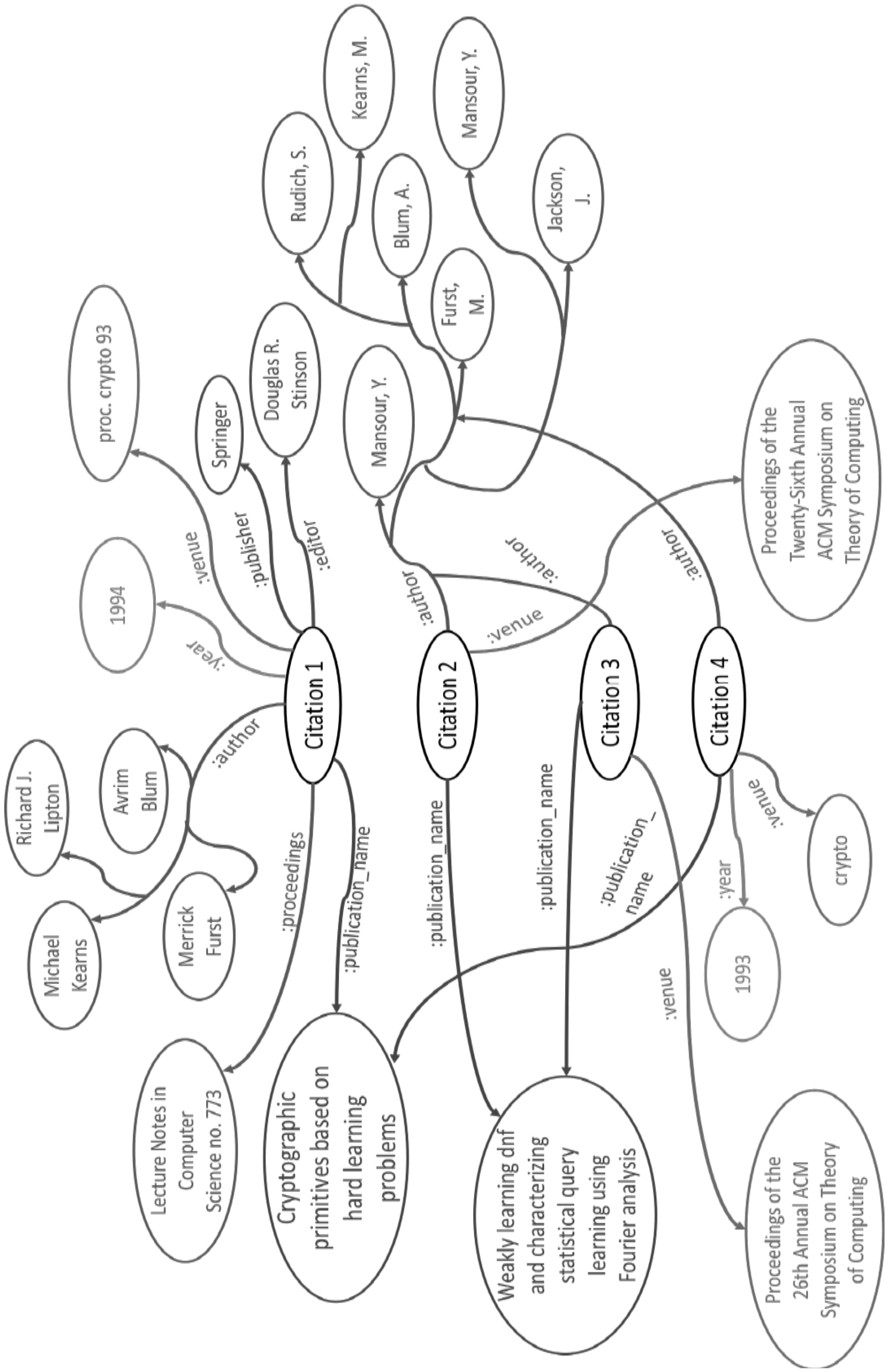
# RECORD LINKAGE

We will use the following running example



a) Group the four citations into clusters of matching records

b) What are some features you could use (think string similarity and quantitative functions) to make such a determination automatically? Is possible, list several such functions and a way in which you could potentially combine them.

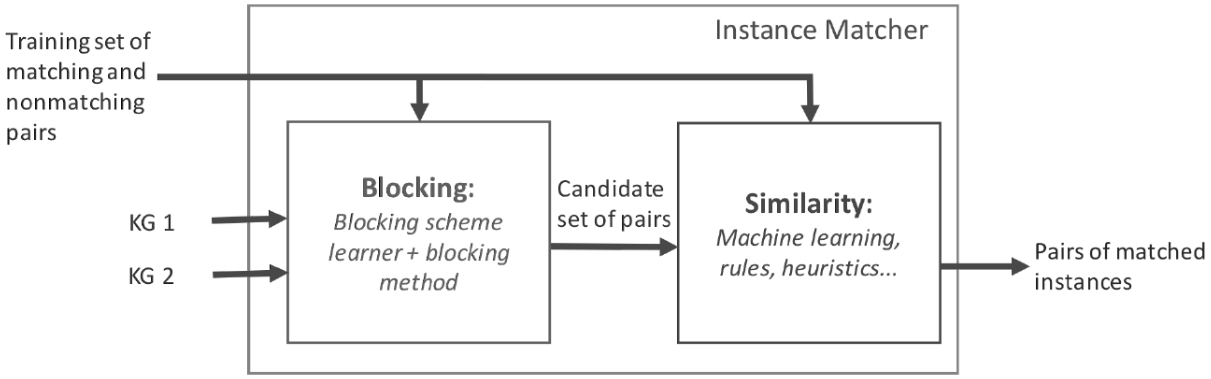
c) Think about cases where your 'algorithm' above could go wrong. List specific examples.

d) Define a pairwise linking function.

e) Define reflexivity, symmetry and transitivity. In real world linking functions, which of the properties are typically fulfilled?

f) List specific reasons why record linkage is challenging. Give examples if possible.

**Review: Two-step record linkage pipeline**


Training set of matching and nonmatching pairs
Instance Matcher
KG 1
KG 2
Blocking:
Blocking scheme learner + blocking method
Candidate set of pairs
Similarity:
Machine learning, rules, heuristics...
Pairs of matched instances

*Questions:*

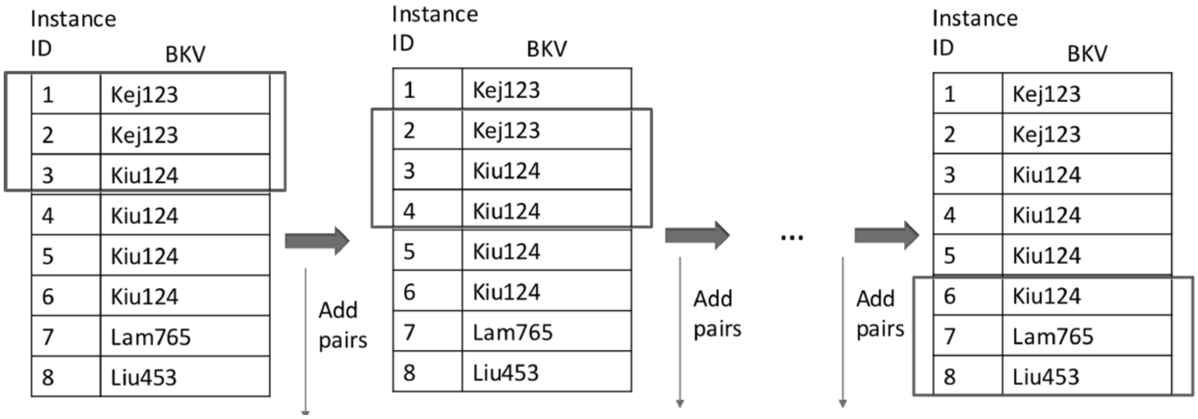
Notice in this pipeline that we refer to the box as 'instance matcher'. What are at least five other terms used to refer to record linkage in the literature?

Describe the blocking step qualitatively, and its purpose.

**Review: Sorted Neighborhood**

Briefly describe the sorted neighborhood blocking algorithm.

What are some limitations of sorted neighborhood? What about traditional blocking?



| Instance ID | BKV |
|---|---|
| 1 | Kej123 |
| 2 | Kej123 |
| 3 | Kiu124 |
| 4 | Kiu124 |
| 5 | Kiu124 |
| 6 | Kiu124 |
| 7 | Lam765 |
| 8 | Liu453 |

| Instance ID | BKV |
|---|---|
| 1 | Kej123 |
| 2 | Kej123 |
| 3 | Kiu124 |
| 4 | Kiu124 |
| 5 | Kiu124 |
| 6 | Kiu124 |
| 7 | Lam765 |
| 8 | Liu453 |

| Instance ID | BKV |
|---|---|
| 1 | Kej123 |
| 2 | Kej123 |
| 3 | Kiu124 |
| 4 | Kiu124 |
| 5 | Kiu124 |
| 6 | Kiu124 |
| 7 | Lam765 |
| 8 | Liu453 |

Using the example above, show how the candidate set of pairs C is getting updated each time we add pairs in the sorted neighborhood algorithm. *Hint: C starts with the empty set before we first start adding pairs.*

Provide the formula for reduction ratio, and explain its purpose. Using the example above, compute the reduction ratio.

**Review: metrics**

**Symbols for all metrics below should be obvious. If not, be sure to review record linkage again*

**Blocking**

$$RR = 1 - \frac{|C|}{|O|}.$$

$$PQ = \frac{|C \cap O_D|}{|C|}.$$

$$PC = \frac{|C \cap O_D|}{|O_D|}.$$

$$FM = \frac{2 \times PC \times RR}{PC + RR}.$$

**Similarity**

$$Precision = \frac{|C_D \cap O_D|}{|C_D|}.$$

$$Recall = \frac{|C_D \cap O_D|}{|O_D|}.$$