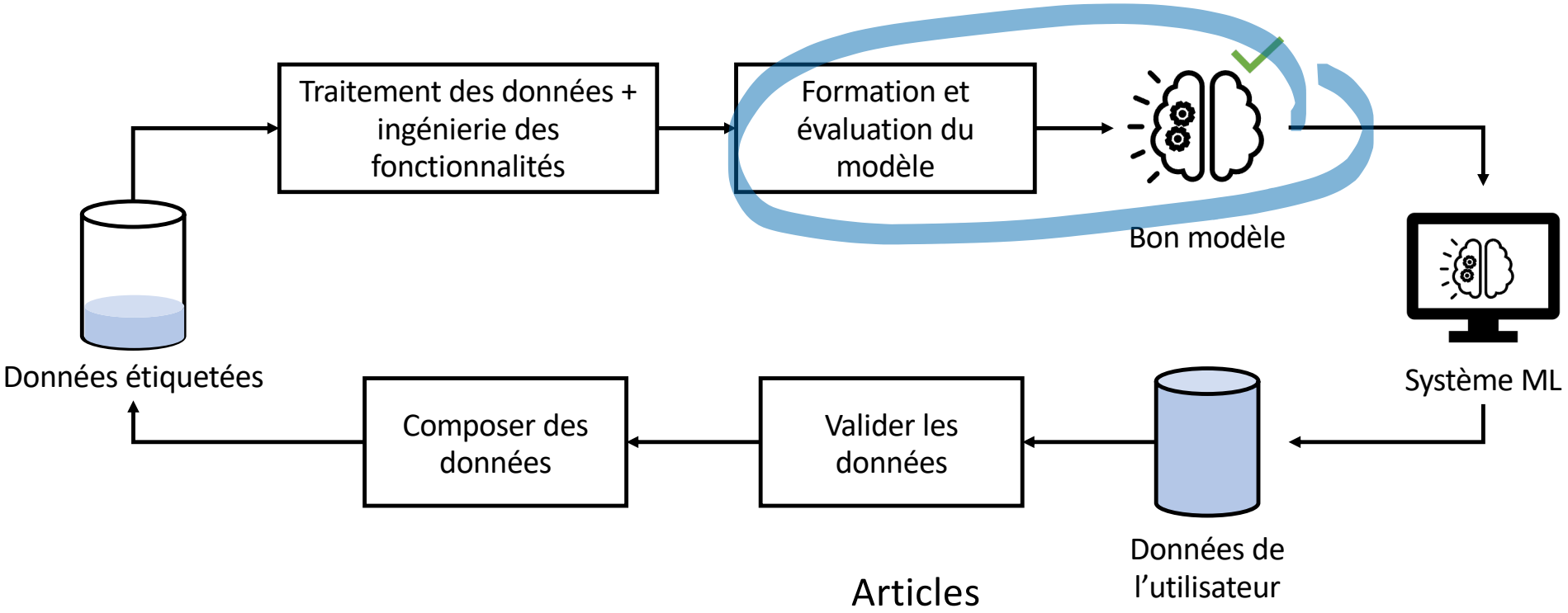
Sélection des modèles

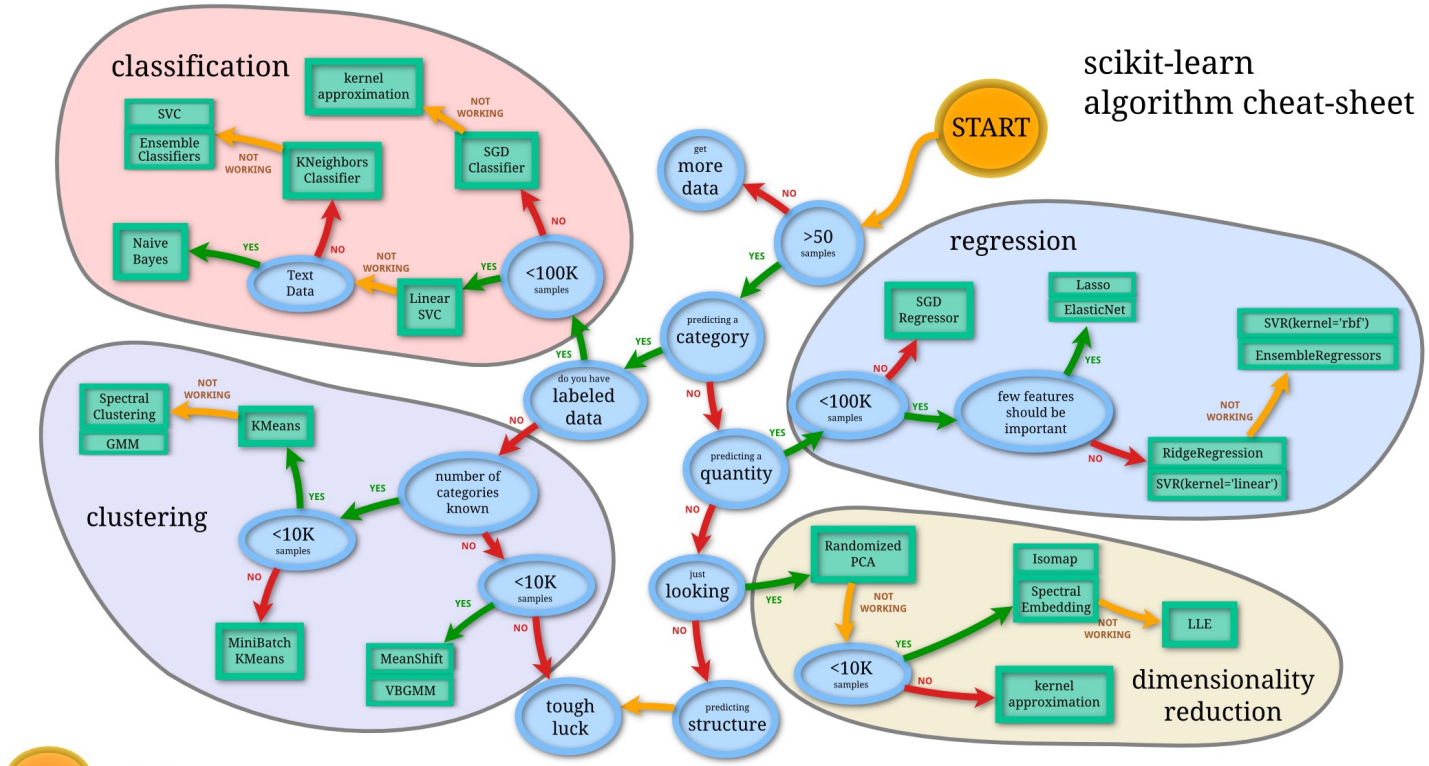
MGL 7320: Ingénierie logicielle des systèmes d'intelligence artificielle

Pipeline de données des systèmes ML



Articles

Comment sélectionner le meilleur modèle?



Certains problèmes choisissent le modèle

Convolutional Neural Networks

Naïve Bayes
Embeddings + Traditional Classifiers
Transformers

Parfois, le modèle est sélectionné en fonction de contraintes

Modèles plus simples avec un temps d'inférence rapide

The sky is the limit

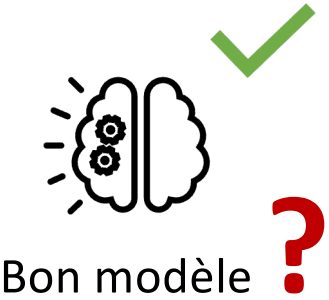
- Modèles d'ensemble
- Modèles classiques
- Apprentissage profond

Considérations de la selection des modèles

- Facteurs à considerer lors de la sélection des modèles (trade-offs)
  - Performance
  - Interprétabilité
  - Robustesse
  - Taille du modèle
  - Temps de formation
  - Temps d'inference

- En pratique, nous essayons de nombreux modèles différents et sélectionnons les meilleurs

Comment évaluer la qualité du modèle?

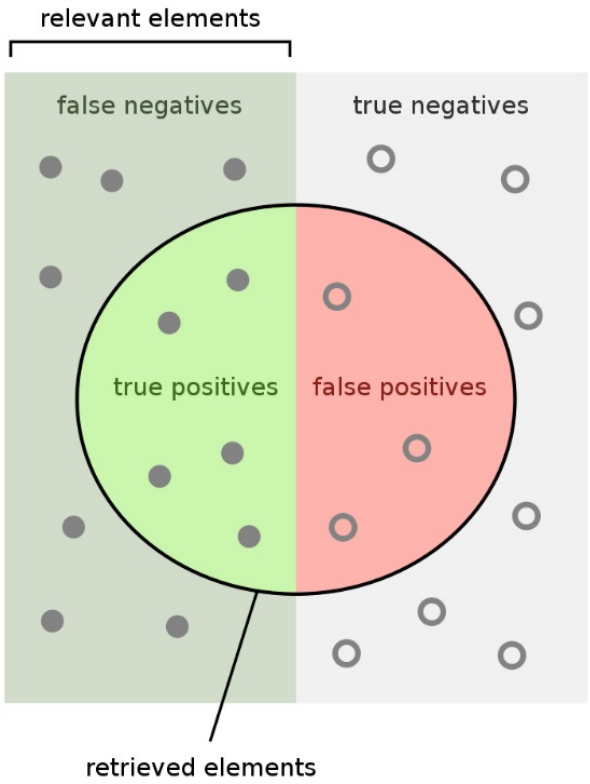


Choisir les mesures d'évaluation appropriés

- Évaluer l'objectif de chaque mesure
  - Sélectionnez celui qui est approprié pour le domaine du système

- Mesures de classification
  - **Exactitude** (*Accuracy*)
  - **Précision** (*Precision*)
  - **Rappel** (*Recall*)
  - **F1-score** (moyenne harmonique entre précision et rappel)
  - **Courbes Précision-Rappel** (*Precision-Recall curves*)
  - **AUC ROC** (*Aire sous la courbe ROC*)

Choisir les mesures d'évaluation appropriés



How many retrieved items are relevant?

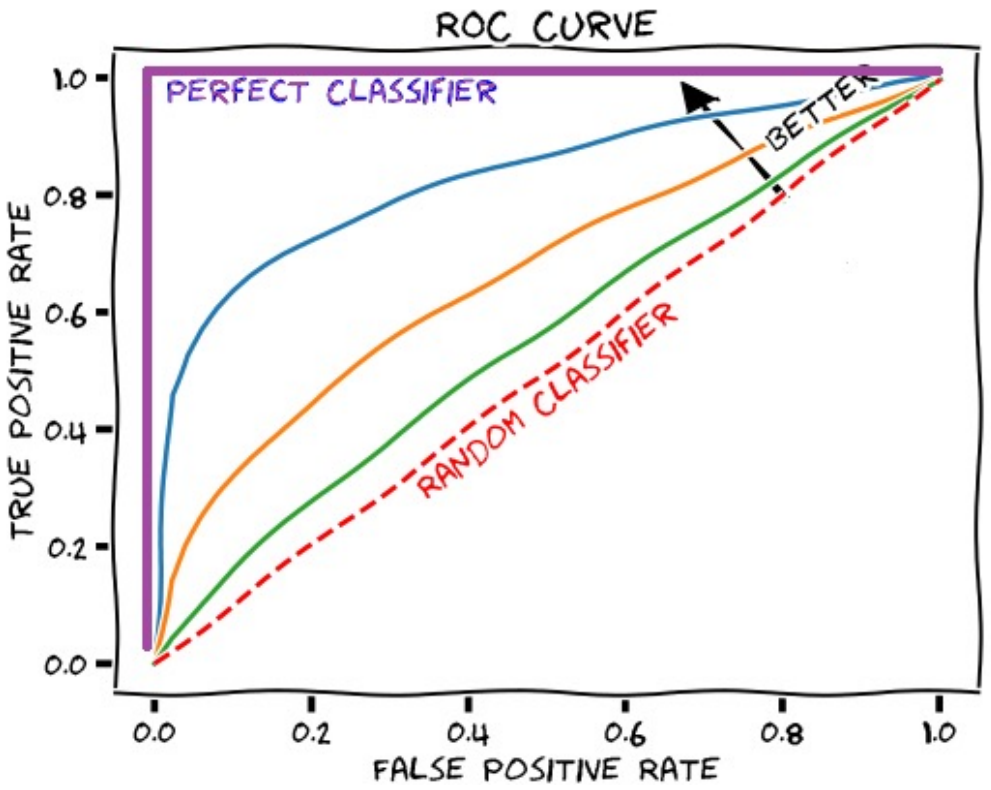
$$\text{Precision} = \frac{\text{true positives}}{\text{retrieved elements}}$$

How many relevant items are retrieved?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

AUC (*Area Under the Curve*)

- **AUC** représente l'aire sous la courbe ROC (***Receiver Operating Characteristic***), utilisée pour évaluer la performance des modèles de classification binaire.
- La courbe ROC trace la **sensibilité** (taux de vrais positifs) en fonction de la **spécificité** (1 - taux de faux positifs).



Sélectionner et optimiser les modèles

- Essayez différents algorithmes de modèle (supervised learning)
  - Logistic Regression
  - Random Forest
  - Support Vector Machines
  - …

- Réglage des hyper-paramètres du modèle (hyper-parameter tuning)
  - Ajuster la taille du modèle
    - Plus grand: plus robuste mais plus de consommation de ressources

À quel point est-ce que c'est assez bon?

- Scénario:
  - "Votre modèle a obtenu une précision de **75%** sur votre test set."

- À quel point est-ce bon?
  - Dépend du domaine
  - Dépend des exigences imposées par le client
  - Dépend des performances des solutions actuelles
  - (…)
- On a besoin d'un moyen d'évaluer les améliorations apportées par l'IA
  - Modèles de base

Modèles de base

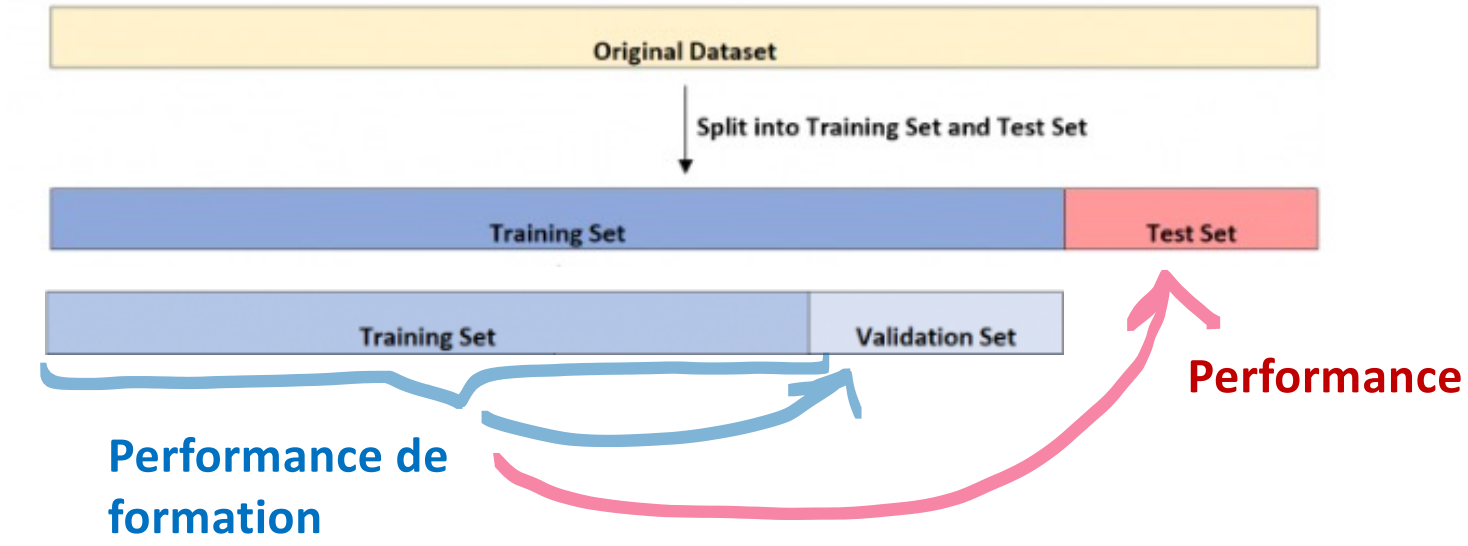
- Utiliser les caractéristiques des données (dummy baseline models).
  - Stratégies: Le plus fréquent, Stratifié, Uniforme, etc.

- Utiliser l'expertise du domaine pour établir un **ensemble de règles**
  - Évaluer la performance de l'ensemble de règles

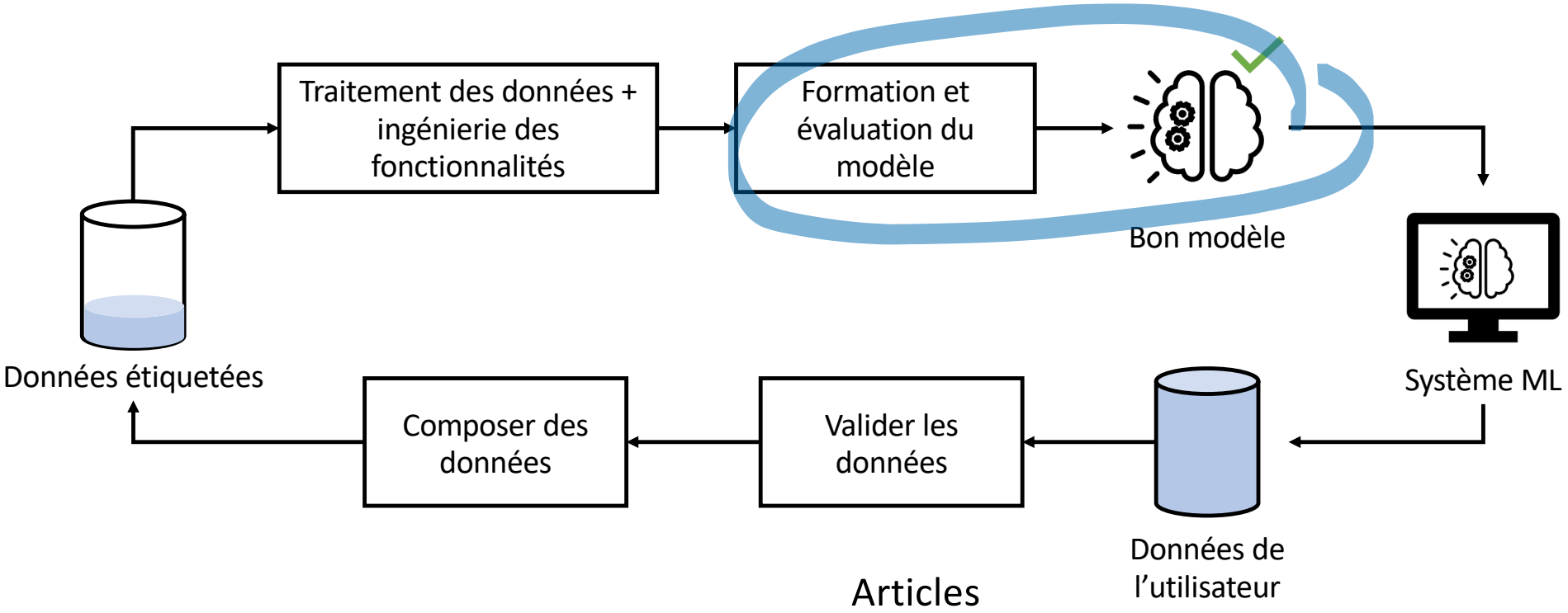
- Comparer le rendement des modèles entraînés à tous les modèles de base
  - "Notre modèle entraîné améliore les performances de classification de **50%**.
    - Notre modèle = 75% | Meilleure base de référence = 50%

Formation et évaluation du modèle

- Deux types de performance
  - Performance de formation (validation set)
  - Performance réelle (test set)



Pipeline de données des systèmes ML



Netflix Recommendations

- Concentrez-vous sur la description des premiers jours de la sélection du modèle Netflix
  - Netflix Prize

Study design

- Il ne s'agit pas d'une étude scientifique

- Rapport de l'industrie
  - Blog
  - Littérature grise
  - Met en valeur l'expérience industrielle
  - Leçons apprises

The Netflix Prize

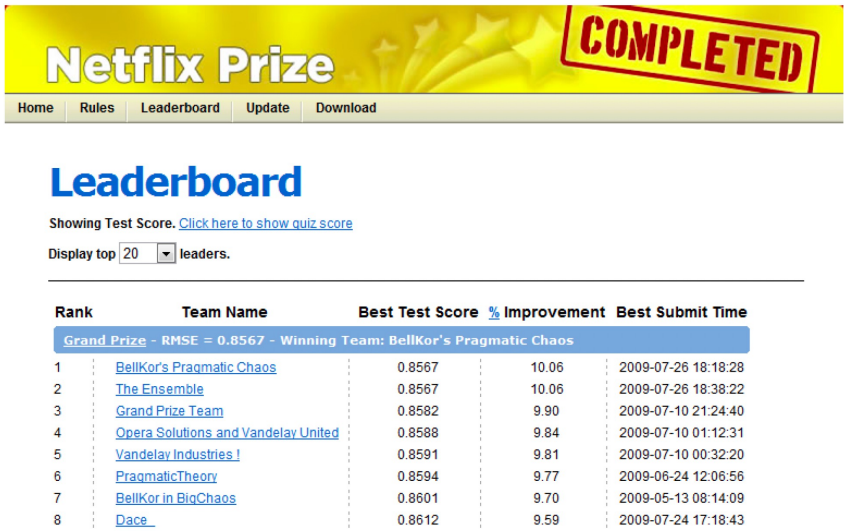
- Concours d'apprentissage automatique et d'exploration de données
  - Prix : 1 million de dollars à quiconque a amélioré la précision de son système (Cinematch) de 10%
  - Prediction des notes données aux films (regression)

La première équipe gagnante

- L'équipe Korbell a gagner le Prix Progress
  - Amélioration de 8.43%
  - Plus de 2000 heures de travail
  - Solution: Ensemble de 107 algorithmes
- Les deux meilleurs algorithmes (RMSE 0.88)
  - Singular Value Decomposition (SVD)
  - Restricted Boltzmann Machines (RBM)
- Grand effort d'ingénierie pour déployer ces modèles à l'échelle de Netflix
  - De 100 mille à 5 milliards de recommandations

L'équipe gagnante finale

- La solution a mélangé des centaines de modèles prédictifs
  - L'effort d'ingénierie pour mettre la solution en production n'en valait pas la peine



Changements dans le paysage

- Netflix a publié son service de streaming
  - Le prix original de Netflix a été prévu pour aider leur système de recommandation de DVD

- Service de streaming
  - **75%** de ce que les gens regardent est basé sur la recommandation

Tout est une recommandation

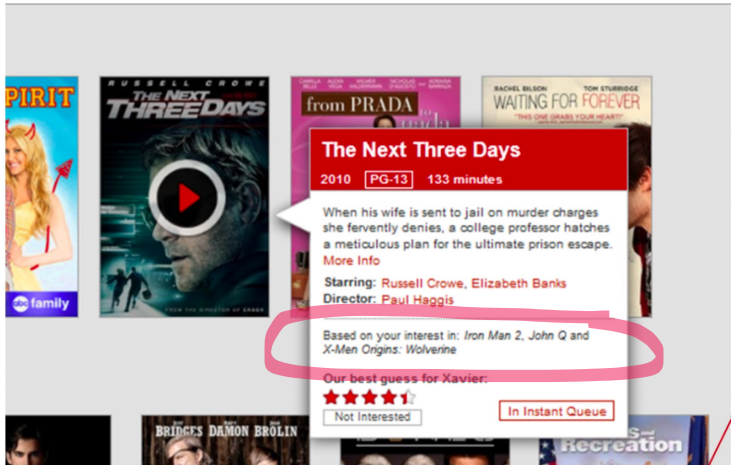
- La recommandation commence par les lignes
  - Groupes de vidéos avec une connexion significative

- E.g., Les 10 vidéos que vous êtes le plus susceptible de regarder

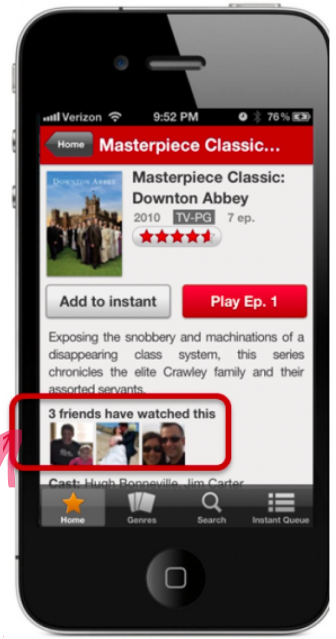
Tout est personnalisé

L’explication est la clé

- Les recommandations doivent fournir une **explication**



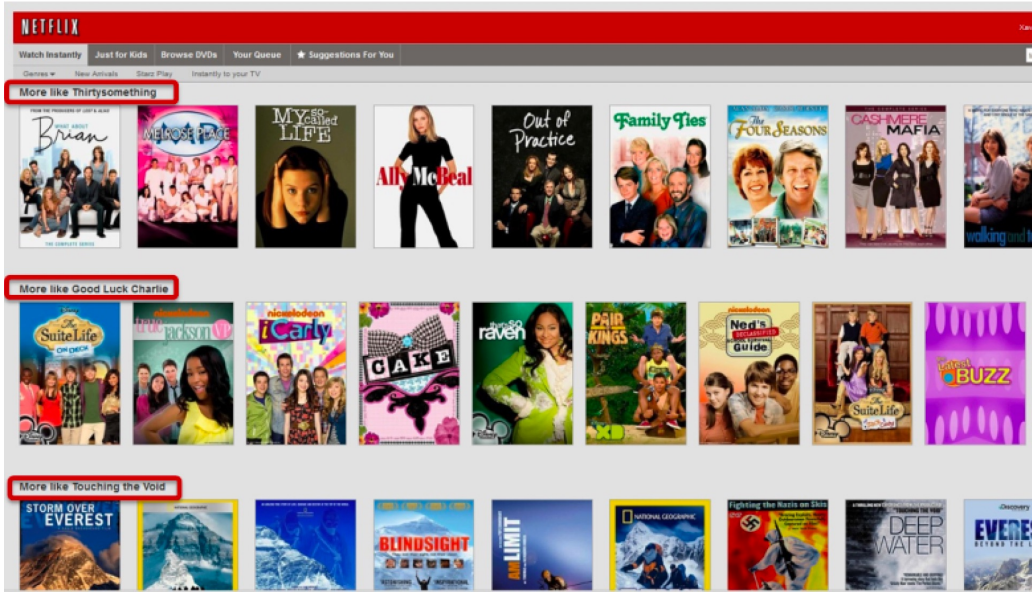
Explication



Soutien social

La similarité est une source importante de personnalisation

- Similarity between movies, members
  - Multiple dimensions: metadata, ratings, viewing data
  - Similarity is blended and **used as features** for models



Le classement est l'un des aspects d'une recommandation efficace

- Le prix Netflix original visait à prédire les notes des film

- Une recommandation efficace doit être prise en compte:
  - Contexte
  - Popularité du titre
  - Intérêts
  - Diversité
  - Fraîcheur
  - …

La popularité n’est pas tout

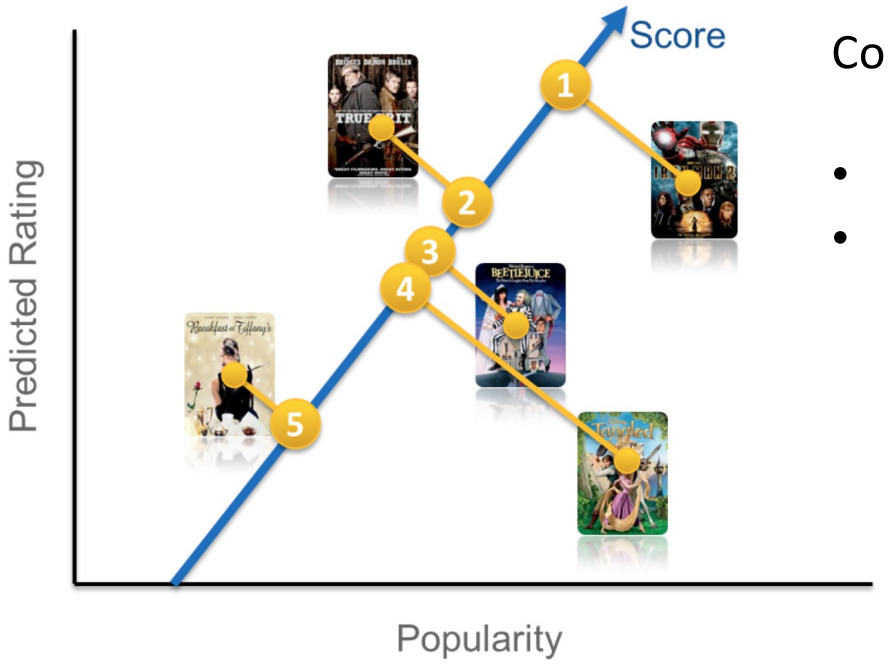
- L’objectif est de recommander les titres que chaque membre est le plus susceptible de jouer et d’apprécier
  - Maximize consumption!
  - Ranking = Scoring + Sorting + Filtering

- Ligne de base : Popularité
  - Des films moins populaires qui ne seront jamais regardés

Équilibrer la popularité et la personnalisation

Popularité Personnalisation

$$\text{frank}(\text{user}, \text{video}) = \mathbf{w1}\ \text{popularity}(\text{video}) + \mathbf{w2}\ \text{rating}(\text{user}, \text{video}) + b$$



Comment trouver les poids?

- A/B testing?
- L'apprentissage automatique!

Sources de données

- Obtenez autant de données (pertinentes) que possible:
  - Des milliards de notes d'articles des membres
  - Popularité de l'article
  - Comportement de l'utilisateur (films regardés, watch list)
  - Métadonnées du titre: acteurs, réalisateurs, genre, ...
  - Données sociales, termes de recherche, …
  - External data: box-office, critic reviews, …

Modèles utilisés chez Netflix

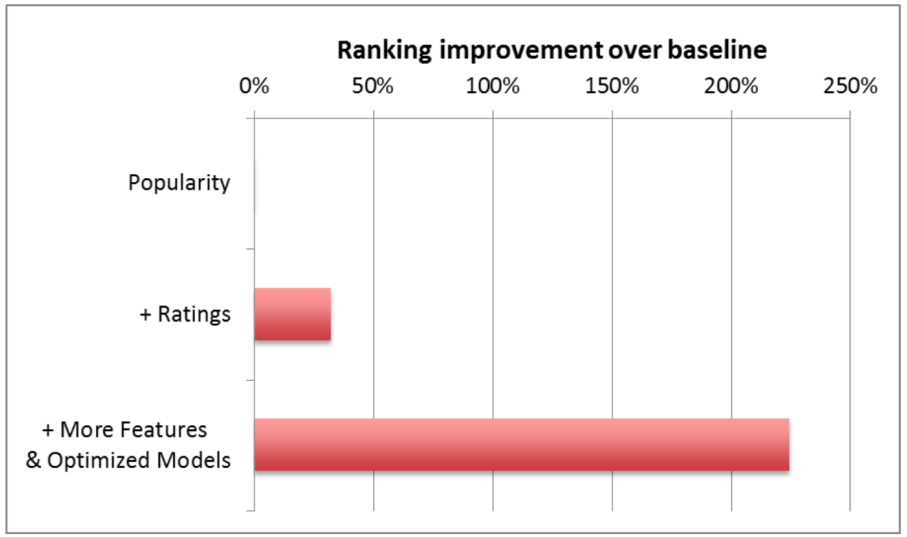
- Compte tenu de la pléthore de données, Netflix utilise plusieurs types de modèles
  - Linear regression
  - Logistic regression
  - Elastic nets
  - Singular Value Decomposition
  - Restricted Boltzmann Machines
  - Association Rules
  - Random Forests
  - Clustering algorithms (k-means)

Utilisez tout ce qui résout leur problème

Amélioration du classement

- Base de référence basée sur la popularité
  - L'utilisation des cotes a déjà amélioré la base de référence d'environ 40 %

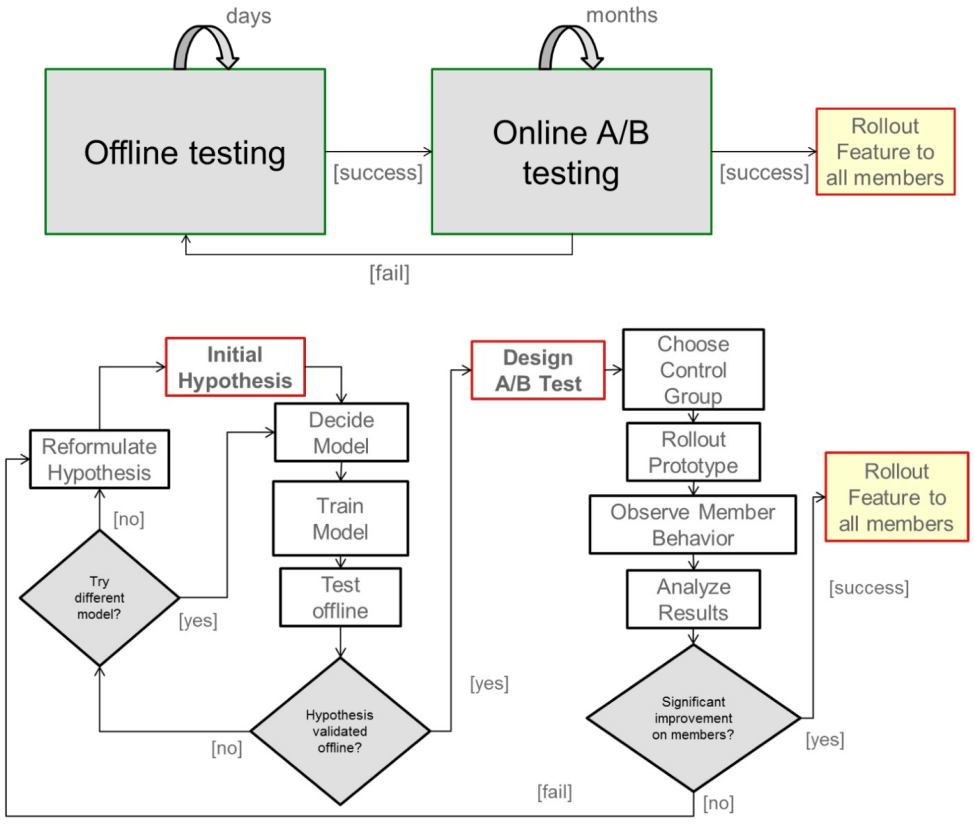
Plus de fonctionnalités + Les modèles optimisés ont amélioré les performances >2x



Approche de la science des données

1. Commencez par une hypothèse
   - Algorithm/feature/ X peut ammelioré l'engagement des utilisateurs des Y%
2. Concevoir un test
   - Prototypage
3. Exécuter le test
4. Laissez les données parler d'elles-mêmes

Offline testing -> A/B testing



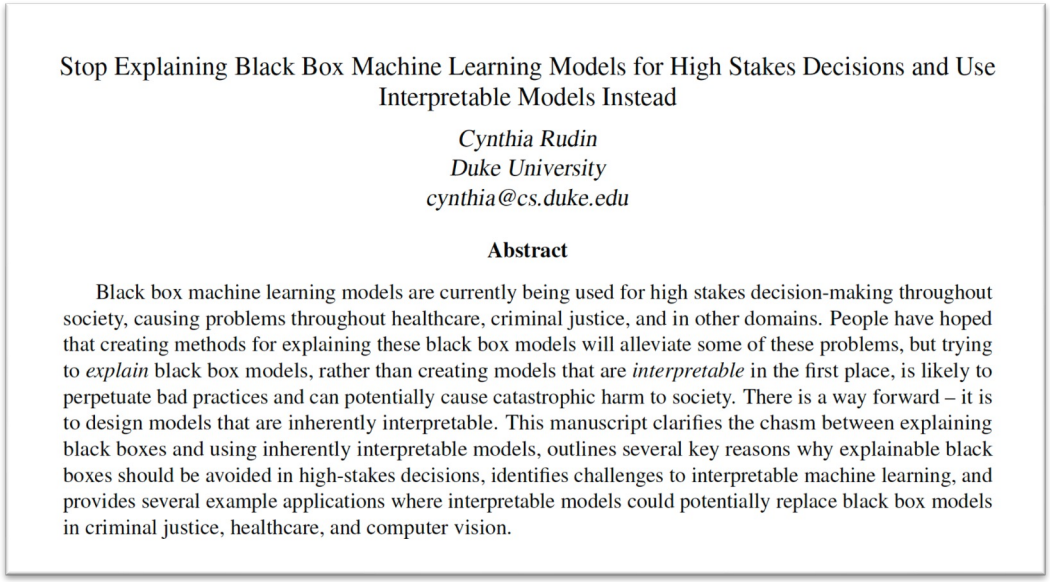
Message à retenir

- Le prix Netflix : prédire les notes du film
  - Pas effectif pour le système de streaming
- La recommandation va bien au-delà des cotes de cinq étoile
  - Maximizer l'engagement d'utilisateur
- Un vrai système utilise souvent:
  - Sources de données multiples
  - De nombreux modèles différents
  - Approche axée sur les données pour faire évoluer le système
    - Offline testing, A/B testing

Use Interpretable Models in High Stakes Decisions

L'auteur fait valoir que :

- Les modèles explicables ne devraient pas être utilisés dans les décisions à enjeux élevés
- Au lieu, nous devrions utiliser des modèles interprétables



Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead

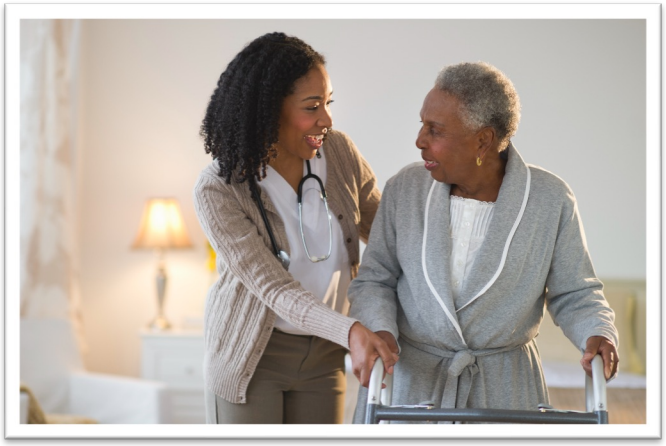
Cynthia Rudin
Duke University
cynthia@cs.duke.edu

Abstract

Black box machine learning models are currently being used for high stakes decision-making throughout society, causing problems throughout healthcare, criminal justice, and in other domains. People have hoped that creating methods for explaining these black box models will alleviate some of these problems, but trying to *explain* black box models, rather than creating models that are *interpretable* in the first place, is likely to perpetuate bad practices and can potentially cause catastrophic harm to society. There is a way forward – it is to design models that are inherently interpretable. This manuscript clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare, and computer vision.

Que sont les décisions à enjeu élevé ?

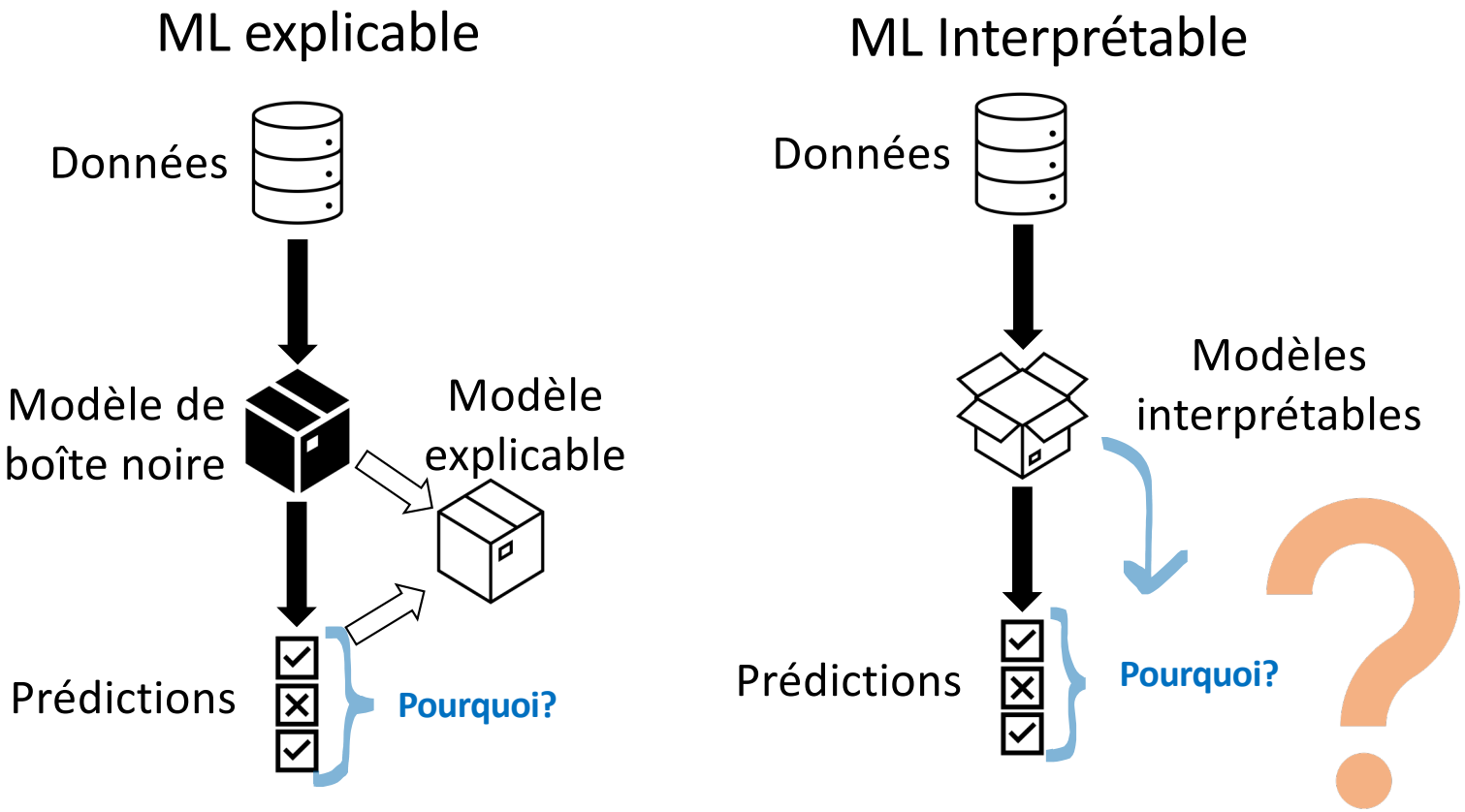
- Applications qui ont un impact profond sur la vie humaine



Santé

Justice pénale

Explicable vs Interpretable modèles



Explicable vs Interpretable modèles (cont.)

- Modèles interprétables
  - Tendance à être des modèles plus simples (e.g., linear, rule-based)
  - Les explications proviennent du modèle
  - Le processus décisionnel est **transparent**

- Modèles explicables
  - Le deuxième modèle (posthoc) est créé pour expliquer le premier modèle de boîte noire plus complexe
  - Les explications sont indirectes
  - Le processus décisionnel est **opaque**

Study Design?

- Article argumentatif (persuasif)
  - Examen des travaux connexes
  - Position d'un expert dans le domaine
  - Anecdotique mais avec des exemples forts

- Publié dans Nature (Machine Intelligence)
  - Depuis 2019, cité près de 3700 fois

Le Problème

- Le manque de **transparence** et de **responsabilisation** des modèles prédictifs a de **graves conséquences**

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

230

Le détenu s'est vu refuser la libération conditionnelle malgré le fait qu'il avait un dossier presque parfait de la réadaptation

En raison d'un système automatisé COMPAS

- Le système était défectueux
- Le détenu devait interjeter appel et prouver que le système était mauvais

https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

Le ML explicable n’est pas suffisant pour les décisions à enjeu élevé

- Les explications ne sont souvent pas fiables et trompeuses

- Nous devrions utiliser le ML interpretable
  - Modèles fondés sur des règles
  - Modèles clairsemés (plus facile d’identifier l’interaction)
  - Obéit à la connaissance structurelle du domaine

Principaux problèmes liés au ML explicable

- Le ML explicable explique un modèle de **boîte noire**
  - Un modèle trop compliqué pour les humains
  - Un modèle propriétaire

Questions clés

1. Mythe du compromis entre l'exactitude et l'interprétabilité
2. Les explications ne sont pas fidèles au modèle original
3. Souvent, les explications n'ont pas de sens
4. Non compatible avec l'évaluation externe des risques
5. Conduit à un chemin de décision trop compliqué

Mythe du compromis entre l'exactitude et l'interprétabilité

- En cas de problèmes avec
  - … des **fonctionnalités significatives** et un
  - … **prétraitement** adéquat des données, …
  - … il n'y a pas de différence significative entre les classificateurs complexes et les classificateurs simples.

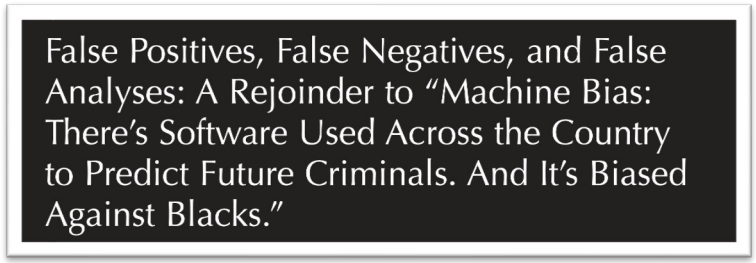
- De petites différences de précision sont souvent **moins importantes** que la capacité à interpréter les résultats.

Le cas du réseau électrique de New York

- Objectif: prévoir les défaillances du réseau électrique à New York
  - Données désordonnées, texte de forme libre, données héritées (1890)
  - Au début, des modèles plus complexes donnent de meilleurs résultats (>1%)
- La capacité d'interpréter les résultats et de réinterpréter les données mène à:
  - Révéler de fausses hypothèses
  - Problèmes de données
- A la fin, la prédiction la plus précise a été faite par des **modèles simples et clairsemés**

Les explications ne sont pas fidèles au modèle

- Les explications **doivent** être fausses (dans une certaine mesure)
  - Sinon, avons-nous besoin du modèle original de la boîte noire?
- Les explications n'utilisent pas (toujours) les mêmes fonctionnalités que les modèles



Utilisation d'un modèle d'explication et accusé COMPAS d'avoir des préjugés raciaux.

Le modèle d'explication était erroné.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
https://www.uscourts.gov/federal-probation-journal/2016/09/false-positives-false-negatives-and-false-analyses-rejoinder

Les explications ne sont pas assez informatives

- Les explications laissent souvent de côté trop d'informations
  - Difficile de comprendre le modèle

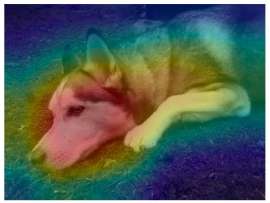
	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps			

Figure 2: Saliency does not explain anything except where the network is looking. We have no idea why this image is labeled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes. Figure credit: Chaofan Chen and [28].

Conduit à un chemin de décision trop compliqué

- Par exemple: COMPAS a 137 facteurs

- Même si une erreur typographique se produit 1% du temps
  - une enquête sur 2 peut avoir des fautes de frappe

https://hdsr.mitpress.mit.edu/pub/7z10o269/release/4

Principaux problèmes liés au ML interprétable

- Les modèles interprétables sont
  - plus simples,
  - et plus faciles à comprendre

- Principaux problèmes liés aux modèles interprétables:
  1. Pas facilement rentable
     - Difficile de protéger la propriété intellectuelle
  2. Nécessite plus d'efforts pour construire
     - Expertise dans le domaine

Les entreprises font des profits sur les modèles de boîte noire

- Les modèles plus simples peuvent être plus facilement copies

- Un autre exemple de l'affaire COMPAS :

Beaucoup plus simple + Performances similaires

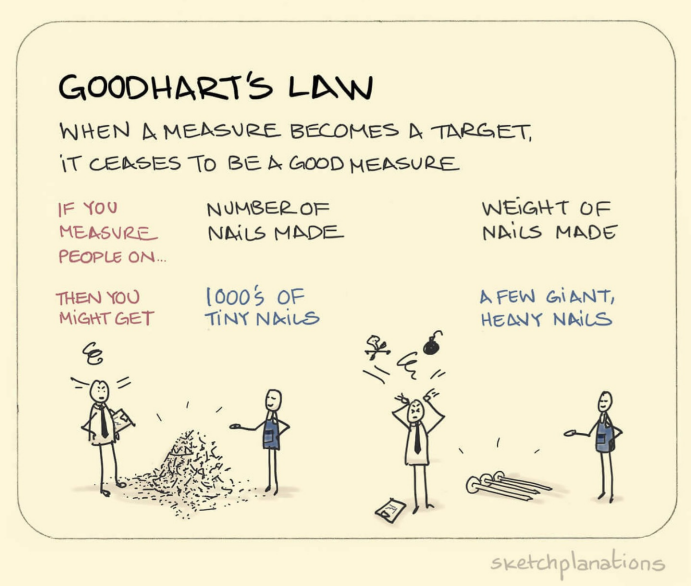
COMPAS	CORELS
black box 130+ factors might include socio-economic info expensive (software license), within software used in U.S. Justice System	full model is in Figure 3 only age, priors, (optional) gender no other information free, transparent

Table 1: Comparison of COMPAS and CORELS models. Both models have similar true and false positive rates and true and false negative rates on data from Broward County, Florida.

Les modèles de boîte noire sont plus difficiles à gamifier

- Les modèles facile sont plus facile à faire l'ingenierie reverse.
  - Un système **mal défini** peut en souffrir

Un système transparent **aidera** les utilisateurs à s'aligner véritablement sur l'objectif global d'amélioration



Les modèles interprétables sont plus difficiles à construire

- L'expertise du domaine est nécessaire pour construire les définitions de **l'interprétabilité** du domaine
  - Il a tendance à être plus facile d'utiliser des modèles de boîte noire pour résoudre des problèmes de calcul difficiles

- Pour les décisions à enjeux élevés
  - Une prédiction erronée **coûte plus cher** que le temps de l'analyste

Encourager la gouvernance d'IA

- **Proposition I** : **Aucune boîte noire** ne devrait être déployée lorsqu'il **existe** un modèle interprétable avec des **performances similaires**
  - Considérée comme de la fausse publicité

- **Proposition II** : Les modèles de boîte noire devraient reporter la performance.
  - Accuracy

Conclusion

- Remettre en question l'hypothèse:
  - les modèles de boîte noire **sont nécessaires** pour des prédictions précises

- Encourager les décideurs **à ne pas accepter** les solutions de boîte noire sans tenter des **variantes interprétables**

- Sensibiliser aux problèmes des modèles interprétables