

INTELLIGENCE ARTIFICIELLE

Intelligence Artificielle • Machine Learning • Deep Learning
Introduction didactique

DÉMARRER PRESENTATION



MACHINE LEARNING

DATA SCIENCE

Sources et qualité des données

JEUX DE DONNÉES LIBRES

Introduction

- Les **jeux de données** font partie intégrante de l'apprentissage automatique. Sans jeux de données, les algorithmes d'apprentissage automatique **ne pourraient pas apprendre** à analyser du texte, à faire de la classification ou de la catégorisation de produits.
- Il y a cinq ou dix ans, il était **très difficile** de trouver des jeux de données adaptés pour le machine ou le deep learning.
- Aujourd'hui, les **jeux de données sont légion** et le problème n'est pas d'en trouver un, mais plutôt de choisir parmi l'offre pléthorique les plus pertinents.
- Nous allons présenter une liste de sources de données libres pour l'apprentissage automatique parmi les catégories suivantes :

- Apprentissage automatique classique
- Traitement du langage naturel
- Finance
- Vision par ordinateur
- Analyse de sentiments
- Apprentissage profond
- Données géométriques
- Données temporelles

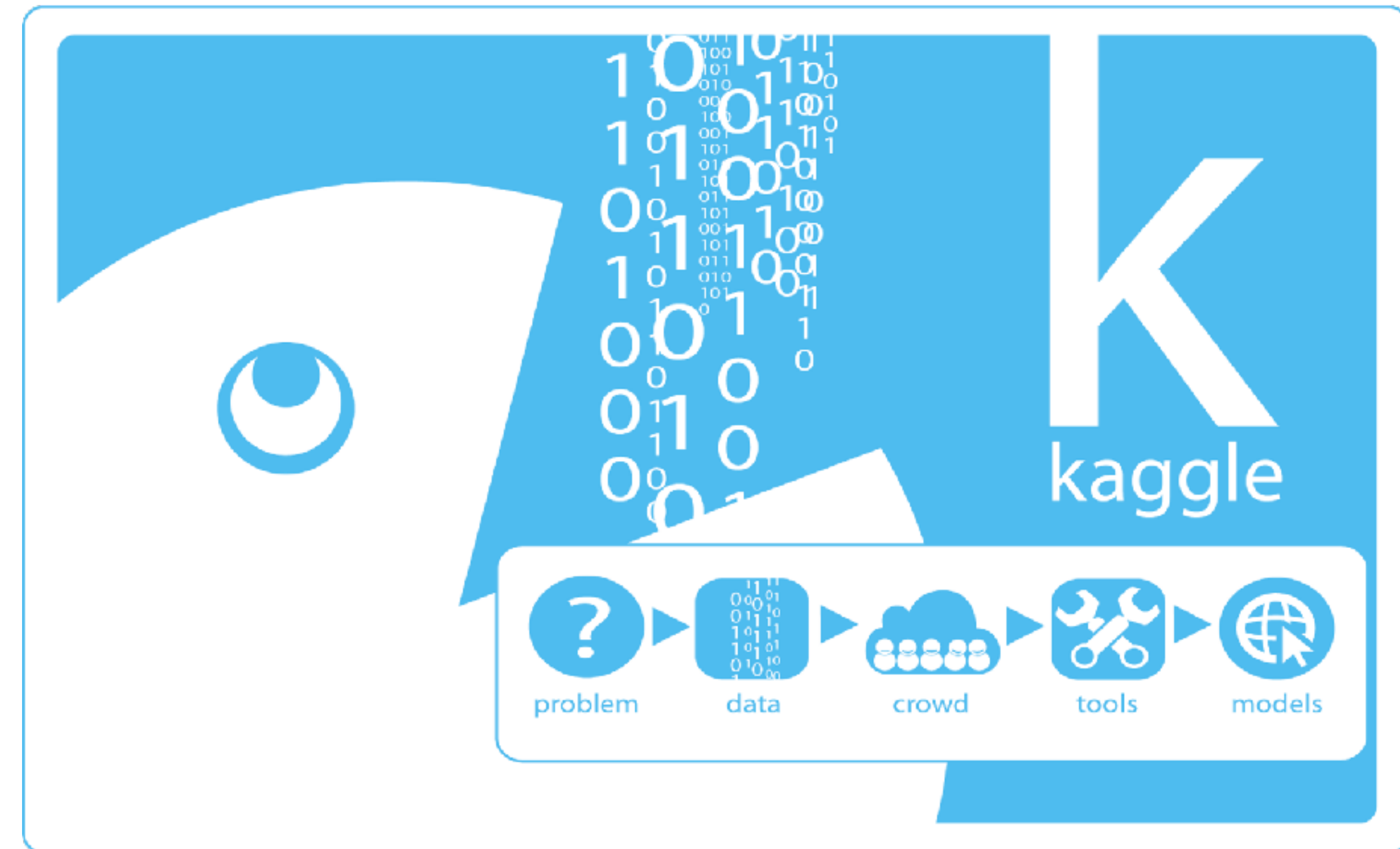


Kaggle

JEUX DE DONNÉES LIBRES

Les bases de données de Kaggle

- **KAGGLE**, appartient à Google, est une communauté en ligne pour les spécialistes en **Data Science** et **Machine Learning**.
- Kaggle permet aux utilisateurs de trouver et de publier des **ensembles de données**, d'explorer et de construire des modèles dans un environnement de sciences des données basé sur le Web, de travailler avec d'autres scientifiques des données et des ingénieurs en apprentissage automatique, et de participer à des concours pour relever des défis en sciences des données.
- Kaggle a vu le jour en proposant des concours d'apprentissage automatique et propose désormais également une **plate-forme de données publiques**, un atelier Cloud pour la science des données et une plate-forme de **formation** en **Intelligence Artificielle**.
- Kaggle contient (mi-novembre 2020) plus de **60 000 bases de données** pour l'apprentissage machine et plus spécifiquement le **Deep Learning**. Pratiquement, tous les domaines sont couverts puisque le nombre de bases augmente au fur et à mesure des nouvelles compétitions et **retours d'expérience**.



QUALITÉ DES DONNÉES - LE BIAIS

La question des données biaisées

- Les biais peuvent essentiellement advenir de deux manières :
 1. d'abord si le codeur, souvent homme, blanc et cisgenre, laisse transparaître ses **propres préjugés** dans les paramètres qu'il établit pour une IA ;
 2. ensuite et de manière plus pernicieuse, si un algorithme de machine learning est entraîné sur un **échantillon d'exemples lui-même biaisé**.
- Les premiers programmes de reconnaissance faciale **identifiaient mal** les individus à peau noire pour deux raisons principales : les critères explicites ne prenaient pas en compte les contrastes de couleur et les données d'entraînement représentaient **essentiellement des blancs**.
- La sociologue française Angèle Christin, enseignante à Stanford, a constaté, explique Koenig, combien les **algorithmes prédictifs** utilisés par la justice pénale pour évaluer les risques posés par les prévenus sont **biaisés contre les Noirs** à tous les stades de la procédure – quand il faut déterminer la probabilité de ne pas se présenter au tribunal, de récidiver de manière violente, ou d'enfreindre les règles d'une libération conditionnelle.
- De ce point de vue, l'IA apparaît comme une force éminemment **conservatrice**, jugeant l'avenir en fonction du passé¹.

1. On parle de biais des « données-rétroiseur ».

QUALITÉ DES DONNÉES - LE BIAIS

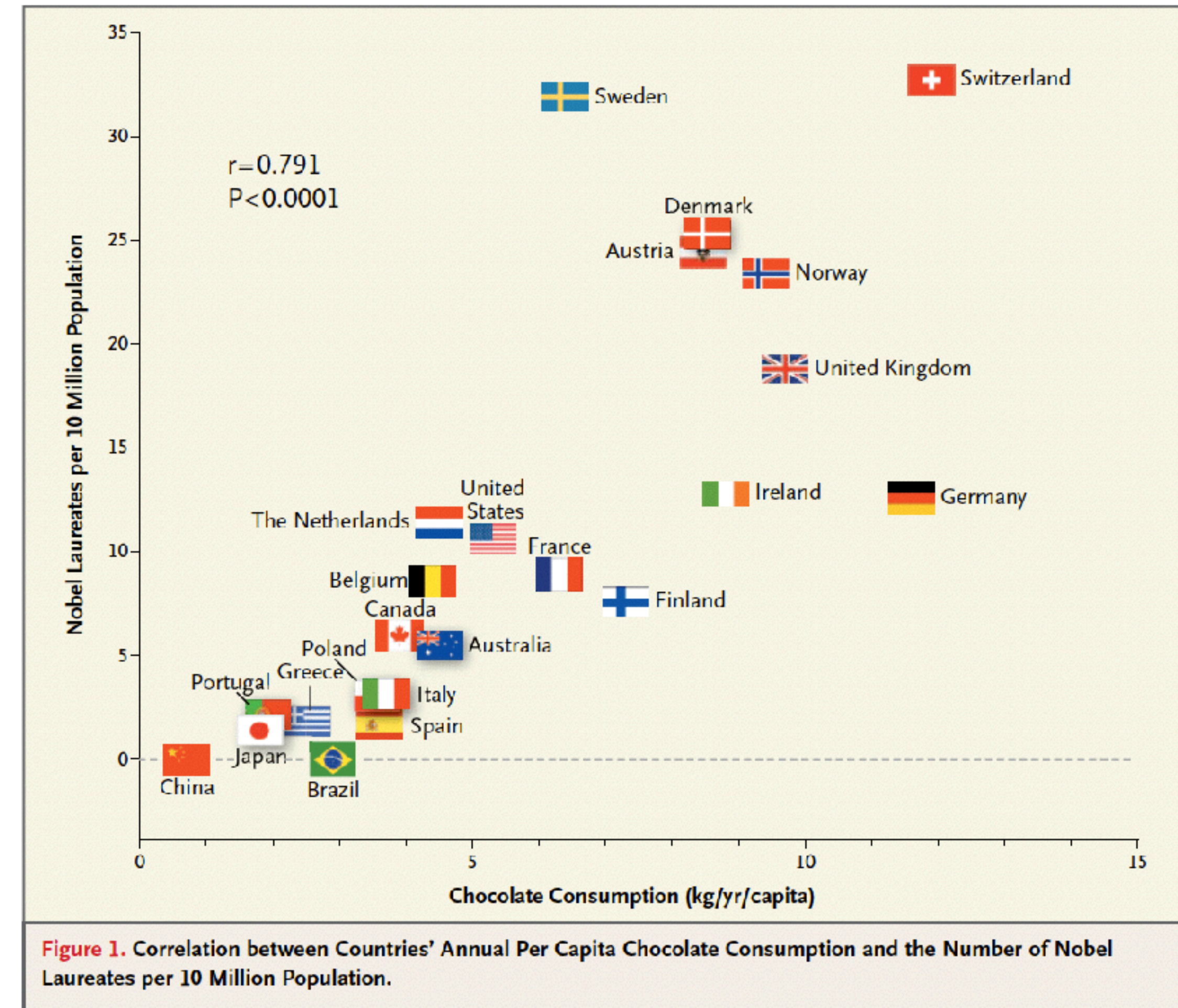
La question des données biaisées

- Le système d'assistance au recrutement d'Amazon avait tendance à **défavoriser les femmes candidates**. La raison ?
- Les données d'entraînement ne comportaient pas **suffisamment de femmes** (elles sont moins représentées dans les domaines techniques).
- Le modèle, entraîné sur une sous-représentation de données féminines, avait tendance à **diminuer les chances** des femmes d'être sélectionnées.
- Pour éviter l'**effet rétroviseur** des données du passé, il est indispensable de disposer de données d'entraînement **équilibrées** et **équitables** (au mieux) selon les critères jugés pertinents.
- Le défaut des modèles d'apprentissage est, d'un certain point de vue, de refléter un peu trop **le monde tel qu'il est** et pas suffisamment celui qui est **souhaitable**.
- Tendre vers le monde souhaitable ne peut être obtenu qu'en injectant un **biais explicite** dans le choix des données afin de contrebalancer un **biais historique et implicite**.

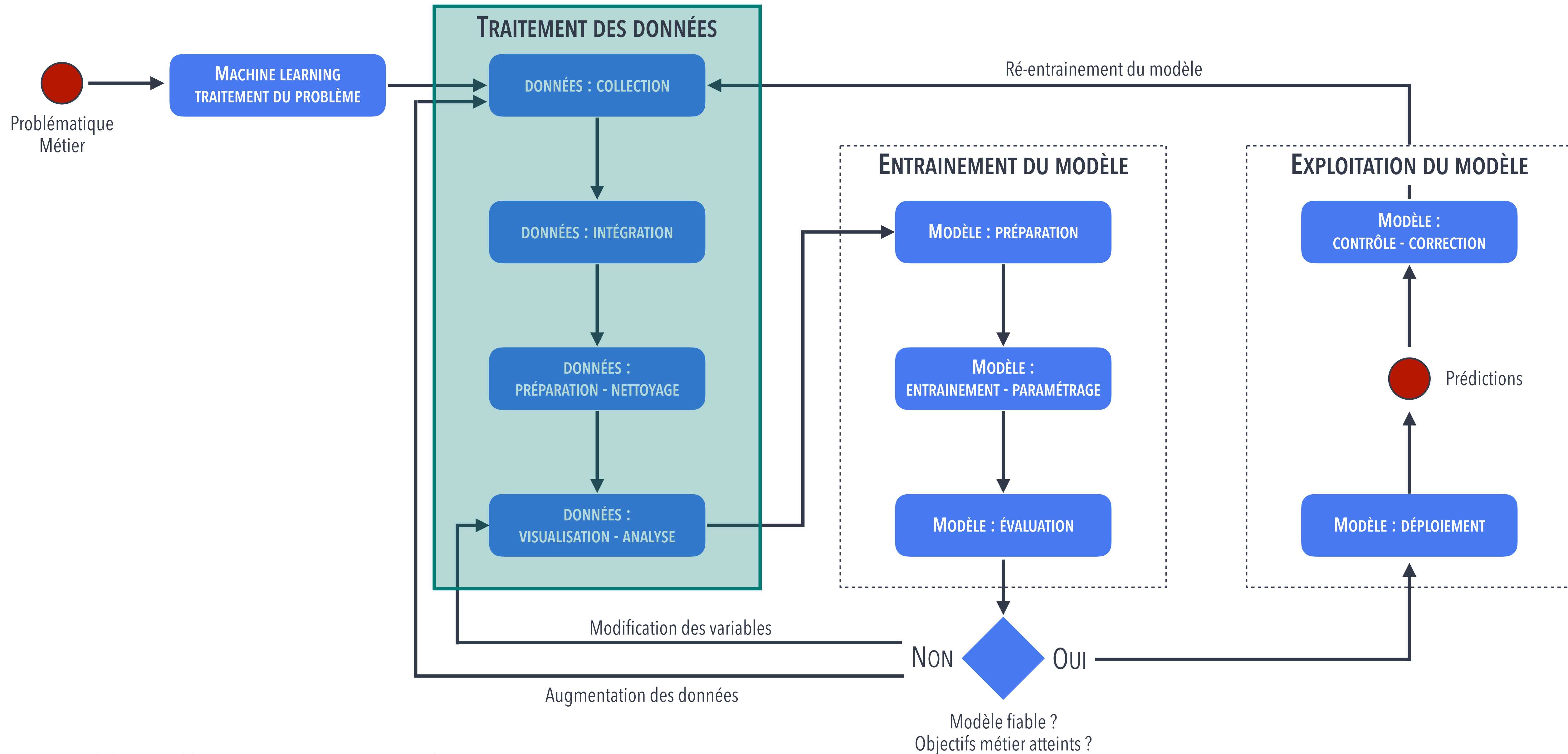
QUALITÉ DES DONNÉES - LE BIAIS

La question des données biaisées

- *Corrélation n'est pas causalité*. Le biais dit de **corrélation** confond les deux.
- Les exemples sont légions ; on peut trouver une corrélation entre la consommation de **chocolat** dans un pays et son nombre de prix **Nobel** par habitant¹.
- Un modèle entraîné uniquement sur des données de consommation de chocolat et de prix Nobel identifiera ce **type de corrélation** alors qu'il existe des **dizaines d'autres paramètres** qui les influencent et qui sont plus pertinents.



QUALITÉ DES DONNÉES - LE BIAIS

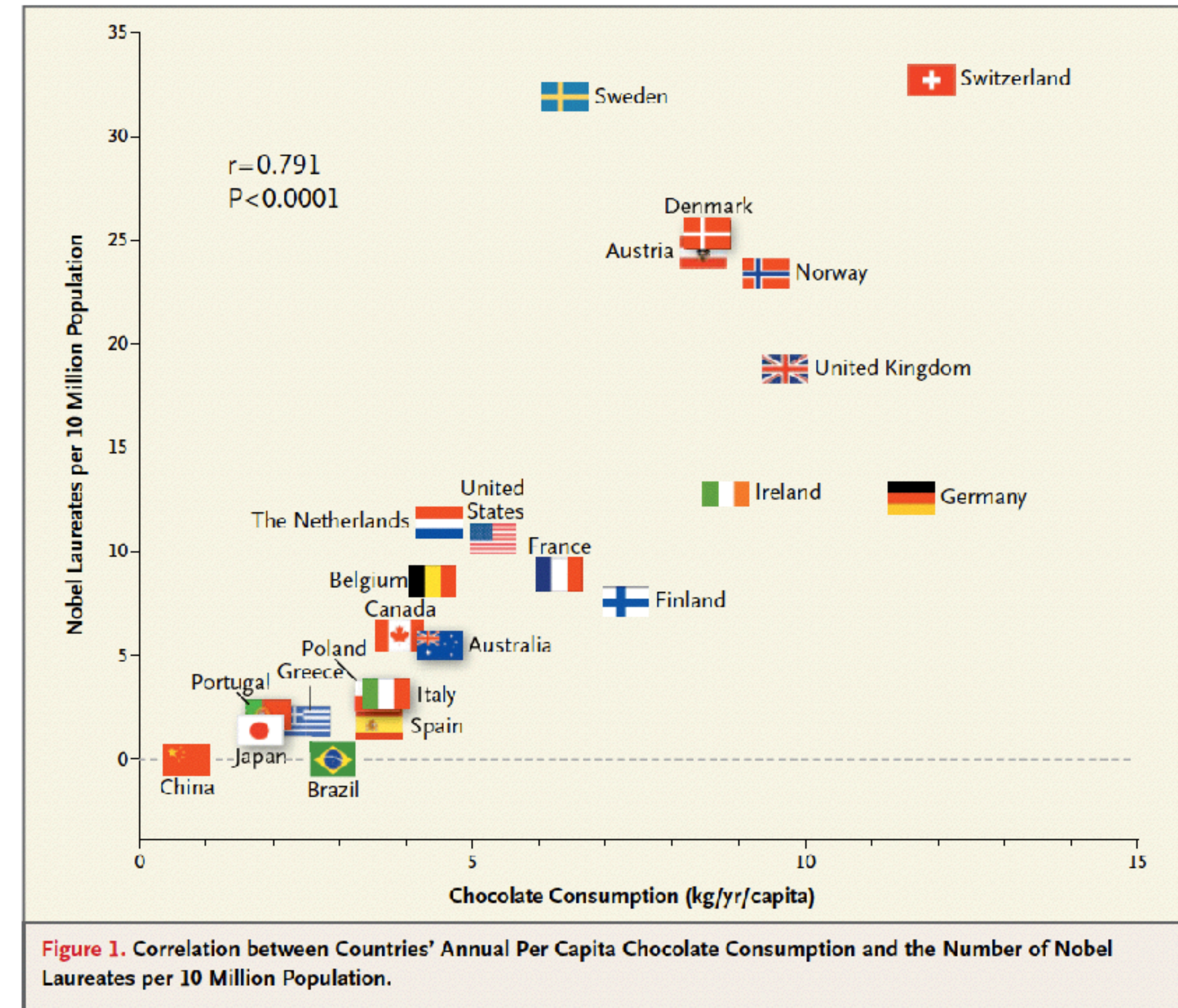


QUALITÉ DES DONNÉES - LE BIAIS

La question des données biaisées

- *Corrélation n'est pas causalité*. Le biais dit de **corrélation** confond les deux.
- Les exemples sont légions ; on peut trouver une corrélation entre la consommation de **chocolat** dans un pays et son nombre de prix **Nobel** par habitant¹.
- Un modèle entraîné uniquement sur des données de consommation de chocolat et de prix Nobel identifiera ce **type de corrélation** alors qu'il existe des **dizaines d'autres paramètres** qui les influencent et qui sont plus pertinents.
- Une **analyse multiparamètres** sera à coup sûr beaucoup plus pertinente.


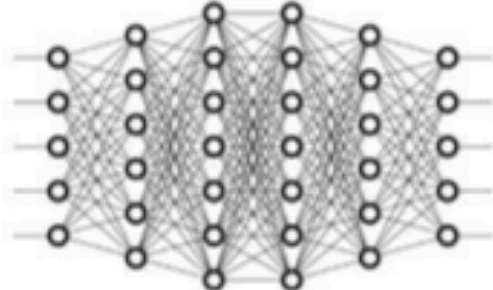
Le choix de la structure des données mises en jeu est essentiel



QUALITÉ DES DONNÉES - LE BIAIS

La question des données biaisées

- Les biais de **généralisation** et de **disponibilité** proviennent d'une insuffisance des données.
- Les biais d'**échantillonnage**, de **confirmation** et de **corrélation** et les biais reflètent les biais humains.

	exemples de biais		
« le manager qui a croisé un client et extrapole son feedback à tous les clients »	biais d'échantillonnage	base d'entraînement de visages blancs ou historique client mono-pays	
« tous les politiciens sont pourris » « les médias mentent »	biais de généralisation	base d'entraînement pas assez grande et ne couvrant pas bien l'espace du possible	
consulter uniquement les sources d'information conformes à nos croyances et opinions	biais de confirmation	Amazon et le recrutement de femmes dans les fonctions techniques, passé « non satisfaisant »	
stéréotypes d'association entre un individu et ses groupes d'appartenance	biais de corrélation	très courants en machine learning, dans la santé, dans le marketing	
on favorise les événements récents, manque d'approche historique	biais de disponibilité	base avec un historique trop limité	

**MERCI DE VOTRE
ATTENTION**

Q & R

