Document de cadrage

DATA VISUALIZATION

Visualisation de données personnelles de la plateforme YouTube

Angéline MARC – Loric GROS – Mathys SAMBET - Thomas HUGUENEL
8 DECEMBRE 2024

Présentation du problème abordé et du besoin auquel nous répondons

Nous avons chacun l'habitude de regarder des vidéos de tout type sur la plateforme YouTube depuis des années mais est-il possible de savoir si notre temps de visionnage a augmenté ou si les contenus que nous regardons ont évolués ? Ce sont des questions auxquelles nous tentons de répondre à l'aide d'une visualisation de nos données personnelles provenant de YouTube. Cette visualisation peut nous servir à assouvir notre curiosité concernant ce sujet et comparer nos habitudes de visionnage de vidéos au cours des années.

Public cible et tâches effectuées

Notre projet s'adresse en partie à nous-mêmes car nous trouvons intéressant la possibilité d'observer à travers une visualisation nos données concernant YouTube. On peut imaginer que des personnes extérieures curieuses pourraient aussi observer la visualisation. Il est également possible que ce projet intéresse des personnes à la recherche de données pour réaliser des études et des statistiques sur un sujet similaire au nôtre.

Parmi les différentes tâches possibles grâce à l'utilisation d'une visualisation, nous en avons retenu 3 principales :

- Analyser l'évolution des contenus regardés et du temps passé sur la plateforme

Nous pouvons, à l'aide de graphiques qui montrent le nombre de vidéos regardées sur une année et ce pour plusieurs années, déterminer si nous regardons plus ou moins de vidéos qu'avant ou si le genre de vidéo majoritaire visionné a pu évoluer entre le lycée et l'université.

- Repérer des corrélations entre les vidéos et la temporalité

Cette tâche peut nous permettre de répondre à des questions larges comme le fait de se demander si notre temps passé sur YouTube est plus important le week-end qu'en semaine. Elle nous permet également de répondre à des questions plus précises telles que « Quel est le type de vidéo que je regarde quand je suis stressé, à la veille d'un examen par exemple ? » ou « Quelle chaîne ai-je l'habitude de regarder pendant mon repas ? ».

- Comparer nos habitudes d'utilisation

Il est possible de placer nos données sur une même visualisation afin de les comparer et déterminer qui a le plus de temps de visionnage, de voir si nous regardons les mêmes créateurs ou encore de repérer les périodes qui concentrent le plus de temps sur la plateforme.

Sources de données choisies

La réalisation de notre projet se base sur nos données personnelles liées à la plateforme YouTube. Nous les avons récupéré grâce à l'outil Google Takeout qui permet de télécharger une copie des données stockées sur les services Google. Cela nous a permis d'obtenir de nombreuses informations telles que nos abonnements ou tous les commentaires postés mais nous avons principalement retenu notre historique de vidéos. Il manque néanmoins des indications sur la durée ou la catégorie des vidéos qui pourraient nous être utiles pour la visualisation.

Etant donné que nous avons déjà chacun récupéré nos informations, nous ne devrions pas être confrontés à un problème imprévu mais dans le cas où cela arriverait, nous pourrions chercher d'autres données en ligne.

Travaux importants liés au projet

Il existe de nombreux projets qui traitent des données personnelles pour effectuer des statistiques ou des visualisations.

Parmi eux, on trouve « Spotify Wrapped », un rapport annuel généré par Spotify pour les utilisateurs et qui présente des statistiques sur les habitudes d'écoute musicale. On trouve notamment des informations sur les artistes et les musiques les plus écoutés, le temps d'écoute total ou le genre préféré de l'utilisateur. Il s'agit d'un outil intéressant par rapport à notre projet car il utilise l'historique pour créer des classements et des mesures esthétiques. Ceci peut être une source d'inspiration pour notre projet mais nous avons pour but d'améliorer cette solution en ajoutant des informations sur la localisation et en utilisant les données sur plusieurs années.

Il existe un autre travail de visualisation similaire permettant d'analyser un dataset contenant des données provenant des plateformes de streaming les plus célèbres (Netflix, Hulu, Prime Video et Disney+)[1]. On peut trouver plusieurs graphes qui montrent le nombre de films regardés pour chaque tranche d'âge, les langues les plus utilisées ou encore les genres les plus célèbres. C'est un projet pertinent par rapport au nôtre car il y a plusieurs graphiques sur les mêmes données qui montrent différentes choses. Néanmoins, on a ici des données sur les films et non sur des utilisateurs comme dans notre cas.

Enfin, un article de recherche publié en 2022 présente une étude qui analyse des critiques d'utilisateurs sur des visualisations de données dans des applications de santé.[2] Elle vise à identifier les types de visualisation les plus utilisées et à comprendre quels sont les problèmes rencontrés par les utilisateurs concernant ces visualisations. Cet article est intéressant car il met en évidence les pratiques les plus adaptées et des méthodes qui pourraient être utilisées.

[1] https://www.kaggle.com/code/sahilchachra/netflix-hulu-disney-prime-data-visualzation
[2] https://arxiv.org/pdf/2202.10620

Organisation

Afin de communiquer facilement entre nous, nous échangeons directement à l'oral durant les journées de cours ainsi que par l'intermédiaire d'un groupe sur Discord. Nous utilisons Github pour le code du projet et la rédaction de notre cahier d'avancement.

Nous prévoyons d'avancer sur le travail chacun de notre côté durant les prochaines semaines en faisant des points réguliers durant lesquels nous pourrons échanger sur l'avancement et les potentiels problèmes rencontrés.

En ce qui concerne notre organisation, nous contribuerons tous à la conception et au code des visualisations de manière égale mais nous avons mis en place une petite répartition des rôles principaux. Loric et Mathys seront en charge du développement D3 et du design de la visualisation. Angéline dirigera la partie du pré-traitement des données tandis que Thomas assurera la bonne tenue du cahier d'avancement.

Scan des esquisses finales

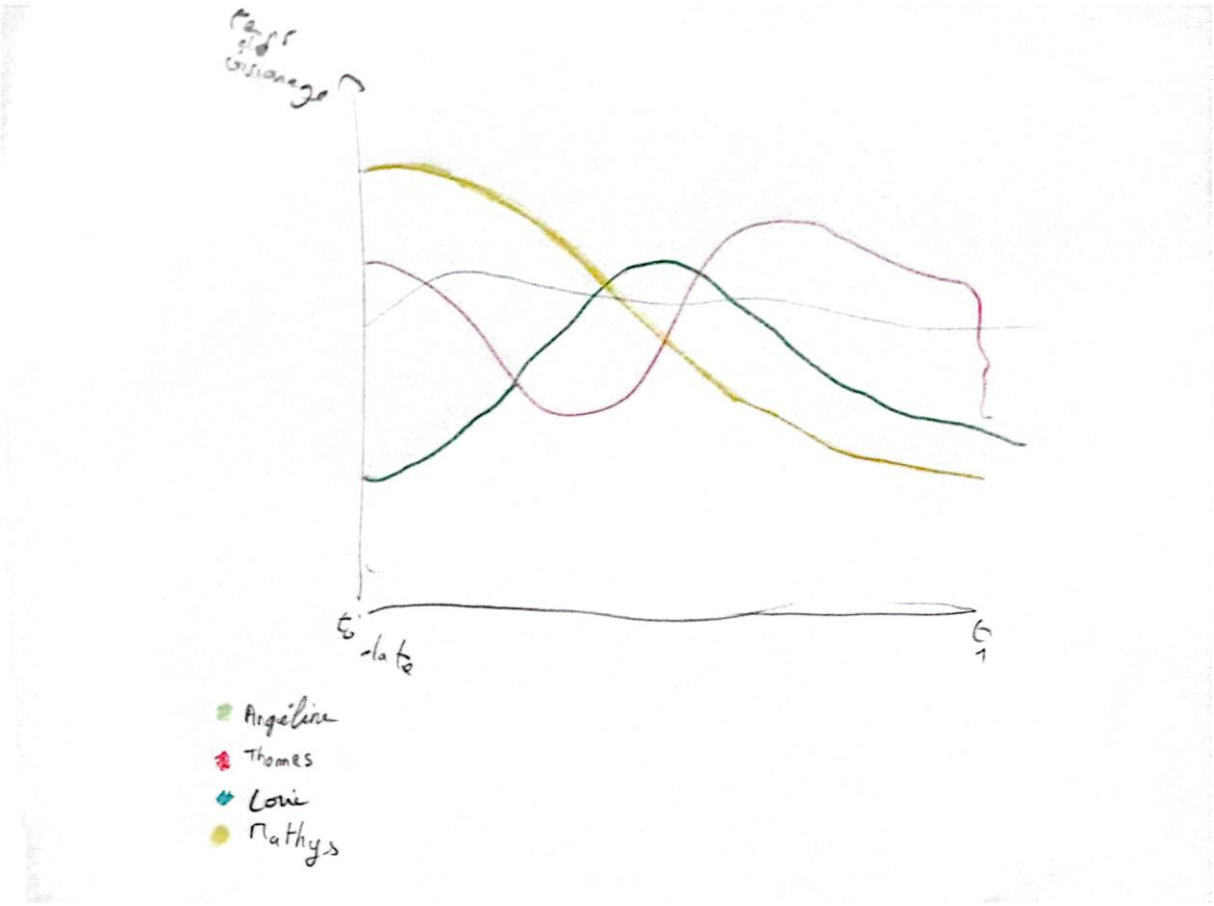
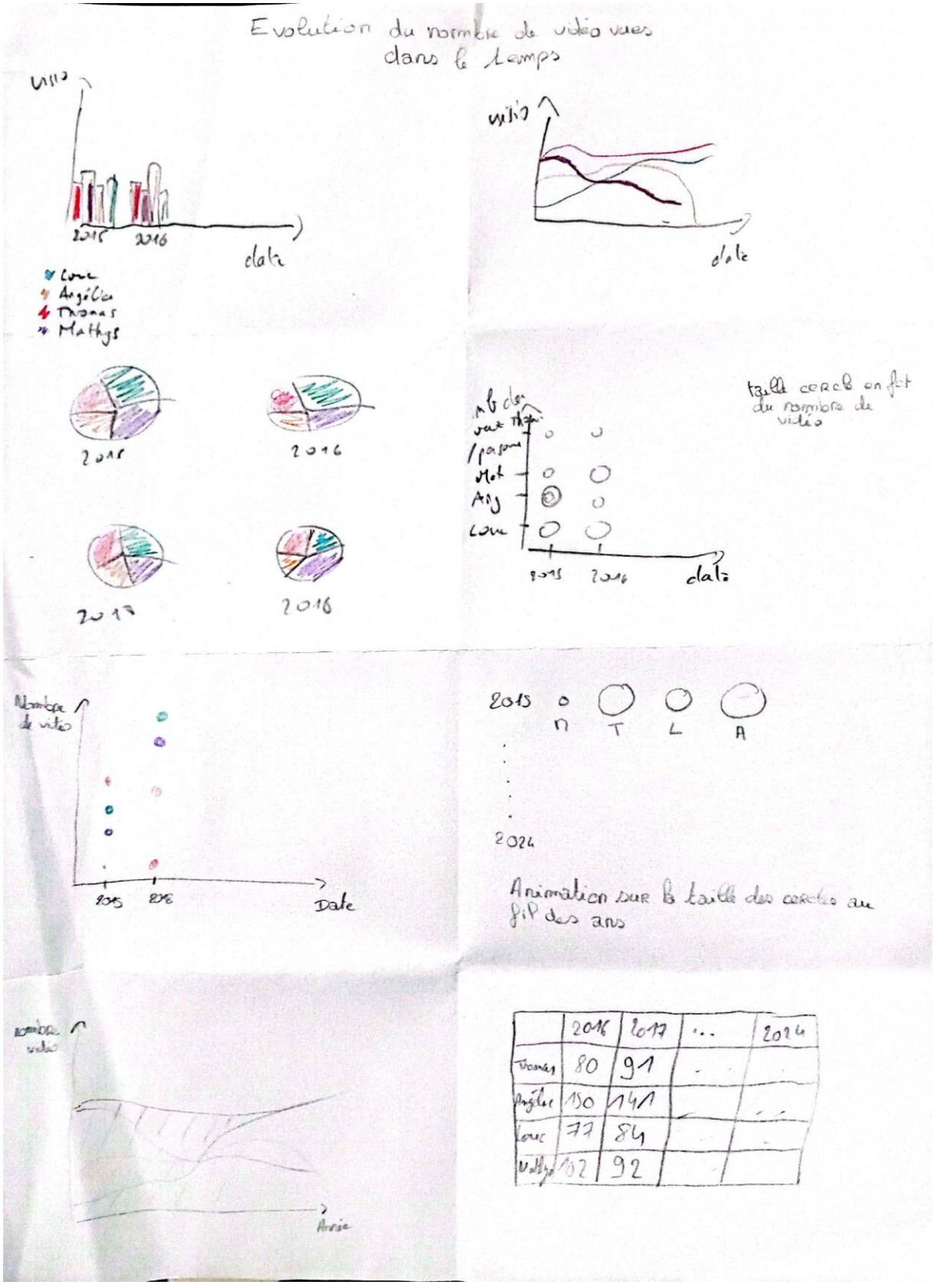


Figure 1 : Graphique montrant l'évolution du temps de visionnage en fonction du temps pour chaque personne



	2016	2017	...	2024
Thomas	80	91		
Angélie	130	141		
Love	77	84		
Matthys	82	92		

Figure 2 : Crazy 8s lié à l'évolution du nombre de vidéos vues dans le temps

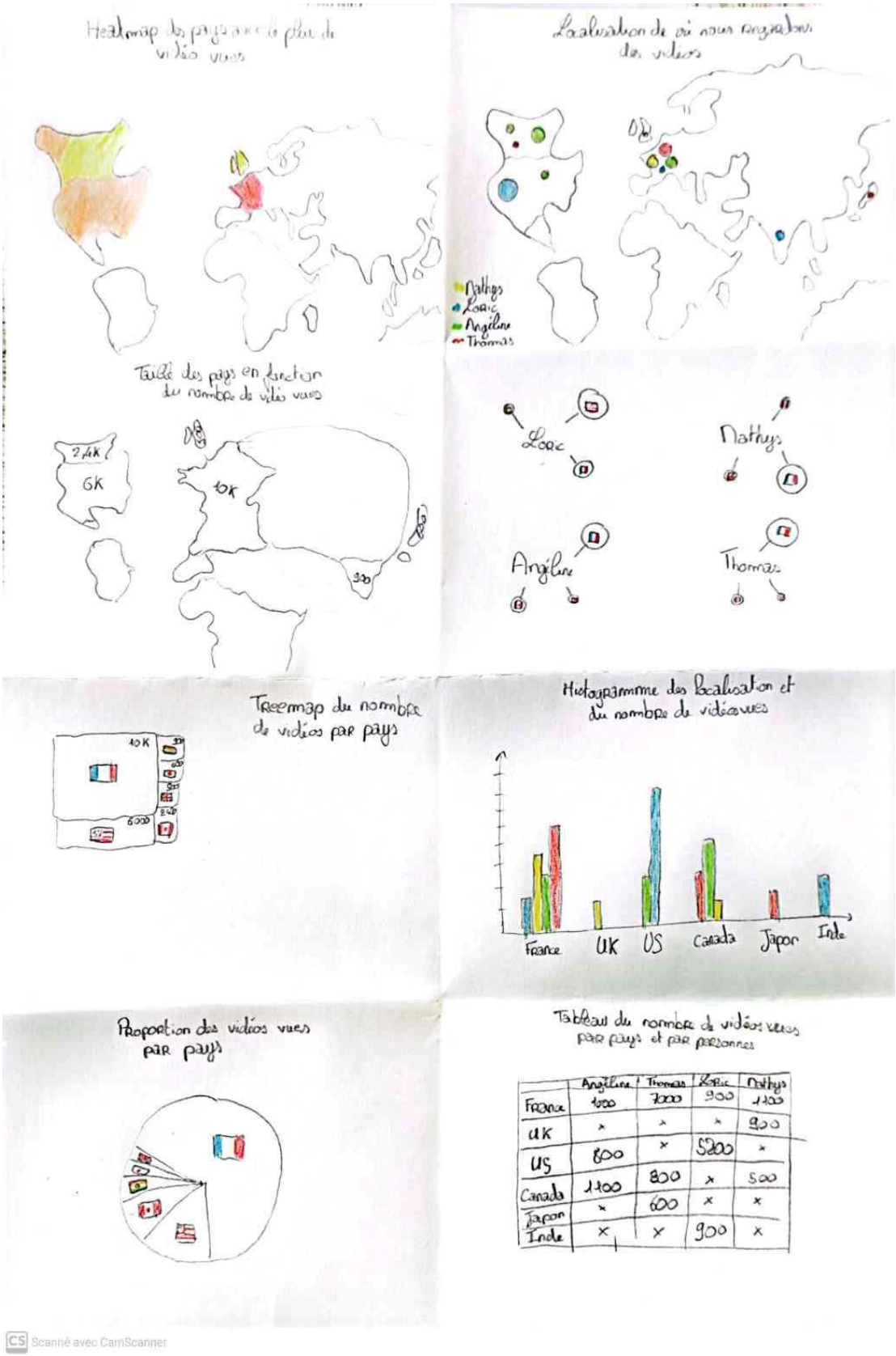


Figure 3 : Crazy 8s lié à la localisation des vidéos visionnées

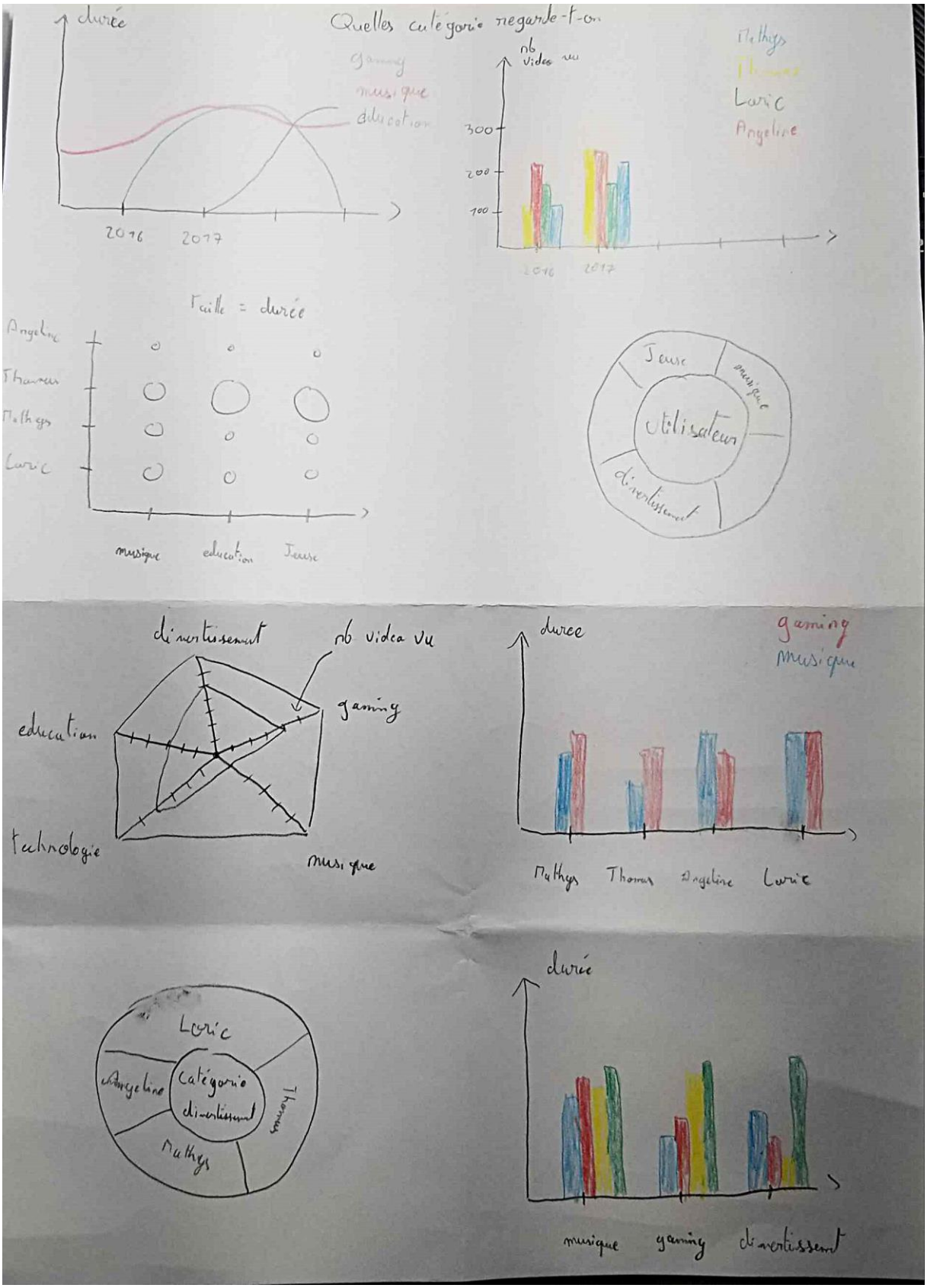


Figure 4 : Crazy 8s lié aux catégories regardées par chacun

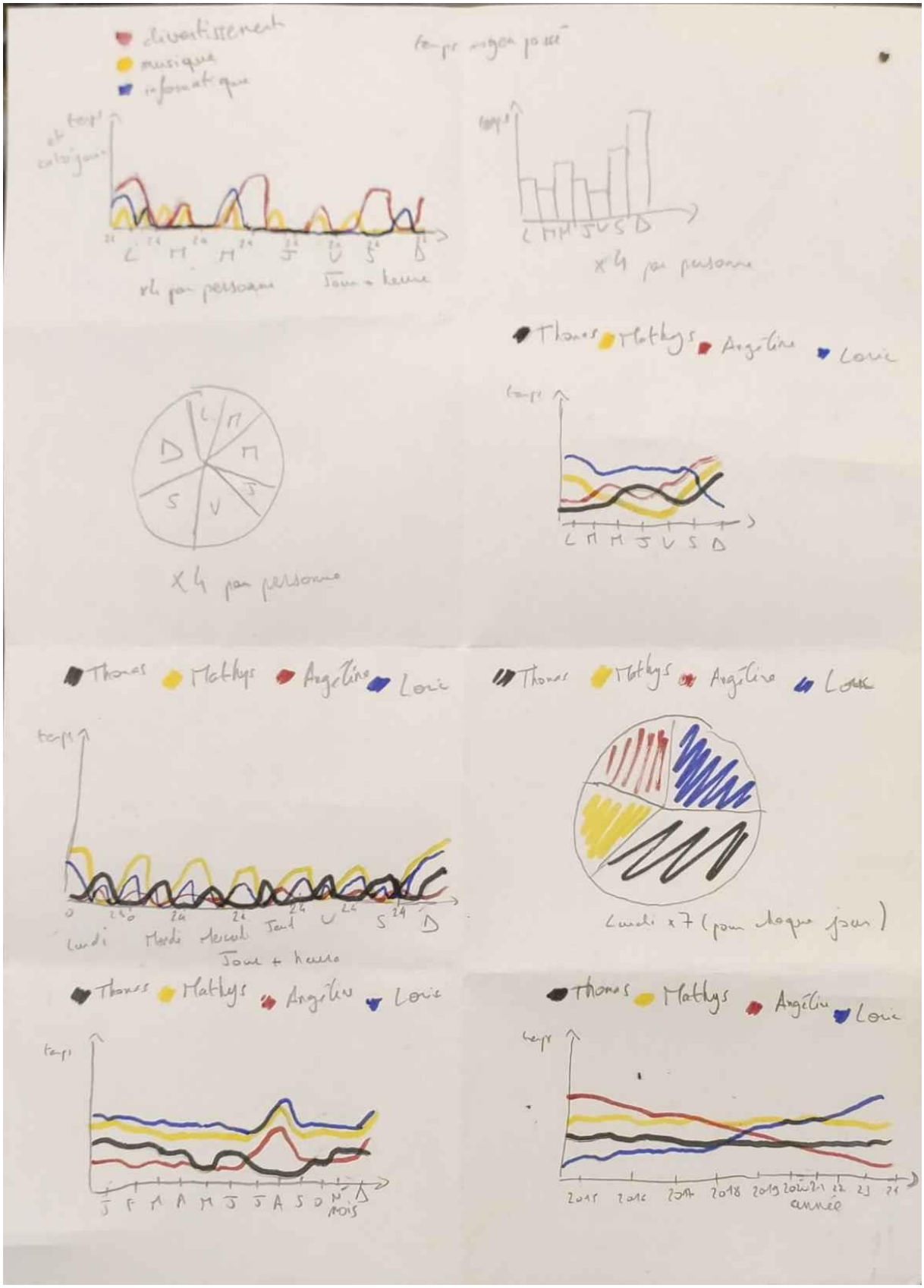


Figure 5 : Crazy 8s lié à la temporalité des vidéos visionnées