Big Data et Fouille de données

Massih-Reza Amini

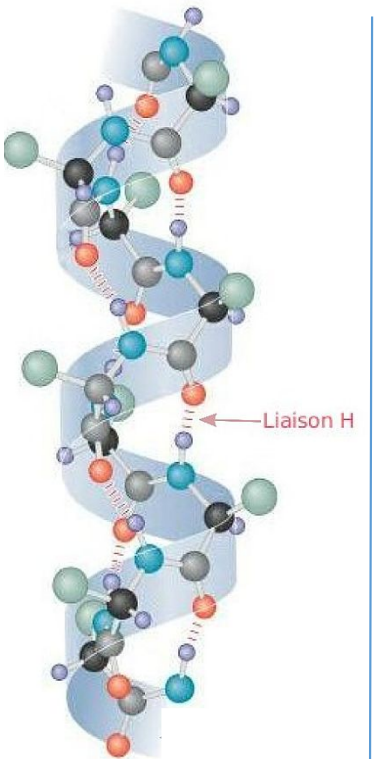
Université Grenoble Alpes

Big Data HPC - 26 avril 2016

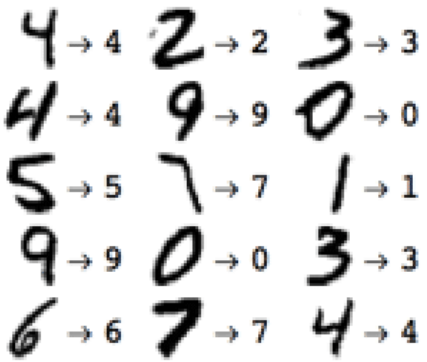
L I G

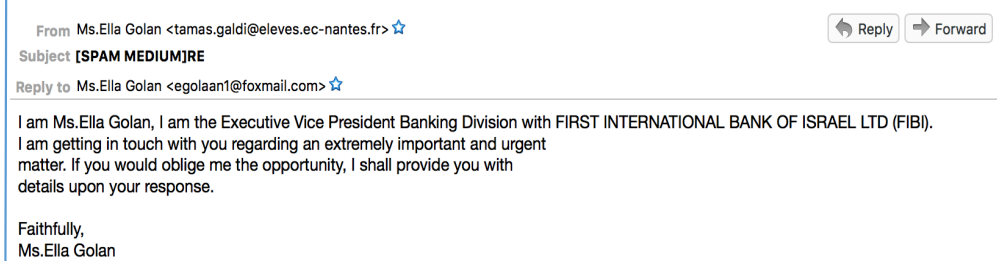
En apprentissage automatique

on cherche à trouver l'association entre les observations et leurs sorties désirées



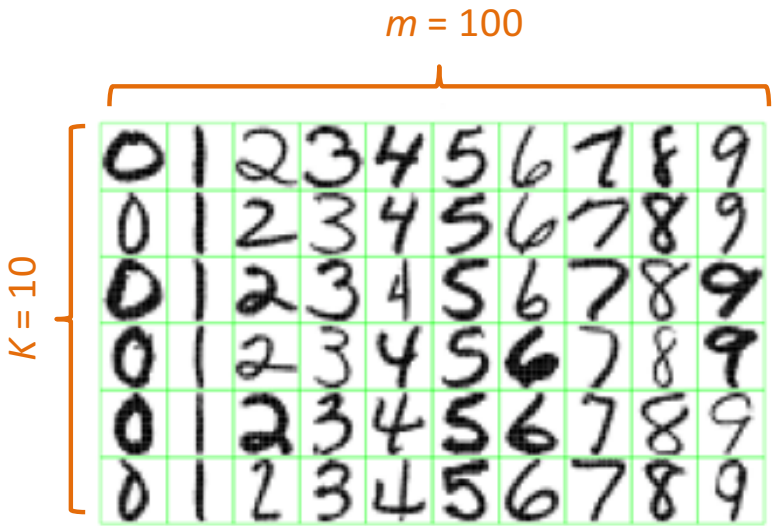
Hélice alpha





Spam

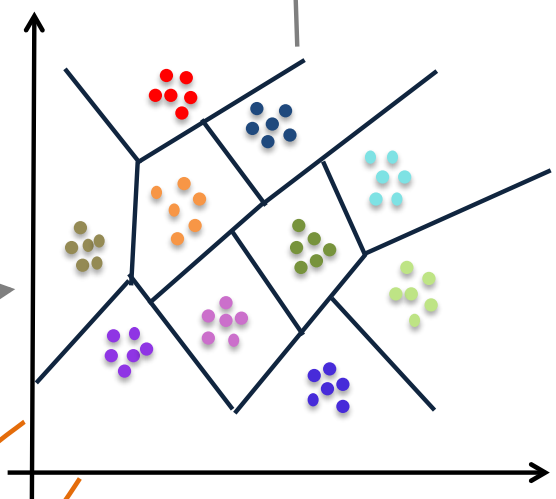
En apprentissage automatique



Base d'apprentissage

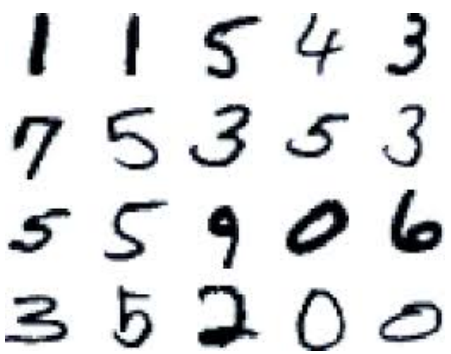
1. Représentation vectorielle

2. Trouver les séparateurs



4. Représentation vectorielle

3. Nouveaux exemples



d = 2

5. Prédire les classes de nouveaux exemples

Cadre classique de l'apprentissage automatique

Nous considérons un espace d'entrée $\mathcal{X} \subseteq \mathbb{R}^d$ et un espace de sortie $\mathcal{Y} = \{1, \ldots, K\}$.

Hypothèse:Les paires d'exemples $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ sont *identiquement* et *indépendamment* distribuées (`i.i.d`) d'après une distribution de probabilité fixe $\mathcal{D}$.

Échantillon:Nous observons une séquence de m paires d'exemples $(\mathbf{x}_i, y_i)$ générées `i.i.d` suivant $\mathcal{D}$.

Objectif: Construire une fonctions de prédiction $f : \mathcal{X} \to \mathcal{Y}$ qui prédit la sortie y d'un nouvel exemple $\mathbf{x}$ avec la plus petite probabilité d'erreur.

Apprentissage supervisé

- Les modèles discriminants trouve directement la fonction de classification $f : \mathcal{X} \to \mathcal{Y}$ à partir d'une classe de fonctions $\mathcal{F}$;
- La fonction trouvée devrait être celle qui minimise la probabilité d'erreur

$$R(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}} L(x,y) = \int_{\mathcal{X}\times\mathcal{Y}} L(f(x),y) d\mathcal{D}(x,y)$$

Où L est la fonction de risque définie comme

$$L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$$

La fonction de risque usuellement considérée est l'erreur de classification:

$$\forall (x,y); L(f(x),y) = \mathbb{1}_{f(x)\neq y}$$

Où $\mathbb{1}_\pi$ est la fonction indicatrice.

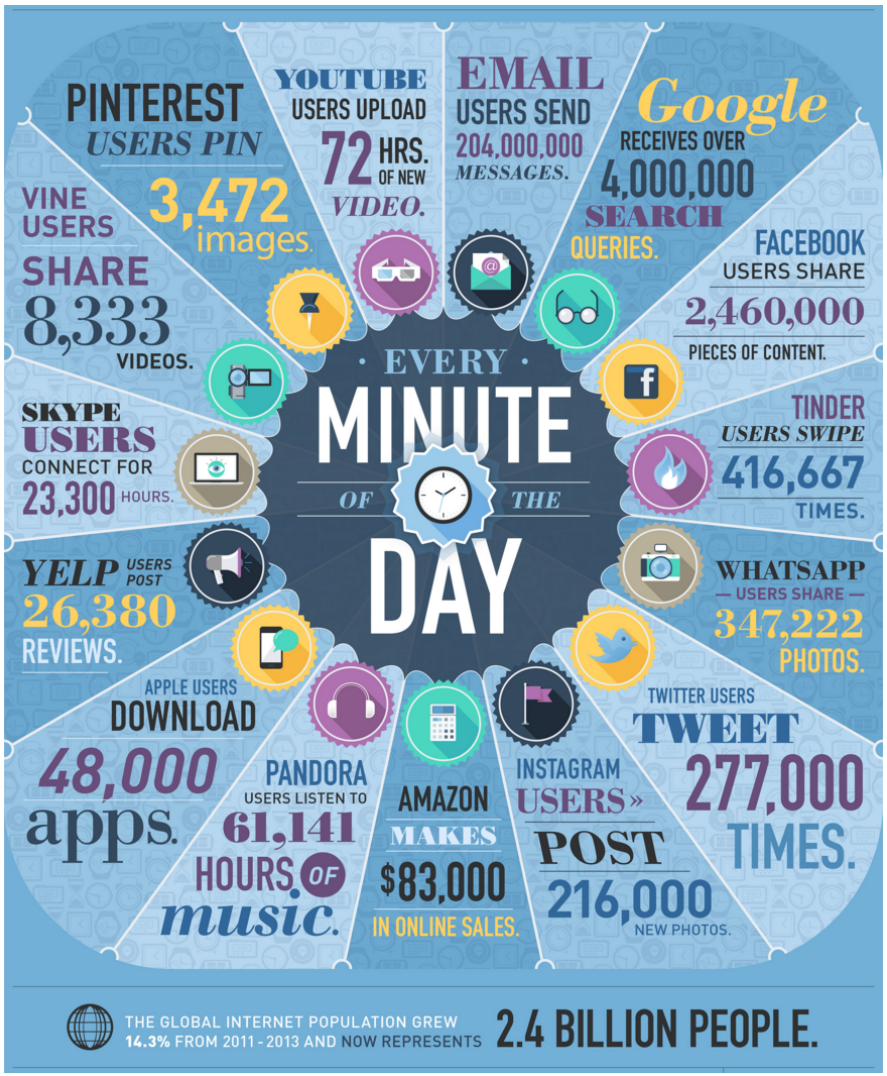
Principe de la minimisation du risque empirique

- Comme la distribution de probabilité $\mathcal{D}$ est inconnue, la forme analytique du risque ne peut pas être calculée, la fonction de prédiction ne peut ainsi être calculée directement de $R(f)$.
- Principe de la minimisation du risque empirique: Trouve f en minimisant l'estimateur non-biaisé de R sur une base d'apprentissage $S = (x_i, y_i)_{i=1}^m$:

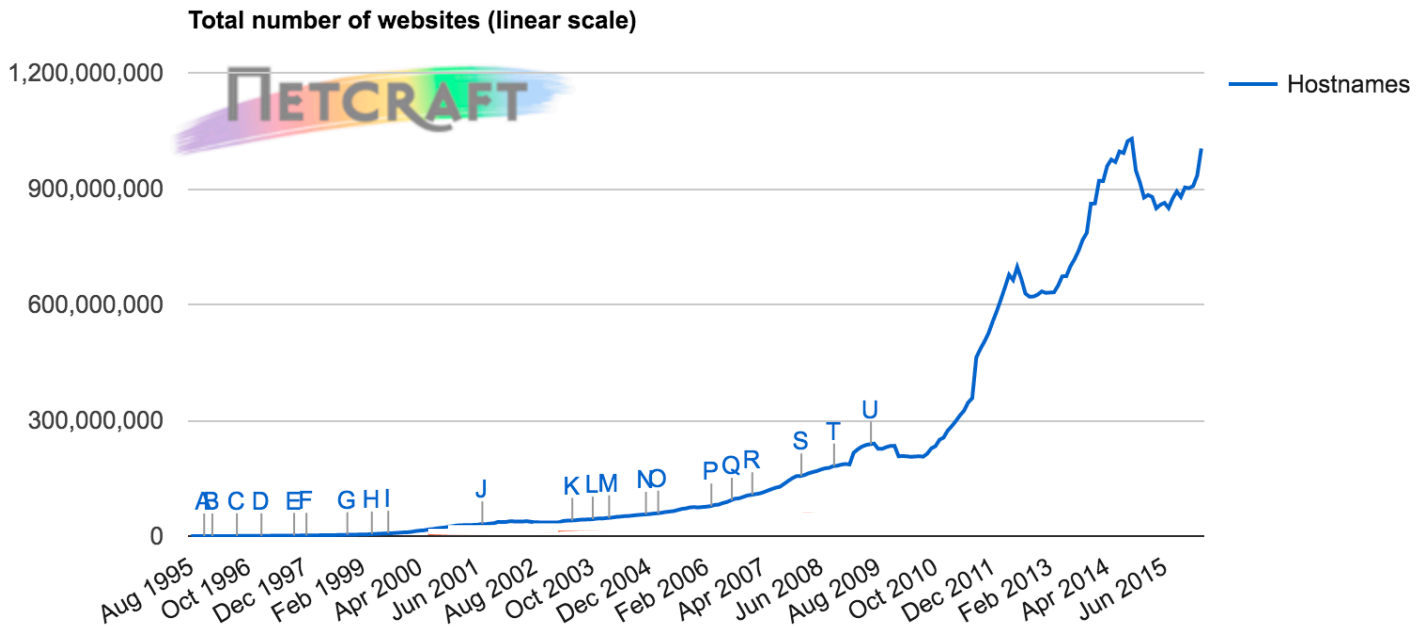
$$\hat{R}_m(f, S) = \frac{1}{m} \sum_{i=1}^{m} L(f(x_i), y_i)$$

- Au meilleur des cas, la complexité de la minimisation de $\hat{R}_m(f, S)$ est $O(d \times m)$, des techniques à base de gradient stochastique ont été proposée pour accelerer l'apprentissage.

Ère de Big Data dans le domaine de la RI



Ère de Big Data dans le domaine de la RI



- D'après les prévisions du projet EMC[1], 90% des données sur le web sont des données textuelles, et en 2020 il y aura 40 zetta octets (40×10^{21} octets) de données non-structurées.

[1] http://www.emc.com/leadership/digital-universe/index.htm

Cadre classique de l'apprentissage machine

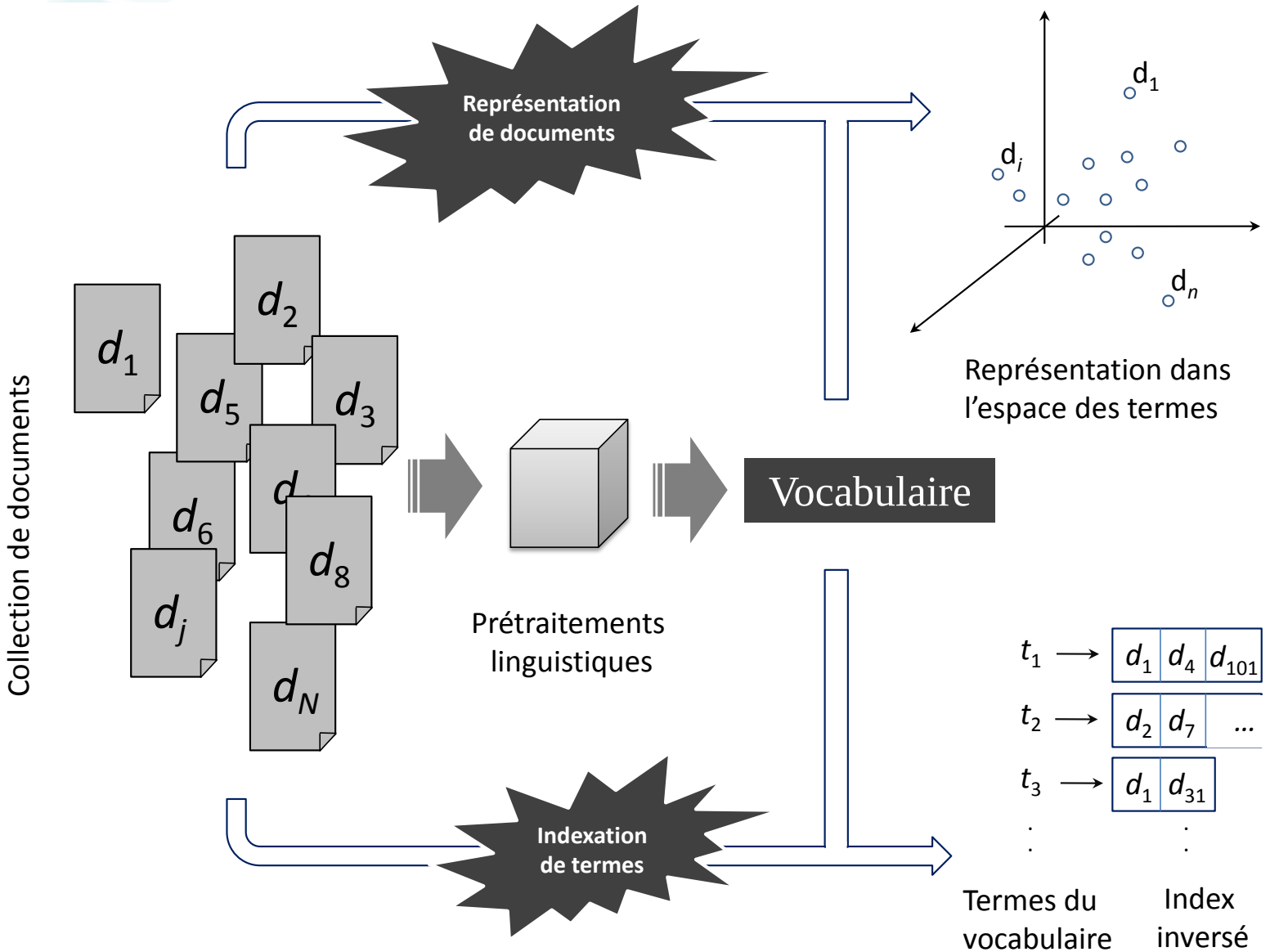
Nous considérons un espace d'entrée $\mathcal{X} \subseteq \mathbb{R}^d$ et un espace de sortie $\mathcal{Y} = \{1, \ldots, K\}$.

Hypothèse:Les paires d'exemples $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ sont *identiquement* et *indépendamment* distribuées (`i.i.d`) d'après une distribution de probabilité fixe $\mathcal{D}$.

Échantillon:Nous observons une séquence de m, $\boldsymbol{m >> 1,}$ paires d'exemples $(\mathbf{x}_i, y_i)$ générées `i.i.d` suivant $\mathcal{D}$.

Objectif: Construire une fonctions de prédiction $f : \mathcal{X} \to \mathcal{Y}$ qui prédit la sortie y d'un nouvel exemple $\mathbf{x}$ avec la plus petite probabilité d'erreur.

Indexation et représentation

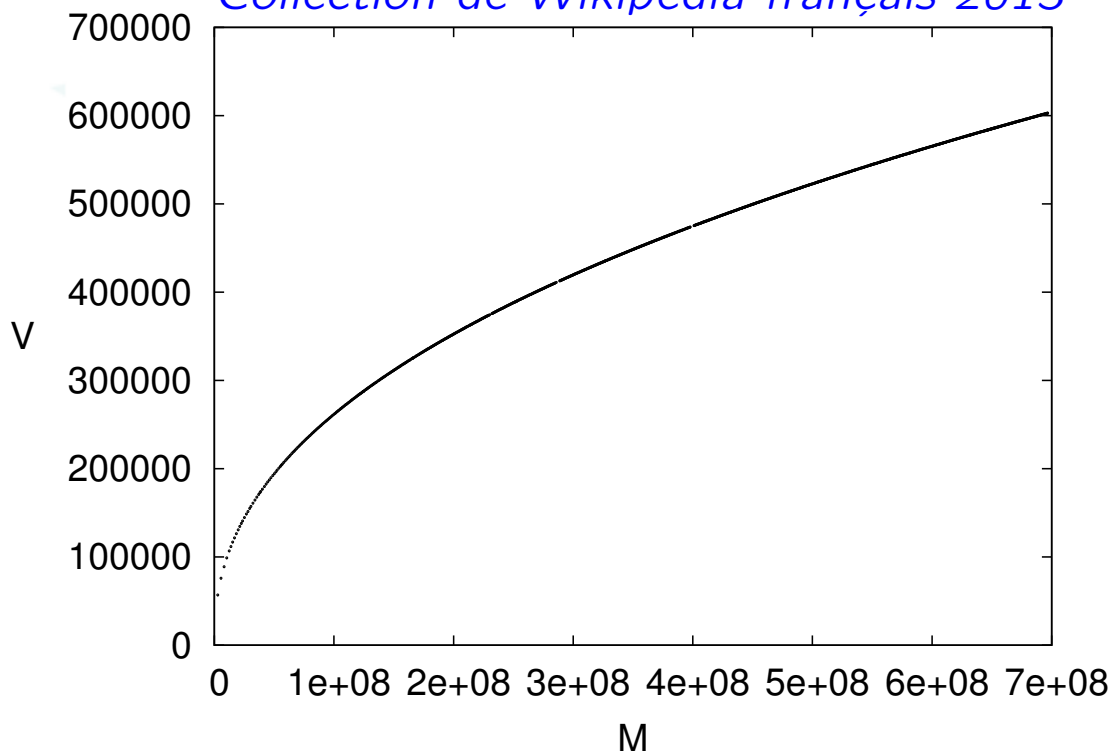


Loi de Heaps (1978)

La loi de Heaps stipule que la taille du vocabulaire (V) a une croissance sous-linéaire en fonction du nombre d'occurrence des mots (M) ou de la taille du corpus.

$$V = K \times M^{\beta}$$

Collection de Wikipédia français 2013



Cadre classique de l'apprentissage machine

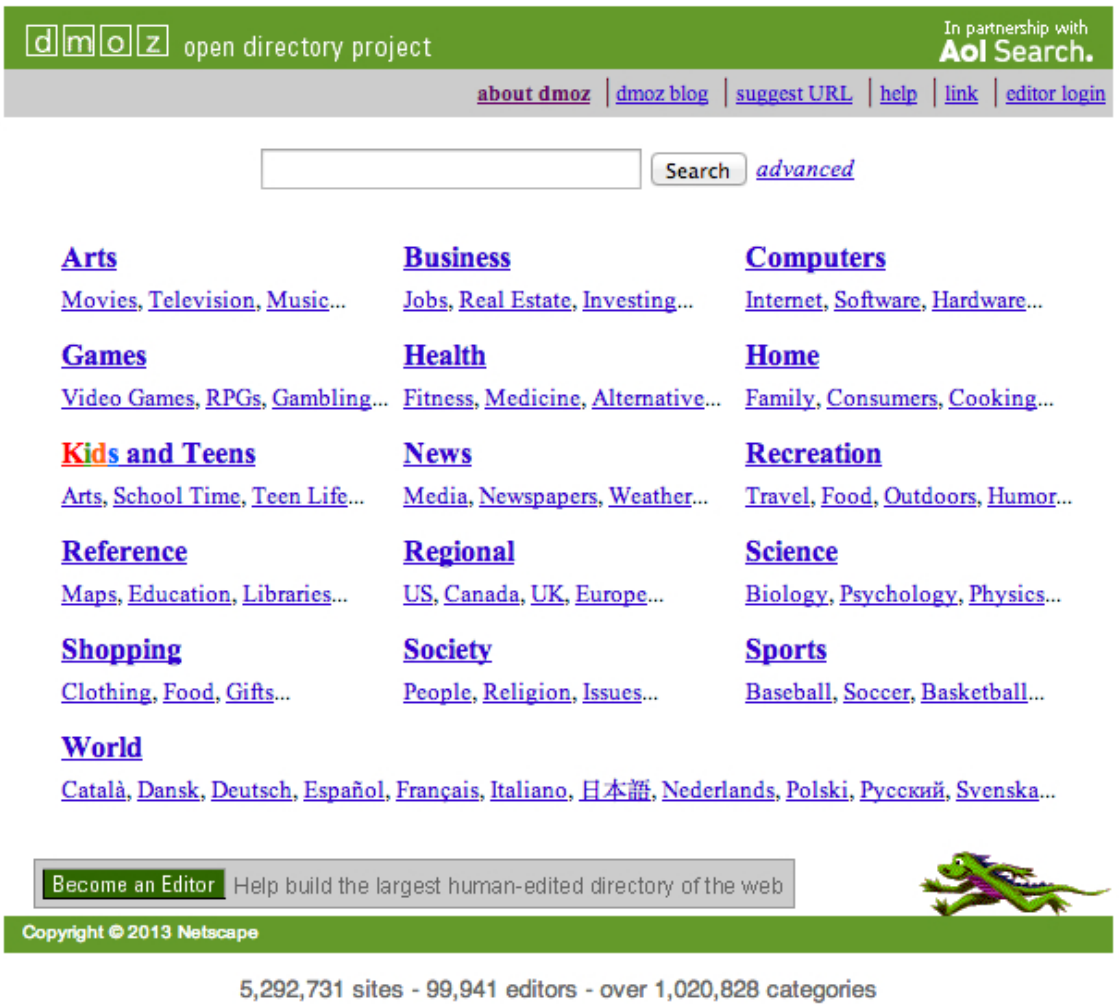
Nous considérons un espace d'entrée $\mathcal{X} \subseteq \mathbb{R}^d$, $\boldsymbol{d >> 1}$, et un espace de sortie $\mathcal{Y} = \{1, \ldots, K\}$.

Hypothèse:Les paires d'exemples $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ sont *identiquement* et *indépendamment* distribuées (`i.i.d`) d'après une distribution de probabilité fixe $\mathcal{D}$.

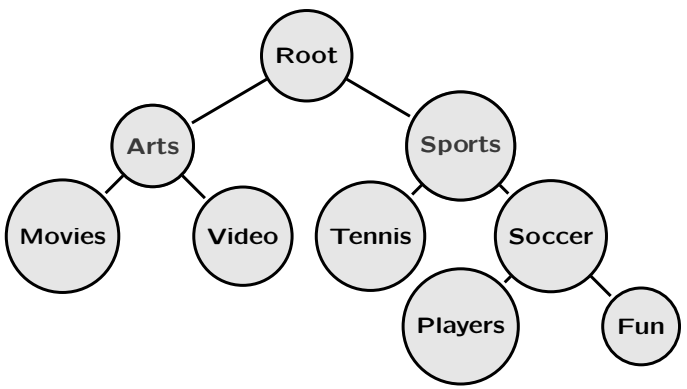
Échantillon:Nous observons une séquence de m, $\boldsymbol{m >> 1}$, paires d'exemples $(\mathbf{x}_i, y_i)$ générées `i.i.d` suivant $\mathcal{D}$.

Objectif: Construire une fonctions de prédiction $f : \mathcal{X} \to \mathcal{Y}$ qui prédit la sortie y d'un nouvel exemple $\mathbf{x}$ avec la plus petite probabilité d'erreur.

Classification à large échelle



- 5×10^6 sites
- 10^6 catégories
- 10^5 editeurs



Cadre classique de l'apprentissage machine

Nous considérons un espace d'entrée $\mathcal{X} \subseteq \mathbb{R}^d$, $\boldsymbol{d >> 1}$, et un espace de sortie $\mathcal{Y} = \{1, \ldots, K\}$, $\boldsymbol{K >> 1}$.

Hypothèse:Les paires d'exemples $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ sont *identiquement* et *indépendamment* distribuées (`i.i.d`) d'après une distribution de probabilité fixe $\mathcal{D}$.

Échantillon:Nous observons une séquence de m, $\boldsymbol{m >> 1}$, paires d'exemples $(\mathbf{x}_i, y_i)$ générées `i.i.d` suivant $\mathcal{D}$.

Objectif: Construire une fonctions de prédiction $f : \mathcal{X} \to \mathcal{Y}$ qui prédit la sortie y d'un nouvel exemple $\mathbf{x}$ avec la plus petite probabilité d'erreur.

Challenges

- Face à ces challenges de nouveaux paradigmes d'apprentissage sont en train de voir le jour.
- Les nouveaux algorithmes travaillent sur des bases distribuées à travers differentes machines et en prenant en compte les caractéristiques du réseau,
- Exemple: Collaboratif Filtering

$R =$ User

	Item			
User/ Item	A	B	C	D
Ted		4		3
Carol	3		2	
Bob		5		2
Alice			4	

Exemple: Filtrage Collaboratif

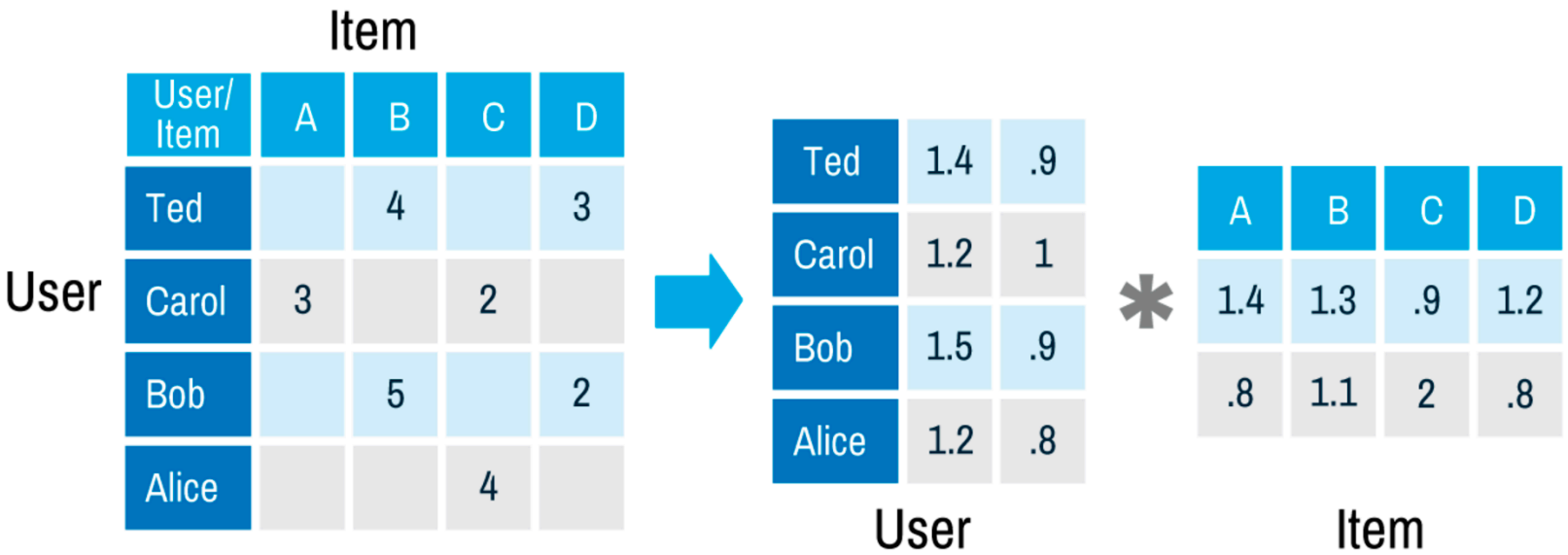
- Les corpus publics

Collection	$\lvert\mathcal{U}\rvert$	$\lvert\mathcal{I}\rvert$	sparsité
ML-100K	943	1682	93.7 %
ML-1M	6040	3952	95.8 %
ML-10M	71567	10681	98.7 %
NetFlix	480189	17770	98.8 %
Yahoo! Music	1000990	624961	99.6%

- Approche gagnante le prix NetFlix à base de la factorisation matricielle

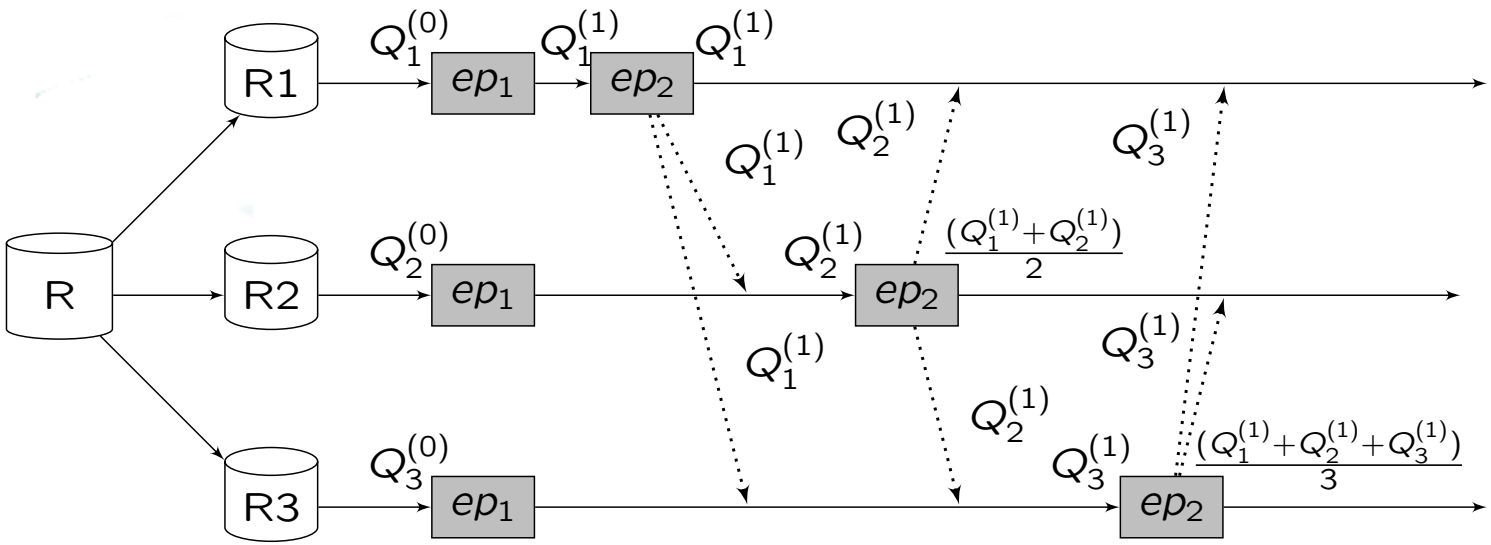
$$\min_{P,Q} \sum_{i,j:r_{ij}\text{existe}} \left(r_{ij} - q_j^\top p_i\right)^2 + \lambda(||p_i||^2 + ||q_j||^2)$$

Filtrage Collaboratif



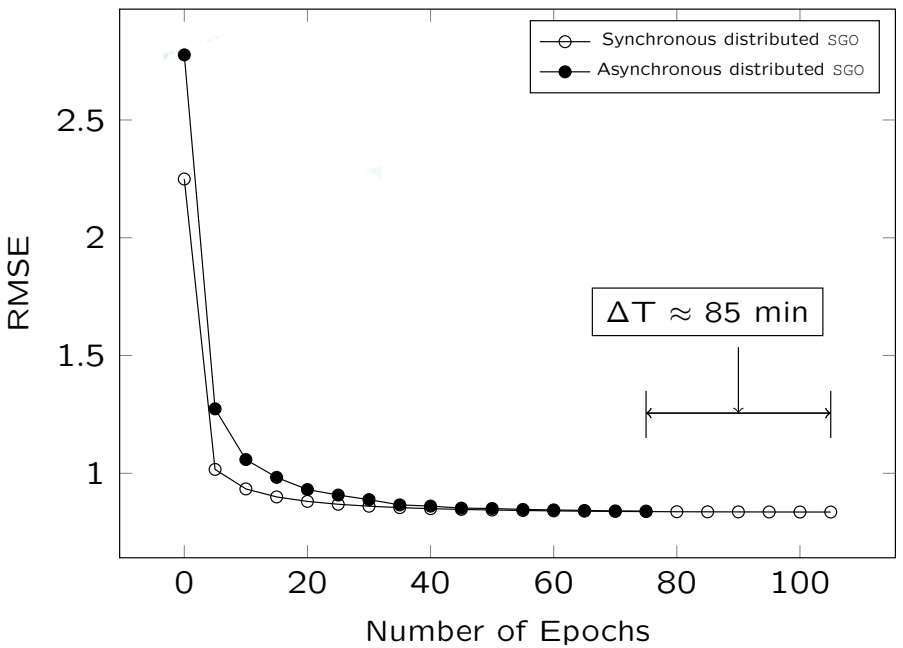
- Une façon de faire est de distribuer la matrice des scores R à travers les machines et de minimiser itérativement la fonction objectif avec la descente de gradient stochastique et puis moyenner les matrices P et Q une fois que les machines ont fait leurs mises à jour.

Factorisation Matricielle avec une approche asynchrone

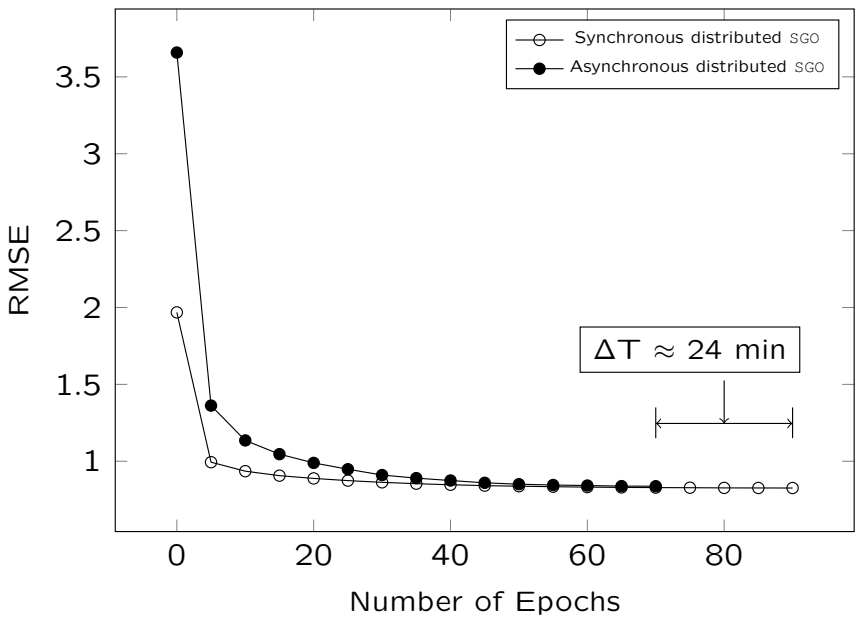


Factorisation Matricielle avec une approche asynchrone: temps de convegence

MovieLens 10M



NetFlix



Conclusion

- Le boom des données numériques se traduit par un accroissement fort des trois dimensions sous-jacents au cadre de l'apprentissage automatique
- Cela a amené de nouvelles problématiques en apprentissage automatique
- Les nouveaux algorithmes d'apprentissage sont de plus conçus en mode distribués et prennent en compte les caractéristiques du réseau pour atteindre de meilleures performances computationnelles.

Références

- M.-R. Amini. **Apprentissage Machine: de la théorie à la pratique**. *Éditions Eyrolles*, 2015.
- M.-R. Amini and É. Gaussier. **Recherche d'Information: applications, modèles et algorithmes**. *Éditions Eyrolles*, 2013.
- I.H. Witten, A. Moffat and T.C. Bell. **Managing Gigabytes**, *Morgn Kaufmann Publishing*, 1999.