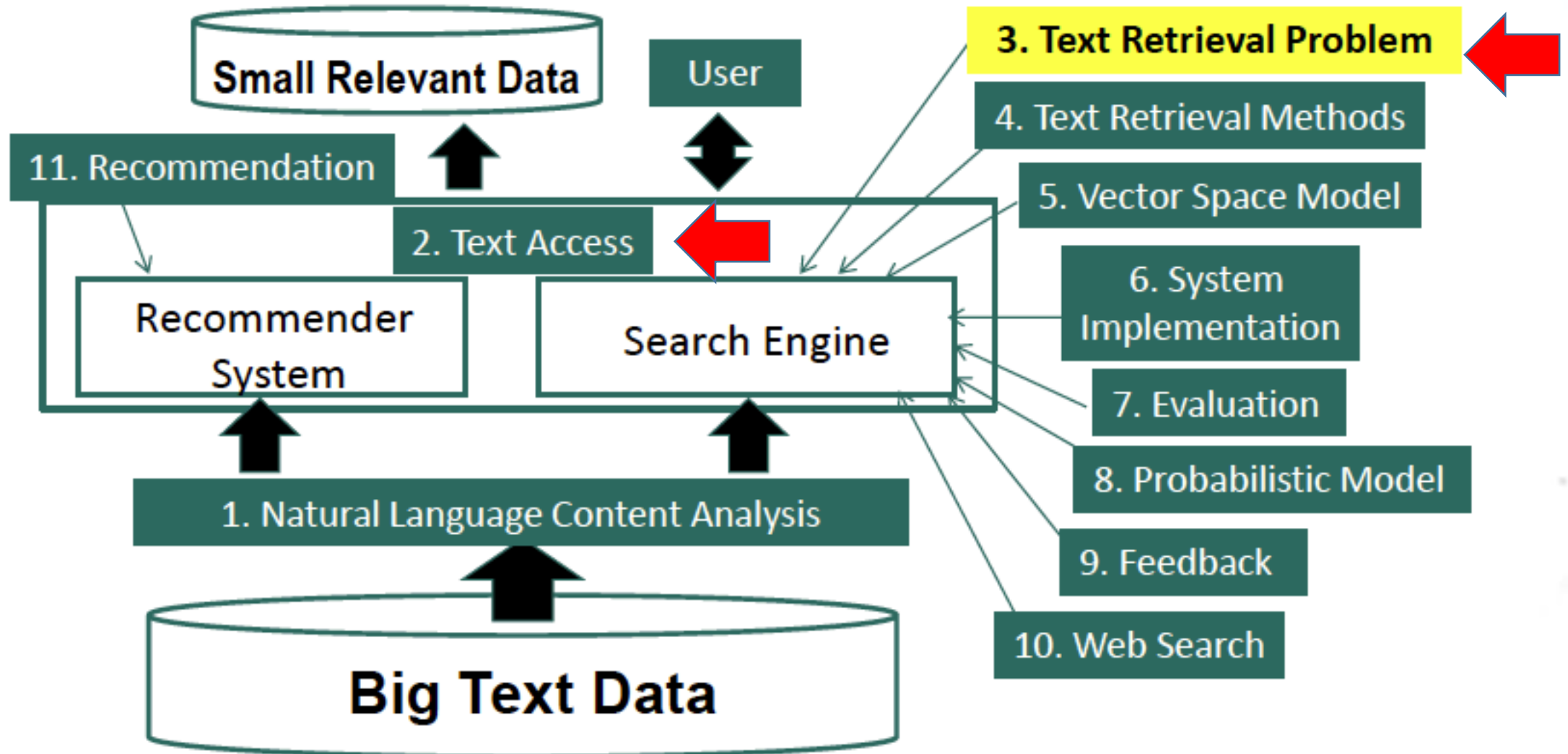


Information Retrieval & Text Mining

Overview of Text Retrieval Methods

Dr. Iqra Safder
Information Technology University

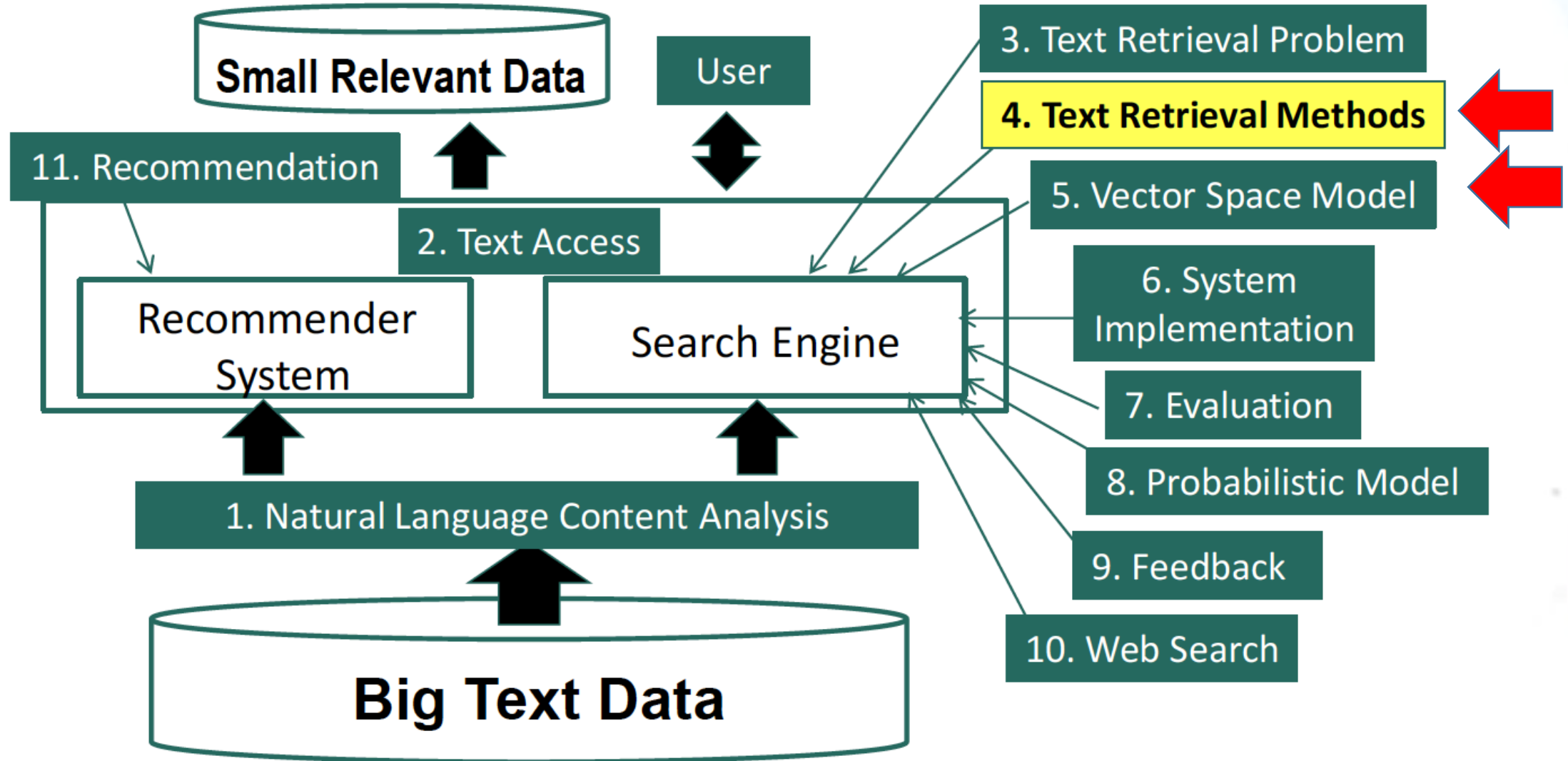
Recap



Recap

- Text retrieval is an empirically defined problem
 - Which algorithm is better must be judged by users
- Document ranking is generally preferred to
 - Help users prioritize examination of search results
 - Bypass the difficulty in determining absolute relevance (users help decide the cutoff on the ranked list)
- Main challenge: design an effective ranking function
 $f(q,d) = ?$

Course Schedule



How to Design a Ranking Function

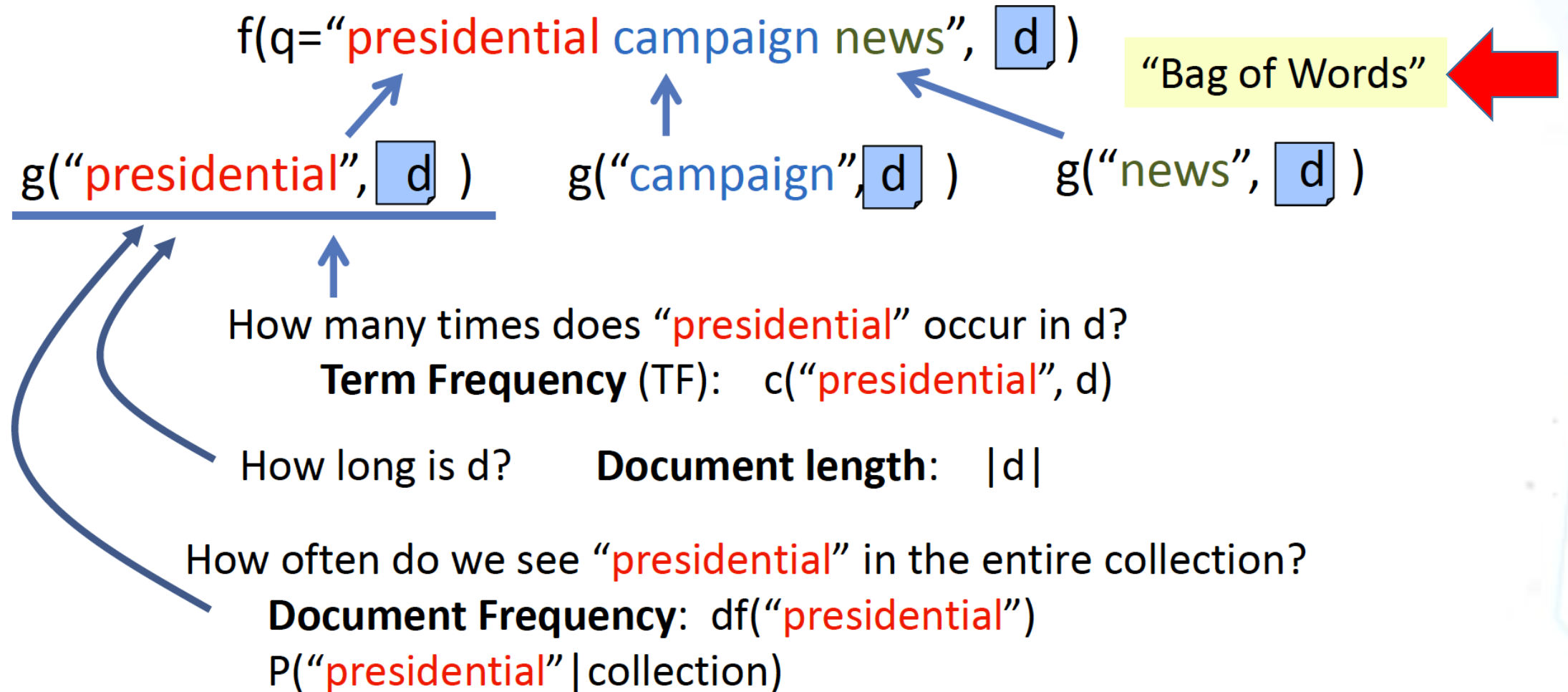
- **Query:** $q = q_1, \dots, q_m$, where $q_i \in V$
- **Document:** $d = d_1, \dots, d_n$, where $d_i \in V$
- **Ranking function:** $f(q, d) \in \mathcal{R}$
- A good ranking function should rank relevant documents on top of non-relevant ones
- Key challenge: how to measure the likelihood that document d is relevant to query q
- **Retrieval model** = formalization of relevance (give a computational definition of relevance)

We have to have a computational definition of the relevance.

Many Different Retrieval Models

- **Similarity-based models:** $f(q,d) = \text{similarity}(q,d)$
 - Vector space model
- **Probabilistic models:** $f(d,q) = p(R=1 \mid d,q)$, where $R \in \{0,1\}$
 - Classic probabilistic model
 - Language model
 - Divergence-from-randomness model
- **Probabilistic inference model:** $f(q,d) = p(d \rightarrow q)$
- **Axiomatic model:** $f(q,d)$ must satisfy a set of constraints
- These different models tend to result in similar ranking functions involving similar variables

Common Ideas in State of the Art Retrieval Models



Matching a rare term in the collection is contributing more to the overall score than matching up common term.

Which Model Works the Best?

- When optimized, the following models tend to perform equally well [Fang et al. 11]:
 - **Pivoted length normalization**
 - **BM25**
 - **Query likelihood**
- BM25 is most popular

Summary

- Design of ranking function $f(q,d)$ pre-requires a computational definition of relevance (retrieval model)
- Many models are equally effective with no single winner
- State of the art ranking functions tend to rely on
 - Bag of words representation
 - Term Frequency (TF) and Document Frequency (DF) of words
 - Document length

There is no single winner yet. Researchers are still active and working on this problem, trying to find a truly optimal retrieval model

Additional Readings

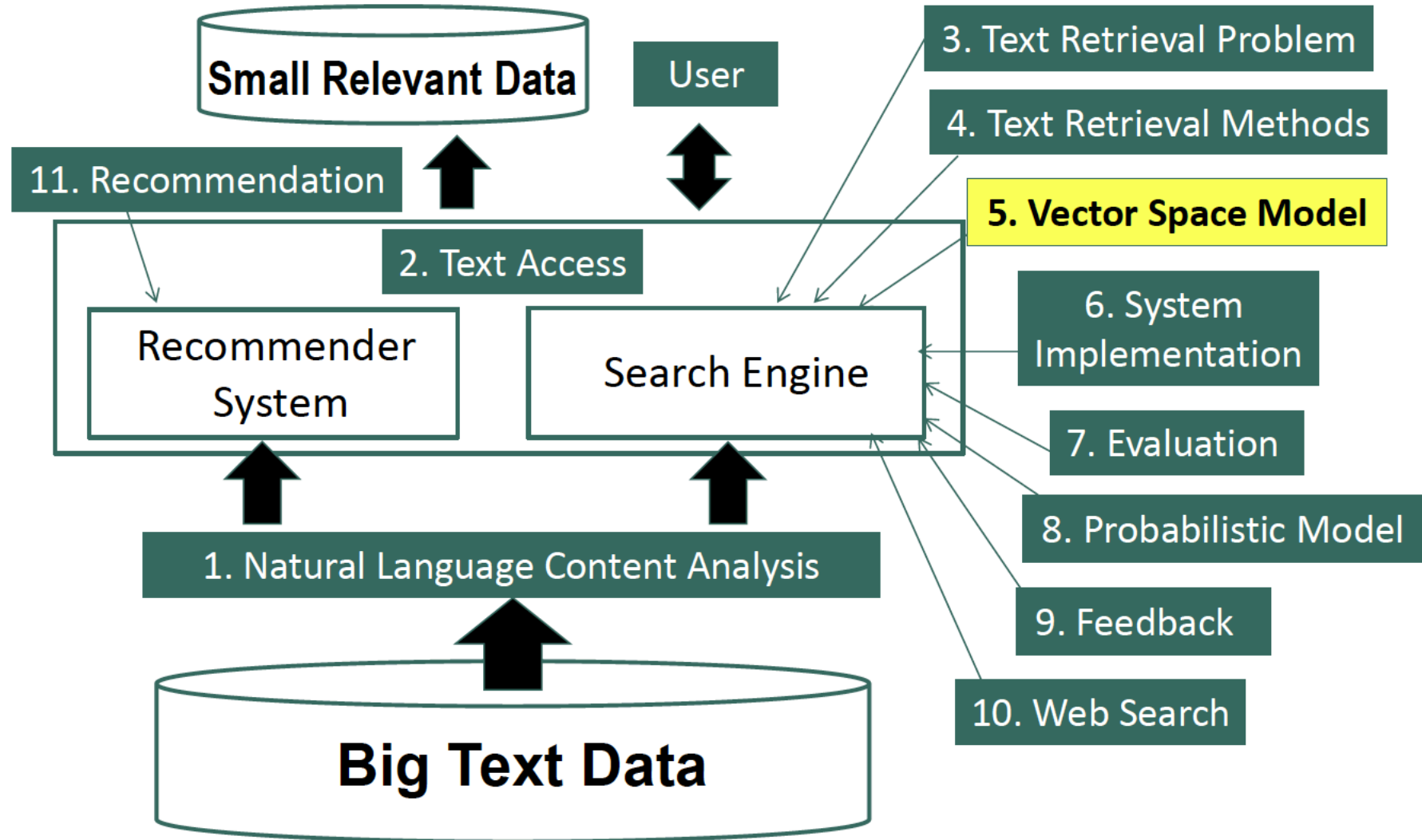
- Detailed discussion and comparison of state of the art models
 - Hui Fang, Tao Tao, and Chengxiang Zhai. 2011. Diagnostic Evaluation of Information Retrieval Models. *ACM Trans. Inf. Syst.* 29, 2, Article 7 (April 2011)
- Broad review of different retrieval models
 - ChengXiang Zhai, *Statistical Language Models for Information Retrieval*, Morgan & Claypool Publishers, 2008. (Chapter 2)

Information Retrieval & Text Mining

Vector Space Retrieval Model Basic Idea

Dr. Iqra Safder
Information Technology University

Course Schedule



Many Different Retrieval Models

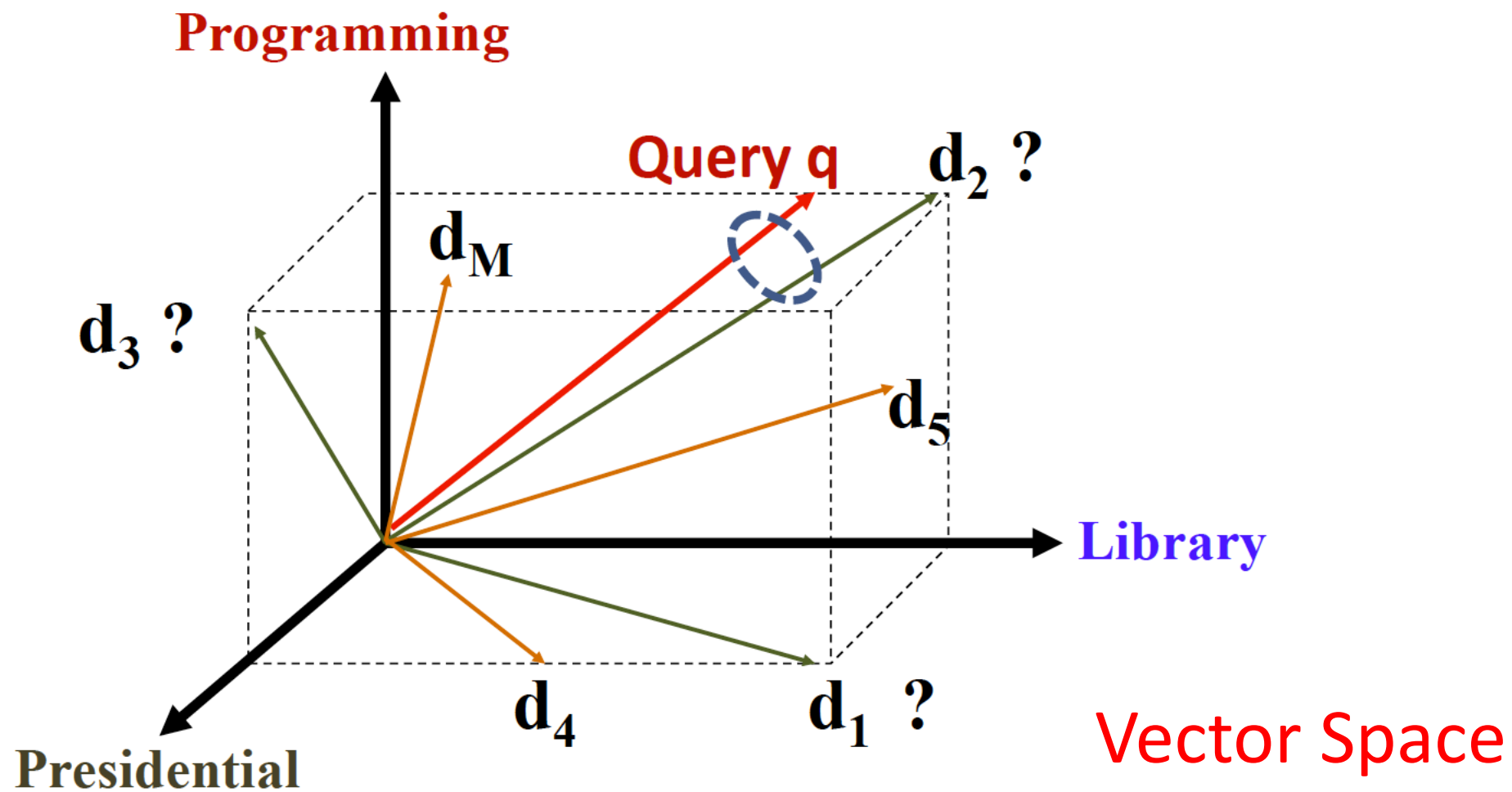
- Similarity-based models: $f(q,d) = \text{similarity}(q,d)$
 - Vector space model

Relevance based method.

Assumptions:

If a first documents is more similar to the query then it is assumed as more relevant document than others.

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: $\mathbf{q}=(x_1, \dots x_N)$, $x_i \in \mathfrak{R}$ is query term weight
 - **Doc** vector: $\mathbf{d}=(y_1, \dots y_N)$, $y_j \in \mathfrak{R}$ is doc term weight
- $\text{relevance}(\mathbf{q}, \mathbf{d}) \propto \text{similarity}(\mathbf{q}, \mathbf{d}) = f(\mathbf{q}, \mathbf{d})$

What VSM Doesn't Say

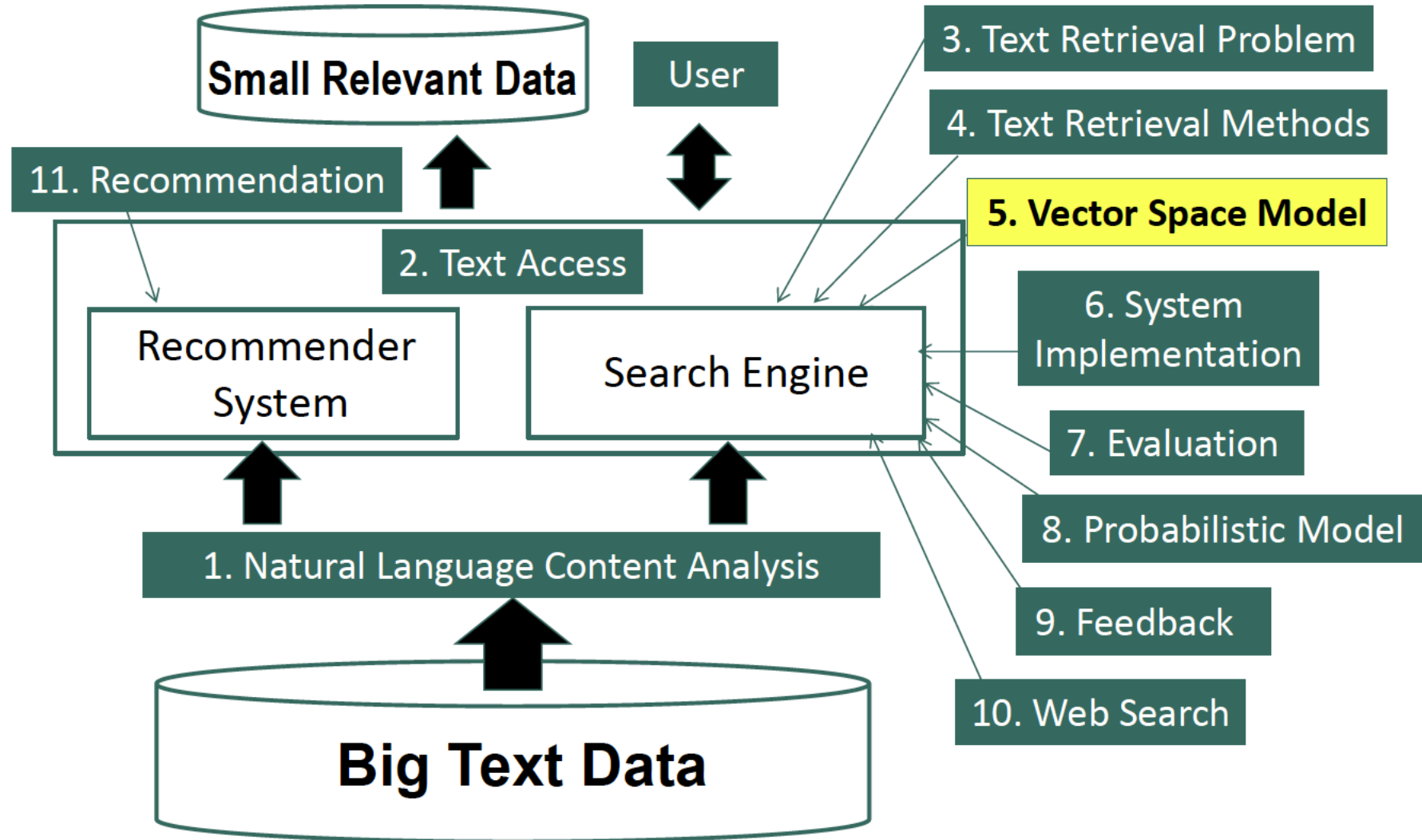
- How to define/select the “basic concept”
 - Concepts are assumed to be orthogonal → Statistically Independent
- How to place docs and query in the space (= how to assign term weights)
 - Term weight in query indicates importance of term
 - Term weight in doc indicates how well the term characterizes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

Information Retrieval & Text Mining

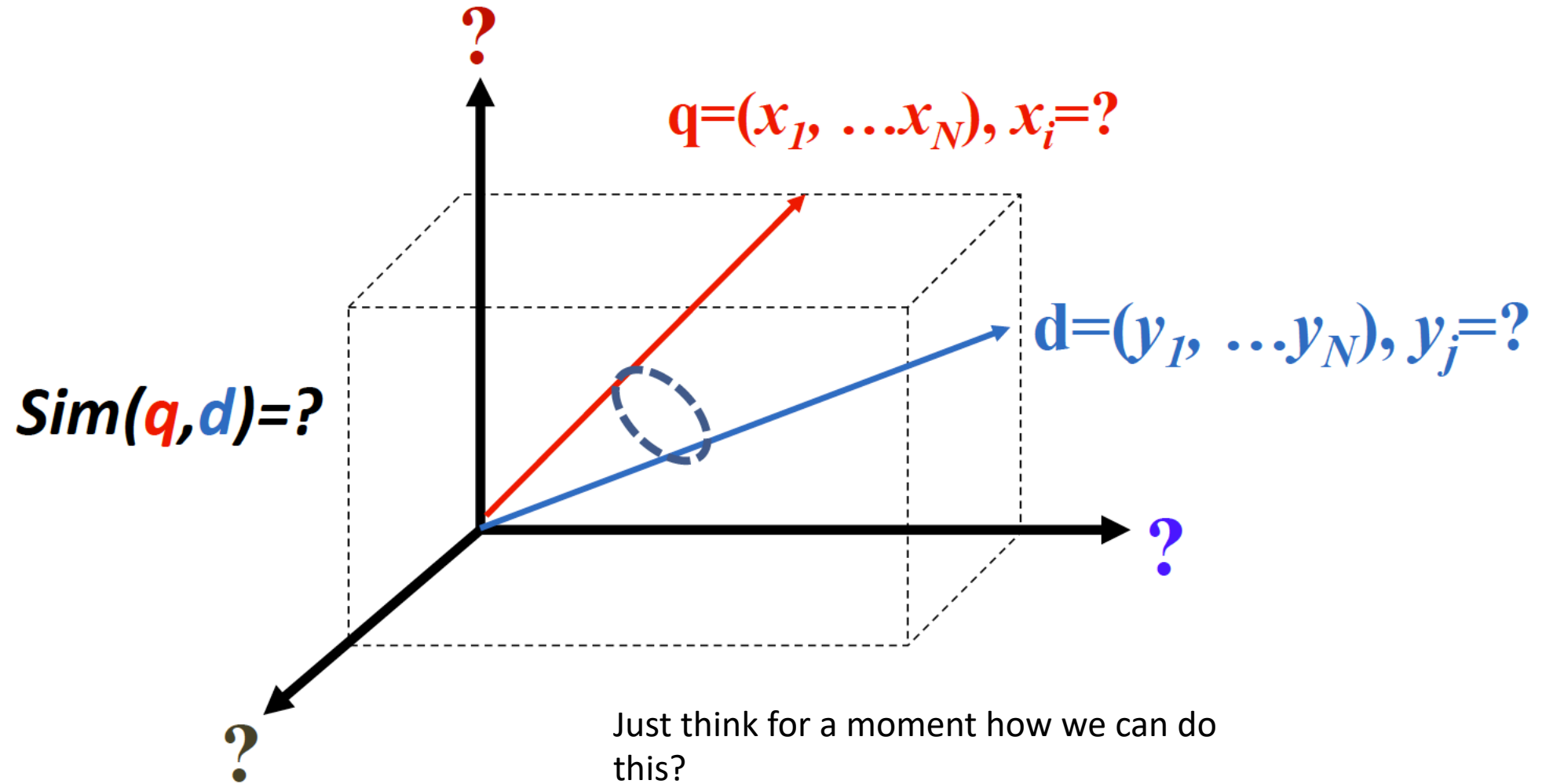
Vector Space Model Simplest Instantiation

Dr. Iqra Safder
Information Technology University

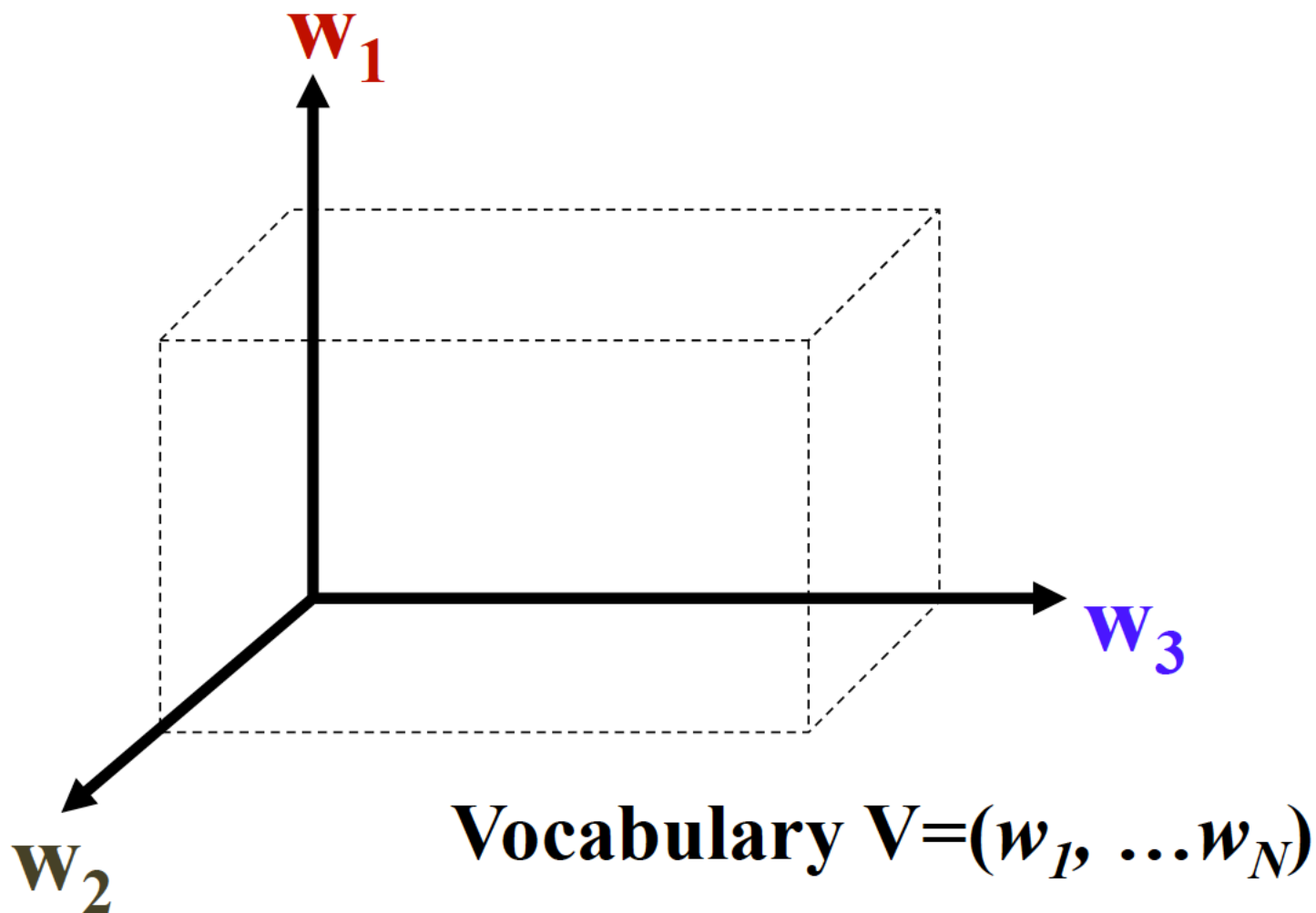
Course Schedule



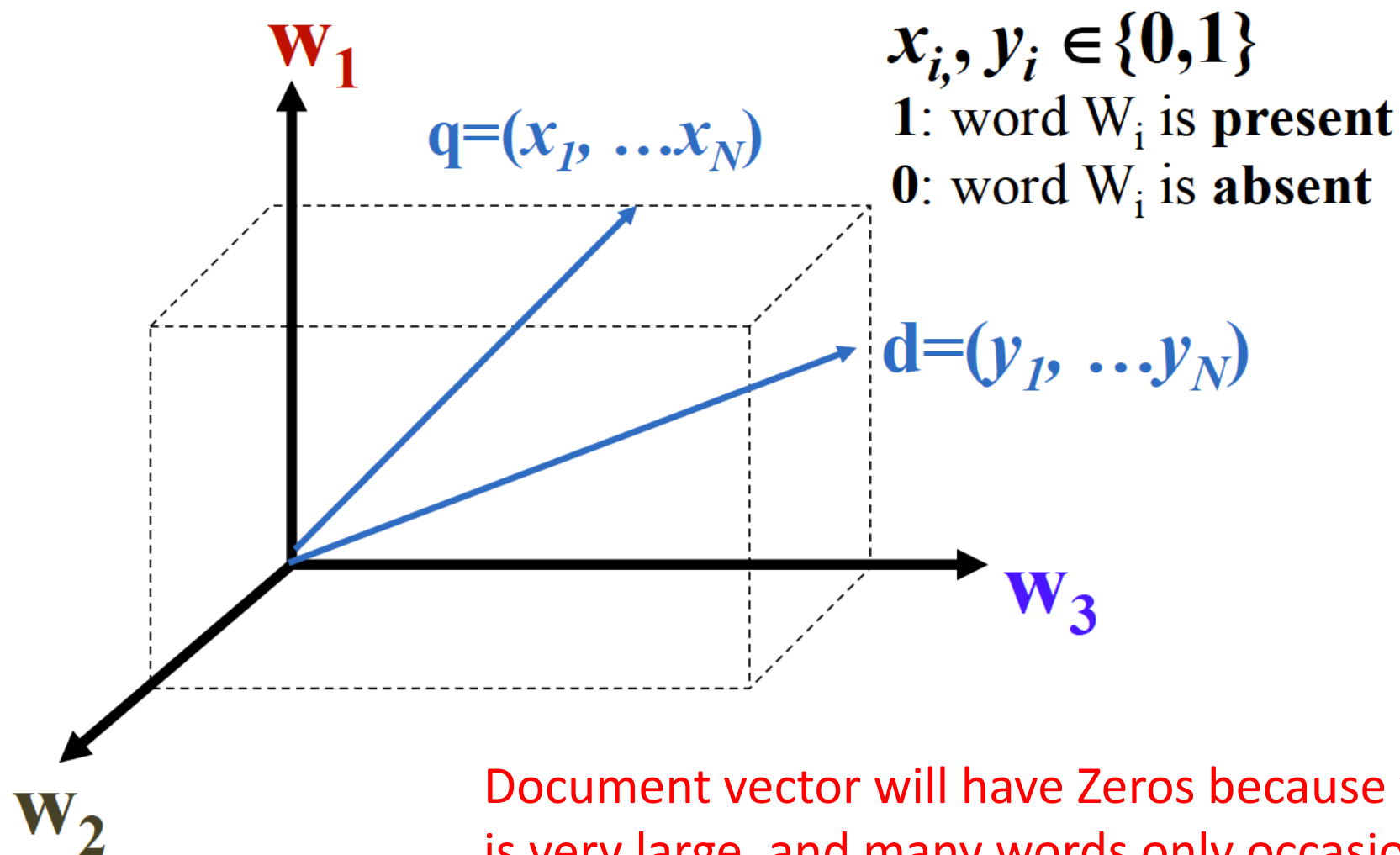
What VSM Doesn't Say



Dimension Instantiation: Bag of Words (BOW)



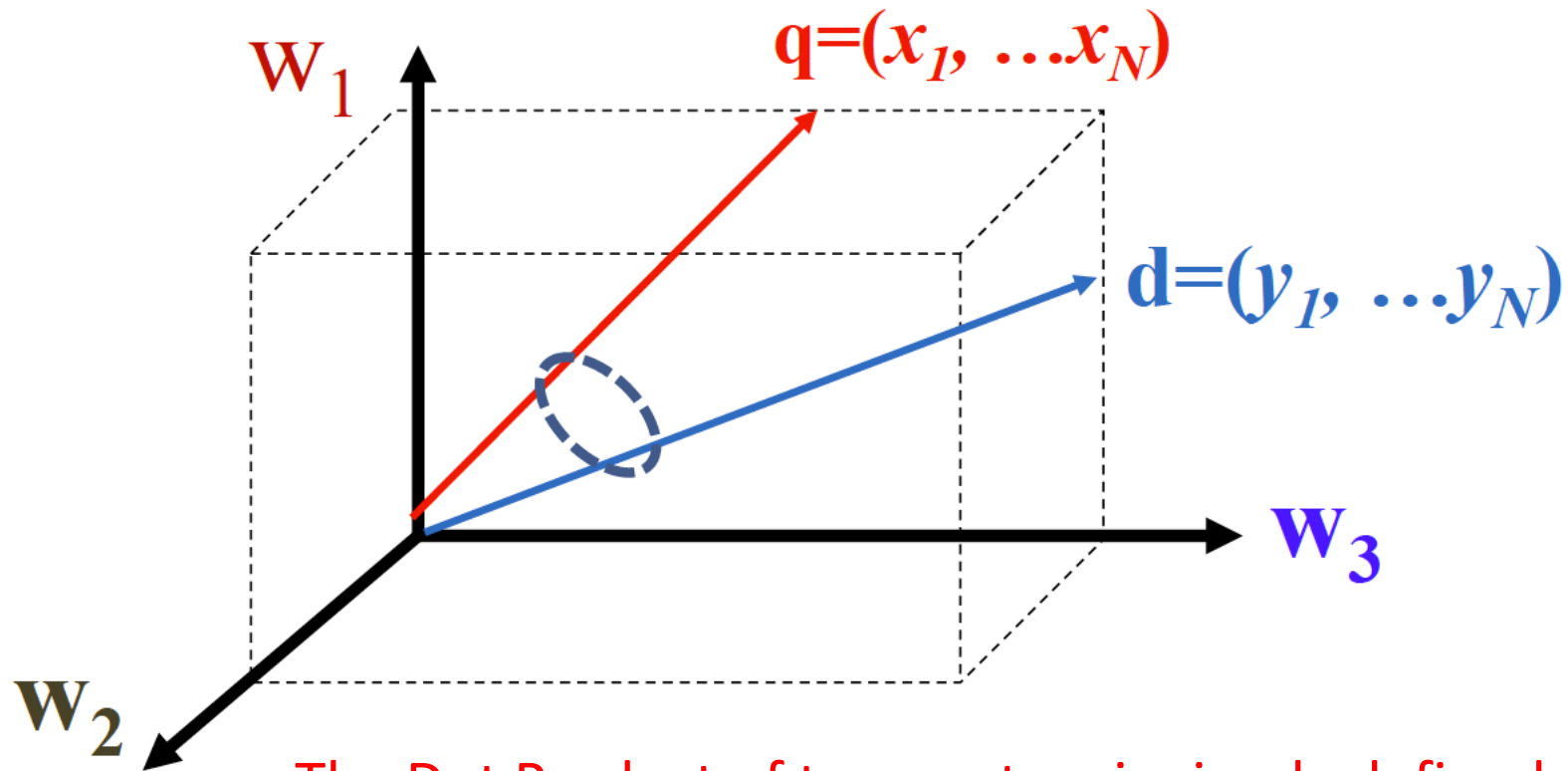
Vector Placement: Bit Vector



Document vector will have Zeros because vocabulary is very large, and many words only occasionally occurs in the document.

Similarity Instantiation: Dot Product

$$\text{Sim}(\mathbf{q}, \mathbf{d}) = \mathbf{q} \cdot \mathbf{d} = x_1 y_1 + \dots + x_N y_N = \sum_{i=1}^N x_i y_i$$



The Dot Product of two vectors is simply defined as the sum of the products of the corresponding elements of the two vectors.

Simplest VSM= Bit-Vector + Dot-Product + BOW

$$\begin{aligned} \mathbf{q} &= (x_1, \dots, x_N) & x_i, y_i &\in \{0, 1\} \\ \mathbf{d} &= (y_1, \dots, y_N) & 1: \text{word } W_i \text{ is present} \\ & & 0: \text{word } W_i \text{ is absent} \end{aligned}$$

$$\text{Sim}(\mathbf{q}, \mathbf{d}) = \mathbf{q} \cdot \mathbf{d} = x_1 y_1 + \dots + x_N y_N = \sum_{i=1}^N x_i y_i$$

What does this ranking function intuitively capture?
Is this a good ranking function?

An Example: How Would You Rank These Documents?

Query = “**news about presidential campaign**”

Ideal Ranking?

d1

... **news about** ...

d2

... **news about** organic food **campaign**...

d3

... **news** of **presidential campaign** ...

d4

... **news** of **presidential campaign** ...
... **presidential** candidate ...

d5

... **news** of organic food **campaign**...
campaign...**campaign**...**campaign**...

An Example: How Would You Rank These Documents?

Query = “ news about presidential campaign ”		Ideal Ranking?
d1	... news about ...	d4 + d3 +
d2	... news about organic food campaign ...	
d3	... news of presidential campaign ...	
d4	... news of presidential campaign presidential candidate ...	d1 - d2 -
d5	... news of organic food campaign ... campaign ... campaign ... campaign ...	d5 -

Ranking Using the Simplest VSM

Query = “**news about presidential campaign**”

d1 ... **news about** ...

d3 ... **news** of **presidential campaign** ...

$V = \{\text{news, about, presidential, campaign, food ...}\}$

$q = (1, 1, 1, 1, 0, \dots)$

$d1 = (1, 1, 0, 0, 0, \dots)$

$f(q, d1) = 1*1 + 1*1 + 1*0 + 1*0 + 0*0 + \dots = 2$

$d3 = (1, 0, 1, 1, 0, \dots)$

$f(q, d3) = 1*1 + 1*0 + 1*1 + 1*1 + 0*0 + \dots = 3$

Is the Simplest VSM Effective?

Query = “news about presidential campaign”

d1	... news about ...	$f(q, d1)=2$
d2	... news about organic food campaign ...	$f(q, d2)=3$
d3	... news of presidential campaign ...	$f(q, d3)=3$
d4	... news of presidential campaign presidential candidate ...	$f(q, d4)=3$
d5	... news of organic food campaign ... campaign ... campaign ... campaign ...	$f(q, d5)=2$

Summary

- VSM instantiation: dimension, vector placement, similarity
- Simplest VSM
 - Dimension = word
 - Vector = 0-1 bit vector (word presence/absence)
 - Similarity = dot product
 - $f(q,d)$ = number of **distinct** query words matched in d

Simple vector space model still doesn't work well, and we need to improve it. And this is a topic that we're going to cover in the next lecture.

Assignment # 1

Information Retrieval & Text Mining

Deadline: 14 Oct, 2021

Submit code file on google classroom

Find me @

Dr. Iqra Safder

Teaching Fellow

iqra.safder@itu.edu.pk

Office: 6th Floor, ITU

Information Technology University (ITU)



<http://ai.itu.edu.pk>