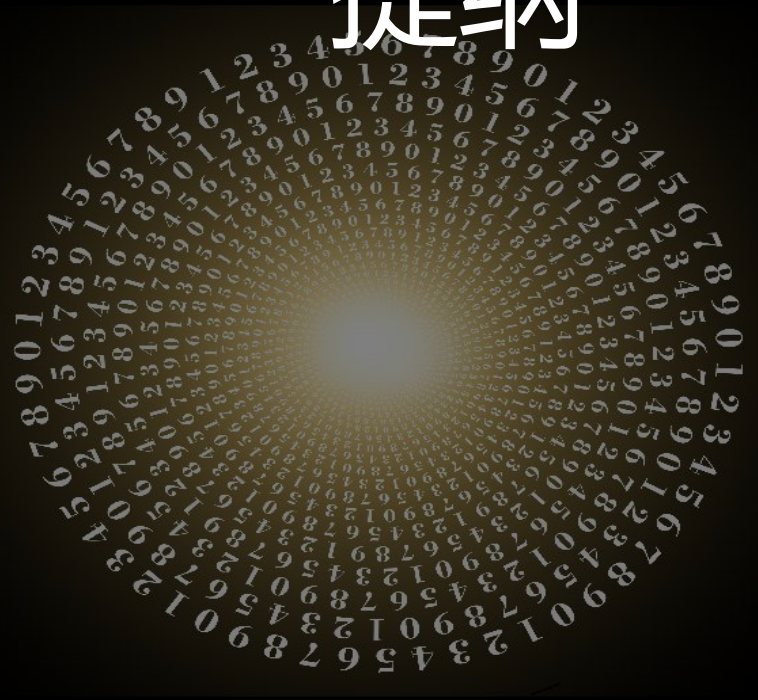
Chapter 4: Finding Similar Items -Locality Sensitive Hashing

崔金华
电子邮箱: jhcui@hust.edu.cn
个人主页: https://csjhcui.github.io/

致谢:课件主要参考《 Mining of Massive Datasets》(Second Edition) J. Leskovec, A. Rajaraman, J. Ullman的相关课件, 特此致谢!!!

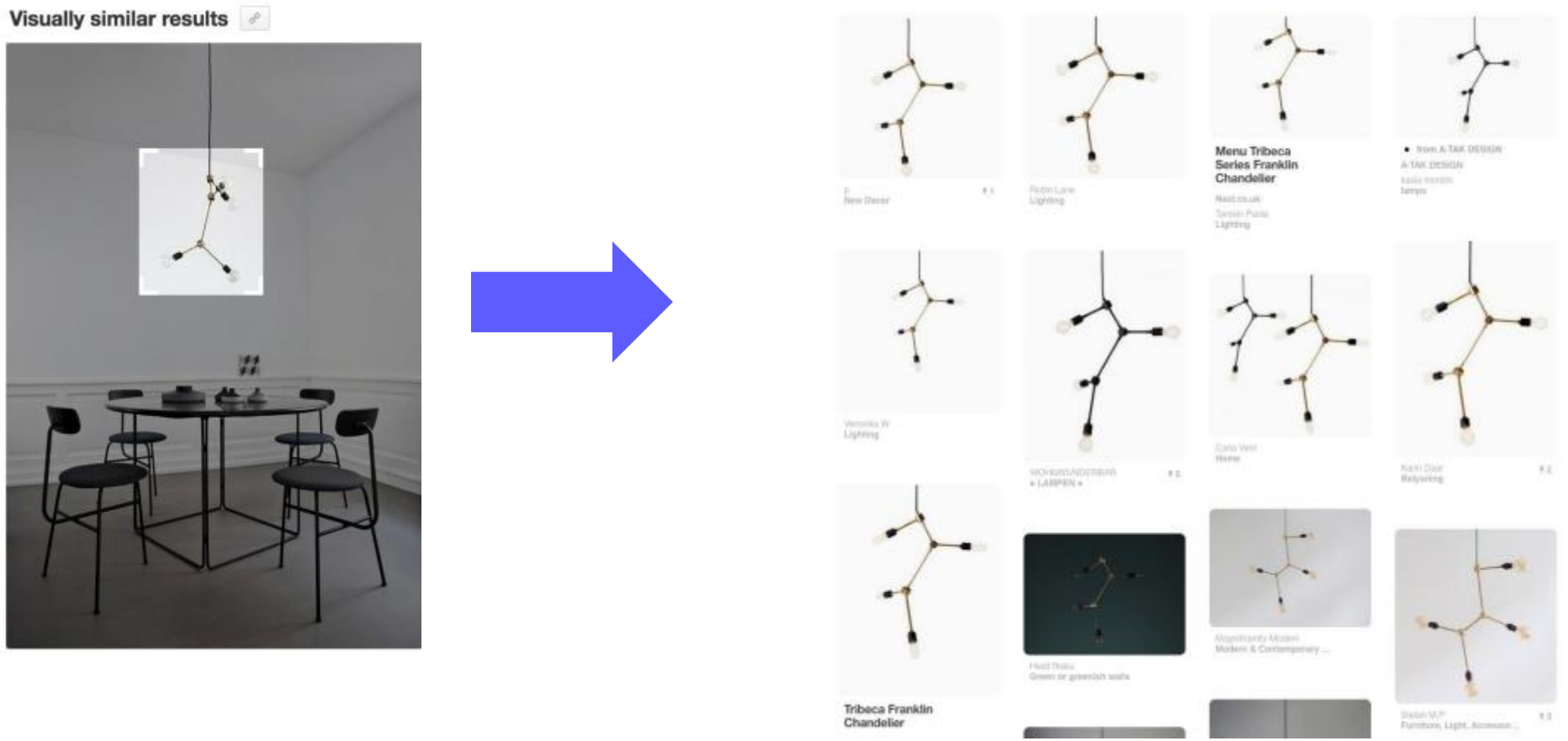
Contents

提纲



4.1 Pinterest Visual Search

- Given a query image patch, find similar images



4.1 Pinterest Visual Search

- Given a query image patch, find similar images

4.1 A Common Metaphor

- **Many problems can be expressed as finding "similar" sets:**
  - **Find near-neighbors in high-dimensional space**

- **Examples:**
  - **Customers who purchased similar products**
    - Products with similar customer sets
  - **Images with similar features**
    - Users who visited similar websites
  - **Pages with similar words**
    - For duplicate detection, classification by topic

4.1 Task: Finding Similar Documents

華中科技大學
计算机科学与技术学院
School of Computer Science &Technology，HUST

- **Goal: Given a large number (N in the millions or billions) of documents, find "near duplicate" pairs**
- **Applications:**
  - Mirror websites, or approximate mirrors
    - Don't want to show both in search results
  - Similar news articles at many news sites
    - Cluster articles by "same story"

4.1 Task: Finding Similar Documents

- **Problems:**
  - Many small pieces of one document can appear out of order in another
  - Too many documents to compare all pairs
  - Documents are so large or so many that they cannot fit in main memory