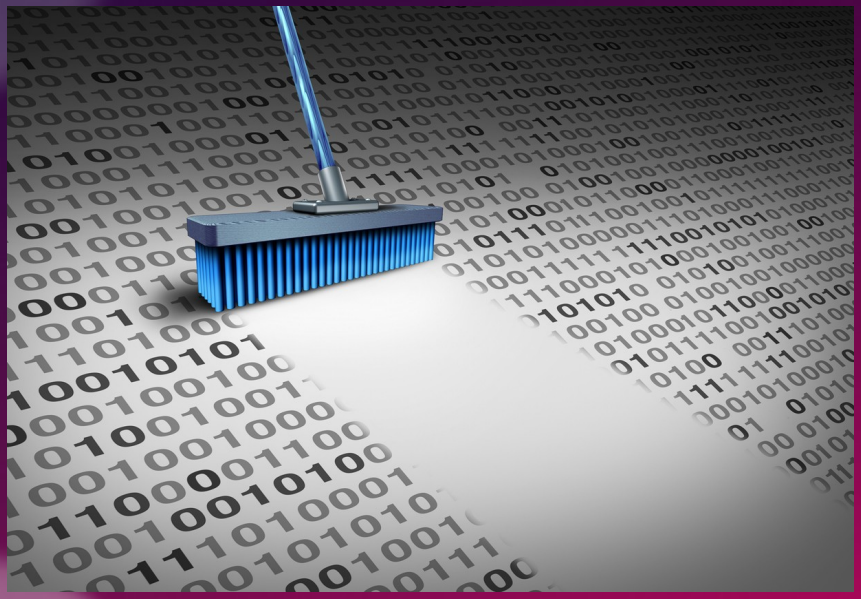
Nettoyer ses données ...

… avec un logiciel de Data-science : *Dataiku*



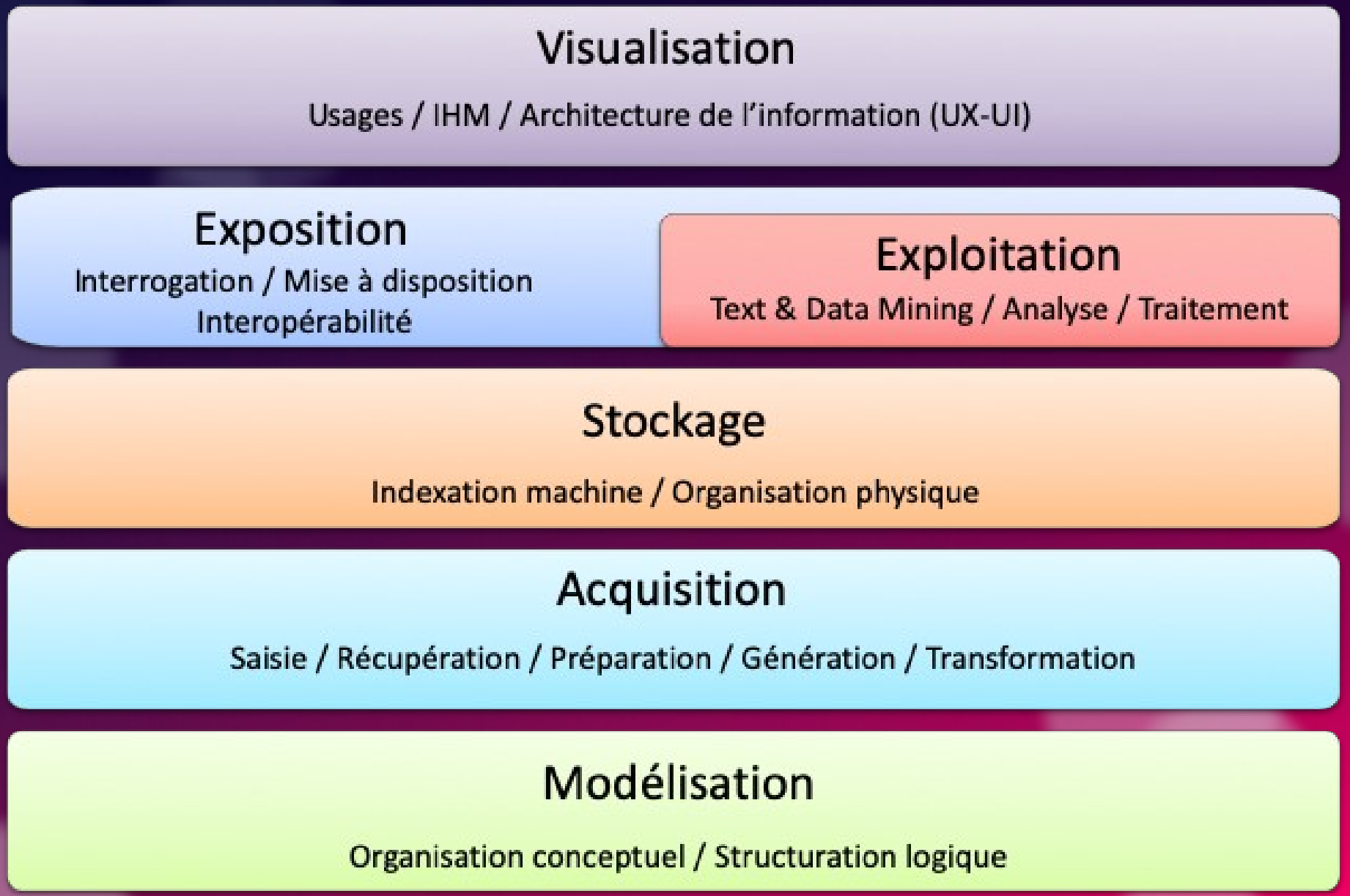
Data — données : késako ?

Document : ensemble logique et fini d'informations, dont les limites sont définies par des caractères physiques

Donnée : le fait que vous soyez en master est une donnée

Métadonnée : le fait que ce master dépende de l'UVSQ est une métadonnée, c'est en effet, une donnée sur la donnée

Cycle de vie de la donnée :



De la modélisation...

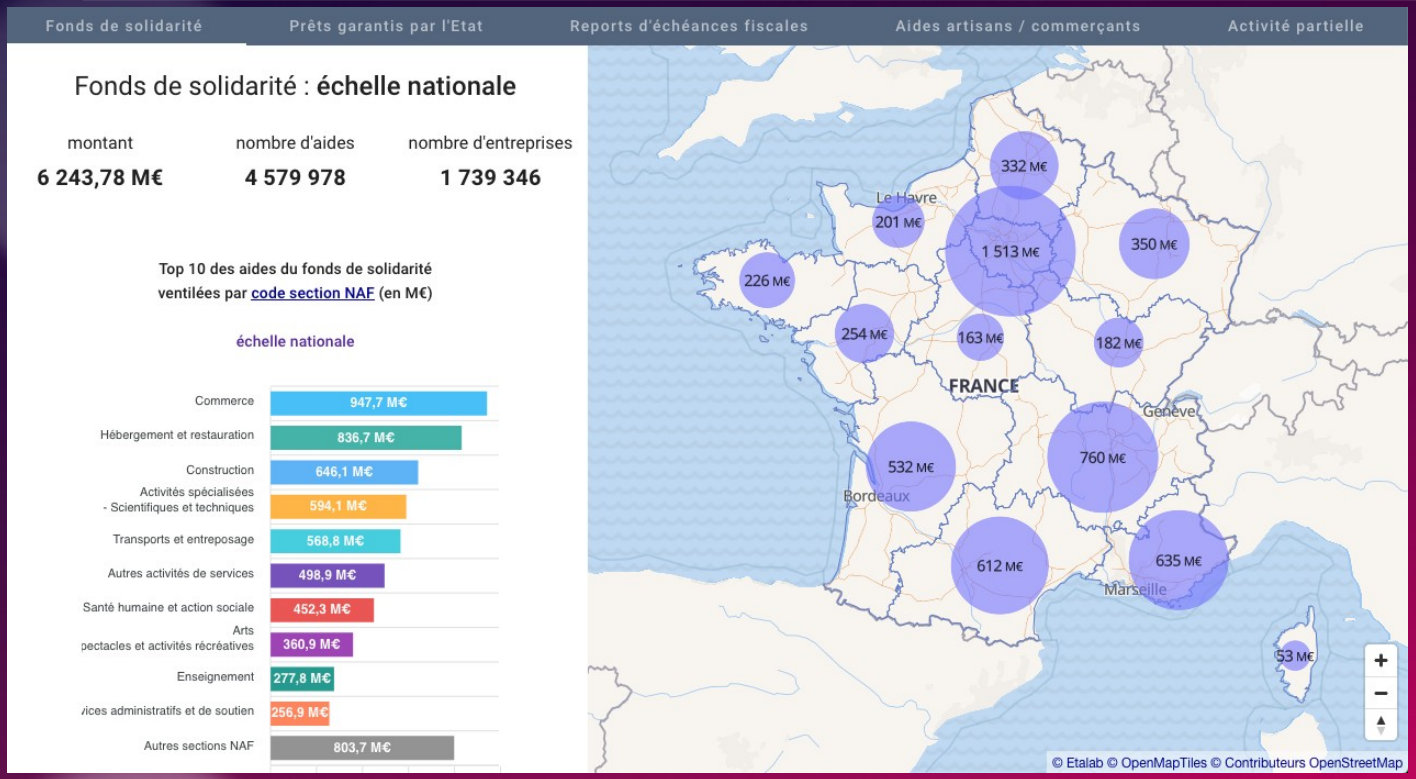
- Modéliser c’est faire le tour d’un « monde »

Exercice par groupe autour d’une donnée … la modélisation en ...

… à la visualisation

- Visualisation de données ou *dataviz'* : désigne les techniques permettant de présenter des données sous forme visuelle afin d'en faciliter la compréhension et/ou l'analyse. (*Source : https://www.1min30.com/dictionnaire-du-web/dataviz*)

A partir de données exportées sur le site data.gouv.fr : https://www.data.gouv.fr/fr/datasets/donnees-relatives-au-fonds-de-solidarite-mis-en-place-dans-le-cadre-de-lepidemie-de-covid-19/

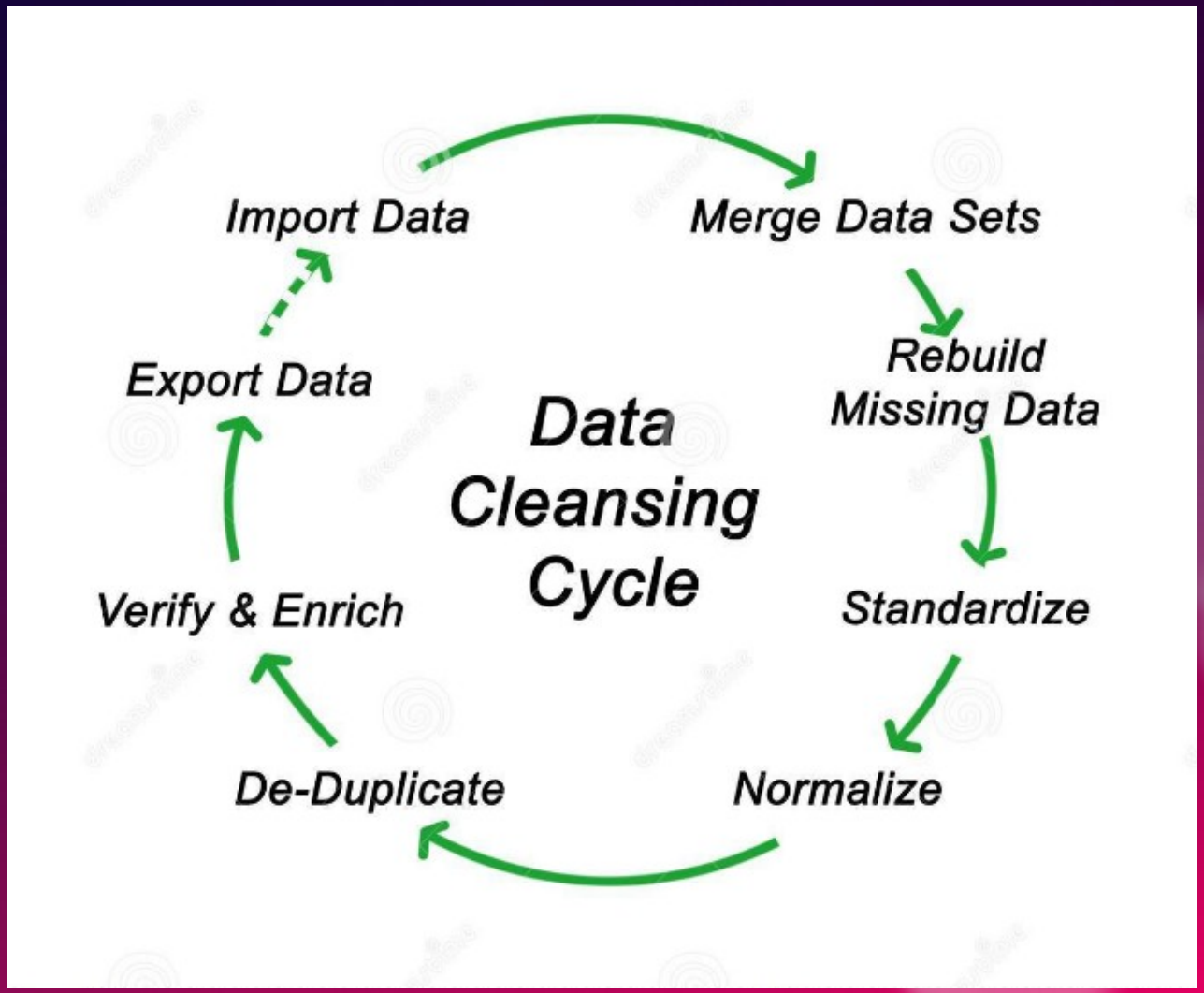


Etalab à manipuler les données pour produire une visualisation : https://aides-entreprises.data.gouv.fr/

Outil de production gratuit de dataviz' : https://public.tableau.com/en-us/s

Nettoyer ses données #1

Pourquoi ?
- Identification
- Correction



Nettoyer ses données #2

- L’identification de données ?

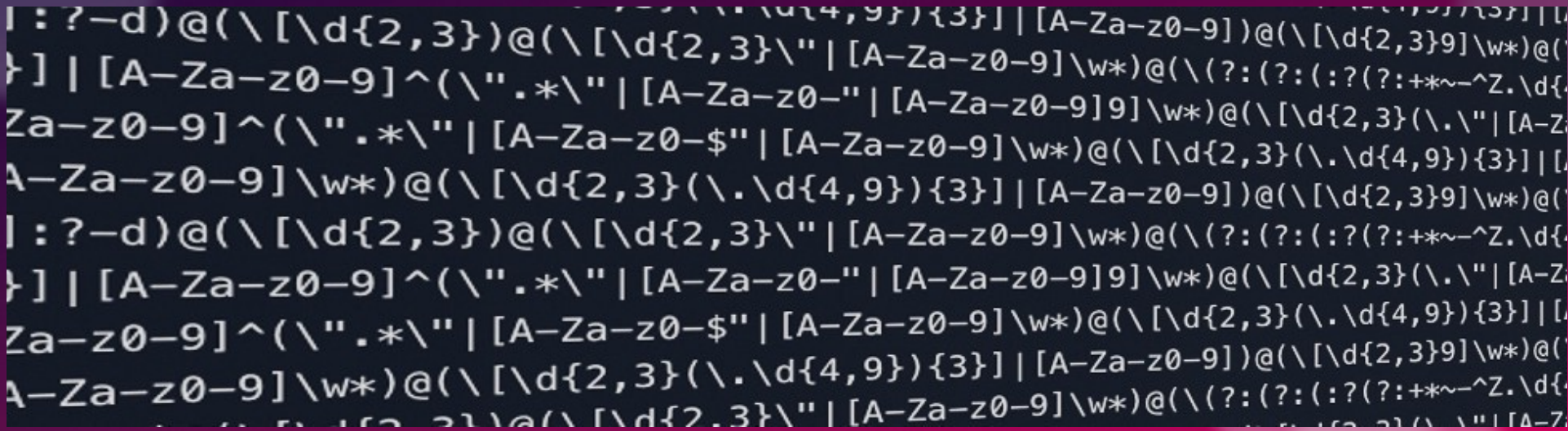
  Permet de connaître le niveau de scientificité des données, c’est à dire savoir si elles sont qualitatives ou non. Si on peut s’en servir ou non pour la recherche.

- Correction de données ?

  Permet de structurer intelligemment, sémantiquement une ou plusieurs données afin de rendre son traitement automatique possible

Passage à la pratique :

- Un monde de possibilités (quasi) infinies grâce aux **RegEx**, ou Regular Expression, ou encore Expressions Régulières ;-)



Memento RegEx #1

- Délimitateurs : un caractère assez rare de préférence comme #, %, / etc.
- Les métas-caractères :
  - ^ : marque un début de *string*
  - $: marque une fin de *string*
  - | : connecteur logique « ou »
  - . : tous les caractères sauf les retours chariots
  - \ : caractère d'échappement (/\ ?/ = « ? » en caractère normal)

Mémento RegEx #2

- Quantificateurs :
  - ? : 0 ou 1 fois
  - + : 1 fois ou plus
  - * : 0, 1 ou plus
  - ( ) : répétition sur plusieurs lignes
  - { } : précise le nombre de répétitions
- Les classes ou intervalles :
  - [ ] : classe de caractères
  - [-] : intervalle de classe
  - [^] : classe à exclure

Mémento RegEx #3

- Classes abrégées :
  - \d : un chiffre = [0-9]
  - \D : ce qui n'est pas un chiffre [^0-9]
  - \w : caractère alphanumérique ou un tiret de soulignement
  - \W : ce qui n'est pas un caractère alphanumérique ou un tiret de soulignement
  - \t : une tabulation
  - \n : une nouvelle ligne
  - \r : un retour chariot
  - \s : un espace blanc
  - \S : ce qui n'est pas un espace blanc
  - . : n'importe quel caractère (autorise tout)

Mémento RegEx #4

- Classes nommées :
  - [:alnum:] : tout ce qui est alphanumérique
  - [:alpha:] : les caractères alphabétiques
  - [:blank:] : les caractères blancs
  - [:ctrl:] : caractères de contrôle (1er du code ASCII)
  - [:digit:] : chiffres
  - [:graph:] : caractères d'imprimerie
  - [:print:] : les caractères imprimables
  - [:punct:] : la ponctuation
  - [:space:] : caractères d'espacement
  - [:upper:] : les majuscules
  - [:xdigit:] : l'hexadécimal

Dataiku :

- A retenir :
  - localhost:
  - Permet de nettoyer ses données
  - IA
  - Recettes
  - Flow
  - Supporte différents langages
  - Permet de réaliser des *mashup*