Big Data Sécurité

Alison PATOU

Patou.alison@gmail.com

Programme

- Introduction
- Focus Big Data
- Cybersécurité & Big Data
- Enoncé projet

A savoir

- Tout est sur mon GitHub : https://github.com/apatou
- (L'essentiel à retenir du cours, les dataset, exercices, …)

- Merci de m'envoyer à chaque fin de séance vos TPs : patou.alison@gmail.com

- 1 projet en groupe

1

Introduction

Introduction

L’un des principaux usages de l’IA dans le monde réel mais aussi sur Internet est dans la cybersécurité

« Deux entreprises sur trois prévoient de déployer des systèmes IA dès 2020 afin renforcer leur défense »

Arrêt de production d’un site pétrochimique en Arabie saoudite à cause du malware Triton

PIXELS

Partage

Renault touché par la cyberattaque de niveau mondial, des sites de production à l'arrêt

Le constructeur est la première institution française à reconnaître avoir été atteinte par l'attaque informatique qui a visé vendredi des dizaines de milliers d'ordinateurs.

Le Monde avec AFP et Reuters • Publié le 13 mai 2017 à 10h07 - Mis à jour le 13 mai 2017 à 13h24

Lecture 2 min.

Arrêt de production de Renault à cause de WannaCry

Introduction

Les sites industriels sont démunis face aux attaques, notamment par le fait qu'historiquement leur informatique industrielle est très scillotée.

Cependant, ces réseaux montrent leur vulnérabilité (au email de pishing, un malware provenant d'une clé USB, …)

Les projets de transformation digitale sont donc en cours afin d'améliorer la sécurité.

C’est là qu’intervient l’IA

Pour aider et contrer les cyberattaques de manière efficace au sein des entreprises, l’IA peut se baser sur les informations au sujet de toutes les cybermenaces connues à l’échelle mondiale afin de détecter instantanément toute tentative d’attaque. Ceci permet aux utilisateurs de réagir bien plus rapidement.

Pour anticiper les crimes en les prédisant et ainsi les éviter en amont

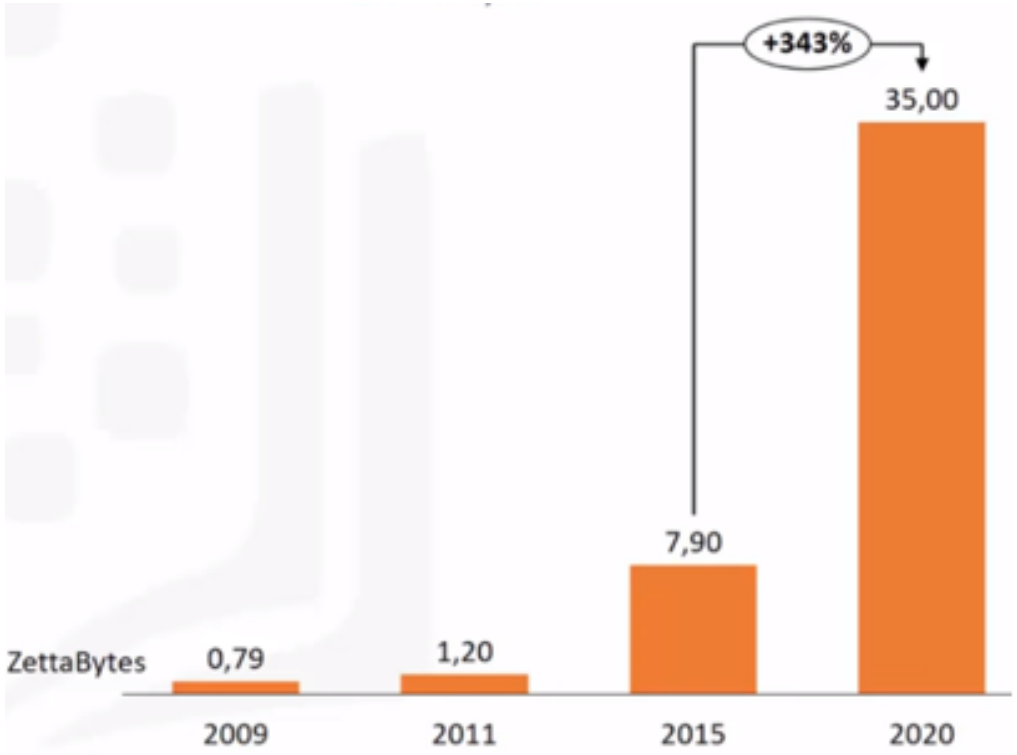
2

Focus Big Data

C’est l’explosion du Big Data qui rend possible l’IA

L’intelligence Artificielle n’aurait pas été possible sans l’aide d’une immense puissance informatique, de coût de stockage réduit et de logiciels customisés.

D’une part le volume des données générées augmente



D’autres part les données ne sont plus seulement structurées

24 PO (24 000 000 000 000 000 000) / j
(environ 30 000 disques durs de PC)

10 000 000 photos / jour
3 000 000 000 ‘like’ / jour

2015 : 500 000 000 tweets / jour

800 000 000 utilisateurs / mois

Particuliers

Entreprises

Objets connectés

Open Data

Les technologies clés

- **Bases de données NoSQL**
- **Hadoop et son environnement**
- **Spark**
- **Technologies sémantiques**

Données structurées/non structurées

Données structurées	Données non structurées
Fichier Excel	Email
Données de capteur	Images satellites
Données des clients	Vidéo Youtube
Etc	Etc

Une approche différente :

- **Au niveau du stockage :**
  - l'unite de stockage dans les SGBDR est la relation.
    Dans le NoSQL, c'est une collection d'objets (un document, une image, etc) référencée par une clé
  - Dans un SGBDR, les tables sont liées par des contraintes référentielles. En NoSQL, il n'y a pas de lien entre les collections d'objets, ce qui facilite la distribution des données
- **Au niveau du schéma des données :**
  - les SBBDR permettent de séparer la couche données de la couche applicative
  - dans le NoSQL, l'absence de schéma oblige le développeur a coder la logique applicative et la cohérence des données dans l'application, pouvant complexifier la mise-a-jour de l'application ou la modification de la base de données
- **Au niveau de la modélisation des données :**
  - dans un SGBDR, la modélisation des données est la première phase du projet (approche projet séquentielle)
  - en NoSQL, la modélisation est plus flexible permettant les approches itératives de developpement,

Différence SQL/NoSQL

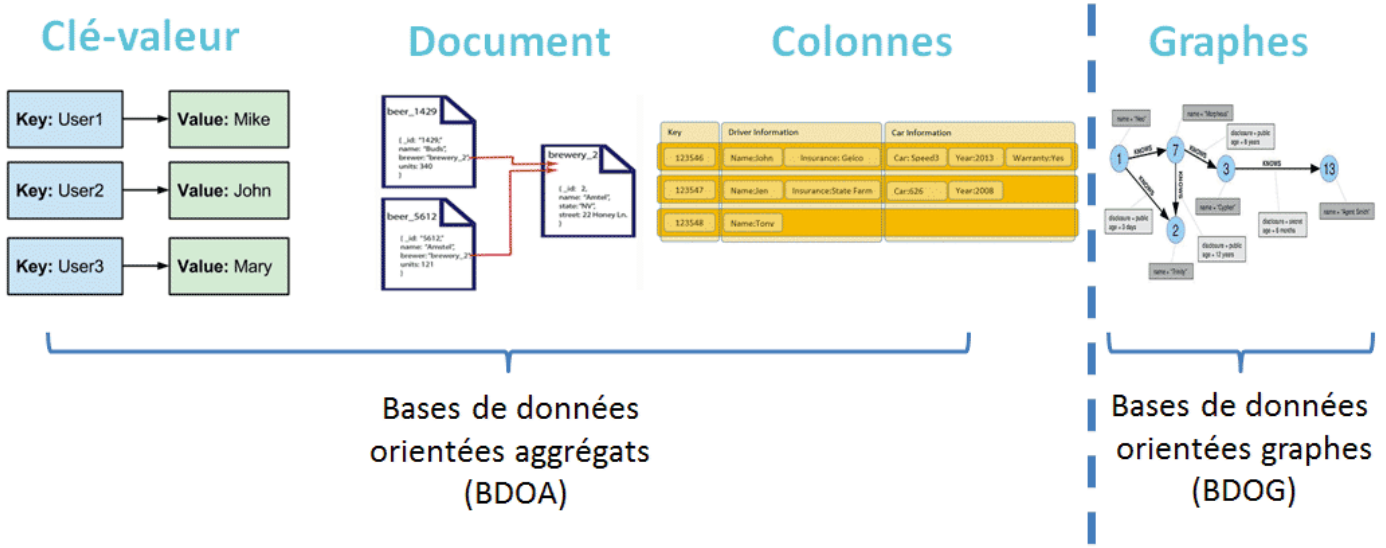
SQL	NoSQL
Modele relationnel : • Fondements theoriques • normalisation	Schéma dynamique Sans schéma
Langage SQL	Langage SQL adapté
OLTP/OLAP	Environnement basé sur Hadoop
Pas d'impédance	Encapsulation des langages facilité

Emergence du NoSQL

Le terme base NoSQL définit une nouvelle génération de produits qui ne suivent pas le modèle relationnel. Mais l'architecture de ces produits varie beaucoup entre eux.

Il existe 4 types de bases de données NoSQL :

- Clé-valeur
- Documents
- Colonnes
- Graphes



Emergence du NoSQL

Hadoop

Projet Open Source (Apache Software Foundation)
Environnement d'execution **distribué**

Sous-projets: MapReduce, HDFS, Hbase, Pig, Hive, Mahout, etc.

Permet de traiter de grande volumétrie de données.

Contient deux composants :

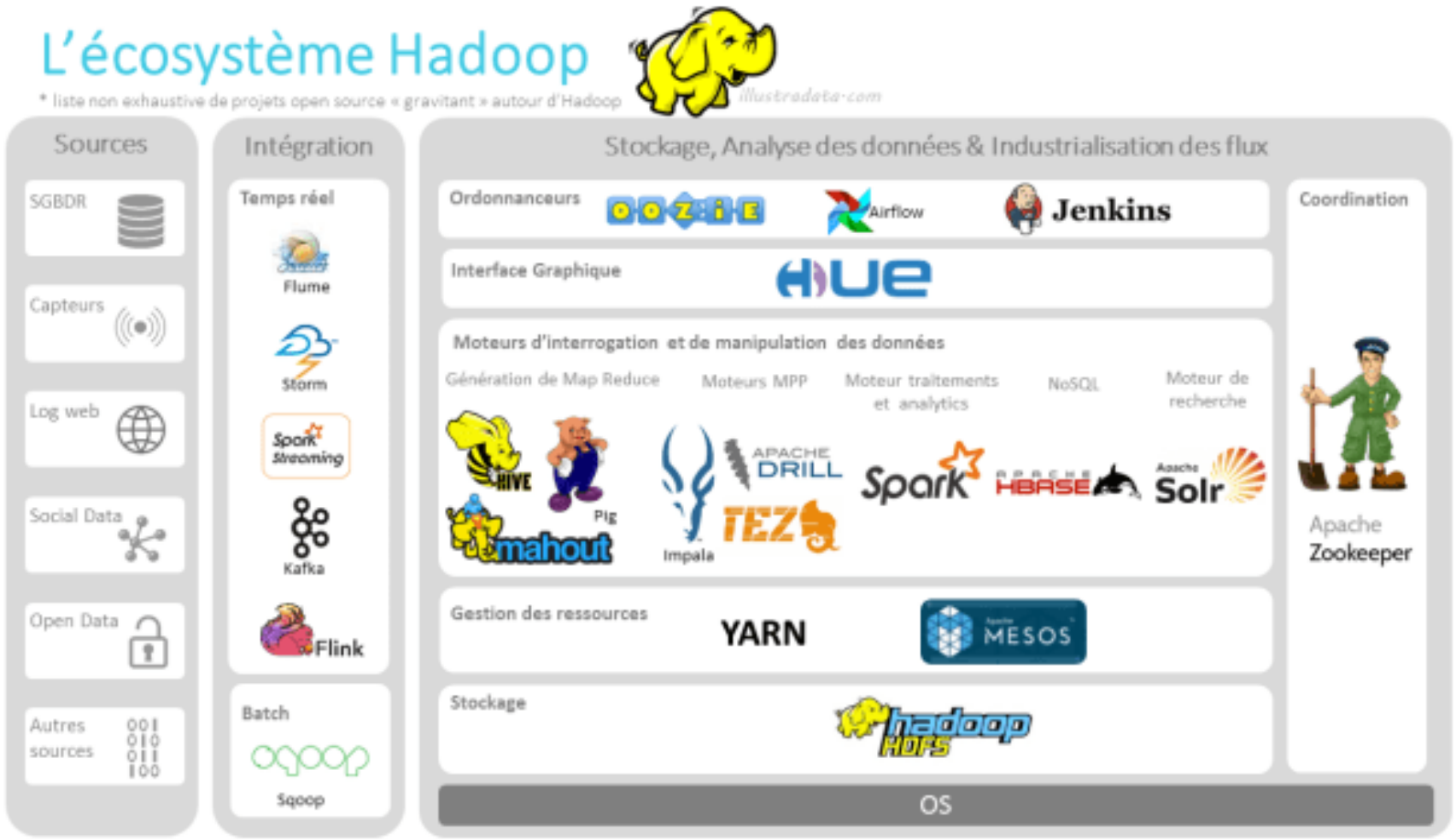
- Le systeme de gestion de fichiers distribue **HDFS**
- Le framework logiciel **MapReduce**

Hadoop

L’objectif principal d’Hadoop est de stocker et traiter des volumes de données très importants, dans des **délais** et a un **coût raisonnable.**

Plus le nombre de nœuds de calcul utilisés est important, plus la puissance de traitement est élevée. Les données et les applications traitées sont protégées contre les échecs hardware. Si **un nœud tombe en panne, les tâches sont directement redirigées vers d’autres nœuds pour s’assurer que le calcul distribué n’échoue pas**. De multiples copies de toutes les données sont stockées automatiquement.

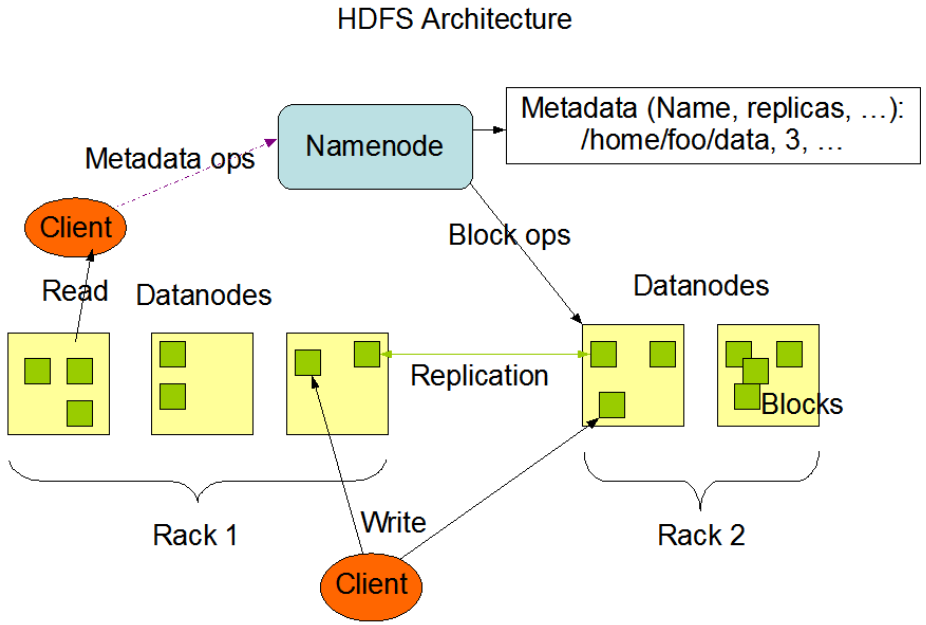
Ecosysteme Hadoop



HDFS

HDFS (Hadoop Distributed File System) : C'est le Système de stockage de fichiers distribué de Hadoop. Il permet de stocker au format natif n'importe quel objets (image, vidéo, fichiers texte etc…) en faisant porter la charge sur plusieurs serveurs

Les performances sont meilleures quand :
– Les fichiers sont de taille importante (> 100 Mo)
– Le nombre de fichiers est reduit (des millions plutot que des milliards)
Répartition des données sur plusieurs machines
Réplication des données pour gérer les pannes et la répartition de charge



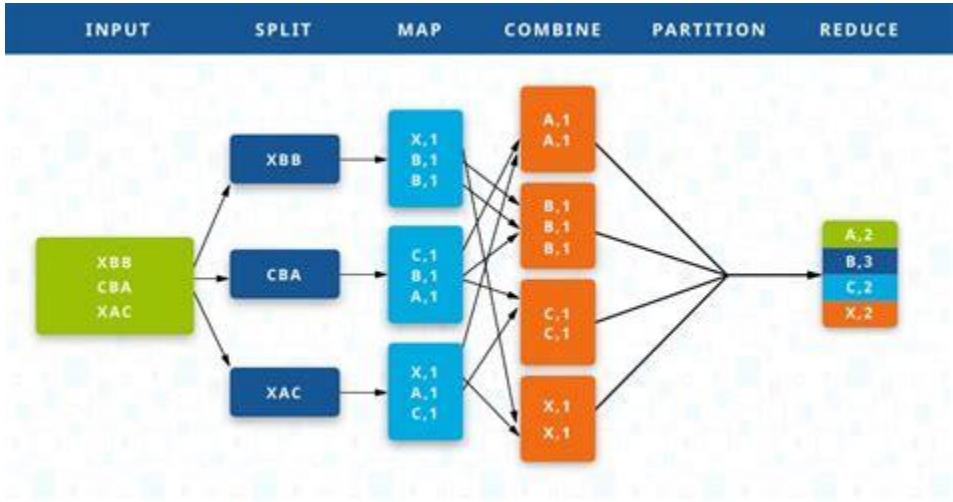
MapReduce

Moteurs d'interrogation et de manipulation des données

MapReduce : Framework open source java, permettant la manipulation des données dans un environnement distribué.
Il est composé de deux étapes principales.

- L'étape de map qui va permettre d'effectuer des actions là ou sont stockées les données et fournir en sortie une liste de clés valeurs.
- L'étape de reduce qui va regrouper les résultats des map en fonctions de clés et effectuer les actions finales (les actions sur les valeurs?).

Les étapes de map et de reduce lisent les données et écrivent leurs résultats sur disque , cela rend le processus stable mais lent.



Spark

En voie de généralisation dans les projets big data

Projet permettant la manipulation des données dans un environnement distribué. Il peut aussi bien faire les traitements sur disque ou tout en mémoire. Il va 10 fois plus vite que MapReduce sur disque et 100 fois plus vite en mémoire. Il est enrichi de librairies notamment MLiB qui contient des algorithmes parallélisés de machine learning, GraphX pour les algorithmes de graphes, SparkSQL pour notamment se connecter au metastore de Hive. Spark peut se plugger sur la majorité des systèmes distribués (NoSQL, Hadoop, MPP …).

3

Cybersécurité et Big Data

La cybersécurité

Protéger les informations sensibles peut sembler plutôt simple. Toutefois, face au volume des données à traiter et à analyser pour prévenir les cyberattaques, la plupart des entreprises sont confrontées à un challenge d'envergure.

Selon Computer World, un réseau de taille moyenne, composé de 20 000 appareils incluant des laptops, des smartphones et des serveurs, transmet plus de 50 terabytes de données sur une période de 24 heures.

De fait, pour détecter les cyberattaques, plus de 5 gigabits de données doivent être analysées chaque seconde.

La sécurité de stockage des données

Le stockage Cloud est de plus en plus utilisé par les entreprises, mais la sécurité du Cloud reste un enjeu majeur. Selon un rapport publié par RedLock, 51% des entreprises qui utilisent le service de stockage Cloud Amazon Web Services (AWS) S3 ont subi **au moins une fuite de données en 2017**.

La protection des données

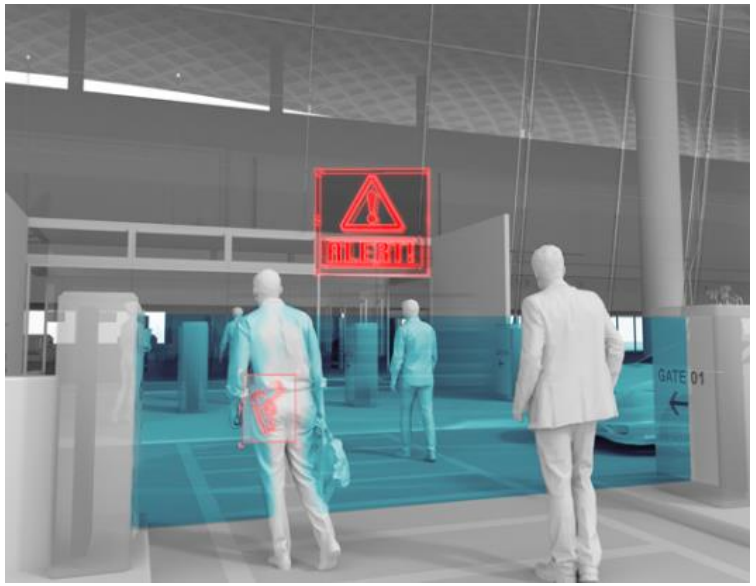
L’évolution du Big Data pose naturellement la question de la protection des données et du respect de la vie privée, comme vu dans le cours précédent.

Rappelez-vous :
Les données personnelles doivent être collectées dans un but bien précis, explicite et surtout légitime. La durée de conservation des données ne doit pas excéder l’atteinte de cet objectif.

Mais le Big Data peut aussi œuvrer pour la sécurité

Anticipation

Liberty Defense est une entreprise dédiée à la détection d'armes qui utilise l'IA pou réduire les crimes à main armée. Son système HEXWARE utilise l'imagerie 3D et l'IA pour détecter les menaces lorsque des groupes de personnes passent près de ses capteurs. Cette technologie peut être utilisée en intérieur ou en extérieur, et peut détecter aussi bien les armes métalliques que non métalliques.



Anticipation

Le logiciel de sécurité OPENALPR, utilisant les caméras de sécurité pour scanner les plaques d'immatriculation et fournir les données sur le véhicule en temps réel. Ce système peut être utilisé pour aider les autorités à détecter les véhicules de criminels. Plus de 9000 caméras compatibles sont installées dans 70 pays.

Anticipation

Hikvision qui représente 21,4% du marché des caméras de surveillance à l'échelle mondiale a commencé à intégrer l'IA à ses équipements. Ses nouvelles caméras peuvent scanner les plaques d'immatriculation, offrent une fonctionnalité de reconnaissance faciale pour chercher les criminels ou les personnes portées disparues, et détectent les anomalies telles que les bagages abandonnés ou même mesurer les températures corporelles

Anticipation

Au Royaume-Uni, la ville de **Durham utilise l'intelligence artificielle pour décider si un suspect peut être relaxé en attendant son procès**. Son programme " Harm Assessment Risk Tool " (Hart) a été nourri de données sur les crimes accumulées pendant cinq ans.

En se basant sur ces données, **l'IA détermine si un individu présente un risque faible, moyen ou élevé**. Le système est testé depuis 2013, et ses prédictions se révèlent exactes dans plus de 90% des cas en moyenne.

Mais se pose la question de la sécurité des IA : quelle légitimité dans la prise de décision? est-ce certifié ?

Le côté légal est juridique a déjà été balayé (d'autant plus avec la RGPD récemment), et que la loi Informatique et Libertés, par exemple, interdit qu'une machine puisse prendre des décisions pouvant avoir des ***« conséquences cruciales pour les personnes »***.

→ Quid de la mise en place d'une entité en charge de la sécurité des IA?

Place au Travaux Dirigés

Enoncé disponible : https://github.com/apatou