# Projet inter-promo 2018


SID

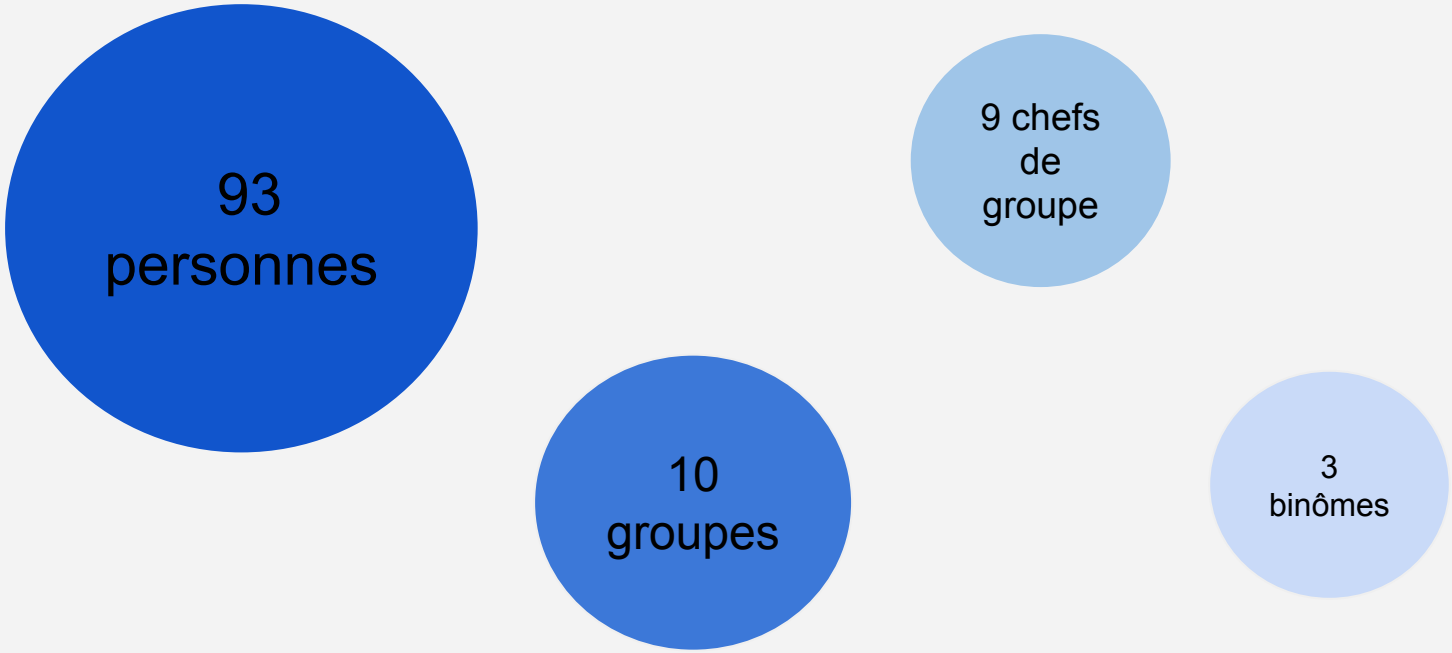
UNIVERSITÉ TOULOUSE III PAUL SABATIER Université de Toulouse

Janvier 2018

# Sommaire

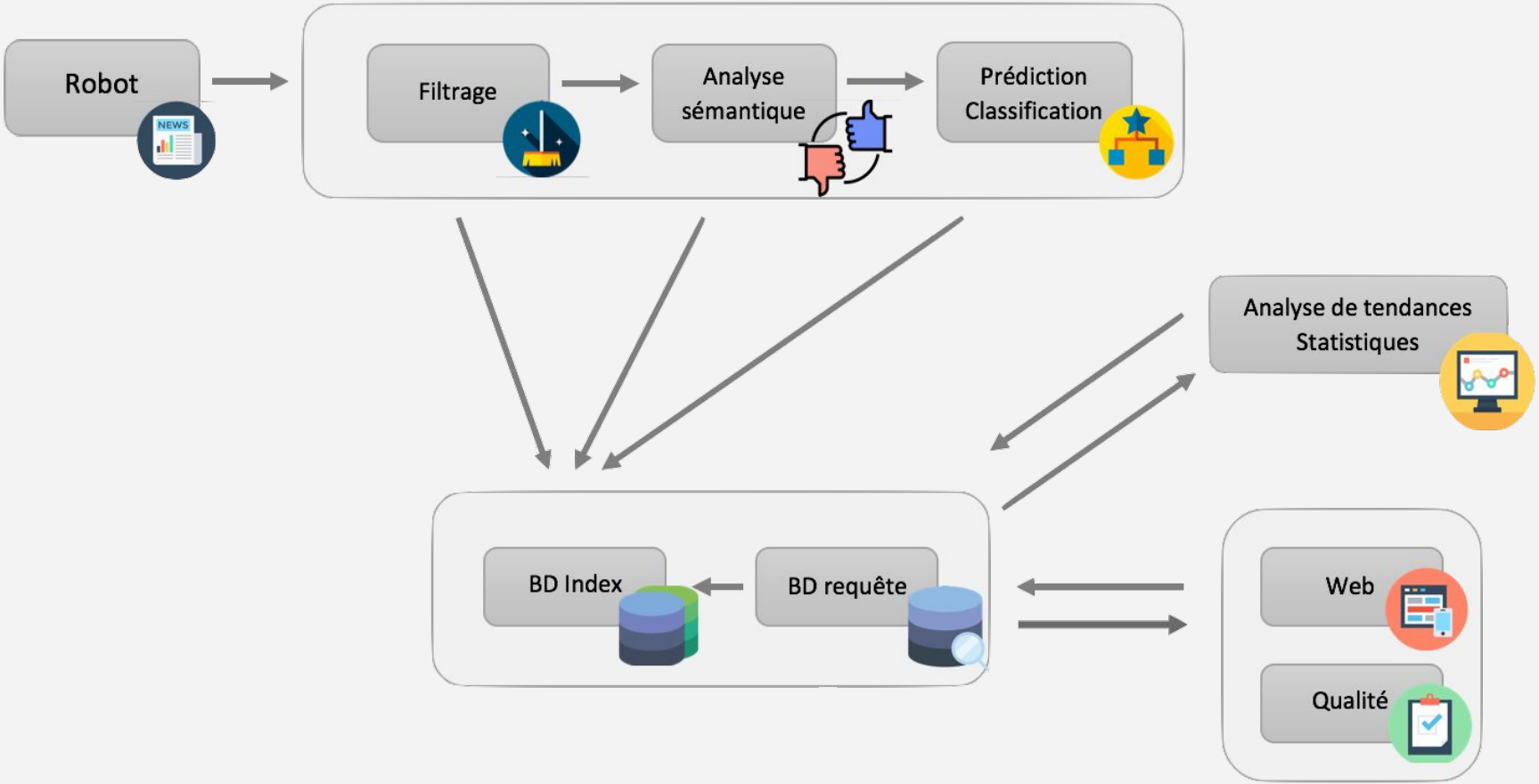
- Présentation du projet
- Les différents groupes
- Bilan global
- Démonstration


# Présentation du projet




*"WatchNews le site qui analyse quotidiennement la presse en ligne"*

# Transfert des données



# Groupe 2 : BD index

# Groupe 2 : BD index

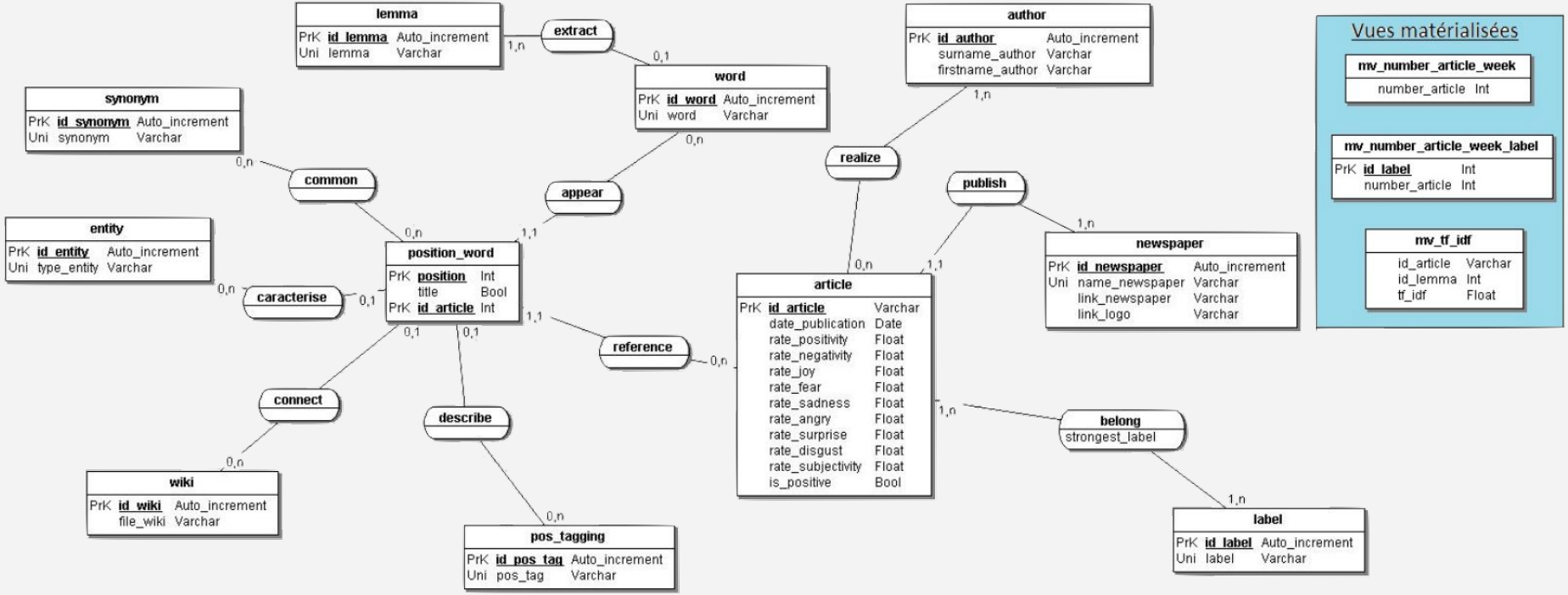
**Objectifs**

- Créer la base (tables, contraintes ...)
- Insérer les données mises à disposition en début de projet
- Gérer l'évolution du schéma
- Implémenter une API REST pour permettre l'insertion de données envoyées par les groupes 5, 6 et 7

**Résultats**

- Scripts python pour récupérer les fichiers json sur le serveur
- BD fonctionnelle contenant des données

# Groupe 2 : BD index



# Groupe 2 : BD index

**Technologies**

MySQL

Laravel

Flask
web development,
one drop at a time

**Difficultés :**

- Mise en place de l’ API
- Changement de spécifications tardifs

# Groupe 3 : BD requêtes

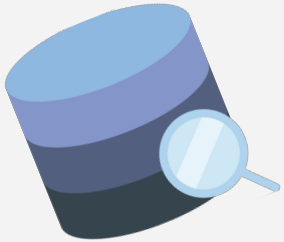
# Groupe 3 : BD requêtes

**Objectifs**

- Ecrire les requêtes/procédures stockées répondant aux Use Cases.
- Implémenter une API REST pour permettre de répondre aux besoins des groupes 8 et 9

**Résultats**

- API fonctionnelle avec le groupe 8 et 9
- Réponses aux besoins des groupe 8 et 9



# Groupe 3 : BD requêtes

**Technologies**

MySQL

Flask
web development,
one drop at a time

python™

**Difficultés**

- Mise en place de l'API
- Mise en place des formats d'échanges
- Modifications récurrentes du MCD
- Manque de données

# Groupe 4 : Robot

# Groupe 4 : Robot

**Objectifs**

- Récupérer le contenu des articles
- Récupérer des métadonnées (l'auteur, la date...)
- Stocker toutes les informations (JSON)


**Résultats**

✓ 15 sources récupérées
✓ Pas de doublon pour les articles


# Groupe 4 : Robot

**Technologies utilisées**

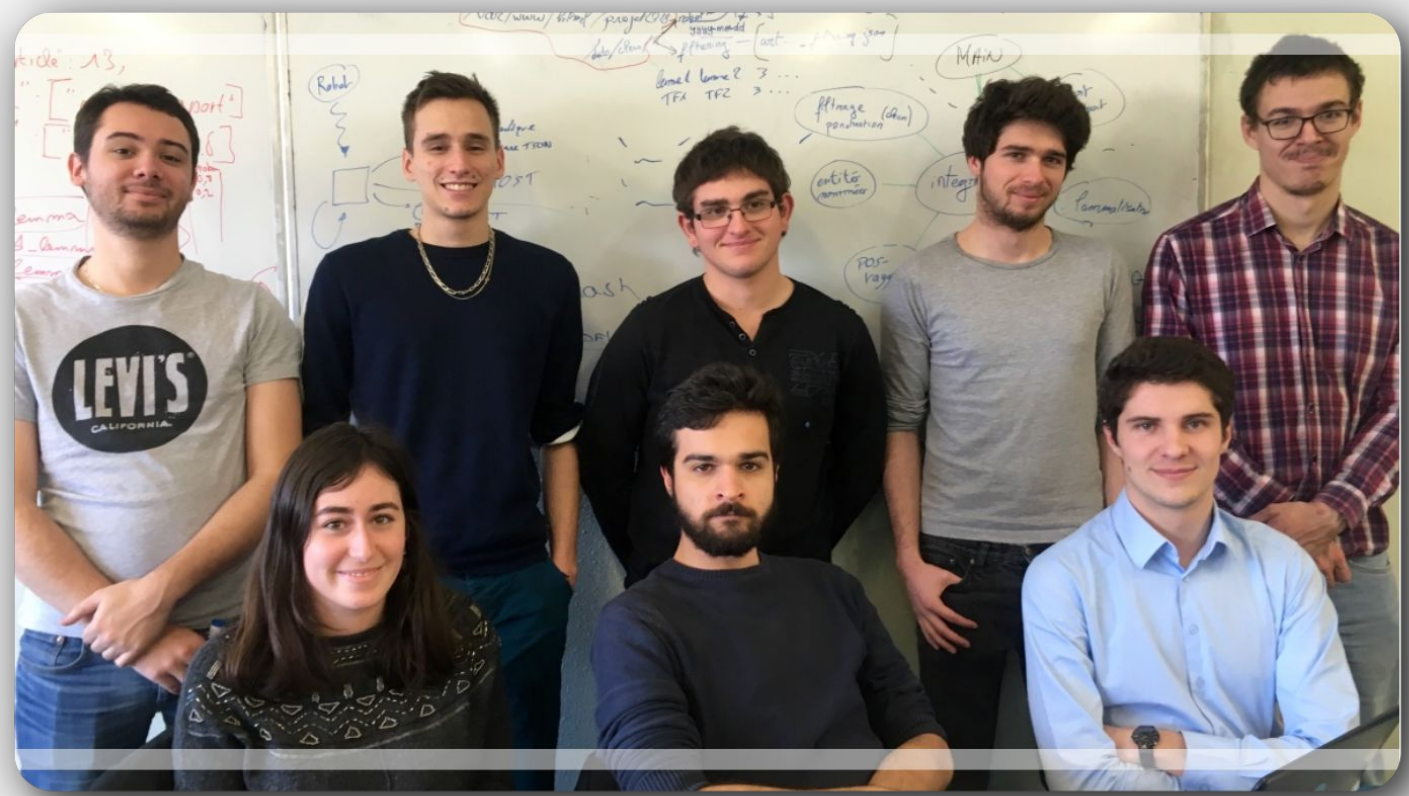
- MD5
- Unidecode
- BeautifulSoup

**Méthodes utilisées**

- Hashage
- Expressions régulières / nettoyage
- Crawler
- Parsage HTML/XML

**Difficultés :** Encodage de certains articles, limite de crawlage

# Groupe 5 : Filtrage



# Groupe 5 : Filtrage

**Objectifs**

- Récupérer les JSON du Robot
- Extraction d'informations (entité nommée, pos-tagging)
- Nettoyer et traiter les articles (NLP)
- Exporter les informations dans la BD
- Stocker toutes les informations (JSON)

**Résultats**

✓ Fichiers traités et envoyés au groupe sémantique
✓ Calcul TF*IDF

# Groupe 5 : Filtrage

**Technologies utilisées**

- Re
- NLTK
- Spacy

**Méthodes utilisées**

- Expressions régulières / nettoyage
- Tokenisation
- Entité nommée
- Pos-tagging
- Suppression des stopwords
- Lemmatisation
- TF*IDF

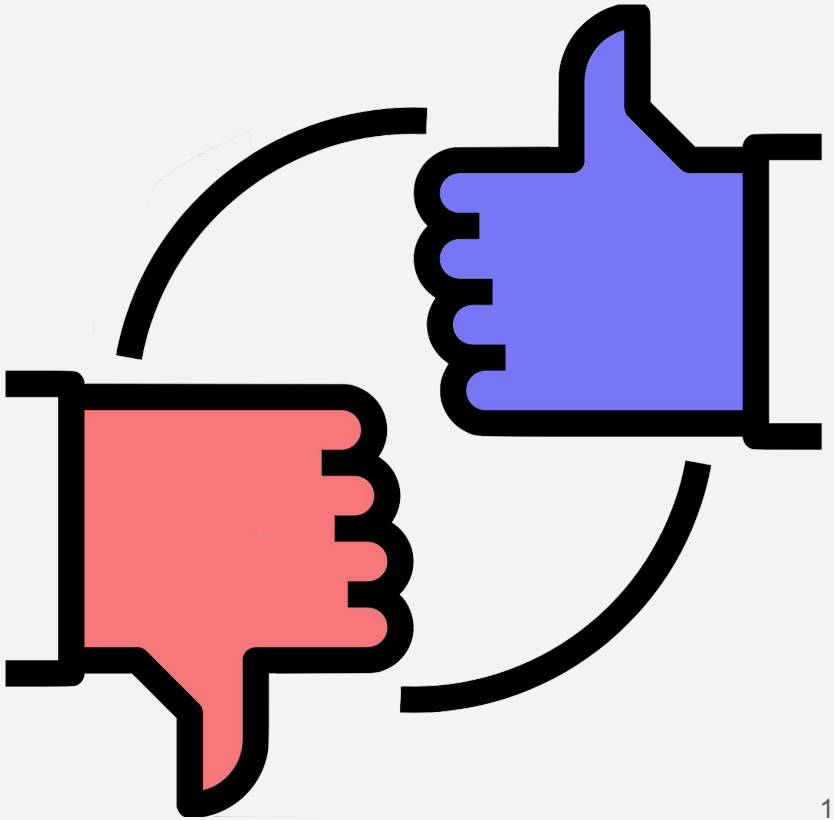
**Difficultés :** Envoi des informations dans la BD

# Groupe 6 : Analyse sémantique

# Groupe 6 : Analyse sémantique

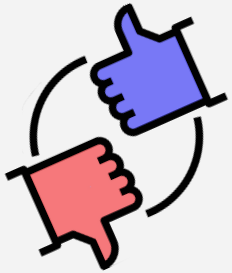
**Objectifs**

- Récupérer les fichiers JSON du groupe Filtrage
- Associer aux mots et aux documents une polarité
- Relier les entités nommées à des pages Wikipedia
- Envoi des données à la BD et au groupe Prédiction



**Résultats**

✓ 9 ratios par document
✓ Bonnes pages Wikipedia
✓ Synonymes
✓ Gravité des mots



# Groupe 6 : Analyse sémantique

**Technologies utilisées**

- French Expanded Emotion Lexicon
- TextBlob
- Package Wikipedia

**Méthodes utilisées**

- Sentiment
- Polarité / Subjectivité
- Page Wikipedia

**Difficultés :** Envoi des données à la BD et temps d'exécution

# Groupe 7 : Prédiction & Classification

# Groupe 7 : Prédiction & Classification

**Objectifs**

- Prédiction de la catégorie thématique de l'article
- Apprentissage supervisée : méthodes de machine learning et deep learning
- Apprentissage non supervisée : prédiction de la catégorie sémantique

**Résultats**

✓ 80% de bonne prédiction en mono-label
✓ 90% de bonne prédiction en multi-label
✓ 5 clusters sémantique
✓ Rapidité du temps d'exécution

# Groupe 7 : Prédiction & Classification

**Démarche**

- Mise en place de la base d'apprentissage : recodage des labels selon les catégories de Google News
- Modèles :
  - Xgboost : package LGBM
  - Réseau de neurones
  - CAH, K-means
- Amélioration du modèle :
  - Bagging et Stacking

**Difficultés rencontrées :** Retard des données pour la partie non supervisée. Répartition des tâches.

# Groupe 8 : Analyse de tendance & Statistiques

# Groupe 8 : Analyse de tendance & Statistiques

**Objectifs**

- Analyse de tendance
- Statistiques : statiques et dynamiques

**Résultats**

- ✓ API flask
- ✓ Des fonctions génériques créées
- ✓ Trois fonctionnalités seulement utilisées par le site
- ✓ Plusieurs fonctionnelles sur le serveur de test

# Groupe 8 : Analyse de tendance & Statistiques

**Méthodes/Technologies**

- Python
- Tendance →
- Moyenne de moyenne TF*IDF
- Polarité
- Modèles basiques

→
- Moyenne mobile
- Tests statistiques
- Décomposition STL
- Droite de régression

**Difficultés**

Coopération avec le groupe Web et BD requêtes.
Manque de données → impossibilité de faire des tests ou du machine learning

# Groupe 9 : Web

# Groupe 9 : Web

**Objectifs**

- Site Web responsive
- Afficher des statistiques/informations pertinentes
- Possibilité de recherches (selon thème ou mot clé)

**Résultats**

✓ Site web présentant différents graphiques et informations

✓ 3 pages :

→ Accueil
→ Thème
→ Recherche

# Groupe 9 : Web

**Méthodes**

- Structure du site et responsive (html, bootstrap, css)
- Interactivité/dynamisme du site (java script, JQuery, ajax)
- Graphiques (librairies javascript : googlechart, jqcloud, justgauge)

**Difficultés**

Communication avec les autres groupes

# Groupe 10 : Qualité & Communication

# Groupe 10 : Qualité & Communication

**Objectifs**

- Visuel du site web
- Communication
- Chartes
- Tests
- Page Html :  description du projet

**Résultats**

- ✓ De nombreuses fonctions testées
- ✓ Un site avec un beau visuel
- ✓ Chartes
- ✓ Site web
- ✓ Read me

# Groupe 10 : Qualité & Communication

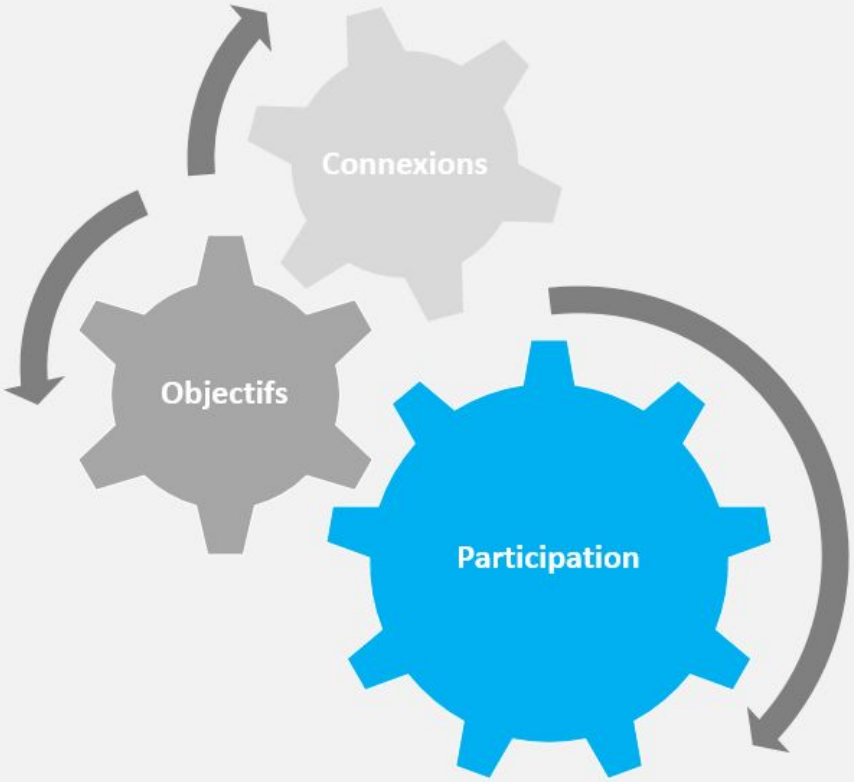
**Méthodes**

- Css, Html, Photoshop
- Python, SQL
- Organisation par groupe

**Difficultés**

Coopération avec le groupe web
Récupération des codes
Faible effectif (et peu de personnes avec les connaissances nécessaires)

# Bilan




- ✓ Les 15 sources
- ✓ Le stockage des données
- ✓ Le traitement des données
- ✓ L’analyse des données
- ✓ La qualité des données
- ✓ La visualisation des données

*Et maintenant…*


Merci de votre attention