

Lexique des notions clés

30 notions - Données et création de valeur

Catégories

Conformité (2 notions)
Fil rouge du cours (1 notion)
Gouvernance (6 notions)
Infrastructure IA (5 notions)
Infrastructure données (3 notions)
Métiers de la donnée (1 notion)
Qualité & Risques (4 notions)
Souveraineté et Dependance (2 notions)
Stratégie (6 notions)

Glossaire - ordre alphabétique

AI Act - Conformité

Réglementation européenne sur l'IA (en application depuis 2024) classifiant les systèmes d'IA par niveau de risque et imposant des obligations proportionnelles.

A noter : Les usages à haut risque (scoring de crédit, sélection RH, décisions médicales) font l'objet d'obligations renforcées.

Algorithme - Infrastructure IA

Ensemble d'instructions mathématiques permettant à un système d'apprendre à partir de données et de produire des prédictions ou recommandations.

A noter : Statut stratégique : commodité. Accessible via API pour quelques centimes (OpenAI, Mistral, Anthropic...).

API - Infrastructure IA

Interface qui permet à un logiciel d'utiliser les fonctionnalités d'un autre système sans en connaître le fonctionnement interne.

A noter : Concrètement : vous appelez un modèle d'IA via une API sans avoir à l'héberger ni à le maintenir vous-même.

Biais algorithmique - Qualité & Risques

Distorsion systématique dans les décisions d'un modèle IA, causée par des biais présents dans les données d'entraînement.

A noter : Exemple : un scoring de crédit entraîné sur des données historiques reproduit les discriminations passées à grande échelle et automatiquement.

Chief Data Officer (CDO) - Gouvernance

Dirigeant en charge de la stratégie données à l'échelle de l'organisation. Arbitre entre les métiers, garant de la gouvernance globale, rend compte à la direction générale.

A noter : Rôle stratégique et décisionnel - pas technique.

Cloud Act - Souveraineté & Dépendance

Loi américaine (2018) autorisant les autorités US à accéder aux données stockées par des opérateurs américains, même sur des serveurs situés hors des États-Unis.

A noter : Des données hébergées chez AWS, Azure ou Google peuvent être accessibles à des autorités américaines sans notification.

Commodité (commoditization) - Stratégie

Processus par lequel un produit autrefois rare et différenciant devient standardisé, largement accessible et peu coûteux.

A noter : Dans l'IA, les algorithmes et la puissance de calcul sont des commodités. La donnée reste rare et propriétaire (cela pourrait d'ailleurs évoluer avec la montée de la donnée synthétique, qui permet de générer artificiellement des données d'entraînement - réduisant potentiellement l'avantage de ceux qui en ont accumulé le plus).

Data catalog - Gouvernance

Répertoire interne centralisant les métadonnées de chaque jeu de données : contenu, localisation, Data Owner, fréquence de mise à jour, droits d'accès, cadre légal.

A noter : Sans catalogue : chaque projet IA démarre en mode archéologie (semaines perdues). Avec catalogue : quelques heures.

Data lake - Infrastructure données

Système de stockage centralisé qui conserve les données brutes dans leur format natif, sans structure imposée à l'avance.

A noter : Différent du data warehouse : le data lake stocke tout, structure ou non, en attendant de savoir comment l'exploiter.

Data lineage (lignée de données) - Qualité & Risques

Traçabilité complète d'une donnée : d'où elle vient, qui l'a créée, comment elle a été transformée, qui l'a modifiée, dans quel but.

A noter : Sans data lineage, un modèle peut apprendre des patterns sans que personne ne sache précisément sur quoi il a été entraîné.

Data moat (fosse défensif) - Stratégie

Avantage compétitif créé par l'accumulation de données uniques et propriétaires, structurellement impossibles à répliquer rapidement, même avec un investissement important.

A noter : Exemples : Amazon (25 ans de comportements d'achat), Doctolib (RDV médicaux en France), Tesla (milliards de km de conduite réelle).

Data Owner - Gouvernance

Responsable métier d'un périmètre de données. Décide quoi collecter et dans quel but, définit les règles de qualité, arbitre les accès.

A noter : Rôle managérial - il vous appartient en tant que directeur d'un domaine, indépendamment de votre intitulé officiel.

Data scientist - Métiers de la donnée

Professionnel qui conçoit et entraîne des modèles d'IA, explore des données pour en extraire de la valeur, et traduit des problèmes métier en problèmes analytiques.

A noter : Réalité terrain : 60 à 80 % du temps est consacré à la préparation des données, pas à la modélisation
NB : ce chiffre tend à être GRANDEMENT diminué avec l'utilisation de Modèles spécialisés (Ex : Claude Code intégré à des outils tel que Databricks)

Data warehouse - Infrastructure données

Base de données structurée et optimisée pour les requêtes analytiques, alimentée par plusieurs systèmes sources après transformation.

A noter : Différent du data lake : le data warehouse impose une structure et un schéma définis à l'avance.

Donnée stratégique - Stratégie

Donnée construite intentionnellement par une organisation - définie, collectée et organisée dans un but précis - par opposition à la donnée subie.

A noter : La différence n'est pas technique. C'est une décision managériale.

Donnée subie - Stratégie

Donnée produite comme sous-produit involontaire de l'activité, sans intention ni organisation préalable. Elle existe mais n'est pas exploitée.

A noter : Symptômes : données éparpillées, pas de responsable identifié, pas de règles de qualité.

ERP (Enterprise Resource Planning) - Infrastructure données

Logiciel de gestion intégrée couvrant les processus financiers, logistiques, RH. Principale source de données structurées dans les grandes entreprises.

A noter : Souvent en silo avec les autres systèmes (CRM, outils métier), source fréquente d'incohérences.

Flywheel (roue d'inertie) - Stratégie

Boucle d'auto-renforcement : les données génèrent de meilleures performances, qui attirent plus d'utilisateurs, qui génèrent plus de données. La roue accélère.

A noter : Exemples : Netflix (temps de visionnage -> recommandations), TikTok (seconde regard -> ajustement temps réel).

Garbage in, garbage out - Qualité & Risques

Principe fondamental : la qualité de la sortie dépend directement de la qualité des données à en entrée. Un modèle entraîné sur de mauvaises données produit de mauvaises décisions.

A noter : Formulé dans les années 1960. Plus pertinent que jamais à l'ère de l'IA qui automatise les décisions à grande échelle.

Gouvernance des données - Gouvernance

Ensemble des règles, rôles et responsabilités qui définissent qui peut accéder à quelle donnée, dans quel état elle doit être, et qui rend des comptes.

A noter : Ce n'est pas un sujet technique. C'est un sujet de décision stratégique et de pouvoir organisationnel.

GPU - Infrastructure IA

Processeur spécialisé dans les calculs parallèles massifs, essentiel pour entraîner des modèles d'IA. Loué à la demande via les fournisseurs cloud.

A noter : Entraîner un LLM pour : des dizaines à centaines de millions de dollars. Utiliser un modèle via API : quelques centimes.

LLM (Large Language Model) - Infrastructure IA

Grand modèle de langage entraîné sur des corpus massifs de texte pour comprendre et générer du langage naturel. Fondement de ChatGPT, Gemini, Claude, Mistral...

A noter : Exemples : GPT (OpenAI), Gemini (Google), Claude (Anthropic), Llama (Meta), Mistral (français).

Métadonnée - Gouvernance

Donnée qui décrit une autre donnée : format, date de création, auteur, source, conditions d'utilisation. Matière première du data catalog.

Moat - Stratégie

Terme de stratégie compétitive (Warren Buffett) désignant ce qui protège durablement l'avantage d'une entreprise contre la concurrence.

A noter : Dans l'IA : la donnée est le moat le plus puissant car impossible à acheter rapidement. Voir aussi : Data moat.

RGPD - Conformité

Règlement européen (2018) encadrant la collecte, le traitement et la conservation des données personnelles des résidents européens. Sanctions jusqu'à 4 % du CA mondial.

A noter : Avant tout projet IA impliquant des données personnelles : vérifier le consentement, la finalité déclarée, la durée de conservation.

Silo organisationnel - Gouvernance

Cloisonnement entre départements qui stockent leurs données dans des systèmes séparés, avec des définitions différentes, sans partage ni vision consolidée.

A noter : Premier obstacle aux projets IA. La solution est une décision de gouvernance - pas un outil technologique.

Souveraineté des données - Souveraineté & Dépendance

Capacité d'une organisation à exercer un contrôle effectif sur ses données : où elles sont stockées, sous quelle juridiction, avec quels droits d'accès pour des tiers.

A noter : Héberger son data moat chez un opérateur étranger, c'est potentiellement perdre le contrôle de son principal avantage compétitif.

Traçabilité - Qualité & Risques

Capacité à retracer l'historique complet d'une donnée depuis sa création jusqu'à son utilisation. Voir aussi : Data lineage.

Valeur x Risque x Dépendance - Fil rouge du cours

Grille d'analyse du cours pour évaluer tout outil ou stratégie IA selon trois dimensions : la valeur créée, le risque exposé, et le niveau de dépendance induit.

A noter : Un manager qui optimise une dimension sans tenir compte des deux autres prend une décision incomplète. Toujours les trois.