



Introduction à la data

Yves-Armel Martin



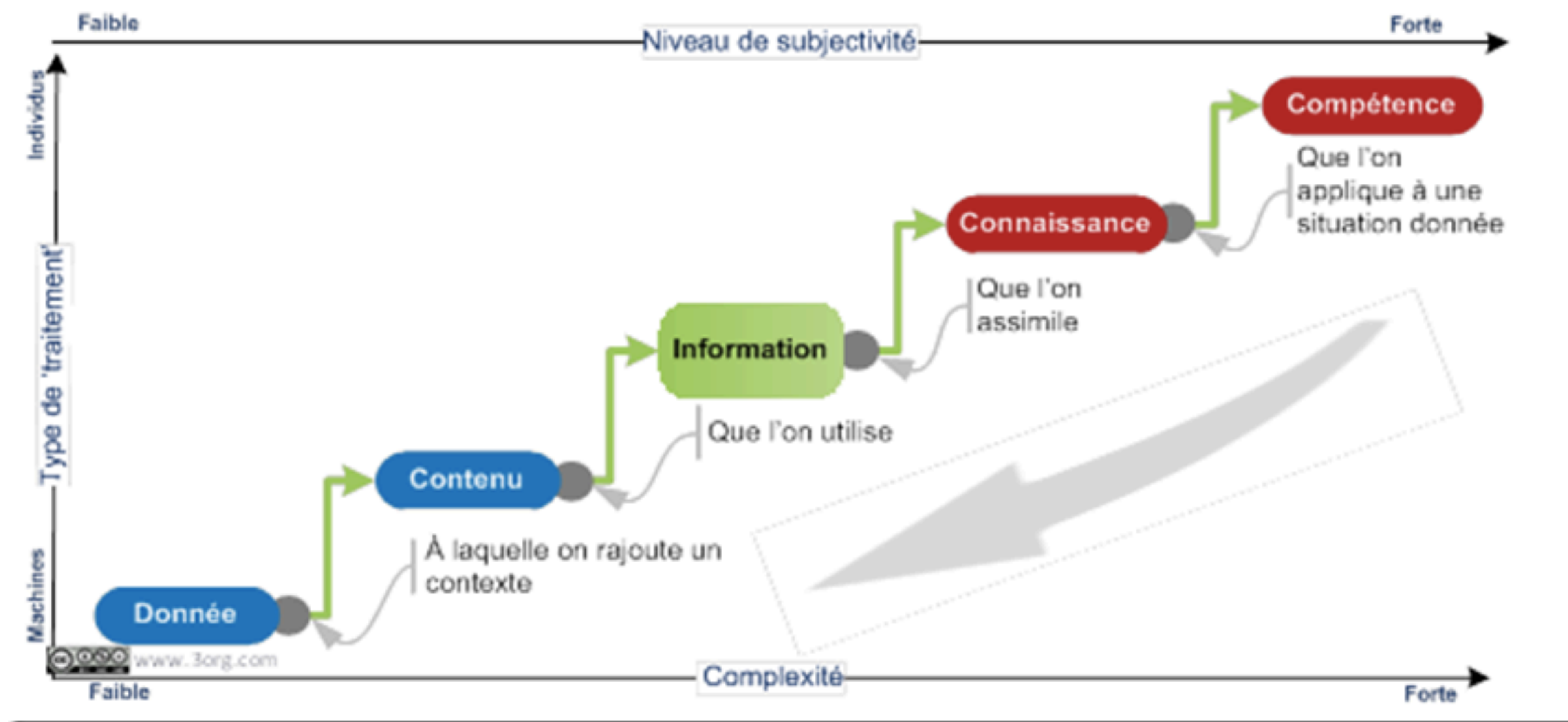
BUREAU

DES

POSSIBLES

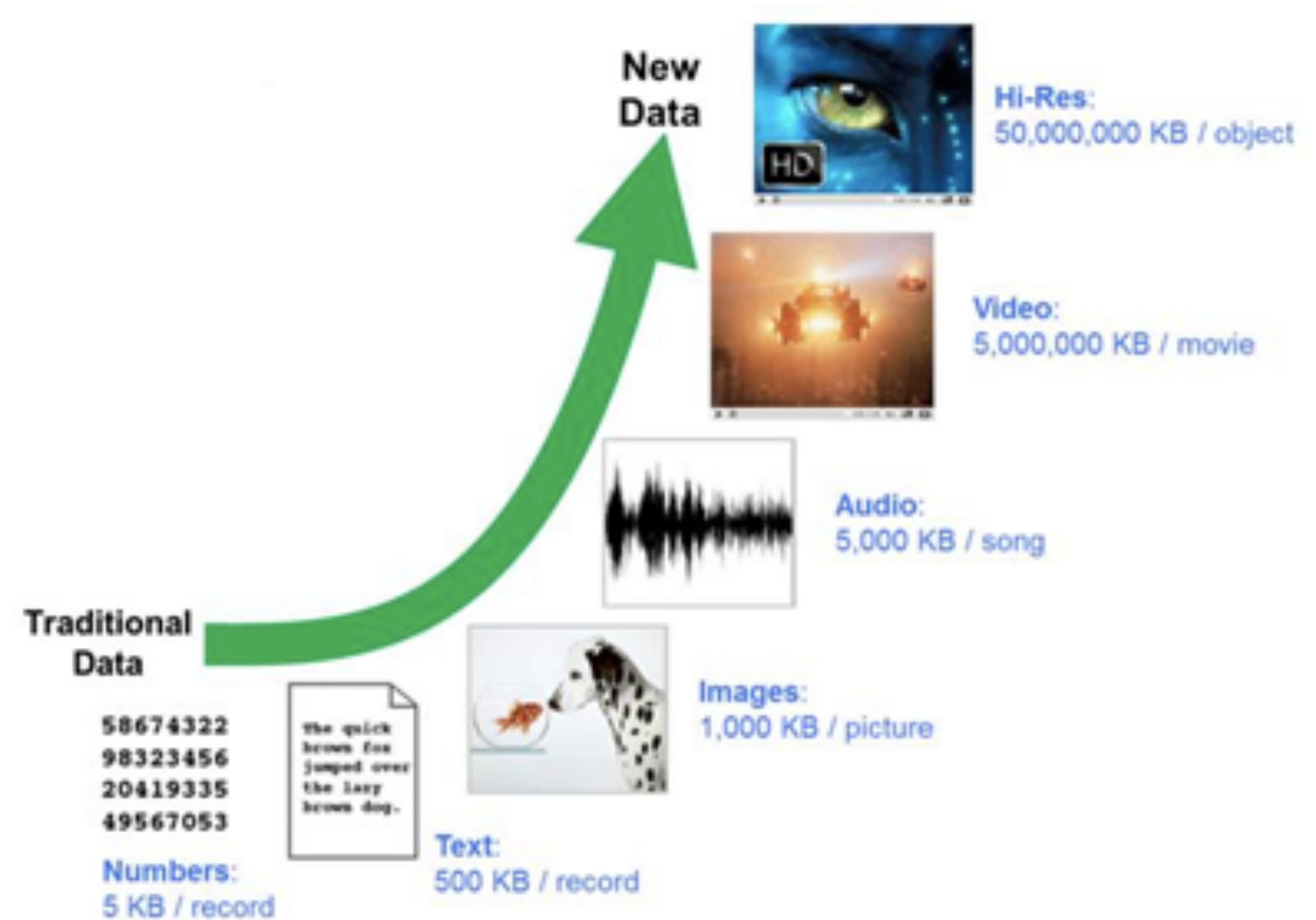
Une donnée

En science de l'information



La donnée en pratique

Des briques d'informations numériques



Produites par

- Des capteurs : ex données de températures, capteurs de santé
- Des machines : fichiers de vidéosurveillance
- Le travail humain : textes, plans, dessins...
- Les réseaux sociaux (travail humain caché)
- Les traces de nos activités (logs sur les serveurs)...
- ...

Non structurées -> brutes et sans liens

Structurées -> reliées et entourées de descripteurs pour les comprendre.



The logo for 'Open Data' features the word 'OPEN' in a dark blue, bold, sans-serif font on the top line. The letter 'O' is stylized as a power button symbol. Below it, the word 'DATA' is written in a teal, bold, sans-serif font. The letter 'I' is stylized as a vertical line with a small white dot at the top, resembling a data point or a vertical bar.

OPEN DATA

Open Data, ou données ouvertes, sont des données auxquelles l'accès est totalement public et libre de droit, au même titre que l'exploitation et la réutilisation.

Données ouvertes

- **Accessibles**
 - Les données doivent être pleinement accessibles. De préférence sur Internet (ou à coût de reproduction raisonnable)
 - **Exploitable**
 - Aisément manipulable par des programmes informatiques (ex : un PDF avec une liste de données n'est pas exploitable...)
 - **Libres de droits**
 - Non soumises au droit d'auteur ou au brevet
 - Droit d'utilisation et de redistribution (incluant le mélange avec d'autres ensembles de données)
 - Pas de restriction à certains secteurs ou usages non commerciaux
 - **Non propriétaires**
 - **Format propriétaire** : Un format propriétaire (ou format fermé) est un format de données dont les spécifications ne sont pas publiques, ou dont l'utilisation est restreinte par son propriétaire. Ses spécifications ne sont pas connues, ou soumise à un accord de non-divulgation.
Exemple : fichier adobe Indesign
-

Service public de la donnée

La loi pour une République Numérique d'octobre 2016 rend notamment **obligatoire** l'ouverture des données – l'open data – pour les collectivités de plus de 3 500 habitants (et les établissements publics et les organismes privés chargés d'un service public) au 1^{er} octobre 2018 (5% le faisait en mars 2018)

- Les administrations au sens large devront publier en ligne dans un standard ouvert leurs principaux documents, y compris leurs codes sources, ainsi que leurs bases de données et les données qui présentent un intérêt économique, social, sanitaire ou environnemental.
 - + mettre à disposition, en vue de faciliter leur réutilisation, les **jeux de données de référence** qui présentent le plus fort impact économique et social (9 à ce jour)
 - Etalab, en charge du portail **data.gouv.fr**, recense l'ensemble des jeux de données de référence.
-

À ce jour, neuf jeux de données, qui couvrent un large champ thématique ont été identifiés comme des données de référence.



Base Adresse Nationale (BAN)

La Base Adresse Nationale est une base de données qui a pour but de référencer l'intégralité des adresses du territoire

🚗 2 🔄 26 ★ 33 📍 France 📍 Point d'Intérêt



Base Sirene des entreprises et de leurs établissements (SIREN, SIRET)

La base Sirene contenant des données à caractère personnel, l'Insee attire votre attention sur les obligations légales qui en

🚗 476 🔄 42 ★ 58 📍 France



Code Officiel Géographique (COG)

Le code officiel géographique rassemble les codes et libellés des communes, des cantons, des arrondissements, des

🚗 19 🔄 6 ★ 17 📍 France 📍 Commune française



Plan Cadastral Informatisé (PCI)

Le plan cadastral est un assemblage d'environ 600 000 feuilles ou planches représentant chacune une section ou une

🚗 3 🔄 2 ★ 51 📍 France



Registre parcellaire graphique (RPG) : contours des parcelles et îlots cultureux et leur groupe de cultures majoritaire

Le registre parcellaire graphique est une base de données géographiques servant de référence à l'instruction des aides de la

🚗 15 🔄 2 ★ 7 📅 2013 à 2016 📍 France



Référentiel de l'organisation administrative de l'Etat

Le décret du 14 mars 2017 a institué le Service Public de la Donnée. Celui-ci met à la disposition du public 9 jeux de données

🚗 4 🔄 4 ★ 14 📅 12/05/2016 à 12/05/2026 📍 France



Référentiel à grande échelle (RGE)

L'État a confié à l'IGN le développement du référentiel à grande échelle (RGE). Pour ce faire, il fait appel à ses moyens propres

🚗 12 🔄 2 ★ 8 📍 France



Répertoire National des Associations (RNA)

Le Répertoire National des Associations (RNA) contient l'ensemble des associations relevant de la loi 1901, à savoir toutes les

🚗 23 🔄 5 ★ 20 📅 01/01/1901 à 05/04/2017 📍 France



Répertoire Opérationnel des Métiers et des Emplois (ROME)

Dans un contexte marqué par de fortes mutations de l'environnement économique et social, le ROME (Répertoire Opérationnel

🚗 5 🔄 2 ★ 8

Qu'entend-on par données publiques ?

- Données collectées par les organismes publics
 - Données non nominatives
 - un jeu de données publiques doit au préalable être anonymisé
 - Données ne relevant pas de la vie privée
 - Données de ne relevant pas de la sécurité
-

Données d'intérêt général

- Issues des **délégations de service public** (art. 10) : lorsqu'un organisme public telle qu'une municipalité choisit de déléguer à une entreprise la gestion d'un service public (eau, déchets, transports...), elle pourra exiger que l'entreprise lui fournisse les données détaillées collectées dans ce cadre.
 - Données essentielles des **conventions de subvention** (art 11) : lorsqu'un organisme public verse une subvention supérieure à 23 000 euros, il devra désormais publier les données essentielles de la convention de subvention (telles que le nom de l'organisme bénéficiaire, le montant, l'objet ou la durée de la convention).
 - Transmission de certaines **bases de données des entreprises** pour les enquêtes de l'Insee (art 12)
-

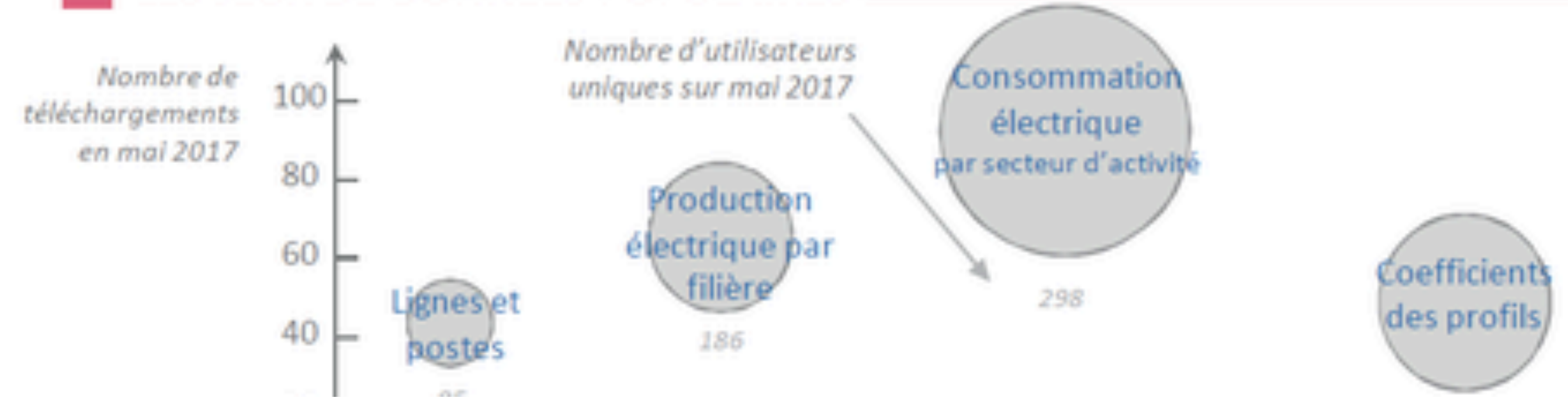


Nombre de données, pour comparaison :



Nombre de jeux de données et nombre d'enregistrements publiés au 1er juin 2017

LES JEUX DE DONNEES POPULAIRES



Bien distinguer :

Les producteurs produisent la donnée de référence et documentent les métadonnées.
Ils traitent les remontées des utilisateurs ou le cas échéant, orientent les utilisateurs vers le service compétent pour traiter ces demandes.
Ils prennent des engagements sur la mise à jour des données.
Ils désignent les diffuseurs pour chacun des jeux de données de référence.

Les diffuseurs mettent à disposition les données avec un haut niveau de qualité.
Ils s'engagent sur des niveaux de performance et de disponibilité.



Plateformes publiques



Elles collectent des données (les leurs ou de tiers), les référencent et les mettent à disposition de tous sur Internet.

Ce sont des plateformes logicielles (certaines en open source, d'autres portées par des éditeurs).

Etalab - data.gouv.fr

Nantes métropole, le département de Loire-Atlantique et la région Pays-de-la-Loire ouvrent leur portail mutualisé à toutes les communes du territoire.

Grand Lyon / Région : deux plateformes en concurrence

Plusieurs départements ont leur plateforme... parfois difficile à faire vivre : Seine St Denis, Côtes d'Armor, Manche, ...



API : interface de programmation.

Elles facilitent l'intégration dans des développements et permettent d'intégrer des mises à jours régulières des données.

► Rendre les données interopérables

- 1 API transport
- 2 API quantified-self
- 3 API trafic
- 4 API maison connectée
- 5 API réseaux des villes
- 6 API énergie





TROUVER **UNE DONNÉE**

ex : chantier, transport, cadastre

RECHERCHER 🔍

RECHERCHE AVANCÉE ▼

TOUTES LES DONNÉES

1048 DONNÉES
DISPONIBLES

Date de publication

OK

**CATALOGUE DE
DONNÉES**

Catégories : 

Charte graphique et gabarit utilisés pour les prestations topographiques de la Métropole de Lyon

Ces documents techniques sont produits par l'unité « Topographie et délimitation du domaine ... »

Fournisseur : *Métropole de Lyon / Direction Innovation*

Catégories :  

Comptage Vélo de la Métropole de Lyon

Informations sur : • les sites, qui permettent de récupérer les informations relatives aux ...

Fournisseur : *Métropole de Lyon / Direction de la voirie (DV)*

Licence : [Licence Ouverte](#)





Téléchargez la liste de l'ensemble des données de Data Grand Lyon

RÉSULTAT DE LA RECHERCHE

14 DONNÉES DISPONIBLES

Date de publication

OK

Catégories :  

Comptage Vélo de la Métropole de Lyon

Informations sur : • les sites, qui permettent de récupérer les informations relatives aux ...

Fournisseur : *Métropole de Lyon / Direction de la voirie (DV)*

Licence : [Licence Ouverte](#)

👁 Consultée 218 fois

Catégories :    

Historique des disponibilités des stations Vélo'v de la Métropole de Lyon


La donnée historique des disponibilités des stations Vélo'v propose un ...

Fournisseur : *Métropole de Lyon / Direction de la voirie (DV),*

JCDécaux

Licence : [Licence Ouverte](#)

👁 Consultée 2247 fois

Catégories : 






Prochain passage temps réel du service Rhônexpress (Métropole de Lyon et Département du Rhône)

Les prochains passages sont les horaires estimés des deux prochaines dessertes de la ...

Fournisseur : *Rhônexpress (Administrateur opendata)*

Licence : [Licence Associée](#)

👁 Consultée 599 fois

Catégories :   



Parcs Relais du réseau TCL (Disponibilités temps réel)

Les parcs relais du réseau TCL sont des objets ponctuels qui localisent les ...

Fournisseur : *SYTRAL*

Licence : [Licence Associée](#)

👁 Consultée 19033 fois

Données temps réel grand lyon - uniquement 14 jeux de données



Big Data / Megadonnées

- Des données extrêmement volumineuses
- Exploitable uniquement par des machines
- Mais porteuses d'informations nouvelles, inaccessibles auparavant





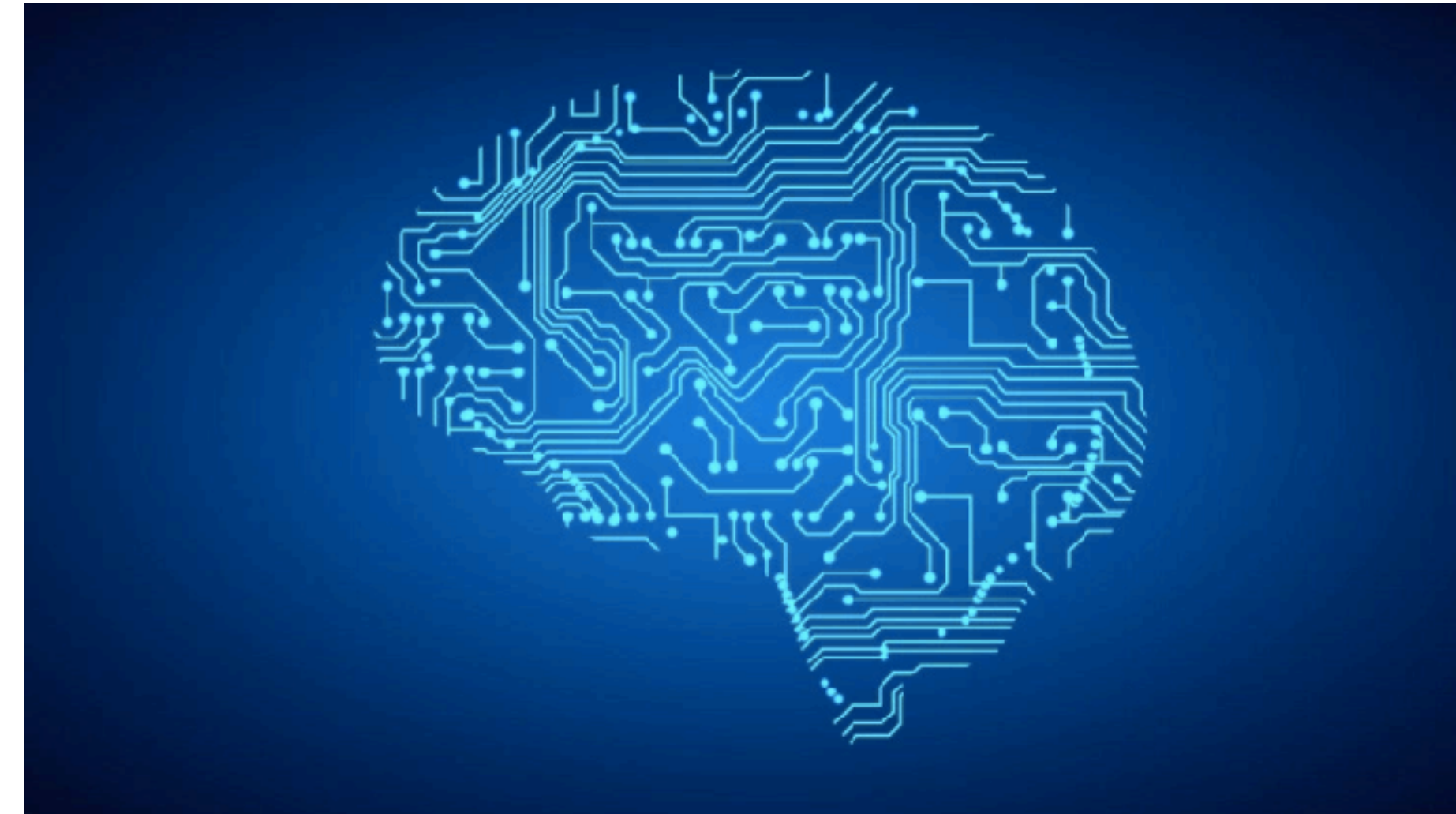
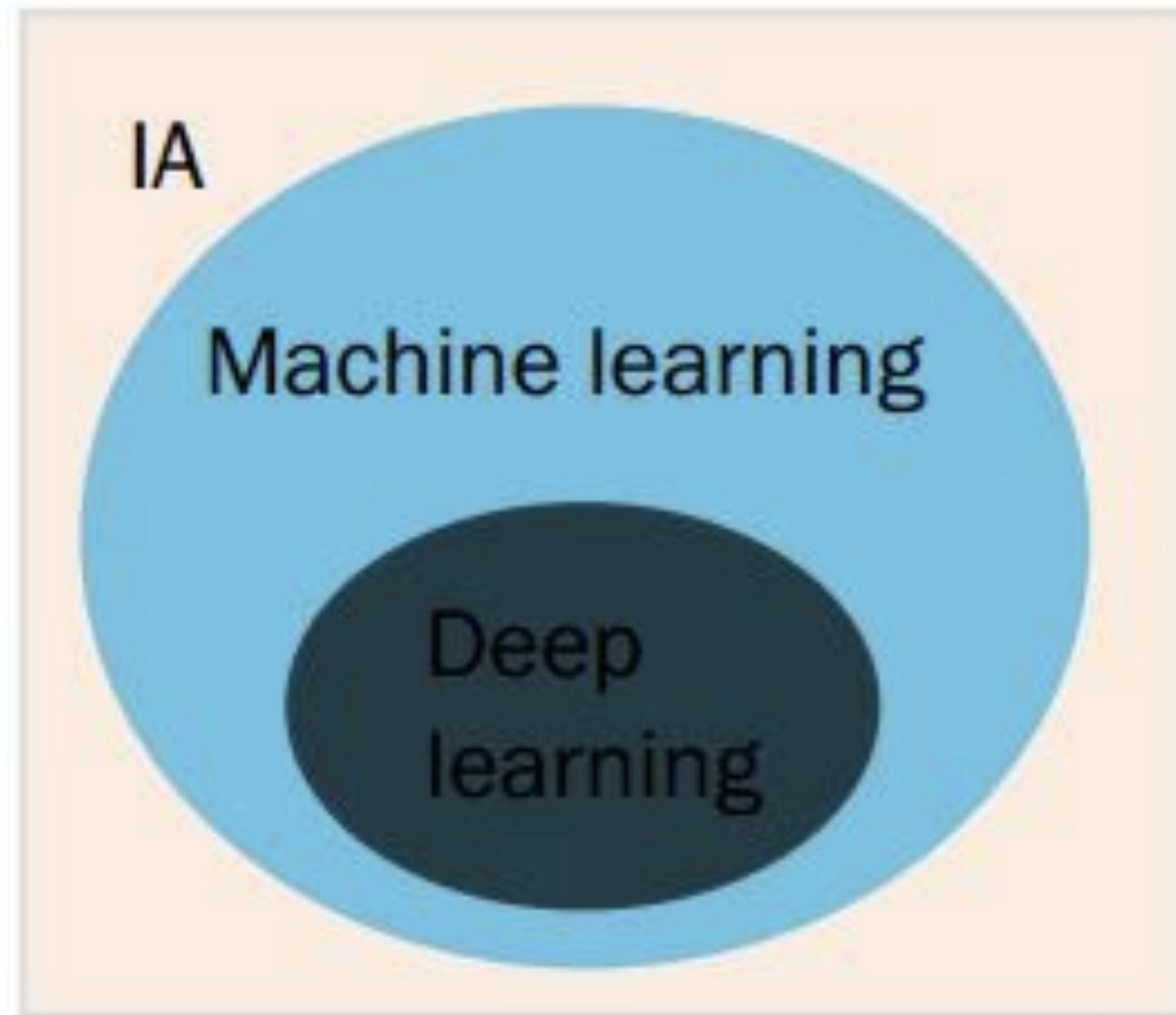
Cognitive computing

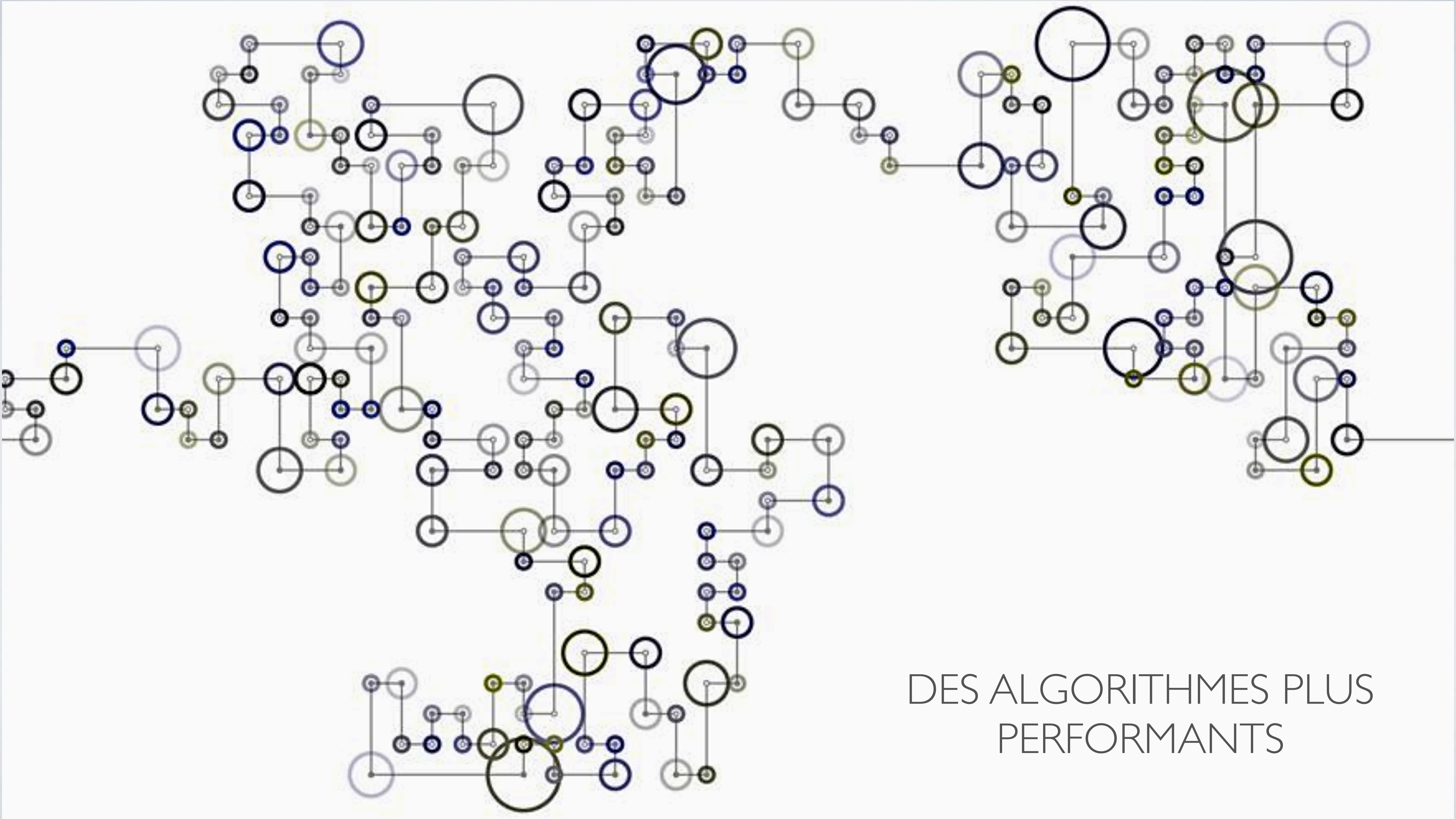
(Alias Intelligence artificielle)



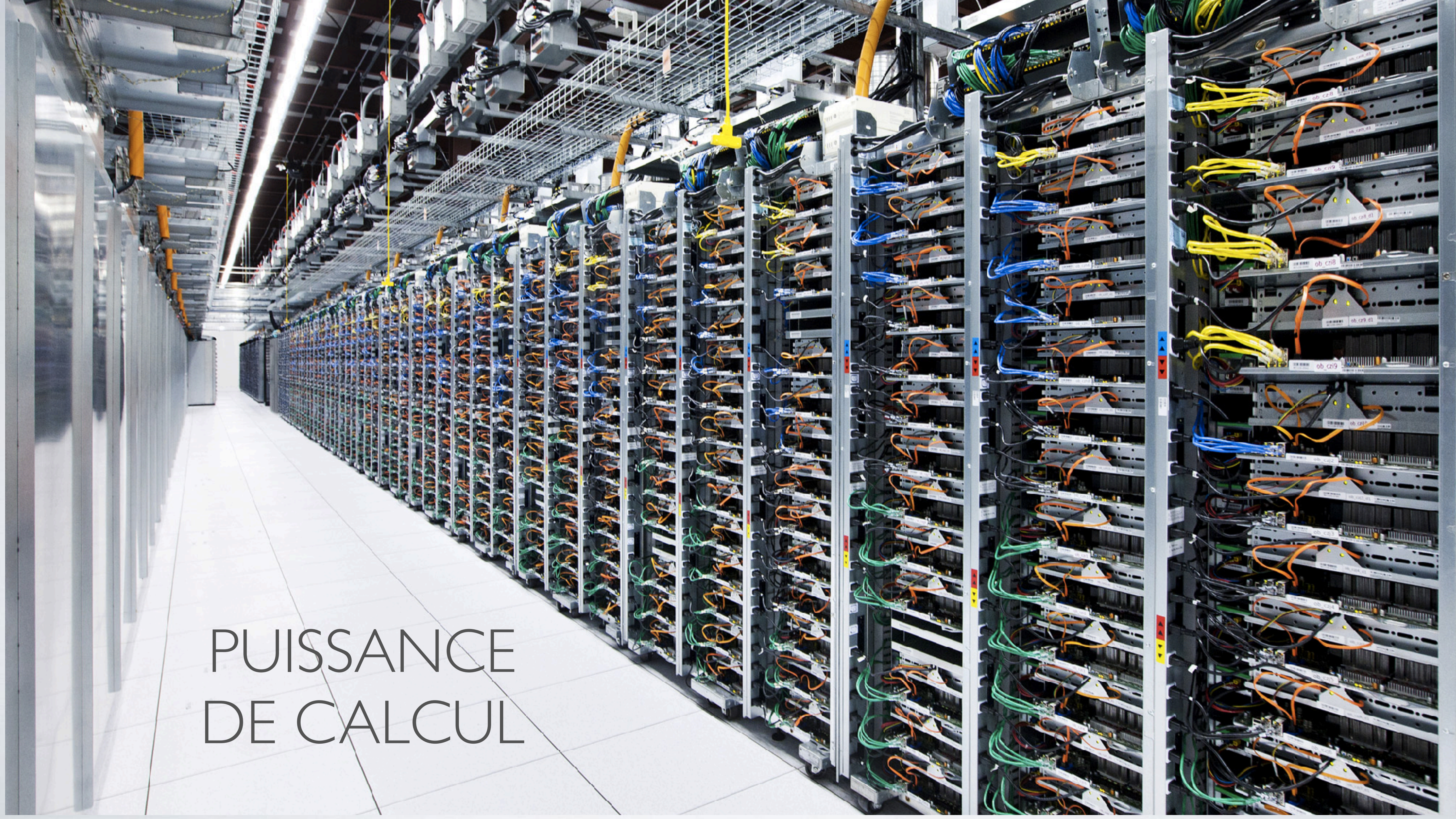
Qu'est-ce que c'est ?

un ensemble de technologies





DES ALGORITHMES PLUS
PERFORMANTS



PUISSANCE
DE CALCUL

An iceberg floating in a blue ocean. The tip of the iceberg is above the water line, while the much larger, submerged part is below. The sky is a lighter blue with some clouds.

< WHAT WE KNOW...

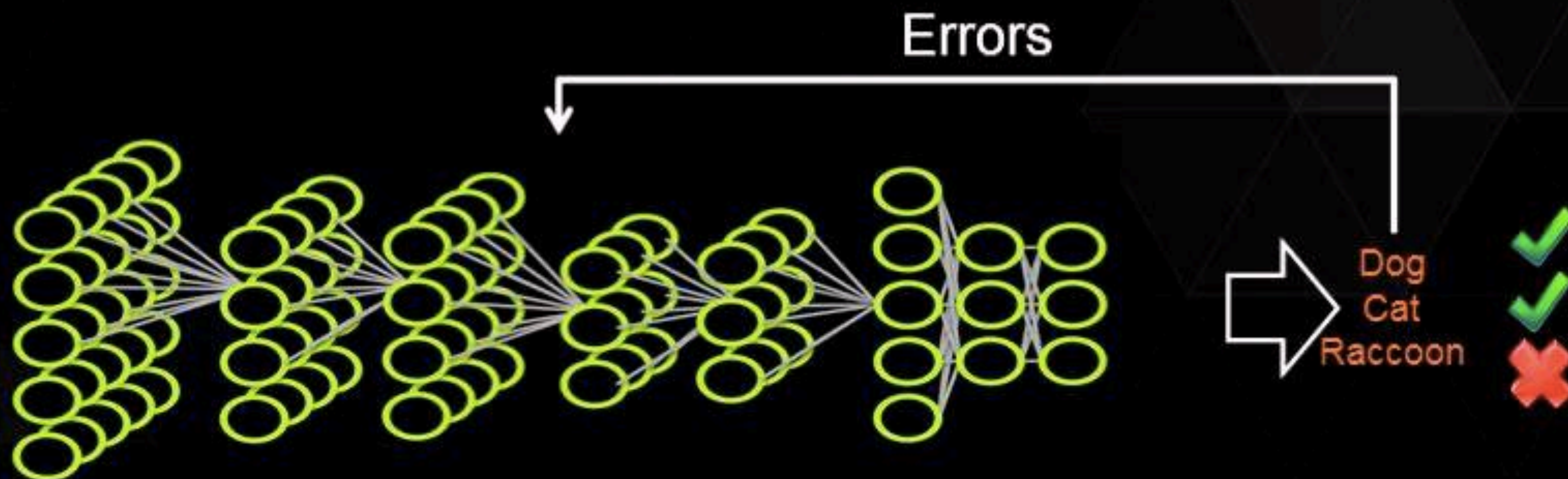
< THE REST...

BIG DATA

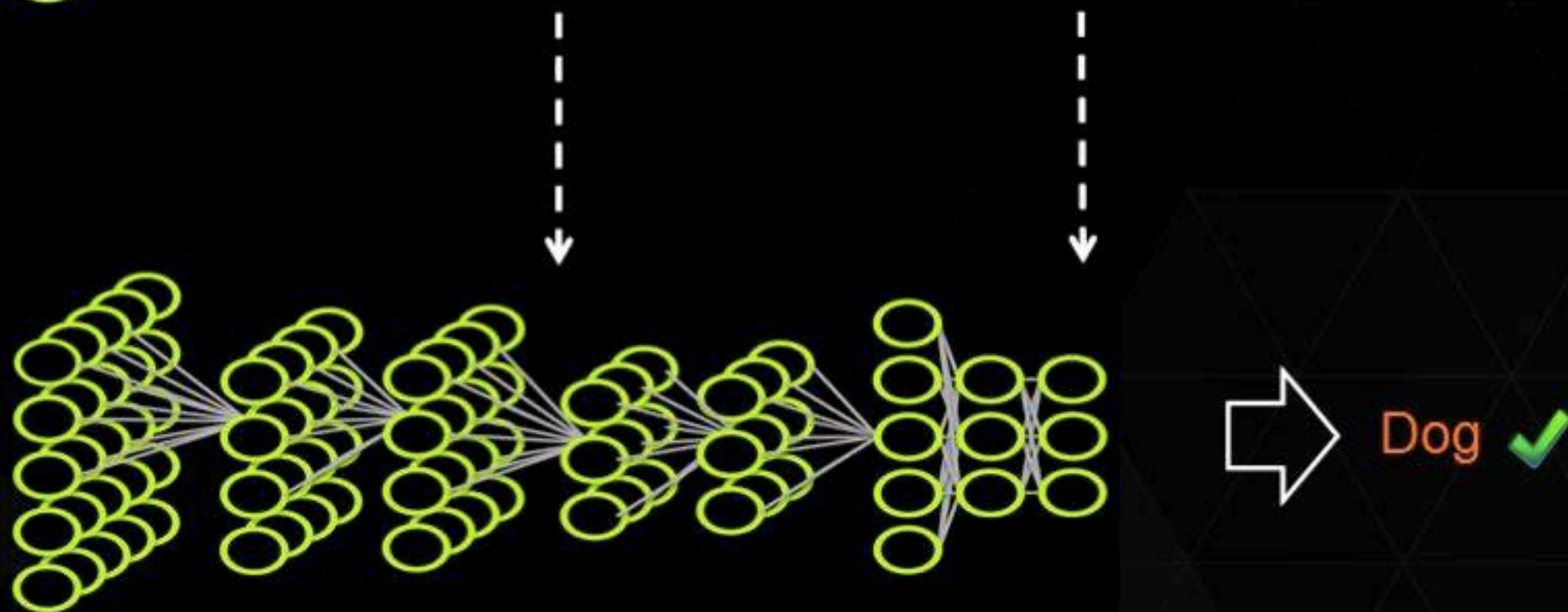
DES DONNÉES

DEEP LEARNING APPROACH

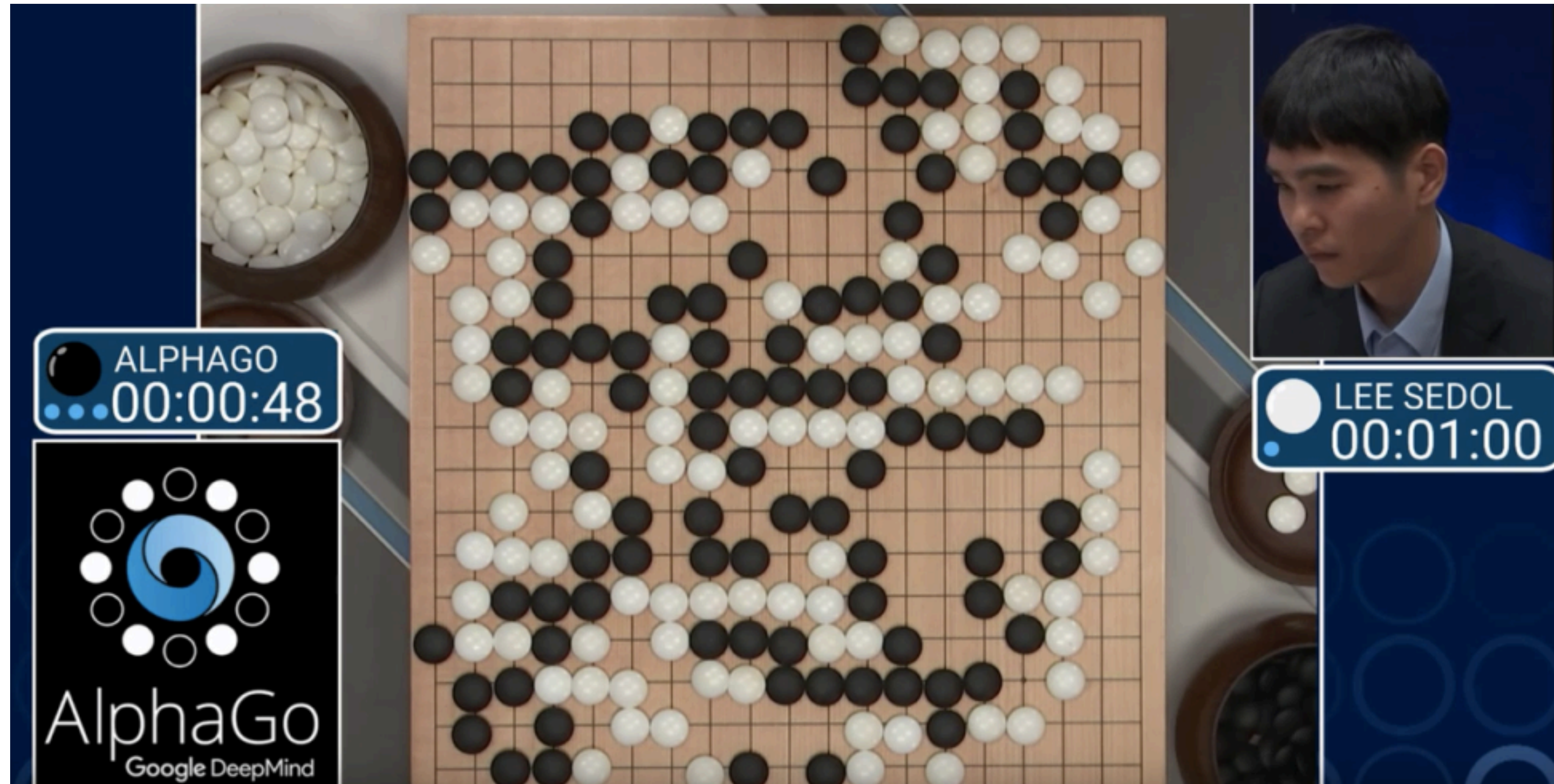
Train:



Deploy:



Une prise de conscience récente

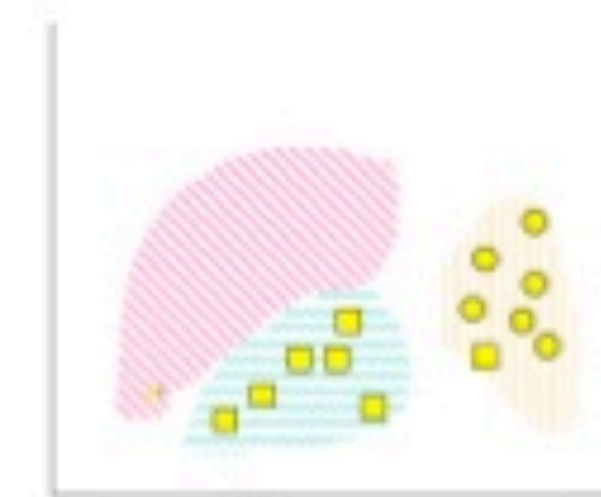
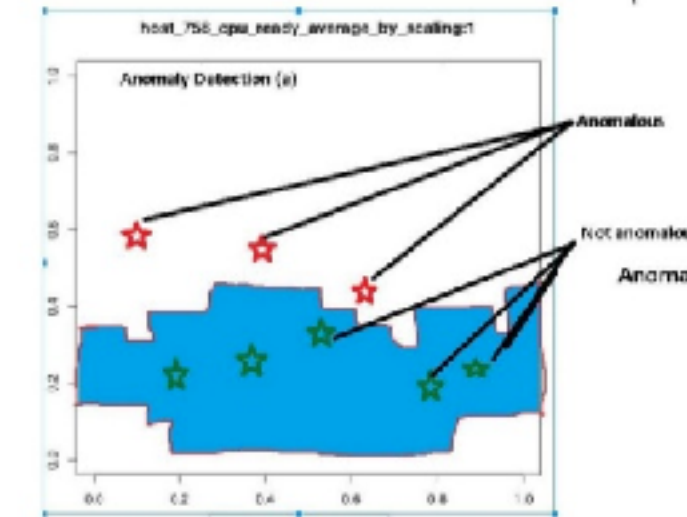
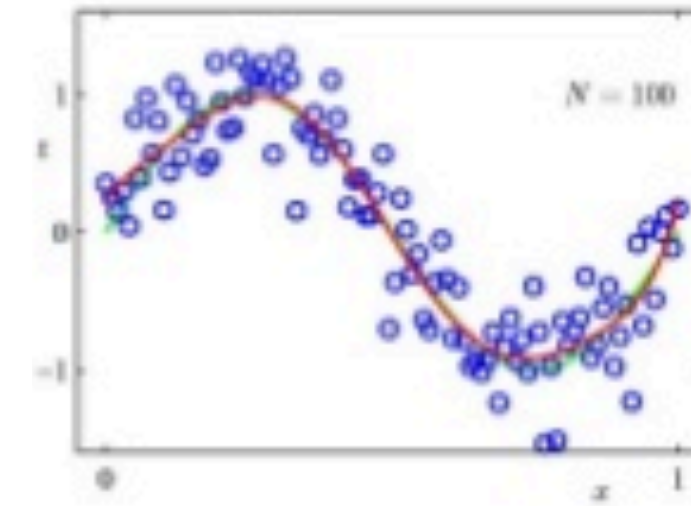


Types d'approches IA



Usages IA & Data

- prédire
- identifier les anomalies
- rapprocher les ressemblances
- classer automatiquement (et simplifier la saisie)
- reconnaître des formes...





A LA MANIÈRE DE

image



flowmachines

A LA MANIÈRE DE

musique

« Daddy's Car » : les beatles continués

Nécessité de qualifier les données

Volume : quelle quantité de données, quel domaine couvert ?

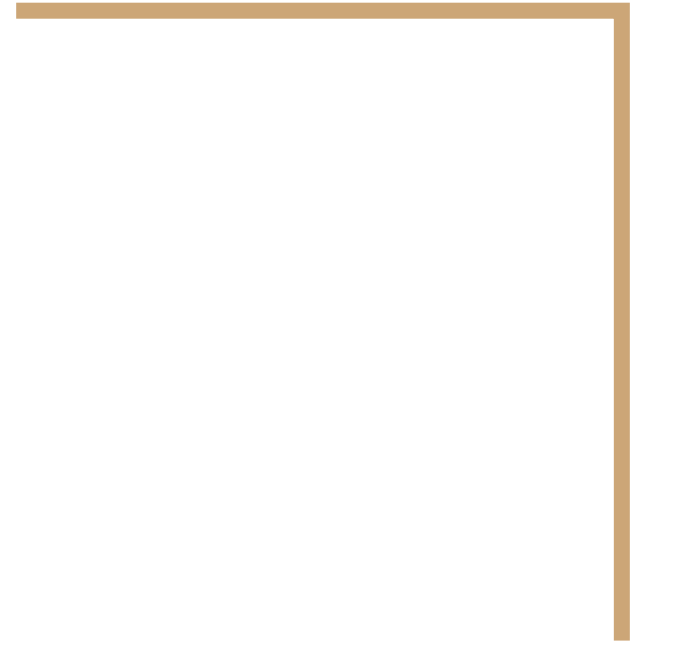
Variété des données : les données sont-elles diverses et de nature différentes (ouvertes ou privées, image, localisation, textes etc).

Vélocité : fréquence à laquelle les données sont générées, capturées, partagées et mises à jour

Véracité : qualité des données et intégrité des données (garantir que les données initiales n'ont pas été modifiées)

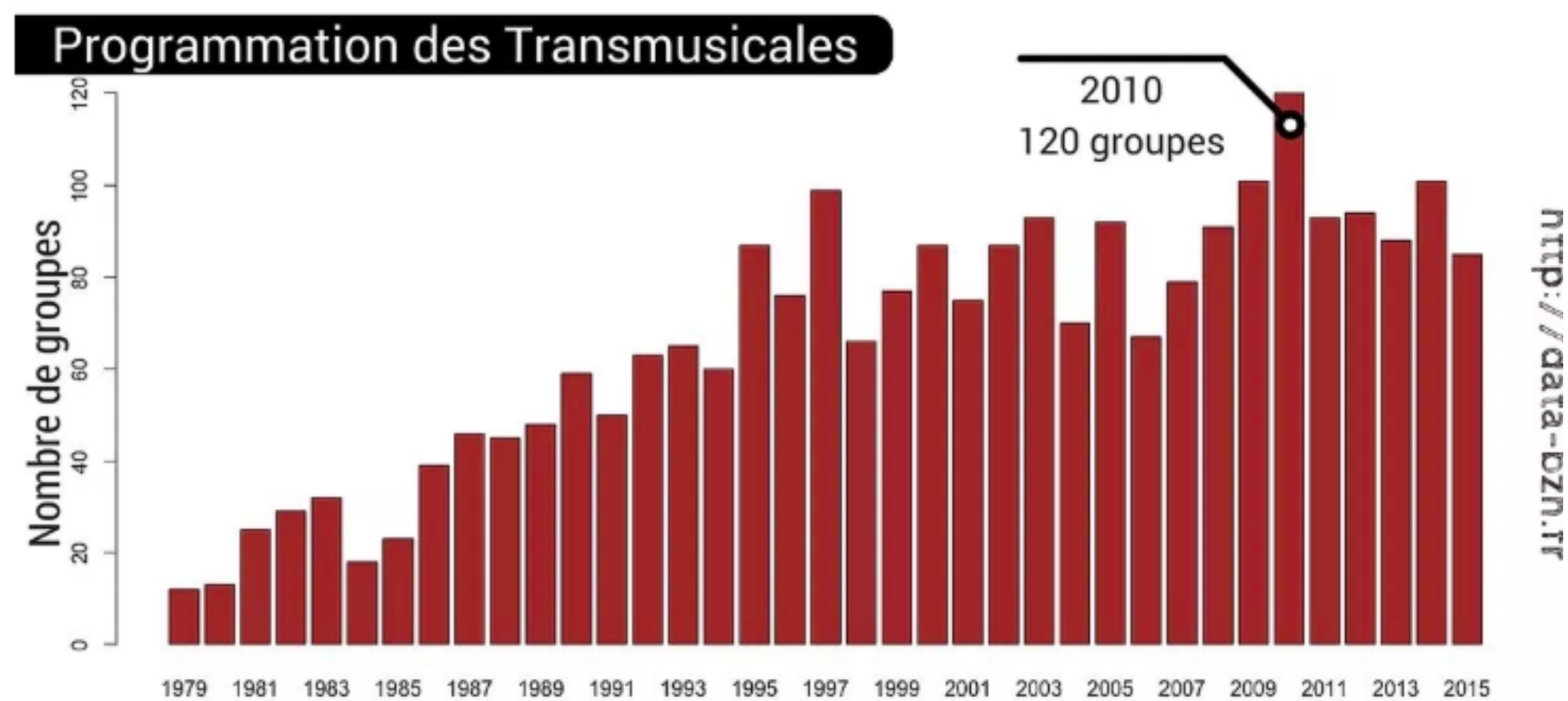


Quelques usages

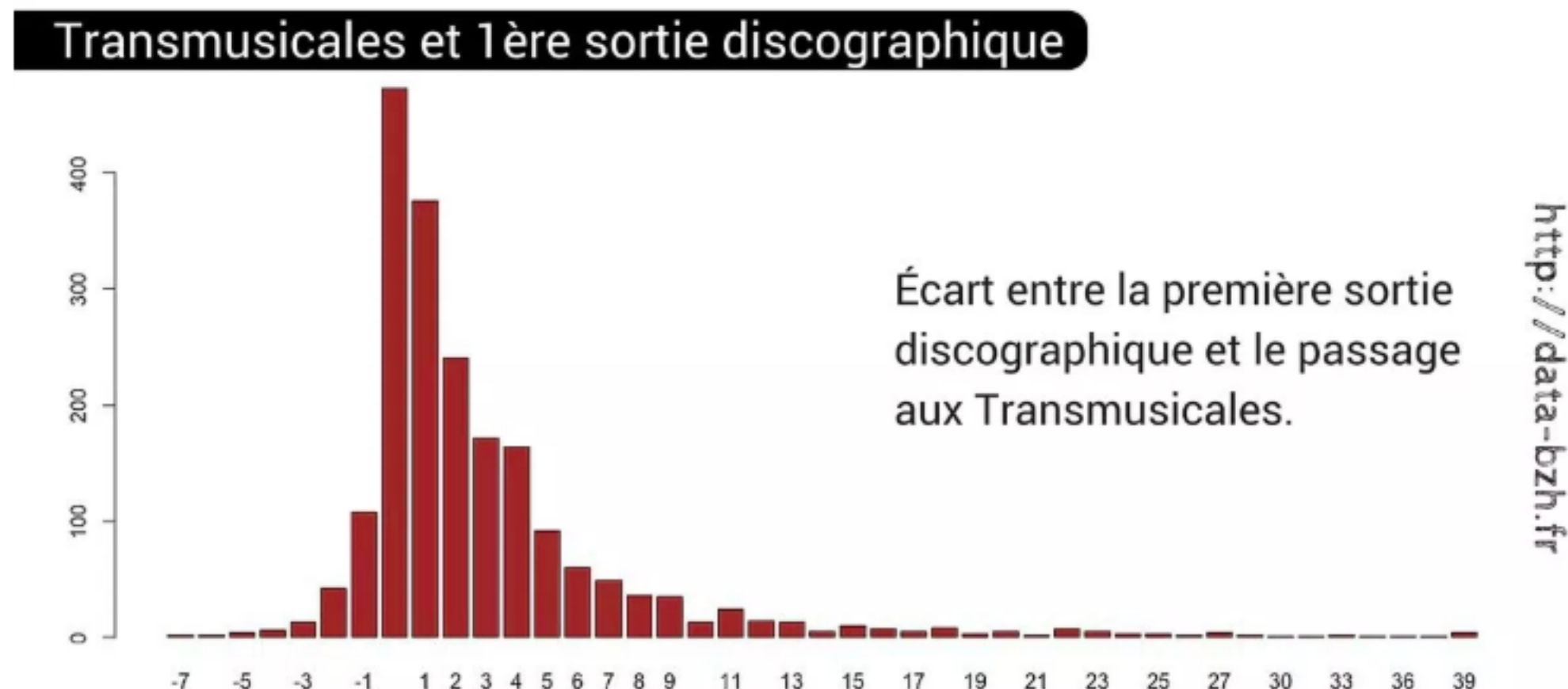


Usage : la dataviz

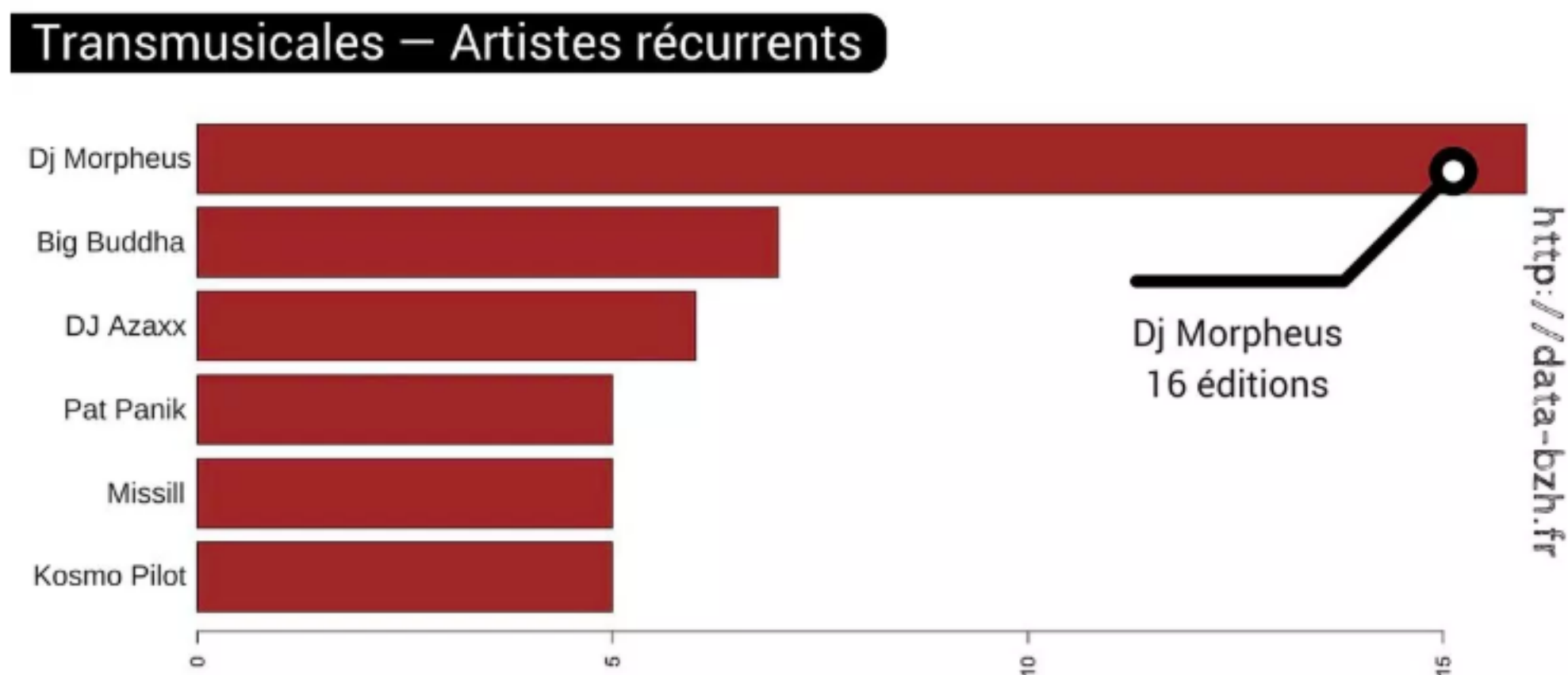
ÉVOLUTION DE LA PROGRAMMATION DES TRANSMUSICALES DE RENNE



ÉCART ENTRE LA PREMIÈRE SORTIE DISCOGRAPHIQUE ET LE PASSAGE AUX TRANSMUSICALES



LES ARTISTES RÉCURRENTS DANS LA PROGRAMMATION DE TRANSMUSICALES



Réutilisation des données fournies par l'asso des Trans Musicale à Etalab, par data- bzh

LA DATAVIZ DE LA DATAVIZ

la fonderie
Agence numérique

Infographie réalisée par
WEDODATA



LES SOURCES

Plusieurs données font du Big Data : l'ensemble, particulier ou non, des données, l'ensemble des données plus ou moins structurées (graphiques, textes, images, vidéos, etc.) de vidéos, de photos, de textes en ligne...

Interviews Open Data
Crowdsourcing Insee Eurostat sondages Enquêtes
statistiques Rapport textes
Administration
données propriétaires Experts Entreprises

LES FORMATS

Il est essentiel de jongler dans les formats de fichiers. Le premier travail du data analyst est de nettoyer tous ces fichiers diversifiés avant de pouvoir faire parler les données.



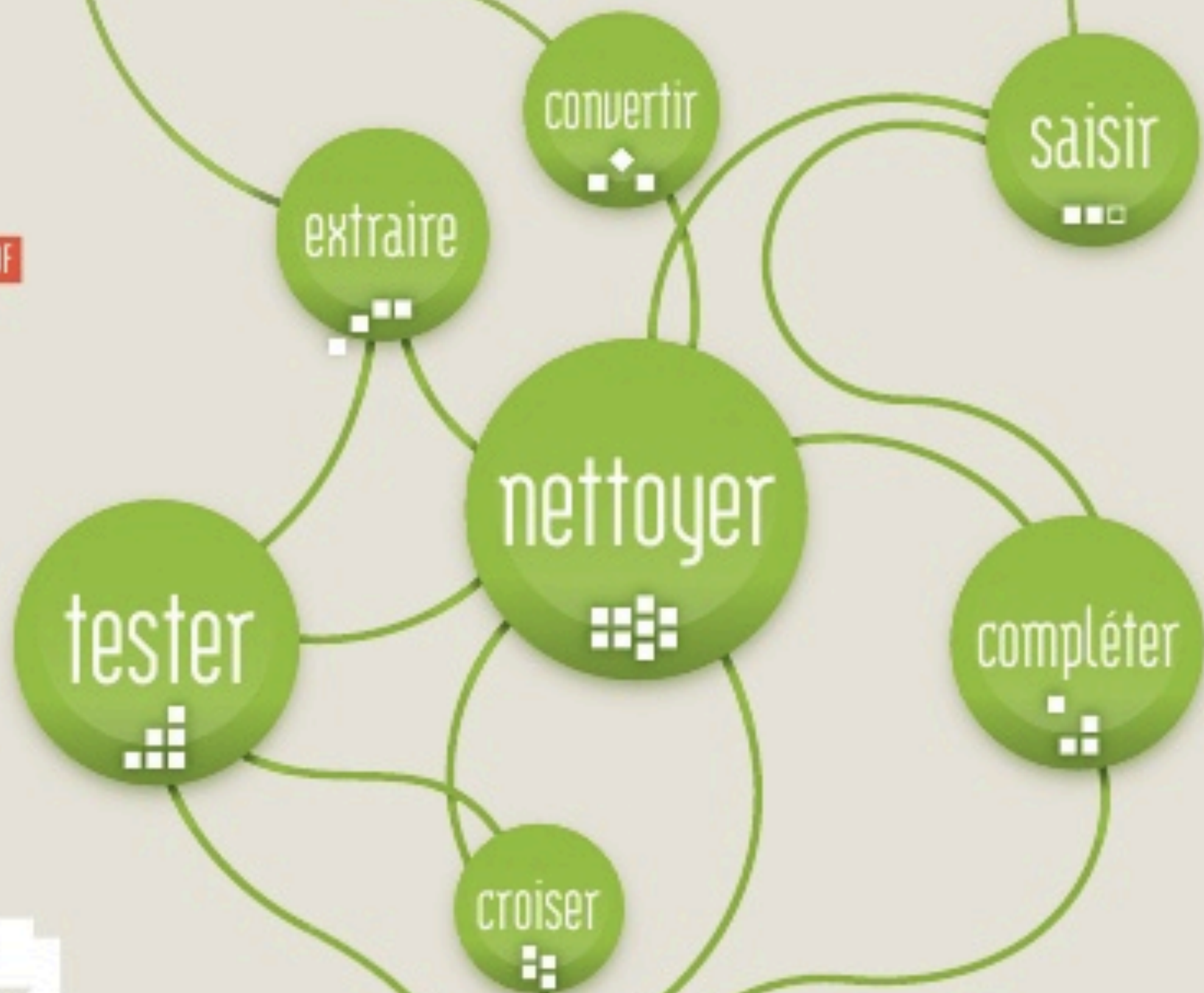
LES OUTILS

LES COMPÉTENCES

Comment travailler avec les données et les analyser ? La question est plus complexe qu'elle n'y paraît. Elle implique de maîtriser les outils et de savoir les utiliser pour exploiter les données de manière optimale.



LES BESOINS



LES OUTILS

Quelle est la forme de la communication d'information de votre projet ?
 Comment présenter des tableaux complexes ou graphiques ?
 Comment faire des données d'accès en ligne ?
 Comment présenter des données complexes, statiques et interactives tout cela ?



croiser

données

choisissez une forme de dataviz

LES BESOINS

Il y a pas de bonne dataviz seule la chose la plus adaptée à votre projet de communication de données. Le graphique peut aussi influencer la lecture.



fixe

Pour une technique de l'infographie de données, la dataviz fixe est une solution privilégiée car elle permet de visualiser les données.

ÉTAPES DE CRÉATION



vidéo

Pour une technique de l'infographie de données, la dataviz vidéo est une solution privilégiée car elle permet de visualiser les données.

SCÉNARIO



photo

Pour une technique de l'infographie de données, la dataviz photo est une solution privilégiée car elle permet de visualiser les données.

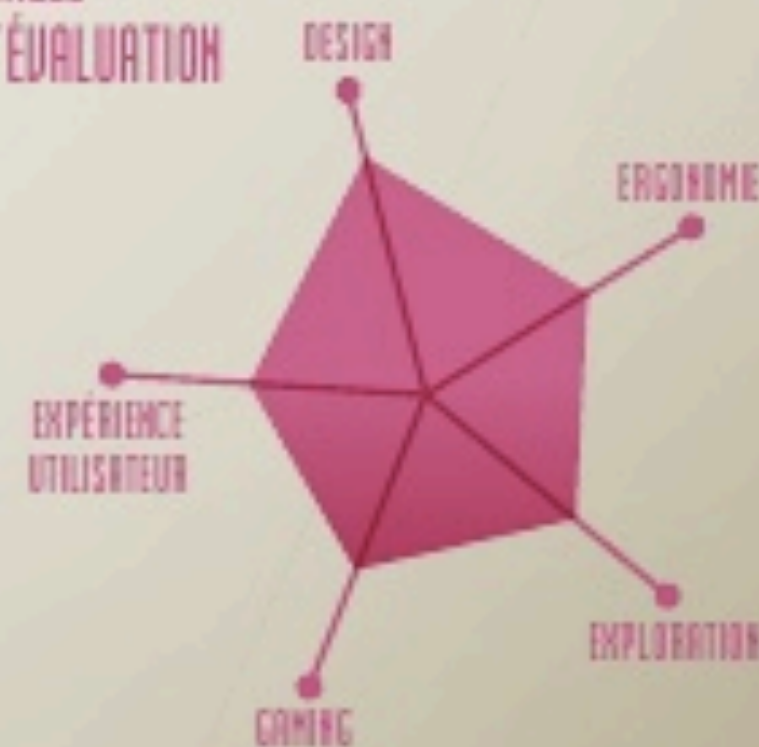
DISPOSITIF DE BASE



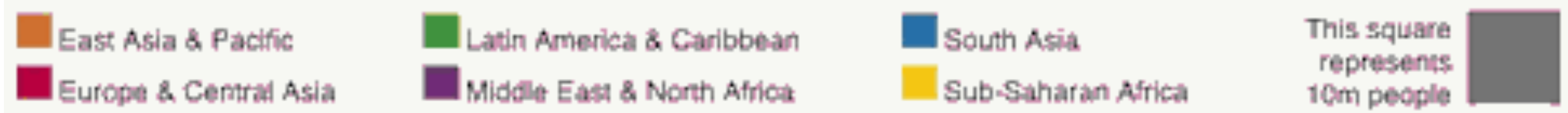
interactive

Pour une technique de l'infographie de données, la dataviz interactive est une solution privilégiée car elle permet de visualiser les données.

GRILLE D'ÉVALUATION



People living in extreme poverty,



||

1980 1993 1998 2002 2005 2008 2010 2011 2012 2013

Usage : Qu'est-ce qu'un datathon ?

Un sprint collaboratif (1, 2 ou 3 jours en général)

A partir de jeux de données du ou des commanditaires

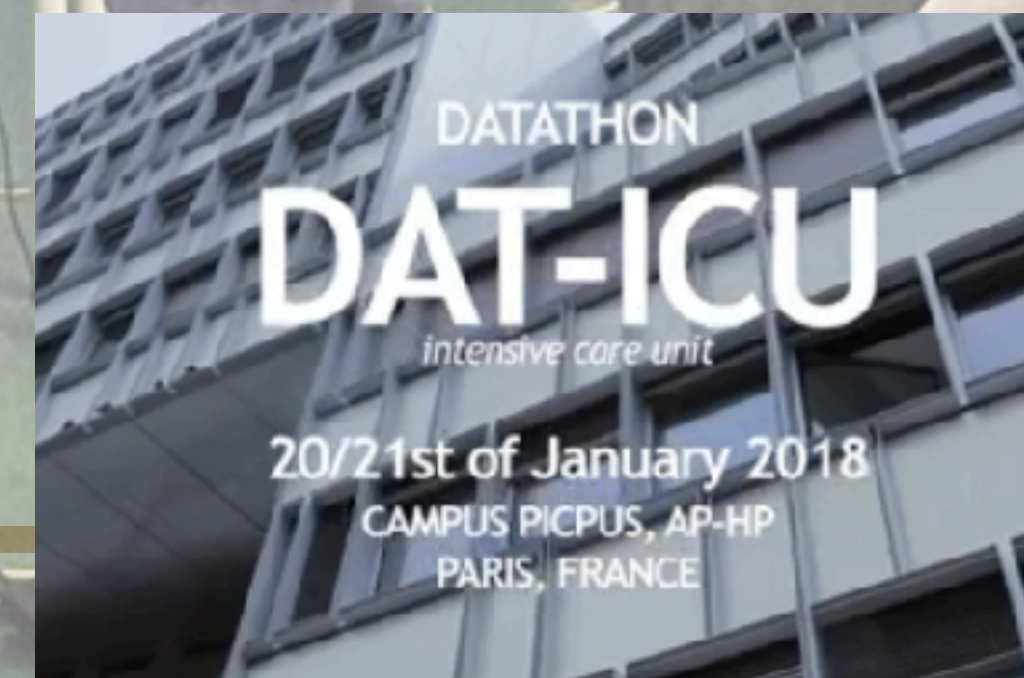
Ouvert à un collectif de volontaires (start up , particuliers, étudiants)

Pour produire des applications, services, dataviz

Variante le Liberathon : un sprint pour produire et libérer des données ou des contenus



Ne pas oublier de mettre des sous titre en anglais...

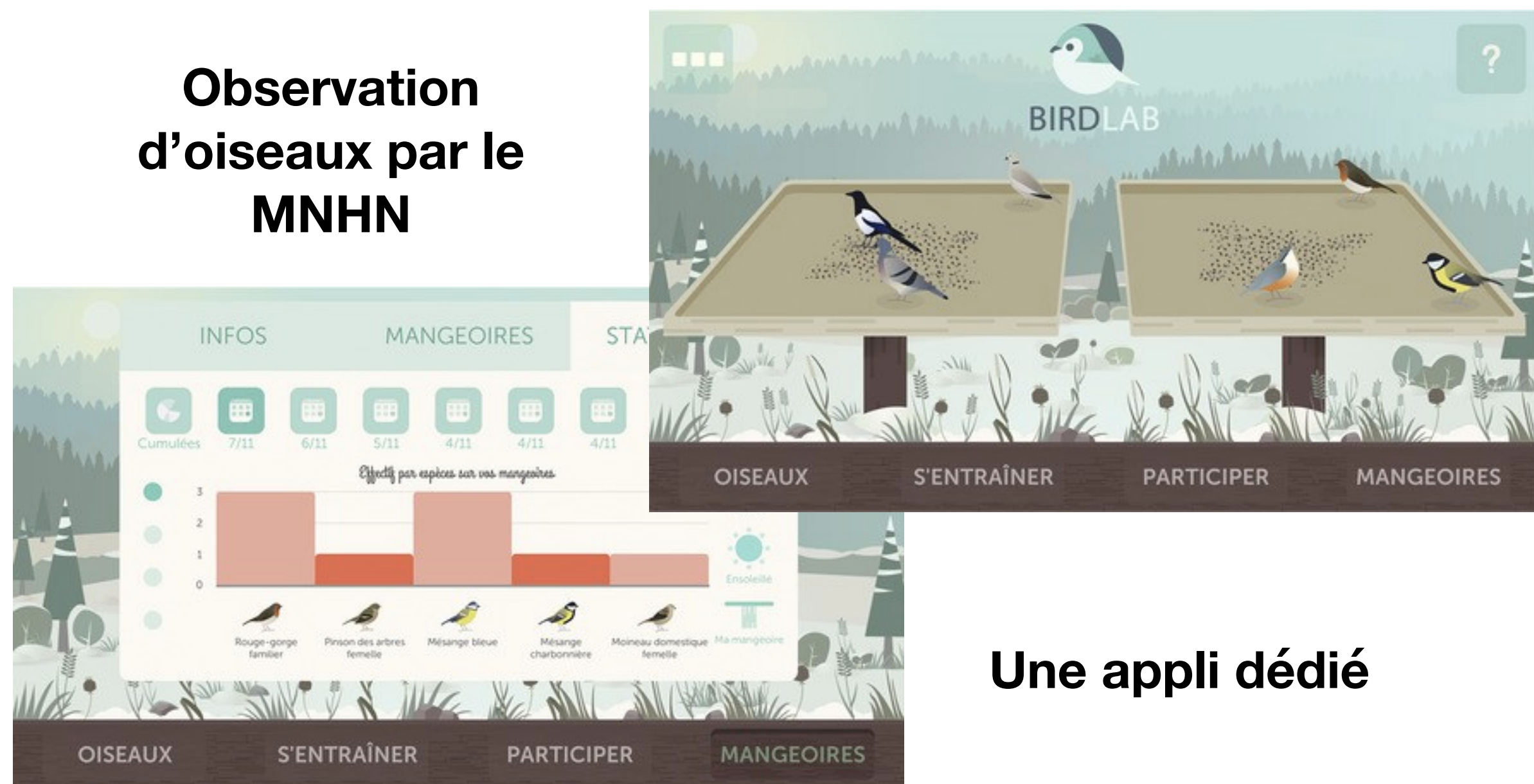


Crowdsourcing et data

Ou production participative.

Inviter les citoyens à produire des datas ouvertes.

Exemples :



Une appli dédié

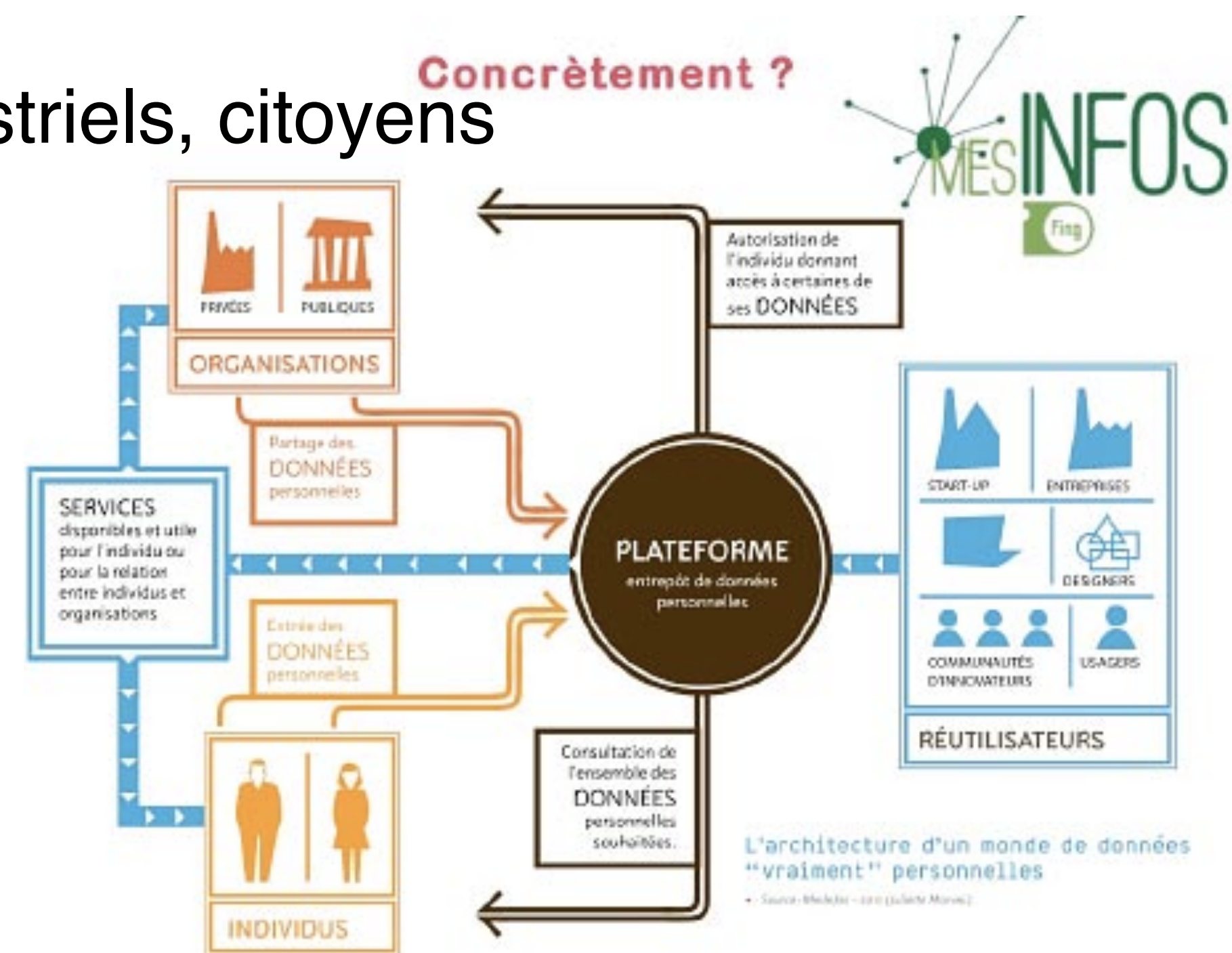


ANFR : une appli pour mesurer et partager le débit des opérateurs mobiles

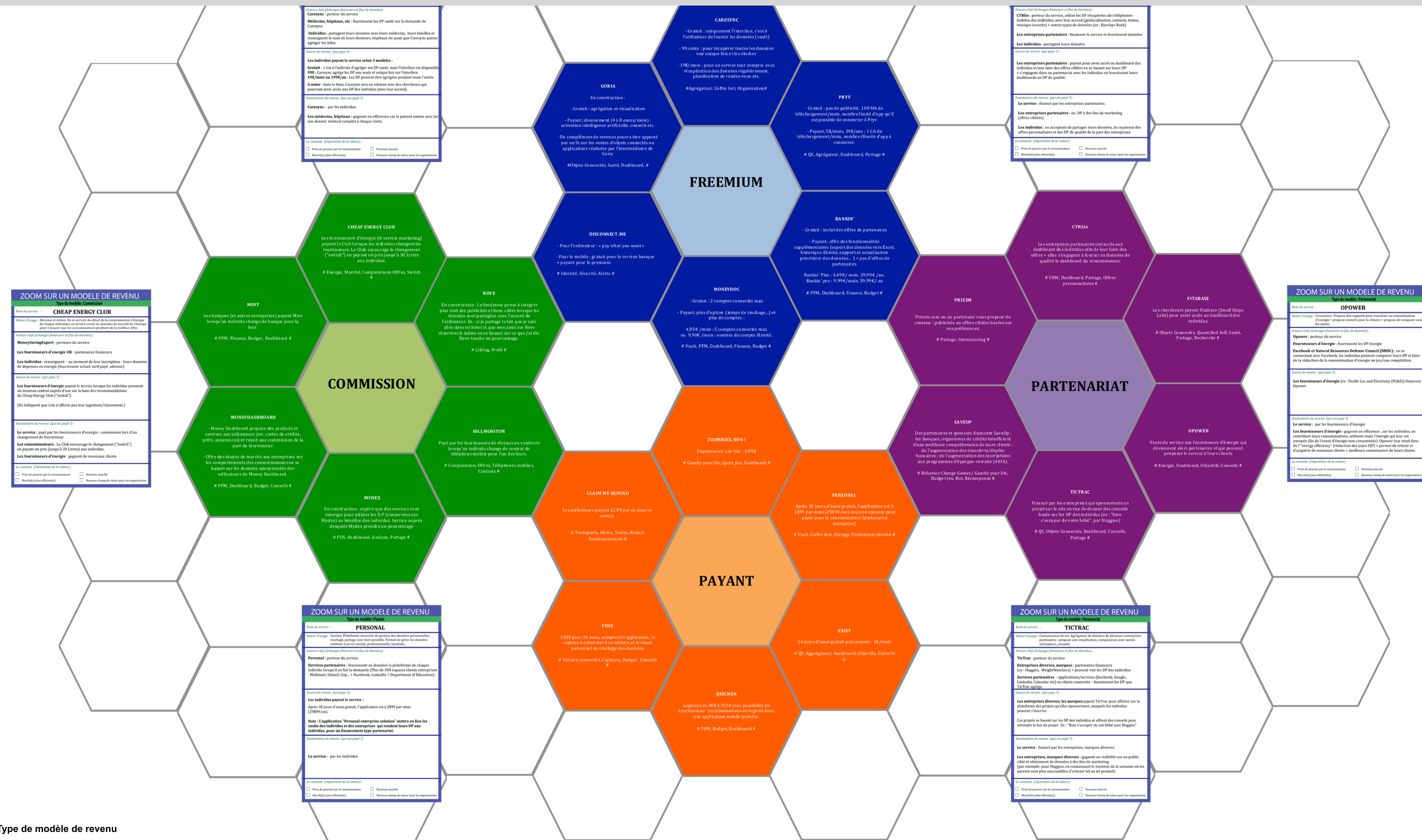
Le self data : devenir propriétaire de ses données

“la production, l’exploitation et le partage de données personnelles par les individus, sous leur contrôle et à leurs propres fins”.

- En réponse aux inquiétudes sur l’utilisation des données personnelles
- Une approche partenariale collectivités, acteurs industriels, citoyens
- Un cadre de confiance
- Un pilote : plateforme Mes Infos par la Fing.



Principaux modèles économiques



Légende :
 Partenariat
 Type de modèle de revenu



Principaux modèles économiques

- Payant : l'utilisateur du service paie pour pouvoir accéder ou utiliser les données ou pour utiliser les services associés aux données.
 - Commission : lors d'une vente suite à l'utilisation de données, prélèvement d'une commission qui rémunère le service de données. (modèle par défaut dans utilisation de datas internes par les plateformes numériques comme Google ou Amazon...).
 - Partenariat : un tiers paie le service et le fait savoir
 - Freemium : une partie du service est gratuit, des services additionnels sont payants.
-



Merci !



yam@bureaudespossibles.fr

BUREAU

DES

POSSIBLES