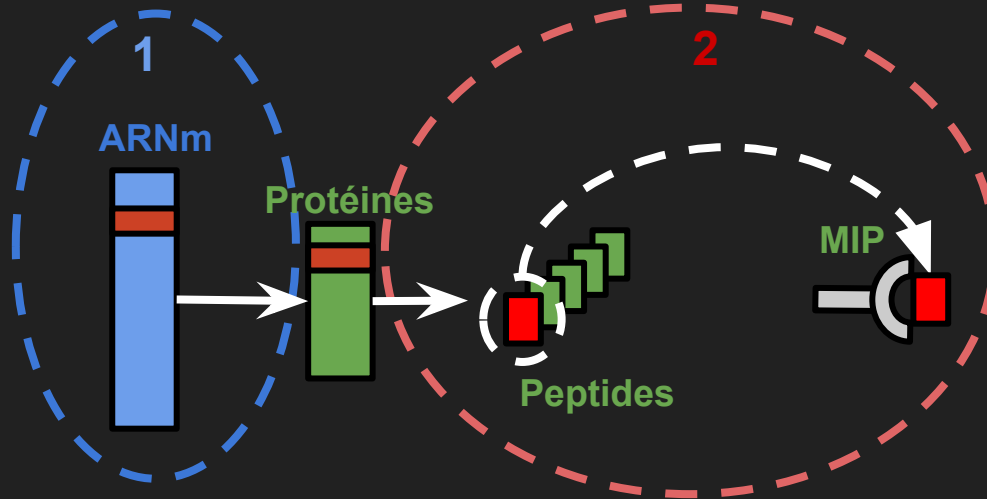


M.Sc

PhD

Apprentissage
Automatique

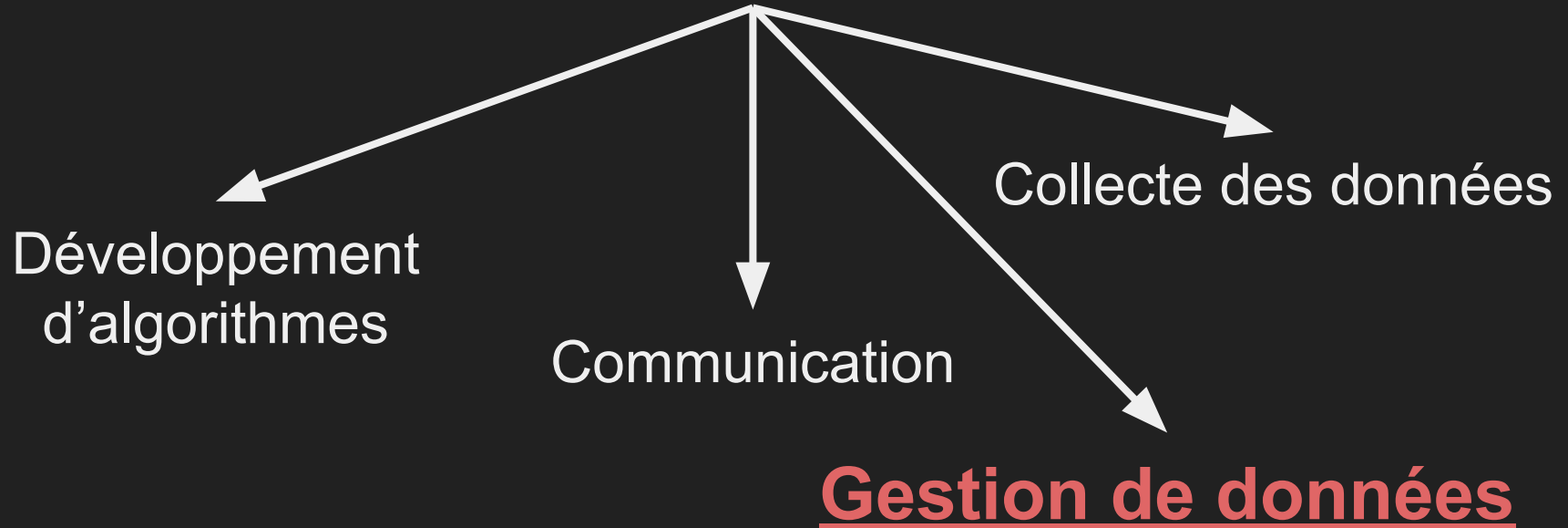
Bio-informatique



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

Défis

Apprentissage + Sciences de la
santé



Comment rendre vos données conviviales pour l'apprentissage automatique

Tariq Daouda

Lab Perreault / Lemieux

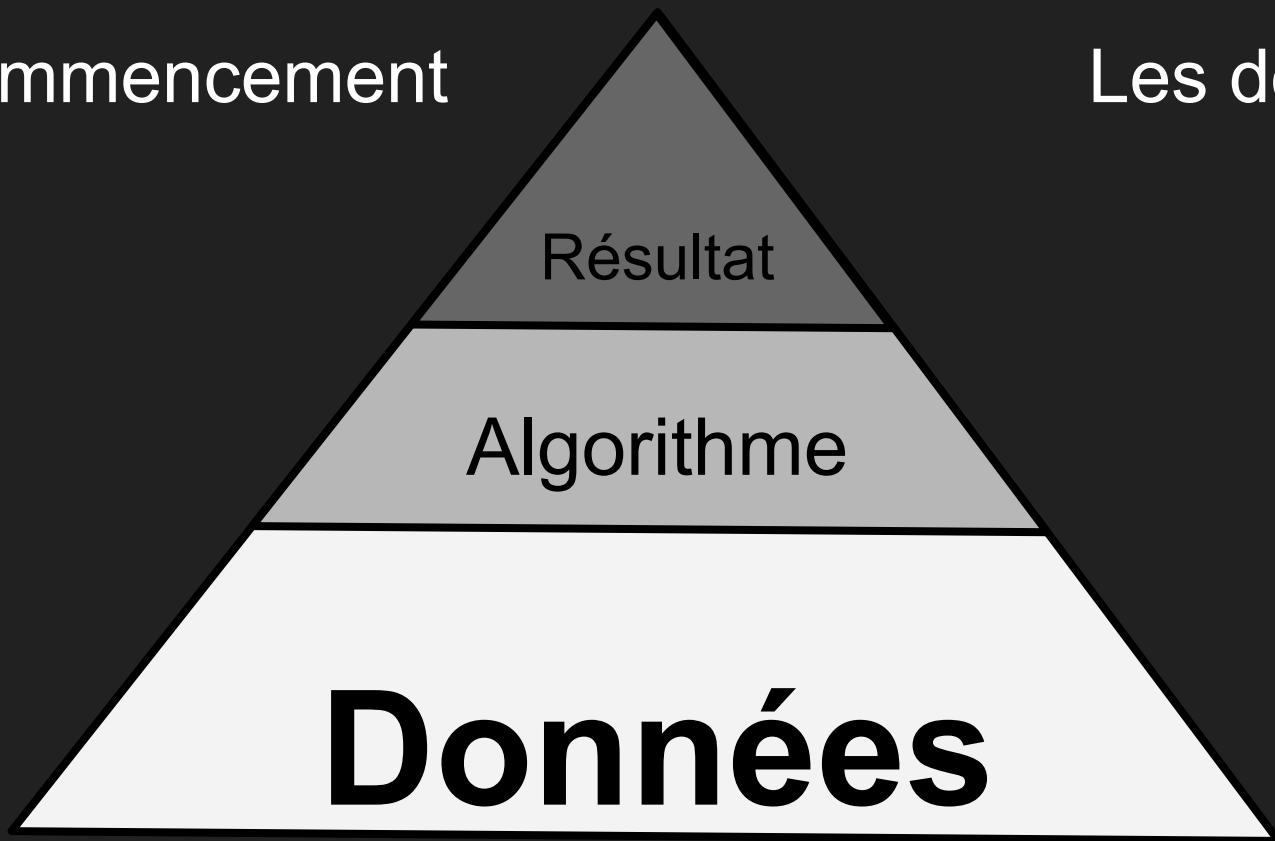
INSTITUT DE RECHERCHE
EN IMMUNOLOGIE ET
EN CANCÉROLOGIE



Université 
de Montréal

Au commencement

Les données

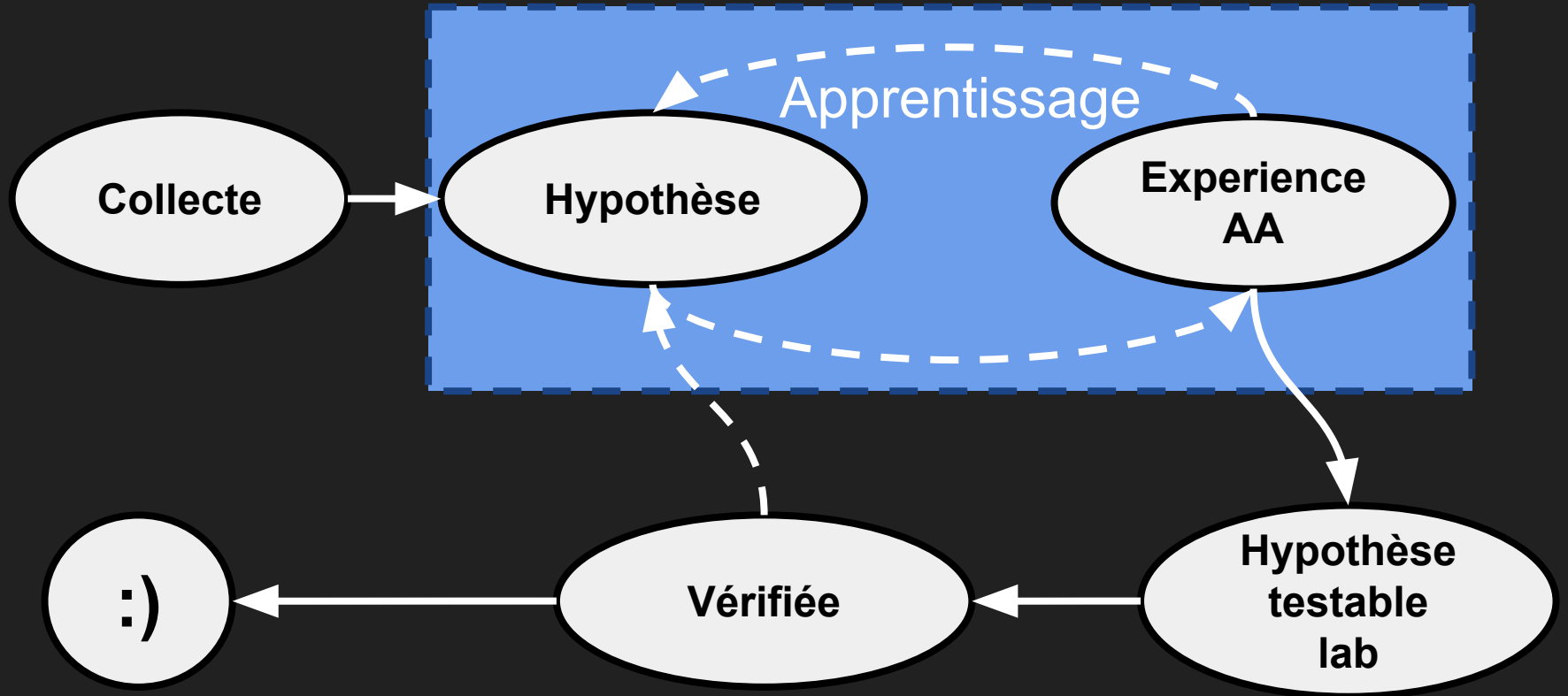


- 10s de Milliers / Millions exemples

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 parmi 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?
- Résultats négatifs : Le problème vient-il des données ou de l'algorithme d'apprentissage?
- Questions sur les données :
 - Provenance ?
 - Date d'acquisition ?
 - Protocole?

Collecte et génération des données

Generation de données

Meta-données importantes

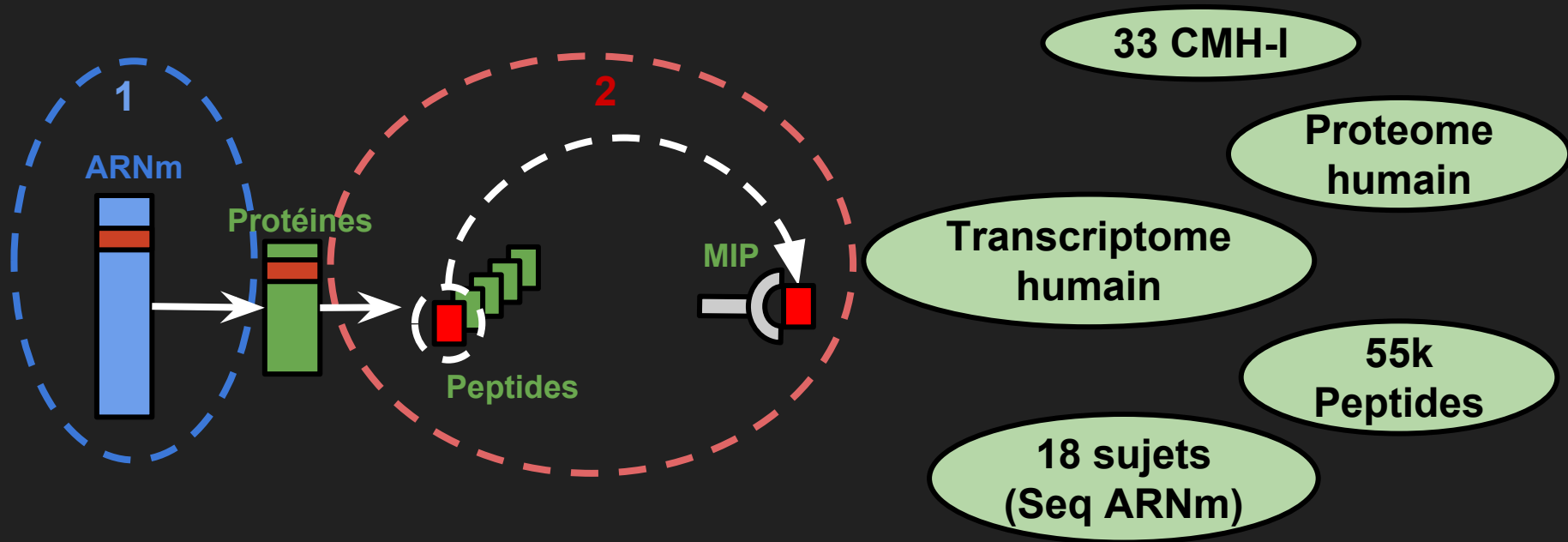
- Personne en charge de la collecte de données (questions)
- Lieu et date d'acquisition
- Protocole

Generation de données

Fortement déconseillé

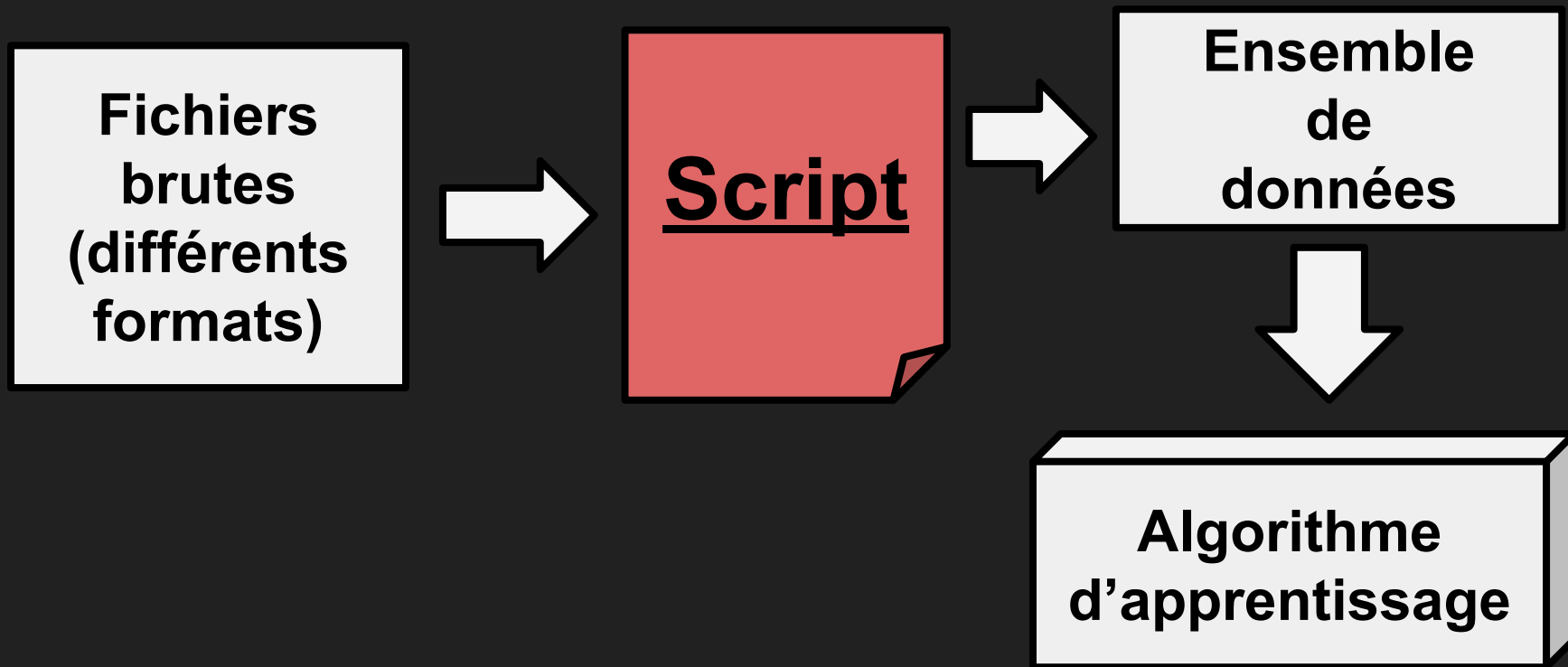
- Changer de nomenclature
 - Garder les mêmes noms de colonnes
 - Les mêmes identifiants de patients
 - Les mêmes identifiants de molécules
 - Terminologie, ex : Undef / None / '?' / NA. Faire un choix

Gestion des données et apprentissage

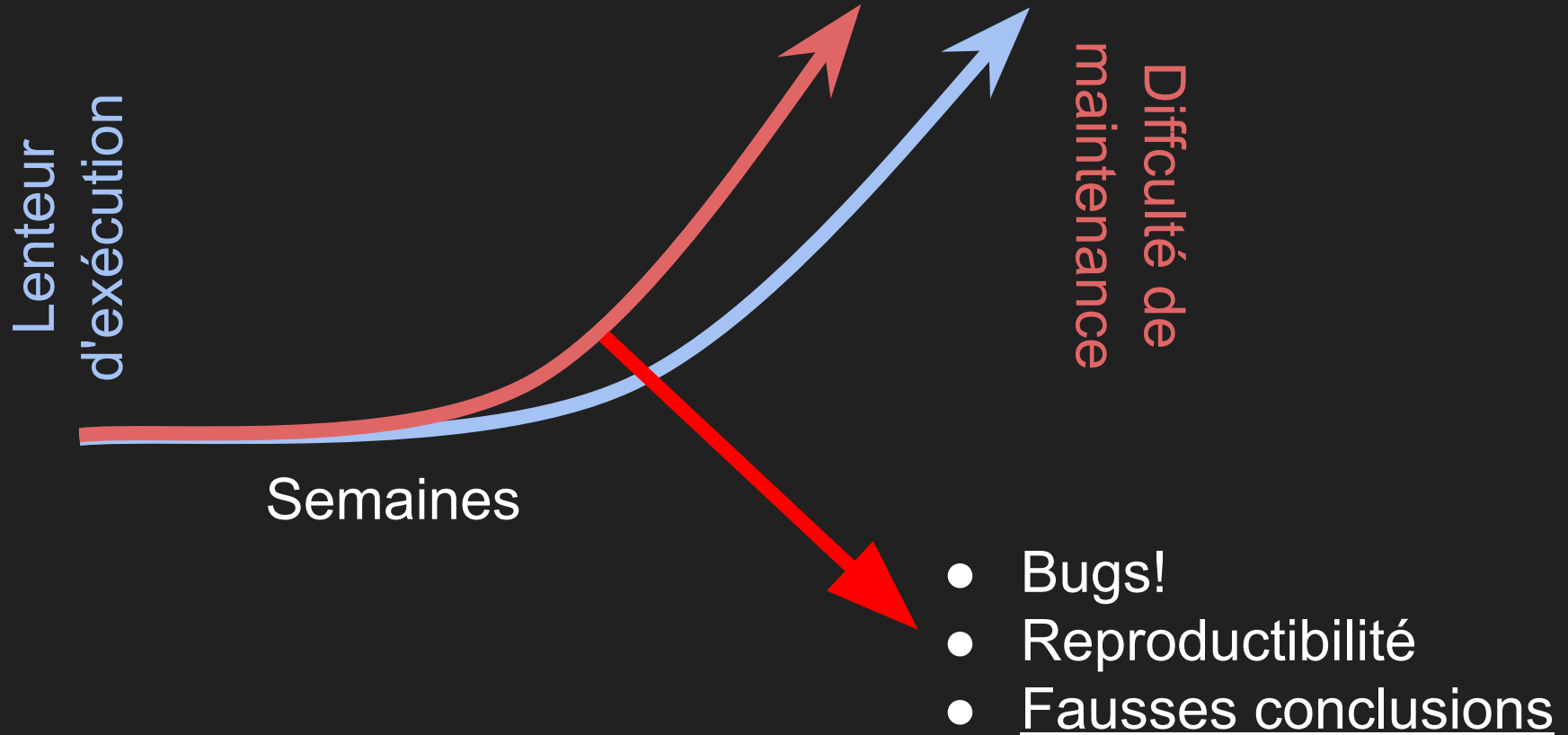


Est-ce que la traduction des protéines influe sur le système immunitaire?

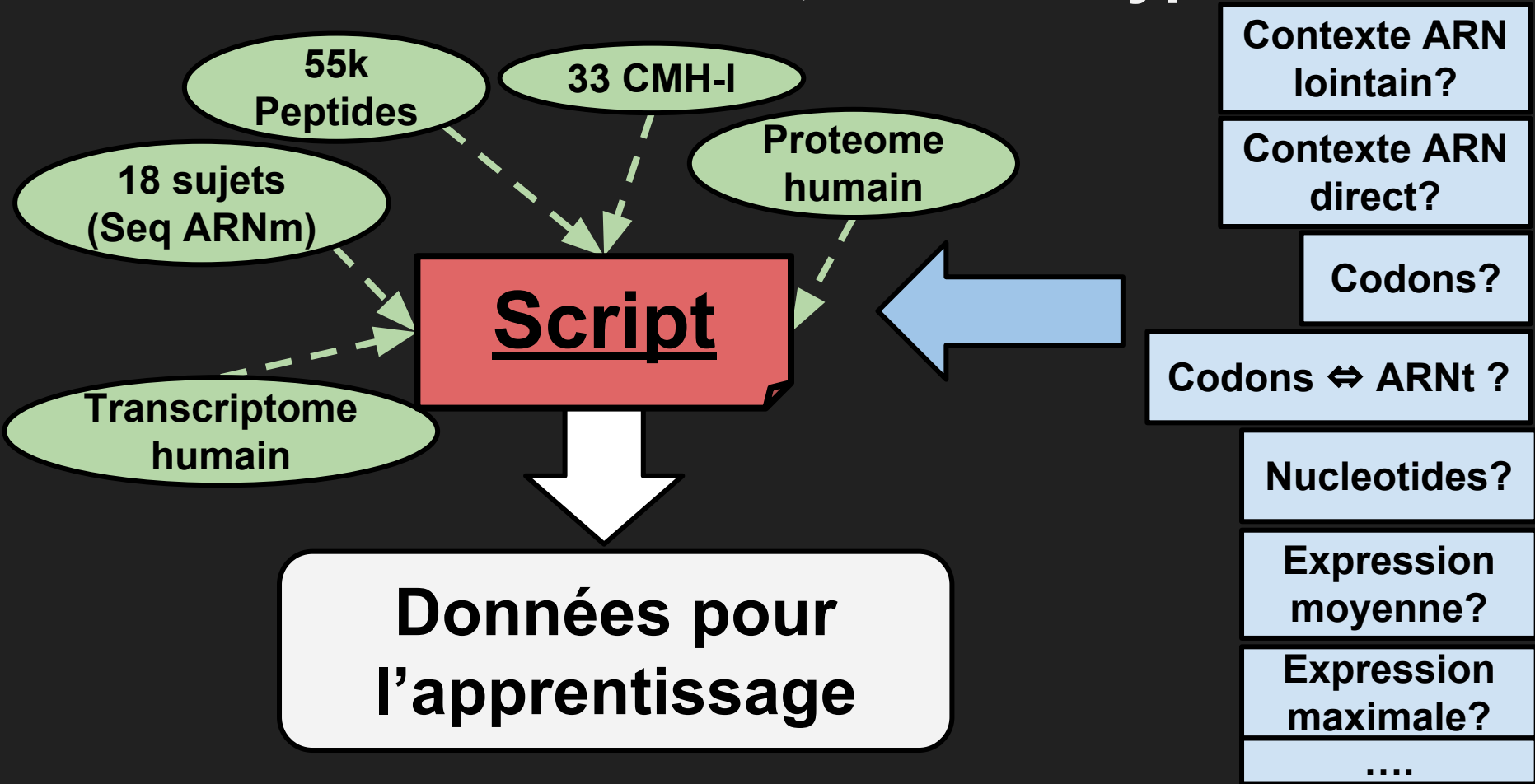
Episode 1 : Approche naive



Limitations



3, 4 ans d'Hypothèses :



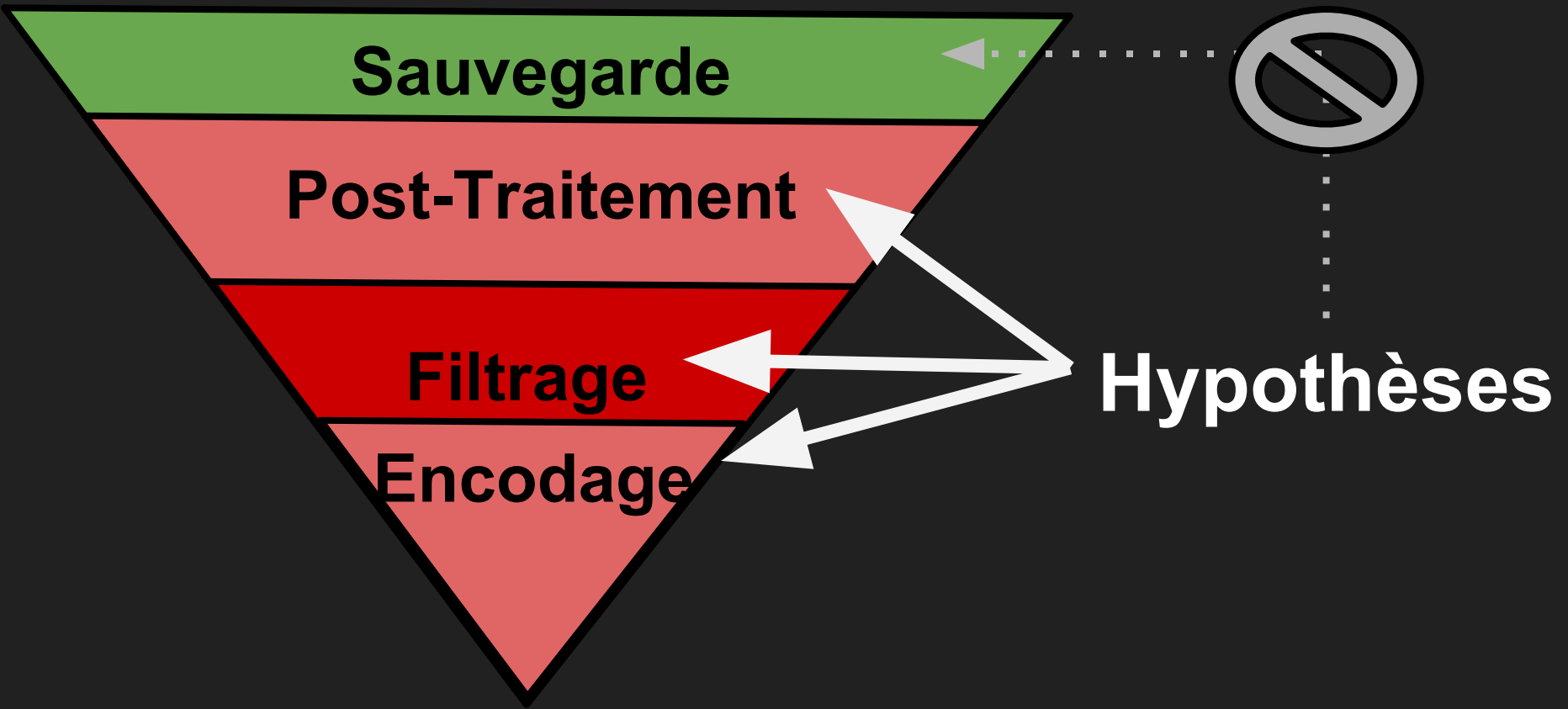
Raisons

Sauvegarde

**Disque
(lent)**

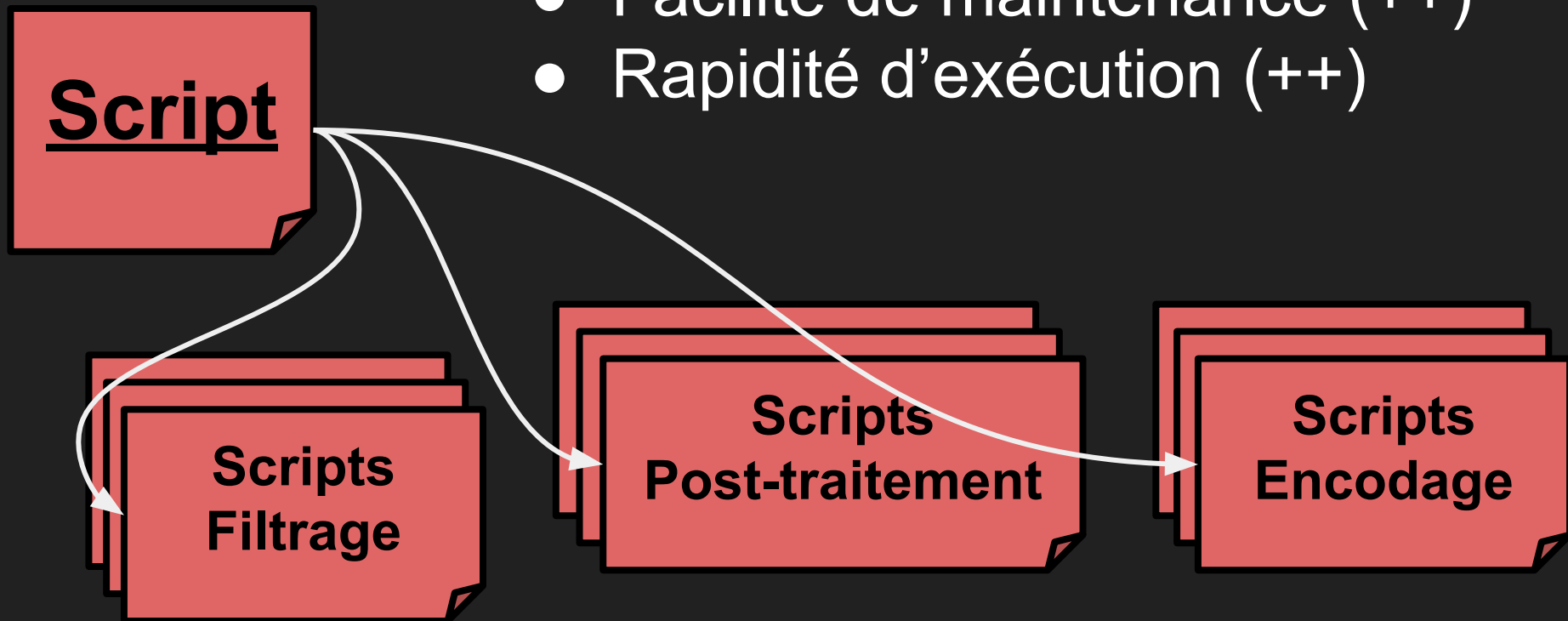
- **Filtrage**
- **Post-traitement**
- **Encodage**

Script

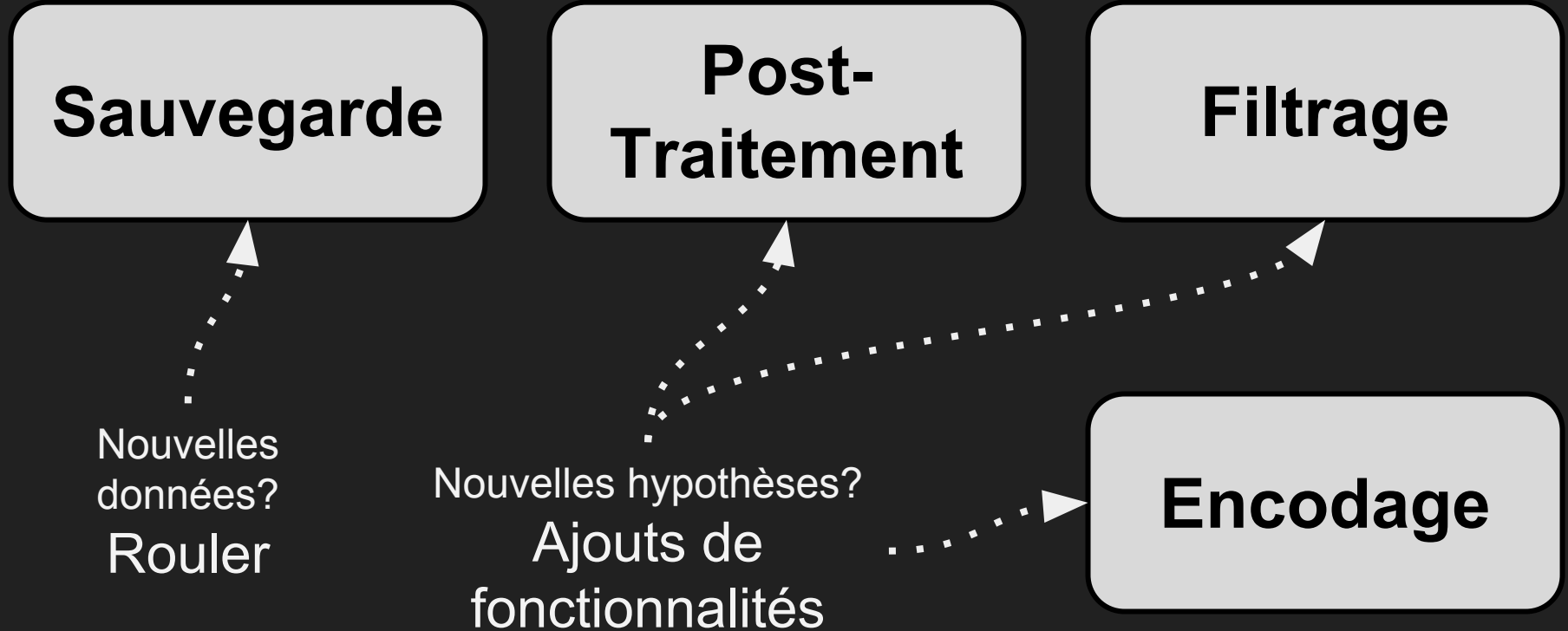


Diviser pour mieux régner

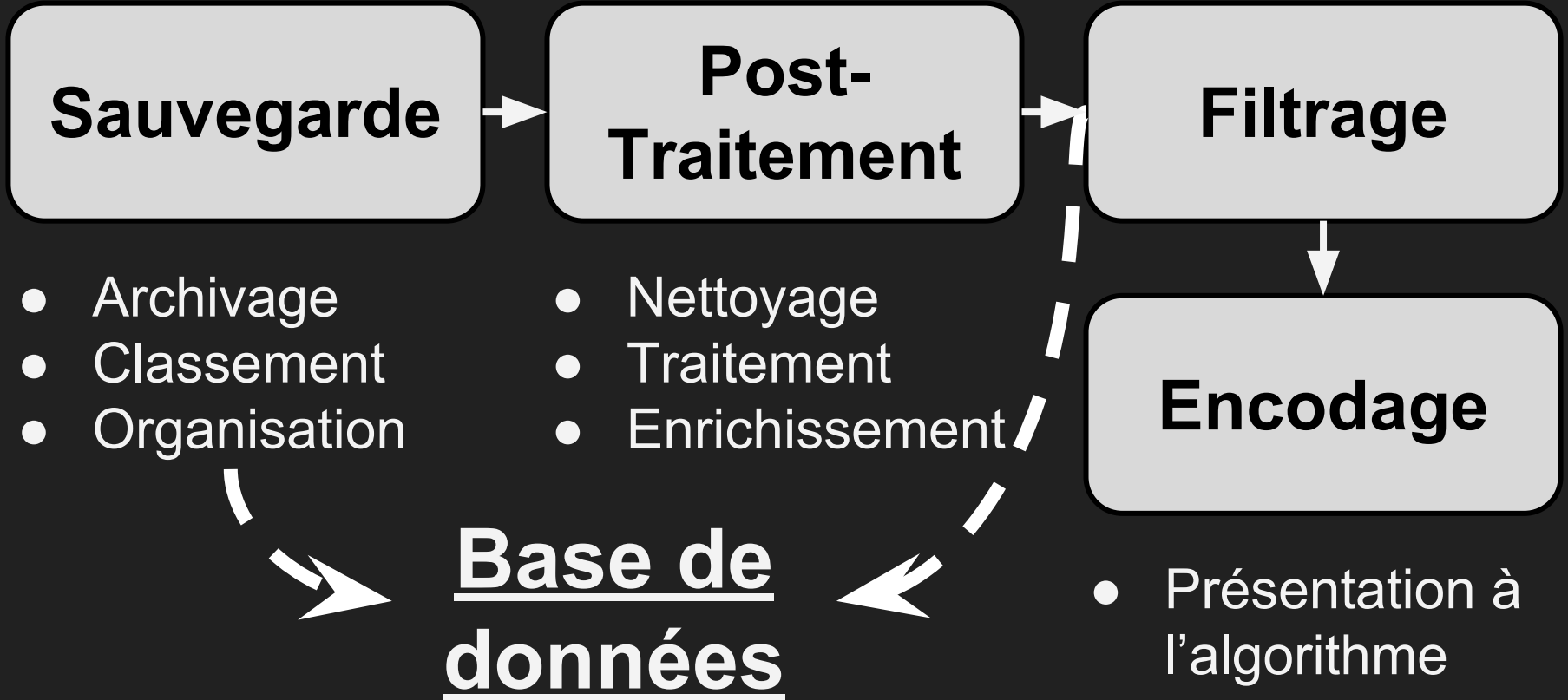
- Facilité de maintenance (++)
- Rapidité d'exécution (++)



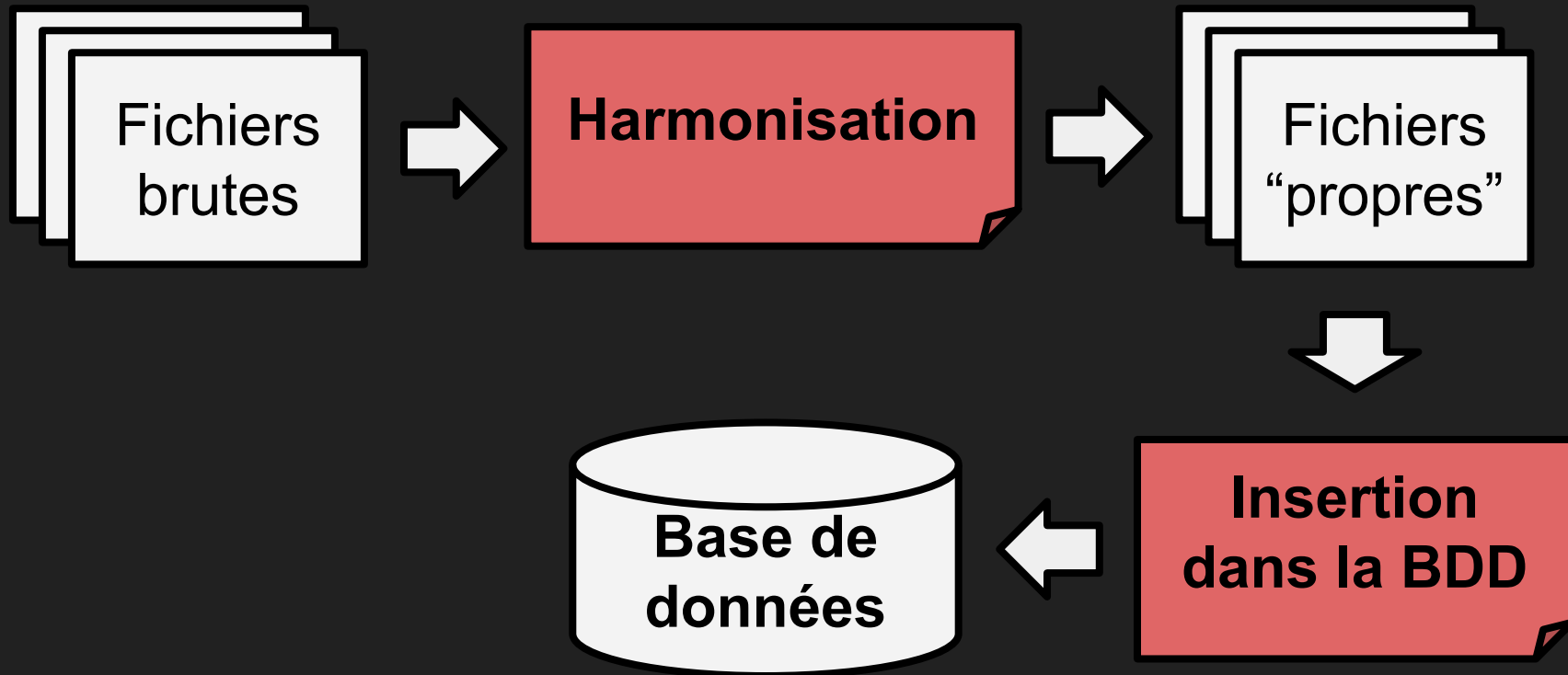
Systeme : S.P.F.E, diviser pour mieux régner



Systeme : S.P.F.E



Sauvegarde



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

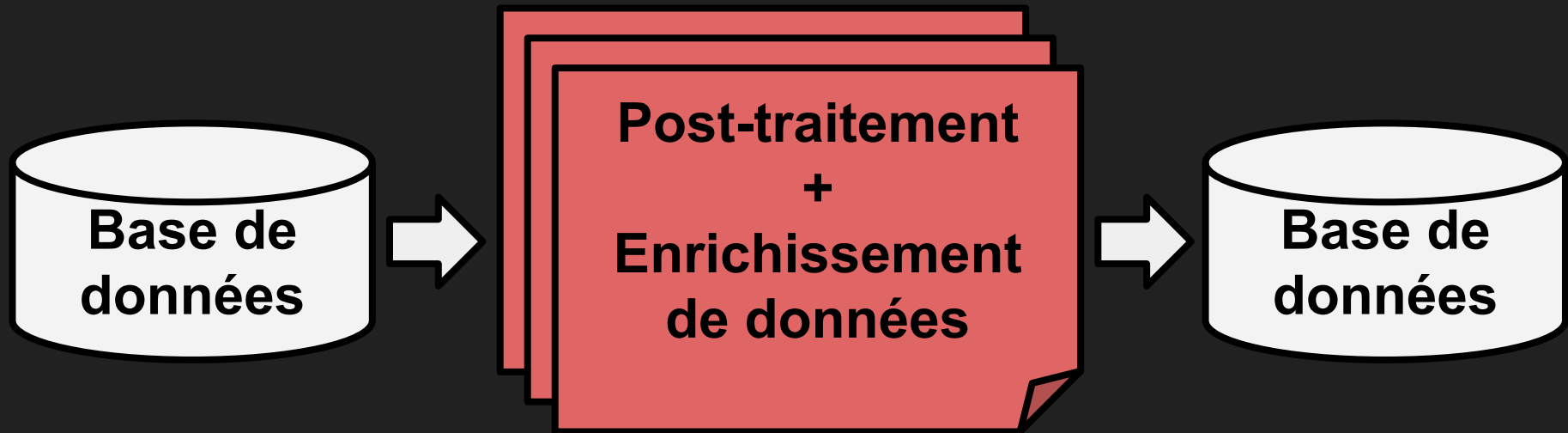
- Insère le contenu complet des fichiers propres dans un système de base de données

**Insertion
dans la BDD**



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

Post-traitement

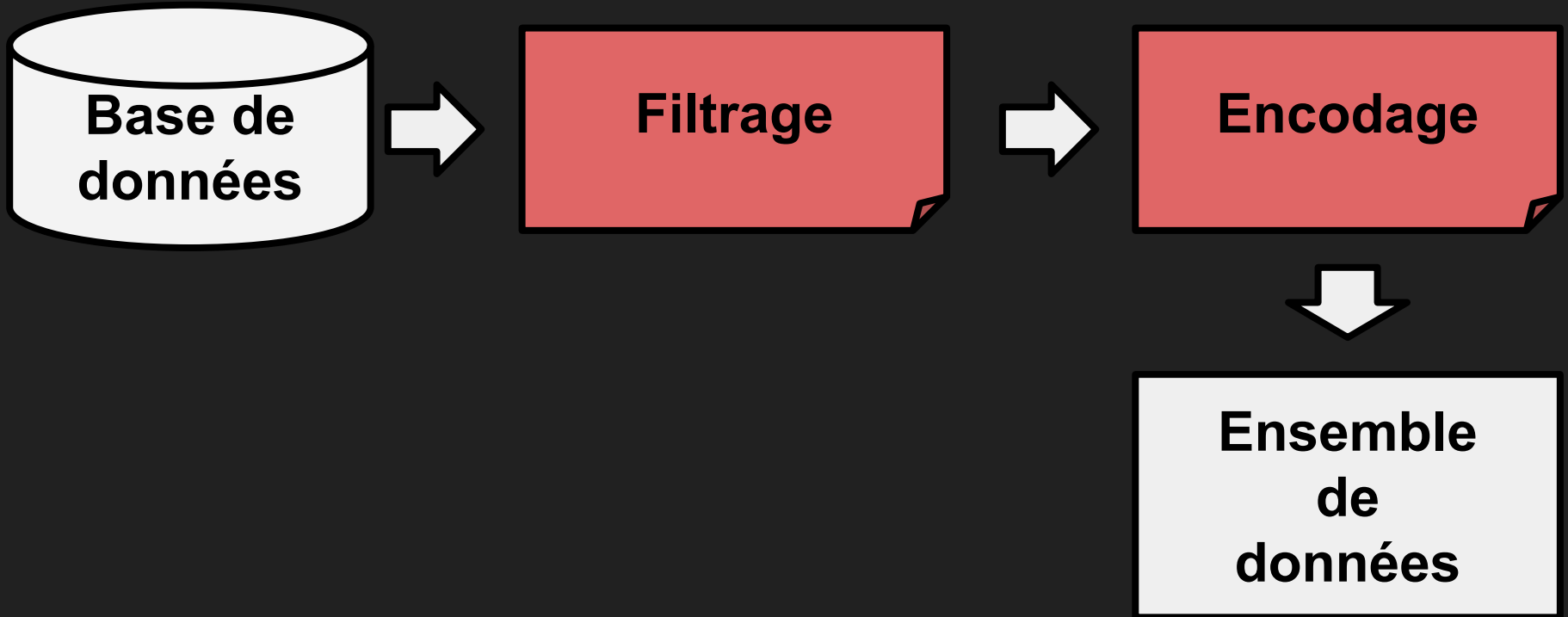


Details : Post-traitement

- 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

Données

Continues

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

Discrètes

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

- Réduire écarts entre valeurs
 - Expressions => $\log(\text{expressions})$

1 Item => 1 entier unique

0	ALL
1	AML
2	CML
...	...

Format standard : tableau numpy (numpy array)

Continues

- Images
- Expressions

float32

Discrètes

- Classes
- Catégories
- Mots

int16

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

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

Détails : Ensemble de données (fichier)

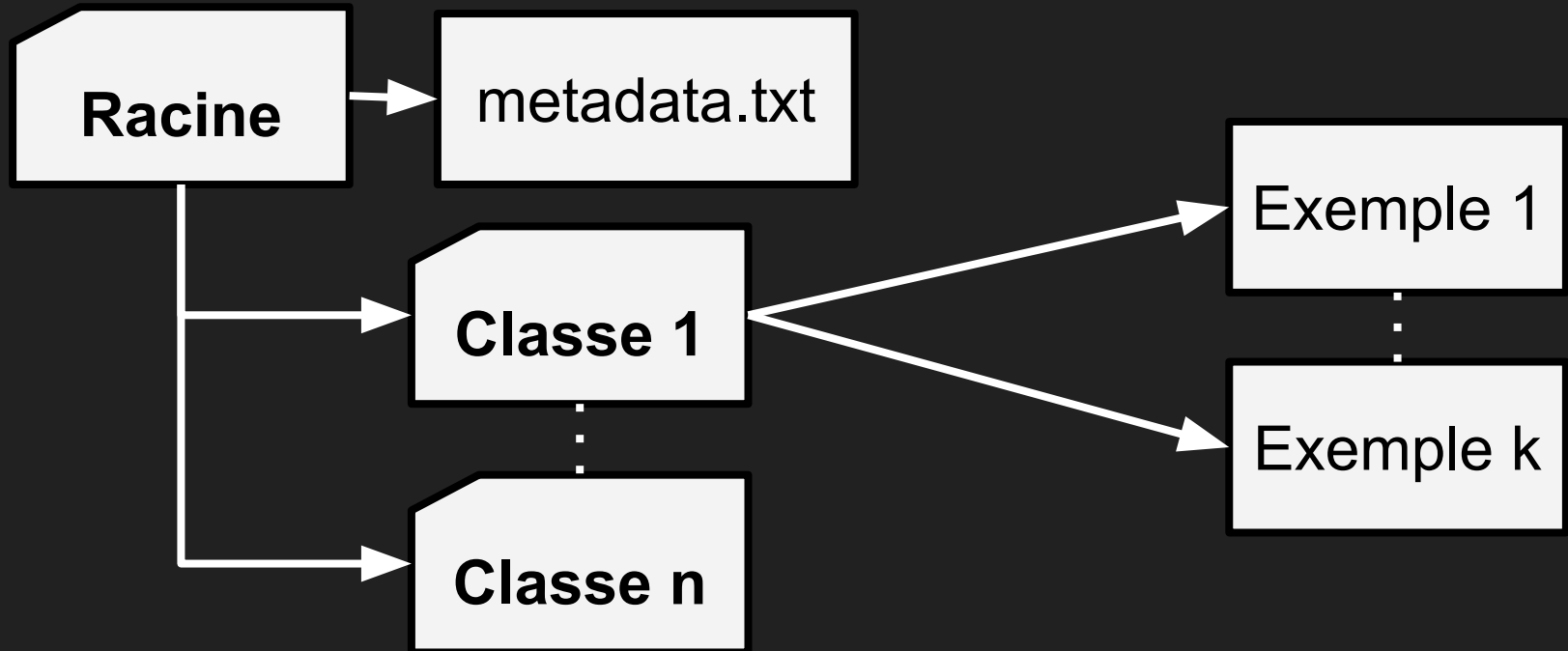
Ensemble de données

- 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

Données

Expressions

Type de leucémie

Vecteur

0.2	0.21	0.45	0.01	0.6	...
-----	------	------	------	-----	-----

0.1	0.02	0.42	0.07	0.4	...
-----	------	------	------	-----	-----

0.26	0.13	0.56	0.5	0.61	...
------	------	------	-----	------	-----

Matrice

Vecteur

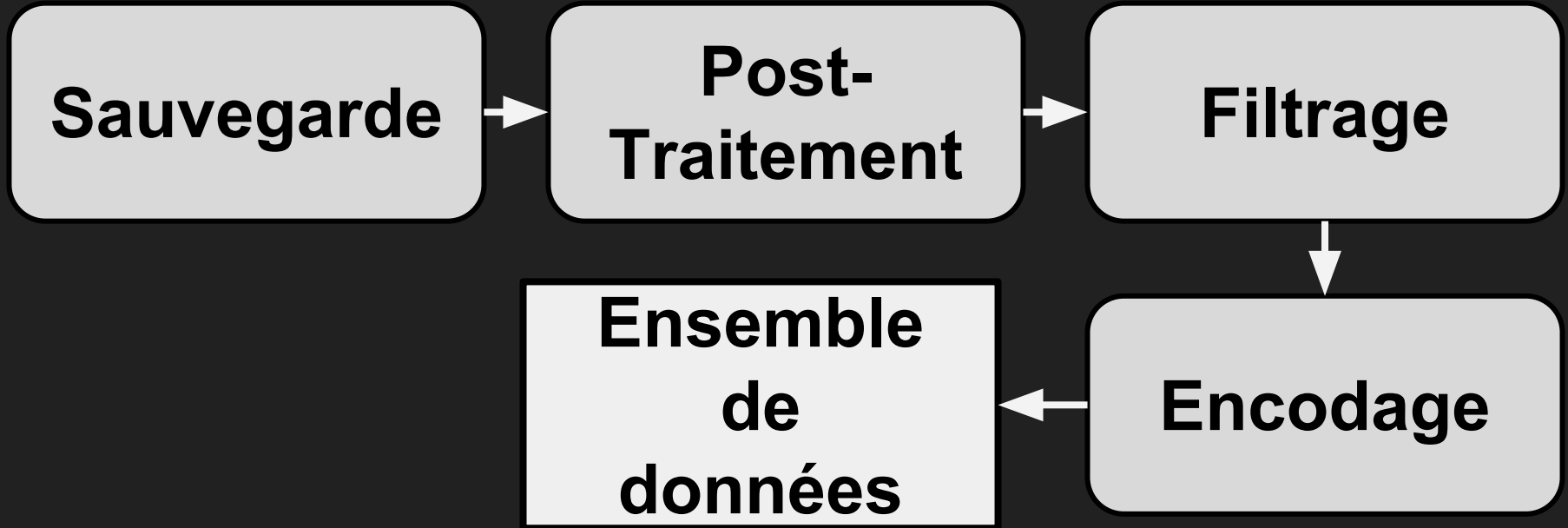
1

0

2

Matrice

Systeme : S.P.F.E, conclusions



- Maintenance plus simple
- Plus de confiance : Reproductibilité
- 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

The screenshot displays the ArangoDB web interface. On the left is a navigation sidebar with options like DASHBOARD, COLLECTIONS, QUERIES, GRAPHS, SERVICES, USERS, DATABASES, LOGS, SUPPORT, and HELP US. The main area shows a query editor with the following code:

```
14
15 for u in packages
16   filter u.Architecture == 'amd64'
17   return u
```

Below the editor, there are buttons for "Remove all results", "Explain", and "Execute". The "Execute" button has been clicked, resulting in a query execution summary: "Query" with "23987 elements" and a duration of "0.275 s". A "JSON" button is visible on the right.

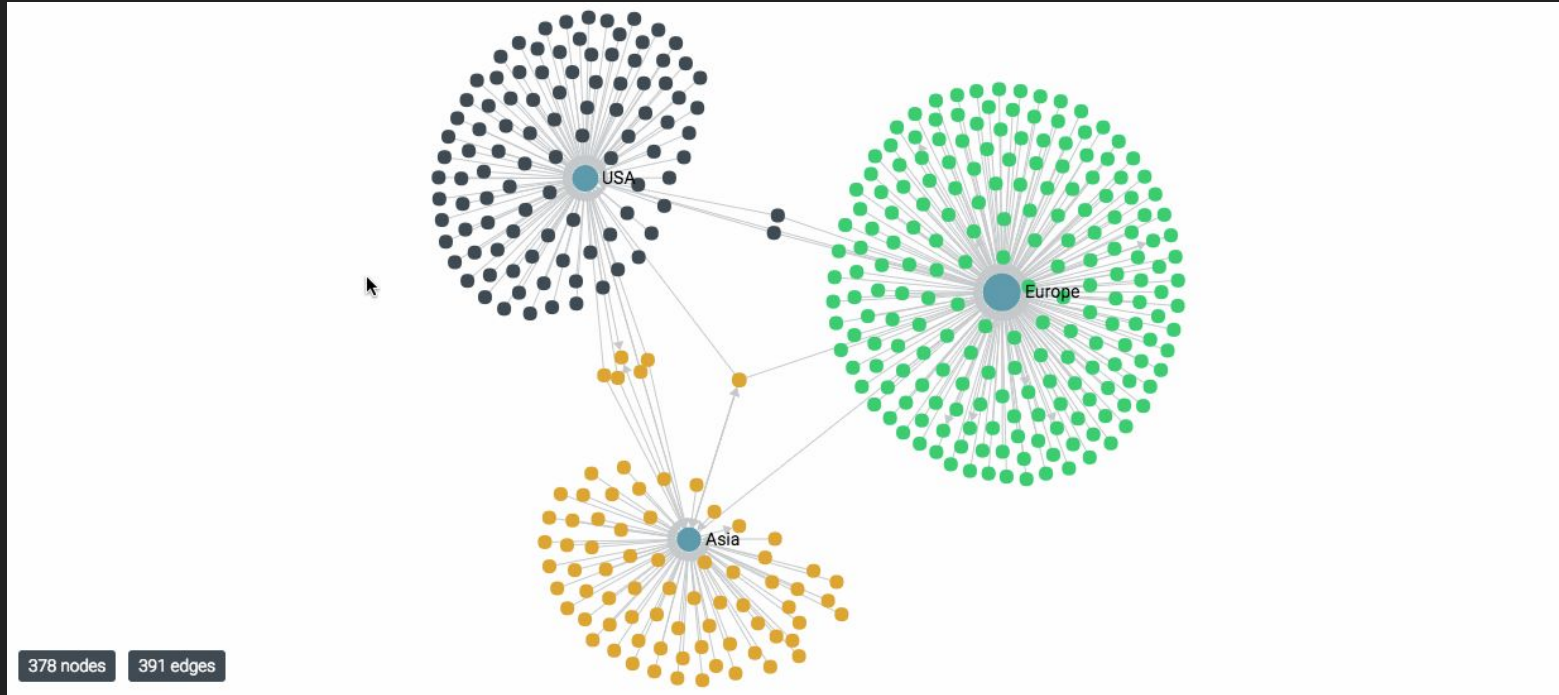
A "Profiling information" window is open, showing a list of execution steps with their durations and descriptions:

Step	Duration	Operation	Description
A	0.002 ms	initializing	startup time for query engine
B	0.035 ms	parsing	query parsing
C	0.005 ms	optimizing ast	abstract syntax tree optimizations
D	0.004 ms	loading collections	loading collections
E	0.016 ms	instantiating plan	instanciation of initial execution plan
F	0.076 ms	optimizing plan	execution plan optimization and permutation
G	274.995 ms	executing	query execution

Below the table is a horizontal bar chart where the longest bar, labeled 'G', is highlighted in green. The chart also shows a legend for 'ABCDEFG'.

At the bottom of the interface, there are "Download" and "Copy To Editor" buttons. The status bar at the bottom left shows "3.1.DEV" and social media icons.

Interface graphique : Visualisation de graphes

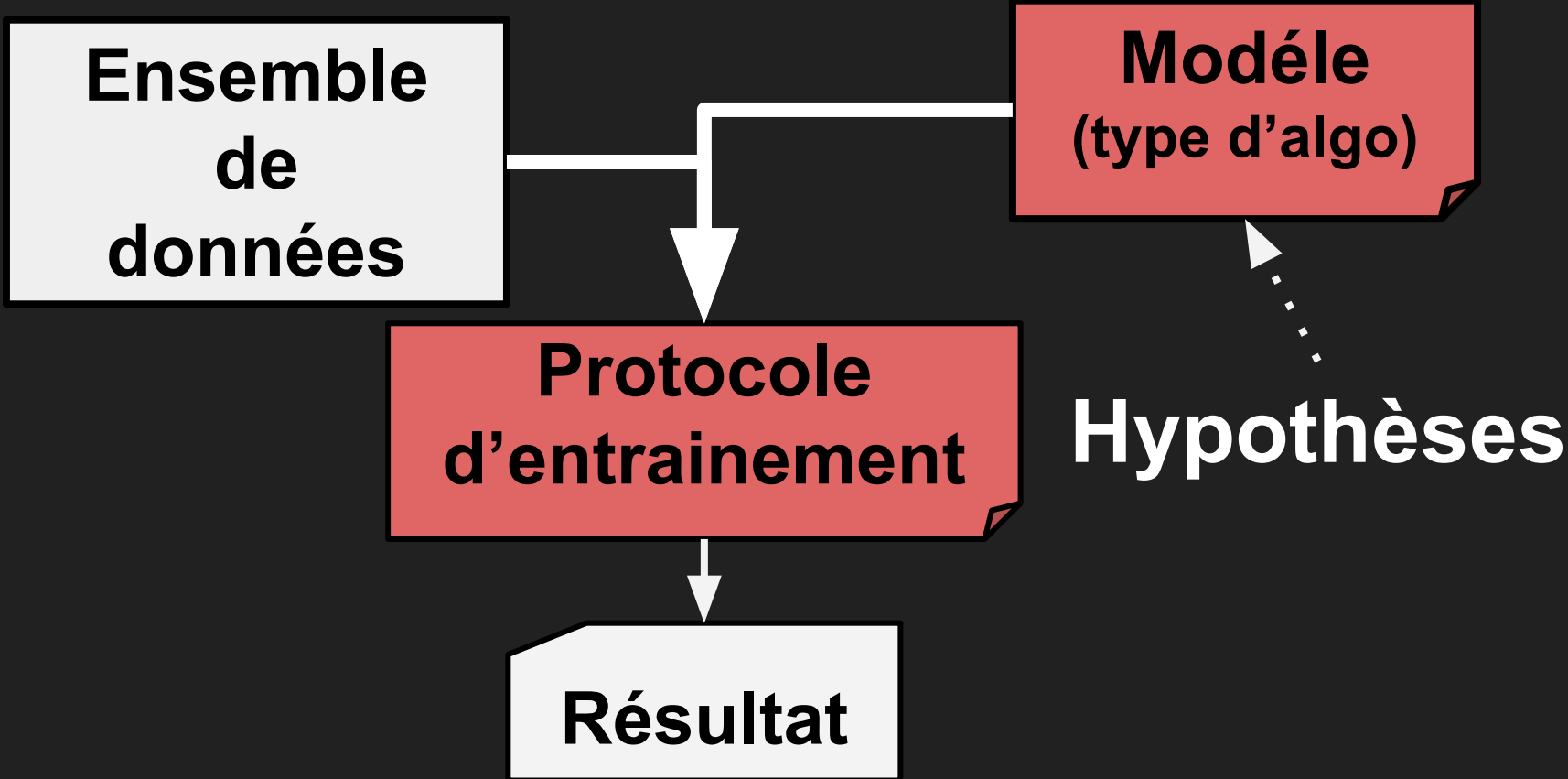


Images et gros fichiers

- 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

Gestion des expériences

Données Modèle Protocole Résultats



Détails

Modèle

- 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)

Résultat

Dossier unique

**Meilleur modèle
pour l'expérience**

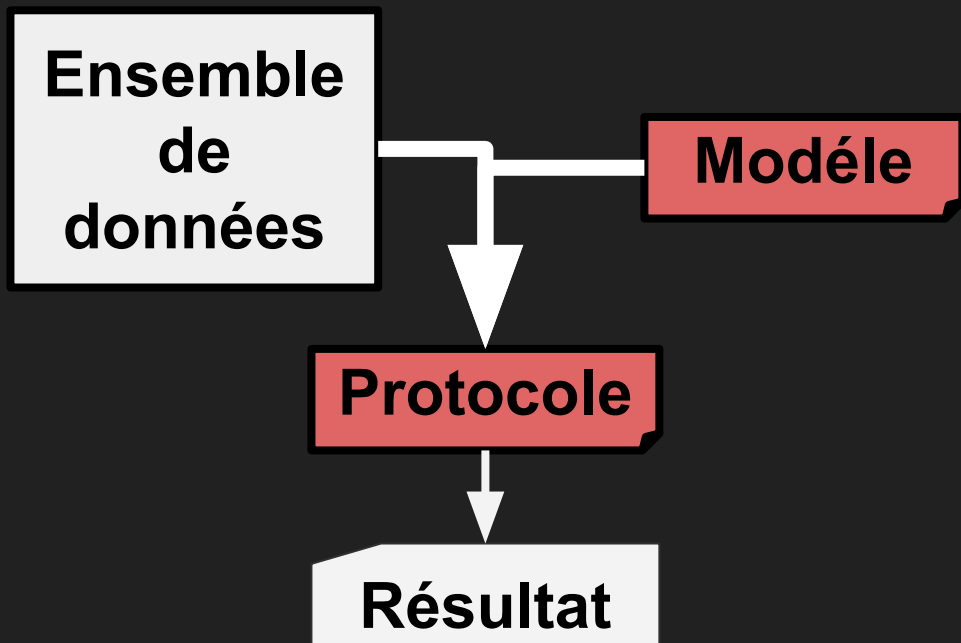
**Historique
d'apprentissage**

**Courbes
d'apprentissage**

Metadata

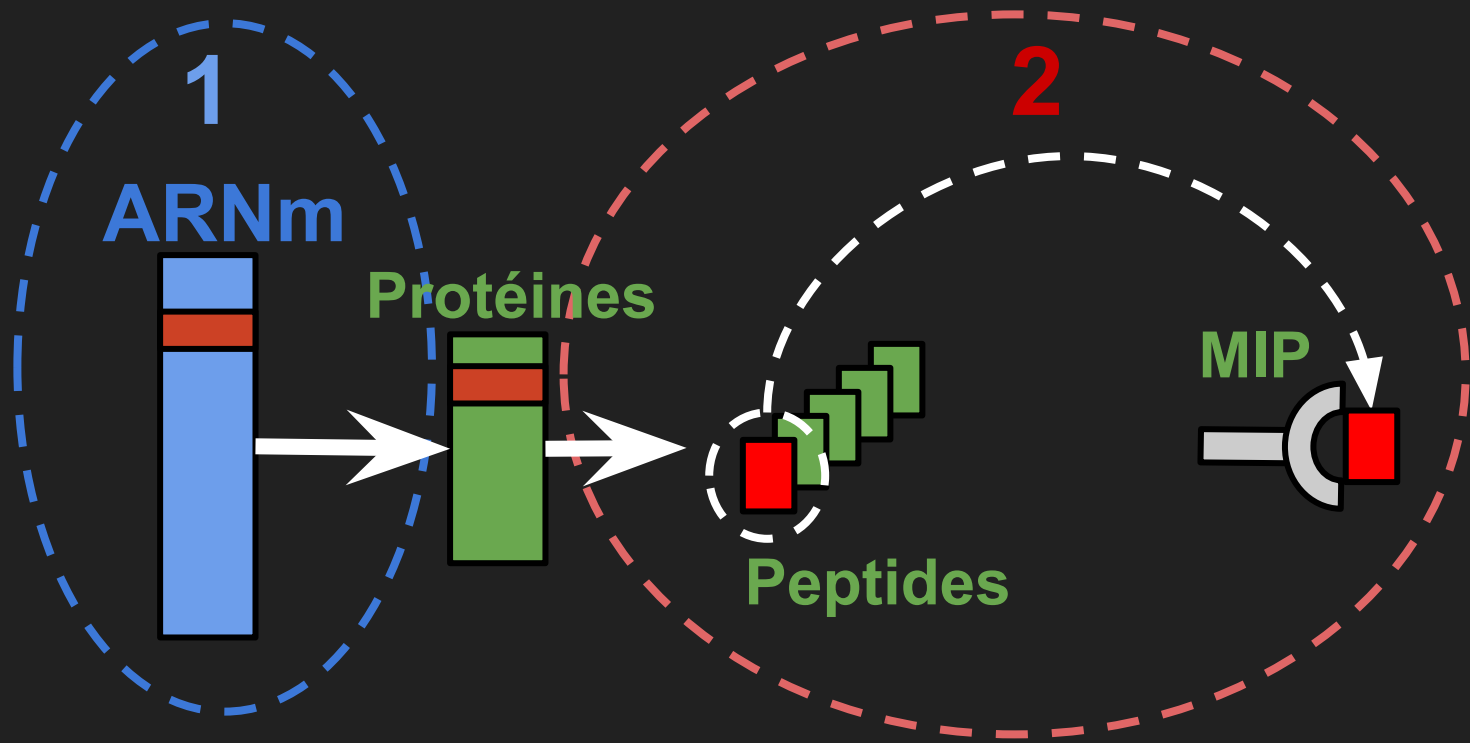
- Architecture du modèle
- Date d'entraînement
- Protocole
 - Hyper-paramètres
 - Minibatch
 - Nb époques
- Nom : Ensemble de données

Conclusions



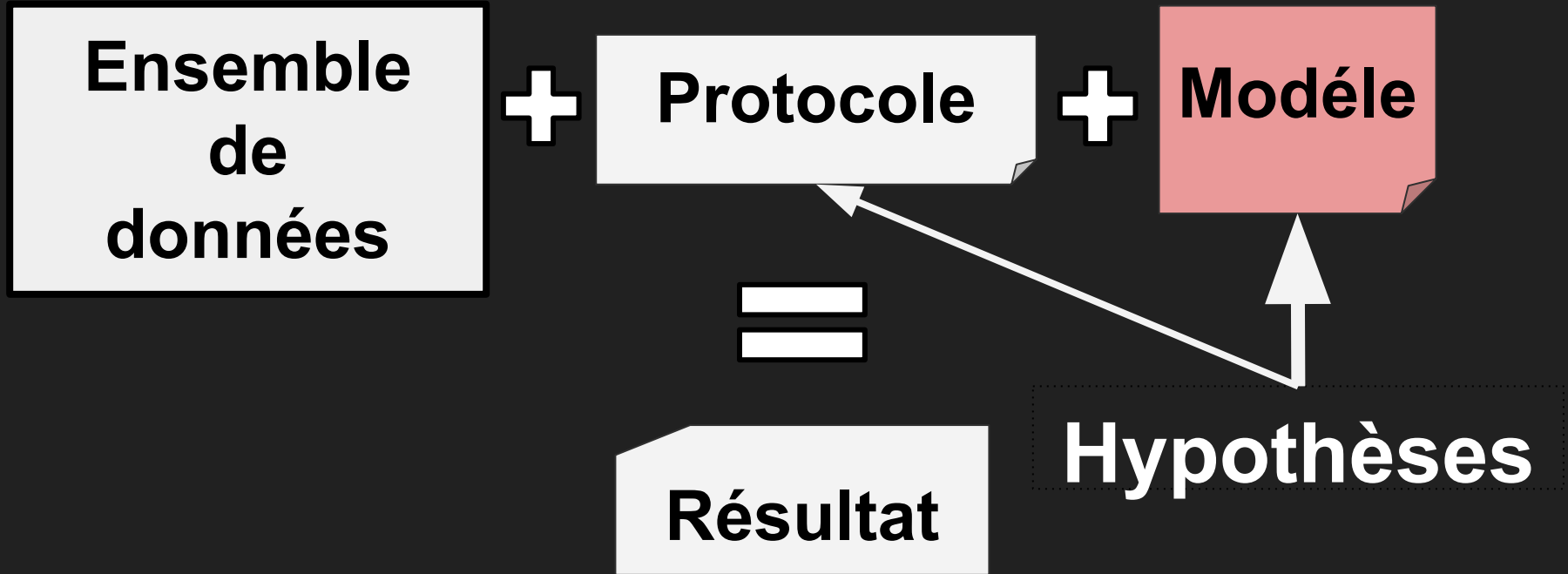
- Facilité de maintenance
 - Moins de Copier-Coller
- Meilleure reproductibilité
 - Plusieurs modèles
 - Même protocole
- Meilleure traçabilité
- Format de résultat standardisé

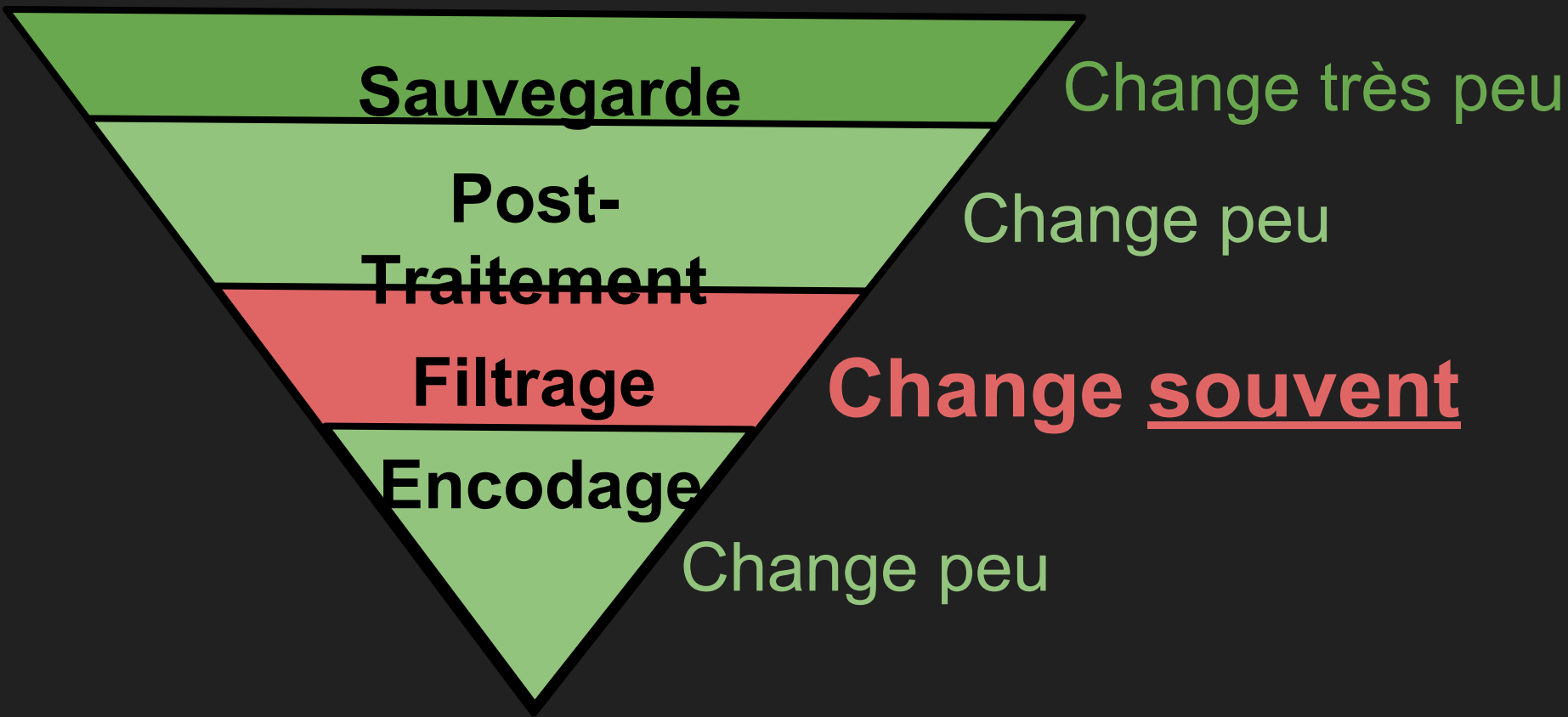
.FIN.



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

Experience (Quel algo d'apprentissage ?)





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

Limitations

