**Did Your Mom Watch This Movie?**



An analysis of baby names and movie data.

Team 034 Benchwarmers

Jonathan Goldstein, Tadeus Rossetti Marchesi, Geoff Mueller, Tamas Gyalay

**Introduction**  
How do parents choose a name? Naming a child is a deeply personal endeavor which has numerous influences. A person’s name could elicit fond memories of movies, songs, friends, or families. What if movies, among other influences, have a direct effect on what parents name their children? This project explored how the names of characters in movies affect generational changes in naming children.

Although there are some initiatives in place, current baby name recommenders and books do not consider social influences of movies. This is evident in the "name recommender" space where contextual information1 and data visualization web applications2 explore historical trends in baby naming, and offer a browsable list of names given a set of inputs. In several academic articles and applications, a correlation between historical figures and baby names were prevelent3. In others, web applications such as 'The Name Voyager' do not offer an interactive way to visualize the effect of cultural references on baby names.

The scientific study of names is not a new theme4,5, and many authors investigated temporal trends of name popularity6,7. It is interesting to note that certain cultural elements, such as popular music8 and historical figures3, appear to correlate to the names given to children. However there has been no notable analysis on the specific impact of movies on names.

**Problem Definition (Heilmeier #1)**  
What impact do popular movie character names have on the popularity of baby names from year to year? This project created a tool to visualize the relationship between popular movies and generational baby names. Currently, no such tool or analysis exists on the scale that this project accomplished. Our innovation shows when the popularity of a baby name trend changed and what popular movies came out just before the changes. The tool also predicts the popularity of the name going forward to help parents understand how unique the name will be.

**Literature Survey**   
In evaluating this problem extensive research was done on the current best methods to gather name data, what factors influenced naming, recommendation engines for naming, and finally predicting name popularity based on social influences.

**1) Data Gathering:**Many of the works surveyed relied on the SSA data for naming information. One potential way to augment this data set is to enrich the data from other sources, such as movies, tweets, or song lyrics. Some papers identified ways to improve the reliability of naming frequency. IBM Research created a module using several heuristics to detect the use of names within natural language text12. Another potential technique was outlined by Wang13, where names were processed by a phonetic hash function enabling the determination of likeness between names. Barucca *et al*.7 found similar sounding names' popularities are found to be correlated within states, so this technique can aid in predictions/recommendations as well. These name correlations were replicated elsewhere.16 An NLP model like TweetNLP14 provides a potential framework to adapt to scraping public websites like Twitter for name use, to determine the context of a name or its frequency.

**2) Contributing Factors to Naming:**One existing paper examined the effect of Billboard Hot 100 songs on some name's SSA ranking, showing a bump in the popularity of a name around the time of a song's peak [6]. Showing a connection between pop culture and naming conventions can either enable a new predictive feature or allow for insight to detect other influential media. Not all cultural elements would affect naming patterns but some of the techniques outlined in [10] for time series change point detection would be useful in doing so.

A potential predictive quality may be the state or culture, as "Cross Correlations of American Baby Names" [14] investigated. There it was found groups of states which naturally emerged over time and shared naming behavior. In this work the baby names were actually used to quantify the evolution of culture over time. This could be used in our project to predict into the future or provide yet another feature to determine a more true popularity of a name over time. In the same vein, [8] investigates correlations between names over time. 'Phenomes' are identified which are related between names and can help identify future popularity.

Location can be useful in predicting likely names in an area due to cultural trends, and to visualize the distribution of names using data blended from the Census Bureau, Zillow, public telephone directories, Researchers at ASU created the *Name Profiler Toolkit* [17], which was used to identify 'hotspots' of certain ethnic groups based on the name distributions throughout the United States.

**3) Recommendation Techniques:**In September 2013 the 15th Discovery Challenge of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases was dedicated to improving baby name recommendations based on data gathered by a service called Nameling, acted as a search engine for names. Users would enter a name, and related names would appear based on a ranking algorithm used with data from the web evaluating the connectedness of related names. Further detailed in Ranking given names: Algorithms and evaluation paradigms9 andamong a series of other papers focused on Nameling, the authors described how similarity metrics among names mentioned in Wikipedia, popularity on Twitter (a topic we visited in “Tweetnlp: Cutting-edge natural language processing for social media”14), and the user's Social Graph were used to create Nameling. In one of the contributions to this challenge, the authors employed two existing recommender systems and adapted it to naming to create an improved system for naming based on Nameling data. In order to improve the system the authors also used contextual information such as time and location.

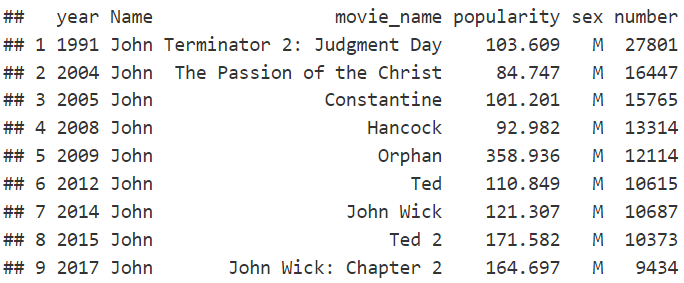
***4*) Predicting Names using the SSA Dataset:**Another service which caught on for naming, dubbed The Name Voyager was an exploratory data analysis tool for users to explore the trends of names over time using the SSA dataset2. The popularity of this tool further reinforces the demand seen for this service, and highlights the features which even brought the attention of folks who were otherwise uninterested in naming children. Naming frequency history (also using the SSA dataset) was also fitted empirically using a combination of Zipf's Law and the Beta function6, showing that the simplest dataset yields still useful results and limits complexity. FiveThirtyEight11 was able to find correlations between someone's likely age and their name. The groupings of names over certain time ranges also correlates these names with each other.

**Proposed Method: (Heilmeier #3)**  
Current approaches show a timeline of a name's popularity if users are lucky. Otherwise they are presented top 100 name popularity lists. Our approach predicts how the name popularity will change over the next few years. The Social Security name record is only updated once a year leaving people to guess if the names they are picking are going to be unique for their baby. In addition to the prediction users can see what popular movies may have influenced a name.

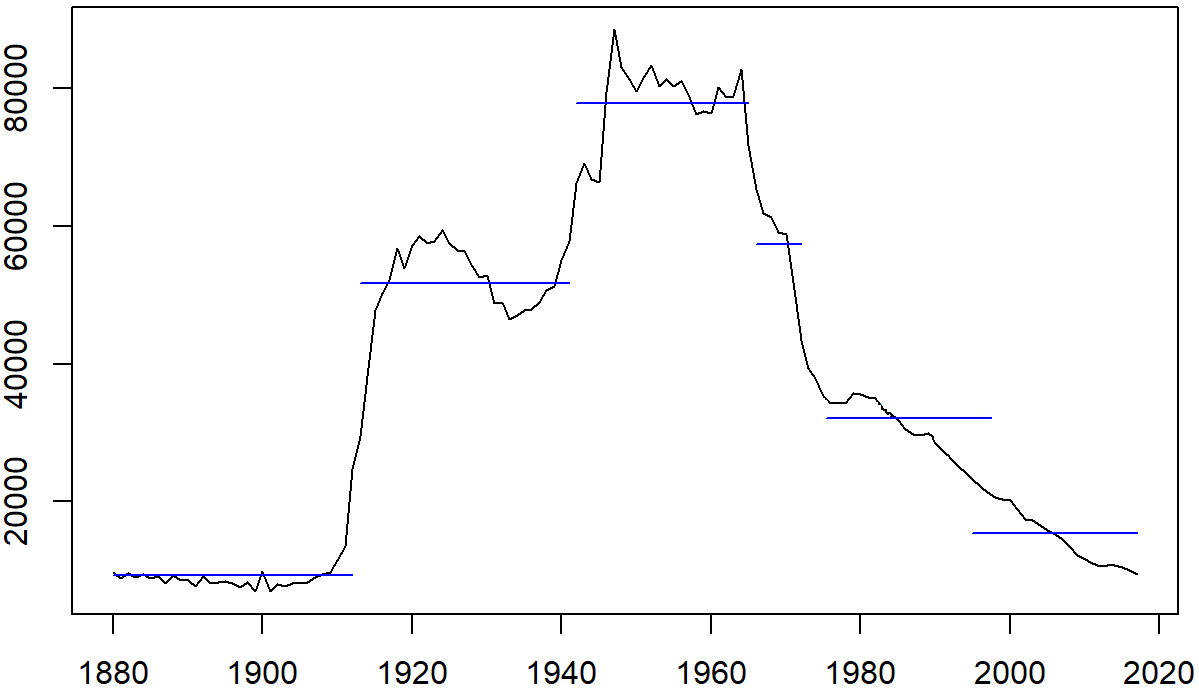
**Proposed Method: Detailed Approach**  
This analysis was done in the following order.

1. **Movie Data:** The top 1,842 most popular TMDB movies were pulled with an API. The data contains movie title, ID, the release date, and cast members names. The films span from 1990 to 2020. Character names had to be pulled separately with another call to TMDB.
2. **Name Data:** A database of SSA names was created. These names come from the Social Security card applications for births in the United States after 1879. The data is available as CSV files by year on the SSA.gov site. All 140 CSV files had to be combined to build the data set. There were 101,338 unique names in the data. Some of them appear in some years but not in others. Zeros were added to fill in the blanks to complete the time series.
3. **Movie Character Names:** A dictionary of release dates and first names of main characters was developed by combining data from multiple API calls against TMDB. The first 5 character names were used and assumed to be the most popular. The character first names were separated out as shown below. This is the field regression tests were run on.

Tabel 1: Movie Name Data



1. **Time Series:** Both the movie and names data sets were converted to time series using the R Forecast package. This is required to prepare the data for univariate time series forecasts, specifically ARIMA.
2. **Change Detection on Names Data:** Binary change point detection was used to perform fast signal segmentation using the R package “ChangePoint” to utilize BinSeg capabilities. This provided a simple method to generate breakpoint partitions within a list of specific years of when a name's popularity changed. Figure 1 shows the name John with 6 changes detected over the 140 year period. Binseg calculated the BIC and Residual Sum of Squares to determine the optimal breakpoints for the change detection.
3. **Visualization of breakpoints, changepoints, and trends of name.**

  
Figure 1:Change Detection for name John.

1. **ARIMA Name (1890 to 2020) with no dependent variable:** ARIMA was chosen to produce the time series forecast. Using the names data and the R “fUnitRoots” package, testing for unit root, seasonality, stationarity, and autocorrelation was performed. The data was adjusted based on the 4 tests. The model was then fit and a “Ljung–Box Q test” was used to view the results.

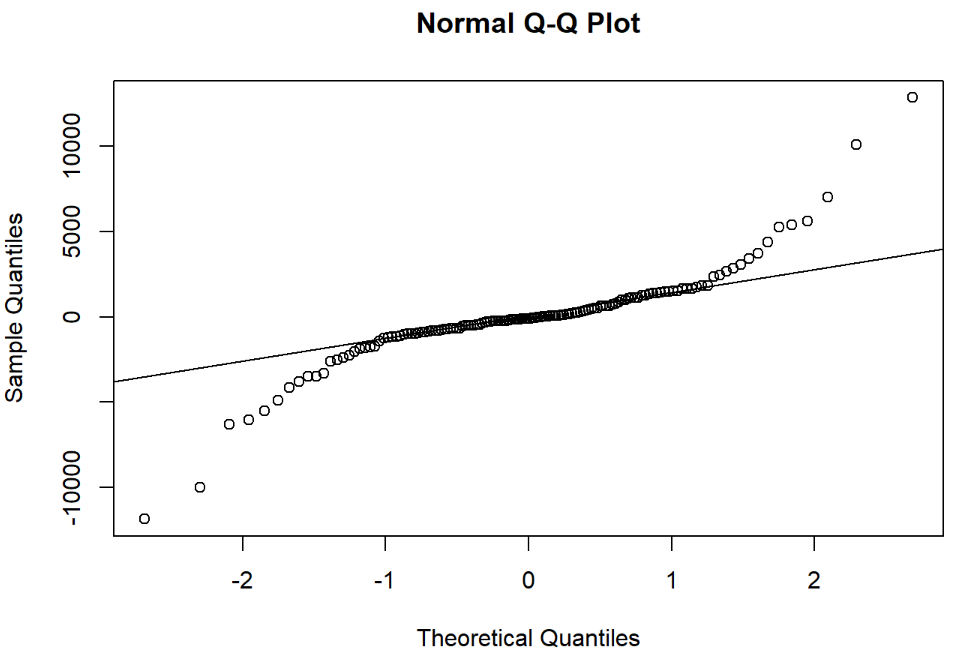
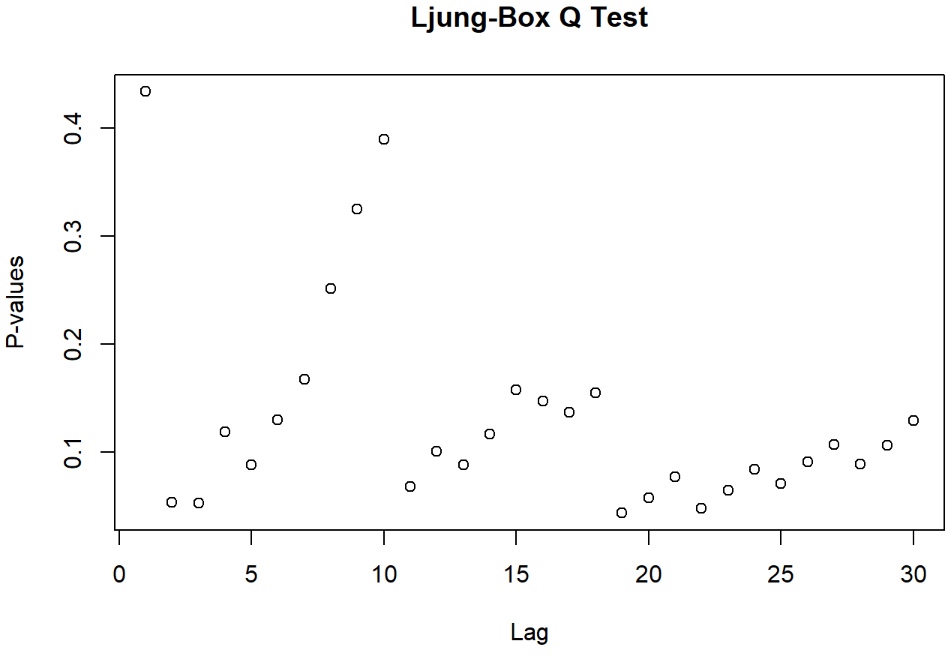


Figure 2:“Ljung–Box Q test” and Normal QQ Plot

To select the best ARIMA model, we used the auto.arima() function in R. It uses a variation of the Hyndman-Khandakar algorithm (Hyndman & Khandakar, 2008), which combines unit root tests, minimisation of the AIC and MLE to obtain an ARIMA model. The arguments to auto.arima() provide for many variations on the algorithm. Finally the ARIMA model is used to predict the next 20 years of the name's popularity, and in the case of “John” (Figure 3), the popularity will most likely remain the same.

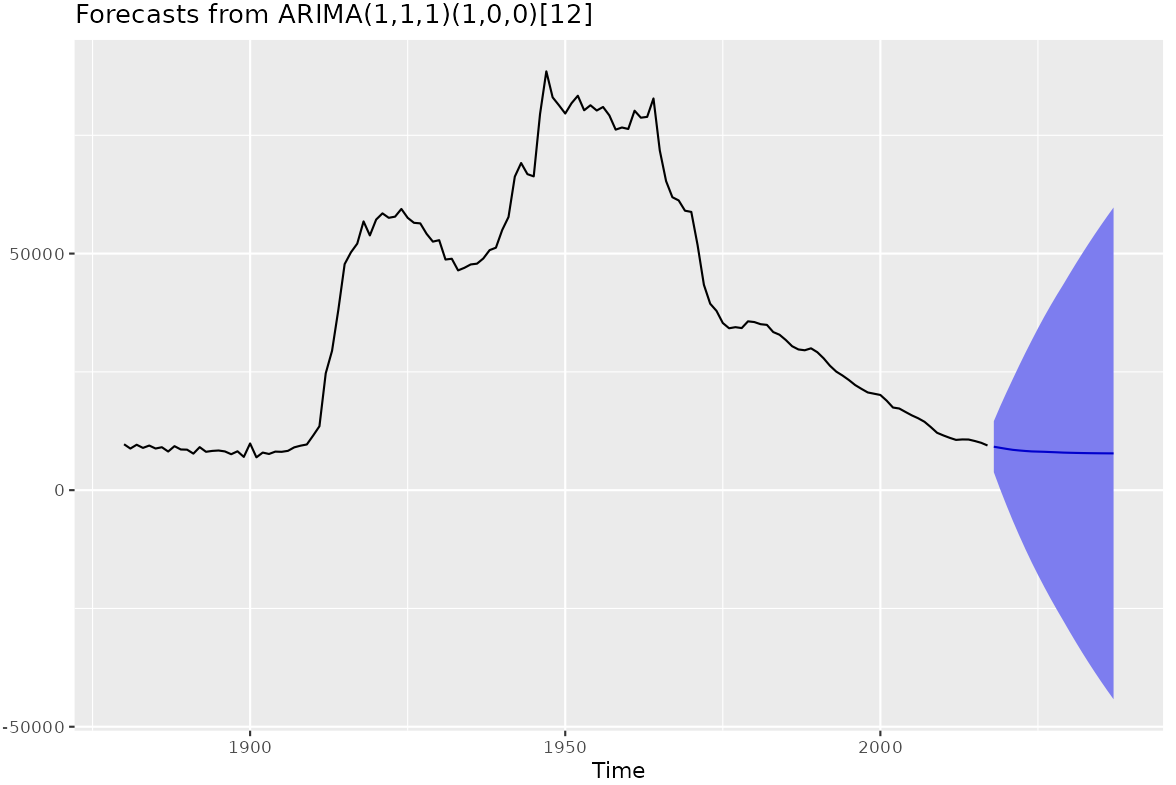


Figure 3:ARIMA Name Popularity Prediction (1890 to 2020, with Prediction to 2050)

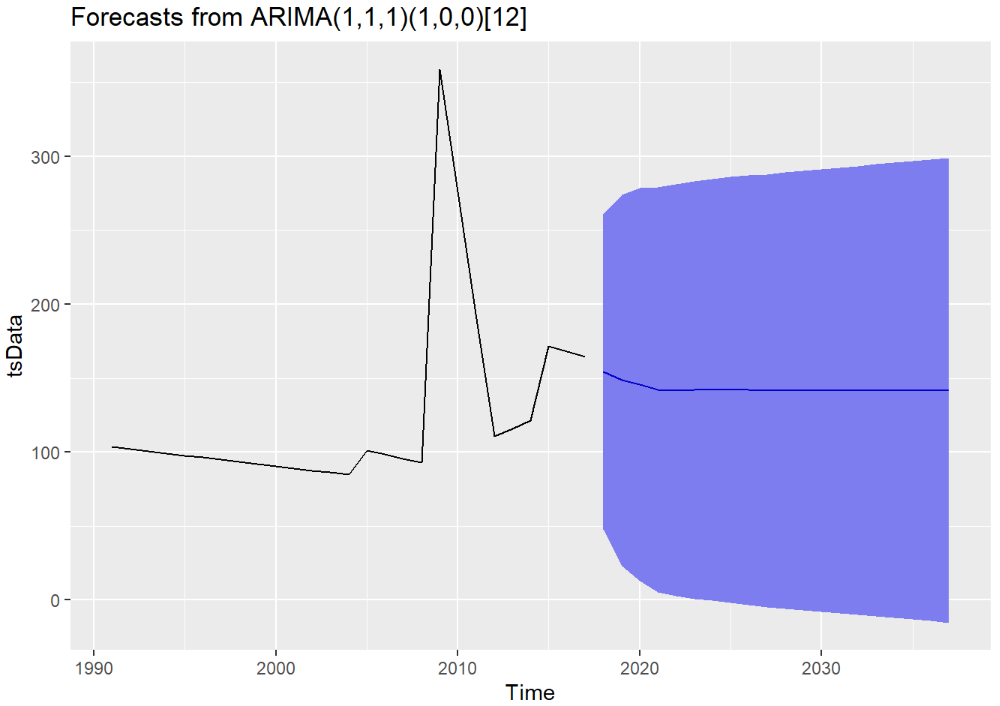
1. **ARIMA Movie (1990 to 2020):** The same process was used to predict the movie popularity if the movie had a character with the selected name. The time series was then adjusted to the movie start date (as shown in Figure 4 - Adjusted to 1990). The model predicted that movies with the name John will continue to have the same popularity, as the forecast is flat for T+20 years.  
   

Figure 4:ARIMA Movie Popularity Prediction (1990 to 2020, with Prediction to 2050)

1. **ARIMA Name Time Adjusted (1990 to 2020):** In order to visualize the forecast of the name within the same time period as part 8 above, a third ARIMA model was produced for the truncated time period. In this model, the popularity of the name John decreased. This is mainly because the model is referencing only 1 breakpoint as opposed to 6.

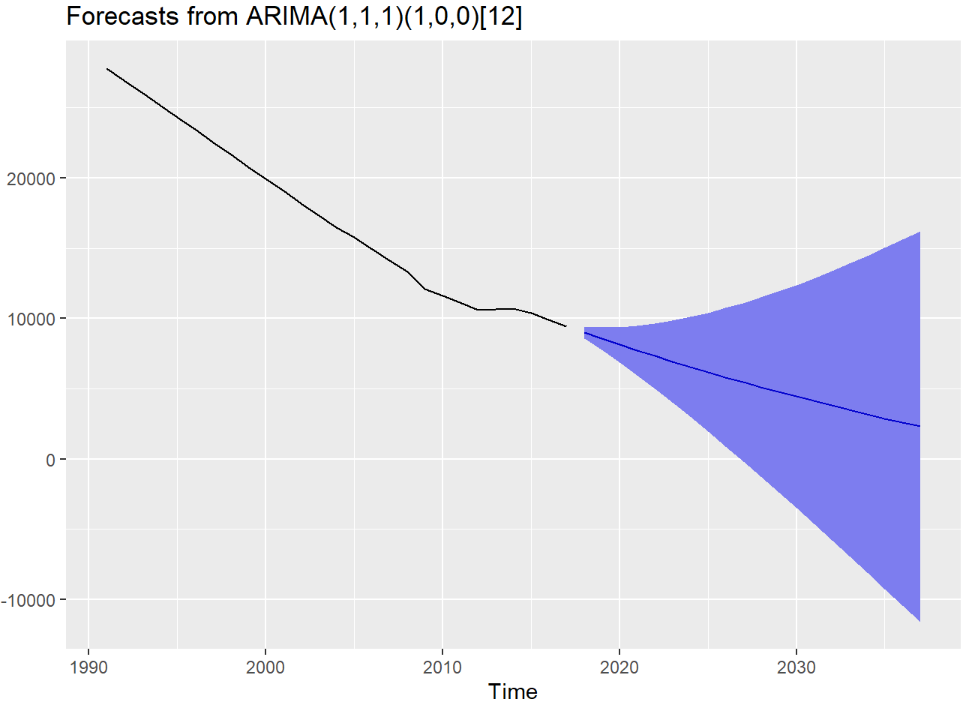


Figure 5:ARIMA Name Popularity Prediction (1990 to 2020, with Prediction to 2050)

1. **Regression:** The output of both ARIMA 1990 forecast models was run in regression to determine the correlation of the movie's popularity on the names popularity. The lm() function in R’s stats package was used to run the regression. The data was run with and without interpolation.   
   -0.074 = Correlation without Interpolation  
   -0.48 = Correlation with Interpolation. As in the case for the name John, movies negatively influence name popularity with a semi strong relationship. However, when running on the broader name group, our conclusion is that the model is not strong enough to provide evidence that movie data directly influences name popularity.

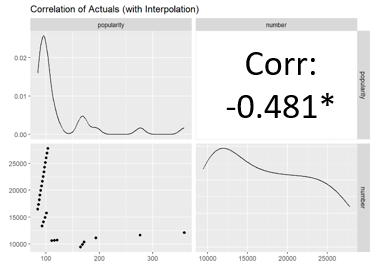


Figure 6:Correlation with Interpolation

1. **User Interface (UI):** Since the model was not accurate enough to find a relationship between name popularity and movie popularity, the framework for the UI was adjusted. The UI shows both the change point detection graph and the name popularity prediction graph. The change point detection graph allows users to see when the name popularity changed. They can then look at the list of movies where a top character had that name and when the movie came out. The second graph shows the ARIMA results of the forecasted popularity of the name going forward. Helping parents determine if they are picking a unique name.

To build the UI, R’s Shiny Package was used. Once a user selects a name and a gender, the R code is executed, and the graphs and table are produced. The typical refresh time is 10 seconds.

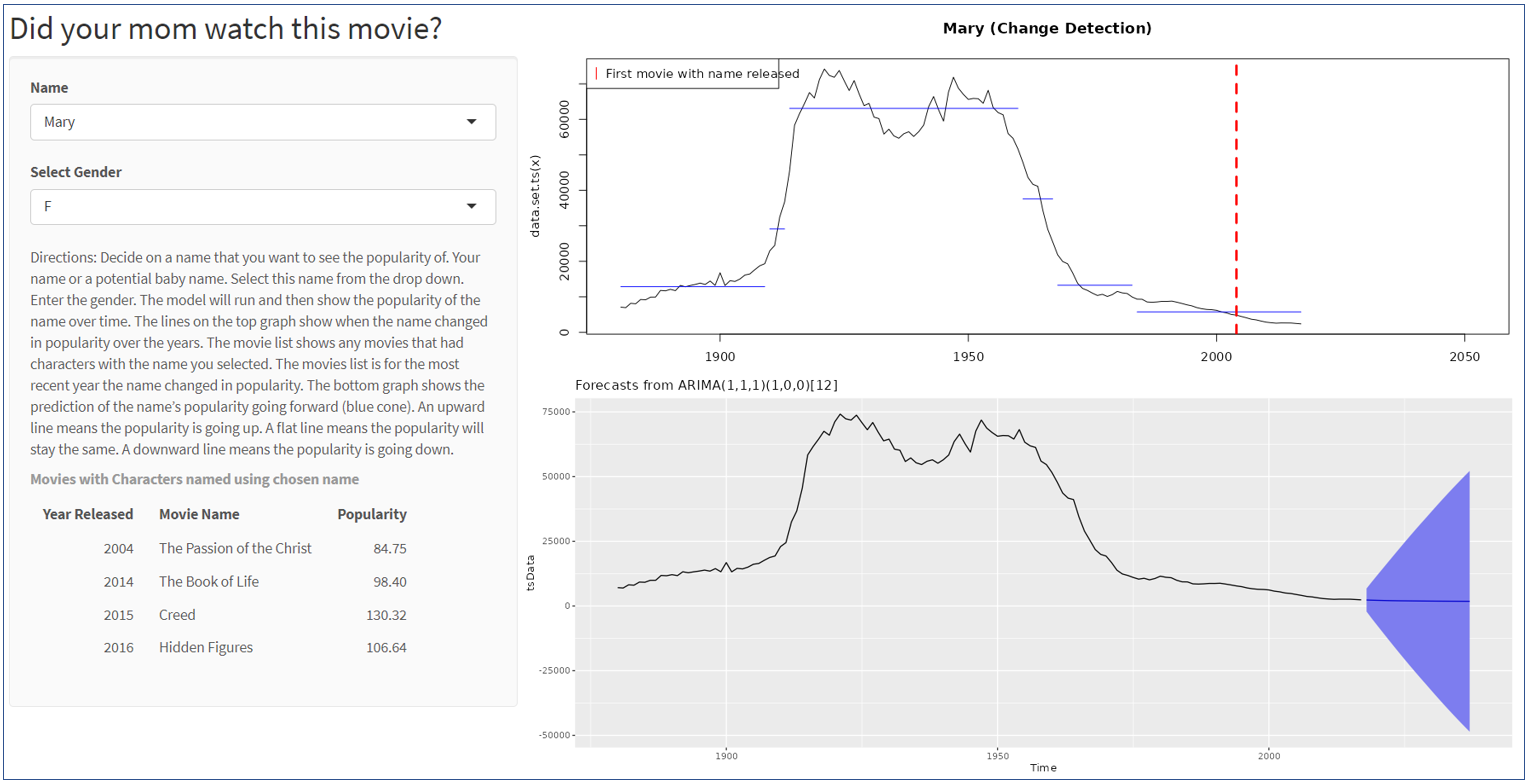


Figure 7:User interface in R Shiny  
<https://bench-warmers.shinyapps.io/baby-names/>

**Experiment Questions and Evaluation:**  
To evaluate the effectiveness of our tool, we wanted to answer the following questions. The experiments we used to do so are listed directly after each question.

**Q 1:** Do the character names in popular movies impact the popularity of those baby names?  
**Exp 1:** We ran a linear regression on both sets of ARIMA results (Names and Movies) and checked the correlation and RMSE. There is no evidence to support the hypothesis that the movie data enhances the name prediction. The cand RMSE ranged from 6.4 to 5,177 (sample population).

**Q2:** Can one reliably predict the popularity of a name in the SSA data a few years into the future?

**Exp 2:** A “bag of 30” approach was used to run 30 names and check the range of the MASE results. The MASE averaged ~0.83, with all sample values below 1. The predictive forecast is better than selecting the last observation in the sample.

**Q 3:** Will people want to see if their name was influenced by a movie?

Exp 3: We had 10 parents use our Web Interface and adjusted it based on their feedback.

* Parents wanted to start by selecting a name, not a movie title. We changed our UI accordingly.
* They said our site was confusing. They weren’t sure what to do. We added directions.
* The column names on the movie table were confusing (i.e. “Change Year”). We adjusted these to be more intuitive.

**Q 4: What change detection technique should we use?** Several approaches for the changepoint detection were tested. **Exp 4.1:** The R package ChangePoint was used to produce a BinSeg signal detection. This provided the best approach as the results were easily isolated and visualized.

**Exp 4.2:** A Bayesian model averaging algorithm called BEAST was also tested, but its results provided more analytical measure than were needed.

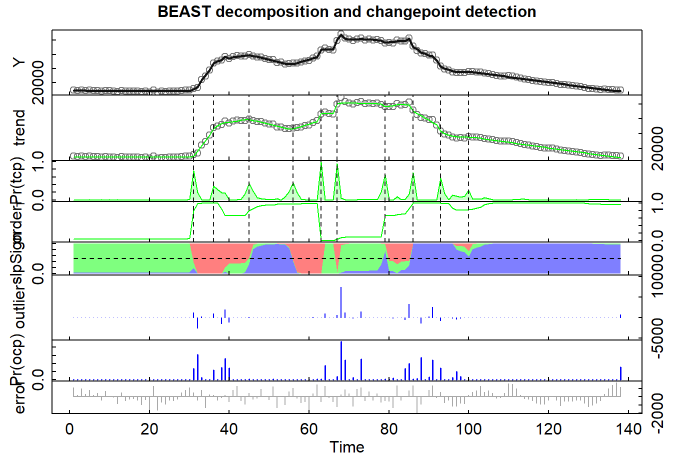


Figure 8:Output of the BEAST method.

**Conclusions:**  
All team members have contributed a similar amount of effort.

No evidence supported the hypothesis that the movie characters' names influence baby naming. We tested the regression model on a sample of 30 names and the results were sporadic. Correlation ranged from -1 to 1 and RMSE ranged from 6.4 to 5,177 (sample population). This implied that our models were unreliable and the data did not suggest a relationship. One of the reasons for the variability of results is mainly derived from the scarcity of data within the TMDB. With more data points, a more accurate model could be produced, and perhaps, a better relationship could be explored.

However, since the SSA data was quite robust (1.9 data points), a predictive analysis of a name's popularity can still be forecast. This allows us to help parents to understand if they are picking a unique name. The ARIMA models for the sample population averaged a MASE of ~0.83, with all sample values below 1. This implies that the predictive forecast is better than selecting the last observation in the sample.

In order to incorporate movies into the visualization, the change detection graph was added to the UI to help users understand movies where a top character had the same name along with the year the movie came out. This allows users to see what movies may have influenced the name popularity changes.

**Discussion & Future Enhancements:**  
Choosing a child’s name is something that parents spend considerable time and effort agonizing over. Our analysis gives parents a unique opportunity to avoid picking names that will become popular in the future.

Future enhancements that could be added:

1. Create word frequency distribution (using phonetic n-gram relationship) to develop correlations. This may help to identify a correlation between name popularity and movie character popularity.
2. Movie Popularity: TMDB shows the current movie popularity as of now. Baby names were picked based on the popularity of the movie at the time the name was picked. Enhancing this analysis with the popularity of the movies in each year may enhance this analysis.

**REFERENCES**

[1] Domingues, M. A., Marcacini, R. M., & Oliveira, S. (2013). Improving the Recommendation of Given Names by Using Contextual Information. ECML PKDD Discovery Challenge, 61.

[2] Wattenberg, M. (2005, October). Baby names, visualization, and social data analysis. In IEEE Symposium on Information Visualization, 2005. INFOVIS 2005. (pp. 1-7). IEEE.

[3] Ludwig, L., Hightower, M., Engels, D. W., & McGee, M. (2019). Analyzing Influences on US Baby Name Trends. SMU Data Science Review, 2(3), 8.

[4] Bredart, S., Brennen, T., & Valentine, T. (2002). The cognitive psychology of proper names. Routledge.

[5] Berger, J., Bradlow, E. T., Braunstein, A., & Zhang, Y. (2012). From Karen to Katie: Using baby names to understand cultural evolution. Psychological science, 23(10), 1067-1073.

[6] Li, W. (2012). Analyses of baby name popularity distribution in US for the last 131 years. Complexity, 18(1), 44-50.

[7] Barucca, P., Rocchi, J., Marinari, E., Parisi, G., & Ricci-Tersenghi, F. (2015). Cross-correlations of American baby names. Proceedings of the National Academy of Sciences, 112(26), 7943-7947.

[8] Napierski-Prancl, M. (2016). Brandy, You're a Fine Name: Popular Music and the Naming of Infant Girls from 1965-1985. Studies in Popular Culture, 38(2), 41-53.

[10] Aksholakova, A. (2014). Proper name as a clue symbol of identity. Procedia-Social and Behavioral Sciences, 112, 465-471.

[9] Mitzlaff, F., & Stumme, G. (2012, December). Ranking given names: Algorithms and evaluation paradigms. In 2012 International Conference on Social Informatics (pp. 185-191). IEEE.

[11] Silver, N., & McCann, A. (2014). How to tell someone’s age when all you know is her name. FiveThirtyEight.

[12] Apte, C., Morgenstern, L., & Hong, S. J. (2000). AI at IBM research. IEEE Intelligent Systems and their applications, 15(6), 51-57.

[13] Wang, A. P., Pullen, B. D., Talburt, C. J., & Wu, D. N. (2015). Applying Phonetic Hash Functions to Improve Record Linking in Student Enrollment Data. In Proceedings of the International Conference on Information and Knowledge Engineering (IKE) (p. 187). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[14] Camacho-Collados, J., Rezaee, K., Riahi, T., Ushio, A., Loureiro, D., Antypas, D., ... & Barbieri, F. (2022). Tweetnlp: Cutting-edge natural language processing for social media. arXiv preprint arXiv:2206.14774.

[15] Mitzlaff, F. & Stumme, G. (2012). Namelings - Discover Given Name Relatedness Based on Data from the Social Web.. In K. Aberer, A. Flache, W. Jager, L. Liu, J. Tang & C. Guéret (eds.), SocInfo (p./pp. 531-534), : Springer. ISBN: 978-3-642-35385-7

[16] Mitzlaff, F., & Stumme, G. (2013). Onomastics 2.0 - The Power of Social Co-Occurrences. ArXiv, abs/1303.0484.