

# Information Retrieval

Modul MI-IR-M



Prof. Dr. Andreas Henrich

Lehrstuhl für Medieninformatik

Fakultät für Wirtschaftsinformatik und Angewandte Informatik

Otto-Friedrich-Universität Bamberg, 96045 Bamberg

Telefon: 0951 863 2850 oder 2851 (Sekretariat)

Email: [andreas.henrich@uni-bamberg.de](mailto:andreas.henrich@uni-bamberg.de)

WWW: <http://www.uni-bamberg.de/minf>

# Literatur

## ■ Grundlage der Vorlesung:

- Croft, W Bruce; Metzler, Donald; Strohan, Trevor: *Search Engines: Information Retrieval in Practice*. International edition. Boston: Addison-Wesley / Pearson, 2009.

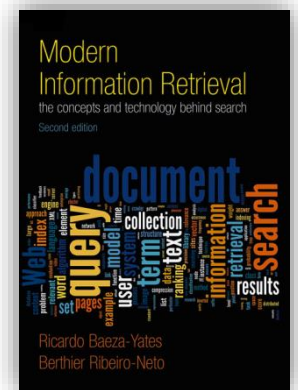
Online unter: <https://ciir.cs.umass.edu/irbook/>

- Henrich, Andreas: *Information Retrieval 1 (Grundlagen, Modelle und Anwendungen)*. Lehrtext, Universität Bamberg, 2008,

Online: [http://www.uni-bamberg.de/minf/ir1\\_buch/](http://www.uni-bamberg.de/minf/ir1_buch/)

## ■ Weitere Literatur zum Thema:

- Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier (2011): *Modern Information Retrieval*. 2. Auflage. Harlow: Pearson Education Limited, 2011.



1. Suchmaschinen und Information Retrieval
2. Architektur von Suchmaschinen
3. Evaluierung von Suchmaschinen
4. Retrieval-Modelle
5. Ranking mit Indexstrukturen
6. Textverarbeitung
7. Anfragen / Benutzerschnittstellen / Interaktion
8. Crawling und Texterfassung
9. Suchmaschinenoptimierung, Werbung, ...
10. Bilder und vertikale Suchlösungen

Grundlage der Vorlesung:

Croft, W Bruce; Metzler, Donald;  
Strohman, Trevor: Search Engines:  
Information Retrieval in Practice.  
International edition. Boston:  
Addison-Wesley / Pearson, 2009.

- Die **Suche** (im Internet) ist eine **tägliche Aktivität** für viele:
  - **Suche** und **Kommunikation** sind die **populärsten** Nutzungsformen des Computers
  - Anwendungen, die **Suchfunktionen** beinhalten, sind **allgegenwärtig** (Websuche, Bibliothekssuche, Desktop-Suche ...)
- Das **Teilgebiet der Informatik**, das sich mit **Forschung und Entwicklung** von Suchmaschinen (nicht nur für das Internet) beschäftigt, nennt man **Information Retrieval (IR)**



- *“Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.” (Salton, 1968)*



Gerard Salton (\* 8. März 1927 in Nürnberg; † 28. August 1995)  
Bildquelle: [http://www.cs.cornell.edu/annual\\_report/1996/beginning/salton.html](http://www.cs.cornell.edu/annual_report/1996/beginning/salton.html)

# IR: Motivation und Sichtweisen



## ■ Motivationen für IR:

- immer größere Datenmengen
- Vielfalt hinsichtlich Struktur und verwendeter Medien
- Nutzung von Information stark durch soziale Interaktionen geprägt

⇒ erhöhte Anforderungen an IR-Systeme

⇒ Bedarf für elaborierte Suchalgorithmen und wissensbasierte, computerlinguistische und medienspezifische Erschließungsverfahren

## ■ Sichtweisen auf IR:

- klassische systemorientierte Sichtweise („Suchmaschinen bauen“)
- Algorithmen (z. B. für die schnelle Suche)
- mathematische Modelle (z. B. für „Relevanz“)
- Ansätze aus den Bereichen der Benutzermodellierung und der Kognitionswissenschaft

<https://fg-retrieval.gi.de/>; Stand: 11.10.2019

# Fragestellungen des IR:



1. Datenmodellierung und Wissensrepräsentation,
2. Modelle für Information Retrieval (kognitive Modelle, Retrievalmodelle),
3. Gestaltung von Benutzerschnittstellen für gezielte und für assoziative Informationssuche,
4. Implementierung von IR-Systemen, insbesondere unter Berücksichtigung neuer technischer Entwicklungen,
5. Benutzernahe Evaluierung von Retrievalverfahren und -anwendungen,
6. Anwendung von IR-Systemen: Probleme, Lösungen, praktische Erfahrungen,
7. ...

<https://fg-retrieval.gi.de/ueber-die-fachgruppe/>; Stand: 11.10.2019

## Was ist ein Dokument?

### ■ Beispiele für Dokumente:

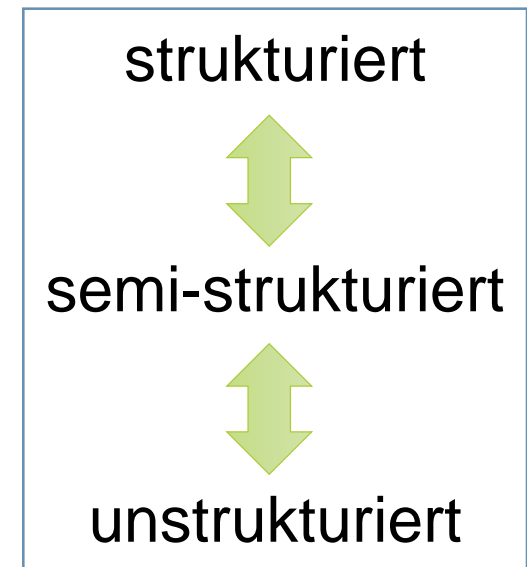
Webseiten, E-Mails, Bücher, Nachrichten, wissenschaftliche Papiere, SMS, Word™-Dokumente, Powerpoint™-Präsentationen, PDFs, Forenbeiträge, Patentschriften, Instant Messaging Dialoge, usw.

### ■ Gemeinsame Eigenschaften

- Signifikanter Textinhalt
- Eine gewisse Struktur  
(z. B. Titel, Autor, Datum für Artikel, Betreff, Absender, Empfänger für E-Mail)



- **Datenbankeinträge** (Tupel in relationalen Datenbanken) bestehen typischerweise aus eindeutigen Feldern (Datenbankattribute)
  - z. B. Bankdaten mit Kontonummern, Kontostand, Namen, Adressen, Geburtsdatum, usw.
- Es ist einfach, Anfragen über Feldern mit **eindeutiger Semantik** durchzuführen
- Bei **Textdokumenten** gestalten sich Anfragen schwieriger



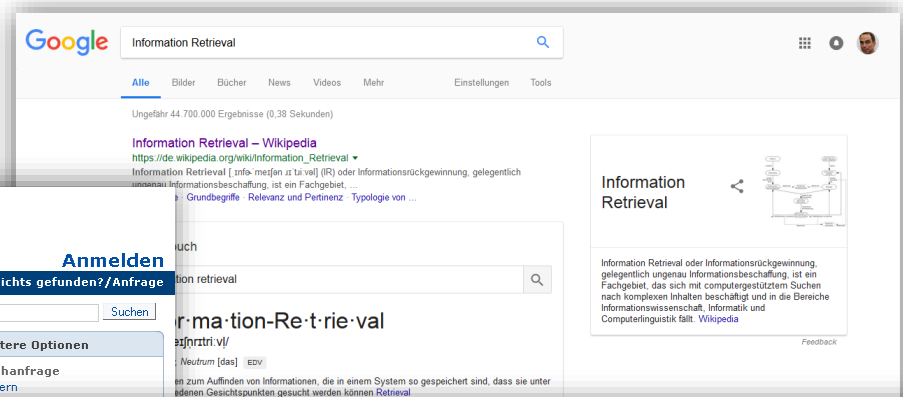
- Kernaspekte des Information Retrieval:
  - Anfrage(text) mit Text des Dokuments „vergleichen“ sowie
  - Entscheidung, was eine gute Übereinstimmung (match) ausmacht
- Exakte Übereinstimmung von Wörtern ist nicht ausreichend:
  - Viele verschiedene Möglichkeiten, etwas in einer natürlichen Sprache wie Englisch oder Deutsch auszudrücken.

- IR ist mehr als
  - nur Text und
  - mehr als nur Websuche,  
obwohl dies zentrale Aspekte sind.
  
- Menschen, die sich mit IR beschäftigen, arbeiten mit
  - verschiedenen Medientypen,
  - verschiedenen Arten von Suchapplikationen und
  - verschiedenen Aufgaben.

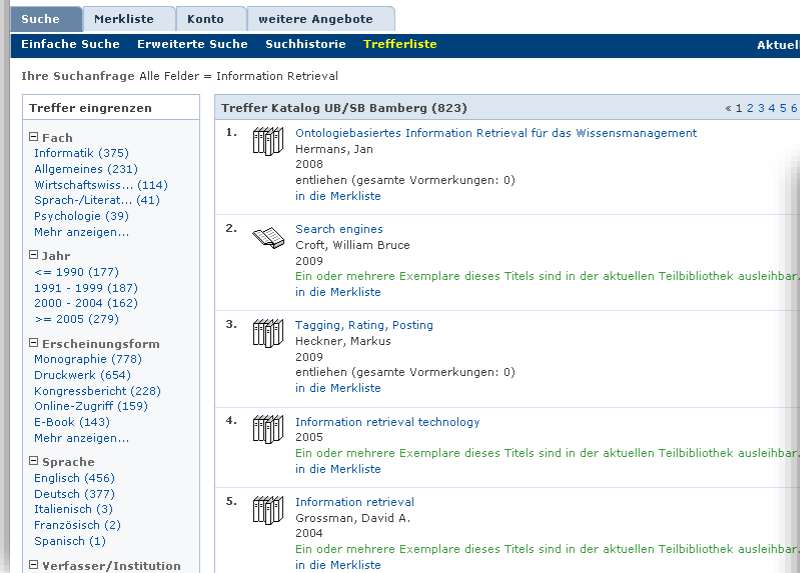
# Suchmaschinen und Information Retrieval



Suchen im Internet:  
Beispiel <http://www.google.de/>



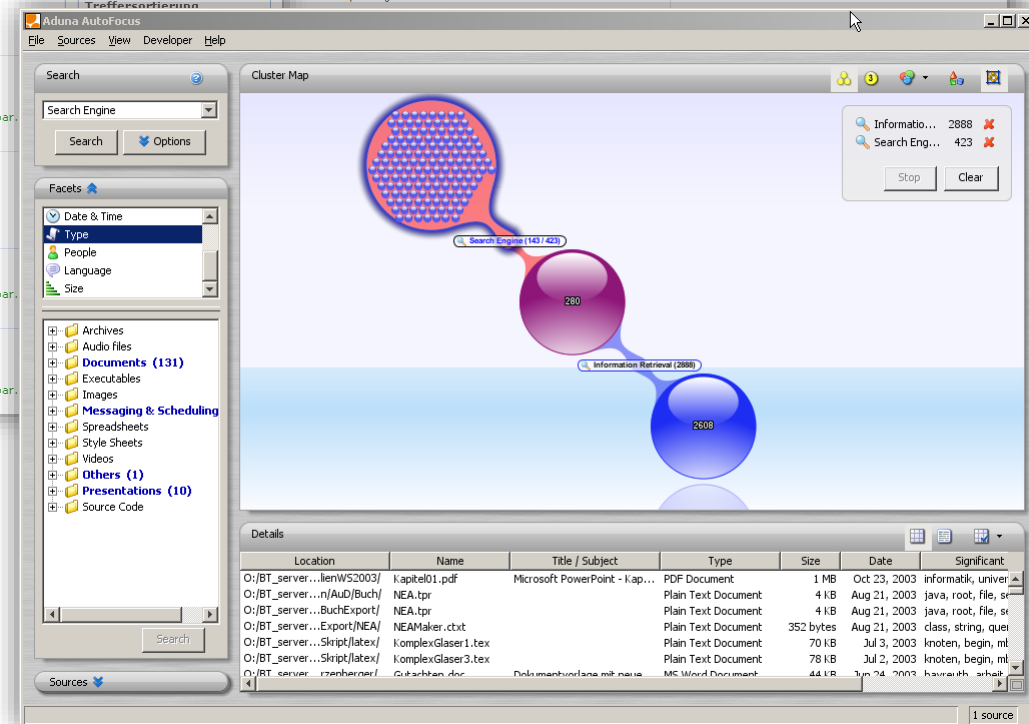
## Bamberger Katalog



Suchen in einem (Bibliotheks-)Katalog:  
Beispiel: Katalog der TB3 der Uni Bamberg

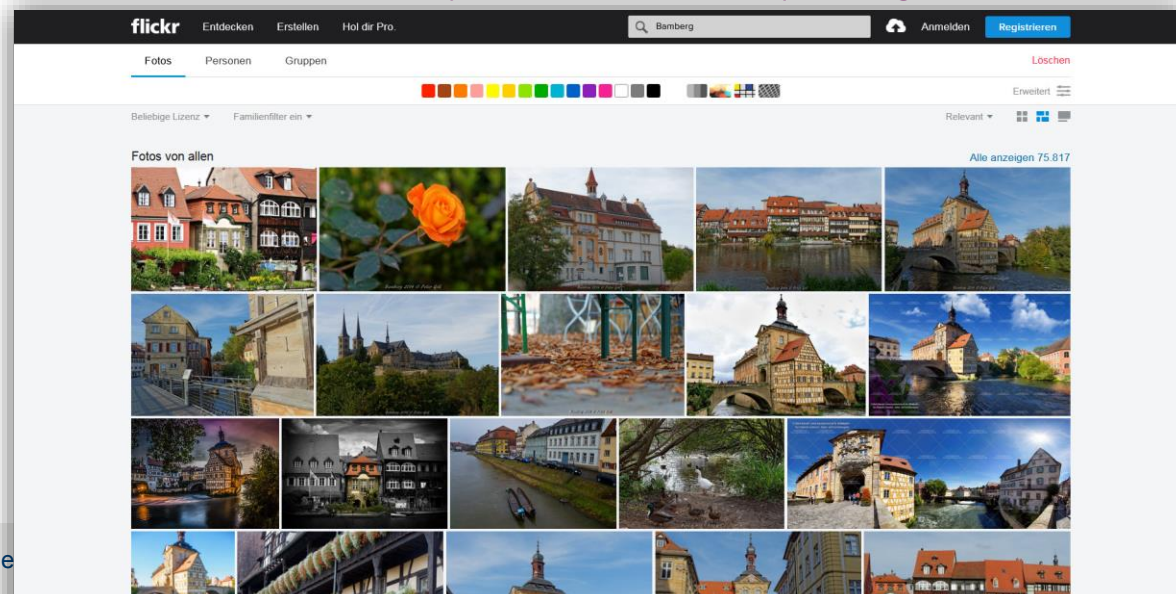
Suchen „auf dem Desktop“

Beispiel: <http://www.aduna-software.com/>



- Neue Applikationen betrachten vermehrt „neue Medien“
  - z. B. Videos, Fotos, Musik, Sprachaufnahmen, 3D-Modelle ...
- Noch stärker als bei Text ist der Inhalt schwierig zu beschreiben und zu vergleichen!
  - Manchmal kann man Text zur Repräsentation nutzen
  - z. B. in Form von Tags, ALT-Texten ...

<http://www.flickr.com/search/?q=Bamberg&ct=0&mt=all&adv=1>



# Die semantische Lücke



{Lehrstuhlflug, Medieninformatik,  
Wiesent, Kentern ...}

{Krise, Managergehälter, Abschwung,  
Chefeinkommen, Wirtschaft ,  
Autoindustrie, Gehaltsreport,  
Branchen, Geld ...}



Kleine semantische Lücke von Text zu  
beschreibenden Schlüsselwörtern



Große  
semantische  
Lücke von Pixeln  
/ Farbwerten zu  
beschreibenden  
Schlüsselwörtern

## Lohnstudie

### Wie die Krise Managergehälter drückt

Der Abschwung hat die Chefeinkommen in der deutschen Wirtschaft geschmälert, vor allem in der Autoindustrie mussten Top-Kräfte Abstriche machen. Der Gehaltsreport des manager magazin zeigt, welche Branchen ebenfalls betroffen waren - und in welchen Positionen noch richtig Geld verdient wird. Von *Anne Spies und Klaus Werle* mehr... [ Forum ]



## Inhalt

Text

Bilder

Video

Gescannte Dokumente

Audio

Musik

3D-Modelle

Bewegungssequenzen

Gensequenzen

## Applikationen

Websuche

Vertikale Suche  
(nur eine Domäne)

Enterprise Search

Desktopsuche

Foren-/Blogsuche

P2P-Suche

Literatursuche

## Aufgaben

Ad-hoc Suche

Filtern

Klassifikation

Beantworten von Fragen



- Ad-Hoc Suche
  - Finde relevante Dokumente für eine beliebige Textanfrage.
  
- Filtern (Bsp.: nächste Folie)
  - Identifiziere relevante Benutzerprofile für ein neues Dokument.



# Filtern: Beispiel Alerts (Filtern von neuen „Seiten“ im Web)



The screenshot shows the Google Alerts interface. At the top left is the Google logo. The main heading is 'Alerts' with the subtitle 'Interessanten neuen Inhalten im Web folgen'. A search bar contains the text 'Information Retrieval'. Below the search bar are several filter options, each with a dropdown menu:

- Häufigkeit: Höchstens einmal täglich
- Quellen: Automatisch
- Sprache: Deutsch
- Region: Beliebige Region
- Anzahl: Nur die relevantesten Ergebnisse
- Senden an: AH.postfach@googlemail.com

At the bottom of the filter section are two buttons: 'Alert erstellen' (highlighted in blue) and 'Optionen ausblenden'. Below the filters is a section titled 'Alert-Vorschau' with the text: 'Für deine Suchanfrage sind keine aktuellen Ergebnisse vorhanden. Unten findest du frühere Ergebnisse, die deiner Suchanfrage entsprechen.' Underneath, there is a section for 'Nachrichten' with a preview of a message: 'Kooperation von Uni Bonn und ZB MED trägt Früchte Informationsdienst Wissenschaft (Pressemitteilung)'. The Google interface also shows a grid icon, a notification bell, and a user profile picture in the top right corner.

<https://www.google.de/alerts>

- Ad-Hoc Suche
  - Finde relevante Dokumente für eine beliebige Textanfrage.
- Filtern (Bsp.: nächste Folie)
  - Identifiziere relevante Benutzerprofile für ein neues Dokument.
- Klassifikation (Bsp.: nächste Folie)
  - Identifiziere **relevante Kennzeichen** eines Dokuments.
  - Bilde Gruppen und verschaffe so einen **Überblick**.



Web News Images Video



Results 1-20 of 29 in Indonesia

[Sources](#) [Sites](#) [Time](#) [Topics](#)

Top 555 Clusters

+ Downloads (52)

+ Tutorial (36)

+ Free (55)

- **Indonesia (29)**

- Police (3)

- + Bali (7)

- Java, Island Of Indonesia (3)

- Magnitude, Tremor off the island (2)

- Major Power (2)

- Aquatic (2)

- Tour, West Java (2)

- Custom (3)

- Research (2)

- Other Topics (5)

+ Database (32)

4)

ments (22)

mming language (14)

+ Java Developer (16)

+ Picture (21)

+ Processing (17)

+ Training, Courses (17)

+ Operator, Stack Overflow (10)

+ Editor (17)

+ Solutions, Software (15)

1. [Java | island, Indonesia | Britannica.com](#) [new window](#) [preview](#)

**Date:** 2019-10-05T05:04:00

**Java**, island of **Indonesia** lying southeast of Malaysia and Sumatra, south of Borneo (Kalimantan), and west of Bali. **Java** is only the fourth largest island in **Indonesia** but contains more than half of the nation's population and dominates it politically and economically. The capital of **Java** and of the <https://www.britannica.com/place/Java-island-Indonesia> - Yippy Index V

2. [Indonesia Sends Police to Restive Region After Racial Slurs Ignite Protests](#) [new window](#) [preview](#)

**Description:** Demonstrations erupted in distant **Indonesian** Papua after a video circulated showing Papuan students on **Java** being insulted as the police surrounded their dormitory. <https://www.nytimes.com/...-papua-indonesia.html?emc=rss&partner=rss> - Yippy News

3. [Indonesian security minister attacked by man with knife: police](#) [new window](#) [preview](#)  
... security minister Wiranto on Thursday, during a visit to the town of Pandeglang in Banten province on the island of **Java**, images from police and television showed.

<https://www.reuters.com/.../worldNews> (Reuters World News) - Yippy News

4. [Indonesia capital, neighboring provinces in Java hit by major power blackout](#) [new window](#) [preview](#)

<https://www.jpost.com/...-in-Java-hit-by-major-power-blackout-597586> - Yippy News Archives

5. [Tsunami warning for parts of Indonesia's Sumatra, Java after strong quake](#) [new window](#) [preview](#)

**Indonesian** authorities issued a tsunami warning in the wake of a magnitude 7 earthquake off the islands of Sumatra and **Java**. <https://www.cnb.com/...g-quake-off-indonesias-sumatra-and-java.html> - Yippy News Archives

6. [4 dead after quake hits off Indonesia's Java island](#) [new window](#) [preview](#)

Quelle:

<https://yippy.com/>

- Ad-Hoc Suche
  - Finde relevante Dokumente für eine beliebige Textanfrage.
- Filtern (Bsp.: nächste Folie)
  - Identifiziere relevante Benutzerprofile für ein neues Dokument.
- Klassifikation (Bsp.: übernächste Folie)
  - Identifiziere relevante Kennzeichen eines Dokuments.
  - Bilde Gruppen und verschaffe so einen Überblick.
- Beantworten von Fragen (Bsp.: überübernächste Folie)
  - Gib eine gezielte **Antwort zu einer Frage**.

Who is the president of germany



[Browse Examples](#) [Surprise Me](#)

Input interpretation:

Germany President

Result:

**Frank-Walter Steinmeier** (from 19. March 2017 to present)

Basic information:

official position	President
country	Germany
start date	19. March 2017 (1 year 6 months 29 days ago)

Sequence:

[More](#)

Sunday, March 19, 2017 to present	Frank-Walter Steinmeier
Sunday, March 18, 2012 to Sunday, March 19, 2017 (5 years 1 day)	Joachim Gauck
Friday, February 17, 2012 to Sunday, March 18, 2012 (1 month 1 day)	Horst Seehofer CSU (acting)
Wednesday, June 30, 2010 to Friday, February 17, 2012 (1 year 7 months 17 days)	Christian Wulff (Christian Democratic Union)
Monday, May 31, 2010 to Wednesday, June 30, 2010 (30 days)	Jens Böhrnsen (Social Democratic Party of Germany) (acting)

Personal information:

full name	Frank-Walter Steinmeier
date of birth	05. January 1956 (age: 62 years)
place of birth	Detmold, North Rhine-Westphalia, Germany

Timeline:



Image:



**DISCOVER WHAT'S POSSIBLE**  
with Wolfram|Alpha

[Take the Tour](#)

Step-by-step calculators for calculus, algebra and more...

Take the integral:  
 $\int \tan^2(x) dx$

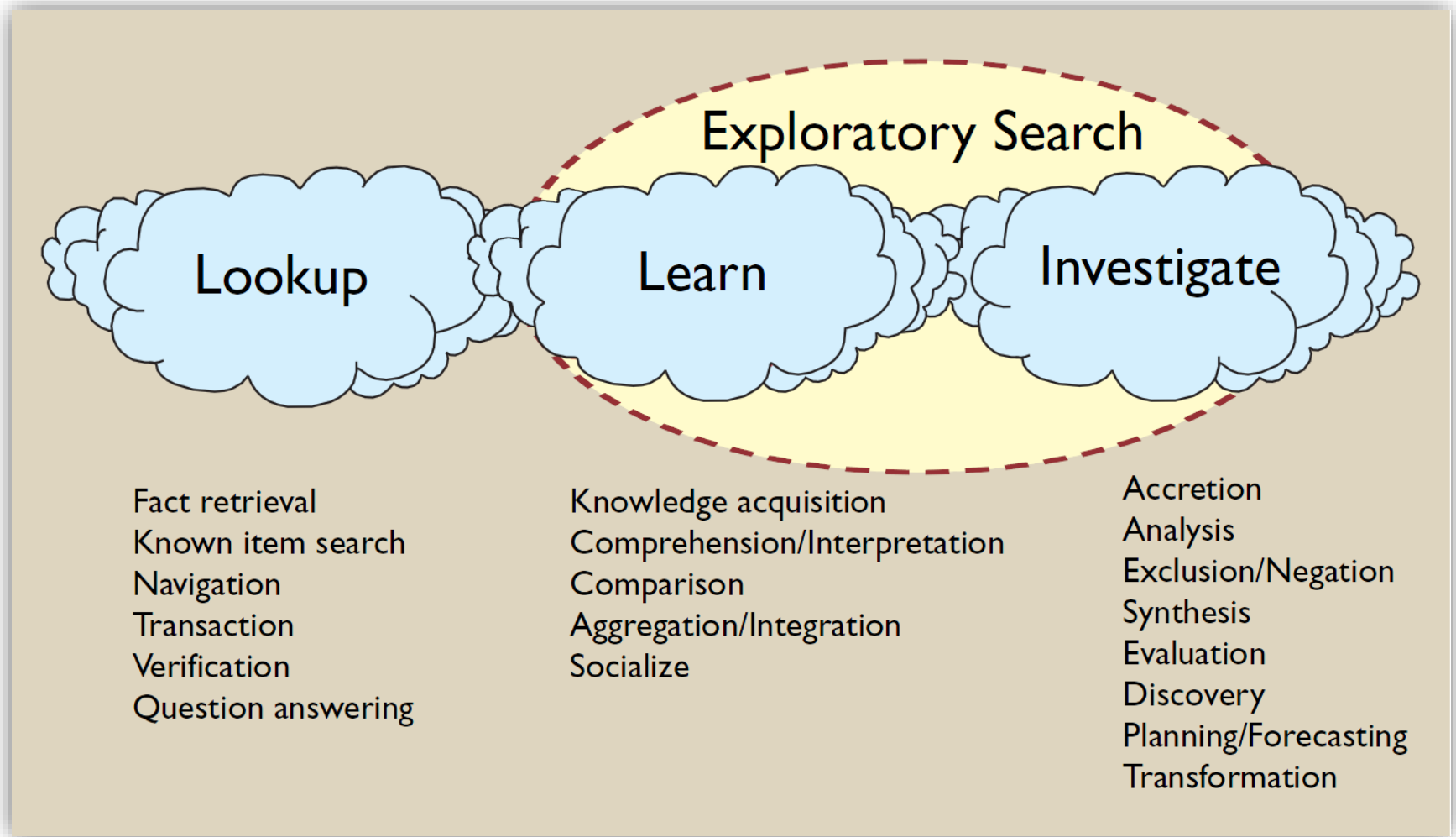
Write  $\tan^2(x)$  as  $\sec^2(x) - 1$ :  
 $= \int (\sec^2(x) - 1) dx$

Integrate the sum term by term and factor out:  
 $= \int \sec^2(x) dx - \int 1 dx$

Your homework problems solved!

[Go Pro Now](#)

Quelle: <https://www.wolframalpha.com/input/?i=Who+is+the+president+of+germany>



## ■ Relevanz

- Was ist das?
- Einfache (und vereinfachende) Definition:  
„Ein relevantes Objekt enthält die Information, nach der eine Person gesucht hat, als sie eine Anfrage an eine Suchmaschine gestellt hat.“
- Viele Faktoren beeinflussen die Entscheidung einer Person bezüglich der Relevanz:
  - z. B. Aufgabe, Kontext, Aktualität, Stil, Vorwissen, Umfang
- Topical relevance (gleiches Thema [näherungsweise objektiv entscheidbar])  
vs. user relevance (alles andere [nur subjektiv entscheidbar])

- Retrievalmodelle definieren eine Sicht auf Relevanz
- Rankingalgorithmen, die in Suchmaschinen eingesetzt werden, basieren auf Retrievalmodellen
- Viele Modelle beschreiben eher statistische Eigenschaften von Texten als linguistische oder semantische
  - D. h. es werden einfache Merkmale des Texts gezählt (wie Wörter) anstatt Sätze zu parsen und zu analysieren
  - Statistische Ansätze zur Textverarbeitung wurden erstmals von Luhn in den 1950er Jahren vorgestellt
  - Linguistische Merkmale können Teil statistischer Modelle sein



# Linguistische Ansätze: Grammatik Browser



- CC Coordinating conjunction
- CD Cardinal number
- DT Determiner
- EX Existential there
- FW Foreign word
- IN Preposition or subordinating conjunction
- JJ Adjective
- JJR Adjective, comparative
- JJS Adjective, superlative
- LS List item marker
- MD Modal
- NN Noun, singular or mass
- NNS Noun, plural
- NNP Proper noun, singular
- NNPS Proper noun, plural
- PDT Predeterminer
- POS Possessive ending
- PRP Personal pronoun
- PRP\$ Possessive pronoun
- RB Adverb
- RBR Adverb, comparative
- RBS Adverb, superlative
- RP Particle
- SYM Symbol
- TO to
- UH Interjection
- VB Verb, base form
- VBD Verb, past tense
- VBG Verb, gerund or present participle
- VBN Verb, past participle
- VBP Verb, non3rd person singular present
- VBZ Verb, 3rd person singular present
- WDT Whdeterminer
- WP Whpronoun
- WP\$ Possessive whpronoun
- WRB Whadverb

## Part-of-Speech:


1 During the Second World War, Turing worked for the Government Code and Cypher School (GC&CS) at Bletchley Park, Britain's codebreaking centre that produced Ultra intelligence.

2 For a time he led Hut 8, the section which was responsible for German naval cryptanalysis.

3 Here he devised a number of techniques for speeding the breaking of German ciphers, including improvements to the pre-war Polish bombe method, an electromechanical machine that could find settings for the Enigma machine.

Quelle: Stanford NLP <http://nlp.stanford.edu:8080/corenlp/process>

## ■ Evaluation

- Experimentelle **Verfahren** und **Maße** um die **Ergebnisse** eines Systems mit den **Erwartungen des Nutzers** zu vergleichen.
  - Auf die Cranfield Experimente in den 1960er Jahren zurückzuführen
- IR-Evaluationsmethoden werden mittlerweile in vielen Bereichen eingesetzt.
- Meist werden **Testkollektionen** von
  - Dokumenten und
  - Anfragen mit
  - bekannter Relevanzentscheidung verwendet.
- Am weitesten verbreitet sind die TREC-Kollektionen (Text REtrieval Conference) 
- **Recall** und **Precision** sind zwei Beispiele für Maße zur Bestimmung der Effektivität

# The Text REtrieval Conference (TREC)



<https://trec.nist.gov/>



Aufgabenstellungen,  
Suchszenarien

Kollektionen mit Dokumenten,  
Anfragen und Relevanzurteilen

- Wie kann man die **Qualität des Ergebnisses** messen?
- Annahme: Unterstellt wird eigentlich eine Ergebnismenge
  - Für eine gegebene Anfrage kann für jedes Objekt **eindeutig** entschieden werden, ob es **relevant** oder **nicht relevant** ist.
- **Qualität des Ergebnisses**
  - Wie **vollständig** ist das Ergebnis?
    - Anteil der **gefundenen relevanten Dokumente** an der **Gesamtzahl** der relevanten Dokumente
    - 50 relevante gefunden; 100 relevante insgesamt  $\Rightarrow$  **Recall** = 50 %
  - Wie **genau** ist das Ergebnis? Enthält es **nur relevante** Einträge?
    - Anteil der **relevanten Dokumente** im Ergebnis
    - 50 relevante gefunden; 75 insgesamt im Ergebnis  $\Rightarrow$  **Precision** = 67 %

Frage:

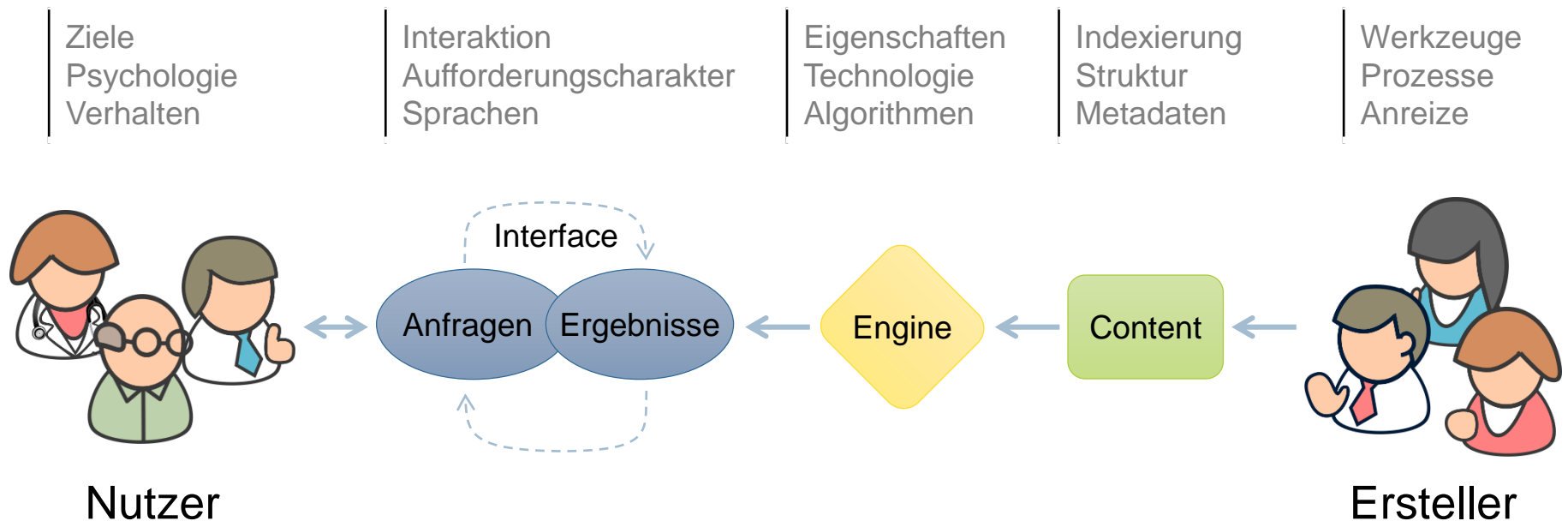
Braucht man *ein* relevantes Dokument (Bsp. QA) oder *alle* relevanten Dokumente (Bsp. Patentrecherche)?

- Nutzer und Informationsbedürfnisse
    - Evaluation von Suchalgorithmen sollte **nutzerorientiert** sein
    - **Anfragen** mit Schlüsselwörtern bilden den **eigentlichen Informationsbedarf** oft **schlecht ab** (nächste Folie)
    - **Interaktion** und **Kontext** sind wichtig, um die **Intention des Nutzers** zu verstehen
    - Technologien zur Anfrageverfeinerung (**query refinement**), wie
      - **query expansion**,
      - Such[term]vorschläge (**query [term] suggestion**) oder
      - **Relevanz Feedback**
- ⇒ verbessern das Ranking

# Informationsbedürfnis vs. Anfrage



- **Information Retrieval (IR)** betrachtet Informationssysteme in Bezug auf ihre Rolle im Prozess des **Wissenstransfers vom menschlichen Wissensproduzenten zum Informations-Nachfragenden**.



In Anlehnung an: Morville, Peter; Callender, Jeffery (2010): Search patterns. Sebastopol, CA: O'Reilly.

- Eine Suchmaschine ist die praktische Anwendung verschiedener IR-Techniken auf große Textkollektionen
- Internetsuchmaschinen sind die bekanntesten Beispiele, es gibt jedoch noch viele andere
  - Open Source Suchmaschinen sind wichtig für Forschung und Entwicklung; z.B.
    - Lucene (<http://lucene.apache.org/>),
    - Solr (<https://lucene.apache.org/solr/>)
    - ElasticSearch (<https://github.com/elastic/elasticsearch>)
    - Lemur/Indri (<http://www.lemurproject.org/>),
    - Galago (<http://www.galagosearch.org/>)
- Wichtige Aspekte von Suchmaschinen schließen wichtige Aspekte des IR ein, es existieren aber weitere Gesichtspunkte



## Information Retrieval

### Relevanz

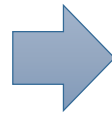
- *Effektives Ranking*

### Evaluierung

- *Testen und Messen*

### Informationsbedürfnisse

- *Benutzerinteraktion*



## Suchmaschinen

### Performanz

- *Effiziente Suche und Indexierung*

### Einpflegen neuer Daten

- *Umfang (coverage) und Aktualität (freshness)*

### Skalierbarkeit

- *wächst mit Daten und Nutzern*

### Anpassbarkeit

- *Tuning für Applikationen*

### Spezielle Probleme

- *z.B. Spam*





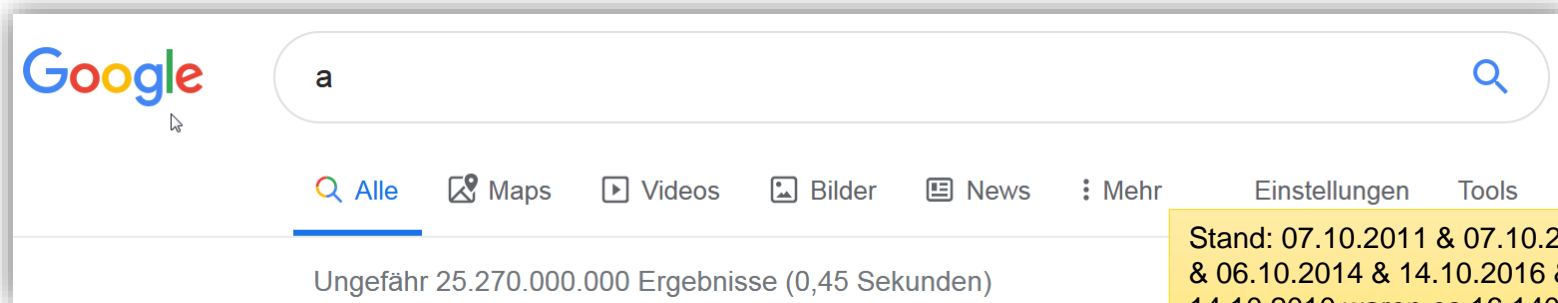
## ■ Performanz

- Das Messen und Verbessern der **Effizienz der Suche**; z.B.
  - Verringern der **Antwortzeiten**,
  - Erhöhen des **Query-Throughput**,
  - Erhöhung der **Indexierungsgeschwindigkeit**
- **Indizes als Datenstrukturen**, die entwickelt wurden, um die Effizienz der Suche zu erhöhen
  - Entwurf und Implementierung von Indexstrukturen sind wichtige Aspekte im Forschungsgebiet der Suchmaschinen
  - Invertierte Listen & Co

- Dynamische Daten
  - Die „Kollektion“ für die meisten realen Anwendungen ist **ständigen Änderungen**, in Form von Updates sowie dem Hinzufügen und Löschen von Dokumenten, **unterworfen**.
    - Z. B. Webseiten
  - Das „Beschaffen“ oder „**Crawlen**“ der Dokumente ist eine Hauptaufgabe
    - Typische **Maße** sind
      - **coverage** – Wie viel wurde bereits indexiert?
      - **freshness** – Wie viel Zeit ist seit der (letzten) Indexierung verstrichen?

## ■ Skalierbarkeit

- Es muss sichergestellt werden, dass die Suchmaschine auch mit Millionen von Nutzern jeden Tag und vielen Terabytes an Dokumenten funktioniert



- Verteilte Verarbeitung ist essentiell (Cloud Computing)

## ■ Anpassbarkeit

- Verändern und Tunen von Suchmaschinenkomponenten, wie des Rankingalgorithmus, der Indexierungsstrategie und des Interfaces für verschiedenen Applikationen

## ■ Spam

- Für Suchmaschinen ist Spam einer der zentralen Aspekte
- Spam **beeinträchtigt die Effizienz** der Suchmaschinen **und**, was noch wichtiger ist, die **Effektivität** der Ergebnisse.
- Viele Arten von Spam (Index-Spamming / spamdexing):
  - Term-Spamming,
  - Link-Spamming,
  - „Optimierung“



# Was will dieser Kurs erreichen?



- Verstehen von Suchmaschinen
- Evaluieren und Vergleichen von Suchmaschinen
- Anpassung von Suchmaschinen für spezielle Anwendungen
  
- Einen breit gefächerten Überblick über wichtige Aspekte des Information Retrieval und Suchmaschinen vermitteln
  - Das schließt die zugrundeliegenden Modelle sowie aktuelle Forschungsrichtungen ein

