AN INTRODUCTION TO INFORMATION RETRIEVAL

Dr Mohamed Abdeldaiem Abdelhadi

ITIS401

2021

Overview

Intro to IR

Information vs knowledge

IR and search engines

The IR process

What is information retrieval

Gathering information from a source(s) based on an information need usually from a query

- *Major assumption - that the information need can be specified*
- *Broad definition of information*
- *Most methods are automated - scaling*

Sources of information

- Other people
- Archived information (libraries, maps, etc.)
- Radio, TV, etc.
- Web (search engines)
- Natural settings

Information retrieval is more than just web search

information retrieval

Information retrieval (IR) is the activity or process of obtaining information resources relevant to an information need from a collection of information resources.

Data mining is the process that attempts to discover patterns in large data sets.

Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents

Data, information, knowledge

Data - Facts, observations, or perceptions.

Information - Subset of data, only including those data that possess context, relevance, and purpose.

Knowledge - **A more simplistic view** considers knowledge as being at the highest level in a hierarchy with data (at the lowest level) and information (at the middle level).

- **Data** refers to bare facts void of context.
  - *–A telephone number.*
- **Information** is data in context.
  - *–A phone book.*
- **Knowledge** is information that facilitates action.
  - *–Recognizing that a phone number belongs to a good client, who needs to be called once per week to get his orders.*

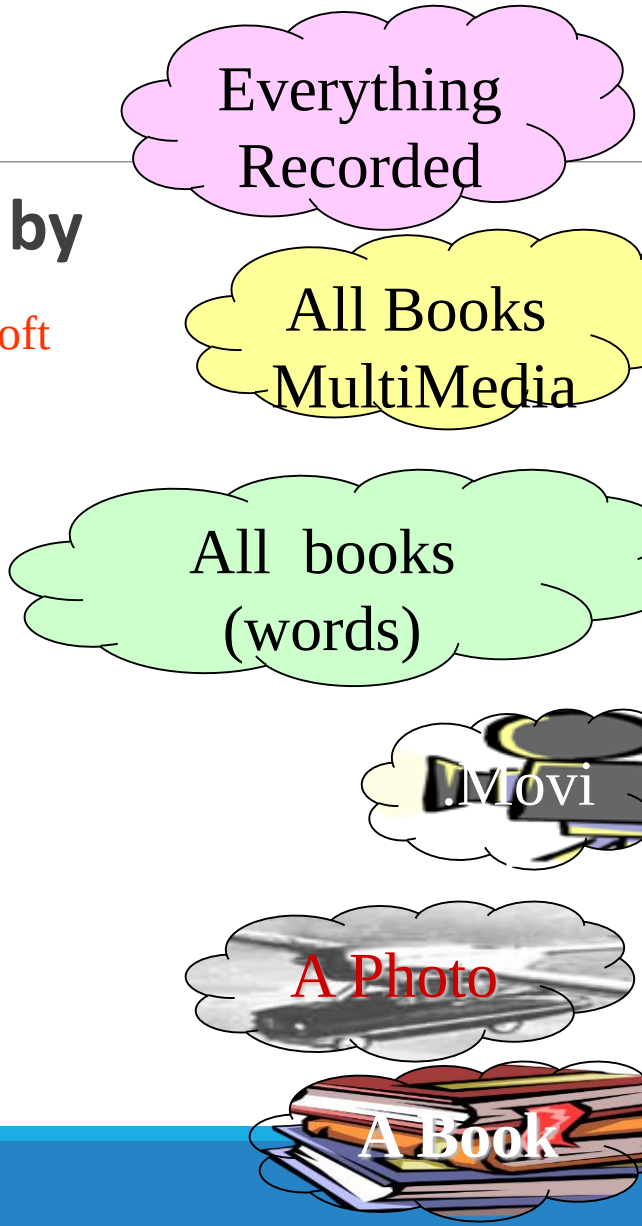
How much information is there?

Soon most everything will be recorded and indexed

Most bytes will never be seen by humans.

Gray - Microsoft

Data summarization, trend detection, anomaly detection, etc. are key technologies



See Mike Lesk:
How much information is there:
http://www.lesk.com/mlesk/ksg97/ksg.html

See Lyman & Varian:

How much information

http://www.sims.berkeley.edu/research/projects/how-much-info/

24 Yecto, 21 zepto, 18 atto, 15 femto, 12 pico, 9 nano, 6 micro, 3 milli

Ideal Information Retrieval

The answer should be:

- what is actually needed (**relevant**)
  - IR is very concerned with relevance
- available when you want it
- available where you want it
- how you want it
  - tailored to the user (personalization)
- your information needs anticipated

What is relevance?

An answer(s) that fits your need.

Definitions of **relevance** on the Web:

- A subjective measure of how well a document satisfies the user's information need. Ideally, your search tool should retrieve all of the documents relevant to your search. However, this is subjective and difficult to quantify.
  www.virtechseo.com/seoglossary.htm
- A measure of how closely a database entry matches a search request. Most search tools on the Web return results based on relevance. The specific algorithm for computing relevance varies from one service to another, but it's often based on the number of times terms in the search expression appear in the document and whether they appear in the appropriate fields.
  www.webliminal.com/search/glossary.htm
- refers to how closely the search engine results appear to match your query. Often it is based upon the number of times your search terms appear in a record. Most search engines attempt to sort and rank your hits by relevance.
  www.lib.ku.edu/research/terms.shtml
- the degree to which a source addresses a research topic (some relevant sources may be more broad or more narrow than the specific research topic.)
  ucblibraries.colorado.edu/about/glossary.htm
- ranking of hits/items/results retrieved from a search. Relevance is a measure of how closely search results match the search request.Search engines vary in the way they determine relevance. In some a document is considered more relevant if the words appear in certain fields, perhaps the title or the summary field. In others relevance is establishedby a percentage determined simply by how many times the keywords appears in the document divided by the total number of words on the page. ...
  mciunix.mciu.k12.pa.us/~spjvweb/glossary.html

How is IR accomplished

Ask someone

Search

- Search for someone to ask
- Search for needed information - library
- Use a search engine

Process of IR - queries or questions

Information to be retrieved

Tacit vs explicit information
- Tacit: in someone's mind
- Explicit: written down

Permanent vs Impermanent information
- Conversation, events
- Documents (in a general sense)
  - Text, tweets
  - Video
  - Files
  - Pictures
  - Data

Both

Assumption: ***it exists!***

What doesn't exist on the web?

The information acquisition process

Know what you want, where it is and go get it

Ask questions to information sources as needed (queries) - manifestation of *SEARCH* - and let them suggest (rank) answers

Have information sent to you on a regular basis based on some predetermined information need or source preference

Push/pull models (RSS)

What is search?

Search vs Information retrieval

Differences

Many definitions of search

- IR (information retrieval)
- CS (computer science)
- Convention

What is SEARCH?

DEFINITIONS FROM THE WEB

the activity of looking thoroughly in order to find something or someone

an investigation seeking answers; "a thorough search of the ledgers revealed nothing"; "the outcome justified the search"

an operation that determines whether one or more of a set of items has a specified property; "they wrote a program to do a table lookup"

the examination of alternative hypotheses; "his search for a move that would avoid checkmate was unsuccessful"

try to locate or discover, or try to establish the existence of; "The police are searching for clues"; "They are searching for the missing man in the entire county"

To request the electronic retrieval of documents based on the presence of specific terms and within other restrictions established (e.g., subject, date, journal, etc.). Search results list The list of documents retrieved as a result of a search request submitted. Settings The record of the personal details related to an individual user, containing information such as, name, address, e-mail, and display preferences (if available), etc. Settings are used to set up a personal profile for the user, and are available only on systems that have user/password authentication.

Intelligently seeking answers to a known or unknown question, often as part of solving a larger problem (AI, planning, strategy, etc.)

What IR is usually not about

Not about structured data (databases)

- Why?
- Grow of structured data?

Retrieval from databases is usually not considered

- Database querying assumes that the data is in a standardized format
- Transforming all information, news articles, web sites into a database format is difficult for large data collections

INTEGRATED IR and database search

- Ex: Craigslist

What is different about IR from other areas, say parts of Computer Science

- Many problems have a right answer
  - How much money did you make last year?
- IR problems usually don′t
  - Find all documents relevant to ″hippos in a zoo″
- Answers are usually “good enough”
- IR defines “good enough”

What an IR system should do

Store/archive information

Provide access to that information

Answer queries with relevant information

Stay current

Future list

- Understand the user's queries
- Understand the user's need
- Acts as an assistant

What is relevance?

In IR relevance is everything

- Defining “good enough”

Relevant information is that suited to your information need.

Dependent on

- User
- Space/time
- Group
- **Context**

Examples?

How good is the IR system

Measures of performance based on what the system returns:

- Relevance
- Coverage
- Recency
- Functionality (e.g. query syntax)
- Speed
- Availability
- Usability
- Time/ability to satisfy user requests

How IR systems work

Algorithms implemented in software

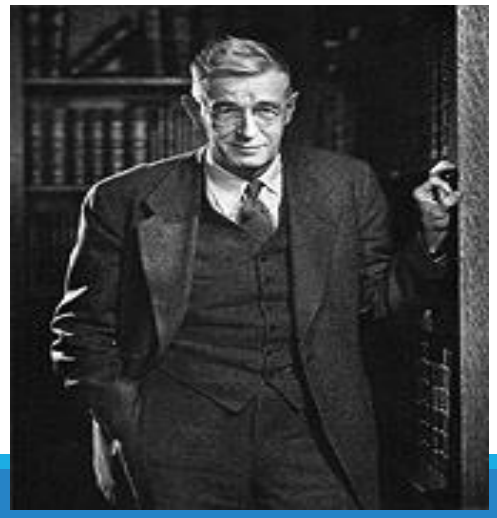
- Gathering of information
- Storage of information
- Processing and indexing
- Interaction
- Evaluation

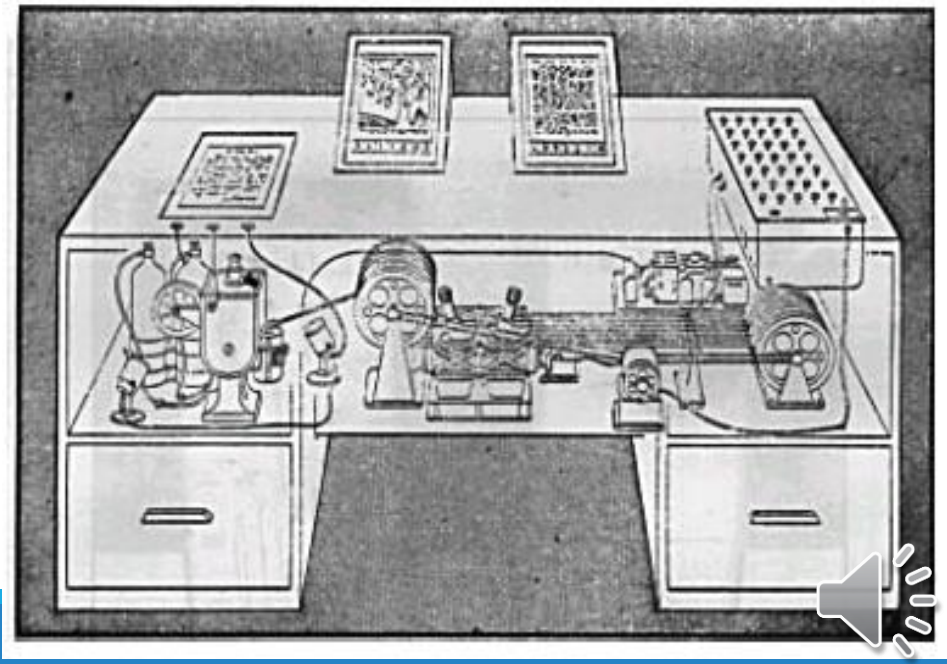
Early ideas of IR-search

Vannevar Bush - Memex - 1945

"A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory."

Bush seems to understand that computers won't just store information as a product; they will transform the process people follow to produce and use information.





Some IR History

- Roots in the scientific "Information Explosion" following WWII
- Interest in computer-based IR from mid 1950's
  - **Key word indexing H.P. Luhn at IBM (1958)**
  - **Probabilistic models at Rand (Maron & Kuhns) (1960)**
  - **Boolean system development at Lockheed ('60s)**
  - **Vector Space Model (Salton at Cornell 1965)**
  - **Statistical Weighting methods and theoretical advances ('70s)**
  - **Refinements and Advances in application ('80s)**
  - **User Interfaces, Large-scale testing and application ('90s)**
- Then came the web and search engines and everything changed
- More History

IR and Search Engines

Search engines are an IR application.

Search engines have become the most popular IR tools.

Why?

Search Engines

Search engines

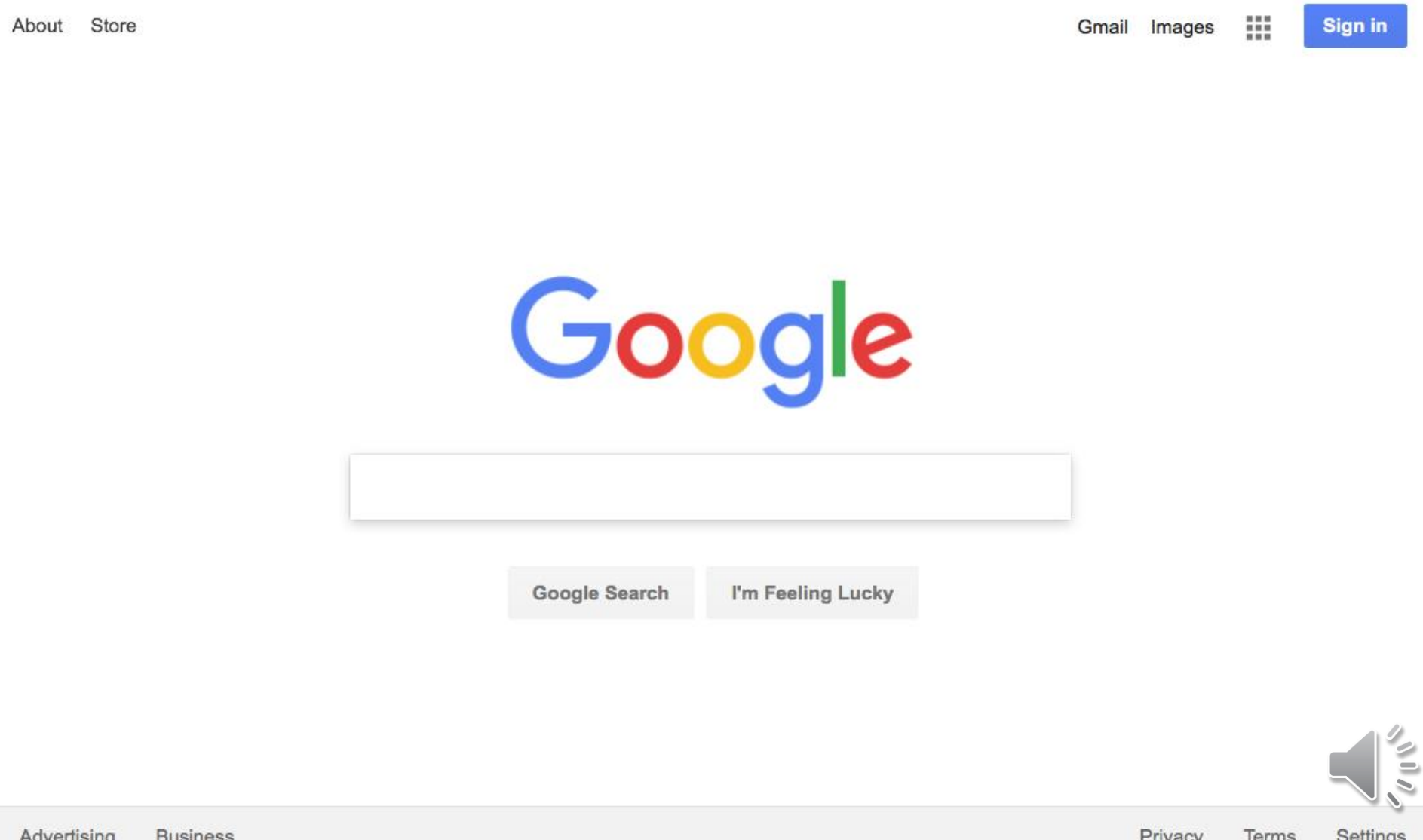
History of search engines

There are many types: Wikipedia list

Many types are not on this list.

- Academic search engines

Existing Popular IR System: Search Engine - Spring 2018



New search engines constantly emerging

Reasonably new search engines

- Bing
- Wolfram Alpha
- Cuil (deceased)
  - July 2008 – Sept 2010
- DuckDuckGo

Which one is best? Ask the search engines!

- best search engine

Impact of search engines

Make the web scale!
- Without search engines, the web probably wouldn't be that important

Unbelievable access to information
- Implications are only just being understood
- Democratization of humankind's knowledge

The online world
- I "googled" him just to see …
- Search is crucial part of many's **everyday** existence and 2nd most popular online activity after email
- Social interactions - blogs

The death of anonymity/privacy
- Nearly everyone is searchable
  - Choicepoint
  - Facebook

Digital divide

What is a Search Engine?

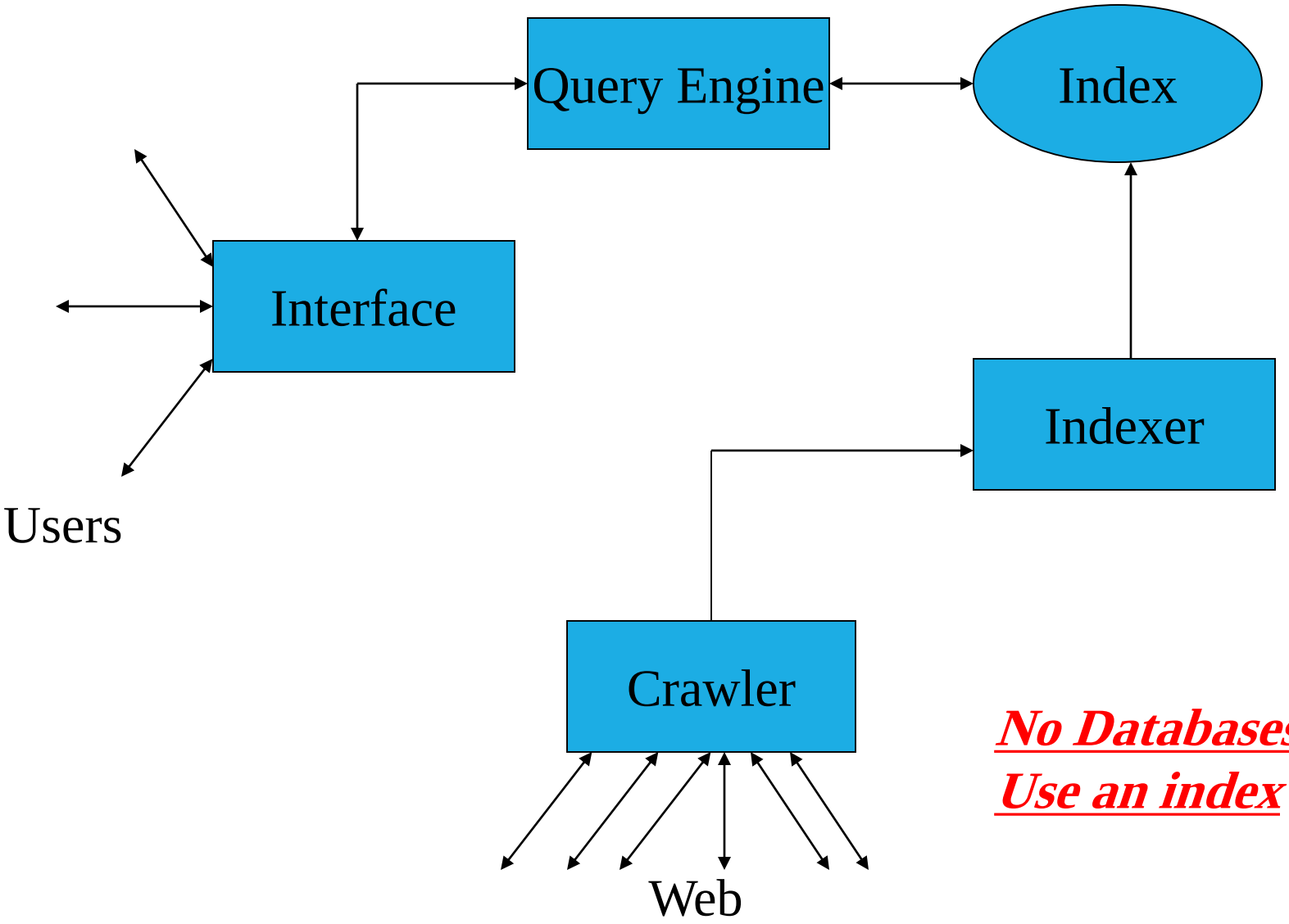
An IR system with an active data harvester to actively collect information,

- Crawler, spider, web bot

Search engines usually collect this information on the web or some part of it.

- Not always – enterprise search

“how search engines work”



No Databases!
Use an index

A Typical Web Search Engine

Google relevance

- Changed everything - 2nd gen search
- 1st gen Search engine relevance - key words
- Google - relevance is popularity
  -who links to you!

Crawlers

Web crawlers (spiders) gather information (files, URLs, etc) from the web.

Primitive IR systems

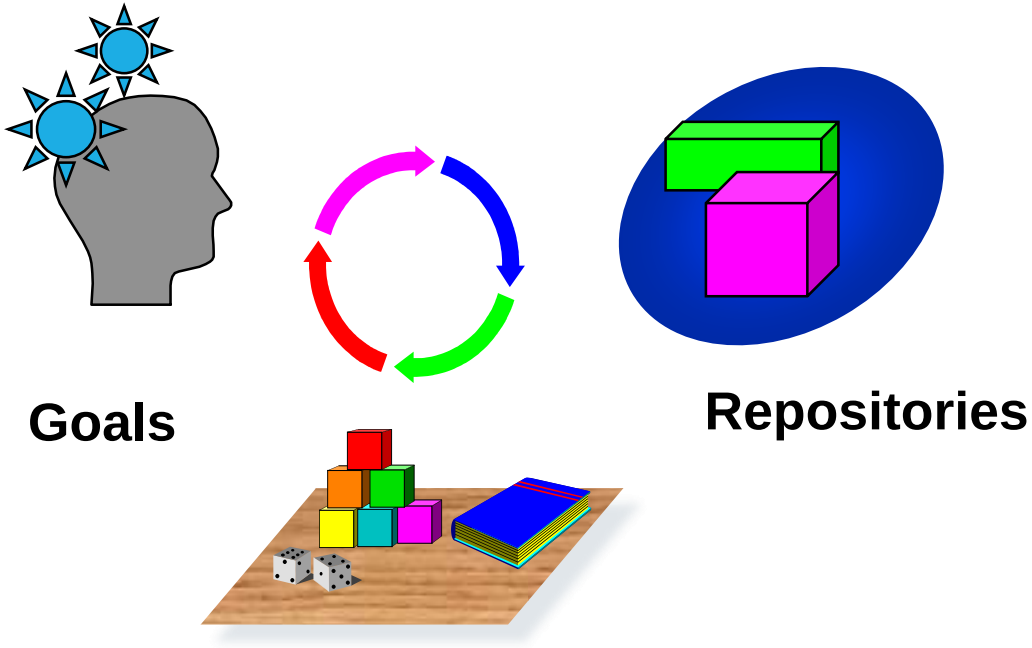
Finding Out About (FOA)
(Reference R. Belew)

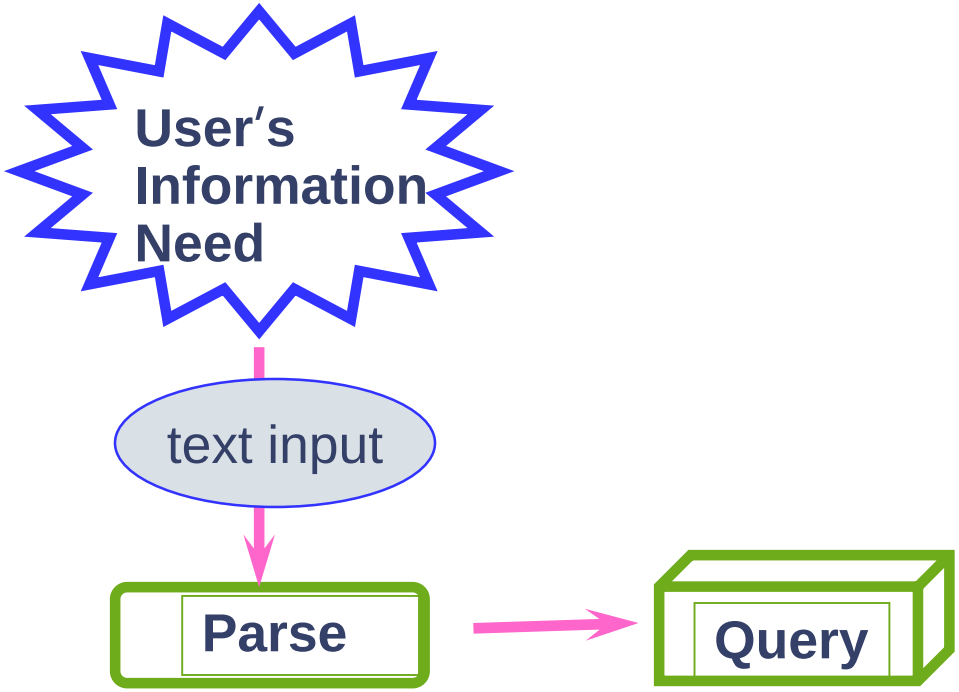
Three phases:

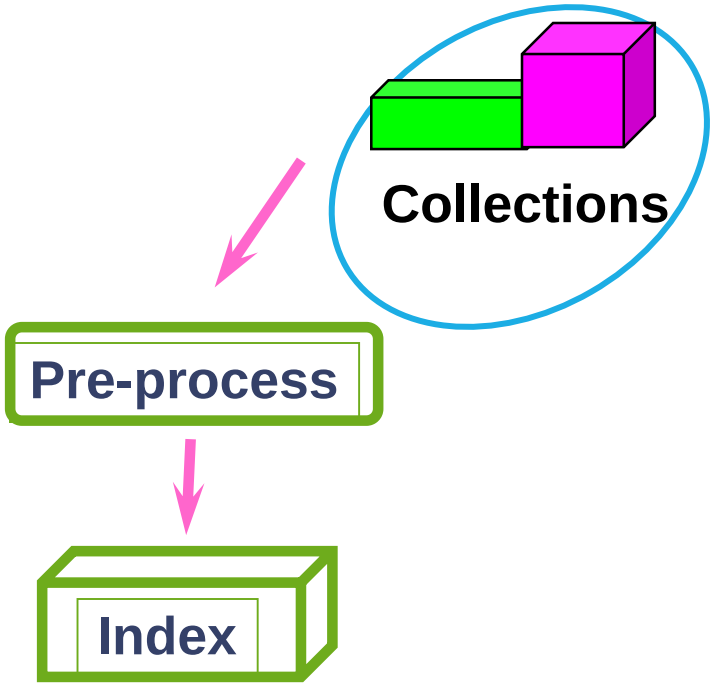
- Asking of a question (the *Information Need*)
- Construction of an answer (IR proper)
- Assessment of the answer (Evaluation)

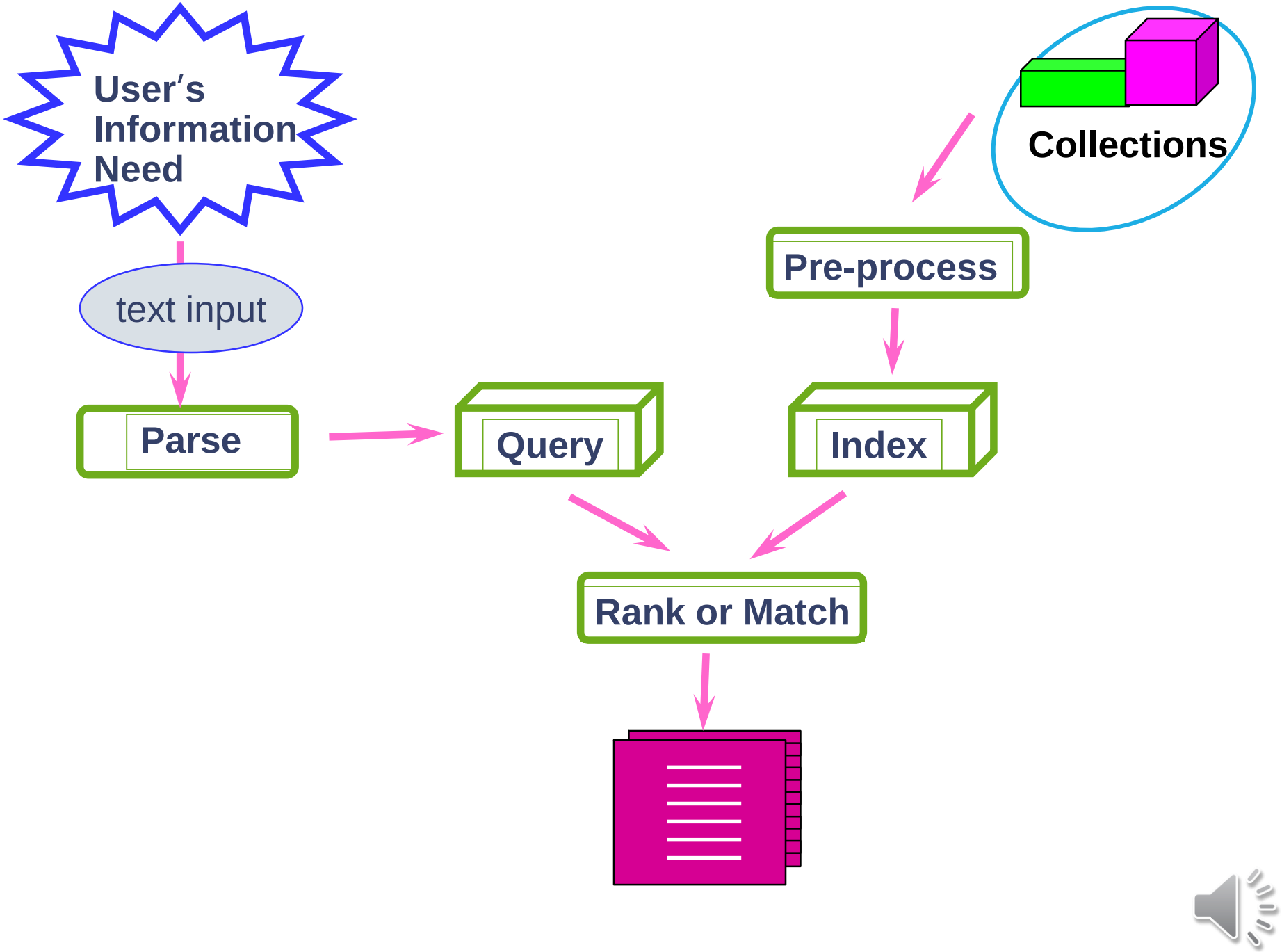
Part of an iterative process

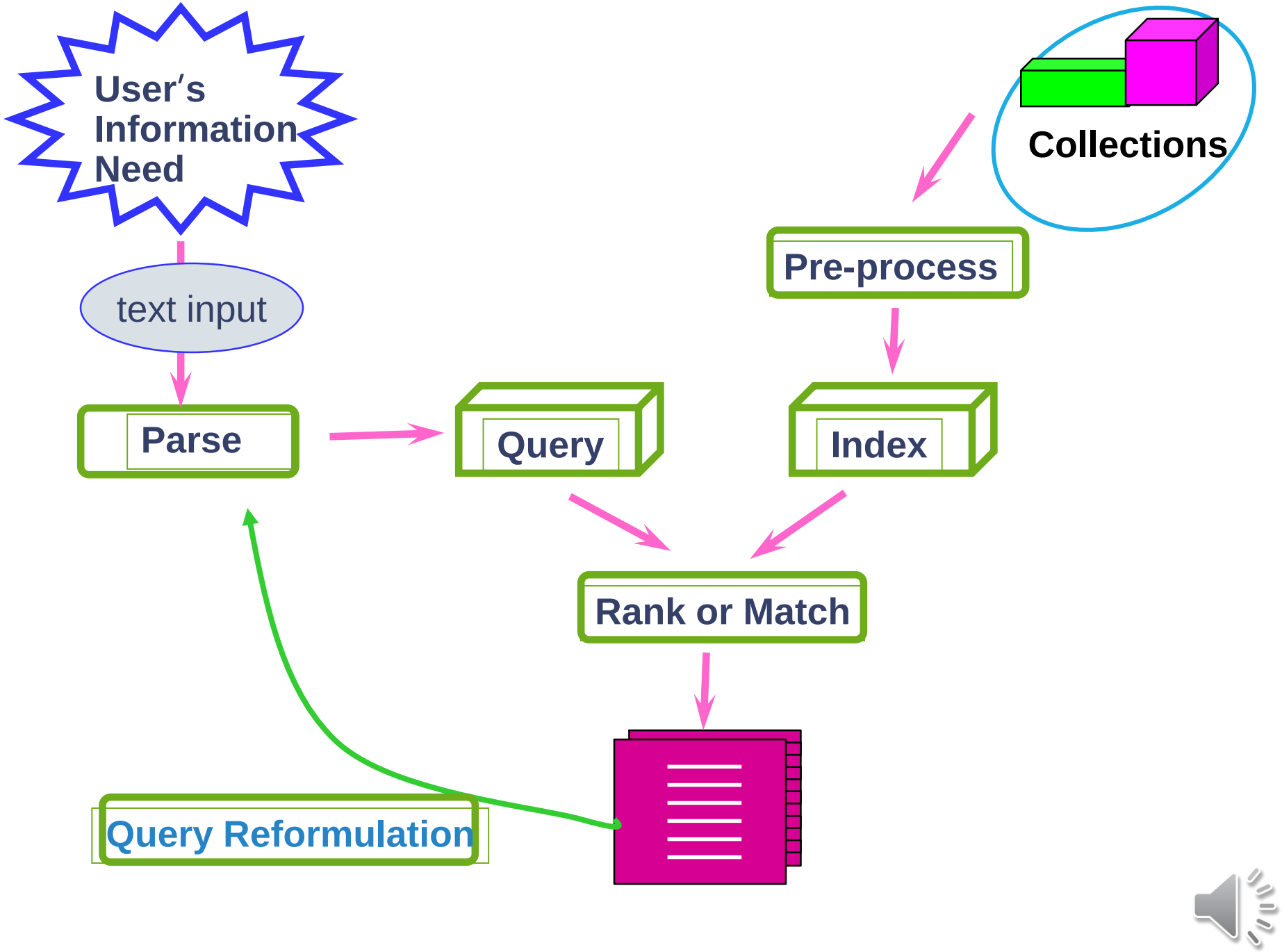
IR is an Iterative Process




User's Information Need
text input
Parse
Query


Collections
Pre-process
Index


User's Information Need
text input
Parse
Query
Collections
Pre-process
Index
Rank or Match


User's Information Need
text input
Parse
Query
Collections
Pre-process
Index
Rank or Match
Query Reformulation

Assessing the Answer to an IR System

How well does it answer the question?

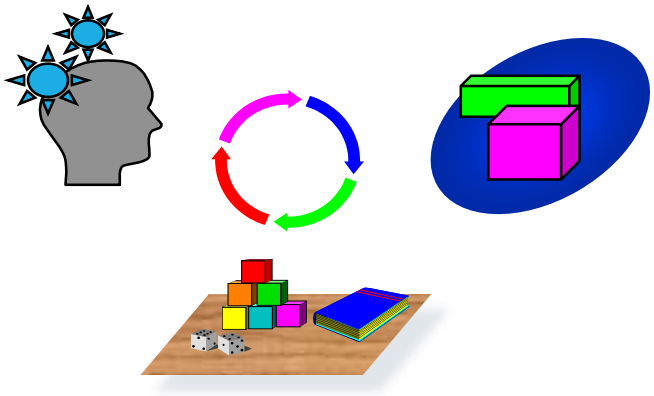
- Complete answer? Partial?
- Background Information?
- Hints for further exploration?

How relevant is it to the user?

Notion of relevance.

What about IBM's Watson?

IR is usually a dialog



- The exchange doesn't end with first answer
- User can recognize elements of a useful answer
- Questions and understanding changes as the process continues.

Information Seeking Behavior

Two parts of the process:

- search and retrieval
- analysis and synthesis of search results

examples?

Information Retrieval

Revised Goal Statement:

Build a system that retrieves documents that users are ***likely*** to find relevant to their queries.

This set of assumptions underlies the field of Information Retrieval.

Measures of performance

How good is that IR system?

BUDLITE SEARCH – never fills you up.

What we covered

IR interacts with users!

Field is relatively old

- Legal search
- Library search

Computers/web changed IR

Little theory – lots of math and data structures

Does not use databases for the most part.

Pioneered foundations of many large scale systems

Research directions in IR

Semantic search, indexing, and retrieval
- Natural language understanding and generation
- Question and answering
- Knowledge graphs

New data to index and search
- Images; video
- Equations, figures, tables, etc

Automate, automate, automate

Data integration

Search and HCI

Privacy, security

Is Information Retrieval?

discovering new knowledge

capturing existing knowledge

sharing knowledge with others

applying knowledge

app for large data

Should we really be studying knowledge retrieval?

Seems to be Google's new goal