Intersectional inequalities in science

Diego Kozlowski[1], Vincent Larivière[2], Cassidy R. Sugimoto[3] and Thema Monroe-White[4]

[1] University of Luxembourg, Luxembourg

[2] Université de Montréal, Montréal, QC, Canada

[3] Georgia Institute of Technology, Atlanta, GA, USA

[2] Berry College, Mount Berry, GA, USA

Mind The Research Gap

Motivation

image by Lina Castellanos/Ecofeminita

- Heart-attack symptoms in women are understudied and more often misdiagnosed by doctors,
- Black women have a higher pregnancy-related mortality rate than White women.
- The effects of COVID vaccines on the menstrual cycle were omitted from the first studies.
- When a specific group is understudied by science, this has direct consequences on the health of that population.
- Why are some topics understudied in science?

Motivation

image by Lina Castellanos/Ecofeminita

- Heart-attack symptoms in women are understudied and more often misdiagnosed by doctors,
- Black women have a higher pregnancy-related mortality rate than White women.
- The effects of COVID vaccines on the menstrual cycle were omitted from the first studies.
- When a specific group is understudied by science, this has direct consequences on the health of that population.
- Why are some topics understudied in science?

Motivation

image by Lina Castellanos/Ecofeminita

- Heart-attack symptoms in women are understudied and more often misdiagnosed by doctors,
- Black women have a higher pregnancy-related mortality rate than White women.
- The effects of COVID vaccines on the menstrual cycle were omitted from the first studies.
- When a specific group is understudied by science, this has direct consequences on the health of that population.
- Why are some topics understudied in science?

Motivation

image by Lina Castellanos/Ecofeminita

- Heart-attack symptoms in women are understudied and more often misdiagnosed by doctors,
- Black women have a higher pregnancy-related mortality rate than White women.
- The effects of COVID vaccines on the menstrual cycle were omitted from the first studies.
- When a specific group is understudied by science, this has direct consequences on the health of that population.
- Why are some topics understudied in science?

Motivation

image by Lina Castellanos/Ecofeminita

- Heart-attack symptoms in women are understudied and more often misdiagnosed by doctors,
- Black women have a higher pregnancy-related mortality rate than White women.
- The effects of COVID vaccines on the menstrual cycle were omitted from the first studies.
- When a specific group is understudied by science, this has direct consequences on the health of that population.
- Why are some topics understudied in science?

Introduction

image by Lina Castellanos/Ecofeminita

- We want to understand the distribution of research topics by race and gender,
- and how research topics that focus on marginalized populations are studied.
- We also want to see the citation gap by race, gender and research topics.
- We will approach these questions with large-scale quantitative tools.

Introduction

image by Lina Castellanos/Ecofeminita

- We want to understand the distribution of research topics by race and gender,
- and how research topics that focus on marginalized populations are studied.
- We also want to see the citation gap by race, gender and research topics.
- We will approach these questions with large-scale quantitative tools.

Introduction

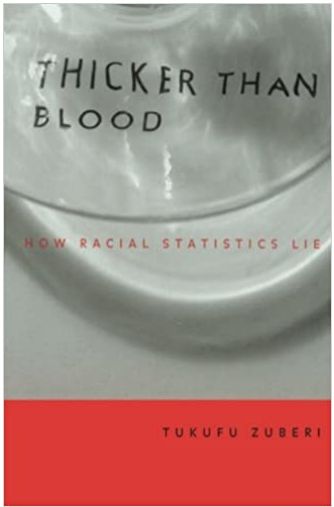
image by Lina Castellanos/Ecofeminita

- We want to understand the distribution of research topics by race and gender,
- and how research topics that focus on marginalized populations are studied.
- We also want to see the citation gap by race, gender and research topics.
- We will approach these questions with large-scale quantitative tools.

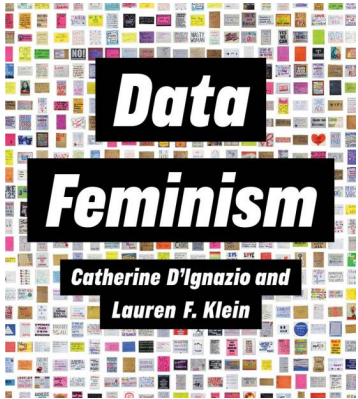
Introduction

image by Lina Castellanos/Ecofeminita

- We want to understand the distribution of research topics by race and gender,
- and how research topics that focus on marginalized populations are studied.
- We also want to see the citation gap by race, gender and research topics.
- We will approach these questions with large-scale quantitative tools.

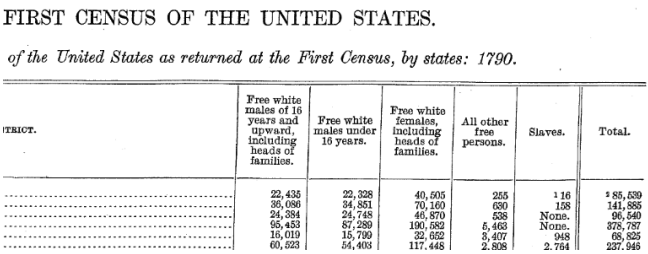


“The racialization of data is an artifact of both the struggles to preserve and to destroy racial stratification.”
(Zuberi 2002)



"Counting and measuring do not always have to be tools of oppression. We can also use them to hold power accountable" (D'Ignazio and Klein 2020)

Census



FIRST CENSUS OF THE UNITED STATES.

of the United States as returned at the First Census, by states: 1790.

ITRICT.	Free white males of 16 years and upward, including heads of families.	Free white males under 16 years.	Free white females, including heads of families.	All other free persons.	Slaves.	Total.
.....	22,435	22,328	40,505	255	[1] 16	[2] 85,539
.....	36,086	34,851	70,160	630	158	141,885
.....	24,384	24,748	46,870	538	None.	96,540
.....	95,453	87,289	190,582	5,463	None.	378,787
.....	16,019	15,799	32,652	5,407	948	68,825
.....	60,523	54,403	117,448	2,808	2,764	237,946

- The census is one of those cases where structural racism gets codified. For example, the first US census (1790) counting "Slaves", "Free White" and "Other free person",
- nevertheless, it is a tool that can be use to help us understand inequalities, in the struggle to overcome them.

Bibliometric Databases

- There is a lot of qualitative evidence of structural discrimination in academia,
- but large scale bibliometric analysis has the power to put a number to those experiences lived by women and racialized people.
- Bibliometric databases don't have information on author's self-perceived race,
- If we could know authors' race, we would be able to unravel multiple dimensions structural racism in academia.

Bibliometric Databases

- There is a lot of qualitative evidence of structural discrimination in academia,
- but large scale bibliometric analysis has the power to put a number to those experiences lived by women and racialized people.
- Bibliometric databases don't have information on author's self-perceived race,
- If we could know authors' race, we would be able to unravel multiple dimensions structural racism in academia.

Bibliometric Databases

- There is a lot of qualitative evidence of structural discrimination in academia,
- but large scale bibliometric analysis has the power to put a number to those experiences lived by women and racialized people.
- Bibliometric databases don't have information on author's self-perceived race,
- If we could know authors' race, we would be able to unravel multiple dimensions structural racism in academia.

Bibliometric Databases

- There is a lot of qualitative evidence of structural discrimination in academia,
- but large scale bibliometric analysis has the power to put a number to those experiences lived by women and racialized people.
- Bibliometric databases don't have information on author's self-perceived race,
- If we could know authors' race, we would be able to unravel multiple dimensions structural racism in academia.

Data Sources

US Census

- 2010 US census data of Family names and racial groups.
- All names with more that 100 appearances,
- 162 253 names (90% of the population).

Web of Science

- WOS US-based first authors,
- between 2008-2019,
- 5.4M articles, 1.6M first authors.

We focus on US because racial categories are a social construction, that can only have meaning in a specific context.

Data & Methods

- Using authors' names to infer their probable race is both a powerful and a limited technique.
- **powerful:** because it opens the possibility of multiple large-scale analysis on racism,
- **limited:** because it cannot account for small populations, reinforcing their invisibilization,

Data & Methods

- Using authors' names to infer their probable race is both a powerful and a limited technique.
- **powerful:** because it opens the possibility of multiple large-scale analysis on racism,
- **limited:** because it cannot account for small populations, reinforcing their invisibilization,

Data & Methods

- Using authors' names to infer their probable race is both a powerful and a limited technique.
- **powerful:** because it opens the possibility of multiple large-scale analysis on racism,
- **limited:** because it cannot account for small populations, reinforcing their invisibilization,

Racial Categories

- Non-Hispanic White Alone (*White*)
- Non-Hispanic Black or African American Alone (*Black*)
- Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander Alone (*Asian*)
- Non-Hispanic American Indian and Alaska Native Alone (*AIAN*)
- Non-Hispanic Two or More Races (*Two or more*)
- Hispanic or Latino origin (*Hispanic/Latinx*)

Limitation

- Given that the method relies on statistical significance of the association between names and groups, smaller groups cannot be properly inferred.
- *AIAN* and *Two or more* account for 0.69% and 1.76% of the WOS authors.
- Excluding racial groups is major limitation, as it reinforces invisibilization.

Racial Categories

- Non-Hispanic White Alone (*White*)
- Non-Hispanic Black or African American Alone (*Black*)
- Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander Alone (*Asian*)
- Non-Hispanic American Indian and Alaska Native Alone (*AIAN*)
- Non-Hispanic Two or More Races (*Two or more*)
- Hispanic or Latino origin (*Hispanic/Latinx*)

Limitation

- Given that the method relies on statistical significance of the association between names and groups, smaller groups cannot be properly inferred.
- *AIAN* and *Two or more* account for 0.69% and 1.76% of the WOS authors.
- Excluding racial groups is major limitation, as it reinforces invisibilization.

Racial Categories

- Non-Hispanic White Alone (*White*)
- Non-Hispanic Black or African American Alone (*Black*)
- Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander Alone (*Asian*)
- Non-Hispanic American Indian and Alaska Native Alone (*AIAN*)
- Non-Hispanic Two or More Races (*Two or more*)
- Hispanic or Latino origin (*Hispanic/Latinx*)

Limitation

- Given that the method relies on statistical significance of the association between names and groups, smaller groups cannot be properly inferred.
- *AIAN* and *Two or more* account for 0.69% and 1.76% of the WOS authors.
- Excluding racial groups is major limitation, as it reinforces invisibilization.

Racial Categories

- Non-Hispanic White Alone (*White*)
- Non-Hispanic Black or African American Alone (*Black*)
- Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander Alone (*Asian*)
- Non-Hispanic American Indian and Alaska Native Alone (*AIAN*)
- Non-Hispanic Two or More Races (*Two or more*)
- Hispanic or Latino origin (*Hispanic/Latinx*)

Limitation

- Given that the method relies on statistical significance of the association between names and groups, smaller groups cannot be properly inferred.
- *AIAN* and *Two or more* account for 0.69% and 1.76% of the WOS authors.
- Excluding racial groups is major limitation, as it reinforces invisibilization.

Racial Categories

- Non-Hispanic White Alone (*White*)
- Non-Hispanic Black or African American Alone (*Black*)
- Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander Alone (*Asian*)
- Non-Hispanic American Indian and Alaska Native Alone (*AIAN*)
- Non-Hispanic Two or More Races (*Two or more*)
- Hispanic or Latino origin (*Hispanic/Latinx*)

Limitation

- Given that the method relies on statistical significance of the association between names and groups, smaller groups cannot be properly inferred.
- *AIAN* and *Two or more* account for 0.69% and 1.76% of the WOS authors.
- Excluding racial groups is major limitation, as it reinforces invisibilization.

Racial Categories

- Non-Hispanic White Alone (*White*)
- Non-Hispanic Black or African American Alone (*Black*)
- Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander Alone (*Asian*)
- Non-Hispanic American Indian and Alaska Native Alone (*AIAN*)
- Non-Hispanic Two or More Races (*Two or more*)
- Hispanic or Latino origin (*Hispanic/Latinx*)

Limitation

- Given that the method relies on statistical significance of the association between names and groups, smaller groups cannot be properly inferred.
- *AIAN* and *Two or more* account for 0.69% and 1.76% of the WOS authors.
- Excluding racial groups is major limitation, as it reinforces invisibilization.

Racial Categories

- Non-Hispanic White Alone (*White*)
- Non-Hispanic Black or African American Alone (*Black*)
- Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander Alone (*Asian*)
- Non-Hispanic American Indian and Alaska Native Alone (*AIAN*)
- Non-Hispanic Two or More Races (*Two or more*)
- Hispanic or Latino origin (*Hispanic/Latinx*)

Limitation

- Given that the method relies on statistical significance of the association between names and groups, smaller groups cannot be properly inferred.
- *AIAN* and *Two or more* account for 0.69% and 1.76% of the WOS authors.
- Excluding racial groups is major limitation, as it reinforces invisibilization.

Racial Categories

- Non-Hispanic White Alone (*White*)
- Non-Hispanic Black or African American Alone (*Black*)
- Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander Alone (*Asian*)
- Non-Hispanic American Indian and Alaska Native Alone (*AIAN*)
- Non-Hispanic Two or More Races (*Two or more*)
- Hispanic or Latino origin (*Hispanic/Latinx*)

Limitation

- Given that the method relies on statistical significance of the association between names and groups, smaller groups cannot be properly inferred.
- *AIAN* and *Two or more* account for 0.69% and 1.76% of the WOS authors.
- Excluding racial groups is major limitation, as it reinforces invisibilization.

Racial Categories

- Non-Hispanic White Alone (*White*)
- Non-Hispanic Black or African American Alone (*Black*)
- Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander Alone (*Asian*)
- Non-Hispanic American Indian and Alaska Native Alone (*AIAN*)
- Non-Hispanic Two or More Races (*Two or more*)
- Hispanic or Latino origin (*Hispanic/Latinx*)

Limitation

- Given that the method relies on statistical significance of the association between names and groups, smaller groups cannot be properly inferred.
- *AIAN* and *Two or more* account for 0.69% and 1.76% of the WOS authors.
- Excluding racial groups is major limitation, as it reinforces invisibilization.

Gender Categories

- Our method also relies on census data and given names to infer the gender of authors,
- we can only infer gender in a binary way, which is another important limitation of our research.
- We do not have any automatic solution for these limitations. Surveys with self-identified race and gender are key to move forward into a more comprehensive discussion.
- Individual research groups we can only partially solve this, with very specific case studies,
- but institutions -such as universities and publishers- can do large scale surveys to address this.

Gender Categories

- Our method also relies on census data and given names to infer the gender of authors,
- we can only infer gender in a binary way, which is another important limitation of our research.
- We do not have any automatic solution for these limitations. Surveys with self-identified race and gender are key to move forward into a more comprehensive discussion.
- Individual research groups we can only partially solve this, with very specific case studies,
- but institutions -such as universities and publishers- can do large scale surveys to address this.

Gender Categories

- Our method also relies on census data and given names to infer the gender of authors,
- we can only infer gender in a binary way, which is another important limitation of our research.
- We do not have any automatic solution for these limitations. Surveys with self-identified race and gender are key to move forward into a more comprehensive discussion.
- Individual research groups we can only partially solve this, with very specific case studies,
- but institutions -such as universities and publishers- can do large scale surveys to address this.

Gender Categories

- Our method also relies on census data and given names to infer the gender of authors,
- we can only infer gender in a binary way, which is another important limitation of our research.
- We do not have any automatic solution for these limitations. Surveys with self-identified race and gender are key to move forward into a more comprehensive discussion.
- Individual research groups we can only partially solve this, with very specific case studies,
- but institutions -such as universities and publishers- can do large scale surveys to address this.

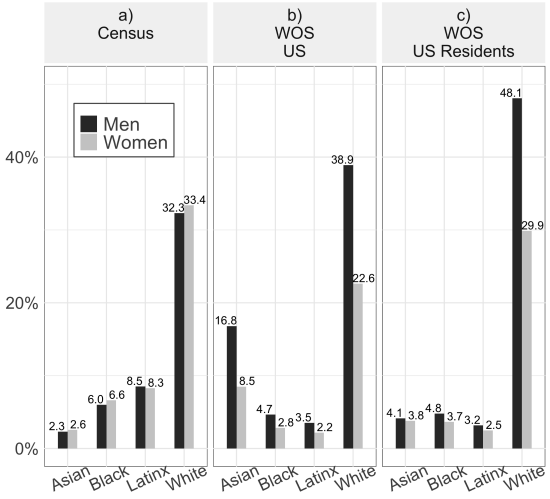
Gender Categories

- Our method also relies on census data and given names to infer the gender of authors,
- we can only infer gender in a binary way, which is another important limitation of our research.
- We do not have any automatic solution for these limitations. Surveys with self-identified race and gender are key to move forward into a more comprehensive discussion.
- Individual research groups we can only partially solve this, with very specific case studies,
- but institutions -such as universities and publishers- can do large scale surveys to address this.

Results

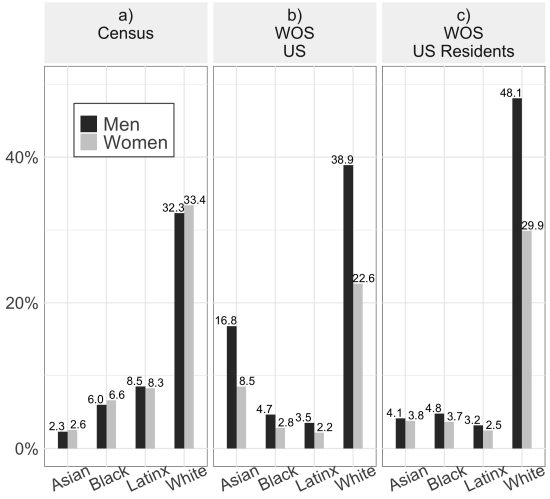
image by Lina Castellanos/Ecofeminita

General distribution



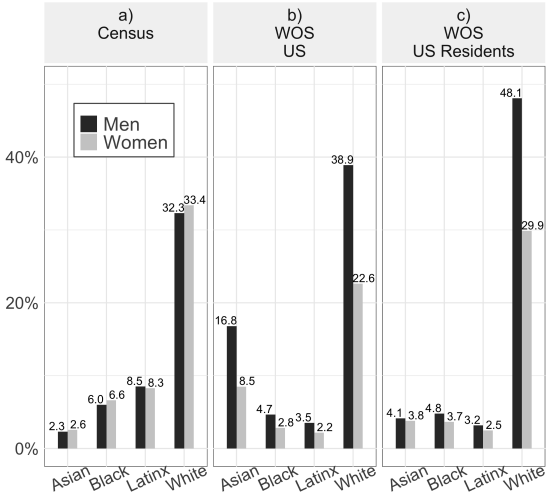
- Overall, there is an over representation of White and Asian men,
- a large proportion of Asian authors are not US residents,
- i.e. the Census is not a perfect benchmark to define *over representation* of this group,
- **women** are underrepresented,
- **Black and Latinx** are underrepresented,
- **Black and Latinx women** are the most underrepresented.

General distribution



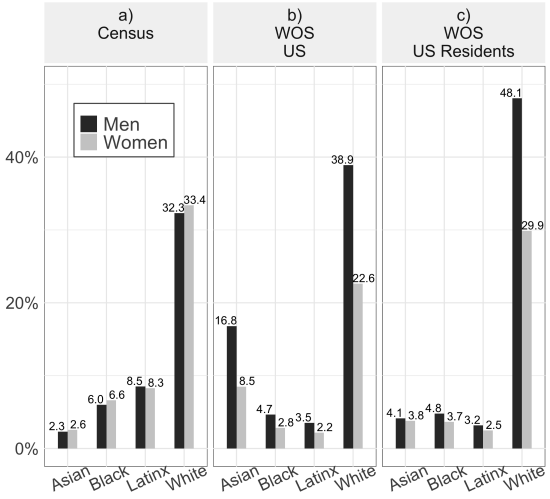
- Overall, there is an over representation of White and Asian men,
- a large proportion of Asian authors are not US residents,
- i.e. the Census is not a perfect benchmark to define *over representation* of this group,
- **women** are underrepresented,
- **Black and Latinx** are underrepresented,
- **Black and Latinx women** are the most underrepresented.

General distribution



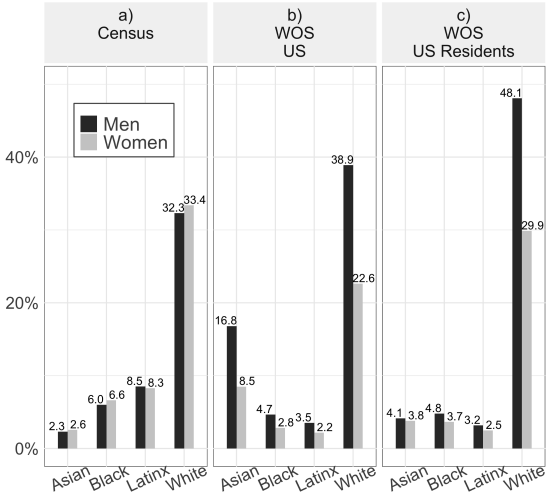
- Overall, there is an over representation of White and Asian men,
- a large proportion of Asian authors are not US residents,
- i.e. the Census is not a perfect benchmark to define *over representation* of this group,
- **women** are underrepresented,
- **Black and Latinx** are underrepresented,
- **Black and Latinx women** are the most underrepresented.

General distribution



- Overall, there is an over representation of White and Asian men,
- a large proportion of Asian authors are not US residents,
- i.e. the Census is not a perfect benchmark to define *over representation* of this group,
- **women** are underrepresented,
- **Black and Latinx** are underrepresented,
- **Black and Latinx women** are the most underrepresented.

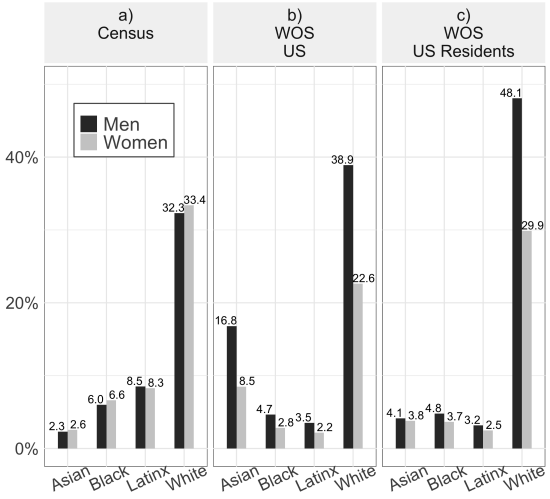
General distribution



- Overall, there is an over representation of White and Asian men,
- a large proportion of Asian authors are not US residents,
- i.e. the Census is not a perfect benchmark to define *over representation* of this group,

- **women** are underrepresented,
- **Black and Latinx** are underrepresented,
- **Black and Latinx women** are the most underrepresented.

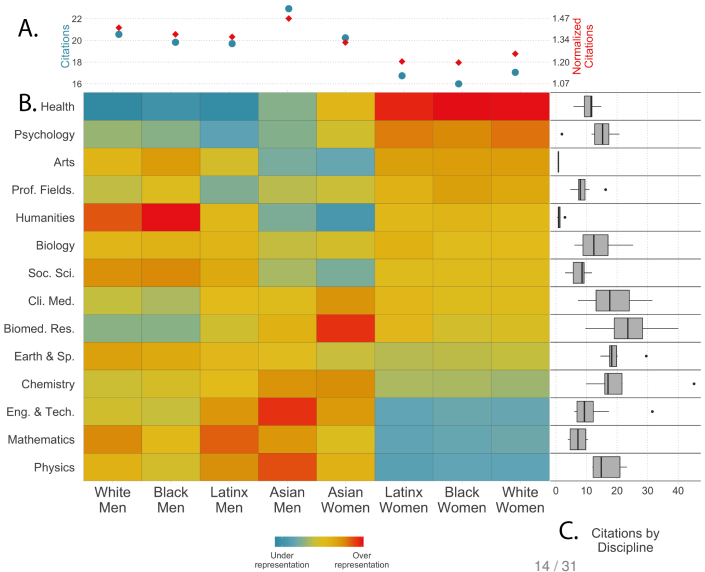
General distribution



- Overall, there is an over representation of White and Asian men,
- a large proportion of Asian authors are not US residents,
- i.e. the Census is not a perfect benchmark to define *over representation* of this group,

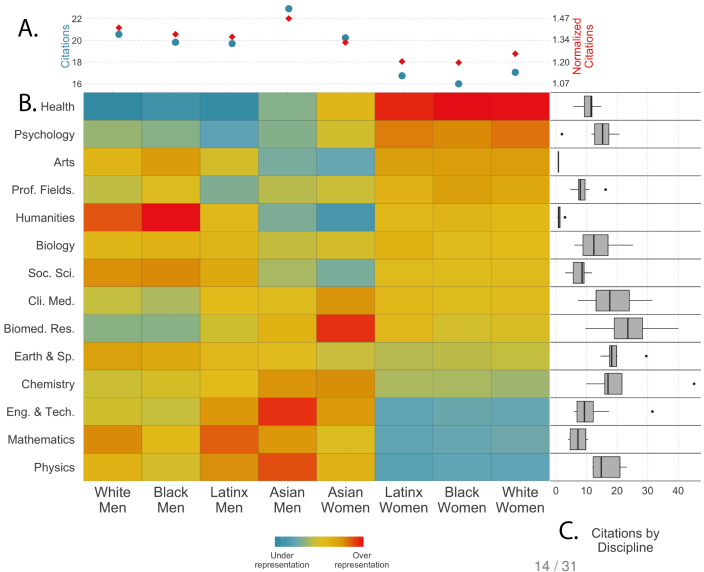
- **women** are underrepresented,
- **Black and Latinx** are underrepresented,
- **Black and Latinx women** are the most underrepresented.

Disciplines' Heterogeneity



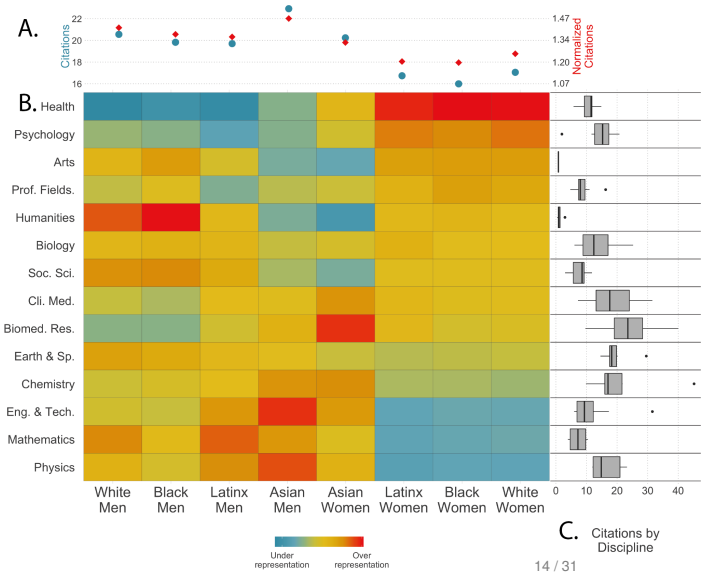
- The distribution by disciplines has a clear pattern by gender, except on Asian authors.
- Women have less citations on average,
- field normalization on citations reduces the gap, but not completely.

Disciplines' Heterogeneity



- The distribution by disciplines has a clear pattern by gender, except on Asian authors.
- Women have less citations on average,
- field normalization on citations reduces the gap, but not completely.

Disciplines' Heterogeneity



- The distribution by disciplines has a clear pattern by gender, except on Asian authors.
- Women have less citations on average,
- field normalization on citations reduces the gap, but not completely.

Disciplines' Heterogeneity

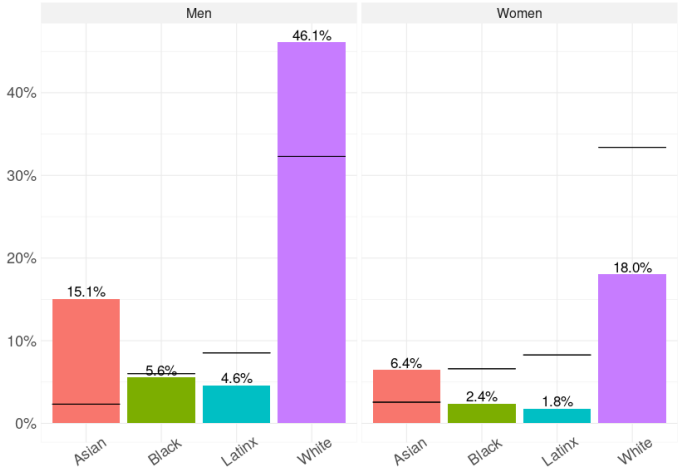


Figure: Surgery

As an example, White and Asian men are overrepresented in Surgery

you can explore other examples at `sciencebias.uni.lu/app`

Disciplines' Heterogeneity

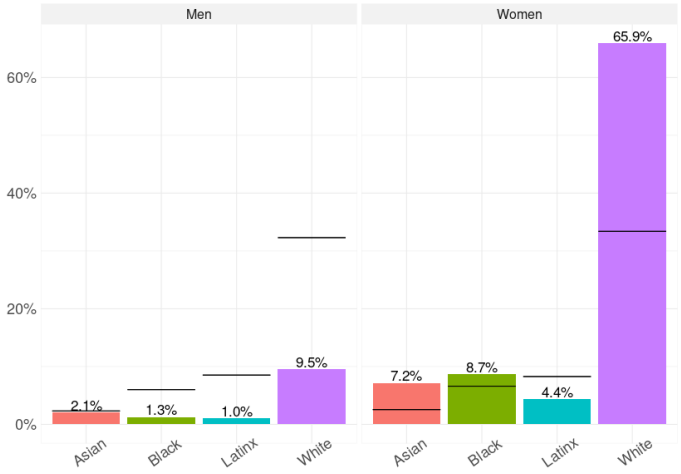


Figure: Nursing

- Women are overrepresented on Nursing
- Black and Latinx authors are usually underrepresented in all fields

Disciplines' Heterogeneity

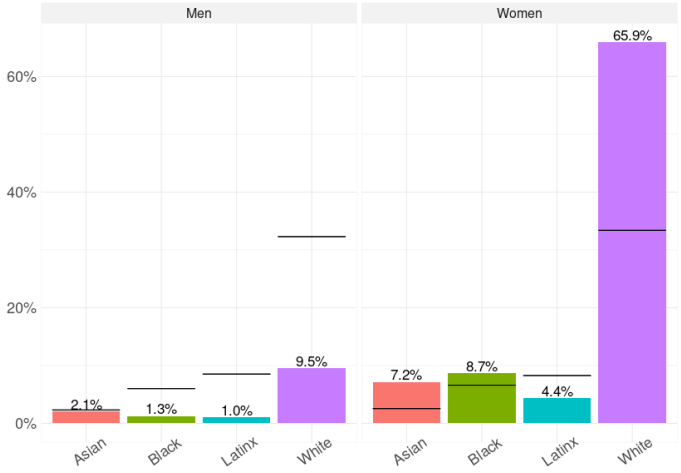


Figure: Nursing

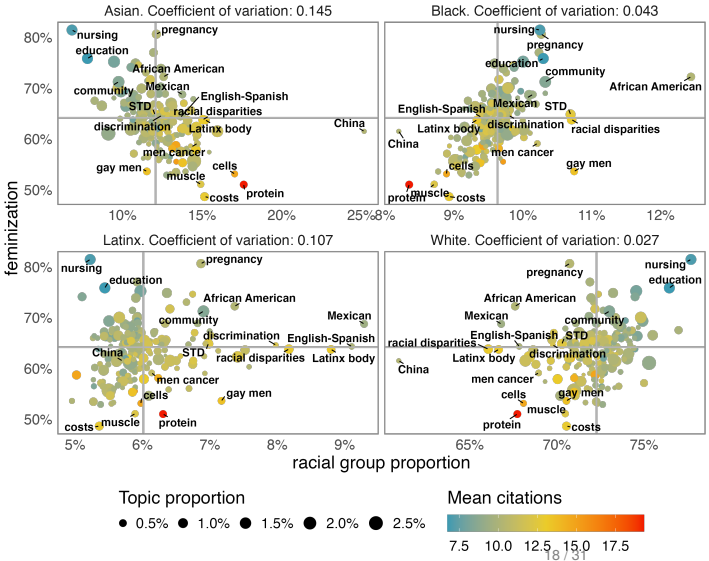
- Women are overrepresented on Nursing
- Black and Latinx authors are usually underrepresented in all fields

- We want to go deeper into the correlation of research topics with race & gender,
- that is why we focus on Health (and Social Sciences), and define research topics using LDA (Blei, Ng, and Jordan 2003).
- For each race & gender we define the average participation on each topic.

- We want to go deeper into the correlation of research topics with race & gender,
- that is why we focus on Health (and Social Sciences), and define research topics using LDA (Blei, Ng, and Jordan 2003).
- For each race & gender we define the average participation on each topic.

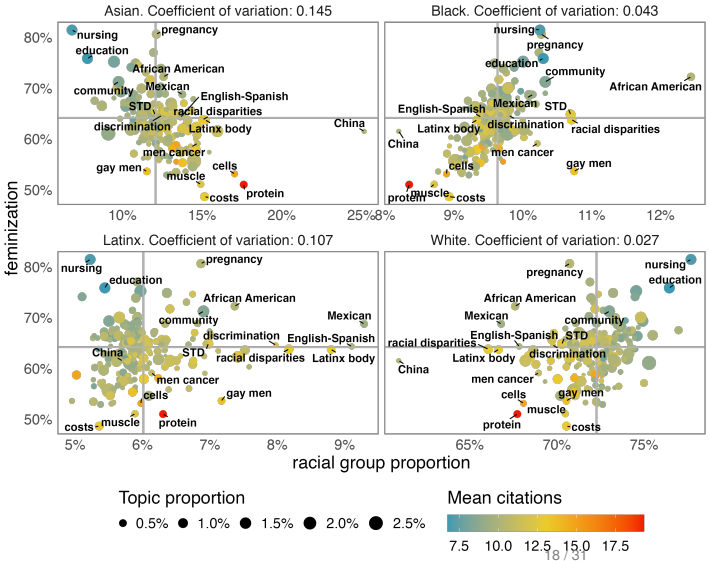
- We want to go deeper into the correlation of research topics with race & gender,
- that is why we focus on Health (and Social Sciences), and define research topics using LDA (Blei, Ng, and Jordan 2003).
- For each race & gender we define the average participation on each topic.

Research topics - Health



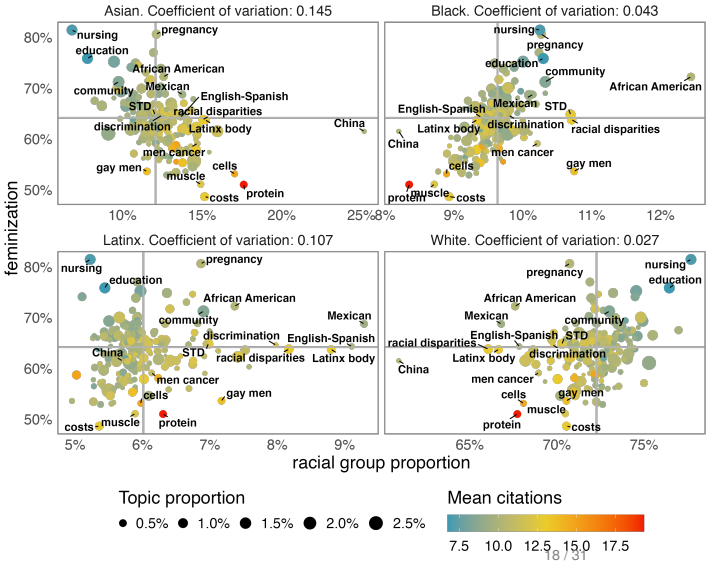
- Women publish more on *nursing*, *pregnancy* and *education*,
- Black authors focus on *African American* and *racial disparities* studies,
- Latinx authors focus on *Mexican* and *Latinx body* studies, but also on language issues,
- Asian authors focus on *China*, while White authors are more evenly distributed across all topics.

Research topics - Health



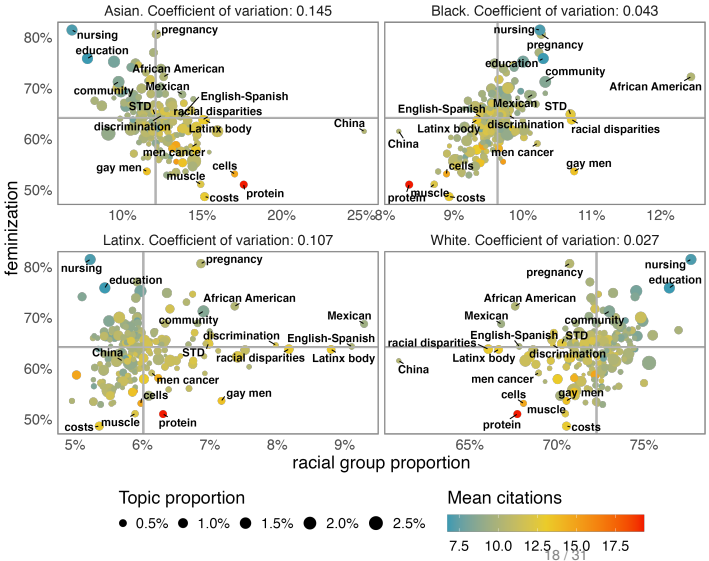
- Women publish more on *nursing*, *pregnancy* and *education*,
- Black authors focus on *African American* and *racial disparities* studies,
- Latinx authors focus on *Mexican* and *Latinx body* studies, but also on language issues,
- Asian authors focus on *China*, while White authors are more evenly distributed across all topics.

Research topics - Health



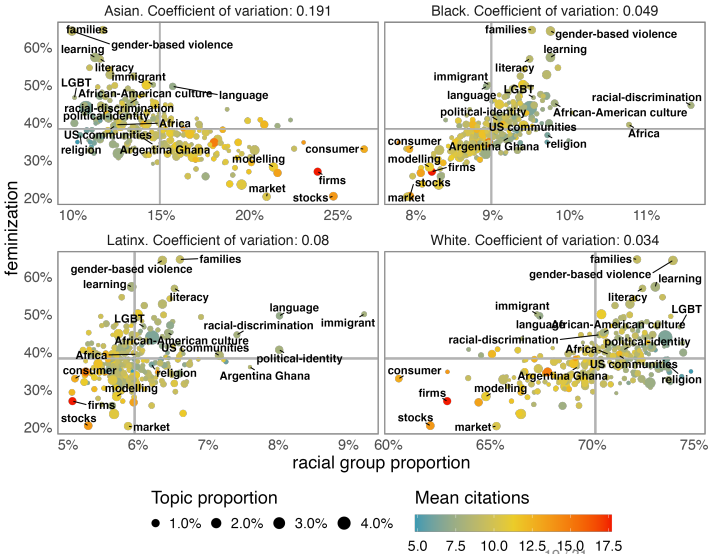
- Women publish more on *nursing*, *pregnancy* and *education*,
- Black authors focus on *African American* and *racial disparities* studies,
- Latinx authors focus on *Mexican* and *Latinx body* studies, but also on language issues,
- Asian authors focus on *China*, while White authors are more evenly distributed across all topics.

Research topics - Health



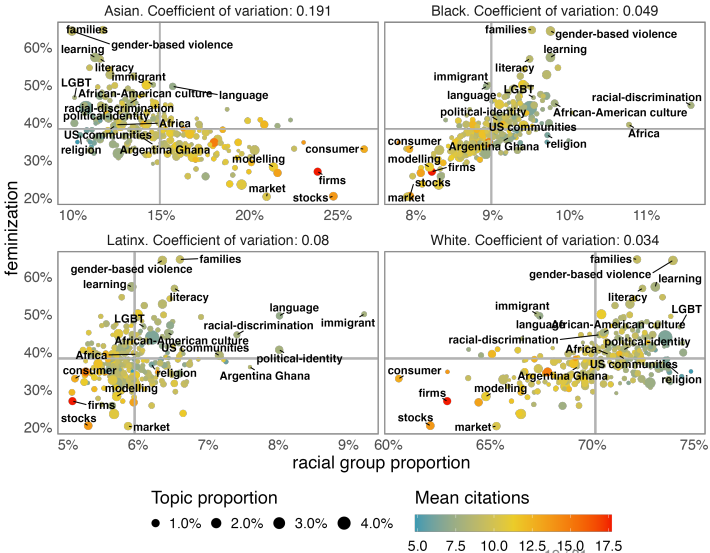
- Women publish more on *nursing*, *pregnancy* and *education*,
- Black authors focus on *African American* and *racial disparities* studies,
- Latinx authors focus on *Mexican* and *Latinx body* studies, but also on language issues,
- Asian authors focus on *China*, while White authors are more evenly distributed across all topics.

Research topics - Social Sciences



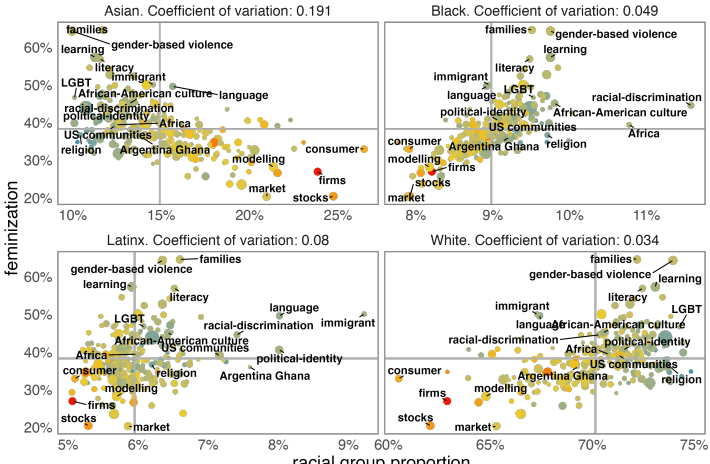
- Women publish more on *families*, *gender-based violence* and *learning*,
- Black authors focus on *racial discrimination*, *African-American culture* and *Africa*,
- Latinx authors focus on *migration*, *language* and *political identity*,
- Asian authors focus on *economics*, while White authors are (again) more evenly distributed across all topics.

Research topics - Social Sciences



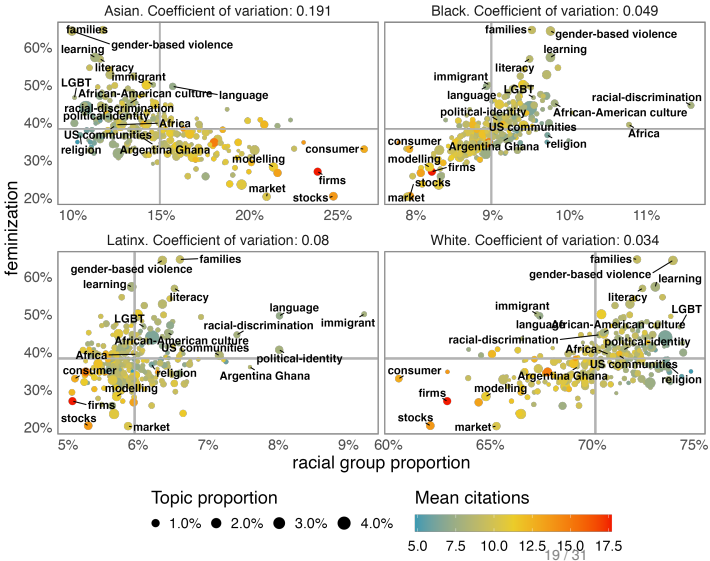
- Women publish more on *families*, *gender-based violence* and *learning*,
- Black authors focus on *racial discrimination*, *African-American culture* and *Africa*,
- Latinx authors focus on *migration*, *language* and *political identity*,
- Asian authors focus on *economics*, while White authors are (again) more evenly distributed across all topics.

Research topics - Social Sciences



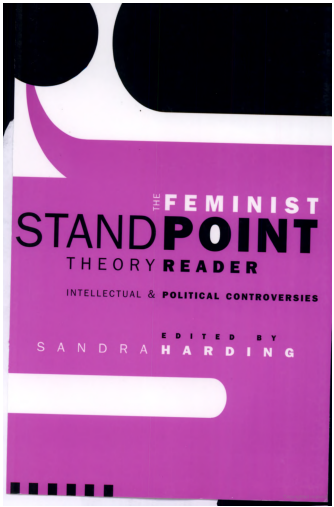
- Women publish more on *families*, *gender-based violence* and *learning*,
- Black authors focus on *racial discrimination*, *African-American culture* and *Africa*,
- Latinx authors focus on *migration*, *language* and *political identity*,
- Asian authors focus on *economics*, while White authors are (again) more evenly distributed across all topics.

Research topics - Social Sciences



- Women publish more on *families*, *gender-based violence* and *learning*,
- Black authors focus on *racial discrimination*, *African-American culture* and *Africa*,
- Latinx authors focus on *migration*, *language* and *political identity*,
- Asian authors focus on *economics*, while White authors are (again) more evenly distributed across all topics.

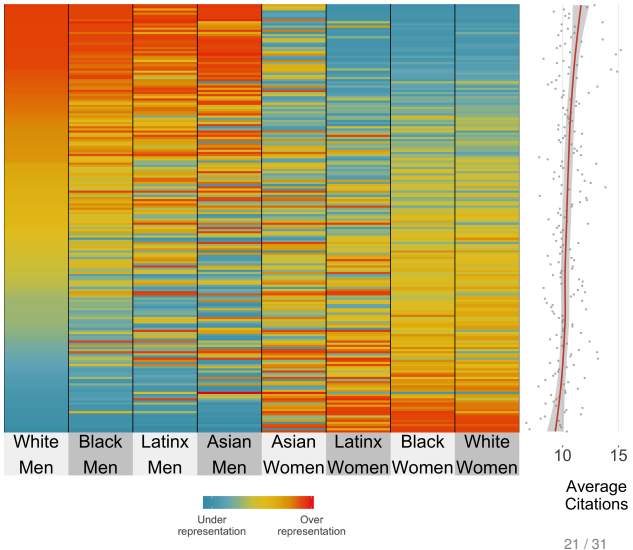
Standpoint



(Harding 2004)

- *"Standpoint theories map how a social and political disadvantage can be turned into an epistemological, scientific, and political advantage."*
- *"It cannot be overemphasized that the epistemic privilege oppressed groups possess is by no means automatic."*
- On the topics distribution, we find a mix of *gender stereotypes* and topics that are most dare to specific communities,
- If White authors have no specific topics, is because they do not hold any epistemic privilege,
- but other types of economic and symbolic privileges, which gives them a general advantage over all topics.

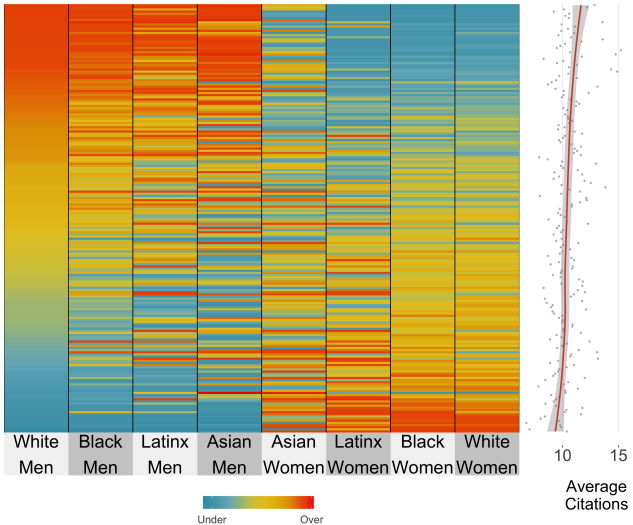
Topics and citations - Health



How do these topics affect citations gaps?

- If we sort topics by White men's participation, this positively correlates with the average number of citations by topic.
- This means that White men tend to do research on more highly cited topics.
- We can also see the gender patterns across research topics.

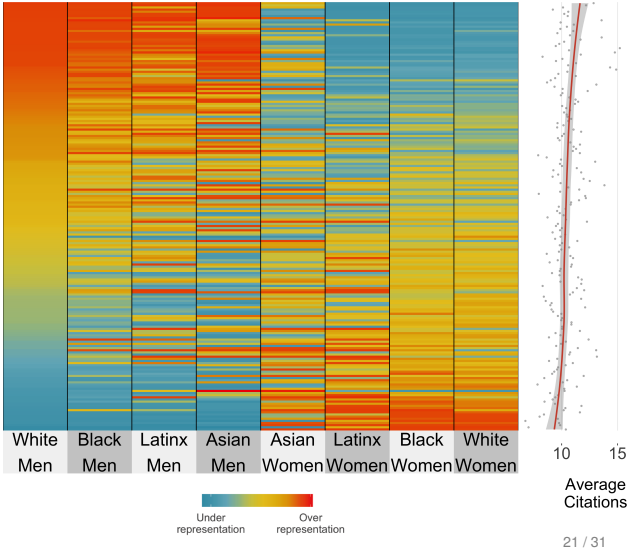
Topics and citations - Health



How do these topics affect citations gaps?

- If we sort topics by White men's participation, this positively correlates with the average number of citations by topic.
- This means that White men tend to do research on more highly cited topics.
- We can also see the gender patterns across research topics.

Topics and citations - Health

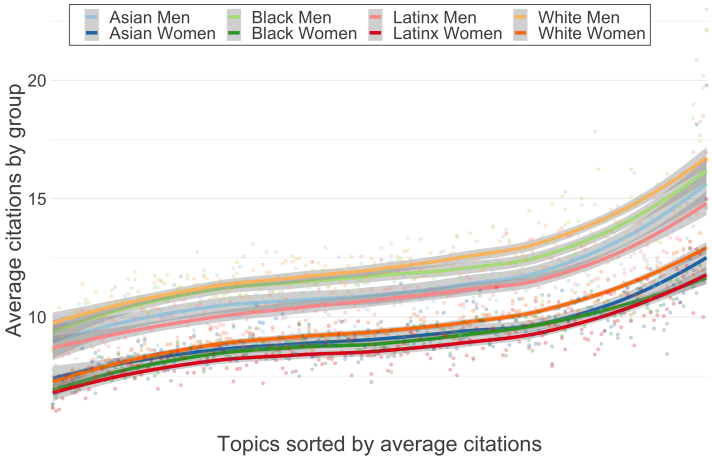


How do these topics affect citations gaps?

- If we sort topics by White men's participation, this positively correlates with the average number of citations by topic.
- This means that White men tend to do research on more highly cited topics.
- We can also see the gender patterns across research topics.

Topics and citations - Health

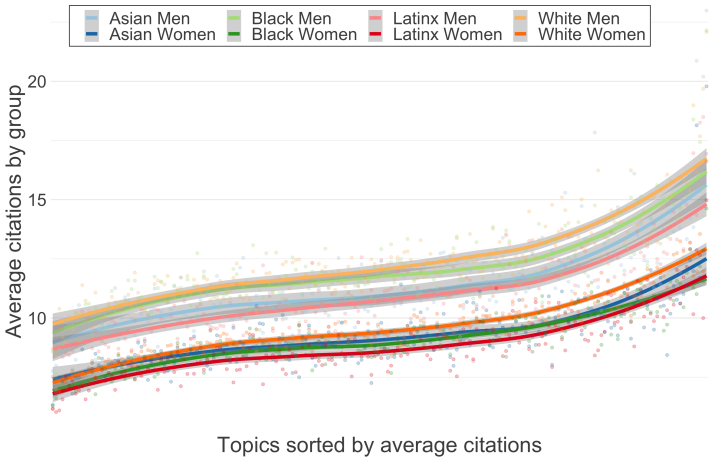
Are citations gaps exclusively explained by research topics?



- If we sort topics by their average citations, and model the citation distribution by groups, we can see that
- Men are more cited in both high and low-cited topics.

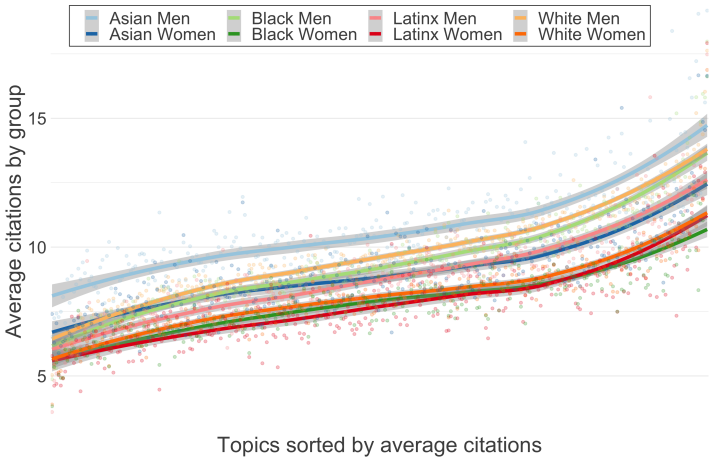
Topics and citations - Health

Are citations gaps exclusively explained by research topics?



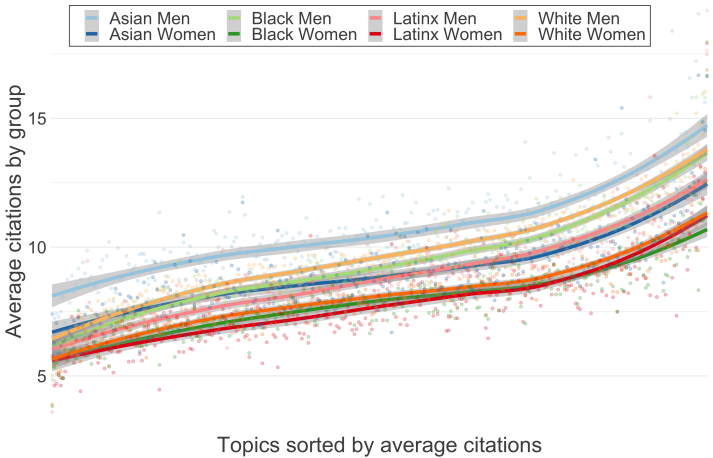
- If we sort topics by their average citations, and model the citation distribution by groups, we can see that
- Men are more cited in both high and low-cited topics.

Topics and citations - Social Sciences



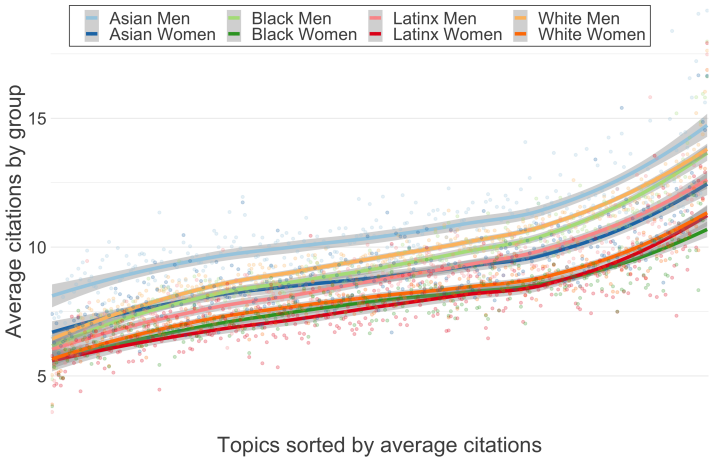
- In the case of Social Sciences, Asian men are have more citations across topics,
- then White and Black men,
- Latinx men and Asian women,
- White, Black and Latinx women have the less citations across topics.
- There is both a inter-topic and intra-topic bias.

Topics and citations - Social Sciences



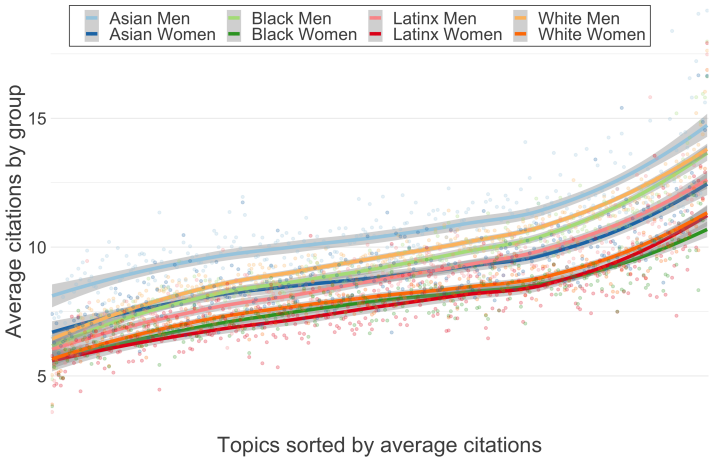
- In the case of Social Sciences, Asian men are have more citations across topics,
- then White and Black men,
- Latinx men and Asian women,
- White, Black and Latinx women have the less citations across topics.
- There is both a inter-topic and intra-topic bias.

Topics and citations - Social Sciences



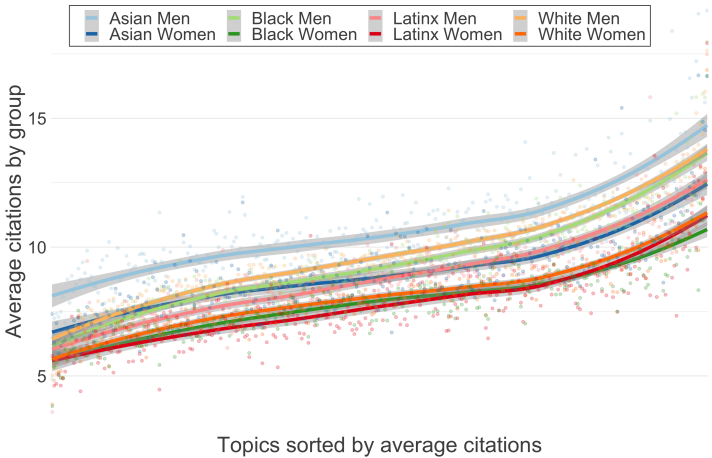
- In the case of Social Sciences, Asian men are have more citations across topics,
- then White and Black men,
- Latinx men and Asian women,
- White, Black and Latinx women have the less citations across topics.
- There is both a inter-topic and intra-topic bias.

Topics and citations - Social Sciences



- In the case of Social Sciences, Asian men are have more citations across topics,
- then White and Black men,
- Latinx men and Asian women,
- White, Black and Latinx women have the less citations across topics.
- There is both a inter-topic and intra-topic bias.

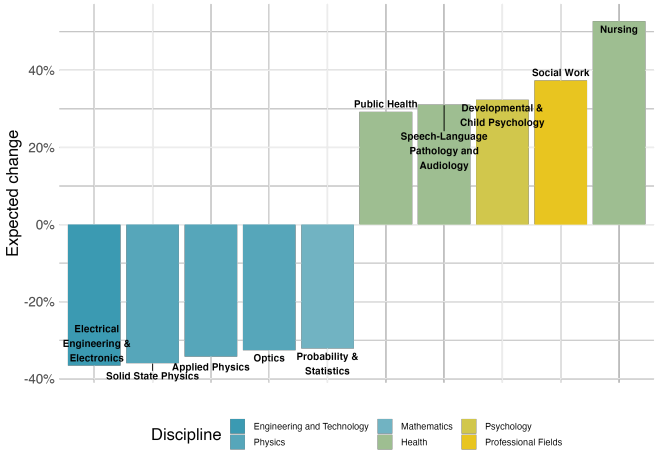
Topics and citations - Social Sciences



- In the case of Social Sciences, Asian men are have more citations across topics,
- then White and Black men,
- Latinx men and Asian women,
- White, Black and Latinx women have the less citations across topics.
- There is both a inter-topic and intra-topic bias.

Counterfactual analysis

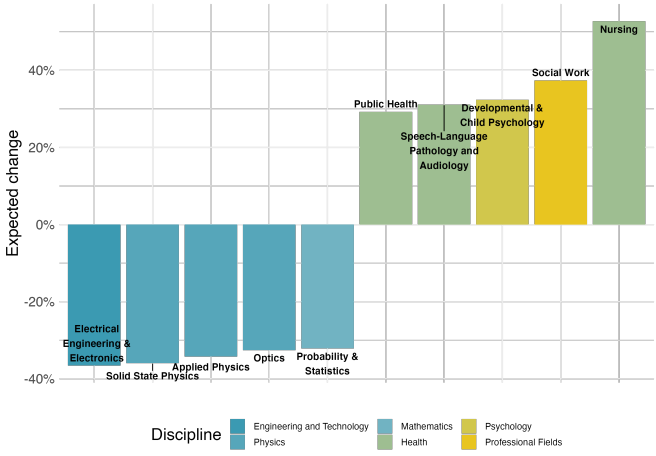
What would happen if the proportion of authors by race & gender was that of the 2010 US Census?



- Assuming constant productivity, this figure shows the expected cumulative change in number of publications that would occur by discipline,
- a 53% more articles in Nursing, 37.4% more in Social Work, and 29% in Public Health,
- while Engineering, Physics and Math would have the greatest decreases.

Counterfactual analysis

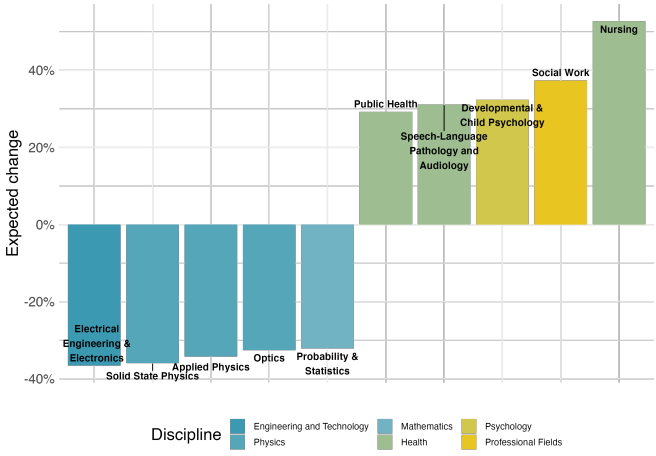
What would happen if the proportion of authors by race & gender was that of the 2010 US Census?



- Assuming constant productivity, this figure shows the expected cumulative change in number of publications that would occur by discipline,
- a 53% more articles in Nursing, 37.4% more in Social Work, and 29% in Public Health,
- while Engineering, Physics and Math would have the greatest decreases.

Counterfactual analysis

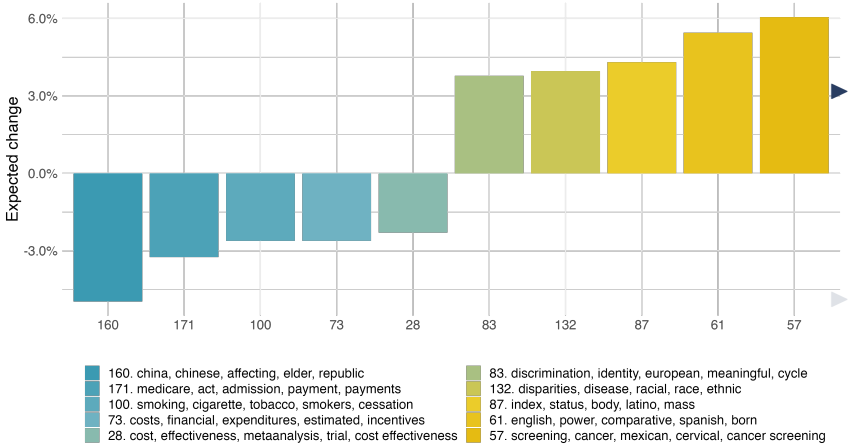
What would happen if the proportion of authors by race & gender was that of the 2010 US Census?



- Assuming constant productivity, this figure shows the expected cumulative change in number of publications that would occur by discipline,
- a 53% more articles in Nursing, 37.4% more in Social Work, and 29% in Public Health,
- while Engineering, Physics and Math would have the greatest decreases.

Counterfactual analysis

What would happen if the proportion of authors by race & gender was that of the 2010 US Census?

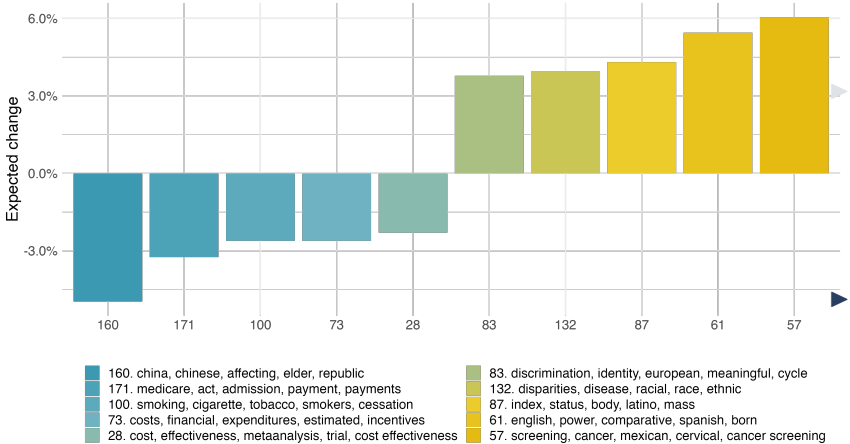


In Health, topics with the highest increase are related to latinx population, racial disparities and discrimination,

while topics related to health costs or tobacco would decrease the most.

Counterfactual analysis

What would happen if the proportion of authors by race & gender was that of the 2010 US Census?



In Health, topics with the highest increase are related to latinx population, racial disparities and discrimination,

while topics related to health costs or tobacco would decrease the most.

Conclusions

- There is an **underrepresentation** of marginalized groups (at the intersection of race & gender),
- these groups have specific research interests, which creates a **research gap of understudied topics.**
- There is also a **citation gap** due to both the field and topic distribution, and within topics bias.

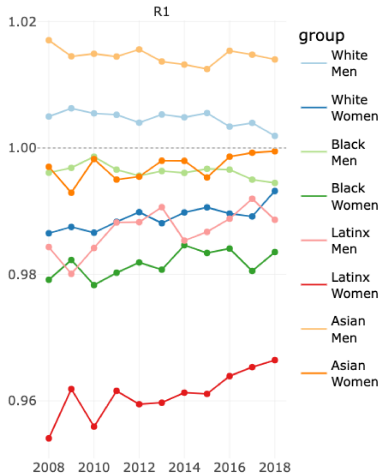
Conclusions

- There is an **underrepresentation** of marginalized groups (at the intersection of race & gender),
- these groups have specific research interests, which creates a **research gap of understudied topics.**
- There is also a **citation gap** due to both the field and topic distribution, and within topics bias.

Conclusions

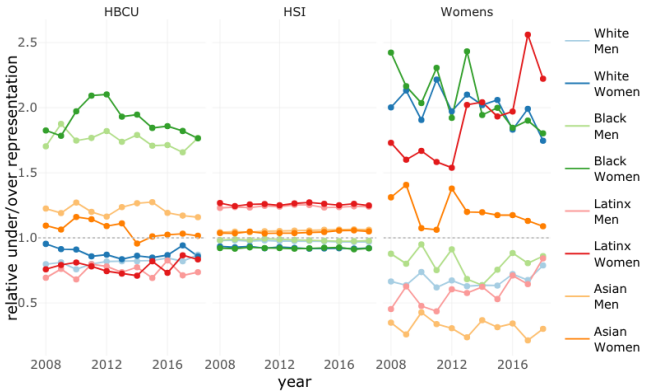
- There is an **underrepresentation** of marginalized groups (at the intersection of race & gender),
- these groups have specific research interests, which creates a **research gap of understudied topics.**
- There is also a **citation gap** due to both the field and topic distribution, and within topics bias.

Moving Forwards



- To start thinking about how to move forwards into a more inclusive science, we need to consider the role of institutions.
- Research intensive -R1- institutions produce the large majority of research articles,
- and within these institutions, White and Asian men are overrepresented,
- while women -and specially Black and Latinx women- are underrepresented.

Moving Forwards



- Other institutions, like Historically Black Colleges, Hispanic Serving Institutions, and Womens' Colleges are focus on students of specific populations,
- and our data shows that they also play an important role for inclusion of authors from their target population.

Thank You!

Questions?

@Diego_Koz
diego.kozlowski@uni.lu
sciencebias.uni.lu/app

Acknowledgement

The Doctoral Training Unit **Data-driven computational modelling and applications** (DRIVEN) is funded by the Luxembourg National Research Fund under the PRIDE programme (PRIDE17/12252781).

https://driven.uni.lu

Bibliography

[1] Tukufu Zuberi. “Thicker Than Blood: How Racial Statistics Lie”. In: 31 (2002), p. 529. ISSN: 0094-3061. DOI: 10.2307/3090025.

[2] Catherine D’Ignazio and Lauren Klein. “4. “What Gets Counted Counts””. In: *Data Feminism*. https://data-feminism.mitpress.mit.edu/pub/h1w0nbqp. Mar. 16, 2020. URL: https://data-feminism.mitpress.mit.edu/pub/h1w0nbqp.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation”. In: *the Journal of machine Learning research* 3 (2003), pp. 993–1022.

[4] Sandra G Harding. *The feminist standpoint theory reader: Intellectual and political controversies*. Psychology Press, 2004.