DATASCIENCE, LEARNING AND APPLICATIONS

DALAS - Les biais en ML

11 mars 2024

Laure Soulier - Nicolas Baskiotis

SCIENCES SORBONNE UNIVERSITÉ

ISIR INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE

MLIA Machine Learning & Deep Learning for Information Access

Vidéo : la journée d'Inaja

https://youtube.com/watch?v=Fn83SEdToE4&ab_channel=CN22021G7
https://www.youtube.com/watch?v=UG_X_7g63rY&ab_channel=TED

Quel impact ?

Language (Technology) is Power : A Critical Survey of ?Bias ? in NLP. Blodgett et al., ACL 2020

- "Allocation harms arise when an automated system allocates resources (e.g., credit) or opportunities (e.g., jobs) unfairly to different social groups"
  - Les filles sont nulles en sport $\rightarrow$ Elle ne pédalera pas assez vite, on ne la recrute pas
- "Representational harms arise when a system (e.g., a search engine) represents some social groups in a less favorable light than others, demeans them, or fails to recognize their existence altogether."
  - "The nurse" dans la vidéo
- Invisibilisation : les accents peu représentés, on ne la comprend pas dans la vidéo

Les enjeux

- Interprétabilité/explicabilité des algorithmes
- Transparence/audit des algorithmes
- Responsabilité des algorithmes

Les types de biais

- Biais des données : "Garbage in - Garbage out"
- Biais des variables omises / biais de sélection
- Biais d'endogénéité (cas où une des variables explicatives est corrélée avec le terme d'erreur) : difficulté d'anticiper les changements de comportement et donc de les prendre en compte

Les types de biais

- **Biais d'échantillonnage.**
  - Manque de représentativité.
  - Exemple d'une entreprise qui cherche à prédire les pannes sur ses machines. Si elle collecte majoritairement des informations sur les erreurs, l'algorithme ne sera pas en capacité d'identifier suffisamment précisément le fonctionnement normal de l'équipement.
- **Biais de mesure.**
  - Absence de mesure ou d'enregistrement précis des données qui ont été sélectionnées.
  - Exemple : le salaire peut y avoir des différences de traitements (primes, avantages ?), ou des différences régionales dans les données.

Les types de biais

- **Biais d'exclusion.**
  - Provient de données qui sont retirées de manière inappropriée de la source de données.
  - Suppression des doublons lorsque les éléments de données sont réellement distincts.
- **Biais d'enregistrement.**
  - Enregistrer seulement certaines données
  - Enregistrement pas continu (capteurs) / données manquantes
  - Acte d'observation biaisé

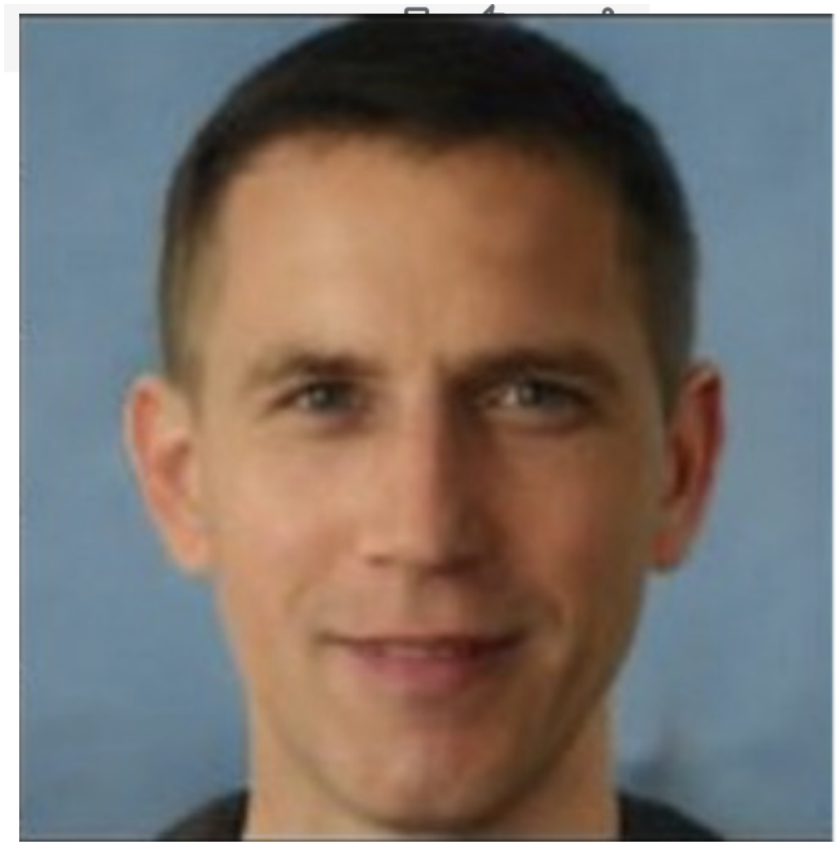
Les types de biais

- **Biais liés aux préjugés.**
  - Surtout lors de l'utilisation d'historiques qui sont eux mêmes soumis à des préjugés (recrutement)
  - Distortion : le biais de "du mouton de Parnurge" qui peut conduire le programmeur à suivre des modélisations qui sont populaires sans s ?assurer de leur exactitude.
- **Biais de confirmation.**
  - Désir de sélectionner uniquement les informations qui soutiennent ou confirment quelque chose que vous connaissez déjà, plutôt que des données qui pourraient suggérer quelque chose qui va à l'encontre d'idées préconçues.
- **Effet de mode.** surreprésenter un phénomène

Quelques exemples supplémentaires

Quelques exemples supplémentaires

Quelques exemples supplémentaires



Les solutions

- Identifier les sources potentielles (utiliser le liste des types de biais)
- Compléter l'information, ré-échantillonnage, redressement
- Ajout de variables auxiliaires
- Surveiller les modèles déployés en production (peut-être différent que le comportement en test)

Pour qui ?

- Les concepteurs des modèles
- Les entreprises qui réutilisent/raffinent les modèles
- Les utilisateurs

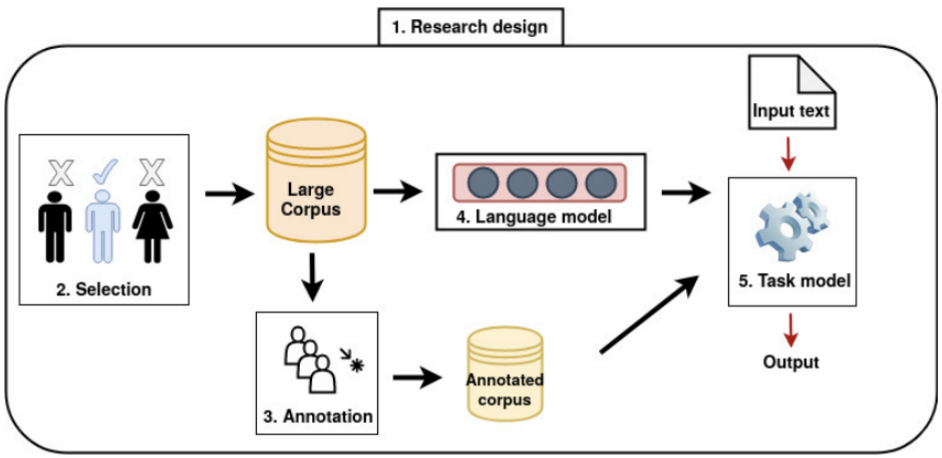
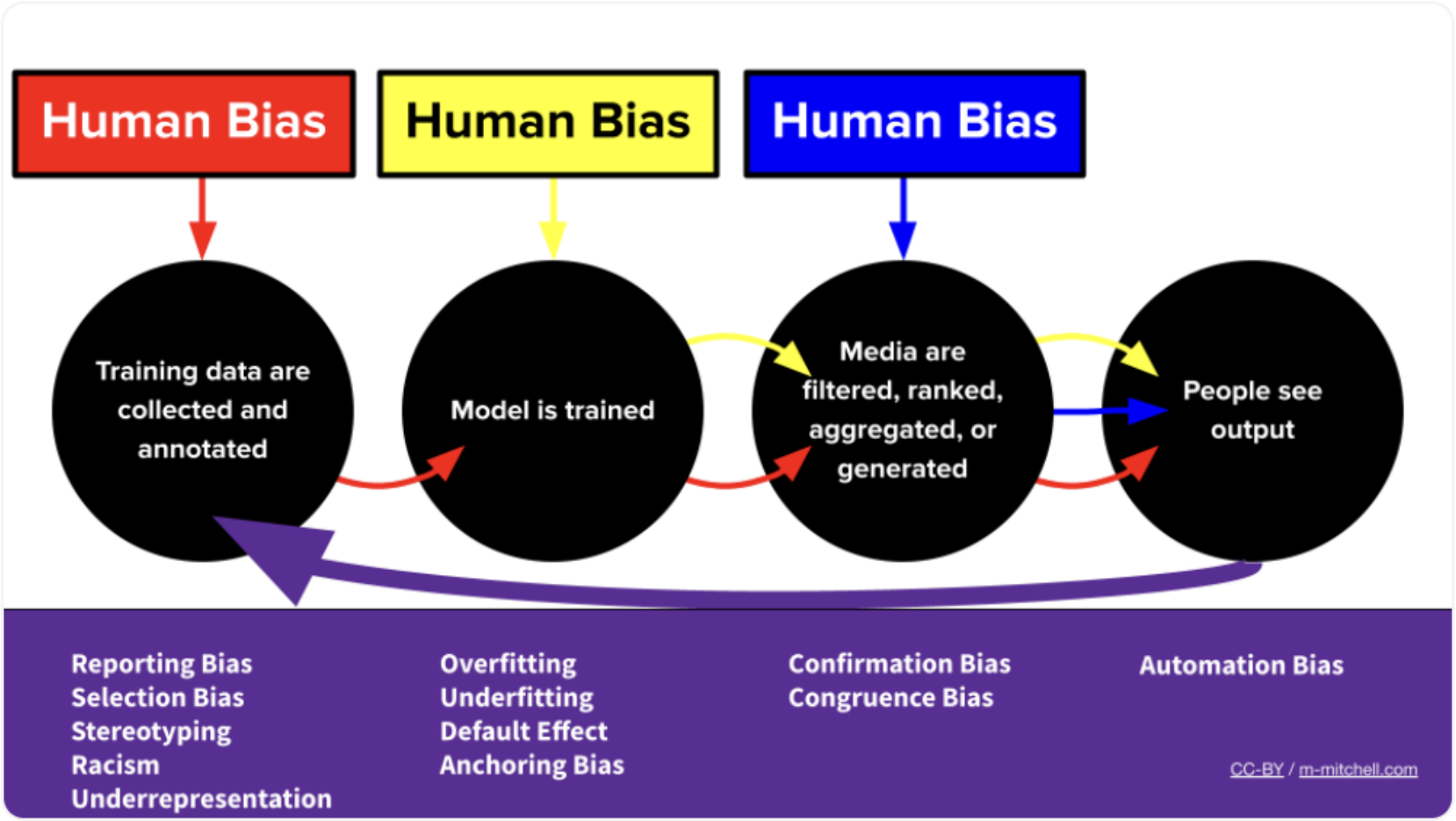


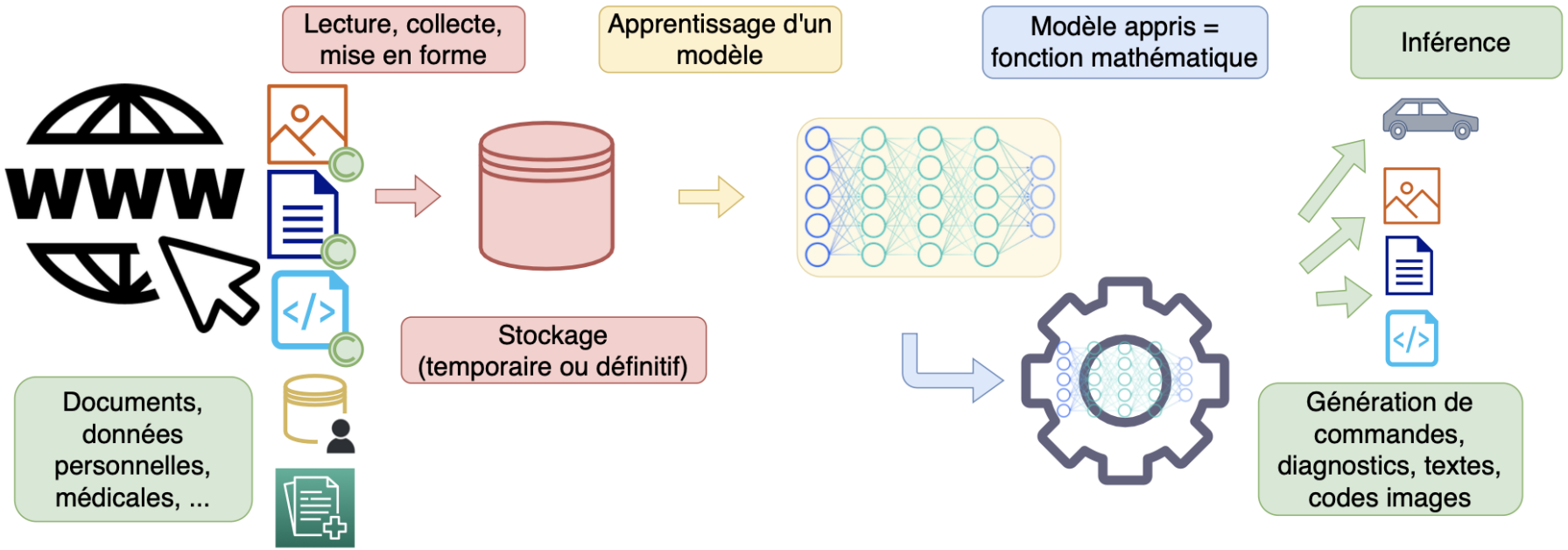
Figure 1 – (c) Aurélie Névéol - from [Hovy and Prabhumoye, 2021]

Pour qui ?



The Bias ML Pipeline by Meg

Enjeux dans la pipeline



Droit d'auteur, droit des bases de données

Droit de collecte, droit de copie, consentement

Droit d'utiliser les données dans un algorithme

Modèle = émanation des données ?

Cambridge Analytica

Clearview.ai

Reproductions d'extraits non traçables

Régulation des usages

Responsabilité en cas d'erreur

Le mot de la fin : et il n'y a pas que les algorithmes qui sont biaisés...

SCIENCES SORBONNE UNIVERSITÉ

Plus 25% de rapidité, 12% de productivité et 40% d'amélioration : ChatGPT booste les performances des consultants... surtout les pires

Antoine Crochet-Damais
JDN

Mis à jour le 10/10/23 06:55

L'enquête a été réalisée par Harvard et le MIT. Ils se sont penchés sur l'efficacité des consultants du Boston consulting Group, avec ou sans ChatGPT.

Selon une étude dirigée par Havard et le MIT, les consultants du Boston Consulting Group (BCG) voient la qualité de leur travail exploser de 40% avec l'utilisation de ChatGPT. L'étude en question a été réalisée auprès de 758 consultants de l'entreprise de conseil américaine. Un groupe utilisant ChatGPT a été comparé à un groupe témoin n'ayant pas recours à l'IA conversationnelle d'OpenAI. Le différentiel de 40% est mesuré au regard du niveau de qualité du travail fourni (voir le graphique ci-dessous).

Sources - lectures supplémentaires

- https://www.telecom-paris.fr/wp-content-EvDsK19/uploads/2019/02/Algorithmes-Biais-discrimination-equite.pdf
- https://members.loria.fr/KFort/files/fichiers_cours/EthiqueBiais.pdf
- https://huggingface.co/blog/ethics-soc-2
- https://quorumlanguage.com/lessons/code/Data/Lesson6.html