Data Mining

Keyword Based Search Engines (Part B)

Dr. Jason T.L. Wang, Professor

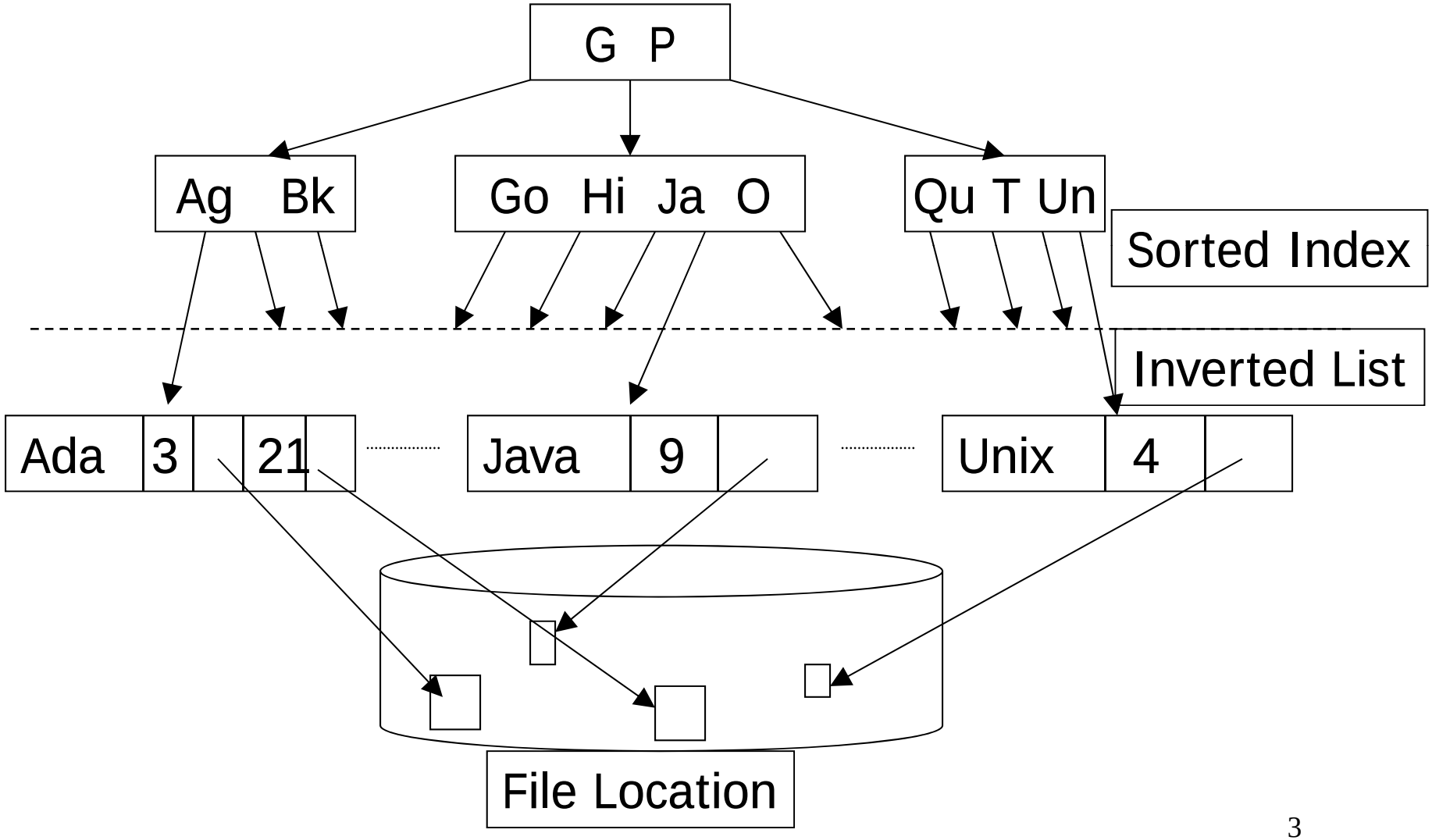
Department of Computer Science

New Jersey Institute of Technology

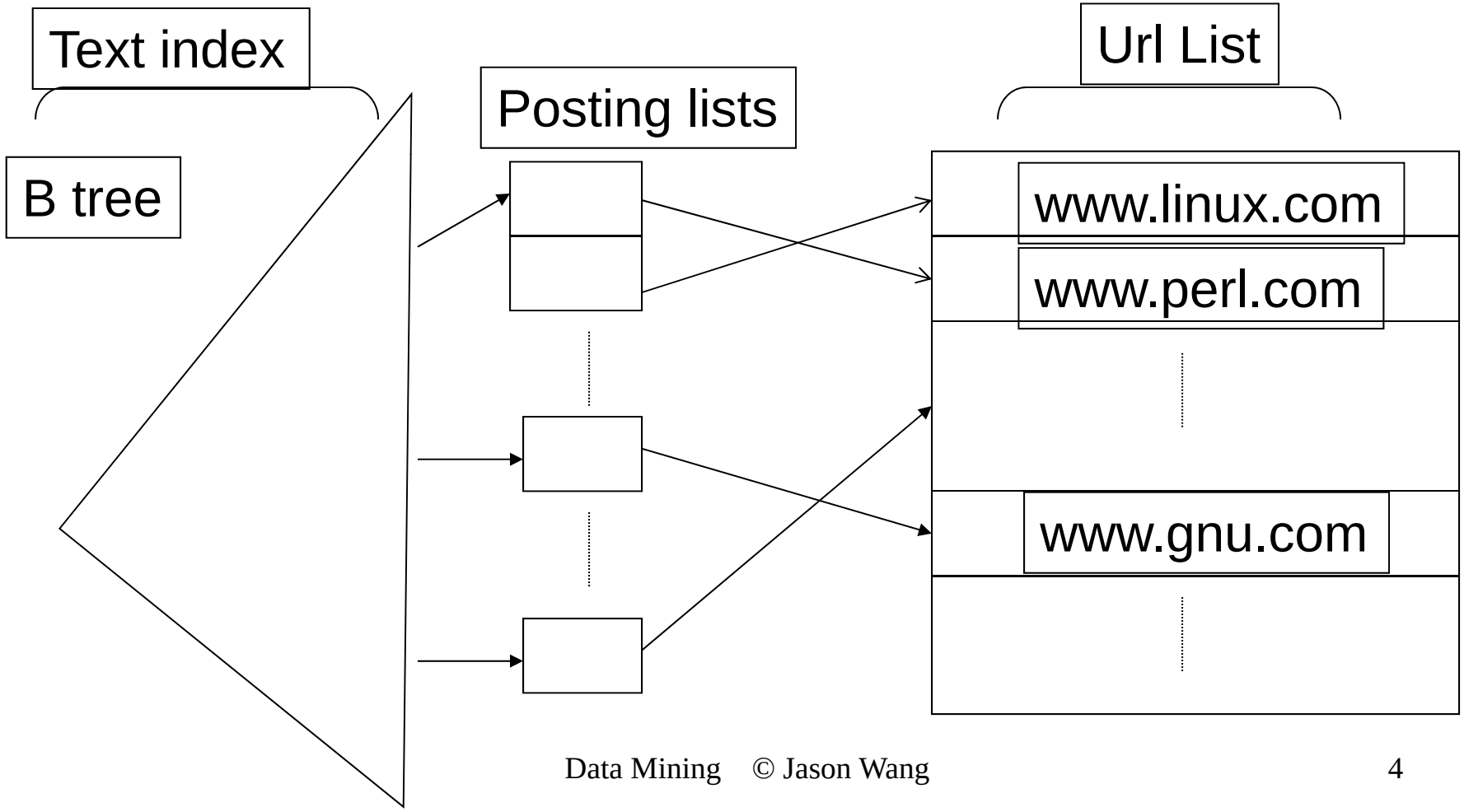
Search Index

- **Inverted File**
- Suffix Tree
- Suffix Array
- Signature File

Inverted File with a B-tree



Index Structure of a Web Search Engine



Techniques for Reducing Index Size

Case folding - converts everything to lower case.
E.g., "Data Mining" becomes "data mining".

Stemming - reduces words to their morphological root.
E.g., "compression" and "compressed" become "compress".

Stop word removal - removes common or semantically insignificant words.
E.g., "the", "a" "an" are removed.

Text compression - reduces the inverted file size.

Web Crawlers

Web crawlers (or agents, robots, spiders) are programs that work continuously behind the scenes, locating information on the Web and retrieving it for indexing.

• Fact: Research shows that Web coverage of a search engine ranges between 5% to 30% and the union of 11 major search engines covers less than 50% of the Web.

• Search engines using crawlers to automatically create their index include Google, Northern Light, Direct Hit, Inktomi and Fast Search.

• Yahoo! depends on humans to create its directory (information submitted by various sites or generated by human editors).

Yahoo! directory categories

Arts & Humanities	News & Media
Business & Economy	Recreation & Sports
Computer & Internet	Reference Education
Regional Entertainment	Science
Government	Social Science
Health	Society & Culture

Meta-Search Engines

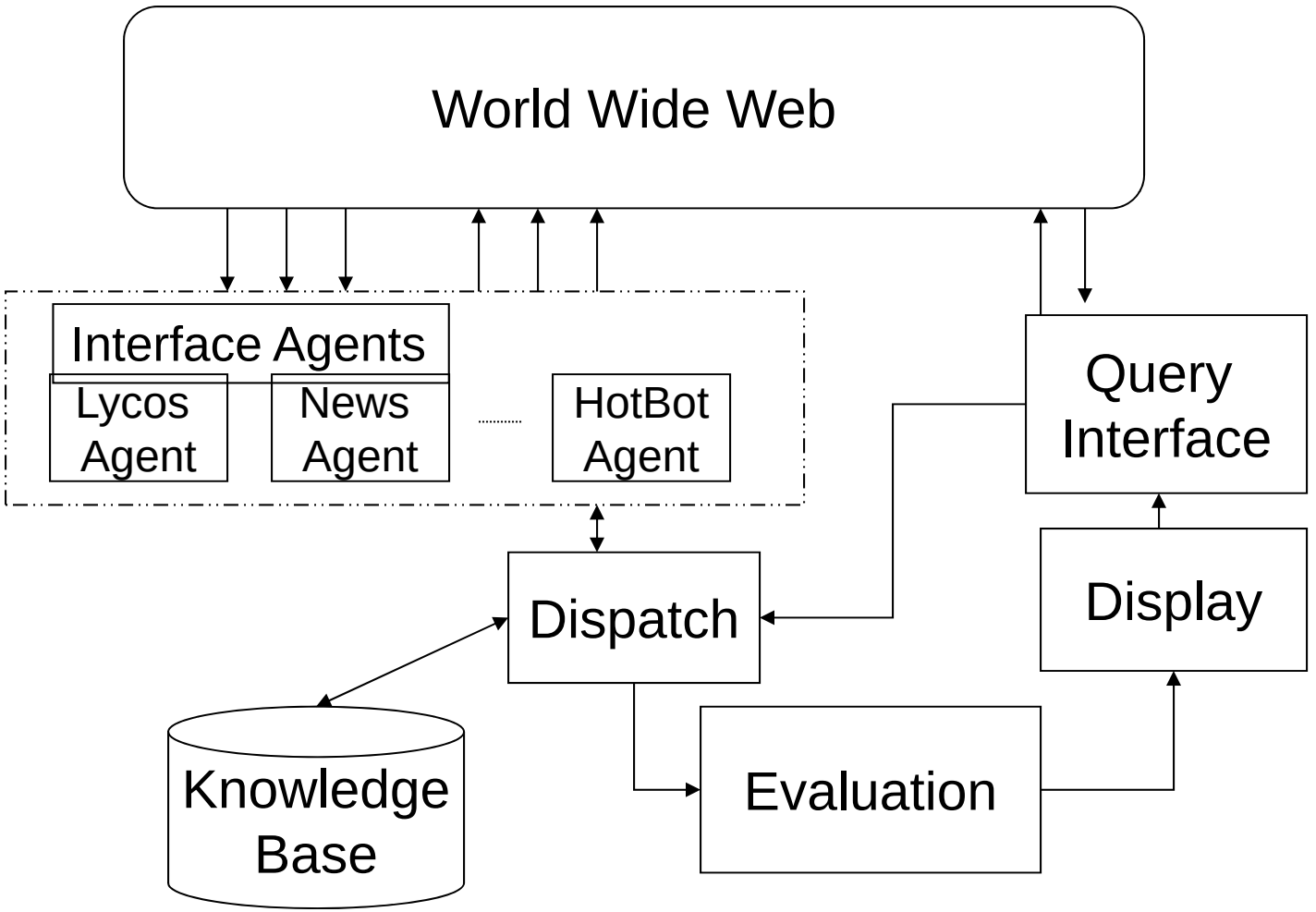
Facts:

- A single search engine that has a recall of 45%, returns only 45% of relevant results
- Research shows that the coverage of a typical search engine is between only 5% and 30% of the Web.

Principle: A dozen search engines is better than one.

Example engines: MetaCrawler, SherlockHound, SavvySearch, Inquirus

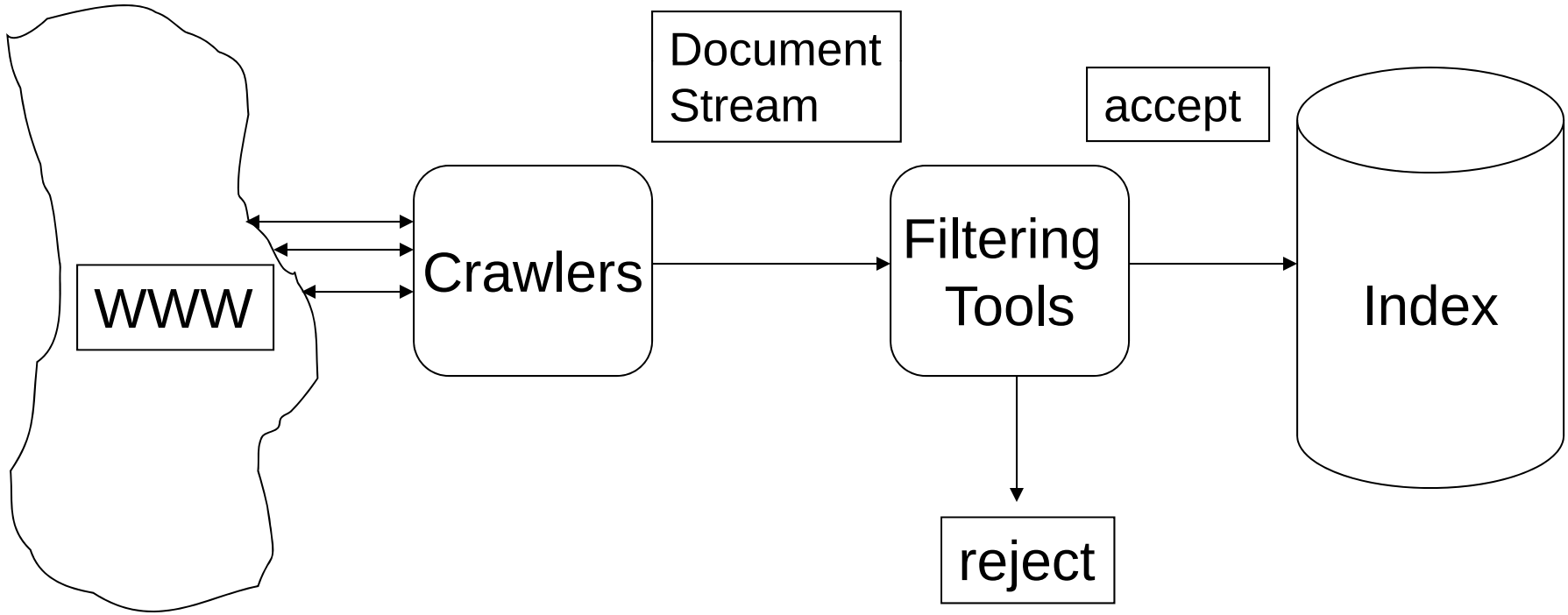
Architecture of A Meta-Search Engine



Information Filtering

Filtering is essential to build topic-specific search engines
- DejaNews: specializes in Usenet news articles
- BioCrawler: for biological information search
- Cora: allows to search for computer science research papers in Postscript format from universities and labs

Information Filtering Process



End of Keyword Based Search Engines Module (Part B)