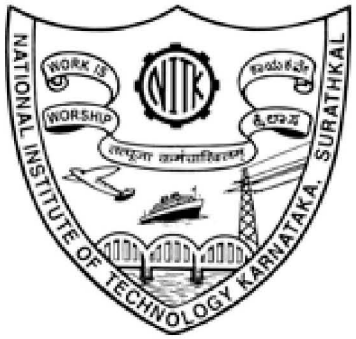
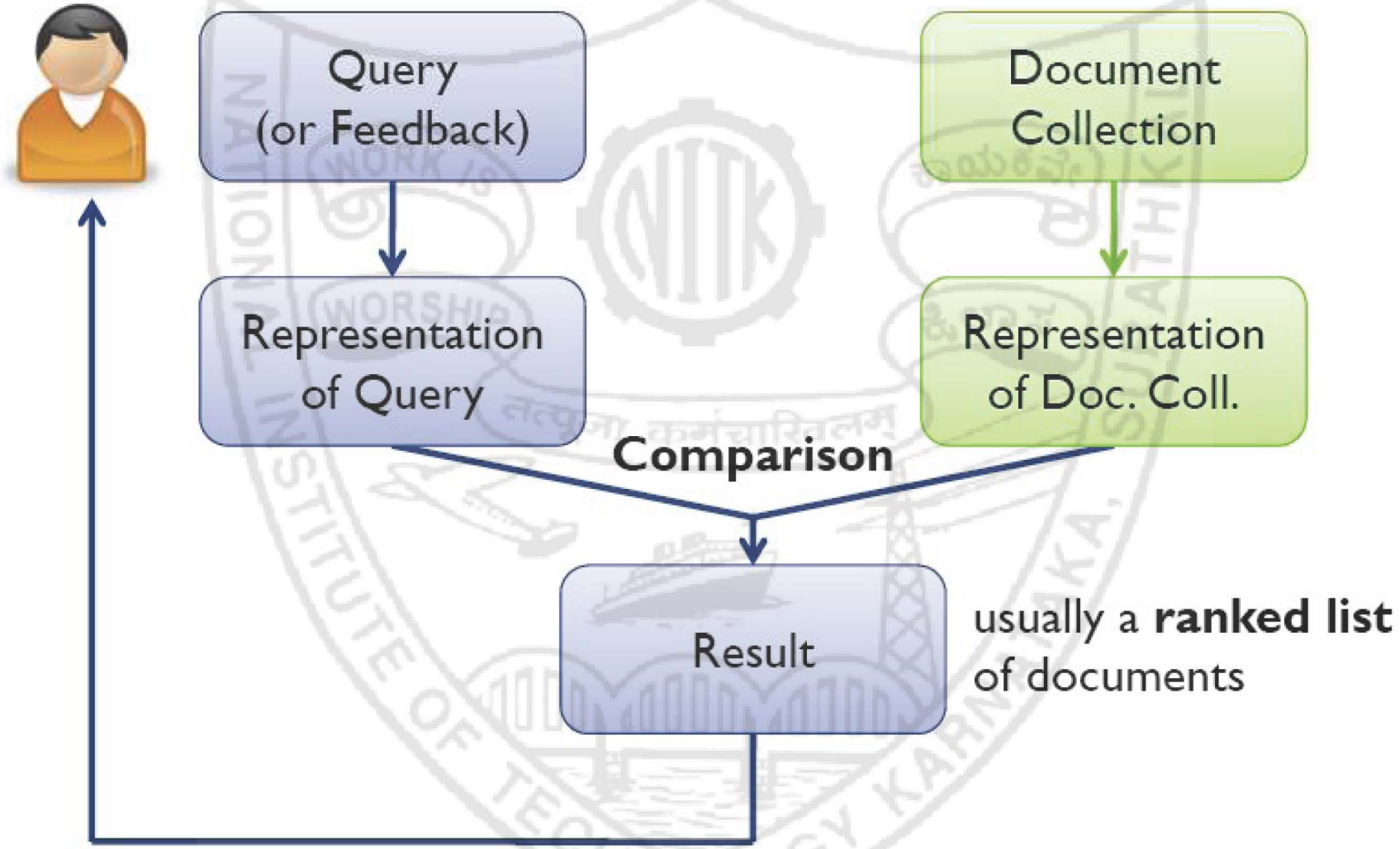
Jul – Nov 2022

IT458




# IR Models

# Generic IR Systems – High level Architecture



# Information Retrieval as a Process

- Document Representation **(Indexing)**
  - Given the documents, identify the concepts that describe the content and how well they describe it.

# Information Retrieval as a Process

- Document Representation **(Indexing)**
  - Given the documents, identify the concepts that describe the content and how well they describe it.
- Representing Information Need **(Query Formulation)**
  - describe and refine information needs as explicit queries

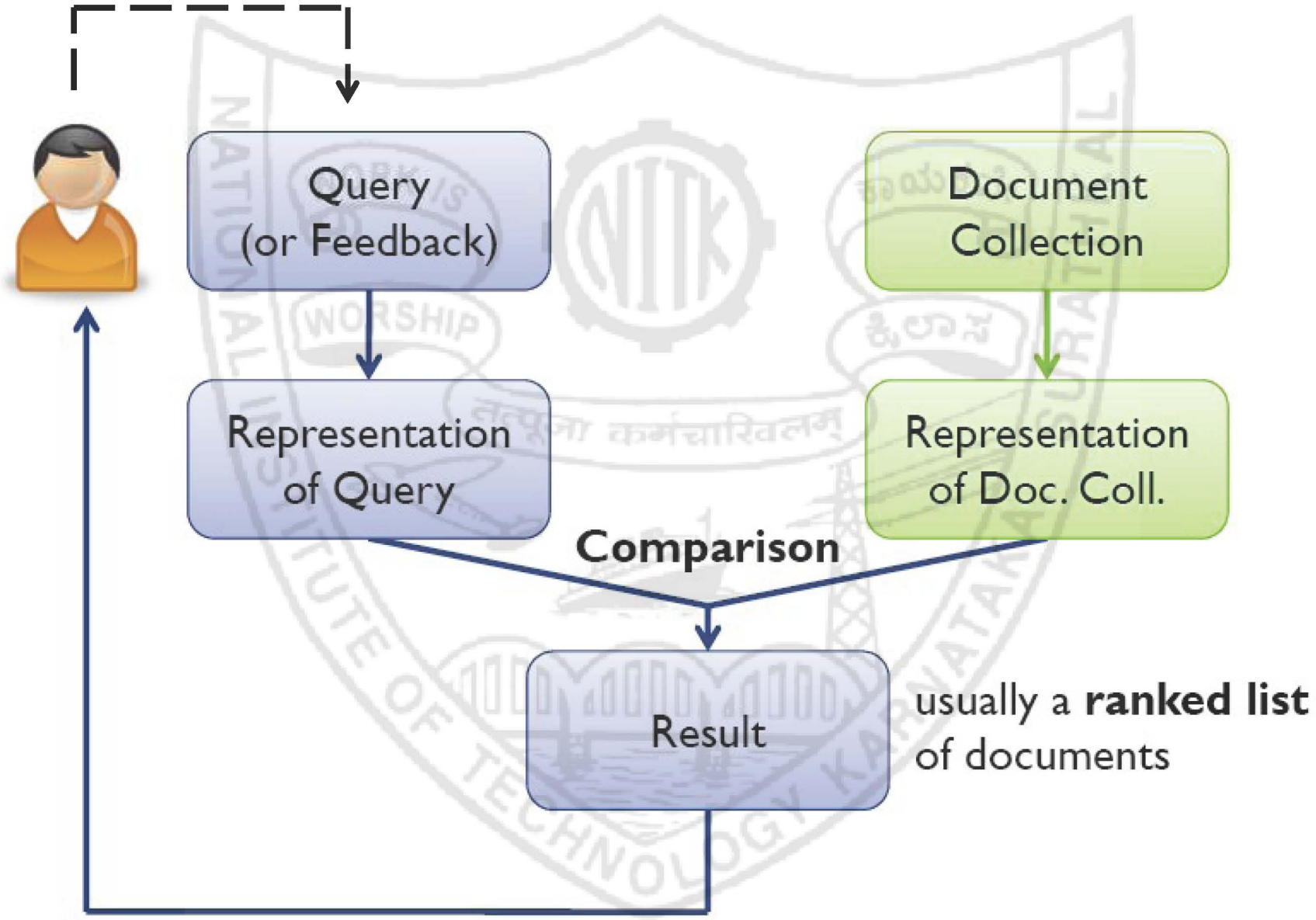
# Information Retrieval as a Process

- Document Representation **(Indexing)**
  - Given the documents, identify the concepts that describe the content and how well they describe it.
- Representing Information Need **(Query Formulation)**
  - describe and refine information needs as explicit queries
- Comparing Representations **(Retrieval)**
  - compare text and query representations to determine which documents are potentially relevant

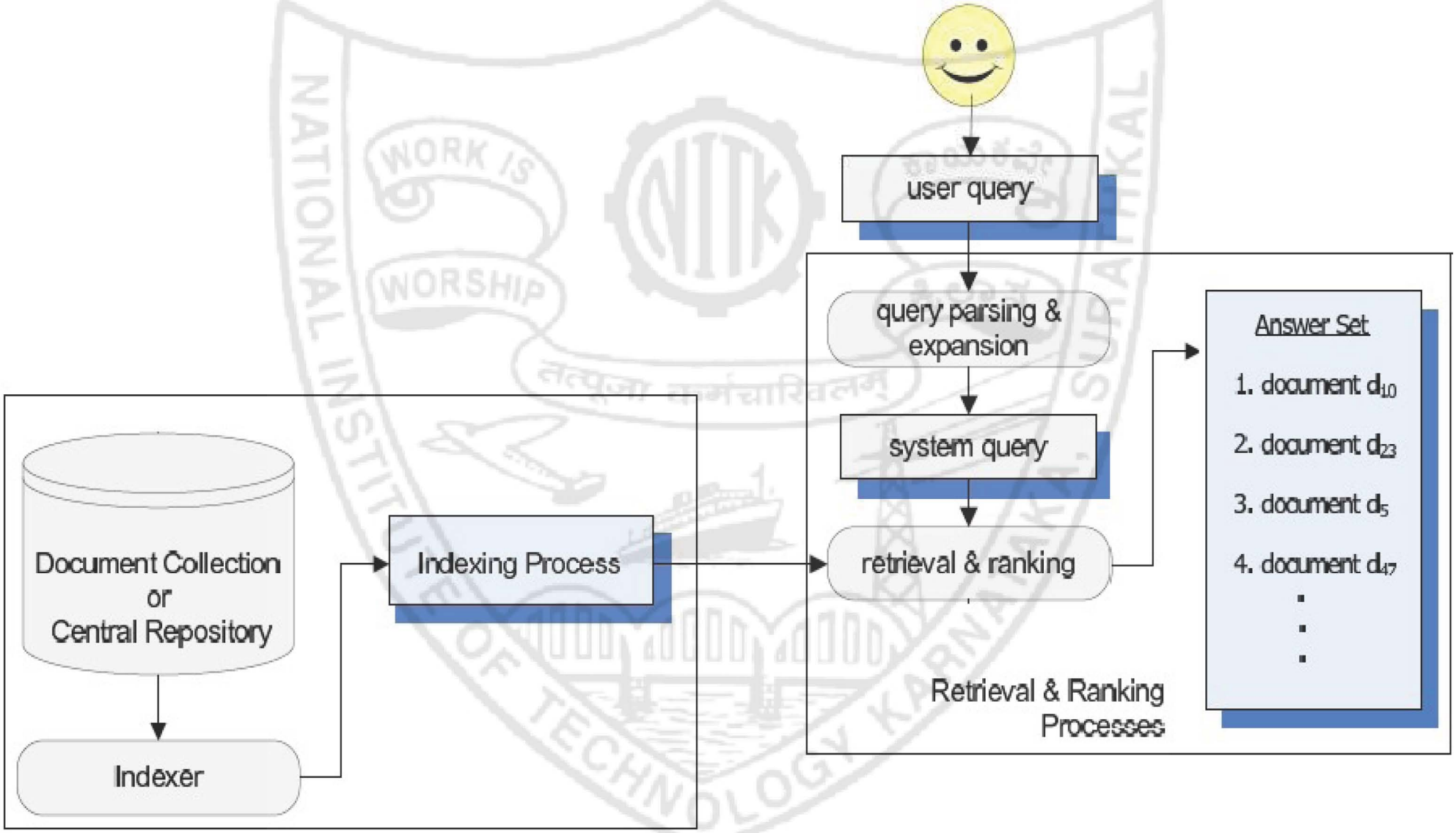
# Information Retrieval as a Process

- Document Representation **(Indexing)**
  - Given the documents, identify the concepts that describe the content and how well they describe it.
- Representing Information Need **(Query Formulation)**
  - describe and refine information needs as explicit queries
- Comparing Representations **(Retrieval)**
  - compare text and query representations to determine which documents are potentially relevant
- Evaluating Retrieved Text **(Relevance Feedback)**
  - present documents to user and modify query based on feedback

# Generic IR Systems – High level Architecture



# Generic IR Systems – Detailed view



# Modeling in IR

- consists of two main tasks:
  - Conception of a logical framework for representing documents and queries.
  - Defining a **ranking** function
    - quantifies the similarities among documents and queries.
    - A function that assigns scores to documents with regard to a given query.

* Every IR application is based on a specific/hybrid IR model.

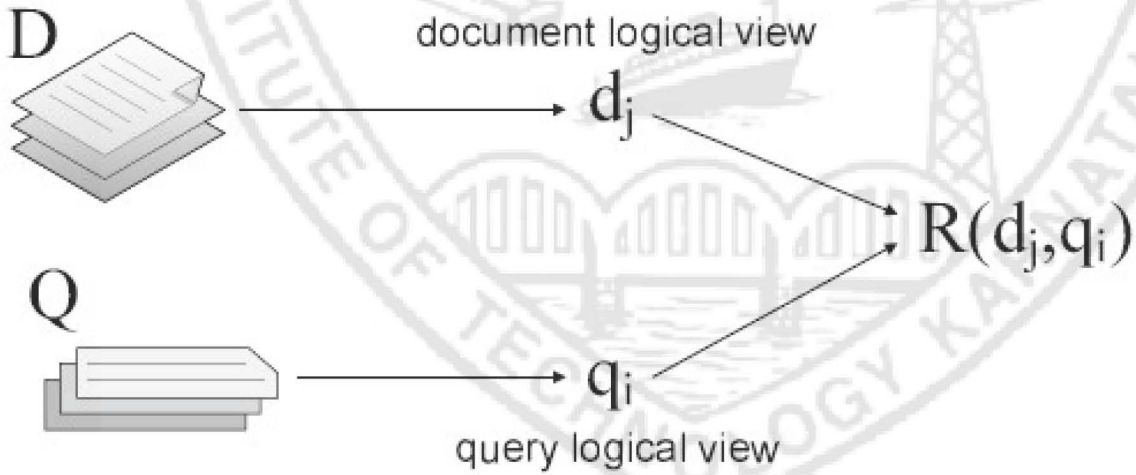
# IR Model

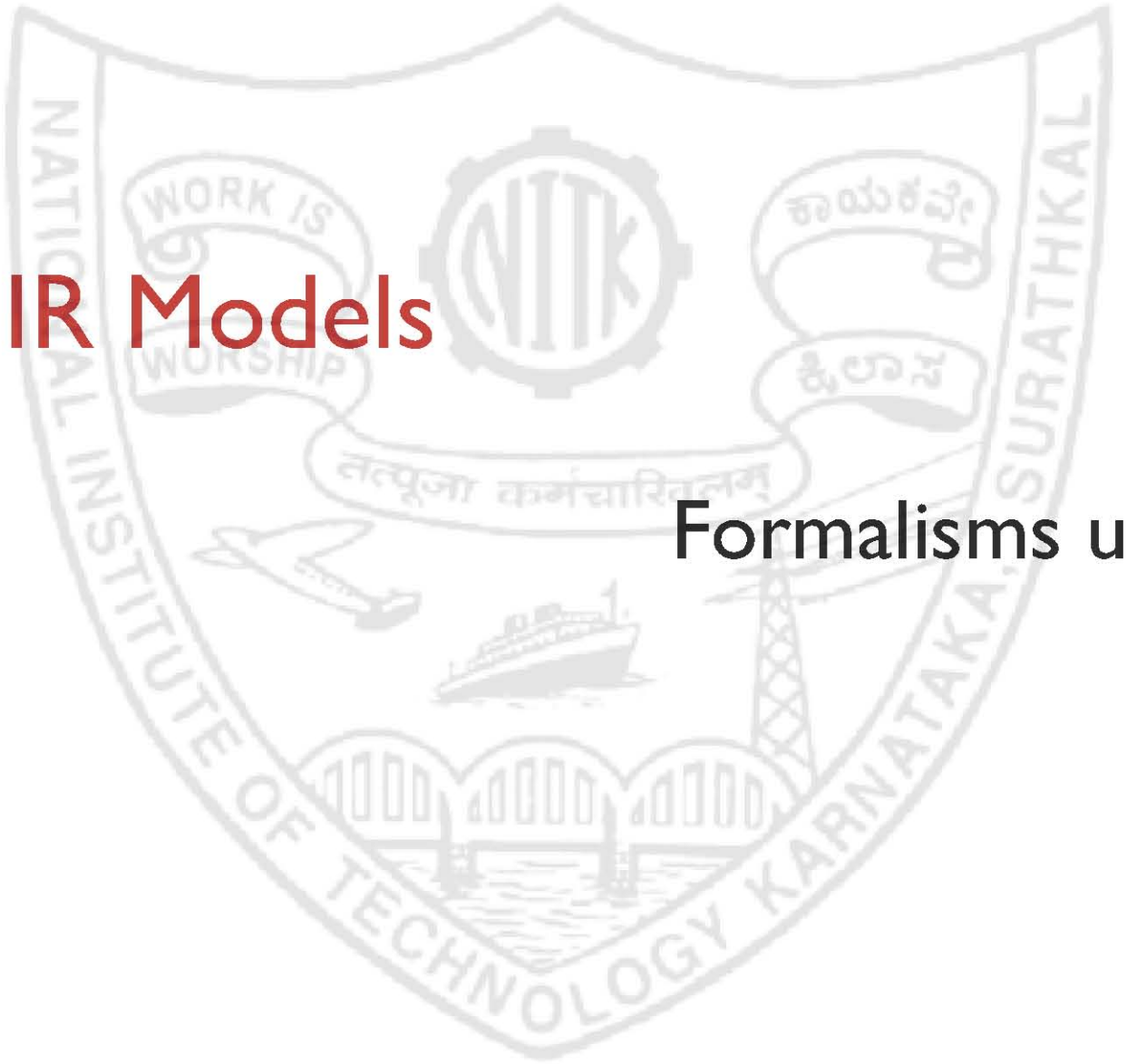
- The IR model defines …
  - a query language,
  - an internal representation of queries,
  - an internal representation of documents,
  - a ranking function which associates a real number with respect to each query–document pair.
  - *Optional*: A mechanism for relevance feedback
    - Notion of relevance can be *binary* or *continuous* (i.e. ranked retrieval)

# IR Model – Formal representation

An **IR model** is a quadruple $[\mathbf{D}, \mathbf{Q}, \mathcal{F}, R(q_i, d_j)]$ where

1. $\mathbf{D}$ is a set of logical views for the documents in the collection
2. $\mathbf{Q}$ is a set of logical views for the user queries
3. $\mathcal{F}$ is a framework for modeling documents and queries
4. $R(q_i, d_j)$ is a ranking function






# IR Models

## Formalisms used

# IR Model Formalisms - Index terms and Ranking

- **Index term:** used to index and retrieve documents
  - In a restricted sense:
    - it is a keyword that has some meaning on its own; usually plays the role of a noun.
  - In a more general form:
    - it is any word that appears in a document (full text representation)

# IR Model Formalisms - Index terms and Ranking

- **Ranking:**
  - an ordering of the documents that (hopefully) reflects their relevance w.r.t a user query.

# IR Model Formalisms - Vocabulary

Let,

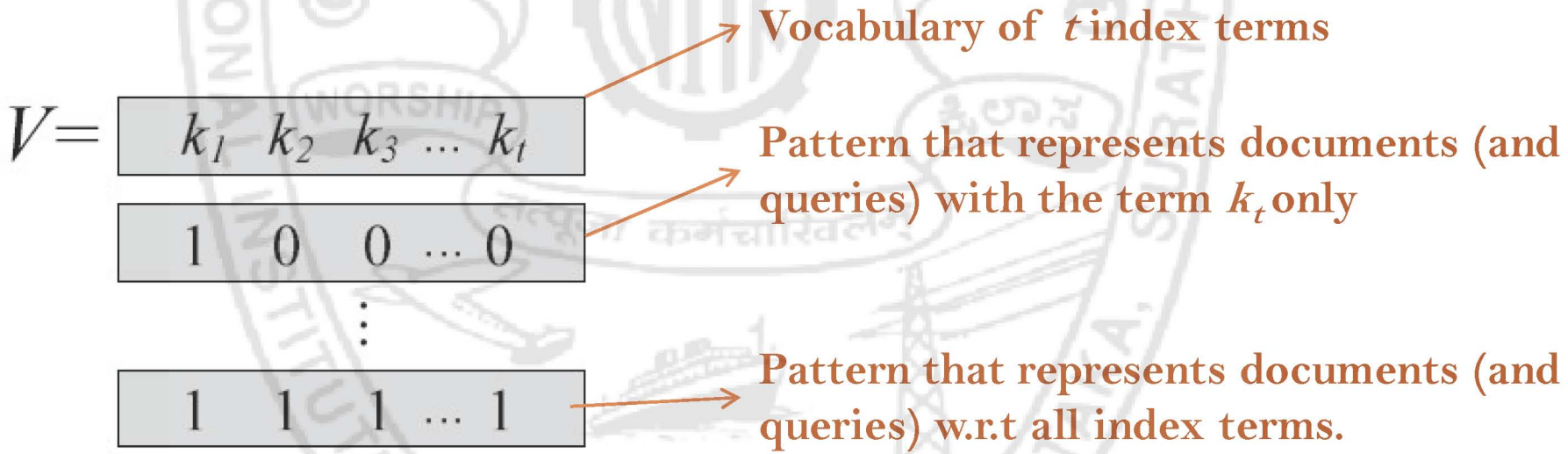
- $t$ be the **number of index terms** in the document **corpus**
- $k_i$ be a generic **index term** ($i$= 0 to $t$)

- Then, the **vocabulary** $V = \{k_1, \ldots, k_t\}$ is the set of all distinct index terms in the collection.

$$V = \boxed{k_1 \;\; k_2 \;\; k_3 \;\; \ldots \;\; k_t} \rightarrow \textbf{Vocabulary of } \boldsymbol{t} \textbf{ index terms}$$

# IR Model Formalisms – Term Co-occurrences

- Documents and queries can be represented by patterns of **term co-occurrences/term incidence.**




- Each of these patterns is called a **term conjunctive component.**

# IR Model Formalisms– *Term Document Matrix*

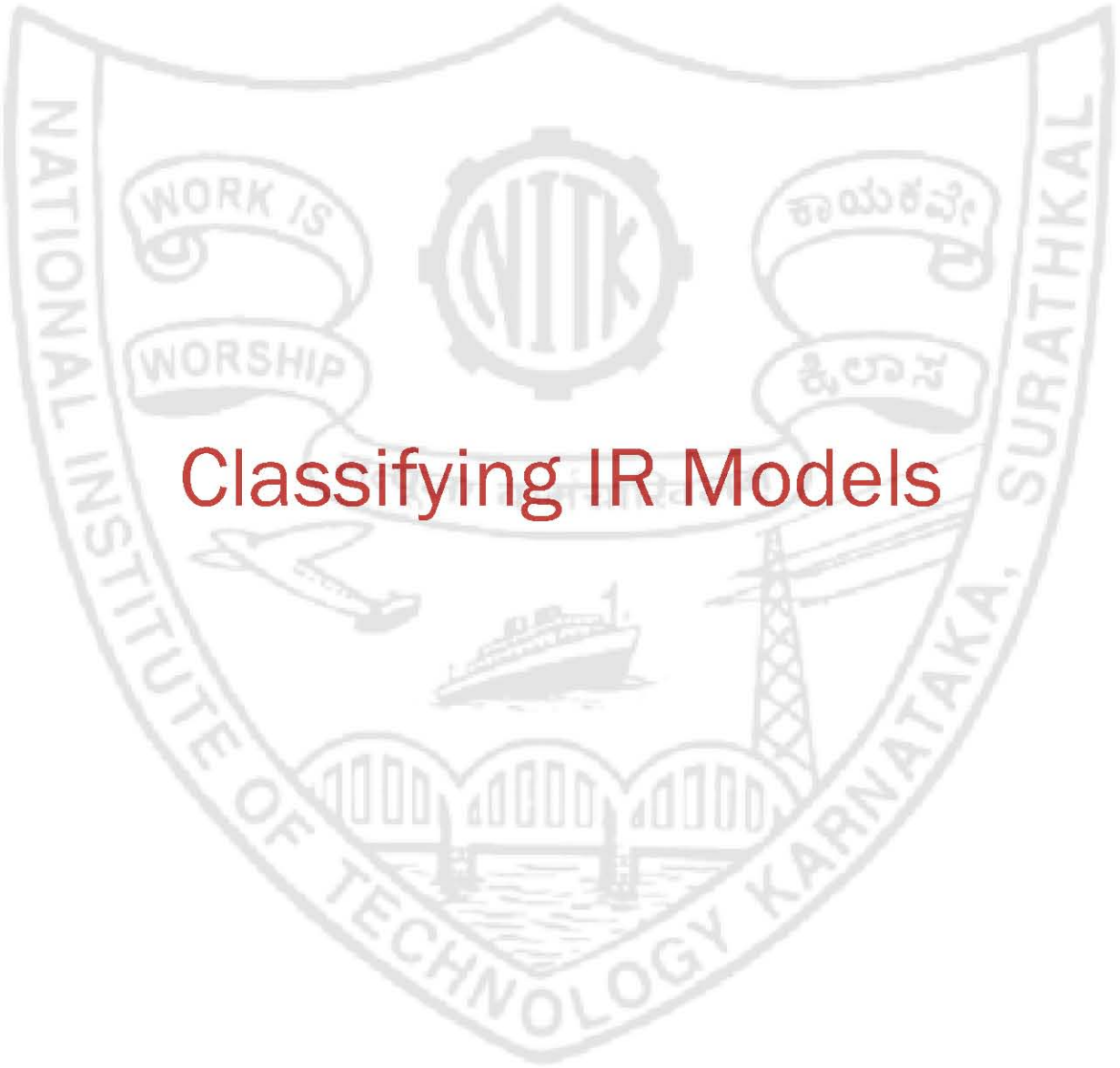
- If term $k_i$ occurs in a document $d_j$, then –
  - A **term-document relation** exists between $k_i$ and $d_j$.
  - captured by the frequency of the term in the document.

# IR Model Formalisms– *Term Document Matrix*

- In matrix form, this can written as

$$\begin{array}{c} \\ k_1 \\ k_2 \\ k_3 \end{array} \begin{array}{c} \begin{array}{cc} d_1 & d_2 \end{array} \\ \begin{bmatrix} f_{1,1} & f_{1,2} \\ f_{2,1} & f_{2,2} \\ f_{3,1} & f_{3,2} \end{bmatrix} \end{array}$$

  - where each $f_{i,j}$ element stands for the frequency of term $k_i$ in document $d_j$.




# Classifying IR Models

# Classifying IR Models

- Classic IR Models (unstructured text)
  - Bag of Words
  - Boolean
  - Vector Space
  - Probabilistic

# Classifying IR Models

- Classic IR Models (unstructured text)
  - Bag of Words
  - Boolean
  - Vector Space
  - Probabilistic

# Classifying IR Models

- Classic IR Models (unstructured text)
  - Bag of Words
  - Boolean
  - Vector Space
  - Probabilistic
- Semi-structured Text
  - Proximal nodes
  - XML based

# Classifying IR Models

- Classic IR Models (unstructured text)
  - Bag of Words
  - Boolean
  - Vector Space
  - Probabilistic
- Semi-structured Text
  - Proximal nodes
  - XML based
- Link Analysis based (Web)
  - Hubs and Authorities
  - PageRank

# Classifying IR Models

- Classic IR Models (unstructured text)
  - Bag of Words
  - Boolean Model
  - Vector Space
  - Probabilistic
- Semi-structured Text
  - Proximal nodes
  - XML based
- Link Analysis based (Web)
  - Hubs and Authorities
  - PageRank
- Multimedia Retrieval
  - Image/Audio/Video Retrieval models (adaptations of classical models)