THE UNIVERSITY *of* EDINBURGH

Text Technologies for Data Science

INFR11145

Definitions

Instructor:
Youssef Al Hariri

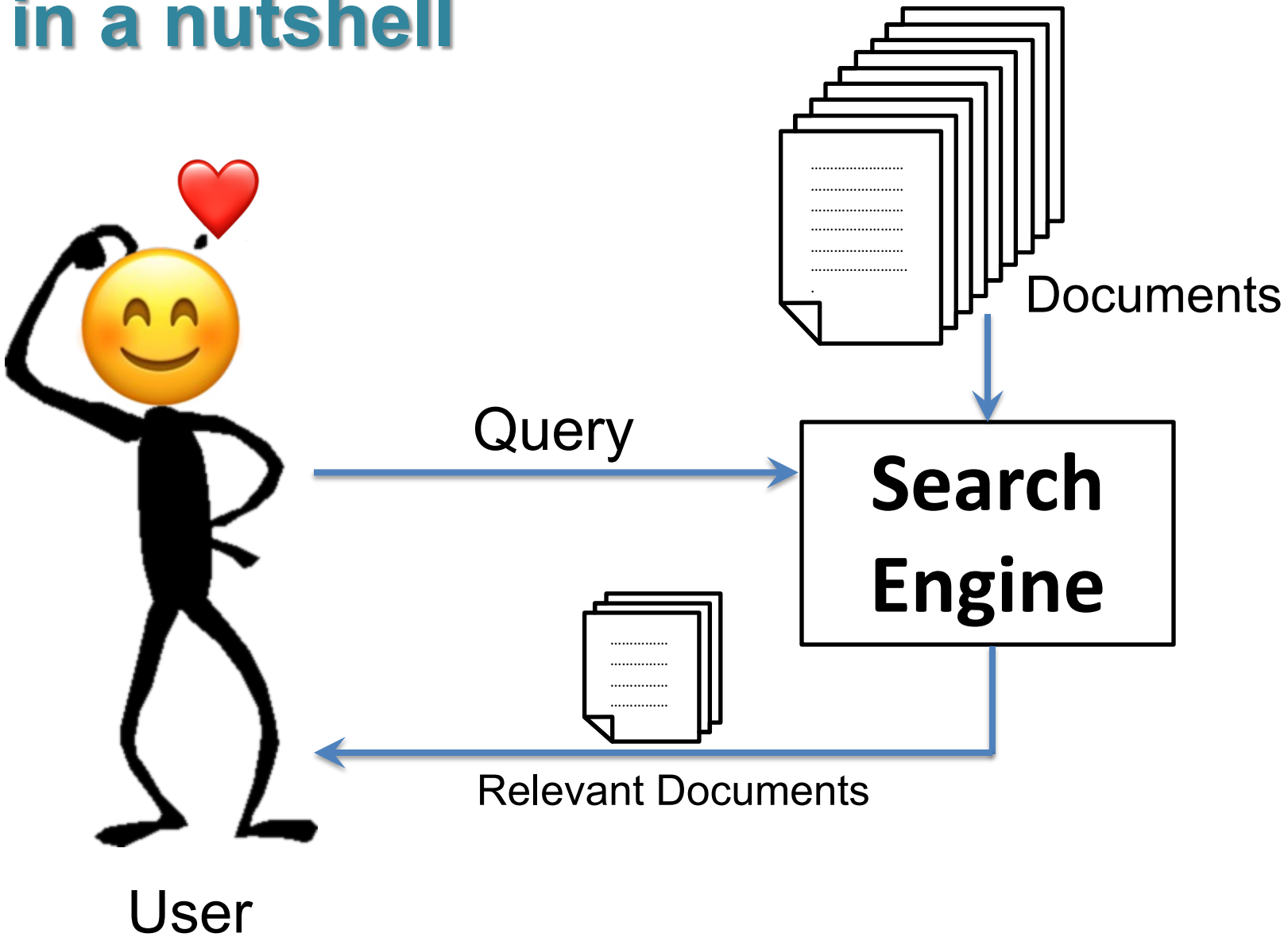
20/09/2023

slides credit to Dr Walid Magdy

Lecture Objectives

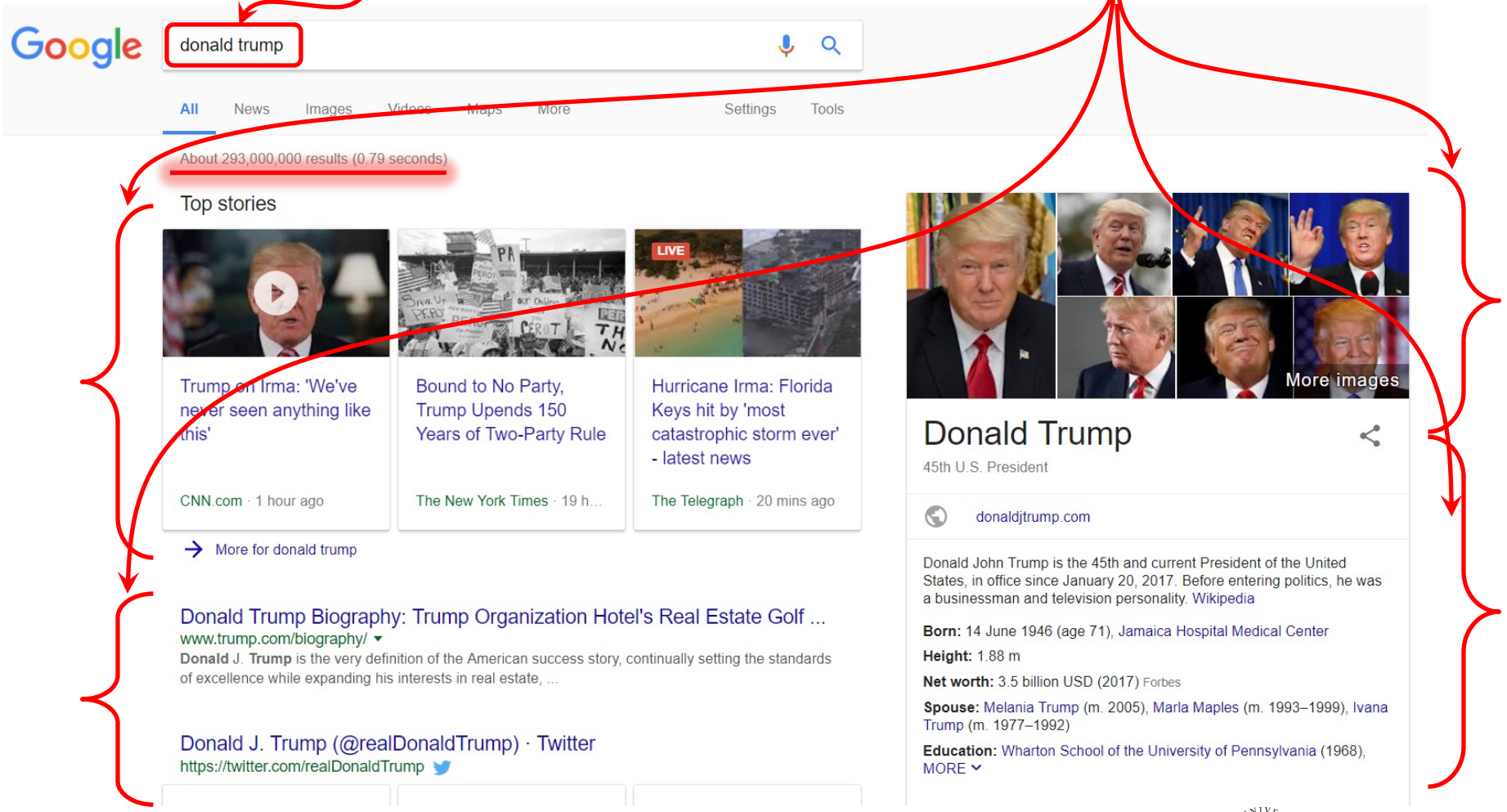
- Learn about main concepts in IR
  - Document
  - Information need
  - Query
  - Index
  - BOW

IR in a nutshell



IR, basic form

- Given Query **Q**, find relevant documents **D**



Two main Issues in IR

About 293,000,000 results (0.79 seconds)

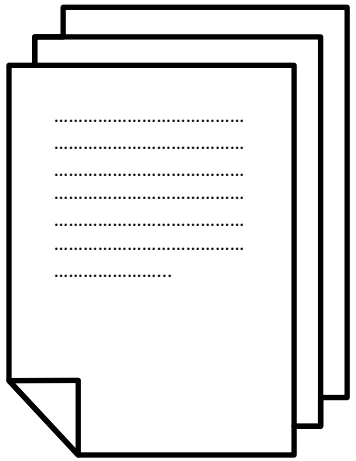
- Effectiveness
  - need to find **relevant** documents
  - needle in a haystack
  - very different from relational DBs (SQL)

- Efficiency
  - need to find them quickly
  - vast quantities of data (100's billions pages)
  - thousands queries per second (Google, 99,000)
  - data constantly changes, need to keep up
  - compared with other NLP areas, IR is **very** fast

IR main components

- Documents
- Queries
- Relevant documents

Documents



- The element to be retrieved
  - Unstructured nature
  - Unique ID
  - *N* documents → Collection

- web-pages, emails, book, page, sentence, tweets
- photos, videos, musical pieces, code
- answers to questions
- product descriptions, advertisements
- may be in a different language
- may not have words at all (e.g. DNA)

Queries

- Free text to express user's information need
- Same information need can be described by multiple queries
  - Latest news on the hurricane in the US
  - North Carolina storm
  - Florence
- Same query can represent multiple information needs
  - Apple
  - Jaguar




Queries – different forms

- Web search → keywords, narrative …
- Image search → keywords, sample image
- QA → question
- Music search → humming a tune
- Filtering/recommendation → user's interest/history
- Scholar search → structured (author, title ..)

- Advanced search

  #wsyn(0.9 **#field** (title, **#phrase** (homer,simpson)) 0.7 **#and** (**#>** (pagerank,3), **#ow3** (homer,simpson)) 0.4 **#passage** (homer, simpson, dan, castellaneta))

Relevance

- At an abstract level, IR is about:
  - does item *D* ***match*** item *Q*? …or…
  - is item *D* ***relevant*** to item *Q*?
- Relevance a tricky notion
  - will the user like it / click on it?
  - will it help the user achieve a task? (satisfy information need)
  - is it novel (not redundant)?
- *Relevance = what is the topic about?*
  - i.e. *D*, *Q* share similar “meaning”
  - about the same topic / subject / issue

What is the challenge in relevance?

- No clear semantics, contrast:
  - *"William Shakespeare"*
  - Author history's? list of plays? a play by him?
- Inherent ambiguity of language:
  - synonymy: "Edinburgh festival" = "The fringe"
  - polysemy: "Apple", "Jaguar"
- Relevance highly subjective
  - Relevance: yes/no
  - Relevance: perfect/excellent/good/fair/bad
- On the web: counter SEOs / spam

Relevant Items are Similar

- Key idea:
  - Use similar vocabulary → similar meaning
  - Similar documents relevant to same queries

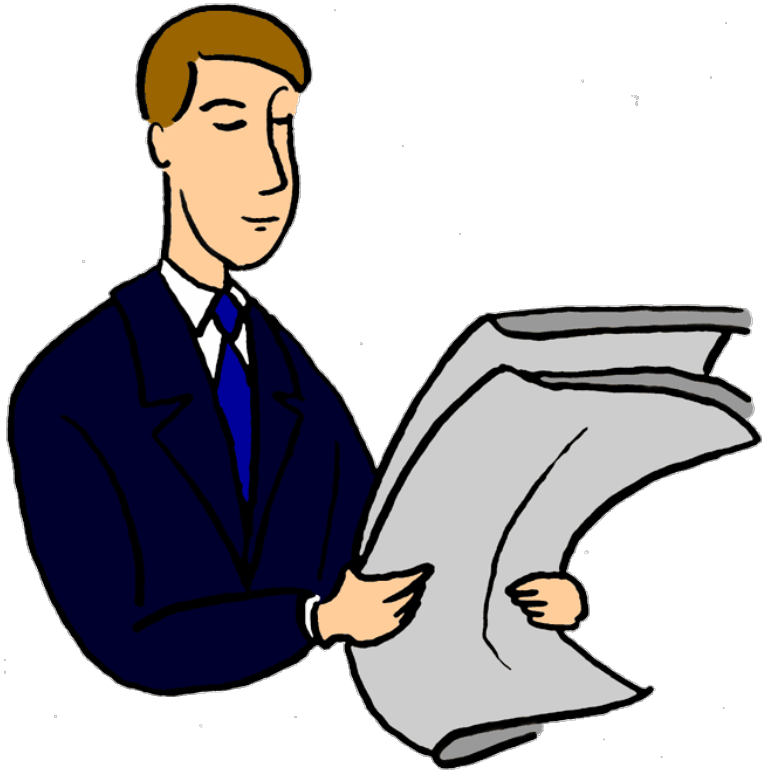
- Similarity
  - String match
  - Word overlap
  - P(D|Q) → retrieval model

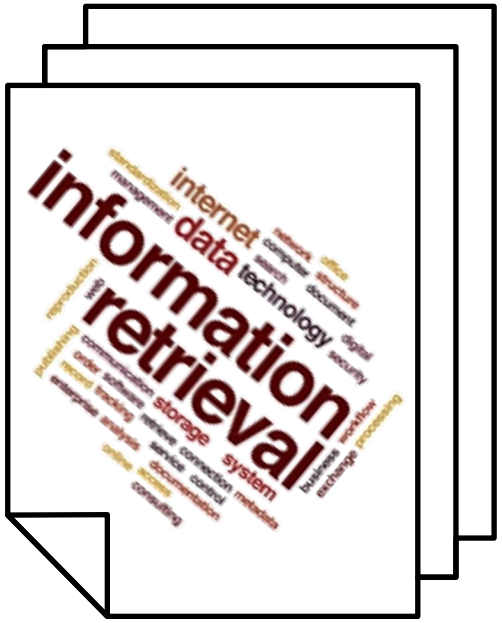
IR vs. DB

	Databases	IR
What we're retrieving	Structured data. Clear semantics based on a formal model.	Mostly unstructured. Free text with some metadata.
Queries we're posing	Formally-defined (relational algebra, SQL). Unambiguous.	Free text ("natural language"), Boolean
Results we get	Exact (always "correct")	Imprecise (need to measure relevance)
Interaction with system	One-shot queries.	Interaction is important.

Tamer Elsayed, QU

How IR sees documents?





Bag-of-words trick

- Can you guess what this is about:
  - per is salary hour €25,000 Mbappe's
  - obesity French is of full cause and fat fries
- Re-ordering doesn't destroy the topic
  - individual words – "building blocks"
  - "bag" of words: a "composition" of "meanings"

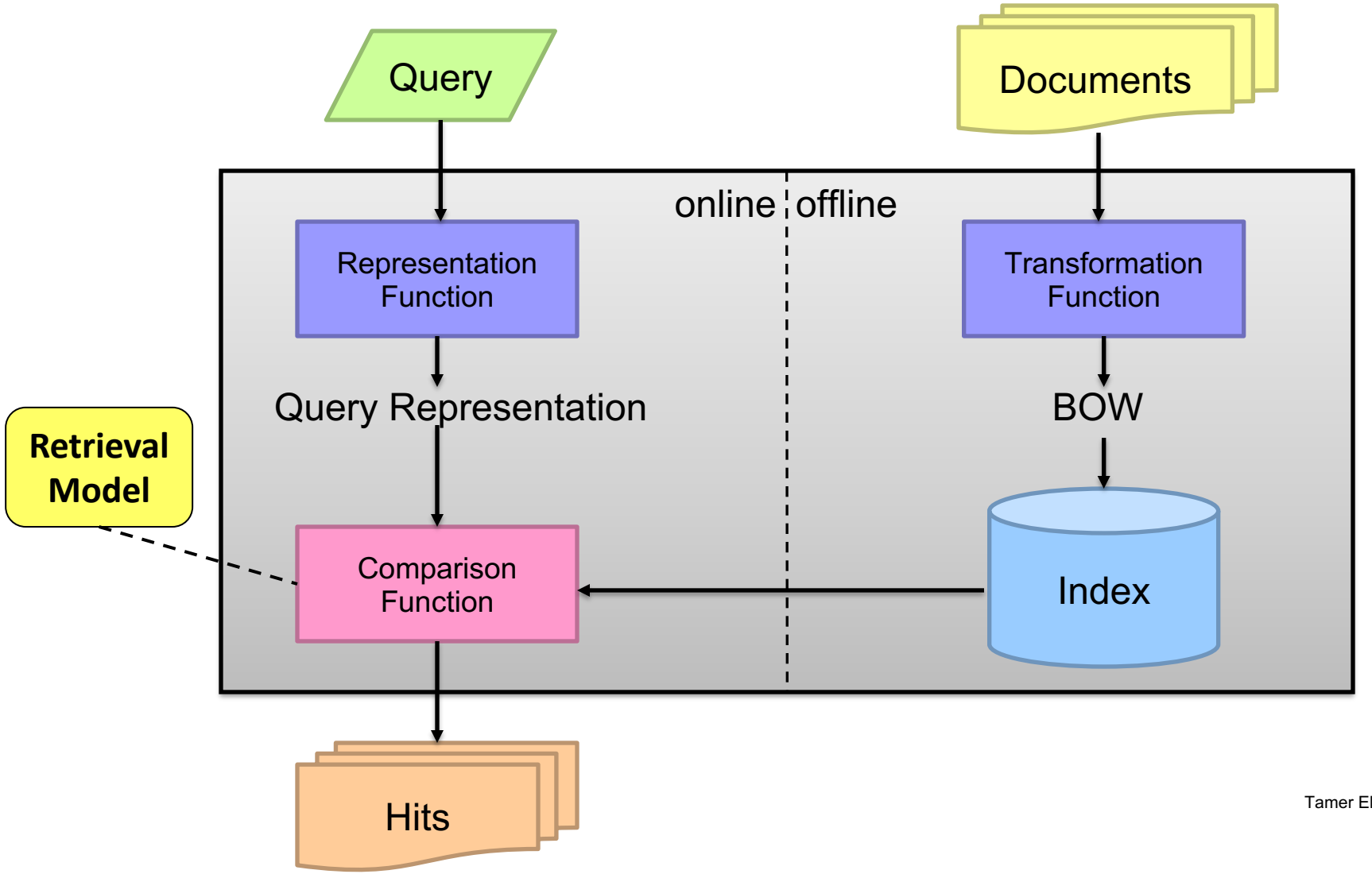
Bag-of-words trick

- Most search engines use BOW
  - treat documents, queries as bags of words
- A “bag” is a set with repetitions
  - match = “degree of overlap” between D,Q
- Retrieval models
  - statistical models (function) that use words as features
  - decide which documents most likely to be relevant
- What should be the top results for Q?
  - BOW makes these models tractable

Bag-of-words: Criticism

- word meaning lost without context
  - True, but BOW doesn't really discard context
- what about negations, etc.?
  - {no, climate change is real} vs. {climate change is no real}
- does not work for all languages
  - No natural "word" unit for Chinese, images, music
  - Solve by "segmentation" or "feature induction"

IR Black Box

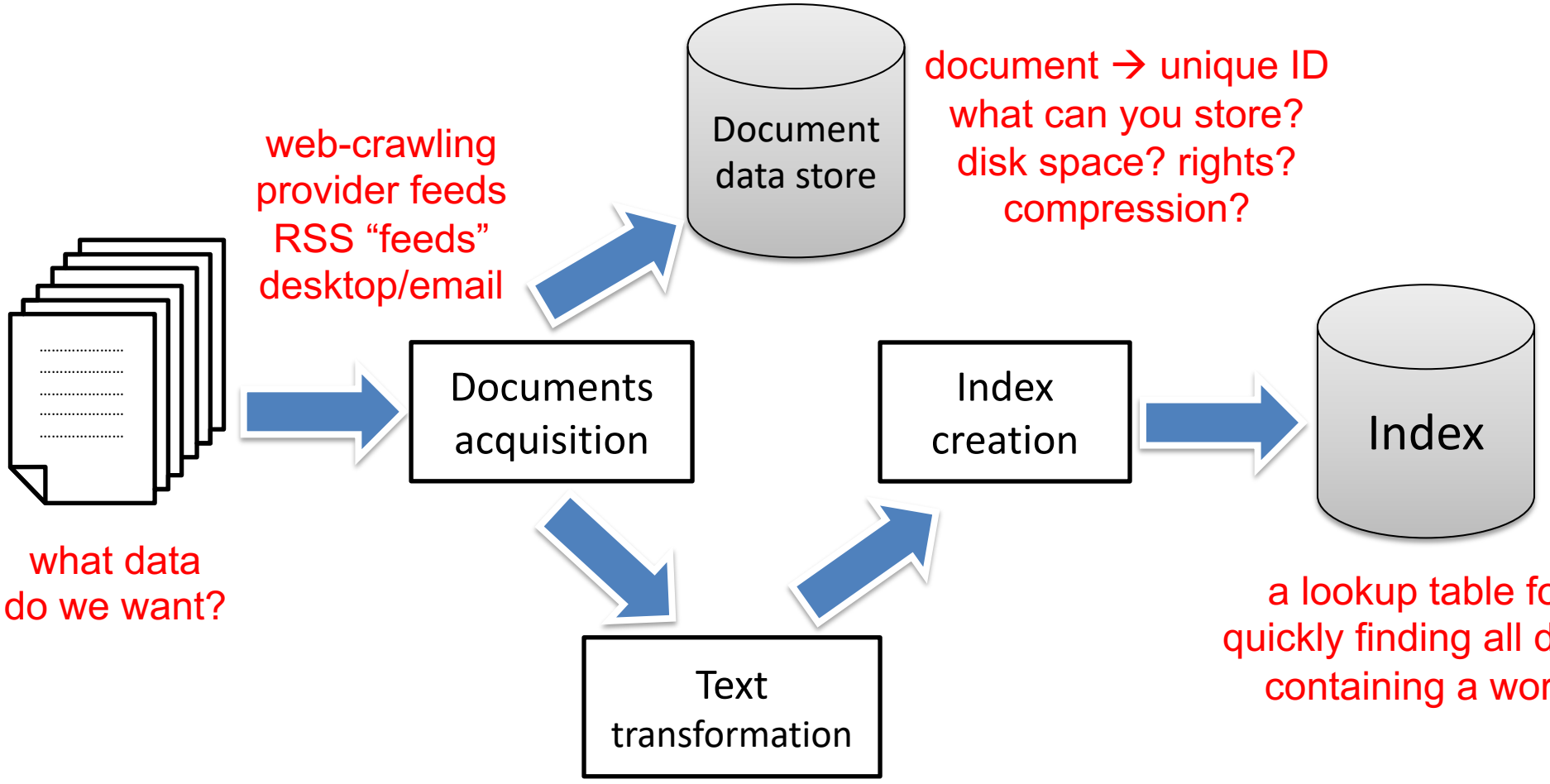


Tamer Elsayed, QU

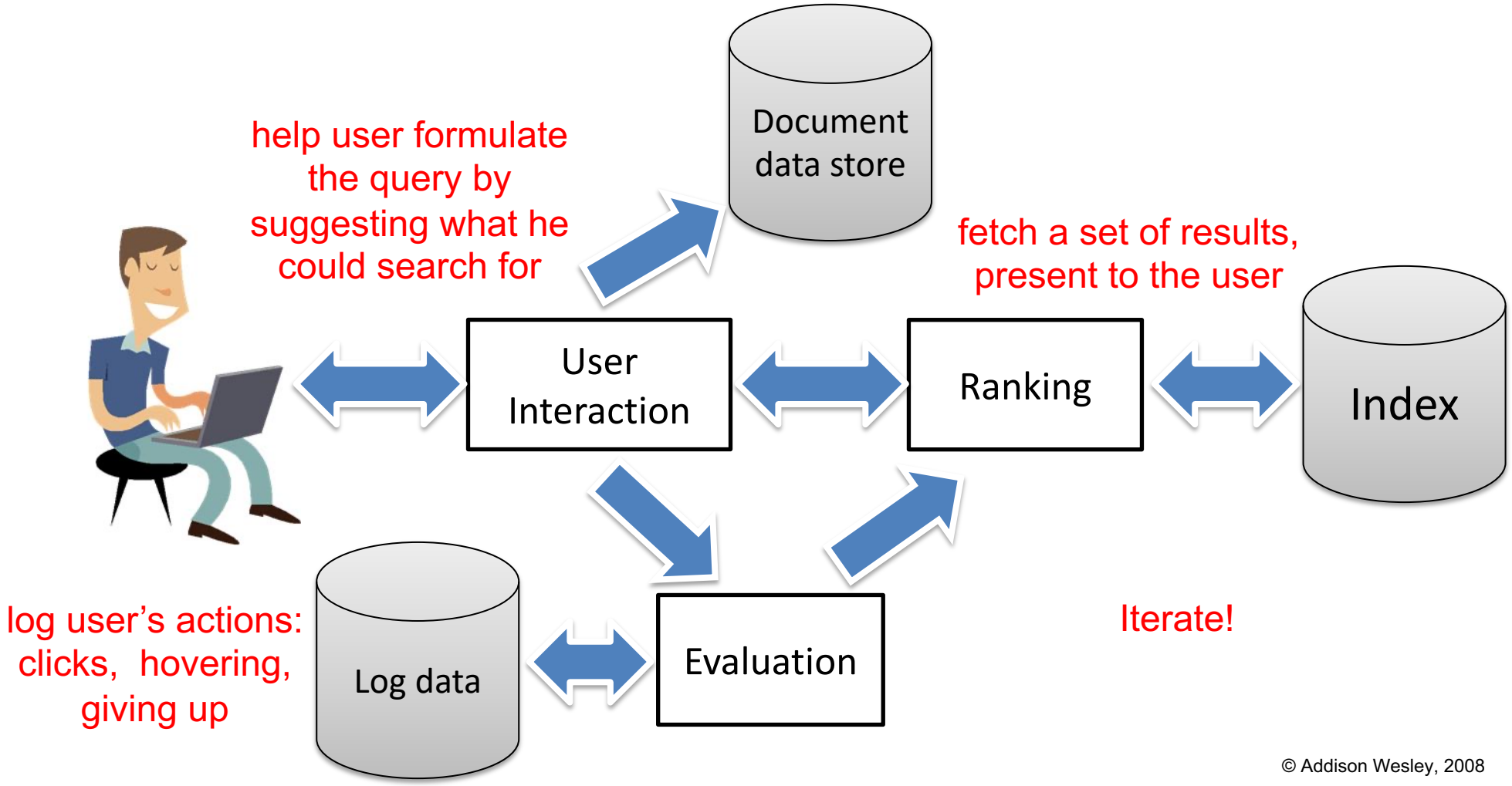
Systems perspective on IR

- Indexing Process: *(offline)*
  → get the data into the system
  - acquire the data from crawling, feeds, etc.
  - store the originals (if needed)
  - transform to BOW and “index”
- Search (retrieval) Process: *(online)*
  → satisfy users’ requests
  - assist user in formulating query
  - retrieve a set of results
  - help user browse / re-formulate
  - log user’s actions, adjust retrieval model

Indexing Process



Search Process



Summary

- Information Retrieval (IR): core technology
  - selling point: IR is very fast, provides context
- Main issues: effectiveness and efficiency
- Documents, queries, relevance
- Bag-of-words trick
- Search system architecture:
  - indexing: get data into the system
  - searching: help users find relevant data

Resources

- Search Engines: Information Retrieval in Practice, chapter 1 & 2

- Lab 0:
  - You have to be confident doing it!
  - If you have trouble finishing it, think twice before committing to the course

Questions

- Next time:
  - Laws of text (Zipf ….)
  - Vector space models
- Skill to learn by next time:
  - Read text file from disk
  - Read word by word
- Videos:
  - The Zipf Mystery, Vsauce
- Tools:
  - Regular expressions: https://www.w3schools.com/python/python_regex.asp