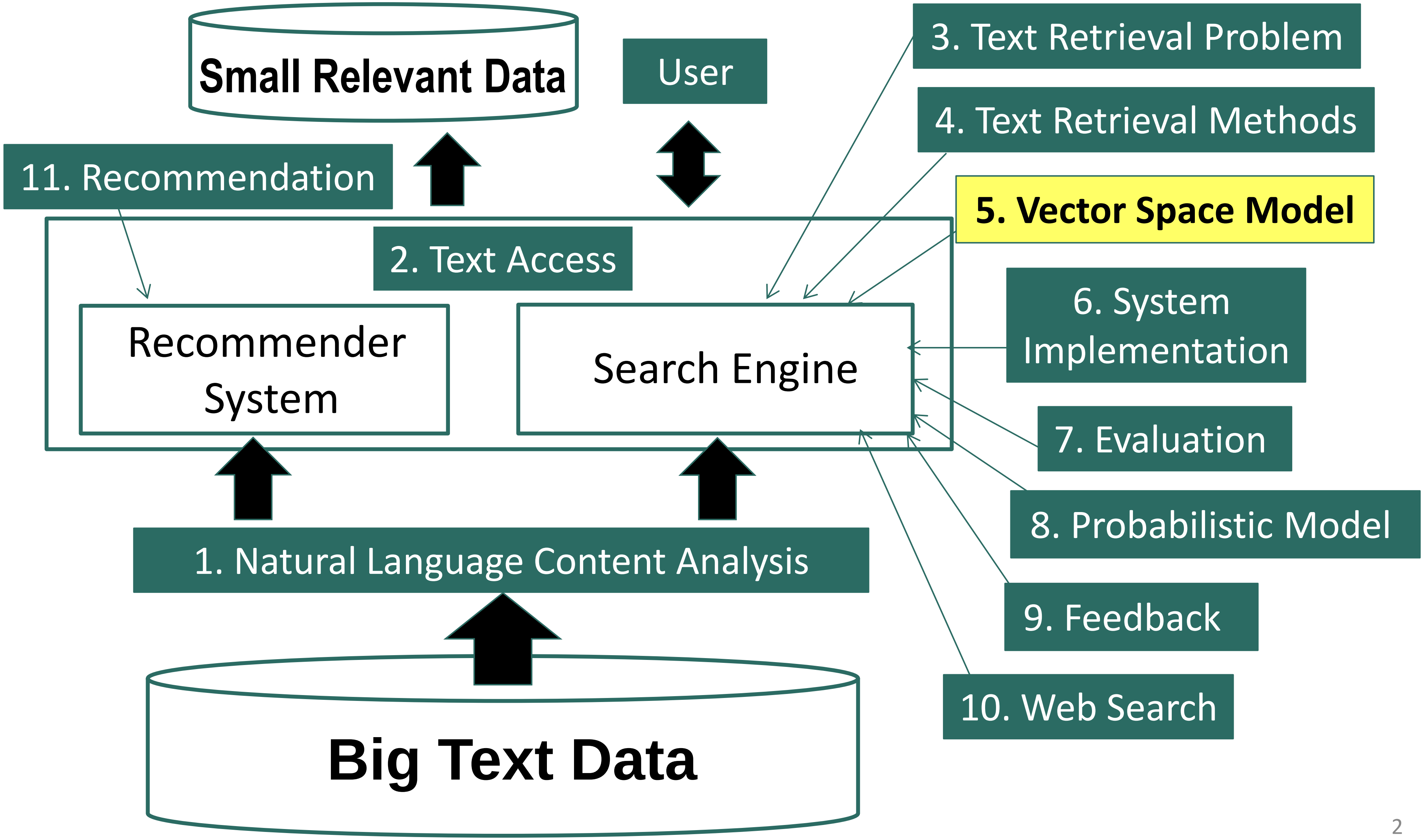
Text Retrieval and Search Engines

Vector Space Retrieval Model: Basic Idea

ChengXiang "Cheng" Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

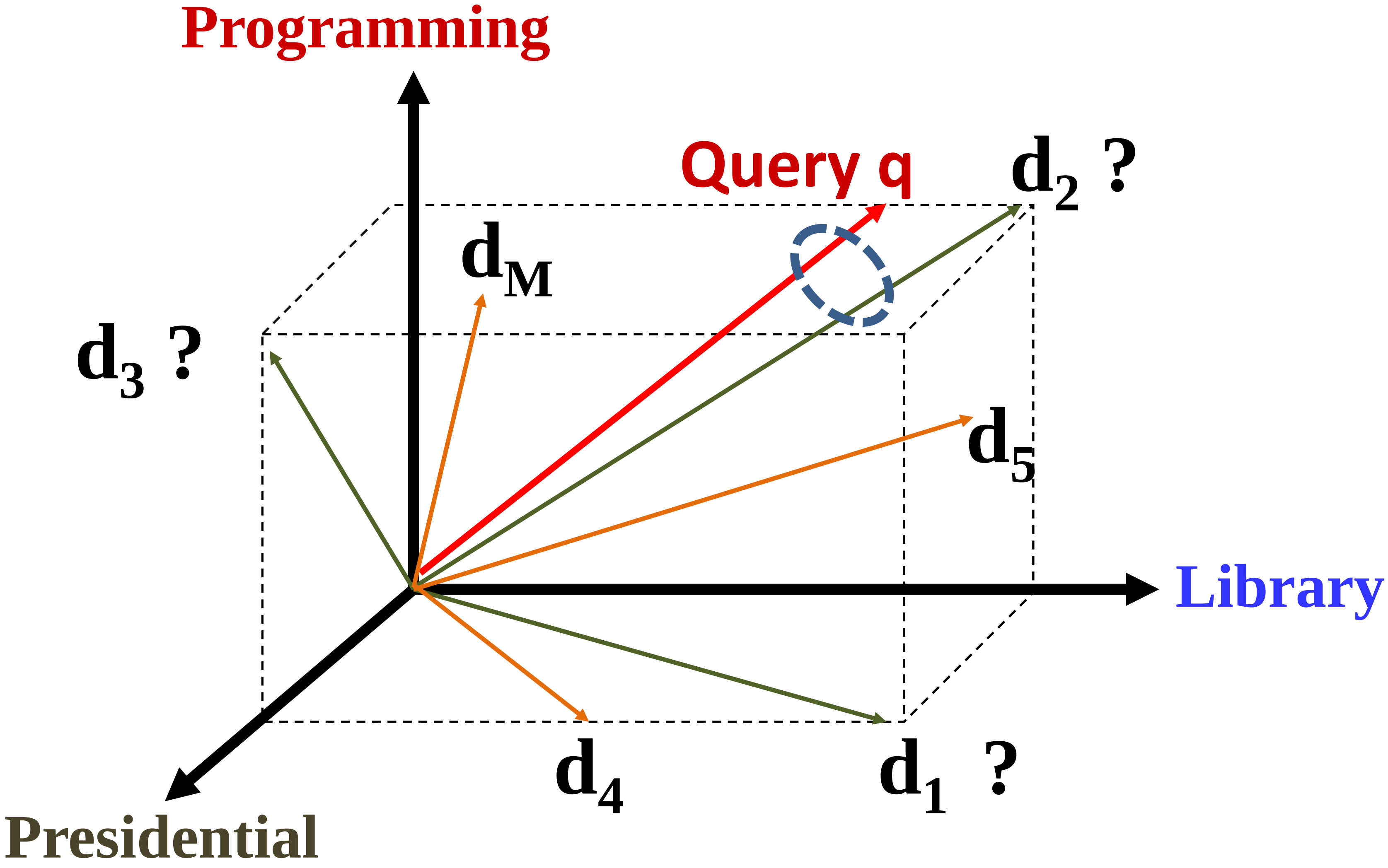
Course Schedule



Many Different Retrieval Models

- Similarity-based models: f(q,d) = similarity(q,d)
  - Vector space model

Vector Space Model (VSM): Illustration



VSM Is a Framework

- Represent a doc/query by a term vector
  - **Term**: basic concept, e.g., word or phrase
  - Each term defines one dimension
  - N terms define an **N-dimensional space**
  - **Query** vector: **q**=(x_1, …x_N), $x_i \in \Re$ is query term weight
  - **Doc** vector: **d**=(y_1, …y_N), $y_j \in \Re$ is doc term weight

- relevance(q,d) $\propto$ similarity(**q**,**d**) =f(q,d)

What VSM Doesn’t Say

- How to define/select the “basic concept”
  - Concepts are assumed to be orthogonal
- How to place docs and query in the space (= how to assign term weights)
  - Term weight in query indicates importance of term
  - Term weight in doc indicates how well the term characterizes the doc
- How to define the similarity measure