CNRS - INP - UT3 - UT1 - UT2J

Institut de Recherche en Informatique de Toulouse

Système de recommandation pour l'aide à l'élaboration de processus d'analyse

Gabriel Ferrettini

Encadrants : Chantal Soulé-Dupuy and Julien Aligon

Université de Toulouse Capitole, Institut de Recherche en Informatique de Toulouse.

I. Introduction

II. Recommandation

III. Explication

Contexte et problématique

- Le machine learning connaît un intérêt croissant dans de très nombreux domaines

- Problème : une utilisation efficace nécessite une certaine expertise
  - Employer un expert coûte cher
  - Se former prends du temps

Evolution de la recherche des termes « machine learning » sur google au cours du temps

Des outils insuffisants?

- De nombreux outils existent déjà :
  - Weka
  - Knime
  - RapidMiner
  - Orange
  - What if?

- Visent d'avantage à aider des experts qu'à guider des débutants.

Quelques définitions

- Machine learning/Apprentissage automatique :
  - un champ d'étude de l'***intelligence artificielle*** qui vise à donner aux ordinateurs la capacité d' « apprendre » à partir de données, c'est-à-dire devenir capable de résoudre des tâches sans être explicitement programmés pour chacune.

- Dataset/Jeu de données :
  - Ensemble de données présentées sous la forme d'une ***matrice***, organisée en ***instances***, caractérisées par des ***attributs***, Chaque instance est ainsi caractérisée par un vecteur d'attributs.

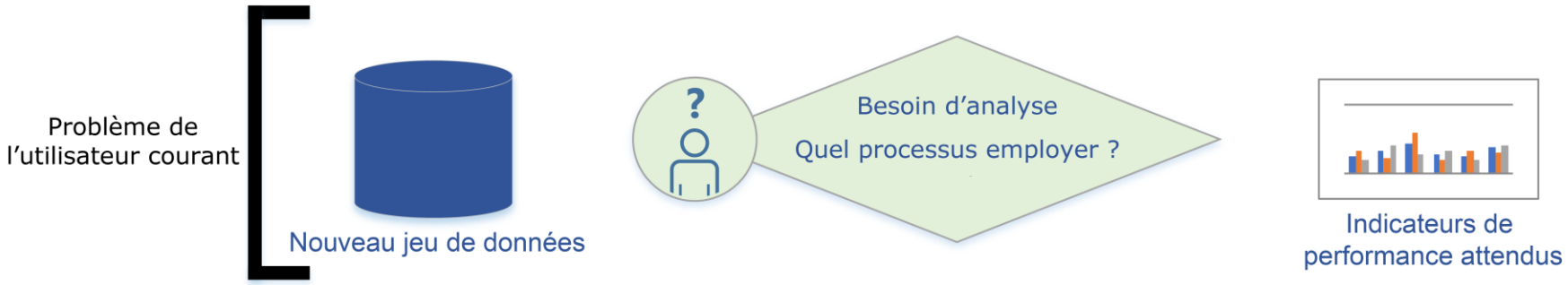
- Workflow/Chaîne de traitement :
  - Ensemble d'opérations d'analyse de données résultant en un modèle d'apprentissage entraîné. Ces opérations peuvent par exemple être un prétraitement des données, suivie d'une optimisation des paramètres d'entraînement d'un modèle puis de l'entraînement du modèle lui-même.

I. Introduction

II. Recommandation

III. Explication

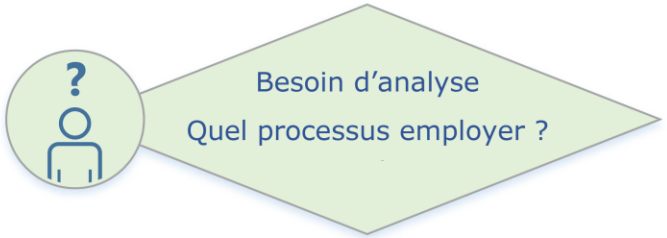
Recommandation de chaîne de traitement

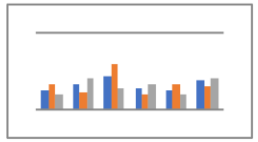


Recommandation de chaîne de traitement

Problème de l'utilisateur courant

Nouveau jeu de données

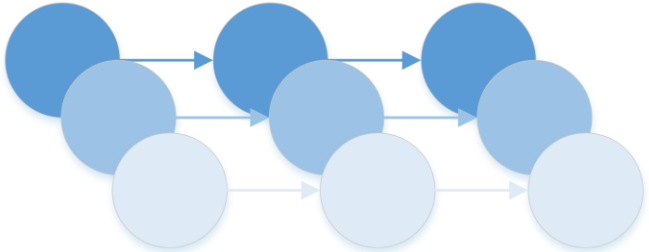




Indicateurs de performance attendus

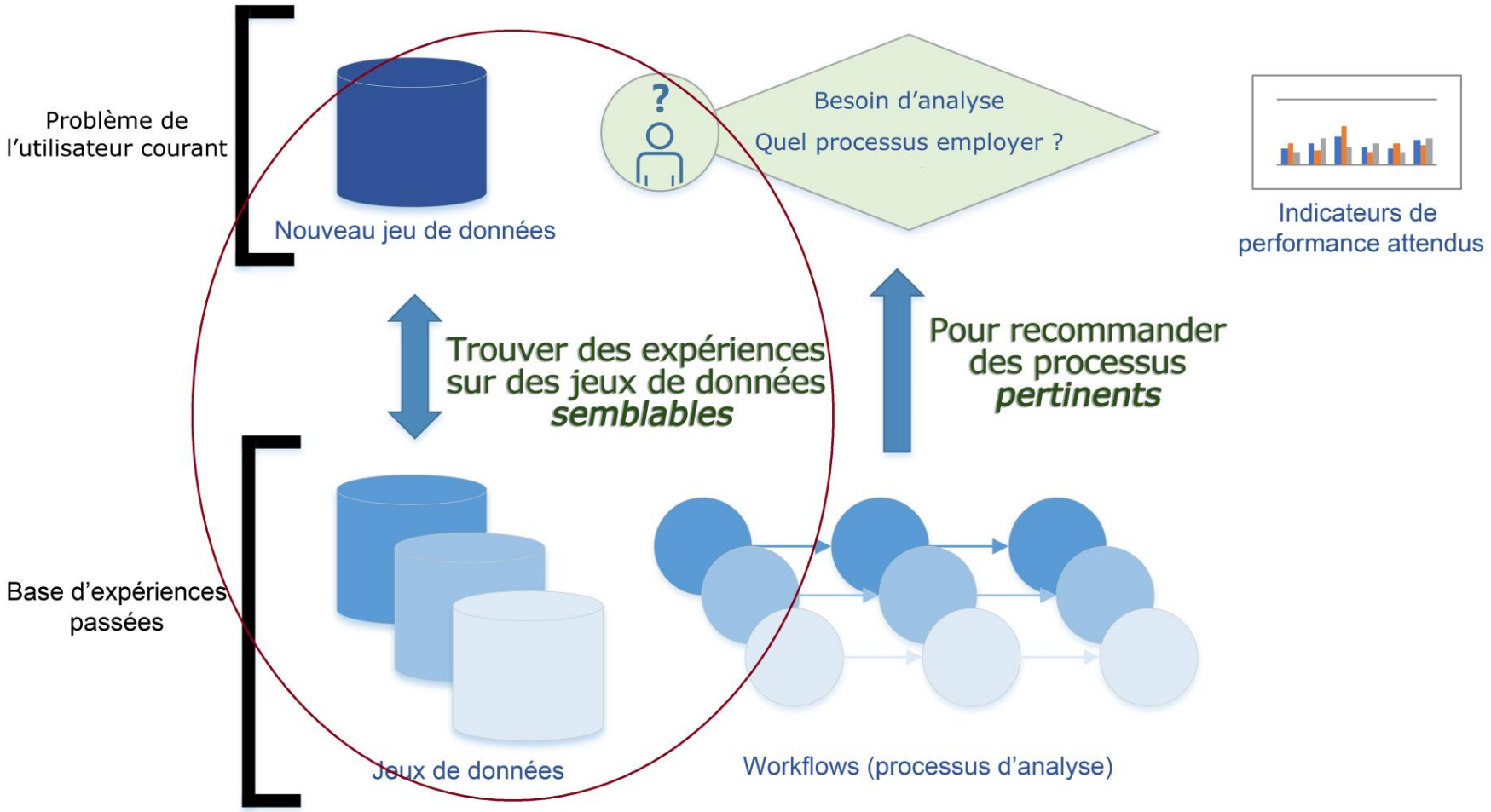
Base d'expériences passées

Jeux de données

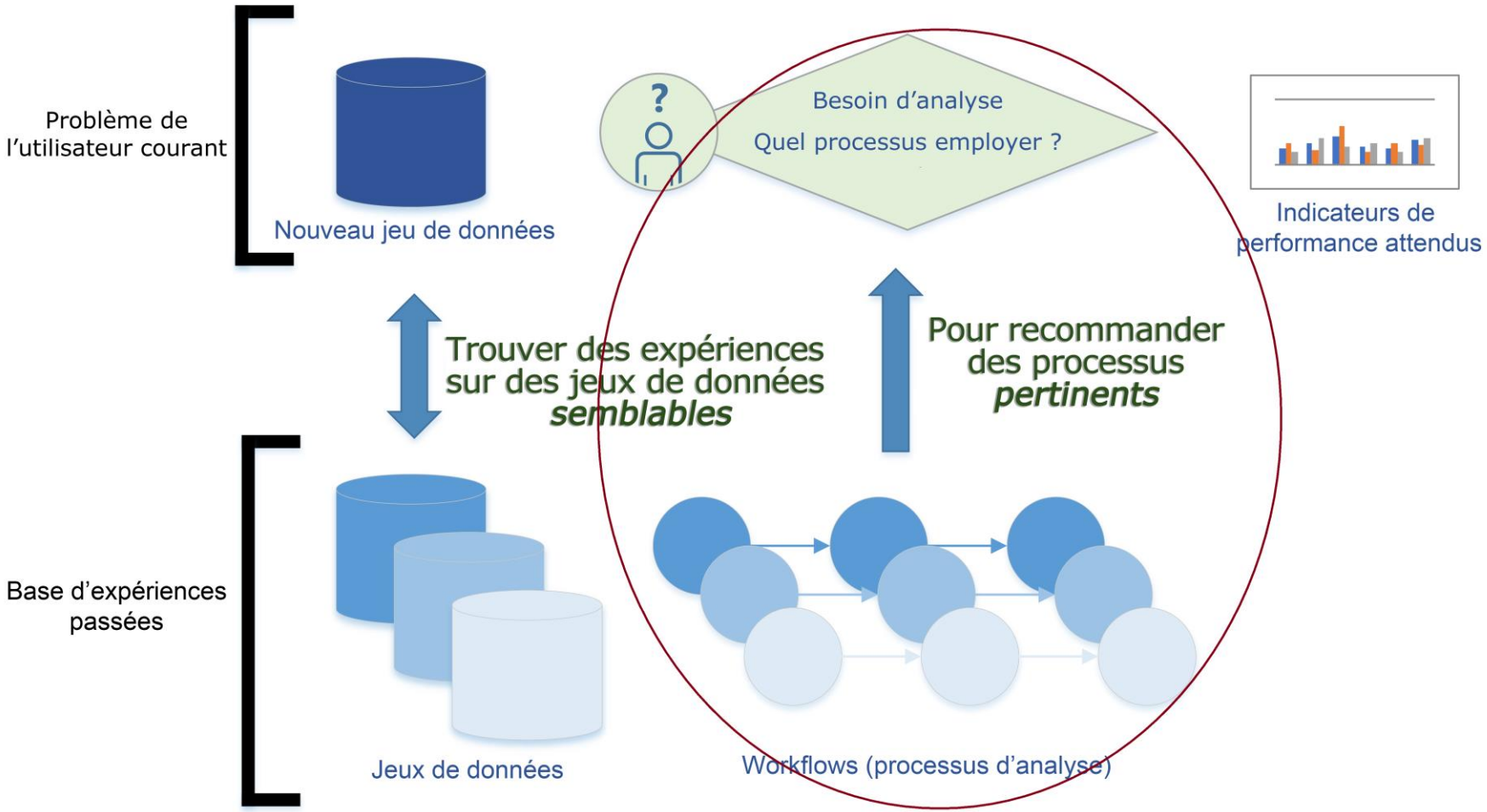


Workflows (processus d'analyse)

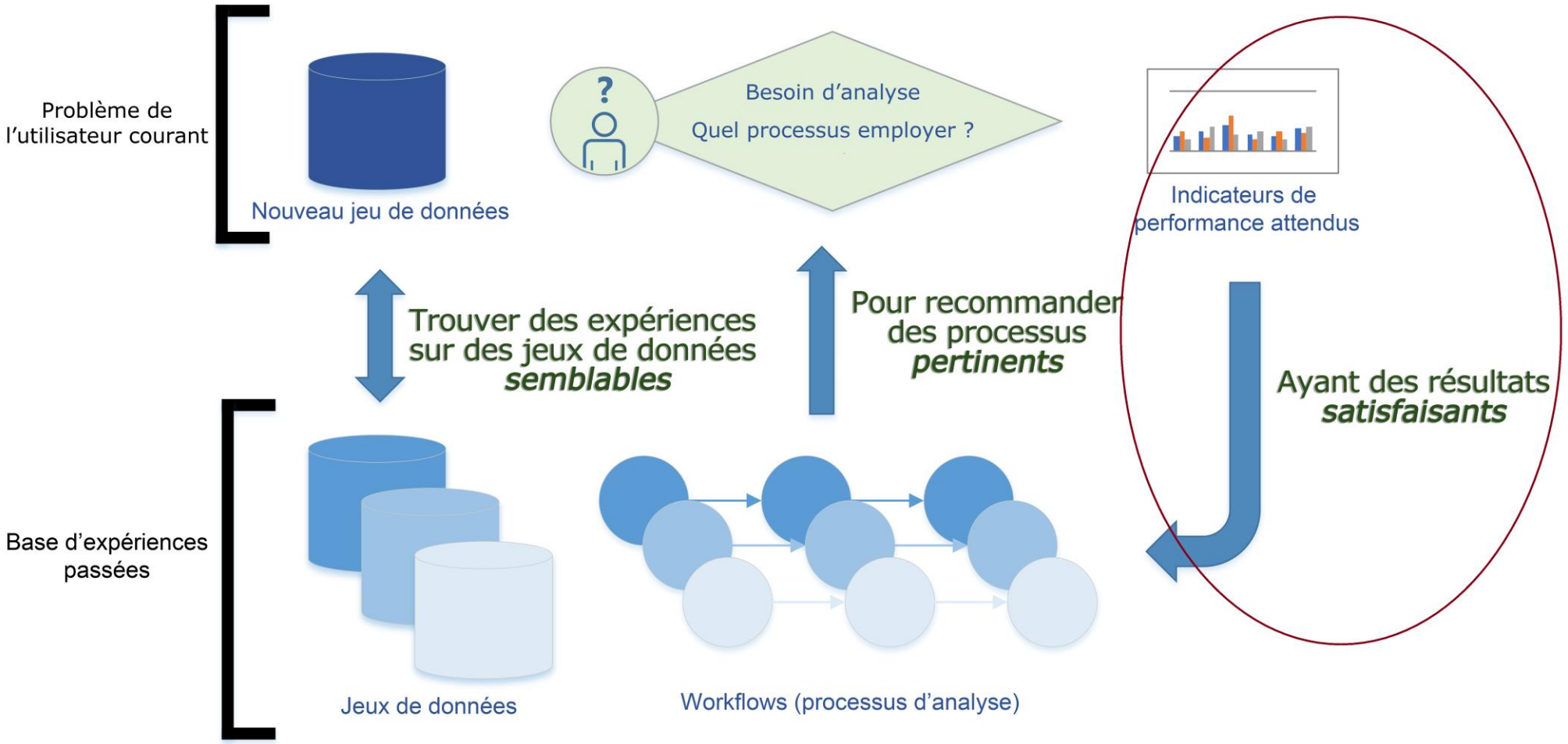
Recommandation de chaîne de traitement



Recommandation de chaîne de traitement

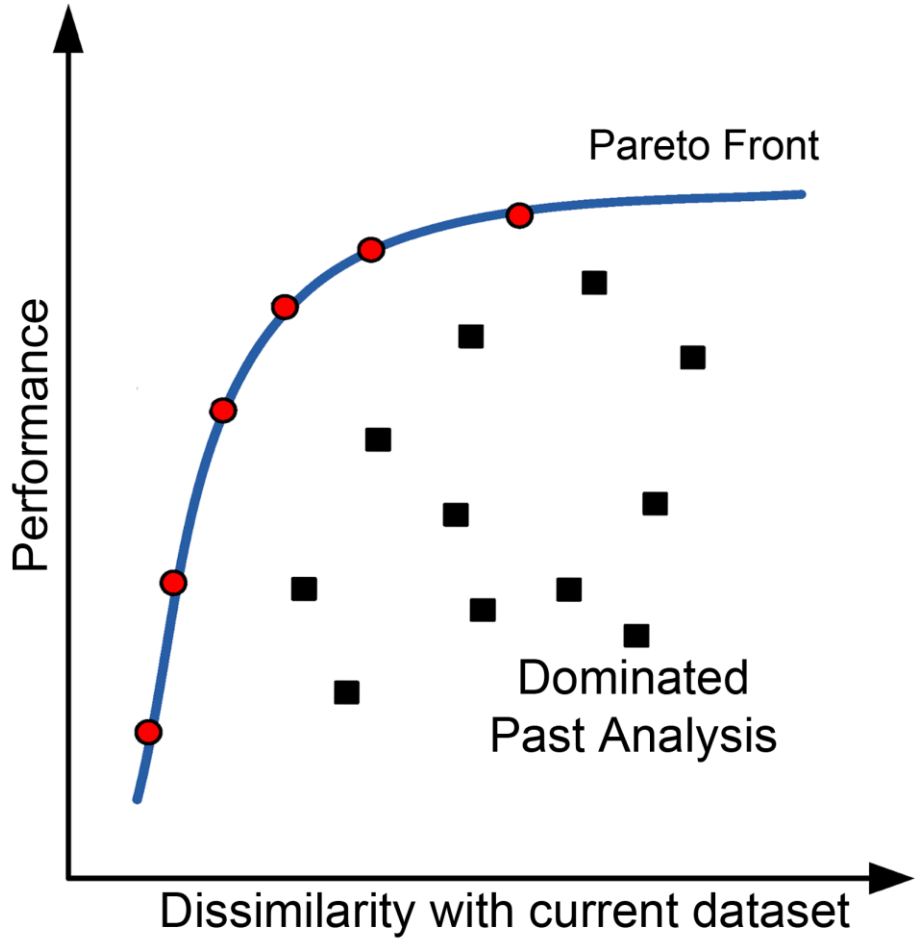


Recommandation de chaîne de traitement



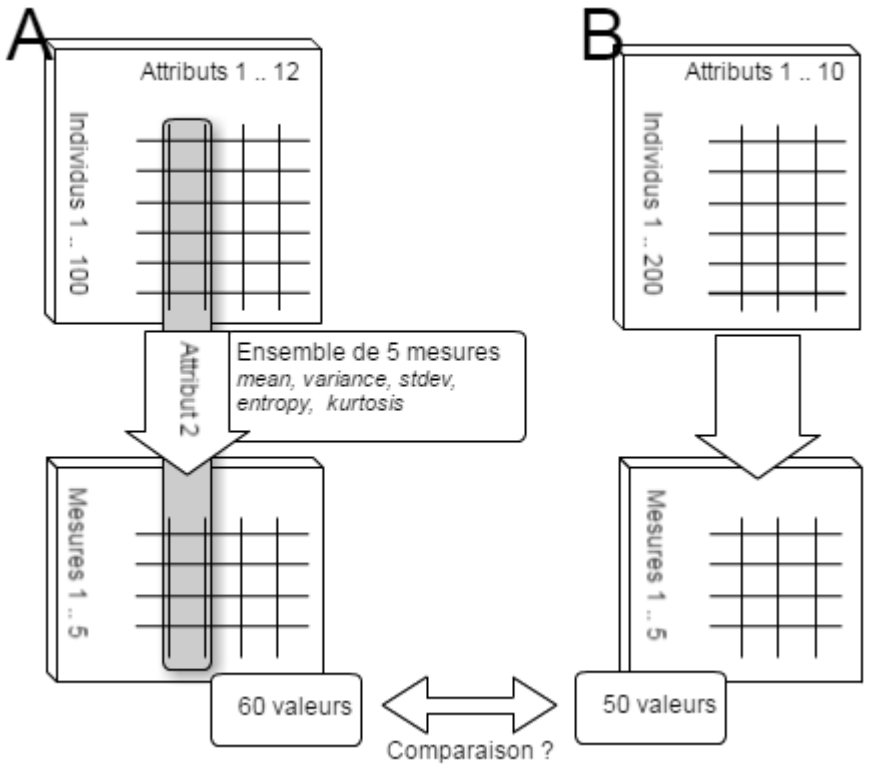
Recommandation de chaîne de traitement

- Sélection sur deux critères distincts



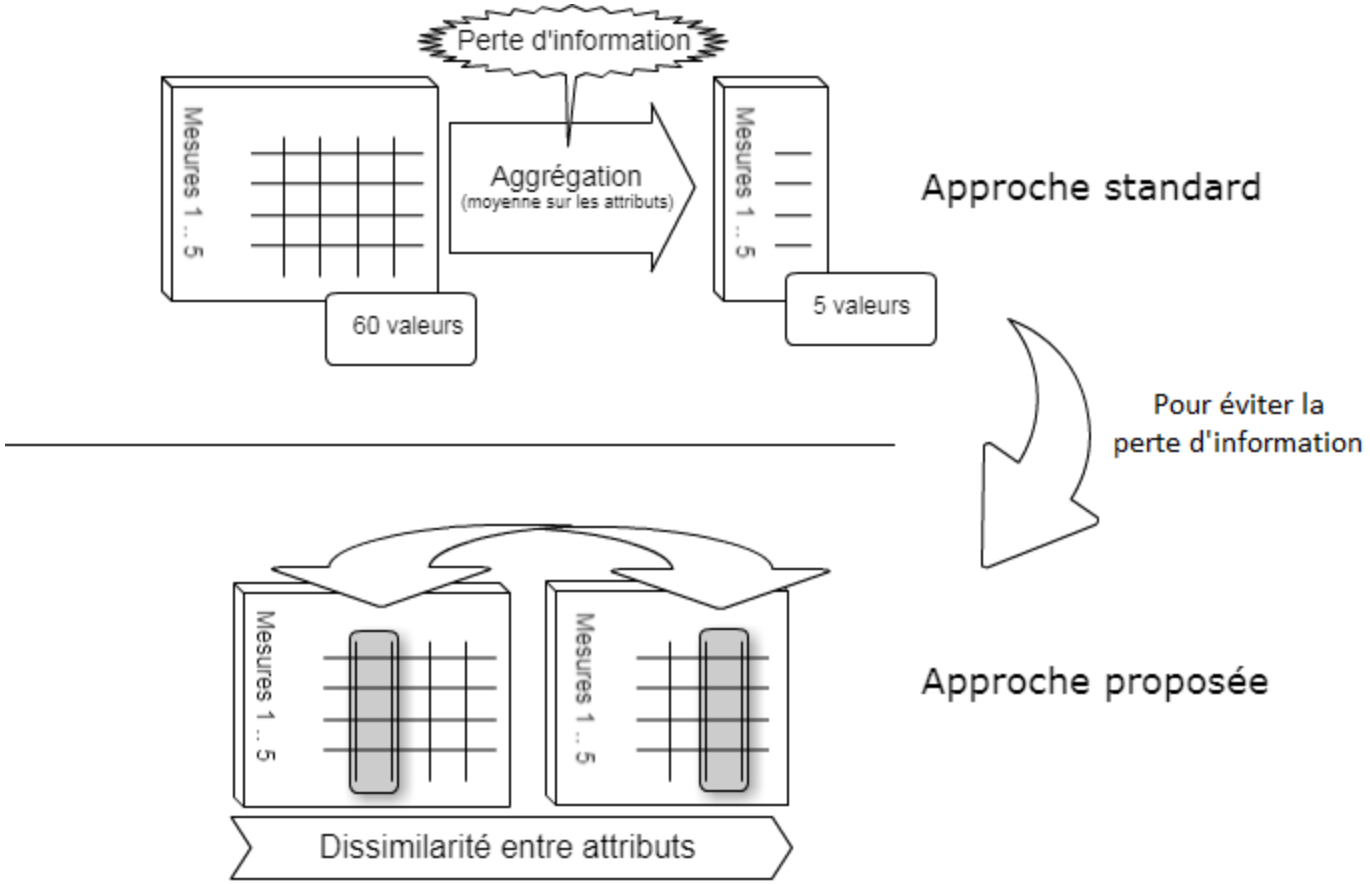
Dissimilarité entre datasets

- Comment comparer deux datasets?



Dissimilarité entre datasets

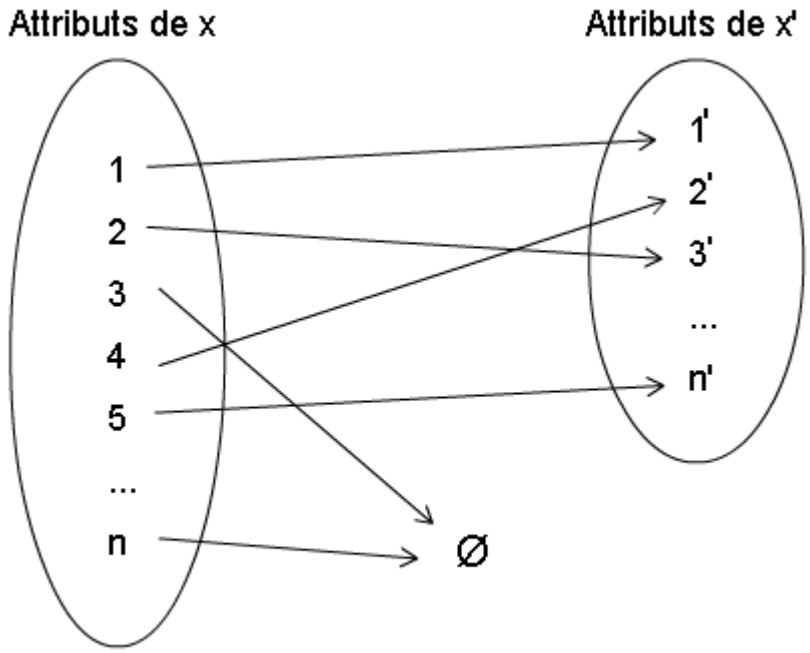
- Comment comparer deux datasets?



Au final un problème d'optimisation

- Appairage des attributs de manière à minimiser la dissimilarité totale

- Somme des dissimilarités = dissimilarité totale



I. Introduction

II. Recommandation

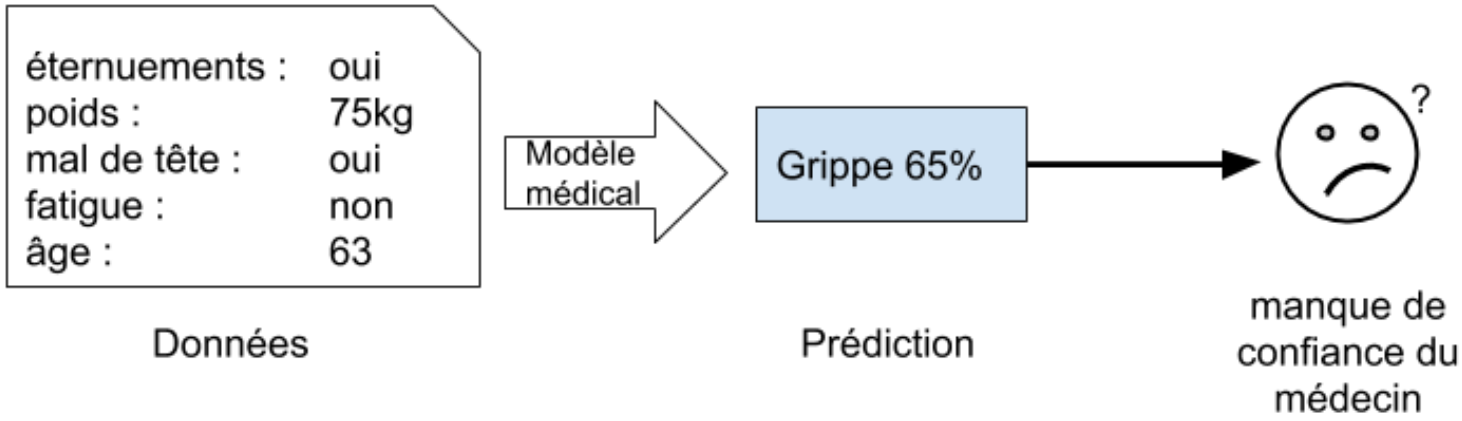
III. Explication

a) Mise en œuvre

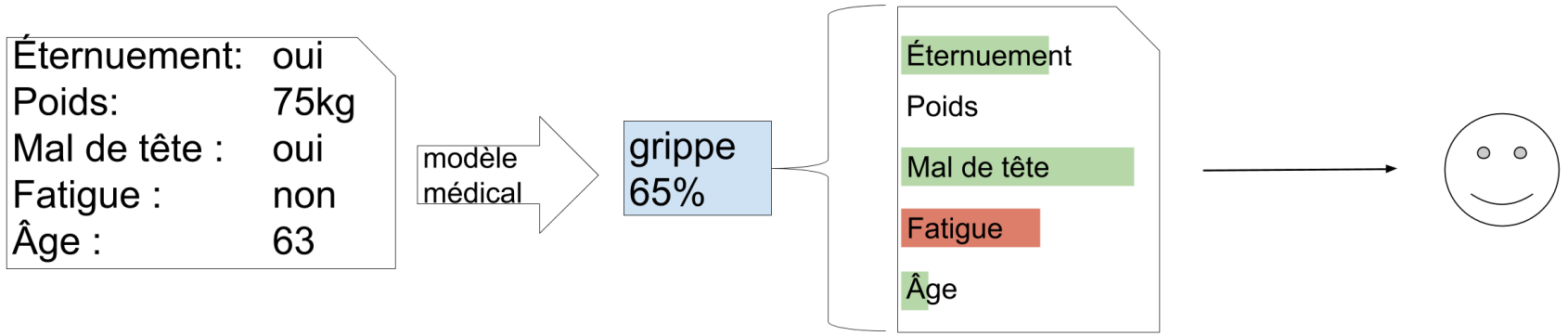
b) Applications

Contrecarrer l'effet « boîte noire »

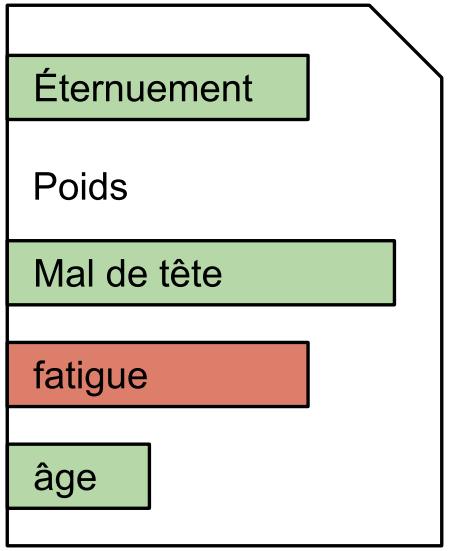
- Un utilisateur d'un modèle prédictif peut avoir du mal à placer sa confiance dans les prédictions produites



- Solution : Présenter une explication à l'utilisateur



Génération des explications



- Soit l'attribut $a \in A$, son influence sur la prediction du modèle f pour l'instance x correspond à :

$$Inf_x(a) = f(x) - f(x \setminus a)$$

Simuler l'absence d'un attribut

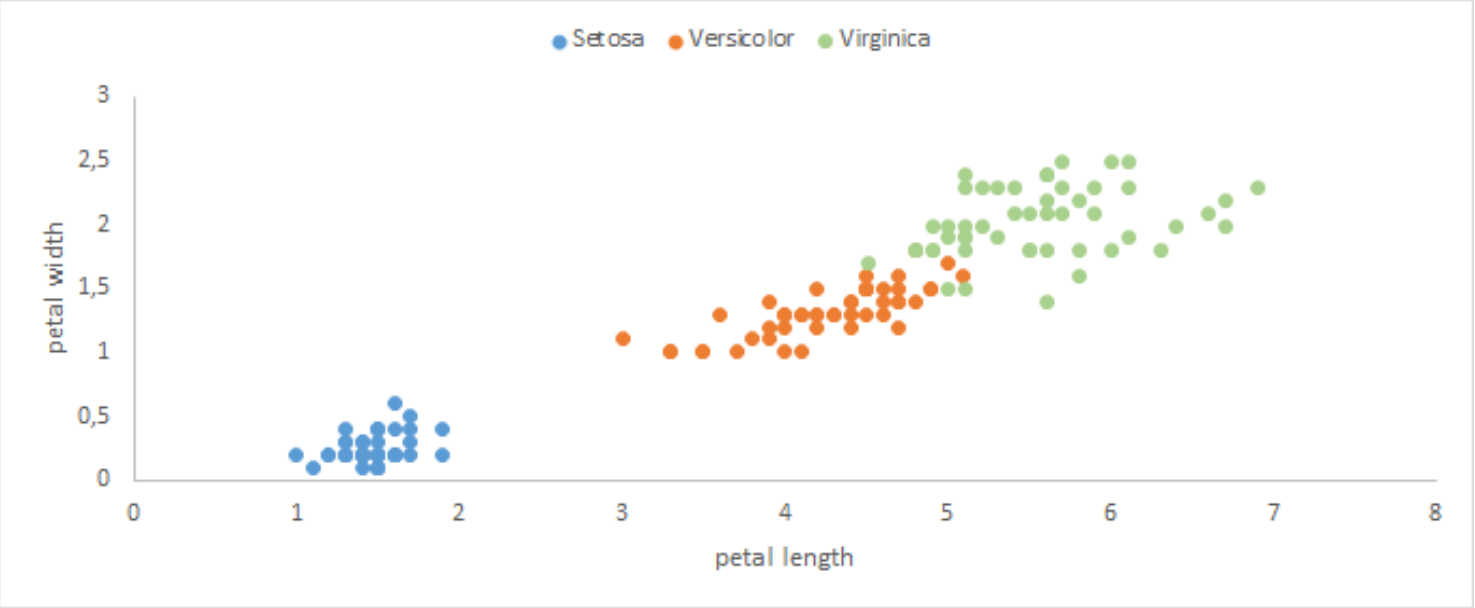
Trois méthodes possibles :

- Définir l'attribut comme « manquant »
  - + :Simple
  - - : implique de n'utiliser que des algorithmes fonctionnant avec des attributs manquants

- Réentraîner le modèle sans l'attribut
  - + : Générique et proche de l'intuition originelle
  - - : Coûteux en calculs

- Procéder par une moyenne statistique
  - + : Mise en place simple et peu coûteuse
  - - : Nécessite une connaissance à priori du dataset pour un fonctionnement optimal.

Non indépendance des attributs



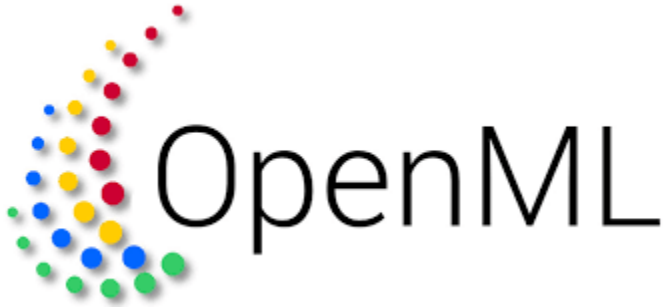
- Il faut tenir compte, non pas uniquement des attributs seuls, mais aussi des groupes d'attributs.

Coût calculatoire

- Calculer toutes les influences de toutes les combinaisons de groupes d'attributs serait trop cher à mettre en place
- Plusieurs possibilités pour alléger la charge calculatoire :
  - Ne calculer que jusqu'à un certain nombre d'attributs dans chaque groupes
  - Trouver les groupes d'attributs effectivement liés et ne prendre en compte que ceux là

- Besoin de vérifier ces méthodes par une expérience.

Openml



- Base de données rassemblant des datasets et des workflows
- Open source et disposant de nombreuses APIs
- Permet de lancer de nouvelles runs et d’ajouter de nouvelles données très facilement

I. Introduction

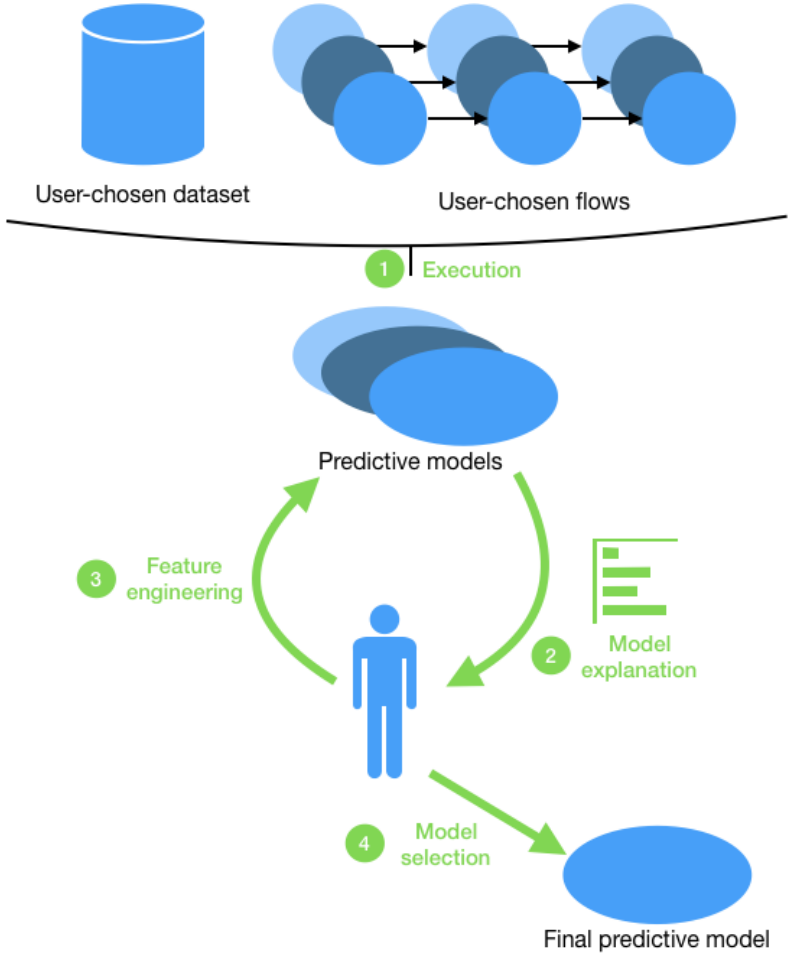
II. Recommandation

III. Explication

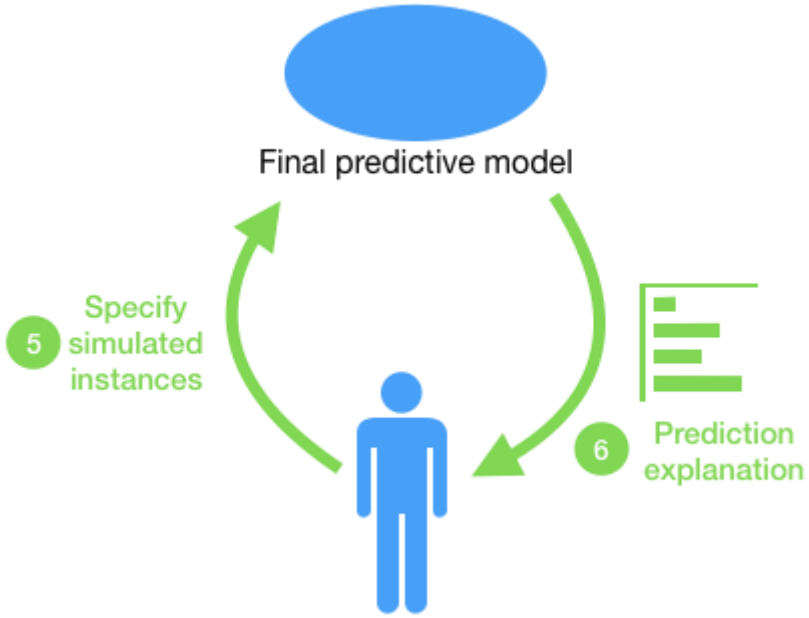
a) Mise en œuvre

b) Applications

Entraînement d'un modèle



Exploitation d’un modèle



Merci pour votre attention

Avez-vous des questions?