

Search Technology

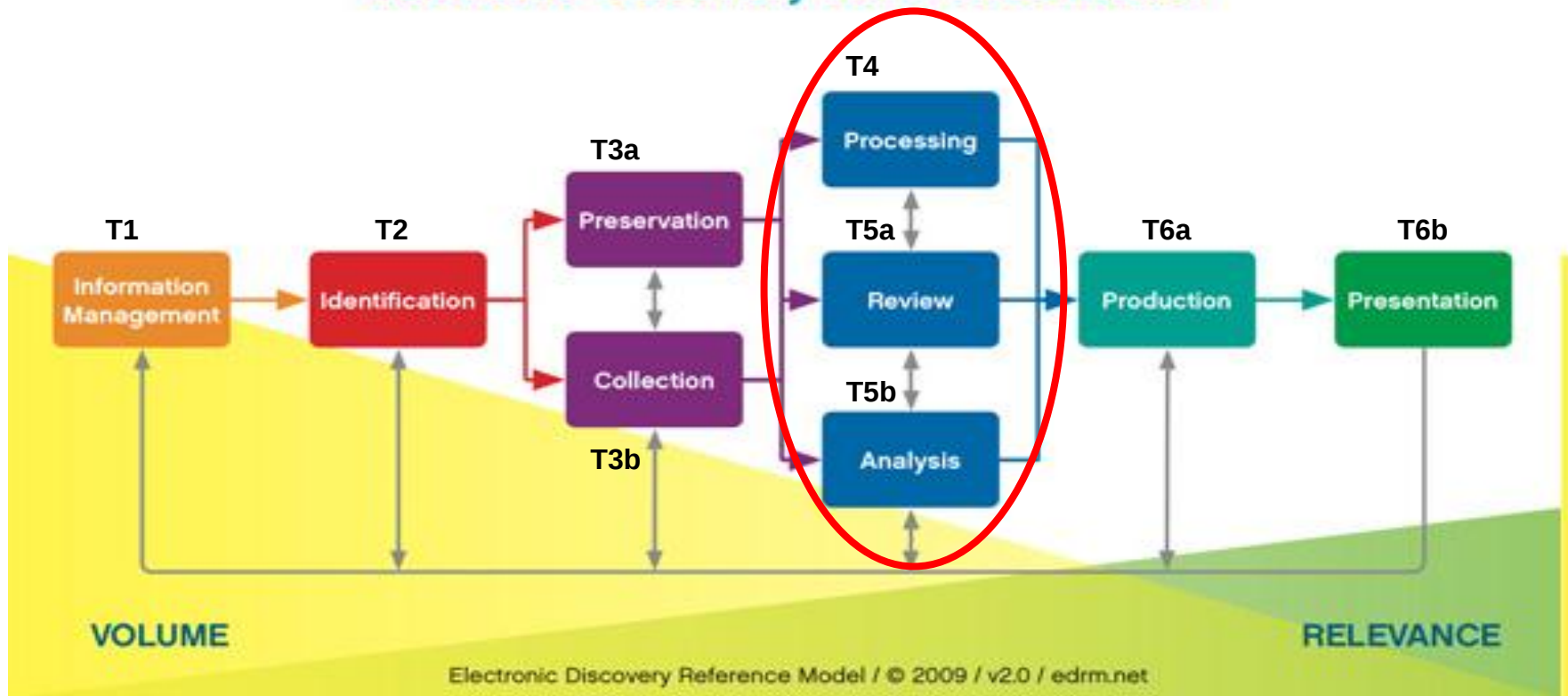
LBSC 708X/INFM 718X

Week 5

Doug Oard

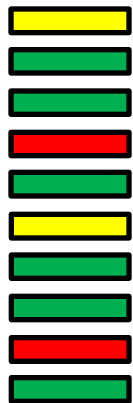
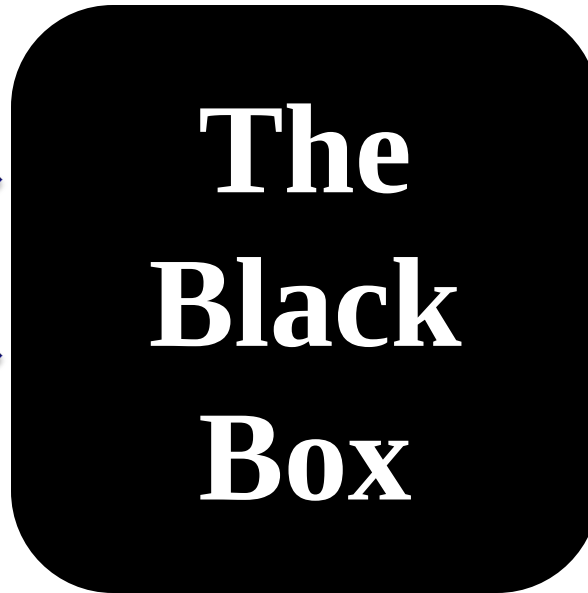
Where Search Technology Fits

Electronic Discovery Reference Model

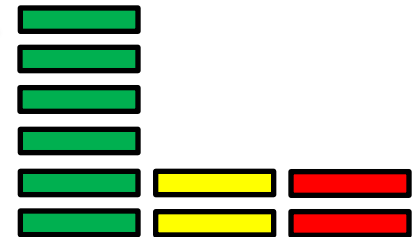


Document Review

Case Knowledge



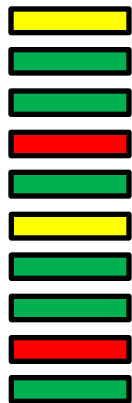
Unprocessed
Documents



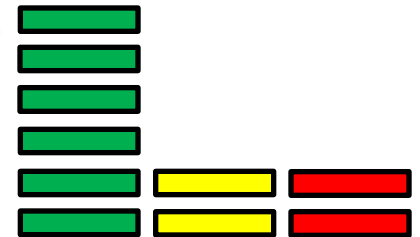
Coded
Documents

Inside Yesterday's Black Box

Case Knowledge



**Unprocessed
Documents**



**Coded
Documents**

“Linear Review”

PLANNING AND PROJECT MANAGEMENT

- Ensure a project plan is tailored to the specifications of counsel and consistent with best practices
- Deliver a key set of documents that govern the execution and project management of the review process

TEAM SELECTION AND TRAINING

- Develop specific job descriptions and define a detailed protocol for recruiting, testing, and selection
- Conduct reference and background checks, and a conflicts check, where necessary
- Employ team members previously used on similar projects
- Ensure the review team receives comprehensive substantive and platform training

WORKFLOW

- Design processes, assignments and quality assurance steps specifically geared to the project's requirements
- Demonstrate compliance with key security and quality standards while maintaining acceptable pace

QUALITY CONTROL

- Develop quality control processes to achieve key project goals
- Implement controls to manage privilege designation and preparation/validation of results for production
- Test first review work product using sampling, targeted re-review, and validations searches
- Employ formal statistics to ensure the highest quality end result
- Maintain performance tracking for all reviewers

COMMUNICATION

- Develop a formal schedule of communications with counsel
- Calibrate initial review results, seeking counsel's guidance to confirm or correct results and to conform review protocol and training materials to insights gained

REPORTING

- Deliver regular, comprehensive reports to monitor progress and quality and to assist counsel in managing the review process

PRODUCTIONS AND PRIVILEGE LOGS

- Isolate and validate producible documents for counsel's imprimatur
- Prepare privilege logs in accordance with specifications set by counsel

Is it reasonable?

- Yes, if we followed a reasonable process.
 - Staffing
 - Training
 - Quality assurance

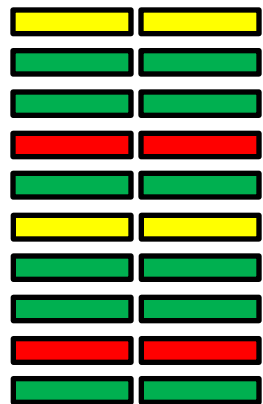
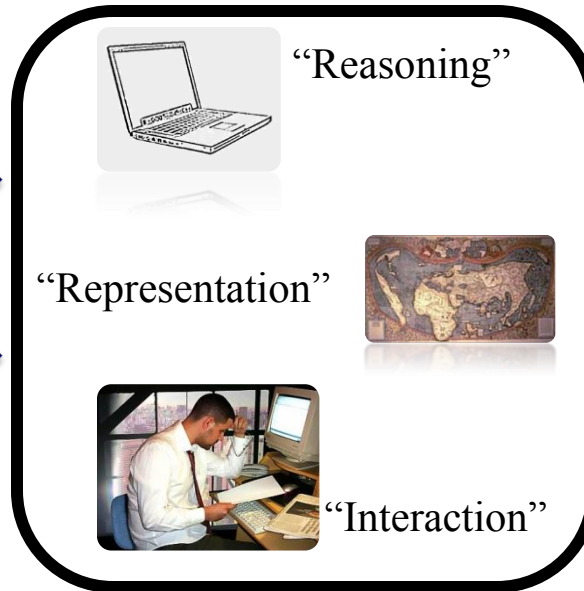


Linear Review

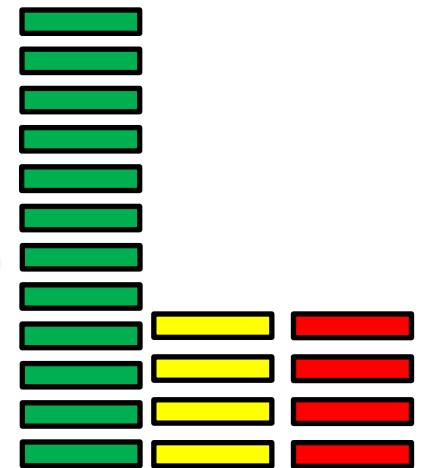
Inside Today's Black Box

Case Knowledge

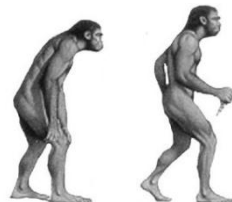
Keyword Search & Linear Review



Unprocessed
Documents



Coded
Documents



Example of Boolean search string from *U.S. v. Philip Morris*

- (((master settlement agreement OR msa) AND NOT (medical savings account OR metropolitan standard area)) OR s. 1415 OR (ets AND NOT educational testing service) OR (liggett AND NOT sharon a. liggett) OR atco OR lorillard OR (pmi AND NOT presidential management intern) OR pm usa OR rjr OR (b&w AND NOT photo*) OR phillip morris OR batco OR ftc test method OR star scientific OR vector group OR joe camel OR (marlboro AND NOT upper marlboro)) AND NOT (tobacco* OR cigarette* OR smoking OR tar OR nicotine OR smokeless OR synar amendment OR philip morris OR r.j. reynolds OR ("brown and williamson") OR ("brown & williamson") OR bat industries OR liggett group)

Is it reasonable?

- Yes, if we followed a reasonable process.
 - Indexing
 - Query design
 - Sampling



Linear Review

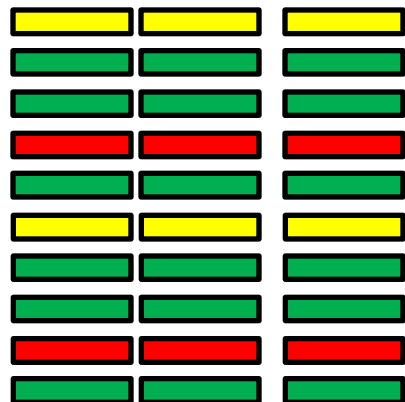
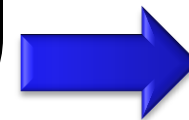
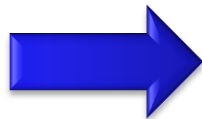


•Keyword Search
•Linear Review

Inside Tomorrow's Black Box

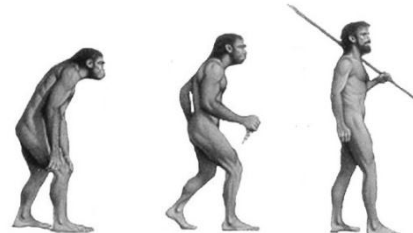
Case Knowledge

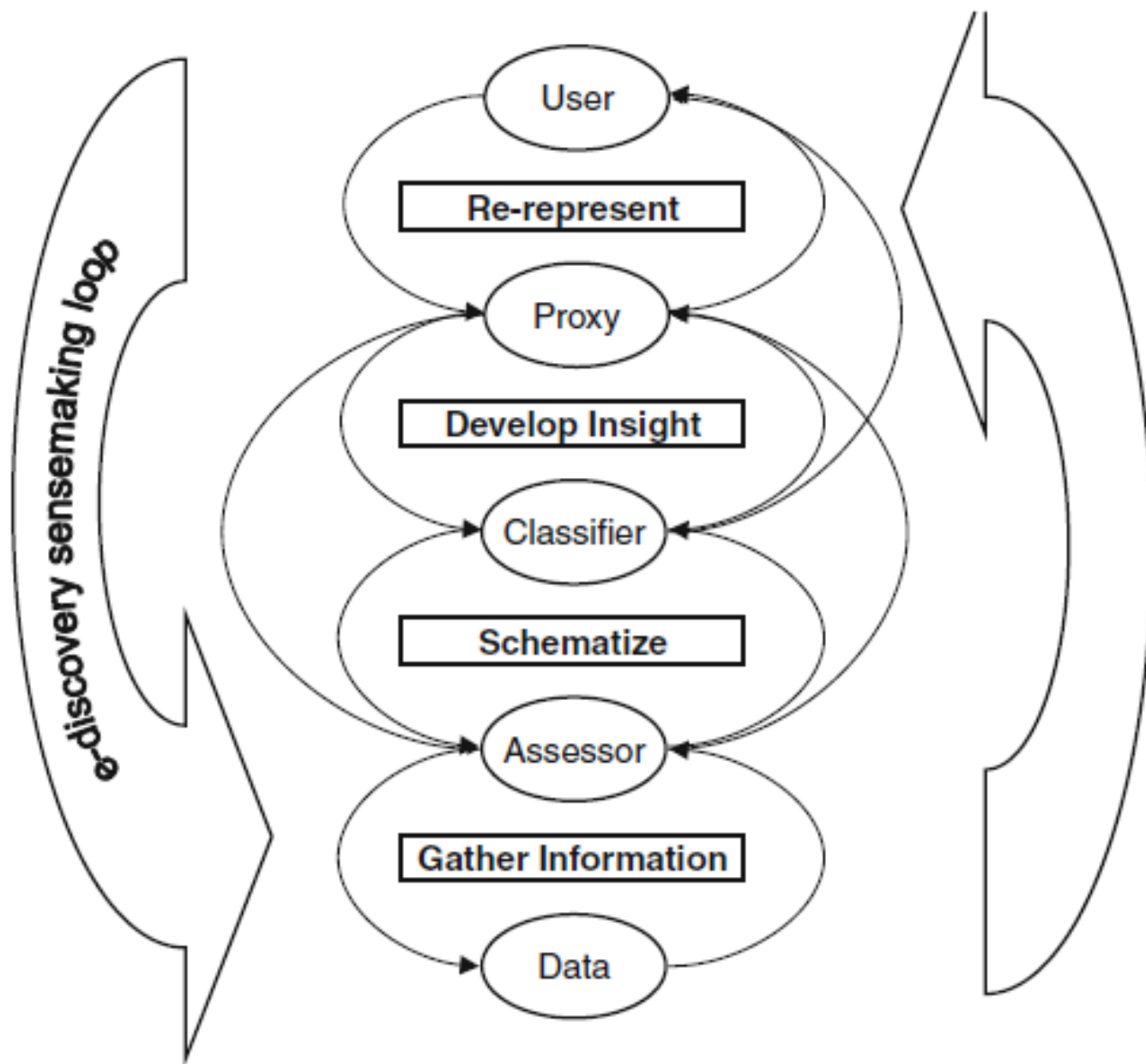
Technology Assisted Review



Unprocessed
Documents

Coded
Documents





Is it reasonable?

- Yes, if we followed a reasonable process.
 - Rich representation
 - Explicit & example-based interaction
 - Process quality measurement



Agenda

- Three generations of e-discovery

➤ Design thinking

- Content-based search example
- Putting it all together

Databases vs. IR

	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 (mathematically) defined queries. Unambiguous.	Vague, imprecise information needs (often expressed in natural language).
Results we get	Exact. Always correct in a formal sense.	Sometimes relevant, often not.
Interaction with system	One-shot queries.	Interaction is important.
Other issues	Concurrency, recovery, atomicity are all critical.	Issues downplayed.

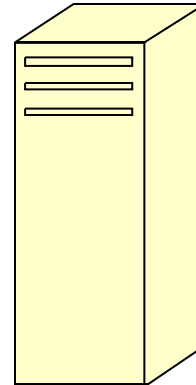
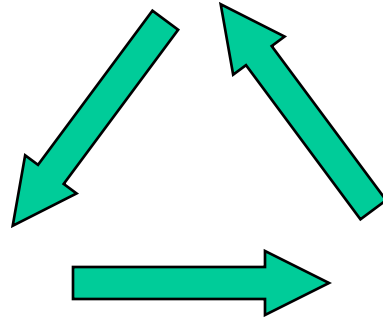
Design Strategies

- Foster human-machine synergy
 - Exploit complementary strengths
 - Accommodate shared weaknesses
- Divide-and-conquer
 - Divide task into stages with well-defined interfaces
 - Continue dividing until problems are easily solved
- Co-design related components
 - Iterative process of joint optimization

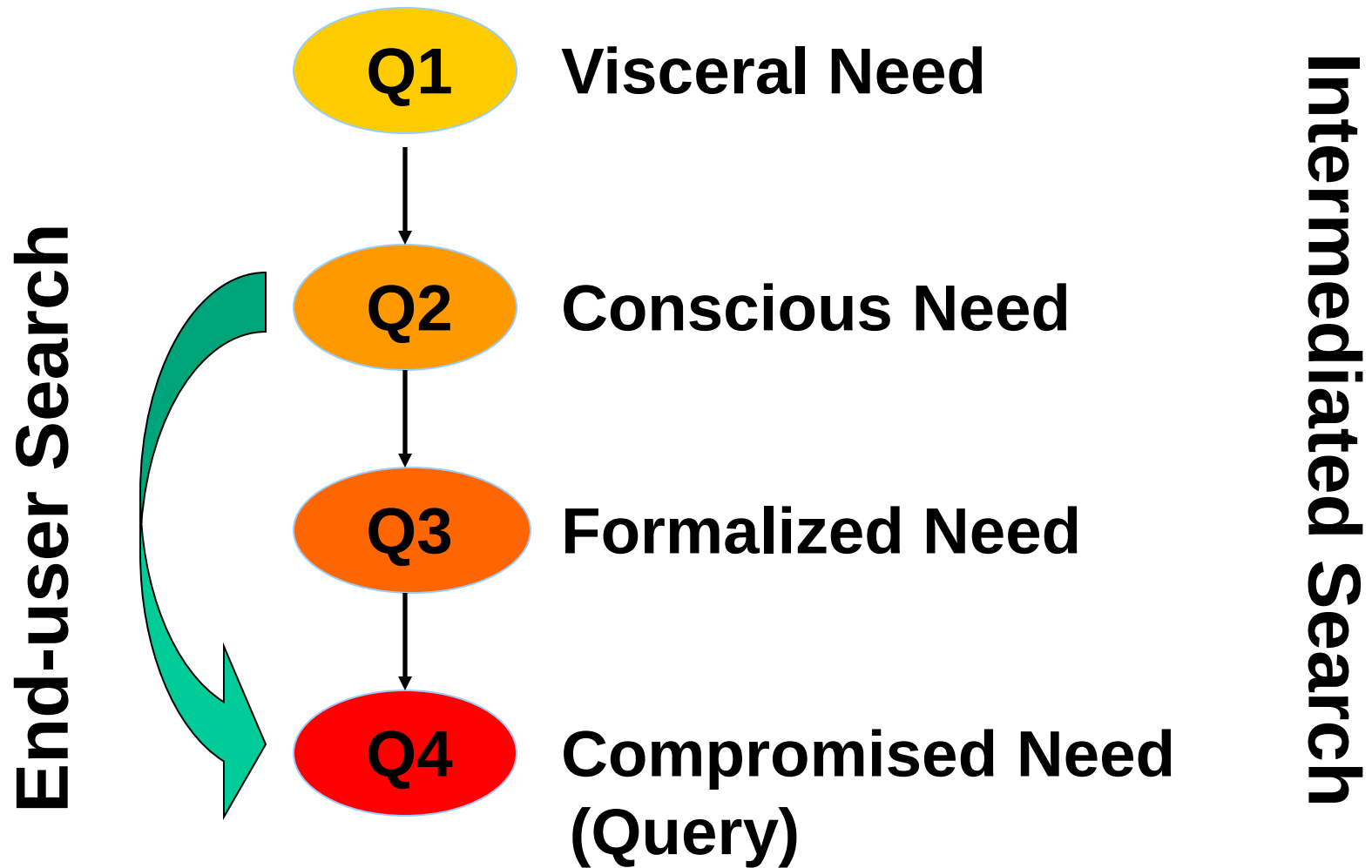
Human-Machine Synergy

- Machines are good at:
 - Doing simple things accurately and quickly
 - Scaling to larger collections in sublinear time
- People are better at:
 - Accurately recognizing what they are looking for
 - Evaluating intangibles such as “quality”
- Both are pretty bad at:
 - Mapping consistently between words and concepts

Process/System Co-Design

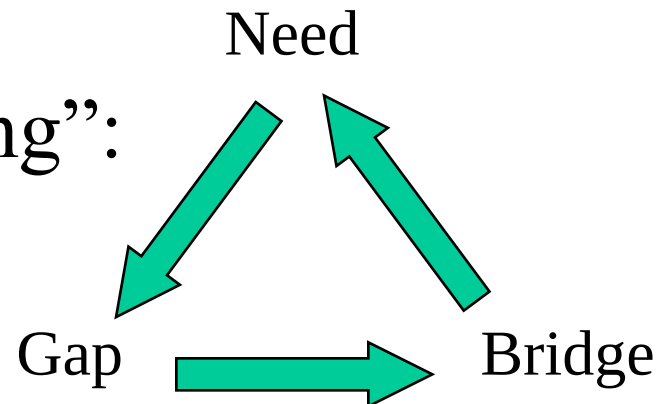


Taylor's Model of Question Formation



Iterative Search

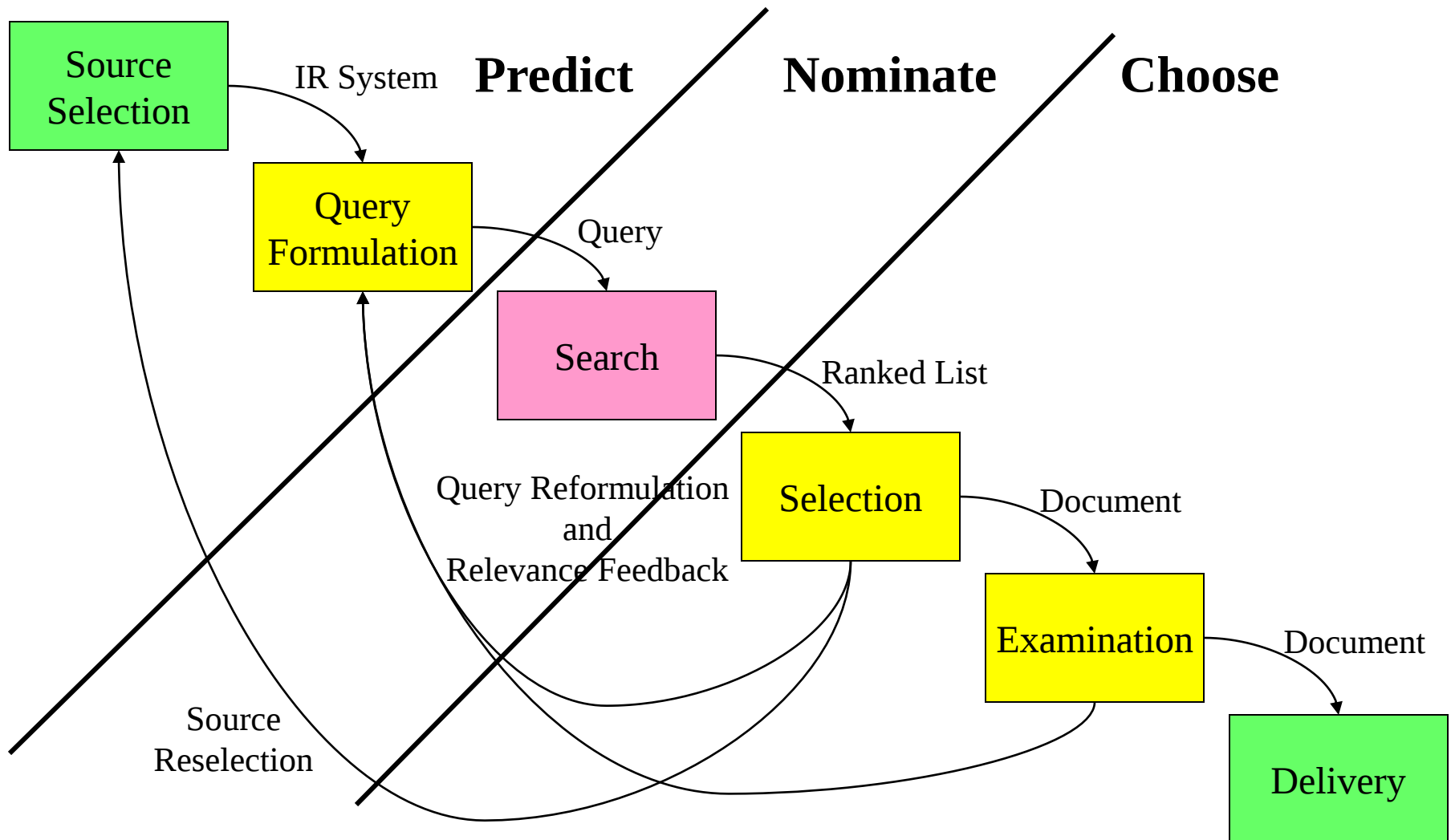
- Searchers often don't clearly understand
 - What actually happened
 - What evidence of that might exist
 - How that evidence might best be found
- The query results from a clarification process
- Dervin's "sense making":



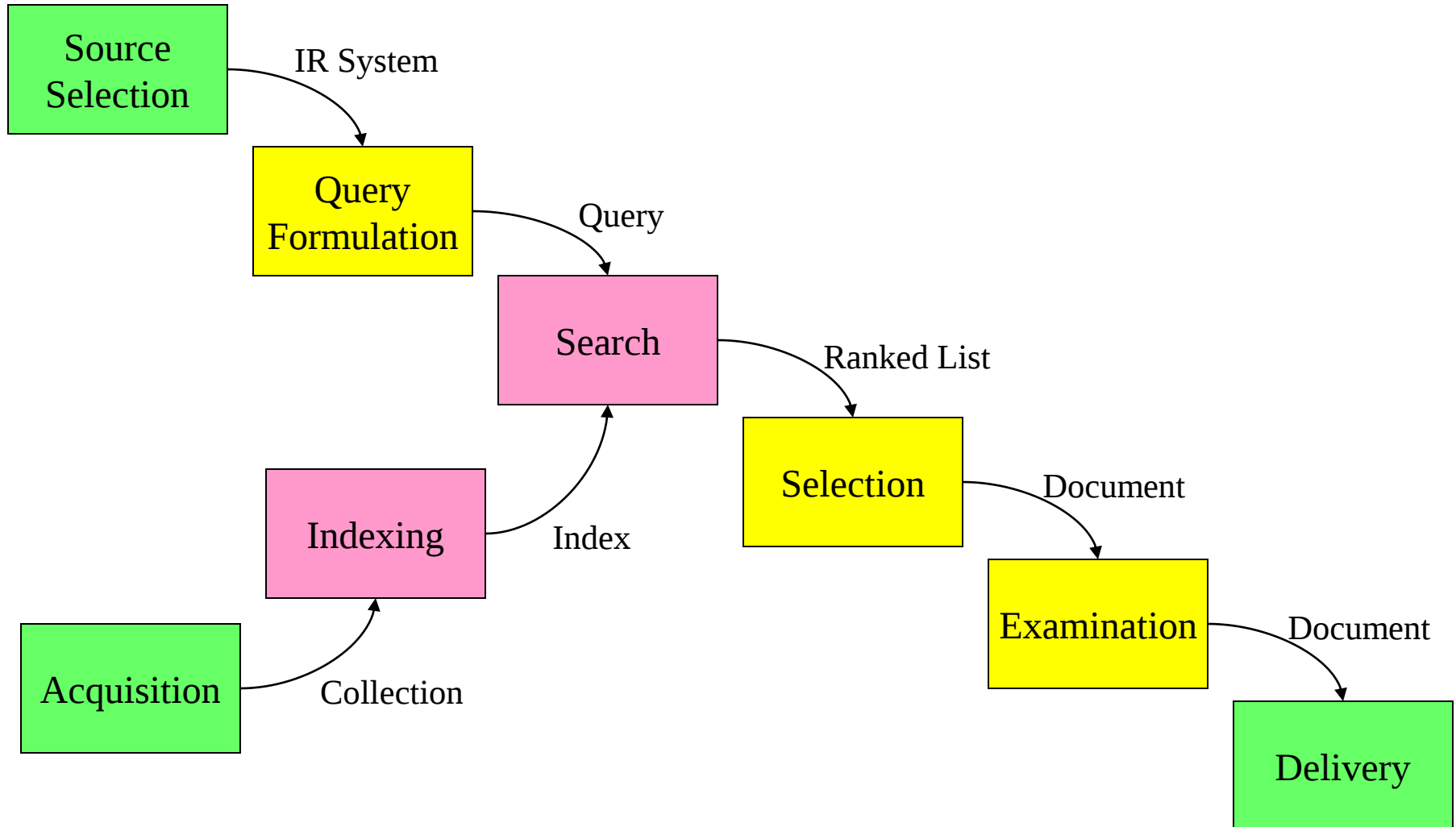
Divide and Conquer

- Strategy: use encapsulation to limit complexity
- Approach:
 - Define interfaces (input and output) for each component
 - Define the functions performed by each component
 - Build each component (in isolation)
 - See how well each component works
 - Then redefine interfaces to exploit strengths / cover weakness
 - See how well it all works together
 - Then refine the design to account for unanticipated interactions
- Result: a hierarchical decomposition

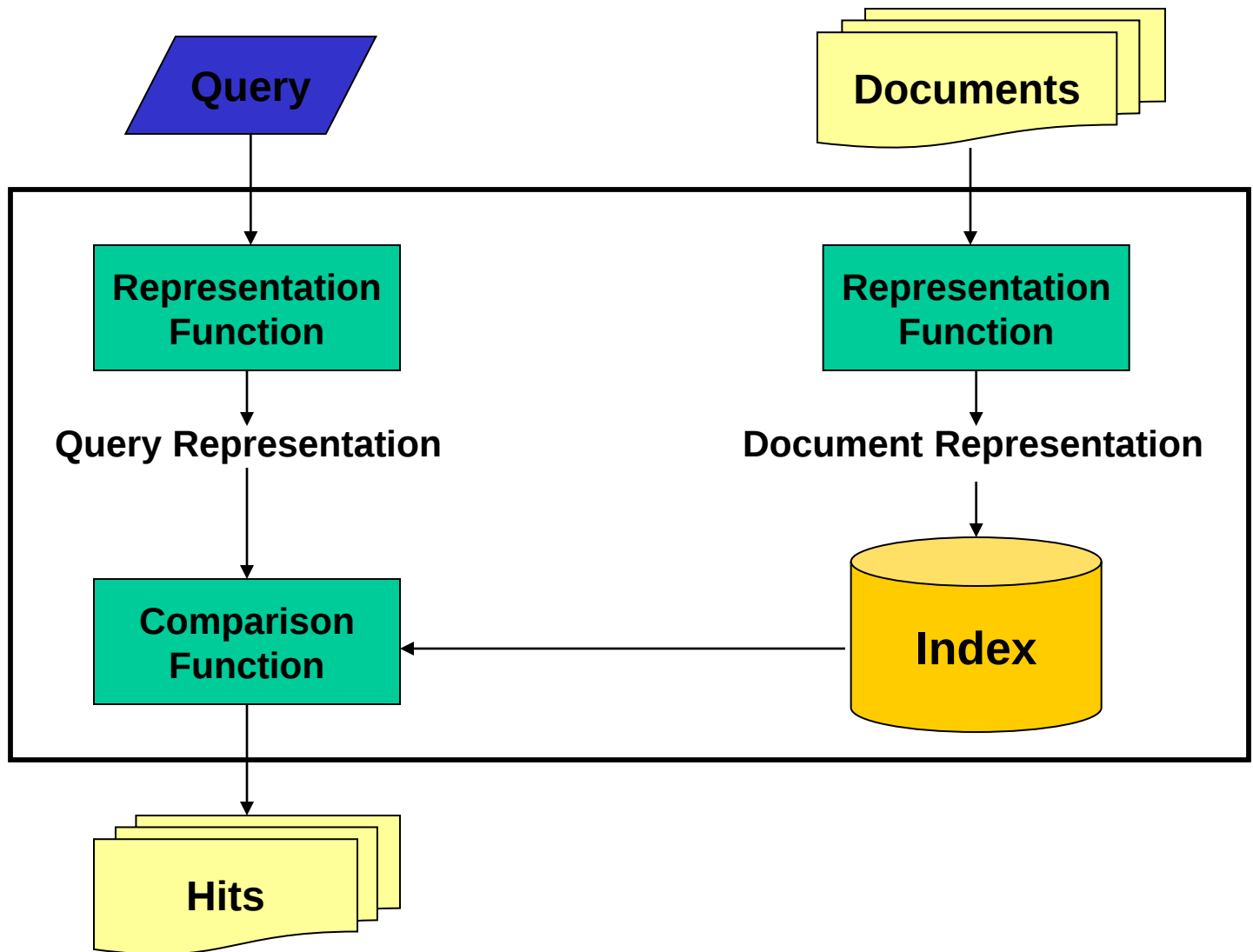
Supporting the Search Process



Supporting the Search Process



Inside The IR Black Box



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

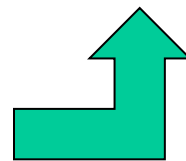
8 × new

6 × company, french, nutrition

5 × food, oil, percent, reduce,

taste, Tuesday

...



"Bag of Words"

Agenda

- Three generations of e-discovery
- Design thinking
 - Content-based search example
- Putting it all together

A “Term” is Whatever You Index

- Token
- Word
- Stem
- Character n-gram
- Phrase
- Named entity
- ...

ASCII

- Widely used in the U.S.
 - American Standard Code for Information Interchange
 - ANSI X3.4-1968



0 NUL	32 SPACE	64 @	96 `	
1 SOH	33 !	65 A	97 a	
2 STX	34 "	66 B	98 b	
3 ETX	35 #	67 C	99 c	
4 EOT	36 \$	68 D	100 d	
5 ENQ	37 %	69 E	101 e	
6 ACK	38 &	70 F	102 f	
7 BEL	39 '	71 G	103 g	
8 BS	40 (	72 H	104 h	
9 HT	41)	73 I	105 i	
10 LF	42 *	74 J	106 j	
11 VT	43 +	75 K	107 k	
12 FF	44 ,	76 L	108 l	
13 CR	45 -	77 M	109 m	
14 SO	46 .	78 N	110 n	
15 SI	47 /	79 O	111 o	
16 DLE	48 0	80 P	112 p	
17 DC1	49 1	81 Q	113 q	
18 DC2	50 2	82 R	114 r	
19 DC3	51 3	83 S	115 s	
20 DC4	52 4	84 T	116 t	
21 NAK	53 5	85 U	117 u	
22 SYN	54 6	86 V	118 v	
23 ETB	55 7	87 W	119 w	
24 CAN	56 8	88 X	120 x	
25 EM	57 9	89 Y	121 y	
26 SUB	58 :	90 Z	122 z	
27 ESC	59 ;	91 [	123 {	
28 FS	60 <	92 \	124	
29 GS	61 =	93]	125 }	
30 RS	62 >	94 ^	126 ~	
31 US	63 ?	95 _	127 DEL	

Unicode

- Single code for all the world's characters
 - ISO Standard 10646
- Separates “code space” from “encoding”
 - Code space extends ASCII (first 128 code points)
 - And Latin-1 (first 256 code points)
 - UTF-7 encoding will pass through email
 - Uses only the 64 printable ASCII characters
 - UTF-8 encoding is designed for disk file systems

Tokenization

- Words (from linguistics):
 - Morphemes are the units of meaning
 - Combined to make words
 - Anti (disestablishmentarian) ism
- Tokens (from Computer Science)
 - Doug 's running late !

Stemming

- Conflates words, usually preserving meaning
 - Rule-based suffix-stripping helps for English
 - {destroy, destroyed, destruction}: *destr*
 - Prefix-stripping is needed in some languages
 - Arabic: {alselam}: *selam* [Root: SLM (peace)]
- Imperfect: goal is to usually be helpful
 - Overstemming
 - {centennial, century, center}: *cent*
 - Understemming:
 - {acquire, acquiring, acquired}: *acquir*
 - {acquisition}: *acquis*

“Bag of Terms” Representation

- Bag = a “set” that can contain duplicates
 - “The quick brown fox jumped over the lazy dog’s back” →
{back, brown, dog, fox, jump, lazy, over, quick, the, the}
- Vector = values recorded in any consistent order
 - *{back, brown, dog, fox, jump, lazy, over, quick, the, the}* →
[1 1 1 1 1 1 1 1 2]

Bag of Terms Example

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

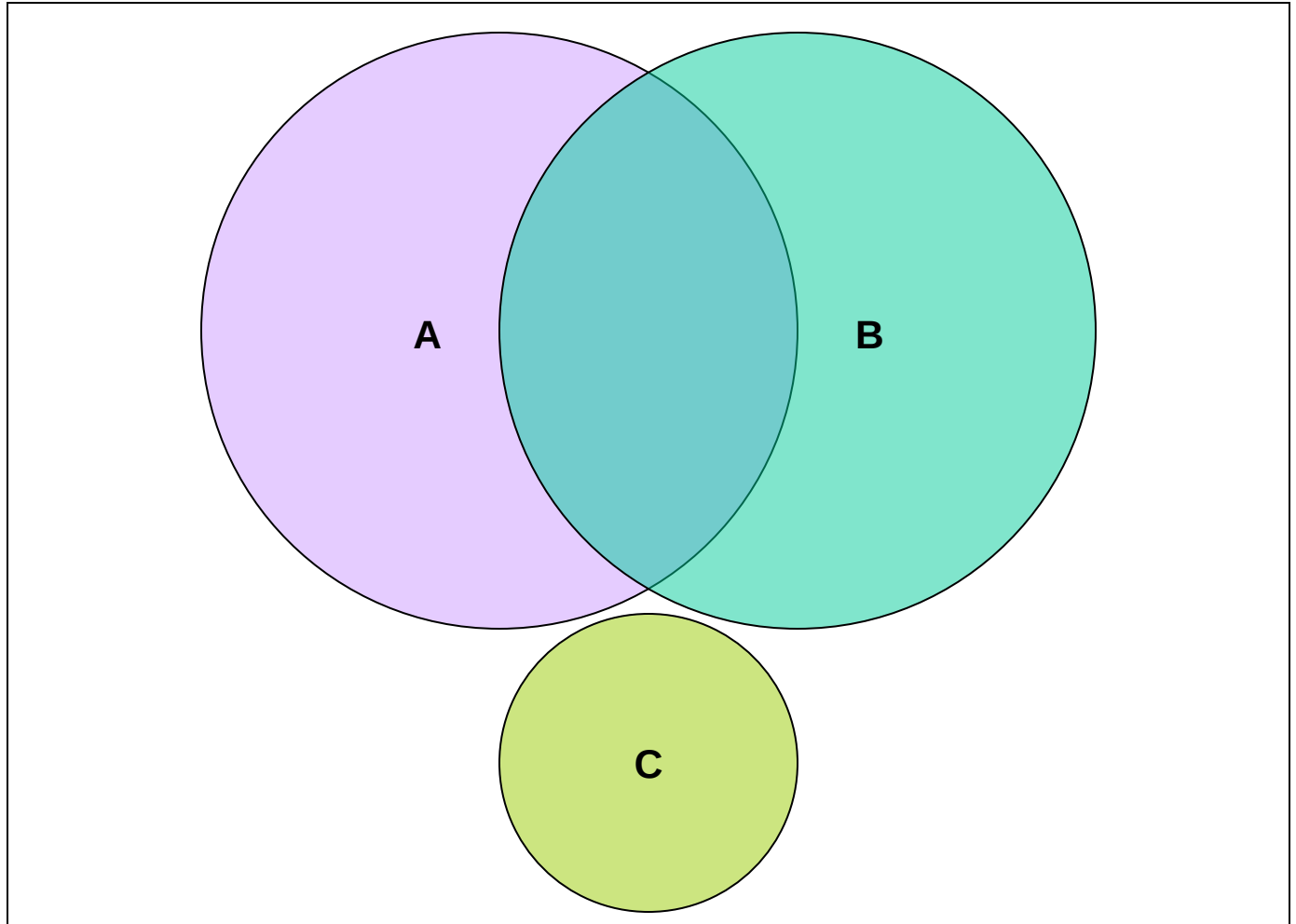
for
is
of
the
to

Boolean “Free Text” Retrieval

- Limit the bag of words to “absent” and “present”
 - “Boolean” values, represented as 0 and 1
- Represent terms as a “bag of documents”
 - Same representation, but rows rather than columns
- Combine the rows using “Boolean operators”
 - AND, OR, NOT
- Result set: every document with a 1 remaining

AND/OR/NOT

All documents



Boolean Operators

A OR B

A \ B	0	1
0	0	1
1	1	1

NOT B

B	0	1
	1	0

A AND B

A \ B	0	1
0	0	0
1	0	1

A NOT B
(= A AND NOT B)

A \ B	0	1
0	0	0
1	1	0

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
- NOT can discover alternate meanings
 - Democratic party

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
- Easy to implement, but less efficient
 - Store a list of positions for each word in each doc
 - Warning: stopwords become important!
 - Perform normal Boolean computations
 - Treat WITH and NEAR like AND with an extra constraint

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.

Ranked Retrieval

- Terms tell us about documents
 - If “rabbit” appears a lot, it may be about rabbits
- Documents tell us about terms
 - “the” is in every document -- not discriminating
- Documents are most likely described well by rare terms that occur in them frequently
 - Higher “term frequency” is stronger evidence
 - Low “document frequency” makes it stronger still

Ranking with BM-25 Term Weights

$$\sum_{e \in Q} \left[\log \frac{(N - df(e) + 0.5)}{(df(e) + 0.5)} \right] \left[\frac{(2.2 * tf(e, d_k))}{(0.3 + 0.9 * \frac{dl(d_k)}{avdl} + tf(e, d_k))} \frac{8 * qtf(e)}{7 + qtf(e)} \right]$$

document frequency

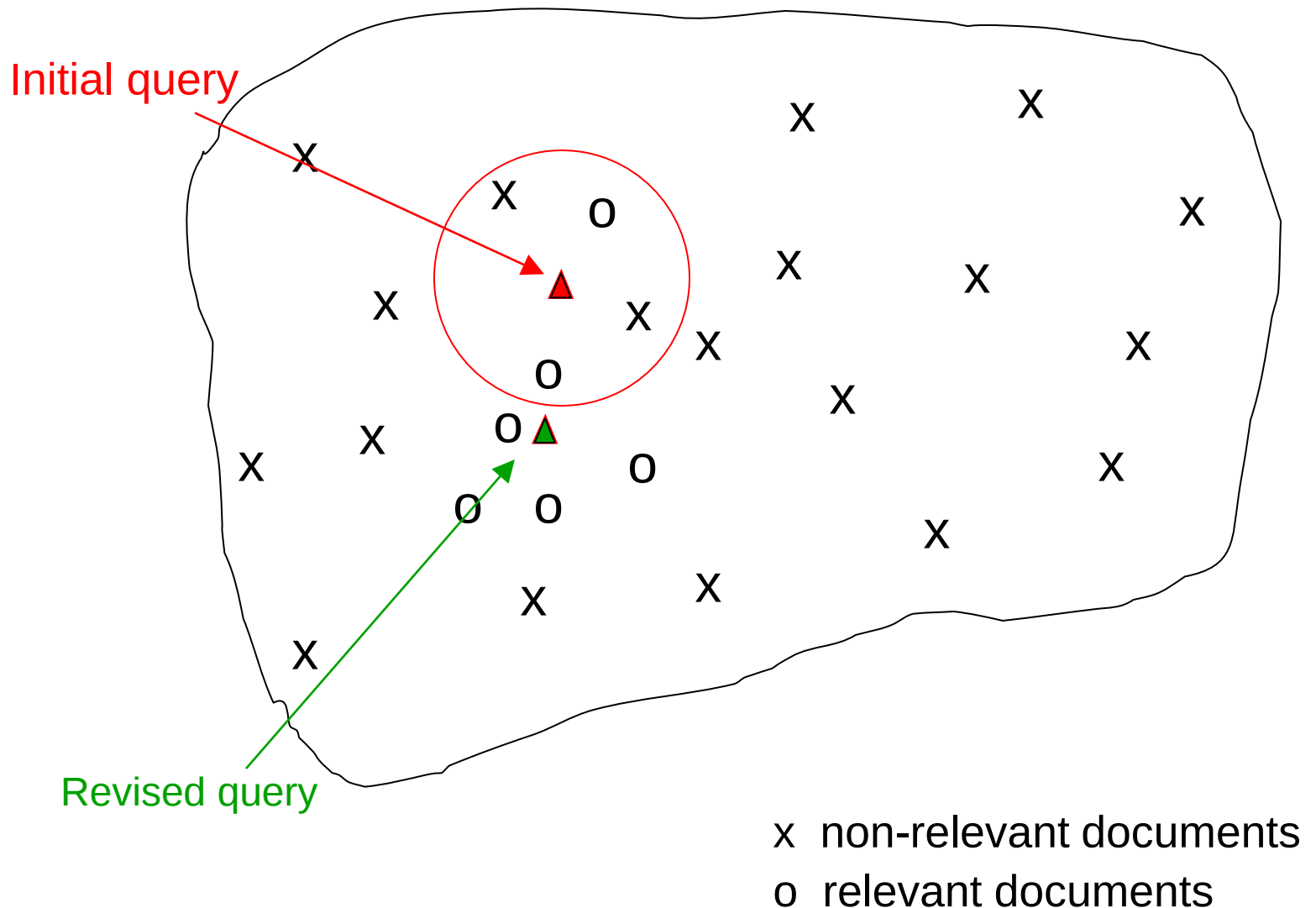
term frequency

document length

“Blind” Relevance Feedback

- Perform an initial search
- Identify new terms strongly associated with top results
 - Chi-squared
 - IDF
- Expand (and possibly reweight) the query

Visualizing Relevance Feedback



Problems with “Free Text” Search

- Homonymy
 - Terms may have many unrelated meanings
 - Polysemy (related meanings) is less of a problem
- Synonymy
 - Many ways of saying (nearly) the same thing
- Anaphora
 - Alternate ways of referring to the same thing

Machine-Assisted Indexing

- Goal: Automatically suggest descriptors
 - Better consistency with lower cost
- Approach: Rule-based expert system
 - Design thesaurus by hand in the usual way
 - Design an expert system to process text
 - String matching, proximity operators, ...
 - Write rules for each thesaurus/collection/language
 - Try it out and fine tune the rules by hand

Machine-Assisted Indexing Example

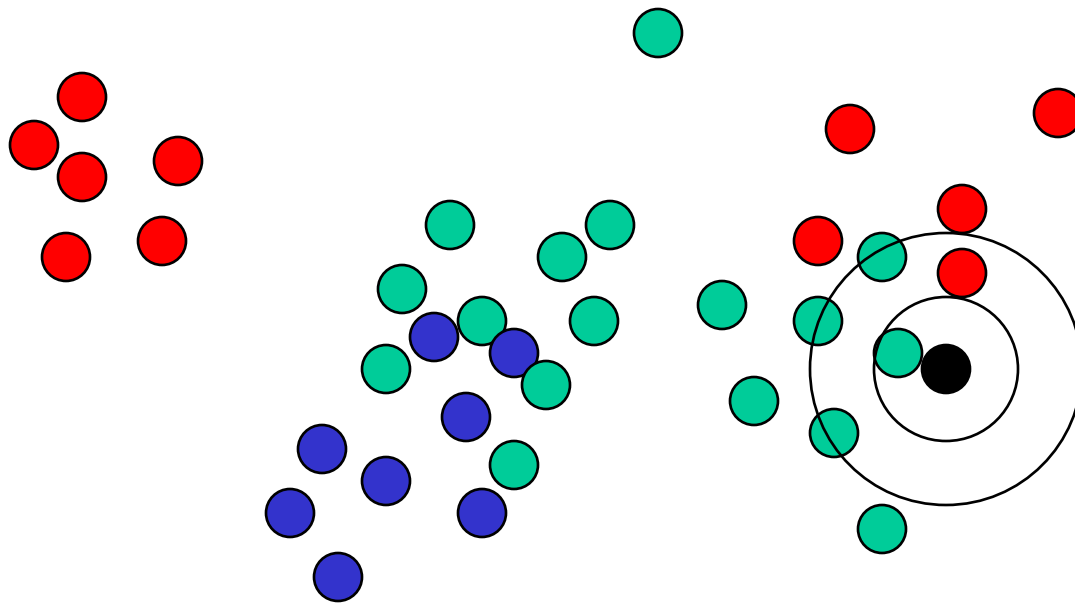
Access Innovations system:

```
//TEXT: science
IF (all caps)
    USE research policy
    USE community program
ENDIF
IF (near "Technology" AND with "Development")
    USE community development
    USE development aid
ENDIF
```

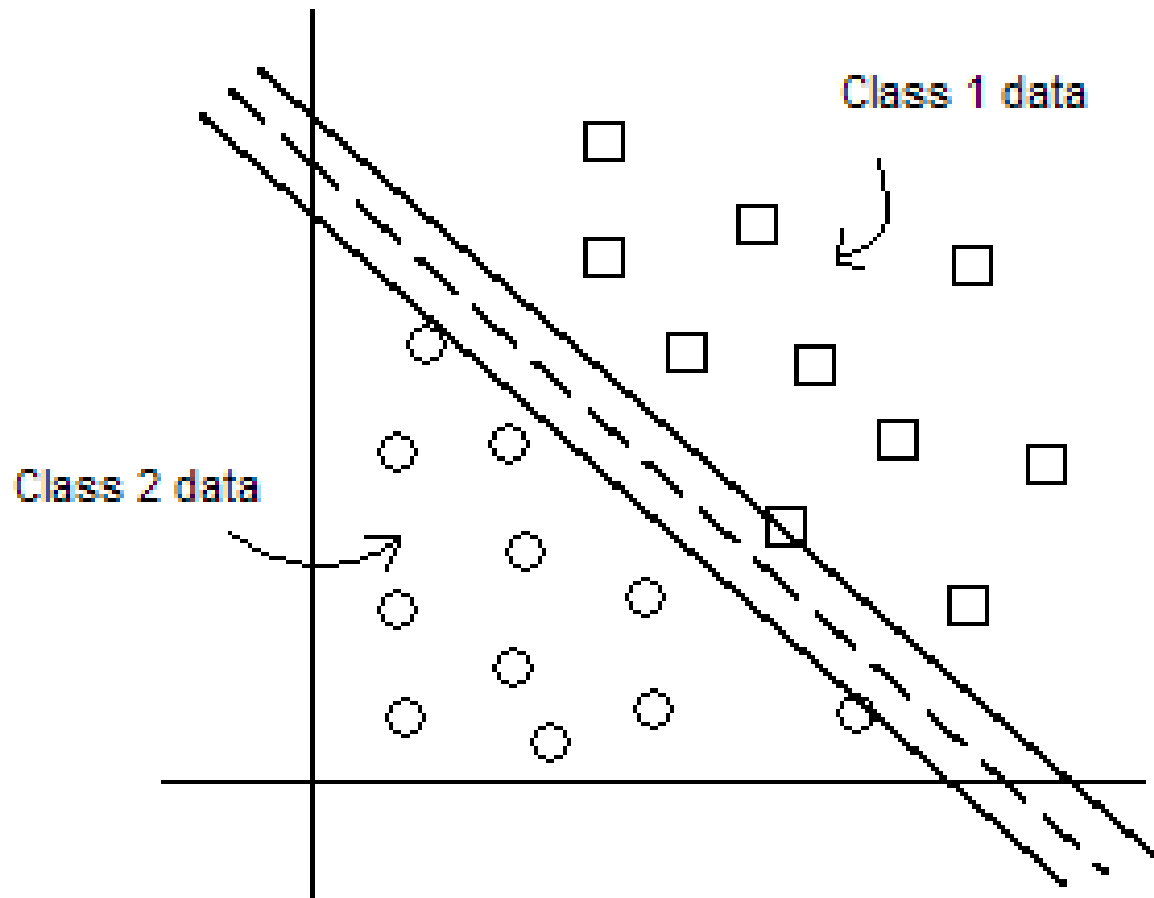
near: within 250 words

with: in the same sentence

Machine Learning: kNN Classifier



Support Vector Machine (SVM)



“Named Entity” Tagging

- Machine learning techniques can find:
 - Location
 - Extent
 - Type
- Two types of features are useful
 - Orthography
 - e.g., Paired or non-initial capitalization
 - Trigger words
 - e.g., Mr., Professor, said, ...

Normalization

- Variant forms of names (“name authority”)
 - Pseudonyms, partial names, citation styles
- Acronyms and abbreviations
- Co-reference resolution
 - References to roles, objects, names
 - Anaphoric pronouns
- Entity Linking

Entity Linking

0.47

فوفقاً لأرقام مضابط البرلمان الم
زيارة إسرائيلية الصنع المضادة
للدبابات وأكدت المصادر أن
الصفقة كانت أحد الموضوعات
الرئيسية على مائدة المفاوضات
بين شارون ونظيره البريطاني
توني بلير.
كانت وزارة الدفاع البريطانية
قد أبرمت صفقة مبدئية مع
الحكومة الإسرائيلية في عام
2001 اشترت بمقتضاها عدداً

0.62

WIKIPEDIA

The Free Encyclopedia

Main page

Contents

Featured content

Current events

Random article

Donate to Wikipedia

Interaction

Help

About Wikipedia

Community portal

Recent changes

Contact Wikipedia

Toolbox

General report

Languages

Afrikaans

العربية

Asturianu

Azərbaycanca

Бân-lâm-gú

Беларуская

Беларуская (тарашкевіца)

Bosanski

Brezhoneg

Буряадска

Català

Česky

Cymraeg

Dansk

Deutsch

Eesti

Tony Blair

From Wikipedia, the free encyclopedia

For other uses, see Tony Blair (disambiguation).

Anthony Charles Lynton Blair

(born 6 May 1953)^[1] is a former British Labour Party politician who served as the Prime Minister of the United Kingdom from 2 May 1997 to 27 June 2007. He was the Member of Parliament (MP) for Sedgefield from 1983 to 2007 and Leader of the Labour Party from 1994 to 2007. He resigned from all of these positions in June 2007.

Tony Blair

Tony Blair was elected Leader of the Labour Party in the leadership election of July 1994, following the sudden death of his predecessor, John Smith. Under his leadership, the party adopted the term "New Labour"^[2] and moved away from its traditional left wing position towards the centre ground.^{[3][4]} Blair subsequently led Labour to a landslide victory in the 1997 general election. At 43 years old, he became the youngest Prime Minister since Lord Liverpool in 1812. In the first years of the New Labour government, Blair's government implemented a number of 1997 manifesto pledges, introducing the minimum wage, Human Rights Act and Freedom of Information Act, and carrying out regional devolution, establishing the Scottish Parliament, the National Assembly for Wales, and the Northern Ireland Assembly.

Blair's role as Prime Minister was particularly visible in foreign and security policy, including in Northern Ireland, where he was involved in the 1998 Good Friday Agreement. From the start of the War on Terror in 2001, Blair strongly

The Right Honourable

Tony Blair



Blair at the World Economic Forum in Davos, Switzerland (29 January 2008)

Prime Minister of the United Kingdom

In office

2 May 1997 – 27 June 2007

Monarch Elizabeth II

Deputy John Prescott

Preceded by John Major

Succeeded by Gordon Brown

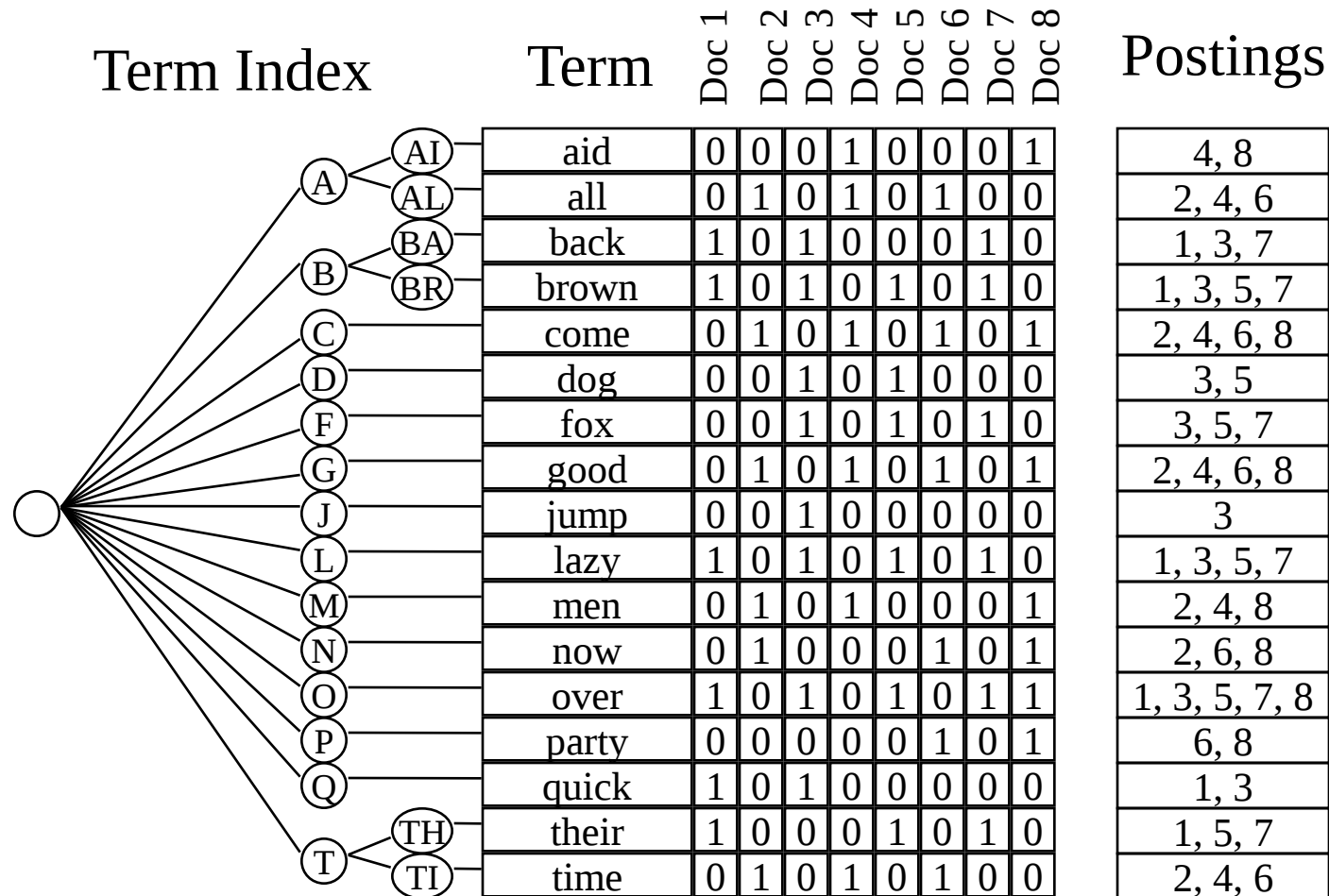
Leader of the Opposition

In office

Desirable Index Characteristics

- Very rapid search
 - Less than $\sim 100\text{ms}$ is typically impercievable
- Reasonable hardware requirements
 - Processor speed, disk size, main memory size
- “Fast enough” creation

An “Inverted Index”

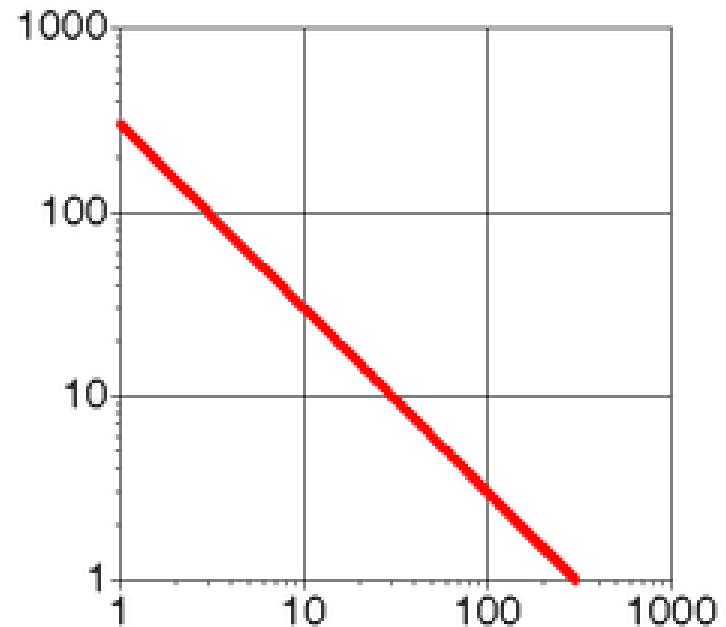
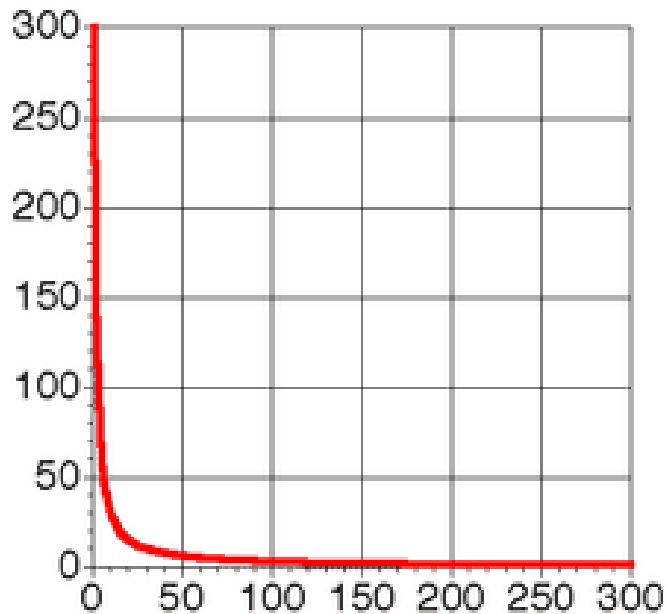


Word Frequency in English

Frequency of 50 most common words in English
(sample of 19 million words)

the	1130021	from	96900	or	54958
of	547311	he	94585	about	53713
to	516635	million	93515	market	52110
a	464736	year	90104	they	51359
in	390819	its	86774	this	50933
and	387703	be	85588	would	50828
that	204351	was	83398	you	49281
for	199340	company	83070	which	48273
is	152483	an	76974	bank	47940
said	148302	has	74405	stock	47401
it	134323	are	74097	trade	47310
on	121173	have	73132	his	47116
by	118863	but	71887	more	46244
as	109135	will	71494	who	42142
at	101779	say	66807	one	41635
mr	101679	new	64456	their	40910
with	101210	share	63925		

Zipfian Distribution: The “Long Tail”

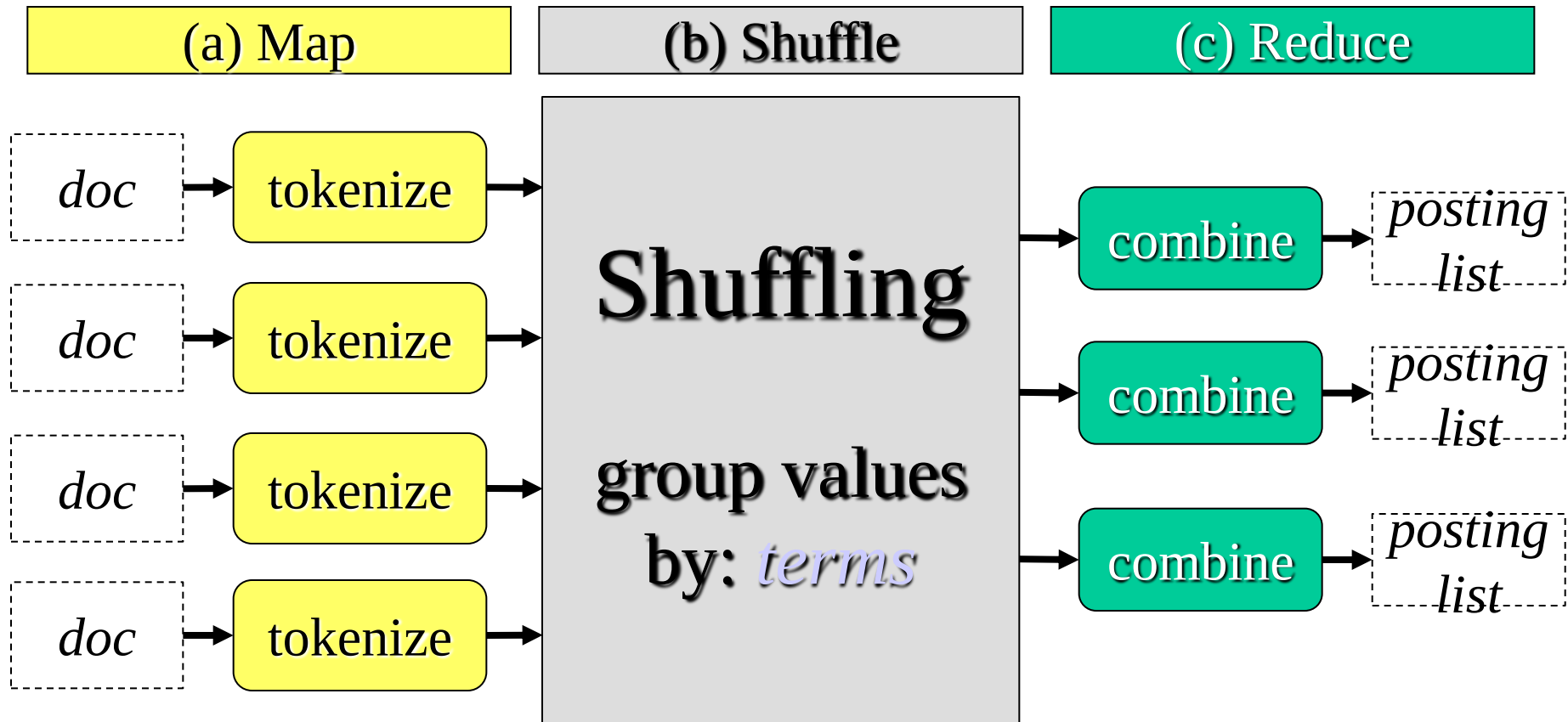


- A few elements occur very frequently
- Many elements occur very infrequently

Index Compression

- CPU's are much faster than disks
 - A disk can transfer 1,000 bytes in ~20 ms
 - The CPU can do ~10 million instructions in that time
- Compressing the postings file is a big win
 - Trade decompression time for fewer disk reads
- Key idea: reduce redundancy
 - Trick 1: store relative offsets (some will be the same)
 - Trick 2: use an optimal coding scheme

MapReduce Indexing



Agenda

- Three generations of e-discovery
- Design thinking
- Content-based search example
- Putting it all together

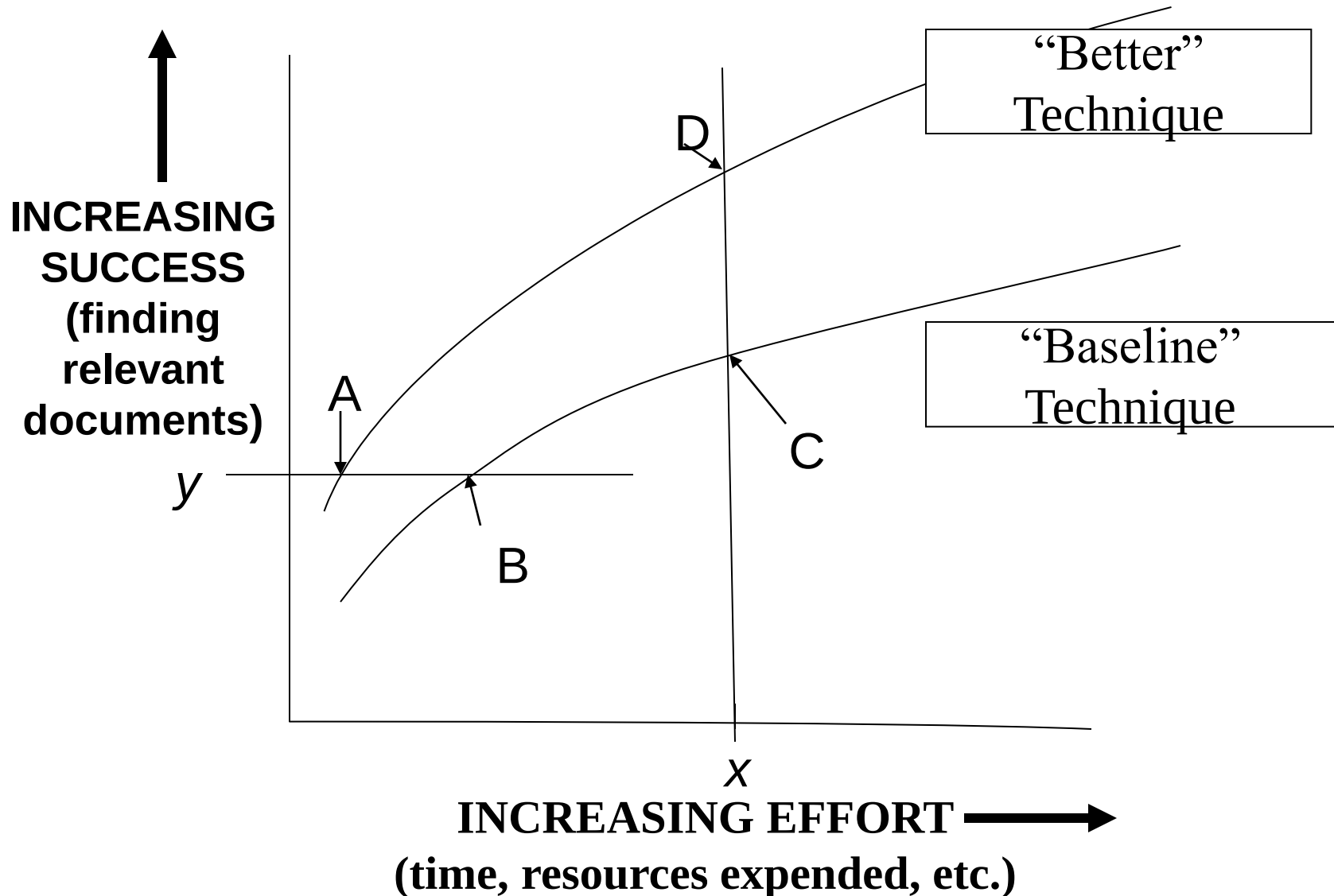
Indexable Features

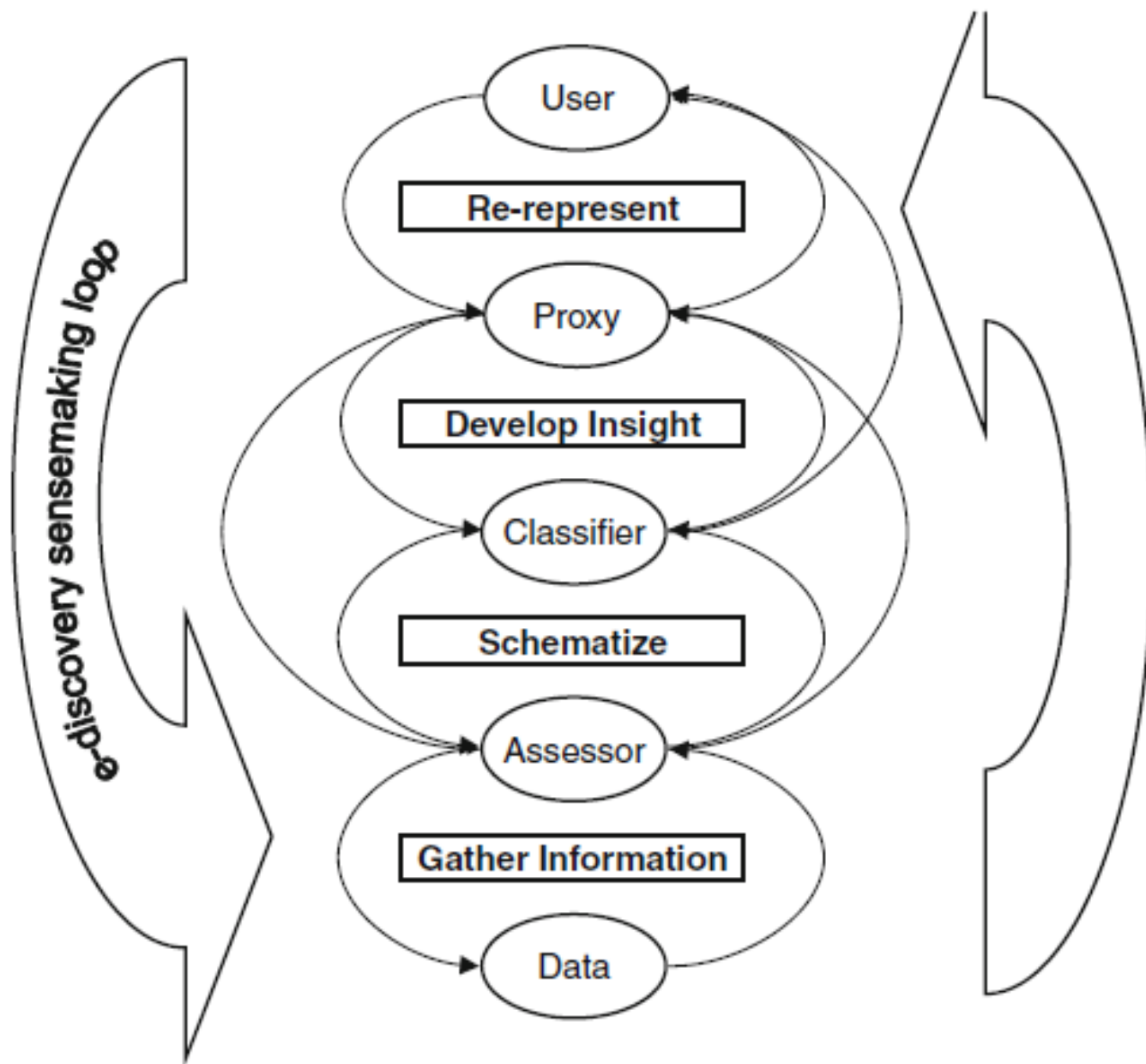
- Content
 - Stems, named entities, ...
- Context
 - Sender, time, ...
- Description
 - Subject line, anchor text, ...
- Behavior
 - Most recent access time, incoming links, ...

Technology-Assisted Review

- Understand the task
 - Analyze and clarify the production request
- Find a sufficient set of seed documents
 - Adequate diversity, adequate specificity
- Iteratively improve the classifier
 - Judge samples for training and for evaluation
- Stop when benefit exceeds cost

What Does “Better” Mean?





Responsiveness vs. Privilege

- Very large review set
 - Topical
 - False positive risks
harmful disclosure
- Much smaller review set
 - Non-topical
 - False negative risks
harmful disclosure
 - Last chance to catch errors!