

Handling Data Quality in Entity Resolution

Hector Garcia-Molina

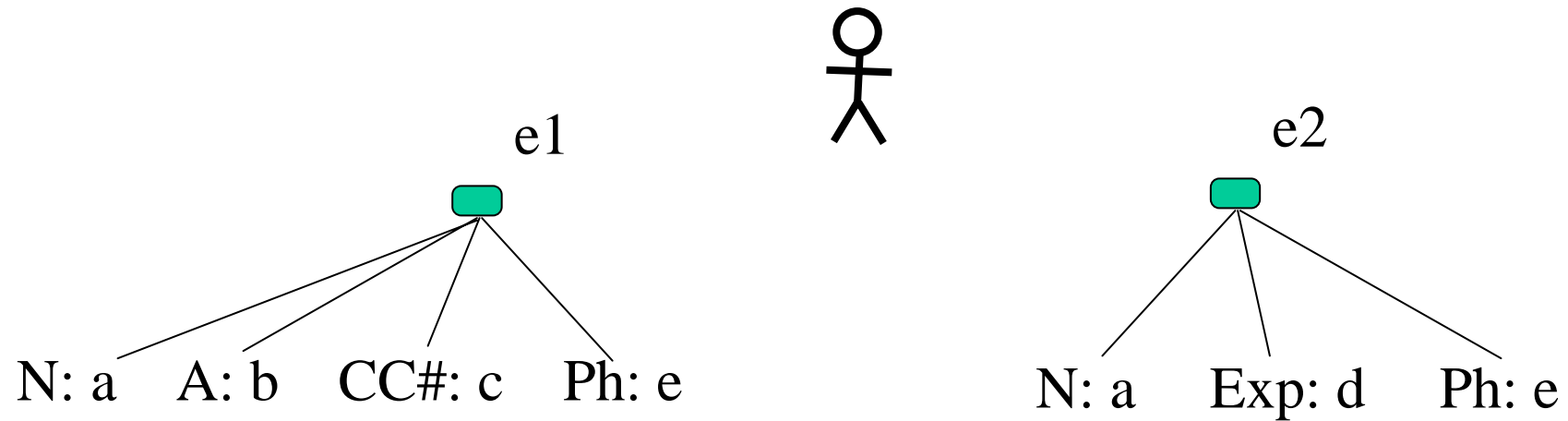
Stanford University

Work with: Omar Benjelloun, Qi Su, Jennifer
Widom, Tyson Condie, Nicolas Pombourcq

“Reverse Talk”

- Entity Resolution Problem
- Confidences
- Two Ideas
- Ask YOU for ideas!

Entity Resolution





- Applications:

- mailing lists, customer files, counter-terrorism, ...

Challenges (1)

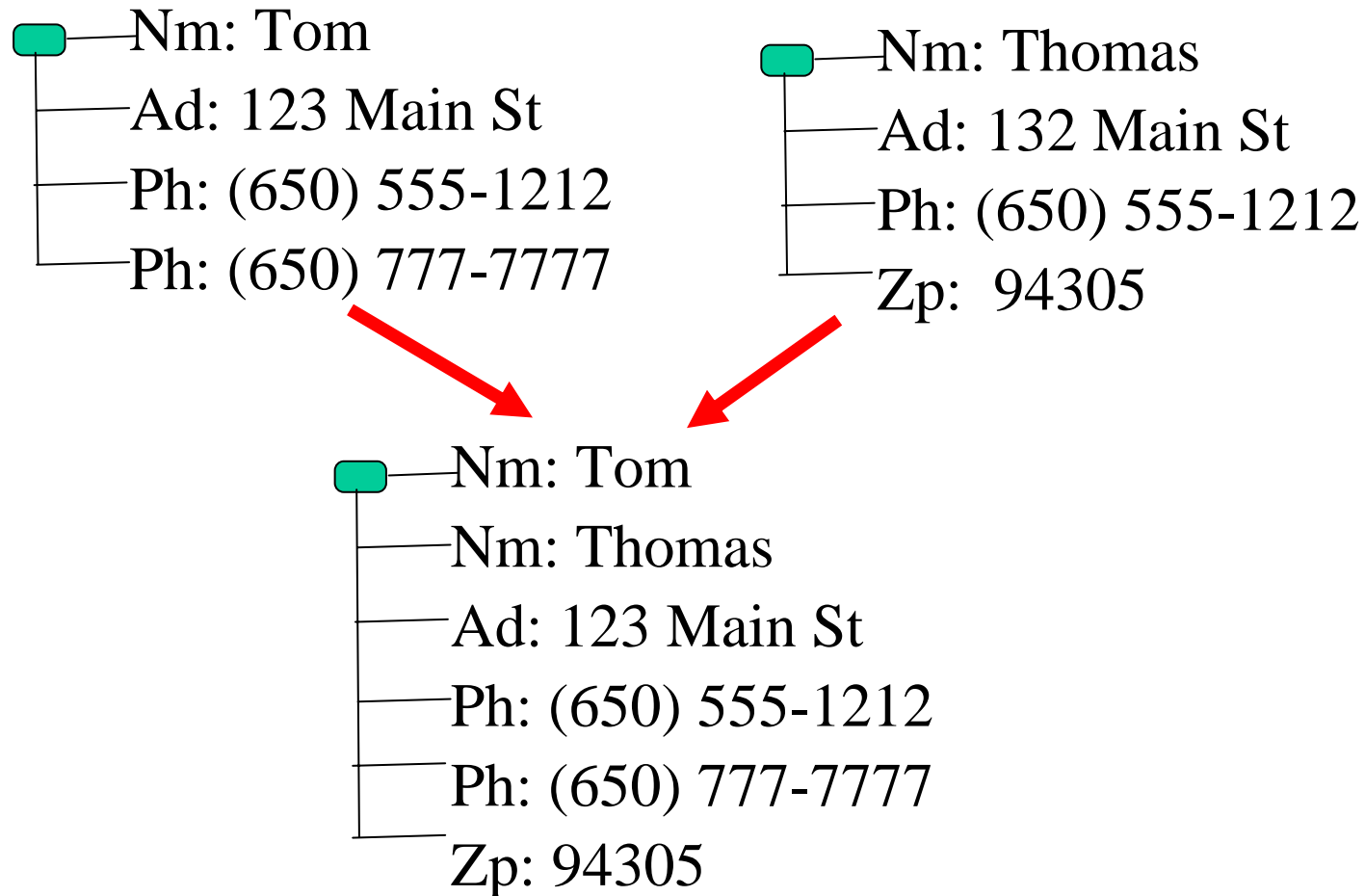
- No keys!
- Value matching
 - “Kaddafi”, “Qaddafi”, “Kadafi”, “Kaddaffi”...
- Record matching

—Nm: Tom
—Ad: 123 Main St
—Ph: (650) 555-1212
—Ph: (650) 777-7777

—Nm: Thomas
—Ad: 132 Main St
—Ph: (650) 555-1212

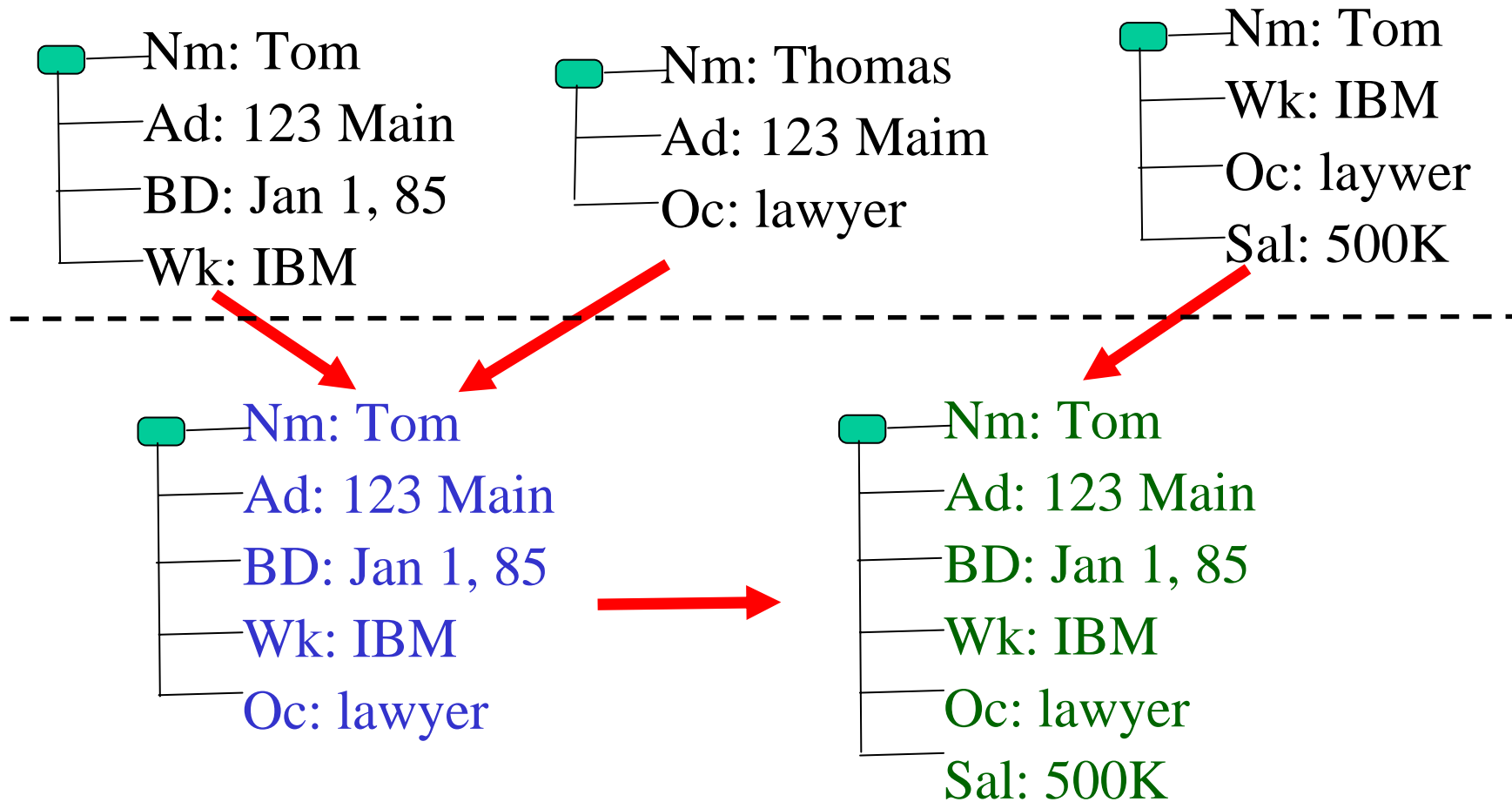
Challenges (2)

- Merging records



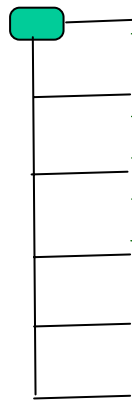
Challenges (3)

- Chaining



Challenges (4)

- Un-merging

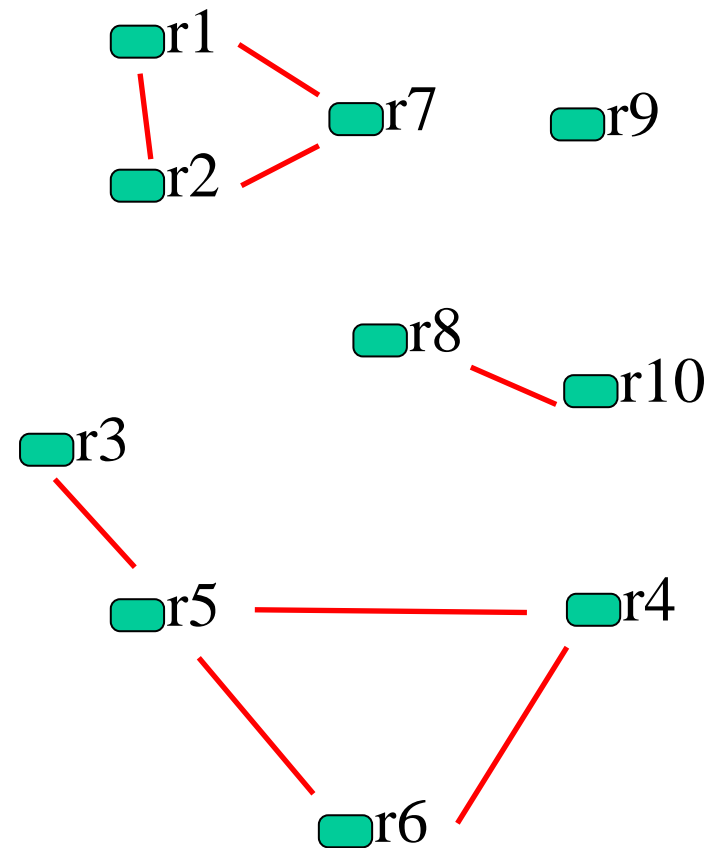
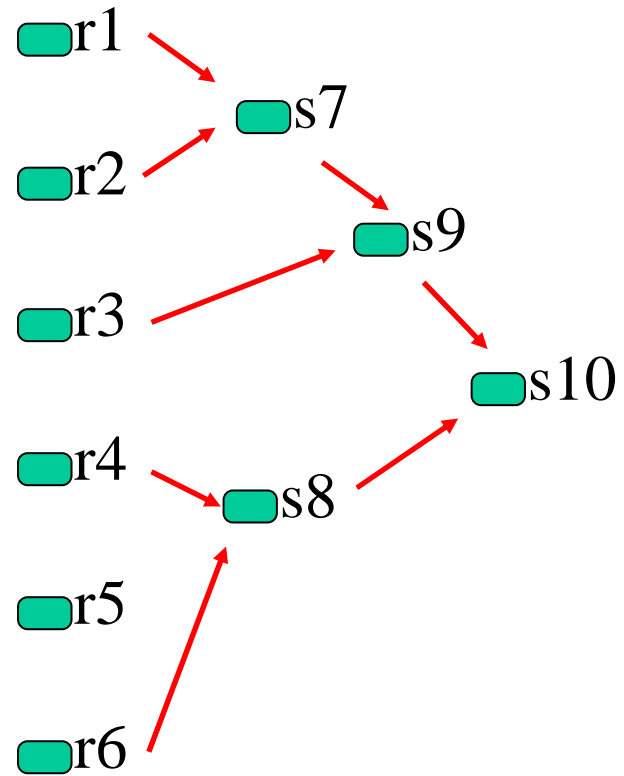
 Nm: Tom
Ad: 123 Main
BD: Jan 1, 85
Wk: IBM
Oc: lawyer
Sal: 500K

too young to make
500K at IBM!!

Taxonomy

- Pairwise snaps vs. clustering
- De-duplication vs. fidelity enhancement
- Schema differences
- Relationships
- Exact vs. approximate
- Generic vs application specific
- Confidences

Pair-Wise Snaps vs. Clustering



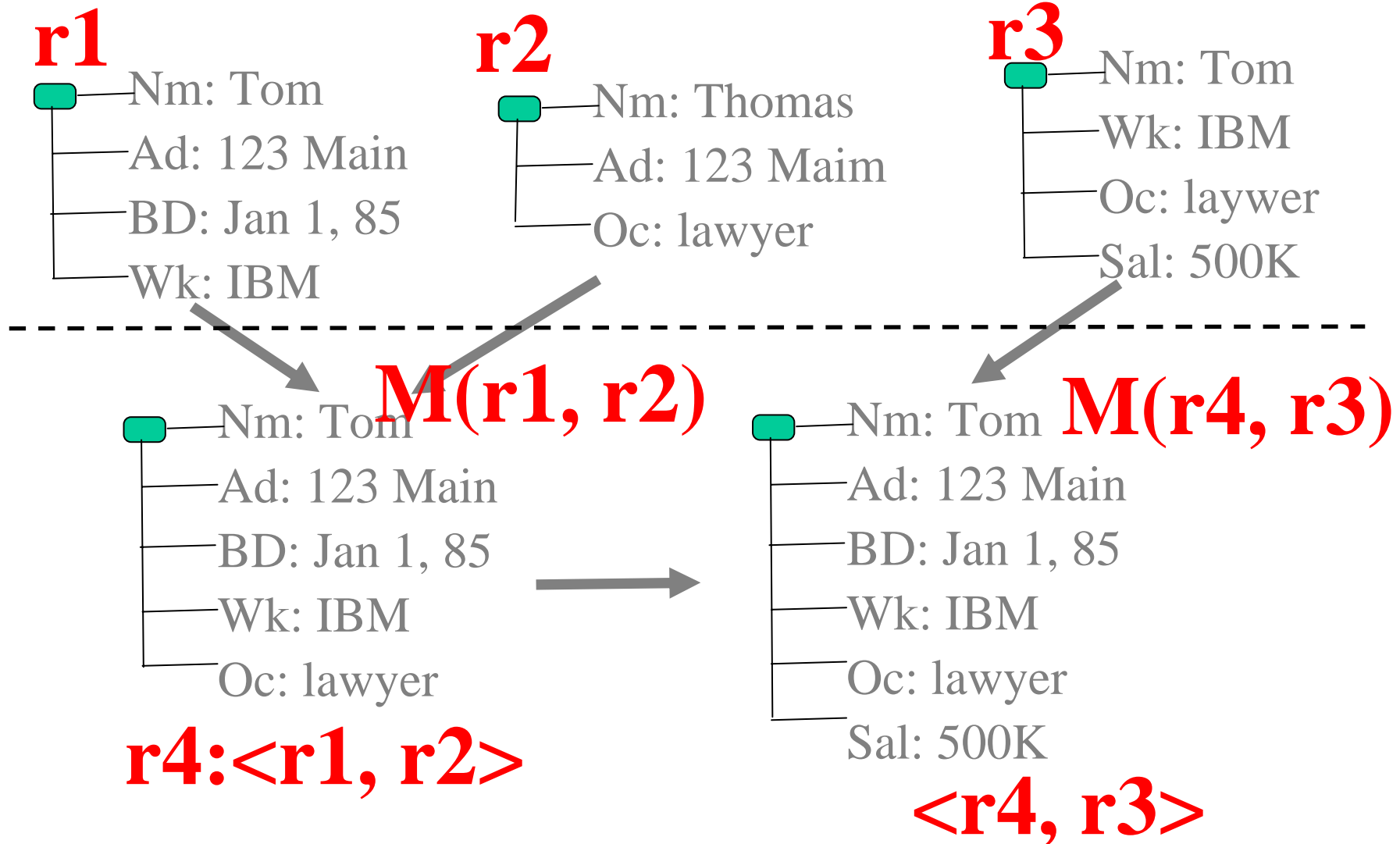
Taxonomy

- Pairwise snaps vs. clustering
- De-duplication vs. fidelity enhancement
- Schema differences
- Relationships
- Exact vs. approximate
- Generic vs application specific
- Confidences

Taxonomy

- Pairwise snaps vs. clustering
- De-duplication vs. fidelity enhancement
- Schema differences No
- Relationships No
- Exact vs. approximate
- Generic vs application specific
- Confidences ... later on

Model



Example

- Records:
 - $r1 = [a: \{1, 2\}, b:2, c:\{5,6\}]$,
 - $r2 = [a:3, b:\{1, 2\}, c:5, d:8], \dots$
- Features:
 - $F1 = \{a, b\}, F2 = \{c\}$
- Match function:
 - $M(r1, r2) = M_{F1} \vee M_{F2}$
- Merge Function:
 - $\langle r1, r2 \rangle = [a:\{1, 2, 3\}, b:\{1, 2\}, c:\{5,6\}, d:8]$

Question

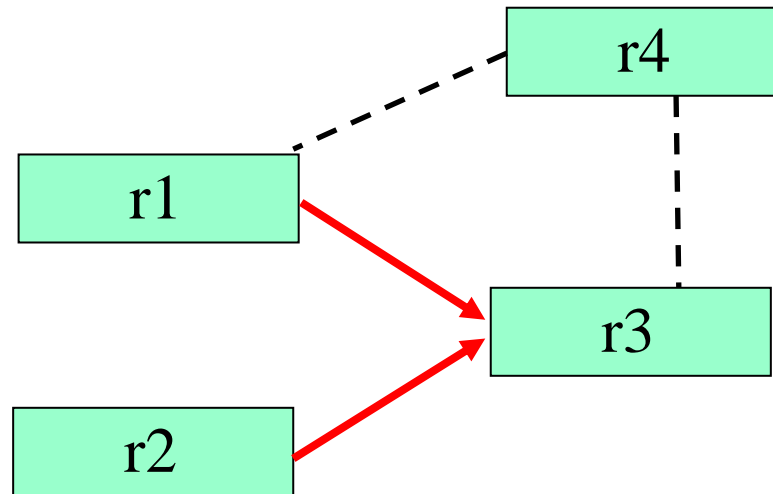
- What is best sequence of match, merge calls that give us right answer?

Properties

- Commutativity:
 - $M(r1, r2) = M(r2, r1)$
 - $\langle r1, r2 \rangle = \langle r2, r1 \rangle$
- Idempotence:
 - $M(r1, r1) = \text{true}; \langle r1, r1 \rangle = r1$
- Associativity
 - $\langle r1, \langle r2, r3 \rangle \rangle = \langle \langle r1, r2 \rangle, r3 \rangle$

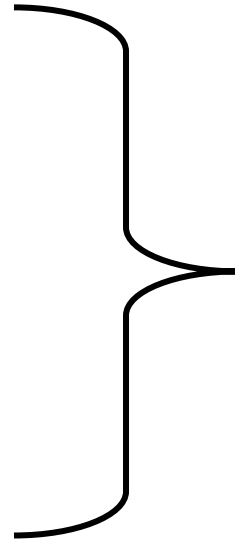
More Properties

- Representativity
 - If $\langle r1, r2 \rangle = r3$, then
for any $r4$ such that $M(r1, r4)$ is true
we also have $M(r3, r4) = \text{true}$.



4 Properties → Efficiency

- Commutativity
- Idempotence
- Associativity
- Representativity



- ER result is unique
- ER result independent of processing order

Example

- Feature F1: {a}
- [a: v1, b: w1]
- [a: v2, b: w2]
- [a: v3, b: w3]
- ...
- [a: vn, b: wn]

$M(r_i, r_j) = \text{True}$

answer: [a:{v1, ...,vn}, b:{w1, ..., wn}]

Brute Force Algorithm

not_done := true

while not_done do

[not_done := false;

RP := RN := empty set;

for each (r_i, r_j) in R, s.t. $r_i \neq r_j$ do

if M (r_i, r_j) then

not_done := true;

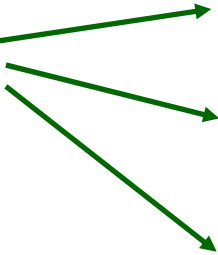
RP = RP union $\{ \langle r_i, r_j \rangle \}$

RN = RN union $\{ r_i, r_j \}$

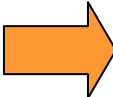

R := R union RP minus RN]

RP = new records
RN = no longer needed
records

Back to Example...

- [a: v1, b: w1]
 - [a: v2, b: w2]
 - [a: v3, b: w3]
 - [a: v4, b: w4]
- 
- [a: {v1,v2}, b: {w1,w2}]
 - [a: {v1,v3}, b: {w1,w3}]
 - [a: {v1,v4}, b: {w1,w4}]
 - [a: {v2,v3}, b: {w2,w3}]
 - [a: {v2,v4}, b: {w2,w4}]
 - [a: {v3,v4}, b: {w3,w4}]

Example Continued...

- [a:{ v1,v2 }, ...]
 - [a:{ v1,v3 }, ...]
 - [a:{ v1,v4 }, ...]
 - [a:{ v2,v3 }, ...]
 - [a:{ v2,v4 }, ...]
 - [a:{ v3,v4 }, ...]
- 
- [a:{ v1,v2,v3 }, ...]
 - [a:{ v1,v2,v4 }, ...]
 - [a:{ v2,v3,v4 }, ...]
 - [a:{ v1,v2,v4 }, ...]
- 
- [a:{ v1,v2,v3,v4 }, ...]

... A lot of useless work!

Swoosh Algorithms

- Record Swoosh
 - Merges records as soon as they match
 - Optimal in terms of record comparisons
- Feature Swoosh
 - Remembers values seen for each feature
 - Avoids redundant value comparisons

Swoosh Example

- [a: v1, b: w1]

- [a: v2, b: w2]

- [a: v3, b: w3]

- [a: v4, b: w4]

- $M(r1, r2) \rightarrow$

[a: {v1, v2}, ...]

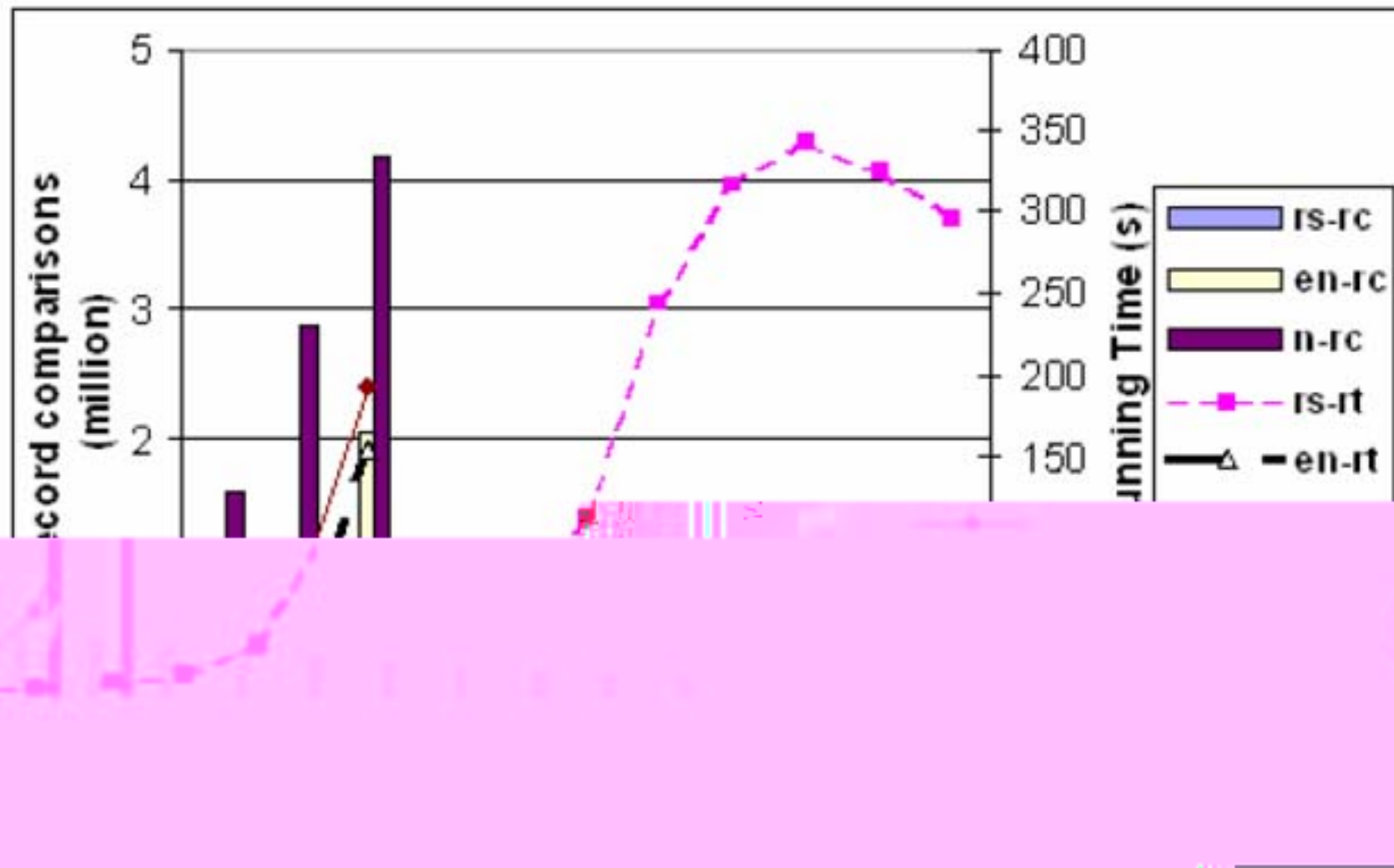
- $M(r3, r12) \rightarrow$

[a: {v1, v2, v3}, ...]

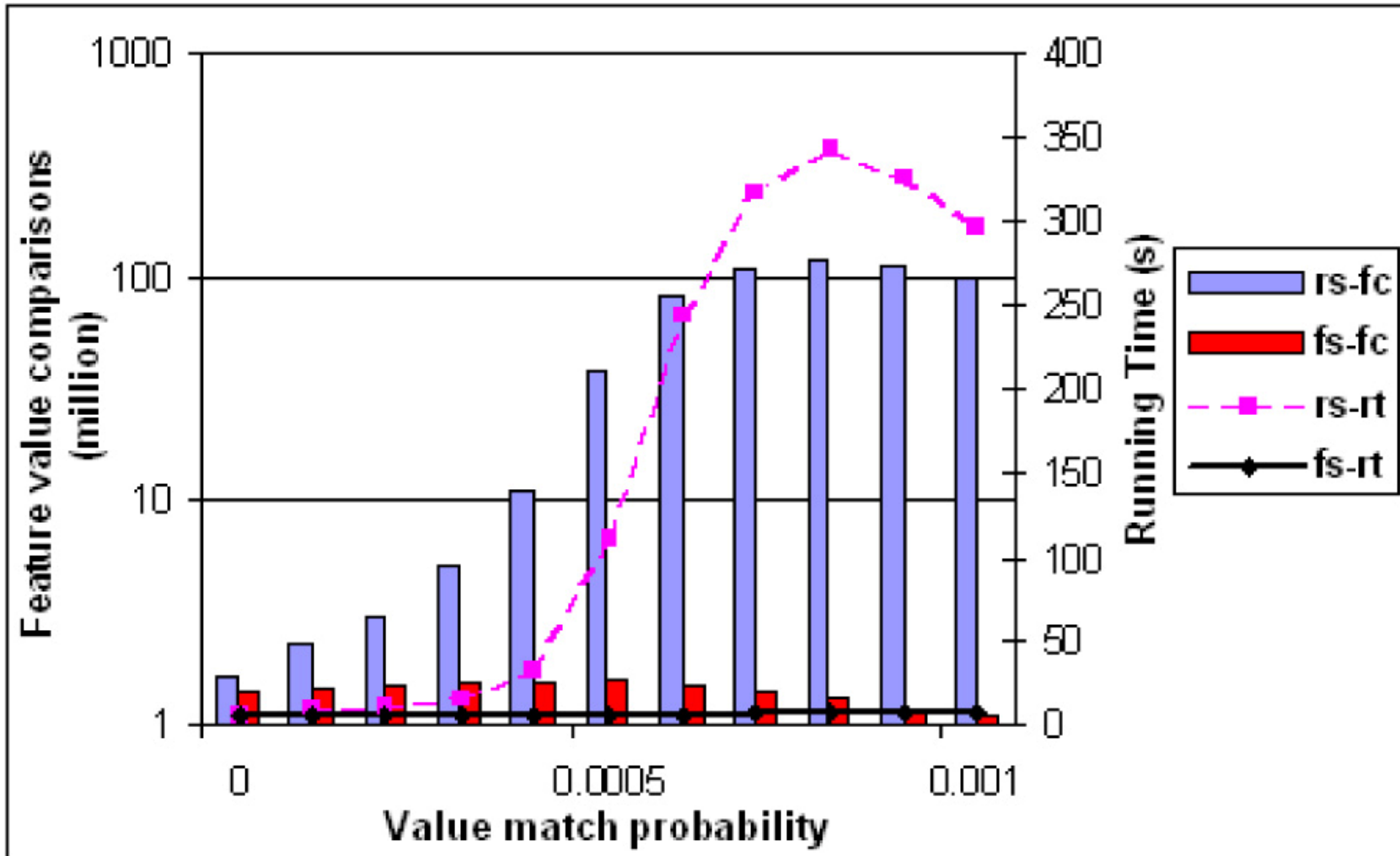
- $M(r4, r123) \rightarrow$

[a: v1, a: v2, a: v3, a: v4, ...]

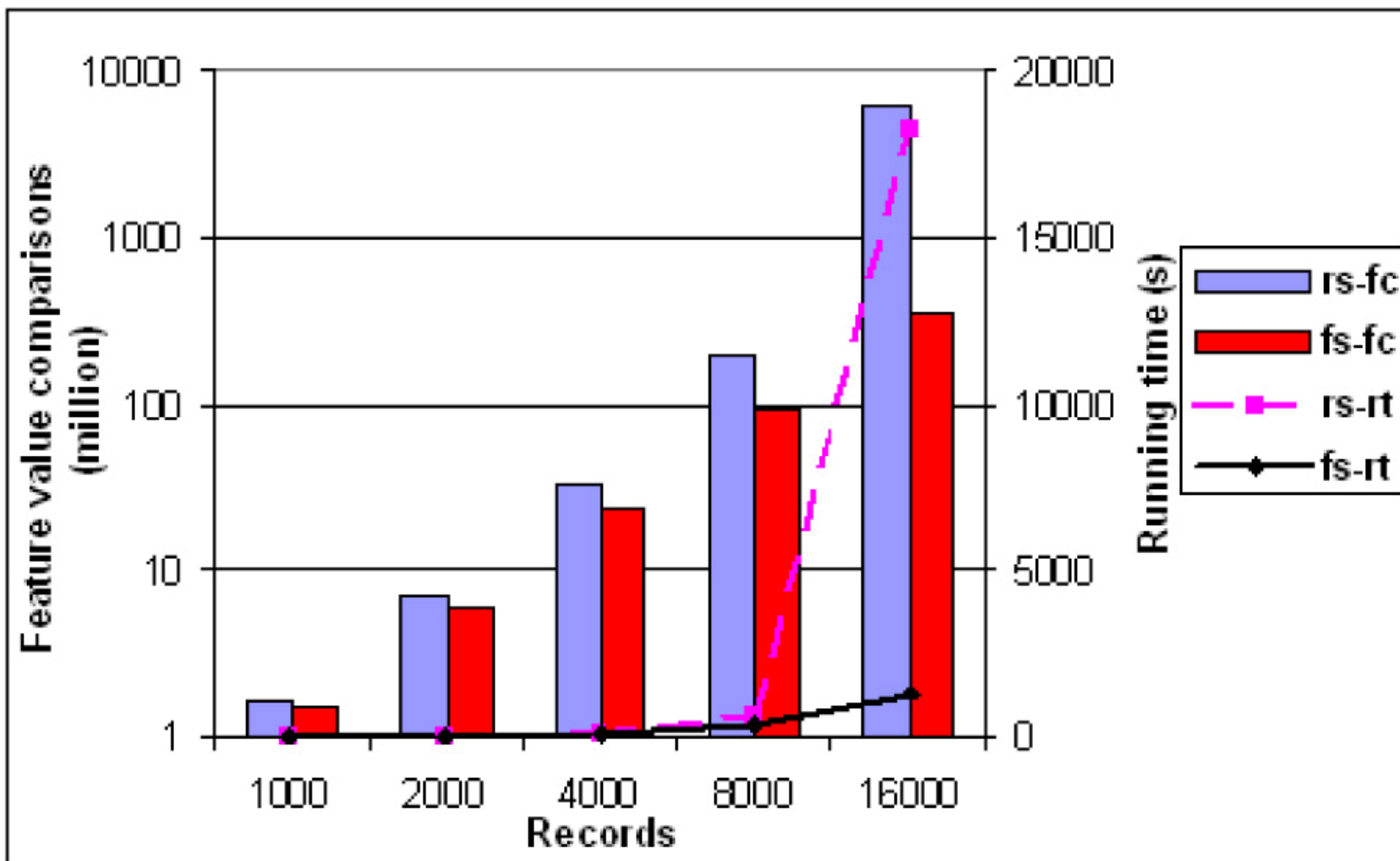
Swoosh Performance (I)



Swoosh Performance (II)

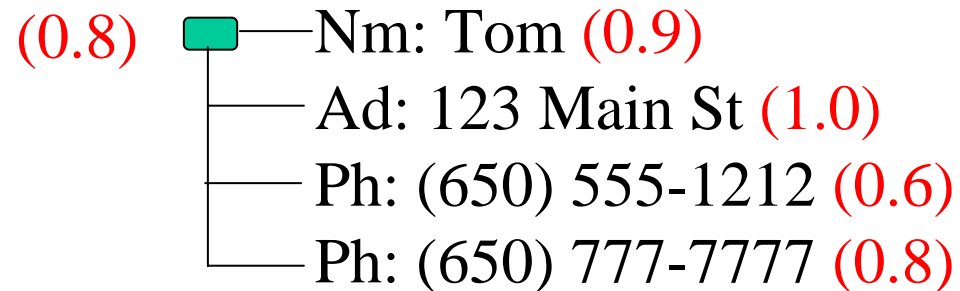


Swoosh Performance (III)



Confidences

- In data



- In value matching:

$$\text{sim}(\text{Qadafi}, \text{Kadafi}) = 0.95$$

- In match rules:

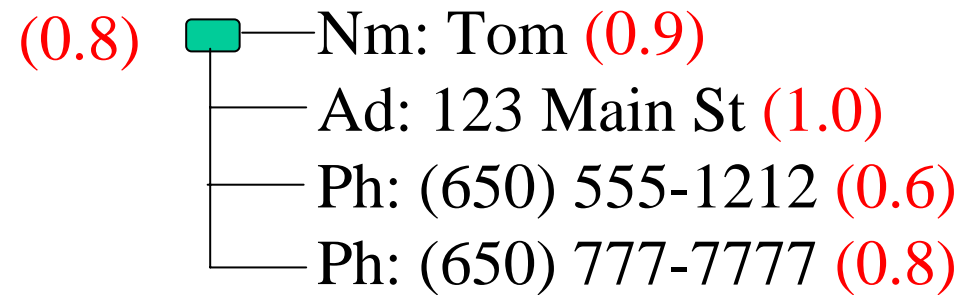
$$M(\text{r1}, \text{r2}) = T (0.9)$$

- In merge/fusion:

$$\text{Merge}(\text{Héctor}, \text{Ettore}) = \text{Hector} (0.8)$$

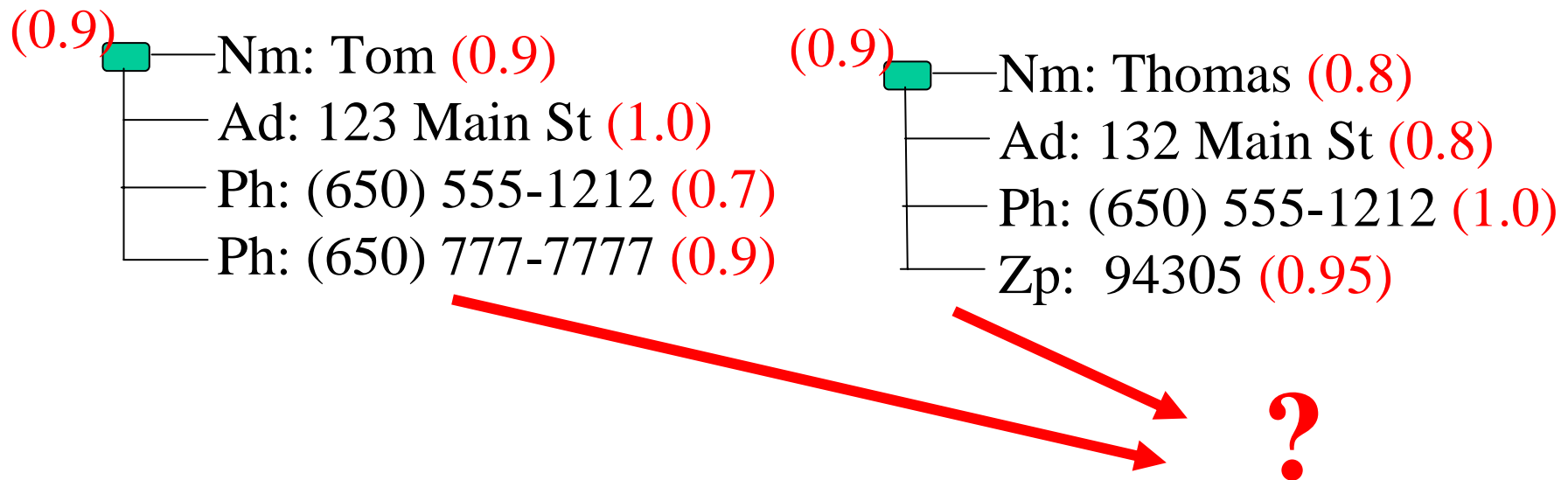
Challenges

- What do confidences mean?



Challenges

- How do we operate on confidences?



$$\text{sim}(\text{Tom}, \text{Thomas}) = 0.92$$

$$M_{F\{\text{Nm}, \text{Ad}\}} = 0.88$$

...

One Confidence Model



[id1, a, b, c, d]



[id2, a, c, e]



[id3, a, b, f, g]

[id1, a, b, c, d]

[id1, a, b, d]

[id1, a, x]

[id1, b, y]

[id2, a, b, c]

[id2, a, c, e]

[id2, a, c, e]

[id2, a, c, e]

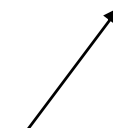
[id3, a, b, c]

[id3, a, b, d]

[id3, a, b, f, g]

[id3, a, b, f, g]

shorthand



Records Are Evidence



[id1, a, b, c, d]

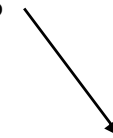
[id1, a, b, c, d]

[id1, a, b, d]

[id1, a, x]

[id1, b, y]

not 0.25

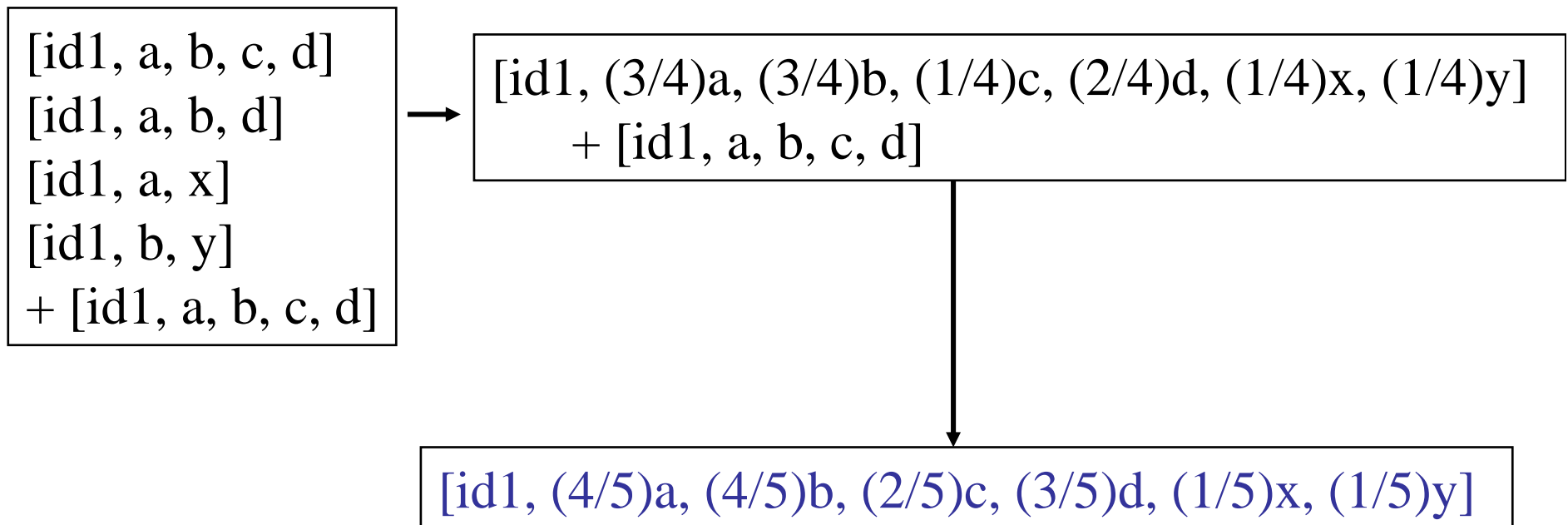


[id1, (3/4)a, (3/4)b, (1/4)c, (2/4)d, (1/4)x, (1/4)y]

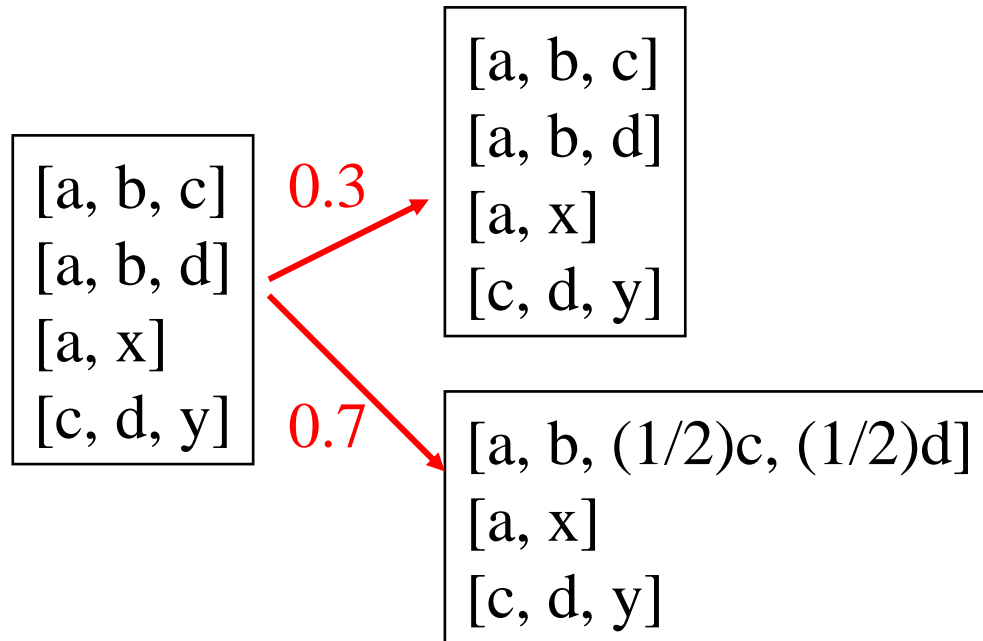
New Evidence



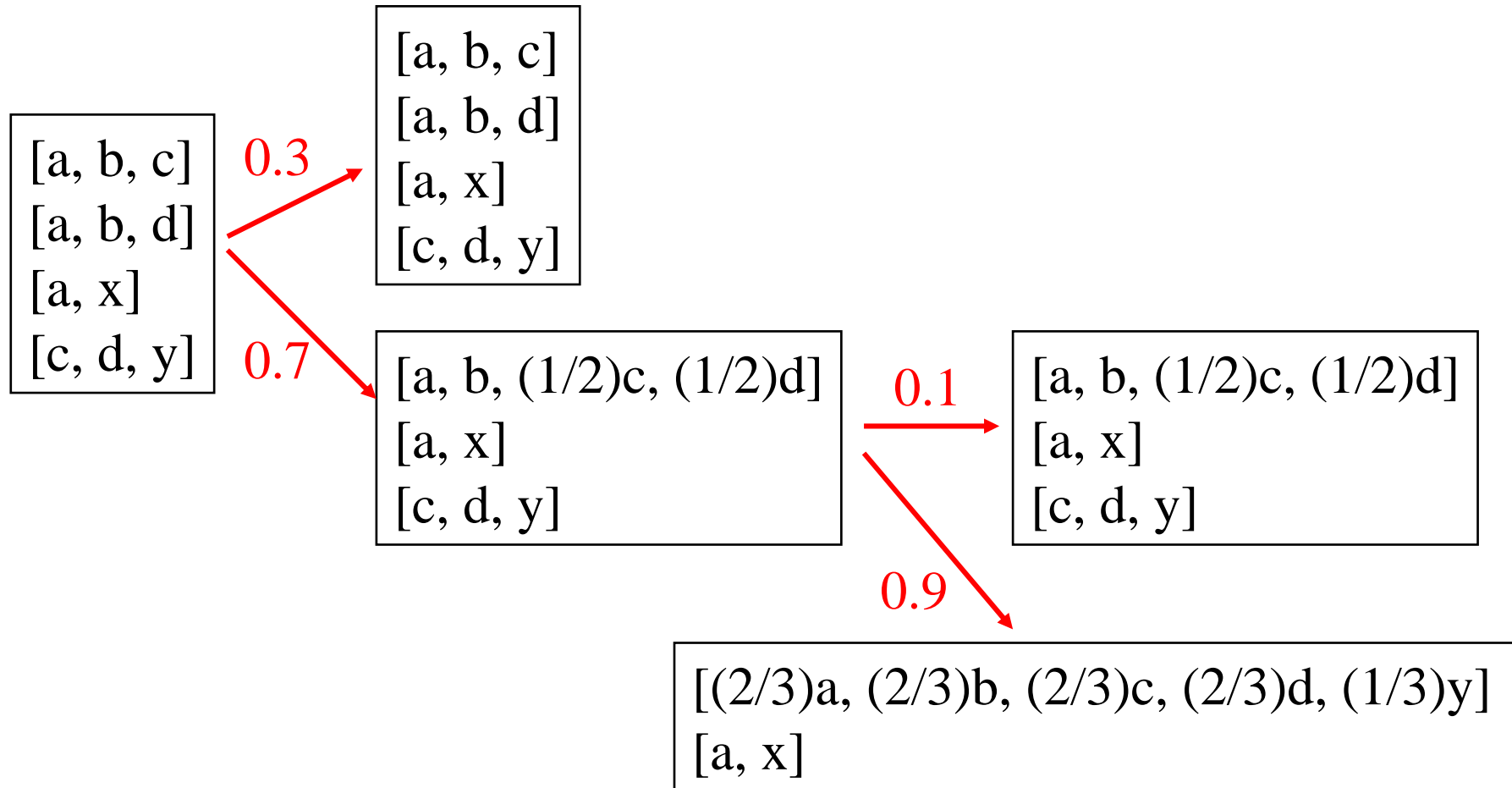
[id1, a, b, c, d]



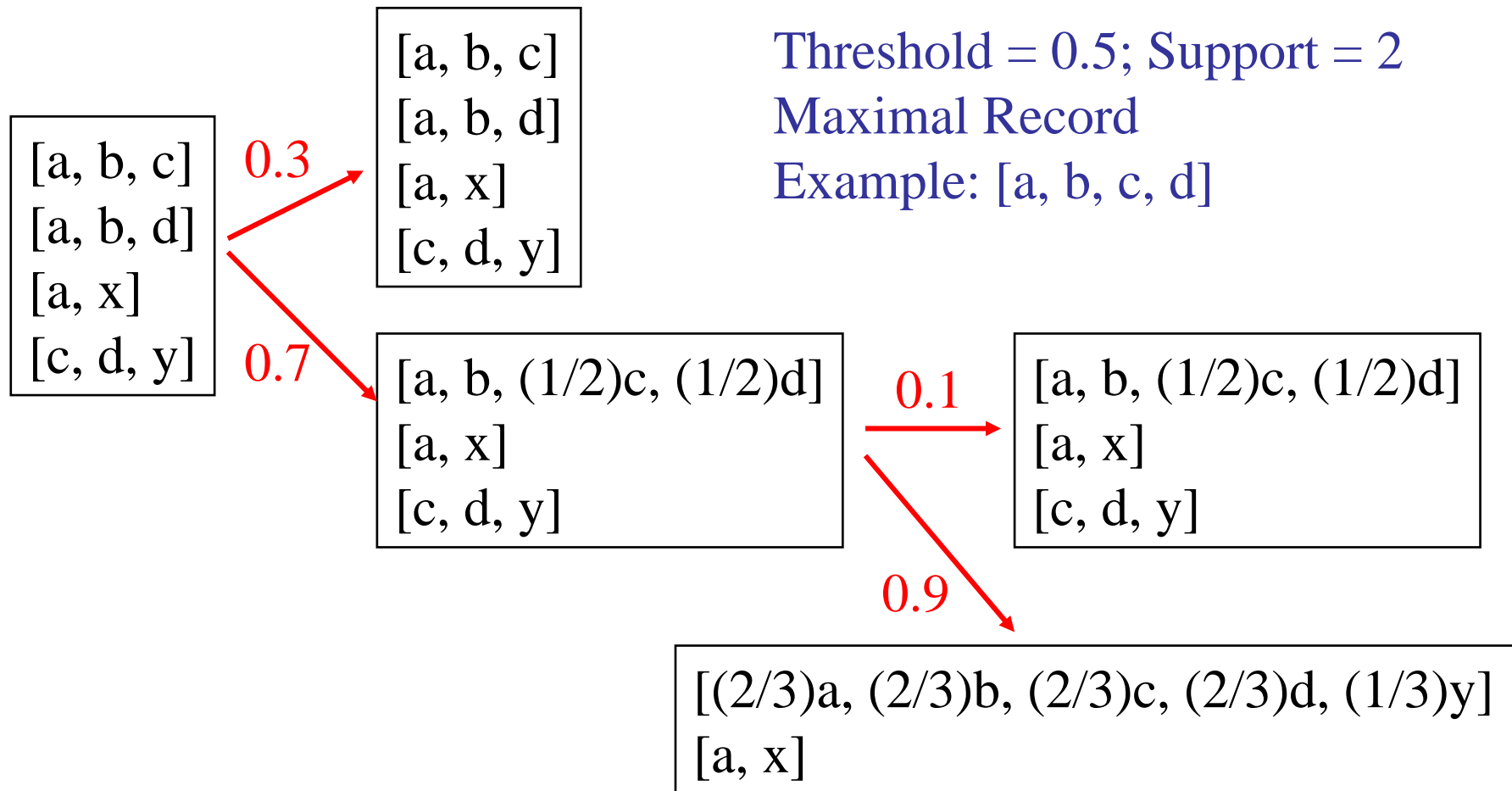
No Ids



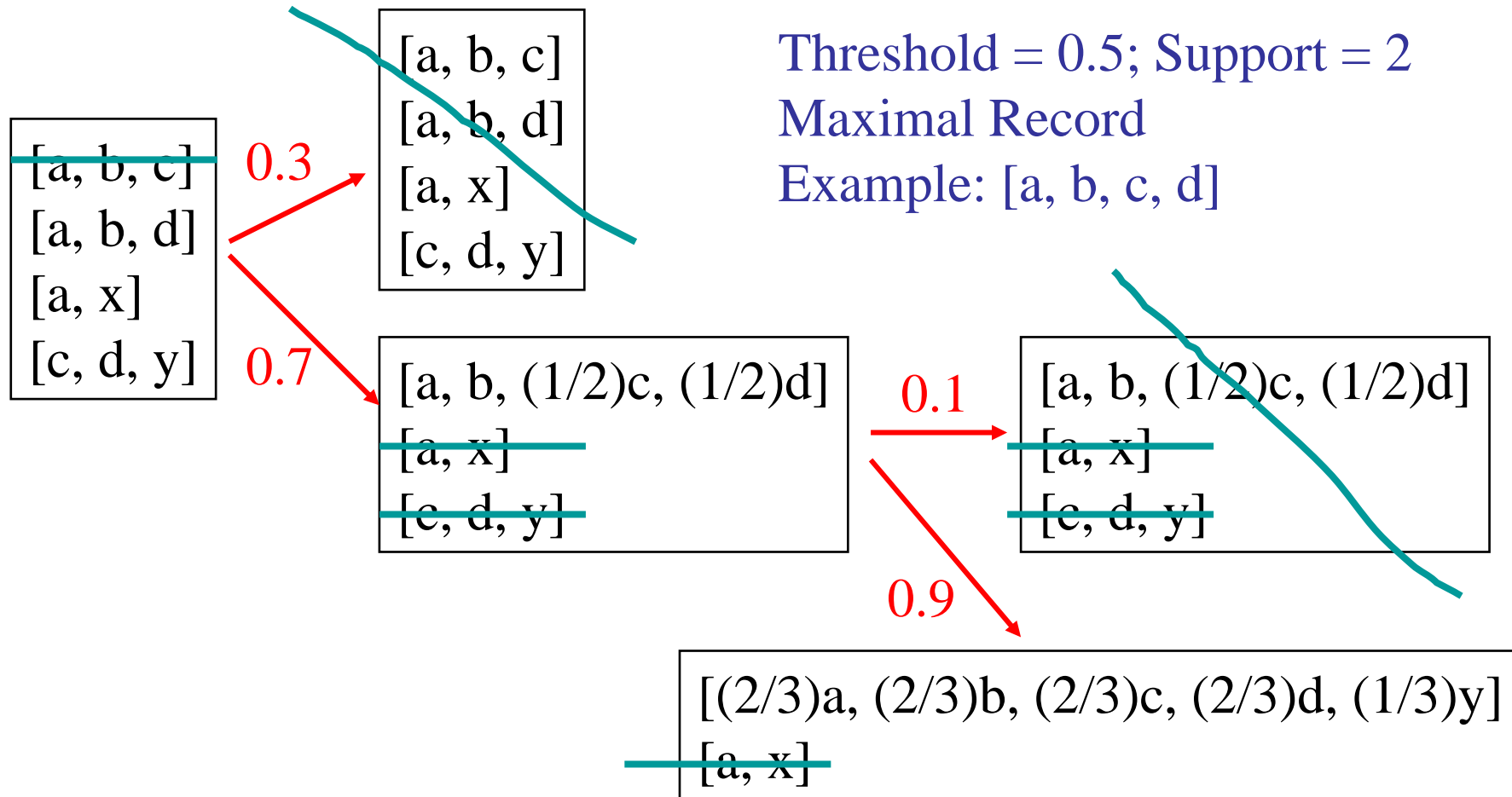
No Ids



Queries?



Queries?

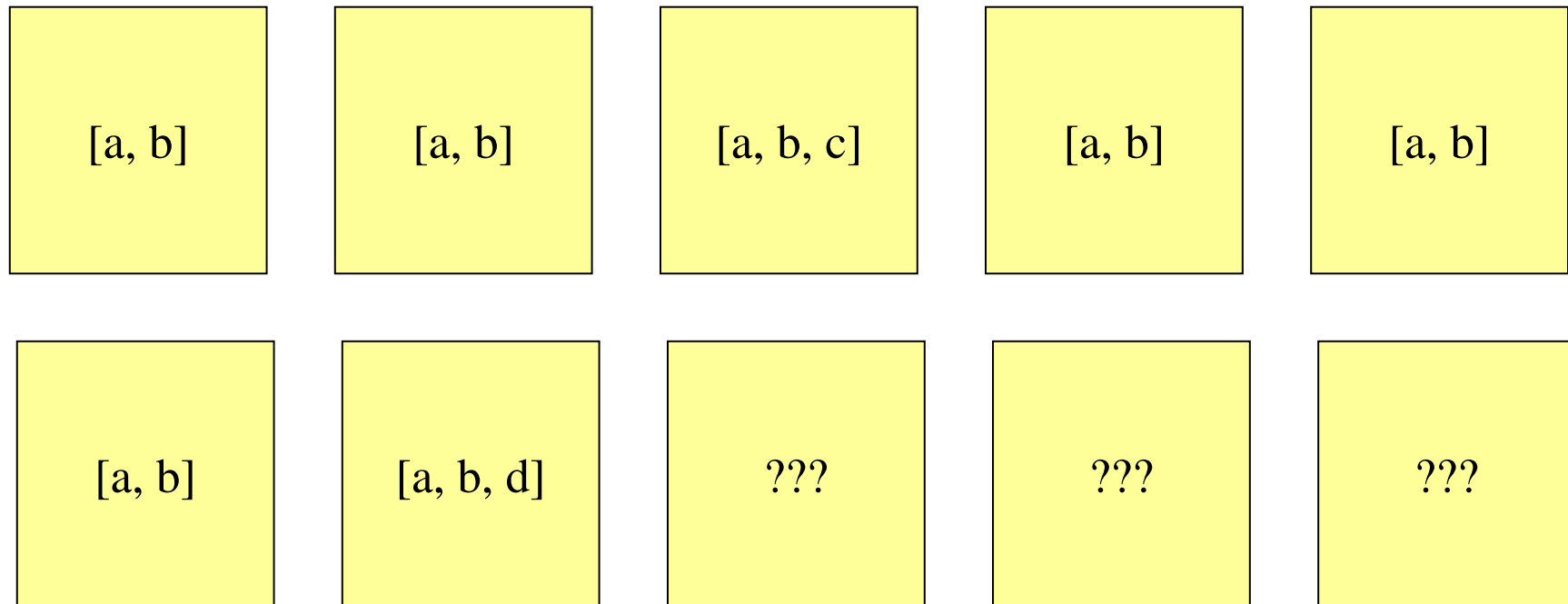


Need Simpler Model?

Simple Confidence Model

- 0.7 [a, b]

Alternate Worlds:



Rules

- $0.7[a, b, c], 0.7[a, b, c]$
 $\Rightarrow 0.7 [a, b, c]$
- $0.7 [a, b], 0.5 [a, b]$
 $\Rightarrow 0.7 [a, b]$
- $0.7 [a, b, c], 0.5 [a, b]$
 $\Rightarrow 0.7 [a, b, c]$
- $0.7 [a, b, c], 0.9[a, b]$
 $\Rightarrow 0.7 [a, b, c], 0.9[a, b]$
- etc

Matches

0.9[a, b, c]

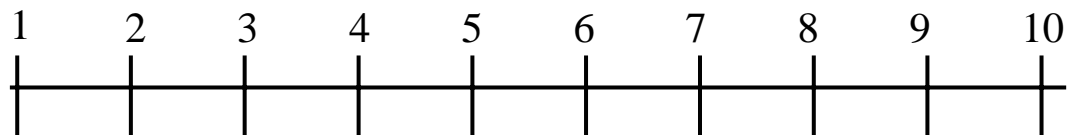
0.8[a, b, d]

[a, x]

[c, d, y]

Match with confidence 0.5

worlds

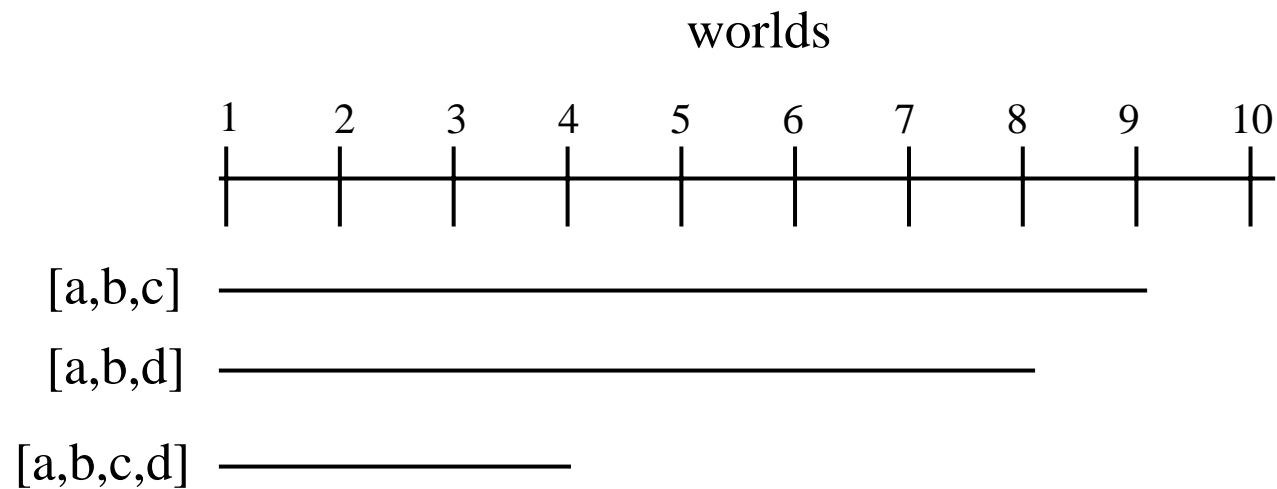
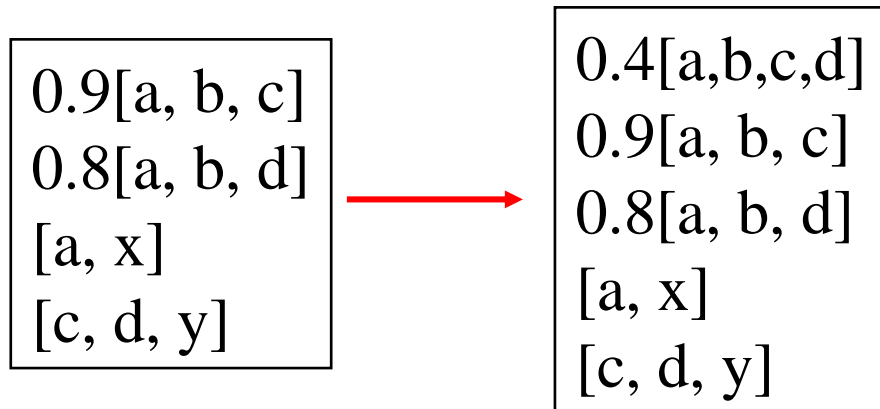


[a,b,c] _____

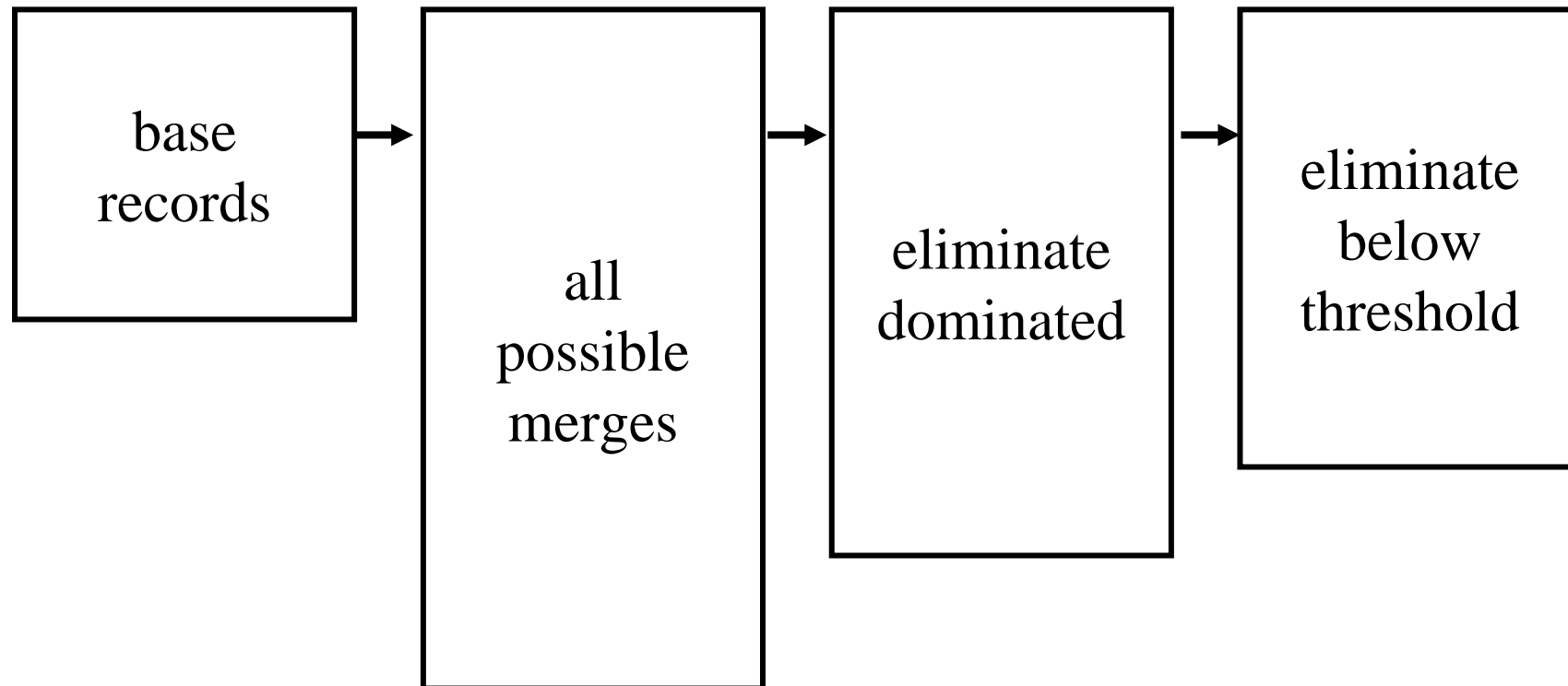
[a,b,d] _____

[a,b,c,d] _____

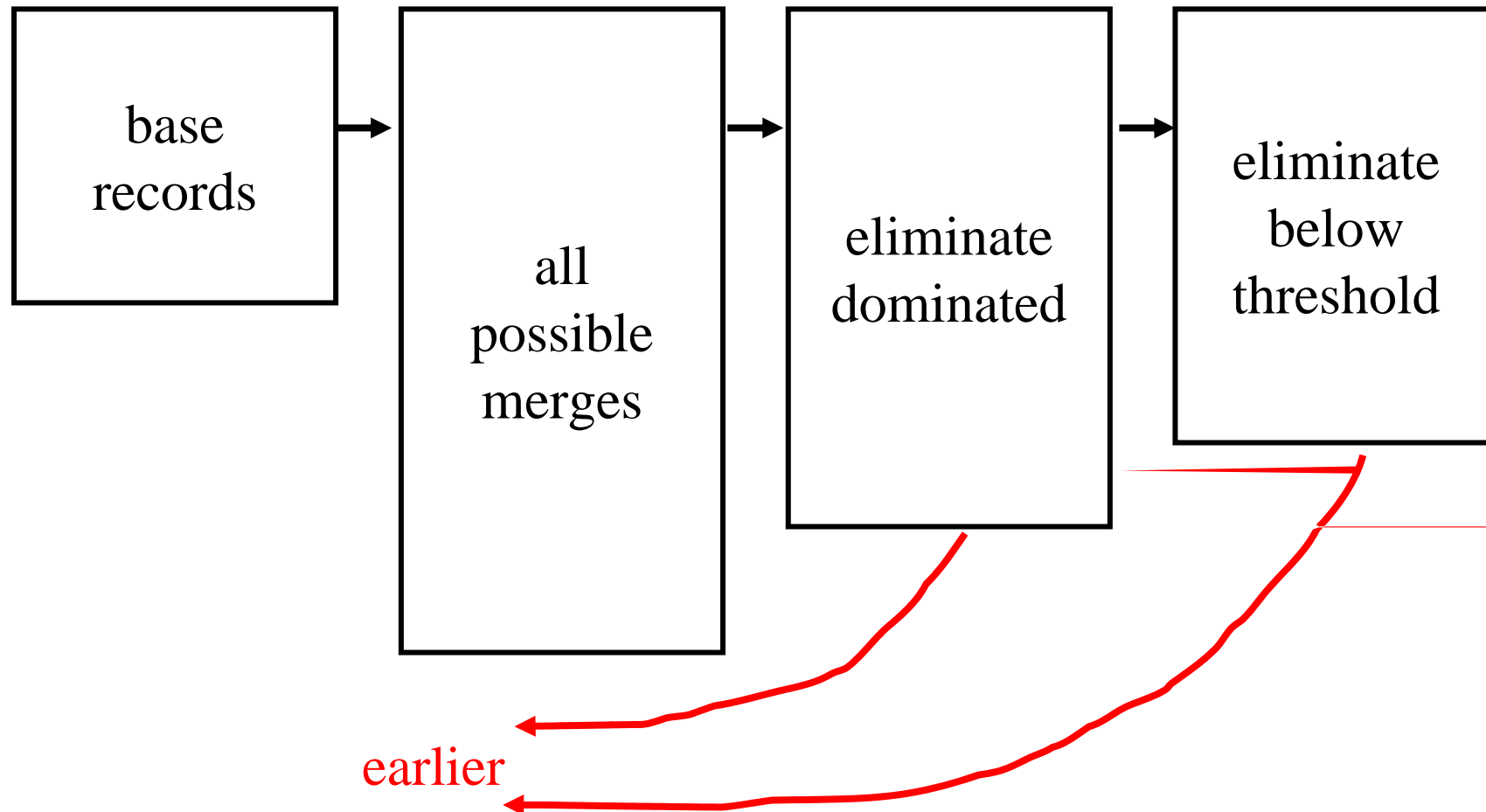
Matches



Goal: C-Swoosh



Goal: C-Swoosh



Questions

- Each model has drawbacks...
- Is there a better confidence model??

Summary

- Entity resolution is critical
 - Efficient resolution important
 - Confidences are important, but how?
 - ER is key aspect of info privacy
- check www-db.stanford.edu for
Swoosh paper & forthcoming paper

Thanks.