



Generic Entity Resolution with Negative Rules

Steven Whang

Hector Garcia-Molina

Stanford University

Omar Benjelloun

Google Inc.

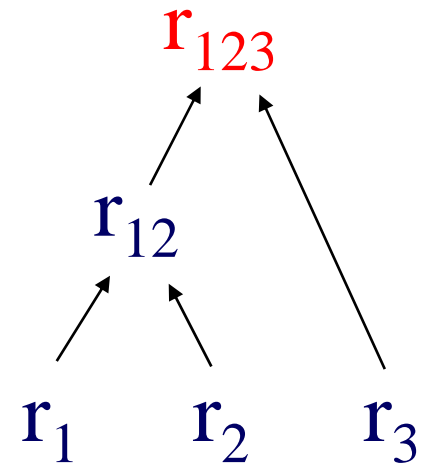
Entity Resolution

	Name	SSN	Gender
r_1	Pat	999-04-1234	
r_2	Patricia		F
r_3	Pat	999-04-1234	M

- $M(r_1, r_2) = T$, merge $\langle r_1, r_2 \rangle = r_{12}$
- $M(r_3, r_{12}) = T$, merge $\langle r_3, r_{12} \rangle = r_{123}$

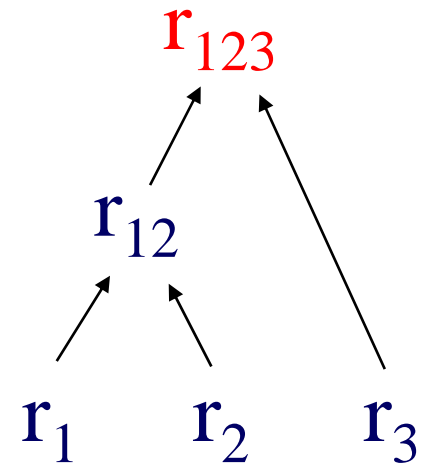
Entity Resolution

	Name	SSN	Gender
r_1	Pat	999-04-1234	
r_2	Patricia		F
r_3	Pat	999-04-1234	M
r_{12}	{Pat, Patricia}	999-04-1234	F
r_{123}	{Pat, Patricia}	999-04-1234	{F, M}



Entity Resolution

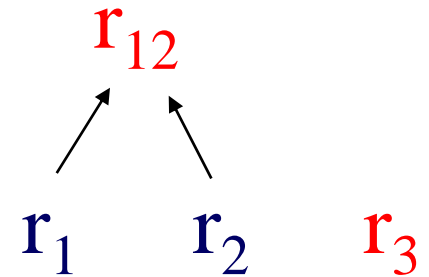
	Name	SSN	Gender
r_1	Pat	999-04-1234	
r_2	Patricia		F
r_3	Pat	999-04-1234	M
r_{12}	{Pat, Patricia}	999-04-1234	F
r_{123}	{Pat, Patricia}	999-04-1234	{F, M}



Negative Rules

Entity Resolution

	Name	SSN	Gender
r_1	Pat	999-04-1234	
r_2	Patricia		F
r_3	Pat	999-04-1234	M
r_{12}	{Pat, Patricia}	999-04-1234	F



Negative Rules

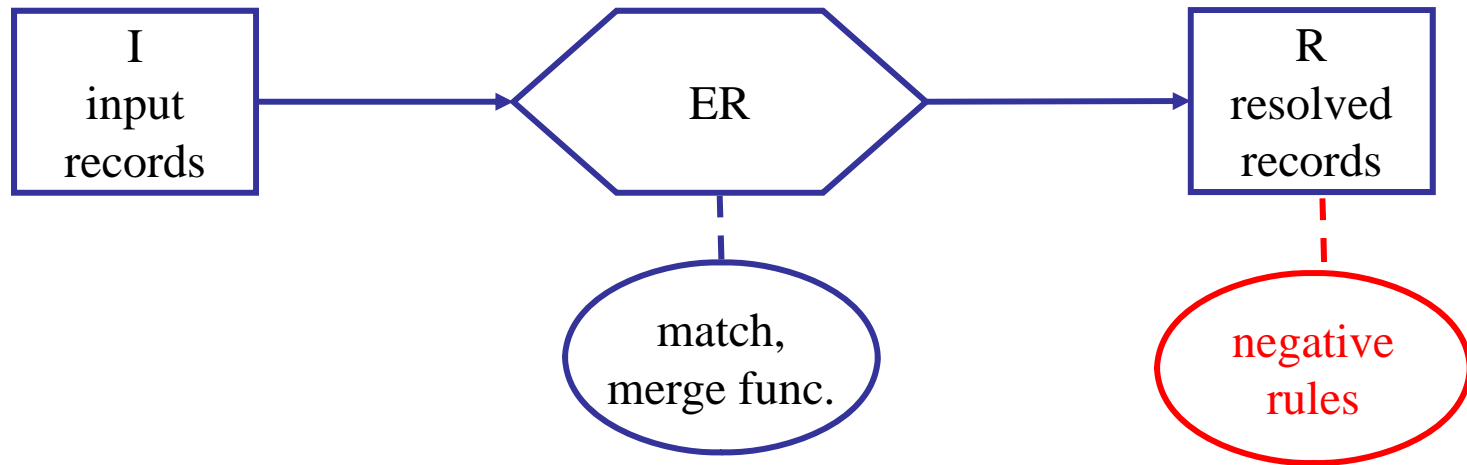
Entity Resolution

	Name	SSN	Gender
r_1	Pat	999-04-1234	
r_2	Patricia		F
r_3	Pat	999-04-1234	M

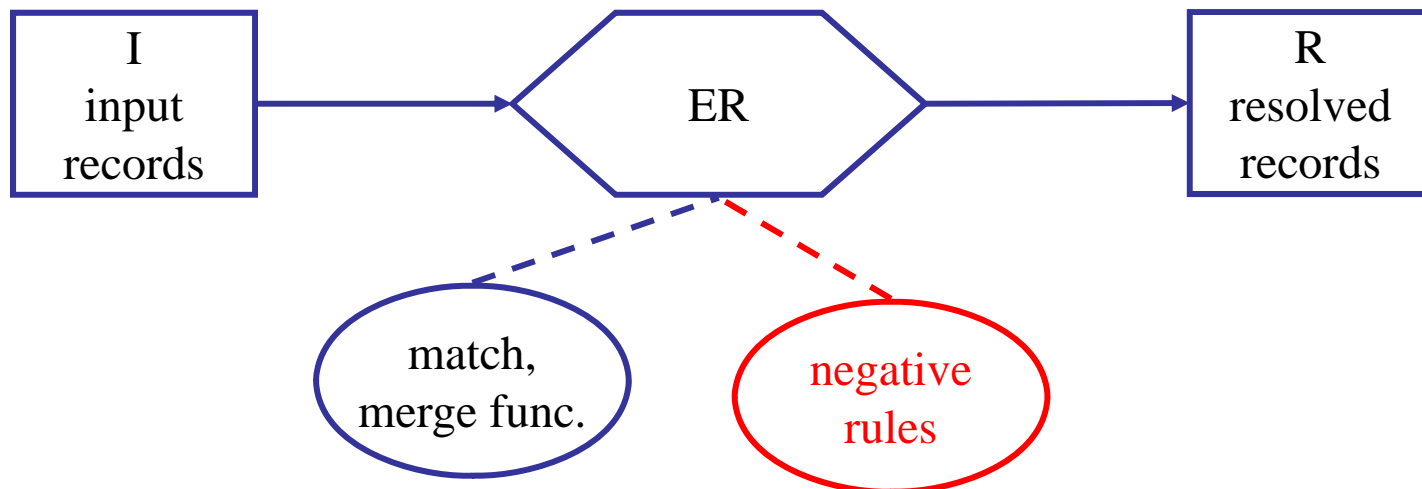
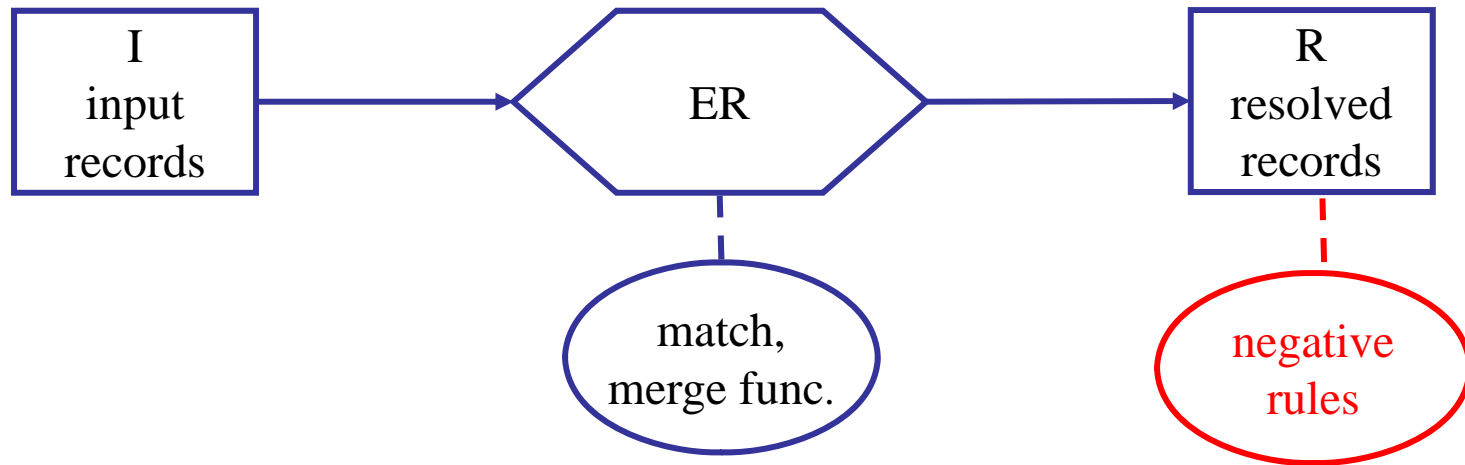
Solutions: $\{r_1, r_2\}$ or $\{r_1, r_3\}$

Undesirable: $\{r_1, r_3\}$

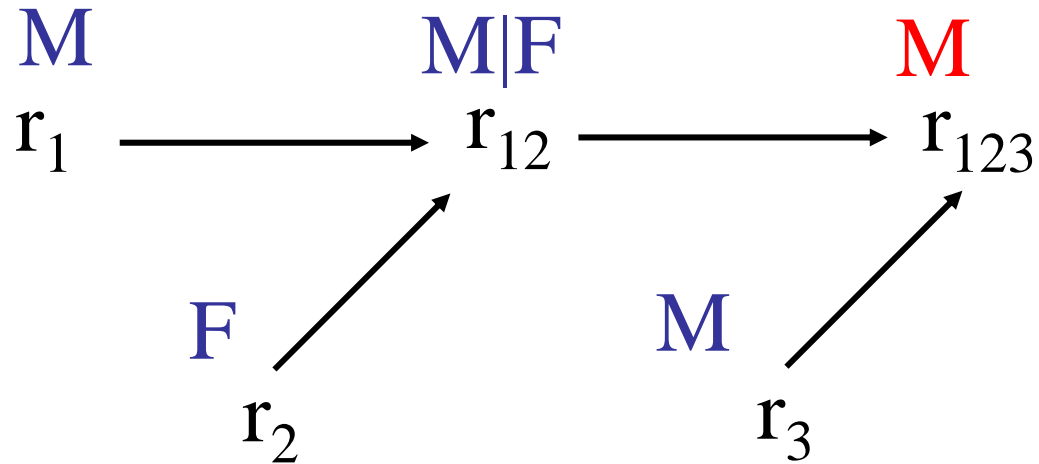
Negative Rules



Negative Rules

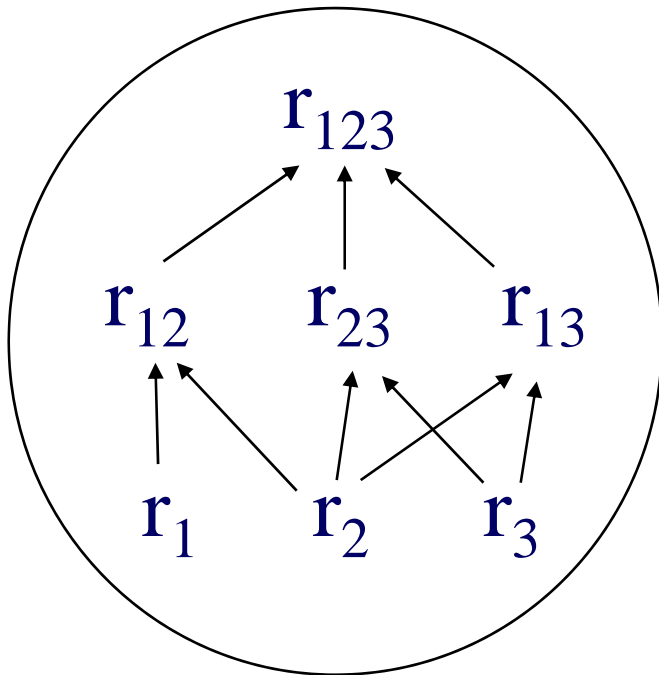


Why not simply extend match func.?



Algorithm

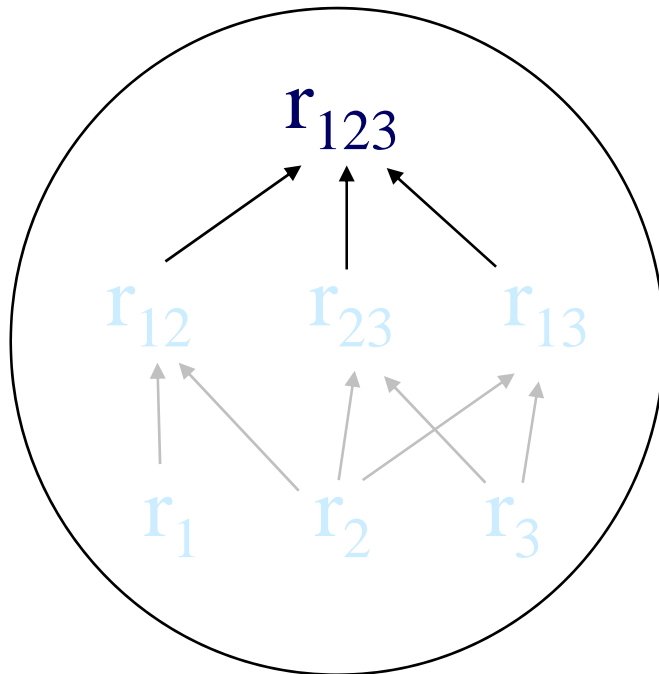
	Name	SSN	Gender
r_1	Pat	999-04-1234	
r_2	Patricia		F
r_3	Pat	999-04-1234	M



Solution

Algorithm

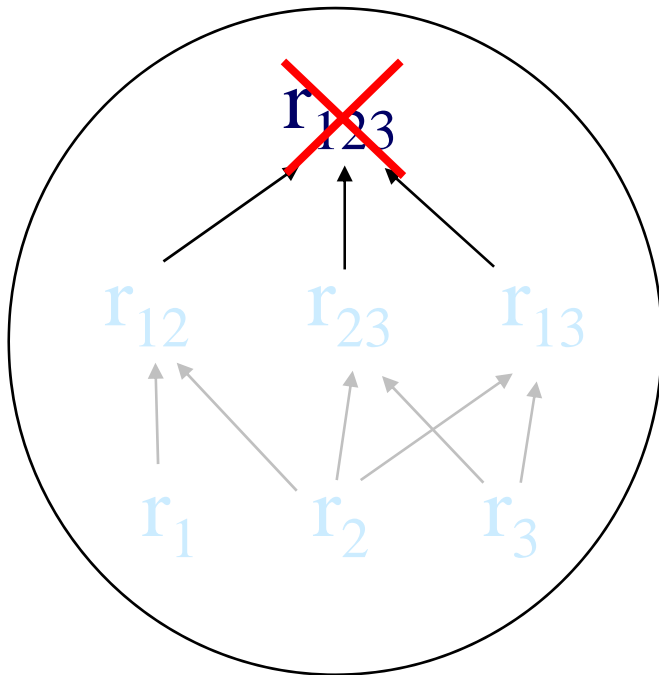
	Name	SSN	Gender
r_1	Pat	999-04-1234	
r_2	Patricia		F
r_3	Pat	999-04-1234	M



Solution

Algorithm

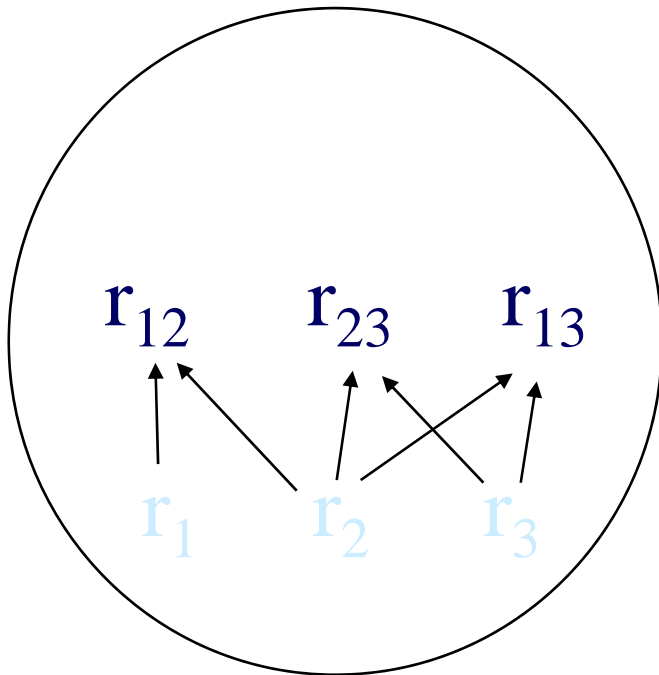
	Name	SSN	Gender
r_1	Pat	999-04-1234	
r_2	Patricia		F
r_3	Pat	999-04-1234	M



Solution

Algorithm

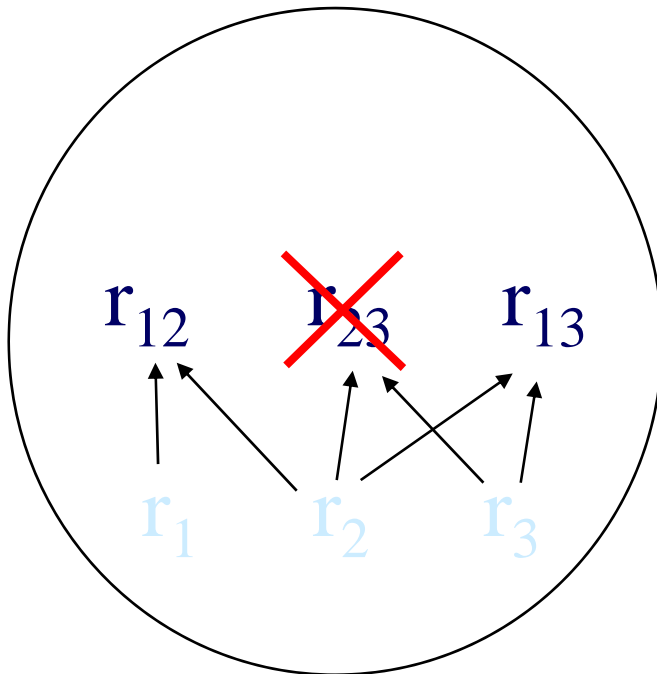
	Name	SSN	Gender
r_1	Pat	999-04-1234	
r_2	Patricia		F
r_3	Pat	999-04-1234	M



Solution

Algorithm

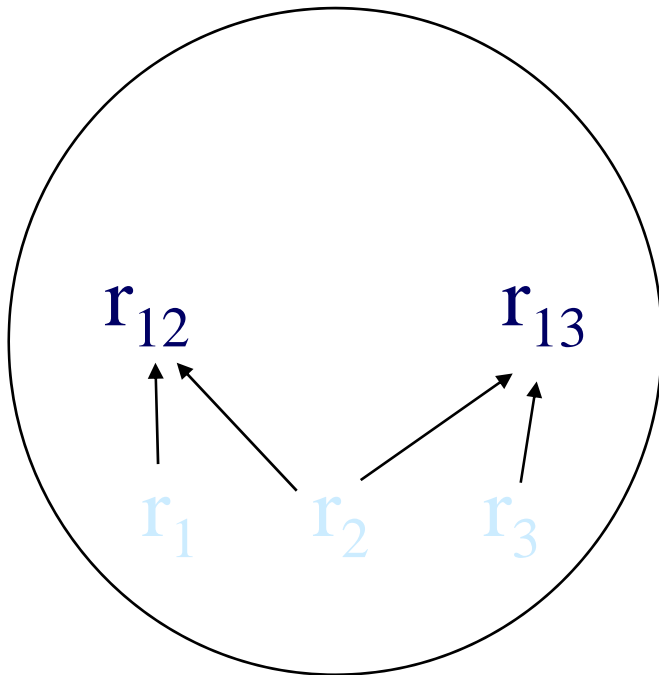
	Name	SSN	Gender
r_1	Pat	999-04-1234	
r_2	Patricia		F
r_3	Pat	999-04-1234	M



Solution

Algorithm

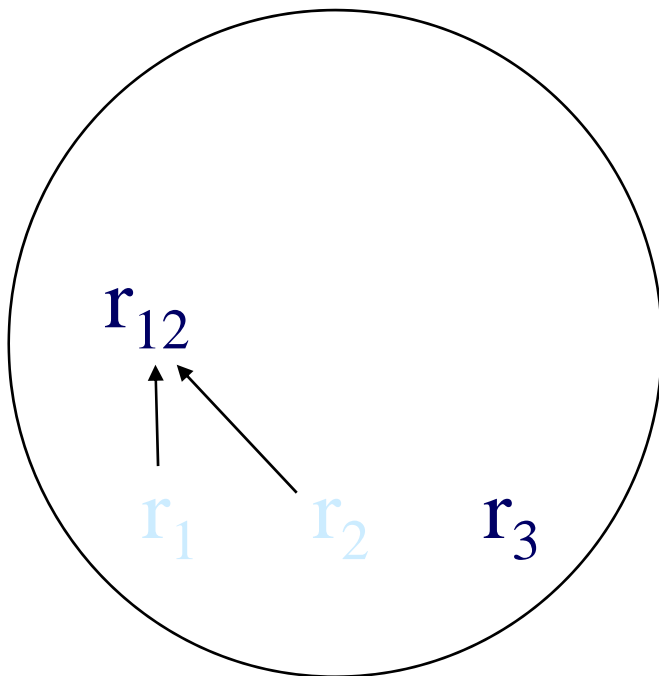
	Name	SSN	Gender
r_1	Pat	999-04-1234	
r_2	Patricia		F
r_3	Pat	999-04-1234	M



Solution

Algorithm

	Name	SSN	Gender
r_1	Pat	999-04-1234	
r_2	Patricia		F
r_3	Pat	999-04-1234	M

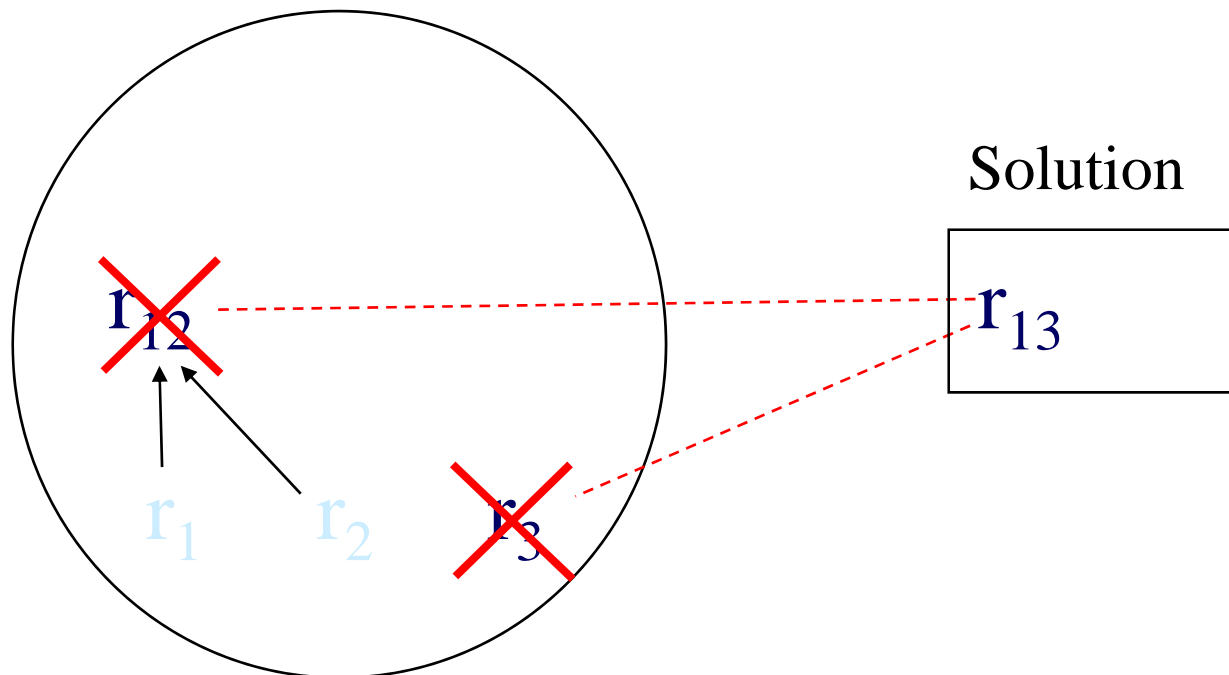


Solution

r_{13}

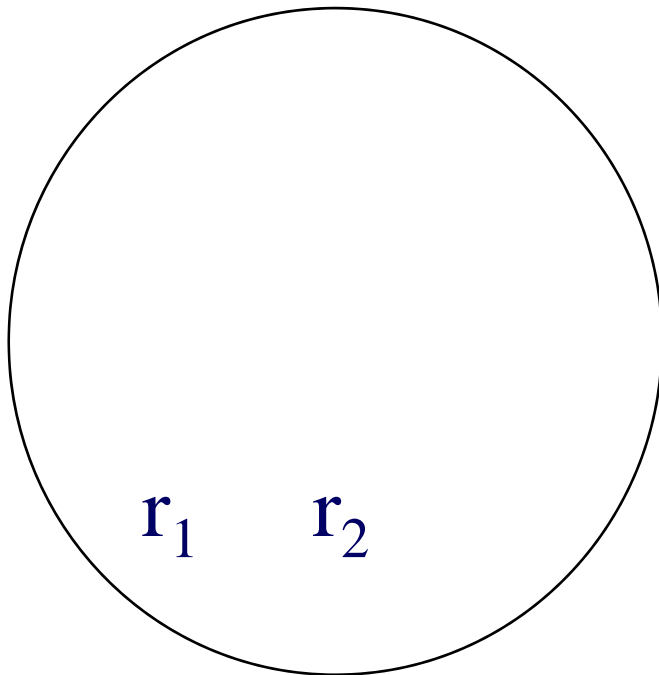
Algorithm

	Name	SSN	Gender
r_1	Pat	999-04-1234	
r_2	Patricia		F
r_3	Pat	999-04-1234	M



Algorithm

	Name	SSN	Gender
r_1	Pat	999-04-1234	
r_2	Patricia		F
r_3	Pat	999-04-1234	M

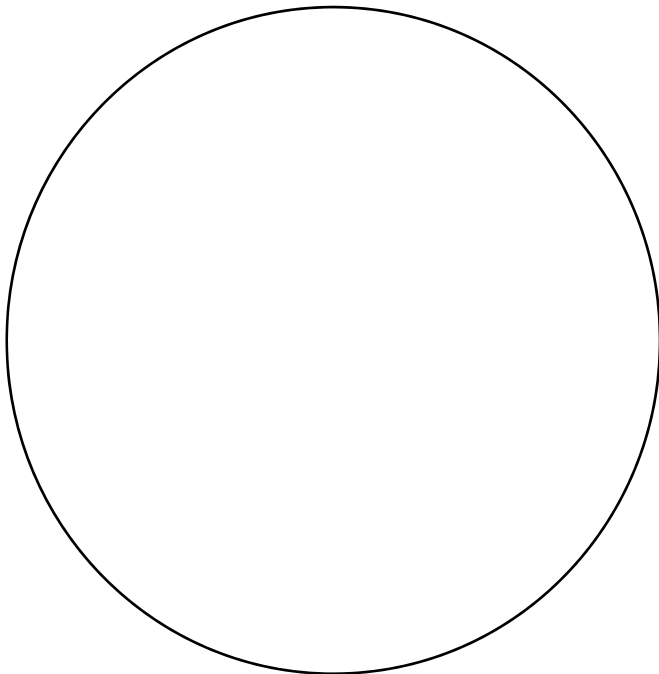


Solution

r_{13}

Algorithm

	Name	SSN	Gender
r_1	Pat	999-04-1234	
r_2	Patricia		F
r_3	Pat	999-04-1234	M



Solution

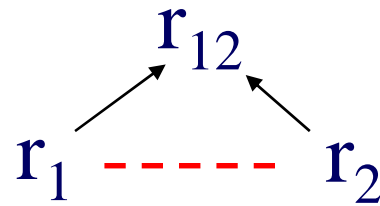
r_{13} r_2

Resolving Inconsistencies

Discard



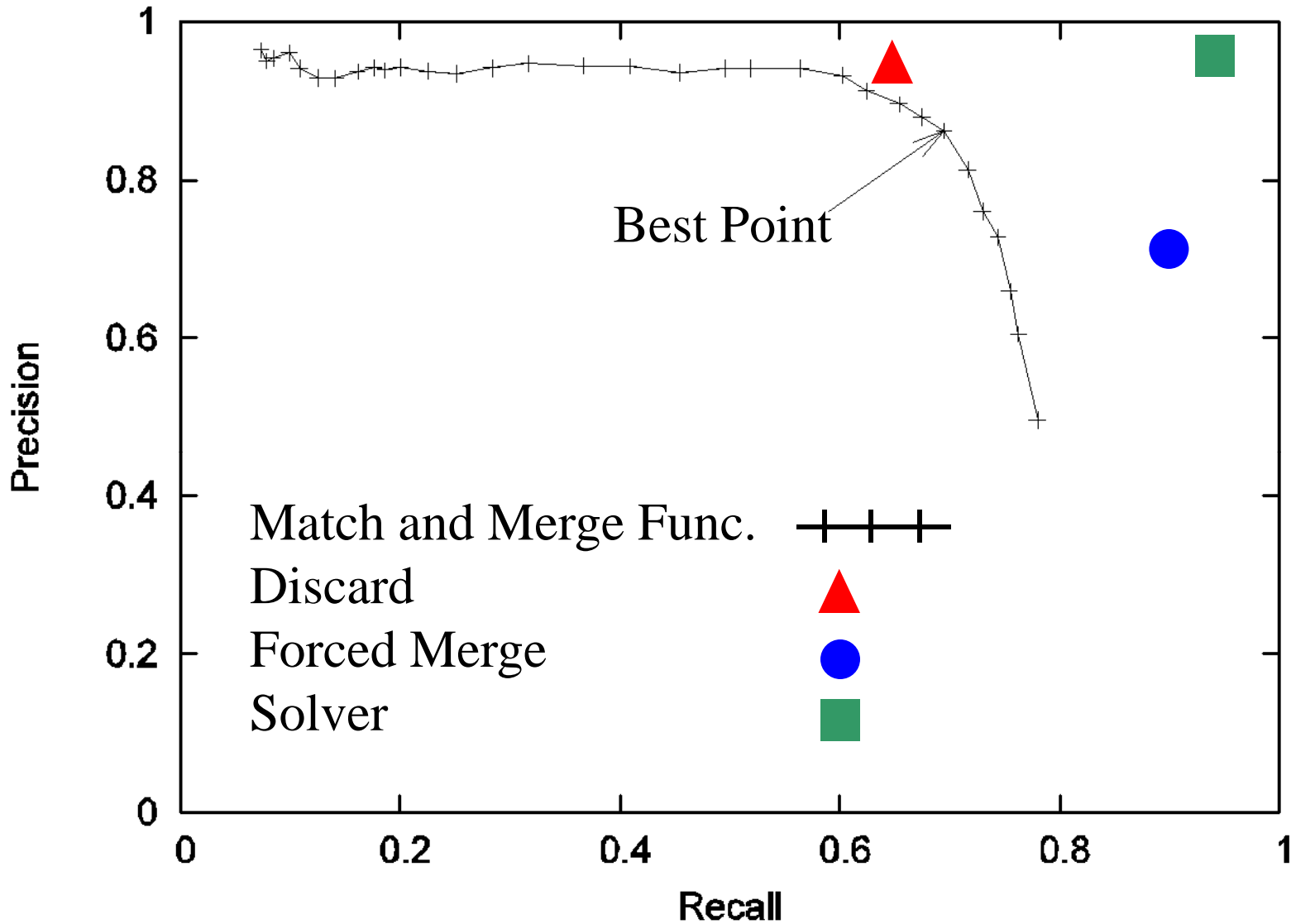
Forced Merge



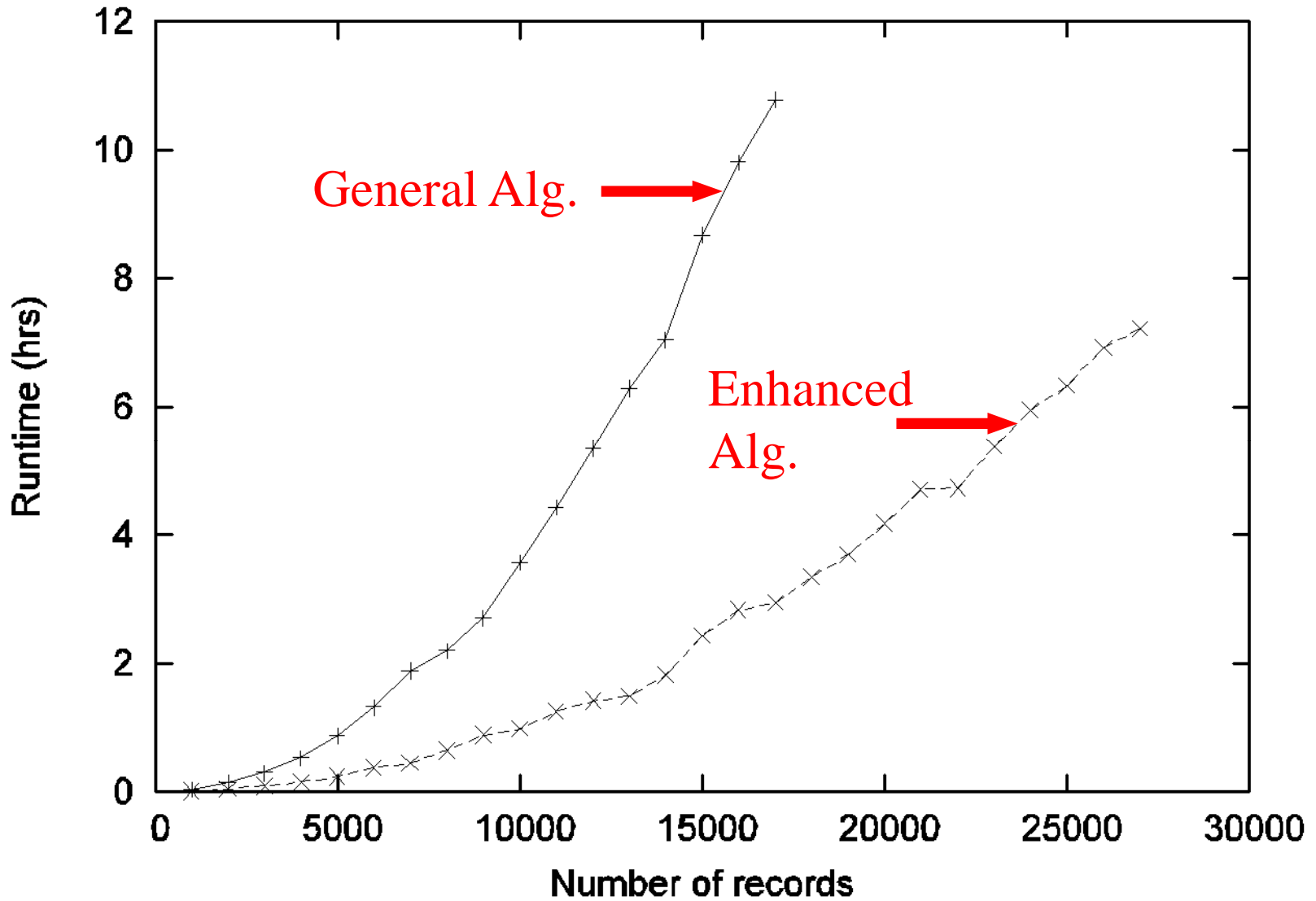
Override



Precision and Recall



Runtime

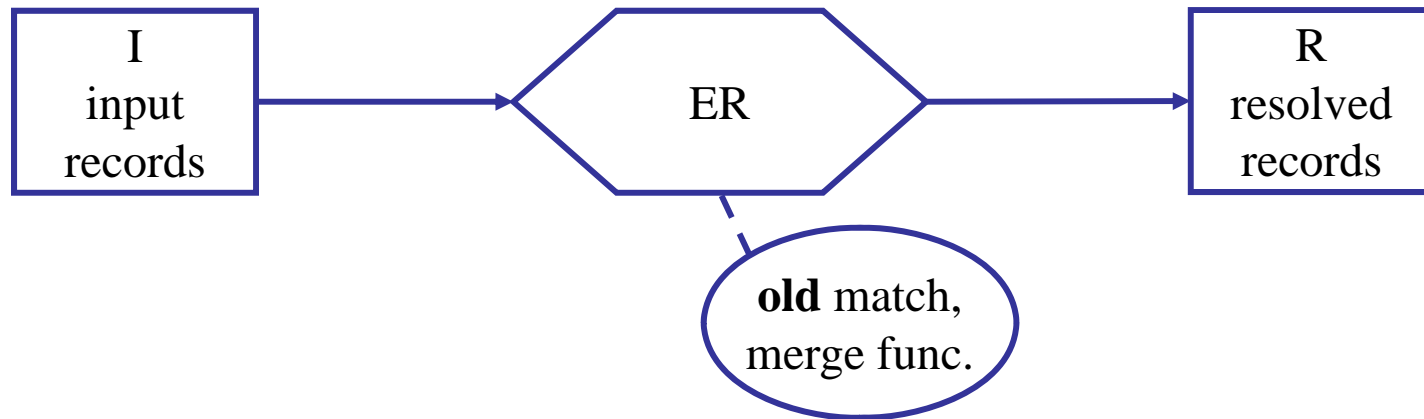


Negative Rules Summary

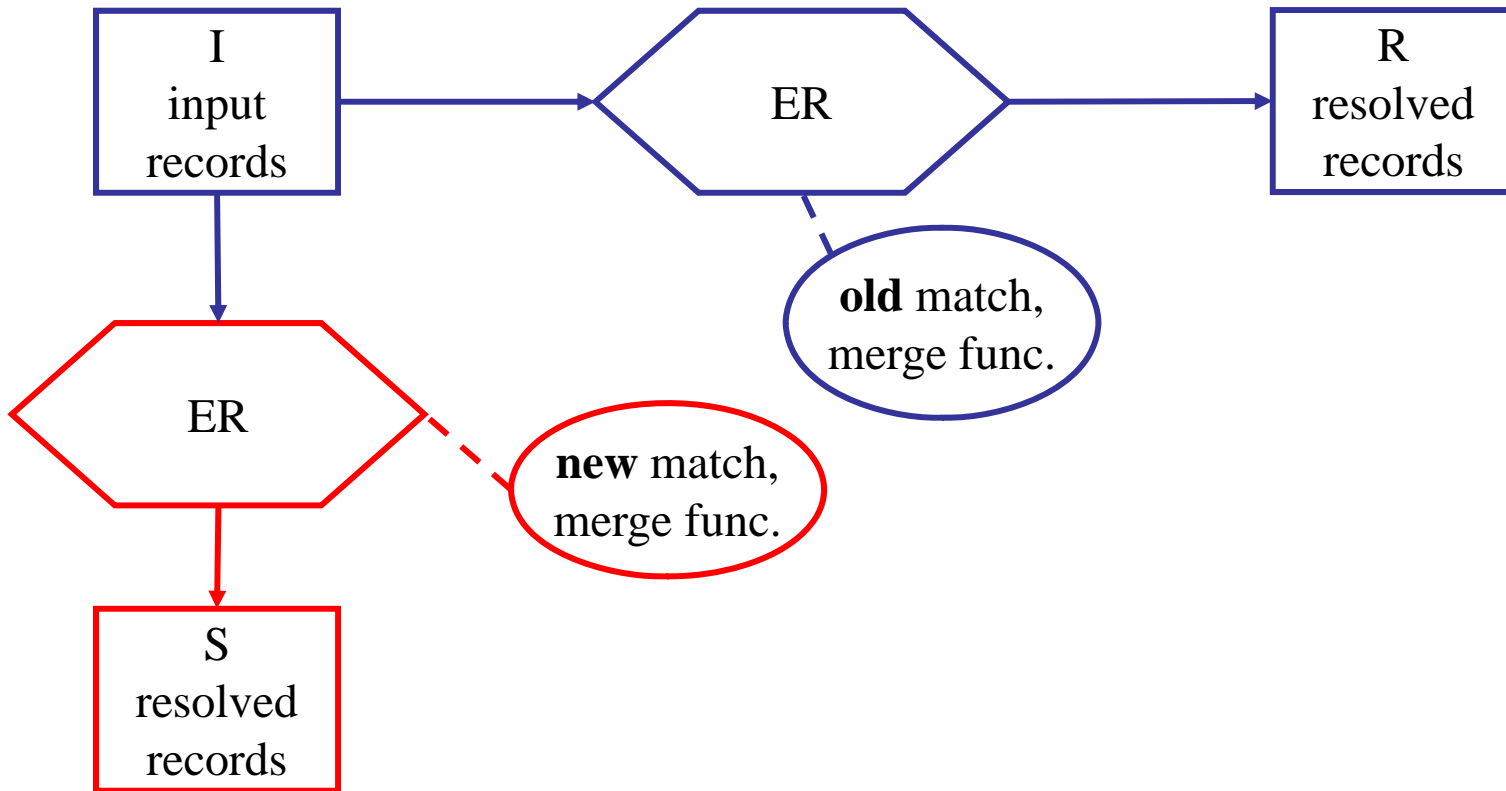
Negative Rules can improve the precision and recall of Entity Resolution

Entity Resolution with Negative Rules is very expensive and should be used within buckets after blocking

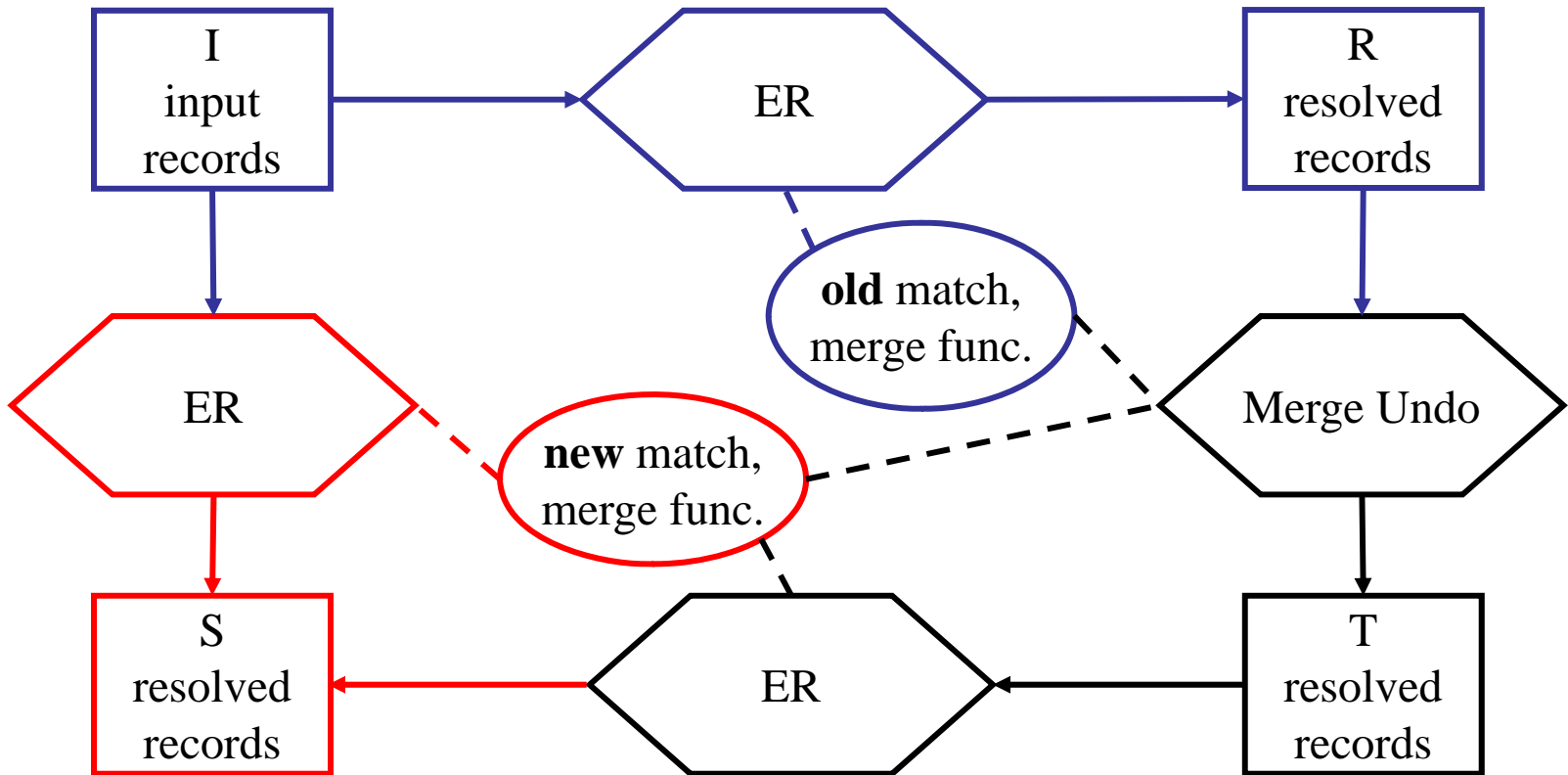
Evolving Rules



Evolving Rules



Evolving Rules



ER in the InfoLab

- Generic ER
- Confidences
- Distributed ER
- Negative Rules
- Evolving Rules
- Blocking