User-Centered and System-Centered IR

Information Retrieval

Lecture 2

User tasks
Role of the system
Document view and model

What is Information Retrieval?

IR is the study of the

- Representation
- Storage
- Organization
- Access

of information items

- articles, books, web pages, CDs, movies ...

for the people who are interested in them.

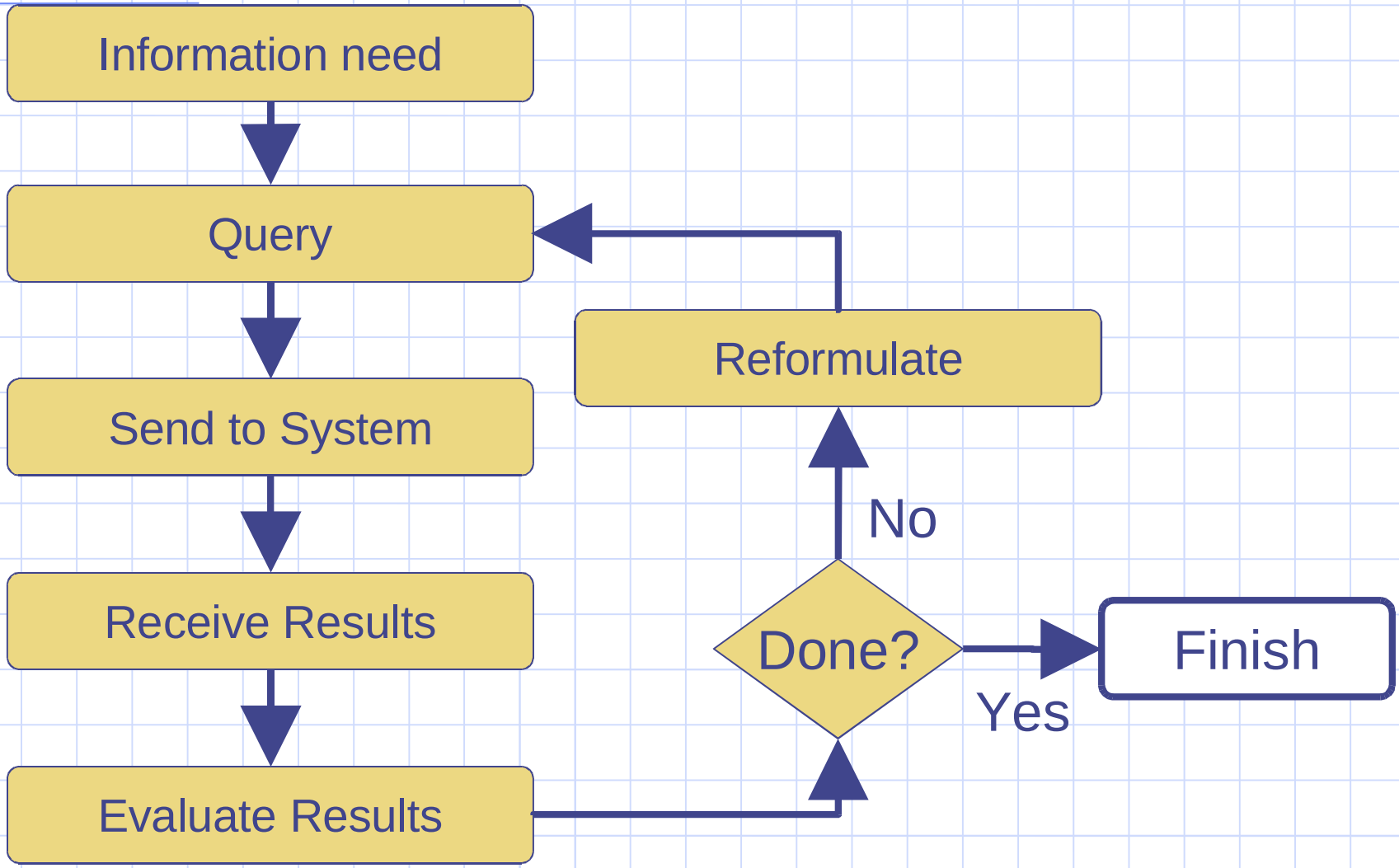
The User Task

- A person has a *goal* to accomplish:
  - find a plumber
  - keep informed about a business competitor
  - write a scholarly article
  - investigate an allegation of fraud
- Along the way, he needs to find information to accomplish the goal.

Three Kinds of User Tasks

1. Monitoring a well-known topic over time
2. Following a plan of searches to learn a specific thing
3. Undirected exploration to gain basic knowledge on a topic

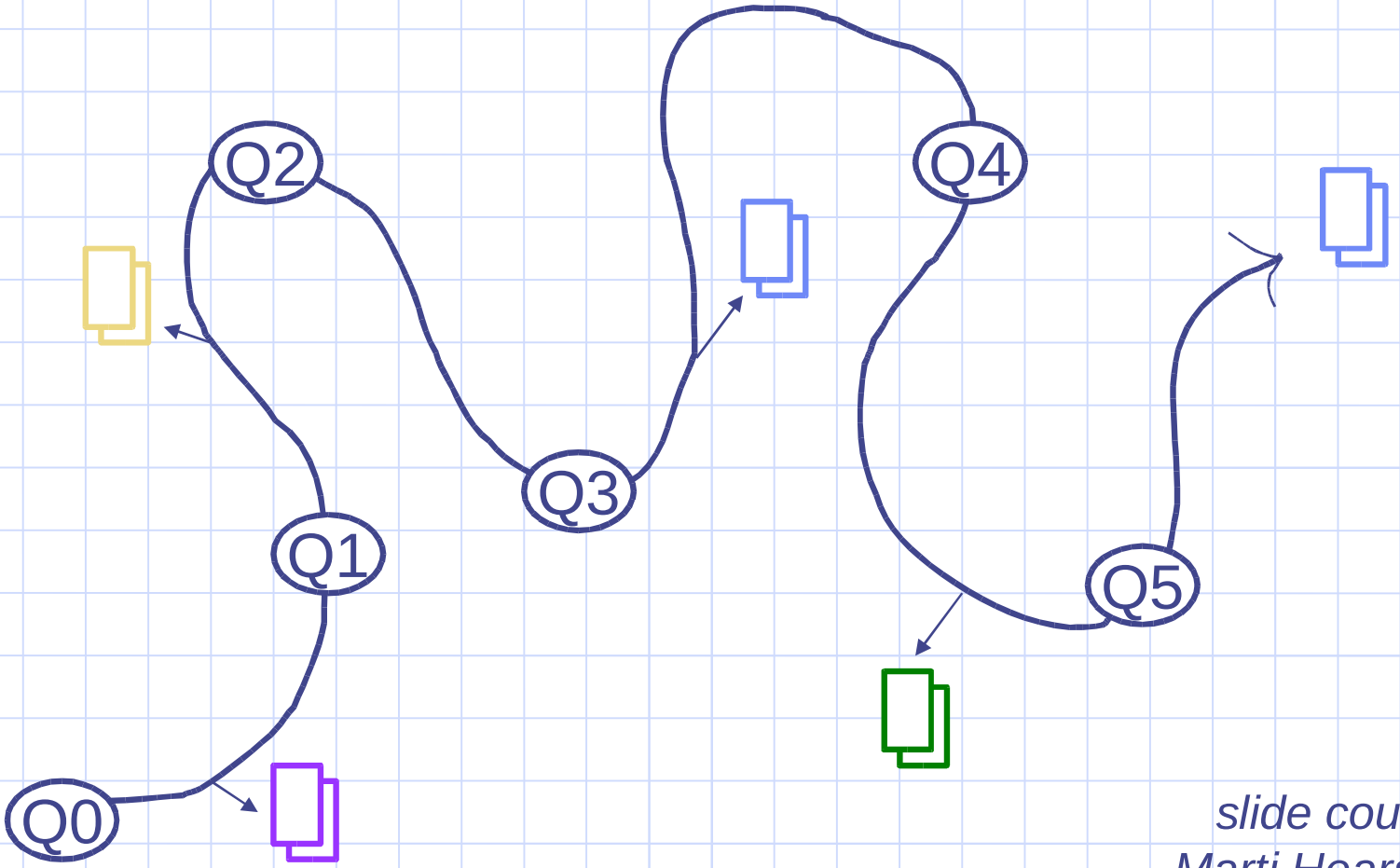
Classical Model of Info Seeking



"Berry-Picking" Model

- Users learn as they search
  - causes need to change
  - causes queries to shift around, not refine
  - one goal leads to another
- Information needs not satisfied by a single set of documents
  - really by bits and pieces found along the way

A sketch of a searcher… "moving through many actions towards a general goal of satisfactory completion of research related to an information need." (after Bates 89)



slide courtesy of Marti Hearst, UCB

Role of the IR System

- Berry-picking model is more realistic
  - User's goal is not to search
  - Searching is part of pursuit of the goal.
- Thus, IR engine is only a *tool* to support searching
  - Only one part of the toolbox
  - Don't let the only tool be a hammer!
  - But make your tool effective, efficient, and flexible

IR vs. DB Retrieval

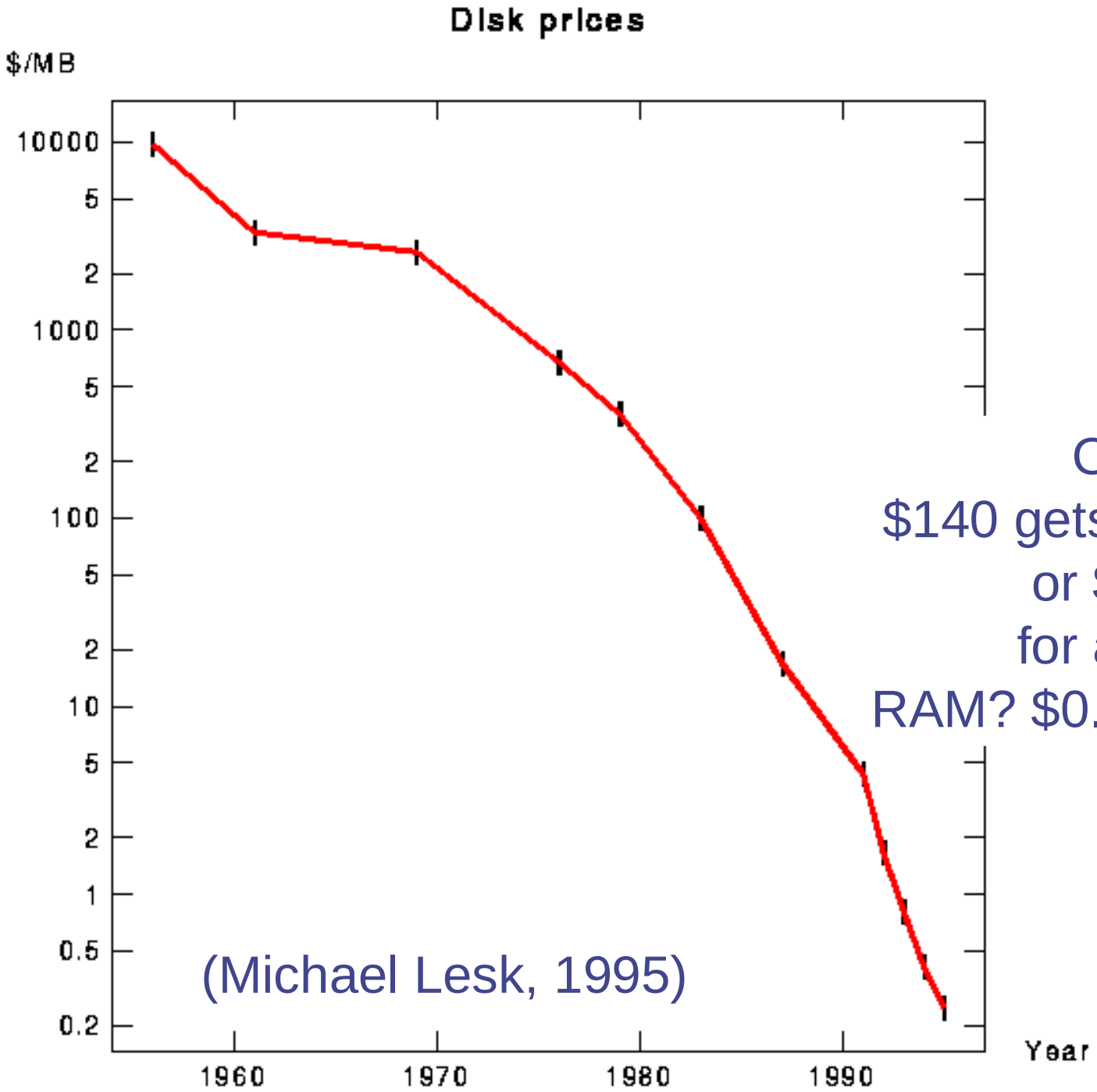
- IR is very different from databases!
- Best match, rather than exact match
  - Inference is inductive, not deductive
  - Models are generally "probabilistic"
  - Relevant results instead of matching results
- Query language natural and informal
  - Query specification incomplete
- Query processing less sensitive to bad queries
  - However, users still sensitive to bad results

Representing Documents

- Logical view of a document
  - how the system represents a document
- Might include
  - structural information
  - multiple types of information (multimedia)
  - metadata
- Any logical view loses some information

Full Text representation

- Historical representation
  - set of key words or index terms
  - assigned automatically or by hand
- Modern computing hardware allows the use of "full-text" representations
  - set of all words contained in a document
  - easier to automate than assigning index terms



On 9/3/2002,
$140 gets you 80 GB
or $0.0018/MB
for an IDE disk.
RAM? $0.30-0.40/MB

(Michael Lesk, 1995)

All the words?

So, what are the words, anyway?

- Alphabetic sequence of characters
- Punctuation? Is hyphenation 1 or 2 words?
- Dates? Prices?
- Is whitespace important?
- In some languages, *segmentation* is harder
  - for example, Chinese

All the words?

- Sometimes, you don't want *all* the words
- Stop list
  - some words contain less information
  - words like *and, the, in, of, which, that*
- Stemming
  - some words relate to the same concept
  - computer, computation, computing: *comput*
  - computer, pc, von-neumann architecture?

Just the words?

- So far, we have a "bag of words"
- No information about lexical structure
- N-grams
  - sequences of *n* adjacent words in text
  - can also have sequences of *n* characters
- Identify noun groups or phrases

Document Structure

- Most documents are not simply streams of bytes, but have a structure
  - Fixed: set of fields
  - Hierarchical: chapter, section, title, figure
  - Hypertext: directed graph
- The structure leads to a *document model* which defines the granularity of a search

Bringing it together

- With
  - a set of documents
  - a document model defining what can be searched
  - a document logical view or representation defining what to look for

we can begin to construct the text database and index.