



College of Information Studies

University of Maryland Hornbake Library Building College Park, MD 20742-4345

Search Engines

Session 5

INST 301

Introduction to Information Science

THE WORLD'S CAPACITY TO STORE INFORMATION

This chart shows the world's growth in storage capacity for both analog data (books, newspapers, videotapes, etc.) and digital (CDs, DVDs, computer hard drives, smartphone drives, etc.)

In gigabytes or estimated equivalent

1986
ANALOG
2.62 billion

DIGITAL
0.02 billion

ANALOG STORAGE

DIGITAL

2000

2007
ANALOG
18.86 billion gigabytes

Paper, film, audiotape and vinyl: 6.2%

Analog videotapes: 93.8%

Other digital media: 0.8%*

Portable media players, flash drives: 2%

Portable hard disks: 2.4%

CDs and minidisks: 6.8%

Computer servers and mainframe hard disks: 8.9%

Digital tape: 11.8%

DVD/Blu-ray: 22.8%

PC hard disks: 44.5%

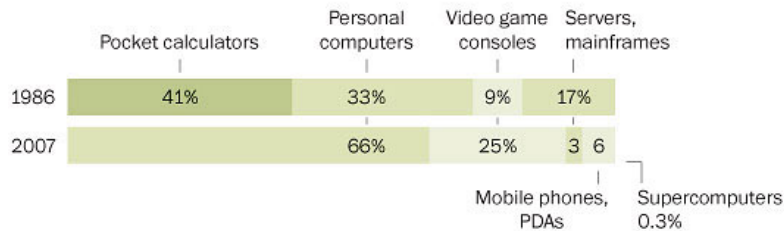
123 billion gigabytes

*Other includes chip cards, memory cards, floppy disks, mobile phones/PDAs, cameras/camcorders, video games

COMPUTING POWER

In 1986, pocket calculators accounted for much of the world's data-processing power.

Percentage of available processing power by device:



2007
DIGITAL
276.12 billion gigabytes

so what is a

Search Engine?

[All](#)[Shopping](#)[Images](#)[Videos](#)[News](#)[More ▾](#)[Search tools](#)

About 173,000,000 results (0.55 seconds)

Commercial cat food[edit] Most store-bought cat food comes in either dry form, also known in the US as kibble, or wet **canned** form. Some manufacturers sell frozen **raw** diets and premix products to cater to owners who feed **raw**.



[Cat food - Wikipedia, the free encyclopedia](#)

https://en.wikipedia.org/wiki/Cat_food Wikipedia ▾

Feedback

[Cat Food Brands: Grain Free, Organic Cat & Kitten Food ...](#)

www.petsmart.com/cat/food/cat-36-catid-200004 ▾ PetSmart ▾

Give your **cat food** designed for her. We can help you find grain free, organic and natural **cat food** brands that meet her unique nutritional needs.

[Dry Food](#) · [Canned Food](#) · [Food Pouches](#) · [Milk Replacers](#)

[Cat Food: Grain Free, Natural, Organic Cat Food | petco.com](#)

www.petco.com/shop/en/petcostore/cat/cat_food ▾ Petco ▾

[All](#)[Shopping](#)[Images](#)[Videos](#)[News](#)[More ▾](#)[Search tools](#)

About 173,000,000 results (0.55 seconds)

Commercial cat food[edit] Most store-bought cat food comes in either dry form, also known in the US as kibble, or wet **canned** form. Some manufacturers sell frozen **raw** diets and premix products to cater to owners who feed **raw**.



[Cat food - Wikipedia, the free encyclopedia](#)

https://en.wikipedia.org/wiki/Cat_food Wikipedia ▾

Feedback

[Cat Food Brands: Grain Free, Organic Cat & Kitten Food ...](#)

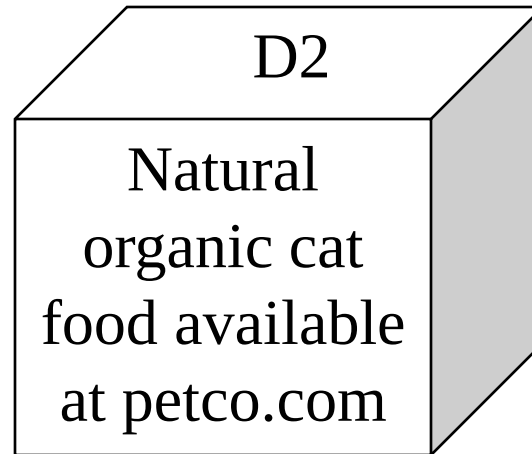
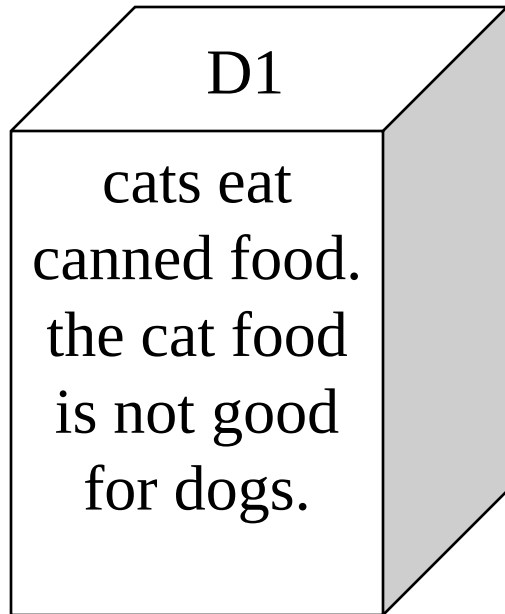
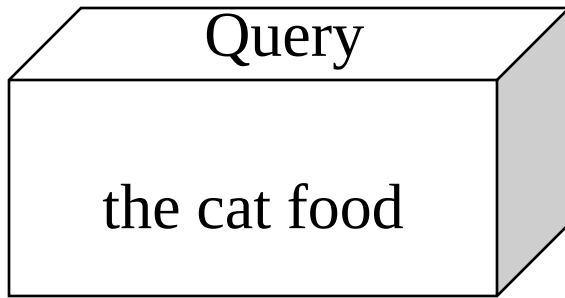
www.petsmart.com/cat/food/cat-36-catid-200004 ▾ PetSmart ▾

Give your **cat food** designed for her. We can help you find grain free, organic and natural **cat food** brands that meet her unique nutritional needs.

[Dry Food](#) · [Canned Food](#) · [Food Pouches](#) · [Milk Replacers](#)

[Cat Food: Grain Free, Natural, Organic Cat Food | petco.com](#)

www.petco.com/shop/en/petcostore/cat/cat_food ▾ Petco ▾





Find all the browser extensions and No Index

components of, 173
emphasize intangibles, 177–80
incentive pay, 169–72
keep pay low, 165–68
merit pay, 169
principles of, 165–80
principles of executive pay, 173
Concern reflex, 102
Confirmation
defined, 128
using to advance Radical
Demotivation™, 147–48
Cook, C., 121, 228
Core values, 54–58, 60–64
and cynicism, 55
avoid socially transcendent, 59
burden of authenticity, 58
defined, 54
guidelines for developing, 59–61
introduction to the organization,
61–63
reinforcing in daily life, 63–64
Covey, S., 64, 228
Crawford, R., 96, 228
Crowe, M., 200, 228
Csikszentmihalyi, M., 39, 228
Culture of blame, 118
Custer's Bluster™, 129

D

Deal, T., 53, 65, 66, 71, 227, 228
Deci, E., 118, 169, 171, 228
Demotivation Vortex, 49, 50
DeNicholas, M., 99, 139, 227

defined, 128
elements of, 132
strategem to confront the Noble
Employee Myth, 128–31
strategies
for expressing disqualification,
146–47
for expressing imperviousness,
143–46
for expressing indifference,
133–43
Dweck, C., 116, 228

E

Eisenberg, E., 7, 60, 228
Eisenhardt, K., 195, 197, 203, 228
Elliott, E., 116, 228
Ellis, D., 187, 230
Employees
and infantile rage, 106
and psychological obesity, 127
artificially motivated, 44
impact of, 45
as paradigmatic failures, 79
authentically motivated, 44
critical of motivational
programs, 45
impervious to Radical
Demotivation™, 46
bad attitudes, 105–7
become poorer value over time,
152
believe others have an unfair
advantage, 105

How about here

- This is what indexing does
- Makes data accessible in a structured format, easily accessible through search.

Building Index

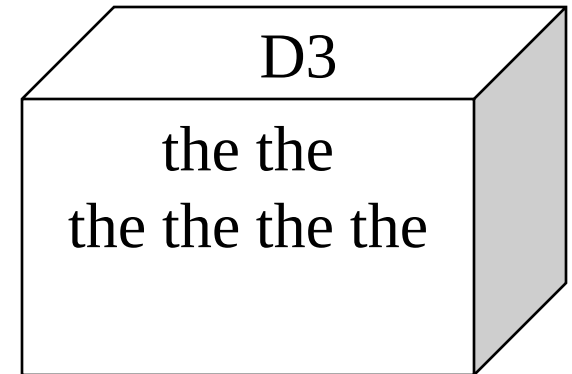
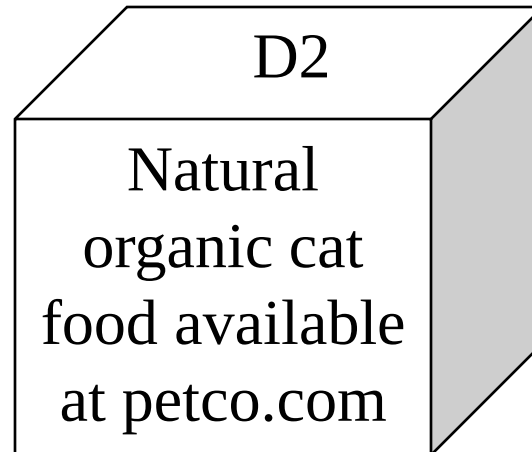
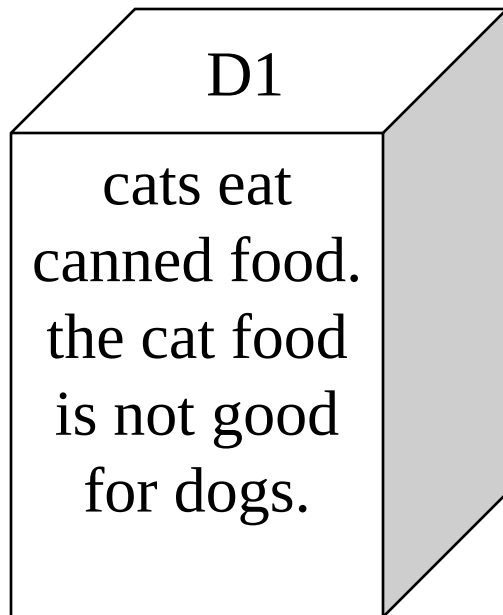
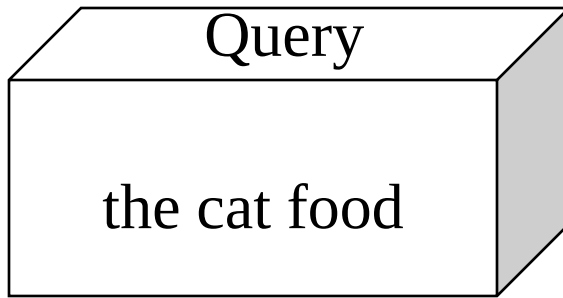
Documents:

1: cats eat canned food. the cat food is not good for dogs.

2: natural organic cat food available at petco.com

Term – Document Index Matrix

TERM	D1	D2
available	0	1
canned	1	0
cat	2	1
dog	1	0
eat	?	?
food	?	?
...	...	...



**Some terms are more informative
than others**

How Specific is a Term?

TERM (t)	Document Frequency of term t (df_t)	Inverse Document Frequency of term t ($idf_t = (N/df_t)$)	Log of Inverse Document Frequency of term t [$\log(idf_t)$]
cat	1	1,000,000	
petco.com	100	10,000	
food	1000	1000	
canned	10,000	100	
good	100,000	10	
the	1,000,000	1	

How Specific is a Term?

TERM (t)	Document Frequency of term t (df_t)	Inverse Document Frequency of term t ($idf_t = (N/df_t)$)	Log of Inverse Document Frequency of term t [$\log(idf_t)$]
cat	1	1,000,000	
petco.com	100	10,000	
food	1000	1000	
canned	10,000	100	
good	100,000	10	
the	1,000,000	1	

Magnitude of increase

How Specific is a Term?

TERM (t)	Document Frequency of term t (df_t)	Inverse Document Frequency of term t ($idf_t = (N/df_t)$)	Log of Inverse Document Frequency of term t [$\log(idf_t)$]
cat	1	1,000,000	6
petco.com	100	10,000	4
food	1000	1000	3
canned	10,000	100	2
good	100,000	10	1
the	1,000,000	1	0

Putting it all together

- To rank, we obtain the weight for each term using tf-idf
- The tf-idf weight of a term is the product of its tf weight and its idf weight

$$Weight(t) = tf_t \times \log(N / df_t)$$

- Using the term weights, we obtain the document weight



the earthquake



All

News

Videos

Images

Maps

More ▾

Search tools

About 118,000,000 results (0.62 seconds)

In the news



Powerful earthquake rocks southern Taiwan

Al Jazeera - 3 hours ago

At least 23 people were killed and scores injured when the quake struck the city of Tainan ...

Taiwan struck by earthquake: skyscrapers flattened

Sydney Morning Herald - 52 mins ago

Rescue Efforts Continue as Toll Rises in Taiwan Earthquake

New York Times - 17 hours ago

[More news for the earthquake](#)

Finding based on MetaData or Description

- A type of “document expansion”
 - Terms near links describe content of the target
- Works even when you can’t index content
 - Image retrieval, uncrawled links, ...

[Bean - "And that's the way we tried to do every rock. Because you always had the gnomon. And then we took a photo afterwards."]

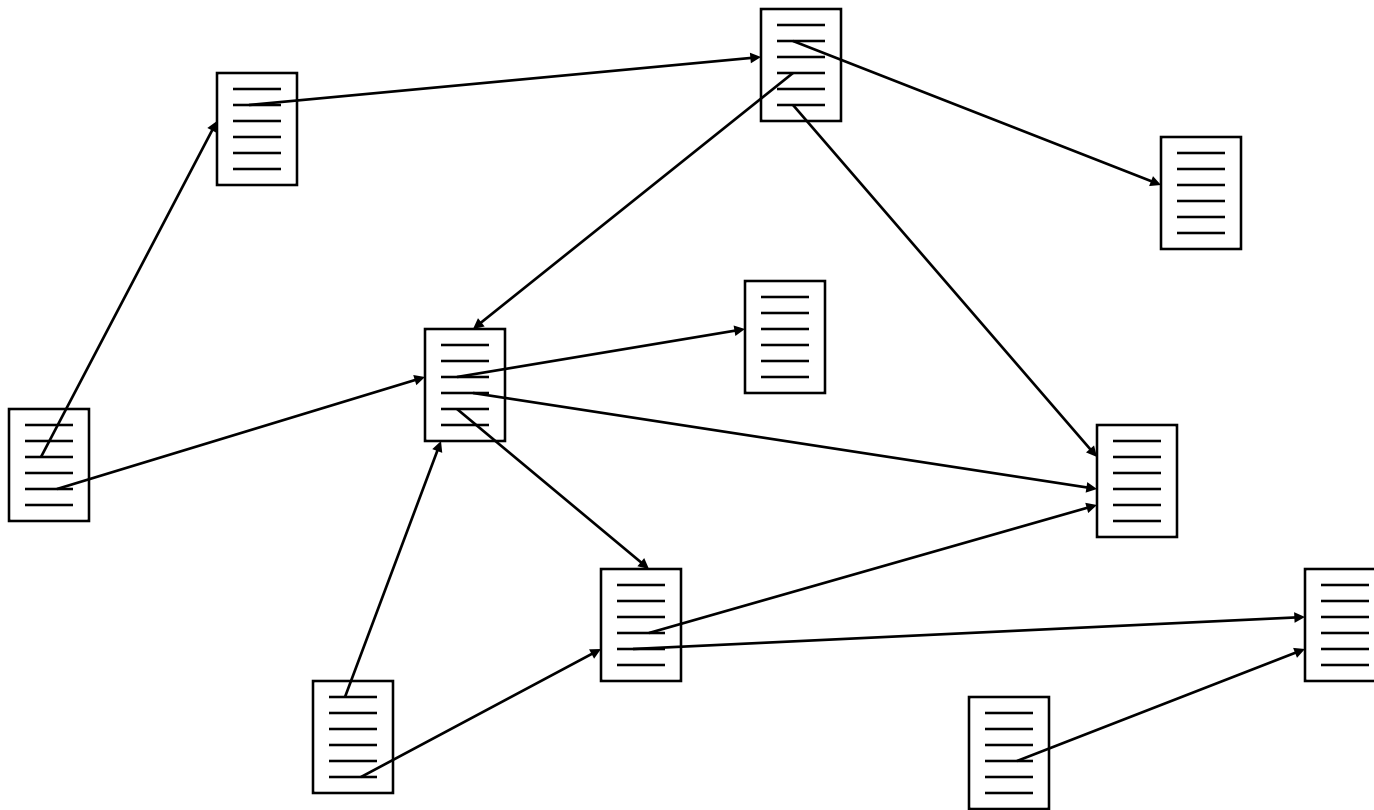
[Conrad - "We practiced this...I started out by just laying rocks around on the floor. One of the things was setting the camera deal; we had the three (focus) distances. And what we did was actually take pictures to calibrate ourselves. They developed that film in training to make sure we stood the right distance."]



Ways of Finding Information

- Searching content
 - Characterize documents by the words the contain
- Searching behavior
 - Find similar search patterns
 - Find items that cause similar reactions
- Searching description
 - Anchor text

Crawling the Web



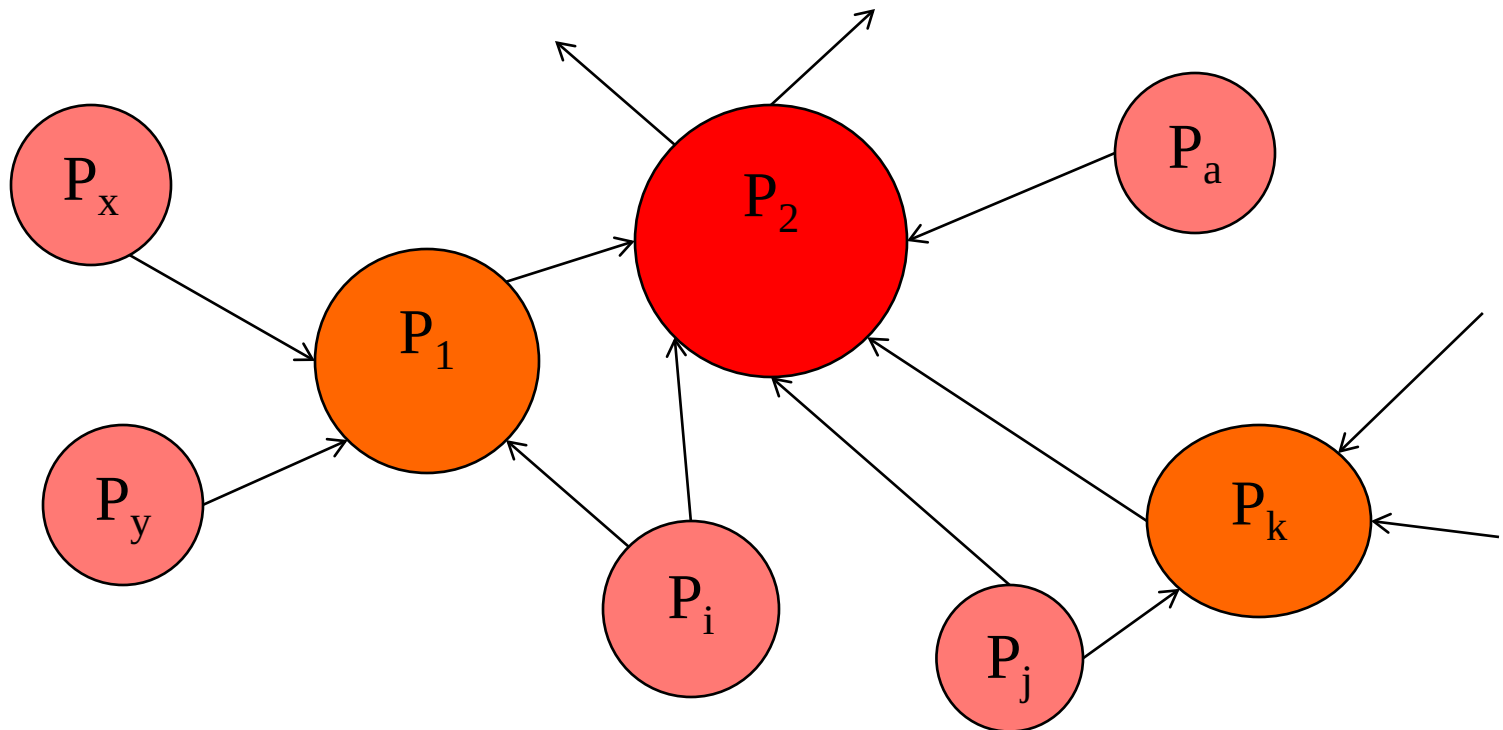
Web Crawl Challenges

- Adversary behavior
 - “Crawler traps”
- Duplicate and near-duplicate content
 - 30-40% of total content
 - Check if the content is already index
 - Skip document that do not provide new information
- Network instability
 - Temporary server interruptions
 - Server and network loads
- Dynamic content generation

How does Google PageRank work?

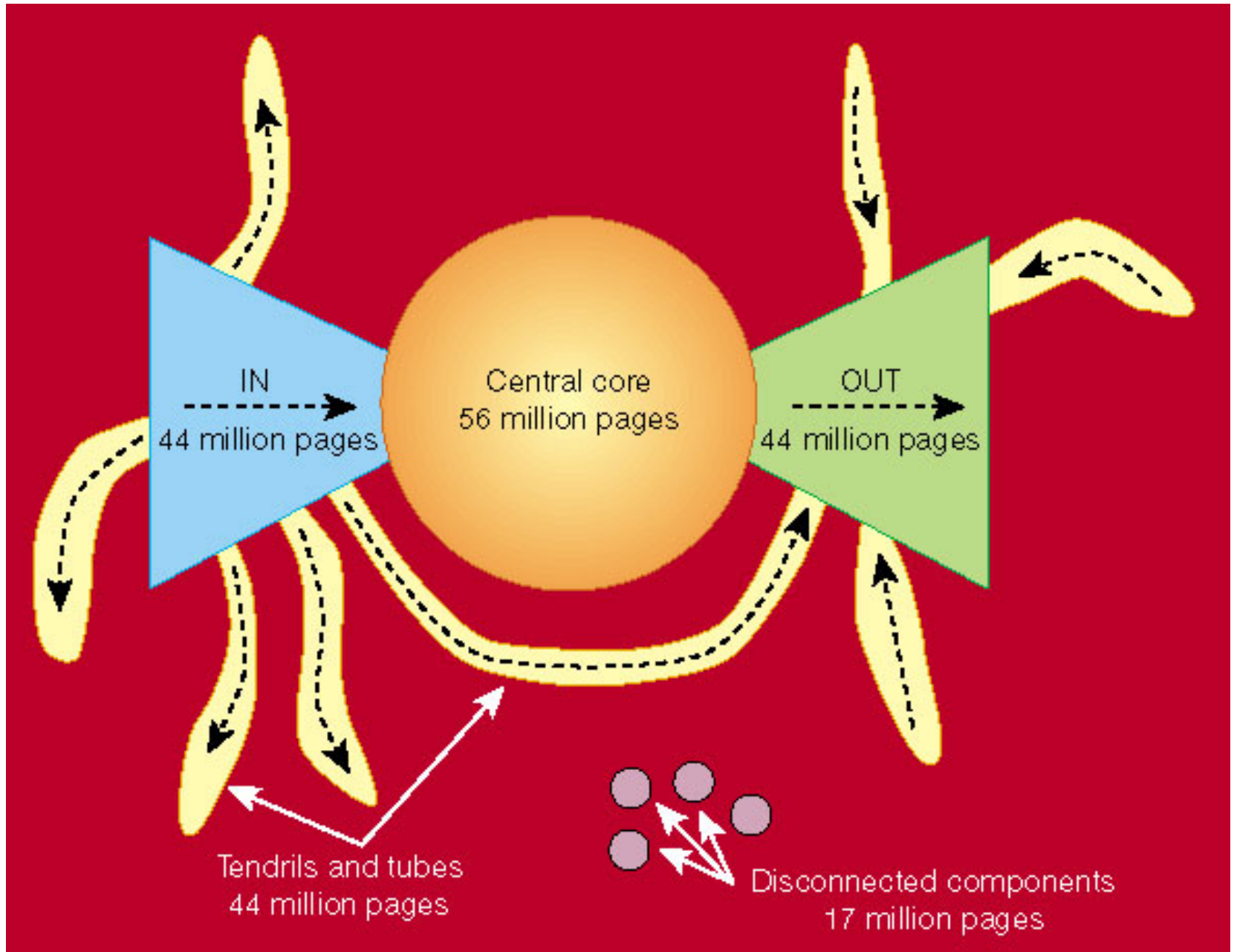
Objective - estimate the importance of a webpage

- Inlinks are “good” (like recommendations)
- Inlinks from a “good” site are better than inlinks from a “bad” site



Link Structure of the Web

Nature 405, 113 (11 May 2000) | doi:10.1038/35012155



So,

A Web search engine is an application composed of ;

CRAWLING component

- important to define a search space

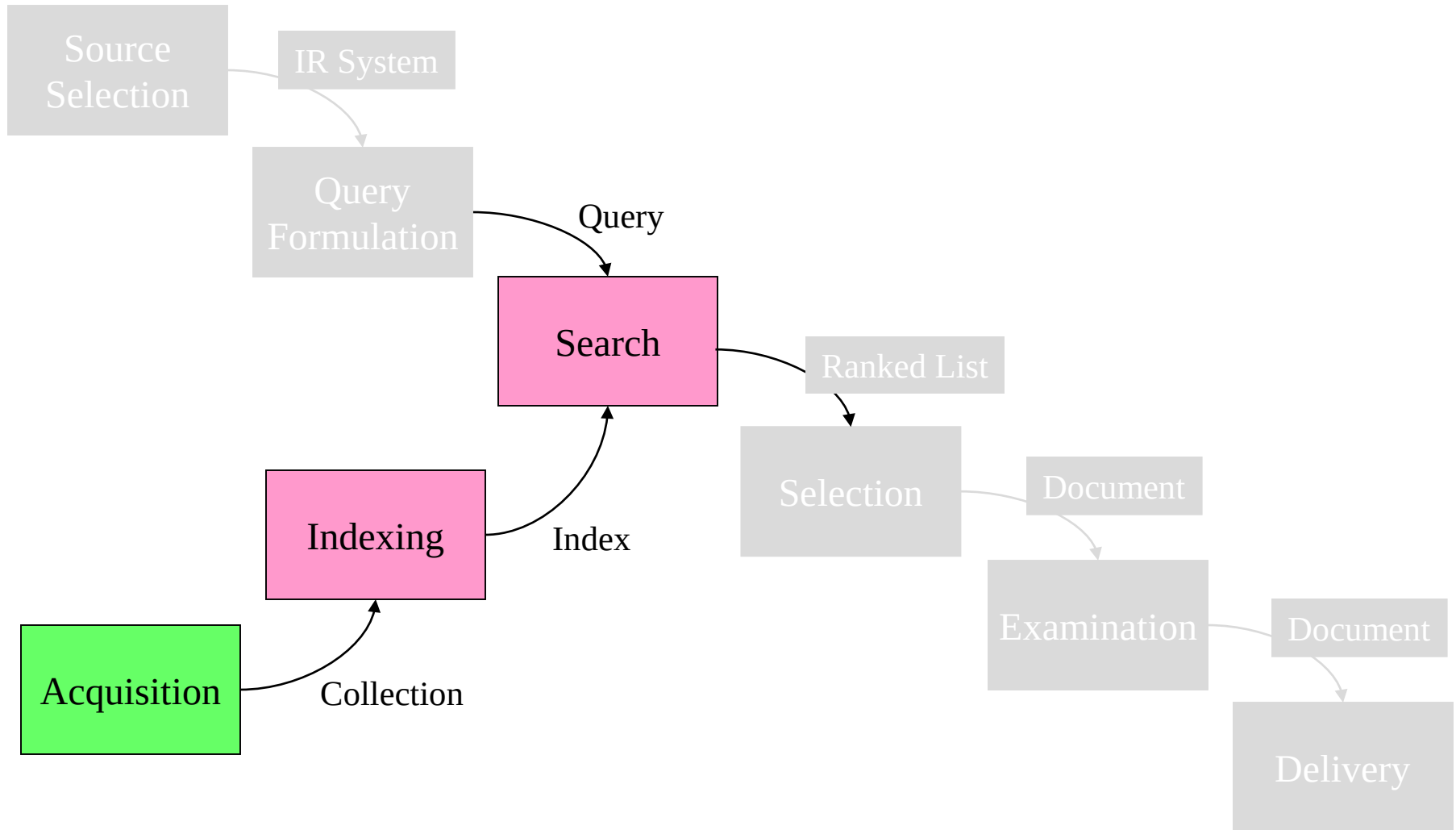
INDEXING component

- of importance to developers AND content-centric

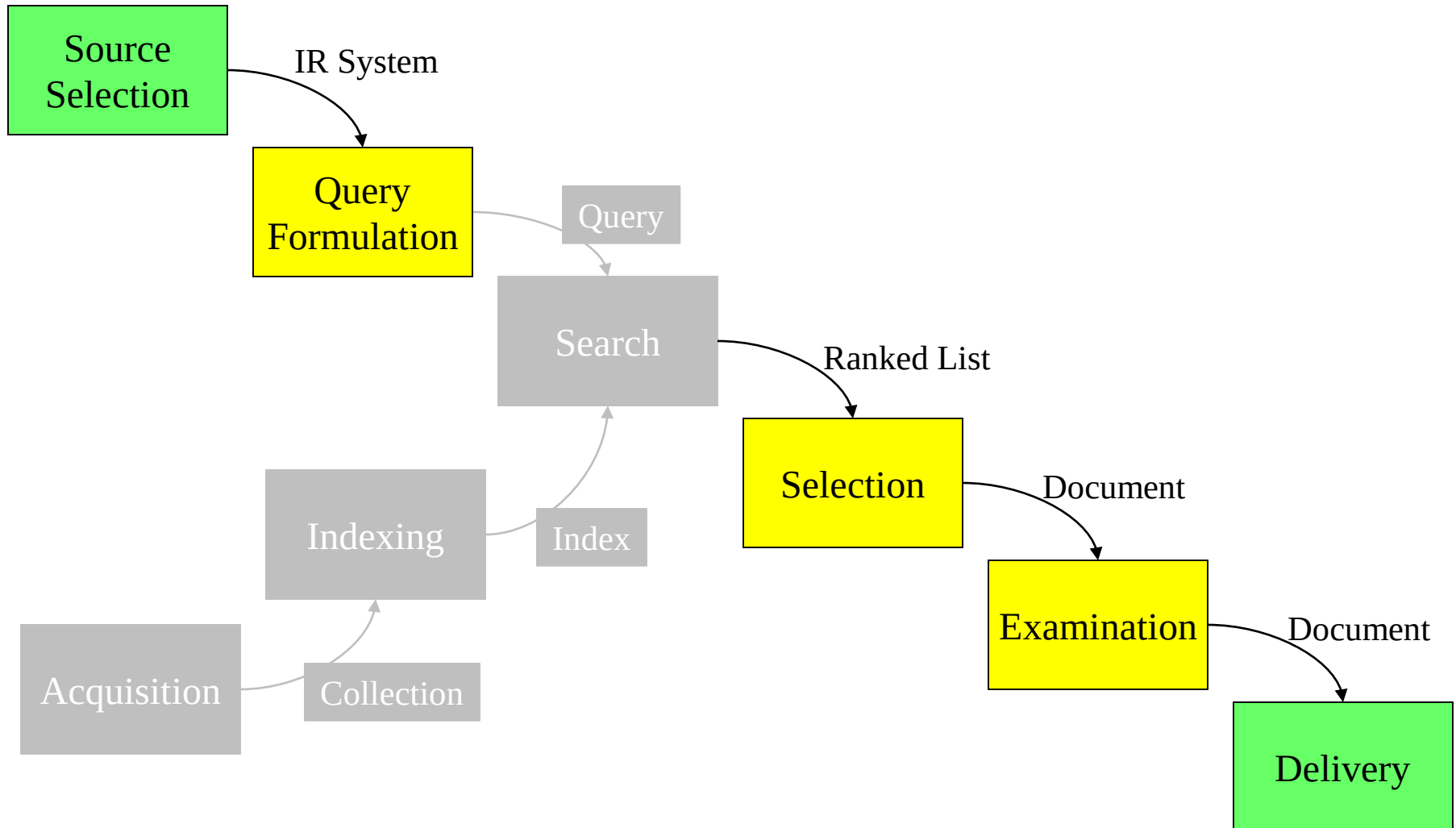
SEARCH component

- of importance to the users AND user-centric

Today: The “Search Engine”



Next Session: “The Search”



Before You Go

- Assignment H2

On a sheet of paper, answer the following
(ungraded) question (no names, please):

What was the muddiest point in
today's class?