
# Information Retrieval

## 01: Einführung und Motivation

Philipp Schaer, Technische Hochschule Köln, Cologne, Germany

Version: 2022-03-28

Technology
Arts Sciences
TH Köln

# Was ist eigentlich Retrieval?








# Was ist eigentlich Retrieval?

In diesem Kurs geht es um Information **Retrieval** und wir benutzen folgende, **vorläufige Arbeitsdefinition**:

„*Gegeben eine **Anfrage** und einen **Dokumentenkorpus**, finde **relevante** Dokumente.*“

- **Anfrage**: Eine Anfrage ist die Beschreibung eines **Informationsbedürfnisses**, die an das IR-System geschickt wird. Kann natürlichsprachig oder formal (Anfragesprache) sein.
- **Korpus**: Eine Sammlung von durchsuchbaren Dokumenten / Ressourcen. In unserem Falle meistens Textdokumente.
- **Relevanz**: Befriedigung des Informationsbedürfnisses eines Benutzers.

# Weitere Definitionen von IR

Drei Definitionen (von vielen)

- „Information retrieval (IR) is **finding material** (usually documents) of an **unstructured nature** (usually text) that **satisfies an information need** from within **large collections** (usually stored on computers).“ (Manning et al., 2008)
- „Information retrieval (IR) is a field concerned with the design, development, and evaluation of interactive systems that help users find information.“ (Arguello, 2017)
- “Information retrieval is a field concerned with the structure, analysis, organization, storage, and retrieval of information.“ (Salton, 1968)

# Weitere Definitionen von IR

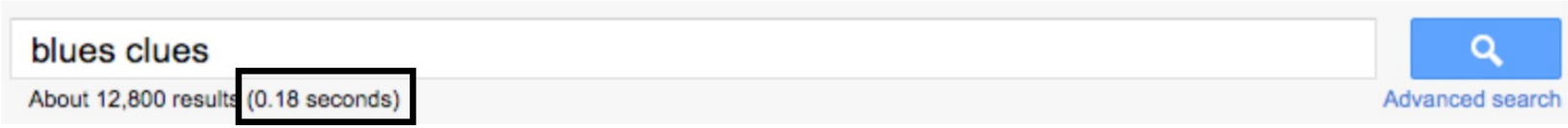
Besser!

Drei Definitionen (von vielen)

- „Information retrieval (IR) is **finding material** (usually documents) of an **unstructured nature** (usually text) that **satisfies an information need** from within **large collections** (usually stored on computers).“ (Manning et al., 2008)
- „Information retrieval (IR) is a field concerned with the design, development, and evaluation of interactive systems that help users find information.“ (Arguello, 2017)
- “Information retrieval is a field concerned with the structure, analysis, organization, storage, and retrieval of information.“ (Salton, 1968)

# Es geht also um die Nadel im ...

# Was gehört noch zum Thema Retrieval?




## Zwei zentrale Größen

- **Effizienz**: Wir wollen das Ergebnis noch vor Feierabend (oder noch besser: in 0,18 Sekunden)
- **Effektivität**: Liefere nur Ergebnisse, die Informationsbedürfnisse der Nutzer befriedigen

- Später: Klarer Fokus auf das Thema Effektivität.
- Allerdings: Wir werden uns auch darüber unterhalten, wie Suchmaschinen Ergebnisse möglichst schnell liefern können.

# Effizienz von Suche: `grep`

- Lorem ipsum dolor sit amet,
- consetetur sadipscing elitr,
- sed diam nonumy eirmod tempor,
- invidunt ut labore et dolore magna
- aliquyam erat, sed diam voluptua.
- At vero eos et accusam et justo
- duo dolores et ea rebum.

Finde alle Zeilen mit „et“.

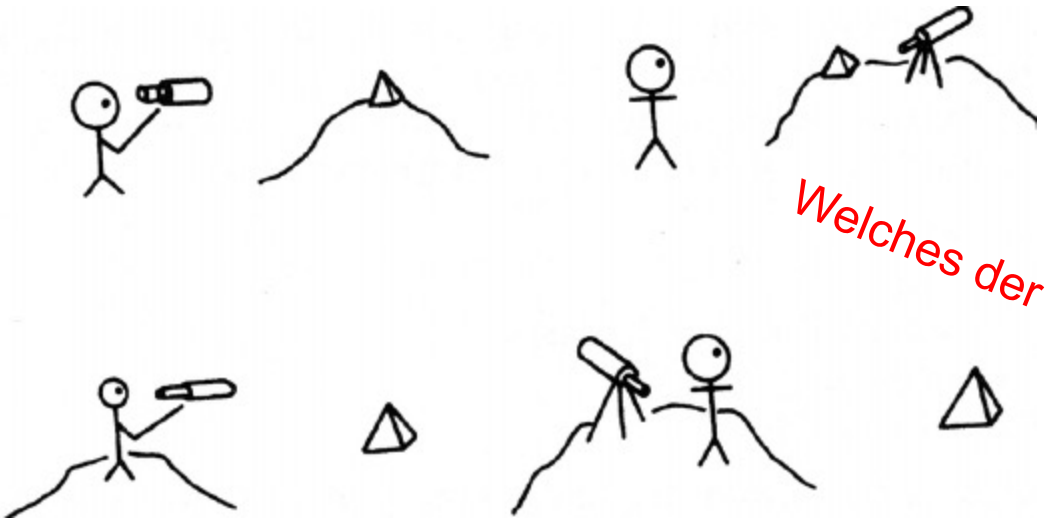
- Wie viele Schritte braucht `grep`?
- Ist das ein guter Weg zu suchen? Geht das nicht besser?

# Effektivität: Was ist eigentlich Relevanz?

Schwierig...

„The man saw the pyramid on the hill with the telescope.“

- Viele Interpretationen dieses Satzes sind denkbar...



Welches der Bilder ist relevant?

Bild aus Vickery & Vickery, 2004

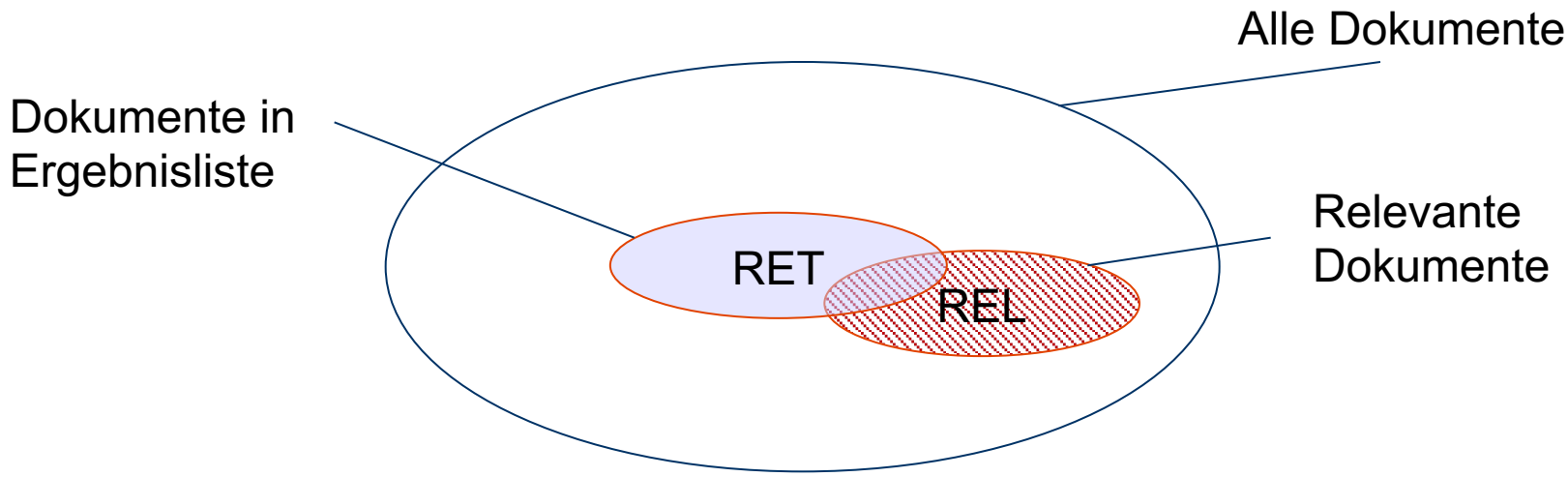
# Maßzahlen für die Evaluation

- **Precision** (Treffergenauigkeit)

$$\mathcal{P} = \frac{|\mathrm{RET} \cap \mathrm{REL}|}{|\mathrm{RET}|}$$

- **Recall** (Treffervollständigkeit)

$$\mathcal{R} = \frac{|\mathrm{RET} \cap \mathrm{REL}|}{|\mathrm{REL}|}$$



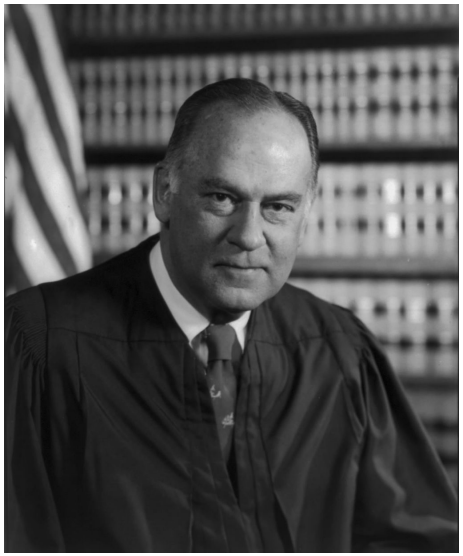
# Precision und Recall: Ein Beispiel

| | Relevant | Nicht relevant |
|---|---|---|
| Gefunden | 30 | 12 |
| Nicht gefunden | 14 | 44 |

- Precision $P = 30 / (30 + 12) \approx 0{,}714$
- Recall $R = 30 / (30 + 14) \approx 0{,}681$

# Was ist eigentlich Information?

Schwierig...



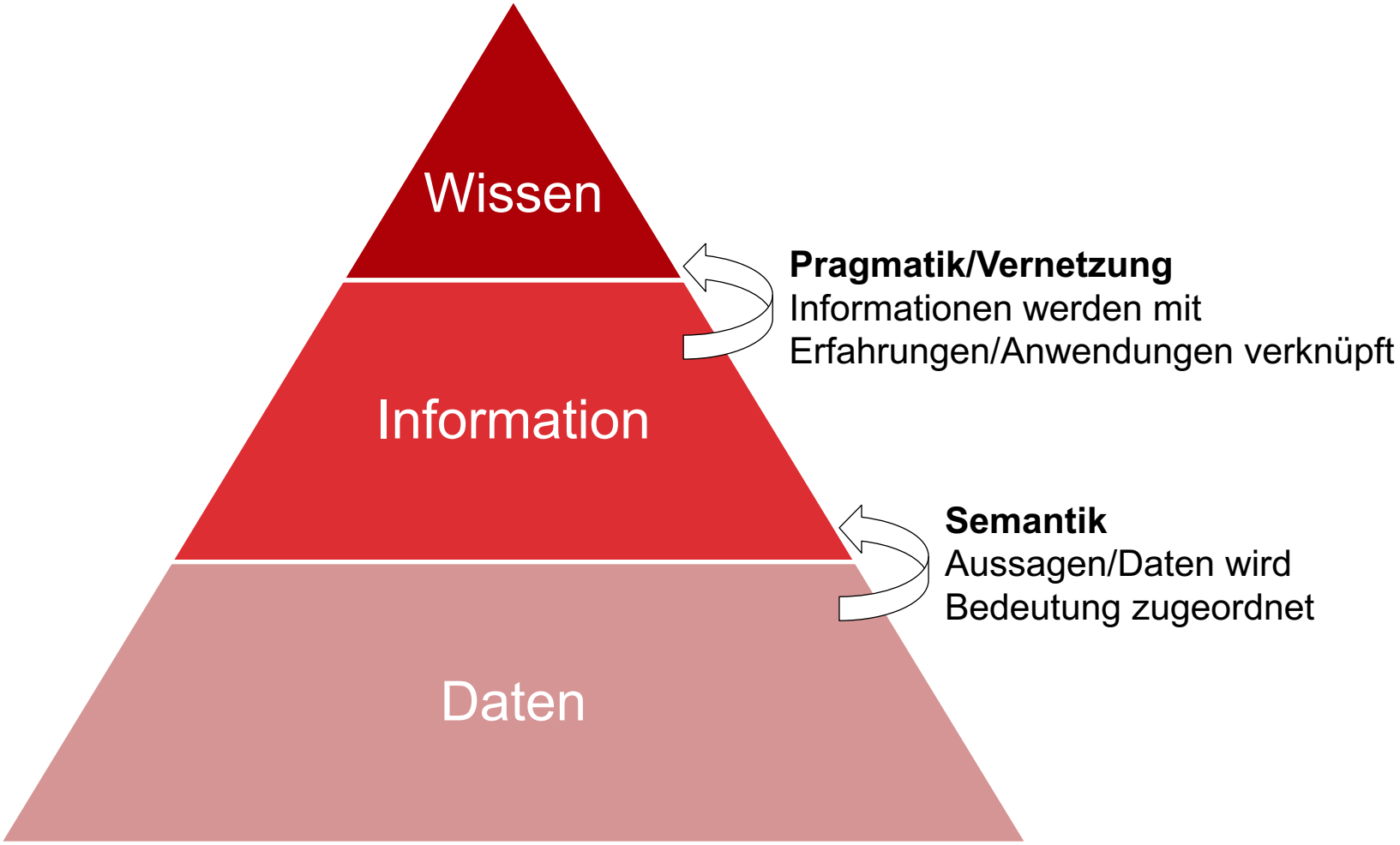
Häufig verwendet man den Begriff Information ganz intuitiv (und leider falsch)

- Analog zu Potter Stewart (1964): „I know it, when I see it“

Aufteilung in drei zentrale Begriffe:

- Daten
- Information
- Wissen

# Daten – Information – Wissen



# Strukturierte Daten

Strukturierte Daten sind z.B. Tabellendaten

| Angestellter | Boss | Gehalt |
|---|---|---|
| Berthold Heisterkamp | Bernd Stromberg | 50000 |
| Ulf Steinke | Bernd Stromberg | 60000 |
| Sinan Turçulu | Timo Becker | 50000 |

- Numerische Anfragen und Exact Match sind möglich, bspw.: ***Gehalt < 60000 AND Boss = Timo Becker***

- Toll, aber meistens nicht das was wir im Information Retrieval vorfinden → **Wir suchen in unstrukturierte Daten!**

# Unstrukturierte Daten…

Log in / create account

WIKIPEDIA
The Free Encyclopedia

Article | Discussion | Read | Edit | View history

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

## 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.

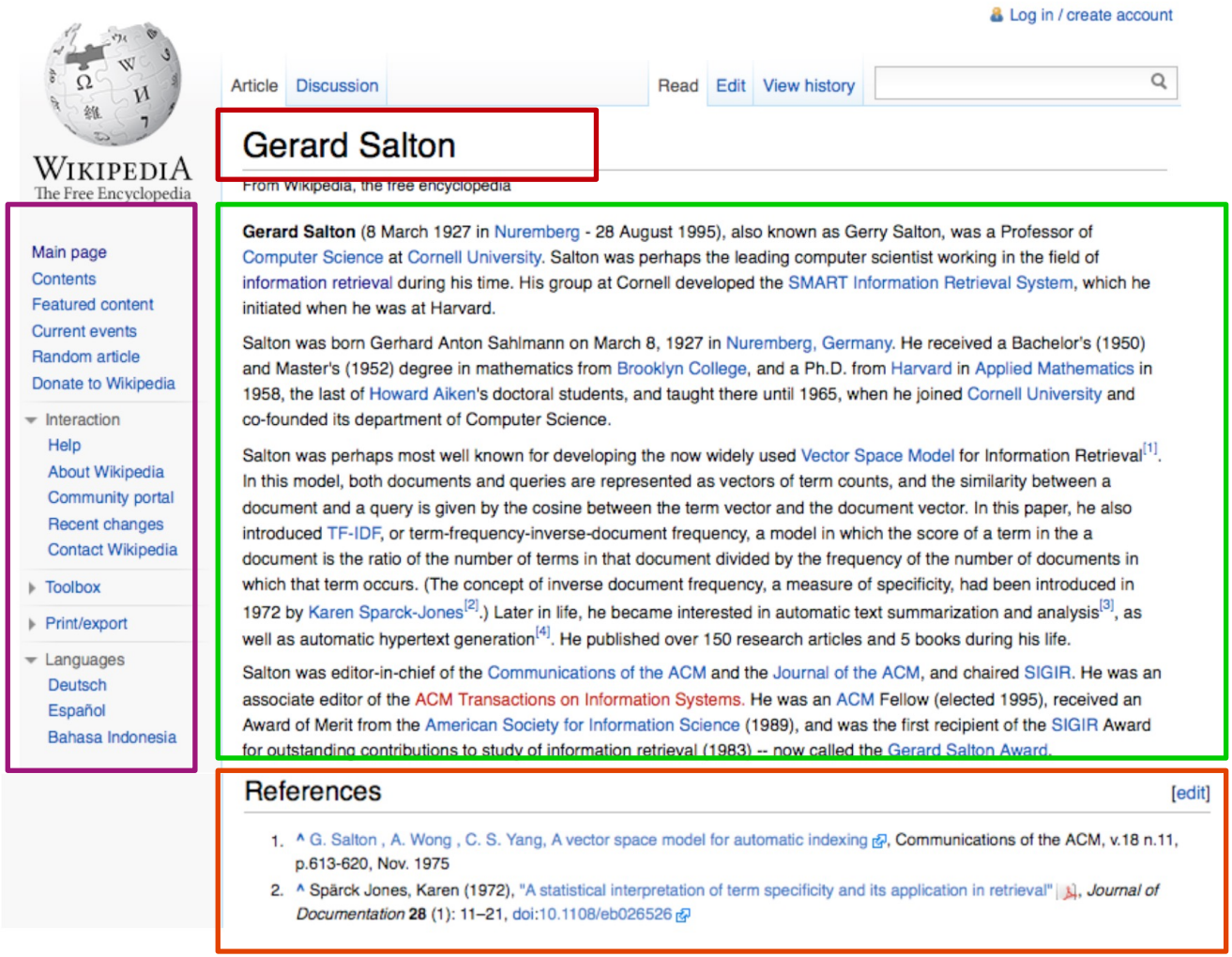
Salton was perhaps most well known for developing the now widely used Vector Space Model for Information Retrieval[1]. In this model, both documents and queries are represented as vectors of term counts, and the similarity between a document and a query is given by the cosine between the term vector 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.

## References [edit]

1. ^ G. Salton , A. Wong , C. S. Yang, A vector space model for automatic indexing, Communications of the ACM, v.18 n.11, p.613-620, Nov. 1975
2. ^ Spärck Jones, Karen (1972), "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation* **28** (1): 11–21, doi:10.1108/eb026526

# Das Dokument ist teilstrukturiert…



Log in / create account

WIKIPEDIA
The Free Encyclopedia

Article Discussion Read Edit View history

## Gerard Salton

From 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

**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.

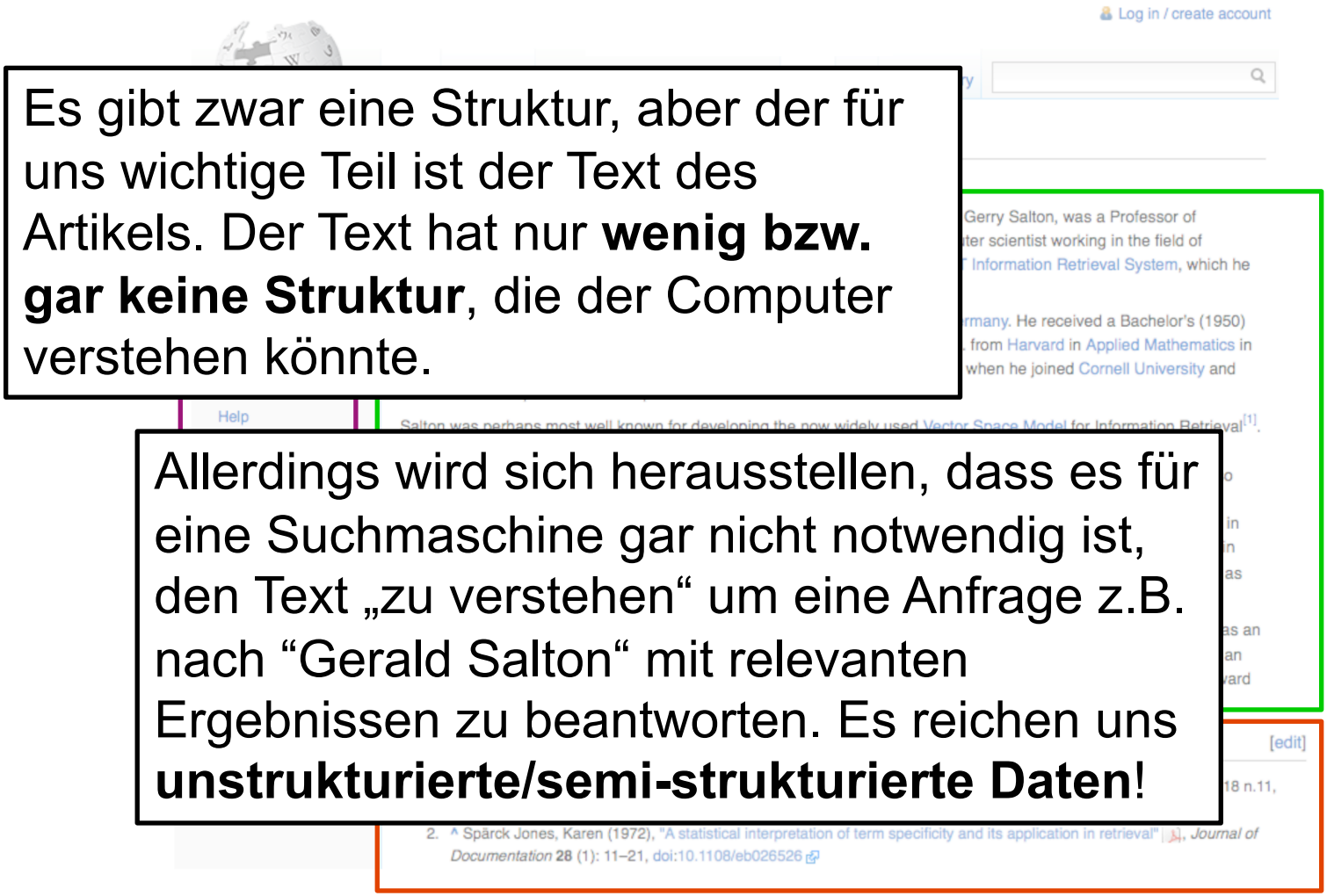
Salton was perhaps most well known for developing the now widely used Vector Space Model for Information Retrieval[1]. In this model, both documents and queries are represented as vectors of term counts, and the similarity between a document and a query is given by the cosine between the term vector 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.

### References [edit]

1. ^ G. Salton , A. Wong , C. S. Yang, A vector space model for automatic indexing, Communications of the ACM, v.18 n.11, p.613-620, Nov. 1975
2. ^ Spärck Jones, Karen (1972), "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation* **28** (1): 11–21, doi:10.1108/eb026526

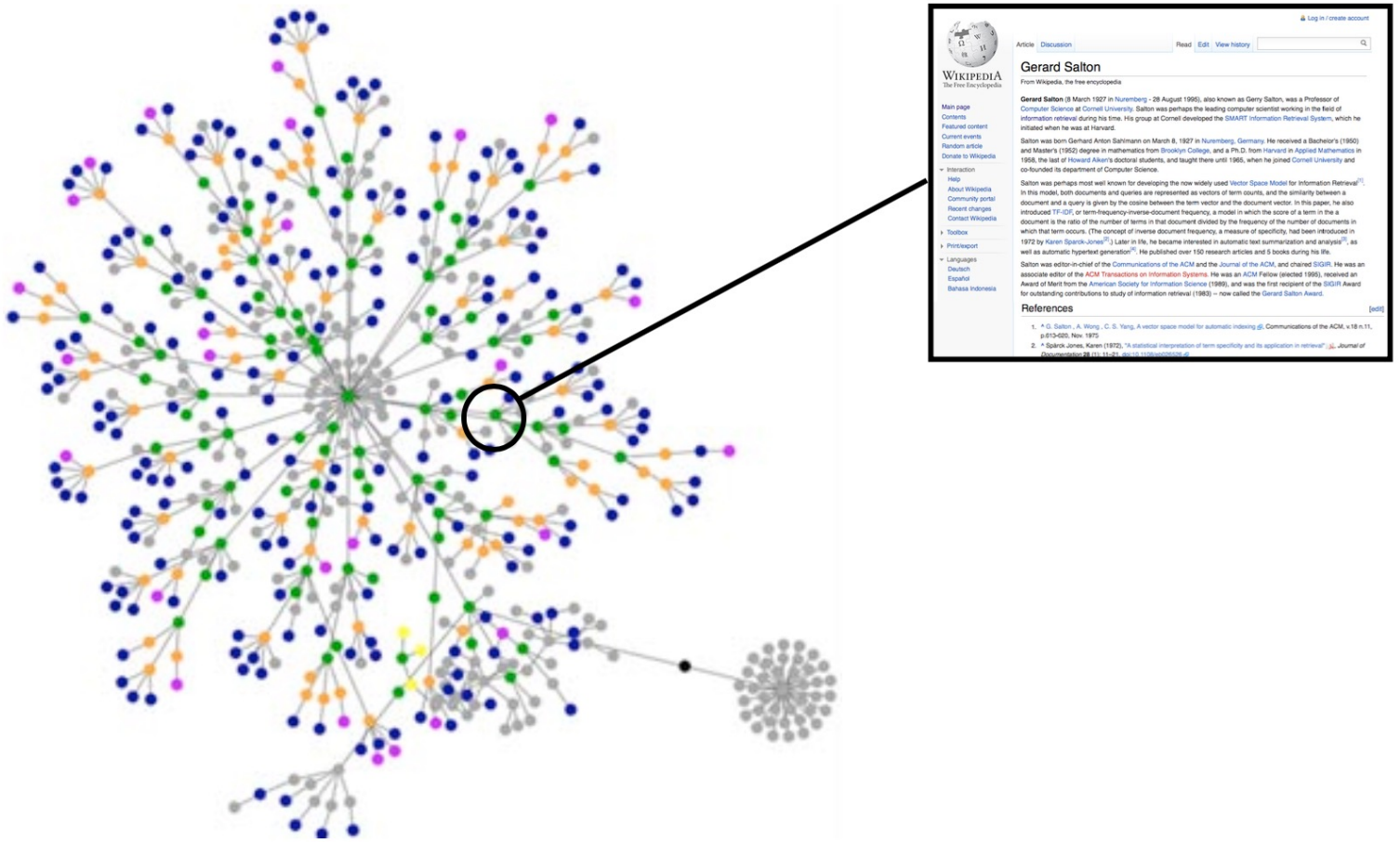
# Dokumentenstrukturen



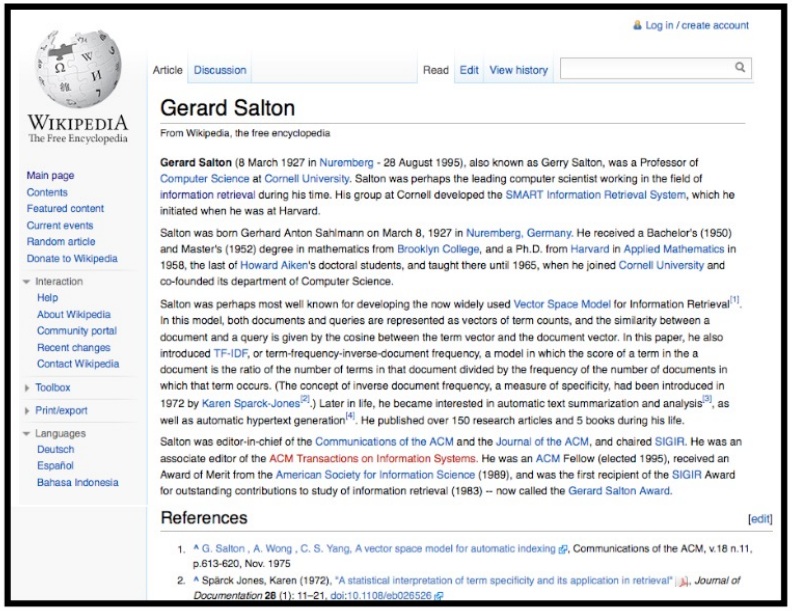

Es gibt zwar eine Struktur, aber der für uns wichtige Teil ist der Text des Artikels. Der Text hat nur **wenig bzw. gar keine Struktur**, die der Computer verstehen könnte.

Allerdings wird sich herausstellen, dass es für eine Suchmaschine gar nicht notwendig ist, den Text „zu verstehen“ um eine Anfrage z.B. nach “Gerald Salton“ mit relevanten Ergebnissen zu beantworten. Es reichen uns **unstrukturierte/semi-strukturierte Daten**!

# Kollektionsstrukturen



# Strukturen durch Klassifikationen



Log in / create account

Article | Discussion | Read | Edit | View history

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

## 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.

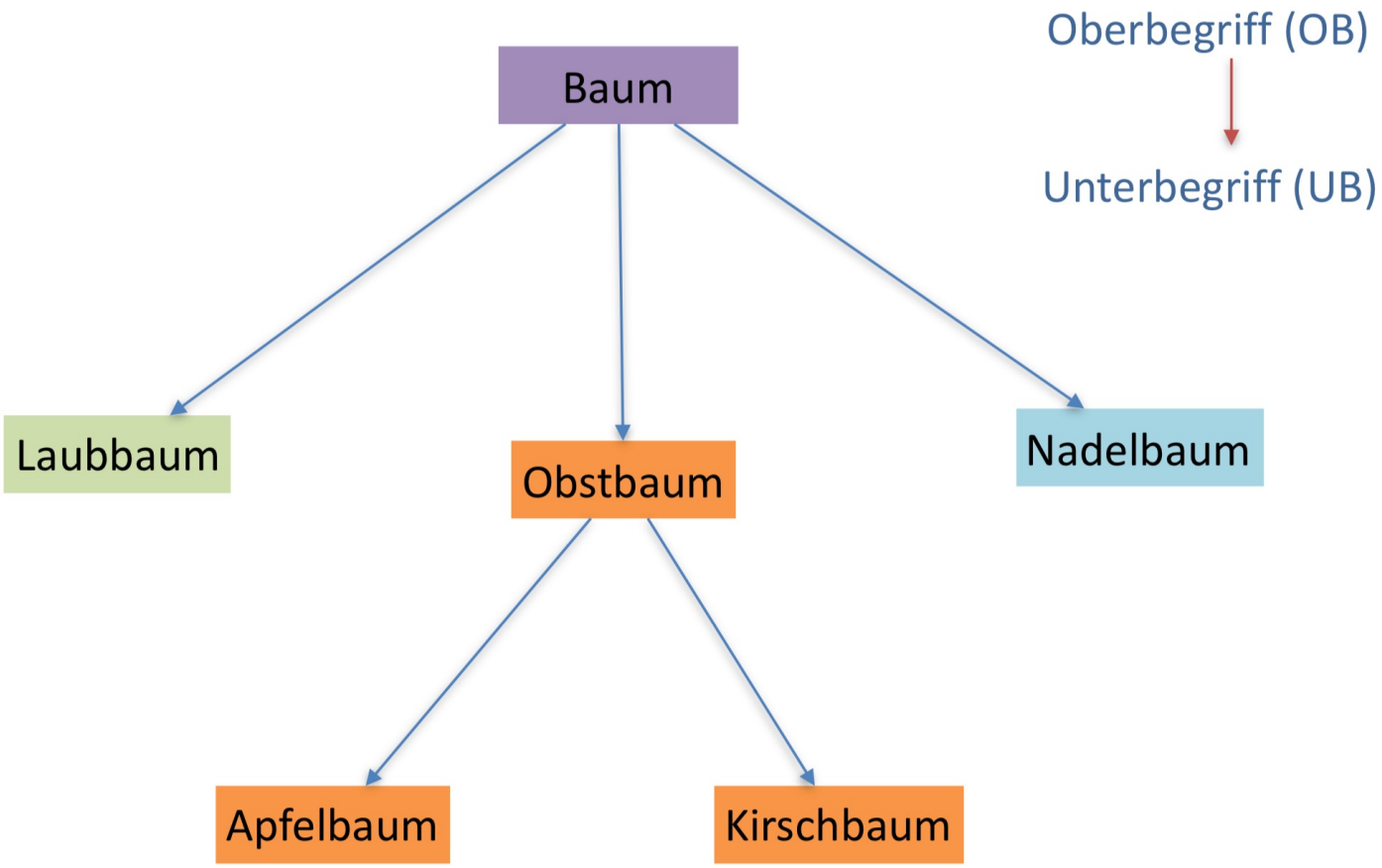
Salton was editor-in-chief of the Communications of the ACM and the Journal of the ACM, and chaired SIGIR. He was an associate editor of the ACM Transactions on Information Systems. He was an ACM Fellow (elected 1995), received an Award of Merit from the American Society for Information Science (1989),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

### References [edit]

1. ^ G. Salton , A. Wong , C. S. Yang, A vector space model for automatic indexing, Communications of the ACM, v.18 n.11, p.613-620, Nov. 1975
2. ^ Spärck Jones, Karen (1972), "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation* **28** (1): 11–21, doi:10.1108/eb026526

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

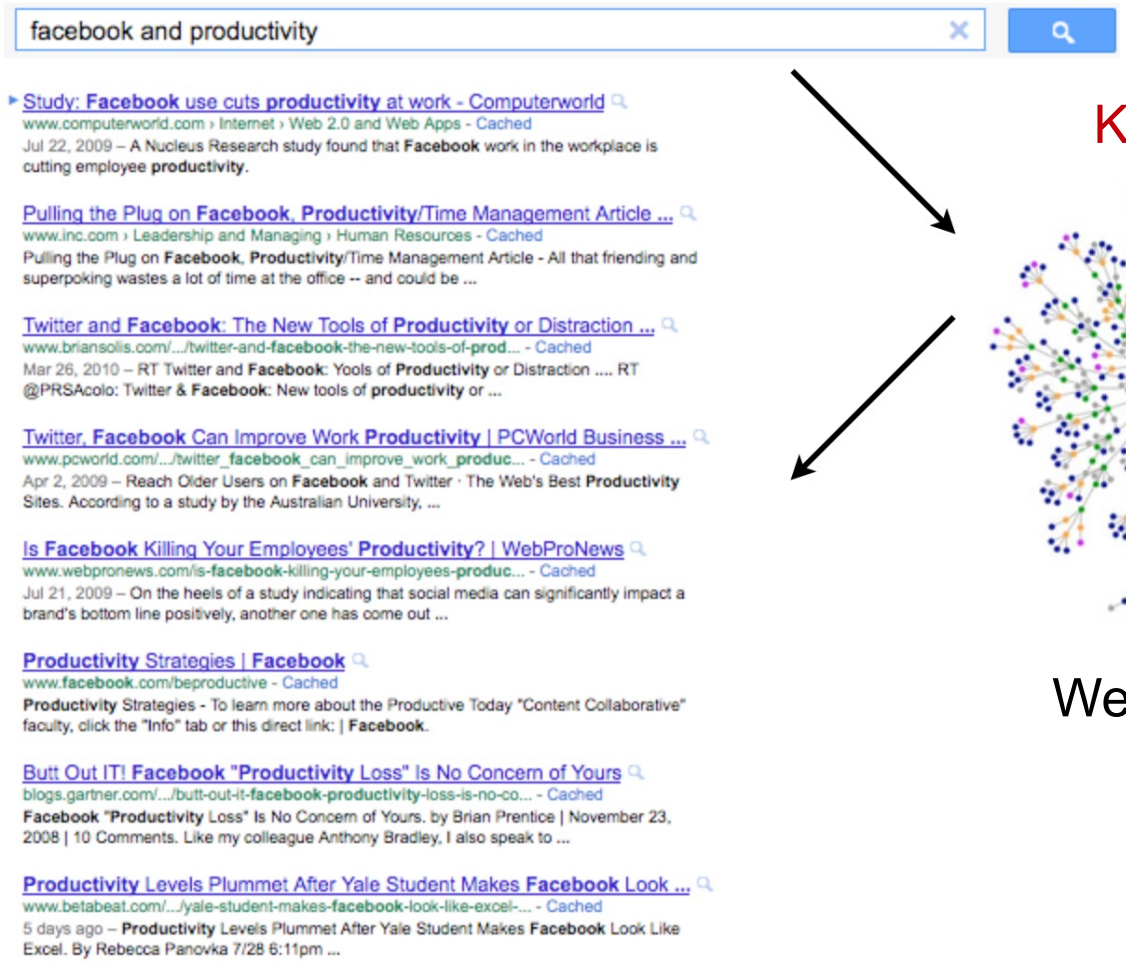
# Weitere Strukturen...



# IR für Webseiten

Anfrage

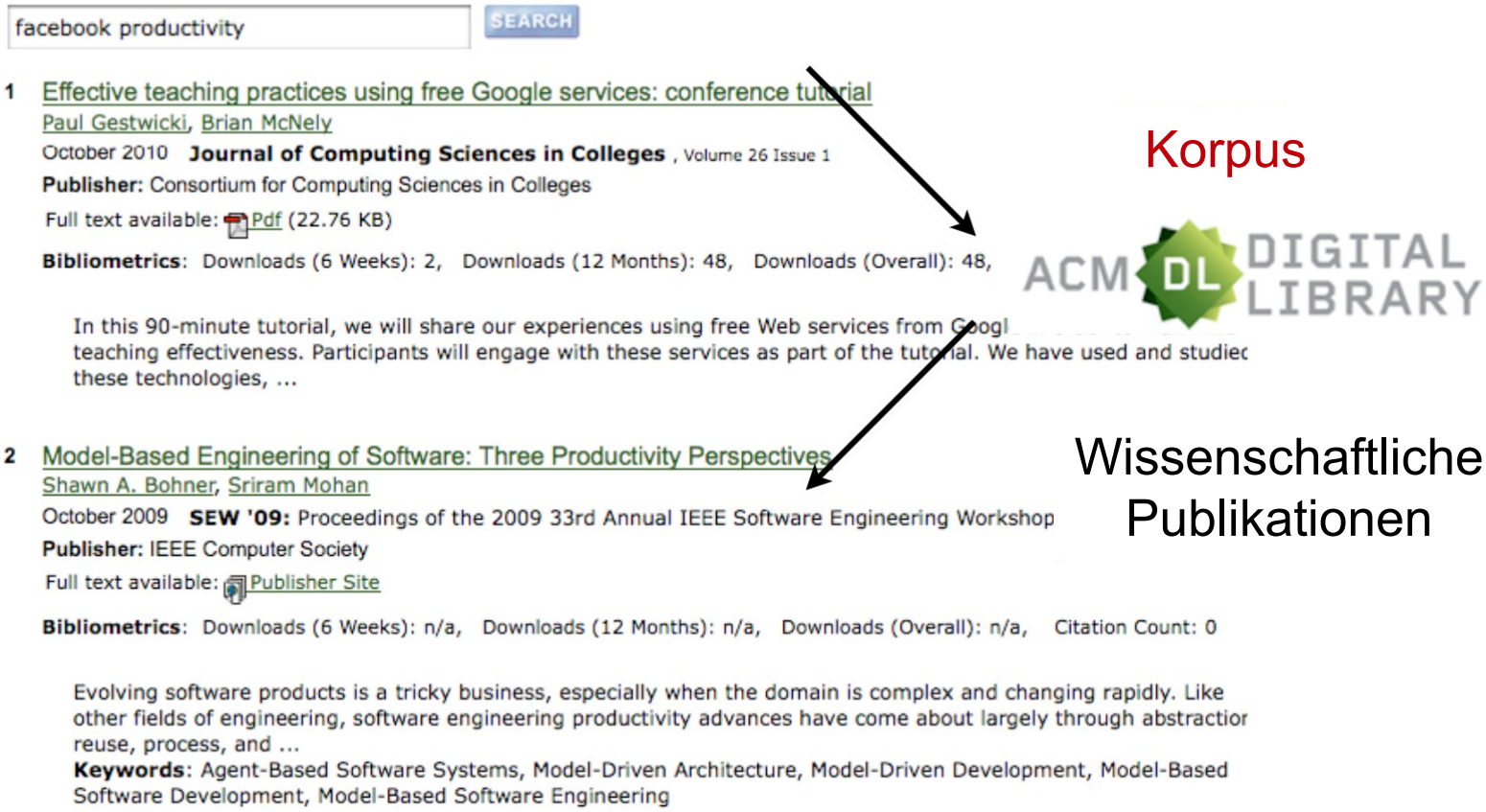
Ergebnis



# IR für digitale Bibliotheken

Anfrage

Ergebnis




Korpus

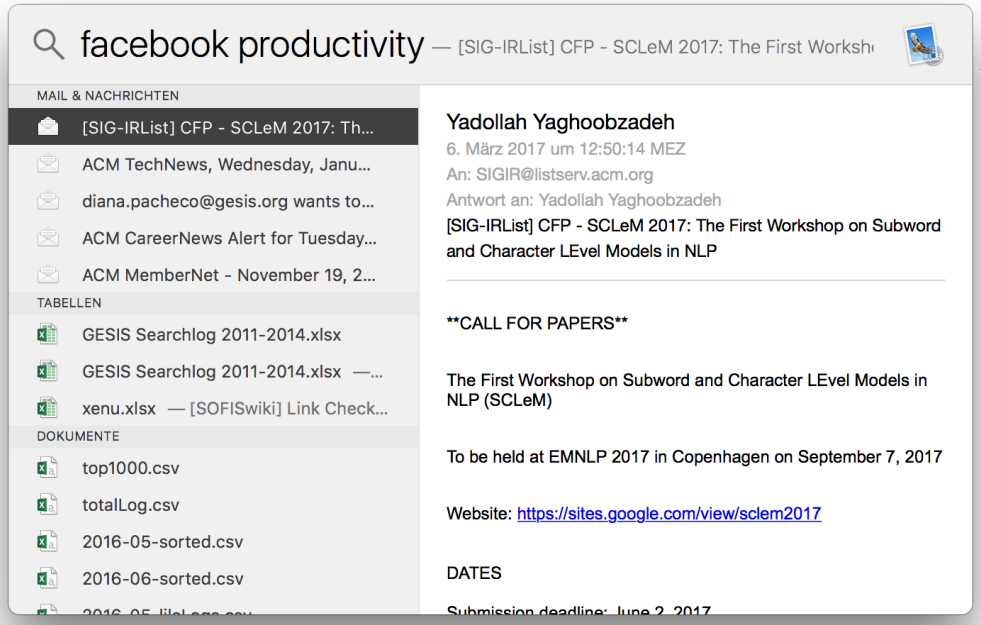
ACM DL DIGITAL LIBRARY

Wissenschaftliche Publikationen

# IR für den Desktop

Anfrage

Ergebnis




Korpus

Dateien auf meinem Computer

# Anfragen sind nur eine Krücke

Eine Suchanfrage ist nur eine verkürzte Beschreibung des **Informationsbedürfnisses** des Nutzers.

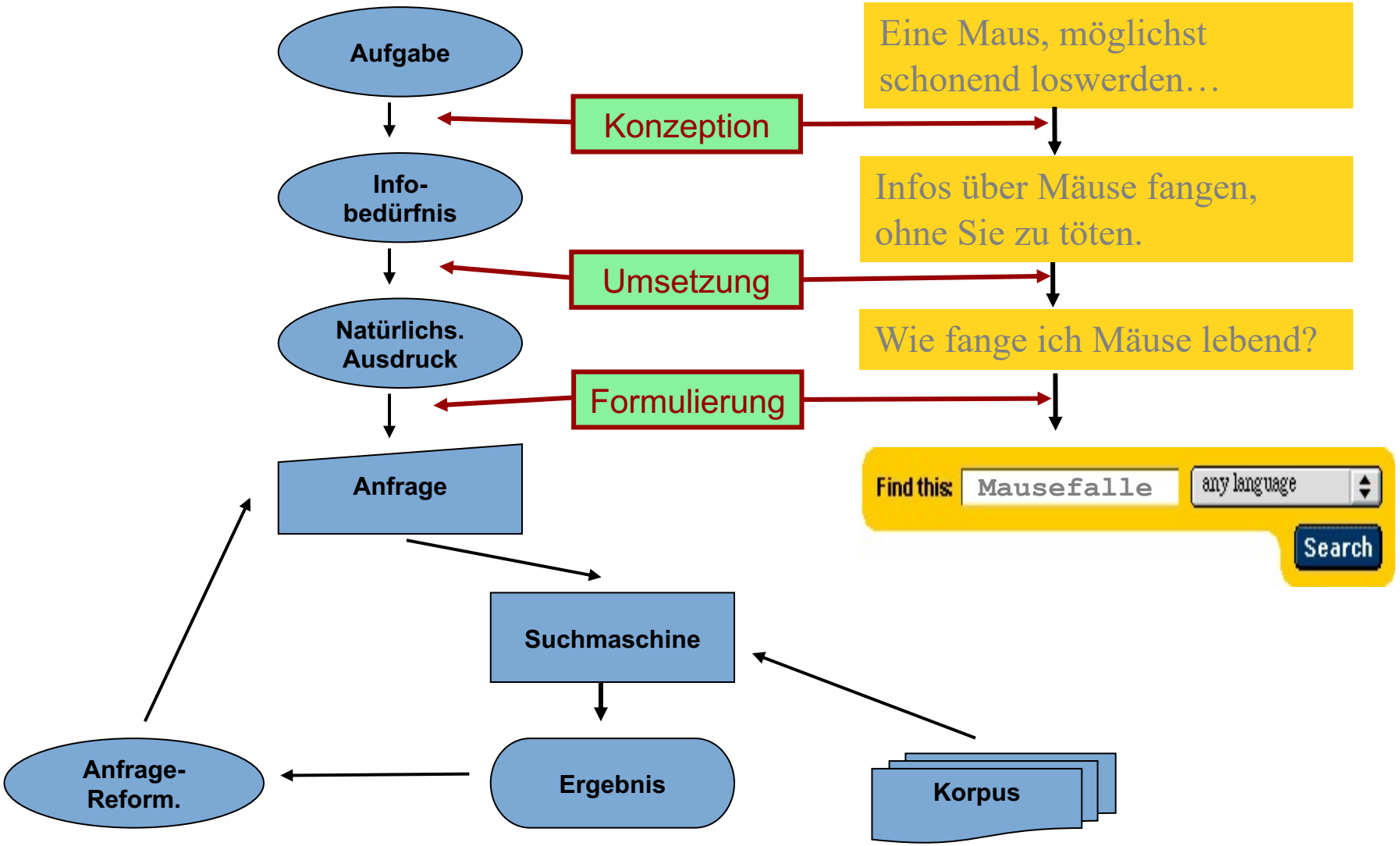
- Mehrdeutig und eigentlich nur für den Nutzer selbst verständlich
- Eine Anfrage und ein Informationsbedürfnis setzt eine konkrete Aufgabe voraus
- Der Nutzer befindet sich in einem konkreten Kontext, den wir nicht kennen

# AOL-Logdateien – User 1515830

*chai tea calories*
*calories in bananas*
*aftermath of incest*
*how to tell your family you're a victim of incest*
*pottery barn*
*curtains*
*surgical help for depression*
*oakland raiders comforter set*
*can you adopt after a suicide attempt*
*who is not allowed to adopt*
*i hate men*
*medication to enhance female desire*
*jobs in denver colorado*
*teaching positions in denver colorado*
*how long will the swelling last after my tummy tuck*
*divorce laws in ohio*
*free remote keyloggers*
*baked macaroni and cheese with sour cream*
*how to deal with anger*
*teaching jobs with the denver school system*
*marriage counseling tips*
*anti psychotic drugs*

http://www.cnet.com/news/aols-disturbing-glimpse-into-users-lives/

# Anfragen sind nur eine Krücke

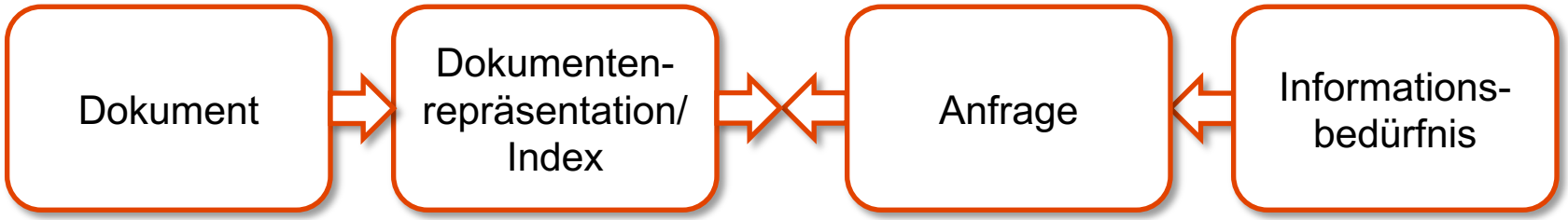


# Klassisches Information Retrieval-Modell

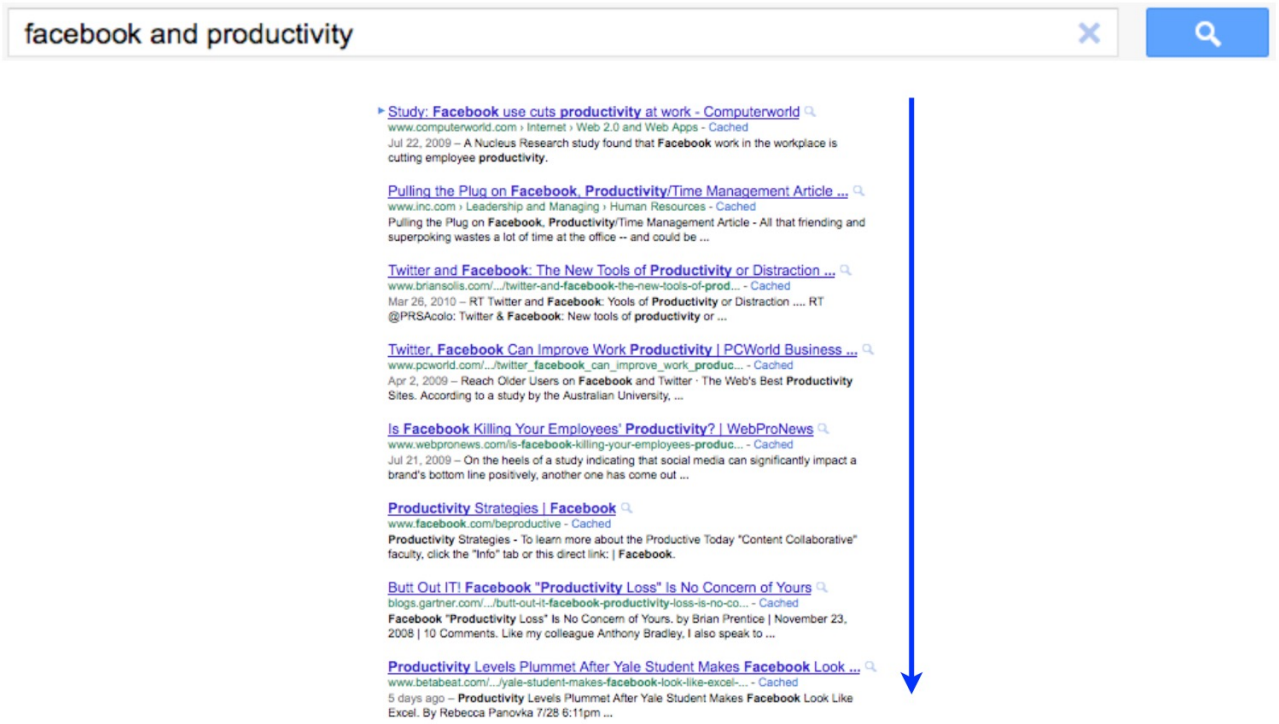
Das klassische **Ad-Hoc-Retrieval** basiert auf Abgleich von

- Dokumenttermen (Document Representation) und
- Anfragetermen (Query).

Im klassischen Information Retrieval-Modell sind das Informationsbedürfnis als auch die Anfrage starr und verändern sich nicht.



# Retrieval als Suchaufgabe




- **Ausgabe**: Ein **Ranking** von Dokumenten, in absteigender Reihenfolge Ihrer geschätzten Relevanz (macht es einfacher!).
- **Annahme**: Der Benutzer schaut sich die **ersten paar Dokument** an und ist zufrieden, wenn er etwas passendes gefunden hat.

# Ranking – Welches ist besser?


- Relevante Treffer sind rot markiert.

# Warum ist IR eine schwierige Aufgabe?

Information Retrieval ist ein Prozess mit **Unsicherheiten**...

- Benutzer **wissen nicht was sie eigentlich wollen**
- Benutzer wissen nicht, wie sie das was sie suchen **ausdrücken sollen**
- Computer können Nutzer keine **Kontextinformationen** entlocken, wie es z.B. ein menschlicher Bibliothekar könnte
- Computer verstehen **keine natürliche Sprache**
- Suchmaschinen müssen **erraten, was relevant ist**
- Suchmaschinen müssen erraten, wann ein Benutzer **zufrieden** ist
- ...

# Suchanfragen, Relevanz, Ranking...

Dies alles macht Information Retrieval so **schwer** – und **faszinierend**!

- Eigentlich müssten wir viel mehr „verstehen“, was der Nutzer will.

Im IR geht es aber nicht darum Text zu verstehen, sondern darum „zu bestimmen“/„zu berechnen“/„vorherzusagen“ ob ein Dokument zu einer Anfrage relevant ist – oder nicht.

- Das ist in vielen Fälle **viel einfacher**, als „verstehen“.
- Es ermöglicht uns, das Problem an **Computer abzugeben**, z.B. in Form von Suchmaschinen!
- Computer arbeiten mit **Modellen**, um Informationen zu finden, z.B. dem **booleschen Modell**...
- **Reicht aber oft genug aus!**

# Zusammenfassung

Das Ziel des Information Retrieval ist es Informationssuchenden die **passenden Informationen** zu liefern, die sich **benötigen**!

- IR beinhaltet die **Analyse**, **Speicherung** und die eigentliche **Suche**
- Es gibt viele **verschieden Arten** von Suchmaschinen
- Es gibt viel **Unwägbarkeiten** und **Unsicherheiten** im IR
- Daher machen IR-Systeme viele **Annahmen**, **Schätzungen** und **Vermutungen**, um ein passendes Ergebnis zu liefern
- Nutzer erwarten ganz unterschiedliche Dinge als Ergebnis, abhängig von ihrer **Suchaufgabe** und ihrem **Kontext**
- Die **Evaluation von IR-Systemen** ist komplex und setzt ein Verständnis des Nutzerbedürfnisses voraus

Mein Ziel ist es, Ihnen in dieser Veranstaltung all diese näher zu bringen und Sie davon zu überzeugen, dass IR spannend ist!