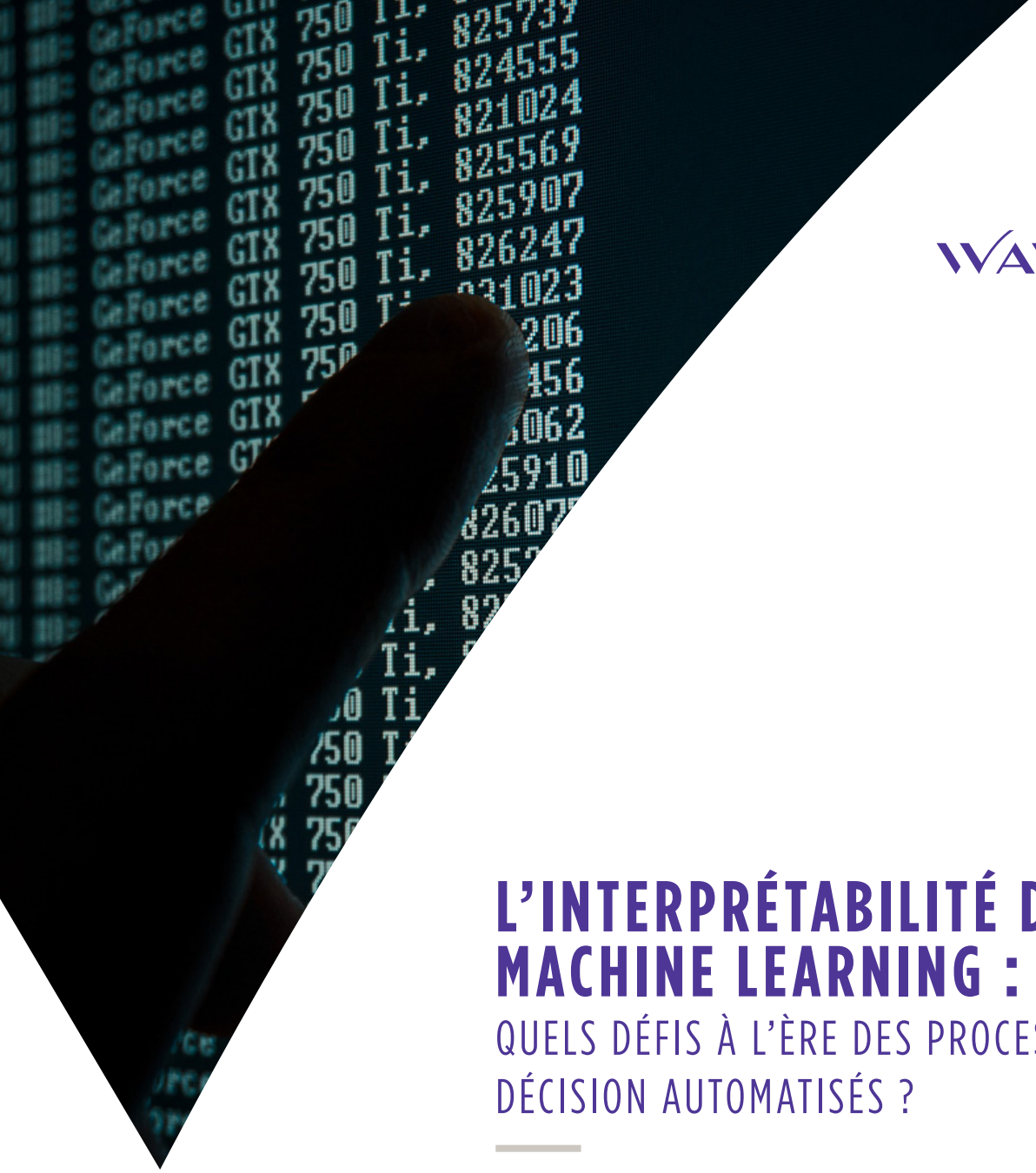


The Positive Way

WAVESTONE

L'INTERPRÉTABILITÉ DU MACHINE LEARNING :

QUELS DÉFIS À L'ÈRE DES PROCESSUS DE DÉCISION AUTOMATISÉS ?

AUTEURS

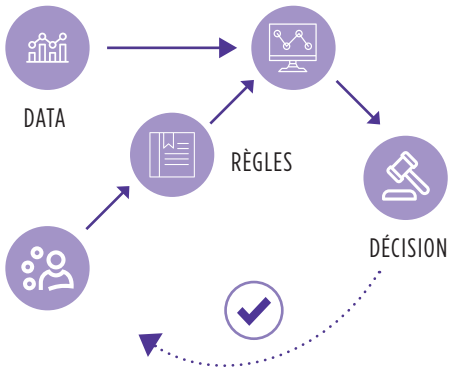
ALEXANDRE VÉRINE
alexandre.verine@wavestone.com

STÉPHAN MIR
stephan.mir@wavestone.com

L'évolution de l'intelligence artificielle depuis les années 1970 a bouleversé la conception des processus décisionnels. Le Machine Learning permet aujourd'hui d'apprendre directement des données plutôt que des connaissances humaines, avec un niveau accru de précision. Le manque d'interprétabilité et l'introduction de biais potentiels ont cependant généré des problématiques d'ordre éthique et juridique. L'UE a pris une série de mesures à travers le Règlement Général sur la Protection des Données, tandis que l'interprétabilité des algorithmes de Machine Learning suscite des inquiétudes croissantes.

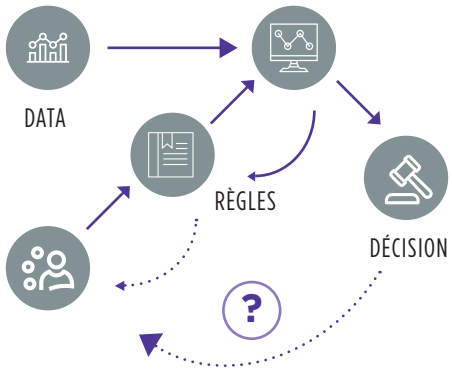
Par ailleurs, l'interprétabilité peut aider les entreprises à tirer parti des nouveaux outils d'IA pour mieux comprendre leurs processus décisionnels.

Prise de décision automatisée basée sur un processus déterministe



L'HOMME DÉTERMINE LES RÈGLES ET PEUT DONC EXPLIQUER LA DÉCISION QUI A ÉTÉ PRISE

Prise de décision basée sur le Machine Learning



LES OUTILS D'INTERPRÉTABILITÉ PERMETTENT D'EXPLIQUER LES RÈGLES OU LA DÉCISION DANS DES TERMES COMPRÉHENSIBLES PAR L'HOMME

IA : APRÈS LA BOITE NOIRE, L'INTERPRÉTABILITÉ ?

La principale différence entre l'IA des années 1970 et celle d'aujourd'hui est qu'il ne s'agit plus d'une approche purement déterministe. Les règles de décision et les équations mathématiques étaient autrefois définies par l'homme et implémentées pour être traitées automatiquement. Aujourd'hui, grâce aux algorithmes de Machine Learning, ces règles peuvent directement être définies à partir des données. La prise de décision automatisée a ainsi perdu son interprétabilité inhérente, car les règles peuvent facilement être masquées par la complexité des algorithmes ou la multiplicité des entrées.

L'interprétabilité, i.e. la capacité d'expliquer ou de présenter des informations dans des termes humainement compréhensibles, a été définie comme une exigence minimale pour certains processus automatisés.

Depuis 2018, le Règlement Général de l'Union européenne sur la Protection des Données (RGPD) exige que toute décision importante ou de nature juridique puisse être expliquée. La personne concernée peut demander une intervention humaine pour contester la décision prise. Il existe d'autres réglementations dans des domaines plus spécifiques, telles que le *Code of Federal Regulations* aux États-Unis, qui stipule que chaque décision liée à un crédit sous-tend un droit d'explication bien établi.

Sur le plan scientifique, le CNRS étudie les limites et les conséquences de telles réglementations et s'emploie à proposer des solutions. Aux États-Unis, la DARPA (*Defense Advanced Research Projects Agency*) dispose d'un budget annuel de 400 millions de dollars pour le projet *Next Generation AI*, qui inclut le projet XAI (*eXplainable AI*) visant à améliorer l'interprétabilité du Machine Learning.

INTERPRETABILITÉ OU EXPLICABILITÉ ?

L'interprétabilité désigne généralement la capacité d'expliquer ou de présenter une information dans des termes humainement compréhensibles. Cela dit, une autre définition peut être donnée en différenciant l'interprétation de l'explication, afin d'établir la différence entre ces deux échelles d'observation. L'interprétation répond à la question « comment » un algorithme prend-il une décision , tandis que l'explication vise à savoir « pourquoi » il a pris une telle décision.

En d'autres termes, « l'interprétation » désigne l'évaluation globale d'un processus de prise de décision. Elle vise à représenter l'importance relative de chaque variable. De son côté, « l'explication » fournit des informations ciblées sur les variables qui ont été déterminantes pour une décision particulière.

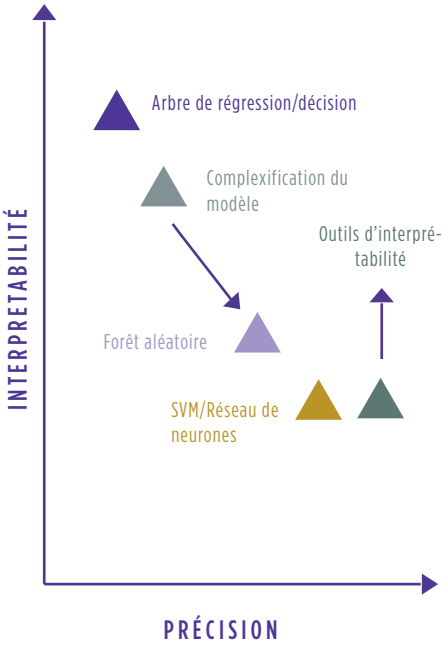
RGPD, Raison 71

« La personne concernée devrait avoir le droit de ne pas faire l'objet d'une décision, qui peut comprendre une mesure, impliquant l'évaluation de certains aspects personnels la concernant, qui est prise sur le seul fondement d'un traitement automatisé et qui **produit des effets juridiques la concernant** ou qui, de façon similaire, **l'affecte de manière significative**, tels que le rejet automatique d'une demande de crédit en ligne ou des pratiques de recrutement en ligne sans aucune intervention humaine. [...] En tout état de cause, un traitement de ce type devrait être assorti de garanties appropriées, qui devraient comprendre une information spécifique de la personne concernée ainsi que **le droit d'obtenir une intervention humaine**, d'exprimer son point de vue, **d'obtenir une explication quant à la décision prise** à l'issue de ce type d'évaluation et de contester la décision. »

FAUT-IL PRIVILÉGIER L'INTERPRÉTABILITÉ À LA PRÉCISION ?

En matière de méthodes et d'algorithmes de Machine Learning, les niveaux d'interprétabilité peuvent varier considérablement. Certaines méthodes sont plus adaptées à l'homme, car facilement interprétables. D'autres sont trop complexes pour être appréhendées et nécessitent donc des méthodes *ad hoc* pour obtenir une interprétation.

Tendance générale du compromis précision/interprétabilité



Le nombre important de variables et la forte dimensionnalité liés à l'utilisation du Big Data complexifient davantage la compréhension de la méthode décisionnelle. Un arbre de décision, par exemple, constitue une séquence de décisions afin de scinder les données. Ces décisions sont faciles à comprendre si la séquence n'est pas trop longue. La forêt aléatoire, quant à elle, forme un ensemble d'arbres de décision. L'homme n'a donc pas la capacité de visualiser chaque séquence. Le Deep Learning, qui s'appuie sur un réseau de neurones artificiels, constitue un très grand nombre de liens par addition et multiplication avec des éléments non linéaires. Il est ainsi difficile d'identifier quels calculs sont pertinents.

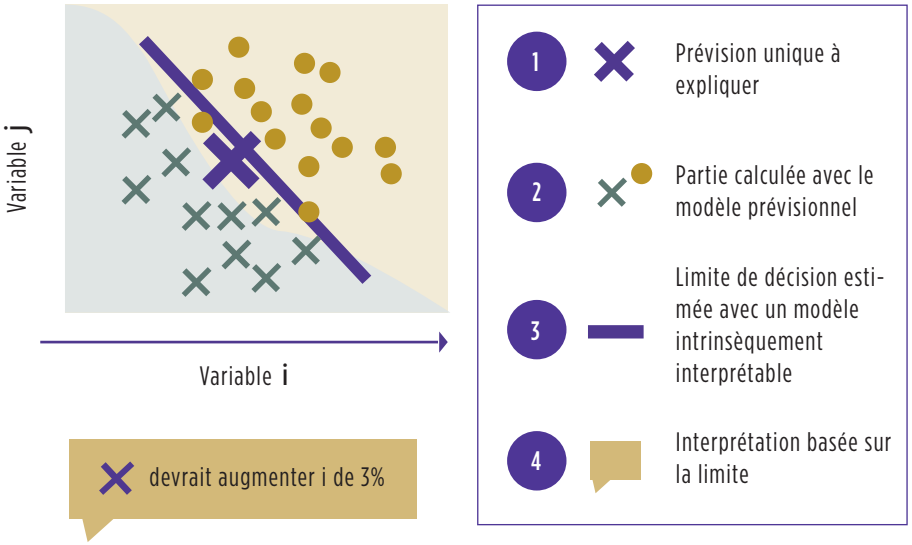
Pourquoi les modèles non interprétables sont-ils donc si peu utilisés ? Parce que la complexité introduite dans les modèles de Machine Learning a permis d'améliorer les performances dans la plupart des domaines. La recherche de précision a ainsi pris le pas sur l'interprétabilité dans certains cas.

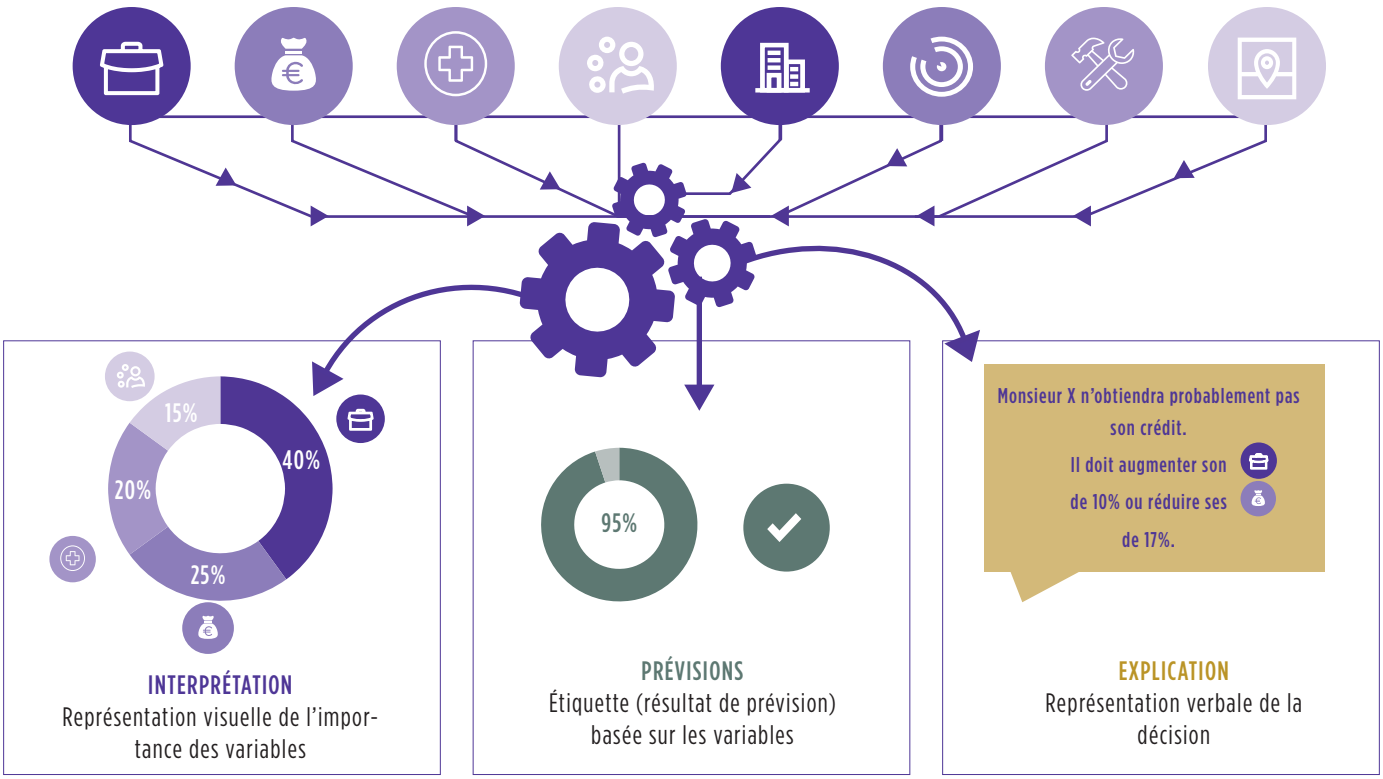
FAUT-IL CHOISIR ENTRE STABILITÉ DES RÉSULTATS, TEMPS DE CALCUL ET SPÉCIALISATION DE L'ALGORITHME ?

Même pour les algorithmes de haute précision, il existe des méthodes permettant d'obtenir une interprétation ou une explication. Cependant, aucune ne peut être appliquée de façon sûre à tous les modèles de Machine Learning avec un résultat stable. Chaque méthode a ses avantages et ses inconvénients.

Un nouveau compromis s'est dessiné entre la stabilité des résultats, le temps de calcul et la spécialisation de l'algorithme. L'importance des variables à partir d'un arbre (« *Tree-Specific Feature Importance* »), par exemple, est une méthode dérivée de la théorie de l'information. Elle évalue dans quelle mesure chaque variable sépare les données pour prendre la décision et produire ainsi une interprétation globale. Rapide à calculer, elle est toutefois instable. Elle ne fonctionne qu'avec les arbres de décision, et donc la forêt aléatoire. La « méthode SHAP », quant à elle, est plus longue à calculer mais très stable. Basée sur la théorie des jeux, elle évalue comment chaque groupe de variables affecte les résultats, puis donne une interprétation et une explication pour chaque modèle. Autre exemple, la méthode « LIME » permet de réaliser une prévision unique avec un modèle géométrique simple, hautement interprétable. Elle est rapide à calculer. Cependant, le domaine d'une prévision unique est mal défini et le résultat peut donc varier considérablement.

Méthode «Lime»





SIMPLE COMME « INTERPRÉTABILITÉ » ?

La première étape consiste à déterminer le besoin réel d'interprétabilité. Par exemple, les analystes quantitatifs sont légalement tenus d'utiliser des modèles hautement interprétables adaptés à une gamme spécifique d'algorithmes qui sont eux-mêmes facilement interprétables. À l'inverse, le tri postal n'aura aucun effet juridique ni significatif sur la personne concernée. Les réseaux complexes de neurones profonds exécutent ainsi la tâche décisionnelle. Dans tous les cas, le besoin d'interprétabilité est variable et doit être évalué. Si l'on prend l'exemple de Wavestone, la moitié des projets récemment menés ont nécessité une interprétabilité d'après le RGPD.

Ensuite, la méthode d'interprétation doit être choisie en fonction du résultat souhaité (interprétabilité ou explicabilité) et de l'implémentation actuelle de l'algorithme de Machine Learning.

Une fois la méthode choisie, il est important de l'appliquer avec précaution, principalement en raison du manque de fiabilité de certaines d'entre elles. Enfin, l'interprétation et l'explication peuvent être représentées de nombreuses manières. Certains préfèrent s'appuyer sur un tableau pour faciliter le traitement automatisé, d'autres préfèreront qu'un message soit directement envoyé au client ou voudront obtenir une représentation visuelle.

Cependant, l'interprétation et l'explication se basent souvent sur un certain ensemble de données et sur une partie spécifique de l'espace de données, ce qui augmente le risque d'interprétation erronée. Certaines méthodes d'interprétation omettent des corrélations entre les variables ou n'offrent qu'une seule explication contrefactuelle lorsque plusieurs auraient pu être données. Malgré ces limites, les outils sont suffisamment puissants pour offrir une interprétabilité conforme au RGPD.

L'INTERPRÉTABILITÉ : LE DÉFI 2020 DES DSI

Indépendamment du RGPD et de son application, l'interprétation d'un algorithme d'apprentissage complexe permet également d'optimiser le modèle dans son ensemble. Dans le cas d'une demande de crédit, le fait de prédéterminer les variables importantes permet de suivre directement les clients qui correspondent, et ainsi d'optimiser les coûts d'acquisition et de marketing.

L'analyse de l'exigence d'interprétabilité et l'optimisation du problème par le biais de méthodes d'interprétabilité constitueront bientôt une étape essentielle de tout cas d'utilisation du Machine Learning.