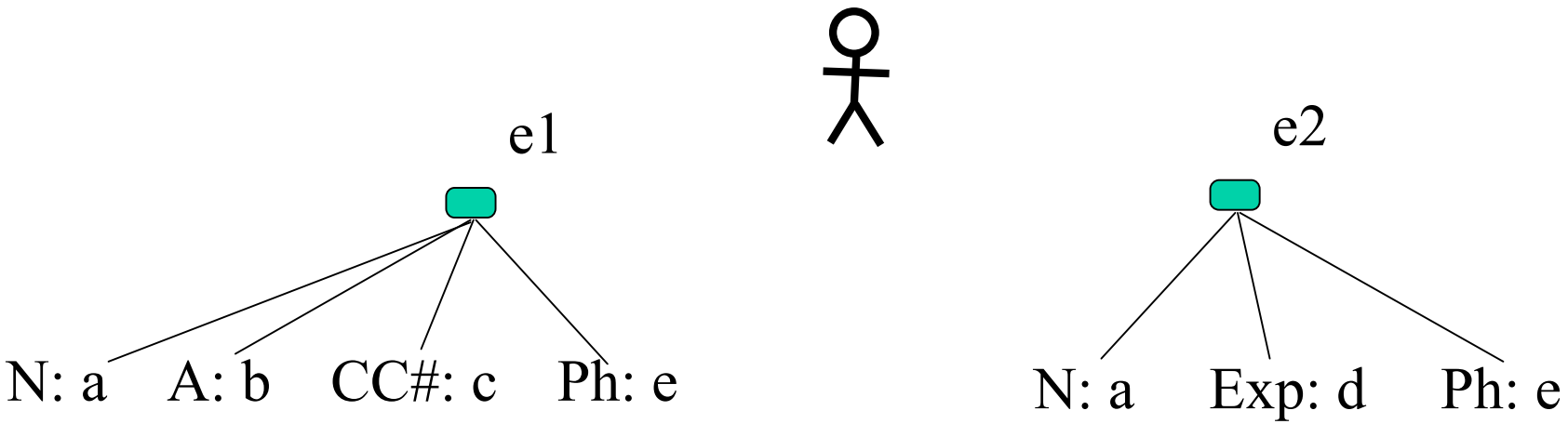
Entity Resolution with Evolving Rules

Steven Whang, Hector Garcia-Molina

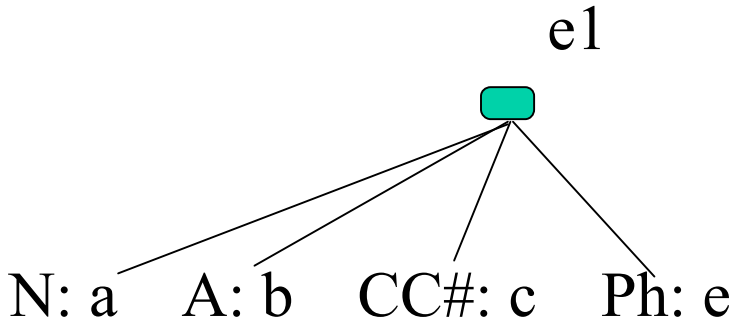
Stanford University

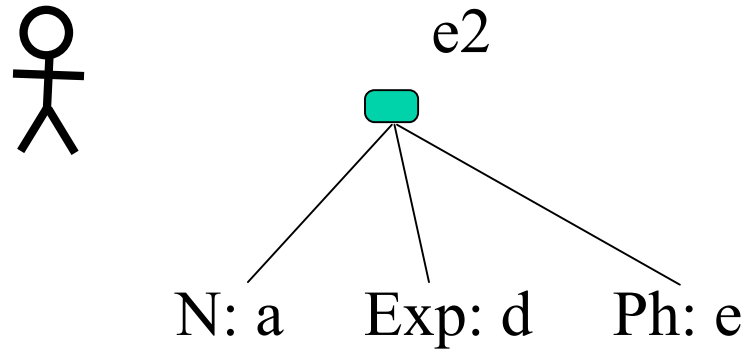
Entity Resolution



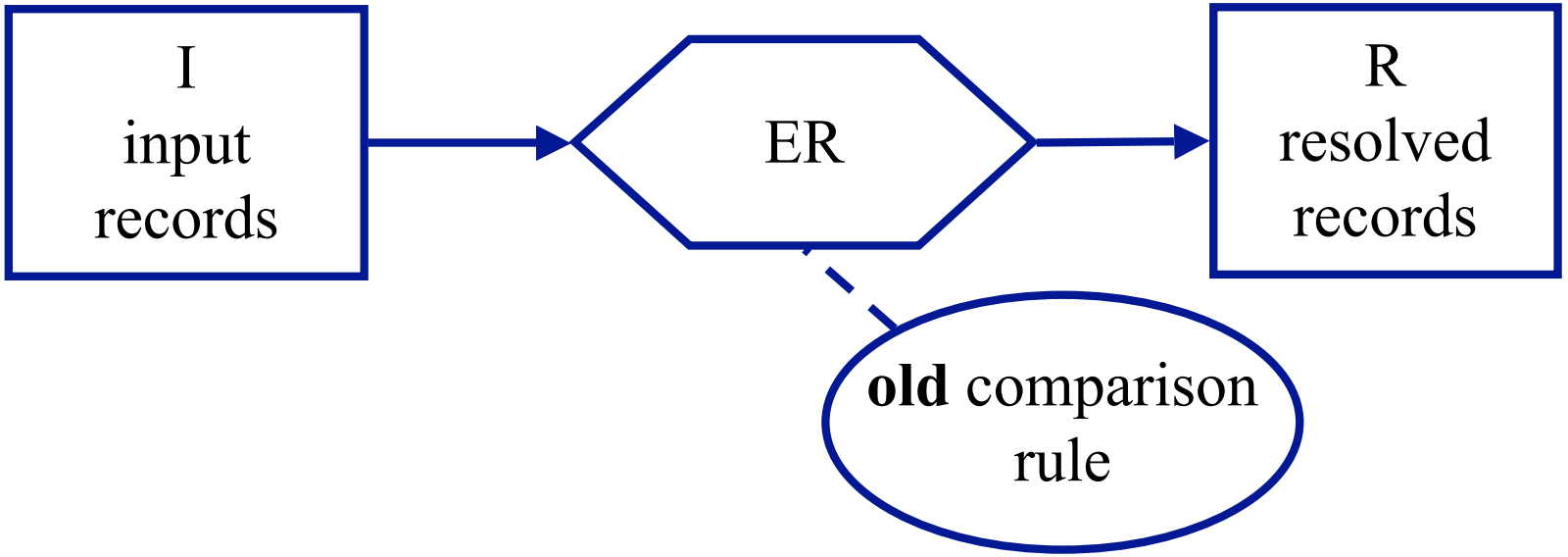
Applications

- comparison shopping
- mailing lists
- classified ads
- customer files
- counter-terrorism

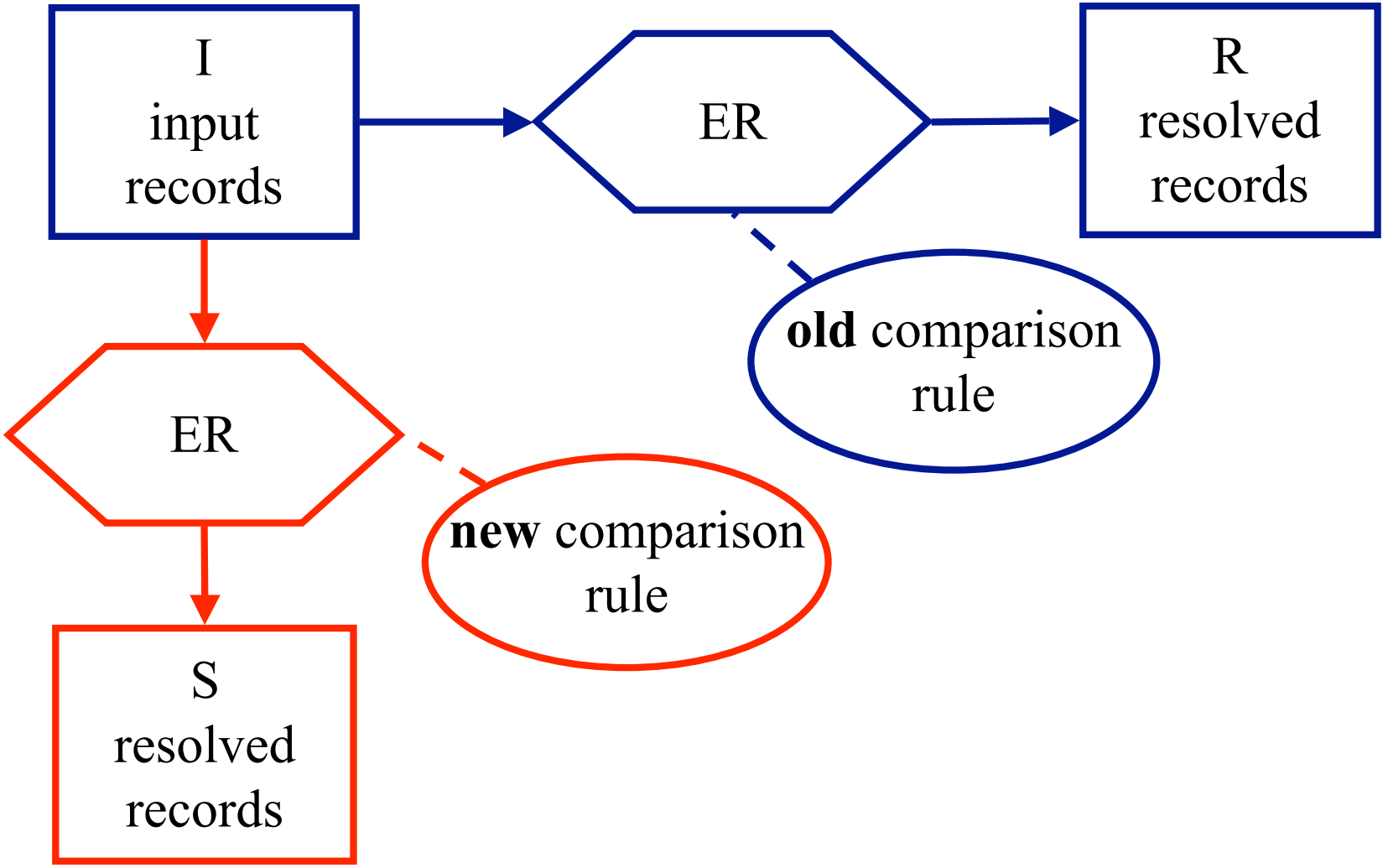




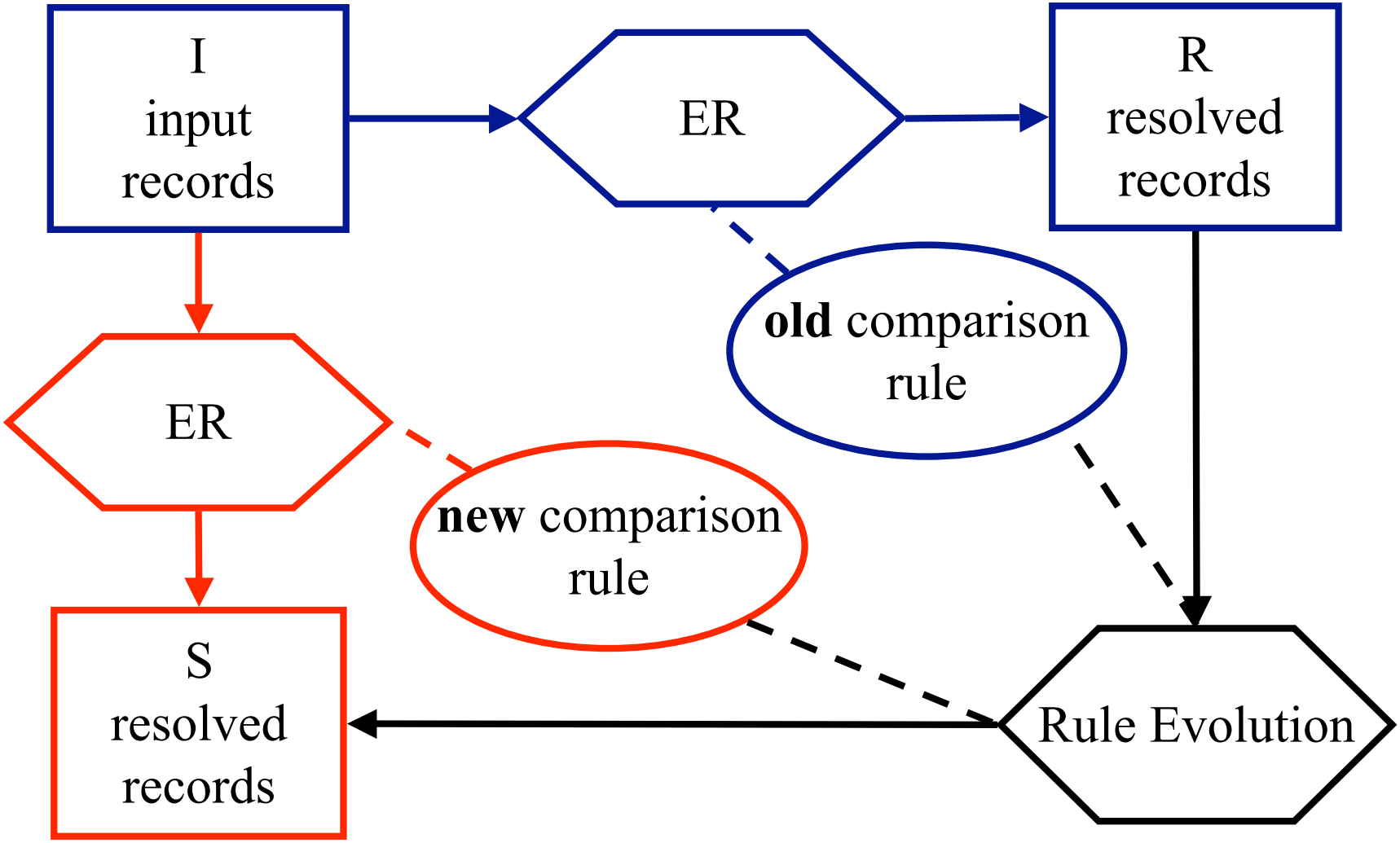
Evolving Rules



Evolving Rules

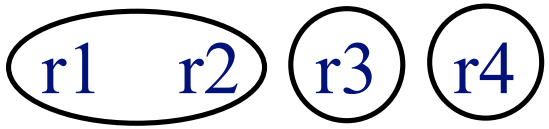


Evolving Rules



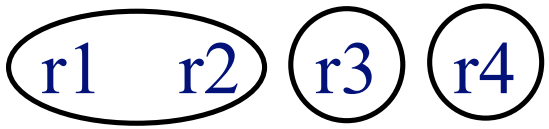
Example

Record	Name	Zip	Phone
r1	John	54321	123
r2	John	54321	987
r3	John	11111	987
r4	Bob	null	121

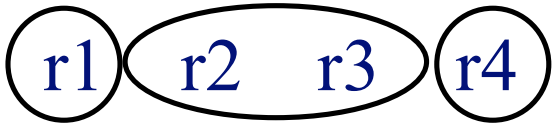
N & Z  (6 comps)

Example

Record	Name	Zip	Phone
r1	John	54321	123
r2	John	54321	987
r3	John	11111	987
r4	Bob	null	121

N & Z 

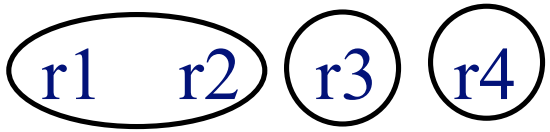
(6 comps)

N & P 

(6 comps)

Rule Evolution

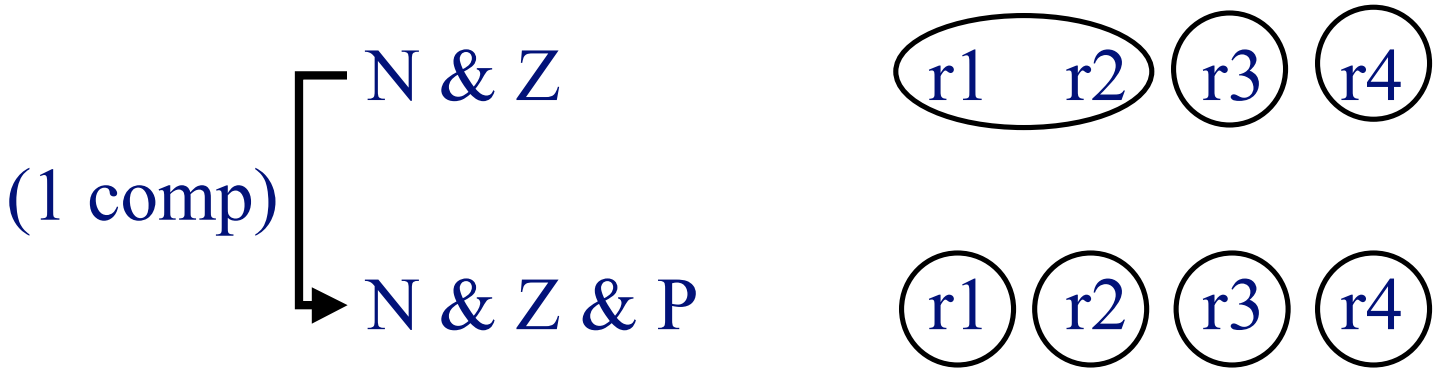
N & Z



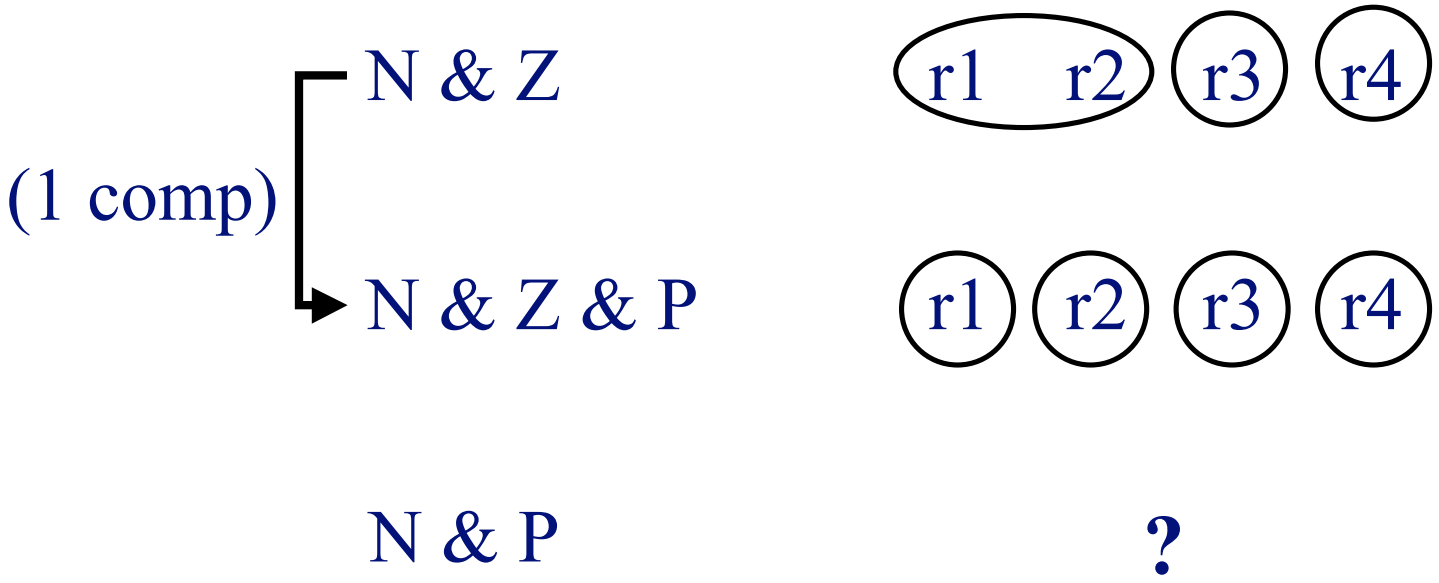
N & Z & P

r1 r2 r3 r4

Rule Evolution



Rule Evolution

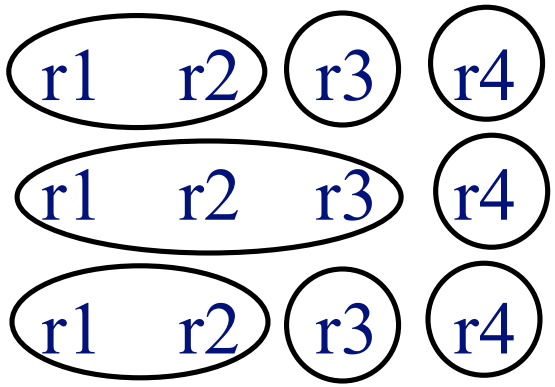


Materialization

N & Z

N

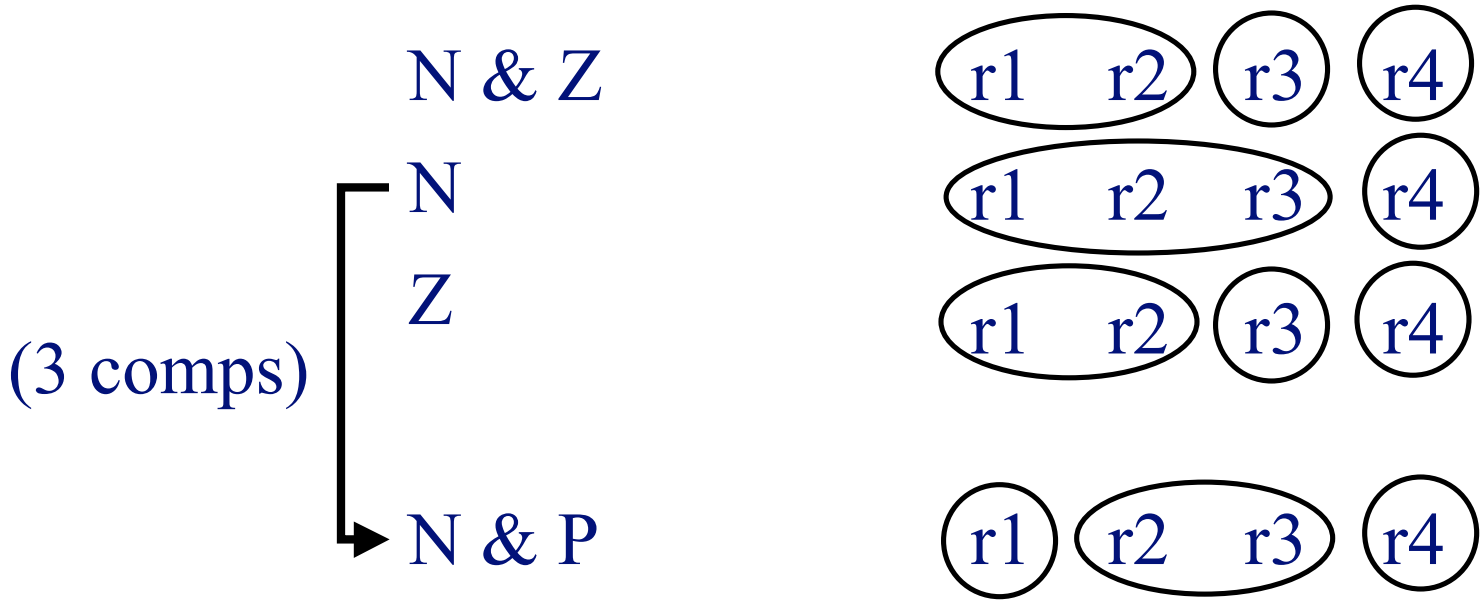
Z



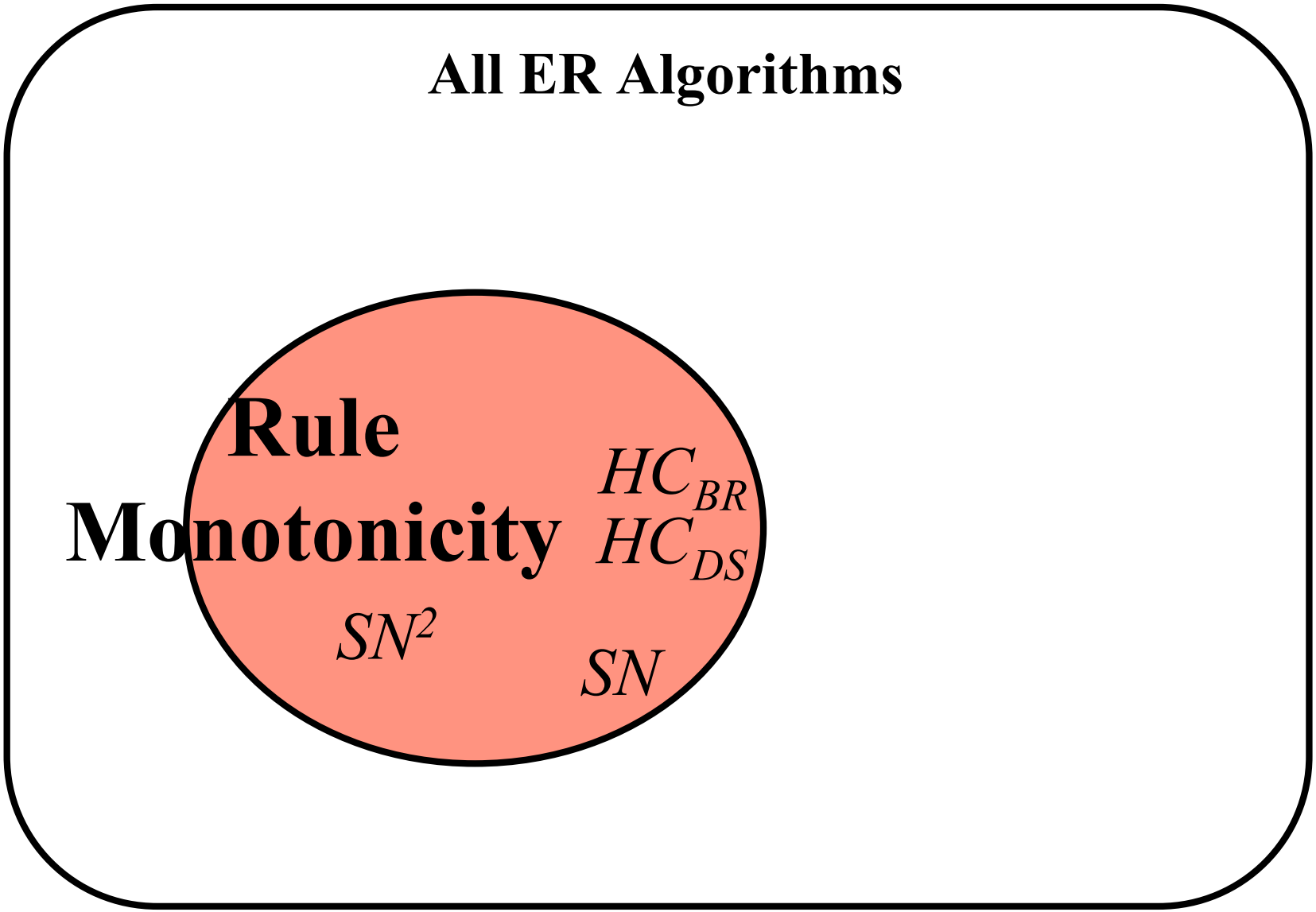
N & P

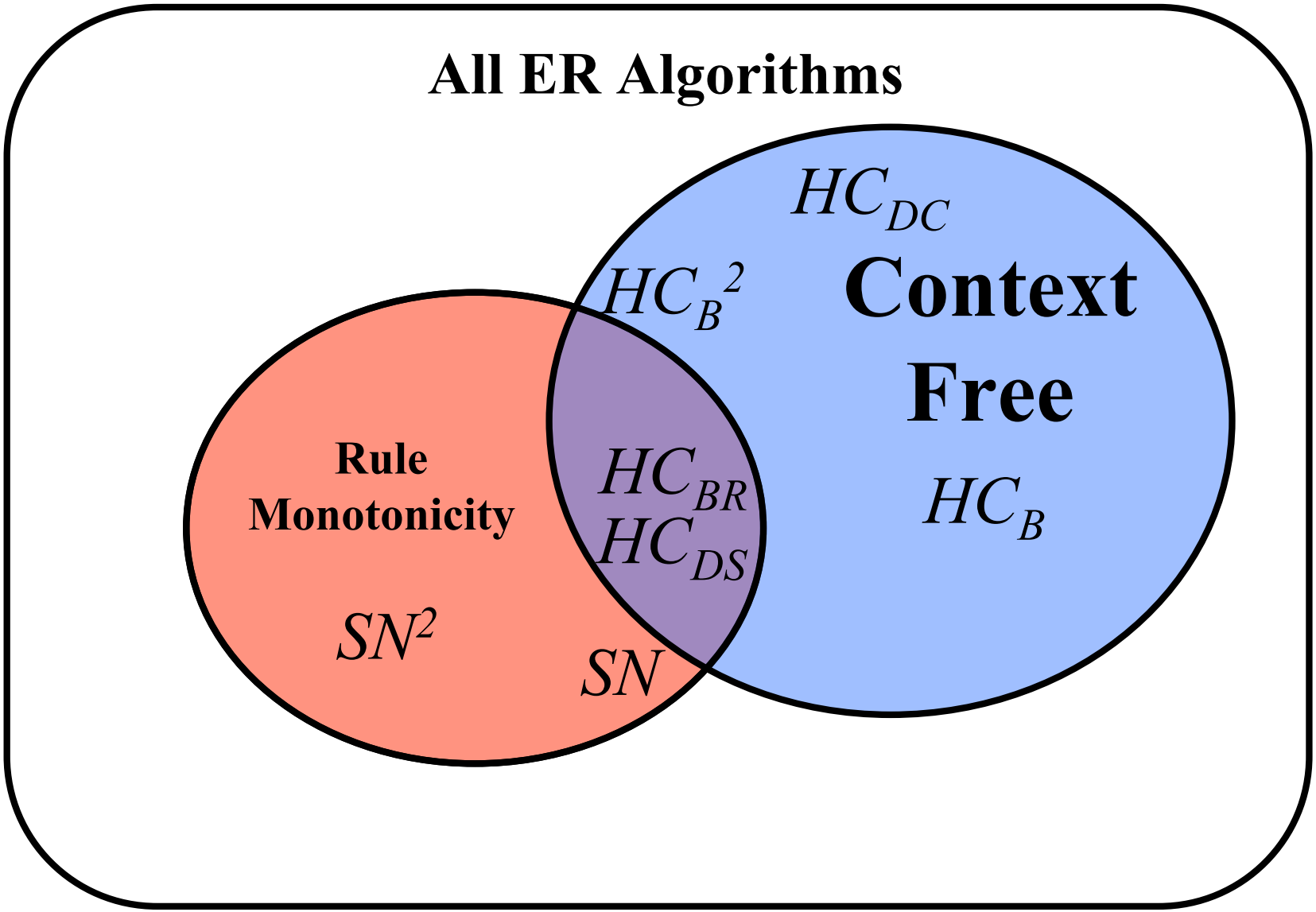
r1 r2 r3 r4

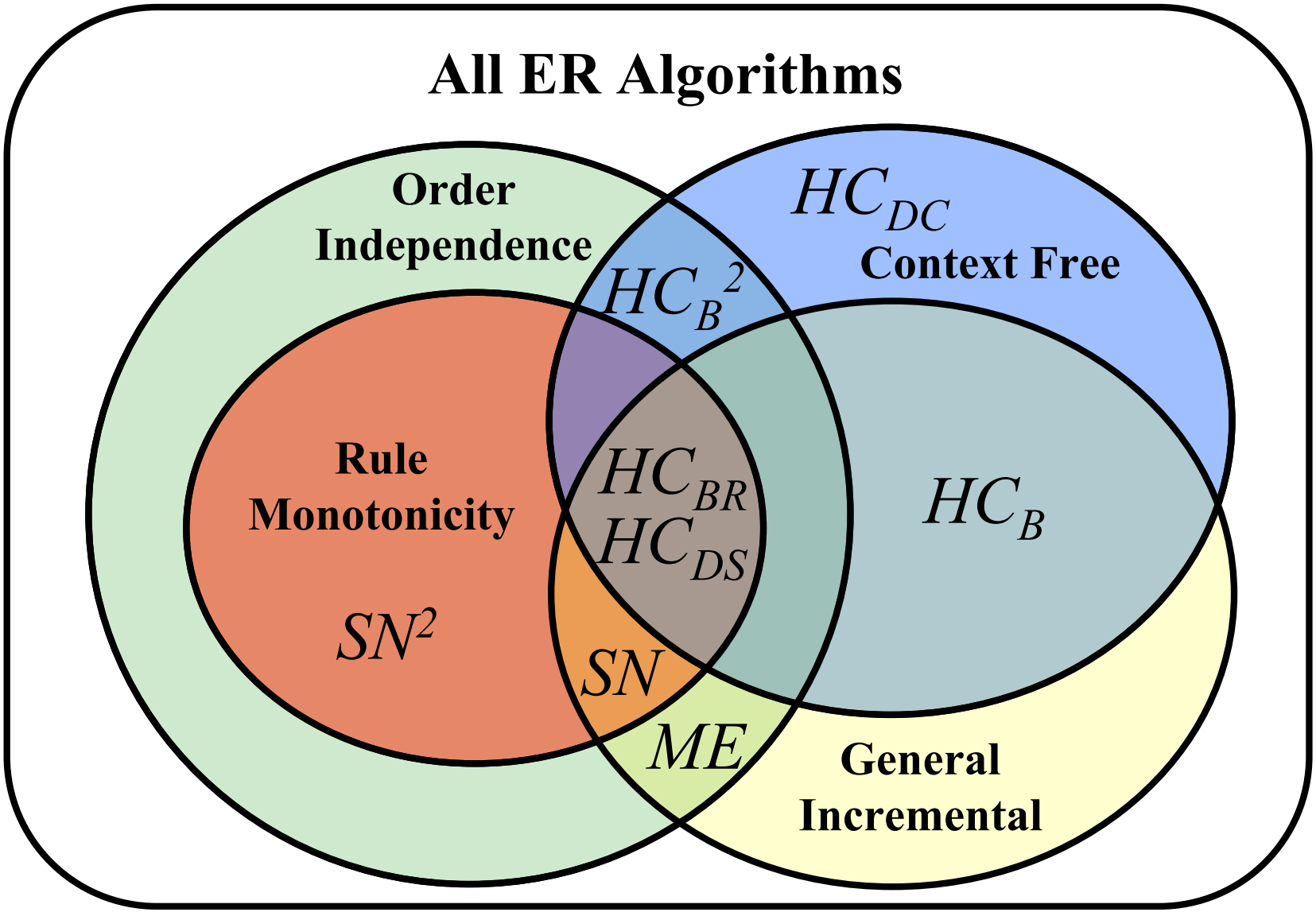
Materialization

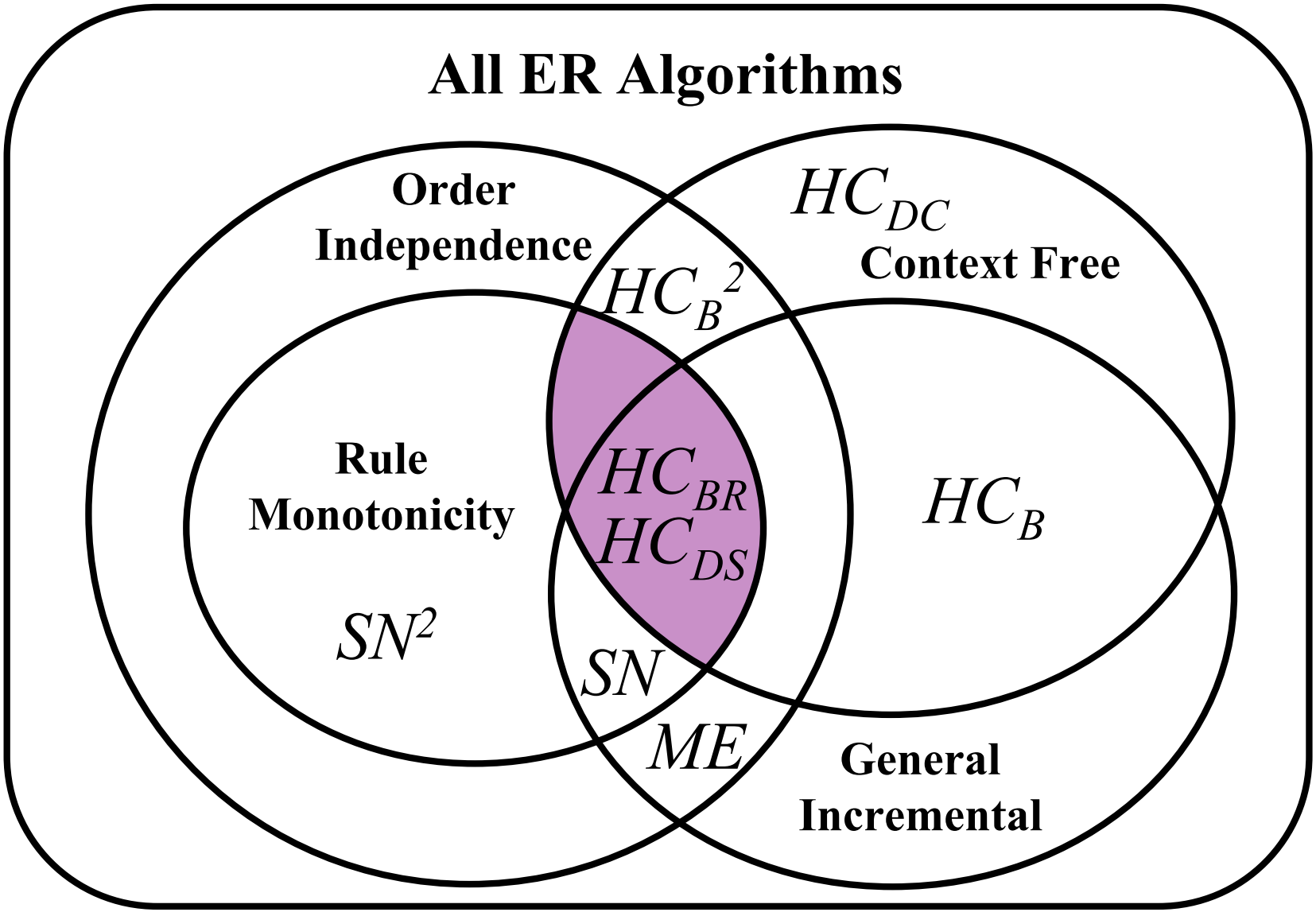


All ER Algorithms


All ER Algorithms
Rule Monotonicity
HC_{BR}
HC_{DS}
SN^2
SN


All ER Algorithms
HC_{DC}
$HC_B{}^2$
Context
Free
HC_B
Rule
Monotonicity
HC_{BR}
HC_{DS}
SN^2
SN


All ER Algorithms
Order Independence
HC_{DC}
Context Free
$HC_B{}^2$
Rule Monotonicity
HC_{BR}
HC_{DS}
HC_B
SN^2
SN
ME
General Incremental


All ER Algorithms
Order Independence
HC_{DC}
Context Free
$HC_B{}^2$
Rule Monotonicity
HC_{BR}
HC_{DS}
HC_B
SN^2
SN
ME
General Incremental

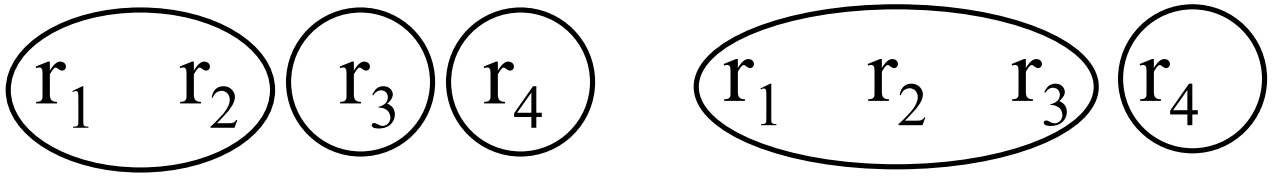
Strictness

- $B1 \leq B2$

  e.g., $N\&Z \leq N$

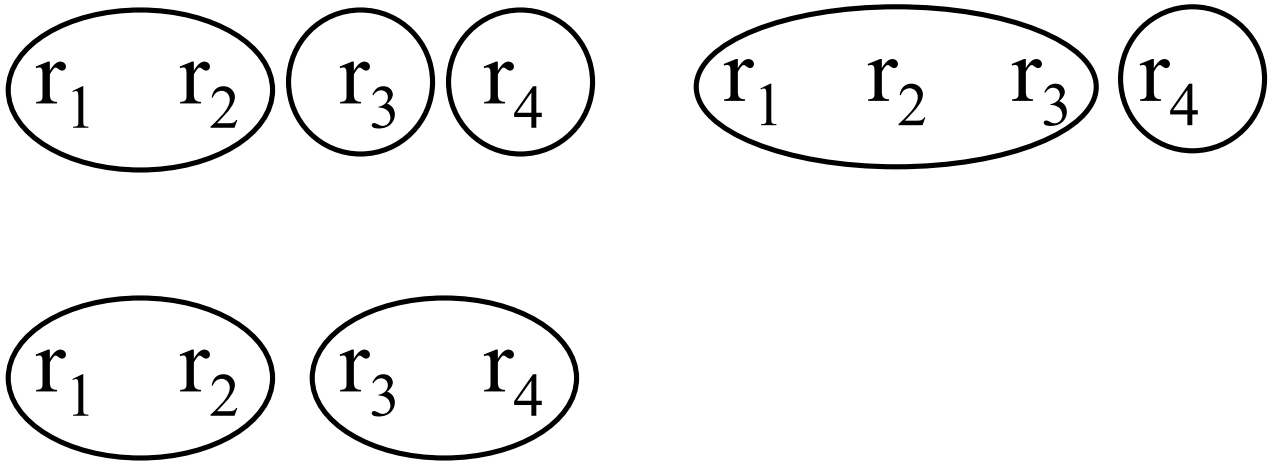
Domination

- ER result 1 $\leq$ ER result 2



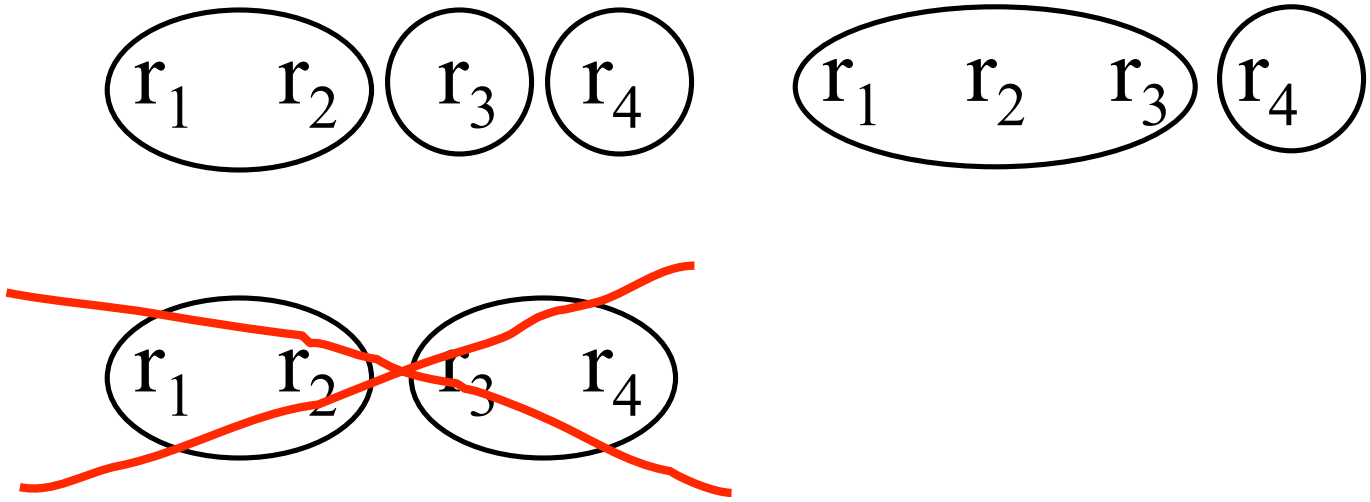
Domination

- ER result 1 ≤ ER result 2



Domination

- ER result 1 ≤ ER result 2



Definition: Rule Monotonicity (RM)

- If B1 $\leq$ B2

  Then ER(R, B1) $\leq$ ER(R, B2)

Definition: Rule Monotonicity (RM)

- If B1 $\leq$ B2

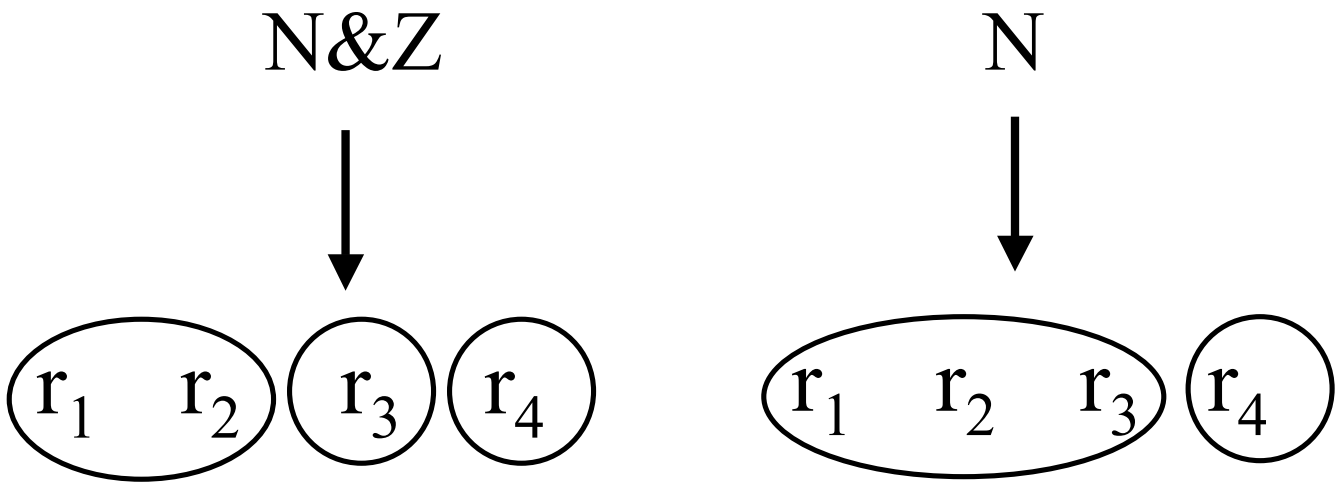
  Then ER(R, B1) $\leq$ ER(R, B2)

N&Z N

Definition: Rule Monotonicity (RM)

- If B1 $\leq$ B2

  Then ER(R, B1) $\leq$ ER(R, B2)



Definition: Context Free (CF)

- If $ER(R1 \cup R2, B) \leq \{R1, R2\}$

  Then $ER(R1 \cup R2, B)$

  $= ER(R1, B) \cup ER(R2, B)$

Definition: Context Free (CF)

- If $ER(R1 \cup R2, B) \leq \{R1, R2\}$

  Then $ER(R1 \cup R2, B)$

  $= ER(R1, B) \cup ER(R2, B)$

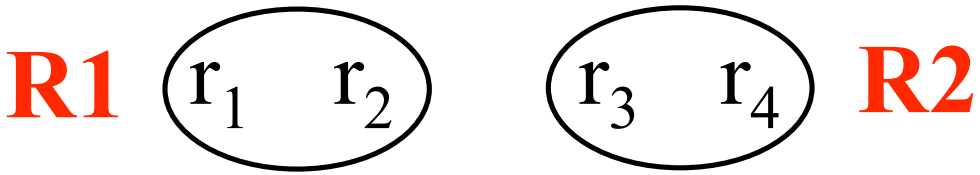
r_1 r_2 r_3 r_4

Definition: Context Free (CF)

- If ER(R1∪R2, B) ≤ {R1, R2}

  Then ER(R1∪R2, B)

  = ER(R1, B) ∪ ER(R2, B)

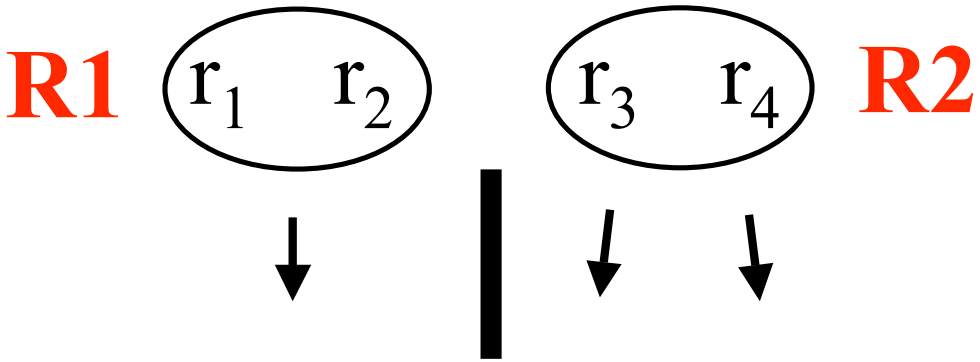


Definition: Context Free (CF)

- If ER(R1∪R2, B) ≤ {R1, R2}

  Then ER(R1∪R2, B)

  = ER(R1, B) ∪ ER(R2, B)

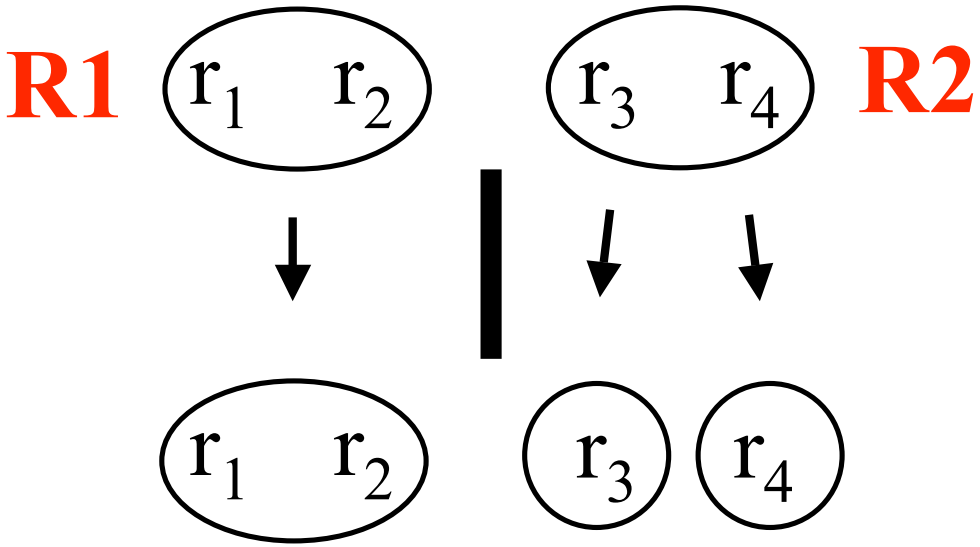


Definition: Context Free (CF)

- If ER(R1∪R2, B) ≤ {R1, R2}

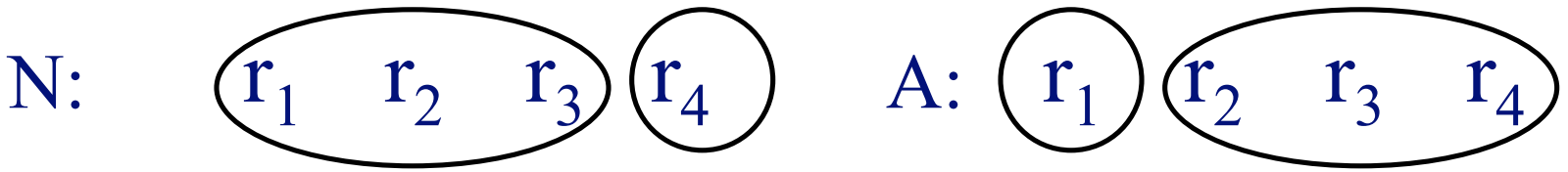
  Then ER(R1∪R2, B)

  = ER(R1, B) ∪ ER(R2, B)



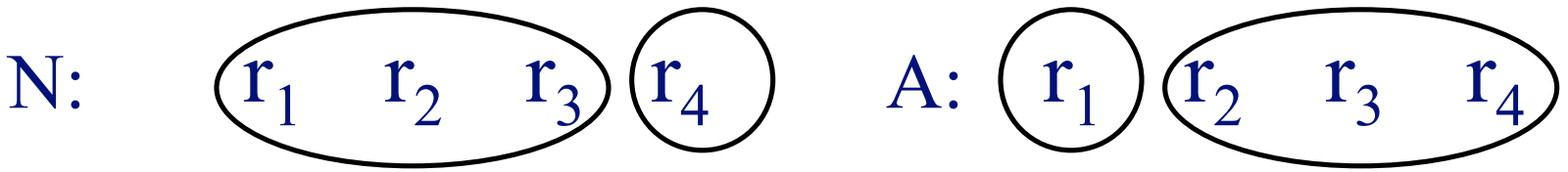
Algorithm exploiting RM & CF

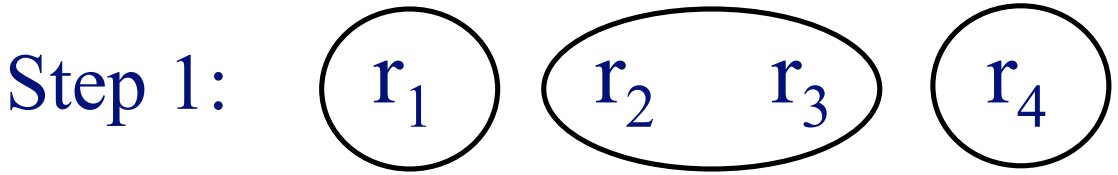
N & A & Z => N & A & P



Algorithm exploiting RM & CF

N & A & Z => N & A & P

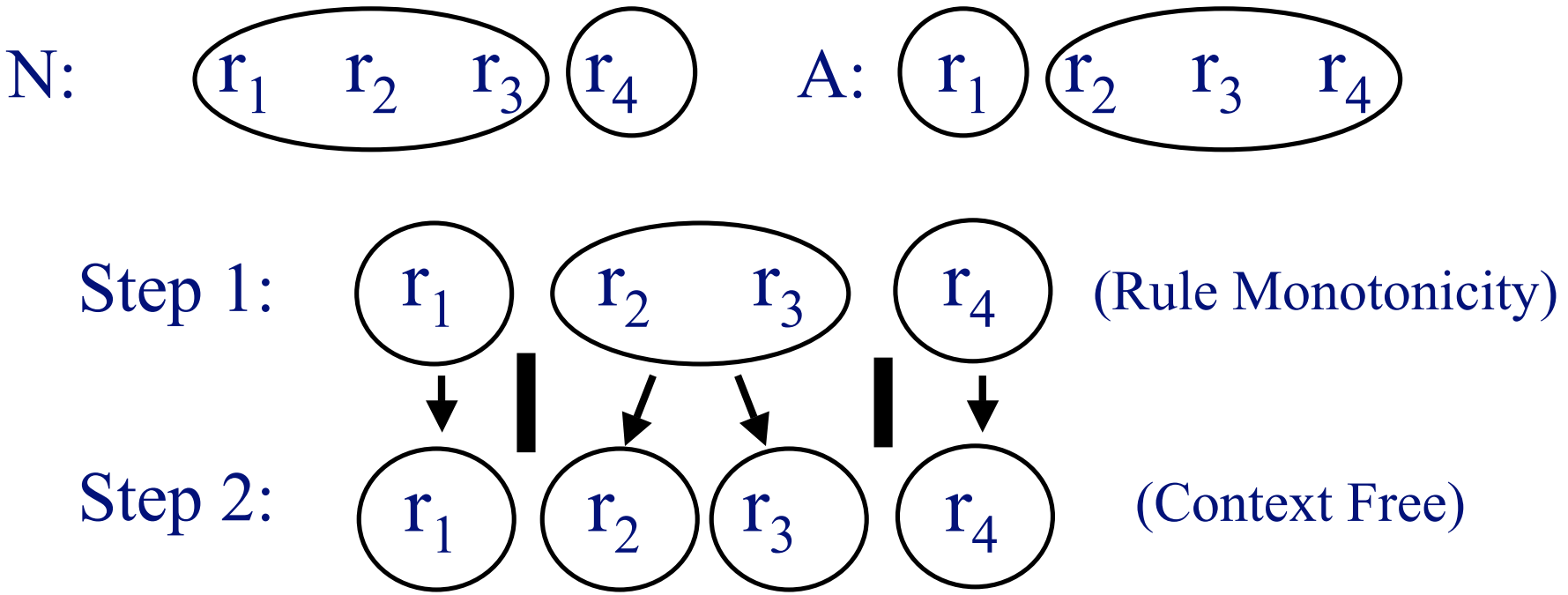


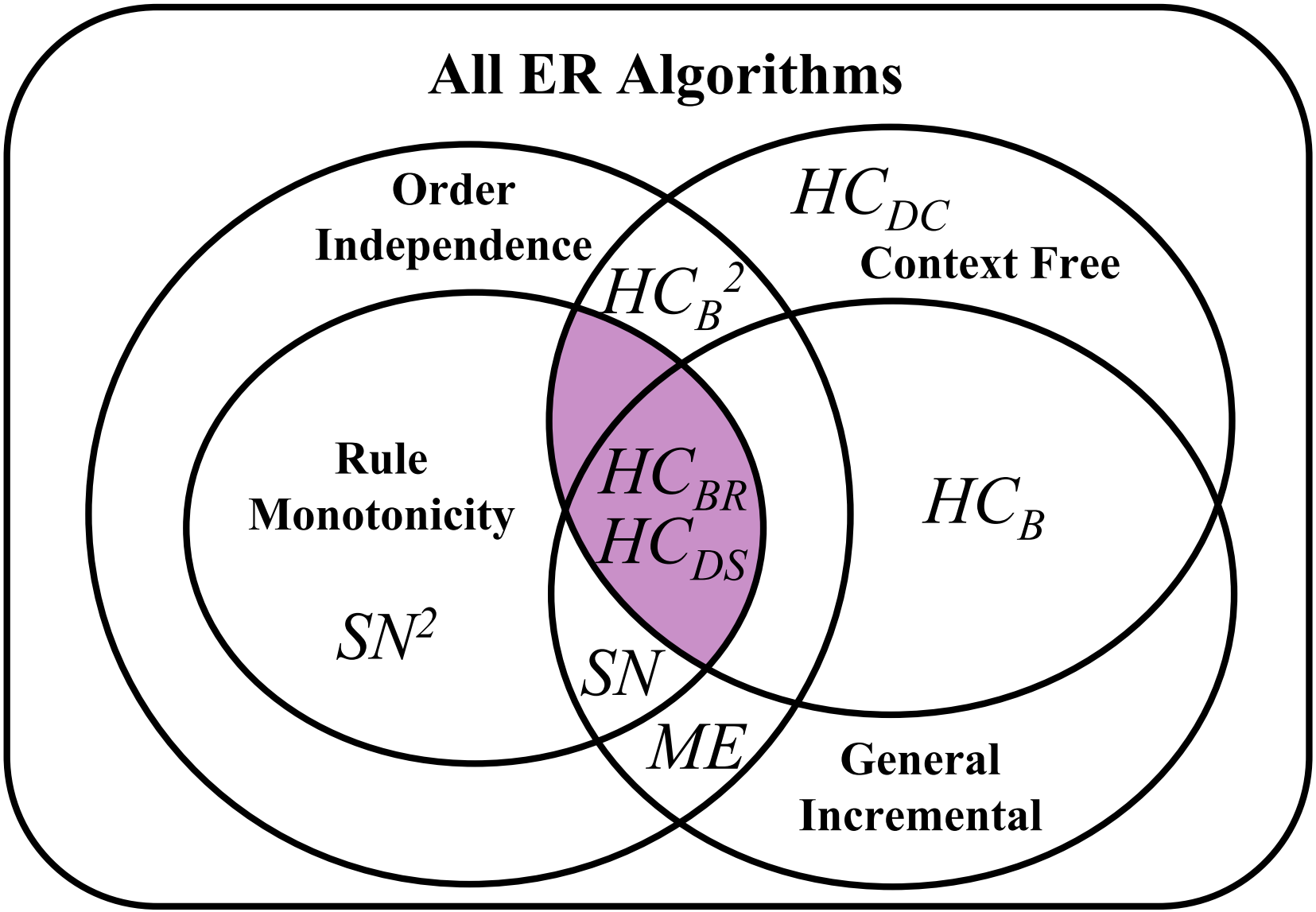


(Rule Monotonicity)

Algorithm exploiting RM & CF

N & A & Z => N & A & P

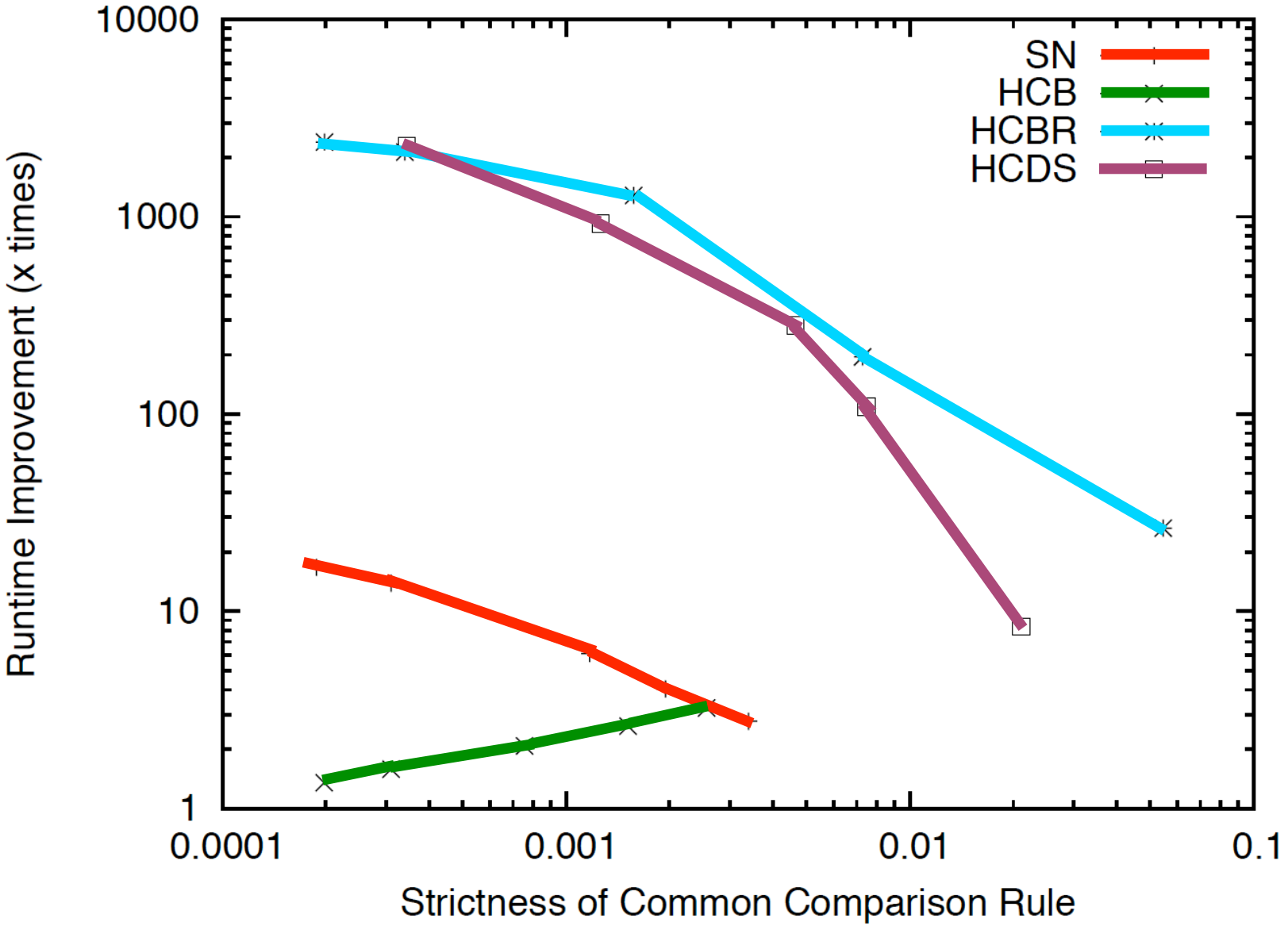



All ER Algorithms
Order Independence
HC_{DC}
Context Free
$HC_B{}^2$
Rule Monotonicity
HC_{BR}
HC_{DS}
HC_B
SN^2
SN
ME
General Incremental

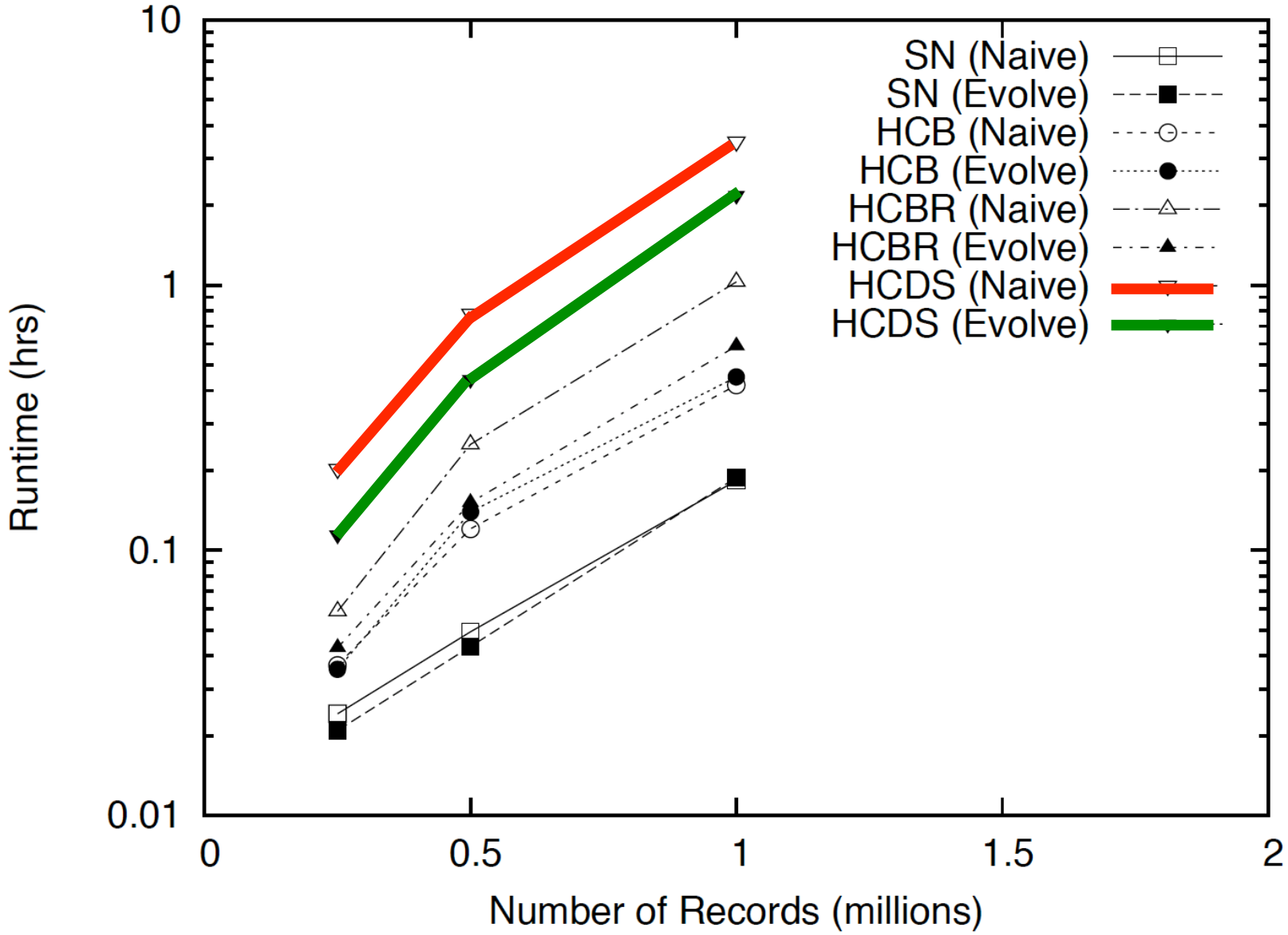
ER models

- Match-based clustering
- Distance-based clustering

Rule Evolution Speedup



Total Runtime



Conclusion

- Proposed a rule evolution framework that exploits ER properties and materialization
- Showed that rule evolution can significantly enhance the runtime of the naive approach

Thanks!