Validation et gestion des données

MGL 7320: Ingénierie logicielle des systèmes d'intelligence artificielle

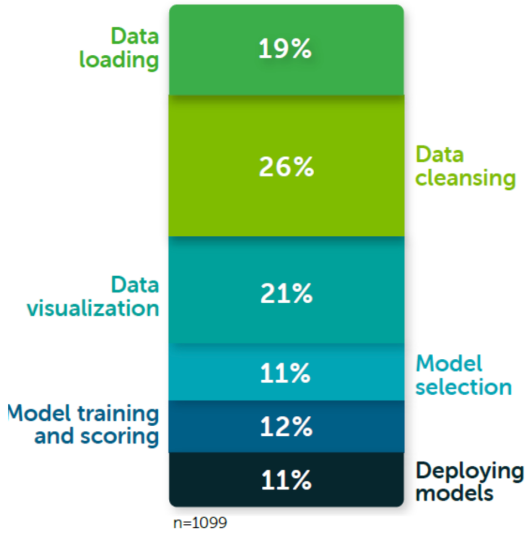
G I G O

Garbage In Garbage Out

La qualité des données est primordiale pour la qualité ML

- Un modèle d'apprentissage automatique est limité par la qualité des données en entrée

- Les scientifiques des données passent beaucoup plus de temps à assurer **la qualité des données** qu'à travailler avec les modèles



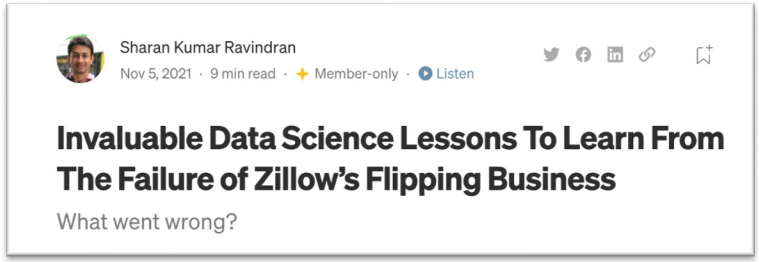
How data scientists spend their time (Image courtesy Anaconda "2020 State of Data Science: Moving From Hype Toward Maturity.")

Les mauvaises données sont pires que l'absence de données

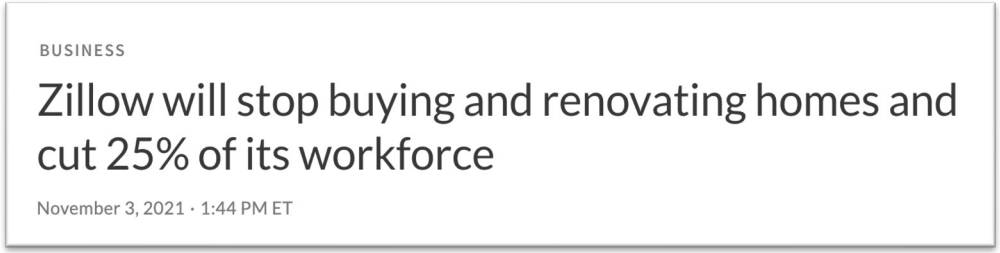
- Zillow
  - Prédire les prix des maisons
  - Achetez des maisons à un prix inférieur aux prévisions
  - Vendre au prix prévu
  - Profit

La précision du modèle a commencé à se degrader

- Problèmes de qualité des données
- Augmentation de l'erreur de prédiction
- **Perte de 300 millions de dollars**



https://towardsdatascience.com/invaluable-data-science-lessons-to-learn-from-the-failure-of-zillows-flipping-business-25fdc218a62



https://www.npr.org/2021/11/03/1051941654/zillow-will-stop-buying-and-renovating-homes-and-cut-25-of-its-workforce

Les mauvaises données sont pires que l'absence de données (cont.)

- Amazon système d'embauche
  - Les entreprises reçoivent des tonnes de CV
  - Utilisez ML pour classer les CV
  - Choisissez les meilleurs CV pour l'entrevue
  - Profit

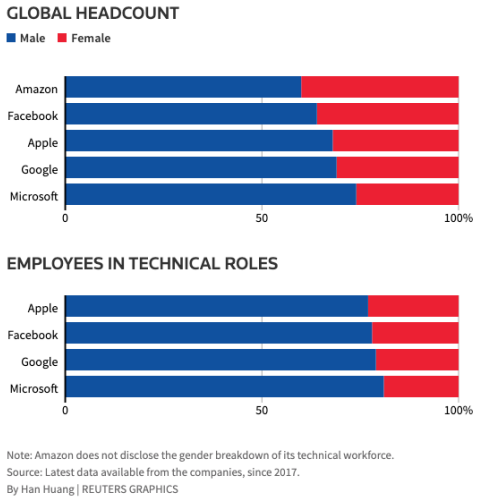
WORLD

Amazon ditches AI recruiting tool that didn't like women

By Jeffrey Dastin · **Reuters**
Posted October 10, 2018 6:46 am

https://globalnews.ca/news/4532172/amazon-jobs-ai-bias/

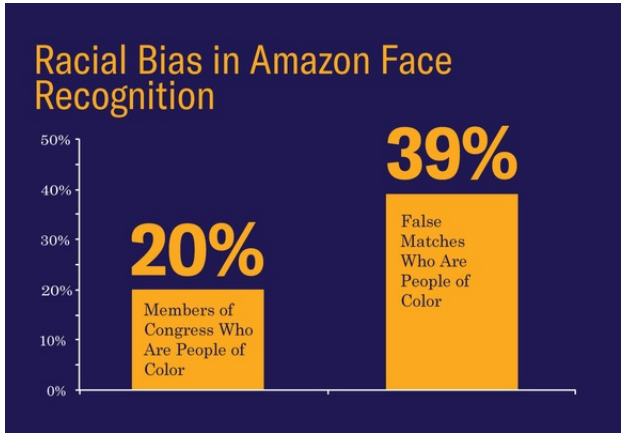
Qualité des données



Les mauvaises données sont pires que l'absence de données (cont.)

- Amazon Rekognition scan
  - Entrez une image
  - Faire correspondre l'image aux bases de données mugshot
  - Profit (?)

Qualité des données

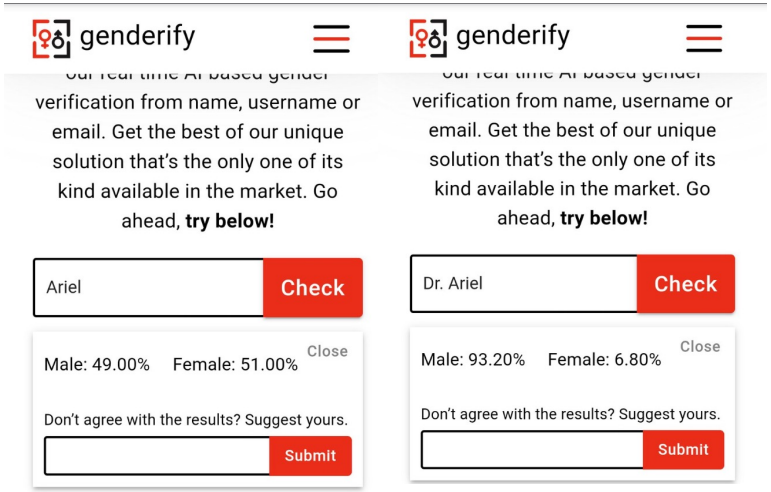


NEWS & COMMENTARY

Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots

https://www.aclu.org/news/privacy-technology/amazons-face-recognition-falsely-matched-28

Plusieurs autres exemples



https://twitter.com/fisadev/status/1288327018522779648/photo/2

https://deepchecks.com/top-10-ml-model-failures-you-should-know-about/

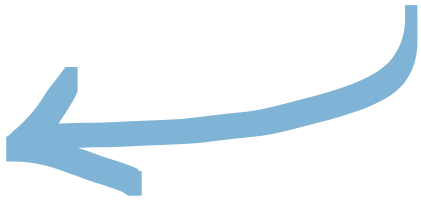
https://www.prolific.co/blog/data-quality-and-ai-safety

Les mauvaises données ne génèrent pas d'erreurs

- Les erreurs logiciels sont identifier pour:
  - Stack traces
  - Error logs
  - etc.

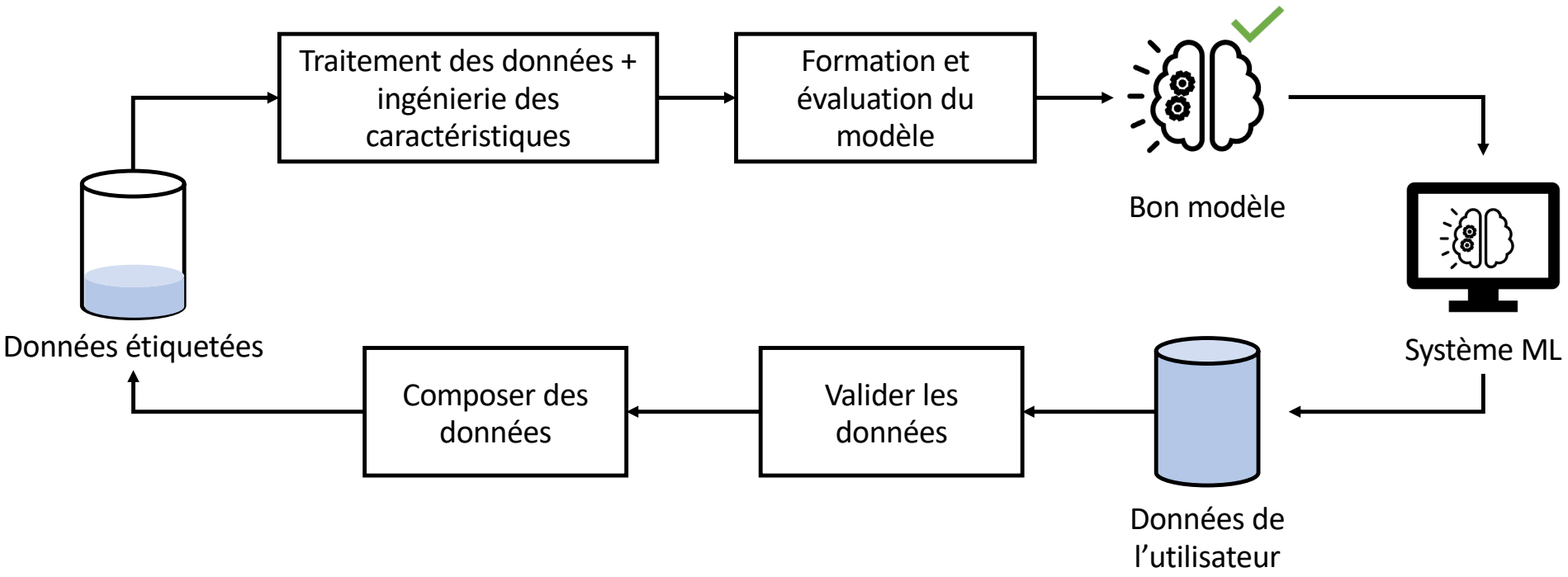
Comment identifier les mauvaises données?

Notre objectif

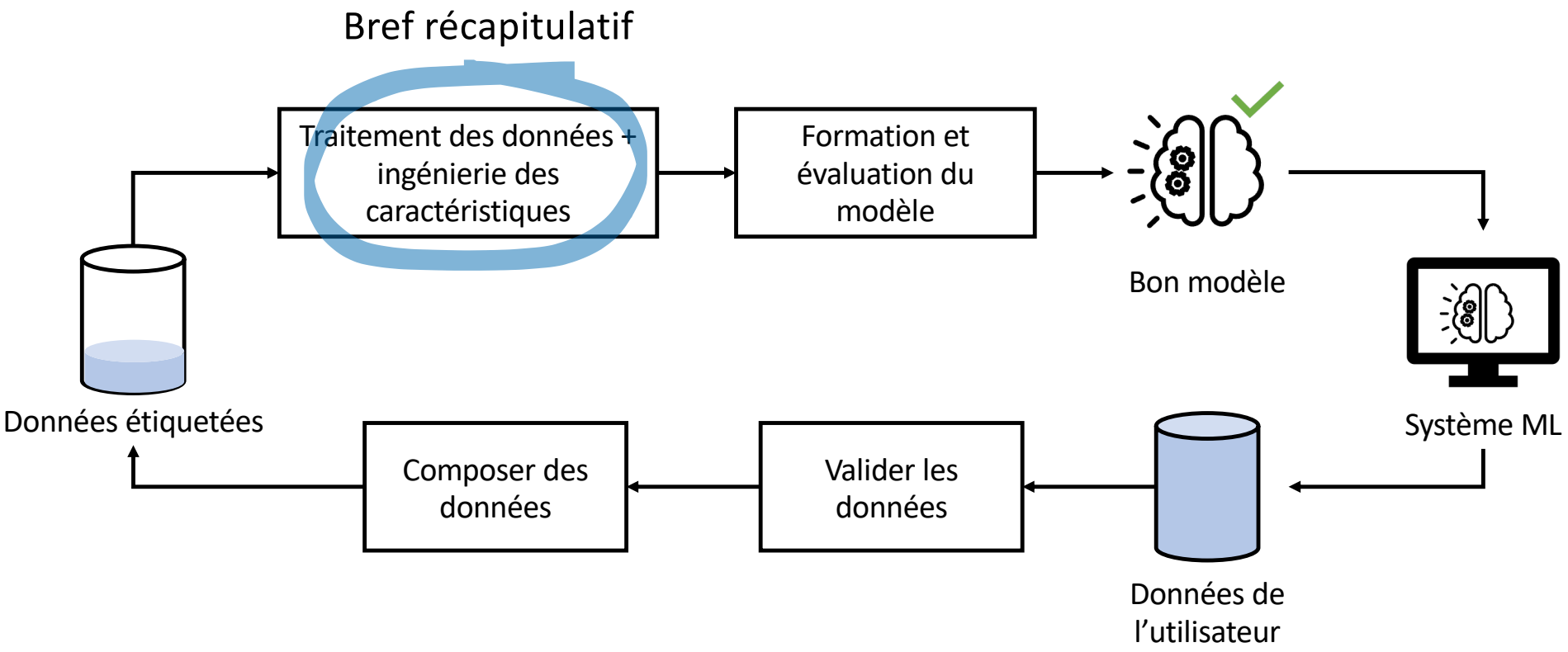


- Le meilleur scénario:
  - Erreurs dans le pipeline de traitement
- Scénario typique:
  - ???

Pipeline de données des systèmes ML



Pipeline de données des systèmes ML



Traitement des données (un bref récapitulatif)

Exploration des données

Identifier les problèmes dans les données:

- Valeurs manquantes
- Distribution asymétrique (possibilité de biais)
- Valeurs non documentées
- Mauvaise couverture des données

Identifier le type de problème:

- Distribution des variables cibles
  - E.g., Problème de classification équilibré ou déséquilibré?

Traitement des données

Problème: Les données brutes sont bruitées

- Plus difficile à modéliser, à prédire et à expliquer

Des solutions:

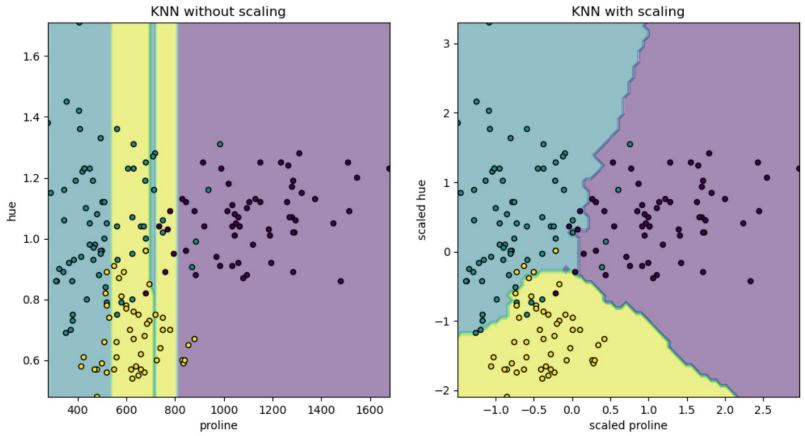
- Gérer les valeurs manquantes
- Anonymiser les données (données sensibles à la confidentialité)
- Supprimer les valeurs aberrantes
- Spécifier la mise en forme des données

Transformation des données

Problème: Les caractéristiques sont présentes dans différentes gammes et distributions

- Affecter la performance des modèles
- Plus difficile à expliquer

- Des solutions
  - Feature scaling
  - Feature normalization
  - Dimensionality reduction



Ingénierie des caractéristiques (*features engineering*)

- **Problème**: Les caractéristiques brutes ne sont peut-être pas idéales pour notre problème de domaine

- Des solutions: Ajouter les connaissances du domaine
  - Regroupez les données
    - E.g., Numerical age into age groups (teenagers, young adult, …)
  - Feature encoding
    - Coder des caractéristiques catégorielles (one-hot encoding)
  - Créer de nouvelles caractéristiques
    - D'après des données supplémentaires

Sélection des caractéristiques

- **Problème**: Toutes les caractéristiques ne sont pas pertinentes pour notre modèle
  - Augmenter la difficulté de maintenir les données

- Des solutions: Supprimer les caractéristiques excessives (Scikit learn methods)
  - Supprimer les entités invariantes
  - Supprimer les entités corrélées
    - Si deux entités sont fortement corrélées, supprimez l'une d'entre elles
  - Sélectionnez les meilleures caractéristiques
    - Utiliser le modèle + la formation et la validation

Division des données (Data Split)

Fractionner les données

- Ensemble de formation et de validation
  - Former et évaluer l'aptitude du modèle
- Ensemble de tests
  - Évaluer le rendement des modèles entraînés

Décision

- Plus de données d'entraînement: mieux pour la performance de la formation
- Plus de données de test: une évaluation plus approfondie

Attention à la division des données

Les données de test **ne doivent jamais** influencer les données d'entraînement

- **Rule of thumb:** diviser les données avant le traitement des données
- Utiliser les paramètres des données d'entraînement pour éclairer la transformation des données de test