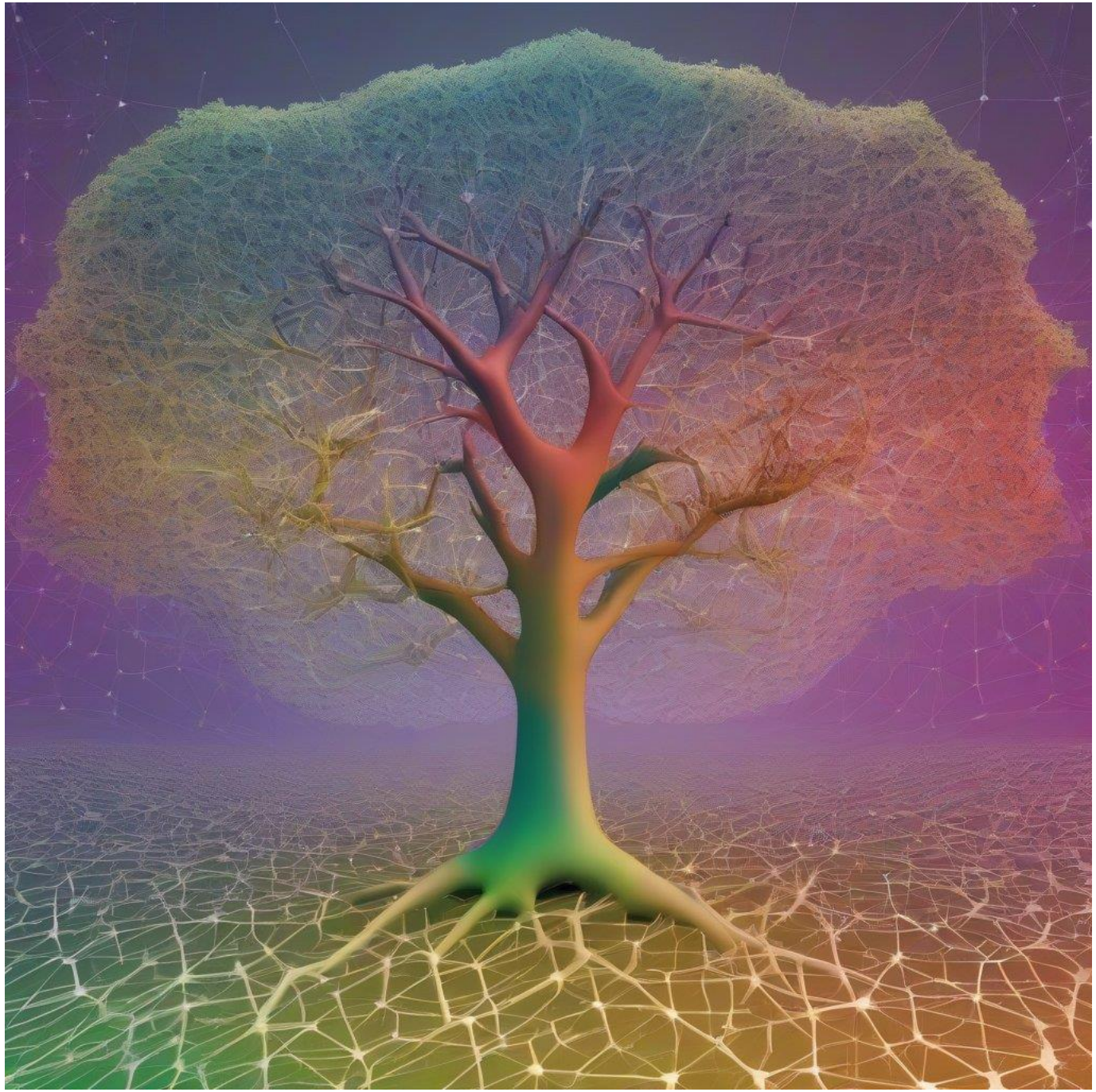
AI VIDENCE

Expliquer les IA : de l'IA statistique à l'extraction de connaissance via une approche « régionale »

David Cortés

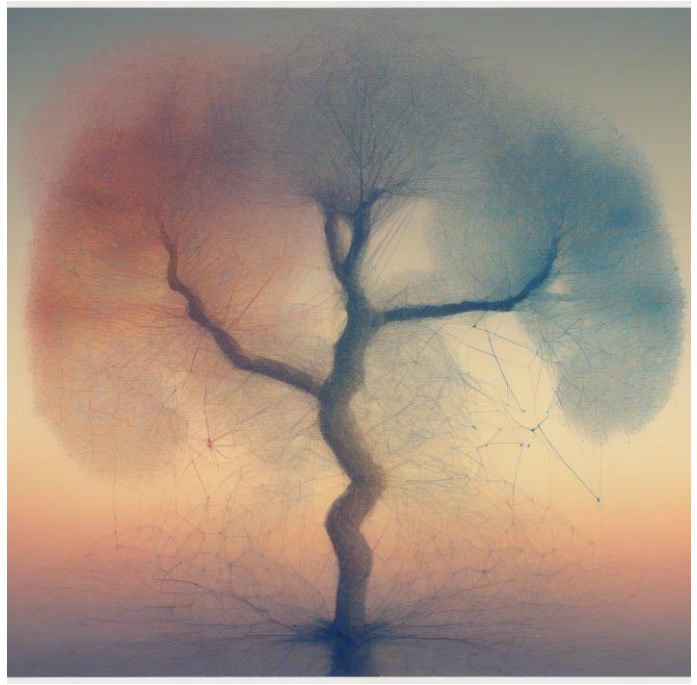


Arbre de la connaissance hybride
selon Stable Diffusion

IAs, réconcilier les sœurs ennemies ?

1956

Conférence de Darmouth
Les 2 IAs…



IA logique : raisonnement, « compréhensible par nature »

IA statistique : machine à prédire, « singe savant », mais « meilleure en performance »

Des freins communs aux 2 IAs

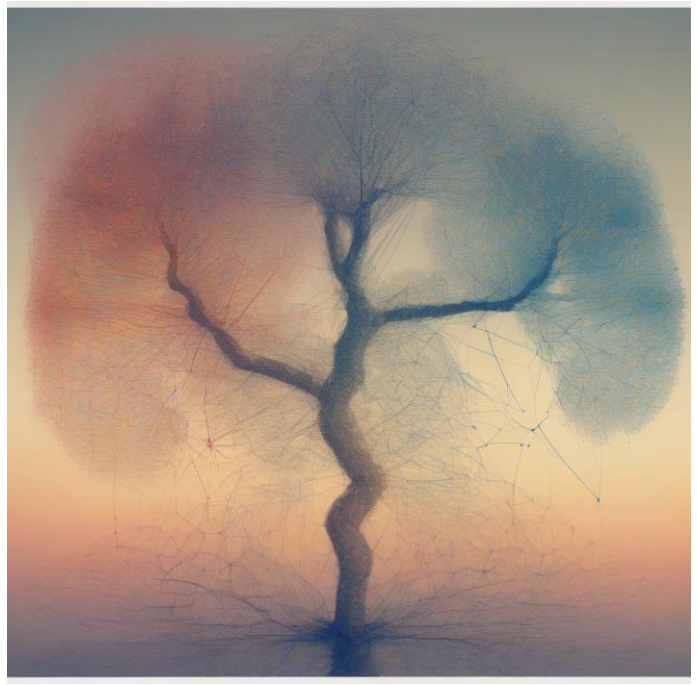
1956

Conférence de Darmouth
+

Loi de Miller
La règle des « 7 +/- 2… »

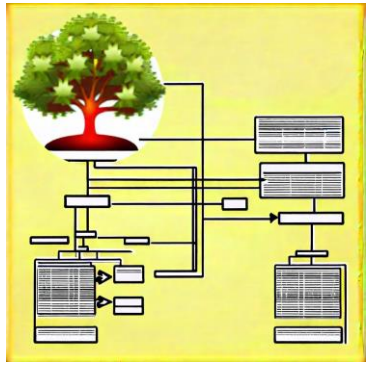
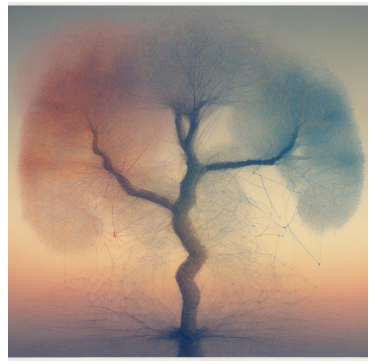
« **L'obsolescence de l'homme** »
Gunther Anders

1958 …. '**The Human Condition**'
Hannah Arendt



Complexification et automatisation → Défiance, manque de contrôle

IAs, un ennemi commun : la complexité

Complexité du *phénomène*	Simple	Moyenne	Élevée
			
Performance			
 Logique	+++	++	+
Statistique	+++	+++	+++ 
Compréhension			
 Logique	+++	++	+
Statistique	+	- -	- - - 

IA statistique : exacerbe l'automatisation

L' « apprentissage » est automatique

→ Fait mieux en performance

… même si l'on ne sait pas pourquoi !

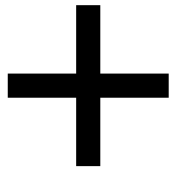
→ Les performances se dégradent dans le temps (dès la mise en production…)

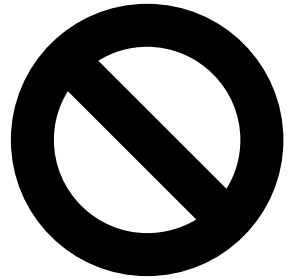
→ On ne sait pas ce qu'elle a « appris » !

Comportement aux limites ? Responsabilité ?

Vos équipes n’ont plus confiance en l’IA ?

AI VIDENCE





90%
des projet IA
ne vont pas en prod

50%
des modèles IA en prod
sont débranchés

→ l'IA de confiance représente 60% des investissements IA

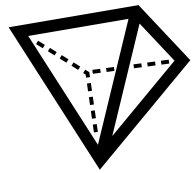
AI VIDENCE

Vous avez rencontré trop de problèmes avec l'IA ?

Opaque

Pas d'adoption

Modèles sont **discriminants** ou **peu éthiques**



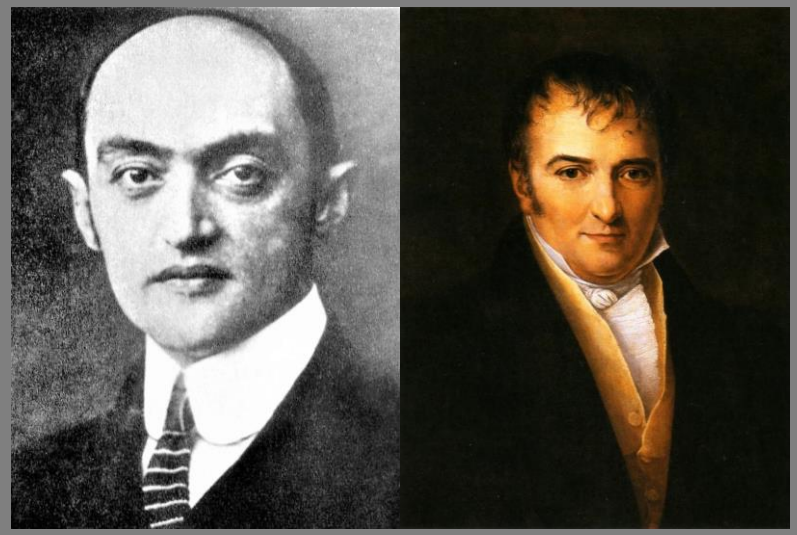
Robustesse? Quid aux limites ?

Cf hallucinations des LLMs

Impératifs d’adoption et de mise en conformité : créer la CONFIANCE

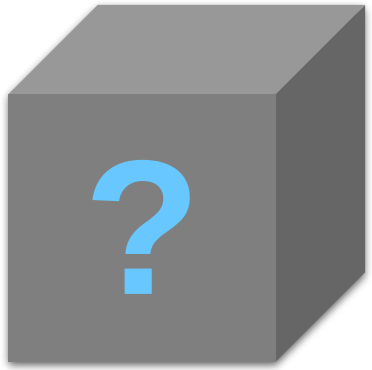
- **Éthique**
  (« neutralité technologique »)
- **Robustesse**

- **Adoption opérationnelle**




AI VIDENCE
rend vos IA explicables

Expliquer ? C’est s’adapter à chacun en fonction de ses enjeux et de sa culture



Modèle opaque (« boîte noire »), entraîné sur des données

Pour qui ?

- Le régulateur, les auditeurs
- Les équipes conformité
- Les utilisateurs du modèle
- Les clients
- Le data-scientist
- Les experts métier

Pour quoi ?

- **Audits**
- **AI Act** et régulations sectorielles
- **Adoption**
- **Acceptation**
- **Amélioration** du *modèle*
- **Augmentation** de l’expertise (ex: segmentation, élasticité)

AI VIDENCE

Un grand pouvoir ... Pour le concepteur du « modèle »

Le Data Scientist :

- Ne comprend pas le fonctionnement fin de son modèle
- Ne maitrise pas les limites du modèle

→ il ne peut pas

- rendre compte,
- donner des garanties aux parties prenantes.

Comment ? établir des échanges plus productifs entre utilisateur et concepteur du modèle

Data Scientist +

Client	Explication locale	Information / Justification **contextualisée**
Collaborateur	Explication « **régionale** » **+** globale	Interprétation / compréhension
Régulateur Superviseur	Preuve	**+ de** Garanties de fonctionnement

Expliquer à une échelle humaine

AI VIDENCE

Une réponse ?

1637

Discours de la Méthode
René Descartes

4 étapes :
- Exploration « évidente »
- Découpe en parcelles
- Reformulation et substitution
- Exhaustivité

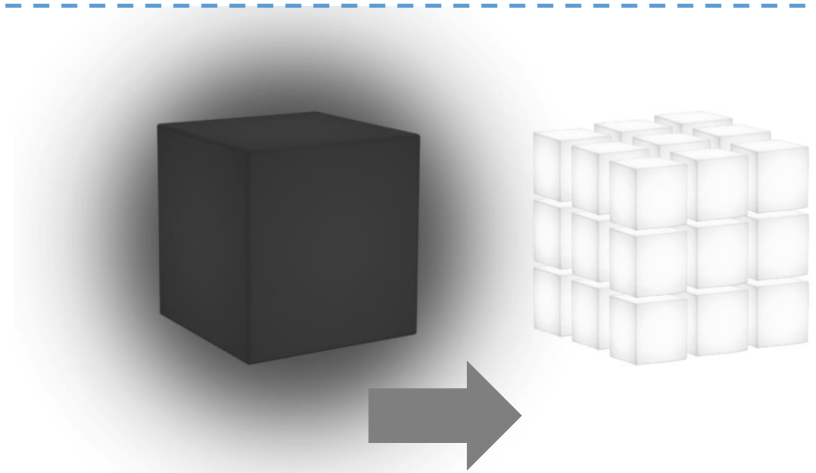
« René Descartes devant un arbre de connaissance hybride intelligence artificielle en mosaïque »

Notre proposition pour : améliorer l'interprétabilité et apporter des garanties

AntakIA

A New Tool to Acquire Knowledge from AI

Explorer une IA → **Echanger, documenter** → **Substituer**



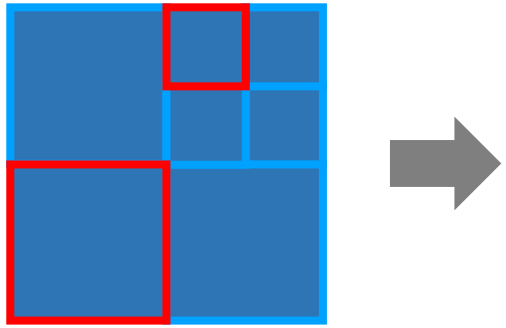
AntakIA en détail… explicabilité « régionale » :

explorer et substituer collaborativement à une échelle « parlante »

1

Explication « régionale »

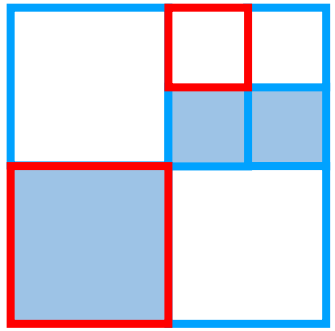
observations ET
explications semblables



2

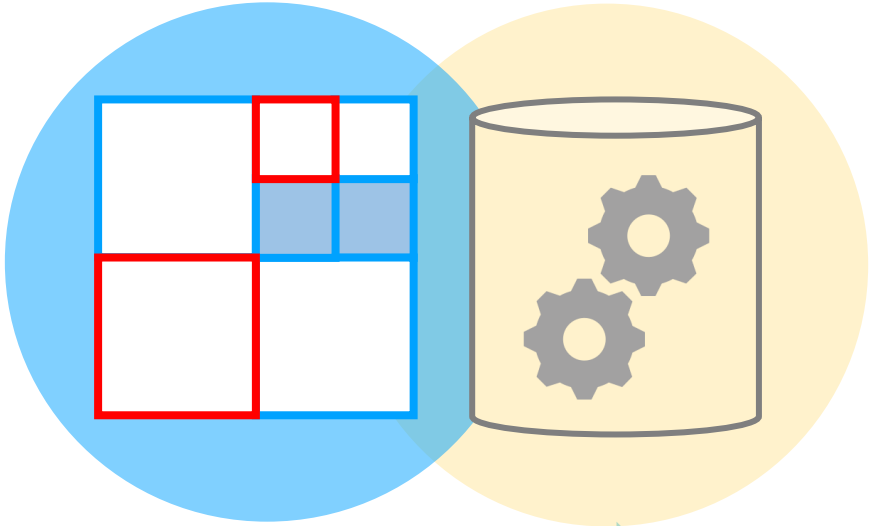
Substitution

par des modèles
explicables ou causaux



3

Hybridation du modèle…



Cas d'usage à **« haut risque »**

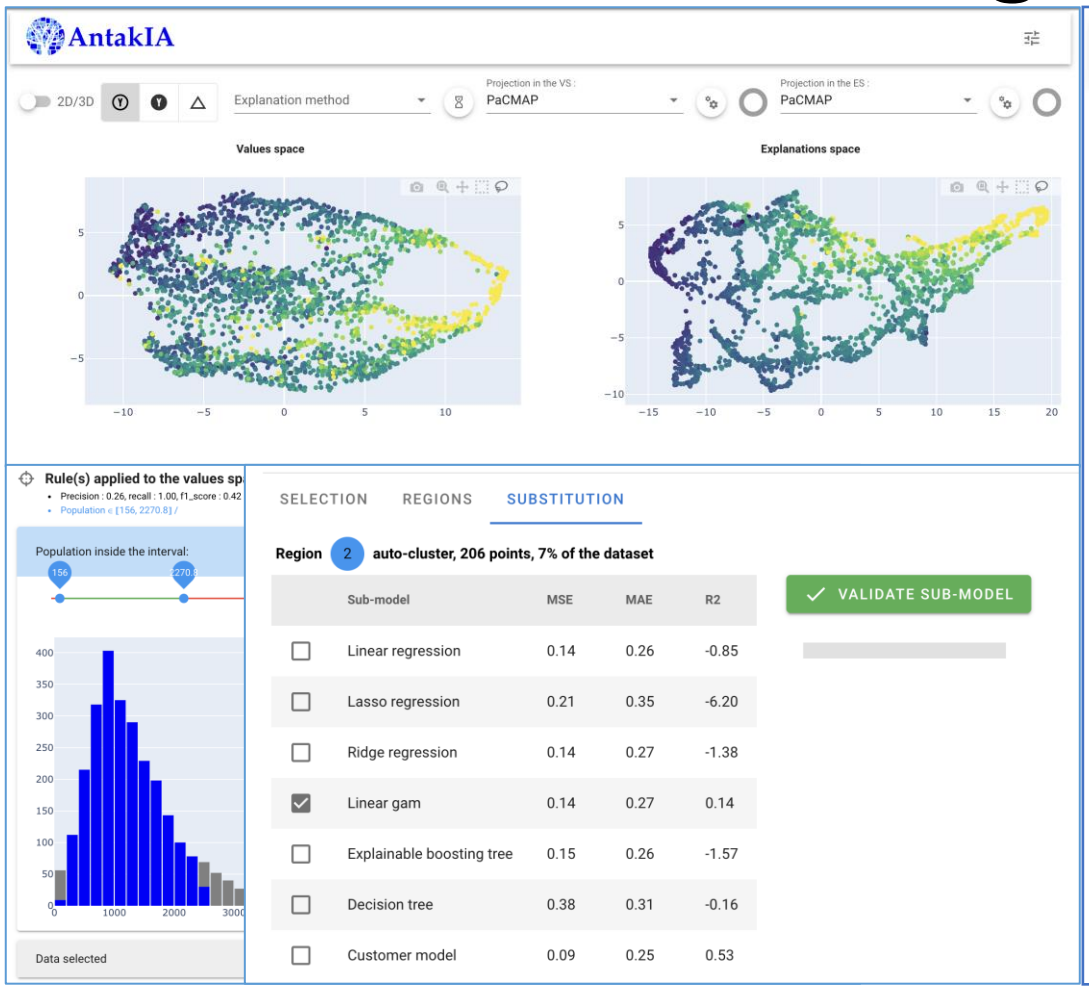
4

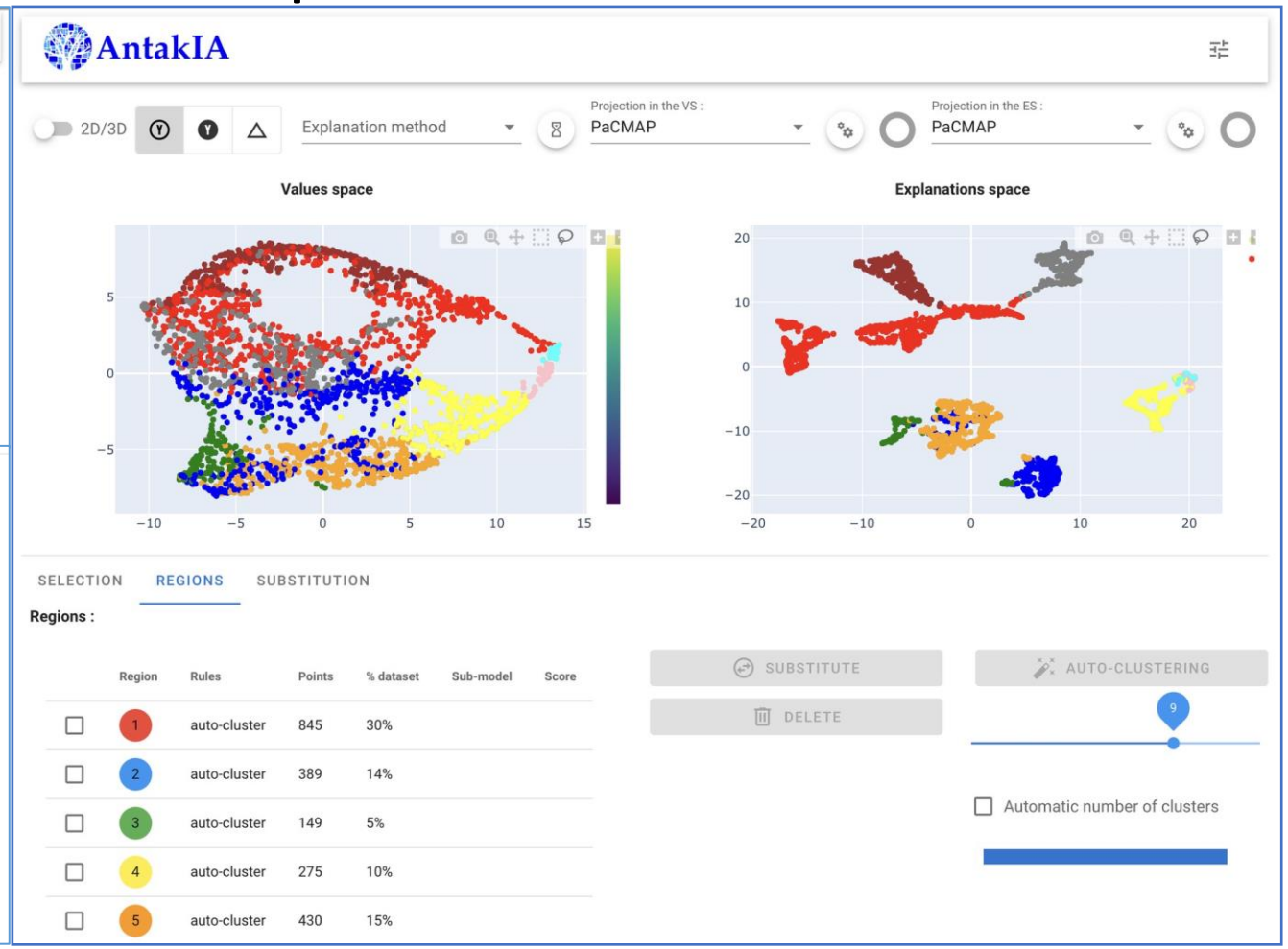
Collaboration avec le métier

5

Suivi dans le temps (dérive)

Comment ? Exploration « Dyadique » : construction de régions explicables





En bref : une IA Cartésienne ++…

Qui utilise l'IA (statistique) pour **remplacer** tout ou partie de l'IA initiale **par une IA hybride**

- la plus logique, partout où c'est possible, et **sur les zones à risque**

- la plus performante et compréhensible ailleurs

Dès la mise en production :

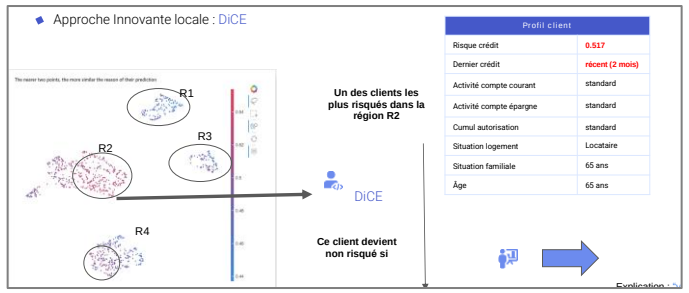
- Suivi dans le temps
- Correction des biais de sélection

Source : « Descartes » selon Stable Diffusion, 2023

AntakIA : audit, modèle explicable par construction…

Éthique

Octroi de crédits

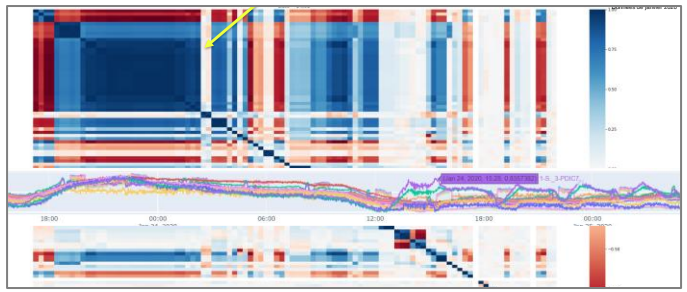


- **Audit externe** de boîtes noires
- **Modèle agnostique**

→ Approche régionale plébiscitée
→ Une palette d'outils de ML pour illustrer et contextualiser

Robustesse

Détection d'anomalies

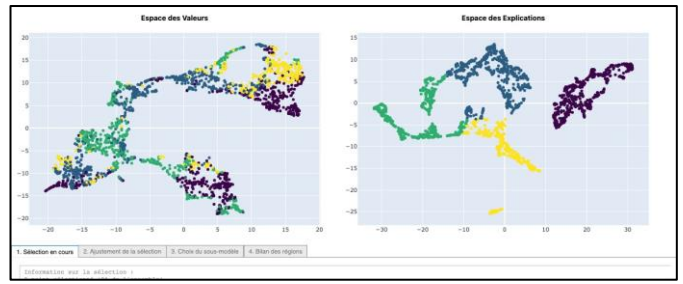


- **Réduction de faux positifs**
- **Superviser la complexité** (env. 3 000 séries temporelles)

→ Approche multi-échelle pour contextualisation
→ Substitution via indicateurs agrégés compréhensibles

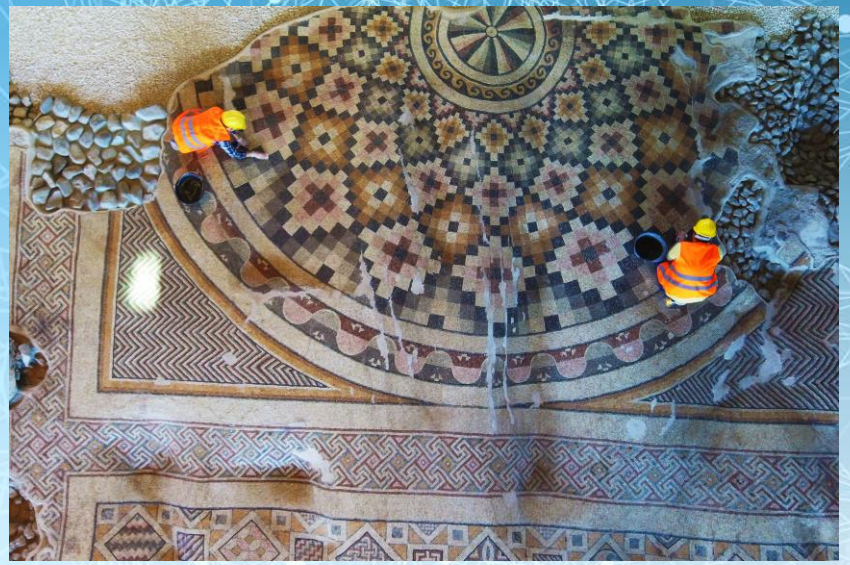
Adoption

Tout cas d'usage



- **Exploration** d'un cas nouveau
- **Substitution d'un modèle existant**

→ Modélisation Causale
→ Gain en performance et maintenabilité (dérive dans le temps)



*Small Is Beautiful**

(*) A Study Of ~~Economics~~ *Machine Learning* As If People Mattered

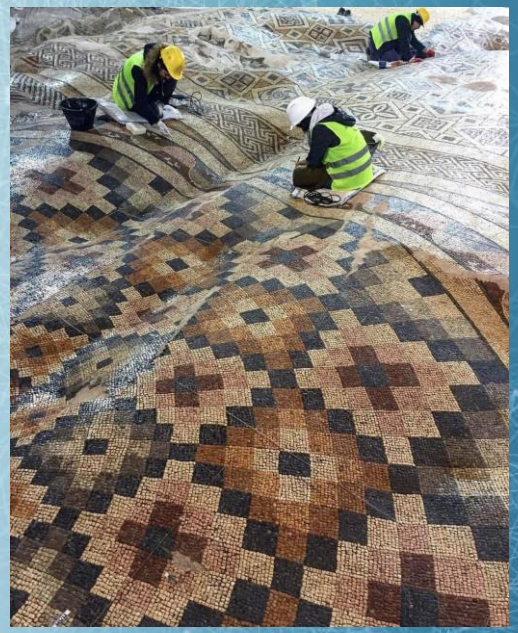
Région ? « parcelle » doublement homogène et « tesselle » de substitution

- **Nous définissons 2 espaces :**
  - **Espace des Valeurs (EV)**: les « observations »
  - **Espace des Explications (EE)**: l'importance des prédicteurs pour le modèle

- **Une parcelle est un segment doublement cohérent :**
  - *Observations similaires* (EV) **ET**
  - *Explications similaires* (EE)

- **Une tesselle est un modèle simple et explicable par parcelle**

→ Le modèle d'IA resultant est une mosaïque explicable, et plus aisément certifiable par construction



Méthodologie – Discours de la méthode, 1637

- **Le premier** était de ne recevoir jamais aucune chose pour vraie que je ne la connusse évidemment être telle; c'est-à-dire, d'éviter soigneusement la précipitation et la prévention, et de ne comprendre rien de plus en mes jugements que ce qui se présenterait si clairement et si distinctement à mon esprit, que je n'eusse aucune occasion de le mettre en doute.
- **Le second**, de diviser chacune des difficultés que j'examinerais, en autant de parcelles qu'il se pourrait, et qu'il serait requis pour les mieux résoudre.
- **Le troisième**, de conduire par ordre mes pensées, en commençant par les objets les plus simples et les plus aisés à connaître, pour monter peu à peu comme par degrés jusqu'à la connaissance des plus composés, et supposant même de l'ordre entre ceux qui ne se précèdent point naturellement les uns les autres.
- Et **le dernier**, de faire partout des dénombrements si entiers et des revues si générales, que je fusse assuré de ne rien omettre.

Visualisation dyadique (dy.) (= valeurs et explications)

Segmentation (dy.)

Exploration dy. et **Modèle le plus simple explicable** par segment

Analyse de **densité** et **d'exhaustivité**

Approche AntakIA

Illustration

Illustration sur données simulées

Visualisation dyadique (dy.) (= valeurs et explications)

Segmentation (dy.)

Exploration dy. et Modèle le plus simple explicable par segment

Analyse de **densité** et **d'exhaustivité**

AntakIA

Retrouver « à l'aveugle » les parcelles d'un modèle segmenté

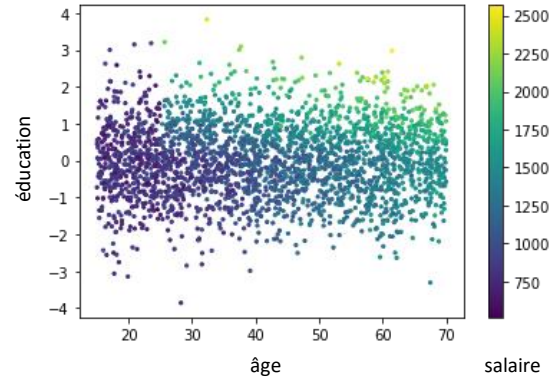
- **Modèle de prévision de salaire avec âge, éducation, genre**
  - Aléatoire avant 25 ans
  - Croissant avec l'âge
  - Bonus Homme après 40 ans

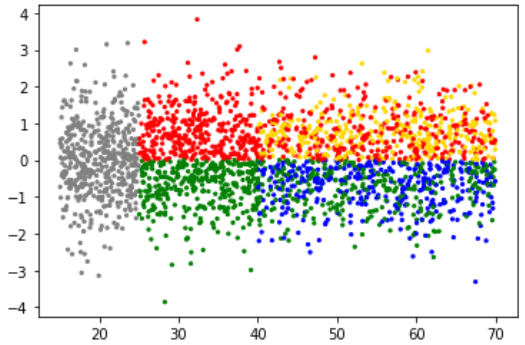
→ 5 segments !

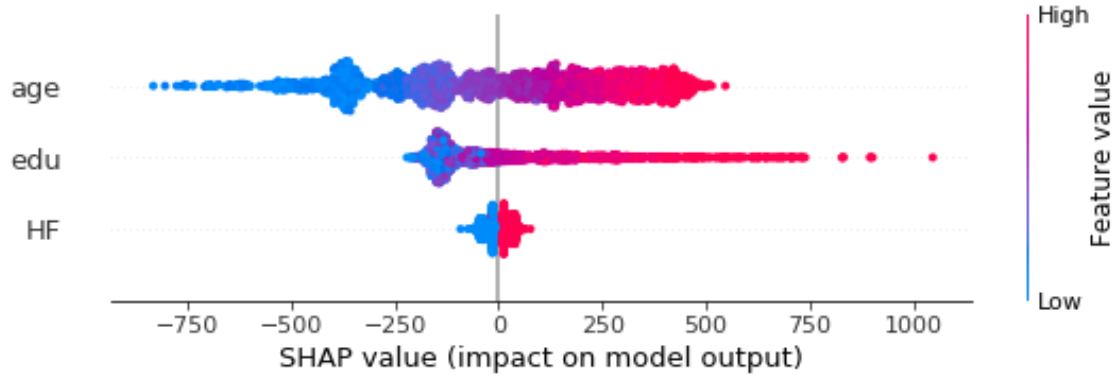
1. grey: age<25
2. gold: age>40, edu>0, H
3. blue: age>40, edu<0, H

else:

4. red: age>25, edu>0
5. green: age>25, edu<0







Méthode : illustration

Visualisation dyadique (dy.) (= valeurs et explications)

Segmentation (dy.)

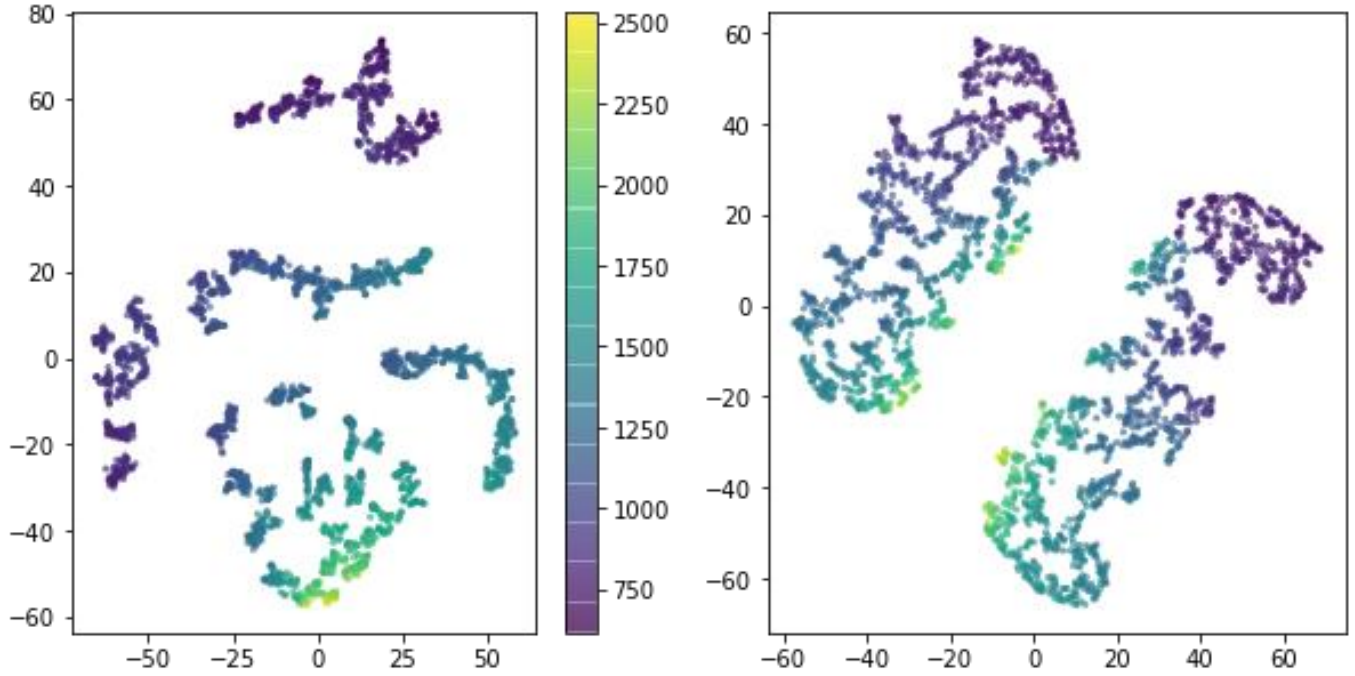
Exploration dy. et Modèle le plus simple explicable par segment

Analyse de **densité** et **d'exhaustivité**

AntakIA

Retrouver « à l'aveugle » les parcelles d'un modèle segmenté

- **EDA « Dyadique »**
  - **Classique**, sur EV (droite)
  - **Complément** : Sur EE (gauche)



→ Exploration conjointe des 2 espaces, entre DS et BO

Méthode : illustration

Visualisation dyadique (dy.) (= valeurs et explications)

Segmentation (dy.)

Exploration dy. et Modèle le plus simple explicable par segment

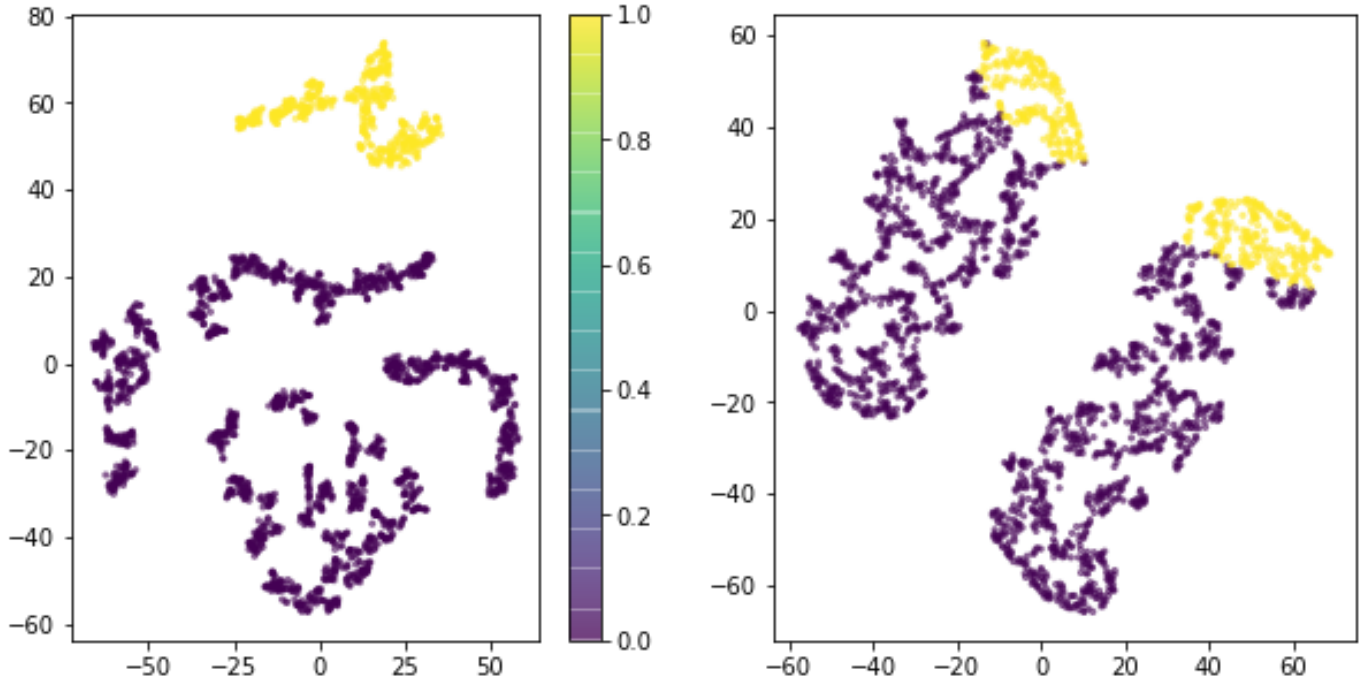
Analyse de **densité** et **d'exhaustivité**

AntakIA

Retrouver « à l'aveugle » les parcelles d'un modèle segmenté

- **EDA « Dyadique »**
  - **Exploration dyadique** (simultanément dans les 2 espaces)
  - **Aide à la description** (ici, via Skope Rules)

→ Suggestion par le DS au BO : pertinent ?



number of points selected: 480
average prediction: 796

Rules
rule: age <= 25
precision: 0.998
recall: 0.996

Méthode : illustration

Visualisation dyadique (dy.) (= valeurs et explications)

Segmentation (dy.)

Exploration dy. et Modèle le plus simple explicable par segment

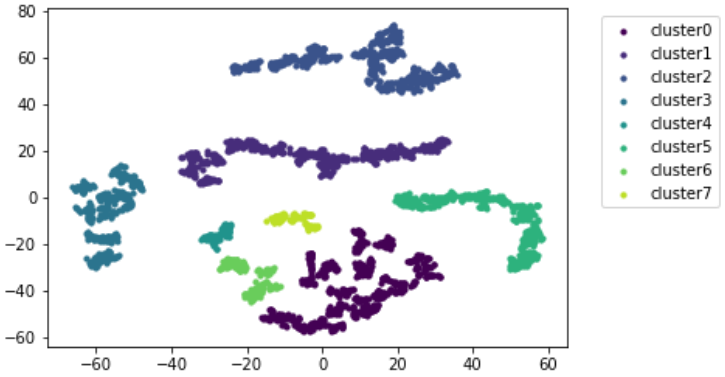
Analyse de **densité** et **d'exhaustivité**

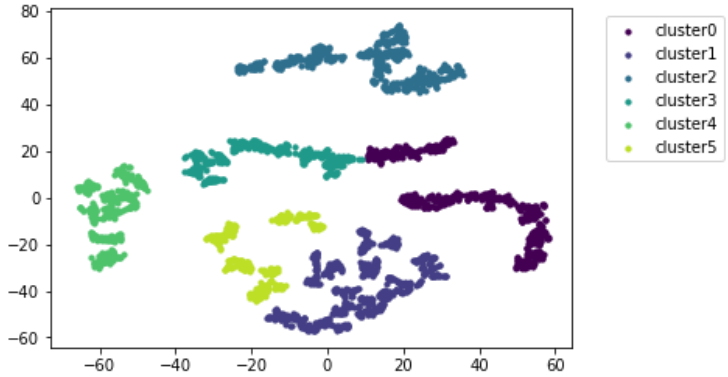
AntakIA

Retrouver « à l'aveugle » les parcelles d'un modèle segmenté

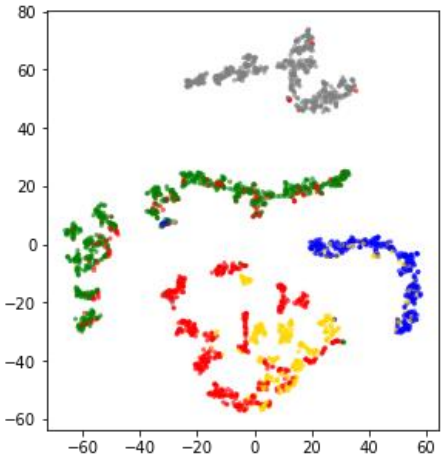
- **Segmentation « Dyadique »**
  - **Classification automatique non supervisée** (simultanément dans les 2 espaces

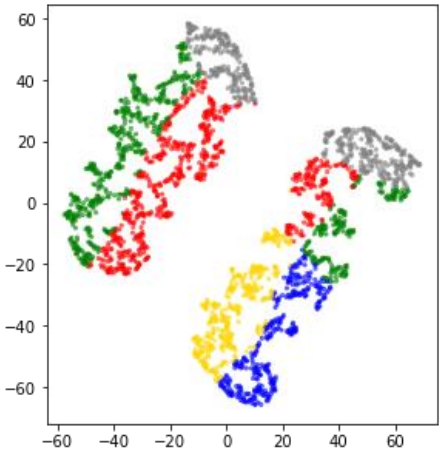
→ Suggestion par le DS au BO : pertinent ?





(Modèle théorique)





Méthode : illustration

Visualisation dyadique (dy.) (= valeurs et explications)

Segmentation (dy.)

Exploration dy. et **Modèle le plus simple explicable** par segment

Analyse de **densité** et **d'exhaustivité**

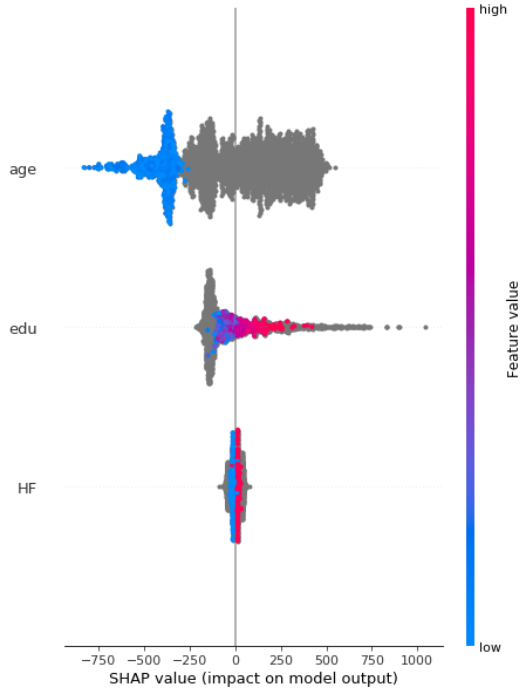
AntakIA

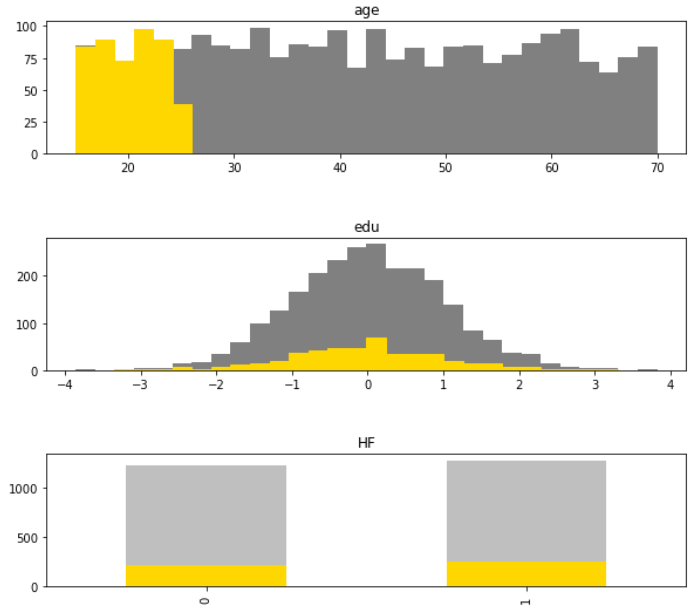
Retrouver « à l'aveugle » les parcelles d'un modèle segmenté

- **Segmentation « Dyadique »**
  - **Affinage par prédicteur** (simultanément dans les 2 espaces)

→ Suggestion par le DS au BO : pertinent ?

- Dès accord : **choix d'un modèle de substitution simple,** réduit aux variables les plus importantes





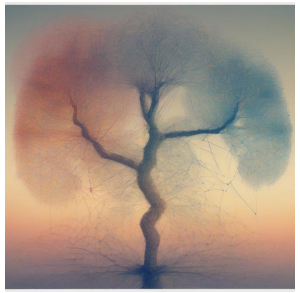
Nos travaux futurs

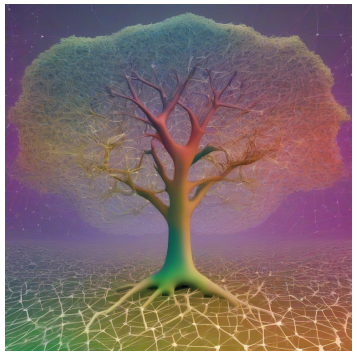
- Approfondir et étendre :
  - l'approche causale (analyse, illustration et prédiction, contrefactuels) et la formalisation de connaissances (circuits logiques ?) pour gain en robustesse
  - Les analyses topologiques (bi-clustering sur SHAP, LIME, etc.)
  - « L'équité » (individuelle, sur une région, entre régions)
  - Les dérives dans le temps (avec Telecom Paris)
- Expliquer les structures latentes :
  - Physique statistique (après apprentissage, pendant l'apprentissage) et l'approche multi-échelles
  - Distillation (matrices de connexion)

- Développer les interfaces d'explication sur des cas concrets !

→ + Lab AI-vidence : Partenariats bienvenus !

En conclusion





- L’approche cartésienne : découper un grand sujet en + petits évidents
- Structuration par les risques : des comptes à rendre … (→ AI act)
- Hybrider ?
  - **apprentissage statistique** → **déterministe sur les zones à hauts risques**, et naturellement régulières; **statistique** ailleurs.
    → connaissances accrues sur le phénomène

  - **LLM**, **IA génératives** : même approche, mais sur les modèles eux-mêmes :
    - Après apprentissage
    - Pendant l’apprentissage

Pilotes, collaborations ?

Contactez-nous !

AI VIDENCE

Merci !

David CORTÉS
06 14 173 173
david@ai-videnc e.com

Laurent MICHEL
06 61 96 60 32
laurent@ai-videnc e.com