Intersectional inequalities in science

Diego Kozlowski[1]

[1] DRIVEN DTU-FSTM, University of Luxembourg, Luxembourg

METASCIENCE 2021

Name-based racial inference

Race and gender inequalities in US authors

What are we missing? (case example)

Name-based racial inference

Introduction

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

- We want to study how the cultural construct of *race* influences US academy.
- As bibliometric databases don't have information on authors self-perceived race, we first need to infer it from their names,
- We focus on how to avoid introducing new biases when inferring race.

Introduction

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

- We want to study how the cultural construct of *race* influences US academy.
- As bibliometric databases don't have information on authors self-perceived race, we first need to infer it from their names,
- We focus on how to avoid introducing new biases when inferring race.

Introduction

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

- We want to study how the cultural construct of *race* influences US academy.
- As bibliometric databases don't have information on authors self-perceived race, we first need to infer it from their names,
- We focus on how to avoid introducing new biases when inferring race.

Information retrieval

Each author with a *given* and *family* name has two associated probability distributions:

Type	*Name*	*Asian*	*Black*	*Latinx*	*White*
Given	Juan	1.5%	0.5%	93.4%	4.5%
	Doris	3.4%	13.5%	6.3%	76.7%
	Andy	38.8%	1.6%	6.4%	53.2%
Family	Rodriguez	0.6%	0.5%	94.1%	4.8%
	Lee	43.8%	16.9%	2.0%	37.3%
	Washington	0.3%	91.6%	2.7%	5.4%

If we aim to assign a single label, we need to choose:

1. which probability or combination of probabilities to use,
2. How to define the threshold/assignation.

Information retrieval

Each author with a *given* and *family* name has two associated probability distributions:

Type	*Name*	*Asian*	*Black*	*Latinx*	*White*
Given	Juan	1.5%	0.5%	93.4%	4.5%
	Doris	3.4%	13.5%	6.3%	76.7%
	Andy	38.8%	1.6%	6.4%	53.2%
Family	Rodriguez	0.6%	0.5%	94.1%	4.8%
	Lee	43.8%	16.9%	2.0%	37.3%
	Washington	0.3%	91.6%	2.7%	5.4%

If we aim to assign a single label, we need to choose:

1. which probability or combination of probabilities to use,
2. How to define the threshold/assignation.

Proposed algorithms

Based on this, we built multiple algorithms

1. using Family names,
2. using Given names,
3. using a mixture of both distributions, and
4. normalizing Given names to match the census distribution

because information on given names from mortgage data (Tzioumis 2018) has a different underlying distribution.

Racial group	*Family (Census)*	*Given (Mortgage)*
Asian	5.0%	6.3%
Black	12.4%	4.2%
Latinx	16.5%	6.9%
White	66.1%	**82.6%**

Alternatively we use **fractional counting:** considering the full distribution for each author. In this way, authors are not uniquely identified, but on aggregate gives an unbiased estimate.

Proposed algorithms

Based on this, we built multiple algorithms

1. using Family names,
2. using Given names,
3. using a mixture of both distributions, and
4. normalizing Given names to match the census distribution

because information on given names from mortgage data (Tzioumis 2018) has a different underlying distribution.

Racial group	*Family (Census)*	*Given (Mortgage)*
Asian	5.0%	6.3%
Black	12.4%	4.2%
Latinx	16.5%	6.9%
White	66.1%	**82.6%**

Alternatively we use **fractional counting:** considering the full distribution for each author. In this way, authors are not uniquely identified, but on aggregate gives an unbiased estimate.

Proposed algorithms

Based on this, we built multiple algorithms

1. using Family names,
2. using Given names,
3. using a mixture of both distributions, and
4. normalizing Given names to match the census distribution

because information on given names from mortgage data (Tzioumis 2018) has a different underlying distribution.

Racial group	*Family (Census)*	*Given (Mortgage)*
Asian	5.0%	6.3%
Black	12.4%	4.2%
Latinx	16.5%	6.9%
White	66.1%	**82.6%**

Alternatively we use **fractional counting:** considering the full distribution for each author. In this way, authors are not uniquely identified, but on aggregate gives an unbiased estimate.

Proposed algorithms

Based on this, we built multiple algorithms

1. using Family names,
2. using Given names,
3. using a mixture of both distributions, and
4. normalizing Given names to match the census distribution

because information on given names from mortgage data (Tzioumis 2018) has a different underlying distribution.

Racial group	*Family (Census)*	*Given (Mortgage)*
Asian	5.0%	6.3%
Black	12.4%	4.2%
Latinx	16.5%	6.9%
White	66.1%	**82.6%**

Alternatively we use **fractional counting:** considering the full distribution for each author. In this way, authors are not uniquely identified, but on aggregate gives an unbiased estimate.

Proposed algorithms

Based on this, we built multiple algorithms

1. using Family names,
2. using Given names,
3. using a mixture of both distributions, and
4. normalizing Given names to match the census distribution

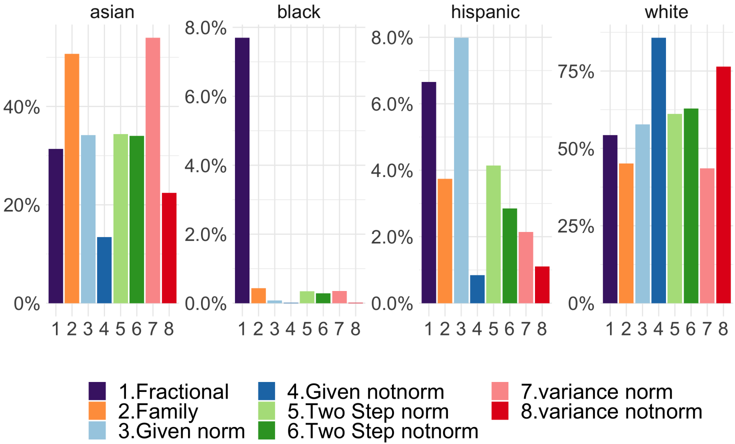
because information on given names from mortgage data (Tzioumis 2018) has a different underlying distribution.

Racial group	*Family (Census)*	*Given (Mortgage)*
Asian	5.0%	6.3%
Black	12.4%	4.2%
Latinx	16.5%	6.9%
White	66.1%	**82.6%**

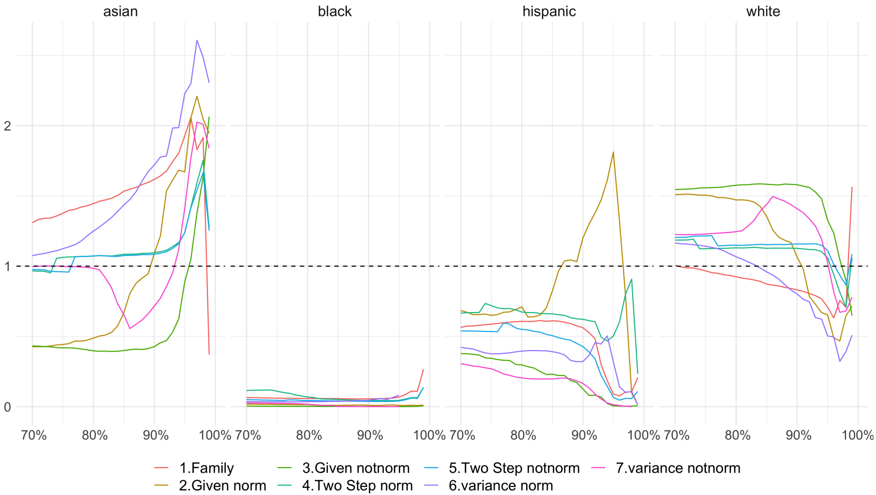
Alternatively we use **fractional counting:** considering the full distribution for each author. In this way, authors are not uniquely identified, but on aggregate gives an unbiased estimate.

90% threshold in WOS

If we take a 90% threshold, all models underestimate Black authors with respect to **fractional counting**, and almost always underestimate Latinx authors



Moving Threshold



Changing the threshold does not solve the bias.

Imputation

Finally, for those names that do not appear on the Census, we would like to impute a probability that doesn't add bias

Racial group	*US census aggregate*	*US census "All other names"*	*US WoS (fractional counting)*
Asian	5.0%	8.2%	24.5%
Black	12.4%	8.8%	7.2%
Latinx	16.5%	14.1%	5.4%
White	66.1%	68.8%	59.4%

If we use the US Census category "All other names", we would introduce a bias, because the distribution on WOS differs from that of the Census.

Summary

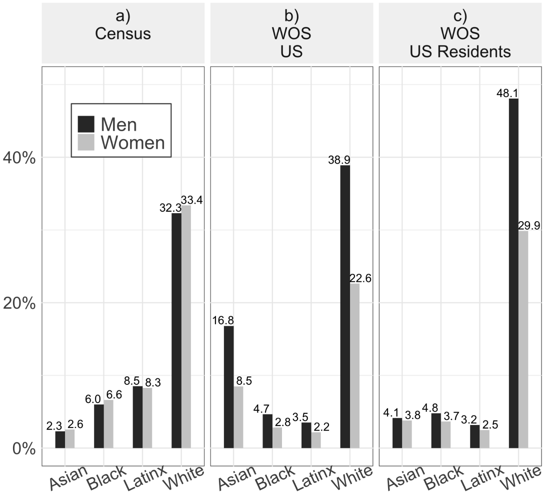
In summary, our recommendation is:

Topic	*Do's*	*Don'ts*
Given Names	Use only family names	Given names might have a biased distribution
Thresholding	Use the **fractional counting**	Do not use a threshold
Imputation	Impute by your own data average	Do not use 'All other names' from Census

- **Always consider the historical context of your data:** These racial categories only make sense within contemporary US, because they are a product of this society.
- Naming practices are also a product of society, and in the case of US they go back to the times of slavery. This historical context needs to be acknowledge when we build such name-based algorithms.
- Using a full distribution instead of a simple label might imply more work, but is the best way to avoid biases.

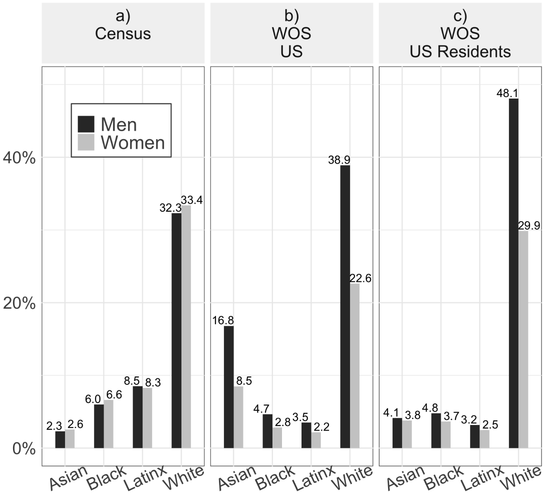
Race and gender inequalities in US authors

Results. General distribution



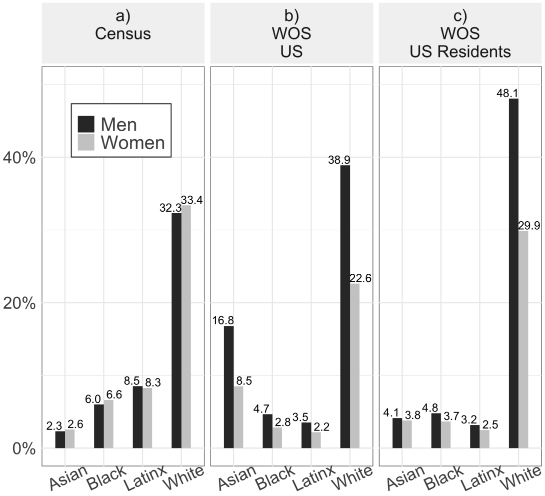
- Overall, there is an overrepresentation of White and Asian men,
- a large proportion of Asian authors are not US residents,
- i.e. the Census is not a perfect benchmark to define *overrepresentation* of this group,
- **women** are underrepresented,
- **Black and Latinx** are underrepresented,
- **Black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

Results. General distribution



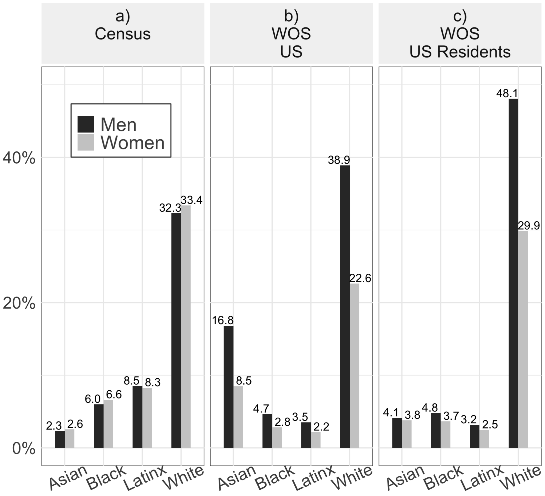
- Overall, there is an overrepresentation of White and Asian men,
- a large proportion of Asian authors are not US residents,
- i.e. the Census is not a perfect benchmark to define *overrepresentation* of this group,
- **women** are underrepresented,
- **Black and Latinx** are underrepresented,
- **Black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

Results. General distribution



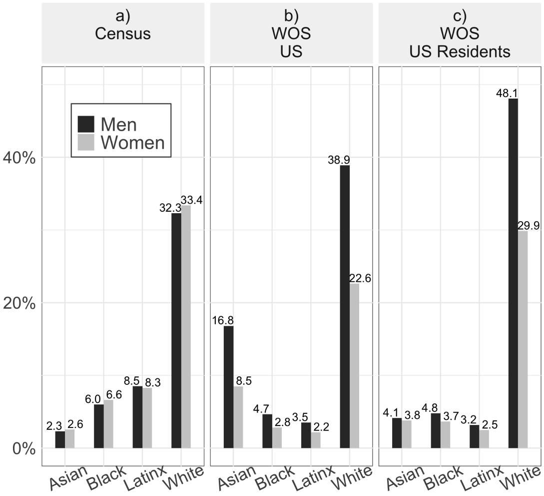
- Overall, there is an overrepresentation of White and Asian men,
- a large proportion of Asian authors are not US residents,
- i.e. the Census is not a perfect benchmark to define *overrepresentation* of this group,
- **women** are underrepresented,
- **Black and Latinx** are underrepresented,
- **Black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

Results. General distribution



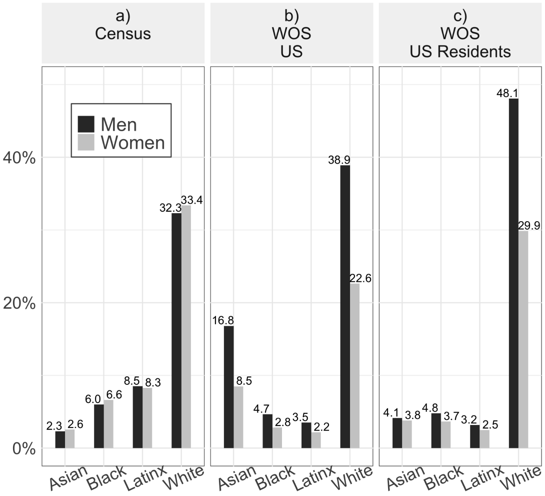
- Overall, there is an overrepresentation of White and Asian men,
- a large proportion of Asian authors are not US residents,
- i.e. the Census is not a perfect benchmark to define *overrepresentation* of this group,
- **women** are underrepresented,
- **Black and Latinx** are underrepresented,
- **Black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

Results. General distribution



- Overall, there is an overrepresentation 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 *overrepresentation* of this group,
- **women** are underrepresented,
- **Black and Latinx** are underrepresented,
- **Black and Latinx women** are the most underrepresented.

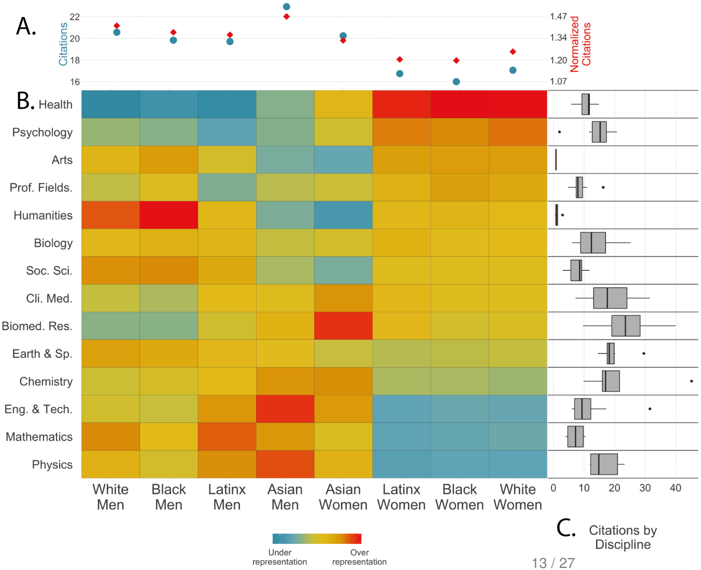
Results. General distribution



- Overall, there is an overrepresentation 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 *overrepresentation* of this group,

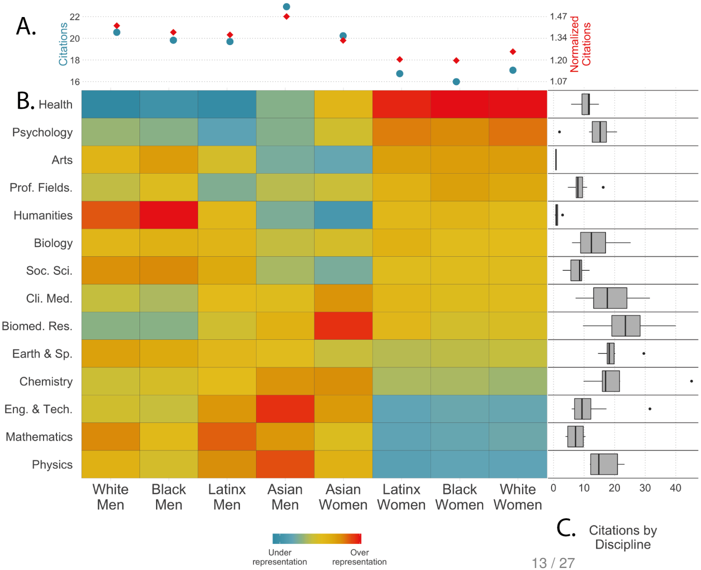
- **women** are underrepresented,
- **Black and Latinx** are underrepresented,
- **Black and Latinx women** are the most underrepresented.

Results. Disciplines' Heterogeneity



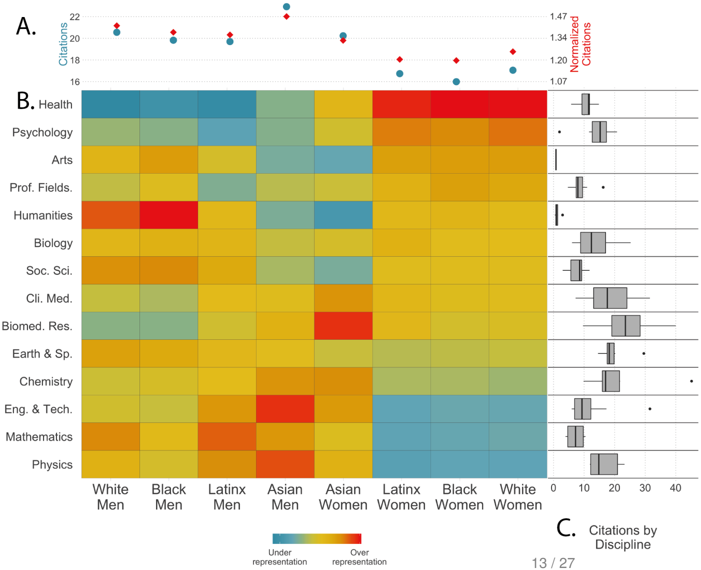
- The distribution by disciplines has a clear pattern by gender, except on Asian authors.
- Women have less citations on average,
- field normalization on citations reduces the gap, but it does not go away.

Results. 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 it does not go away.

Results. Disciplines' Heterogeneity



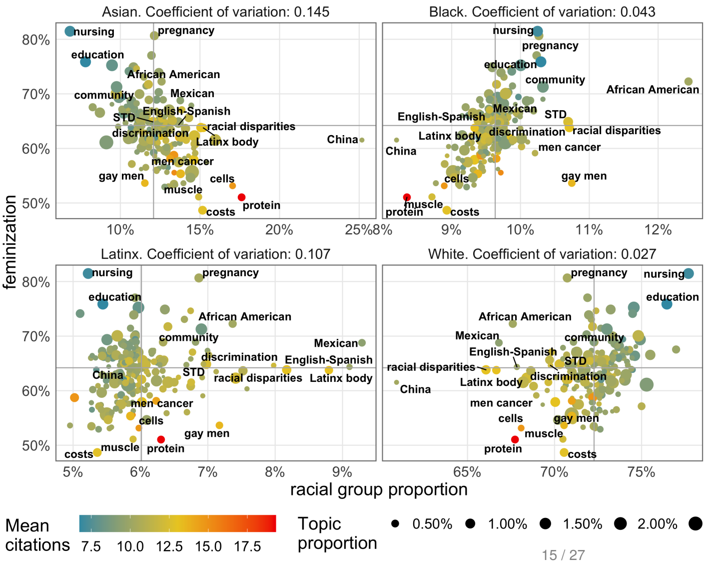
- The distribution by disciplines has a clear pattern by gender, except on Asian authors.
- Women have less citations on average,
- field normalization on citations reduces the gap, but it does not go away.

- We want to go deeper into the correlation of research topics with race & gender,
- for this, we focus on *health* and define 200 specific 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,
- for this, we focus on *health* and define 200 specific topics, using LDA (Blei, Ng, and Jordan 2003).
- For each race & gender we define the average participation on each topic.

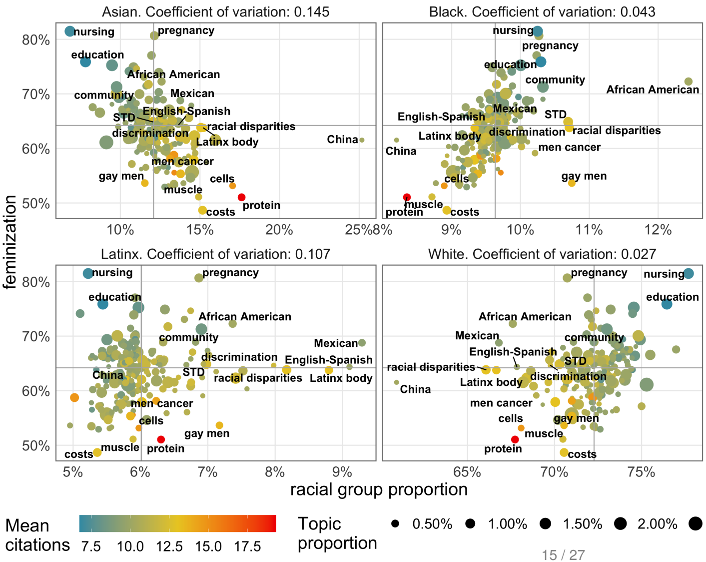
- We want to go deeper into the correlation of research topics with race & gender,
- for this, we focus on *health* and define 200 specific topics, using LDA (Blei, Ng, and Jordan 2003).
- For each race & gender we define the average participation on each topic.

Research topics



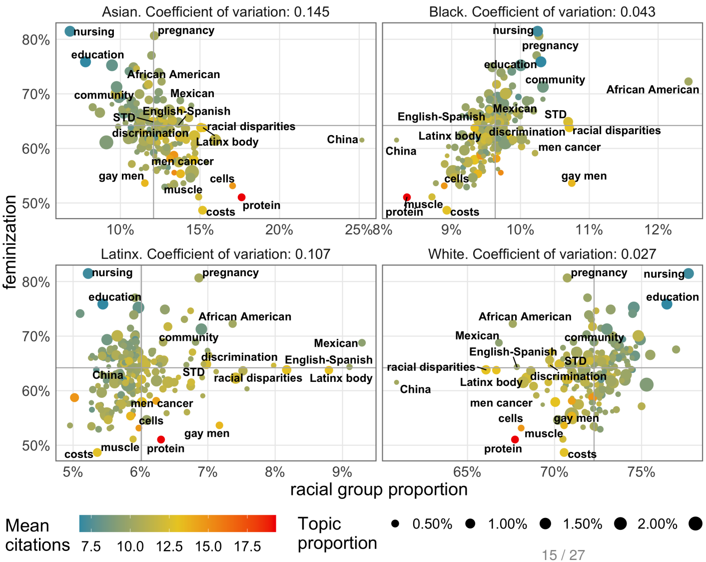
- Women publish more on *nursing*, *pregnancy* and *education*,
- Black authors focus on *African American* and *racial disparities* studies,
- Latinx authors focus on *Mexican* and *Latinx body* studies, but also on language issues,
- Asian authors focus on *China*, while White authors are more evenly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

Research topics



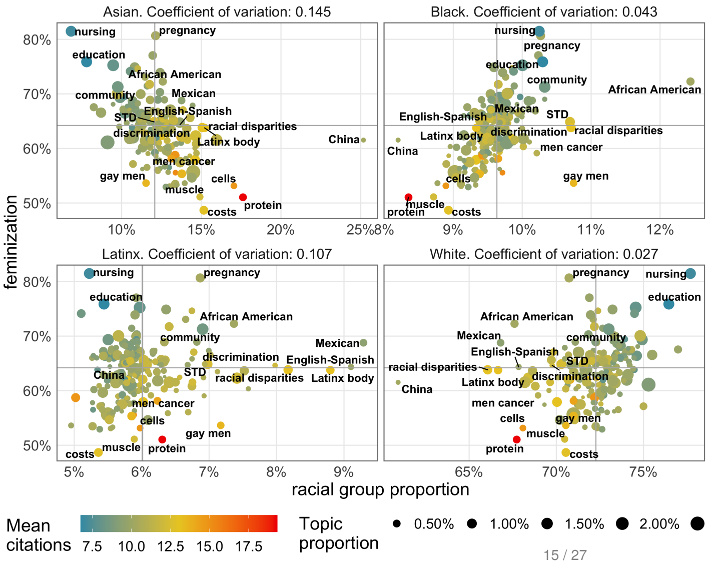
- Women publish more on *nursing*, *pregnancy* and *education*,
- Black authors focus on *African American* and *racial disparities* studies,
- Latinx authors focus on *Mexican* and *Latinx body* studies, but also on language issues,
- Asian authors focus on *China*, while White authors are more evenly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

Research topics



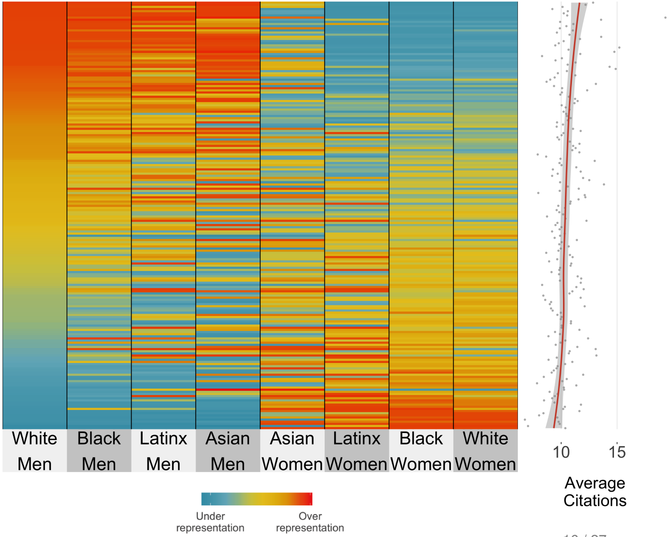
- Women publish more on *nursing*, *pregnancy* and *education*,
- Black authors focus on *African American* and *racial disparities* studies,
- Latinx authors focus on *Mexican* and *Latinx body* studies, but also on language issues,
- Asian authors focus on *China*, while White authors are more evenly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

Research topics



- 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.

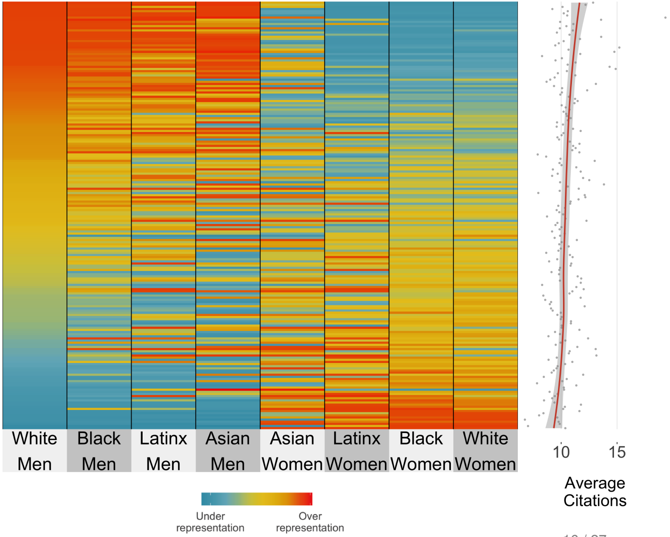
Topics and citations



How does this 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.

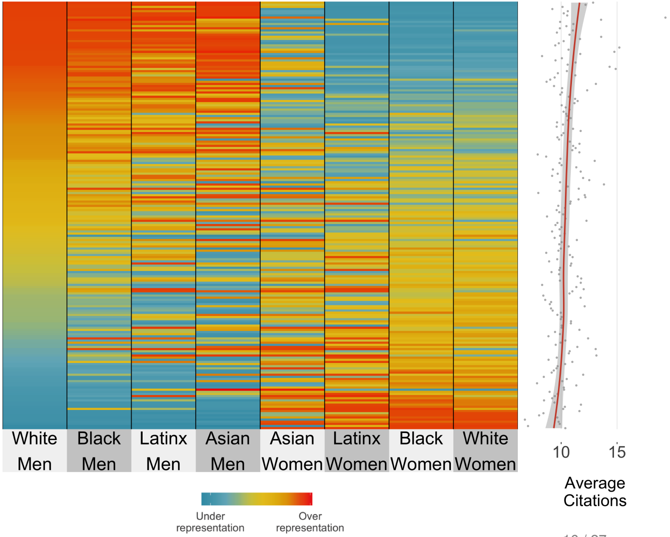
Topics and citations



How does this 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

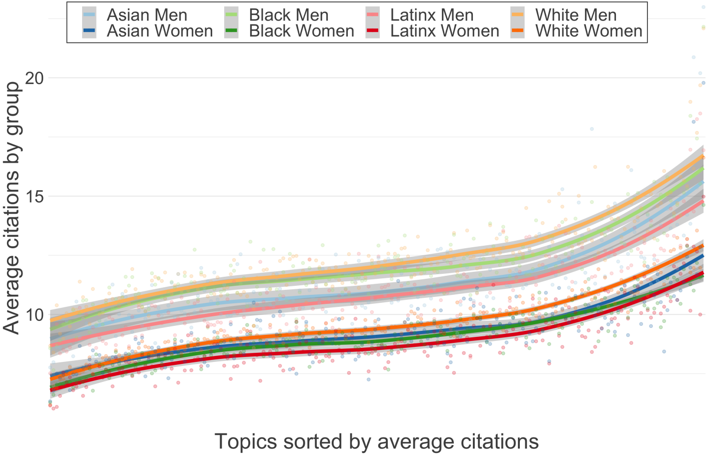


How does this topics affect citations gaps?

- 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

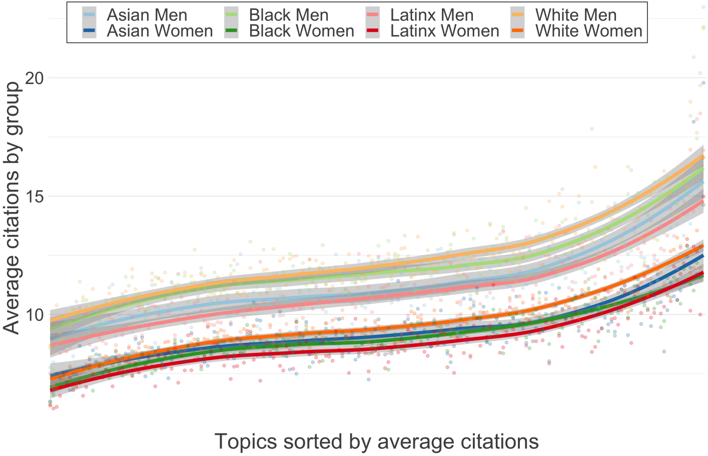
Are citations gaps exclusively explained by research topics? Not only



- 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*.
- there is both a inter-topic and intra-topic bias.

Topics and citations

Are citations gaps exclusively explained by research topics? Not only



- 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*.
- there is both a inter-topic and intra-topic bias.

Conclusions

- There is an under representation of marginalized groups (at the intersection of race & gender),
- these groups have specific research interests,
- therefore, there are relevant under-studied topics in science, that mainly affect marginalized groups.
- Also, women tend to be less cited. This is due to both the field and topic distribution, and within topics bias.

Conclusions

- There is an under representation of marginalized groups (at the intersection of race & gender),
- these groups have specific research interests,
- therefore, there are relevant under-studied topics in science, that mainly affect marginalized groups.
- Also, women tend to be less cited. This is due to both the field and topic distribution, and within topics bias.

Conclusions

- There is an under representation of marginalized groups (at the intersection of race & gender),
- these groups have specific research interests,
- therefore, there are relevant under-studied topics in science, that mainly affect marginalized groups.
- Also, women tend to be less cited. This is due to both the field and topic distribution, and within topics bias.

Conclusions

- There is an under representation of marginalized groups (at the intersection of race & gender),
- these groups have specific research interests,
- therefore, there are relevant under-studied topics in science, that mainly affect marginalized groups.
- Also, women tend to be less cited. This is due to both the field and topic distribution, and within topics bias.

What are we missing? (case example)

- As we conclude, there are relevant under-studied topics in science, that mainly affect marginalized groups.
- this can also reflect on **missing datasets**,
- As the authors of *Data Feminism* explain:

  > *"the things that we do not or cannot collect data about are very often perceived to be things that do not exist at all." (D'Ignazio and Klein 2018)*

- How can we make evidence-based public policy over marginalized topics?
- We are going to present the case of data on abortion in clandestine situations in Argentina.
- In times of increasing persecution against abortion, how can we understand this practice based on evidence rather than personal beliefs?

- As we conclude, there are relevant under-studied topics in science, that mainly affect marginalized groups.
- this can also reflect on **missing datasets**,
- As the authors of *Data Feminism* explain:

> *"the things that we do not or cannot collect data about are very often perceived to be things that do not exist at all." (D'Ignazio and Klein 2018)*

- How can we make evidence-based public policy over marginalized topics?
- We are going to present the case of data on abortion in clandestine situations in Argentina.
- In times of increasing persecution against abortion, how can we understand this practice based on evidence rather than personal beliefs?

- As we conclude, there are relevant under-studied topics in science, that mainly affect marginalized groups.
- this can also reflect on **missing datasets**,
- As the authors of *Data Feminism* explain:

  > *"the things that we do not or cannot collect data about are very often perceived to be things that do not exist at all." (D'Ignazio and Klein 2018)*

- How can we make evidence-based public policy over marginalized topics?
- We are going to present the case of data on abortion in clandestine situations in Argentina.
- In times of increasing persecution against abortion, how can we understand this practice based on evidence rather than personal beliefs?

- As we conclude, there are relevant under-studied topics in science, that mainly affect marginalized groups.
- this can also reflect on **missing datasets**,
- As the authors of *Data Feminism* explain:

  “the things that we do not or cannot collect data about are very often perceived to be things that do not exist at all.” (D’Ignazio and Klein 2018)
- How can we make evidence-based public policy over marginalized topics?
- We are going to present the case of data on abortion in clandestine situations in Argentina.
- In times of increasing persecution against abortion, how can we understand this practice based on evidence rather than personal beliefs?

- As we conclude, there are relevant under-studied topics in science, that mainly affect marginalized groups.
- this can also reflect on **missing datasets**,
- As the authors of *Data Feminism* explain:

  > *"the things that we do not or cannot collect data about are very often perceived to be things that do not exist at all." (D'Ignazio and Klein 2018)*

- How can we make evidence-based public policy over marginalized topics?
- We are going to present the case of data on abortion in clandestine situations in Argentina.
- In times of increasing persecution against abortion, how can we understand this practice based on evidence rather than personal beliefs?

- As we conclude, there are relevant under-studied topics in science, that mainly affect marginalized groups.
- this can also reflect on **missing datasets**,
- As the authors of *Data Feminism* explain:

  > *"the things that we do not or cannot collect data about are very often perceived to be things that do not exist at all." (D'Ignazio and Klein 2018)*

- How can we make evidence-based public policy over marginalized topics?
- We are going to present the case of data on abortion in clandestine situations in Argentina.
- In times of increasing persecution against abortion, how can we understand this practice based on evidence rather than personal beliefs?

- As we conclude, there are relevant under-studied topics in science, that mainly affect marginalized groups.
- this can also reflect on **missing datasets**,
- As the authors of *Data Feminism* explain:

  "the things that we do not or cannot collect data about are very often perceived to be things that do not exist at all." (D'Ignazio and Klein 2018)

- How can we make evidence-based public policy over marginalized topics?
- We are going to present the case of data on abortion in clandestine situations in Argentina.
- In times of increasing persecution against abortion, how can we understand this practice based on evidence rather than personal beliefs?

Abortion in Argentina

- Argentina legalized abortion in 2020,
- one of the main arguments against the legalization was the distress it causes.
- There was no proof for this, and because it was illegal, no studies were conducted.

Abortion in Argentina

- Argentina legalized abortion in 2020,
- one of the main arguments against the legalization was the distress it causes.
- There was no proof for this, and because it was illegal, no studies were conducted.

Abortion in Argentina

- Argentina legalized abortion in 2020,
- one of the main arguments against the legalization was the distress it causes.
- There was no proof for this, and because it was illegal, no studies were conducted.

Abortion in Argentina

- To get evidence-based knowledge, two feminist organizations joined forces:
- a grass-root organization of female doctors that helps people to have the safest conditions when practicing clandestine abortions,
- a feminist organization that works on data analysis.
- This goes beyond traditional research institutions. Sometimes, only grass-root organizations are able to access the most relevant data (D'Ignazio and Klein 2018)

LaRe vuelta.
Colectiva Feminista
Neuquén

larevuelta.com.ar

Abortion in Argentina

- To get evidence-based knowledge, two feminist organizations joined forces:
- a grass-root organization of female doctors that helps people to have the safest conditions when practicing clandestine abortions,
- a feminist organization that works on data analysis.
- This goes beyond traditional research institutions. Sometimes, only grass-root organizations are able to access the most relevant data (D'Ignazio and Klein 2018)

larevuelta.com.ar

Abortion in Argentina

- To get evidence-based knowledge, two feminist organizations joined forces:
- a grass-root organization of female doctors that helps people to have the safest conditions when practicing clandestine abortions,
- a feminist organization that works on data analysis.
- This goes beyond traditional research institutions. Sometimes, only grass-root organizations are able to access the most relevant data (D'Ignazio and Klein 2018)

larevuelta.com.ar

ecofeminita.com

Abortion in Argentina

- To get evidence-based knowledge, two feminist organizations joined forces:
- a grass-root organization of female doctors that helps people to have the safest conditions when practicing clandestine abortions,
- a feminist organization that works on data analysis.
- This goes beyond traditional research institutions. Sometimes, only grass-root organizations are able to access the most relevant data (D’Ignazio and Klein 2018)

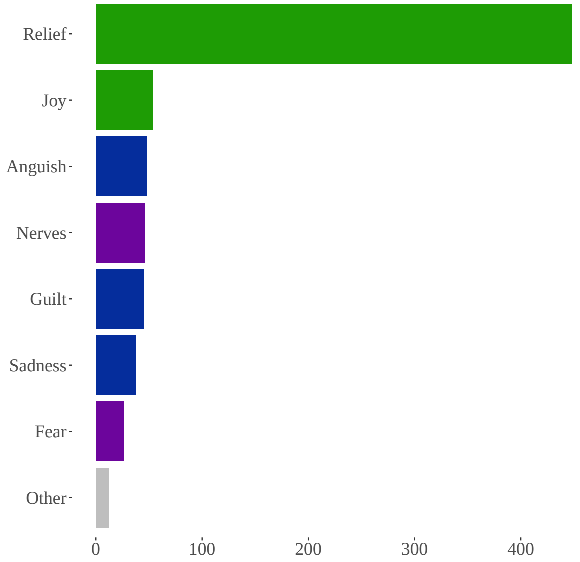
larevuelta.com.ar

ecofeminita.com

Abortion in Argentina

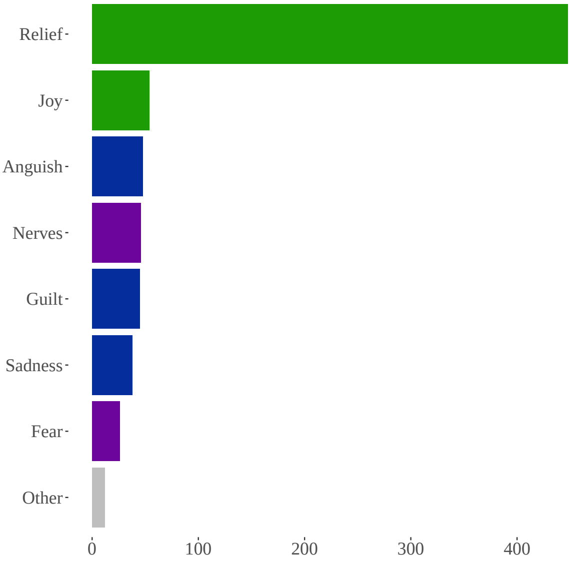
- They interviewed more than 400 people who were previously accompanied by La Revuelta, asking about which where the principal emotions they felt after practicing the abortion.
- The data shows that the most common emotion is actually **relief**,
- this evidence goes against the more or less generalized belief that abortion produces anguish and permanent trauma, an idea popularized by the anti-abortion organizations.

Abortion in Argentina



- They interviewed more than 400 people who were previously accompanied by La Revuelta, asking about which where the principal emotions they felt after practicing the abortion.
- The data shows that the most common emotion is actually **relief**,
- this evidence goes against the more or less generalized belief that abortion produces anguish and permanent trauma, an idea popularized by the anti-abortion organizations.

Abortion in Argentina



- They interviewed more than 400 people who were previously accompanied by La Revuelta, asking about which where the principal emotions they felt after practicing the abortion.
- The data shows that the most common emotion is actually **relief**,
- this evidence goes against the more or less generalized belief that abortion produces anguish and permanent trauma, an idea popularized by the anti-abortion organizations.

Conclusion

- Understudied research topics also appear in the form of *missing data*, that is also necessary for public policy.
- In order to move towards a more inclusive science, we also need to include grass-root organizations.
- In US, there is a rise of laws that persecute organizations that help people who practice abortions,
- this does not only generate a more unsafe environment for them, but will also restrict the possibility of research on this topic.

Conclusion

- Understudied research topics also appear in the form of *missing data*, that is also necessary for public policy.
- In order to move towards a more inclusive science, we also need to include grass-root organizations.
- In US, there is a rise of laws that persecute organizations that help people who practice abortions,
- this does not only generate a more unsafe environment for them, but will also restrict the possibility of research on this topic.

Conclusion

- Understudied research topics also appear in the form of *missing data*, that is also necessary for public policy.
- In order to move towards a more inclusive science, we also need to include grass-root organizations.
- In US, there is a rise of laws that persecute organizations that help people who practice abortions,
- this does not only generate a more unsafe environment for them, but will also restrict the possibility of research on this topic.

Conclusion

- Understudied research topics also appear in the form of *missing data*, that is also necessary for public policy.
- In order to move towards a more inclusive science, we also need to include grass-root organizations.
- In US, there is a rise of laws that persecute organizations that help people who practice abortions,
- this does not only generate a more unsafe environment for them, but will also restrict the possibility of research on this topic.

Thank You!

Questions?

@Diego_Koz
diego.kozlowski@uni.lu

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`

DTU DRIVEN

Fonds National de la Recherche Luxembourg

Bibliography

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

[2] Konstantinos Tzioumis. “Demographic aspects of first names”. In: 5 (2018). ISSN: 2052-4463. DOI: 10.1038/sdata.2018.25.

[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] Catherine D’Ignazio and Lauren Klein. “Chapter One: Bring Back the Bodies”. In: *Data Feminism*. https://mitpressonpubpub.mitpress.mit.edu/pub/zrlj0jqb. Nov. 1, 2018.