Music Recommender

Nohan Budry
Pierre Dominguez
Rémy Vuagniaux

Cours BDA
Salle C37
10.07.2022

Sommaire

Description du dataset

Description du projet

Description du modèle

Questions annexes

Clustering

Market basket analysis

Ratio torrent/achat

Conclusion

Questions/réponses

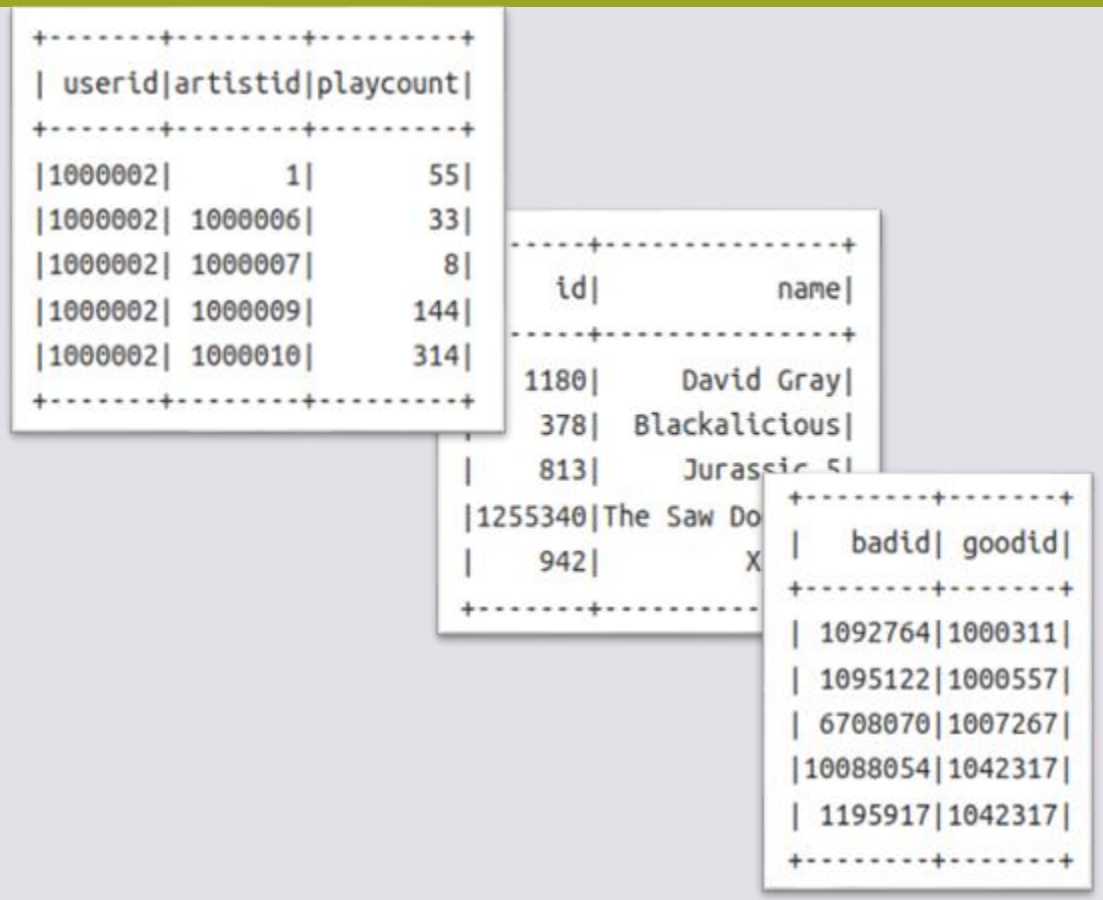
Description du dataset

- **user_artist_data.txt** : 24’296’858

  3 columns: userid, artistid, playcount
- **artist_data.txt** : 1’848’281

  2 columns: artistid, artist_name
- **artist_alias.txt** : 190’892

  2 columns: badid, goodid



Description du projet

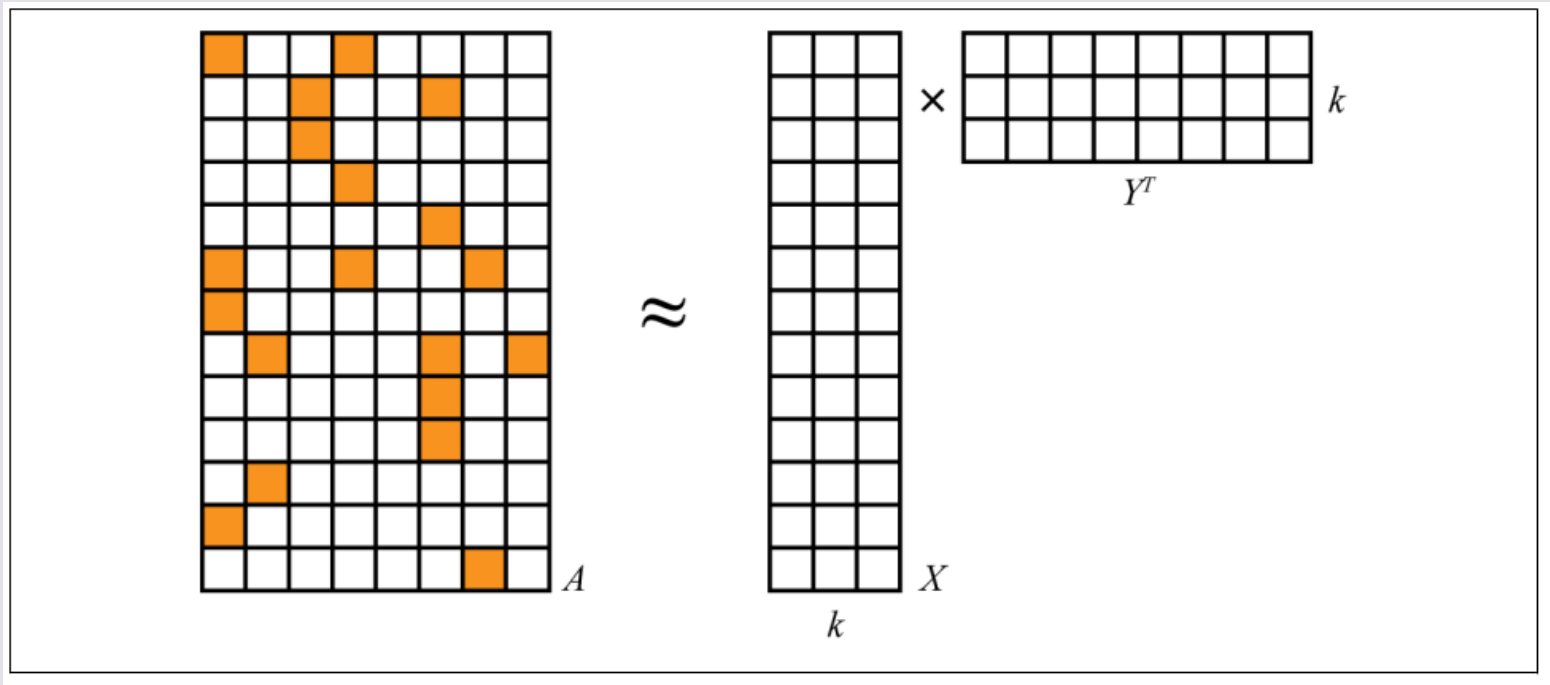
- **Audio recommender**
- **Réaliser des prédictions personnalisées par utilisateurs**
- **Evaluer le modèle et l'optimiser**

Description du modèle

- **Collaborative filtering**
- **Item: artists**
- **User: user**

$$A_i Y (Y^T Y)^{-1} = X_i$$

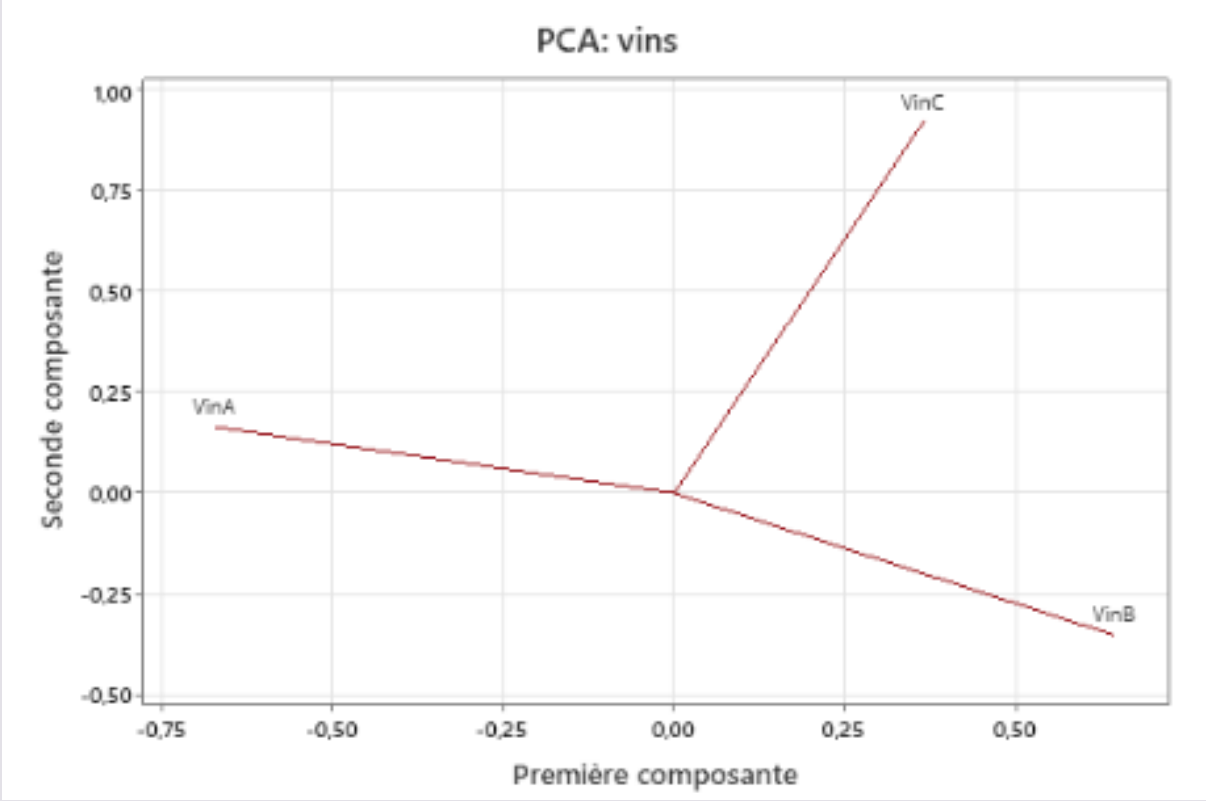
$$\left| A_i Y (Y^T Y)^{-1} - X_i \right|$$



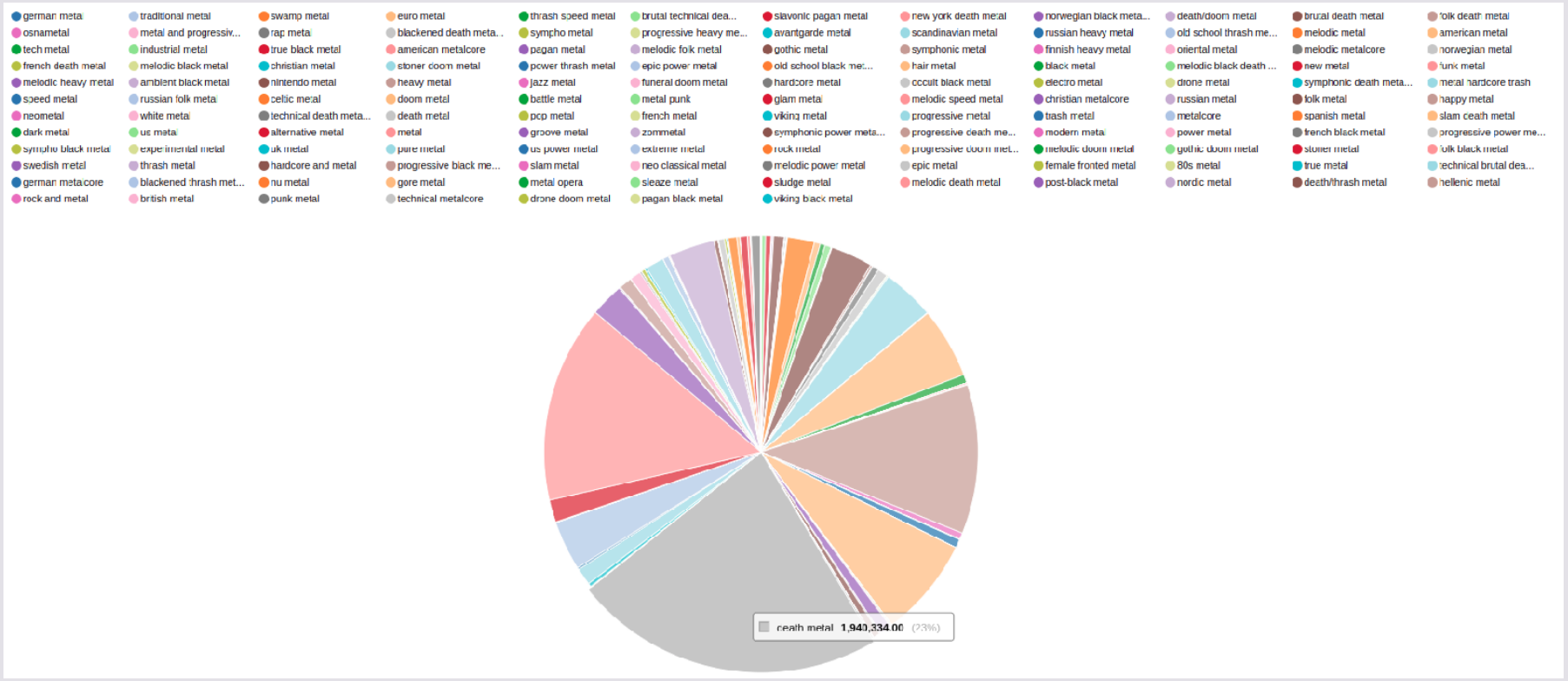

Clustering

Clustering

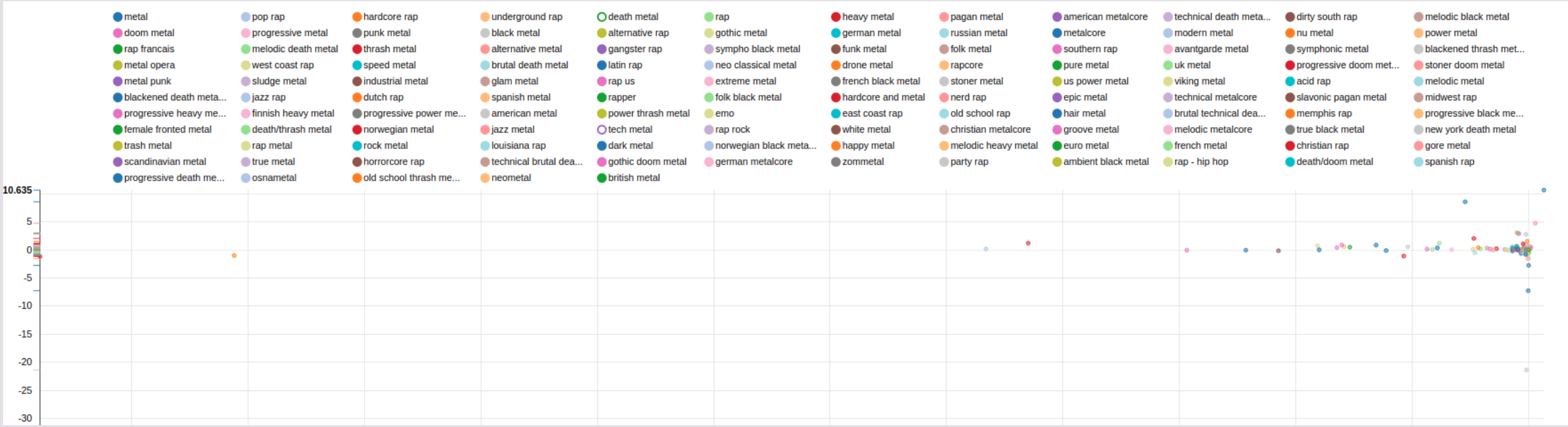
- **Réduction de la dimentionnalité**
- **Features: utilisateurs**
- **Items: artists**



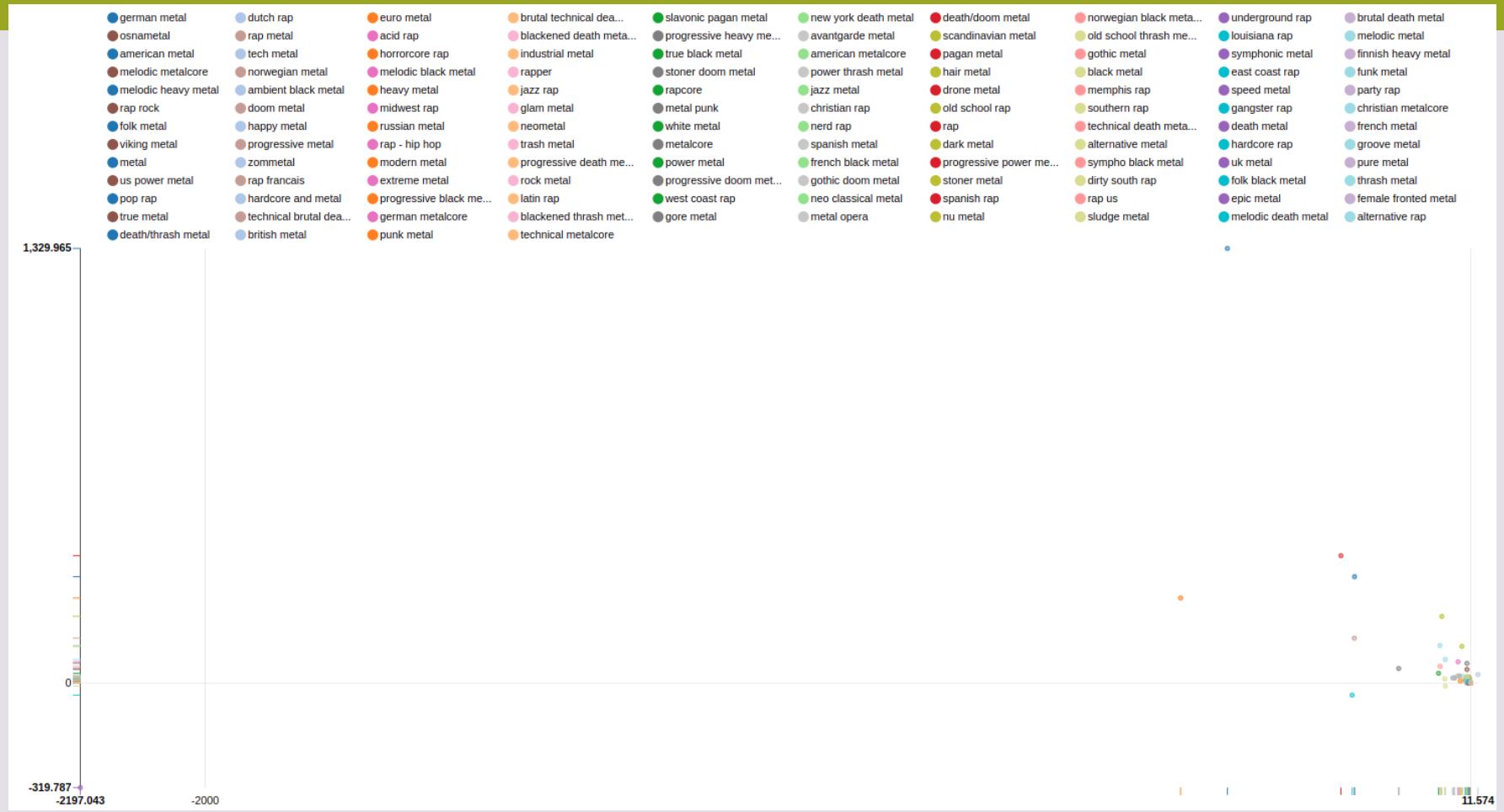
Clustering: Style Dataset



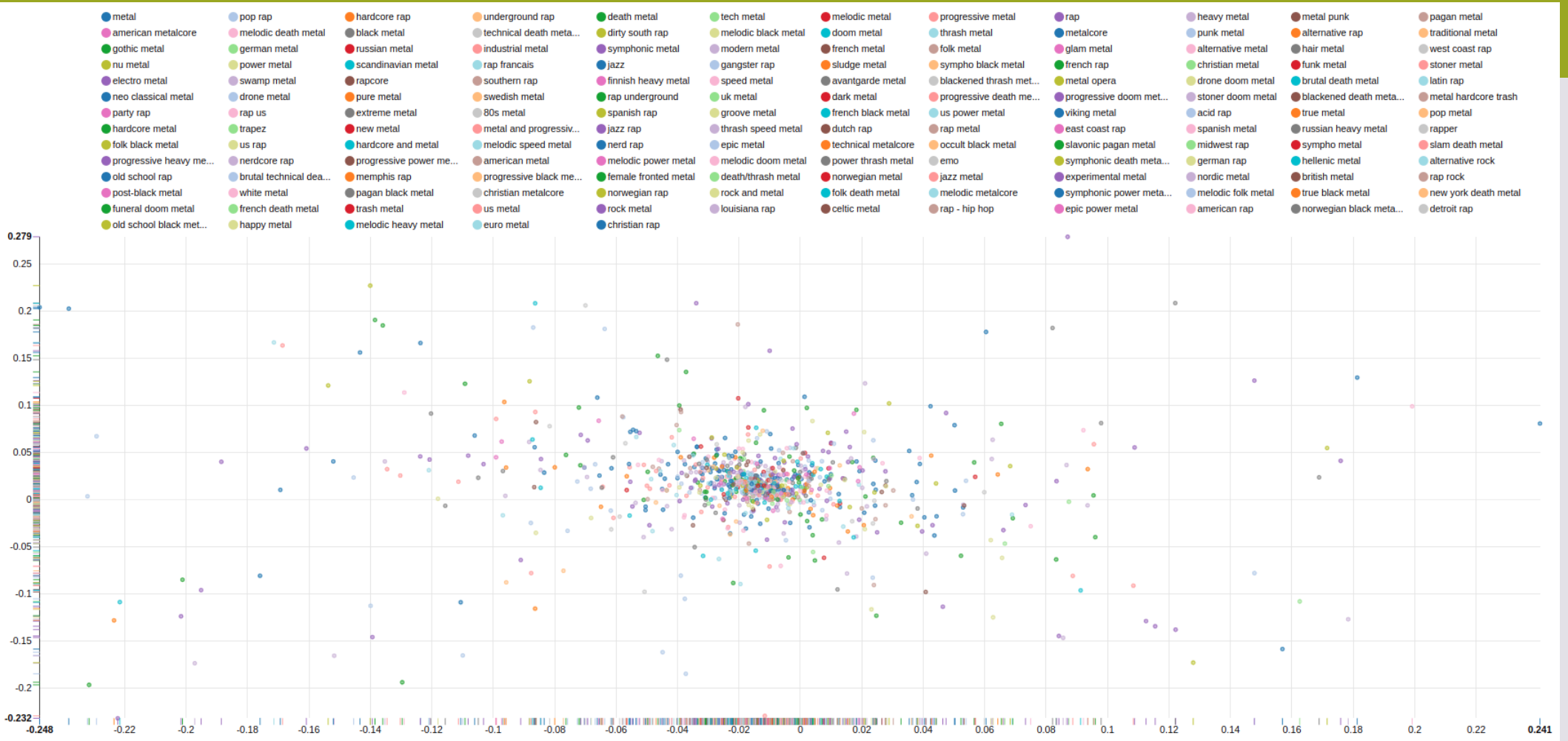
Clustering: Resultat



Clustering: Resultat



Clustering: Word2Vec



Market basket analysis

Market basket analysis

- **FP-Growth**
- **Règles d'associations**
  - **[A, B, C] => D**
  - **Confiance et Support**

Market basket analysis

- **Transformation des données**
  - User 1: Coldplay
  - User 2: U2
  - User 3: Coldplay
  - User 2: The Beatles
  - User 3: Mika

  - User 1: Coldplay
  - User 2: U2, The Beatles
  - User 4: Coldplay, Mika

  - ~150000 transactions
  - ~164 (min: 1, max: 6836) artists

- groupBy("user").agg(collect_set("artist"))

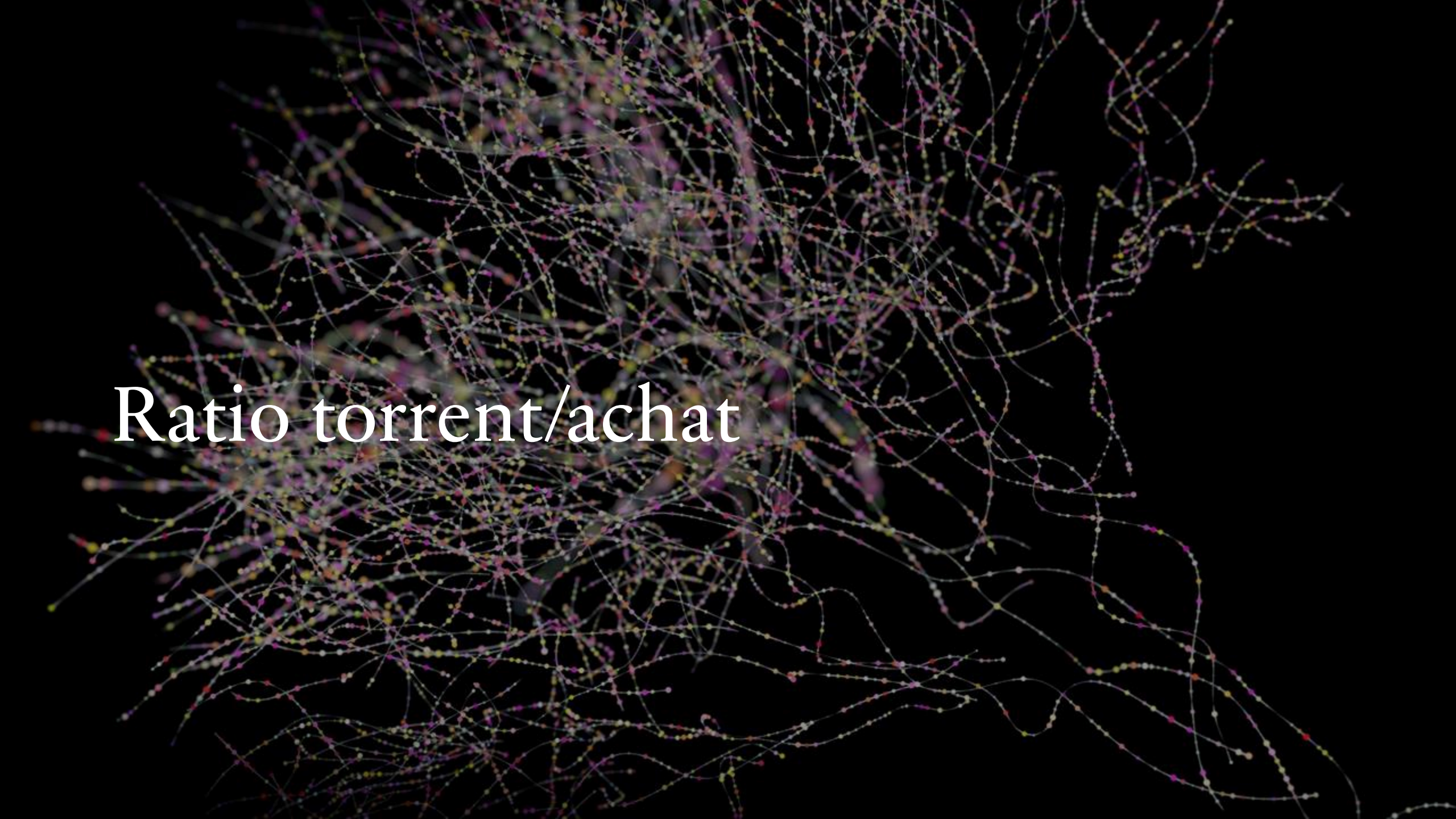
Market basket analysis

- **Résultats**

antecedent	consequent	confidence	lift	support
[1001412]	[4267]	0.7014503673008099	2.367488898919633	0.1257165234182471 3
[1000052]	[4267]	0.695210141228518	2.34642730049215	0.1103429184868105 6
[793]	[1000113]	0.689974880109614	2.4974429841373214	0.1019978259548581 8
[234]	[979]	0.6892902609802674	2.237376064958325	0.1169798326930477 8
[231]	[979]	0.6734475374732334	2.1859519663094034	0.1061703722208343 8
[606]	[1000113]	0.6722663408674404	2.4333449012003583	0.1040233338509631 3
[1001646]	[979]	0.6717144975113841	2.180326669754736	0.1284779658499368 7
[1001909]	[979]	0.6712015361359643	2.178661641872317	0.1109235640836939 8
[352]	[979]	0.6561588858779542	2.129834511204683	0.1291531351486385 2
[15]	[979]	0.6537388879205159	2.1219794089150894	0.1012888981912214 4
[1233770]	[979]	0.6530135088326983	2.119624891666004	0.1018290336301827 7

Market basket analysis

- **Résultats**
  - The Offspring => Green Day
  - Led Zeppelin => Red Hot Chili Peppers
  - Bob Dylan => The Beatles
- Confiance max: 0.7
- Support max: 0.1


Ratio torrent/achat

Ratio torrent/achat

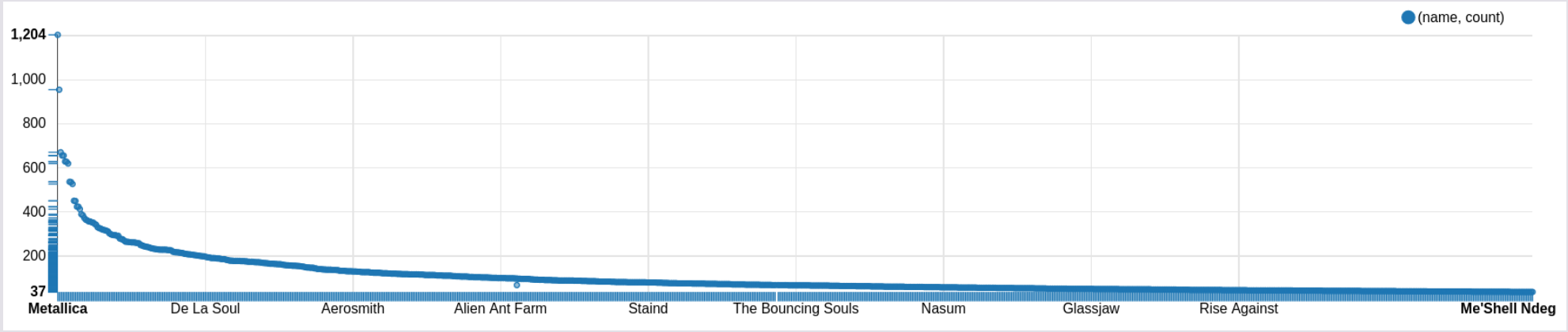
Contexte : année 2000, téléchargement de masse

Assomption : mauvaise écriture => torrent

Ratio torrent/achat

- **MisspelledArtist/OverAllArtist : 1.216%**

count	name
1204	Metallica
955	[unknown]
671	Pink Floyd
656	South Park
656	Linkin Park
629	The Beatles
628	Eminem
620	Radiohead
537	Iron Maiden
536	Nirvana



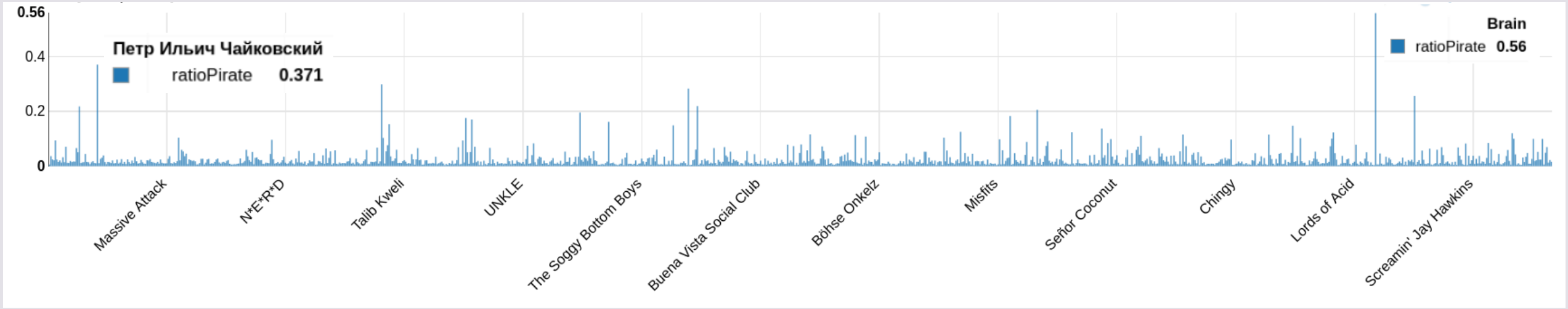
Ratio torrent/achat

artistid	artist	ratioPirate
7016732	Hopee	2.4
9946978	Christoph Poppen ...	2.0
6670976	b0x	2.0
1293381	Peter Silk	2.0
6874890	Enríquez de Valde...	2.0
7023311	The Frankie Orteg...	2.0
6854794	La famille Dembélé	2.0
6856395	Brahima Dembélé	2.0
6841333	Joten	2.0
6757930	Vortechtral	1.6666666666666667

count	name
1	Club Spice
1	Bobcat Goldthwait
1	ハナレグミ
1	David Singer
1	The Ambush
1	Steve Rhyner
1	Dumdum boys
1	Midnight Configur...
1	What Price, Wonde...
1	Emre Altuğ

Ratio torrent/achat

artistid	artist	ratioPirate
1000024	Metallica	0.03698470234072618
1034635	[unknown]	0.025337613753946567
82	Pink Floyd	0.02179987004548408
1003694	South Park	0.09449726303658888
1854	Linkin Park	0.024572969733293377
1000113	The Beatles	0.015317926113532864
930	Eminem	0.02128524945770065
979	Radiohead	0.013562584766154788
1000107	Iron Maiden	0.033005531653349726
976	Nirvana	0.014526138919753923



Conclusion

Conclusion

- **Appréhender le “big data” et “Spark”**
- **Bien connaître le dataset**
- **Extraction de features pertinentes**