racial lines:

race, ethnicity and dialogue in 780 hollywood films, 1970-2014

vicky svaikovsky, anne meisner, eve kraicer & matthew sims

McGill

.txtLAB COLLABORATIONS

Introduction

The lack of racial and ethnic diversity within Hollywood cinema has emerged as a broad public concern. As a recent paper by the Media, Diversity and Social Change Initiative at the USC Annenberg School for Communication and Journalism has shown, there remains a consistent bias towards able-bodied, straight, cisgender, white men in contemporary films.

As scholars have long pointed out, American cinema began with the deployment of racial stereotypes. The first feature length American film, W. W. Griffith's *The Birth of a Nation*, is a narrativization of a Black man, Gus, played by a white man in blackface, who causes the death a white woman, leading to the justification and glorification of the rise of the Klu Klux Klan. Recent debates surrounding Academy Awards nominations, along with social media campaigns like #OscarsSoWhite, have helped make this issue more explicit within public consciousness.

Our paper aims to contribute to these debates using the tools of data science in two principal ways: first, we provide historical context to better understand the scope of this problem with respect to the present. When we talk about underrepresentation in Hollywood, what are the historical patterns that have helped shape where we are today? Second, we try to provide a more nuanced discussion of what it means to be "represented" in a visual medium like film. Building on the work of Maryann Erigha, we measure representation in three ways:

1. **Quantitative Representation**

How many actors from different racial and ethnic groups are seen and heard in our data set? Representation is understood here as a question of distribution: how well distributed are roles among different groups? Measuring quantitative forms of representation are important because they help combat the problem of t*okenization*, when a single example (whether an actor or a film), is held up as a solution to the problem of diversity.

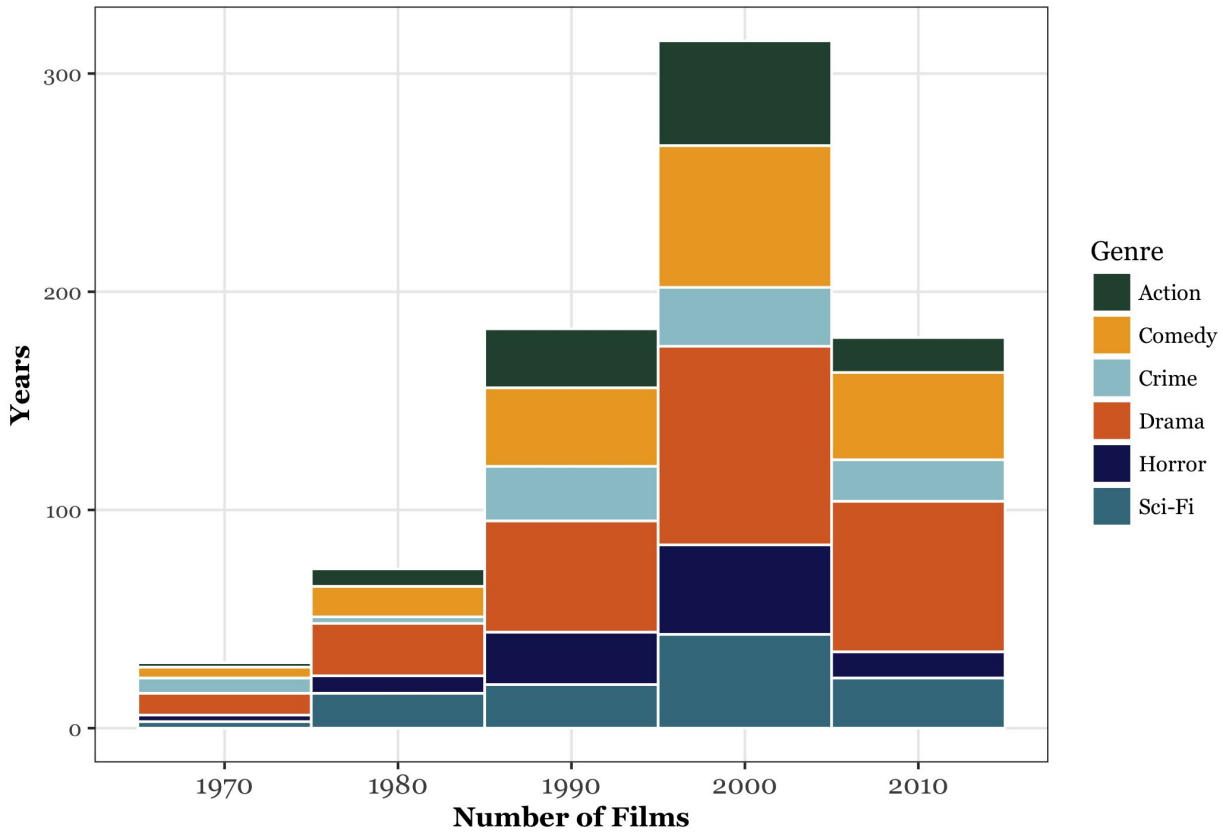
2. **Centralizing Representation**

What is the overall prominence of actors from different racial and ethnic groups? More does not necessarily equal better. Here, we explore who plays a more central role in films. It is not enough to have more actors from underrepresented groups if those actors have less prominent roles in films.

3. **Qualitative Representation**

Finally, we look at the type of language associated with actors from different racial and ethnic groups. When actors do appear on screen, what *kinds* of roles are they playing? How can we use new techniques in language analysis to assess diversity with respect to roles that actors play? Do actors from different racial and ethnic groups have the same opportunities for on-screen creative expression?

Figure 1: Genre Distribution of Films by 10 Year Segments



Our research is based on a data set of screenplays of 780 films, which were drawn from a prior research study on gender and dialogue conducted by Matt Daniels and Hanah Anderson at The Pudding. These films cover 44 years, from 1970-2014 (**Fig. 1**). After downloading our scripts, we first hand-labeled every film by genre, where a film could only be in a single genre. We then automatically extracted dialogue for every character and matched characters to their respective actors using IMDB. Keeping only characters who spoke more than 250 words in a movie (roughly equivalent to a minimum of two minutes of screen-time), we hand-labeled the actor's race / ethnicity according to seven categories: Black, East Asian, Indigenous, Latinx, Near Eastern, South Asian and white. In cases where one parent was white and the other not white, the actor was labelled as the non-white race or ethnicity. In these cases, data on parentage was recorded and analyzed separately.

In all, our data consisted of 4,058 characters whose dialogue made up 3,136,683 words. **Table 1** shows the total number of characters and unique actors per group. All of our data and code is available here.

	Black	East Asian	Indigenous	Latinx	Near Eastern	South Asian	White
# Characters	385	44	8	66	9	21	3,525
# Unique Actors	279	39	8	51	9	18	2,148

Table 1: Overview of character data by race / ethnic group

Limitations

Before presenting our findings, we would like to address the limitations of this project:

Sample Bias: Our data was drawn from a previous study. It cannot be taken as fully representative of all Hollywood movies from this period. However, there are as of yet no existing definitive, large-scale collections of of Hollywood screenplays from this period. Inspecting our sample, however, we feel that it provides a good representation of major Hollywood films with a reasonable distribution by genre.

Screenplay Bias: Our data is drawn from available screenplays of films, which are not identical with what is spoken on-screen. Our attention to language also ignores important visual cues that shape a character's identity on-screen. Both of these areas deserve further analysis.

Error: As we use an automated process to extract character dialogue from screenplays, some data is lost in this process; however, we valued losing dialogue (and some characters, who then dropped under 250 words), over using a process that could falsely attribute dialogue or descriptive information to characters.

Reception: Our data does not account for viewer reception, but only looks at the conditions of representation. How audiences respond to our forms of representation is an open and valuable research question.

Findings

Quantity

Quantity allows us to measure how many actors of different racial and ethnic groups are seen and heard in our data set. It can tell us how visible or invisible different groups are with respect to the real world. We use the U.S. Census data from 2000 (the mean year of our data set) to better understand the discrepancies between population size, character counts and dialogue distribution.

Character Distribution

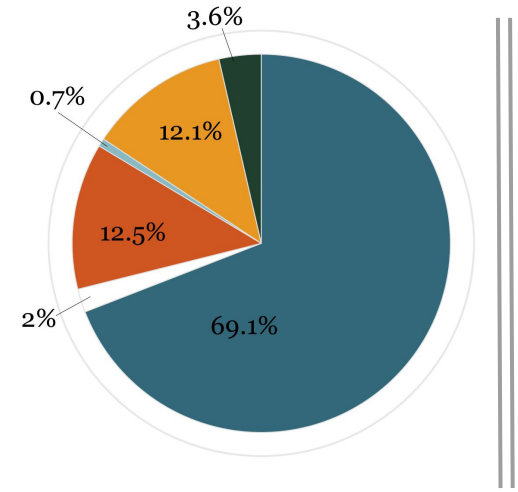


Figure 2a: Distribution of population by Race / Ethnicity in 2000 U.S. census

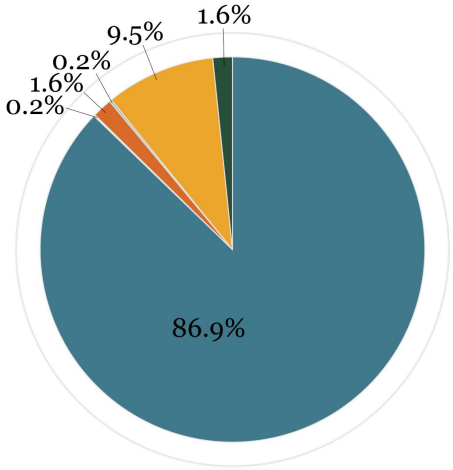


Figure 2b: Distribution of characters by Race / Ethnicity in dataset

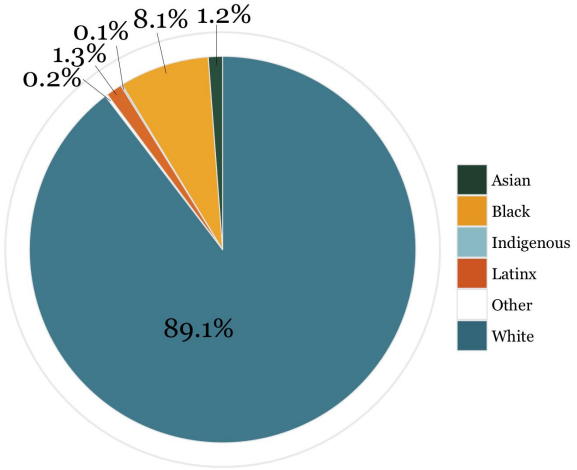


Figure 2c: Distribution of words spoken by Race / Ethnicity in dataset

White people are considerably overrepresented in Hollywood films. White people comprised 69.1% of the U.S. population in 2000 (**Fig. 2a**). Comparatively, white characters comprise 87% of all characters in the data set and 89% of words spoken in our dialogue files (**Fig. 2b** and **2c**). In other words, white people are almost 3 times more likely to appear as characters in movies and just over 3 and a half times more likely to speak than their population size would predict, leading to the underrepresentation of all other groups.

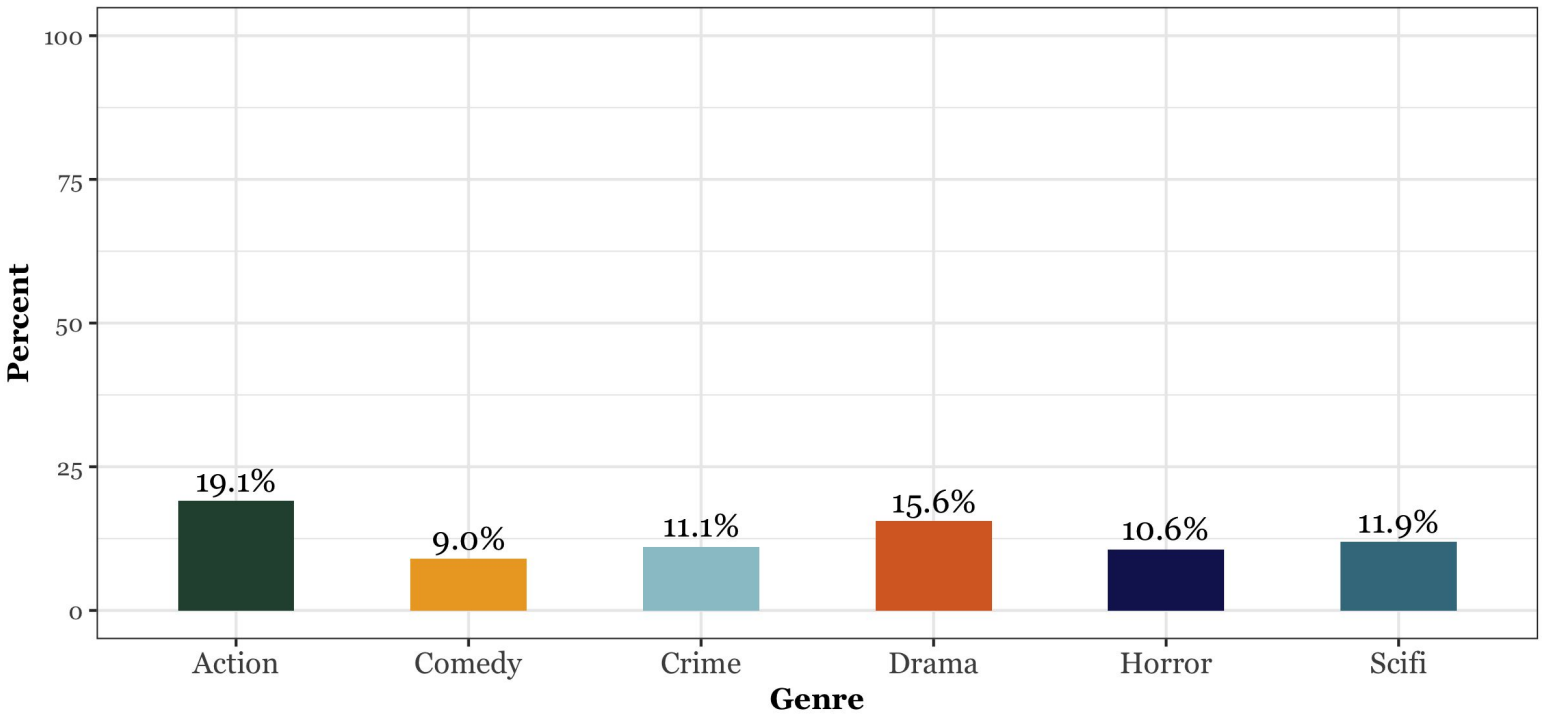
How many of our 780 films would you need to watch to see all representations of:

White Characters?	761 movies (98%)
Black Characters?	228 movies (29%)
Latinx Characters?	57 movies (7%)
East Asian Characters?	32 movies (4%)
South Asian Characters?	14 movies (1.8%)
Indigenous Characters?	7 movies (0.9%)
Near Eastern Characters?	6 movies (0.8%)

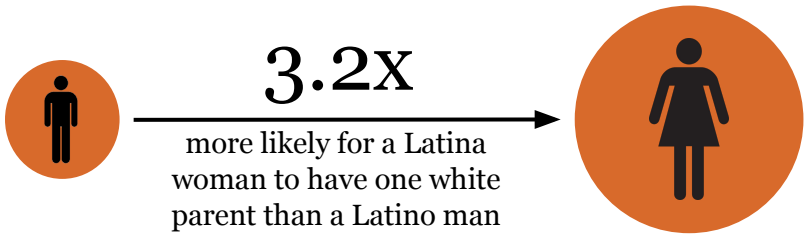
Race and Genre

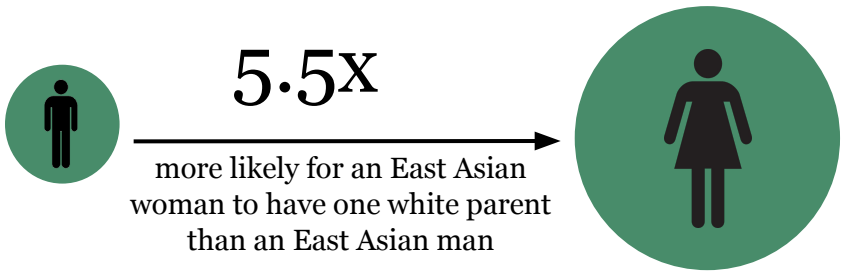
Not all genres are created equal. With respect to distribution of words spoken by group within each genre, we can see that, although all genres exhibit poor diversity, some genres, like comedy, require more work than others (**Fig. 3**).

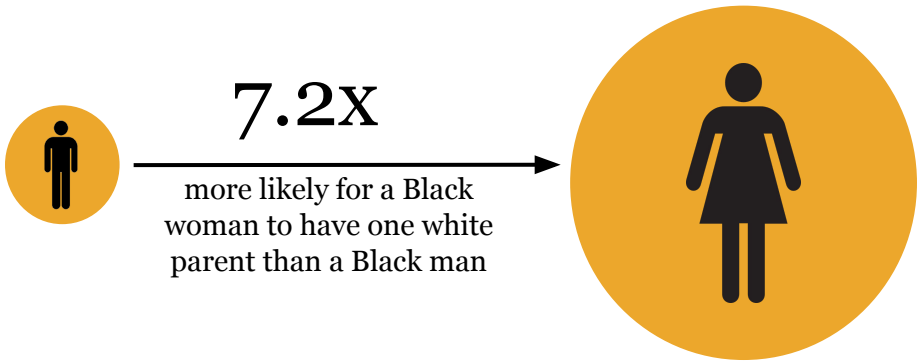
Figure 3: Percentage of Non-White Characters by Genre



Does Hollywood prioritize whiteness particularly with respect to women? A troubling finding in this research indicates that there is a significant difference between men of colour and women of colour in terms of racial and ethnic parentage. Using annotations on actor race (as opposed to previous measures which use character race) In our data, we find that women of colour are over 6.7 times more likely to have one white parent than men of colour. Although this occurs in all groups, it operates to different degrees.







In a recent paper, Marta Holliday discusses the resurgence of the 'Tragic Mulatta' a racist trope of a woman with one Black parent and one white parent. This character is often highly eroticized, and tied to themes of both submission and violence.

A NOTE ON RACIAL AND ETHNIC IDENTITY — Racial identity is complex, nuanced, and personal. We do not aim to make any claims about how these actors identify, nor are we stating that they are passing as white. We do, however, see a pattern in the ways intersectional identities function which may suggest biases in casting. We hope more research can be done to understand how, where, and why, these numbers exist in the ways that they do.

Centrality

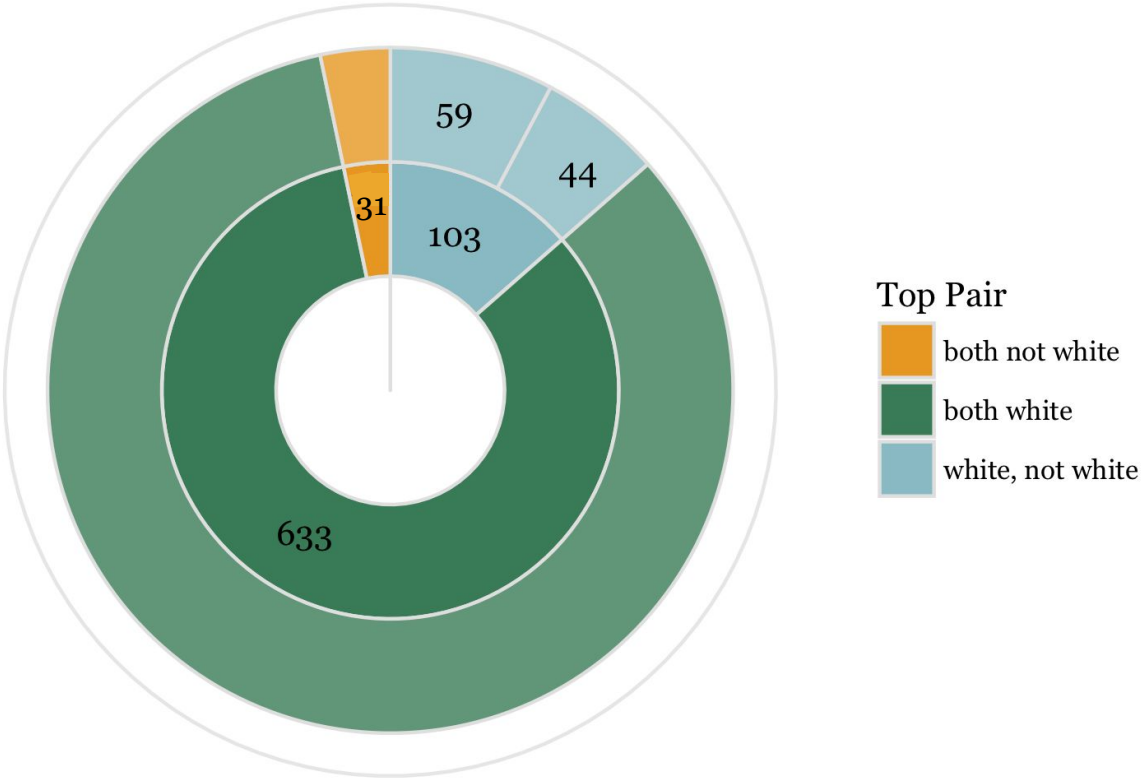
In the quantitative analysis above, we treat all appearances by characters as equal, which is not an accurate reflection of a film. In this section, we examine the centrality of characters by their racial and ethnic identities. We do so in two ways: first, by observing the top two characters for every film, and second, by analyzing quartiles of all characters per film.

Top Billed Characters

Seeing two non-white characters in leading roles is very rare. White pairs dominate the top-billed characters of our data set; in pairs with only one white character, they most often speak more (**Fig. 4**).

In other words, two white characters are **111 times more likely** to make up the top pair than two non-white characters. Given a pair containing one white and one non-white character, the white character is **1.8 times more likely** to speak more.

Figure 4: Who Speaks the Most? Race and Ethnicity of Top Two Characters

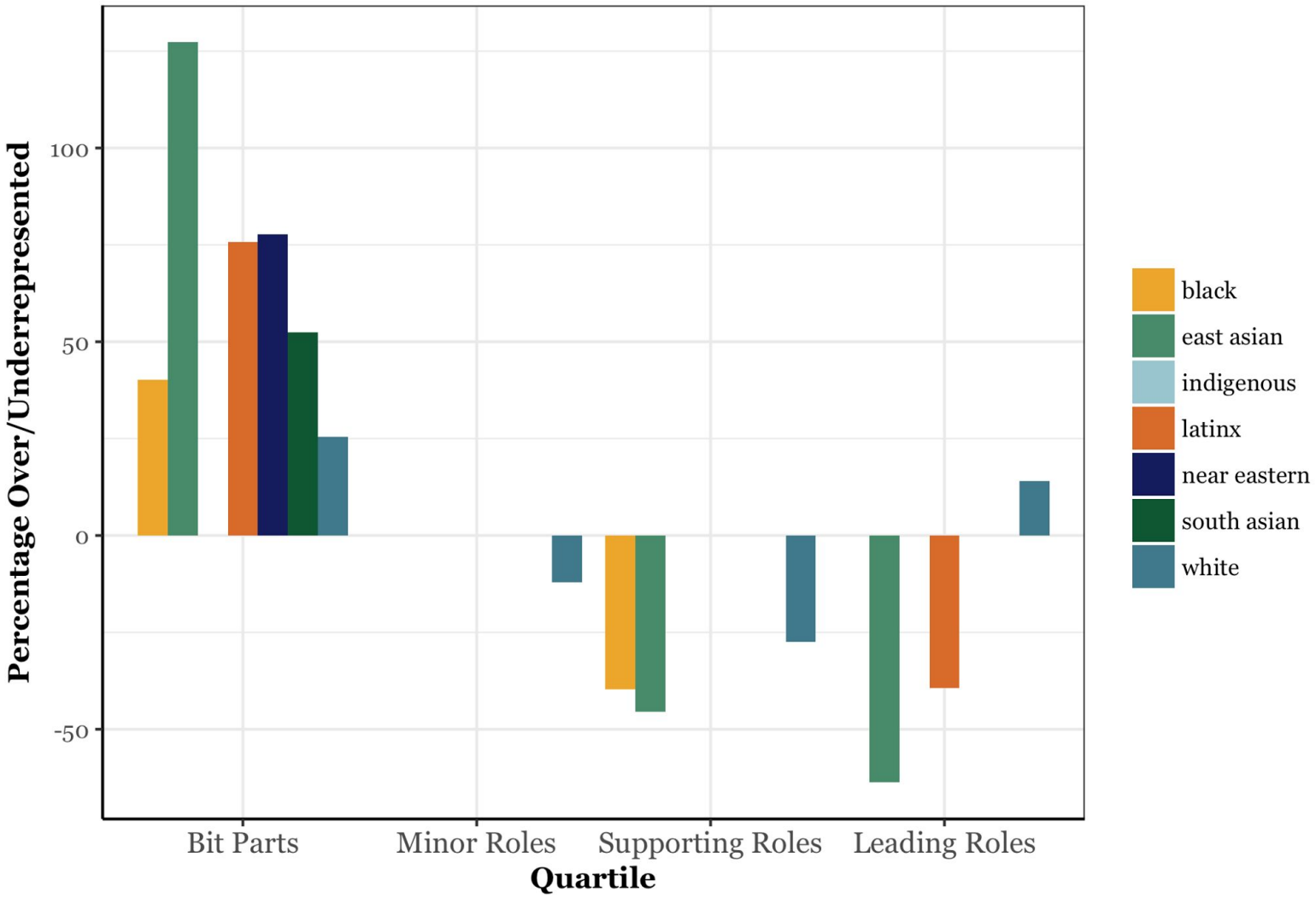


Distribution of Dialogue

We used word counts to create a distribution of characters, effectively dividing bit roles and top-billed characters among quartiles. This allowed us to visualize where different groups are over- or underrepresented. The following figure only shows significant over- and underrepresentation per quartile.

Non-white characters are considerably more likely to be in the bottom quartile of dialogue. Breaking down the appearances of characters by number of words spoken in film, we can see a higher likelihood of characters who aren't white appearing in the bottom quartile (**Fig. 5**). Instead of getting roles with varying amounts of screen time, minority characters are far more likely to get bit roles. Only white characters are significantly overrepresented in the top quartile.

Figure 5: Significant Over- and Underrepresentation of Characters



Quality

Given how few and how decentralized the appearance of characters of colour are in these films, those words which they do say hold a significant amount of weight in creating our conceptions of how these groups behave and speak. To study this, we analyzed the dialogue of our characters, looking for linguistic signals that would give us insights into how these characters are represented in our dataset.

We approach this through the concept of geographic tokenization. If a character of a certain racial or ethnic group appears, is the dialogue more likely to be associated with a particular geographic region? **Are actors that belong to specific ethnicities, in other words, more likely to be cast for roles associated with the underlying geographic regions associated with those identities?**

Mapping Race / Ethnicity

Is it possible to predict a character's race by simply looking at the geographic places they say? To test this, we extract location references from all of our dialogue files, and collapse them into eight distinct regions (see Appendix for full list of countries included in each region):

1. Caribbean
2. **East Asia**
3. *Europe*
4. **North Africa and the Middle East**
5. *North America and Australasia*
6. **Latin America**
7. **South Asia**
8. Sub-Saharan Africa

We only report our findings for the bolded locations. For locations in grey, our sample sizes were too small to confidently make any statistical claims. For italicized regions, there was normal variation across all racial and ethnic groups, meaning those places were so ubiquitous that we can not glean meaningful information about individual groups' references.

Our working hypothesis was that when conditioning only on these locations, groups would be much more likely to say a location connected to their racial or ethnic identity – i.e. East Asian characters would be more likely to mention a place in East Asia than a white character. This would suggest to us that non-white characters may be cast in certain roles because the narrative requires that racial or ethnic group as part of the narrative. This is one way to capture how minority characters may be tokenized when they appear in a film, no matter the quartile.

What are the odds that a reference to a geographic location was made by a character who's race or ethnicity is associated with that location, compared to a white character?

30x more likely for a South Asian character to reference South Asia than a white character

TOKENIZATION — the use of a particular body specifically because of the stereotypes that body holds. This type of casting and scripting allow for minorities to appear in film, but instead of representing an individual, the character takes on an entire cultural stereotype. This serves only to reify existing, oppressive typecasts.

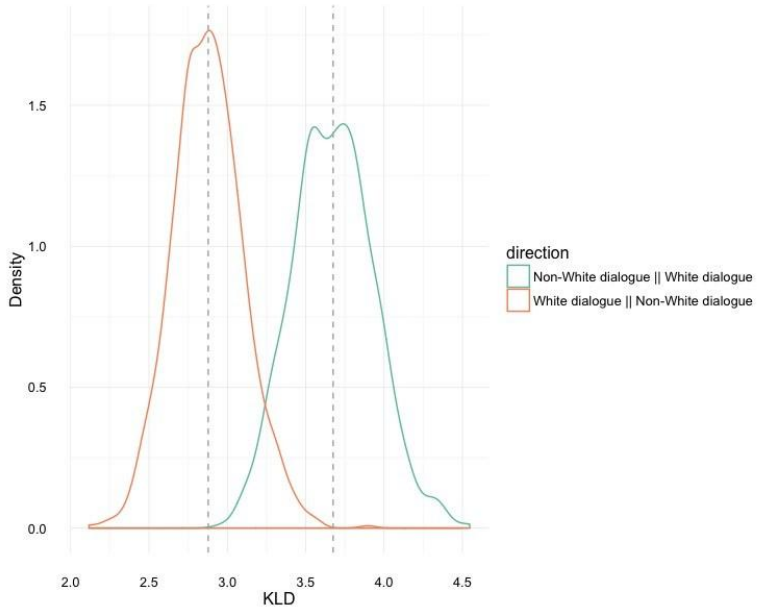
Asymmetries

How similar do white characters and characters of colour sound to each other? Are they drawing from roughly the same distribution of words or is one group marked off as distinctive from the other? If so, how?

Using the measure Kullback-Leibler Divergence, we calculate the difference between the language of White characters and characters of colour. KLD captures the amount of surprise between two probability distributions. A score of 0 would indicate that the two vocabularies are identical, while a higher score indicates greater divergence.

KLD is useful because it allows us to measure difference in two directions. How much more surprising is the language of White characters from the perspective of characters of colour than the other way around?

Figure 6: Distribution of KLD Scores



Sampling 100 characters at a time and running this 1,000 times, we see a very strong difference in the amount of divergence between the two models (Fig. 6).

Contrary to our expectations that characters of colour would simply make up a subset of the speech of white characters, who are a much larger group, we find that the dialogue of White characters is a far worse predictor of the dialogue of characters of colour than the other way around. **Actors of colour are more often cordoned off into separate linguistic zones.** We see this as a form of linguistic segregation, which could lead to the maintenance of cultural stereotypes.

Conclusion

Throughout this research, we have looked at the systematic ways representation functions across 47 years of Hollywood films. We continuously see the same troubling, although perhaps unsurprising, pattern: white characters, in each of our measures, are predominant. This comes at a cost. **Not only does it privilege the narratives and casting of white folks, it also limits the number of visible and audible minorities in these films.**

Fixing this systemic underrepresentation is incredibly complex, and also encompasses issues beyond the scope of this paper. Ultimately, any solution needs to consider not only who is seen and heard in these movies, but how that representation affects our perception of different lived experiences. We hope that our paper demonstrates the valuable role data can play in assessing the cultural practices that have such a profound impact on who we see as projections of ourselves and our society.

Acknowledgements

This project could not have been completed without the generous support of many organizations and people. We would like to thank McGill University and the Social Sciences and Humanities Research council for funding this project.

We would also like to thank Alayne Moody and Beata Skazinetsky for their expertise at the lab. We're equally indebted to all the Research Assistants at .txtLAB who provided us with advice, critiques and snacks while we worked on this. To Professor Richard So for his invaluable insights, and finally to Professor Andrew Piper for his guidance, bad puns and unconditional patience throughout this entire process.

Works Cited

Anderson, Hanah and Matt Daniels. *Film Dialogue: The Largest Ever Analysis of Film Dialogue*. The Pudding. 2016.

Diawara, Manthia. *Black American Cinema*. Routledge, 2012.

Erigha, Maryann. "Race, Gender, Hollywood: Representation in Cultural Production and Digital Media's Potential for Change." *Sociology Compass* 9, no. 1 (2015): 78-89.

Gray, J. R. a. T. (2016). Diversity in Hollywood: Failure of Inclusion Plagues the Entire Industry. *Variety Magazine*. New York City: Variety. Web.

Holiday, Marta. (2017) Halle Berry as the Modernized Tragic Mulatta. *The Projector: A Journal on Film, Media and Culture* 17(1).

Smith, Stacy L., Marc Choueti, Katherine Pieper. "Inequality in 900 Popular Films: Examining Portrayals of Gender, Race/Ethnicity, LGBT, and Disability from 2007-2016" (2017).