# ISyE 6740 – Spring 2025
# Project Proposal

**Team Member Names:** Jeff Triemer

**Project Title:** The Evolving Faces of Genre Through Time

## Contents

## Problem Statement

At the point of sale, frequently books are either sold or left behind based on what the cover communicated to the reader and whether it seemed to fit their preferences. In such a crowded and competitive marketplace, we must assume that publishers and authors make every effort to advertise their product as effectively as possible. Book covers are a crucial tool in the publishers' arsenal to do just that. To that end, I wonder what publishers have done over time to design their book covers in this aim. My expectation is that covers have been optimized to quickly communicate to readers a given subject or genre for the book in question. As time has passed and publishers have become more competitive, has the cream separated from the crop? Have publishers refined their approach and settled on a clear and consistent approach? Do some publishers appear to have a different approach from others? Or alternatively, have books led efforts to differentiate themselves from their competition, leading to increasingly diverse bookshelves by genre?

My suspicion is that all publishers alike have converged on very similar approaches to book covers to effectively communicate their substance. In fact, this appears to have been backed up by some researchers. Some researchers have effectively used information from the covers of books to effectively predict their genre, reporting a 94% top 3 accuracy, when combining book title and cover data- 59% using just the book cover imagery alone (Kjartansson & Ashavsky, 2017). In this paper, we used a technique that will utilize dimensionality reduction, image clustering, and simple linear regression to attempt to determine how book covers have become similar over time and use some analysis and visualization to interpret why or how that might be.

## Methodology

I used use OpenLibrary's API to pull covers by their release years and then apply our own initial preprocessing, namely RGB normalization and image padding (to keep the aspect ratio unaltered). Next, we apply dimensionality reduction to the images, using UMAP as PCA will likely wash out some context that may be important to our use case. After this, I was ready to apply the clustering.

First, I determined the number of clusters using some exploratory analysis and the number of genres. I found the most success with a GMM model, so we will use this. Note that the 59% accuracy referenced from the research (Kjartansson & Ashavsky, 2017) includes some logic to extract title from a book's cover- I will likely miss this context in our simplified approach and find a worse performance, but this will show the overall fit of our clusters by genre. Instead, I compare the clustering to some random estimation of genre. This excludes any temporal split so that we can first understand the imagery cluster's relationships with book genre, and I have a graph of covers for visual exploratory data analysis.

Then I used the new clusters and some temporal data, after first doing some exploratory analysis into what temporal data would be the best fit for this use case. It may be that we group the data by decade, pre-determined publishing "eras" or just year. On inspection, I found the best fit to be simply using the year.

Next, once we have our clusters over time, we can analyze their results. My hypothesis is that publishers have become more "conformist" or targeted in their cover designs, and now I can see the nature of this relationship using the accuracy of my clustering over time. That said, these results may rely heavily on the quality of our data, dimensionality reduction, and overall cluster fit to book genre. I believe the best approach, particularly for its interpretability, was to use simple linear regression with published year or decade as an independent variable and accuracy as the dependent variable. Once I check the assumptions of the linear regression and verify the overall fit, I can use the beta of our temporal feature

and its value to interpret the magnitude, direction, and significance of the relationship between time and the accuracy of our temporal book cover clusters.

## Data source

I pulled a sample of ~300 books and their covers from OpenLibrary's API. This includes features like title, release date (either by edition or original book release), average rating and number of reviews, genre, subject(s), description text, and finally of course the cover image

Our data quality will be crucial to the success of the unsupervised learning we are using, and further the final simple linear regression. I had to do a bit of work to set up a pull for the image of each cover for a sampled book, but some additional work revealed that, while thousands of editions is tempting to include, it turns out that many of the additional editions came with an added baggage of data quality concerns, an example of which we see in the plotting of the book image clusters.

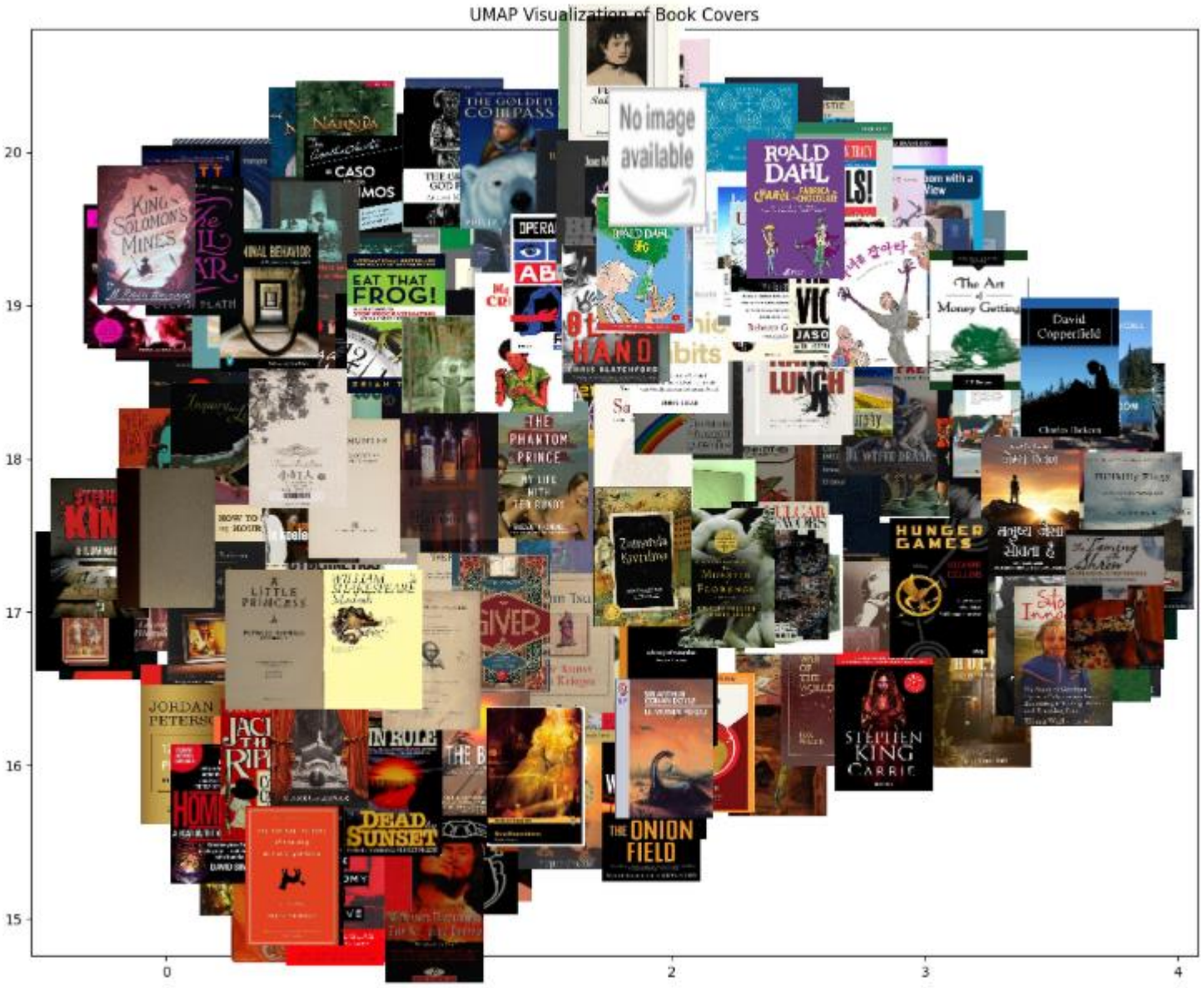
It is also worth noting that we have titles and book descriptions from this data. I approached this project with the knowledge that I could add to our clustering approach and add the data from the language in the title and book descriptions, which may further improve the clustering as it has in the aforementioned research. This will be a lever to pull if we find our initial example clustering lackluster or are just interested in taking the additional step. This will alter the eventual evaluation of the results, but it should all the same indicate how publishers have refined this aspect of their marketing approach over time.

## Proposed Evaluation and Final Results

As mentioned briefly in the proposed methodology, we have several outputs from our process that we can use to interpret and evaluate our results. First, we have the benchmark of random genre assignment to use in comparison with our clustering results. We plotted a confusion matrix here for easy interpretation.

Finally, I can use the outputs of our simple linear regression to first determine if it's a good fit for the data; it could be that simple linear regression is not the most appropriate form of regression. Once we have a decent fit, we can interpret the results to reject the null hypothesis that there is no significant relationship between the published date and the fit of our clusters.

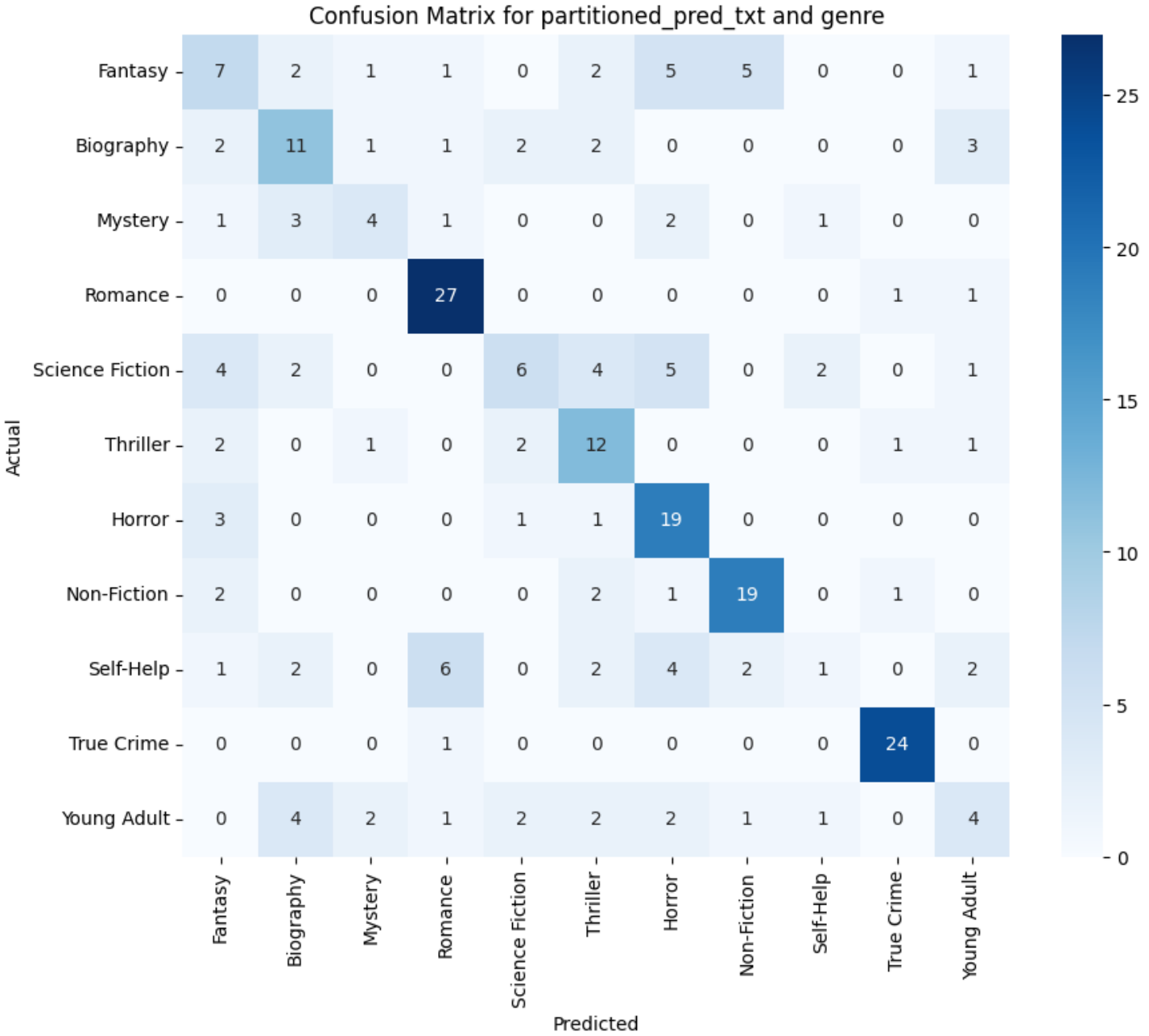
## Results




Much more representative of the cover's color rather than genre. We can also see there is some difficulty ensuring data quality when we select an image in OpenLibrary's api. The classic 'no image available' iconography from amazon appears present for one sample. This clustering model predicts genre about 3% better than a random assignment. This suggests there is at least some predictive value in the color of the cover, but very little!

Given that we are not going to tune a model to extract the text or iconography from the covers, we will have to pivot, applying natural language processing to book descriptions and titles in combination with the image features to see how effectively these can predict genre when coupled together.

Applying all image, description and title features together with some natural language processing, I have much more success clustering the books into their respective genres. This model appears to perform roughly 35% better than a randomized genre assignment. Looking at the results, I grow even more confident in the model's performance. This is due to the specific pitfalls the model appears to be falling prey to when we plot the model results in a confusion matrix.




There is some genuine diagonal dominance, with some genres being pain points. Young Adult, Self-Help, Mystery, Science Fiction and Fantasy being the genres that struggled the most. Most interesting to me is the Self-Help confusion. For one, it appears that language in the Self-Help genre has significant overlap with Romance. This to me suggests that the model has identified specific language like 'love' as

indicative of a Romance book, and is confused when Self-Help models speak of concepts like self-love, care, etc. Even more surprising is the overlap between Self-Help and Horror. My best guess is that there may be some overlap in language that is deployed to frighten Horror readers and relate to depressed readers with negative internal dialogues.

Young Adult, Mystery, Science Fiction and Fantasy all have results that suggest a similar pattern to me. These genres highlight the importance of the soft clustering method applied here. It appears likely to me that a given novel could be described as every one of these genres at once. I can think of plenty of books that are young adult / mystery / science fiction, for example.

As mentioned before, comparing these clustering results to some random assignment of genre, we see a ~**35%** improvement in performance. To apply some improved measurement of performance, we could also apply a 'top 3' clustering performance as well. This may be a good improvement for the future, however for our use case in utilizing the model's accuracy as a measure of how concise publishers' cover images, titles, and descriptions appear to be over time.

Next, we can use the model's accuracy in a new supervised model to see whether the published year has a relationship with our model's accuracy. I initially fit a regression model predicting accuracy percentage per decade, but this appeared to have too few rows left for any real inference. Instead, I built a logistic regression model to use the published year to predict whether the model would accurately cluster the book by its genre. This appeared to fit much better and gave us results we can easily interpret! With a beta p value of less than .01, we can say with 99% confidence that the clustering model predicts better for each additional year, but the result is a smaller magnitude than expected. Specifically, the model states that for each year later the book is published, the logistic regression model expects that a given book is .03% more likely to accurately cluster the book into its appropriate genre. For a year, this is relatively small at first glance. My first reaction is that while this might suggest my hypothesis that publishers have become more concise is accurate, I was expecting a more dramatic shift over time. For a more concrete example, if we compare two given books, like for example a book released in 1970 with one released in 2020, this model expects that the 2020 book is 1.5% more likely to be accurately clustered (50 years, expected clustering accuracy increased by .03% each year).

For further analysis, there are a few improvements to these models that would ultimately improve the ultimate conclusions we are able to draw. For example, it occurs to me that the relationship between publication year and clustering model accuracy might have a non-linear relationship that would be better modeled, some other way. I also think this methodology might be improved by performing more work related to the overall book genre and publication year imbalance. In this training data, I would love to have more overall total sample, as well as a balance of books over each year and genre. There also may be room for improvement in the regression model's inference variable. A WCSS metric might perform better here than simply did the clustering model get it correct. This is because, while an indicator of correct predictions allows us to model the data over time, WCSS might be a better fit to define 'publisher conciseness.' Finally, this clustering model primarily uses descriptions and text for predictions, despite the image data contributing too. That said, I still feel the overall image might have more to offer to this process. I feel that a more robust image classification model that can extract more specific features from the book cover would allow for a much more nuanced and interesting discussion on publishing techniques over time.

## Citations

Kjartansson, S., & Ashavsky, A. (2017). *Can you Judge a Book by its Cover?* https://cs231n.stanford.edu/. https://cs231n.stanford.edu/reports/2017/pdfs/814.pdf