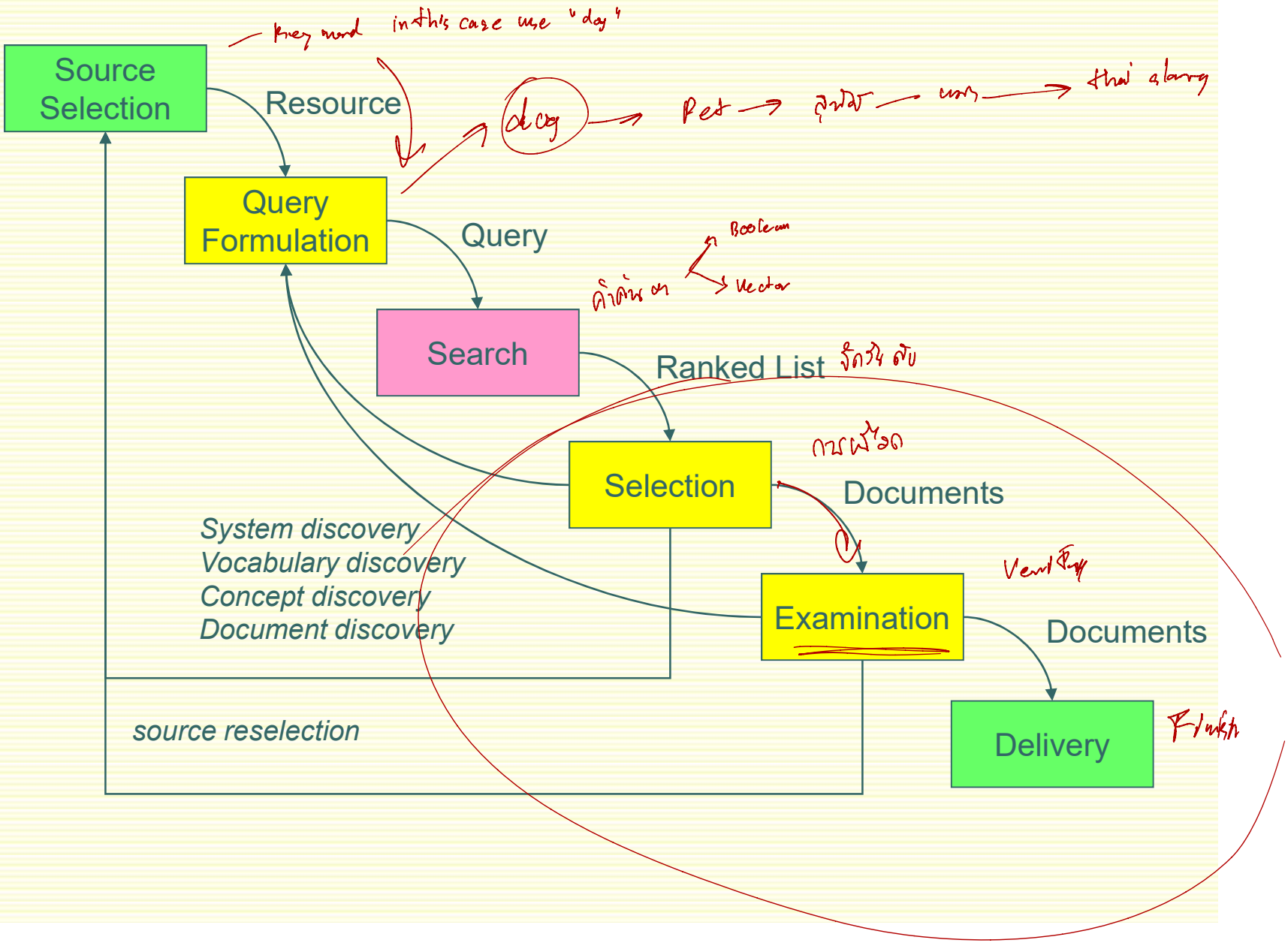
The Information Retrieval Cycle

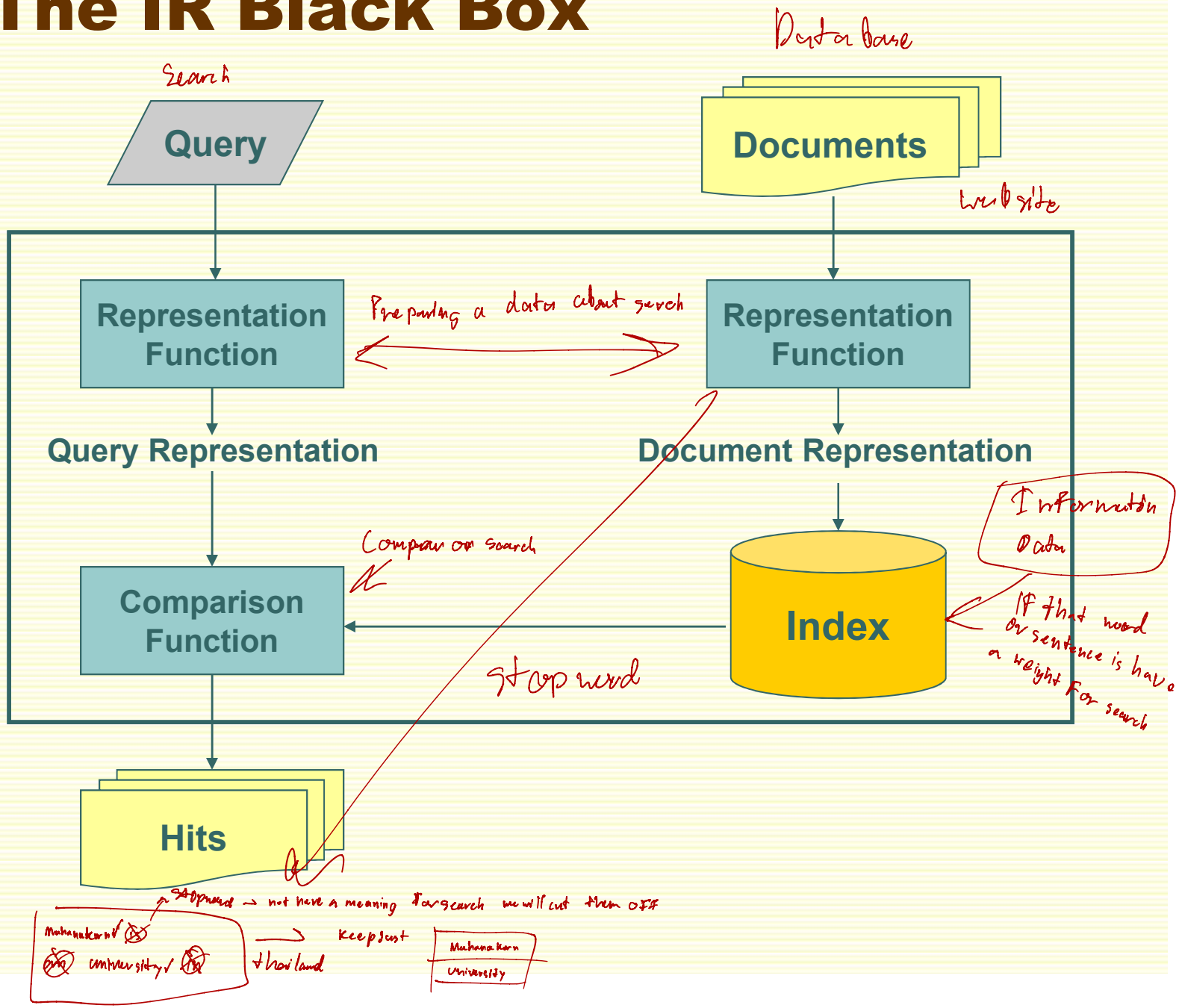


What is a model?

- A model is a construct designed help us understand a complex system
  - A particular way of “looking at things”
- Models inevitably make simplifying assumptions
  - What are the limitations of the model?
- Different types of models:
  - Conceptual models
  - Physical analog models
  - Mathematical models
  - …

model that use

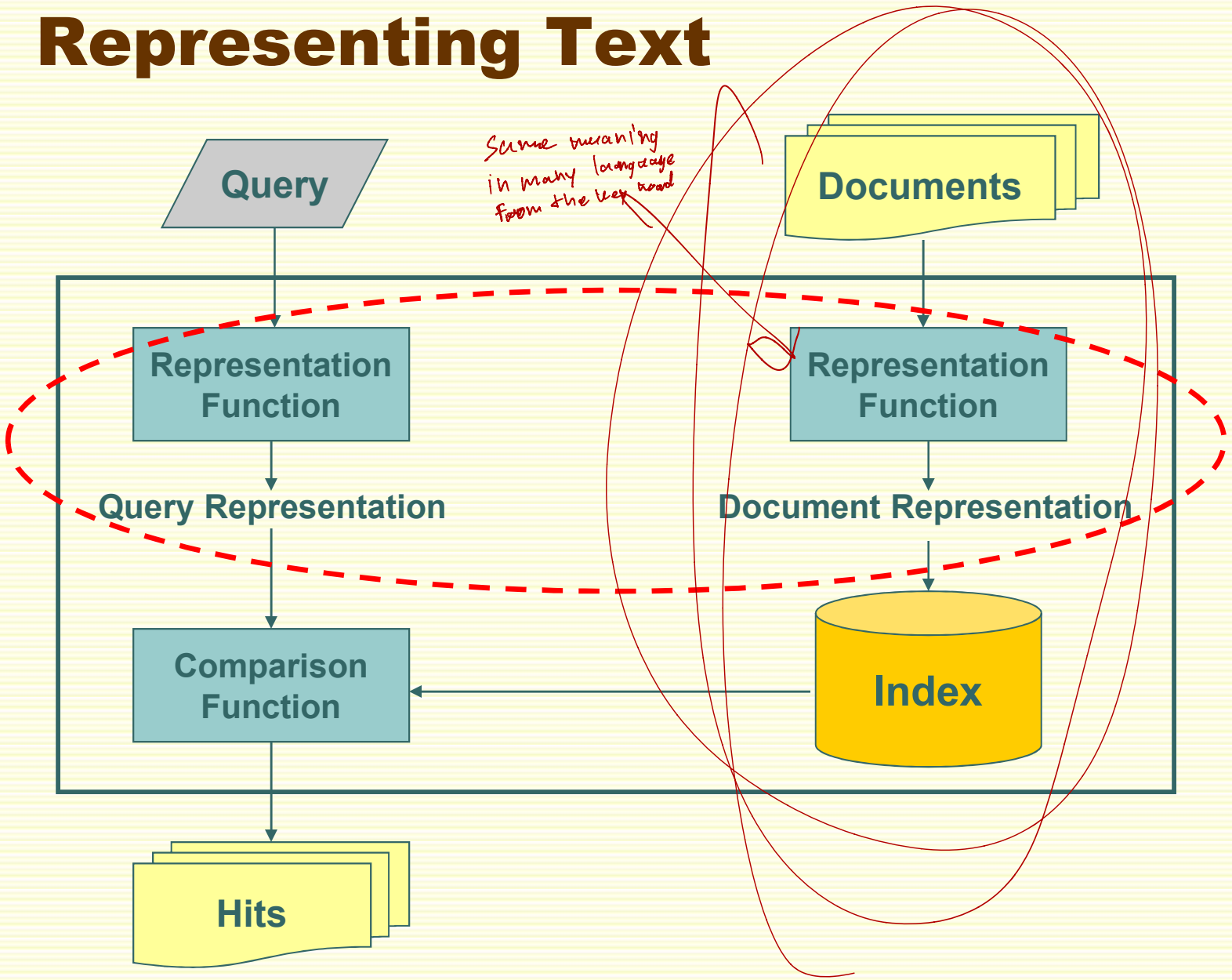
The IR Black Box



Today's Topics

- Boolean model
  - Based on the notion of sets
  - Documents are retrieved *only* if they satisfy Boolean conditions specified in the query
  - Does not impose a ranking on retrieved documents
  - Exact match
- Vector space model
  - Based on geometry, the notion of vectors in high dimensional space
  - Documents are ranked based on their similarity to the query (ranked retrieval)
  - Best/partial match

Representing Text



How do we represent text?

- How do we represent the complexities of language?
  - Keeping in mind that computers don't "understand" documents or queries
- Simple, yet effective approach: "bag of words"
  - Treat all the words in a document as index terms for that document
  - Assign a "weight" to each term based on its "importance"
  - Disregard order, structure, meaning, etc. of the words

What's a "word"? We'll return to this in a few lectures…

Sample Document

McDonald's slims down spuds

Fast-food chain to reduce certain types of fat in its french fries with new cooking oil.

NEW YORK (CNN/Money) - McDonald's Corp. is cutting the amount of "bad" fat in its french fries nearly in half, the fast-food chain said Tuesday as it moves to make all its fried menu items healthier.

But does that mean the popular shoestring fries won't taste the same? The company says no. "It's a win-win for our customers because they are getting the same great french-fry taste along with an even healthier nutrition profile," said Mike Roberts, president of McDonald's USA.

But others are not so sure. McDonald's will not specifically discuss the kind of oil it plans to use, but at least one nutrition expert says playing with the formula could mean a different taste.

Shares of Oak Brook, Ill.-based McDonald's (MCD: down $0.54 to $23.22, Research, Estimates) were lower Tuesday afternoon. It was unclear Tuesday whether competitors Burger King and Wendy's International (WEN: down $0.80 to $34.91, Research, Estimates) would follow suit. Neither company could immediately be reached for comment.

…

16 × said

14 × McDonalds

12 × fat

11 × fries

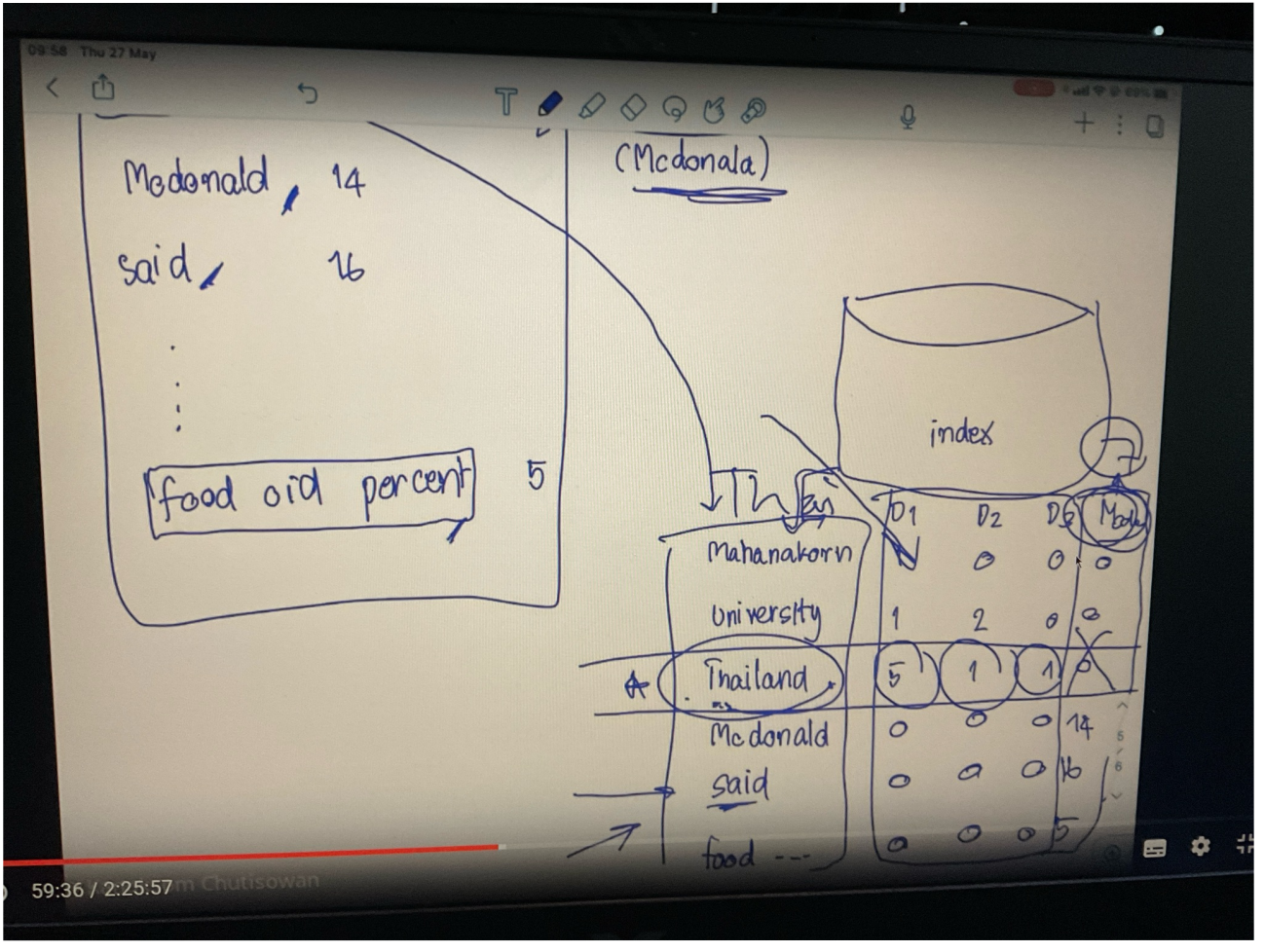
8 × new

6 × company french nutrition

5 × food oil percent reduce taste Tuesday

…

“Bag of Words”


Mcdonald 14
said 16
food oid percent 5
(Mcdonala)
index
Mahanakorn
University
Thailand
Mcdonald
said
food ---
D1 D2 D6 Mcdo
1 2
5 1 1
14
16
5
59:36 / 2:25:57

Vector Representation

- "Bags of words" can be represented as vectors
  - Why? Computational efficiency, ease of manipulation
  - Geometric metaphor: "arrows"
- A vector is a set of values recorded in any consistent order

"The quick brown fox jumped over the lazy dog's back"

→ [1 1 1 1 1 1 1 1 2]

1st position corresponds to "back"
2nd position corresponds to "brown"
3rd position corresponds to "dog"
4th position corresponds to "fox"
5th position corresponds to "jump"
6th position corresponds to "lazy"
7th position corresponds to "over"
8th position corresponds to "quick"
9th position corresponds to "the"

Representing Documents

Document 1

The quick brown fox jumped over the lazy dog's back.

Document 2

Now is the time for all good men to come to the aid of their party.

Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

Stopword List

for
is
of
the
to

Boolean View of a Collection

Term	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8
aid	0	0	0	1	0	0	0	1
all	0	1	0	1	0	1	0	0
back	1	0	1	0	0	0	1	0
brown	1	0	1	0	1	0	1	0
come	0	1	0	1	0	1	0	1
dog	0	0	1	0	1	0	0	0
fox	0	0	1	0	1	0	1	0
good	0	1	0	1	0	1	0	1
jump	0	0	1	0	0	0	0	0
lazy	1	0	1	0	1	0	1	0
men	0	1	0	1	0	0	0	1
now	0	1	0	0	0	1	0	1
over	1	0	1	0	1	0	1	1
party	0	0	0	0	0	1	0	1
quick	1	0	1	0	0	0	0	0
their	1	0	0	0	1	0	1	0
time	0	1	0	1	0	1	0	0

Each column represents the view of a particular document: What terms are contained in this document?

Each row represents the view of a particular term: What documents contain this term?

To execute a query, pick out rows corresponding to query terms and then apply logic table of corresponding Boolean operator

Proximity Operators

- More "precise" versions of AND
  - "NEAR n" allows at most *n-1* intervening terms
  - "WITH" requires terms to be adjacent and in order
  - Other extensions: within *n* sentences, within *n* paragraphs, etc.
- Relatively easy to implement, but less efficient
  - Store position information for each word in the document vectors
  - Perform normal Boolean computations, but treat WITH and NEAR as extra constraints

Proximity Operator Example

Term	Doc 1	Doc 2
aid	0	1 (13)
all	0	1 (6)
back	1 (10)	0
brown	1 (3)	0
come	0	1 (9)
dog	1 (9)	0
fox	1 (4)	0
good	0	1 (7)
jump	1 (5)	0
lazy	1 (8)	0
men	0	1 (8)
now	0	1 (1)
over	1 (6)	0
party	0	1 (16)
quick	1 (2)	0
their	0	1 (15)
time	0	1 (4)

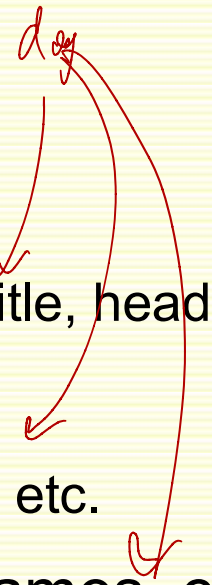
time AND come → Doc 2

time (NEAR 2) come → empty

quick (NEAR 2) fox → Doc 1

quick WITH fox → empty

Other Extensions



- Ability to search on fields
  - Leverage document structure: title, headings, etc.
- Wildcards
  - lov* = love, loving, loves, loved, etc.
- Special treatment of dates, names, companies, etc.

Why Boolean Retrieval Works

- Boolean operators *approximate* natural language
  - Find documents about a good party that is not over
- AND can discover relationships between concepts
  - good party
- OR can discover alternate terminology
  - excellent party, wild party, etc.
- NOT can discover alternate meanings
  - Democratic party

Why Boolean Retrieval Fails

- Natural language is way more complex
- AND "discovers" nonexistent relationships
  - Terms in different sentences, paragraphs, …
- Guessing terminology for OR is hard
  - good, nice, excellent, outstanding, awesome, …
- Guessing terms to exclude is even harder!
  - Democratic party, party to a lawsuit, …

Strengths and Weaknesses

- Strengths
  - Precise, if you know the right strategies
  - Precise, if you have an idea of what you're looking for
  - Efficient for the computer
- Weaknesses
  - Users must learn Boolean logic
  - Boolean logic insufficient to capture the richness of language
  - No control over size of result set: either too many documents or none
  - When do you stop reading? All documents in the result set are considered "equally good"
  - What about partial matches? Documents that "don't quite match" the query may be useful also