Respect de la vie privée et problématiques éthiques à l'ère des données massives

Sébastien Gambs
Chaire de recherche du Canada en analyse respectueuse de la vie privée et éthique des données massives
Université du Québec à Montréal (UQAM)

gambs.sebastien@uqam.ca

4 novembre 2019

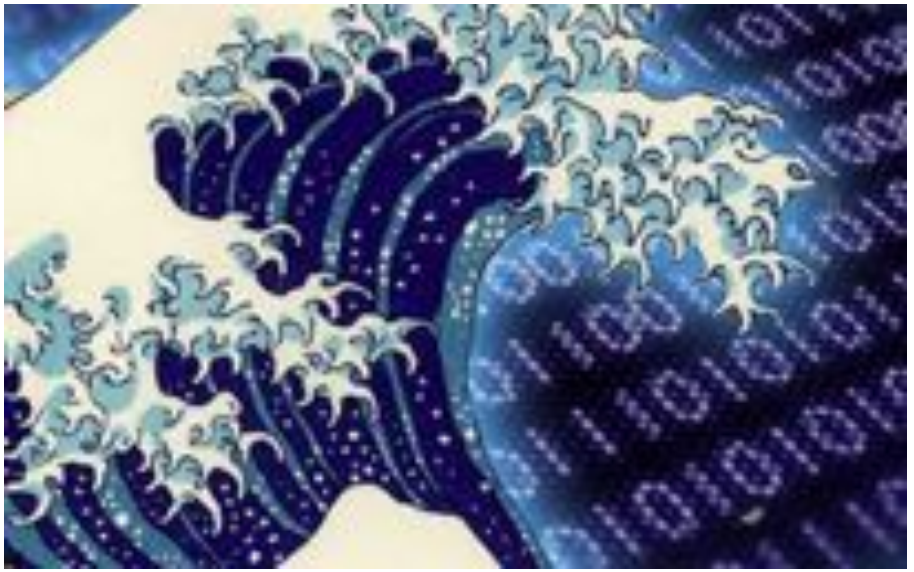
Respect de la vie privée

- Le respect de la vie privée est un droit fondamental de tout individu :
  - Déclaration universelle des droits de l'homme à l'assemblée des nations unies (article 12), 1948.
  - Loi sur la protection des données personnelles dans le secteur privé, Québec.
  - Règlement général européen sur la protection des données, RGPD (voté en 2016, devenu effectif en mai 2018).

- Risques : collecte et utilisation des traces numériques et des données personnelles à des fins frauduleuses.
- Exemples : pourriel ciblé, usurpation d'identité, profilage, discrimination, ...

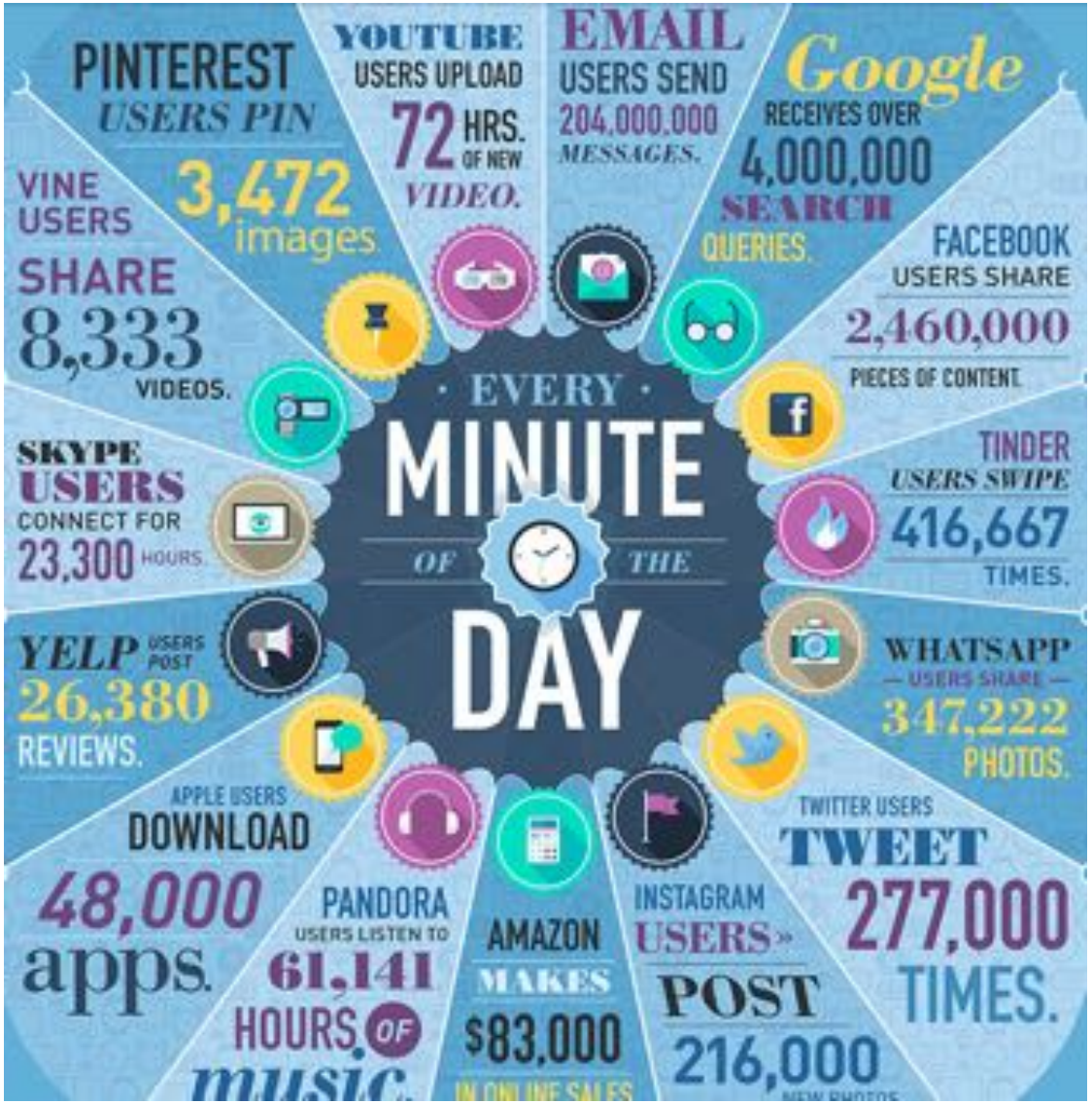
Données massives (*"Big Data"* en anglais)

- Fait référence de manière très large à l'accroissement de la quantité et de la diversité de données collectées et disponibles.



- Caractérisiation technique : souvent défini en terme des 5 "V" (*Volume*, *Variété*, *Vélocité*, *Variabilité* and *Veracité*).
- Promesse principal des données massives couplées à l'intelligence artificielle : offrir la possibilité de réaliser des inférences avec un niveau inédit de détails et de précision.

Un aperçu du “Big personal Data”



Impact des avancées en IA sur le respect de la vie privée

1. Magnification des risques de vie privée dû à l'augmentation du volume et de la diversité des données personnelles collectées et de la puissance de calcul pour les traiter.
2. Souvent les données sont "ré-utilisées" pour un but complètement différent que celui pour lequel on a demandé le consentement à l'utilisateur.
3. Les inférences qui sont possibles avec les données massives sont beaucoup plus fines et précises que précédemment.
4. Une divulgation massive de données sans prendre en compte le respect de la vie privée $\Rightarrow$ fuite majeure de données

- Une fois qu'une donnée est rendue publique, elle est là pour toujours (pas de *droit à l'effacement* absolu).

5. Éthique de l'inférence : quelles sont les inférences qui sont acceptables pour la société et celles qui ne le sont pas ?

Exemple d'inférence sensible : police prédictive

THE VERGE TWEET SHARE

Struggling to reduce its high murder rate, the city of Chicago has become an incubator for experimental policing techniques. Community policing, stop and frisk, "interruption" tactics — the city has tried many strategies. Perhaps most controversial and promising has been the city's futuristic "heat list" — an algorithm-generated list identifying people most likely to be involved in a shooting.

The hope was that the list would allow police to provide social services to people in danger, while also preventing likely shooters from picking up a gun. But a new report from the RAND Corporation shows nothing of the sort has happened. Instead, it indicates that the list is, at best, not even as effective as a most wanted list. At worst, it unnecessarily targets people for police attention, creating a new form of profiling.

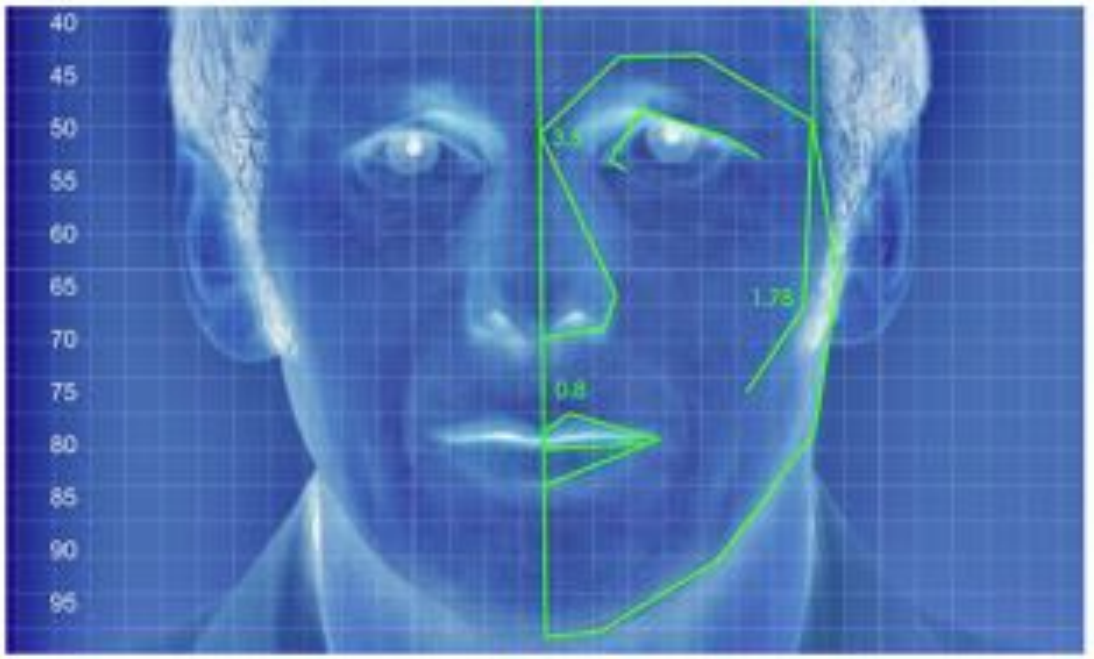
IT UNNECESSARILY TARGETS PEOPLE FOR POLICE ATTENTION

Funded through a $2 million grant from the National Institute of Justice, the list's algorithm identifies people by looking not only at arrests, but also whether someone is socially connected with a known shooter or shooting victim. The program also has a kind of pre-crime feature in which police visit people on the list before any crime has been committed.

Autre exemple d'inférence sensible : prédiction de l'orientation sexuelle à partir de photos

New AI can guess whether you're gay or straight from a photograph

An algorithm deduced the sexuality of people on a dating site with up to 91% accuracy, raising tricky ethical questions



An illustrated depiction of facial analysis technology similar to that used in the experiment. Illustration: Alamy

Artificial intelligence can accurately guess whether people are gay or straight based on photos of their faces, according to new research that suggests machines can have significantly better "gaydar" than humans.

Technologies de protection de la vie privée

Technologies de protection de la vie privée (*Privacy Enhancing Technologies* ou *PETs* en anglais) : ensemble de techniques pour protéger la vie privée d'un individu et lui offrir un meilleur contrôle sur ses données personnelles.

Exemple de PETs : réseau de communication anonyme, chiffrement homomorphe.

Deux principes fondamentaux :

- Minimisation des données : seule l'information nécessaire pour compléter une tâche particulière devrait être collectée.
- Souveraineté des données : permettre à un utilisateur de garder un contrôle sur ses données, en particulier sur comment elles sont collectées et disséminées.

Piste de solution possible : anonymisation de données

- Objectif : modifier les données avant publication pour limiter les risques en terme de respect de la vie privée (ré-identification, inférence d'attribut sensible).

- Nombreuses difficultés : question scientifique difficile (nombreux exemples d'anonymisation ayant été cassées), pas de technique universel fonctionnant pour tout type de données, le croisement de données possible grâce aux données massives augmente les risques de ré-identification, . . .

Opacité des algorithmes d'apprentissage

- L'apprentissage machine joue un rôle central dans la plupart des systèmes personnalisés.
- Opacité : difficulté de comprendre et d'expliquer leur décision due à leur conception complexe.
- Exemple : le classifieur produit par un algorithme d'apprentissage profond est typiquement composé un réseau de neurones composé de nombreuses couches.
- Risque de la "dictature algorithme" (Rouvroy) : perte de contrôle des individus sur leur vie numérique suite à des décisions automatiques contre lesquelles il n'existe aucune procédure de recours.

Droit à l'équité et à la transparence (Règlement général sur la protection des données)

Afin d'assurer un traitement équitable et transparent à l'égard de la personne concernée, [...], le responsable du traitement devrait utiliser des procédures mathématiques ou statistiques adéquates aux fins du profilage, appliquer les mesures techniques et organisationnelles appropriées pour faire en sorte, en particulier, que les facteurs qui entraînent des erreurs dans les données à caractère personnel soient corrigés et que le risques d'erreur soit réduit au minimum, [...] et qui prévienne, entre autres, les effets discriminatoires à l'égard des personnes physiques fondées sur la l'origine raciale ou ethnique, les opinions politiques, la religion ou les convictions, l'appartenance syndicale, le statut génétique ou l'état de santé, ou l'orientation sexuelle, ou qui se traduisent par des mesures produisant un tel effet.

Difficulté : absence de consensus sur la "bonne" notion d'équité

Translation tutorial:
21 fairness definitions and their politics

Arvind Narayanan
(Computer scientist, Princeton University)

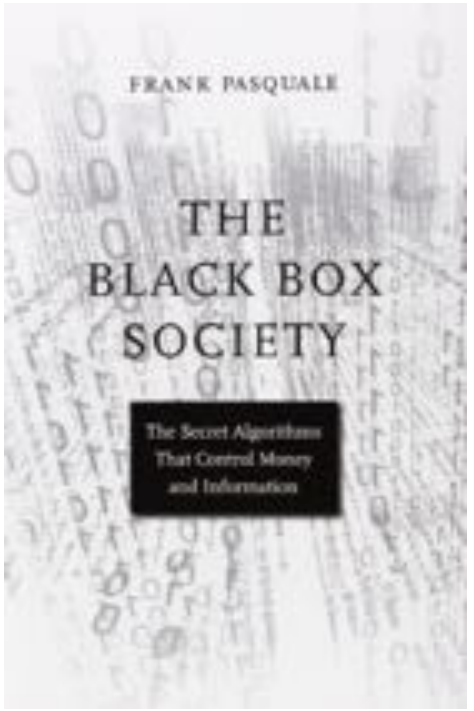
Computer scientists and statisticians have devised numerous mathematical criteria to define what it means for a classifier or a model to be fair. The proliferation of these definitions represents an attempt to make technical sense of the complex, shifting social understanding of fairness. Thus, these definitions are laden with values and politics, and seemingly technical discussions about mathematical definitions in fact implicate weighty normative questions. A core component of these technical discussions has been the discovery of trade-offs between different (mathematical) notions of fairness; these trade-offs deserve attention beyond the technical community.

This tutorial has two goals. The first is to explain the technical definitions. In doing so, I will aim to make explicit the values embedded in each of them. This will help policymakers and others better understand what is truly at stake in debates about fairness criteria (such as individual fairness versus group fairness, or statistical parity versus error-rate equality). It will also help computer scientists recognize that the proliferation of definitions is to be celebrated, not shunned, and that the search for one true definition is not a fruitful direction, as technical considerations cannot adjudicate moral debates.

FIGURE – Exemple d'un tutoriel donné par Arvind Narayanan à la conférence FAT*18

La transparence comme une première étape

- Asymétrie de l'information : forte différence entre ce que le système sait d'une personne et ce que la personne sait à propos du système.



- Le manque de transparence conduit au manque de confiance.
- Besoin fort d'améliorer la transparence des algorithmes d'apprentissage.

Possible approches pour la transparence

1. Approches par la régulation pour forcer les companies à laisser les utilisateurs examiner et corriger les informations collectées à leur propos.
2. Méthodes pour améliorer la transparence en “ouvrant la boîte noire”.

- Outils afin d’atteindre la transparence par conception.
- Exemples : publication du code source, utilisation d’un modèle interprétable en apprentissage machine.

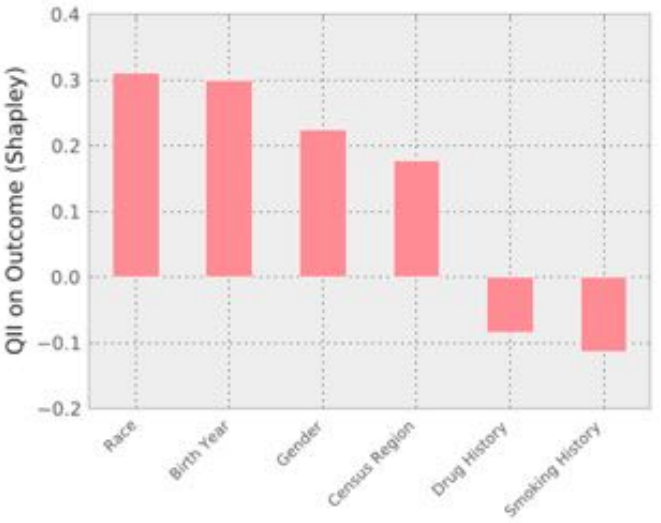
Investiguer l'origine des biais

1. Problème dans la collecte des données dû à une erreur ou au fait que les données sont biaisées de manière inhérentes.
- Exemples : erreur dans le profil d'un utilisateur, ensemble de données qui reflète des décisions discriminatoires contre une population particulière.
2. Imprécision dû à l'algorithme d'apprentissage.
- Exemple : l'algorithme est très précis, à l'exception de 1% des individus.

Mesurer la discrimination

- Exemple : mesure quantitative de l'influence des entrées.

(a) Mr. Z's profile



(b) Transparency report for Mr. Z's positive classification

- Défi : possibilité d'une *discrimination indirecte* dans le cas où l'attribut protégé est inféré à partir d'autres attributs.
- Exemple : même si l'origine ethnique n'est pas demandé à l'utilisateur, dans certains pays cette information est très fortement corrélée avec le code postal.

Améliorer l'équité

- Objectif ultime : pouvoir améliorer l'équité tout en limitant l'impact sur la précision.
- Exemples d'approches possibles :
  - Échantillonner les données d'entrée afin d'enlever le biais originel,
  - Modifier l'algorithme afin qu'il devienne conscient de la discrimination par conception,
  - Adapter la sortie produite par l'algorithme (par exemple le classifieur) pour réduire la discrimination.
- Sujet actif de recherche mais toujours dans son enfance, beaucoup de travail reste à faire.

Vers la responsabilité algorithmique

- Limite : la transparence n'est pas forcément synonyme d'interprétabilité ou d'imputabilité.
- Exemple : la structure d'un classifieur pourrait être publique mais trop complexe pour être appréhendé par un humain.
- Besoin fort pour le développement d'outils qui peuvent analyser et certifier un classifieur.
- Objectif : verifier que les décisions prises par une intelligence artificielle ou un algorithme d'apprentissage corresponde au comportement intentionnel ou aux valeurs éthiques qui sont attendues.

Propriétés attendues d'un système intelligent

- Loyauté (notion très présente en Europe) : le système se comporte comme déclaré à l'utilisateur.
- Équité / non-discrimination.
- Transparence : comprendre les données utilisées par le système et son fonctionnement, être capable d'expliquer ses décisions.
- Responsabilité : être capable d'avoir une autorité à qui s'adresser en cas de dysfonctionnement ou de constatation de la prédiction par l'utilisateur.
- Conformité : le système respecte les lois en vigueur ainsi que sa spécification

Importance du respect de l'autonomie

- Importance d'expliquer comment fonctionne le système, en particulier si celui-ci peut avoir un impact significatif sur un humain.
- Transparence par conception sur les données collectées et l'usage du système.
- Le consentement pourrait être demandé à l'utilisateur pour participer au projet mais aussi séparément pour contribuer aux systèmes appris sur les données qui seront potentiellement partagés.
- Si possible, il faut garder un humain dans la boucle (pas de prise de décision automatique se basant simplement sur les prédictions du système dans des cas critiques).
- Défi : Équilibre à trouver entre interprétabilité, limitation des biais et précision du système.

Conclusion

- Observation 1 : la capacité à enregistrer et stocker les informations personnelles augmentent de manière régulière.
- Observation 2 : les données massives résultent en de plus en plus de données devenant disponibles $\Rightarrow$ accroissement des possibilités d'inférence.
- Observation 3 : le mouvement d'ouverture des données va conduire à la divulgation de grosses quantités de données $\Rightarrow$ aggrave l'impact sur la vie privée (observation 2).
- Les avancées en intelligence artificielle et l'avènement des données massives magnifient les risques de vie privée qui existaient auparavant mais soulèvent aussi de nouvelles questions éthiques.
- Défi principal : trouver un équilibre entre les bénéfices sociétaux et économiques de l'intelligence et le respect de la vie ainsi que les droits fondamentaux des individus (liens forts avec les valeurs d'autonomie, justice et démocratie).

Emergence of initiative for the ethical development of AI

Examples :

- General Data Protection Regulation (GDPR) in Europe.
- Montréal Declaration for a Responsible Development of Artificial Intelligence (2018).

Consequence : strong pressure for companies for the ethically-aligned development of systems with respect to fairness or interpretability.

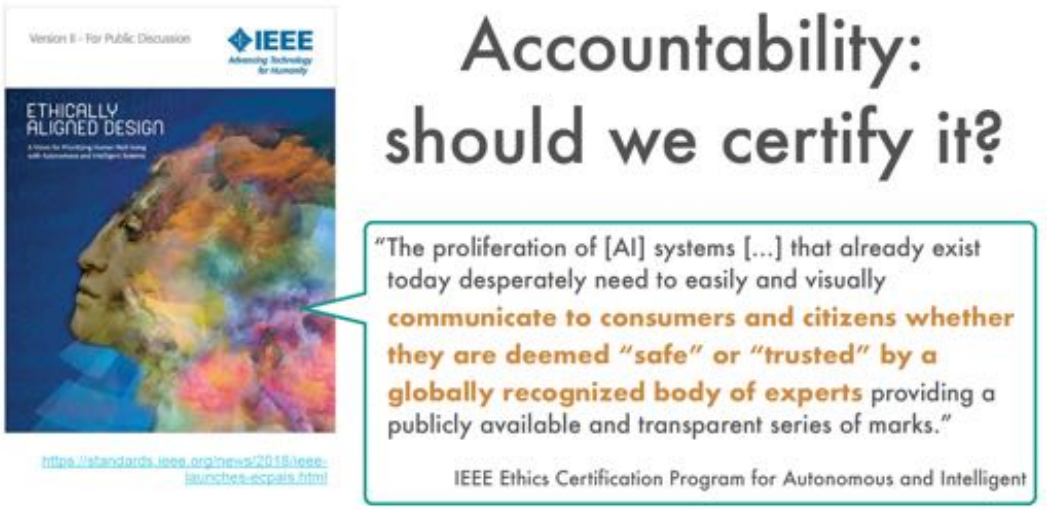


FIGURE – Excerpt a talk from Roel Dobbe.

"Right to *an* explanation" (GDPR)

The data subject should have tthe right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention. In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.

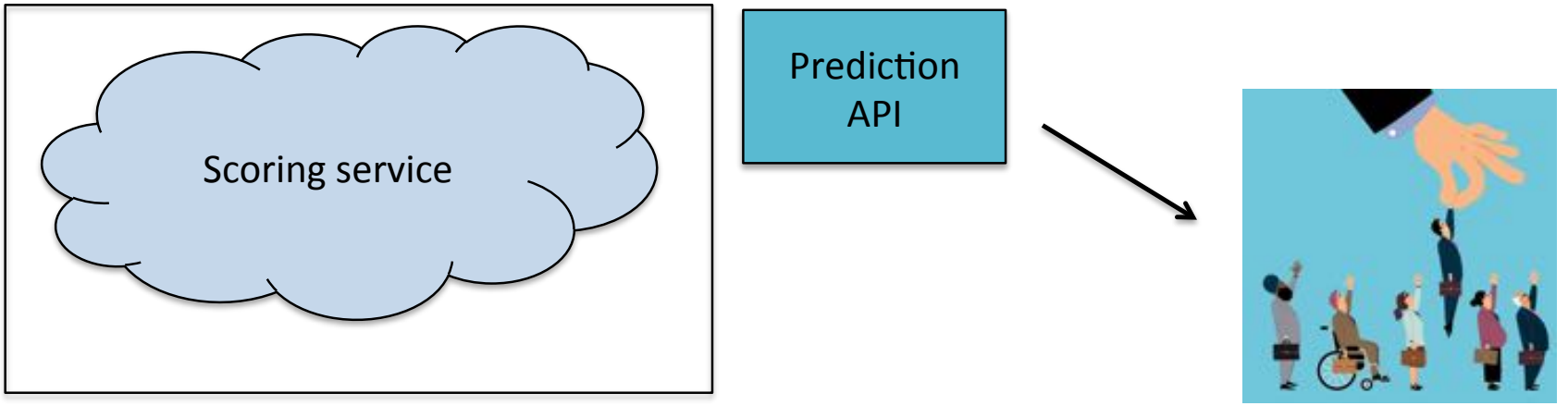
Fairwashing (to be published at ICML 2019)

- Fairwashing (in machine learning) : promoting the perception that a machine learning model respects some ethical values (ex : fairness) while it might not be the case.

- Main objective of the work : Raise the awareness on the risk of fairwashing by showing that an unfair model is explained in such a way that the underlying decisions seem more fair than they actually.
- Method followed : demonstrate that one can **systematically** found an interpretable model to rationalize biased decisions of a black-box model.
- Rationalization : process that can be used *a posteriori* to explain globally the decision of a black-box model (*i.e.*, rationalization of model explanation) as well as its decision for

Illustration of fairwashing through rationalization

In reality :



After rationalization :