



---

# Projet 2 Data Analyst : Création d'un système de recommandation de films (WildSalto)

Par Maëlle, Sabrina, Vincent et Laëtitia  
28 Janvier 2022

---



# Sommaire :

- **Objectif du projet**
- **Présentation de l'équipe**
- **Planning**
- **Etapas**
  - Etape 1 : Analyse
  - Etape 2 : Sélection
  - Etape 3 : Algorithme de films

# Objectif du projet :

Nous sommes Data Analyst freelance. Un cinéma en perte de vitesse situé dans la Creuse nous contacte.

Il nous demande de créer un moteur de recommandation de films qui à terme, enverra des notifications aux clients via une application web.

Au départ, aucun client n'a renseigné ses préférences, nous sommes dans une situation de cold start. Mais heureusement, le cinéma fournit une base de données de films basée sur la célèbre plateforme IMDb.

Nous disposons de **5 semaines** pour réaliser ce projet.

## Outils utilisés :

**Trello** : suivi de projet  
**Jupyter, Visual Studio Code** : notebooks  
**Suite Google** : réunions, présentation  
**Slack** : communication  
**Datasets IMDb** : [lien](#)

The IMDb logo is displayed in a bold, black, sans-serif font. It is centered within a bright yellow rectangular background that has rounded corners.

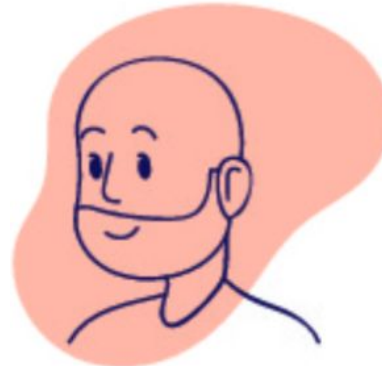
# Présentation de l'équipe :



**Maëlle**  
Data Analyst



**Sabrina**  
Data Analyst



**Vincent**  
Data Analyst



**Laëtitia**  
Data Analyst

# Planning :



## Semaines 1-2

Appropriation et première exploration des données  
(Pandas, NumPy, Matplotlib)

## Semaine 4

Modèles de Machine Learning, recommandations (Scikit-Learn)

Décembre

Janvier

## Semaines 2-3

Jointures, filtres, nettoyage, recherche de corrélations  
(Pandas, Seaborn)

## Semaine 5

Affinage, présentation et Demo Day



## Étapes :

- Analyse complète
- Sélection
- Algorithme de recommandation de films

# Étape 1 :

# Analyse complète des bases de données

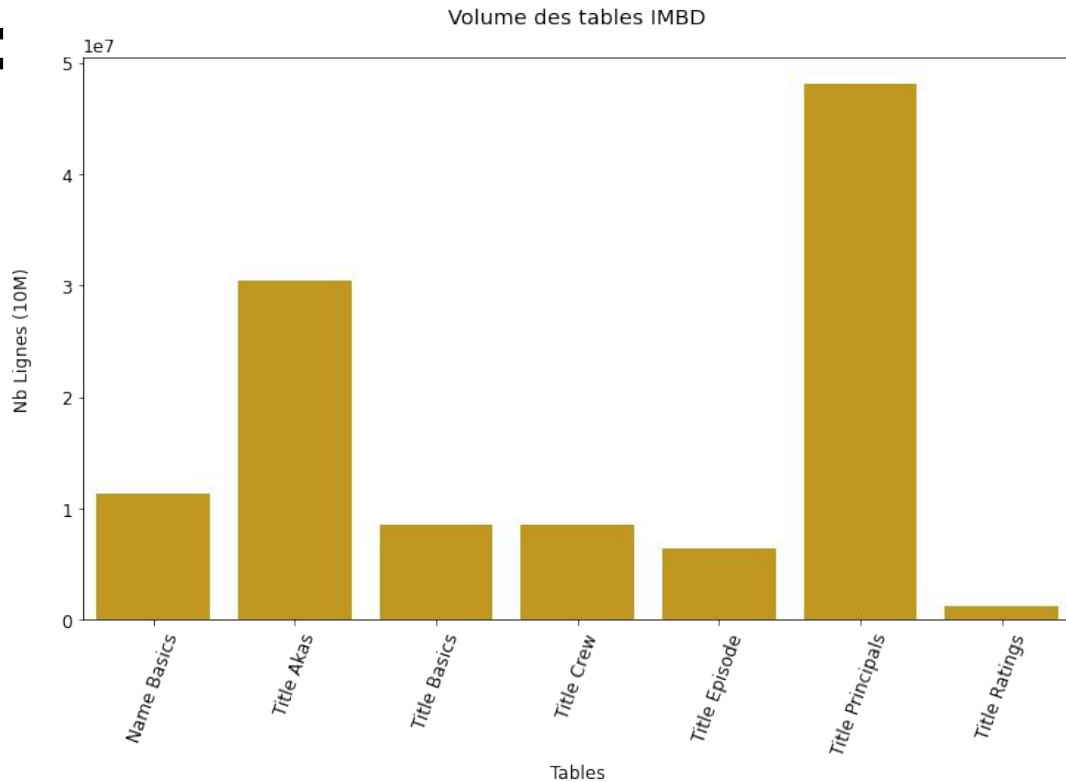


## Objectif :

Prise de connaissance des contenus de chaque table, puis sortie de KPI sous forme de graphiques pertinents.

Grâce à cette étape, nous pourrons ensuite spécialiser notre cinéma.

# 7 tables :



## Volume des tables

Quantité des données par table



# Analyse des tables :

NAME BASICS : ACTEURS

nconst	primaryName	birthYear	deathYear	primaryProfession	knownForTitles
nm0000001	Fred Astaire	1899	1987	soundtrack,actor,miscellaneous	tt0072308,tt0050419,tt0053137,tt0031983
nm0000002	Lauren Bacall	1924	2014	actress,soundtrack	tt0117057,tt0075213,tt0038355,tt0037382
nm0000003	Brigitte Bardot	1934	\N	actress,soundtrack,music_department	tt0049189,tt0054452,tt0057345,tt0056404
nm0000004	John Belushi	1949	1982	actor,soundtrack,writer	tt0080455,tt0072562,tt0077975,tt0078723
nm0000005	Ingmar Bergman	1918	2007	writer,director,actor	tt0050976,tt0050986,tt0060827,tt0083922

TITLE BASICS

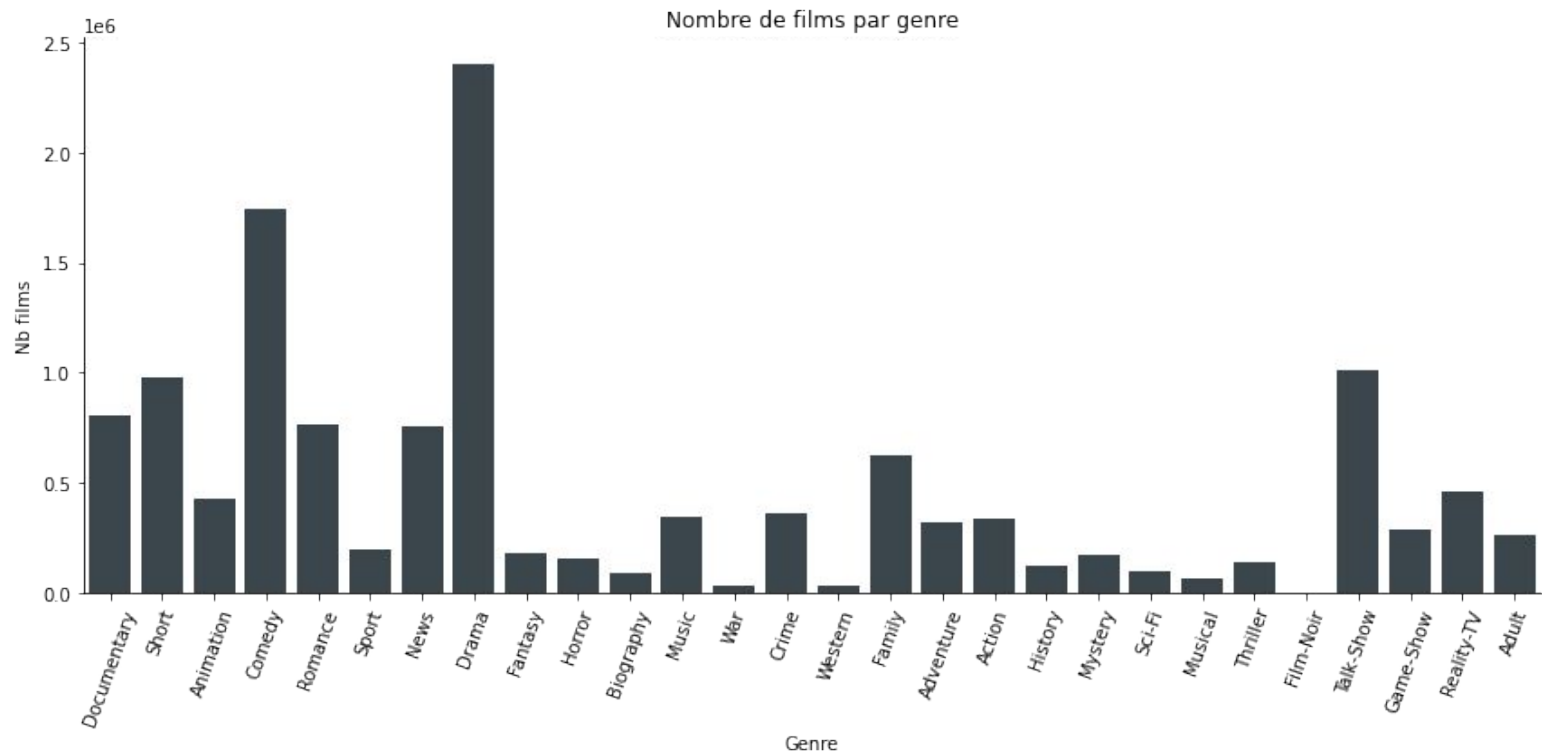
tconst	titleType	primaryTitle	originalTitle	isAdult	startYear	endYear	runtimeMinutes	genres
tt0000001	short	Carmencita	Carmencita	0	1894	\N	1	Documentary,Short
tt0000002	short	Le clown et ses chiens	Le clown et ses chiens	0	1892	\N	5	Animation
tt0000003	short	Pauvre Pierrot	Pauvre Pierrot	0	1892	\N	4	Animation
tt0000004	short	Un bon bock	Un bon bock	0	1892	\N	12	Animation
tt0000005	short	Blacksmith Scene	Blacksmith Scene	0	1893	\N	1	Comedy

TITLE RATINGS

tconst	averageRating	numVotes
tt0000001	5.7	1847
tt0000002	6.0	238
tt0000003	6.5	1609
tt0000004	6.0	155
tt0000005	6.2	2432

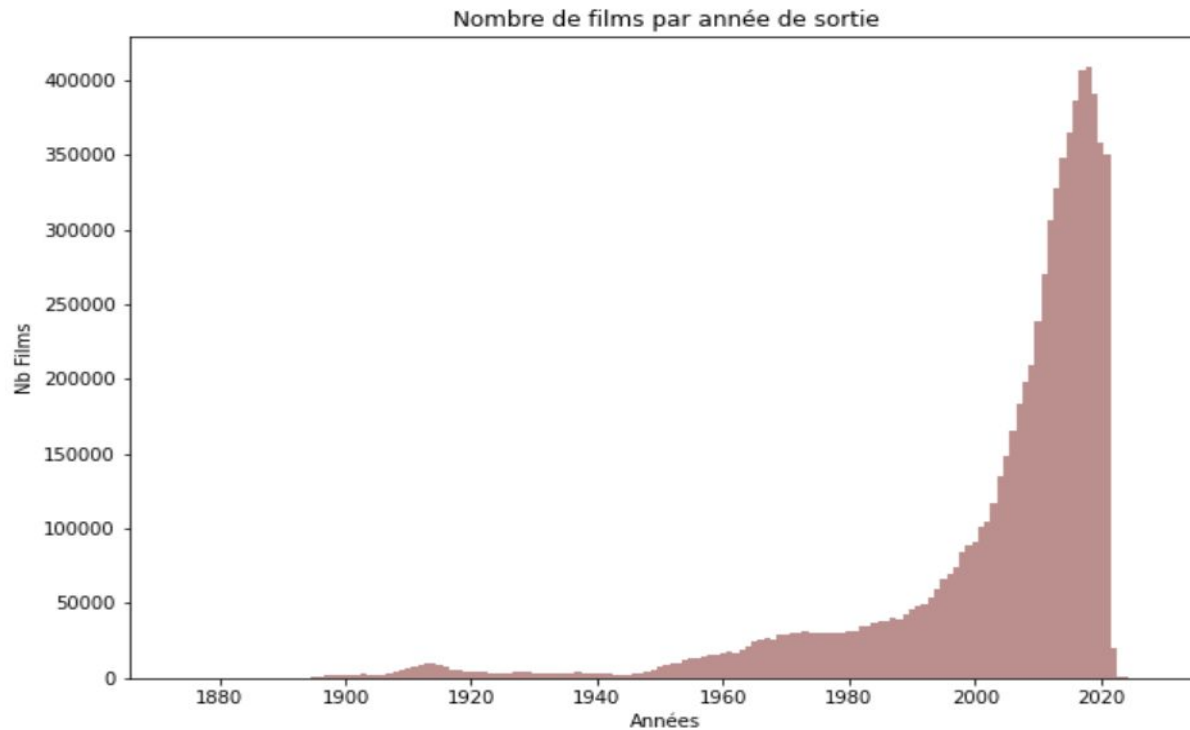
TITLE AKAS : TITRES FILMS

titleId = tconst	ordering	title	region	language	types	attributes	isOriginalTitle
tt0000001	1	Карменцита	UA	\N	imdbDisplay	\N	0
tt0000001	2	Carmencita	DE	\N	\N	literal title	0
tt0000001	3	Carmencita - spanyol tánc	HU	\N	imdbDisplay	\N	0
tt0000001	4	Карμενοίτα	GR	\N	imdbDisplay	\N	0
tt0000001	5	Карменсита	RU	\N	imdbDisplay	\N	0



## Nombre de films par genre

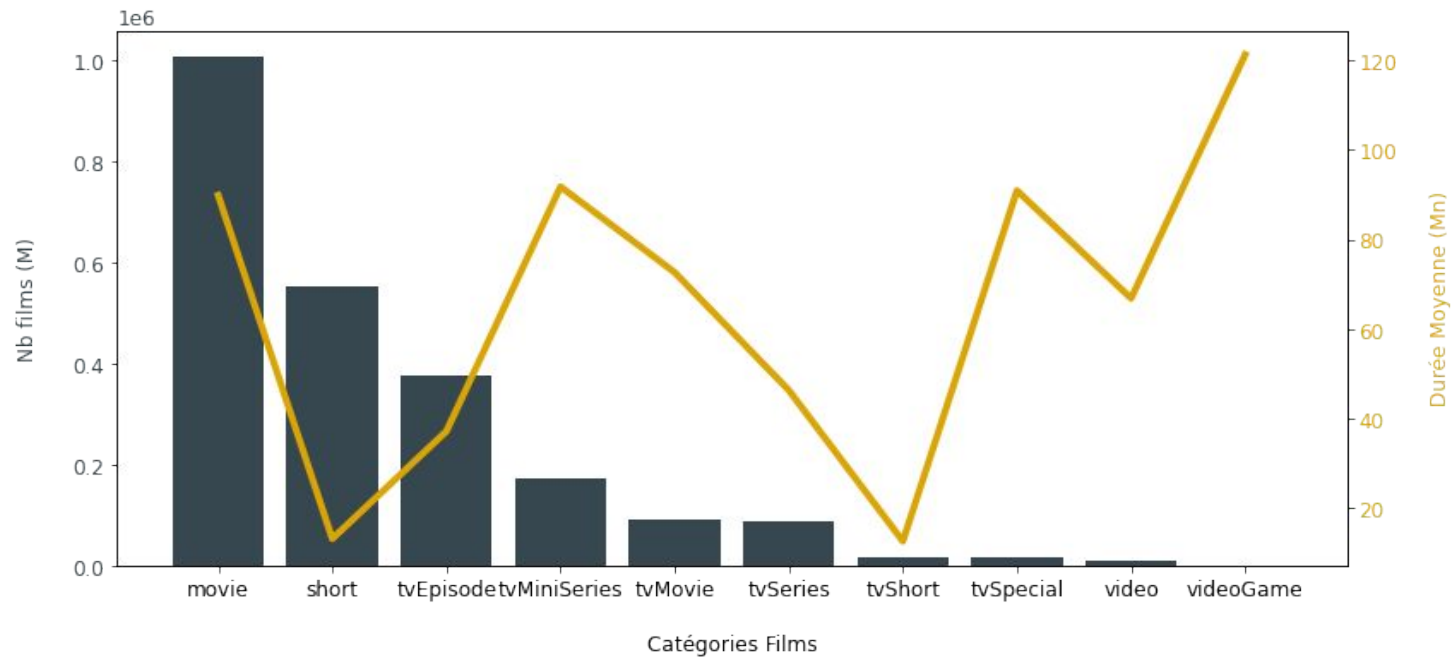
Sélection des genres Comédie, Drame, Romance



## Nombre de films par année de sortie

Sélection : 1950 - 2022

Nombre de films vs durée moyenne par catégories



Nombre de films par type

Sélection : movie

## “Anomalies” observées durant l’exploration des données :

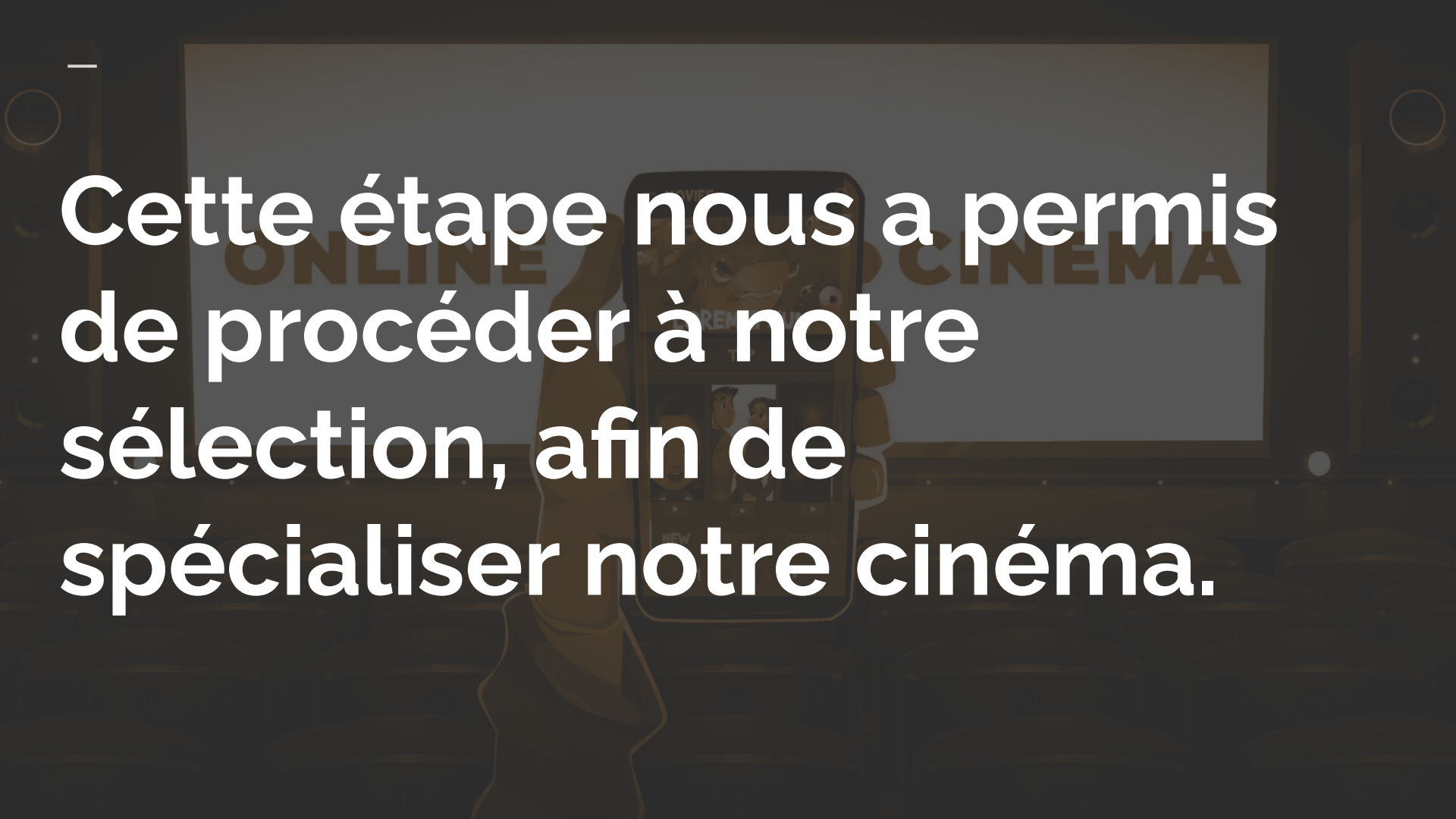
- C. Daveillans, né en 1962, mort en 1936 !
- Dennis Reddy : mort en 1967 au lieu de 1947 !
- Michael Hook : né en 1946, mort en 1913 !
- Titus Livius : né en -59 avant JC est présent !
- Felipe Villanueva (1862-1863) apparaît dans des films des années 1980 !
- Jakob Haringer (1948-1948) était scénariste en 1994...
- Aristote : -384 / -322 avant JC est présent !
- Charline Arthur (1929-1929) : apparaît dans un doc de 2001 !
- Yuki Kawamura, née en 1998, connue pour des films de 1968 et 1972 !
- **Le saviez-vous ?** Dexter The Kitten (“A Cat’s Life”) était un chaton victime de cruauté animale... :(
- Une ligne d’une série dont les éléments d’une colonne sont décalées (ie: le type dans le runtime :-D)



Inconvénient du NoSQL nous pouvons insérer n’importe quoi : il y a beaucoup d’erreurs...

—

Cette étape nous a permis  
de procéder à notre  
sélection, afin de  
spécialiser notre cinéma.

The background features a dark, stylized illustration of a person's profile holding a smartphone. The phone screen displays a movie application interface with various movie posters and the text 'ONLINE CINEMA'. The overall aesthetic is modern and digital.

# Étape 2 :

## Sélection de films



### Objectif :

Dans le cadre de la spécialisation du cinéma. Puis, nouvelle sortie de KPI sous forme de graphiques pertinents pour notre sélection.

# Notre sélection **IMDb** :

Elle se compose de films :

- tous publics
- toutes nationalités



Genres de film :

- Comedy
- Drama
- Romance

Dates de sortie de films comprises entre :

- 1950 et 2022

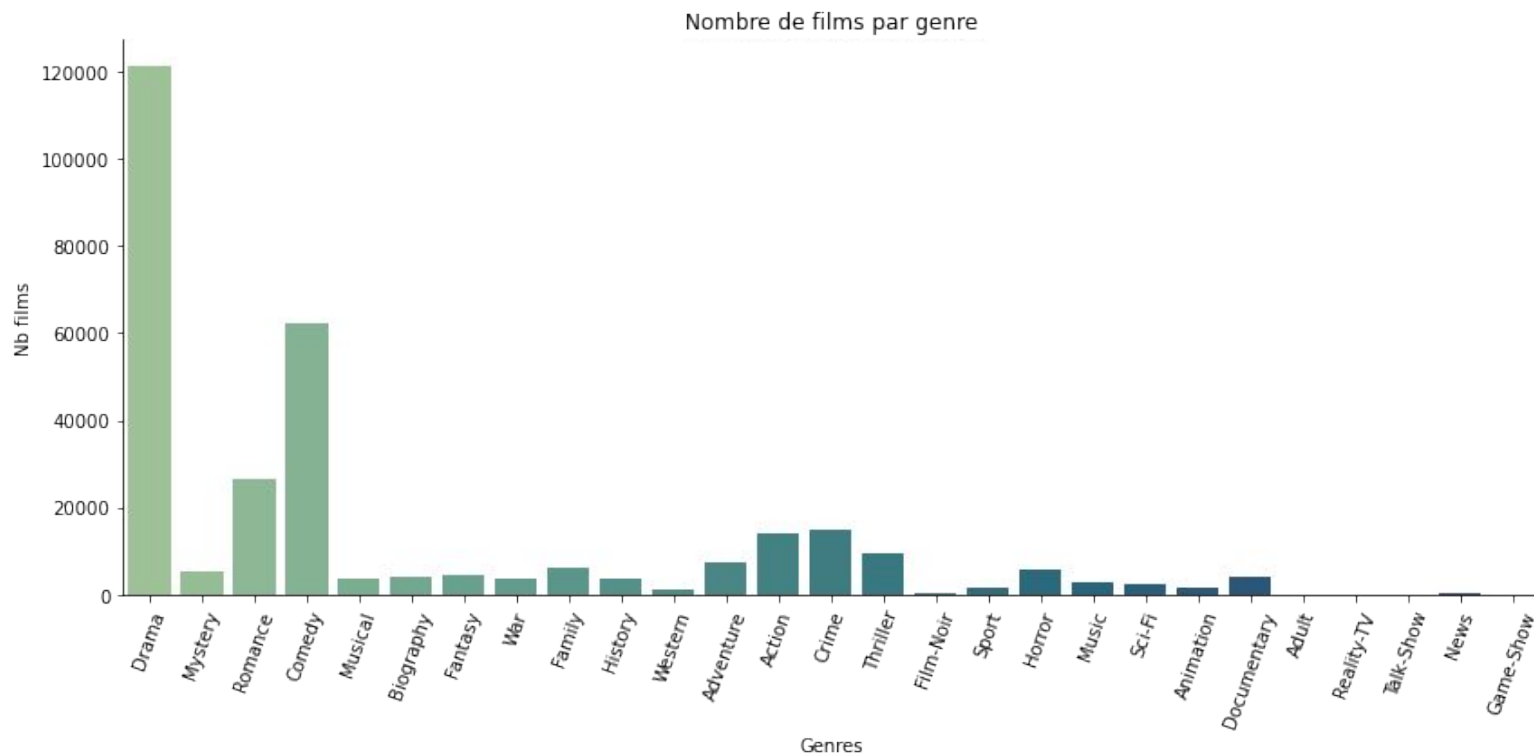
Durées de films comprises entre :

- 69 et 240 min

Elle compte :

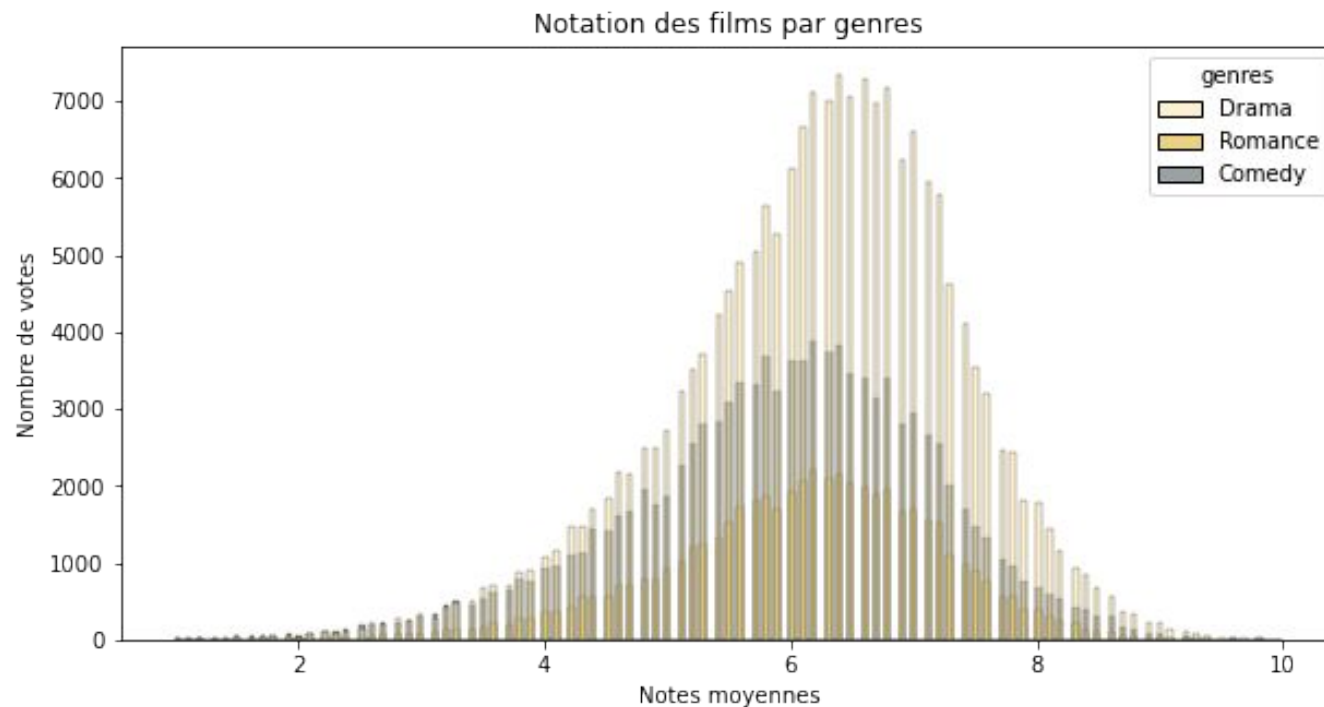
**169 929 films !**





## Nombre de films par genre

Genres principaux Drame, Romance, Comédie

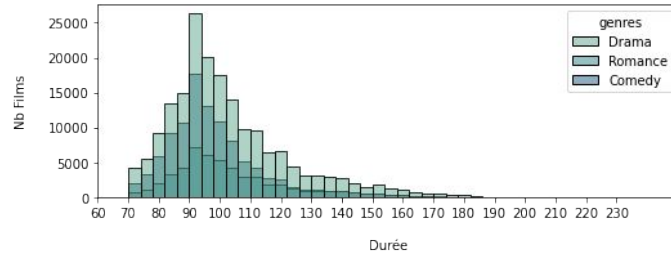


## Notation et nombre de votes par genres

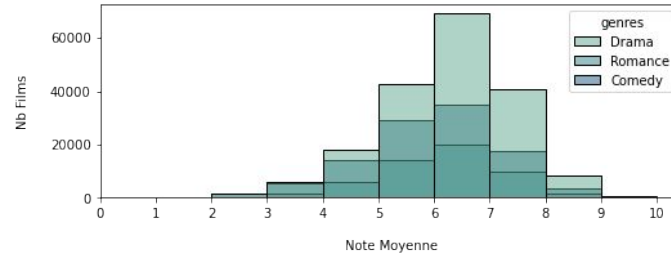
Genres principaux : Drame, Romance, Comédie

# **Analyse Durée vs Notation de la sélection.** **Sur la catégorie principale du film Comédie ou Dramatique ou Romance.**

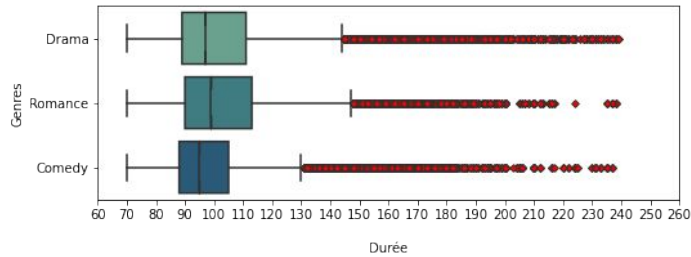
Nombre de films par durée



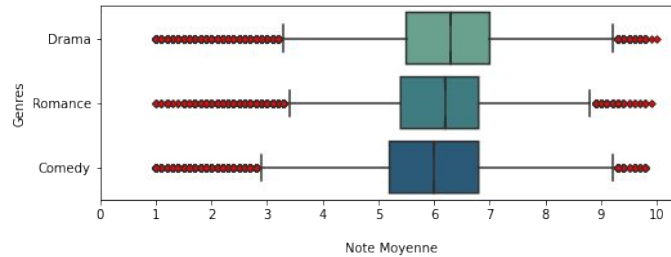
Nombre de films par note



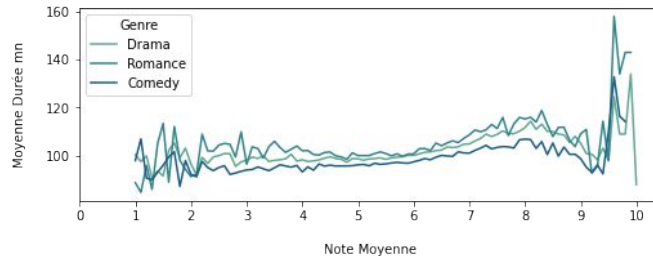
Répartition nombre de films selon leur durée



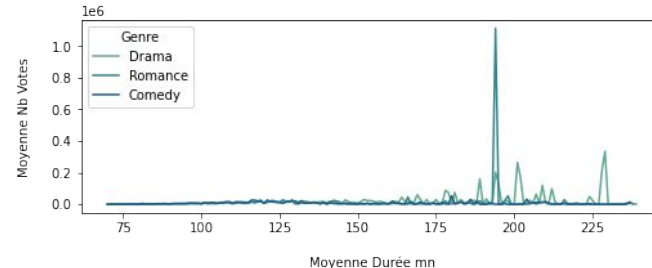
Répartition nombre de films selon leur notation



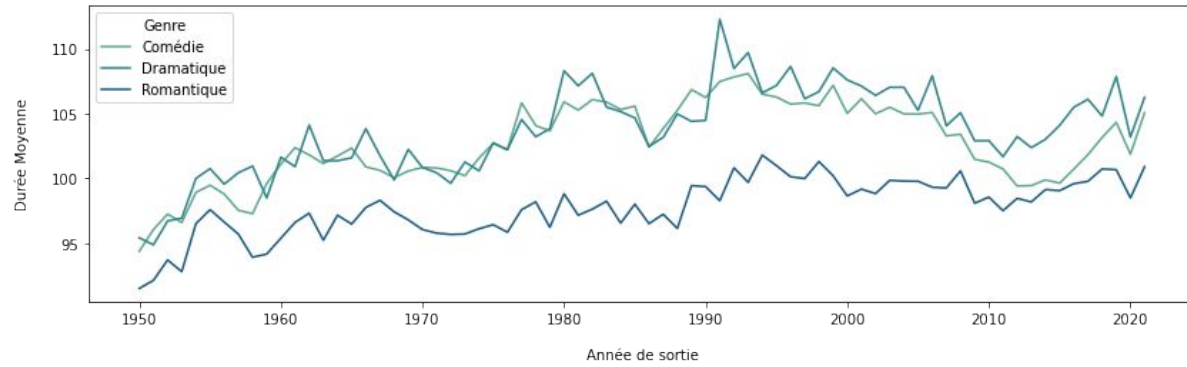
Note vs Durée



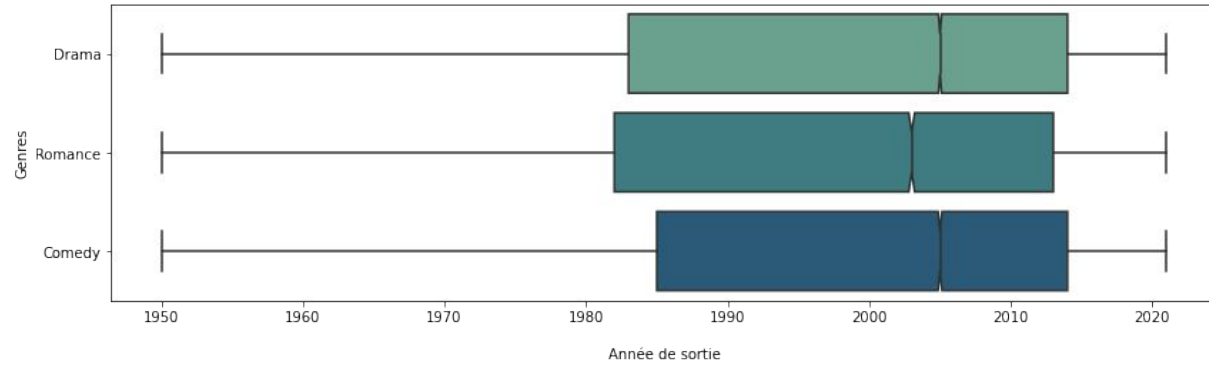
Nombre de votes vs Durée



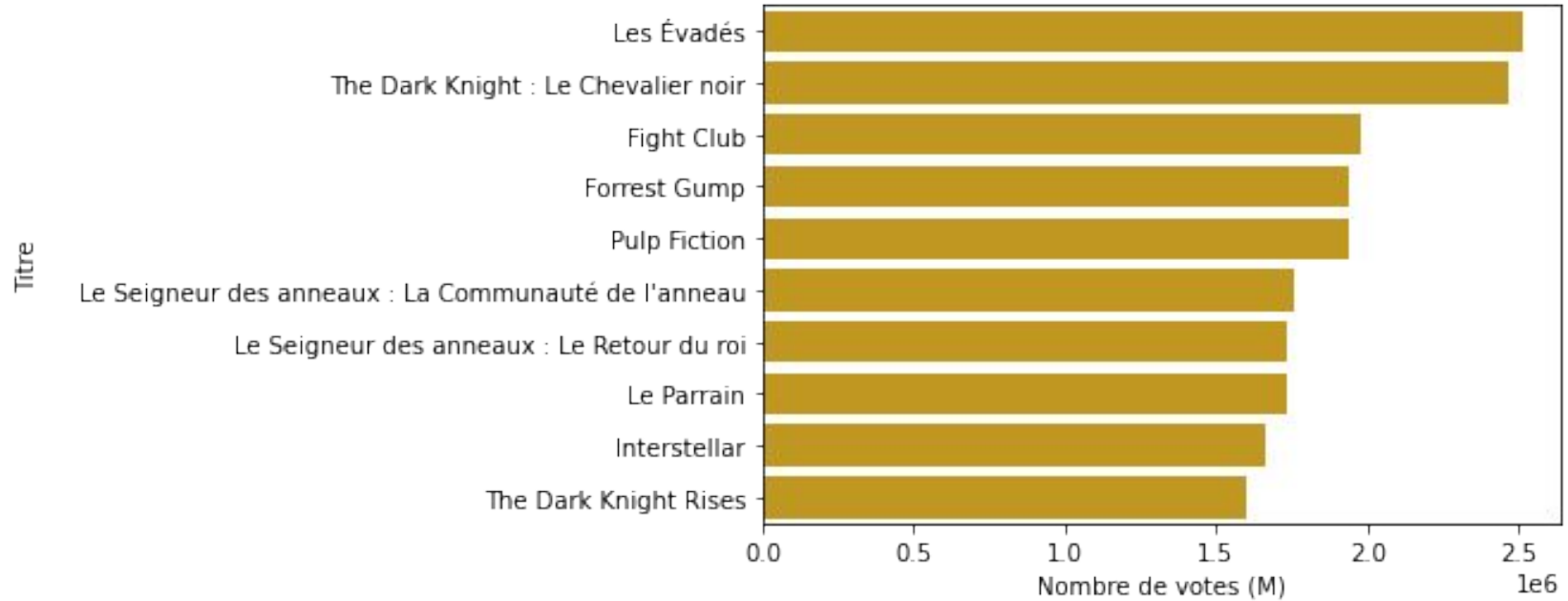
Evolution Durée Par Date de Sortie



Répartition nombre de films selon leur date de sortie



Les 10 films ayant eu le plus de votes



**Top 10 des films ayant obtenu le plus de votes**

# Étape 3 :

# Utilisation d'un algorithme de recommandation de films

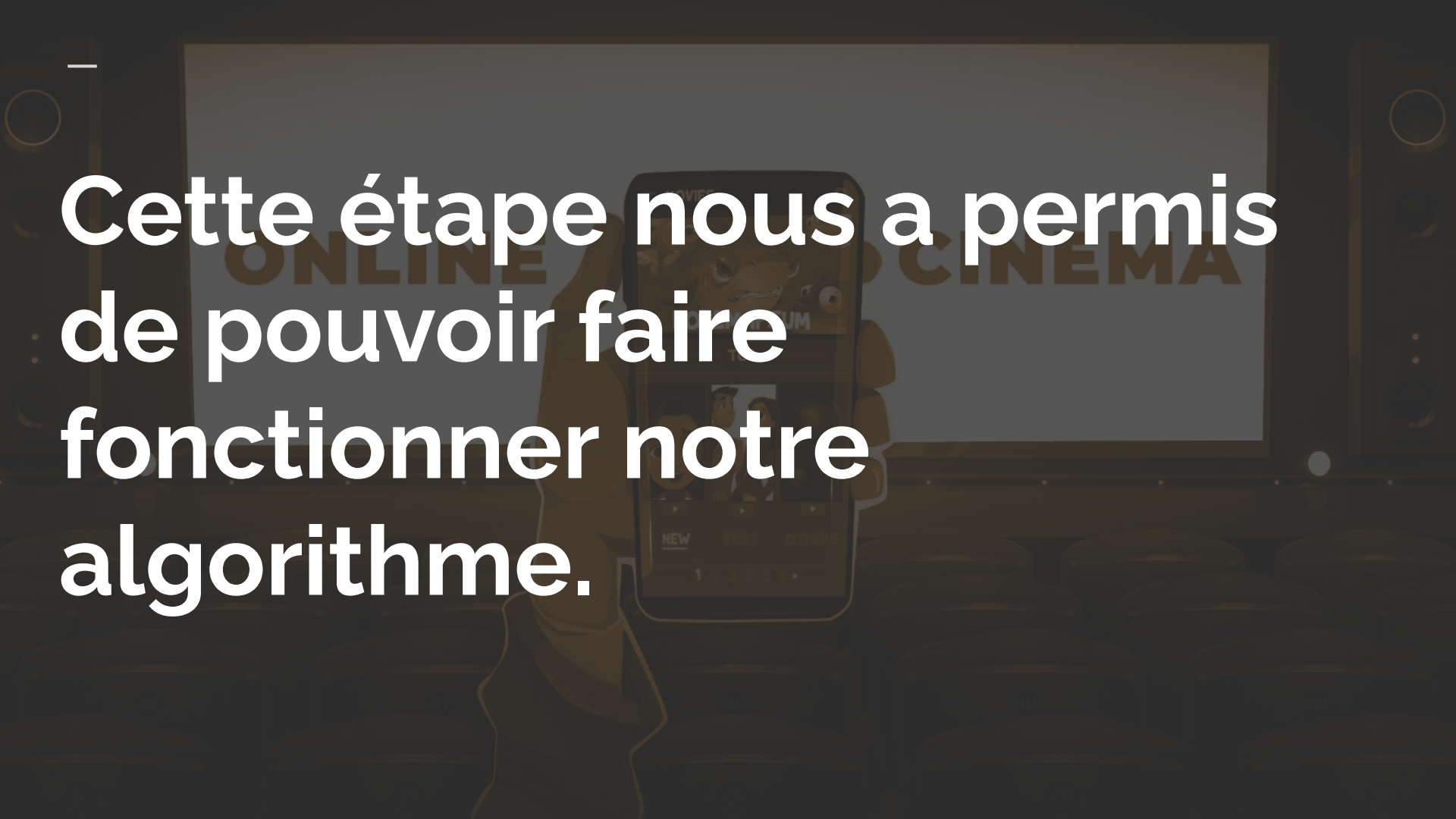


## Objectif :

Réalisation et application de modèles de Machine Learning à partir de la sélection pour la recommandation de films aux clients.

—

Cette étape nous a permis  
de pouvoir faire  
fonctionner notre  
algorithme.

The background is a dark, stylized illustration. In the center, a hand holds a smartphone. The phone's screen shows a movie application interface with a top banner for 'COMICS' featuring a character, a 'NEW' section with a movie poster, and a bottom navigation bar with 'NEW', 'BEST', and 'OTHERS' tabs. In the background, a large screen displays the words 'ONLINE' and 'CINEMA' in a bold, sans-serif font. The overall aesthetic is modern and tech-oriented.

## Définition des features :

- Nombre de traduction par film :
  - nombre de régions par titre (tconst)
- Langues dans lesquelles le film a été traduit :
  - Français, Anglais GB, Anglais US
- Titres traduits en Français :
  - Pour la recherche par nom de films
- Nombre de genre par film :
  - On peut avoir jusqu'à 3 genres
- Nombre d'acteurs, d'actrices, de "self" (acteurs ayant joué leur propre rôle)
- Création de colonnes (+ de 4000) pour chaque réalisateur dont les films sont les mieux notés (>8)



—

**Cette étape permet au client d'avoir des recommandations de films en fonction d'un titre.**

The background features a stylized illustration of a person's head and shoulders in profile, holding a smartphone. The phone screen displays a movie recommendation interface with the text 'MOVIES' at the top and 'ONLINE CINEMA' in large letters below. The overall aesthetic is dark and modern, with decorative circular elements in the corners.



## Point sur notre modèle

Points positifs	Points d'amélioration
<ul style="list-style-type: none"><li>● Le film recherché est trouvé (quand il est dans la base)</li><li>● Message lorsque le film cherché n'est pas dans la base</li><li>● Proposition de films si plusieurs films du même titre</li></ul>	<ul style="list-style-type: none"><li>● Ajouts de features :<ul style="list-style-type: none"><li>- % d'actrices</li><li>- Acteurs célèbres</li></ul></li><li>● Permettre de faire des recherches en fonction du titre original</li><li>● Réduire le nombre de colonnes de réalisateurs (test d'un top 100)</li></ul>