Re-evaluation of IR Systems

Yaşar Tonta

Hacettepe Üniversitesi

tonta@hacettepe.edu.tr

yunus.hacettepe.edu.tr/~tonta/

DOK324/BBY220 Bilgi Erişim İlkeleri

Note: Slides are taken from Prof. Ray Larson's web site (www.sims.berkeley.edu/~ray/

Evaluation of IR Systems

- Precision vs. Recall
- Cutoff Points
- Test Collections/TREC
- Blair & Maron Study

Evaluation

- Why Evaluate?
- What to Evaluate?
- How to Evaluate?

Why Evaluate?

- Determine if the system is desirable
- Make comparative assessments
- Others?

What to Evaluate?

- How much of the information need is satisfied.
- How much was learned about a topic.
- Incidental learning:
  - How much was learned about the collection.
  - How much was learned about other topics.
- How inviting the system is.

Relevance

- In what ways can a document be relevant to a query?
  - Answer precise question precisely.
  - Partially answer question.
  - Suggest a source for more information.
  - Give background information.
  - Remind the user of other knowledge.
  - Others ...

Relevance

- How relevant is the document
  - for this user for this information need.
- Subjective, but
- Measurable to some extent
  - How often do people agree a document is relevant to a query
- How well does it answer the question?
  - Complete answer? Partial?
  - Background Information?
  - Hints for further exploration?

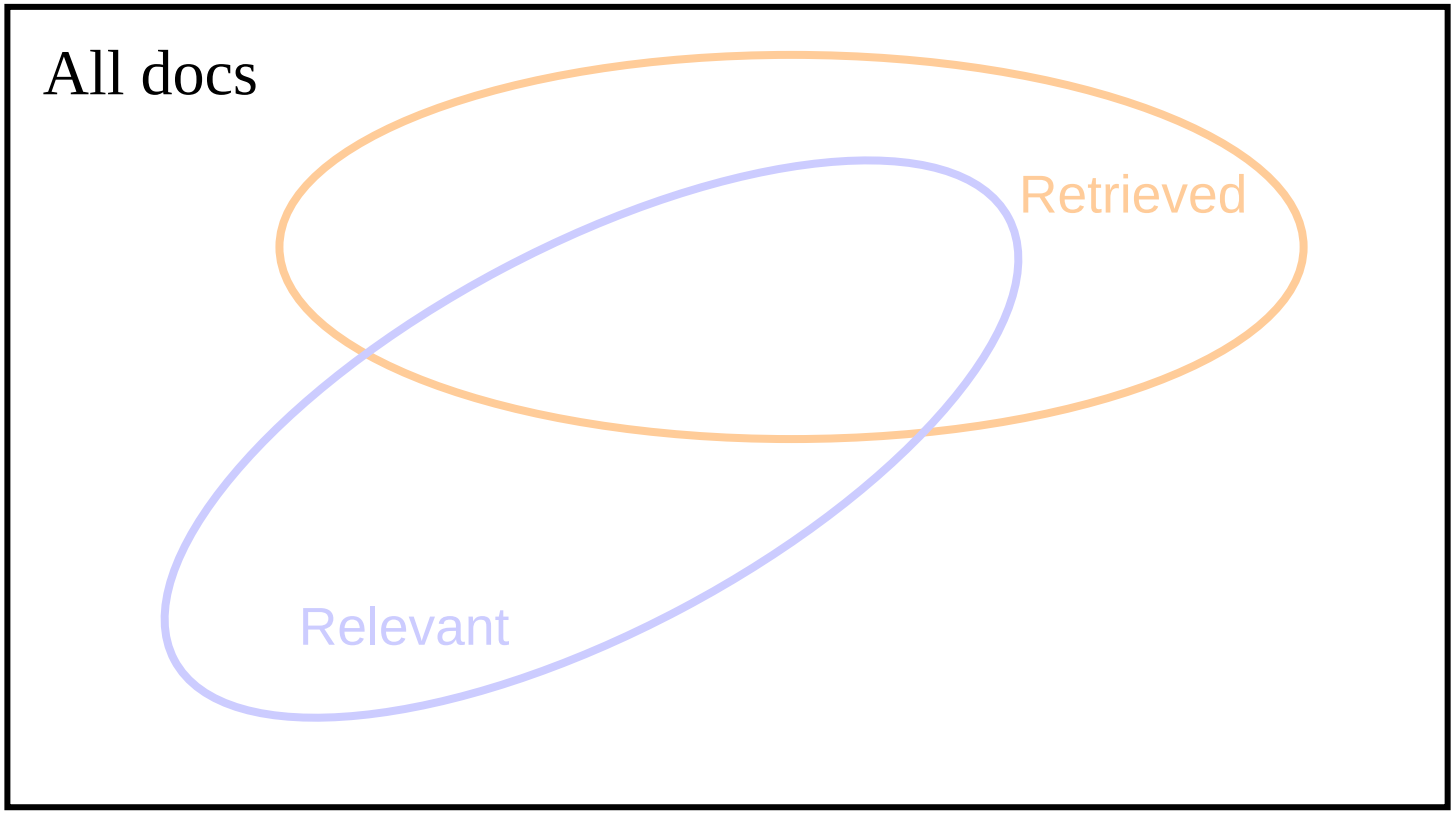
What to Evaluate?

What can be measured that reflects users' ability to use system? (Cleverdon 66)

- Coverage of Information
- Form of Presentation
- Effort required/Ease of Use
- Time and Space Efficiency
- Recall
  - proportion of relevant material actually retrieved
- Precision
  - proportion of retrieved material actually relevant

effectiveness

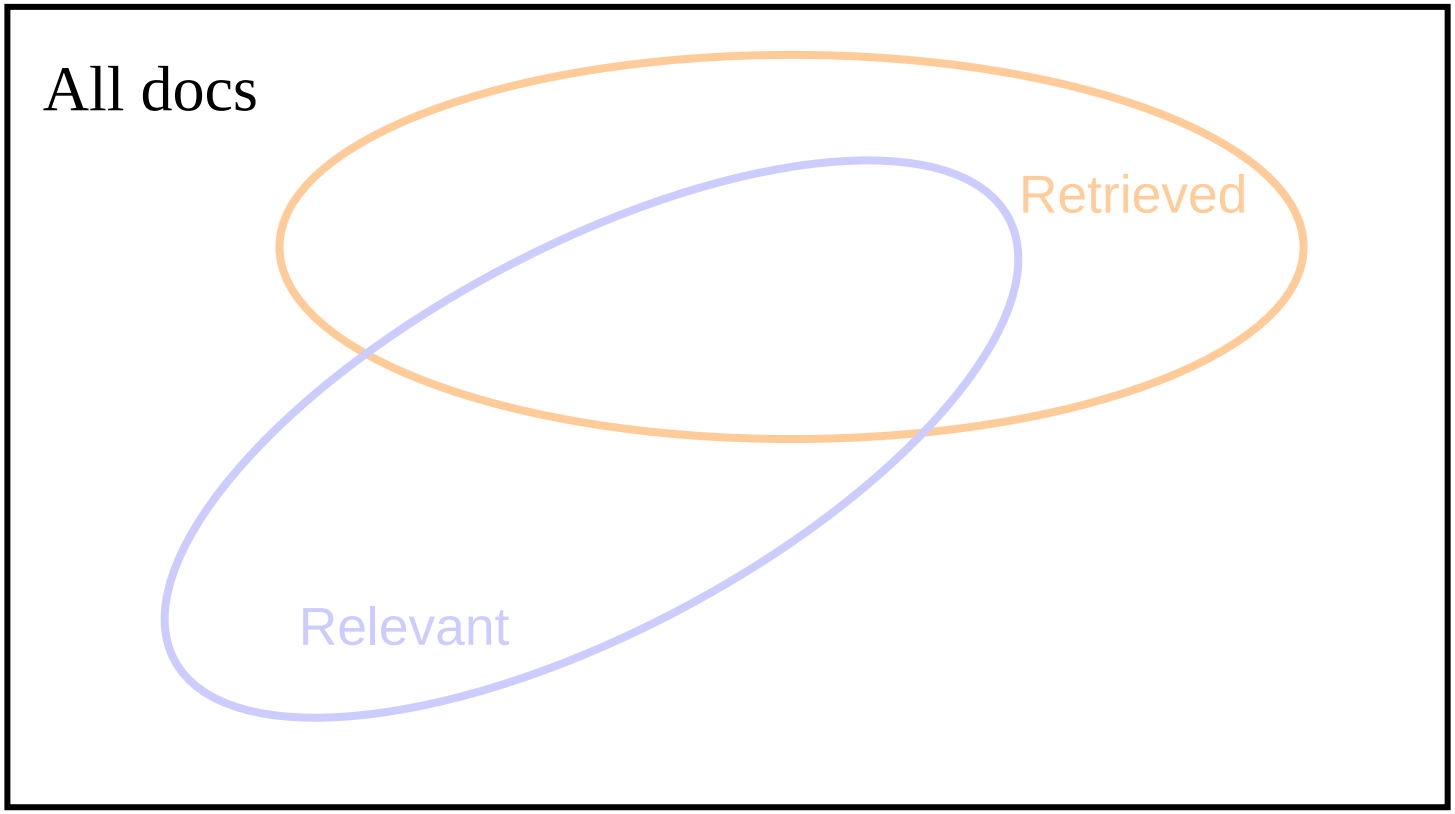
Relevant vs. Retrieved



Precision vs. Recall

$$\text{Precision} = \frac{|\,\text{RelRetrieved}\,|}{|\,\text{Retrieved}\,|}$$

$$\text{Recall} = \frac{|\,\text{RelRetrieved}\,|}{|\,\text{Rel in Collection}\,|}$$

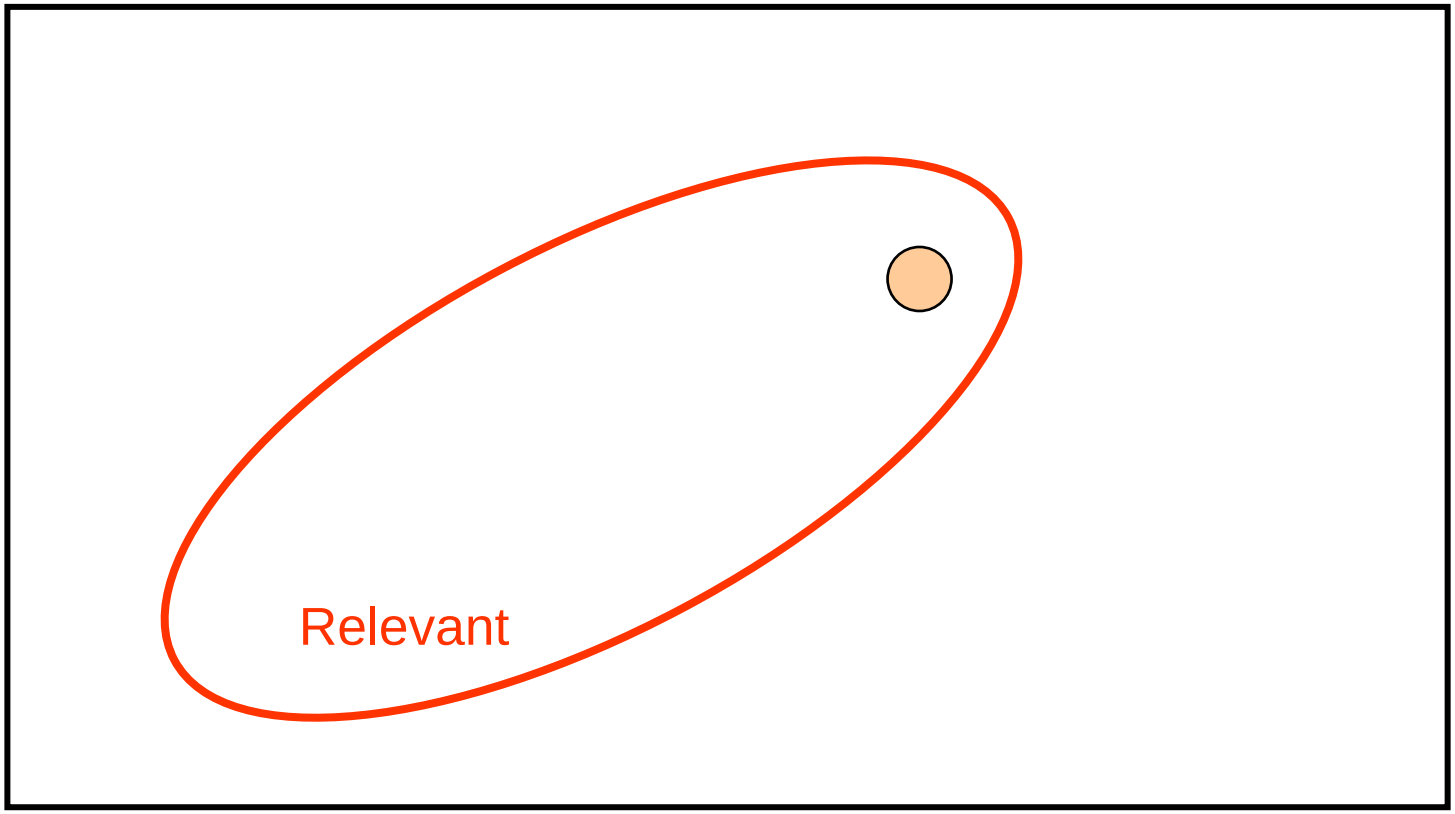


Why Precision and Recall?

Get as much good stuff while at the same time getting as little junk as possible.

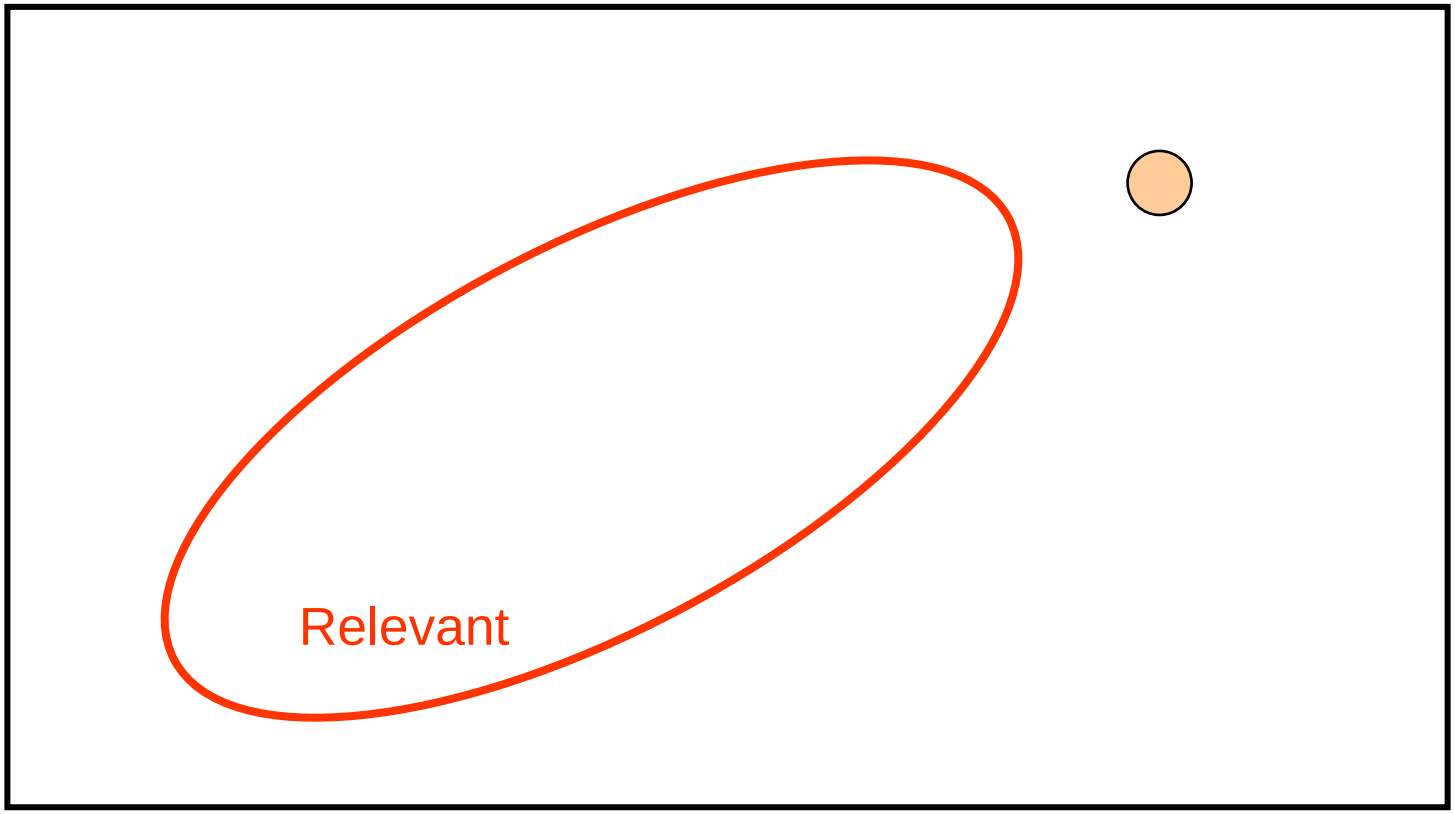
Retrieved vs. Relevant Documents

Very high precision, very low recall



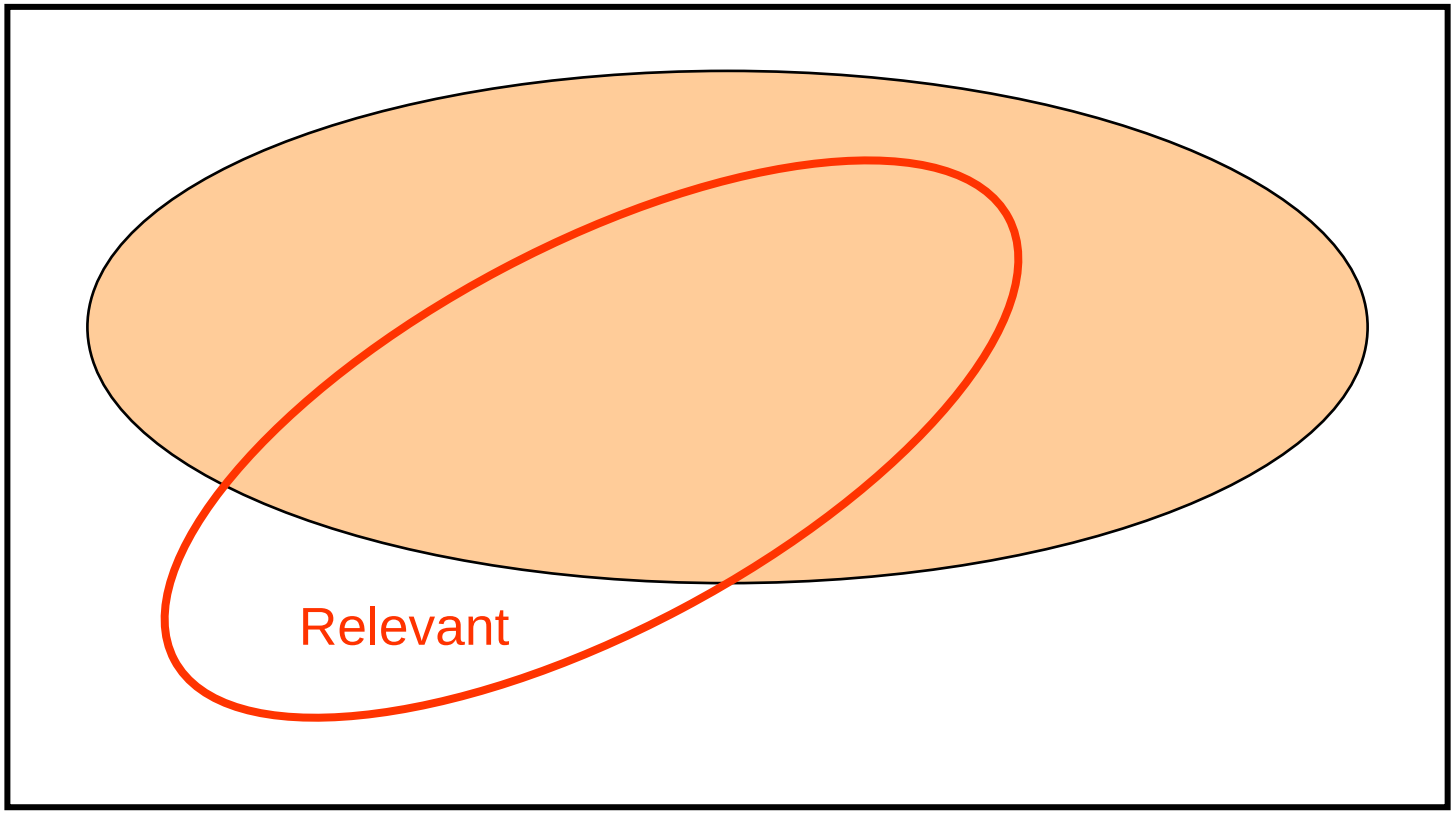
Retrieved vs. Relevant Documents

Very low precision, very low recall (0 in fact)



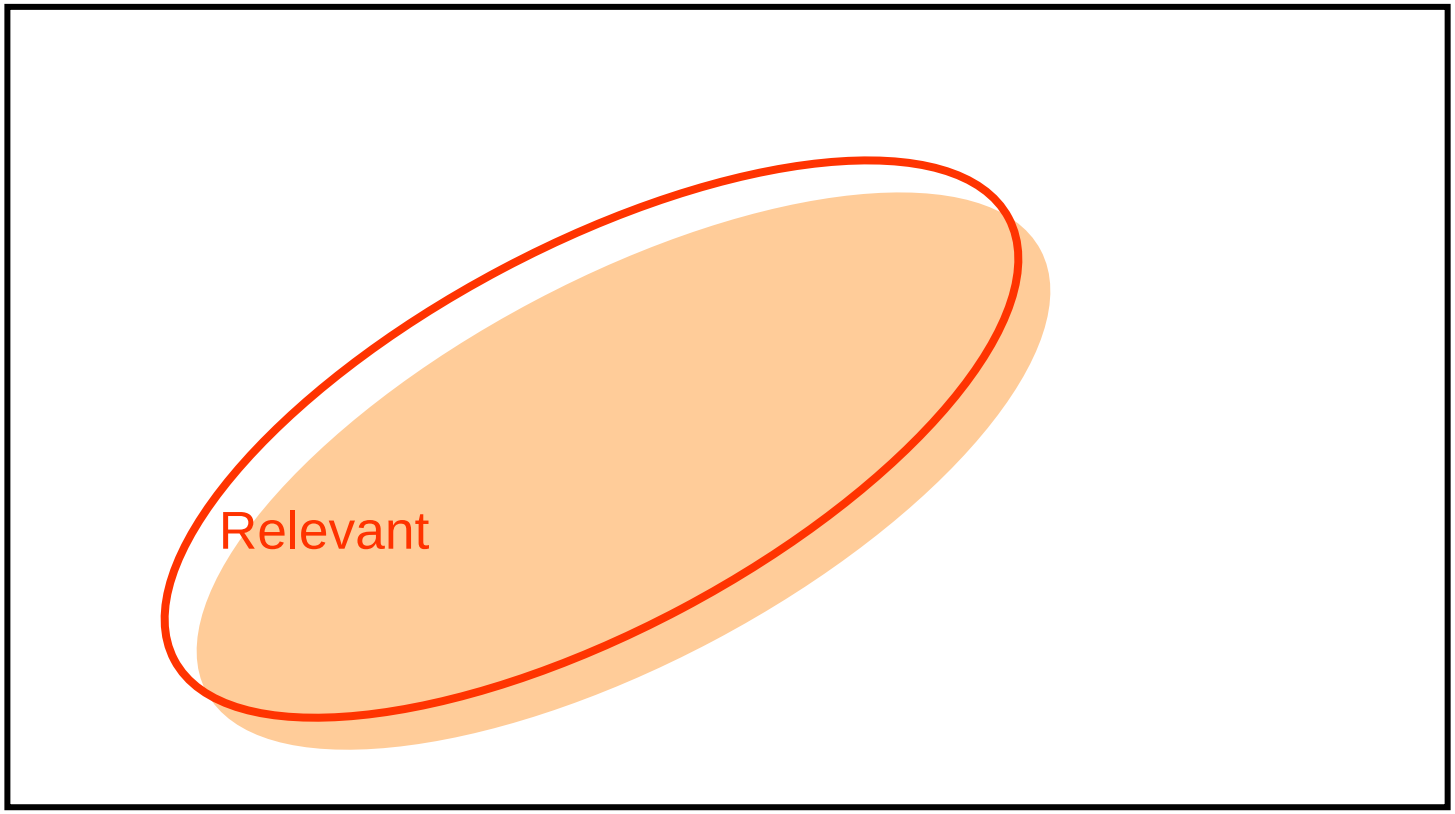
Retrieved vs. Relevant Documents

High recall, but low precision



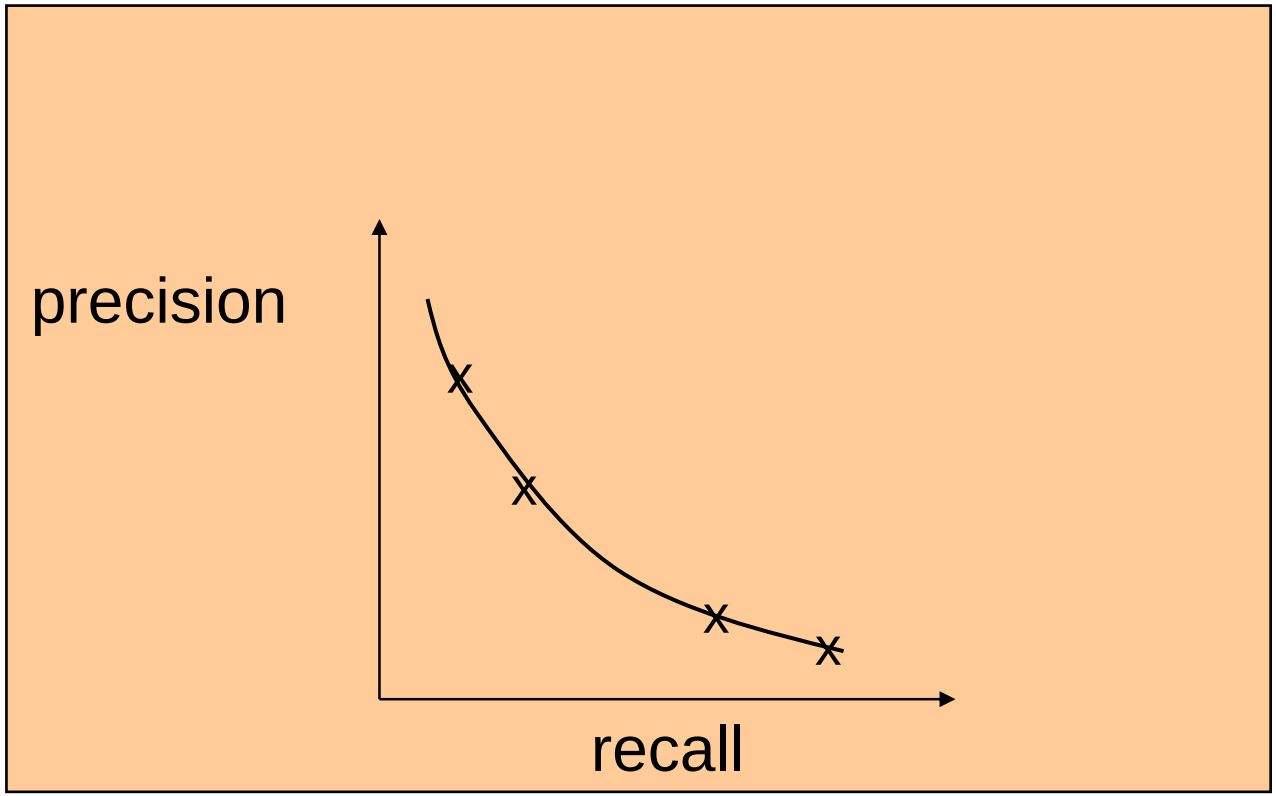
Retrieved vs. Relevant Documents

High precision, high recall (at last!)



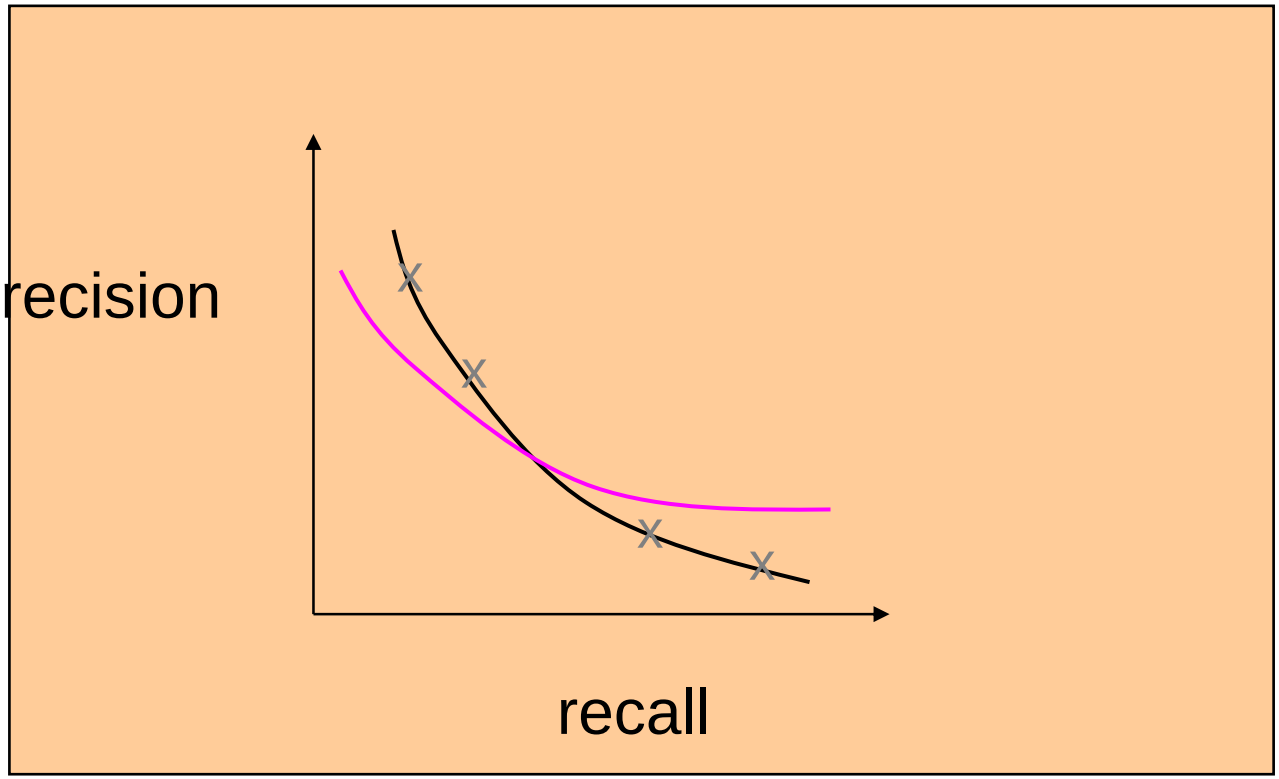
Precision/Recall Curves

- There is a tradeoff between Precision and Recall
- So measure Precision at different levels of Recall
- Note: this is an AVERAGE over MANY queries

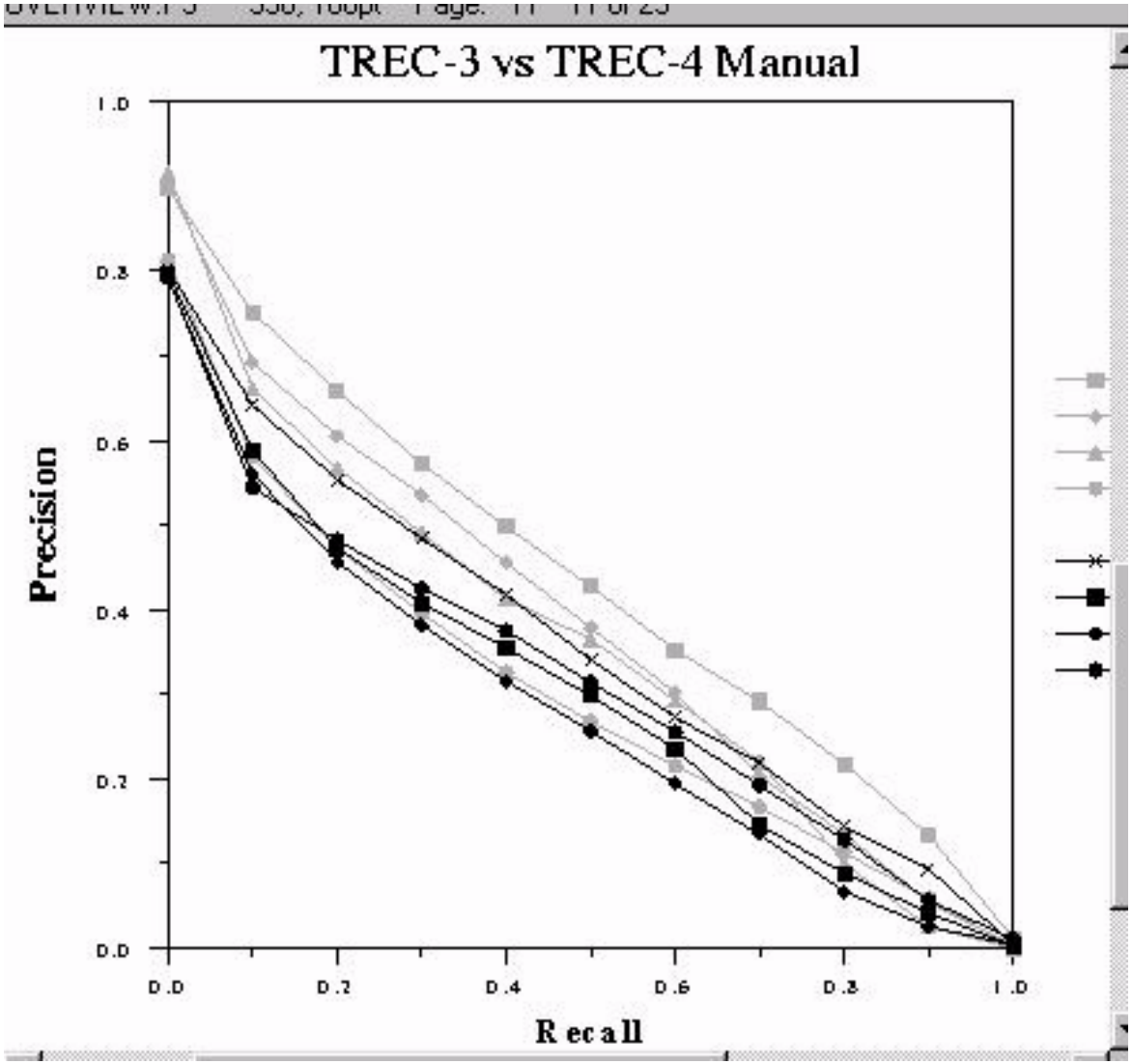


Precision/Recall Curves

- Difficult to determine which of these two hypothetical results is better:



Precision/Recall Curves



Document Cutoff Levels

- Another way to evaluate:
  - Fix the number of documents retrieved at several levels:
    - top 5
    - top 10
    - top 20
    - top 50
    - top 100
    - top 500
  - Measure precision at each of these levels
  - Take (weighted) average over results
- This is a way to focus on how well the system ranks the first k documents.

Problems with Precision/Recall

- Can’t know true recall value
  - except in small collections
- Precision/Recall are related
  - A combined measure sometimes more appropriate
- Assumes batch mode
  - Interactive IR is important and has different criteria for successful searches
  - We will touch on this in the UI section
- Assumes a strict rank ordering matters.

Relation to Contingency Table

	Doc is Relevant	Doc is NOT relevant
Doc is retrieved	a	b
Doc is NOT retrieved	c	d

- Accuracy: (a+d) / (a+b+c+d)
- Precision: a/(a+b)
- Recall: ?
- Why don't we use Accuracy for IR?
  - (Assuming a large collection)
  - Most docs aren't relevant
  - Most docs aren't retrieved
  - Inflates the accuracy value

The E-Measure

Combine Precision and Recall into one number

(van Rijsbergen 79)

$$E = 1 - \frac{1 + b^2}{\frac{b^2}{R} + \frac{1}{P}}$$

$$E = 1 - \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\frac{1}{R}}$$

$$\alpha = 1/(\beta^2 + 1)$$

P = precision

R = recall

b = measure of relative importance of P or R

For example,

b = 0.5 means user is twice as interested in precision as recall

How to Evaluate?
Test Collections

TREC

- Text REtrieval Conference/Competition
  - Run by NIST (National Institute of Standards & Technology)
  - 2001 was the 10th year - 11th TREC in November
- Collection: 5 Gigabytes (5 CRDOMs), >1.5 Million Docs
  - Newswire & full text news (AP, WSJ, Ziff, FT, San Jose Mercury, LA Times)
  - Government documents (federal register, Congressional Record)
  - FBIS (Foreign Broadcast Information Service)
  - US Patents

TREC (cont.)

- Queries + Relevance Judgments
  - Queries devised and judged by “Information Specialists”
  - Relevance judgments done only for those documents retrieved -- not entire collection!
- Competition
  - Various research and commercial groups compete (TREC 6 had 51, TREC 7 had 56, TREC 8 had 66)
  - Results judged on precision and recall, going up to a recall level of 1000 documents

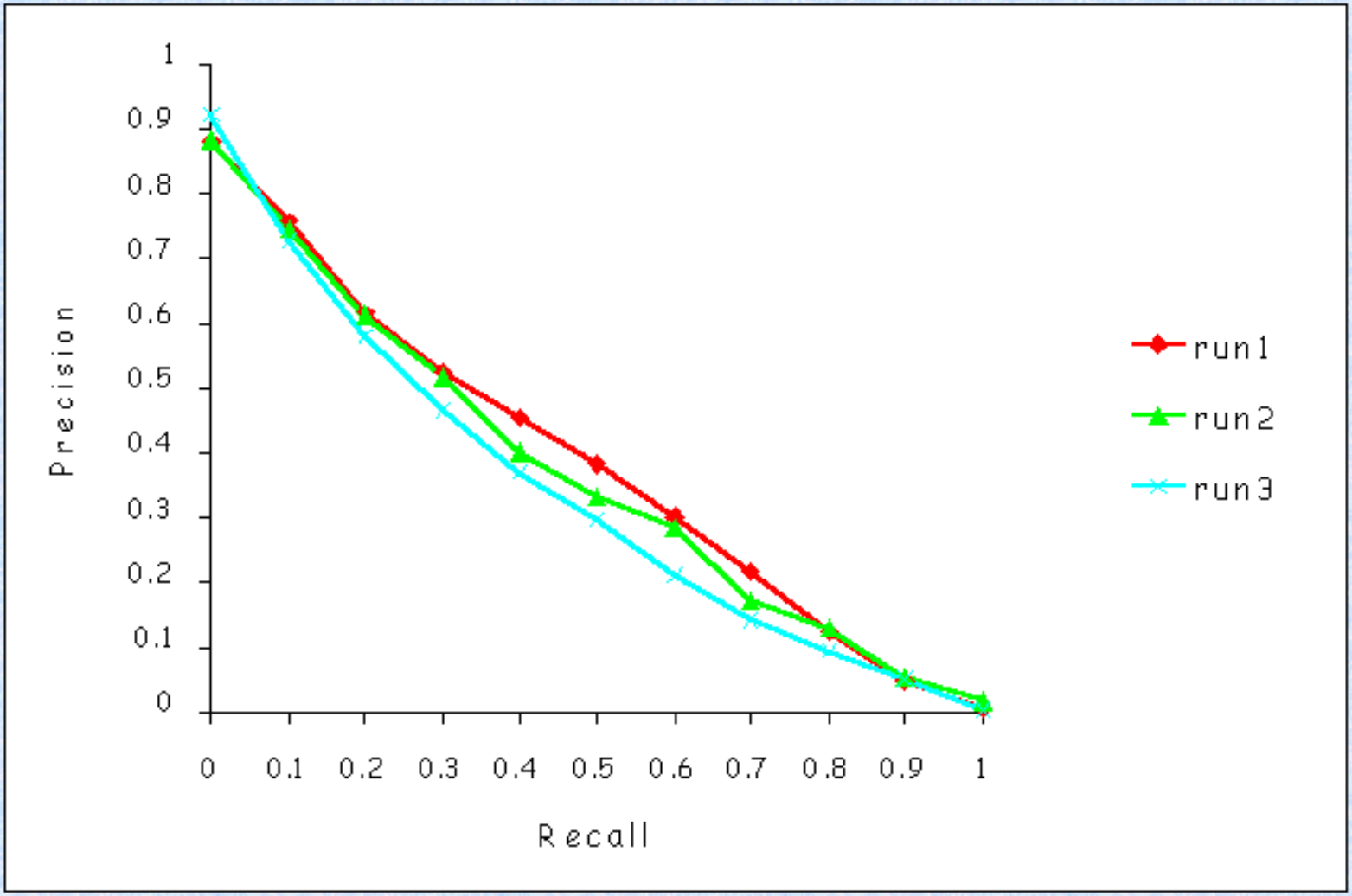
Sample TREC queries (topics)

<num> Number: 168
<title> Topic: Financing AMTRAK

<desc> Description:
A document will address the role of the Federal Government in financing the operation of the National Railroad Transportation Corporation (AMTRAK)

<narr> Narrative: A relevant document must provide information on the government's responsibility to make AMTRAK an economically viable entity. It could also discuss the privatization of AMTRAK as an alternative to continuing government subsidies. Documents comparing government subsidies given to air and bus transportation with those provided to aMTRAK would also be relevant.

Recall-Precision Graph



TREC

- Benefits:
  - made research systems scale to large collections (pre-WWW)
  - allows for somewhat controlled comparisons
- Drawbacks:
  - emphasis on high recall, which may be unrealistic for what most users want
  - very long queries, also unrealistic
  - comparisons still difficult to make, because systems are quite different on many dimensions
  - focus on batch ranking rather than interaction
    - There is an interactive track.

TREC is changing

- Emphasis on specialized “tracks”
  - Interactive track
  - Natural Language Processing (NLP) track
  - Multilingual tracks (Chinese, Spanish)
  - Filtering track
  - High-Precision
  - High-Performance
- http://trec.nist.gov/

TREC Results

- Differ each year
- For the main track:
  - Best systems not statistically significantly different
  - Small differences sometimes have big effects
    - how good was the hyphenation model
    - how was document length taken into account
  - Systems were optimized for longer queries and all performed worse for shorter, more realistic queries

What to Evaluate?

- Effectiveness
  - Difficult to measure
  - Recall and Precision are one way
  - What might be others?

How Test Runs are Evaluated

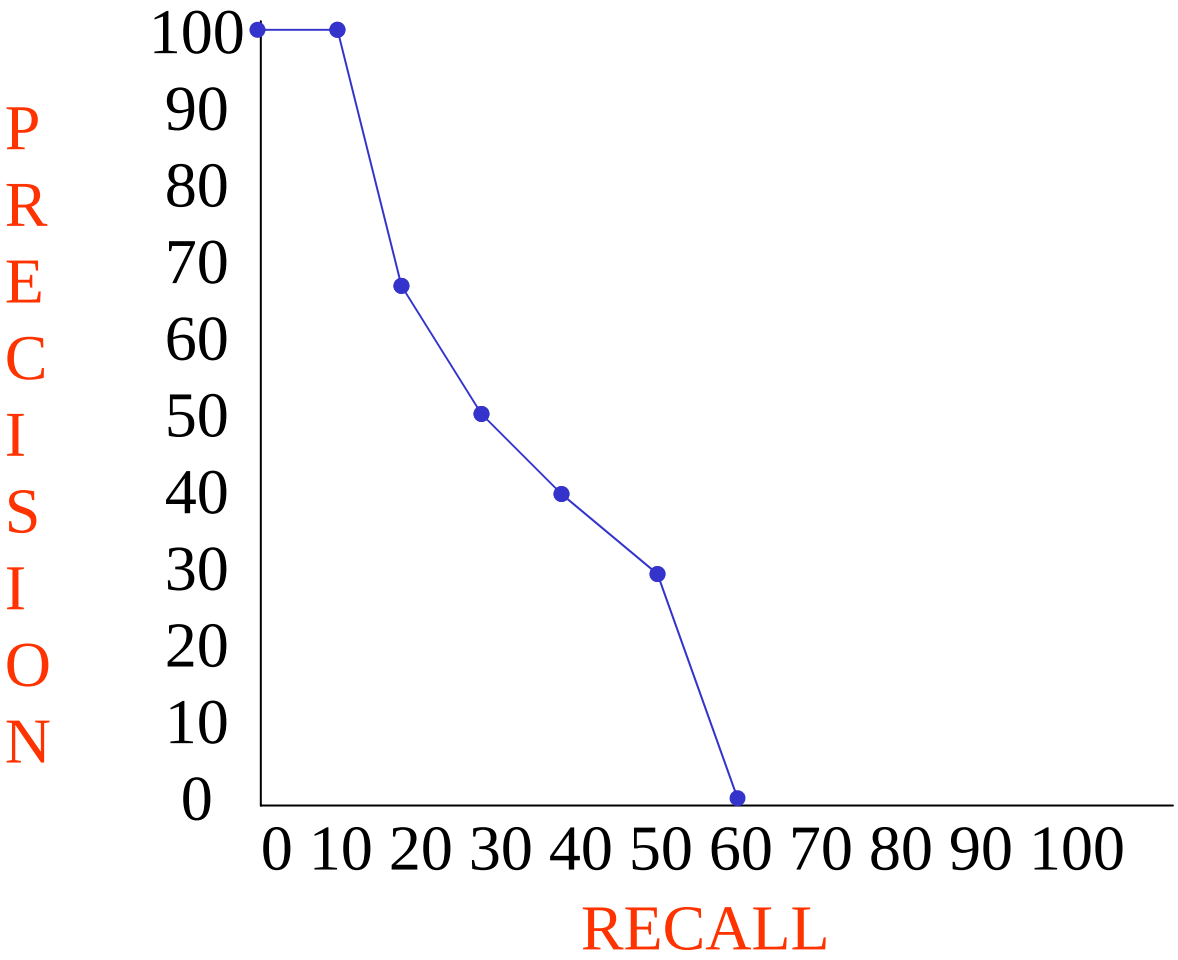
$R_q=\{d_3,d_5,d_9,d_{25},d_{39},d_{44},d_{56},d_{71},d_{89},d_{123}\}$: 10 Relevant

1. d_{123}*
2. d_{84}
3. d_{56}*
4. d_6
5. d_8
6. d_9*
7. d_{511}
8. d_{129}
9. d_{187}
10. d_{25}*
11. d_{38}
12. d_{48}
13. d_{250}
14. d_{113}
15. d_3*

- First ranked doc is relevant, which is 10% of the total relevant. Therefore Precision at the 10% Recall level is 100%
- Next Relevant gives us 66% Precision at 20% recall level
- Etc….

Examples from Chapter 3 in Baeza-Yates

Graphing for a Single Query



Averaging Multiple Queries

$$\overline{P}(r) = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q}$$

$\overline{P}(r)$ is the average Precision at Recall level r

N_q is the number of queries

$P_i(r)$ is the Precision at Recall level r for the i - th query

Interpolation

$R_q=\{d_3,d_{56},d_{129}\}$

1. d_{123}*	9. d_{187}
2. d_{84}	10. d_{25}*
3. d_{56}*	11. d_{38}
4. d_6	12. d_{48}
5. d_8	13. d_{250}
6. d_9*	14. d_{113}
7. d_{511}	15. d_3*
8. d_{129}	

- First relevant doc is 56, which is gives recall and precision of 33.3%
- Next Relevant (129) gives us 66% recall at 25% precision
- Next (3) gives us 100% recall with 20% precision
- How do we figure out the precision at the 11 standard recall levels?

Interpolation

$$r_j,\ j \in \{0,1,2,...,10\}$$

is a reference to the j - th standard recall level

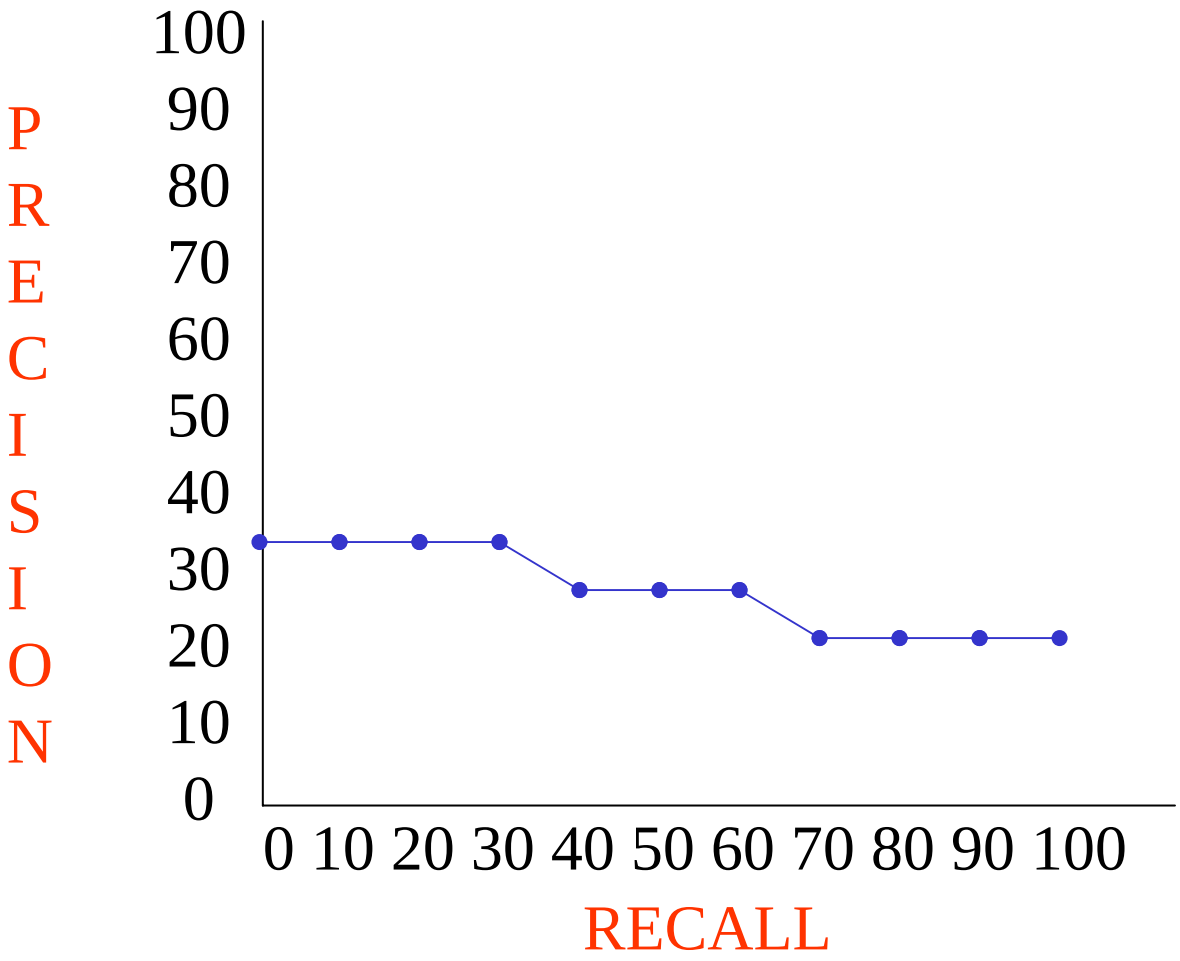
$$P(r_j) = \max r_j \le r \le r_{j+1} P(r)$$

I.e., The Maximum known Precision at any recall level between the j - th and the $(j+1)$ - th

Interpolation

- So, at recall levels 0%, 10%, 20%, and 30% the interpolated precision is 33.3%
- At recall levels 40%, 50%, and 60% interpolated precision is 25%
- And at recall levels 70%, 80%, 90% and 100%, interpolated precision is 20%
- Giving graph…

Interpolation



Problems with Precision/Recall

- Can’t know true recall value
  - except in small collections
- Precision/Recall are related
  - A combined measure sometimes more appropriate
- Assumes batch mode
  - Interactive IR is important and has different criteria for successful searches
  - We will touch on this in the UI section
- Assumes a strict rank ordering matters.

Blair and Maron 1985

- A classic study of retrieval effectiveness
  - earlier studies were on unrealistically small collections
- Studied an archive of documents for a legal suit
  - ~350,000 pages of text
  - 40 queries
  - focus on high recall
  - Used IBM's STAIRS full-text system
- Main Result:
  - The system retrieved less than 20% of the relevant documents for a particular information need; lawyers thought they had 75%
- But many queries had very high precision

Blair and Maron, cont.

- How they estimated recall
  - generated partially random samples of unseen documents
  - had users (unaware these were random) judge them for relevance
- Other results:
  - two lawyers searches had similar performance
  - lawyers recall was not much different from paralegal's

Blair and Maron, cont.

- Why recall was low
  - users can't foresee exact words and phrases that will indicate relevant documents
    - "accident" referred to by those responsible as:

    "event," "incident," "situation," "problem," …
    - differing technical terminology
    - slang, misspellings
  - Perhaps the value of higher recall decreases as the number of relevant documents grows, so more detailed queries were not attempted once the users were satisfied

Blair and Maron, cont.

- Why recall was low
  - users can't foresee exact words and phrases that will indicate relevant documents
    - "accident" referred to by those responsible as:

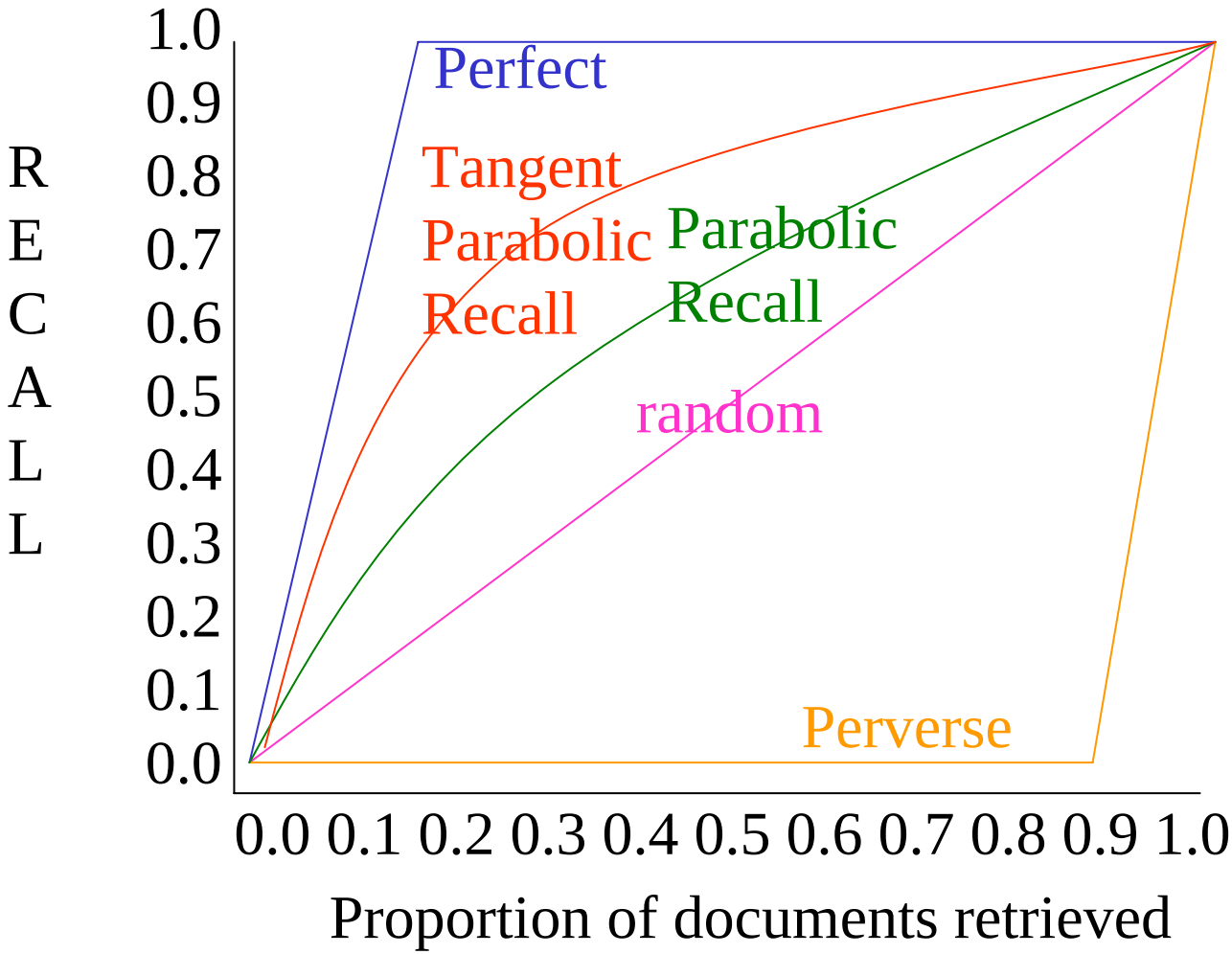
    "event," "incident," "situation," "problem," …
    - differing technical terminology
    - slang, misspellings
  - Perhaps the value of higher recall decreases as the number of relevant documents grows, so more detailed queries were not attempted once the users were satisfied

Relationship between Precision and Recall

	Doc is Relevant	Doc is NOT relevant	
Doc is retrieved	$N_{ret \cap rel}$	$N_{ret \cap \overline{rel}}$	N_{ret}
Doc is NOT retrieved	$N_{\overline{ret} \cap rel}$	$N_{\overline{ret} \cap \overline{rel}}$	$N_{\overline{ret}}$
	N_{rel}	$N_{\overline{rel}}$	N_{tot}

Buckland & Gey, JASIS: Jan 1994

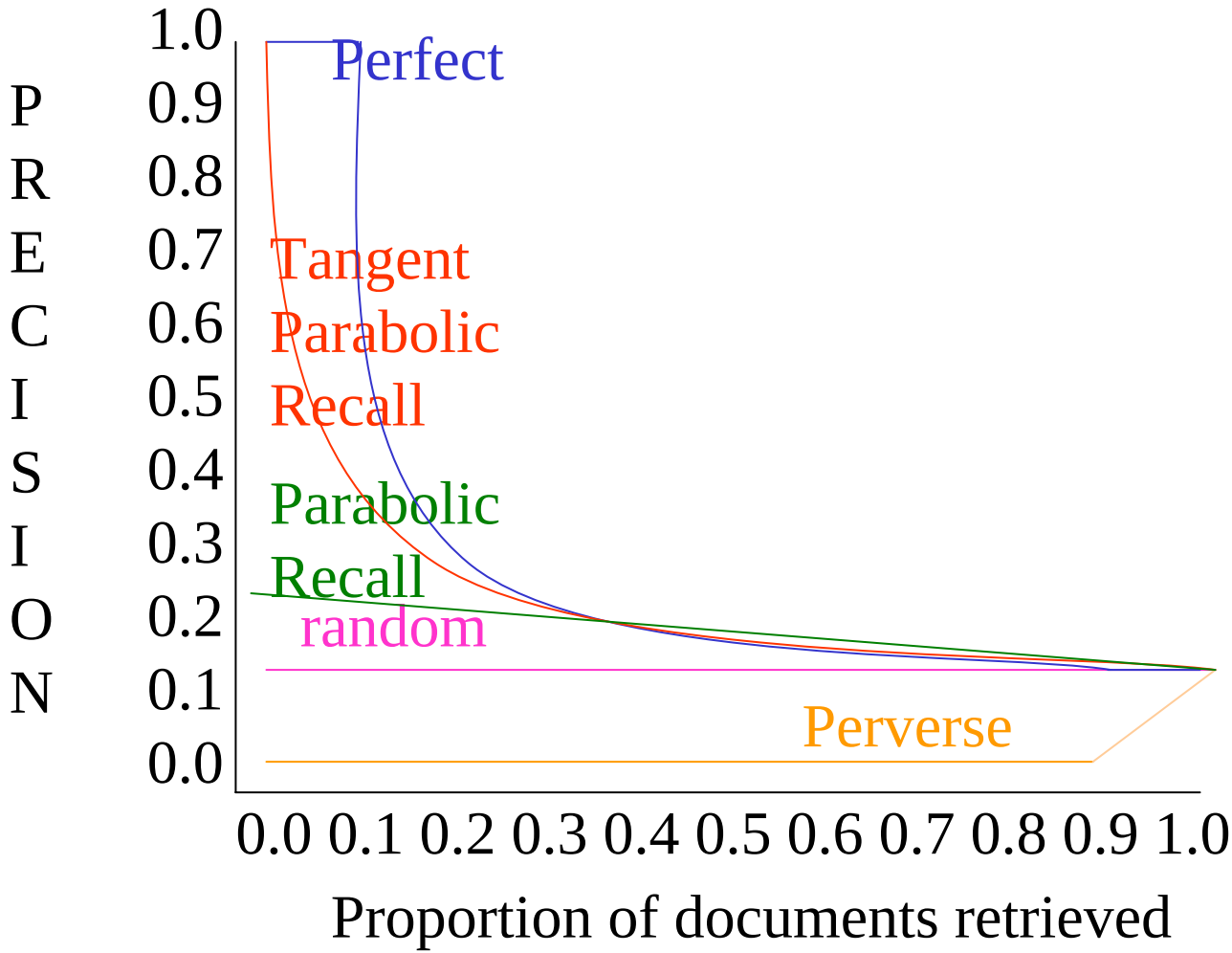
Recall Under various retrieval assumptions



1000 Documents
100 Relevant

Buckland & Gey, JASIS: Jan 1994

Precision under various assumptions



1000 Documents
100 Relevant

What to Evaluate?

- Effectiveness
  - Difficult to measure
  - Recall and Precision are one way
  - What might be others?

Other Ways of Evaluating

- “The primary function of a retrieval system is conceived to be that of saving its users to as great an extent as possible, the labor of perusing and discarding irrelevant documents, in their search for relevant ones”

William S. Cooper (1968) “Expected Search Length: A Single measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems” *American Documentation*, 19(1).

Other Ways of Evaluating

- If the purpose of retrieval system is to rank the documents in descending order of their probability of relevance for the user, then maybe the sequence is important and can be used as a way of evaluating systems.
- How to do it?

Query Types

- Only one relevant document is wanted
- Some arbitrary number *n* is wanted
- All relevant documents are wanted
- Some proportion of the relevant documents is wanted
- No documents are wanted? (Special case)

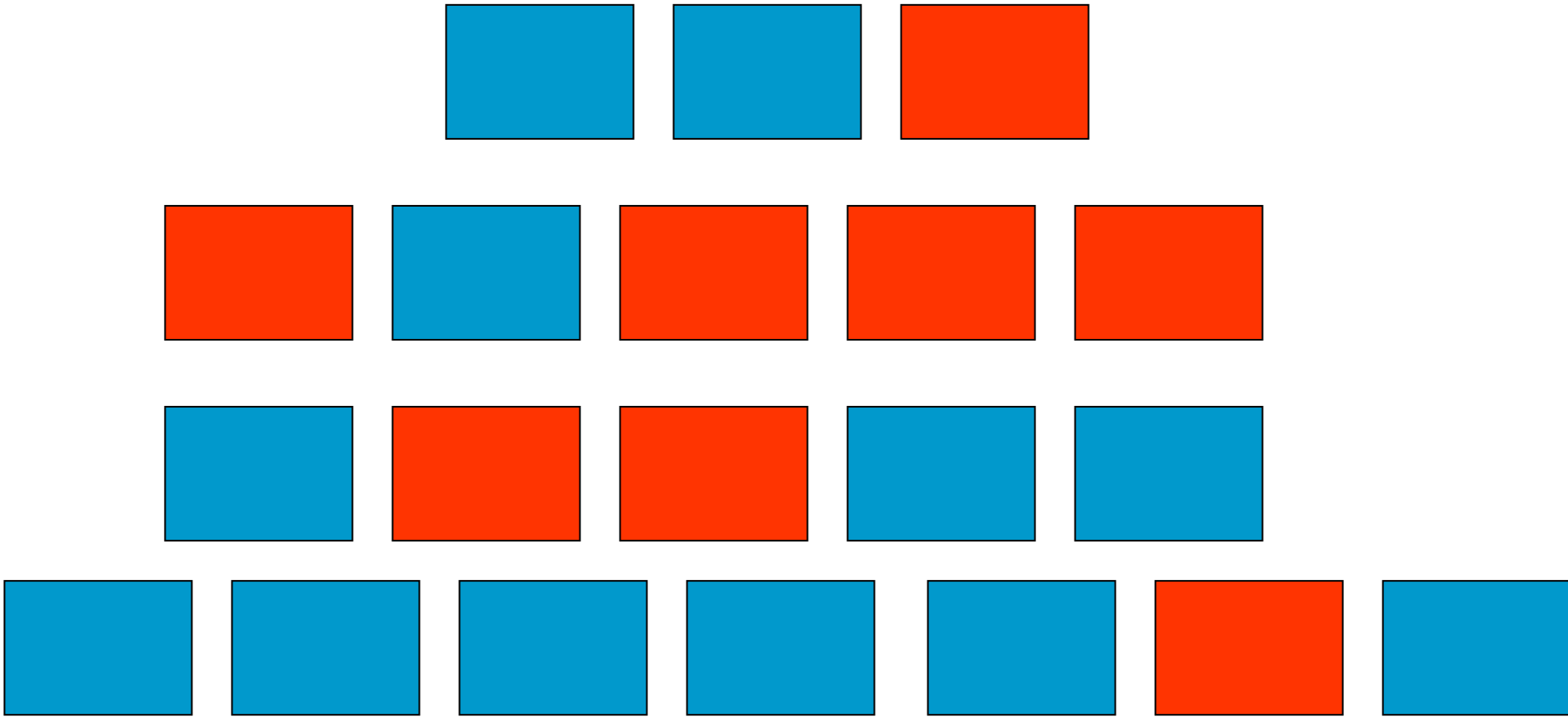
Search Length and Expected Search Length

- Work by William Cooper in the late ’60s
- Issues with IR Measures:
  - Usually not a single measure
  - Assume “retrieved” and “not retrieved” sets without considering more than two classes
  - No built-in way to compare to purely random retrieval
  - Don’t take into account how much relevant material the user actually needs (or wants)

Weak Ordering in IR Systems

- The assumption that there are two sets of “Retrieved” and “Not Retrieved” is not really accurate.
- IR Systems usually rank into many sets of equal retrieval weights
- Consider Coordinate-Level ranking…

Weak Ordering



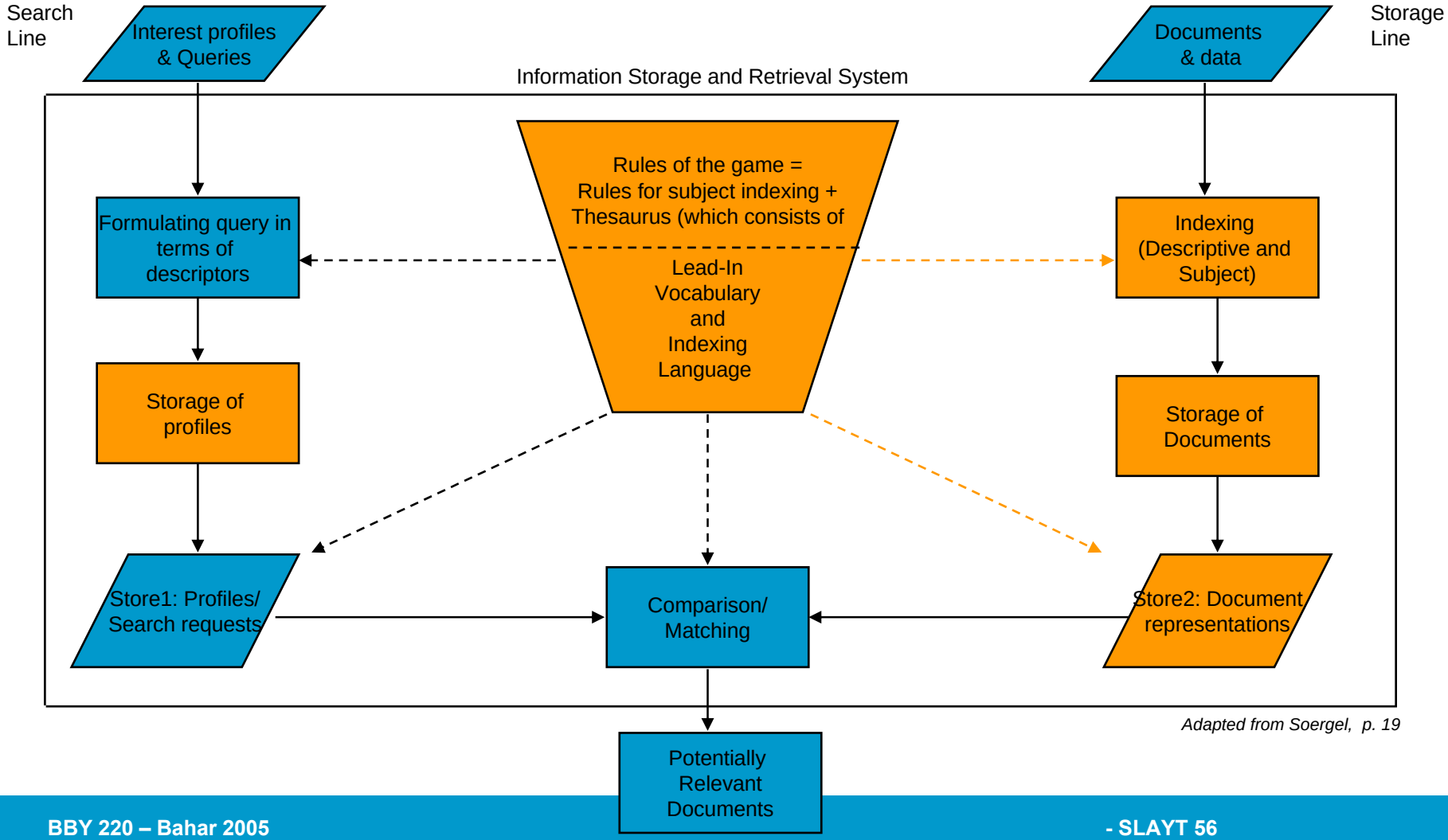
Filtering

- Characteristics of Filtering systems:
  - Designed for unstructured or semi-structured data
  - Deal primarily with text information
  - Deal with large amounts of data
  - Involve streams of incoming data
  - Filtering is based on descriptions of individual or group preferences – profiles. May be negative profiles (e.g. junk mail filters)
  - Filtering implies *removing* non-relevant material as opposed to selecting relevant.

Filtering

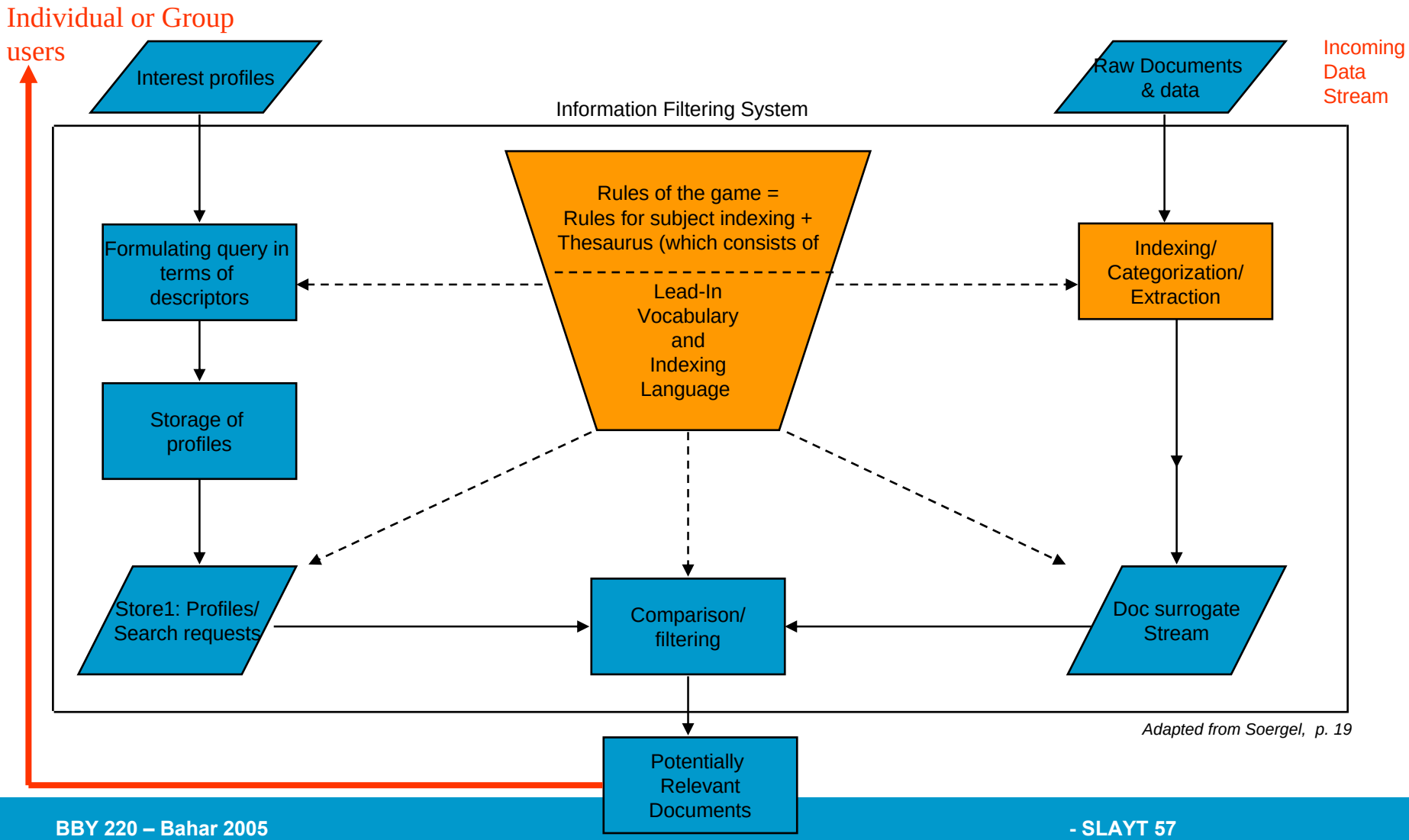
- Similar to IR, with some key differences
- Similar to Routing – sending relevant incoming data to different individuals or groups is virtually identical to filtering – with multiple profiles
- Similar to Categorization systems – attaching one or more predefined categories to incoming data objects – is also similar, but is more concerned with static categories (might be considered information extraction)

Structure of an IR System



Adapted from Soergel, p. 19

Structure of an Filtering System



Adapted from Soergel, p. 19

Major differences between IR

- IR concerned with single uses of the system
- IR recognizes inherent faults of queries
  - Filtering assumes profiles can be better than IR queries
- IR concerned with collection and organization of texts
  - Filtering is concerned with distribution of texts
- IR is concerned with selection from a static database.
  - Filtering concerned with dynamic data stream
- IR is concerned with single interaction sessions

Contextual Differences

- In filtering the *timeliness* of the text is often of greatest significance
- Filtering often has a less well-defined user community
- Filtering often has privacy implications (how complete are user profiles?, what to they contain?)
- Filtering profiles can (should?) adapt to user feedback
  - Conceptually similar to Relevance feedback

Methods for Filtering

- Adapted from IR
  - E.g. use a retrieval ranking algorithm against incoming documents.
- Collaborative filtering
  - Individual and comparative profiles

TDT: Topic Detection and Tracking

- Intended to automatically identify new topics – events, etc. – from a stream of text

Topic Detection and Tracking

Introduction and Overview

- **The TDT3 R&D Challenge**
- **TDT3 Evaluation Methodology**

Slides from “Overview NIST Topic Detection and Tracking
-Introduction and Overview” by G. Doddington
-http://www.itl.nist.gov/iaui/894.01/tests/tdt/tdt99/presentations/index.htm

TDT Task Overview*

5 R&D Challenges:

- Story Segmentation
- Topic Tracking
- Topic Detection
- *First-Story Detection*
- *Link Detection*

TDT3 Corpus Characteristics:†

- Two Types of Sources:
  - Text
  - Speech
- Two Languages:
  - English 30,000 stories
  - Mandarin 10,000 stories
- 11 Different Sources:

8 English		3 Mandarin
ABC	CNN	VOA
PRI	VOA	XIN
NBC	MNB	ZBN
APW	NYT	

* see http://www.itl.nist.gov/iaui/894.01/tdt3/tdt3.htm for details

† see http://morph.ldc.upenn.edu/Projects/TDT3/ for details

Preliminaries

A topic is …

a seminal **event** or activity, along with all directly related events and activities.

A story is …

a topically cohesive segment of news that includes two or more DECLARATIVE independent clauses about a single event.

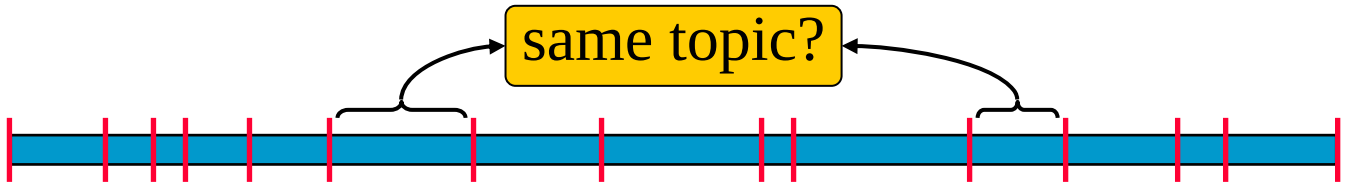
Example Topic

Title: Mountain Hikers Lost

- WHAT: 35 or 40 young Mountain Hikers were lost in an avalanche in France around the 20th of January.
- WHERE: Orres, France
- WHEN: January 1998
- RULES OF INTERPRETATION: 5. Accidents

The Link Detection Task

To detect whether a pair of stories discuss the same topic.



- The topic discussed is a free variable.
- Topic definition and annotation is unnecessary.
- The link detection task represents a basic functionality, needed to support all applications (including the TDT applications of topic detection and tracking).
- The link detection task is related to the topic tracking task, with Nt = 1.

Latent Semantic Indexing

- Latent Semantic Indexing (LSI)
- Issues in IR

LSI Rationale

- The words that searchers use to describe the their information needs are often not the same words used by authors to describe the same information.
- I.e., index terms and user search terms often do NOT match
  - Synonymy
  - Polysemy
- Following examples from Deerwester, et al. *Indexing by Latent Semantic Analysis*. JASIS 41(6), pp. 391-407, 1990

LSI Rationale

	Access	Document	Retrieval	Information	Theory	Database	Indexing	Computer	REL	M
D1	x	x	x			x	x		R	
D2				x*	x			x*		M
D3			x	x*				x *	R	M

Query: IDF in computer-based information lookup

Only matching words are “information” and “computer”
D1 is relevant, but has no words in the query…

LSI Rationale

- Problems of synonyms
  - If not specified by the user, will miss synonymous terms
  - Is automatic expansion from a thesaurus useful?
  - Are the *semantics* of the terms taken into account?
- Is there an underlying semantic *model* of terms and their usage in the database?

LSI Rationale

- Statistical techniques such as *Factor Analysis* have been developed to derive underlying meanings/models from larger collections of observed data
- A notion of semantic similarity between terms and documents is central for modelling the patterns of term usage across documents
- Researchers began looking at these methods that focus on the proximity of items within a space (as in the vector model)

LSI Rationale

- Researchers (Deerwester, Dumais, Furnas, Landauer and Harshman) considered models using the following criteria
  - Adjustable representational richness
  - Explicit representation of both terms and documents
  - Computational tractability for large databases

Clustering and Automatic Classification

- Clustering
- Automatic Classification
- Cluster-enhanced search

Classification

- The grouping together of items (including documents or their representations) which are then treated as a unit. The groupings may be predefined or generated algorithmically. The process itself may be manual or automated.
- In document classification the items are grouped together because they are likely to be wanted together
  - For example, items about the same topic.

Automatic Indexing and Classification

- Automatic indexing is typically the simple deriving of keywords from a document and providing access to all of those words.
- More complex Automatic Indexing Systems attempt to select controlled vocabulary terms based on terms in the document.
- Automatic classification attempts to automatically group similar documents using either:
  - A fully automatic clustering method.
  - An established classification scheme and set of documents already indexed by that scheme.

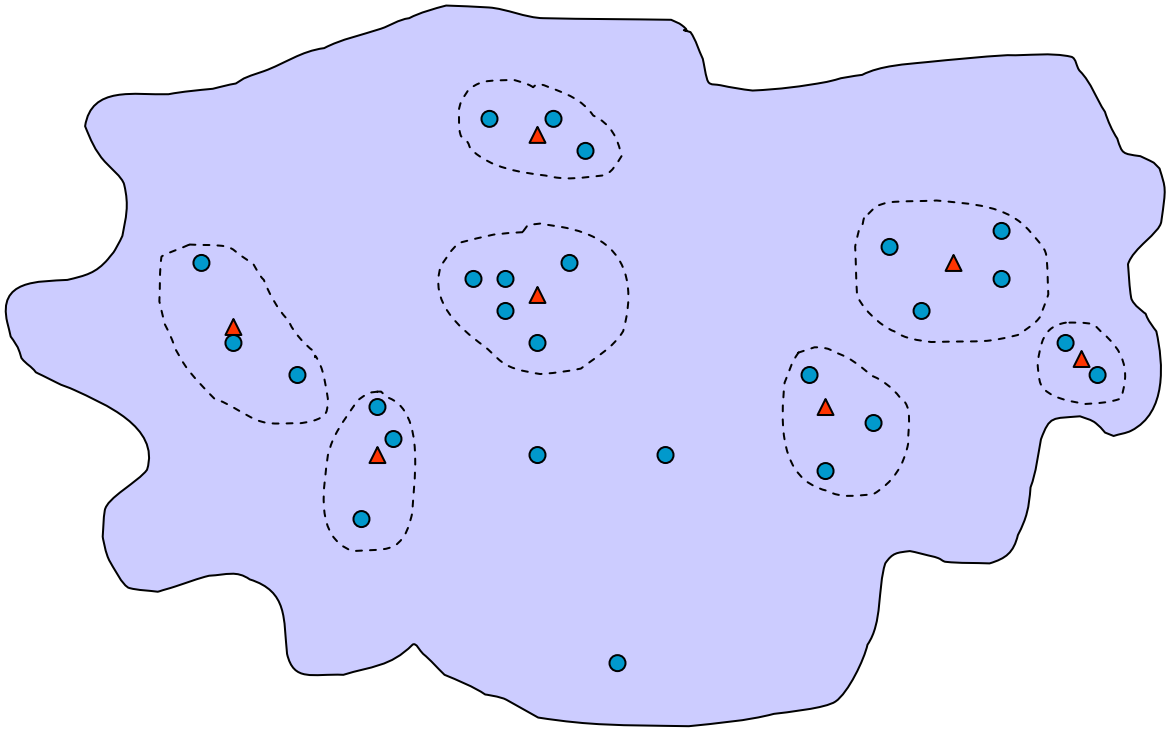
Background and Origins

- Early suggestion by Fairthorne
  - "The Mathematics of Classification"
- Early experiments by Maron (1961) and Borko and Bernick(1963)
- Work in Numerical Taxonomy and its application to Information retrieval Jardine, Sibson, van Rijsbergen, Salton (1970's).
- Early IR clustering work more concerned with efficiency issues than semantic issues.

Document Space has High Dimensionality

- What happens beyond three dimensions?
- Similarity still has to do with how many tokens are shared in common.
- More terms -> harder to understand which subsets of words are shared among similar documents.
- One approach to handling high dimensionality:Clustering

Vector Space Visualization



Cluster Hypothesis

- The basic notion behind the use of classification and clustering methods:
- “Closely associated documents tend to be relevant to the same requests.”
  - C.J. van Rijsbergen

Classification of Classification Methods

- Class Structure
  - Intellectually Formulated
    - Manual assignment (e.g. Library classification)
    - Automatic assignment (e.g. Cheshire Classification Mapping)
  - Automatically derived from collection of items
    - Hierarchic Clustering Methods (e.g. Single Link)
    - Agglomerative Clustering Methods (e.g. Dattola)
    - Hybrid Methods (e.g. Query Clustering)

Classification of Classification Methods

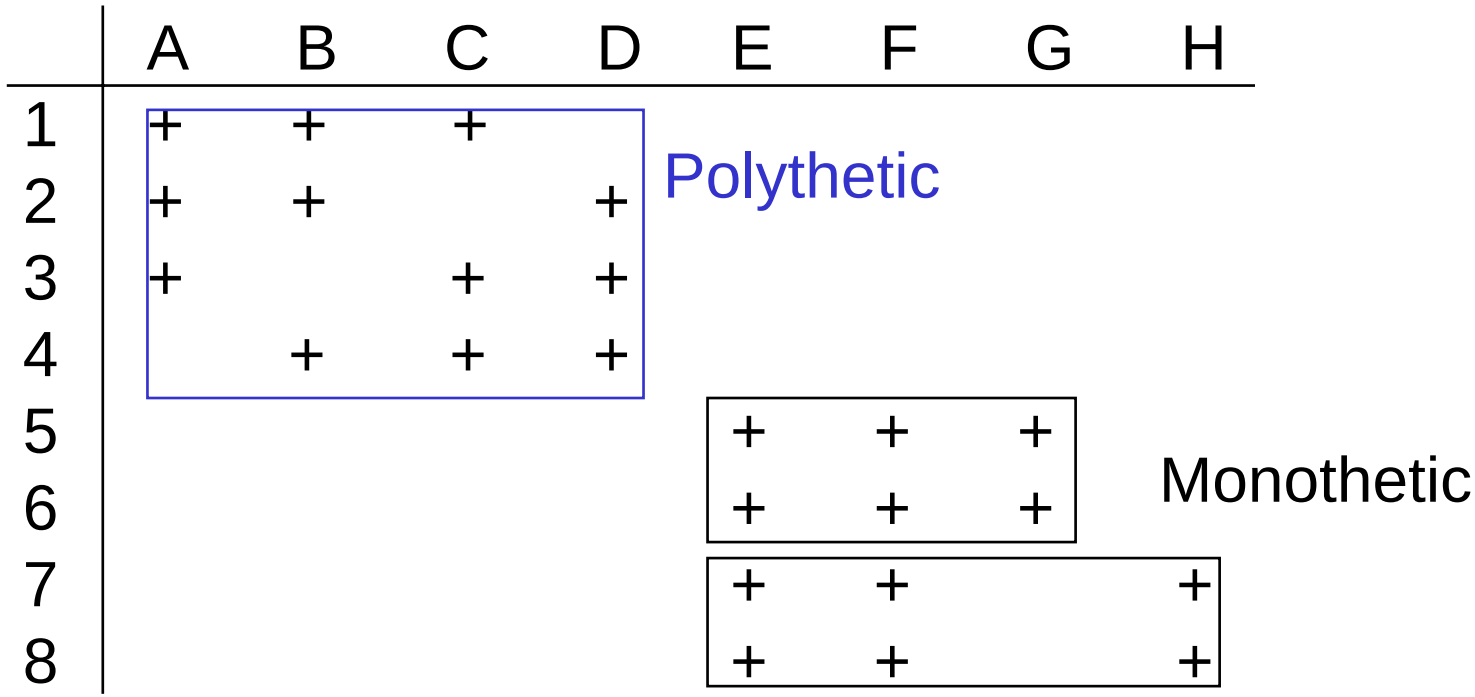
- Relationship between properties and classes
  - monothetic
  - polythetic
- Relation between objects and classes
  - exclusive
  - overlapping
- Relation between classes and classes
  - ordered
  - unordered

Adapted from Sparck Jones

Properties and Classes

- Monothetic
  - Class defined by a set of properties that are both *necessary* and *sufficient* for membership in the class
- Polythetic
  - Class defined by a set of properties such that to be a member of the class some individual must have some number (usually large) of those properties, and that a large number of individuals in the class possess some of those properties, and no individual possesses all of the properties.

Monothetic vs. Polythetic



Adapted from van Rijsbergen, ‘79

Exclusive Vs. Overlapping

- Item can either belong exclusively to a single class
- Items can belong to many classes, sometimes with a “membership weight”

Ordered Vs. Unordered

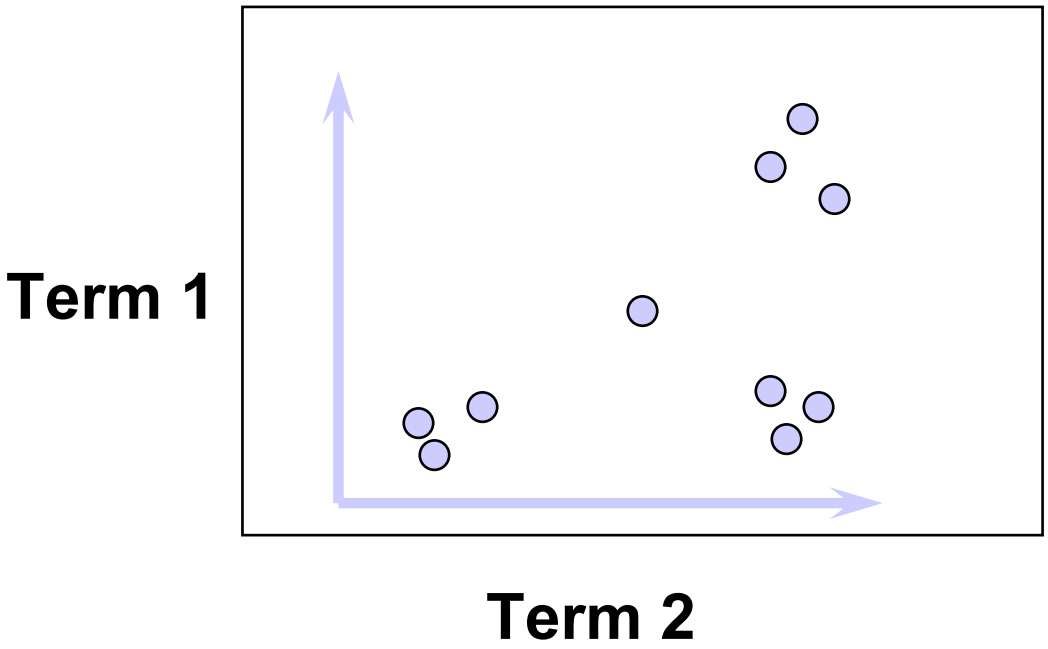
- Ordered classes have some sort of structure imposed on them
  - Hierarchies are typical of ordered classes
- Unordered classes have no imposed precedence or structure and each class is considered on the same “level”
  - Typical in agglomerative methods

Text Clustering

Clustering is

"The art of finding groups in data."

-- Kaufmann and Rousseeu

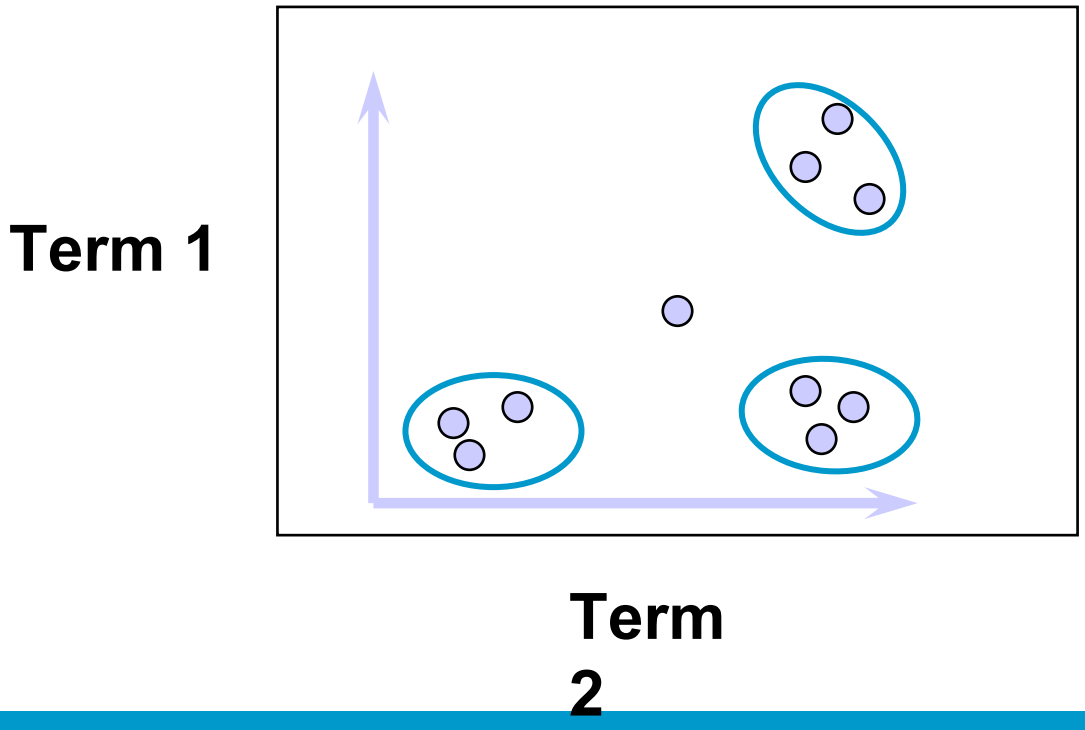


Text Clustering

Clustering is

"The art of finding groups in data."

-- Kaufmann and Rousseeu



Text Clustering

- Finds overall similarities among groups of documents
- Finds overall similarities among groups of tokens
- Picks out some themes, ignores others

Coefficients of Association

$$|A \cap B|$$

- Simple

$$2\frac{|A \cap B|}{|A| + |B|}$$

- Dice's coefficient

$$\frac{|A \cap B|}{|A \cup B|}$$

- Jaccard's coefficient

$$\frac{|A \cap B|}{\sqrt{|A|} + \sqrt{|B|}}$$

- Cosine coefficient

$$\frac{|A \cap B|}{\min(|A|, |B|}$$

- Overlap coefficient

Pair-wise Document Similarity

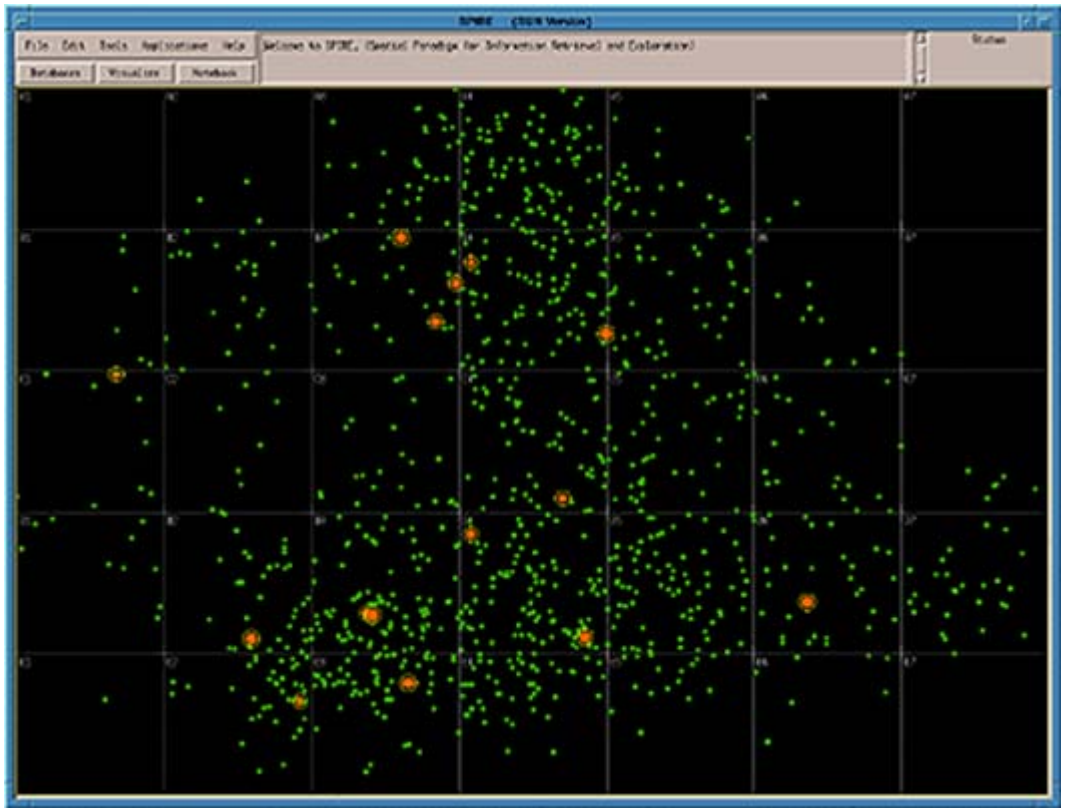
	nova	galaxy	heat	h'wood	film	role	diet	fur
A	1	3	1					
B	5	2						
C				2	1	5		
D				4	1			

How to compute document similarity?

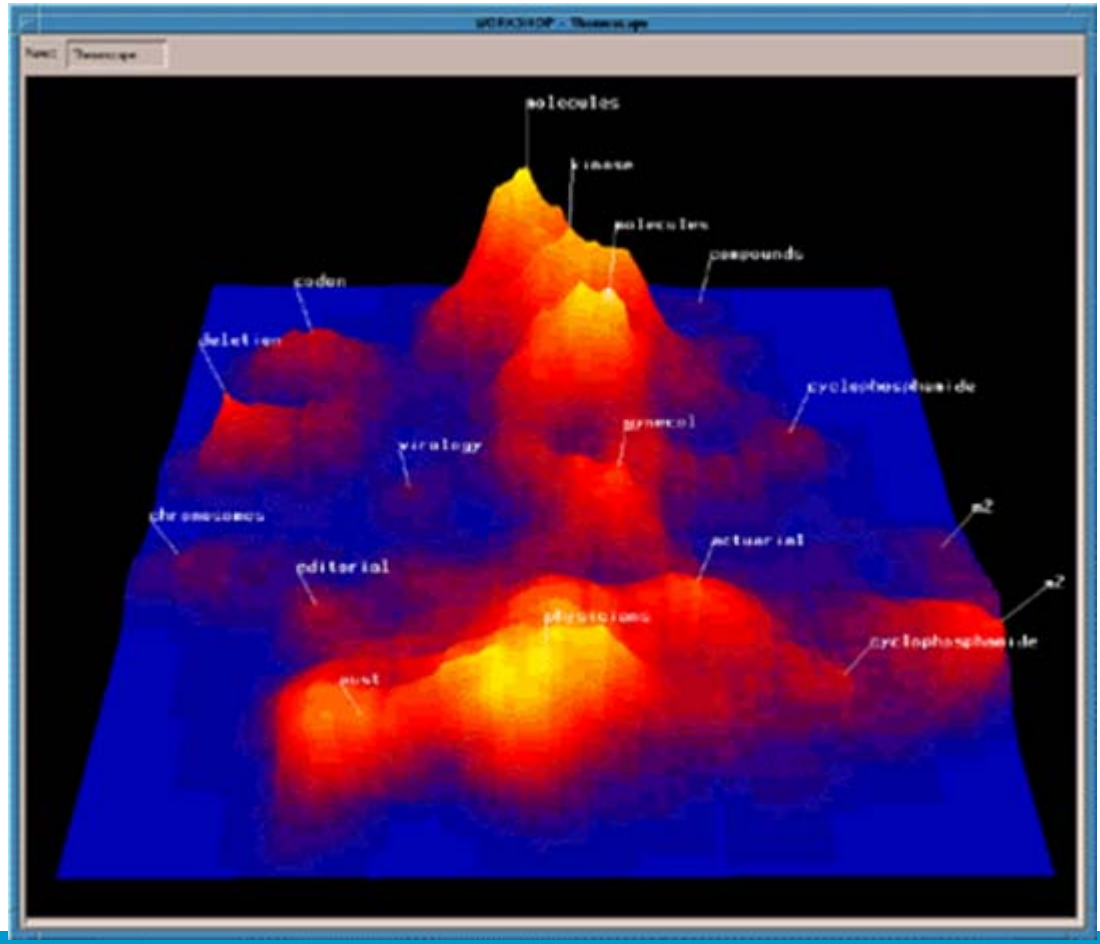
Another use of clustering

- Use clustering to map the entire huge multidimensional document space into a huge number of small clusters.
- “Project” these onto a 2D graphical representation:

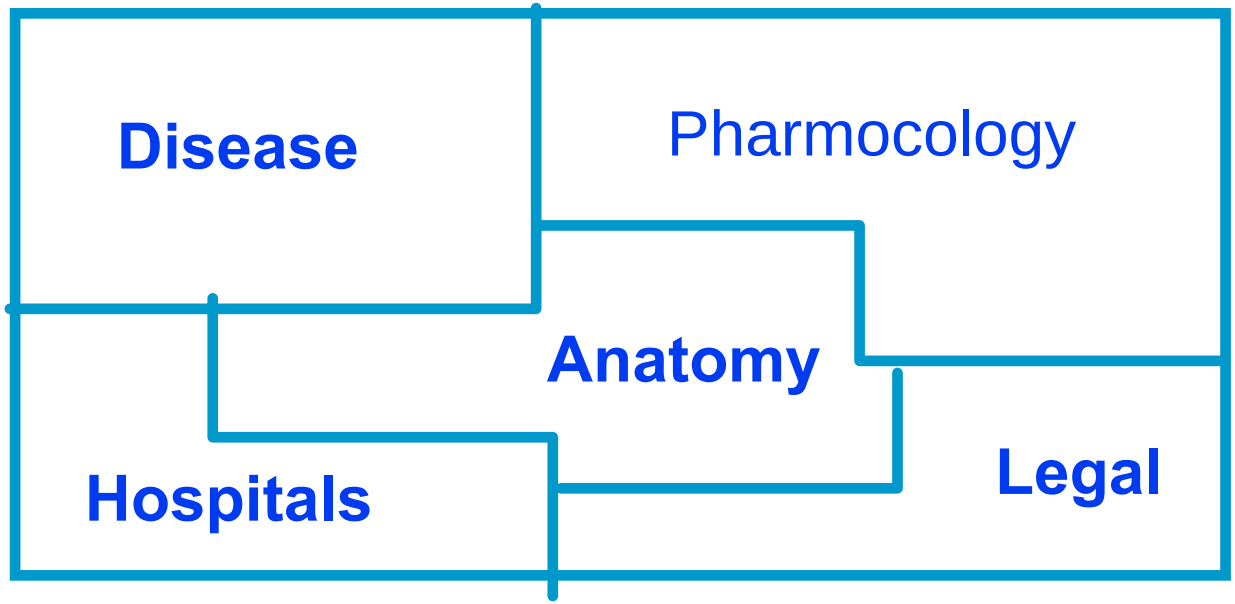
Clustering Multi-Dimensional Document Space



Clustering Multi-Dimensional Document Space



Concept "Landscapes"



(e.g., Lin, Chen, Wise et al.)

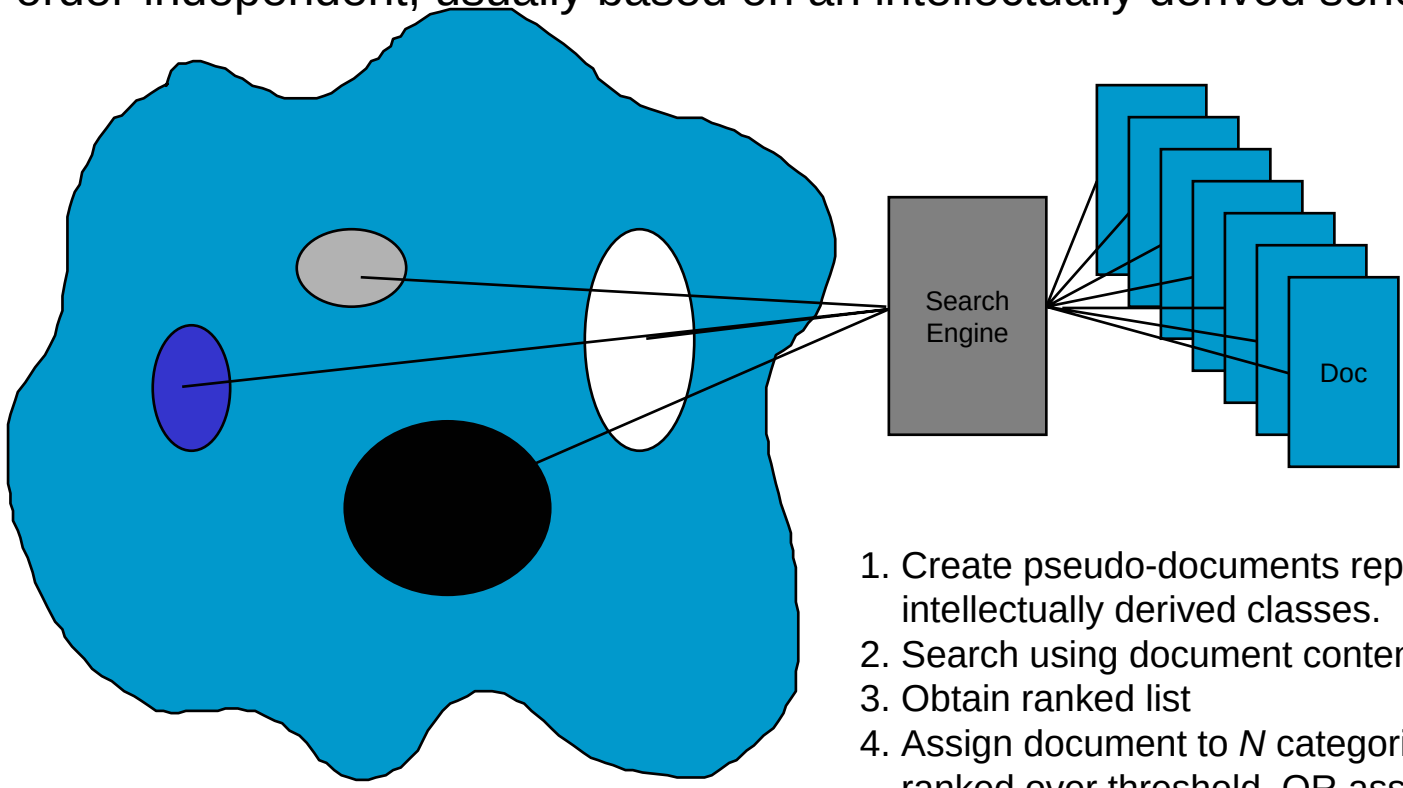
- Too many concepts, or too coarse
- Single concept per document
- No titles
- Browsing without search

Clustering

- Advantages:
  - See some main themes
- Disadvantage:
  - Many ways documents could group together are hidden
- Thinking point: what is the relationship to classification systems and facets?

Automatic Class Assignment

Automatic Class Assignment: Polythetic, Exclusive or Overlapping, usually ordered clusters are order-independent, usually based on an intellectually derived scheme



1. Create pseudo-documents representing intellectually derived classes.
2. Search using document contents
3. Obtain ranked list
4. Assign document to *N* categories ranked over threshold. OR assign to top-ranked category