

## Project Goals

- ❖ Determine the accuracy of numeric representation of racial and ethnic groups in American movies by number of characters and amount of dialogue
- ❖ Use centrality measures and quality measures in order to better understand the representation that is present in the movies
- ❖ Determine whether dialogue itself is racialized

## Abstract

Previous studies have found that white characters dominate film and television, leaving little room for non-white characters. This overrepresentation of whiteness, however, is not limited to the visual space. Minority characters are even more underrepresented by the amount of lines spoken than by number of characters, and when the dialogue itself is considered, we find that non-white characters are more restricted to a geographical space than white characters. This study centres on three ways to measure, analyze and understand representation in American film.

## Data

- ❖ 855 American films released between 1970 to 2014.
- ❖ 4,188 characters who each say at least 250 words (4,144,200 words total)

With the support and work of

Eve Kraicer & Anne Meisner



SSHRC CRSH

# Underrepresentation of Race in Film:

## The visual and auditory (pre)dominance of white characters in Hollywood

Victoria Svaikovsky (B.A. Linguistics, French Literature) & Andrew Piper (Professor, Director of .txtLAB @ McGill)

.txtLAB  
@mcgill

## 3. Quality

We have shown that this film corpus includes only a very small set of non-white characters. If these roles present any stereotypes, visual or based in dialogue, the effect of the stereotype is augmented by the lack of counterexamples. The quality measure of representation analyzes linguistic variability and dialogue to better understand what roles non-white characters hold when they are included, highlighting the importance of assessing multiple types of representation.

### ❖ Unique Words

A type-token ratio measures diversity of language by taking into account the number of unique words over all words. By randomly sampling 750-word chunks 1,000 times, we found no significant differences between groups except for Indigenous women versus other women (Indigenous women had fewer unique words).

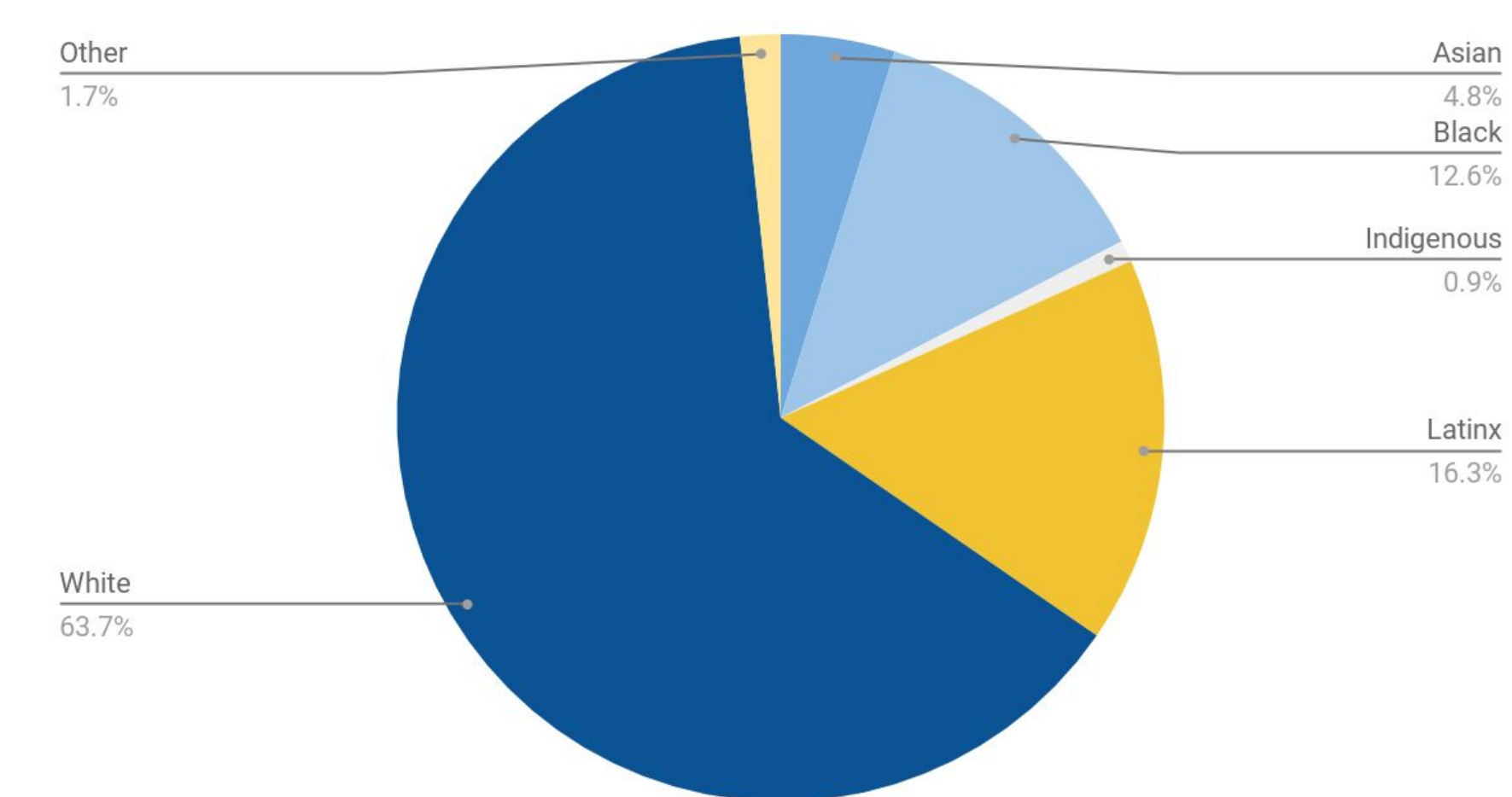
### ❖ Geo-tagging

We were interested to find whether, when minority characters are included in a film, it is done so because the race is important to the film rather than as a neutral choice. Using a Bayesian analysis of references to places and Fisher's Odds Ratio, we found how much more likely a non-white group is to make a reference to a place in the region often associated with the racial or ethnic group than white characters.

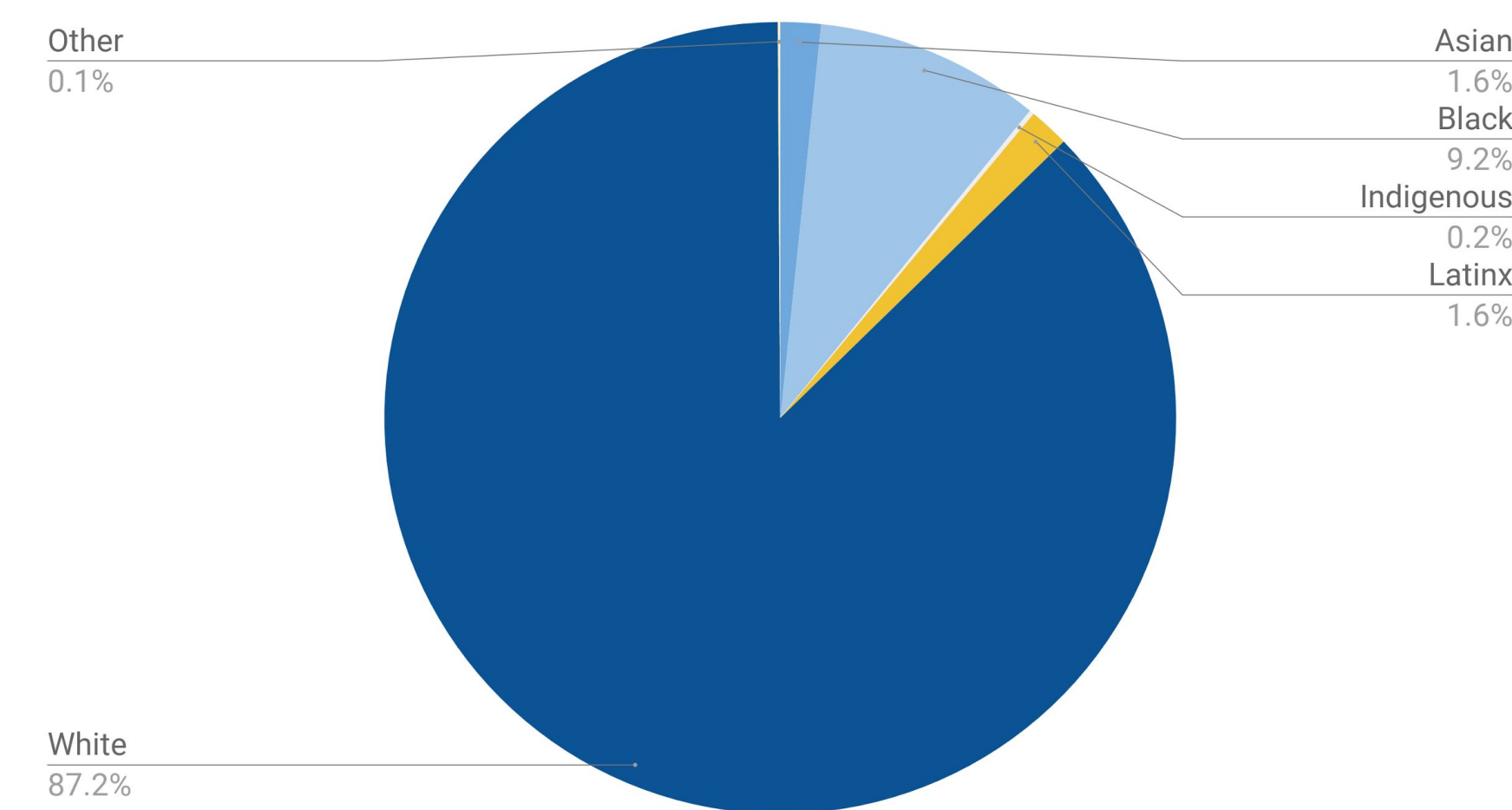
Latinx characters are **2.5 times more likely** to reference Latin America than white characters, East Asian characters are **8 times more likely** to refer to East Asia, South Asian characters are **33 times more likely** to refer to South Asia, and Near Eastern (Middle East + Northern Africa) characters are **43 times more likely** to refer to the Near East.

What does it mean for a group to be represented in film? Previous studies have found that white characters dominate our visual space while non-white characters barely make it onto our screens, with the brunt of the disproportionality carried by hispanic characters<sup>1</sup>. Drawing from the work of Erigha (2015)<sup>2</sup>, we studied three components of representation: numerical representation (visibility and audibility), centrality (the context of the visibility), and quality (to what extent that visibility is a nuanced conception of a character).

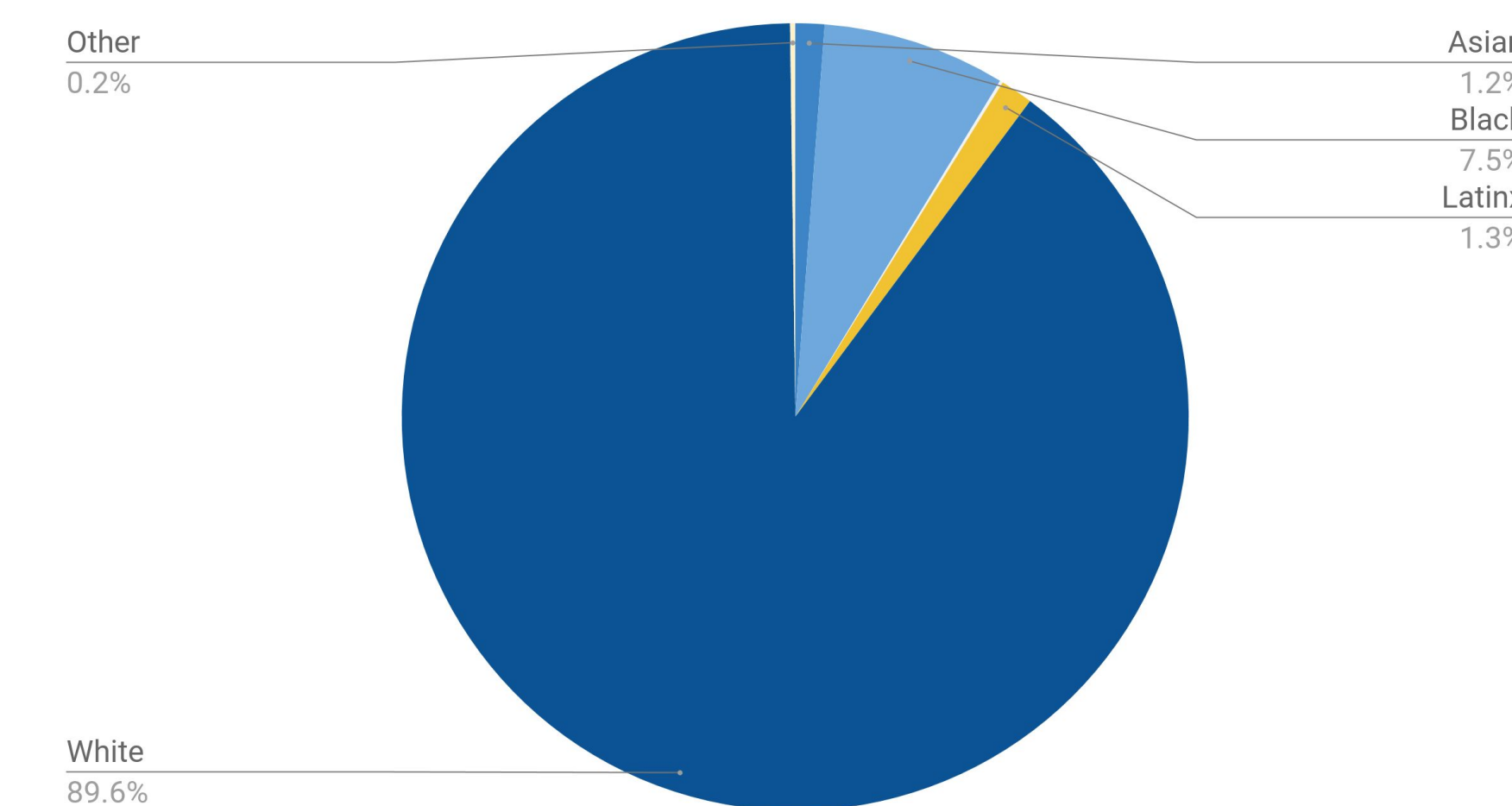
2010 US Census Data



Percent of characters of each group



Percent of words spoken by each group



## 1. Numeric

### Proportionality of Representation

The US population is not evenly divided amongst the ethnic and racial groups that it comprises; therefore, it would be unreasonable to expect the characters in our corpus to divide evenly into those groups. Our intention was to find how representative these films are of the US population as of the 2010 Census. Limited by Census data, we can only take into account five racial and ethnic groups: Asian (East Asian and South Asian together), Black, Indigenous, Latinx, and white.

A Fisher's Odds Ratio determines the proportionality of representation of a group by comparing the number of characters or the number of words spoken to the group's US population size. The tables to the right indicate the odds ratio (how overrepresented a group is) as well as the inverse odds ratio (necessary to understand underrepresentation). An odds ratio of 1 indicates proportional representation.

Both by number of characters and words spoken, white characters are **4-5 times overrepresented** while all other groups are consistently underrepresented. Latinx characters are almost **12 times underrepresented** by number of characters and are more than **14 times underrepresented** by words spoken. Black characters, although always underrepresented, are the closest to proportional representation.

### Gender Split

Women face a larger disparity: white women are **4.5 times overrepresented** by number of characters and **5.5 times overrepresented** by number of words. Asian women: **3.2-4.8 times underrepresented**  
Black women: **1.8-2.4 times underrepresented**  
Indigenous women: **6.1-7.7 times underrepresented**  
Latina women: **11.8 times underrepresented** by characters and **8.1 times underrepresented** by words.

### Part-white measures

In addition to the gender disparity in proportionality between population and words or characters, non-white women are also **4.2 times more likely to be part white** (have one white parent) than non-white men (p-value = 1.78 e-07).

By number of characters:

	# char.	% char.	% pop.	Over rep.	Under rep.	p-value
Asian	69	1.64%	4.8%	0.33	3.01	0
Black	386	9.2%	12.6%	0.70	1.42	7.04 e-12
Indigenous	8	0.19%	0.9%	0.21	4.74	1e-08
Latinx	69	1.65%	16.3%	0.08	11.62	0
White	3,643	87.2%	63.7%	3.61	0.26	0

By words spoken:

	# words	% words	% pop.	Over rep.	Under rep.	p-value
Asian	49,499	1.19%	4.8%	0.24	4.17	0
Black	311,705	7.52%	12.6%	0.56	1.77	0
Indigenous	5,293	0.13%	0.9%	0.14	7.10	0
Latinx	55,387	1.33%	16.3%	0.07	14.37	0
White	3,715,588	89.6%	63.7%	4.91	0.20	0

### "Best" and Worst Genres:

	genre	% non-white characters	genre	% words spoken by a non-white character	
1.	action	19.0%	1.	action	15.3%
2.	drama	15.5%	2.	drama	12.4%
3.	overall	13.0%	3.	overall	10.7%
4.	sci-fi	12.1%	4.	horror	10.3%
5.	crime	11.1%	5.	sci-fi	10.2%
6.	horror	10.8%	6.	crime	9.4%
7.	comedy	8.8%	7.	comedy	7.7%

## Conclusions & Next steps

→ Previous work has found that white characters disproportionately dominate Hollywood, but our study has found that audibly (the number of words spoken), the disproportionality is augmented. When non-white characters are included, they are often relegated to minor, restricted, background roles. The results of the geo-tagging measures indicate that their race or ethnicity is often integral to the role; that is, the decision to include an actor of colour was not neutral and not based solely on the talent of the actor.

→ Our next steps include further linguistic analysis as well as developing a rating system for films based on the diversity of characters and dialogue. We will use a perplexity measure to determine how similarly to a given model each group speaks. For example, if we build a model of criminal language (i.e., from crime TV shows), is the dialogue of one race group more "criminalized" than that of another?

<sup>1</sup>Smith, Stacy L., Marc Choueiri, and Katherine Pieper. "Race/Ethnicity in 600 popular films: Examining on screen portrayals and behind the camera diversity." Media, Diversity, & Social Change Initiative (2014).

<sup>2</sup>Erigha, Maryann. "Race, Gender, Hollywood: Representation in Cultural Production and Digital Media's Potential for Change." *Sociology Compass* 9.1 (2015): 78-89.

## 2. Centrality

The numerical measures demonstrate how disproportionately small the set of non-white characters within our corpus is. Centrality measures serve to better understand where this smaller set resides. We analyzed the top roles and the top pairs of each movie by amount of dialogue.

	# top roles	# of men in group with top role	# of women in group with top role
white	712	542	170
black	62	52	10
s. asian	5	4	1
latinx	3	1	2
e. asian	2	1	1
indigenous	1	1	0
n. east	1	1	0

Out of 855 top roles (the character with the most amount of dialogue in each film), white characters hold 83% of them, which is to be expected with the character distributions we saw earlier.

When analyzing the top two characters of each film (chart right), we found that almost **82% of top pairs were two white characters**, and when there was a mixed pair, **white characters are 1.99 times more likely** to be the top character. It was **142.99 times more likely** that a top pair was 2 white characters than 2 non-white characters.

