SAPIENZA
UNIVERSITÀ DI ROMA

MoVis

Presentation of the final project

Visual Analytics

A.Y. 2021/2022

Giovanni **Pecorelli**,
Jacopo **Rossi**,
Giacomo **Venneri**

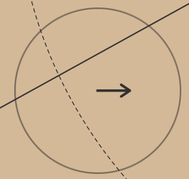
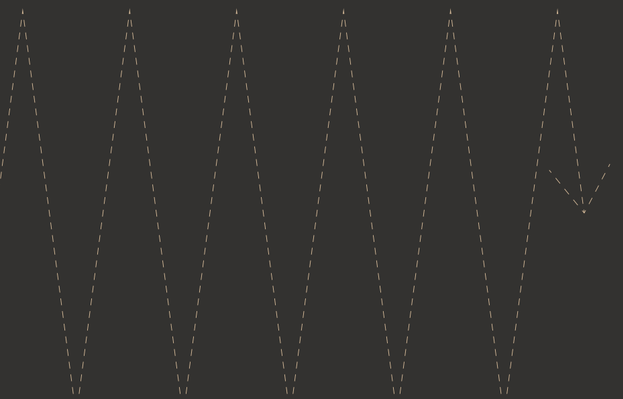


Table of contents

Introduction

The **movie industry** put together
a very extensive catalogue in the last century.
The whole collection of movies can be found on the **IMDB.com**
website.

MoVis is a tool to **explore visually** the IMDB dataset,
and more generally the whole history of the cinema industry
production.
The final tool is an **interactive visualization** built in **D3.js**
where different parts work together to give the user an
accurate **overview** of the whole dataset, which can be used to
highlight patterns, find **outliers** and more.



Objectives

This tool is **directed** to **movies enthusiasts**, and can be used in different ways:

To **discover** new movies based on personal preferences.
It is possible to apply filters to different movies' characteristics, and then explore the dataset to find a new movie to watch;

To **find** movies similar to the ones you like: a user can find his favourite films and thanks to a dimensionality reduction visualization ideally it will be able to find out similar ones;

To **explore** the current state and the history of the movies industry as a whole, and find trends, insights, statistics and curiosities.

Related works

[1] Max Wasserman, Satyam Mukherjee, Konner Scott, Xiao Han T. Zeng, Filippo Radicchi and Luís A. N. Amaral.
Correlations Between User Voting Data, Budget, and BoxOffice for Films in the Internet Movie Database (2015)

[2] Canet, F., Valero, M.A. & Codina, L.
Quantitative approaches for evaluating the influence of films using the IMDb database (2016)

[3] Laia Blasco-Soplon, Josep Grau-Valldosera, Julia Minguillon
Visualization of Enrollment Data using Chord Diagrams (2015)

[4] Ben Shneiderman
The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations (1996)

Dataset

The dataset used was created by joining and manipulating **Kaggle Movie Dataset**, **MovieLens 25M Dataset** and from a web scraping of **IMDb.com**.

- **Actual budget and Actual revenue:** calculated from budget and revenue keeping into account the **inflation**;
- **Genres:** *Documentary, Western, Musical, Animation, Adventure, Fantasy, Horror, Sci-Fi, Comedy, Drama, Thriller, Action;*
- **Connections:** web scraping of the *imdb.com/title/imdbID/movieconnections* page of every movie.

Dataset

Imdb_id: 0076759

Title: Star Wars

Year: 1977

Budget: 11.0

Actual_budget: 50.6

Revenue: 775.4

Actual_revenue: 3566.84

Runtime: 121

Vote_average: 8.1

Vote_count: 6778

Popularity: 42.149697

Movielens_id: 260

Tmdb_id: 11

Genres: Action|Adventure|Sci-Fi

Director: George Lucas

In_connections: 7037

Out_connections: 28

Tot_connections: 7065

Connected_movies:

0080684|0086190|0120915|0121765
0121766|2488496|2527336|2527338
2199603|1490713|0076759

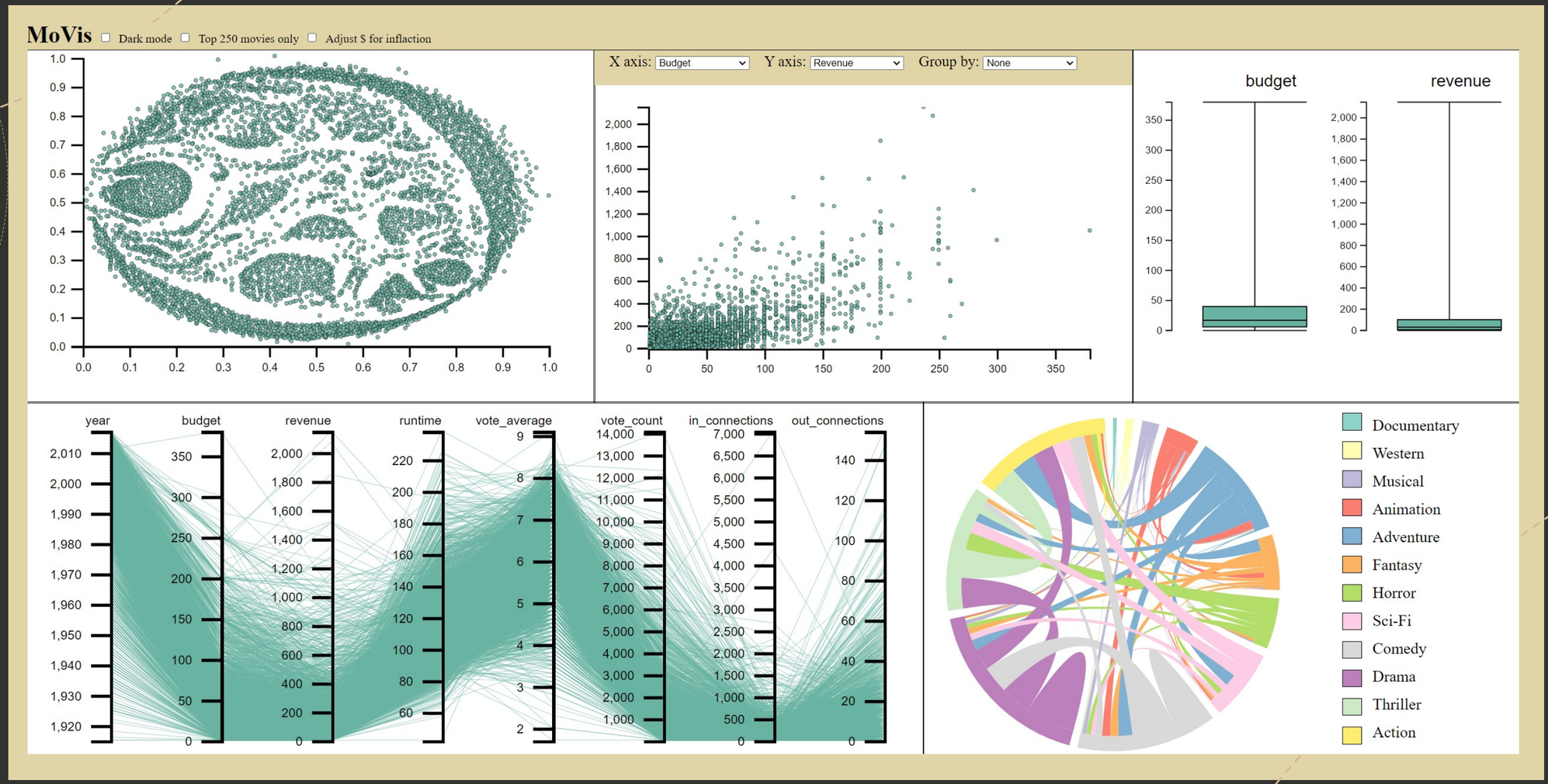
Dimensionality reduction

Done as a preprocessing phase where we compute a **dissimilarity matrix** between all the movies.

MDS is our dimensionality reduction technique because it allows to define and use **different distance functions** for different types of attributes, such as Jaccard distance for categorical data and euclidean distance for numerical data.

All of the matrices are then summed into one final matrix, with **different weights** to give more importance to certain attributes, and to cancel the differences between the attribute domains.

Visualization

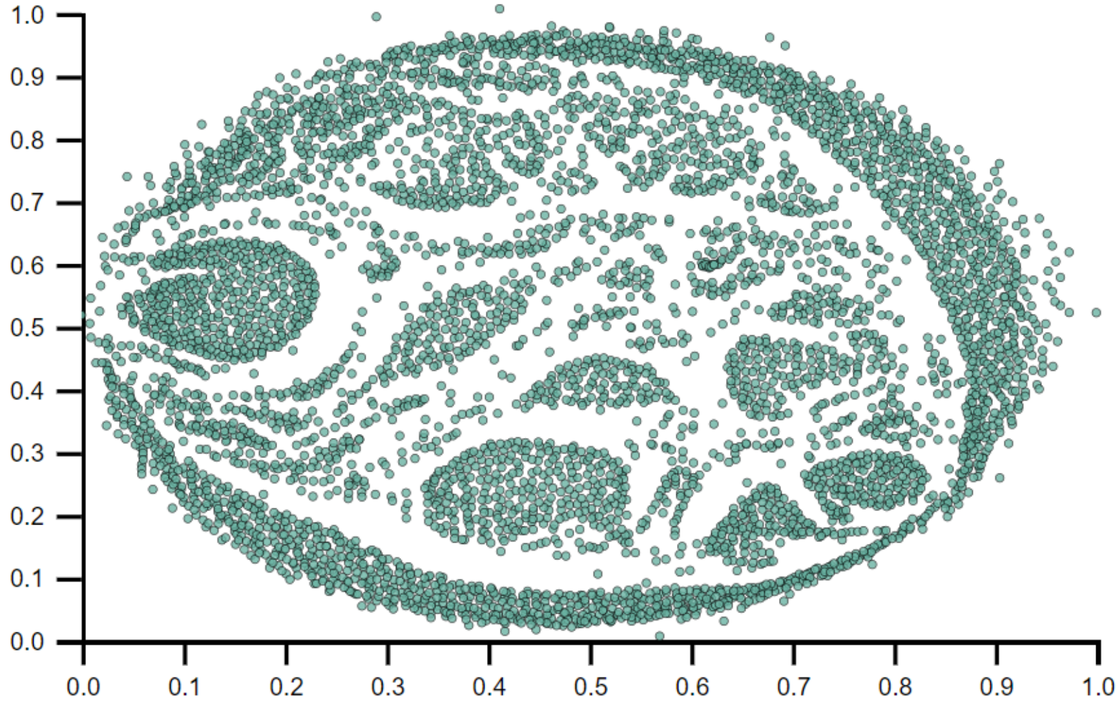


Scatterplot

In the scatterplot we visualize the results of the **MDS**.

Every movie is a circle.
The closeness of two movies indicates their **similarity**.

When the user **hovers** on a specific circle a tooltip appears.
The chart can also be **zoomed** and **moved** with the mouse.

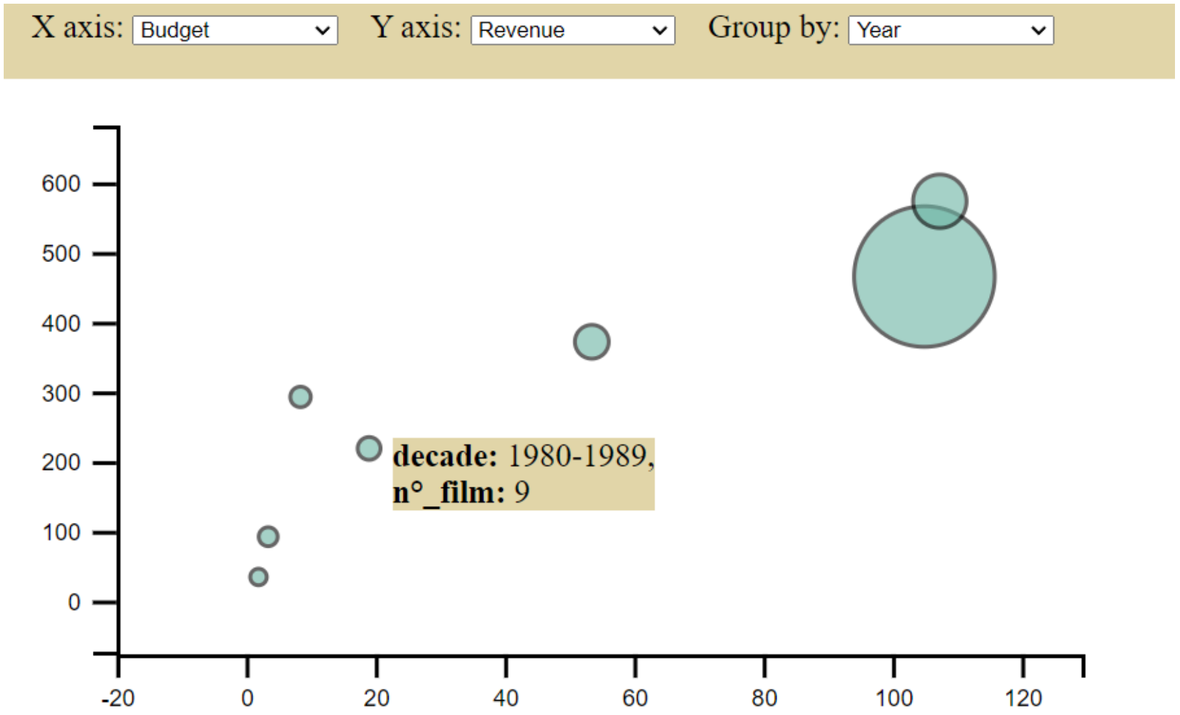


Bubbleplot

In the bubbleplot we can visualize nearly every attribute.

It’s possible to choose the attribute to use on both the **X** and **Y axes**, as well as the attribute used to **aggregate** the data.

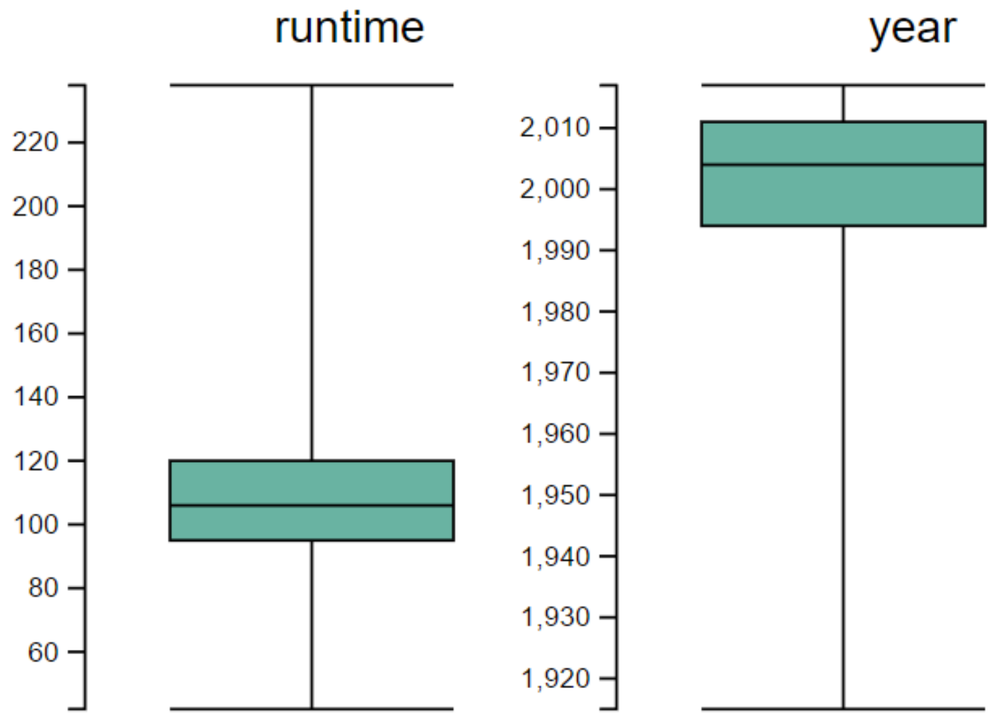
In this chart too it is possible for the user to **hover** on the circles and **click** them.



Boxplots

The two boxplots are used to show the **distribution** of the two features momentarily selected on the **X** and **Y** **axes**.

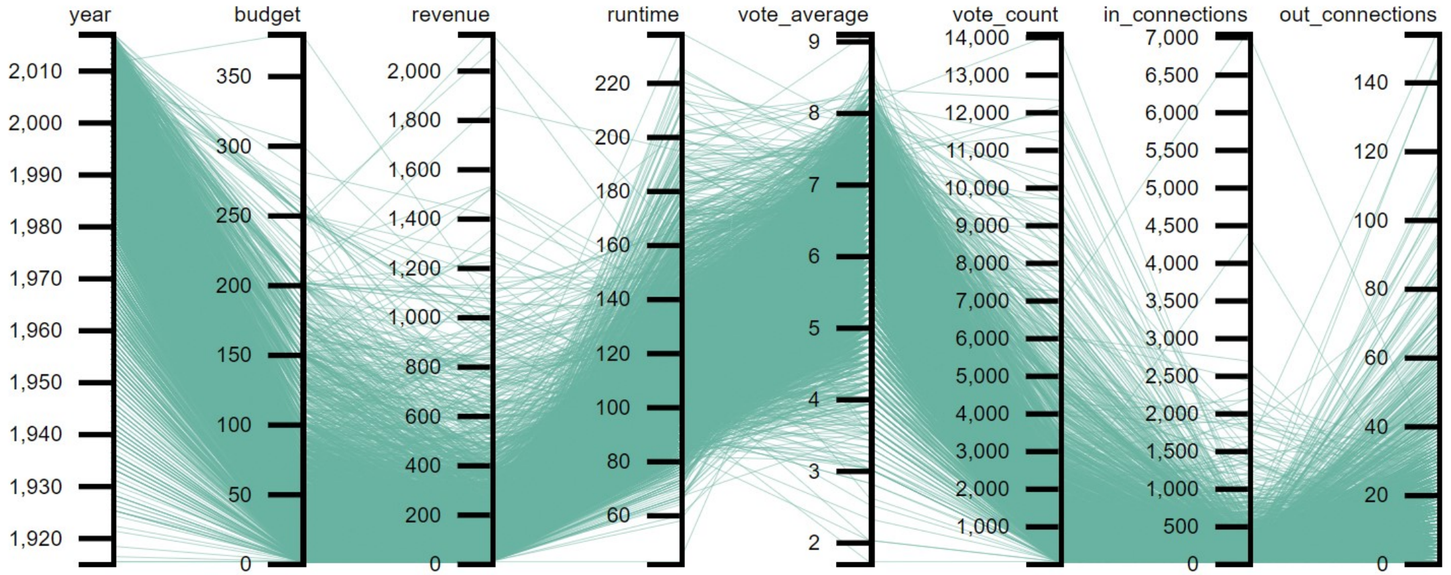
They show a five-number summary of the data: the minimum, the maximum, the sample median, and the first and third quartiles.



Parallel coordinates

To represent our high-dimensional dataset we used a parallel coordinates visualization.
It shows all the **8** continuous (not categorical) **features**.

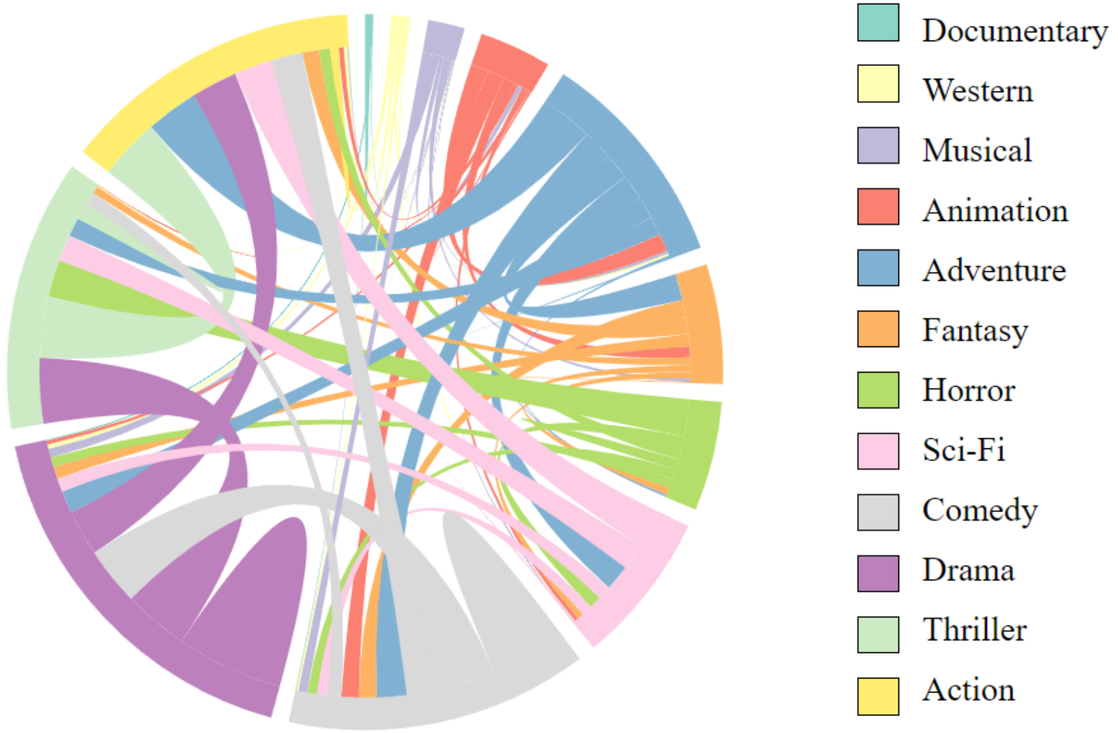
The user can interact with this visualization by **brushing** one or more columns at the same time.



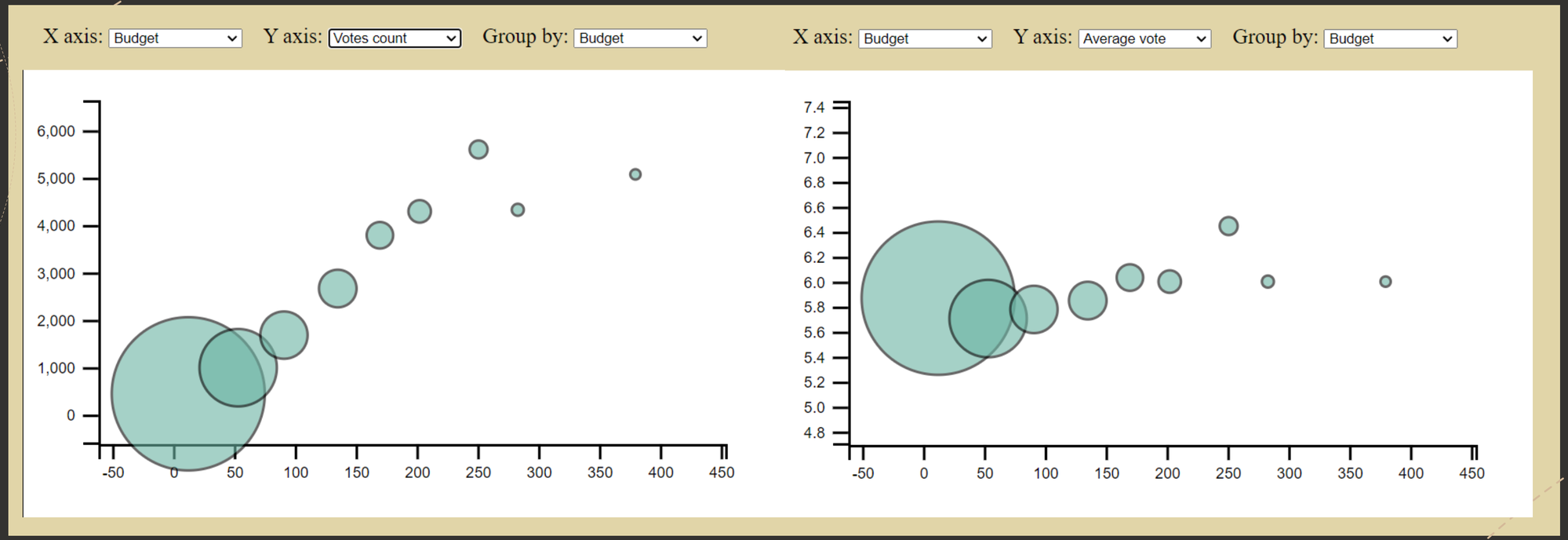
Chord

A chord diagram is a way of displaying the inter-relationships between data. In our case it is used to show the movies' **genres** and their occurrences.

It can be interacted with by **clicking** on one or more genres in the legend or **clicking** on a single arc.
Also, arcs can be **hovered** to reveal additional info.



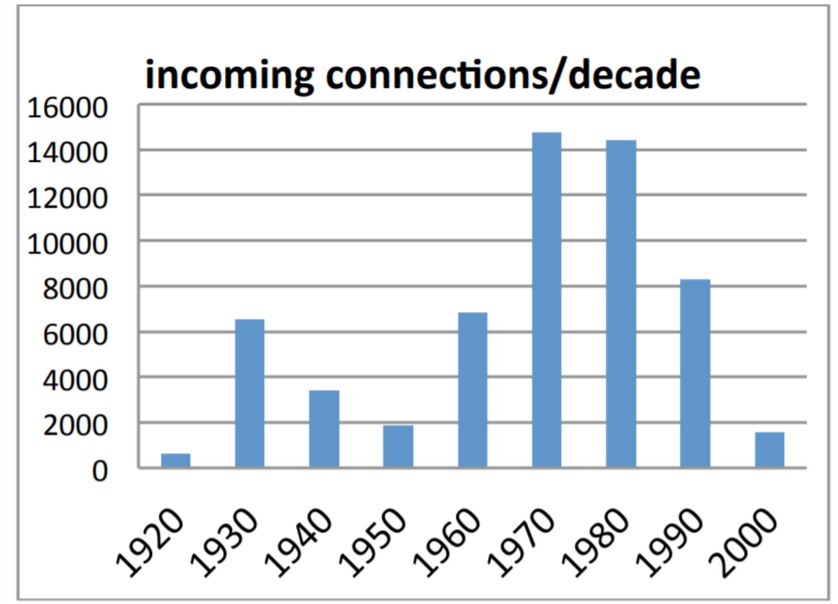
Insights

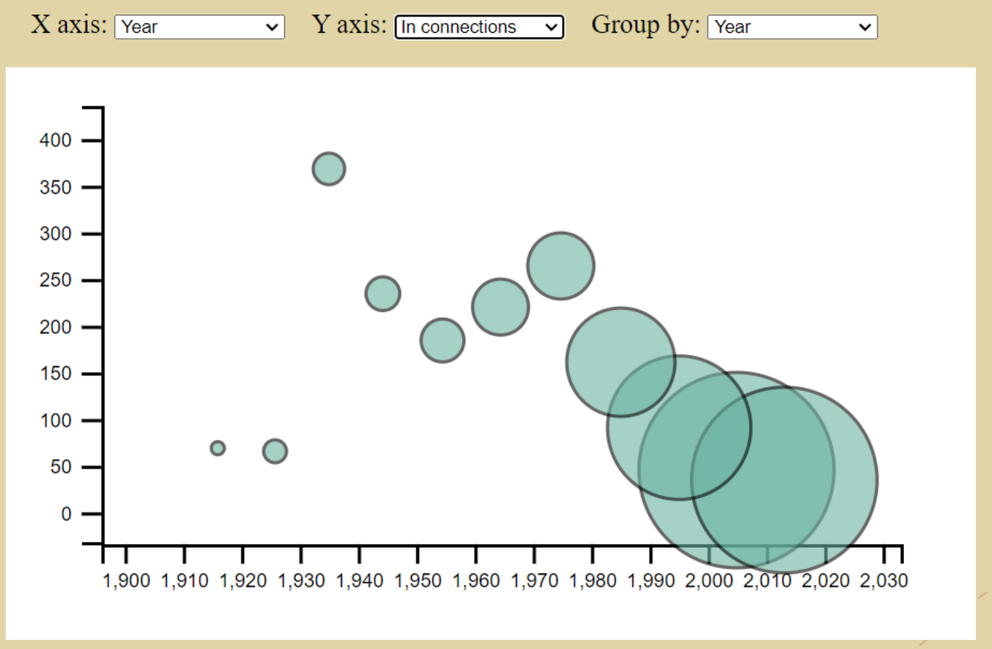


"[...] we find a strong correlation between number of **user votes** and **budget**. Remarkably, we find no evidence for a correlation between **number of votes** and **average user rating**..."

→

Insights

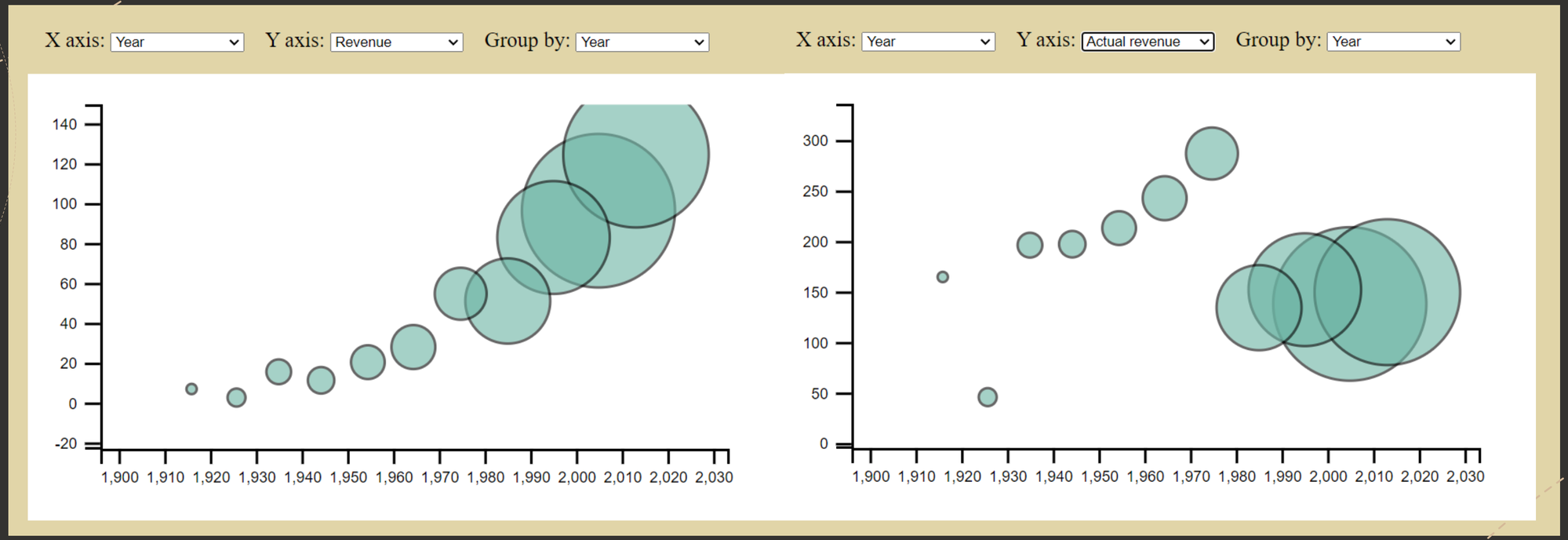




The first image comes from a paper, showing which decades were the ones receiving the greatest amount of incoming connections.
The same shape can be found in our visualization when grouping by year and plotting incoming connections over years.

→

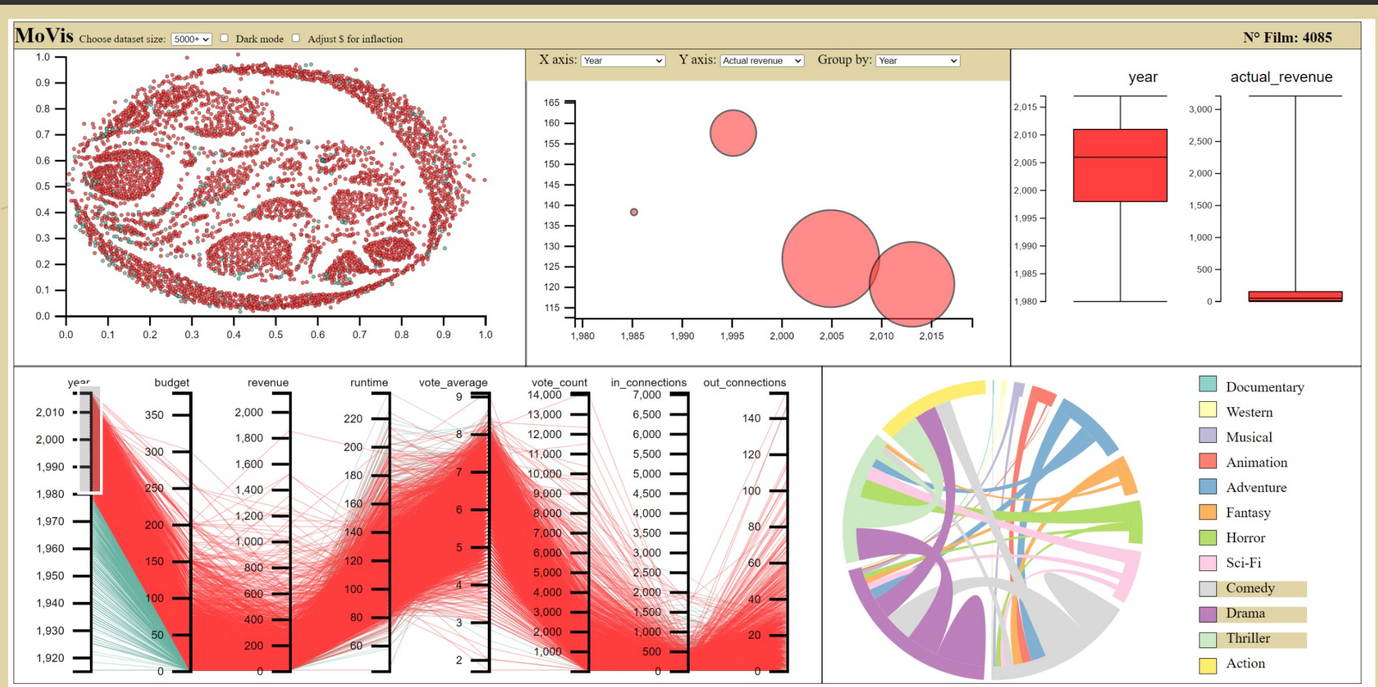
Insights

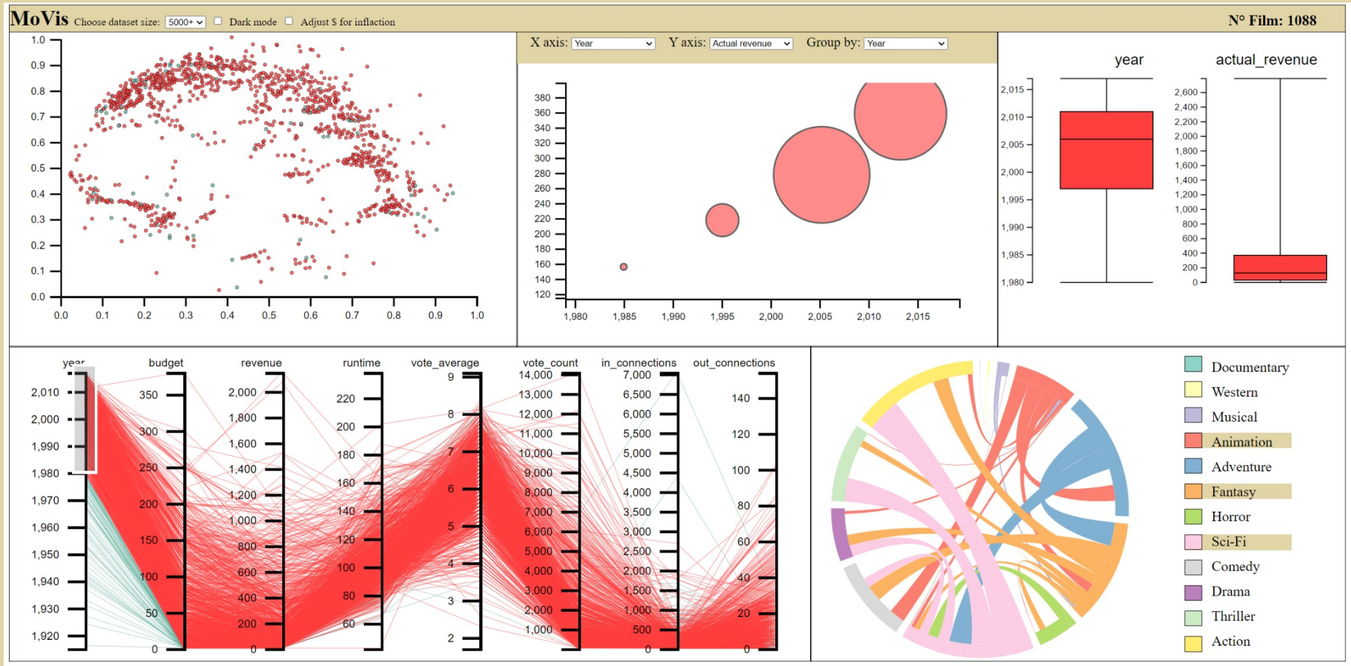


A visual confirmation of the correctness of adjusting for inflation.

→

Insights





The bubbleplot of these two different collection of **genres** shows opposite trends in **revenues**.

DEMO