Fairness et confidentialité en IA pour l'éducation : risques et opportunités

Jill-Jênn Vie

26 janvier 2023

Découvert l'algorithmique par les compétitions de programmation (Prologin)

Entraîneur de l'X au ICPC

Fondé Girls Can Code! en 2014 (toujours via Prologin)

Stages de prog° pour filles

Milité en faveur d'une agrégation d'informatique

1re édition en 2022

Sif
Société informatique de France

MINISTÈRE DE L'ÉDUCATION NATIONALE ET DE LA JEUNESSE

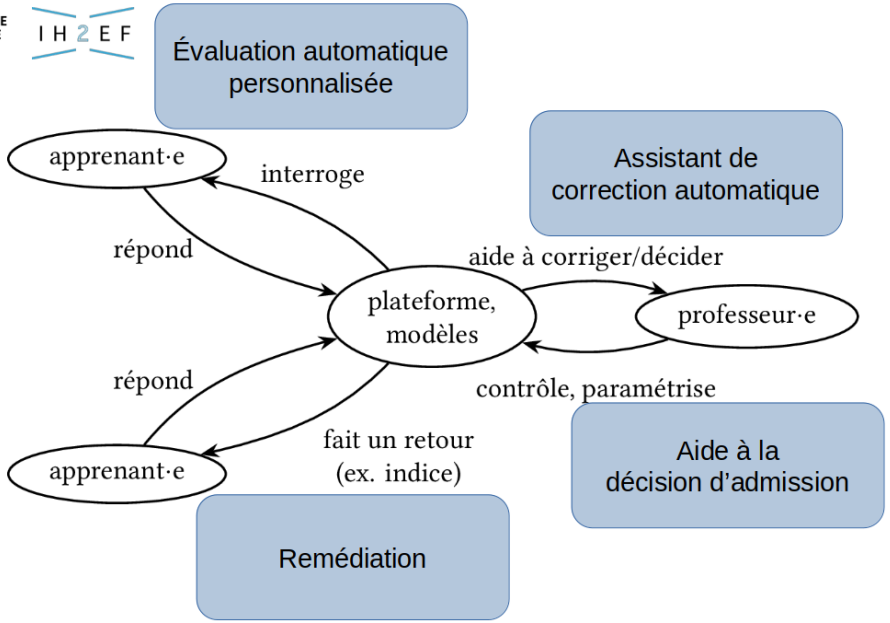
Liberté
Égalité
Fraternité

Sujets de recherche dans l'équipe Soda

Machine learning sur des données d'humains

- Données manquantes, inférence causale
- Représentations de bases de données
  - trajectoires de patients dans un hôpital (ex. AP-HP)
  - trajectoires d'apprenants sur une plateforme
- Applications en santé et éducation

Nos ingénieurs de recherche sont les développeurs principaux de la bibliothèque `scikit-learn`

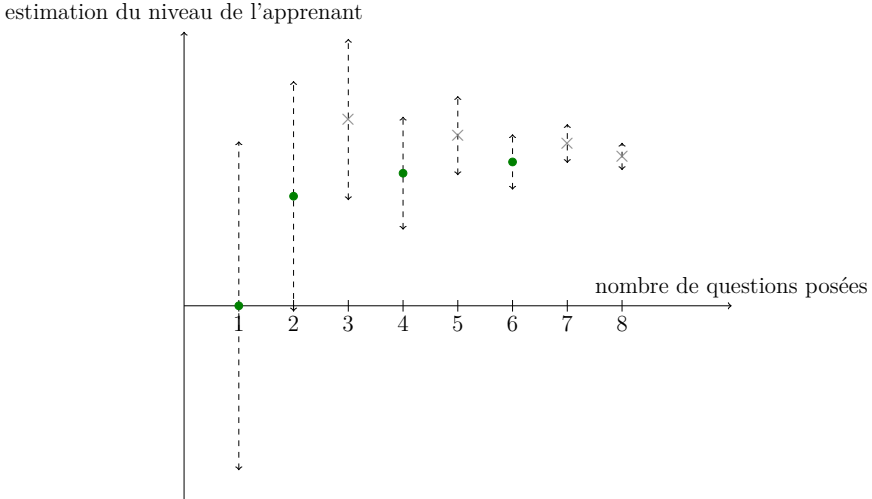

Évaluation automatique personnalisée
Assistant de correction automatique
apprenant·e
interroge
répond
aide à corriger/décider
plateforme, modèles
professeur·e
répond
contrôle, paramétrise
Aide à la décision d'admission
fait un retour (ex. indice)
apprenant·e
Remédiation

Mesurer les connaissances des apprenants à un instant donné

Théorie de la réponse à l'item (Rasch, 1961) (Lord, 1986) et un peu (Binet, 1905)

Tests adaptatifs → premières évaluations personnalisées par ordinateur (1970-1980)

Compromis entre bien mesurer et poser peu de questions

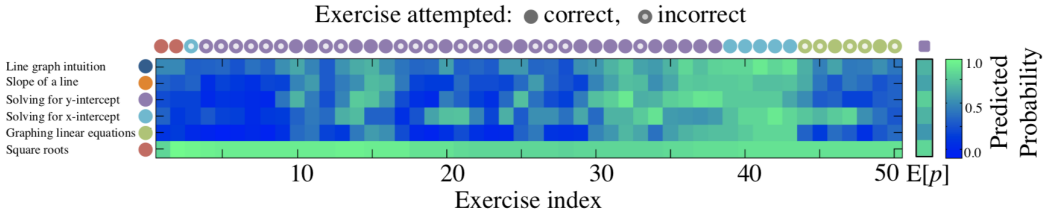


Tracer les connaissances au cours du temps : prédire la performance

Apprentissage d'une langue (jeux de données de Duolingo)

	PRON	VERB	PRON	NOUN	CONJ	PRON	VERB	PRON	NOUN
correct:	She	is	my	mother	and	he	is	my	father
student:	she	is		mader	and	he	is		fhader
label:	O	O	✖	✖	O	O	O	✖	✖

Exercices de maths



Recommandations de la Commission européenne (*guidelines*)

IA & données pour l'éducation et la formation

1. Facteur humain et supervision
2. Transparence
3. Diversité, non discrimination et *fairness* (impartialité)
4. Bien-être sociétal et environnemental
5. Confidentialité et gouvernance des données
6. Robustesse technique et sécurité
7. Responsabilité

Diversité, non discrimination et *fairness* (impartialité)

- Le système est-il accessible pour tous sans barrière ?
- Modes d'interaction appropriés pour les personnes à besoins spéciaux / interfaces appropriées
- Y a-t-il des procédures pour s'assurer que l'IA n'induira pas un traitement discriminatoire ou injuste pour ses utilisateurs ?
- La documentation du système ou son procédé d'entraînement indique-t-elle des biais potentiels dans les données ?

Fairness

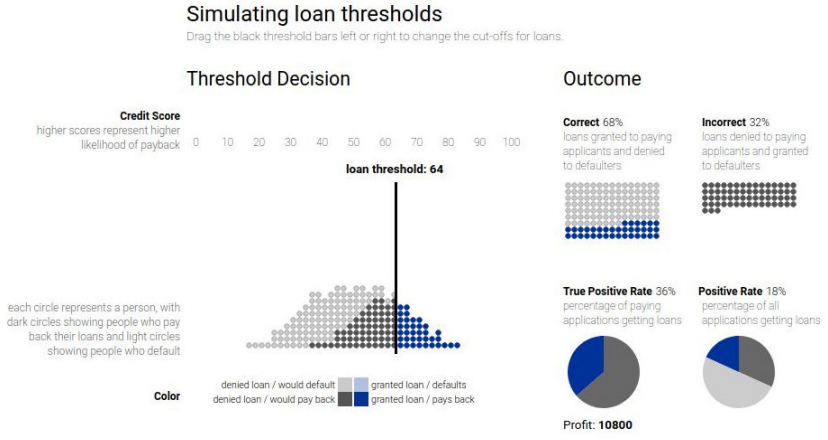
"Different models with the same reported accuracy can have a very different distribution of error across population" (Hardt, 2017)

Fairness

"Different models with the same reported accuracy can have a very different distribution of error across population" (Hardt, 2017)

Scores de criminalité (regardez la série *Psycho-Pass*):

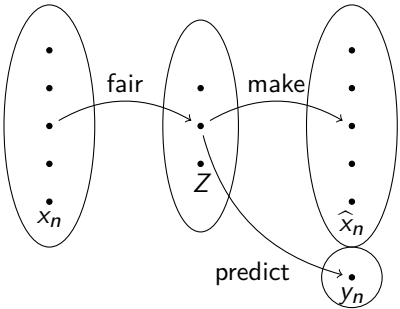
Beaucoup de définitions de *fairness*, parfois contradictoires

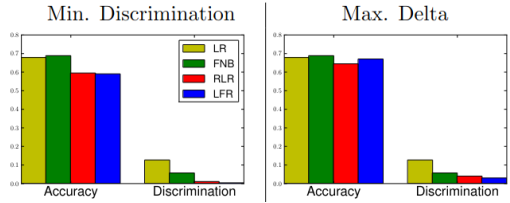


Moritz Hardt, Eric Price et Nati Srebro (2016). “Equality of opportunity in supervised learning”. In : *Advances in neural information processing systems*. T. 29

Leur visualisation interactive : *Attacking discrimination with smarter machine learning*

Apprendre des représentations “justes”



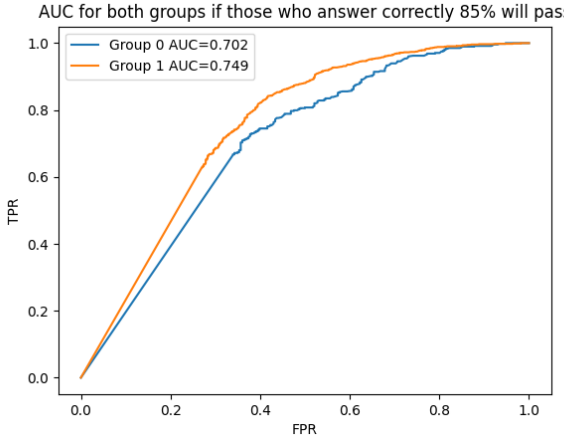


Rich Zemel et al. (2013). “Learning fair representations”. In : *International conference on machine learning*. PMLR, p. 325-333

Voir aussi

Ben Hutchinson et Margaret Mitchell (2019). “50 years of test (un) fairness: Lessons for machine learning”. In : *Proceedings of the conference on fairness, accountability, and transparency*, p. 49-58

Classifieurs différents selon la catégorie de population



Voir aussi Josh Gardner, Christopher Brooks et Ryan Baker (2019). “Evaluating the fairness of predictive student models through slicing analysis”. In : *Proceedings of the 9th international conference on learning analytics & knowledge*, p. 225-234

Confidentialité et gouvernance des données

- Des mécanismes sont-ils en place pour s'assurer que les données sensibles sont anonymisées et protégées pour en limiter l'accès aux personnes nécessaires ?
- Les données sont-elles traitées dans le seul but pour lequel elles ont été collectées ?
- Les enseignants ont-ils un moyen de signaler des problèmes quant à la confidentialité ou la protection des données ? En sont-ils informés ?
- Simplement : est-ce que le système respecte la RGPD ? Les paramètres de confidentialité sont-ils modifiables ?

Intérêt pour les données synthétiques

- Il est difficile d'accéder à des données sensibles (procédures très longues pour la recherche)
- Un jeu de données qui est ouvert peut être archivé pour toujours
- Pourquoi ne pas avoir plutôt accès à :
  - des statistiques (cf. DEPP)
  - des modèles pré-entraînés
  - des jeux de données synthétiques ? (ne serait-ce que pour la reproductibilité)

Les faits

La pseudonymisation, ce n'est pas suffisant

Narayanan et Shmatikov (2008) ont réussi à dé-anonymiser le jeu de données pseudonymisé du prix Netflix de films vus et notés, avec les données publiques d'IMDb

Les données de grande dimension sont rarement k-anonymisables

- 4 points espace-temps sont suffisant pour caractériser de façon unique 95% des trajectoires d'individus dans un jeu de données de 1,5 millions de lignes (De Montjoye et al., 2013)
- 15 données démographiques sont suffisantes pour réidentifier 99,96% des Américains (Rocher, Hendrickx et De Montjoye, 2019)

Les grands modèles de langage se souviennent des données d'entraînement

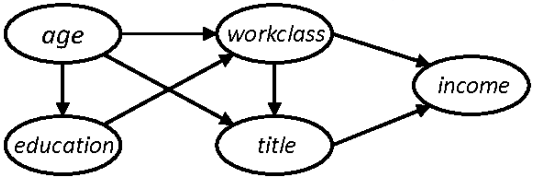
Nicholas Carlini et al. (2021). "Extracting training data from large language models". In : *30th USENIX Security Symposium (USENIX Security 21)*, p. 2633-2650

Modèles génératifs préservant la confidentialité

Confidentialité différentielle (*differential privacy*)

La sortie de l'algorithme doit être quasi indistinguable de selon si une personne manque dans le jeu de données d'entraînement.

Réseaux bayésiens PrivBayes (Zhang et al., 2017)

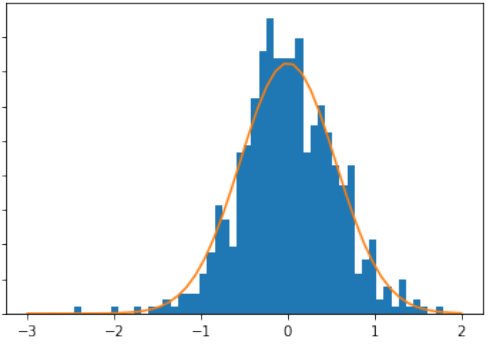


Générer des données individuelles à partir de données agrégées

Angeela Acharya et al. (2022). "GenSyn: A Multi-stage Framework for Generating Synthetic Microdata using Macro Data Sources". In : *Proceedings of the BigData 2022 conference*, in press

Intuition

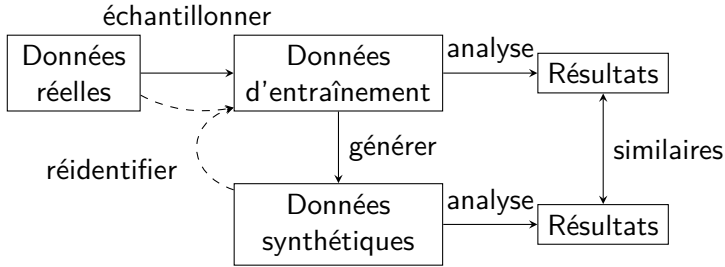
Échantillonner les données sensibles selon la distribution



Schéma

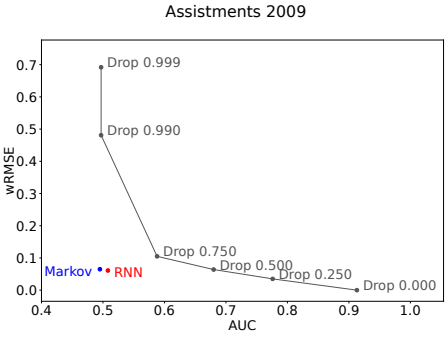
Utilité On doit pouvoir déduire des analyses similaires à partir du jeu de données réel et à partir du jeu de données synthétique

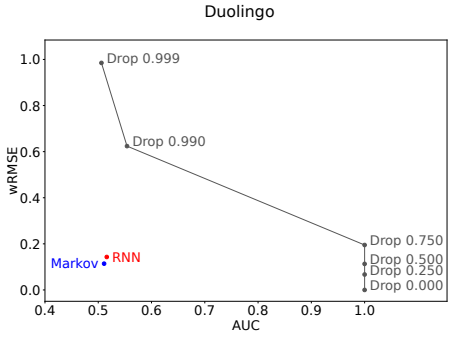
Réidentification Il faut empêcher que la réidentification soit facile / le jeu de données synthétique ne doit pas compromettre la confidentialité des participants



Jill-Jênn Vie*, Tomas Rigaux* et Sein Minn (2022). "Privacy-Preserving Synthetic Educational Data Generation". In : *Proceedings of EC-TEL 2022. Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption*, p. 393-406. ISBN : 978-3-031-16290-9. DOI : 10.1007/978-3-031-16290-9_29. URL : https://hal.archives-ouvertes.fr/hal-03715416

Résultats quantitatifs

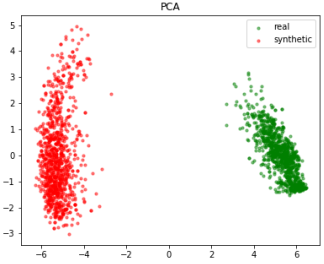


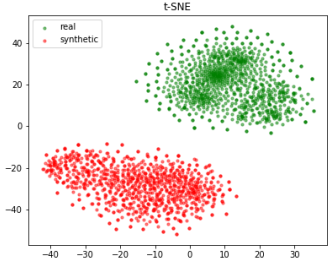


$\leftarrow$ réidentification (aussi bas que possible)

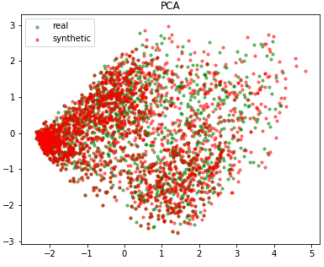
$\downarrow$ différence entre résultats (aussi bas que possible)

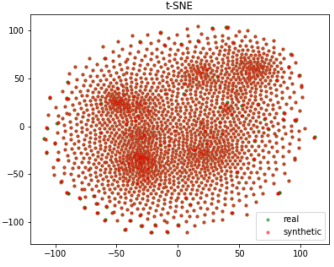
Génération fidèle





Génération non fidèle





Conclusion

- Ouvrons massivement les données de gens qui n'existent pas
- Il faut mesurer les discriminations pour réduire les inégalités
  - (donc regarder plus d'une métrique)

Merci ! Questions ?

Ces slides sur jjv.ie/slides/relia.pdf

- Acharya, Angeela et al. (2022). "GenSyn: A Multi-stage Framework for Generating Synthetic Microdata using Macro Data Sources". In : *Proceedings of the BigData 2022 conference*, in press.
- Carlini, Nicholas et al. (2021). "Extracting training data from large language models". In : *30th USENIX Security Symposium (USENIX Security 21)*, p. 2633-2650.
- De Montjoye, Yves-Alexandre et al. (2013). "Unique in the crowd: The privacy bounds of human mobility". In : *Scientific reports* 3.1, p. 1-5.
- Gardner, Josh, Christopher Brooks et Ryan Baker (2019). "Evaluating the fairness of predictive student models through slicing analysis". In : *Proceedings of the 9th international conference on learning analytics & knowledge*, p. 225-234.
- Hardt, Moritz, Eric Price et Nati Srebro (2016). "Equality of opportunity in supervised learning". In : *Advances in neural information processing systems*. T. 29.
- Hutchinson, Ben et Margaret Mitchell (2019). "50 years of test (un) fairness: Lessons for machine learning". In : *Proceedings of the conference on fairness, accountability, and transparency*, p. 49-58.

- Narayanan, Arvind et Vitaly Shmatikov (2008). "Robust de-anonymization of large sparse datasets". In : *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, p. 111-125.
- Rocher, Luc, Julien M Hendrickx et Yves-Alexandre De Montjoye (2019). "Estimating the success of re-identifications in incomplete datasets using generative models". In : *Nature communications* 10.1, p. 1-9.
- Vie*, Jill-Jênn, Tomas Rigaux* et Sein Minn (2022). "Privacy-Preserving Synthetic Educational Data Generation". In : *Proceedings of EC-TEL 2022. Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption*, p. 393-406. ISBN : 978-3-031-16290-9. DOI : 10.1007/978-3-031-16290-9_29. URL : https://hal.archives-ouvertes.fr/hal-03715416.
- Zemel, Rich et al. (2013). "Learning fair representations". In : *International conference on machine learning*. PMLR, p. 325-333.
- Zhang, Jun et al. (2017). "PrivBayes: Private data release via Bayesian networks". In : *ACM Transactions on Database Systems (TODS)* 42.4, p. 1-41.