Major Crawling Strategies

- Breadth-First is common (balance server load)
- Parallel crawling is natural
- Variation: focused crawling
  - Targeting at a subset of pages (e.g., all pages about “automobiles”)
  - Typically given a query
- How to find new pages (they may not linked to an old page!)
- Incremental/repeated crawling
  - Need to minimize resource overhead
  - Can learn from the past experience (updated daily vs. monthly)
  - Target at : 1) frequently updated pages; 2) frequently accessed pages