Introduction à la Science de Données

Présenté par : Rodrique Kafando
Doctorant en Science de Données & IA
Email : kafando.rodrique@gmail.com

Décembre 2020

SOMMAIRE

Généralité

- Introduction
- Problématique
- Objectifs
- Quelles questions se poser lors d'une étude en DS?

Introduction

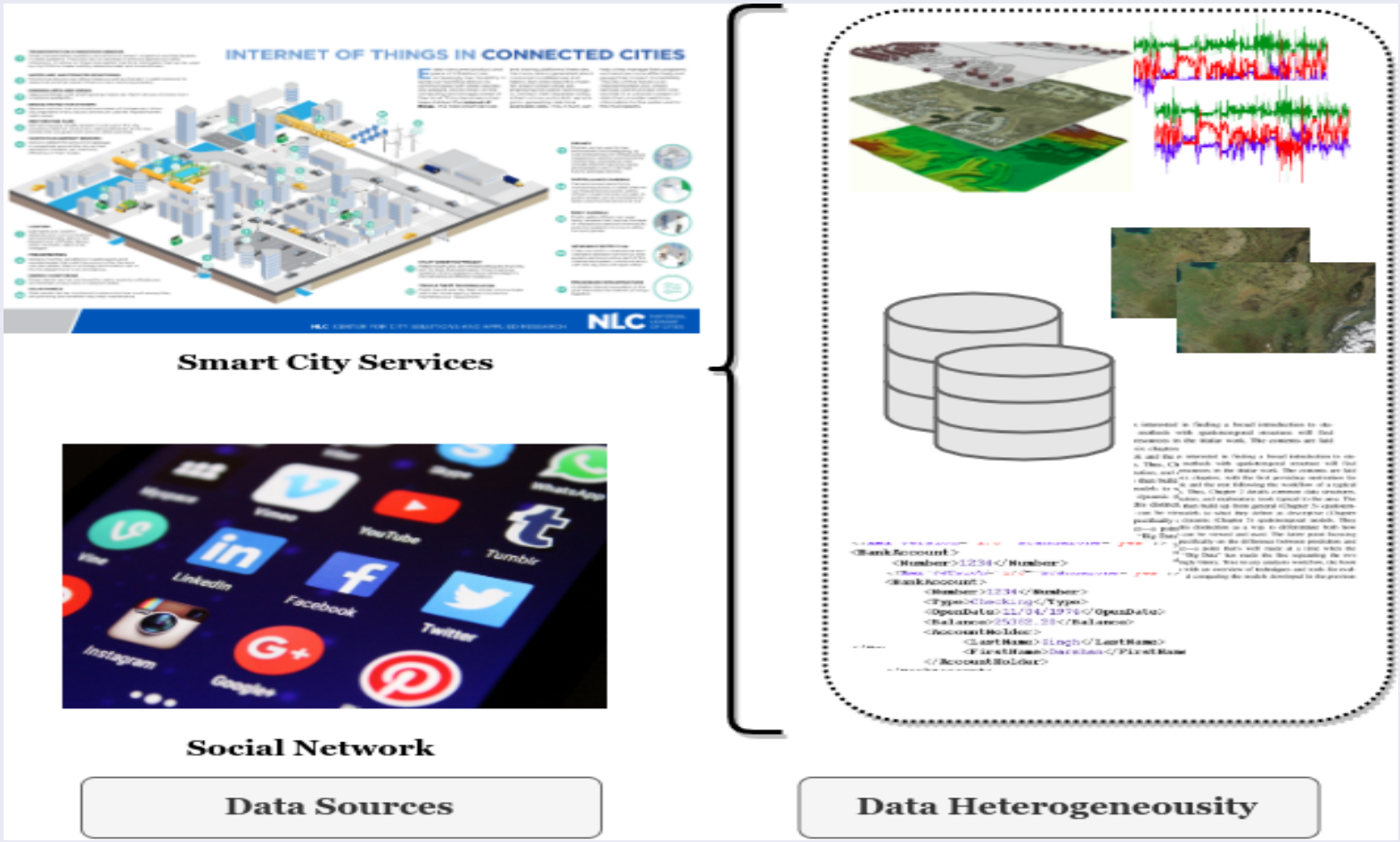

Figure 1: Sources des Données et Complexités liées à leur Analyse

Introduction

- une forte croissance des données avec l'évolution des nouvelles technologies (NTIC)
- chaque service (réponse à un besoin) produit un grand nombre de données
- les données sont de sources diverses, stockées différemment, avec une forte hétérogénéité dans leur structure
- d'où le concept de Données massives (Big Data)

Introduction

- Pourquoi il est si important d'exploiter les données ?
  - Traiter les masses de données pour extraire de l'information
    - essentielle et non évidente pour comprendre un comportement ou pour faciliter la prise de décision
    - comprendre le fonctionnement ou l'évolution d'une entité (diagnostic)
    - Ex: médecine (détecter l'émergence d'une nouvelle maladie,suivi de patients), marketing (connaître ses clients), etc.
  - internet => Banque de données
    - Google > 47 milliards de pages indexées en 2016
    - Intagram > 80 millions de photos par jour en 2015
    - Facebook > 30+ Petabytes d'info (2012)
    - E-mail spams 183 milliards par jour (Mars 2010)
    - TomTom intègre 5.5 milliard de nouvelles données GPS quotidiennement

Problématique

- Pb1 : Vue d'ensemble d'un processus en Science de Données

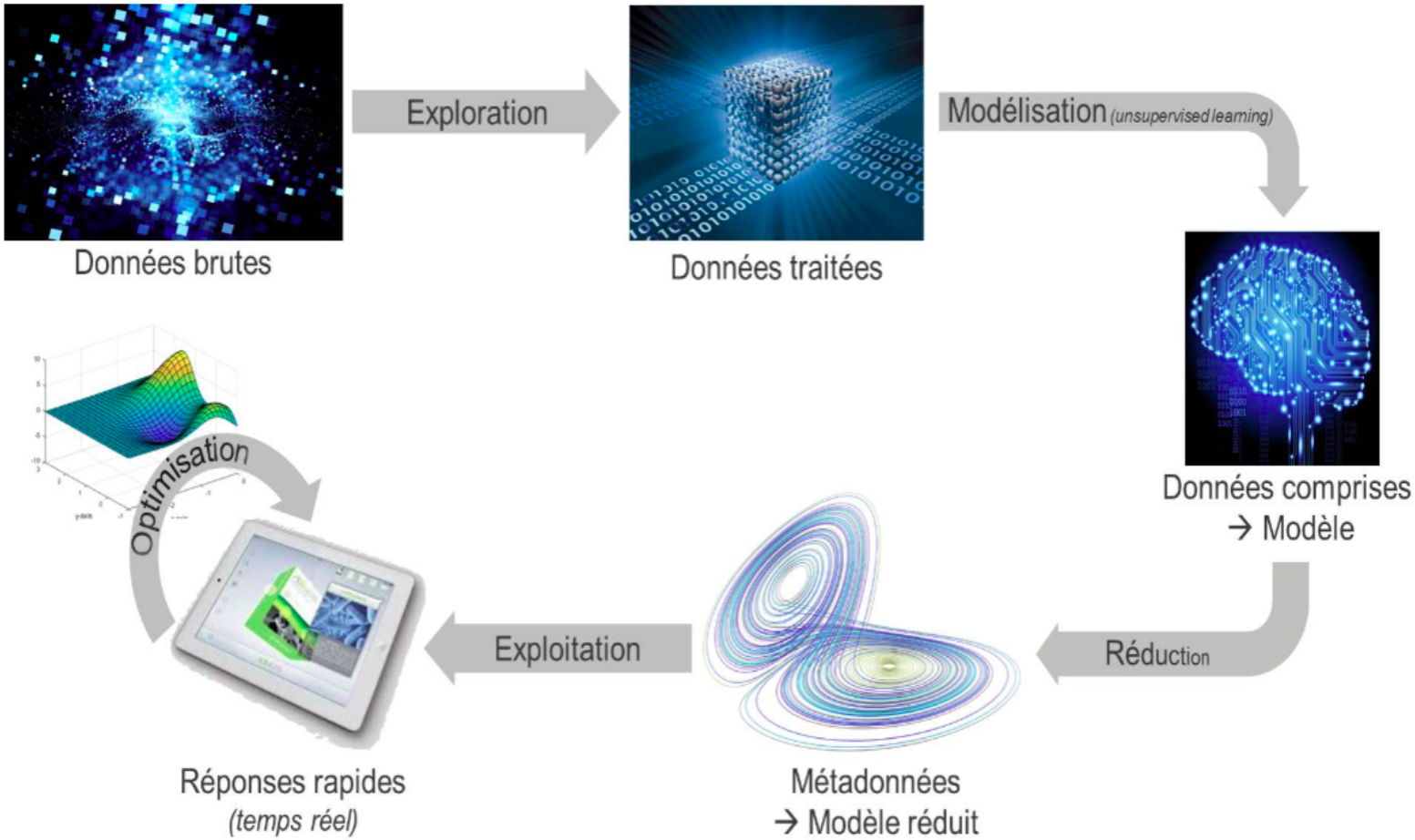

Figure 2: Vu d'ensemble d'un processus SD[?]

Problématique

- Pb1 : Vue d'ensemble d'un processus en Science de Données
  - On dispose initialement de **donnée brutes** sur lesquelles on dispose de plus ou moins de connaissances ;
  - une première étape consiste à explorer ces données, i.e. à essayer de mieux les appréhender, les connaître pour les transformer en **données traitées**, i.e. en données faisant sens pour nous ;
  - on peut alors entreprendre de modéliser ces données pour les transformer en données comprises ce qui, pour nous, signifie que ces données peuvent être représentées par un **modèle**.
  - Un modèle pouvant être volumineux, il peut être nécessaire de le réduire, i.e. de le décrire à l'aide d'un nombre réduit de variables ou **méta-données**
  - enfin, disposant d'un modèle réduit, i.e. capable de fournir une bonne approximation du modèle initial dans un temps court.

Problématique

- Pb2 : La science de données intégrée dans un SI

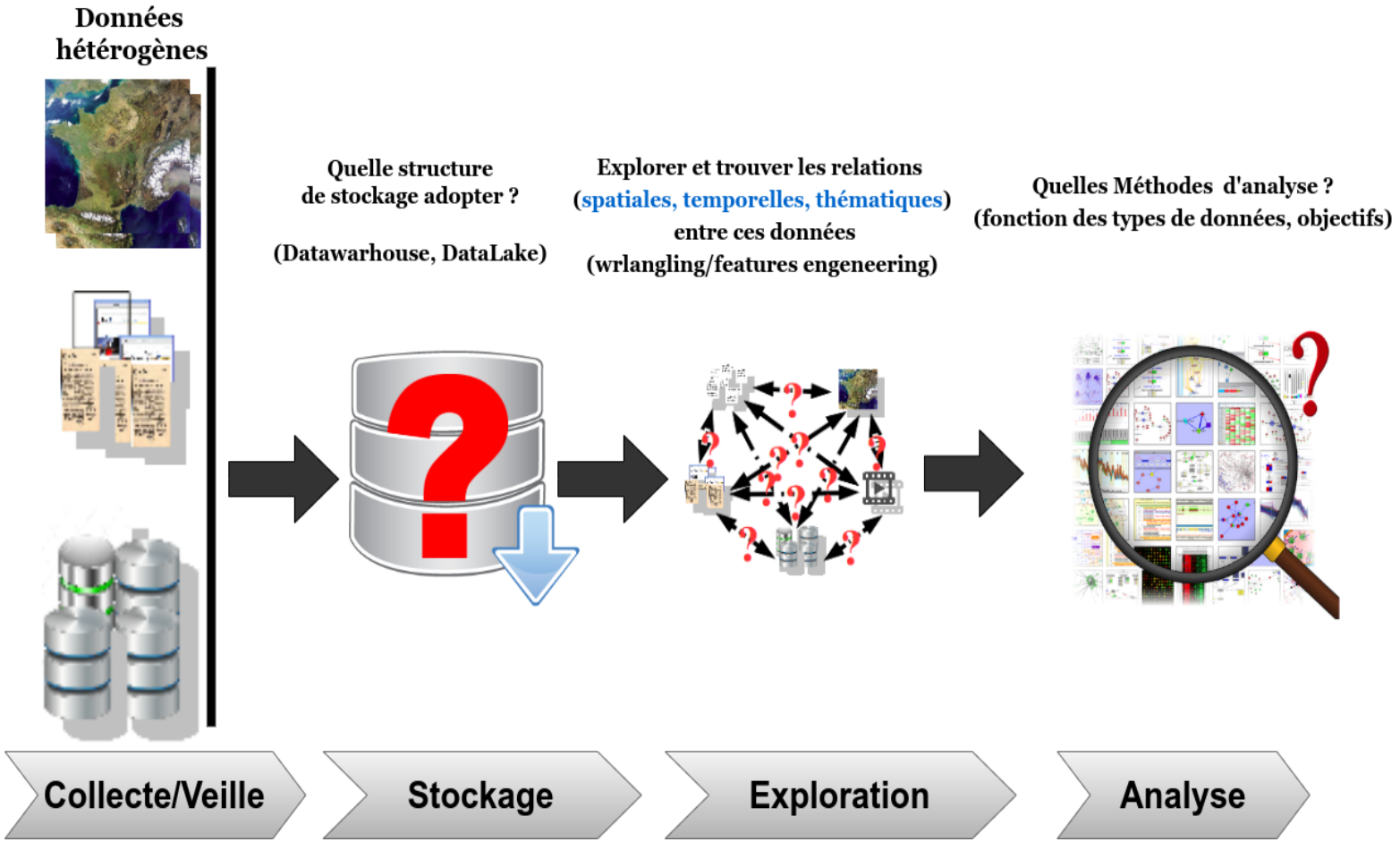

Figure 3: SD intégré dans un SI général

Généralité

Objectifs du module

- Comprendre les notions de bases sur la Science de données
- Apprendre à explorer les variables d'un ensemble de données
- Apprendre à traiter de gros volumes de données
- Manipuler les données avec le langage Python

Quelles questions se poser lors d'une étude en SD ?

- descriptive :
  - quels sont les principaux groupes ?
  - nettoyer les données (variables/individus) ?
- expérimentale :
  - établir un lien de causalité
  - émission de carbone et la température
- comparative/prédictive : y'a t'il une différence entre deux groupes ?
  - ce médicament est-il efficace ?

Questions éthiques

- l'éthique de la méthode
- l'éthique des données
  - confidentialité
  - mode d'obtention
- L'étique des usages : quelle est la question ?

Généralité

Quelles questions se poser lors d'une étude en SD ?

Les données servent à répondre à une question

- poser la question
- récupérer les données / les nettoyer (pré traitement)
- modéliser poser un modèle et des hypothèses (statistique)
- interpréter les données, (inférence statistique et vérif. des hypothèses)

Nature des variables statistiques

- Exemple de données
- Définitions sur les variables
- Transformations sur les variables

Nature des variables statistiques

Exemple de données

nom de la variable	*exemple*	aléatoire	discrète	Type
Matricule	242 335 AD	Non	-	-
nom	Peutu	Non	-	-
prénom	Stefen	Non	-	-
sexe	M	Oui	discrète	qualitative
age	38	Oui	continue	quantitative
département de résidence	76	Oui	discrète	qualitative
PCS	Employé	Oui	discrète	qualitative
total des avoirs	11 240	Oui	continue	quantitative
max des entrées	1 570	Oui	continue	quantitative
taux d'endettement	31 %	Oui	continue	quantitative
nombre de visites	2	Oui	discrète	quantitative

Figure 4: Exemple de données - étude de variables

Trois questions fondamentales

- variable qualitative ou quantitative,
- discrète ou continue,
- variable aléatoire ou non (déterministe)

* Définitions sur les variables

Domaine d'une variable

Le **domaine d'une variable** est l'ensemble des valeurs que cette variable peut prendre.

$$\Omega_{sexe} = \{M, F\} \tag{1}$$

Variable discrète

une **variable est discrète** si le cardinal de son domaine est dénombrable
ex : *sexe, département, PCS, nombre de visites...*

Variable quantitative

une **variable discrète est quantitative** si l'ensemble de ses modalités est comparable[a]. Toutes les variables continues sont quantitatives.
ex : *age, nombre de visites...*

[a] $\forall$ x, y deux modalités quelconques, soit x $<$ y soit x $\geqslant$ y

Nature des variables statistiques

Transformation des variables

- continue -> continue (log)
- continue -> discrète (quantification)
- discrète -> continue (calcul d'une moyenne sur un age par exemple)
- continue -> discrète (ordre)

Observation	2	-5,5	1	3,4
Rang	3	1	2	4

Table 1: Exemple de transformation (c -> d)

Séance de TD

- Environnement de travail
- cas d'étude - exploration de données

MERCI POUR VOTRE ATTENTION!