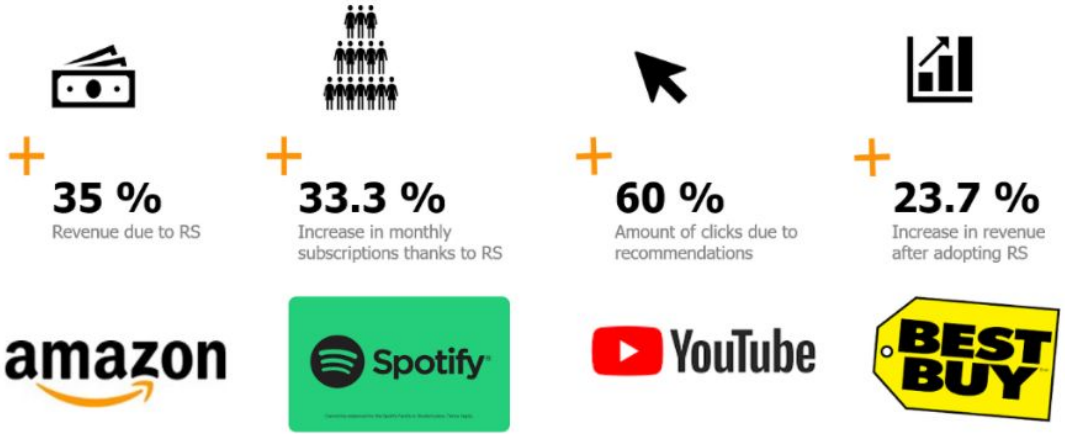
VISIUM

User clustering for movie recommendation

Quick POC for Technical Test

Constraints & Requirements

Executive summary



Recommender system business impact

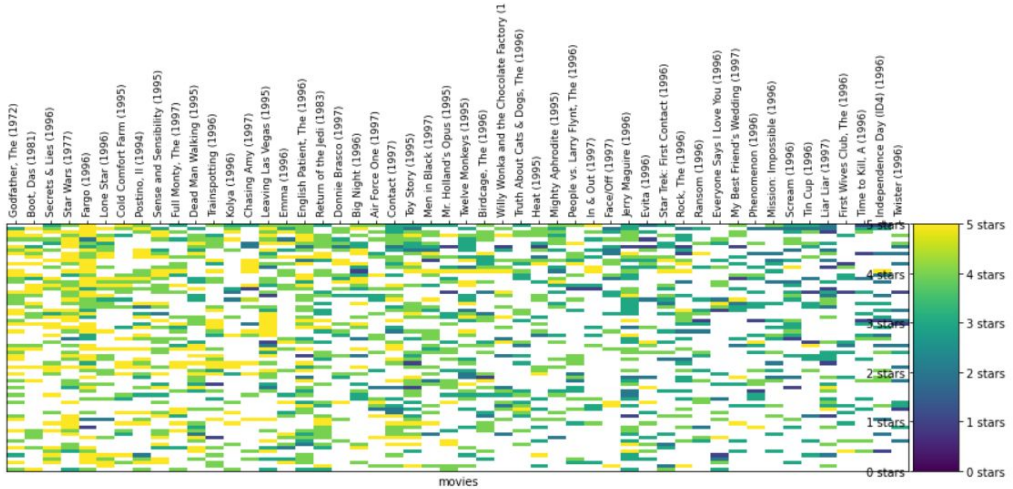
- Recommender systems aim to predict users' interests and recommend product items that quite likely are interesting for them.
- E-commerce and divertissement companies are **leveraging the power of data and boosting sales by implementing** recommender systems on their websites..
  - ➢ E-commerce companies that use the recommender system of visium get >**30% increase in the average price of the user basket.**
  - ➢ Visium recommender system provides recommendations that enable to make **optimal earning from customer behaviour.**

Dataset & constraints

- **Movielens 100k Dataset**
  - 100 000 Ratings (1 to 5)
  - 1682 Movies
  - 987 Users
  - User informations : Age, Occupation
  - 18 Movies Categories

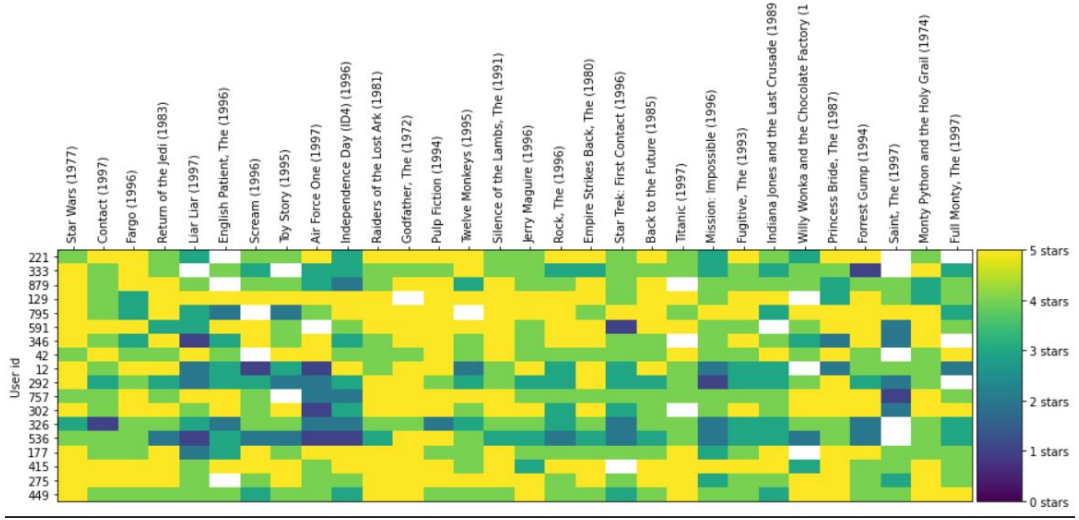
- **Data Constraint**
  - Sparse Dataset

    (Ratings, movies)



Selection of the benchmark :

Rating Features for Clustering

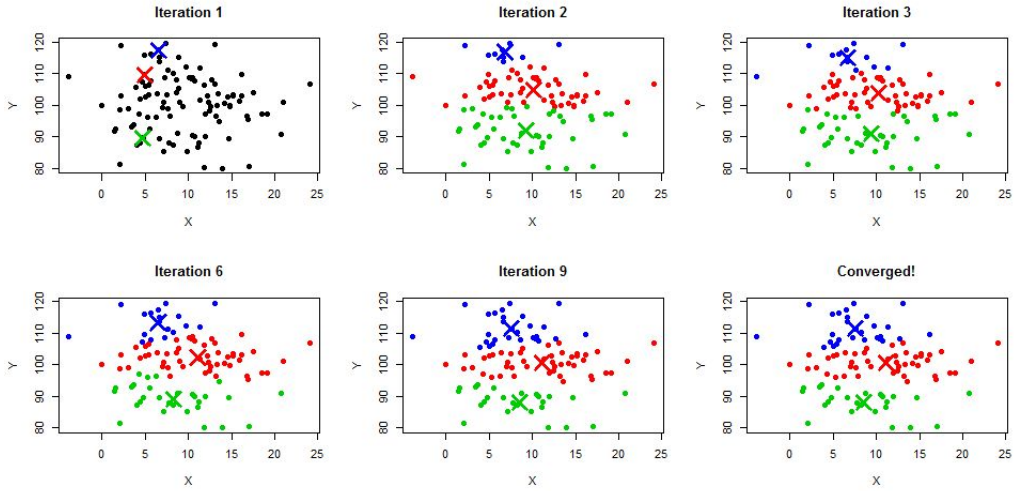


The 30 Most-rated movies & the 18 users with most ratings

- We will make clusters based on the the 1000 most rated movies & corresponding users (out of +9000 in the dataset).
  - ➢ **More dense** and **understandable** than the entire dataset
  - ➢ White cells corresponds to not rated movies for corresponding users. We will still have to deal with sparsity.

First modelling strategy

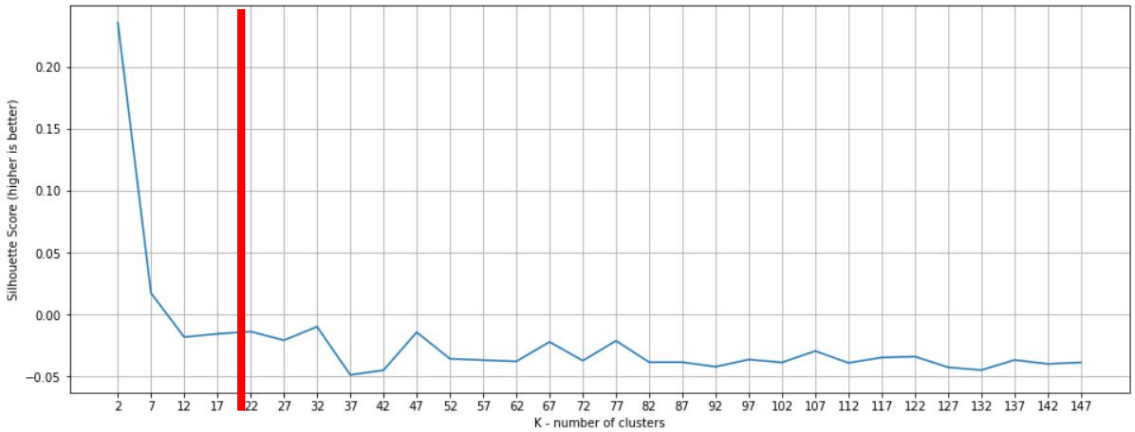
Using Clustering : K-Means to optimize recommendations



K-Means scheme

- K-Means method put **n observations (users) into k clusters** in which each observation belongs to the cluster with the nearest (cluster centroid)..
  - ➢ The second step is to create **new centroids by taking the mean value of all of the samples assigned to each previous centroid.** The border between the old and the new centroids are computed and the algorithm repeats these last two steps until this value is less than a threshold (until the centroids do not move significantly).
  - ➢ The **K-Means** will automatically cluster regarding users ratings..
  - ➢ We measure user distance with a **similarity metric (users = vectors)**

K-Means clustering optimization

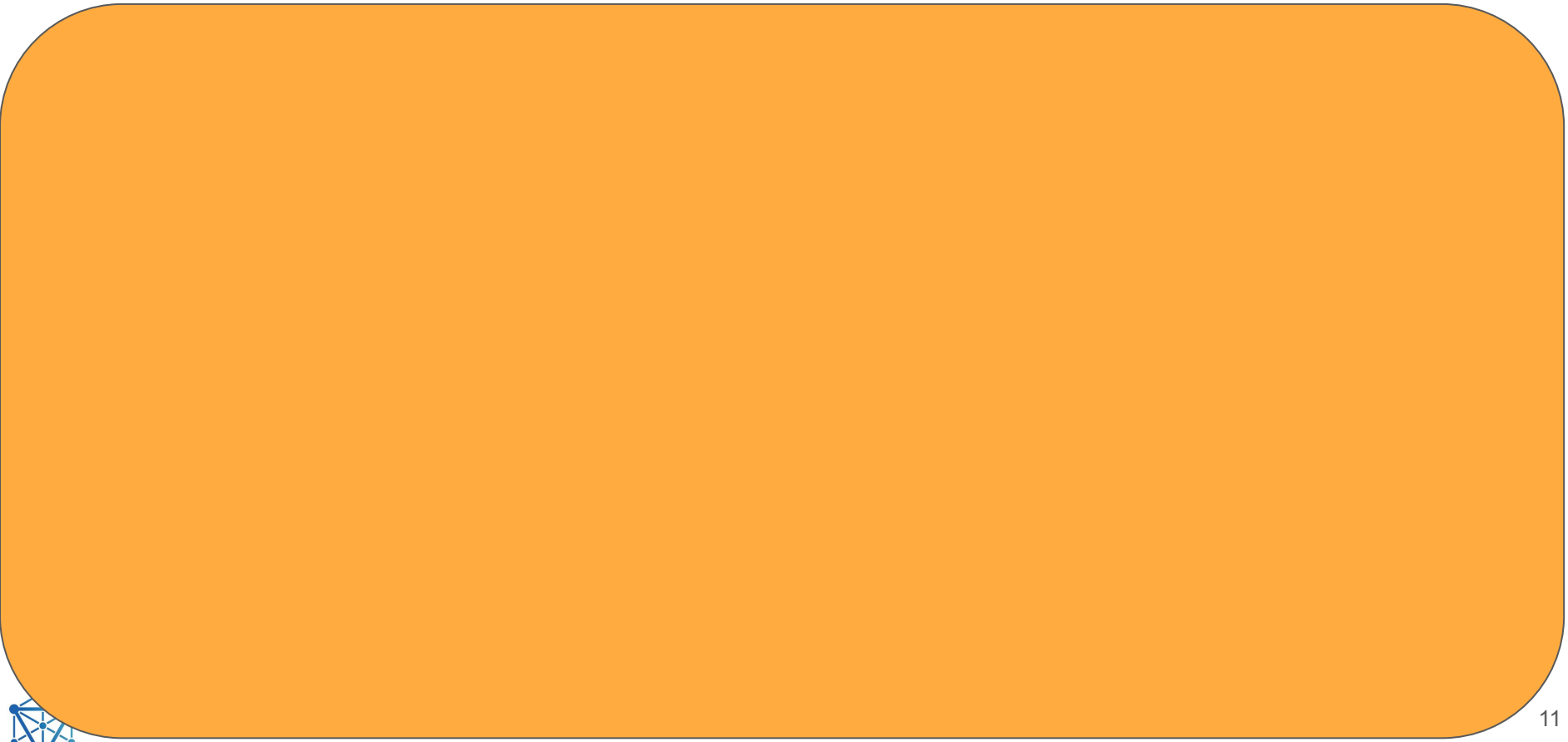


Silhouette score vs number of clusters

- Backtesting the silhouette score of K-Means for a lot of cluster numbers K to evaluate the performance.
  - ➢ **We want to have the one** with the best silhouette score for the highest cluster number.
  - ➢ We will use this K to **have the best clustering algorithm**
  - ➢ We will chose **the optimal K on the inflection point :** here we chose **K = 20**.

Performances

Cluster analysis



Business analysis

Recommendation

Contact

Saussaye Matthieu, Project Manager
Email : saussayematthieu50@gmail.com
Mob. : +33 6 46 04 69 09