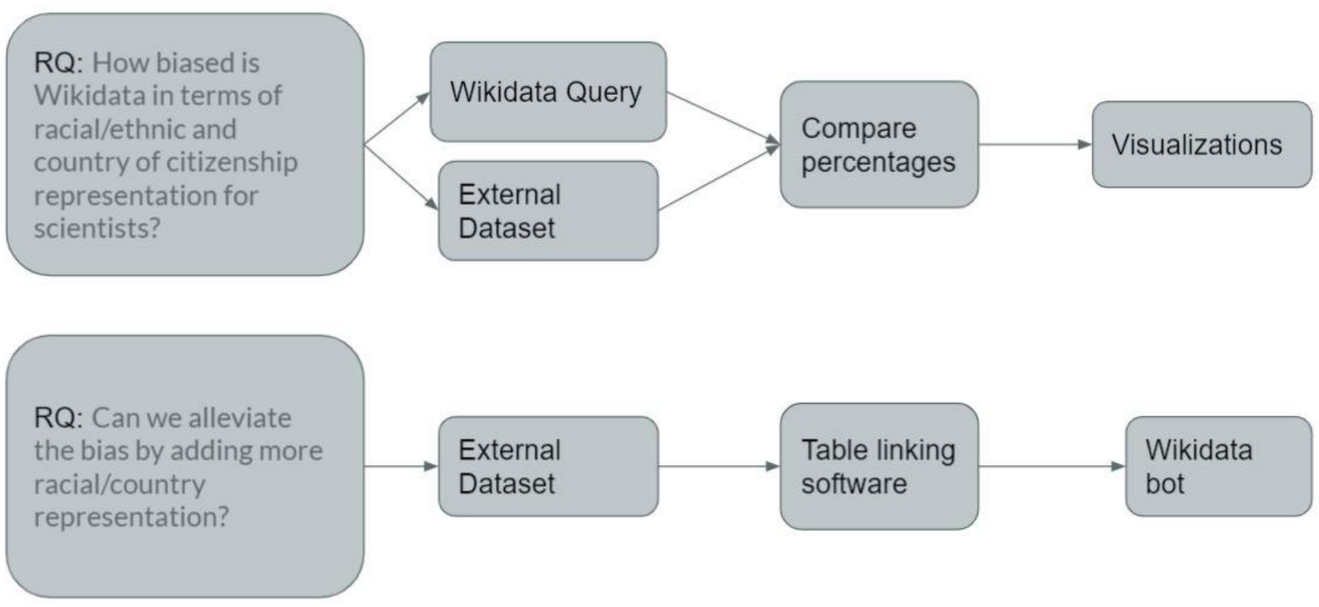
Race and Country Bias in Wikidata

Zaina Shaik, Filip Ilievski, Fred Morstatter

Problem Statement

- Because of the few active contributors to Wikidata and existing bias in STEM, there is race and country bias present in Wikidata. Comparison and integrating information can decrease bias.
- Research Questions:
  - How biased is Wikidata in terms of racial/ethnic and country of citizenship representation in general and for scientists, software developers, and engineers?
  - Can we alleviate the bias by adding more race/country representation with a bot?

Method Overview



- Found Wikidata Queries by writing code in SPARQL and using the Knowledge Graph Toolkit (KGTK)
- Looked at top 50 (most frequent) results in query and manually grouped them into their respective Race and Continent Categories
- Distributions from queries were compared to external datasets* (Race and Country Populations of World) to analyze bias
- Bias Measurement: difference of percentage between the Wikidata query and external dataset representation
- Connected Black Scientists dataset* to Wikidata/Wikipedia with table linking software

- *Challenges: Dataset Retrieval, Data Sparsity, Information Integration*
  - Future steps:
    - Run table linking with other scientist datasets* (Asian, Black, Middle Eastern, Hispanic/Latino)
    - Implement Wikidata bot to insert minority data into Wikidata.

*https://bit.ly/3AdZ80f

Race Results

TABLE I
COMPARISON OF RACE DISTRIBUTIONS BETWEEN WIKIDATA (WD) AND REAL-WORLD DATA.

	Total		Scientists	Software	Engineers
	WD	real	WD	WD	WD
White	**37.63**	**17.80**	**83.95**	**44.08**	**70.74**
Black	18.60	14.83	9.05	16.95	14.72
Asian	39.35	31.14	1.10	20.34	5.97
Middle Eastern	2.37	24.57	5.46	15.25	6.75
Indigenous to America	0.82	3.71	0.00	1.69	0.61
Hispanic/Latino	1.15	n/a	0.44	0.00	1.15
Pacific Islander	0.18	n/a	0.00	1.69	0.00
Other	n/a	7.95	n/a	n/a	n/a

- Wikidata is skewed towards the white race while underrepresenting all other races

Country of Citizenship Results

TABLE II
COMPARISON OF COUNTRY OF CITIZENSHIP DISTRIBUTIONS BETWEEN WIKIDATA (WD) AND REAL-WORLD DATA.

	Total		Scientists	Software	Engineers
	WD	real	WD	WD	WD
Asia	18.06	55.26	7.47	5.46	5.36
Europe	**57.10**	**8.80**	**71.06**	**68.09**	**66.82**
Middle East	1.97	5.55	1.30	1.34	1.32
Africa	0.00	11.90	0.00	0.00	0.00
North America	**16.47**	**5.45**	**15.43**	**19.88**	**19.51**
South America	3.62	7.38	2.76	3.46	5.25
Pacific Islands	2.76	5.66	1.98	1.77	1.74

- Wikidata is skewed towards European and North American countries while underrepresenting all other continents

Table Linking Accuracy

- Table linking software took information from dataset and ranked five possible matching Wikidata Qnodes for each person and occupation.
- Candidate Ranking Error: found correct Qnode, but not first
- Candidate Generation Error: did not find correct Qnode

Accuracy	Person	Occupation
Correct	**93.75%**	59.17%
Candidate Ranking Error	4.47%	35.00%
Candidate Generation Error	1.79%	5.83%

If interested contact <Zaina Shaik> zainas@isi.edu
Work performed under REU Site program
supported by NSF grant #2051101

USC Viterbi
School of Engineering
Information Sciences Institute