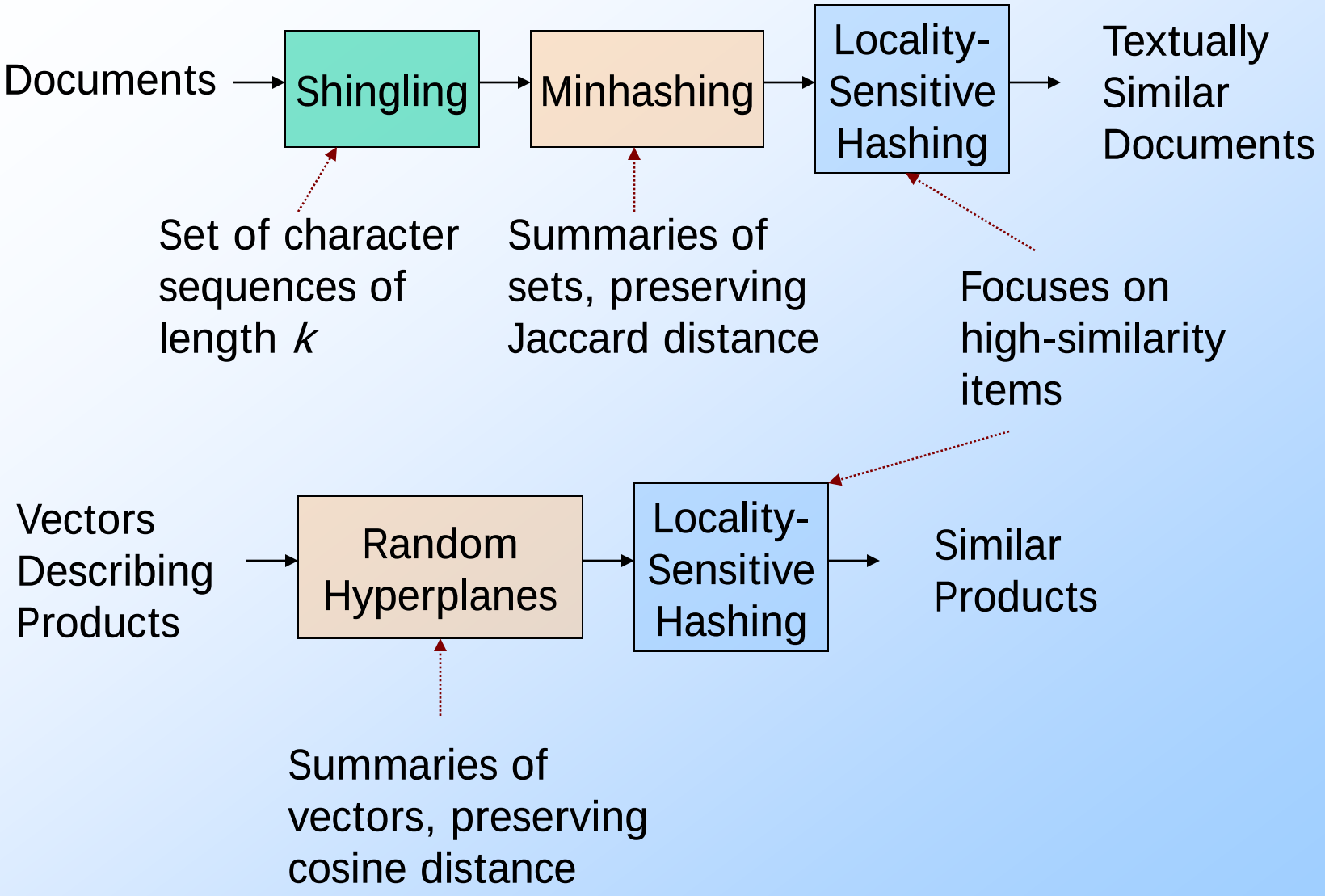
Near-Neighbor Search

Finding Variants of a News Article

Finding Similar Restaurants


Documents
Shingling
Minhashing
Locality-Sensitive Hashing
Textually Similar Documents
Set of character sequences of length k
Summaries of sets, preserving Jaccard distance
Focuses on high-similarity items
Vectors Describing Products
Random Hyperplanes
Locality-Sensitive Hashing
Similar Products
Summaries of vectors, preserving cosine distance

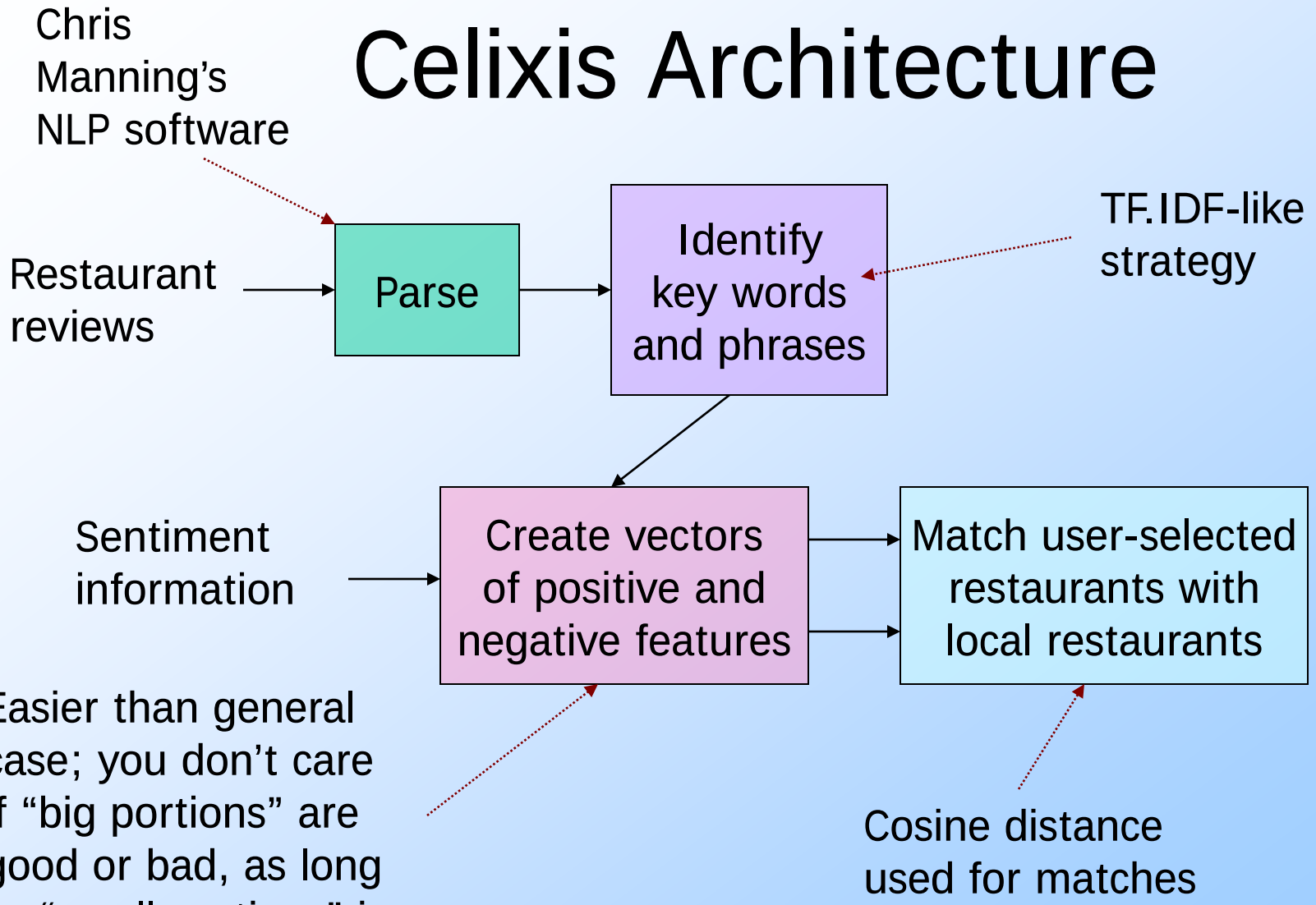
Organizing News Articles

- Applications such as Google News see thousands of on-line news articles.
- Many come from the same story, with modifications by the publisher.
- Need to cluster by underlying story.
- Perfect for shingling – minhashing – locality-sensitive hashing.

Restaurant Advisor

- Celixis is a Stanford startup trying to create “advisors.”
  - First application: restaurant advisor.
- Uses data from restaurant reviews plus one or more restaurants you select as examples of what you like.

Celixis Architecture



Further Reading

- A new book by Anand Rajaraman and Jeff Ullman, titled *Mining of Massive Datasets*, covers these topics.
- See it on-line at infolab.stanford.edu/~ullman/pub/book.pdf