- Click here Slide(pdf))
- LSH: Locality-Sensitive Hashing
  - a way to find "similiar" sets.
  - upside: only a small faction of points are ever examined
  - downside: exists false negatives
- Steps for similiar docs
  - fisrt. Shingling
    - 1. hash shingles to a few bytes
    - 2. Compute Sim(C1, C2)
  - second. Min-hashing
    - Find a hash function h(x) make that:
    - The function to our taste is Min-hashing
    - Then we get signatures.
  - third. LSH :
    - Focus on pairs of signatures likely to be from similiar documents
    - There uses bands and rows to reduce the error.
    - Use hashing to find candidate pairs of simiarity >= s