

IA de confiance et données

Interpréter (local/global/régional)

Assurer la **robustesse** (Biais, identité de distribution)

Cas d'usage et valeur économique

par **AI**-vidence

David Cortés – CEO



AI VIDENCE

Ordre du jour

1. AI-vidence
2. Contexte : « IA de Confiance » ?
 1. 1^{ère} illustration...
 2. Plus généralement, « IA »
 3. Start up, grands groupes et recherche ?

Bonus...

4. Des applications
5. Le contexte réglementaire européen
6. Des solutions AI-vidence...
 1. Expliquer et substituer *régionalement*
 2. Traiter les hypothèses IID... *dérive des distributions* et *causalité*
7. La librairie *AntakIA*

1

Notre start up, AI-vidence

Observation: despite significant gains, there is (still) a dark side to AI...



90%

of AI projects don't go in production¹

even in Gen AI²

+



”

50%

of AI models are rapidly unplugged

”

- Stuart Russel, 2024

Source : (1) Mc Kinsey 2022, (2) Artefact 2024

Notre histoire...

July 2021 - TechSprint ACPR

- Innovation in Explanability of AI ('regional' level)



January 2022 – Confiance.AI

- Substitution by simpler models
- Multi-level analyses of TS system
- Operationnalizing ethics



2023 – 2024

- Data Drift, Selection Bias
- Causality
- Frugal Deep Learning



Air Liquide
La Plateforme



Qui nous sommes ? L'équipe fondatrice



David
Président

X (97), Telecom Madrid, Stanford
Ex conseil (Dir. IA @ PwC) et
entrepreneuriat

Stratégie & partenariats



Laurent
DG

Telecom Paris, MBA, Actuaire
Ex conseil (Roland Berger),
Investissements d'avenir @Matignon

Produit & ingénierie



Pierre
Resp. data-sciences

X(13), MSc Rech. IA à Montréal
Ex conseil (PwC) et lead DS et data
startup

Maths & data-sciences

Notre comité scientifique

Directeur Scientifique
CRIL, Université d'Artois



Pierre Marquis –
Chercheur pionnier en « IA »

Programmation logique, IA formelle et hybride, injection/extraction de connaissances humaines dans les modèles de ML

Actuaire, Chercheur
Rennes, UQAM, IVADO



Arthur Charpentier
Expert en actuariat, biais, explicabilité, non-discrimination

ENSAE

... + Marianne Clausel
Responsable PePR AI

Enseignant Chercheur en IA
Telecom Paris



Stéphan Cléménçon
Professeur LTCI/S2A
chaire industrielle "Data Science and AI for Digitalized Industry and Services"

Telecom Paris

Our clients, partners an ecosystem



Air Liquide



... ?



2

IA de Confiance ? Le problème

AI VIDENCE



GROK X1

Observation: despite significant gains, there is (still) a dark side to AI...



+



90%

of AI projects don't go in
production¹

even in Gen AI²

”

50%

of AI models are rapidly
unplugged

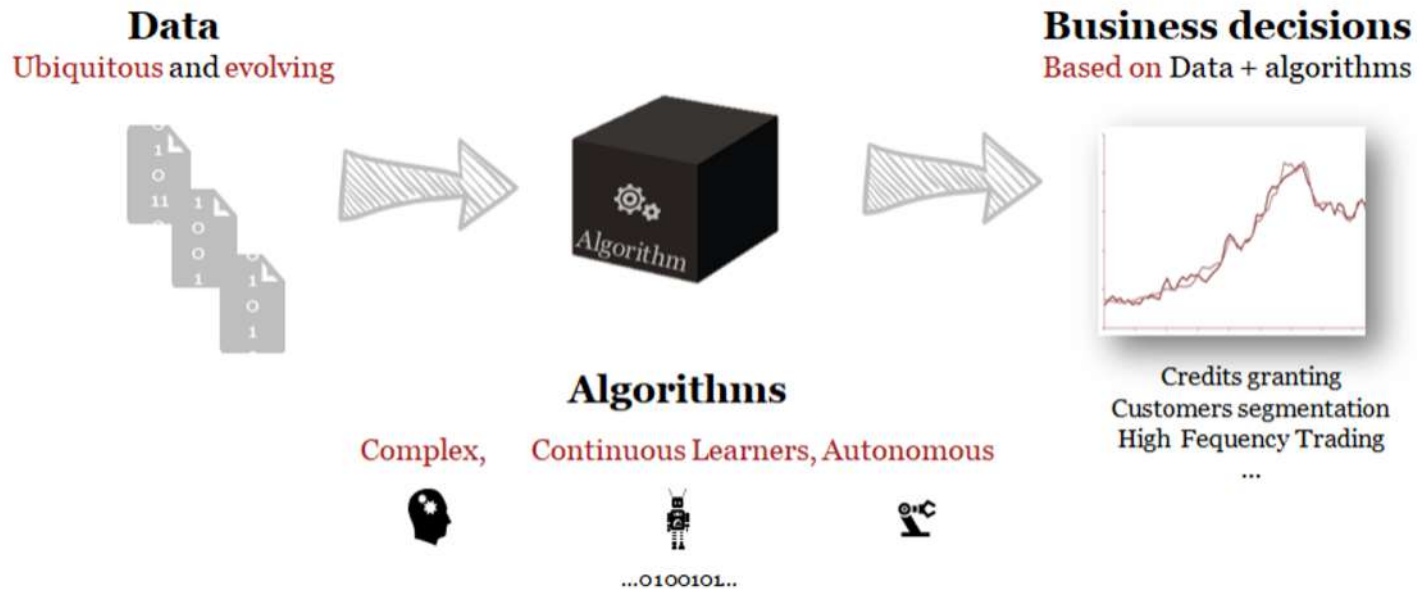
”

- Stuart Russel, 2024

Source : (1) Mc Kinsey 2022, (2) Artefact 2024

Le problème : les algos d'AI et machine learning

Nouveaux problèmes liés à l'apprentissage machine



Too complex for
transparency to be
enough

And *Correlation is
not Causality*

Passer de l'intuition à l'intelligence pour maîtriser et adopter des IA...

Systèmes humains¹

Le « système 1 »

Intuition, réflexes



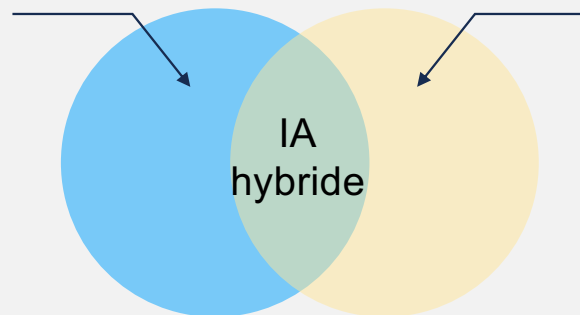
Le « système 2 »

Raisonnement, logique

Systèmes artificiels

L'IA connexionniste

apprentissage statistique automatique²



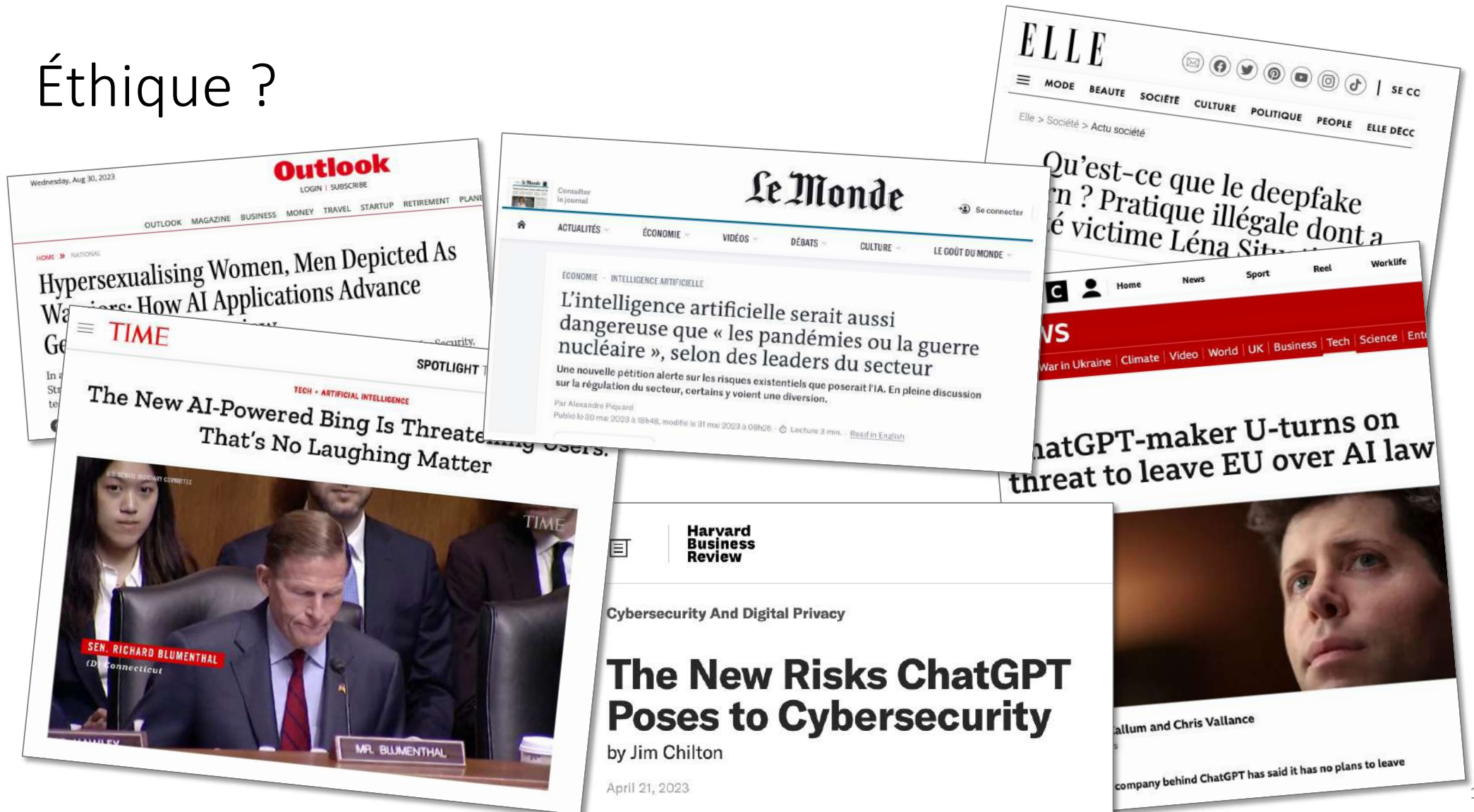
L'IA symbolique

Moteur de règles métier,
liens causaux
graphes de connaissances
...

(1) D'après les travaux de Daniel Kahneman, publiés en 2011 dans le livre « *Thinking, Fast and Slow* »

(2) « *machine learning* » en anglais

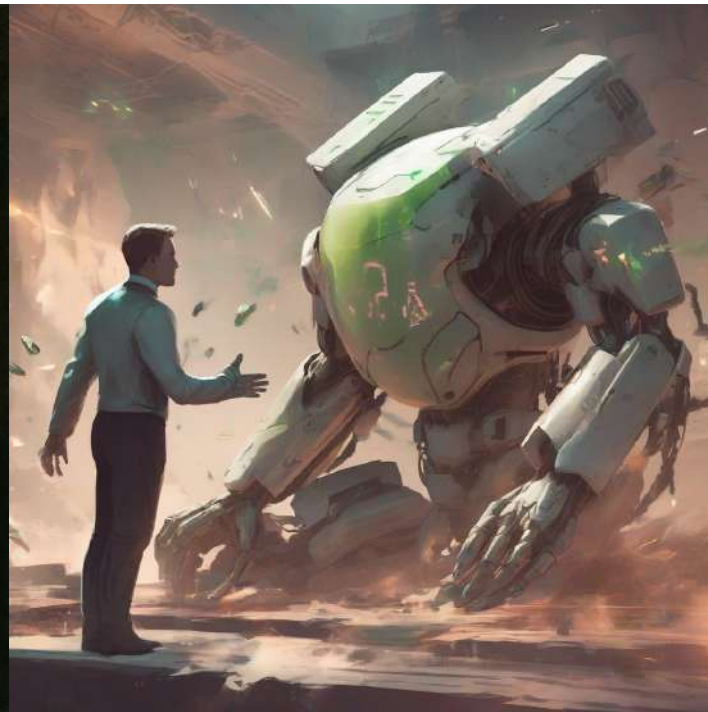
Éthique ?



Robustesse ?



Adoption ?



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

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



Three hurdles to successful AI deployment



Opacity



Models are **discriminating**
or **unethical**



Robustness ?
Liability ?

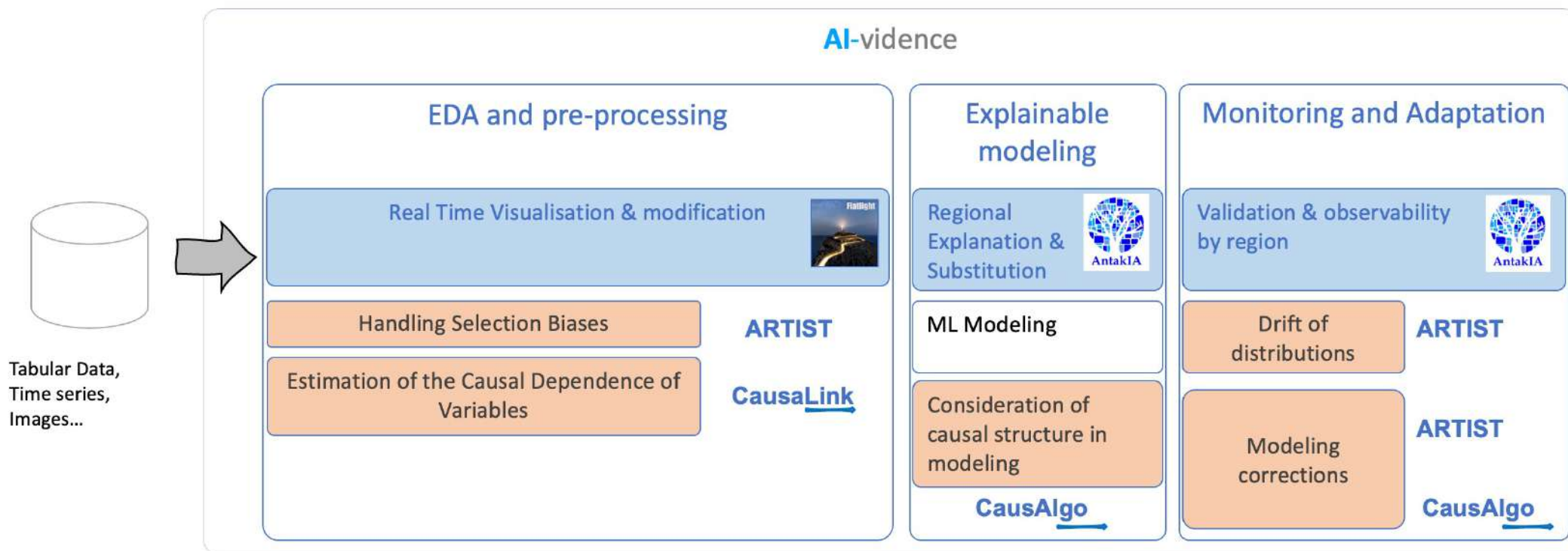


Adoption ?



Meet and maintain expectations in
ethics and **performance ?**

An Open Source processing base : Freemium along with Components under IP



The team building ARTIST



David Cortés
CEO & co-founder
École polytechnique – Telecom Madrid – Stanford

Data&IA Director at PwC – 15y in Telco

Stephan Cléménçon, PhD
Sr. Advisor
Institut polytechnique
Statistics Professor at Telecom Paris



Pierre Hulot
Dir. Data Science & co-founder
École polytechnique – BsC Polytechnique Montréal

Data&IA Consultant at PwC

Amir Dir, PhD
Data Science and Code
PhD from Institut Polytechnique



Data&IA and development Consultant



The team building CausaLink and CausAlgo



David Cortés
CEO & co-founder
École polytechnique – Telecom Madrid – Stanford

Data&IA Director at PwC – 15y in Telco

Marianne Clausel
Senior Researcher

Centrale Lyon, Agrégation in Mathematics,
DEA Dauphine
Univ. of Lorraine
Professor

Responsible for the PEPR AI : CAUSAL-IT-AI



Pierre Hulot
Dir. Data Science & co-founder
École polytechnique – BsC Polytechnique Montréal

Data&IA Consultant at PwC

Emilie Devijver
Senior Researcher

PhD from Univ. Paris-Sud
Post-doc from Leuven Univ.
CNRS Researcher



Our secret sauce :



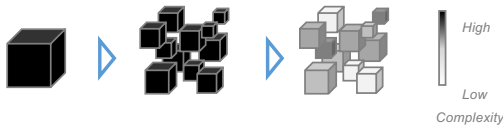
15+y R&D from french Top Data Science academics,
30+ level-A publication

- A patented bunch of **versatile scoring algorithms**
- State of the art **causal approaches** (PEPR CAUSAL-IT-AI)



Award winning Innovative methodologies to :

- **Monitor and Control** : *Vast Time Series System*
- **Detect and correct** : *quantified drift, selection biases*
- **Score** : fault detection, predictive maintenance, etc.
- **Explain** : **AntakIA[®]** (*subsystem-, operating mode-wise*)
local, 'regional', global explanations



A black and white photograph of Hannah Arendt, a woman with short, dark, curly hair, wearing a dark jacket with a circular brooch. She is looking upwards and to the right with a thoughtful expression. The background is a plain, light-colored wall.

“ Il se pourrait que nous ne soyons plus jamais capables de **comprendre**, c'est-à-dire de penser et d'exprimer, les choses que nous sommes cependant capables de faire. ”

Hannah Arendt - Condition de l'Homme Moderne
1958

Expliquer quand même...

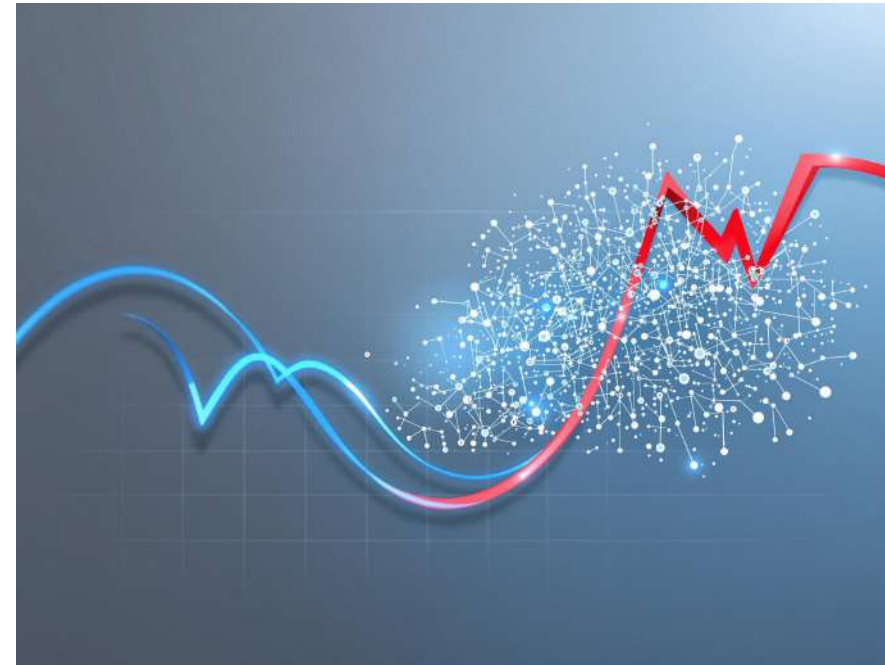
Accéder à la valeur des données...

- ~~Données~~ → « (re)construites » : 90% du temps passé
 - Découverte ou Rol ?
 - Nettoyage de données
 - Analyse exploratoire
 - Extraction et présentation de connaissances...
- Apprentissage **automatique** → hybride...
 - Itérations et approches multidisciplinaires, explications - Annotations, imputations sur données manquantes
 - Les hyperparamètres ! Itérer...
 - Expliquer : quelle valeur effective ? Quelles garanties ?
 - Gestion de la connaissance



IA de confiance ? Freins techniques, et opérationnels

- Garanties... non-respect des hypothèses de base de l'apprentissage statistique
 - **Identité des distributions** (I)
 - **Indépendance des variables** (II)
- Adoption, gain opérationnel autre (III)
- Court terme :
 - **Dérive des données**, correction des **biais** (I)
quantification des incertitudes
 - **Explication/substitution/simplification** (I, II)
- Moyen terme :
 - **Approches causales** (I, II), notamment sur système de séries temporelles
 - **Approche hiérarchique** (gestion de la complexité) (III)
- Long terme :
 - **frugalité ?** (III)



Lab d'IA, Start up et Grands groupes

- Rythme effréné d'innovations, obsolescence extrêmement rapide
→ structure agile
- Lien avec des labos de R&D, voire entre R&D et 'BU'
→ concilier un calendrier opérationnel et un calendrier R&D
→ permettre un développement concret de solutions logicielles
- Lancer des projets innovants avec soutien des pouvoirs publics
- Start up et R&D ? Anticiper les besoins à 2-4 ans





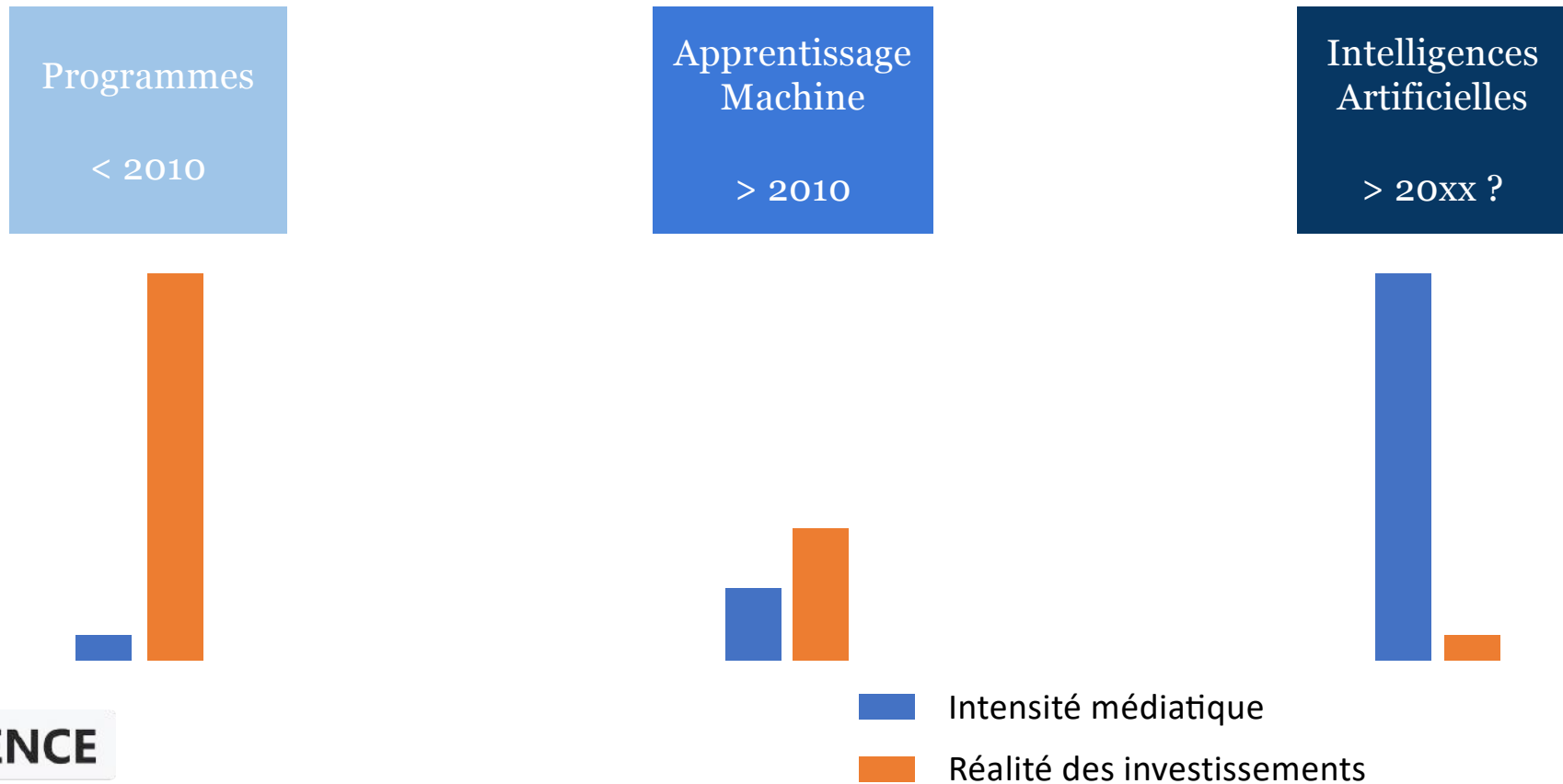
Annexes Bonus...

3 Des cas d'applications

- Commençons par les données
- Puis les algorithmes



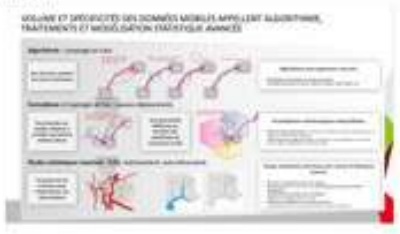
IA **générative** : attention à l'emballage mimétique...



Données - algorithmes de 'ML' - Réseaux profonds

1^{ère} illustration... Données volumineuses

• ... Conduite du changement ?



Quelques cas d'usage



FRUGAL and trustworthy AI PROJECT

A collaboration between AI-vidence and CNRS
December 5th 2024



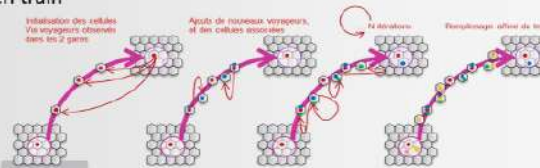
1^{ère} illustration... Données volumineuses

- ... Conduite du changement ?

VOLUME ET SPÉCIFICITÉS DES DONNÉES MOBILES APPELLENT ALGORITHMIE, TRAITEMENTS ET MODÉLISATION STATISTIQUE AVANCÉE

Algorithmie : comptage en train

Des données mobiles de nature statistique



Algorithmes auto-apprenant, récurrents :

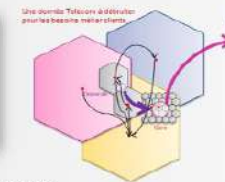
- Identification automatique du mode de transport
- Comptage journalier sur trains, sillons ferroviaires, autoroutiers, etc.

Formalisme et topologie ad hoc : pauses-déplacements

Des données sur mailles Telecom à corriger aux besoins métiers clients



Une granularité différente en fonction des spécificités de couverture locale



Formalisations mathématiques industrialisées :

- Matrices Origine/Destination via méta-formalisation en pauses-déplacements, et suppression des artefacts réseaux
- Études nationale, régionale, locale via adaptation des topologies aux données clients (région, département, commune, IRIS)

Études statistiques massives : PdM, redressement, auto-découverte

Un gisement de données dont l'exploitation est industrialisée



Études statistiques intensives pour calculs d'indicateurs avancés :

- Récurrente, régularité au jour, à la semaine
- Redressement fin via données SFR, calcul de parts de marché, de multi-équipements
- Identification automatique des corrélations (groupes de voyageurs dans un train)
- Affinement probabiliste des données mesurées (caractérisation par entropie, prédictibilité, chaînes de Markov)

Quelques cas d'usage

Time Series System (Reducing false positive rates)



1. Auto-identification of sub-systems (among 1000 sensors)
2. Multi-scale scoring and centralization method (downtime, sub-systems, sensors)
3. Filtering false anomalies due to heterogeneous data quality
4. Explainable anomaly detection (identifying errors, fault types)

Onboarding of the data center by the business of customer

Strong results: 80% less time and cost for data center operations

Enhanced adoption by business experts

Customer: **ALSTOM**

Causal forecasting and predictive maintenance



1. Causal modeling
2. Root Cause Analysis
3. Detection of regime change
4. Anomaly Detection and explanation
5. Predictive maintenance
6. Multi-scale analysis
7. Scenario: Sub-system, sub-regions

Score, identify and correct your biases and drifts

Onboarding of the data center by the business of customer

Customer: **ALSTOM**

Substituting any AI with simpler, explainable ones



1. Explore model and/or data through explanation methods
2. Identify and design early explainable systems
3. Challenge existing clusters
4. Substitute on each cluster with simpler models
5. Assess performance, stability and biases
6. Monitor simply your new AI in time

Onboarding of the data center by the business of customer

Customer: **ALSTOM**

Score, Identify and correct your biases and drifts



1. The performance scoring with explanation interface
2. Identify bias selection and quantify it
3. Export your model with low computational cost
4. Identify drift in data
5. Use Anomaly to identify potential abnormal, and explain at segment level

Onboarding of the data center by the business of customer

Customer: **ALSTOM**

Auditing, explaining black boxes (credit granting)



- Adapt the explanations to the volume and dates
- Use state of the art methods
- Provide the visualisation
- Reflect new knowledge
- Monitor the changes
- Monitor simply your segments, operation by design

Onboarding of the data center by the business of customer

Customer: **ALSTOM**

Predictive maintenance and knowledge discovery ?



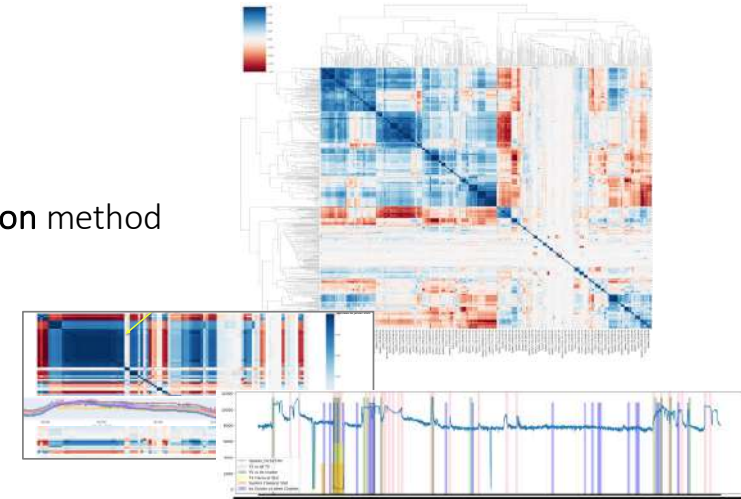
- What we suggest :
 - Causal graph of a huge system of sensors
 - Detect/Define sub-systems
 - Discover interactions between sub-systems, each element and sub-systems
 - Interpretable Causal Modeling (ICM)
- Expected results :
 - Enhanced knowledge of root causes
 - Simulation of interventions
 - Better adoption and easier conduct of change

Onboarding of the data center by the business of customer

Customer: **ALSTOM**

Time Series System (Reducing false positive rates)

1. **Auto-identification of sub-systems**
(among ~3000 sensors)
2. **Multi-scale scoring and contextualization** method
(System, Sub-systems, sensor)
3. **Filtering false anomalies**
due to heterogeneous data quality
4. **Explainable anomaly detection**
(identifying errors typology)



Co-Winner of the AMI start-up
by the **Confiance.AI** collective



- Better results than brute force and black box Deep Learning approaches
- Enhanced adoption by business experts

Confidentiel – propriété Alvidence

Confiance.ai feedback from : **NAVAL GROUP**³²



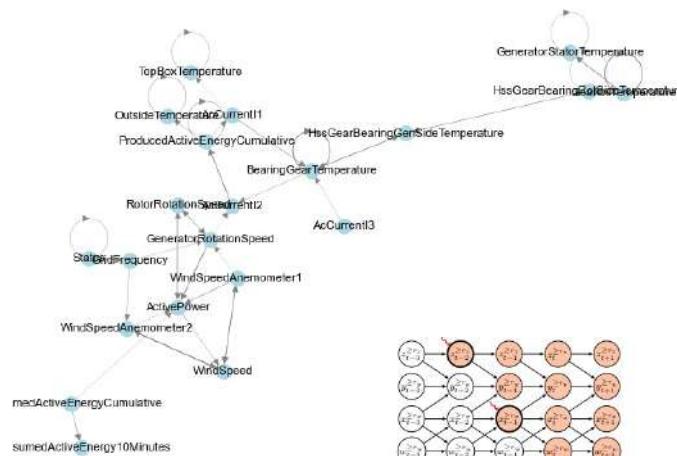
Causal forecasting and predictive maintenance



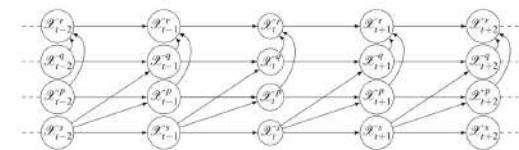
1. Guided Causal modeling
2. Root Cause Analysis
3. Detection of regime change
4. Anomaly Detection and explanation
5. Predictive Maintenance
6. Multi-level analyses
(System, Sub-system, auto-regressive)

Easier :

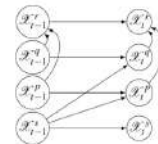
- adoption,
- scaling and
- multi-site deployment



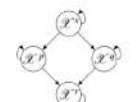
(a) Finite Threshold-based Full-time Causal Graph, T-F-TCG. Orange vertices are anomalies, thick border vertices are root causes, and the variable i is displayed by squiggly red arrows when equal to L .



Full Time Causal Graph (a)



Window Causal Graph (b)

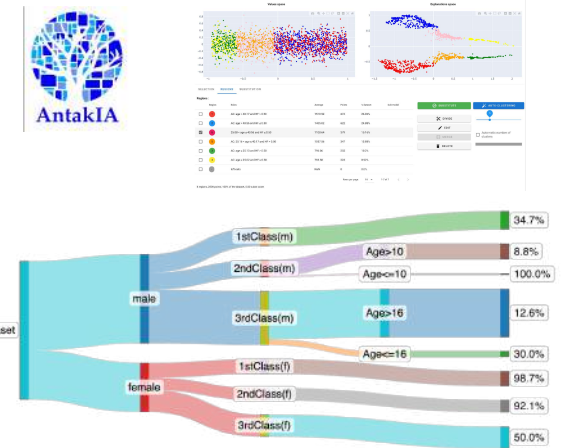


Summary Causal Graph (c)

Substituting any AI with simpler, explainable ones



1. Explore model and/or data through explanation methods
2. Identify and design easily explainable clusters or/
3. Challenge existing clusters
4. Substitute on each cluster with simpler models
5. Assess performance, stability and biases
6. Monitor simply your new AI in time



On Titanic dataset : (initial model XGBoost)

6 'regions' identified through AntakIA methodology and tool

1. Fare ≤ 9.7 and male and Embarked = S
 - 11% (Survival rate) *average baseline (best surrogate model)*
2. Pclass = 2 or 3 and Fare > 10.3 and male
 - 21% *customer model*
3. Pclass = 2 or 3 and female and not Embarked = Q
 - 63% *linear regression - using : Pclass, Age, SibSp*
4. Pclass = 1 and male
 - 45% *average baseline*
5. $48.3 < \text{Fare} \leq 188.1$ and female
 - 97% *average baseline*
6. Fare ≤ 59.6 and Embarked = Q
 - 36% *linear regression - using only : male, Parch*

→ +3% in performance, transparent collection of simpler models

On Iconic Titanic dataset,



enhanced performance
with a transparent model



Confidentiel – propriété Alvidence

Score, Identify and correct your biases and drifts

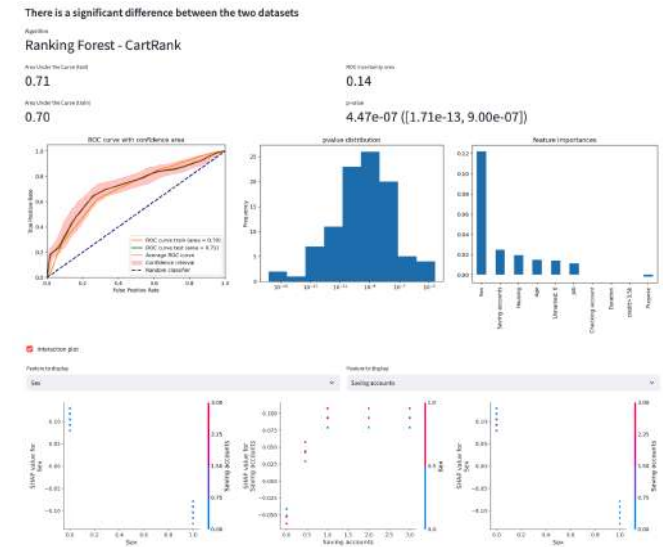
1. Top performance scoring with explanation interfaces
2. Identify bias selection and quantify it
3. Correct your model with few complementary data
4. Identify drift in data quantified
5. Use AntakIA to identify piecewise approach, and explain at segment level



On Credit dataset,



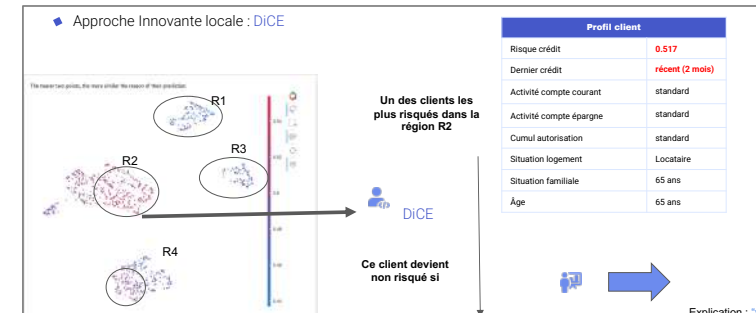
quantification of biases and top performance scoring



Auditing, explaining black boxes (credit granting)



- Adapt the explanations to the culture and stakes
- Use state of the art methods
- Provide the robustness that regulators demand
- Extract new knowledge thanks to the AI model
- Correct incrementally your model if context changes
- Monitor simply your segments, explainable by design



On credit granting :



Co-winner of the techsprint “Explainability of credit granting models”.

Innovative and versatile ‘regional’ explainability methods



Predictive maintenance and knowledge discovery ?

- What we suggest :

Causal graph of a huge system of sensors :

- Detect/Define sub-systems,
- Discover Interactions between sub-systems, each element and sub-systems
- Hierarchical Causal modeling (D.Blei)

Set common R&D partnerships to adopt state of the art solutions

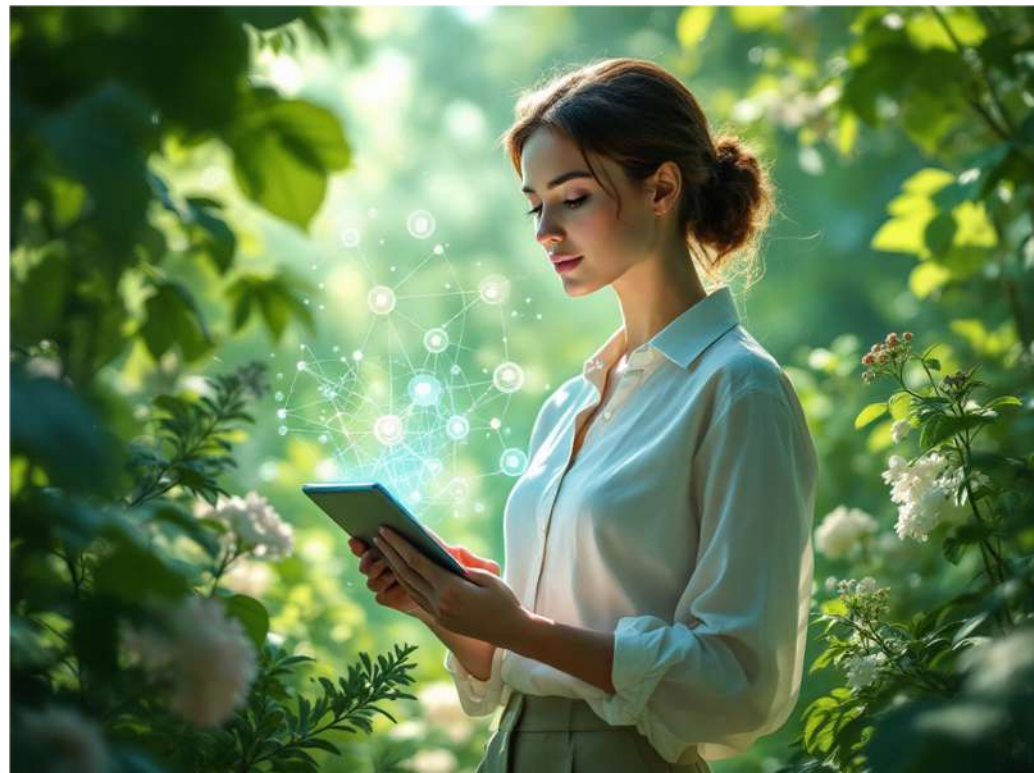
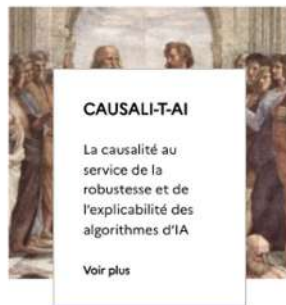
- Expected results :

- Enhanced knowledge of root causes
- Simulation of interventions
- Better adoption and easier conduct of change



FRUGAL and trustworthy AI PROJECT

A collaboration between **AI-vidence** and **CNRS**
December 5th 2024





Frugal
(and trustworthy...)
AI project ?

- **Reduce the size of large deep learning models** (incl. LLM, CV)
- **Gain in inference time!**
 - Improve user experience.
 - Allows real-time applications:
 - Reduce environmental cost

with

- **no loss of accuracy**
- **provable a priori performance...**

thanks to *cutting edge*
mathematical approaches

5

Conformité : Le cadre réglementaire Européen



Explicabilité et réglementation

- 1980/1990 : explicabilité des systèmes experts
- 2018 : rapport Villani
- 2018 : Le RGPD
- 2019 : Convention 108 du Conseil de l'Europe
- 2019 : Règlement européen P2B
- 2021-22 : “RGPD” V2 pour les modèles d'IA : *proposition de régulation par la CE.*

Entre certification / labellisation et bac à sable de régulation

Explicabilité et réglementation européenne



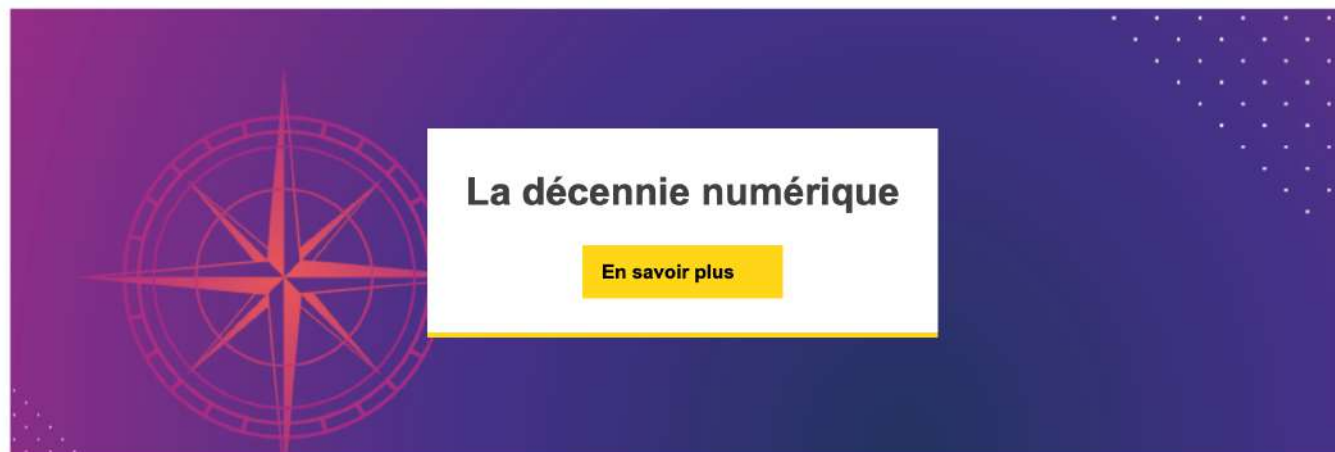
 français

Search

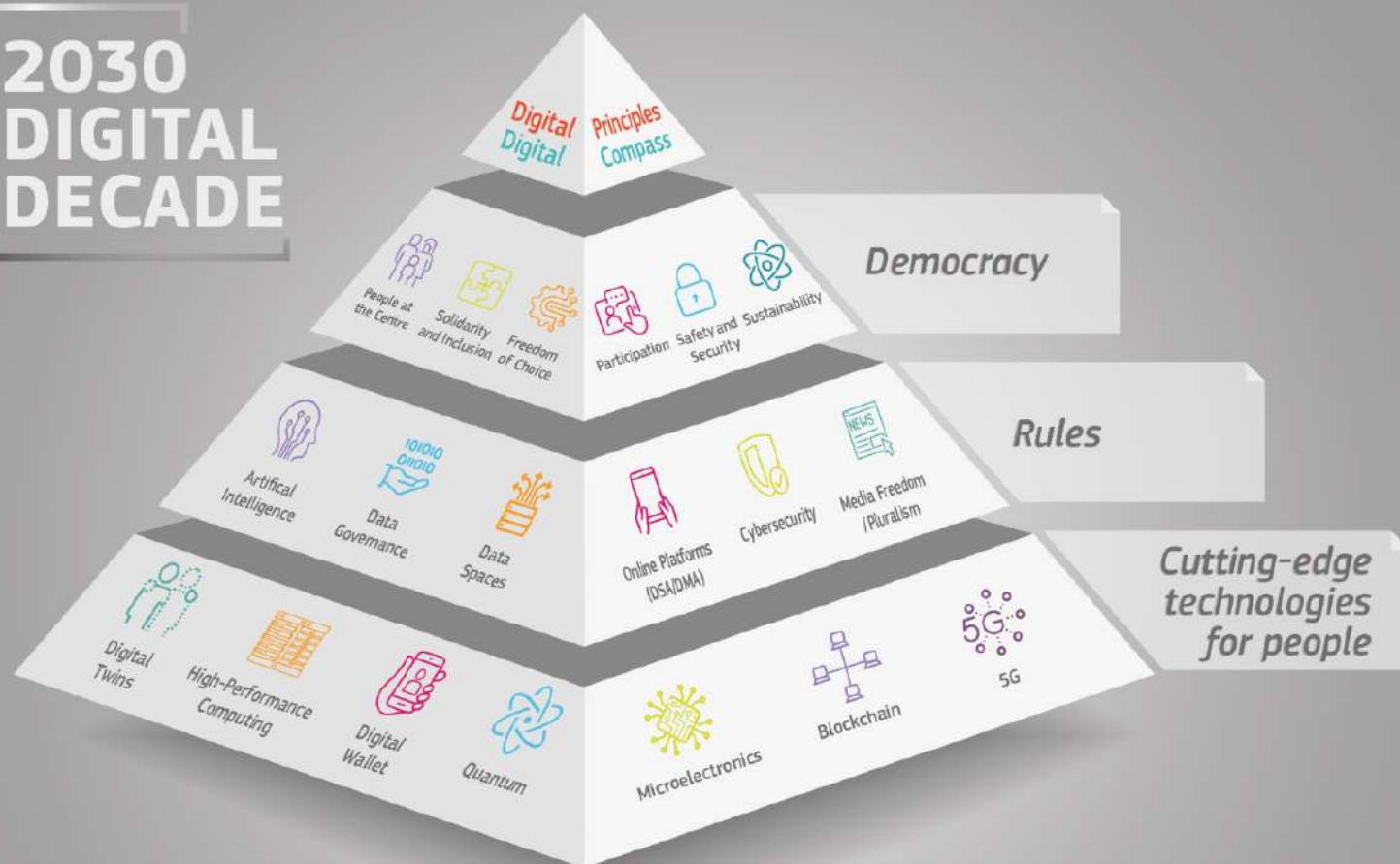
Bâtir l'avenir numérique de l'Europe

[Accueil](#) | [Politiques](#) | [Activités](#) | [Actualités](#) | [Bibliothèque](#) | [Financements](#) | [Calendrier](#) | [Consultations](#) | [AI Office](#)

[Accueil](#) > [Politiques](#)



2030 DIGITAL DECADE



#DigitalEU

Cibles

Programme politique

Projets

Droits et principes

Programme décennie numérique

Digital Decade Policy Program

**Digital Markets
Act (DMA)**

**Digital Services
Act (DSA)**

AI Act (AIA)

AI Liability
Directive

**Data Governance
Act (DGA)**

Data Act (DA)

European Data
Spaces

e-identity (eIDAS)

cybersecurity

connectivity

skills

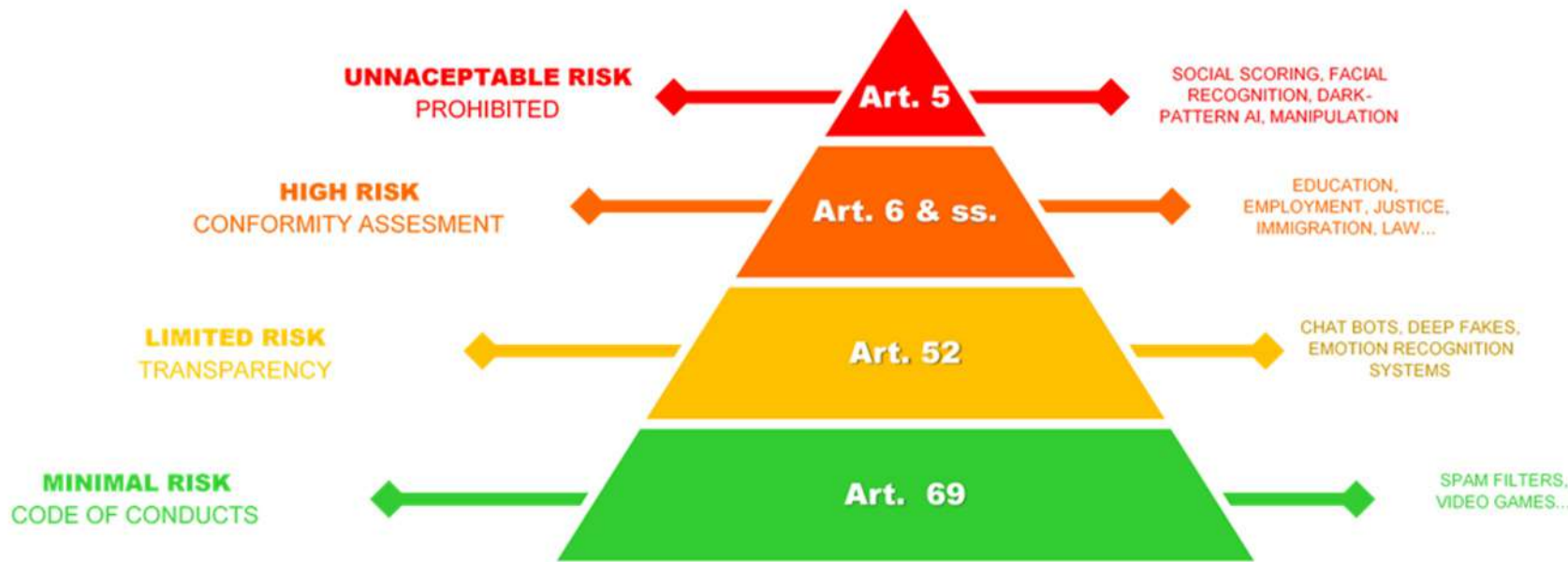
high performance
computing (HPC)

Normes, standards et spécifications techniques

Réglementation existante (RGPD,
régulation sectorielle...)

Adopté ; suivi par France Digitale

EU proposal : une approche par les risques



Data and Data Governance
(exhaustivité, qualité, **biais**,)
Transparency for Users
(infos techniques + “**intention**”,...)

Human Oversight
(**quand “débrancher” l’IA...**)

Accuracy, Robustness and Cybersecurity
(**métrique de précision**, plan de continuité, ...)
Traceability and Auditability
(documentation, **validation**, ...)

→ 7% ou 35M€, 4% ou 20M€

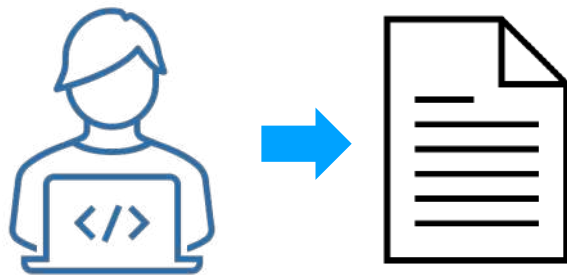
L'explicabilité : tout au long des étapes de mise en conformité

Les 10 étapes de mise en conformité des systèmes d'IA à haut risque

- | | | |
|--------------------------|---|--|
| ☐ | Système de gestion des risques | J'adopte des mesures de gestion des risques appropriées et ciblées pour répondre aux risques identifiés. |
| ☐ | Données et gouvernance des données | J'utilise des données d'entraînement de qualité, je respecte les pratiques appropriées de gouvernance des données, je m'assure que les jeux de données sont pertinents et non-biaisés. |
| ☐ | Documentation technique | J'inclus les éléments minimums spécifiés dans l'Annexe IV. |
| ☐ | Traçabilité | Je m'assure que des archives sont disponibles tout au long de la durée de vie du système d'IA, avec un suivi conçu pour la traçabilité et leur transparence. |
| ☐ | Supervision humaine | J'incorpore des outils d'interface homme-machine pour prévenir ou minimiser les risques en amont, permettant ainsi aux utilisateurs de comprendre, d'interpréter et d'utiliser en confiance ces outils. |
| ☐ | Exactitude, robustesse et sécurité | J'assure une exactitude, une robustesse et des mesures de cybersécurité continues, tout au long du cycle de vie du SIA, avec des métriques de précision déclarées, une résilience contre les erreurs et des mesures appropriées pour traiter les biais potentiels. |
| ☐ | Système de gestion de la qualité | J'établis et je documente un système de gestion de la qualité couvrant la conformité réglementaire, la conception, le développement, les tests, la gestion des risques, la surveillance post-commercialisation, la déclaration d'incidents, la communication, la gestion des données, la conservation des enregistrements, la gestion des ressources et la responsabilité. |
| ☐ | Déclaration de conformité de l'UE | Je rédige la déclaration de conformité, lisible et signée, pour chaque système d'IA à haut risque, affirmant la conformité avec les exigences du Chapitre 2, je la maintiens à jour pendant 10 ans, je sou mets des copies aux autorités nationales, et je la mets à jour quand nécessaire. |
| ☐ | Marquage CE | Je m'assure que le marquage CE est apposé de manière visible, lisible et indélébile, ou numériquement accessible pour les systèmes numériques, indiquant ainsi la conformité aux principes généraux et aux lois de l'Union applicables. |
| ☐ | Enregistrement | Avant de mettre sur le marché ou de mettre en service la solution d'IA, j'inscris l'entreprise ainsi que le système dans la base de données de l'UE mentionnée à l'Article 60. |

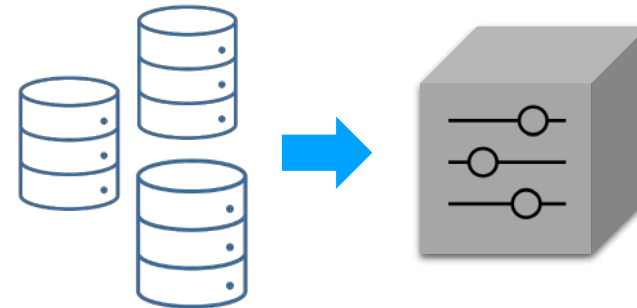
La révolution de l'IA a un prix : l'opacité des modèles et la défiance associée

Avant



- Il suffisait de lire ce programme pour comprendre le fonctionnement du logiciel

Aujourd'hui



- Moins besoin de programmer MAIS beaucoup plus difficile de comprendre

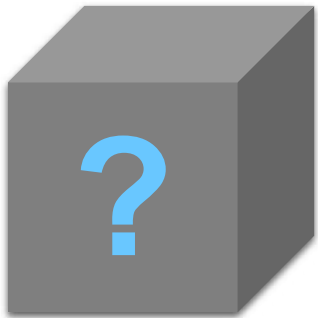


Défi : Expliquer

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

Pour qui ?

Pour quoi ?



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

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

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



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 la bonne échelle

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

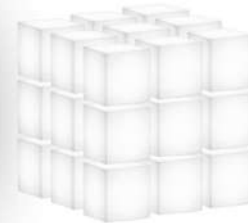


A New Tool to Acquire Knowledge from AI

Explorer une IA

Echanger, documenter

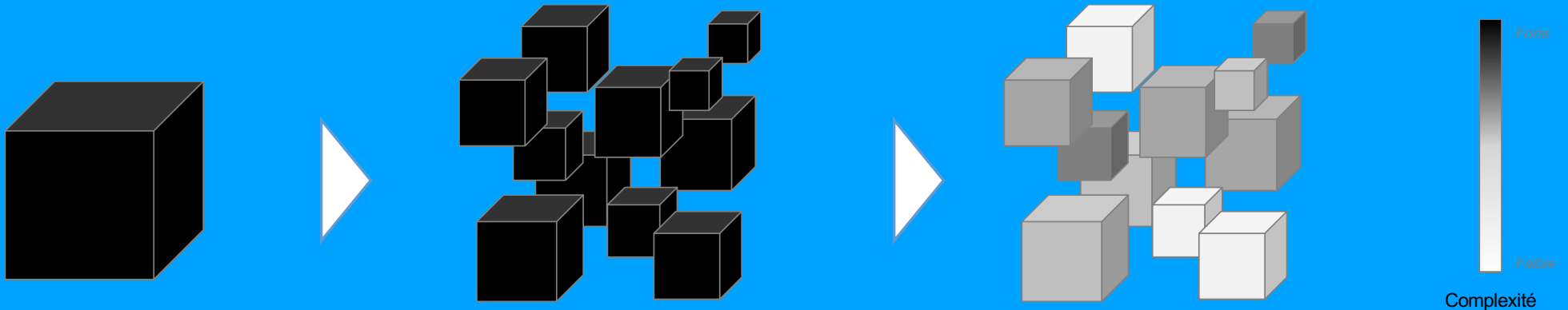
Substituer





6 Expliquer, substituer : AntakIA

- Expliquer à une échelle compréhensible
- Simplifier : ajuster la complexité



Our solutions by **AI-vidence** :



Interpretable
by design



Control of the
biases



Optimal and monitored
performance



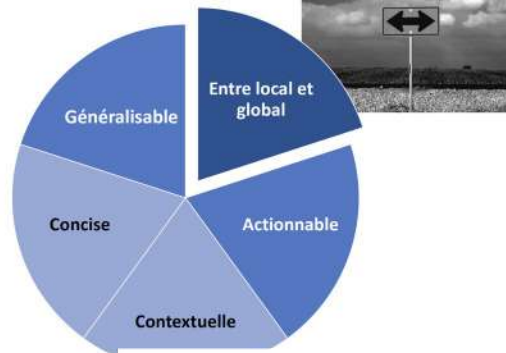
ARTIST

User Friendly SaaS solutions and methodologies

Optimal for *explainable AI* (esp. scoring, anomaly detection), *data drifts*, etc...

La solution : le besoin **métier** d'approche régionale

Points d'attention ressortis du Tech Sprint pour une bonne explication de modèle



Principaux enseignements

- Importance de la **pluridisciplinarité** et de la polyvalence
- La question du "**build vs. buy**" se pose aussi en ML
- **R&D avancée** dans le secteur
 - Shapash (MAIF)
 - Active Coalition Values (QuantMetry)
 - Skope-rules (BPCE)
- **Simplicité ≠ explicabilité**



“L'approche pure boîte noire a montré ses limites et la construction de modèle de ML doit être itérative, agile, multidisciplinaire et documentée ; les explications doivent se faire en présence d'experts métiers et de data science.
-Un participant au Tech Sprint

Entre le local et le global : *clustering de valeurs SHAP*

Explications régionales



Echelle globale
les effets importants en "général"



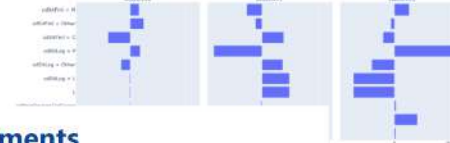
Echelle régionale
segment de clientèle
Régime de fonctionnement



Echelle locale
centrée sur un client



Profils-types



Exemples d'explications produites pour l'auditeur

Objectifs : catégoriser les clients, comprendre l'ensemble du processus et l'objectif du modèle, ses limites, comprendre si le modèle fait ce pour quoi il est conçu

Modèle de substitution : difficile à définir, encore plus à entraîner (surtout en audit externe)

Méthode post-hoc : explications globales + focus sur les régions de « stress »





Les obstacles à l'interprétation : la complexité...

- La **diversité** des explications !
- **Le mur de la complexité** (7 ± 2 , G.Miller)... besoin de :
 - **synthétiser** après l'analyse...
 - Adapter une **échelle compréhensible par les humains**
- **La robustesse/fiabilité** : i.i.d. !
 - **Dérive des distributions** : plus facile par approche « par morceaux »
 - **Indépendance** : gérer les interactions !! Idem...

Diversité des explications : sous-échantillonnages ?

On the overlooked issue of defining explanation objectives for local-surrogate explainers

Rafael Poyiadzi^{*1} Xavier Renard^{*2} Thibault Laugel² Raul Santos-Rodriguez¹ Marcin Detyniecki^{2,3,4}

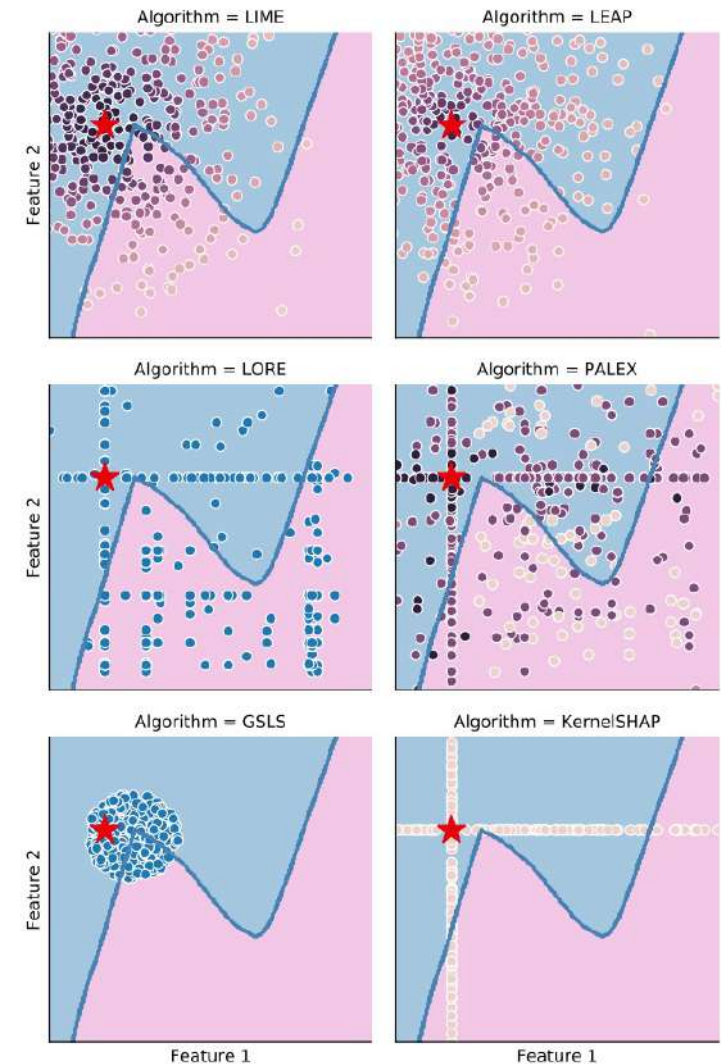
Abstract

Local surrogate approaches for explaining machine learning model predictions have appealing properties, such as being model-agnostic and flexible in their modelling. Several methods exist that fit this description and share this goal. However, despite their shared overall procedure, they set out different objectives, extract different information from the black-box, and consequently produce diverse explanations, that are -in general- incomparable. In this work we review the similarities and differences amongst multiple methods, with a particular focus on what information they extract from the model, as this has large impact on the output: the explanation. We discuss the implications of the lack of agreement, and clarity, amongst the methods' objectives on the research and practice of explainability.

fore, a careful analysis of the use-case should be carried out to define the interpretability objective of each situation and choose the most appropriate interpretability method, as it is unlikely that a one-size-fits-all solution exists.

Yet, we argue that the current literature on model surrogates to explain a prediction lacks the clarity needed for a practitioner to make an informed choice on which method to use, given explanation needs. Existing approaches usually lack transparency with regards to the explanation needs they propose to solve, on their specifications and ultimately on their formal objectives. This situation (1) fuels a disseminated research with propositions that are difficult to compare and (2) prevents a sound development of the explainability practice.

In this paper, we propose a study to highlight the diversity amongst the approaches categorized under the same vague objective of “explaining a prediction with a model surrogate”. This work is based on a theoretical analysis of proposed solutions and an experiment to illustrate the





Mieux expliquer ? À chacun selon ses enjeux,
pour permettre l'interprétation

Data Scientist +

Client	Explication locale	Information / Justif. contextualisée
	Explication globale + « régionale »	Interprétation / compréhension
	Preuve	+ de Garanties de fonctionnement



Expliquer à une échelle humaine



Compréhension humaine de phénomène complexe

1637

Discours de la Méthode

René Descartes

4 étapes :

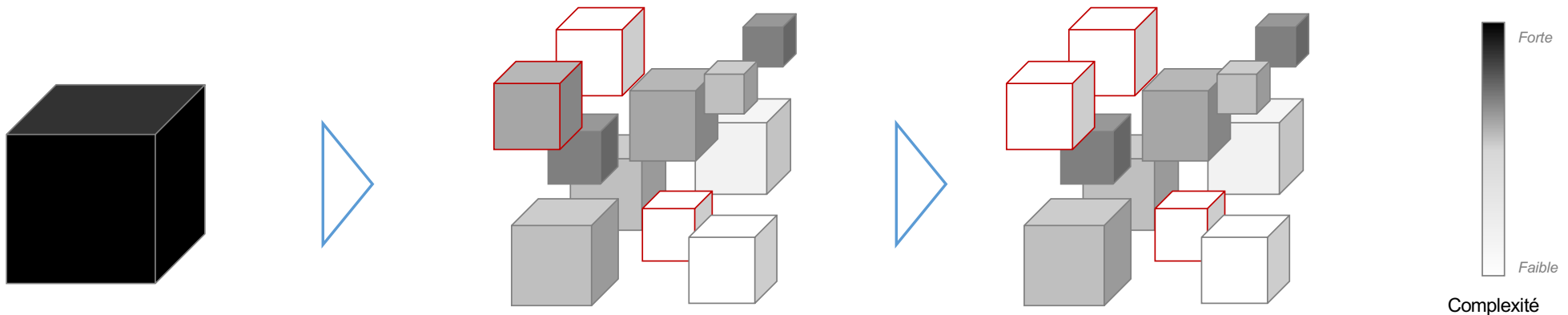
- Découpe en parcelles
- Explication « évidente »
- Reformulation et substitution
- Exhaustivité

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





Notre approche de l'explicabilité...



Une **découpe explicable par construction** :

- Création d'un **espace des explications** (Shapley, LIME...)
- Analyse **topologique des régularités** sur les 2 espaces :
 - **Exploration libre**
 - **Aide au clustering** (ou Auto-clustering)
- **Description simple** automatique via système de règles

Un **ajustement de la modélisation** par morceau selon :

- la **complexité résiduelle observée**
- le **niveau de risque**
 - **risqué** : si nécessaire, **compromis explicabilité**
- les **critiques de l'expert métier**
 - **Accroissement de connaissances**
 - **Correction des apprentissages erronés**

Output :

Concrete tools to address **adoption** and **regulatory** challenges 'beyond testing'

Ethical alignment

Simpler and **frugal** models

<https://github.com/AI-vidence/antakia>





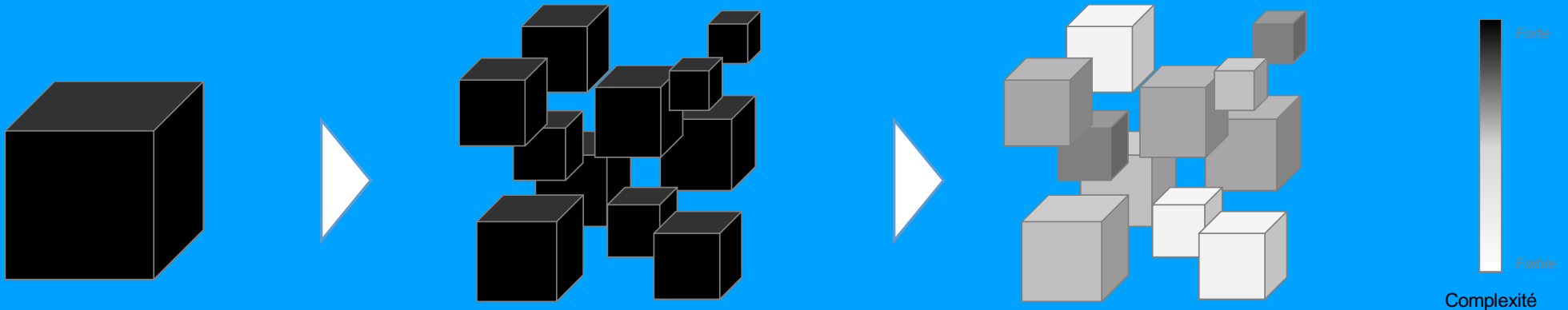
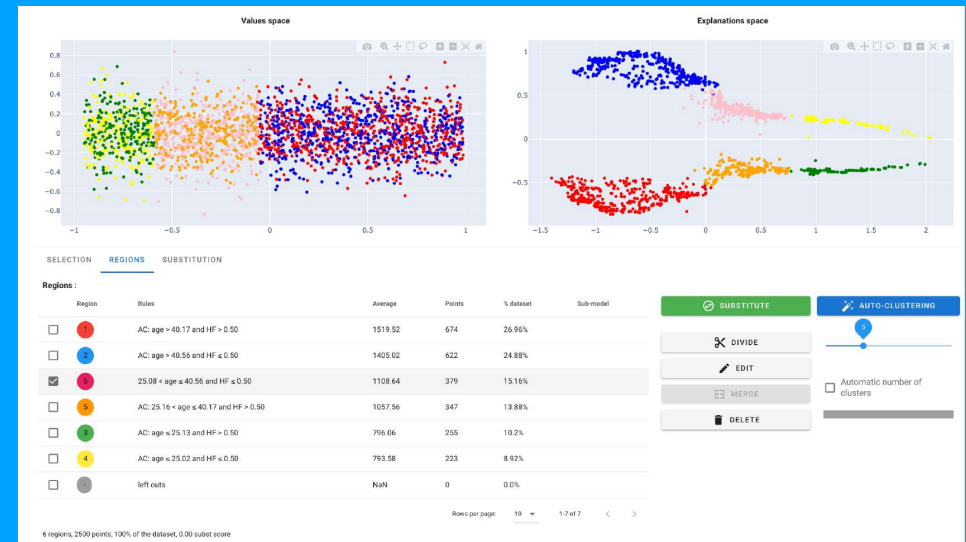
Les LLMs ? Principales réserves

- Les modèles **les plus complexes**... (~1000 milliards de paramètres)
- Sujet aux « **hallucinations** » : aucune garantie de fidélité/précision
- **Confidentialité** des données d'entrées (*'prompt'*)
- **Contrats** léonins avec les GAFAM
- Utilisation « clandestine » par vos équipes (*'BYOD'*)
- **Excès de confiance**/défaut de vérification
- **Obsolescence extrêmement rapide** : urgent d'attendre avant une solution « cousue main » ?

A New Tool to Acquire Knowledge from artificial intelligence

7 La librairie AntakIA

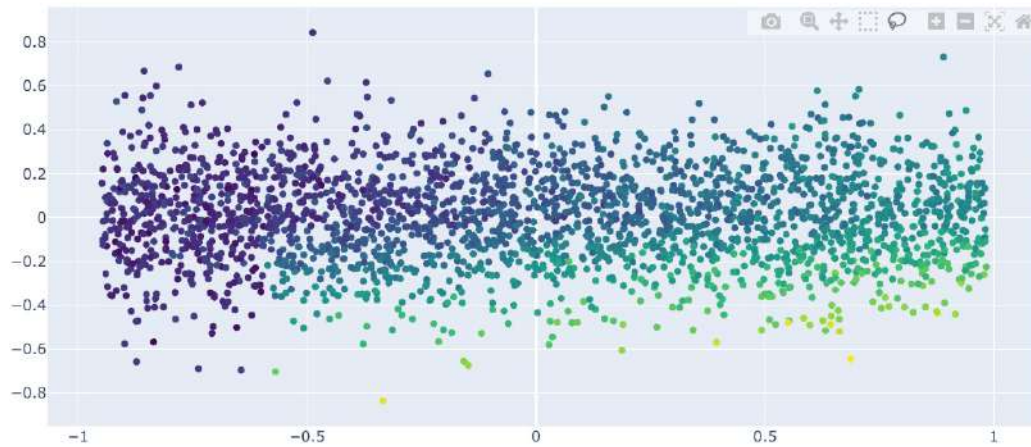
- Expliquer à une échelle compréhensible
- Simplifier : ajuster la complexité



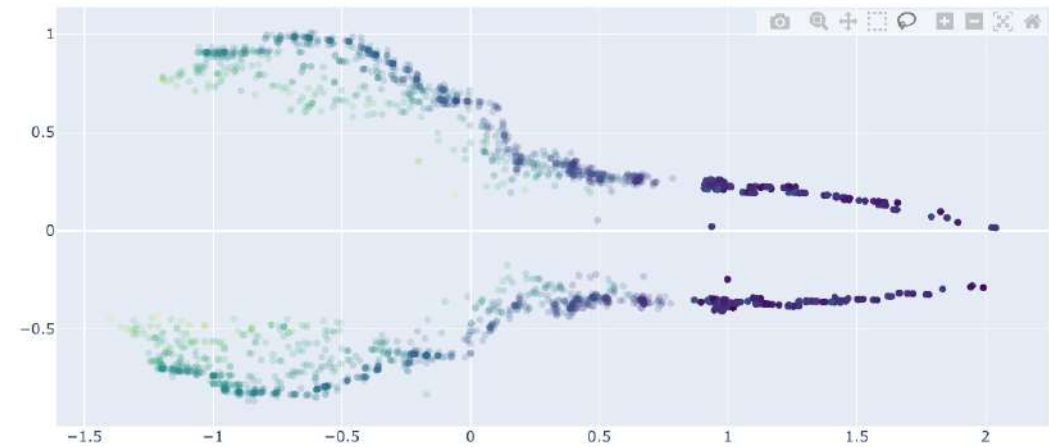
Outil pour définition de régions explicables et substitutions



Values space



Explanations space



SELECTION REGIONS SUBSTITUTION

Creating new region :

- 476 point selected
- 19.04% of the dataset

FIND RULES

CANCEL

UNDO

VALIDATE RULES

No rule to display for the VS space

- Precision : 0.19, recall : 1.00, f1_score : 0.32, target_avg : 1226.08
- N/A

age - True

14.48

70.03

No rule to display for the ES space

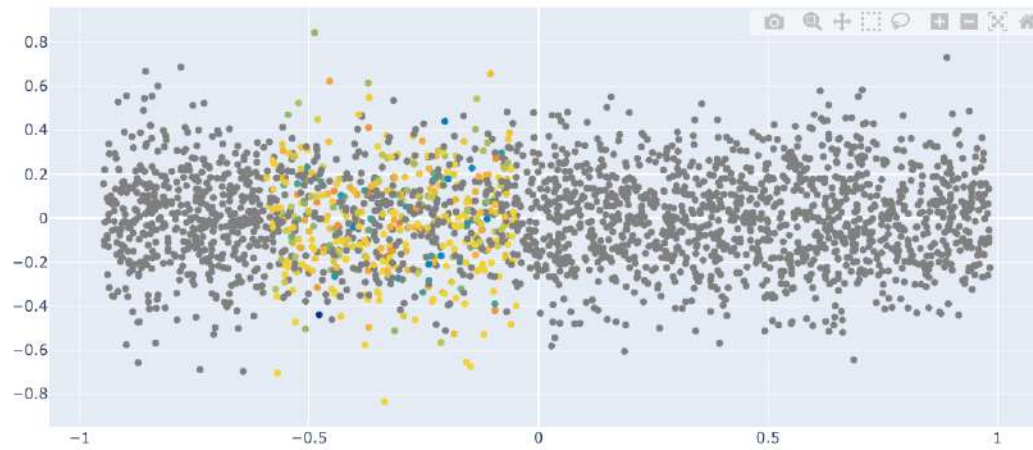
- Precision : 0.19, recall : 1.00, f1_score : 0.32, target_avg : 1226.08
- N/A

age - True

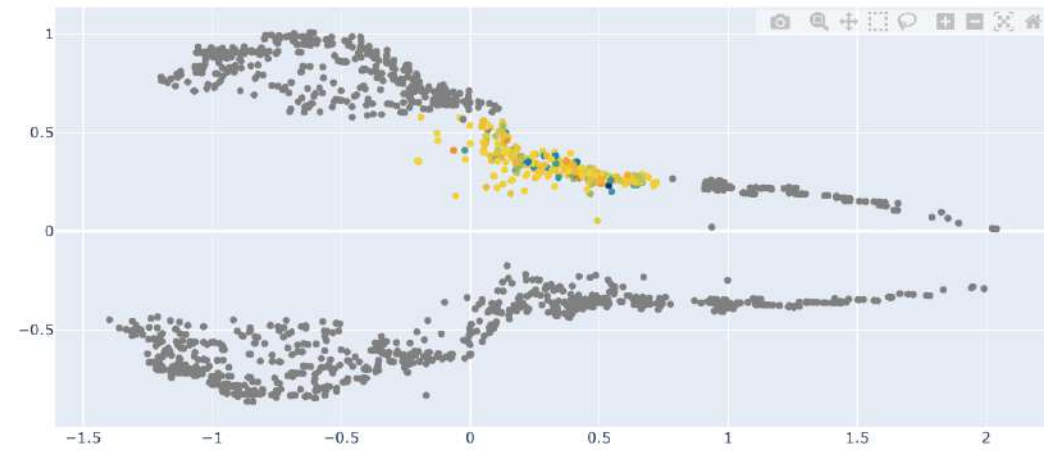
edu - True

HF - True

Values space



Explanations space

SELECTION REGIONS SUBSTITUTION

Region 7 25.08 < age ≤ 40.56 and HF ≤ 0.50, 379 points, 15.2% of the dataset

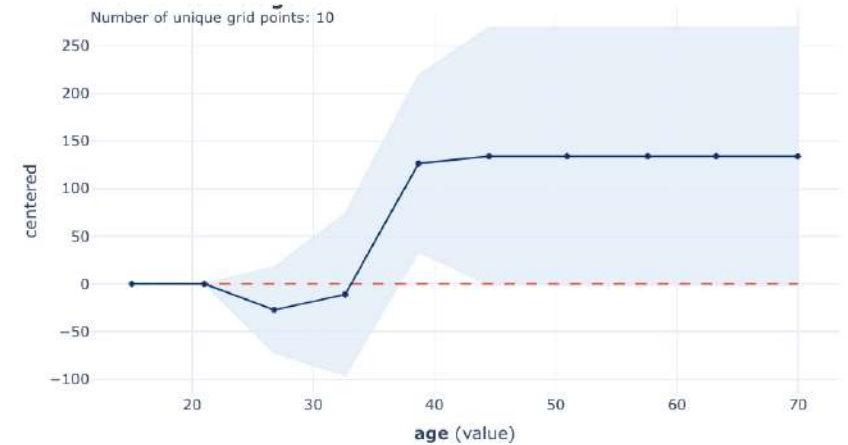
	Sub-model	MSE	MAE	R2	delta
☒	Decision Tree	10132.64	67.00	0.91	-43.03
☐	Original Model	10175.66	79.99	0.91	0.00
☐	Explainable Boosting Tree	12350.69	88.82	0.89	2175.02
☐	Linear Gam	14506.21	97.43	0.87	4330.55
☐	Linear Regression	36351.42	149.47	0.66	26175.75
☐	Lasso Regression	36371.43	149.48	0.66	26195.76
☐	Ridge Regression	36396.69	149.50	0.66	26221.03
☐	Average Baseline	108261.62	237.30	-0.01	98085.95

Rows per page: 10 1-8 of 8

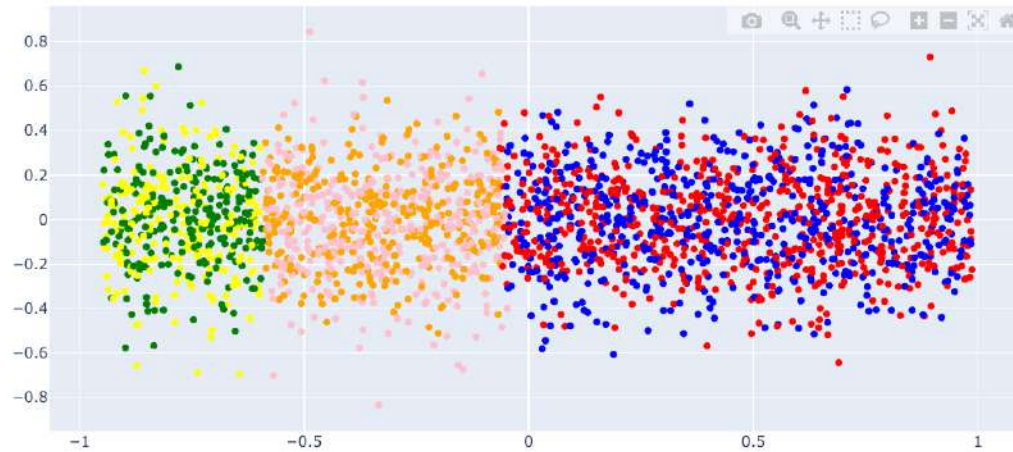
✓ VALIDATE SUB-MODEL

FEATURE IMPORTANCE PARTIAL DEPENDENCY

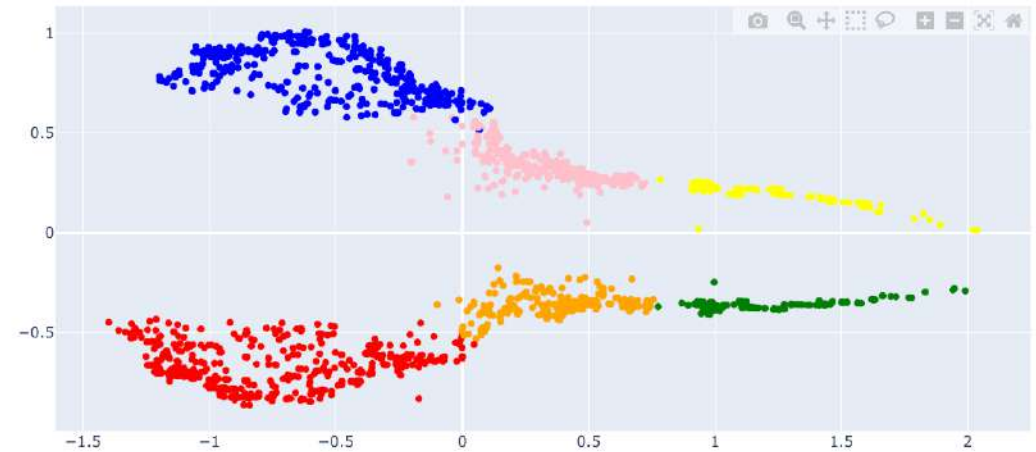
age



Values space



Explanations space

SELECTION **REGIONS** SUBSTITUTION

Regions :

	Region	Rules	Average	Points	% dataset	Sub-model
☐	1	AC: age > 40.17 and HF > 0.50	1519.52	674	26.96%	
☐	2	AC: age > 40.56 and HF ≤ 0.50	1405.02	622	24.88%	
☒	5	25.08 < age ≤ 40.56 and HF ≤ 0.50	1108.64	379	15.16%	
☐	5	AC: 25.16 < age ≤ 40.17 and HF > 0.50	1057.56	347	13.88%	
☐	3	AC: age ≤ 25.13 and HF > 0.50	796.06	255	10.2%	
☐	4	AC: age ≤ 25.02 and HF ≤ 0.50	793.58	223	8.92%	
☐	-	left outs	NaN	0	0.0%	

Rows per page: 10 1-7 of 7 < >

6 regions, 2500 points, 100% of the dataset, 0.00 subst score

SUBSTITUTE

AUTO-CLUSTERING

DIVIDE

EDIT

MERGE

DELETE

☐ Automatic number of clusters

Editing Region 6

- 379 point selected
- 15.16% of the dataset

FIND RULES

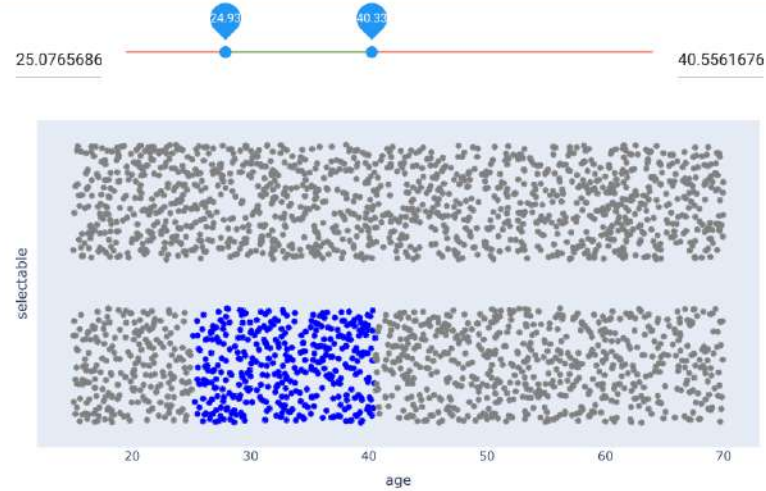
CANCEL

UNDO

VALIDATE RULES

Rule(s) applied to the values space

- Precision : 1.00, recall : 1.00, f1_score : 1.00, target_avg : 1108.64
- $25.08 < \text{age} \leq 40.56$ and $\text{HF} \leq 0.50$

 $25.08 < \text{age} \leq 40.56$  $\text{HF} \leq 0.50$

edu - True

Data selected

age	edu	HF
28.39	-0.91	0
26.77	0.53	0
39.86	0.16	0

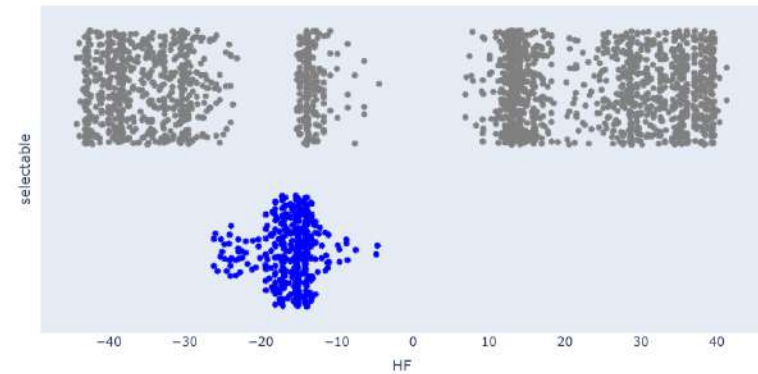
No rule to display for the ES space

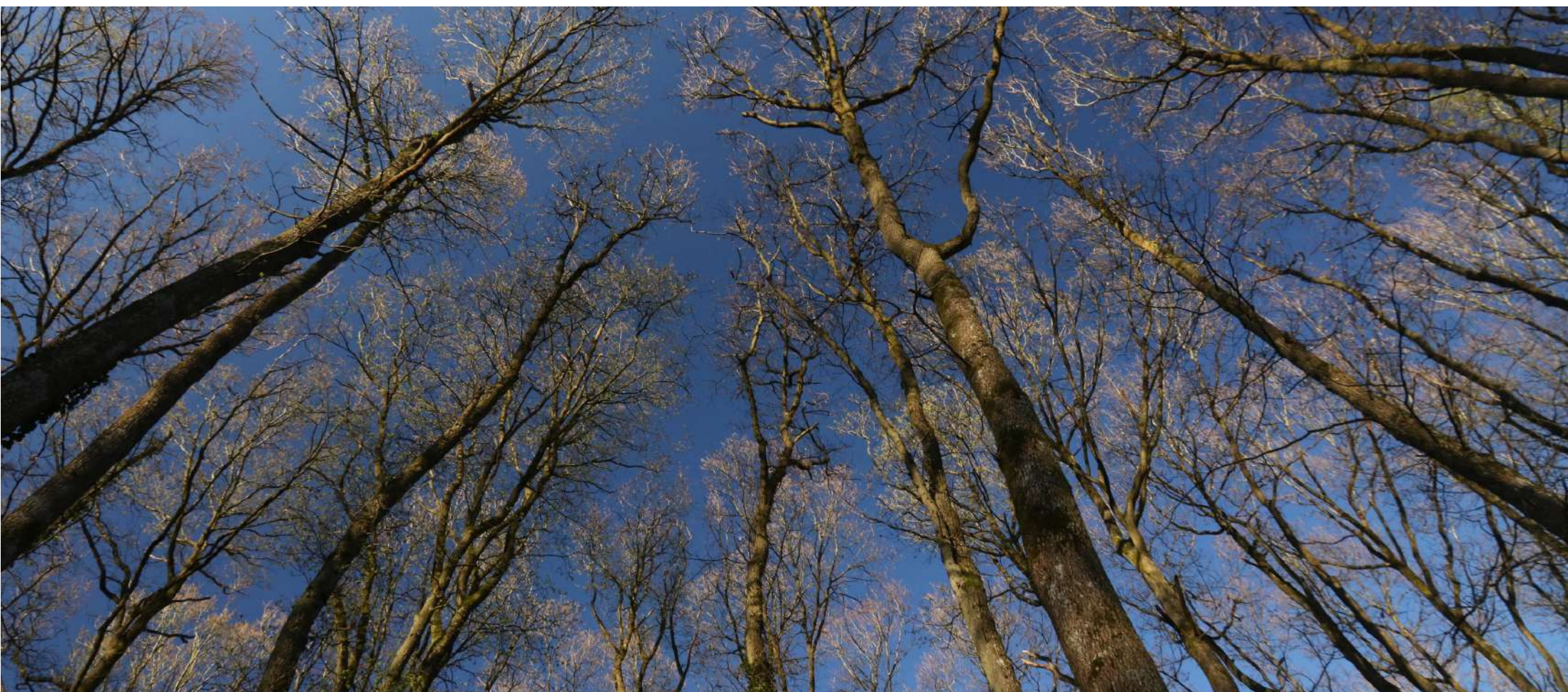
- Precision : 0.15, recall : 1.00, f1_score : 0.26, target_avg : 1226.08
- N/A

age - True

edu - True

HF - True





Merci !

Nous contacter :

David Cortés – Président Fondateur - david@ai-vidence.com – tel : 06 14 173 173

Pierre Hulot – Dir. Science des Données – pierre@ai-vidence.com

