

Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media

Juhi Kulshrestha

joint work with

Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar,
Saptarshi Ghosh, Krishna P. Gummadi and Karrie Karahalios



MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

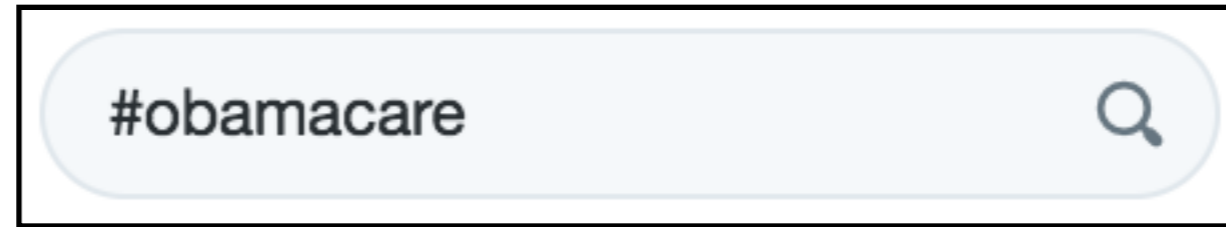


Social media as "search" platform


- Rich source of news & information
- **Search** is how users follow news about events & people
- **Hashtags** - recommended **queries**






Social media as "search" platform








Social media as "search" platform

#obamacare 

 **Mark Meadows**  @RepMarkMeadows · Feb 23
I support @RandPaul and @RepSanfordSC's #Obamacare replacement plan -- a plan that will lower costs and put the focus back on the patient.
117 290 971


 **Still With Her** @craftyme25 · Feb 23
Whoa. Arkansas is pissed. Woman says her husband will die without #ObamaCare. Asks @SenTomCotton "What kind of insurance do YOU have?" #CNN
27 25

 **Ryan**  @Politicalry · Feb 22
Support for Obamacare growing! We do not need to #RepealAndReplace #Obamacare, fix it as is or keep it for the sake of the people's welfare.
6 24 56

 **Charles Gaba**   @charles_gaba · Feb 22
...but not everyone has to buy a Lamborghini, as Ted Cruz falsely claimed was the case under #Obamacare.
3 128 932

Ranked list
(according to
importance)

Potential bias in search results

Mark Meadows 
@RepMarkMeadows

 **Follow** 

I support [@RandPaul](#) and [@RepSanfordSC](#)'s [#Obamacare](#) replacement plan -- a plan that will lower costs and put the focus back on the patient.





Potential bias in search results

 **Mark Meadows** 
@RepMarkMeadows Follow

I support @RandPaul and @RepSanfordSC's #Obamacare replacement plan -- a plan that will lower costs and put the focus back on the patient.



 **Ryan** 
@Politicalry Follow

Support for Obamacare growing! We do not need to #RepealAndReplace #Obamacare, fix it as is or keep it for the sake of the people's welfare.



Search can shape user opinion

- Users place **greater trust in higher ranked** items *[Pan et al., 2007]*
- Biased search results can **influence voting** patterns *[Epstein & Robertson, 2015]*

Search bias in the headlines

Search engine bias: What search results are telling you (and what they're not)

How Google Shapes the News You See About the Candidates

Who would Google vote for? An analysis of political bias in internet search engine results

Donald Trump Accuses Google of Bias in Search Engine Results

How Google's search algorithm spreads false information with a rightwing bias

Search bias in the headlines

Search engine bias: What search results are telling you (and what they're not)

How Google Shapes the News You See About the Candidates

Who would Google vote for? An analysis of political bias in internet search engine results

Donald Trump Accuses Google of Bias in Search Engine Results

How Google's search algorithm spreads false information with a rightwing bias

**What does bias of a search system mean?
How can we quantify it?**



Identify
sources of
search bias



Quantify
bias of each
source



Study bias of
political
searches in
Twitter



Identify
sources of
search bias



Quantify
bias of each
source

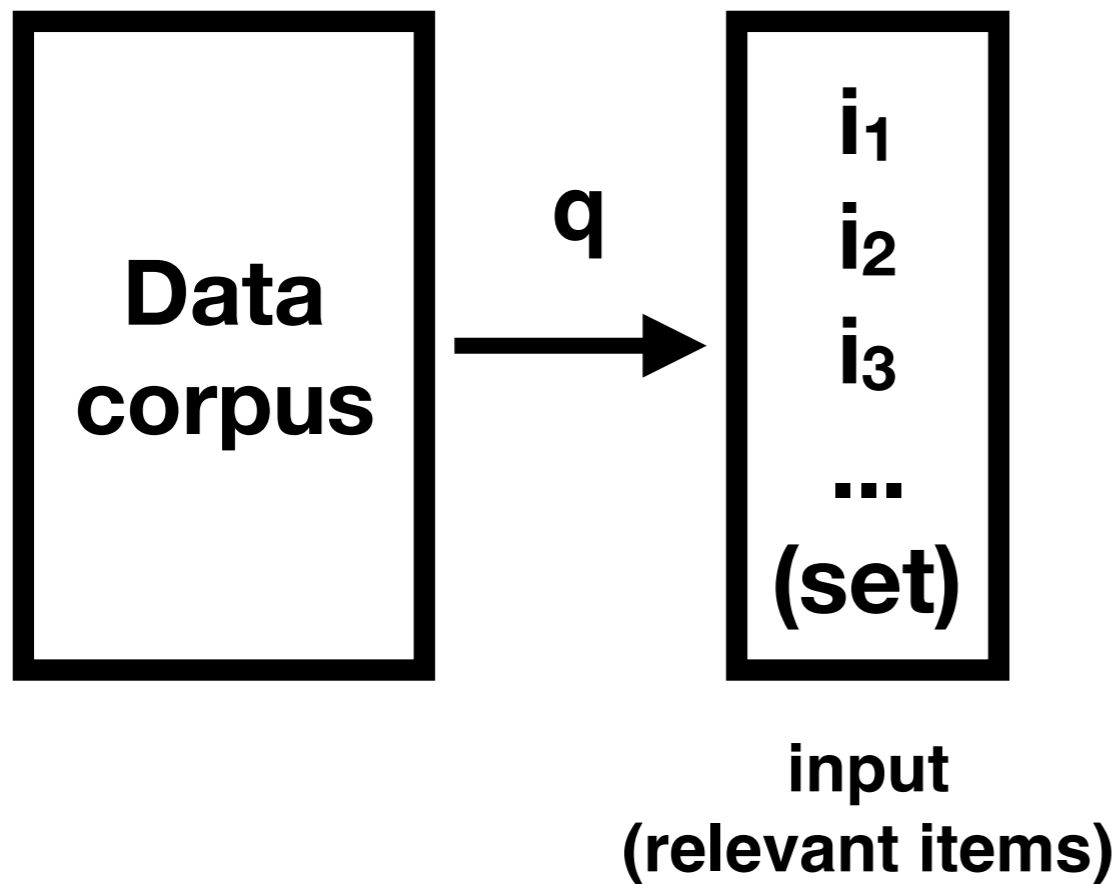


Study bias of
political
searches in
Twitter

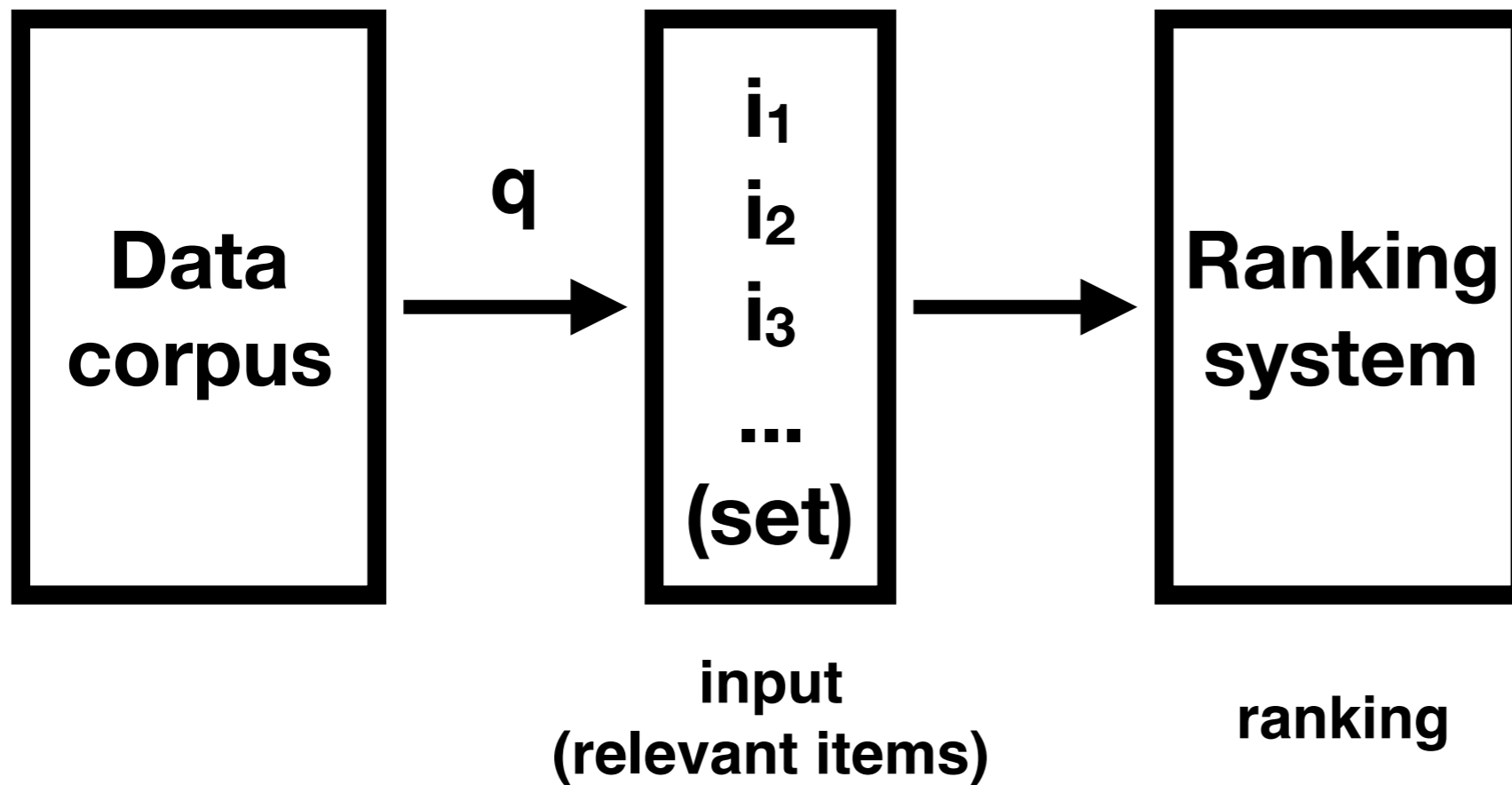
Social media search engine design

**Data
corpus**

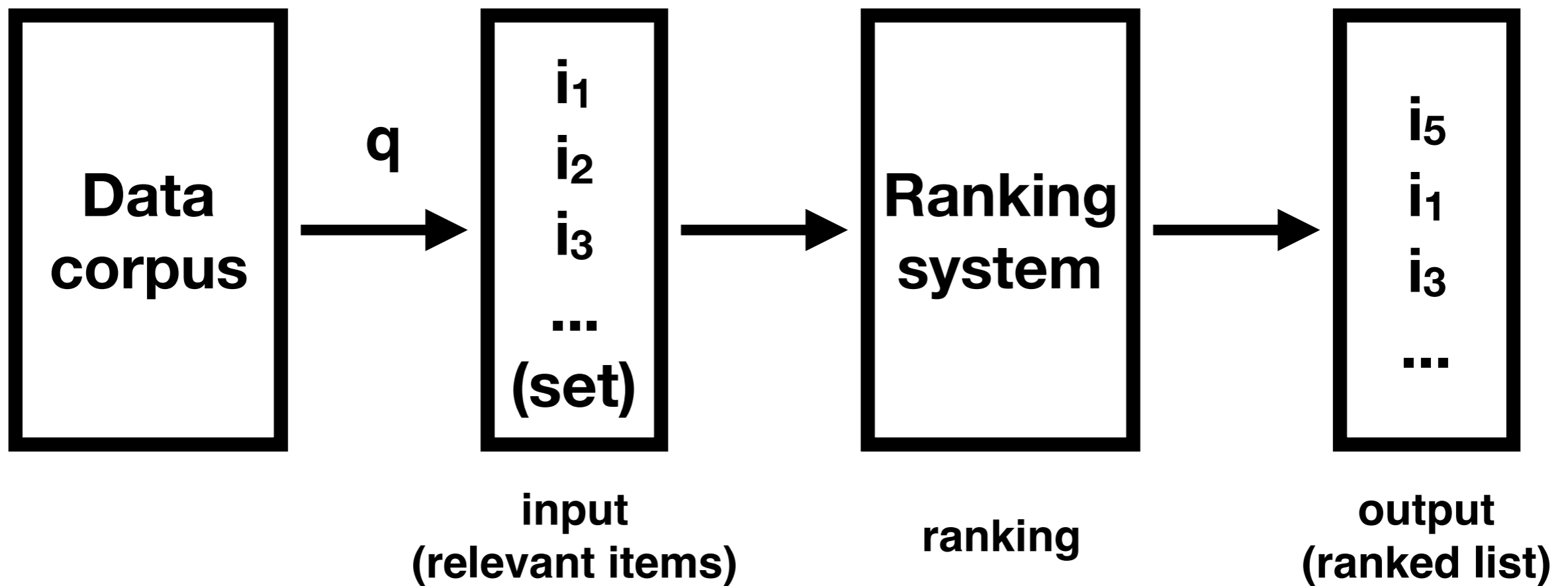
Social media search engine design



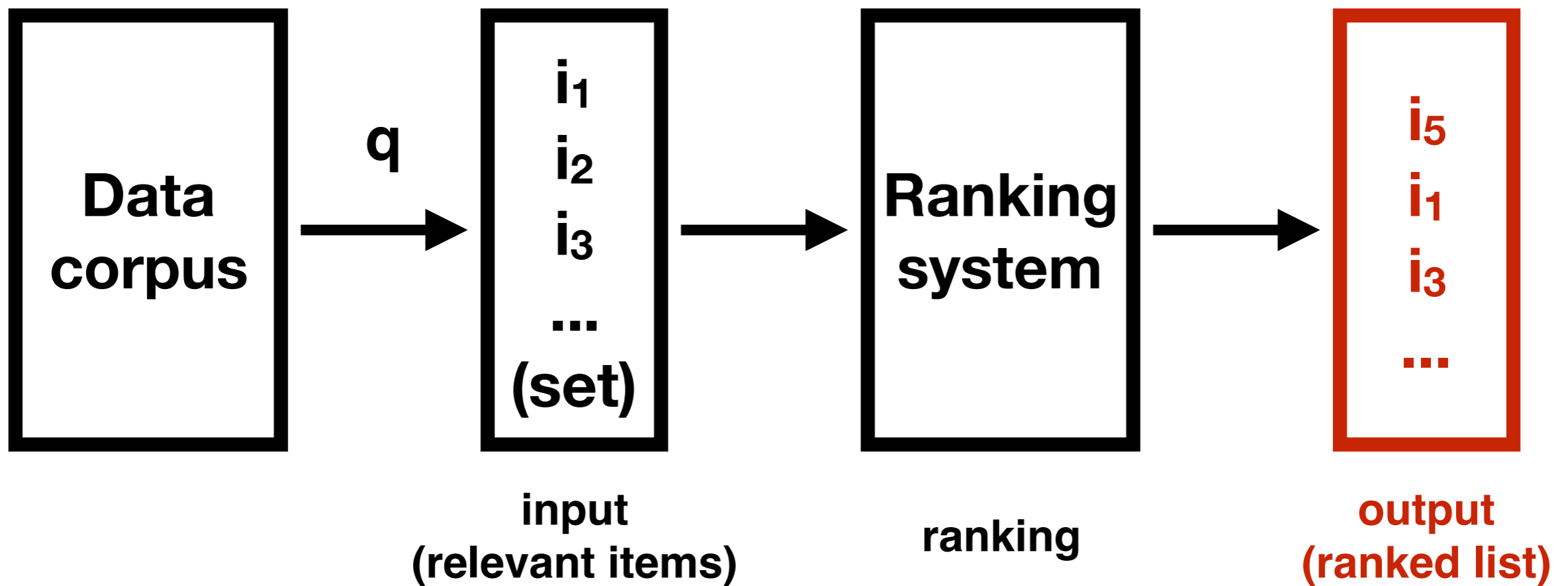
Social media search engine design



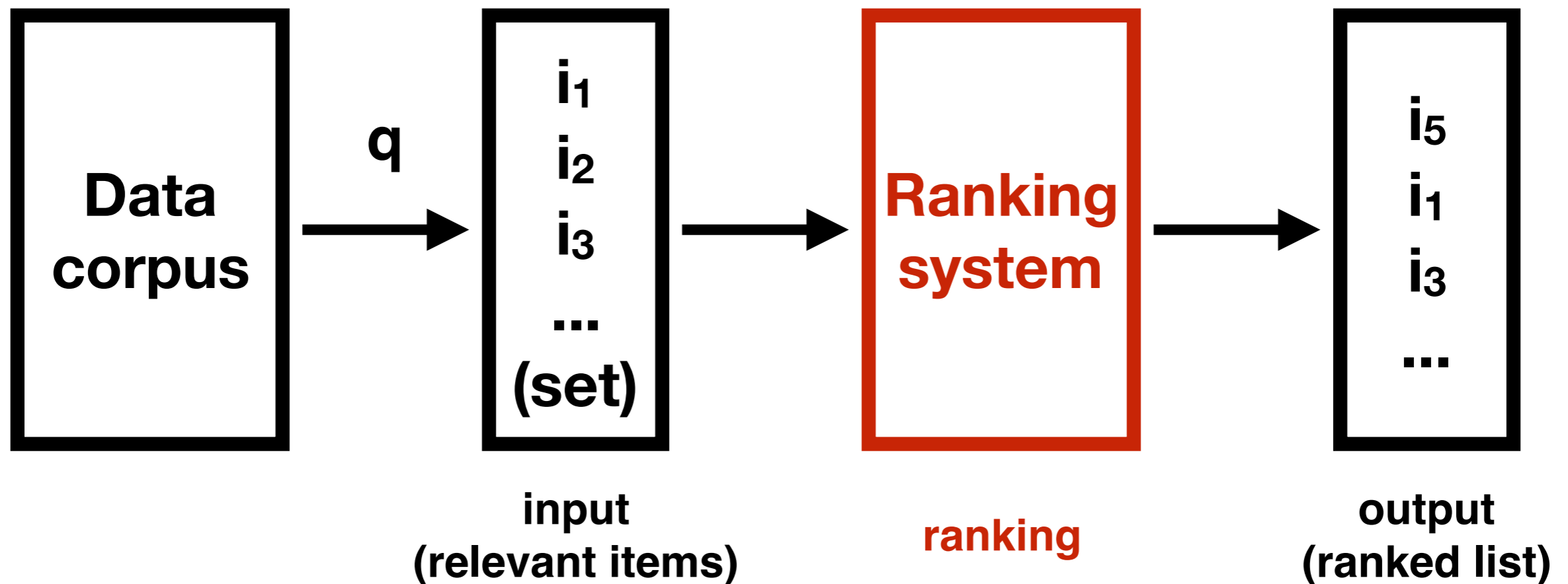
Social media search engine design



Social media search engine design



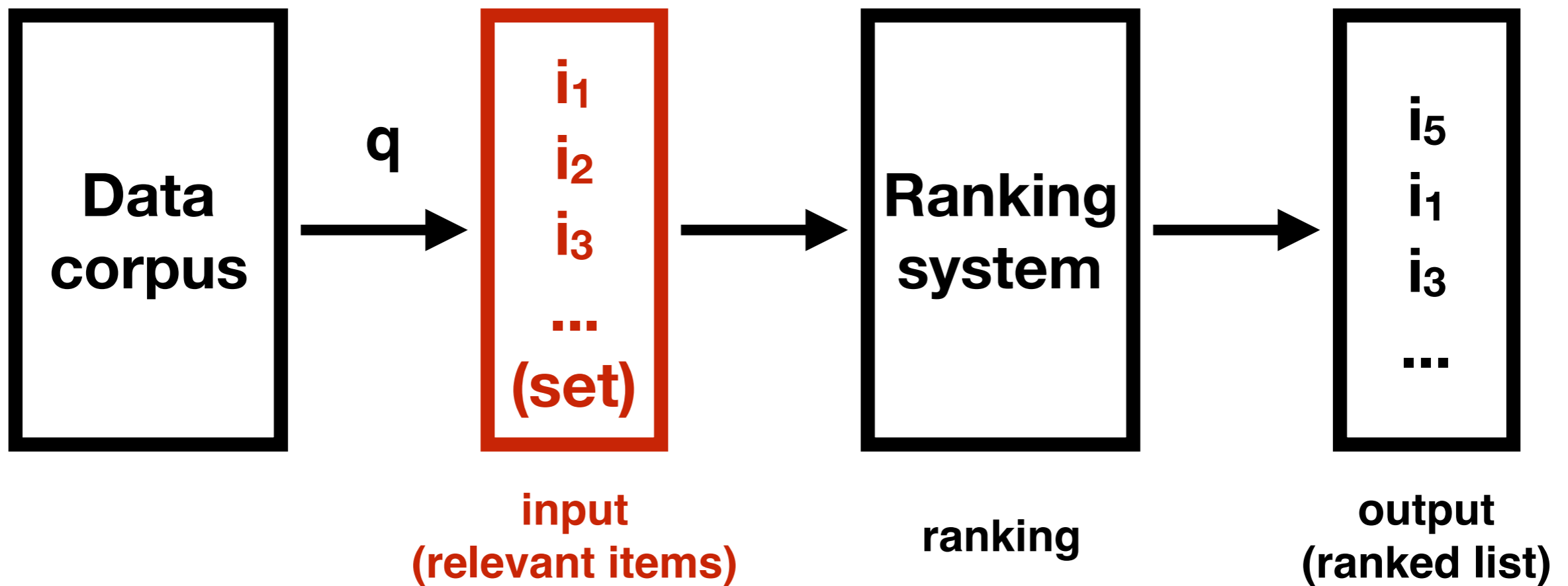
Social media search engine design



Output bias may stem from

- Bias introduced by the **ranking** system

Social media search engine design



Output bias may stem from

- Bias introduced by the **ranking** system
- Bias in the **input** relevant item set



Identify
sources of
search bias

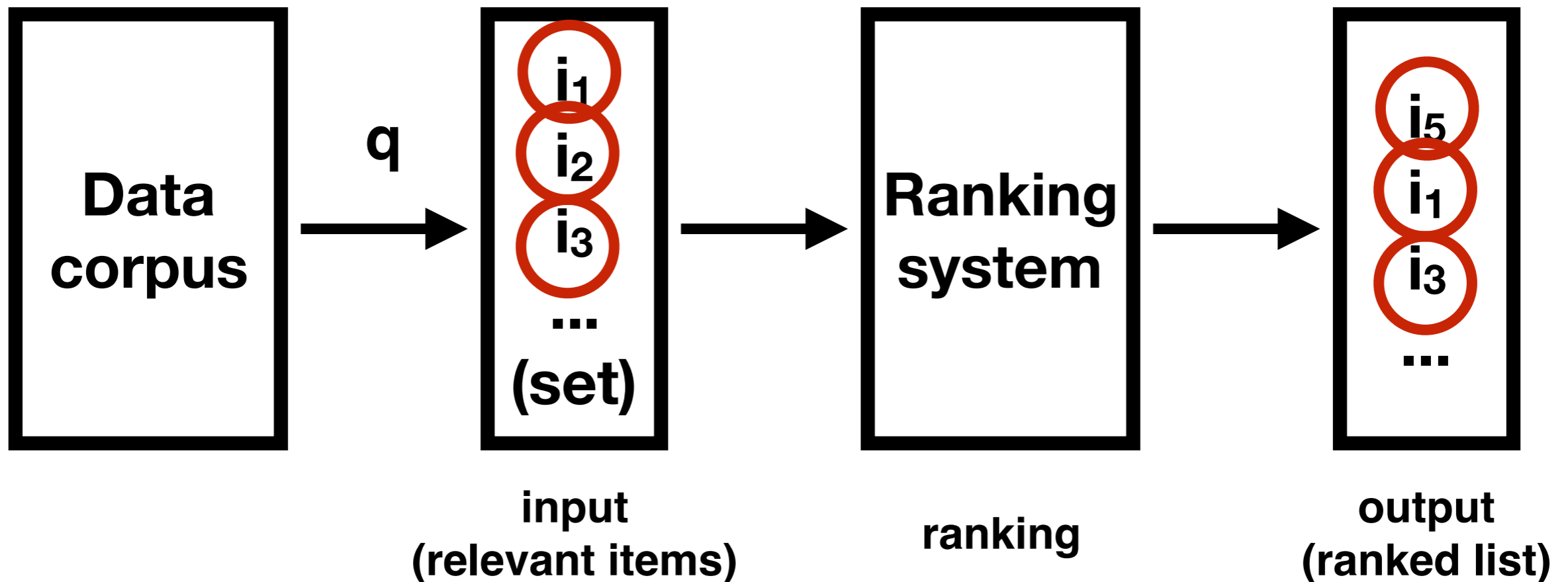


Quantify
bias of each
source



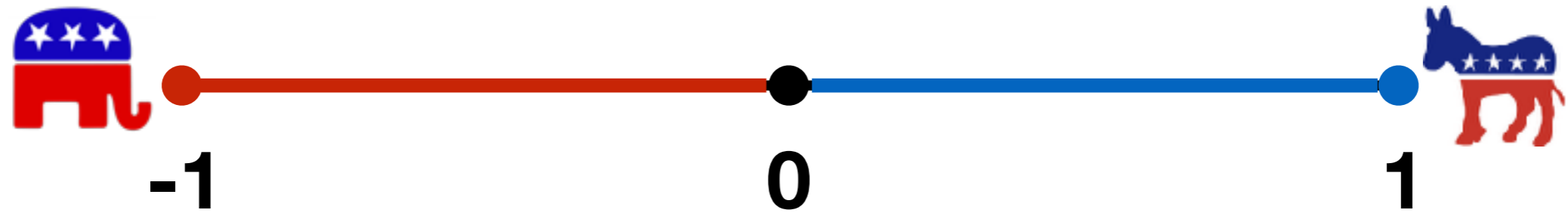
Study bias of
political
searches in
Twitter

Quantifying bias of each source

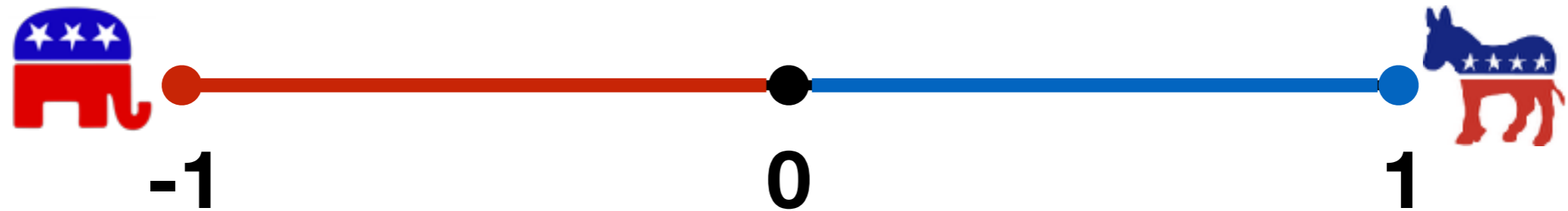


Step 1: Quantify bias of an individual item

Step 1: Quantifying bias of a single items

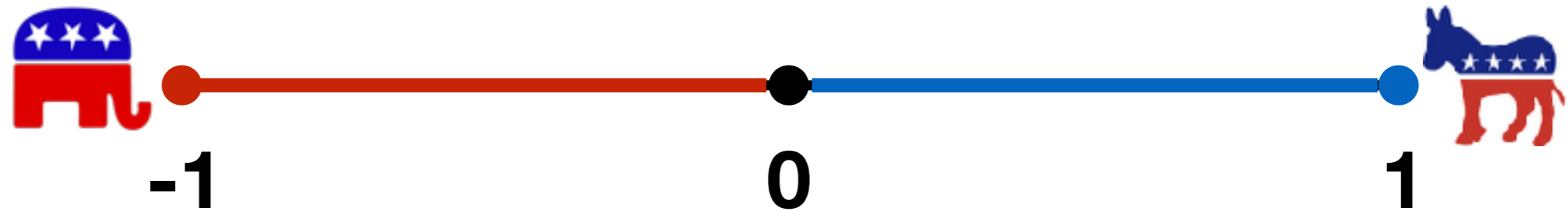


Step 1: Quantifying bias of a single items



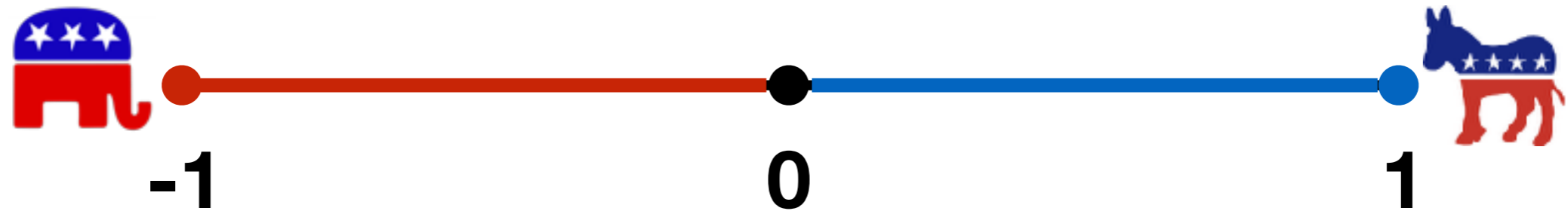
- We use **source bias** as a proxy
 - Infer bias of each individual item from the bias of the author
 - High scalability

Step 1: Quantifying bias of a single items



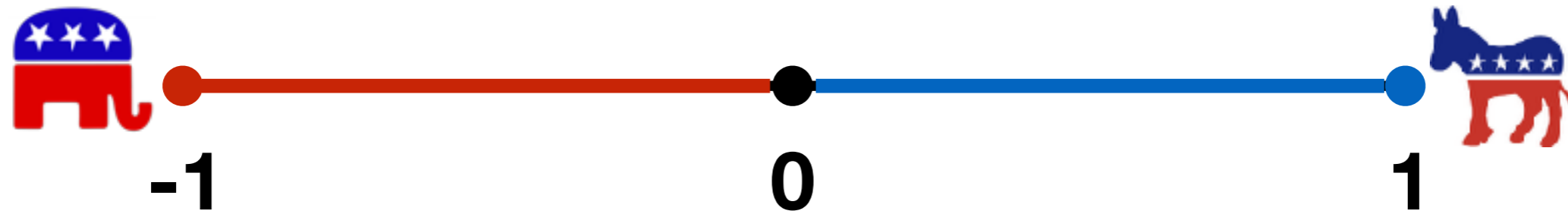
- We use **source bias** as a proxy
 - Infer bias of each individual item from the bias of the author
 - High scalability
- Prior work on inferring **content bias**
 - Could be plugged into our bias quantification framework

Step 1: Quantifying bias of a single items



Evaluation: High agreement between source and content bias (75% or more)

Step 1: Quantifying bias of a single items



Evaluation: High agreement between source and content bias (75% or more)

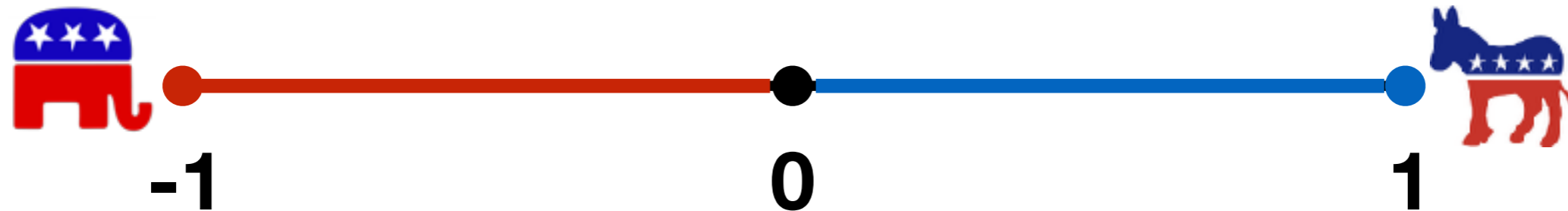


Mark Meadows 
@RepMarkMeadows


[Follow](#)



I support [@RandPaul](#) and [@RepSanfordSC](#)'s [#Obamacare](#) replacement plan -- a plan that will lower costs and put the focus back on the patient.

Step 1: Quantifying bias of a single items





Evaluation: High agreement between source and content bias (75% or more)



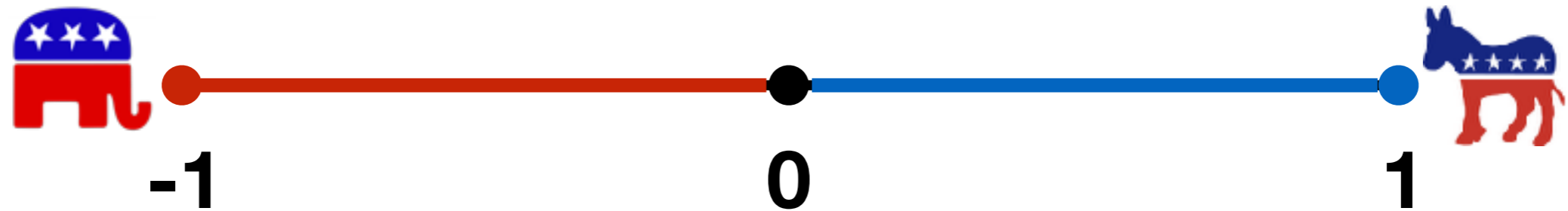
 **Mark Meadows** 
@RepMarkMeadows

*"R-NC 11th District,
Republican party"*


 Follow 

I support @RandPaul and @RepSanfordSC's #Obamacare replacement plan -- a plan that will lower costs and put the focus back on the patient.

Step 1: Quantifying bias of a single items

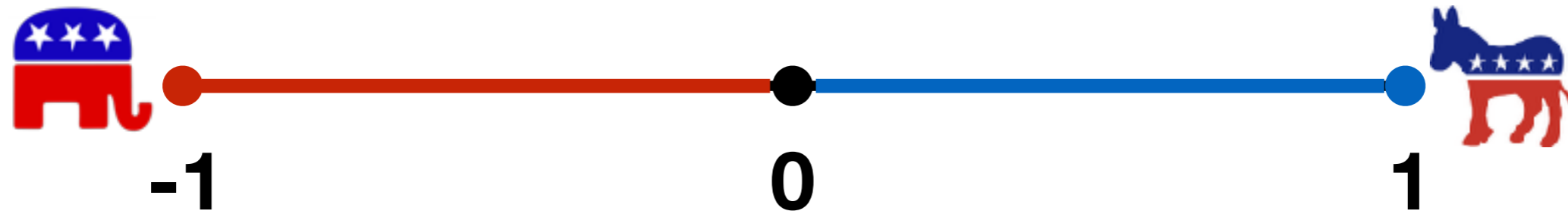


Evaluation: High agreement between source and content bias (75% or more)

 **Mitt Romney** 
@MittRomney + Follow 

If Trump had said 4 years ago the things he says today about the KKK, Muslims, Mexicans, disabled, I would NOT have accepted his endorsement

Step 1: Quantifying bias of a single items



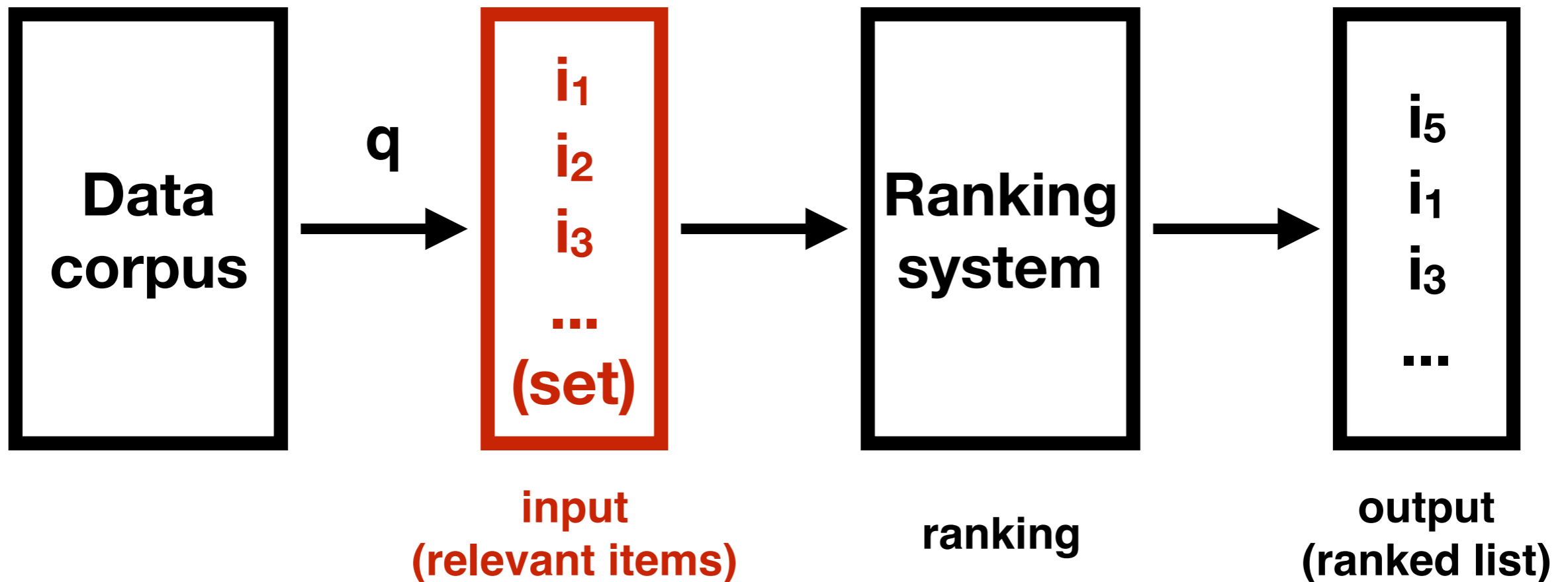
Evaluation: High agreement between source and content bias (75% or more)

  **Mitt Romney** 
@MittRomney

"Republican party nominee for Presidential elections"  

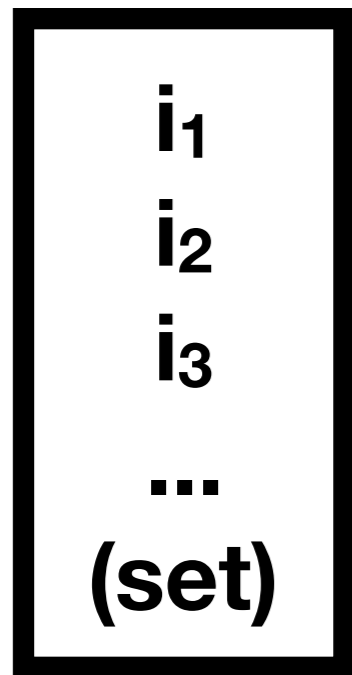
If Trump had said 4 years ago the things he says today about the KKK, Muslims, Mexicans, disabled, I would NOT have accepted his endorsement

Quantifying bias of each source



Step 2: Quantify bias of a set of items

Step 2: Quantifying bias of set of items

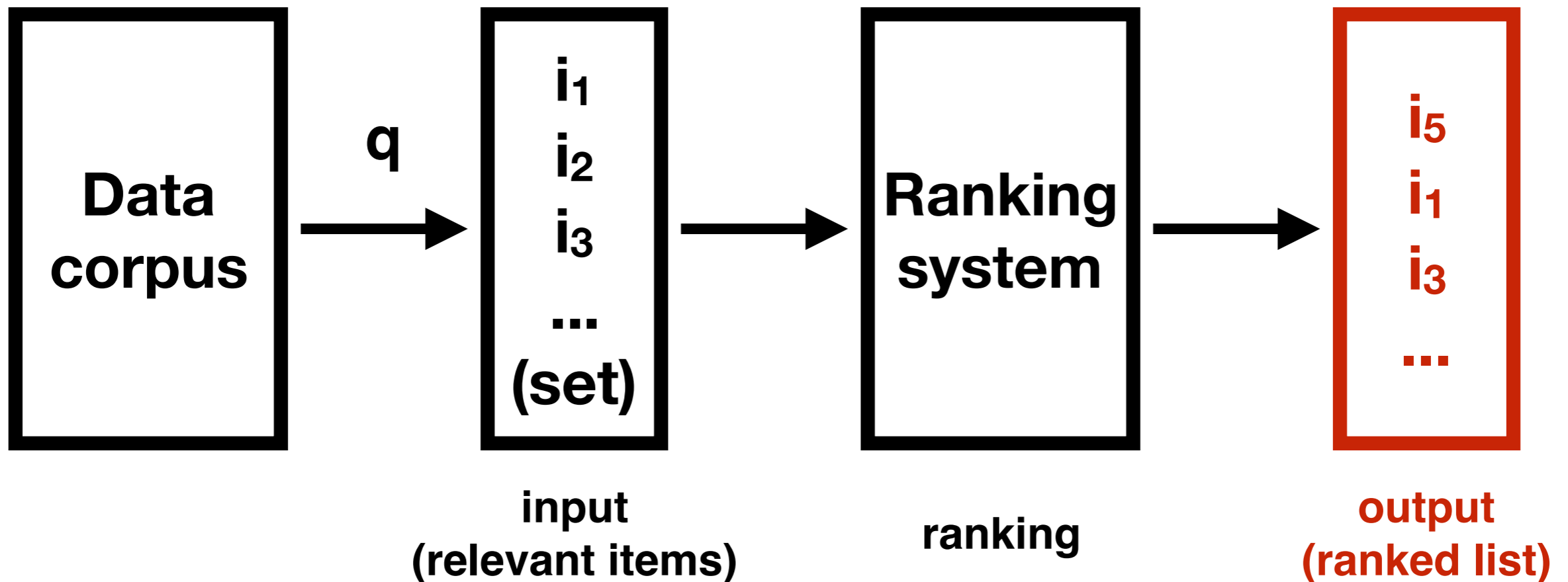


input
(relevant items)

- Compute bias score (s) for each item
- Take the average over the whole set

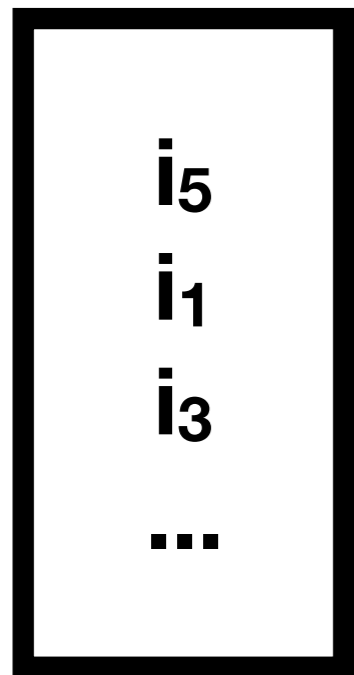
$$IB(q) = \frac{\sum_{i=1}^n s_i}{n}$$

Quantifying bias of each source



Step 3: Quantify bias of a ranked list of items

Step 3: Quantifying bias of ranked list of items



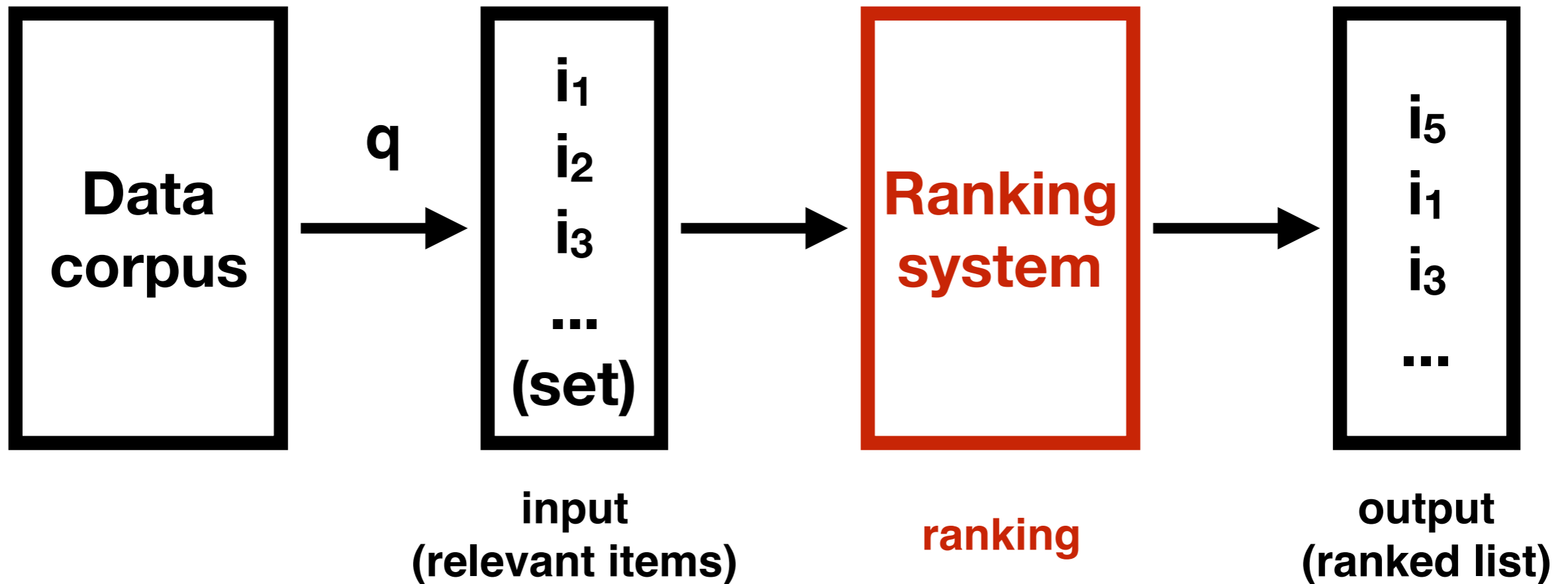
output
(ranked list)

- MAP-style measure

- Bias till rank r
$$B(q, r) = \frac{\sum_{i=1}^r s_i}{r}$$

- Output bias
$$OB(q, r) = \frac{\sum_{i=1}^r B(q, i)}{r}$$

Quantifying bias of each source



Step 4: Quantify bias introduced by ranking system

Step 4: Quantifying bias introduced by ranking



$$\text{Ranking bias} = \text{Output bias} - \text{Input bias}$$



Identify
sources of
search bias



Quantify
bias of each
source



Study bias of
political
searches in
Twitter

Studying bias for political searches in Twitter

- Search queries for
 - 2016 Democratic and Republican presidential primary debates (**#demdebate, rep debate, ...**)
 - Presidential candidates (**Hillary Clinton, Donald Trump, ...**)

Studying bias for political searches in Twitter

- Search queries for
 - 2016 Democratic and Republican presidential primary debates (**#demdebate, rep debate, ...**)
 - Presidential candidates (**Hillary Clinton, Donald Trump, ...**)
- Data collected
 - **Output:** Twitter top search snapshots every 10 mins
 - **Input:** All tweets containing the query, using streaming api

Studying bias for political searches in Twitter

- Search queries for
 - 2016 Democratic and Republican presidential primary debates (**#demdebate, rep debate, ...**)
 - Presidential candidates (**Hillary Clinton, Donald Trump, ...**)
- Data collected
 - **Output:** Twitter top search snapshots every 10 mins
 - **Input:** All tweets containing the query, using streaming api

Non-personalized search data

Studying bias for political searches in Twitter

- Search queries for
 - 2016 Democratic and Republican presidential primary debates (**#demdebate, rep debate, ...**)
 - Presidential candidates (**Hillary Clinton, Donald Trump, ...**)
- Data collected
 - **Output:** Twitter top search snapshots every 10 mins
 - **Input:** All tweets containing the query, using streaming api
- **Computed input, ranking, and output bias** for each of the 25 queries

Bias in Twitter search: Input bias vs. Ranking bias

Bias in Twitter search: Input bias vs. Ranking bias

- Does input bias matter?

Impact of input bias

Query	Output Bias	Input Bias
Bernie Sanders	0.71	0.55
Martin O'Malley	0.64	0.57
Rand Paul	-0.37	-0.18
John Kasich	-0.09	-0.13
dem debate	0.52	0.29
#demdebate	0.57	0.56
republican debate	0.53	0.27
rep debate	0.31	0.40

Impact of input bias

Query	Output Bias	Input Bias
Bernie Sanders	0.71	0.55
Martin O'Malley	0.64	0.57
Rand Paul	-0.37	-0.18
John Kasich	-0.09	-0.13
dem debate	0.52	0.29
#demdebate	0.57	0.56
republican debate	0.53	0.27
rep debate	0.31	0.40

Input bias matters!

Impact of input bias

Query	Output Bias	Input Bias
Bernie Sanders	0.71	0.55
Martin O'Malley	0.64	0.57
Rand Paul	-0.37	-0.18
John Kasich	-0.09	-0.13
dem debate	0.52	0.29
#demdebate	0.57	0.56
republican debate	0.53	0.27
rep debate	0.31	0.40

Input bias varies across queries

Effect of query phrasing

Query	Input Bias
dem debate	0.29
#demdebate	0.56
republican debate	0.27
rep debate	0.40

Effect of query phrasing

Query	Input Bias
dem debate	0.29
#demdebate	0.56
republican debate	0.27
rep debate	0.40

Even for the same event, query phrasing can greatly effect the bias

Bias in Twitter search: Input bias vs. Ranking bias

- Does input bias matter?
 - Input bias does matter
 - Can vary significantly based on the query
 - Even for the same event, different phrasings of queries have widely differing biases

Bias in Twitter search: Input bias vs. Ranking bias

- Does input bias matter?
 - Input bias does matter
 - Can vary significantly based on the query
 - Even for the same event, different phrasings of queries have widely differing biases
- Does the ranking bias exist?

Examining ranking bias

Query	Ranking Bias
Hillary Clinton	0.18
Bernie Sanders	0.16
Martin O'Malley	0.07
Donald Trump	0.10
Ted Cruz	-0.37
Marco Rubio	-0.29
Ben Carson	0.26

Examining ranking bias

Query	Ranking Bias
Hillary Clinton	0.18
Bernie Sanders	0.16
Martin O'Malley	0.07
Donald Trump	0.10
Ted Cruz	-0.37
Marco Rubio	-0.29
Ben Carson	0.26

Ranking bias does exist...

Examining ranking bias

Query	Ranking Bias
Hillary Clinton	0.18
Bernie Sanders	0.16
Martin O'Malley	0.07
Donald Trump	0.10
Ted Cruz	-0.37
Marco Rubio	-0.29
Ben Carson	0.26

... but no evidence of systemic bias

Bias in Twitter search: Input bias vs. Ranking bias

- Does input bias matter?
 - Input bias does matter
 - Can vary significantly based on the query
 - Even for the same event, different phrasings of queries have widely differing biases
- Does the ranking bias exist?
 - Yes and varies across queries
 - No evidence of systemic bias

Open challenge: How to address search bias?

Open challenge: How to address search bias?

- Modify ranking system to account for bias
 - Might lead to reduction in quality of results

Open challenge: How to address search bias?

- Modify ranking system to account for bias
 - Might lead to reduction in quality of results
- Make the bias transparent
 - Keep the current ranking
 - Inform the users about the bias they are seeing
 - Make biases related to query phrasing transparent



Ted Cruz

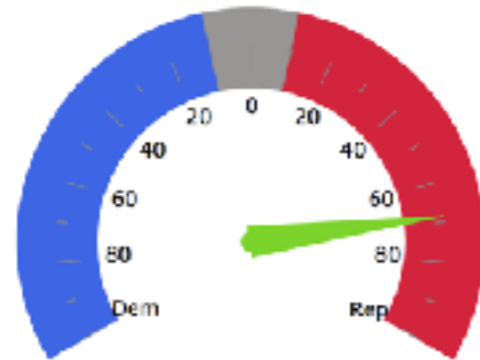
Houston, Texas

Father of two, [@heidiscruz](#)'s husband, fighter for liberty. Representing the great state of Texas in the U.S. Senate.

tedcruz.org

Joined Mar 2009

Ted Cruz is inferred to be republican leaning



<https://tinyurl.com/bias-users>



Ted Cruz ✓

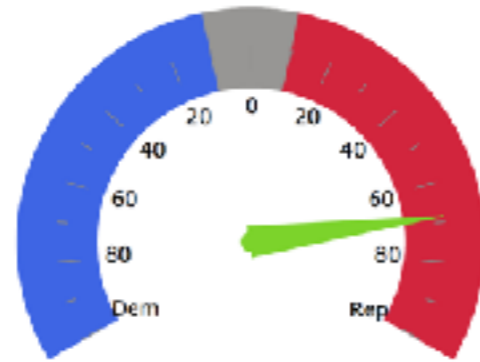
📍 Houston, Texas

👤 Father of two, @heidiscruz's husband, fighter for liberty. Representing the great state of Texas in the U.S. Senate.

🌐 tedcruz.org

📅 Joined Mar 2009

Ted Cruz is inferred to be republican leaning



John Oliver ✓

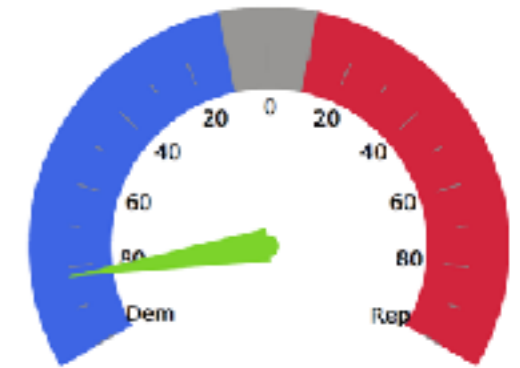
📍 New York

👤 Comedian. @LastWeekTonight, @TheDailyShow, The Bugle Podcast (@hallobuglers)

🌐 iamjohnoliver.com

📅 Joined Jun 2011

John Oliver is inferred to be democrat leaning



<https://tinyurl.com/bias-users>



Ted Cruz

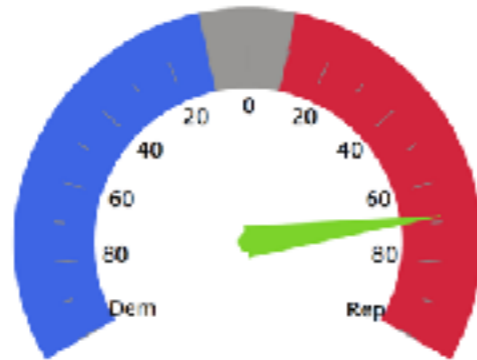
Houston, Texas

Father of two, @heidiscruz's husband, fighter for liberty. Representing the great state of Texas in the U.S. Senate.

tedcruz.org

Joined Mar 2009

Ted Cruz is inferred to be republican leaning



John Oliver

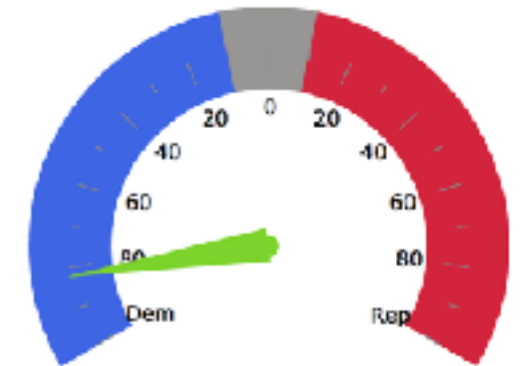
New York

Comedian. @LastWeekTonight, @TheDailyShow, The Bugle Podcast (@hallobuglers)

ianjohnoliver.com

Joined Jun 2011

John Oliver is inferred to be democrat leaning



Sean P. Goggins

Columbia, MO

Sociotechnical Data Scientist

seangoggins.net

Joined Aug 2014

Sean P. Goggins is inferred to be democrat leaning



<https://tinyurl.com/bias-users>