# Information Retrieval and Web Search

Introduction to Information Retrieval

# Information Retrieval (IR)

- Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information (Salton, 1968)
- Primary focus: text
  - Scholarly papers, books, news, email, Web pages
- Other media:
  - Audio, images, video

# Types of IR

- **Search**
- Filtering or tracking
  - Detecting documents of interest
- Classification
  - Labeling documents with pre-existing labels
- Question answering
  - Beyond searching for documents, searching for answers

# Types of Search Engines

- **Web search**
- Vertical search
- Enterprise search
- Desktop search
- Peer-to-peer search

# Database Engines vs. Search Engines

- "Find accounts whose balance is greater than $100"
  - Requires data to be well-structured and well-defined
  - Typically involves not just text but numbers
  - Requires use of formal language and logic

- "Find the Canvas system of UCI" or
  "Find how to merge two lists in Python"
  "Find the graph of y=log(x)"
  - Vague and varied information needs
  - Informal and ad-hoc queries
  - Requires engine to infer **meaning** of words and sentences

# Core Issue: **Relevance**

- Textual similarity [databases stop here]
- Document context
  - Origin
  - Author
  - Popularity
  - …
- User context
  - Geographic location
  - Prior queries
  - Age
  - Preferred language
  - …
- Query context
  - Special symbols

# Retrieval Models

- Conceptual designs of what to pay attention to when matching a query and a document
  - E.g. grammar vs. raw text, what parts of context, etc.
- Good retrieval model: finds documents relevant to the person who submitted the query
- Basis for *ranking algorithms*

# Classic IR Assumptions

- Corpus: fixed document collection
- Goal: retrieve information content relevant to the information need

# Classic IR Goal

- “Relevance”
  - For each query Q, and stored document D, there exists a relevance score R(Q, D)
  - Maximize R(Q, D)
    - Context is ignored
    - User is ignored
    - Corpus is static

# Web IR

- The Web is huge
- The Web changes all the time
- There is information to avoid (adversarial IR)
- One interface for hugely divergent needs

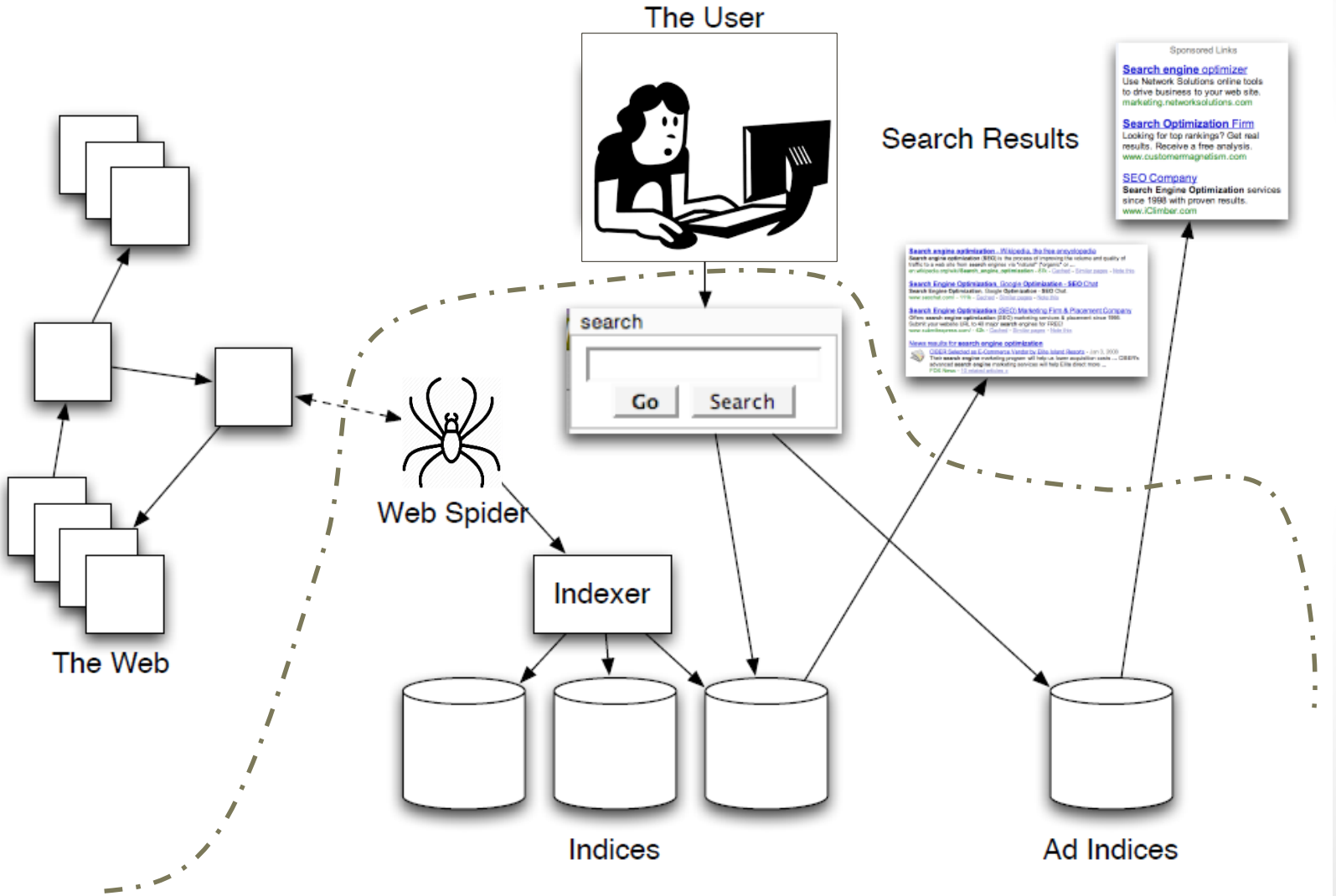
# Web Search Engines

- Practical and useful applications of IR
- Must crawl terabytes of web pages and provide sub-second response to queries
- Big Issues in the design, additionally to IR issues:
  - Performance
    - Response time
    - Query throughput
    - Indexing speed
  - Speed in discovery and integration of new documents
  - Coverage
  - Freshness
  - Spam

# Search Engineers

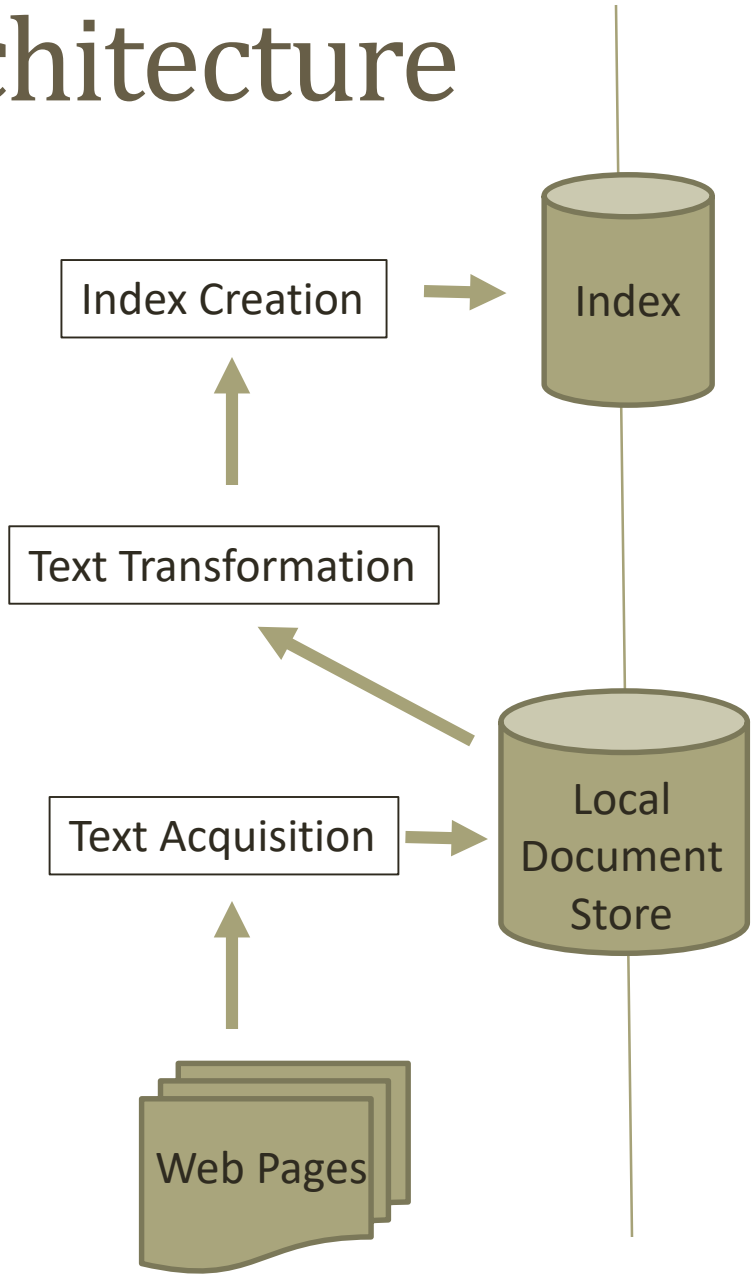
- Develop and maintain search engines
- Design or optimize content for search engines

# Workflow

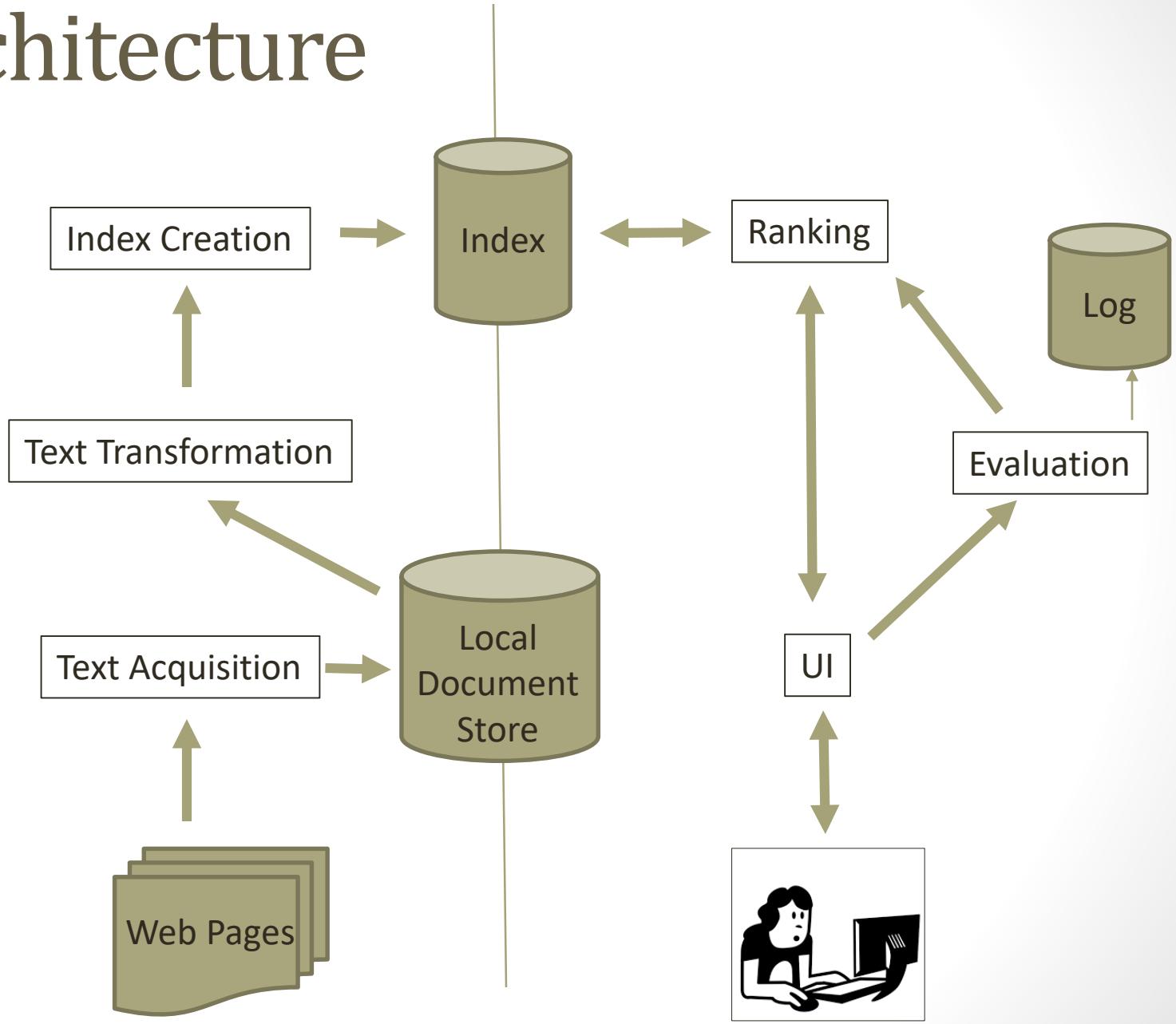

The User
Search Results
Sponsored Links
Search engine optimizer
Use Network Solutions online tools to drive business to your web site.
marketing.networksolutions.com
Search Optimization Firm
Looking for top rankings? Get real results. Receive a free analysis.
www.customermagnetism.com
SEO Company
Search Engine Optimization services since 1998 with proven results.
www.iClimber.com
search
Go
Search
Web Spider
Indexer
The Web
Indices
Ad Indices

# Architecture

Preprocessing Steps



# Architecture




Preprocessing Steps

Querying Process

# Text Acquisition

- Crawler
- Feeds
- Text conversion
- Document Store

# Text Transformation

- Parser = Tokenizer + Structure
- Stopping
- Stemming
- Link Analysis
- Information Extraction (text structure)
- Classifier (topic, non-content, spam, etc.)

# Index Creation

- Corpus statistics
- Term weighting
- Index inversion (doc → term ➔ term → doc)
- Index distributions (for large indexes)

# User Interaction

- Query input
- Query transformation
- Results output

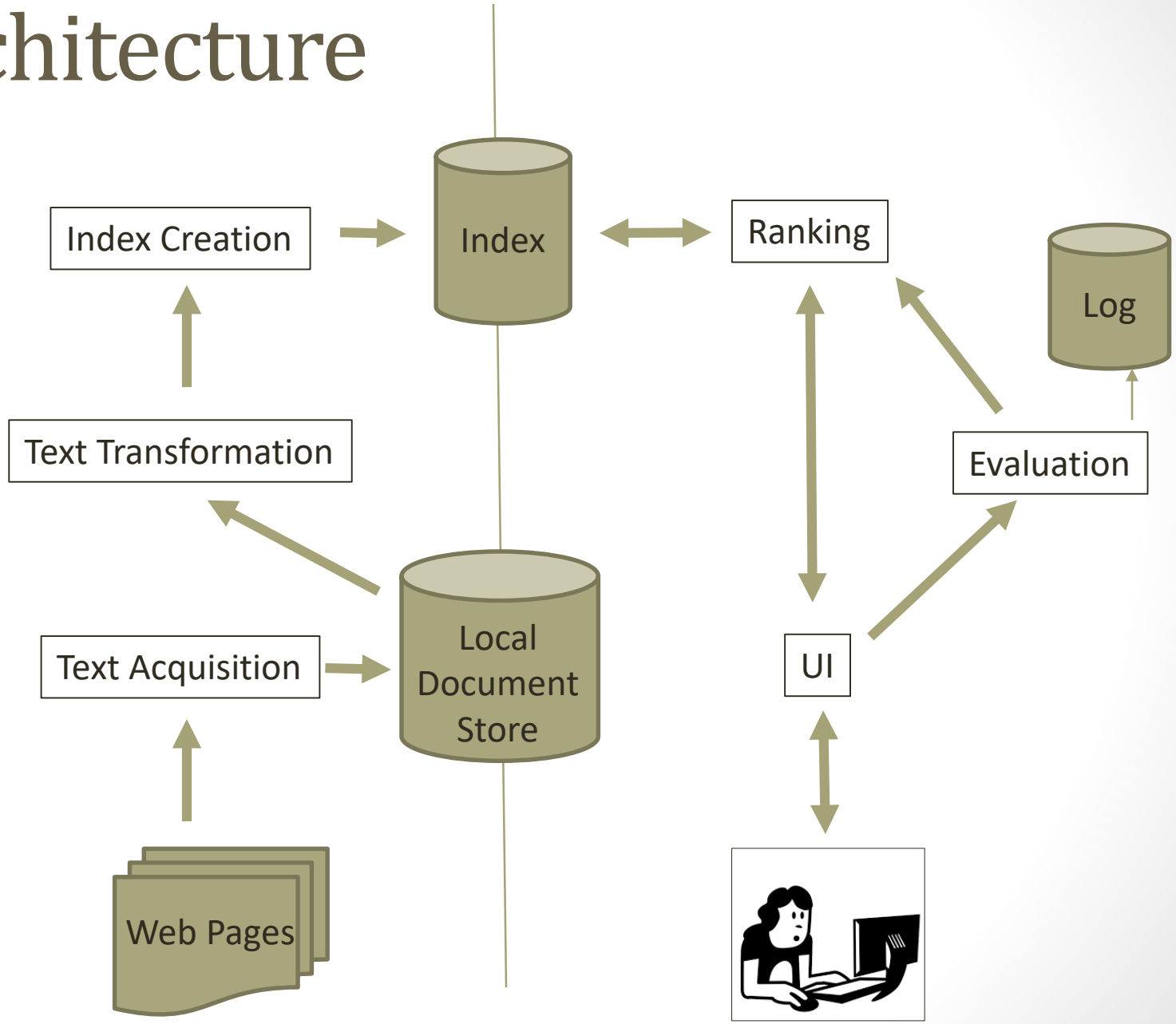
# Ranking

- Scoring: how well each doc matches the query
- Performance optimization
- Distribution

# Evaluation

- Logging
- Ranking analysis
- Performance analysis

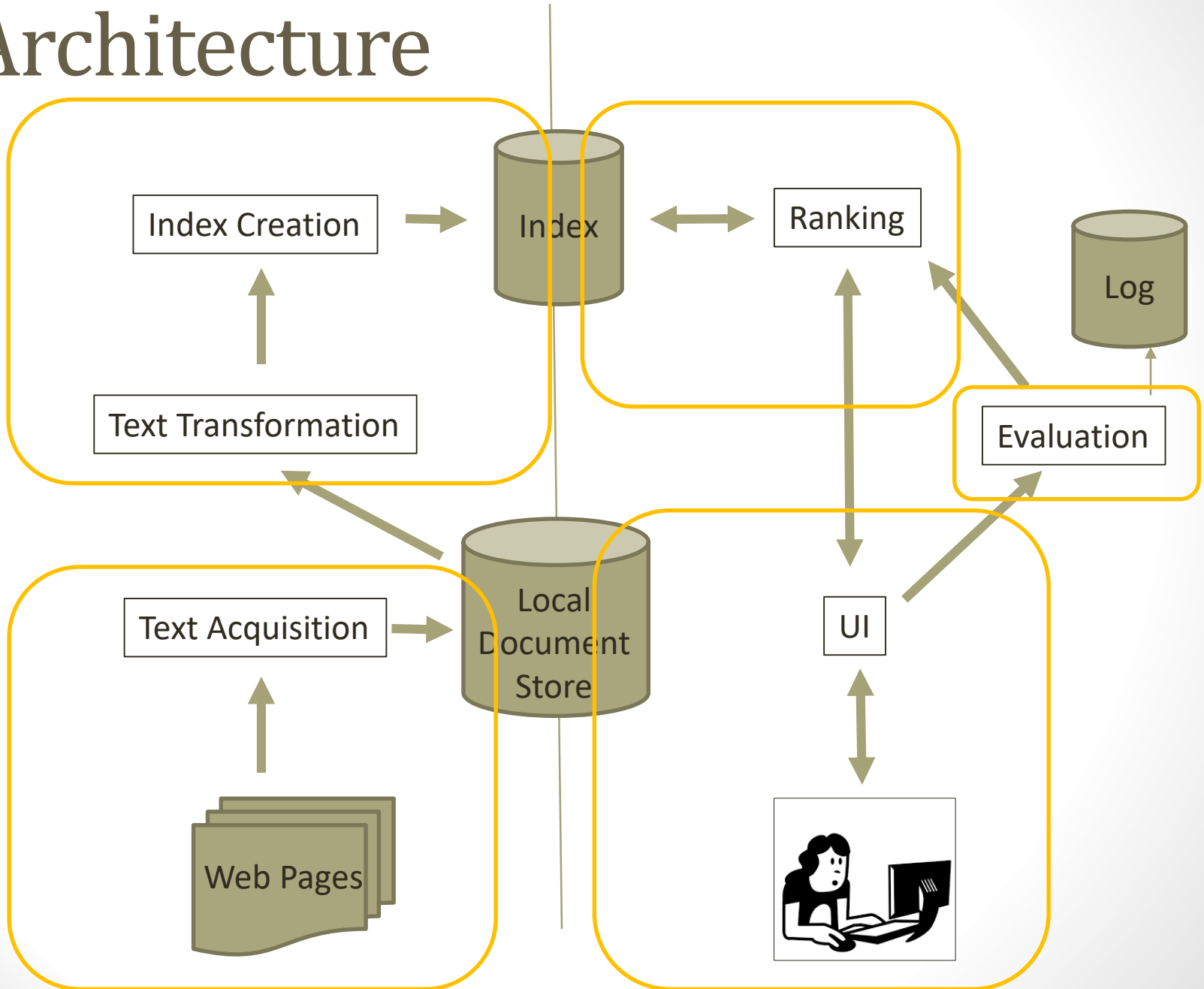
# Architecture




Preprocessing Steps

Querying Process

# Architecture




Preprocessing Steps

Querying Process