

# **Biais des données : Diagnostic, Symptômes et Remèdes**

---

Yannick Guyonvarch et Nathan Noiry  
datacraft, le 20 mai 2021.

## Au programme

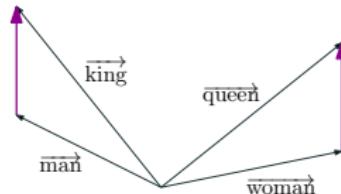
- Qu'entend-on par biais ?
- D'où viennent les biais ?
- Peut-on détecter un biais ?
- Peut-on corriger un biais ?

**Qu'entend-on par biais ?**

# Un exemple en NLP

Les **words embeddings** capturent certaines relations sémantiques...

mot  $\longrightarrow$  vecteur  $\in \mathbb{R}^d$

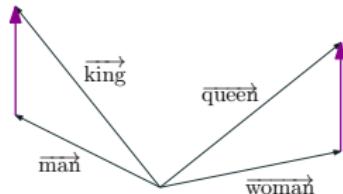


$$\overrightarrow{\text{woman}} + (\overrightarrow{\text{king}} - \overrightarrow{\text{man}}) = \overrightarrow{\text{queen}}$$

# Un exemple en NLP

Les **words embeddings** capturent certaines relations sémantiques...

mot  $\longrightarrow$  vecteur  $\in \mathbb{R}^d$



$$\overrightarrow{\text{woman}} + (\overrightarrow{\text{king}} - \overrightarrow{\text{man}}) = \overrightarrow{\text{queen}}$$

... mais peuvent aussi reproduirent des stéréotypes de genre :

**Man is to Computer Programmer as Woman is to Homemaker ?**

**Debiasing Word Embeddings.**

Bolukbasi, Chang, Zou, Saligrama, Kalai ; 2016, NeurIPS.

$$\overrightarrow{\text{woman}} + (\overrightarrow{\text{computer programmer}} - \overrightarrow{\text{man}}) = \overrightarrow{\text{homemaker.}}$$

## Un exemple en Computer Vision

**Vérification en reconnaissance faciale** : prédire si deux images de visages correspondent à la même personne ou pas.

Exemples d'application : Contrôle d'identité dans les aéroports.

# Un exemple en Computer Vision

**Vérification en reconnaissance faciale** : prédire si deux images de visages correspondent à la même personne ou pas.

Exemples d'application : Contrôle d'identité dans les aéroports.

Deux types d'erreur peuvent se produire :



→ “non”

- Faux négatifs : prédire “non” pour une paire d'images du même individu.



$$FNMR = \frac{\text{Prédits négatifs}}{\text{Vraies paires}}.$$



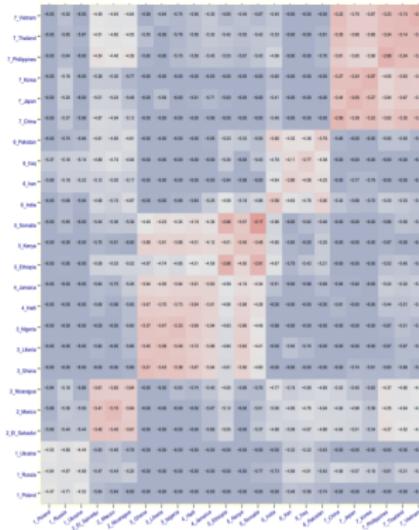
→ “oui”

- Faux positifs : prédire “oui” pour une paire d'images de deux individus différents.

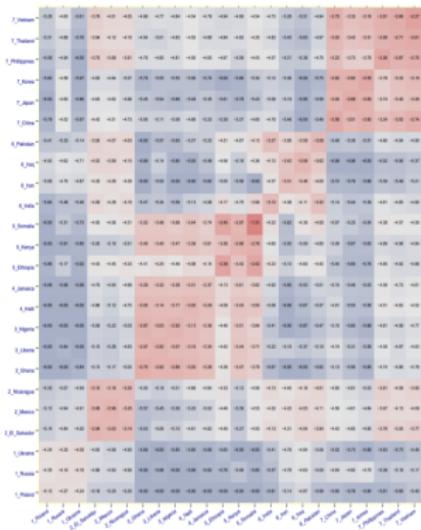


$$FMR = \frac{\text{Prédits positifs}}{\text{Fausses paires}}.$$

# Un exemple en Computer Vision



False Matching Rate on Men.



False Matching Rate on Women.

**Face Recognition Vendor Test, Part 3 : Demographics Effects.**  
Grother, Ngan, Hanaoka ; 2019, NIST.

~ Différentes performances selon les sous-groupes de la population.

## Un exemple en économie

**Enquêtes sectorielles annuelles menées chaque année par l'INSEE auprès de 160 000 entreprises françaises.**

Objectif : dresser un panorama de la santé économique des entreprises.

## Un exemple en économie

**Enquêtes sectorielles annuelles menées chaque année par l'INSEE auprès de 160 000 entreprises françaises.**

Objectif : dresser un panorama de la santé économique des entreprises.

Le taux de réponse est très variable selon la taille des entreprises :

- Les plus grandes entreprises sont toutes interrogées et répondent massivement.
- Les plus petites sont tirées aléatoirement. Parmi celles retenues, le taux de réponse est faible  $\approx 55\%$  !

~~> Très peu probable que la non-réponse n'ait pas d'impact.

## Un exemple en économie

**Enquêtes sectorielles annuelles menées chaque année par l'INSEE auprès de 160 000 entreprises françaises.**

Objectif : dresser un panorama de la santé économique des entreprises.

Le taux de réponse est très variable selon la taille des entreprises :

- Les plus grandes entreprises sont toutes interrogées et répondent massivement.
- Les plus petites sont tirées aléatoirement. Parmi celles retenues, le taux de réponse est faible  $\approx 55\%$  !

~~> Très peu probable que la non-réponse n'ait pas d'impact.

Source : **La correction de la non-réponse par repondération**, Thomas Deroyon ; 2017.

## Un dernier exemple sportif



# Tentative de définition



Une école de l'INSA



## Algorithmes : biais, discrimination et équité

Patrice Bertail, David Bounie, Stephan Cléménçon et Patrick Waelbroeck

## Algorithmes : biais, discrimination et équité.

Bertail, Bounie, Cléménçon,  
Waelbroeck ; 2019.

“Les biais pourraient être alors définis comme une **déviation** par rapport à un résultat censé être neutre, loyal ou encore équitable.”

# Un concept aux multiples facettes

On peut distinguer plusieurs catégories de biais.

- **Les biais cognitifs.** (Kahneman et Tversky, 70').

Exemple : tendance à détecter des corrélations inexistantes.

- **Les biais algorithmiques.**

Exemple : mauvais choix de métrique.

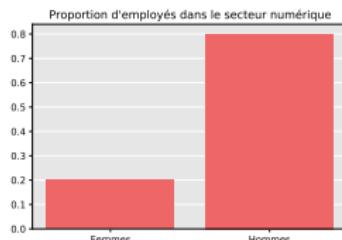
- **Les biais statistiques (ou biais des données).**

Problème de **représentativité**, inadéquation entre les données d'entraînements et la population sur laquelle on souhaite déployer l'algorithme.

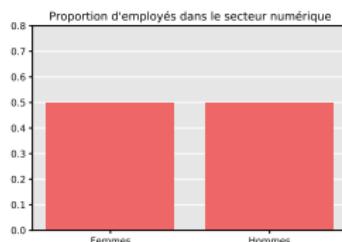
# Biais inhérents v.s. Biais de représentativité

Population (pas si) fictive : employés du secteur numérique constitués à 80% d'hommes et 20% de femmes.

Lors de la collecte des données d'entraînement d'un algorithme, on distinguera deux situations :



- Mêmes proportions que la population : les données sont représentatives. On peut cependant considérer que la situation même présente un **biais inhérent** de genre.



- Proportions différentes de la population : les données ne sont pas représentatives. On parle de **biais de représentativité** ou de **biais des données**.

## Mise en garde sur l'équité

Rappel : en Machine Learning, les méthodes de **fairness** consistent à se fixer un ou des attributs discriminants, et à évaluer les performances d'un algorithme au regard de ces attributs.

Si le biais est inhérent à la population, il est illusoire de penser qu'un algorithme résoudra seul le problème : l'enjeu est avant tout politique.

~~> Nous nous concentrerons uniquement sur les biais de représentativité.

⚠ Ces derniers **ne sont pas dissociés** des problématiques d'équité.

# Apprentissage statistique : cadre théorique

## Notations :

- $X \in \mathbb{R}^d$  : vecteur de **variables explicatives** de dimension  $d \geq 1$ ,
- $Y$  : **label** que l'on souhaite prédire.

Classification :  $Y \in \{0, 1\}$ . Régression  $Y \in \mathbb{R}$ .

# Apprentissage statistique : cadre théorique

## Notations :

- $X \in \mathbb{R}^d$  : vecteur de **variables explicatives** de dimension  $d \geq 1$ ,
- $Y$  : **label** que l'on souhaite prédire.

Classification :  $Y \in \{0, 1\}$ . **Régression**  $Y \in \mathbb{R}$ .

**Hypothèse de loi** :  $Z := (X, Y)$  est une variable aléatoire de loi  $P$ .

# Apprentissage statistique : cadre théorique

## Notations :

- $X \in \mathbb{R}^d$  : vecteur de **variables explicatives** de dimension  $d \geq 1$ ,
- $Y$  : **label** que l'on souhaite prédire.

Classification :  $Y \in \{0, 1\}$ . **Régression**  $Y \in \mathbb{R}$ .

**Hypothèse de loi** :  $Z := (X, Y)$  est une variable aléatoire de loi  $P$ .

**Hypothèse de métrique** : La performance d'une fonction de prédiction  $f$  est mesurée à l'aide du risque :

$$\mathcal{R}(f) = \mathbb{E}_P [(Y - f(X))^2].$$

# Apprentissage statistique : cadre théorique

$$\mathcal{R}(f) = \mathbb{E}_P [(Y - f(X))^2].$$

**Objectif idéal :** Construire une fonction de prédiction  $f_*$  appartenant à une classe de fonctions  $\mathcal{F}$ , minimisant le risque au sein de cette classe :

$$f_* \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}(f).$$

# Apprentissage statistique : cadre théorique

$$\mathcal{R}(f) = \mathbb{E}_P [(Y - f(X))^2].$$

**Objectif idéal :** Construire une fonction de prédiction  $f_*$  appartenant à une classe de fonctions  $\mathcal{F}$ , minimisant le risque au sein de cette classe :

$$f_* \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}(f).$$

**Hypothèse d'observation :** La loi  $P$  est inconnue.

En lieu et place de  $P$ , on dispose de  $n \geq 1$  observations indépendantes :

$$Z_1 = (X_1, Y_1), Z_2 = (X_2, Y_2), \dots, Z_n = (X_n, Y_n) \stackrel{\text{i.i.d.}}{\sim} P.$$

~ Le vrai risque est remplacé par son approximation empirique :

$$\mathcal{R}_n(f) := \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2.$$

# Apprentissage statistique : cadre théorique

$$\mathcal{R}(f) = \mathbb{E}_P [(Y - f(X))^2] \quad \text{et} \quad \mathcal{R}_n(f) = \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2.$$

**Objectif pratique :** On remplace l'objectif idéal de minimisation par son approximation empirique :

$$f_* \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}(f) \quad \rightsquigarrow \quad \hat{f}_n \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}_n(f).$$

# Apprentissage statistique : cadre théorique

$$\mathcal{R}(f) = \mathbb{E}_P [(Y - f(X))^2] \quad \text{et} \quad \mathcal{R}_n(f) = \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2.$$

**Objectif pratique :** On remplace l'objectif idéal de minimisation par son approximation empirique :

$$f_* \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}(f) \quad \rightsquigarrow \quad \hat{f}_n \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}_n(f).$$

**Pourquoi ça marche ?** Par la loi des grands nombres,

$$\mathcal{R}_n(f) := \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2 \approx \mathbb{E}_P [(Y - f(X))^2] = \mathcal{R}_P(f).$$

Donc

$$\hat{f}_n \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}_n(f) \approx \mathcal{R}_P(f) \quad \Rightarrow \quad \hat{f}_n \approx f_*.$$

# Apprentissage statistique : cadre théorique

$$\mathcal{R}(f) = \mathbb{E}_P [(Y - f(X))^2] \quad \text{et} \quad \mathcal{R}_n(f) = \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2.$$

**Objectif pratique :** On remplace l'objectif idéal de minimisation par son approximation empirique :

$$f_* \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}(f) \quad \rightsquigarrow \quad \hat{f}_n \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}_n(f).$$

**Pourquoi ça marche ?** Par la loi des grands nombres,

$$\mathcal{R}_n(f) := \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2 \approx \mathbb{E}_P [(Y - f(X))^2] = \mathcal{R}_P(f).$$

↑ Uniquement si  
 $(X_i, Y_i) \sim P$  !

Donc

$$\hat{f}_n \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}_n(f) \approx \mathcal{R}_P(f) \Rightarrow \hat{f}_n \approx f_*.$$

# Apprentissage statistique : cadre théorique

$$\mathcal{R}(f) = \mathbb{E}_P [(Y - f(X))^2] \quad \text{et} \quad \mathcal{R}_n(f) = \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2.$$

**Objectif pratique :** On remplace l'objectif idéal de minimisation par son approximation empirique :

$$f_* \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}(f) \quad \rightsquigarrow \quad \hat{f}_n \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}_n(f).$$

**Pourquoi ça marche ?** Par la loi des grands nombres,

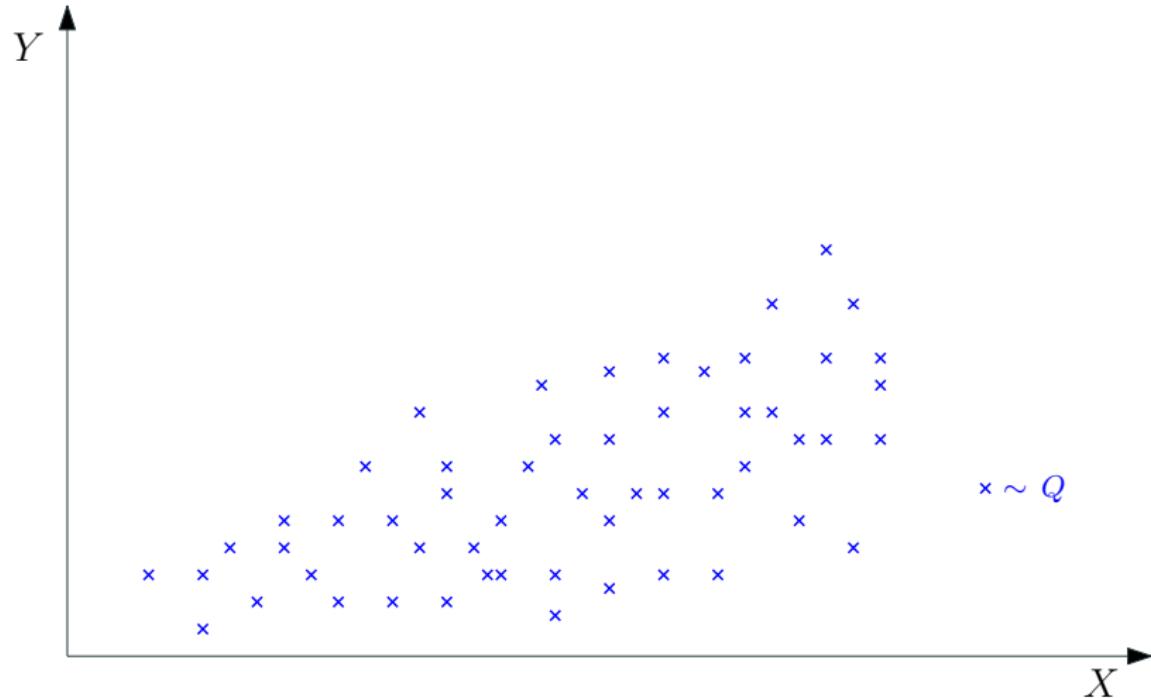
$$\mathcal{R}_n(f) := \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2 \approx \mathbb{E}_Q [(Y - f(X))^2] = \mathcal{R}_Q(f).$$

Donc

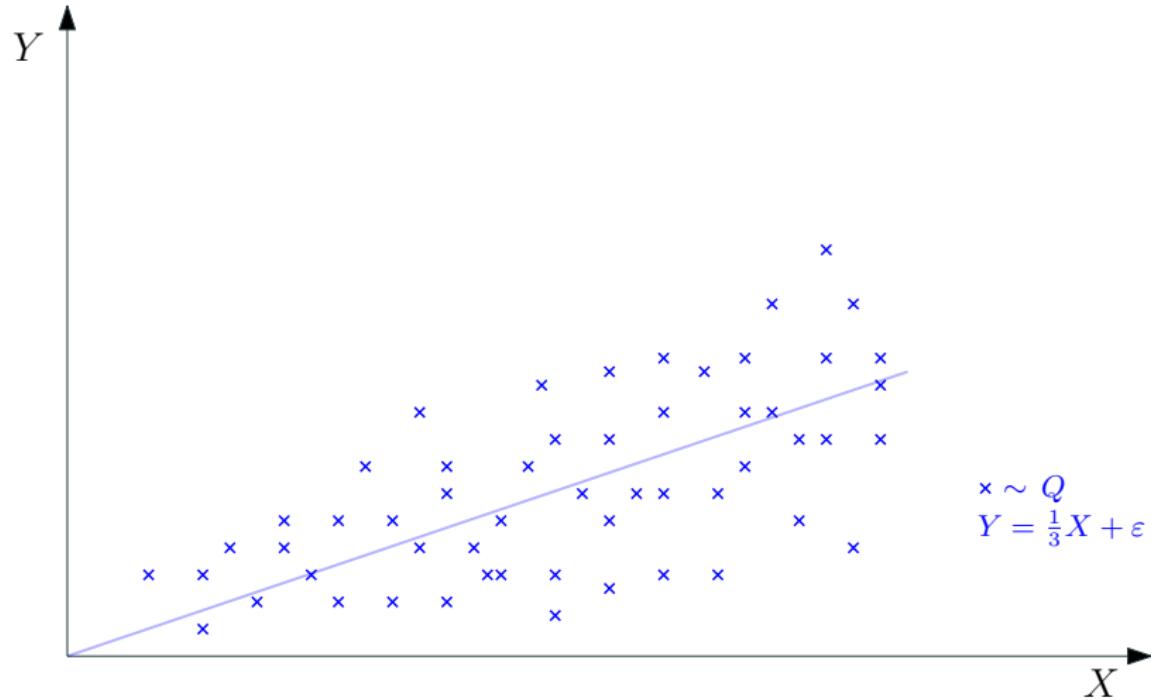
Si  $(X_i, Y_i) \sim Q$   
avec  $Q \neq P$  ?

$$\hat{f}_n \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}_n(f) \approx \mathcal{R}_Q(f) \Rightarrow \hat{f}_n \neq f_*$$

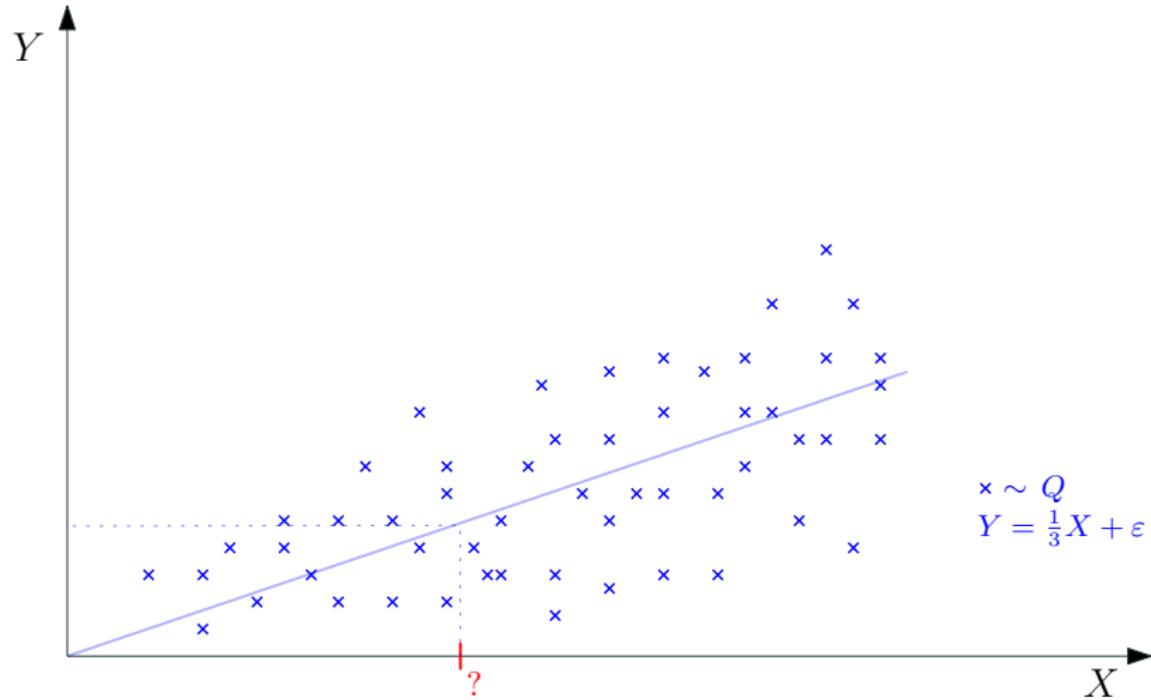
# Illustration



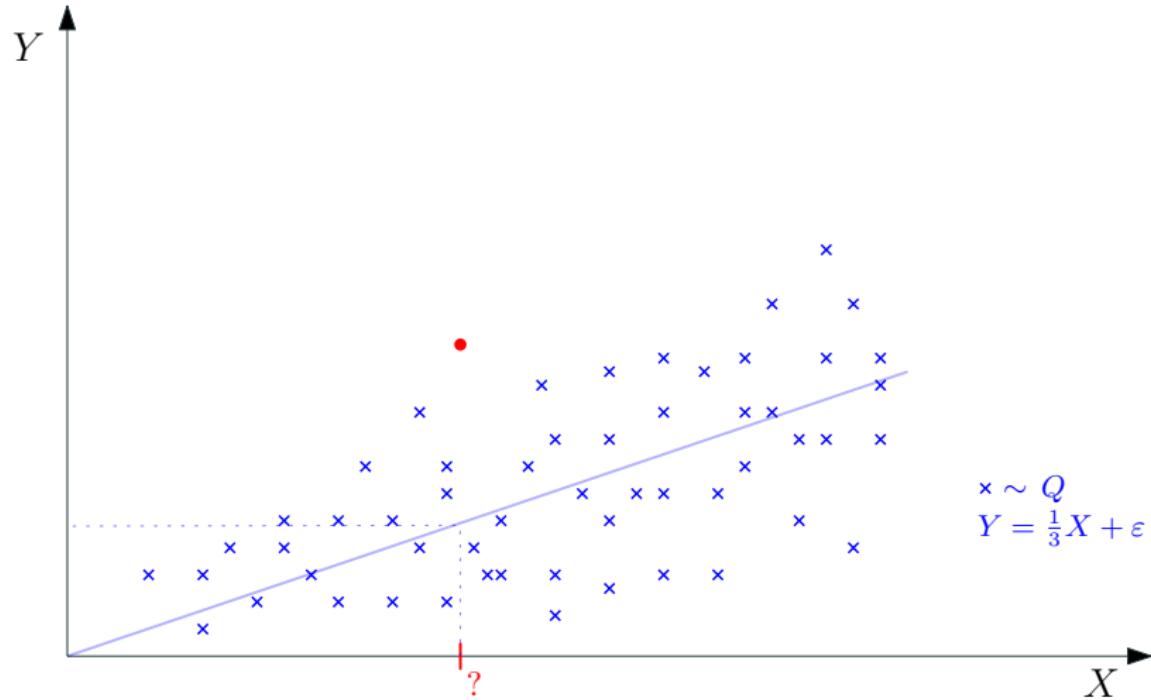
# Illustration



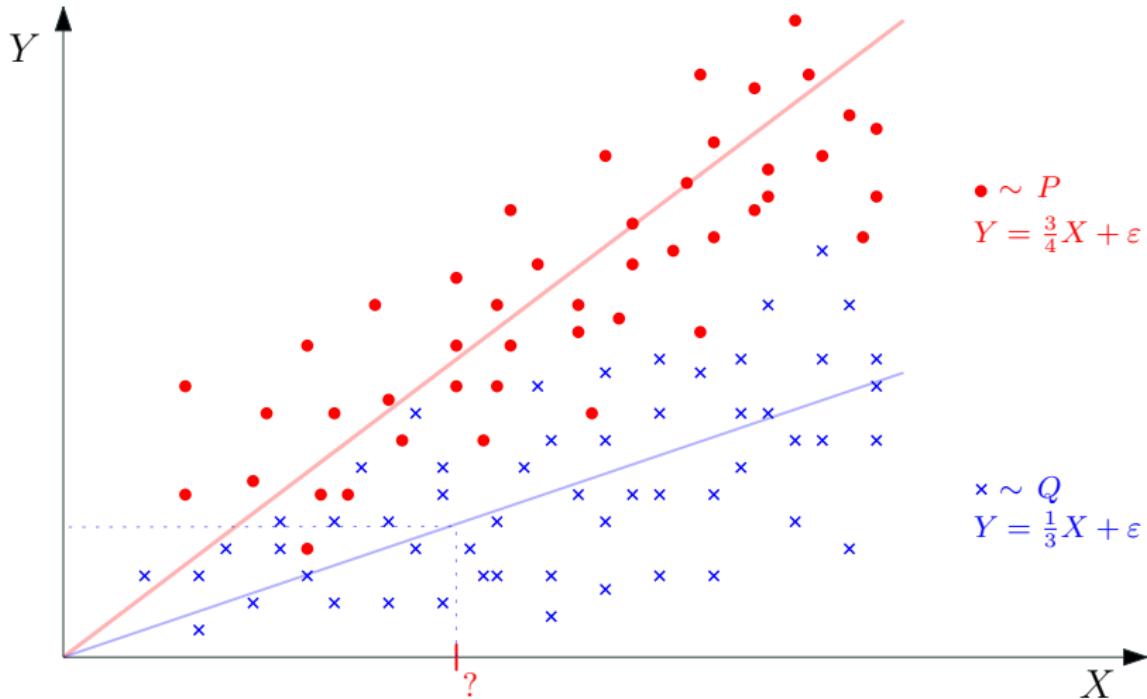
# Illustration



# Illustration



# Illustration



**D'où viennent les biais ?**

# Big Data : Un changement de paradigme

La possibilité de **collecter** et de **stocker** des volumes toujours plus massifs de données a modifié nos manières de les exploiter.

Schématiquement :

# Big Data : Un changement de paradigme

La possibilité de **collecter** et de **stocker** des volumes toujours plus massifs de données a modifié nos manières de les exploiter.

Schématiquement :

## Avant :



Pour une problématique donnée, les praticiens établissent un plan d'expérience adapté afin d'obtenir un échantillon représentatif.

# Big Data : Un changement de paradigme

La possibilité de **collecter** et de **stocker** des volumes toujours plus massifs de données a modifié nos manières de les exploiter.

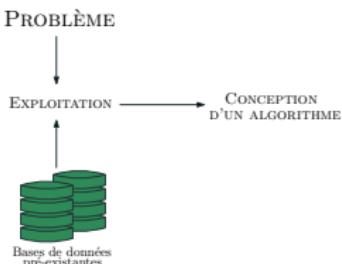
Schématiquement :

## Avant :



Pour une problématique donnée, les praticiens établissent un plan d'expérience adapté afin d'obtenir un échantillon représentatif.

## Après :

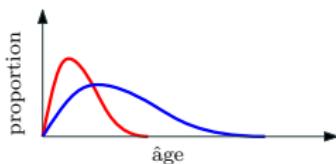


Les données sont disponibles avant même de se poser une question.  
~~ Le praticien n'a donc **aucun contrôle sur le processus d'acquisition des données**, qui peuvent présenter de nombreux biais.

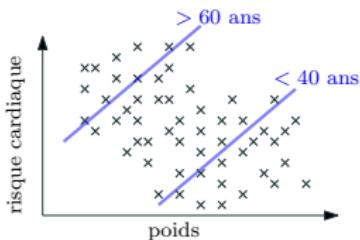
# Petit tour d'horizon des biais des données

|          | $Q_1$ | $Q_2$ | $Q_3$ | $\dots$ |
|----------|-------|-------|-------|---------|
| $I_1$    | ✓     | ✓     | ✗     |         |
| $I_2$    | ✓     | ✓     | ✓     |         |
| $I_3$    | ✗     | ✗     | ✓     |         |
| $\vdots$ |       |       |       |         |

- ABSENCE DE RÉPONSE



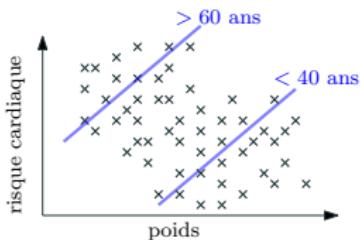
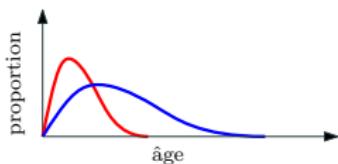
- DISTORSION CONTINUE DES DONNÉES



- EXISTENCE DE VARIABLES OMISES

# Petit tour d'horizon des biais des données

|          | $Q_1$ | $Q_2$ | $Q_3$ | $\dots$ |
|----------|-------|-------|-------|---------|
| $I_1$    | ✓     | ✓     | ✗     |         |
| $I_2$    | ✓     | ✓     | ✓     |         |
| $I_3$    | ✗     | ✗     | ✓     |         |
| $\vdots$ |       |       |       |         |



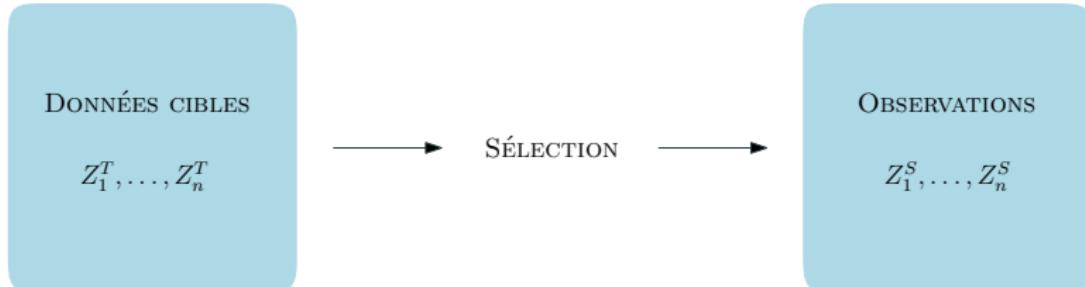
- ABSENCE DE RÉPONSE

NOTION DE  
SÉLECTION

- DISTORSION CONTINUE DES DONNÉES

- EXISTENCE DE VARIABLES OMISES

# Qu'est-ce que la sélection ?



**Modélisation :** Il est important de bien distinguer :

- Les données cibles que l'on aimerait traiter, réalisations i.i.d. d'une variable aléatoire  $Z^T = (X^T, Y^T)$  de loi  $P_T$ ,
- Les données observées dont on dispose, réalisations i.i.d. d'une variable aléatoire  $Z^S = (X^S, Y^S)$  de loi  $P_S$ .

À propos des notations :  $S$  pour *source* et  $T$  pour *target*.

# Deux grandes familles de sélection

## La sélection stricte :

On observe les données (ou une partie des données)  $Z_i^T$  de l'individu  $i$  si et seulement si celles-ci vérifient un critère  $c : \mathbb{R}^d \rightarrow \{0, 1\}$  (1 pour "oui" et 0 pour "non") fixé à l'avance :

$$Z_i^S = \begin{cases} Z_i^T & \text{si } c(Z_i^T) = 1, \\ \times & \text{si } c(Z_i^T) = 0. \end{cases}$$

**Exemple classique : Troncature.** Pour des raisons de confidentialité, on observe un salaire  $Z^T$  si et seulement si celui-ci est inférieur à un certain seuil  $s : c(Z^T) = 1_{\{Z^T \leq s\}}$

## Deux grandes familles de sélection

### La sélection stricte logistique :

Quand  $Z^T$  contient plusieurs variables, il est pratique de modéliser la sélection de la manière suivante :  $c(Z^T) = 1_{\{Z^{T'}\theta + \eta > 0\}}$  avec  $\eta$  une variable aléatoire de loi logistique.

## Deux grandes familles de sélection

### La sélection stricte logistique :

Quand  $Z^T$  contient plusieurs variables, il est pratique de modéliser la sélection de la manière suivante :  $c(Z^T) = 1_{\{Z^T \theta + \eta > 0\}}$  avec  $\eta$  une variable aléatoire de loi logistique.

Dans ce cas, on peut montrer que la probabilité d'observer un individu vérifie

$$\mathbb{P}(c(Z^T) = 1 \mid Z^T) = \frac{1}{1 + e^{-(Z^T)' \theta}}.$$

## Deux grandes familles de sélection

### La sélection stricte logistique :

Quand  $Z^T$  contient plusieurs variables, il est pratique de modéliser la sélection de la manière suivante :  $c(Z^T) = 1_{\{Z^T \theta + \eta > 0\}}$  avec  $\eta$  une variable aléatoire de loi logistique.

Dans ce cas, on peut montrer que la probabilité d'observer un individu vérifie

$$\mathbb{P}(c(Z^T) = 1 \mid Z^T) = \frac{1}{1 + e^{-(Z^T)' \theta}}.$$

**Interprétation :** Si  $Z^T = \text{âge}$  et  $\theta = -1$ , alors

$$\mathbb{P}(c(\text{âge}) = 1 \mid \text{âge}) = \frac{1}{1 + e^{\text{âge}}}$$

↔ les personnes âgées ont tendance à être moins observées que les autres.

## Deux grandes familles de sélection

### La sélection stricte (fin) :

La question des données manquantes se pose mécaniquement lorsque l'on évalue l'impact d'un traitement (e.g. médical). En effet, si l'on note  $Y(0)$  (*resp.*  $Y(1)$ ) l'état de santé sans (*resp.* avec) traitement, on remarque

|        | groupe de contrôle | groupe de traitement |
|--------|--------------------|----------------------|
| $Y(0)$ | observé            | manquant             |
| $Y(1)$ | manquant           | observé              |

## Deux grandes familles de sélection

### La sélection stricte (fin) :

La question des données manquantes se pose mécaniquement lorsque l'on évalue l'impact d'un traitement (e.g. médical). En effet, si l'on note  $Y(0)$  (*resp.*  $Y(1)$ ) l'état de santé sans (*resp.* avec) traitement, on remarque

|        | groupe de contrôle | groupe de traitement |
|--------|--------------------|----------------------|
| $Y(0)$ | observé            | manquant             |
| $Y(1)$ | manquant           | observé              |

L'évaluation de traitement est une question omniprésente dans le domaine très actif de l'apprentissage statistique dit "causal".

~~ Encadrement d'un groupe d'étudiants en stage à Air Liquide sur ce sujet.

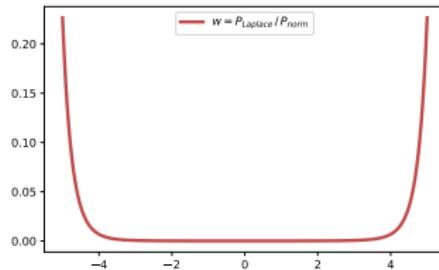
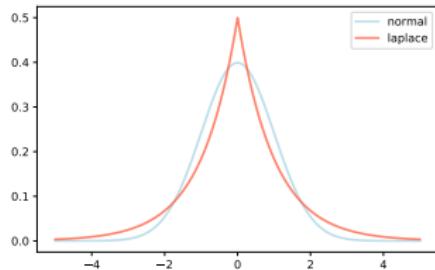
# Deux grandes familles de sélection

## La sélection molle :

Les données sont parfaitement observées... mais leur loi est une déformation de la loi cible :

$$P_T(z) \approx \omega(z) P_S(z).$$

La fonction  $\omega(\cdot)$  contient l'information nécessaire pour passer de la loi source à la loi cible.



# **Comment détecter les biais ?**

## Cas difficile : aucune information a priori sur la sélection

**Cadre :** On observe un échantillon d'apprentissage  $(Z_i^S)_{i=1}^n$  i.i.d. de loi  $P_S$ , sans valeur manquante. On souhaite vérifier que  $P_S = P_T$ .

## Cas difficile : aucune information a priori sur la sélection

**Cadre :** On observe un échantillon d'apprentissage  $(Z_i^S)_{i=1}^n$  i.i.d. de loi  $P_S$ , sans valeur manquante. On souhaite vérifier que  $P_S = P_T$ .

Aucune observation sous  $P_T \rightsquigarrow$  besoin d'information extérieure sur  $P_T$  !

## Cas difficile : aucune information a priori sur la sélection

**Cadre :** On observe un échantillon d'apprentissage  $(Z_i^S)_{i=1}^n$  i.i.d. de loi  $P_S$ , sans valeur manquante. On souhaite vérifier que  $P_S = P_T$ .

Aucune observation sous  $P_T$  ↠ besoin d'information extérieure sur  $P_T$  !

**Hypothèse :** On suppose connaître l'espérance  $\mathbb{E}[Z^T]$  de  $Z^T$ , i.e. un vecteur de moments sous la loi  $P_T$ .

## Cas difficile : aucune information a priori sur la sélection

**Cadre :** On observe un échantillon d'apprentissage  $(Z_i^S)_{i=1}^n$  i.i.d. de loi  $P_S$ , sans valeur manquante. On souhaite vérifier que  $P_S = P_T$ .

Aucune observation sous  $P_T \rightsquigarrow$  besoin d'information extérieure sur  $P_T$  !

**Hypothèse :** On suppose connaître l'espérance  $\mathbb{E}[Z^T]$  de  $Z^T$ , i.e. un vecteur de moments sous la loi  $P_T$ .

**En pratique :** On a accès à un estimateur de  $\mathbb{E}[Z^T]$  issu d'une source d'information extérieure fiable (recensement, Google trends...). Celles-ci fournissent de l'information agrégée sur un grand nombre de variables, mais n'autorisent pas un accès à des données fines.

## Cas difficile : aucune information a priori sur la sélection

**Cadre :** On observe un échantillon d'apprentissage  $(Z_i^S)_{i=1}^n$  i.i.d. de loi  $P_S$ , sans valeur manquante. On souhaite vérifier que  $P_S = P_T$ .

Aucune observation sous  $P_T$  ↠ besoin d'information extérieure sur  $P_T$  !

**Hypothèse :** On suppose connaître l'espérance  $\mathbb{E}[Z^T]$  de  $Z^T$ , i.e. un vecteur de moments sous la loi  $P_T$ .

**En pratique :** On a accès à un estimateur de  $\mathbb{E}[Z^T]$  issu d'une source d'information extérieure fiable (recensement, Google trends...). Celles-ci fournissent de l'information agrégée sur un grand nombre de variables, mais n'autorisent pas un accès à des données fines.

**Exemple :** L'INSEE communique le salaire moyen des français mais pas la base de données des salaires individuels.

## Cas difficile : aucune information a priori sur la sélection

**Stratégie :** On va comparer la moyenne empirique issue de notre échantillon  $\overline{Z^S} := (1/n) \sum_{i=1}^n Z_i^S$  à  $\mathbb{E}[Z^T]$ .

## Cas difficile : aucune information a priori sur la sélection

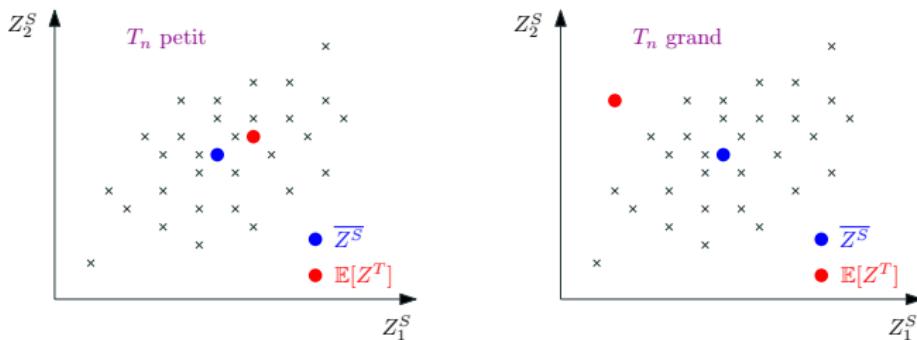
**Stratégie :** On va comparer la moyenne empirique issue de notre échantillon  $\overline{Z^S} := (1/n) \sum_{i=1}^n Z_i^S$  à  $\mathbb{E}[Z^T]$ . On aimerait résumer l'information avec un critère simple *univarié*.

# Cas difficile : aucune information a priori sur la sélection

**Stratégie :** On va comparer la moyenne empirique issue de notre échantillon  $\overline{Z^S} := (1/n) \sum_{i=1}^n Z_i^S$  à  $\mathbb{E}[Z^T]$ . On aimerait résumer l'information avec un critère simple *univarié*. Critère possible :

$$T_n := n \left( \overline{Z^S} - \mathbb{E}[Z^T] \right)' W_n^{-1} \left( \overline{Z^S} - \mathbb{E}[Z^T] \right),$$

avec  $W_n := (1/n) \sum_{i=1}^n Z_i^S Z_i^{S'} - \overline{Z^S} \overline{Z^S}'$ .



**Remarque :** Si la dimension de  $Z^S$  vaut  $p$ , la loi de  $T_n$  est proche de celle du chi-deux à  $p$  degrés de liberté.

## Cas plus simple : information a priori sur la sélection

**Cadre :** On observe  $(Z_{1,i}^S, Z_{2,i}^T, D_i)_{i=1}^n$  tel que  $Z_{1,i}^S$  est égal à  $Z_{1,i}^T$  si  $D_i = 1$  et est égal à une valeur manquante si  $D_i = 0$ .

|         |         |         |
|---------|---------|---------|
| $D = 1$ | $Z_1^T$ | $Z_2^T$ |
| $D = 0$ | X       | $Z_2^T$ |

## Cas plus simple : information a priori sur la sélection

**Cadre :** On observe  $(Z_{1,i}^S, Z_{2,i}^T, D_i)_{i=1}^n$  tel que  $Z_{1,i}^S$  est égal à  $Z_{1,i}^T$  si  $D_i = 1$  et est égal à une valeur manquante si  $D_i = 0$ .

|         |         |         |
|---------|---------|---------|
| $D = 1$ | $Z_1^T$ | $Z_2^T$ |
| $D = 0$ | X       | $Z_2^T$ |

Seuls les individus tels que  $D_i = 1$  sont exploitables. Ces derniers sont-ils similaires à ceux pour qui  $D_i = 0$ ? Si non : problème de biais!

## Cas plus simple : information a priori sur la sélection

**Cadre :** On observe  $(Z_{1,i}^S, Z_{2,i}^T, D_i)_{i=1}^n$  tel que  $Z_{1,i}^S$  est égal à  $Z_{1,i}^T$  si  $D_i = 1$  et est égal à une valeur manquante si  $D_i = 0$ .

|         |         |         |
|---------|---------|---------|
| $D = 1$ | $Z_1^T$ | $Z_2^T$ |
| $D = 0$ | X       | $Z_2^T$ |

Seuls les individus tels que  $D_i = 1$  sont exploitables. Ces derniers sont-ils similaires à ceux pour qui  $D_i = 0$ ? Si non : problème de biais!

Pour répondre à cette question, deux possibilités (complémentaires) :

- utiliser méthode précédente sur l'échantillon  $D_i = 1$ .
- utiliser seulement  $(Z_{2,i}^T)_{i=1}^n$  (car toujours observé) et comparer les sous-groupes  $(Z_{2,i}^T)_{i:D_i=0}$  et  $(Z_{2,i}^T)_{i:D_i=1}$ .

## Cas plus simple : information a priori sur la sélection

**Comment comparer  $(Z_{2,i}^T)_{i:D_i=0}$  et  $(Z_{2,i}^T)_{i:D_i=1}$  ?**

## Cas plus simple : information a priori sur la sélection

**Comment comparer  $(Z_{2,i}^T)_{i:D_i=0}$  et  $(Z_{2,i}^T)_{i:D_i=1}$  ?**

Quand  $Z_2^T$  est scalaire, de nombreux tests statistiques existent : tests de Kolmogorov-Smirnov, de Cramer-von Mises, de rangs...

## Cas plus simple : information a priori sur la sélection

**Comment comparer  $(Z_{2,i}^T)_{i:D_i=0}$  et  $(Z_{2,i}^T)_{i:D_i=1}$  ?**

Quand  $Z_2^T$  est scalaire, de nombreux tests statistiques existent : tests de Kolmogorov-Smirnov, de Cramer-von Mises, de rangs... Quand  $Z_2^T$  est multivarié, cela se complique.

## Cas plus simple : information a priori sur la sélection

**Comment comparer**  $(Z_{2,i}^T)_{i:D_i=0}$  **et**  $(Z_{2,i}^T)_{i:D_i=1}$  ?

Quand  $Z_2^T$  est scalaire, de nombreux tests statistiques existent : tests de Kolmogorov-Smirnov, de Cramer-von Mises, de rangs... Quand  $Z_2^T$  est multivarié, cela se complique.

Contribution récente : **A Kernel Two-Sample Test.** Gretton, Borgwardt, Rasch, Schölkopf, Smola ; 2012, JMLR.

## Focus sur le test à noyau de Gretton et al.

Supposons que les  $n_1$  premiers individus soient tels que  $D_i = 0$ . Gretton et al. proposent le critère suivant :

$$T_{n,k} :=$$

$$\left( \frac{1}{n_1^2} \sum_{1 \leq i, j \leq n_1} k_{i,j} - \frac{2}{n_1(n-n_1)} \sum_{\substack{1 \leq i \leq n_1 \\ n_1+1 \leq j \leq n}} k_{i,j} + \frac{1}{(n-n_1)^2} \sum_{n_1+1 \leq i, j \leq n} k_{i,j} \right)^{1/2}$$

avec  $k_{i,j} := k(Z_{2,i}^T, Z_{2,j}^T)$  une fonction appelée un *noyau*.

## Focus sur le test à noyau de Gretton et al.

Supposons que les  $n_1$  premiers individus soient tels que  $D_i = 0$ . Gretton et al. proposent le critère suivant :

$$T_{n,k} :=$$

$$\left( \frac{1}{n_1^2} \sum_{1 \leq i, j \leq n_1} k_{i,j} - \frac{2}{n_1(n-n_1)} \sum_{\substack{1 \leq i \leq n_1 \\ n_1+1 \leq j \leq n}} k_{i,j} + \frac{1}{(n-n_1)^2} \sum_{n_1+1 \leq i, j \leq n} k_{i,j} \right)^{1/2}$$

avec  $k_{i,j} := k(Z_{2,i}^T, Z_{2,j}^T)$  une fonction appelée un *noyau*.

Exemple classique : noyau gaussien  $k_{i,j} = \exp(-||Z_{2,i}^T - Z_{2,j}^T||^2/\sigma^2)$ .

## Focus sur le test à noyau de Gretton et al.

Supposons que les  $n_1$  premiers individus soient tels que  $D_i = 0$ . Gretton et al. proposent le critère suivant :

$$T_{n,k} :=$$

$$\left( \frac{1}{n_1^2} \sum_{1 \leq i, j \leq n_1} k_{i,j} - \frac{2}{n_1(n-n_1)} \sum_{\substack{1 \leq i \leq n_1 \\ n_1+1 \leq j \leq n}} k_{i,j} + \frac{1}{(n-n_1)^2} \sum_{n_1+1 \leq i, j \leq n} k_{i,j} \right)^{1/2}$$

avec  $k_{i,j} := k(Z_{2,i}^T, Z_{2,j}^T)$  une fonction appelée un *noyau*.

Exemple classique : noyau gaussien  $k_{i,j} = \exp(-||Z_{2,i}^T - Z_{2,j}^T||^2/\sigma^2)$ .

**Intuition derrière  $T_{n,k}$**  : si l'on note  $Q_0$  (*resp.*  $Q_1$ ) la loi de  $Z_2^T \mid D = 0$  (*resp.* de  $Z_2^T \mid D = 1$ ), alors

$$\begin{aligned} T_{n,k}^2 &\approx \mathbb{E}_{(U_0, U'_0) \sim Q_0 \times Q_0} [k(U_0, U'_0)] + \mathbb{E}_{(U_1, U'_1) \sim Q_1 \times Q_1} [k(U_1, U'_1)] \\ &\quad - 2\mathbb{E}_{(U_0, U_1) \sim Q_0 \times Q_1} [k(U_0, U_1)]. \end{aligned}$$

## Tester l'indépendance entre $D$ et $Z_2^T$

Au lieu de comparer  $(Z_{2,i}^T)_{i:D_i=0}$  et  $(Z_{2,i}^T)_{i:D_i=1}$ , on peut construire un test d'indépendance des lois de  $D$  et  $Z_2^T$  à l'aide de  $(D_i, Z_{2,i}^T)_{i=1}^n$ .

## Tester l'indépendance entre $D$ et $Z_2^T$

Au lieu de comparer  $(Z_{2,i}^T)_{i:D_i=0}$  et  $(Z_{2,i}^T)_{i:D_i=1}$ , on peut construire un test d'indépendance des lois de  $D$  et  $Z_2^T$  à l'aide de  $(D_i, Z_{2,i}^T)_{i=1}^n$ .

Si l'indépendance est rejetée, on conclut que  $Z_2^T$  explique (en partie) l'existence de variables manquantes ↵ données non manquantes sont biaisées.

# Tester l'indépendance entre $D$ et $Z_2^T$

Au lieu de comparer  $(Z_{2,i}^T)_{i:D_i=0}$  et  $(Z_{2,i}^T)_{i:D_i=1}$ , on peut construire un test d'indépendance des lois de  $D$  et  $Z_2^T$  à l'aide de  $(D_i, Z_{2,i}^T)_{i=1}^n$ .

Si l'indépendance est rejetée, on conclut que  $Z_2^T$  explique (en partie) l'existence de variables manquantes ↪ données non manquantes sont biaisées.

## Méthode :

1. Effectuer un clustering en  $K$  classes de  $(Z_{2,i}^T)_{i=1}^n$ .
2. Soit  $C_i$  la classe d'appartenance de la  $i$ -ème observation. Faire un test du khi-deux d'indépendance à partir de  $(D_i)_{i=1}^n$  et  $(C_i)_{i=1}^n$ .

## Extensions

**Pour comparer**  $(Z_{2,i}^T)_{i:D_i=0}$  **et**  $(Z_{2,i}^T)_{i:D_i=1}$  : généralisations multivariées  
des tests de Kolmogorov-Smirnov, Cramer-von Mises ou de rangs  
~~> Travaux en cours avec Myrto Limnios et Stéphan Cléménçon.

## Extensions

**Pour comparer**  $(Z_{2,i}^T)_{i:D_i=0}$  **et**  $(Z_{2,i}^T)_{i:D_i=1}$  : généralisations multivariées des tests de Kolmogorov-Smirnov, Cramer-von Mises ou de rangs  
~~~ Travaux en cours avec Myrto Limnios et Stéphan Cléménçon.

**Pour tester l'indépendance entre  $D$  et  $Z_2^T$**  : il est possible d'apprendre des poids de re-pondération à partir de  $(Z_{2,i}^T)_{D_i=1}$  (*cf. section suivante*). Si les poids varient beaucoup entre les observations, alors  $D$  et  $Z_2^T$  ne sont pas indépendants.

**Peut-on corriger les biais ?**

# Combattre la sélection stricte ?

## L'imputation :

|    | Q1 | Q2 | Q3 | Q4 | Q5 | ... |
|----|----|----|----|----|----|-----|
| I1 | ✓  | ✓  | ✗  | ✓  | ✓  |     |
| I2 | ✓  | ✓  | ✓  | ✗  | ✓  |     |
| I3 | ✓  | ✓  | ✓  | ✓  | ✓  |     |
| I4 | ✓  | ✓  | ✓  | ✓  | ✓  |     |
| I5 | ✓  | ✗  | ✓  | ✓  | ✓  |     |
| I6 | ✓  | ✓  | ✓  | ✓  | ✓  |     |
| ⋮  |    |    |    |    |    |     |

Lorsque l'absence de réponse est **parcimonieuse**, la stratégie d'imputation consiste à substituer aux données manquantes des réponses *vraisemblables*.  
~~> Matching, Random Matrices Completion, ...

**Autre approche pour l'absence de réponse parcimonieuse :** entraîner un algorithme de ML prenant en compte l'absence de réponse.  
~~> NeuMiss networks, Le Morvan and al. (2020).

# Combattre la sélection stricte ?

## La re-pondération :

Lorsqu'on a uniquement accès aux données entièrement observées  $Z_1^S, Z_2^S, \dots, Z_n^S$ , on cherche à corriger le biais

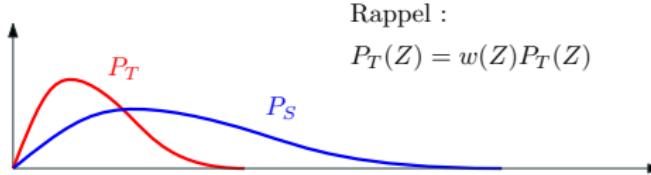
|          | $Q1$         | $Q2$         | $Q3$         | $Q4$         | $Q5$         | $\dots$ |
|----------|--------------|--------------|--------------|--------------|--------------|---------|
| $I1$     | $\times$     | $\times$     | $\times$     | $\times$     | $\times$     |         |
| $I2$     | $\times$     | $\times$     | $\times$     | $\times$     | $\times$     |         |
| $I3$     | $\checkmark$ | $\checkmark$ | $\checkmark$ | $\checkmark$ | $\checkmark$ |         |
| $I4$     | $\checkmark$ | $\checkmark$ | $\checkmark$ | $\checkmark$ | $\checkmark$ |         |
| $I5$     | $\times$     | $\times$     | $\times$     | $\times$     | $\times$     |         |
| $I6$     | $\checkmark$ | $\checkmark$ | $\checkmark$ | $\checkmark$ | $\checkmark$ |         |
| $\vdots$ |              |              |              |              |              |         |

$$\frac{1}{n} \sum_{i=1}^n \delta_{Z_i^S} \neq P_T.$$

Pour cela, on calcule des poids  $w_i$  de sorte que

$$\sum_{i=1}^n w_i \delta_{Z_i^S} \approx P_T.$$

# Combattre la sélection molle ?



Rappel :  
 $P_T(Z) = w(Z)P_S(Z)$

Les approches existantes supposent avoir accès à :

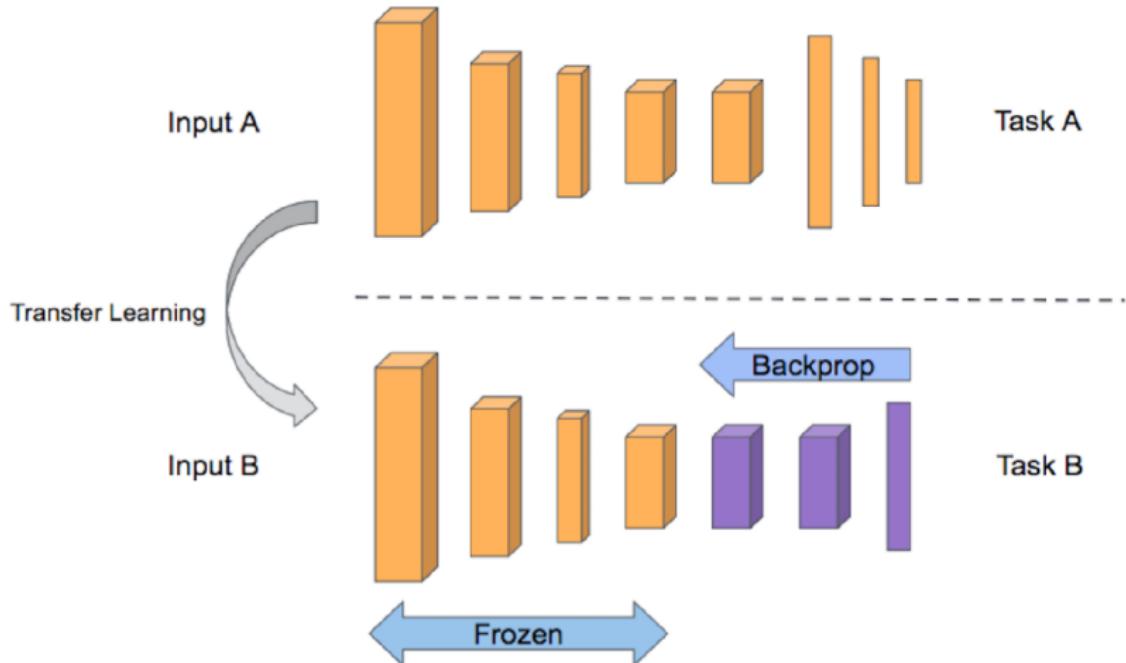
- un grand échantillon sous la loi source  $P_S$  :  $Z_1^S, \dots, Z_n^S$  ;
- un petit échantillon sous la loi cible  $P_T$  :  $Z_1^T, \dots, Z_m^T$ .

↔ Ceci permet d'apprendre la fonction de lien  $w$ . Plusieurs algorithmes :

- **Kernel Mean Matching. Correcting Sample Selection Bias by Unlabeled Data**, NeurIPS 2007, Huang and al. ;
- **KLIEP. Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation**, NeurIPS 2007, Sugiyama and al.

Plutôt adapté pour l'augmentation de données.

# Un mot sur le transfer learning



# Notre contribution

Méthode de re-pondération en absence de données cibles :

**Learning from Biased Data : A Semi-Parametric Approach** ; Bertail,  
Cléménçon, G., N.; 2021, ICML.

---

## Learning from Biased Data: A Semi-Parametric Approach

---

Patrice Bertail<sup>\*1</sup> Stephan Cléménçon<sup>\*2</sup> Yannick Guyonvarch<sup>\*2</sup> Nathan Noiry<sup>\*2</sup>

### Abstract

We consider risk minimization problems where the (source) distribution  $P_S$  of the training observations  $Z_1, \dots, Z_n$  differs from the (target) distribution  $P_T$  involved in the risk that one seeks to minimize. Under the natural assumption that  $P_S$  dominates  $P_T$ , i.e.  $P_T \ll P_S$ , we develop a semi-parametric framework in the situation where we do not observe any sample from  $P_T$ , but rather have access to some auxiliary information at the target population scale. More precisely, assuming that the Radon-Nikodym derivative  $dP_T/dP_S(z)$  belongs to a parametric class  $\{g(z, \alpha), \alpha \in \mathcal{A}\}$  and that some (generalized) moments of  $P_T$  are available to the learner, we propose a two-step learning procedure to perform the risk minimization task. We first select  $\hat{\alpha}$  so as to match the moment constraints as closely as possible and then reweight each (biased) training observation  $Z_i$  by  $g(Z_i, \hat{\alpha})$  in the final ERM algorithm. We establish a  $O_p(1/\sqrt{n})$  generalization bound proving that, remarkably, the solution to the weighted ERM problem thus constructed achieves a learning rate of the same order as that attained in absence of any sampling bias. Beyond these theoretical guarantees, numerical results providing strong empirical evidence of the relevance of the approach promoted in this article are displayed.

### problem

$$\inf_{\theta \in \Theta} \mathcal{R}_F(\theta), \quad (1)$$

one assumes that a training dataset  $\mathcal{D}_n = [Z_1, \dots, Z_n]$  composed of  $n \geq 1$  independent copies of the generic r.v.  $Z$  is available. A natural learning procedure consists in replacing the unknown risk by its empirical version based on the  $Z_i$ 's and solving next

$$\inf_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \ell(Z_i, \theta). \quad (2)$$

The accuracy of this procedure, referred to as *Empirical Risk Minimization* (ERM in abbreviated form), is usually assessed by establishing upper confidence bounds for the risk excess of empirical minimizers, that is the difference between the risk of solutions  $\hat{\theta}_n$  to (2) and the minimum risk attained over the class  $\Theta$ , under suitable assumptions on the loss function  $\ell$  and the parameter space  $\Theta$ , see e.g. (Devroye et al., 1996). Such results offer statistical guarantees regarding the generalization capacity of the predictive rule encoded by the learned parameter  $\hat{\theta}_n$ , when applied to a new/test observation  $Z_{\text{test}}$  with distribution  $P$ .

The usual validity framework for ERM crucially relies on the assumption that the distributions of the random variables involved in the training and test/prediction stages are the same. However, this assumption is now highly arguable in a wide variety of situations. Whereas, in the recent past, data collection was expensive and still performed by means of carefully elaborated experimental designs through surveys and questionnaires, practitioners have more and more often

## Notre contribution

**Et si aucun échantillon sous la loi cible n'est disponible ?**

**Objectif :** construire un algorithme uniquement à partir des données sources  $Z_1^S, \dots, Z_n^S$ , capable de généraliser sous la loi cible  $P_T$ .

**Hypothèse :** On remplace l'accès à des données de loi  $P_T$  par une connaissance *a priori* sur  $P_T$ . Plus précisément, on suppose connaître un certain nombre de moments de cette loi :

$$M_l = \mathbb{E}_{P_T}[m_l(Z)], \quad l = 1, \dots, p.$$

**Égalité fondamentale :** Par définition de la dérivée de Radon-Nikodym,

$$\mathbb{E}_{P_T}[m_l(Z)] = \mathbb{E}_{P_S}[w(Z)m_l(Z)].$$

## Notre contribution

$$M_l = \mathbb{E}_{P_T}[m_l(Z)] = \mathbb{E}_{P_S}[w(Z)m_l(Z)].$$

**Stratégie :** On cherche une approximation  $\hat{w}$  de la fonction de lien  $w$ , satisfaisant au mieux les conditions marginales, *i.e.*

$$\forall l \in \{1, \dots, p\}, \quad \underbrace{\frac{1}{n} \sum_{i=1}^n \hat{w}(Z_i^S) m_l(Z_i^S)}_{\approx \mathbb{E}_{P_S}[\hat{w}(Z)m_l(Z)]} \approx M_l.$$

# Notre contribution

## Algorithme en deux étapes :

1. On sélectionne

$$\hat{w} \in \operatorname{argmin}_{g \in \mathcal{W}} \sum_{l=1}^p \left( \frac{1}{n} \sum_{i=1}^n g(Z_i^S) m_l(Z_i^S) - M_l \right)^2.$$

2. On entraîne ensuite notre algorithme de ML préféré sur les données re-pondérées, ce qui revient à résoudre le problème suivant :

$$\hat{f}_n \in \operatorname{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n \hat{w}(Z_i^S) \ell(Z_i^S, f).$$

# Notre contribution

## Garantie théorique sur la généralisation :

**Théorème (Bertail, Cléménçon, Guyonvarch, N. ; 2021)**

*Si  $w \in \mathcal{W}$ , alors avec probabilité au moins égale à  $1 - \delta$ ,*

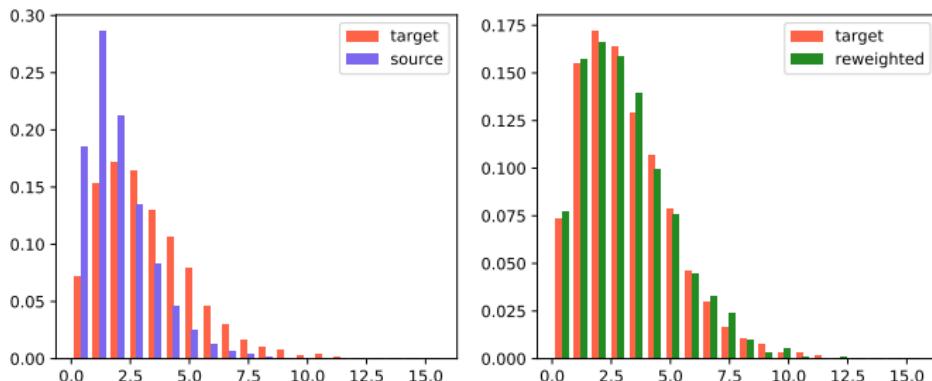
$$\begin{aligned} \mathcal{R}_{P_T}(\hat{f}_n) - \inf_{f \in \mathcal{F}} \mathcal{R}_{P_T}(f) \\ \leq \|w\|_\infty \|\ell\|_\infty \frac{C(\delta)}{\sqrt{N}} + 2\|w\|_\infty \left( 2R_N + \|\ell\|_\infty \sqrt{\frac{2 \log(4/\delta)}{N}} \right). \end{aligned}$$

où

$$\mathcal{R}_{P_T}(f) = \mathbb{E}_{P_T}[\ell(Z, f)] \quad \text{et} \quad R_N = \mathbb{E}_\sigma \left[ \sup_{\theta \in \Theta} \frac{1}{n} \left| \sum_{i=1}^N \sigma_i \ell(\theta, Z_i) \right| \right].$$

# Notre contribution

Objectif du notebook :



| ALGORITHM | Rw-ERM( $P_S$ )                 | ERM( $P_T$ )  | ERM( $P_S$ )  |
|-----------|---------------------------------|---------------|---------------|
| OLS       | <b><math>3.8 \pm 0.4</math></b> | $3.8 \pm 0.4$ | $6.3 \pm 0.7$ |
| SVR       | <b><math>1.5 \pm 0.5</math></b> | $1.2 \pm 0.3$ | $2.8 \pm 0.8$ |
| RF        | <b><math>1.7 \pm 0.2</math></b> | $1.6 \pm 0.2$ | $2.5 \pm 0.4$ |

## Ce qu'il faut retenir

- Les biais sont multiples. À notre petite échelle de chercheur / ingénieur / data scientist, nous pouvons uniquement espérer avoir un impact sur les **biais de représentativité**  $\neq$  biais inhérents.
- Toute tentative de correction de biais repose sur un **modèle**.  
~~> Nécessité d'un bon choix de modélisation (sélection stricte/soft).
- Associés à la connaissance d'**informations auxiliaires**, ces modèles permettent de :
  - déceler l'existence de certains biais à l'aide de tests d'hypothèses ;
  - le cas échéant, corriger ces biais avec des méthodes d'imputation / re-pondération.

**Merci de votre attention !**