

Document Representation

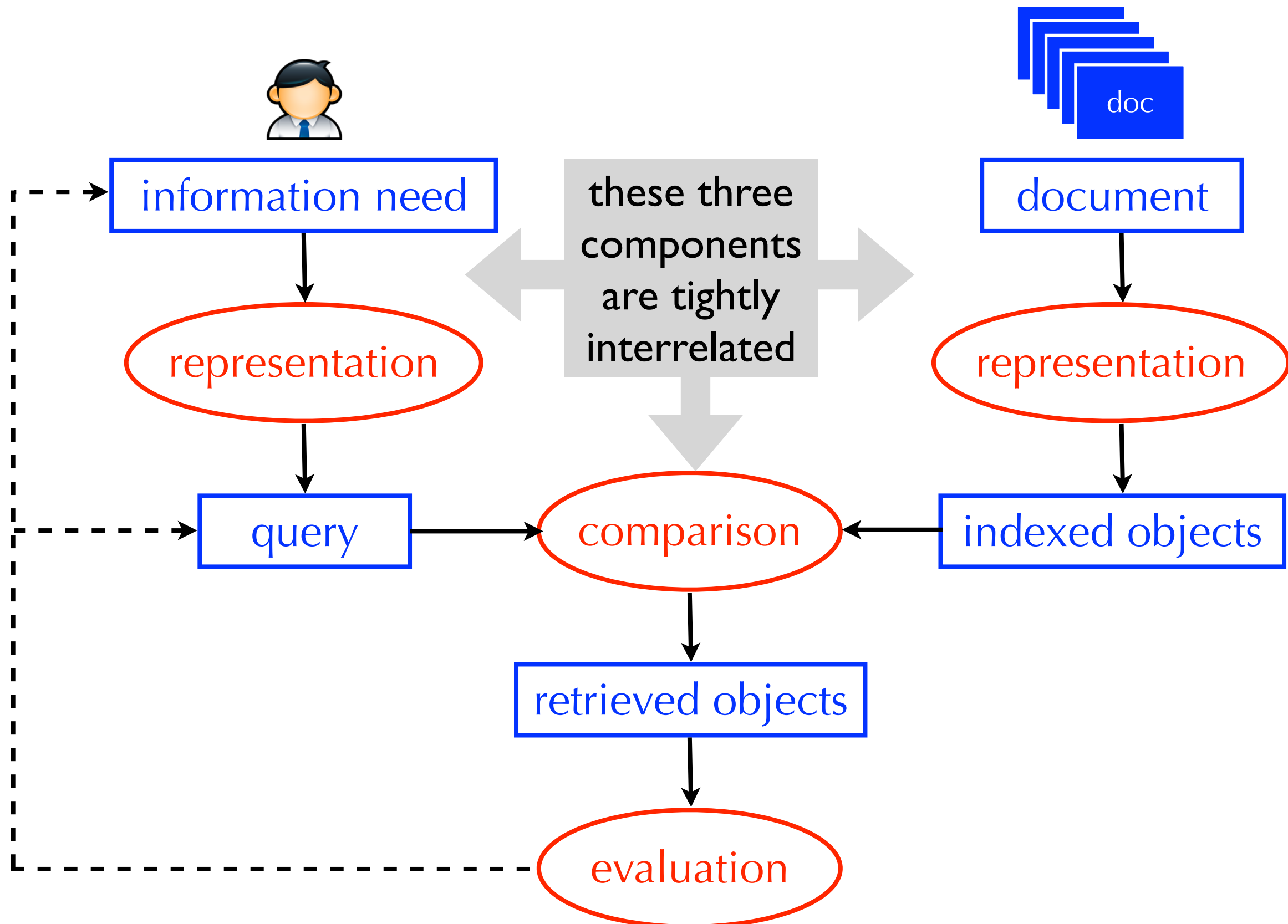
Jaime Arguello

INLS 509: Information Retrieval

jarguell@email.unc.edu

February 8, 2016

Document Representation



Document Representation

- How should this document be represented?



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](#)

► [Print/export](#)

▼ Languages
[Deutsch](#)
[Español](#)
[Bahasa Indonesia](#)

[Log in / create account](#)

Article [Discussion](#) [Read](#) [Edit](#) [View history](#)

Gerard Salton

From Wikipedia, the free encyclopedia

Gerard Salton (8 March 1927 in [Nuremberg](#) - 28 August 1995), also known as Gerry Salton, was a Professor of [Computer Science](#) at [Cornell University](#). Salton was perhaps the leading computer scientist working in the field of [information retrieval](#) during his time. His group at Cornell developed the [SMART Information Retrieval System](#), which he initiated when he was at Harvard.

Salton was born Gerhard Anton Sahlmann on March 8, 1927 in [Nuremberg, Germany](#). He received a Bachelor's (1950) and Master's (1952) degree in mathematics from [Brooklyn College](#), and a Ph.D. from [Harvard](#) in [Applied Mathematics](#) in 1958, the last of [Howard Aiken](#)'s doctoral students, and taught there until 1965, when he joined [Cornell University](#) and co-founded its department of Computer Science.

Salton was perhaps most well known for developing the now widely used [Vector Space Model](#) for Information Retrieval^[1]. In this model, both documents and queries are represented as vectors of term counts, and the similarity between a document and a query is given by the cosine between the term vector and the document vector. In this paper, he also introduced [TF-IDF](#), or term-frequency-inverse-document frequency, a model in which the score of a term in the a document is the ratio of the number of terms in that document divided by the frequency of the number of documents in which that term occurs. (The concept of inverse document frequency, a measure of specificity, had been introduced in 1972 by [Karen Sparck-Jones](#)^[2].) Later in life, he became interested in automatic text summarization and analysis^[3], as well as automatic hypertext generation^[4]. He published over 150 research articles and 5 books during his life.

Salton was editor-in-chief of the [Communications of the ACM](#) and the [Journal of the ACM](#), and chaired [SIGIR](#). He was an associate editor of the [ACM Transactions on Information Systems](#). He was an [ACM Fellow](#) (elected 1995), received an Award of Merit from the [American Society for Information Science](#) (1989), and was the first recipient of the [SIGIR Award](#) for outstanding contributions to study of information retrieval (1983) -- now called the [Gerard Salton Award](#).

Elements of a Document Representation

- Document attributes (metadata)
 - ▶ source, publication date, language, length, etc.
- Controlled vocabulary index terms
- Free-text index terms
 - ▶ terms selected from the document text itself
 - ▶ may also include text from outside the document (e.g., anchor text)
 - ▶ lots of room for creativity!

Elements of a Document Representation



controlled-
vocabulary index
terms

Categories: 1927 births | 1995 deaths | American computer scientists | Computer pioneers | Harvard University alumni | Harvard University faculty | Cornell University faculty | Fellows of the Association for Computing Machinery | Guggenheim Fellows

Elements of a Document Representation

anchor text
(nearby terms?)



This screenshot shows the Wikipedia article for Gerard Salton. The page includes a sidebar with navigation links, a top navigation bar with tabs for 'Article', 'Discussion', 'Read', 'Edit', and 'View history', and a search bar. The main content area contains a brief biography of Gerard Salton, mentioning his birth in 1927, his work at Cornell University, and his role in developing the SMART Information Retrieval System. A red box highlights a sentence in the text: "Amit continued his studies at Cornell University in Ithaca, New York and received a Ph.D. degree in 1996. At Cornell Amit studied with Gerard Salton, a pioneer in the field of information retrieval, the academic discipline which forms the foundation of modern search. John Battelle, in his book 'The Search' calls Gerard Salton 'the father of digital search.'"



This screenshot shows the Wikipedia article for Amit Singhal. The page includes a sidebar with navigation links, a top navigation bar with tabs for 'Article', 'Discussion', 'Read', 'View source', and 'View history', and a search bar. The main content area contains a brief biography of Amit Singhal, mentioning his work at Google Inc. and his role as a Google Fellow. A red box highlights a sentence in the text: "Amit continued his studies at Cornell University in Ithaca, New York and received a Ph.D. degree in 1996. At Cornell Amit studied with Gerard Salton, a pioneer in the field of information retrieval, the academic discipline which forms the foundation of modern search. John Battelle, in his book 'The Search' calls Gerard Salton 'the father of digital search.'"

Text Processing

gerard salton 8 march 1978 in nuremberg 28 august 1995 also know as gerry salton was professor of computer science at cornell university salton was perhaps the leading computer scientist working in the field of information retrieval during his time his group at cornell developed the smart information retrieval system which he initiated when he was at harvard

- Our goal is to describe content using content
- After mark-up removal, down-casing, and tokenization, what we have is a sequence of terms
- What are the most descriptive words?



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](#)
[Print/export](#)

[Languages](#)
[Deutsch](#)
[Español](#)
[Bahasa Indonesia](#)

ArticleDiscussion

ReadEditView history

Gerard Salton

From Wikipedia, the free encyclopedia

Gerard Salton (8 March 1927 in [Nuremberg](#) - 28 August 1995), also known as Gerry Salton, was a Professor of [Computer Science](#) at [Cornell University](#). Salton was perhaps the leading computer scientist working in the field of information retrieval during his time. His group at Cornell developed the [SMART Information Retrieval System](#), which he initiated when he was at Harvard.

Salton was born Gerhard Anton Sahlmann on March 8, 1927 in [Nuremberg, Germany](#). He received a Bachelor's (1950) and Master's (1952) degree in mathematics from [Brooklyn College](#), and a Ph.D. from [Harvard](#) in [Applied Mathematics](#) in 1958, the last of [Howard Aiken's](#) doctoral students, and taught there until 1965, when he joined [Cornell University](#) and co-founded its department of Computer Science.


Salton was perhaps most well known for developing the now widely used [Vector Space Model](#) for Information Retrieval^[1]. In this model, both documents and queries are represented as vectors of term counts, and the similarity between a document and a query is given by the cosine between the term vector and the document vector. In this paper, he also introduced TF-IDF, or term-frequency-inverse-document frequency, a model in which the score of a term in the a document is the ratio of the number of terms in that document divided by the frequency of the number of documents in which that term occurs. (The concept of inverse document frequency, a measure of specificity, had been introduced in 1972 by [Karen Sparck-Jones](#)^[2].) Later in life, he became interested in automatic text summarization and analysis^[3], as well as automatic hypertext generation^[4]. He published over 150 research articles and 5 books during his life.

Salton was editor-in-chief of the [Communications of the ACM](#) and the [Journal of the ACM](#), and chaired SIGIR. He was an associate editor of the [ACM Transactions on Information Systems](#). He was an [ACM Fellow](#) (elected 1995), received an [Award of Merit](#) from the [American Society for Information Science](#) (1989), and was the first recipient of the [SIGIR Award](#) for outstanding contributions to study of information retrieval (1983) -- now called the [Gerard Salton Award](#).

Term-Frequencies

top 20

rank	term	freq.	rank	term	freq.
1	the	34	11	as	9
2	of	29	12	he	9
3	a	20	13	vector	8
4	in	20	14	an	8
5	and	19	15	s	7
6	salton	18	16	term	7
7	model	15	17	for	7
8	was	12	18	automatic	7
9	information	11	19	paper	6
10	retrieval	10	20	gerard	6



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](#)
[Print/export](#)

[Languages](#)
[Deutsch](#)
[Español](#)
[Bahasa Indonesia](#)

ArticleDiscussion

ReadEditView history

Gerard Salton

From Wikipedia, the free encyclopedia

Gerard Salton (8 March 1927 in [Nuremberg](#) - 28 August 1995), also known as Gerry Salton, was a Professor of [Computer Science](#) at [Cornell University](#). Salton was perhaps the leading computer scientist working in the field of information retrieval during his time. His group at Cornell developed the [SMART Information Retrieval System](#), which he initiated when he was at Harvard.

Salton was born Gerhard Anton Sahlmann on March 8, 1927 in [Nuremberg, Germany](#). He received a Bachelor's (1950) and Master's (1952) degree in mathematics from [Brooklyn College](#), and a Ph.D. from [Harvard](#) in Applied Mathematics in 1958, the last of [Howard Aiken's](#) doctoral students, and taught there until 1965, when he joined [Cornell University](#) and co-founded its department of Computer Science.

Salton was perhaps most well known for developing the now widely used [Vector Space Model](#) for Information Retrieval^[1]. In this model, both documents and queries are represented as vectors of term counts, and the similarity between a document and a query is given by the cosine between the term vector and the document vector. In this paper, he also introduced TF-IDF, or term-frequency-inverse-document frequency, a model in which the score of a term in the a document is the ratio of the number of terms in that document divided by the frequency of the number of documents in which that term occurs. (The concept of inverse document frequency, a measure of specificity, had been introduced in 1972 by [Karen Sparck-Jones](#)^[2].) Later in life, he became interested in automatic text summarization and analysis^[3], as well as automatic hypertext generation^[4]. He published over 150 research articles and 5 books during his life.

Salton was editor-in-chief of the [Communications of the ACM](#) and the [Journal of the ACM](#), and chaired SIGIR. He was an associate editor of the [ACM Transactions on Information Systems](#). He was an [ACM Fellow](#) (elected 1995), received an [Award of Merit](#) from the [American Society for Information Science](#) (1989), and was the first recipient of the SIGIR Award for outstanding contributions to study of information retrieval (1983) -- now called the [Gerard Salton Award](#).

Term-Frequencies

top 20

rank	term	freq.	rank	term	freq.
1	the	34	11	as	9
2	of	29	12	he	9
3	a	20	13	vector	8
4	in	20	14	an	8
5	and	19	15	s	7
6	salton	18	16	term	7
7	model	15	17	for	7
8	was	12	18	automatic	7
9	information	11	19	paper	6
10	retrieval	10	20	gerard	6

IMDB Corpus

term-frequencies

rank	term	frequency	rank	term	frequency
1	the	1586358	11	year	250151
2	a	854437	12	he	242508
3	and	822091	13	movie	241551
4	to	804137	14	her	240448
5	of	657059	15	artist	236286
6	in	472059	16	character	234754
7	is	395968	17	cast	234202
8	i	390282	18	plot	234189
9	his	328877	19	for	207319
10	with	253153	20	that	197723

Stopwords

- A stopword is a term that is discarded from the document representation
- Typically the same set of stopwords is used in processing all documents in the collection
- Stopwords are typically function words: determiners (a, the), prepositions (on, above), conjunctions (and, but)
- May also be corpus-specific: “plot” in the IMDB corpus
- **Assumption:** stopwords are unimportant because they are frequent in every document

Lemur Stopword List

first 60 (sorted alphabetically)

a	all	amongst	anywhere	become	besides
about	almost	an	apart	becomes	between
above	alone	and	are	becoming	beyond
according	along	another	around	been	both
across	already	any	as	before	but
after	also	anybody	at	beforehand	by
afterwards	although	anyhow	av	behind	can
again	always	anyone	be	being	can
against	am	anything	became	below	cannot
albeit	among	anyway	because	beside	canst



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

Print/export

Languages

Deutsch

Español

Bahasa Indonesia

ArticleDiscussion

ReadEditView history

Gerard Salton

From Wikipedia, the free encyclopedia

Gerard Salton (8 March 1927 in [Nuremberg](#) - 28 August 1995), also known as Gerry Salton, was a Professor of [Computer Science](#) at [Cornell University](#). Salton was perhaps the leading computer scientist working in the field of information retrieval during his time. His group at Cornell developed the [SMART Information Retrieval System](#), which he initiated when he was at Harvard.

Salton was born Gerhard Anton Sahlmann on March 8, 1927 in [Nuremberg, Germany](#). He received a Bachelor's (1950) and Master's (1952) degree in mathematics from [Brooklyn College](#), and a Ph.D. from [Harvard in Applied Mathematics](#) in 1958, the last of [Howard Aiken's](#) doctoral students, and taught there until 1965, when he joined [Cornell University](#) and co-founded its department of Computer Science.

Salton was perhaps most well known for developing the now widely used [Vector Space Model](#) for Information Retrieval^[1]. In this model, both documents and queries are represented as vectors of term counts, and the similarity between a document and a query is given by the cosine between the term vector and the document vector. In this paper, he also introduced TF-IDF, or term-frequency-inverse-document frequency, a model in which the score of a term in the a document is the ratio of the number of terms in that document divided by the frequency of the number of documents in which that term occurs. (The concept of inverse document frequency, a measure of specificity, had been introduced in 1972 by [Karen Sparck-Jones](#)^[2].) Later in life, he became interested in automatic text summarization and analysis^[3], as well as automatic hypertext generation^[4]. He published over 150 research articles and 5 books during his life.

Salton was editor-in-chief of the [Communications of the ACM](#) and the [Journal of the ACM](#), and chaired SIGIR. He was an associate editor of the [ACM Transactions on Information Systems](#). He was an [ACM Fellow](#) (elected 1995), received an [Award of Merit](#) from the [American Society for Information Science](#) (1989), and was the first recipient of the [SIGIR Award](#) for outstanding contributions to study of information retrieval (1983) -- now called the [Gerard Salton Award](#).

Term-Frequencies

after stopwords removal

rank	term	freq.	rank	term	freq.
1	salton	18	11	paper	6
2	model	15	12	document	6
3	information	11	13	acm	6
4	retrieval	10	14	1975	4
5	vector	8	15	frequency	4
6	s	7	16	science	4
7	term	7	17	cornell	4
8	automatic	7	18	award	3
9	gerard	6	19	0	3
10	space	6	20	8	3

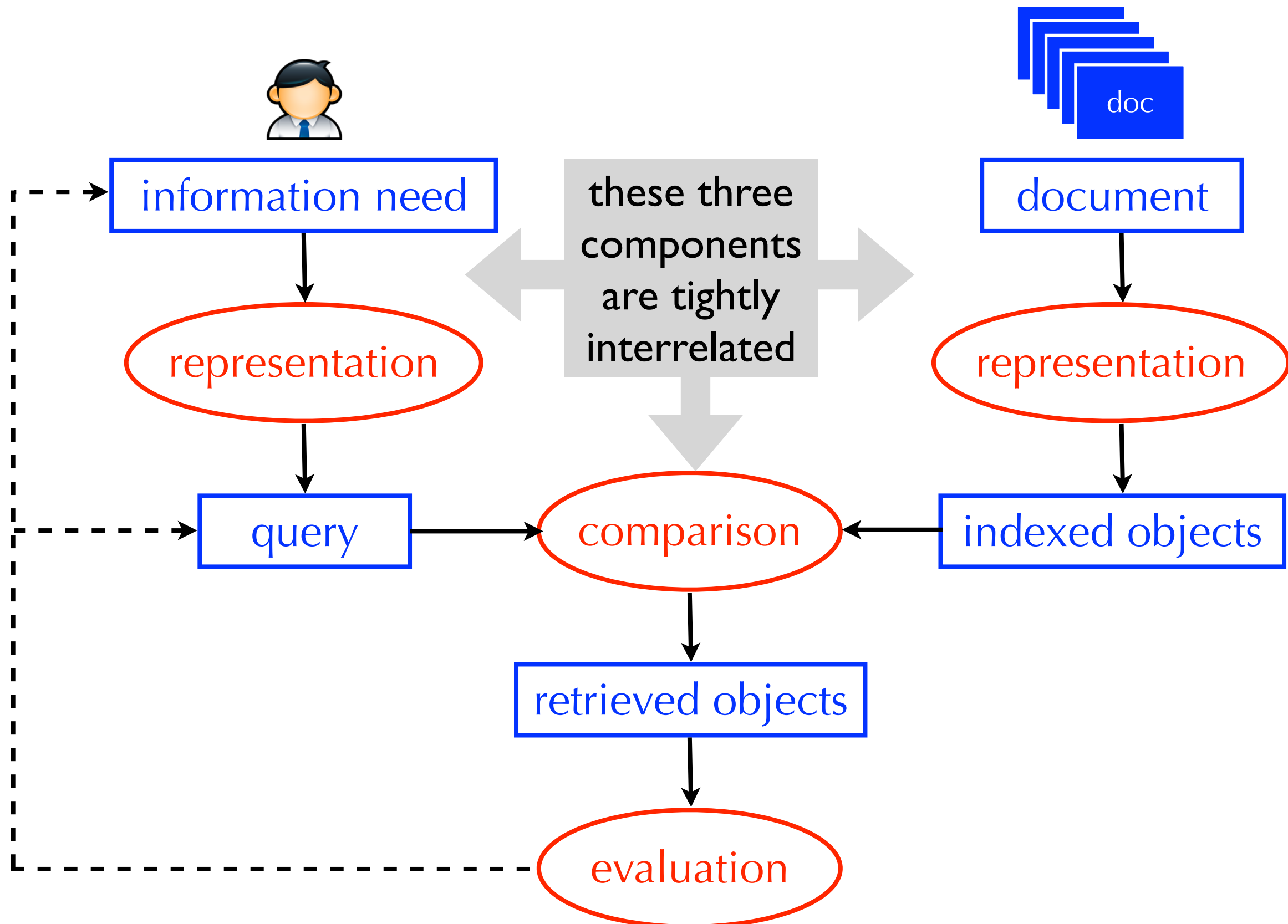
Creating a Stopword List

- Sort vocabulary based on frequency in the corpus
- Examine the most frequent words
- Examine a query-log to see which frequent terms may be important
- 38% of unique AOL queries contain at least one stopword (may or may not be important - more later)
- 0.025% of unique AOL queries are nothing but stopwords
 - ▶ **longest:** i want something else to get me through this (third eye blind lyrics)

Trends in Stopword Removal

- The earliest systems used stopwords lists of 200-300 terms
- To improve efficiency and effectiveness
- Very frequent terms were problematic for early retrieval models (e.g, **OR** operations in ranked boolean)
- Web search engines generally do not remove stopwords
- The latest trend is to index stopwords and (possibly) ignore them at query-time if they seem unimportant
- Newer retrieval models are better at handling very frequent terms (later lecture)

Document Representation



AOL Query-Log Examples

stopword removal

wrong lyrics

am i wrong lyrics
i was wrong lyrics
wrong again lyrics
where did i go wrong lyrics
wrong lyrics
got me wrong lyrics
what went wrong lyrics

buy house

who will buy my house
buy a house
buy my house
buy house
we buy house
how to buy a house

change

be the change you want in others
how can i change me
change
where is my change
i want my change
never change

calculate bmi

calculate bmi
calculate my bmi
how to calculate your bmi
how to calculate bmi

Morphological Analysis

Morphology

- the study and description of word formation (as inflection, derivation, and compounding) in language

Merriam-Webster Dictionary

Morphology

- **Inflectional morphology:** changes to a word that encode its grammatical usage (e.g., tense, number, person)
 - ▶ say vs. said, cat vs. cats, see vs. sees
- **Derivational morphology:** changes to a word to make a new word with related meaning
 - ▶ organize, organization, organizational
- **Compounding:** combining words to form new ones
 - ▶ shipwreck, outbound, beefsteak
 - ▶ more common in other languages (e.g., german)
 - ▶ lebensversicherungsgesellschaftangestellter

Morphological Analysis

in information retrieval

- **Basic question:** words occur in different forms. Do we want to treat different forms as different index terms?
- **Conflation:** treating different (inflectional and derivational) variants as the same index term

Morphological Analysis

in information retrieval

- **Conflation:** treating different (inflectional and derivational) variants as the same index term

<i>image</i>	<i>images</i>	<i>imaging</i>	<i>imag*</i> (root form)
<i>df=6</i>	<i>df=4</i>	<i>df=3</i>	<i>df=6</i>
1, 4	1, 4	1, 4	1, 12
10, 1	10, 5	10, 5	10, 11
15, 2	16, 1	16, 1	15, 2
16, 1	68, 1		16, 3
33, 5			33, 5
68, 7			68, 8

docid , term frequency

Morphological Analysis

in information retrieval

- What are we trying to achieve by conflating morphological variants?
- **Goal:** help the system ignore unimportant variations of language usage



Morphological Analysis in information retrieval

repairing computer



[Guide to Computer Troubleshooting and Repair - PC ...](#)

[www.daileyint.com/hmdpc/manual.htm](#) - Cached

PC's are actually much easier to **repair** these days than in the early 90's when I wrote my original guide for technicians I was training. The number of discrete ...

[Online Computer Training Courses - For all beginners and experts ...](#)

[www.beyourownit.com/](#) - Cached

Want to learn about **computers**? You've found the right place! On this website, you'll be able to find everything from simple **computer repair** articles and **computer ..**

[Repairing basic computer hard ware problem \(system disk failure ...](#)

[www.instructables.com/.../Repairing-basic-computer-hard-ware-and...](#) - Cached

May 10, 2008 – THIS GUIDE IS NOT YET FINISHED, I WILL ADD MORE INFORMATION WHEN I GET A CHANCE.If you need any help with **fixing** a **computer** ...

[Computer-Repair Technicians](#)

[www.collegeboard.com > ... > Majors & Careers Central > Profiles](#) - Cached

Computer-repair technicians maintain and **repair computers**, scanners, printers, monitors, and other computer equipment. Learn more about this career at ...

[Computer repair NYC | Laptop repair ny | PC repair NY](#)

[ifixny.com/](#) - Cached

Computer repair ny. Data recovery nyc. We offers a full range of **computer fix** and technical support with free diagnostics and estimates, also iPhone BlackBerry ...

Morphological Analysis

in information retrieval

- The query “computer repairs” will match all combinations of:

computer
computers
computing
computation
computational
::

and

repair
repairs
repaired
repairing
repairable
::

Morphological Analysis

in information retrieval

- In English, conflating morphological variants is usually done using a stemmer
- **Stemming**: automatic suffix-stripping
- English word variations occur at the end of a word
- root/stem + suffix
 - ▶ repair + s/ed/ing/able
- A stemmer conflates different variations by reducing them to a common root/stem

Morphological Analysis

in information retrieval

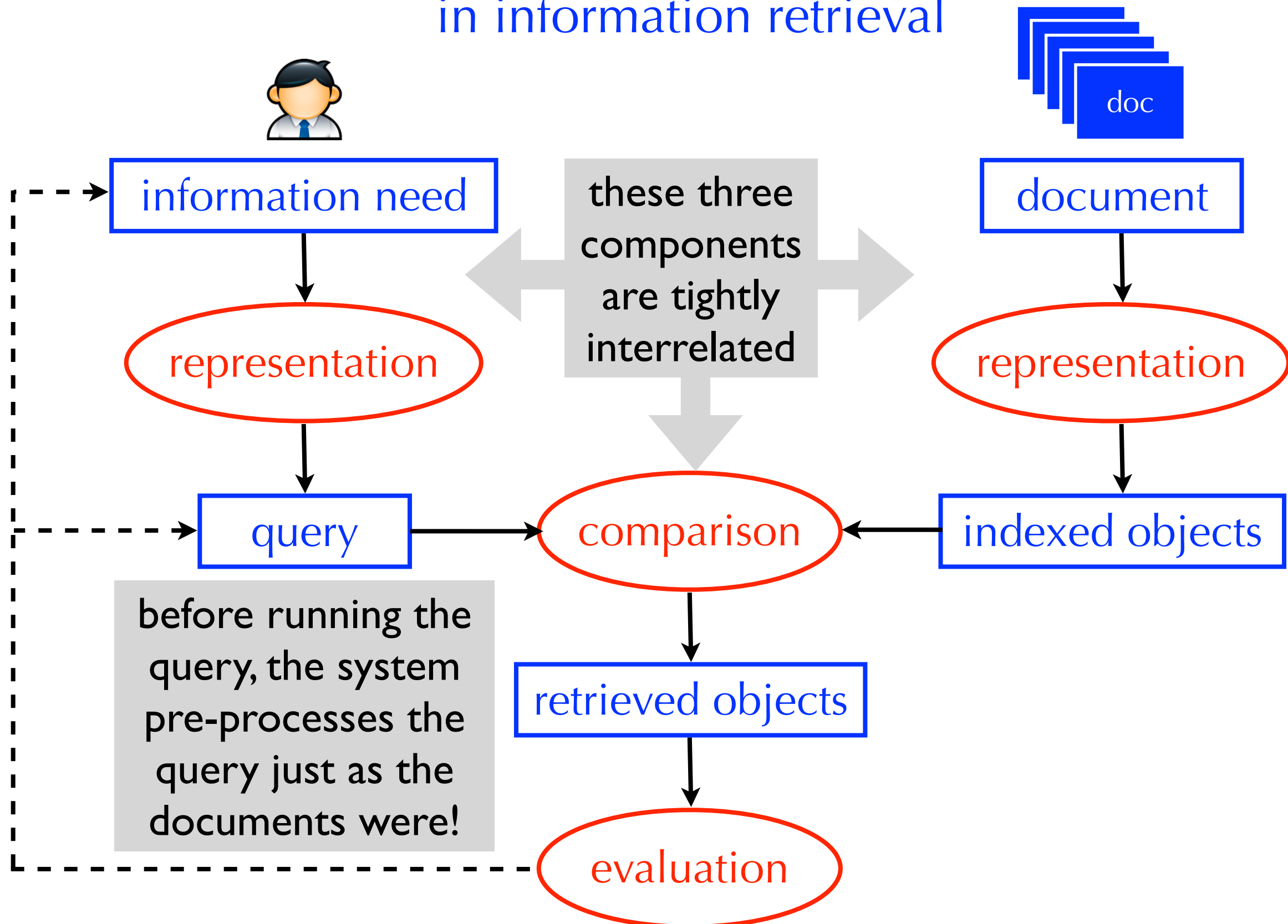
- In some cases, whatever is left after suffix-stripping is not even a word (e.g., **comput**)
- Is this a problem?

computer
computers
computing
computation
computational
::

repair
repairs
repaired
repairing
repairable
::

Morphological Analysis

in information retrieval



Morphological Analysis

the porter stemmer (porter '80)

- A long list of rules that are applied in sequence
 - ▶ apply the rule that removes the longest suffix
 - ▶ check to see that the stem is likely to be a root (replac+ement vs. c+ement)
- Fast, effective, and, therefore, very popular

Martin Porter's Home Page

No doubt you came here out of idle curiosity from the [Porter Stemming Algorithm](#) page. Before you hastily return, you are welcome to look at the following.

This (jerkily) spinning can is the work of [Philip Holmes Esquire](#), ingenious graphic designer and inventor of visual puns. I could never have thought up anything so clever. (Apologies to the Dr Pepper people!)



Morphological Analysis

the porter stemmer (porter '80)

- Example step (1 of 5)

Step 1a:

- Replace *sses* by *ss* (e.g., stresses → stress).
- Delete *s* if the preceding word part contains a vowel not immediately before the *s* (e.g., gaps → gap but gas → gas).
- Replace *ied* or *ies* by *i* if preceded by more than one letter, otherwise by *ie* (e.g., ties → tie, cries → cri).
- If suffix is *us* or *ss* do nothing (e.g., stress → stress).

Step 1b:

- Replace *eed*, *eedly* by *ee* if it is in the part of the word after the first non-vowel following a vowel (e.g., agreed → agree, feed → feed).
- Delete *ed*, *edly*, *ing*, *ingly* if the preceding word part contains a vowel, and then if the word ends in *at*, *bl*, or *iz* add *e* (e.g., fished → fish, pirating → pirate), or if the word ends with a double letter that is not *ll*, *ss* or *zz*, remove the last letter (e.g., falling → fall, dripping → drip), or if the word is short, add *e* (e.g., hoping → hope).
- Whew!

Morphological Analysis

the porter stemmer (porter '80)

- Original Text

gerard salton 8 march 1978 in nuremberg 28 august 1995 also know as gerry salton was professor of computer science at cornell university salton was perhaps the leading computer scientist working in the field of information retrieval during his time his group at cornell developed the smart information retrieval system which he initiated when he was at harvard

- Stemmed Text

gerard salton 8 march 1978 in nuremberg 28 august 1995 also know as gerri salton wa professor of comput scienc at cornel univers salton wa perhap the lead comput scientist work in the field of inform retriev dure hi time hi group at cornel develop the smart inform retriev system which he initi when he wa at harvard

Morphological Analysis

the porter stemmer (porter '80)

- **false positives:** two words conflated to the same root when they shouldn't have been

organization/organ
generalization/generic
numerical/numerous
policy/police
university/universe
addition/additive
negligible/negligent
execute/executive
past/paste
ignore/ignorant
special/specialized
head/heading

Morphological Analysis

the porter stemmer (porter '80)

- **false negatives:** two words not conflated to the same root word when they should have been

european/europe

cylinder/cylindrical

matrices/matrix

urgency/urgent

create/creation

analysis/analyses

useful/usefully

noise/noisy

decompose/decomposition

sparse/sparsity

resolve/resolution

triangle/triangular

AOL Query-log Examples

stemmed queries

russian translat

russian translations
russian translator
russian translation
russian translate

secret

secret
secretions
secrets
secretion

stock for sale

stockings for sale
stocking for sale
stocks for sale

smokei mountain nation park

smokey mountains national park
smokey mountain national park
smokey mountains national parks

cat fenc

cat fencing
cat fences
cat fence

strawberri plant

strawberry planting
strawberry plants
strawberries planting

AOL Query-log Examples

stopped + stemmed queries

bui comput

buy a computer
buying a computer
we buy computers
how to buy a computer
buying computers

rid raccoon

get rid of raccoons
how to get rid of raccoons
how to get rid of a raccoon
what to use to get rid of raccoons
how do i get rid of a raccoon

auto repair

auto repairables
how to auto repairs
auto repair do it yourself
do it yourself auto repair
auto repair .com
do it yourself auto repairs
auto repair

water diet

the water diet
the all water diet
water and diet
water diet
water diets

AOL Query-log Examples

stopped + stemmed queries

planet orbit sun

why is there only one planet in each orbit around the sun
why do the planets orbit the sun
planets that orbit the sun

plant shade

plant shade
plants for shade
plants that do well in shade
plants that like shade
plants shade
planting in the shade

univers

universalism
universism
other universe
university
our universe
across the universe
the universe within
universities

Morphological Analysis

evaluation results

- Stemming
 - ▶ English: 0-5% improvements
 - ▶ Finnish: 30% improvement
 - ▶ Spanish: 10% improvement
- Compound Splitting
 - ▶ German: 15% improvements
 - ▶ Swedish: 25% improvement

(Hollink *et al.*, 2004)

Morphology Across Languages

European Parliament Corpus

- Number of unique terms (remember, these are translations of the same text):
 - ▶ English: 150,725
 - ▶ Spanish: 213,486
 - ▶ Portuguese: 219,121
 - ▶ Danish: 367,282
 - ▶ Finnish: 709,049
 - ▶ German: 401,929

To Stem or Not To Stem

small corpus large corpus

users care
more about
recall

users care
more about
precision

?	?
?	?

To Stem or Not To Stem

small corpus large corpus

users care
more about
recall

users care
more about
precision

yes	maybe
maybe	maybe

- Google seems to be doing stemming. They must think it helps

Big Picture

- Text-processing requires making decisions about what to store in the index
- Two big decisions: stopword-removal and stemming
- My own recommendation (take it, leave it, question it)
 - ▶ remove stopwords only if you have to (don't have enough disk-space)
 - ▶ off-load the job to query-processing (removing stopwords from the query)
 - ▶ stem depending on the importance of recall and the size of the collection

What about homonyms?
(words that are spelled the same, but have different meaning)

Words often have multiple senses

- *bank* (noun)
 1. the rising ground bordering a lake, river, or sea
 2. a mound, pile, or ridge above the surrounding level
 3. a steep slope (as in “bank of a hill”)
 4. an establishment for the custody, loan, exchange, and issue of money
 5. a supply of something held in reserve
 6. the lateral inward tilt of a vehicle (as an airplane) when turning

(Merriam-Webster Dictionary)

Word Sense Disambiguation

- Given a **word** in a particular **context**, automatically predict its correct **sense** from a finite set (bank 1-6)?

“I stopped by the **bank** to deposit some cash.”



An establishment for the custody, loan, and exchange of money

“I stopped by the food **bank** to donate some food.”



A supply of something held in reserve

- An active area of research since the 1950's
- How would you do this?

Word Sense Disambiguation

- Predict the sense whose **definition** contains terms that co-occur often with those in the **surrounding context**

“I stopped by the **bank** to deposit some cash.”



An establishment for the custody, loan, and exchange of money

mutual information from IMDB corpus	money	raise	2.686
	debt	money	2.578
	dollars	money	2.567
	money	cash	2.546
	buy	money	2.471
	money	gambling	2.436
	money	pay	2.427
	money	bank	2.387
	insurance	money	2.117
	money	paid	2.018

Word Sense Disambiguation

in information retrieval

1. Expand the indexed vocabulary so that each **sense** of a word is a different index term
2. Automatically predict the correct sense for each word in the collection (e.g, **bank**¹, **bank**² , ... , **bank**⁶)
 - ▶ lots of context (i.e., surrounding text)
3. Index the collection as usual
4. At query-time, predict the correct word sense in the query (e.g., “drive-through **bank**⁴ carrboro”)
 - ▶ more difficult, not much context
5. Retrieve documents as usual

Word Sense Disambiguation in information retrieval

- Does it improve (average) retrieval effectiveness?

Word Sense Disambiguation in information retrieval

- Not much. Why not?

(Sanderson, 1996)

Word Sense Disambiguation

in information retrieval

- Not really a problem for long-queries (other query terms disambiguate the ambiguous ones)
- In theory, can improve performance for short queries
- However, these are precisely the queries for which disambiguation is the most difficult (not much context)

(Sanderson, 1996)

Word Sense Disambiguation in information retrieval

- There is another reason. What is it?

Word Sense Disambiguation

in information retrieval

united bank	national bank south carolina
union bank california	national bank oneida
union bank	national bank omaha
tyra banks show	national bank marin
star bank	national bank alaska
republic bank	national bank
pnc bank	merchants bank
people bank	loans bank account
outer banks north carolina	hotels outer banks nc
outer banks nc	hotels outer banks
online banking bank america	guaranty bank
national bank texas	freedom bank
commerce bank	farmers merchants bank

Word Sense Disambiguation

in information retrieval

- Wait for it..., Wait for it...

Word Sense Disambiguation

in information retrieval

- Word senses also (more or less) follow Zipf's law: a few are very frequent and most a rare

united bank
union bank california
union bank
tyra banks show
star bank
republic bank
pnc bank
people bank
outer banks north carolina
outer banks nc
online banking bank america
national bank texas
commerce bank

national bank south carolina
national bank oneida
national bank omaha
national bank marin
national bank alaska
national bank
merchants bank
loans bank account
hotels outer banks nc
hotels outer banks
guaranty bank
freedom bank
farmers merchants bank

Word Sense Disambiguation

in information retrieval

No. of senses	Size of set	Most common sense (%)	
2	3145	92	{50}
3	1697	85	{33}
4	1046	79	{25}
5	640	72	{20}
6	448	68	{17}
7	275	63	{14}
8	200	60	{13}
9	141	60	{11}
10	93	53	{10}

Table 10. Percentage of occurrences accounted for by the most common sense of a word.

The figures in brackets (shown for comparison) is the percentage that would result if senses occurred in equal amounts. Measurements made on the SEMCOR corpus.

(Sanderson, 1996)