Machine learning, data science et intelligence artificielle

Table of Contents

Première version : 2018-06-20

1. Expliquer le machine learning en termes simples

a. Une comparaison avec les statistiques classiques

Nous allons comparer le machine learning (**en français: apprentissage automatique ou apprentissage machine**) à un exemple classique de la statistique: calculer une ligne de régression pour identifier une tendance dans un nuage de points. Pour illustrer, nous prenons quelques données sur les budgets de marketing et les chiffres de vente dans la période correspondante :

```
<iframe width="600" height="371" seamless frameborder="0" scrolling="no"
```

```
src="https://docs.google.com/spreadsheets/d/e/2PACX-
1vS8dKfwxvgz3ALH8Y1FzxWk9lZtiVBlQdZYUrKJqRXNqBFRjKIP3LUvv29QSIBbGx2-
ray5nK8cALMH/pubchart?oid=1075418595&format=interactive"></iframe>
```

Les "statistiques régulières" permettent, entre autres :

1. trouver la relation numérique entre les deux séries, basée sur un modèle formel pré-établi (par exemple, la méthode des moindres carrés ordinaire).

→ on voit que les ventes sont corrélées avec les dépenses marketing. Il est probable que plus de dépenses de marketing entraîne plus de ventes.

2. prédire, sur la base de ce modèle:

→ en traçant la ligne plus loin (en utilisant le modèle formel), nous pouvons prédire l'effet de plus de dépenses de marketing

Les "statistiques régulières" sont le domaine de travail de scientifiques qui:

1. sont hautement qualifiés en mathématiques

→ leur objectif est de trouver l'expression mathématique exacte définissant la situation, dans des conditions rigoureuses

→ une approche clé est **l'inférence**: en définissant un **échantillon des données** de la taille juste, nous pouvons arriver à des conclusions valables pour l'ensemble de données entier.

2. n'ont pas de formation en informatique

→ ils ne sont pas concernés par la difficulté d'éxecution de leurs modèles sur des ordinateurs, en termes de calculs à effectuer.

→ puisqu'ils se concentrent sur **l'échantillonnage** des données, ils ne sont pas concernés par le traitement de jeux de données entiers avec les problèmes informatiques que cela pose : coût, lenteur, complexité technique de mise en oeuvre…

Le **machine learning** fait des choses similaires à la statistique, mais d'une manière légèrement différente :

- on met l'accent sur la bonne prédiction, sans se soucier d'identifier le modèle mathématique sous-jacent
- la prévision doit être réalisable dans le temps disponible, avec les ressources informatiques disponibles
- les données d'intérêt sont dans un format / dans un volume qui n'est pas couramment traité par les outils de statistiques standards (par exemple : images, vidéos, observations comportant des centaines de paramètres…)

L'apprentissage automatique est le domaine de travail de scientifiques qui sont typiquement :

1. hautement qualifiés en statistiques (les statistiques "classiques" que nous avons vu ci-dessus)

2. avec une formation ou une expérience en informatique, familiers du travail sur données non structurées / big data
3. habitués au travail dans des environnements (industriels, militaires, ...) où les aspects opérationnels du problème sont des déterminants clés (données non structurées, limites sur les ressources informatiques)

L'apprentissage automatique met l'accent sur les techniques qui sont «adéquates en termes de calculs» :

- qui ont besoin des opérations algébriques minimales / les plus simples à exécuter : la meilleure technique est sans valeur si elle est trop longue ou trop chère à calculer.
- qui peut être exécuté de telle sorte que plusieurs ordinateurs travaillent en parallèle (simultanément) pour le résoudre.

(note de bas de page: donc l'apprentissage automatique, à mon avis, partage l'esprit de "faire avancer les choses" comme l'était la recherche opérationnelle pendant la Seconde Guerre mondiale)

La recherche en statistiques traditionnelles n'ignore pas la question de la charge de calcul ("computational effiency") - elle est considérée comme une propriété souhaitable - mais dans l'apprentissage automatique, c'est en grande partie un pré-requis.

b. Une illustration: le cas des GPUs

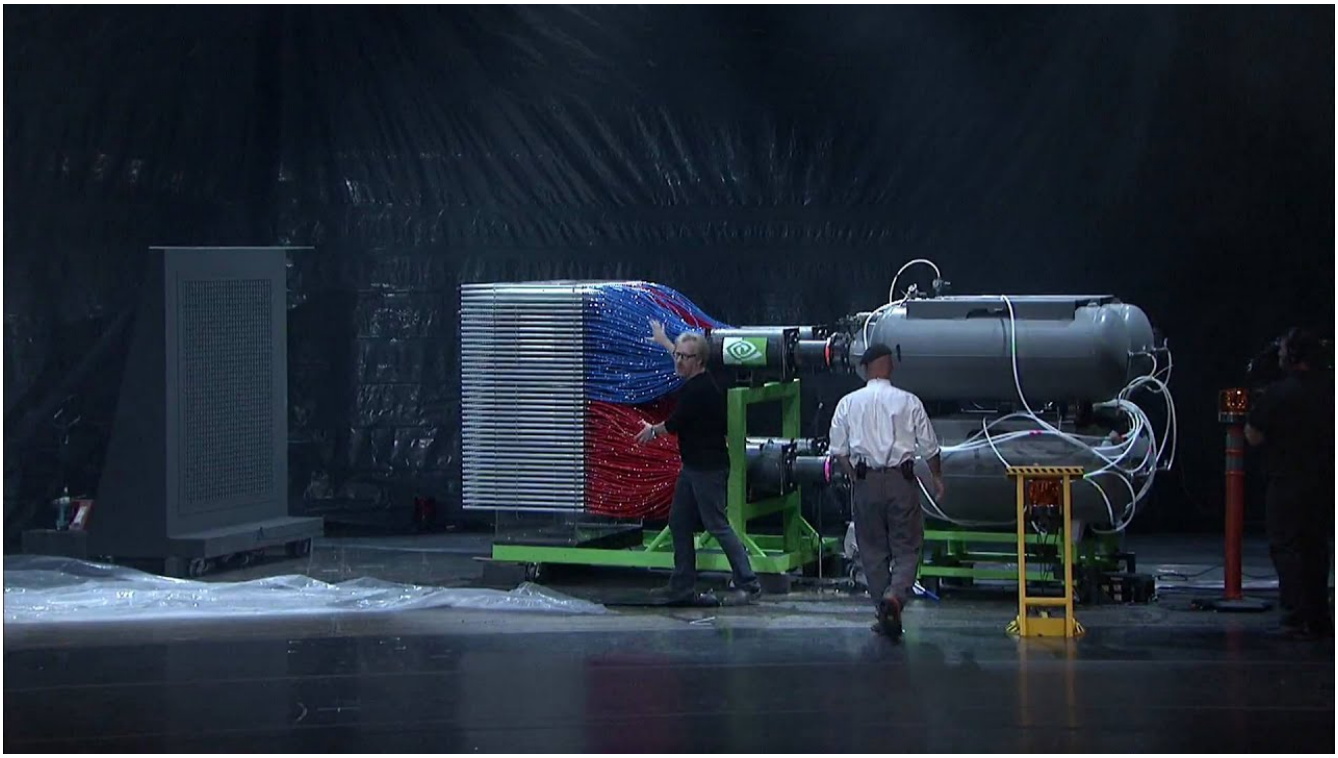
Une illustration clé de la différence entre les statistiques et l'apprentissage automatique peut être fournie avec l'utilisation de **cartes graphiques**. Les cartes graphiques (ou GPU: unités de traitement graphique) sont ces composants électroniques pleins de puces trouvées dans un ordinateur, qui sont utilisées pour l'affichage d'images et de vidéos sur des écrans :

Figure 1. Une carte graphique vendue par NVidia- un des principaux fabricants

Dans les années 1990, le jeu vidéo s'est beaucoup développé, des consoles et arcades aux ordinateurs de bureau. Les développeurs de jeux ont créé des jeux informatiques montrant des scènes et des animations de plus en plus complexes. (voir une évolution des graphiques en jeux vidéo, et les jeux graphiques avancés en 2017). Ces jeux vidéo ont besoin de puissantes cartes vidéo (aussi appelés processeurs graphiques ou GPU) pour restituer des scènes complexes dans les moindres détails - avec des calculs sur les effets de lumière et les animations **réalisés en temps réel**. Cela a poussé au développement de **GPUs** plus puissants. Leurs caractéristiques sont qu'ils peuvent calculer des opérations simples pour changer les couleurs des pixels, **pour chacun des millions de pixels de l'écran en parallèle**, de sorte que la prochaine séquence de l'image peut être affichée en millisecondes.

Des millions d'opérations simples se déroulent en parallèle pour le prix d'un GPU (quelques centaines de dollars), pas le prix de douzaines d'ordinateurs fonctionnant en parallèle (peut être des dizaines de milliers de dollars)?

→ C'est intéressant pour les calculs sur les big data! Si un problème statistique de prédiction peut être décomposé en opérations simples pouvant être exécutées sur un GPU, alors un grand ensemble de données peut être analysé en secondes ou en minutes sur un ordinateur portable, au lieu d'un cluster d'ordinateurs. Pour illustrer la différence de vitesse entre une opération mathématique exécutée sans ou avec un **GPU**:



Le problème est le suivant : pour utiliser un GPU pour les calculs, vous devez conceptualiser le problème pour qu'il soit :

- décomposé en une très grande série…
- … d'opérations très simples (fondamentalement, des sommes ou des multiplications, rien de complexe comme des racines carrées ou des polynômes)

- … qui peuvent fonctionner indépendamment les uns des autres.

→ alors, les calculs vont pouvoir se faire sur un GPU, ce qui peut accélérer le traitement par 10x, 100x ou plus.

→ L'apprentissage automatique ou machine learning prête attention à ces dimensions du problème dès la phase de conception des modèles et des techniques, là où les statistiques "classiques" ne considèrent généralement pas le problème, ou seulement en aval : non pas au stade de la conception mais à la phase de mise en œuvre - ce qui est souvent trop tard.

Maintenant que nous avons vu comment les statistiques et l'apprentissage machine diffèrent dans leur approche, nous devons encore comprendre comment l'apprentissage automatique obtient de bons résultats, s'il ne repose pas sur la modélisation / l'échantillonnage des données comme le font les statistiques.

L'apprentissage automatique peut être catégorisé en 3 familles :

2. Trois familles d'apprentissage automatique

a. L'apprentissage non supervisé

Apprentissage non supervisé désigne les méthodes qui utilisent un jeu de données nouveau et y trouvent des modèles intéressants, **sans que cela ne soit par apprentissage sur de précédents ensembles de données similaires**.

Comment l'apprentissage supervisé fonctionne-t-il ? Prenons un exemple. Dans une réception de mariage, comment asseoir des gens avec des intérêts similaires aux mêmes tables?

Les données initiales du problème :

- une liste de 100 invités, et pour chaque invité, une liste de 3 goûts que vous connaissez d'eux
- 10 tables avec 10 sièges chacune.
- une mesure de similitude entre 2 invités: 2 invités ont une similitude de 0% s'ils partagent 0 goût, 33% s'ils partagent 1 goût, 66% avec 2 goûts en commun, 100% avec trois intérêts correspondants.
- une mesure de similitude au niveau d'une table : la somme des similitudes entre toutes les paires d'invités à la table (45 paires possibles pour une table de 10).

Une solution possible au problème peut être apportée en utilisant une approche non supervisée :

1. Sur un ordinateur, assigner au hasard les 100 invités aux 10 tables.
2. prendre une table :
    - mesurer le degré de similitude des goûts pour la table
    - échanger le siège de 1 personne à cette table, avec le siège d'une personne à une table

différente.

- mesurer à nouveau le degré de similarité de la table: si elle s'est améliorée (parce que maintenant, les personnes à cette table ont plus de goûts en commun), alors garder les nouvelles assises. Sinon, annuler l'échange de place et revenir à la situation avant l'échange.

3. Répéter l'étape 2 pour toutes les tables, plusieurs fois, jusqu'à ce que plus aucun échange de sièges n'améliore le degréé de similitude à aucune table. Lorsque cette étape est atteinte, nous disons que le modèle a "**convergé**".

Cette approche permet d'identifier des groupes de personnes qui ont des points communs. C'est évidemment d'une grande utilité pour organiser des données, depuis une segmentation de clientèle ou de prospects, jusqu'à une classification de produits en catégories à des fins d'évaluation ou de gestion de portefeuille.

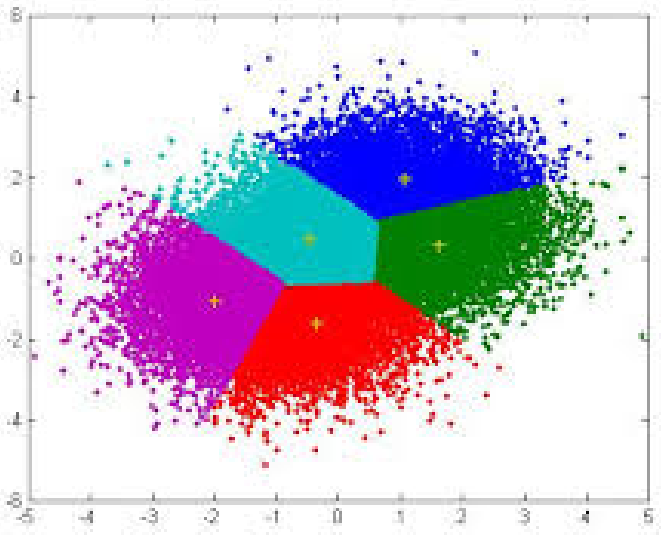


Figure 2. K-means, une approche d'apprentissage non supervisée

b. L'approche d'apprentissage supervisé

L'**apprentissage supervisé** est l'approche consistant à calibrer un modèle basé sur l'histoire des expériences passées afin de deviner / prédire une nouvelle occurrence de la même expérience. Prenons l'exemple suivant : comment faire pour qu'un ordinateur "devine" si une image représente un chat ou un chien? Pour cela, en approche supervisée, nous allons commencer par récolter 50000 images ou plus de chats et de chiens, avec leurs légendes associées, comme ceci:

- une image d'un chat, avec la légende "chat"
- une image d'un chien, avec la légende "chien"
- une autre image d'un chat, avec la légende "chat"

etc....

- Ces 50000 images et leur légende s'appelle le *training set*..

- Ceci est aussi appelé un ensemble de données **annotées**, ce qui signifie que nous avons une étiquette décrivant chacune des observations (en anglais : *labelled set*).

Dans un jeu de données libellé, d'où viennent les étiquettes?

- les étiquettes peuvent être fournies par les utilisateurs d'un service. Par exemple, les photos sur Instagram légendées par des hashtags sont exactement cela: une image avec une étiquette. L'étiquetage est fait par les utilisateurs d'Instagram affichant les photos et en écrivant les hashtags ci-dessous. Instagram est un service gratuit, mais les jeu de données libellées qu'il crée sont d'une grande valeur pour une entreprise comme Instagram (et pour Facebook, qui a racheté Instagram).
- ils peuvent être produits par des travailleurs humains. En pratique, les humains sont payés quelques centimes par image qu'ils doivent étiqueter (est-ce un chat? Est-ce un chien? Etc.). Une grande industrie et un marché du travail associé se développent pour effectuer une variété de tâches de ce genre. Une main-d'œuvre croissante fournit leur travail numérique aux entreprises qui ont besoin de **l'annotation des données**(données, annotation des données ou de **nettoyer, classer ou qualifier les données**. Voir le travail de Antonio Casilli sur ces sujets.

La tâche est la suivante: si nous donnons à notre ordinateur une nouvelle image d'un chat *sans étiquette*, pourra-t-il deviner l'étiquette "chat"?

La méthode:

- prendre une liste de coefficients aléatoires (en pratique, la liste est un vecteur, ou une matrice).
- pour chacune des 50 000 photos de chiens et de chats:
  - appliquer les coefficients à l'image à portée de main (disons que nous avons un chien ici)
  - Si le résultat est "chien", ne faites rien, ça marche!
  - Si le résultat est "chat", modifiez légèrement les coefficients.
  - passer à l'image suivante
- Après avoir parcouru 50 000 images en boucle, les paramètres ont été ajustés et réglés. C'était **l'entraînement du modèle**.

Maintenant, lorsque vous présentez une nouvelle image au logiciel que vous venez d'entraîner, l'application du modèle devrait produire une prédiction correcte ("chat" ou "chien").

L'apprentissage supervisé est actuellement la famille d'apprentissage automatique la plus populaire et obtient d'excellents résultats notamment en reconnaissance d'image, même si certains cas restent difficiles à résoudre:

Figure 3. Un cas de test difficile pour l'apprentissage supervisé

(source)

C'est donc ce qu'on appelle l'apprentissage **supervisé** car l'apprentissage est guidé, dirigé, encadré par des exemples passés.

Trois conditions à retenir sur l'apprentissage supervisé :

- pour que l'apprentissage supervisé soit possible, **il est nécessaire de disposer de grands ensembles de données pour la phase d'entraînement**. Sans ces données, pas d'apprentissage supervisé.
- l'apprentissage supervisé **permet d'analyser des situations similaires à celles représentées dans le jeu de données sur lequel l'apprentissage a été entraîné**. Un modèle entraîné sur 50,000 photos de chats et de chiens ne saura pas reconnaître un dauphin.
- les données d'apprentissage doivent être spécifiques. Si l'on souhaite apprendre à un algo à reconnaître un chihuhua, le training set doit être fait de chihuahuas - plutôt que des chiens de toutes races.

Ce dernier point est explicité par Maryne Cotty-Eslous, fondatrice de Lucine, une app de reconnaissance et d'analyse de la douleur:

[] | *https://img.youtube.com/vi/tL7ojiOTQho?t=16m31s/maxresdefault.jpg*

c. L’approche de l’apprentissage par renforcement (reinforcement learning)

Pour comprendre l’apprentissage par renforcement, nous pouvons penser intuitivement comment les animaux peuvent apprendre rapidement en **ignorant** les comportements indésirables et en **récompensant** les comportements souhaitables.

C’est facile et ne prend que quelques secondes. La vidéo suivante montre B.F. Skinner, figure centrale de la psychologie comportementale dans les années 1950-1970, qui apprend à un pigeon à faire un tour sur lui-même. Pour cela, Skinner procède simplement en récompensant le pigeon par des graines, dès que le pigeon fait des mouvements de rotation. A la fin, le pigeon finit par faire un tour complet sur lui-même, car il a appris que cela allait lui donner une récompense.

[] | *https://img.youtube.com/vi/TtfQlkGwE2U/maxresdefault.jpg*

Outre les pigeons, l’apprentissage par renforcement peut être appliqué à tout type d' "agents experts". Prenons le cas d’un jeu vidéo comme Super Mario Bros:

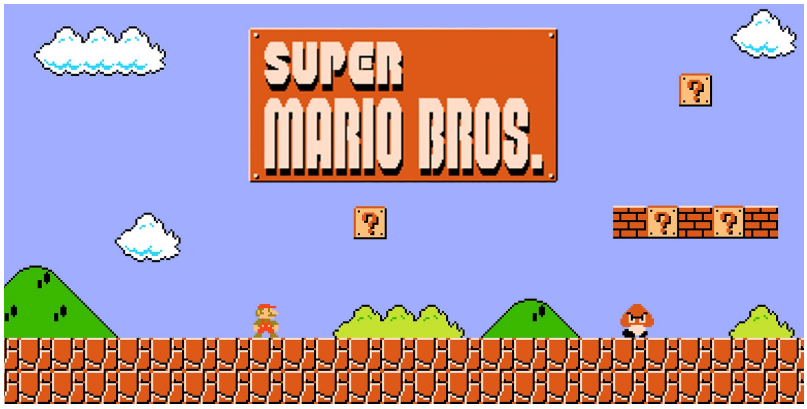


Figure 4. Mario Bros - un jeu vidéo populaire

Structure du jeu / de la tâche:

- But de la tâche : Mario doit collecter des pièces d’or et compléter le jeu en atteignant l’extrême droite de l’écran.
- Résultat négatif à éviter : se faire tuer par des ennemis ou en tombant dans des trous.
- Point de départ : Mario Bros est debout au début du jeu.
- Actions possibles : se déplacer à droite, à gauche, sauter, s’accroupir, tirer en avant.

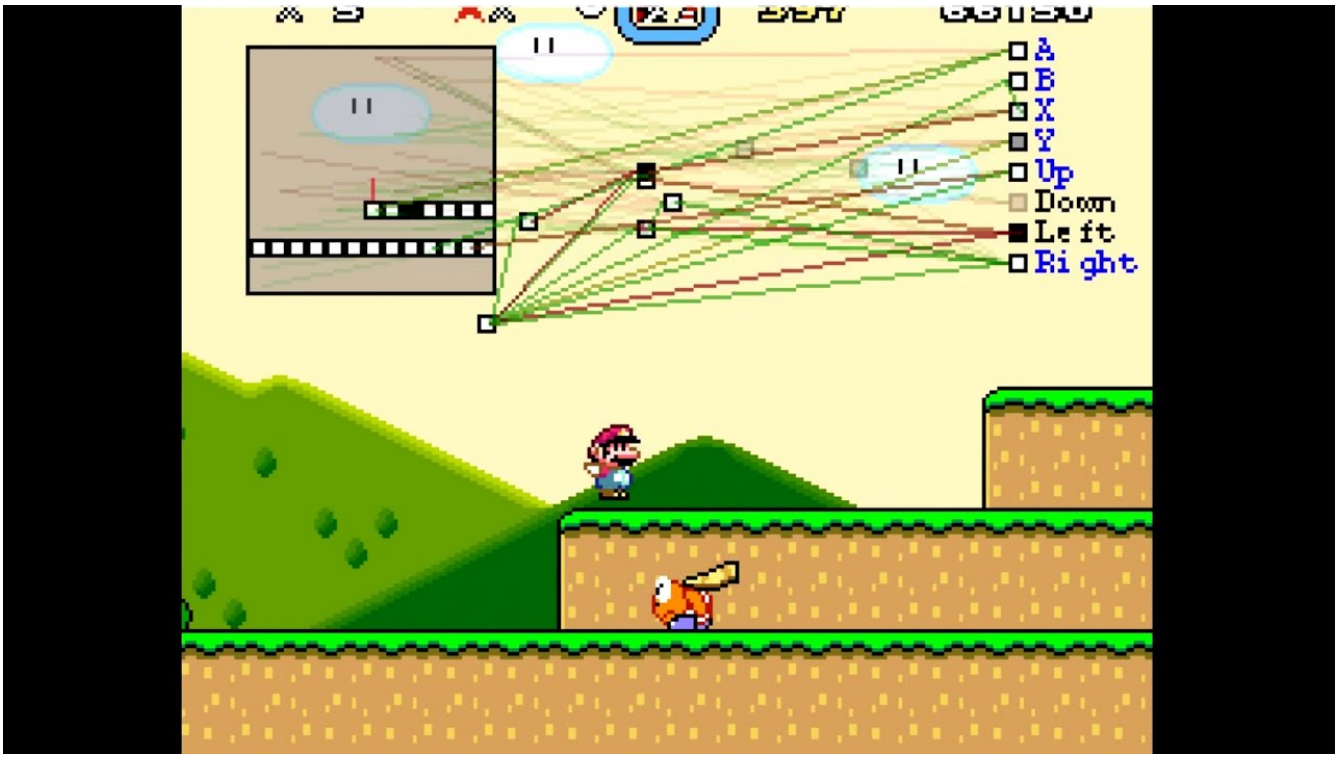
L’apprentissage par renforcement fonctionne de la manière suivante :

1. Faire faire à Mario une nouvelle action aléatoire ("essayer quelque chose"), par exemple: "déplace-toi à droite".
2. Le jeu se termine (Mario a bougé à droite, a été touché par un ennemi et est mort)
3. Ce résultat est stocké quelque part :

- se déplacer à droite ⇒ bien (on a progressé dans le jeu, même si c'est très peu). A refaire!
- marcher près d'un ennemi et être touché par celui-ci ⇒ mauvais. A éviter! Déclenchons une autre action à proximité d'un ennemi (comme "sauter en avançant", par exempe).

4. Le jeu recommence (retour à l'étape 1) avec une combinaison de :
   - actions qui ont été enregistrées comme positives lors de l'étape précédente
   - essais de nouvelles choses (sauter, tirer?) à proximité d'une situation associée à un résultat négatif au tour précédent.

Après avoir répété les étapes de 1. à 4. des milliers de fois, et enregistré à chaque fois les combinaisons d'actions favorables à répéter, et défavorables à éviter, Mario finit par arriver au bout du jeu, sans qu'aucun joueur humain ne tienne les commandes :



L'apprentissage par renforcement est perçu comme correspondant à un aspect important de l'apprentissage humain / de l'intelligence humaine (axé sur les buts, «essai et erreur»).

Maintenant, imaginons que nous créons une situation dans laquelle deux machines apprenantes se font concurrence: l'une qui contrôle Mario Bros, l'autre qui contrôle un personnage ennemi dans le jeu, et qui essaie de faire échouer Mario Bros. En les faisant combattre l'une contre l'autre des milliers de fois, les deux machines vont adapter leur comportement en apprenant de leurs erreurs. Ainsi, elles vont apprendre beaucoup plus vite et se perfectionner beaucoup plus. Ce type d'intelligence artificielle s'appelle "réseaux antagonistes génératifs" et beaucoup d'observateurs y voient une voie de progrès futur majeur pour l'IA.

d. Quand le machine learning est-il utile? Quand ne l'est-il pas?

L'utilisation de l'apprentissage automatique peut être un gaspillage de ressources, lorsque des statistiques bien connues peuvent être facilement appliquées. Des indices que la modélisation statistique "classique" (peut-être aussi simple qu'une régression linéaire) devrait suffire:

- L'ensemble de données n'est pas grand (moins de 50k observations), l'apprentissage supervisé ne fonctionnera pas
- Les données sont parfaitement structurées (données tabulaires)
- Les points de données ont peu de dimensions (chaque observation a peu d'attributs - il y a peu de "colonnes" dans une représentation sous forme de tableau)

Enfin, il existe une situation dans laquelle **le machine learning n'est absolument pas la bonne solution**. Si la question est celle du rôle de tel ou tel facteur dans la détermination d'un résultat, le machine learning reste silencieux sur ce sujet. Reprenons l'exemple des images de chats et chiens:

- l'apprentissage supervisé est très efficace pour arriver à "deviner" si une image est celle d'un chat ou d'un chien, après entraînement sur des milliers d'images de chats ou de chiens.
- mais rien ne nous dit "comment" l'apprentissage supervisé a fait. Est-ce que la longueur des moustaches, la forme des oreilles, la couleur du poil... aide à classer une image comme celle d'un chat ou d'un chien? L'apprentissage supervisé ne répond pas à ces questions.
- des travaux ont été publiés pour rendre intelligible comment l'apprentissage supervisé détermine le résultat "chat" ou "chien". Cependant ce type de travaux reste assez peu courant. L'apprentissage supervisé reste très largement une boîte noire.

3. Machine learning et data science

Le machine learning est une seule des étapes dans la longue chaîne du traitement et de l'analyse des données. Le processus du traitement et de l'analyse des données a été formalisé dans les années 1980 sous le nom de "data mining", "exploration des données", "fouille de données," ou kdd: Knowledge Discovery in Databases.

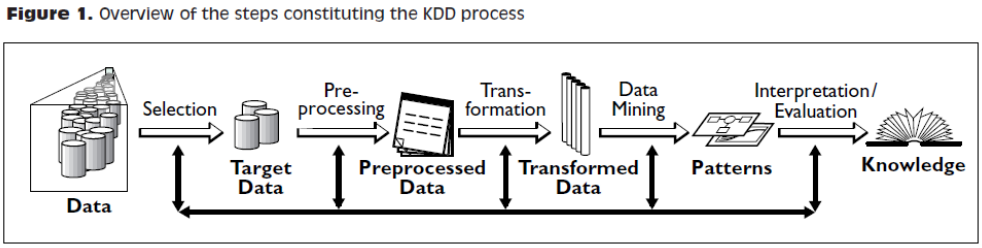


Figure 5. KDD - découverte des connaissances dans les bases de données

Des représentations plus récentes des étapes du traitement des données ont été suggérées, laissant place au rôle de la visualisation de données :

→ voir le processus de conception de l'information par Ben Fry et ce workflow de visualisation des données par Moritz Stefaner :

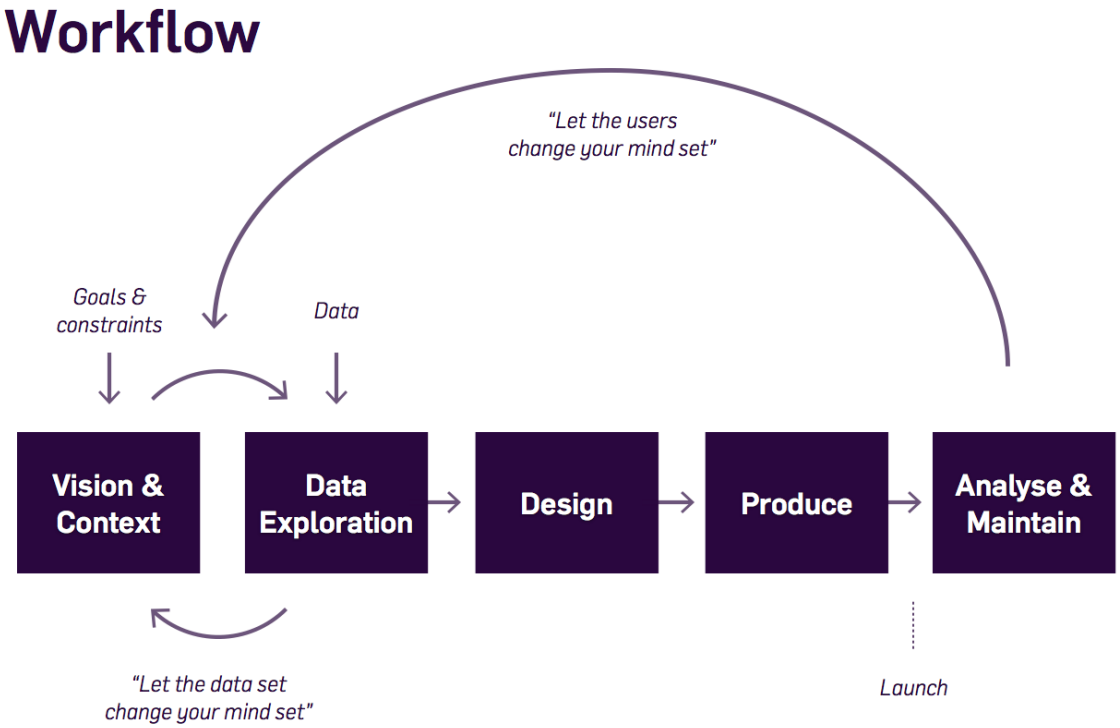


Figure 6. workflow de visualisation des données par Moritz Stefaner

- **L'apprentissage machine** est l'une des techniques intervenant à l'étape du "Data mining".
- **La data science** a) dans un sens restrictif: est synonyme de "data mining" ou b) désigne la totalité de la chaîne de traitement des données.

Pour effectuer toute la chaîne de traitement de données, une grande variété de compétences est nécessaire :

- capacité à mettre en place et gérer l'infrastructure informatique permettant de collecter, stocker et accéder à de gros volumes de données (types de compétences "ingénieur base de données", "back-end").
- capacité à appliquer des données mathématiques et des modèles statistiques aux données (compétences "data scientist", "data mining")
- capacité à communiquer efficacement les résultats (compétences en "visualisation de données", types de compétences "front-end") Les compétences d'une équipe de data scientist sont souvent représentées comme la réunion de trois domaines distincts :

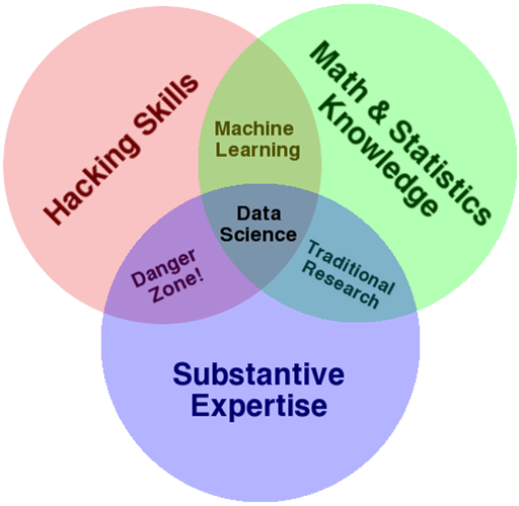


Figure 7. Le Diagramme de Venn de la science des données par Drew Conway

Ce diagramme met en évidence un point important : la science des données n'est pas simplement un ensemble de compétences en informatique et en mathématiques. Une "connaissance du domaine" ou "expertise métier" est également requise: des personnes connaissant parfaitement le contexte organisationnel et la problématique métier sous-jacente. En pratique, cet équilibre de compétences est rarement trouvé chez un seul individu. On passe donc le plus souvent par la création d' **équipes de data science** comprenant des informaticiens, des analystes et des représentants des métiers de l'entreprise.

4. Rapport entre machine learning et intelligence artificielle (faible et forte)

IA faible désigne des programmes informatiques capables de surpasser les humains dans des tâches complexes sur un domaine étroitement et précisément délimité (comme jouer aux échecs). L'IA faible fonctionne grâce à des systèmes experts ou des techniques de machine learning vues ci-dessus. L'IA que nous voyons fonctionner aujourd'hui est une IA faible: reconnaissance d'image, aide à la conduite et véhicules autonomes, chatbots, ordinateurs capables de battre des humains au jeu de GO ou à Mario, …

IA forte est une intelligence qui serait capable de résoudre des problèmes de portée générale, capable de fixer son propre but, d'être consciente d'elle-même, ou de résoudre des problèmes variés et originaux. Aujourd'hui, rien ne s'approche de cela et le consensus dit que les techniques de machine learning actuelles ne sont pas adaptées à la mise au point de ce type d'intelligence.

Donc l'IA est synonyme d'IA faible aujourd'hui, et couvre les trois familles de machine learning présentées ci-dessus.

5. Vidéos sur machine learning et intelligence artificielle

- La qualité de la donnée, un enjeu pour le machine learning : https://youtu.be/tL7ojiOTQho?t=972
- Intelligence artificielle faible et forte : quels impacts sur les métiers? https://youtu.be/xO8c257G4ms
- Laurent Alexandre sur les enjeux sociétaux de l'IA : https://youtu.be/rJowm24piM4

Pour aller plus loin

Retrouvez le site complet : ici.

Clement Levallois

Découvrez mes autres cours et projets : https://www.clementlevallois.net

Ou contactez-moi via Twitter: @seinecle