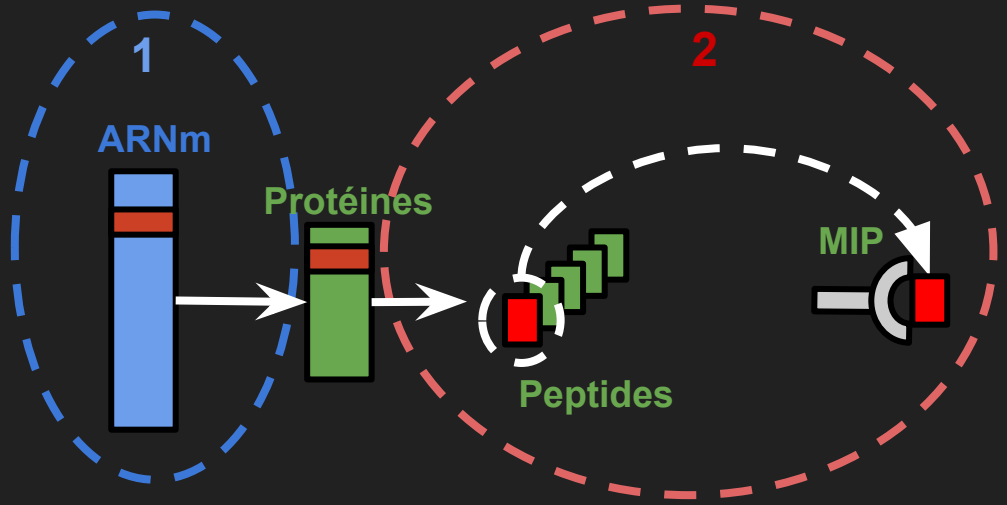
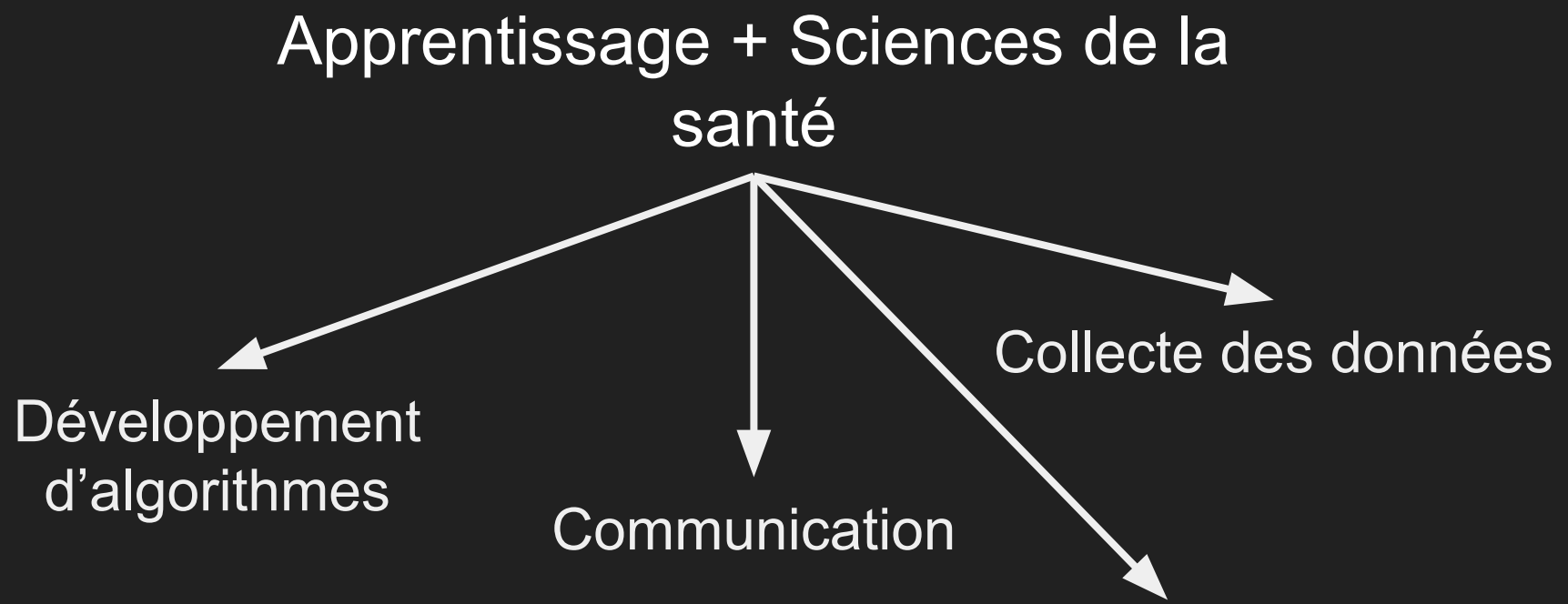
M.Sc
Apprentissage
Automatique

PhD
Bio-informatique




Est-ce que la traduction des protéines influe sur le système immunitaire?

# Défis


Apprentissage + Sciences de la santé
Développement d'algorithmes
Communication
Collecte des données
Gestion de données

# Comment rendre vos données conviviales pour l'apprentissage automatique

Tariq Daouda

Lab Perreault / Lemieux

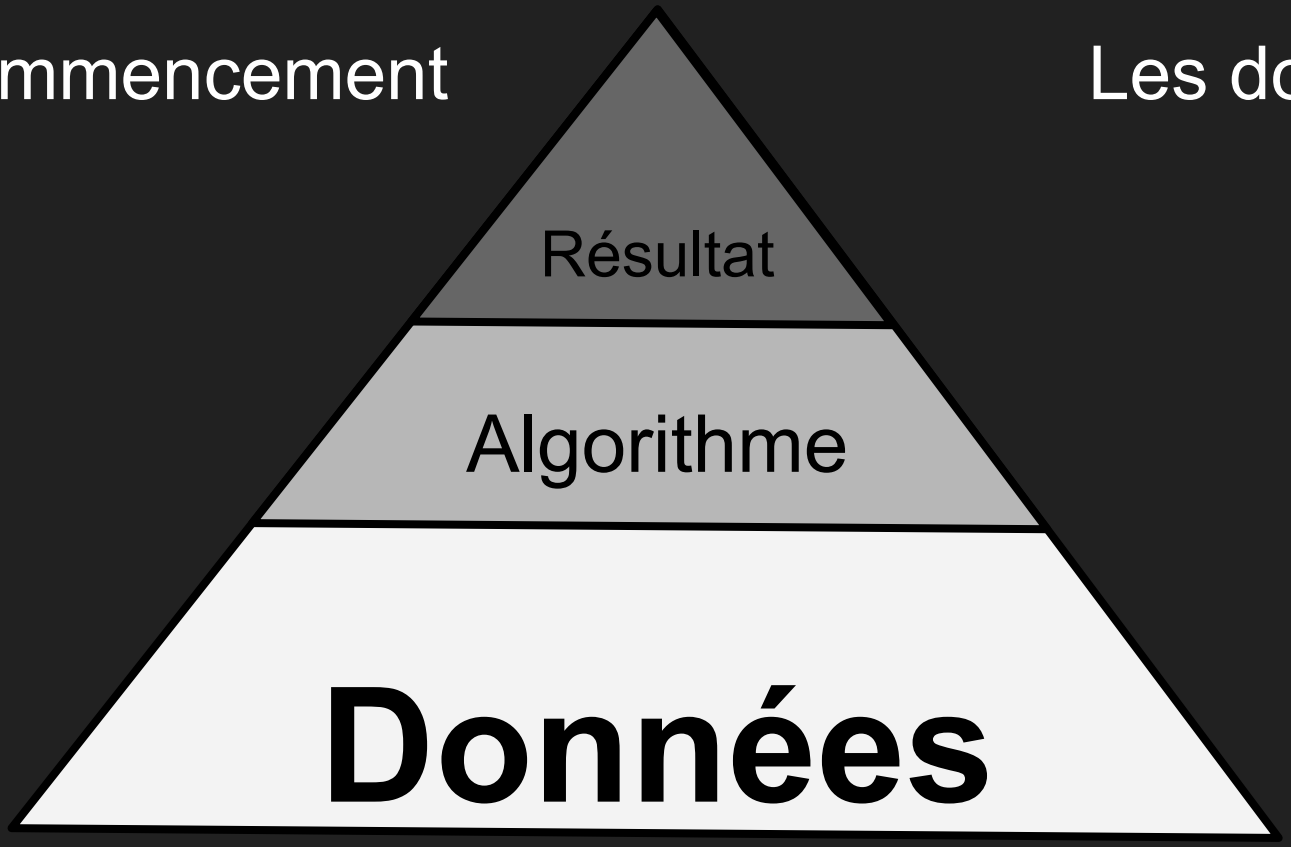
INSTITUT DE RECHERCHE
EN IMMUNOLOGIE ET
EN CANCÉROLOGIE

IRIC

Université de Montréal

Au commencement

Les données

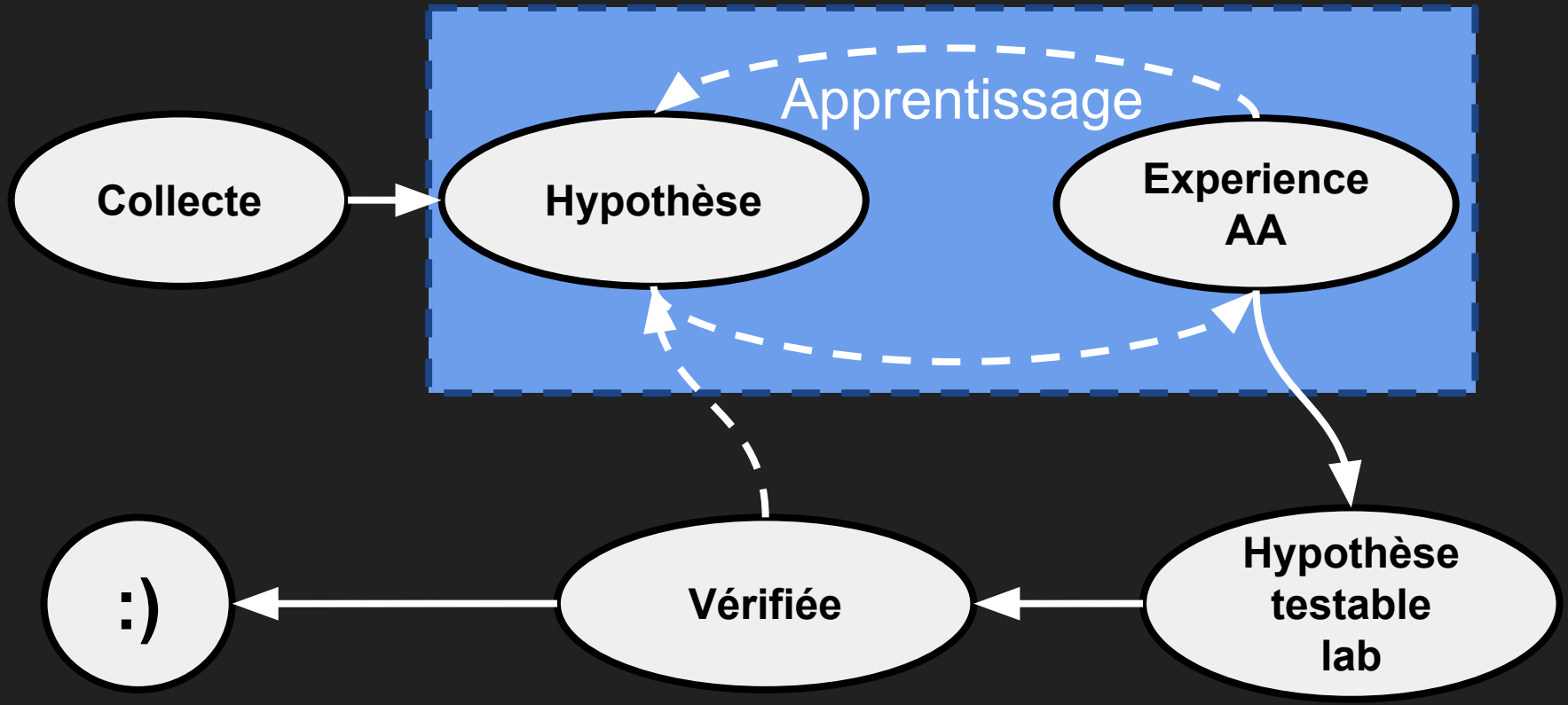



- 10s de Milliers / Millions examples

# Science de la santé : “Données vivantes”

- Répondre à des questions de recherche en biologie
- Donnée changeante (en cours d’acquisition)
- Expériences évoluent avec hypothèses de recherche
- Apprentissage Automatique est un outil parmis d’autres
- Analyse automatique de gros ensembles de données
  - Tester des hypothèses rapidement
  - Tester à moindre coût

# Processus intégré



# Science de la santé : “Données vivantes”

- Quelles sont les données pertinentes (filtrage)?
- Peut-on facilement enrichir les données (ajouts)?
- Les nouvelles données sont-elle compatibles avec les anciennes?
- Résultat négatifs : Le problème vient-il des données ou de l'algorithme apprentissage?
- Questions sur les données :
  - Provenance ?
  - Date d’acquisition ?
  - Protocole?

# Collecte et génération des données

# Generation de données

## Meta-données importantes

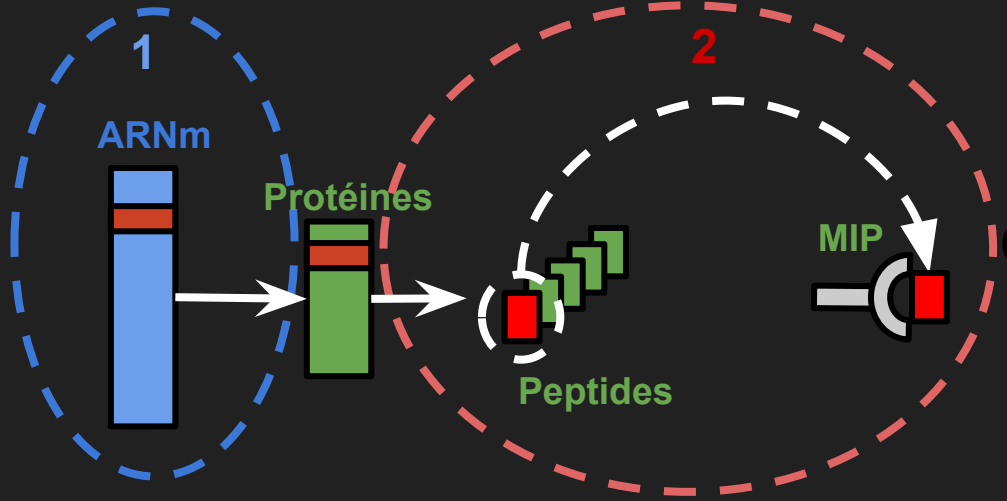
- Personne en charge de la collecte de données (questions)
- Lieu et date d'acquisition
- Protocole

# Generation de données

Fortement déconseillé

- Changer de nomenclature
  - Garder les mêmes noms de colonnes
  - Les mêmes identifiants de patients
  - Les mêmes identifiants de molécules
  - Terminologie, ex : Undef / None / ‘?’ / NA. Faire un choix

# Gestion des données et apprentissage




33 CMH-I

Proteome humain

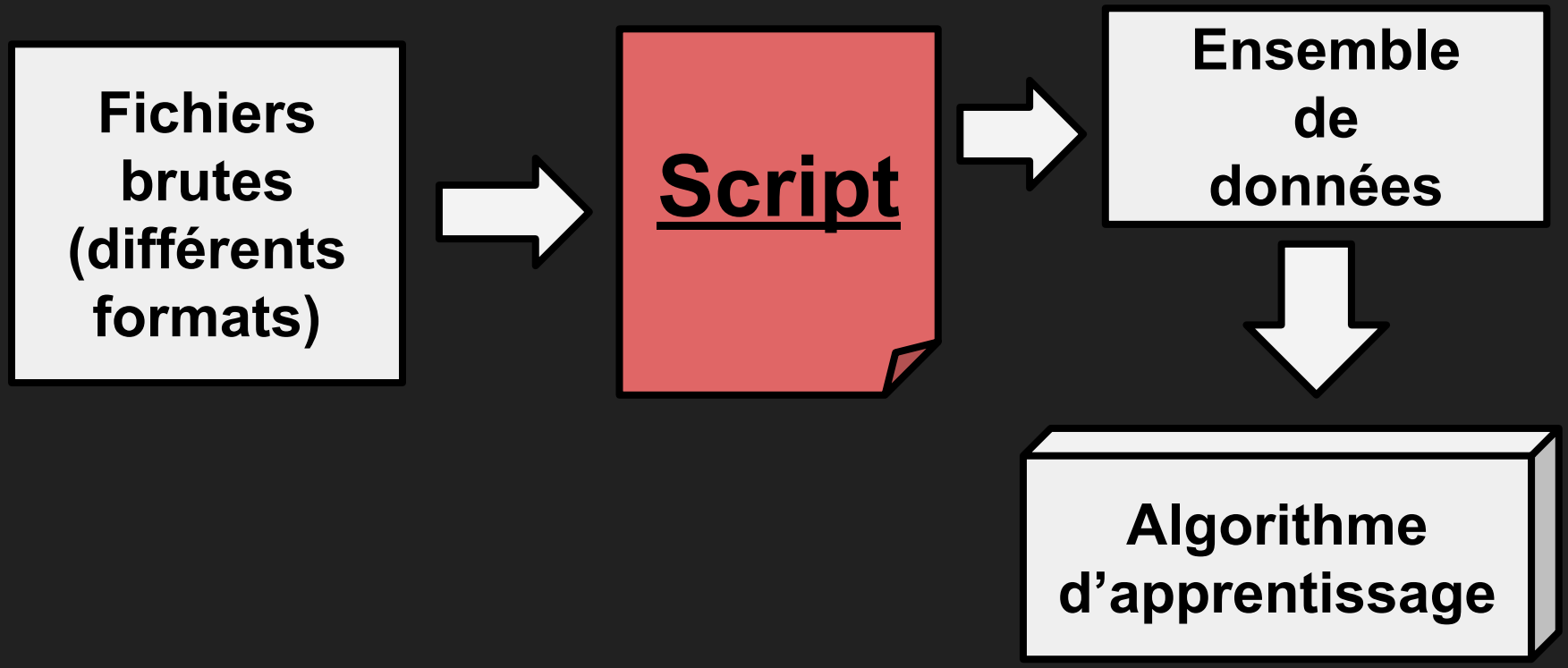
Transcriptome humain

55k Peptides

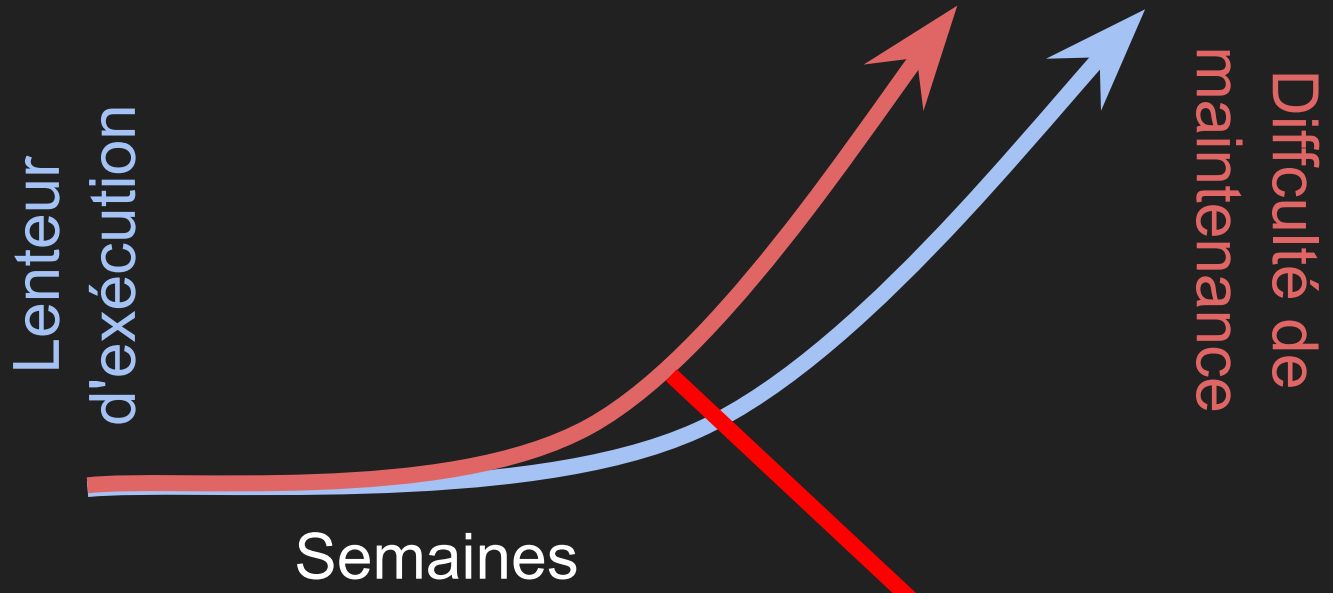
18 sujets (Seq ARNm)

Est-ce que la traduction des protéines influe sur le système immunitaire?

# Episode 1 : Approche naive

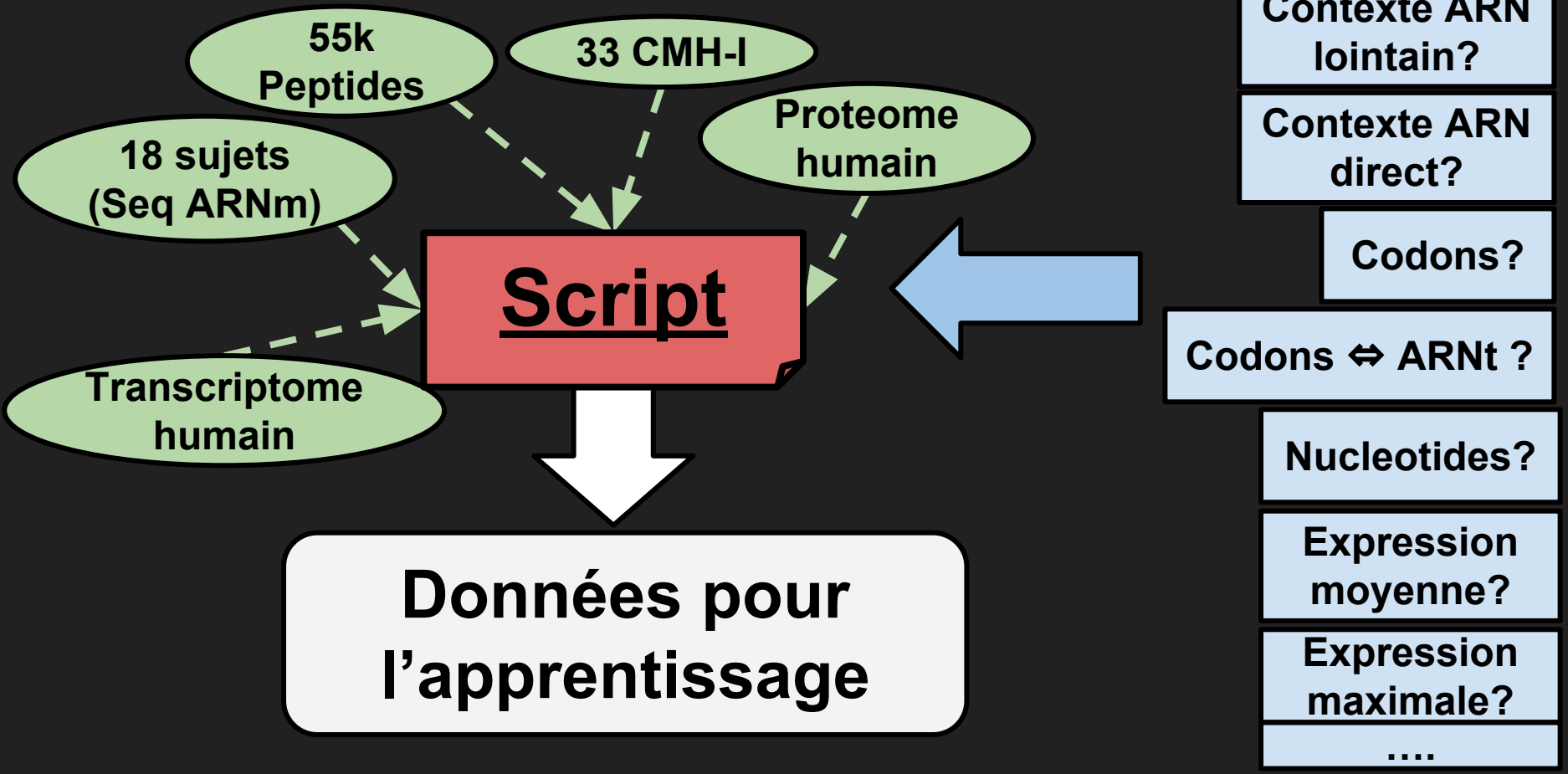


# Limitations

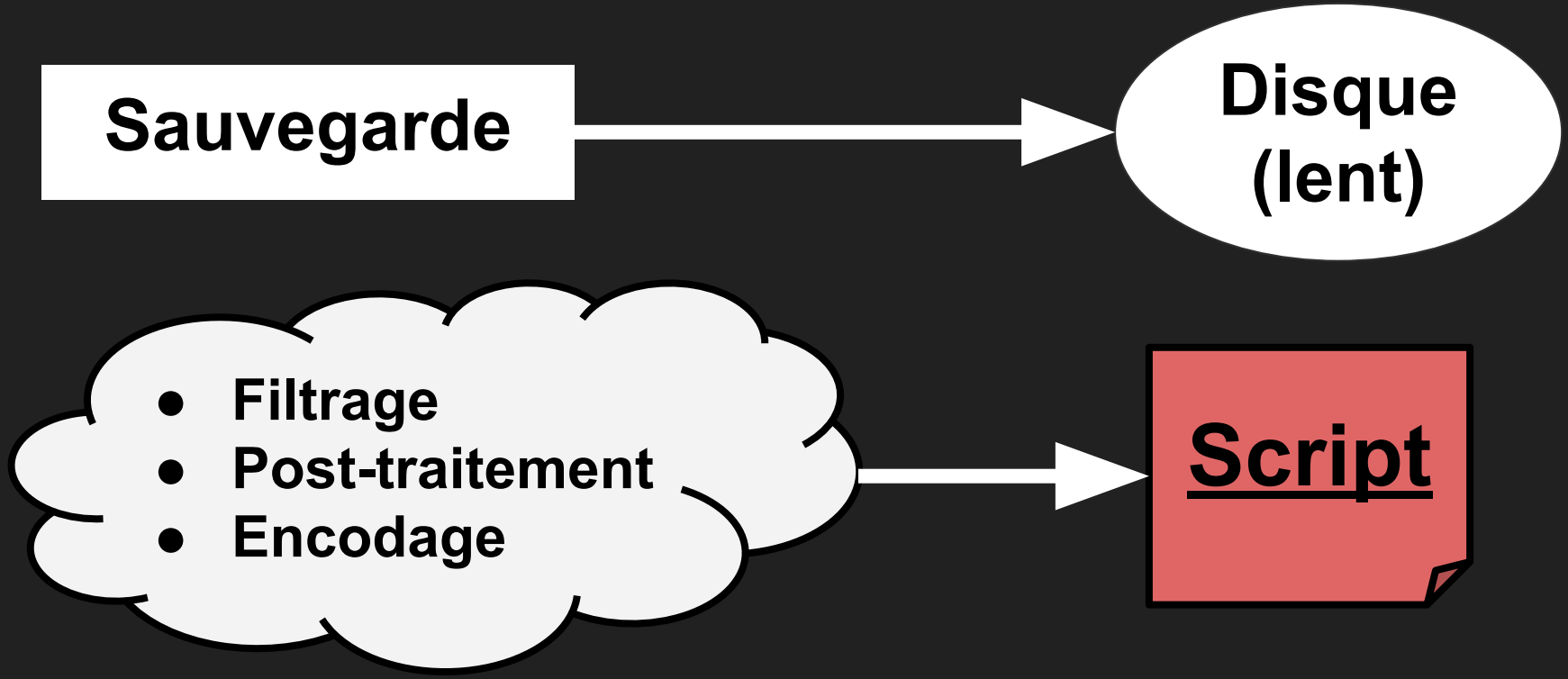


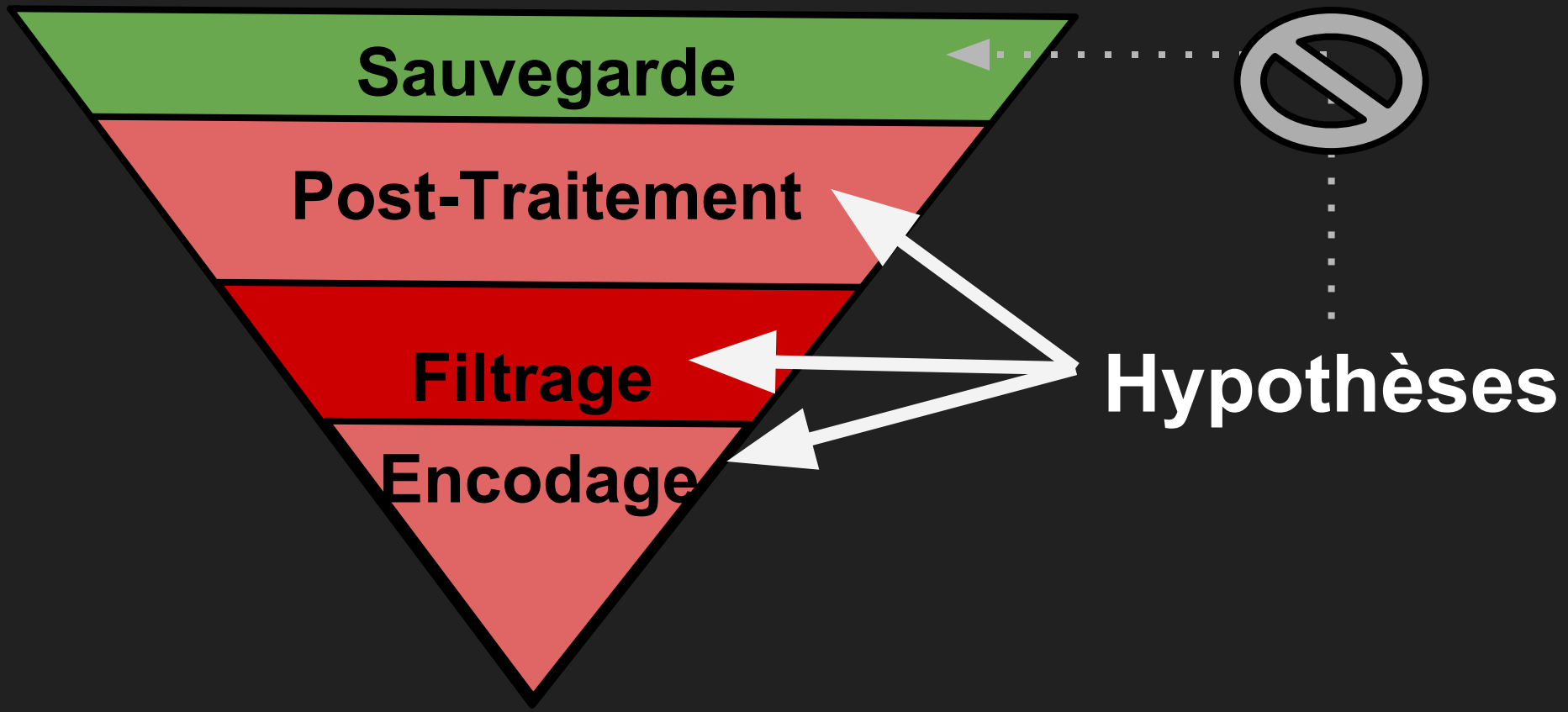

- Bugs!
- Reproductibilité
- Fausses conclusions

# 3, 4 ans d'Hypothèses :



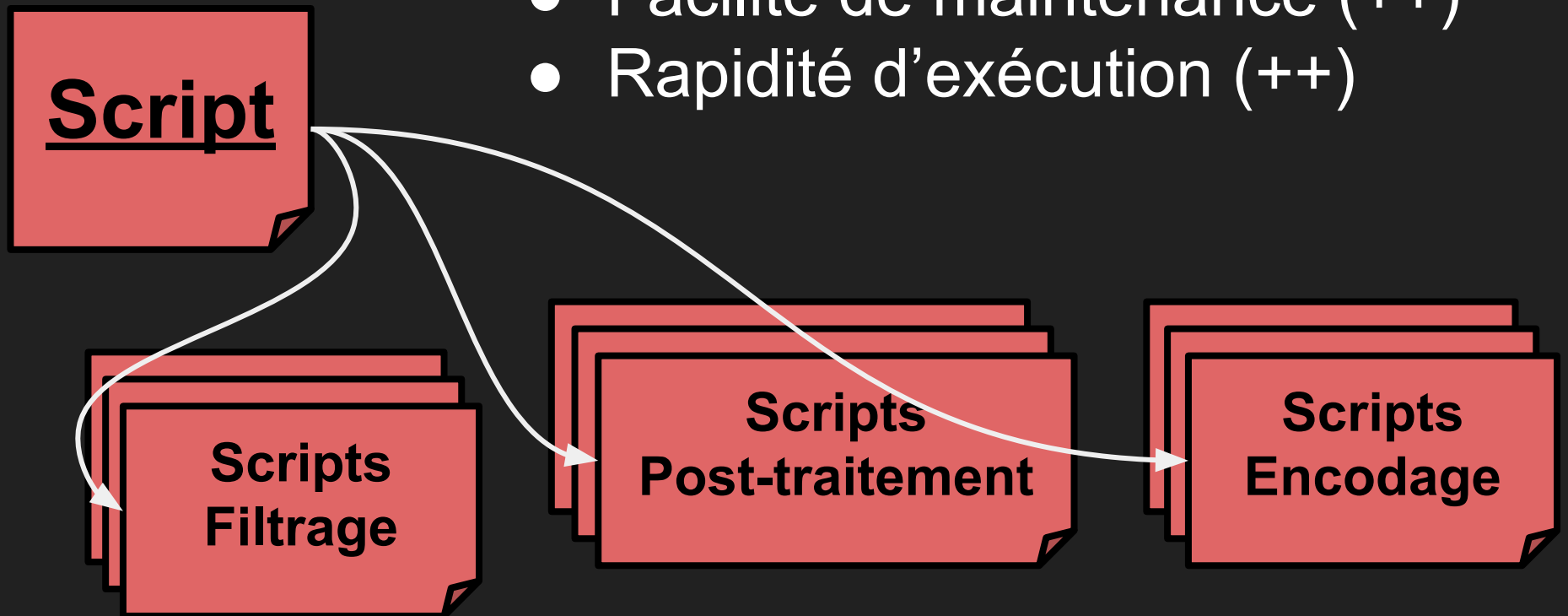
# Raisons


Sauvegarde
Disque
(lent)
Filtrage
Post-traitement
Encodage
Script

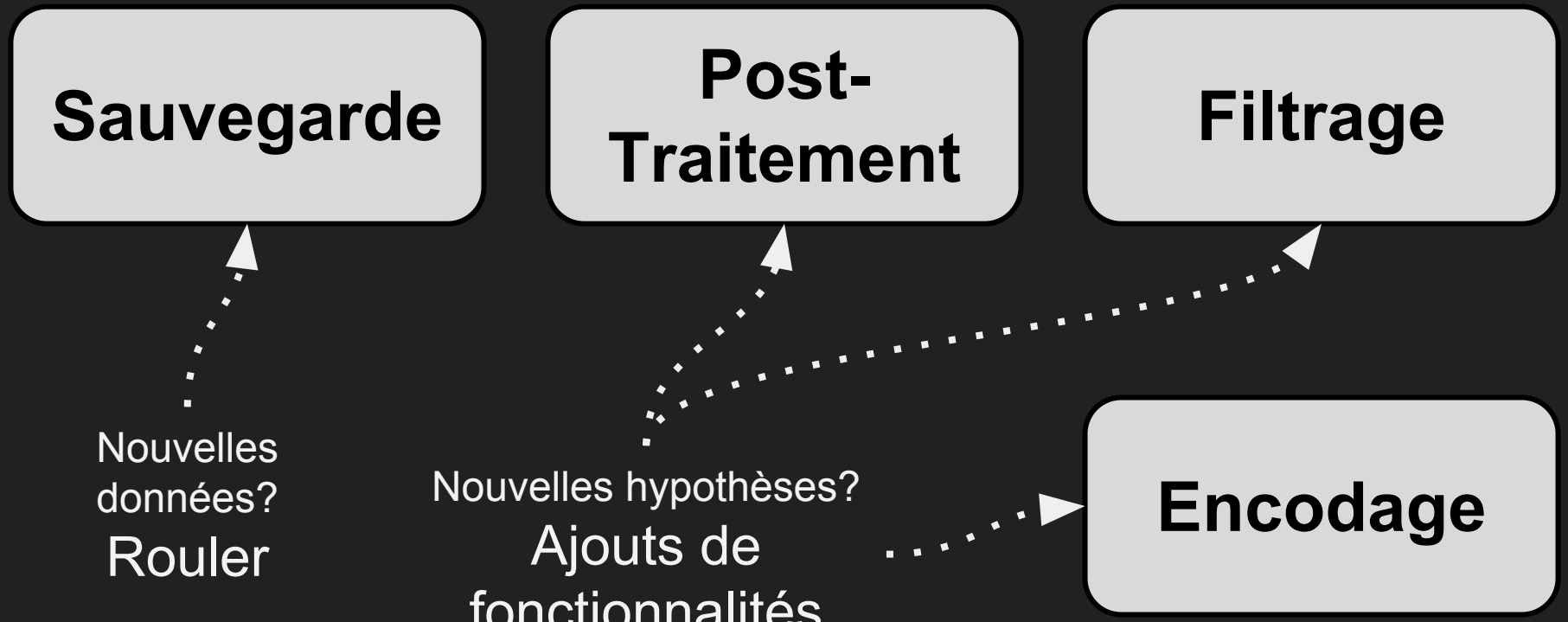

Sauvegarde
Post-Traitement
Filtrage
Encodage
Hypothèses

# Diviser pour mieux régner

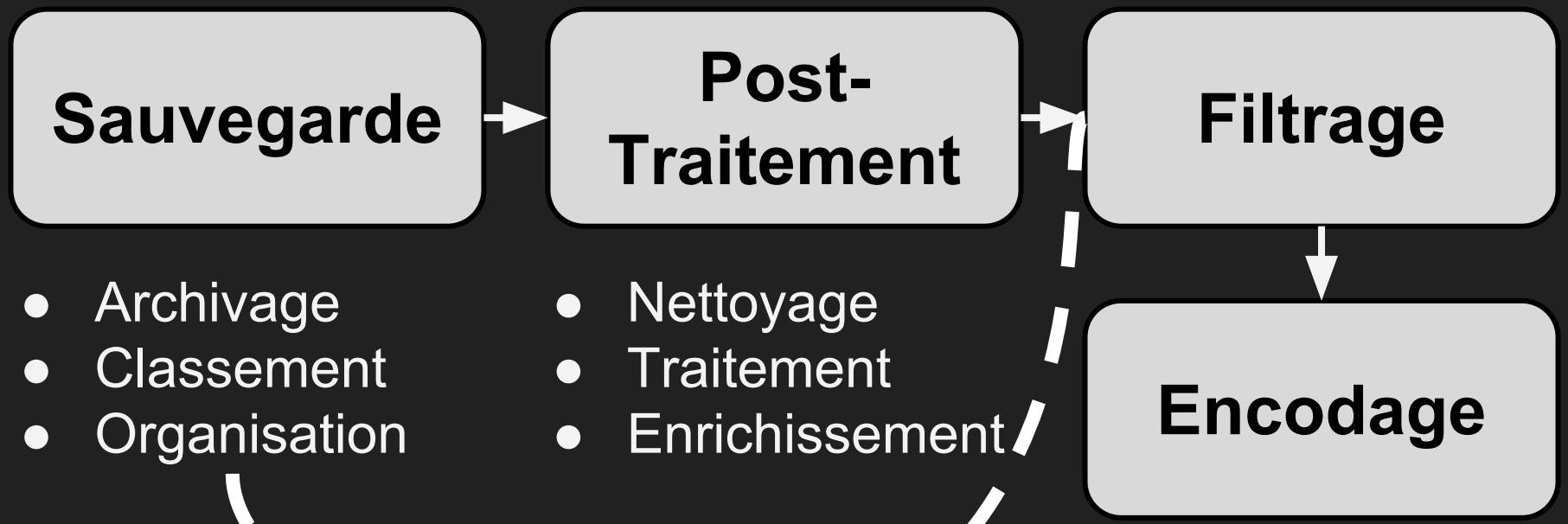
- Facilité de maintenance (++)
- Rapidité d’exécution (++)



# Système : S.P.F.E, diviser pour mieux régner



Système : S.P.F.E



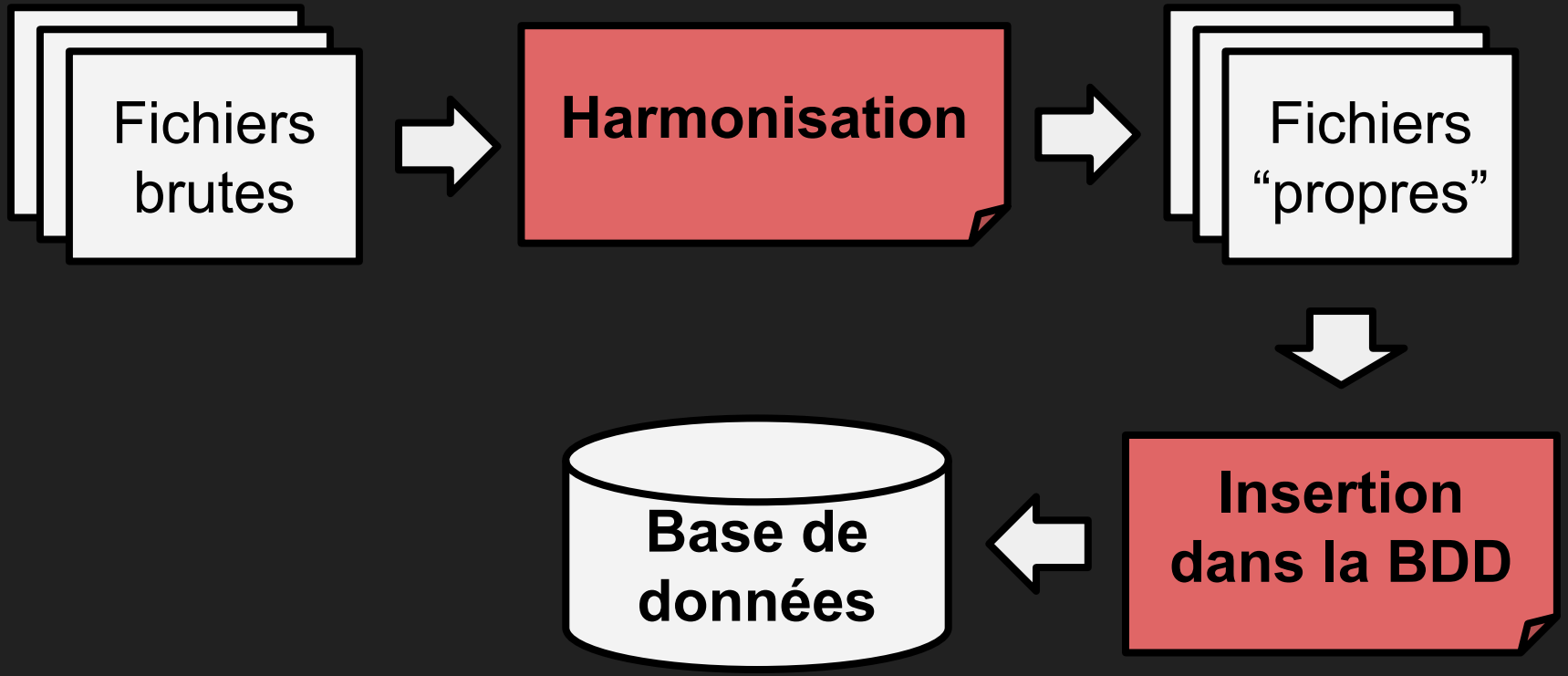

- Archivage
- Classement
- Organisation

- Nettoyage
- Traitement
- Enrichissement

- Selection des données

- Présentation à l'algorithme

# Sauvegarde


Fichiers brutes
Harmonisation
Fichiers “propres”
Insertion dans la BDD
Base de données

# Details : Sauvegarde

- IMMUABLES, ne modifier sous aucun prétexte
- Toujours être capable de retourner aux versions originales

Fichiers brutes

**Harmonisation**

- Rétablir noms de colonnes
- Normaliser la nomenclature. Ex : ‘Undef’ / ‘?’
- Ne rien effacer

- Exact même format que les fichiers bruts
- Communication avec collaborateurs

Fichiers “propres”

# Details : Sauvegarde

- Insère le contenu complet des fichiers propres dans un système de base de données

**Insertion dans la BDD**


- Moderne, rapide (indexes)
  - Archivage
- (+) Interface graphique

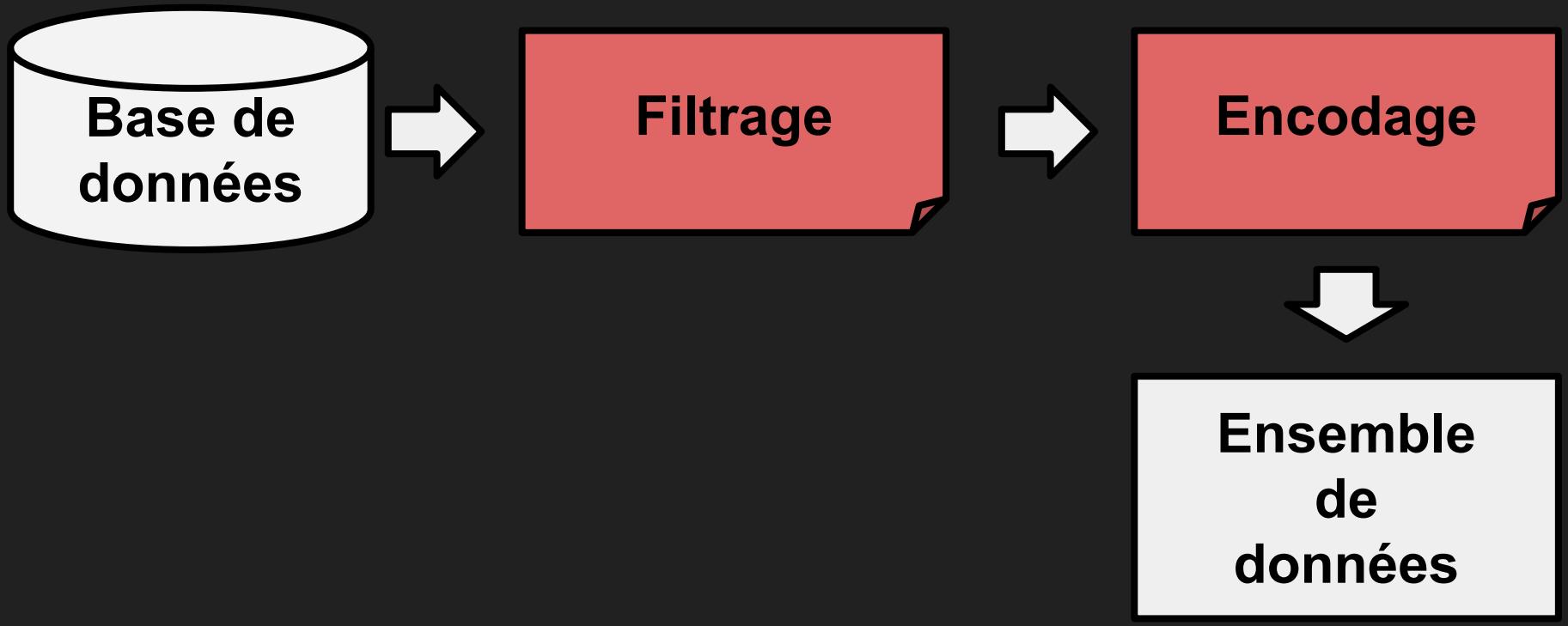
# Post-traitement

# Details : Post-traitement

- Ajouts à la base de donnée
- Informations utilisés pour le filtrage
- Métriques importantes
  - Distances
  - Expressions moyennes (sujets)
- Transformations
  - Normalisations
  - Log

**Post-traitement**
**+**
**Enrichissement de données**

# Filtrage / Encodage



# Details : Filtrage / Encodage

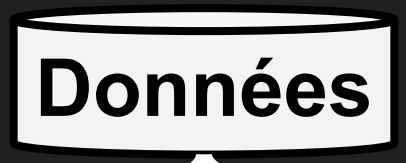
**Filtrage**

- Sélection des données
- Utiliser les fonctionnalités de la BDD
  - Rapide (index)
  - Fiable, reproductible
  - Language adapté (SQL, AQL) Vs Boucles
  - Plus facile à maintenir

**Encodage**

- Présentation à l’algorithme
- Flottants / Entiers
- Identifiants uniques (mots)

# Details : Encodage




**Continues**

- **Images**
- **Expressions**

**Discrètes**

- **Classes**
- **Catégories**
- **Mots**

- Réduire écarts entre valeurs
  - Expressions => log(expressions)

**1 Item => 1 entier unique**

| | |
|---|---|
| 0 | ALL |
| 1 | AML |
| 2 | CML |
| ... | ... |

# Format standard : tableau numpy (numpy array)

**Continues**

- **Images**
- **Expressions**

**Discrètes**

- **Classes**
- **Catégories**
- **Mots**

**float32**

**int16**

```
data = numpy.asarray(data, dtype='int16')
```

```
data = numpy.asarray(data, dtype='float32')
```

# Détails : Ensemble de données (fichier)

**Ensemble de données**

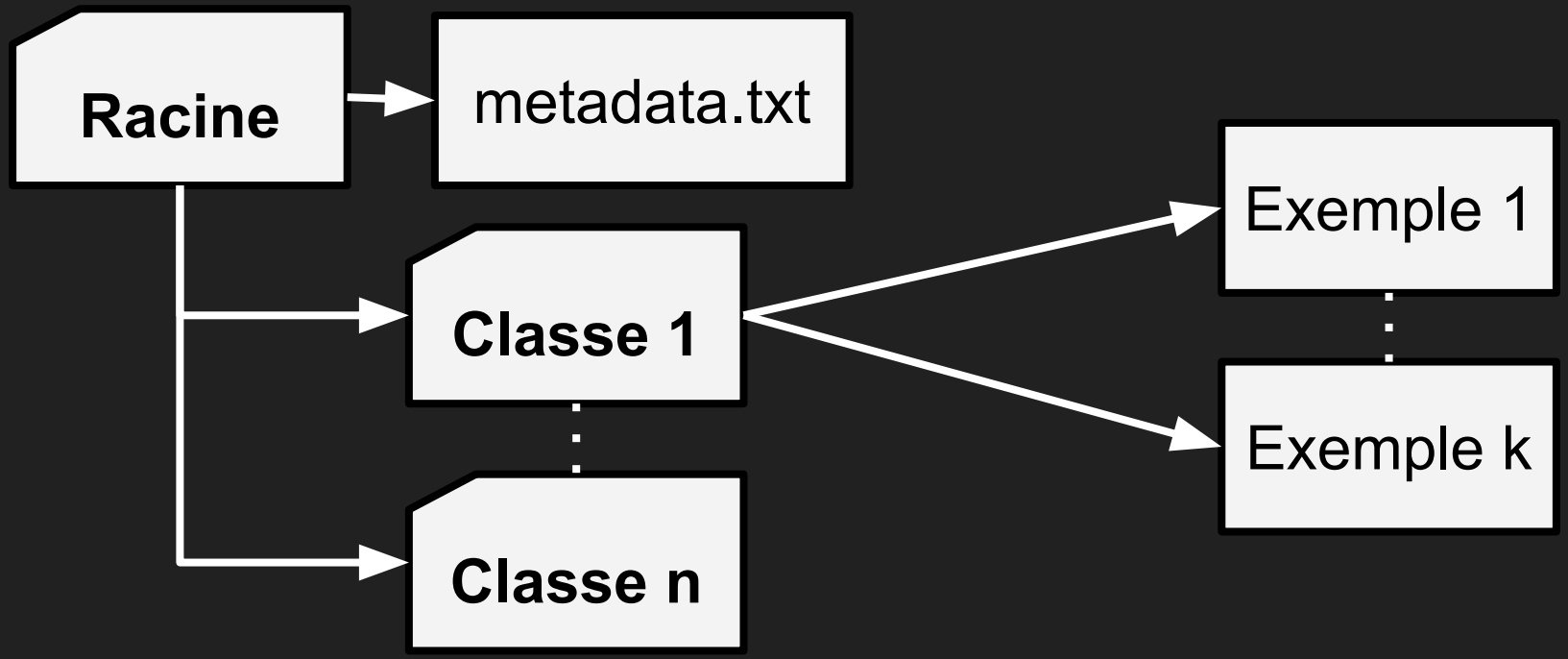
- Peut être effacé et régénéré au besoin
- Format facile à digérer
- Ex : Dictionnaire Pickle (nom des champs explicites)
- Inclure metadata : Filtres, normalisations

- Nom de fichier unique :
  - Filtres utilisés
  - Date
- Fichier texte accompagnant
  - Metadata
  - Date de génération
  - Notes

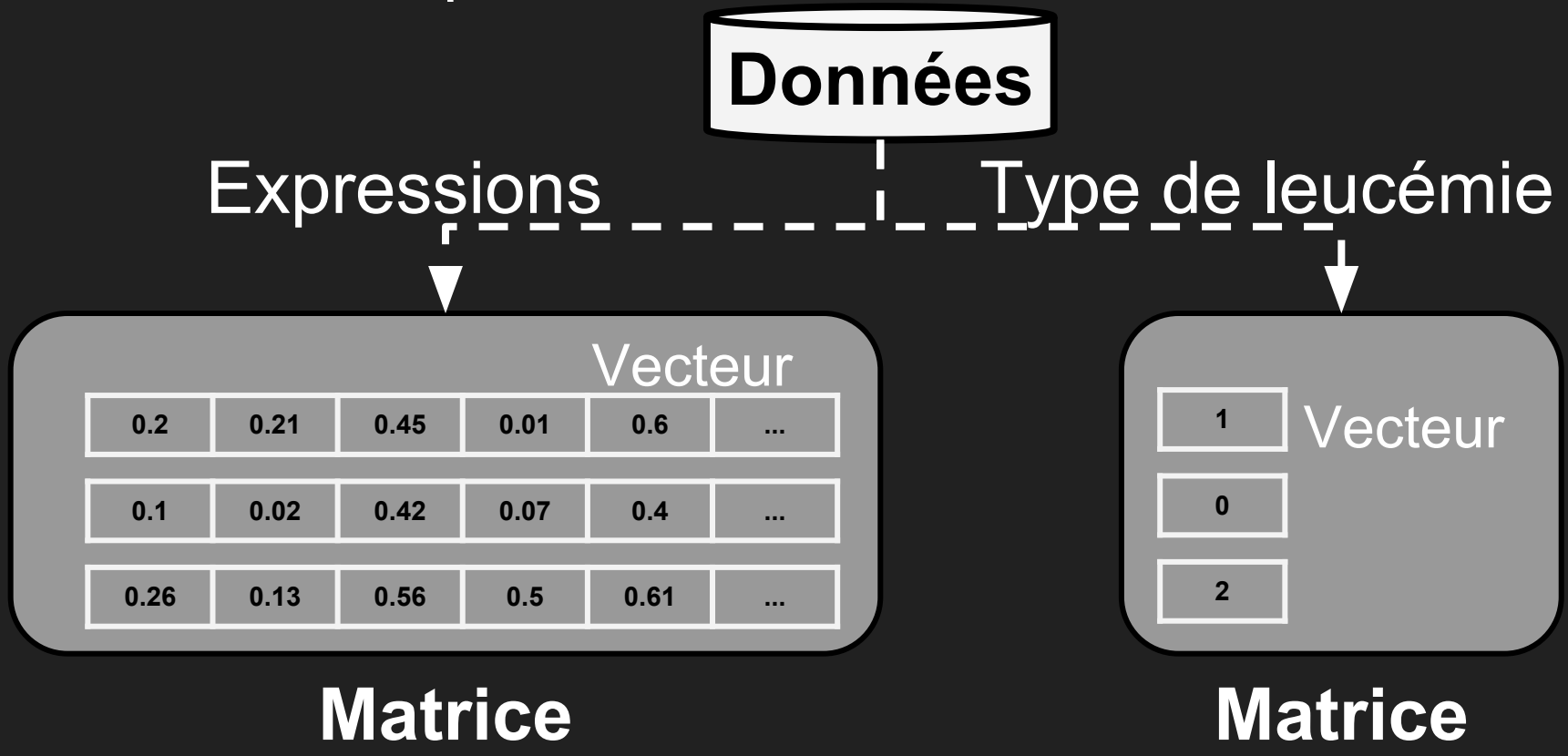
```
516  res = {
517
518      "data" : {
519          "train": ...,
520          "test": ...,
521          "validation": ...,
522      }
523      "meta-data" {
524          "filters": ...
525          "normalization": ...
526          "generation-date": ...
527      }
528  }
```

# Détails : Ensemble de données

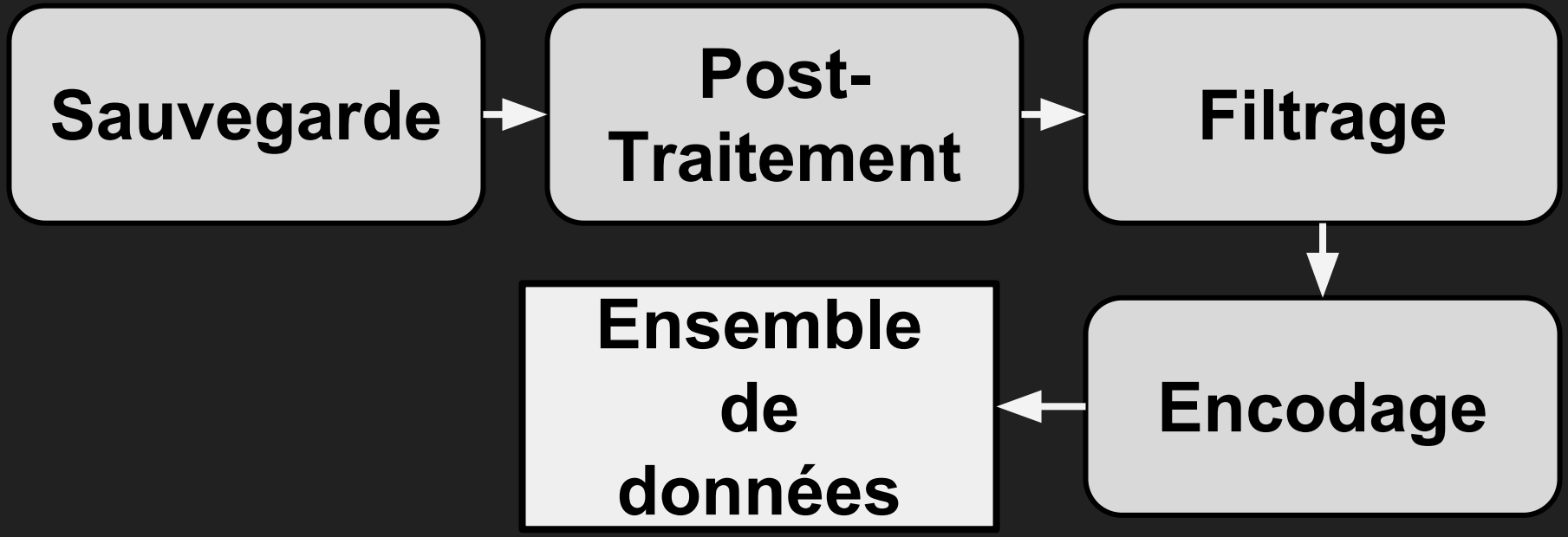
- Hiérarchie de dossiers



# Ex : Ensemble pour classification



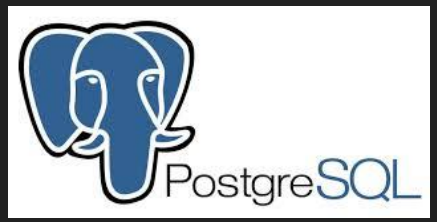
# Système : S.P.F.E, conclusions




- Maintenance plus simple
- Plus de confiance : Reproducibilité
- Meilleure vitesse d’exécution

# Quel système de base de données utiliser?

- SQL
  - Standard
  - Plus rigide
  - Grande protection de l'intégrité des données
- NoSQL
  - Plus de flexibilité
  - Plus de responsabilités



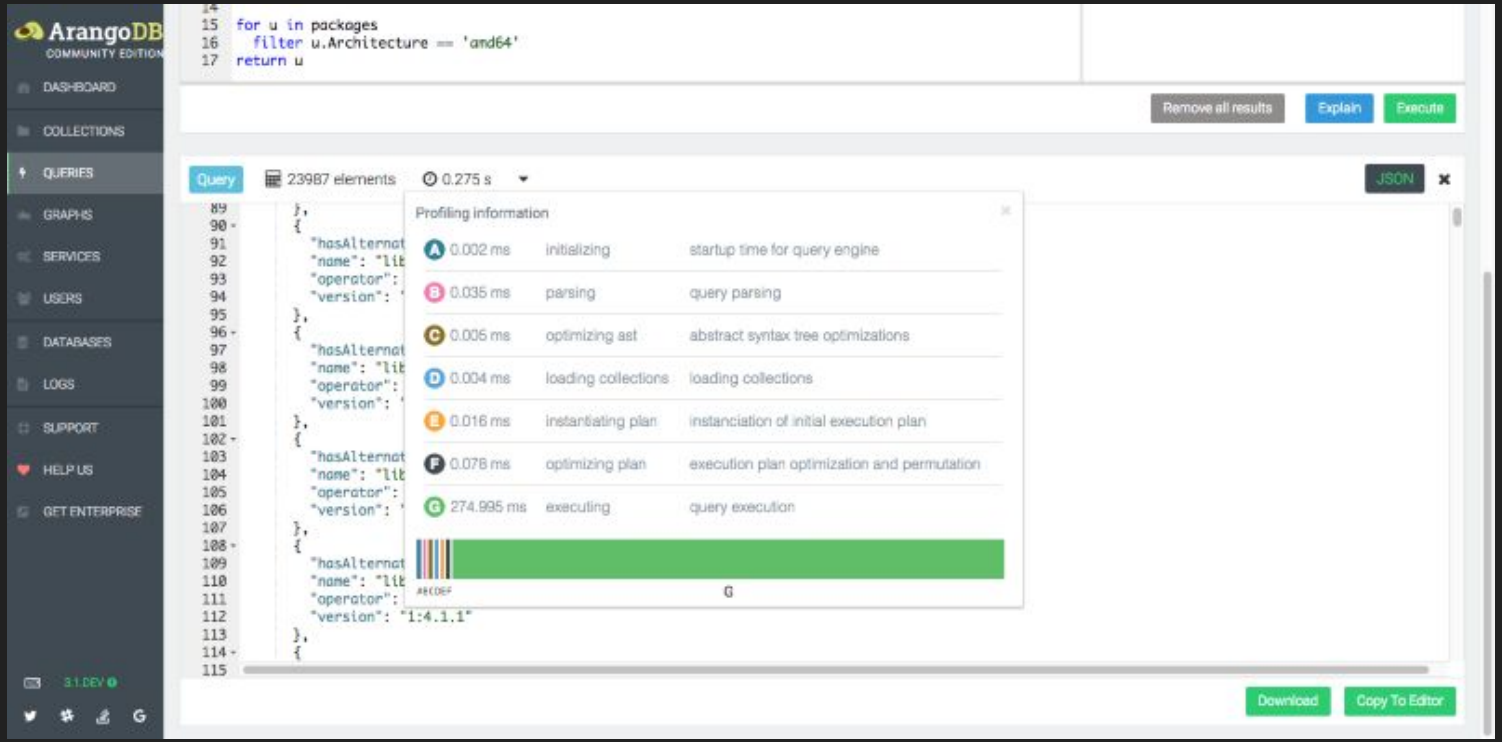



# Mon choix : ArangoDB

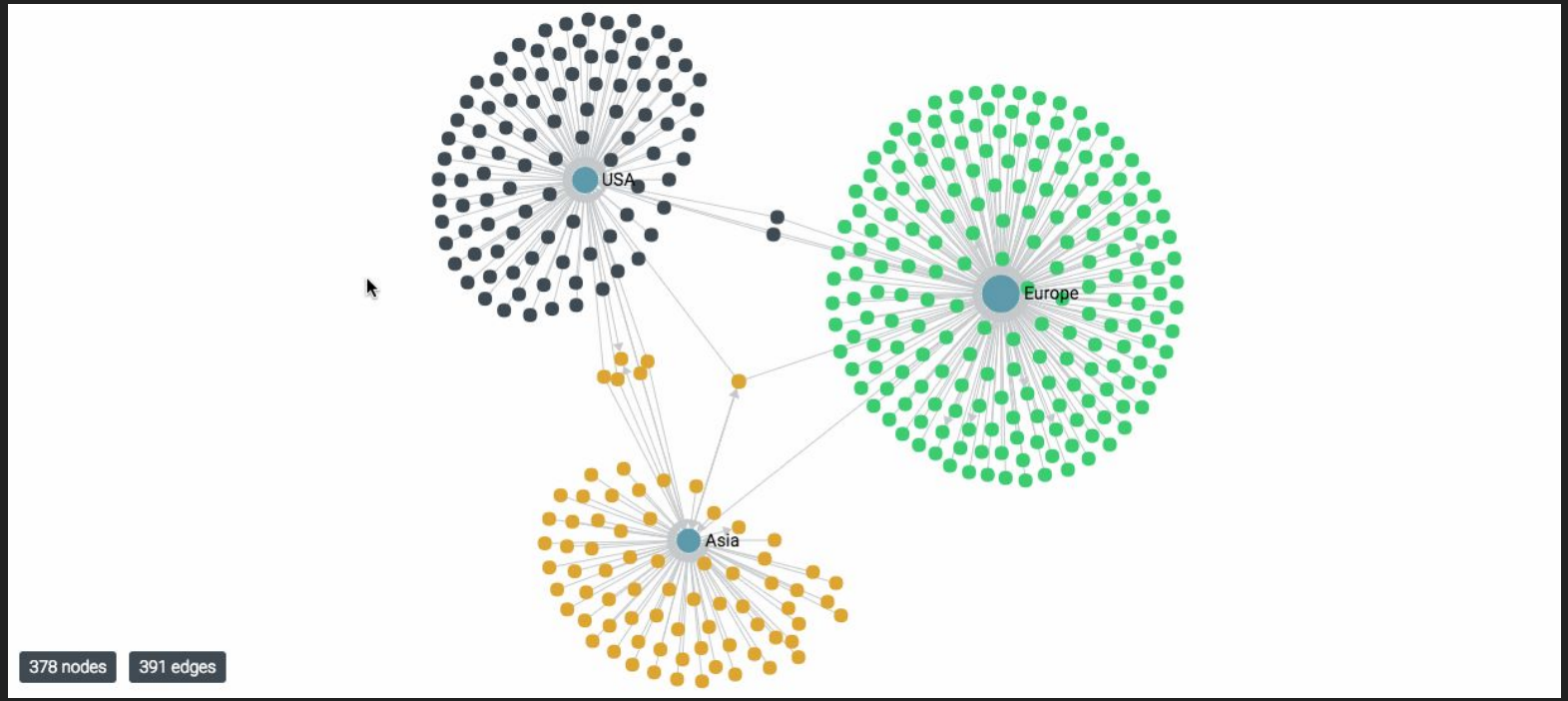

- Rapide
- Multimodale
  - Graphe
  - Documents
  - Clef-valeur

- Language de requêtes clair : AQL
- Interface web
  - Tests requêtes
  - Collaboration
  - Exploration

# Interface graphique : test de requêtes



# Interface graphique : Visualisation de graphes

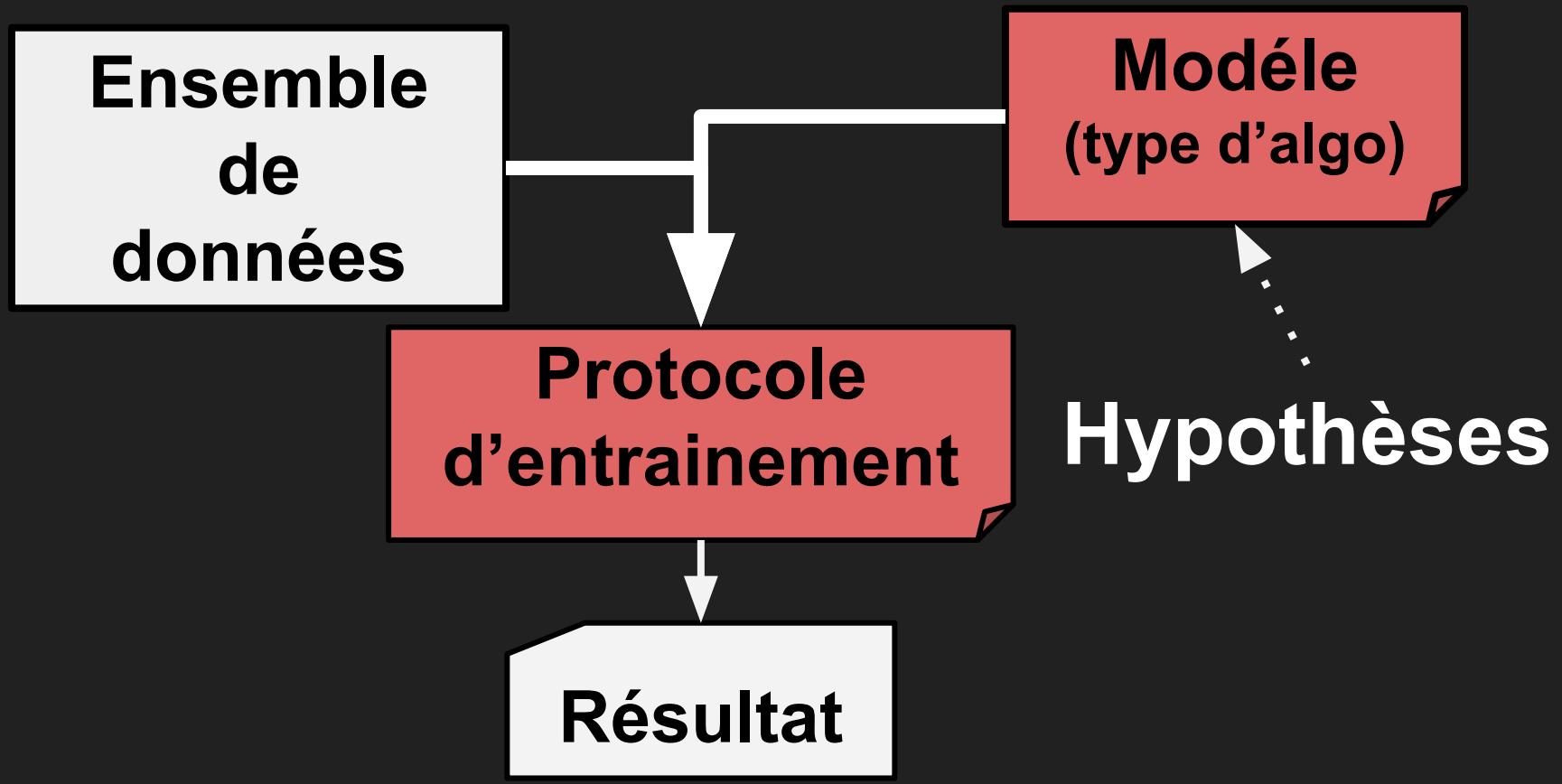


# Images et gros fichiers

- Ne pas les mettre dans la Base de donnée
- Enregistrer sur le disque
- Enregistrer le chemin vers le fichier dans la BDD
- Utiliser plusieurs dossiers si nécessaire

# Géstion des expériences

Données Modèle Protocole Résultats



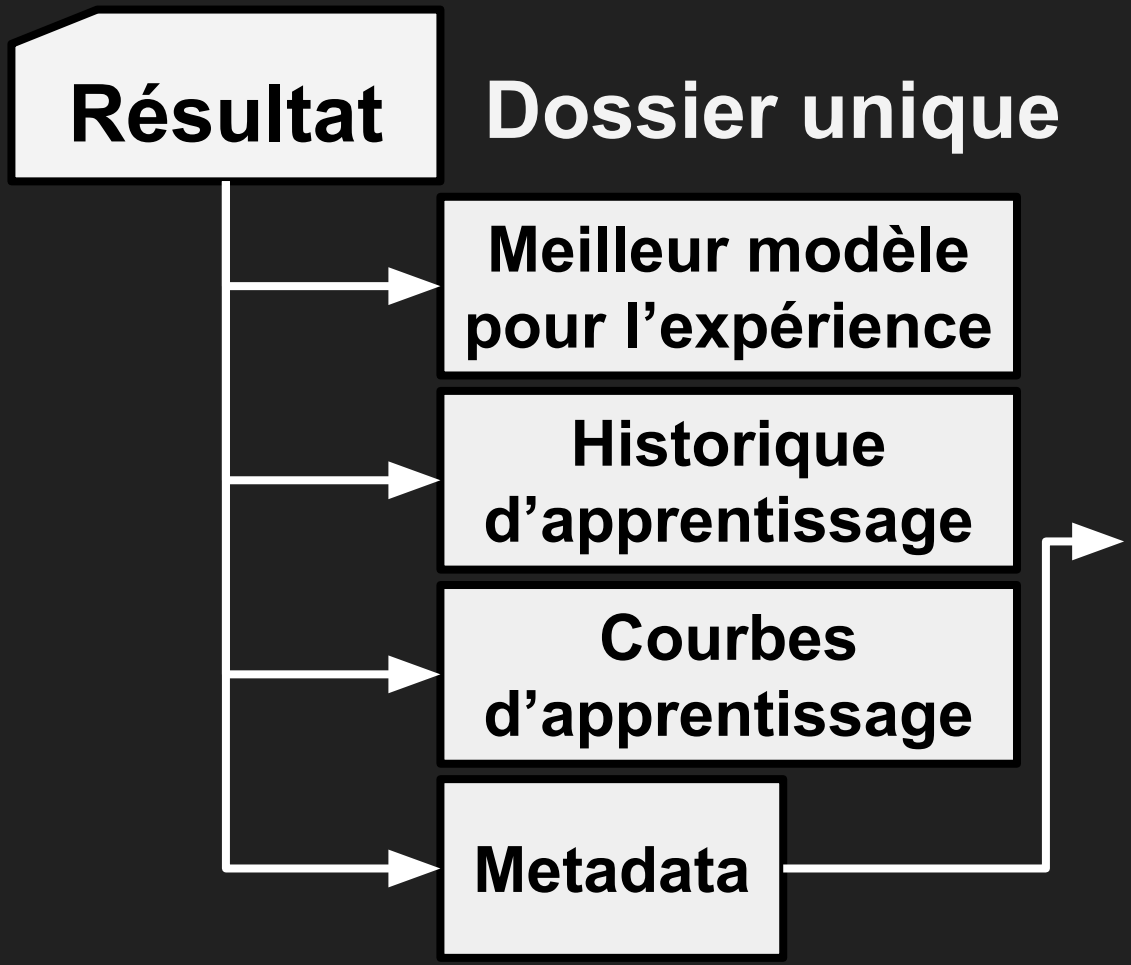
# Détails

**Modéle**

- Défini l’algorithme d’apprentissage
- Ex : architecture du réseau de neurone

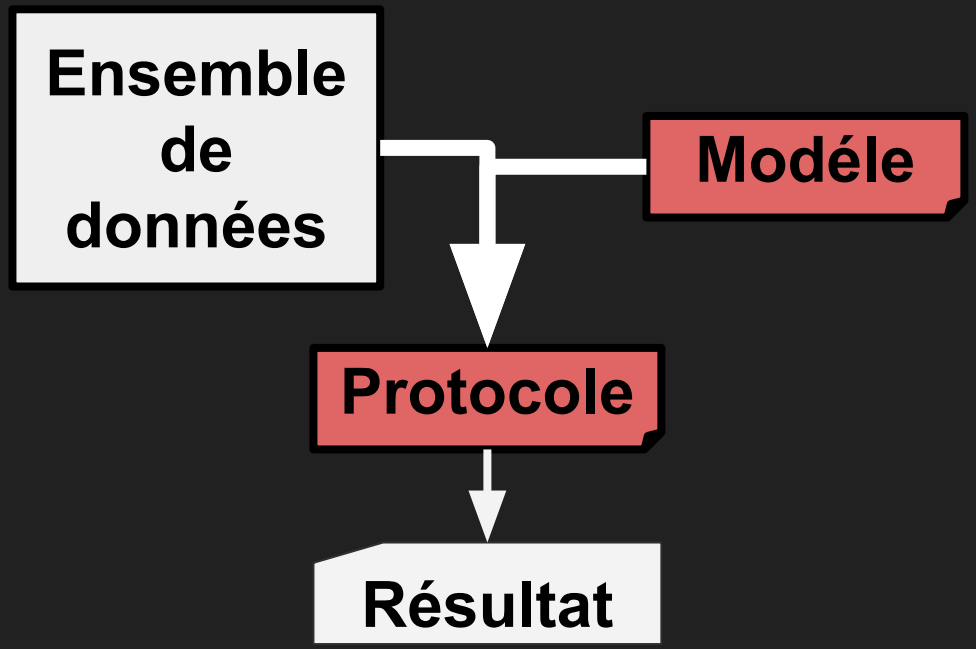
**Protocole**

- Met en relation ensemble de données et modèle
- Comment entrainer
- Boucle d’apprentissage
  - Nombre epoques
  - Taille de mini-batch
- Sauvegarde des résultats (dossier unique)




- Architecture du modèle
- Date d'entraînement
- Protocole
  - Hyper-paramètres
  - Minibatch
  - Nb epoques
- Nom : Ensemble de données

# Conclusions




- Facilité de maintenance
  - Moins de Copier-Coller
- Meilleure reproductibilité
  - Plusieurs modèles
  - Même protocole
- Meilleure traçabilité
- Format de résultat standardisé

.FIN.