BIG DATA ET SCIENCE DE DONNÉES

INTRODUCTION

Master 1 Intelligence Artificielle

Université de M'sila, Département d'Informatique

Dr Mehenni Tahar

2020-2021

Définition de “donnée”

- **Définition de Donnée (Data)**

Data are raw symbols that represent the properties of objects and events and in this sense data has no meaning of itself, it simply exists (Russell L. Acko, 1989).

- **Exemple.** “John”, “Smith”, 30000

- **Information**: Donnée + sens (meaning).

(first name, “John”), (last name, “Smith”), (salary, 30000)

- **Définition de Dataset (ensemble de données)**

A dataset is a collection of data.

Données structurées

- Les données peuvent être identifiées selon leur structure.

- **Définition (Données Structurées)**

Une donnée structurée décrit une propiété (e.g., nom, adresse, Numéro de carte de crédit) d'une entité (e.g., client, produit) selon un modèle (ou template) fixé.

- **Exemples:**
- Données stockées dans des feuilles (e.g., Fichier Excel).
- Enregistrements stockés dans les tables d'une base de données relationnelle.
- Chaque propriété est distinguée facilement des autres.
- Elle correspond à une unité de la structure (e.g., colonne de la table).

Données non structurées

- **Définition (Données Non Structurées)**
- Une donnée non structurée décrit une entité qui ne poosède pas une structure à cause de ses propriétés qui ne peuvent pas être distinguées les unes des autres..

- Un texte est non structuré.
- Description des propriétés d’une entité noyée dans un contexte riche.
- Aucun accès direct à ces propriétés.

Données semi-structurées

- **Définition (Données Semi-structurées)**
- Une donnée semi-structurée possède une structure où les entitées et leurs propriétés peuvent être facilement distinguées, MAIS l'organisation de la structure n'est pas rigoureuse comme celle de la table de la base de données.

- **Exemples:** documents XML, JSON, HTML.
- **Exemple (document XML)**

```
<book id="bk101">
    <author>Gambardella, Matthew</author>
    <title>XML Developers Guide</title>
    <genre>Computer</genre>
    <price>44.95</price>
    <publish_date>2000-10-01</publish_date>
</book>
```

Niveaux de Structuration des Données

Niveau de structuration	Modèle de données	Exemples	Facilité de traitement
Structuré	Système de données relationnel objet/colonne	Base de données d'entreprise...	Facile (indexé)
Semi-structuré	XML, JSON, CSV, logs	API Google, API Twitter, web, logs...	Facile (non indexé)
Non structuré	Texte, image, vidéo	web, e-mails, documents...	Complexe

Source: Data Science: fondamantaux et études de cas, Biernat et Lutz, 2015

Définition du “Big Data”

- **Définition (Big Data)**
- Le terme Big Data se réfère à une accumulation de données très larges et très complexes pour être traitées par les outils classiques de gestion des bases de données.

- *“The term Big Data refers to an accumulation of data that is too large and complex for processing by traditional database management tools.”*

- Le terme Big Data se réfère:
  - Aux solutions hardware et software développées pour la gestion de données volumineuses.
  - A la branche de l’informatique qui étudie les solutions de gestion de données volumineuses.

Disciplines du Big Data

- **Informatique distribuée:**
  - Paradigme de programmation *Map-Reduce: «amener les codes de calcul sur les noeuds de données» «traitements large échelle » sur cluster Hadoop, …*
- **Informatique parralèle:**
  - Paradigmes du Calcul à Haute Performance (*HPC*): pour accélérer les algorithmes de « data analytics » ou de « machine learning », sur cluster de calcul intensif, sur GPU, sur superCalculateurs...
- **Bases de Données « NoSQL » : Not Only SQL**

-BdD plus simples mais à très large échelle

-Plusieurs types de BdD NoSQL: stockage distribué, interrogation sur les données.

Les 3 Vs du Big Data

- Big data is high-Volume, high-Velocity and high-Variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making (Gartner).

- Volume: La taille du dataset.
- Vélocité: La nécessité de traitement des données à leur arrivée.
- Variété: La nature hétérogène des données (structurées, semi-structurées, non structurées).

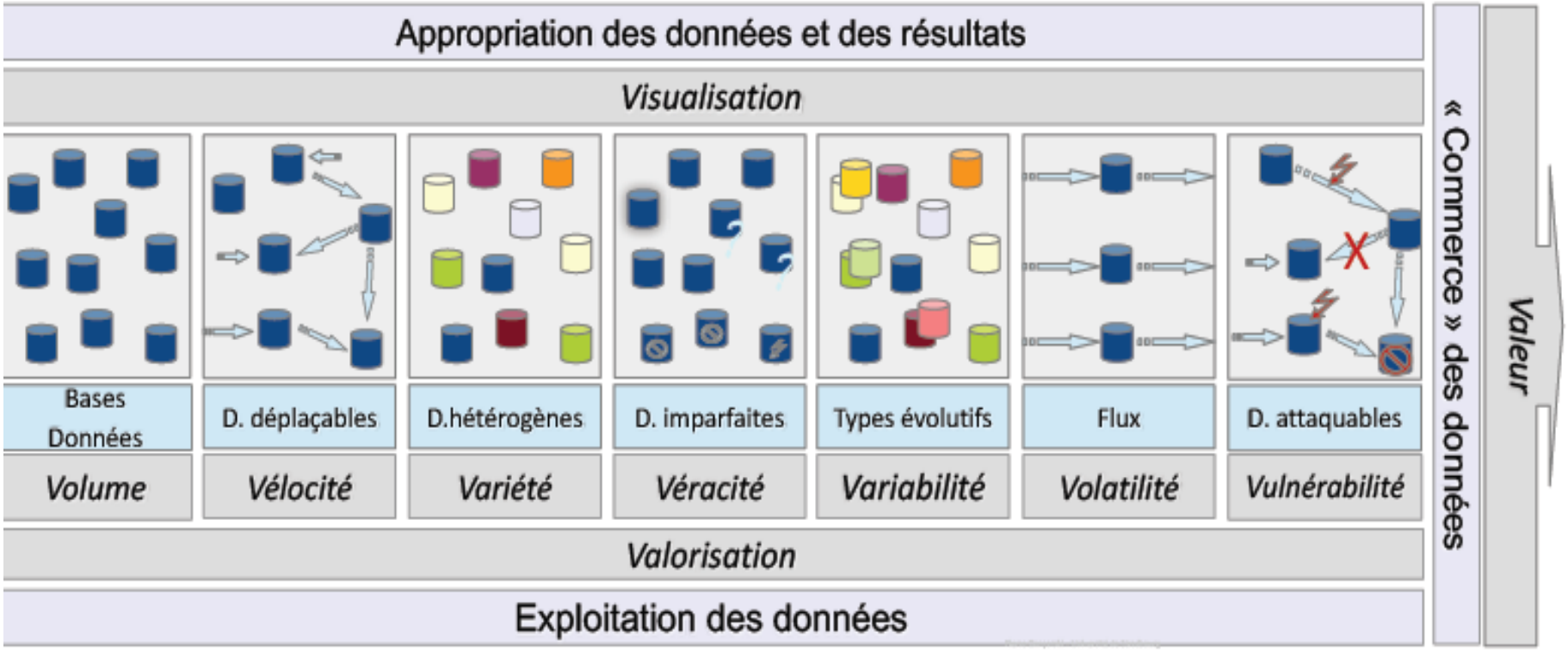
Les 4 Vs du Big Data

- Big Data consists of extensive datasets primarily in the characteristics of Volume, Variety, Velocity, and/or Variability that require a scalable architecture for efficient storage, manipulation, and analysis (NIST).
- **Différence entre variété et variabilité.**
- *Variété:* un boulanger qui vend dix types de pain.
- *Variabilité:* un boulanger qui vend un seul type de pain qui change de goût chaque jour.
- **Scalabilité:** la capacité d'une architecture du système de gérer la taille grandissante des données, sans aucune réduction de sa performance.

Exemple des 4 Vs du Big Data

- **Système d'analyse des Sentiments** qui traite les tweets afin de trouver l'ambiance générale du candidat politique.
- **Analyse du Langage:** sentiment positif/négatif/neutre?
- **Volume:** Millions de tweets.
- **Vélocité:** flot constant de données (7,500 tweets/second).
- **Variété:** Textes, images et liens pages Web.
- **Variabilité:** Le sens de chaque mot change selon le contexte.
  - Je suis profondément satisfait du candidat
  - Je suis profondément offencé par le candidat

Caractéristiques des Big Data



Source: Sciences de données (Big Data), iCube, 2017

Défis du Big Data

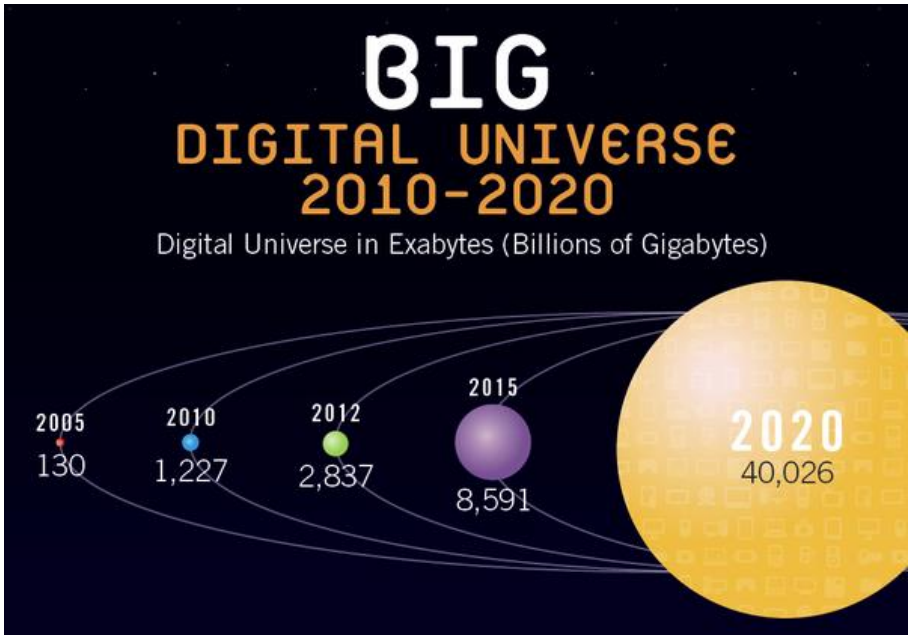
- Le Big Data ne peut pas être manipulé au niveau d'une seule machine.
- Il y a deux défis principaux du Big Data: traitement et stockage

Traitement:

- Parallélisation du calcul sur les machines.
- Bases de données parallèles.
- Frameworks de traitement distribué (e.g., Hadoop MapReduce/Spark)

Stockage:

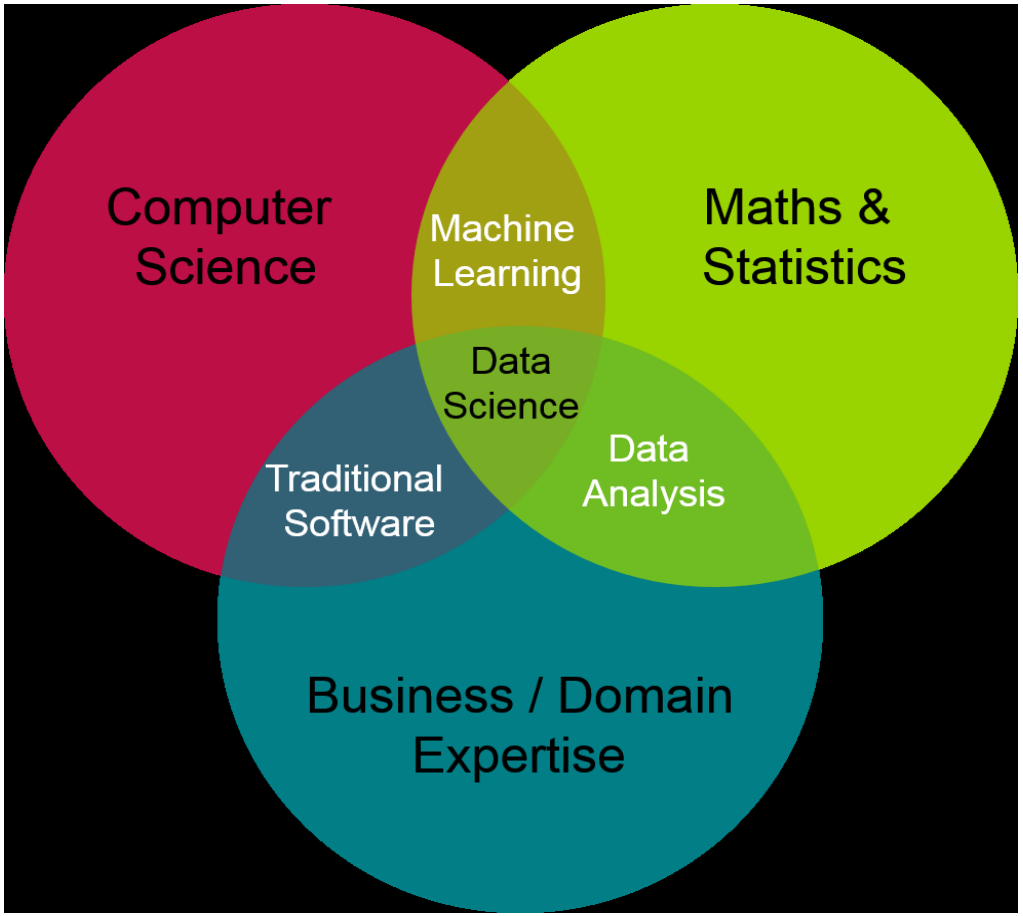
- Systèmes Distribués.
- Bases de données (relationnelles/NoSQL) distribuées.



Data Science (Science de Données)

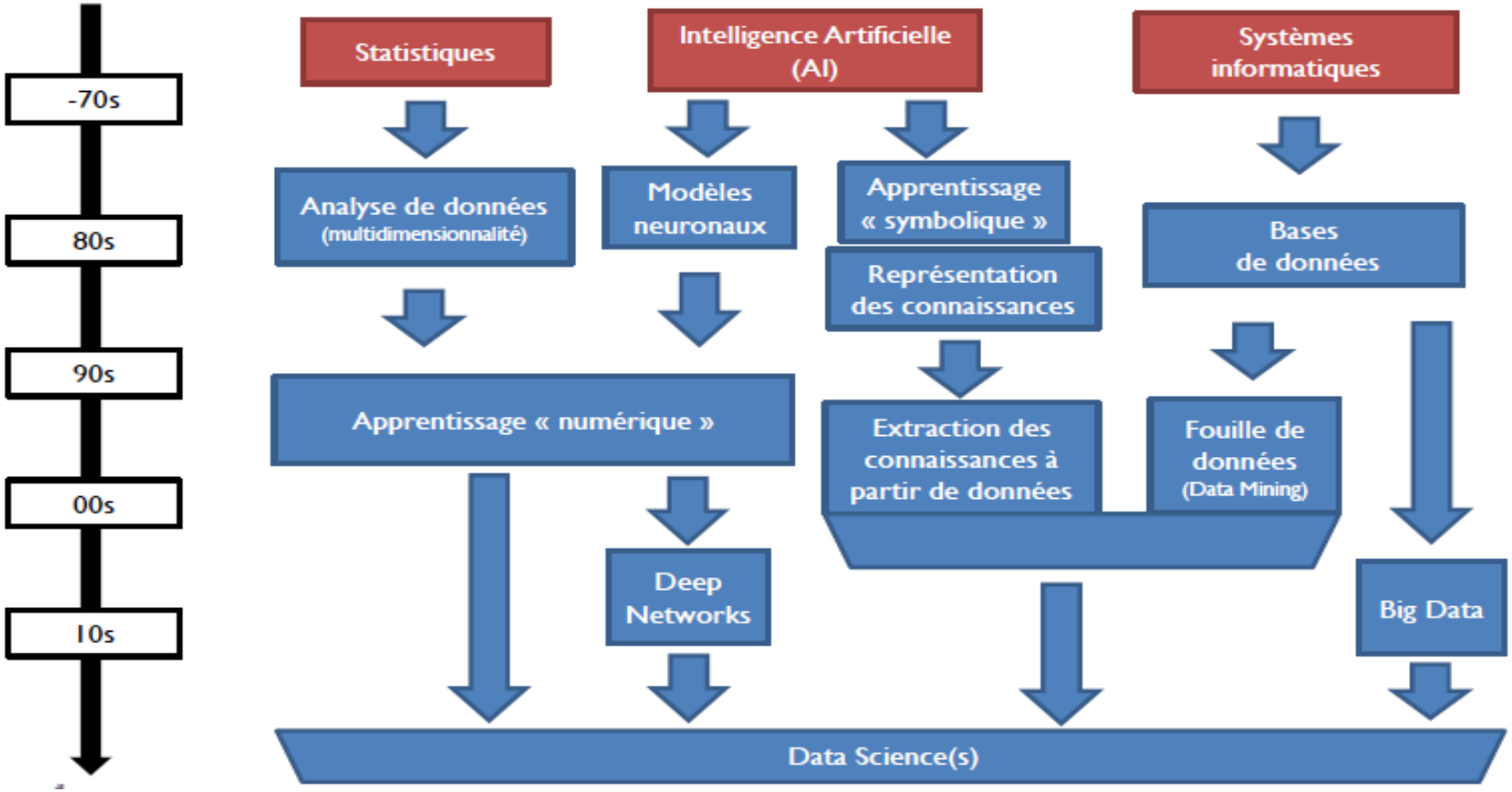
- *Data science* involves using methods to analyze massive amounts of data and extract the knowledge it contains. Cielen (2016).
- Une extension de l'analyse de données vers d'autres champs techniques, tels que l'informatique et l'expertise métier, qui sont nécessaires pour récolter, manipuler et exploiter les données disponibles dans nos environnements professionnels et personnels. Cleveland (2001).
- Un bon *data scientist* doit savoir naviguer entre ces différentes disciplines : statistique, algorithmie, informatique, sans a priori théorique.
- La différence entre un statisticien et un Data Scientist est que ce dernier possèle la capacité de travailler avec les Big Data, tout en maîtrisant le machine learning, l'informatique et l'algorithmique.

Diagramme de Venn de la *data science*



Source: https://www.freepng.fr/

Origines du Big Data et Data Science



Source: Mineure Data Science, Pinnerath

Objectifs du Big Data et Data Science

Détecter et optimiser :

– Croisement en temps réel d'un grand nombre de données diversifiées : meilleure connaissance des activités, de l'environnement, l'écosystème d'affaires.

– Aide au pilotage et à la prise de décision :

- Trouver la meilleure localisation pour des des éoliennes à partir de données météorologiques et géospatiales, des phases de la lune et de la marée, d'images satellites

  …

- Prévision de sinistralité, de détection de la fraude
- Mémoire territoriale

Objectifs du Big Data et Data Science

Tracer et cibler :

– Analyser la situation et le contexte de milliers de personnes en temps réel.

– Meilleure compréhension des réactions du marché pour la proposition de messages et d'offres personnalisés.

• Des systèmes GPS et télématiques permettant de collecter et monétiser une multitude de données sur les habitudes de conduite d'une population afin de proposer des primes d'assurance adaptées

• Usage-based-insurance : Connected xx (car, health, home)

- Analyse de graphes : fouille de réseaux sociaux, recherche de relations, …

Objectifs du Big Data et Data Science

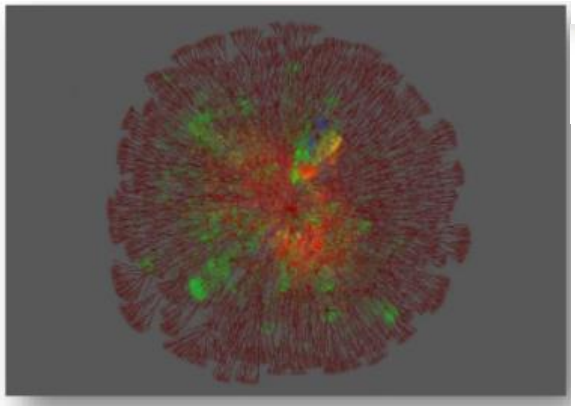
Prévoir et prédire :

- Analyse prédictive : projections pour identifier des nouvelles sources d'opportunités ou de menaces

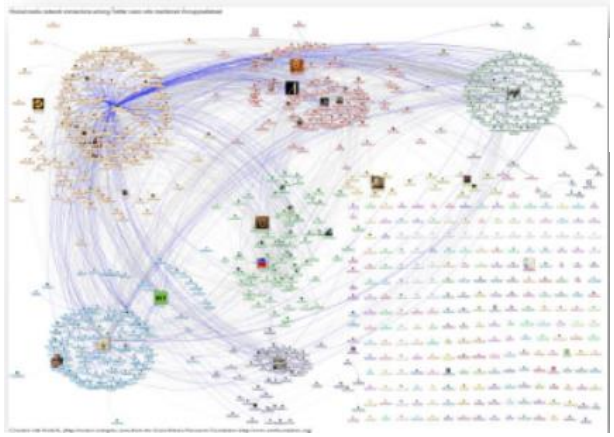
- Risques climatiques

- Prédiction de comportements

- Analyse de signaux de capteurs : maintenance prédictive, prévention des risques, …



Visualisation de 100000 pages de Wikipedia (source : Tulip/INRIA)



Utilisateurs de Twitter faisant référence à #occupiedwallstreet (source: Marc Smith/Wikimedia)

Domaines d’application de Big Data et Data Science

- **Entreprises commerciales:**
- Aperçus de leurs clients, processus, personnel, achèvement et produits.
- Offrir aux clients une meilleure expérience utilisateur, cross-sell, up-sell, personnalisation des offres.
- AdSense: collecte les données des internautes afin que les messages commerciaux pertinents puissent correspondre à la personne qui navigue sur Internet.
- MaxPoint (http://maxpoint.com/us): publicité personnaliée au temps réel.
- Professionnels des Gestion des Ressources Humaines (GRH) utilisent l'analyse des personnes et le text mining pour le recrutement, la surveillance de l'humeur des employés et l’étude des réseaux informels entre collègues.
- People analytics est le thème central du livre *Moneyball: The Art of Winning an Unfair Game*.

Domaines d’application de Big Data et Data Science

- **Institutions financières**: prédire les marchés boursiers, déterminer le risque de prêt et apprendre à attirer de nouveaux clients pour leurs services.
- **Organisations gouvernementales:** découvrir des informations précieuses sur la détection des fraudes, les activittés criminelles ou l'optimisation financement de projets.
- Un exemple bien connu: Edward Snowden, qui a divulgué comment l'Agence Américaine de Sécurité Nationale et du gouvernement britannique ont utilisé la science des données et le big data pour surveiller des millions d'individus. Ces organisations ont collecté 5 milliards d’enregistrements de données provenant d'applications célèbres telles que Google Maps, Angry Birds, e-mail et SMS, parmi de nombreuses autres sources de données.

Domaines d’application de Big Data et Data Science

- **Organisations non gouvernementales (ONG):** collecter des fonds et défendre leurs causes. Le Fonds mondial pour la nature (WWF), par exemple, emploie des data scientists pour augmenter l'efficacité des effors de collecte de fonds. DataKind en est un groupe de scientifiques de données qui consacre son temps au bénéfice de l'humanité.