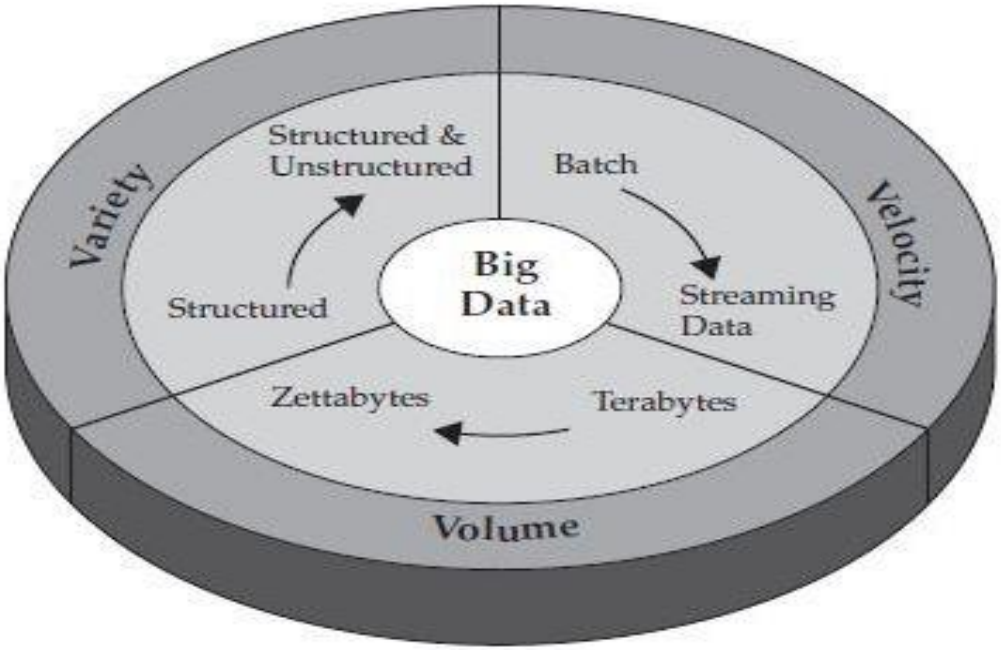
Scalable and Emerging Information System Techniques

Big Data

- Big Data applies to information that can't be processed or analyzed using traditional processes or tools. (3V)



IBM characterizes Big Data by its volume, velocity, and variety—or simply,

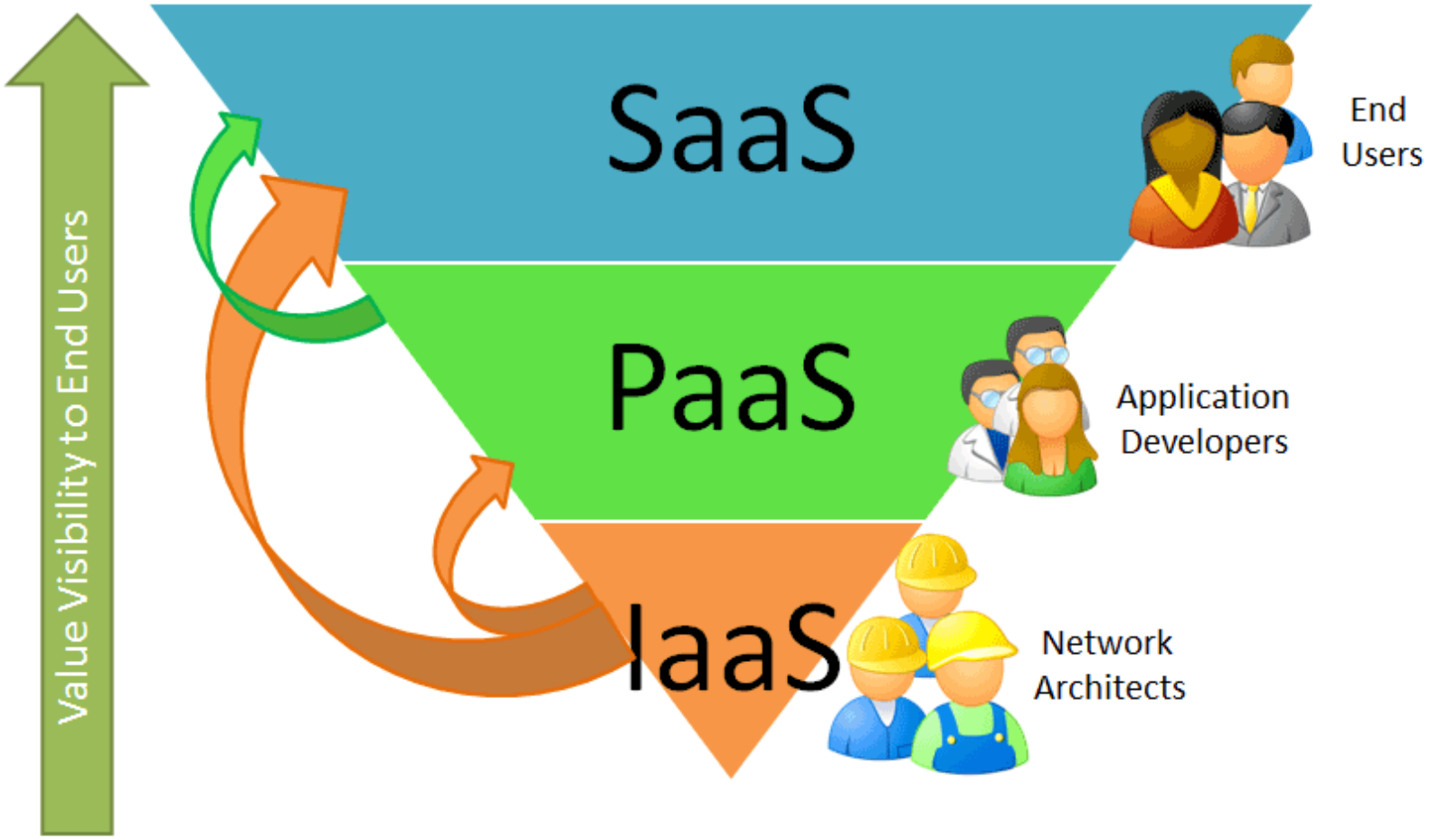
Techniques for Voluminous Data

- Cloud Computing is an efficient method to balance between dealing with voluminous data and keeping costs competitive, is designed to deliver IT services consumable on demand, is scalable as per user need and uses a pay-per-use model.
- Business houses are progressively turn towards retaining core competencies, and shedding the non-core competencies for on-demand technology, business innovation and savings.

Cloud Computing

- Cloud computing consists of hardware and software resources made available on the internet as managed third-party services.
- These services typically provide access to advanced software applications and high-end networks of server computers.
- Cloud computing is comparable to grid computing, a type of computing where unused processing cycles of all computers in a network are connect to solve problems too intensive for any stand-alone machine.

Cloud Computing Layers



Types of Cloud Computing

SaaS (Software as a Service): This is the idea of providing a given application to multiple tenants, typically using the browser which supports business applications of host and delivery type as a service. End Customers for instance Google Doc, Myspace.com

Common features of SaaS:

a) User applications run on cloud infrastructure

b) Accessible by users through web browser

c) Suitable for CRM (Customer Resource Management) applications

d) Supports multi-tenant environment

Types of Cloud Computing

PaaS (Platform as a Service): This is a variant of SaaS. You run your own applications but you do it on the cloud provider's infrastructure. Provides a comprehensive stack for developers to create Cloud-ready business applications.

Developers for instance Force.com, Google App Engine, Azure and Salesforce.com etc

Features of PaaS are:

a) Supports web-service standards

b) Dynamically scalable as per demand

c) Supports multi-tenant environment

Types of Cloud Computing

IaaS (Infrastructure as a Service): These are virtual storage and server options that organizations can access on demand, even allowing the creation of a virtual data center. Delivers computing hardware like Servers, Network, Storage, etc. For instance Rackspace.com, GoGrid.com etc.

Typical features are:

a)Users use resources but have no control of underlying cloud infrastructure

b)Users pay for what they use

c)Flexible scalable infrastructure without extensive pre-planning

Benefit of Cloud Computing

- Reduced Cost : Cloud technology is paid incrementally, saving organizations money.
- Increased Storage: Organizations can store more data than on private computer systems.
- Highly Automated: No longer do IT personnel need to worry about keeping software up to date.
- Flexibility: Cloud computing offers much more flexibility than past computing methods.
- More Mobility: Employees can access information wherever they are, rather than having to remain at their desks.
- Allows IT to Shift Focus: No longer having to worry about constant server updates and other computing issues, government organizations will be free to concentrate on innovation.

MapReduce

- MapReduce is a software framework that allows developers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand-alone computers.
- It was developed at Google for indexing Web pages and replaced their original indexing algorithms and heuristics in 2004.

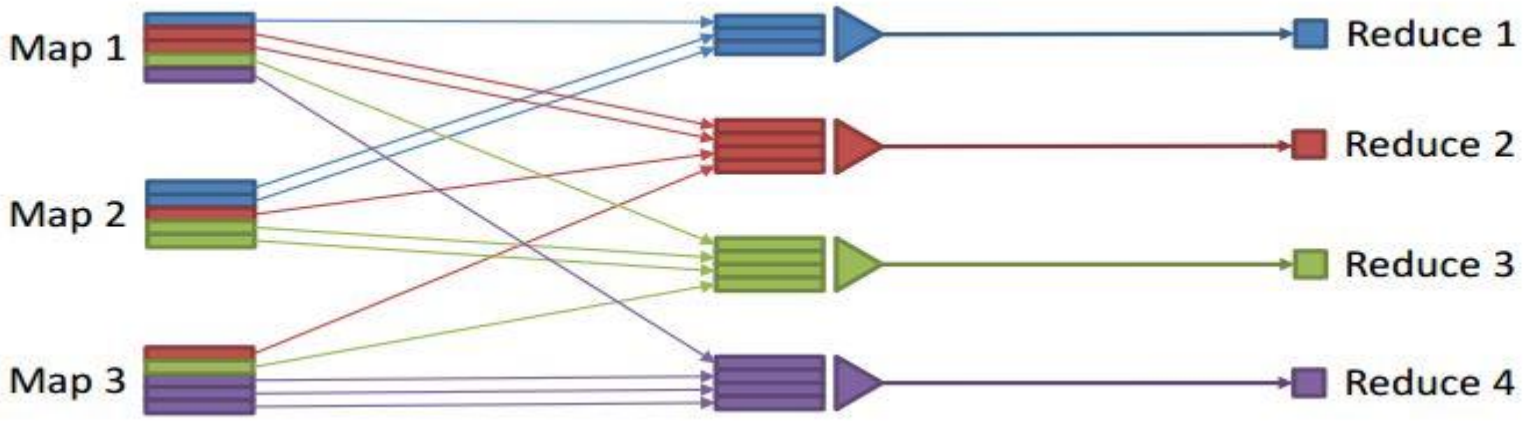
MapReduce and Hadoop Systems

- MapReduce is the heart of Hadoop.
- MapReduce allows data to be distributed across a large cluster, and can distribute out tasks across the data set to work on pieces of it independently, and in parallel.
- This allows big data to be processed in relatively little time.
- Apache has produced an open source MapReduce platform called **Hadoop**.

Framework is divided into two parts

- **Map**, a function that parcels out work to different nodes in the distributed cluster.
- **Reduce**, another function that collates the work and resolves the results into a single value.
- The MapReduce framework is fault-tolerant because each node in the cluster is expected to report back periodically with completed work and status updates. If a node remains silent for longer than the expected interval, a master node makes note and re-assigns the work to other nodes.

MapReduce Interaction



Map functions create a user-defined “index” from source data

- Reduce functions compute grouped aggregates based on index
- Flexible framework
  - users can cast raw original data in any model that they need
  - wide range of tasks can be expressed in this simple framework

Hadoop System

- Developed by Apache as an open source distributed MapReduce platform, based off of Google's MapReduce.
- Runs on a Java architecture framework that supports the processing of large data sets in a distributed computing environment.
- Hadoop allows businesses to process large amounts of data quickly by distributing the work across several nodes.
- Good for Big data sets and on large cluster.

Hadoop - A Key Business Tool

Hadoop System is used by Large Content-Distribution Companies, such as...

Yahoo

- Hadoop is used for many of their tasks, and over 25,000 computers are running Hadoop.

Amazon

- Hadoop is good for Amazon, they have lots of product data, as well as user-generated content to index, and make searchable.

New York Times

- Hadoop is used to perform large-scale image conversions of public domain articles.

Hadoop - A Key Business Tool

Used by Non-Content-Distribution Companies, such as

- Facebook
- eHarmony

Other early adopters include anyone with big data:

- medical records
- tax records
- network traffic
- large quantities of data

Wherever there is a lot of data, a Hadoop cluster can generally process it relatively quickly.

Data Management in the Cloud

- Data management applications are potential candidates for deployment in the cloud
  - Industry: enterprise database system have significant up-front cost that includes both hardware and software costs
  - Academia: manage, process and share mass-produced data in the cloud
- Many “Cloud Killer Apps” are in fact data-intensive
  - Batch Processing as with MapReduce
  - Online Transaction Processing (OLTP) as in automated business applications
  - Offline Analytical Processing (OLAP) as in data mining or machine learning

Data Management in the cloud

- A database system must implement for it to run well in the cloud, in potential database applications to consider for cloud deployment.
- Data management applications are best suited for deployment on top of cloud computing infrastructure.
- Data management is the proper management of a data resource for an organization. Data management consists of a set of theories, concepts, principles, and techniques for properly managing data.
- The primary objective is to support the business information as needs of the organization.

Data Management in Cloud

There are three characteristics of a cloud computing environment.

- Compute power is elastic, but only if workload is parallelizable.
- Data is stored at untrusted host.
- Data is replicated, often across large geographic distances

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”.

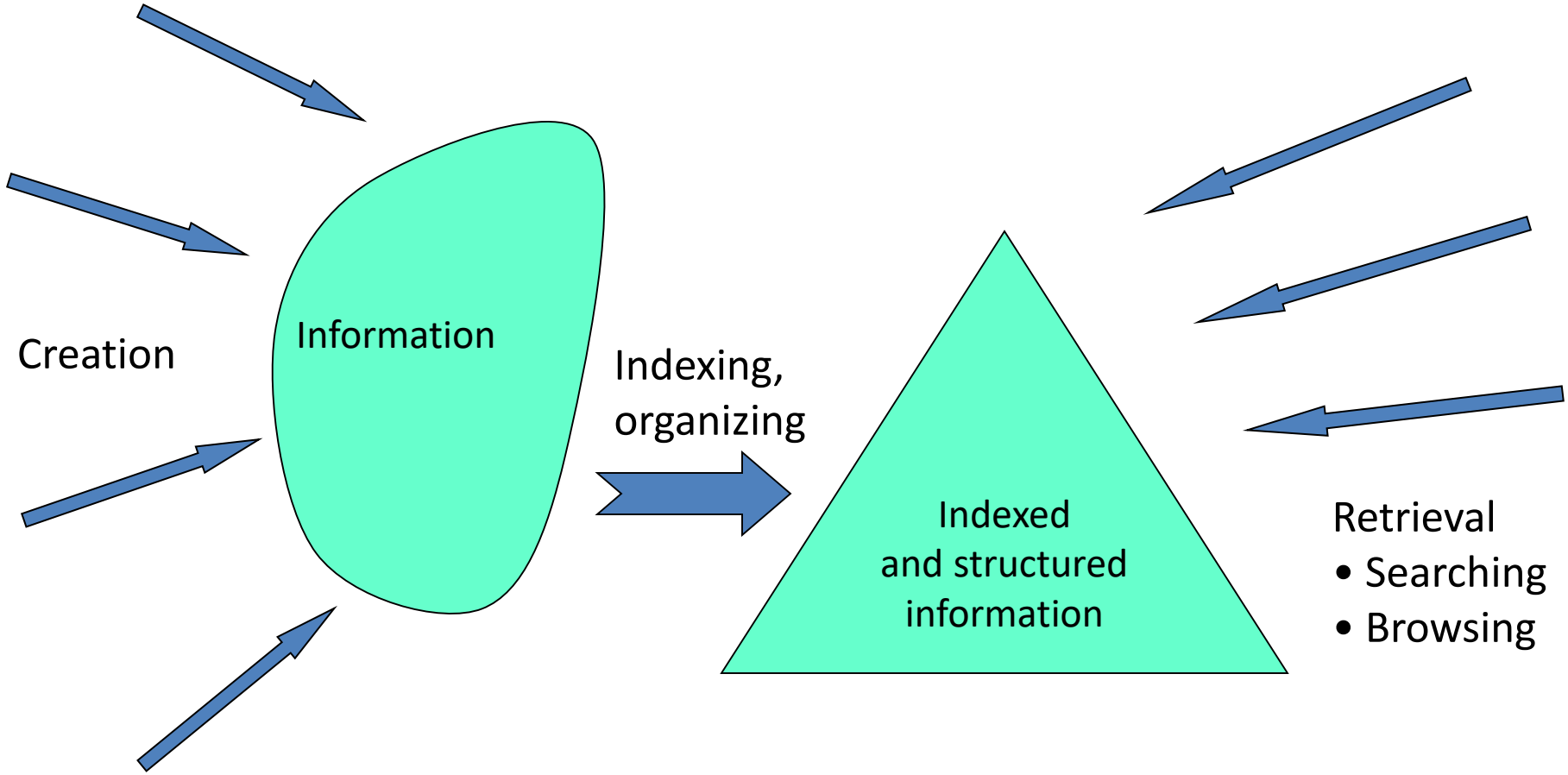
Information Retrieval in the Cloud

- 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.
- Example of index: index at last page of books

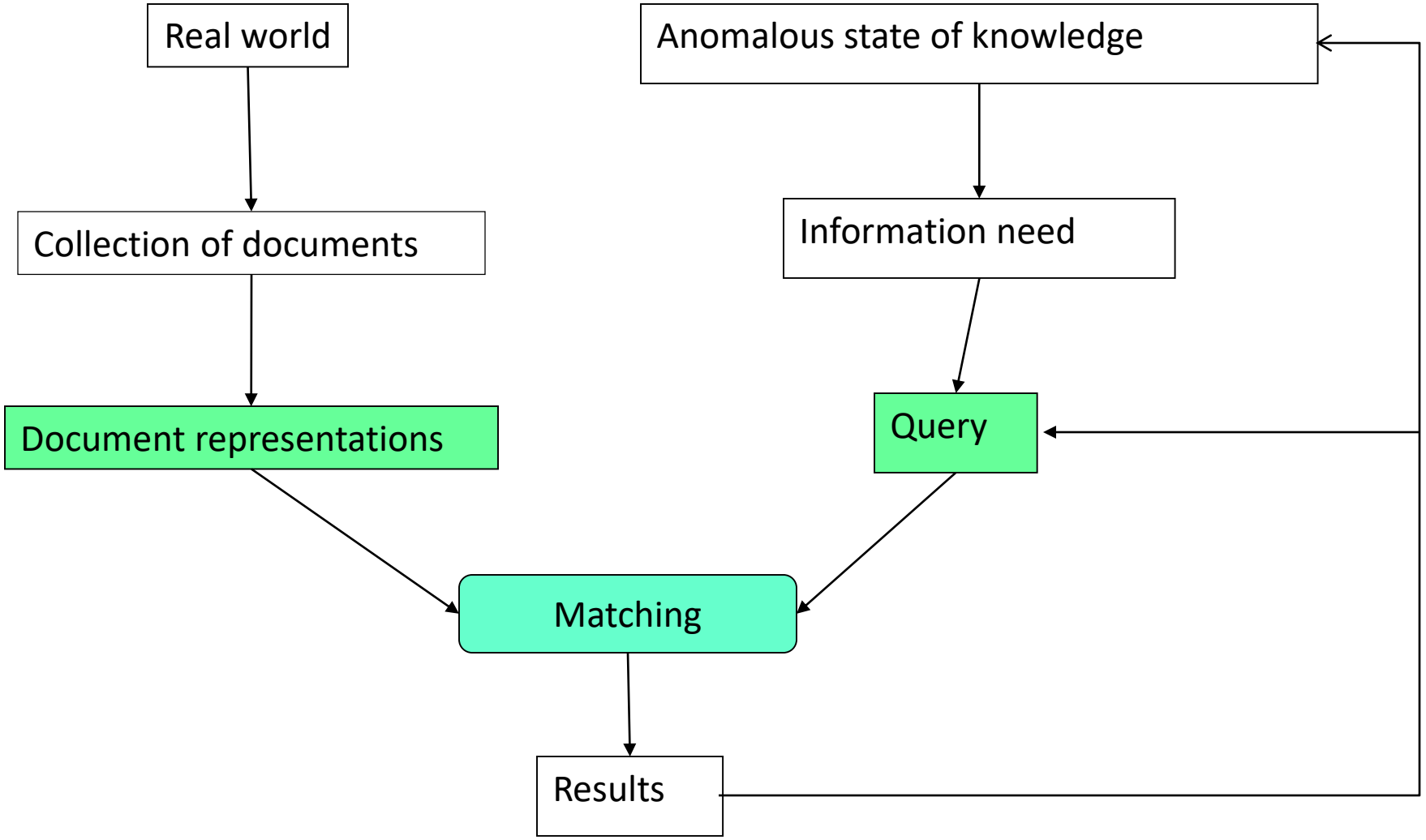
Information Retrieval System

- 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



The formalized IR process



Stopwords / Stoplist

- Function words do not bear useful information for IR
  of, in, about, with, I, although, …
- Stoplist: contain Stopwords, not to be used as index
  - Prepositions
  - Articles
  - Pronouns
  - Some adverbs and adjectives
  - Some frequent words (e.g. document)

- The removal of stopwords usually improves IR effectiveness
- A few “standard” stoplists are commonly used.

Stemming

- Reason:
  - Different word forms may bear similar meaning (e.g. search, searching): create a “standard” representation for them
- Stemming:
  - Removing some endings of word

computer
compute
computes
computing
computed
computation

comput

Link Analysis in Cloud Setup

- The web is not just a collection of documents – its hyperlinks are important!
- A link from page *A* to page *B* may indicate:
  - *A* is related to *B*, or
  - *A* is recommending, citing, voting for or endorsing *B*
- Links are either
  - referential – *click here and get back home*, or
  - Informational – *click here to get more detail*
- Links effect the ranking of web pages and thus have commercial value.

- See More on Chapter 7 for Link Analysis

Thank you