*En résumé, dans ce module, vous avez appris à :*

• Reconnaître comment l'apprentissage automatique et l'apprentissage profond font partie de l'intelligence artificielle

• Décrire la terminologie de l'intelligence artificielle et de l'apprentissage automatique

• Identifier comment l'apprentissage automatique peut être utilisé pour résoudre un problème d'entreprise

• Décrire le processus d'apprentissage automatique

• Énumérer les outils disponibles pour les data scientists

• Identifier quand utiliser l'apprentissage automatique plutôt que les méthodes traditionnelles de développement logiciel

