Bachelor`s thesis defense

Semi-supervised Cleansing of Web-based Argument Corpora

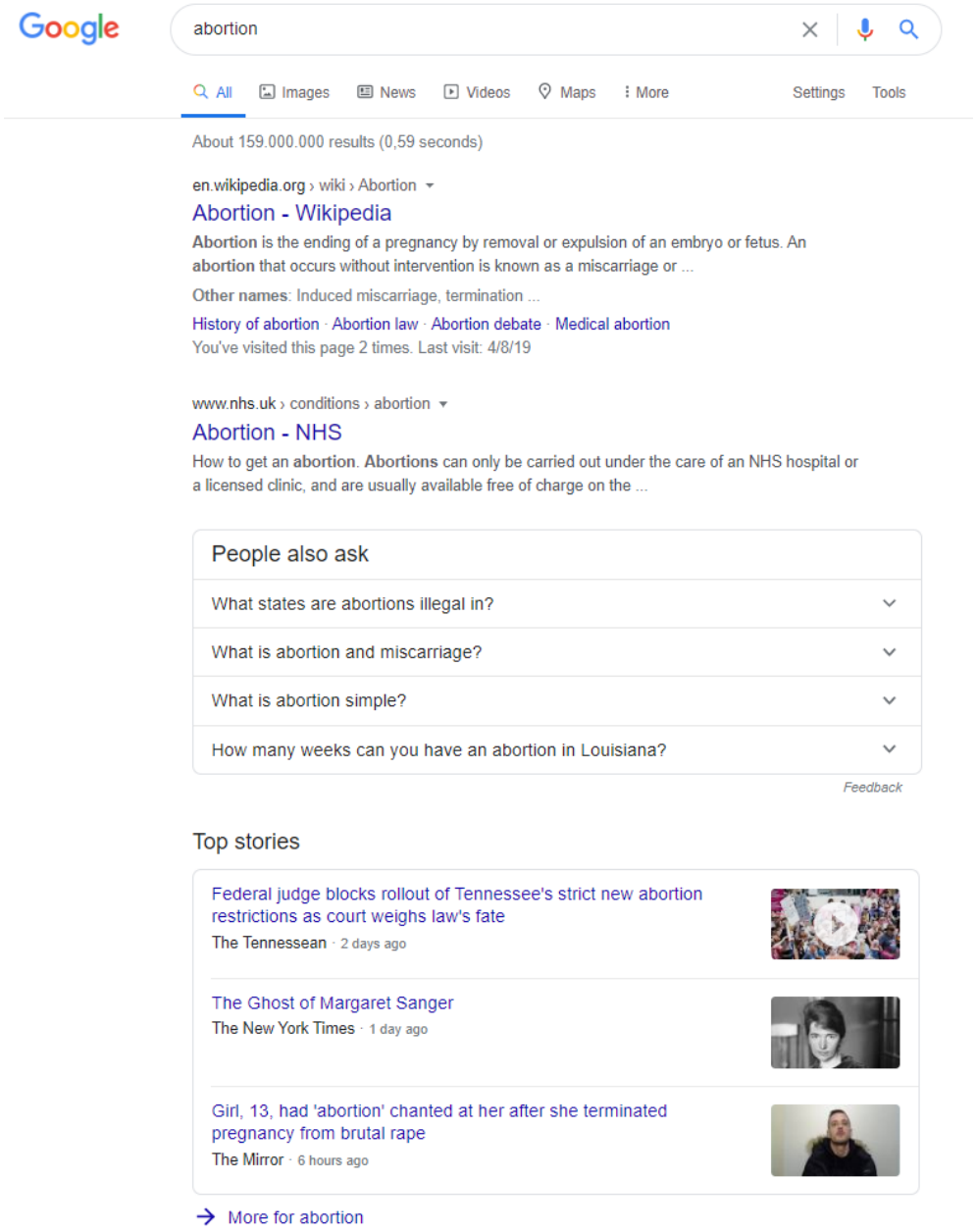
Jonas Dorsch

September 09, 2020

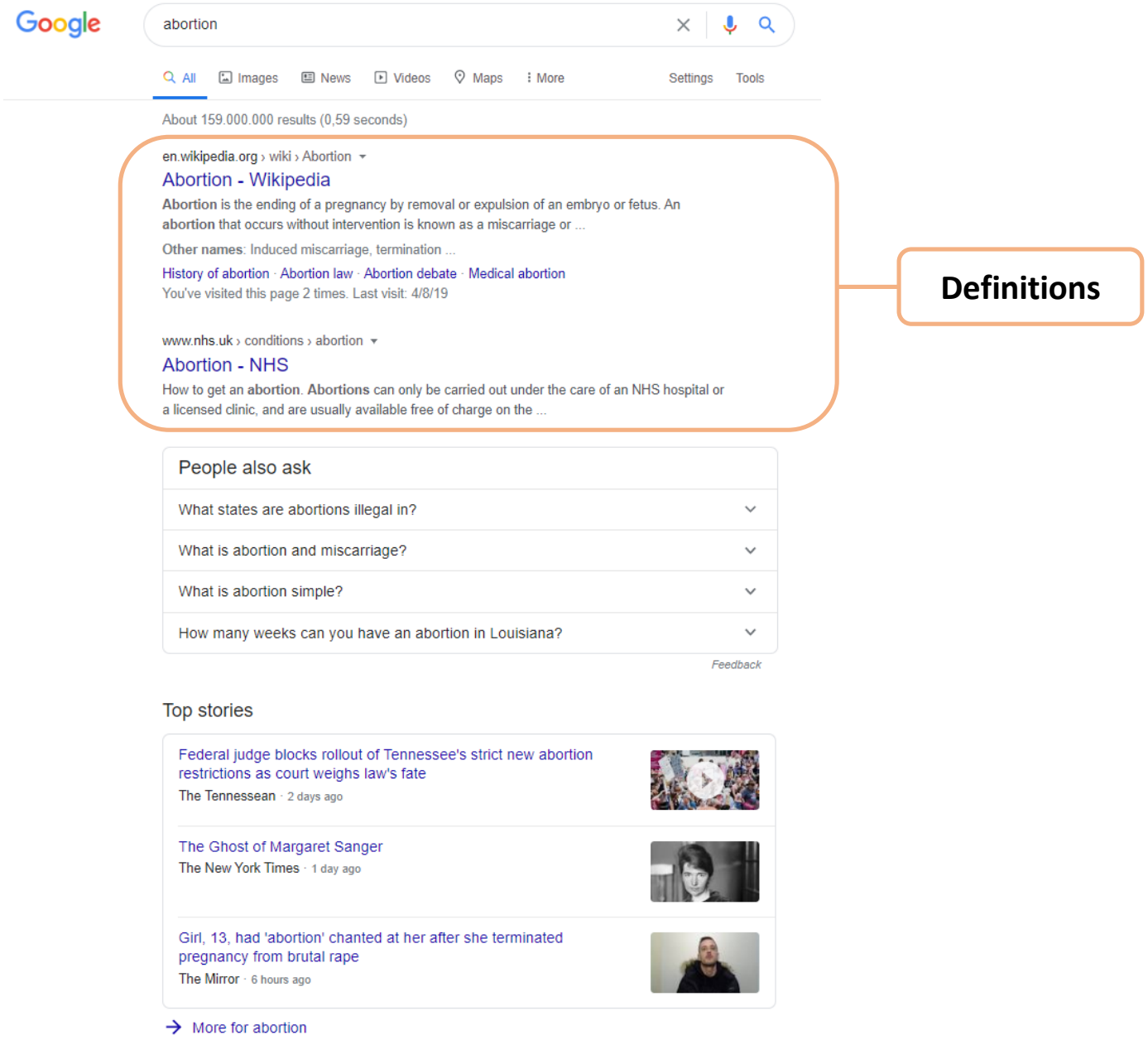
Contents

- Introduction
- Approach
- Evaluation
- Conclusion

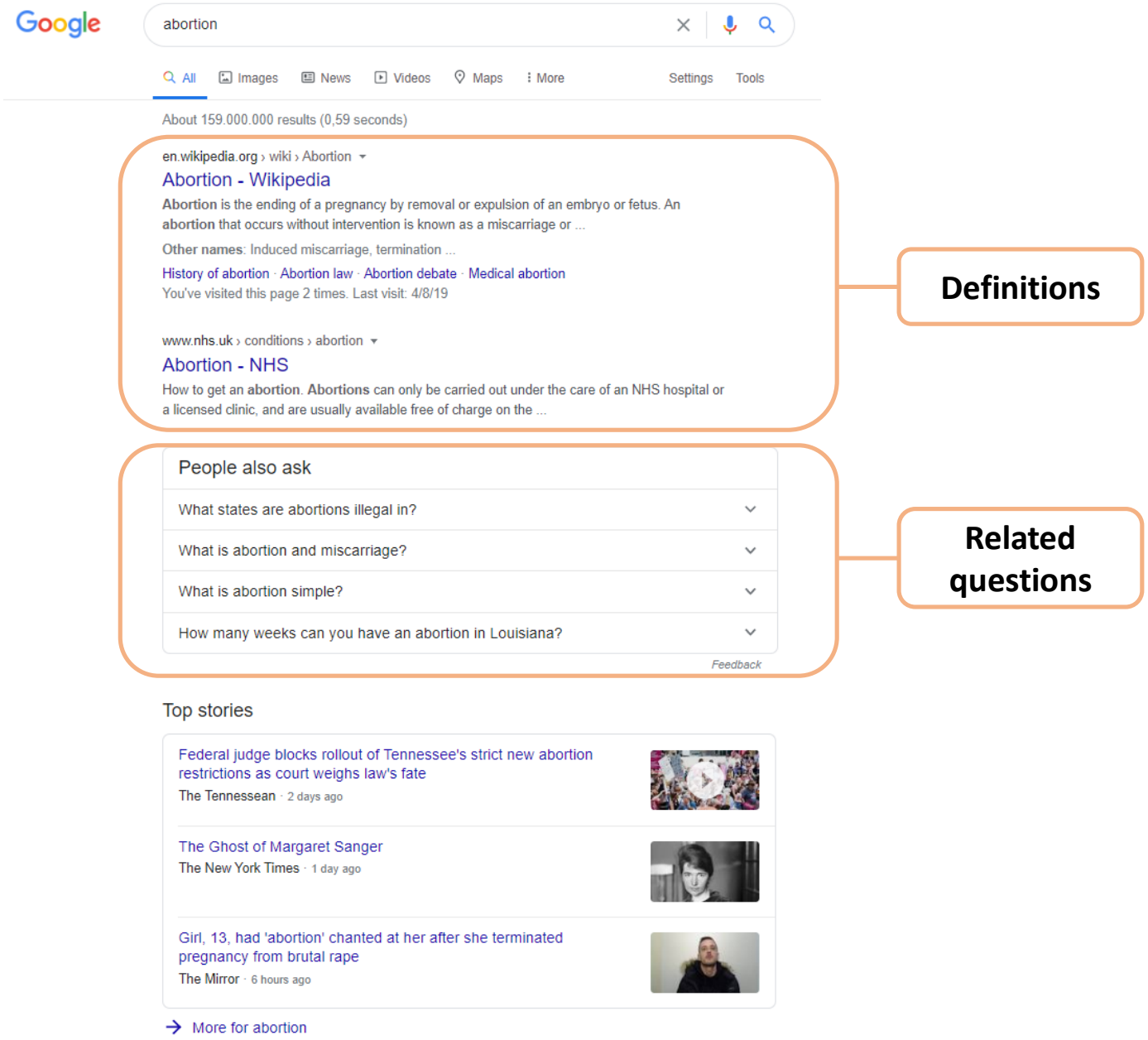
Introduction



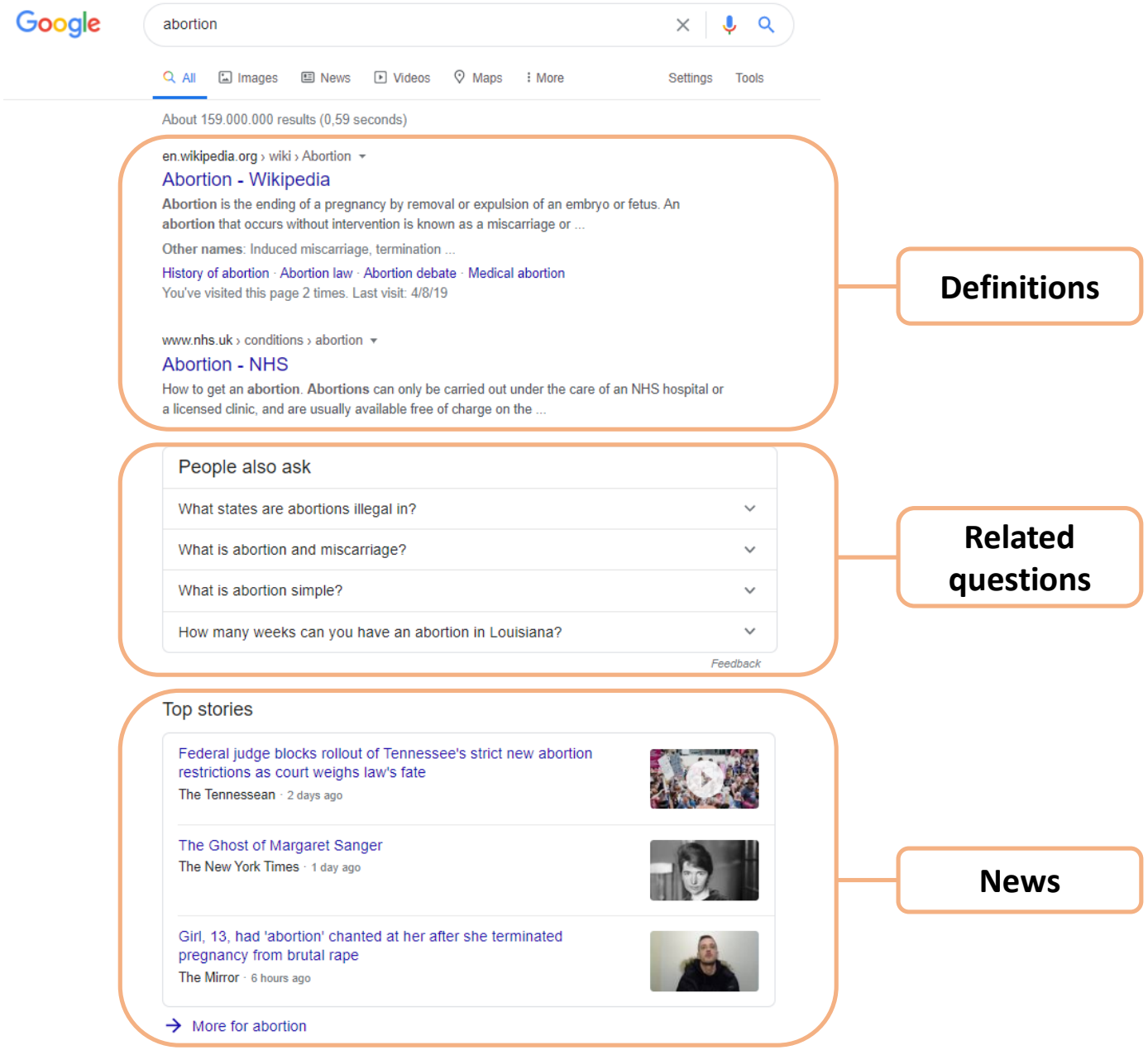
Introduction



Introduction



Introduction



Introduction

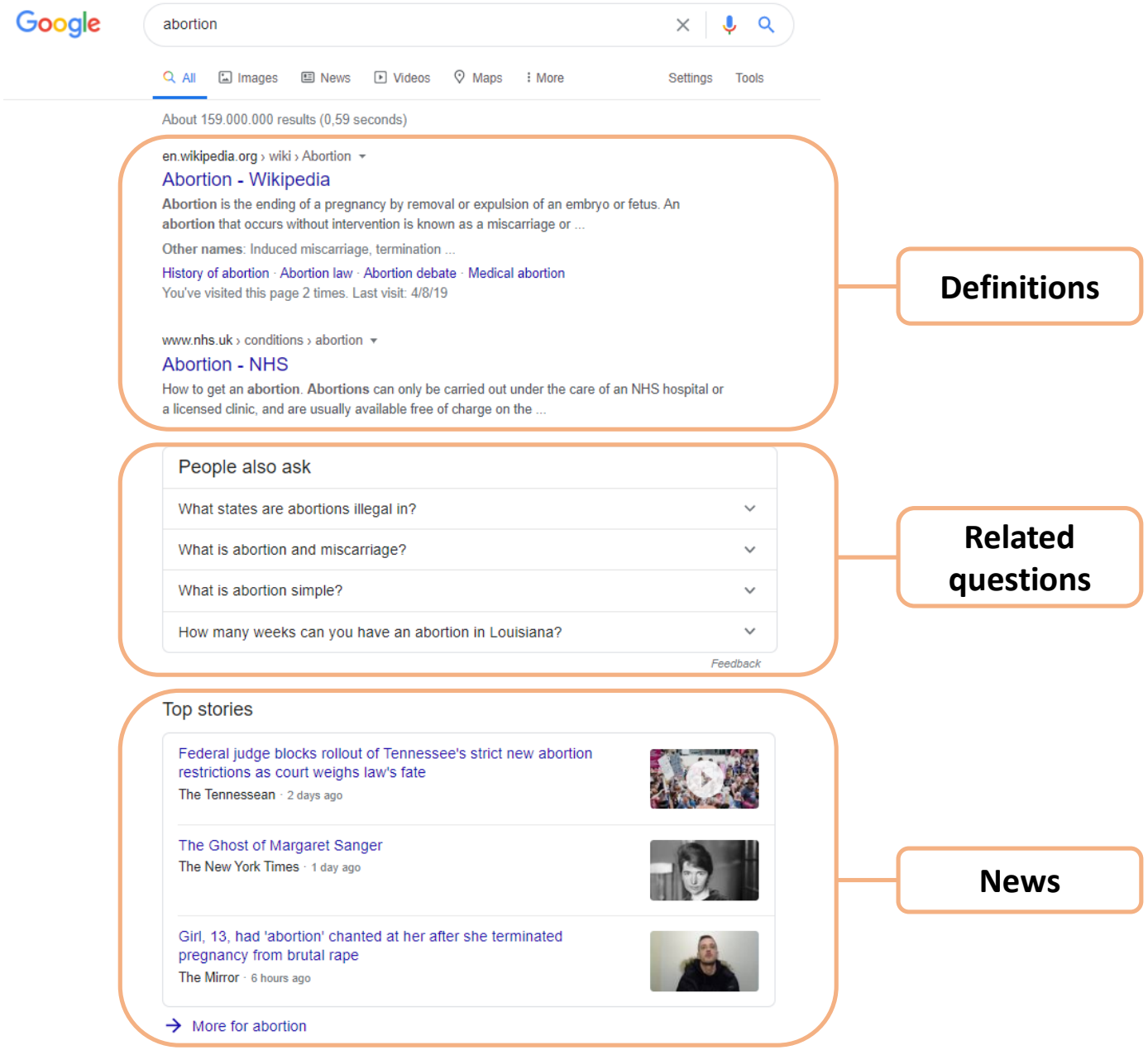
Google

Good at:

Presenting factual data

Bad at:

Presenting argumenative data



Do we need argumentation?

Introduction – Do we need argumentation?

- Essential part of:
  - Politics

Source: https://en.wikipedia.org/wiki/European_Parliament#/media/File:Bundesarchiv_B_145_Bild-F023908-0002,_Stra%C3%9Fburg,_Tagung_des_Europarates.jpg

Introduction – Do we need argumentation?

- Essential part of:
  - Politics
  - Work
  - Education

Source: https://pixabay.com/photos/workplace-team-business-meeting-1245776/

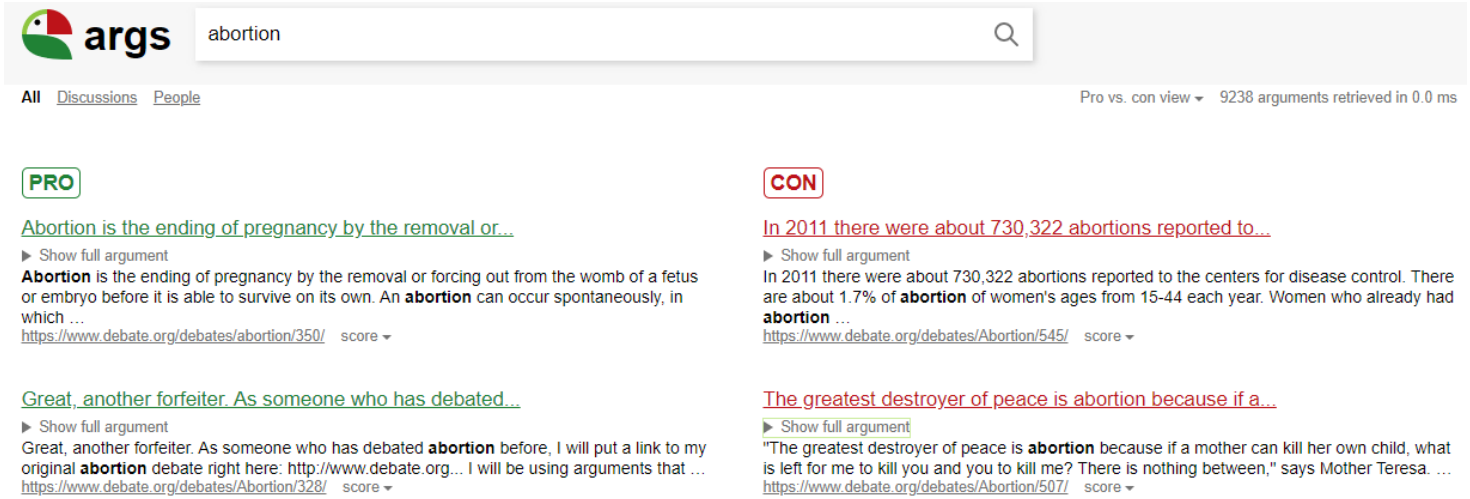
Introduction – Do we need argumentation?

- Essential part of:
  - Politics
  - Work
  - Education
  - Relationships
  - Social life
  - etc.

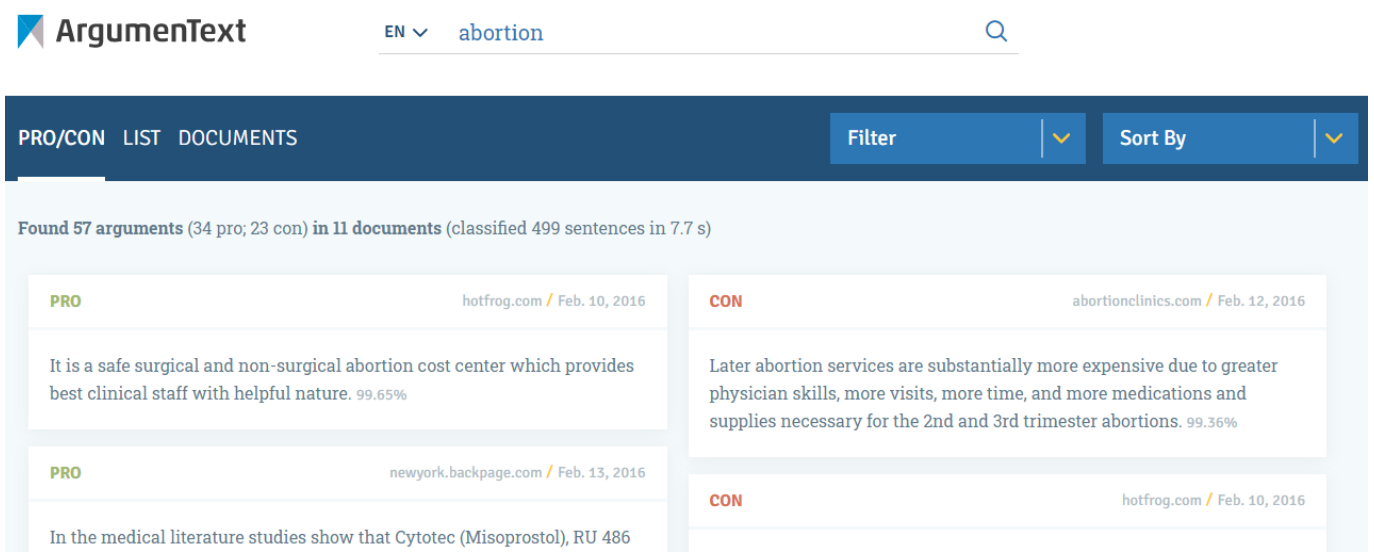
Source: https://pixabay.com/illustrations/conversation-talk-talking-people-799448/

So why don´t make arguments accessible online?

Introduction – Argument search engines



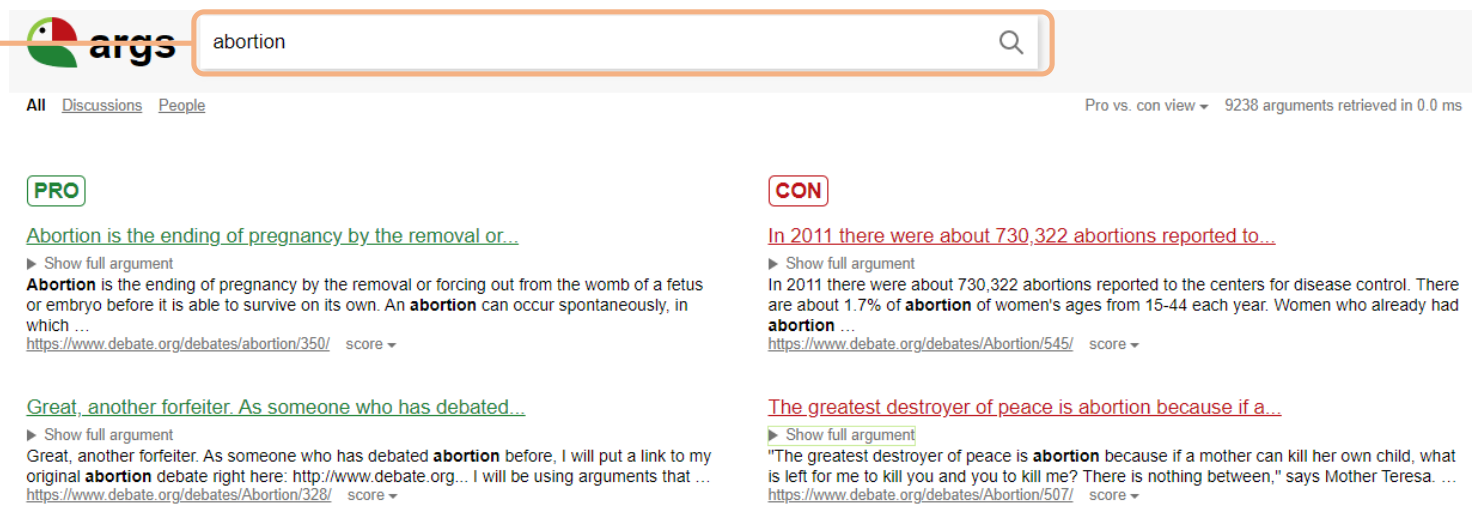
Source: https://www.args.me/index.html



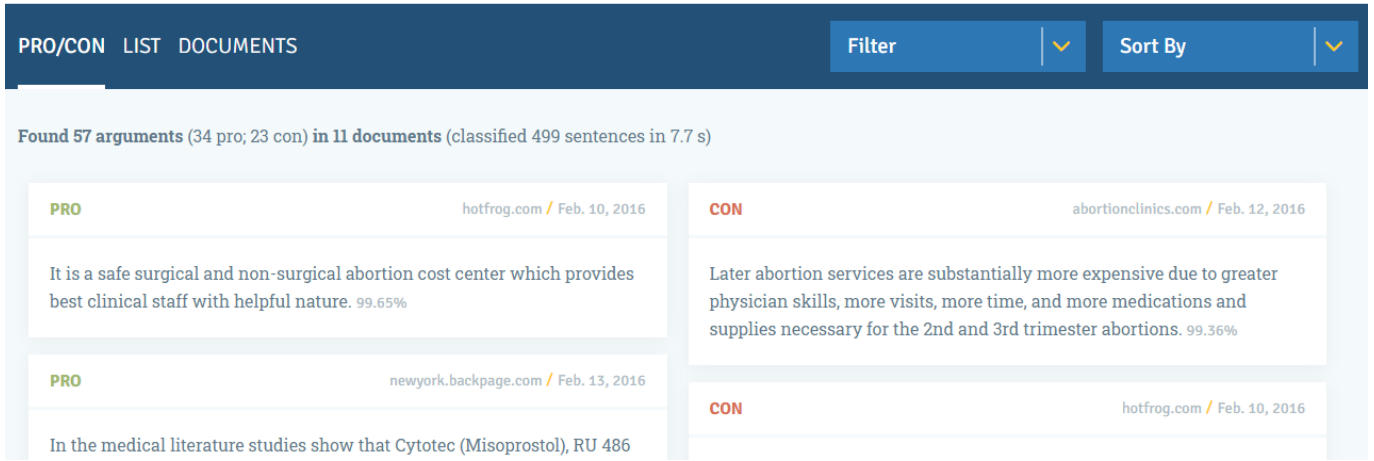
Source: https://www.argumentsearch.com/

Introduction – Argument search engines

Search Field

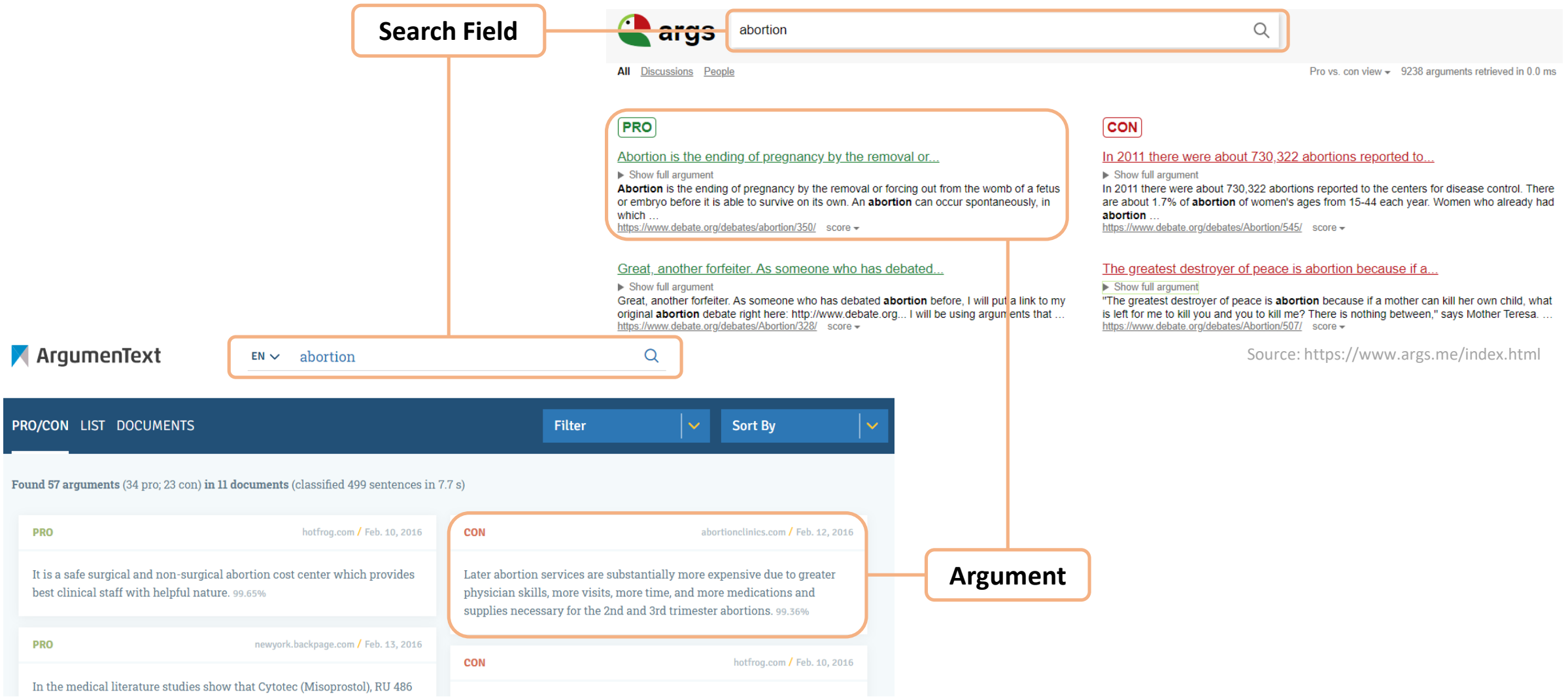


Source: https://www.args.me/index.html



Source: https://www.argumentsearch.com/

Introduction – Argument search engines

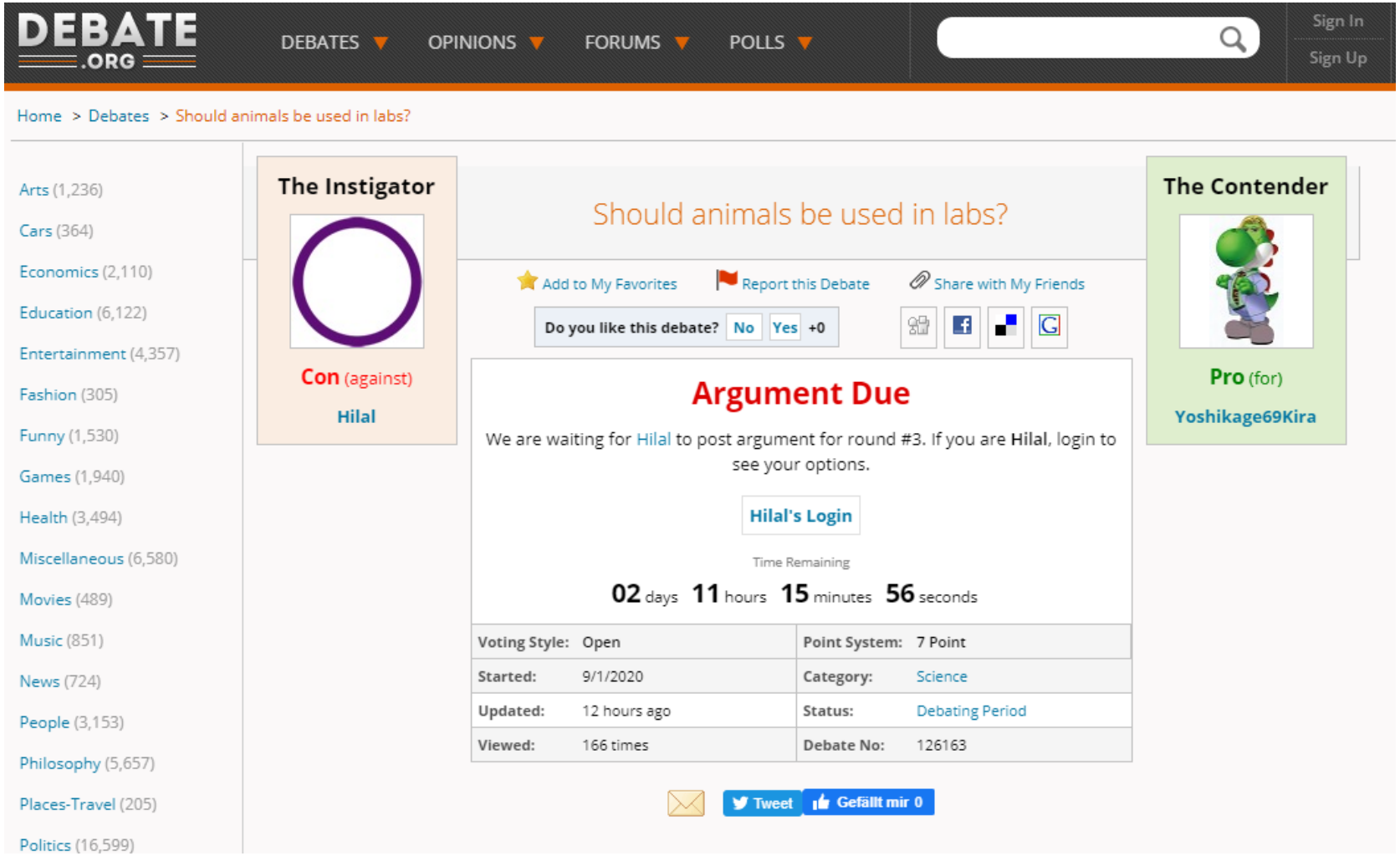


Source: https://www.args.me/index.html

Source: https://www.argumentsearch.com/

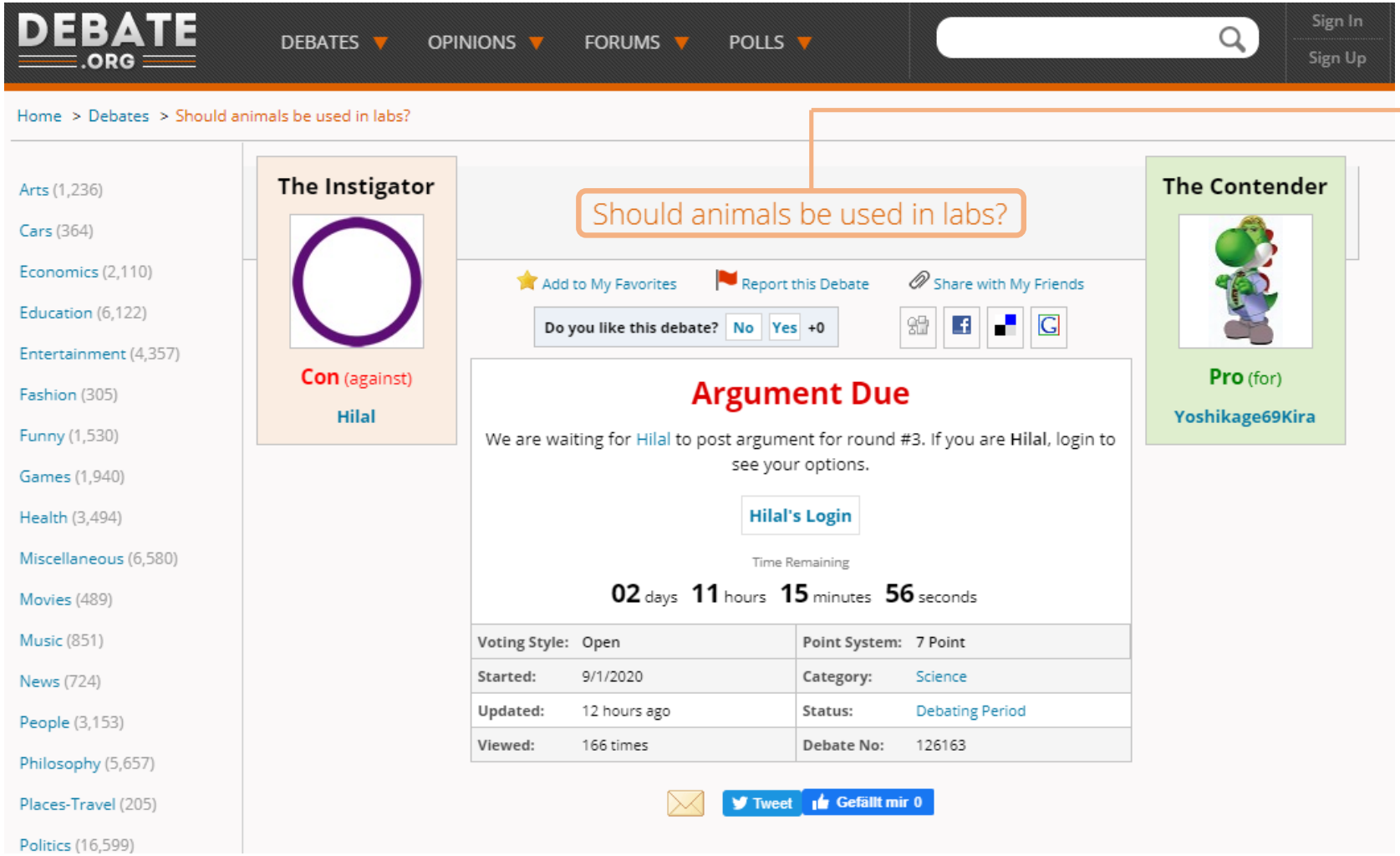
So where does the data come from?

Introduction – Argument Corpora



Source: https://www.debate.org/debates/Should-animals-be-used-in-labs/1/

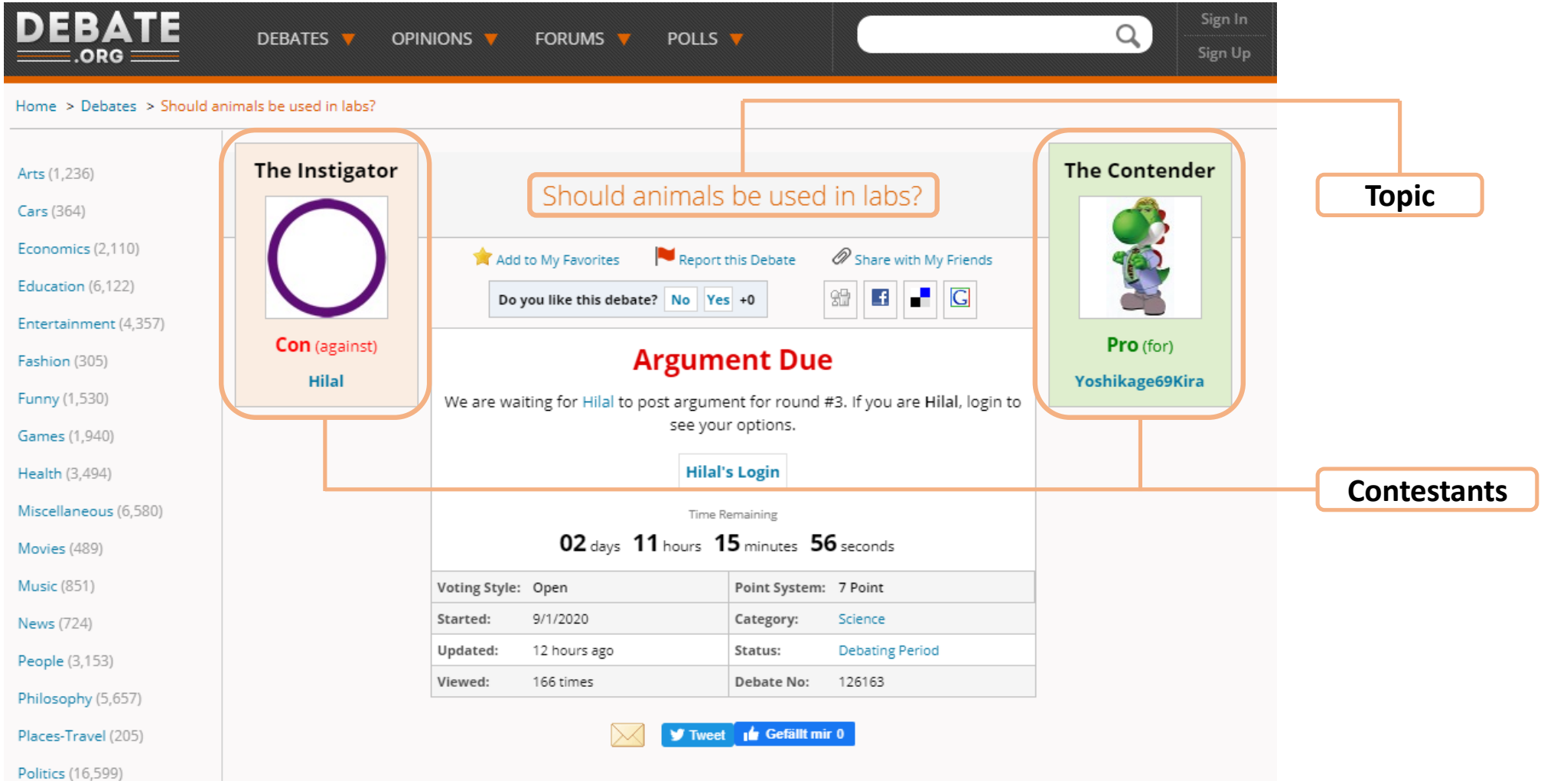
Introduction – Argument Corpora



Topic

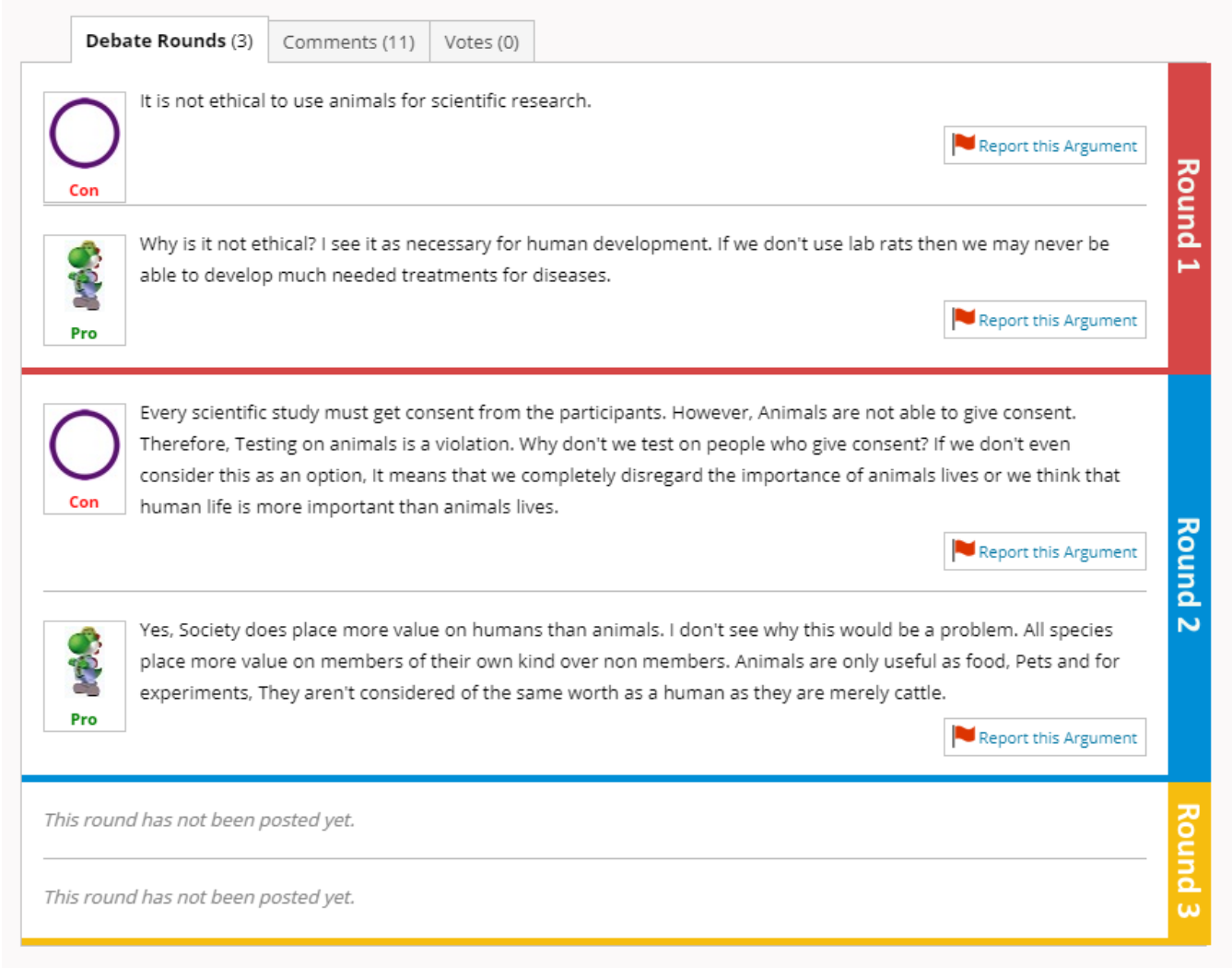
Source: https://www.debate.org/debates/Should-animals-be-used-in-labs/1/

Introduction – Argument Corpora



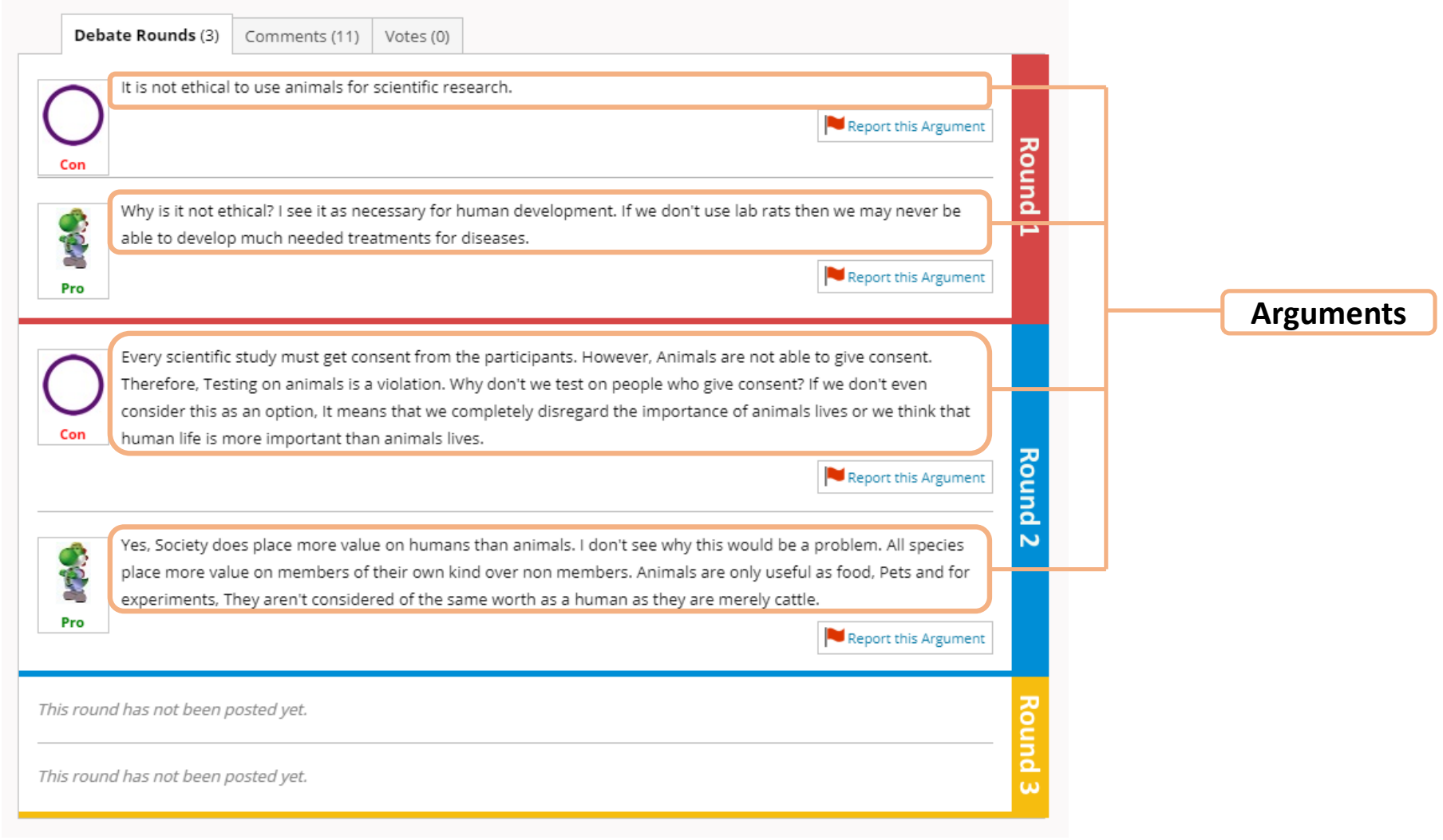
Source: https://www.debate.org/debates/Should-animals-be-used-in-labs/1/

Introduction – Argument Corpora



Source: https://www.debate.org/debates/Should-animals-be-used-in-labs/1/

Introduction – Argument Corpora



Source: https://www.debate.org/debates/Should-animals-be-used-in-labs/1/

Introduction – The args.me corpus

- Published by Ajjour et al. [2019]
- 4 debate portals
- 390k arguments
- 60k debates

(a) Query	Absolute	Relative
climate change	251	3.5%
feminism	193	2.7%
abortion	158	2.2%
trump	146	2.0%
brexit	128	1.8%
death penalty	73	1.0%
google	58	0.8%
vegan	57	0.8%
nuclear energy	56	0.8%
donald trump	47	0.7%

(b) Conclusion	Absolute	Relative
Abortion	2 838	0.7%
Gay Marriage	1 558	0.4%
Rap Battle	1 396	0.4%
Death Penalty	972	0.3%
God exists	790	0.2%
Gun control	719	0.1%
Ivf debate	432	0.1%
Animal testing	372	0.1%
I will not contradict myself	357	0.1%
Euthanasia	331	0.1%

(a) Top 10 queries from args.me query log (from Sept 2017 to May 2019).
(b) Top 10 conclusions of arguments in the args.me corpus.
Presented in Ajjour et al. [2019]

Source: https://www.debate.org/debates/Should-animals-be-used-in-labs/1/

So why do we have to cleanse the corpus?

Source: https://www.debate.org/debates/Should-animals-be-used-in-labs/1/

Introduction – Corpus cleansing

Argumentative debate post pro *"abortion"*

https://www.debate.org/debates/Abortion/526/

Thank you for bringing up this interesting debate topic, and I look forward to an exciting debate.

This has been repeated a million times but I'm going to say it again because it must act as the foundation of this debate. Abortion is not murder, especially when it is performed before the fetus has developed into a human being. When an abortion is performed, no baby is killed or murdered. Think of it as destroying the seed before it becomes a plant. Performing an abortion is by definition not committing murder. With this fact as a base or foundation, throughout the next few rounds I will argue that abortion does not demean the value of human life, but protects it -- using facts and rational arguments.

I will also argue that mother"s, being the carrier of the fetus, have the right and responsibility to decide and weigh the advantages and disadvantages of bringing a human being into the world. Let me ask you. How is it better to have the child and allow it to lead a poor and neglected life, rather than never existing in the first place? If you value life, and the sanctity of living, then you will accept abortion.
You don"t have to practice it, advocate for it, understand it, or even respect it. But if you respect life, then you will accept abortion and allow other women to have abortions.

Thank you, I look forward to your response.

Source: https://www.debate.org/debates/Should-animals-be-used-in-labs/1/

Introduction – Corpus cleansing

Argumentative debate post pro *"abortion"*

https://www.debate.org/debates/Abortion/526/

Irrelevant Sentences

Thank you for bringing up this interesting debate topic, and I look forward to an exciting debate.

This has been repeated a million times but I'm going to say it again because it must act as the foundation of this debate. Abortion is not murder, especially when it is performed before the fetus has developed into a human being. When an abortion is performed, no baby is killed or murdered. Think of it as destroying the seed before it becomes a plant. Performing an abortion is by definition not committing murder. With this fact as a base or foundation, throughout the next few rounds I will argue that abortion does not demean the value of human life, but protects it -- using facts and rational arguments.

I will also argue that mother"s, being the carrier of the fetus, have the right and responsibility to decide and weigh the advantages and disadvantages of bringing a human being into the world. Let me ask you. How is it better to have the child and allow it to lead a poor and neglected life, rather than never existing in the first place? If you value life, and the sanctity of living, then you will accept abortion.
You don"t have to practice it, advocate for it, understand it, or even respect it. But if you respect life, then you will accept abortion and allow other women to have abortions.

Thank you, I look forward to your response.

Source: https://www.debate.org/debates/Should-animals-be-used-in-labs/1/

Approach

Approach – Goal

Remove as much irrelevant sentences as possible without loosing any relevant information.

Approach – Goal

Remove as much irrelevant sentences as possible without loosing any relevant information.

1. High precision

Approach – Goal

Remove as much irrelevant sentences as possible without loosing any relevant information.

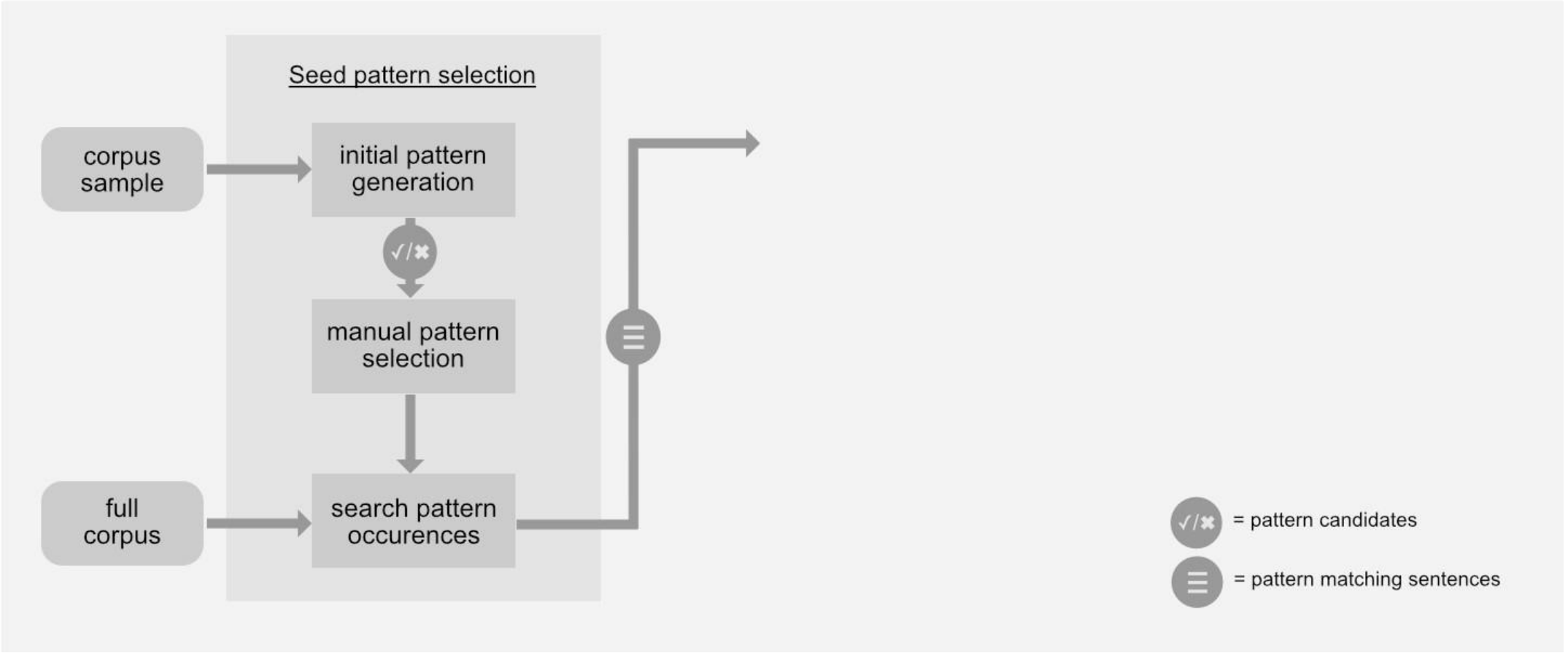
1. High precision
2. High recall

Approach – Goal

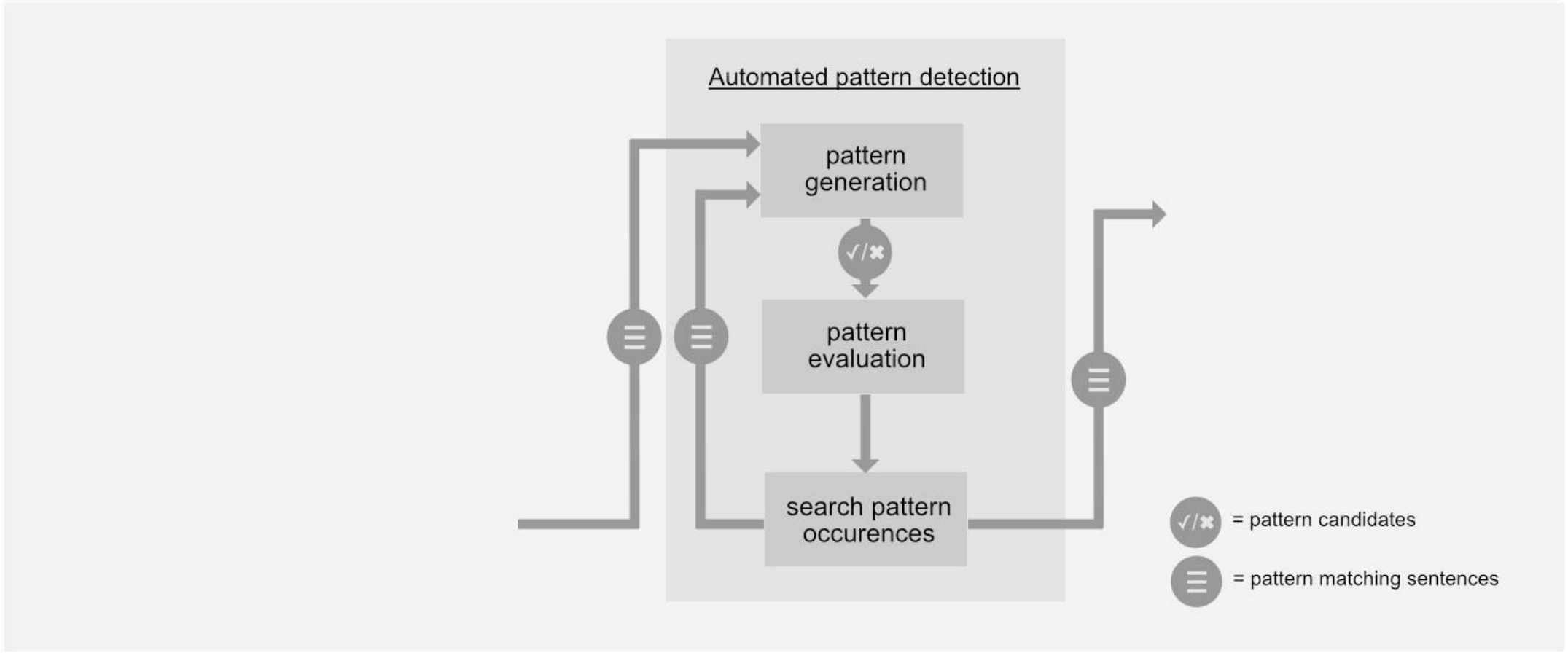
Remove as much irrelevant sentences as possible without loosing any relevant information.

1. High precision
2. High recall
3. Automated

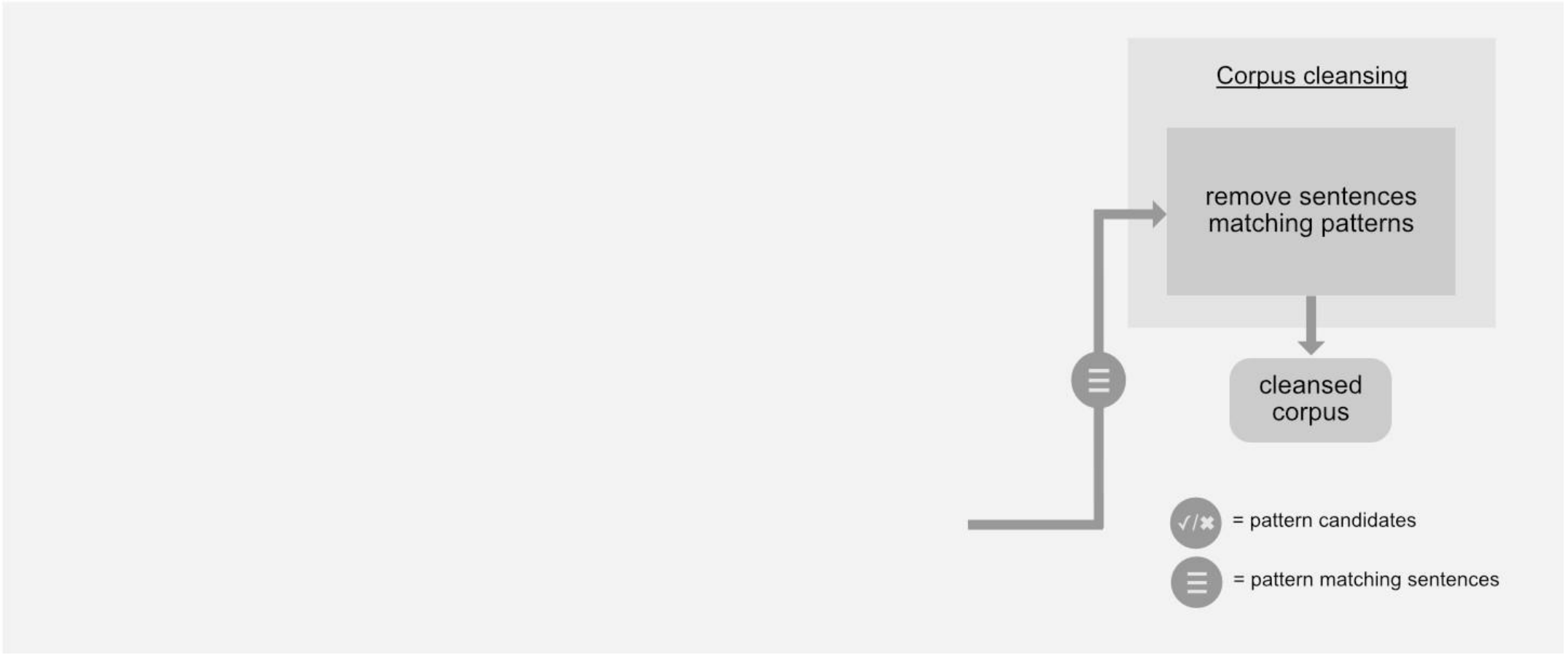
Approach – Concept



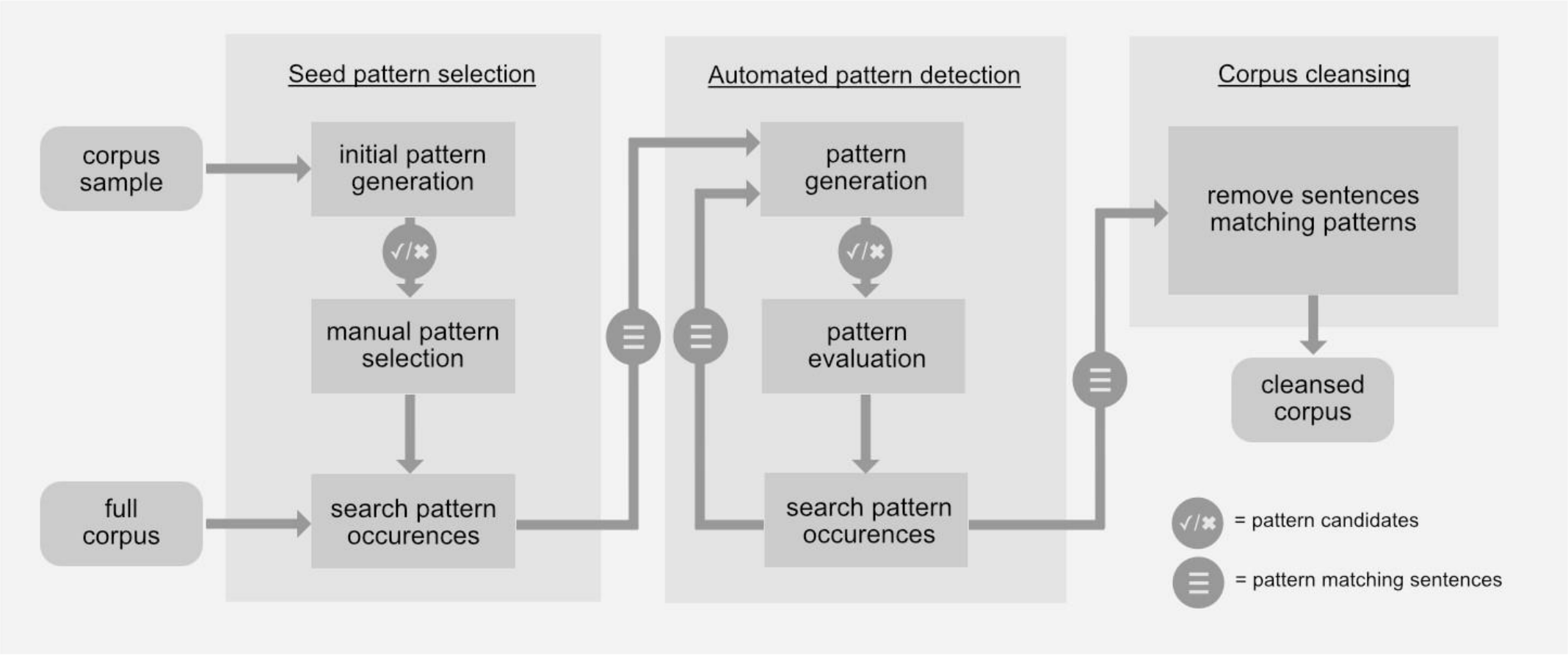
Approach – Concept



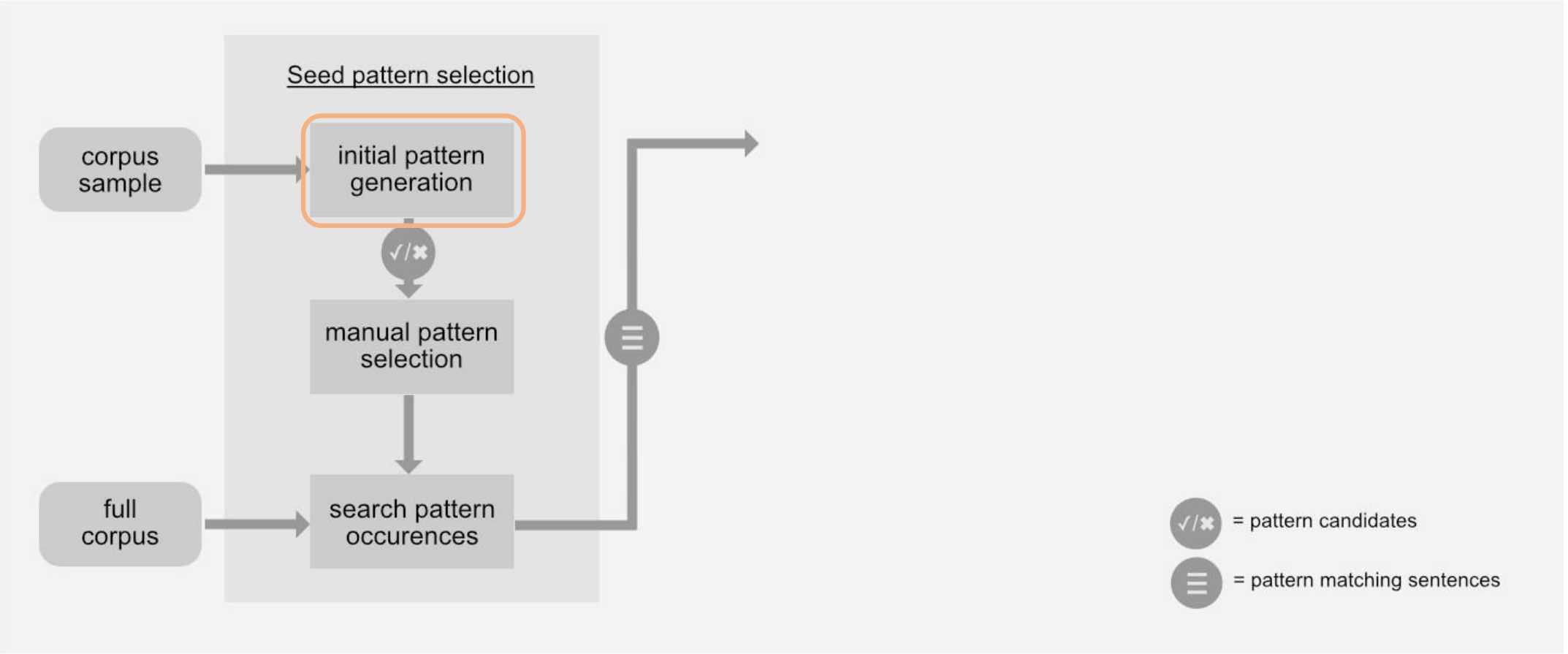
Approach – Concept



Approach – Concept



Approach – Seed pattern selection



Approach – Seed pattern selection

		Patterns of Relevant Sentences		Patterns of Irrelevant Sentences	
Pattern type	**n-gram**	**Most frequent pattern**	**Score**	**Most frequent pattern**	**Score**
Counts w/ Stopwords	1-gram	people	36 287	opponent	29 088
	2-gram	the world	6 210	my opponent	27 149
	3-gram	the fact that	4 206	my opponent s	3 983
	4-gram	in the united states	977	thank my opponent for	1251
	5-gram	has nothing to do with	494	i thank my opponent for	682
Counts w/o Stopwords	1-gram	people	36 287	opponent	29 088
	2-gram	united states	3 906	thank opponent	1 494
	3-gram	big bang theory	251	first round acceptance	617
	4-gram	life liberty pursuit happiness	102	would like thank opponent	359
	5-gram	make law respecting establishment religion	69	round 1 acceptance round 2	82
TF-IDF w/ Stopwords	1-gram	chronicled	1.00	—	—
	2-gram	and weaponry	0.56	actually forfeiting	1.00
	3-gram	an infinite regression	1.00	—	—
	4-gram	abusive education and domestic	1.00	—	—
	5-gram	acceptance of metapyhsical space that	1.00	—	—
TF-IDF w/o Stopwords	1-gram	abortions	1.00	—	—
	2-gram	americans like	1.00	—	—
	3-gram	able kill others	1.00	—	—
	4-gram	accidentally killing equivalent purposely	1.00	—	—
	5-gram	able disprove evolution instead creationists	1.00	—	—

Approach – Seed pattern selection

		Patterns of Relevant Sentences		Patterns of Irrelevant Sentences	
Pattern type	**n-gram**	**Most frequent pattern**	**Score**	**Most frequent pattern**	**Score**
Counts w/ Stopwords	1-gram	people	36 287	opponent	29 088
	2-gram	the world	6 210	my opponent	27 149
	3-gram	the fact that	4 206	my opponent s	3 983
	4-gram	in the united states	977	thank my opponent for	1251
	5-gram	has nothing to do with	494	i thank my opponent for	682
Counts w/o Stopwords	1-gram	people	36 287	opponent	29 088
	2-gram	united states	3 906	thank opponent	1 494
	3-gram	big bang theory	251	first round acceptance	617
	4-gram	life liberty pursuit happiness	102	would like thank opponent	359
	5-gram	make law respecting establishment religion	69	round 1 acceptance round 2	82
TF-IDF w/ Stopwords	1-gram	chronicled	1.00	—	—
	2-gram	and weaponry	0.56	actually forfeiting	1.00
	3-gram	an infinite regression	1.00	—	—
	4-gram	abusive education and domestic	1.00	—	—
	5-gram	acceptance of metapyhsical space that	1.00	—	—
TF-IDF w/o Stopwords	1-gram	abortions	1.00	—	—
	2-gram	americans like	1.00	—	—
	3-gram	able kill others	1.00	—	—
	4-gram	accidentally killing equivalent purposely	1.00	—	—
	5-gram	able disprove evolution instead creationists	1.00	—	—

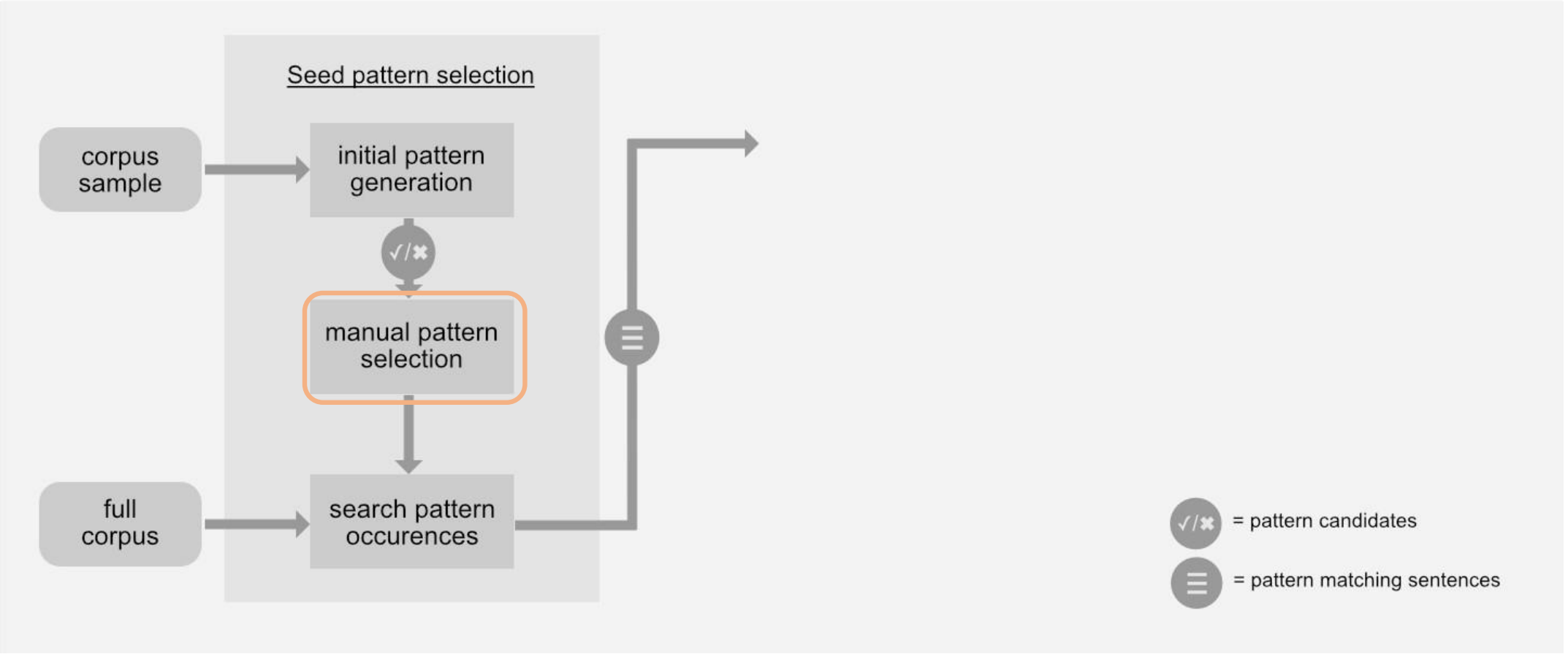
Approach – Seed pattern selection

		Patterns of Relevant Sentences		Patterns of Irrelevant Sentences	
Pattern type	**n-gram**	**Most frequent pattern**	**Score**	**Most frequent pattern**	**Score**
Counts w/ Stopwords	1-gram	people	36 287	opponent	29 088
	2-gram	the world	6 210	my opponent	27 149
	3-gram	the fact that	4 206	my opponent s	3 983
	4-gram	in the united states	977	thank my opponent for	1251
	5-gram	has nothing to do with	494	i thank my opponent for	682
Counts w/o Stopwords	1-gram	people	36 287	opponent	29 088
	2-gram	united states	3 906	thank opponent	1 494
	3-gram	big bang theory	251	first round acceptance	617
	4-gram	life liberty pursuit happiness	102	would like thank opponent	359
	5-gram	make law respecting establishment religion	69	round 1 acceptance round 2	82
TF-IDF w/ Stopwords	1-gram	chronicled	1.00	—	—
	2-gram	and weaponry	0.56	actually forfeiting	1.00
	3-gram	an infinite regression	1.00	—	—
	4-gram	abusive education and domestic	1.00	—	—
	5-gram	acceptance of metapyhsical space that	1.00	—	—
TF-IDF w/o Stopwords	1-gram	abortions	1.00	—	—
	2-gram	americans like	1.00	—	—
	3-gram	able kill others	1.00	—	—
	4-gram	accidentally killing equivalent purposely	1.00	—	—
	5-gram	able disprove evolution instead creationists	1.00	—	—

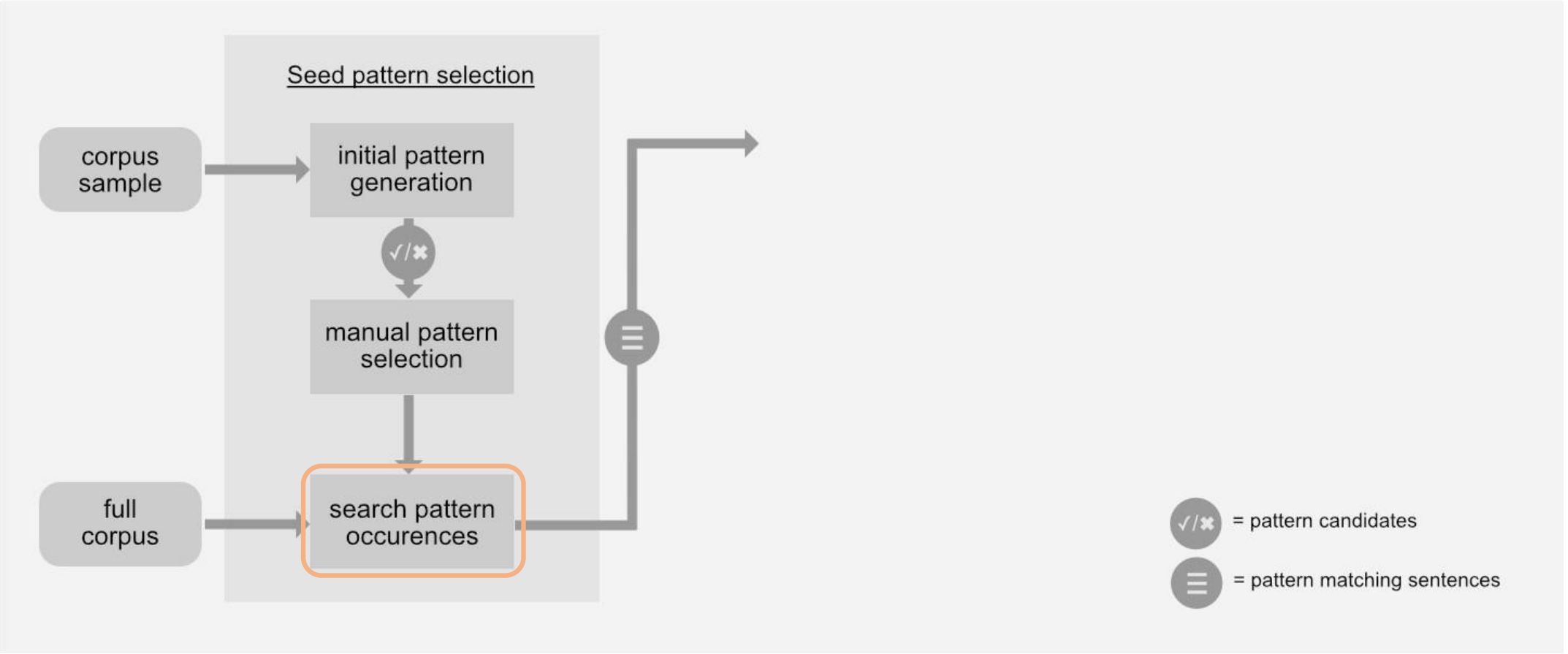
Approach – Seed pattern selection

		Patterns of Relevant Sentences		Patterns of Irrelevant Sentences	
Pattern type	**n-gram**	**Most frequent pattern**	**Score**	**Most frequent pattern**	**Score**
Counts w/ Stopwords	1-gram	people	36 287	opponent	29 088
	2-gram	the world	6 210	my opponent	27 149
	3-gram	the fact that	4 206	my opponent s	3 983
	4-gram	in the united states	977	thank my opponent for	1251
	5-gram	has nothing to do with	494	i thank my opponent for	682
Counts w/o Stopwords	1-gram	people	36 287	opponent	29 088
	2-gram	united states	3 906	thank opponent	1 494
	3-gram	big bang theory	251	first round acceptance	617
	4-gram	life liberty pursuit happiness	102	would like thank opponent	359
	5-gram	make law respecting establishment religion	69	round 1 acceptance round 2	82
TF-IDF w/ Stopwords	1-gram	chronicled	1.00	—	—
	2-gram	and weaponry	0.56	actually forfeiting	1.00
	3-gram	an infinite regression	1.00	—	—
	4-gram	abusive education and domestic	1.00	—	—
	5-gram	acceptance of metapyhsical space that	1.00	—	—
TF-IDF w/o Stopwords	1-gram	abortions	1.00	—	—
	2-gram	americans like	1.00	—	—
	3-gram	able kill others	1.00	—	—
	4-gram	accidentally killing equivalent purposely	1.00	—	—
	5-gram	able disprove evolution instead creationists	1.00	—	—

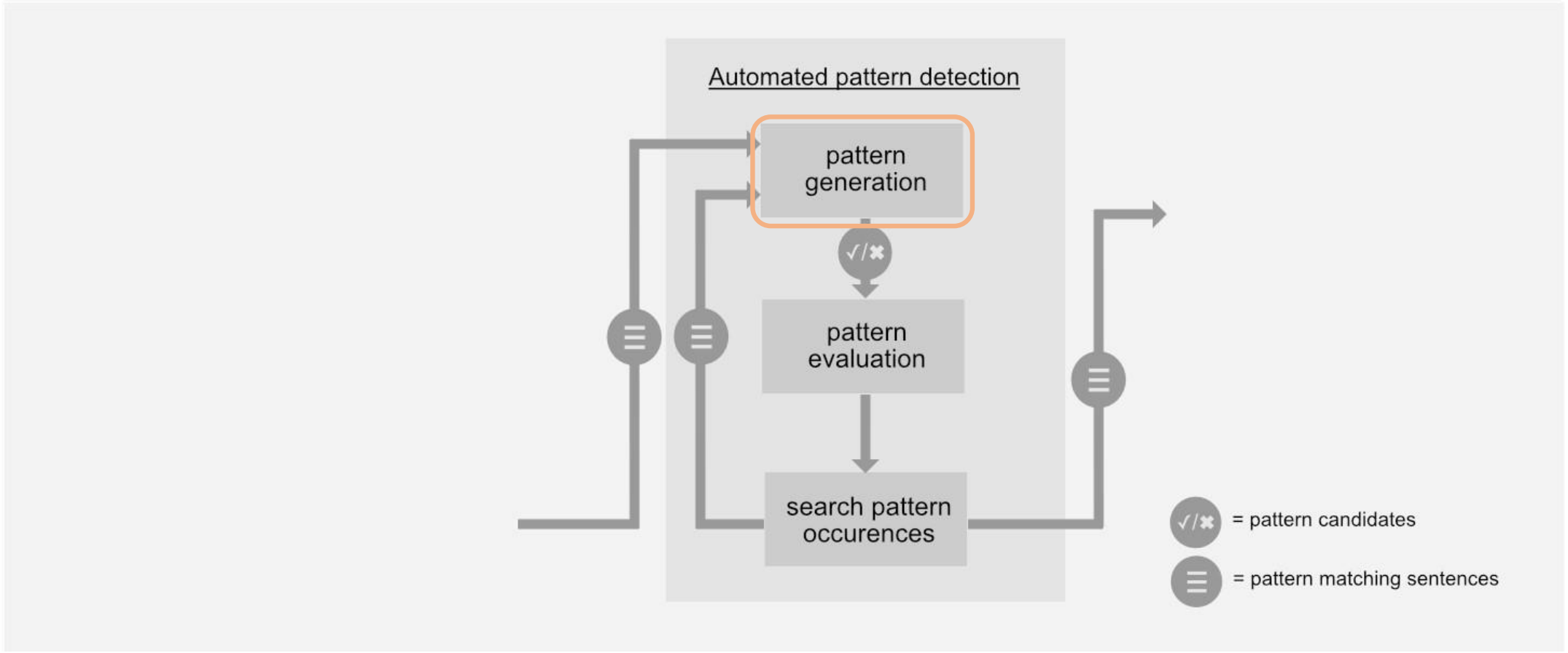
Approach – Seed pattern selection



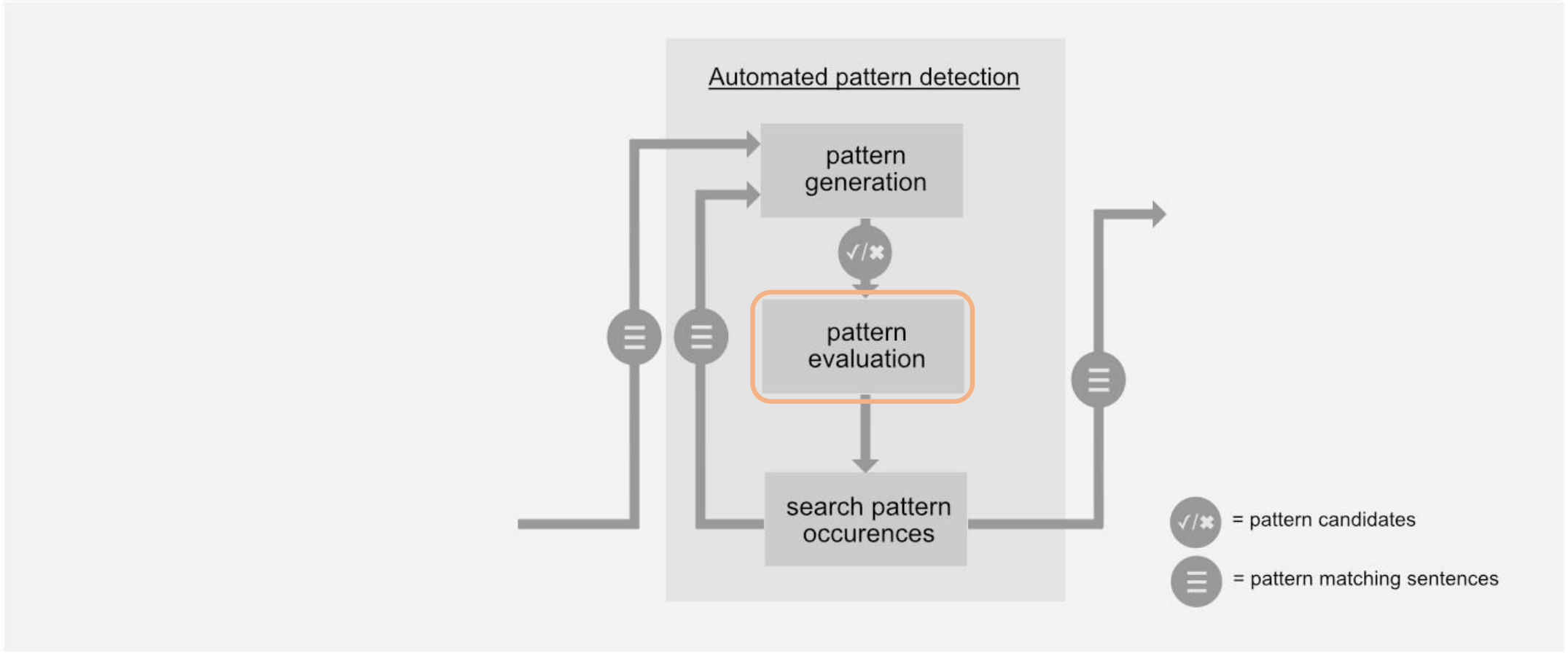
Approach – Seed pattern selection



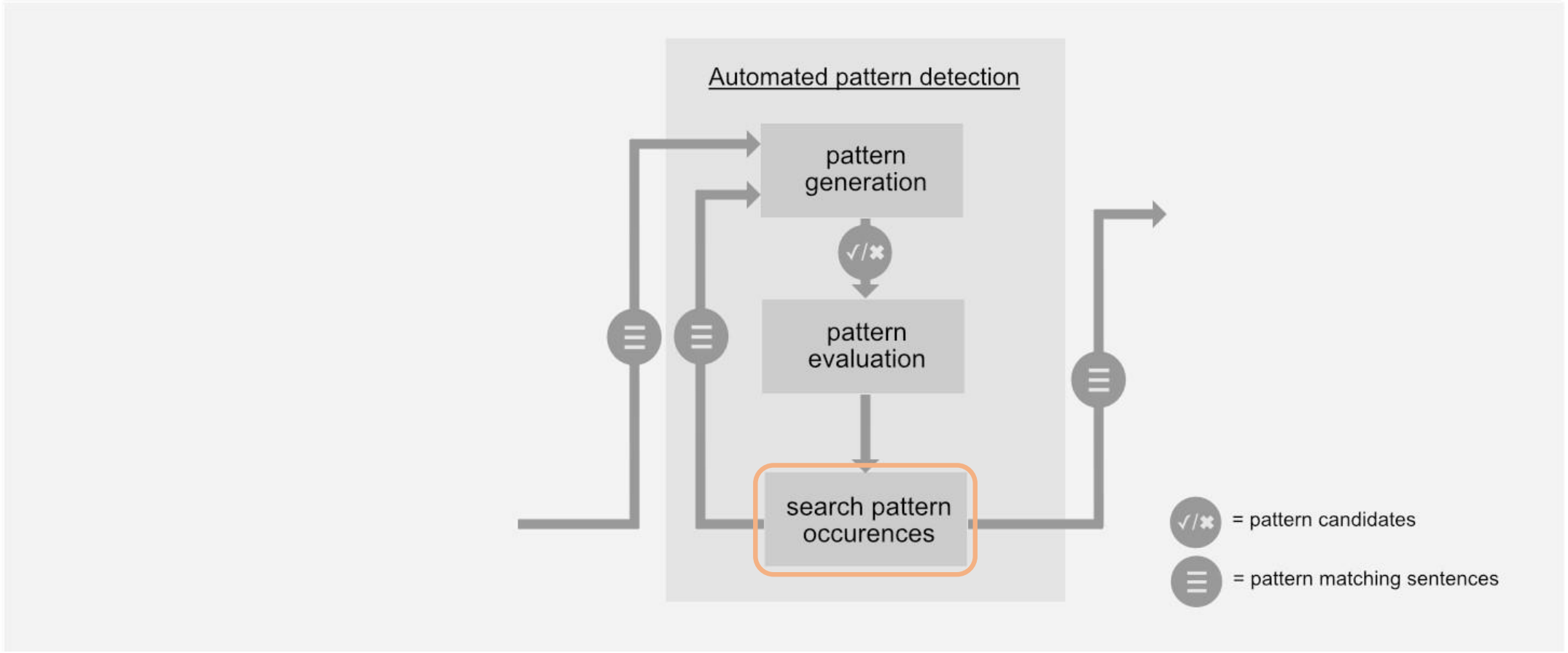
Approach – Automated pattern detection



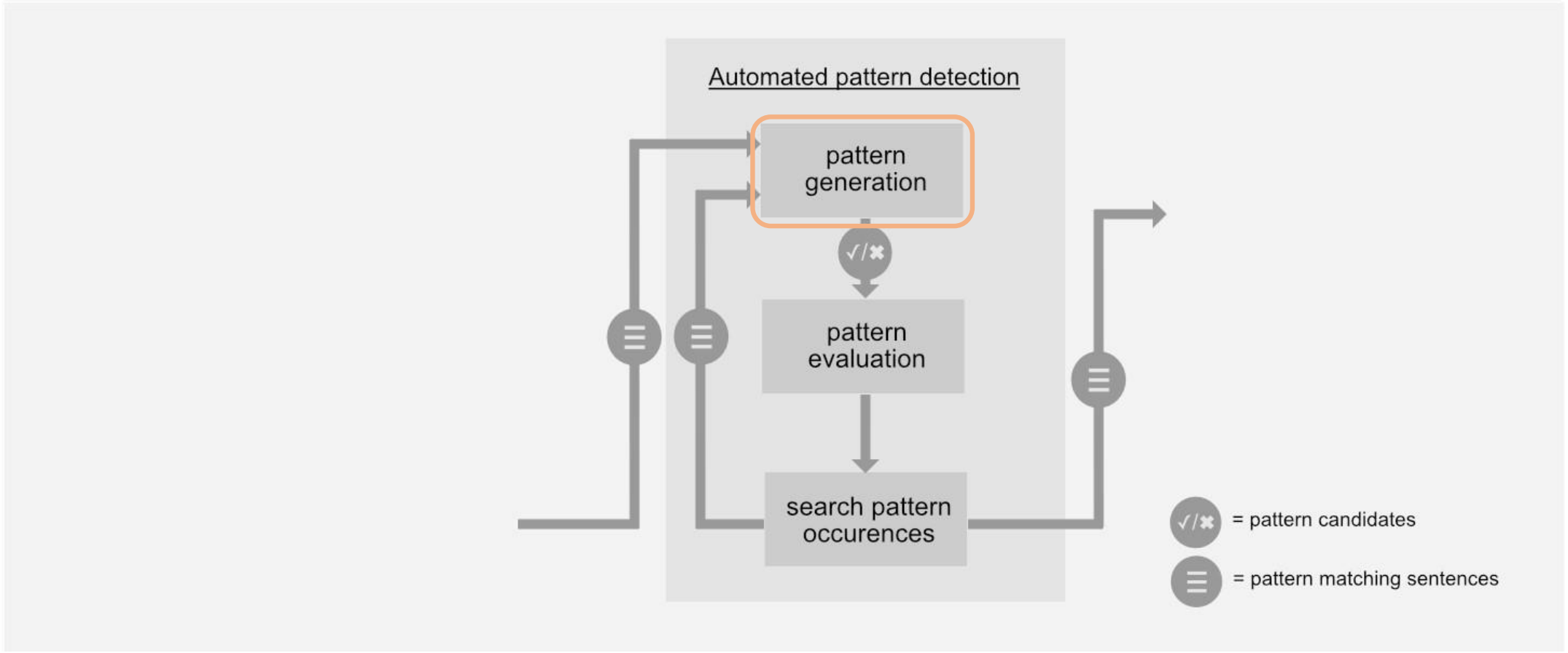
Approach – Automated pattern detection



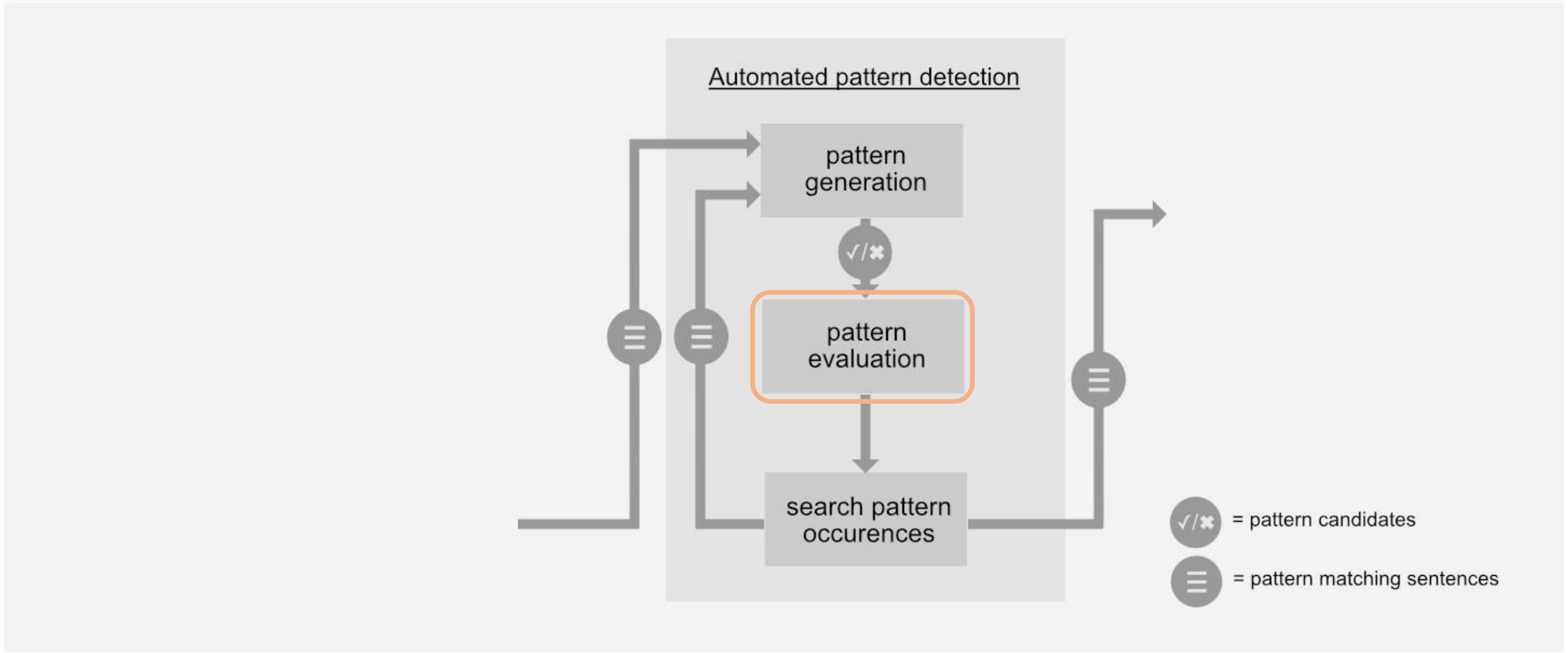
Approach – Automated pattern detection



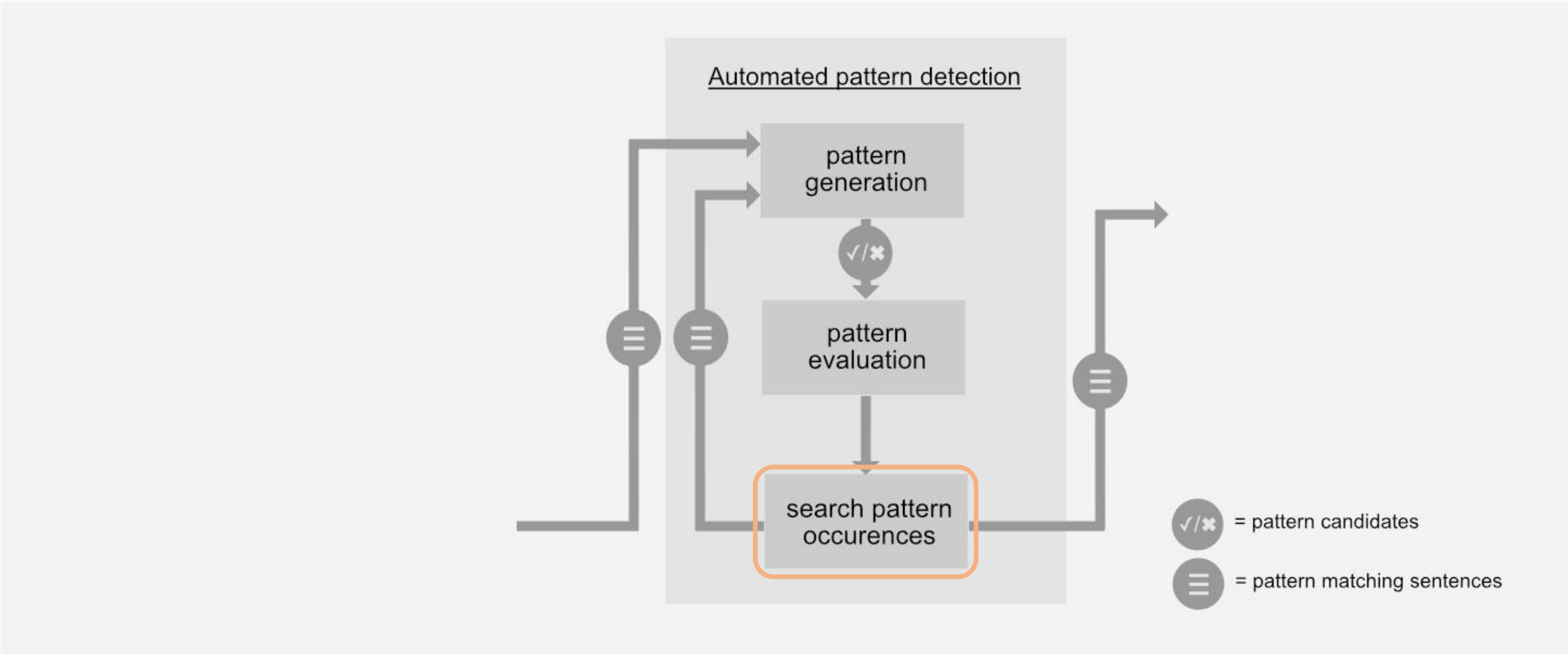
Approach – Automated pattern detection



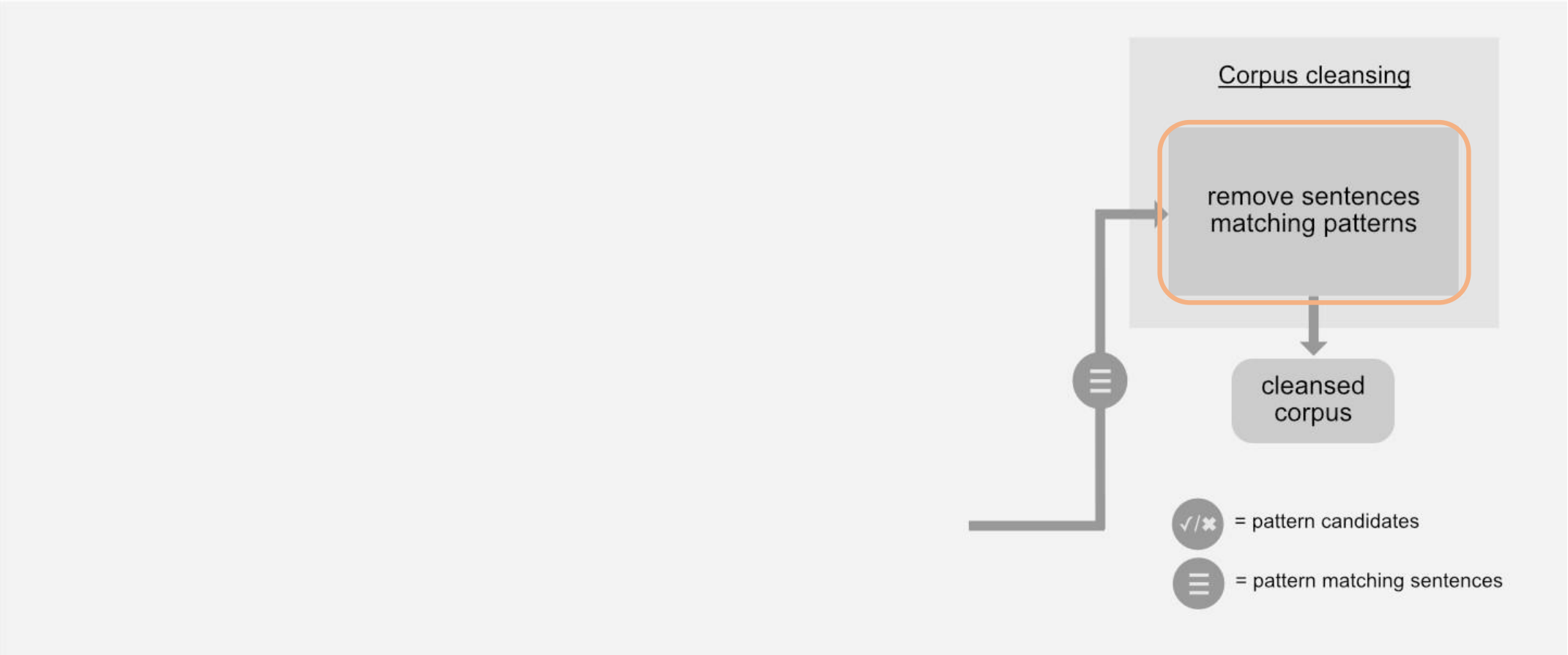
Approach – Automated pattern detection



Approach – Automated pattern detection



Approach – Corpus cleansing



Approach – Corpus cleansing

Apply special metrics?

Approach – Corpus cleansing

Apply special metrics?

Only remove sentences clearly irrelevant.

Approach – Corpus cleansing

Apply special metrics?

Only remove sentences clearly irrelevant.

Assumption: Most irrelevant sentences are located around the beginning and end of an argument.

Evaluation – Automated process

	Positive Patterns			
Iteration	**Patterns**	**Sentences**	**Auto Prec.**	**Manual Prec.**
Seed	38	600469	0.997	n/a
1	10	7602	0.987	n/a
2	0	-57	n/a	n/a
3	...	-15	...	n/a
4	...	-10	...	n/a
5	...	-6	...	n/a
Total	48	607983	0.984	n/a

	Negative Patterns			
Iteration	**Patterns**	**Sentences**	**Auto Prec.**	**Manual Prec.**
Seed	17	41619	0.972	1.0
1	74	5849	0.983	1.0
2	19	3606	0.982	1.0
3	4	956	0.973	0.96
4	3	594	0.980	0.97
5	5	225	0.983	0.88
Total	122	52849	0.980	0.97

Evaluation – Automated process

	Positive Patterns			
Iteration	**Patterns**	**Sentences**	**Auto Prec.**	**Manual Prec.**
Seed	38	600469	0.997	n/a
1	10	7602	0.987	n/a
2	0	-57	n/a	n/a
3	...	-15	...	n/a
4	...	-10	...	n/a
5	...	-6	...	n/a
Total	48	607983	0.984	n/a

	Negative Patterns			
Iteration	**Patterns**	**Sentences**	**Auto Prec.**	**Manual Prec.**
Seed	17	41619	0.972	1.0
1	74	5849	0.983	1.0
2	19	3606	0.982	1.0
3	4	956	0.973	0.96
4	3	594	0.980	0.97
5	5	225	0.983	0.88
Total	122	52849	0.980	0.97

Evaluation – Automated process

	Positive Patterns			
Iteration	**Patterns**	**Sentences**	**Auto Prec.**	**Manual Prec.**
Seed	38	600469	0.997	n/a
1	10	7602	0.987	n/a
2	0	-57	n/a	n/a
3	...	-15	...	n/a
4	...	-10	...	n/a
5	...	-6	...	n/a
Total	48	607983	0.984	n/a

	Negative Patterns			
Iteration	**Patterns**	**Sentences**	**Auto Prec.**	**Manual Prec.**
Seed	17	41619	0.972	1.0
1	74	5849	0.983	1.0
2	19	3606	0.982	1.0
3	4	956	0.973	0.96
4	3	594	0.980	0.97
5	5	225	0.983	0.88
Total	122	52849	0.980	0.97

Evaluation – Automated process

	Positive Patterns			
Iteration	**Patterns**	**Sentences**	**Auto Prec.**	**Manual Prec.**
Seed	38	600469	0.997	n/a
1	10	7602	0.987	n/a
2	0	-57	n/a	n/a
3	...	-15	...	n/a
4	...	-10	...	n/a
5	...	-6	...	n/a
Total	48	607983	0.984	n/a

	Negative Patterns			
Iteration	**Patterns**	**Sentences**	**Auto Prec.**	**Manual Prec.**
Seed	17	41619	0.972	1.0
1	74	5849	0.983	1.0
2	19	3606	0.982	1.0
3	4	956	0.973	0.96
4	3	594	0.980	0.97
5	5	225	0.983	0.88
Total	122	52849	0.980	0.97

Evaluation – Automated process

	Positive Patterns			
Iteration	**Patterns**	**Sentences**	**Auto Prec.**	**Manual Prec.**
Seed	38	600469	0.997	n/a
1	10	7602	0.987	n/a
2	0	-57	n/a	n/a
3	...	-15	...	n/a
4	...	-10	...	n/a
5	...	-6	...	n/a
Total	48	607983	0.984	n/a

	Negative Patterns			
Iteration	**Patterns**	**Sentences**	**Auto Prec.**	**Manual Prec.**
Seed	17	41619	0.972	1.0
1	74	5849	0.983	1.0
2	19	3606	0.982	1.0
3	4	956	0.973	0.96
4	3	594	0.980	0.97
5	5	225	0.983	0.88
Total	122	52849	0.980	0.97

Evaluation – Manual Annotation

- Three experts
- 100 Sentences per round

  +100 Sentences from seed

	Positive Patterns			
Iteration	**Patterns**	**Sentences**	**Auto Prec.**	**Manual Prec.**
Seed	38	600469	0.997	n/a
1	10	7602	0.987	n/a
2	0	-57	n/a	n/a
3	...	-15	...	n/a
4	...	-10	...	n/a
5	...	-6	...	n/a
Total	48	607983	0.984	n/a

	Negative Patterns			
Iteration	**Patterns**	**Sentences**	**Auto Prec.**	**Manual Prec.**
Seed	17	41619	0.972	1.0
1	74	5849	0.983	1.0
2	19	3606	0.982	1.0
3	4	956	0.973	0.96
4	3	594	0.980	0.97
5	5	225	0.983	0.88
Total	122	52849	0.980	0.97

Evaluation – Manual Annotation

	Annotator Precision					
	Individual			Combined		
Iteration	**1**	**2**	**3**	**Full**	**Majority**	**At least one**
Seed	1.0	0.99	1.0	0.99	1.0	1.0
1	0.98	0.98	1.0	0.96	1.0	1.0
2	1.0	0.95	0.99	0.94	1.0	1.0
3	0.96	0.96	0.95	0.93	0.96	0.98
4	0.95	0.95	0.97	0.93	0.97	0.97
5	0.84	0.85	0.93	0.79	0.88	0.95
Total	0.955	0.946	0.973	0.92	0.97	0.98

Evaluation – Manual Annotation

	Annotator Precision					
	Individual			Combined		
Iteration	**1**	**2**	**3**	**Full**	**Majority**	**At least one**
Seed	1.0	0.99	1.0	0.99	1.0	1.0
1	0.98	0.98	1.0	0.96	1.0	1.0
2	1.0	0.95	0.99	0.94	1.0	1.0
3	0.96	0.96	0.95	0.93	0.96	0.98
4	0.95	0.95	0.97	0.93	0.97	0.97
5	0.84	0.85	0.93	0.79	0.88	0.95
Total	0.955	0.946	0.973	0.92	0.97	0.98

Evaluation – Manual Annotation

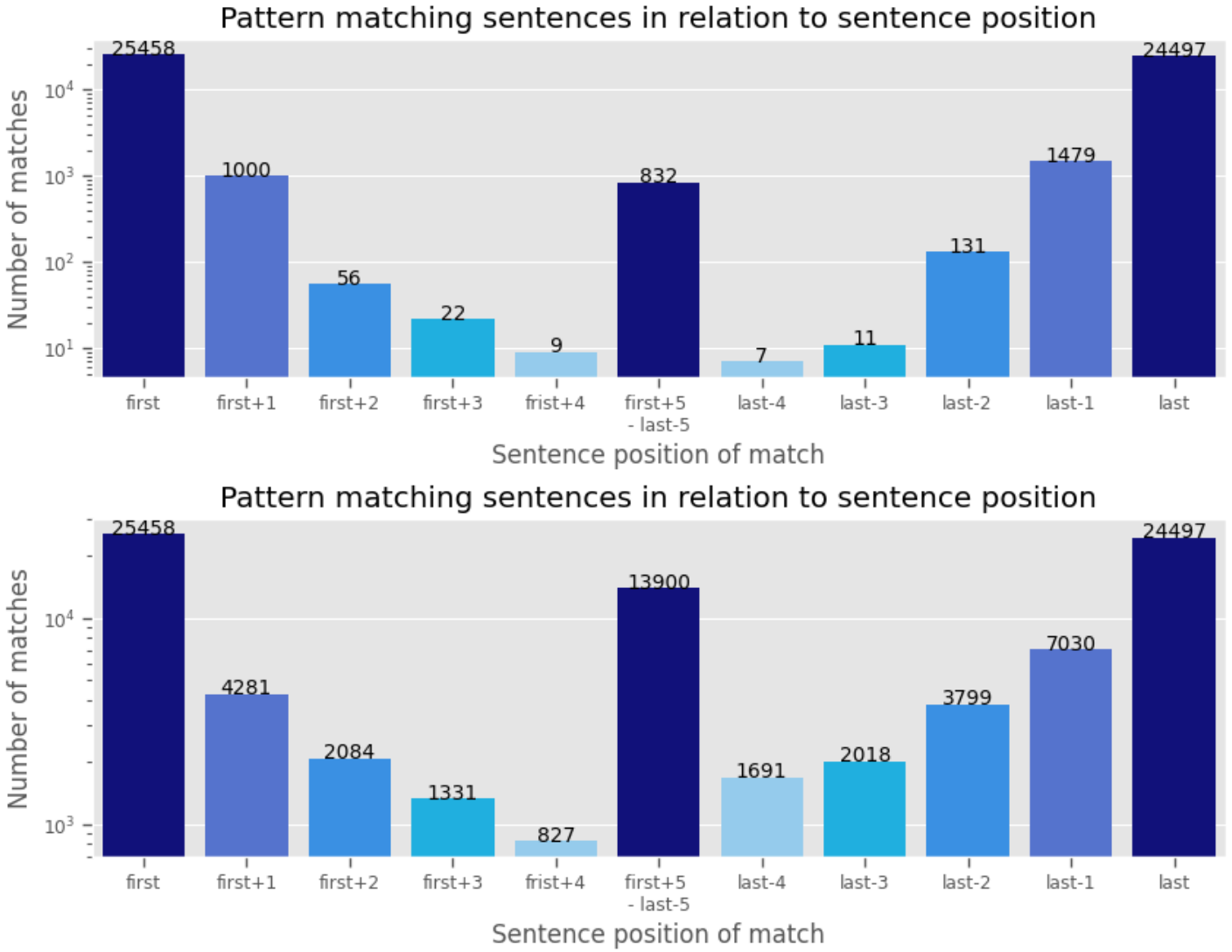
	Annotator Precision					
	Individual			Combined		
Iteration	**1**	**2**	**3**	**Full**	**Majority**	**At least one**
Seed	1.0	0.99	1.0	0.99	1.0	1.0
1	0.98	0.98	1.0	0.96	1.0	1.0
2	1.0	0.95	0.99	0.94	1.0	1.0
3	0.96	0.96	0.95	0.93	0.96	0.98
4	0.95	0.95	0.97	0.93	0.97	0.97
5	0.84	0.85	0.93	0.79	0.88	0.95
Total	0.955	0.946	0.973	0.92	0.97	0.98

Just remove every matched sentence?

Evaluation – Sentence Position

Assumption: Most irrelevant sentences are located around the beginning and end of an argument

Evaluation – Sentence Position



Conclusion – Overall performance

Overall Performance:

- 53 502 sentences from 48 089 arguments removed
  -> more then 10% of arguments filtered (from complete corpus)

Conclusion – Overall performance

Overall Performance:

- 53 502 sentences from 48 089 arguments removed
  -> more then 10% of arguments filtered (from complete corpus)
- Average precision of 97% according to manual evaluation

Conclusion – Overall performance

Overall Performance:

- 53 502 sentences from 48 089 arguments removed
  -> more then 10% of arguments filtered (from complete corpus)
- Average precision of 97% according to manual evaluation

However: Some improvements and further analysis required

Conclusion – Future Improvements

Improvements:

- Improve tokenization

Conclusion – Future Improvements

Improvements:

- Improve tokenization
- Experiment with different pattern extraction methods

Conclusion – Future Improvements

Improvements:

- Improve tokenization
- Experiment with different pattern extraction methods
- Evaluate sentence position relevance

Thank you for your attention!

Any questions?

References

- Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast,Matthias Hagen, and Benno Stein. Data acquisition for argument search:The args.me corpus. InProceedings of the 42nd Edition of the GermanConference on Artificial Intelligence, page 48â59, 2019. 2.3, 3, 4