Tutoriel

L'éthique de l'Intelligence Artificielle

Nadia Abchiche-Mimouni (équipe IRA2) & Farida Zehraoui (équipe ARO@S)

Laboratoire IBISC univ. Evry université Paris-Saclay

Plate-Forme Intelligence Artificielle Saint-Étienne, France - 27 Juin 2022
https://ci.mines-stetienne.fr/pfia2022/tutoriels/?p=programme#t1

PLAN

Partie I

Éthique, Transparence et Interprétabilité en IA.

Motivations/Définitions

Partie II

Transparence et Explicabilité en IA : état de l’art

Partie III

Démonstration

QUESTIONNAIRE 1

Partie I
Ethique, Transparence et Interprétabilité en IA : Motivations/Définitions

Définitions à partir de la philosophie morale

Morale

- Ensemble de valeurs et de principes qui permettent de différencier le bien du mal, le juste de l'injuste, l'acceptable de l'inacceptable, et auxquels il faudrait se conformer
- Ensemble des normes propres à un individu, un groupe social ou un peuple, à un moment précis de son histoire
- Notion de droits, de devoirs et d'interdits
- Notion de bonne/mauvaise action

Déontologie

- Règles et devoirs qui encadrent une profession ou un groupe de personnes

Éthique

- Ce n'est pas un ensemble de valeurs ni de principes en particulier
- Il s'agit d'une réflexion argumentée en vue du bien-agir
- Relativise la notion de la bonne/mauvaise action

Collective

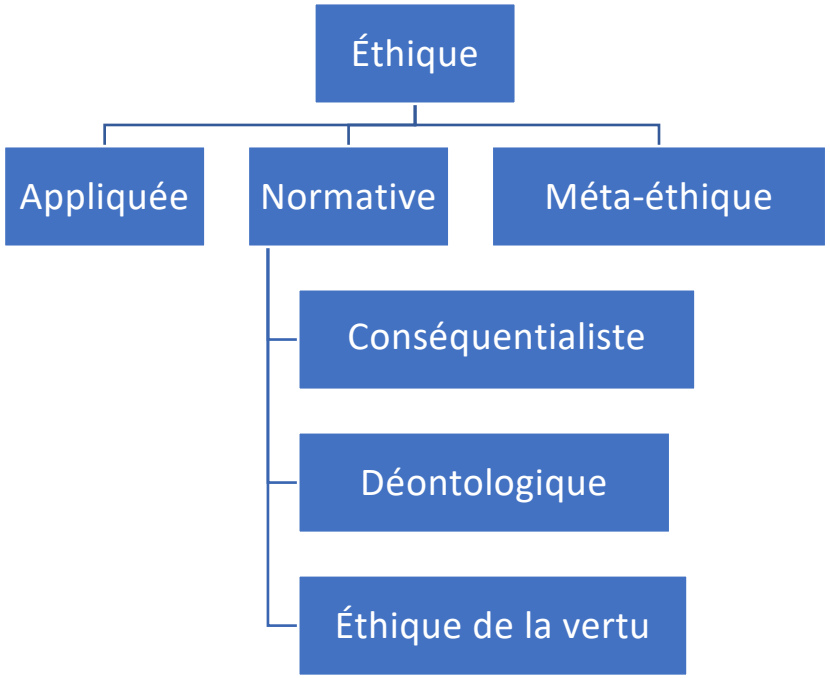
Interpersonnelle

Individuelle

- Science de la conduite et de la morale
- réponse à la question : que dois-je faire ?

Éthique

De façon générale, l’éthique propose de s'interroger sur les valeurs morales et les principes moraux qui devraient orienter nos actions, dans différentes situations, dans le but d'agir conformément à ceux-ci.



Éthique appliquée

Analyse de situations concrètes qui soulèvent des questions éthiques :

- Accent mis sur le soutien à la prise de décision face à des enjeux concrets.
- Ne concerne le point de vue de la forme et du processus décisionnel que du point de vue substantiel, c'est-à-dire des valeurs et principes en jeu et de leurs rapports entre eux.

Exemples :

- Bioéthique : procréation artificielle, manipulations génétiques...
- Éthique de l'environnement : développement durable, responsabilité face aux générations futures, biodiversité...

Éthique normative (1/2)

Proposition de règles pour évaluer une action d'un point de vue moral.
Ne permet pas de déterminer, entre deux actions, laquelle est moralement meilleure.

1. Éthique conséquentialiste (ou conséquentialisme)
   - S'intéresse à l'ensemble des conséquences
   - Une action est bonne si ses conséquences sont bonnes
   - Exemple : aveu adultère
2. Éthique déontologique (ou déontologisme)
   - Notion de devoir, d'obligation et d'impératif moral
   - Un acte est moralement bon ou mauvais indépendamment de ses conséquences
   - Exemple : devoir de parent
3. Éthique de la vertu
   - Traits de caractère dont témoignent les actions
   - Exemple : l'aveu est toujours associé à la vertu d'honnêteté

 Une conception du bien : visée du bon

Éthique normative (2/2)

Une conception du bien

- Quelles sont les meilleures conséquences ?
- Quelles sont mes obligations morales ?
- Comment savoir quelles vertus adopter?
- Une conception des bonnes conséquences, des devoirs moraux fondamentaux, vertus à privilégier
- Théories morales incluses dans ces trois grandes approches

 Méta-éthique

Méta-éthique

- Philosophie qui analyse les modes de fonctionnement et les valeurs d’un système éthique
- Porte sur :
  - La nature même des jugements moraux
  - Les propriétés morales que l'on prête aux actions, aux personnes et aux traits de caractère
  - La définition des fondements de l’éthique normative

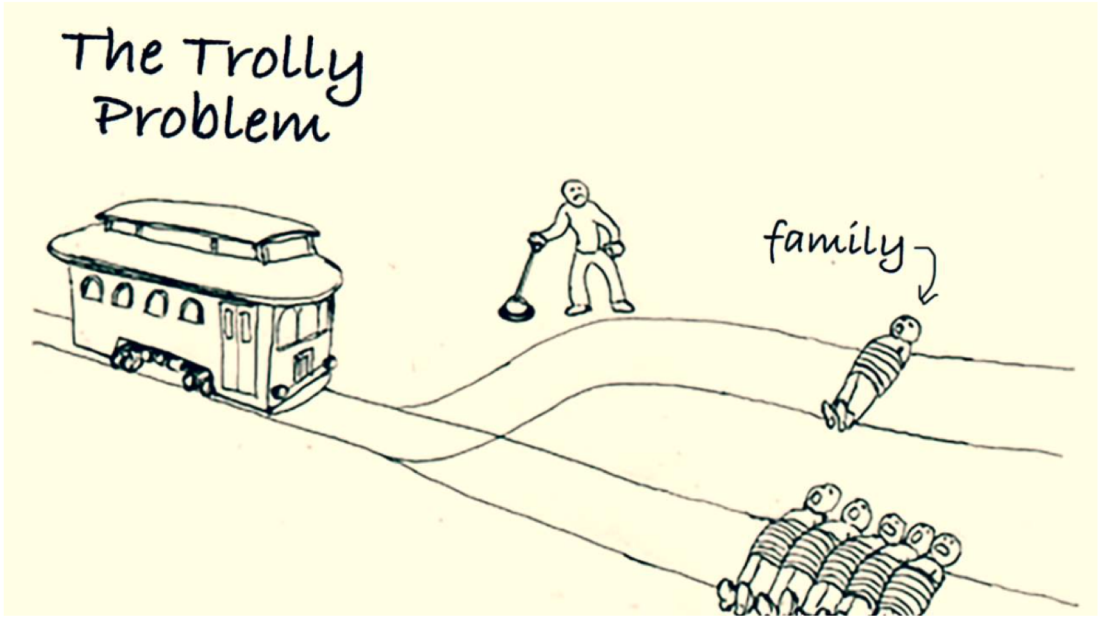
Dilemmes éthiques (1/2)

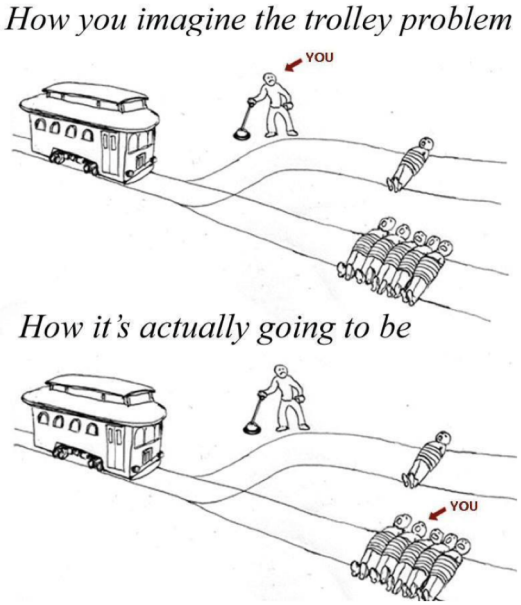
- Situations dans lesquelles tout choix conduit à transgresser des principes éthiques acceptés ET une décision doit absolument être prise (Kirkpatrick, 2015).
- Un principe éthique est incapable de donner une préférence entre deux options : chaque option est cautionnée par une règle éthique, sachant qu'exécuter les 2 options n'est pas possible (McConnell 2014).

Exemple : Livre de la république de Platon

- Cephalus définit la justice comme le fait de dire la vérité et de toujours rendre un bien emprunté.
- Socrate réfute en suggérant que rendre à une personne une arme qu'elle nous a prêtée pose problème si on sait que cette personne a des problèmes psychiatriques et risque de faire une mauvaise utilisation de l'arme.
- Dilemme morale entre 2 normes morales :
  1. Rendre un objet emprunté
  2. Protéger les autres de criminels potentiels

Dilemmes éthiques (2/2)





… et si l'éthique de l'IA avait une histoire ?

« *La machine analytique [ordinateur] n'a nullement la prétention de créer quelque chose par elle-même. Elle peut juste exécuter tout ce que nous saurons lui ordonner d'exécuter […] Son rôle est de nous aider à effectuer ce que nous savons déjà dominer […] des opérations* ***numériques*** *et aussi* ***symboliques*** »
https://www.adalovelaceinstitute.org/, accédée le 25/06/22

Ada Lovelace, 1840

« *Je produirai une poésie mathématiques plus philosophique et plus élevée que celle que nous connaissons. Y a-t-il des vérités qui ne peuvent être obtenues que par l'analyse mathématique et pas autrement ? Est-ce que toutes les expressions analytiques abstraites représentent quelques chose de réel ?* »

Selon Nicolas Witkowski, physicien et auteur de l'essai d'histoire des sciences au féminin *Trop belles pour le Nobel* (Le Seuil, 2016), consacré à Ada Lovelace : « Ce qu'elle cherche, c'est plus une métaphysique qu'une science ou une technique. Une façon d'être, de vivre. »

Problèmes et questions posés par l'IA

- Atteinte à la vie privée et aux libertés
- Contrôle et risque d'asservissement des personnes
- Risques liés à l'utilisation des réseaux sociaux
- Espionnage, cyberattaques et criminalité
- Systèmes autonomes (Robots, armes létales…)
- Prolétarisation
- Atomisation de la société
- Perte de la solidarité
- …

Intelligence Artificielle Symbolique Vs Numérique

IA Numérique : Machine Learning (ML)

✓ Construit des modèles basés sur les données (données d'apprentissage) pour réaliser des tâches spécifiques

✓ Capacités de prédiction s'améliorent au fur et à mesure de leur apprentissage

✓ Prédiction de nouvelles situations

- Entièrement guidée par les données
- Les modèles les plus performants sont opaques

IA Symbolique

✓ Repose sur le raisonnement logique et formel

✓ À l'origine des premiers systèmes experts

✓ Permet de tracer le processus de raisonnement

- N'explique pas les raisons des décisions (solution/échec à trouver une solution)
- Basée sur une expertise limitée

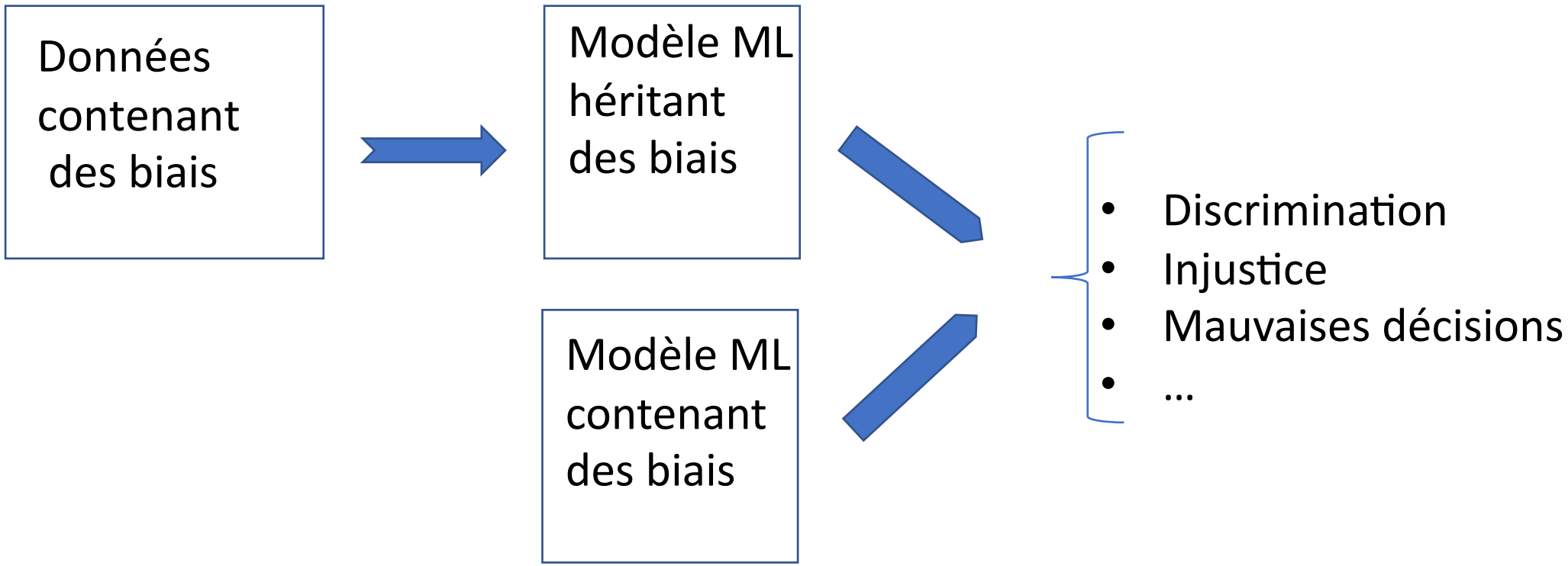
Initiatives pour limiter les risques liés à Intelligence Artificielle

- Lettre ouverte en 2015 initiée par Stephen Hawking et Elon Musk pour interdire les robots tueurs (signée par plus de 1000 chercheurs)
- RGPD, 2016
- Yoshua Bengio en 2018 appelle les gouvernements pour réglementer avec des lois en vue du développement de l'IA pour le bien social
- Déclaration de Montréal en 2018, pour une IA responsable (signée par : Yoshua Bengio et Yann LeCun)
- ONU : Appel du 15 septembre 2021 pour l'instauration d'un moratoire sur certaines applications de l'IA
- Conférence mondiale à l'UNESCO en 2019 : « Principes pour l'Intelligence Artificielle : vers une approche humaniste ? »
- International Telecommunication Union (ITU) & AI for good Foundation : "*What if AI were developped to serve humanity rather than commerce* ? »
- Global AI Ethics Institute : think tank mondial sur l'éthique appliquée à l'IA
- Lettre ouverte : *Research Priorities for Robust and Beneficial Artificial Intelligence* ([Stuart Russell & al., 2015])
- *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*
- The Ethics and Governance of Artificial Intelligence Initiative : projet commun au MIT Media Lab et the Harvard Berkman-Klein Center for Internet and Society (fairness, human autonomy, and justice**)**
- Global Catastrophic Risk Institute
-

Partie II
Transparence et Explicabilité en IA : état de l'art

Transparence et Explicabilité en IA numérique

ML : Biais dans les données/modèles



ML : Biais dans les données/modèles

Biais implicite : Discrimination ou préjugé à l'encontre d'une personne ou d'un groupe (minorités non représentées, utilisation de caractéristiques favorisant la discrimination, …)

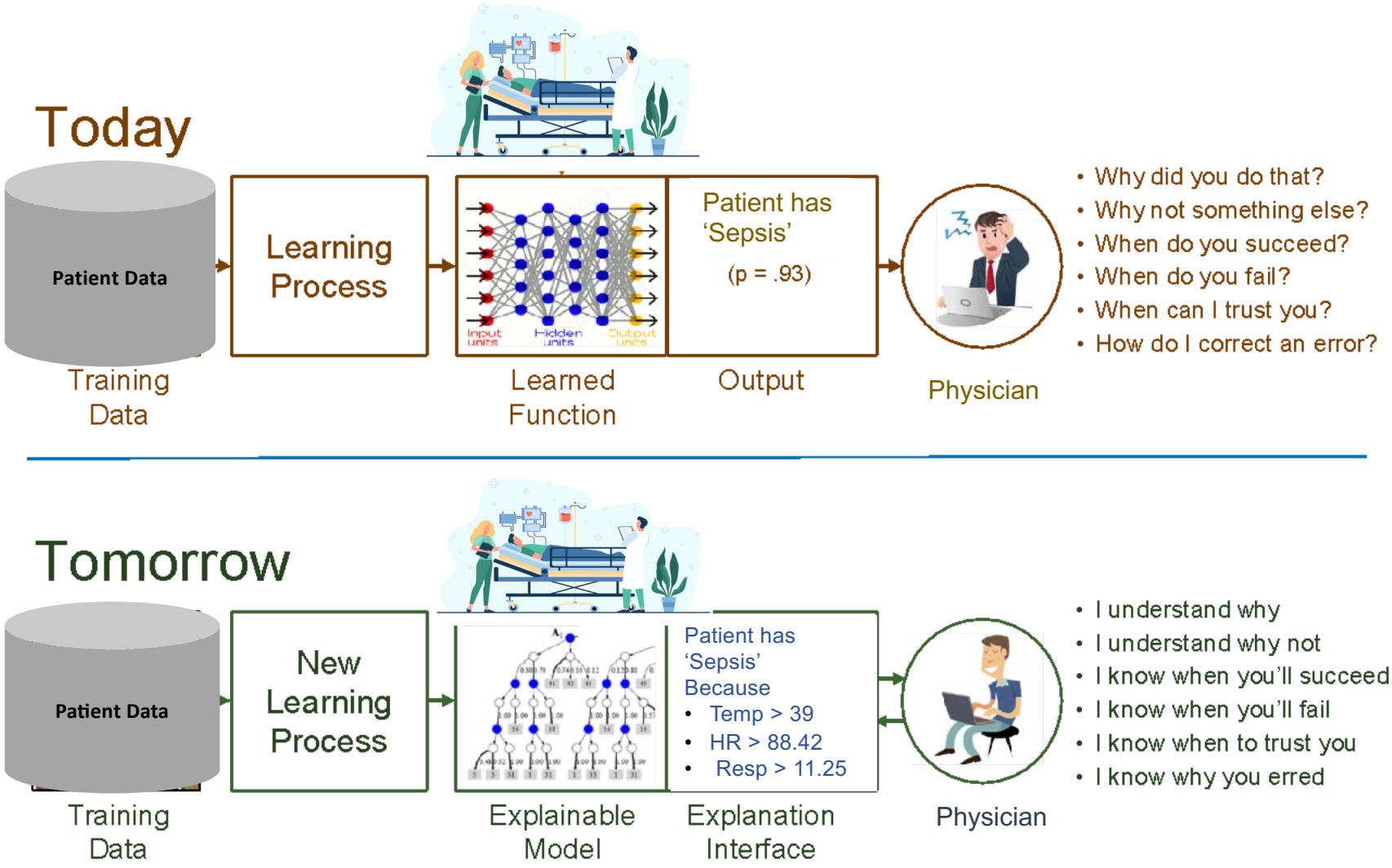
Biais d'échantillonnage : Utilisation d'un échantillon de données non représentatif de la distribution de la population.

Biais temporel : Construction de modèles performants pour le moment présent, mais qui le seront moins dans le futur à cause de la non prise en compte des éventuels changements futurs lors de la création du modèle.

Sur-ajustement aux données d'entraînement : Cela se produit lorsque le modèle d'IA peut prédire avec précision les valeurs de l'ensemble de données d'apprentissage, mais ne peut pas prédire avec précision les nouvelles données.

…

Explicabilité : Motivation



Source: https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

Explicabilité : Motivation

L'application de méthodes d'IA dans des domaines sensibles est problématique
Les mauvaises décisions peuvent être dangereuses

Voitures autonomes

- Risque de mauvaises décision
- Responsabilité

Quel angle de braquage la voiture autonome doit choisir pour éviter un accident?

Qui est responsable en cas d'accident?

A prototype of Waymo's self-driving car, navigating public streets in Mountain View, California in 2017

Explicabilité : Motivation

Qui est responsable ?

L'application de méthodes de ML dans des domaines sensibles est problématique

Les mauvaises décisions peuvent être dangereuses

Premier accident d'une voiture autonome

Décès d'Elaine Herzberg (2 août 1968 - 18 mars 2018) causé par une voiture autonome en Mars 2018.

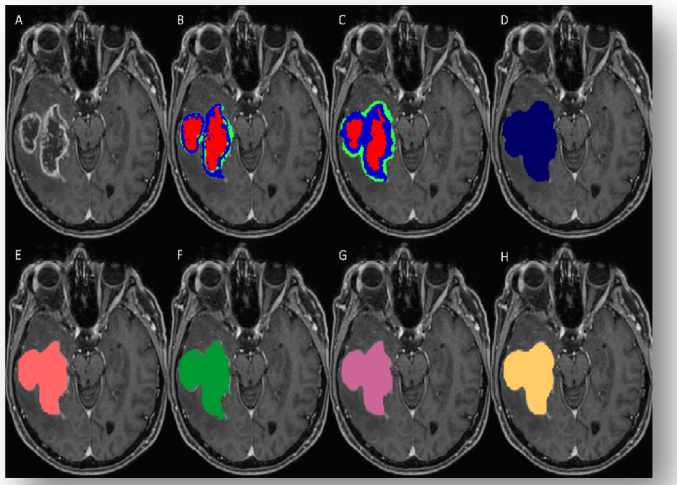
Herzberg poussait un vélo sur une route à quatre voies à Tempe, en Arizona, aux États-Unis, lorsqu'elle a été heurtée par un véhicule d'essai Uber, qui fonctionnait en mode autonome avec un conducteur de secours humain assis au siège du conducteur.

- Nouvelles recommandations du National Transportation Safety Board (NTSB)
- Les tests sur les véhicule autonomes ont été suspendus en Arizona jusqu'en décembre 2018

Explicabilité : Motivation

Domaine de la Santé

- Risque de mauvaises décisions
- Responsabilité
- Confidentialité
- Biais dans les données
- Validation du modèle



Quel est le stade de cancer du patient?

Explicabilité : Motivation

Biais dans les données des algorithmes d'IA

1. Tests auditifs pour déterminer le stade précoce de la maladie d'Alzheimer : **précision > 90 %**

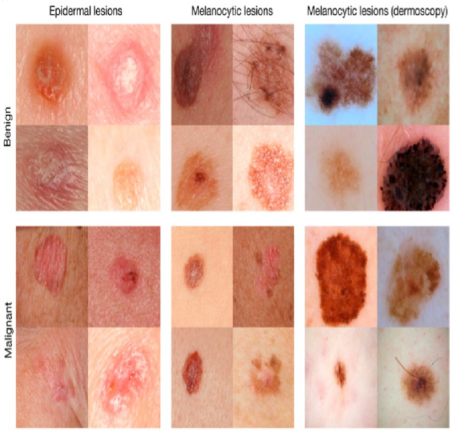
Les données se composaient d'échantillons de locuteurs natifs anglais uniquement.

➡ Pour les non-anglophones : les pauses et les erreurs de prononciation identifiées comme des indicateurs de la maladie.

2. Algorithme pour diagnostiquer les lésions cutanées malignes à partir d'images.

Les données disponibles majoritairement sur des peaux claires.

➡ Efficacité pour diagnostiquer uniquement les personnes à la peau claire



A Esteva *et al. Nature* 1–4 (2017) doi:10.1038/nature21056

Explicabilité : Motivation

Domaine de la justice

- Discrimination
- Biais dans les données
- Confidentialité
- Validation du modèle

Plusieurs tribunaux utilisent des outils *prédictifs* pour décider du sort des prisonniers : détention provisoire, libération conditionnelle, etc.

Explicabilité : Motivation

Domaine de la justice

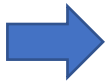
En 2015, a vu le jour le logiciel COMPAS.
On a également réalisé que cet logiciel pouvait être…**raciste**.
En prédisant la probabilité d'une future activité criminelle, COMPAS s'appuie sur des événements passés comme les arrestations.
Or, des observations sur les pratiques policières montrent que les personnes issues de la diversité ont, à crime égal, bien plus de risques d'être arrêtées que les personnes blanches…

L'IA reproduit souvent les biais des humains…

[Hutson, 2018]

Règlement général sur la protection des données (RGPD)

Nouveau règlement général sur la protection des données dans l'Union européenne



Droit à l'explication

L'interprétabilité des modèles de ML garantit la conformité à la législation

Explicabilité en ML

L’explicabilité est un puissant outil de :

- Détection des biais de données
- Vérification du bon comportement du modèle
- Optimisation modèle / architecture
- Meilleure compréhension du problème

Interprétabilité

Définitions

"Décider quelle est la signification voulue de quelque chose" (Dictionnaire de Cambridge)

"Interprétation : Explication du sens qu'on peut donner à un texte"
(Dictionnaire de l'académie française)

Interprétabilité

2015

''L'interprétabilité est la mesure dans laquelle un humain peut prédire de manière cohérente le résultat du modèle '' [Kim & al. 2015]

2016

''Dans l'interprétabilité indépendante du modèle, le modèle est traité comme une boîte noire. Les modèles interprétables peuvent également être plus souhaitables lorsque l'interprétabilité est beaucoup plus importante que la précision, ou lorsque les modèles interprétables formés sur un petit nombre de caractéristiques soigneusement conçues sont aussi précis que les modèles de boîte noire." [Ribeiro et al. 2016]

2019

"Nous définissons l'apprentissage automatique interprétable comme l'utilisation de modèles d'apprentissage automatique pour l'extraction de connaissances pertinentes sur les relations de domaine contenues dans les données..." [Murdoch et al., 2019].

L'interprétabilité consiste à extraire les connaissances pertinentes compréhensibles par l'homme

Explicabilité

Définitions

"Expliquer : Rendre quelque chose clair ou facile à comprendre en le décrivant ou en donnant des informations à son sujet" (Dictionnaire de Cambridge)

"Explication : Ce qui donne la cause, le motif, la raison, d'un fait, d'un phénomène'" *(Dictionnaire de l'académie française)*

Explicabilité

2017 “L'explication est considérée comme étroitement liée au concept d'interprétabilité; Les systèmes sont interprétables si leurs opérations peuvent être comprises par un humain, soit par introspection, soit par une explication produite” [Biran and Cotton, 2017].

2017 ‘’ L'interprétabilité est la capacité d'expliquer ou de donner un sens en termes compréhensibles pour un humain’’[Doshi-Velez et Kim, 2017]

2019 L'interprétabilité est “ Le degré auquel l'humain peut comprendre la cause de la décision ’’ , “J'assimile l'interprétabilité à l'explicabilité” [Miller, 2019]

L'explication est composée de suffisamment de concepts que l'humain peut facilement comprendre.

?

Explicabilité ⟺ Inteprétabilité

Transparence

Définitions

"Transparent : Clair et facile à comprendre" (Dictionnaire de Cambridge)

'' Caractère de ce qui est transparent. *La transparence de l'eau, du verre''*
(Dictionnaire de l'académie française)

Transparence

2012

"La transparence décrit clairement la structure du modèle, les équations, les valeurs des paramètres et les hypothèses pour permettre aux parties intéressées de comprendre le model" [Briggs et al., 2012]

2016

" De manière informelle, la transparence est le contraire de l'opacité ou de la boîte noire. Cela implique un certain sens de la compréhension du mécanisme par lequel le modèle fonctionne. Nous considérons la transparence au niveau du modèle, au niveau des composants individuels (par exemple, les paramètres) (décomposabilité) et au niveau de l'algorithme d'apprentissage (transparence algorithmique)" [Lipton,2016].

2018

"La transparence est un niveau auquel un système fournit des informations sur son fonctionnement ou sa structure interne" et "l'explicabilité et la transparence sont importantes pour améliorer l'interprétabilité du créateur". [Tomsett et al., 2018]

La transparence doit satisfaire des critères de compréhension à plusieurs niveaux : modèle, composants, paramètres, algorithme d'apprentissage, …

Intelligibilité

Définitions

"Intelligible : Assez clair pour être compris" (Dictionnaire de Cambridge)

"Intelligible : Qui peut être compris, dont le sens se comprend aisément.'' *(Dictionnaire de l'académie française)*

Intelligibilité

2001

"Intelligibilité : les systèmes sensibles au contexte qui cherchent à agir sur ce qu'ils infèrent sur le contexte doivent être capables de représenter à leurs utilisateurs ce qu'ils savent, comment ils le savent et ce qu'ils font à ce sujet" [Bellotti and Edwards, 2001].

2009

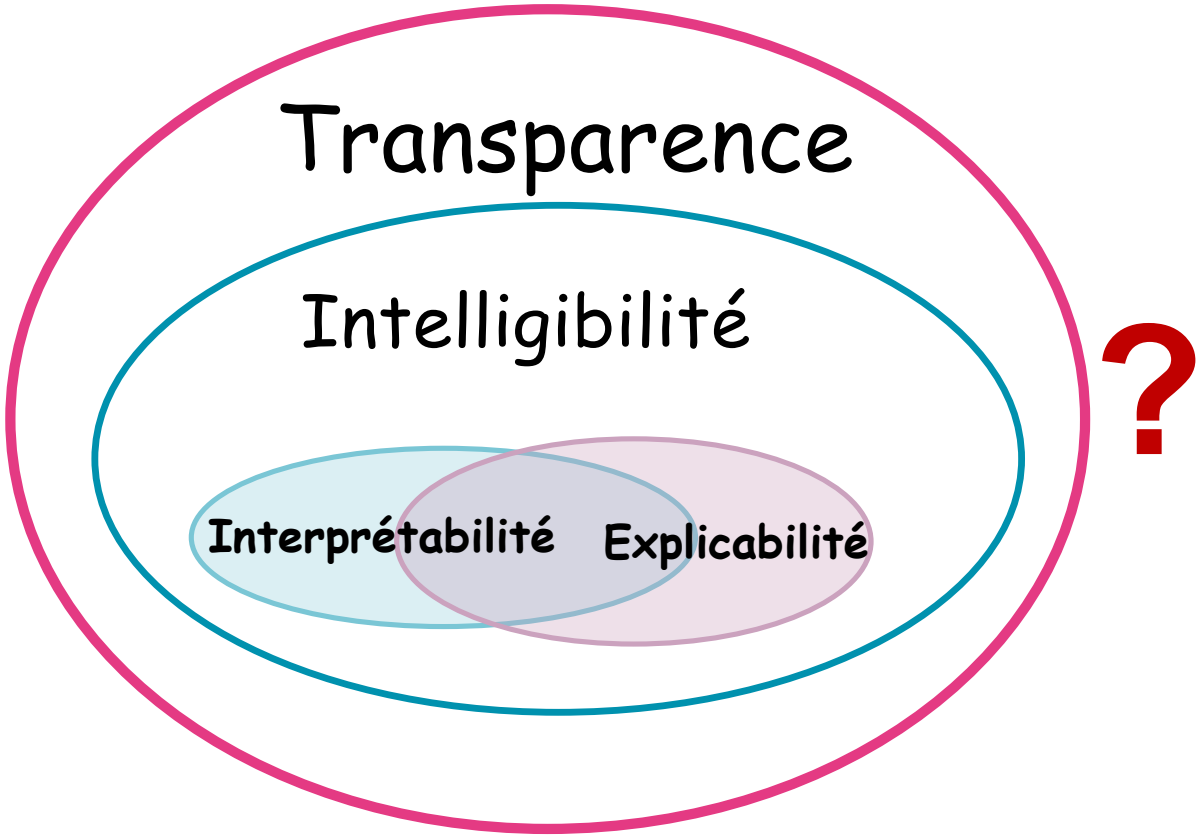
"L'intelligibilité peut aider à exposer le fonctionnement interne et les entrées des applications sensibles au contexte qui ont tendance à être opaques pour les utilisateurs en raison de leurs actions implicites" [Lim and Dey, 2009].

2018

"Il reste remarquablement difficile de spécifier ce qui rend un système intelligible ; Le principal défi pour concevoir une IA intelligible est de communiquer un processus de calcul complexe à un humain." [Weld and Bansal, 2018].

L'intelligibilité permet d'exposer le fonctionnement interne et les entrées des applications sensibles au contexte

Explicabilité en IA : Définitions



[Clinciu & Hastie, 2019]

Transparence et Explicabilité en IA numérique

Exemple : Cas d’étude

Prédiction de la maladie « Sepsis » à partir des variables :

Cliniques

HR Heart rate (beats/min)
O2Sat Pulse oximetry (%)
SBP Systolic BP (mm Hg)
MAP Mean arterial pressure (mm Hg)
Resp Respiration rate (breaths/min)

Démographique

Age Age (yr)
Gender Female (0) or male (1)

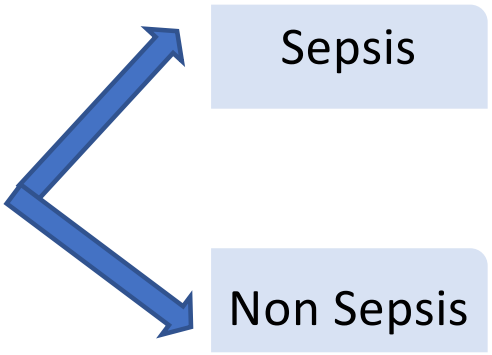
https://www.biomerieux.com/corp/fr/blog/sepsis/quick-facts-about-sepsis-.html

Exemple : Cas d'étude

Prédiction de la maladie « Sepsis »

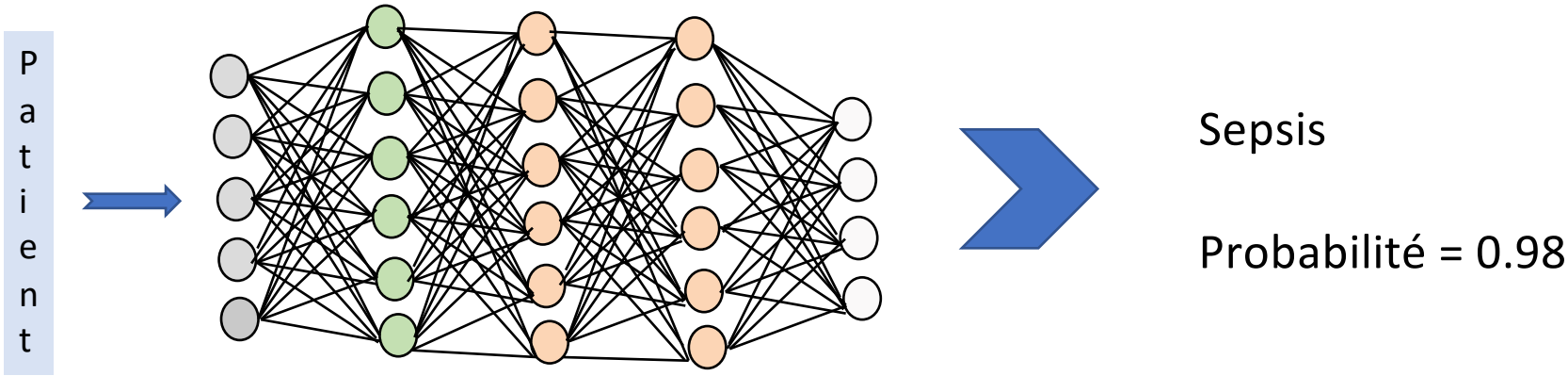
Données

	HR	O2Sat	SBP	MAP	Resp	Age	Gender
0	82.48	97.0	121.0	78.00	18.0	53.20	0.0
1	82.48	97.0	121.0	78.00	18.0	66.00	0.0
2	135.00	95.0	130.0	90.67	26.0	79.65	0.0
3	68.00	100.0	126.0	67.33	18.0	77.33	0.0
4	76.00	100.0	159.0	95.00	24.0	75.50	1.0



Modèles opaques

- Machines à vecteurs de support (SVM), Méthodes d'ensembles, …
- Deep Learning



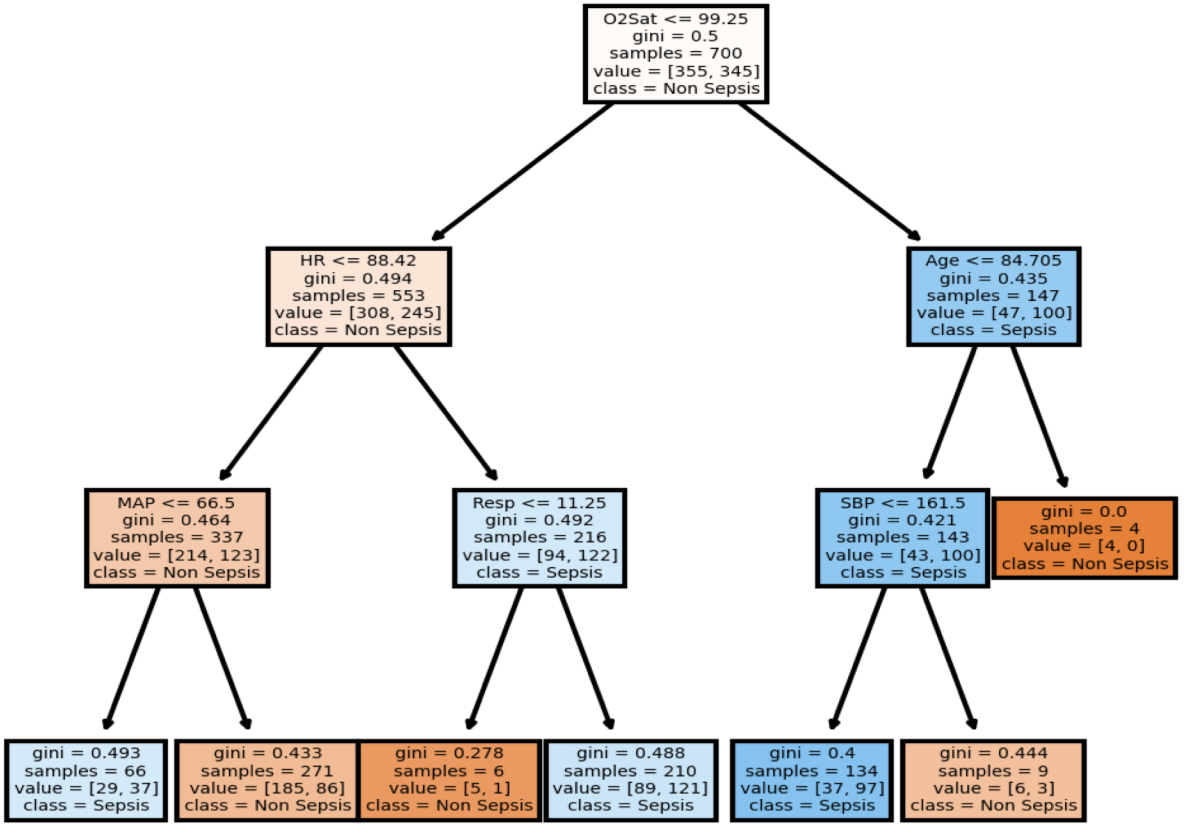
Accuracy = 0,90

Modèles interprétables

Arbres de décision

Un arbre de décision est un arbre où :

- Chaque nœud intermédiaires teste un attribut
- Chaque branche correspond à une valeur d'attribut
- Chaque feuille est étiquetée par une classe



Accuracy = 0,60

Modèles interprétables

Règles logiques

- if (O2Sat <= 99.25) and (HR <= 88.42) and (MAP > 66.5) then class: Non Sepsis (proba: 68.27%) | based on 271 samples
- if (HR > 88.42) and (Resp > 11.25) then class: Sepsis (proba: 57.62%) | based on 210 samples
- if (O2Sat > 99.25) and (Age <= 84.705) and (SBP <= 161.5) then class: Sepsis (proba: 72.39%) | based on 134 samples
- if (O2Sat <= 99.25) and (HR <= 88.42) and (MAP <= 66.5) then class: Sepsis (proba: 56.06%) | based on 66 samples

Les arbres de décision peuvent être considérés comme un cas particulier de règles logiques.

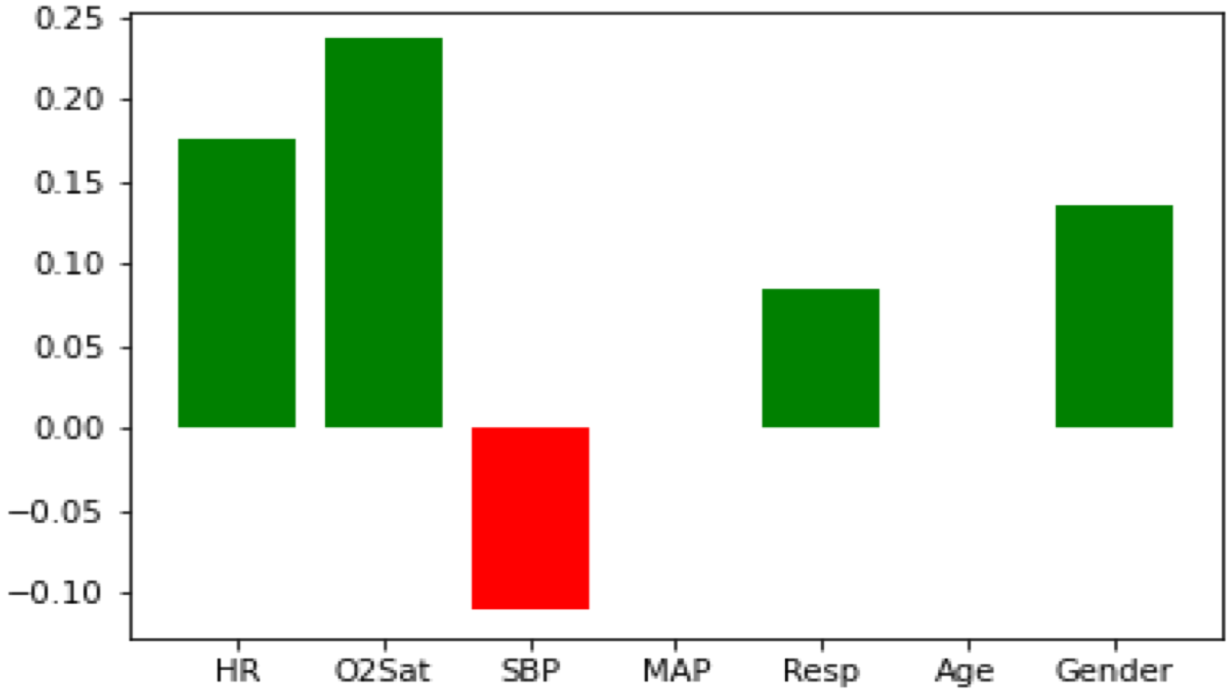
Accuracy = 0,57

Modèles interprétables

Utiliser des modèles interprétables capables d'imiter le comportement des modèles « boîte noire »

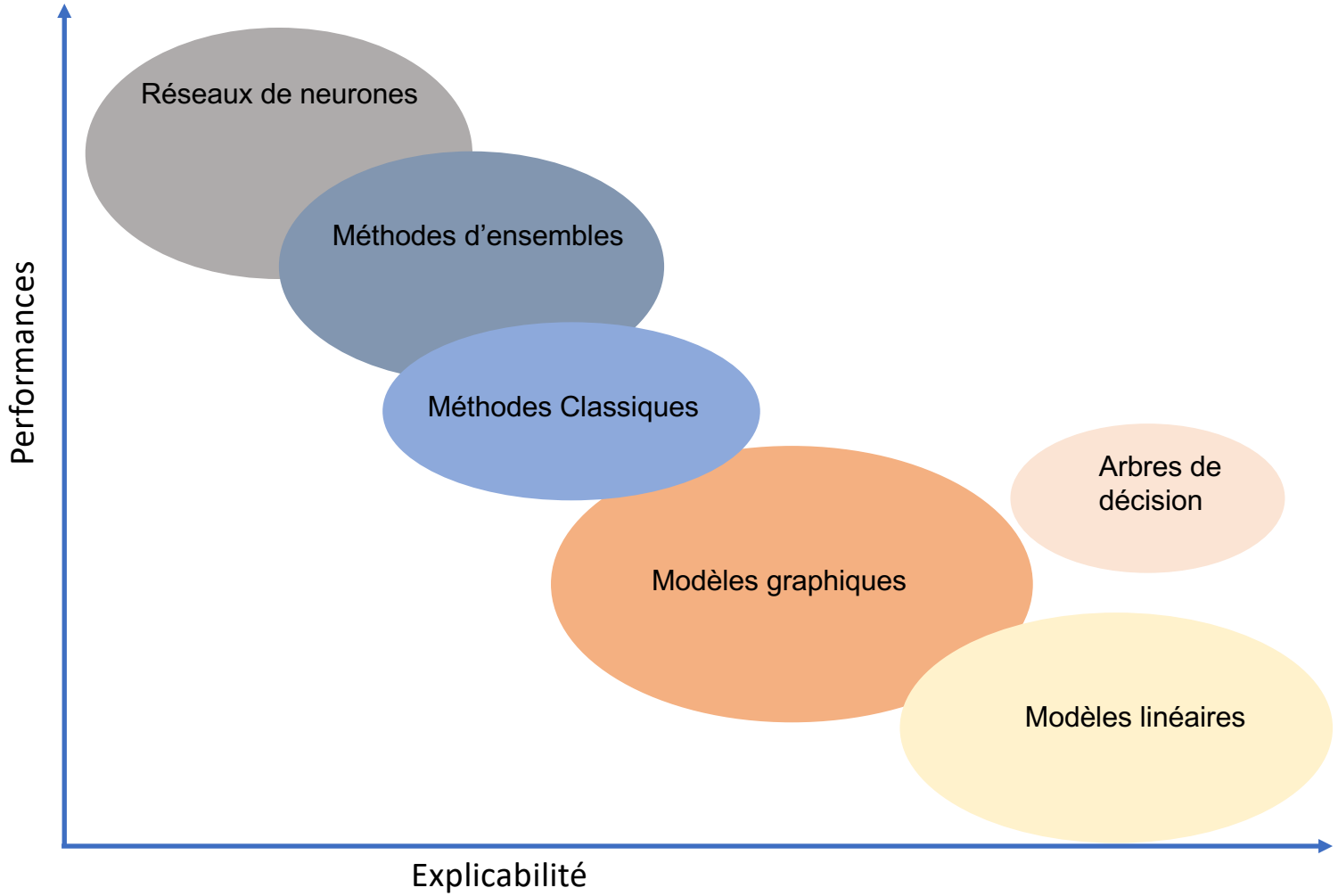
Fonctions linéaires

$$f(X) = \sum_{i=1}^{m} \alpha_i X_i + \beta$$

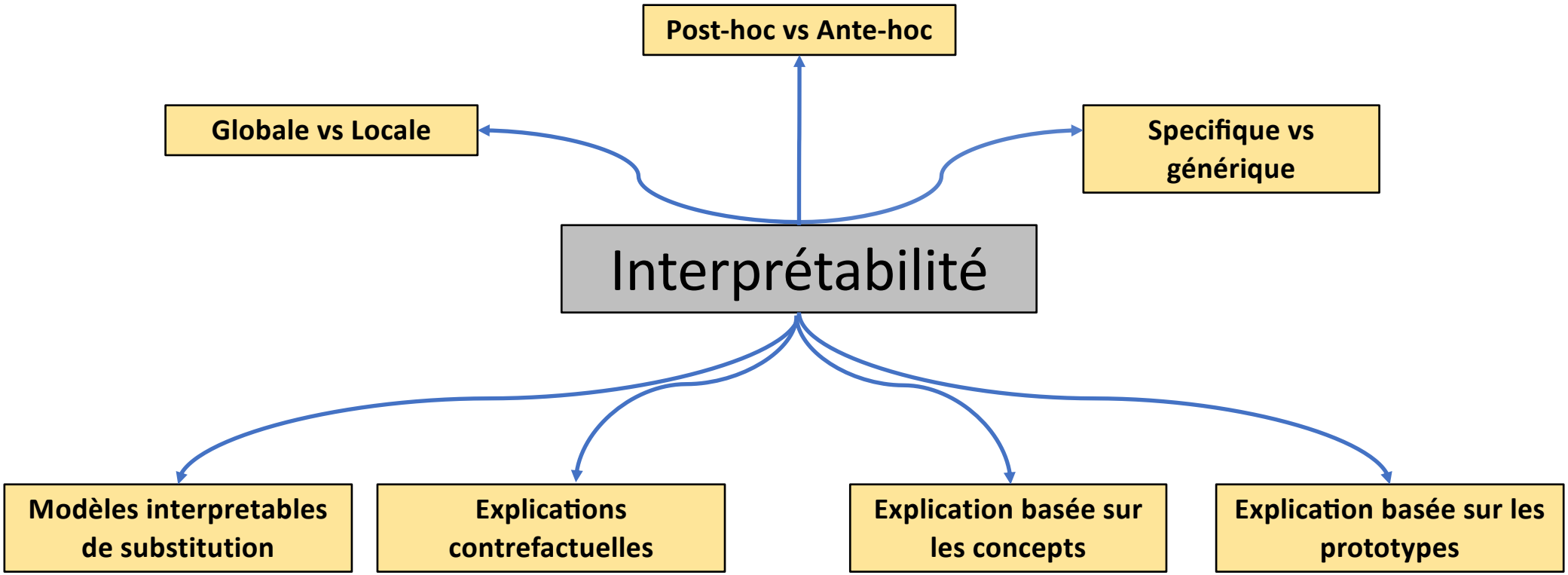


Accuracy = 0,56

Explicabilité vs Performance



Taxonomie de l'interprétabilité en ML



[Doshi-Velez et Kim, 2017] [Guidotti et al. 2018][Lipton, 2018] [Arrieta et al., 2020] [Vilone & Longo, 2020][Zhang et al., 2021][Linardatos et al.2021]…

Explication basée sur les concepts/variables

Méthodes d'attribution

Calculent des valeurs de pertinences (ou de contributions) pour toutes les caractéristiques (variables) d'entrée et pour certains composants internes (Exp. neurones).

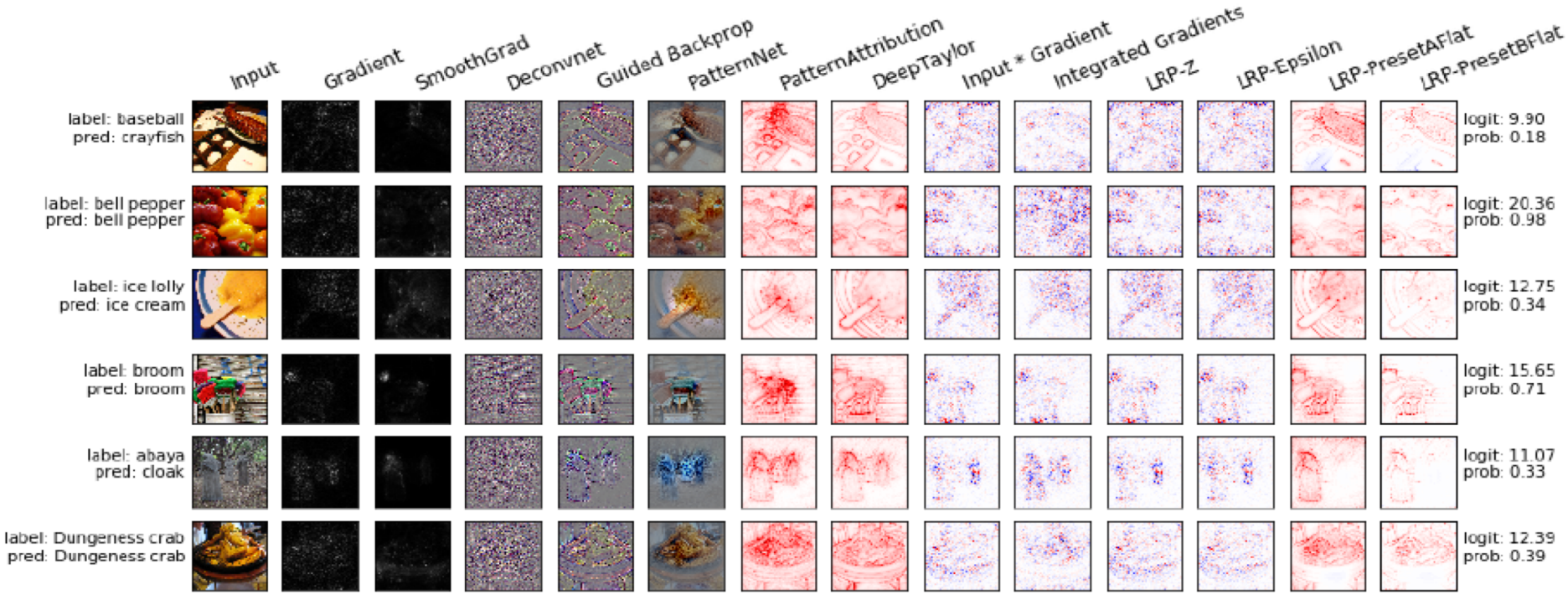
Analyse de sensibilité : Sensibilité du modèle par rapport aux variables d'entrée ou leur abstraction [Fong et al., 2019][Ancona et al. 2017]…

Méthodes de perturbation : Perturbation des variables d'entrée et analyse des changements dans les sorties [Kindermans et al., 2019][Ivanovs et al. 2021]…

Méthodes de décomposition : Décomposition du signal de la sortie jusqu'à l'entrée [Bach et al., 2015] [Montavon, 2018]][Ancona et al. 2018]…

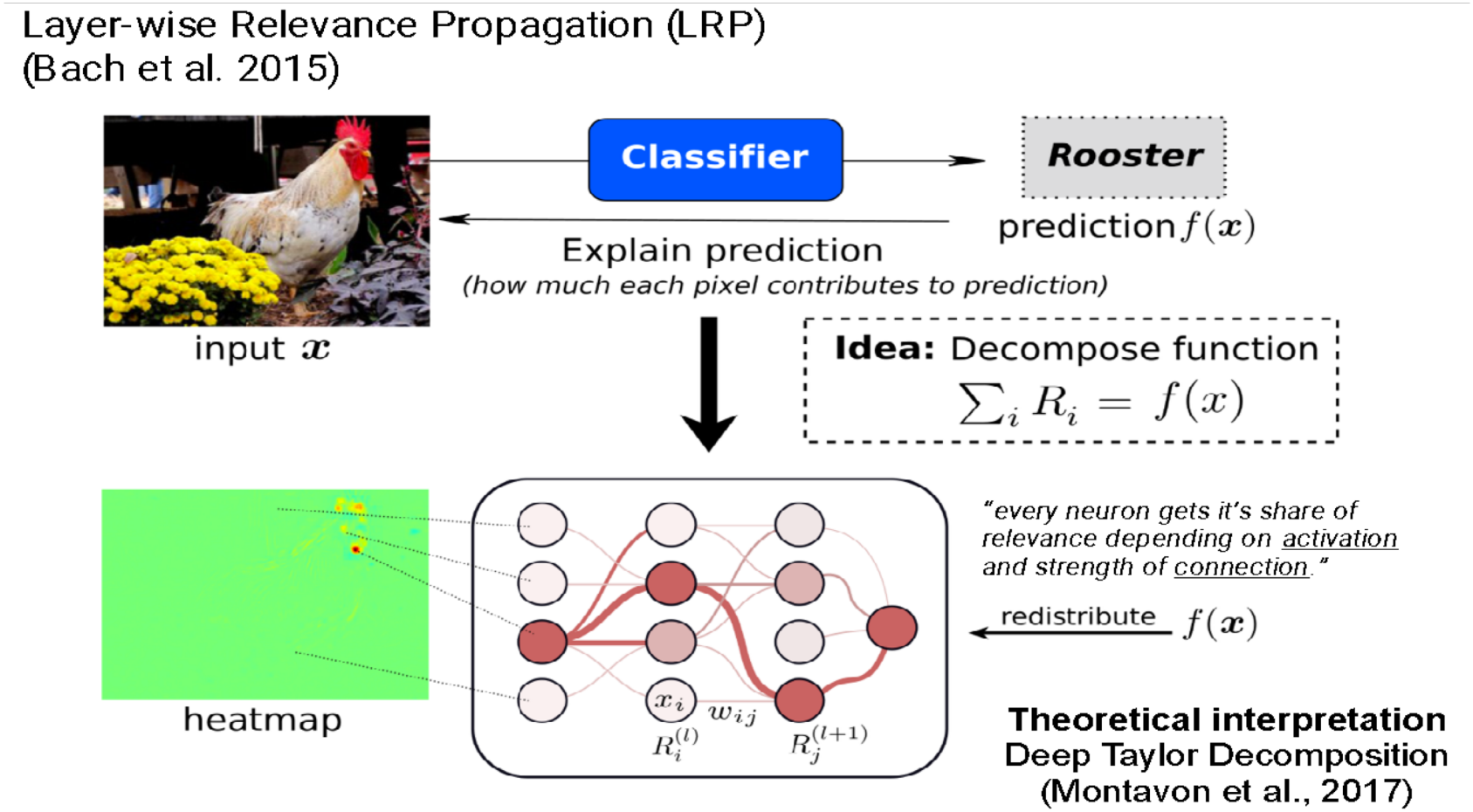
[Ancona et al. 2018]

Méthodes basées sur le Gradient



[Alber et al., 2019]

Layer-wise Relevance Propagation (LRP)



[Bach et al., 2015][Montavon, 2017]

Explication contrefactuelle

Explication contrefactuelle

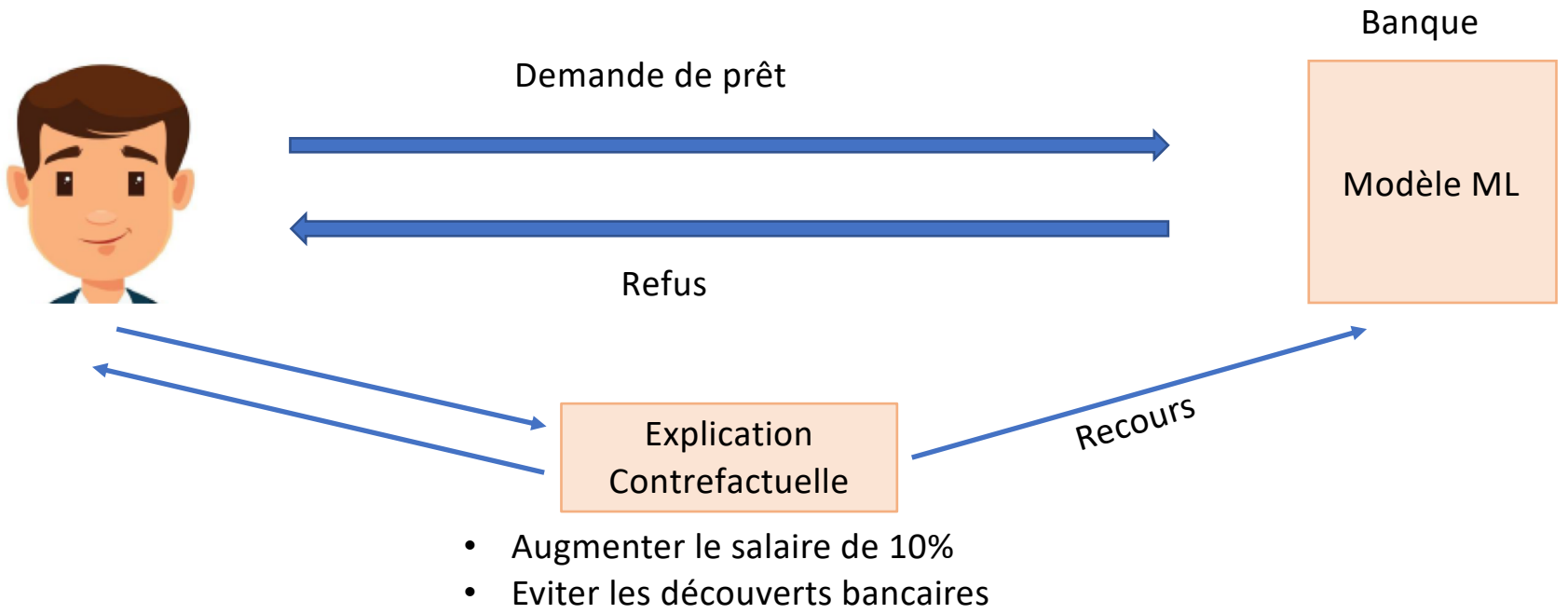
Identification des liens de causalité entre les entrées et les sorties

- Ne répond pas au « Pourquoi » de la décision
- Peut être utilisée quand la sortie désirée n'est pas obtenue

Identification des changements dans les variables d'entrée qui mènent au changement de la sortie vers une sortie désirée.

Explication contrefactuelle

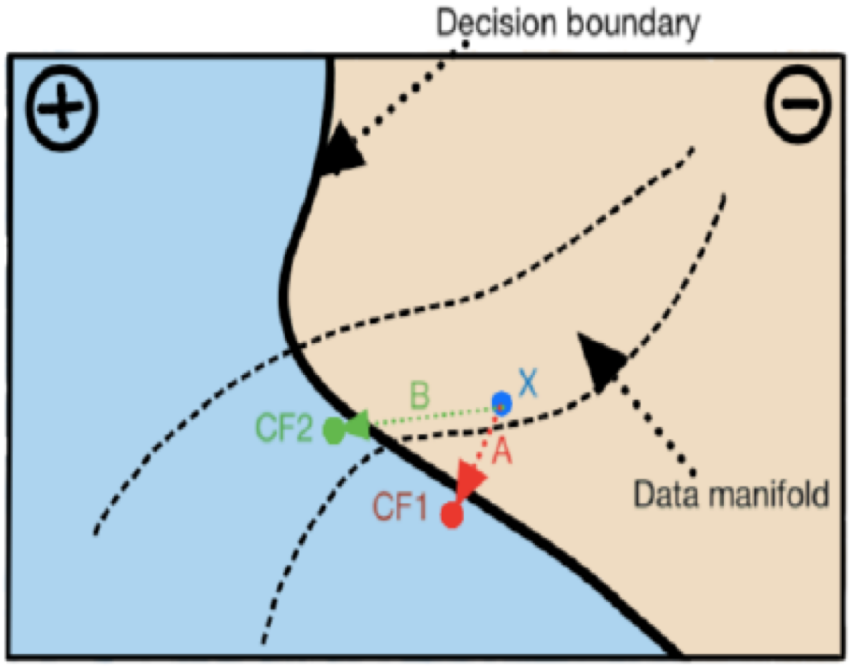
Identification des liens de causalité entre les entrées et les sorties



Explication contrefactuelle

(x, y) ----------->. (x’,y’)

Comment choisir la solution contrefactuelle ?



Explication contrefactuelle : contraintes

Problème d'optimisation [Verma et al., 2020][Guidotti, 2022]

$$\arg\min_{x' \in \mathcal{A}} \max_{\lambda} \lambda(f(x') - y')^2 + d(x, x') + g(x' - x) + l(x'; \mathcal{X})$$

- Validité : Minimiser la distance entre x' et x pour atteindre y'
- Parcimonie : Choisir un petit ensemble de caractéristiques
- Proximité : Choisir des solutions réalistiques dans l'espace des données d'apprentissage
- Actionnabilité : Considérer uniquement les caractéristique mutables
- Causalité : Maintenir des relations de causalité entre les caractéristiques

Explication basée sur les exemples/prototypes

Explication basée sur les exemples/prototypes

Exemples contrefactuels : Permettent de comprendre comment le modèle fait ses prédictions, ceci aide à expliquer les prédictions individuelles.

Exemples contradictoires : Utilisés pour tromper les modèles d'apprentissage automatique. L'accent est mis sur l'inversion de la prédiction et non sur l'explication.

Instances influentes : Instances de données d'apprentissage qui ont le plus influencé les paramètres d'un modèle de prédiction ou les prédictions elles-mêmes.

k-plus proches voisins : Sélection des k instances les plus proches de l'instance à expliquer.

Prototypes : Sélection d'instances ou d'abstraction d'instances représentatives des données

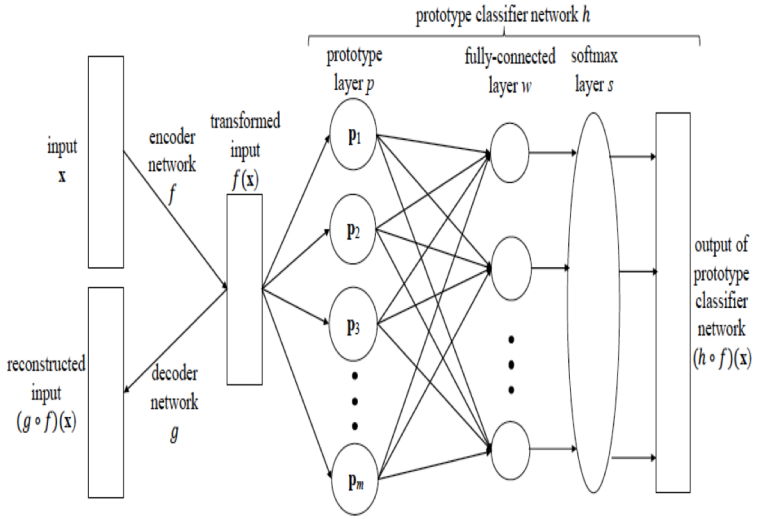
Explication basée sur les exemples/prototypes

Fournir des exemples d'individus/objets proches de l'entrée comme explication [Caruana, 2000][Li et al., 2018]

- Systèmes hybrides : Jumeaux ANN-CBR [Kenny & Keane , 2019]

- Méthodes basées sur le 'raisonnement à base de cas' [Li et al., 2018]

  - construction d'une couche de prototype à l'aide d'un auto-encodeur
  - Visualisation du prototype à l'aide du décodeur.

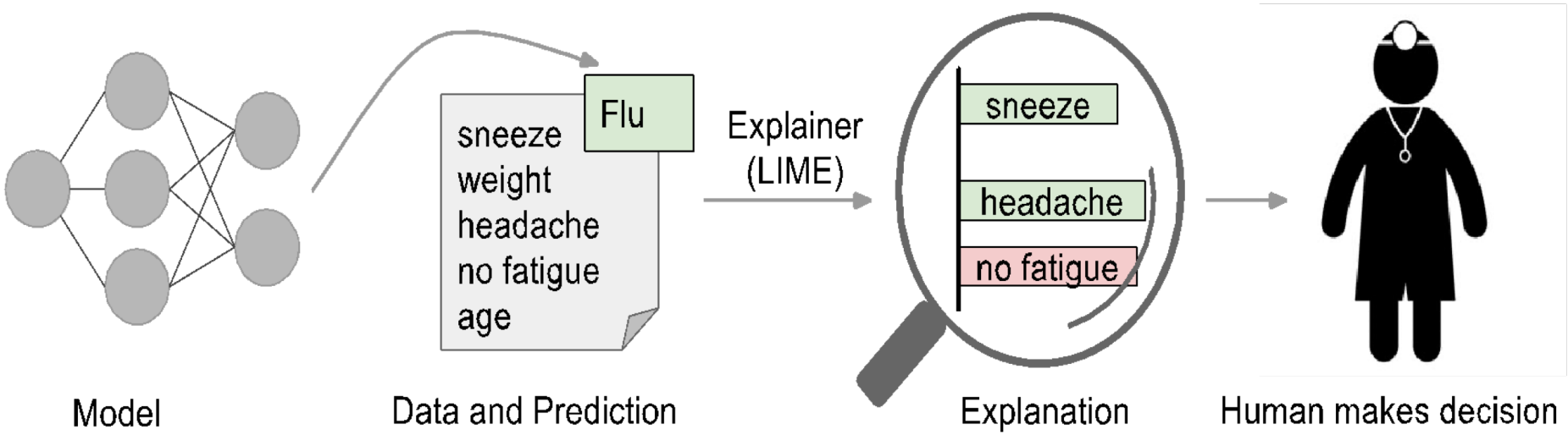


[Li et al., 2018]

Modèles interprétables de substitution

Fonctions linéaires : LIME (Local Interpretable Model-Agnostic Explanations)

LIME : peut expliquer les prédictions de n'importe quel classificateur en l'approximant localement avec un modèle linéaire interprétable.



[Ribeiro et al, 2016]

Fonctions linéaires: LIME

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Différentes familles d'explications G, fonctions de fidélité L et mesures de complexité Ω.
Focus sur les modèles linéaires parcimonieux (sparses) comme explications

$$g(z') = w_g \cdot z'$$

[Ribeiro et al, 2016]

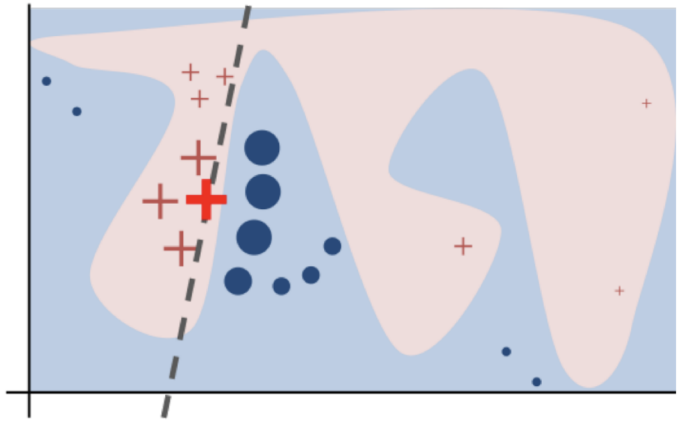


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

Extraction de règles à partir de réseaux de neurones

Représentation de comportement de réseaux de neurones comme règles de décision [Zilke et al., 2016][Hailesilassie et al, 2016][Zarlenga et al., 2021]

Les règles de décision sont compréhensibles et interprétables :

- Les décisions peuvent être expliquées par des règles
- Le modèle basé sur des règles peut être inspecté : découverte des relations entre les entrées et les sorties
- Les experts peuvent vérifier les règles critiques
- La règle révèle les attributs pertinents et les exemples d'apprentissage à partir desquels la règle a été apprise

The treatment for the patient x is class=1 R1:
(Age<55)(SystolicBloodPressure>137.027)
(DiastolicBloodPressure > 83.95) Class = 1 (387)

Types de règles extraites

Règles **SI-ALORS** : facilement compréhensibles. La forme générale de la règle SI-ALORS est : SI X ALORS Y=y

Règles **M-sur-N** : l'expression est satisfaite lorsque M sur N conditions sont satisfaits. La règle a la forme suivante : SI M de {N} ALORS Z

Arbre de décision : ce modèle classe une instance en commençant à la racine de l'arbre et en descendant jusqu'aux branches jusqu'à la fin.

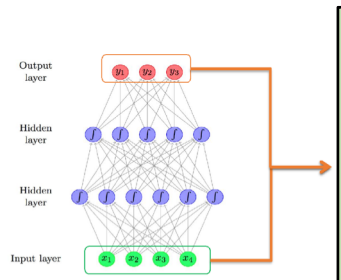
Extraction de règles à partir de réseaux de neurones

Approche pédagogique (Pedagogical approach): mappe la relation entrées-sorties

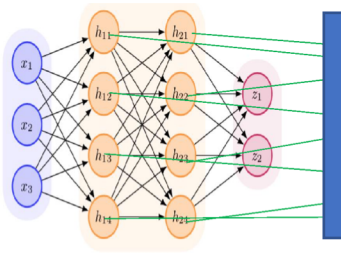
Approche décompositionnelle (Decompositional approach): analyse les connections des couches cachées

Approche Eclectique (Eclectic approach): combination of decompositional and pedagogical approaches.

[Hailesilassie et al, 2016]

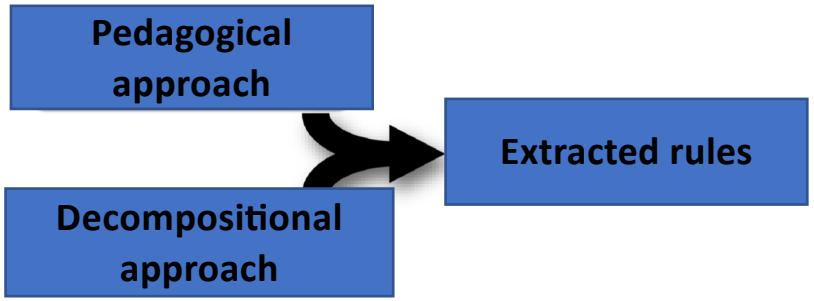


Rule 1: (sbp_v0 > 119) (dbp_v2 < 77) Class = 1 (410)
Rule 2: (dbp_v1 > 78) (sbp_v2 < 123) Class = 6 (720)
Rule 3: (sbp_v0 < 108) (dbp_v4 < 87) (sbp_v4 > 170) Class = 1 (513)
Rule 4: (sbp_v2 < 197) (sbp_v0 < 163) (dbp_v4 < 98) Class = 3 (382)
Rule 6: (sbp_v0 > 197) (sbp_v3 < 170) Class = 2 (25)

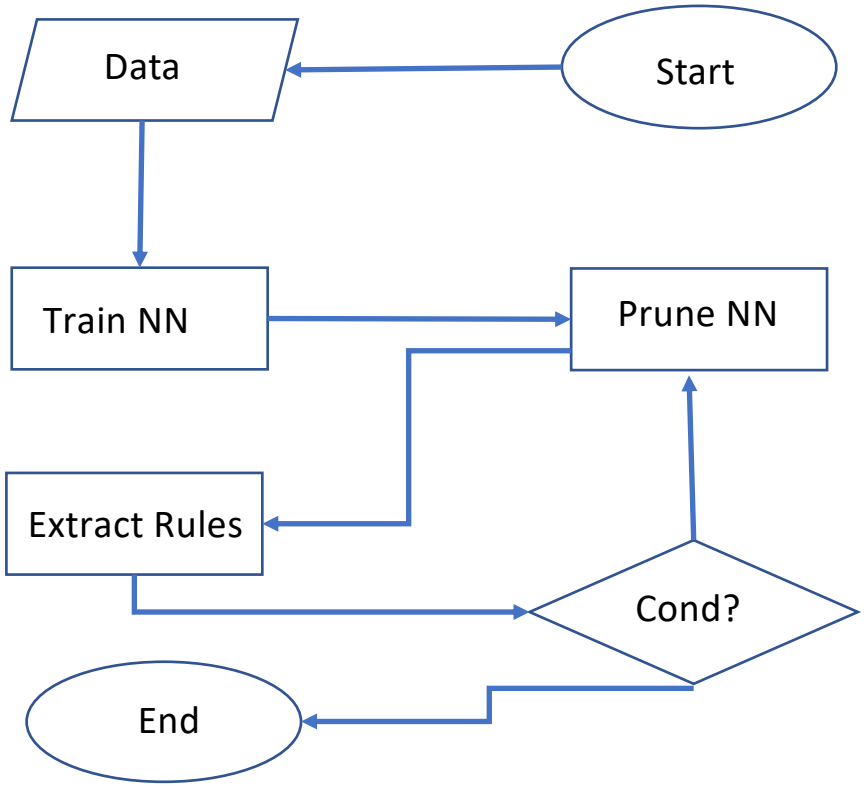


Rule extraction algorithm

Rule 1: (sbp_v1 > 110) (dbp_v3 < 81) Class = 3 (250)
Rule 2: (sbp_v1 > 71) (dbp_v2 < 113) Class = 1 (522)
Rule 3: (sbp_v1 > 110) (dbp_v3 < 81) Class = 3 (150)
Rule 4: (sbp_v1 > 110) (dbp_v3 < 81) Class = 3 (25)



Extraction de règles à partir de réseaux de neurones



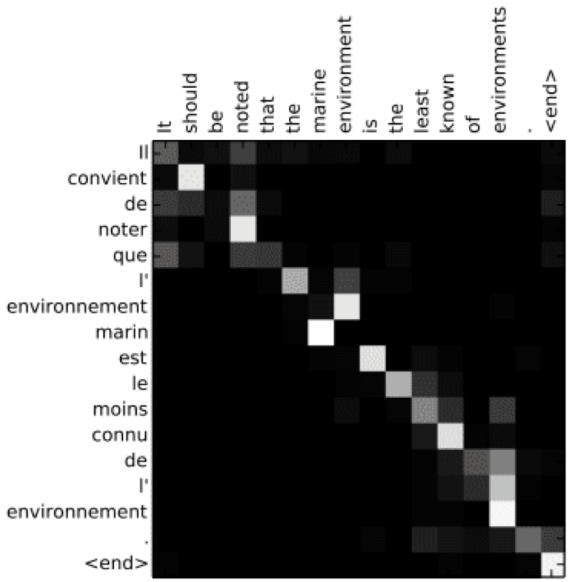
1. Préparation des données
2. Entrainement du RN
3. Elagage du RN
4. Extraction de règles de RN

Méthodes auto-explicantes (Ante-hoc)

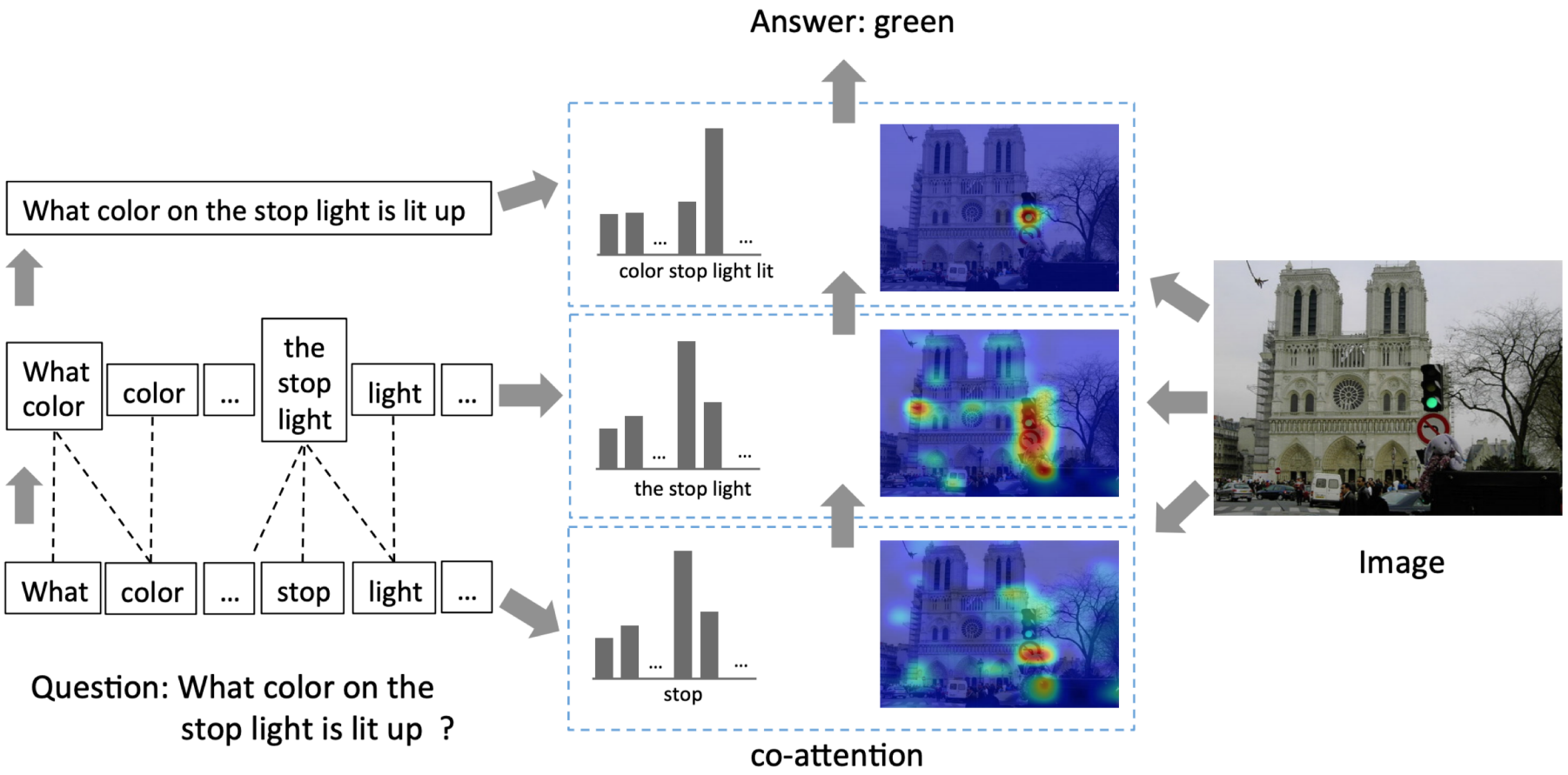
Mécanisme d’Attention

Introduit pour la première fois pour la traduction automatique (2015)
Composant essentiel des architectures neuronales pour plusieurs d'applications dans le traitement du langage naturel, la parole et la vision par ordinateur.

Mécanisme d’attention : [Vaswani et al., 2017][Wiegreffe & Pinter, 2021][Chaudhari e 2021]
Permet au modèle de se focaliser sur les concepts/ caractéristiques importants



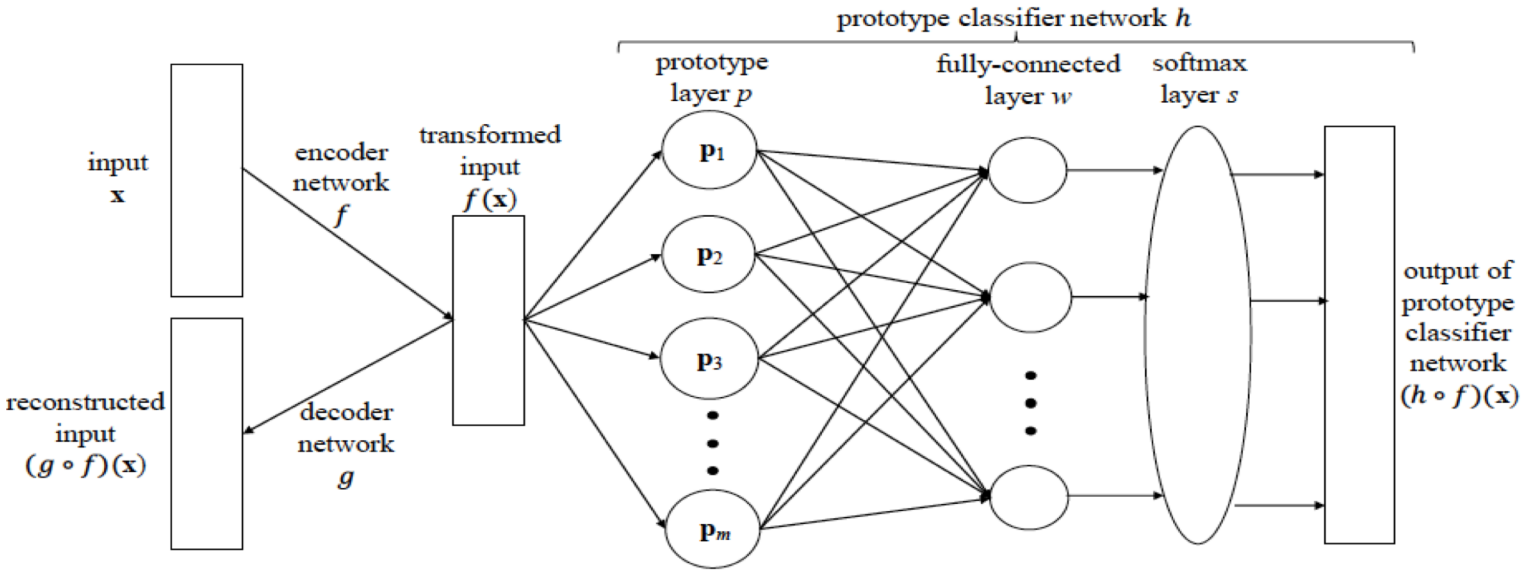
Mécanisme d'attention : Application



[Chaudhari et al., 2021]

Méthode basée du le raisonnement à base de cas

Exemple de méthode auto-expliquante basée sur le raisonnement à base de cas

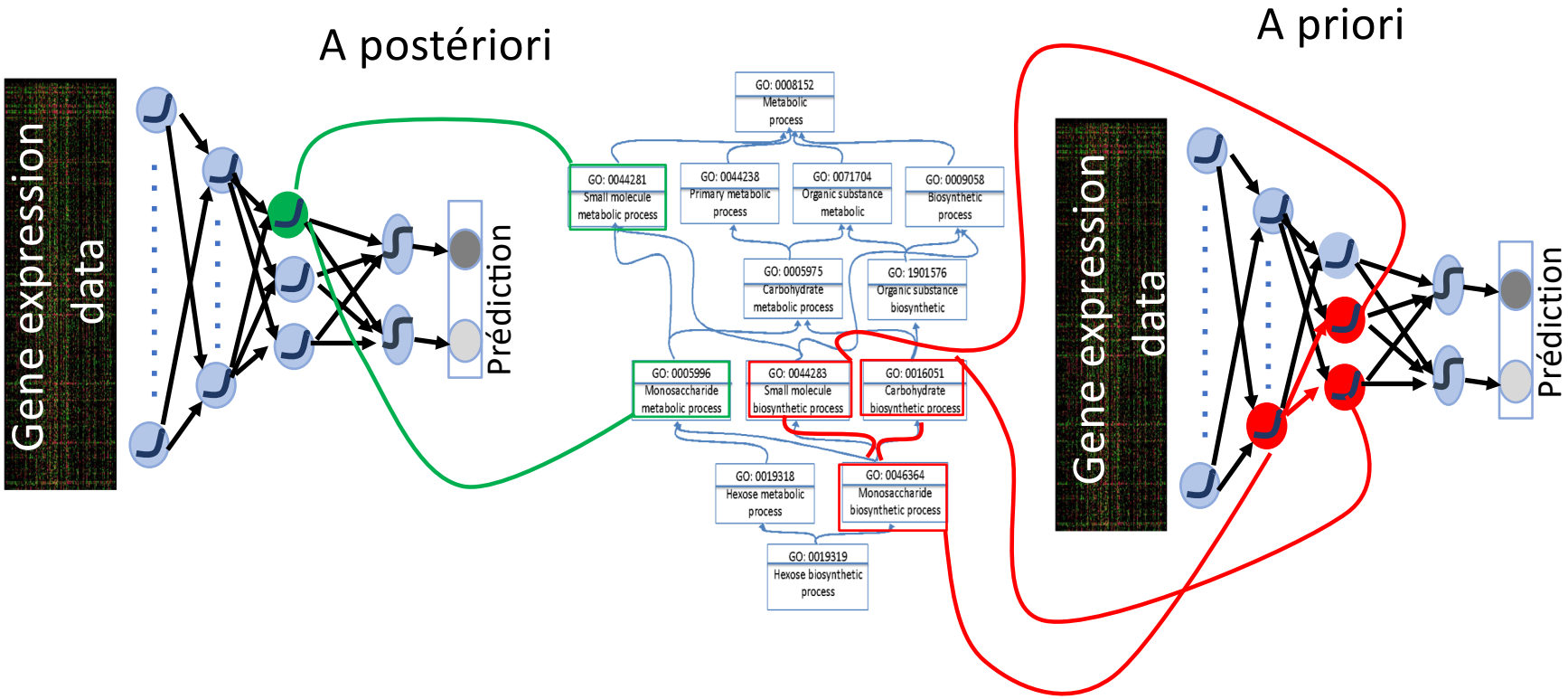


[Li et al., 2018]

Intégration de connaissances

Intégration de connaissances

Connaissances Sous forme de graphes



[Hanczar et al., 2019]…

[Bourgeais et al., 2020…]

Intégration de connaissances

Graphes de connaissances : Utilisation des réseau de neurones à base de graphes « Graph Neural Networks (GNN) »

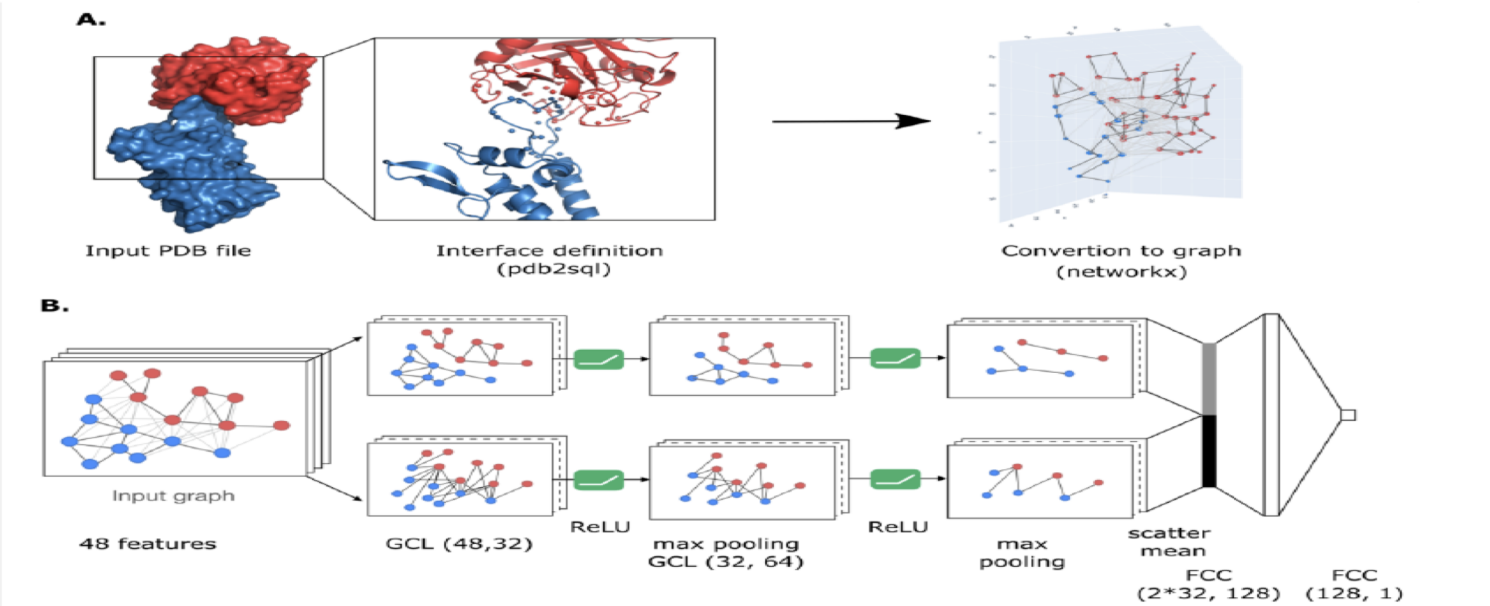


Fig. 1. Overview of the DeepRank-GNN framework. (A) DeepRank-GNN identifies interface residues and converts them into an interface graph. Internal edges are defined between residues from the same chain having heavy atoms within a 3 Å distance cutoff from each other, while external edges are defined between residues from different chains having heavy atoms within the 8.5 Å cutoff. (B) Architecture of a Graph Interaction Network (GINet). The graph representation of a PPI is split into 2 sub-graphs, i.e. the internal graph connecting atoms from the same protein and the external graph connecting atoms from distinct proteins. The 2 sub-graphs are sequentially passed to 2 consecutive convolution/activation/pooling layers. The two final graph representations are flattened using a scatter mean operation and merged before applying two fully connected layers. GCL: graph convolution layer; FCC: fully connected layer.

[Réau et al. 2021][Renaud et al. 2021] [Lin et al., 2020][Zhou et al., 2020]

Interprétabilité pour garantir l'équité

Équité (Fairness) est un sous domaine de l'interprétabilité [Calmon et al.,2017][Yu et al., 2018] [Linardatos et al.2021]…

➢ Plusieurs méthodes ont été proposées pour :

- Protéger certaines minorités du biais social et assurer l'allocation des ressources
- Éviter les discriminations

➢ Ces méthodes concernent

- La manipulation des données avant l'apprentissage
- La modifications de l'algorithme d'apprentissage
- L'ajustements post-hoc des modèles.

Transparence et Explicabilité en IA symbolique

L’IA symbolique est utilisée actuellement pour :

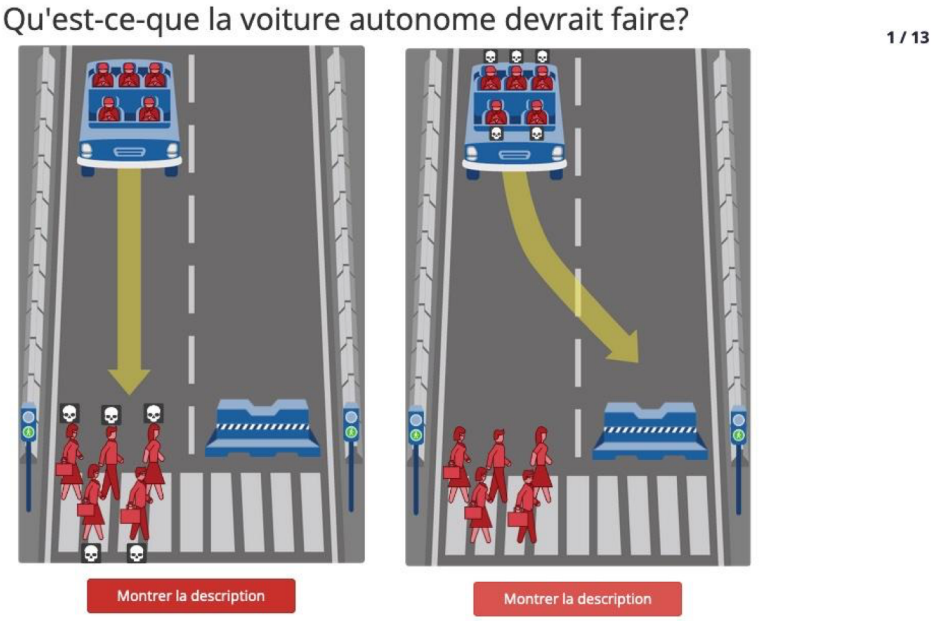
- Rendre les approches numériques plus transparentes et explicables
- Implémenter et expliciter des principes éthiques :
  - Dilemmes éthiques
  - Frameworks d’aide à la décision éthique monoagent
  - Frameworks d’aide à la décision collective

Dilemmes éthiques

Exemples de dilemmes éthiques :

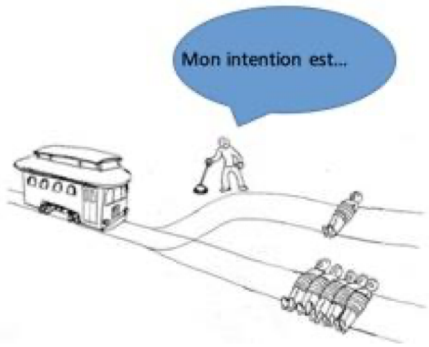
- GenEth ethical dilemma analyzer ([Anderson & Anderson, 2014]) : dialogue avec des éthiciens
- Moral décision Making (MoralDM) and analogy [Blass & Forbus 2015]: Instantanéité de la décision morale
- Moral Machine project : questionnaire http://moralmachine.mit.edu/) : classification des réponses selon des valeurs :

  - Sauver plus de vies
  - Protéger les passagers
  - Faire respecter la loi
  - Éviter une éventuelle intervention
  - Selon genre
  - Selon la race
  - Selon l'âge
  - Selon certaines valeurs sociales



Frameworks monoagent explicitant l’éthique

- Vérification formelle du respect de valeurs morales dans les SMA [T. Vallée et al. 2018]
  - ➢ Montrer qu'un agent respecte une règle éthique selon un système de valeurs
- Combining Game theory and Machine Learning for moral decision [Conitzer et al. 2017]
- Event-Based and Scenario-Based Causality for Computational Ethics [Fiona et al. 2018]
  - ➢ Représentation de la causalité des actions, matrice de causalité (justifier, inhiber)
- Formal definitions of Blameworthiness, Intention and Moral Responsibility [Halpern & Kleiman-Weiner 2018]
  - ➢ Degré de responsabilité morale :
  - Expliciter l'intention (si un agent avait fait telle action, cela aurait produit tel effet)
  - Probabilité sur les liens et l'effet des actions



- Responsible Autonomy (Virginia Dignum 2017)

Délibération éthique	Implications pour la conception du système d'IA	ART Accountability, Responsibility, Transparency
Contrôle humain	• Raisonnement temps réel • Sensibilité utilisateur aux situations • Capacités d'explication • Fournir les états internes à l'utilisateur sous un mode compréhensible	• Délégué à l'utilisateur
Régulation	• Lien formel des valeurs aux normes et aux comportements • Définir les institutions pour la surveillance et le contrôle • Raisonnement moral peut être effectué off-line	• **A** : institutionnelle • **R** : institutionnelle • **T** : Système (selon le besoin)
AMA Agents Moraux Artificiels	• Lien formel entre :valeurs, normes et comportements • Définir les règles de raisonnement • Apprentissage supervisé de la moralité • Raisonnement temps réel	• **A** : institutionnelle (par explication) • **R** : institutionnelle (par délibération) • **T** : Système (selon le besoin)

 Deux questions fondamentales :

- Qui est responsable de la décision (niveau d'autonomie de la décision) ?
- Comment les décisions dépendent-elles de différentes valeurs morales et sociales ?

Frameworks collectifs explicitant l’éthique

- Dealing With Ethical Conflicts In Autonomous Agents And Multi-Agent Systems [Bellonni et al. 2015]
  - Percevoir les dilemmes, attribuer une causalité et une responsabilité, expliquer et justifier ses actions et décisions ainsi que celles des autres agents, vérifier formellement l'éthique d'un agent.
- Ethical Judgment of Agents' Behaviors in Multi-Agent Systems [Cointe et al. 2016]
  - Représentation explicite de l'éthique normative, choix entre plusieurs principes éthiques (notamment en cas de dilemme), conciliation des désirs, de la morale et des capacités (BDI)
- Coalition-based multiagent approach for implementing ethics: an assistive application case-study [Abchiche-Mimouni & Colle 2019]
- Elessar: Ethics in Norm-Aware Agents [Ajmeri & Guo 2020]
  - Agent SIPA (Socially Intelligent Personnal Agent) : agit en conformité avec les normes sociales. Exception : conflit avec les valeurs préférées de l'utilisateur. Approche multiagent et multicritères pour identifier un consensus.

Exemple (Elessar) : partager sa localisation via un réseau social

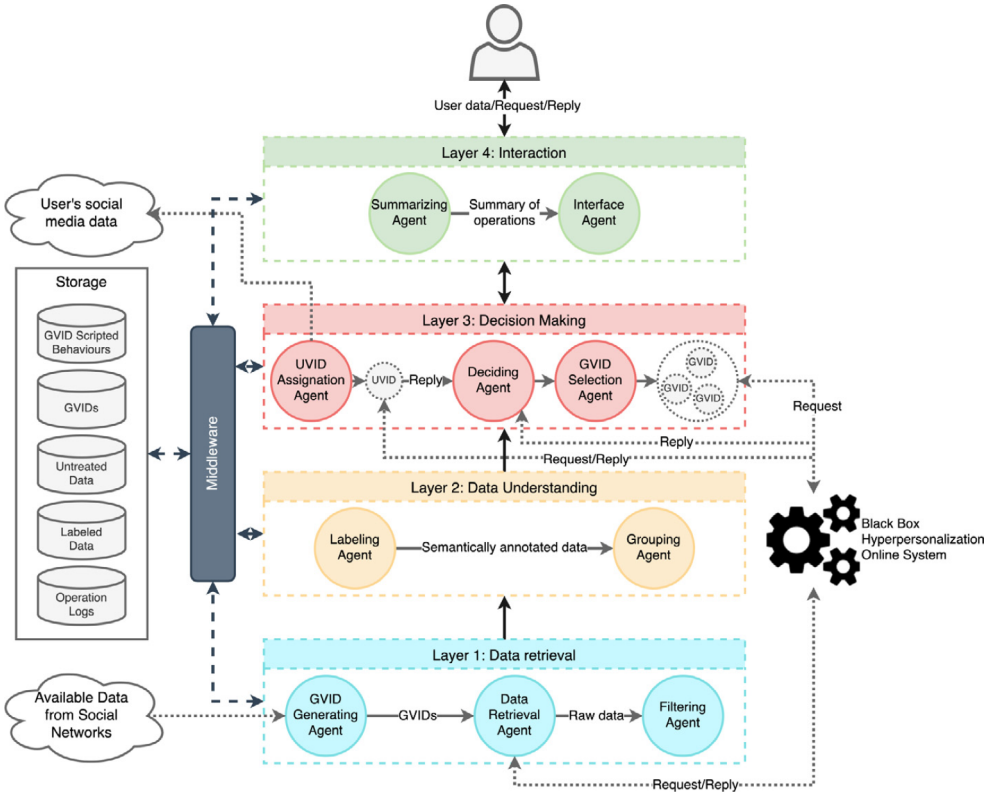
Frank, est un étudiant qui trouve du plaisir (valeur) à utiliser Gimli. Il apprécie également la reconnaissance sociale. Frank s'est engagé (norme) envers sa mère Grace à partager sa localisation avec elle.
Cela satisfait son engagement envers Grace mais porte atteinte à sa vie privée.
Les valeurs de plaisir, de reconnaissance et de sécurité de Frank peuvent être favorisées ou défavorisées selon l'endroit où il se trouve et la politique de partage choisie.

Olympiade. Frank se rend à Yale pour participer à des olympiades de maths. En partageant publiquement le fait que Frank est à Yale pour les Olympiades, Gimli :
Satisfait l'engagement (norme) de Frank envers sa mère ;
Favorise le plaisir (valeur) et la reconnaissance sociale (valeur) de Frank ;
Mais :
Compromet la sécurité et la vie privée de Frank (valeur).
Partager qu'avec des amis satisfait l'engagement de Frank envers Grace et échange le plaisir et la reconnaissance contre la sécurité et la vie privée.

Times Square. Frank visite Times Square et y rencontre son oncle Harold. Harold accorde de l'importance à la vie privée et interdit (norme) à Frank de partager publiquement l'emplacement de Harold. Gimli partage seulement avec Grace le fait que Frank est à Times Square avec Harold, satisfaisant ainsi les normes d'engagement et d'interdiction applicables.
Ainsi, Gimli favorise la vie privée de Harold plutôt que le plaisir et la reconnaissance sociale de Frank.

- Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective [Burton et al. 2020]
  - Expliciter l'intention des fonctionnalités dans les spécifications
  - Avoir le control nécessaire sur l'action en vue de refléter intentions/désirs de la personne
  - Avoir eu les connaissances pertinentes pour comprendre l'action et ses possibles conséquences

- A hierarchical multi-agent architecture based on virtual identities to explain black-box personalization policies [Amador-Domínguez et al. 2021]



- Explaining Multi-Criteria Decision [Labreuche & Fossier 2018]
  - Déterminer la priorité entre navires en termes de risques d'activités illégales (incohérence entre données du système d'identification et la détection radar, suspicion de trafic de drogue/d'êtres humains, vitesse, proximité à la côte)

Quels sont les attributs les plus importants (en moyenne) ?	Interprétabilité
Quels changements (valeurs d'attributs) augmenteraient le niveau de priorité de manière la plus significative ?	Sensibilité
Pourquoi le niveau de priorité de ce navire est-il élevé ? Pourquoi le niveau de priorité de ce navire a-t-il considérablement augmenté au cours des dernières minutes ?	Explication

- Explainable Multi-Agent Systems Through Blockchain Technology [Calvaresi et al. 2019]
  - SMA explicable et de confiance grâce à une authentification par la Blockchain
  - Mechanism d'appartenance
  - Protocol consensuel pour le maintien d'un registre partagé, immutable et transparent
  - Importance de la confiance (Trust) pour l'explicabilité

> "*Trust is the subjective probability by which an agent A expects that another agent B performs a given action on which its welfare depends*".

QUESTIONNAIRE 2

Partie III
Démonstration : utilisation de l'argumentation multiagents

Diverses approches hybrides combinent IA numérique et IA symbolique pour rendre les systèmes intelligents plus transparents et plus explicables.

L'Argumentation

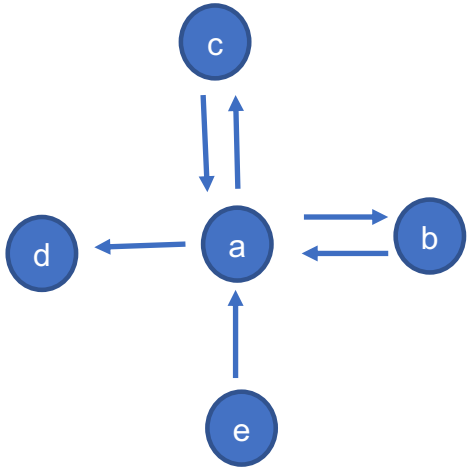
Généralités sur l'argumentation

- Raisonnement non monotone
- Prise de décision malgré des controverses

- Argumentation Dialogue Games (ADG) est à l'origine des dialogues argumentatifs (Levin & Moore 1977)
- Framework d'argumentation introduit par Dung en 1995

Framework d'argumentation F = <*A*, *R*>

A : ensemble d'arguments
$R \subseteq A \times A$: une relation d'attaque
$(x, y) \in R$: x attaque y



L'argumentation

Notions Fondamentales : Simari & Rahwan Ed. 2009

- Différents types de relations entre arguments
- Diverses sémantiques d'acceptabilité
- Framework d'Argumentation
- Argumentation structurée
- Gestion des préférences

Notions avancées : Dionysios Kontarinis, thèse de doctorat 2014

- Dialogues argumentatifs
- Systèmes de persuasion
- Protocoles pour le dialogue
- Dialogues multilatéraux
- Argumentation multi-agents

Argumentation Multiagents

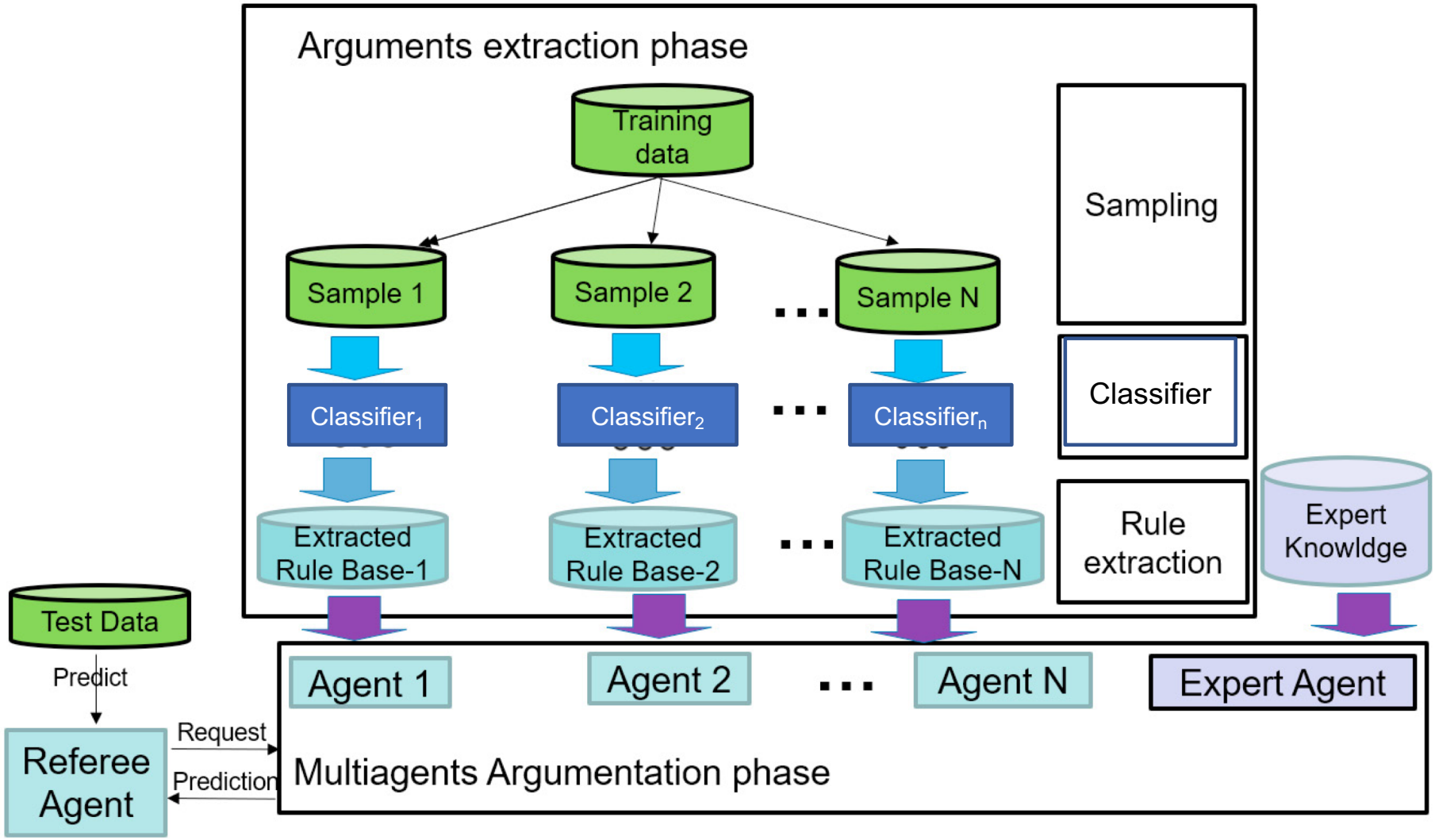
- Gestion des préférences entre agents
- Prise de décisions collectives sur base d'arguments et contre arguments
- Prise de décision en environnement incertain
- Possibilité de révision des connaissances (perceptions de l'environnement…)
- Formes d'interactions complexes (convaincre l'utilisateur ou les autres agents)
- Protocoles d'interaction dédiés
- Intégration de données, connaissances et modèles hétérogènes
- …

Approches hybrides basées sur l'argumentation

Quelques exemples :

- Explainable AI meets persuasiveness: Translating reasoning results into behavioral change advice, AI in Medicine Journal 2020
- Resolving Conflicts in Clinical Guidelines using Argumentation, AAMAS 2019
- Towards an Argumentation System for Supporting Patients in Self-Managing their Chronic Conditions, AAAI Workshop 2018
- Towards a Transparent Deep Ensemble Method Based on Multiagent Argumentation, EXTRAAMAS, 2019

Approche hybride pour expliquer une méthode d'ensemble



Cas d'étude

Prédiction de la maladie « Sepsis » à partir des variables :

Cliniques

HR Heart rate (beats/min)

O2Sat Pulse oximetry (%)

SBP Systolic BP (mm Hg)

MAP Mean arterial pressure (mm Hg)

Resp Respiration rate (breaths/min)

Démographique

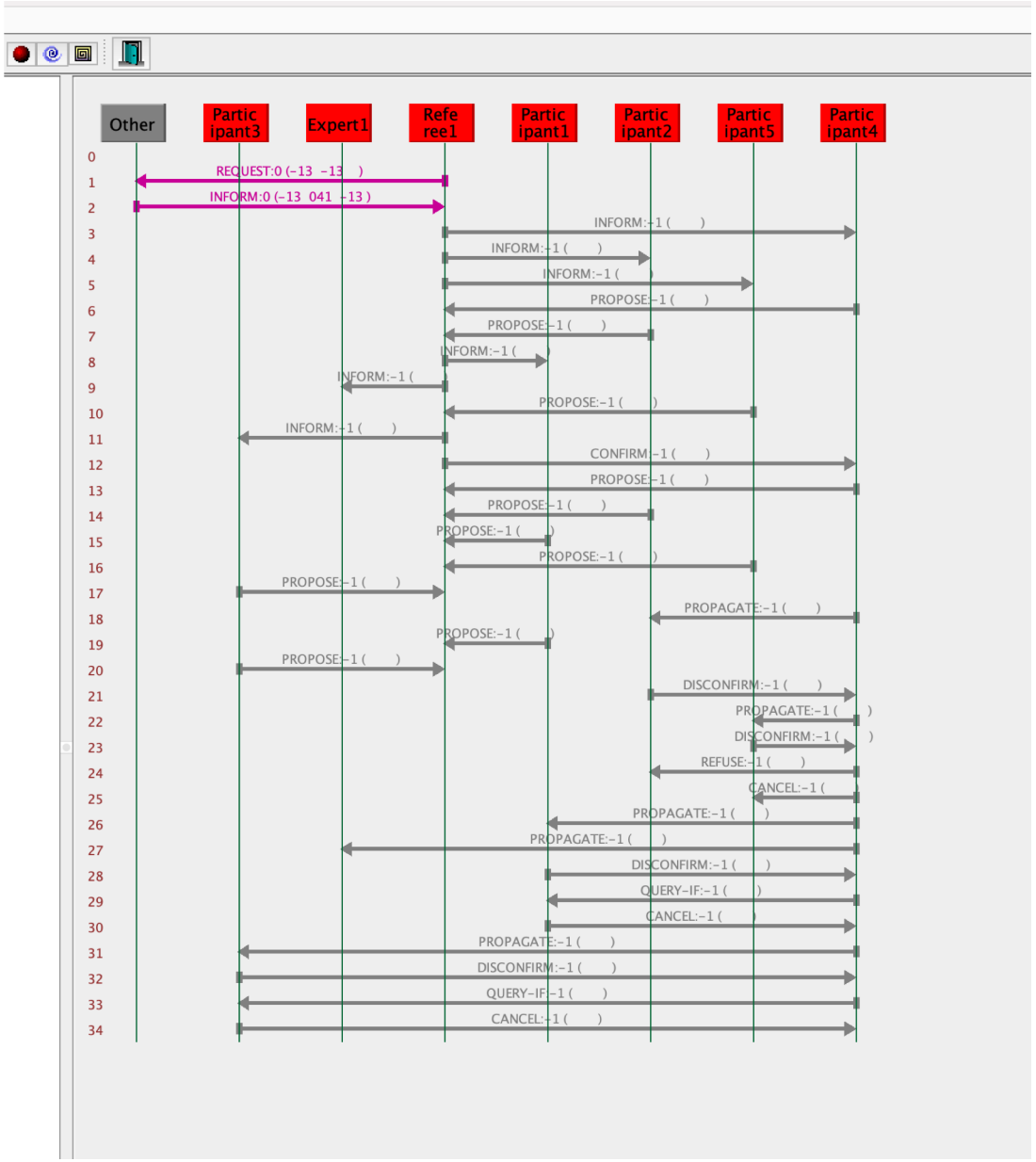
Age (yr)

Gender Female (0) or male (1)

https://www.biomerieux.com/corp/fr/blog/sepsis/quick-facts-about-sepsis-.html

Démonstration

HR, O2Sat, SBP, MAP, Resp



Démonstration disponible en ligne

- Une seul instance de la démo à la fois sur chaque machine virtuelle !
- Avant de relancer :
  - -) Vider le dossier des règles générées par le ML: ressources/data/rules_tree_<i>.clp (i>0)
  - -) Conserver les fichiers official_guidelines !
  - -) Supprimer d'éventuels fichiers cachés (Ex. .DStore) créés par l'OS

https://demo-ethos.ibisc.univ-evry.fr/demo1

Dans le terminal, il faut taper :

export JADE_PORT=$UID

Puis :

Démo avec paramètres par défaut : ./rdemo-1_tuto.sh

Démo avec paramètres : ./rdemo_tuto.sh <nbr agents> <nombre individus>

Vidéo illustration : https://www.youtube.com/watch?v=keBFgUtrWAU

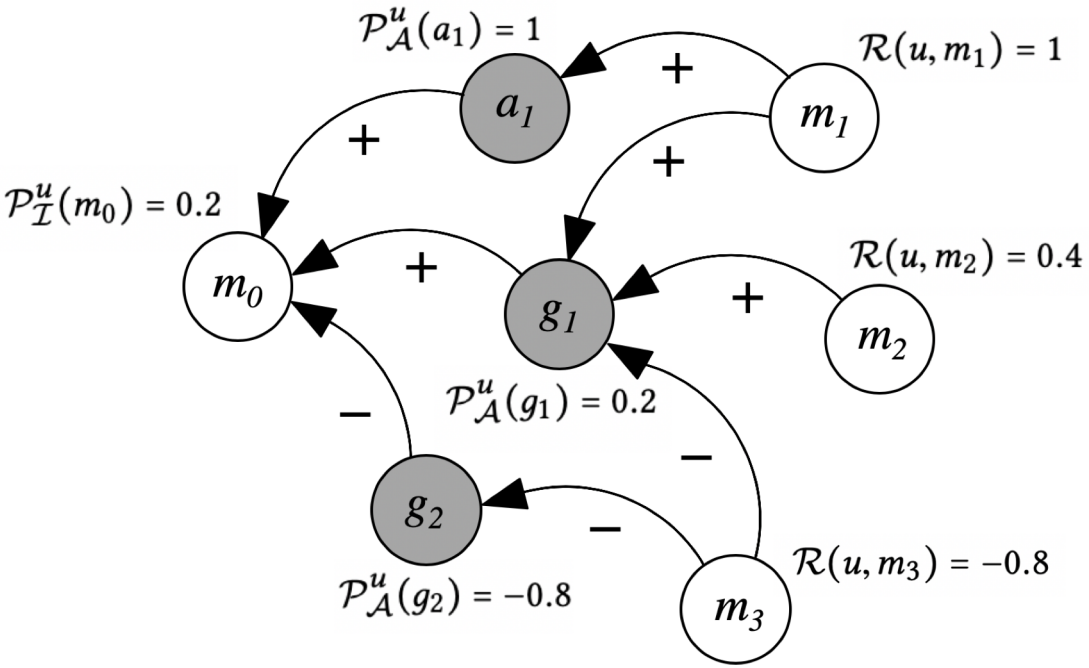
Autres approches hybrides basées sur l'argumentation

- Explainable and Argumentation-based Decision Making with Qualitative Preferences for Diagnostics and Prognostics of Alzheimer's Disease [Zeng et al. 2020]

  - Problème applicatif : déterminer si un patient est à haut risque par rapport à la maladie d'Alzheimer. Le risque est lié à la génétique et aussi variable en fonction du sexe (plus important pour une femme).
  - Framework de décision formel : Bases de connaissances, buts, décision et attributs d'un agent
  - Préférences qualitatives : relativiser l'importance (priorités) des buts ou d'attributs d'un agents :
    - Décisions reliées aux objectifs
    - Gestion des priorités des objectifs
    - Priorités apprises à partir des données patients (différents ordres de priorités pour différents contextes)

- Argumentation as a Framework for Interactive Explanations for Recommendations [Rago et al. 2020]

- Système de recommandations
- Argumentation bipolaire
- Extraction d'arguments pour les recommandations tirées des évaluations prédites
- Relation entre arguments orientés en indiquai la note prédite d'un élément affecte la note d'un autre



- Applying Metalevel Argumentation Frameworks to Support Medical Decision Making [Kökciyan et al. 2021]

- Argumentation structurée
- Méta-argumentation
- Hypertension artérielle