LSH for MinHash

2017-03-20

Number of pairs problem

- We know how to compare 2 documents in a short (for H fixed = O(1)) time

But still:

- with 10M documents we have ~10M^2 pairs = few days of computation
- Solution: approximate NNS with LSH

LSH: Candidate pairs

- Get set of candidate pairs

  Number of candidate pairs << number of all pairs

- candidate pair = a pair of elements whose similarity must be evaluated

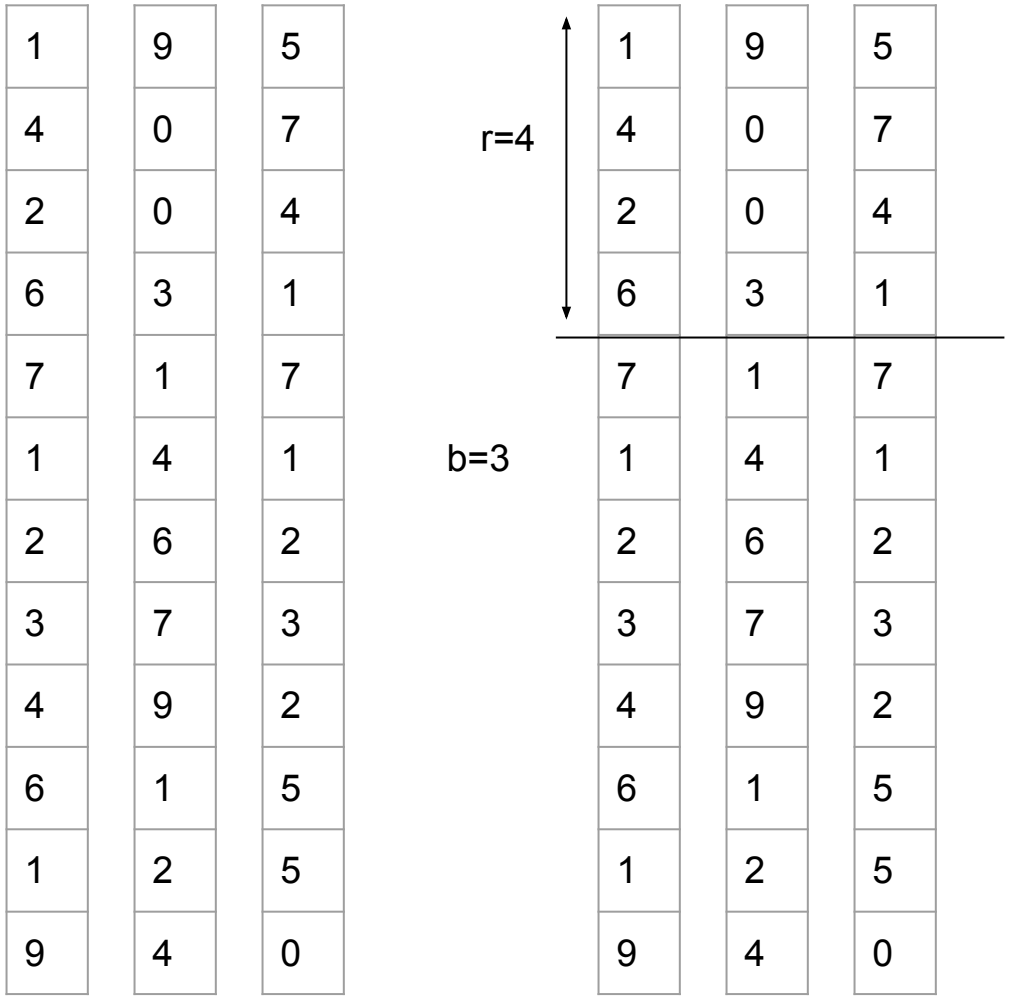
  (to check if they are really similar)

- For some **similarity threshold t** we want
  - Almost all pairs S1, S2 with s(S1,S2) >= t
  - Almost none pairs with s(S1,S2) < t

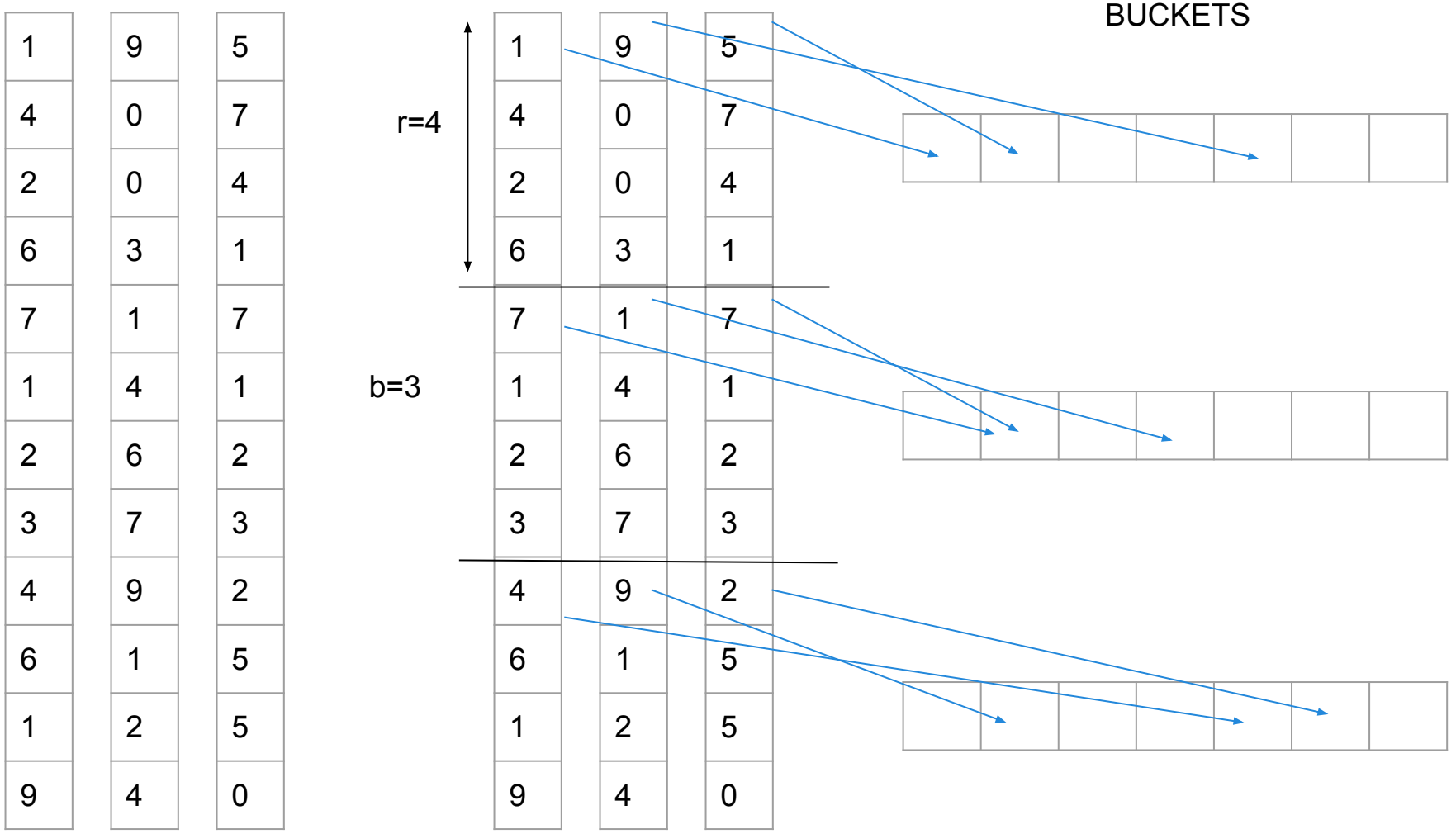
One document = one vector

1	9	5
4	0	7
2	0	4
6	3	1
7	1	7
1	4	1
2	6	2
3	7	3
4	9	2
6	1	5
1	2	5
9	4	0

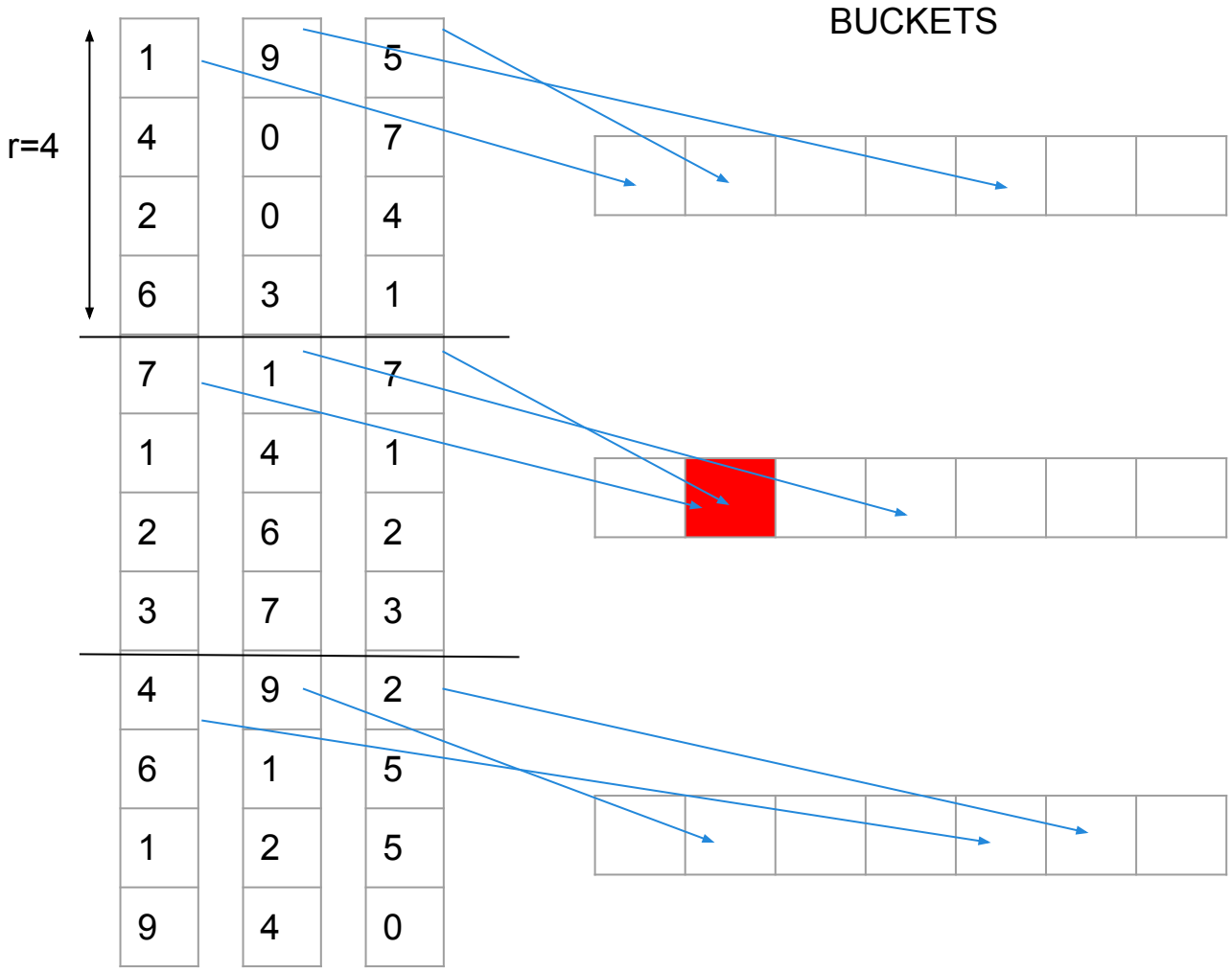
b bands of r rows



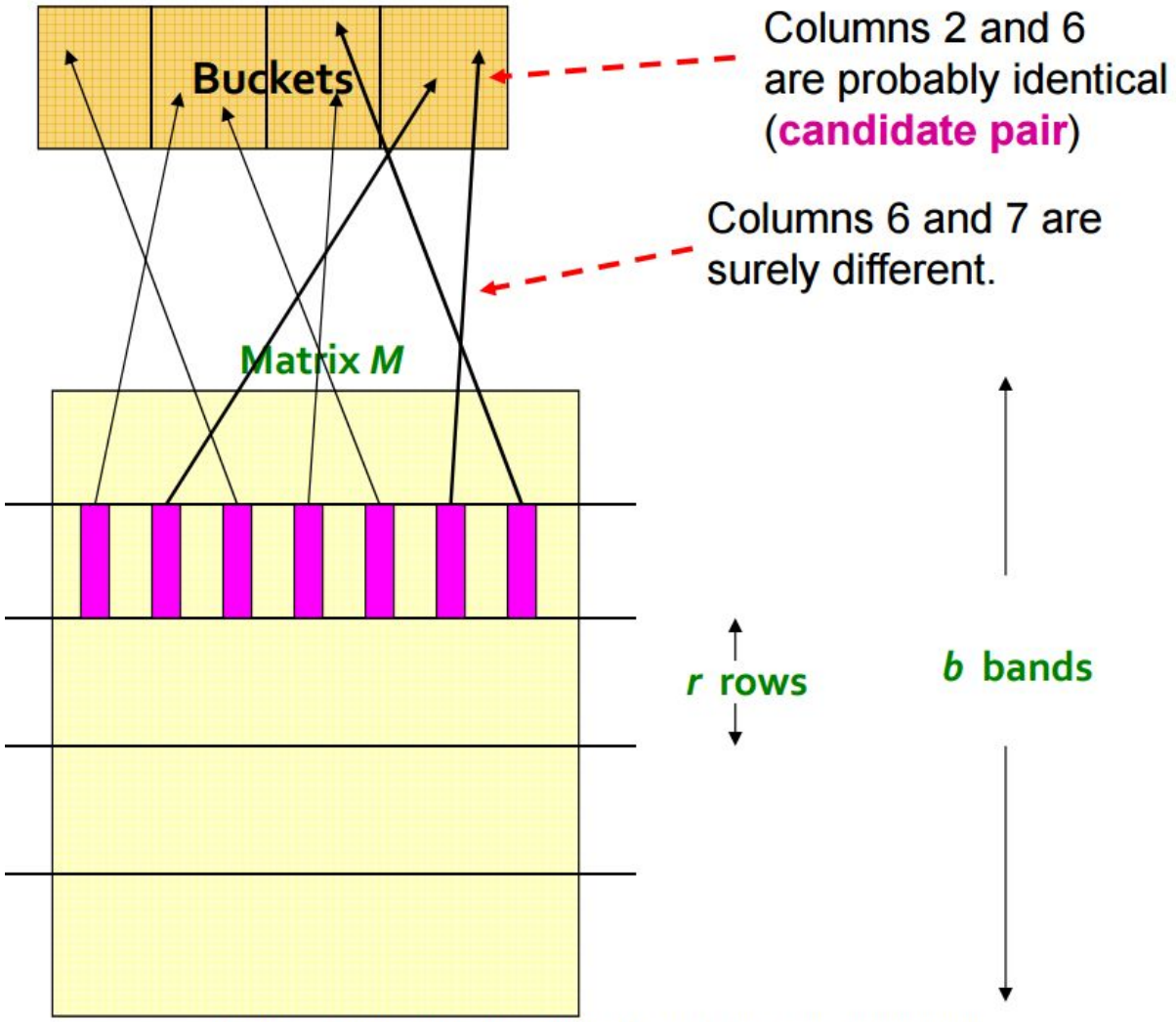
Hashing bands



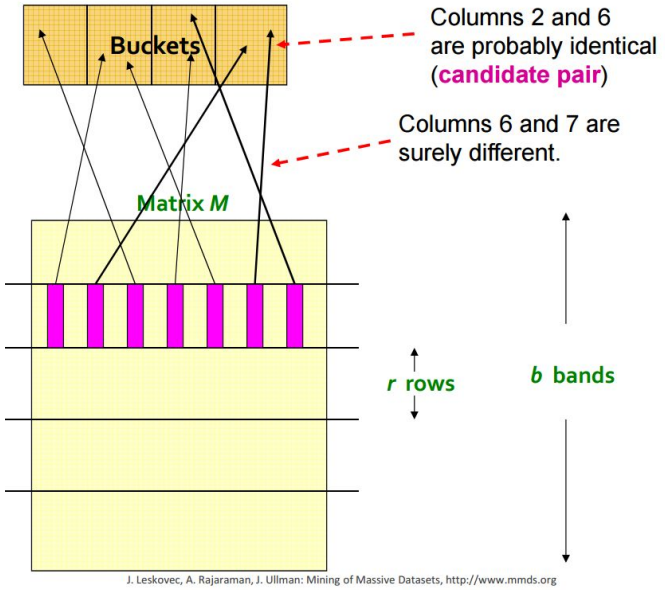
Hashing bands



Another example



Overview & algorithm



- Split M into b bands with r rows each
- For each doc (column):
  - For each band:
    - Hash band signature part into one of buckets
- Go over all buckets for all bands:
  - If there are two docs in one bucket
    -> candidate pair