Apprentissage

# Estimation du nombre d’utilisateurs d’un shell

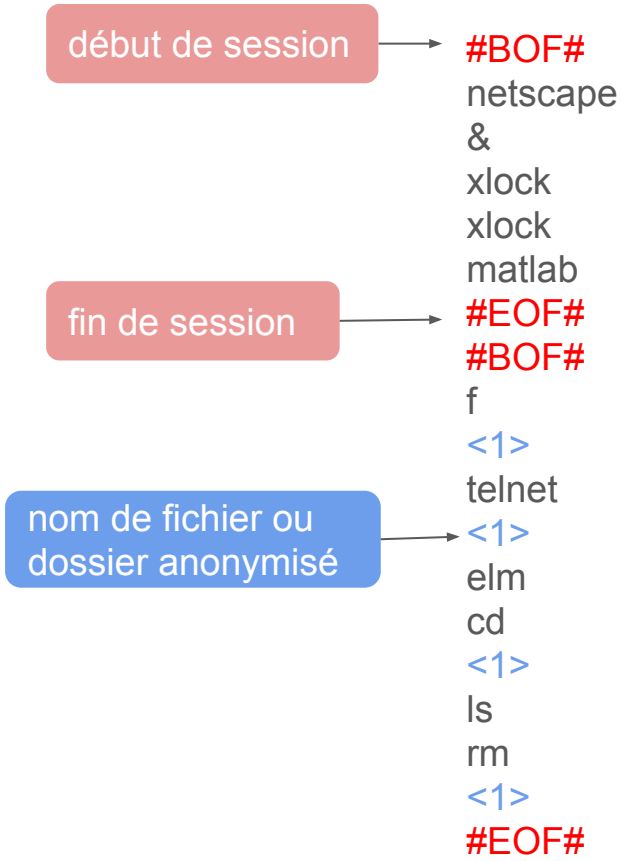
Julien LOUIS - Chloé MERCIER - Perceval WAJSBURT

# PROBLÈME

On dispose de l'historique d'un shell UNIX :

- historique des commandes sur plusieurs années
- sessions concaténées par ordre chronologique (pas de timestamp)
- débuts et fins de sessions indiqués

⇒ Combien d'utilisateurs différents ?



# Approche retenue, modèle génératif

## Modèle utilisateur

**Un profil de compétence par utilisateur (moyenne)**

Alan:
[80% maths et 20% réseau]

Mike:
[50% C++ et 50% ML ]

## Modèle session

**Une session est un tirage autour du profil utilisateur**

Session n°11 de Alan:
[85% maths + 15% réseau]

Session n°15 de Mike:
[55% C++ et  45% ML]

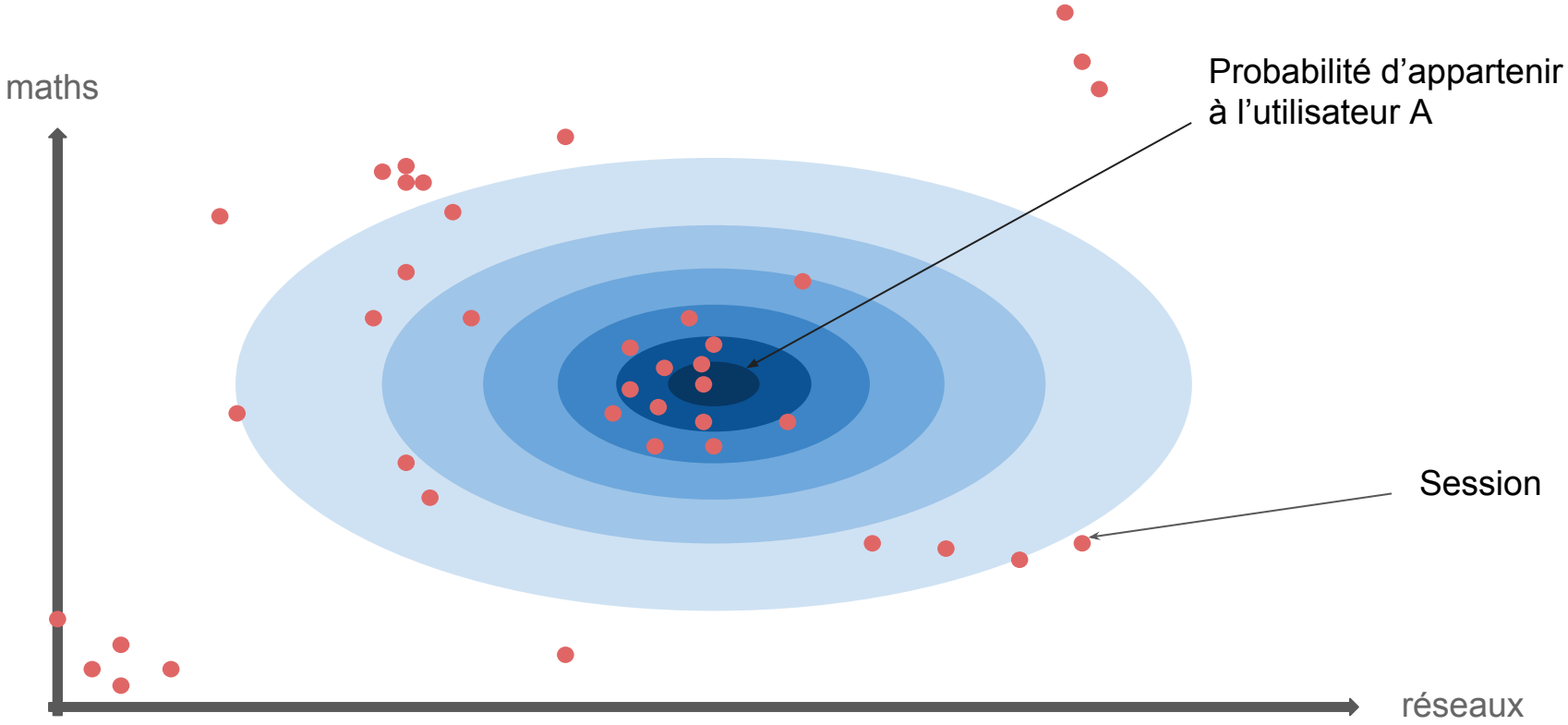
## Tirage des mots

**Pour chaque mot, tirage selon le profil session**

⇒ ‘matlab’, ‘r’

⇒ ‘g++’, ‘tensorflow’, ‘make’

# Approche retenue, modèle génératif

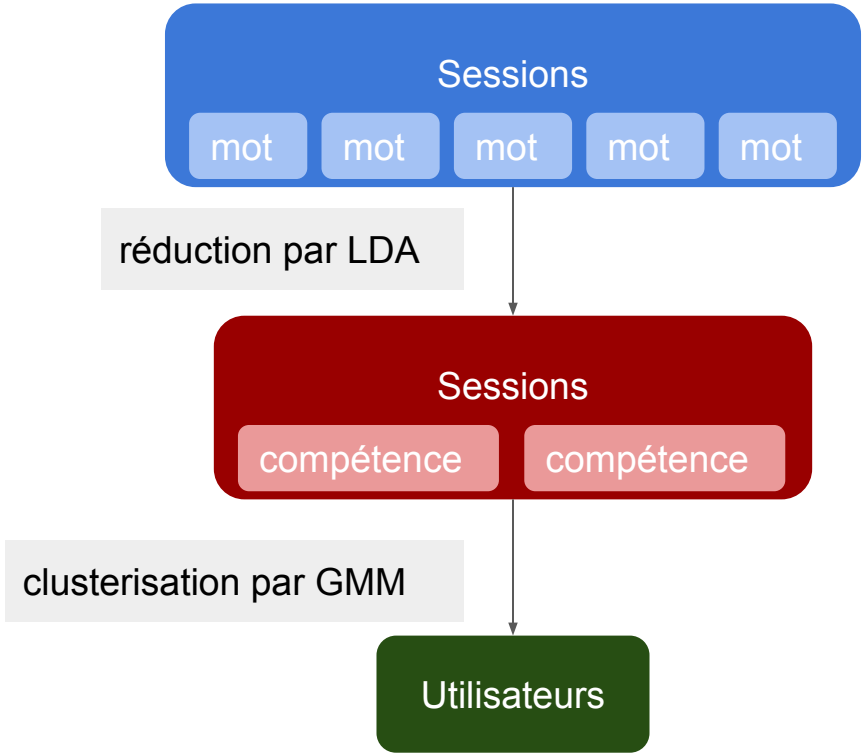


# Prétraitement des données

- découpage par session (balises #BOF# et #EOF#)
- découpage par mot (approche Bag of Words)
- remplacement de certains patterns :

| | |
|---|---|
| ❏ nombres à plus d'un chiffre<br>❏ flags<br>❏ extensions de fichiers<br>❏ fautes de frappe | ❏ _bignum_<br>❏ _dashedarg_<br>❏ _extension_cpp_<br>❏ _typo_ |

# La construction du modèle



# Réduction en compétences

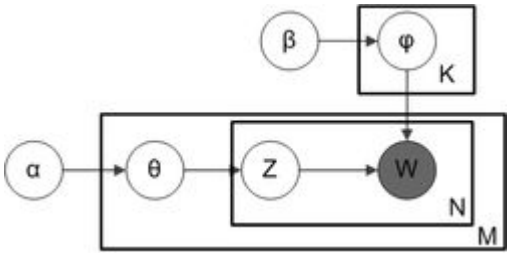
**Pourquoi ?**

- Trop de dimensions par session: autant que la taille du voc !
- Comment généraliser utilisation de mots proche ?

**Modèle**

- Chaque session a une distribution de mots à elle
- Besoin d'une distribution (session), sur les distributions (mots) !

  ⇒ LDA pour inférer param loi Dirichlet




**Critère: perplexité**

- Capacité du modèle à représenter les mots d'une session
- Cross validation pour ne pas overfitter

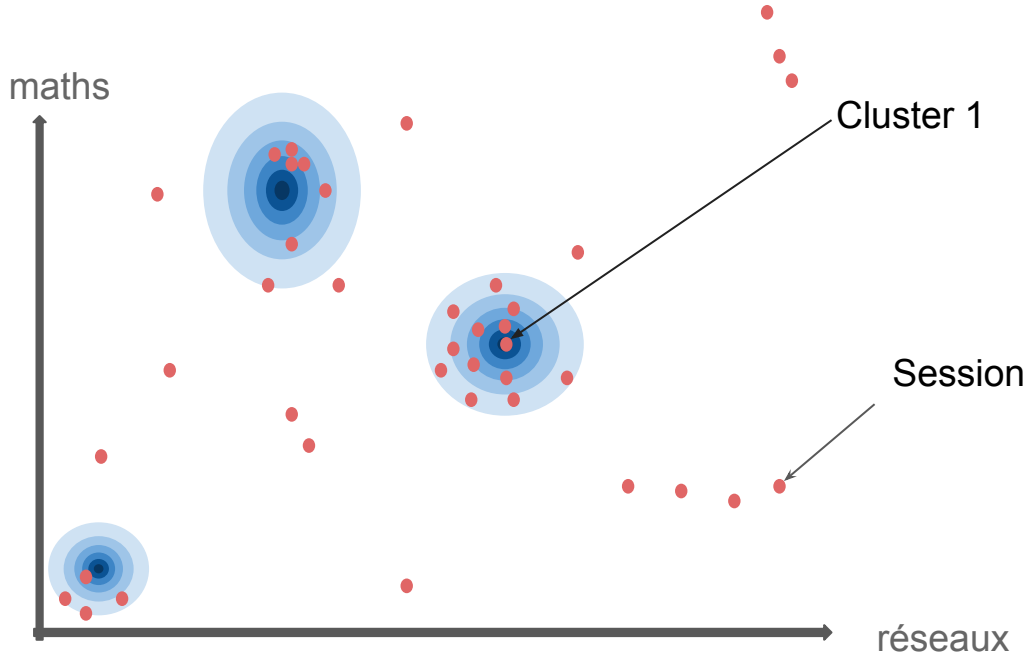
# Clusterisation : GMM

**Motivation:**

- Utilisateur aura souvent des comportements proches
- Comment trouver ces comportements similaires ?

**Modèle:**

- Mélange de gaussiennes



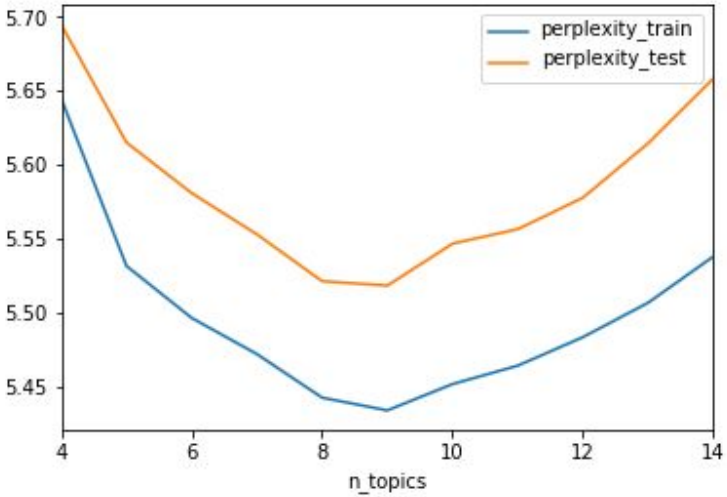
# Clusterisation : GMM

Détermination du nombre de clusters - critère de sélection :

| | |
|---|---|
| **Log-vraisemblance** | • Cohérence entre le modèle et les observations<br>• Favorise les grands nombres de clusters |
| **Calinsky-Harabaz** | • Ratio entre les dispersions intra-cluster et inter-cluster<br>• Bon score pour clusters bien séparables |
| **Silhouette** | • Combine la distance entre exemples intra-cluster et la distance avec le cluster différent le plus proche |
| **Bayesian Information Criterion** | • Log-vraisemblance des données, pénalisée par la taille du modèle |

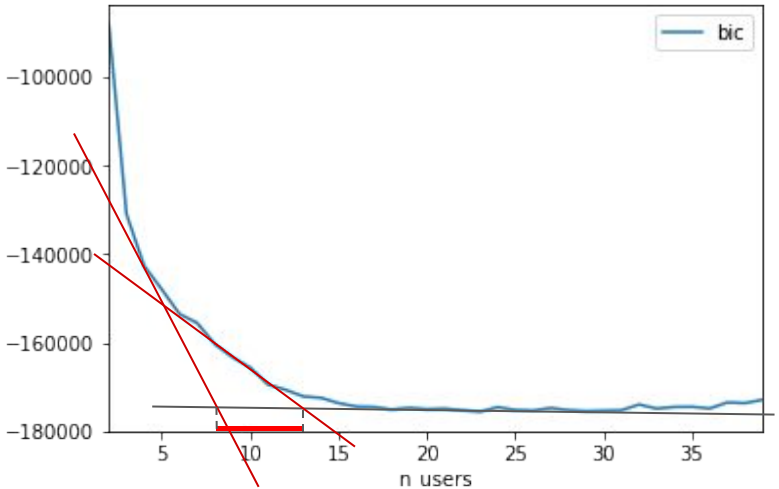
# Résultats du modèle

Perplexité des train et test sets




Nombre de compétences = 9

BIC du modèle




Nombre d’utilisateurs : entre 8 et 13

# Limites du modèle et critique des résultats

Profils session estimés par LDA ont une forte variance → Nombre d'utilisateur estimé par GMM varie entre 8 et 12

Représentation par compétences et hypothèse de profils utilisateurs mixes → Un utilisateur utilisant exclusivement une de ses compétences par session est non-modélisable

Espace d'entrée de GMM n'est pas linéaire: hyperplan positif → Les sessions d'utilisateurs aux profils extrêmes ne sont pas gaussiennes