History of Cinema

A Process Book by
Simon Maulini, Sebastien Chevaleyre and Giacomo Alliata

Introduction:

History of Cinema is an interactive visual experience that aims at exploring the cinema's evolution over the years, from three main axes: the people who worked in the industry, the most successful movies and the genres of films. Specifically, we focus on the actors and the film directors and show who has worked with who, as well as on the genres of movies, both to highlight the most popular ones over time as well as the preferred genres of actors and film directors.

The intended audience is quite large, since no prior knowledge of cinema is really needed to understand or appreciate our data story, but it is probably true that recognizing some actors or movies can spark more interest in the user.

In this process book, we quickly present our project, spending a few words on the dataset and on the website hosting our visualisations, and then present the path we have taken to create it, focusing on the design choices we have made and difficulties we have encountered. Finally, a peer assessment presents who has worked on which part of this project.

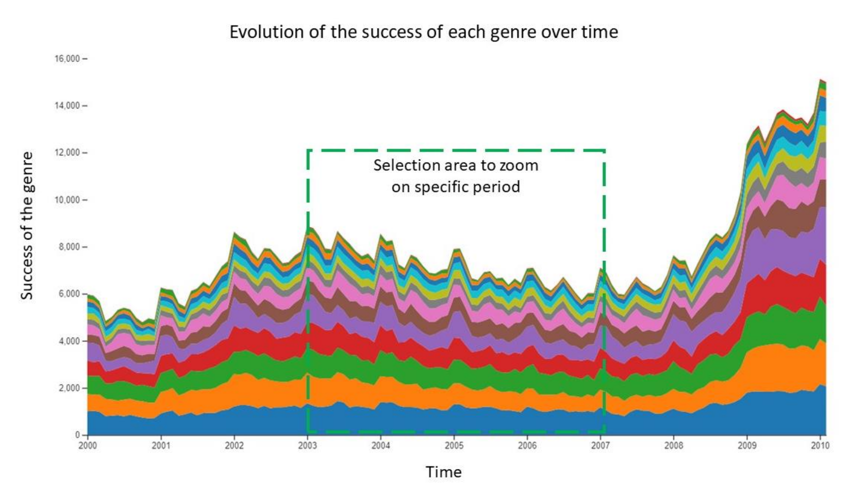
The Dataset:

To build such a visualisation, the first step is of course to find some data. We spend some time first discussing whether it would be more interesting to look at some specific movies, perhaps even considering the scripts and using NLP to extract some meaningful information, or to look at the bigger picture, with a global view of cinema. We have chosen the latter approach and we have specifically chosen this historical aspect as a general lens through which to present the data. One could say that the general question we want to answer thanks to this visualisation is how cinema has evolved through time, and we have found a dataset that would help us create a visualisation project to address such a problem. The dataset can be found on Kaggle at the following link and it is a subset of 45000 movies of the Full MovieLens Dataset. It corresponds to an ensemble of data collected from TMDB and GroupLens. This dataset has many information for each movie, so we definitely had to make a choice on what we wanted to focus on. The challenge was to have a structured and interesting yet comprehensive story of cinema to present. In the end we have chosen to focus our work on two main axes: the evolution of genres over time and the people who have worked in the industry. This of course meant that we had to leave out some interesting aspects that future works could work on, such as ratings or tags used on the movies. In particular, an aspect that we wanted to look at but later decided to skip is the question of gender in cinema. At first, we wanted to explore whether cinema was more or less "sexist" over time, but unfortunately this dataset did not provide enough information that could be used as a measure of sexism in cinema.

The Visualisations:

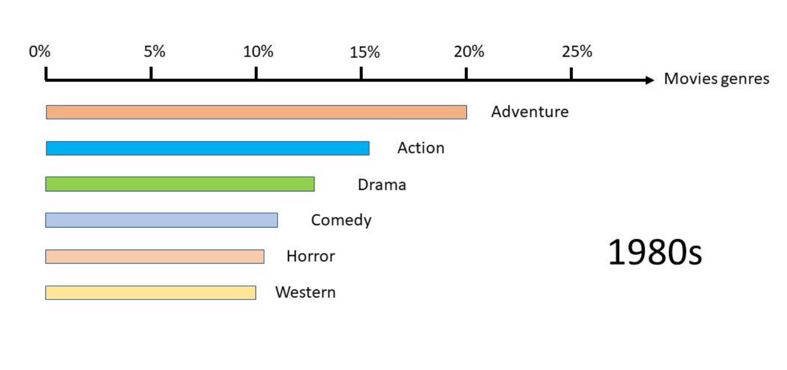
Once we had chosen which axis focus our work on, we had to find a meaningful and interesting way to show the data. An important point for us was to let the user have some control over the visualisation, by selecting a time span for instance rather than always looking at the entire history of cinema. With the data at hand, we started exploring which visualisations would best help us present it.

The first axis we decided to focus on is the genres of movies, and in particular their evolution over time. For this, we wanted to have two complementary views: a dynamic one and a static one. Thus, users could look at the big picture at a glance with the static view or follow the evolution year by year with the dynamic one. Several possibilities offered to us, but we ended up choosing a stacked area chart for the static view and a bar chart race for the dynamic one. The following sketch presents our initial idea for the stacked area chart, where each colour would correspond to a genre. The x-axis is simply the time dimension, year by year, and the y-axis measures the success of a given genre. Here as well, different measures for success could be designed but we have simply taken the number of films that are assigned to that genre. Initially we wanted to measure the success by grossing, but as in our database a lot of movies were lacking this data, we chose to only consider the sheer number of movies.



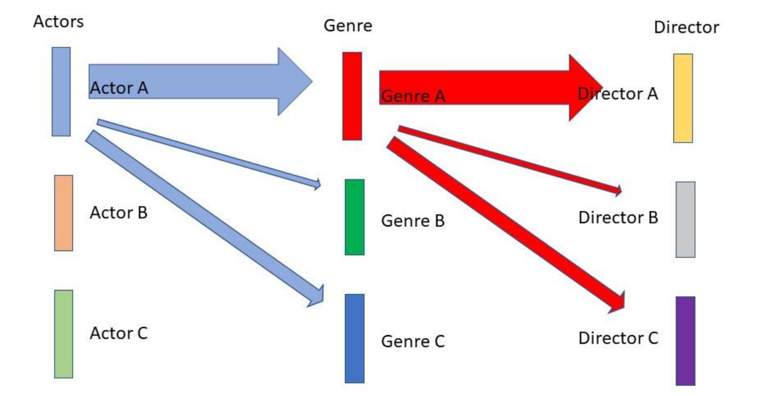
It is important to note that a movie can be assigned to multiple genres, meaning that the sum over all values at a specific time is bigger than the number of films produced that year. Nonetheless, this still quickly shows which genres were more popular at a given time, even though it means it is not possible to normalize by the number of films produced. Users could also have the possibility to select a time span and zoom in on it, since the overall period of time considered is quite large and thus yearly variations could appear smaller than what they actually are.

For the dynamic view, the following sketch gives the idea of the bar chart race. Here we follow year by year the number of films produced, cumulatively (meaning that at time x we show all films produced up to time x). Since there are many different genres, only the top 12 are shown, with new entries and new exits each year. This allows us to have a visualisation complete enough without being cluttered.



Bar race of genre across time by Count/Grossing

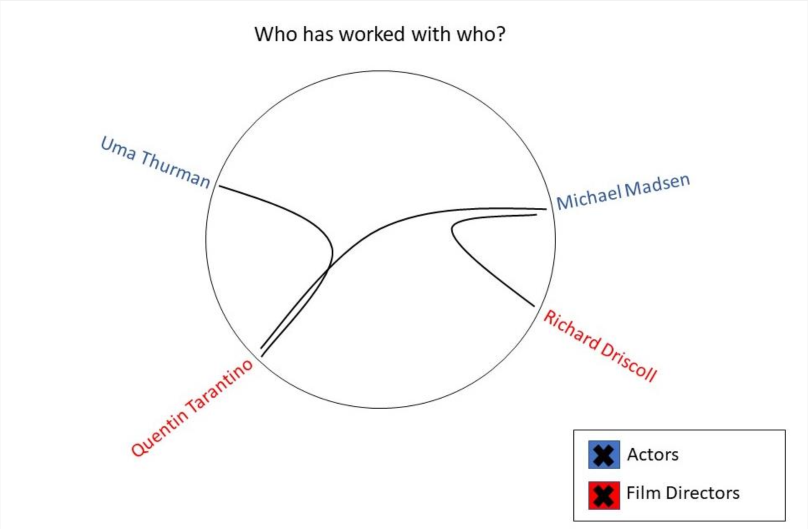
Finally, we also wanted a visualisation to bridge to the second axis: the people. To do so, we have selected a Sankey diagram with three levels: the actors on the left, the film directors on the right and the genres in the centre, as shown in the following sketch.



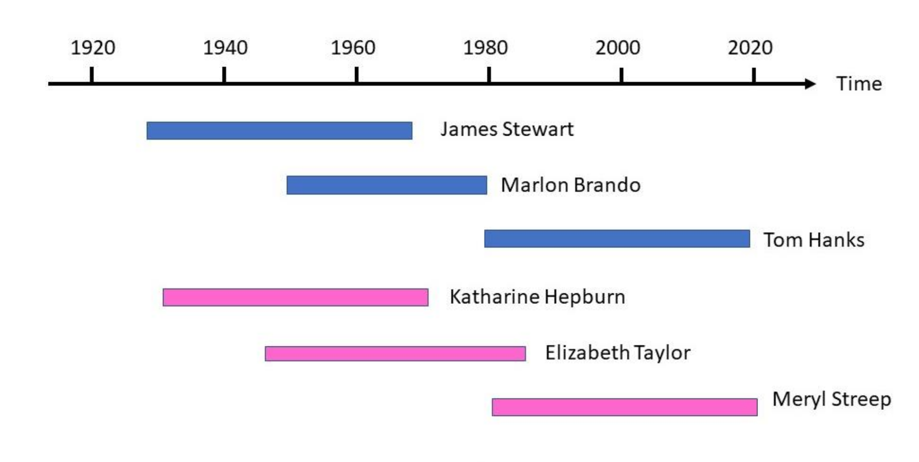
Sankey Diagram of Main Actors (option for to Genre) to Directors for most prolific ones.

Here again of course not all people can be shown, so we have selected the 15 most prolific people in each category. A connection is made between each people and the genre corresponding to a movie they have worked on.

For the second axis, the people, we immediately thought of using graphs to show who has worked in the industry, and with whom. Of course, the industry of cinema is huge, especially when considered over such a period, and therefore a selection had to be made. We decided to focus on two main categories: the actors and the film directors. Even though all people working on a movie are important, it is quite understandable to consider those two categories as the prominent ones, and this would also ensure to have some familiar names that, even without being a huge fan of cinema, users could already know and recognize. The first idea we had was to use a circular graph to connect actors and film directors that have worked with each other. A colour code could be used to easily recognize both categories. Such a representation can be quite interesting and can be used to find out if some actors have worked together on numerous movies or if there are some specific pairs of actors and directors. One could for instance think of famous collaborations such as Tim Burton with Johnny Depp or Martin Scorsese with Robert De Niro. An interesting challenge to build such a visualisation is of course to make it readable for the user, as it can quickly get cluttered. With thousands of actors in our dataset, we had to select only a few of them. Many criteria can of course be used here but we have decided to remain simple and simply take the 50 actors and film directors with the largest number of movies. This also ensures that we indeed have many connections, since having people that have only worked with one or two other people might not be that interesting in the end. The following sketch summarises our initial idea for this visualisation: Who has Worked with Who.



We also wanted to show who were the most prolific actor and actress of all times. This was challenging as we wanted to show the rise and fall of many of those actors while keeping the visualization as clear as possible. We thus decided to filter the actors by only keeping the ones that have worked on movies of one of the four principal genres (action, drama, comedy and thriller). Because there would still be way too many, we decided to include only those whose timeline would exceed 10 years. The user can sort the visualisation either by time or by genre. We went from distinguishing male/female to distinguishing movie genre in our draft so that we could offer a more diverse and meaningful representation of the succession of actors and actress.



Career Span of top actors and actress (activity/grossing), sorted by genre or date of rise to fame.

In the end, we also included an actors' bar race, to offer a more dynamic view of the evolution of the most prolific actors over time. To reduce the overall amount of information displayed, we have split the actors by main genre of their movies, thus offering four races.

The Overall Design:

Once the individual visualisations had been chosen, we had to find a way to present them together. This meant starting to discuss the overall design of the website. Since we already had many colours in our visualisations, especially the ones with genres, we have settled on a neutral colour scheme for the rest of the website, inspiring us from old black and white movies. Here are some images that we used as inspiration for the colour scheme for instance.

Peer Assessment:

For this project, we have discussed together on which direction to take in the beginning and for the main design choices later. More specifically, Simon and Sebastien have handled the technical aspect of the website while Giacomo has worked more on the process book and has found inspiration material for the overall design. The three of us have all worked on the implementation of some elements of the visualisations and on the pre-processing of the data.