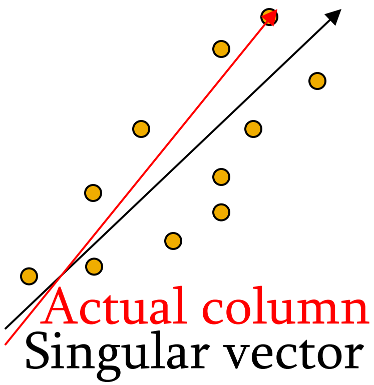
CUR: Pros & Cons

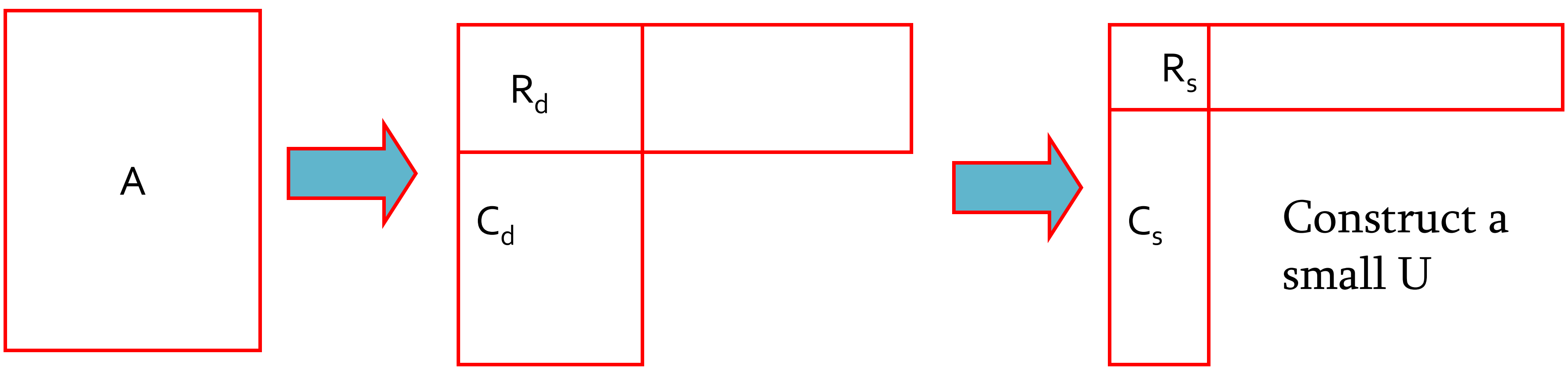
+ **Easy interpretation**
  - Since the basis vectors are actual columns and rows



+ **Sparse basis**
  - Since the basis vectors are actual columns and rows
- **Duplicate columns and rows**
  - Columns of large norms will be sampled many times

Solution

- **If we want to get rid of the duplicates:**
  - Throw them away
  - Scale (multiply) the columns/rows by the square root of the number of duplicates



SVD vs. CUR

sparse and small

SVD: $A = U \Sigma V^T$

Huge but sparse   Big and dense

dense but small

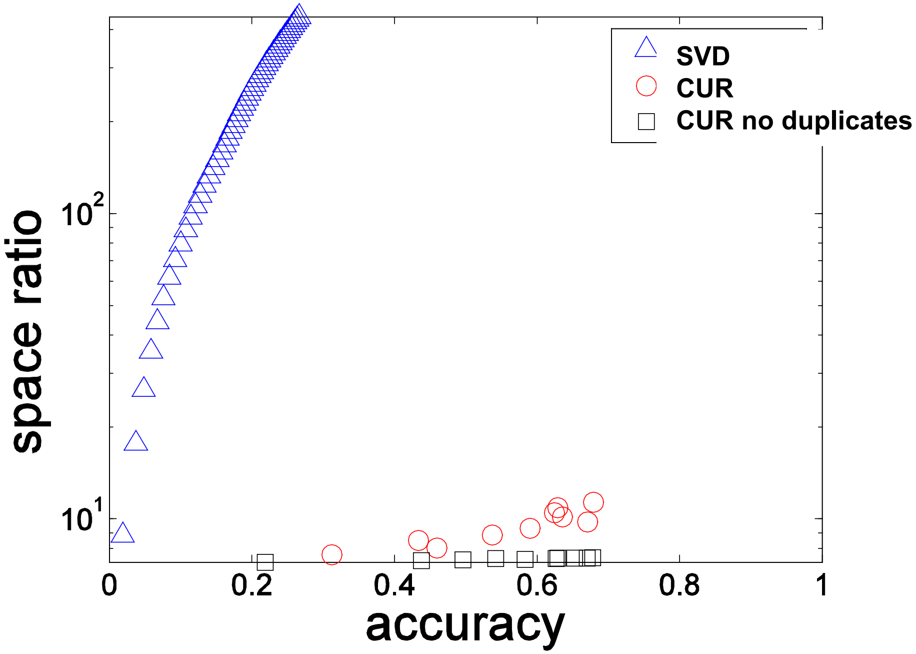
CUR: $A = C U R$

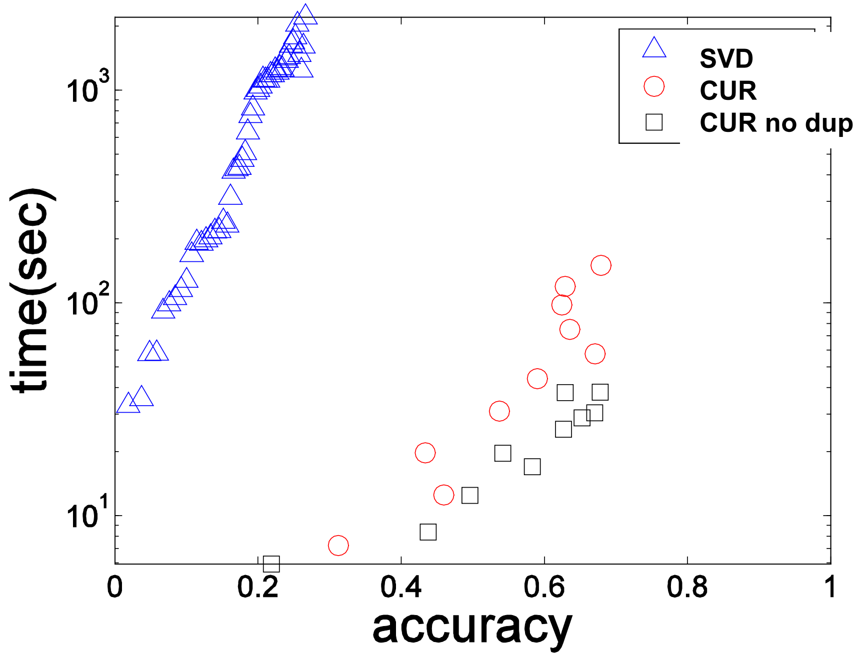
Huge but sparse   Big but sparse

Simple Experiment

- **DBLP bibliographic data**
  - Author-to-conference big sparse matrix
  - A_{ij}: Number of papers published by author *i* at conference *j*
  - 428K authors (rows), 3659 conferences (columns)
    - **Very sparse**
- **Want to reduce dimensionality**
  - How much time does it take?
  - What is the reconstruction error?
  - How much space do we need?

Results: DBLP- big sparse matrix





- **Accuracy:**
  - 1 – relative sum squared errors
- **Space ratio:**
  - #output matrix entries / #input matrix entries
- **CPU time**

Sun, Faloutsos: *Less is More: Compact Matrix Decomposition for Large Sparse Graphs*, SDM '07.