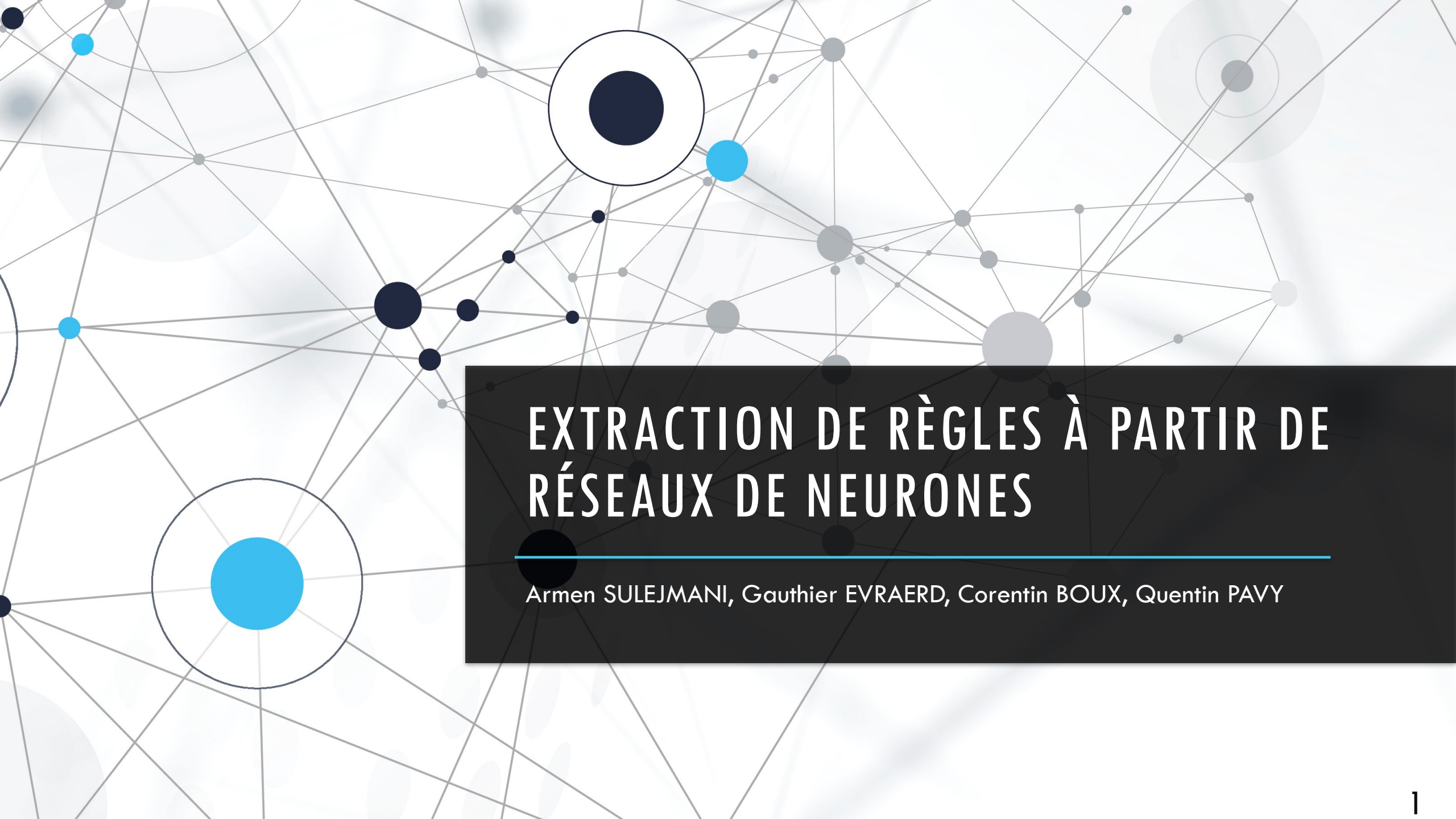
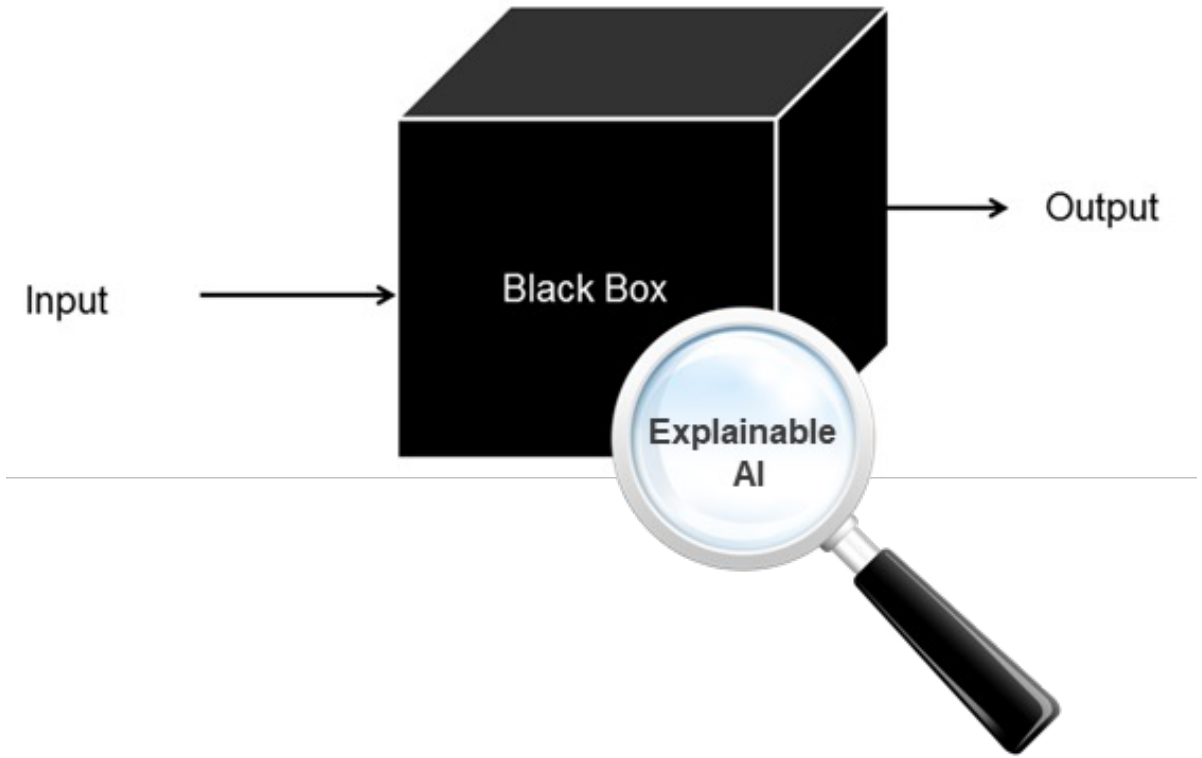


# EXTRACTION DE RÈGLES À PARTIR DE RÉSEAUX DE NEURONES

Armen SULEJMANI, Gauthier EVRAERD, Corentin BOUX, Quentin PAVY

# INTRODUCTION

**Le problème de la boîte noire**



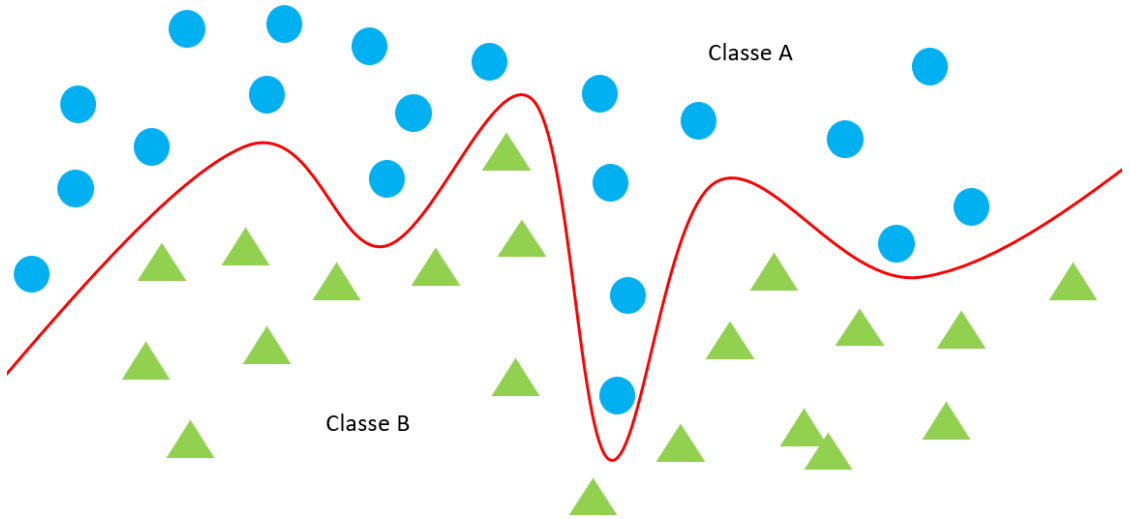

- Problème dans des contextes critiques en matière de sécurité (santé…)
- Rends difficile leur adoption à grande échelle

**Apparition de méthodes pour expliquer ces boîtes noires**

# GLOBALE VS LOCALE

## Globale:

- Comprendre le comportement sur l'ensemble du jeu de données
- Frontière très complexe
- Difficile à expliquer de manière simple le comportement

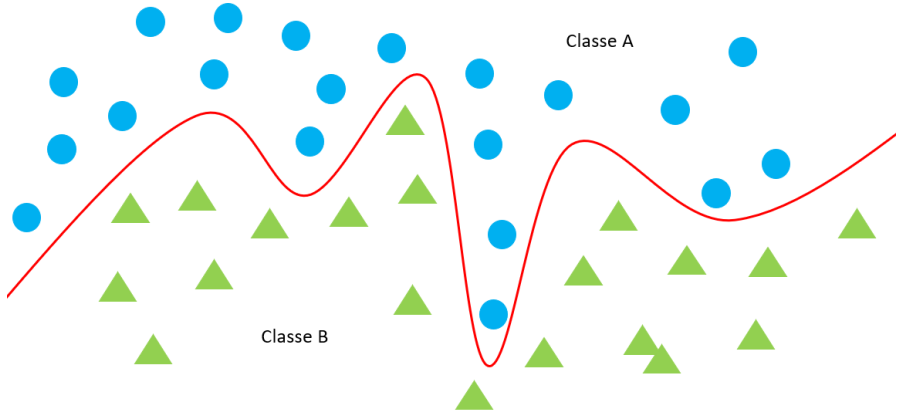



*Frontière de décision globale*

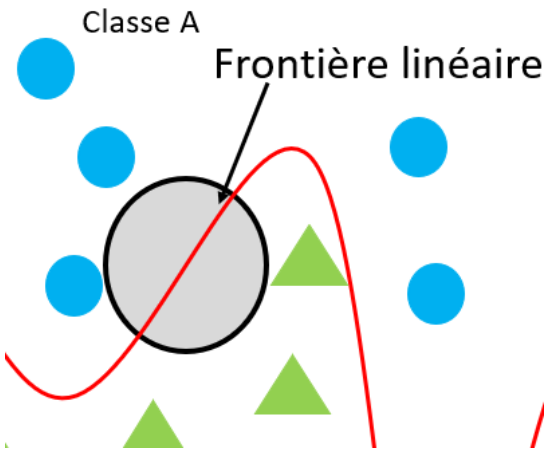
# GLOBALE VS LOCALE

## Locale:

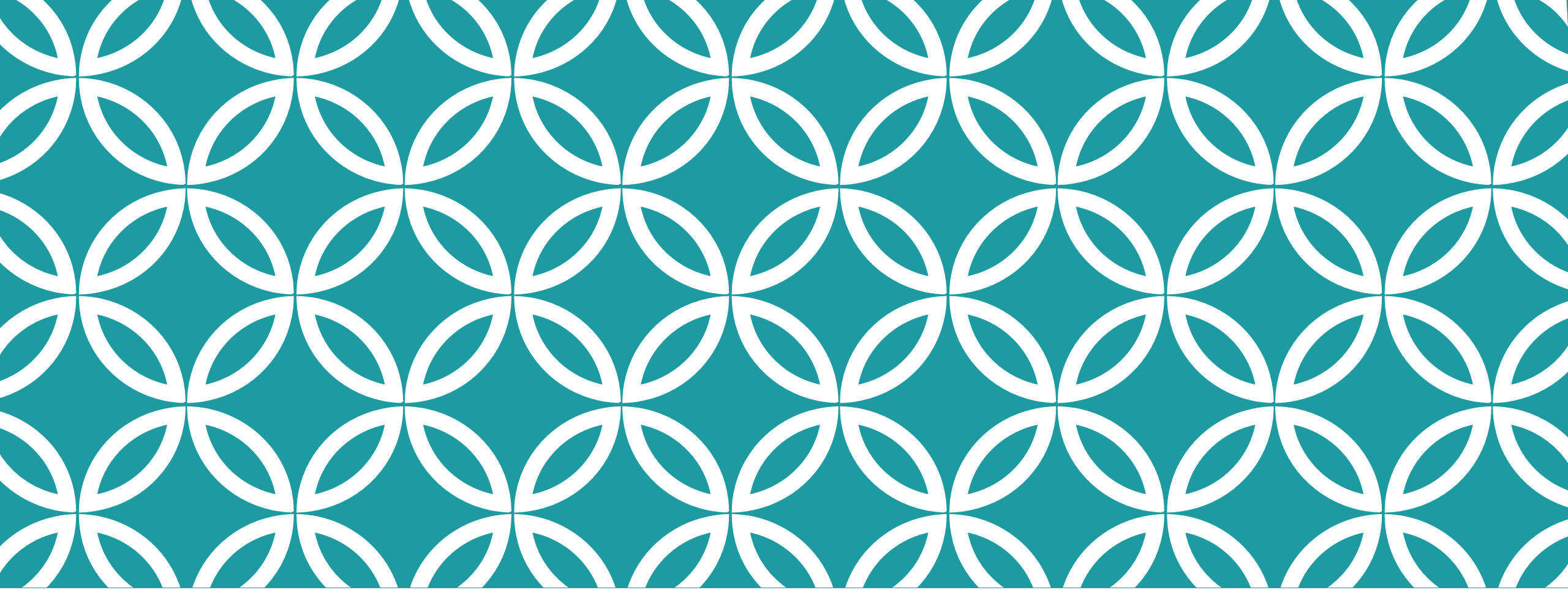
- Comprendre le comportement pour une entrée spécifique
- Frontière souvent simple
- Comportement facile à expliquer (arbre de décision)




*Frontière de décision globale*




*Frontière de décision locale*



# MÉTHODE GLOBALE : GLOCALX

# GLOCALX

- Méthode Agnostique
- A partir d'explication locale d'un modèle
- Obtenir une explication globale de ce modèle
- Forme:

$$e = \langle r = P \rightarrow y \rangle$$

- r : règle de décision
- P : Ensemble de prémisses
- y : La prédiction de la règle
- Exemple:

$$e = \langle r = \{age \geq 25, job = unemployed, amount \leq 10k\} \rightarrow deny \rangle$$

# GLOCALX: SIMILARITY

- Cover(e,x): e couvre x si P satisfaits par x
- Cover(E,x): E couvre x si e∈E,
  Cover(e,x) => Vrai
- Coverage(E,X) => $\{x_1,\ldots,x_m\}\in X$,
  Cover(E,$x_i$) => Vrai
- Covered(x,E) => $\{e_1,\ldots,e_m\}\in E$,
  Cover($e_i$,x)

$$similarity_X(E_i, E_j) = \frac{|coverage(E_i, X) \cap coverage(E_j, X)|}{|coverage(E_i, X) \cup coverage(E_j, X)|}$$

# GLOCALX: ALGORITHME

**Algorithm 1** GLocalX($\mathbb{E}, \alpha$)

**Input:** $\mathbb{E}$ explanation theories, $\alpha$ filter threshold
**Output:** $E$ explanation theory

$\mathrm{E} \leftarrow \emptyset$
**repeat**
  $\mathbb{Q} \leftarrow \mathrm{SORT}(\mathbb{E})$ ▷ sort pairs of theories by similarity
  *merged* $\leftarrow$ `False`
  $X' \leftarrow \mathrm{batch}(\mathrm{X})$
  **while** $\neg$ *merged* $\wedge$ $\mathbb{Q} \neq \emptyset$ **do**
    $E_i, E_j \leftarrow \mathrm{POP}(\mathbb{Q})$ ▷ select most similar theories
    $E_{i+j} \leftarrow \mathrm{MERGE}(E_i, E_j, X')$ ▷ merge theories
    **if** $\mathrm{BIC}(E_{i+j}) \leq \mathrm{BIC}(E_i \cup E_j)$ **then** ▷ verify improvement
      *merged* $\leftarrow$ `True`
      **break**
  **if** *merged* **then** ▷ merge occurred
    $\mathbb{E} \leftarrow \mathrm{UPDATE}(E_i, E_j, E_{i+j})$ ▷ update hierarchy
**until** $|\ \mathrm{E}\ | > 1 \wedge$ *merged* ▷ until the merge is successful
$E \leftarrow \mathrm{FILTER}(E, \alpha)$ ▷ Filter final theory
**return** $E$

# GLOCALX: JOIN ⊕

$$P \oplus Q = \{P_1 + Q_1, ..., P_m + Q_m\}$$

$$P_i + Q_i = \begin{cases} P_i \cup Q_i & \text{intersection non nulle} \\ (min\{P_i \cup Q_i\}, max\{P_i \cup Q_i\}) & \text{intersection vide} \\ \emptyset & P_i \text{ ou } Q_i \text{ est vide} \end{cases}$$

$$e_1 = \{age \geq 50, job = \text{office clerk}\} \rightarrow deny$$
$$e_2 = \{age \geq 40\} \rightarrow deny$$

$$e_1' = \{age \geq 40\} \rightarrow deny$$

# GLOCALX: CUT⊖

$$P \ominus Q = \{P_1 - Q_1, ..., P_m - Q_m\}$$

$$P_i - Q_i = \begin{cases} \{P_i, \emptyset\} & Q_i \text{ vide} \\ \{P_i, Q_i \backslash P_i\} & \text{sinon} \end{cases}$$

$$e_1 = \{age \geq 25, job = \text{unemployed}, amount \geq 10k\} \rightarrow deny$$
$$e_2 = \{age \geq 20, job = \text{manager}, amount > 8k\} \rightarrow accept$$

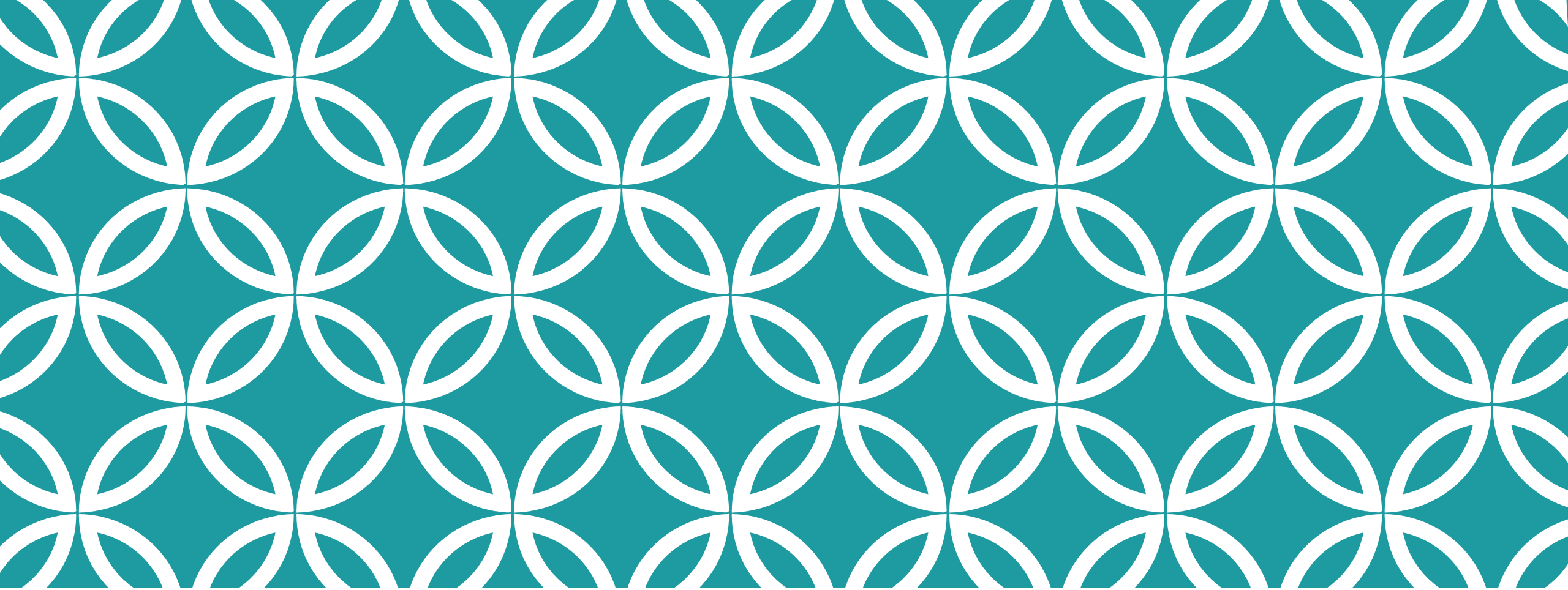
$$e_1 = \{age \geq 25, job = \text{unemployed}, amount \geq 10k\} \rightarrow deny$$
$$e_1' = \{age \in [20, 25], job = \text{manager}, amount \in [8k, 10k] \rightarrow accept$$

# GLOCALX: MERGE

**Algorithm 2** MERGE$(E_i, E_j, X)$

**Input:** $E_i, E_j$ explanation theories, $X$ batch
**Output:** $E_{(i+j)}$ explanation theory

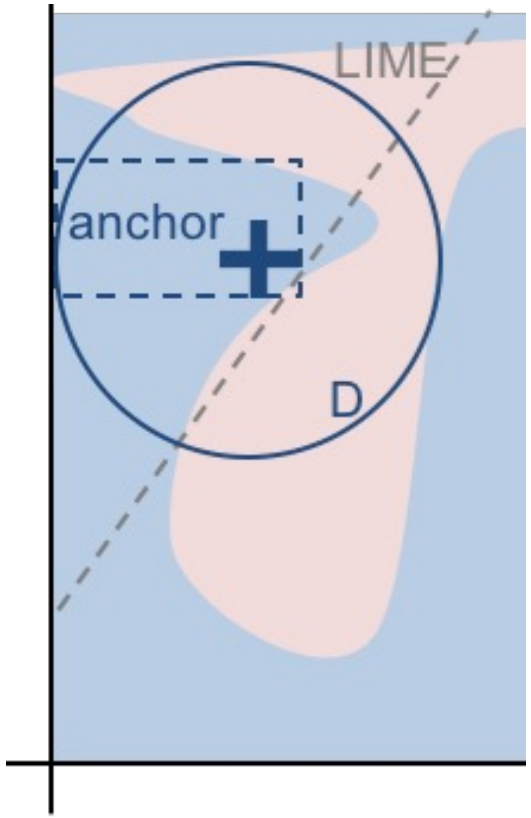
1: $\mathrm{E} \leftarrow E_i \cup E_j$
2: **for** $x \in X$ **do**
3: $C_i \leftarrow$ COVERED$(x, E_j)$ ▷ retrieve rules in $E_i$ covering $x$
4: $C_j \leftarrow$ COVERED$(x, E_j)$ ▷ retrieve rules in $E_j$ covering $x$
5: $C_= \leftarrow$ NON-CONFLICTING$(x, C_i, C_j)$ ▷ non-conflicting rules in $C_i, C_j$ and covering $x$
6: $C_\neq \leftarrow$ CONFLICTING$(x, C_i, C_j)$ ▷ non-conflicting rules in $C_i, C_j$ covering $x$
7: $\mathrm{E} \leftarrow \mathrm{E} \setminus (E^i \cup E^j)$
8: $E_= \leftarrow$ JOIN$(C_=)$
9: $E_\neq \leftarrow$ CUT$(C_\neq, X)$
10: $\mathrm{E} \leftarrow \mathrm{E} \cup E_= \cup E_\neq$
11: **return** $E$

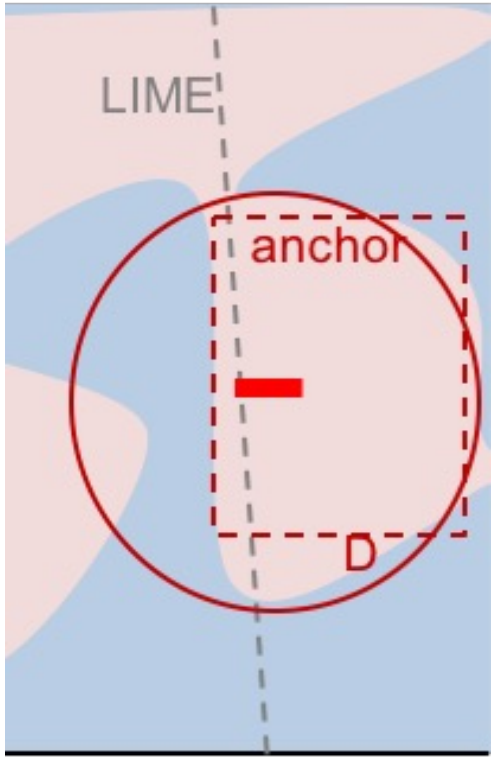


# MÉTHODES LOCALES : ANCHORS

# Anchors

- Méthode Agnostique
- Créer un modèle de prédiction interprétable sur le voisinage d'un individu
- A partir de ce modèle, il trouve une explication
- Exemple : critique de film



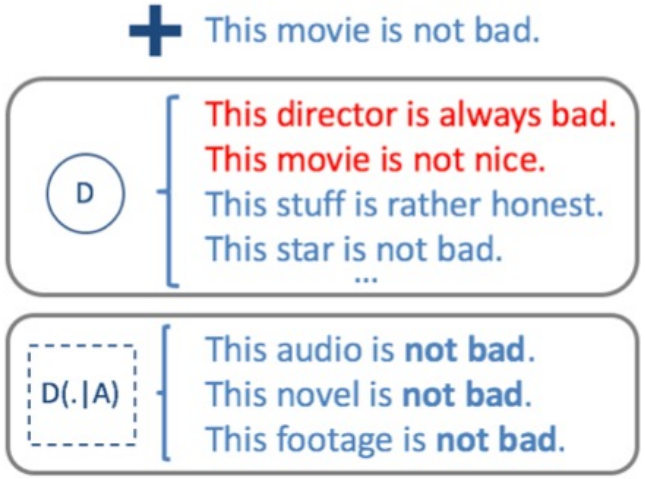



# Anchors : Exemple critique de film

"J'ai adoré ce film ! L'intrigue était captivante, les acteurs étaient excellents."

Prédicat d'ancrage simple

If 'adoré' Then 'positif'




{"not", "bad"} → Positive

# Anchors : Exemple sur données tabulaires

| | If | Predict |
|---|---|---|
| adult | No capital gain or loss, never married | $\leq$ 50K |
| | Country is US, married, work hours $>$ 45 | $>$ 50K |
| rcdv | No priors, no prison violations and crime not against property | Not rearrested |
| | Male, black, 1 to 5 priors, not married, and crime not against property | Re-arrested |
| lending | FICO score $\leq$ 649 | Bad Loan |
| | 649 $\leq$ FICO score $\leq$ 699 and \$5, 400 $\leq$ loan amount $\leq$ \$10, 000 | Good Loan |

L'état matrimonial assez déterminant

Mais

D 'autres Nécessitent une exploration plus approfondie

# Anchors

$$\mathbb{E}_{D(z|A)}\left[1[f(x) = f(z)]\right] \geq \tau, \quad A(x) = 1$$

Espérance de la Distribution par l'ancre A

Indicateur: 1si la prédiction du modèle f pour l'exemple x est égale à la prédiction pour l'exemple z, et 0 sinon.

Seuil minimum $\tau$ (tau)
Niveau de précision souhaité

# Anchors

Garantir une précision exacte = Très complexe

$$P(\text{prec}(A) \geq \tau) \geq 1 - \delta$$

Une précision satisfaisante avec une probabilité élevé

Probabilité minimale requise pour que la précision de l'ancrage A soit supérieure ou égale à $\tau$

# Anchors : GenerateCands(A,c)

- Création de l'ensemble d'ancres A.
- Ajout itératif de candidats ancrés en étendant A.
- Choix des candidats basé sur l'amélioration de la couverture.
- Préférence pour une haute couverture avec des prédicats significatifs.
- Optimisation de l'explication en privilégiant les candidats à haute couverture.

**Algorithm 1** Identifying the *Best* Candidate for Greedy

**function** GenerateCands($\mathcal{A}, c$)
  $\mathcal{A}_r = \emptyset$
  **for all** $A \in \mathcal{A}; a_i \in x, a_i \notin A$ **do**
    **if** $\text{cov}(A \wedge a_i) > c$ **then** {Only high-coverage}
      $\mathcal{A}_r \leftarrow \mathcal{A}_r \cup (A \wedge a_i)$ {Add as potential anchor}
  **return** $\mathcal{A}_r$ {Candidate anchors for next round}

# Anchors : BestCand($\mathcal{A}, \mathcal{D}, \epsilon, \delta$)

- Estimation initiale des précisions & Sélection de règle A initial
- Itération pour améliorer la précision
- Génération d'échantillons
- Actualisation des estimations
- Sélection de la règle optimale
- Renvoi de la règle optimale

**function** BestCand($\mathcal{A}, \mathcal{D}, \epsilon, \delta$)
  **initialize** prec, $\text{prec}_{ub}$, $\text{prec}_{lb}$ estimates $\forall A \in \mathcal{A}$
  $A \leftarrow \arg\max_A \text{prec}(A)$
  $A' \leftarrow \arg\max_{A' \neq A} \text{prec}_{ub}(A', \delta)$ {$\delta$ implicit below}
  **while** $\text{prec}_{ub}(A') - \text{prec}_{lb}(A) > \epsilon$ **do**
    **sample** $z \sim \mathcal{D}(z|A), z' \sim \mathcal{D}(z'|A')$ {Sample more}
    **update** prec, $\text{prec}_{ub}$, $\text{prec}_{lb}$ for $A$ and $A'$
    $A \leftarrow \arg\max_A \text{prec}(A)$
    $A' \leftarrow \arg\max_{A' \neq A} \text{prec}_{ub}(A')$
  **return** $A$

# Anchors : BeamSearch($f$, $x$, $\mathcal{D}$, $\tau$ )

Initialisation : $A*$ à null, $A_0$ à un ensemble vide.

- Recherche par faisceau :

Itération pour les règles candidates.
Génération de nouveaux candidats en considérant la meilleure règle actuelle $A*$.
Sélection des meilleurs candidats avec LUCB.

- Si aucun candidat, fin de la recherche.

- Filtrage des candidats sur la précision supérieure

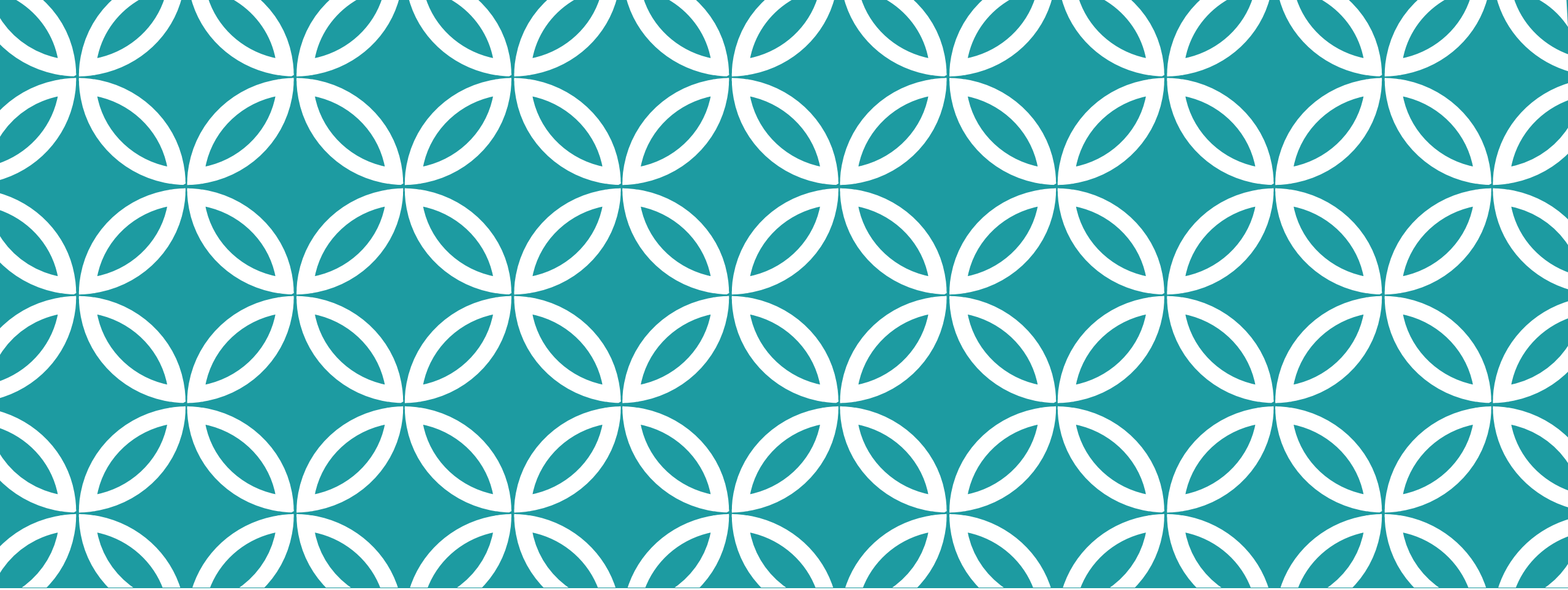
- Mise à jour de la meilleure règle

Résultat : Meilleure règle $A*$

Cette fonction effectue une recherche progressive pour trouver la meilleure règle candidate, en considérant la couverture et la précision, et s'arrête lorsque le seuil de précision est atteint.

**Algorithm 2** Outline of the Beam Search

**function** BeamSearch($f, x, \mathcal{D}, \tau$)
  **hyperparameters** $B, \epsilon, \delta$
  $A^* \leftarrow$ **null**, $\mathcal{A}_0 \leftarrow \emptyset$ {Set of candidate rules}
  **loop**
    $\mathcal{A}_t \leftarrow$ GenerateCands($\mathcal{A}_{t-1}$, cov($A^*$))
    $\mathcal{A}_t \leftarrow$ B-BestCand($\mathcal{A}_t, \mathcal{D}, B, \delta, \epsilon$) {LUCB}
    **if** $\mathcal{A}_t = \emptyset$ **then break loop**
    **for all** $A \in \mathcal{A}_t$ s.t. $\text{prec}_{lb}(A, \delta) > \tau$ **do**
      **if** cov($A$) > cov($A^*$) **then** $A^* \leftarrow A$
  **return** $A^*$



# MÉTHODES LOCALES: LORE

# Local Rule-Based Explanations (LORE)

- Méthode Agnostique
- Créer un modèle de prédiction interprétable sur le voisinage d'un individu
- A partir de ce modèle, LORE trouve une explication
- Résultat :

$$e = \langle r = p \rightarrow y, \varphi \rangle$$

Règle de décision

Règles contrefactuelles

**Algorithm 1:** $LORE(x, b)$

**Input** : $x$ - instance to explain, $b$ - black box, $N$ - # of neighbors
**Output**: $e$ - explanation of $x$

1 $G \leftarrow 10$; $pc \leftarrow 0.5$; $pm \leftarrow 0.2$; // init. parameters
2 $Z_{=} \leftarrow GeneticNeigh(x, fitness^{x}_{=}, b, N/2, G, pc, pm)$ // generate neigh.
3 $Z_{\neq} \leftarrow GeneticNeigh(x, fitness^{x}_{\neq}, b, N/2, G, pc, pm)$ // generate neigh.
4 $Z \leftarrow Z_{=} \cup Z_{\neq}$; // merge neighborhoods
5 $c \leftarrow BuildTree(Z)$; // build decision tree
6 $r = (p \rightarrow y) \leftarrow ExtractRule(c, x)$; // extract decision rule
7 $\Phi \leftarrow ExtractCounterfactuals(c, r, x)$; // extract counterfactuals
8 **return** $e = \langle r, \Phi \rangle$;

# Local Rule-Based Explanations (LORE)

On commence par créer deux populations à l'aide d'un algorithme génétique:

- $Z_{=}$ où $\forall z \in Z_{=}, b(z) = b(x)$
- $Z_{\neq}$ où $\forall z \in Z_{\neq}, b(z) \neq b(x)$
- $b(z)$ : la prédiction du modèle

**Algorithm 2:** *GeneticNeigh*$(x, fitness, b, N, G, pc, pm)$

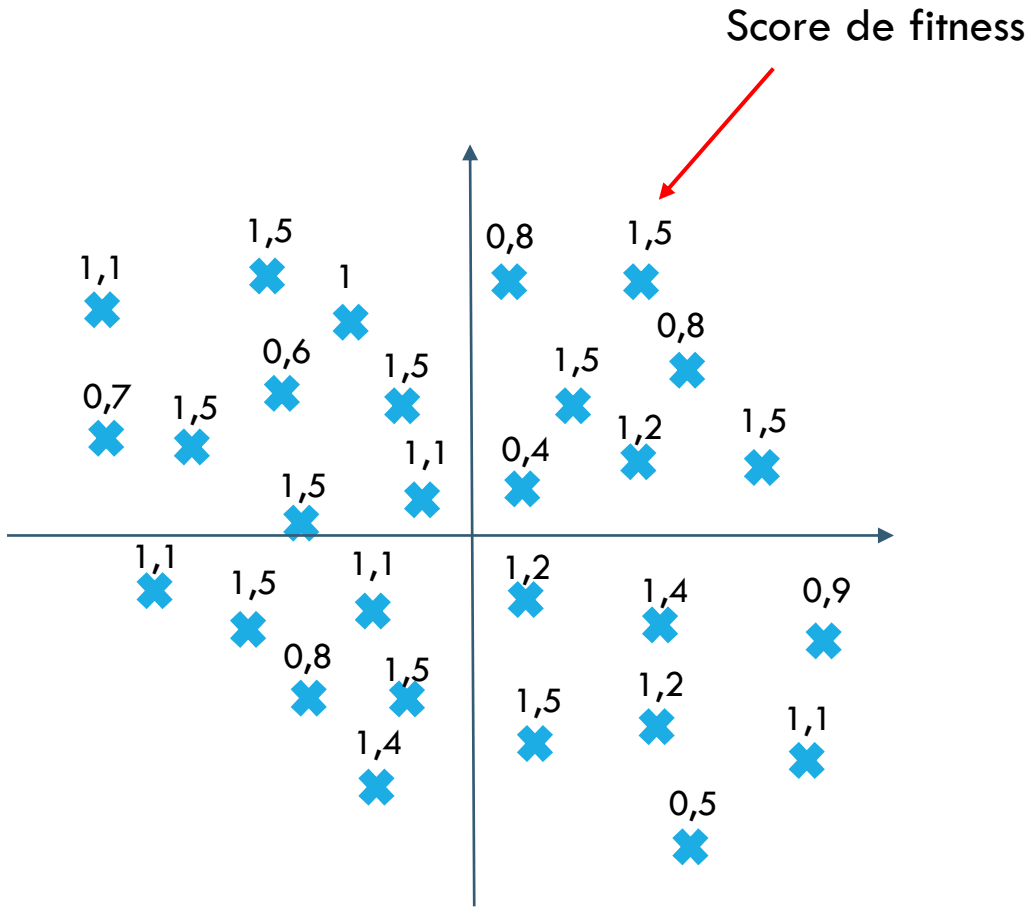
**Input** : $x$ - instance to explain, $b$ - black box, *fitness* - fitness function, $N$ - population size, $G$ - # of generations, $pc$ - crossover probability, $pm$ - mutation probability

**Output**: $Z$ - neighbors of $x$

1 $P_0 \leftarrow \{x | \forall 1 \ldots N\}; i \leftarrow 0;$ // population init.
2 *evaluate*$(P_0, fitness, b);$ // evaluate population
3 **while** $i < G$ **do**
4 $\quad P_{i+1} \leftarrow select(P_i);$ // select sub-population
5 $\quad P'_{i+1} \leftarrow crossover(P_{i+1}, pc);$ // mix records
6 $\quad P''_{i+1} \leftarrow mutate(P'_{i+1}, pm);$ // perform mutations
7 $\quad evaluate(P''_{i+1}, fitness, b);$ // evaluate population
8 $\quad P_{i+1} = P''_{i+1}; i \leftarrow i + 1$ // update population
9 **end**
10 $Z \leftarrow P_i$ **return** $Z;$

*Algorithme génétique*

# ALGORITHME GÉNÉTIQUE: EVALUATION




Fonction de fitness :

$$fitness_{=}^{x}(z) = I_{b(x)=b(z)} + \big(1 - d(x,z)\big) - I_{x=z}$$

$$fitness_{\neq}^{x}(z) = I_{b(x)\neq b(z)} + \big(1 - d(x,z)\big) - I_{x=z}$$

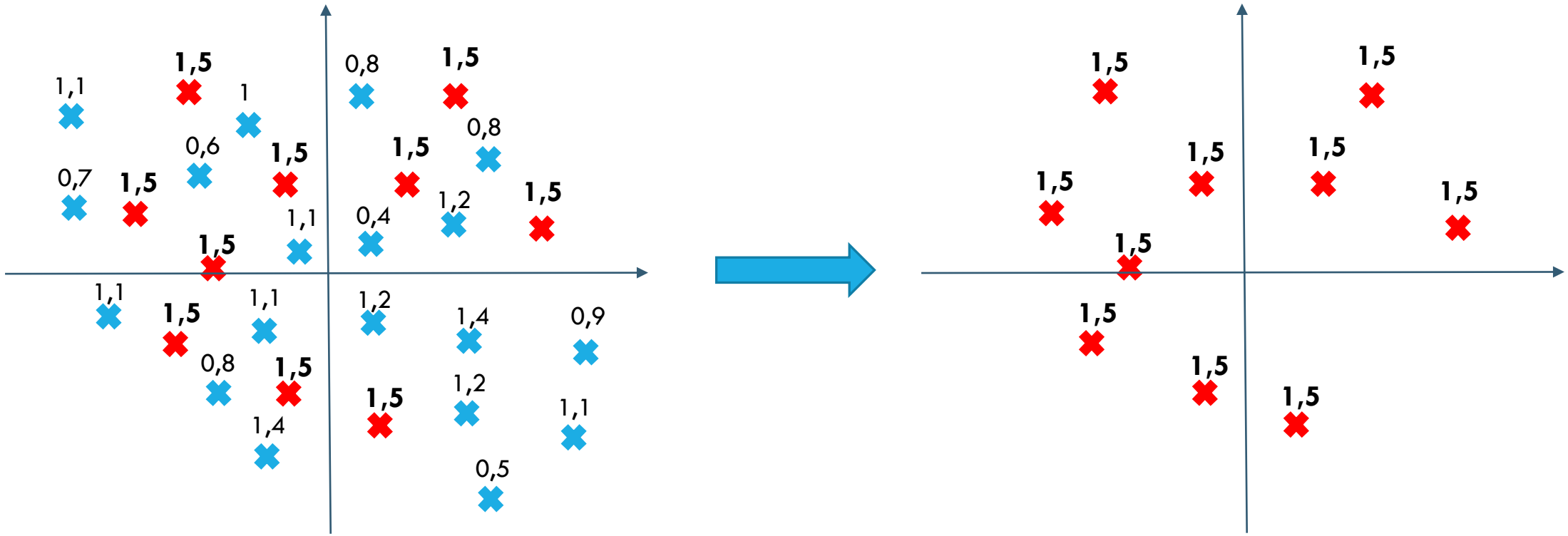
Valeur de I:

$$I_{true} = 1 \qquad I_{false} = 0$$

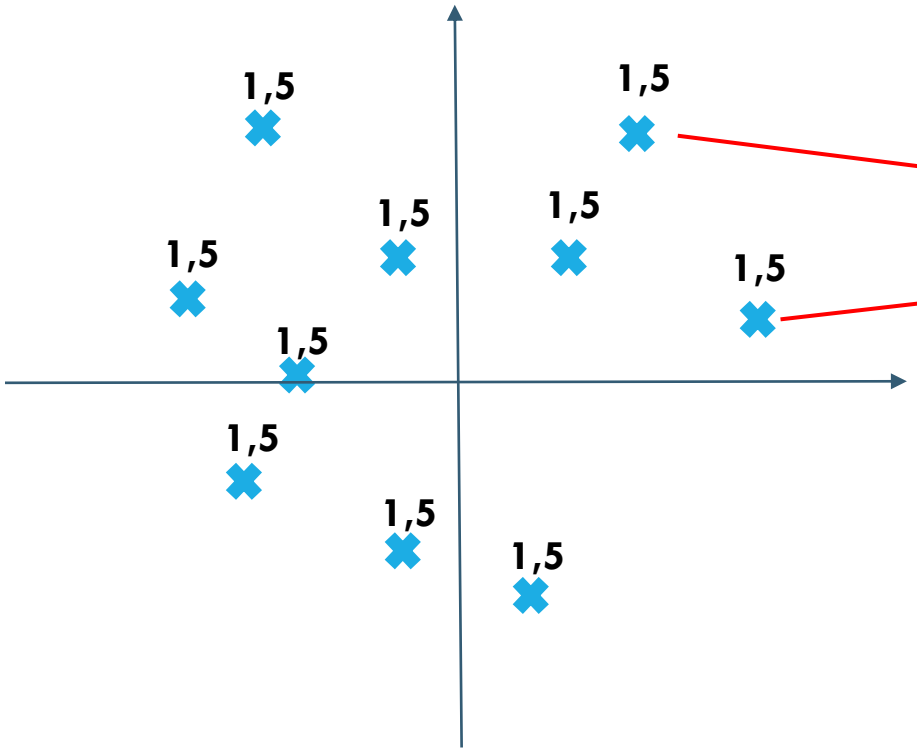
Pour calculer la distance :

$$d(x,z) = \frac{h}{m} \cdot SimpleMatch(x,z) + \frac{m-h}{m} \cdot NormEuclid(x,z)$$

# ALGORITHME GÉNÉTIQUE: SÉLECTION



# ALGORITHME GÉNÉTIQUE: CROSSOVER

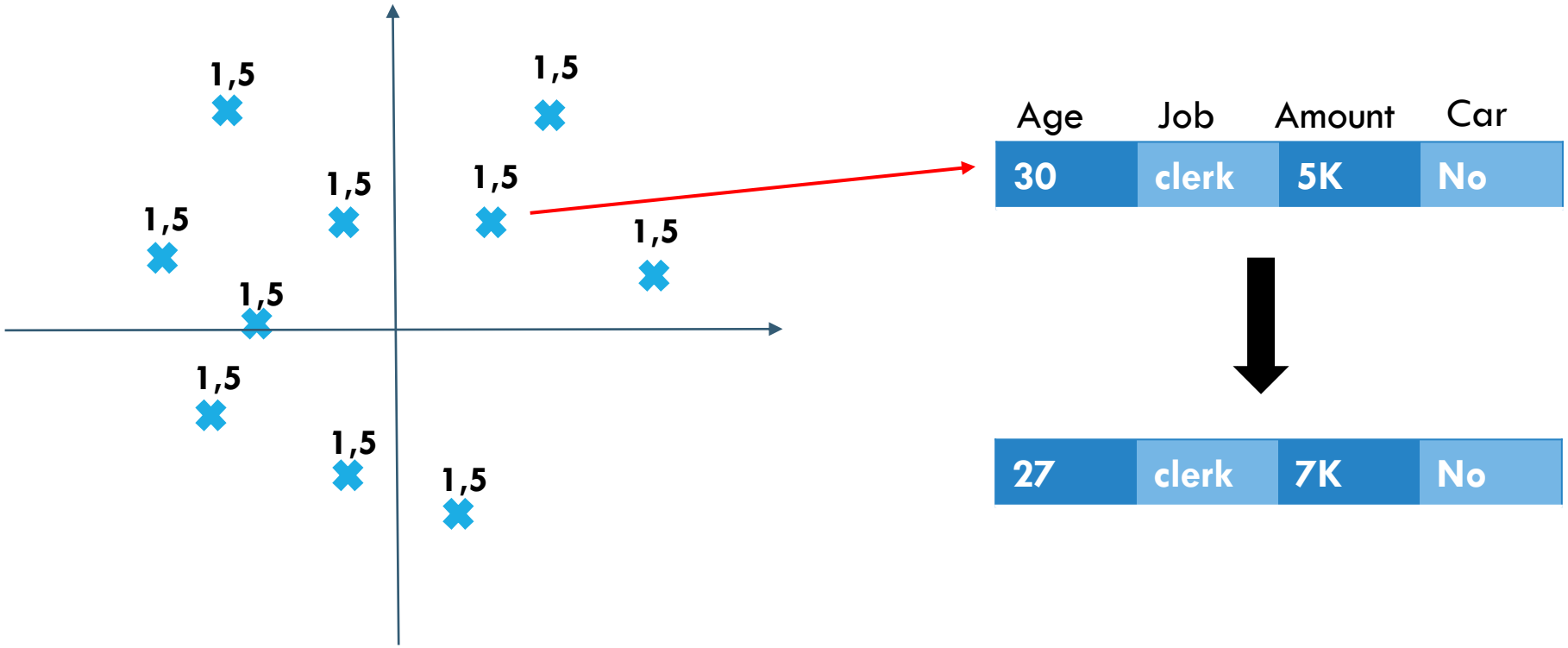



| Age | Job | Amount | Car |
|---|---|---|---|
| 25 | Clerk | 10K | Yes |
| 30 | Other | 5K | No |

| | | | |
|---|---|---|---|
| 25 | other | 5K | Yes |
| 30 | clerk | 10K | No |

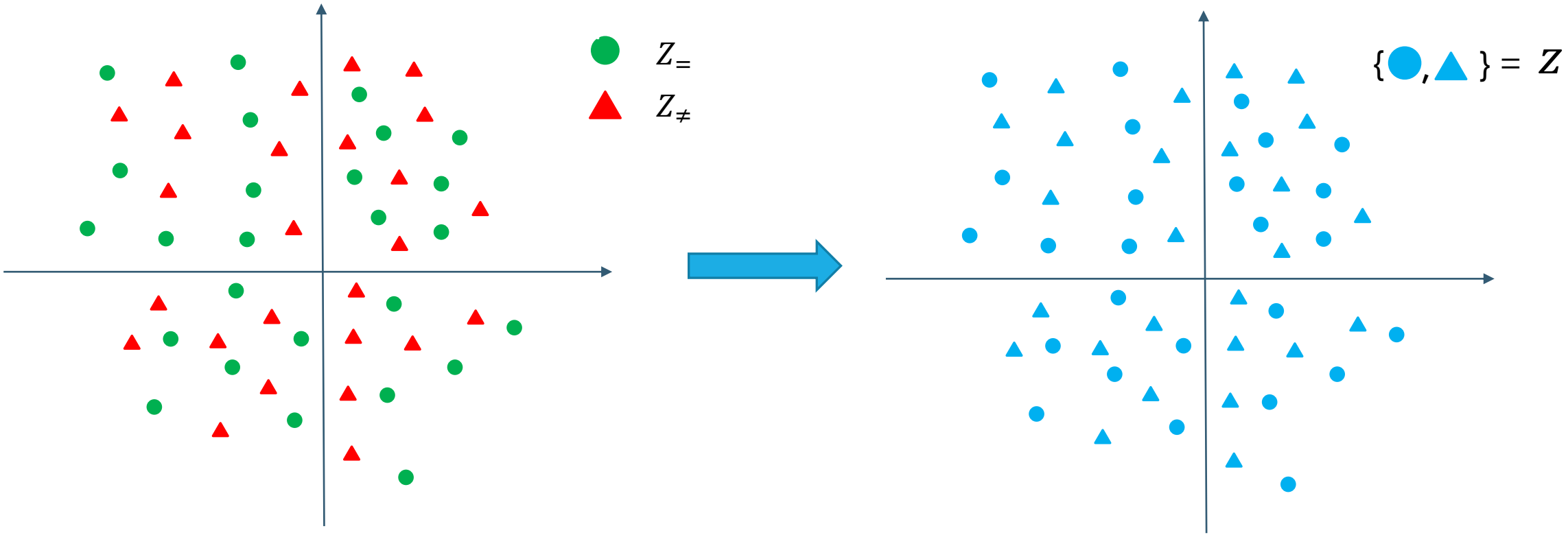
Two-point crossover

# ALGORITHME GÉNÉTIQUE: MUTATION

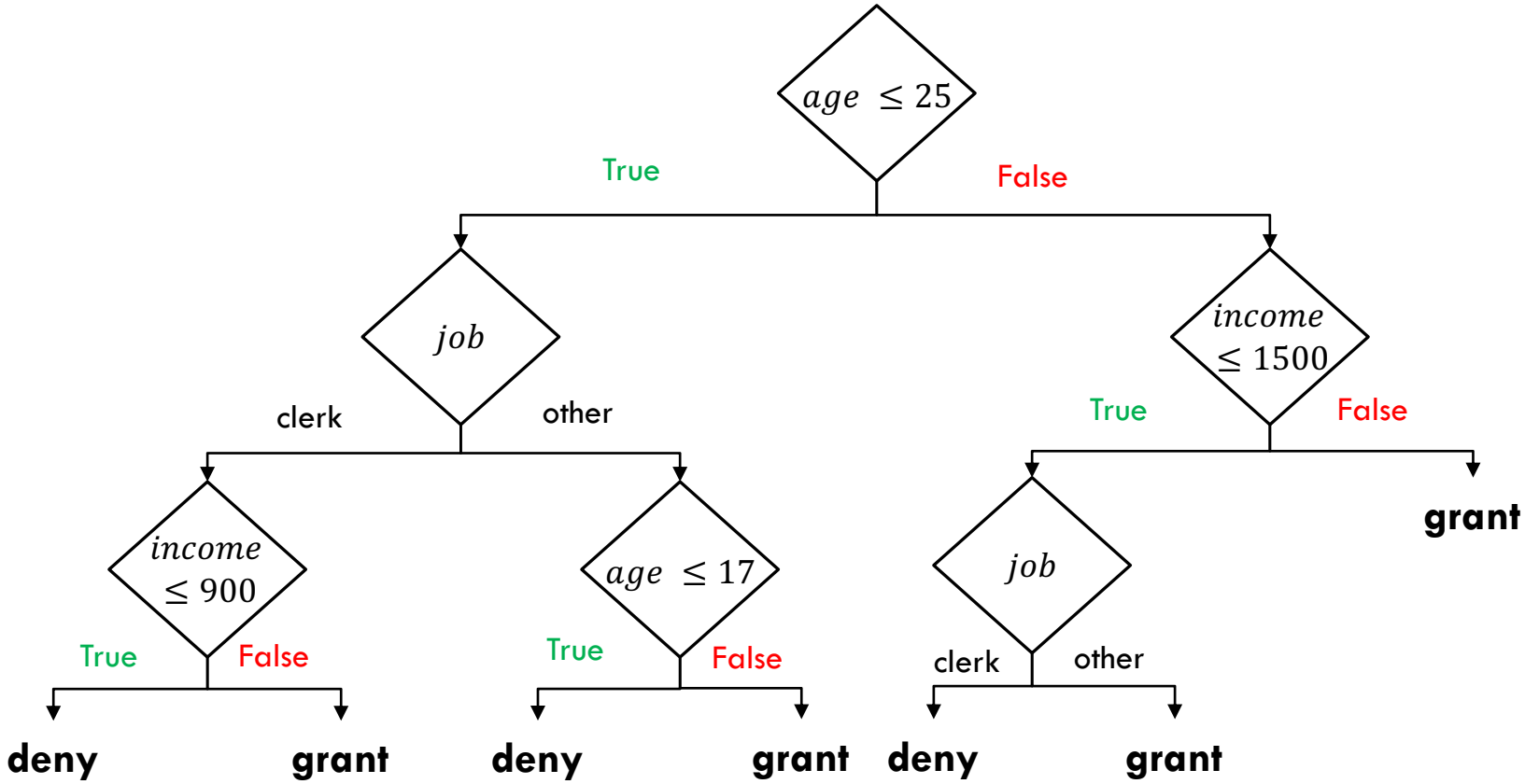


# ALGORITHME GÉNÉTIQUE: ÉVALUATION

# LORE: LE VOISINAGE (Z)

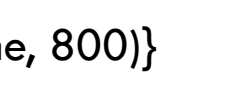


# LORE: MODÈLE INTERPRÉTABLE



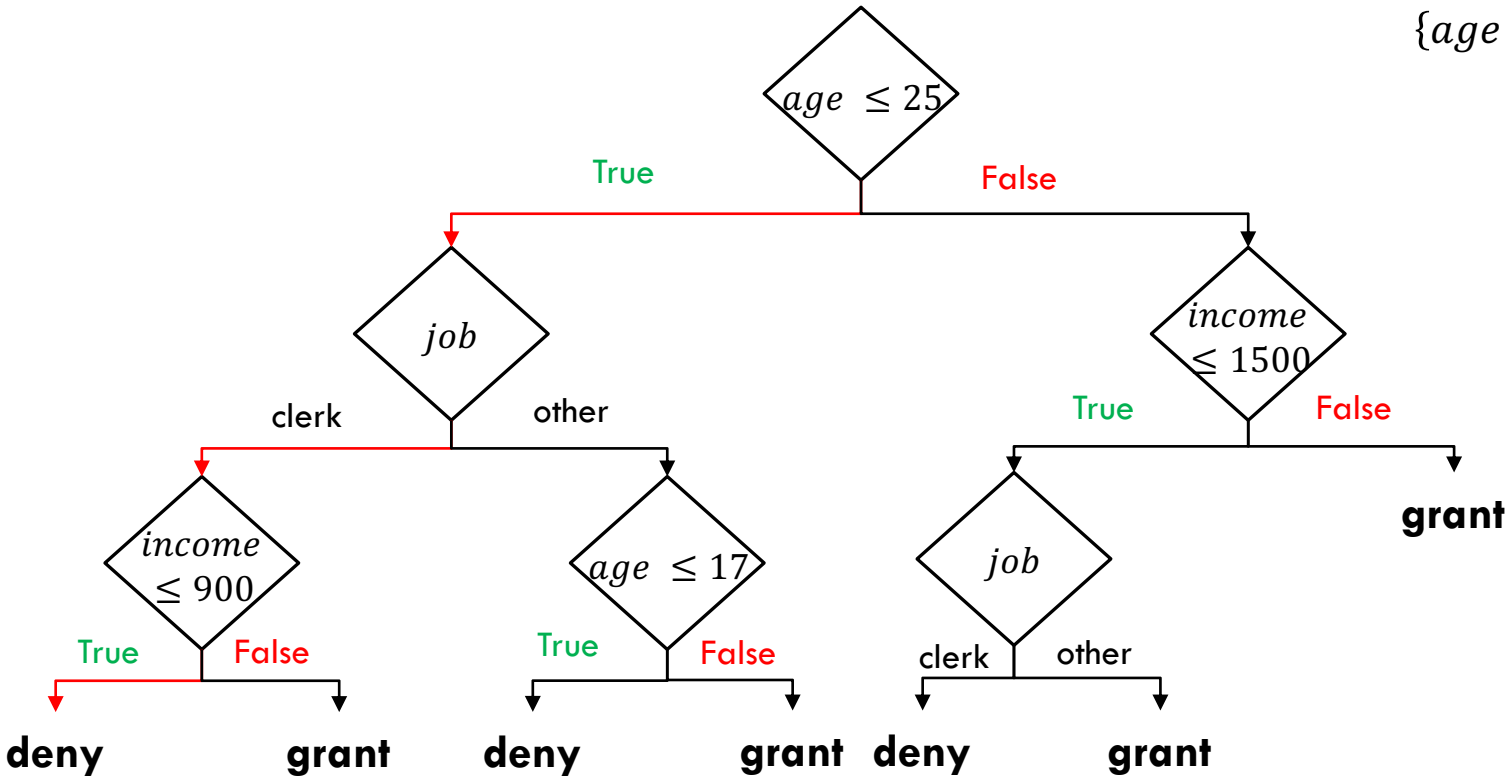
# LORE: EXTRACTION DE LA RÈGLE

x = {(age, 22), (job, clerk), (income, 800)}

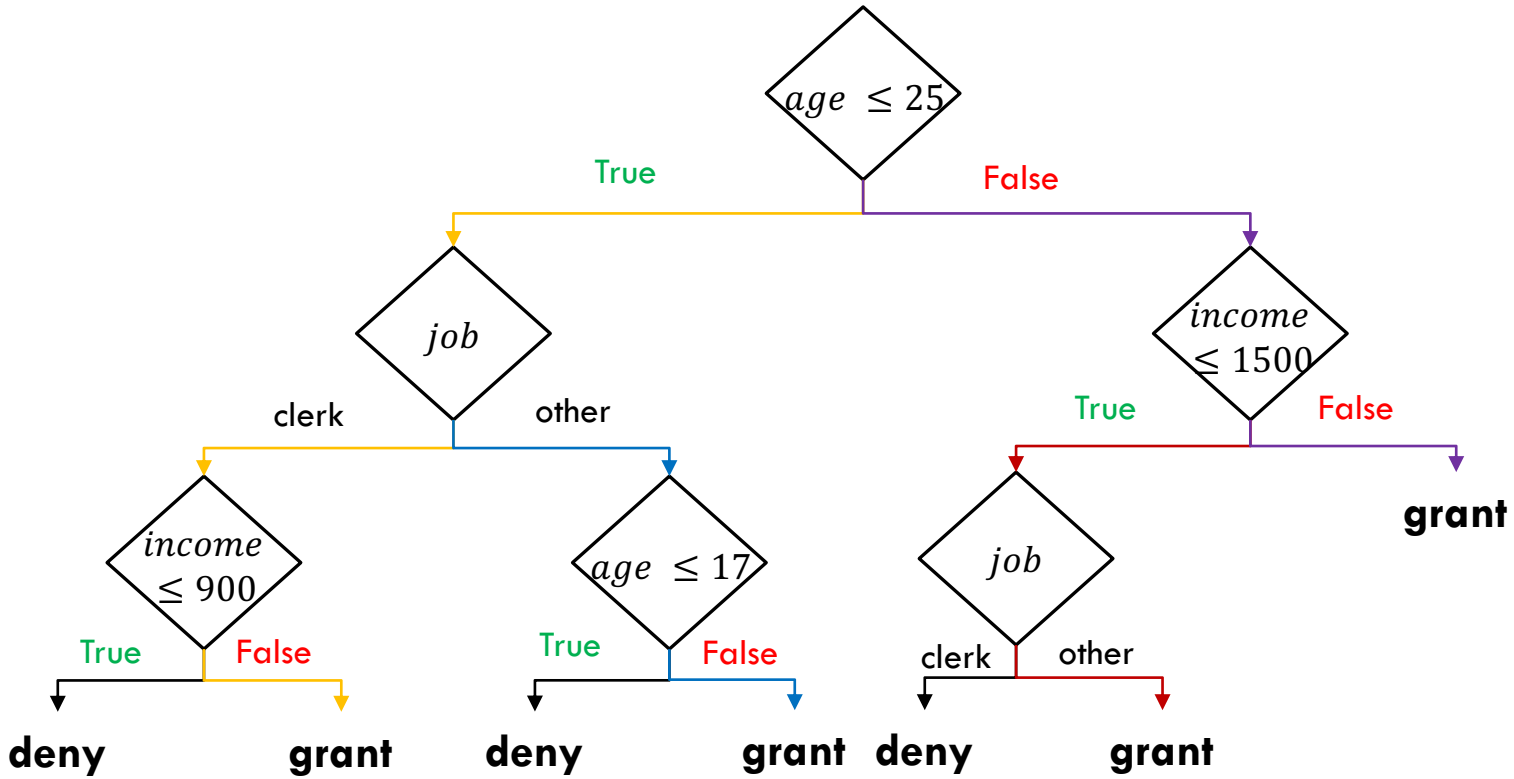


Règle extraite:

$\{age \leq 25, job = clerk, income \leq 900\} \rightarrow deny$



# LORE: EXTRACTION DES RÈGLES CONTREFACTUELLES




$q1 = \{age \leq 25, job = clerk, income > 900\}$

$q2 = \{17 < age \leq 25, job = other\}$

$q3 = \{age > 25, income \leq 1500, job = other\}$

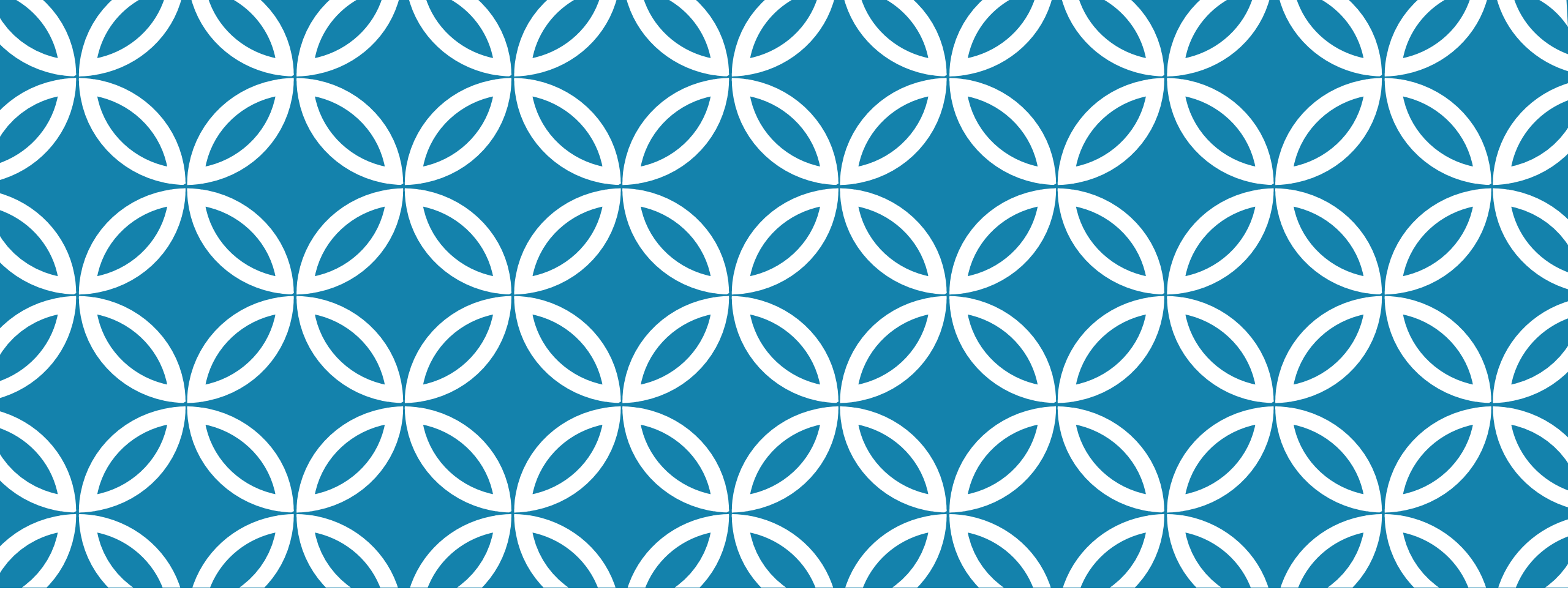
$q4 = \{age > 25, income > 1500\}$

**LORE ne prend que les minimales**

$$\boldsymbol{\varphi = \{q1 \rightarrow grant, q2 \rightarrow grant\}}$$

# COMPARAISON LORE/ANCHORS

| LORE | ANCHORS |
| --- | --- |
| Règles contrefactuelles | Pas de règles contrefactuelles explicites |
| Couverture plus élevée | Couverture plus faible |
| Précision plus faible | Précision plus élevée |
| Plus stable face aux perturbations | Moins stable face aux perturbations |
| Seulement sur des données tabulaires | Possible aussi sur des images et du texte |



# ÉVALUATION DES MÉTHODES

# PRÉSENTATION DES MÉTRIQUES

Completeness $\frac{c}{N}$

C = nombre d'instances couverte
N = nombre d'instances total

Fidelity $\frac{f}{N}$

f = nombre d'instances où les prédictions du modèle et des règles sont identiques

Correctness $\frac{r}{N}$

r = nombre d' instances correctement classifier

Robustness $\frac{\sum_{n=1}^{N} f(x_n) - f(x_n + \delta)}{N}$

δ = la perturbation et
(f(xn)) = prédiction du model

# PRÉSENTATION DES MÉTRIQUES

Number of rules $|R|$

Fraction of classes $\frac{1}{|C|}\sum_{c'<=C} 1(\exists r = (s,c) \in R | c = c')$

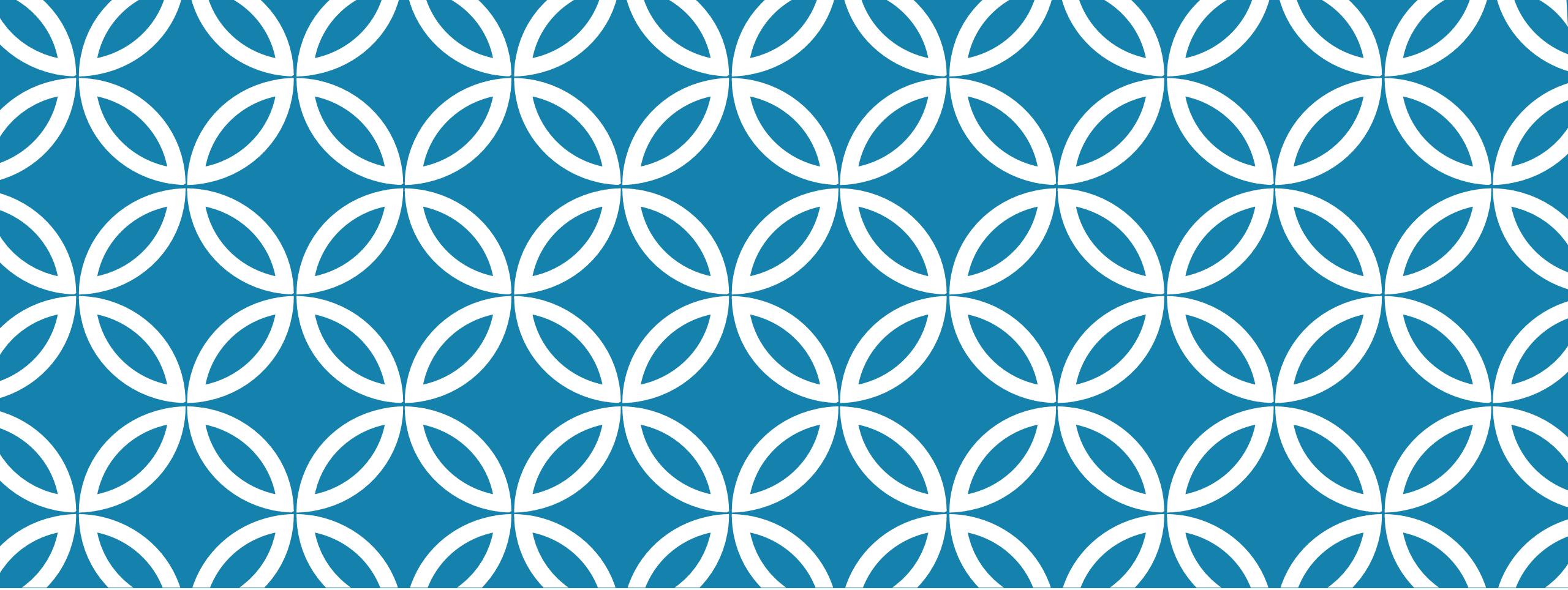
R = liste des règles
|C| = nombre de classe

Average rule length $\frac{\sum_{i=1}^{R} a_i}{|R|}$

ai = nombre d'antécédents de la règle i

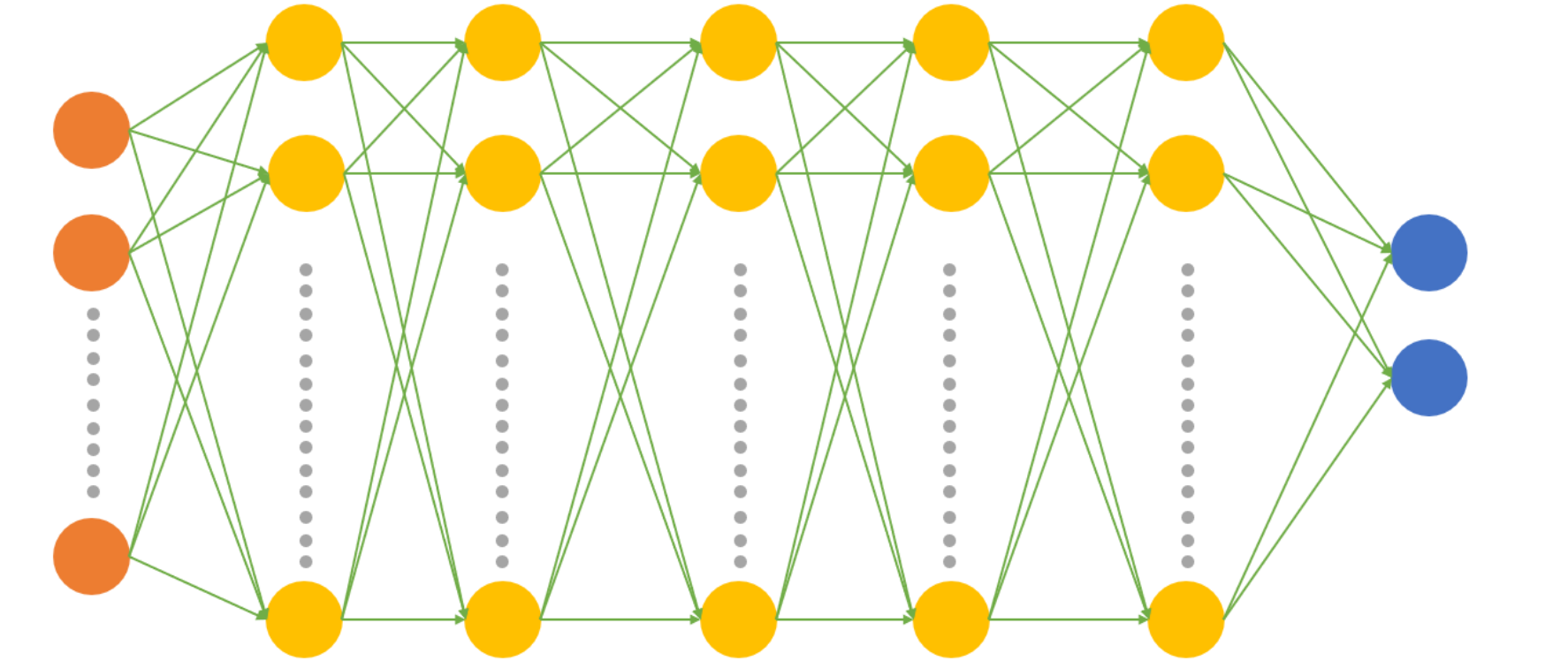
Fraction overlap $\frac{2}{|R|(|R|-1)}\sum_{r_i,r_j,i<=j} \frac{overlap(r_i,r_j)}{N}$

overlap renvoie le nombre d'antécédents sur la même classe et dont les conditions sont remplies sur une instance



# NOTRE EXPÉRIMENTATION

# NOTRE MODÈLE



# ENTRAINEMENT ET RÉSULTATS

Mise en place d'un GridSearch sur les hyperparamètres :

- Batch size
- Dropout
- Learning rate

Entraînement sur 4 jeux de données : Diabètes, Breast-cancer, Heart et Covid-19

Précision déterminée sur l'ensemble de test :

| | **Diabètes** | **Breast-cancer** | **Heart** | **Covid-19** |
|---|---|---|---|---|
| Précision | 78,81 % | 95,65 % | 83,61 % | 74,25 % |

*Précision obtenus par notre modèle*

# COMPARAISON AVEC XGBOOST

- XGBoost est un modèle réputé pour ces performances sur des données tabulaire
- Nous utilisons les mêmes jeux d'entraînement, validation et de test

| | **Diabètes** | **Breast-cancer** | **Heart** | **Covid-19** |
|---|---|---|---|---|
| Notre modèle | 78,81 % | 95,65 % | 83,61 % | 74,25 % |
| XGBoost | 83,12 % | 96,49 % | 80,33 % | 72,80 % |

*Comparaison de la précision entre notre modèle et xgboost*

# PRÉSENTATION DES RÉSULATS : LOCAL

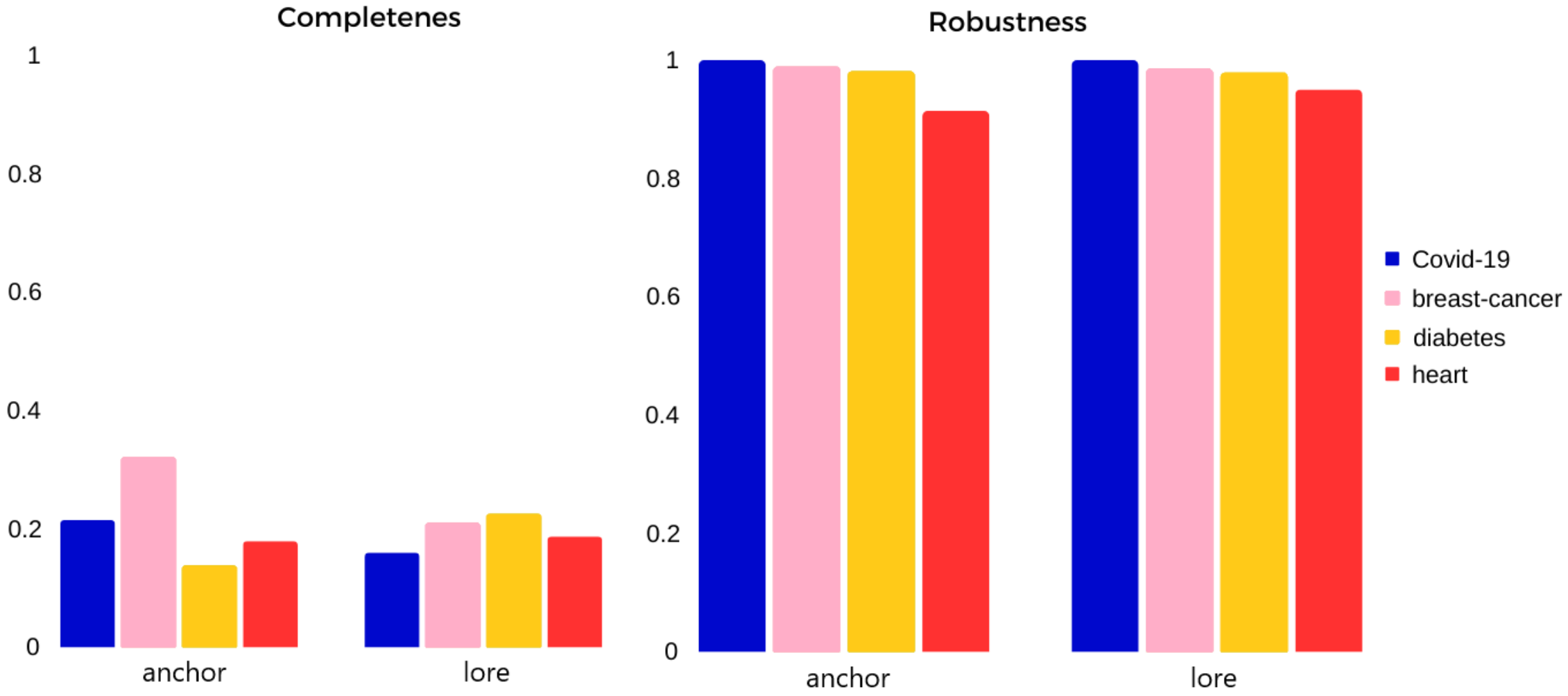
Métrique d'anchor

| | Covid-19 | breast-cancer | diabetes | heart |
|---|---|---|---|---|
| Completeness | 0.2150 | 0.3215 | 0.1382 | 0.1794 |
| Correctness | 0.5 | 0.3796 | 0.3398 | 0.4785 |
| Fidelity | 0.7705 | 0.3567 | 0.3398 | 0.4983 |
| Robustness | 0.9999 | 0.9894 | 0.9817 | 0.9141 |
| Number of rules | 140 | 114 | 154 | 61 |
| Average rule length | 2.8642 | 1.1578 | 2.1623 | 2.1803 |
| Fraction of classes | 0.3809 | 0.3333 | 1 | 0.8461 |
| Fraction overlap | 0.6718 | 0.1867 | 0.1708 | 0.0617 |

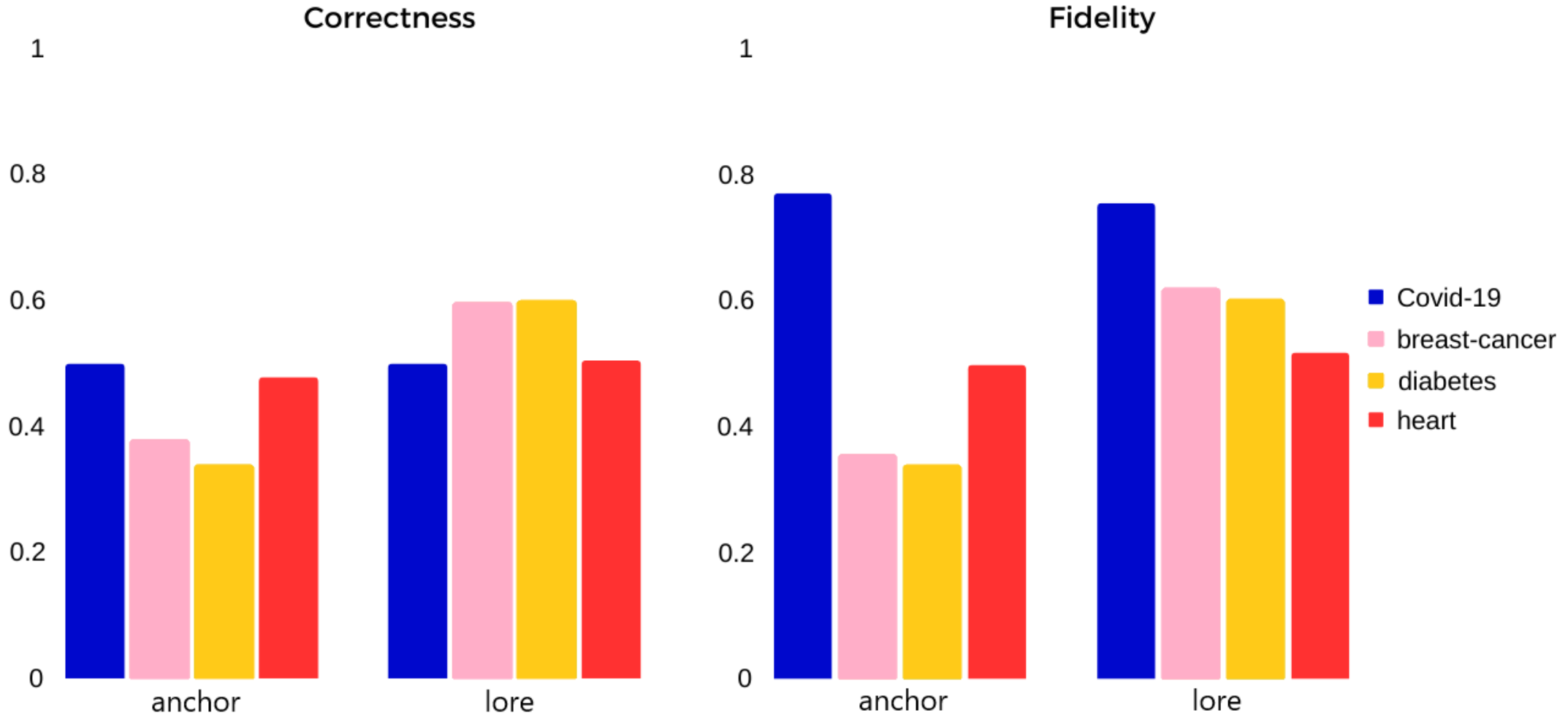
Métrique de lore

| | Covid-19 | breast-cancer | diabetes | heart |
|---|---|---|---|---|
| Completeness | 0.1599 | 0.2105 | 0.2258 | 0.1873 |
| Correctness | 0.4999 | 0.5975 | 0.6002 | 0.5049 |
| Fidelity | 0.7552 | 0.6203 | 0.6028 | 0.5181 |
| Robustness | 0.9999 | 0.9859 | 0.9791 | 0.9504 |
| Number of rules | 140 | 114 | 154 | 61 |
| Average rule length | 3.7214 | 2.6140 | 3.6493 | 3.7704 |
| Fraction of classes | 1 | 0.7333 | 1 | 1 |
| Fraction overlap | 0.3066 | 0.0460 | 0.0786 | 0.0831 |

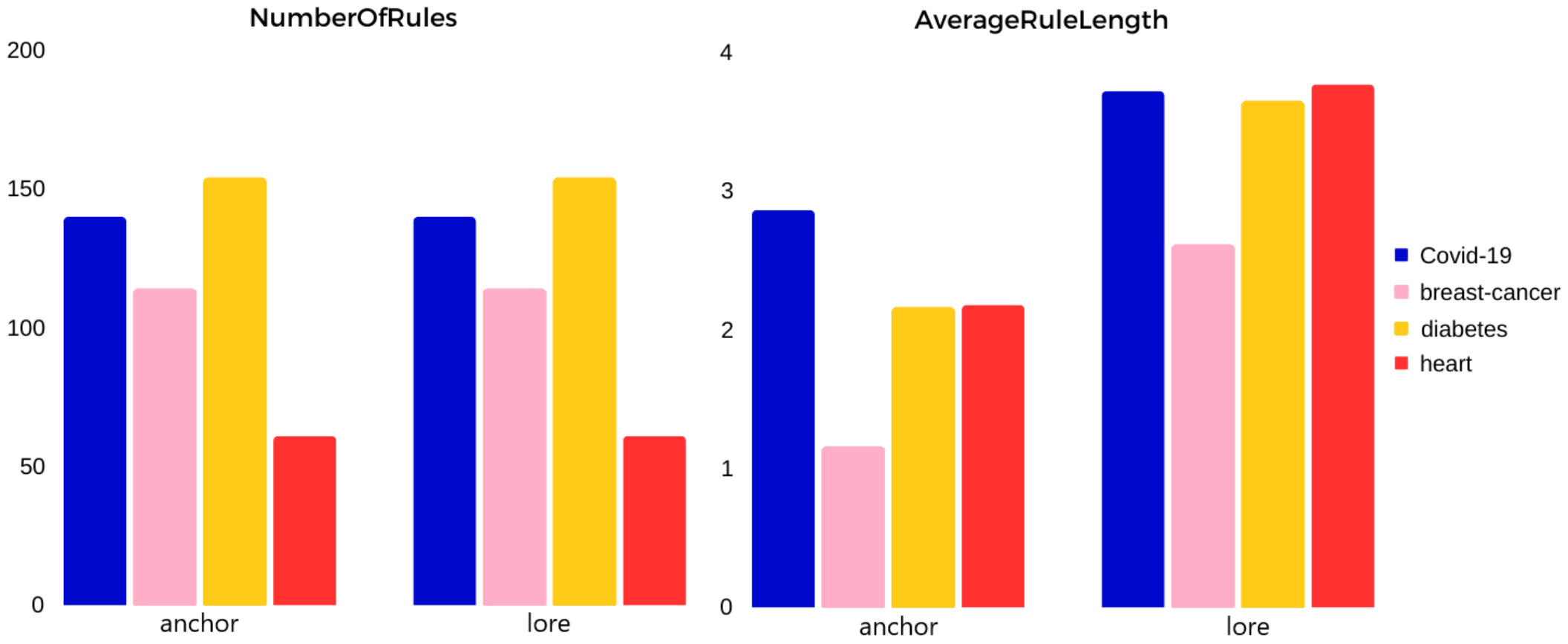
# PRÉSENTATION DES RÉSULATS : LOCAL



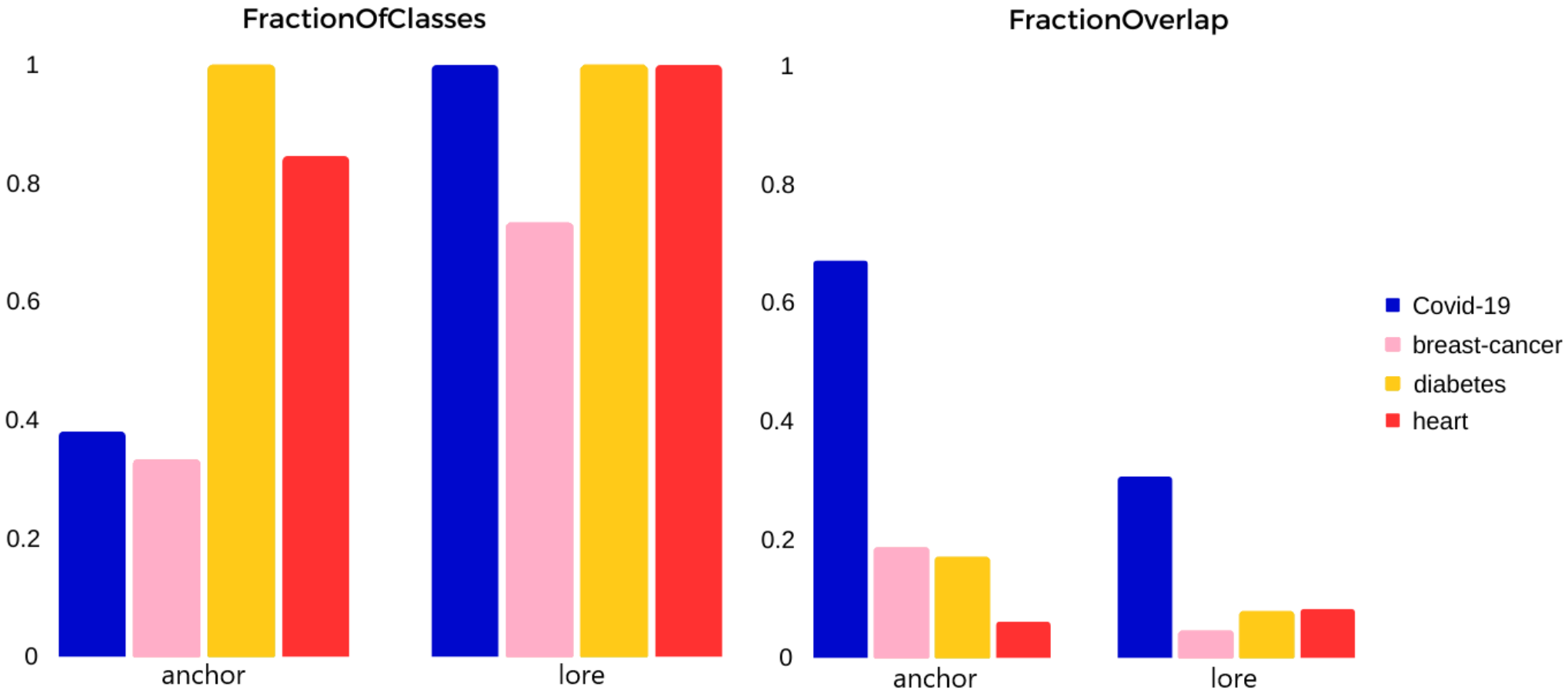
# PRÉSENTATION DES RÉSULATS : LOCAL



# PRÉSENTATION DES RÉSULATS : LOCAL



# PRÉSENTATION DES RÉSULATS : LOCAL



# PRÉSENTATION DES RÉSULATS : GLOBAL

GlocalX-anchor

| | Covid-19 | breast-cancer | diabetes | heart |
|---|---|---|---|---|
| Completeness | 1 | 1 | 0.8932 | 1 |
| Correctness | 0.4999 | 0.6133 | 0.4960 | 0.4587 |
| Fidelity | 0.2293 | 0.6362 | 0.5221 | 0.4389 |
| Robustness | 0.9999 | 0.9894 | 0.8463 | 0.9702 |
| Number of rules | 9 | 8 | 23 | 6 |
| Average rule length | 2.5555 | 2 | 3.8260 | 2.1666 |
| Fraction of classes | 0.3333 | 0.2666 | 1 | 0.5384 |
| Fraction overlap | 0.6382 | 0.0439 | 0.0061 | 0.1355 |

GlocalX-lore

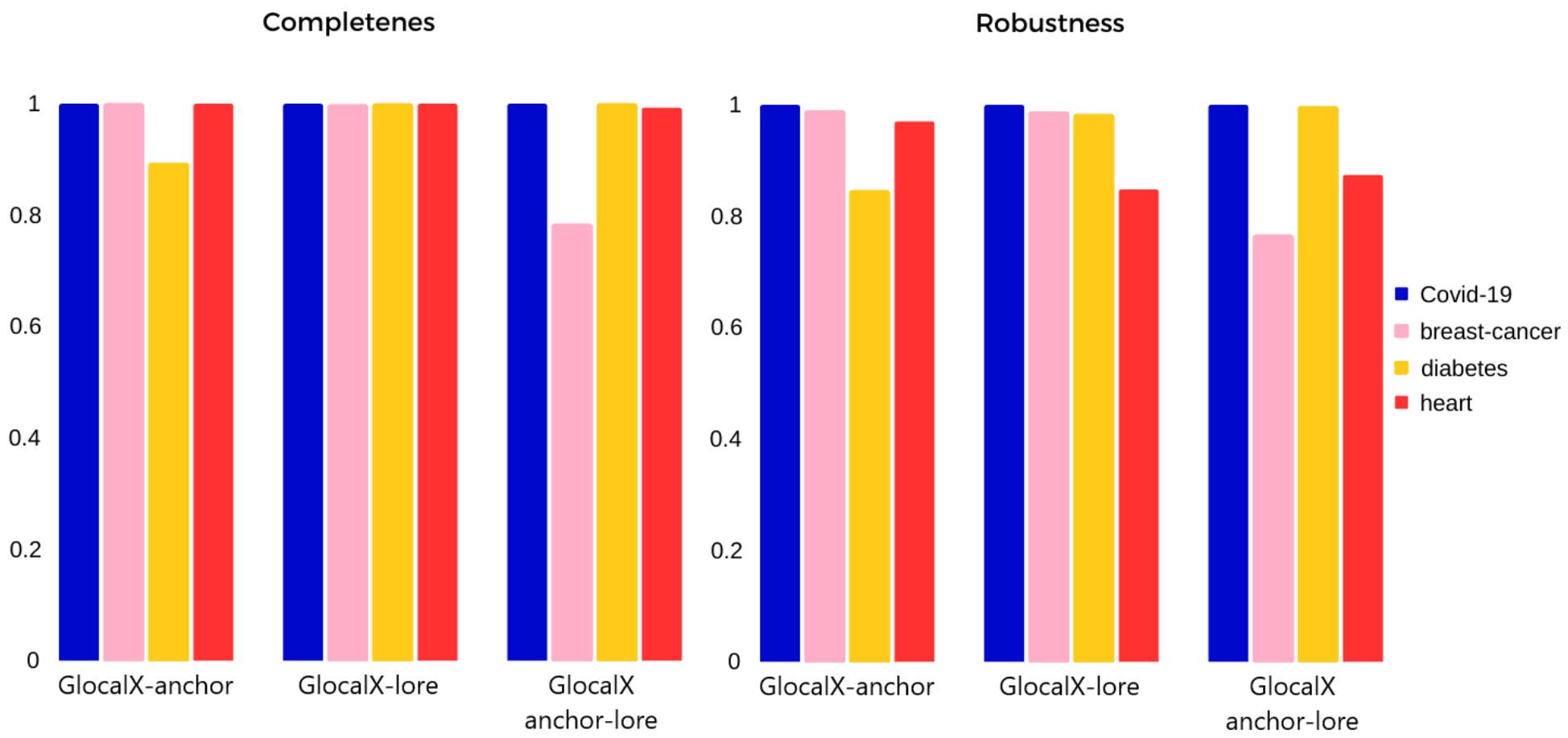
| | Covid-19 | breast-cancer | diabetes | heart |
|---|---|---|---|---|
| Completeness | 1 | 0.9982 | 1 | 1 |
| Correctness | 0.4999 | 0.6115 | 0.6145 | 0.4488 |
| Fidelity | 0.2293 | 0.6344 | 0.6276 | 0.4290 |
| Robustness | 0.9999 | 09876 | 0.9830 | 0.8481 |
| Number of rules | 29 | 14 | 26 | 10 |
| Average rule length | 5 | 3.6428 | 4.0769 | 3.2 |
| Fraction of classes | 0.9523 | 0.3 | 1 | 0.7692 |
| Fraction overlap | 0.1646 | 0.0246 | 0.0121 | 0.0885 |

# PRÉSENTATION DES RÉSULATS : GLOBAL

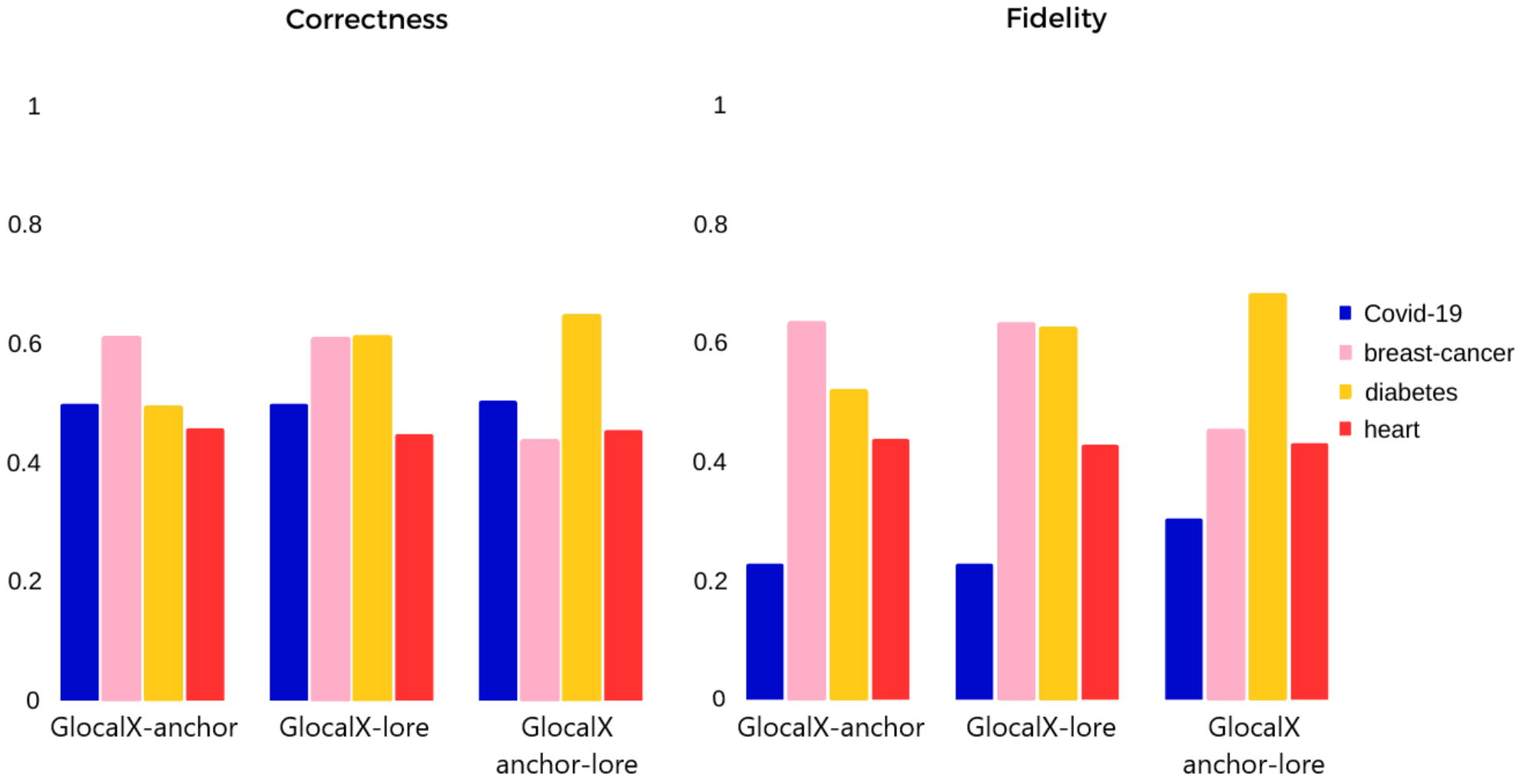
GlocalX-anchor-lore

| | Covid-19 | breast-cancer | diabetes | heart |
|---|---|---|---|---|
| Completeness | 1 | 0.7838 | 1 | 0.9933 |
| Correctness | 0.5046 | 0.4393 | 0.6497 | 0.4554 |
| Fidelity | 0.3055 | 0.4551 | 0.6835 | 0.4323 |
| Robustness | 0.9999 | 0.7662 | 0.9973 | 0.8745 |
| Number of rules | 66 | 16 | 40 | 57 |
| Average rule length | 4.4696 | 4.1875 | 4.3750 | 3.3157 |
| Fraction of classes | 1 | 0.5 | 1 | 0.8461 |
| Fraction overlap | 0.0423 | 0.0114 | 0.0067 | 0.0110 |

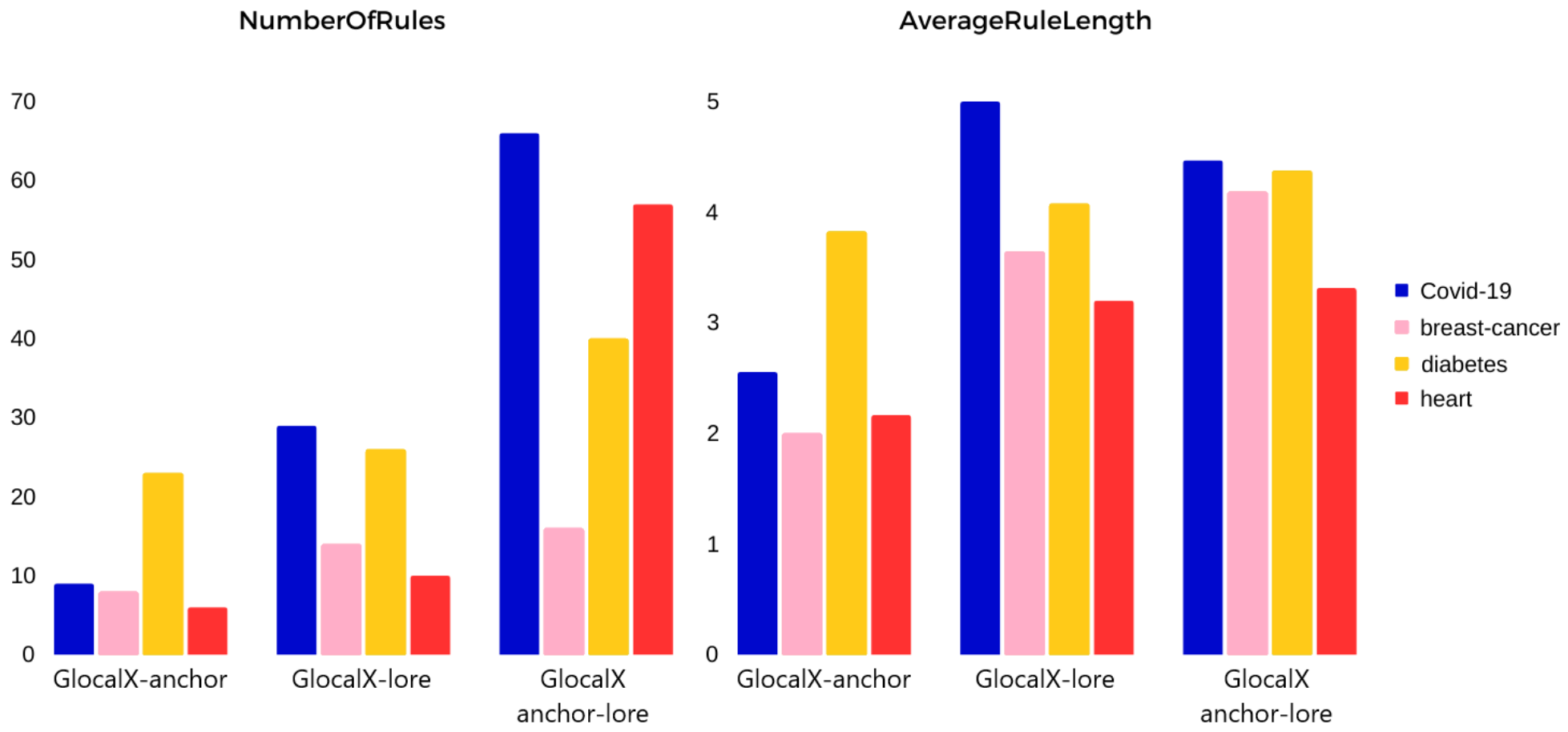
# PRÉSENTATION DES RÉSULATS : GLOBAL



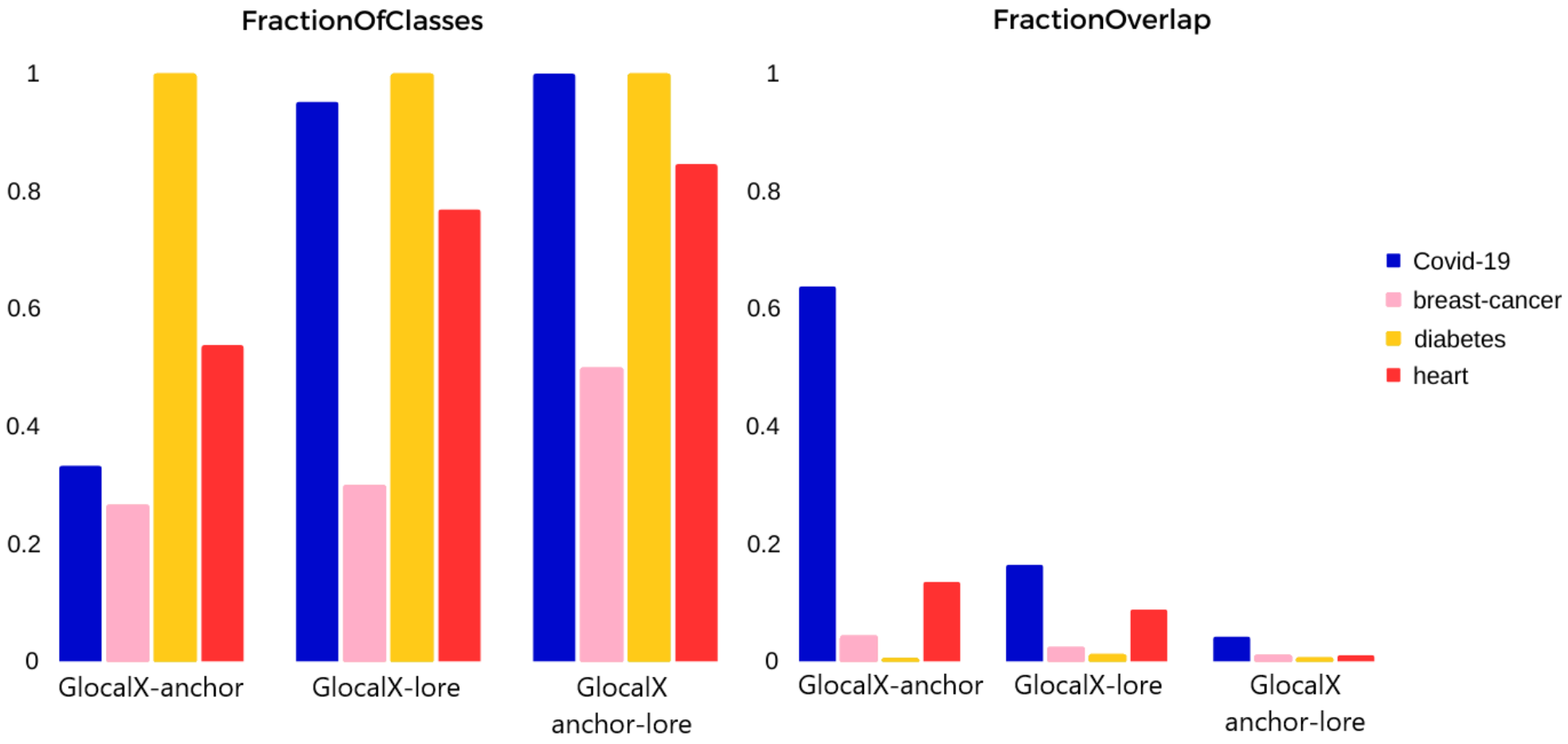
# PRÉSENTATION DES RÉSULATS : GLOBAL

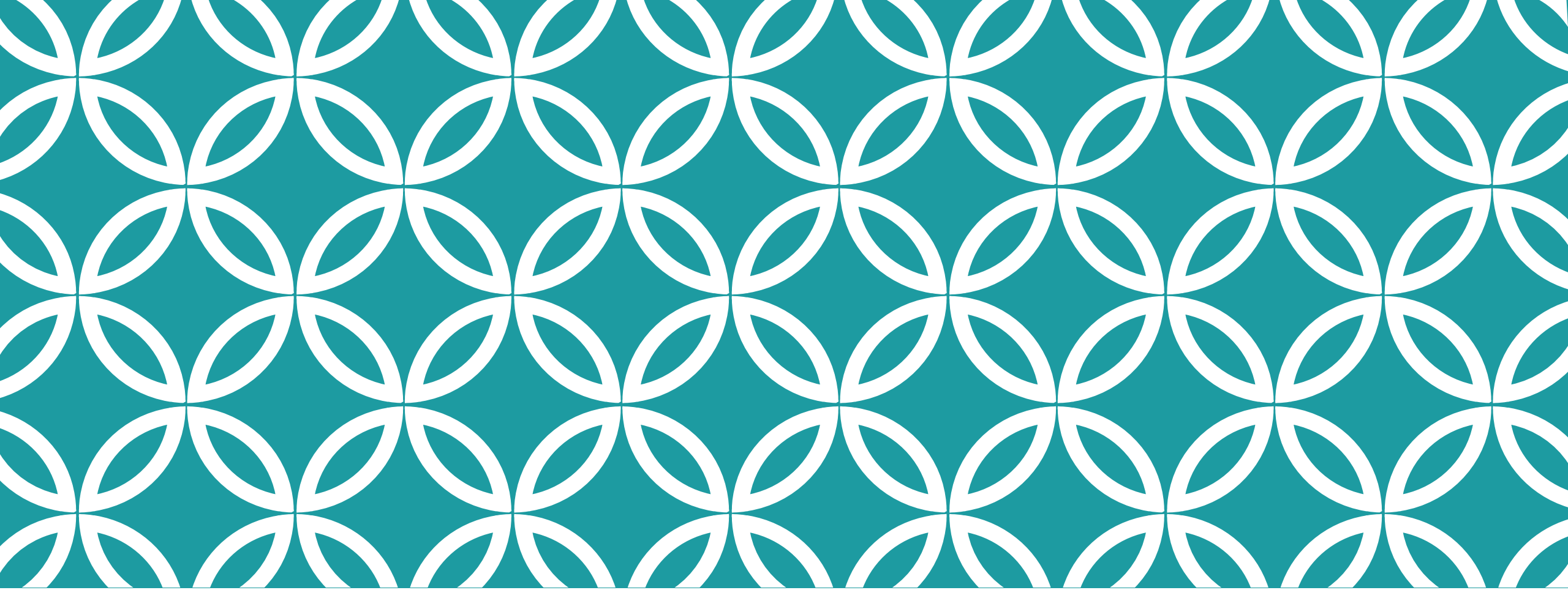


# PRÉSENTATION DES RÉSULATS : GLOBAL



# PRÉSENTATION DES RÉSULATS : GLOBAL





# CONCLUSION

# RÉFÉRENCES

[1] Riccardo Guidotti et al. « Local Rule-Based Explanations of Black Box Decision Systems ». In : (2018). arXiv : 1805.10820 [cs.AI]. url : https://arxiv.org/abs/1805.10820.

[2] Marco Tulio Ribeiro, Sameer Singh et Carlos Guestrin. « Anchors: High-Precision Model-Agnostic Explanations ». In : (2018).

[3] Mattia Setzu et al. « GLocalX - From Local to Global Explanations of Black Box AI Models ». In : Artificial Intelligence 294 (2021), p. 103457. issn : 0004-3702. doi : https://doi.org/10.1016/j.artint. 2021.103457. url : https://www.sciencedirect.com/science/article/pii/S0004370221000084.

[4] Giulia Vilone et Luca Longo. « A Quantitative Evaluation of Global, Rule-Based Explanations of Post-Hoc, Model Agnostic Methods ». In : Frontiers in Artificial Intelligence 4 (2021). issn : 2624-8212. doi : 10.3389/frai.2021.717899. url : https://www.frontiersin.org/articles/10.3389/frai. 2021.717899.