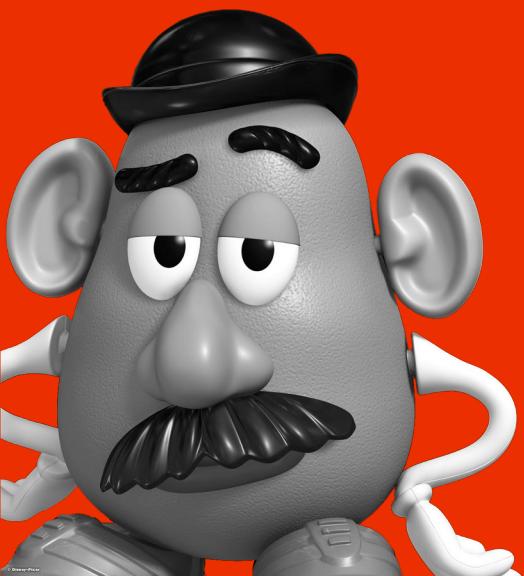


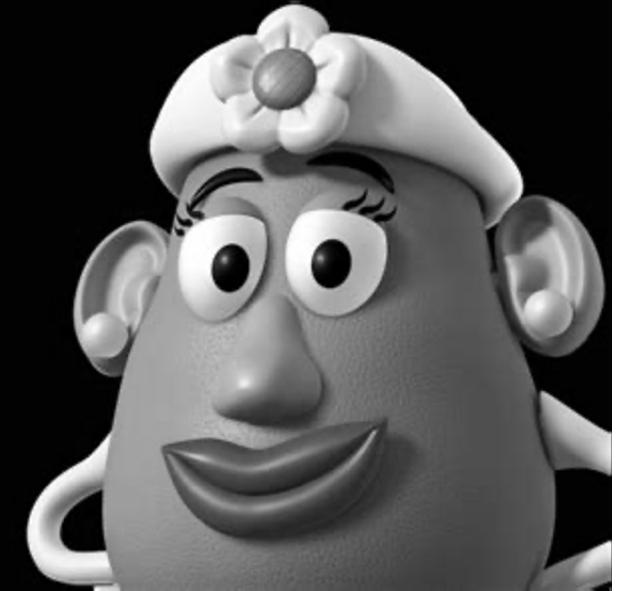
INTERPRÉTABILITÉ ou EXPLICABILITÉ : POTAYTO, PATAHTO ?



Sandrine Blais-Deschênes (elle)

PLAN DE LA PRÉSENTATION

- Introduction
- Définitions
- Pièges de l'explicabilité
- Gagner en interprétabilité
- Perdre de l'interprétabilité
- Conclusion



INTRODUCTION



Potayto, patahto

- Let's call the whole thing off
 - George et Ira Gershwin
- Différence, distinction ou correction négligeable, triviale ou non importante
- « Même différence » (*Same difference*)
- « Bonnet blanc, blanc bonnet »
 - Présentées comme différentes
 - Mais très similaires

https://en.wikipedia.org/wiki/Let%27s_Call_the_Whole_Thing_Off

<https://idioms.thefreedictionary.com/potato+potato>

https://en.wiktionary.org/wiki/potayto,_potahto

https://fr.wiktionary.org/wiki/bonnet_blancl,_blanc_bonnet

<https://www.expressio.fr/expressions/c-est-bonnet-blanc-et-blanc-bonnet>



NON !

- Interprétabilité et explicabilité
 - Semblent similaires
 - En réalité différents

INTRODUCTION

Apprentissage automatique

Individus	Age			Age % 6	
Maxime	33	7,0	Oui	3	
Camille	54	2,5	Non	0	
Félix	26	1,5	Non	2	
Dominique	72	0,5	Oui	0	

INTRODUCTION

Apprentissage automatique

Exemples [

Individus	Age			Age % 6	
Maxime	33	7,0	Oui	3	
Camille	54	2,5	Non	0	
Félix	26	1,5	Non	2	
Dominique	72	0,5	Oui	0	

INTRODUCTION

Apprentissage automatique

Attributs (*features*)

Exemples [

Individus	Age			Age % 6	
Maxime	33	7,0	Oui	3	
Camille	54	2,5	Non	0	
Félix	26	1,5	Non	2	
Dominique	72	0,5	Oui	0	

INTRODUCTION

Apprentissage automatique supervisé

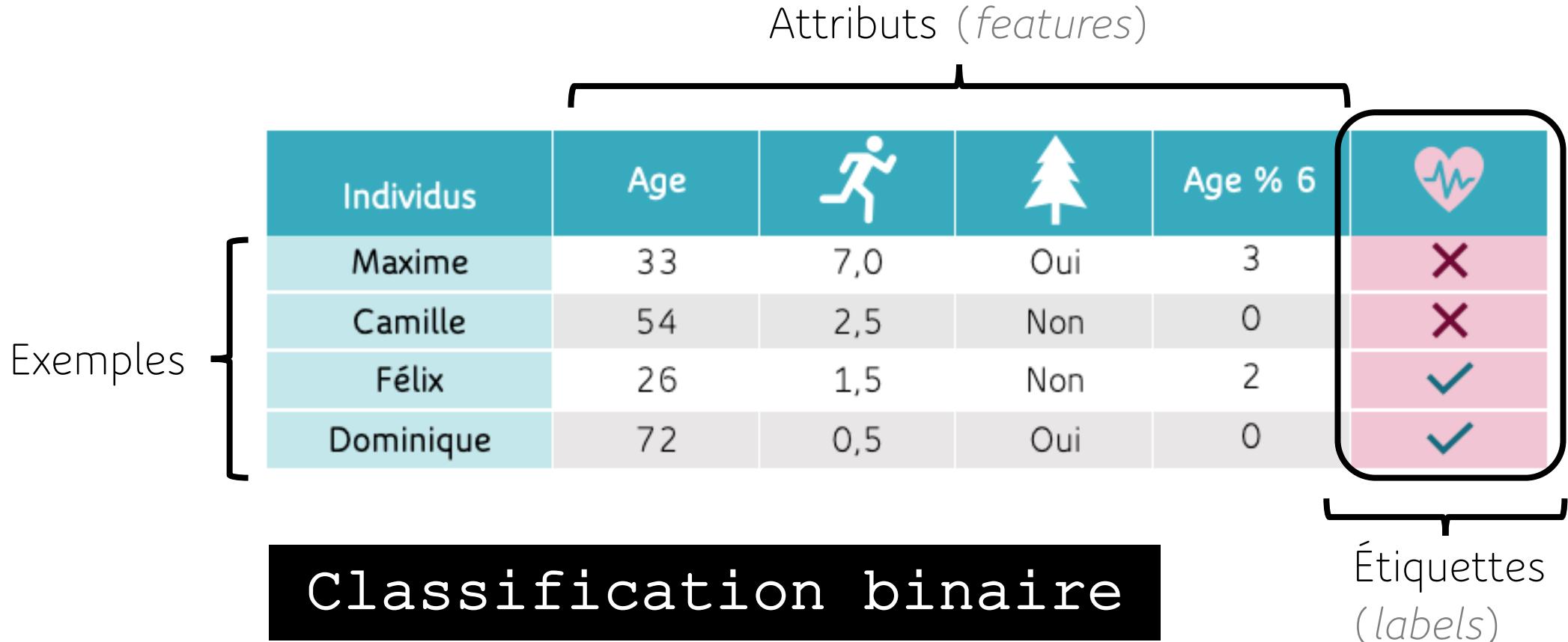
Attributs (*features*)

Exemples [

Individus	Age	🏃	🌲	Age % 6	Etiquettes (labels)
Maxime	33	7,0	Oui	3	X
Camille	54	2,5	Non	0	X
Félix	26	1,5	Non	2	✓
Dominique	72	0,5	Oui	0	✓

INTRODUCTION

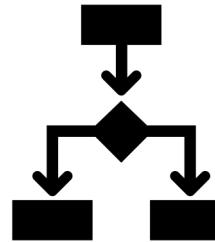
Apprentissage automatique supervisé



INTRODUCTION

Prédiction

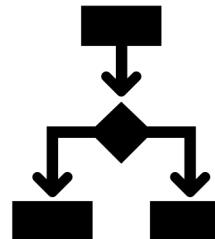
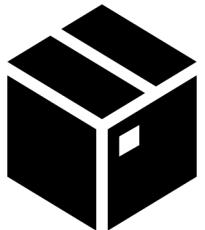
Dominique	72	0,5	Oui	0
-----------	----	-----	-----	---



INTRODUCTION

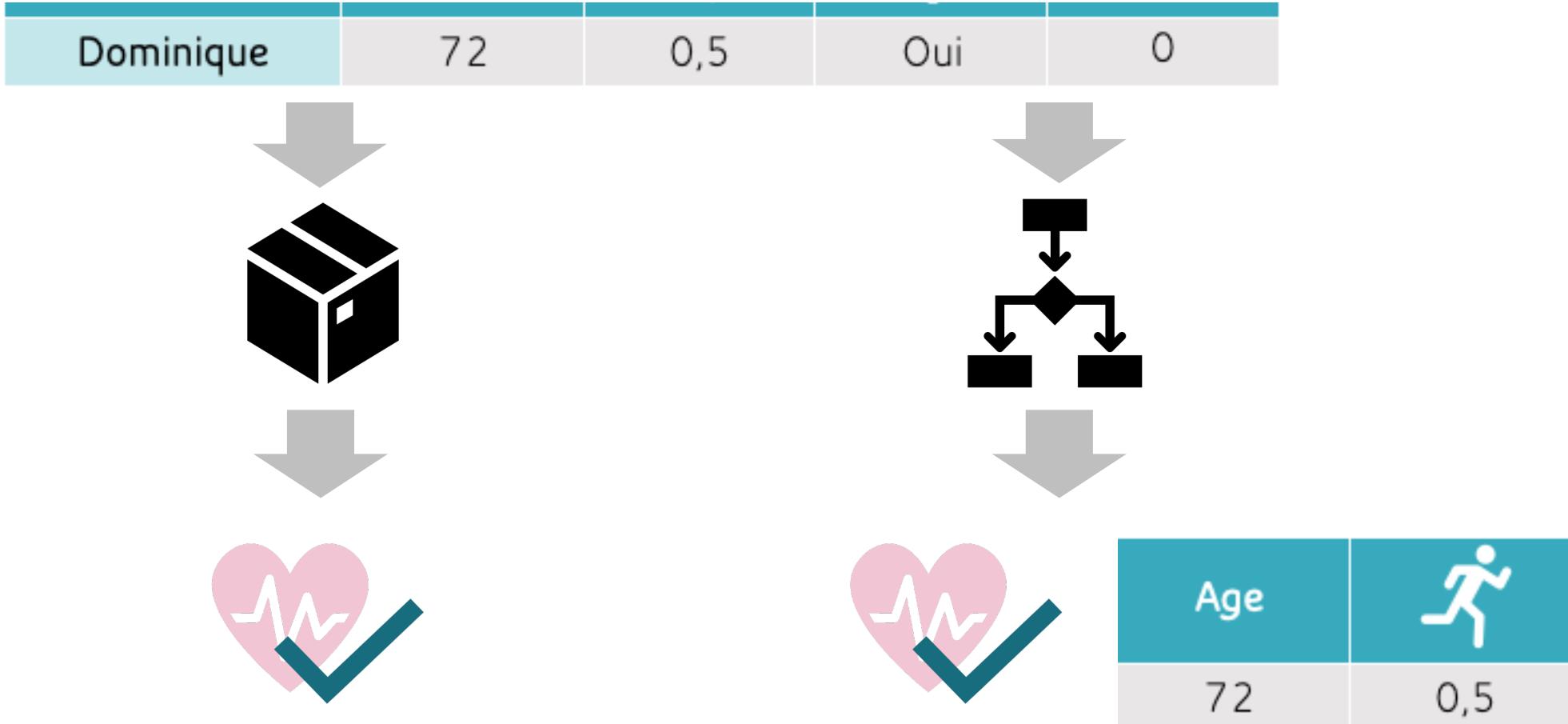
Prédiction

Dominique	72	0,5	Oui	0
-----------	----	-----	-----	---



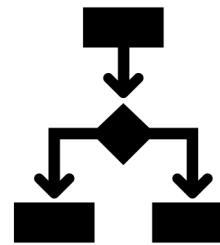
INTRODUCTION

Prédiction



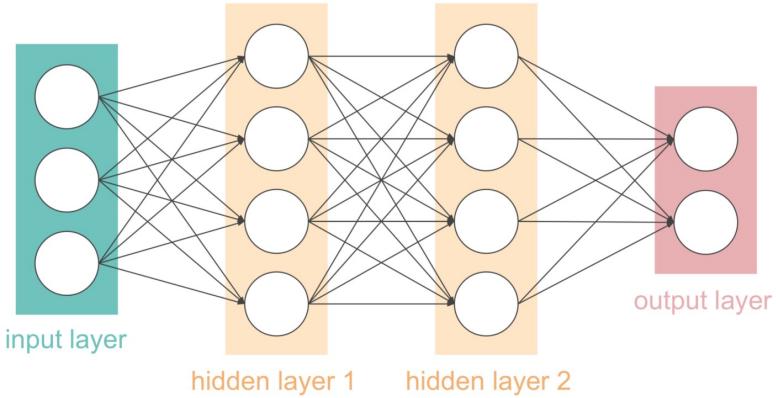
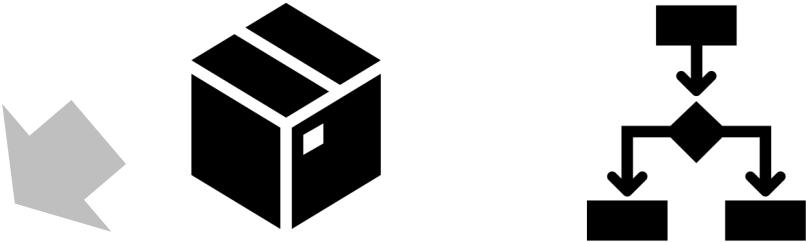
INTRODUCTION

Explication



INTRODUCTION

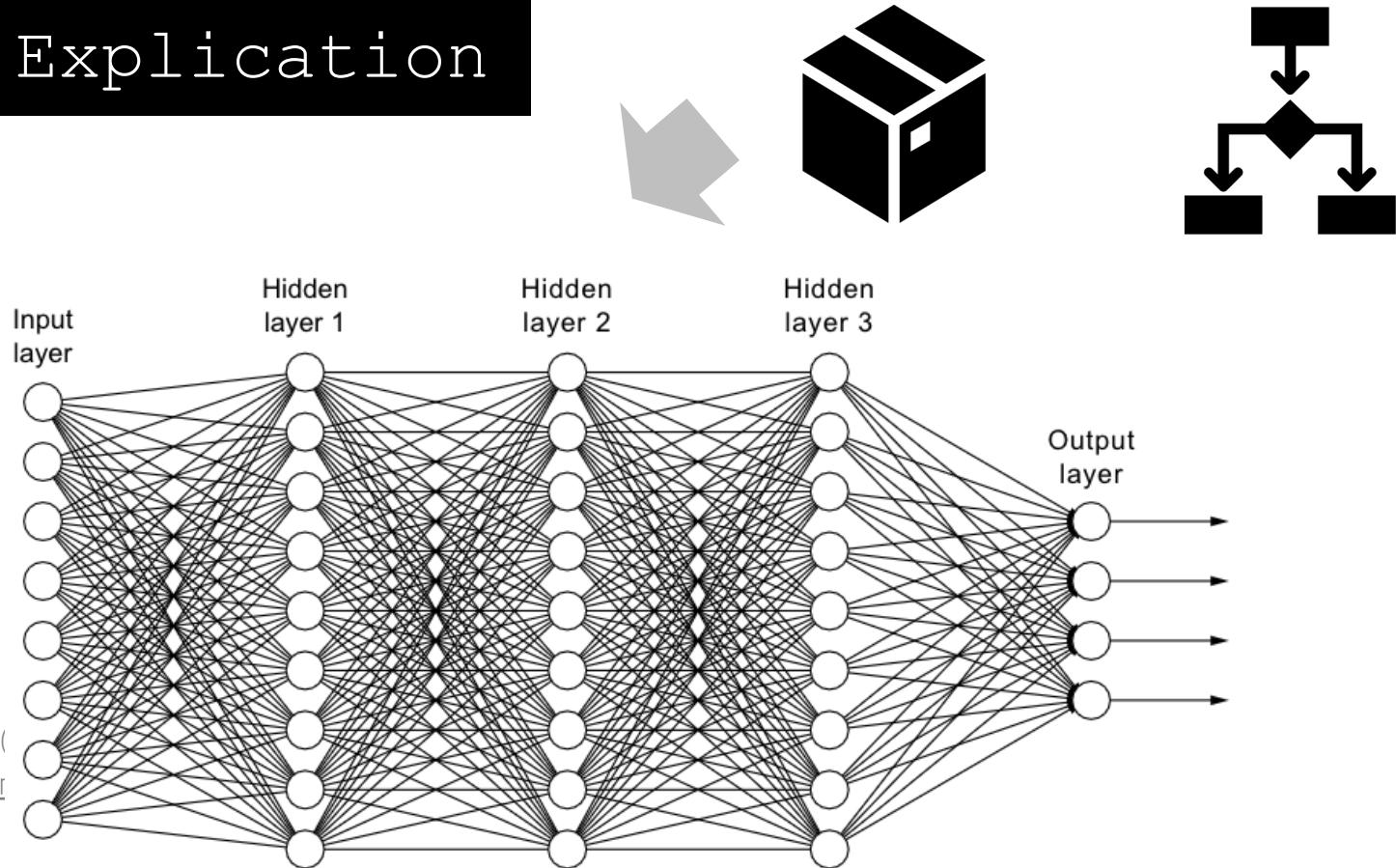
Explication



(<https://datawow.io/blogs/interns-explain-basic-neural-network-ebc555708c9> , mars 2022)

INTRODUCTION

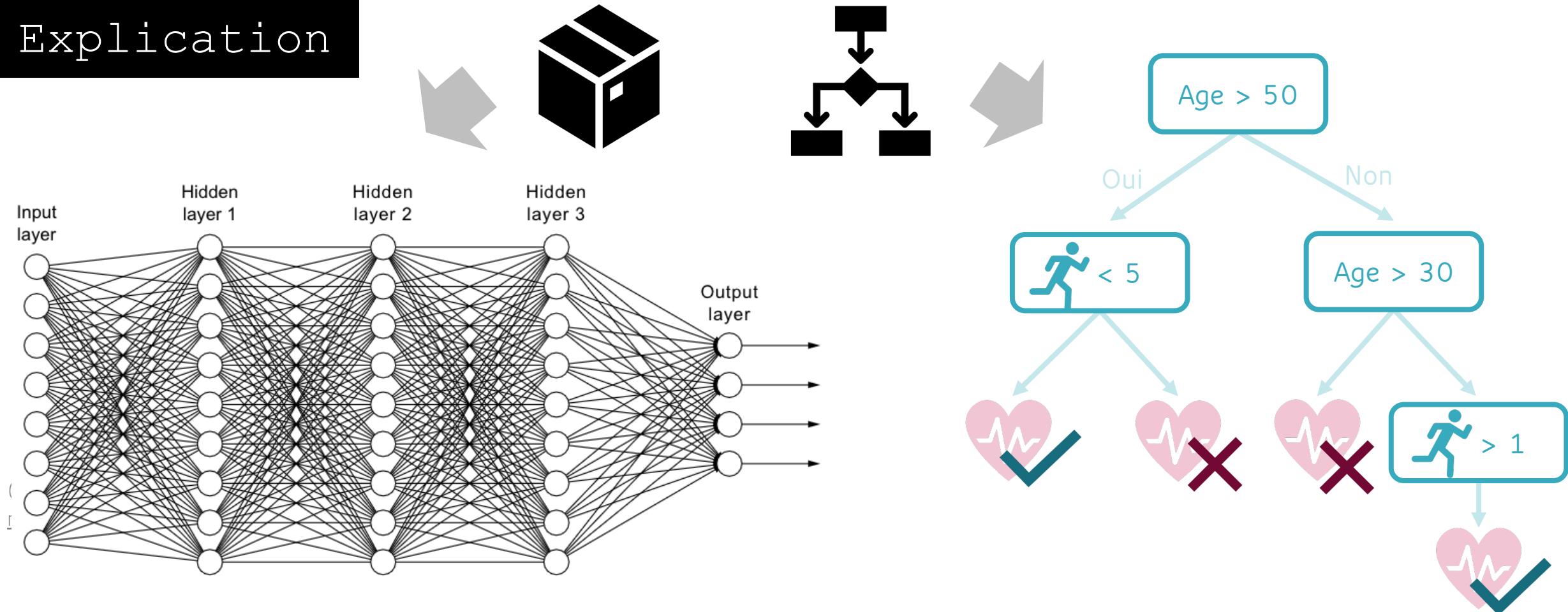
Explication



(<https://manningbooks.medium.com/neural-network-architectures-74527000a798> , mars 2022)

INTRODUCTION

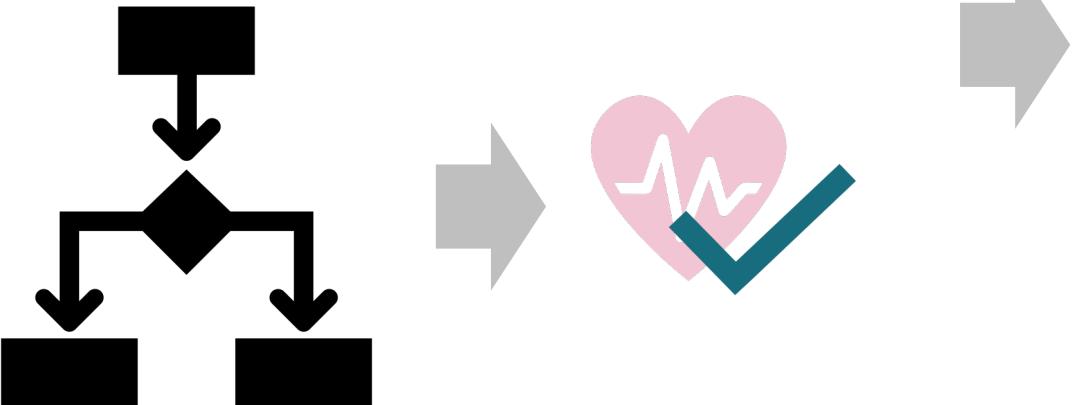
Explication



(<https://manningbooks.medium.com/neural-network-architectures-74527000a798> , mars 2022)

INTRODUCTION

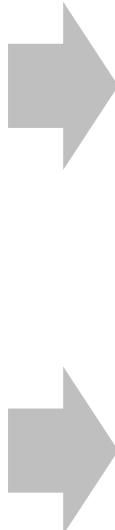
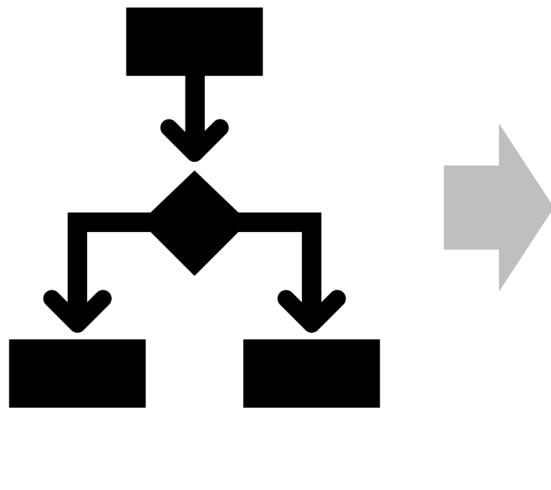
Confiance



Age	Run
72	0,5

INTRODUCTION

Confiance

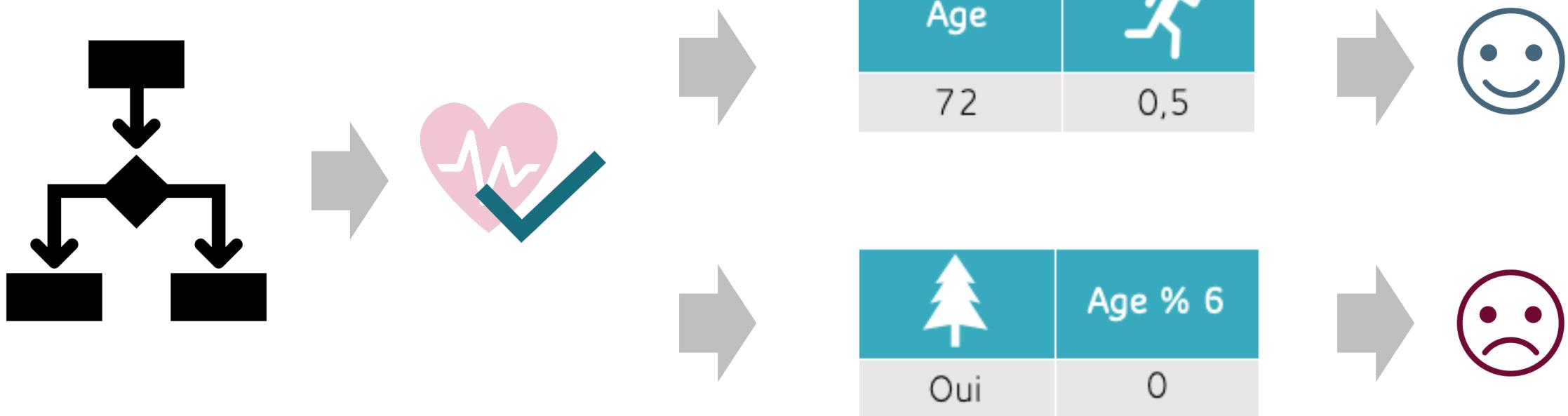


Age	Running icon
72	0,5

Tree icon	Age % 6
Oui	0

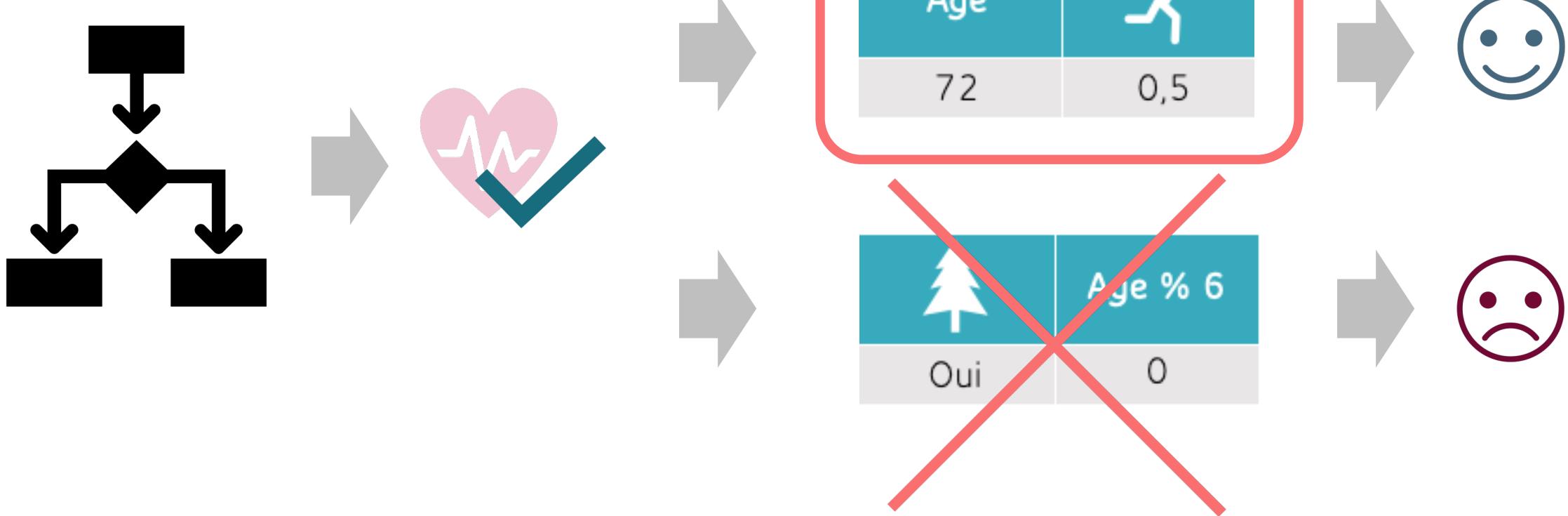
INTRODUCTION

Confiance



INTRODUCTION

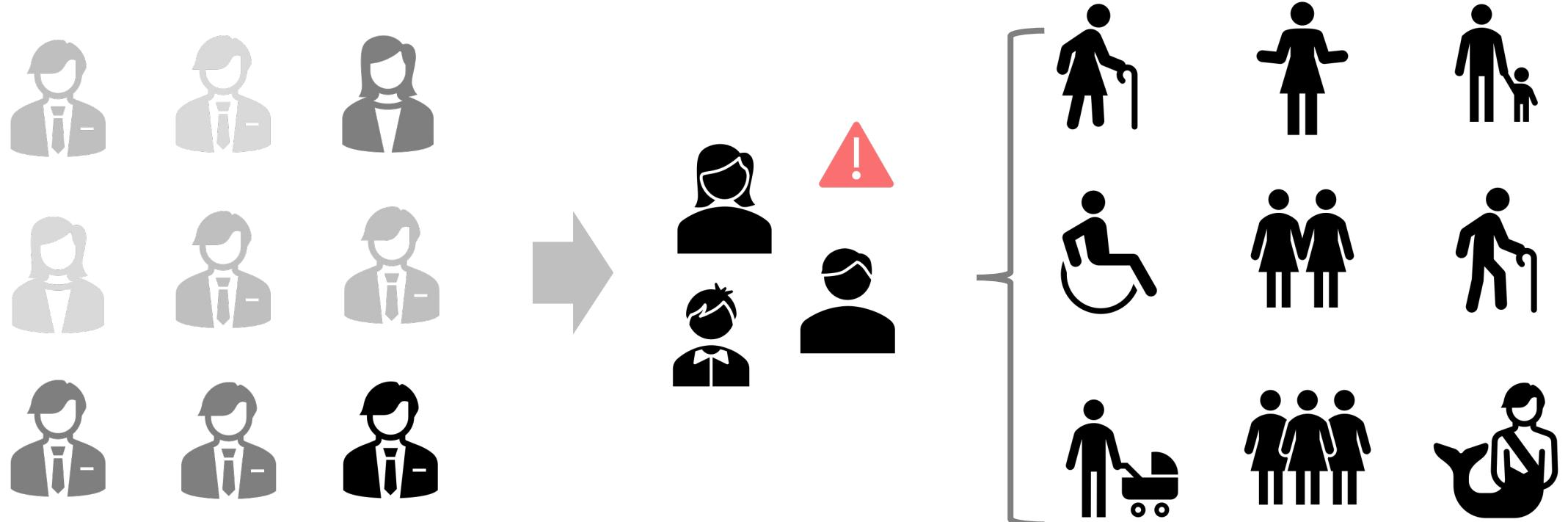
Confiance



INTRODUCTION

PROBLÈME

Les données



(Inspiré de Buolamwini, et al., 2019)

PLAN DE LA PRÉSENTATION

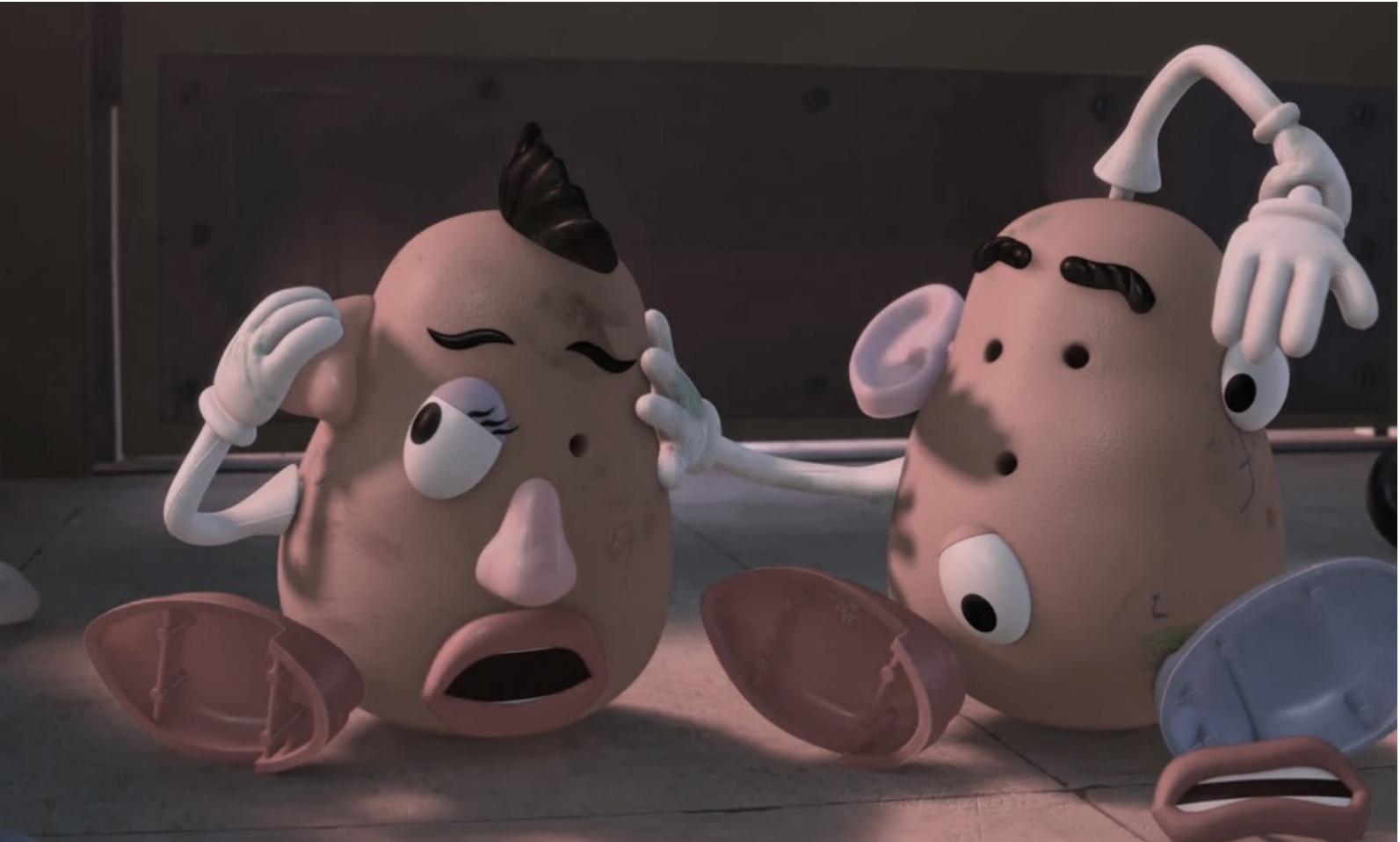
- Introduction
- Définitions
- Pièges de l'explicabilité
- Gagner en interprétabilité
- Perdre de l'interprétabilité
- Conclusion



DÉFINITIONS

Confusion

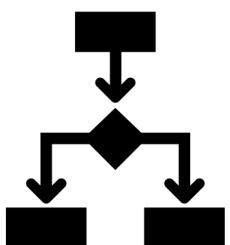
- Normale
- Rigueur scientifique
- Risques éthiques



INTRODUCTION

Interprétabilité

- Modèle
 - Compréhensible pour les humains
 - Transparent
- . . . Intrinsèquement (par design)
- Raisons de la prédiction
- Domaine ancien (années 1950)
 - Arbres de décision, SCM



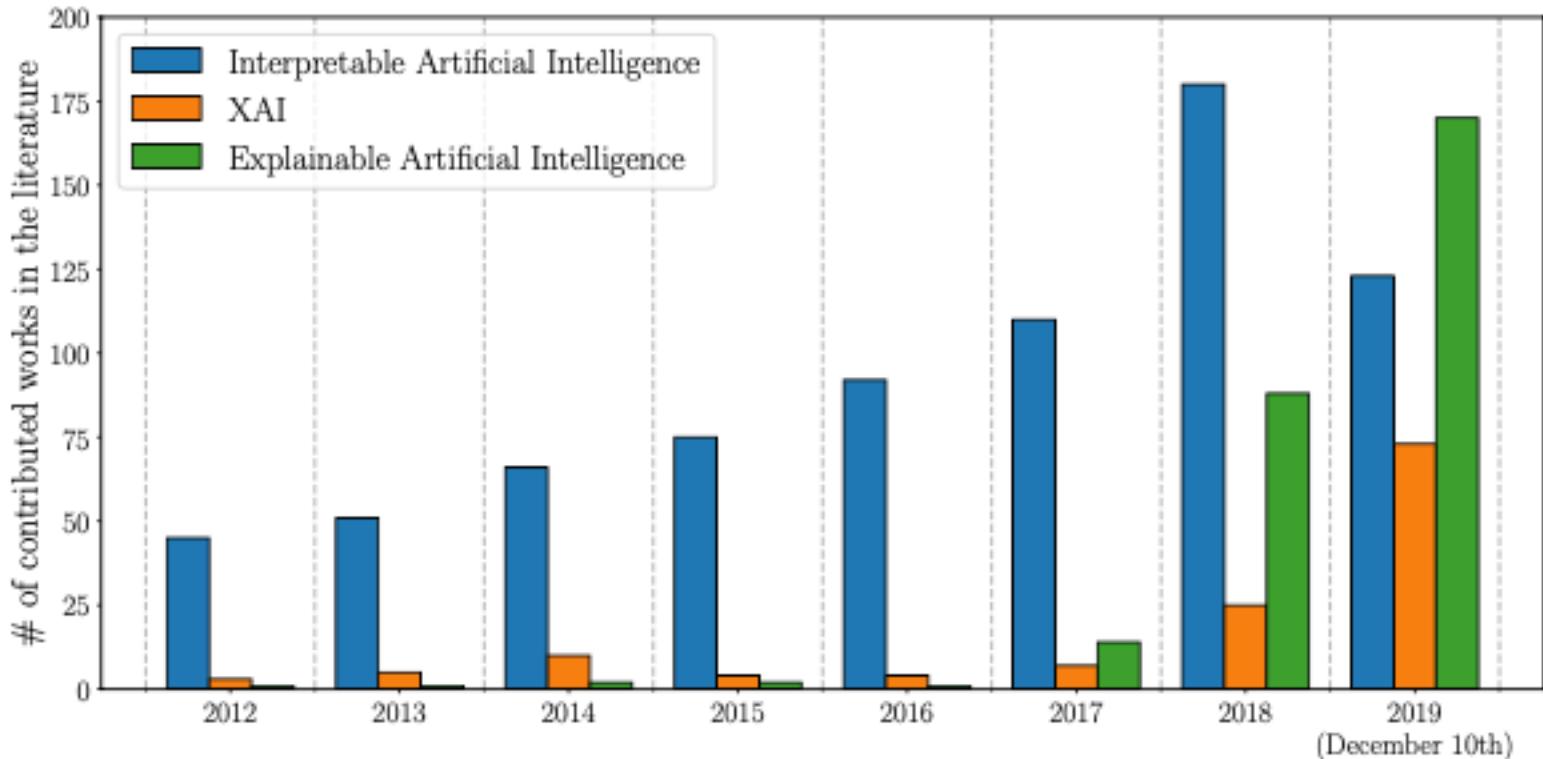
Explicabilité

- Expliquer une boîte noire en utilisant
 - un modèle d'approximation
 - dérivées, mesures d'importance des variables (ou autres statistiques)
 - explication *post hoc*
- Mécanisme de la prédiction
- Domaine récent
 - Réseaux de neurones
- Méthodes courantes
 - LIME
 - Valeurs de Shapley
 - Cartes de protubérance (*saliency maps*)



DÉFINITIONS

Distinction



(Arrieta, 2021)

DÉFINITIONS

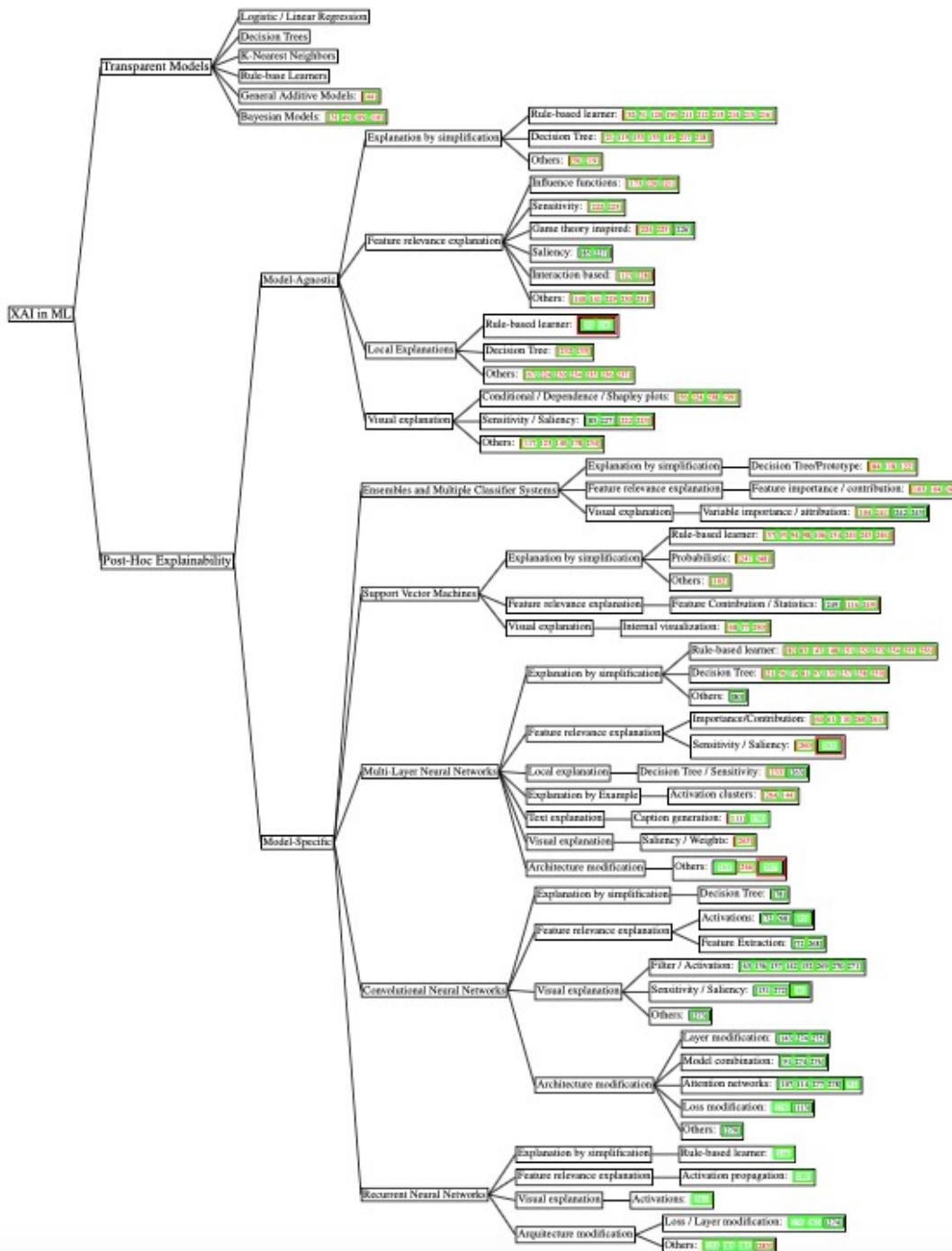
Distinction



(Rudin, 2021)

DÉFINITIONS

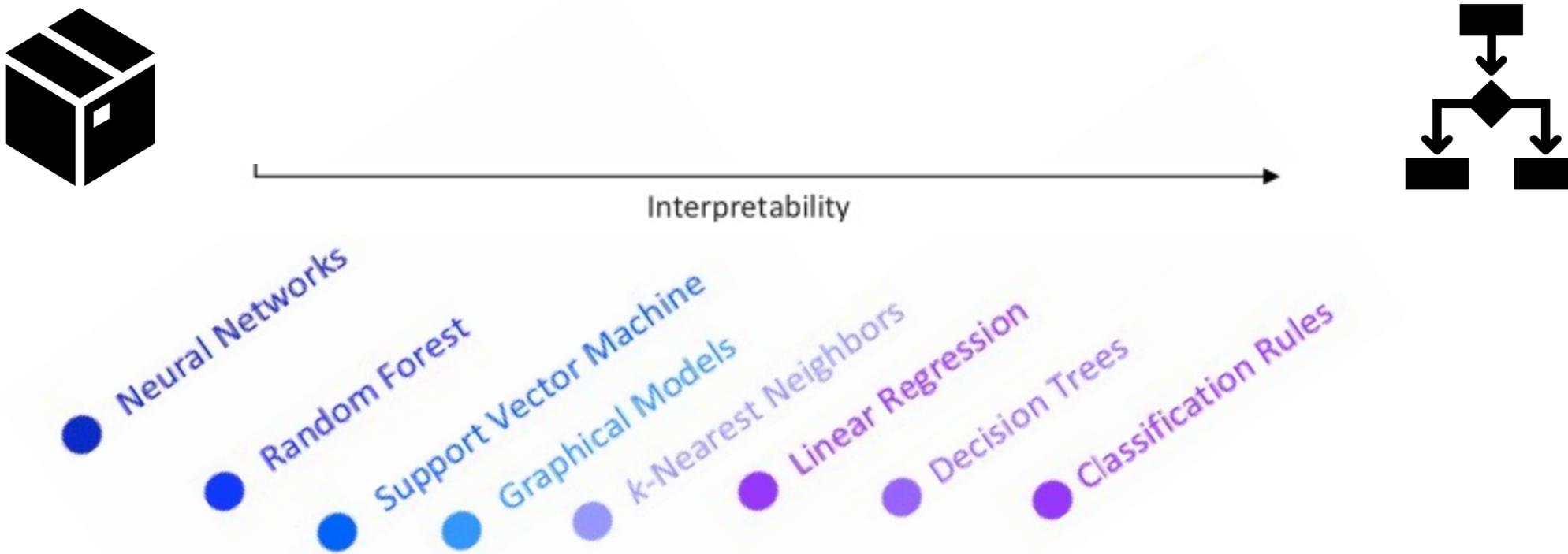
Classification



(Arrieta, 2019)

INTRODUCTION

Continuum



(Inspiré de Morocho-Cayamcela et al., 2019)

PLAN DE LA PRÉSENTATION

- Introduction
- Définitions
- Pièges de l'explicabilité
- Gagner en interprétabilité
- Perdre de l'interprétabilité
- Conclusion

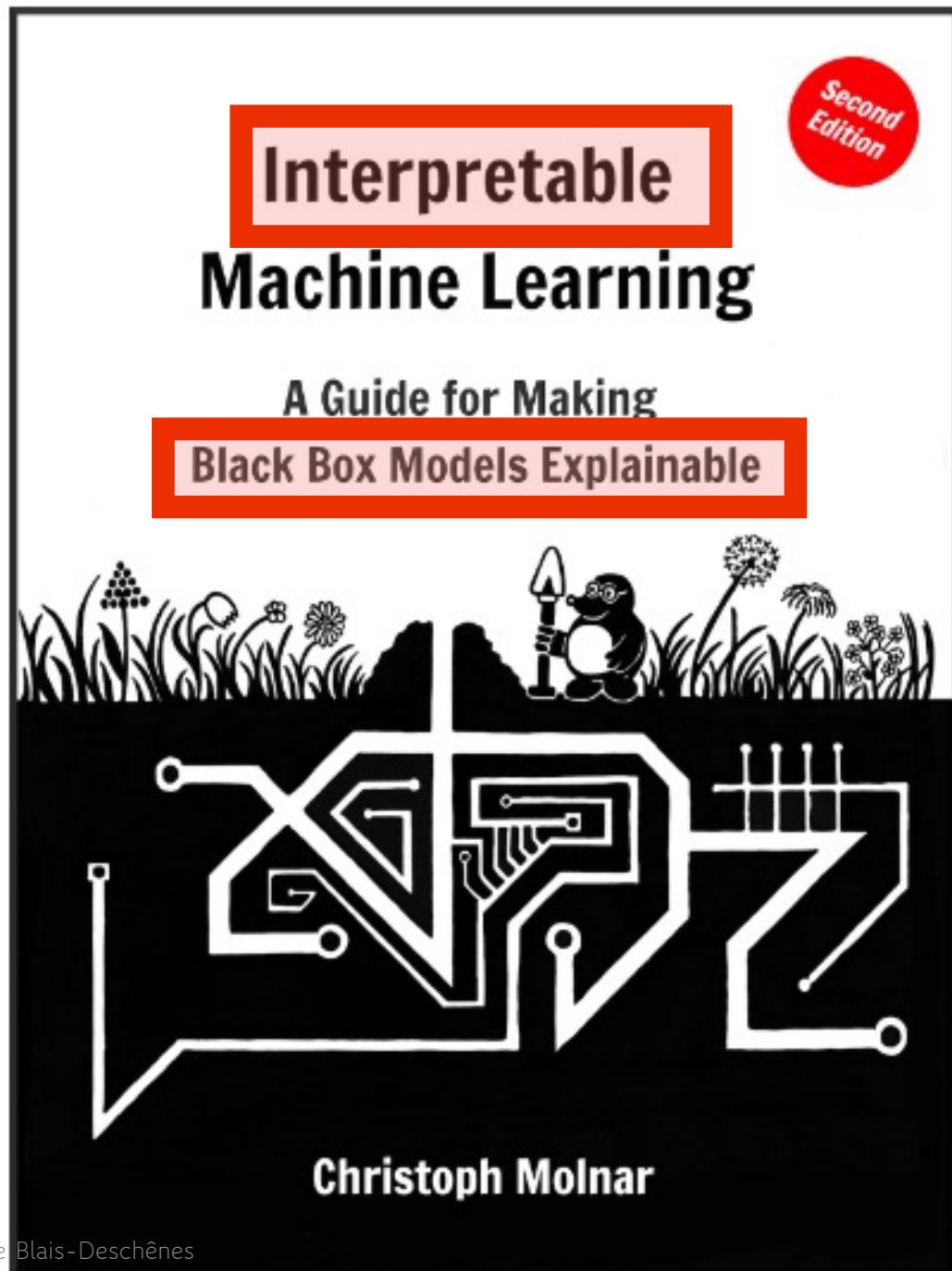


EXPLICABILITÉ

Classification

- Modèles intrinsèquement interprétables
- Explications *post hoc*
 - Modèle agnostique
 - Global
 - Approximation/substituts (*surrogate*)
 - Importance des attributs (*feature importance*)
 - Local
 - Valeurs de Shapley / SHAP
 - LIME
 - Contrefactuels
 - Spécifique au modèle
 - Réseaux de neurone (images)
 - Cartes de protubérances (*saliency maps*)

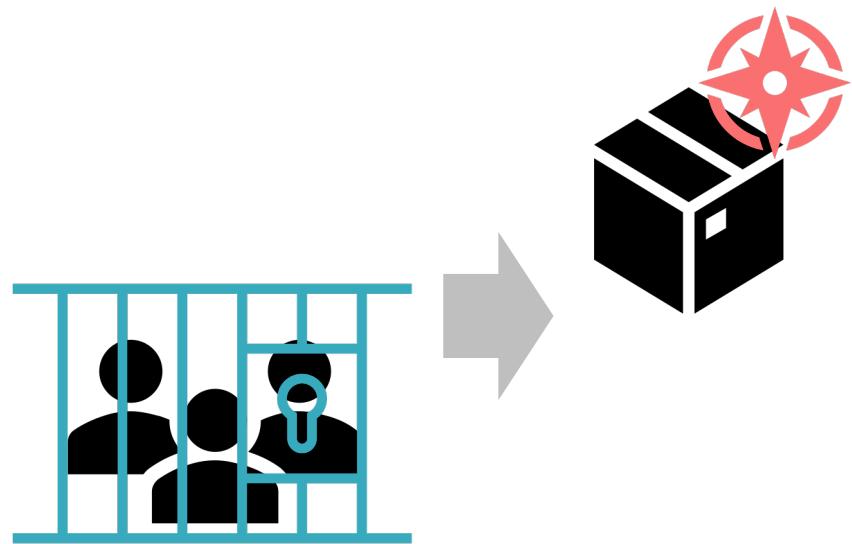
(Molnar, 2022)



EXPLICABILITÉ

Modèle agnostique global

Approximation



Difficulté d'approximation

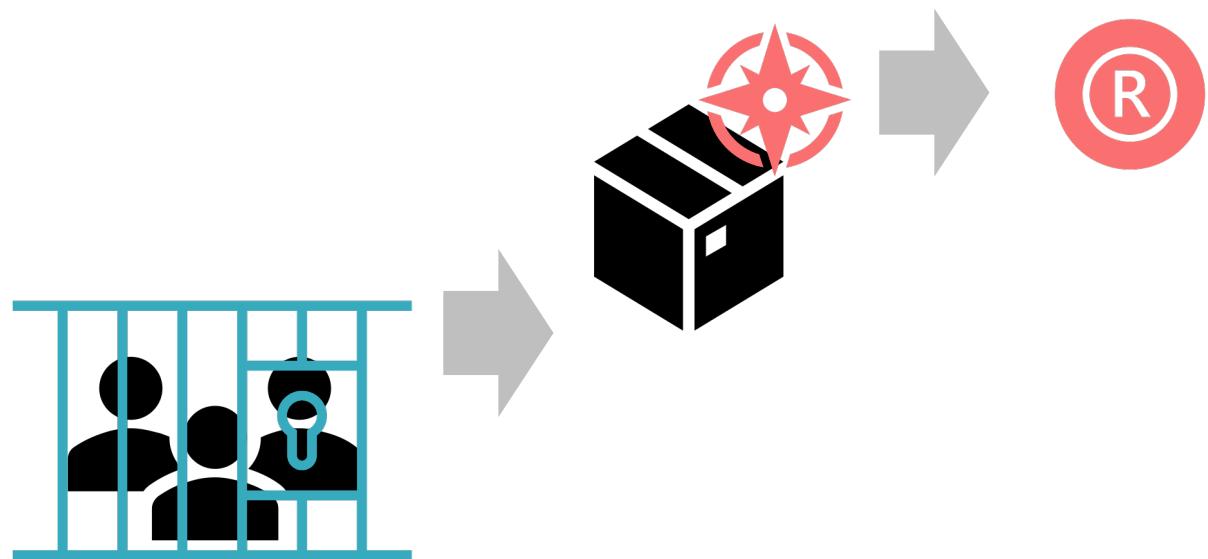
■ COMPAS

- Boîte noire (brevetée)
- Risque de récidive criminelle

EXPLICABILITÉ

Modèle agnostique global

Approximation



Difficulté d'approximation

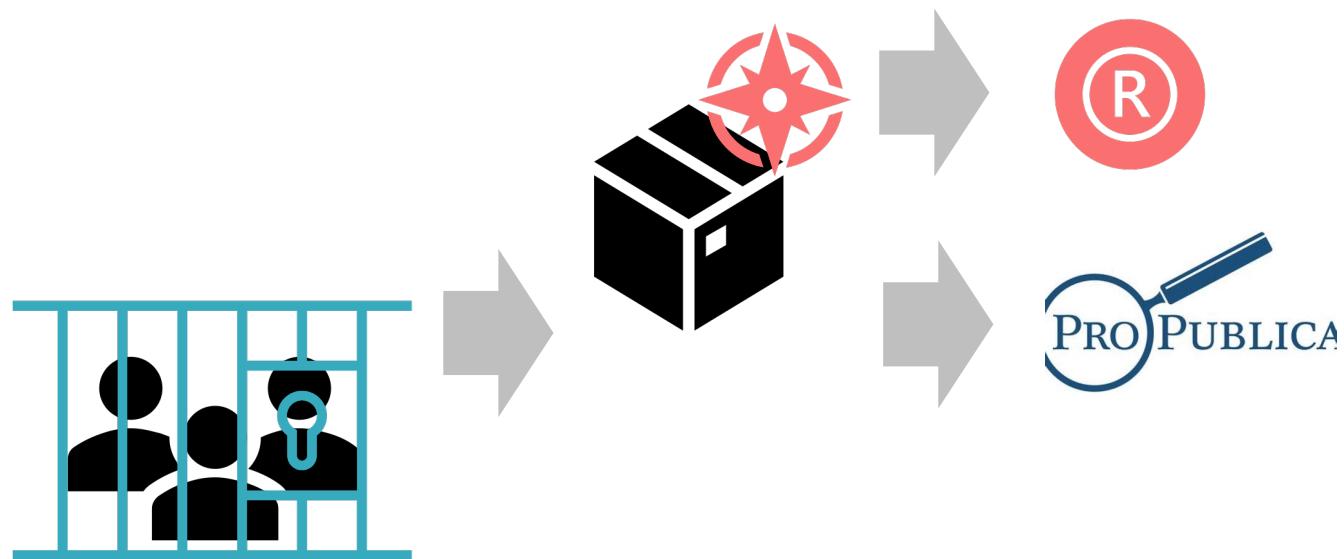
■ COMPAS

- Boîte noire (brevetée)
- Risque de récidive criminelle

EXPLICABILITÉ

Modèle agnostique global

Approximation



Difficulté d'approximation

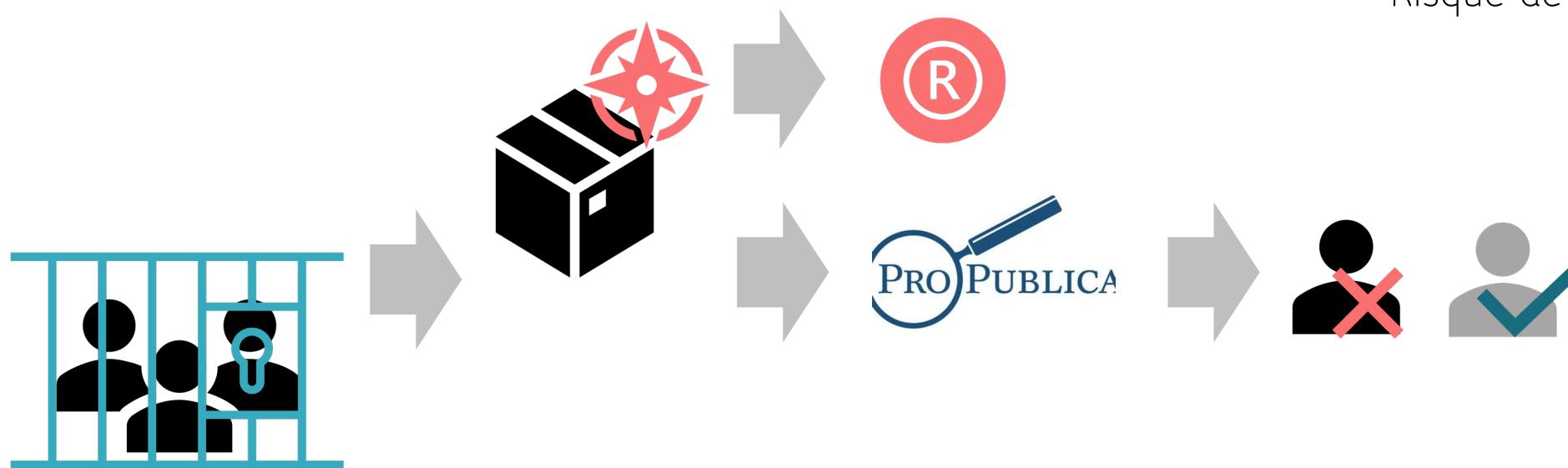
■ COMPAS

- Boîte noire (brevetée)
- Risque de récidive criminelle

EXPLICABILITÉ

Modèle agnostique global

Approximation



Difficulté d'approximation

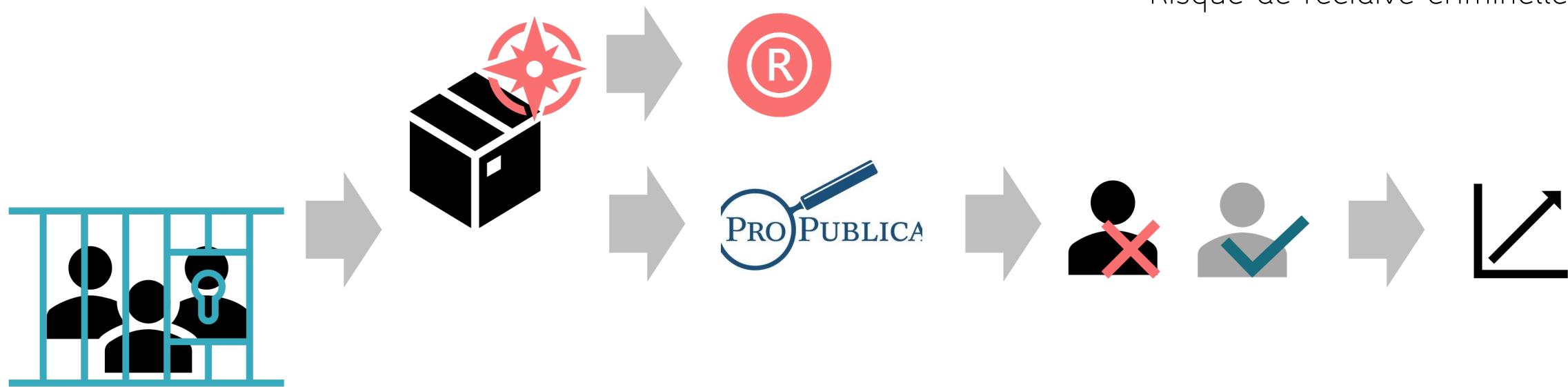
■ COMPAS

- Boîte noire (brevetée)
- Risque de récidive criminelle

EXPLICABILITÉ

Modèle agnostique global

Approximation



Difficulté d'approximation

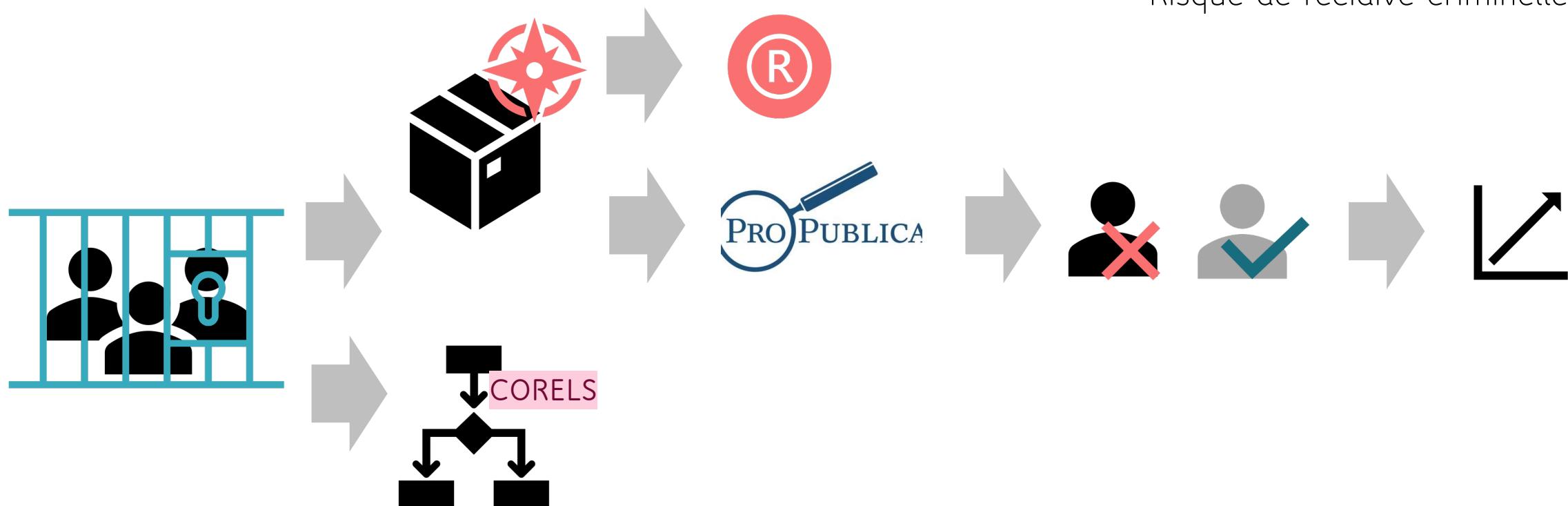
■ COMPAS

- Boîte noire (brevetée)
- Risque de récidive criminelle

EXPLICABILITÉ

Modèle agnostique global

Approximation



Difficulté d'approximation

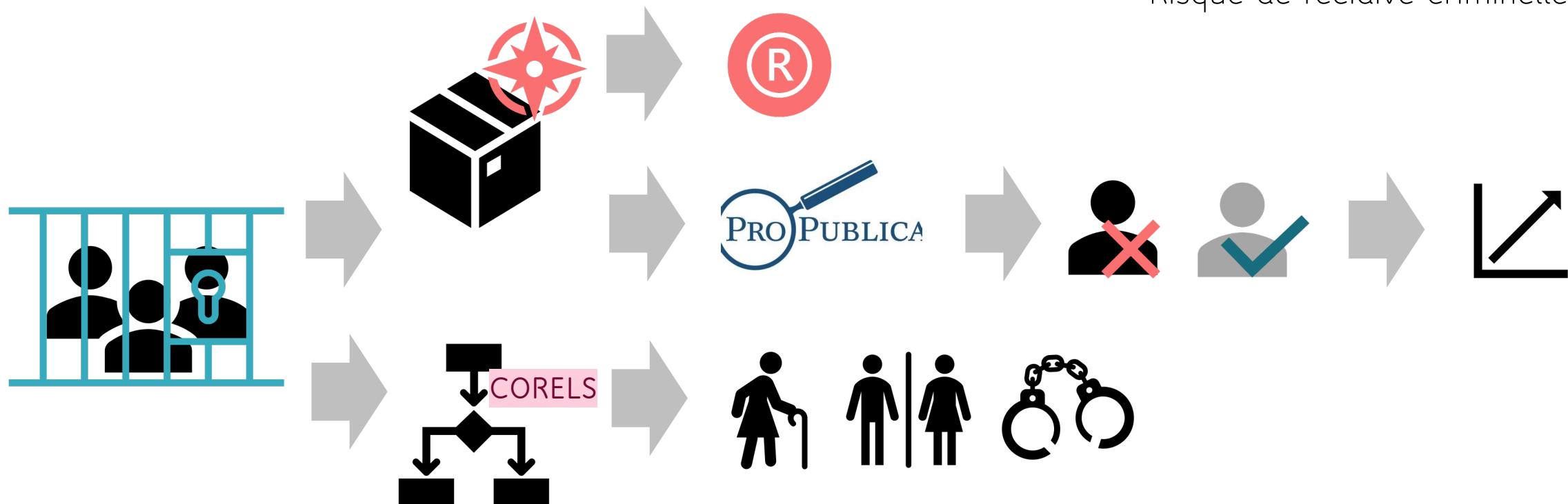
■ COMPAS

- Boîte noire (brevetée)
- Risque de récidive criminelle

EXPLICABILITÉ

Modèle agnostique global

Approximation



(Angelino, et al., 2021)

Difficulté d'approximation

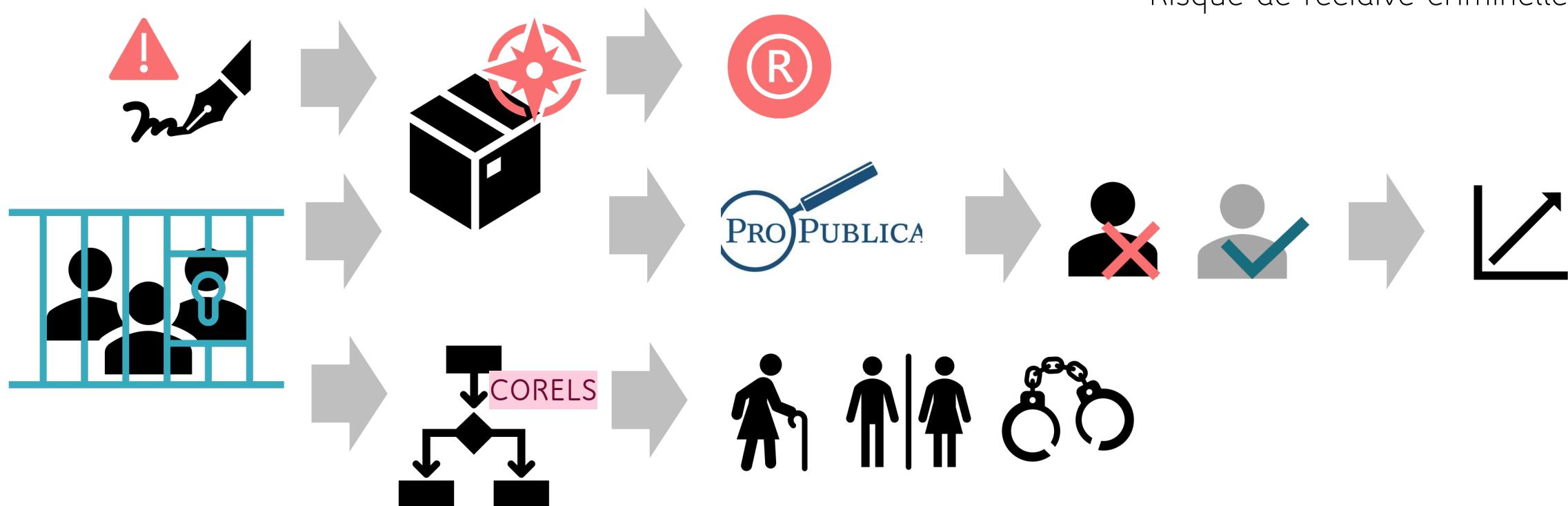
■ COMPAS

- Boîte noire (brevetée)
- Risque de récidive criminelle

EXPLICABILITÉ

Modèle agnostique global

Approximation



(Angelino, et al., 2021)

Difficulté d'approximation

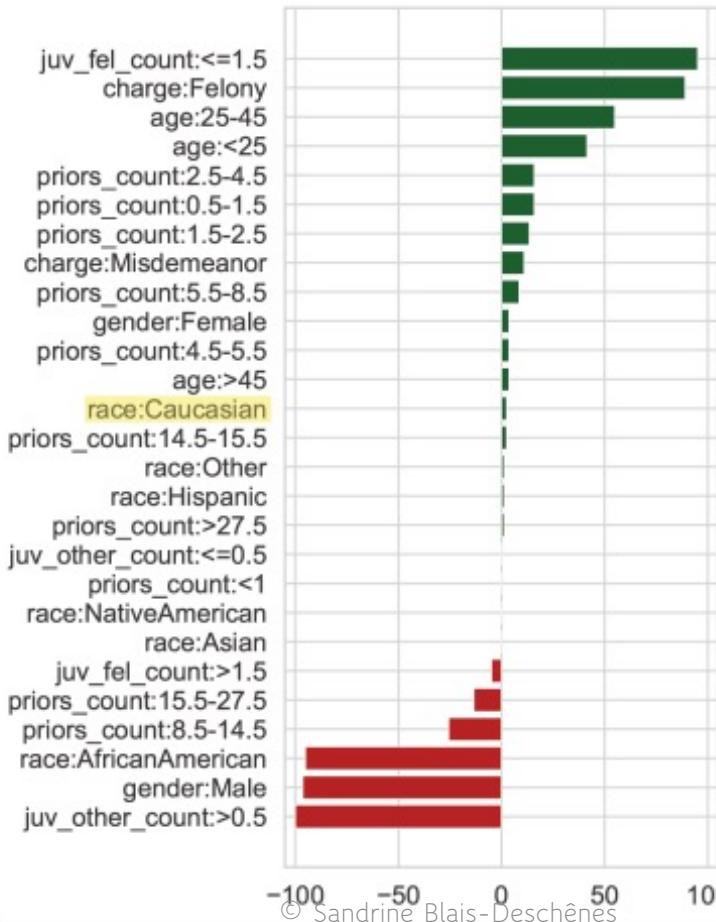
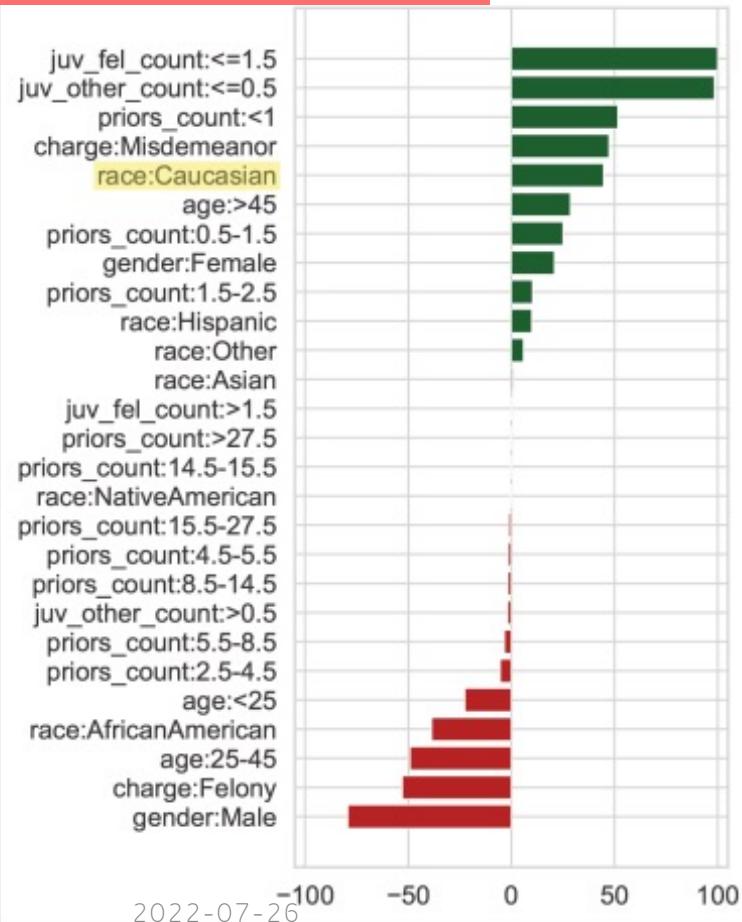
■ COMPAS

- Boîte noire (brevetée)
- Risque de récidive criminelle

EXPLICABILITÉ

Modèle agnostique global

Approximation



Blanchiment éthique

■ COMPAS

- ... Explication (gauche)
 - FairML (Adebayo et al, 2012)
- Approximation (droite)
 - LaundryML (Aïvodji, et al, 2012)

```
if prior_count: 15.5–27.5 then
    recidivate:True
else if prior_count: 8.5–14.5 then
    recidivate:True
else if age:>45 then
    recidivate:False
else if juv_other_count:>0.5 then
    recidivate:True
else
    recidivate:False
end if
```

(Aïvodji, et al., 2021)

EXPLICABILITÉ

Modèle agnostique global

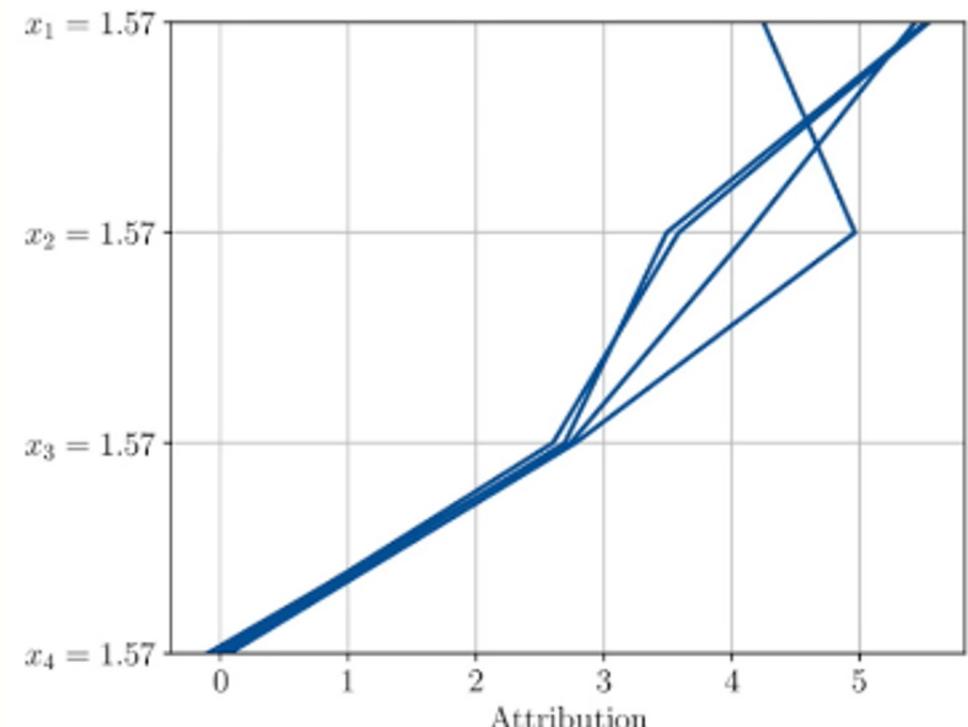
Importance des attributs

Principe:

- Pour chaque attribut
 - Permuter les valeurs d'un attribut
 - Importance plus l'erreur augmente

Piège

- Dépend de l'aléatoire
 - Importance varie selon les itérations
 - Ensemble Rashomon (*Rashomon set*)



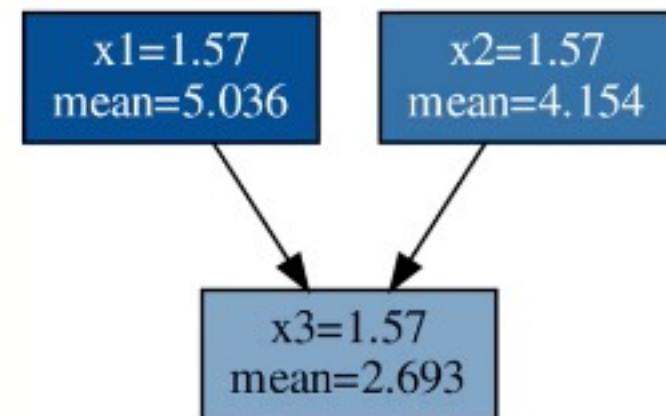
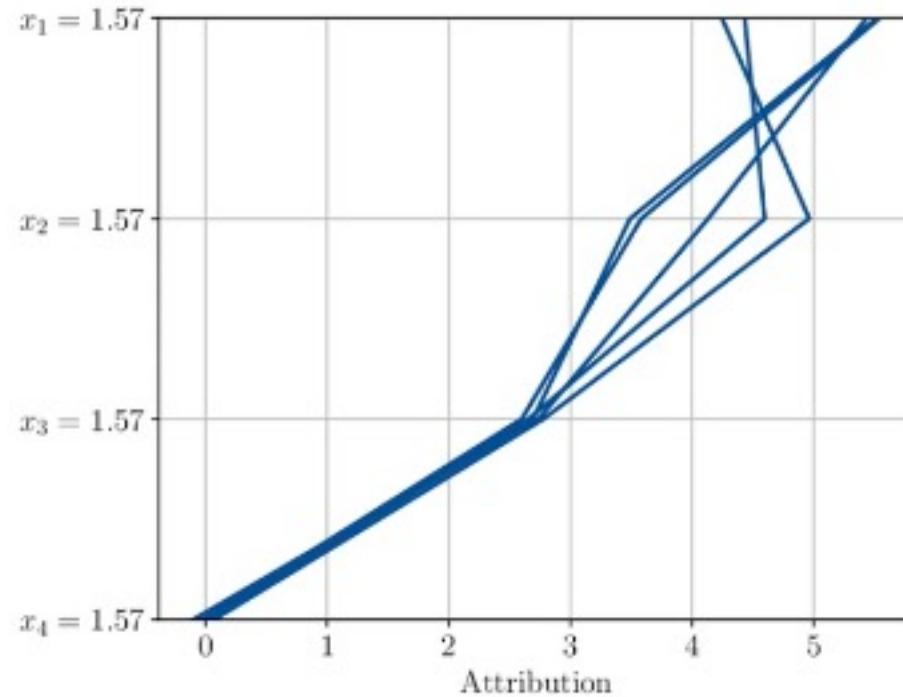
EXPLICABILITÉ

Modèle agnostique global

Importance des attributs

Piège

- Difficile d'interpréter
 - Dépendances attributs
 - Difficulté du consensus
- Hypothèses et simplification
 - Rouler plus vite
 - Ex indépendance des variables
- Erreur ou mauvaise mesure
 - Aller lire le papier



EXPLICABILITÉ

Modèle agnostique local

Valeurs de Shapley / SHAP

SHapely Additive exPlanations

- 6 méthodes de contribution des attributs

Principe

- Contribution de chaque attribut
 - Selon toutes les combinaisons
 - Échantillonnage aléatoire

Pièges

- Computationnellement très lourd
- Explication plus complexe
- Score peu intuitif
- Ignore corrélation entre les attributs

$$v(c) = \begin{cases} 80, & \text{if } c = \{A\} \\ 56, & \text{if } c = \{B\} \\ 70, & \text{if } c = \{C\} \\ 80, & \text{if } c = \{A, B\} \\ 85, & \text{if } c = \{A, C\} \\ 72, & \text{if } c = \{B, C\} \\ 90, & \text{if } c = \{A, B, C\} \end{cases}$$

π	δ_π^G
(A, B, C)	(80, 0, 10)
(A, C, B)	(80, 5, 5)
(B, A, C)	(24, 56, 10)
(B, C, A)	(18, 56, 16)
(C, A, B)	(15, 5, 70)
(C, B, A)	(18, 2, 70)

- Contribution marginale de A
 - $(80+80+56+16+5+70)/6 = 51.17$
- Total de toutes les contributions marginales = 90

<https://www.kdnuggets.com/2019/12/interpretability-part-3-lime-shap.html>

EXPLICABILITÉ

Modèle agnostique local

LIME

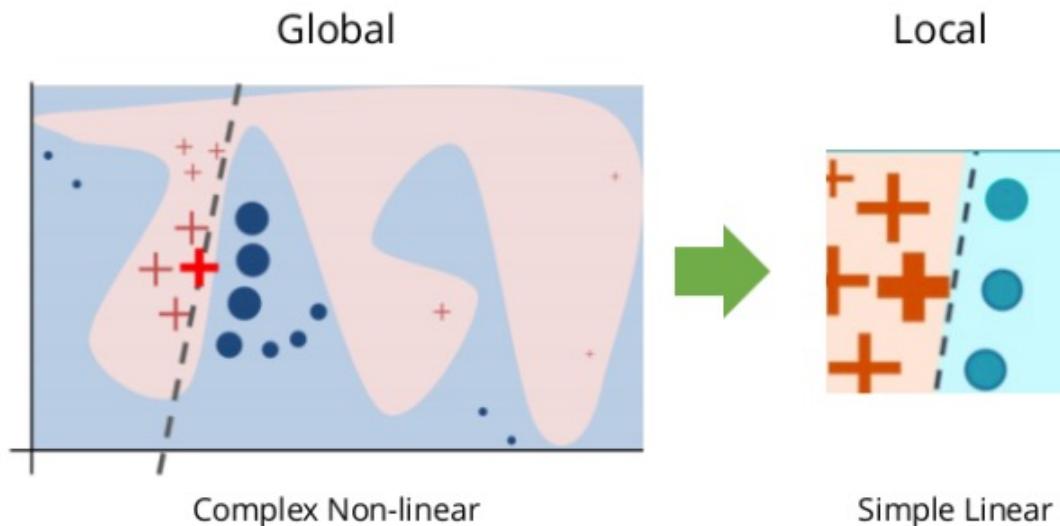
Local Interpretable Model-agnostic Explanations (LIME)

Principe

- Perturbation attributs
- Un point à l'étude
- Création d'un nouveau jeu de données

Pièges

- Définition du voisinage
- Importance des permutations
- Approximation par un modèle linéaire
- Explication instable
- Possible cacher les biais



<https://www.kdnuggets.com/2019/12/interpretability-part-3-lime-shap.html>

EXPLICABILITÉ

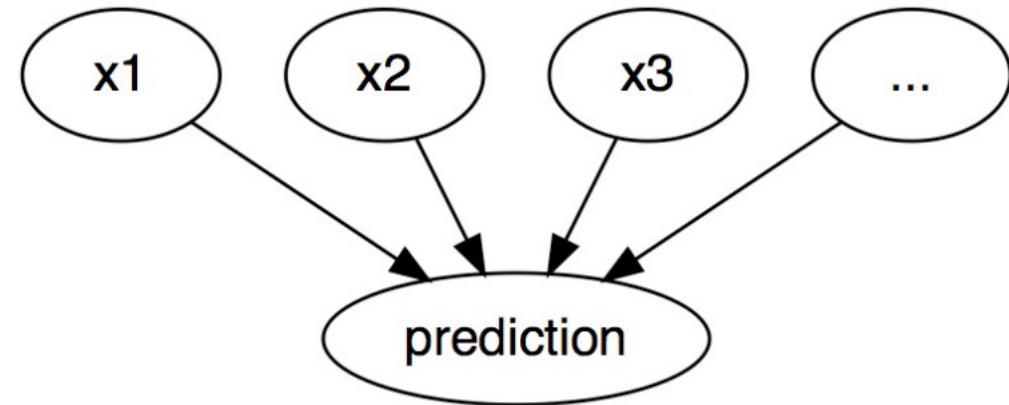
Modèle agnostique local

Contrefactuels

Et si ... ?

Principe

- Changer des éléments
- Observer le changement dans la prédiction
- Proche du raisonnement humain



Pièges

- Plusieurs explications (effet Rashomon)
- Explication locale et dépendante des questions de recherche

EXPLICABILITÉ

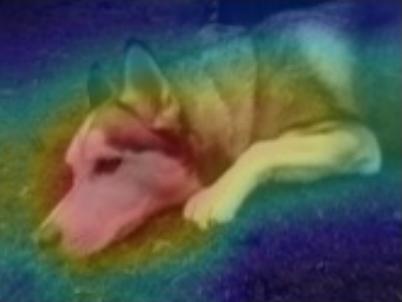
Spécifique au modèle

Carte de protubérances (*Saliency map*)

- Importance des zones d'une image

- Utile pour les zones ignorées

Information incomplète

	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps			

(Rudin, et al., 2019)

EXPLICABILITÉ

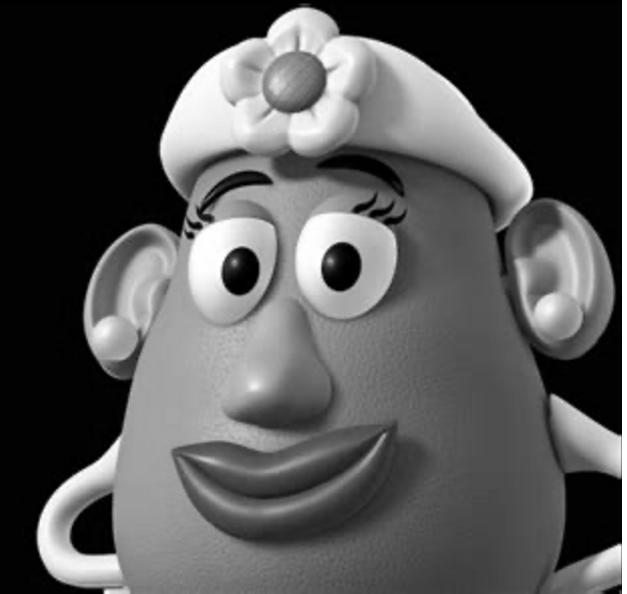
Boîtes noires

```
25 import xgboost  
26  
27 # Fit black box model  
28 model = xgboost.XGBClassifier()  
29 model.fit(X, y)  
30  
31 import shap  
32  
33 # Explain predictions on first 100 instances  
34 explainer = shap.explainers.Exact(model.predict_proba, X)  
35 shap_values = explainer(X[:100])
```

(Laberge, Mobilit.ai 2022)

PLAN DE LA PRÉSENTATION

- Introduction
- Définitions
- Pièges de l'explicabilité
- Gagner en interprétabilité
- Perdre de l'interprétabilité
- Conclusion

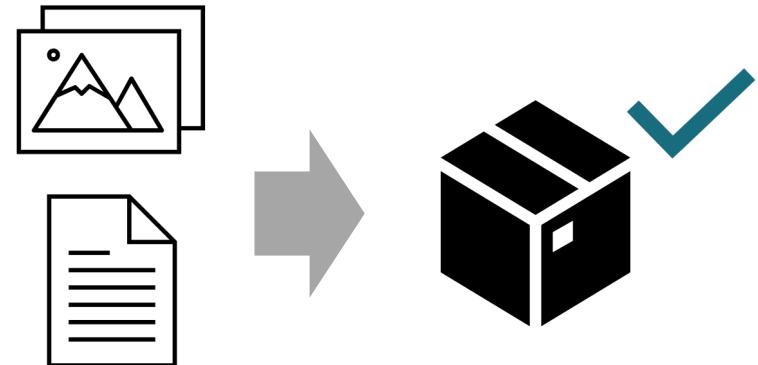


INTERPRÉTABILITÉ

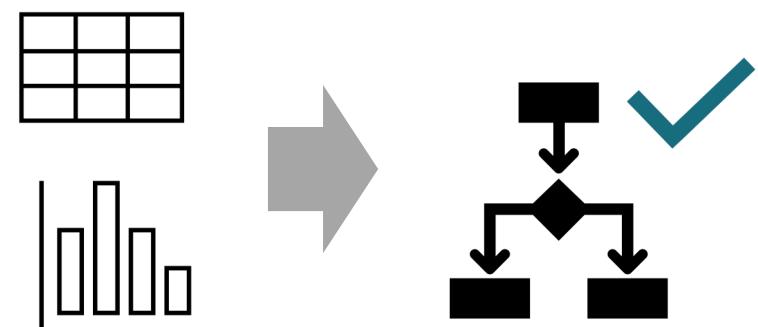
Compromis

- Compromis interprétabilité et exactitude 
 - Fausse dichotomie
 - Aucune preuve scientifique
- Interprétabilité mènerait même à une meilleure exactitude
 - Utile pour déboguer/améliorer (*troubleshooting*)
- Compromis parcimonie et exactitude 
 - Parcimonie $\neg=$ interprétabilité
 - Association forte
 - Une composante parmi d'autres

Données brutes (images, texte)



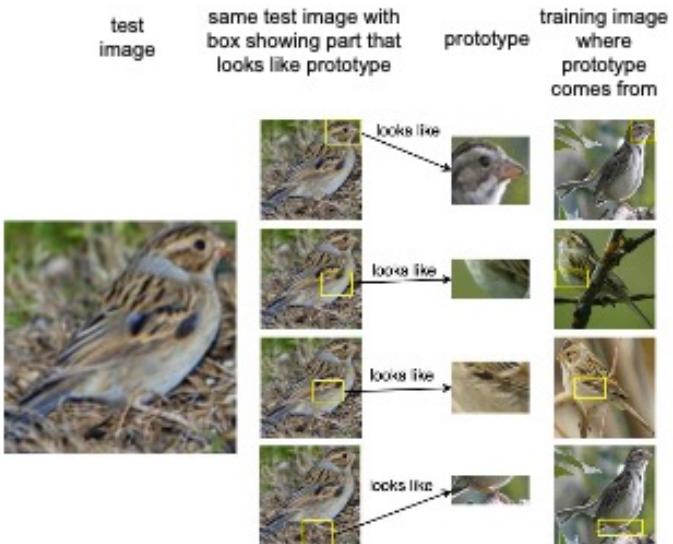
Données tabulaires



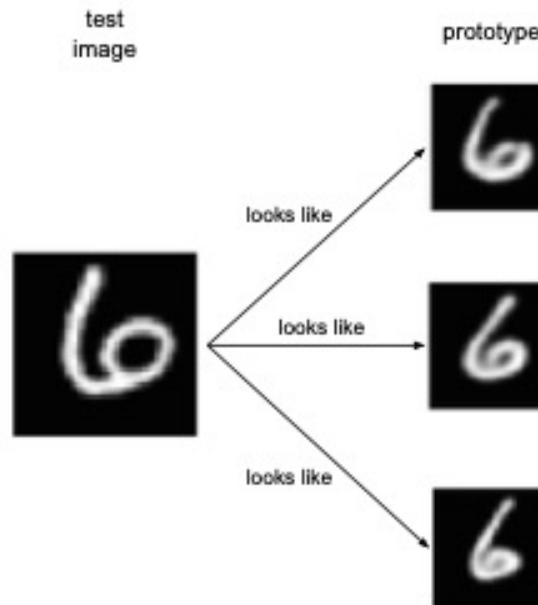
INTERPRÉTABILITÉ

Réseaux de neurones

- Proche du raisonnement humain
 - Cas de base
 - Prototypes



Ex. Cas de bas dans un réseau de neurones

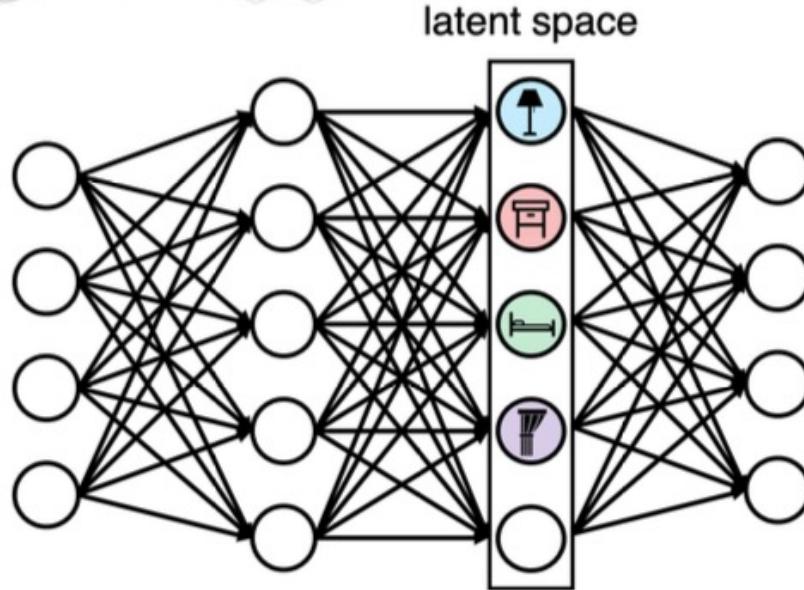


Ex. Prototype dans un réseau de neurones

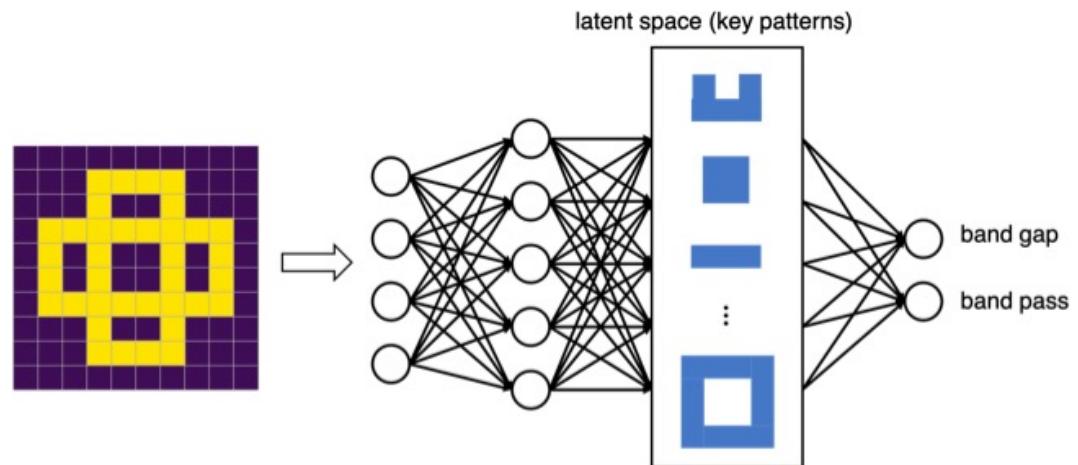
INTERPRÉTABILITÉ

Réseaux de neurones

- Désenchevêtrement (*disentanglement*)
 - Manière dont l'information voyage dans un réseau de neurones
 - Séparer l'information selon les concepts
 - Chaque neurone représente un concept humainement interprétable
- Supervisé
 - Les spécialistes spécifient les concepts
- Non supervisé
 - L'algorithme choisit les concepts d'intérêt
 - Biais dans les images étiquetées
 - Entités étiquetées sont spécifiques à certaines tâches
 - Ignore information pertinente



Ex. Désenchevêtrement supervisé de l'espace latent d'un réseau de neurones

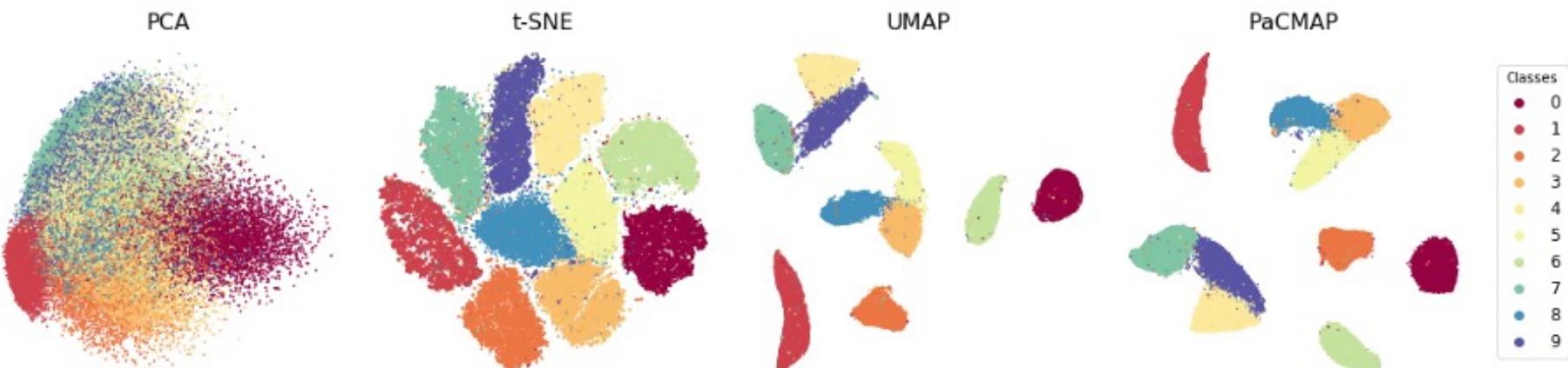


Ex. Désenchevêtrement non supervisé de l'espace latent d'un réseau de neurones

INTERPRÉTABILITÉ

Réduction de dimensions

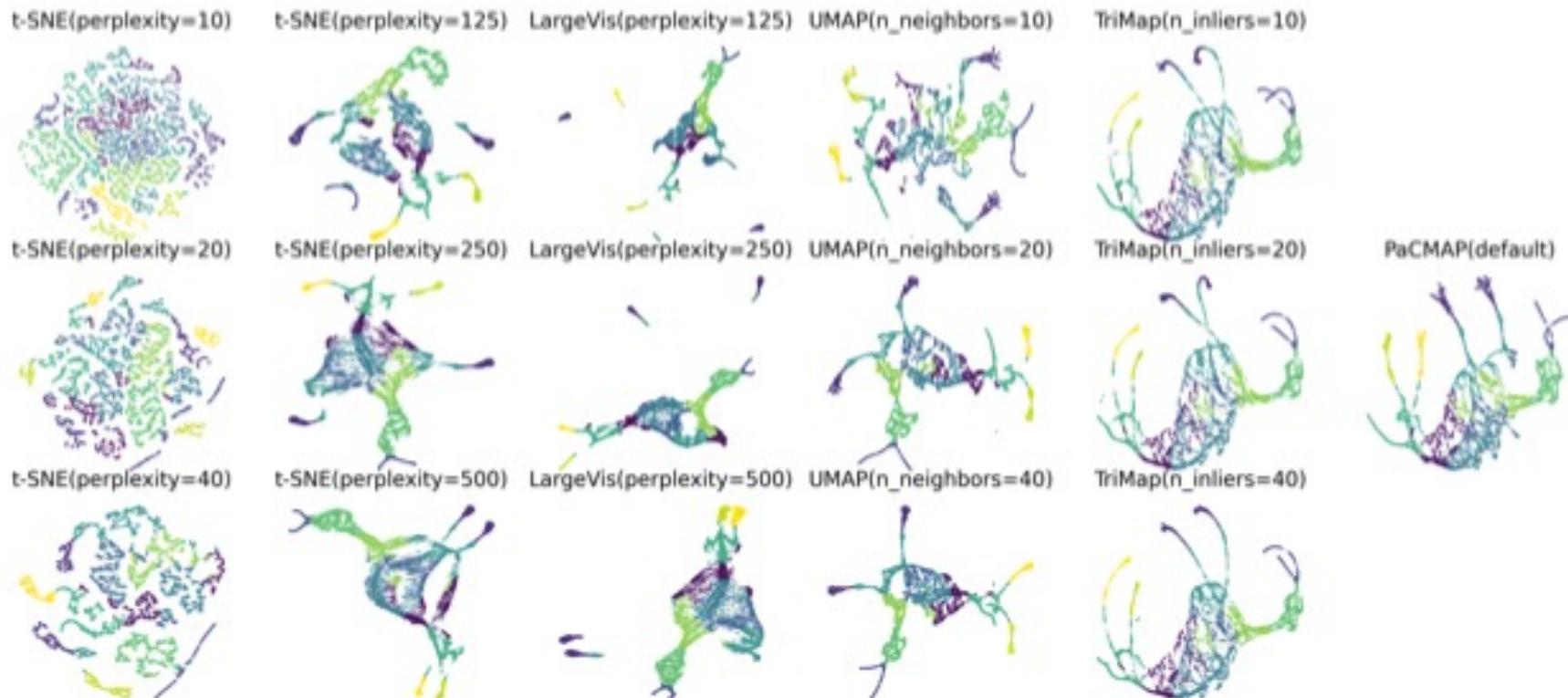
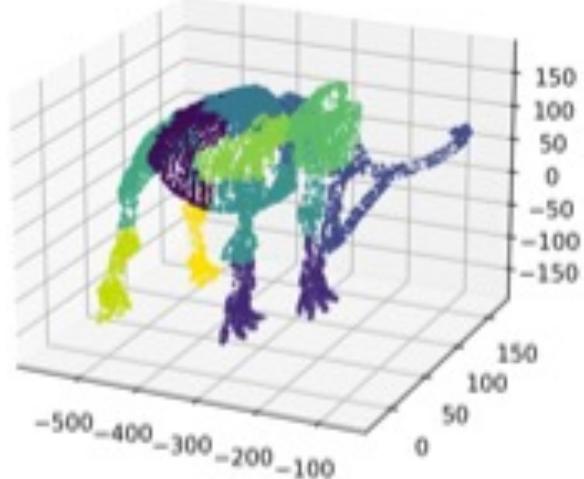
- *Mapping* espace de plus petite dimension
- Préserver une notion de distance en 2D (visualisation)
- Fonctions de perte différent
- Méthodes
 - Globale
 - Distance entre toutes les paires de points
 - Locale
 - Préserve le voisinage
 - Fonction de perte combine les deux (Böhm et al., 2020)



INTERPRÉTABILITÉ

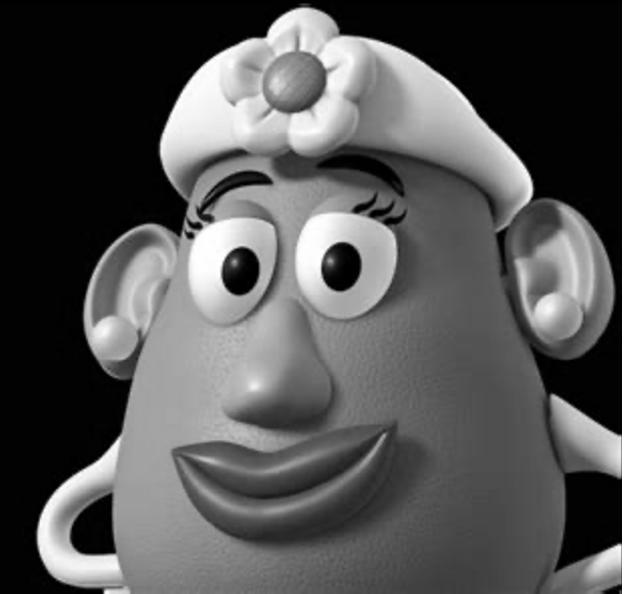
Réduction de dimensions

Original Mammoth



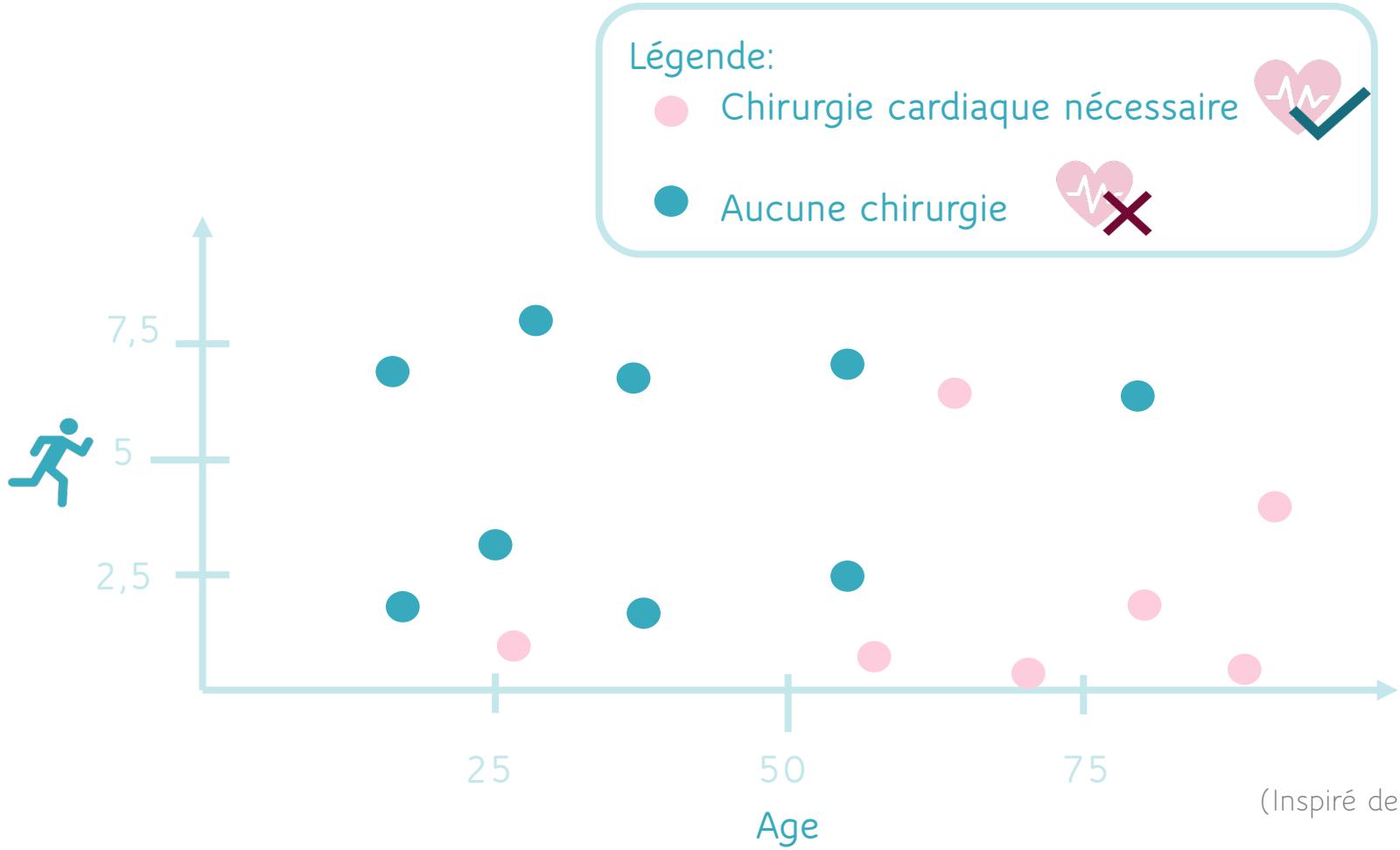
PLAN DE LA PRÉSENTATION

- Introduction
- Définitions
- Pièges de l'explicabilité
- Gagner en interprétabilité
- Perdre de l'interprétabilité
- Conclusion



INTERPRÉTABILITÉ ↓

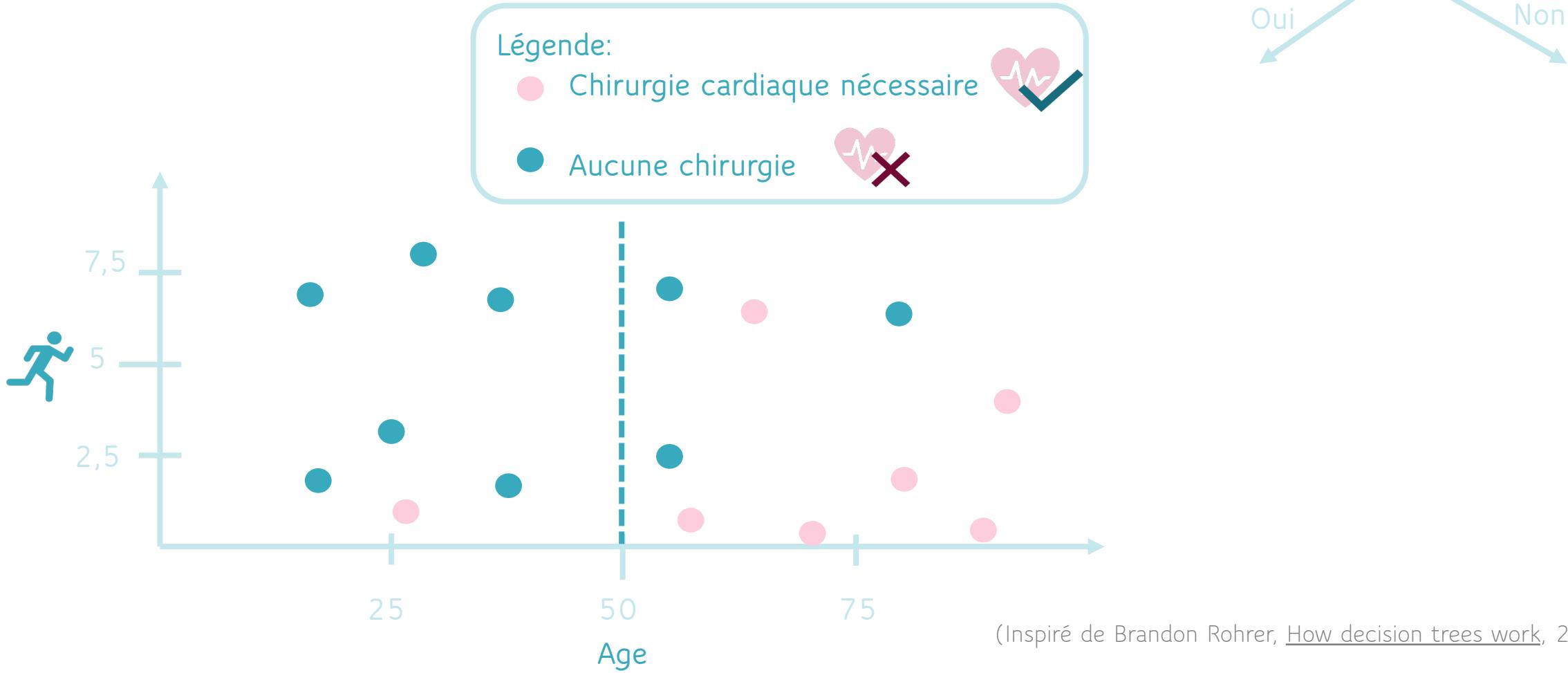
Modèle logique



(Inspiré de Brandon Rohrer, [How decision trees work](#), 2021)

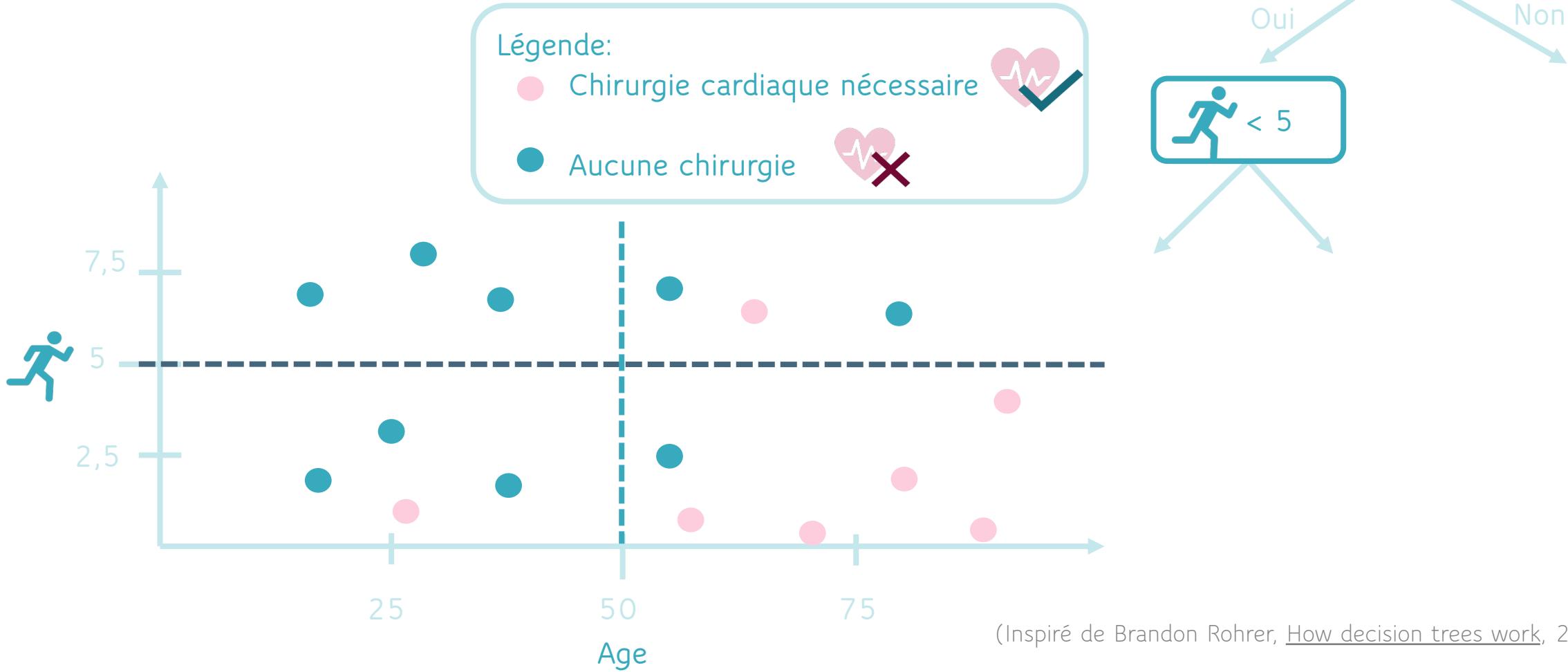
INTERPRÉTABILITÉ ↓

Modèle logique



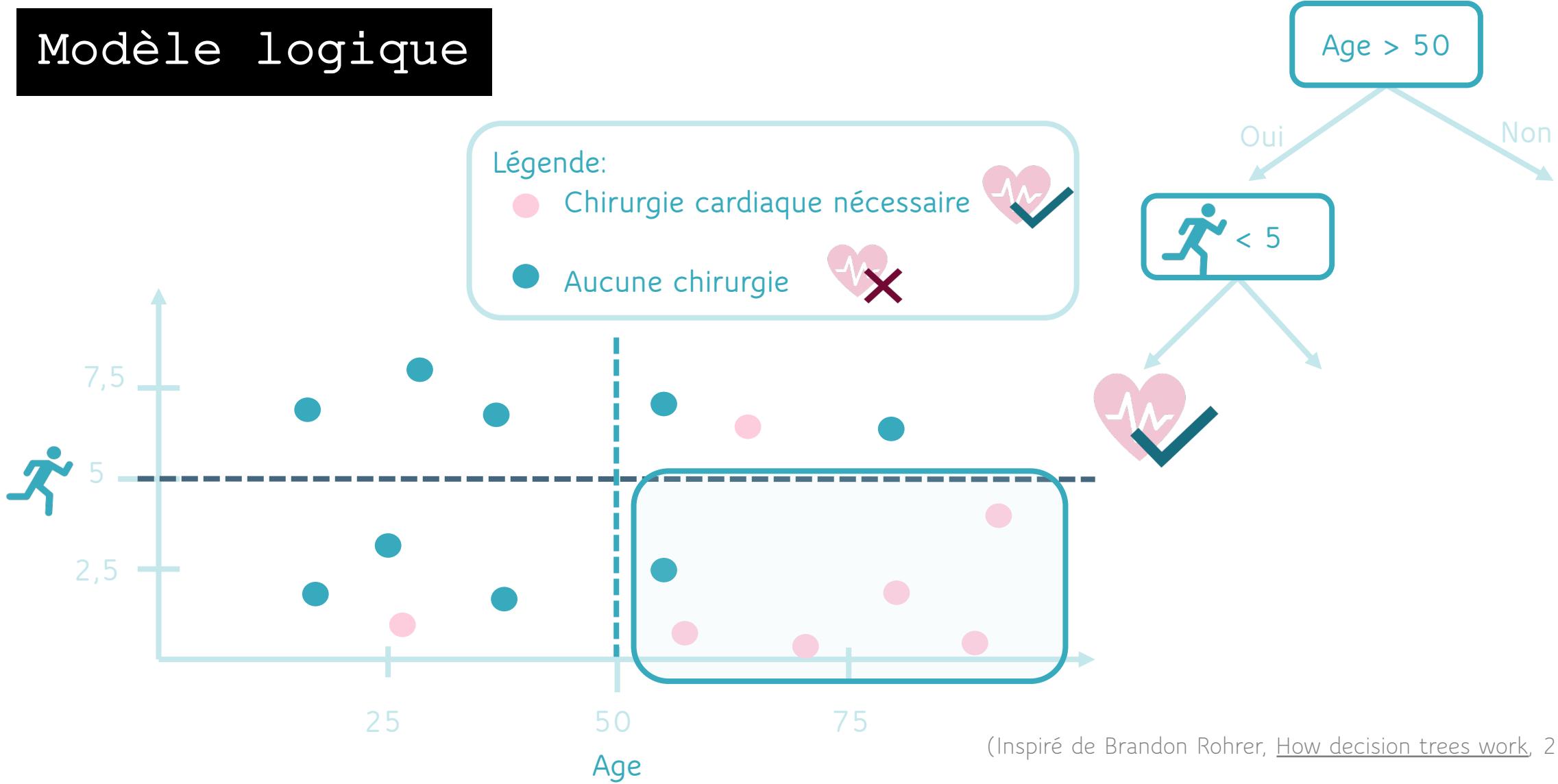
INTERPRÉTABILITÉ ↓

Modèle logique



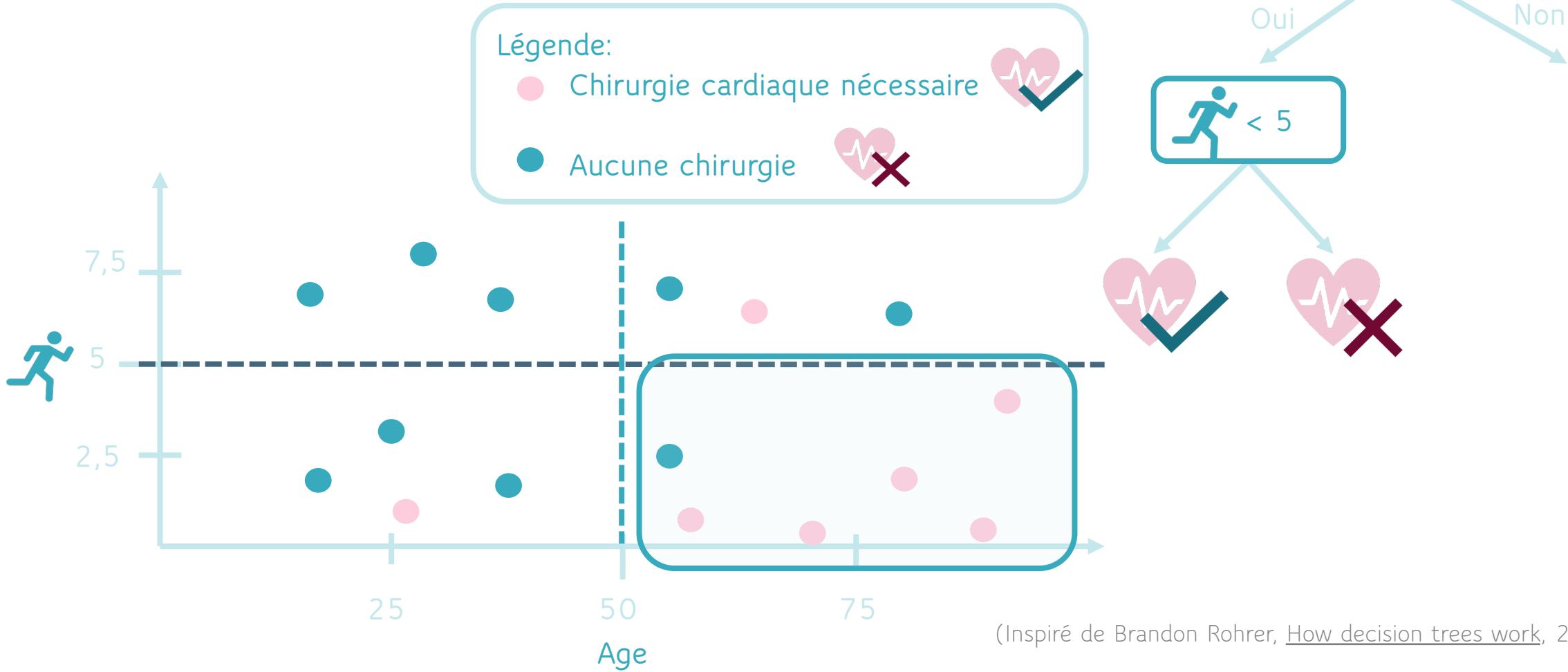
INTERPRÉTABILITÉ ↓

Modèle logique



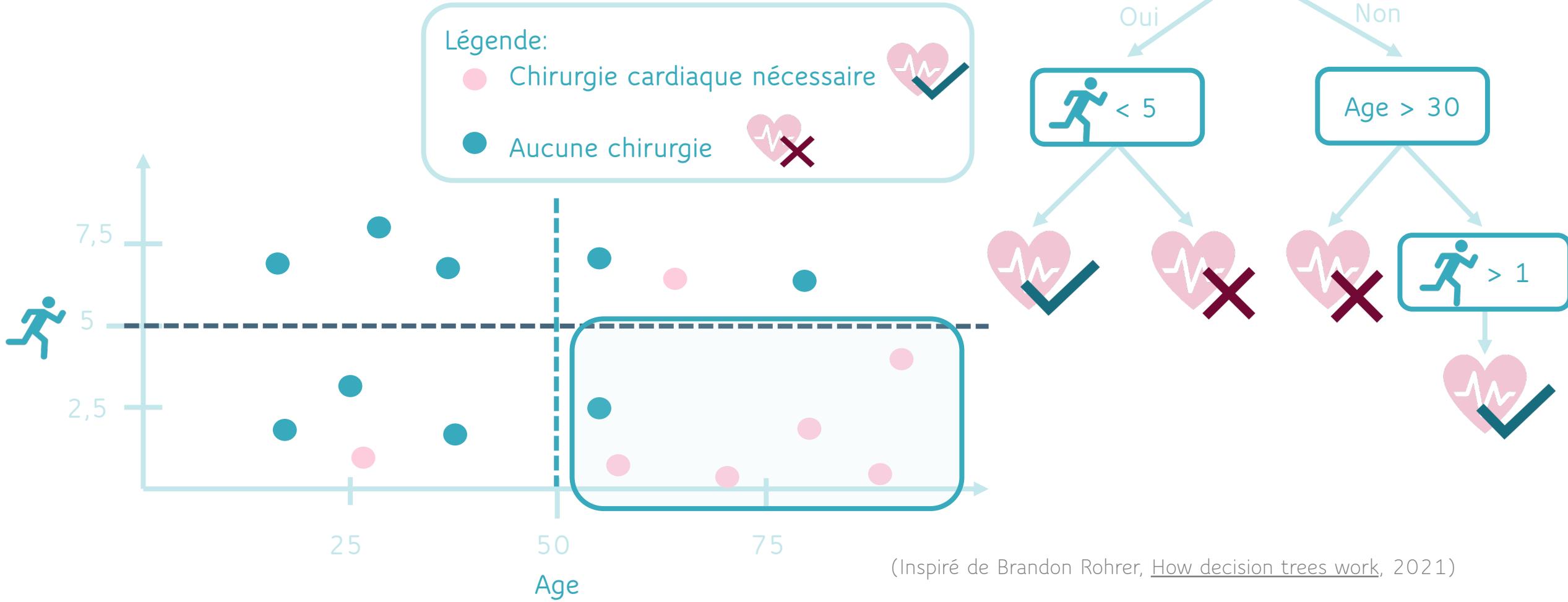
INTERPRÉTABILITÉ ↓

Modèle logique



INTERPRÉTABILITÉ ↓

Modèle logique

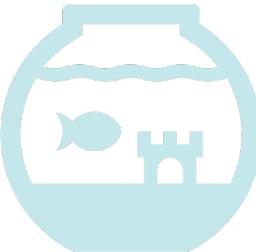
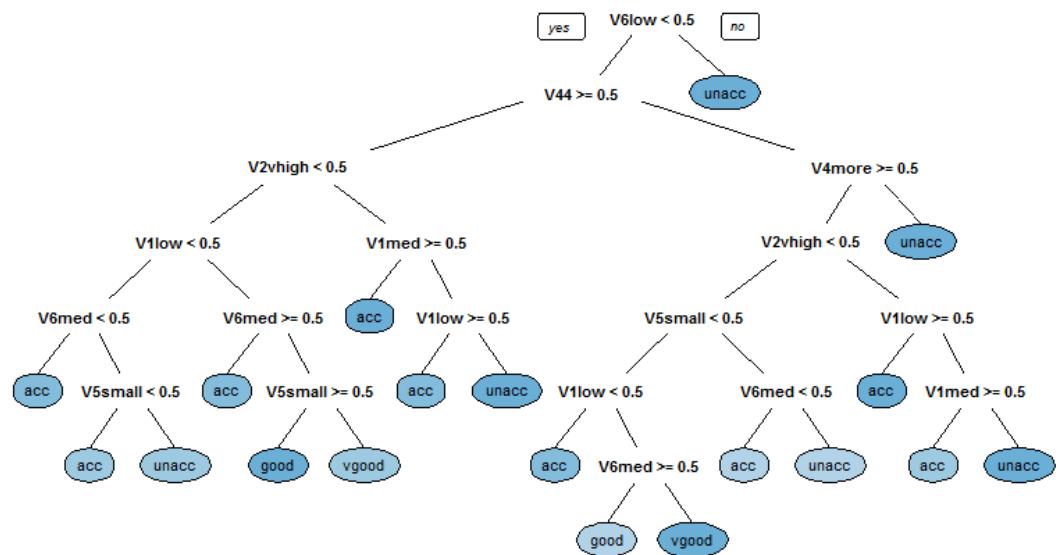


INTERPRÉTABILITÉ ↓

Parcimonie

Arbres de décision

- Limiter la profondeur de l'arbre



(<https://dataaspirant.com/decision-tree-classifier-implementation-in-r/>, mars 2022)

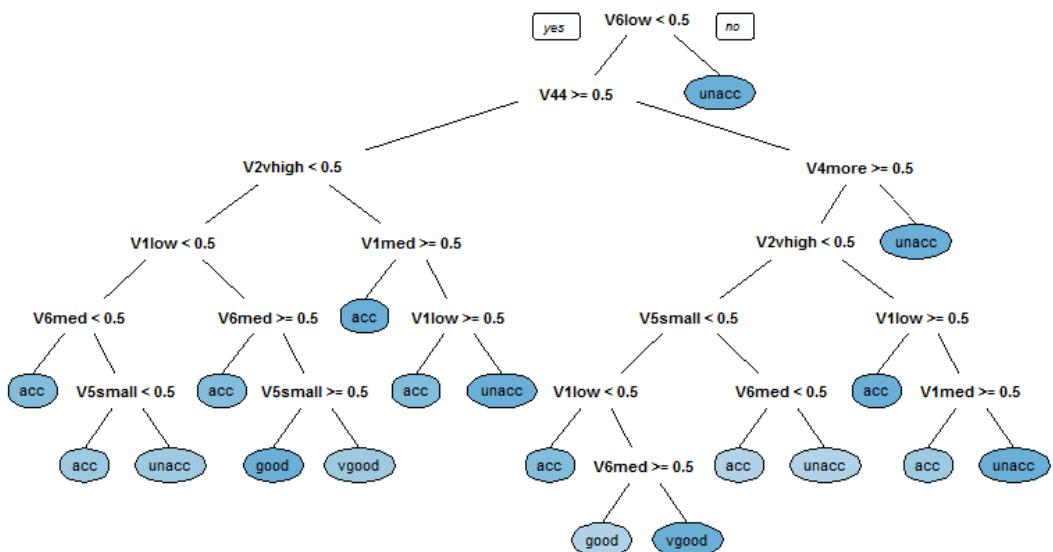
INTERPRÉTABILITÉ



Parcimonie

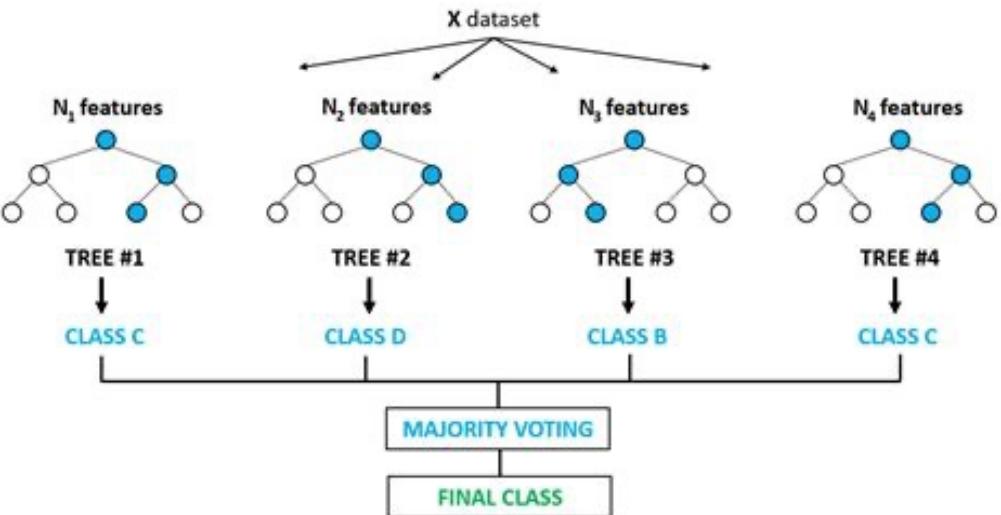
Arbres de décision

- Limiter la profondeur de l'arbre



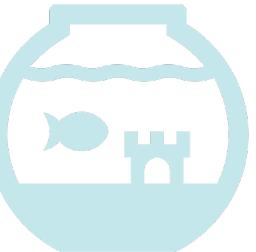
Forêts aléatoires

- Impossible de limiter le nombre d'arbres



(<https://medium.com/@ar.ingenious/applying-random-forest-classification-machine-learning-algorithm-from-scratch-with-real-24ff198a1c57>, mars 2022)

(<https://dataaspirant.com/decision-tree-classifier-implementation-in-r/>, mars 2022)



INTERPRÉTABILITÉ ↓

Type de règles

- Catégorie



INTERPRÉTABILITÉ



Type de règles

- Catégorie



- Seuil



INTERPRÉTABILITÉ



Type de règles

- Catégorie



- Seuil



- Intervalle

$20 < \text{Age} < 50$

INTERPRÉTABILITÉ



Type de règles

- Catégorie



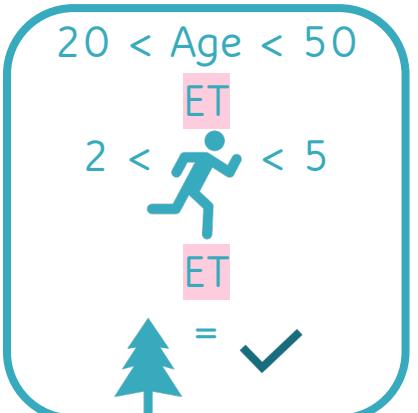
- Seuil



- Intervalle

$20 < \text{Age} < 50$

- Parallélépipède



INTERPRÉTABILITÉ



Type de règles

- Catégorie



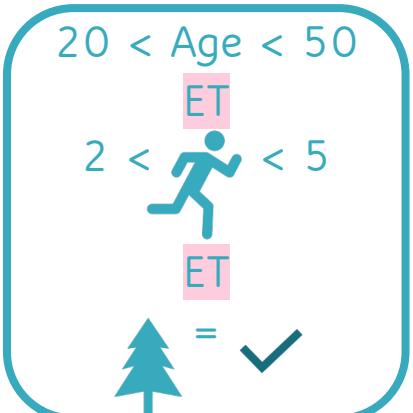
- Seuil



- Intervalle

$20 < \text{Age} < 50$

- Parallélépipède



- Oblique



INTERPRÉTABILITÉ



Type de règles

- Catégorie



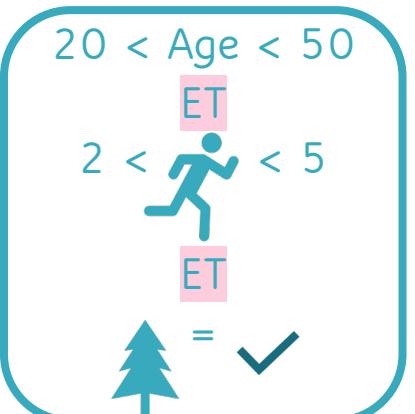
- Seuil



- Intervalle

$$20 < \text{Age} < 50$$

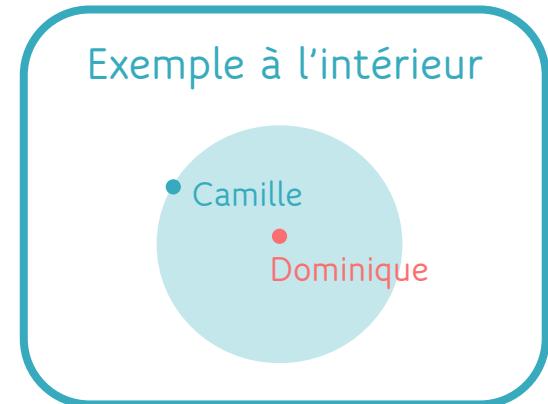
- Parallélépipède



- Oblique



- Boule



INTERPRÉTABILITÉ



Type de règles

Interprétable



- Catégorie



- Seuil



- Intervalle

$20 < \text{Age} < 50$

Interprétable



- Parallélépipède

$20 < \text{Age} < 50$
ET
 $2 < \text{Age} < 5$
ET
= ✓

- Oblique

2 + Age < 100

- Boule

Exemple à l'intérieur

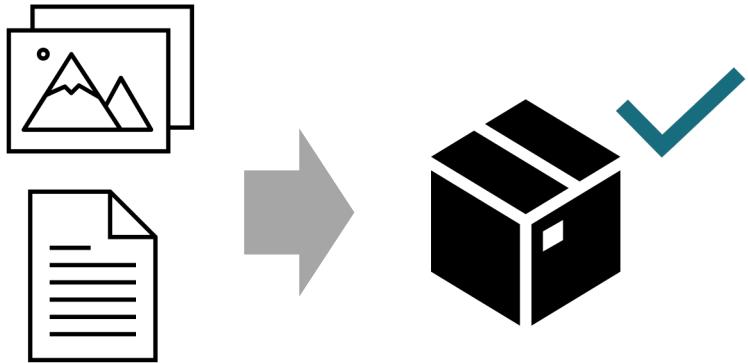
Camille
Dominique

INTERPRÉTABILITÉ



Type d'attributs

Données brutes (images, texte)

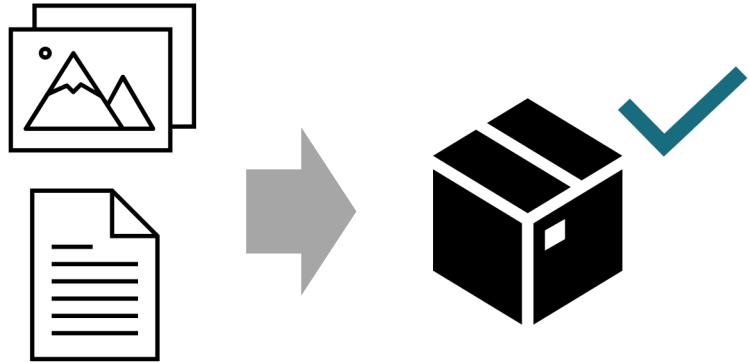


INTERPRÉTABILITÉ

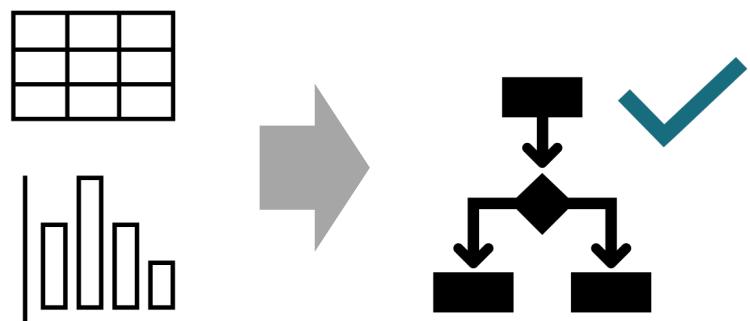


Type d'attributs

Données brutes (images, texte)



Données tabulaires

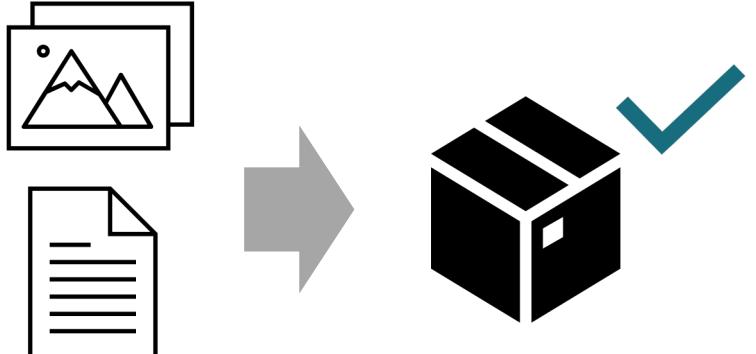


INTERPRÉTABILITÉ

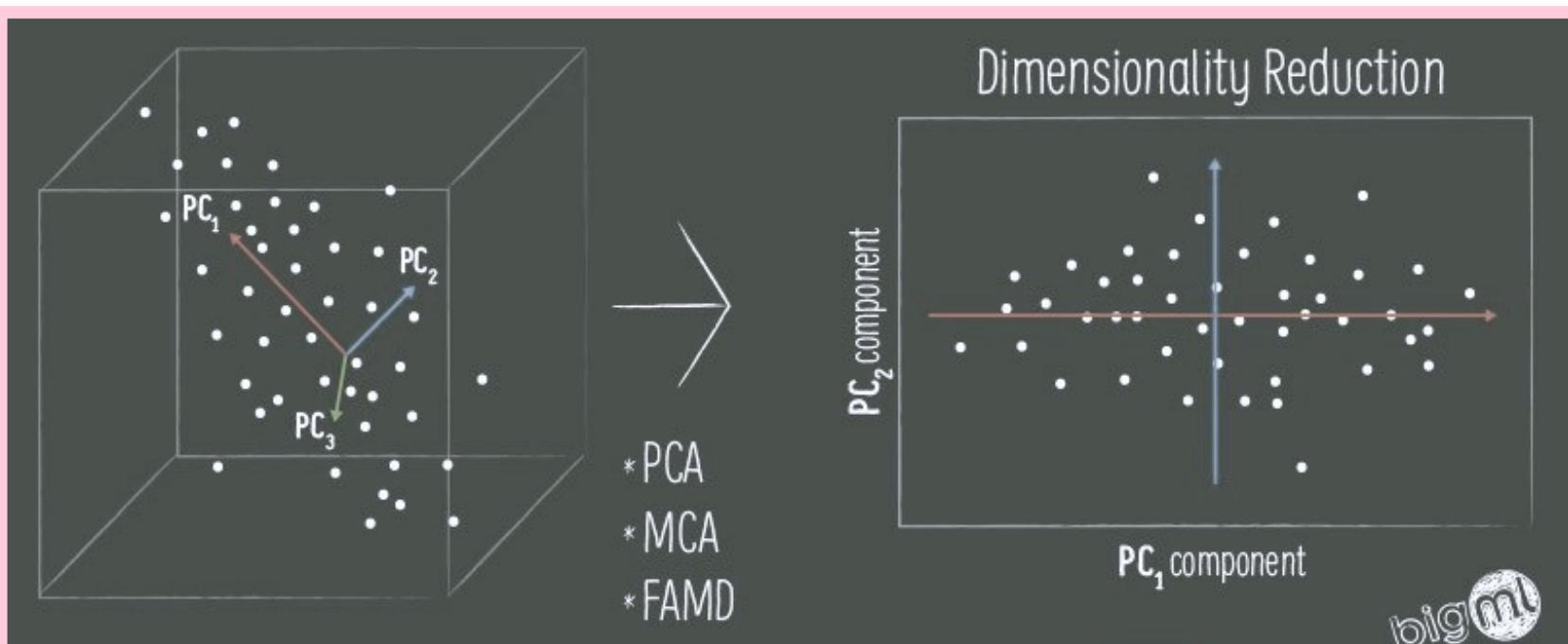
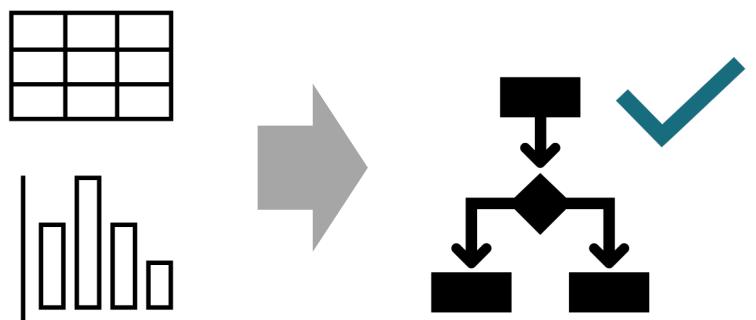


Type d'attributs

Données brutes (images, texte)



Données tabulaires



(<https://medium.com/analytics-vidhya/guide-to-principal-component-analysis-ab04a8a9c305>, mars 2022)

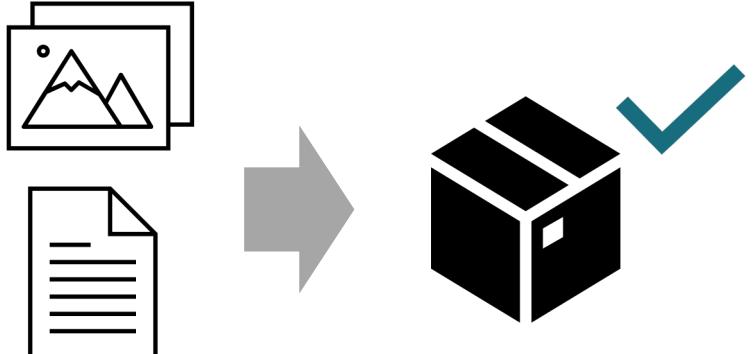
Pré traitement

- Réduction de dimensions
 - Ex: PCA
 - Attributs dénaturés

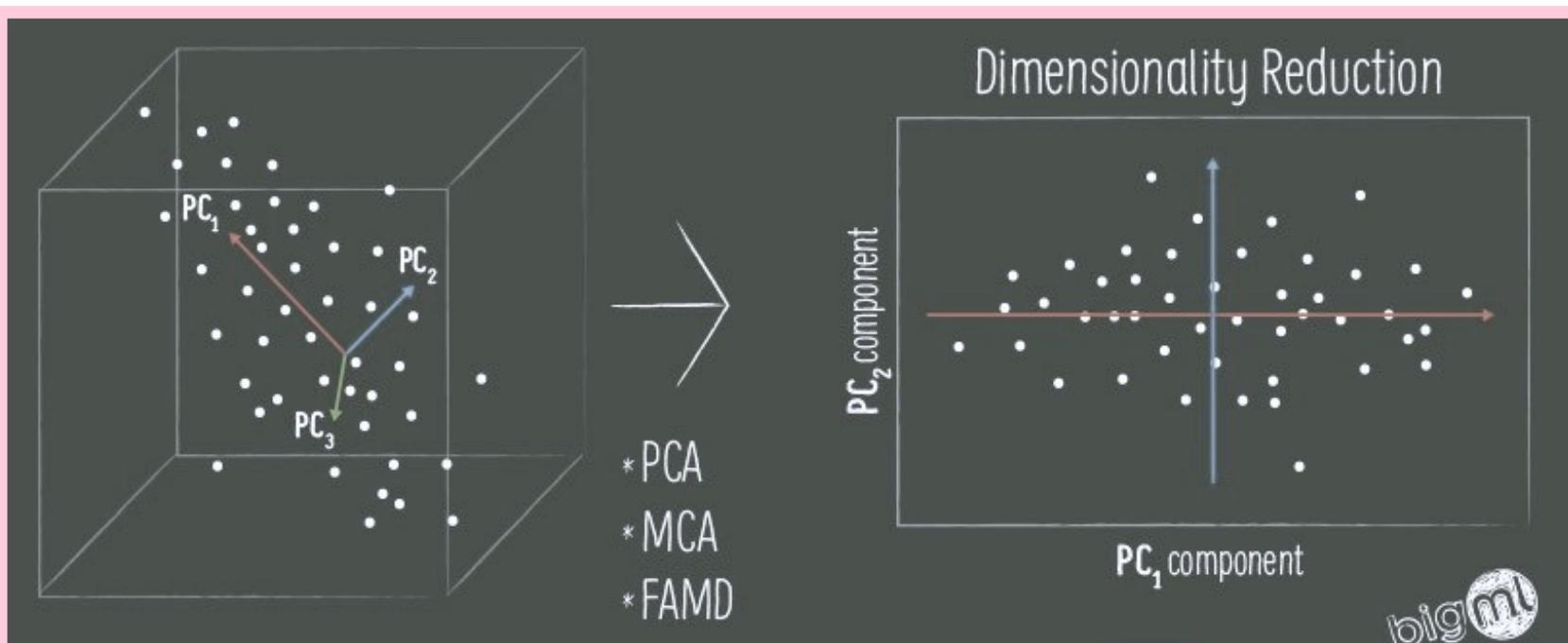
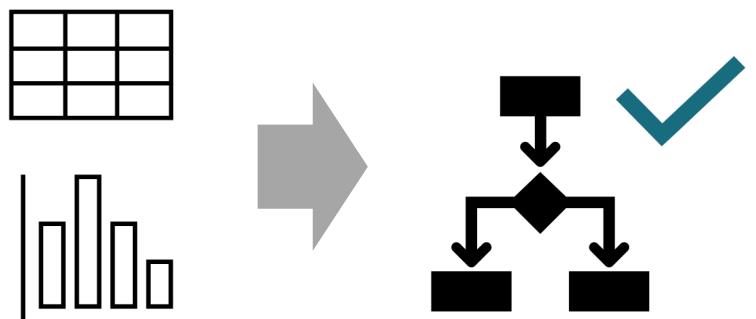
INTERPRÉTABILITÉ

Type d'attributs

Données brutes (images, texte)



Données tabulaires



(<https://medium.com/analytics-vidhya/guide-to-principal-component-analysis-ab04a8a9c305>, mars 2022)

Pré traitement

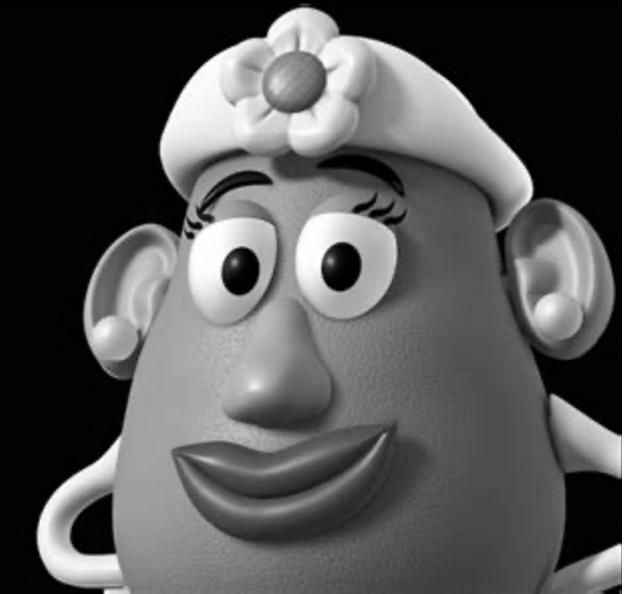
- Réduction de dimensions
 - Ex: PCA
 - Attributs dénaturés

Interprétable



PLAN DE LA PRÉSENTATION

- Introduction
- Définitions
- Pièges de l'explicabilité
- Gagner en interprétabilité
- Perdre de l'interprétabilité
- Conclusion



CONCLUSION

En résumé

■ Risques

- Utilisation de **boîtes noires** (sans interprétabilité ni explicabilité)
- Confondre interprétabilité et explicabilité
- Limites aux méthodes explicabilité
- Possible de gagner ou de perdre de l'interprétabilité

■ Interprétabilité

- Réponse de l'IA contre les biais (discriminatoires)
- Réfère au modèle en soi

■ Explicabilité

- Cherche à expliquer comment une prédiction a été effectuée



CONCLUSION

Perspectives

- Responsabilité en IA est partagée
 - Concepteurs
 - Entraîneurs
 - Utilisateurs
- Interprétabilité : intérêt commun
 - Spécialistes de l'IA
 - Spécialistes de l'éthique

Explication

- Bientôt obligatoire
 - Projet de loi 64
 - RGPD
- Bonne explication
- Pour qui ?



MERCI



sandrine.blais-deschenes.1@ulaval.ca

Présentation disponible au : <https://github.com/SandrineBD/Beneva/>

RÉFÉRENCES

- Aïvodji, U., Arai, H., Gambs, S., & Hara, S. (2021). Characterizing the risk of fairwashing. *ArXiv*.
- Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2018). Learning Certifiably Optimal Rule Lists for Categorical Data. *arXiv:1704.01701 [cs, stat]*. <http://arxiv.org/abs/1704.01701>
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable Artificial Intelligence (XAI) : Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *arXiv:1910.10045 [cs]*. <http://arxiv.org/abs/1910.10045>
- Aylwin, F.-A. (2020, novembre 22). *Projet de loi 64 : Quel est l'impact concret du projet de loi sur les ordres professionnels?* Lexology. <https://www.lexology.com/library/detail.aspx?g=652cc798-0828-4023-8781-0eb218e1aee6>
- Buolamwini, J., & Gebru, T. (2018). Gender Shades : Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77-91. <https://proceedings.mlr.press/v81/buolamwini18a.html>
- Gilmore, E., Estivill-Castro, V., & Hexel, R. (2021). More Interpretable Decision Trees. Dans H. Sanjurjo González, I. Pastor López, P. García Bringas, H. Quintián, & E. Corchado (Éds.), *Hybrid Artificial Intelligent Systems* (p. 280-292). Springer International Publishing. https://doi.org/10.1007/978-3-030-86271-8_24
- Marchand, M., & Shawe-Taylor, J. (2002). The Set Covering Machine. *Journal of Machine Learning Research*, 3(4-5), 723-746. <https://eprints.soton.ac.uk/259011/>
- Morocho-Cayamcela, M. E., Lee, H., & Lim, W. (2019). Machine Learning for 5G/B5G Mobile and Wireless Communications : Potential, Limitations, and Future Directions. *IEEE Access*, 7, 137184-137206. <https://doi.org/10.1109/ACCESS.2019.2942390>
- Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable-ml-book/
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215. <https://doi.org/10.1038/s42256-019-0048-x>
- Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2021). Interpretable Machine Learning : Fundamental Principles and 10 Grand Challenges. *arXiv:2103.11251 [cs, stat]*. <http://arxiv.org/abs/2103.11251> © Sandrine Blais-Deschênes

Images



- <https://www.galeriaplakatu.com/motywacyjne/typographic-potato-art-print>
- <https://i.pinimg.com/200x150/43/02/84/4302842ae9394c827b5dc2cbf5737261.jpg>
- https://www.pngfind.com/download/iTRITmh_seor-patata-toy-story-para-imprimir-mr-potato/
- <https://flickr.com/photos/elycefeliz/4287879280/in/photostream/>
- <https://www.infinitehollywood.com/category/mr-potato-head/>
- https://mrpotatohead.fandom.com/wiki/Mr_Potato_Head_Wiki?file=Mr+Mrs+Potato+Head+TS3.png
- https://www.pngfind.com/mpng/imwTxJ_mrs-potato-head-mrs-potato-head-toy-story/