Étude de cas
en statistiques et informatique décisionnelle

STID Villetaneuse

Novembre 2021

Cours	Détails	Objectifs
Introduction à la Data Science	[présentation] Éléments de vocabulaire ; Workflow d'analyse des données ; Introduction au Machine Learning ; Etapes d'un projet de ML ;	Donner un aperçu des principaux concepts autour de la Data ; comprendre les grandes étapes de construction d'un projet de ML.
Data Cleaning	Handling Missing Values ; Scaling and Normalisation ; Parsing Dates ; Character encoding ; Inconsitant data entry.	Préparer et nettoyer les données afin d'être traitées par un alogrithme de ML.
Feature engineering	Information mutuelle ; Création de features ; Variables catégorielles ; PCA ; Target encoding.	Extraire des caractéristiques à partir de données brutes.
Machine learning	Construire un modèle de ML ; Validation de modèle ; Underfitting et Overfitting.	Construire un modèle de ML, comment valider un modèle, comprendre et éviter le sous-apprentissage et le sur-apprentissage.
Classification des données avec SVM	Appliquer le SVM pour classifier les données.	Utiliser le SVM pour classifier les données.
Clustering des données avec k-means	Appliquer le k-means pour clusteriser les données.	Utiliser le k-means pour clusteriser les données.
Visualisation des données	Visualiser les données et les résultats.	Manipuler une ou plusieurs biliothèques pour visualiser les données et les résultats.
Industrialisation des modèles	Industrialiser et mettre en production un modèle de ML.	Industrialiser et mettre en production un modèle de ML.

Étude de cas – plan de la présentation

- Introduction à la Data Science
  - Éléments de vocabulaire : qu’est-ce que la Data Science, gestion des données
  - Apprentissage Automatique (supervisé, non-supervisé, désambiguïsation)
  - Quiz
- Workflow d'analyse des données
  - Composants d’un workflow d’analyse de données
  - Relation entre les différents domaines de la data science
- Étapes d’un projet Machine Learning

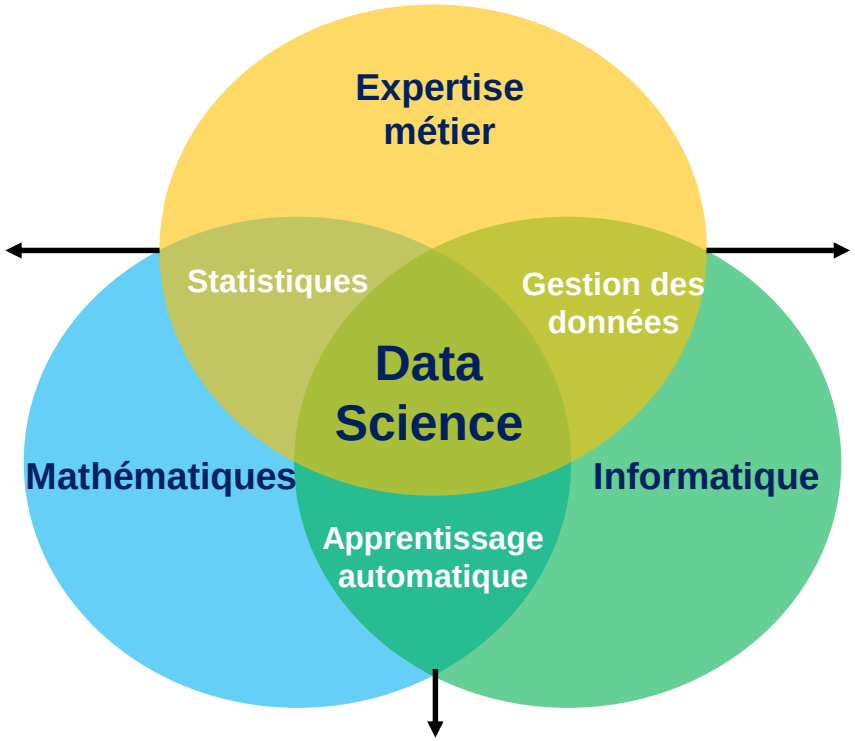
Introduction à la Data Science

Introduction à la Data Science

Éléments de vocabulaire : qu'est-ce que la Data Science

La Data Science est un domaine **multidisciplinaire**, à l'intersection de **l'informatique**, des **mathématiques** et de **l'expertise métier**. Un data scientist formalise une **approche analytique à partir d'un besoin métier**, analyse la **performance** et **industrialise** une solution analytique.

Les statistiques sont à l'**origine** des méthodes d'apprentissage. C'est à partir des **théorèmes statistiques** que sont créés les méthodes qui permettront à une machine **d'apprendre et d'extrapoler**.



Les données sont généralement stockées dans des bases de données.

Ces données seront sélectionnées et constitueront des **échantillons, nécessaires à l'apprentissage**.

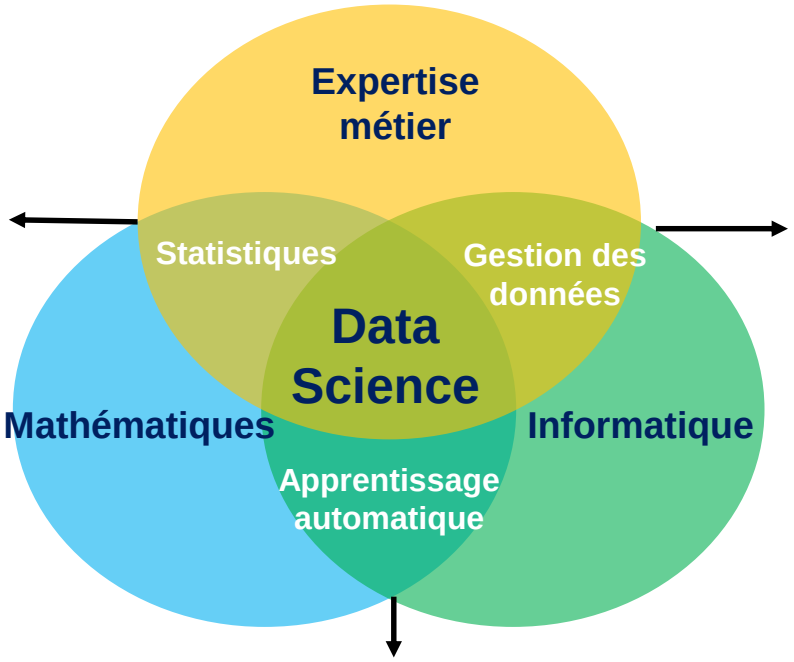
L'apprentissage est l'étape clef d'un projet de data science. On **apprend des données** de l'échantillon et on **applique les logiques sur de nouvelles données.**

Introduction à la Data Science

Éléments de vocabulaire : Cas d'usage d'un projet de data science

Concrètement, la science des données permet d'utiliser les données de façon créative pour générer une valeur pour les entreprises. Par exemple, Notflix cherche à définir les séries qui seront les plus populaires. Pour cela, ils essaient de comprendre ce qui suscite l'intérêt des utilisateurs, et donc les prochaines séries les plus visionnées.

Les statistiques sont à l'origine des méthodes d'apprentissage utilisées ici : fréquences, probabilités, théories mathématiques etc.



Notflix « track » le parcours utilisateur de chacun de ses clients (i.e. leurs activités de visionnage)

Ces données seront stockées dans des bases de données internes.

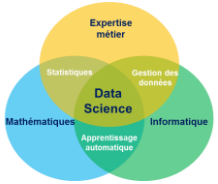
Une partie de ces données sera extraite et utilisée pour entrainer le modèle de prédiction.

Les données sont utilisées afin d'en déduire des tendances de visionnage.

Cela peut être par des règles directes par ex. « Les utilisateurs féminins regardent à 80% des séries comiques » ou par des tendances d'utilisations qu'une machine apprend par ex. «Si l'utilisateur a regardé la série A en peu de temps et a commencé la B sans la finir alors il aimera la série C »

Introduction à la Data Science

Éléments de vocabulaire : Gestion des données



Jeux de données

Collection de données associées entre elles et accessibles individuellement ou de façon combinée, ou gérées comme une entité.

Elles peuvent venir de différentes sources telles que les bases de données ou les fichiers Excel.

Echantillon

Ensemble d'éléments issus du jeu de données.

L'échantillonnage vise à obtenir une meilleure connaissance de la population initiale par l'étude d'un nombre d'éléments jugé statistiquement représentatif.

Features

Chaque élément de l'échantillon possède des caractéristiques appelées features.

Ces features peuvent être issues des données brutes (ex: les colonnes de la BDD ou du fichier) ou de transformations initiées par le data scientist ou de l'algorithme lui-même.

Label

Un label correspond au résultat qu'on s'attend à recevoir du modèle.

Parfois, l'échantillon qui servira à l'entrainement sera labellisé. Les labels seront utilisés par le modèle pour lier les features importantes à chacune des valeurs du label possible et extrapoler sur les prochains éléments.

Données structurées

Informations contrôlées par des référentiels et présentées dans des cases permettant leur interprétation et leur traitement facile par des machines

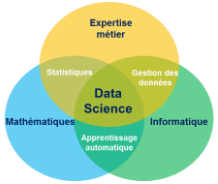
Données non structurées

Données qui ne respectent aucune organisation.

Par exemple: *les images, documents textuels, fichiers audio, vidéo*

Introduction à la Data Science

Éléments de vocabulaire : Gestion des données



Problématique : Notflix souhaite savoir sur quelle série investir pour le mois d'Octobre ?

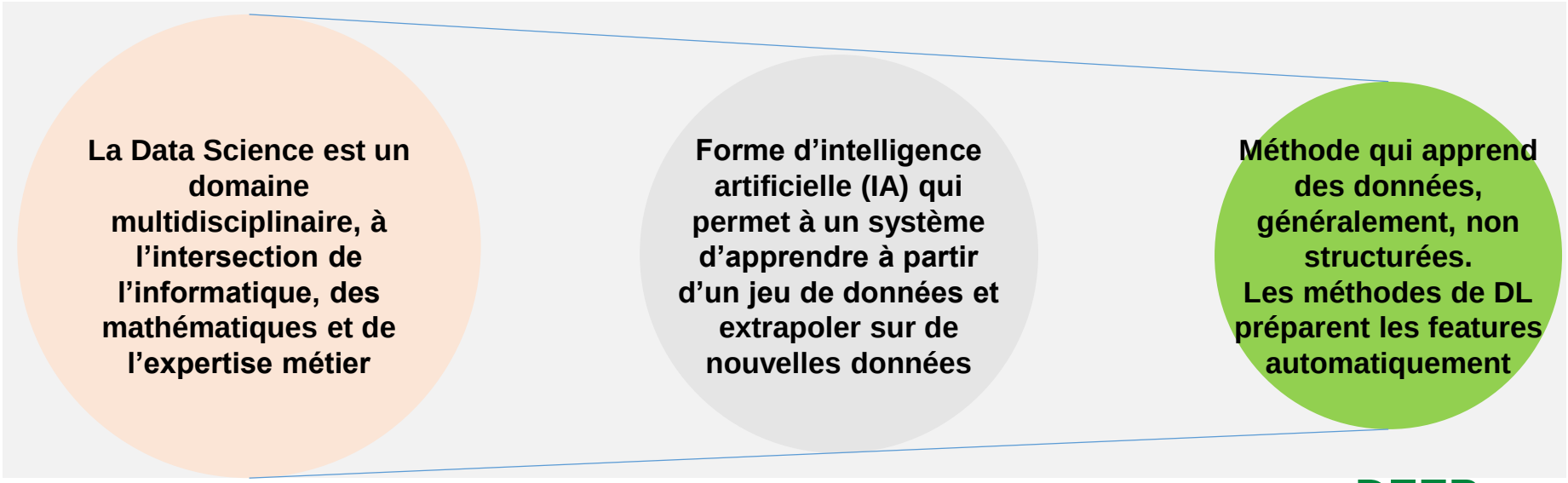
Pour répondre à cette problématique, Notflix doit répondre à la question suivante: Quelle sera la série la plus visionnée en Octobre à partir des données de visionnage de Septembre ?

Jeux de données	Echantillon	Features	Label
Notflix récolte l'ensemble des données historiques de visionnage : activités indexées par date et client en Septembre	Notflix sélectionne un ensemble de clients représentatif: date de naissance, ville de visionnage, sexe, dates et activité de visionnage.	Les data scientists de Notflix transforment les données brutes de Septembre: - Durée pour visionner la série en entier - Age à partir de la date de naissance - Temps de visionnage - Nombre de séries visionnées	Pour chacun des clients, le label correspondra à la prochaine série qu'il visionnera en octobre
• Dataset • Train set, Training set • Blind set • Validation set • Test set	• Sample • Jeu de documents représentatif	• Caractéristiques • Critères • Transformations	• Classe de documents • Valeur des champs • Etiquette • Y

SYNONYMES

Introduction à la Data Science

Apprentissage Automatique



DATA SCIENCE

MACHINE LEARNING

EXEMPLES d'application :

- Classification (ex :présence ou non d'indicateurs pour aider au diagnostic médical)
- Clustering (ex : segmenter une BDD utilisateurs)
- Régressions (ex : prédire le chiffre d'affaire)

DEEP LEARNING

(réseaux de neurones)

EXEMPLES d'application :

- Extraction de texte
- Reconnaissance d'images
- Traduction automatique
- Assistant vocal
- Voitures autonomes

OUTPUT
MODELE MACHINE LEARNING

Représentation de ce qu'un système ML a appris à partir des données d'apprentissage.
Modèle = Algorithme + Data + Apprentissage.

Introduction à la Data Science

Interactions entre les différents composants et leviers d'action

DONNEES

+

APPRENTISSAGE AUTOMATIQUE

+

STATISTIQUES

Il existe 2 grandes familles d'apprentissage:

- **L'apprentissage supervisé :** données **labellisées** au préalable
- **L'apprentissage non supervisé** : données sont **non labellisées.**

Derrière chaque algorithme de data science, se cache des **méthodes statistiques**. Ces méthodes sont **ajustables** via :

Des **hyper-paramètres** qui sont des leviers utilisables **par l'humain :**

- Choisis par l'homme
- Peuvent être fixés par un contexte métier

Des **paramètres** qui sont des leviers utilisables **par la machine :**

- Appris par la machine lors de la phase d'apprentissage

Introduction à la Data Science

L'apprentissage supervisé

On **entraîne** un modèle à **reproduire approximativement** un comportement humain en lui demandant de réaliser une tâche précise. Pour ce faire, on lui présente un **grand nombre de résultats attendus.**

Après entraînement, on présente au modèle une image qu'il n'a jamais vue :
- Il est capable de prédire le type d'animal qui apparait sur la photo
- mais avec un risque d'erreur (pas de règles précises = pas de certitude)

Le modèle **n'apprend pas seul** : il ne connaît rien du monde en dehors des **exemples** qu'on lui a **montrés lors de la phase d'entraînement**

On **améliore progressivement** les performances en **augmentant le nombre d'exemples d'entraînement** pour constituer un **panel représentatif de tous les cas** possibles

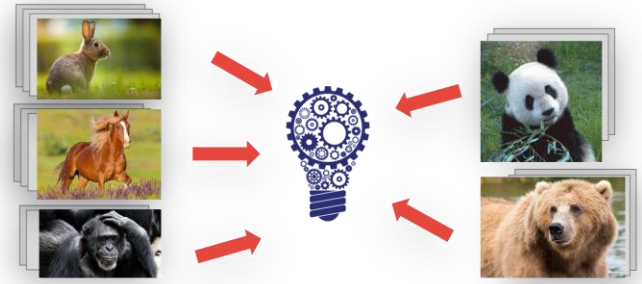
Ex : ici, on demande au modèle d'apprendre le type d'animal présent sur une image, avec les deux résultats attendus possibles : chat ou chien

CHAT
90%
CHIEN
10%

Ex : ici, le modèle prédit avec une probabilité de confiance de 90%, qu'il s'agit d'un chat

CHAT
55%
CHIEN
45%

Ex : ici, le modèle tente de rapprocher l'image du cheval à un chien ou à un chat mais avec un taux de confiance bas



Ex : ici, on donne le maximum d'images d'animaux labellisées pour entraîner le modèle pour qu'il puisse reconnaitre un maximum d'animaux

Introduction à la Data Science

L'apprentissage non-supervisé

Lors de l'entrainement, on donne au modèle un nombre important d'images **sans aucune indication permettant de les différencier**

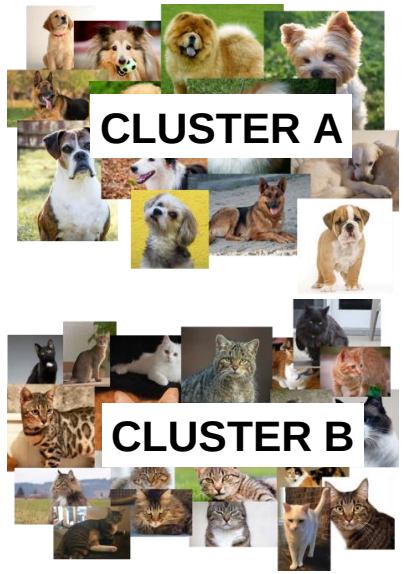
Le modèle fait de lui-même, des regroupements (clusters) d'images similaires qui selon lui présentent des caractéristiques communes

Après entraînement, on donne au modèle une image qu'il n'a jamais vue :

- Il est capable de rapprocher l'image **d'un cluster qu'il a identifié au préalable**
- Mais avec un risque d'erreur (pas de règles précises = pas de certitude)

Le modèle **n'apprend pas seul** : il ne connaît rien du monde en dehors des **exemples** qu'on lui a **montrés lors de la phase d'entraînement.**
Il donnera **toujours** une prédiction pour l'image présentée, avec un risque d'erreur

On **améliore progressivement** les performances du modèle en **augmentant le nombre d'exemples d'entraînement** pour constituer un **panel représentatif de tous les cas** possibles



CLUSTER B

90%

CLUSTER A

10%

CLUSTER B

55%

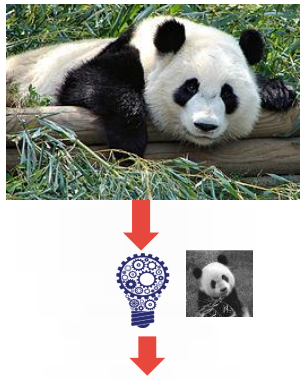
CLUSTER A

45%

Introduction à la Data Science

La désambiguïsation

Le modèle **rapproche l'image soumise des exemples les plus similaires** vus pendant la phase d'entraînement pour **sélectionner les résultats les plus probables**



90% PANDA
8% OURS
1% SINGE
1% LAPIN

Si des **exemples trop similaires** sont présentés pour deux types d'animaux, le modèle ne **peut pas désigner avec confiance** l'une des deux catégories

58% LAPIN
40 % CHAT
2% PANDA

50% LAPIN
45% CHAT
5% PANDA

Le modèle *peut* alors proposer **plusieurs suggestions** à l'utilisateur

Dans certains cas, si les probabilités sont trop proches, on peut demander à l'utilisateur de choisir via la désambiguisation

58% LAPIN
40 % CHAT
2% PANDA

50% LAPIN
45% CHAT
5% PANDA

Désambiguïsation

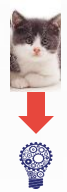
Sur la base des retours utilisateurs, le modèle va être **progressivement amélioré** en ajoutant de **nouveaux exemples d'entrainement**

Analystes métiers

Analyse des données d'utilisation

- Enrichissement de la base d'exemples
- Suivi des erreurs
- Test des modifications

74% LAPIN
25% CHAT
1% PANDA



78% CHAT
20% LAPIN
2% PANDA

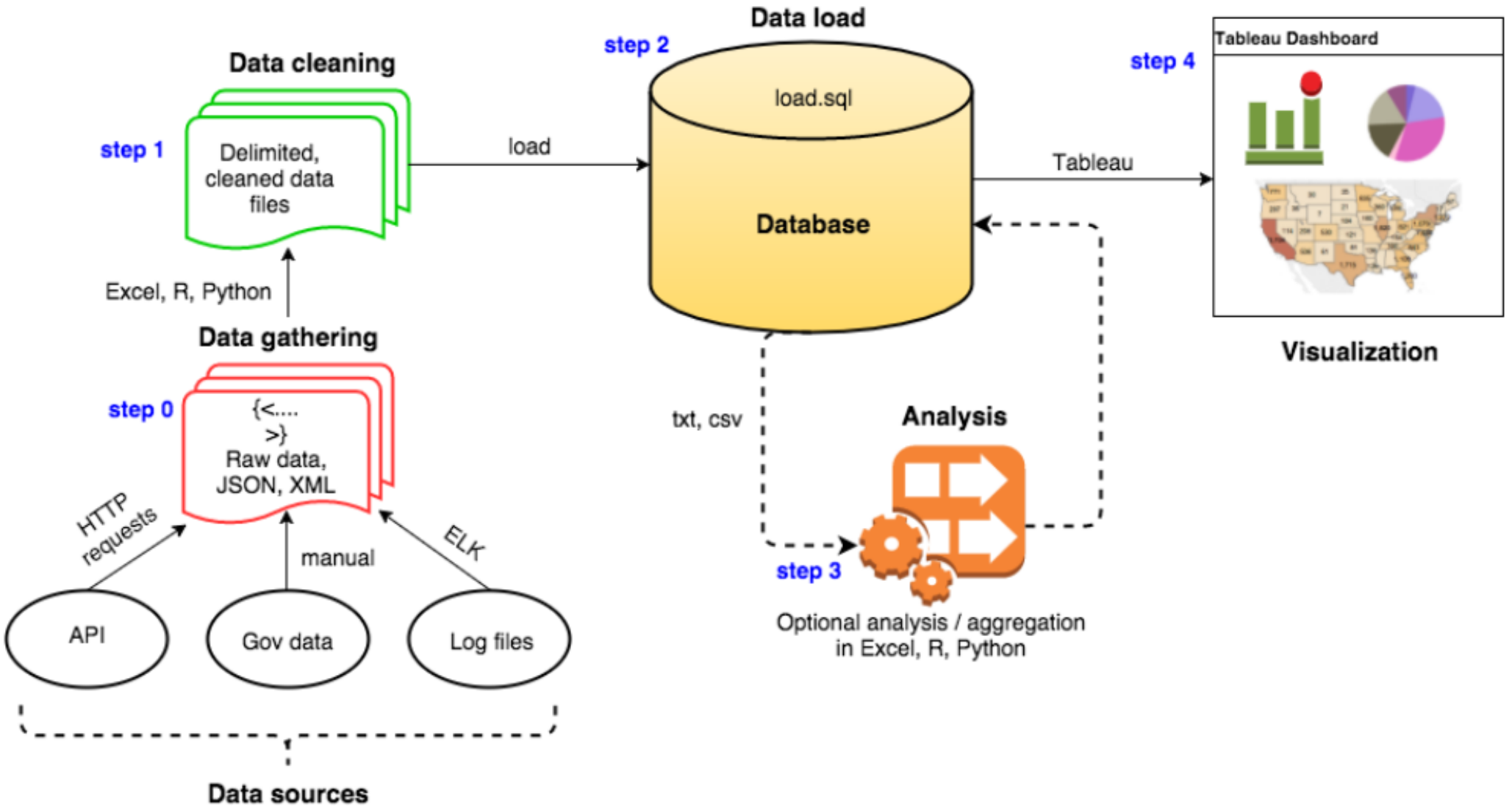
Amélioration du taux de confiance

Quiz
Apprentissage supervisé ou non-supervisé ?

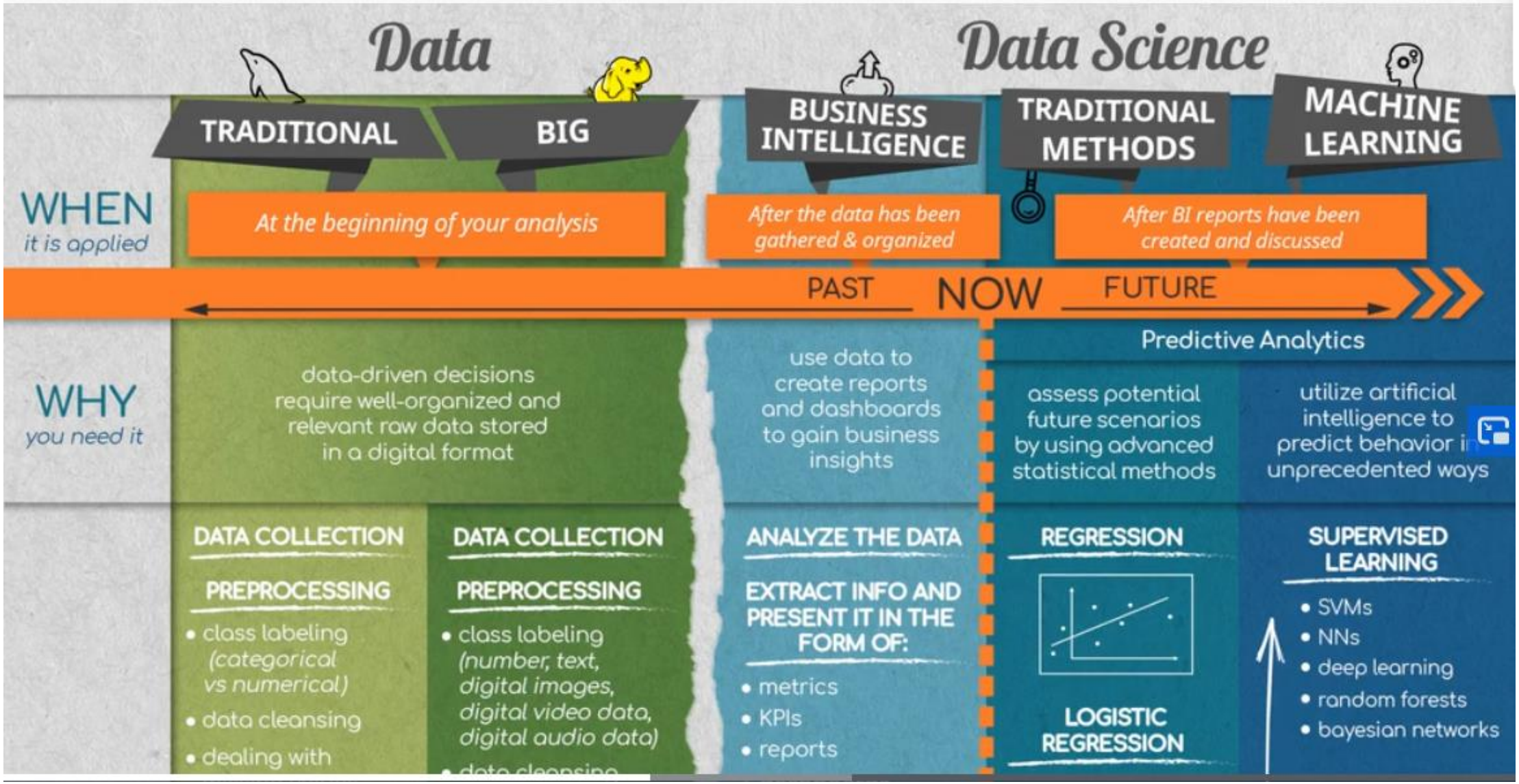
1. Prédiction du chiffre d'affaire d'une entreprise
2. Détection de fraude
3. Segmentation de clients
4. Classification de documents (RIB, CNI, Facture, etc.)
5. Systèmes de recommandation
6. Traduction automatique de langue
7. Self-driving cars

Workflow d'analyse des données

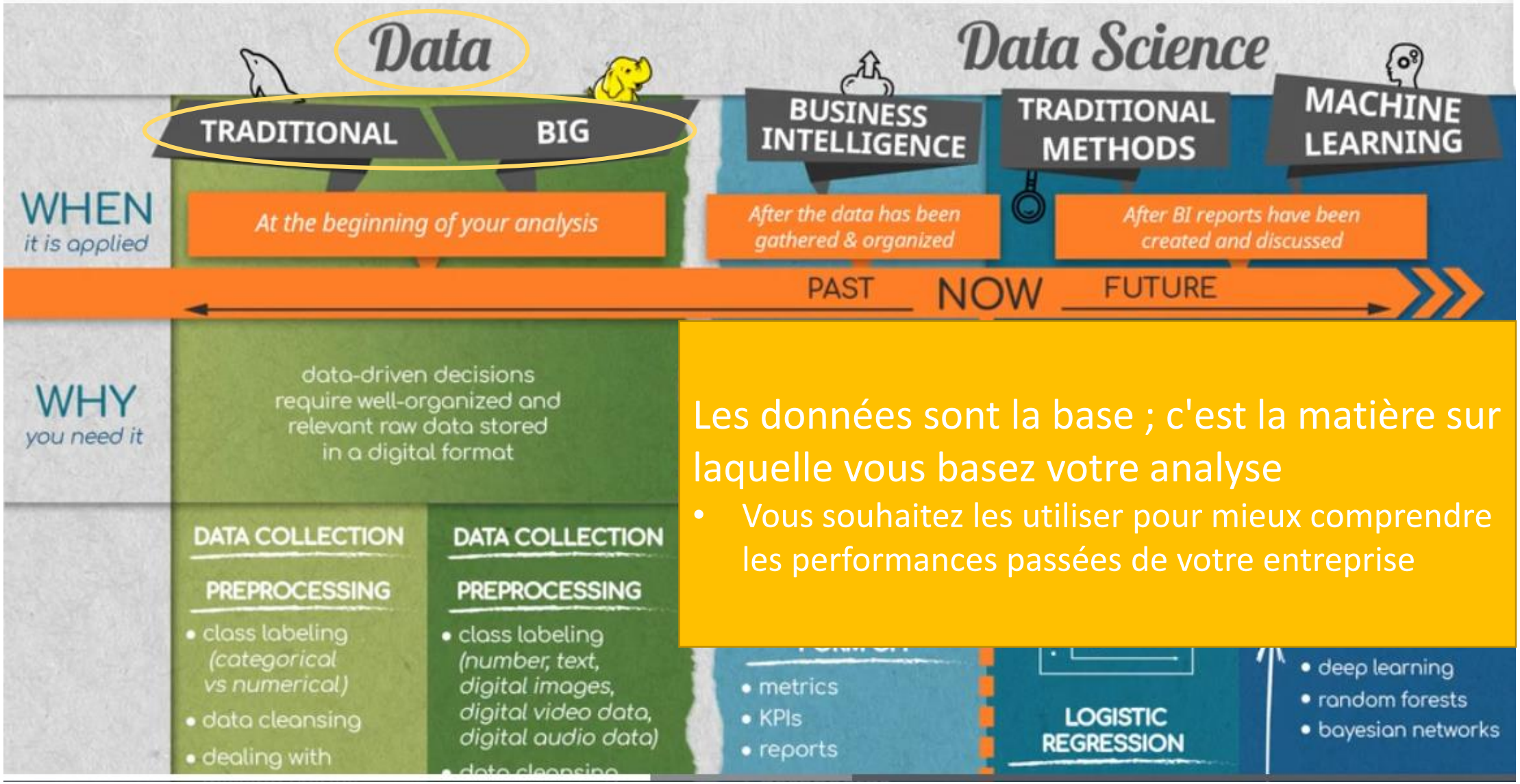
Composants d’un workflow d’analyse de données



Relation entre les différents domaines de la data science



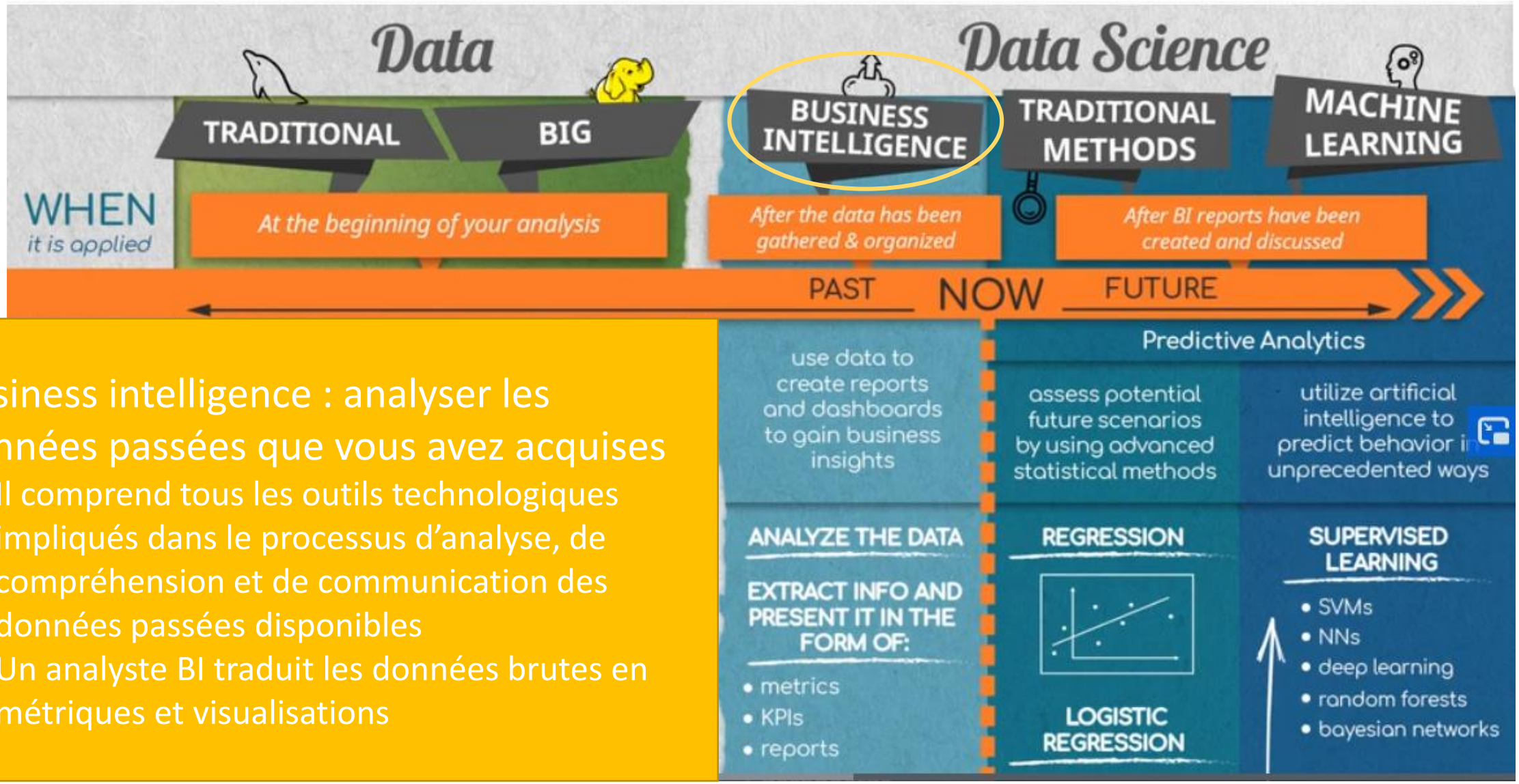
Relation entre les différents domaines de la data science



Les données sont la base ; c'est la matière sur laquelle vous basez votre analyse

- Vous souhaitez les utiliser pour mieux comprendre les performances passées de votre entreprise

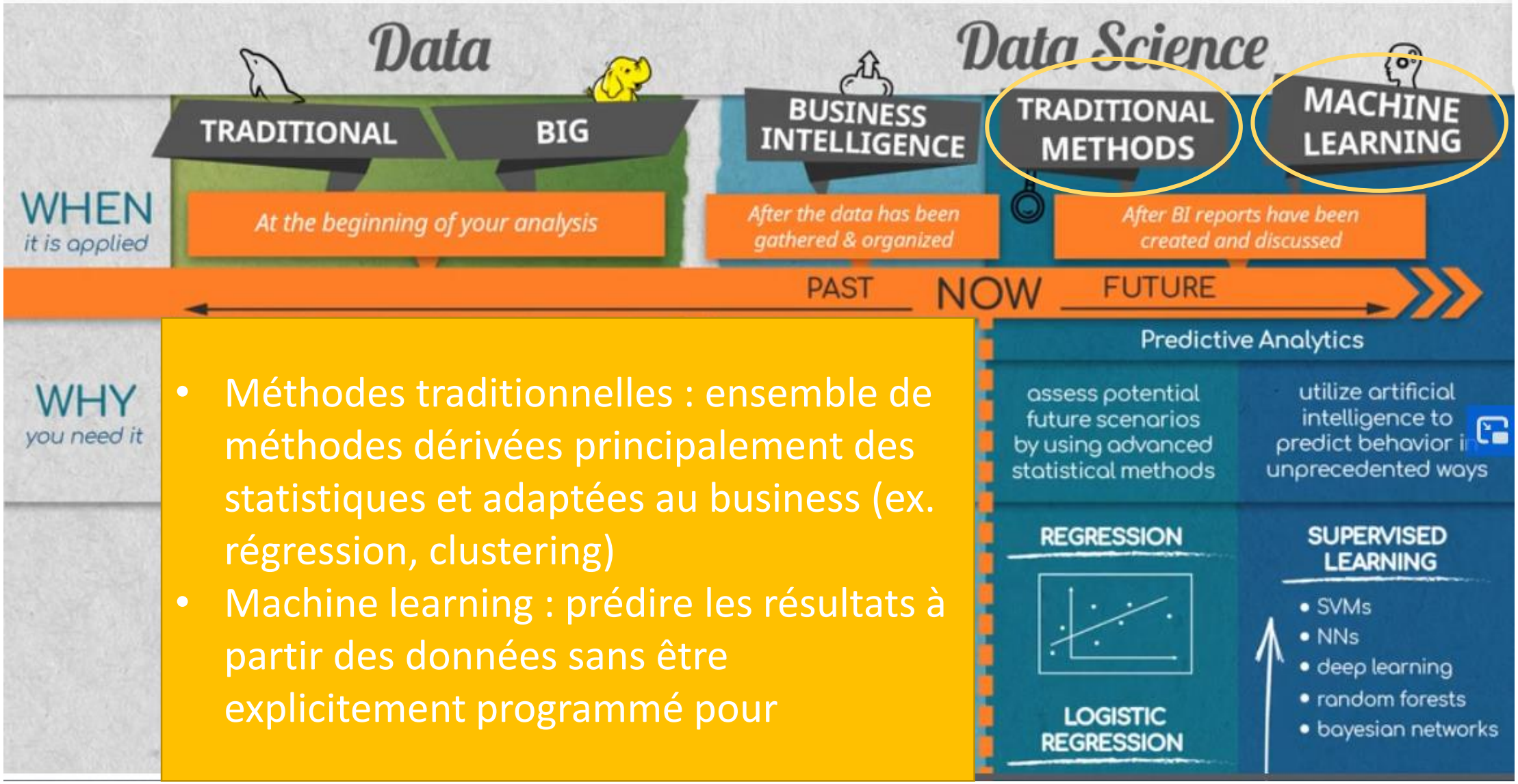
Relation entre les différents domaines de la data science



Business intelligence : analyser les données passées que vous avez acquises

- Il comprend tous les outils technologiques impliqués dans le processus d'analyse, de compréhension et de communication des données passées disponibles
- Un analyste BI traduit les données brutes en métriques et visualisations

Relation entre les différents domaines de la data science



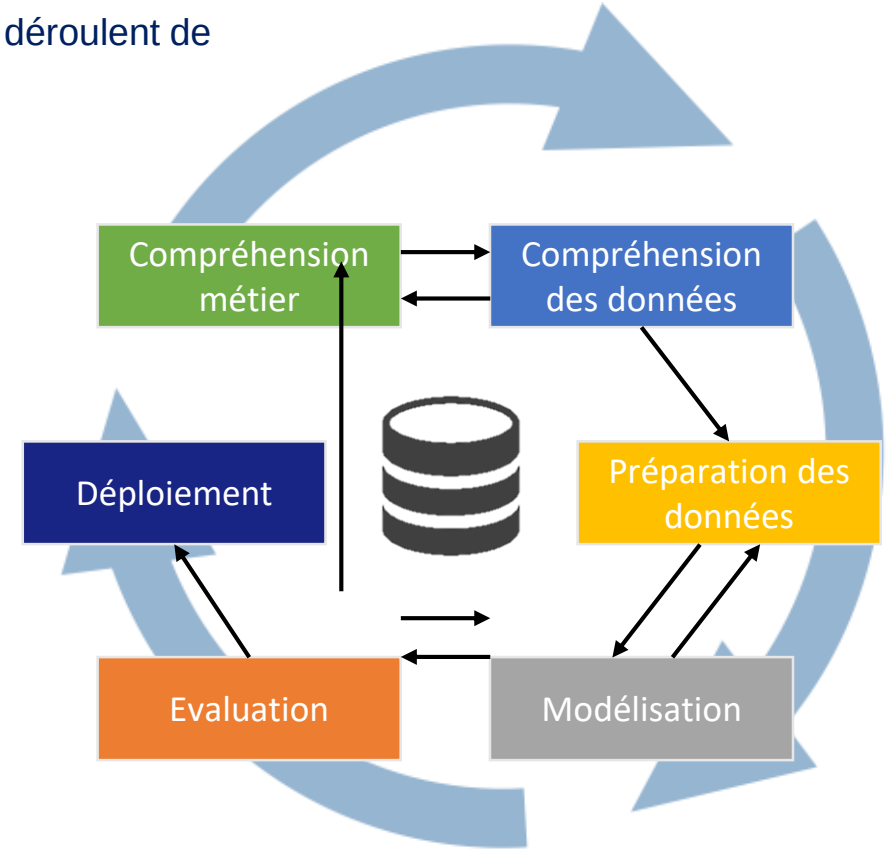
- Méthodes traditionnelles : ensemble de méthodes dérivées principalement des statistiques et adaptées au business (ex. régression, clustering)
- Machine learning : prédire les résultats à partir des données sans être explicitement programmé pour

Étapes d’un projet Machine Learning

Étapes d'un projet Machine Learning

Chaque projet **d'apprentissage automatique** suit **6 étapes** principales qui se déroulent de manière **itérative** afin d'atteindre les meilleurs résultats.

Compréhension métier	**Comprendre les éléments métiers** et les problématiques qu'on veut résoudre ou améliorer
Compréhension des données	**Identifier les données** à analyser et comprendre leur signification d'un point de vue métier
Préparation des données	**Construire, nettoyer et recoder les données brutes** pour les rendre compatibles avec les algorithmes qui seront utilisés dans l'étape suivante
Modélisation	**Choisir, paramétrer et tester** les différents algorithmes.
Evaluation	**Vérifier que les modèles** obtenus répondent aux objectifs formulés dans la première étape
Déploiement	**Mettre en production** les modèles obtenus



Généralement, les modèles s'inscrivent dans une stratégie long terme.

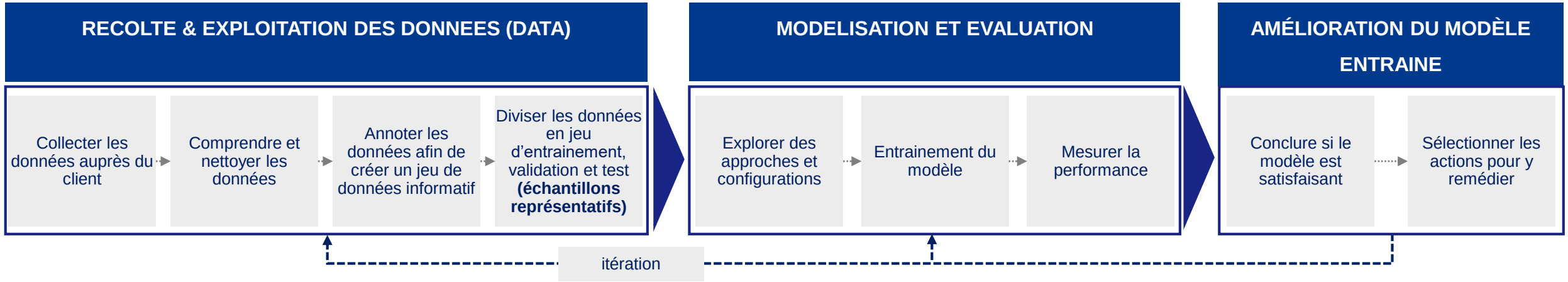
Des analystes suivent régulièrement les performances afin de les maintenir. Lorsqu'ils observent des déviations, ils analysent les causes des déviations et prennent des mesures adaptées incluant une éventuelle refonte du modèle:

- Alimentation par de nouvelles données
- Nouvelles transformations des données
- Ré entrainement du modèle (nouveaux paramètres, algorithmes)

Étapes d'un projet Machine Learning

Cycle de conception d'un modèle de ML

- Avant de démarrer un projet de Machine Learning, il est nécessaire d'explorer les différentes approches ML et de les prioriser en fonction de leur potentiel. Les autres suggestions seront conservées et éventuellement explorées afin d'améliorer le modèle entrainé.
- Une fois le modèle satisfaisant trouvé, il sera mis en production et monitoré à l'occasion de l'amélioration continue.



Description

L'objectif est **d'obtenir 3 jeux de données représentatifs.**

Ces données seront utilisées pour:
- **Entrainer** / paramétrer un modèle
- **Valider** le modèle sur des données représentatives
- **Tester** le modèle afin de s'assurer que les performances obtenues seront consistantes en production

Les données initiales seront annotées par le métier.

La modélisation consiste à produire un modèle qui, **à partir des données annotées initiales,** permet de prédire:
- La classe de l'image
- Le texte contenu dans cette image
- La correspondance entre le texte et les champs attendus

Un critère d'évaluation sera défini et servira à mesurer la précision du modèle en comparant les annotations avec l'output de ce modèle.

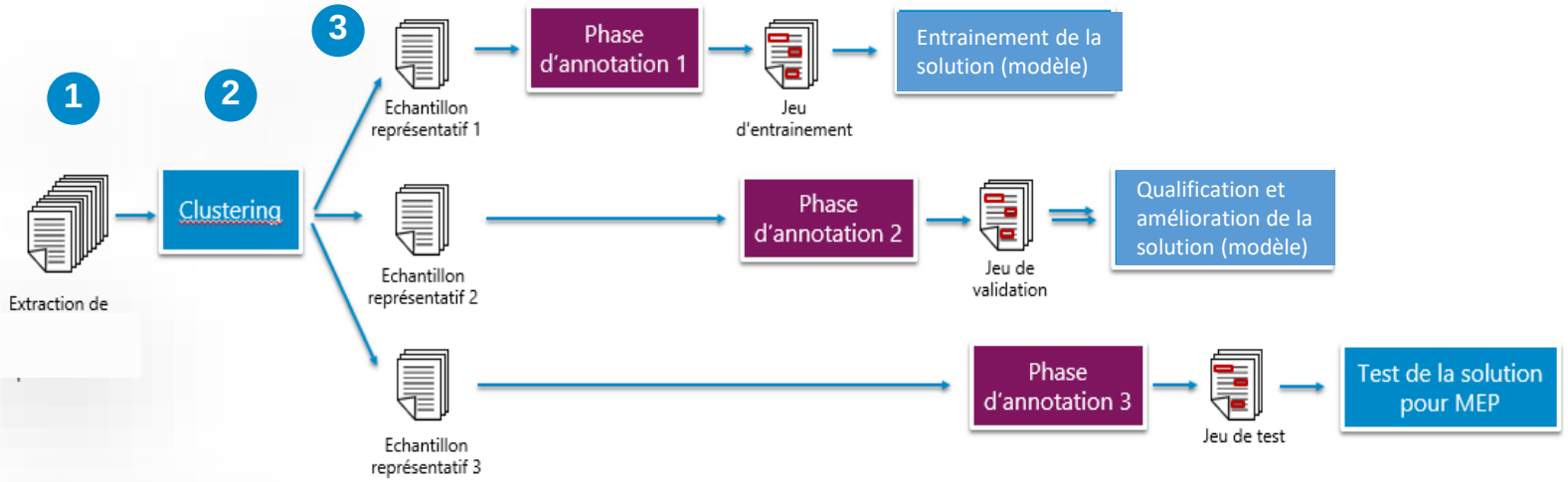
Si la performance du modèle est satisfaisante, mise en production du modèle sélectionné lors de la phase de modélisation.

Si la performance du modèle n'est pas satisfaisante des solutions seront proposées pour améliorer le modèle et une **nouvelle itération** du cycle de conception sera menée.

Étapes d'un projet Machine Learning

Récolte & exploitation des données| Les jeux de données

Le schéma ci-après représente les principales étapes de constitution des jeux de données:



1. Documents fournis par le client ou issus de la production
2. Le clustering consiste à regrouper ces documents en fonction de leurs caractéristiques communes (classe, modèle de document, qualité de numérisation, etc.). Ex : 100 clusters
3. Extraction aléatoire de quelques documents de chaque cluster pour constituer trois jeux de données représentatifs des documents en production

Ci-dessous, les spécificités de chacun des jeux de données:

	Jeu d'entraînement (Train Set = TS)	Jeu de validation (Validation Set = VS)	Jeu de test (Blind Set = BS)
Définitions	• Le jeu d'entrainement va permettre de développer le projet OCR en **entrainant la solution** (que ce soit par le biais d'un paramétrage de mots clé ou par du Machine Learning). Il s'agit du seul jeu pré-requis pour le démarrage des travaux OCR	• Le jeu de validation servira à **qualifier la version du projet OCR** avec des documents n'ayant pas été utilisés pour l'entrainement, afin d'avoir des résultats plus proches de la réalité. Des itérations d'optimisation de l'OCR seront effectuées à partir de cet échantillon.	• Le jeu de test sera utilisé **pour vérifier que les taux obtenus sur ces nouveaux documents sont conformes** à l'attendu et permettent une mise en production.
Utilisations	• Utilisé par les analystes pour générer des règles (solution programmée) OU • Utilisé par l'algorithme de Machine Learning (e.g. Tesseract) pour produire le modèle (solution entraînée)	• Utilisé par les analystes pour valider un modèle en cours de développement	• Utilisé par les analystes pour confirmer la validation pour mise en recette/production une fois le modèle satisfaisant. (Il sert également à confirmer que l'on n'a pas biaisé le modèle sur le jeu de validation)

Étapes d'un projet Machine Learning

Modélisation | Les différentes étapes

Avant et post mise en production d'un modèle, **trois niveaux d'indicateurs** peuvent être contrôlés (1).Les priorités sont définies selon les besoins du projet :

EXHAUSTIVITÉ de Classification

Taux de classification par classe de document = $\frac{\text{Nb pages classées}}{\text{Nb total de pages à classer}}$

Taux d'extraction moyen par classe de document = $\frac{\text{Nb champs extraits}}{\text{Nb total champs extractables (2)}}$

PRECISION de Classification

Taux d'erreur de la classification par classe de document = $\frac{\text{Nb pages « mal classées » (3)}}{\text{Nb pages classées}}$

Taux d'erreur de l'extraction par classe de document = $\frac{\text{Nb champs « mal extraits »}}{\text{Nb total champs extraits}}$

TEMPS D'INFÉRENCE : temps nécessaire en secondes au modèle pour produire un output

Exemple

Objectifs	Classe 1 (ex: contrat prêt)	Classe 2 (ex: contrat assur.)	Classe 3 (ex: carte identité)
Taux de classification année 1	90%	90%	90%
Taux de classification année 2	90%	90%	92%
Taux de classification année 3	89%	89%	94%
Taux d'extraction année 1	80%	80%	
Taux d'extraction année 2	75%	75%	
Taux d'extraction année 3	70%	70%	
Taux erreur classification année 1	1%	2%	1%
Taux erreur classification année 2	1%	2%	1%
Taux erreur classification année 3	2%	2%	2%
Taux erreur extraction année 1	6%	3%	
Taux erreur extraction année 2	6%	3%	
Taux erreur extraction année 3	7%	4%	

(1) Ces métriques sont calculées en appliquant le modèle d'apprentissage sur le Jeu de Test.

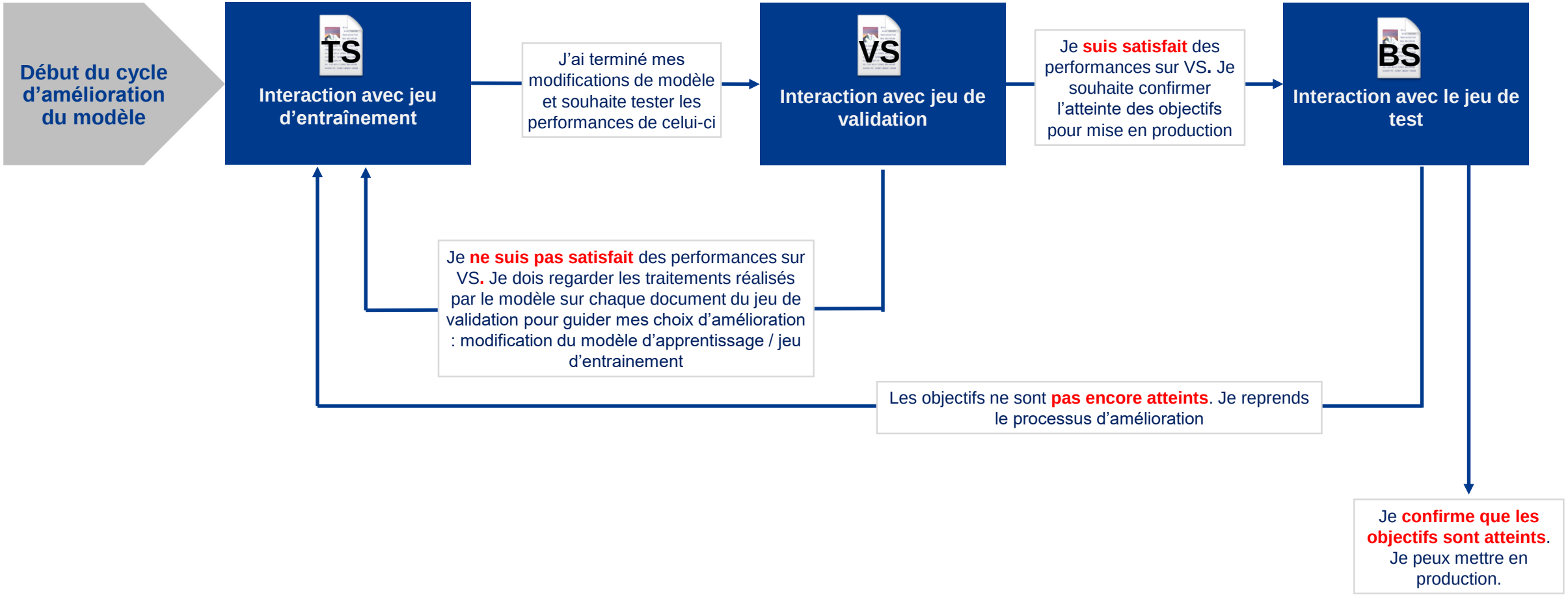
(2) Selon comparaison entre le modèle d'apprentissage et les annotations manuelles sur le Jeu de Test

(3) Selon le nombre de champs de la vérité terrain, non le nombre de champs demandés (ex : sur une CNI scannée seulement au recto, 2 champs sont annotés, même si les règles métiers demandent 3 champs)

Étapes d'un projet Machine Learning

Amélioration du modèle entrainé

Comment améliorer le modèle entrainé?



Les points clefs à retenir

Les points clefs à retenir

1. La Data Science est un domaine **pluridisciplinaire**
2. Un modèle = données + algorithme + entrainement
3. La machine apprend **uniquement** à partir des exemples qu'on lui donne
4. La **qualité, quantité et représentativité** des données d'entrainement sont clefs pour la performance d'un modèle
5. Un modèle de Machine Learning est probabiliste : il donne un **taux de confiance**. Un modèle fiable à **100%** n'existe pas !
6. L'utilisation de techniques d'apprentissage automatique permet de faciliter les projets (réutilisation des modèles, moins d'exploration manuelle, etc.)
7. Trois jeux de données pour un modèle : **entrainement**, **validation** et de **test**
8. L'amélioration continue est un élément clef dans un projet d'apprentissage automatique.