# 100% ACTUAIRES & 100% DATA SCIENCE

INSTITUT DES ACTUAIRES


## 29 / NOV / 2019

Hôtel Marriott Rive Gauche
Paris 14ème

100% ACTUAIRES &
100% DATA SCIENCE

INSTITUT DES ACTUAIRES

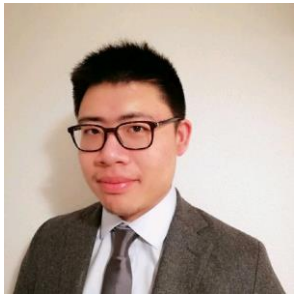
# Apprentissage actif pour la détection des catégories dans des champs textuels

4, rue Chauveau-Lagarde - 75008 Paris - tél. 01 44 51 72 72 - fax 01 44 51 72 73
www.institutdesactuaires.com

# Qui sommes-nous ?

**Marc JUILLARD**,
Directeur du DataLab de Société Générale Assurances

**Romuald ELIE**,
Professeur de Mathématiques à l'UPEM et professeur associé à l'ENSAE



**François HU**,
Doctorant CIFRE à la Société Générale Assurances et à l'ENSAE

# Sommaire

1. Introduction

2. Apprentissage automatique sur les données textuelles

3. Apprentissage actif

4. Différentes stratégies d'apprentissage actif

5. Expérimentations

6. Conclusion

7. Prochaines étapes

# Contexte et objectif



## Objectif ?

Intégrer les retours des Clients chez Société Générale Assurances.

## Problème

1. Impossibilité pour les experts humain d'analyser toutes ces quantités
2. Données non labellisées

## Solution

1. Volume : application d'algorithme d'IA
2. Labels : active Learning

# Avantages de l'active learning



## L'active learning c'est quoi ?

- Sélectionner les meilleurs données à faire labelliser

## Pourquoi faire de l'active learning ?

- Diminue le volume de données nécessaire
- Diminue le temps demandé aux experts
- Cadre de rentabilité pour chaque nouvelle donnée

## Les données Net Promoter Score (NPS)

- Collection de verbatims des assurés afin de **mesurer la satisfaction client**
- Environ 100 000 verbatims

# Processus d'apprentissage NLP





Le dictionnaire permet de définir le corpus de mot sur lequel sera bâti notre algorithme prédictif. Il est donc nécessaire en phase d'apprentissage mais surtout pour le déploiement du modèle.

- Tokenization : processus qui sépare une séquence (les textes) en une liste de tokens (mots)
- Nettoyer si besoin le corpus (faut d'orthographe, synonyme, stop-words,..)

**Exemple :**

"Je suis très satisfait par le service rendu. Conseiller très agrébale,……"

→

[Je, être, satisfait, ~~par~~, ~~le~~ ,service, rendu,…., conseiller, très, agréable,…]

Dictionnaire de taille $D$

# Processus d'apprentissage NLP





Les modèles de machine learning travaillant avec des nombres il est nécessaire d'encoder les données textuelles. Plusieurs approches sont envisageables :

- Encodage binaire
- Encodage fréquentiel
- Encodage TF-IDF
- Encodage par méthodes « embedding »

## Exemple :

"Je suis très satisfait par le service rendu. Conseiller très agrébale,……"

→ [Je, être, satisfait, ~~par~~, ~~le~~ ,service, rendu,…., conseiller, très, agréable,…]

Dictionnaire de taille $D$

### Exemple représentation binaire

On crée $D$ variables (une pour chaque mot du dictionnaire).
Chaque document est alors représenté par le simple comptage des mots présents

| agréable | avoir | conseiller | cible | … | très | train | satisfait |
|---|---|---|---|---|---|---|---|
| X = [1, | 0, | 1, | 0, | …, | 1, | 0, | 1 ] |

Vecteur de taille $D$

# Processus d'apprentissage NLP





Construire et entraîner notre modèle d'apprentissage :

- Modèles de Machine Learning « classique » (arbre, xgboost, …)
- Modèles de Deep Learning (CNN, RNN, …)

# Approche non-supervisée



**Problème : pas de données labellisées et labellisation à la main impossible**

**Idée : appliquer une démarche non supervisée :**

- Encodage des différents verbatims par **doc2vev. Principale avantage : permet d'appliquer une distance entre les documents**
- Application de méthodologies non supervisées **(Kmeans et/ou CAH). Si la première est plus rapide, la deuxième permet de distinguer thèmes et sous-thèmes.**
- Labellisation automatique des 1% les plus proches des clusters.

***Promoteurs*** **: Note de recommandation entre 9 et 10**

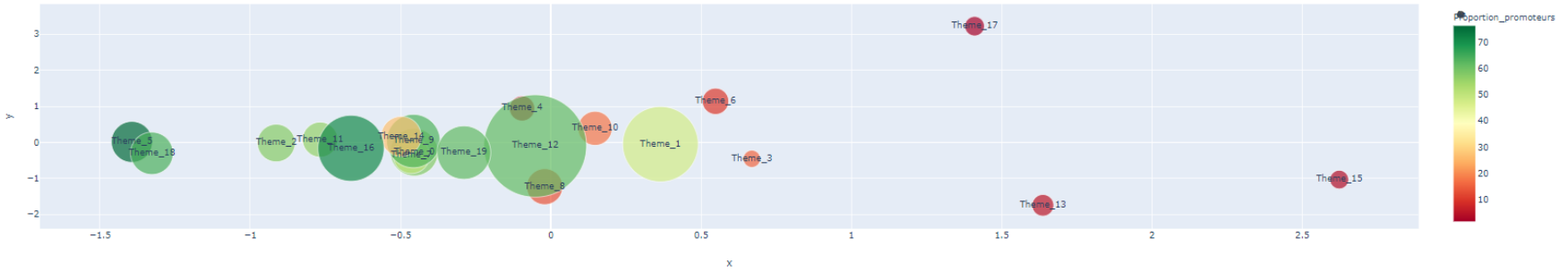
***Passifs*** **: Note de recommandation inferieure entre 7 et 8**

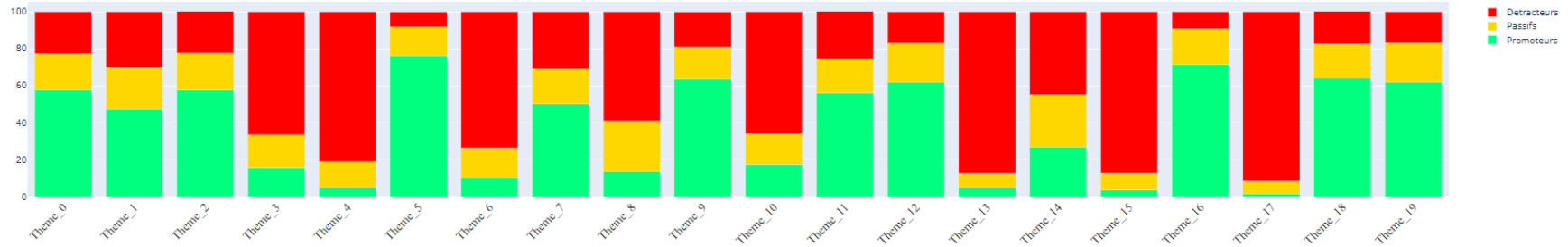
***Detracteurs*** **: Note de recommandation inferieure a 6**

## DETECTION DE THEMES DANS LES VERBATIMS NPS

Analyse de 80 000 verbatims - Periode d'Avril 2018 - Juin 2019

Segmentation des 80 000 verbatims en 20 groupes (themes)




ANALYSE DES THEMES

SELECTIONNEZ UN THEME

0

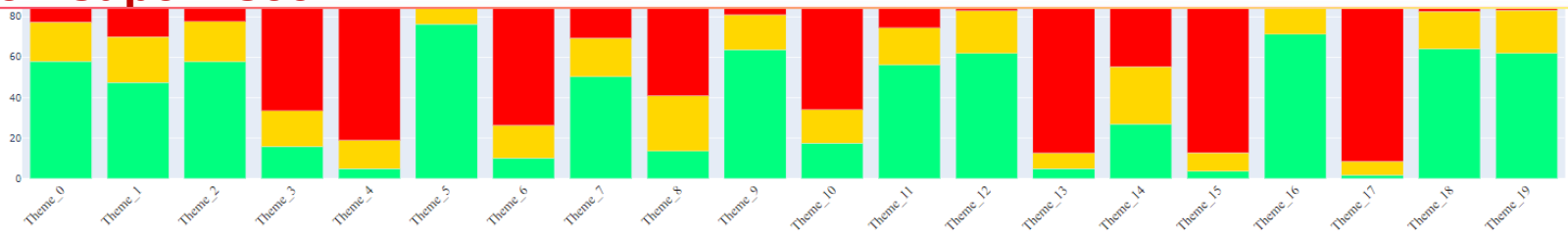
LE **THEME 0** REGROUPE **3954** verbatims

VERBATIMS REPRESENTANTS DU THEME 0

***Promoteurs*** **: Note de recommandation entre 9 et 10**

***Passifs*** **: Note de recommandation inferieure entre 7 et 8**

***Detracteurs*** **: Note de recommandation inferieure a 6**




ANALYSE DES THEMES

SELECTIONNEZ UN THEME

0

## LE **THEME 0** REGROUPE **3954** verbatims

VERBATIMS REPRESENTANTS DU THEME 0

20

1 20


```
{
  "Verbatims_representants" : [
    0 : "INTERLOCUTRICE TRES SYMPA ET SERVIABLE"
    1 : "Ton accusateur et inquisiteur de la personne"
    2 : "L'échange avec le conseiller a été très cordiale et compréhensive."
    3 : "L'écoute de mon interlocutrice était rassurante."
    4 : "La conseillère à été trés attentive et m'a aidé"
    5 : "La personne "
    6 : "Personne "
    7 : "Accueil chaleureux et conseillère à l'écoute"
    8 : "Conseillère très humaine, très compréhensive et à l'écoute."
    9 : "Efficacité et réactivité de mon interlocutrice"
    10 : "personne n'est parfait.."
    11 : "Une personne competente "
    12 : "Personne à l'écoute"
    13 : "Personne très serviable et à l ecoute"
    14 : "pour l'amabilité et l'écoute de la personne"
    15 : "il n'y a personne ."
    16 : "Conseiller très a l'écoute et reactive"
    17 : "Parce que mon interlocutrice a parfaitement fait son job."
    18 : "Conseillère très pro "
    19 : "La personne à été patiente et tres explicite "
  ]
}
```

# Sommaire

1. Introduction

2. Apprentissage automatique sur les données textuelles

3. Apprentissage actif

4. Différentes stratégies d'apprentissage actif

5. Expérimentations

6. Conclusion

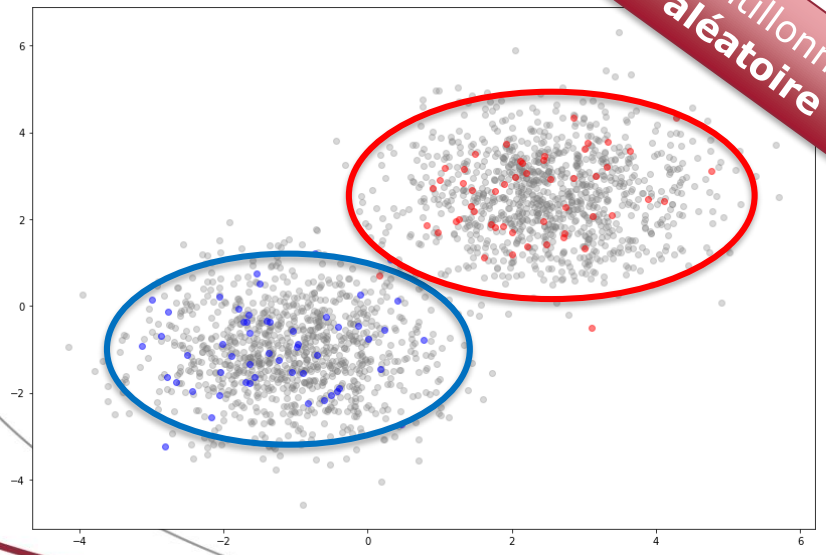
7. Prochaines étapes

INSTITUT DES ACTUAIRES

**Apprentissage actif (*Active Learning* en anglais) :**
Construire un modèle de Machine Learning **plus performant** si nous avons la possibilité de **choisir les données à étiqueter / entraîner**

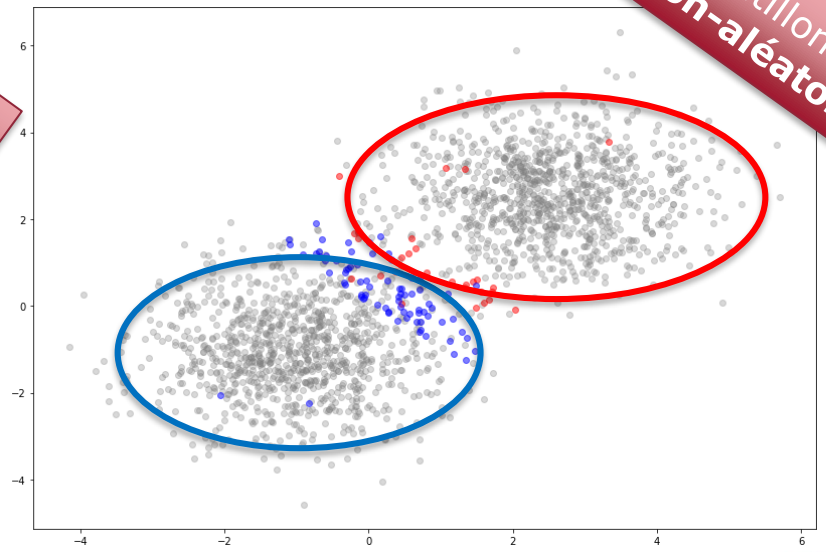
**Passive Learning**

Échantillonnage **aléatoire**



**Active Learning**

Échantillonnage **non-aléatoire**



# Quelques notations et une itération classique d'apprentissage actif

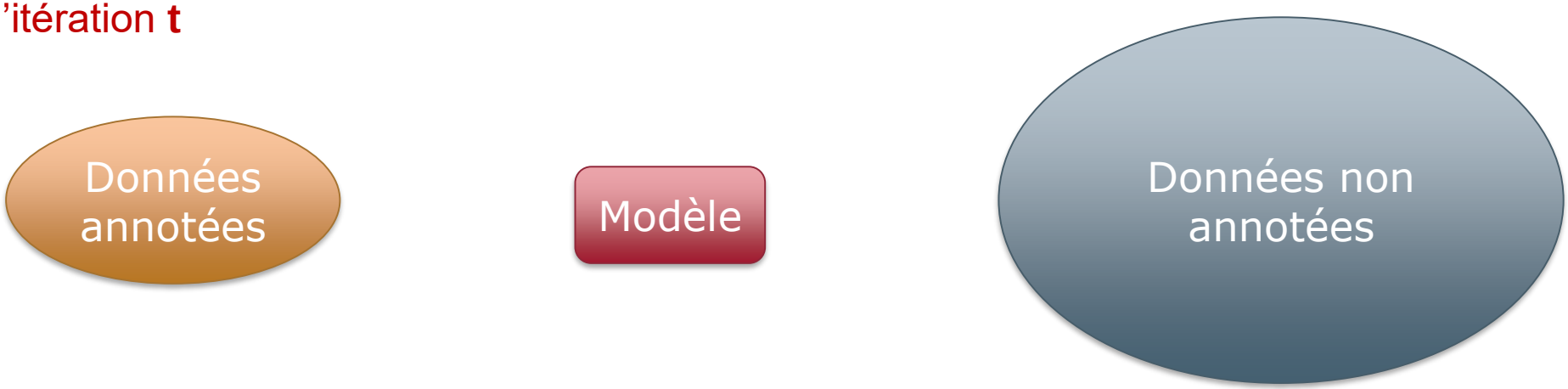


$L$ $\theta$ $U$

# Quelques notations et une itération classique d'apprentissage actif



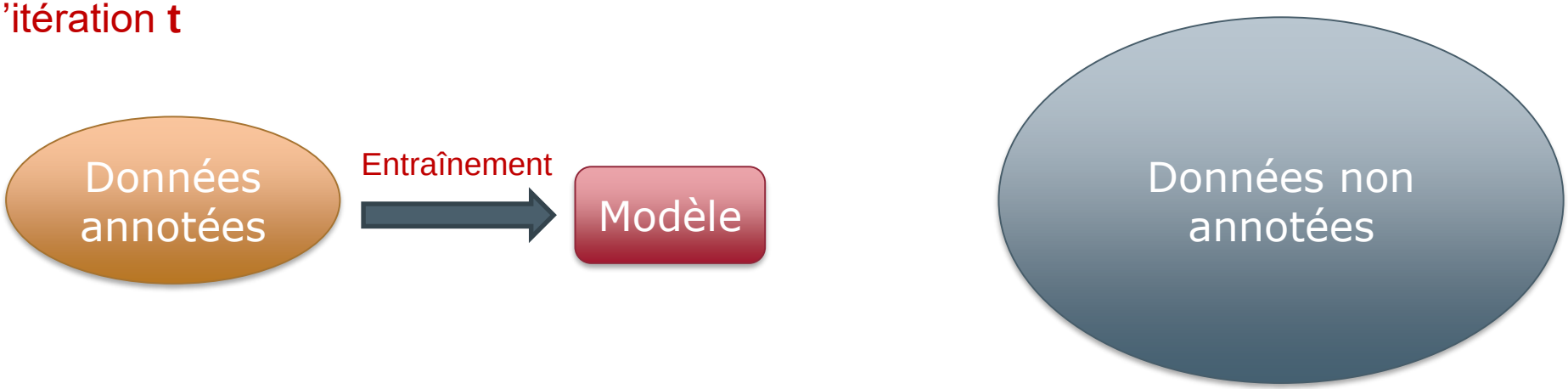
À l'itération **t**



# Quelques notations et une itération classique d'apprentissage actif



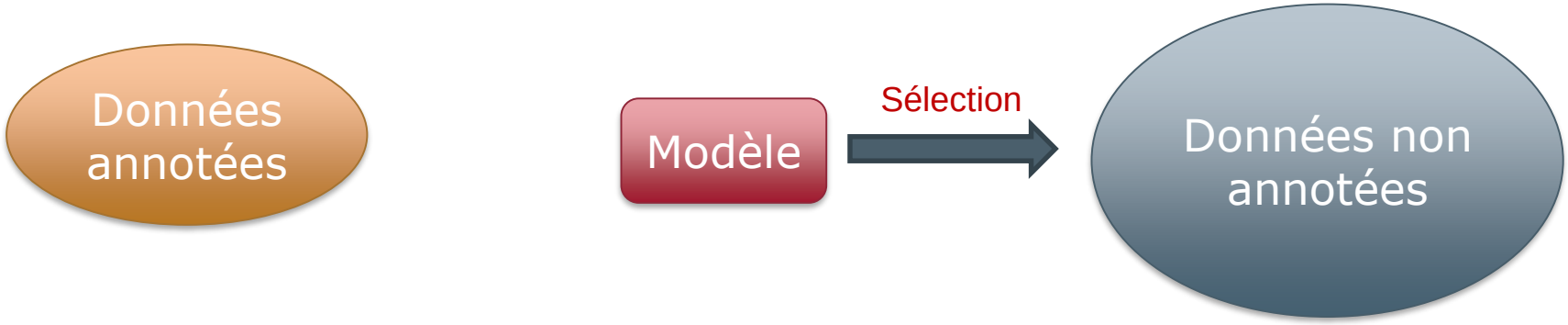
À l'itération **t**



# Quelques notations et une itération classique d'apprentissage actif



À l'itération **t**



# Quelques notations et une itération classique d'apprentissage actif



À l'itération **t**

Données annotées

Modèle

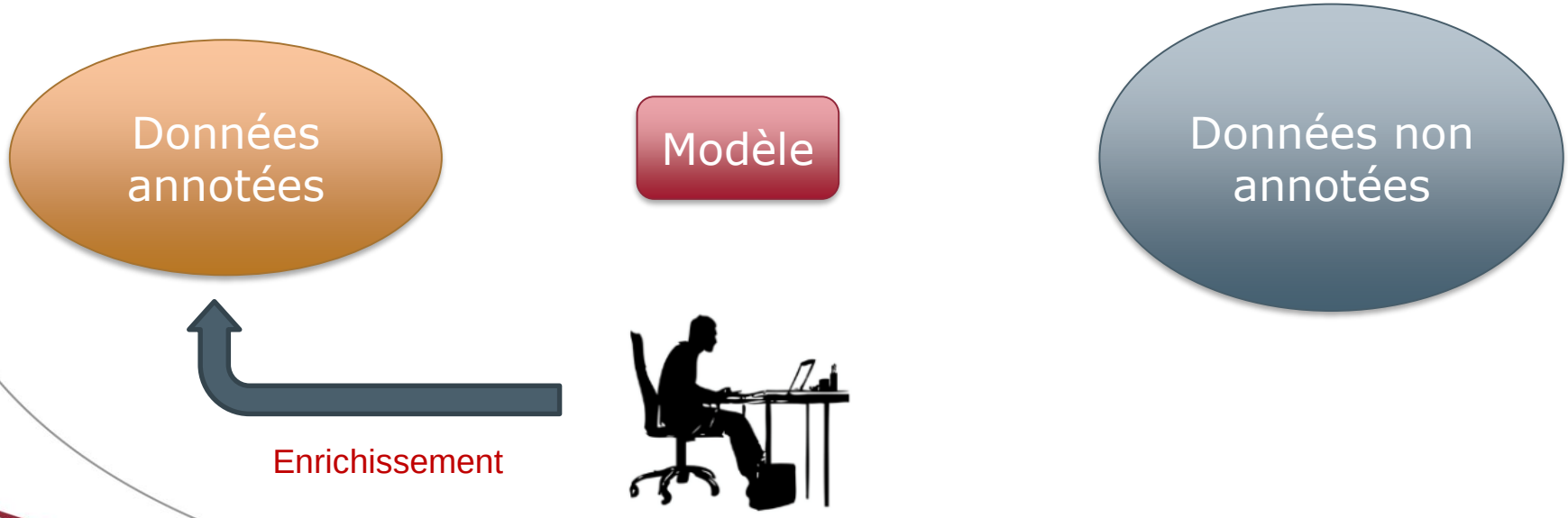
Données non annotées

Étiquetage des données

# Quelques notations et une itération classique d'apprentissage actif



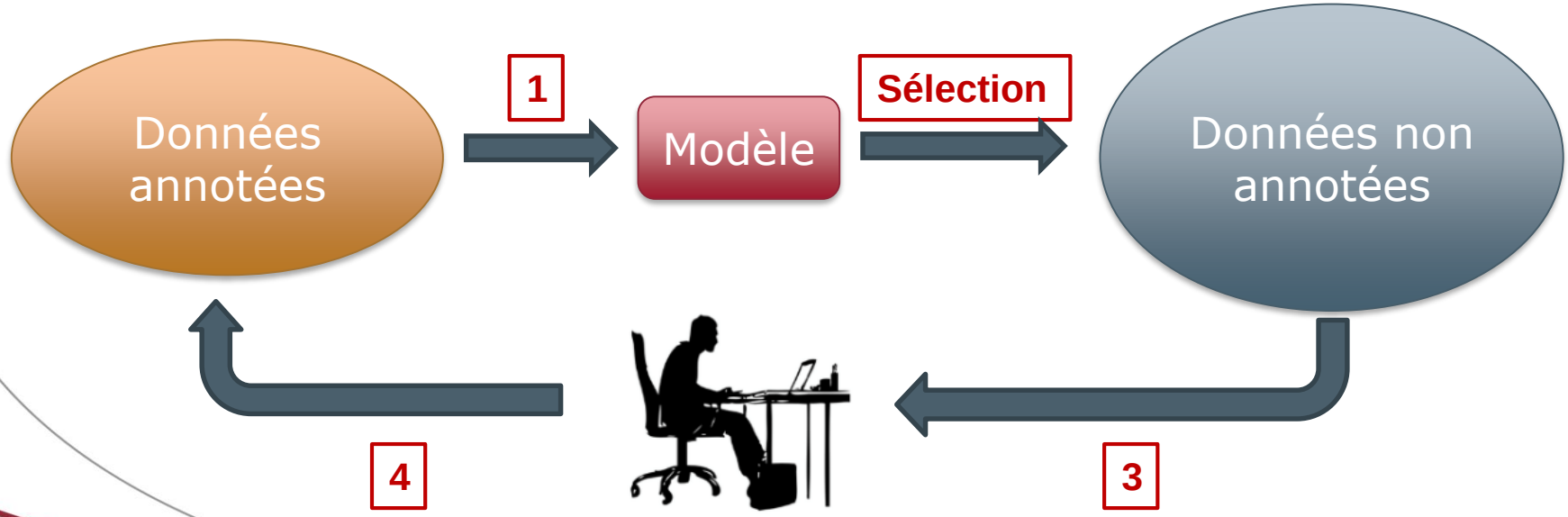
À l'itération **t**



# Quelques notations et une itération classique d'apprentissage actif



À l'itération **t+1 …**



# Sommaire

1. Introduction

2. Apprentissage automatique sur les données textuelles

3. Apprentissage actif

4. Différentes stratégies d'apprentissage actif

5. Expérimentations

6. Conclusion

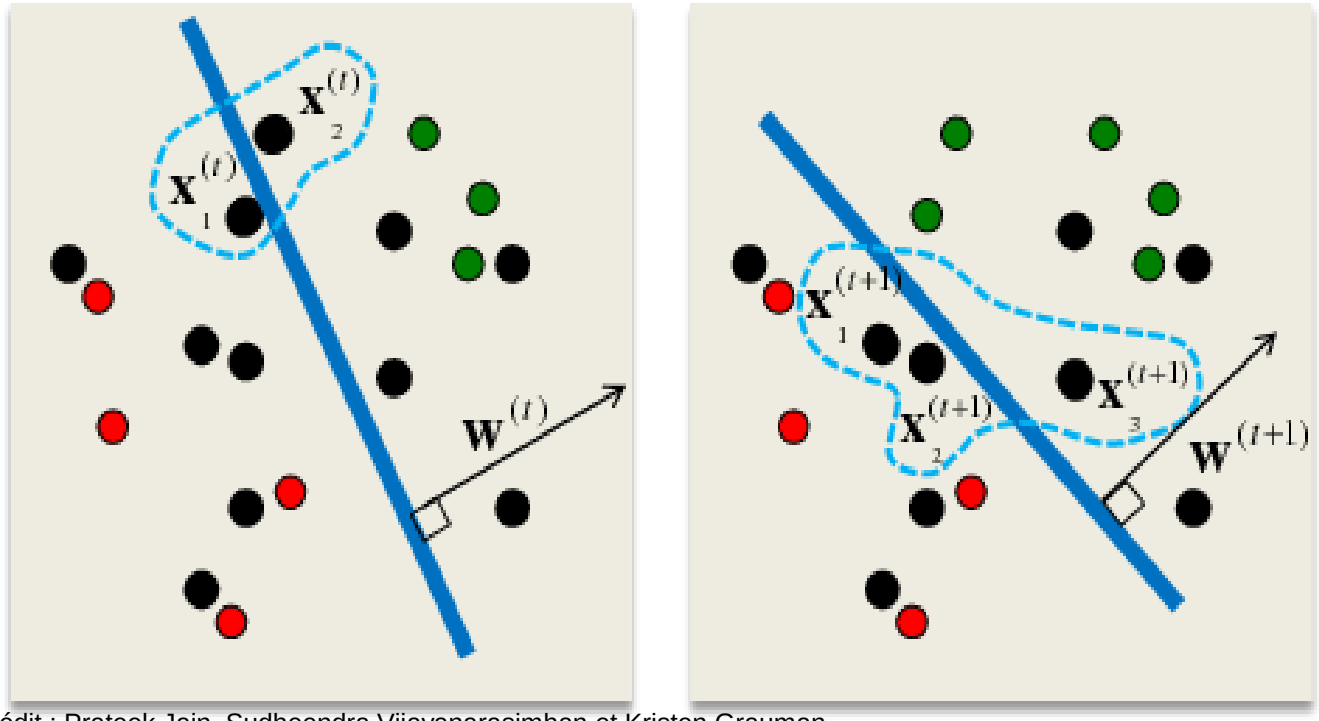
7. Prochaines étapes

# Différentes stratégies d'Active Learning



- Échantillonnage incertain
- Requête par votes
- Changement de modèle prévu
- Méthode de densité pondérée

# Échantillonnage incertain




Crédit : Prateek Jain, Sudheendra Vijayanarasimhan et Kristen Grauman

**Questions :**

- **comment choisir la région d'incertitude ?**

**Les sous-stratégies :**

- Échantillonnage « **Least Confident** »
- Échantillonnage **avec marge**
- Échantillonnage **par entropie**

# Échantillonnage « Least Confident »

| | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| $p_\theta(y_1 \mid x)$ | 0.62 | 0.16 | 0.10 |
| $p_\theta(y_2 \mid x)$ | 0.36 | 0.10 | 0.70 |
| $p_\theta(y_3 \mid x)$ | 0.01 | 0.14 | 0.09 |
| $p_\theta(y_4 \mid x)$ | 0.01 | 0.60 | 0.11 |

$$x^*_{LC} = \arg\max_x \{ 1 - P_\theta(\hat{y} \mid x) \}$$

où $\hat{y} = \arg\max_y \{P_\theta(y \mid x)\}$ est la classe la plus probable sous le modèle

# Échantillonnage « Least Confident »

$$
\begin{array}{l}
p_\theta(y_1 \mid x) \\
p_\theta(y_2 \mid x) \\
p_\theta(y_3 \mid x) \\
p_\theta(y_4 \mid x)
\end{array}
\quad
\begin{bmatrix} 0.62 \\ 0.36 \\ 0.01 \\ 0.01 \end{bmatrix}
\quad
\begin{bmatrix} 0.16 \\ 0.10 \\ 0.14 \\ 0.60 \end{bmatrix}
\quad
\begin{bmatrix} 0.10 \\ 0.70 \\ 0.09 \\ 0.11 \end{bmatrix}
$$

$x_1$ $x_2$ $x_3$

**Donnée à annoter**

$$x^*_{LC} = \arg\max_x \{\, 1 - P_\theta(\hat{y} \mid x) \,\}$$

où $\hat{y} = \arg\max_y \{P_\theta(y \mid x)\}$ est la classe la plus probable sous le modèle

# Échantillonnage « Least Confident »

$$
\begin{array}{l}
p_\theta(y_1 \mid x) \\
p_\theta(y_2 \mid x) \\
p_\theta(y_3 \mid x) \\
p_\theta(y_4 \mid x)
\end{array}
\quad
\begin{bmatrix} 0.62 \\ 0.36 \\ 0.01 \\ 0.01 \end{bmatrix}
\quad
\begin{bmatrix} 0.16 \\ 0.10 \\ 0.14 \\ 0.60 \end{bmatrix}
\quad
\begin{bmatrix} 0.10 \\ 0.70 \\ 0.09 \\ 0.11 \end{bmatrix}
$$

$x_1$ $x_2$ $x_3$

Donnée à annoter

$$x^*_{LC} = \arg\max_x \{ 1 - P_\theta(\hat{y} \mid x) \}$$

- **Problème** : nous regardons seulement la distribution **d'un seul label**

# Échantillonnage avec marge

$$\begin{matrix} p_\theta(y_1 \mid x) \\ p_\theta(y_2 \mid x) \\ p_\theta(y_3 \mid x) \\ p_\theta(y_4 \mid x) \end{matrix} \begin{bmatrix} 0.62 \\ 0.36 \\ 0.01 \\ 0.01 \end{bmatrix} \begin{bmatrix} 0.16 \\ 0.10 \\ 0.14 \\ 0.60 \end{bmatrix} \begin{bmatrix} 0.10 \\ 0.70 \\ 0.09 \\ 0.11 \end{bmatrix}$$

$x_1$ $x_2$ $x_3$

Donnée à annoter

$$x_M^* = \arg\max_x \{ P_\theta(\hat{y}_1 \mid x) - P_\theta(\hat{y}_2 \mid x) \}$$

où $\hat{y}_i$ est la classe la i-ème plus probable sous le modèle

## Échantillonnage avec marge

$$\begin{matrix} p_\theta(y_1 \mid x) \\ p_\theta(y_2 \mid x) \\ p_\theta(y_3 \mid x) \\ p_\theta(y_4 \mid x) \end{matrix} \begin{bmatrix} 0.62 \\ 0.36 \\ 0.01 \\ 0.01 \end{bmatrix} \begin{bmatrix} 0.16 \\ 0.10 \\ 0.14 \\ 0.60 \end{bmatrix} \begin{bmatrix} 0.10 \\ 0.70 \\ 0.09 \\ 0.11 \end{bmatrix}$$

$x_1$ $x_2$ $x_3$

Donnée à annoter

$$x_M^* = \arg\max_x \{ P_\theta(\hat{y}_1 \mid x) - P_\theta(\hat{y}_2 \mid x) \}$$

- **Problème** : nous continuons à ignorer la distribution des sorties pour les classes restantes

## Échantillonnage par entropie

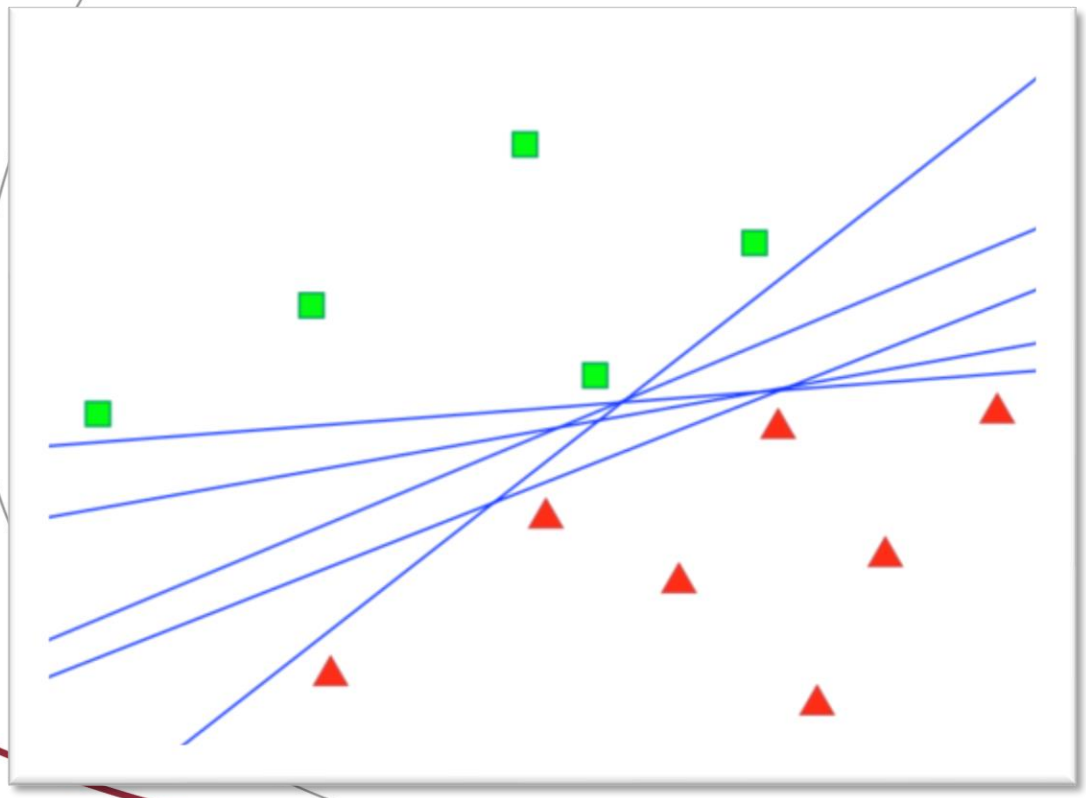
$$\begin{matrix} p_\theta(y_1 \mid x) \\ p_\theta(y_2 \mid x) \\ p_\theta(y_3 \mid x) \\ p_\theta(y_4 \mid x) \end{matrix} \begin{bmatrix} 0.62 \\ 0.36 \\ 0.01 \\ 0.01 \end{bmatrix} \begin{bmatrix} 0.16 \\ 0.10 \\ 0.14 \\ 0.60 \end{bmatrix} \begin{bmatrix} 0.10 \\ 0.70 \\ 0.09 \\ 0.11 \end{bmatrix}$$

$x_1 \qquad x_2 \qquad x_3$

- Stratégie d'échantillonnage incertain **plus générale** (Shannon, 1948) utilise **l'entropie** comme mesure d'incertitude

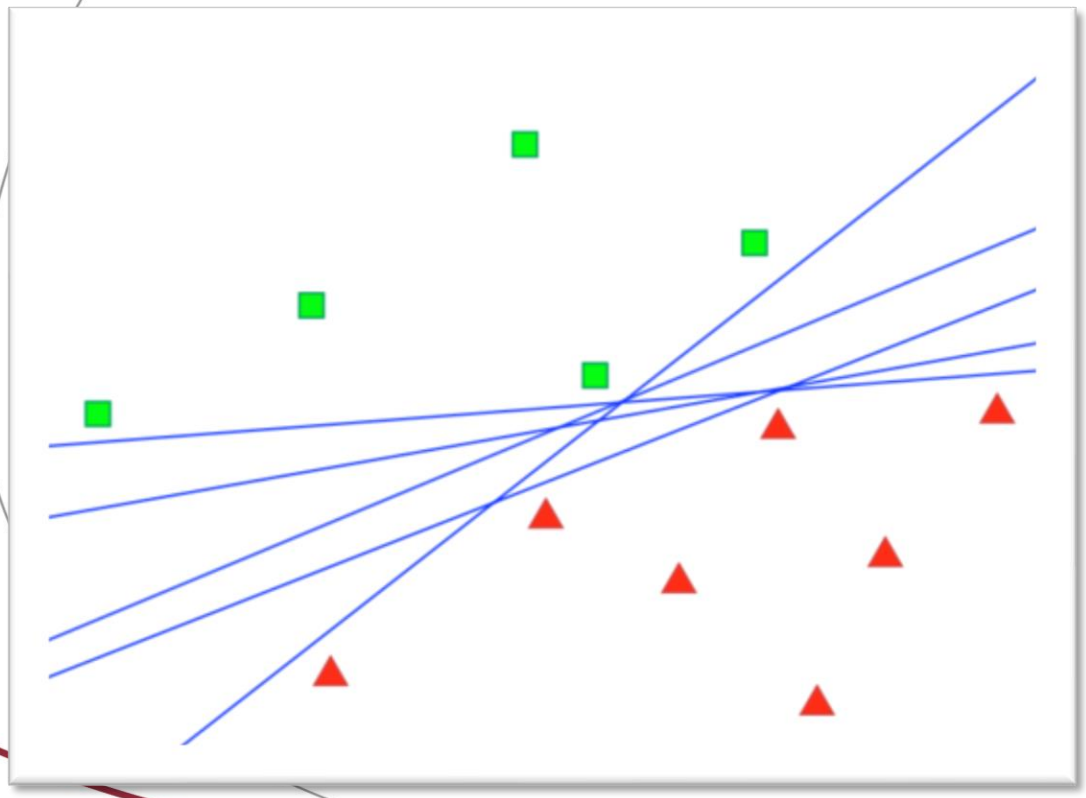
$$x_H^* = \arg\max_x \left\{ -\sum_i P_\theta(y_i \mid x) \log P_\theta(y_i \mid x) \right\}$$

# Requête par votes



- Sélectionner l'instance où les modèles sont **le plus en désaccord**

# Requête par votes



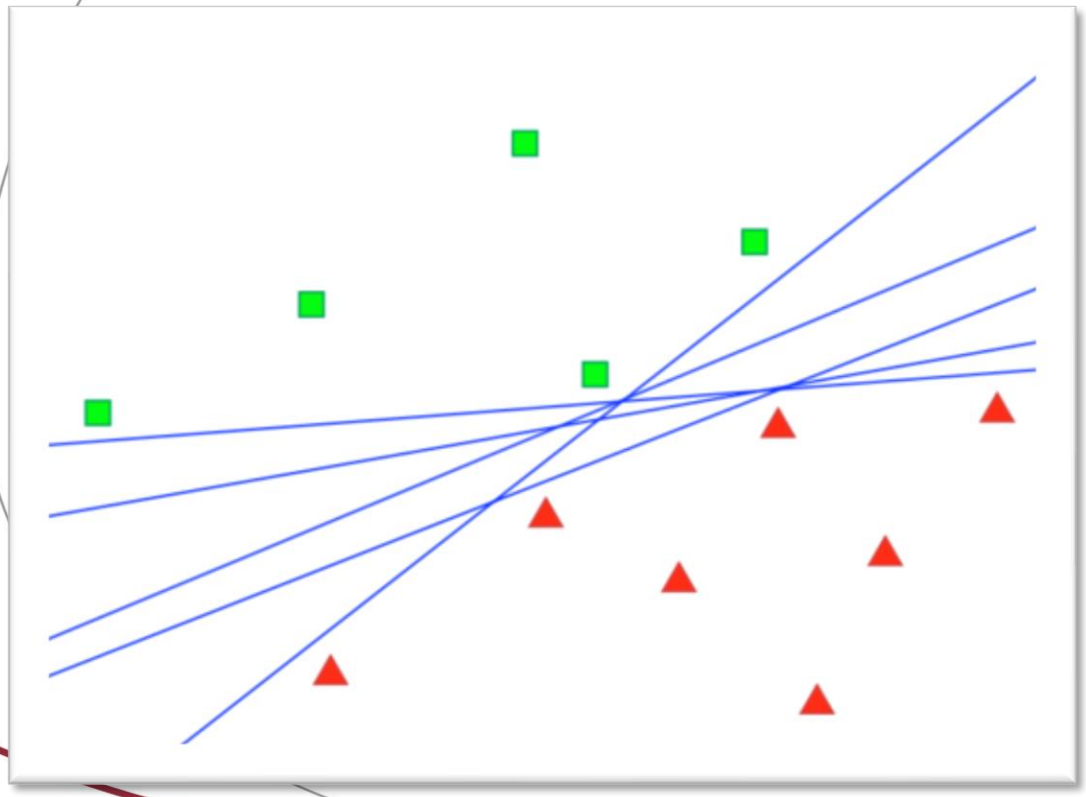
- Sélectionner l'instance où les modèles sont **le plus en désaccord**

**Questions :**

**1. Quel comité de modèles choisir ?**

**2. Quelle mesure de désaccord choisir ?**

# Requête par votes



- Sélectionner l'instance où les modèles sont **le plus en désaccord**

**Questions :**

1. **Quel comité de modèles choisir ?**
2. **Quelle mesure de désaccord choisir ?**

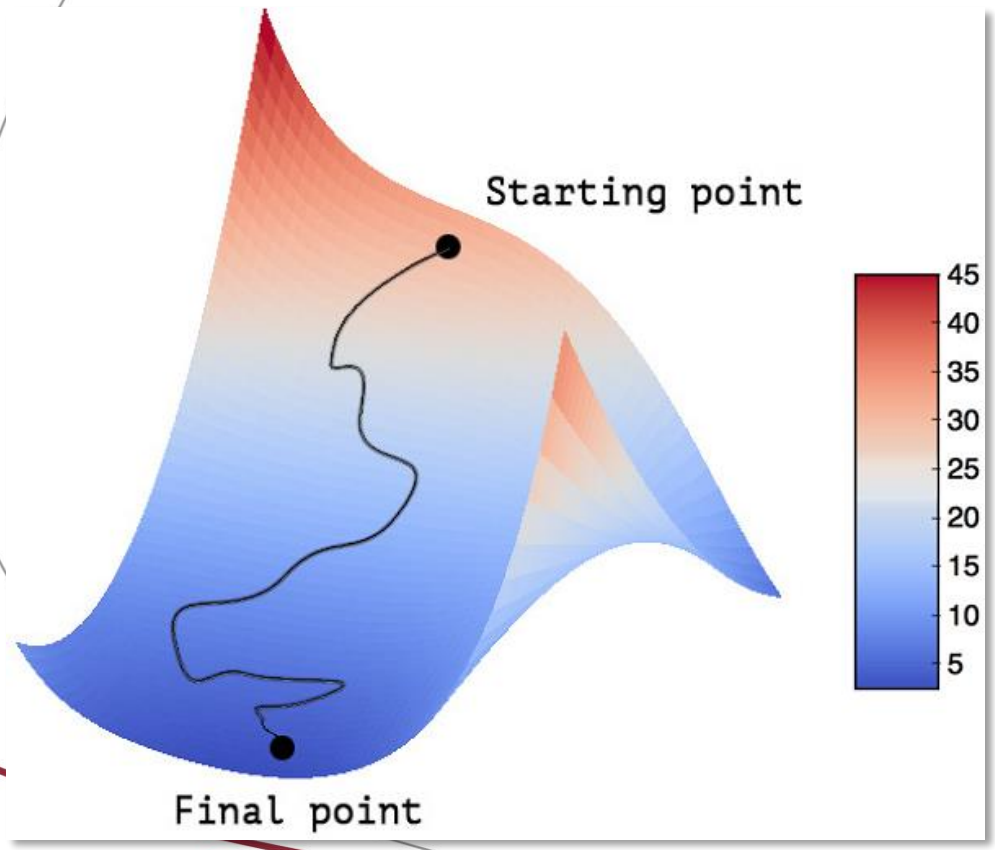
**Solutions possibles :**

1. **Une méthode par bagging**
2. $$x^*_{VE} = \arg\max_{x \in \mathcal{U}} \left\{ -\sum_{i=1}^{|\mathcal{C}|} \frac{V(y_i)}{N} \log \frac{V(y_i)}{N} \right\}$$

Nombre de vote pour la classe

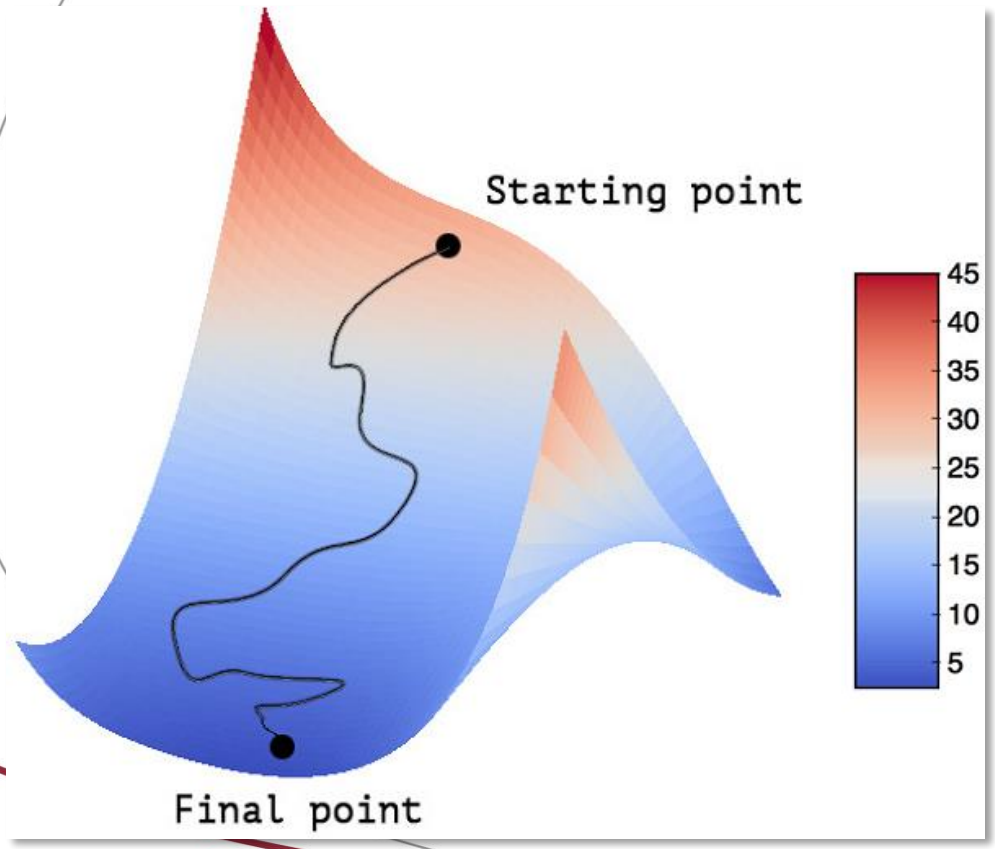
Taille de la comité

# Changement de modèle prévu




- Sélectionner l'instance qui, si nous connaissons son label, présente le **plus grand impact sur les paramètres** de notre modèle

# Changement de modèle prévu




- Sélectionner l'instance qui, si nous connaissons son label, présente le **plus grand impact sur les paramètres** de notre modèle

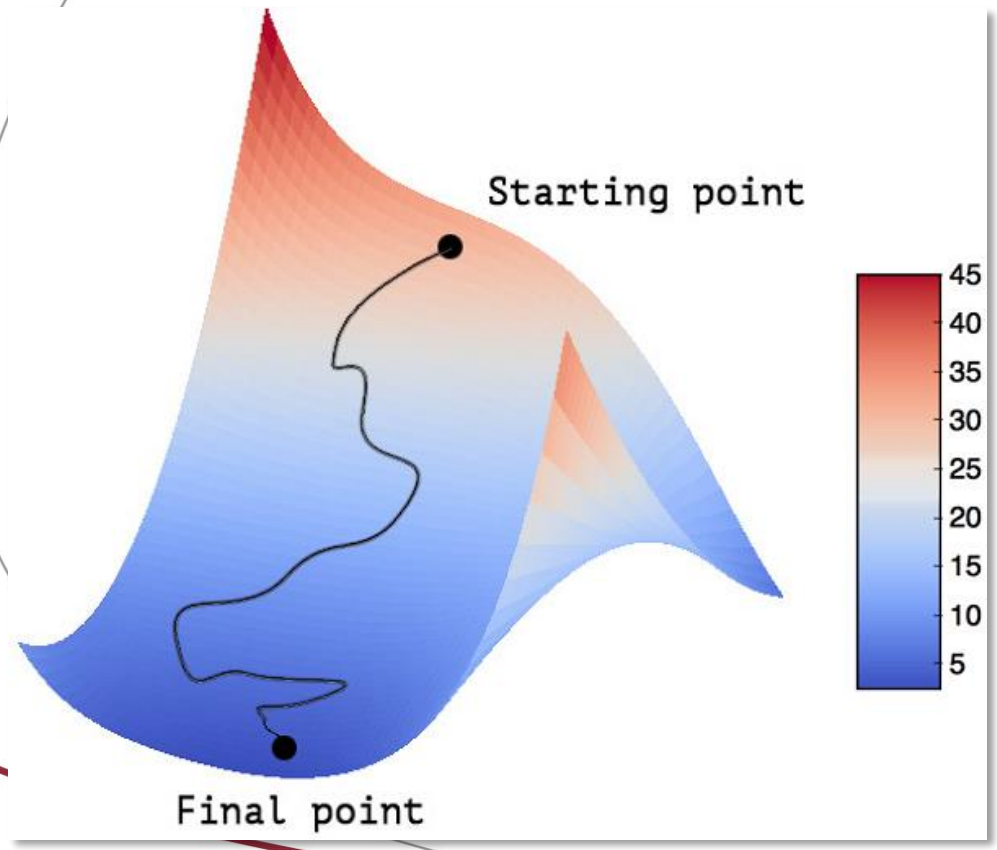
**Exemple de stratégie :**

- **Longueur prévue du gradient** :

$$x^*_{EGL} = \arg\max_{x \in \mathcal{U}} \left\{ \sum_{i=1}^{|C|} P_\theta(y_i \mid x) \cdot ||\nabla l_\theta (\mathcal{L} \cup < x, y_i >)||_2 \right\}$$

Fonction objectif

# Changement de modèle prévu




- Sélectionner l'instance qui, si nous connaissons son label, présente le **plus grand impact sur les paramètres** de notre modèle

**Exemple de stratégie :**

- **Longueur prévue du gradient** :

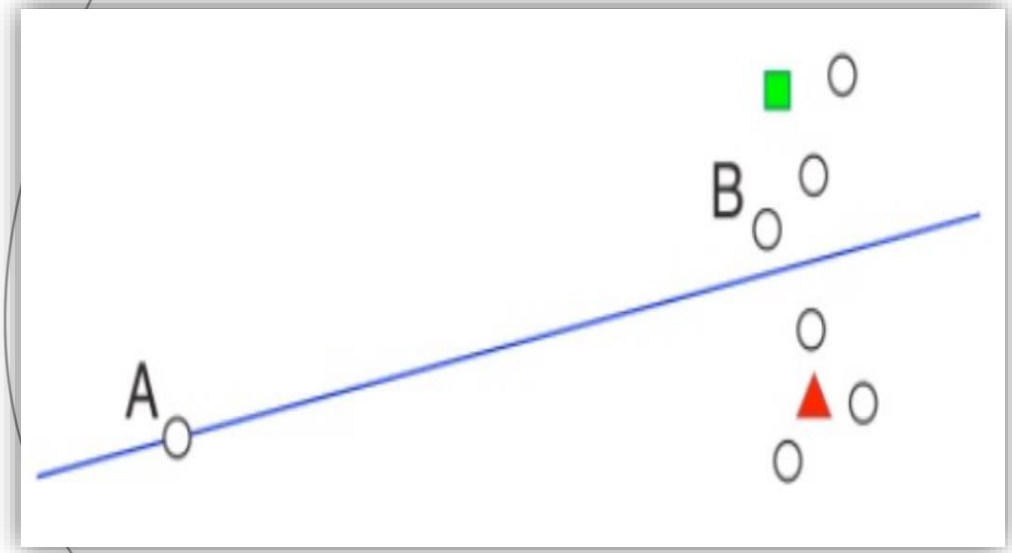
$$x^*_{EGL} = \arg\max_{x \in \mathcal{U}} \left\{ \sum_{i=1}^{|C|} P_\theta(y_i \mid x) \cdot ||\nabla l_\theta (\mathcal{L} \cup < x, y_i >)||_2 \right\}$$

Fonction objectif

- **Problème 1 : coûteux en calcul**

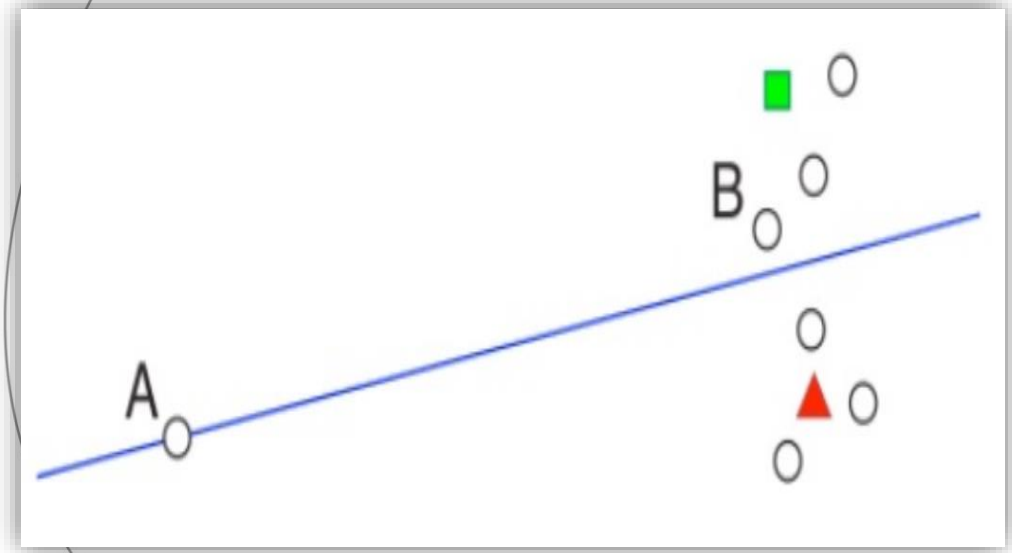
- **Problème 2 :** peu efficace si une des dimensions a une magnitude supérieure aux autres car le gradient peut **surestimer les légères variations** de cette dimension

# Méthode de densité pondérée



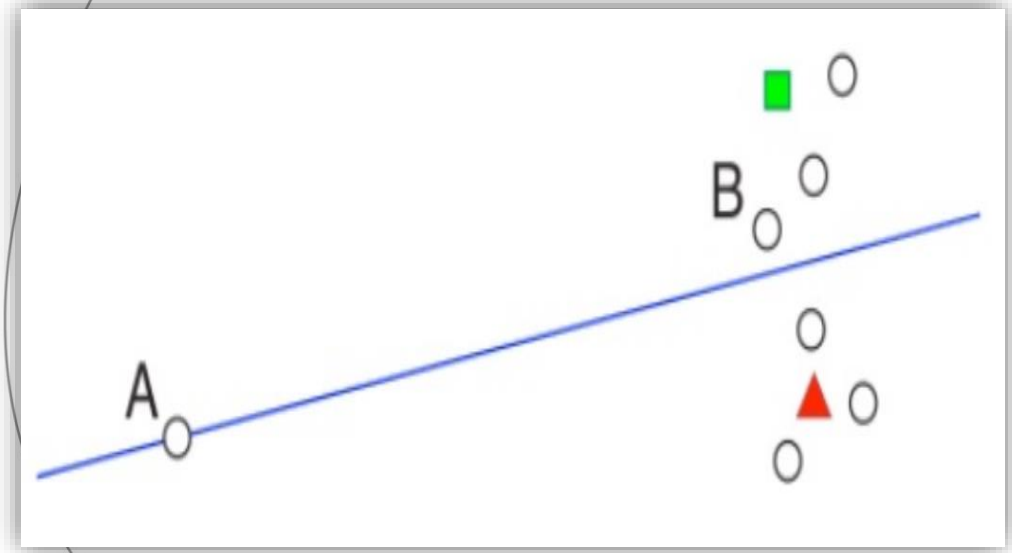

- Il est parfois intéressant de connaître le label d'une **donnée plus représentative** de la distribution sous-jacente

# Méthode de densité pondérée




- Il est parfois intéressant de connaître le label d'une **donnée plus représentative** de la distribution sous-jacente
- Sélectionner l'instance qui est :
  - À la fois **incertaine** pour le modèle (cf. échantillonnage incertain p.ex.)
  - À la fois **proche d'une région dense**

# Méthode de densité pondérée




- Il est parfois intéressant de connaître le label d'une **donnée plus représentative** de la distribution sous-jacente

- Sélectionner l'instance qui est :
  - À la fois **incertaine** pour le modèle (cf. échantillonnage incertain p.ex.)
  - À la fois **proche d'une région dense**

**Exemple de stratégie (Settles et al, 2008) :**

$$x^*_{ID} = \arg\max_{x \in \mathcal{U}} \left\{ \phi_A(x) \times \left( \frac{1}{|\mathcal{U}|} \sum_{u=1}^{|\mathcal{U}|} sim(x, x^{(u)}) \right)^{\beta} \right\}$$

Quantité d'information de x en fonction d'une stratégie A

Fonction de similarité

1. Introduction

2. Apprentissage automatique sur les données textuelles

3. Apprentissage actif

4. Différentes stratégies d'apprentissage actif

5. Expérimentations

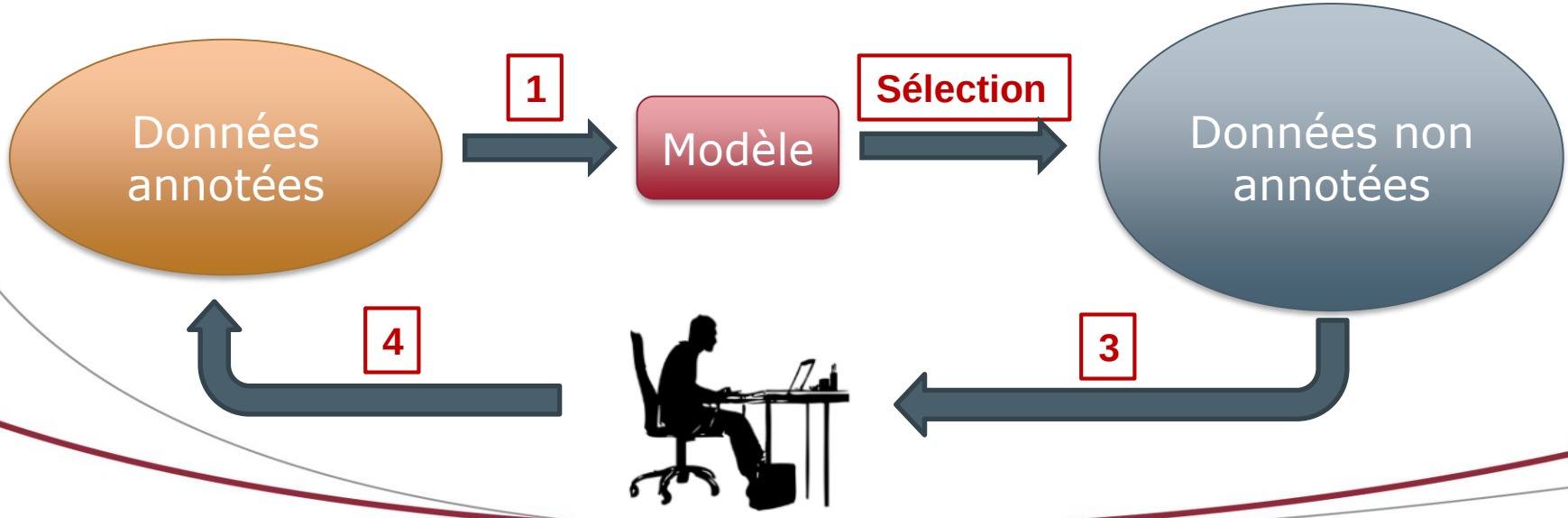
6. Conclusion

7. Prochaines étapes

# Processus d'apprentissage actif par mini-batch

**Algorithme de sélection :** tant que nous n'avons pas atteint une condition d'arrêt

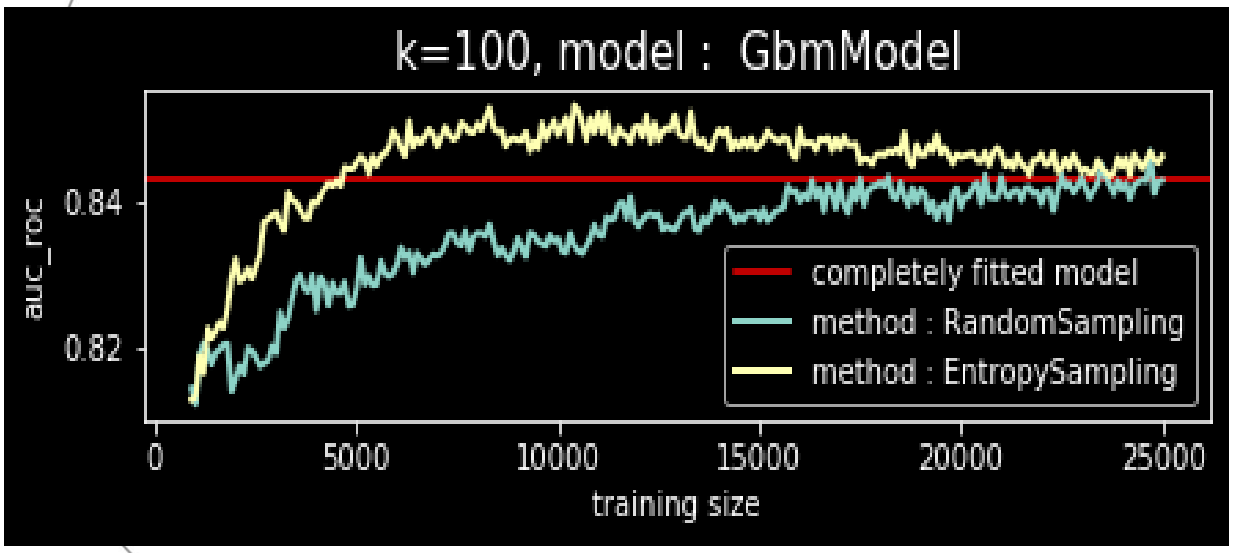
1. Entraîner notre modèle $\theta$ sur le **train-set** $L$ et évaluer sa performance sur le **test-set** $T$
2. Sélectionner les k échantillons $\{x_1^*, x_2^*, \dots, x_k^*\}$ les plus informatifs du **pool-set** $U$
3. Mettre à jour $U \leftarrow U - \{x_1^*, x_2^*, \dots, x_k^*\}$
4. Mettre à jour $L \leftarrow L \cup \{x_1^*, x_2^*, \dots, x_k^*\}$



# Échantillonnage aléatoire

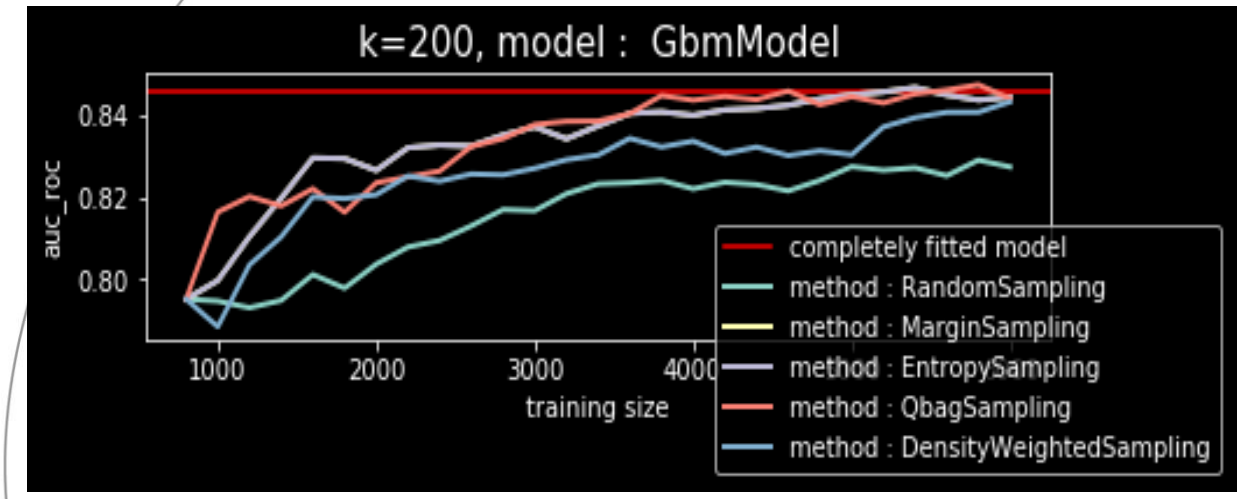

- Modèle(s) :
  - XGBoost
- Stratégie(s) : Échantillonnage
  - Aléatoire
- Taille des données annotées initiales (L) / d'un mini-batch (k) :
  - 200 / 100
- Condition d'arrêt :
  - 25 000

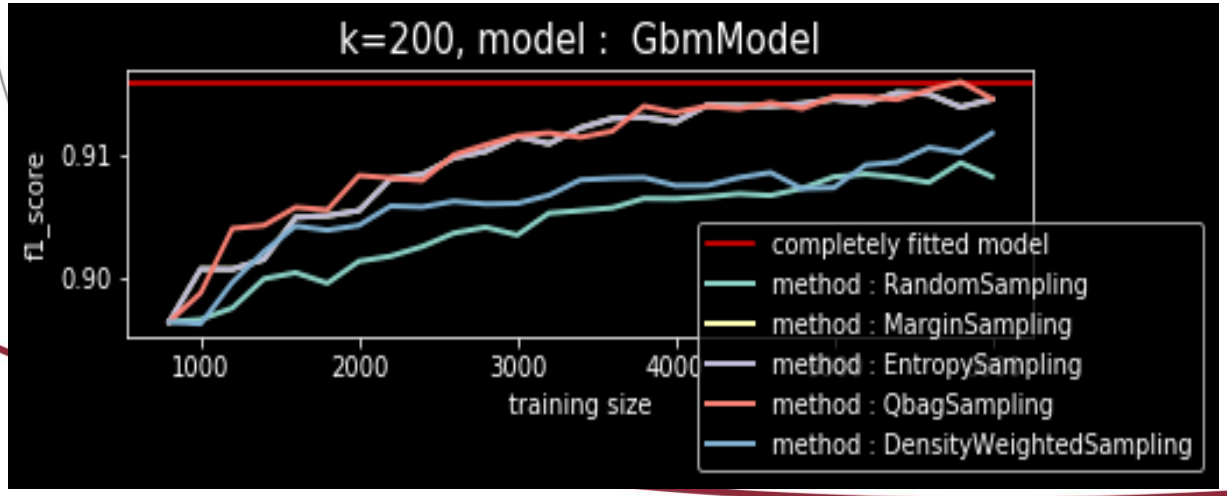
# Échantillonnage aléatoire VS par entropie




- Modèle(s) :
  - XGBoost
- Stratégie(s) : Échantillonnage
  - Aléatoire
  - par entropie
- Taille des données annotées initiales (L) / d'un mini-batch (k) :
  - 800 / 100
- Condition d'arrêt :
  - 25 000

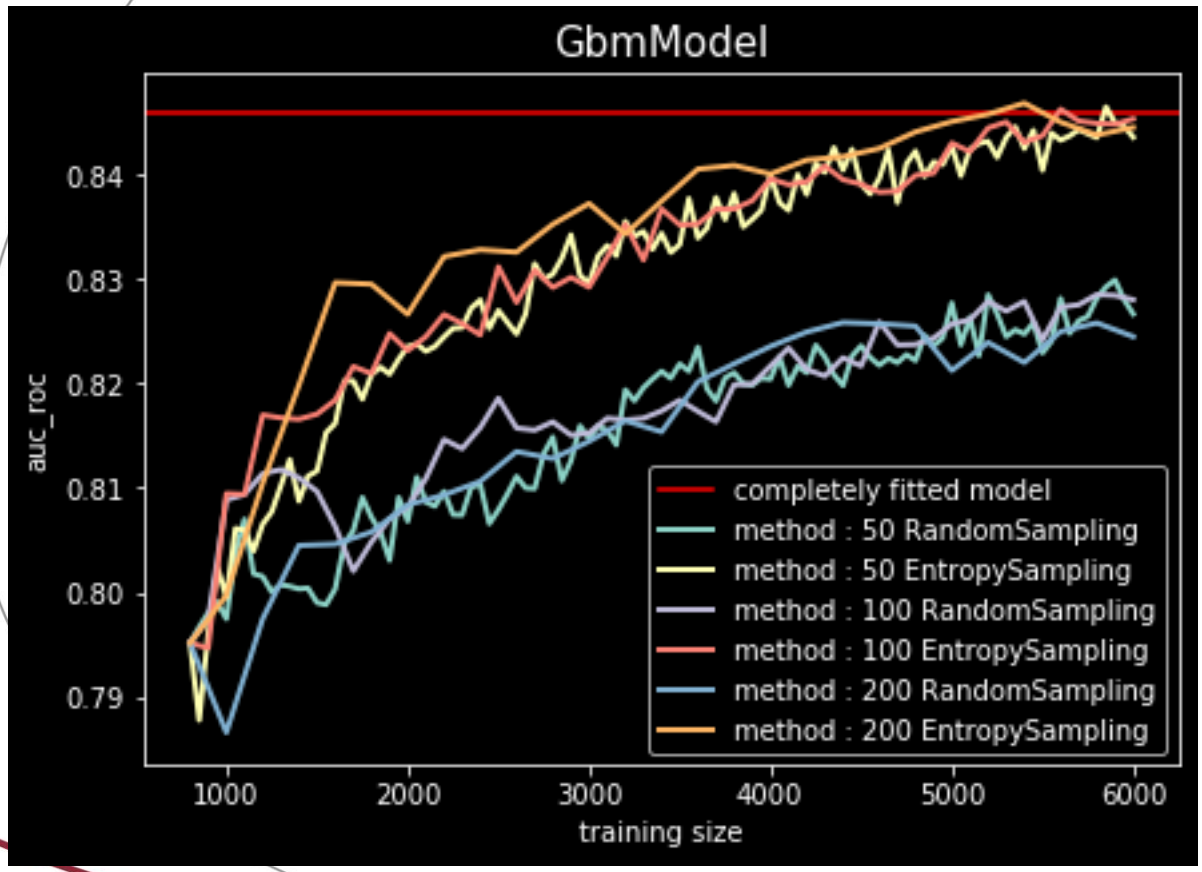
# Différentes stratégies d'apprentissage actif



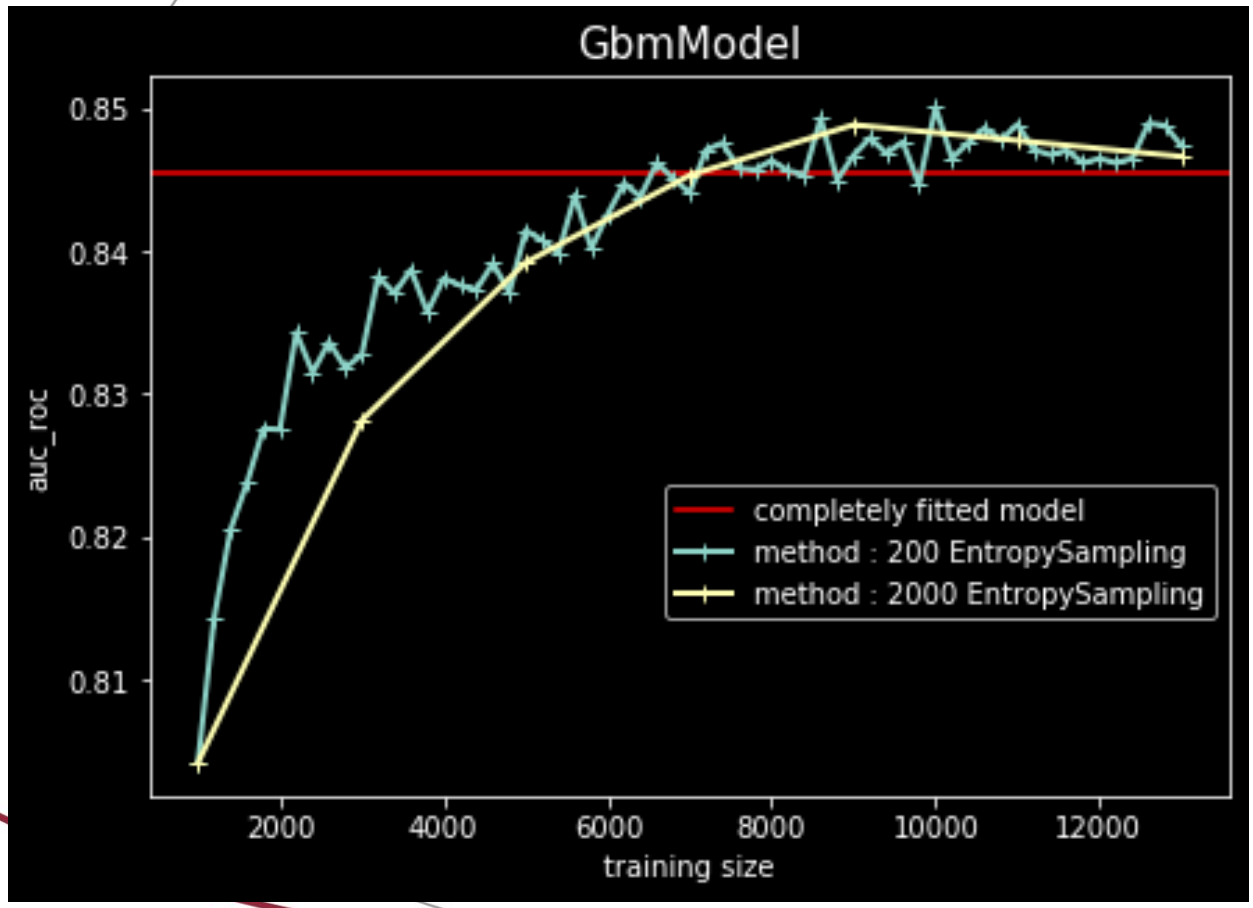

- Modèle(s) :
  - XGBoost
- Stratégie(s) : Échantillonnage
  - Aléatoire
  - Par entropie
  - Avec mage
  - Par requête de votes
  - Par densité pondérée avec l'échantillonnage par entropie
- Taille des données annotées initiales (L) / d'un mini-batch (k) :
  - 800 / 200
- Condition d'arrêt :
  - 6 000

# Échantillonnage par entropie + différentes tailles de mini-batch




- Modèle(s) :
  - XGBoost
- Stratégie(s) : Échantillonnage
  - Aléatoire
  - Par entropie
- Taille des données annotées initiales (L) / d'un mini-batch (k) :
  - 800 / **(50, 100, 200)**
- Condition d'arrêt :
  - 6 000

# Échantillonnage par entropie + différentes tailles de mini-batch




- Modèle(s) :
  - XGBoost
- Stratégie(s) : Échantillonnage
  - par entropie
- Taille des données annotées initiales (L) / d'un mini-batch (k) :
  - 1 000 / **(200, 2 000)**
- Condition d'arrêt :
  - 13 000

1. Introduction

2. Apprentissage automatique sur les données textuelles

3. Apprentissage actif

4. Différentes stratégies d'apprentissage actif

5. Expérimentations

6. Conclusion

7. Prochaines étapes

# Conclusion

- Possibilité de **catégoriser automatiquement les données textuelles** si nous avons la possibilité d'avoir une base de données annotées

- Possibilité de **réduire considérablement le coût d'annotation** avec l'apprentissage actif. Cela est vraie pour :
  - Des données suivant la loi gaussienne
  - Des données réelles assurantielles (type NPS)

- Section suivante : des **améliorations possibles** pour ces stratégies d'apprentissage actif

1. Introduction

2. Apprentissage automatique sur les données textuelles

3. Apprentissage actif

4. Différentes stratégies d'apprentissage actif

5. Expérimentations

6. Conclusion

7. Prochaines étapes

# Prochaines étapes

## Quelques problèmes à soulever :

- **Des approches Deep Learning :**
  - Quel est l'apport des modèles séquentiels (RNN, LSTM, …) dans la performance de notre modèle ?

- **Générer artificiellement des données textuelles :**
  - Quel est l'apport des méthodes génératives (GAN par exemple) dans la performance de notre modèle ?

- **Cas des bases très déséquilibrées :**
  - A l'initialisation, comment peut-on détecter rapidement les catégories très rares (<< 1%) ?
  - Est-ce que les méthodes d'apprentissage actif présentées ci-dessus marchent toujours ?

- **Accélération du temps de traitement :**
  - Séquentiellement, au lieu de faire de l'apprentissage « hors-ligne », peut-on faire de l'apprentissage « en ligne » ?