Module I: introduction à l'IA appliquée au secteur agro-alimentaire

Co-funded by
the European Union

Résumé

L'intelligence artificielle (IA) est une technologie transformatrice qui modifie rapidement le mode de fonctionnement des entreprises. Il devient de plus en plus important pour les gestionnaires et les PDG de comprendre l'IA, ses capacités et ses limites. En outre, l'intelligence artificielle et son sous-domaine, l'apprentissage automatique, ont récemment occupé le devant de la scène dans l'industrie (4.0) avec l'IA générative et les clones numériques. Aujourd'hui déjà, un nombre croissant de processus industriels sont améliorés par l'utilisation de l'IA. Si l'utilisation de l'IA peut changer la donne pour le secteur agroalimentaire, il est important de comprendre **ce qu'elle peut faire et ce qu'elle ne peut pas faire**. L'intelligence artificielle utilise de grandes quantités de données pour « former » des algorithmes afin qu'ils puissent « faire » des prédictions. Les prédictions peuvent porter sur la classification, le sens d'une phrase (compréhension du langage naturel), le contenu d'une image ou d'une vidéo, ou encore le mot suivant dans une phrase qui a un sens (comme pour l'IA générative, telle que les grands modèles de langage). Dans ce matériel de formation, nous fournirons une introduction à l'IA pour les personnes non techniques et expliquerons comment l'IA peut contribuer à la transition circulaire. Nous présenterons également les conditions préalables à l'intelligence artificielle, à savoir la gestion des données. Comme les données doivent être collectées et mises à la disposition de l'algorithme à former, nous verrons comment l'architecture des données, l'ingénierie des données et la gouvernance des données sont les fondements de l'intelligence artificielle. Enfin, nous explorerons quelques cas d'utilisation industrielle de l'IA, en particulier dans les domaines du traitement du langage naturel (NLP), des prévisions et du traitement d'images.

Objectifs d’apprentissage

Cette formation vise à donner des connaissances et des compétences concrètes aux professionnels de l'agroalimentaire par le biais d'un cours en ligne sur l'IA. Cette formation a les objectifs suivants :

- Comprendre ce qu'est l'IA, ce qu'elle peut et ne peut pas faire, et donner des définitions aux termes utilisés dans le domaine de l'IA.
- Comprendre quelles sont les conditions préalables à l'intelligence artificielle.
- Obtenir une bonne connaissance de certaines des technologies d'IA les plus pertinentes à utiliser dans le domaine du secteur agroalimentaire.

Session 1: Introduction à l'IA

Aujourd'hui, la quantité de données générées, tant par les humains que par les machines, dépasse de loin la capacité des humains à absorber, interpréter et prendre des décisions complexes sur la base de ces données. L'intelligence artificielle est à la base de tout apprentissage informatique et représente l'avenir de toute prise de décision complexe. Les organisations capables d'exploiter les données peuvent se développer plus rapidement, acquérir des clients à moindre coût et mieux les fidéliser.

Comprendre les termes utilisés dans l'IA

Commençons par définir trois concepts interconnectés qui diffèrent par leur objectif et leur portée : l'intelligence artificielle, l'apprentissage automatique et l'apprentissage profond.

L'**intelligence artificielle (IA)** est un concept plus large que l'apprentissage automatique qui implique le développement de machines intelligentes capables d'effectuer des tâches qui requièrent normalement l'intelligence humaine, telles que la perception, le raisonnement, l'apprentissage et la prise de décision.

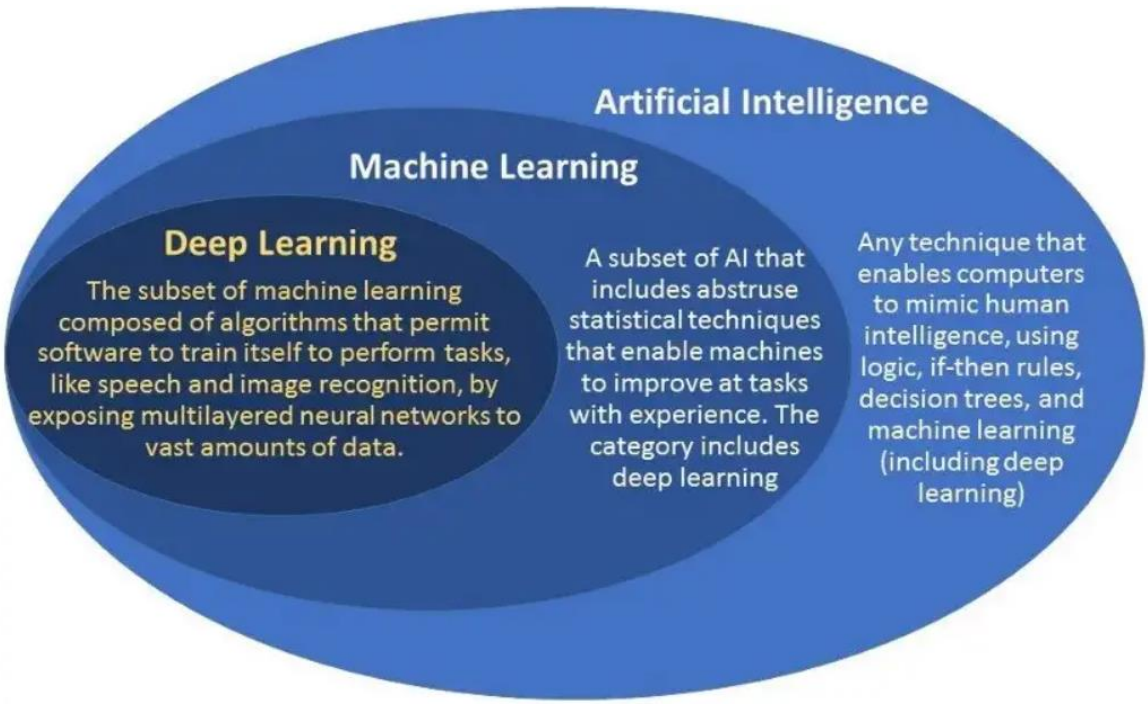


Figure 1. Relations entre IA / ML and DL[1]

[1] Source: https://www.unite.ai/machine-learning-vs-deep-learning-key-differences/

Le **machine learning** est un sous-domaine de l'intelligence artificielle qui se concentre sur le développement d'algorithmes et de modèles capables d'apprendre à partir de données et d'améliorer leurs performances au fil du temps. Il implique l'utilisation de techniques statistiques et informatiques pour analyser de grands ensembles de données et identifier des modèles ou des relations entre les variables. L'objectif de l'apprentissage automatique est de construire des modèles prédictifs capables de faire des prédictions ou de prendre des décisions précises sur la base de nouvelles données.

Le deep learning est un sous-ensemble de l'apprentissage automatique qui imite les processus cognitifs du cerveau humain en apprenant à partir de la manière dont les données sont structurées, plutôt qu'à partir d'un algorithme programmé pour faire une chose spécifique.

Le deep learning utilise des réseaux neuronaux artificiels, c'est-à-dire un ensemble d'algorithmes utilisés pour trouver les relations entre de grandes quantités de données par le biais d'un processus qui imite le fonctionnement du cerveau humain.

Un neurone dans un réseau neuronal est une fonction mathématique (telle qu'une fonction d'activation) dont le travail consiste à rassembler et à classer les informations en fonction d'une structure particulière.

Les réseaux neuronaux sont utiles pour la reconnaissance des formes, comme la reconnaissance de l'écriture manuscrite, la reconnaissance des visages, la reconnaissance vocale, la traduction de textes, le diagnostic médical et les solutions pour les grandes quantités de données.

La science des données est un autre terme largement utilisé dans le domaine de l'IA. Délimitons les frontières de ces deux termes.

La science des données implique l'utilisation de diverses techniques statistiques et informatiques pour extraire des idées et des connaissances à partir de grandes quantités de données.

En résumé, la science des données se concentre sur l'extraction d'informations à partir des ,

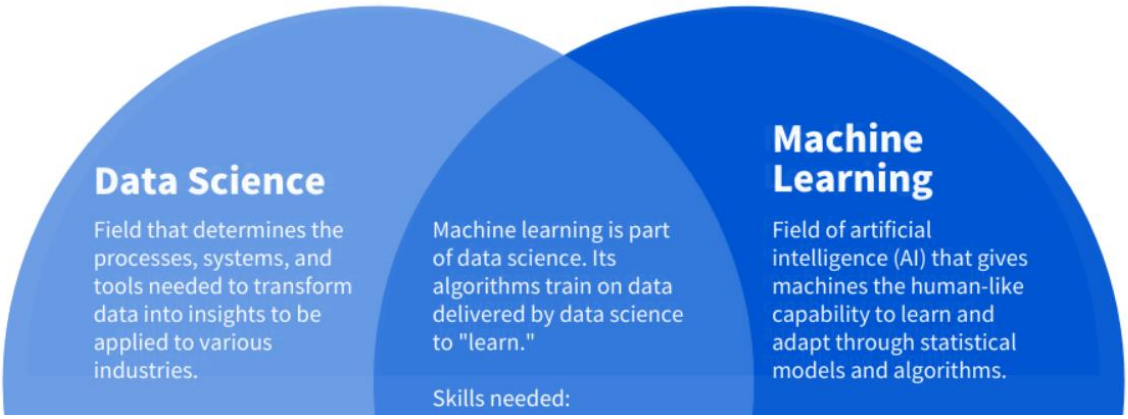


Figure 2. Relations entre Data Science and ML[2]

[2] Source: https://www.coursera.org/articles/data-science-vs-machine-learning

La Programmation manuelle désigne tout programme créé manuellement (système à base de règles) qui utilise des données d'entrée et s'exécute sur un ordinateur pour produire un résultat. Dans un système à base de règles, les règles (ensemble de règles « si » et « alors ») sont élaborées par des experts du domaine modélisé afin d'automatiser les processus de prise de décision et de faire des déductions logiques et des inférences à propos d'un problème spécifique.

Dans la programmation du **machine learning**, les données d'entrée et de sortie sont transmises à un algorithme pour créer un programme. Il s'agit d'une technique qui permet aux ordinateurs d'apprendre sans être programmés.

Par exemple, si vous voulez prédire qui paiera ses factures en retard, identifiez les données d'entrée (clients et factures) et les données de sortie (payer en retard ou non), et laissez l'apprentissage automatique utiliser ces données pour créer votre modèle, c'est-à-dire pour déterminer si un nouveau client paiera en retard ou non. Ce programme est appelé modèle prédictif.

Les systèmes basés sur des règles peuvent être utilisés dans un grand nombre de domaines. Par exemple, un système à base de règles peut être utilisé pour diagnostiquer des conditions médicales, détecter des transactions frauduleuses ou contrôler le fonctionnement d'une installation de fabrication.

L'un des avantages des systèmes basés sur des règles est qu'ils sont transparents et explicables. Les règles sont explicites et peuvent être examinées par des experts pour s'assurer qu'elles sont exactes et appropriées. Toutefois, les systèmes basés sur des règles peuvent être limités par la qualité et l'exhaustivité des règles élaborées. En outre, ils peuvent avoir du mal à gérer l'ambiguïté ou les situations nuancées dans lesquelles plusieurs règles peuvent s'appliquer.

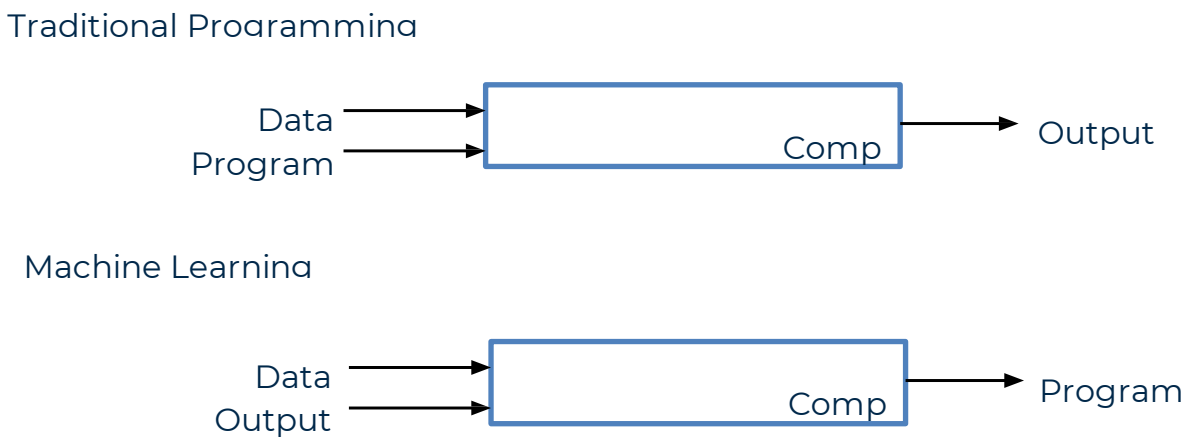


Figure 3. Différence entre la programmation manuelle et le machine learning

Terminologie IA[3]

Précision : La précision est un système de notation dans la classification binaire (c'est-à-dire la détermination si une réponse ou un résultat est correct ou non) et se calcule comme suit : (Vrais positifs + Vrais négatifs) / (Vrais positifs + Vrais négatifs + Faux positifs + Faux négatifs).
Algorithme : Ensemble de règles qu'une machine peut suivre pour apprendre à effectuer une tâche.
Intelligence artificielle : Il s'agit du concept général de machines agissant d'une manière qui simule ou imite l'intelligence humaine. L'IA peut présenter diverses caractéristiques, telles qu'une communication ou une prise de décision semblables à celles de l'homme.
Biais: Hypothèses formulées par un modèle qui simplifient le processus d'apprentissage de la tâche qui lui est assignée. La plupart des modèles d'apprentissage automatique supervisé sont plus performants lorsque le biais est faible, car ces hypothèses peuvent avoir une incidence négative sur les résultats.
Big data: Les ensembles de données qui sont trop volumineux ou trop complexes pour être utilisés par les applications traditionnelles de traitement des données.
Corpus: Un grand ensemble de données écrites ou orales qui peut être utilisé pour former une machine à effectuer des tâches linguistiques.
Data mining: Le processus d'analyse des ensembles de données afin de découvrir de nouveaux modèles susceptibles d'améliorer le modèle.
Sciences des données : S'inspirant des statistiques, de l'informatique et des sciences de l'information, ce domaine interdisciplinaire vise à utiliser une variété de méthodes, de processus et de systèmes scientifiques pour résoudre des problèmes impliquant des données.
Dataset: A collection of related data points, usually with a uniform order and tags.
Deep learning: Une collection de points de données liés, généralement avec un ordre et des étiquettes uniformes.
Caractéristique: : Dans le domaine de l'intelligence artificielle et de l'apprentissage automatique, une caractéristique est une propriété ou une caractéristique individuelle mesurable d'un phénomène. Le choix de caractéristiques informatives, discriminantes et indépendantes est un élément crucial des algorithmes efficaces de reconnaissance des formes, de classification et de régression. Elle est comparable à une colonne dans un tableau de données.
Ingénierie des fonctionnalités : Il s'agit du processus consistant à utiliser les connaissances du domaine pour sélectionner et transformer les variables les plus pertinentes à partir des données brutes lors de la création d'un modèle prédictif utilisant l'apprentissage automatique ou la modélisation statistique.
IA Générative: Algorithmes (tels que ChatGPT) qui peuvent être utilisés pour créer de nouveaux contenus, y compris de l'audio, du code, des images, du texte, des simulations et des vidéos.
Hyperparamètre: Parfois utilisé de manière interchangeable avec le paramètre, bien que les termes présentent quelques différences subtiles. Les hyperparamètres sont des valeurs qui

[3] https://www.telusinternational.com/insights/ai-data/article/50-beginner-ai-terms-you-should-know

affectent la manière dont votre modèle apprend. Ils sont généralement définis manuellement en dehors du modèle**.**

Inférence : Ce terme fait souvent référence à la prédiction qu'un algorithme entraîné (parfois appelé « modèle ») fournit pour de nouvelles données.

Étiquette : Une partie des données d'apprentissage qui identifie la sortie souhaitée pour ce morceau de données.

Machine learning (ML): Sous-ensemble de l'IA axé sur le développement d'algorithmes qui aideront les machines à apprendre et à changer en réponse à de nouvelles données, sans l'aide d'un être humain.

MLOps or ML Ops: Paradigme qui vise à déployer et à maintenir des modèles d'apprentissage automatique en production de manière fiable et efficace.

Modèle: Terme général désignant le produit de l'apprentissage de l'IA, créé par l'exécution d'un algorithme de ML sur des données d'apprentissage.

Réseau neuronal : Également appelé réseau neuronal, un réseau neuronal est un système informatique conçu pour fonctionner comme le cerveau humain. Bien que les chercheurs travaillent encore à la création d'un modèle machine du cerveau humain, les réseaux neuronaux existants peuvent effectuer de nombreuses tâches liées à la parole, à la vision et à la stratégie des jeux de société.

Génération de langage naturel (NLG) : Il s'agit du processus par lequel une machine transforme des données structurées en texte ou en discours compréhensibles par l'homme.**Traitement du langage naturel (TLN)**: Terme générique désignant la capacité d'une machine à effectuer des tâches conversationnelles, telles que la reconnaissance de ce qui lui est dit, la compréhension du sens voulu et la réponse intelligible.

Compréhension du langage naturel (NLU): Sous-ensemble du traitement du langage naturel, la compréhension du langage naturel consiste à aider les machines à reconnaître le sens voulu du langage, en tenant compte de ses nuances subtiles et de toute erreur grammaticale.

Ajustement excessif (overfitting) : Un terme important de l'IA, l'overfitting, est un symptôme de la formation à l'apprentissage automatique dans lequel un algorithme n'est capable de travailler ou d'identifier que des exemples spécifiques présents dans les données d'apprentissage. Un modèle fonctionnel devrait être capable d'utiliser les tendances générales des données pour travailler sur de nouveaux exemples.

Paramètre: Une variable à l'intérieur du modèle qui l'aide à faire des prédictions. La valeur d'un paramètre peut être estimée à l'aide de données et n'est généralement pas fixée par la personne qui exécute le modèle.

Reconnaissance des formes: La distinction entre la reconnaissance des formes et l'apprentissage automatique est souvent floue, mais ce domaine s'intéresse essentiellement à la recherche de tendances et de modèles dans les données.

Analyse prédictive: Type d'analyse combinant l'exploration de données et l'apprentissage automatique, conçu pour prévoir ce qui se passera dans un délai donné sur la base de données historiques et de tendances.

Python: Un langage de programmation populaire utilisé pour la programmation générale.

Données de test: Les données non étiquetées utilisées pour vérifier qu'un modèle d'apprentissage automatique est capable d'effectuer la tâche qui lui a été assignée.

Entraînement: Il s'agit du processus consistant à alimenter un algorithme avec des données historiques afin qu'il « apprenne » à partir de celles-ci et soit en mesure de fournir des prédictions pour de nouvelles données.

Données de formation: Il s'agit de toutes les données utilisées au cours du processus de formation d'un algorithme d'apprentissage automatique, ainsi que de l'ensemble de données spécifique utilisé pour la formation plutôt que pour le test.

Apprentissage par transfert: Cette méthode d'apprentissage consiste à passer du temps à apprendre à une machine à effectuer une tâche connexe, puis à lui permettre de revenir à son travail initial avec une précision accrue. Un exemple potentiel de cette méthode est de prendre un modèle qui analyse le sentiment dans les commentaires sur les produits et de lui demander d'analyser les tweets pendant une semaine.

Données de validation: Structurées comme des données de formation avec une entrée et des étiquettes, ces données sont utilisées pour tester un modèle récemment formé par rapport à de nouvelles données et pour analyser les performances, en mettant l'accent sur la vérification de l'ajustement excessif.

Variance: La quantité de changements dans la fonction prévue d'un modèle d'apprentissage automatique au cours de sa formation. Bien qu'ils soient flexibles, les modèles à forte variance sont sujets à l'overfitting et à une faible précision prédictive, car ils dépendent de leurs données d'apprentissage.

Variation: Également appelées requêtes ou énonciations, elles fonctionnent en tandem avec les intentions pour le traitement du langage naturel. La variation est ce qu'une personne pourrait dire pour atteindre un certain objectif.

Ce que l'IA peut faire et ne pas faire - Principales classifications du ML et cas d'usages [4]

Si l'intelligence artificielle peut faire beaucoup de choses, comme la maintenance prédictive, comprendre le texte et la parole, identifier et compter des objets dans des images et des vidéos, ou même générer de nouvelles images et de nouveaux textes, elle ne peut pas encore raisonner comme un être humain. Cela signifie qu'il y a des tâches et des compétences qui ne peuvent pas être numérisées ou automatisées.

À l'heure actuelle, l'IA est surtout prédictive et peu prescriptive. Cela signifie que les organisations peuvent prédire les résultats potentiels aujourd'hui, mais qu'il y a encore un effort important à faire pour essayer de contrôler ces résultats.

Cependant, sur la base des résultats de l'analyse prédictive, l'analyse prescriptive vise à comprendre quelles variables peuvent être manipulées pour atteindre le résultat souhaité et comment. Elle exige des data scientists une compréhension approfondie des causes et des effets des variables afin qu'elles puissent être affinées pour obtenir les résultats souhaités par les organisations.

L'apprentissage automatique peut être utilisé pour prédire les résultats commerciaux (fournir des réponses aux questions commerciales) dans toutes les situations où l'on dispose de

[4] https://www.ibm.com/topics/artificial-intelligence
https://www.oracle.com/be/artificial-intelligence/what-is-ai/
https://www.coe.int/en/web/artificial-intelligence/what-is-ai#

données d'entrée et de sortie historiques. Le cas d'utilisation le plus répandu de l'apprentissage automatique est celui des moteurs de recommandation pour le commerce électronique, qui segmentent les clients sur la base de leurs données d'utilisateur et de leurs modèles comportementaux (tels que l'historique des achats et de la navigation, les goûts ou les critiques) et les ciblent avec des suggestions personnalisées de produits et de contenu.

Au cours des dernières années, l'apprentissage automatique nous a permis d'obtenir des voitures autonomes, la reconnaissance d'images et de la parole, des modèles de prévision de la demande, des recherches utiles sur le web et diverses autres applications.

En fonction des types de données disponibles, les professionnels de l'information sélectionnent des types d'apprentissage automatique (algorithmes) pour ce qu'ils veulent prédire à partir des données :

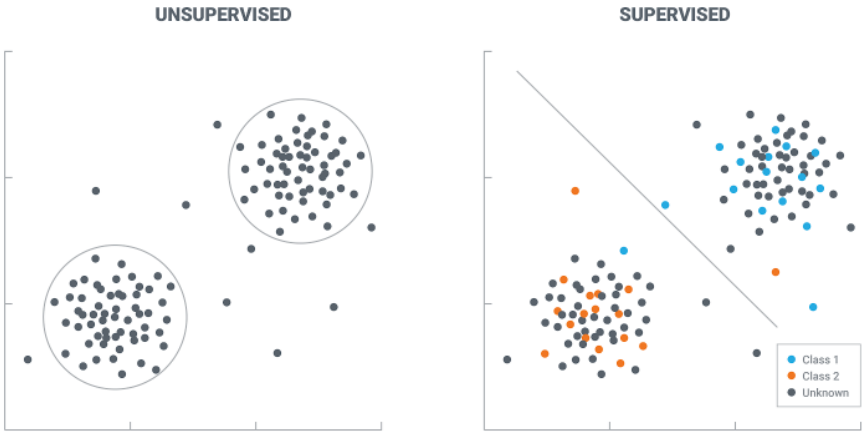


Figure 4. Clustering contre classification

● Apprentissage non supervisé : Ce type d'apprentissage comprend des algorithmes qui s'entraînent (font des déductions) sur des données non étiquetées en analysant des ensembles de données pour en tirer des corrélations significatives. Par exemple, l'une des méthodes est l'analyse de cluster qui utilise l'analyse exploratoire des données pour obtenir des modèles ou des groupes cachés ou groupés dans les ensembles de données. Elle est utilisée pour le clustering, c'est-à-dire pour trouver des aspects communs au sein des clients afin d'appliquer la segmentation de la clientèle et la recommandation de produits.

● Apprentissage supervisé : Dans ce type d'apprentissage, les experts en données fournissent des données de formation étiquetées aux algorithmes et définissent des variables aux algorithmes pour accéder et trouver des corrélations. L'entrée et la sortie de l'algorithme sont toutes deux définies. Ce type de ML est utilisé pour les problèmes de classification, par exemple, le tri des aliments ou la sécurité alimentaire et le contrôle de la qualité.

● Apprentissage par renforcement : Apprendre à une machine informatique à remplir un processus à plusieurs étapes pour lequel il existe des règles clairement définies. Ici, les programmeurs conçoivent un algorithme pour effectuer une tâche et lui donnent des signaux

positifs et négatifs pour agir au fur et à mesure que l'algorithme s'exécute pour accomplir la tâche. Finalement, l'algorithme détermine de lui-même la meilleure action à entreprendre.

IA et Economie circulaire

L'IA peut jouer un rôle crucial dans la transition circulaire en permettant aux entreprises d'optimiser l'utilisation des ressources, de réduire les déchets et d'améliorer la durabilité.
Par exemple, l'IA peut être utilisée pour optimiser l'utilisation de l'eau, des pesticides et des engrais dans l'agriculture. En analysant les données provenant de capteurs et d'autres sources, les algorithmes d'IA peuvent identifier les possibilités d'économie d'eau et de ressources et recommander des changements dans les pratiques agricoles. L'IA peut également être utilisée pour surveiller et gérer les flux de déchets, identifier les possibilités de réduction des déchets et soutenir le développement de chaînes d'approvisionnement circulaires.

Session 2: Pré requis 1ère partie

Comme indiqué précédemment, l'IA s'appuie fortement sur les données existantes. Pour que les données puissent être utilisées dans les algorithmes d'IA, il existe de nombreuses conditions préalables liées aux données.

Cycle de vie d'un projet de Machine learning

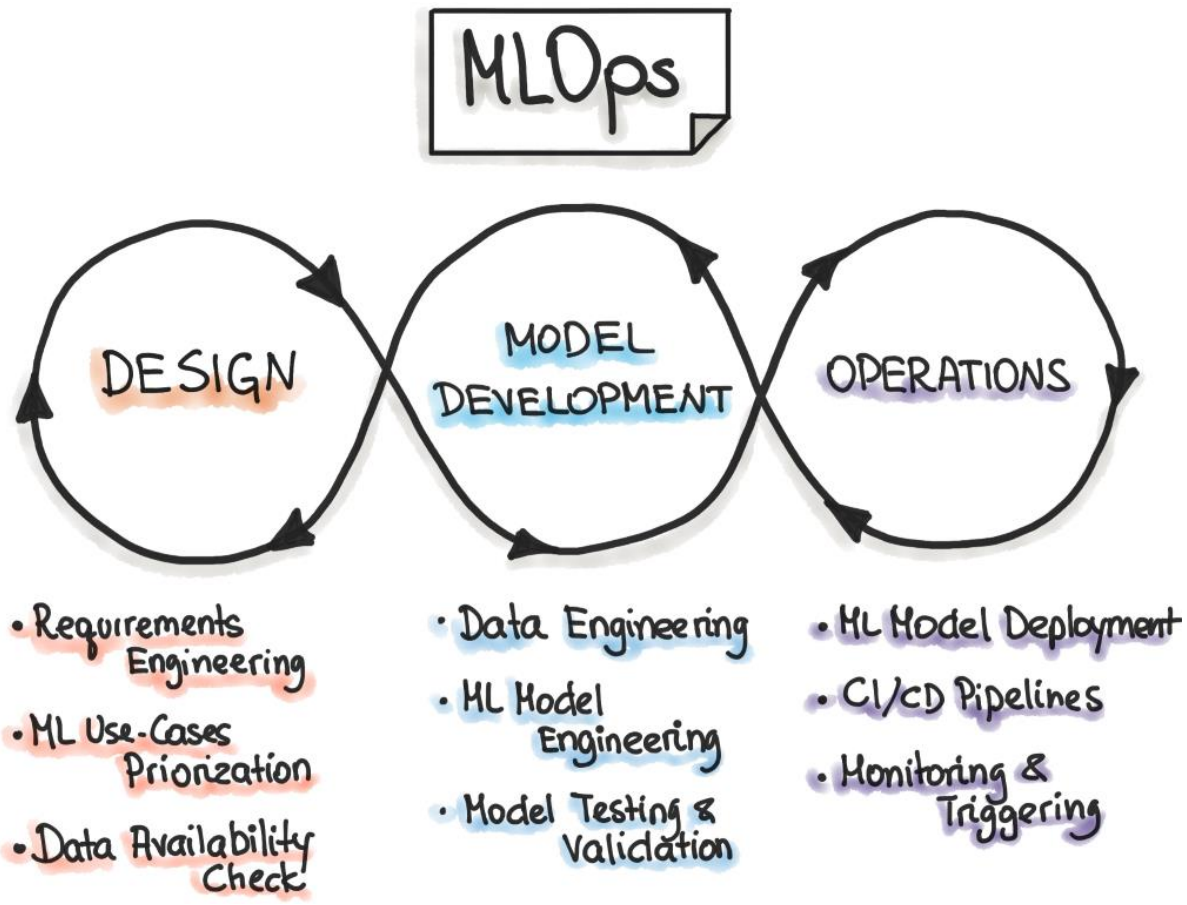


Figure 5. ML Ops: processus de développement de l'apprentissage automatique de bout en bout

[5]

Avec (MLOps), nous voulons fournir un processus complet de développement de l'apprentissage automatique pour concevoir, construire et gérer des logiciels reproductibles, testables et évolutifs basés sur l'apprentissage automatique.

Les MLOps reposent sur l'automatisation de l'entraînement et du réentraînement des modèles d'IA tout en fournissant une observabilité robuste de ces modèles.

La première phase est consacrée à la compréhension de l'activité, à la compréhension des données et à la conception du logiciel basé sur la ML. Au cours de cette étape, nous identifions notre utilisateur potentiel, concevons la solution d'apprentissage automatique pour résoudre son problème et évaluons le développement ultérieur du projet.
Dans un premier temps, nous définissons les cas d'utilisation de l'apprentissage automatique et les classons par ordre de priorité. La meilleure pratique pour les projets d'apprentissage automatique est de travailler sur un cas d'utilisation de l'apprentissage automatique à la fois. En outre, la phase de conception vise à inspecter les données disponibles qui seront nécessaires pour entraîner notre modèle et à spécifier les exigences fonctionnelles et non fonctionnelles de notre modèle de ML. Nous devrions utiliser ces exigences pour concevoir l'architecture de l'application de ML, établir la stratégie de service et créer une suite de tests pour le futur modèle de ML.

La phase suivante « Expérimentation et développement du ML » est consacrée à la vérification de l'applicabilité du ML à notre problème en mettant en œuvre la preuve de concept du modèle ML. Ici, nous exécutons itérativement différentes étapes, telles que l'identification ou le perfectionnement de l'algorithme de ML adapté à notre problème, l'ingénierie des données et l'ingénierie du modèle. L'objectif principal de cette phase est de fournir un modèle de ML de qualité stable que nous utiliserons en production.

L'objectif principal de la phase « ML Operations » est de livrer le modèle de ML précédemment développé en production en utilisant les pratiques DevOps établies telles que les tests, le versionnage, la livraison continue et la surveillance.

Les trois phases sont interconnectées et s'influencent mutuellement. Par exemple, la décision de conception pendant la phase de conception se propagera dans la phase d'expérimentation et influencera finalement les options de déploiement pendant la phase d'exploitation finale.

Types de données / Sources de données

Dans la plupart des cas, les données sont créées dans des systèmes opérationnels, c'est-à-dire des logiciels qui fournissent des services pour soutenir les activités de l'entreprise. Ces données peuvent prendre plusieurs formes, structurées, semi-structurées ou non structurées.

[5] Source: https://ml-ops.org/content/mlops-principles

● Les données structurées font référence aux données qui suivent un modèle. Il s'agit généralement de données pouvant tenir dans un tableau, dont les colonnes définissent la manière dont les données doivent être organisées.

● Les données non structurées font référence aux données qui ne peuvent pas s'inscrire dans un modèle, un schéma ou une structure commune. Il s'agit généralement de vidéos, de langage naturel, de sons, d'images.

● Les données semi-structurées sont une combinaison de données structurées et non structurées. Il peut s'agir d'un texte auquel sont attachées des métadonnées : un courriel est typiquement une donnée semi-structurée : l'expéditeur, le destinataire et l'objet sont des données structurées, tandis que le corps du courriel peut être considéré comme non structuré.

Les données peuvent également provenir de capteurs matériels, tels que des caméras, des détecteurs, des thermomètres, des microphones, etc.

Management des data[6]

Management de la qualité des données

La gestion de la qualité des données fait référence aux activités de planification, de mise en œuvre et de contrôle qui appliquent des techniques de gestion de la qualité aux données, afin de s'assurer qu'elles sont adaptées à la consommation et qu'elles répondent aux besoins des consommateurs de données.

La qualité des données est l'un des piliers les plus importants de la gestion des données car

● Des données de mauvaise qualité peuvent entraîner une grave déformation de la situation et conduire à des modèles d'IA biaisés utilisés pour soutenir des décisions importantes.

● L'ensemble du coût du stockage des données sera gaspillé si les données ne sont pas de haute qualité.

● Les données doivent être exactes, actuelles et accessibles à tous au moment et à l'endroit où ils en ont besoin afin qu'ils puissent prendre des décisions fondées sur des données, pour créer des avantages concurrentiels sur le marché.

Gouvernance des données

Définit les règles, les processus et la structure organisationnelle nécessaires à la gestion des données au sein d'une organisation.

La gouvernance des données vise à maximiser l'utilisation des données, à améliorer la connaissance des données et à fournir un cadre pour rendre les données utilisables.

La gouvernance des données peut être mise en œuvre en suivant de multiples schémas :

● Décentralisée : La collaboration repose sur des comités ad hoc. Difficulté à définir la propriété des données.

● Fédérée : Stratégie centralisée avec exécution décentralisée. Activité gérée dans une perspective à l'échelle de l'entreprise.

● En réseau : La collaboration repose sur une organisation plus formelle. Les responsabilités n'ont pas d'impact sur les organigrammes.

● Centralisé : Modèle le plus formel et le plus mature. Tout ce qui concerne les données appartient à l'organisation de gestion de l'information.

[6] https://www.ibm.com/downloads/cas/YD5R1XLB

Management des metadonnées

Les métadonnées sont souvent appelées « données sur les données » ou « informations sur les informations » : Il s'agit d'informations structurées qui décrivent, expliquent et localisent des données, afin de faciliter leur extraction, leur utilisation et leur gestion.

La gestion des métadonnées, comme son nom l'indique, est le processus de gestion des métadonnées. Elle s'articule autour des services suivants :
● Glossaire métier : une liste de termes et de leurs définitions que les magasins de données.
● Catalogue de données : liste des données effectivement disponibles et leur lien avec les termes métier.
● Linéaire de données : vue sur le cycle de vie des données, de l'endroit où elles sont créées à celui où elles sont utilisées, avec les étapes de transformation.
● Rapports sur la qualité des données : aperçu de la qualité des données disponibles.
● Emplacement des données : indication sur l'emplacement des données et sur la manière de les obtenir.

Session 3: Prérequis de l'IA 2e partie

Architecture des données[7]

L'architecture des données vise à répondre aux questions suivantes :
● Quelles sont les données dont je dispose ?
● Où se trouvent les données ?
● Où se trouvent les données de référence ?
● Comment exploiter les données pour l'IA et l'intelligence économique ?
● Comment gérer la cohérence entre les copies ?
● Comment productiser les modèles d'IA ?
● Qu'en est-il des processus de gouvernance des données ?
● Comment être en conformité avec la réglementation sur les données (GDPR, Solvabilité II, ...) ?

L'architecture des données ne définit pas la manière dont les logiciels et les services doivent être construits. Elle traite uniquement des données et de la manière de les utiliser correctement dans d'autres cas d'utilisation que celui dans lequel elles ont été créées.

L'architecture des données implique la conception et la structuration des actifs de données d'une entreprise, y compris la définition des modèles de données, des flux et des systèmes de gestion pour stocker, gérer, déplacer et analyser les données. La conception de l'architecture des données doit être basée sur les exigences et les contraintes de l'entreprise, et les architectes et ingénieurs des données utilisent ces exigences pour créer le modèle de données et les structures sous-jacentes qui le soutiennent.

[7] https://www.ibm.com/topics/data-architecture

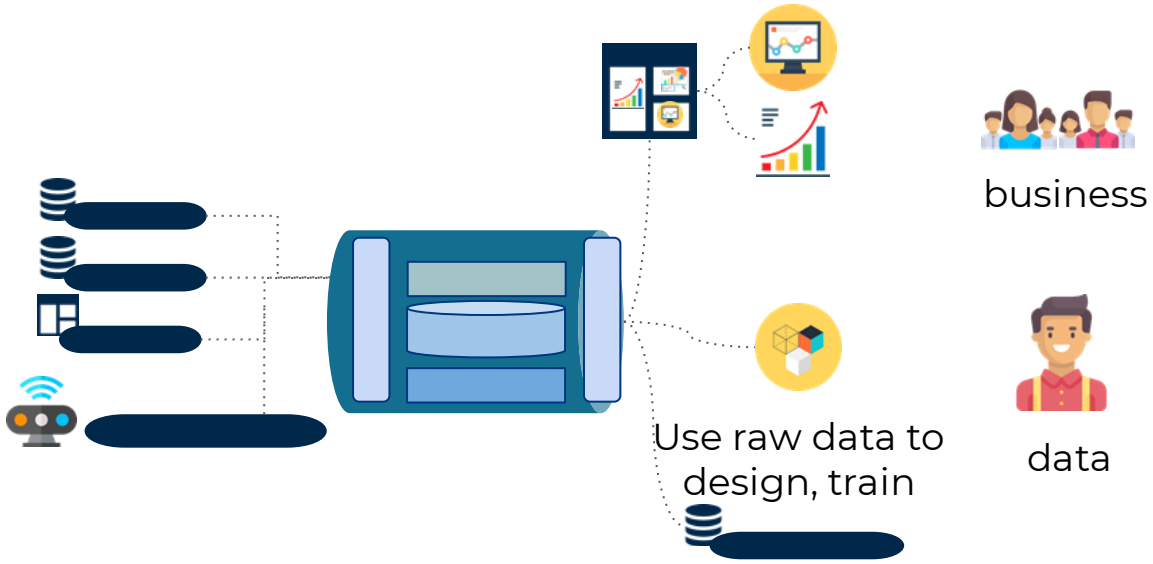


Figure 6. exemple de la platforme d'architecture des données digazua [8]

Temps réél ou architectures par lot[9]

La pratique de l'architecture des données fournit plusieurs modèles pour intégrer les données entre les systèmes. Selon le cas d'utilisation, les données peuvent être diffusées en temps réel pour permettre aux consommateurs (modèles d'IA ou tableaux de bord de veille stratégique) d'obtenir les informations les plus récentes. Une architecture de données « stream first » sélectionnera le type d'outils et de plateforme pour répondre à ces exigences.

Dans d'autres cas, un modèle d'architecture de données basé sur le traitement par lots permettra un traitement moins gourmand en ressources. Là encore, l'architecture de données définira les outils et la plateforme nécessaires pour collecter, stocker et distribuer les données par lots.

[8] www.digazu.com
[9] https://www.confluent.io/learn/batch-vs-real-time-data-processing/

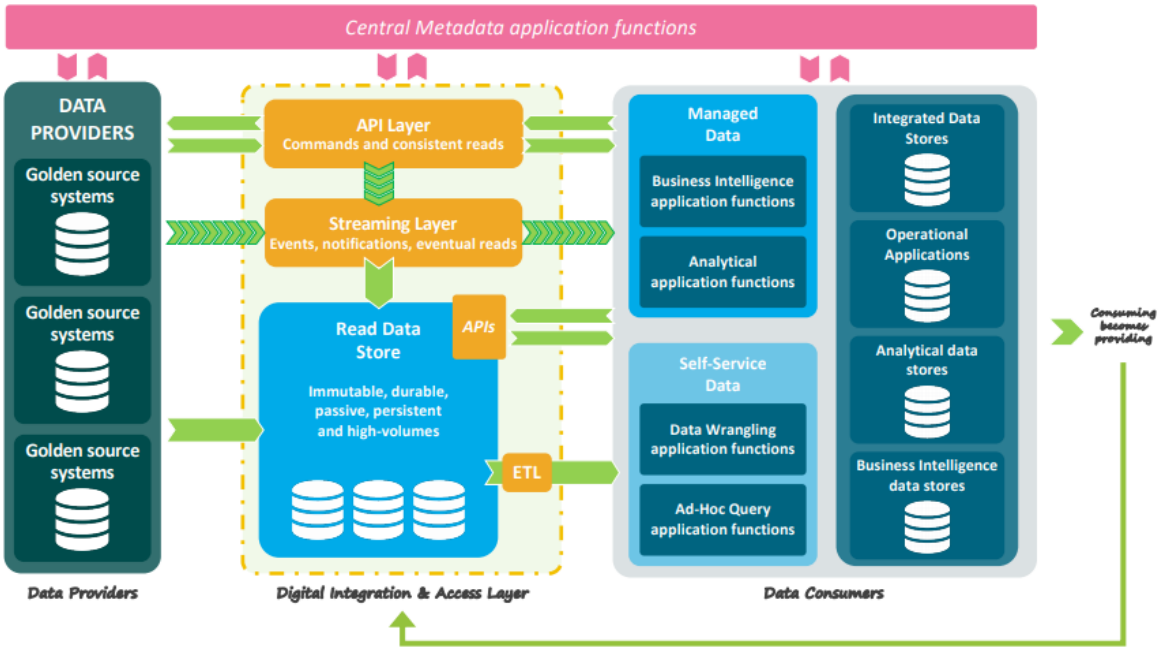
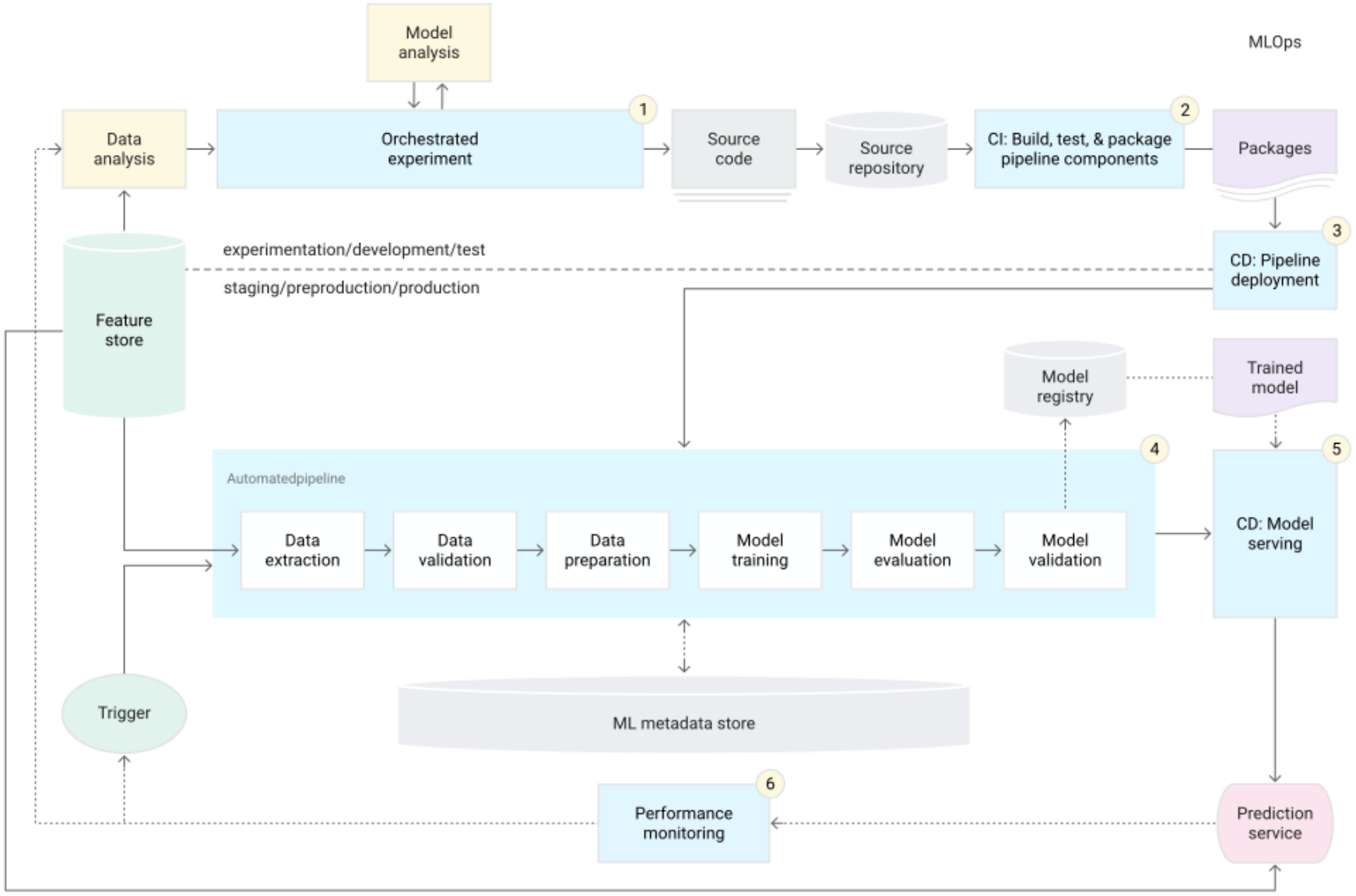


Figure 7. Répartition des données data en mode stream ou batch

Laboratoire d'analyse de données et plateformes MLOps[1]

Dans les architectures de données modernes, qui sont conçues pour soutenir les applications basées sur les données et l'intelligence artificielle, il y a un composant crucial qui est le laboratoire d'analyse de données. Ce système logiciel est conçu pour stocker les données et les mettre à la disposition des scientifiques des données, leur permettant ainsi d'explorer et de construire des modèles d'intelligence artificielle. Une partie importante de ce laboratoire de données est son intégration avec un pipeline de production d'IA et une plateforme de service, qui fournit la capacité d'héberger des modèles d'intelligence artificielle dans un environnement de production.

Comme décrit précédemment, MLOps est un paradigme qui tente d'appliquer les meilleures pratiques de la gestion du cycle de vie du développement logiciel moderne à l'apprentissage automatique. Google définit l'architecture d'une plateforme MLOps comme suit[2]:


MLOps
Model analysis
Data analysis
Orchestrated experiment
1
Source code
Source repository
CI: Build, test, & package pipeline components
2
Packages
3
CD: Pipeline deployment
experimentation/development/test
staging/preproduction/production
Feature store
Model registry
Trained model
4
5
Automatedpipeline
Data extraction
Data validation
Data preparation
Model training
Model evaluation
Model validation
CD: Model serving
Trigger
ML metadata store
6
Performance monitoring
Prediction service

Session 4: Prérequis de l'IA 3ème partie

L'ingénierie des données[10] [11]

L'ingénierie des données fait référence au processus de construction et de maintenance des systèmes et de l'infrastructure nécessaires à la collecte, au stockage et au traitement de grandes quantités de données. En d'autres termes, l'ingénierie des données est le processus de mise en œuvre de l'architecture des données.

La conception, ou l'architecture (de données), de ces systèmes se fait en tenant compte des contraintes de sécurité et de conformité, d'évolutivité et d'efficacité, de fiabilité et de fidélité, de flexibilité et de portabilité. Comme le système est généralement le fournisseur de données pour les data scientists et leurs modèles d'apprentissage automatique, il devrait également être en mesure d'exploiter et d'entraîner en continu des modèles d'apprentissage automatique préexistants. Les systèmes d'ingénierie des données constituent également l'épine dorsale des applications de veille stratégique. Les données doivent donc être faciles à intégrer et de grande qualité.

Ingénierie des données et Economie circulaire

L'industrie agroalimentaire est un exemple d'entreprise où l'ingénierie des données joue un rôle essentiel. Elle est utilisée pour apporter les données aux outils qui permettent à l'équipe de prendre des décisions basées sur les données et d'optimiser les opérations.

Il existe de nombreuses applications de l'ingénierie des données dans l'industrie agroalimentaire. En voici quelques exemples :

- Dans l'agriculture de précision, l'ingénierie des données peut être utilisée pour collecter et analyser des données provenant de capteurs et d'autres sources afin d'optimiser les rendements des cultures et de réduire le gaspillage.
- Dans la gestion de la chaîne d'approvisionnement, l'ingénierie des données peut être utilisée pour collecter et analyser les données provenant des fournisseurs, des distributeurs et des clients afin d'optimiser la logistique et de réduire les déchets.
- En matière de sécurité alimentaire et de contrôle de la qualité, l'ingénierie des données peut être utilisée pour collecter et analyser des données provenant de capteurs et d'autres sources afin de détecter les défauts, la contamination et d'autres problèmes.
- L'ingénierie des données peut être utilisée pour surveiller et analyser la consommation d'énergie et d'eau

[10] https://cloud.google.com/learn/certification/data-engineer
[11] https://www.dataquest.io/blog/what-is-a-data-engineer/

Ingénierie des données dans l’industrie agoalimentaire

Dans le contexte de la transition circulaire, l'ingénierie des données joue un rôle crucial dans la gestion et l'analyse efficaces des données relatives aux pratiques circulaires. Elle facilite la collecte, le stockage et l'analyse des données essentielles, rendant possibles des améliorations, notamment :

● L'approvisionnement en matériaux : En capturant et en analysant les données relatives à l'approvisionnement en matériaux, l'ingénierie des données fournit des indications précieuses sur l'origine, la composition et la disponibilité des ressources. Ces informations aident à prendre des décisions éclairées concernant la sélection des matériaux pour une meilleure conception et durabilité des produits.
● Cycles de vie des produits : L'ingénierie des données permet le suivi et l'analyse complets des cycles de vie des produits. Elle englobe les données relatives aux étapes de fabrication, d'utilisation, de maintenance et de fin de vie. En comprenant l'état, les modes d'utilisation et le potentiel de réparation ou de réutilisation, les entreprises peuvent prendre des décisions éclairées sur l'allongement de la durée de vie des produits et la réduction de la production de déchets.
● La gestion des déchets : Une gestion efficace des déchets est un aspect crucial de la transition circulaire. L'ingénierie des données donne aux organisations les moyens de collecter et d'analyser des données sur la production de déchets, l'élimination et les processus de recyclage. En exploitant ces informations, les entreprises peuvent identifier les opportunités de minimiser les déchets, d'optimiser les pratiques de recyclage et de réduire les coûts associés, contribuant ainsi à une durabilité accrue.

En outre, l'ingénierie des données peut intégrer d'autres sources de données pertinentes, telles que le retour d'information des clients, les informations sur la chaîne d'approvisionnement et les évaluations de l'impact sur l'environnement. Cette approche globale améliore la prise de décision en tenant compte de multiples facteurs et en alignant les pratiques sur les principes de l'économie circulaire.

Introduction aux sessions de Modèles d'IA

Les sessions 5, 6 et 7 couvriront les principaux modèles d'IA, en fonction des types de données. La session 5 couvrira la régression et la classification (données tabulaires), la session 6 couvrira les modèles de traitement neuronal du langage (NLP pour les données textuelles) et enfin la session 7 couvrira les modèles de vision par ordinateur (images et vidéos).

Session 5: Modèles d'IA 1: Régression / classification[12]

La régression et la classification sont deux types d'algorithmes d'apprentissage automatique supervisé utilisés dans l'analyse et la prédiction des données.

L'algorithme de régression est utilisé pour prédire des valeurs numériques continues. Il tente de trouver la relation entre les variables indépendantes et les variables dépendantes. Dans la régression, la variable dépendante est une valeur numérique continue. Il existe de nombreux types d'algorithmes de régression tels que la régression linéaire, la régression polynomiale, la régression logistique, etc.

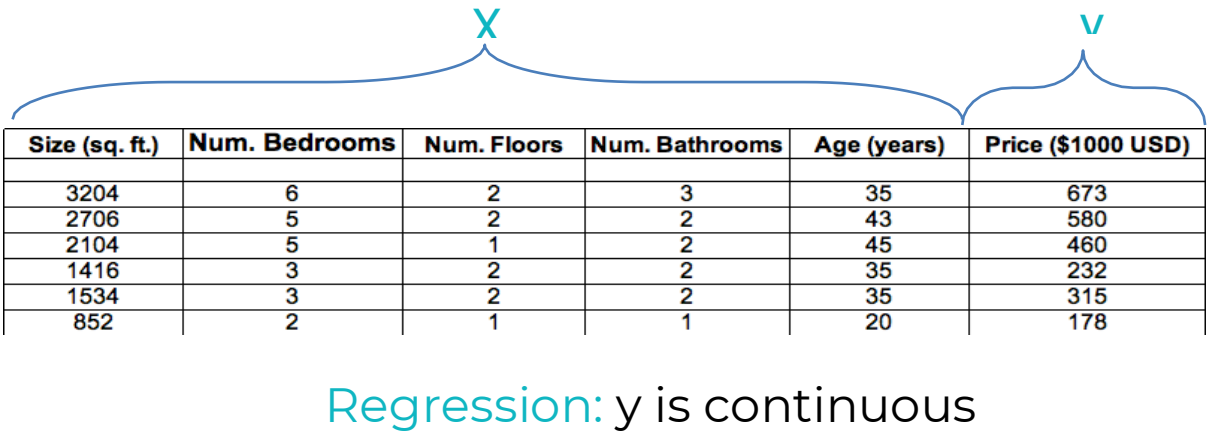


Figure 8. Exemple de prédiction de la valeur d'un bien immobilier

Les algorithmes de régression sont largement appliqués pour prédire les résultats, prévoir les données, analyser les séries temporelles et trouver les dépendances entre les variables.

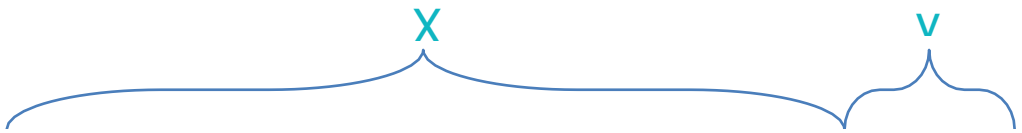
Dans l'industrie agroalimentaire, l'analyse de régression peut être utilisée pour prédire les défauts des produits et améliorer la qualité des produits dans les chaînes de production[13], en contrôlant davantage les paramètres de qualité grâce à l'analyse des données.
Voici quatre exemples d'applications :
- Contrôle statistique prédictif des processus
- Prédiction de la durée de vie des produits

[12] https://themlbook.com/

[13] https://www.sartorius.com/en/knowledge/science-snippets/four-uses-of-data-analytics-in-the-food-and-beverage-industry-603058

- Mesure des attributs de qualité critiques
- Modélisation de la contrefaçon (prévention de la fraude alimentaire)

D'autre part, un algorithme de classification est utilisé pour prédire les résultats discrets mesurables. Il trouve la limite de décision pour classer les données d'entrée dans différentes catégories. Le résultat peut être binaire (oui ou non), ordinal (élevé, moyen, faible) ou multi-classes (classe A, classe B, classe C, etc.). Il existe de nombreux types d'algorithmes de classification, tels que l'arbre de décision, le voisinage le plus proche, la forêt aléatoire, etc.



Age	Has_Job	Own_House	Credit_Rating	Class
young	false	false	fair	No
young	false	false	good	No
young	true	false	good	Yes
young	true	true	fair	Yes
young	false	false	fair	No
middle	false	false	fair	No
middle	false	false	good	No
middle	true	true	good	Yes
middle	false	true	excellent	Yes
middle	false	true	excellent	Yes
old	false	true	excellent	Yes
old	false	true	good	Yes
old	true	false	good	Yes
old	true	false	excellent	Yes
old	false	false	fair	No

Classification: y is discrete

Figure 9. Exemple de modèle de prise de décision basé sur un algorithme de classification

En résumé, la régression prédit des valeurs numériques continues, tandis que la classification prédit des valeurs discrètes.

La prévision consiste à utiliser des données et des modèles statistiques pour prédire les tendances et les résultats futurs. L'IA peut être utilisée pour améliorer la précision et la fiabilité des prévisions, en analysant de grandes quantités de données et en identifiant des modèles et des tendances qui pourraient ne pas être visibles pour les analystes humains.

Par exemple, dans l'industrie manufacturière, l'IA peut être utilisée pour prévoir la demande de produits, optimiser les niveaux de stocks et améliorer la gestion de la chaîne d'approvisionnement. Dans le secteur financier, l'IA peut être utilisée pour prévoir le cours des actions, identifier les opportunités d'investissement et gérer les risques.

Session 6: Modèles d'IA 2 : NLP (données textuelles)[14][15]

Le traitement du langage naturel (TLN) est un domaine de l'informatique qui traite de l'interaction entre les ordinateurs et les humains utilisant le langage naturel. Il s'agit de techniques informatiques qui permettent aux ordinateurs de comprendre, d'interpréter et de générer du langage humain. La PNL vise à créer des machines capables d'analyser, d'interpréter et d'agir sur des entrées en langage naturel, telles que du texte, de la parole et des données.

L'objectif principal du NLP est de développer des algorithmes et des modèles qui permettent aux ordinateurs de comprendre le sens du langage humain d'une manière qui imite la compréhension humaine. Le NLP est un domaine complexe qui comprend plusieurs sous-domaines, notamment la génération de langage naturel, la compréhension du langage naturel et la reconnaissance vocale.

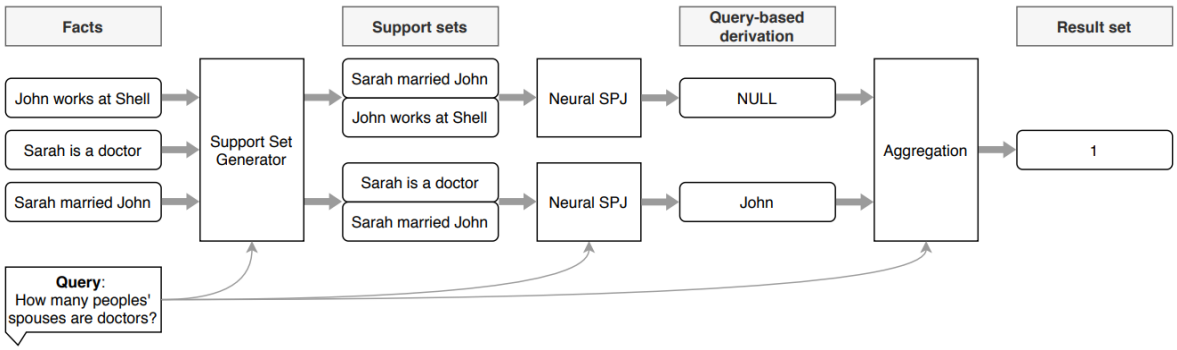


Figure 10. Exemple d'un modèle NLP répondant à une requête

La génération de langage naturel implique la création de texte ou de discours à partir de données structurées ou non structurées. C'est l'inverse de la compréhension du langage naturel, dont l'objectif est de créer une représentation du langage humain qui soit compréhensible par les ordinateurs.

Multimodal Neural Language Models, Kiros et al., 2013

Show and Tell: A Neural Image Caption Generator, Vinyals et al. 2014

Figure 11. Exemple d'un modèle NLG générant des légendes d'images

[14] https://github.com/keon/awesome-nlp
[15] https://themlbook.com/

La compréhension du langage naturel est le processus qui consiste à donner un sens à l'entrée du langage humain. Il s'agit de comprendre la grammaire, la syntaxe et la sémantique du langage humain, ainsi que le contexte et l'intention qui se cachent derrière les mots et les phrases utilisés. Cet aspect est crucial pour le développement de systèmes intelligents capables d'interpréter avec précision le langage humain.

La reconnaissance vocale est la capacité d'un ordinateur à reconnaître et à transcrire le langage parlé. Bien que la reconnaissance vocale soit souvent utilisée de manière interchangeable avec la PNL, elle n'est qu'un aspect du domaine plus vaste de la PNL.

Les applications de la PNL sont vastes et diverses. Les chatbots, les logiciels de reconnaissance vocale, la traduction automatique, l'analyse des sentiments et le résumé de texte sont quelques-unes des applications courantes du NLP.

Les chatbots sont des programmes informatiques conçus pour simuler une conversation humaine. Ils utilisent des techniques de traitement du langage naturel pour comprendre les messages des utilisateurs et y répondre. Les chatbots sont couramment utilisés dans le service client, où ils peuvent fournir des réponses rapides et automatisées aux questions les plus courantes.

Les logiciels de reconnaissance vocale sont utilisés pour transcrire la parole humaine en texte. Ils sont utilisés dans des applications telles que les assistants virtuels, les logiciels de dictée et les outils d'apprentissage des langues.

La traduction automatique consiste à traduire automatiquement un texte d'une langue à une autre. Elle est couramment utilisée dans des applications telles que les services de traduction en ligne.

L'analyse des sentiments est le processus d'analyse d'un texte pour déterminer le sentiment ou l'émotion sous-jacente. Elle est utilisée dans des applications telles que la surveillance des médias sociaux, où les entreprises peuvent suivre le sentiment des clients à l'égard de leur marque.

Le résumé de texte est le processus qui consiste à résumer de longs documents textuels en des résumés plus courts et significatifs.

Le NLP est un domaine extrêmement complexe et exigeant, qui présente plusieurs défis techniques et pratiques. L'un des principaux défis du NLP est de gérer l'ambiguïté du langage naturel. Le langage est souvent ambigu, avec plusieurs significations possibles pour la même phrase. Les systèmes de TAL doivent être en mesure d'identifier et de désambiguïser ces cas afin de comprendre avec précision les entrées en langage naturel.

Un autre défi majeur du NLP est de gérer la variabilité du langage humain. Les gens peuvent exprimer la même idée de différentes manières, en utilisant des mots, des expressions idiomatiques et des structures syntaxiques différentes. Les systèmes de TAL doivent être capables de reconnaître ces variations et de s'y adapter pour interpréter correctement le langage humain.

Le PNA est également confronté à des défis liés au contexte et au flux conversationnel. Dans le langage naturel, le sens d'une phrase peut dépendre du contexte dans lequel elle est utilisée. Les systèmes de PNA doivent intégrer le contexte pour interpréter avec précision les entrées en langue naturelle.

Dans l'ensemble, le NLP est un domaine qui progresse rapidement et qui a le potentiel de révolutionner la façon dont nous interagissons avec la technologie. Au fur et à mesure que les systèmes de PNA se développent et s'améliorent, on peut s'attendre à des applications de plus en plus sophistiquées dans des domaines tels que les soins de santé, la finance et l'éducation.

Session 7: Modèles d'IA 3: Vision par ordinateur (images / videos)[16] [17]

La vision par ordinateur est un domaine d'étude qui permet aux ordinateurs d'interpréter et de comprendre le monde visuel qui les entoure, en utilisant des algorithmes et des modèles mathématiques pour identifier et analyser les caractéristiques des images ou des vidéos de manière similaire à la vision humaine. Les applications de la vision par ordinateur sont diverses, allant de la conduite automobile autonome à la reconnaissance des défauts dans les processus de fabrication. Dans le contexte de l'économie circulaire, la vision par ordinateur peut jouer un rôle essentiel dans la transition vers un modèle de production et de consommation plus durable et plus économe en ressources.

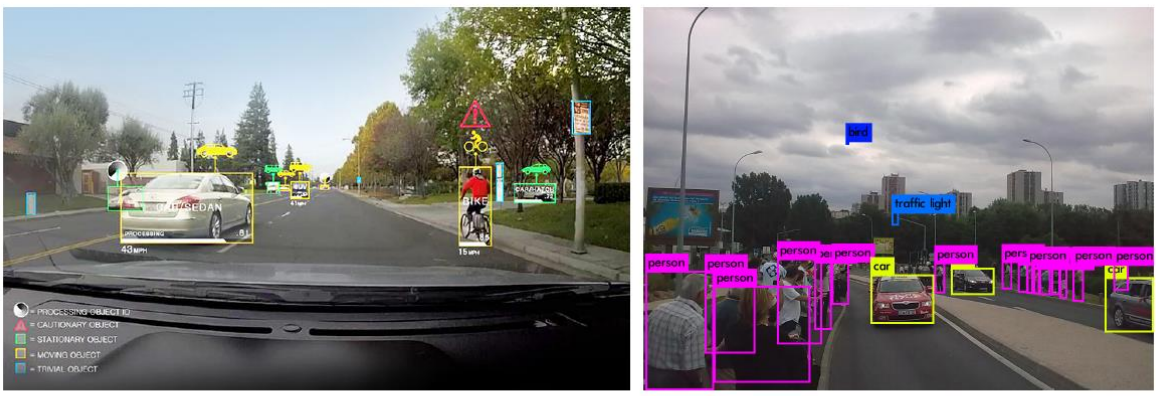


Figure 12. Exemples d'applications de la vision par ordinateur

[16] https://github.com/jbhuang0604/awesome-computer-vision
[17] https://themlbook.com/

Vision par ordinateur et économie circulaire

L'une des applications potentielles de la vision par ordinateur dans l'économie circulaire est la gestion des déchets. Le modèle linéaire actuel d'élimination des déchets, dans lequel les matériaux sont utilisés puis mis au rebut, n'est pas durable et nuit à l'environnement. En revanche, l'économie circulaire souligne l'importance de la prévention, de la réutilisation et du recyclage des déchets. Toutefois, le processus de recyclage est souvent entravé par l'absence de technologies de tri et de séparation efficaces, ce qui entraîne une contamination et une utilisation inefficace des ressources.

C'est là que la vision par ordinateur entre en jeu. En utilisant des algorithmes d'apprentissage automatique, il est possible de développer des systèmes de tri intelligents capables d'identifier différents types de matériaux et de les séparer en conséquence. Par exemple, les bouteilles en plastique peuvent être distinguées des bouteilles en verre ou des emballages en carton, et séparées en différents flux pour le recyclage. La précision et la rapidité du processus de tri peuvent être grandement améliorées, ce qui réduit le risque de contamination et augmente la récupération de matériaux précieux.

Un autre exemple d'utilisation de la vision par ordinateur dans l'économie circulaire concerne la surveillance et l'analyse des processus industriels. De nombreuses entreprises manufacturières mettent en œuvre des stratégies circulaires pour réduire les déchets et accroître l'efficacité, comme les systèmes de production en boucle fermée ou les processus de refabrication. Toutefois, ces stratégies nécessitent une mesure et un suivi minutieux des intrants et des extrants du processus de production, ainsi que de la qualité des produits finaux.

Les outils de vision par ordinateur peuvent être utiles à cet égard, en fournissant des données en temps réel sur divers aspects du processus de production. Par exemple, des caméras peuvent être utilisées pour surveiller le flux de matériaux et l'état des machines, ce qui permet d'identifier les problèmes potentiels avant qu'ils ne deviennent critiques. De même, la vision par ordinateur peut être utilisée pour analyser la qualité des produits finaux, par exemple pour détecter les défauts dans les articles manufacturés ou pour identifier les variations de couleur ou de texture des matériaux recyclés.

En conclusion, la vision par ordinateur peut jouer un rôle crucial dans la mise en place d'une économie plus durable et circulaire en permettant une gestion intelligente des déchets et un contrôle des processus en temps réel. En exploitant la puissance de l'apprentissage automatique et de la reconnaissance d'images, les entreprises et les décideurs politiques peuvent créer de nouvelles opportunités pour l'efficacité des ressources, réduire les déchets et la pollution, et créer un modèle économique plus résilient. À mesure que la technologie de la vision par ordinateur continue de se développer, elle deviendra probablement un outil de plus en plus indispensable dans la transition vers un avenir plus durable.

Conclusions and principaux messages à retenir

L'intelligence artificielle est un terme largement utilisé pour décrire l'automatisation des tâches humaines à l'aide de diverses technologies telles que l'apprentissage automatique et la science des données. La plupart des techniques d'intelligence artificielle existantes sont utilisées pour prédire ou déduire un résultat à partir de nouvelles données : objet dans une image, sens d'une phrase, probabilité d'une rupture de stock, prochain mot significatif dans une phrase, etc.
Cependant, nous avons vu que pour développer l'intelligence artificielle à grande échelle tout en respectant la réglementation, nous avons besoin de données, de bonnes données et de données gérées. La condition préalable à l'intelligence artificielle et à l'apprentissage automatique est une architecture de données solide, des pipelines de données bien conçus, une gestion de la qualité des données et une gouvernance des données.

Enfin, nous avons détaillé les grandes familles de technologies d'intelligence artificielle :
● Régression/Classification : apprentissage de groupes de données existants afin de prédire le groupe auquel appartiennent les nouvelles données. Elle peut être utilisée pour prévoir la demande, regrouper les consommateurs et prédire les prochaines actions
● NLP : le traitement du langage naturel comprend les phrases écrites ou prononcées par un humain et est capable de générer des phrases cohérentes compte tenu de sa compréhension.
● Vision par ordinateur : apprendre à un ordinateur à identifier des objets à partir d'une image d'une vidéo.