



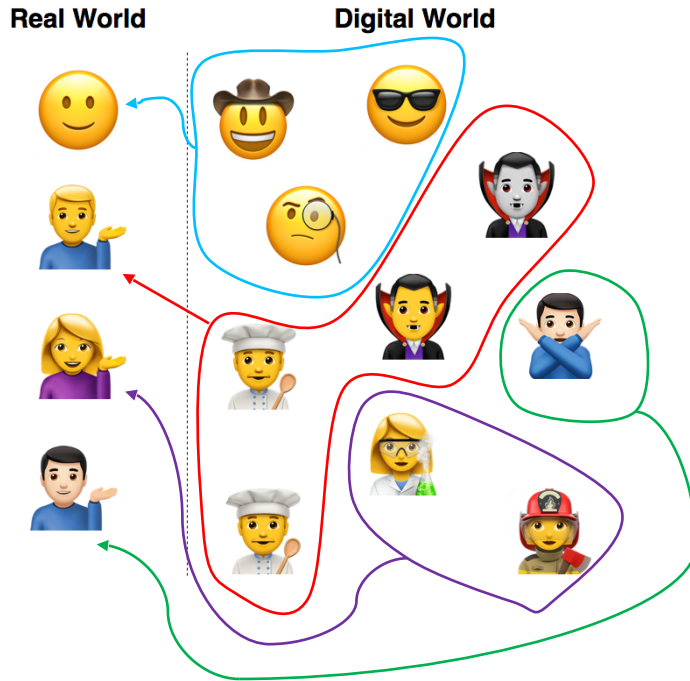
Universität Stuttgart

Towards task-based parallelization for entity resolution

Leonardo Gazzarri, Melanie Herschel

IPVS/DE

Entity Resolution



Entity Resolution (ER) is the task of finding which virtual profiles of entities refer to the same real world entity.

Different profiles of the same entity

Example of «schema-free» data

Example

$e_1 = \{ \textit{Title}: \text{Inception} ; \textit{Starring}: \text{DiCaprio}; \textit{Country}: \text{USA}; \textit{Language}: \text{English} \}$

$e_2 = \{ \text{Titanic, starring L. DiCaprio and K. Winslet, directed by Cameron, U.S. (English)} \}$

$e_3 = \{ \textit{Name}: \text{Inception (USA, 2010)}, \textit{Actors}: \text{DiCaprio}, \textit{Director}: \text{Nolan} \}$

$e_4 = \{ \textit{Name}: \text{The Terminator}; \textit{Directed by}: \text{Cameron}; \textit{Country}: \text{U.S.A}, \textit{Starring}: \text{Schwarzenegger} \}$

$e_5 = \{ \text{The Lord of the Rings – The Fellowship of the Ring, USA 2001} \}$

Different profiles of the same entity

Example of «schema-free» data

Example

$e_1 = \{ \text{Title: Inception ; Starring: DiCaprio; Country: USA; Language: English} \}$

$e_2 = \{ \text{Titanic, starring L. DiCaprio and K. Winslet, directed by Cameron, U.S. (English)} \}$

$e_3 = \{ \text{Name: Inception (USA, 2010), Actors: DiCaprio, Director: Nolan} \}$

$e_4 = \{ \text{Name: The Terminator; Directed by: Cameron; Country: U.S.A, Starring: Schwarzenegger} \}$

$e_5 = \{ \text{The Lord of the Rings – The Fellowship of the Ring, USA 2001} \}$

Different attribute names.

Free text

No fixed schema

No common standards for representation

Different profiles of the same entity

Expected result:

1

$e_1 = \{ \textit{Title: Inception ; Starring: DiCaprio; Country: USA; Language: English} \}$

$e_3 = \{ \textit{Name: Inception (USA, 2010), Actors: DiCaprio, Director: Nolan} \}$

2

$e_2 = \{ \textit{Titanic, starring L. DiCaprio and K. Winslet, directed by Cameron, U.S. (English)} \}$

3

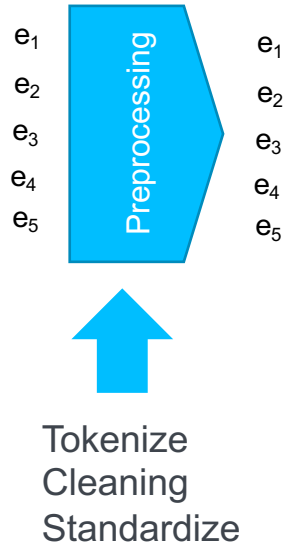
$e_4 = \{ \textit{Name: The Terminator; Directed by: Cameron; Country: USA, Starring: Schwarzenegger} \}$

4

$e_5 = \{ \textit{The Lord of the Rings – The Fellowship of the Ring, USA 2001} \}$

ER workflow [1,2]

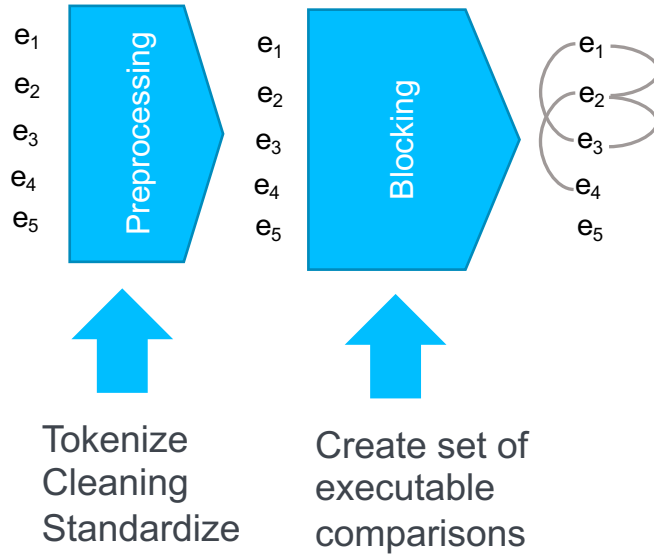
$e_1 = \{ \textit{Title: Inception ; Starring: DiCaprio; Country: USA; Language: English} \}$
 $e_2 = \{ \textit{Titanic, starring L. DiCaprio and K. Winslet, directed by Cameron, U.S. (English)} \}$
 $e_3 = \{ \textit{Name: Inception (USA, 2010), Actors: DiCaprio, Director: Nolan} \}$
 $e_4 = \{ \textit{Name: Terminator; Directed by: Cameron; Country: USA, Starring: Schwarzenegger} \}$
 $e_5 = \{ \textit{The Lord of the Rings – The Fellowship of the Ring, USA 2001} \}$



e.g. remove attributes, perform standardizations (U.S. and U.S.A. to USA)
 $e_2 = \{ \textit{Titanic, DiCaprio, Winslet, Cameron, USA, English} \}$

ER workflow [1,2]

$e_1 = \{ \text{Title: Inception ; Starring: DiCaprio; Country: USA; Language: English} \}$
 $e_2 = \{ \text{Titanic, starring L. DiCaprio and K. Winslet, directed by Cameron, U.S. (English)} \}$
 $e_3 = \{ \text{Name: Inception (USA, 2010), Actors: DiCaprio, Director: Nolan} \}$
 $e_4 = \{ \text{Name: Terminator; Directed by: Cameron; Country: USA, Starring: Schwarzenegger} \}$
 $e_5 = \{ \text{The Lord of the Rings – The Fellowship of the Ring, USA 2001} \}$

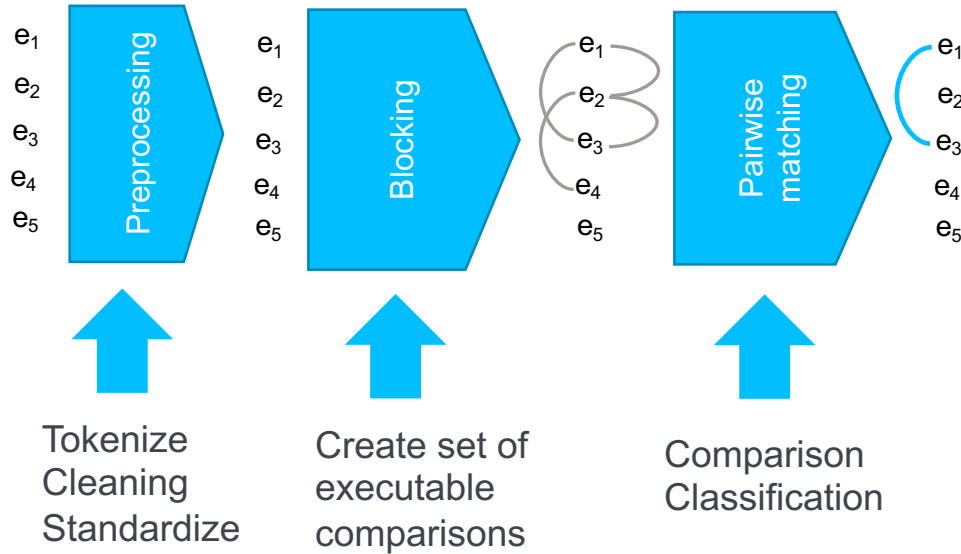


efficiency related goal: limit the quadratic number of total comparisons

4 vs 10 comparisons

ER workflow [1,2]

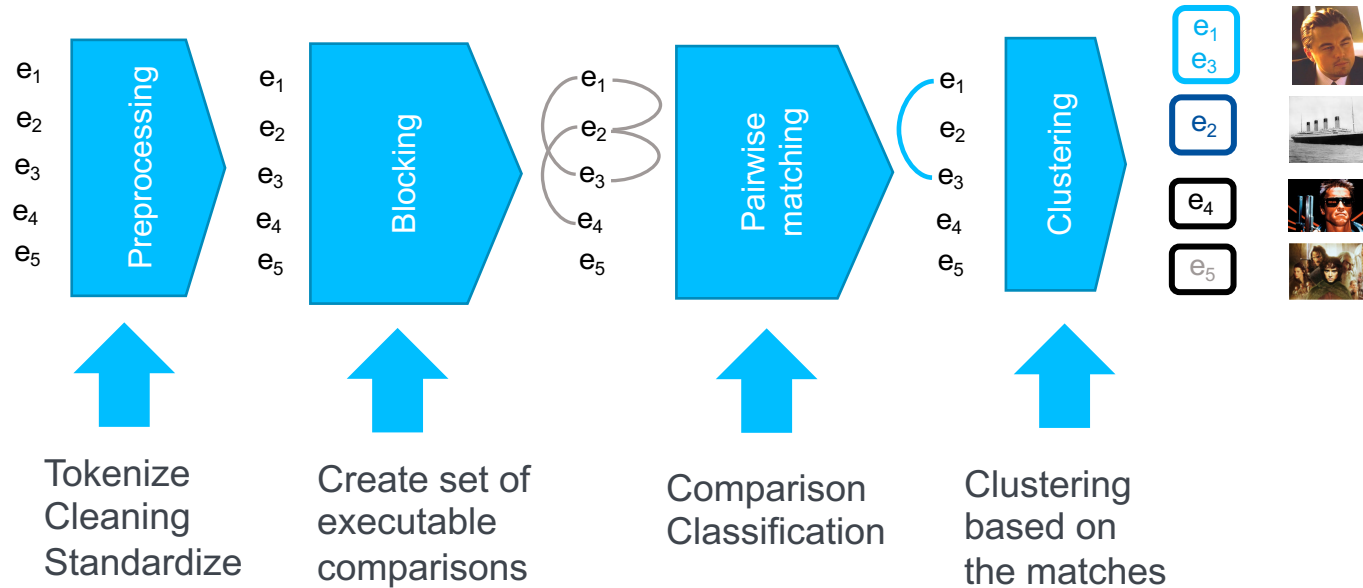
$e_1 = \{ \textit{Title: Inception} ; \textit{Starring: DiCaprio} ; \textit{Country: USA} ; \textit{Language: English} \}$
 $e_2 = \{ \textit{Titanic, starring L. DiCaprio and K. Winslet, directed by Cameron, U.S. (English)} \}$
 $e_3 = \{ \textit{Name: Inception (USA, 2010), Actors: DiCaprio, Director: Nolan} \}$
 $e_4 = \{ \textit{Name: Terminator; Directed by: Cameron; Country: USA, Starring: Schwarzenegger} \}$
 $e_5 = \{ \textit{The Lord of the Rings – The Fellowship of the Ring, USA 2001} \}$



e.g. Jaccard similarity between tokenized representations
and classification based on a similarity score threshold

ER workflow [1,2]

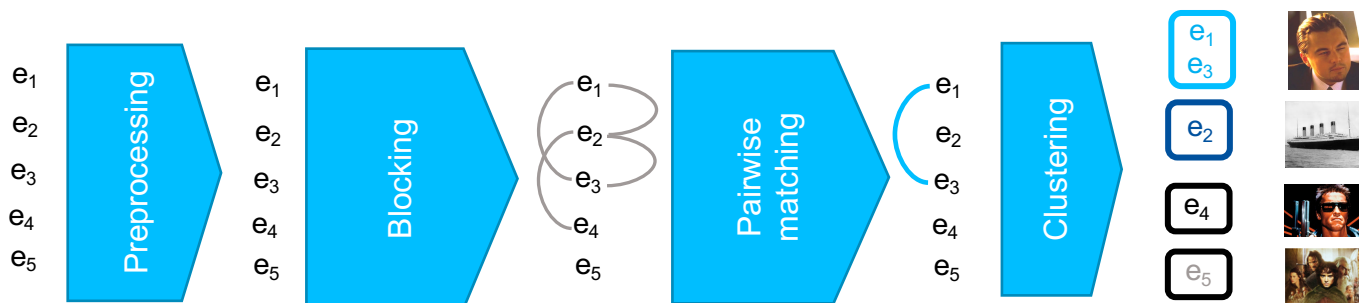
$e_1 = \{ \text{Title: Inception ; Starring: DiCaprio; Country: USA; Language: English} \}$
 $e_2 = \{ \text{Titanic, starring L. DiCaprio and K. Winslet, directed by Cameron, U.S. (English)} \}$
 $e_3 = \{ \text{Name: Inception (USA, 2010), Actors: DiCaprio, Director: Nolan} \}$
 $e_4 = \{ \text{Name: Terminator; Directed by: Cameron; Country: USA, Starring: Schwarzenegger} \}$
 $e_5 = \{ \text{The Lord of the Rings – The Fellowship of the Ring, USA 2001} \}$



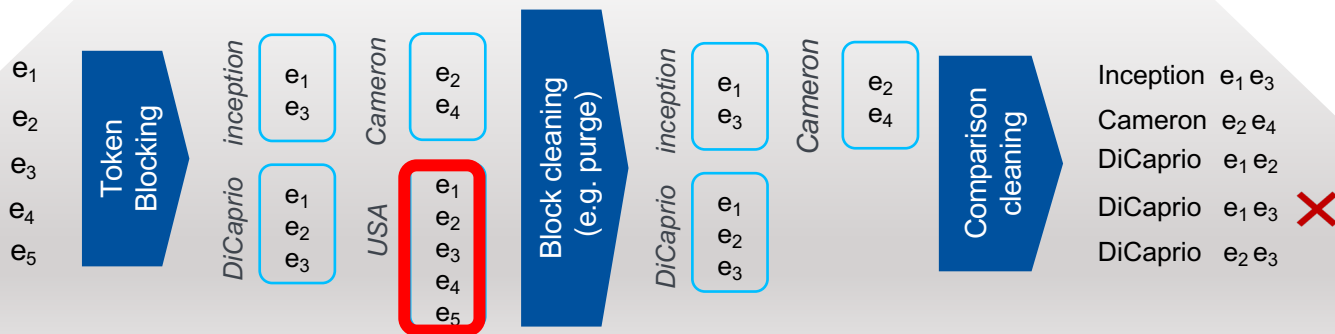
e.g. connected components

ER workflow [1,2]

$e_1 = \{ \text{Title: Inception ; Starring: DiCaprio; Country: USA; Language: English} \}$
 $e_2 = \{ \text{Titanic, starring L. DiCaprio and K. Winslet, directed by Cameron, U.S. (English)} \}$
 $e_3 = \{ \text{Name: Inception (USA, 2010), Actors: DiCaprio, Director: Nolan} \}$
 $e_4 = \{ \text{Name: Terminator; Directed by: Cameron; Country: USA, Starring: Schwarzenegger} \}$
 $e_5 = \{ \text{The Lord of the Rings – The Fellowship of the Ring, USA 2001} \}$



zoom in on blocking



Too big!

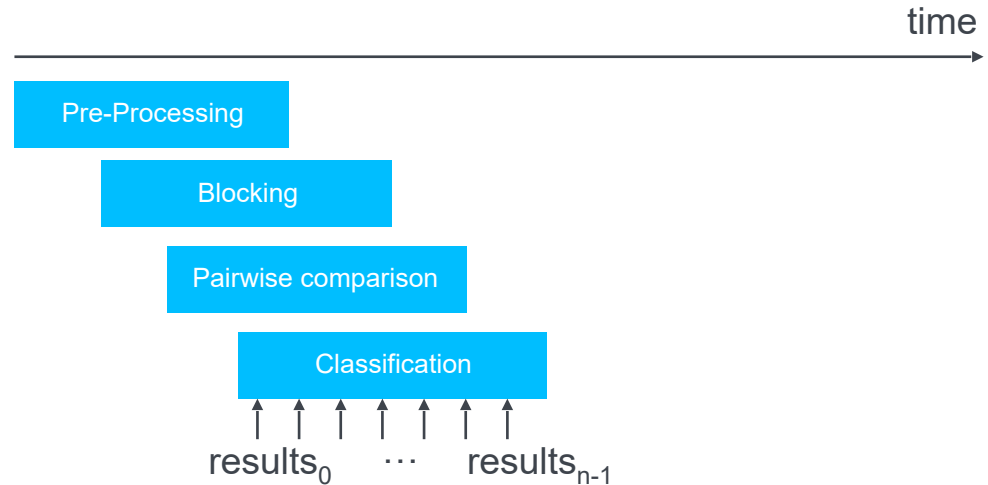
Motivation

Support inter-step parallelism in ER [3,4]

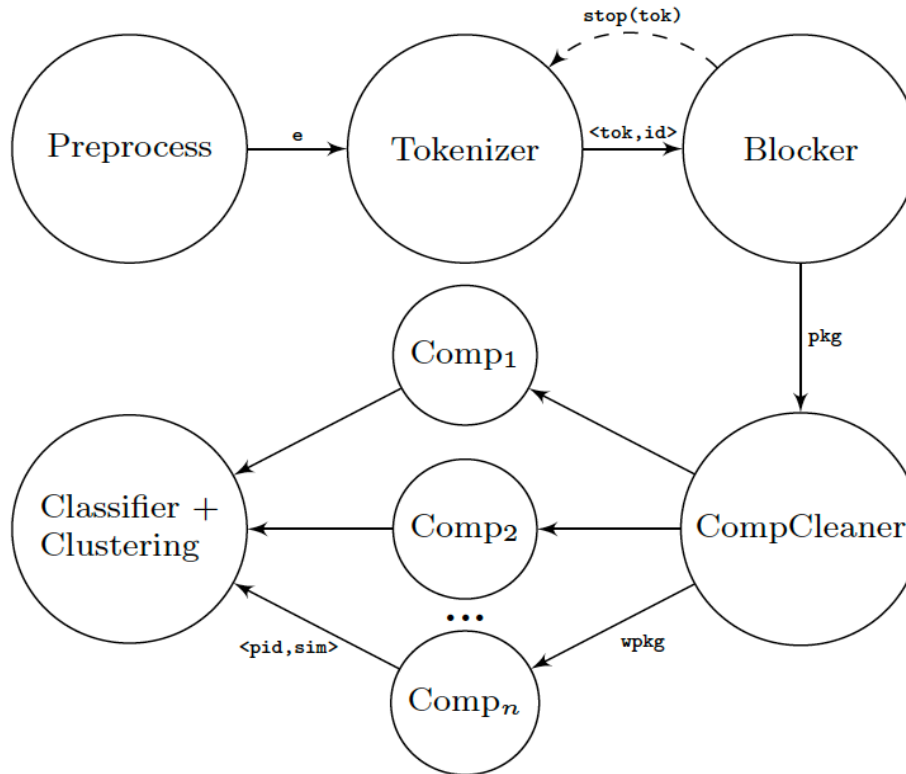
(i) Sequential



(ii) Inter-step

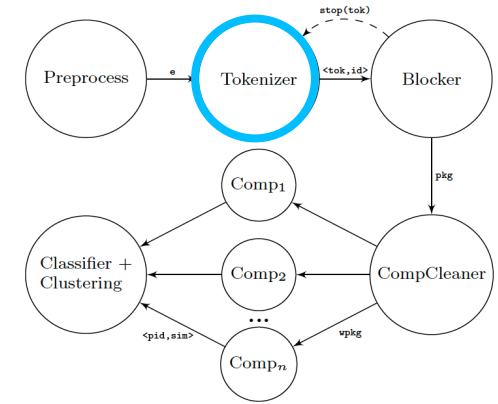
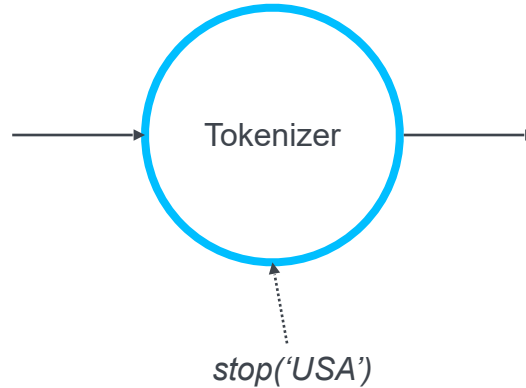


Parallel activities



Preprocess and Tokenizer stages

$e_4 = \{ \textit{Name: The Terminator; Directed by: Cameron; Country: U.S.A., Starring: Schwarzenegger} \}$



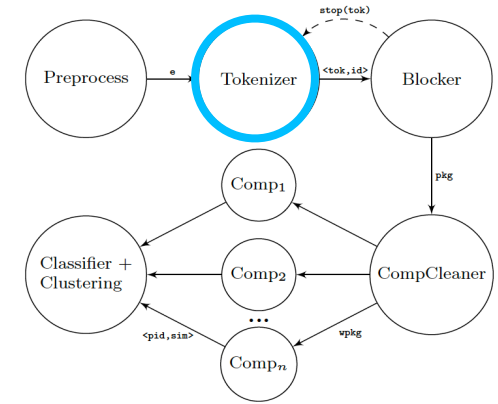
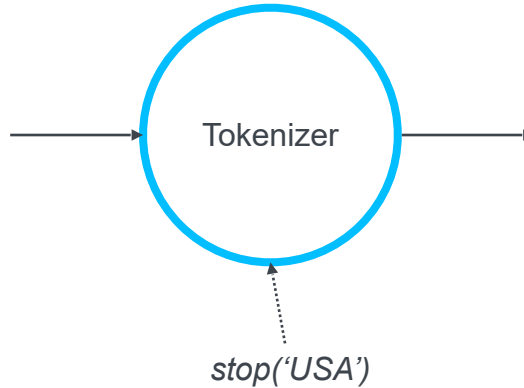
Preprocess and Tokenizer stages

$e_4 = \{ \textit{Name: The Terminator; Directed by: Cameron; Country: U.S.A., Starring: Schwarzenegger} \}$



preprocess + tokenize

$e_4 = \{ \textit{Terminator, Cameron, USA, Schwarzenegger} \}$



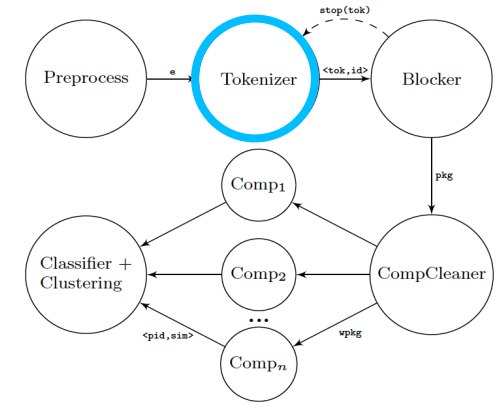
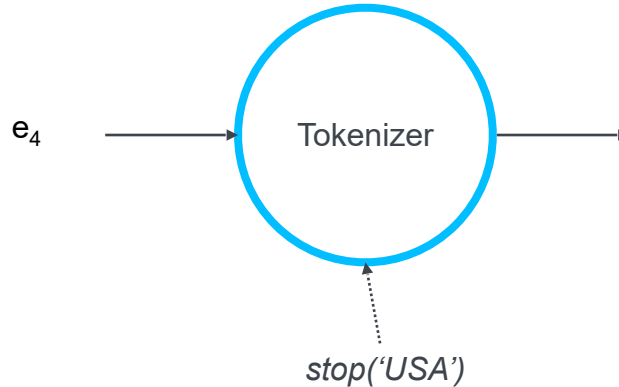
Preprocess and Tokenizer stages

$e_4 = \{ \textit{Name: The Terminator; Directed by: Cameron; Country: U.S.A., Starring: Schwarzenegger} \}$



preprocess + tokenize

$e_4 = \{ \textit{Terminator, Cameron, USA, Schwarzenegger} \}$



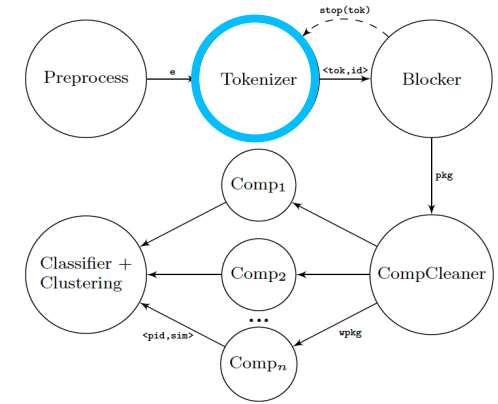
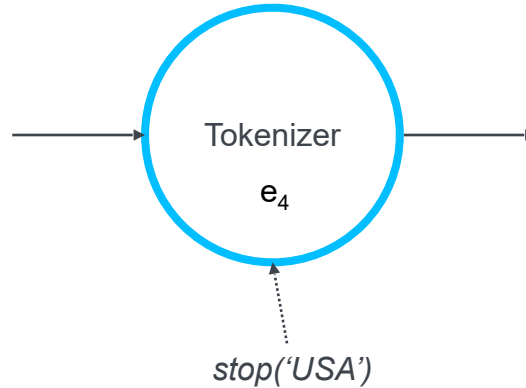
Preprocess and Tokenizer stages

$e_4 = \{ \textit{Name: The Terminator; Directed by: Cameron; Country: U.S.A., Starring: Schwarzenegger} \}$



preprocess + tokenize

$e_4 = \{ \textit{Terminator, Cameron, USA, Schwarzenegger} \}$



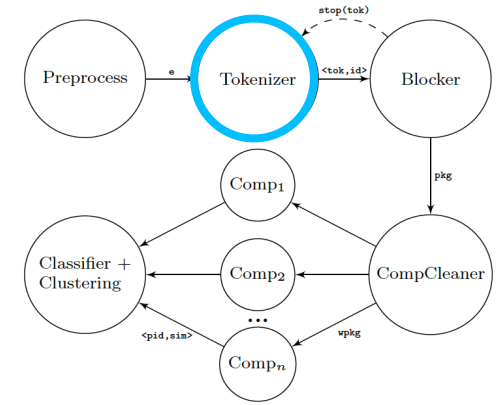
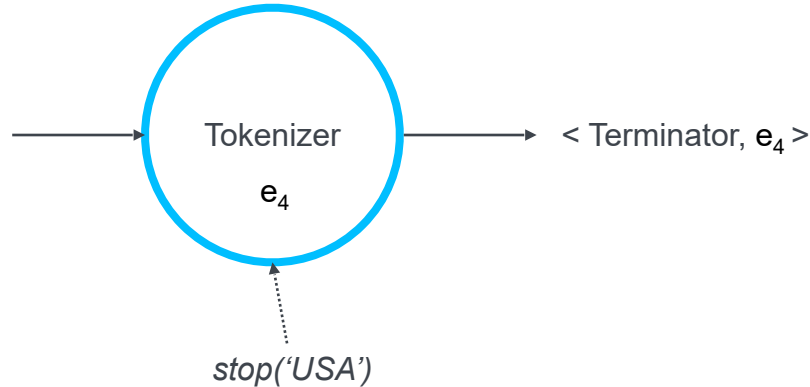
Preprocess and Tokenizer stages

$e_4 = \{ \text{Name: The Terminator; Directed by: Cameron; Country: U.S.A., Starring: Schwarzenegger} \}$



preprocess + tokenize

$e_4 = \{ \text{Terminator, Cameron, USA, Schwarzenegger} \}$



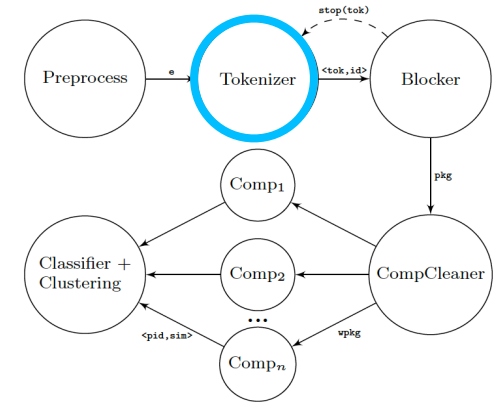
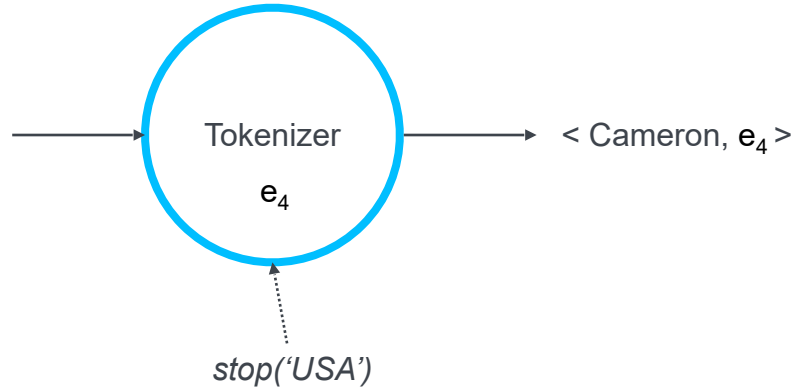
Preprocess and Tokenizer stages

$e_4 = \{ \textit{Name: The Terminator; Directed by: Cameron; Country: U.S.A., Starring: Schwarzenegger} \}$



preprocess + tokenize

$e_4 = \{ \textit{Terminator, Cameron, USA, Schwarzenegger} \}$



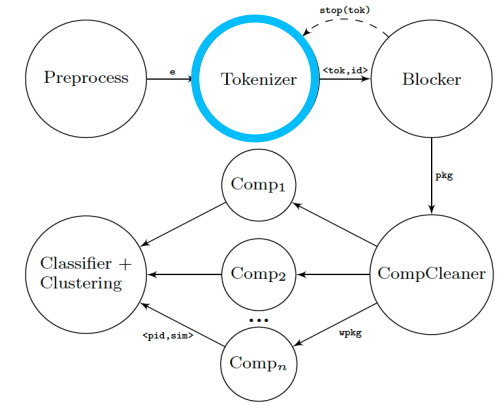
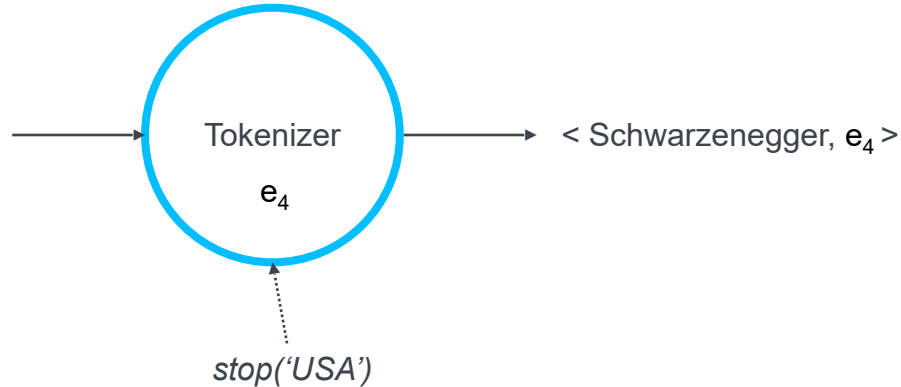
Preprocess and Tokenizer stages

$e_4 = \{ \text{Name: The Terminator; Directed by: Cameron; Country: U.S.A., Starring: Schwarzenegger} \}$



preprocess + tokenize

$e_4 = \{ \text{Terminator, Cameron, USA, Schwarzenegger} \}$



Blocker stage

From Tokenizer:

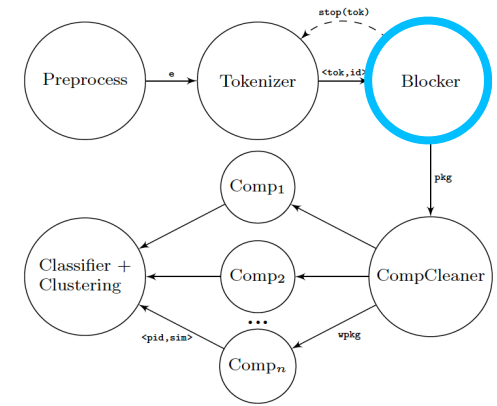
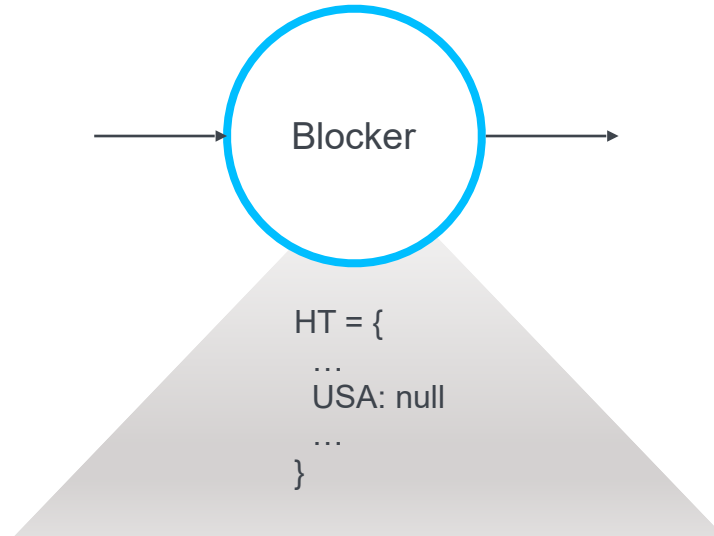
< USA, e_5 >

< USA, e_4 >

< USA, e_3 >

< USA, e_2 >

< USA, e_1 >



Blocker stage

From Tokenizer:

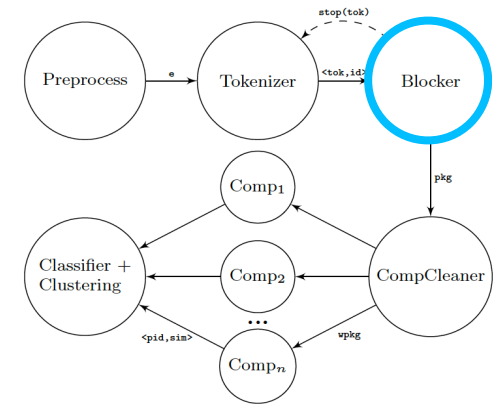
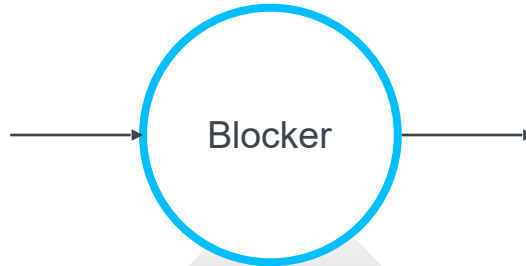
< USA, e_5 >

< USA, e_4 >

< USA, e_3 >

< USA, e_2 >

< USA, e_1 >



Blocker stage

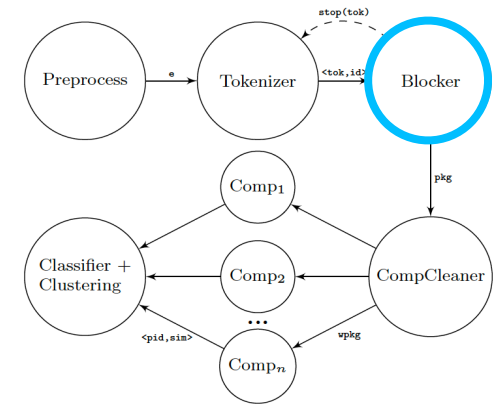
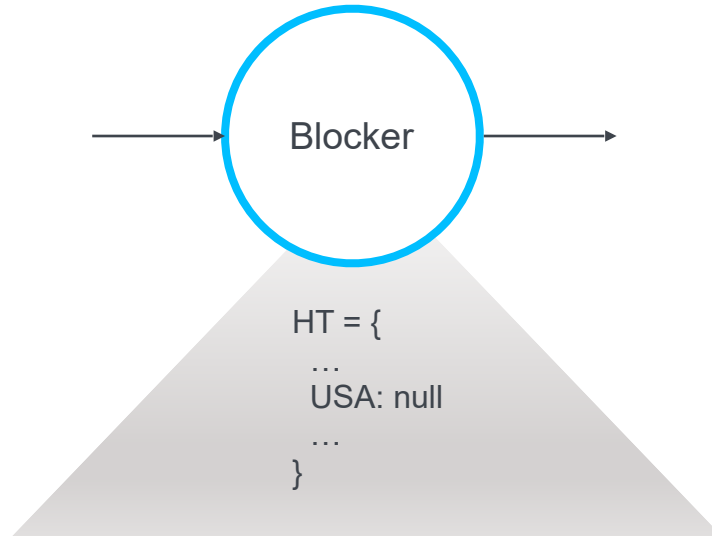
From Tokenizer:

< USA, e_5 >

< USA, e_4 >

< USA, e_3 >

< USA, e_2 >



Blocker stage

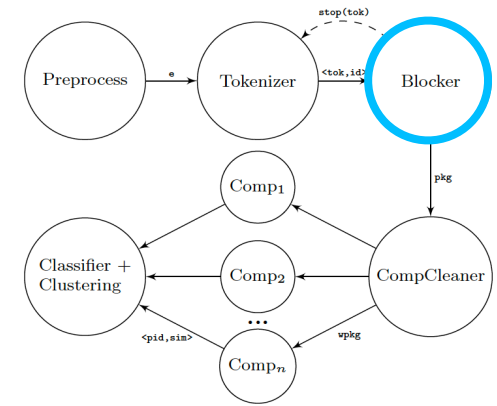
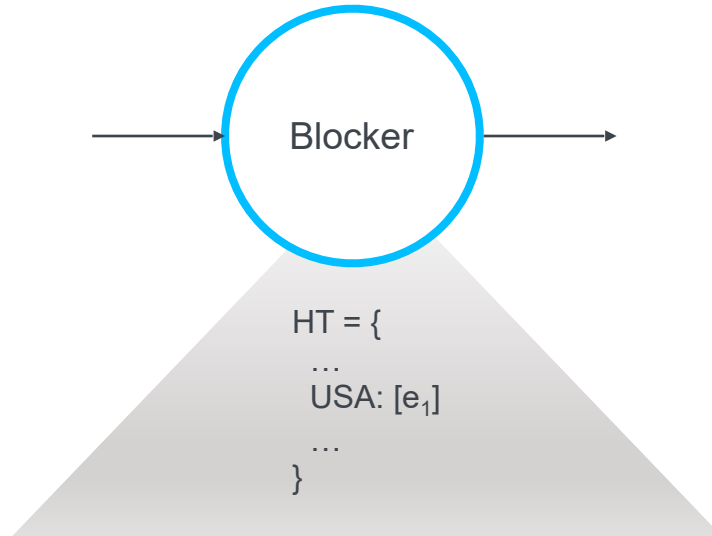
From Tokenizer:

< USA, e_5 >

< USA, e_4 >

< USA, e_3 >

< USA, e_2 >



Blocker stage

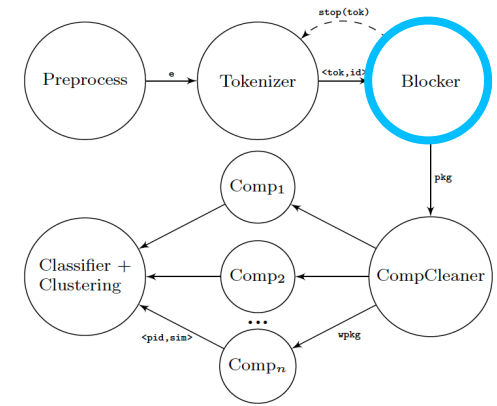
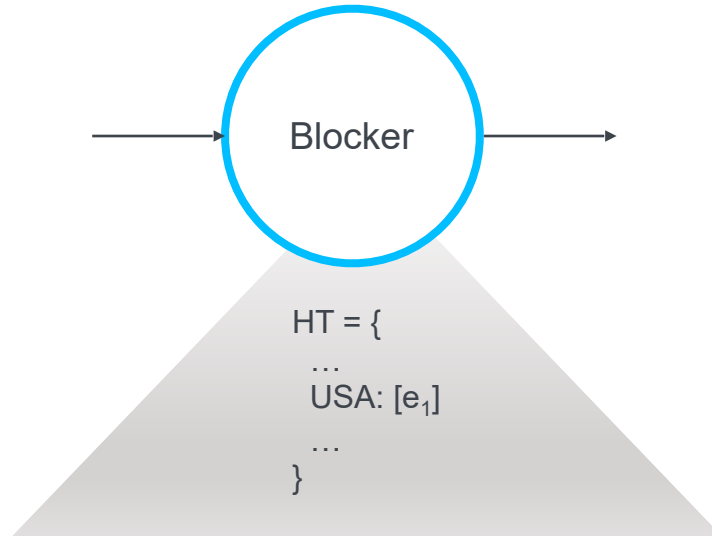
From Tokenizer:

< USA, e_5 >

< USA, e_4 >

< USA, e_3 >

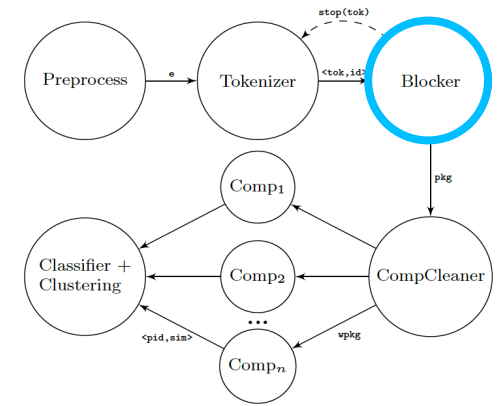
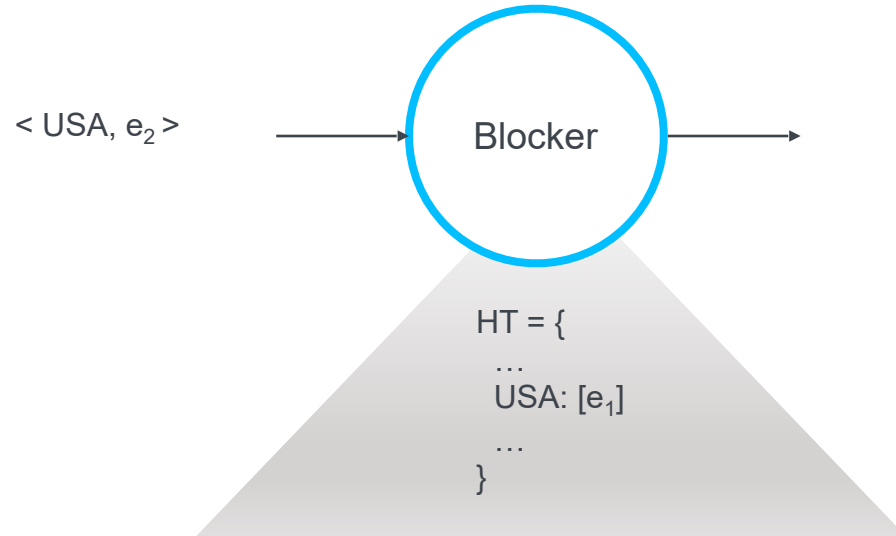
< USA, e_2 >



Blocker stage

From Tokenizer:

< USA, e_5 >
< USA, e_4 >
< USA, e_3 >



Blocker stage

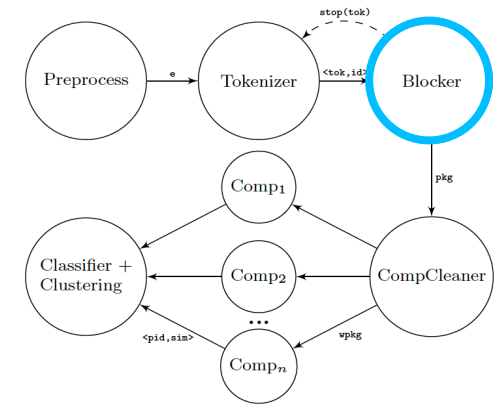
From Tokenizer:

< USA, e_5 >
< USA, e_4 >
< USA, e_3 >

< USA, e_2 >

Blocker

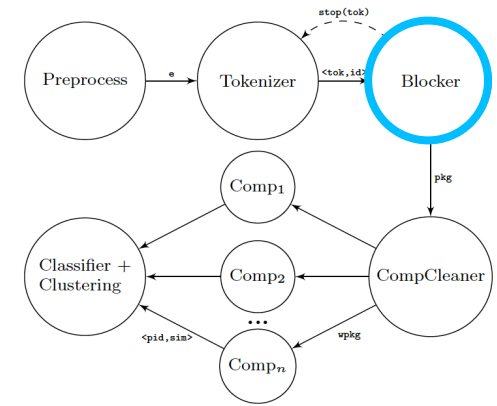
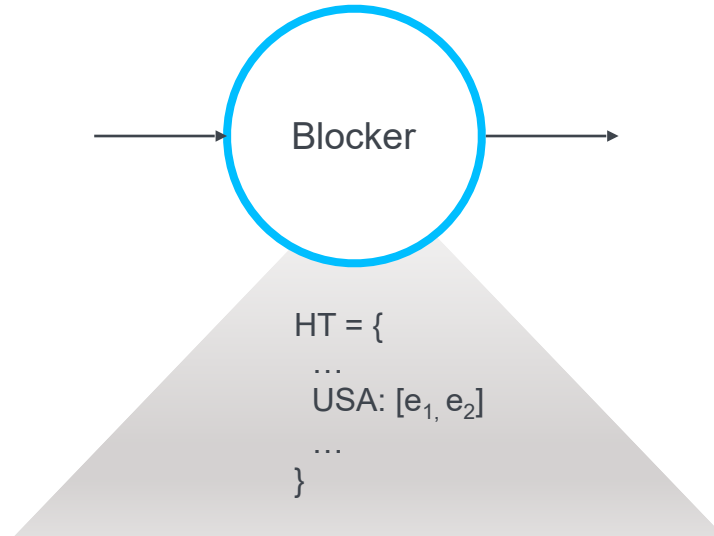
HT = {
...
USA: [e_1 , e_2]
...
}



Blocker stage

From Tokenizer:

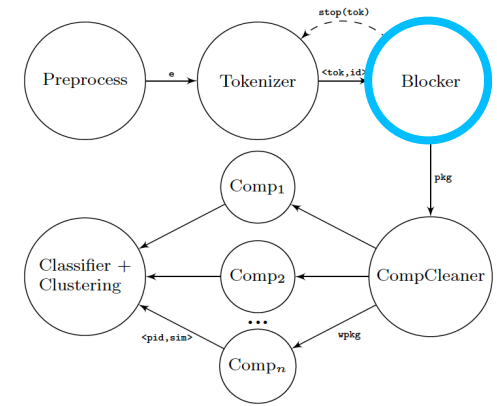
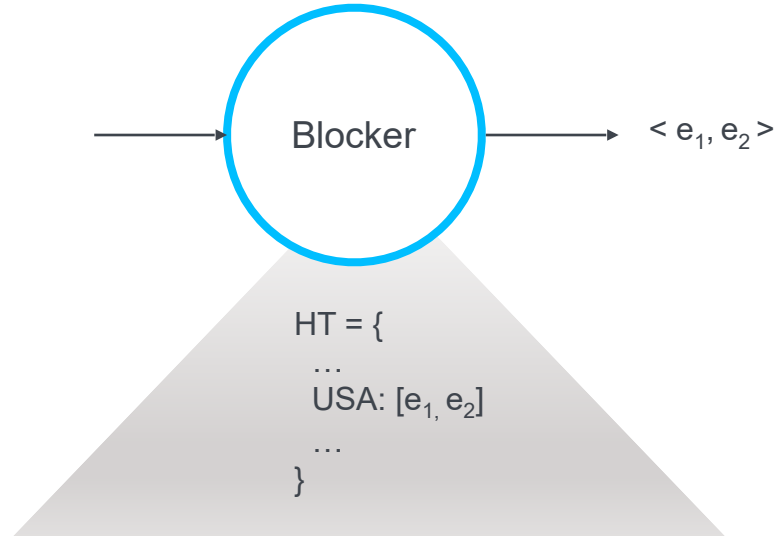
< USA, e_5 >
< USA, e_4 >
< USA, e_3 >



Blocker stage

From Tokenizer:

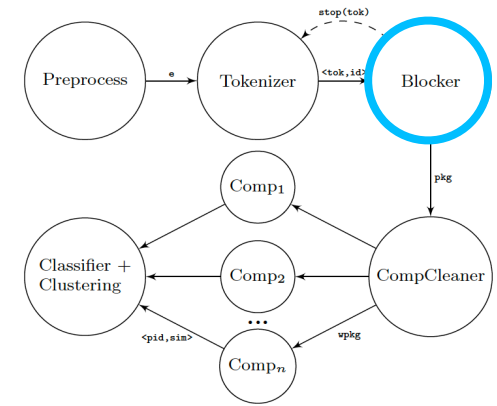
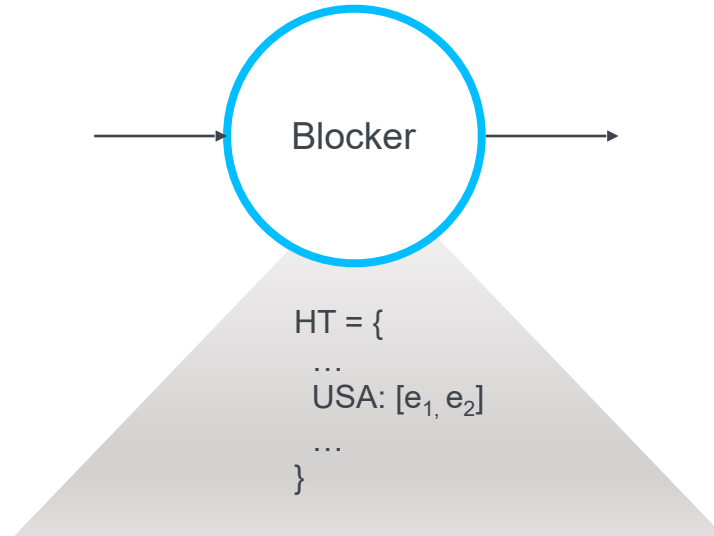
$\langle \text{USA}, e_5 \rangle$
 $\langle \text{USA}, e_4 \rangle$
 $\langle \text{USA}, e_3 \rangle$



Blocker stage

From Tokenizer:

< USA, e_5 >
< USA, e_4 >
< USA, e_3 >



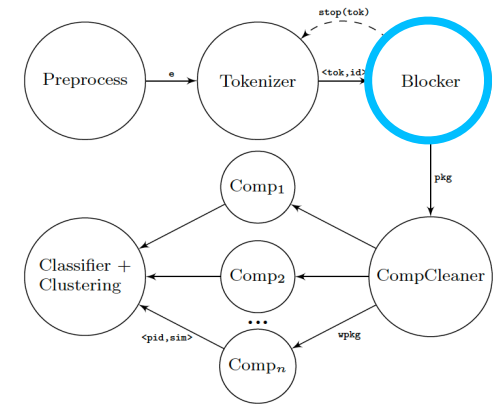
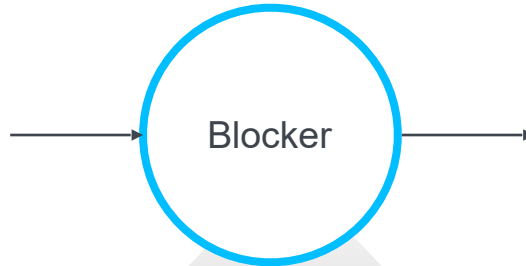
Blocker stage

From Tokenizer:

$\langle \text{USA}, e_5 \rangle$

$\langle \text{USA}, e_4 \rangle$

$\langle \text{USA}, e_3 \rangle$



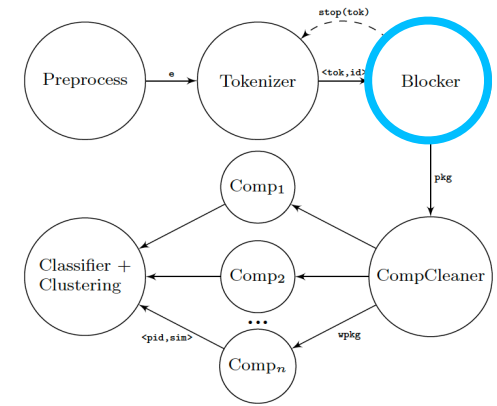
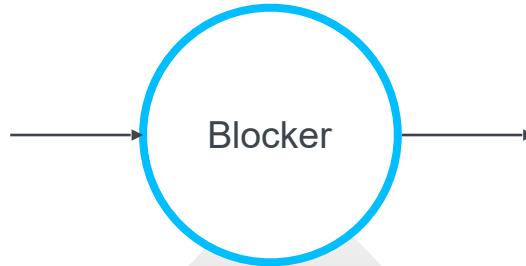
Blocker stage

From Tokenizer:

< USA, e_5 >

< USA, e_4 >

< USA, e_3 >

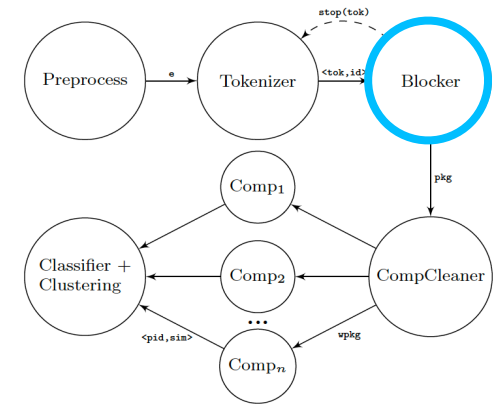
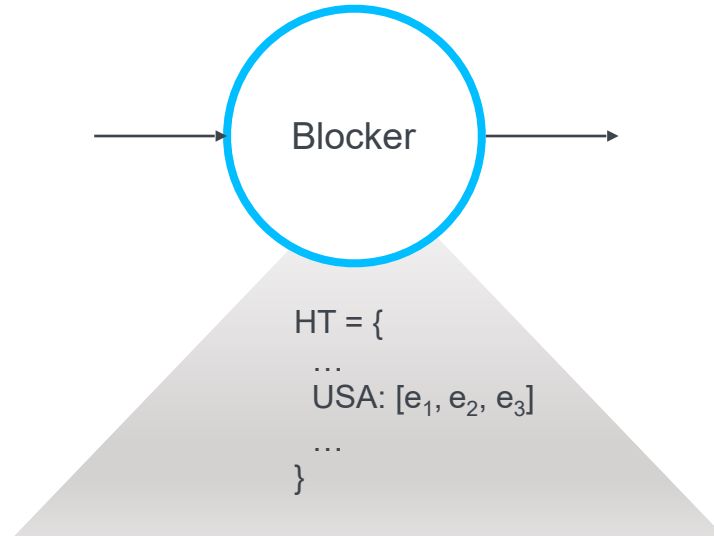


Blocker stage

From Tokenizer:

$\langle \text{USA}, e_5 \rangle$

$\langle \text{USA}, e_4 \rangle$

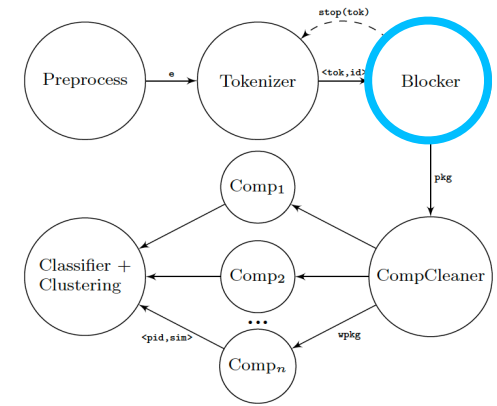
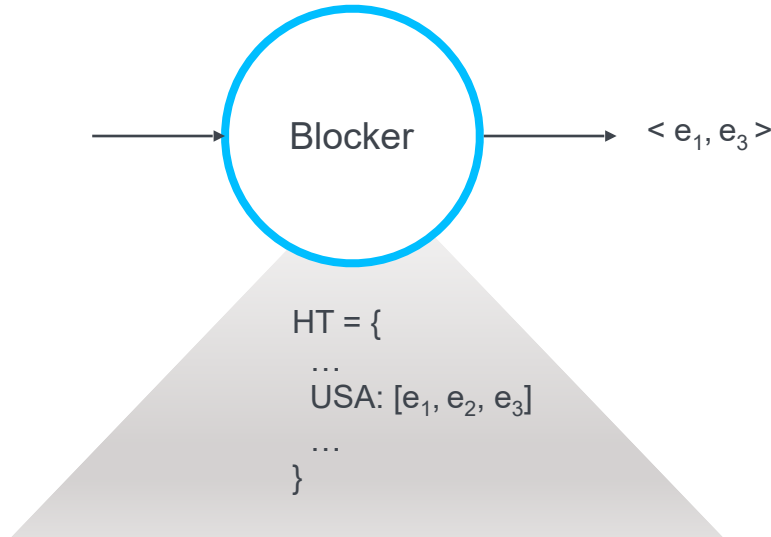


Blocker stage

From Tokenizer:

$\langle \text{USA}, e_5 \rangle$

$\langle \text{USA}, e_4 \rangle$

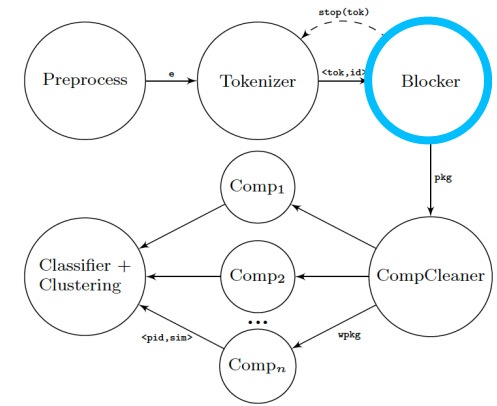
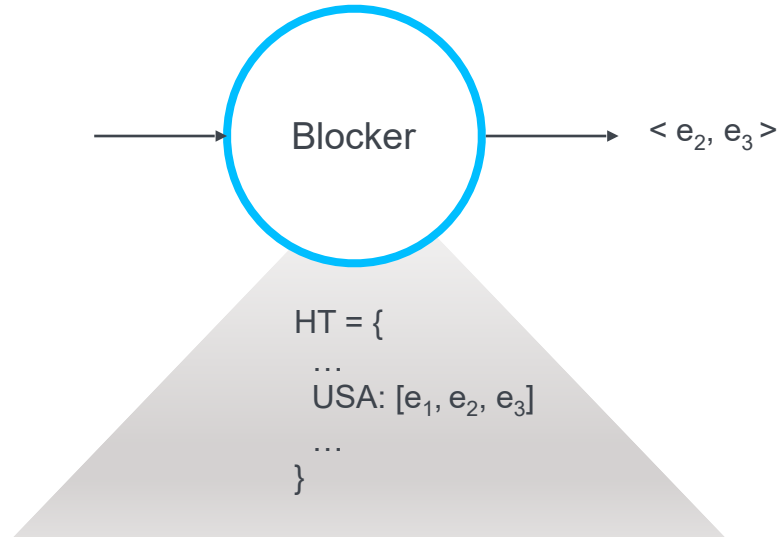


Blocker stage

From Tokenizer:

$\langle \text{USA}, e_5 \rangle$

$\langle \text{USA}, e_4 \rangle$

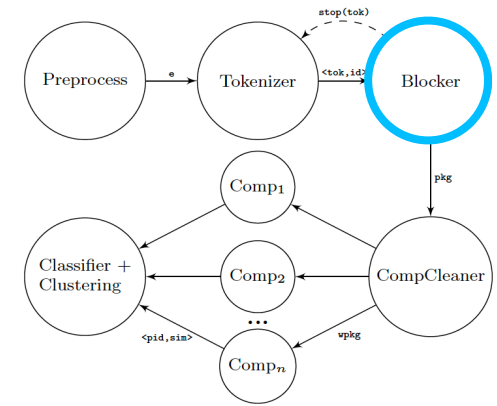
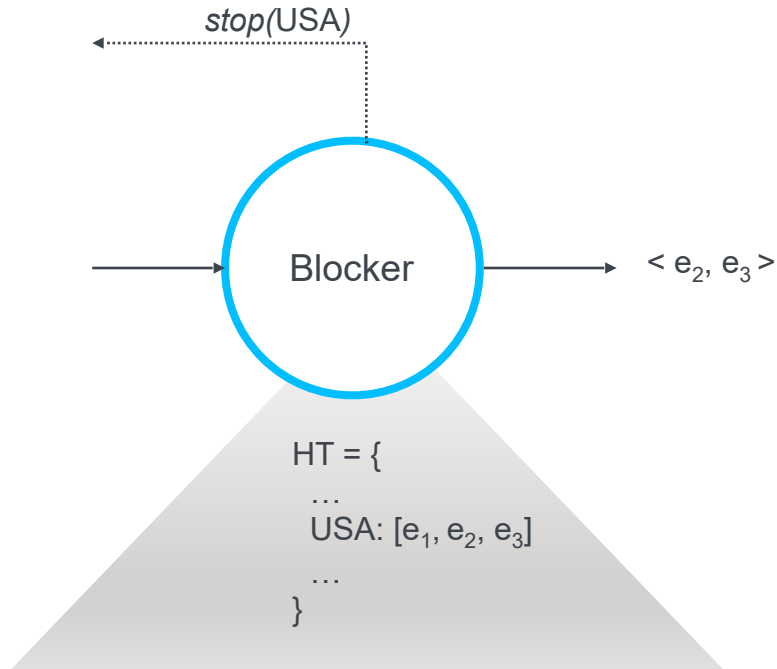


Blocker stage

From Tokenizer:

$\langle \text{USA}, e_5 \rangle$

$\langle \text{USA}, e_4 \rangle$

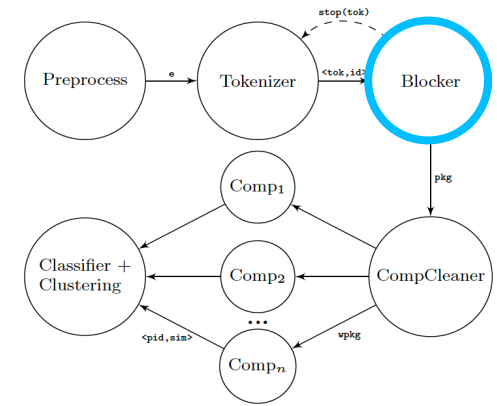
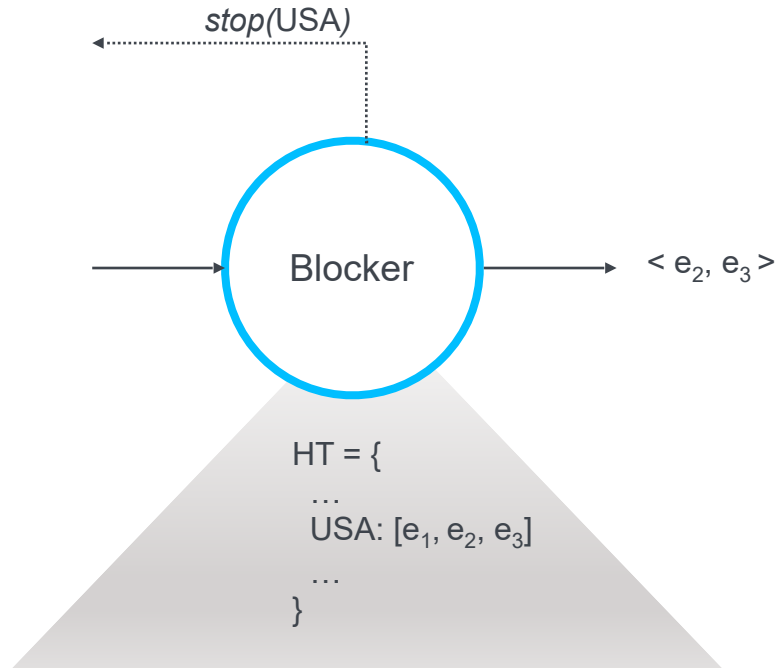


Blocker stage

From Tokenizer:

~~< USA, e₅ >~~

~~< USA, e₄ >~~



CompCleaner stage

From Blocker:

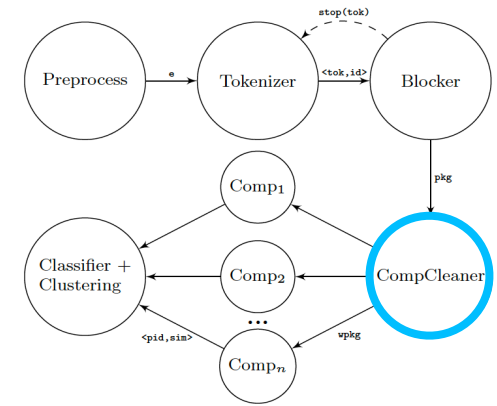
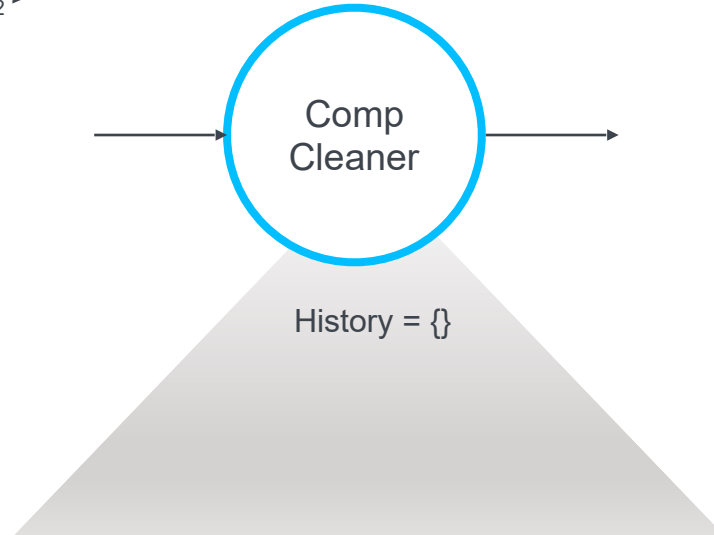
...

USA < e_1 , e_3 >

Inception < e_1 , e_3 >

USA < e_1 , e_2 >

DiCaprio < e_1 , e_2 >



CompCleaner stage

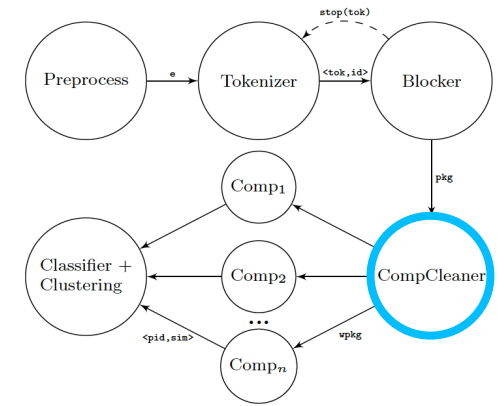
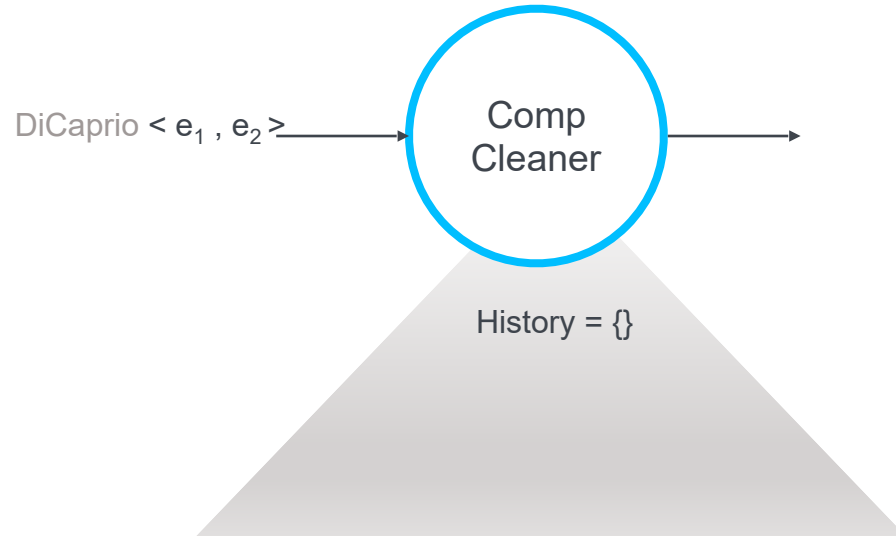
From Blocker:

...

USA < e_1 , e_3 >

Inception < e_1 , e_3 >

USA < e_1 , e_2 >



CompCleaner stage

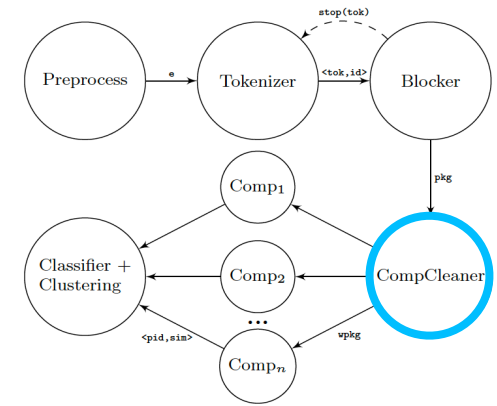
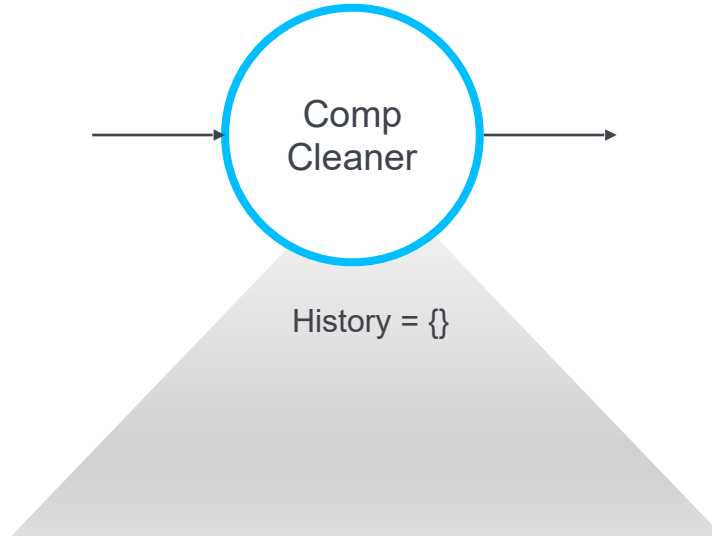
From Blocker:

...

USA < e_1 , e_3 >

Inception < e_1 , e_3 >

USA < e_1 , e_2 >



CompCleaner stage

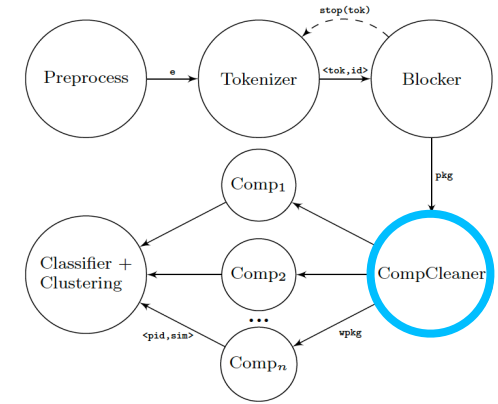
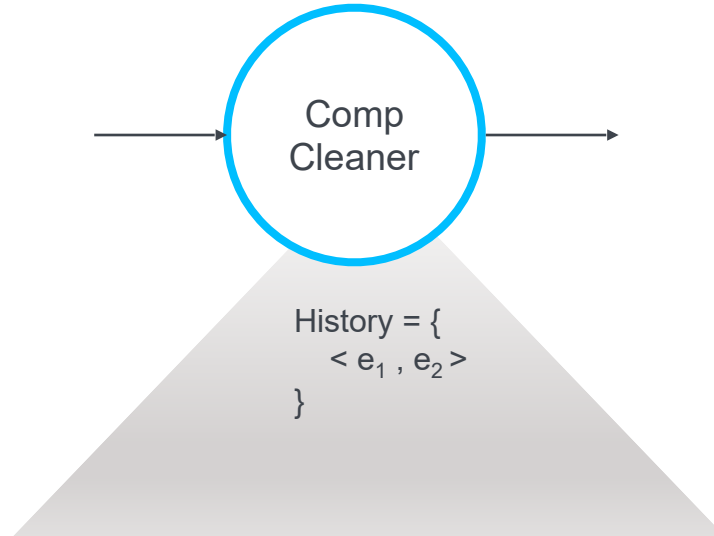
From Blocker:

...

USA < e_1 , e_3 >

Inception < e_1 , e_3 >

USA < e_1 , e_2 >



CompCleaner stage

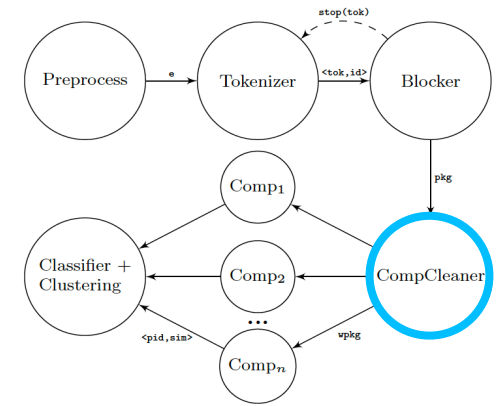
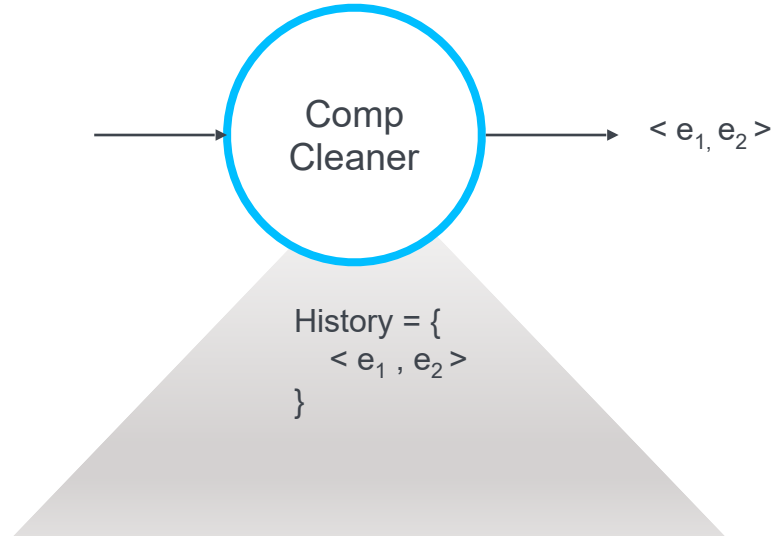
From Blocker:

...

USA $\langle e_1, e_3 \rangle$

Inception $\langle e_1, e_3 \rangle$

USA $\langle e_1, e_2 \rangle$



CompCleaner stage

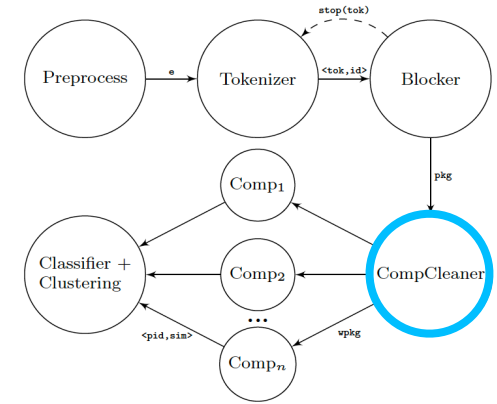
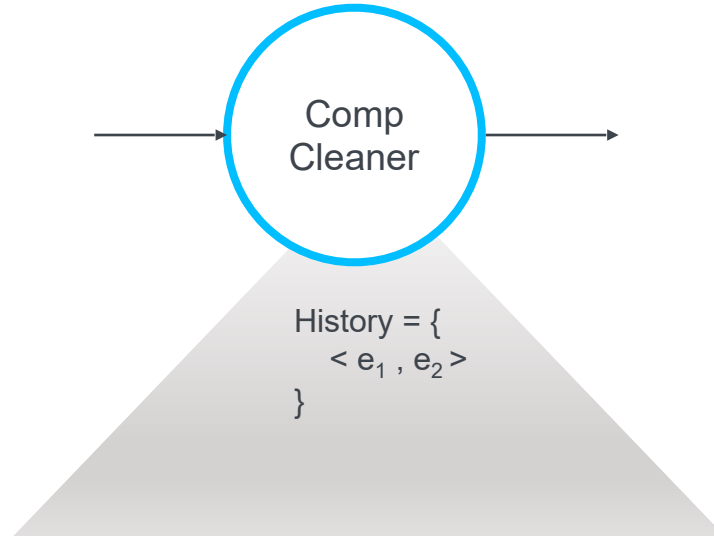
From Blocker:

...

USA < e_1 , e_3 >

Inception < e_1 , e_3 >

USA < e_1 , e_2 >



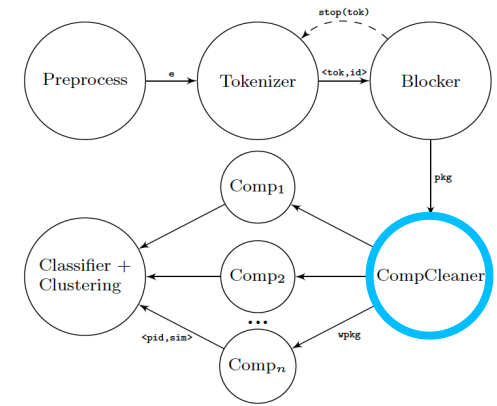
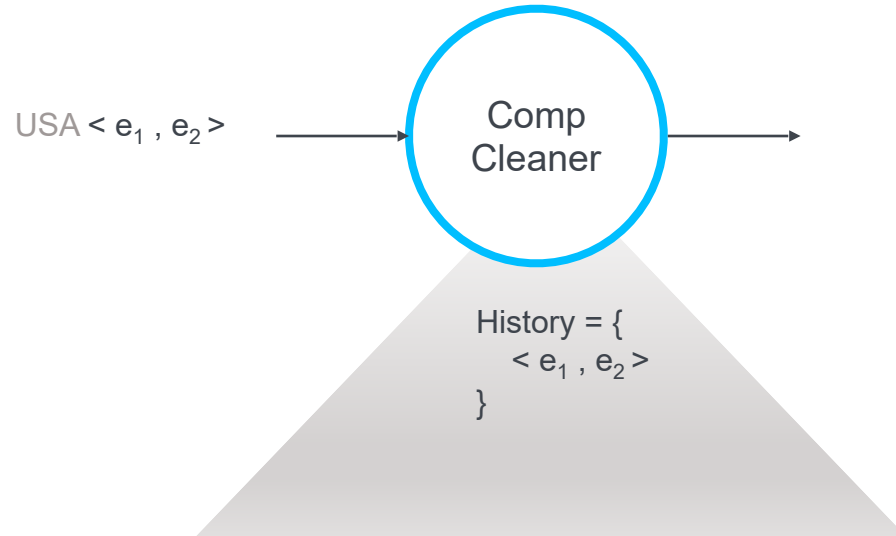
CompCleaner stage

From Blocker:

...

USA $\langle e_1, e_3 \rangle$

Inception $\langle e_1, e_3 \rangle$



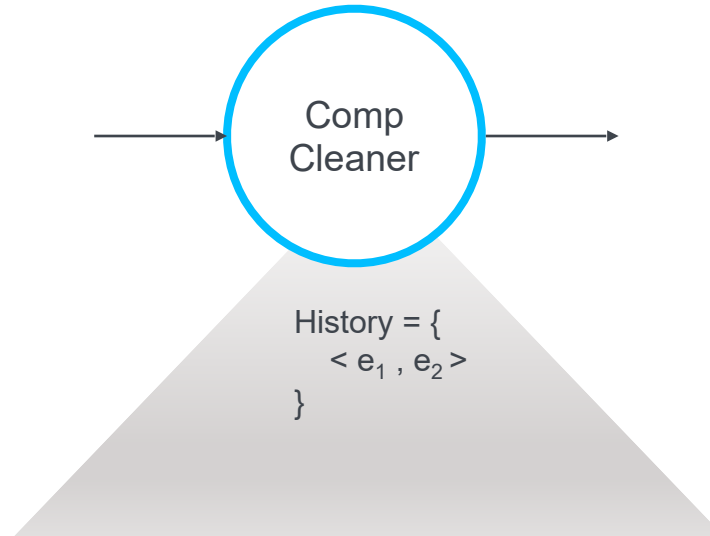
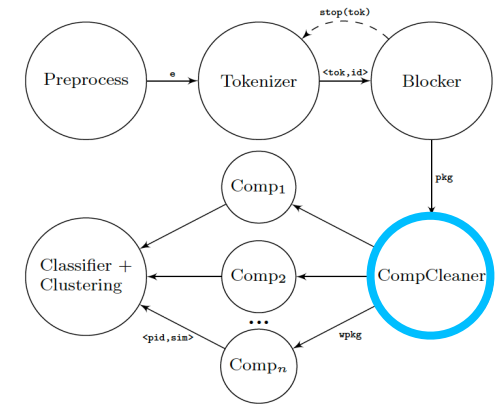
CompCleaner stage

From Blocker:

...

USA $\langle e_1, e_3 \rangle$

Inception $\langle e_1, e_3 \rangle$



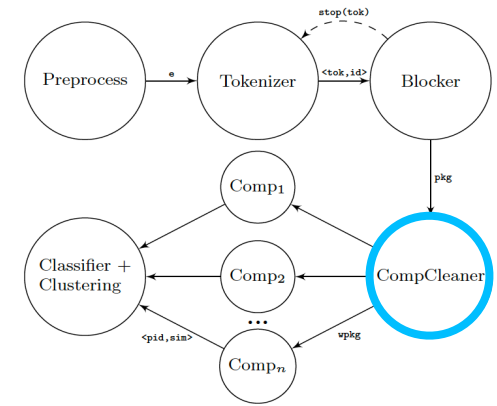
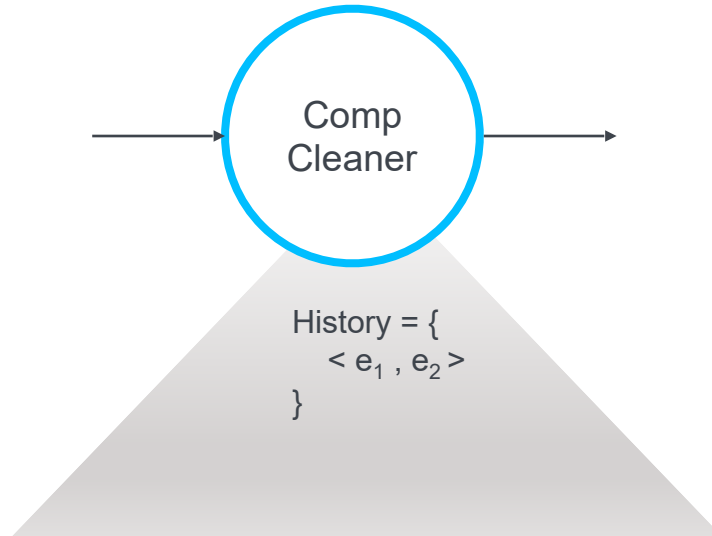
CompCleaner stage

From Blocker:

...

USA < e_1 , e_3 >

Inception < e_1 , e_3 >

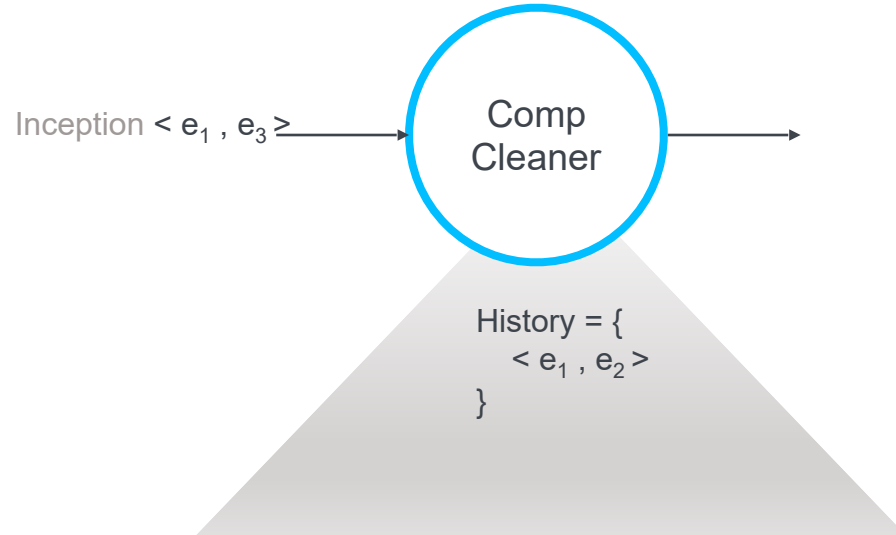
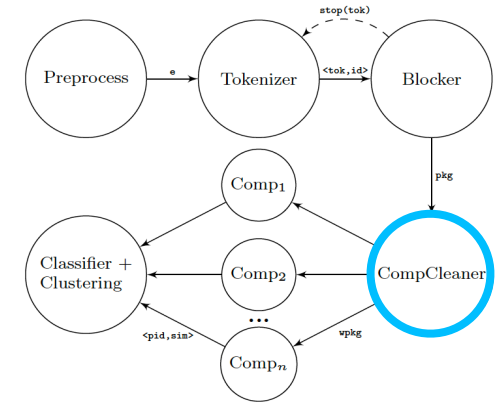


CompCleaner stage

From Blocker:

...

USA < e_1 , e_3 >

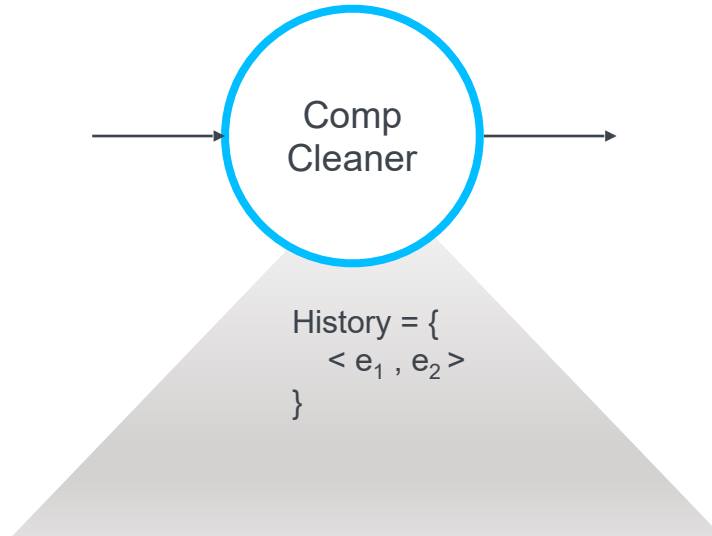
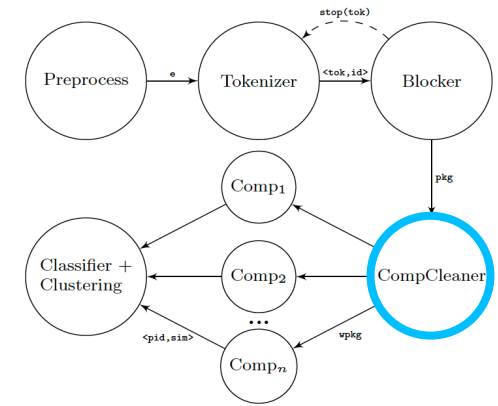


CompCleaner stage

From Blocker:

...

USA < e_1 , e_3 >

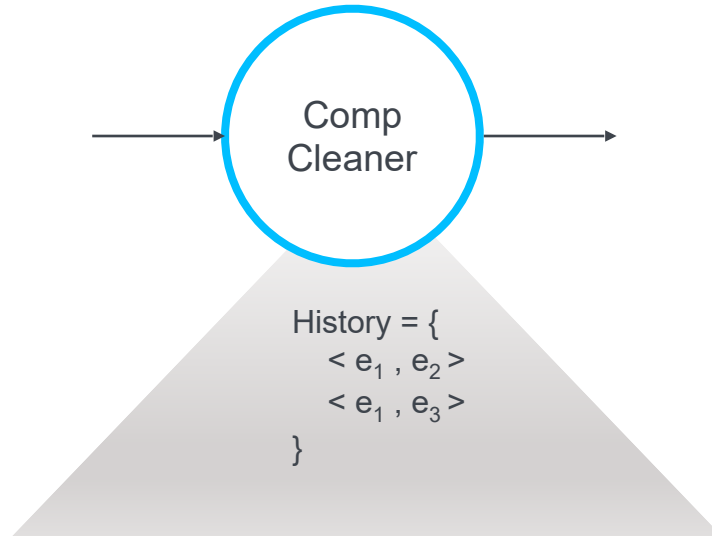
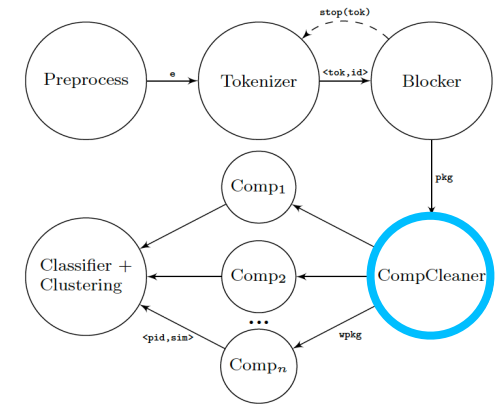


CompCleaner stage

From Blocker:

...

USA < e_1 , e_3 >

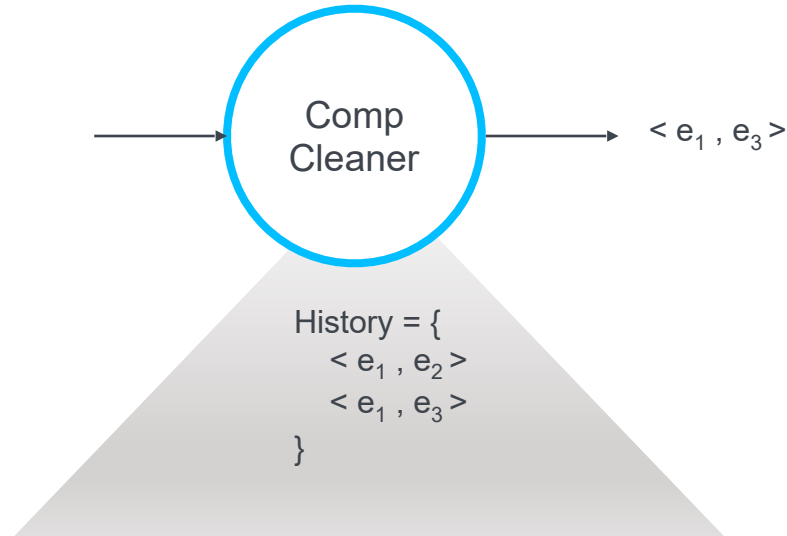
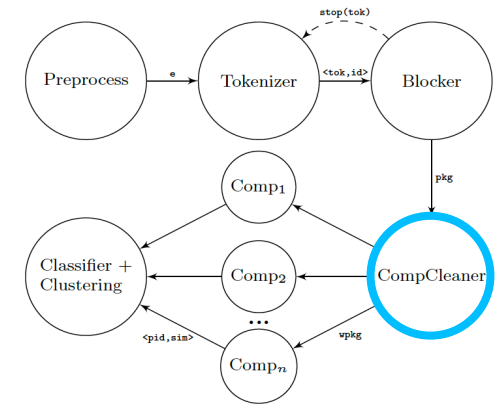


CompCleaner stage

From Blocker:

...

USA $\langle e_1, e_3 \rangle$

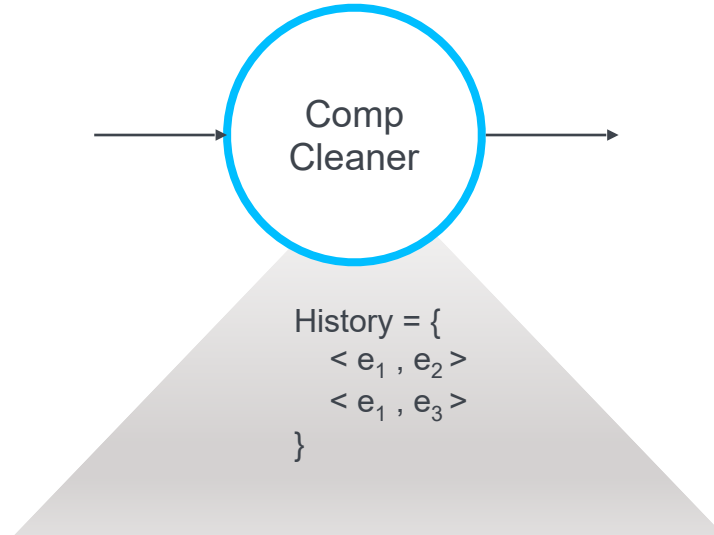
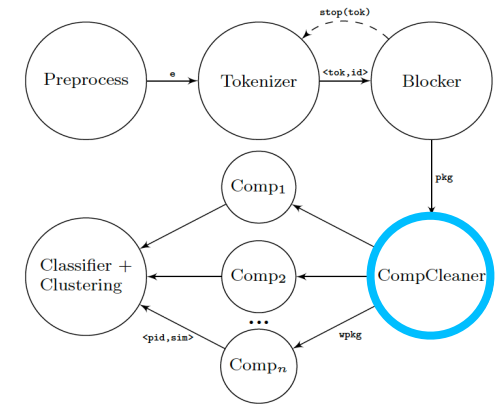


CompCleaner stage

From Blocker:

...

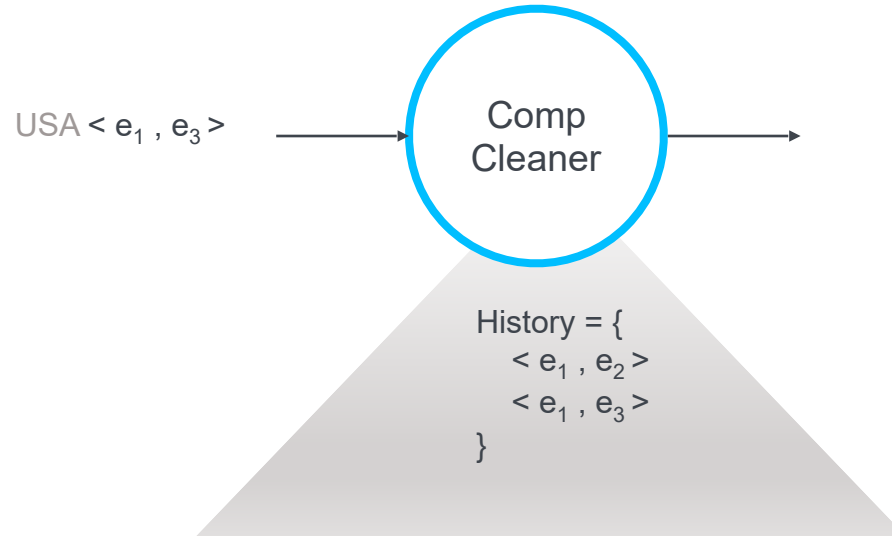
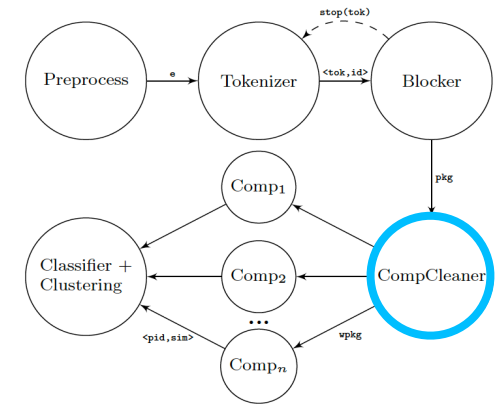
USA < e_1 , e_3 >



CompCleaner stage

From Blocker:

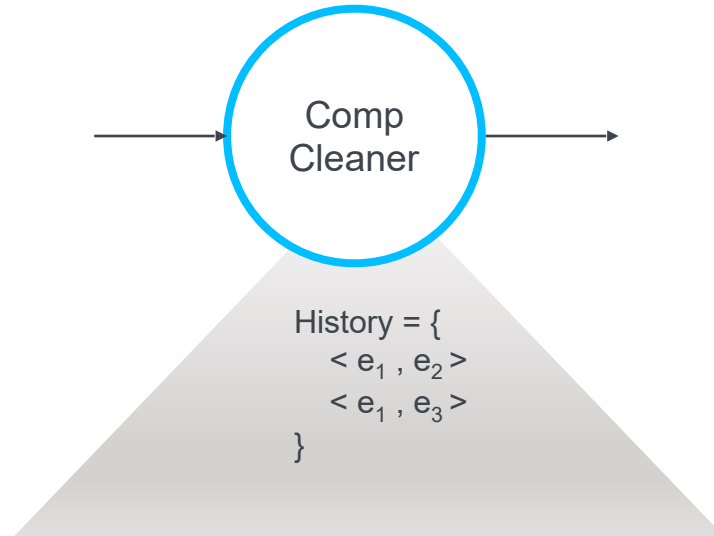
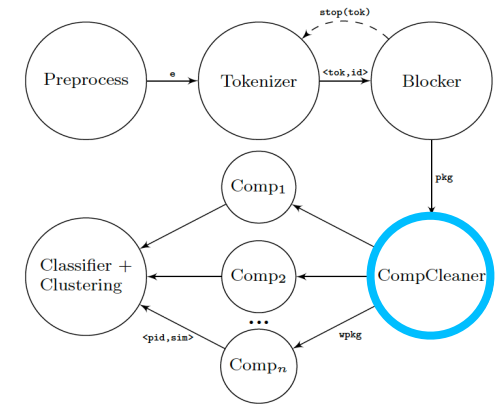
...



CompCleaner stage

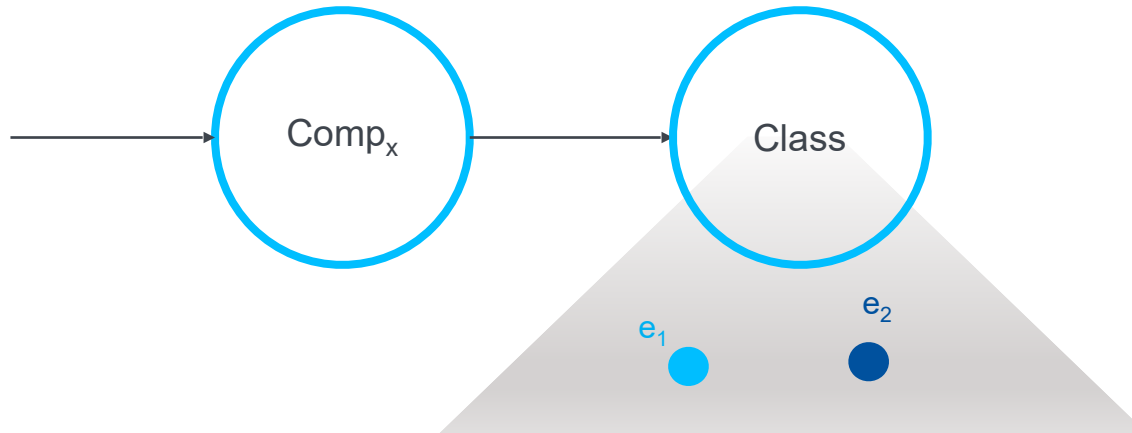
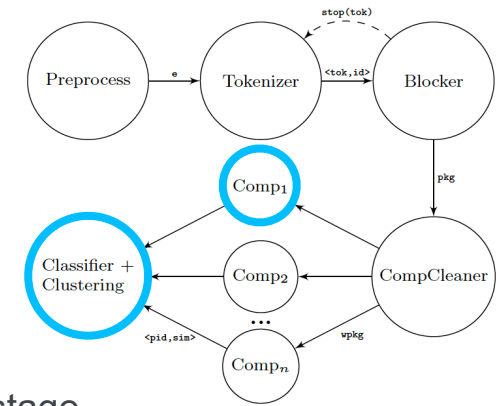
From Blocker:

...



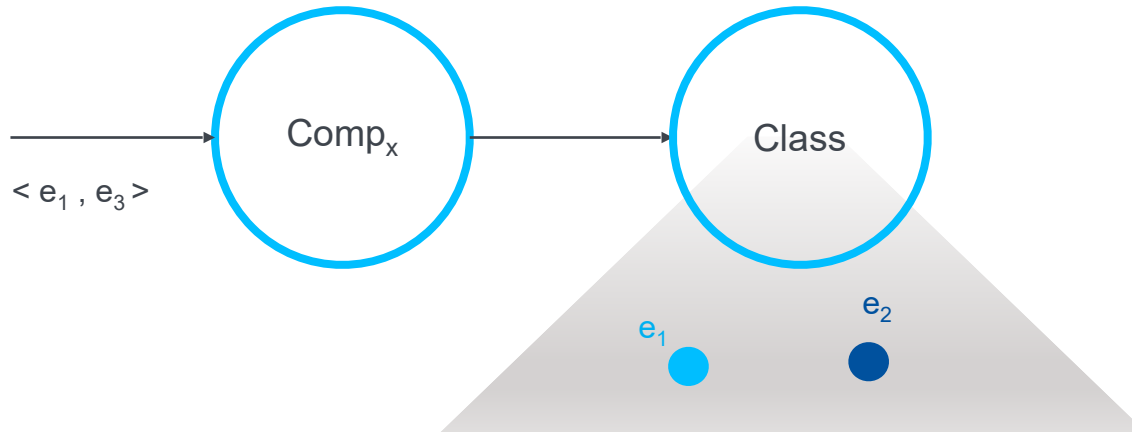
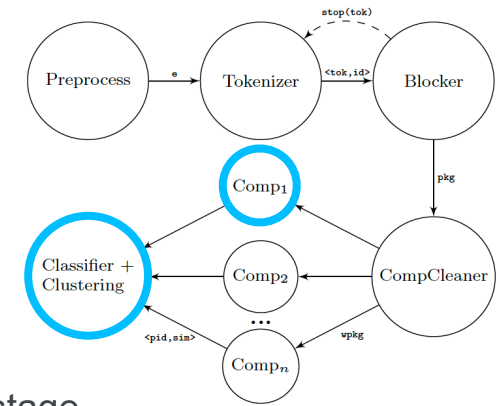
Comparator and classifier

- Comparator stage
 - Performs comparisons for each pair received
 - Outputs results (similar scores associated to the pairs) to the next stage
- Classifier stage
 - Performs incremental clustering-based approaches
 - E.g. incremental transitivity checks for matches considering thresholds



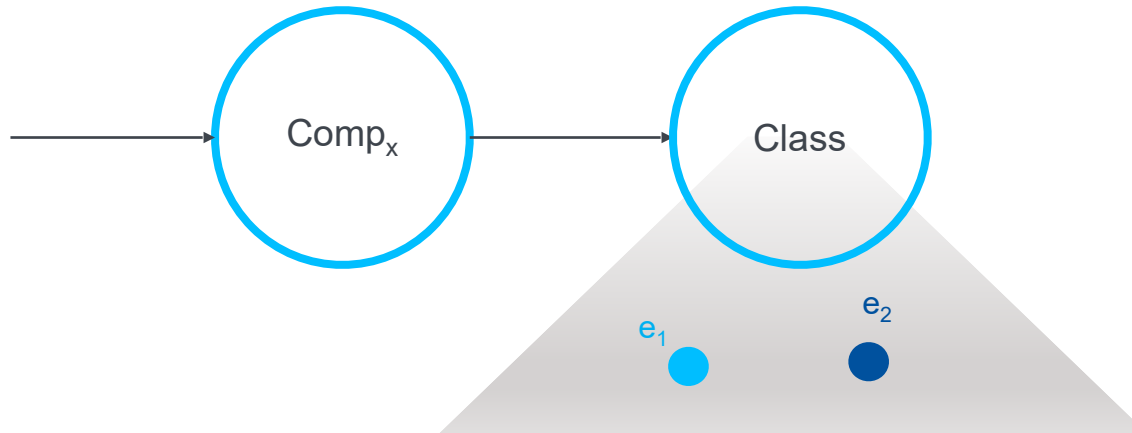
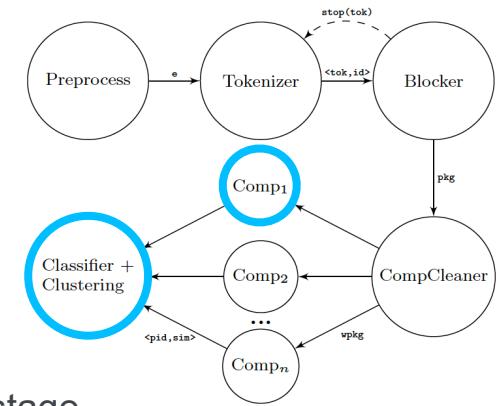
Comparator and classifier

- Comparator stage
 - Performs comparisons for each pair received
 - Outputs results (similar scores associated to the pairs) to the next stage
- Classifier stage
 - Performs incremental clustering-based approaches
 - E.g. incremental transitivity checks for matches considering thresholds



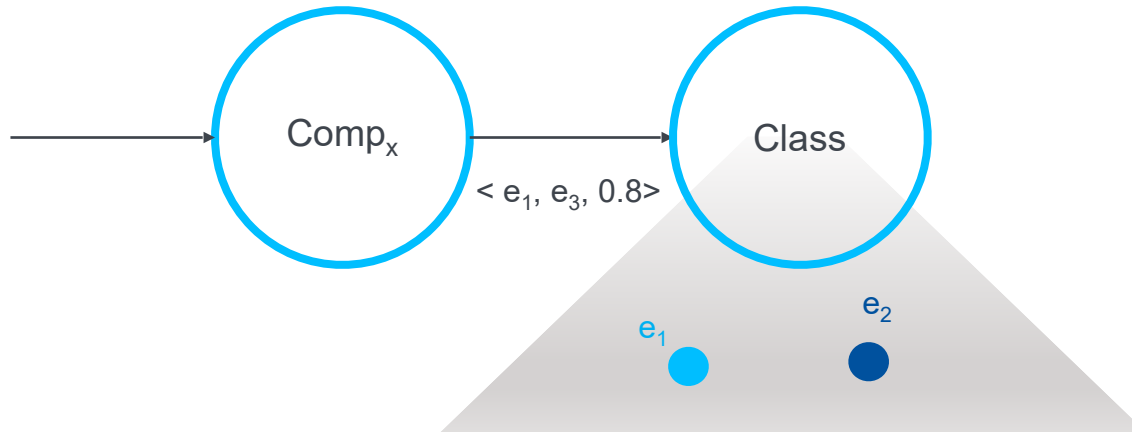
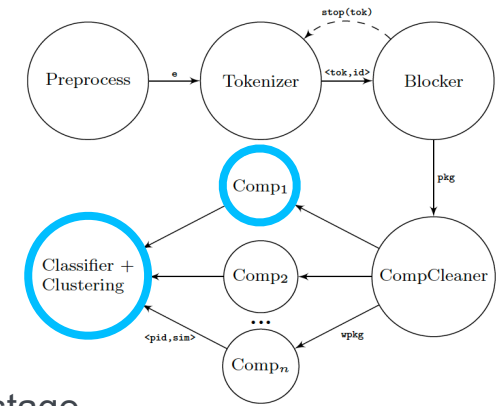
Comparator and classifier

- Comparator stage
 - Performs comparisons for each pair received
 - Outputs results (similar scores associated to the pairs) to the next stage
- Classifier stage
 - Performs incremental clustering-based approaches
 - E.g. incremental transitivity checks for matches considering thresholds



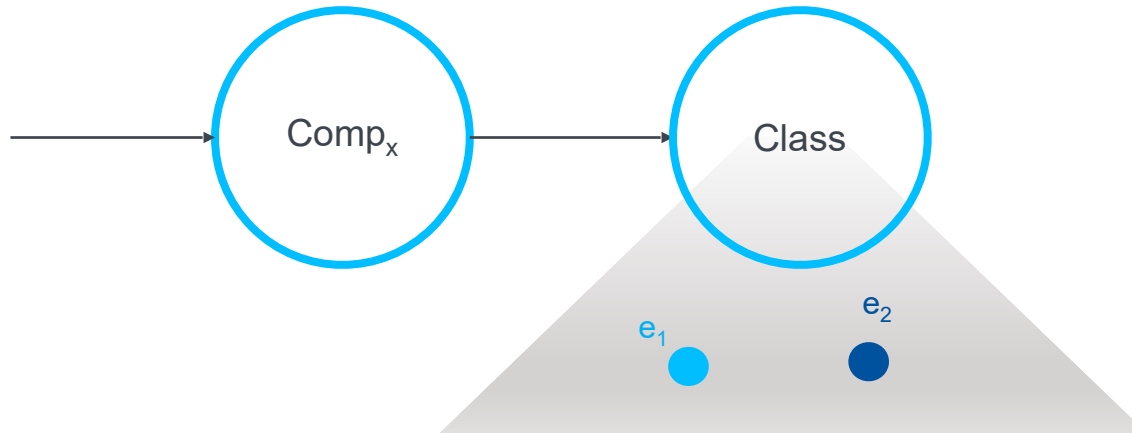
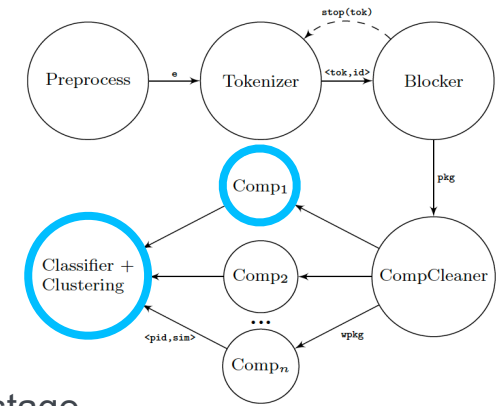
Comparator and classifier

- Comparator stage
 - Performs comparisons for each pair received
 - Outputs results (similar scores associated to the pairs) to the next stage
- Classifier stage
 - Performs incremental clustering-based approaches
 - E.g. incremental transitivity checks for matches considering thresholds



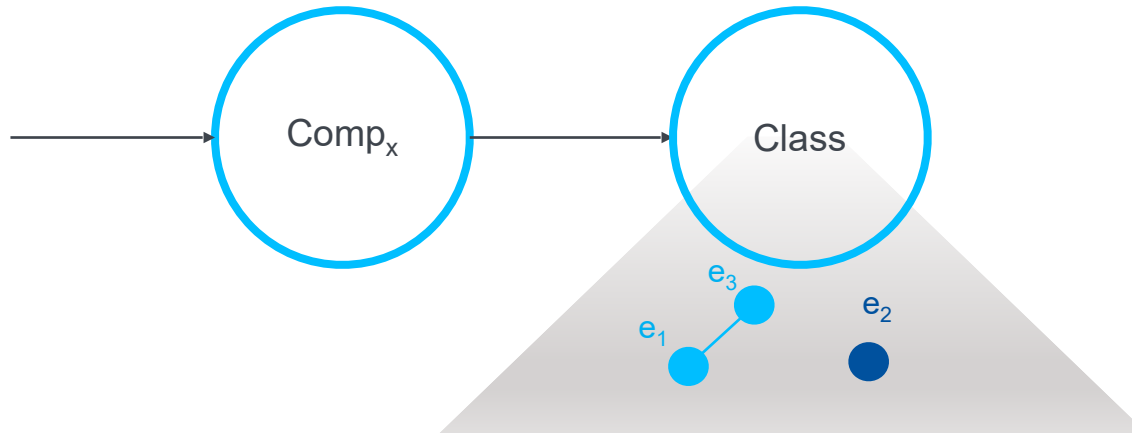
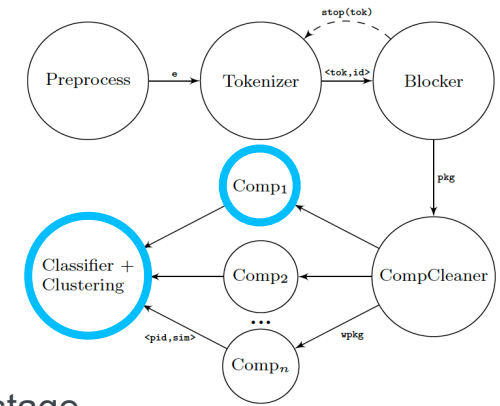
Comparator and classifier

- Comparator stage
 - Performs comparisons for each pair received
 - Outputs results (similar scores associated to the pairs) to the next stage
- Classifier stage
 - Performs incremental clustering-based approaches
 - E.g. incremental transitivity checks for matches considering thresholds



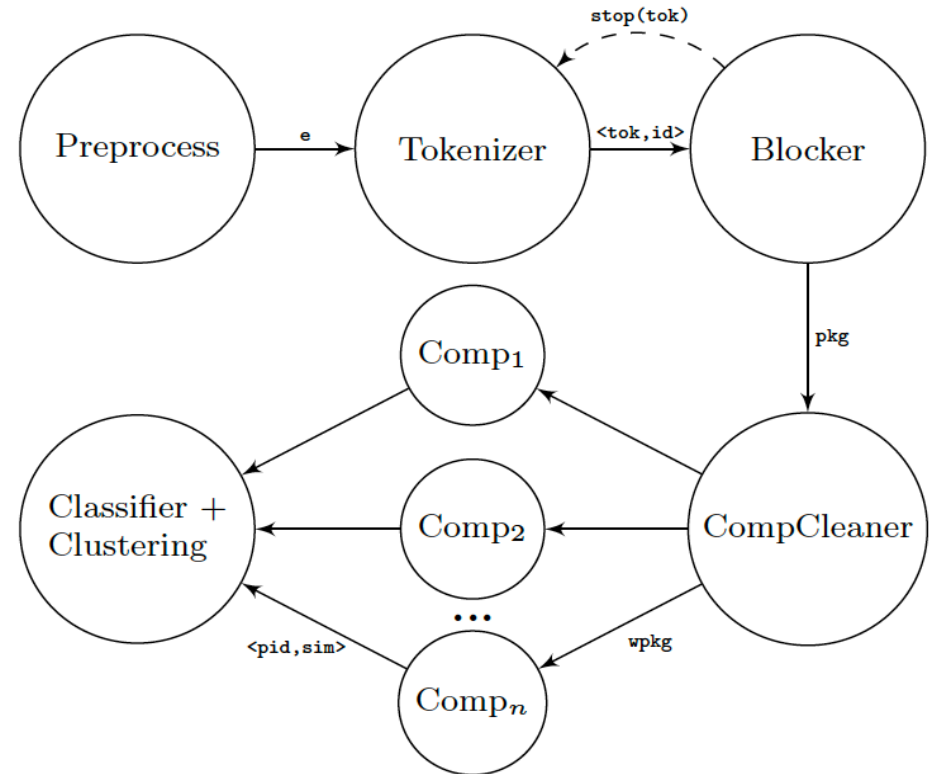
Comparator and classifier

- Comparator stage
 - Performs comparisons for each pair received
 - Outputs results (similar scores associated to the pairs) to the next stage
- Classifier stage
 - Performs incremental clustering-based approaches
 - E.g. incremental transitivity checks for matches considering thresholds



Conclusion

- Task-based parallelization for ER that allows to process heterogeneous data
- The solution allows us to support incremental ER where results are continuously returned



Future work

- Implement the ideal framework integrating the BF-based solutions and the baselines
- Study of different implementations for different computational architectures
 - Multi-core vs multi-computer
- Investigate further incremental Block Cleaning and CompCleaning solutions
 - E.g. Block cutting is a limited form of block purging

References

- [1] Christen, Peter. "A survey of indexing techniques for scalable record linkage and deduplication." *IEEE transactions on knowledge and data engineering* 24.9 (2012): 1537-1555.
- [2] Papadakis, George, et al. "Comparative analysis of approximate blocking techniques for entity resolution." *Proceedings of the VLDB Endowment* 9.9 (2016): 684-695.
- [3] Chen, Xiao, Eike Schallehn, and Gunter Saake. "Cloud-scale entity resolution: current state and open challenges." *Open Journal of Big Data (OJBD)* 4.1 (2018): 30-51.
- [4] Santos, Walter, et al. "A scalable parallel deduplication algorithm." *19th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD'07)*. IEEE, 2007.



Universität Stuttgart

Thank you!



Leonardo Gazzarri

E-Mail leonardo.gazzarri@ipvs.uni-stuttgart.de

Telefon +49 (0) 711 685-60888

Universität Stuttgart

IPVS - DE

Universitätsstraße 38, D-70569 Stuttgart