Assisted Curation: Does Text Mining Really Help?

(Alex et al. 2008)

by Benedict Fehringer

Seminar: „Unlocking the Secrets of the Past: Text Mining for Historical Documents“
Supervisor: Dr. Caroline Sporleder (and Martin Schreiber)

23.02.2012

Outline

- Introduction
- Related Work
- Assisted Curation
- Text Mining Pipeline
- Curation Experiments
- Discussion and Conclusion
- References

Outline

- Introduction
- Related Work
- Assisted Curation
- Text Mining Pipeline
- Curation Experiments
- Discussion and Conclusion
- References

Basic study elements
- Content -

- Curation of biomedical literature
  - For example, protein-protein interaction recognition:
    1. Which protein are there?
    2. If two proteins are named, are they in interaction?

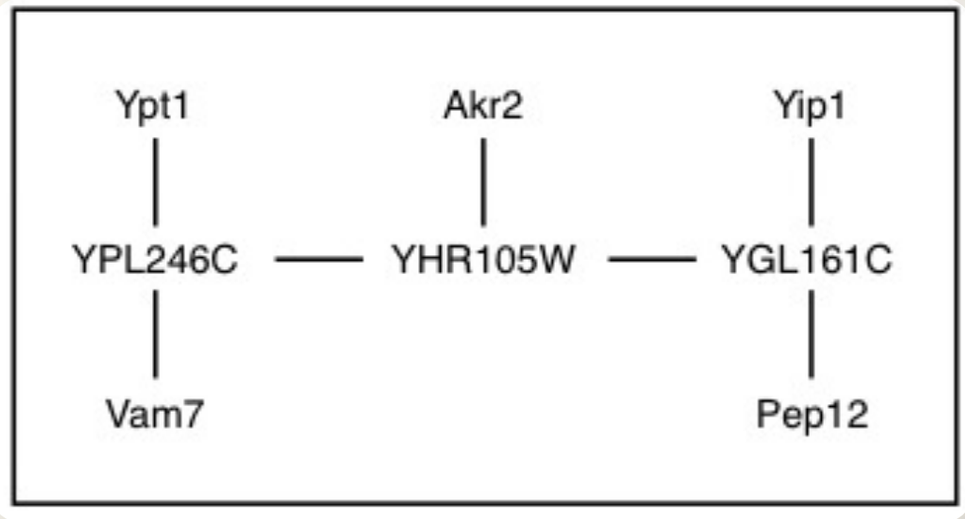
Example for protein-protein interaction recognition

[...] An example is YHR105W, which interacts with one protein involved in vesicular transport, Akr2, and with YGL161C, an uncharacterized protein that interacts with two transport proteins, Yip1 and Pep12. YHR105W also interacts with YPL246C, another uncharacterized protein that interacts with Ypt1 and Vam7, proteins implicated in vesicular transport and membrane fusion, respectively. [...]

Source: Schwikowski, Uetz, & Fields (pp. 1259, 2000)

1. Which proteins are there?

2. If two proteins are named, are they in interaction?



Basic study elements
- Research Question -

- Curation of biomedical literature
  - For example, protein-protein interaction recognition:
    1. Which protein are there?
    2. If two proteins are named, are they in interaction?
- Task should be supported by text mining

Related Work

- Increasing development of information extraction systems (spurred on by BioCreAtIvE II competition; Krallinger, Leitner, & Valencia, 2007)
  - studies suggest reduction of curation time

- But: lack of user studies for extrinsically evaluation
  - no validation by curator feedback about affecting their work and usefulness

Basic study elements
- Evaluation -

- Curation of biomedical literature
  - For example, protein-protein interaction recognition:
    1. Which protein are there?
    2. If two proteins are named, are they in interaction?
- Task should be supported by text mining
- Evaluation by:
  - objective performance metrics (e.g. speed improvement, number of records)
  - focusing on user feedback, too

Outline

- Introduction
- Related Work
- Assisted Curation
- Text Mining Pipeline
- Curation Experiments
- Discussion and Conclusion
- References

Curation Scenario
- General -

- Goal: Curators should identify protein-protein interactions (PPIs)
- Initial step: Providing set of matching papers
- Middle step: Filtering papers into candidates

Curation Scenario
- General -

- Goal: Curators should identify protein-protein interactions (PPIs)
- Initial step: Providing set of mat
- Middle step: Filtering papers in

How can NLP help the curator work?

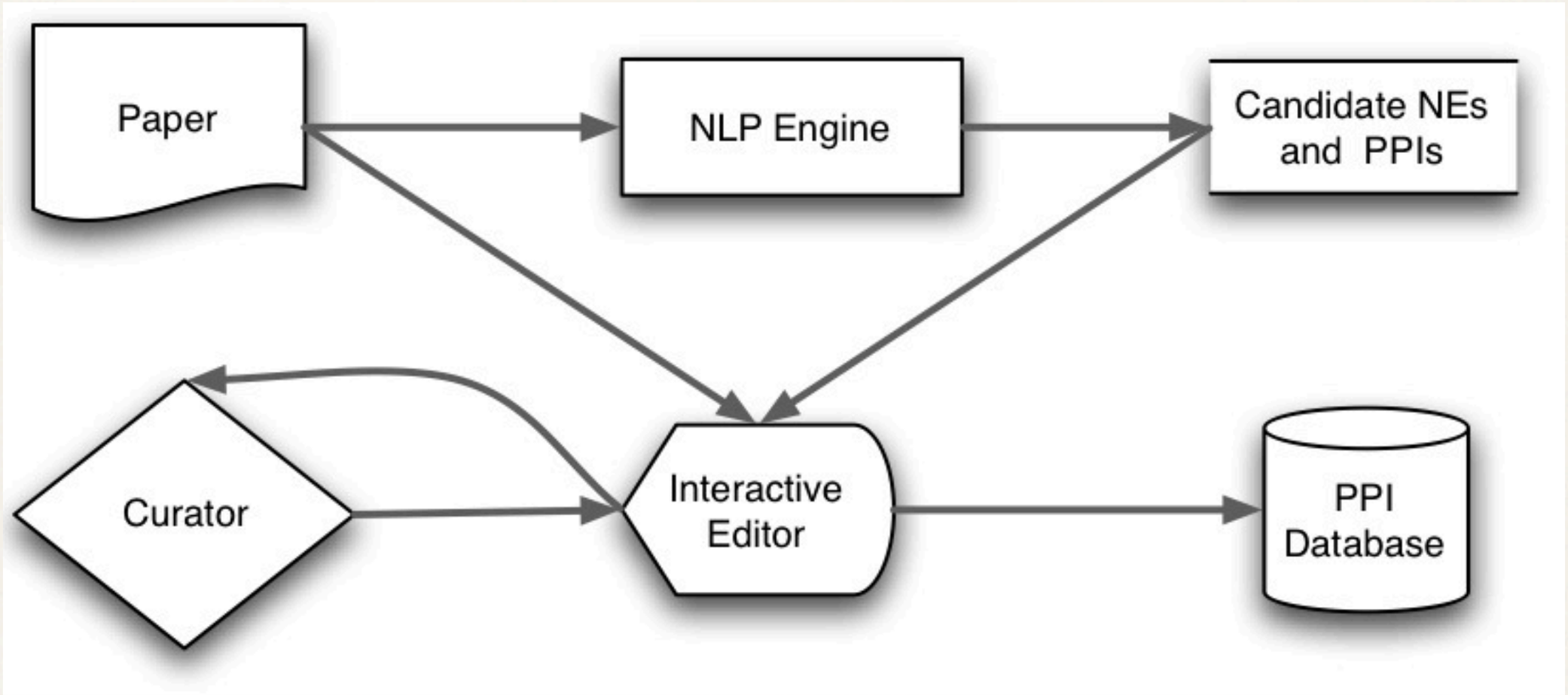
Curation Scenario
- General -

- Goal: Curators should identify protein-protein interactions (PPIs)
- Initial step: Providing set of matching papers
- Middle step: Filtering papers into candidates
- Basic Assumption: Information Extraction (IE) techniques are likely effective in identifying entities and relations
  - More specific: NLP can propose candidate PPIs

Curation Scenario
- General -

- Goal: Curators should identify protein-protein interactions (PPIs)
- Initial step: Providing set of matching papers
- Middle step: Filtering papers into candidates
- Basic Assumption: Information Extraction (IE) techniques are likely effective in identifying entities and relations
  - More specific: NLP can propose candidate PPIs

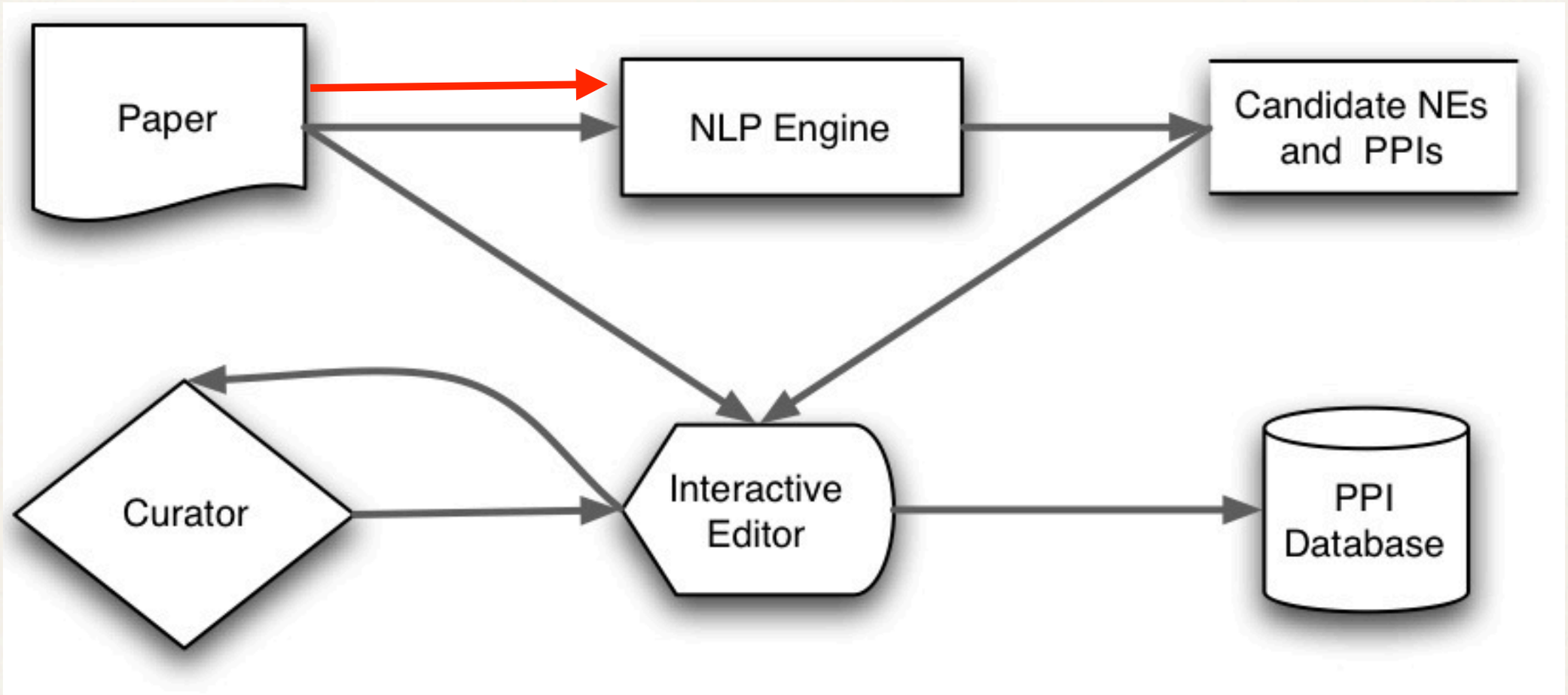
Curation Scenario
- Concrete -



Information Flow in the Curation Process

Source: Alex et al. (p. 558, 2008)

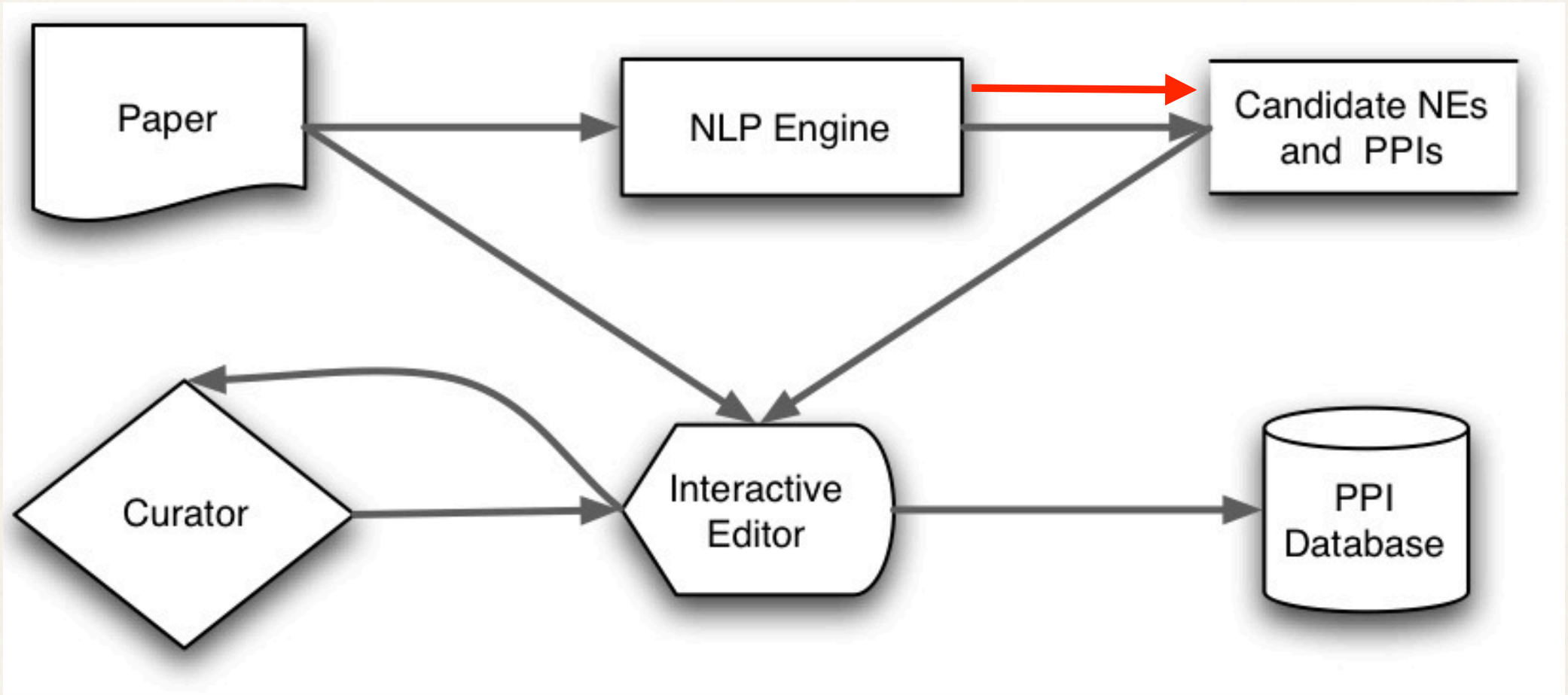
Curation Scenario
- Concrete -



Information Flow in the Curation Process

Source: Alex et al. (p. 558, 2008)

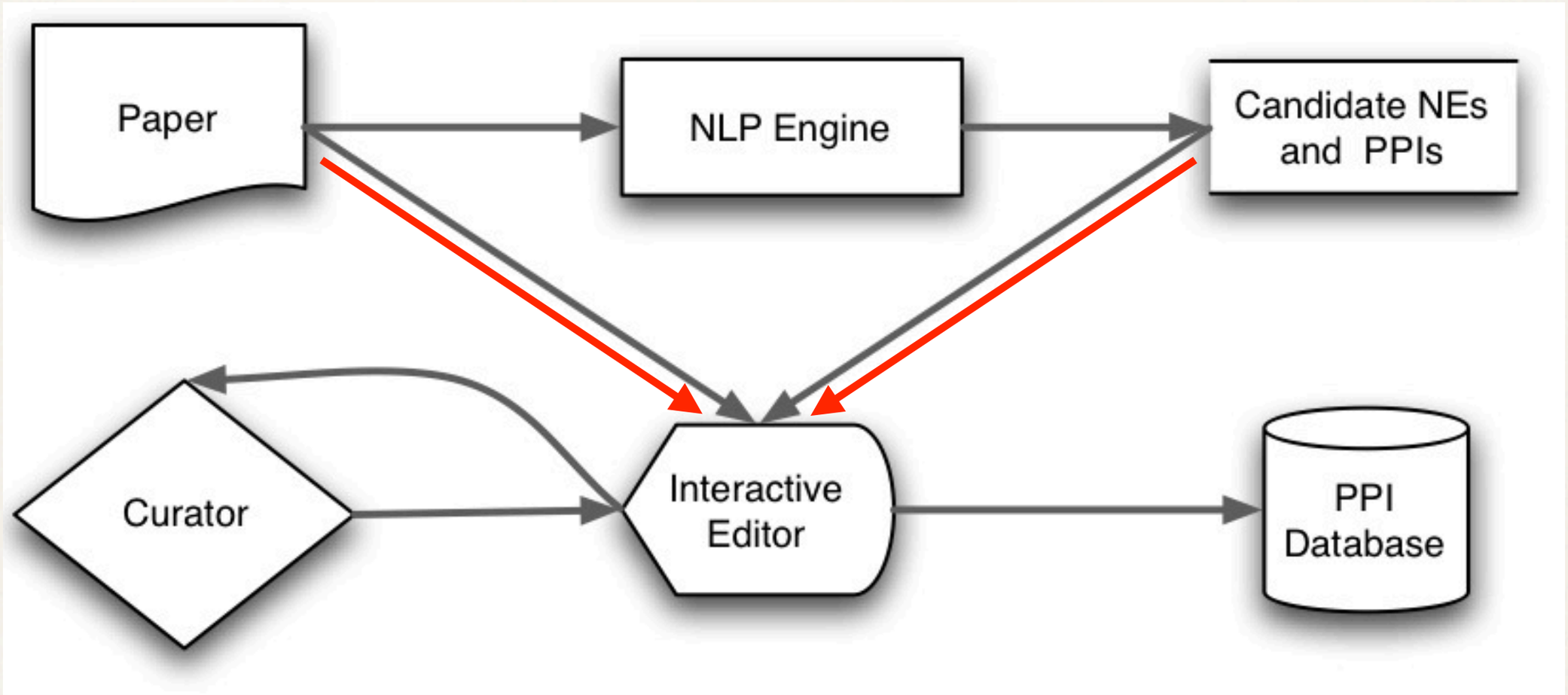
Curation Scenario
- Concrete -



Information Flow in the Curation Process

Source: Alex et al. (p. 558, 2008)

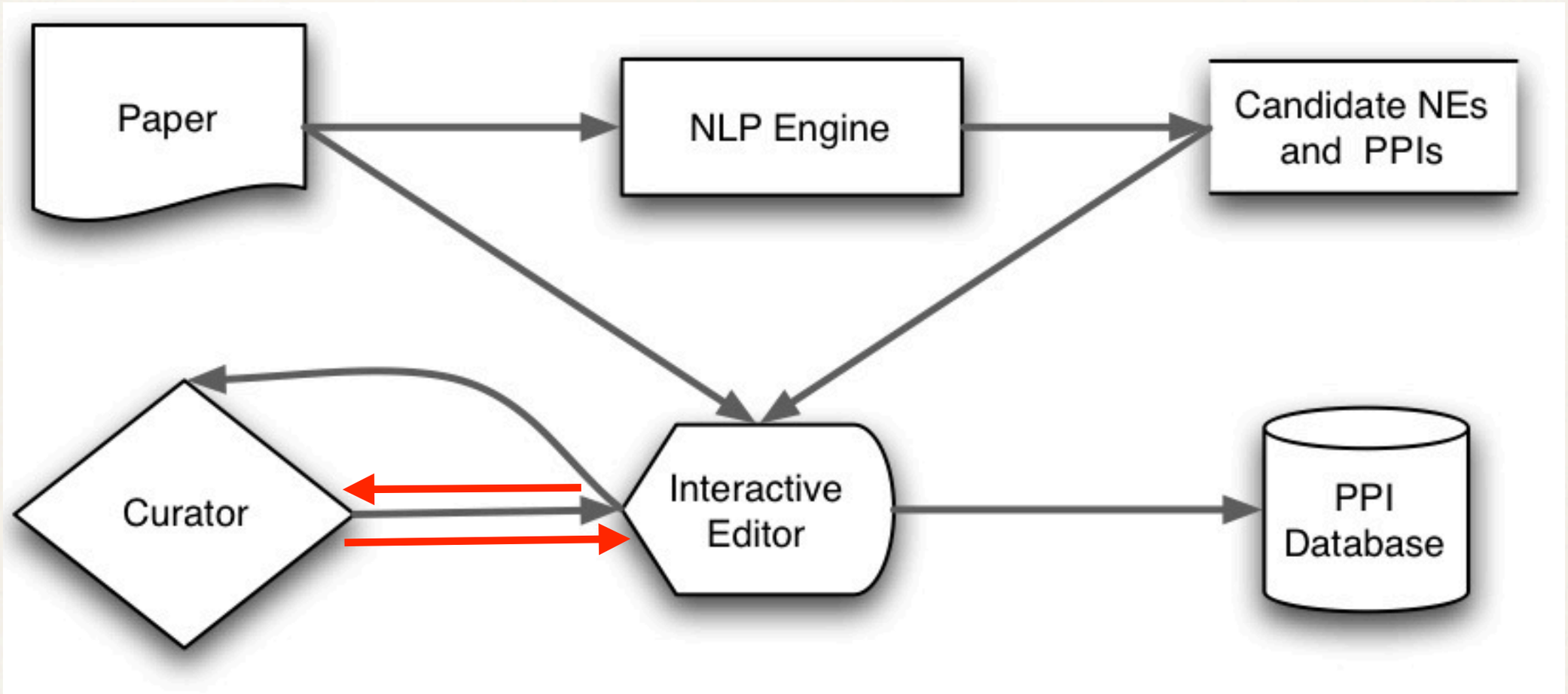
Curation Scenario
- Concrete -



Information Flow in the Curation Process

Source: Alex et al. (p. 558, 2008)

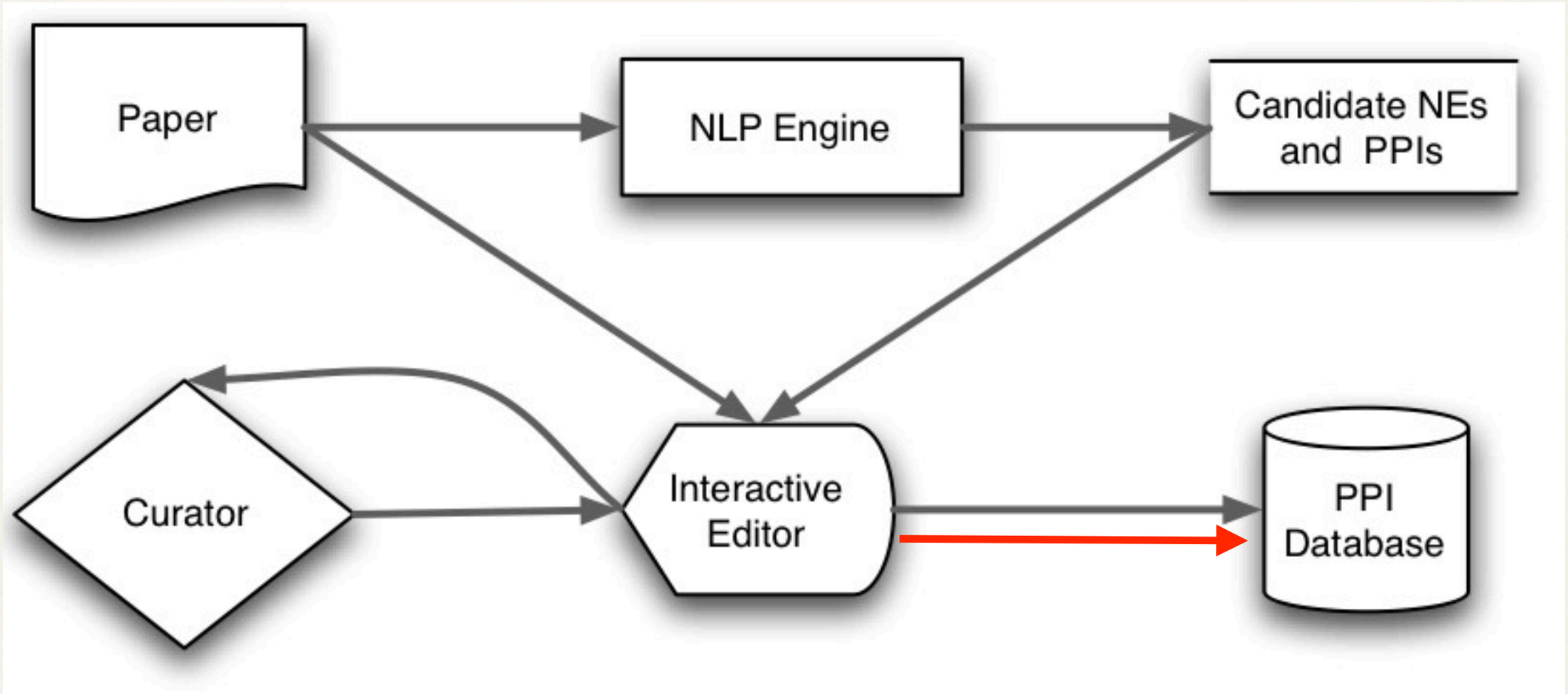
Curation Scenario
- Concrete -



Information Flow in the Curation Process

Source: Alex et al. (p. 558, 2008)

Curation Scenario
- Concrete -



Information Flow in the Curation Process

Source: Alex et al. (p. 558, 2008)

NLP Engine
- Main Components -

Concrete Subtasks

1. Exists protein‘s name in sentence?
2. Which protein do they name?
3. If two proteins are named, are they in interaction?

NLP-Components

1. Named Entity Recognition
2. Term Identification
3. Relation Extraction

NLP Engine
- Creation details -

- How should the interface design look like?

NLP Engine
- Creation

For example:

To decide which species is associated with which protein should be quite simple for an expert but not necessarily for the software.

- How should the
- How should the labour be divided between human and the software?

NLP Engine
- Creation details -

- How should the interface
- How should the labour b nd the software?
- Which functional characteristics of the NLP engine would be optimal?

For example:

Should recall or precision be improved?

NLP Engine
- Creation details -

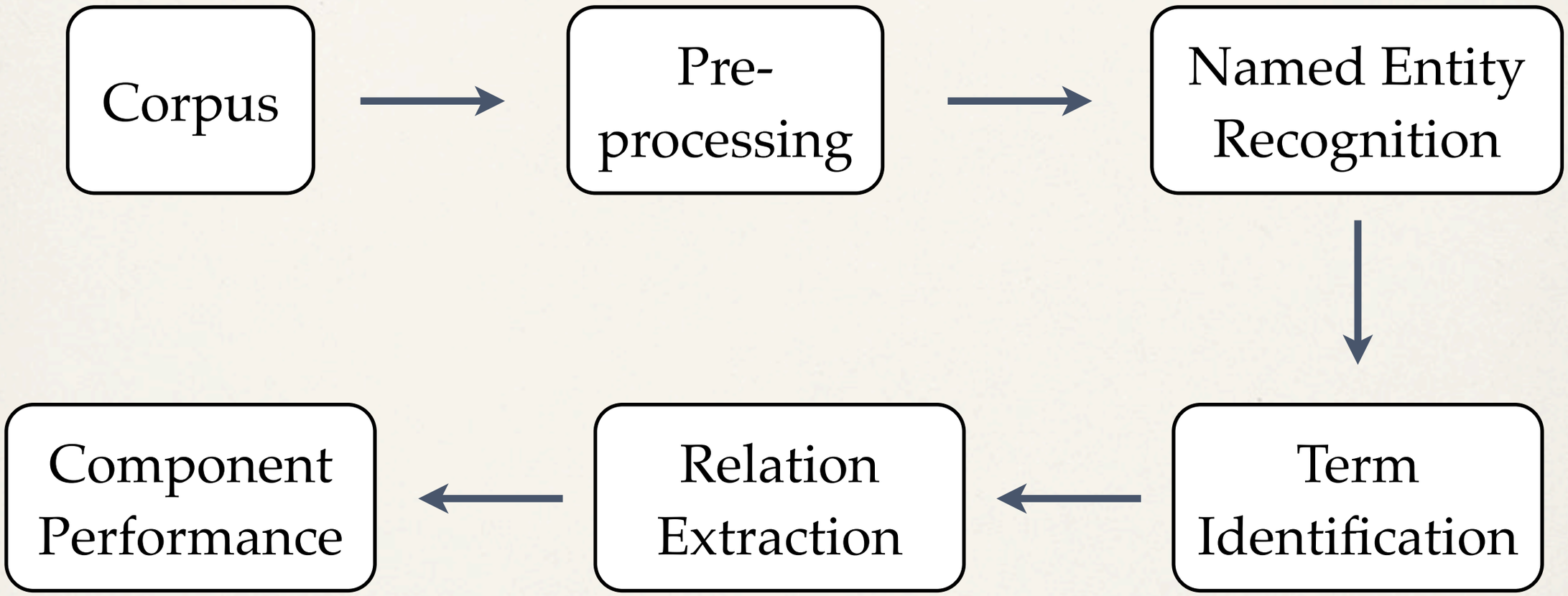
- How should the interface design look like?
- How should the labour be divided between human and the software?
- Which functional characteristics of the NLP engine would be optimal?

The focus will be on the third question.

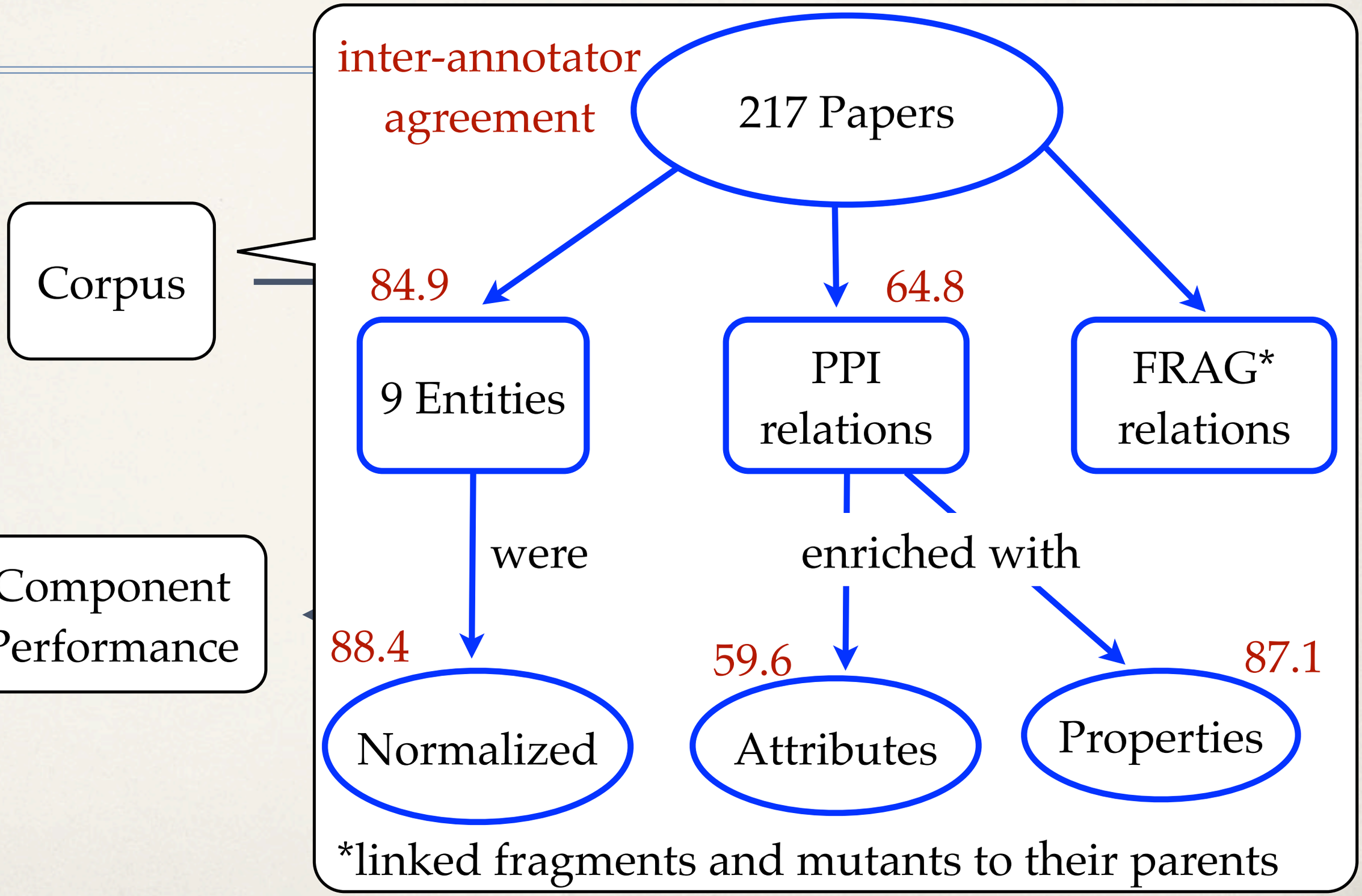
Outline

- Introduction
- Related Work
- Assisted Curation
- Text Mining Pipeline
- Curation Experiments
- Discussion and Conclusion
- References

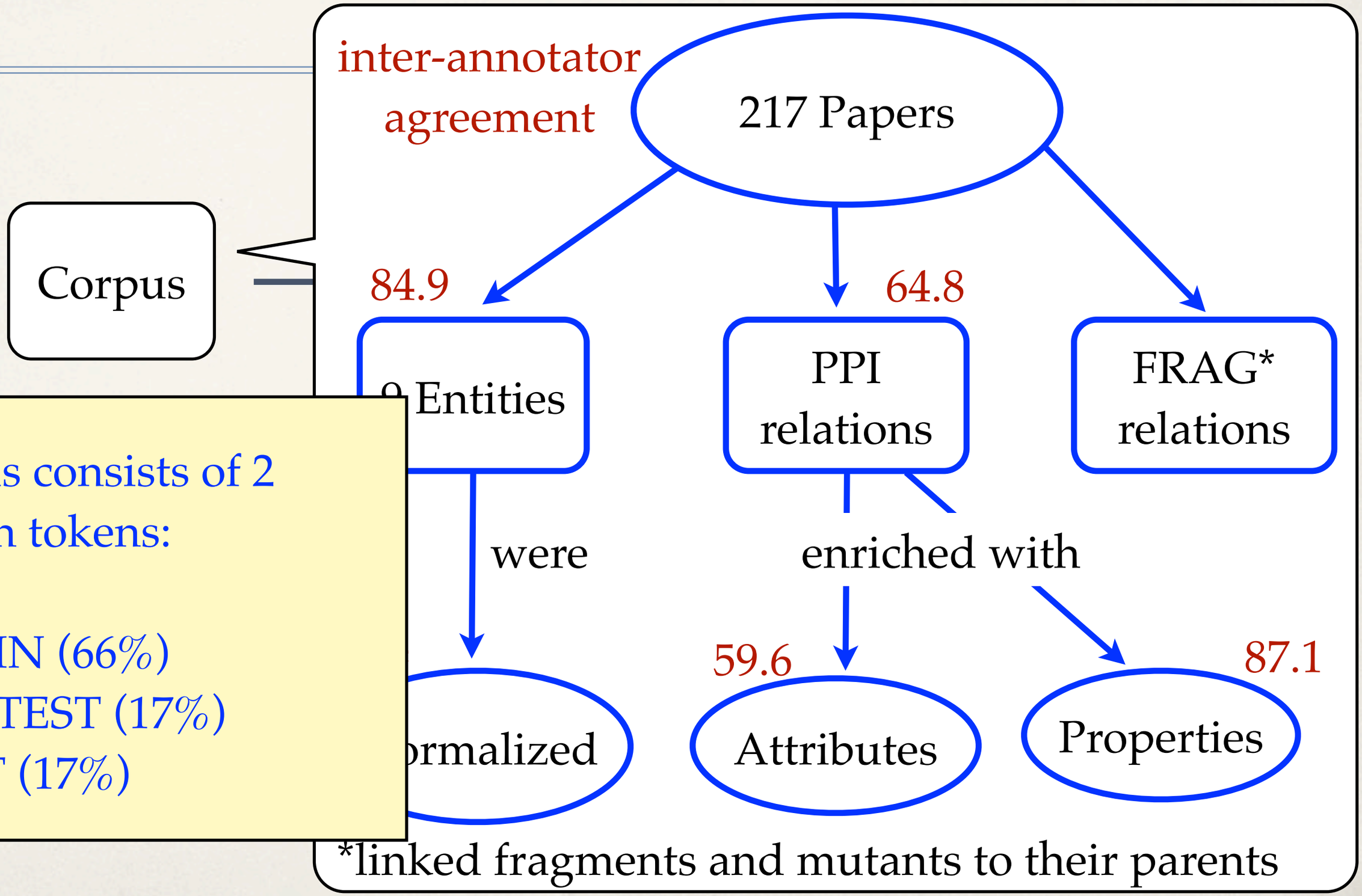
Pipeline-Components



Pipeline-Components



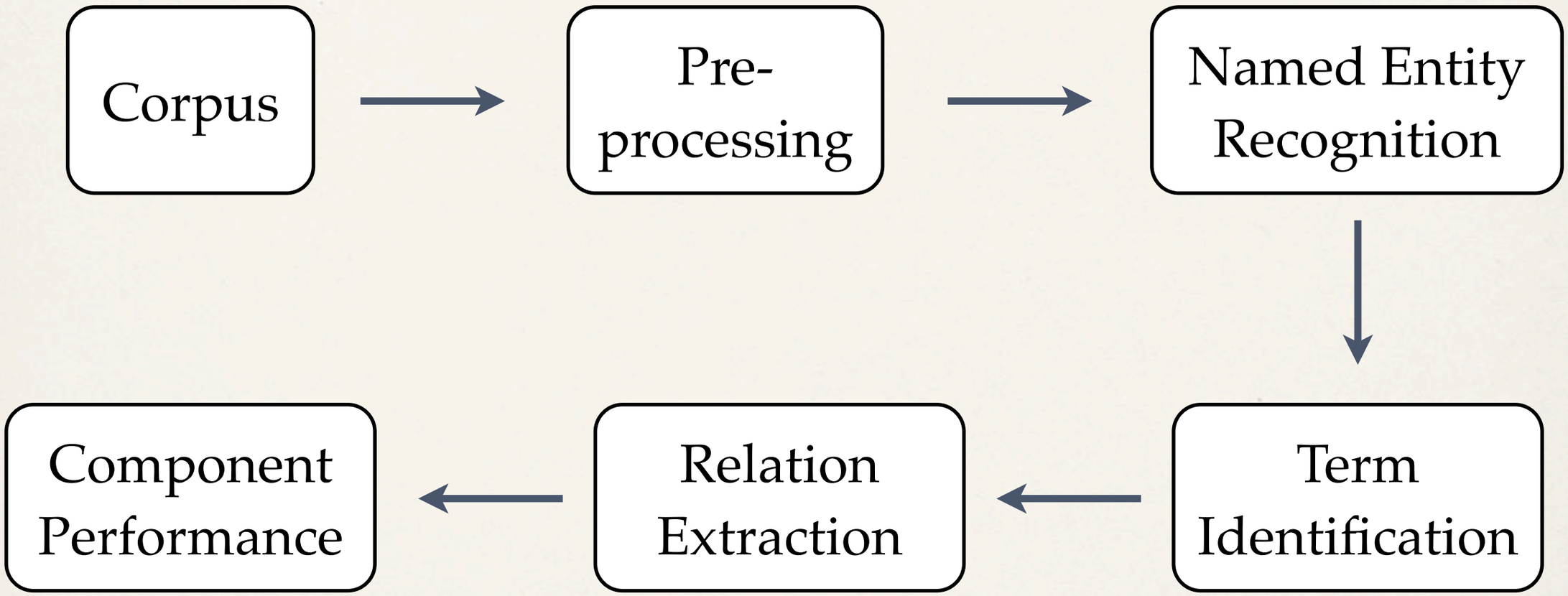
Pipeline-Components



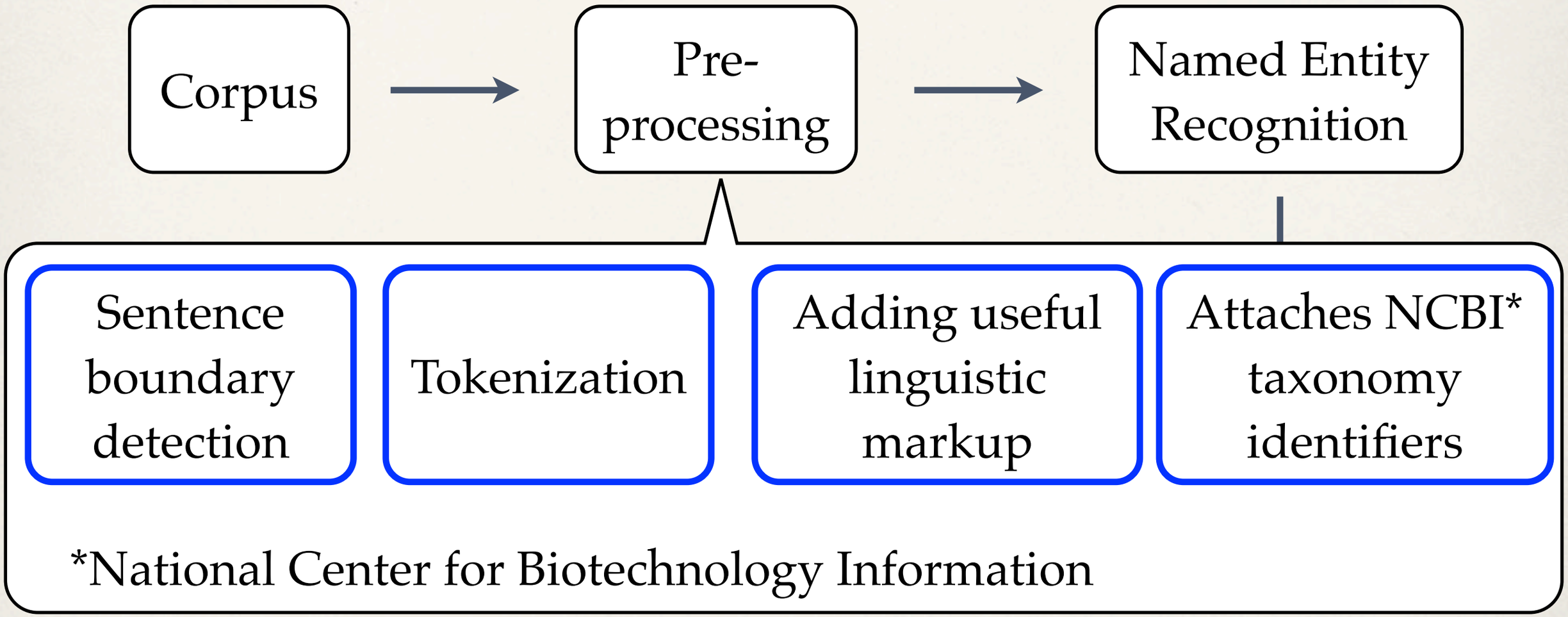
Corpus consists of 2 million tokens:

- TRAIN (66%)
- DEVTEST (17%)
- TEST (17%)

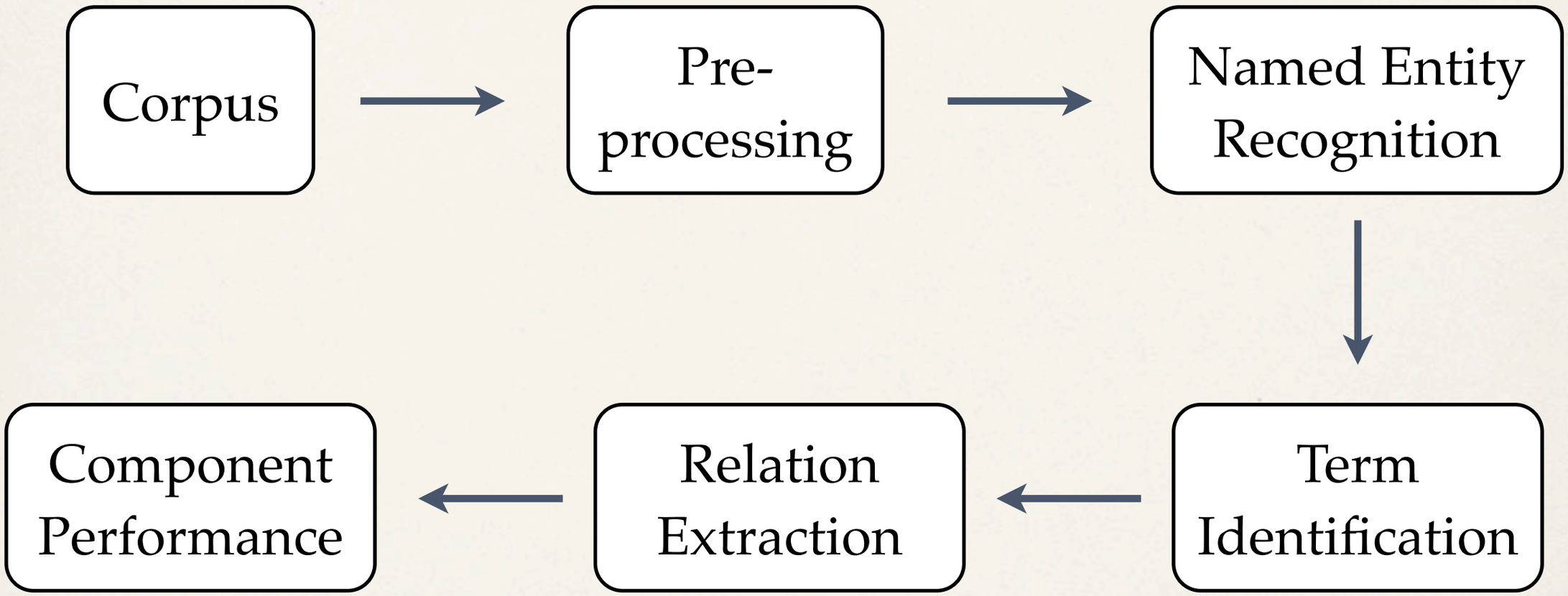
Pipeline-Components



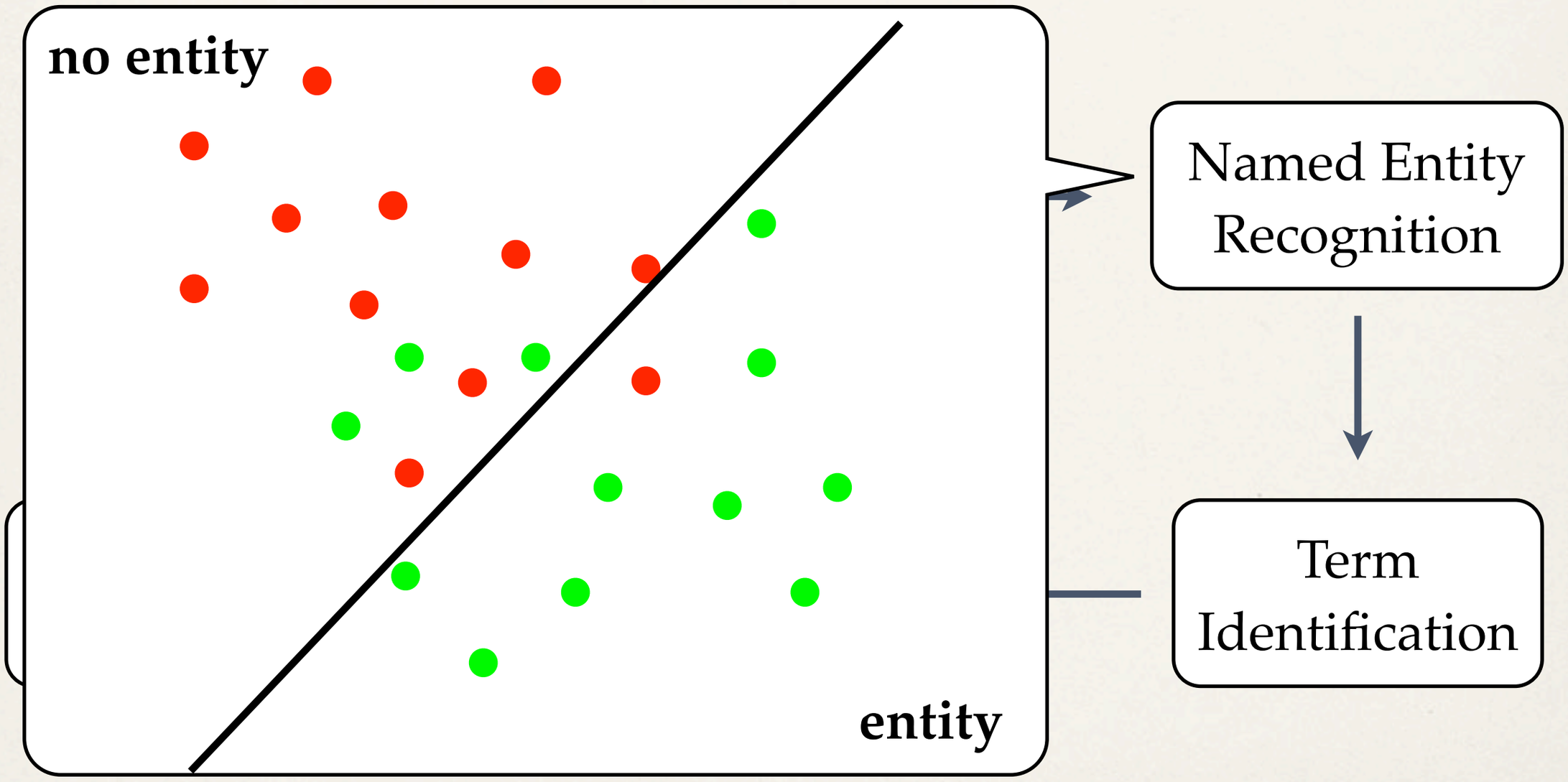
Pipeline-Components



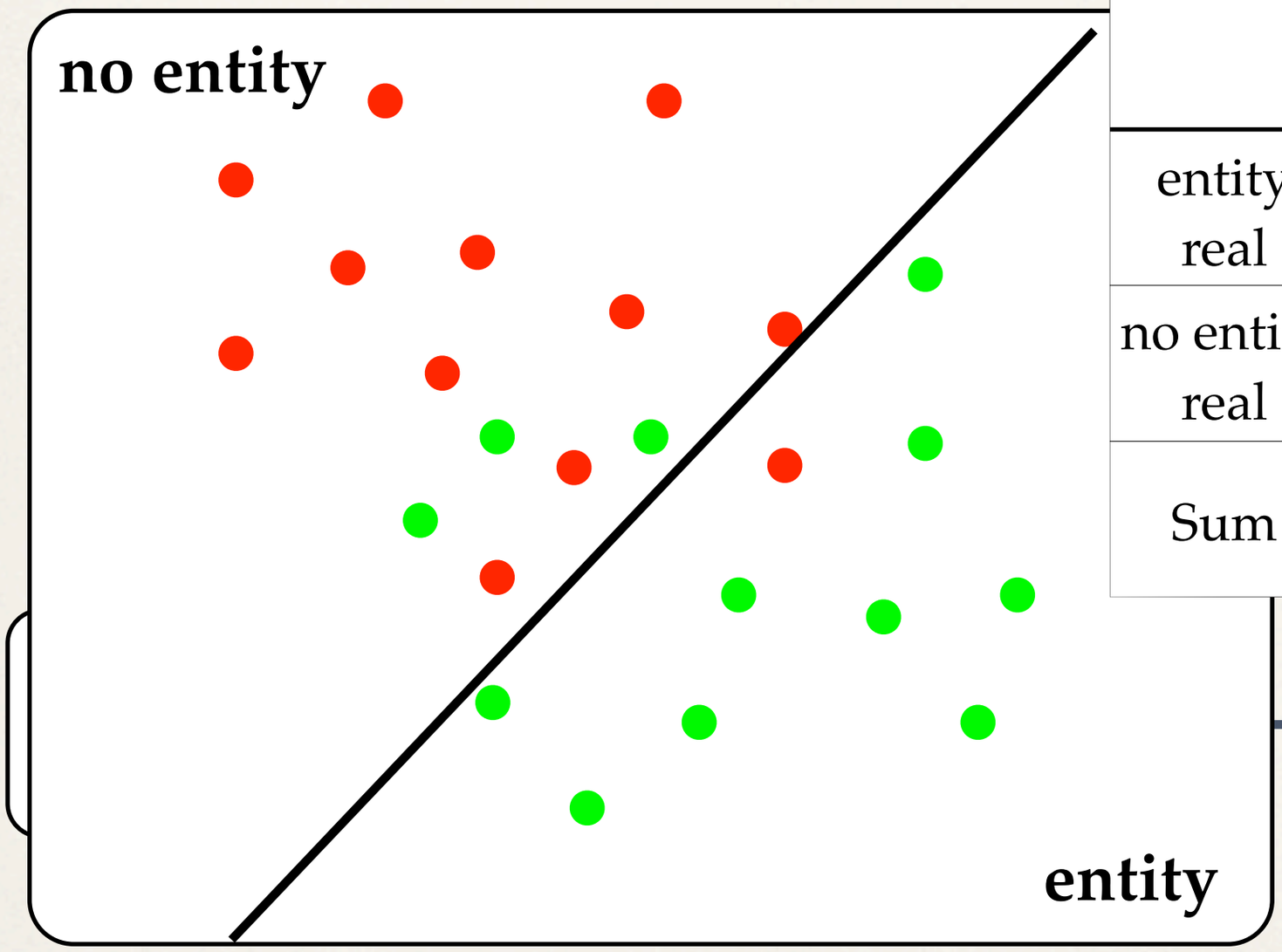
Pipeline-Components



Pipeline-Components



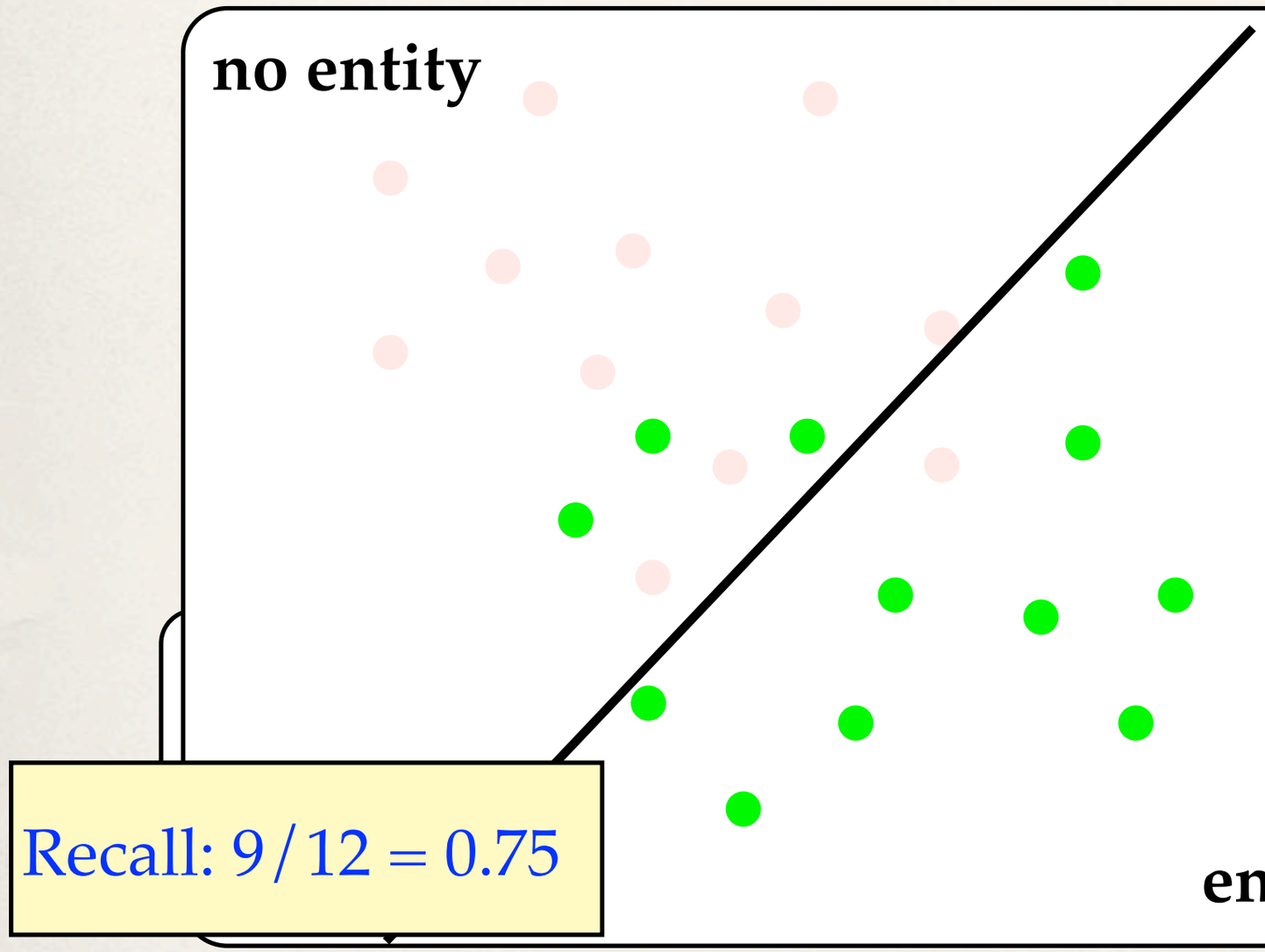
Pipeline-Components



	entity pred	no entity pred	Sum
entity real	9	3	12
no entity real	1	11	12
Sum	10	14	24

Term Identification

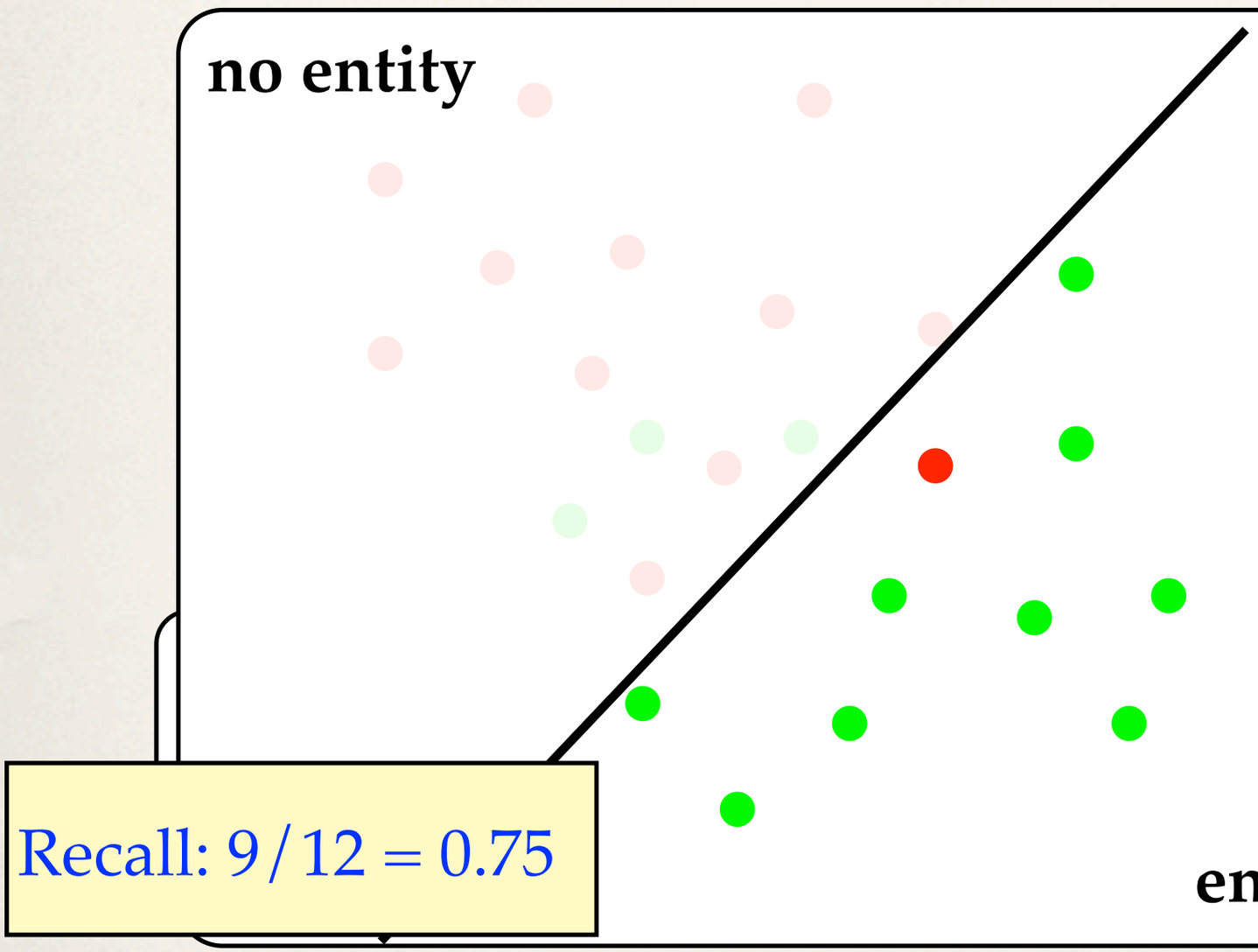
Pipeline-Components



	entity pred	no entity pred	Sum
entity real	9	3	12
no entity real	1	11	12
Sum	10	14	24

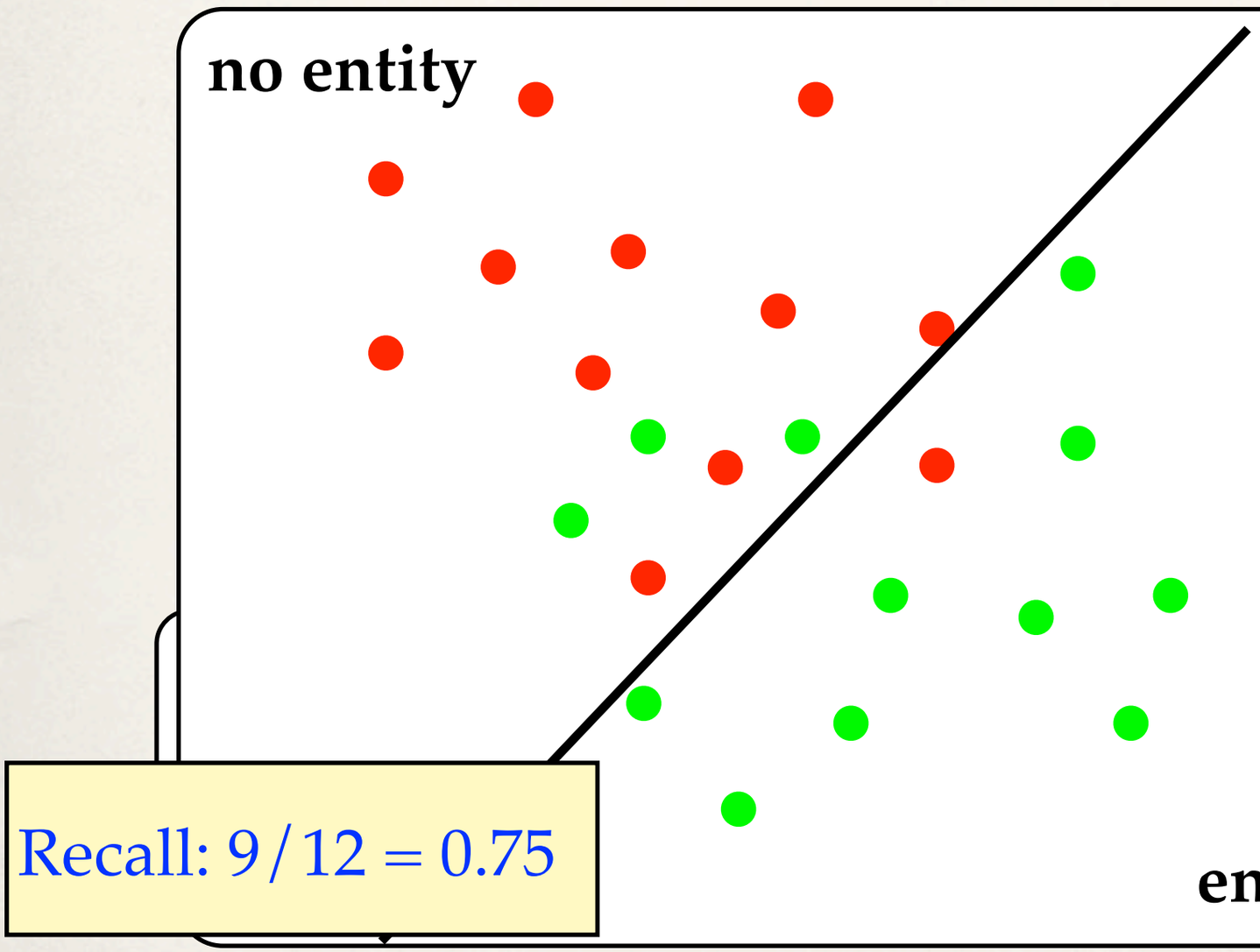
Term Identification

Pipeline-Components



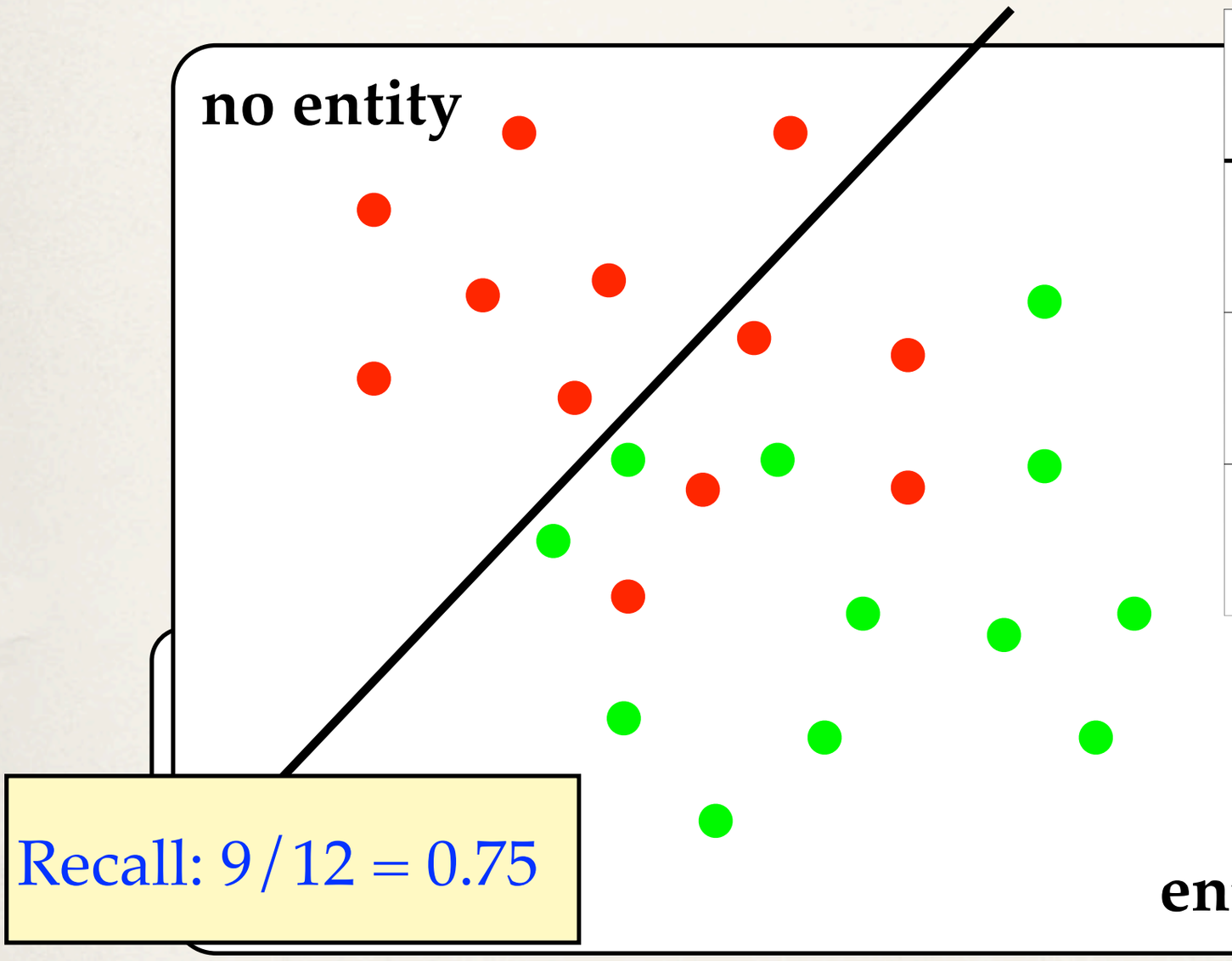
	entity pred	no entity pred	Sum
entity real	9	3	12
no entity real	1	11	12
Sum	10	14	24

Pipeline-Components



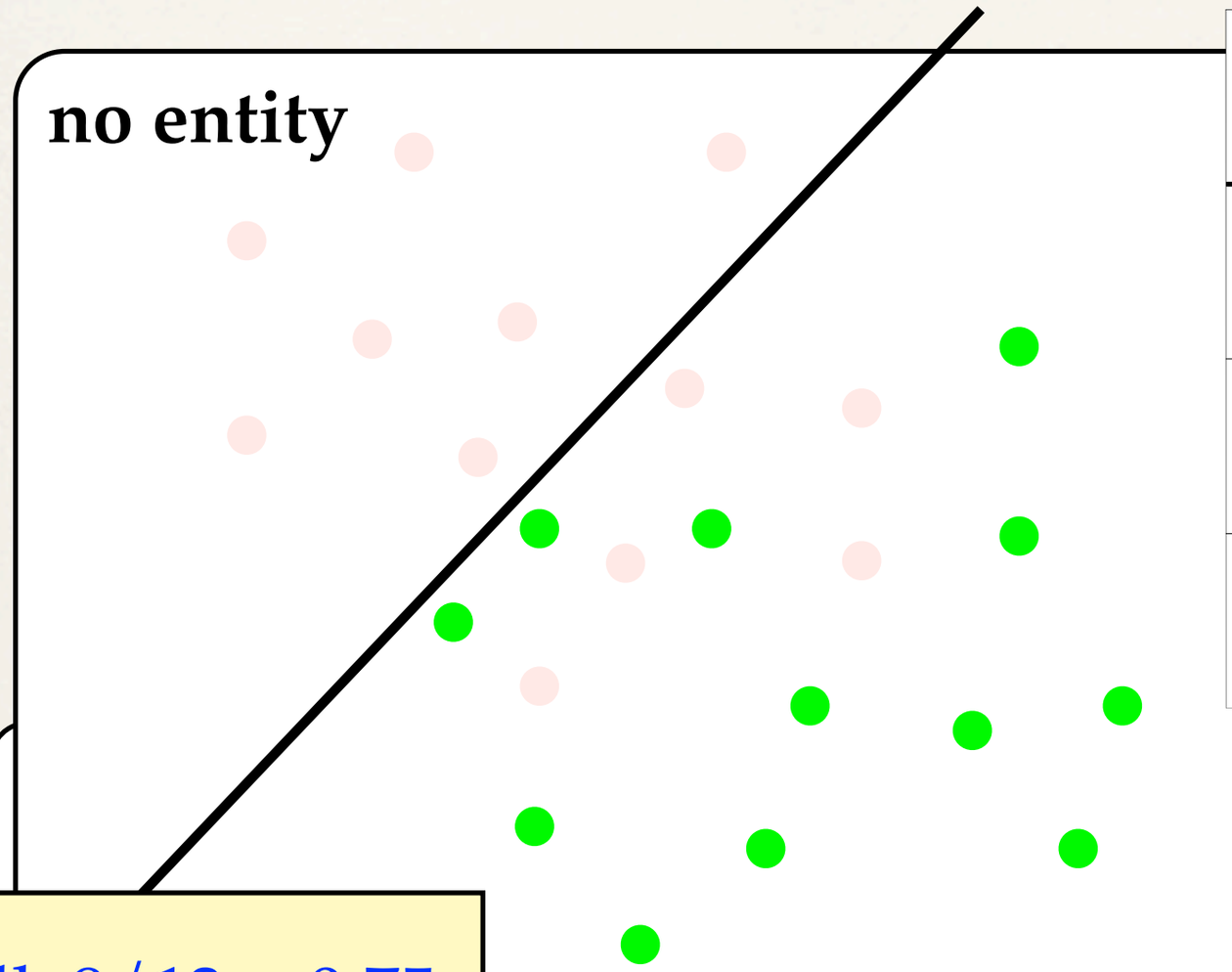
	entity pred	no entity pred	Sum
entity real	9	3	12
no entity real	1	11	12
Sum	10	14	24

Pipeline-Components



	entity pred	no entity pred	Sum
entity real	12	0	12
no entity real	5	7	12
Sum	17	7	24

Pipeline-Components



	entity pred	no entity pred	Sum
entity real	12	0	12
no entity real	5	7	12
Sum	17	7	24

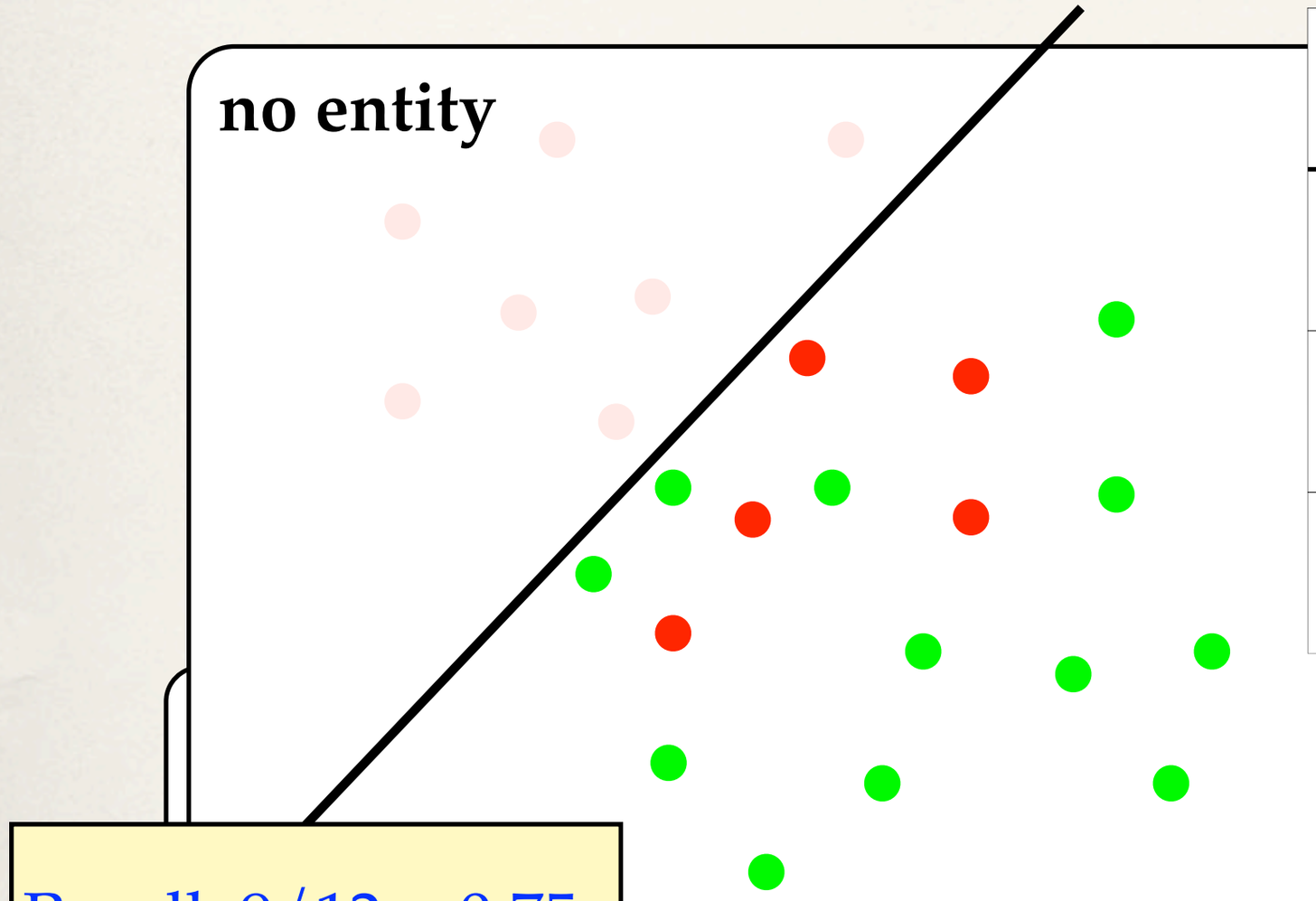
Term

Recall: 9/12 = 0.75

Precision: 9/10 = 0.9

Recall: 12/12 = 1

Pipeline-Components



	entity pred	no entity pred	Sum
entity real	12	0	12
no entity real	5	7	12
Sum	17	7	24

Term

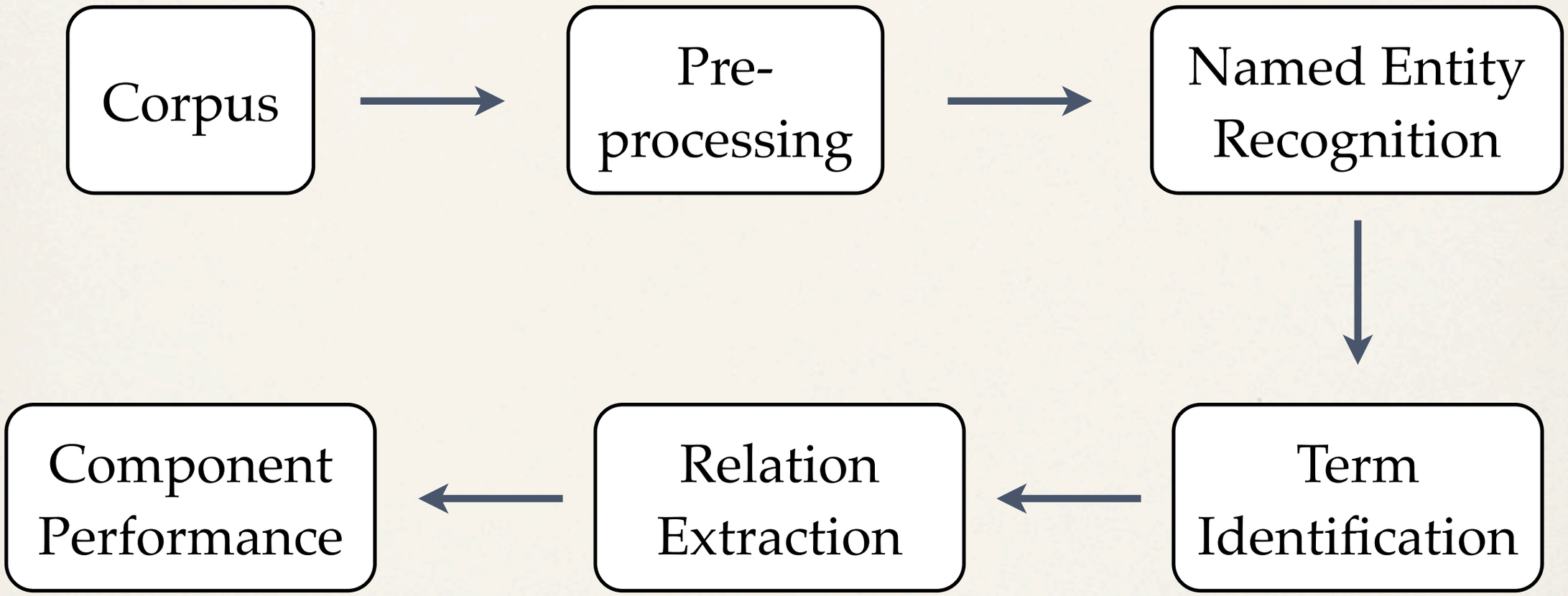
Recall: 9/12 = 0.75

Precision: 9/10 = 0.9

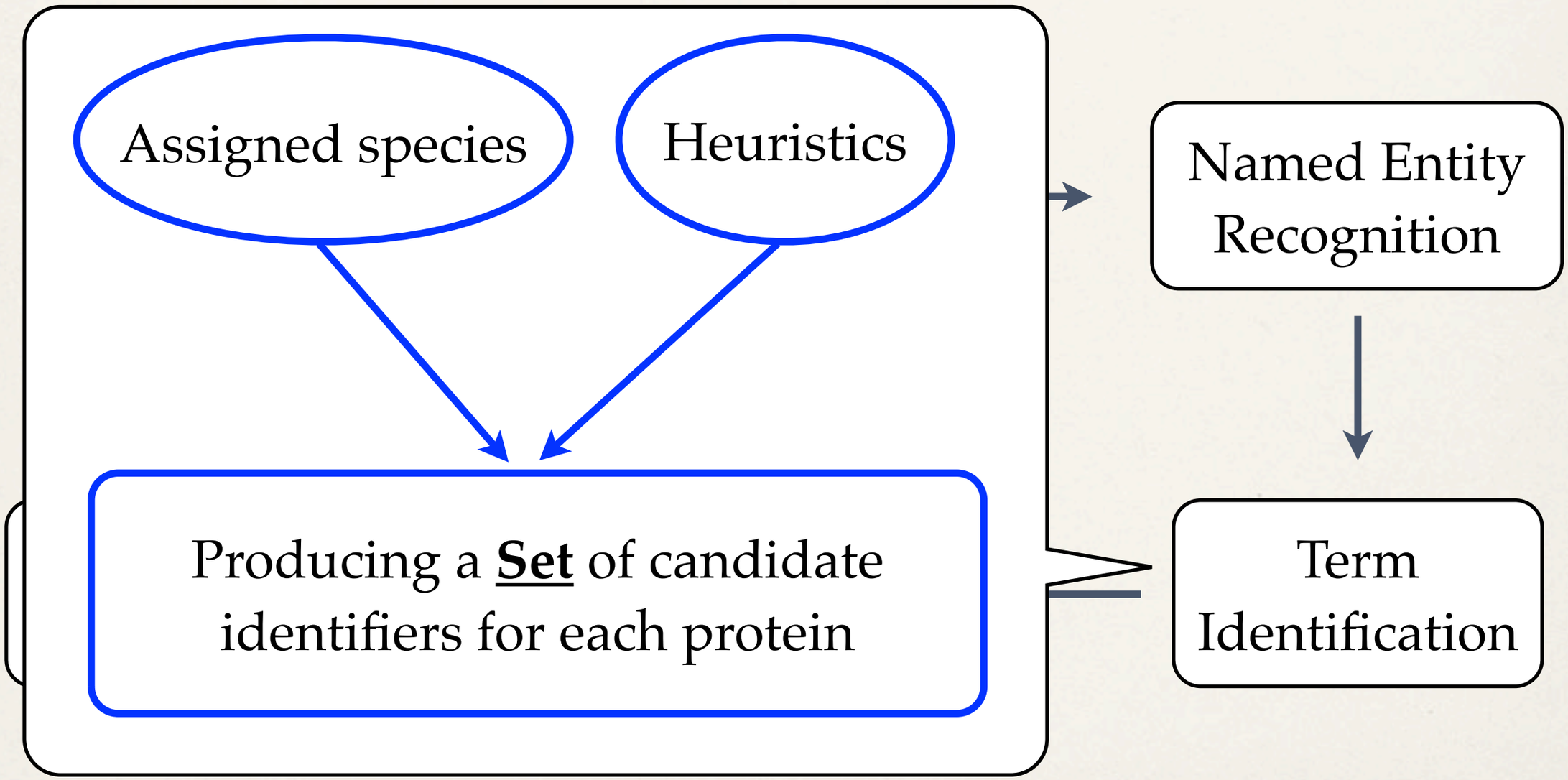
Recall: 12/12 = 1

Precision: 12/17 = 0.71

Pipeline-Components

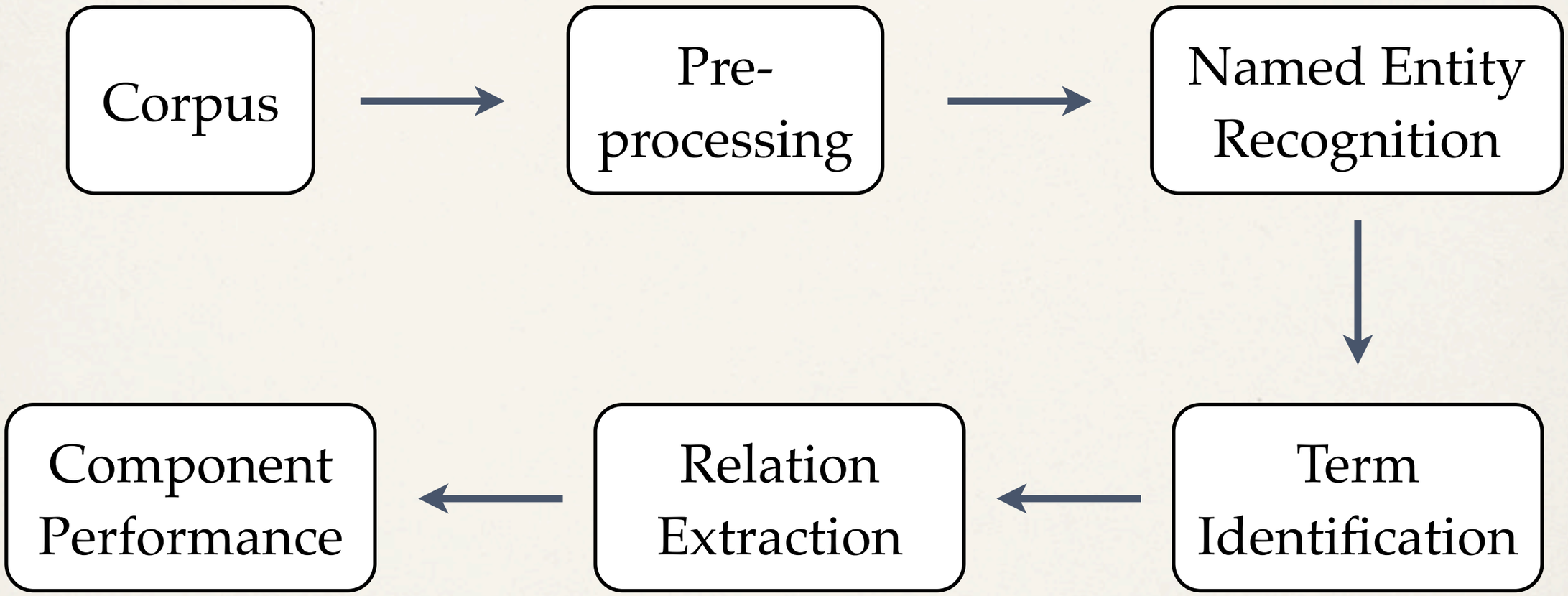


Pipeline-Components

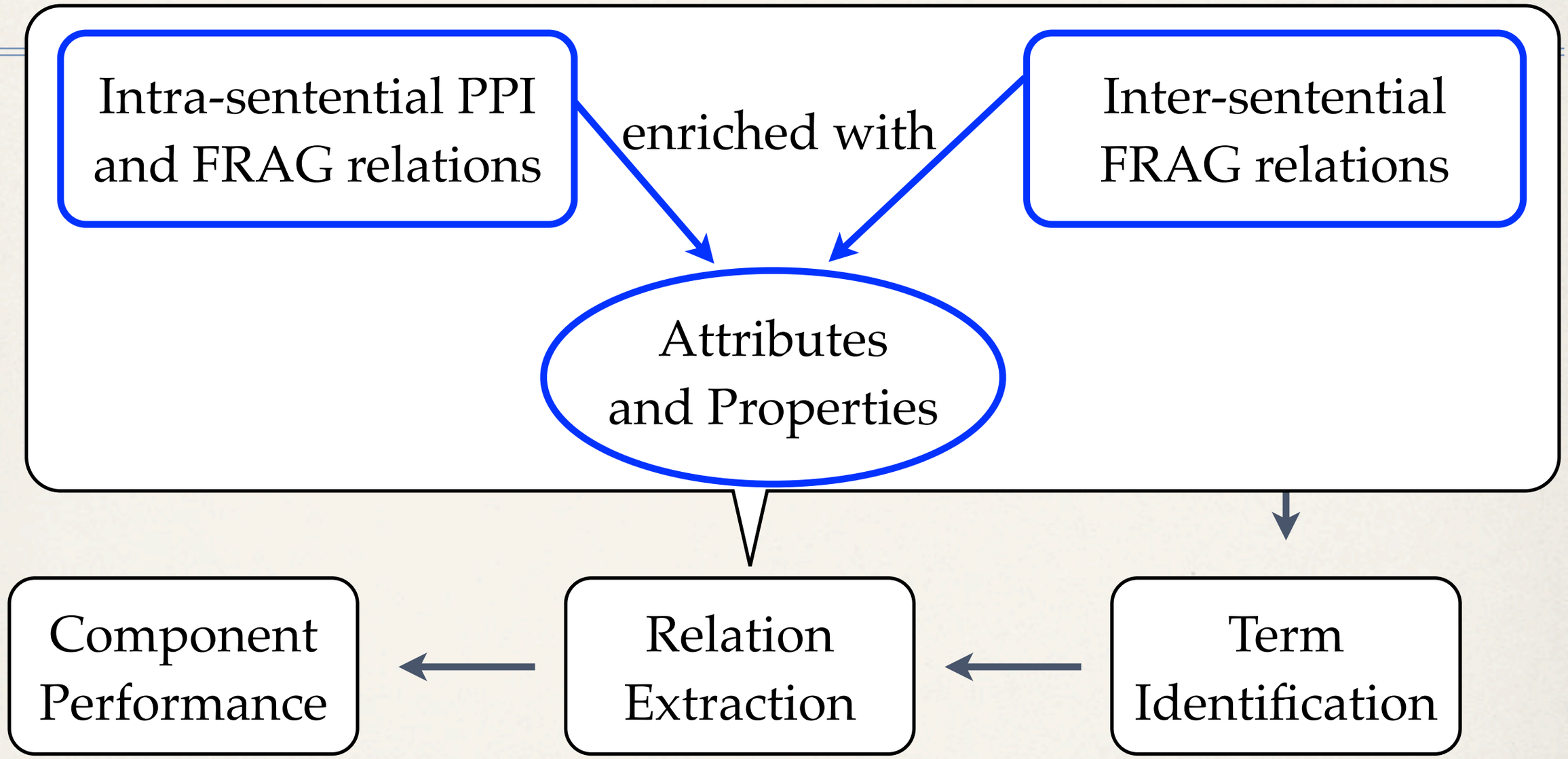


Bag accuracy as evaluation metric

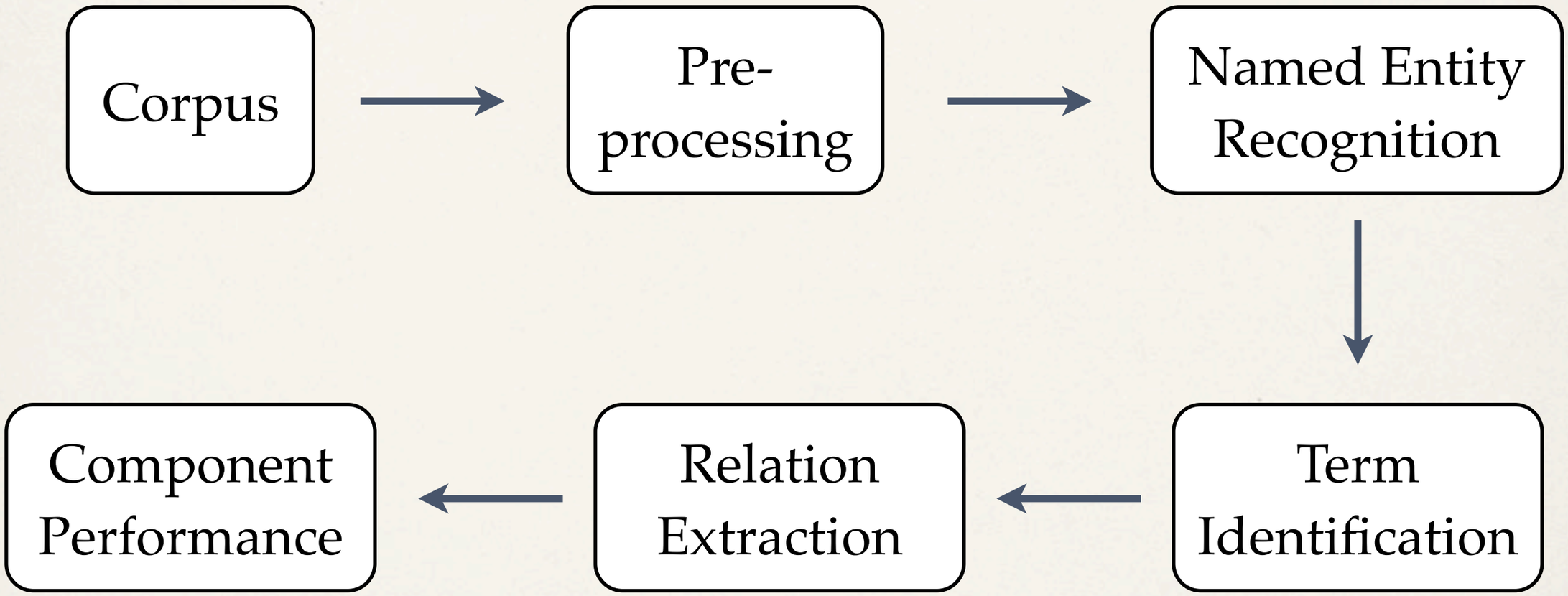
Pipeline-Components



Pipeline-Components



Pipeline-Components



Pipeline-Components

Component	TP	FP	FN	Precision	Recall	F1
NER (micro-average)	19,925	5,964	7,755	76.96	71.98	74.39
RE (PPI)	1,208	1,173	1,080	50.73	52.80	51.75
RE (FRAG)	1,699	963	1,466	63.82	53.68	58.31
RE (properties micro-average)	3,041	567	579	84.28	84.01	84.14
RE (attributes micro-average)	483	822	327	37.01	59.63	45.67

Component	TP	FP	FN	Precision	Recall	Bag Acc.
TI (micro-average)	9,078	91,396	2,843	9.04	76.15	76.15

Component Performance

F1 = 2 * (precision * recall) / (precision + recall)

Extraction

Identification

DEVTEST and trained on TRAIN

Pipeline-Components

Component	TP	FP	FN	Precis	F1
NER (micro-average)	19,925	5,964	7,755	76.9	74.39
RE (PPI)	1,208	1,173	1,080	50.7	51.75
RE (FRAG)	1,699	963	1,466	63.8	58.31
RE (properties micro-average)	3,041	567	579	84.2	84.14
RE (attributes micro-average)	483	822	327	37.0	45.67

Component	TP	FP	FN	Precision	Recall	Bag Acc.
TI (micro-average)	9,078	91,396	2,843	9.04	76.15	76.15

inter-annotator agreement:

84.9/88.4

64.8

87.1

59.6

Component Performance

F1 = 2 * (precision * recall) / (precision + recall)

Extraction

Identification

DEVTEST and trained on TRAIN

Outline

- Introduction
- Related Work
- Assisted Curation
- Text Mining Pipeline
- Curation Experiments
- Discussion and Conclusion
- References

Experiment 1: Manual vs. Assisted Curation

- 4 curators
- 4 papers
- 3 conditions:
  - Manual: **without** assistance
  - GSA-assisted: with integrated **gold standard** annotation
  - NLP-assisted: with integrated **NLP pipeline** output

Experiment 1: Results

		Time per record	
Condition	Records	Average	StDev
MANUAL	121	312s	327s
GSA	170	205s	52s
NLP	141	243s	36s

Total number of records and average curation speed per record

Statement	GSA	NLP
NLP was helpful in curating this documents	2.75	3.25
NLP speeded up the curation of this paper	3.75	3.75
NE annotations were useful for curation	2.50	3.00
Normalizations of NEs were useful for curation	2.75	2.75
PPIs were useful for curation	3.50	3.25

Scores range from (1) for „strongly agree“ to (5) for „strongly disagree“

Experiment 1: Results

		Time per record	
Condition	Records	Average	StDev
MANUAL	121	312s	327s
GSA	170	205s	52s
NLP	141	243s	36s

Total number of records and average curation speed per record

Statement	GSA		NLP
NLP was helpful in curating this documents	2.75	<	3.25
NLP speeded up the curation of this paper	3.75	=	3.75
NE annotations were useful for curation	2.50	<	3.00
Normalizations of NEs were useful for curation	2.75	=	2.75
PPIs were useful for curation	3.50	>	3.25

Scores range from (1) for „strongly agree“ to (5) for „strongly disagree“

Experiment 2: NLP Consistency

- 1 curator
- 10 papers
- 2 conditions:
  - Consistency 1: **all** recognized named entities (NEs) were propagated (5 papers)
  - Consistency 2: **only the most frequent** recognized NEs were propagated (5 papers)

Experiment 2: Results I

	Time per record	
Condition	Average	StDev
CONSISTENCY1	128s	43s
CONSISTENCY2	92s	22s

Total number of records and average curation speed per record

Experiment 2: Results II

Statement	CONSISTENCY1	CONSISTENCY2
Questionnaire 1		
NLP output was helpful for curation	1.6	2.6
NLP output speeded up curation	1.8	3.2
NEs were useful for curation	1.4	4.0
Normalizations of NEs were useful for curation	3.2	4.0
PPIs were useful for curation	3.6	4.2
Questionnaire 2		
A was more useful for curation than B would have been	2.6	4.0
A speeded up the curation process more than B would have	3.0	4.0
A appeared more accurate than B	2.6	4.2
A missed important information compared to B	4.4	1.8
A contained too much information compared to B	3.6	4.6

Scores range from (1) for „strongly agree“ to (5) for „strongly disagree“

A: consistent NLP output (Consistency 1/2)

B: baseline NLP

Experiment 2: Results II

Statement	CONSISTENCY1	CONSISTENCY2
Questionnaire 1		
NLP output was helpful for curation	1.6	2.6
NLP output speeded up curation	1.8	3.2
NEs were useful for curation	1.4	4.0
Normalizations of NEs were useful for curation	3.2	4.0
PPIs were useful for curation	3.6	4.2
Questionnaire 2		
A was more useful for curation than B would have been	2.6	4.0
A speeded up the curation process more than B would have	3.0	4.0
A appeared more accurate than B	2.6	4.2
A missed important information compared to B	4.4	1.8
A contained too much information compared to B	3.6	4.6

Scores range from (1) for „strongly agree“ to (5) for „strongly disagree“

A: consistent NLP output (Consistency 1/2)

B: baseline NLP

Experiment 3: Optimizing for Precision or Recall

- 1 curator
- 10 papers
- 3 conditions:
  - High R: NLP output with high **recall** (5 papers)
  - High P: NLP output with high **precision** (5 papers)
  - High F1: NLP output with high **F1-score** (subsequent all papers; only viewing)

$$F1 = 2 * (precision * recall) / (precision + recall)$$

Experiment 3
Results I

Setting	TP	FP	FN	P	R	F1
High F1	20,091	6,085	7,589	76.75	72.58	**74.61**
High P	11,836	1,511	15,844	**88.68**	42.76	57.70
High R	21,880	20,653	5,800	51.44	**79.05**	62.32

Comparison between High F1, High P and High R

TP: true positive

FP: false positive

FN: false negative

Experiment 3
Results II

Statement	HighP NER	HighR NER
Questionnaire 1		
NLP output was helpful for curation	3.0	2.2
NLP output speeded up curation	3.4	2.4
NEs were useful for curation	3.0	2.0
PPIs were useful for curation	3.2	2.5
Questionnaire 2		
A was more useful for curation than B would have been	4.2	2.6
A speeded up the curation process more than B would have	4.2	3.0
A appeared more accurate than B	4.4	2.8
A missed important information compared to B	1.4	3.2
A contained too much information compared to B	4.8	3.8

Scores range from (1) for „strongly agree“ to (5) for „strongly disagree“

A: High P/High R

B: High F1

Experiment 3
Results II

Statement	HighP NER	HighR NER
Questionnaire 1		
NLP output was helpful for curation	3.0	2.2
NLP output speeded up curation	3.4	2.4
NEs were useful for curation	3.0	2.0
PPIs were useful for curation	3.2	2.5
Questionnaire 2		
A was more useful for curation than B would have been	4.2	2.6
A speeded up the curation process more than B would have	4.2	3.0
A appeared more accurate than B	4.4	2.8
A missed important information compared to B	1.4	3.2
A contained too much information compared to B	4.8	3.8

Scores range from (1) for „strongly agree“ to (5) for „strongly disagree“

A: High P/High R

B: High F1

Outline

- Introduction
- Related Work
- Assisted Curation
- Text Mining Pipeline
- Curation Experiments
- Discussion and Conclusion
- References

Discussion I

- <u>Experiment 1:</u>
  - Maximum time reduction of 1/3 if NLP output is perfectly accurate
  - NLP assistance leads to more records (but the validity has to be proven)
  - In the questionnaire all condition are quite equal

Discussion II

- Experiment 2:
  - Curator prefers consistency with all NEs
    - But: objective metrics suggest that other condition is prefered
- Experiment 3:
  - Curator prefers high recall
    - Must be repeated with other curators (different curation styles)

Conclusion

- Curation time not sufficient measurement for NLP's usefulness
- Closely work with user is necessary
  - Identifying helpful and hindering aspects
- Future work:
  - Further research regarding the merit of high recall and high precision
  - Implementing confidence values of extracted information
  - ... with more curators

Outline

- Introduction
- Related Work
- Assisted Curation
- Text Mining Pipeline
- Curation Experiments
- Discussion and Conclusion
- References

References

- Alex, B., Grover, C., Haddow, B., Kabadjov, M., Klein, E., Matthews, M., Roebuck, S., Tobin, R., Wang, X. (2008). Assisted curation: does text mining really help? In *Pacific Symposium on Biocomputing*. Pacific Symposium on Biocomputing, pp. 556-567.

- Krallinger, M., Leitner, F., & Valencia, A. (2007). Assessment of the second BioCreative PPI task: Automatic extraction of protein-protein interactions. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pp. 41–54, Madrid, Spain.

- Schwikowski, B., Uetz, P., & Fields, S. (2000). A network of protein-protein interactions in yeast. *Nature Biotechnology*, 18, pp. 1257-1261.