Apprentissage décentralisé pour la recherche médicale

Paul Mangold

INRIA Lille, CHU Lille

18 juin 2020

Section 1

Introduction

Introduction

L'essor de l'intelligence artificielle : petite chronologie

- 1957 (Frank Rosenblatt) : perceptron.
- 1980+ : recherche en apprentissage automatique se développe.
- 1990 env. (Yann Le Cun) : réseaux de neurones convolutifs pour MNIST.
- 2010+ : essor du big data.
- 2015+ : apprentissage fédéré/décentralisé.

Introduction

Intelligence Artificielle : efficace grâce aux données

IA : succès en vision, prédiction, explication...

$\rightarrow$ grâce à la disponibilité de données massives.

Mais les données de santé sont cloisonnées (heureusement).

Introduction

Décloisonner les données

Deux cas de figures :

- Health Data Hub (élargissement du Système National des Données de Santé (SNDS)) ;
- études cliniques multi-centriques.

Introduction

Décloisonner les données

Plusieurs problèmes :

- flux de données importants ;
- duplication des données : risque de fuite ;
- contraintes juridiques fortes ;
- risque de dépossession de ses données ;
- dé-identification des données peut aussi détruire leur utilité.

Introduction

Garder ses données sur place

Vision défendue par MAGNET et INCLUDE :

- chaque hôpital, service ou même patient garde ses données ;
- ne sont transmis que les résultats de calculs locaux sur les données.

→ Possible d'analyser les données de façon décentralisée.

Introduction

Garder ses données sur place

Avantages de l'approche décentralisée :

- garanties de confidentialités possibles (même sans désidentification) ;
- distribution du stockage et de la puissance de calcul ;
- possibilité de collaborarer avec plus de centres ?

Introduction

Garder ses données sur place

Questions sur l'approche décentralisée :

- précision des résultats obtenus ?
- fuites des données ?
- difficulté de mise en place des études ?

Introduction

Ce qu'on va regarder aujourd'hui

On regarde comment réaliser deux types d'études :

- études statistiques globales ;
- études du lien entre plusieurs paramètres (par régression logistique).

Introduction

Cas d'études

Cas d'étude 1 :

calcul du volume de ventilation pratiqué au bloc opératoire à l'échelle nationale.

Cas d'étude 2 :

déterminer les correlations entre plusieurs paramètres (chute de tension, médicaments administrés...) lors d'une césarienne sur le pH foetal à l'accouchement.

Introduction

Quelques hypothèses

Aujourd'hui, on s'intéresse à une étude décentralisée où :

- tous les participants peuvent communiquer deux à deux ;
- les communications peuvent être synchronisées ;
- les données suivent la même distribution dans chaque centre ;
- les données restent sur place.

Section 2

Recherche de facteurs de risque

Recherche de facteurs de risque

Introduction

Étude clinique type : groupe témoin et groupe expérimental.

Objectif : recherche des facteurs de risque.

Exemple :

"une chute de tension artérielle à l'accouchement baisse le pH foetal."

$\rightarrow$ Régression logistique.

Recherche de facteurs de risque

Odds Ratio

Cote (odds) d'un événement :

ratio entre la probabilité qu'il se produise et la probabilité qu'il ne se produise pas.

Odds ratio :

ratio des cotes entre un groupe A et un groupe B.

Intuitivement : quantifier l'impact d'une variable sur une autre.

Section 3

Régression logistique

Régression logistique

Rappel

Expliquer chaque observation y en utilisant $X_1, \ldots, X_n$:

$$\log\left(\text{Odds}(y)\right) = \log\left(\frac{\Pr(y=1)}{\Pr(y=0)}\right) \tag{1}$$

$$= \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n. \tag{2}$$

On cherche les $(\beta_k)_{k=0}^n$ qui expliquent le mieux nos données.

Régression logistique

Rappel

Vraisemblance d'une observation :

$$V_\beta(X_1, \ldots, X_n, y) = \begin{cases} \Pr_\beta(y = 0 \mid X_1, \ldots, X_n) & \text{si y} = 0\ ; \\ \Pr_\beta(y = 1 \mid X_1, \ldots, X_n) & \text{si y} = 1. \end{cases} \quad (3)$$

Intuition : plausibilité de β comme explication de nos observations.

Régression logistique

Rappel

On cherche β tel que V_β soit maximal.

Intéressant car

$$\text{OddsRatio}(X_k) = \exp(\beta_k). \tag{4}$$

Régression logistique

Le cas centralisé

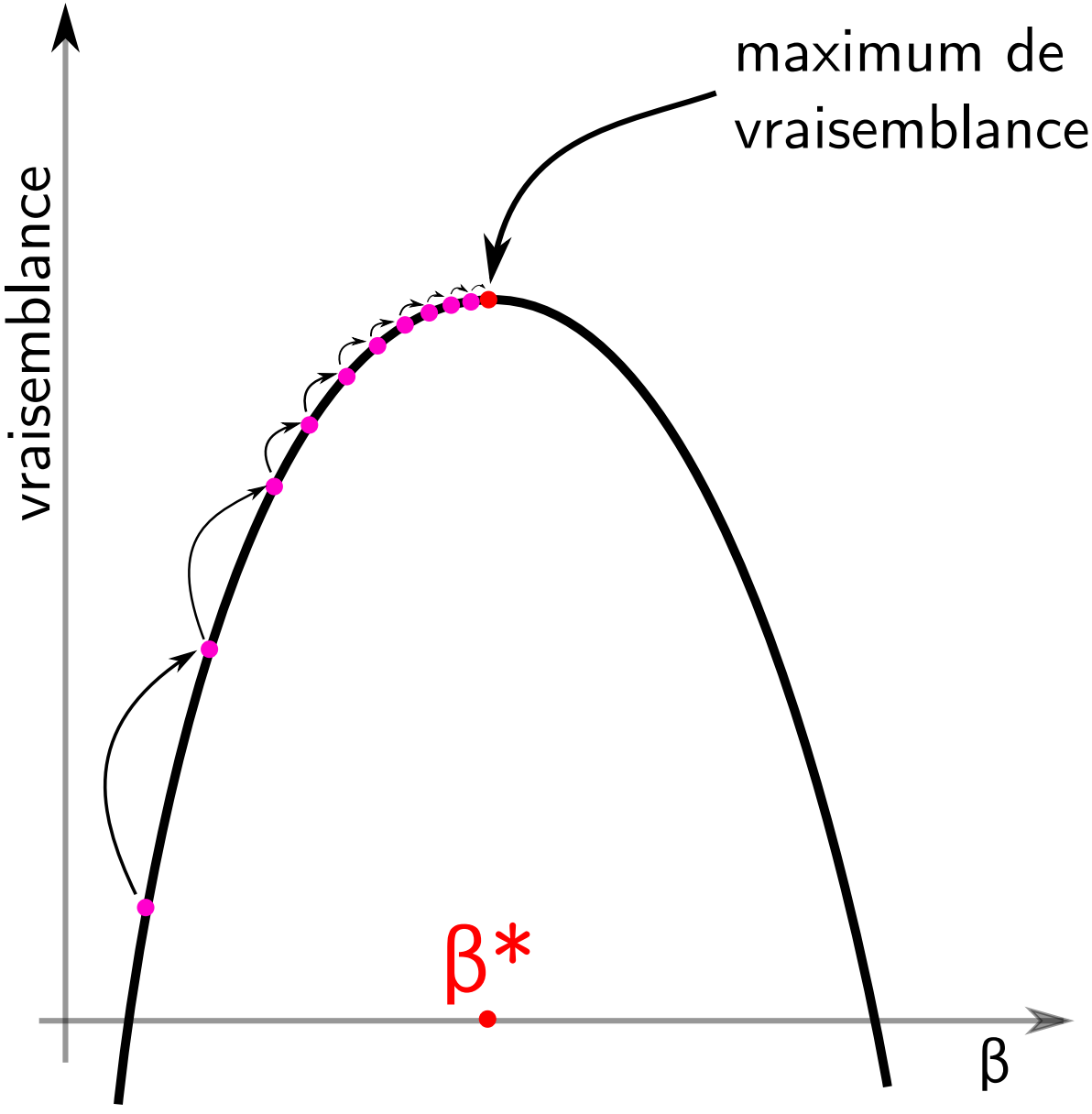
Toutes les données sont sur place.

Algorithme itératif simple (“descente” de gradient) :

- choisir une valeur initiale $\beta(0)$;
- les mettre à jour à partir des données :

$$\beta(t+1) = \beta(t) + \alpha \nabla V_{\beta(t)}(X_1, \ldots, X_n, y) \; ; \qquad (5)$$

- itérer jusqu’à convergence.


maximum de
vraisemblance
vraisemblance
β*
β

Régression logistique

Le cas décentralisé

On ne connaît que nos données... tant pis on fait pareil !

Pour chaque centre c :

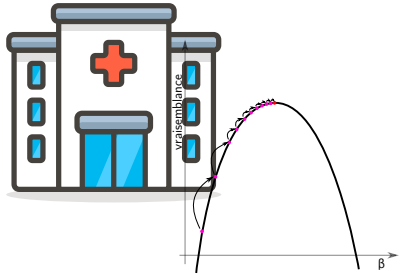
- choisir une valeur initiale $\beta^c(0)$;
- les mettre à jour à partir des données :

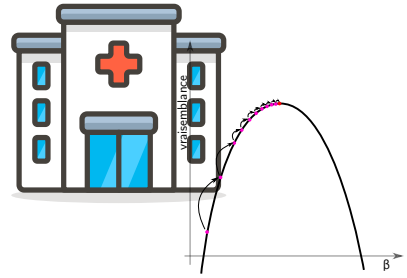
$$\beta^c(t+1) = \beta^c(t) + \alpha \nabla V_{\beta^c(t)}(X_1^c, \ldots, X_n^c, y^c). \qquad (6)$$

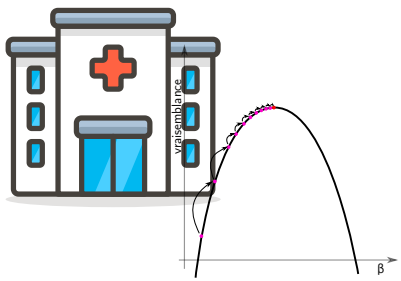
$\rightarrow$ Chaque centre c a une nouvelle valeur $\beta^c(t+1)$!

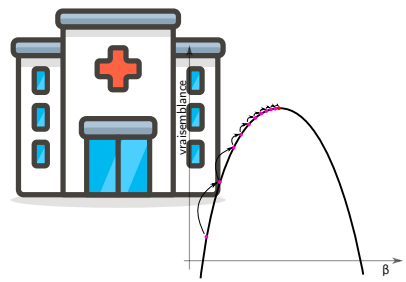
Régression logistique

Le cas décentralisé









Régression logistique

Le cas décentralisé

β^1

β^2

β^3

β^4

Régression logistique

Le cas décentralisé

Chaque centre a une valeur $\beta^c(t+1)$... et l'envoie aux autres !

Pour chaque centre c :

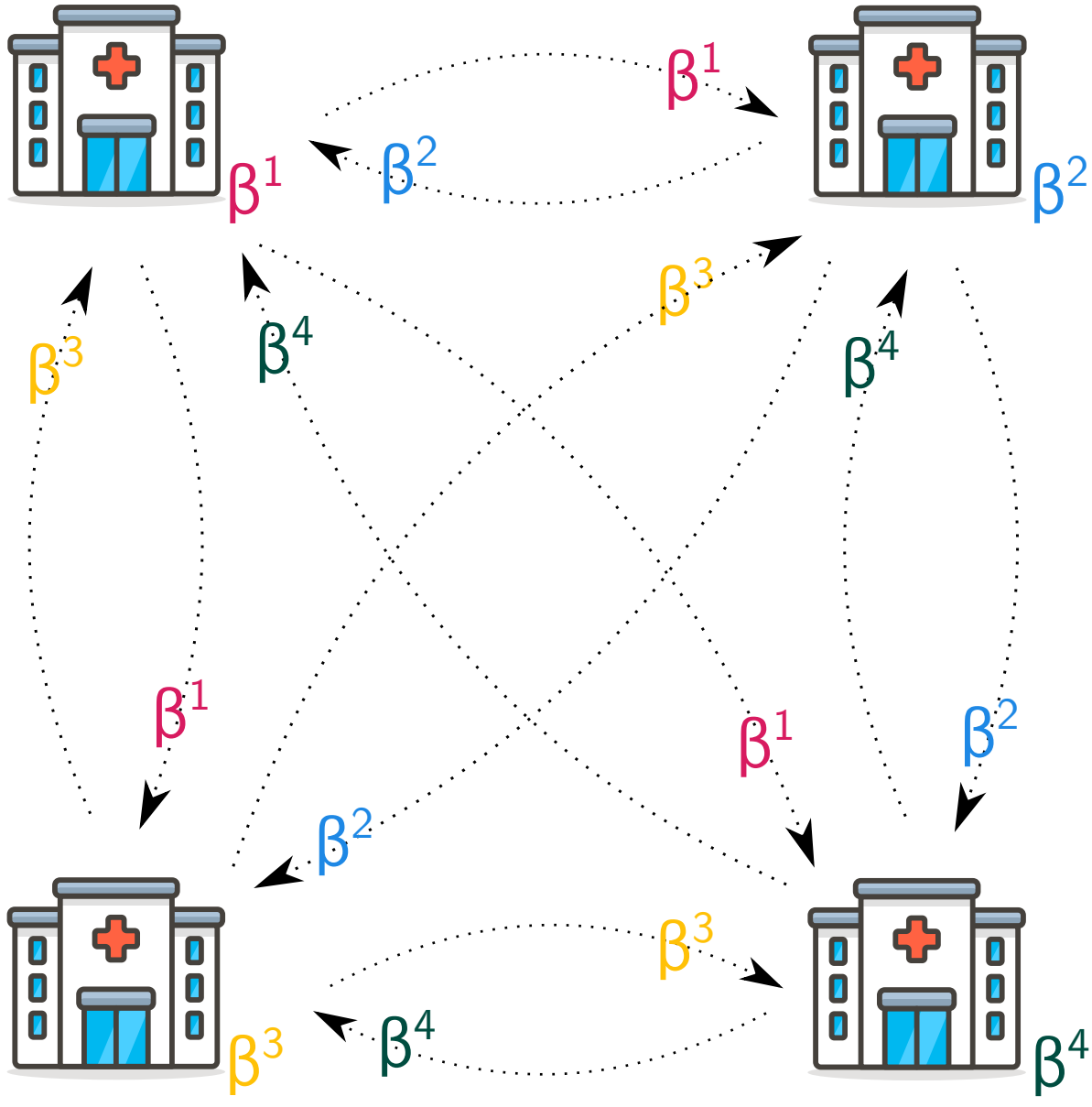
- choisir une valeur initiale de $\beta^c(0)$;
- les mettre à jour à partir des données ;
- calculer la moyenne des valeurs obtenues par chacun :

$$\beta^c(t+1) = \frac{1}{K}\left(\beta^1(t+1) + \beta^2(t+1) + \cdots + \beta^K(t+1)\right), \tag{7}$$

où K est le nombre total de centres.

Régression logistique

Le cas décentralisé



Régression logistique

Le cas décentralisé

$1/4 \times (\beta^1 + \beta^2 + \beta^3 + \beta^4)$

$1/4 \times (\beta^1 + \beta^2 + \beta^3 + \beta^4)$

$1/4 \times (\beta^1 + \beta^2 + \beta^3 + \beta^4)$

$1/4 \times (\beta^1 + \beta^2 + \beta^3 + \beta^4)$

Régression logistique

Le cas décentralisé

Pour chaque centre c :

- choisir une valeur initiale de $\beta^c(0)$;
- les mettre à jour à partir des données ;
- calculer la moyenne des $\beta^1(t+1), \ldots, \beta^K(t+1)$ obtenus par chacun ;
- itérer jusqu'à convergence.

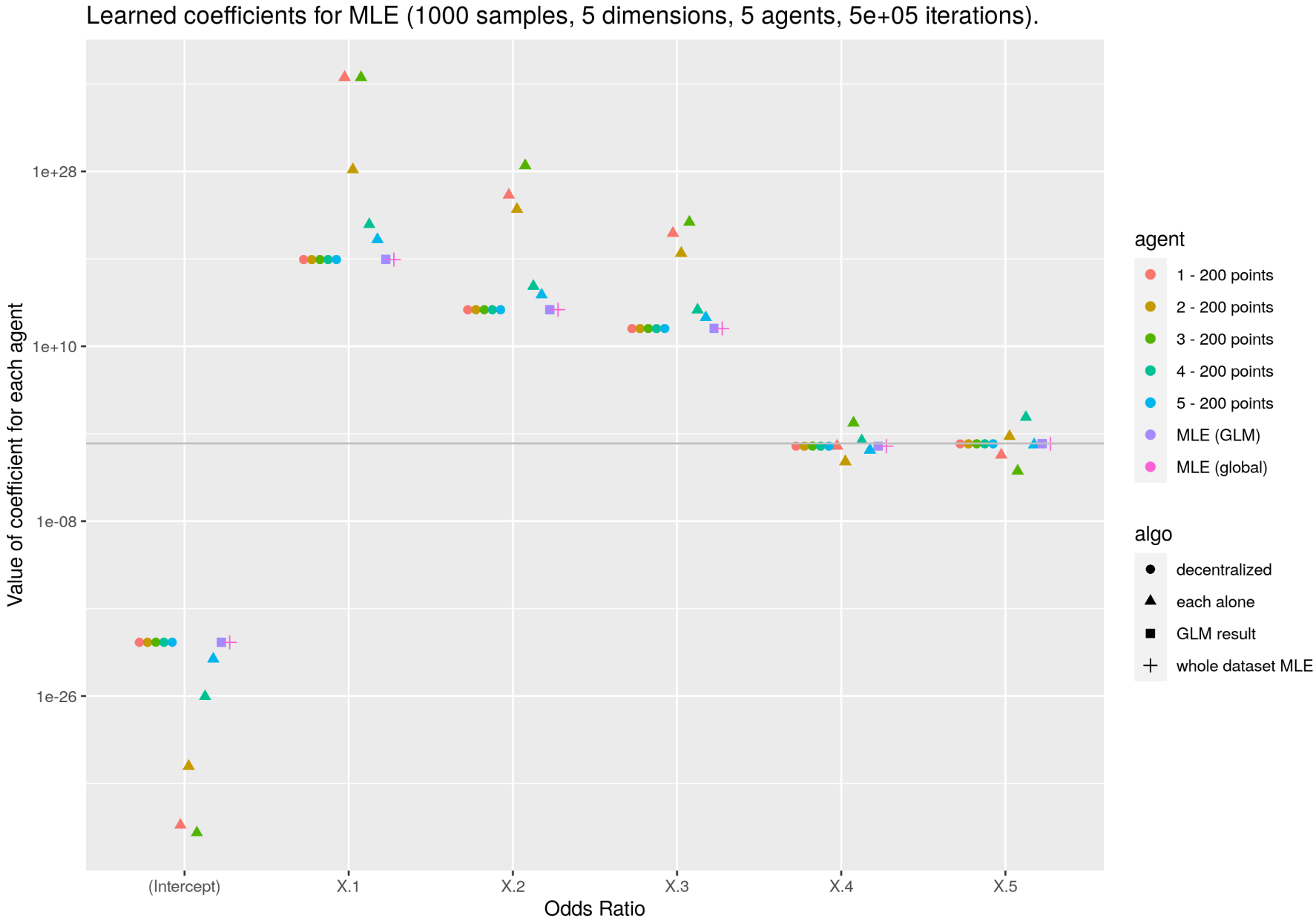
Régression logistique

Le cas décentralisé

On obtient les mêmes résultats que dans le cas centralisé !

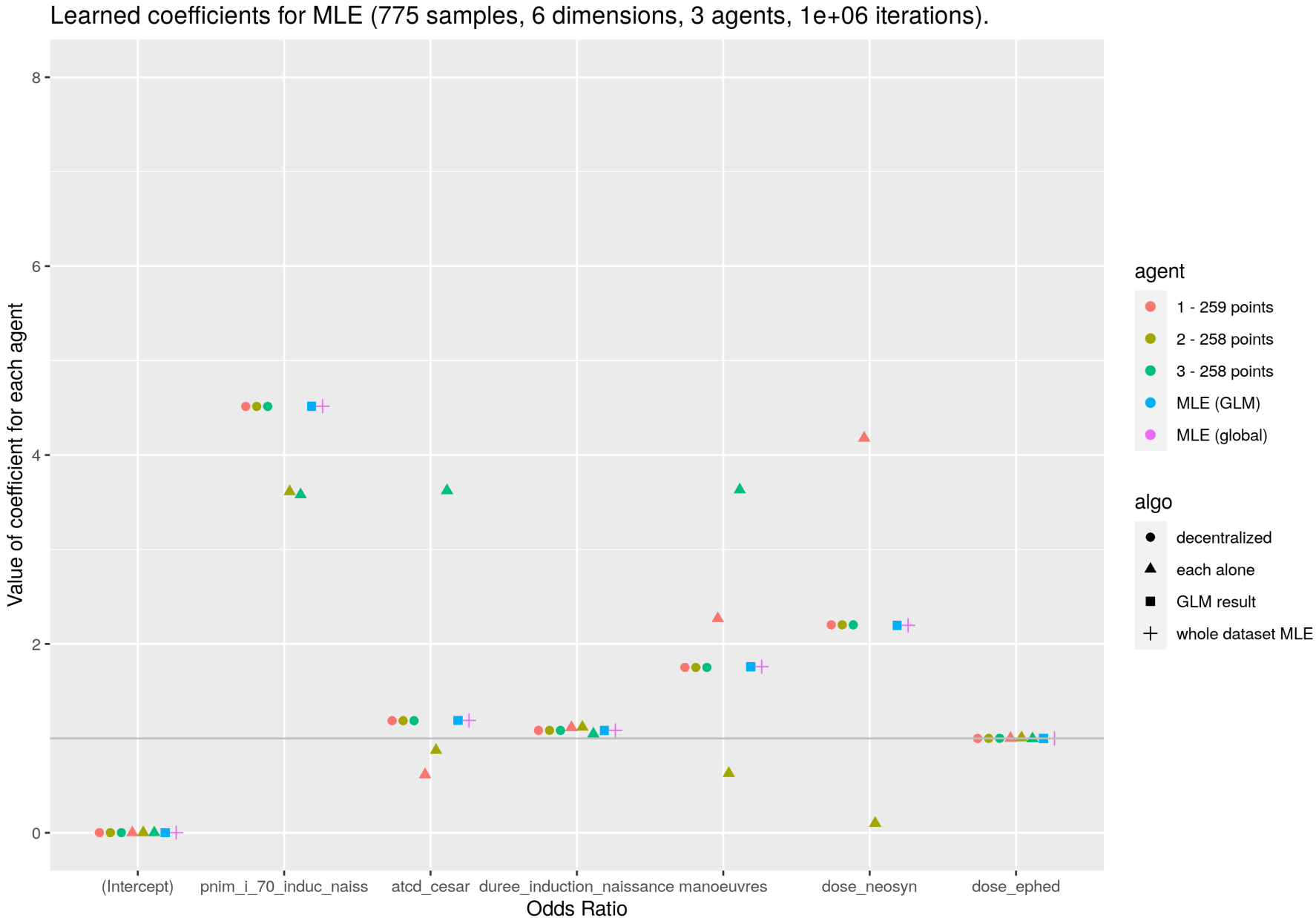
Régression logistique

Le cas décentralisé



Régression logistique

Le cas décentralisé



Régression logistique

Et les intervalles de confiance ?

La matrice de covariance des observations Cov donne

$$\text{ErreurStandard} = \sqrt{diag(Cov)}. \tag{8}$$

D'où $\beta_k^{réel} = \beta_k \pm z(0.025) \times \text{ErreurStandard}_k$.

Régression logistique

Et les intervalles de confiance ?

Remarquons :

$$Cov = M^{-1} \tag{9}$$

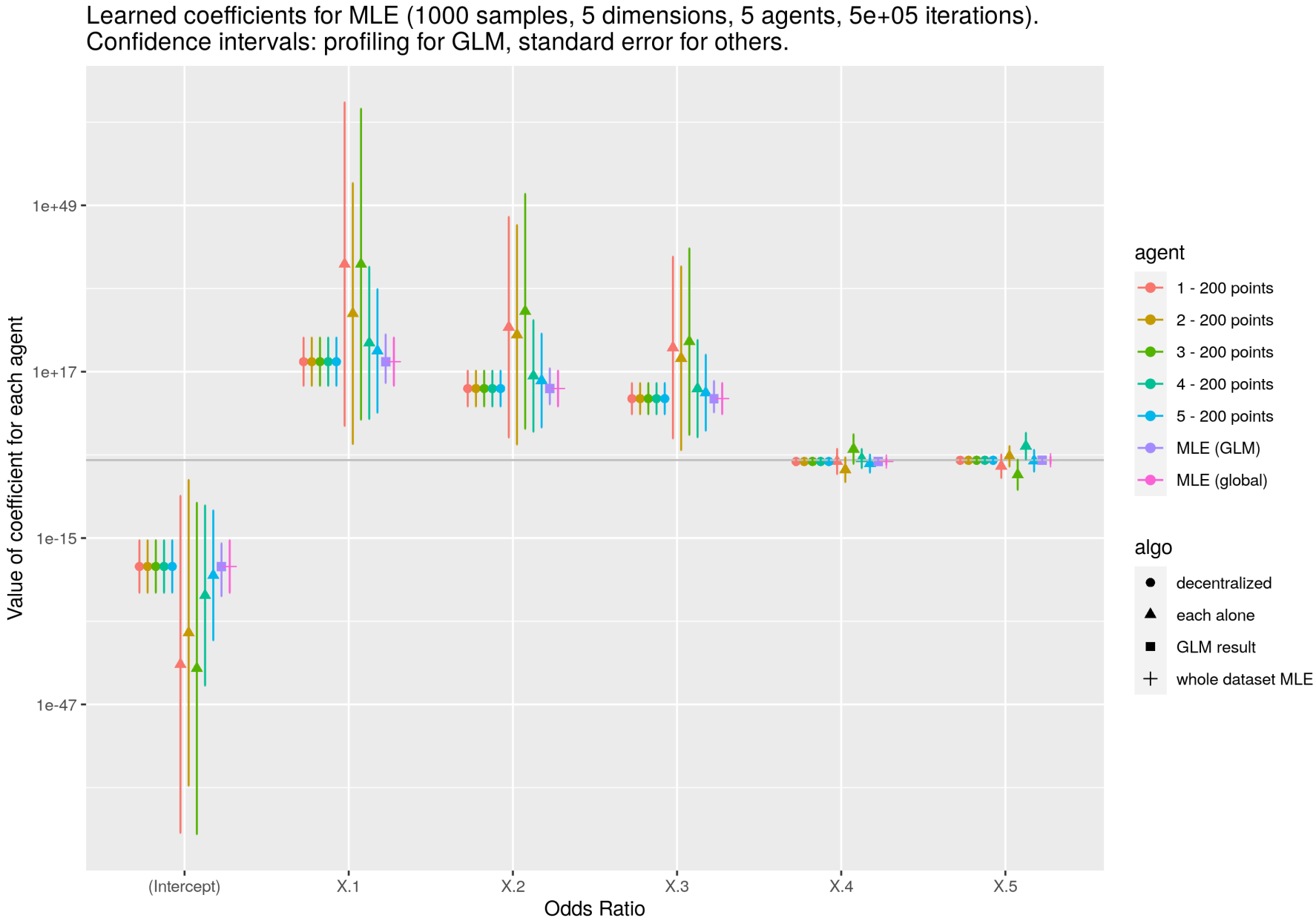
$$= (M^{(\text{centre 1})} + M^{(\text{centre 2})} + \ldots + M^{(\text{centre K})})^{-1}, \tag{10}$$

où chaque $M^{(\text{centre c})}$ est calculé localement.

Conclusion : on peut calculer des intervalles de confiance **sans envoyer les données directement**.

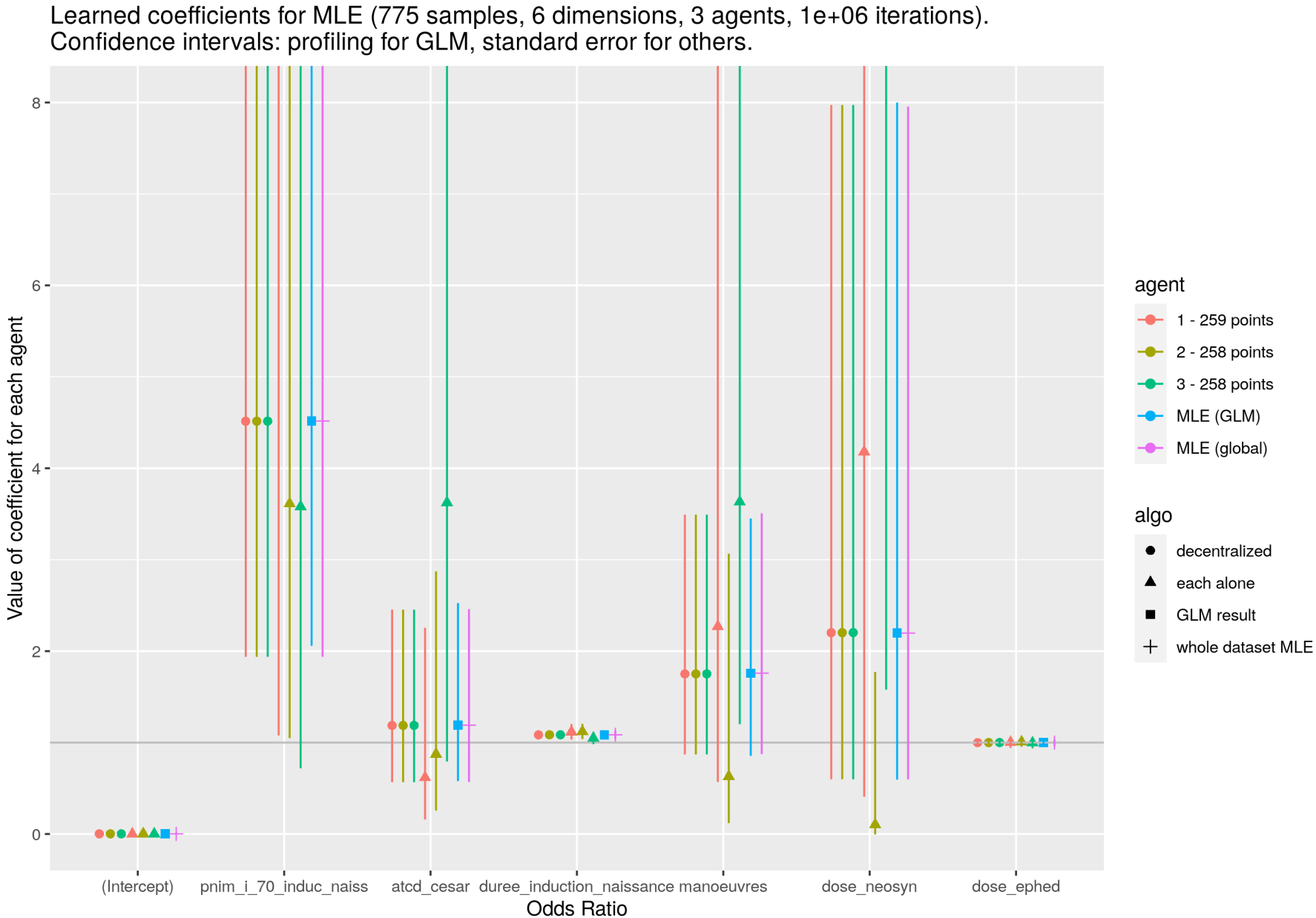
Régression logistique

Et les intervalles de confiance ?



Régression logistique

Et les intervalles de confiance ?



Section 4

Confidentialité des données

Confidentialité des données

Quelles sont les informations envoyées ?

À chaque itération :

- les coefficients β^c.

À la fin :

- les matrices $M^{(\text{centre c})}$ "de covariance".

Ils peuvent contenir des informations sensibles.

Confidentialité des données

Quelles sont les informations envoyées ?

Deux types de protections :

- aggrégation sécurisée : “protéger les sources” ;
- differential privacy : “brouiller les résultats”.

Confidentialité des données

Aggregation sécurisée : “Protéger les sources”

Dans les deux cas, on calcule une somme :

$$\beta = \beta^{(\text{centre 1})} + \beta^{(\text{centre 2})} + \cdots + \beta^{(\text{centre K})} \; ; \tag{11}$$

$$M = M^{(\text{centre 1})} + M^{(\text{centre 2})} + \cdots + M^{(\text{centre K})}. \tag{12}$$

On peut être un peu plus malins... au prix du nombre de communications.

Confidentialité des données

Aggregation sécurisée : “Protéger les sources”

On rajoute du bruit :

- chaque paire de centres c et v s'accorde sur un vecteur aléatoire $s_{c,v} = s_{v,c}$;
- chaque centre c envoie $\beta^{(\text{centre c})} + \sum_{v<c} s_{c,v} - \sum_{v>c} s_{c,v}$.

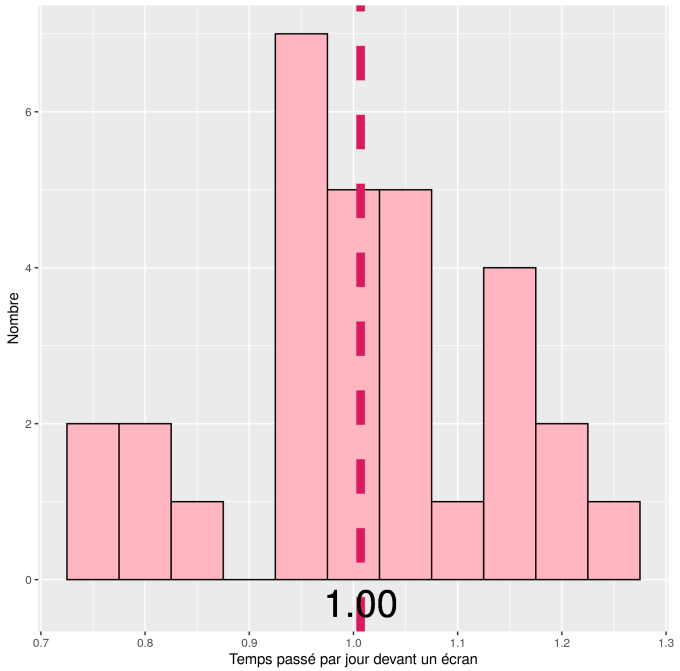
Et tout le bruit s'annule !

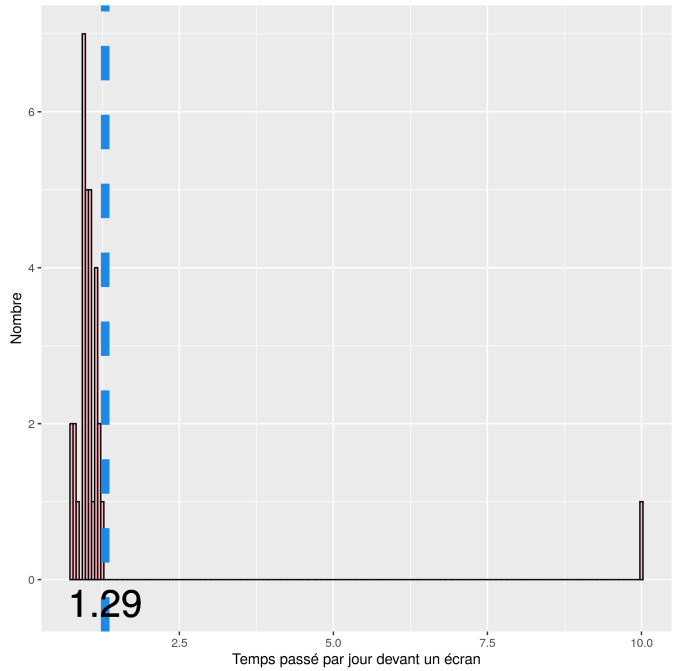
$\rightarrow$ Chaque centre obtient la somme sans connaître aucun des autres termes.

Confidentialité des données

Differential privacy : “brouiller les résultats”

Les données agrégées peuvent contenir des informations sensibles...





Confidentialité des données

Differential privacy : "brouiller les résultats"

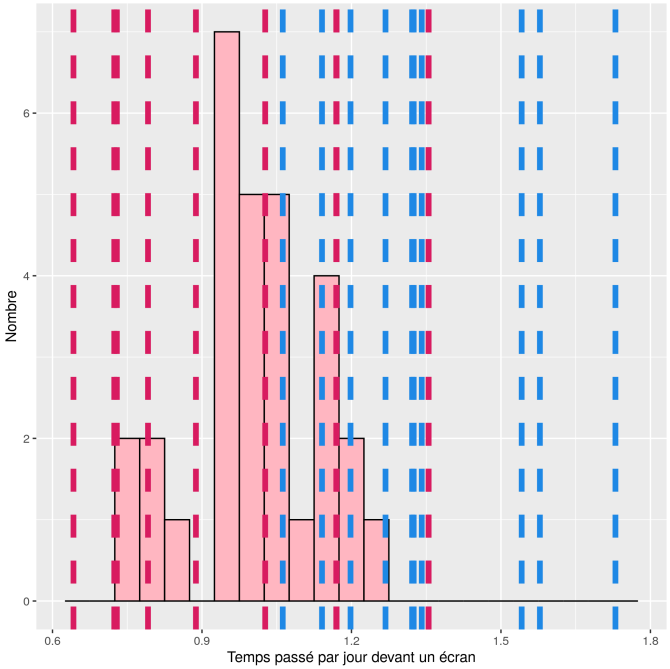
Idée : rajouter du bruit pour cacher la vraie moyenne.

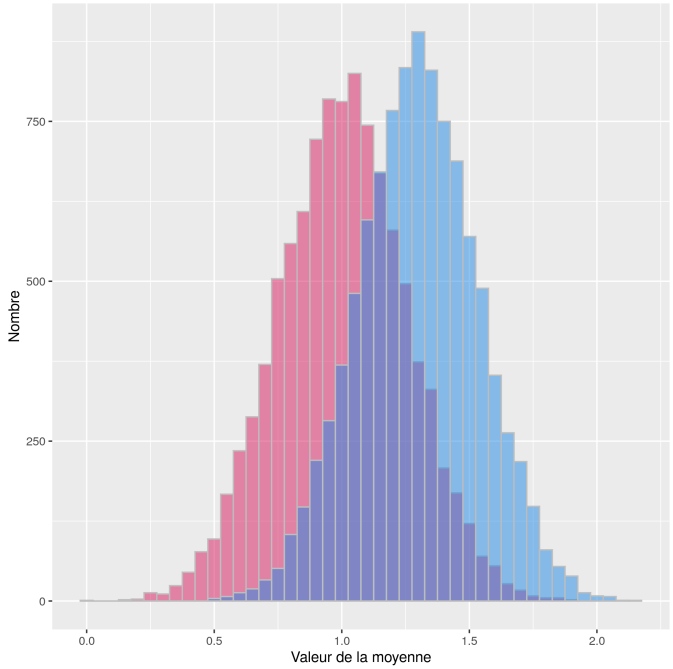
$$\beta = \frac{1}{K} \sum_{u=1}^{K} \beta^u + \text{bruit}. \tag{13}$$

$\rightarrow$ le résultat est moins précis mais il peut rester exploitable.

Confidentialité des données

Differential privacy : “brouiller les résultats”





Confidentialité des données

Differential privacy : “brouiller les résultats”

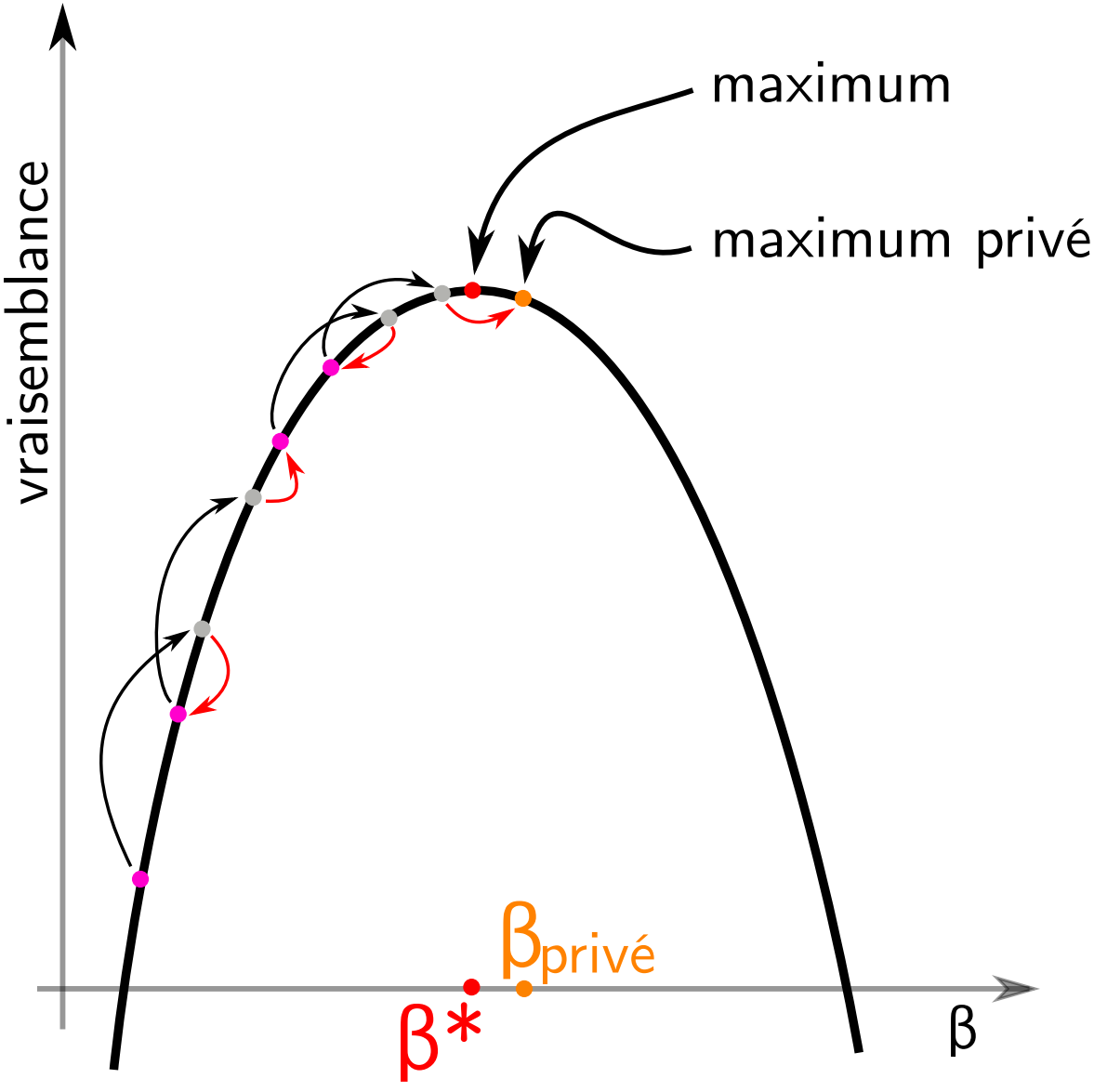
Intuition :

un algorithme est dit “differentially private” si, pour deux jeux de données proches, il donne des résultats proches

→ on bruite les résultats de l'algorithme pour réduire l'influence des données

Confidentialité des données

Differential privacy : “brouiller les résultats”

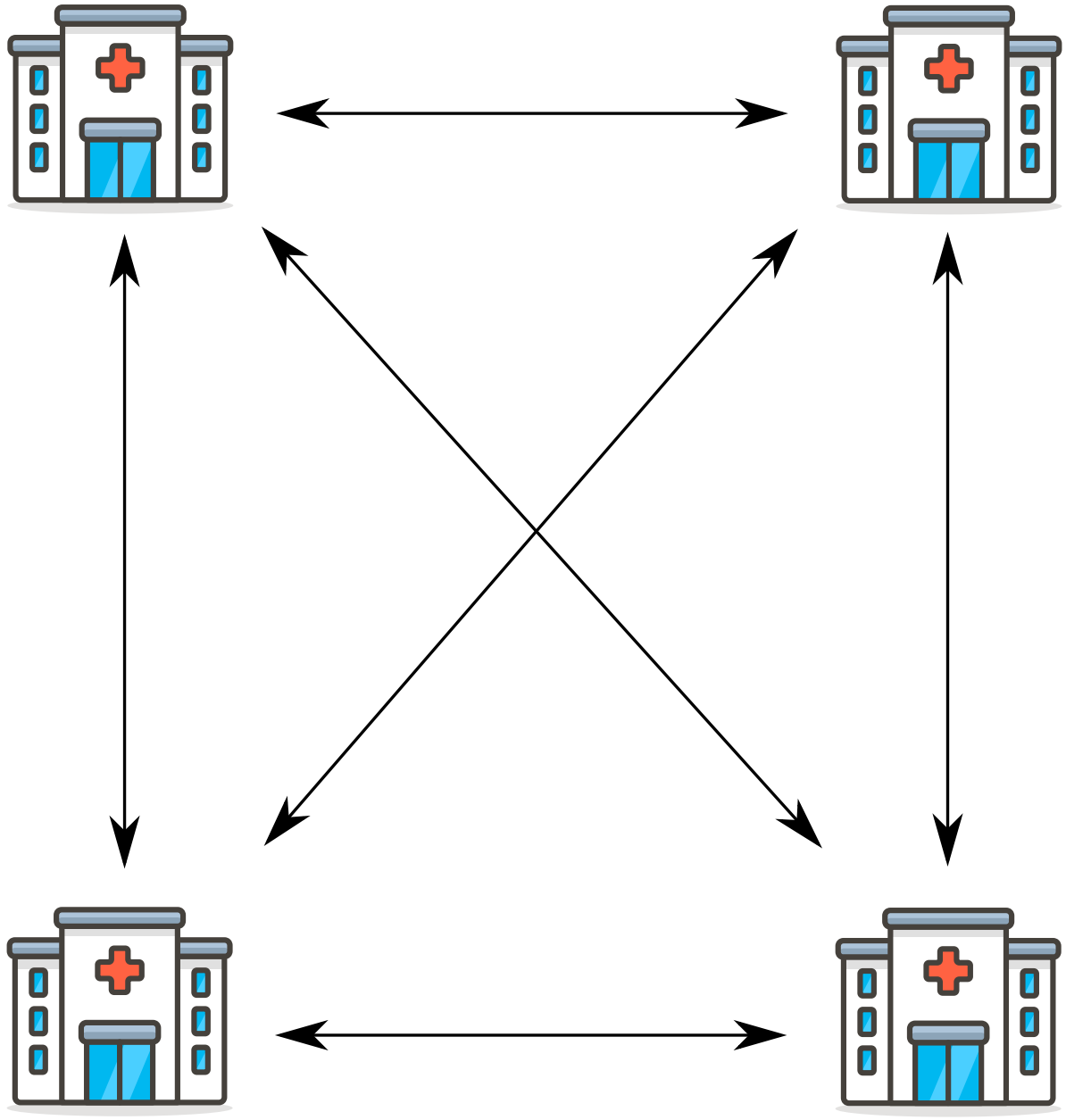


Section 5

Des idées à développer...

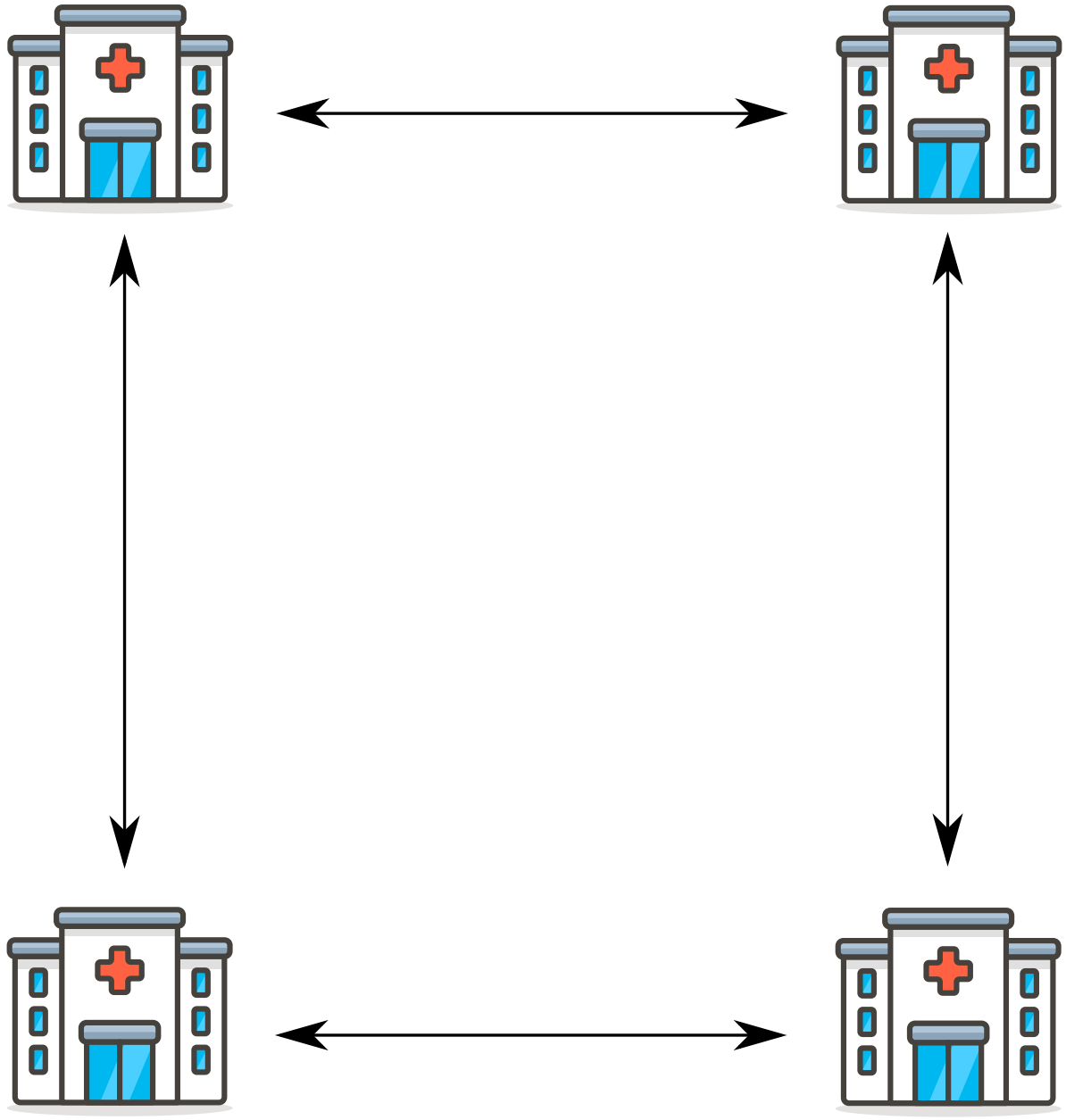
Des idées à développer...

Topologie du réseau



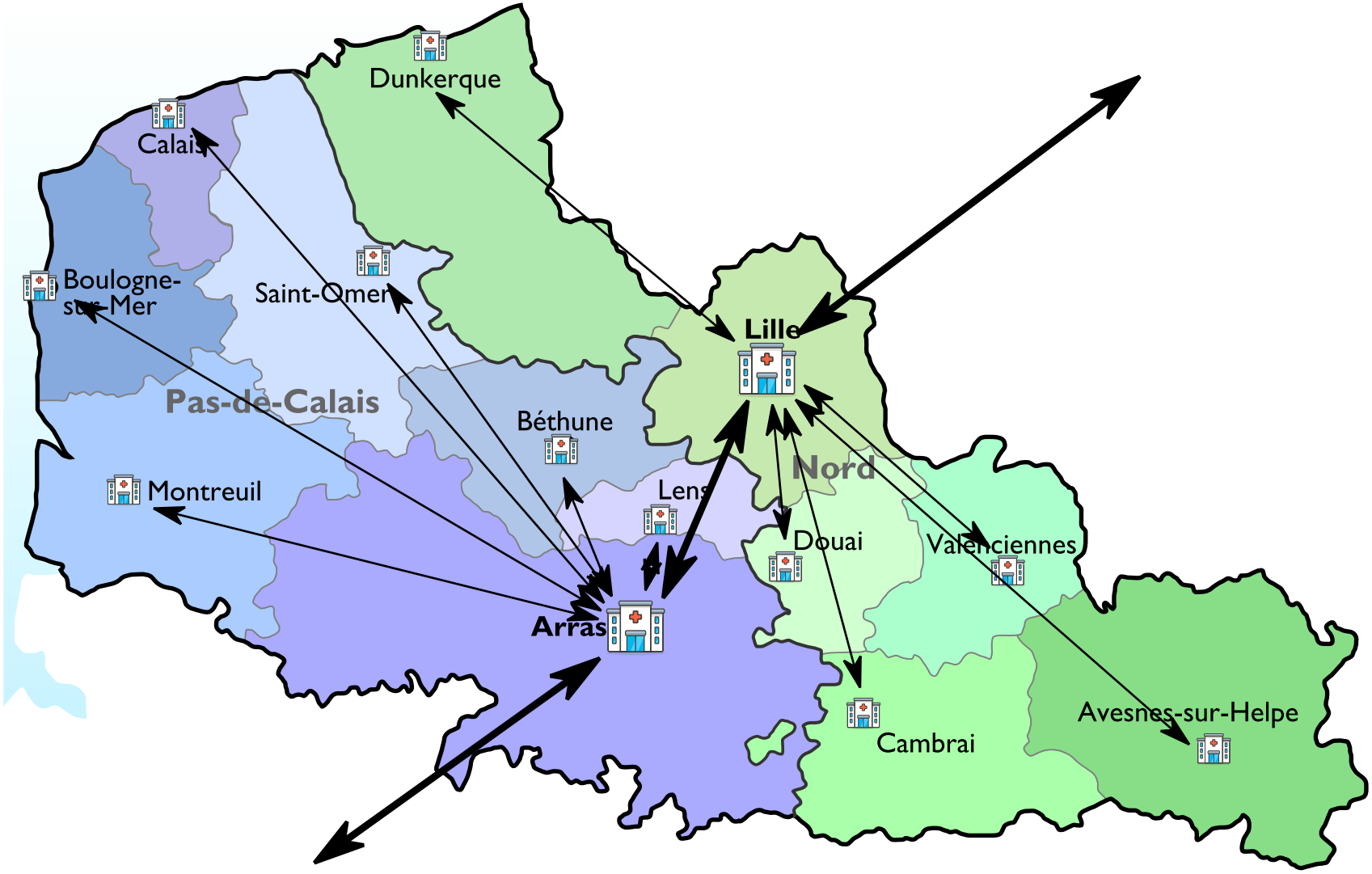
Des idées à développer...

Topologie du réseau



Des idées à développer...

Topologie du réseau



Des idées à développer...

Vis à vis de la loi ?

Les données restent sur place !

Donc si on prouve que l'information identifiante circule de façon très limitée, que dit la loi ?

Des idées à développer...

Études en temps réel

Si la base de données évolue : relancer une étude régulièrement.

Mettre à jour les facteurs de risque en temps réel.

Section 6

Conclusion

Conclusion

- On peut faire des études décentralisées !
- Les données restent à la maison.
- Les informations qui circulent peuvent être très limitées.
- Faciliter la réalisation d'études à grande échelle ?

Conclusion

Il reste des questions à développer (dont certaines existent aussi dans l'approche centralisée) :

- topologie du réseau ?
- contraintes légales (et de confidentialité) plus simples ou pas ?
- communications asynchrones ?
- aspect pratique pour réaliser ces études ?
- distributions de données différentes dans les centres ?
- possibilité pour un centre de trop influencer le résultat ?