Loyauté des décisions algorithmiques
L'IA du quotidien peut-elle être éthique ?

Webinaire INTER-FAIR, Université Paris-Dauphine

Aurélien Garivier
suite à des travaux de et avec Philippe Besse (INSA Toulouse) Céline Castets-Renard (UT1) et Jean-Michel Loubes (UT3)

10 décembre 2020

UMPA
ENS DE LYON

LIP

Plan de l'exposé

1. Nouveaux succès, nouvelles questions
2. La généralisation des règles de décision automatiques apprises
3. La fiabilité en question
4. Explicabilité et interprétabilité
5. Décision automatique et discriminations

Nouveaux succès, nouvelles questions

Des succès spectaculaires

- Reconnaissance d'image

Traitement des langues naturelles

Combinaison des deux

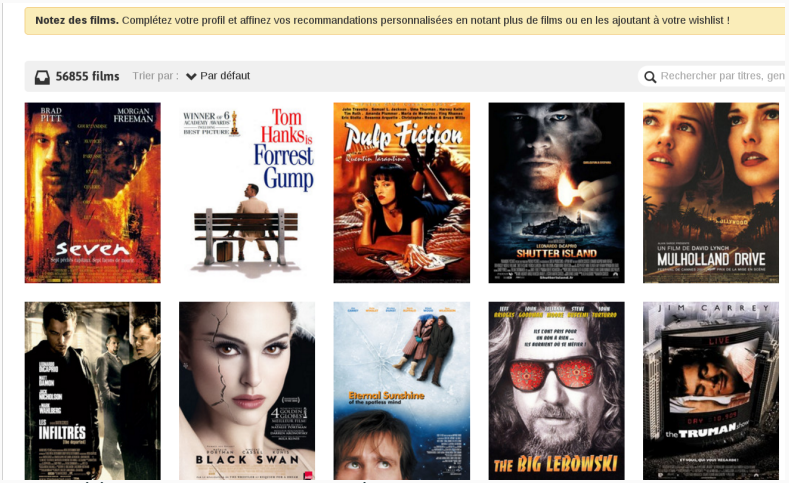
https://link.springer.com/article/10.1007

- Résolution de jeux (stratégie)

AlphaGo

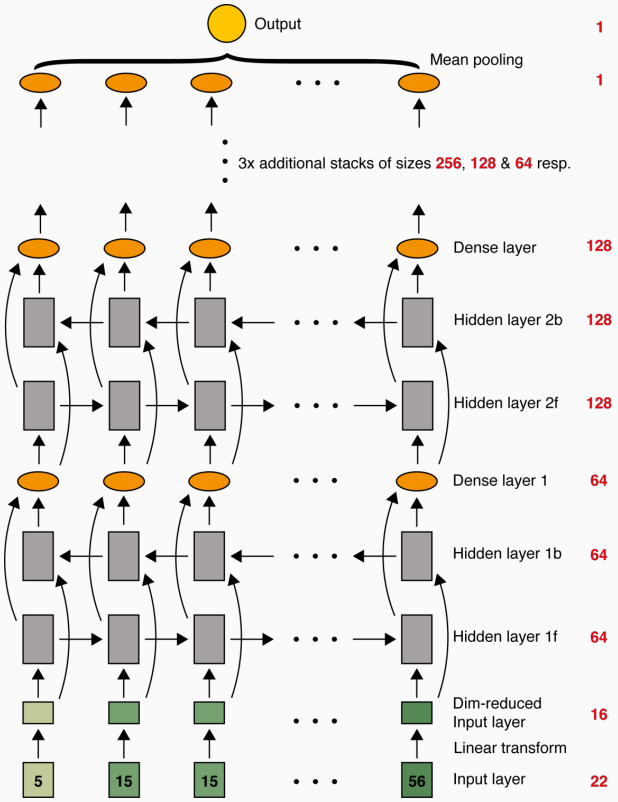
- Véhicules autonomes (combine les deux précédents)

- Systèmes massifs de recommandations : articles, films, produits publicités, etc.



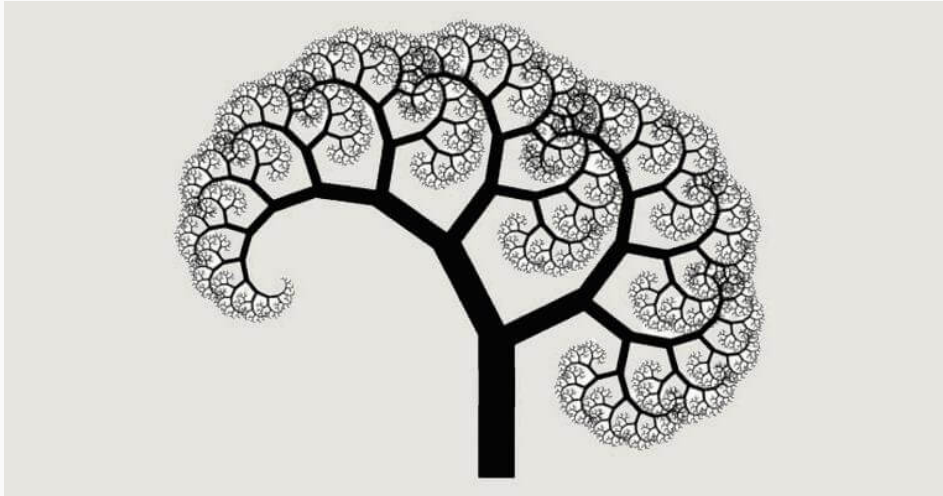
http://www.vodkaster.com/

À quoi ressemblent les algorithmes qui marchent ?



mais aussi

src : https://www.techleer.com/



LightGBM, Light Gradient Boosting Machine

LightGBM is a gradient boosting framework that uses tree based learning algorithms.

combinés et mélangés entre eux...

$\Longrightarrow$ usines à gaz

De la statistique à la science des données

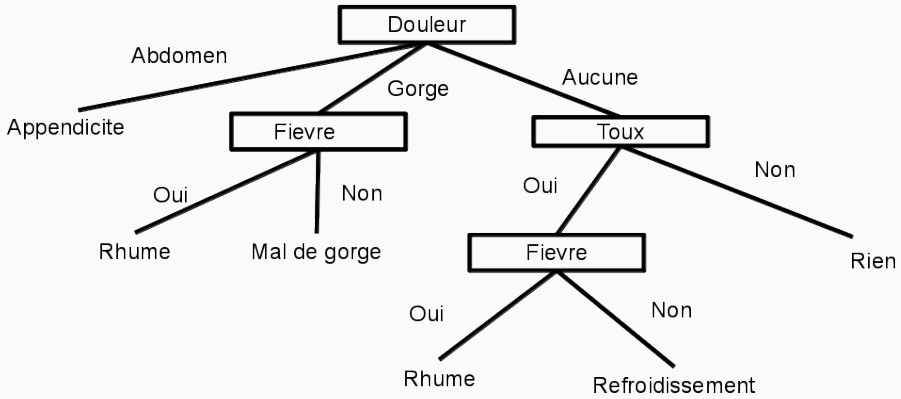
Approche statistique classique : par exemple

Modèle linéaire :

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j} + \epsilon_i$$

β_j = influence du facteur j sur la réponse

Arbres de décision :



http://www.up2.fr/

$\Longrightarrow$ utilisation des données pour trouver la bonne règle de décision parmi un ensemble de fonctions simples

La gouvernance par les nombres est ancienne, mais avec boosting et deep learning sur les données massives la performance moyenne en prédiction est grandement améliorée. Prix : fonctionnement en boîtes noires.

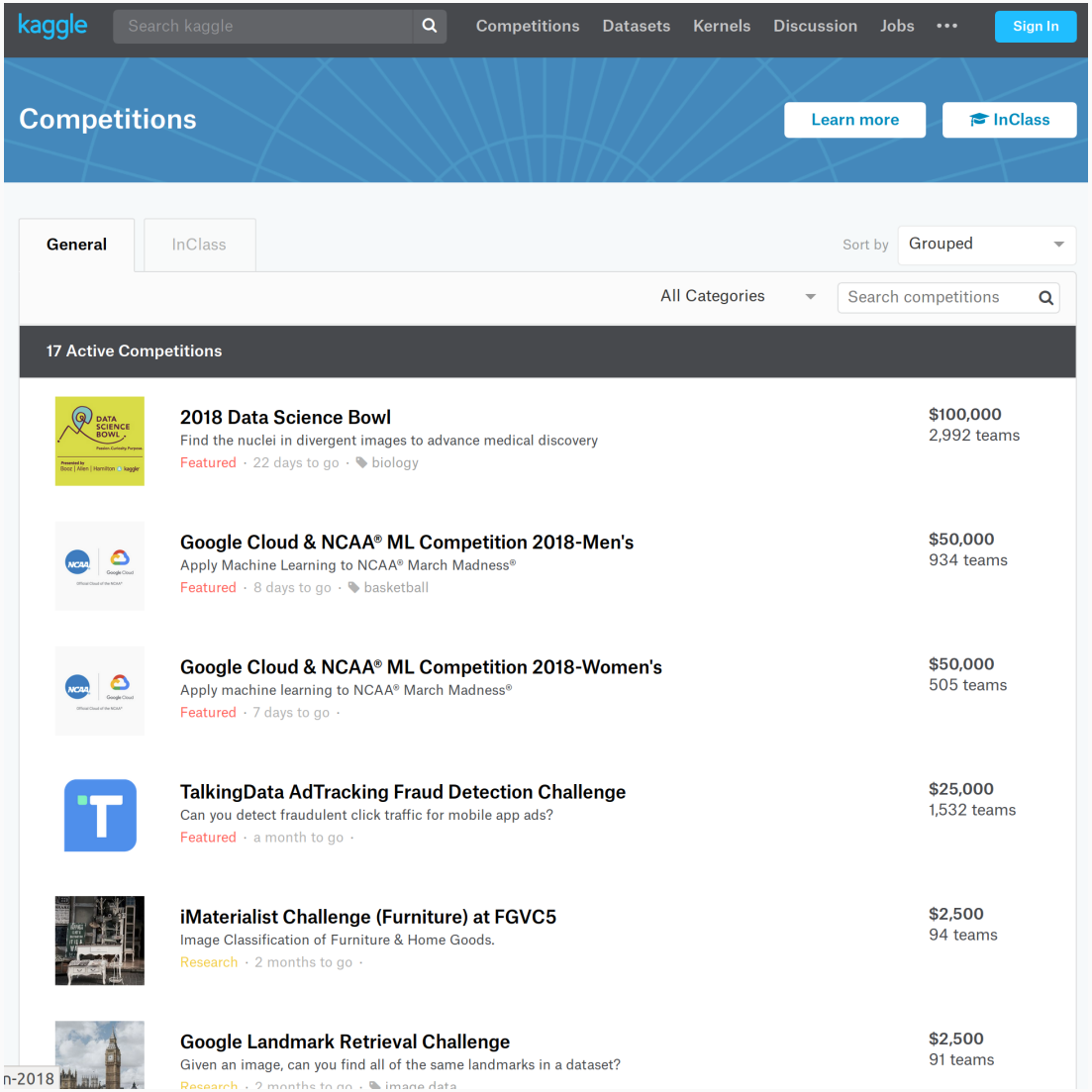
Pourquoi en est-on là ?

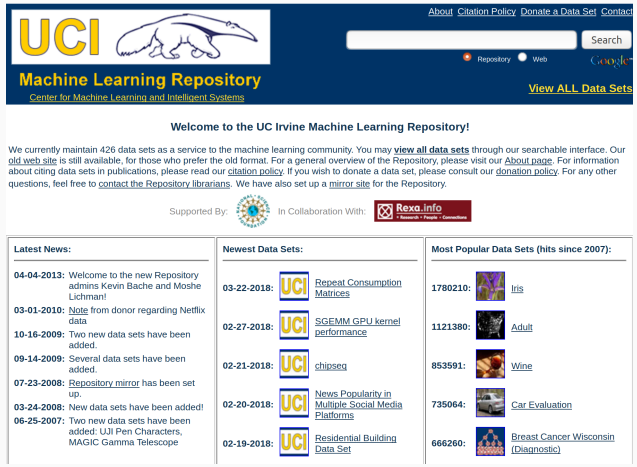
- Prise de conscience de la valeur qu'on peut tirer des données auparavant négligées (phénomène Big Data)
- Communication très efficaces des géants du numérique

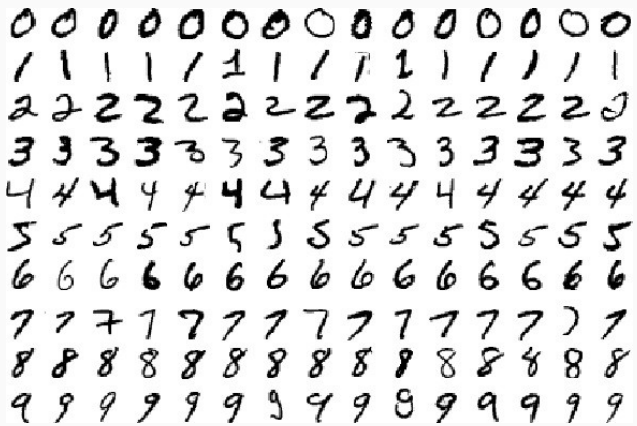
$\Longrightarrow$ Effort de recherche (théorique et appliquée) considérable

- largement porté par la recherche publique
- cf. conférences NIPS, ICML, COLT, ALT, IJCAI, etc.

Benchmarks et Challenges







Orientation de l'effort de recherche

- *Gamification* : fonctionnement par challenges et benchmarks
  - grande efficacité pour dynamiser la recherche
  - fait émerger et diffuser les bonnes idées
  - rend la recherche attrayante et attractive pour les jeunes
- Effort largement orienté vers la pure amélioration de la performance moyenne de systèmes complètement autonomes...
- ... au détriment des autres qualités que l'on pourrait attendre,
- que la statistique classique prenait plus en compte.

Angles morts et questions

- Concurrence, implications épistémologiques, etc : pas évoquées ici
- Généralisation du recueil et de l'utilisation des données pour exploitation par des systèmes de décision automatiques
- Fiabilité, robustesse : nécessité d'une meilleure maîtrise théorique
- Explicabilité et interprétabilité
- Discrimination

La généralisation des règles de décision automatiques apprises

Acceptabilité sociétale des décisions automatiques

- La multiplication des systèmes de décisions automatique pose plus vivement la question de son acceptabilité sociétale.
- Le fait que des firmes commerciales aux pratiques très agressives soient motrices (et les régulateurs souvent à la traîne) suscite de légitimes craintes
- Communication massive sur les "big data" puis "l'intelligence artificielles" $\Longrightarrow$ promesses importantes mais aussi méfiance accrue
- Cf. débat sur APB / ParcourSup (qui pourtant n'implique ni IA, ni big data, ni apprentissage machine)
- Demande de transparence des algorithmes.

Accès aux données privées

Scandale *Cambridge Analytica*

Plainte de David Carroll qui demande l'accès à ses données personnelles Cambridge Analytica (qui travaillait pour Ted Cruz) se vante de vous connaître :

- 10 likes = niveau "collègue de bureau"
- 300 likes = niveau conjoint(e)

$\Longrightarrow$ profilage précis sur des questions comme l'orientation politique, les valeurs, etc.

David Carroll confirme que ça marche... et demande l'accès à *toutes* les données que la firme a sur lui (aspirées sur Facebook ?).

Loyauté des décisions algorithmique

- *trustworthiness :* véracité, crédibilité, pour mériter la confiance,
- *accountability :* responsabilité des décisions, capacité à en rendre compte.

Même statistique ou probabiliste, la décision doit pouvoir être attribuée à un humain qui en assume la responsabilité.

- Elle doit être la plus juste au sens de l'intérêt de la personne concernée et / ou globalement de la communauté ; donc issue d'une meilleure prévision.
- Il faut pouvoir en rendre compte et donc, pouvoir l'expliquer de façon compréhensible (e.g. médecin à son patient).
- Enfin, elle doit éviter tout biais discriminatoire vis-à-vis de minorités et groupes sensibles protégés par la loi.

Justesse (qualité)

Les méthodes de prévisions sont estimées sur les données d'apprentissage, c'est donc la qualité de celle-ci qui est en premier lieu déterminante — garbage in garbage out.

Exemple de Google Flu Trend (2008) qui visait à suivre en temps réel et prédire le déroulement d'une épidémie de grippe à partir du nombre de recherches de certains mots clefs et connaissant la localisation (adresse IP) du questionneur. L'outil a été abandonné par Google (2015) car source de lourdes erreurs de prévision. C'était le battage médiatique de la grippe qui était suivi, pas l'épidémie elle-même. Les données ont été reprises avec de meilleurs résultats par une équipe de Boston (Yanga et al ; 2015) en estimant un modèle autorégressif intégrant une chaîne de Markov cachée et corrigée sur la base des tendances des recherches sur Google.

Au niveau juridique

- Grosses disparités entre Europe et (notamment) USA ! (D. Carroll s'appuie sur le droit anglais, aux USA il ne pourrait pas demander cet accès)
- Adoption en avril 2016 du règlement 2016/679/UE relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données qui entrera en vigueur en mai 2018, RGPD = Règlement Européen sur la Protection des Données
- Le législateur français a consacré des dispositions plus spécifiques sur ces questions dans la loi 1321 "Pour Une République Numérique" du 7 octobre 2016.
- La CNIL, chargée de conduire une réflexion sur les enjeux éthiques soulevés par l'évolution des technologies numériques, a ouvert un débat public sur le thème : Éthique et Numérique.

RGPD et LRN : Accent sur la décision automatique

Article 10 loi 78-17 du 6 janvier 1978

Aucune décision produisant des effets juridiques à l'égard d'une personne ne peut être prise sur le seul fondement d'un traitement automatisé de données destiné à définir le profil de l'intéressé ou à évaluer certains aspects de sa personnalité " Cette règle est donc peu respectée et sa violation ne donne pas lieu à sanction.

Article 22§1 du règlement 2016/679/EU (RGDP)

La personne concernée a le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, y compris le profilage, produisant des effets juridiques la concernant ou l'affectant de manière significative.

$\Longrightarrow$ On passe donc d'une interdiction à la reconnaissance d'un véritable droit subjectif de la personne de ne pas faire l'objet d'une décision automatisée, dès lors qu'elle serait négative.

$\Longrightarrow$ Au moins le droit d'obtenir une intervention humaine de la part du responsable du traitement.

Droits consacrés

- (RGDP, art. 15h) Droit d'être informé de l'existence d'une prise de décision automatisée ;
- (RGDP, art. 22§1) Droit de ne pas faire l'objet d'un traitement automatisé produisant des effets juridiques ou affectant la personne concernée de manière significative ;
- (RGDP, art. 22§3) Droit d'obtenir une intervention humaine de la part du responsable du traitement ;
- (RGDP, art. 22§3) Le droit d'exprimer son point de vue et de contester la décision.

Article 22§2 : exceptions

Des exceptions lorsque la décision :

- est nécessaire à la conclusion ou à l'exécution d'un contrat entre la personne concernée et un responsable du traitement ;
- est autorisée par le droit de l'Union ou le droit de l'État membre auquel le responsable du traitement est soumis et qui prévoit également des mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée ;
- est fondée sur le consentement explicite de la personne concernée.

Au total

Les exceptions appauvrissent substantiellement la règle :

- les activités économiques du numériques peuvent se prévaloir d'un fondement contractuel : l'utilisation des services (sites de e-commerce ou plateformes de mise en relation telles celles des réseaux sociaux) est de fait considérée comme une acceptation des conditions générales d'utilisation et manifestant l'acceptation de l'offre contractuelle ;
- paragraphe 2.c : consentement explicite MAIS est-il éclairé ? Enjeux pas faciles à saisir

En outre, le règlement général sur la protection des données ne concerne pas directement ni véritablement indirectement le principe de transparence algorithmique.

Transparence algorithmique en revanche consacrée dans la LRN

La loi pose deux catégories de règles :

- pour les plateformes numériques,
- pour les administrations

Article 49 de la LRN, codifié à l'article L. 111-7. I. du Code de la consommation

Est qualifiée d'opérateur de plateforme en ligne toute personne physique ou morale proposant, à titre professionnel, de manière rémunérée ou non, un service de communication au public en ligne reposant sur :

1. Le classement ou le référencement, au moyen d'algorithmes informatiques, de contenus, de biens ou de services proposés ou mis en ligne par des tiers ;
2. Ou la mise en relation de plusieurs parties en vue de la vente d'un bien, de la fourniture d'un service ou de l'échange ou du partage d'un contenu, d'un bien ou d'un service.

Pour les plateformes : un régime juridique plutôt pauvre

Article L. 111-7 point II

Tout opérateur de plateforme en ligne est tenu de délivrer au consommateur une information loyale, claire et transparente sur :

1. Les conditions générales d'utilisation du service d'intermédiation qu'il propose et sur les modalités de référencement, de classement et de déréférencement des contenus, des biens ou des services auxquels ce service permet d'accéder ;
2. L'existence d'une relation contractuelle, d'un lien capitalistique ou d'une rémunération à son profit, dès lors qu'ils influencent le classement ou le référencement des contenus, des biens ou des services proposés ou mis en ligne ;
3. La qualité de l'annonceur et les droits et obligations des parties en matière civile et fiscale, lorsque des consommateurs sont mis en relation avec des professionnels ou des non-professionnels.

Pour les plateformes

$\Longrightarrow$ La LRN a essentiellement pour objet d'imposer une obligation d'information (loyauté) sur les modalités de référencement des algorithmes, laquelle s'ajoute aux autres obligations d'information du code de la consommation.

$\Longrightarrow$ cette obligation est utilement complétée par les dispositions préexistantes dans le code de consommation relatives aux pratiques commerciales trompeuses dont les énoncés sont suffisamment larges pour viser et sanctionner les comportements déviants qui pourraient être fondés sur des traitements algorithmiques déloyaux ou faussés.

$\Longrightarrow$ Enfin, la loi pour une république numérique ajoute à l'article 50 que "Les opérateurs de plateformes en ligne dont l'activité dépasse un seuil de nombre de connexions défini par décret élaborent et diffusent aux consommateurs des bonnes pratiques visant à renforcer les obligations de clarté, de transparence et de loyauté".

Pour les administrations

Article R. 311-3-1-2 du code des relations entre le public et l'administration (CRPA), décret 2017-330

L'administration communique à la personne faisant l'objet d'une décision individuelle prise sur le fondement d'un traitement algorithmique, à la demande de celle-ci, sous une forme intelligible et sous réserve de ne pas porter atteinte à des secrets protégés par la loi, les informations suivantes :

- Le degré et le mode de contribution du traitement algorithmique à la prise de décision ;
- Les données traitées et leurs sources ;
- Les paramètres de traitement et, le cas échéant, leur pondération, appliqués à la situation de l'intéressé ;
- Les opérations effectuées par le traitement.

Pour toutes les administrations (y compris territoriales) "sous réserve de ne pas porter atteinte à des secrets protégés par la loi" (et qq exceptions).

Confidentialité différentielle

Il est possible d'apprendre sans compromettre la confidentialité des données avec lesquelles on apprend !

Formalisation mathématique : *differential privacy*

$$x \sim x' \implies \mathbb{P}\big(A(x) \in S\big) \leq \exp(\epsilon)\mathbb{P}\big(A(x') \in S\big)$$

"La loi des sorties de l'algorithmes n'est guère modifiée si on change arbitrairement une des données individuelles"

- constitution actuellement d'une "boîte à outils" d'apprentissage confidentiel
- certaines entreprises vendent ce service avec les garanties *mathématiques* de leur confidentialité
- reste à mieux comprendre ce que l'on perd en termes de précision

La fiabilité en question

Robustesse au changement

- Paradigme classique de l'apprentissage = environnement stable (données iid)
- MAIS souvent l'environnement évolue
  - progressivement
  - ou brutalement
- $\Longrightarrow$ intégrer des systèmes capables
  - d'apprendre la durée de pertinence des données,
  - de détecter des ruptures dans les propriétés statistiques de celles-ci (détection d'anomalies)
  - de raffiner progressivement sa connaissance au fur et à mesure que les données arrivent, tout en oubliant ce qui doit l'être

Robustesse aux données aberrantes

- Objectif = algorithmes d'apprentissage pouvant utiliser des données de qualité variable.
- Notion statistique de robustesse (au sens de Huber).
- Adaptation aux algorithmes de machine learning.

Exemples :

- MONK – Outlier-Robust Mean Embedding Estimation by Median-of-Means ,M. Lerasle, Z. Szabo, G. Lecué, E. Moulines, Gaspar Massiot, (2018), arXiv :1802.04784
- Robust machine learning by median-of-means : theory and practice, M. Lerasle, G. Lecué, (2017), arXiv :1711.10306

Robustesse aux attaques

Problèmes de sécurité généraux des systèmes informatiques mais aussi des risques spécifiques :

- *Attaques par sortie du domaine d'apprentissage* : déstabiliser l'algorithme en présentant des données éloignées des habituelles afin que son comportement soit instable voire prévisiblement mauvais ;
- *Attaques par injection de bruit malicieux* : perturber habilement des données habituelles afin de modifier la réponse du système.

Src : Universal adversarial perturbations

Moosavi-Dezfooli et al., CVPR 2017

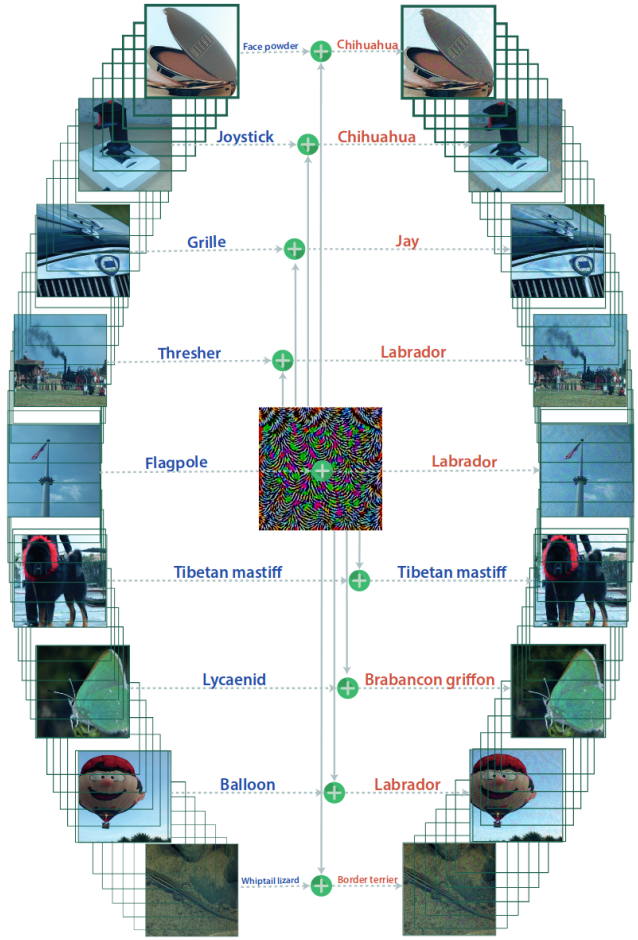
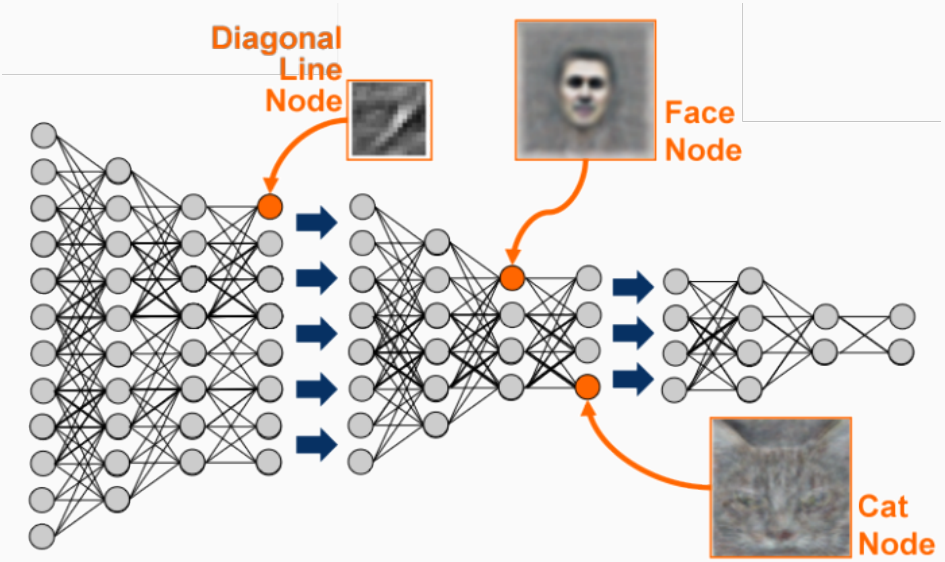


Figure 1: When added to a natural image, a universal perturbation image causes the image to be misclassified by the deep neural network with high probability. *Left images:* Original natural images. The labels are shown on top of each arrow. *Central image:* Universal perturbation. *Right*

Des algorithmes non maîtrisés

- Boîtes noires offrant pour seule garantie la performance passée sur des exemples fixés ;
- Peu de garanties structurelles ;
- Peu de maîtrise du risque, souvent difficile à quantifier ;
- Besoin de travaux théoriques pour :
  - Mieux comprendre les propriétés fines des algorithmes (aidera aussi à calibrer les réseaux) ;
  - Savoir auditer "de l'extérieur" un système d'apprentissage par rapport à des contextes potentiellement changeants.



http://aiehive.com

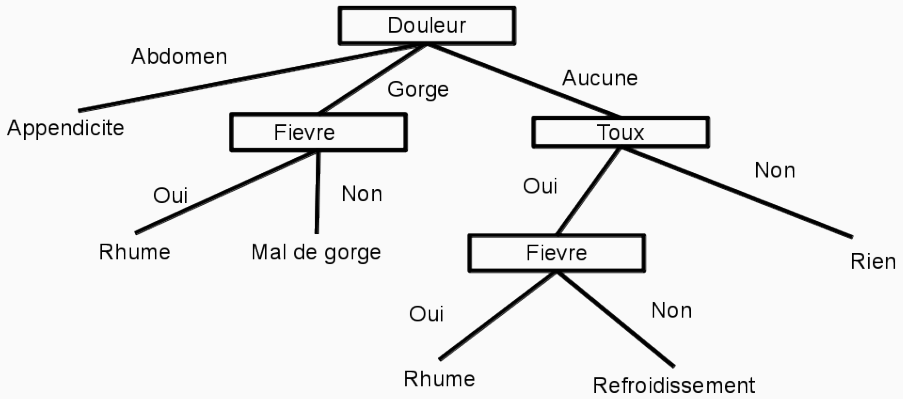
Explicabilité et interprétabilité

Explicabilité versus interprétabilité

Deux notions à distinguer (mais vocabulaire trompeur : on suit ici https://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-159.pdf).

Une règle de décision est dite

interprétable si on comprend comment elle associe une réponse à des observations ; exemple typique : un arbre de décision ;



http://www.up2.fr/

explicable si on comprend sur quels éléments est basée la décision, éventuellement de façon contrefactuelle ; exemple : "si telle variable avait pris telle autre valeur, alors la décision aurait été différente".

Explicabilité versus interprétabilité

Deux notions à distinguer (mais vocabulaire trompeur : on suit ici https://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-159.pdf).

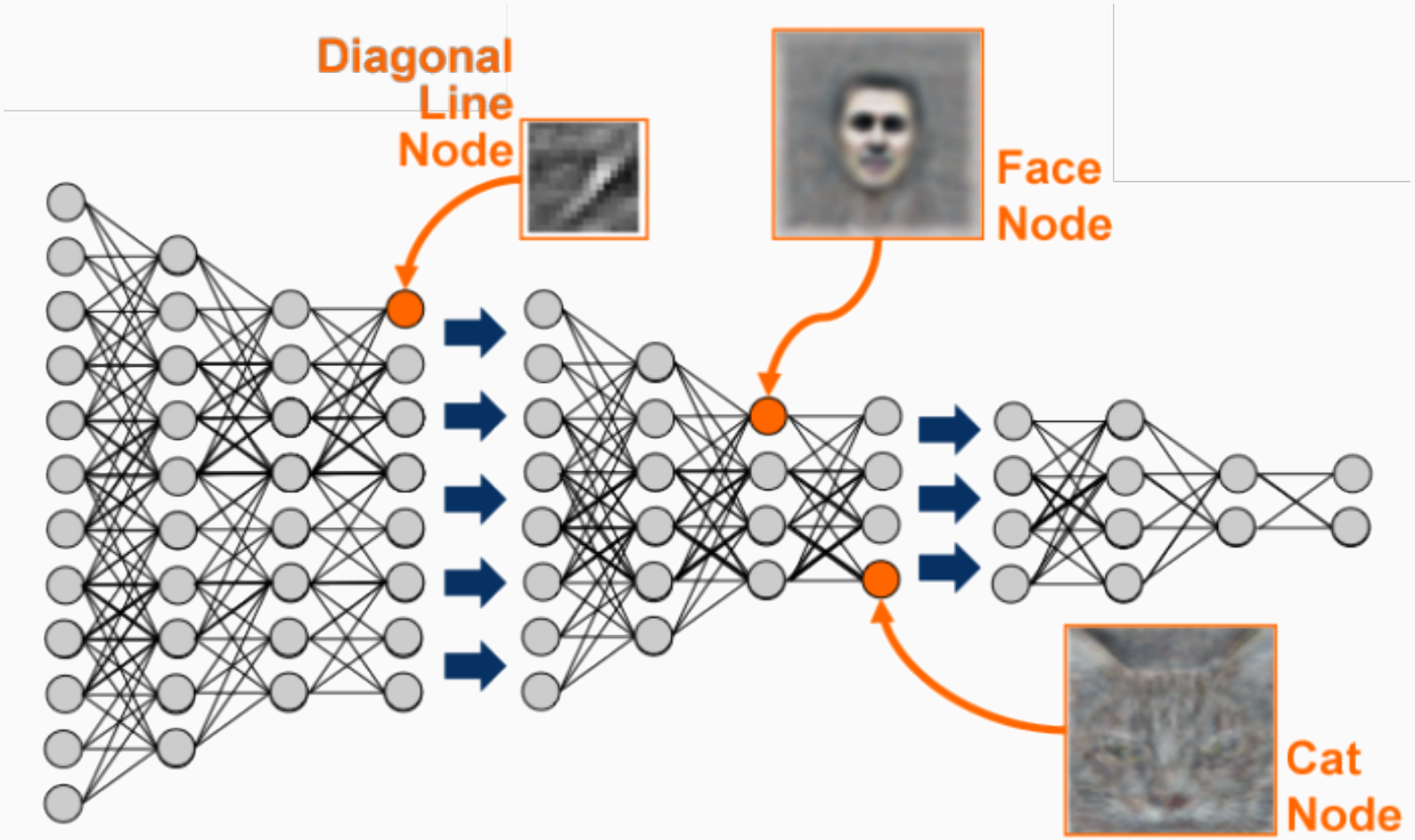
Une règle de décision est dite

interprétable si on comprend comment elle associe une réponse à des observations ; exemple typique : un arbre de décision ;

explicable si on comprend sur quels éléments est basée la décision, éventuellement de façon contrefactuelle ; exemple : "si telle variable avait pris telle autre valeur, alors la décision aurait été différente".

La notion d'explicabilité se rattache à la statistique causale et à l'analyse de sensibilité.

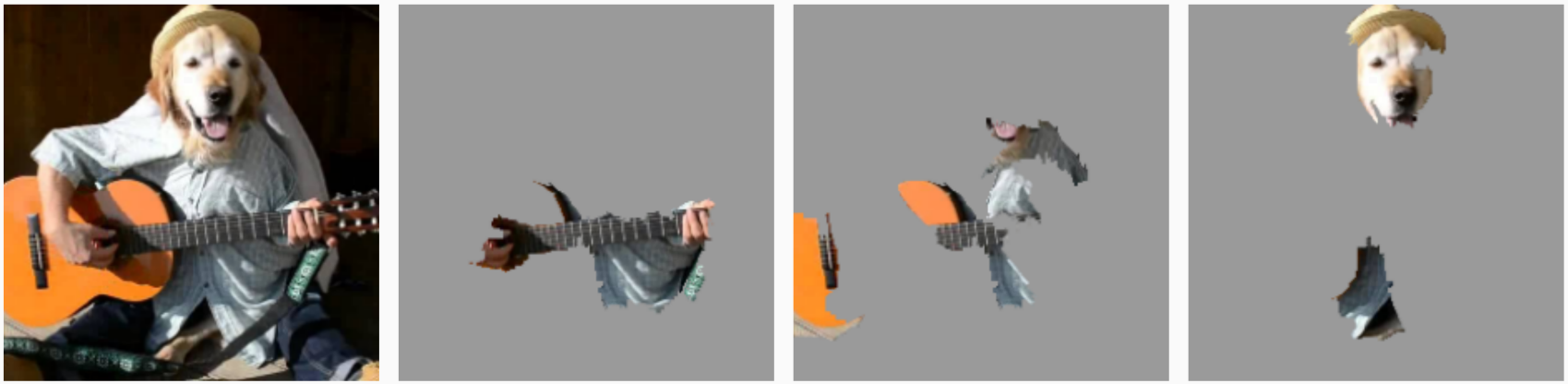
Interpréter un réseau de neurones



http://aiehive.com

Pari scientifique audacieux

Explicabilité : la méthode LIME



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

"Why Should I Trust You ?" Explaining the Predictions of Any Classifier, by Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin.

Explicabilité : la méthode LIME

(a) Husky classified as wolf

(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

"Why Should I Trust You ?" Explaining the Predictions of Any Classifier, by Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin.

Loi pour une République Numérique

- La loi PRN semble pencher vers une exigence d'interprétabilité même si un "droit "subjectif à explication" n'est pas explicitement consacré ;
- Décret d'application : devoir de l'administration d'expliquer précisément ses décisions : expliquer "à la personne faisant l'objet d'une décision individuelle prise sur le fondement d'un traitement algorithmique, à la demande de celle-ci, sous une forme intelligible et sous réserve de ne pas porter atteinte à des secrets protégés par la loi, les informations suivantes :
  1. le degré et le mode de contribution du traitement algorithmique à la prise de décision ;
  2. les données traitées et leurs sources ;
  3. les paramètres de traitement et, le cas échéant, leur pondération, appliqués à la situation de l'intéressé ;
  4. les opérations effectuées par le traitement."

Quelques références

- Interpretable Classification Models for Recidivism Prediction, Zeng J., Ustuny B., Rudin C. (2016). in JRSS A.
- Generating Visual Explanations, Authors : Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell, ECCV 2016
- Datta A., Sen S., Zick Y. (2016). Algorithmic Transparency via Quantitative Input Influence : Theory and Experiments with Learning Systems, in IEEE Symposium on Security and Privacy.

Décision automatique et discriminations

Discriminations : Article 225-1 C. pén.

"Constitue une discrimination toute distinction opérée entre les personnes physiques sur le fondement de leur origine, de leur sexe, de leur situation de famille, de leur grossesse, de leur apparence physique, de la particulière vulnérabilité résultant de leur situation économique, apparente ou connue de son auteur, de leur patronyme, de leur lieu de résidence, de leur état de santé, de leur perte d'autonomie, de leur handicap, de leurs caractéristiques génétiques, de leurs mœurs, de leur orientation sexuelle, de leur identité de genre, de leur âge, de leurs opinions politiques, de leurs activités syndicales, de leur capacité à s'exprimer dans une langue autre que le français, de leur appartenance ou de leur non-appartenance, vraie ou supposée, à une ethnie, une Nation, une prétendue race ou une religion déterminée."

Discriminations : Article 225-1 C. pén.

"Constitue également une discrimination toute distinction opérée entre les personnes morales sur le fondement de l'origine, du sexe, de la situation de famille, de la grossesse, de l'apparence physique, de la particulière vulnérabilité résultant de la situation économique, apparente ou connue de son auteur, du patronyme, du lieu de résidence, de l'état de santé, de la perte d'autonomie, du handicap, des caractéristiques génétiques, des mœurs, de l'orientation sexuelle, de l'identité de genre, de l'âge, des opinions politiques, des activités syndicales, de la capacité à s'exprimer dans une langue autre que le français, de l'appartenance ou de la non-appartenance, vraie ou supposée, à une ethnie, une Nation, une prétendue race ou une religion déterminée des membres ou de certains membres de ces personnes morales."

Discriminations

Article 225-2 C. pén.
"La discrimination définie aux articles 225-1 à 225-1-2, commise à l'égard d'une personne physique ou morale, est punie de trois ans d'emprisonnement et de 45 000 euros d'amende lorsqu'elle consiste :

1. À refuser la fourniture d'un bien ou d'un service ;
2. À entraver l'exercice normal d'une activité économique quelconque ;
3. À refuser d'embaucher, à sanctionner ou à licencier une personne."

Par opposition à discriminatoire, une décision est dite loyale si elle ne se base pas sur l'appartenance d'une personne à une minorité protégée ou la connaissance explicite ou implicite d'une donnée personnelle sensible.

Loyale = pas explicitement basée sur une donnée sensible ?

- Difficile à clarifier ;
- Il ne suffit pas que la variable "sensible" soit inconnue ou supprimée des données d'apprentissage pour que la décision soit sans biais vis-à-vis de ses modalités !
- L'information sensible peut être contenue implicitement, même sans intention de la rechercher, dans les informations non sensibles et ainsi participer au biais de la décision.
- Exemple : des habitudes de consommation, des avis sur les réseaux sociaux, des données de géolocalisation... renseignent sur les orientations de la personne.

Un exemple de controverse justice prédictive

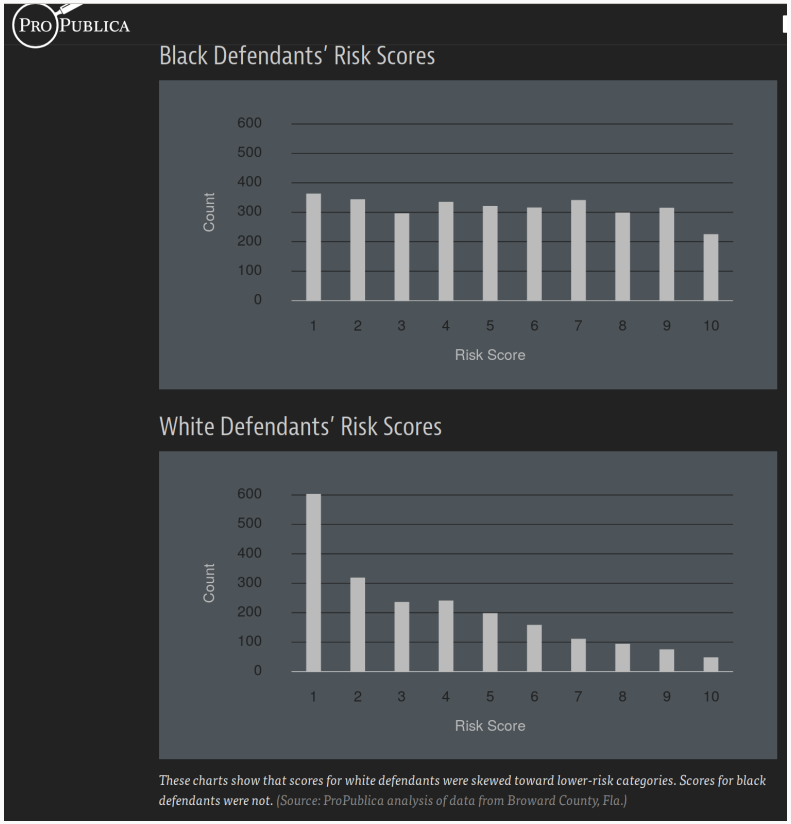
L'ex-société NORTHPOINTE (maintenant EQUIVANT) commercialise l'application COMPAS : Correctional Offender Management Profile for Alternative Sanction) produisant un score de risque de récidive pour les détenus ou accusés lors d'un procès.

- score est estimé sur la base d'un questionnaire détaillé et à partir d'un modèle de durée de vie (modèle de Cox) ;
- qualité de prédiction : assez mauvais (AUC≈ 0.7) ;
- taux d'erreur sur la prévision d'une récidive : autour de 40% pour les blancs comme pour les noirs.

Le site PROPUBLICA (prix Pulitzer 2011) l'accuse d'être discriminatoire.

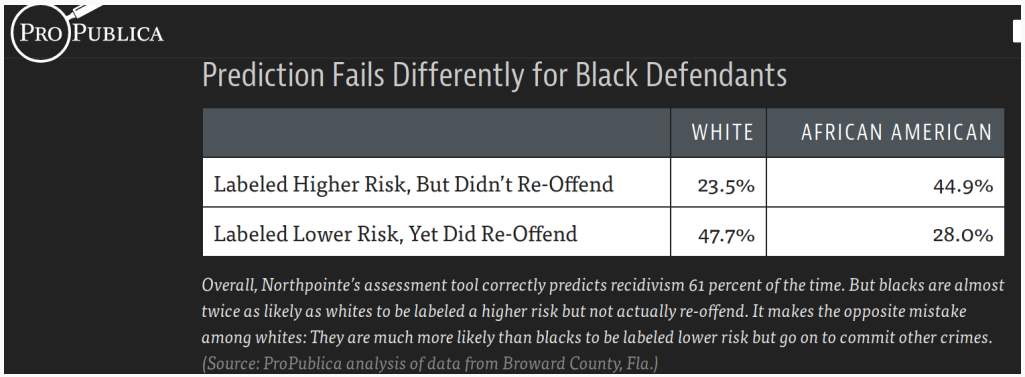
La société Northpointe défend l'impartialité de ce score.

Les accusations de Propublica



https://www.propublica.org

La distribution des scores est différente entre les races.

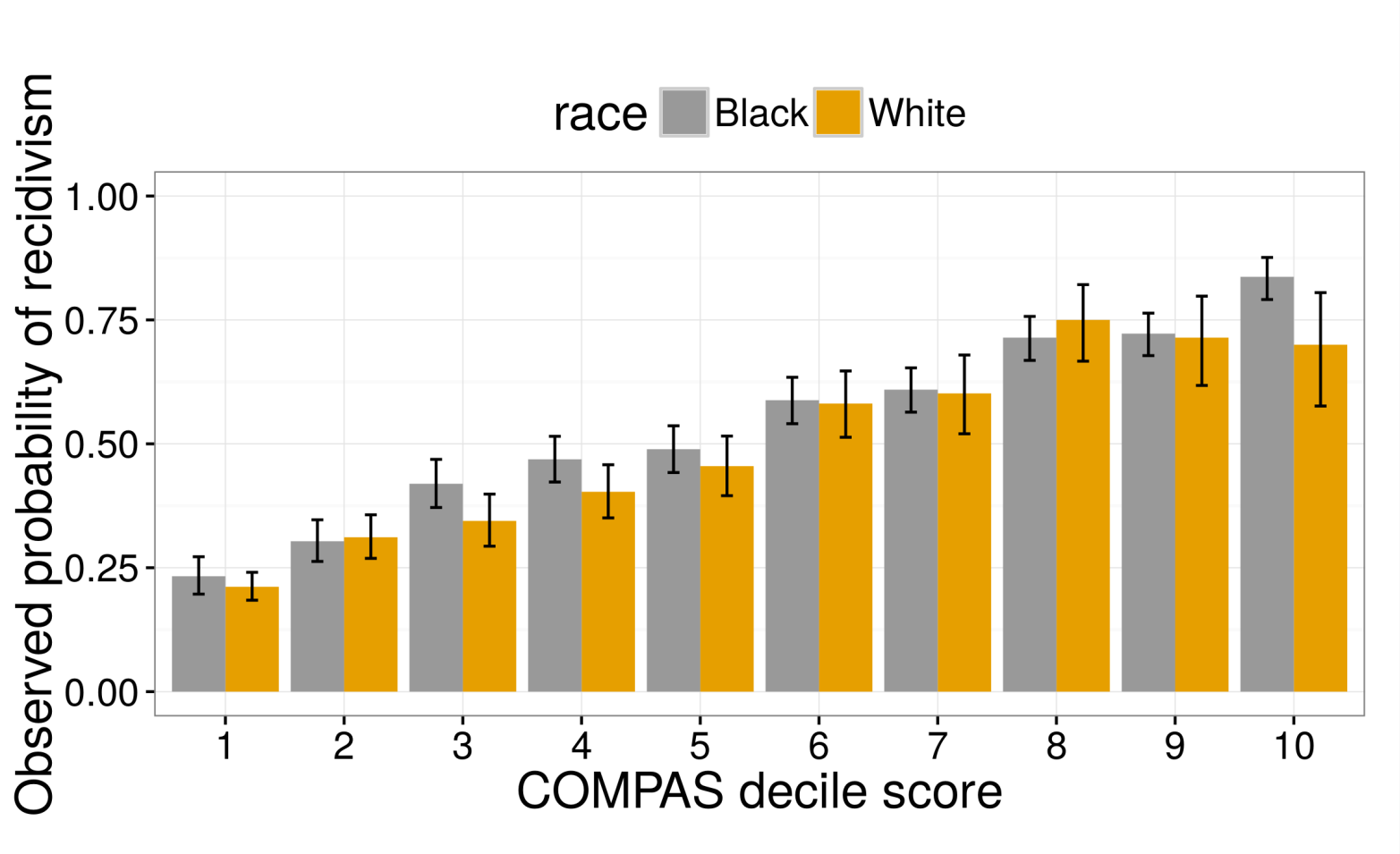


Les erreurs commises ne sont pas les mêmes suivant les races !

Pourtant le score est "fair" (bien calibré)

Chouldechova A. (2016).

Fair prediction with disparate impact : A study of bias in recidivism prediction instruments, FATML 2016 conference.



Pour un score donné, la probabilité de récidive est à peu près la même pour les noirs et pour les blancs.

Un dilemme insoluble

Le taux de récidive des afro-américains est plus élevé.

$\Longrightarrow$ mécaniquement, tout score calibré mène à des taux de faux positifs différents entre noirs et blancs.

Chouldechova A. (2016)

When the recidivism prevalence differs between two groups, a test-fair score cannot have equal false positive and negative rates across those groups.

This observation enables us to better understand why the ProPublica authors observed large discrep- ancies in FPR and FNR between Black and White defendants.

[...] Since the COMPAS RPI approximately satisfies test fairness, we know that some level of im- balance in the error rates must exist.

Solution : pas de décision automatique ?

Chouldechova A. (2016)

In closing, we would like to note that there is a large body of literature showing that **data-driven risk assessment instruments tend to be more accurate than professional human judgements**, and investigating whether human-driven decisions are themselves prone to exhibiting racial bias.

We should not abandon the data-driven approach on the basis of negative headlines.

Rather, we need to work to ensure that the instruments we use are demonstrably free from the kinds of quantifiable biases that could lead to disparate impact in the specific contexts in which they are to be applied.

$\Longrightarrow$ Ce n'est pas le modèle qui est biaisé mais l'échantillon d'apprentissage, reflet des biais sociaux.

$\Longrightarrow$ Risque, dénoncé par O'Neil (2016), de les renforcer !

Voir aussi Fairness Constraints : Mechanisms for Fair Classication, Zafar M., Valera I., Rodriguez M., Gummadi K. (2017), AISTATS

Biais dans les données

`http://www.lemonde.fr/pixels/`

Joy Buolamwini (MIT) a confronté les logiciels de trois entreprises (IBM, Microsoft et Face++) à 1 270 portraits officiels de personnalités politiques (Rwandais, Sénégalais, de Sud-Africains, de Finlandais, d'Islandais et Suédois), afin de déterminer leur genre.

Étude Buolamwini

Résultats globalement bons : 93,7% de taux de réussite pour Microsoft, 90% pour Face++, et 87,9% pour IBM.

MAIS

- Moins bon pour les femmes que pour les hommes, par exemple : Face++ classifie correctement 99,3% des hommes, mais seulement 78,7% des femmes.
- Moins bon pour les peaux foncées que claires : 95% de réponses correctes dans le premier cas, seulement 77,6% pour le second avec le logiciel développé par IBM.
- 93,6% des erreurs faites par Microsoft concernaient ses sujets à la peau foncée, et 95,9% de celles de Face ++ concernaient des femmes !

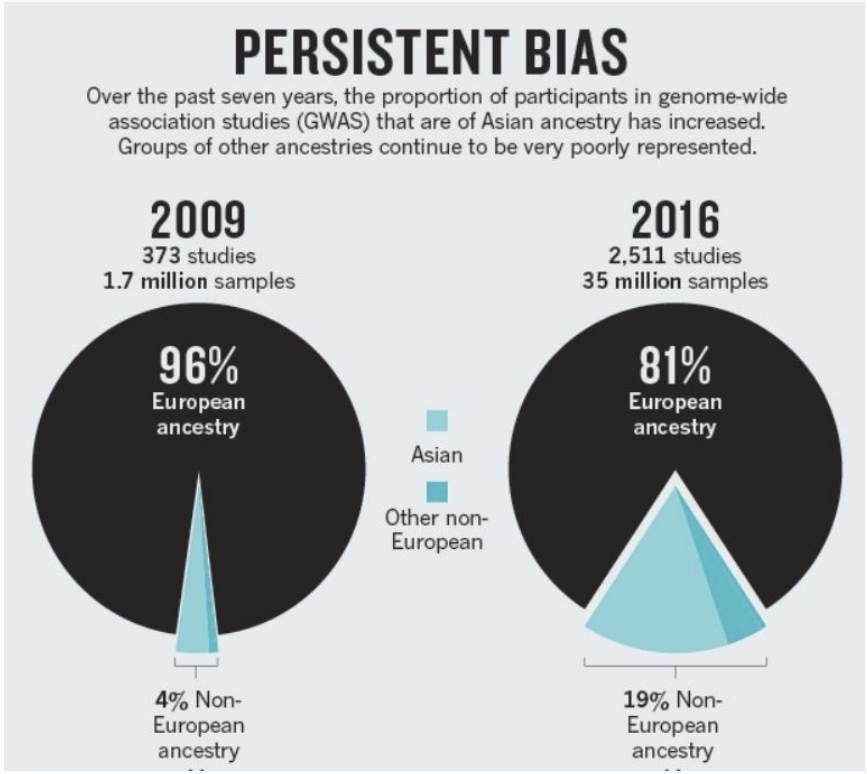
Explication : Biais dans les données !

"Les hommes à la peau claire y sont surreprésentés, et aussi les personnes à la peau claire de manière générale".

Et ça concerne aussi...

- ... les primes d'assurance, offres d'emploi, taux de crédits...

- ... la médecine personnalisée : la grande majorité des études d'association pangénomique ont été faites sur des populations d'ascendance blanche/européenne.
  $\Longrightarrow$ Les facteurs de risque estimés seront potentiellement différents pour un patient d'ascendance africaine ou asiatique !



Popejoy A., Fullerton S. (2016).
Genomics is failing on diversity, Nature 538

Détecter un biais

Détecter une discrimination individuelle : **Testing**

- Idée : modifier seulement une caractéristique "protégée" de l'individu, et regarder si la décision change
- Méthode reconnue par la justice
- Discrimination à l'embauche, bail immobilier, à l'entrée, à l'assurance, etc.

Détecter une discrimination de groupe : trois mesures :

- Disparate Impact (Civil Right Act 1971) : $DI = \frac{\mathbb{P}(\hat{h}_n(X) = 1 | S = 0)}{\mathbb{P}(\hat{h}_n(X) = 1 | S = 1)}$
- Taux d'erreurs conditionnels :
  $\mathbb{P}(\hat{h}_n(X) \neq Y | S = 1) = \mathbb{P}(\hat{h}_n(X) \neq Y | S = 0)$
- Egalité des chances : $\mathbb{P}(\hat{h}_n(X) = 1 | S = 1)$ vs $\mathbb{P}(\hat{h}_n(X) = 1 | S = 0)$

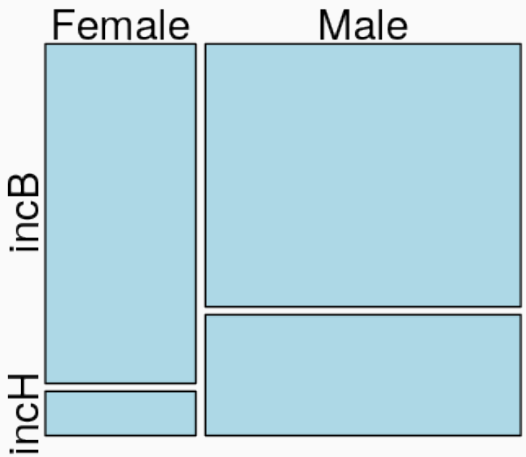
Exemple : Adult Census Dataset voir https://github.com/wikistat

- 48842 US citizens (1994)
- 14 features :

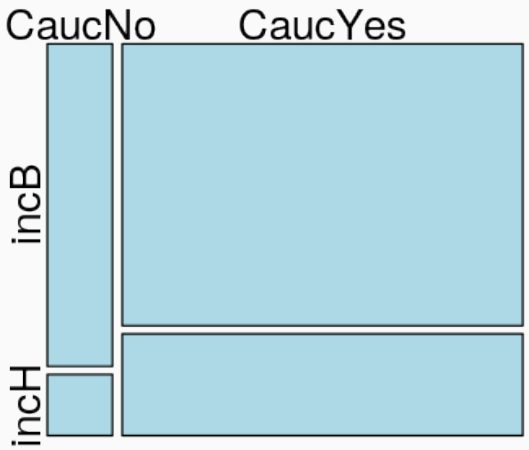
- $Y=$ income threshold ($\pm$ \$50k)
- **age** : continuous.
- **workclass** : Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- **fnlwgt** : continuous.
- **education** : Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- **education-num** : continuous.
- **marital-status** : Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- **occupation** : Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- **relationship** : Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- **race** : White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- **sex** : Female, Male.
- **capital-gain** : continuous.
- **capital-loss** : continuous. hours-per-week : continuous.
- **native-country** : United-States, Cambodia, England, Puerto-Rico, Canada, ...

- On laisse de côté **fnlwgt** (pondération) et **nativ-country** (redondant), **relationship** $\rightarrow$ child, **race** $\rightarrow$ CaucYes / CaucNo

Un biais social évident




```
Confidence interval for the DI
(by delta method)

round(dispImp(datBas[,"sex"],
datBas[,"income"]),3)

0.349 0.367 0.384
```




```
Confidence interval for the DI
(delta method)

round(dispImp(datBas$origEthn,
datBas$income),3)

0.566 0.601 0.637
```

La régression logistique augmente le biais !

```
log.lm=glm(income~.,data=datApp,family=binomial)

# significativity of the parameters
anova(log.lm,test="Chisq")

Df       Deviance          Resid. Df        Resid. Dev       Pr(>Chi)
NULL     NA       NA       35771    40371,72         NA
age      1        1927,29010       35770    38444,43         0,000000e+00
educNum  1        4289,41877       35769    34155,01         0,000000e+00
mariStat         3        6318,12804       35766    27836,88         0,000000e+00
occup    6        812,50516        35760    27024,38         3,058070e-172
origEthn         1        17,04639         35759    27007,33         3,647759e-05
sex      1        50,49872         35758    26956,83         1,192428e-12
hoursWeek        1        402,82271        35757    26554,01         1,338050e-89
LcapitalGain     1        1252,69526       35756    25301,31         2,154522e-274
LcapitalLoss     1        310,38258        35755    24990,93         1,802529e-69
child    1        87,72437         35754    24903,21         7,524154e-21

# Prevision
pred.log=predict(log.lm,newdata=daTest,type="response")
# Confusion matrix
confMat=table(pred.log>0.5,daTest$income)

incB     incH
FALSE    6190     899
TRUE     556      1298

tauxErr(confMat): 16,27

round(dispImp(daTest[,"sex"],Yhat),3) : 0.212 0.248 0.283

# Overall Accuracy Equality?
apply(table(pred.log<0.5,daTest$income,daTest$sex),3,tauxErr)

Female 91.81      Male 79.7
```

Et que fait Random Forest ?

Random Forest améliore significativement la capacité de prédiction...

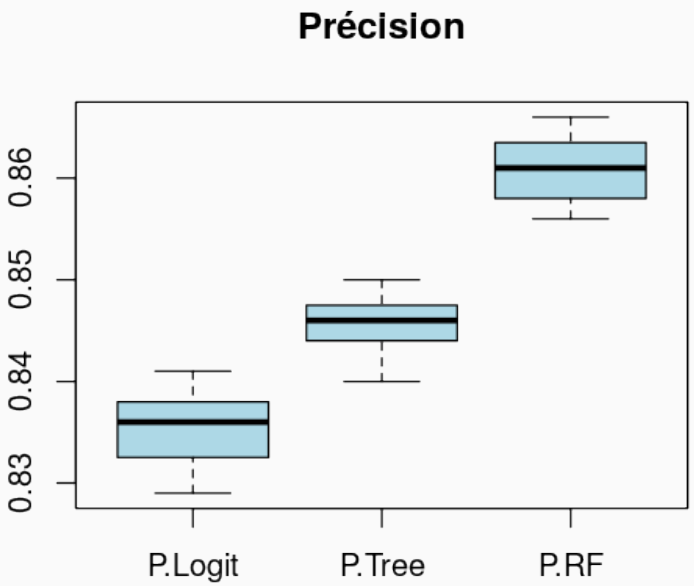
```
rf.mod=randomForest(income~.,data=datApp)
pred.rf=predict(rf.mod,newdata=daTest,type="response")
confMat=table(pred.rf,daTest$income)
confMat
tauxErr(confMat)
```

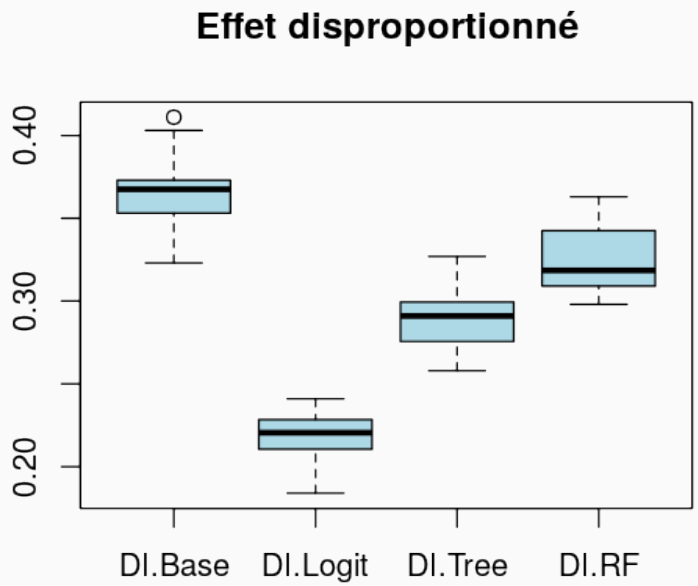
```
pred.rf  incB    incH
incB     6301     795
incH     445     1402

13,87
```

```
round(dispImp(daTest[,"sex"],pred.rf),3)
0.329  0.375  0.42
```

... sans augmenter le biais (ici).

Résumé des résultats par algorithme





$\Longrightarrow$ Random Forest est à la fois plus performant et moins discriminant (MAIS pas interprétable)

$\Longrightarrow$ Ca n'est pas une règle générale ! ça dépend des données.

$\Longrightarrow$ Une approche rigoureuse doit considérer les différents algorithmes et inclure une discussion sur les effets discriminatoires.

Biais individuels : Testing

Est-ce que les prédictions changent si on modifie la valeur du champ "sex" ?

```
daTest2=daTest
# Changement de genre
daTest2$sex=as.factor(ifelse(daTest$sex=="Male","Female","Male"))
# Prevision du 'nouvel' echantillon test
pred2.log=predict(log.lm,daTest2,type="response")
table(pred.log<0.5,pred2.log<0.5,daTest$sex)
```

```
Female
FALSE   TRUE
FALSE   195     0
TRUE    23      2679

Male
FALSE   TRUE
FALSE   1489    155
TRUE    0       4402
```

➔ 178 personnes obtiennent une décision différente, dans le sens attendu...

Approche naïve : supprimer la variable protégée

```
# estimation without the variable "sex"
log_g.lm=glm(income~.,data=datApp[,-6],family=binomial)

# Prevision
pred_g.log=predict(log_g.lm,newdata=daTest[,-8],type="response")
# Confusion Matrix
confMat=table(pred_g.log>0.5,daTest$income)
confMat
```

```
incB incH
FALSE 6157  953
TRUE   523 1310
```

```
tauxErr(confMat)
```

```
16.5
```

```
Yhat_g=as.factor(pred_g.log>0.5)
round(dispImp(daTest[,"sex"],Yhat_g),3)
```

```
0.232 0.269 0.305
```

$\Longrightarrow$ la qualité de prédiction n'est pas altérée, mais le biais reste le même !

Grâce au testing

Facile : utiliser la prédiction la meilleure pour toutes les valeurs de la variable protégée

```
fairPredictGenre=ifelse(pred.log<pred2.log,pred2.log,pred.log)
confMat=table(fairPredictGenre>0.5,daTest$income)
confMat;tauxErr(confMat)

incB    incH
FALSE   6145    936
TRUE    535     1327

16.45


round(dispImp(daTest$sex,as.factor(fairPredictGenre>0.5)),3)
0.24  0.277  0.314

# recall:
round(dispImp(daTest$sex,as.factor(pred.log>0.5)),3)
0.212  0.248  0.283
```

➔ pas d'influence sur la capacité de prédiction

➔ petite réduction du biais, mais ne supprime pas les discriminations de

Adapter le seuil de décision à chaque classe

```
Yhat_cs=as.factor(ifelse(daTest$sex=="Female",pred.log>0.4,pred.log>0.5))
round(dispImp(daTest[,"sex"],Yhat_cs),3)
tauxErr(table(Yhat_cs,daTest$income))

0.293 0.334 0.375

16.55

# Stronger correction forcing the DI to be at least 0.8:

Yhat_cs=as.factor(ifelse(daTest$sex=="Female",pred.log>0.15,pred.log>0.5))
round(dispImp(daTest[,"sex"],Yhat_cs),3)
tauxErr(table(Yhat_cs,daTest$income))

0.796 0.863 0.93

18.57
```

$\Longrightarrow$ déteriore significativement la capacité de prédiction

$\Longrightarrow$ forme de "discrimination positive" qui est un choix "politique"

Construire un classifieur par classe

En régression logistique, cela revient à considérer toutes les interactions de la variable protégée avec les autres

```
yHat=predict(reg.log,newdata=daTest,type="response")
yHatF=predict(reg.logF,newdata=daTestF,type="response")
yHatM=predict(reg.logM,newdata=daTestM,type="response")

yHatFM=c(yHatF,yHatM); daTestFM=rbind(daTestF,daTestM)

# Cumulated errors
table(yHatFM>0.5,daTestFM$income)
incB    incH
FALSE   6150    935
TRUE    530     1328

table(yHat>0.5,daTest$income)
incB    incH
FALSE   6154    950
TRUE    526     1313

tauxErr(table(yHatFM>0.5,daTestFM$income))
16.38

tauxErr(table(yHat>0.5,daTest$income))
16.5

# Bias with an without class separation
round(dispImp(daTestFM[,"sex"],as.factor(yHatFM>0.5)),3)
0.284 0.324 0.365

round(dispImp(daTest[,"sex"],as.factor(yHat>0.5)),3)
0.212 0.248 0.283
```


$\Longrightarrow$ cela réduit le biais

Conclusion

- La classification automatique peut augmenter le biais social
- Tous les algorithmes ne sont pas équivalents face à ce phénomène
- Il faut se méfier particulièrement des classifieurs linéaires
- Random Forest peut (au moins parfois) ne pas en souffrir
- Retirer les variables protégées ne suffit pas
- L'augmentation du biais diminue quand on considère les interactions des variables protégées
- Une solution généralement simple et efficace est d'ajuster des modèles différents pour les différentes classes de variables protégées
- ... si toutefois ces variables protégées sont bien observées !

L'IA du Quotidien peut elle être Éthique ? Loyauté des Algorithmes d'Apprentissage Automatique
Philippe Besse, Céline Castets-Renard, Aurélien Garivier, Jean-Michel Loubes
Statistique et Société vol. 6 (3), Dec. 2018, pp.9-31
en accès libre : `http://statistique-et-societe.fr/article/view/719/762`
ArXiv :1810.01729 et hal-01886699