Présentation du machine learning

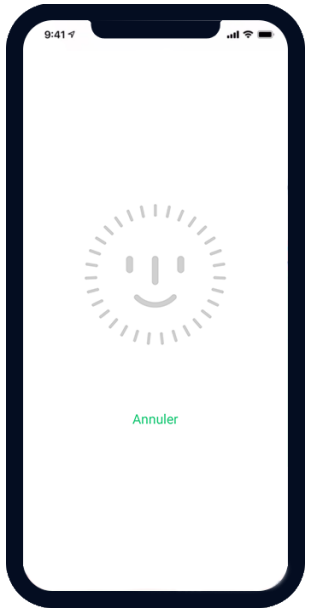
Objectifs:

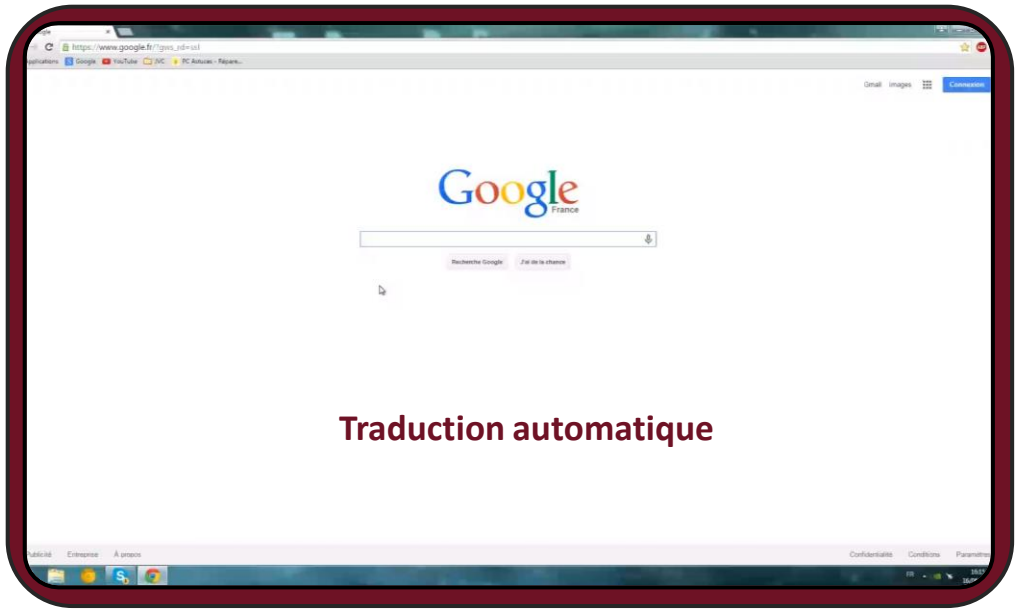
— Définir le machine learning
— Identifier si un problème relève ou non du machine learning
— Donner des exemples de cas concrets
— Les étapes d'un projet data

L'Intelligence Artificielle change tout.

Parfois, il semble que presque chaque site web, application ou outil de productivité cite l'Intelligence Artificielle (IA) comme l'ingrédient secret de leur recette du succès. Il est essentiel de souligner que l'Intelligence Artificielle est actuellement omniprésente dans notre vie quotidienne, jouant un rôle significatif dans divers domaines.

Système de reconnaissance faciale





Traduction automatique

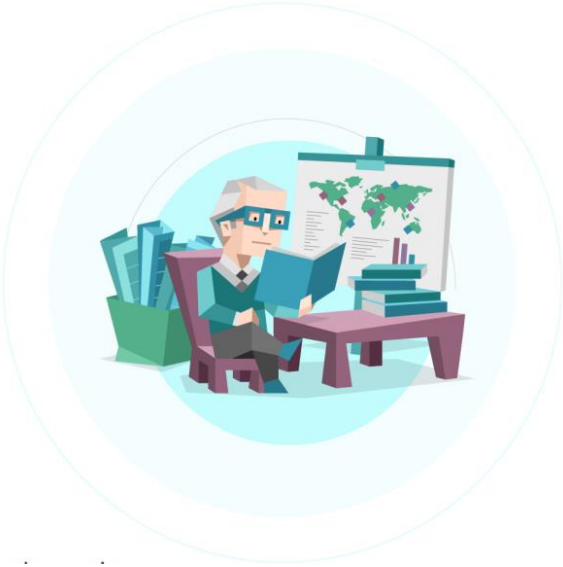
Système de recommandation



L'Intelligence Artificielle

Définition:

L'intelligence artificielle, définie comme l'ensemble des techniques mises en œuvre afin de construire des machines capables de faire preuve d'un comportement que l'on peut qualifier d'intelligent, fait aussi appel aux sciences cognitives, à la neurobiologie, à la logique, à l'électronique, à l'ingénierie et bien plus encore.



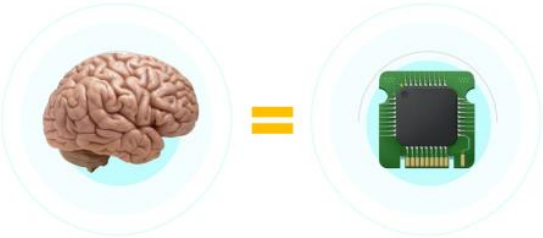
Nous nous appelons **Homo sapiens** (l'homme sage) car notre intelligence nous est si précieuse. Depuis des millénaires, nous avons essayé de comprendre comment nous pensons; c'est-à-dire comment une simple poignée de matière peut percevoir, comprendre, prédire et manipuler un monde bien plus vaste et complexe qu'elle-même. Le domaine de l'Intelligence Artificielle va encore plus loin : il tente non seulement de comprendre, mais aussi de construire des entités intelligentes.

COMMENT DÉFINISSONS-NOUS **L'INTELLIGENCE** ?

Acting humanly: the Turing test! (1950)



A computer passes the test if a human interrogator, after posing some written questions, cannot tell whether the written responses come from a person or from a computer.



Si la machine parvient à convaincre l'examinateur qu'elle est un être humain pendant une durée significative, alors elle est considérée comme ayant passé avec succès le test de Turing. En d'autres termes, elle a démontré une forme d'intelligence indiscernable de celle d'un être humain, du moins dans le contexte spécifique du test.

L'Intelligence Artificielle

Intelligence artificielle

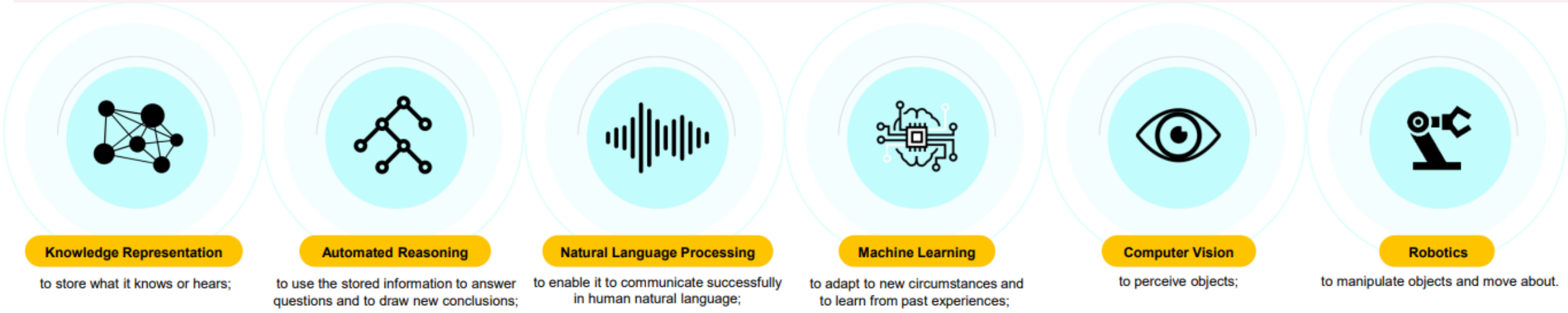
Système d'imitation de la pensée et des actions humaines

Machine Learning

Systèmes qui apprennent sur la base de grands ensembles de données structurés

Deep Learning

Systèmes qui apprennent via des réseaux neuronaux sans être guidés par l'homme



Machine Learning

Définition:

L'apprentissage est une modification d'un comportement sur la base d'une expérience. On parle d'apprentissage automatique, ou machine learning, quand un programme a la capacité d'apprendre sans que cette modification ne soit explicitement programmée.



On peut ainsi opposer un programme classique, qui utilise une procédure et les données qu'il reçoit en entrée pour produire en sortie des réponses, à un programme d'apprentissage automatique, qui utilise les données et les réponses afin de produire la procédure qui permet d'obtenir les secondes à partir des premières.

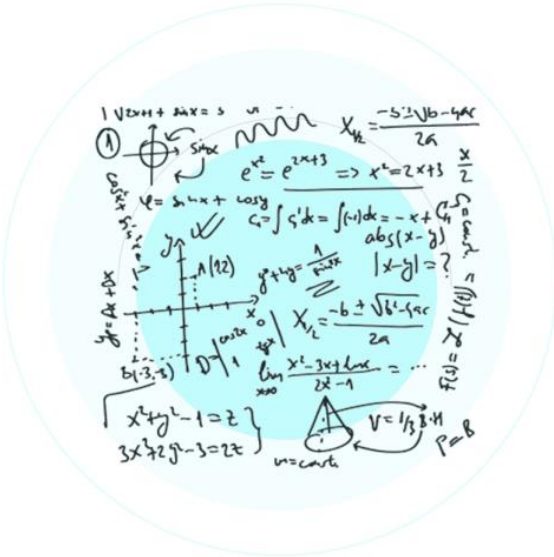
POURQUOI UTILISER LE MACHINE LEARNING ?

Machine Learning

Le machine learning peut servir à résoudre des problèmes
— que l'on ne sait pas résoudre
— que l'on sait résoudre, mais dont on ne sait formaliser en termes algorithmiques comment nous les résolvons (c'est le cas par exemple de la reconnaissance d'images ou de la compréhension du langage naturel) ;
— que l'on sait résoudre, mais avec des procédures beaucoup trop gourmandes en ressources informatiques (c'est le cas par exemple de la prédiction d'interactions entre molécules de grande taille, pour lesquelles les simulations sont très lourdes).

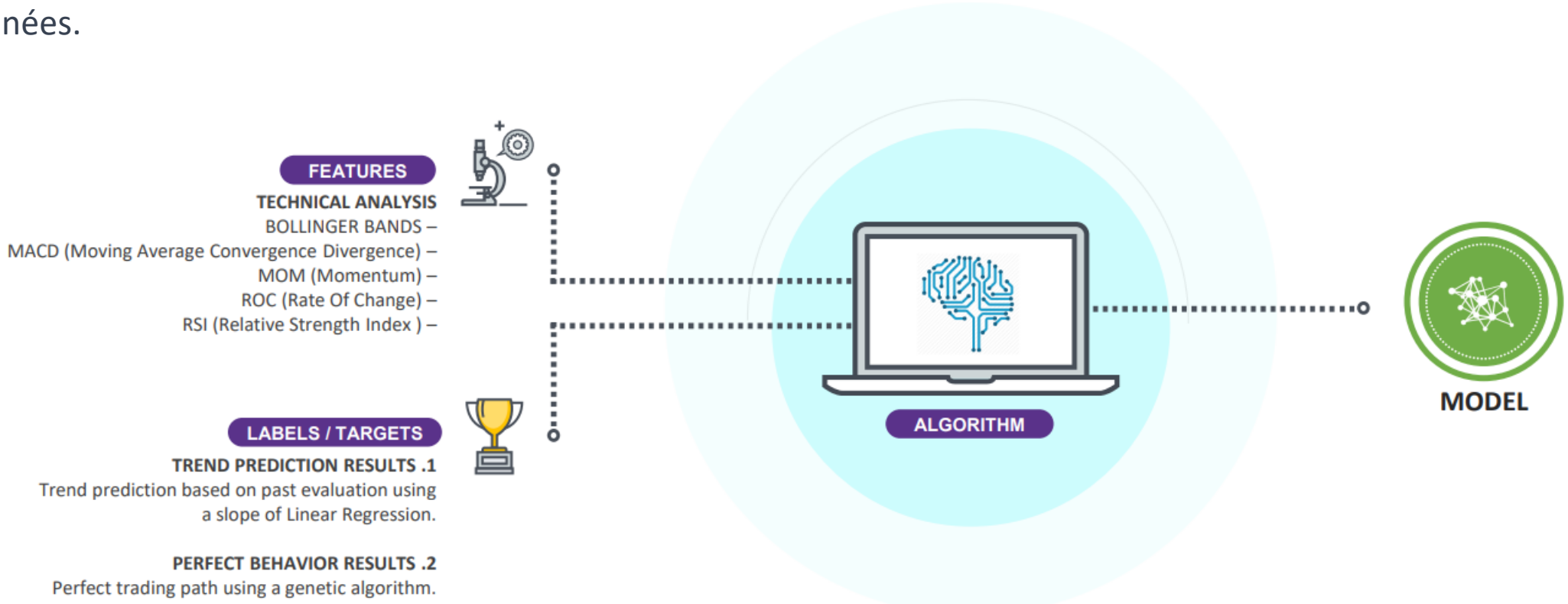






Ingrédients du machine learning

Le machine learning repose sur deux piliers fondamentaux :
— d'une part, les données, qui sont les exemples à partir duquel l'algorithme va apprendre ;
— d'autre part, l'algorithme d'apprentissage, qui est la procédure que l'on fait tourner sur ces données
pour produire un modèle. On appelle entraînement le fait de faire tourner un algorithme d'apprentissage sur un jeu de données.



Domaines d’application



Spam



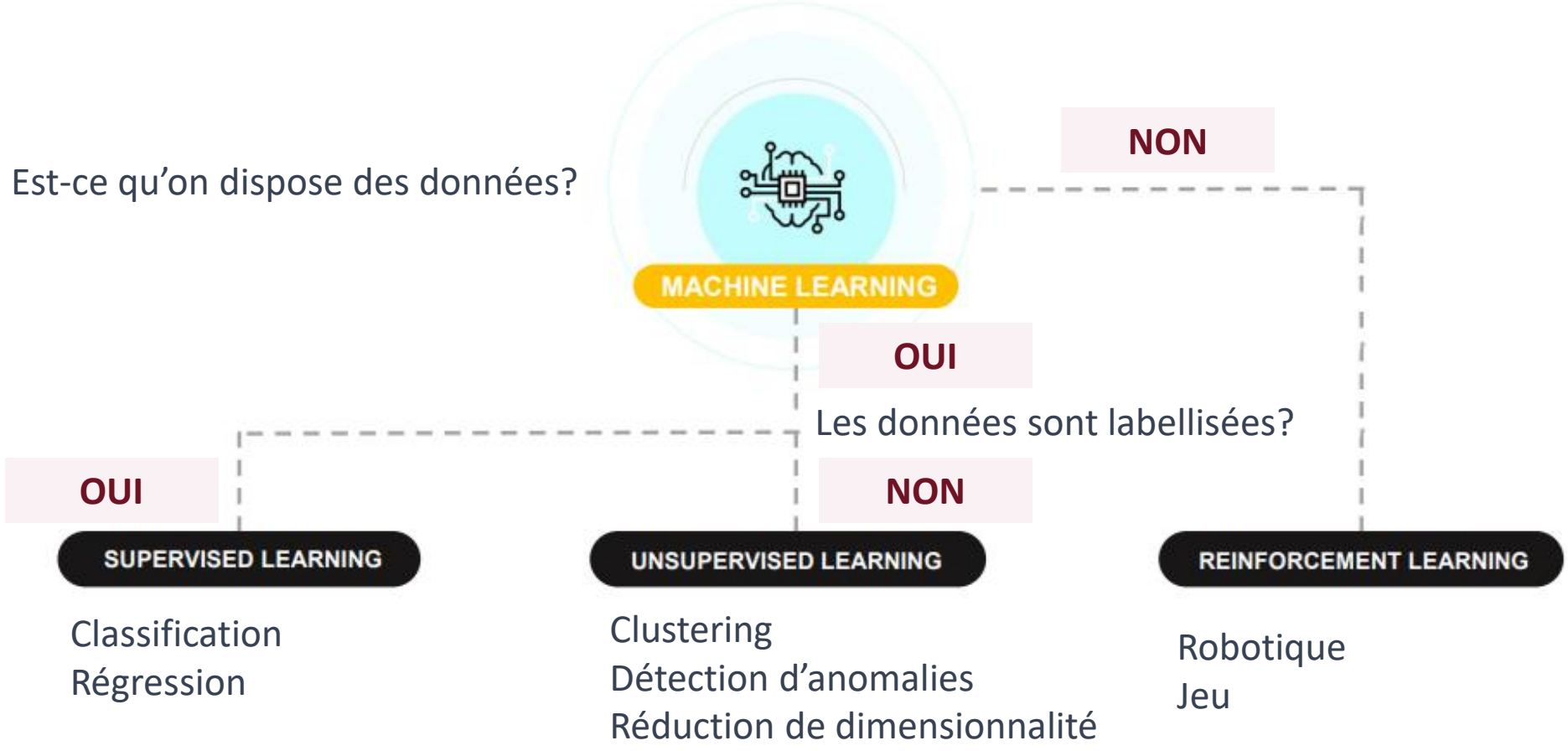
Prédiction des prix



Segmentation

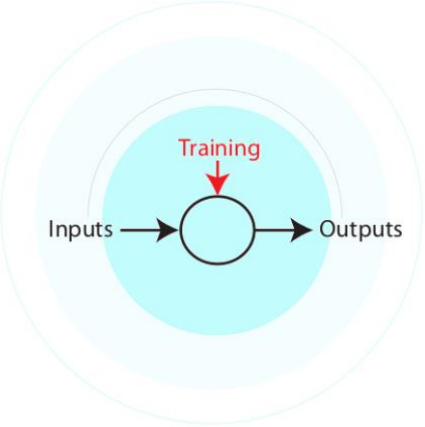
Types de problèmes de machine learning

Le machine learning est un champ assez vaste, et nous dressons dans cette section une liste des plus grandes classes de problèmes:



Apprentissage supervisé

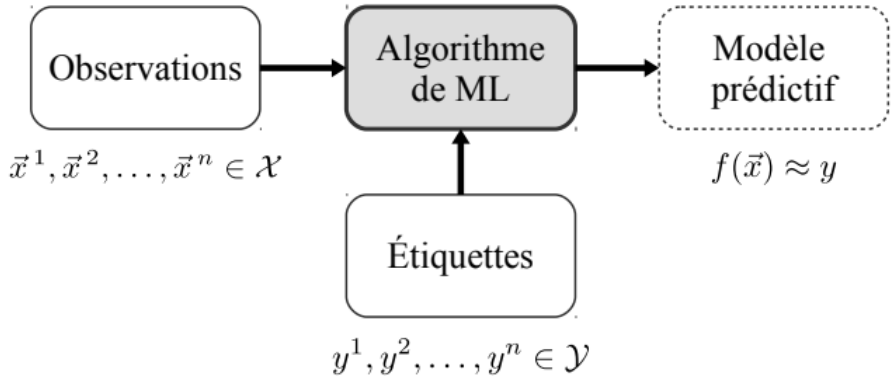
Avec l'apprentissage supervisé, vous alimentez le système avec la sortie de votre algorithme. Cela signifie qu'en apprentissage supervisé, la machine connaît déjà la sortie de l'algorithme avant de commencer à travailler dessus ou à l'apprendre. Un exemple basique de ce concept serait un étudiant apprenant un cours auprès d'un instructeur. L'étudiant sait ce qu'il/elle apprend du cours.



Avec la sortie de l'algorithme connue, tout ce qu'un système doit faire est de déterminer les étapes ou le processus nécessaires pour passer de l'entrée à la sortie.

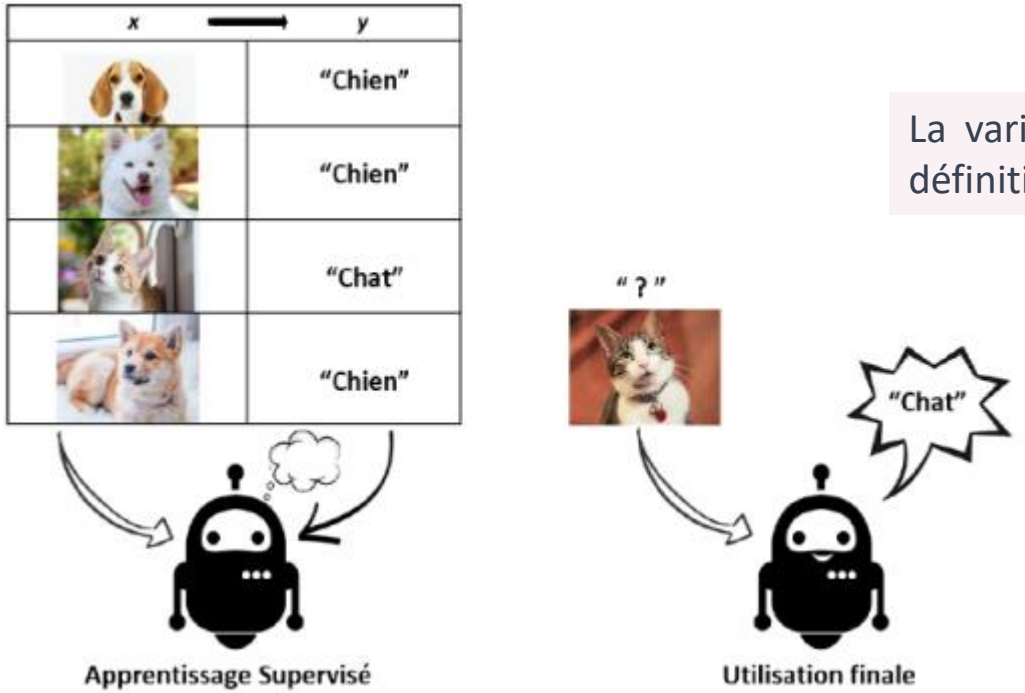
L'algorithme est enseigné à travers un ensemble de données d'entraînement qui guide la machine. Si le processus déraille et que les algorithmes produisent des résultats complètement différents de ce qui devrait être attendu, les données d'entraînement jouent leur rôle pour guider l'algorithme de nouveau sur la bonne voie.

La variable d'entrée (x) est utilisée pour se connecter à **la variable de sortie (y)** grâce à l'utilisation d'un algorithme. Toutes les données d'entrée, de sortie, l'algorithme et le scénario sont fournis par des humains.



Apprentissage supervisé: Classification

Il s'agit de l'apprentissage d'une tâche par un modèle en utilisant des données préalablement annotées, labellisées ou catégorisées. Pour ce faire, les descripteurs extraits de ces données sont introduits en entrée du modèle, et la sortie obtenue est ensuite comparée aux labels associés, qui sont également fournis. Cette comparaison permet d'ajuster les paramètres du modèle afin de les rendre plus adaptés à la tâche spécifique.



La variable cible Y est une variable qualitative dont le domaine de définition est fini.

Apprentissage supervisé: Classification

Classification binaire: L'espace des labels est binaire càd Y={0,1}
Exemple:
• Identifier si un email est un spam ou non
• Identifier si un tableau a été peint par Picasso ou non ;
• Identifier si une image contient ou non une girafe ;
• Identifier si une molécule peut ou non traiter la dépression ;
• Identifier si une transaction financière est frauduleuse ou non.

Classification multi classe : les labels sont discrets càd Y={0,1,2…….C} avec C est le nombre des classes
Exemple:
• Identifier en quelle langue un texte est écrit ;
• Identifier l'expression d'un visage parmi une liste prédéfinie de possibilités (colère, tristesse, joie, etc.) ;
• Identifier à quelle espèce appartient une plante ;
• Identifier les objets présents sur une photographie.

Apprentissage supervisé: Régression

Dans le cadre de la régression, le processus consiste à former un modèle pour prédire une variable continue en utilisant des données préalablement annotées ou mesurées. Les caractéristiques extraites de ces données sont introduites en entrée du modèle, qui tente de modéliser la relation entre ces caractéristiques et la variable cible continue. La sortie obtenue est ensuite comparée aux valeurs réelles de la variable cible, fournies dans les données d'entraînement. Cette comparaison permet d'ajuster les paramètres du modèle afin d'optimiser sa capacité à prédire la variable continue.



La variable cible Y est une variable quantitative dont le domaine de définition est infini.

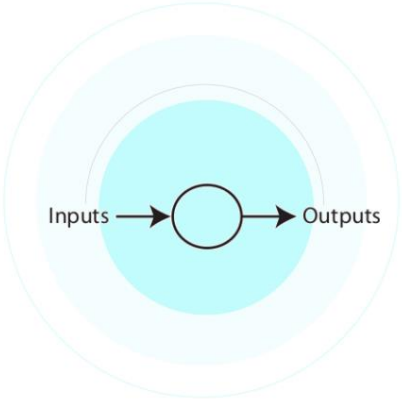
Apprentissage supervisé: Régression

Dans le cas où les étiquettes sont à valeurs réelles, on parle de régression. l'espace des étiquettes est Y = R.
Exemple:

- Prédire le nombre de clics sur un lien ;
- Prédire le nombre d'utilisateurs et utilisatrices d'un service en ligne à un moment donné ;
- Prédire le prix d'une action en bourse ;

Apprentissage non-supervisé

Pendant le processus d'apprentissage non supervisé, le système ne dispose pas de jeux de données concrets: les résultats des observations du jeux de données sont inconnus. En d'autres termes, l'objectif d'apprentissage automatique est aveugle lorsqu'ils entrent en action.



En apprentissage non supervisé, le but est de construire un modèle qui trouve par lui-même des points communs et des interactions entre les données d'entrée.

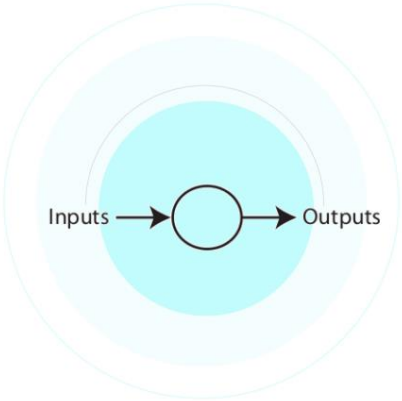
L'objectif principal est de découvrir des motifs, des structures ou des informations cachées au sein des données sans avoir la guidance explicite d'un ensemble de données d'entraînement étiqueté.

La variable d'entrée (x) est utilisée pour construire un ensemble réduit ou des regroupements d'observation grâce à l'utilisation d'un algorithme. Toutes les données d'entrée, l'algorithme et le scénario sont fournis par des humains. Les résultats doivent être interpréter par des humains.

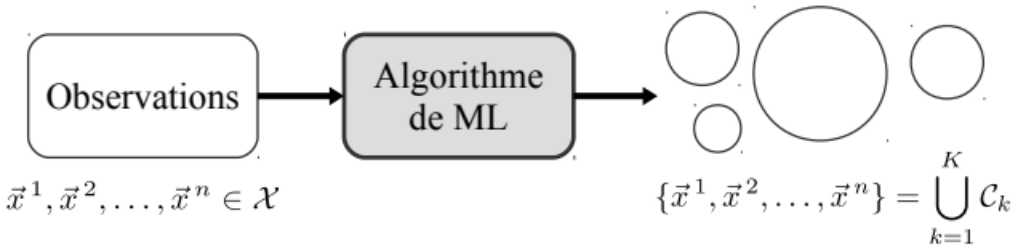


Apprentissage non-supervisé: Clustering

Dans le contexte du clustering, il s'agit de l'apprentissage d'une tâche par un modèle en utilisant des données qui ne sont pas préalablement annotées ou labellisées. Les caractéristiques extraites de ces données sont introduites en entrée du modèle, qui cherche à regrouper les observations similaires dans des ensembles distincts appelés clusters. Contrairement à la classification ou à la régression, le modèle de clustering n'est pas fourni avec des labels spécifiques pour guider le processus.



Le clustering, ou partitionnement, consiste à identifier des groupes dans les données. Cela permet de comprendre leurs caractéristiques générales, et éventuellement d'inférer les propriétés d'une observation en fonction du groupe auquel elle appartient.



Apprentissage non-supervisé: Clustering

Exemple:

• La segmentation de marché consiste à identifier des groupes de clients ayant un comportement similaire. Cela permet de mieux comprendre leur profil, et cibler une campagne de publicité, des contenus ou des actions spécifiquement vers certains groupes.

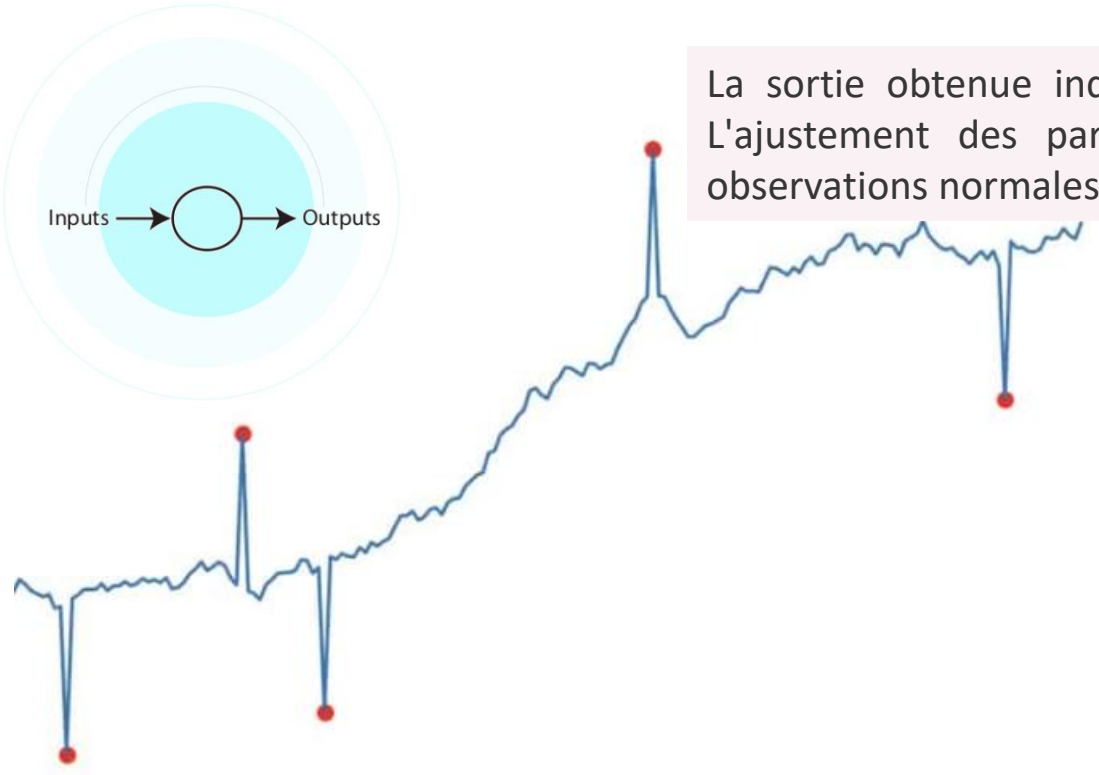
• Identifier des groupes de documents ayant un sujet similaire, sans les avoir au préalable étiqueté par sujet. Cela permet d'organiser de larges banques de textes.

• La segmentation d'image consiste à identifier les pixels d'une image appartenant à la même région.

• Identifier des groupes parmi les patients présentant les mêmes symptômes permet d'identifier des sous-types d'une maladie, qui pourront être traités différemment.

Apprentissage non-supervisé: Détection d'anomalies

Dans le cadre de la détection d'anomalies, il s'agit de l'apprentissage d'une tâche par un modèle en utilisant des données, où l'objectif est d'identifier des points ou des instances inhabituelles ou atypiques. Les caractéristiques extraites de ces données sont introduites en entrée du modèle, qui cherche à détecter les anomalies en se basant sur les schémas ou les comportements normaux présents dans les données.

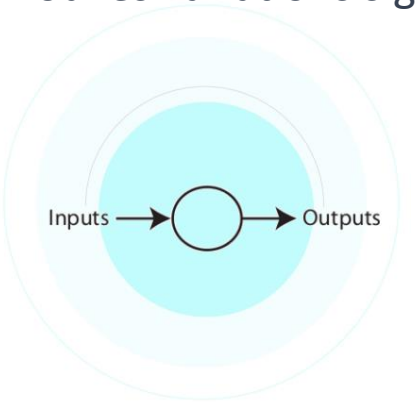


La sortie obtenue indique les points identifiés comme étant des anomalies par le modèle. L'ajustement des paramètres du modèle se concentre sur la capacité à discriminer les observations normales de celles considérées comme des anomalies



Apprentissage non-supervisé: Réduction de dimensionnalité

Dans le contexte de la réduction de dimensionnalité, il s'agit de l'apprentissage d'une tâche par un modèle en utilisant des données, avec pour objectif de diminuer le nombre de caractéristiques tout en préservant au mieux l'information importante contenue dans ces données. Les caractéristiques extraites de ces données, initialement de dimension élevée, sont introduites en entrée du modèle, qui cherche à créer une représentation réduite tout en préservant les structures ou les variations significatives présentes dans les données.

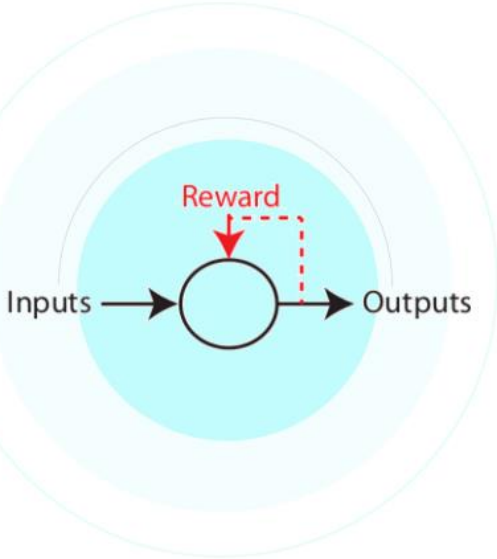


Contrairement à la classification ou à la régression, la réduction de dimensionnalité n'implique pas directement des labels ou des catégories pour guider le processus. Au lieu de cela, le modèle recherche des représentations condensées des données qui capturent l'essentiel de l'information tout en éliminant les redondances ou le bruit.

La sortie obtenue représente la nouvelle représentation des données après réduction de dimensionnalité. L'ajustement des paramètres du modèle se concentre sur la capacité à projeter efficacement les données en entrée dans un espace de dimension réduite, tout en minimisant la perte d'information importante.

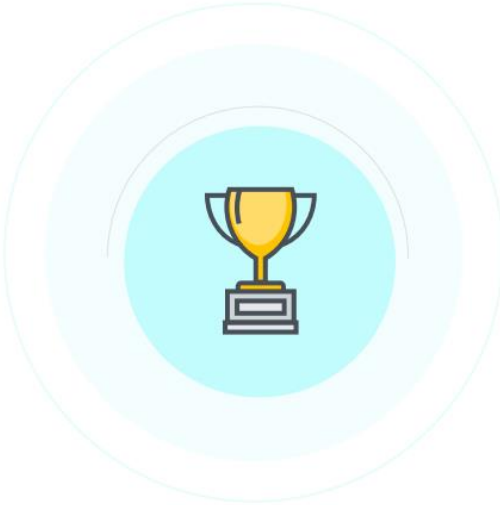


Apprentissage par renforcement

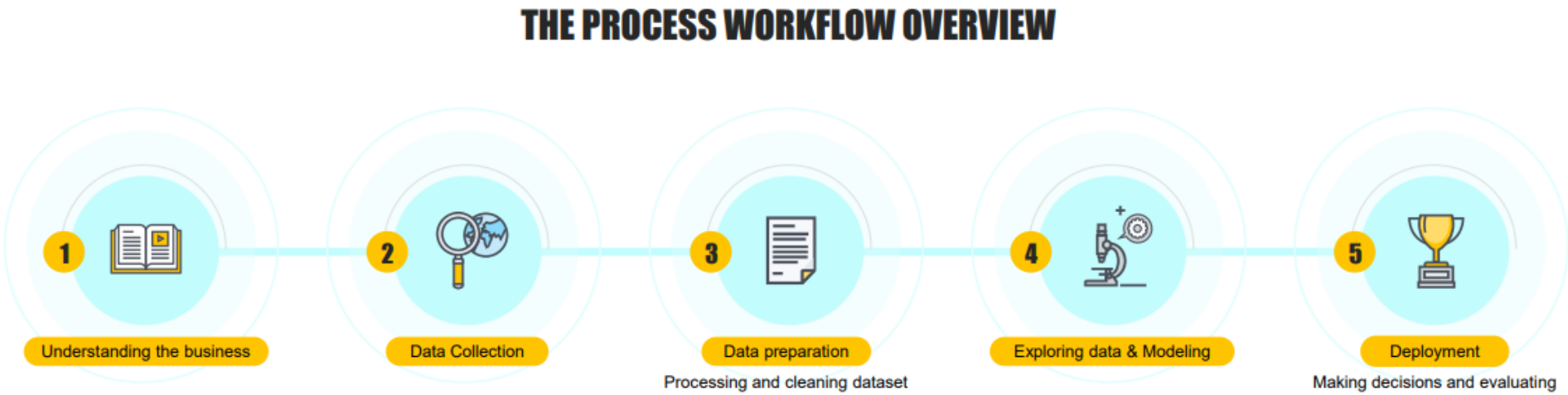


Dans le cadre de l'apprentissage par renforcement, le système d'apprentissage peut interagir avec son environnement et accomplir des actions. En retour de ces actions, il obtient une récompense, qui peut être positive si l'action était un bon choix, ou négative dans le cas contraire. La récompense peut parfois venir après une longue suite d'actions ; c'est le cas par exemple pour un système apprenant à jouer au go ou aux échecs.

Ainsi, l'apprentissage consiste dans ce cas à définir une politique, c'est-à-dire une stratégie permettant d'obtenir systématiquement la meilleure récompense possible. Les applications principales de l'apprentissage par renforcement se trouvent dans les jeux (échecs, go, etc) et la robotique. Ce sujet dépasse largement le cadre de ce cours.



Modélisation d'un problème d'apprentissage



CRISP-DM Methodology

(CRoss-Industry Standard Process for Data Mining)

Modélisation d’un problème d’apprentissage

La science de données est une méthode d’extraction de connaissance à partir de données structurées et non-structurées en utilisant des algorithmes de la statistique et de l’apprentissage automatique.

CRISP-DM: est une méthodologie de planification et de gestion de projet data:
— Elle est très efficace pour les projets Data
— CRISP-DM est une méthodologie Agile itérative. Dans chaque itération on aura des nouvelles information et connaissances métier, ce qui rend le passage à l’étape suivante plus facile.

CRISP-DM: les étapes



— **Business Understanding:** Compréhension des objectifs du projet et des exigences, définition du problème de fouille de données.

— **Data Understanding:** Collecte initiale des données et familiarisation, identification des problèmes de qualité des données.

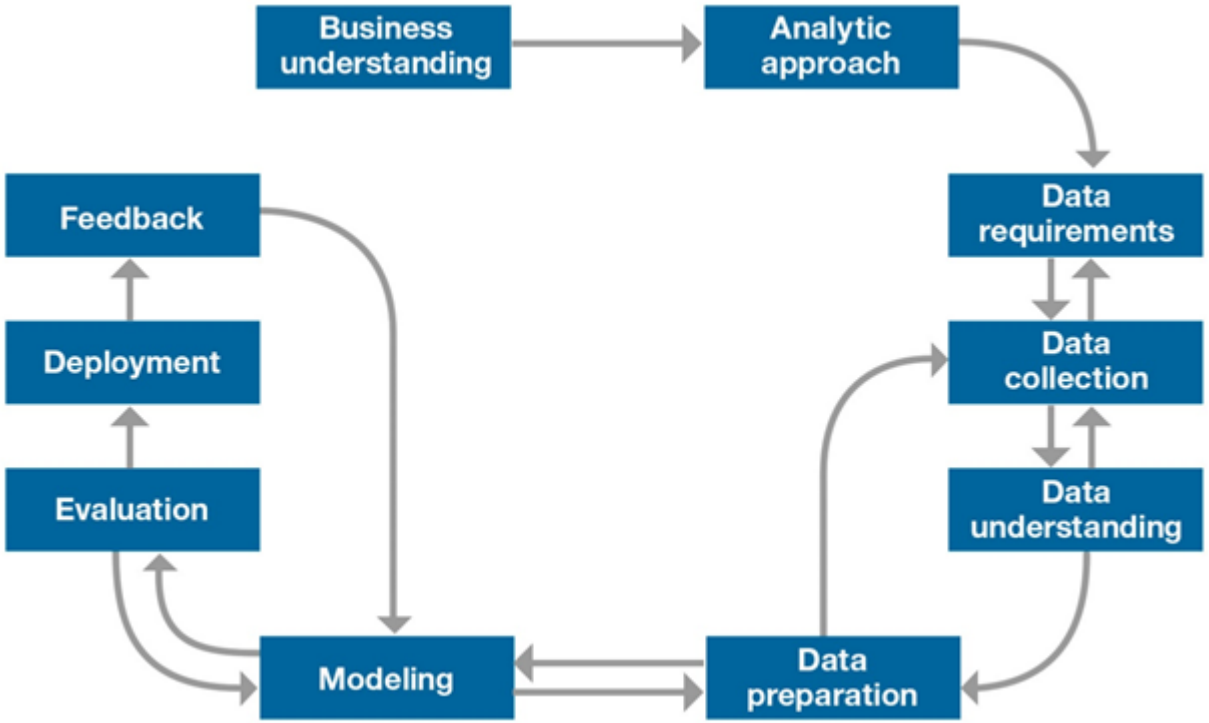
— **Data Preparation:** Sélection des tables, enregistrements et attributs, transformation et nettoyage des données.

— **Modeling:** Sélection et application des techniques de modélisation, Calibration des paramètres.

— **Evaluation:** Évaluation de la réalisation des objectifs commerciaux et des problématiques.

— **Deployment:** Déploiement du modèle de résultats, Mise en œuvre du processus de fouille de données reproductible

IBM Master Plan



— **Business Understanding:** permet de déterminer quelles données seront utilisées pour répondre à la question centrale.
— **Analytic approach:** aide à limiter l'algorithme (ou les algorithmes) qui seront utilisés pendant la modélisation (modèle prédictif / modèle descriptif).
— **Data Requirements:** Identification du contenu, des formats et des sources de données nécessaires pour la collecte initiale des données.
— **Data Understanding:** Représenter les données collectées en fonction du problème que nous souhaitons résoudre.

— **Data Preparation:** De nombreuses opérations telles que le traitement des valeurs manquantes ou invalides et la suppression des doublons. Cette étape prend généralement près de 90 % du temps total du projet.
— **Modeling:** Génération du modèle basé sur l'approche analytique adoptée.
— **Evaluation:** C'est l'étape où nous vérifions si le modèle que nous avons généré répond ou non à la demande initiale.
— **Deployment:** Déploiement du modèle de résultats, Mise en œuvre du processus de fouille de données reproductible

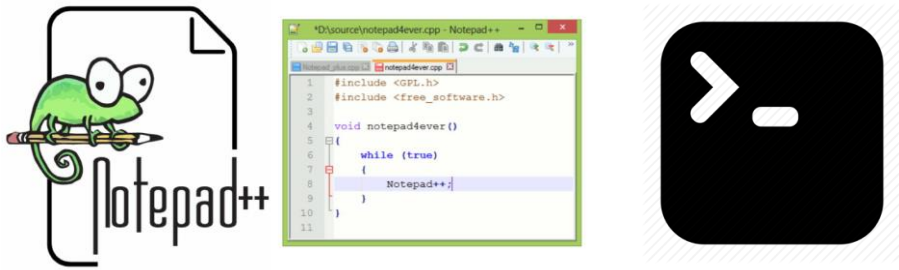
Environnement Python

L'environnement python:
— Anaconda (https://www.anaconda.com/)
— Jupyter Notebook (https://jupyter.org/)

Alternatives:
— Editeur de texte + Ligne de commande
— IDE (Integrated Development Environment): PyCharm, Vscode…







C’est quoi Anaconda?

Anaconda est le logiciel libre le plus facile pour réaliser des opérations de science de données (Python/R) sous Linux, Windows et Mac OS. Avec plus que 19 millions d’utilisateurs.

C’est le logiciel standard dans l’industrie pour développer, tester, et entrainer des modèles sur une seule machine, il permet aux data scientist de:

— Télécharger rapidement plus que 7,500 data science packages,
— Analyser d’une manière évolutive les données, avec une performance élevée en utilisant Dask, Numpy, pandas, et Numba
— Visualiser les résultats en utilisant Matplotlib, Bokeh, Datashader, and Holoviews
— Modéliser et entrainer des algorithmes d’apprentissage automatique avec scikit-learn, tenserflow, and Theano,

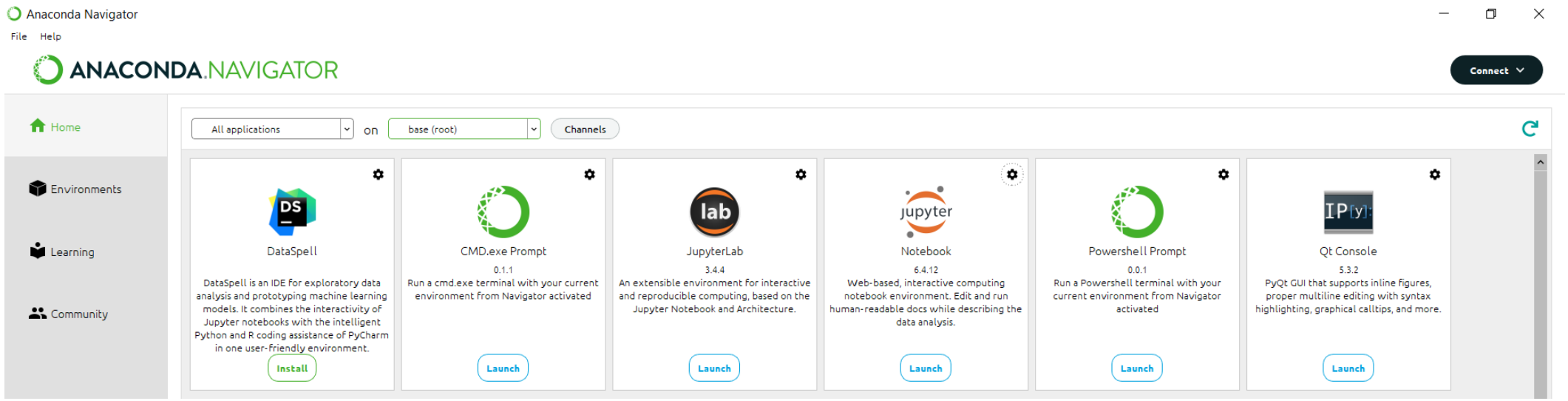


C’est quoi Jupyter?

Jupyter notebook est une application web libre qui permet la création et le partage de documents qui contient des codes sources, équations, visualisation, et du texte.

Les documents peuvent inclure des instructions de data cleaning, des trasformations, des simulations, modélisation statistique, visualisation de données, apprentissage automatique, et autres …

Jupyter notebook est inclus dans Ancaonda



Bibliothèques

Data Processing
— NumPy
— Pandas
— SciPy







Visualisation
— matplotlib
— Seaborn





Apprentissage automatique
— Scikit-learn
— Theano/ Tensorflow / Pytorch

