Chapter 8
Scalable and Emerging Information System Techniques

- Cloud data management is a way to manage data across cloud platforms, either with or instead of on-premises storage.
- The cloud is useful as a data storage tier for disaster recovery, backup, and long-term archiving.
- With data management in the cloud, resources can be purchased as needed.
- Data stored in the cloud has its own rules for data integrity and security.

CLOUD MANAGEMENT COMPONENTS

AUTOMATION AND ORCHESTRATION

- Application migration
- VM images/instances
- Configuration management

SECURITY

- IAM
- Encryption
- Mobile/endpoint security

COST MANAGEMENT

- Cloud instance right sizing
- User chargeback and billing

PERFORMANCE MONITORING

- Storage
- Networks
- Applications
- Compute

GOVERNANCE AND COMPLIANCE

- Risk assessment/ threat analysis
- Audits
- Service and resource governance

ICONS: [illegible]/ADOBE STOCK

- **Cloud Data Management Challenges:**
  - Security - Concern with cloud technology
  - Availability of a Service, Data Confidentiality, Data lock-in, Performance unpredictability
- **Benefits:**
  - Backup, disaster recovery, archiving and analytics
  - some companies also offer Ransomware protection
  - Availability, Scalability, Elasticity, Performance, Fault tolerance, Ability to run in a heterogeneous environment

Data Management in Cloud

There are three characteristics of a cloud computing environment.

- Compute power is elastic
  - Computer resource can be scaled up and down
- Data is stored at untrusted host
  - Subject to local rules and regulations
- Data is replicated, often across large geographic distances

Components of Data Management Market

1. Transactional Data Management
   - Banks, airline reservation, online e-commerce
   - ACID
   - Not ready to move to the cloud for the following reasons:

   – Hard to maintain ACID when data replication are all over the world

   – Enormous risks in storing transactional data on an untrusted host

- **ACID** - acronym used to describe the four properties of an enterprise-level transaction:
- ATOMICITY: a transaction should be done or undone completely. In the event of a failure, all operations and procedures should be undone, and all data should rollback to its previous state.
- CONSISTENCY: a transaction should transform a system from one consistent state to another consistent state.
- ISOLATION: each transaction should happen independently of other transactions occurring at the same time.
- DURABILITY: Completed transactions should remain permanent, even during system failure.

2. Analytical data management

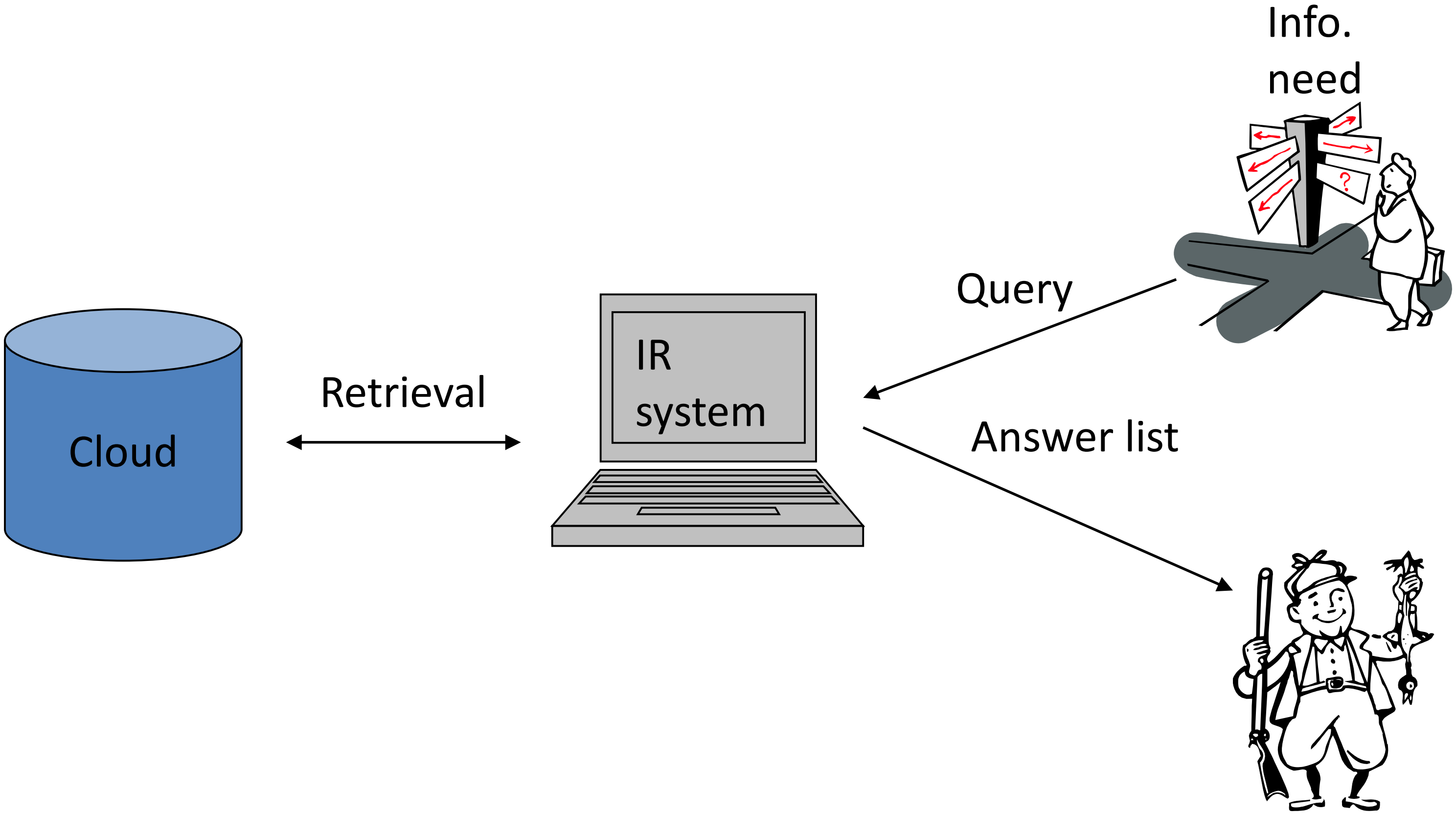
- Business planning, decision support
- well-suited to run in a cloud environment
  - ACID guarantees are typically not needed
  - particularly sensitive data can be left out of the cloud.

Information Retrieval

- **Information retrieval** is the activity of obtaining information resources relevant to an information need from a collection of information resources.
- Searches can be based on metadata or on full-text indexing.
- Automated information retrieval systems are used to reduce what has been called “information overload”.

Information Retrieval from Cloud

- **Goal** = find documents *relevant* to an information need from a Cloud



Info. need

Query

Retrieval

IR system

Cloud

Answer list

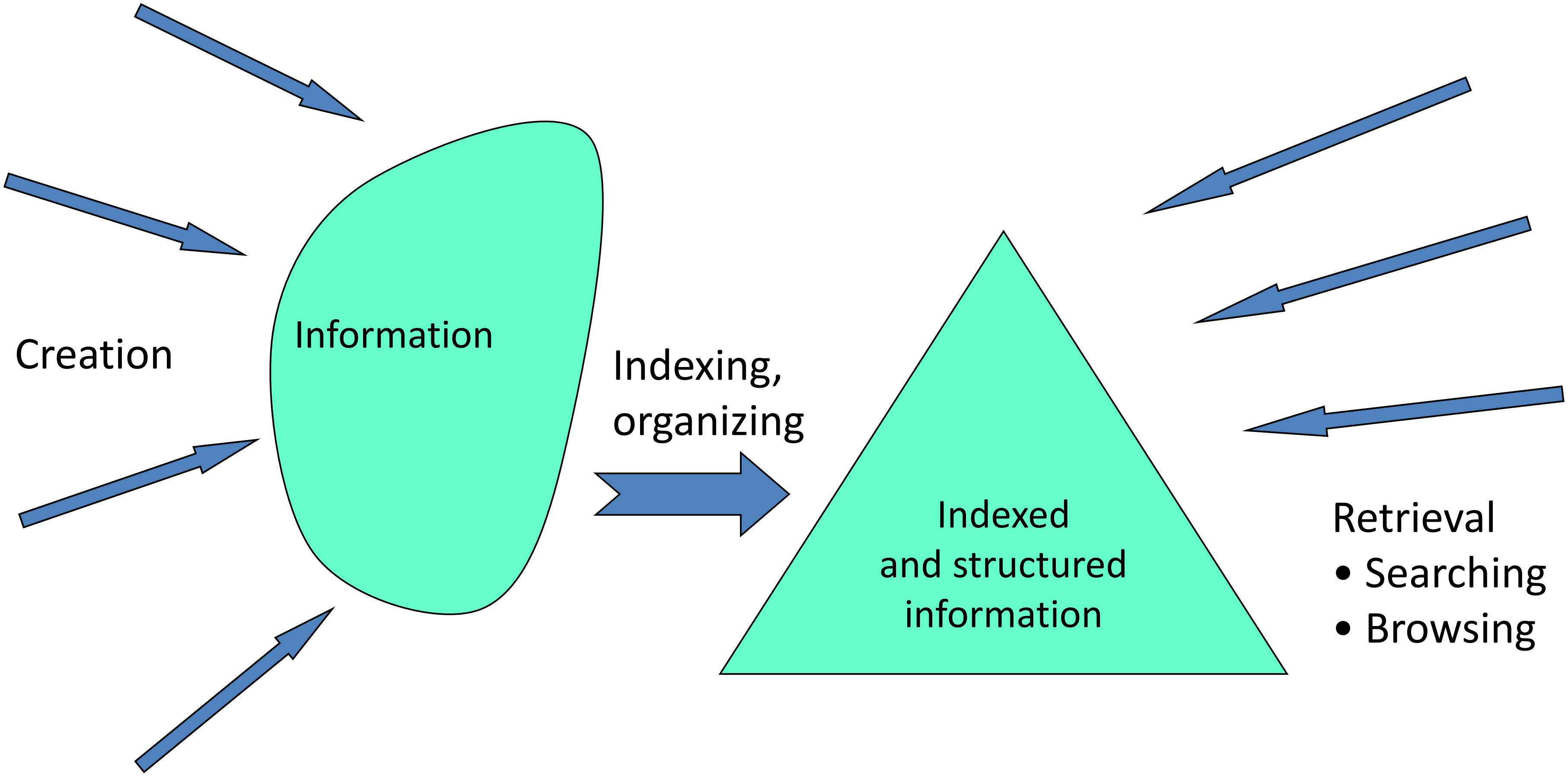
Information Retrieval in the Cloud

- IR user seeks actively information, pulling at it, by means of **querying or browsing**.
- In **tag querying**, user enters one or more tags in the search box to obtain an ordered list of resources which were in relation with these tags.
- When a user is scanning this list, the system also provide a list of related tags (i.e. tags with a high degree of co-occurrence with the original tag), allowing hypertext Browsing.

Information Retrieval System

- Typically it refers to the automatic (rather than manual) retrieval of documents
  - Information Retrieval System (IRS)
- Information Retrieval is a research-driven theoretical and experimental discipline
  - The focus is on different aspects of the information–seeking process, depending on the researcher’s background or interest:
    - Computer scientist – fast and accurate search engine
    - Librarian – organization and indexing of information
    - Cognitive scientist – the process in the searcher’s mind
    - Philosopher – Is this really relevant ?

The stages of IR



Creation

Information

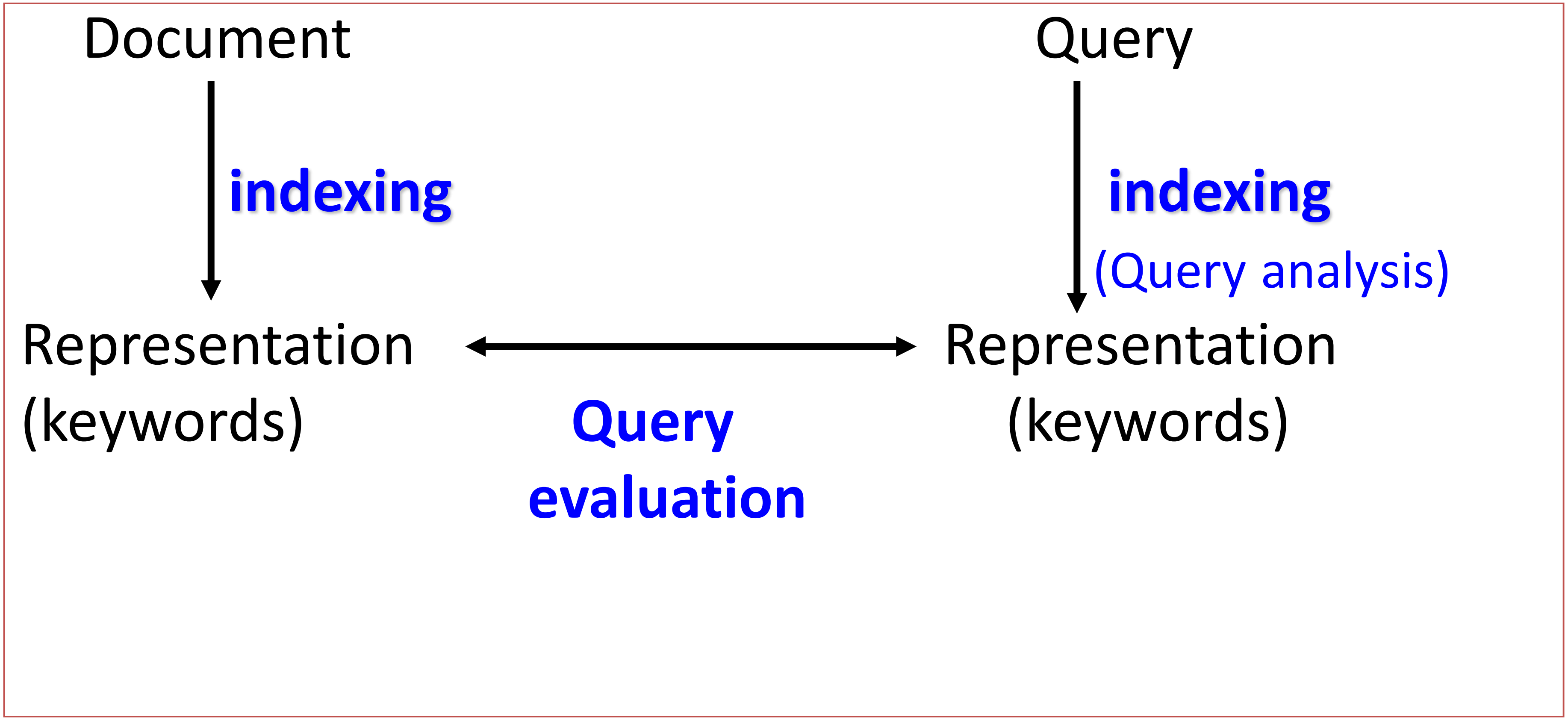
Indexing, organizing

Indexed and structured information

Retrieval

- Searching
- Browsing

Indexing based IR



Document

indexing

Representation (keywords)

Query

indexing

(Query analysis)

Representation (keywords)

Query evaluation

Main problems in IR

- Document and query indexing
  - How to best represent their contents?
- Query evaluation (or retrieval process)
  - To what extent does a document correspond to a query?
- System evaluation
  - How good is a system?
  - Are the retrieved documents relevant? (precision)
  - Are all the relevant documents retrieved? (recall)

Three major components

1. Document subsystem
   - Acquisition, Representation, File organization
2. User sub system
   - Problem, Representation, query
3. Searching/Retrieval subsystem
   - Matching, Retrieved objects

Thank you