## Example 1

**First idea: The influence of the economic crisis (Great Recession in around 2008) on the cinematographic production**

During this crisis, not all countries entered recession at the same time. For example the United State was the first crisis-affected country in around december 2007. What are the influences of the economic crisis on the profits of film production? The people had less money and thus they went less to a cinema and normally the box office of the films decreased in this period. In the cinematographic world, which countries have been more impacted by the economic crisis? We can look at the box office by country and we can compare it before and during the crisis. Has the number of films produced in the countries decreased? The studio of production can have a limited budget and the films with low budget may be canceled and be postponed. And we can look at it by country to see the difference in the impact of the crisis. Does the crisis influenced the stories of the films? After the crisis, has the number of films produced, which talk about the crisis or center the money in history, increased? For example, there is the film “The Wolf of Wall Street” which was released in 2013.

**Second idea: the doublers in the countries**

All the countries don't have the same access to the cinema because many countries don't have enough doublers. For example, most films in Portugal are in French. We can look at each movie which has a version in a particular language. And we can look at the number of films by language. Is there a difference between the continents? Is there a relationship between a version in a particular language and the actors' ethnie of the original version? Is there a relationship between a version’s number of the movie and the genre of this movie?

**Third idea: The age of the actor**

At what age do actors start their careers in cinema? For example, Daniel Radcliff played a role in a telefilm in 1999, he was 10 year old. Is there a difference between women and men? On average, there can be a difference between the men and women actors. Is there a relationship between the starting age of an actor's career and the type of his first movie? Is there a relationship between the starting age of an actor's career and the country of his first movie? To know the starting age of an actor’s career, we can look at the first movie that this actor played. And with that, we can make a average of age by country, by sex, by genre of the movie.

## Example 2

**Idea 1: Can the popularity of public figures be determined?**

Can we determine which politicians are most liked based on the articles they are most connected with? For this, we could perform sentiment analysis using an NLP software on the plain text of both the page of the politician and the articles that are closest to it, using the semantic distance metric proposed in Wikispeedia: An Online Game for Inferring Semantic Distances between Concepts. Can we find whether the most appreciated and hated public figures are linked by the Wikispeedia paths to concepts that are overall positively or negatively connotated respectively? Collection of datasets from the press from different countries (like Quotebank for the US) or even the results of elections could hint at people’s opinion about public figures and could therefore be used to compare the results.

**Idea 2: What is the relation between Damon Hill and the Ku Klux Klan?**

Probably none, although they are related by only a few clicks in the small world network of Wikipedia hyperlinks. The article has highlighted that one of the main strategies of the gamers was to look quickly for hubs, here general concepts that would link many articles, to redirect themselves when the start and endpoint were not related by common sense. Then from some hub the path should narrow down to the concept looked for. It would mean that the concepts after the hubs are related according to common sense. Taking this into account, we use again the semantic distance measure and train a neural network with supervised learning to determine from where the path narrows down again. It gives us a directed network of the finished paths, pondered with the weights given by the semantic distance between them. Note that the asymmetry of the measure does not matter if the network is directed. Then we can perform community detection on this network. Do the clusters found contain notions that are inherently related? Random walks between articles could be performed and their result value for relatedness could be compared to the ones linking two articles inside a cluster. If the values for the random walks indicate less relatedness than the one of the paths inside the clusters, the random walks could give a good null model and it might be an indicator that the clusters make sense.

**Idea 3: Game over?**

What makes people give up a game? Is it because they cannot find relevant hubs, or they do not have enough background knowledge or patience? We try to answer this by looking at the unfinished paths, the time taken to do them and around the point where they stopped. If the last article on which they stopped is semantically far from the endpoint, it could be a lack of knowledge and if it is close but the gamer kept going between concepts related to the endpoint without finding it or if it played multiple times in a short period of time, it could be a lack of patience.

## Example 3

**Idea 1: Culinary evolutions**

A first project idea could be to study the attention shift on food types during the corona virus pandemic: did healthy food became more or less popular in an effort to preserve one’s health? Did the attention to Asian food or other types of food change? First, we would have to classify food-related pages into the categories “healthy” and “unhealthy”, as well as more categories describing their type and origin. We would then find the covid case numbers in another dataset and look for correlations between the page views for the types of food and the covid infections, covid deaths, and the mobility restrictions. To complete the findings, a dataset on the restaurant sales during 2020 could be used to find what types of restaurants were most popular during the crisis. This dataset can be found on Kaggle: https://www.kaggle.com/code/carriech/restaurant-rankings-top-250eda/data?select=Top250.csv

**Idea 2: Reconnection to animals**

While in lockdown, many people expressed the need to reconnect with nature. Many also adopted dogs as an excuse to go outside. The interest in pet-keeping could be studied with this dataset using the page views on animals. We could try to find correlations of covid case numbers and mobility restrictions with the attention shift for animals, and identify the types of animals of interest. We could complement this analysis with a dataset on arrivals and outcomes of animals from shelters, and the types of animals. For example, the Austin Animal shelter is the largest shelter in the United-States that does not kill any animals. They take in animals of all species and publish their intake and outcome data every year: https://data.austintexas.gov/Health-and-Community-Services/Austin-Animal-Center-Intakes/wter-evkm This study could give insights on the attachment to animals while in isolation, and more generally on the connection to nature if we also use page views on plant types.

**Idea 3: Mental health in lockdown**

We could collect page views on different mental illnesses such as depression, anxiety, ADHD during the lockdown period. It could be indeed interesting to see which pathologies gained or lost attention, and how it could relate to the evolution of the pandemic. We could then look for which diseases got diagnosed by medical professionals and treated by using a dataset on medicine consumption in 2020, which we would have to find. This study could tell us about the emotional distress and increase in mental illnesses due to the pandemic, and perhaps the tendency to self-diagnose using the internet.

## Example 4

**Idea 1: Youtube and the rise of consumerism**

Youtube has been recently used to promote and advertise for new products through youtubers and influencers. This is done by either direct advertisement, campaigns or by doing reviews on new products. We can explore this from three different aspects: the trend of videos on product reviews in the period 2005-2019, comparing the number of views, interactions and video production through the years and how it was affected by the 2008 economic crisis. The correlation between this kind of videos and different monetization techniques (eg: marketing for other products in the description). We can study users who interact with these kinds of videos and their behavior and whether they tend to interact oftenly with these types of videos or not.

**Idea 2: “Don’t watch this video!”**

What influences the popularity and views of youtube videos? There are many popular techniques that have been used on youtube to attract more views and interactions. In my proposal we explore three main techniques: Reverse psychology: Videos with descriptions and captions that have statements similar to “Don’t watch this video” as well as videos that have more dislikes and heated comments tend to attract more views and interactions. We can explore this by showing if there’s a correlation between using reverse psychology techniques and the number of views on those videos as well as the channel subscription and video production over time. Mass Intimacy: another way of attracting more views on youtube is by incorporating more personal details about their lives in their content like shooting from their bedrooms, shooting with their families or discussing personal issues. We can explore this by digging into the description of the videos and see the correlation between the rising popularity of these channels by the time they started using the mass intimacy approach. Short-form videos for shorter attention spans: we can explore the effectiveness of this technique by comparing the views and interactions with the videos’ from the same type and with similar content but with different length. We might prove that peoples’ attention span is getting shorter over time as people tend to prefer shorter videos rather than long ones’.

**Idea 3: The rise of the prosumer market**

Prosumerism is the involvement of consumers in the production process. It has always been approached from the perspective of industry and consumption. Youtube has allowed prosumerism to enter media production. The topics and content of youtube videos are highly affected by the viewers opinions and interactions. We can study how the views and interactions on certain topics direct and affect channels content over time. We can explore this for different categories of videos, for example: we can study if views and interactions have affected the popularity of certain music genres over time while other genres that get less interactions become less popular.

## Example 5

**Idea 1: A Data trip to Satellite**

Budgeting is an essential skill for poor graduate students, particularly concerning beer and parties. A common approach to facilitate this type of task (often called the consumer’s constrained choice in microeconomics) is to estimate cost-efficiency ratios in order to reduce the choice problem to an optimization problem. I thus propose ranking the beers sold on the EPFL’s campus (at EPFL's bar, SAT) based on their cost and ratings extracted from SAT's website and from the proposed dataset. I extend this proposal by conditioning the ranking on categorical factors such as origin and type whenever possible. This would enable the construction of rankings and suggestions tailored to common groups, e.g., the ‘Belgian-beer lover,’ the ‘Malt-beer maniac,’ or the ‘National beer explorer.’

**Idea 2: A taxonomy of beers**

Choosing beers from a bar’s menu can be challenging due to the extensive range of choices typically available. For an inexperienced beer drinker, this task is further complicated. Indeed, to learn one’s own tastes, one must explore the ‘beer space,’ often with limited time and economic resources. This is a typical example of the exploration-exploitation trade-off, a common problem in reinforcement learning. I address this problem by noticing that the textual reviews in the proposed dataset have a rich diversity of qualifiers and descriptors, which allow the classification of beers based on their shared review terms (for example, “sweet” beers, “strong” beers, and “corn” beers). I propose to unveil this implicit ‘taxonomy of beers’ by using clustering and NLP techniques on the textual reviews available. I expect that, from the point of view of the unexperienced beer drinker, it would be easier to choose a beer based on color, flavor, and strength than based on type or provenance, in the same way that it is often easier to choose a movie based on synopsis and genre, rather than on technical attributes. Moreover, contrary to the usual beer classification schemes based on fabrication and origin, my proposed ‘taxonomy of beers’ would be based on the human perception of different beverages and, hopefully, constitute a good base for beer recommender systems.

**Idea 3: Biases in Enthusiastic Evaluators Ratings (BEER)**

In the proposed dataset, one can notice that the distribution of reviews by user is highly skewed: the median reviewer gives, in general, only two ratings in total in booth BeerAdvocate and RateBeer, while the 1% most prolific reviewers write at least 800 and 2300 reviews (in BeerAdvocate and Ratebeer, respectively). One can imagine that this distributional form would also be present in review systems of other goods (like video games, food, or movies), in which the consumer population can be separated into a large group of infrequent consumers and a small group of enthusiasts. I suggest using the beer rating (both on multi-aspect parameters or Overall grade) conditioned on the number of reviews by user to identify if (and how) reviewers with high throughput have a preferred brewery, alcohol content, or type of beer that is significantly different from the group of all users.

**Idea 1: YouTube videos' journey: From fresh upload to viral sensation**

Ever wonder why some YouTube videos blow up while others flop? Let's dig into this! We'll look at three things: early signs of a hit, how different types of videos typically perform over time, and how viewer engagement affects a video's spread. We can analyze the correlation between a video's performance in the first 48 hours and its long-term success. By categorizing videos and applying clustering techniques to their time series data, we can identify typical lifecycle patterns for different content types. We can also study how comment activity and sharing behavior influence a video's spread. We'll use the video details from yt\_metadata\_en.jsonl, track performance over time with df\_timeseries\_en.tsv, and look at engagement using youtube\_comments.tsv. This could really help creators figure out how to make their videos take off and might even give us insights into how YouTube's recommendation algorithm works!

**Idea 2: Jack of all trades or master of one? YouTube channel strategies**

Should YouTube creators stick to one topic or mix it up? Let's find out! We can investigate this by analyzing three key aspects: the impact of content diversity on channel growth and engagement, long-term performance comparison between specialized and diversified channels, and common characteristics of successful diversification strategies. Using the yt\_metadata\_en.jsonl file, we can calculate a content diversity index for each channel based on video tags and categories. We can then correlate this index with subscriber growth rates and average view counts from df\_timeseries\_en.tsv. By comparing specialized and diversified channels' long-term growth rates, audience loyalty, and potential revenue, we can identify the pros and cons of each strategy. Case studies of channels that successfully diversified their content could reveal key strategies, such as content relatedness and audience overlap. This could help creators decide whether to branch out or double down on their niche!

**Idea 3: How long should your YouTube video be?**

Is longer better? Or do short and sweet videos win? We'll investigate the best video length for different types of content on YouTube. We can explore this by investigating three aspects: optimal video length for different categories, the impact of video length on viewer engagement and channel growth, and the evolution of audience preferences for video length over time. Using the duration information from yt\_metadata\_en.jsonl and performance data from df\_timeseries\_en.tsv, we can analyze the relationship between video length and metrics like view count and likes for each content category. We'll also look at how comment numbers and patterns (from youtube\_comments.tsv) change with video length. By tracking changes in a channel's average video length against subscriber growth rate, we can assess the impact of length adjustments on channel performance. Finally, we'll analyze long-term trends in video length preferences. Are we witnessing a shift towards shorter content with the rise of platforms like TikTok? Or is there a resurgence of longer, more in-depth videos? This could help creators figure out the perfect length for their videos to keep viewers hooked and grow their channel!