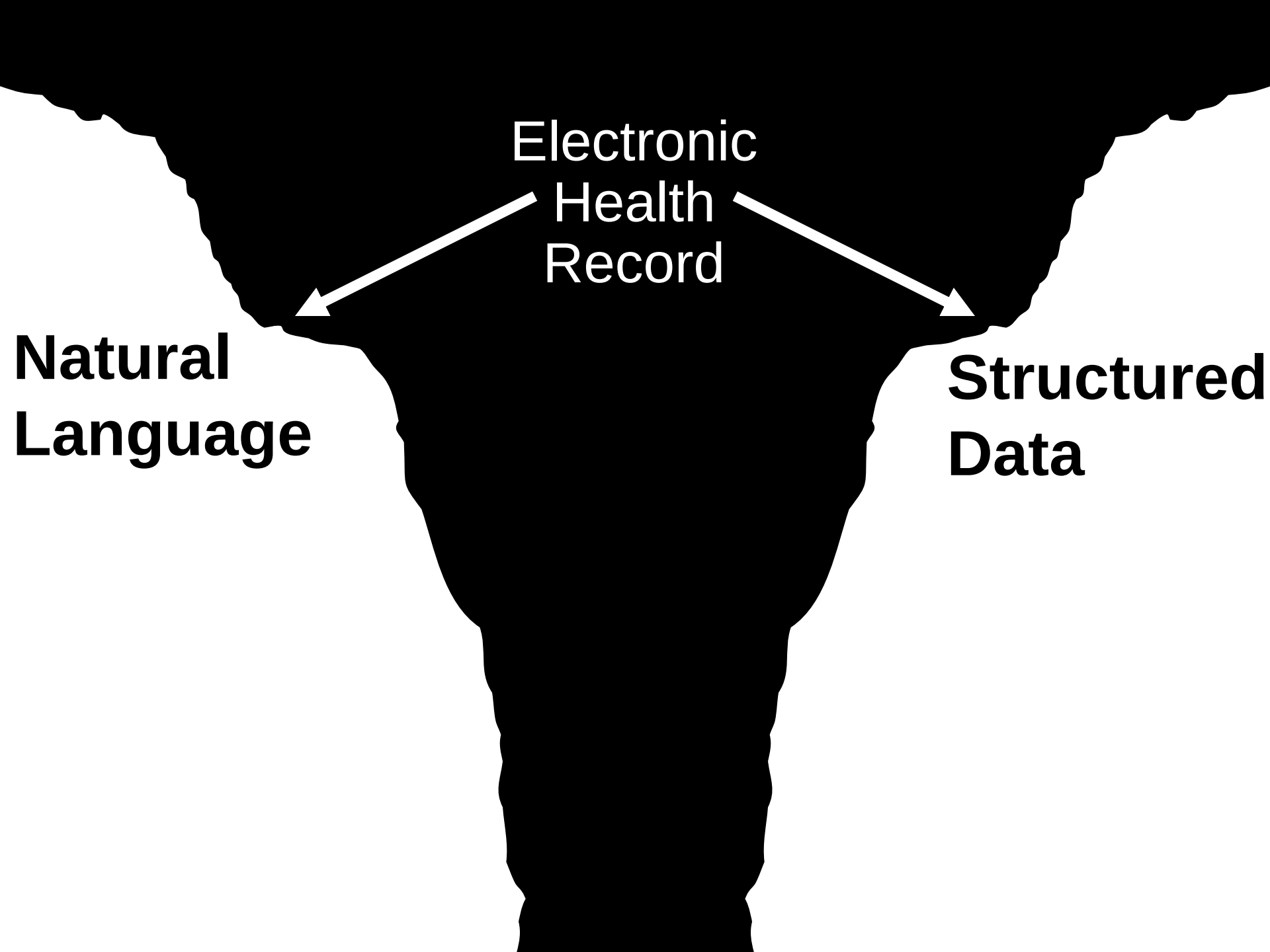
Natural Language Processing, Linguistics and Terminology

… content technologies at the point of care

Stefan Schulz
Department of Med. Informatics
Freiburg University Hospital
Germany

Udo Hahn
Text Knowledge Engineering Lab
Freiburg University
Germany


Electronic
Health
Record
Natural
Language
Structured
Data

Natural Language

Datum	Krankheits-Geschichte	Datum	Krankheits-Geschichte

Datum	Uhrz.	Pflegebericht – Verlaufsbeschreibung Krankenbeobachtung	Hz.
21.4.	19^{30}	Pat. kam mit einer [illegible], hat keine Bettruhe, war schon zur Toilette und hat sich selbstständig gewaschen	
22.4	5^{30}	Pat. hat in d. Nacht sehr wenig geschlafen, sie gab aber ∅ Beschwerden an	Sh
22.4	14^{00}	Pat. lt. Pflegeplan versorgt	
	21^{00}	Pat. lt. Plan versorgt; ist nach Belastung sehr geschafft	
23.4.	5^{00}	Pat saß fast die ganze Nacht im Bett und konnte nicht im Liegen schlafen	Sh
	11^{00}	Pat hatte 3x breiig bis dünnflüssigen Stuhlgang (braun)	
	14^{00}	Versorgg. lt. Plan	
	21^{00}	Pat. lt. Plan versorgt, hält sich nicht an ihre Bettruhe	
24.4.	2^{30}	Pat beim Toilettengang erwischt, hält sich nicht an Bettruhe (sieht keinen Sinn darin)	
	3^{30}	Pat wollte aus dem Zimmer geführt	

Familienanamnese:
Vater verstorben an Bronchial-Karzinom, Mutter verstorben an den Folgen einer Pneumonie. Mutter Diabetes mellitus. 5 gesunde Kinder.

Systemanamnese:
Derzeit Appetitlosigkeit, Trockengewicht um 75 Kg, derzeit 80 Kg. Miktio: gelegentlich Harn-verhalt, gehäuft Harnwegsinfekte, derzeit keine Algurie. Vor Dialyse keine Rest-Diurese. Stuhlgang obstipiert, benutzt regelmäßig Abführmittel. Vor NTX starker Juckreiz, Seit NTX deutlich rückläufig. Kein Husten/Auswurf. Noxen: Nichtraucherin, kein Alkohol.

Soziale Anamnese:
Früher Arbeiterin in der Elektronikbrache, dann Hausfrau, verheiratet, lebt mit dem Ehemann zusammen.
Allergien. Keine bekannt.

Medikation bei Aufnahme:
Ulcogant 1-1-1, Pepdul mit 0-0-0-1, Cellcept 2x1 g, Bayotensin 3 x 1, Cynt 0,2 1x1, Ludiomil 50 mg 1 x 1, Sandimmun 2 x 150 mg, Clexane 0,4 ml 1 x täglich s.c.

Status bei Übenahme:
58-jährige Patientin in vorgealtertem, reduziertem Allgemein- und adipösem Ernährungszu-stand (80 Kg Gewicht bei 160 cm Körpergröße). RR 170/80 mm Hg, Puls 66/Minute, regelmäßig. Punktförmige Depigmentierungen an beiden Unterarmen bei Zustand nach heftigem Kratzen wegen Juckreiz. Keine zervikalen Lymphome. Mundschleimhaut trocken, Zunge weißlich belegt. Rachenschleimhaut reizlos, Tonsillen schlecht einsehbar. Schilddrüse nicht vergrößert. Pulmo: Sonorer Klopfschall und vesikuläres Atemgeräusch. Cor: Spitzenstoß nicht tastbar, leise, reine Herztöne. 3/6. spindelförmiges Systolikum und 1-2/6. Decrescendo-Sofort-dialstolikum über der Aorta mit Fortleitung in die Karotis. Kein abdominales und inguinales Strömungsgeräusch. Abdomen: Bei Adipositas Organgrenzen schlecht beurteilbar, Leber/Milz nicht vergrößert. Reizlose Narbe im Bereich des rechten Unterbauches bei Zustand nach NTX. Dort leichte Druckdolenz. Wirbelsäule nicht klopfschmerzhaft. Bds. Unterschen-kelödeme. Feinschlägiger Tremor beim Arm-Vorhalte-Versuch. Pupillen isokor, Lichtreaktion prompt. Finger-Nase-Versuch bds. unsicher, ataktisch. Reflexe seitengleich.

Burden of infectious diseases in South Asia

Anita K M Zaidi, Shally Awasthi, H Janaka deSilva

Infectious diseases are a major cause of death in South Asia, with children incurring a disproportionate share of the burden. This review discusses the underlying causes of some of the more common diseases and strategies to improve their detection and control

Preventable infections are a major cause of deaths and disabilities in South Asia. Over two thirds of the estimated 3.7 million deaths in children in South Asia in the year 2000 were attributable to infections such as pneumonia, diarrhoea, and measles.[1 2] India now has the second largest population with AIDS and HIV infection in the world, and tuberculosis and chronic hepatitis continue to threaten the lives of millions. Of the overall burden of deaths related to infectious disease in the region, around 63% are in children aged under 5 years.[3] Serious effort should be devoted to the control of infectious disease if South Asian countries are to meet their millennium development goal of two thirds reduction in child mortality by 2015.

Sri Lanka alone among South Asian countries has made remarkable progress in reducing the burden of infectious disease, despite civil war and meagre resources.

This review describes the burden of infectious

Summary points

- Acute respiratory infections, diarrhoea, and neonatal infections remain major child killers
- India has the second highest burden of HIV and AIDS in the world, with 4.58 million people infected with HIV
- Antibiotic misuse has resulted in high rates of antimicrobial resistance
- Only half of all South Asian children receive routine immunisations, and many new vaccines have not been introduced in mass immunisation programmes
- Lack of surveillance systems and poorly

Vyhledávání

článek Hledej

Nemoci podle orgánů

- Břicho
- Hlava
- Hrudník
- Infekční a chronické nemoci
- Končetiny
- Kůže
- Pohlavní orgány a nemoci
- Psychika

Databáze

- Nemoci
- Léky
- Lékové interakce
- Rostliny léčí
- Minerály
- Vitamíny
- Doplatky na léky
- Lékárny
- Lékaři
- Dobré adresy
- Když se řekne...

Kontakt

- Napište nám

Ischemická choroba srdeční

Co je to „ischemie“?

Navzdory poměrně komplikovanému názvu je tato choroba asi většině z Vás dobře známa. Její důsledky pro naše zdraví jsou příliš vážné a výskyt v našem nejbližším okolí příliš častý, než aby ji bylo možné jen tak přehlédnout. Ischemická choroba srdeční je v České republice již několik let prvořadým zdravotním problémem. Je příčinou takřka třetiny úmrtí v našem státě a u stejně početné skupiny kvalitu života významně zhoršuje. O co se tedy jedná?

„Ischemie" znamená nedostatečné prokrvení orgánu. Dostatečné prokrvení, tedy odpovídající příjem živin a kyslíku, je nezbytné pro správné fungování každého orgánu v našem těle. Pokud se tato porucha dotkne natolik důležitého orgánu jakým je srdce, dojde k přímému ohrožení našeho života.

Čím je choroba způsobena

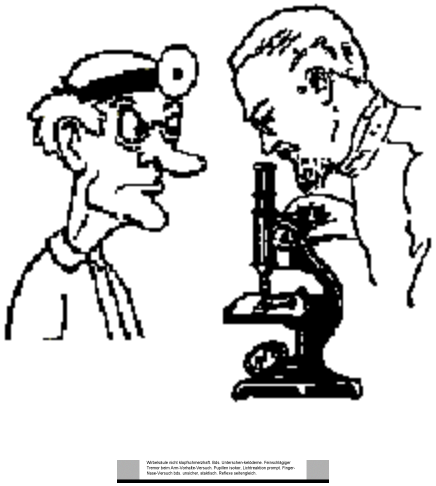
Ischemická choroba srdeční představuje nedostatečné prokrvení části srdečního svalu (ischemii myokardu) v důsledku poruchy věnčitých tepen, které srdeční sval normálně vyživují. Nejčastější příčinou poškození věnčitých tepen je **ateroskleróza**.

Ateroskleróza je onemocnění poškozující cévy ukládáním tuků do jejich stěny. Nejčastější příčinou těchto změn je **zvýšená hladina cholesterolu** v důsledku jeho nadbytečné konzumace v potravě. Ve věnčitých tepnách se vytvářejí tzv. **aterosklerotické pláty** připomínající nánosy bahna a rzi ve starých vodovodních trubkách. Důsledkem tohoto zúžení je pak nedostatečné prokrvení srdečního svalu – ischemie myokardu.

Ke zužování věnčité tepny může docházet postupně a dlouhodobě. Takové zúžení se projeví poprvé až při zvýšené námaze – jedná se o tzv. stabilní **anginu pectoris**. Pokud je však plát měkký a křehký, tzn. nestabilní, může prasknout a vyvolat tak tvorbu krevní sraženiny, která cévu náhle zúží nebo úplně uzavře. Důsledkem je pak náhlý vznik obtíží – **srdeční infarkt**.

Neprokrvená část srdečního svalu odumírá a je postupně nahrazena méněcennou vazivovou tkání - jizvou. Srdce ztrácí část své svalové tkáně a tím i síly k další práci. Rozsah následků závisí na druhu postižené srdeční tepny a na jejím průměru. Čím větší tepna, tím je zasažen větší okrsek svalstva srdce a o to pak i závažnější důsledky pro pacienta. Pokud postihne infarkt větší část srdce – udává se rozsah okolo 40 % - srdce není schopno dále pracovat.

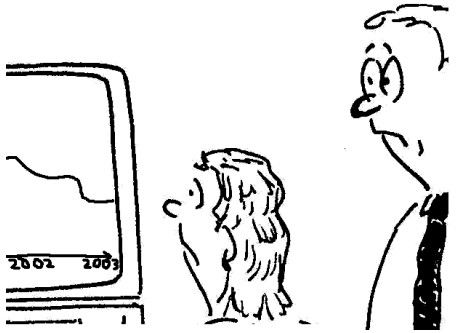
Natural Language



… is the main carrier of

- Communication between health professionals
- Communication between researchers
- Clinical documentation
- Scientific publication
- Dissemination of authorititative medical knowledge for professionals and consumers

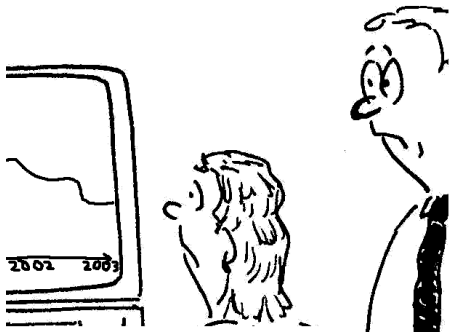
Structured Data



…required for

- Disease reporting
- Patient grouping for Billing, Controlling
- Clinical Trials
- Health Statistics
- Registries, Databases
- Document Indexing and Retrieval

Structured Data



Epidemiology (Morbidity, Mortality)

Causes of Death			Number of Deaths	Number of Deaths at ages (in days)			
ICD 10 Codes	Cause Groupings	Sex	< 1 year	< 1	1-6	7-27	28-364
-	All causes	M	2485	168	703	416	1198
		F	1885	138	448	279	1020
		U	0	0	0	0	0
A00-B99	Infectious and parasitic diseases	M	95	0	1	9	85
		F	93	0	2	9	82
A00-A09	Intestinal infectious diseases	M	65	0	0	6	59
		F	63	0	0	6	57
A37	Whooping cough	M	1	0	0	0	1
		F	0	0	0	0	0
E40-E64	Nutritional deficiencies	M	1	0	0	0	1
		F	0	0	0	0	0
G00-G98	Diseases of the nervous system	M	92	0	2	26	64
		F	61	0	4	16	41
G00,G03	Meningitis	M	37	0	2	18	17
		F	30	0	4	12	14
J00-J98	Diseases of the respiratory system	M	734	0	3	66	665
		F	609	1	3	51	554
J12-J18	Pneumonia	M	705	0	3	65	637
		F	587	1	3	50	533
J10,J11	Influenza	M	0	0	0	0	0
		F	0	0	0	0	0

Billing, Controlling

L XIII. Herzchirurgie

3050

Operative Maßnahmen in Verbindung mit der Herz-Lungen-Maschine zur Herstellung einer extrakorporalen Zirkulation

1850	107,83€	248,01€

3051

Perfusion der Hirnarterien, zusätzlich zur Leistung nach Nummer 3050

1290	75,19€	172,94€

3052

Perfusion der Koronararterien, zusätzlich zur Leistung nach Nummer 3050

1110	64,70€	148,81€

3053

Perfusion von Arterien eines anderen Organs, zusätzlich zur Leistung nach Nummer 3050

1110	64,70€	148,81€

3054

Operative extrathorakale Anlage einer assistierenden Zirkulation

1850	107,83€	248,01€

3055

Überwachung einer assistierenden Zirkulation, je angefangene Stunde

554	32,29€	74,27€

Die Leistung nach Nummer 3055 ist nur während einer Operation berechnungsfähig.

3060

Intraoperative Funktionsmessungen am und/oder im Herzen

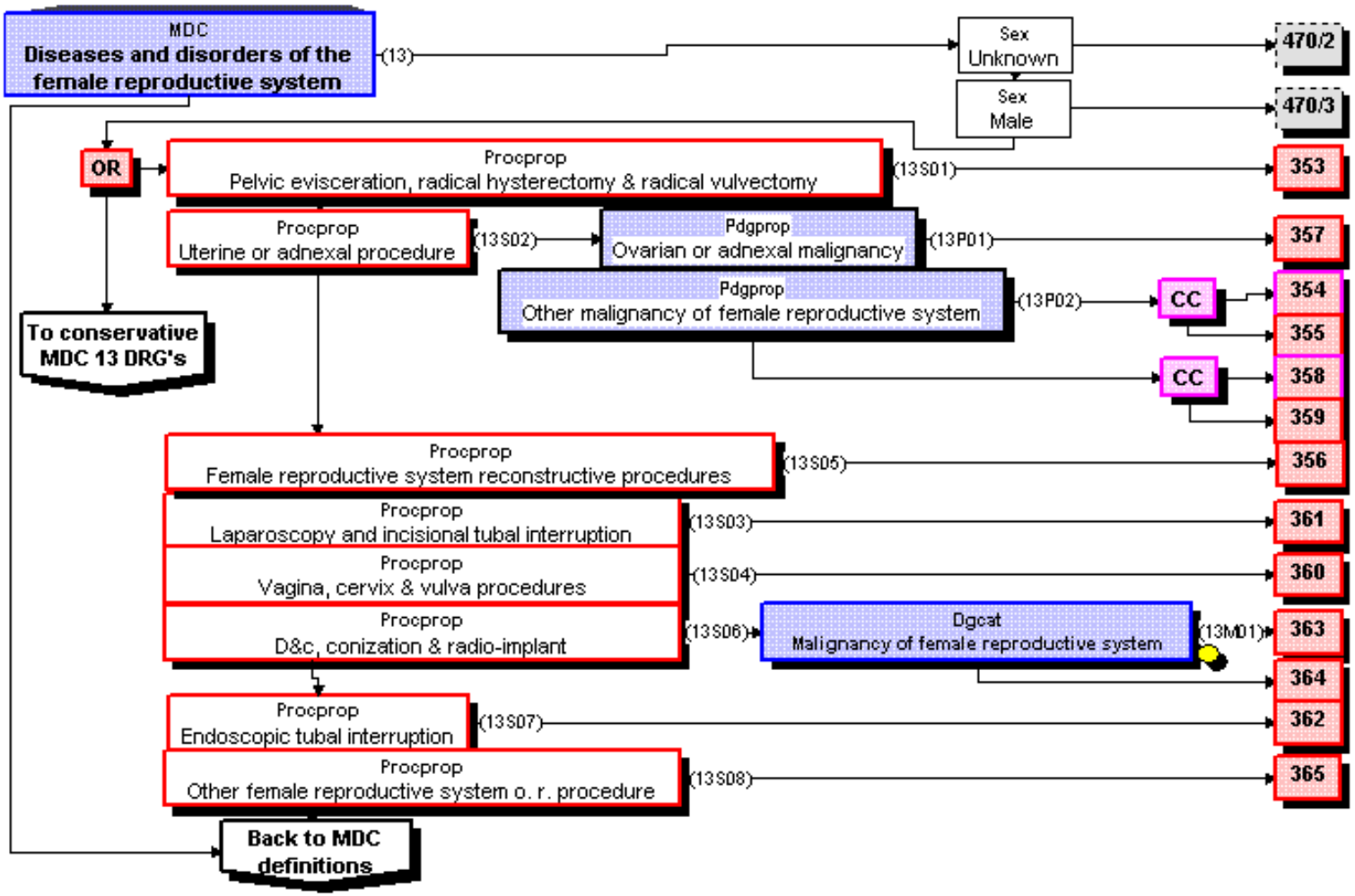
554	32,29€	74,27€

3065

Operation am Perikard, als selbständige Leistung

2000	116,57€	268,13€

DRGs



Cancer Registries

Tabelle 7: **3-, 5-, 10- und 15-Jahres-Überlebenswahrscheinlichkeit und Wahrscheinlichkeit rezidivfreien Überlebens (Sterbetafelverfahren ergänzt nach (9)) für die häufigsten Diagnosen (1981-2001)**
3-, 5-, 10-, and 15-year survival probabilities and event-free survival probabilities (life table method extended according to (9)) for the most common diagnoses (1981-2001)

Diagnoses	Number of cases *	Probabilities							
		Event-free survival				Survival			
		3-	5-	10-	15-year	3-	5-	10-	15-year
Retinoblastoma	461	_	_	_	_	97	97	95	95
Hodgkin's disease	1543	88	86	84	83	96	95	94	93
Nephroblastoma	1830	81	80	80	79	87	86	86	85
Germ cell tumours	1003	82	79	77	75	89	88	85	84
Non-Hodgkin lymphoma	1787	80	79	77	76	84	83	81	80
Lymphoid leukaemia	8136	77	71	68	67	85	81	76	74
Astrocytoma	1987	72	68	63	58	78	76	74	68
Neuroblastoma	2474	61	58	57	56	70	65	62	62
Osteosarcoma	770	63	58	55	54	75	66	61	60
Ewing's sarcoma	584	62	56	54	52	71	64	60	59
Rhabdomyosarcoma	1022	58	55	53	51	69	64	60	58
Acute non-lymphocytic leukaemia	1543	42	40	39	38	49	46	44	42
Primitive neuroectodermal tumours	1254	52	47	41	38	59	53	45	41
All malignancies	27799	70	66	63	62	78	74	71	69

Literature Indexing

TI - CT appearance of primary CNS lymphoma in patients with acquired immunodeficiency syndrome.

PG - 39-44

AB - Cranial CT studies of 32 patients with biopsy-proven AIDS-related primary CNS lymphoma were reviewed retrospectively. A wide variety of different CT appearances were identified. Mass lesions varied in location, size, and number. Most lesions were either iso- or hyperdense and all enhanced with contrast medium. Several different patterns of enhancement were observed. Mass effect and edema were seen in almost all patients. After radiotherapy, most tumors decreased in diameter, became hypodense, and no longer enhanced with contrast medium. Edema and mass effect decreased or resolved in all but one patient. Postradiotherapy CT scans also revealed interval enlargement of the ventricles and cortical sulci. This study demonstrates the wide diversity of CT appearances of AIDS-related primary CNS lymphoma. The CT findings cannot be used in lieu of biopsy for diagnosis of this disorder. The appearance of postradiotherapy CT scans was consistent with regressing lymphoma.

TA - J Comput Assist Tomogr

MH - Acquired Immunodeficiency Syndrome/*complications

MH - Adolescent

MH - Adult

MH - Aged

MH - Brain Neoplasms/etiology/*radiography/radiotherapy

MH - Child

MH - Child, Preschool

MH - Female

MH - Human

MH - Lymphoma/etiology/*radiography/radiotherapy

MH - Male

MH - Middle Aged

MH - Retrospective Studies

MH - *Tomography, X-Ray Computed

EDAT- 1991/01/01

MHDA- 1991/01/01 00:01

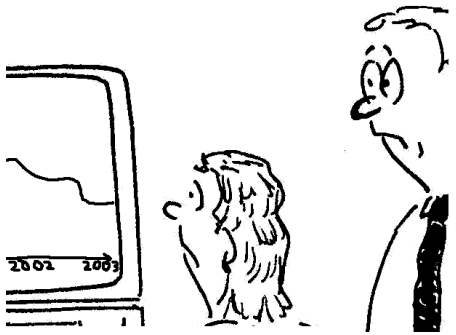
PST - ppublish

SO - J Comput Assist Tomogr 1991 Jan-Feb;15(1):39-44.

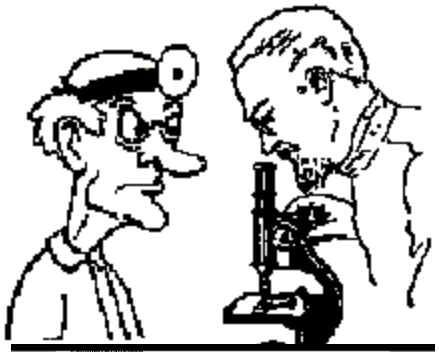
…based on Clinical Terminologies, Classifictions, e.g.

- ICD
- Procedure Codes
- SNOMED
- MeSH
- etc., etc.

Structured Data



Natural Language



Familienanamnese:
Vater verstorben an Bronchial-Karzinom, Mutter verstorben an den Folgen einer Pneumonie. Mutter Diabetes mellitus. 5 gesunde Kinder.

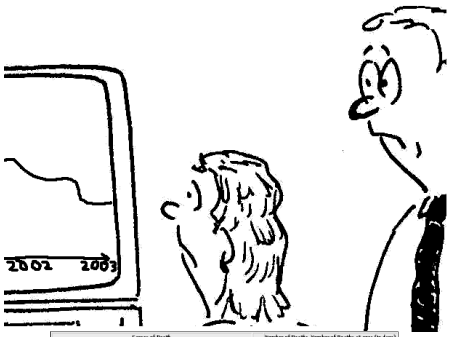
Systemanamnese:
Derzeit Appetitlosigkeit, Trockengewicht um 75 Kg, derzeit 80 Kg. Miktio: gelegentlich Harn-verhalt, gehäuft Harnwegsinfekte, derzeit keine Alguria. Vor Dialyse keine Rest-Diurese. Stuhlgang obstipiert, benutzt regelmäßig Abführmittel. Vor NTX starker Juckreiz, Seit NTX deutlich rückläufig. Kein Husten/Auswurf. Noxen: Nichtraucherin, kein Alkohol.

Soziale Anamnese:
Früher Arbeiterin in der Elektronikbrache, dann Hausfrau, verheiratet, lebt mit dem Ehemann zusammen.
Allergien. Keine bekannt.

Medikation bei Aufnahme:
Ulcogant 1-1-1, Pepdul mit 0-0-0-1, Cellcept 2x1 g, Bayotensin 3 x 1, Cynt 0,2 1x1, Ludiomil 50 mg 1 x 1, Sandimmun 2 x 150 mg, Clexane 0,4 ml 1 x täglich s.c.

Status bei Übenahme:
58-jährige Patientin in vorgealtertem, reduziertem Allgemein- und adipösem Ernährungszu-stand (80 Kg Gewicht bei 160 cm Körpergröße). RR 170/80 mm Hg, Puls 66/Minute, regelmäßig. Punktförmige Depigmentierungen an beiden Unterarmen bei Zustand nach heftigem Kratzen wegen Juckreiz. Keine zervikalen Lymphome. Mundschleimhaut trocken, Zunge weißlich belegt. Rachenschleimhaut reizlos, Tonsillen schlecht einsehbar. Schilddrüse nicht vergrößert. Pulmo: Sonorer Klopfschall und vesikuläres Atemgeräusch. Cor: Spitzenstoß nicht tastbar, leise, reine Herztöne. 3/6. spindelförmiges Systolikum und 1-2/6. Decrescendo-Sofort-diastolikum über der Aorta mit Fortleitung in die Karotis. Kein abdominales und inguinales Strömungsgeräusch. Abdomen: Bei Adipositas Organgrenzen schlecht beurteilbar, Leber/Milz nicht vergrößert. Reizlose Narbe im Bereich des rechten Unterbauches bei Zustand nach NTX. Dort leichte Druckdolenz. Wirbelsäule nicht klopfschmerzhaft. Bds. Unterschen-keldeme. Feinschlägiger Tremor beim Arm-Vorhalte-Versuch. Pupillen isokor, Lichtreaktion prompt. Finger-Nase-Versuch bds. unsicher, ataktisch. Reflexe seitengleich.

Structured Data



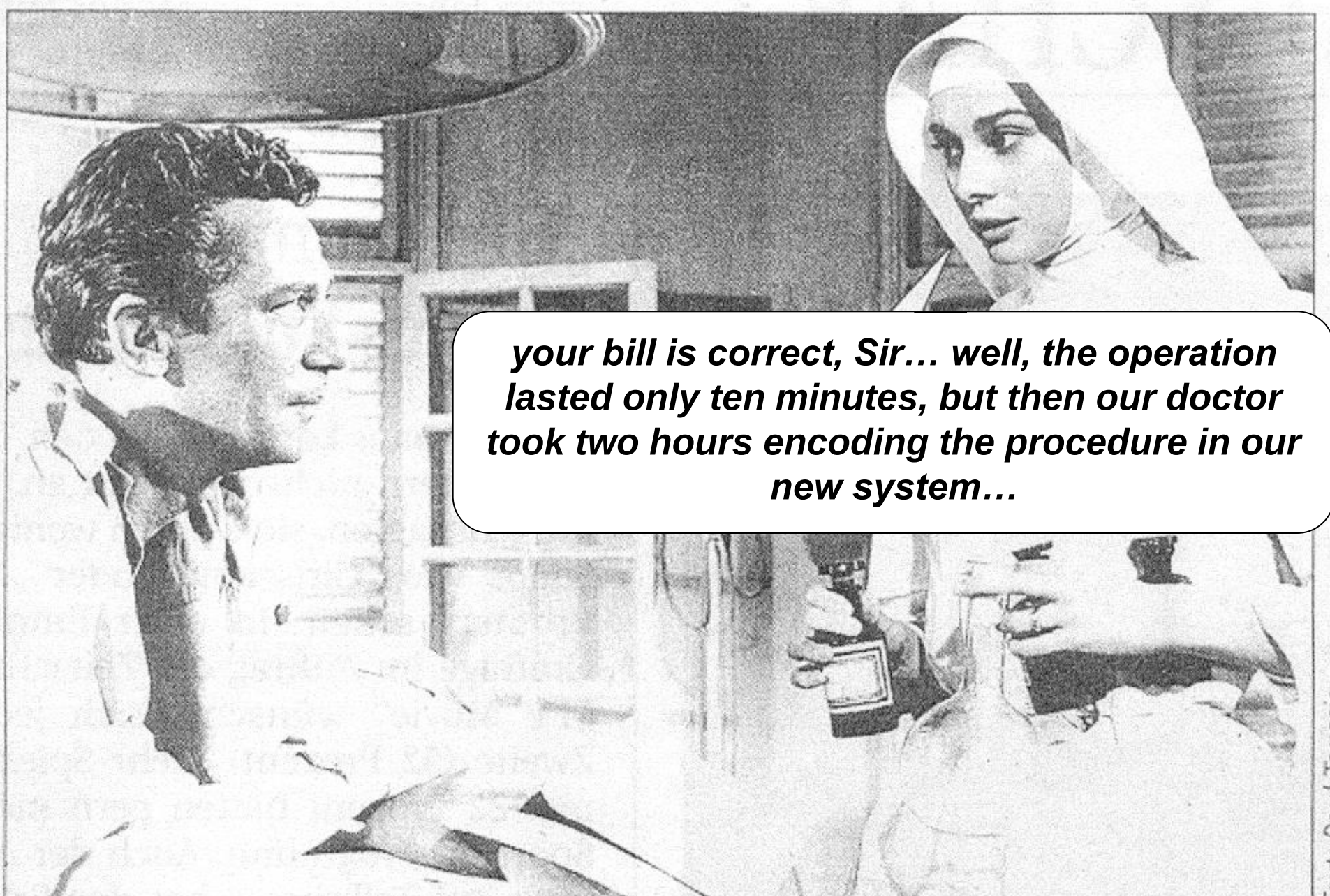
Causes of Death		Number of Deaths	Number of Deaths at ages (in days)			
ICD-10 Codes	Cause Groupings	Sex	< 1 year	1-4	7-27	28-364
	All causes	M	2405	166	703	416	1196
		F	1685	138	448	279	1020
		U	0	0	0	0	0
A00-B99	Infectious and parasitic diseases	M	95	0	1	9	85
		F	93	0	2	9	82
A00-A09	Intestinal infectious diseases	M	65	0	0	6	59
		F	63	0	0	6	57
A37	Whooping cough	M	1	0	0	0	1
		F	0	0	0	0	0
E40-E64	Nutritional deficiencies	M	1	0	0	0	1
		F	0	0	0	0	0
G00-G99	Diseases of the nervous system	M	92	0	2	26	64
		F	61	0	4	16	41
G00,G03	Meningitis	M	17	0	2	18	17
		F	30	0	4	12	14
J00-J99	Diseases of the respiratory system	M	734	0	3	46	685
		F	629	1	3	51	574
J12-J18	Pneumonia	M	705	0	3	45	637
		F	587	1	3	50	533
J10,J11	Influenza	M	0	0	0	0	0
		F	0	0	0	0	0
Q00-Q99	Congenital anomalies	M	494	36	139	160	163
		F	375	35	94	61	185
Q03,Q05	Spina bifida and hydrocephalus	M	71	5	16	5	45
		F	74	2	8	17	47
Q20-Q28	Congenital anomalies of heart and circulatory system	M	222	6	67	97	90
		F	149	9	42	23	75
P00-P96	Certain conditions originating in the perinatal period	M	879	127	55y	187	6
		F	571	97	340	131	3
P10-P15	Birth trauma	M	56	11	35	10	0
		F	34	4	25	5	0

Doctors‘ attitude towards producing...

Natural Language

Structured Data

+	quality	-
-	cost	+


your bill is correct, Sir… well, the operation lasted only ten minutes, but then our doctor took two hours encoding the procedure in our new system…

Attitude of Health Administrators towards analyzing...

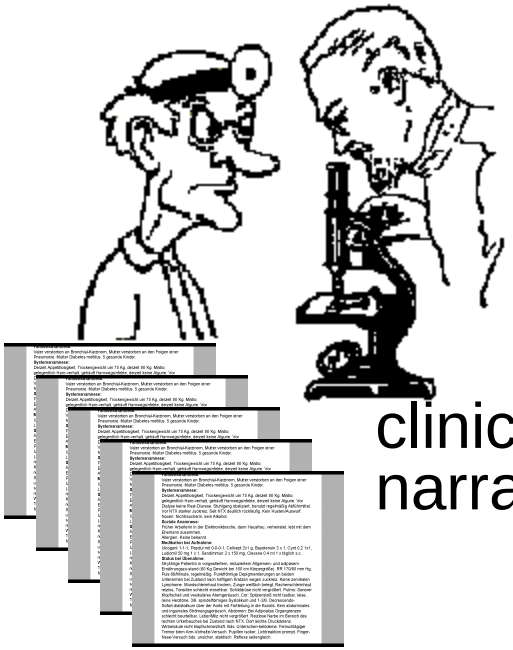
Natural Language

Structured Data

-	quality	+
+	cost	-

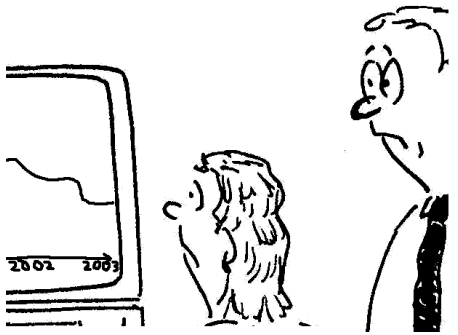
how to bridge this gap...?

Natural Language



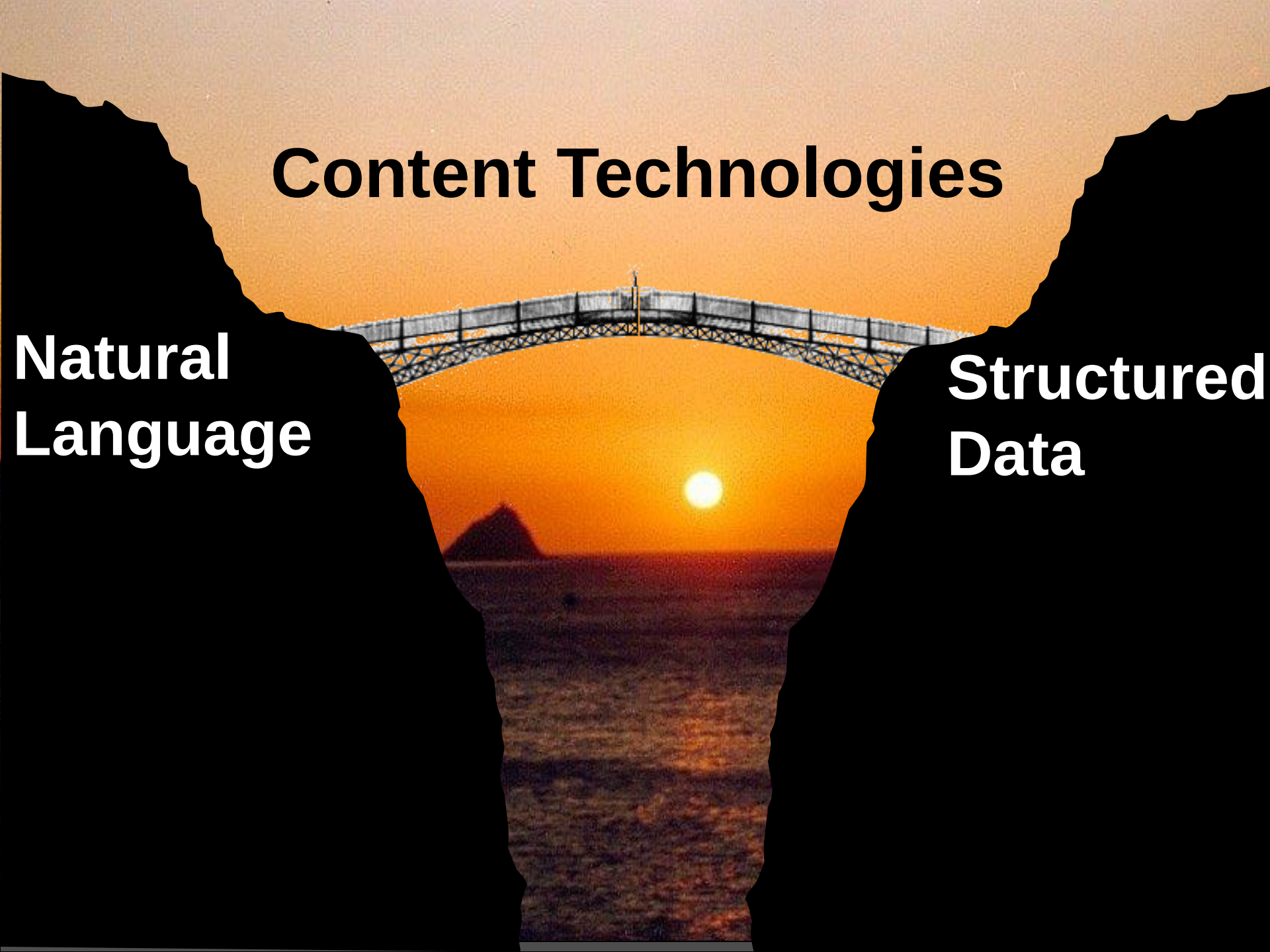
clinical narratives,

Structured Data



clinical terminologies and classifications

Content Technologies



Natural Language Processing, Linguistics and Terminology

… content technologies at the point of care

- Introduction
- **Requirements and Challenges**
- Applications of Content Technologies
  - Text Retrieval
  - Text Summarization
  - Information Extraction
- Where are we now?
- Where are we going to ? Hot Topics.

© Stefan Schulz *stschulz@uni-freiburg.de*

Medical Content Management (I)

Find me relevant documents on this topic!

Find me relevant facts about this issue!

Find me the right classification code !

Find me scientific papers which help treat this patient!

I need more information on my health problem

Medical Content Management (II)

The data is in the system, but I need to fill out a form

Can a have a brief summary of all these documents?

I need to search foreign-language documents

I want to match genomic with patient information

I want to search my health record

Two sides of the same coin

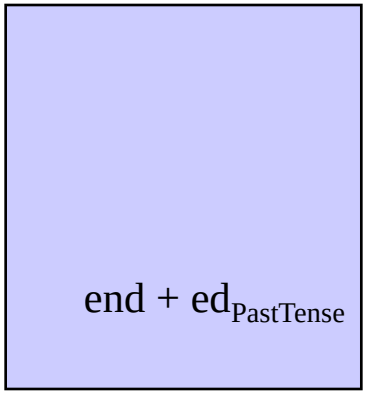
Natural Language

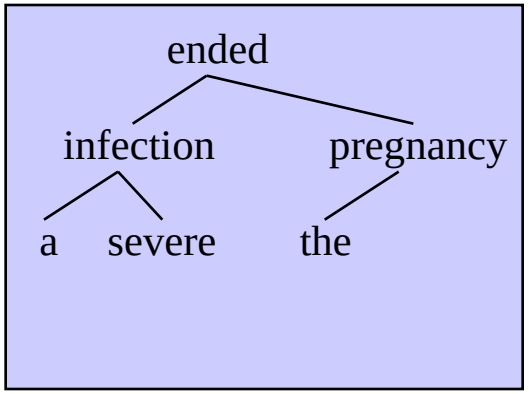
Conceptual Knowledge

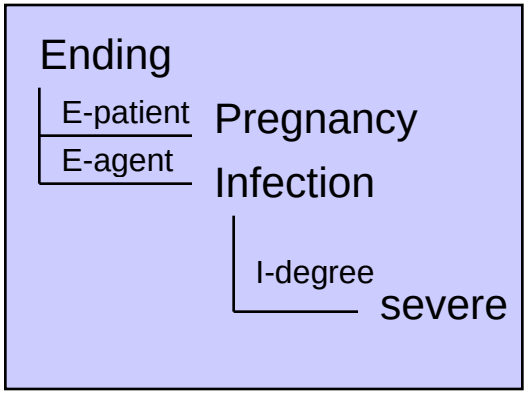
NLP Specification Levels

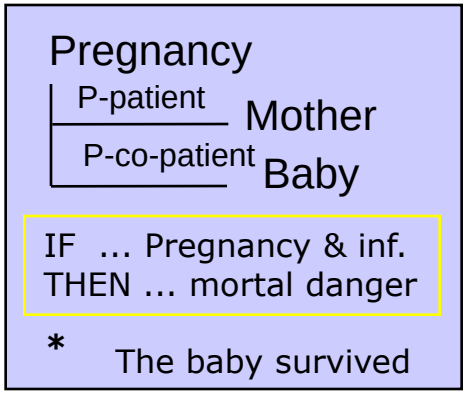
- Structure of Language
  - Morphology: end + ed, infect + ion, infect + ion + s
  - Syntax: A severe infection ended the pregnancy *vs*.
    Did a severe infection end the pregnancy? *vs*.
    The pregnancy infection severe a ended
- Meaning of Language
  - Semantics: A severe infection ended the pregnancy *vs*.
    An abortion ended the pregnancy *vs*.
    An abortion ended the heart attack
  - Understanding: A severe infection ended the pregnancy in the 28th week. The baby, however, survived.

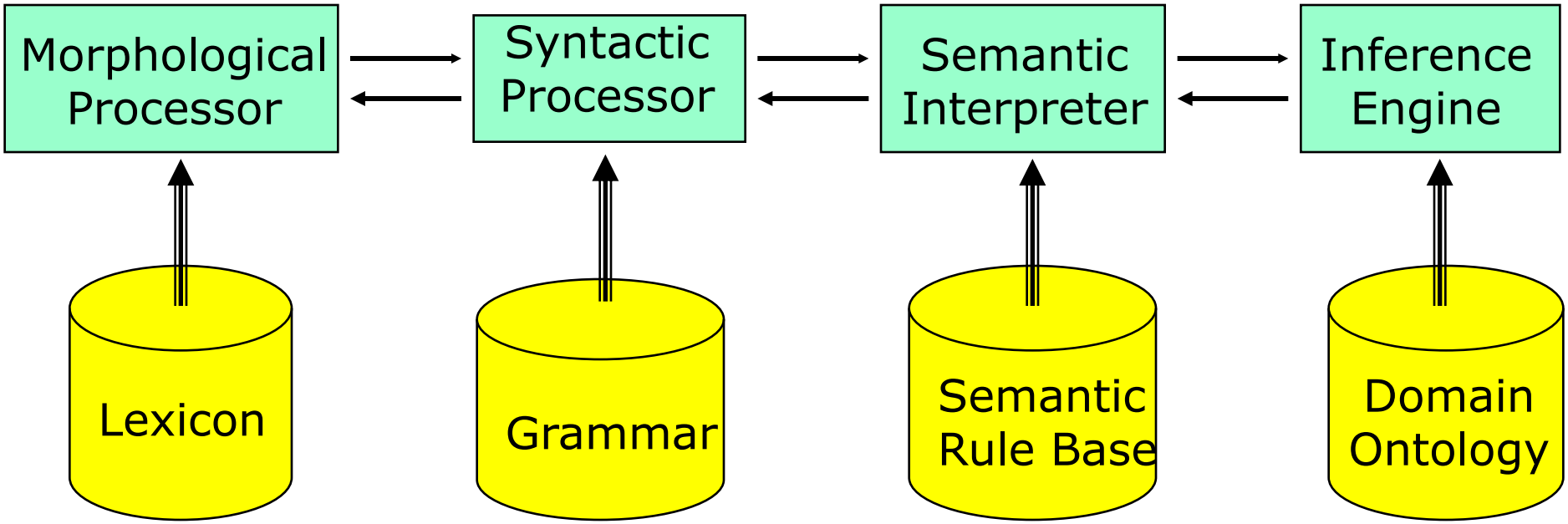
Ideal NLP Architecture











Challenges in Medical NLP

Natural Language

Conceptual Knowledge

Language (1)

- Greek, Latin word stems, Latin inflections
  thyreoglobulin, basofilia, Synechococcus elongatus
- High lexical productivity
  - Compounding
    Knochenmarktransplantation, bedrijfstandheelkunde, hipobetalipoproteinemia
  - Eponyms
    Parkinsonian disease, adenosarcoma mulleriano
  - Acronyms, Neologisms
    ECG, AIDS, ARDS, 5-FU, HWI, psbAI, GGDEF, WDWN

Language (II)

- Extragrammaticality:
  missing conceptualizations, not covered by the language description system
  paciente aidetico (adjective derivated from acronym [AIDS])
- Paragrammaticality:
  specialized meanings, jargon, short-hand utterances
  Kein Anhalt für Malignität. (incomplete sentence)
- Agrammaticality:
  erroneous input
  dictation, typing errors
- Language Mismatch:
  Physicians' vs. lay expressions

Challenges in Medical NLP

Natural Language

Conceptual Knowledge

Ontology (I)

- Proliferation of Biomedical Vocabularies:
  - Unified Medical Language System: 975,354 concepts, 2.4 million terms
  - Open Biological Ontologies:
    41 open-source ontologies
- Big Business: ‘SNOMED CT : $32.4 million contract with the U.S. National Library of Medicine (NLM)
- Vocabularies are designed for specific purposes
  (reporting, billing, indexing,...)

Ontology (II)

- Vocabularies are designed for human, not for machine use
- Computing about conceptual structures requires formally founded ontologies
- Difficulties with **formal ontologies**
  - Concept and relation definitions need do be precise and generally accepted
  - Large-Scale Construction, maintenance and validation are cost-intensive

Natural Language Processing, Linguistics and Terminology

… content technologies at the point of care

- Introduction
- Requirements and Challenges
- **Applications of Content Technologies**
  - Text Retrieval
  - Text Summarization
  - Information Extraction
- Where are we now?
- Where are we going to ? Hot Topics.

 stschulz@uni-freiburg.de

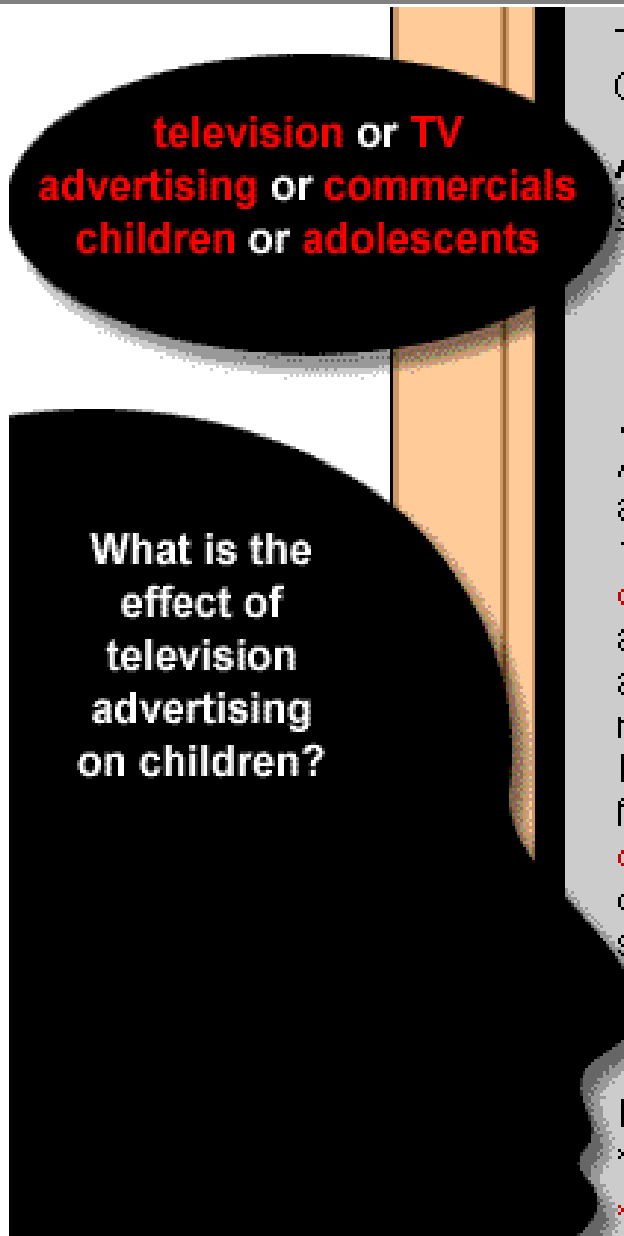
Natural Language Processing, Linguistics and Terminology

… content technologies at the point of care

- Introduction
- Requirements and Challenges
- Applications of Content Technologies
  - **Text Retrieval**
  - Text Summarization
  - Information Extraction
- Where are we now?
- Where are we going to ? Hot Topics.

 stschulz@uni-freiburg.de

Document Retrieval



TITLE
On children's mass media communication.

AUTHOR
Sharma,-Yashini

SOURCE
Psycho-Lingua. 1995 Jan-Jul; Vol25 (1-2): 85-96

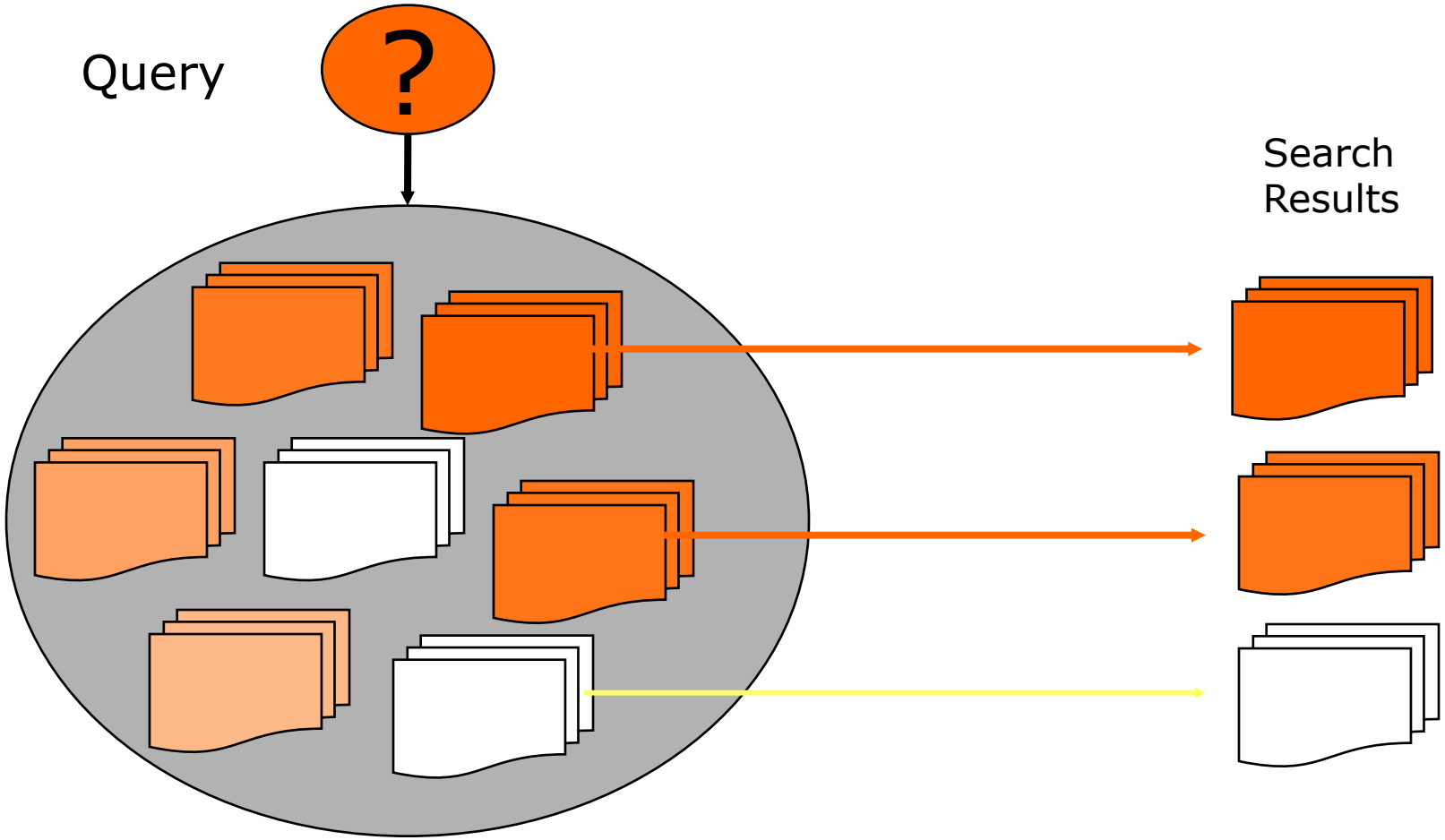
ABSTRACT
Analyzed and interpreted mass media communication that appeared in television commercial advertisements between 1991 and 1994 which were directed at children, of children, by children and only for children. The author employed content analysis for analyzing the behavioral contents of commercial advertisements as well as for children in the ads, problems of measurement, understandability and comprehensibility, language and language-play, disclaimers, etc. The study focuses mainly on disclaimers and their intelligibility in young children. Findings show that understanding of contents of commercial advertisements from the points of view of children's semantics and syntax structures determines their comprehensibility and linguistic competence. ((c)1998 APA/PsycINFO, all rights reserved)

MAJOR DESCRIPTORS
*Childhood-; *Content-Analysis; *Language-Development; *Television-Advertising

Document Retrieval: Basic Approach

- A Document Collection
  $D = \{d_1, d_2, ..., d_n\}$
- A query q
- Two Methods:
  - „Filter“ Split D into two sets D_{relq} and D_{nrelq}
    (D_{relq} = Set of relevant documents for q)
    (D_{nrelq} = Set nonrelevant documents for q)

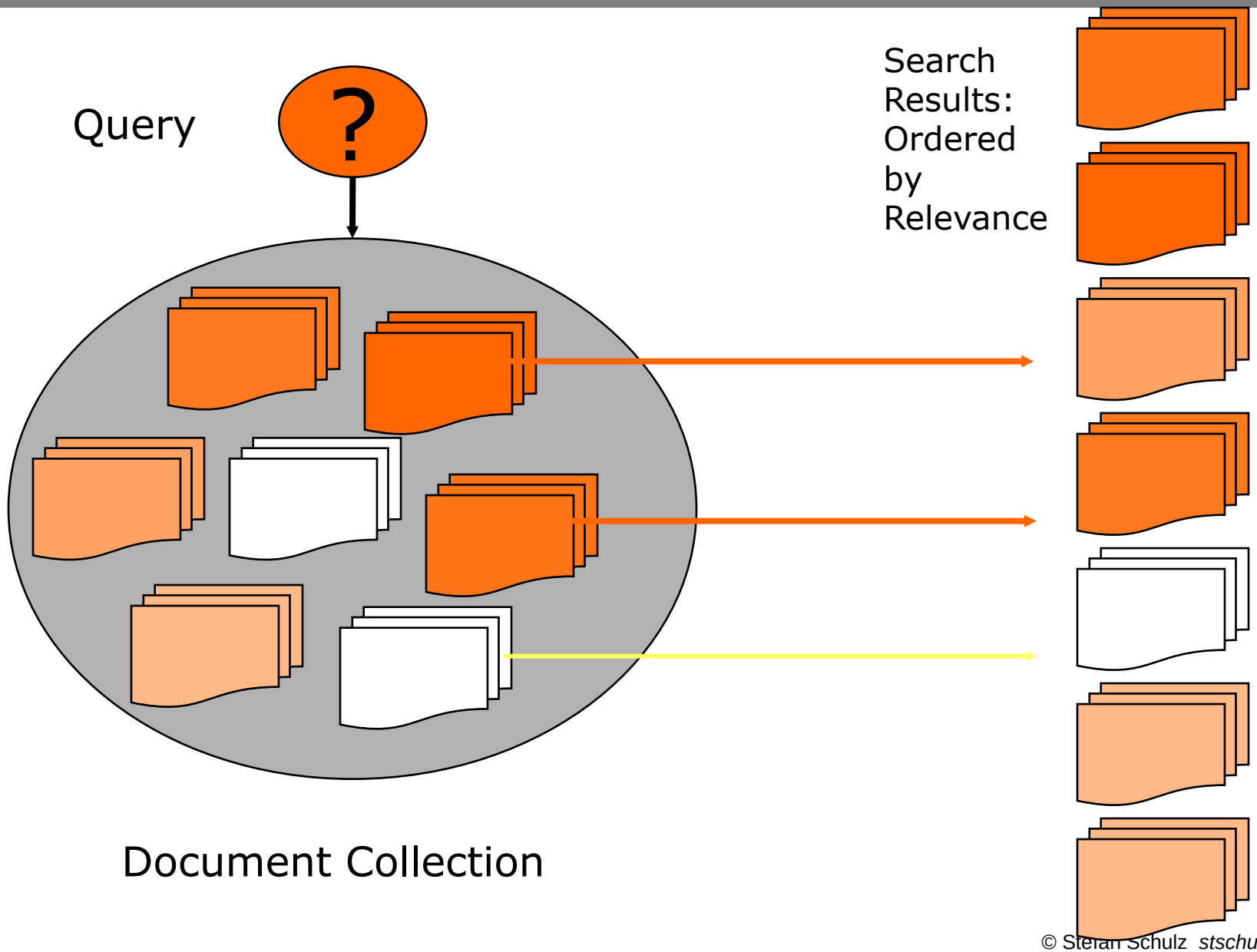
Document Retrieval



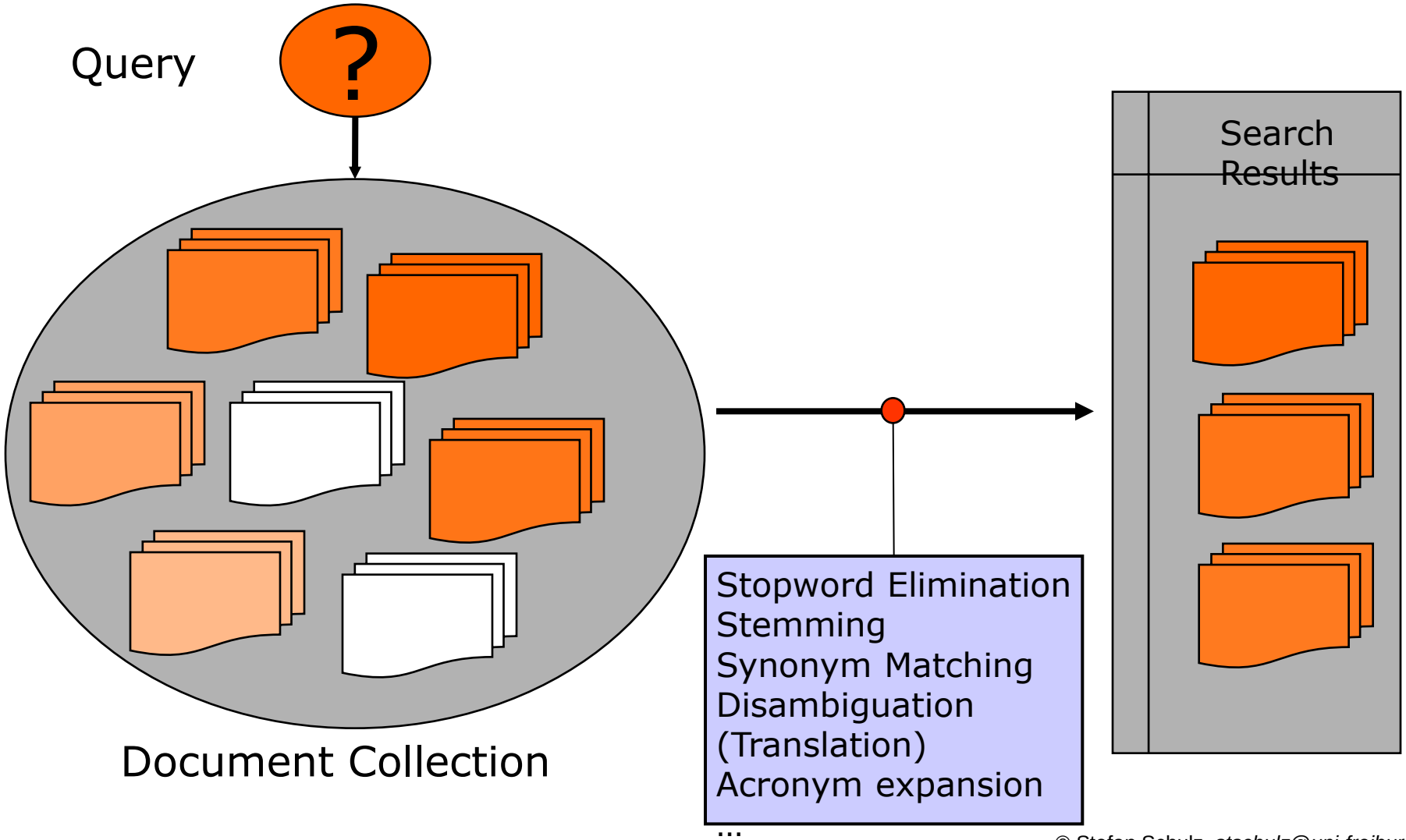
Document Retrieval: Basic Approach

- A Document Collection
  $D = \{d_1, d_2, ..., d_n\}$
- A query q
- Two Methods:
  - „Filter“ Split D into two sets D_{relq} and D_{nrelq}
    (D_{relq} = Set of relevant documents for q)
    (D_{nrelq} = Set nonrelevant documents for q)
  - „Order“ = Order by relevance:
    $D = [d'_1, d'_2, ..., d'_n]$
    with $rel(d'_i) \geq rel(d'_{i+1})$
- Combinations are possible

Document Retrieval



Syntactic / Semantic Preprocessing for Document Retrieval



Evaluation of Text Retrieval Systems

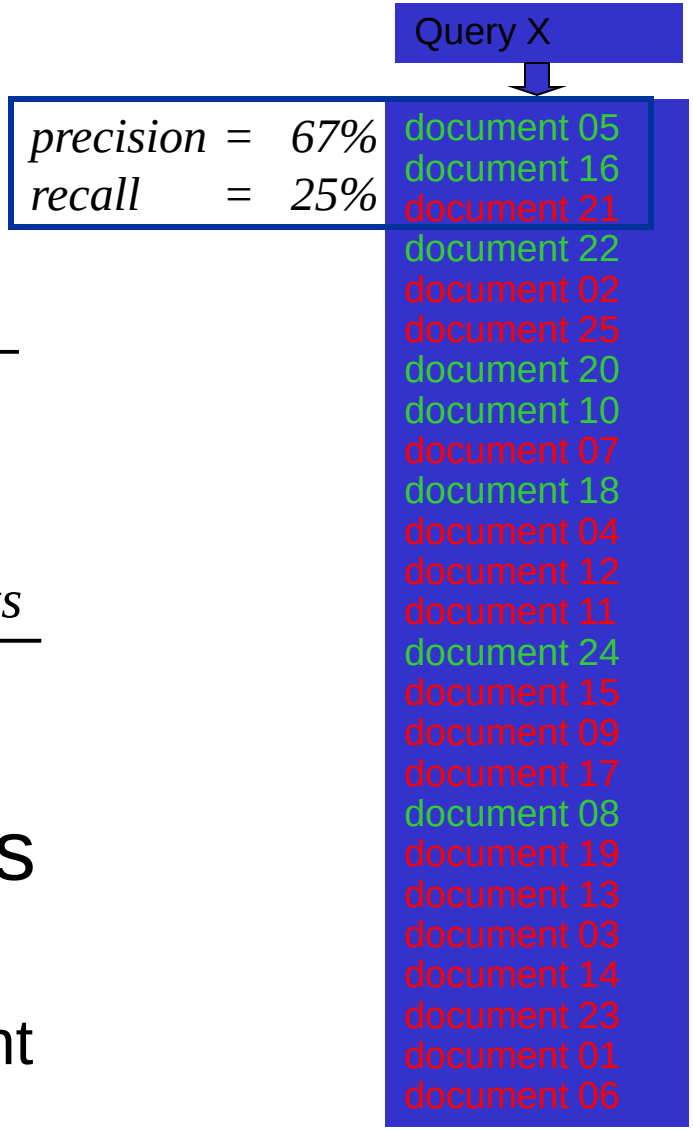
- Parameters

$$precision = \frac{n_{found+relevantDocuments}}{n_{found\_documents}}$$

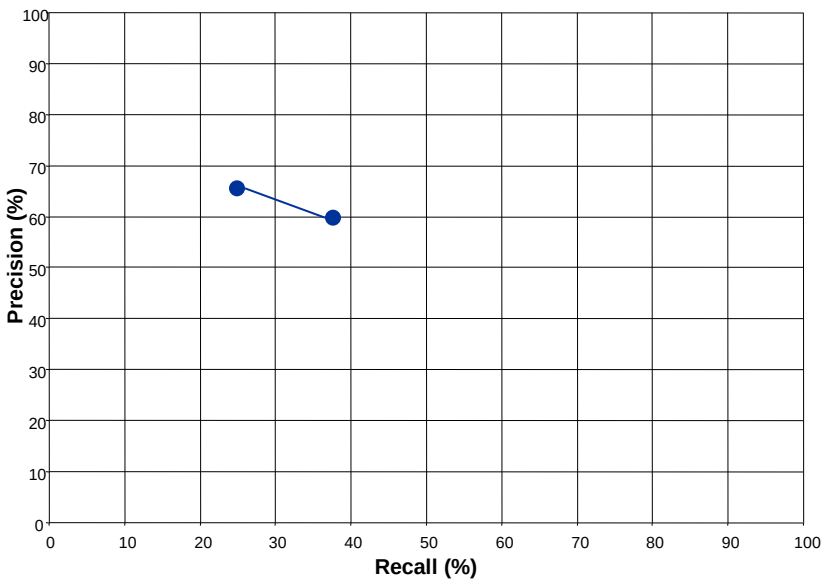
$$recall = \frac{n_{found+relevant\_documents}}{n_{relevant\_documents}}$$

- Precision/Recall-Diagrams with ranked output

  Example: 25 documents, 8 relevant



Evaluation of Text Retrieval Systems



Query X

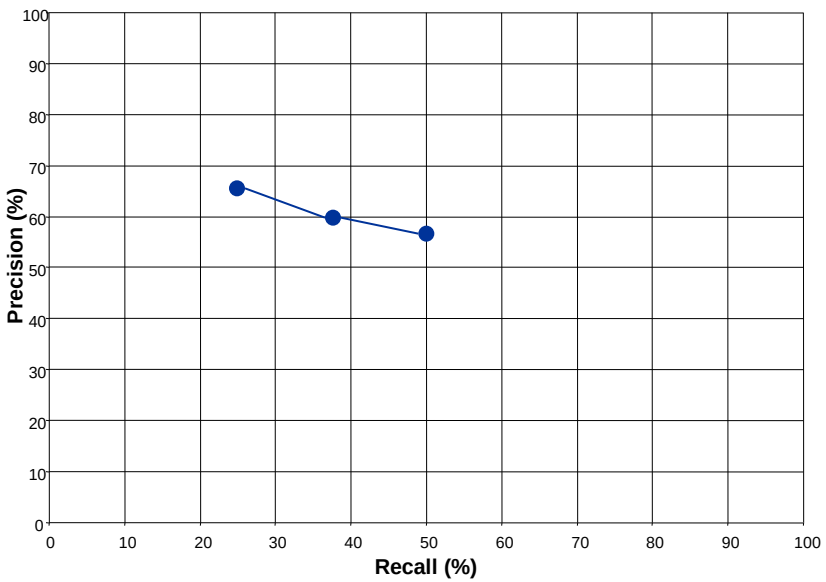
precision = *60%*
recall = *38%*

document 05
document 16
document 21
document 22
document 02
document 25
document 20
document 10
document 07
document 18
document 04
document 12
document 11
document 24
document 15
document 09
document 17
document 08
document 19
document 13
document 03
document 14
document 23
document 01
document 06

- Precision/Recall-Diagrams with ranked output

  Example: 25 documents, 8 relevant

Evaluation of Text Retrieval Systems



Query X

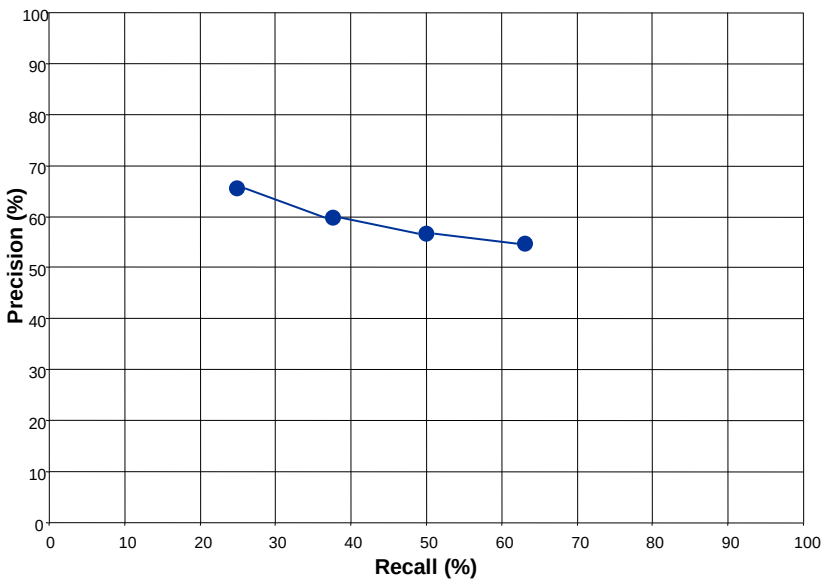
precision = *57%*
recall = *50%*

document 05
document 16
document 21
document 22
document 02
document 25
document 20
document 10
document 07
document 18
document 04
document 12
document 11
document 24
document 15
document 09
document 17
document 08
document 19
document 13
document 03
document 14
document 23
document 01
document 06

- Precision/Recall-Diagrams with ranked output

  Example: 25 documents, 8 relevant

Evaluation of Text Retrieval Systems



Query X

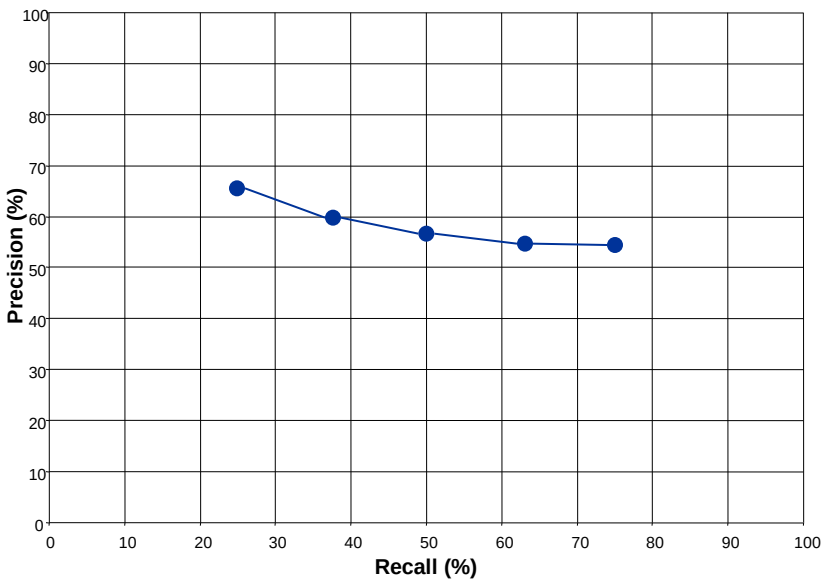
precision = 55%
recall = 63%

document 05
document 16
document 21
document 22
document 02
document 25
document 20
document 10
document 07
document 18
document 04
document 12
document 11
document 24
document 15
document 09
document 17
document 08
document 19
document 13
document 03
document 14
document 23
document 01
document 06

- Precision/Recall-Diagrams with ranked output

  Example: 25 documents, 8 relevant

Evaluation of Text Retrieval Systems



Query X

precision = *54%*
recall = *75%*

document 05
document 16
document 21
document 22
document 02
document 25
document 20
document 10
document 07
document 18
document 04
document 12
document 11
document 24
document 15
document 09
document 17
document 08
document 19
document 13
document 03
document 14
document 23
document 01
document 06

- Precision/Recall-Diagrams with ranked output

  Example: 25 documents, 8 relevant

Aspects of Medical Document Retrieval

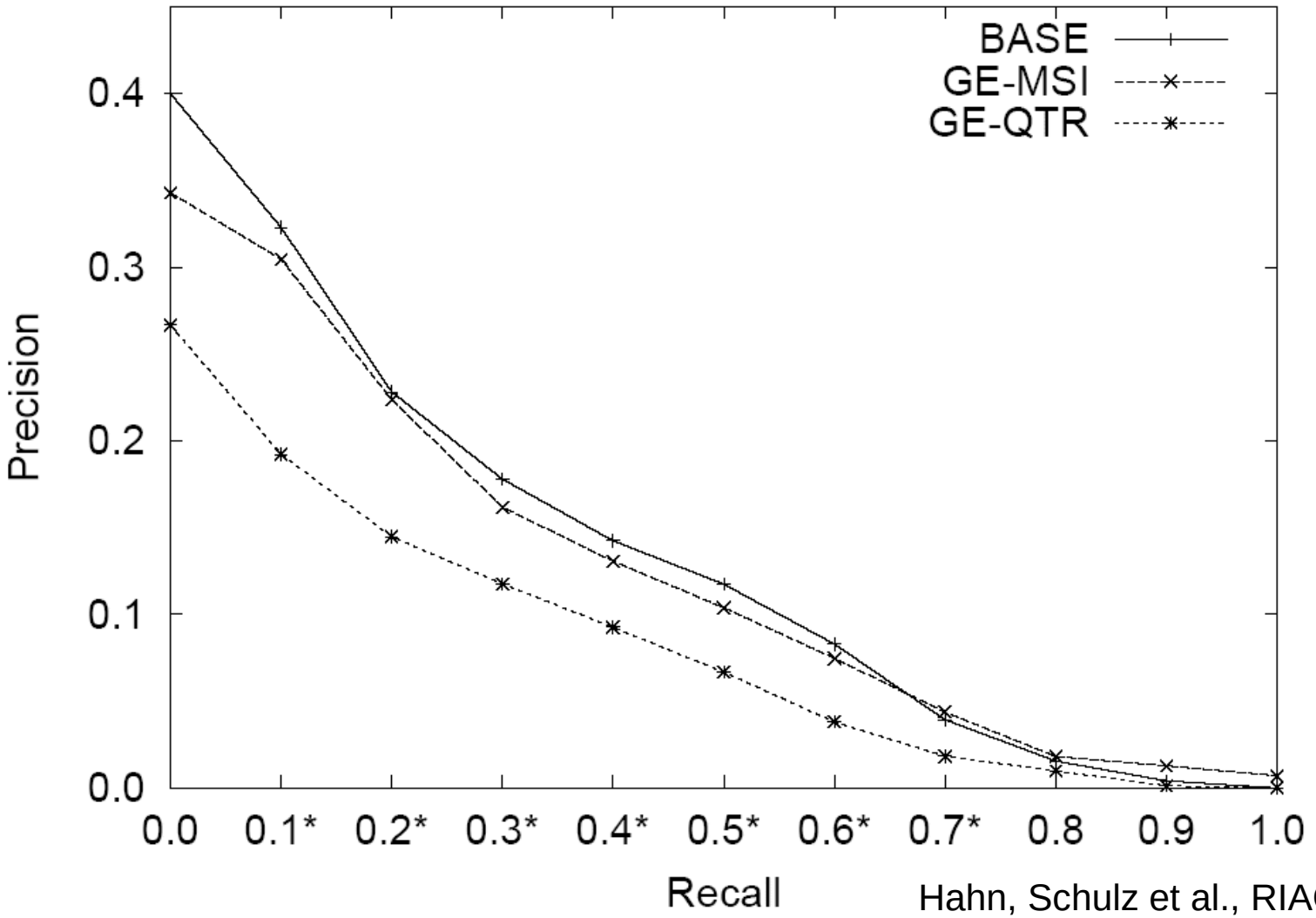
- General Observation: Users don't like operators (Boolean, truncation) in the query
- Two kinds of documents of interest:
  - Manually indexed (MEDLINE, MeSH)
  - Non indexed (EHR, Web,…)
- Automatic indexing: extracting relevant terms (topic descriptors from an indexing Alphabet, e.g. MeSH) from a document
- Retrieval in Medical Vocabularies (disease, procedure encoding)
- Cross-Language Document Retrieval

Example

- MorphoSaurus:
  - extracts meaningful word fragments using a subword lexicon
  - maps them intra- and interlingual synonymy classes (subword thesaurus)

Original Document	**Orthographic Normalization**	**Morphological Segmentation**	**Semantic Normalization**
High TSH values suggest the diagnosis of primary hypothyroidism while a suppressed TSH level suggests hyperthyroidism.	high tsh values suggest the diagnosis of primary hypothyroidism while a suppressed tsh level suggests hyperthyroidism.	high tsh value s suggest the diagnos is of primar y hypo thyroid ism while a suppress ed tsh level suggest s hyper thyroid ism.	**#up# tsh #value#** #suggest# **#diagnost# #primar# #small# #thyre# #suppress# tsh #nivell#** #suggest# **#up# #thyre#** .
Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypothyreose, ein supprimierter TSH-Spiegel spricht dagegen für eine Schilddrüsenüberfunktion.	erhoehte tsh-werte erlauben die diagnose einer primaeren hypothyreose, ein supprimierter tsh-spiegel spricht dagegen fuer eine schilddruesenueberfunktion.	er hoeh te tsh - wert e erlaub en die diagnos e einer primaer en hypo thyre ose, ein supprim iert er tsh - spiegel spricht dagegen fuer eine schilddrues en ueber funktion.	**#up# tsh - #value#** #permit# **#diagnost# #primar# #small# #thyre#** , **#suppress# tsh -** {#mirrorr# **#nivell#**} #speak# **#thyre# #up#** #function# .
A presença de valores elevados de TSH sugere o diagnóstico de hipotireoidismo primário, enquanto níveis suprimidos de TSH sugerem hipertireoidismo.	a presenca de valores elevados de tsh sugere o diagnostico de hipotireoidismo primario, enquanto niveis suprimidos de tsh sugerem hipertireoidismo.	a presenc a de valor es elevad os de tsh suger e o diagnost ico de hipo tireoid ismo primari o, enquanto niveis suprimid os de tsh suger em hiper tireoid ismo.	#actual# **#value# #up# tsh** #suggest# **#diagnost# #small# #thyre# #primar#** , **#nivell# #suppress# tsh** #suggest# **#up# #thyre#** .

MorphoSaurus: Evaluation



Hahn, Schulz et al., RIAO 2004

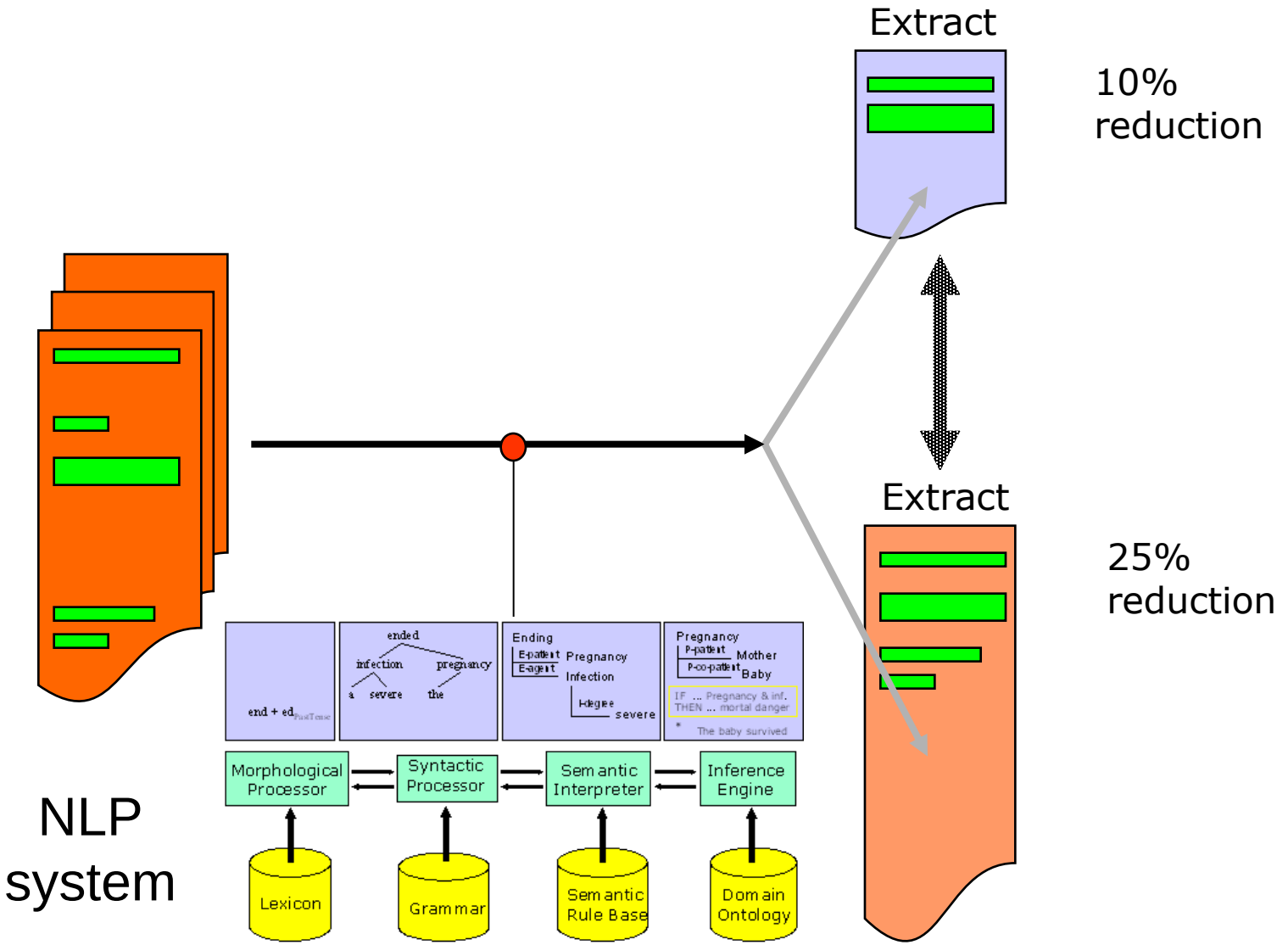
Natural Language Processing, Linguistics and Terminology

… content technologies at the point of care

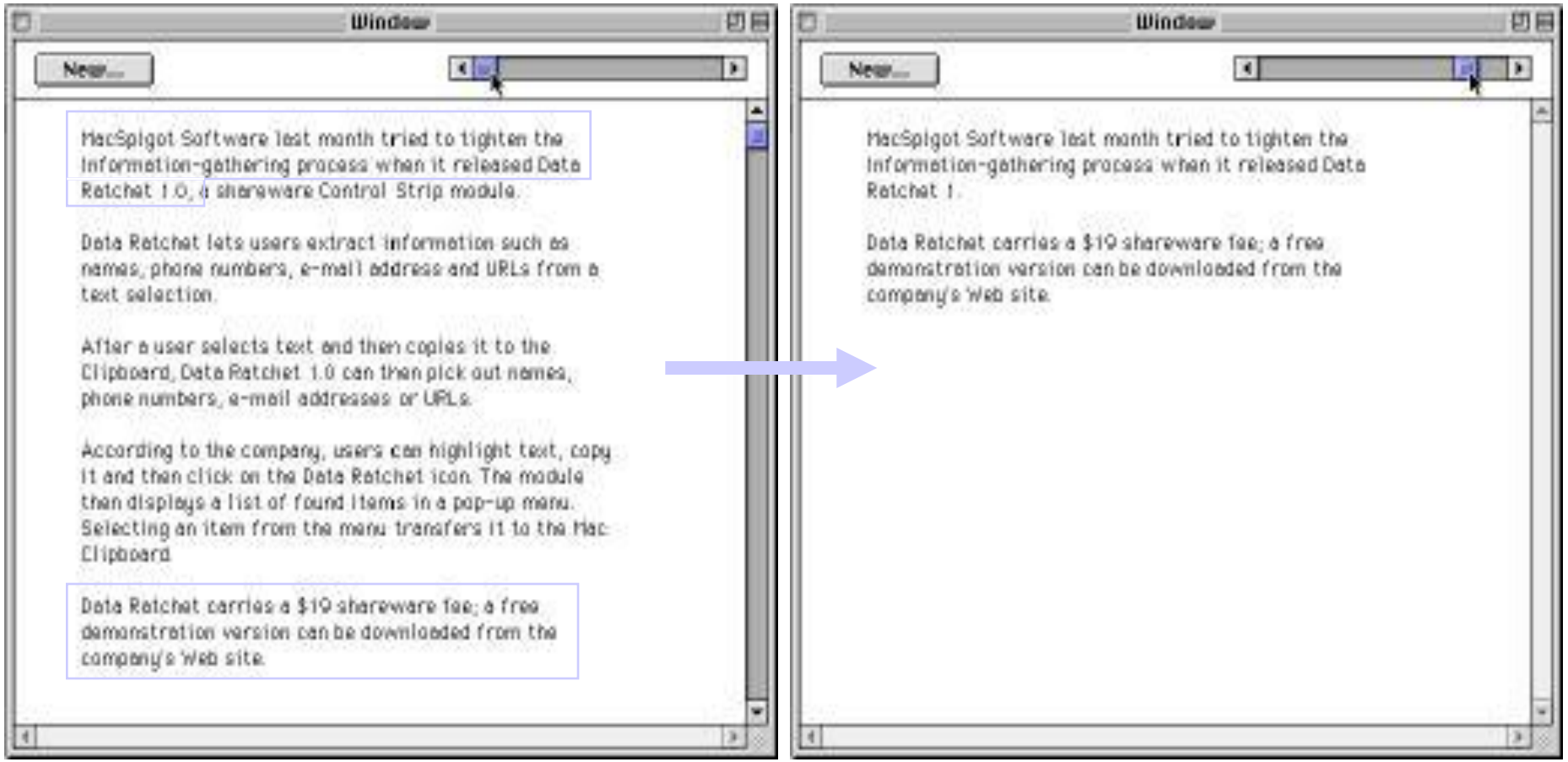
- Introduction
- Requirements and Challenges
- Applications of Content Technologies
  - Text Retrieval
  - **Text Summarization**
  - Information Extraction
- Where are we now?
- Where are we going to ? Hot Topics.

 stschulz@uni-freiburg.de

Text Summarization



MTT Text Summarization System



Perspectives in Biomedical Text Summarization

- Compiling a well-readable and structured digest of the most relevant data of a patient at the point of care
- Semi-automated generation of discharge summaries
- Compiling summaries of a clinical topic across different patients
- Summarize facts from medical literature

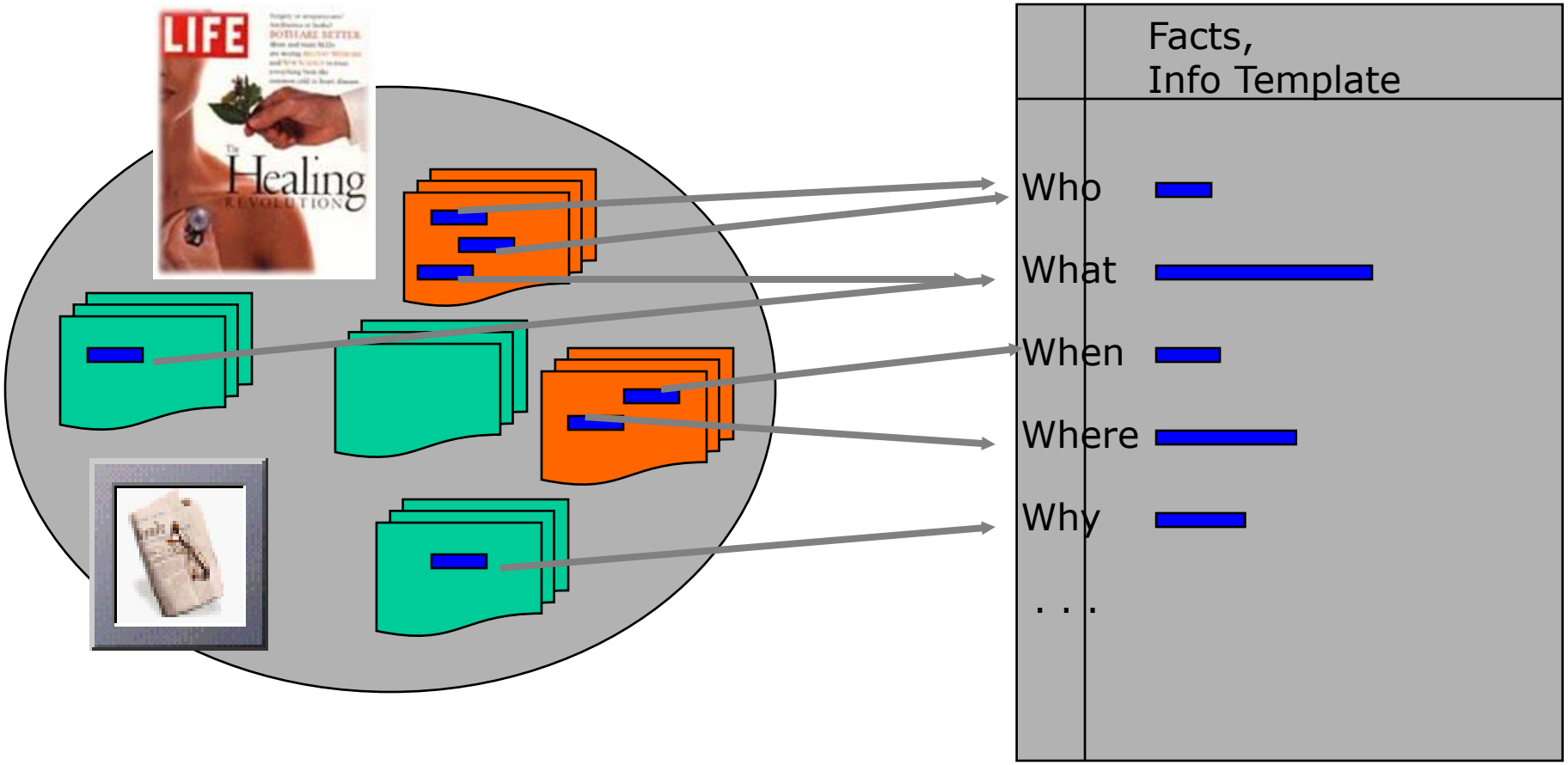
Natural Language Processing, Linguistics and Terminology

… content technologies at the point of care

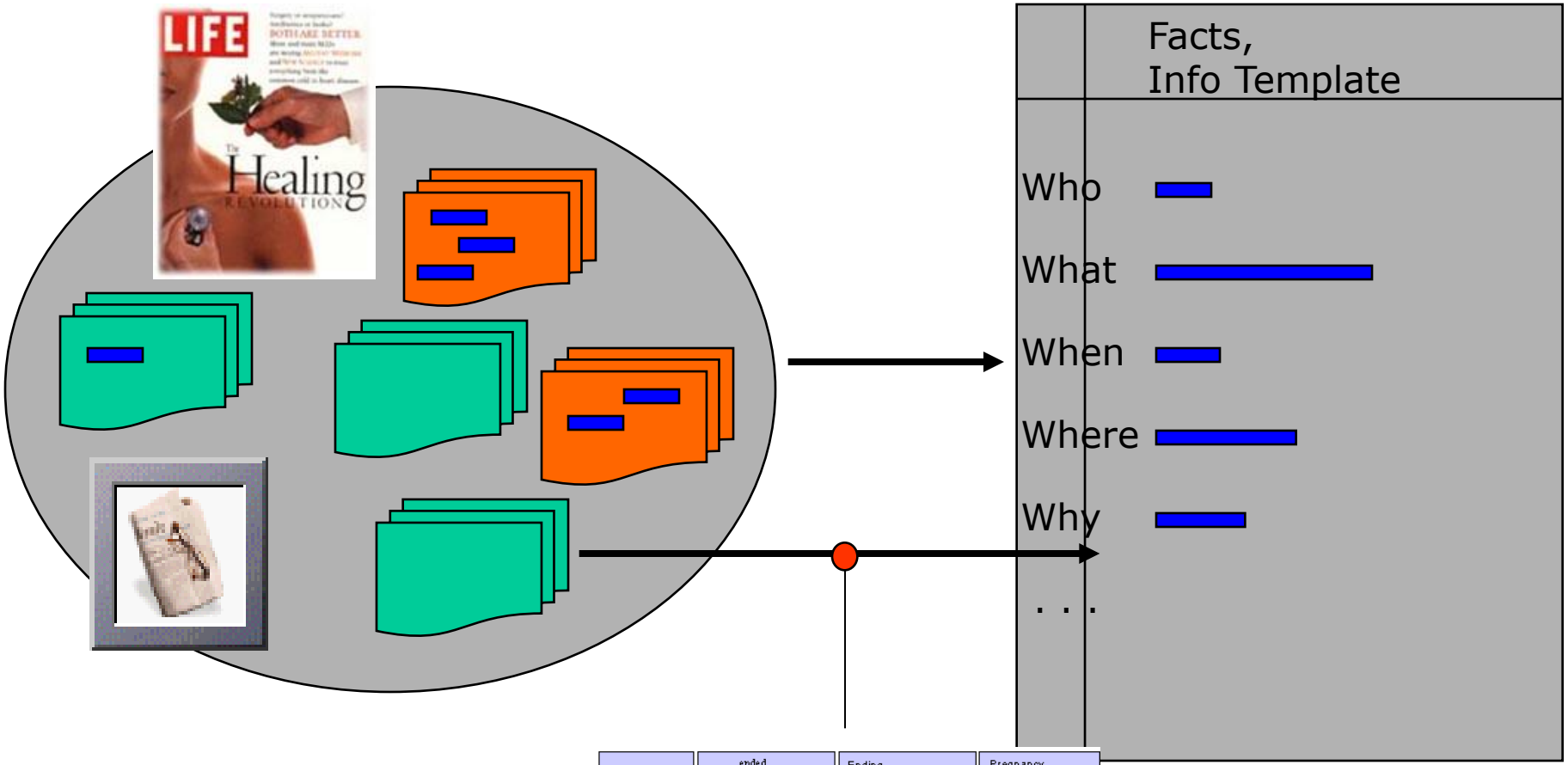
- Introduction
- Requirements and Challenges
- Applications of Content Technologies
  - Text Retrieval
  - Text Summarization
  - **Information Extraction**
- Where are we now?
- Where are we going to ? Hot Topics.

© Stefan Schulz *stschulz@uni-freiburg.de*

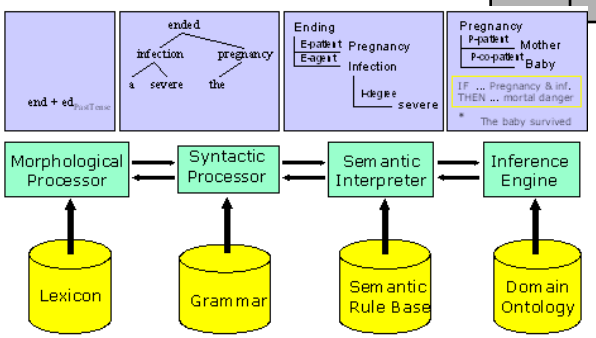
Information Extraction



Information Extraction



NLP system



Information Extraction

Syntactic Analysis & Semantic Tagging

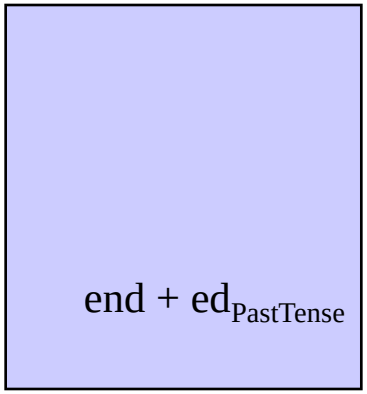
He succeeds Jack Harper, a company founder who was chairman ...

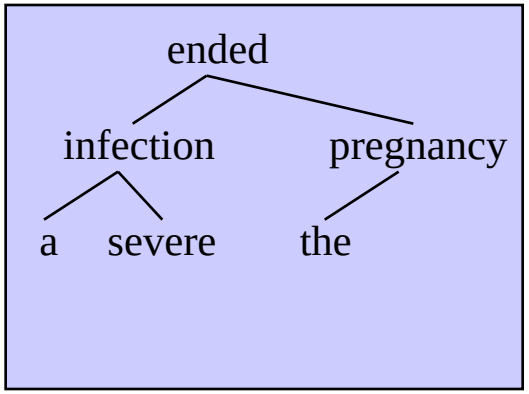
SUBJ: Terms: He;
Classes: <Generic Person>
VERB: Terms: succeeds;
Root: succeed;
Mode: active, affirmative
OBJ: Terms: Jack Harper, a c. f.
Classes: <Person Name>, <Generic Org.>, <Generic P.>
REL-OBJ: Terms: who was chairman
Classes: <Event>

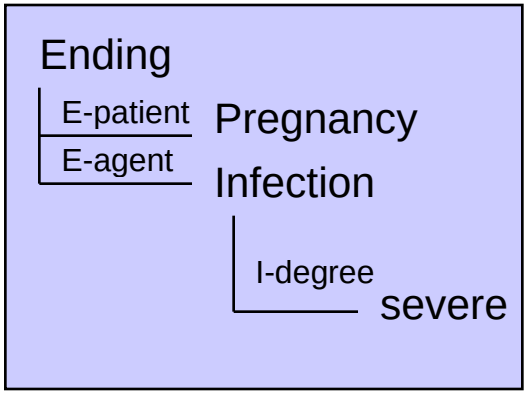
Extraction Rule

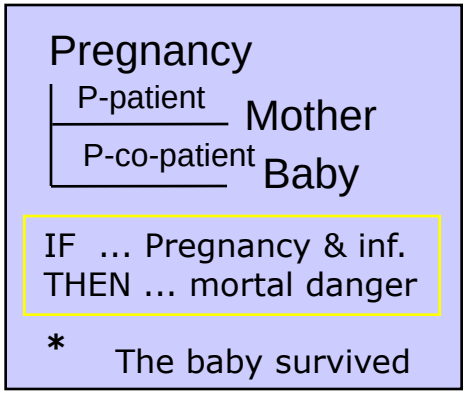
CONCEPT TYPE: *Succession Event*

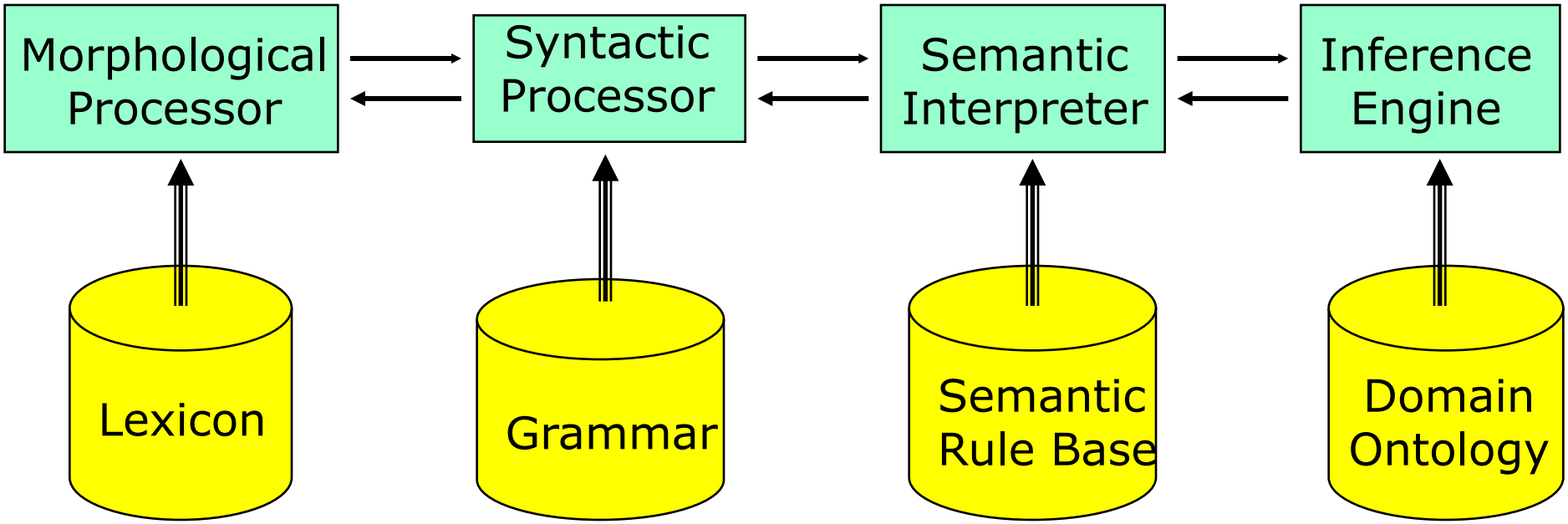
Ideal NLP Architecture











Information Extraction

Syntactic Analysis & Semantic Tagging

He succeeds Jack Harper, a company founder who was chairman ...

SUBJ: Terms: He;
Classes: <Generic Person>
VERB: Terms: succeeds;
Root: succeed;
Mode: active, affirmative
OBJ: Terms: Jack Harper, a c. f.
Classes: <Person Name>, <Generic Org.>, <Generic P.>
REL-OBJ: Terms: who was chairman
Classes: <Event>

Extraction Rule

CONCEPT TYPE: *Succession Event*
CONSTRAINTS:
SUBJ: (extract: Person In)
Classes: <Generic Person>
VERB: Root: succeed;
Mode: active
OBJ: (extract: Person Out)
Classes: <Person Name>

Extracted Template

[*Succession Event*
[Person In: He]
[Person Out: Jack Harper, a company founder]]

Two Approaches to Building Information Extraction Systems

- Classical Knowledge Engineering
  - grammars are hand-crafted
  - domain patterns are discovered by a human expert through introspection or inspection of a corpus
  - laborious tuning and “hill climbing” for maintenance/updates
- Automatically Trainable Systems
  - use statistical methods when possible (need training data!)
  - learn rules from annotated corpora (e.g. FRAMED)
  - learn rules from interaction with the user(s)
  - easy maintainability, extensibility and adaptivity

Perspectives in Biomedical Information Extraction

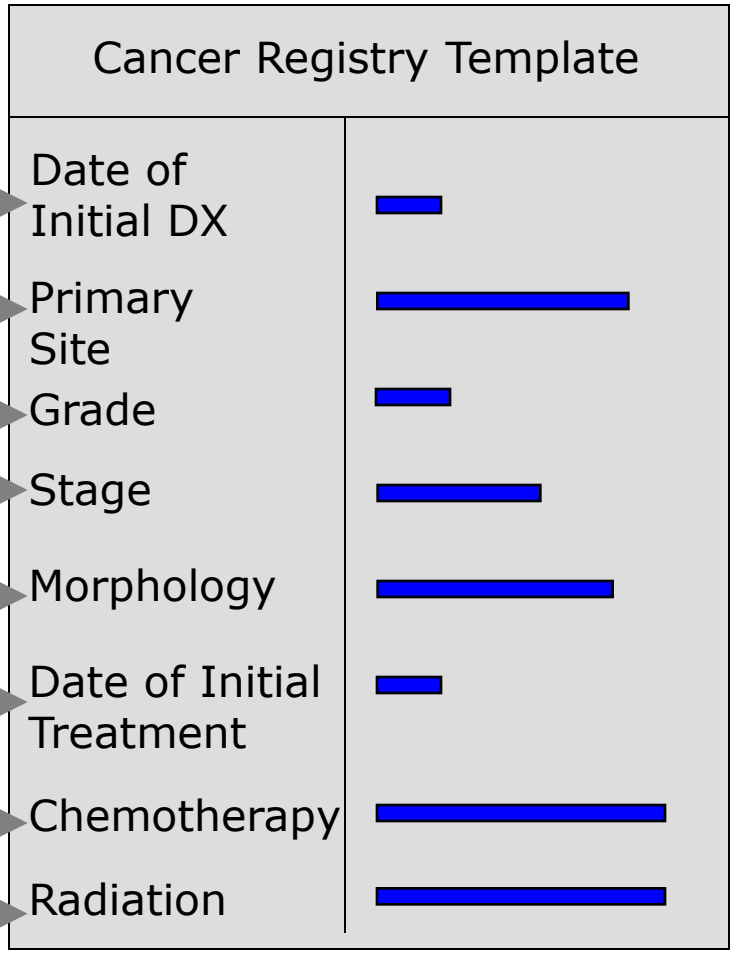
- In Basic Research (Molecular Biology) Information Extraction from Literature, e.g. Medline-Abstracts (e.g. facts about Protein-Protein Interaction)
- Information Extraction from narratives in the Electronic Health Records to meet various needs for structured documentation, e.g. tumor documentation
- Managing metadata by extracting semantic tags for XML markup of medical text (e.g. using CDA)

```
<body>
 <section>
  <section.title>Procedure</section.title>
  <paragraph>
   <healthcare.code identifier="P5-20100"
    name.of.coding.system="SNM3"
    local.coding.system="N">Chest X-Ray
   </healthcare.code>
  </paragraph>
 </section>
 <section>
  <section.title>Findings</section.title>
  <paragraph>RLL nodule</paragraph>
 </section>
 <section>
  <section.title>Impressions</section.title>
  <paragraph>Nodule in the RLL, suggestive of
   malignancy.</paragraph>
 </section>
 <section>
  <section.title>Recommendations</section.title>
  <paragraph>I notified the ordering physician of this
   finding.</paragraph>
 </section>
</body>
</LevelOne>
```

Example: Cancer Documentation

EHR

shadow was pointed out on a routine chest X-ray film, but she had no further examination. Physical examination on admission revealed purpura of the upper and lower extremities, swelling of the gums and tonsils, but no symptoms showing the complication of myasthenia gravis. Hematological tests revealed leucocytosis: WBC count 68 700/μl (blasts 11.5%, myelocytes 0.5%, bands 2.0%, segments 16.0%, monocytes 65.5%, lymphocytes 4.0%, atypical lymphocytes 0.5%), Hb 7.1 g/dl (reticulocytes 12%) and a platelet count of 9.1 × 10^4/μl. Further laboratory examination revealed elevated serum lactic dehydrogenase (589 U/l), vitamin B_{12} (2010 pg/ml) and ferritin (650. 0 ng/ml). Human chorionic gonadotropin and [alpha]-fetoprotein levels were normal. A bone marrow aspiration revealed hypercelllar bone marrow with a decreased number of erythroblasts and megakaryocytes and an increased number of monoblasts that were positive for staining by [illegible] butyrate esterase and negative for staining by naphthol ASD chloroacetate esterase. Chest X-ray upon admission revealed a mediastinal mass and an elevated left diaphragm . Computed tomography (CT) of the chest showed a left anterior mediastinal mass . Based on these findings, the patient was diagnosed with a mediastinal tumor accompanied by AMoL. First, in June 1991, the patient was treated with DCMP therapy: daunorubicin (DNR) (25 mg/m^2, days 1, 2, 3, 4, 6 and 8), cytosine arabinoside (Ara-C) (100 mg/m^2, days 1-9), 6MP-riboside (6-MP) (70 mg/m^2, days 1-9) and prednisolone (PSL) (20 mg/m^2, days 1-9), followed by five courses of consolidation chemotherapy [1, DCMP; 2, ID-Ara-C:adriacin (ADR), vincristine (VCR), Ara-C, PSL; 3, DCMP; 4, ID-Ara-C; 5, A-triple V: Ara-C, VP-16, VCR, vinblastine (VBL)]. After induction chemotherapy, a hematological examination and bone marrow findings had improved to normal, and complete remission was attained. Chest CT scan after chemotherapy in November 1991 revealed regression of the mediastinal tumor . An invasive thymic tumor was suspected and surgery was undertaken in January 1992. The tumor (50 × 45 × 45 mm), located mainly in the anterior mediastinum, was strongly adhered to the adjacent tissues. Resection of the tumor included the left upper lobe of the lung, the phrenic nerve and pericardium. The histological finding was that the tumor cells have large, vesicular nuclei and prominent nucleoli, but keratinazation was unclear . The results of immunohistochemical finding of anti-TdT was negative. From these findings, we diagnosed poorly or moderately differentiated squamous cell carcinoma of the thymus. The postoperative course was uneventful. The patient underwent radiation therapy of the mediastinum and left hilum at doses of 4000 cGy delivered over 4 weeks. She was



Other NLP Techniques

- Text mining:
  discovery by computer of new, previously unknown information, by automatically extracting information from text
- Question answering:
  finding answers from a vast amount of underlying text
- Machine Translation:
  translate text from one natural language into another
- Natural Language Generation
  generating text from an abstract semantic representation, e.g. multilingual generation of patient information
- Speech technologies
  processing spoken language

Natural Language Processing, Linguistics and Terminology

… content technologies at the point of care

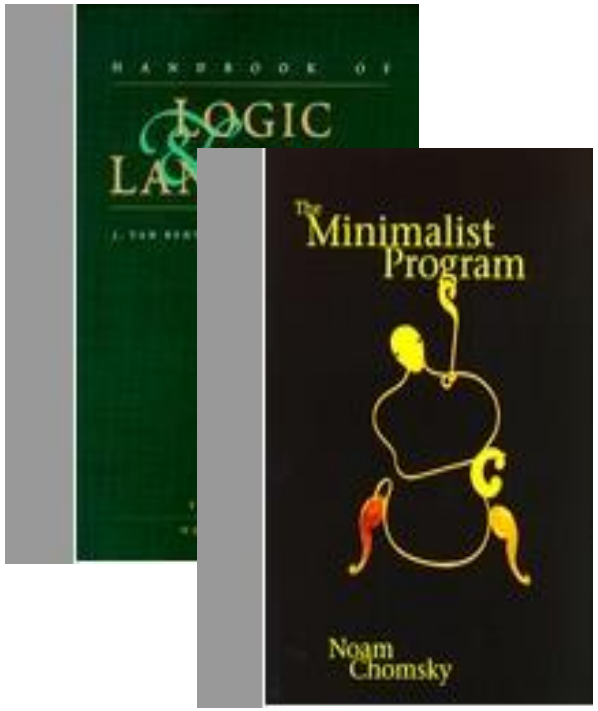
- Introduction
- Requirements and Challenges
- Applications of Content Technologies
  - Text Retrieval
  - Text Summarization
  - Information Extraction
- **Where are we now?**
- Where are we going to ? Hot Topics.

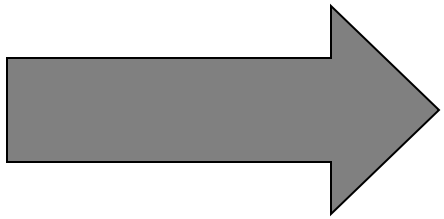
 stschulz@uni-freiburg.de

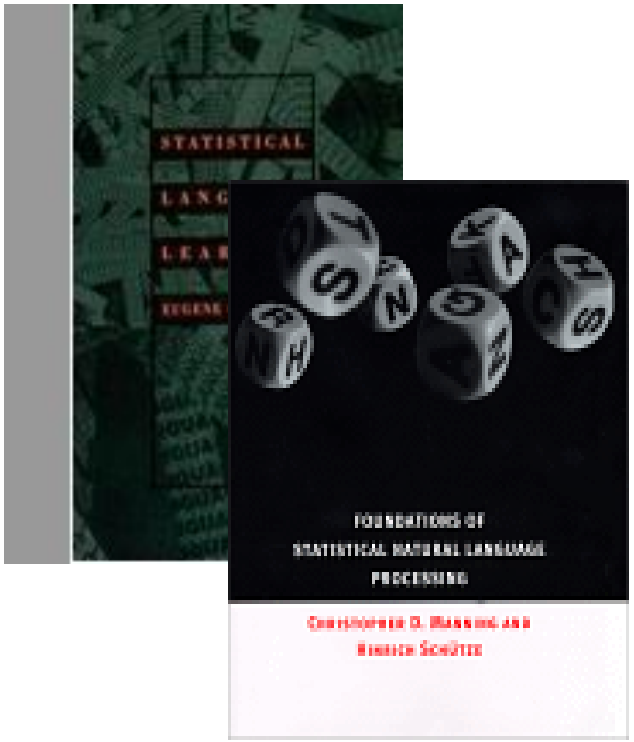
NLP Methodologies and Sources

- Vector Space Model (document retrieval)
- Finite State Automata (information extraction)
- N-gram data (bigrams, trigrams)
- Morphological tools (stemmers)
- Text segmenters (coherent portions of a larger text)
- POS (part-of-speech) taggers
- Enormous amounts of corpora
- Enormous amounts of resources (lexicons, grammars, ontologies)
- Emergence of uniform evaluation standards

Paradigm Shift in Computational Linguistics







Rationalism

Empiricism

Grand Challenges for Content Technologies I

- Ambiguity of Language
  - lexical ambiguity (polysemy, homonymy)
    - [biological] *cell* vs. *cell* [in a monastery, prison]
  - syntactic ambiguity (e.g., prepositional phrase attachment)
    - extraction *[of the transplant [with a scalpel]]*
    - *[extraction]* {of the transplant} *[with a scalpel]*
  - semantic ambiguity (e.g., quantifier scope)
    - *each* sample showed an increased Ph value
      - specific reading: each sample showed *exactly one* increased PH
      - unspecific reading: each sample showed *some* increased PH

Large ambiguity rates lead to excessive demands for computational resources (i.e, intractability)!

Grand Challenges for Content Technologies II

- Computational Complexity
  - worst case complexity for grammars
    - finite state automata / linear grammars O(n)
    - pushdown automata / context-free grammars $O(n^3)$
    - unification grammars, dependency grammars NP-complete
  - decidability of logics
    - propositional logic, monadic first-order predicate logic decidable
    - first-order predicate logic semi-decidable
    - n^{th}-order predicate logic ($n > 1$), modal logics undecidable

Grand Challenges for Content Technologies III

- Dynamics of Real-world Language
  - Morphological productivity
    - {cherry/rice-corn/...}-sized biopsy material
    - plasmacellular infiltration
  - Neologisms, abbreviations high ambiguity, e.g.
    - LCA = leukocyte common antigen
      LCA = left coronary artery
      LCA = lymphocyte common antigen.
  - Erroneous and underspecified input requires robust devices

Special Challenges for Medical Content Technologies (I)

Language engineering

- (Generalized) Quantification
  - *most of* the samples
  - *two from eight* samples contained ...
- Negation (ako Quantification)
  - *no* evidence for X
  - can be *excluded* from further consideration
- Certainty, Strength or Severity Assessments
  - without *significant* findings
  - is *strongly* infiltrated
- Temporal Expressions
  - in the *first two weeks*
  - *First*, take X, *then* do Y (guidelines)
- Coordination
  - *invasive and metastatic, highly differentiated* carcinoma

Special Challenges for Medical Content Technologies (II)

Knowledge Engineering

- Cross-mapping and unification of heterogeneous terminologies (UMLS)
- Providing large terminologies as formal systems
- Use state-of-the art of ontological analysis and engineering
- Bridge semantic differences and reach consensus about precise meaning of terms
- Manage labor-intensive ontology construction and maintenance
- Build / Adapt inference engines (e.g. terminological classifiers) capable of dealing with > 10^5 concepts

Natural Language Processing, Linguistics and Terminology

… content technologies at the point of care

- Introduction
- Requirements and Challenges
- Applications of Content Technologies
  - Text Retrieval
  - Text Summarization
  - Information Extraction
- Where are we now?
- **Where are we going to ? Hot Topics.**

© Stefan Schulz *stschulz@uni-freiburg.de*

What's Hot?

- Focus on Natural Language, Terminology, Knowledge, Ontology:
  - AMIA 2003: 48 out of 156 papers
  - eHealth Projects in the EU 6th Framework Program: 9 out of 16 projects
- The Semantic Web Initiative: Controversy (is a global ontology feasible and desirable ?)
- Google & Co...Traditional IR assumptions are challenged in the context of the web (billions of pages...)

What's Hot?

- Combining *multiple media* (text, speech, graphics, sound, movies, tables, etc.) for summarization, question answering, etc.
- Combining *multiple modalities* (spoken/written language, gestures, tactile and haptic movements) in a versatile user interface
- *Crosslingual* and *multilingual* document retrieval, summarization, question answering
- Multilingual Dictionaries

What's Hot?

- A new generation of formally founded biomedical ontologies
- The Semantic Electronic Health Record
- Named entity recognition challenged by the deluge of new proper names from the bio domain
- Use huge (Terabyte !) medical corpora (from all sources including anonymized EHR data) for the discovery of domain and linguistic knowledge
- Use content technologies to match genotype information (Bio-DBs) with phenotype information (EHR).

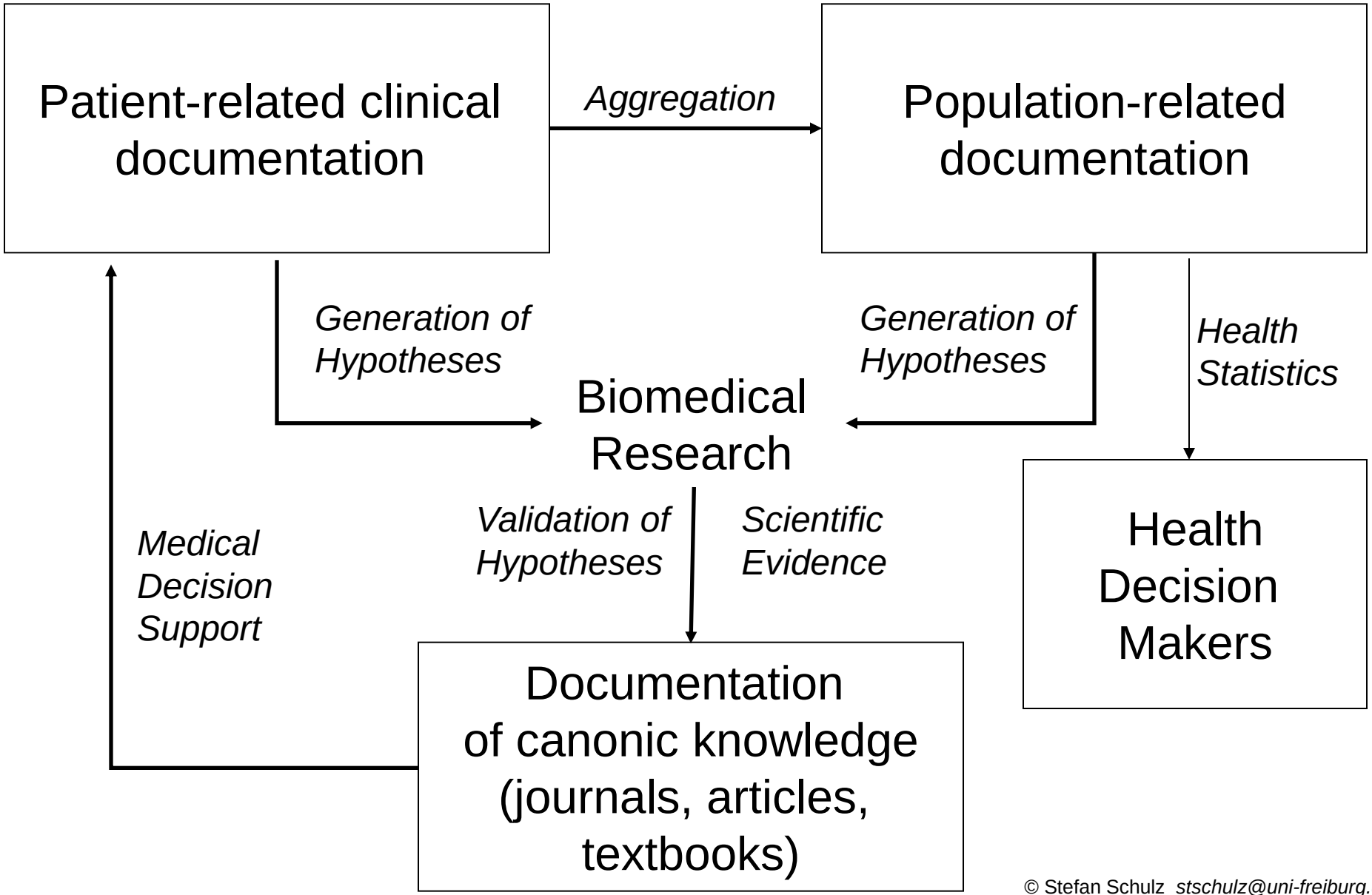
Activities

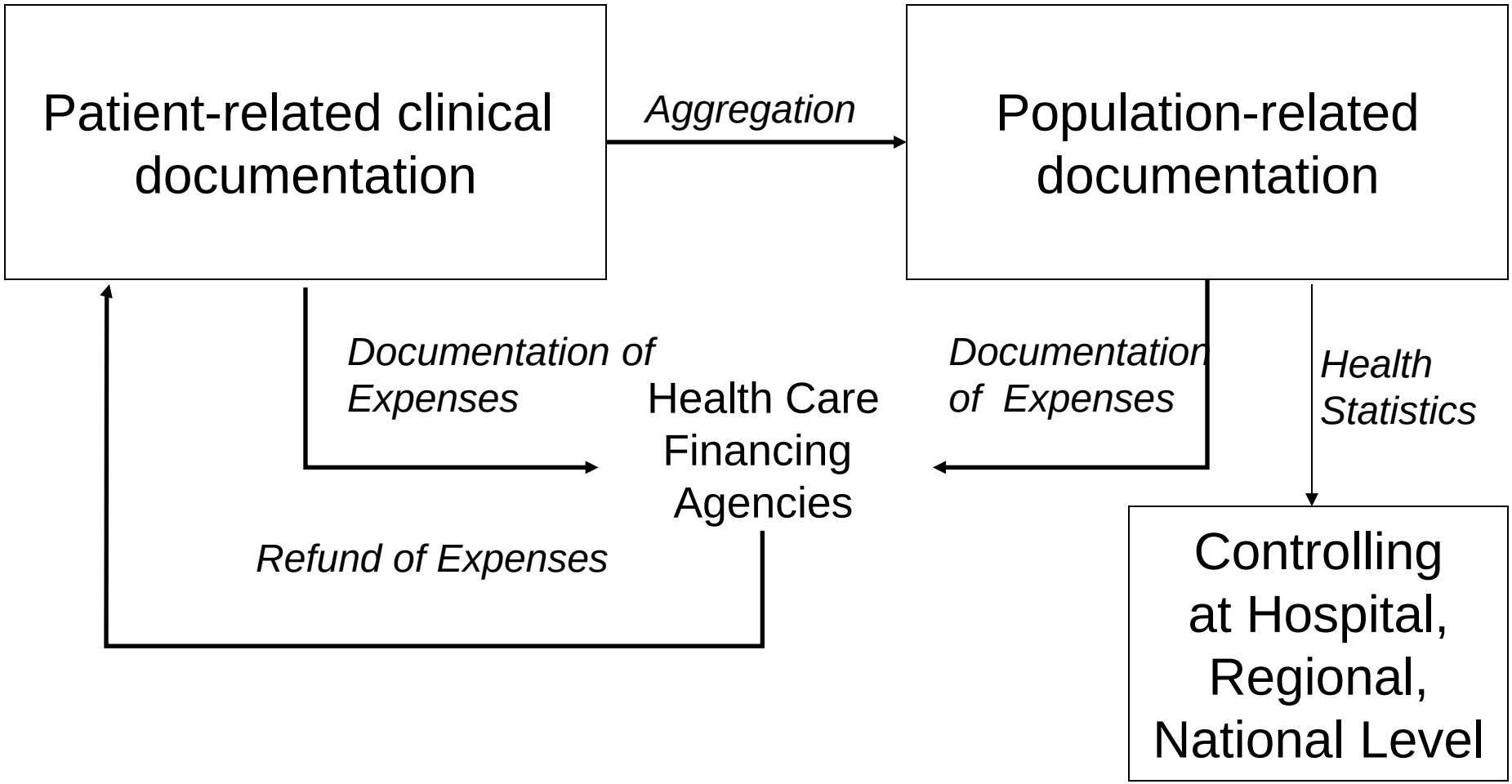
- EU 6th Framework Program:
  Network of Excellence “SemanticMining”
  (Semantic Interoperability and Data Mining in Biomedicine): 2004 – 2006, 25 Partners
  www.semanticmining.org
- AMIA Special Interest Group KR-SIG
  “Formal (Bio)medical Knowledge Representation”, founded 2003
- Workshop KR-MED on 1 June in Whistler/Canada
  www.coling.uni-freiburg.de/pub/kr-med

Děkuju / Thank You

Stefan Schulz

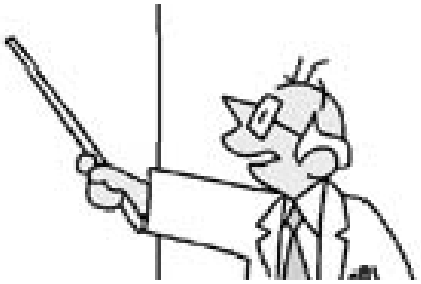
Department of Medical Informatics
Freiburg University Hospital
http://www.imbi.uni-freiburg.de/medinf
stschulz@uni-freiburg.de



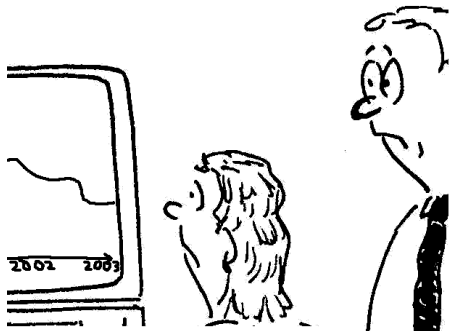


Adressaten ärztlicher Dokumente

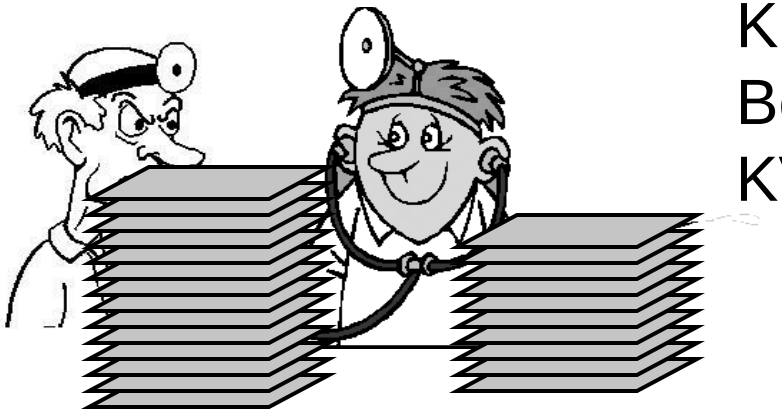
Andere Ärzte



Forschung/Lehre



Klinikverwaltung
Behörden, Kassen, KVen



Pflege

Apotheken

Justiz

Patienten

Arten ärztlicher Dokumente

…von Ärzten für „Verwaltung“

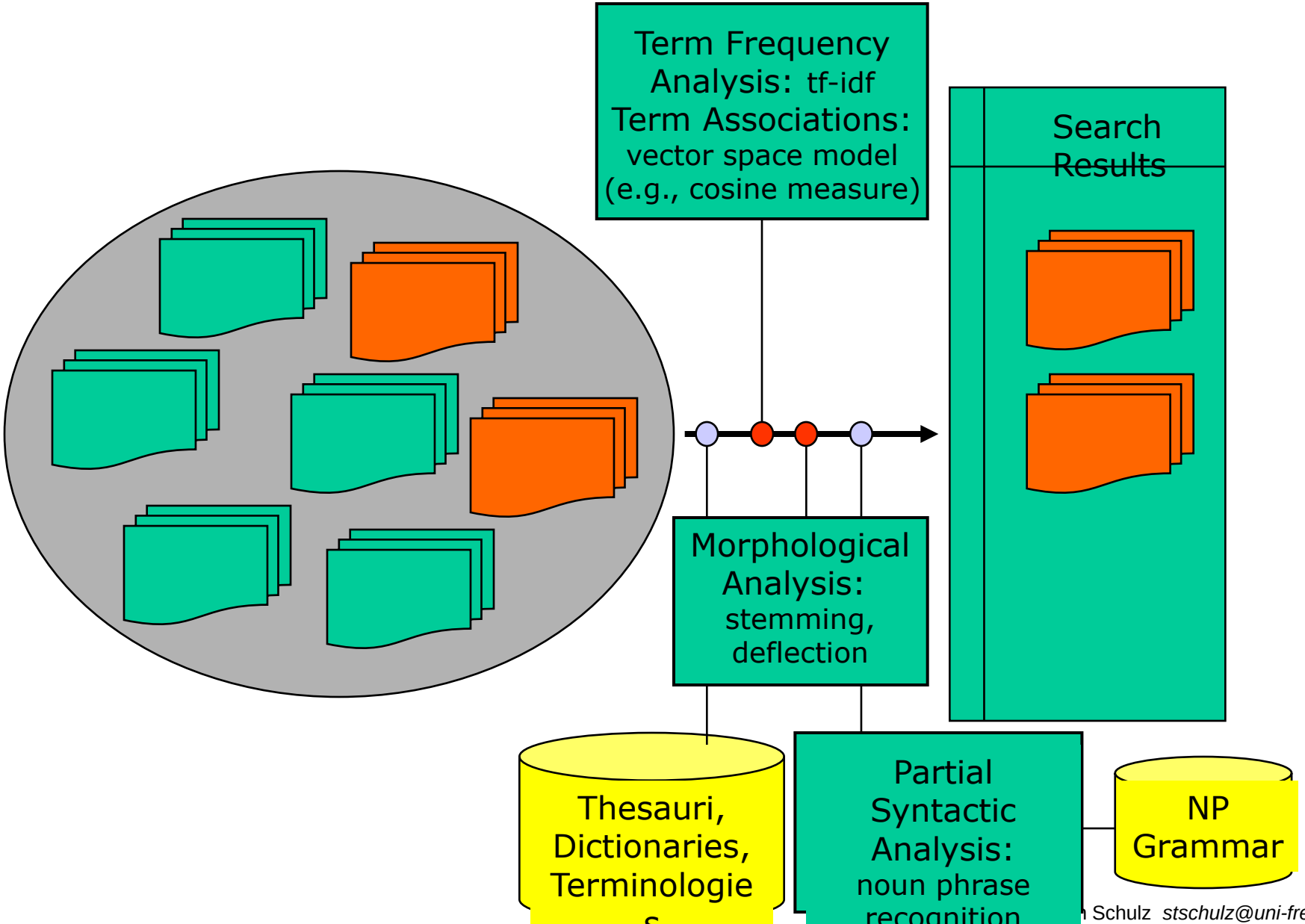
- Basis[illegible]ntation
- Spez[illegible]mentationen
  - z.B[illegible]

*…von Ärzten für Arzte**

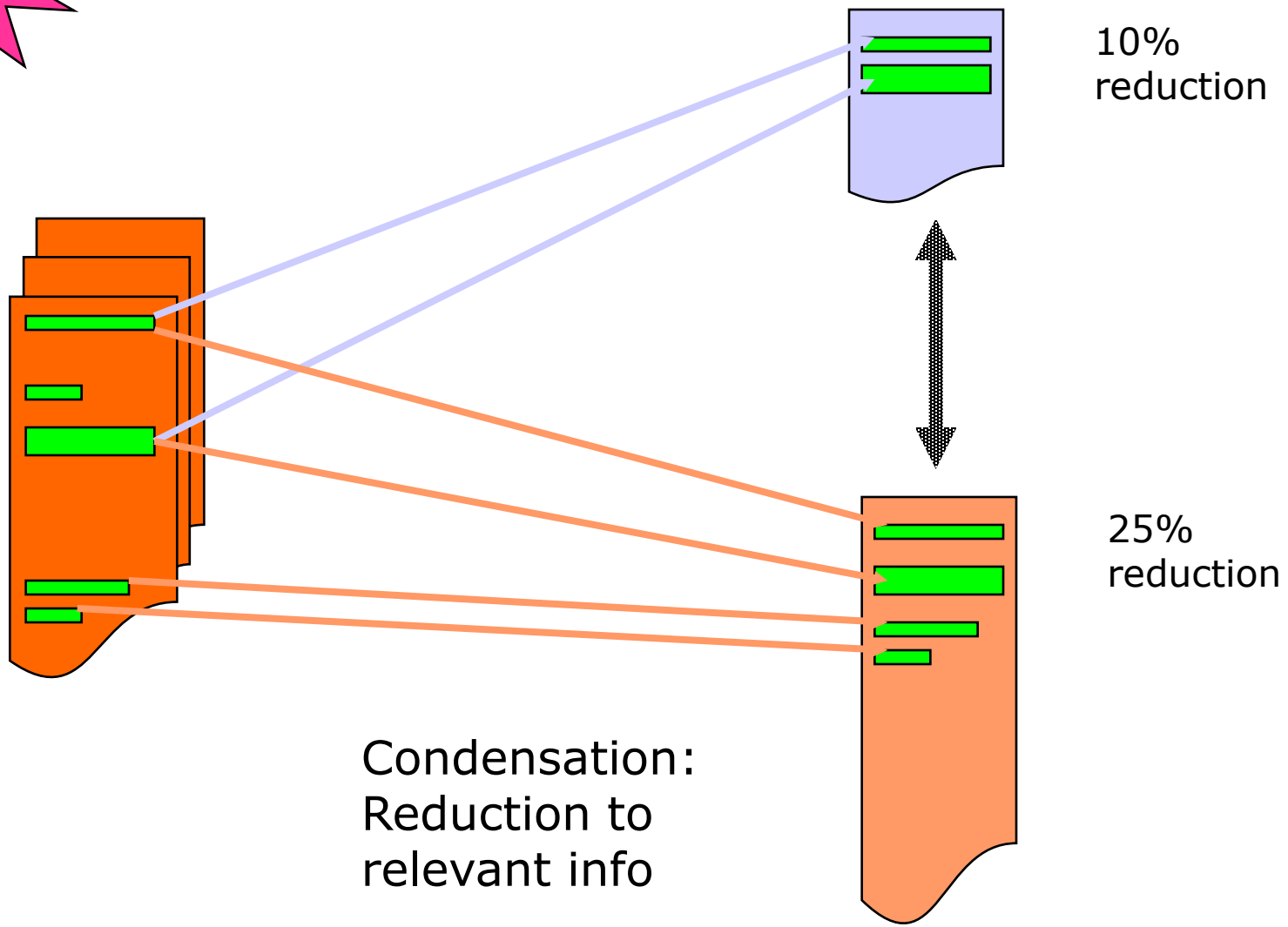
- Krankengeschichte
- Arztbrief
- Befunddokumentation

*und Vertreter anderer Medizinberufe

Document Retrieval

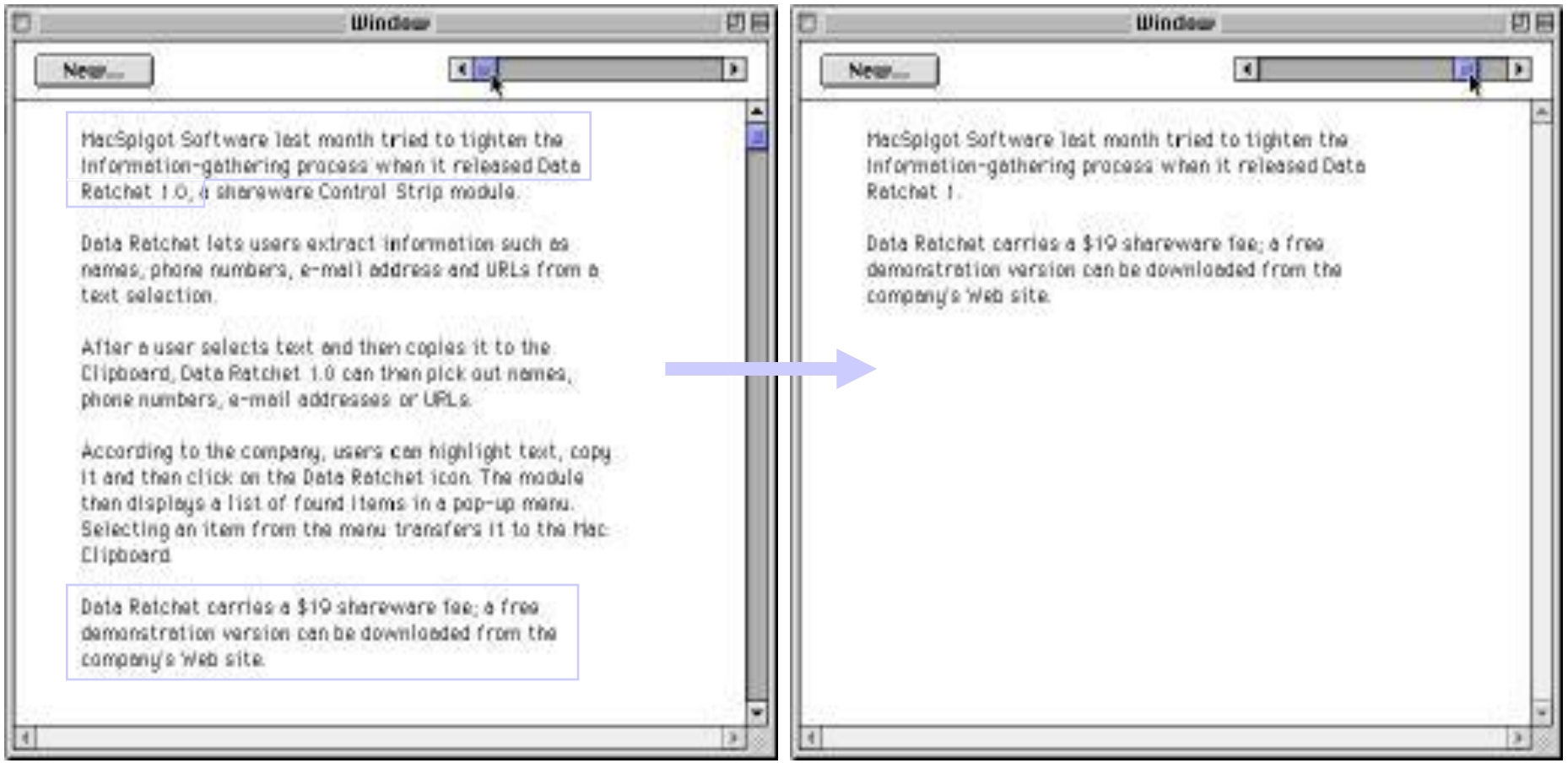


Text Summarization

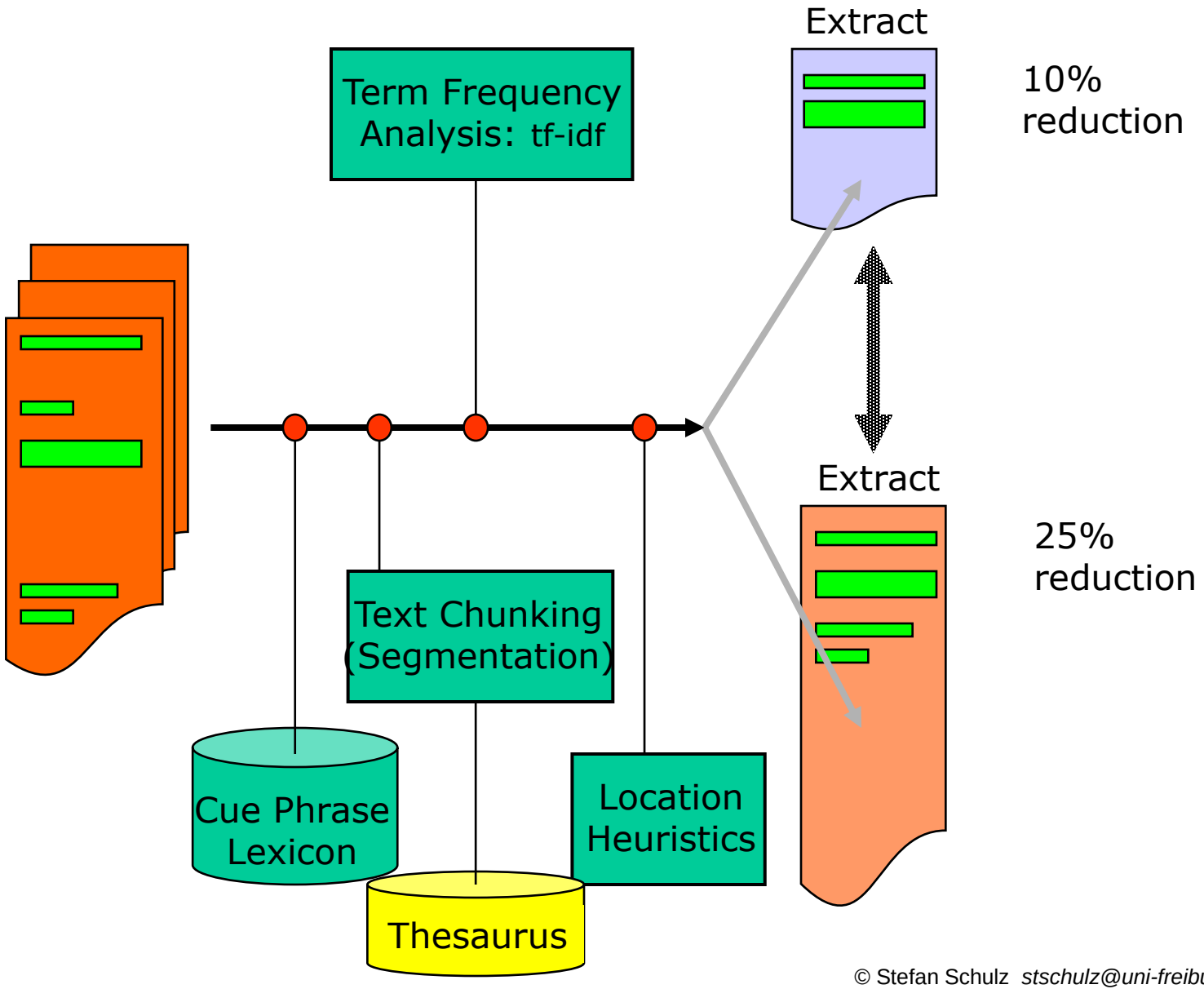


Condensation:
Reduction to
relevant info

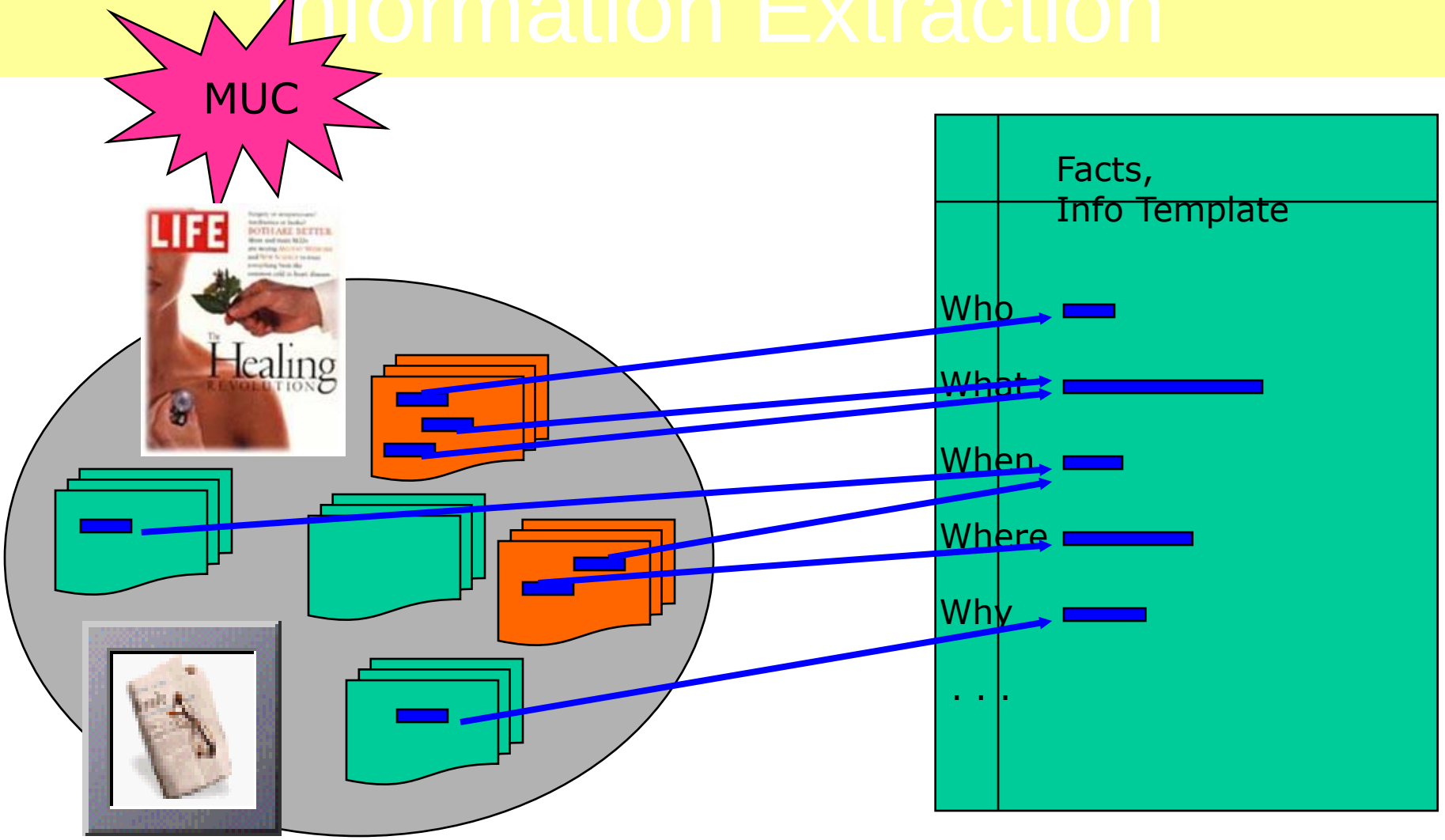
MTT Text Summarization System



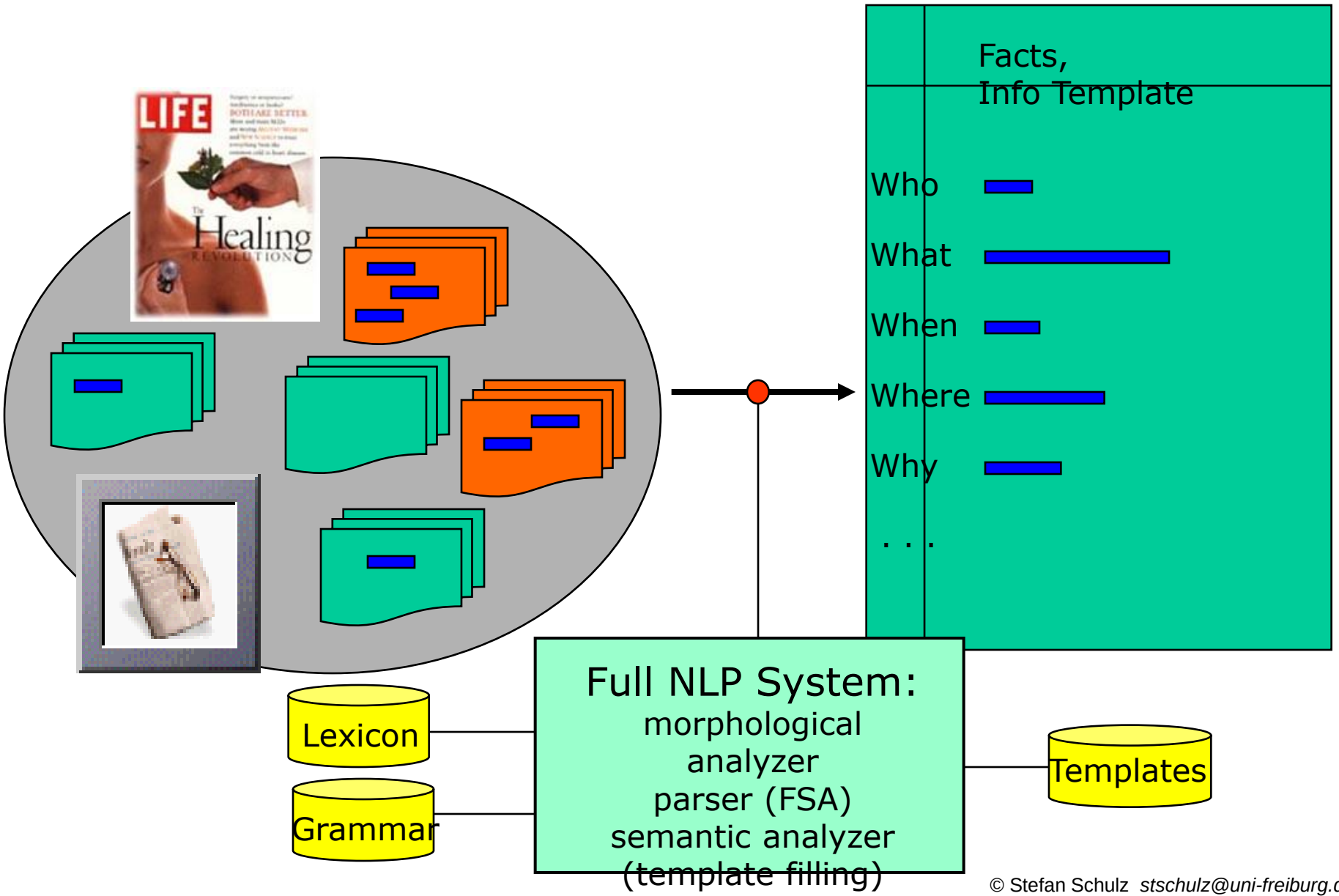
Text Summarization
(based on sentence extraction)



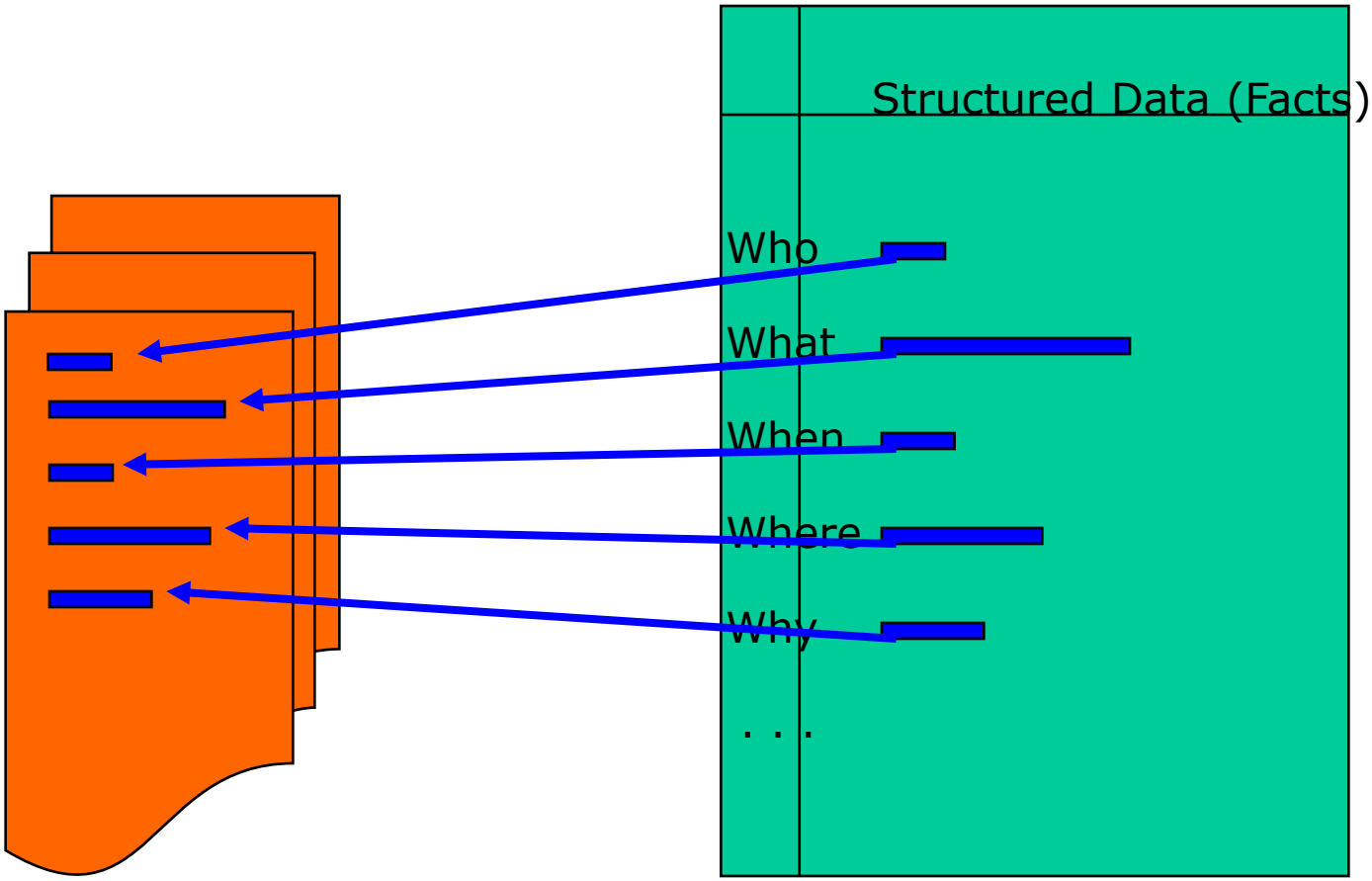
Information Extraction



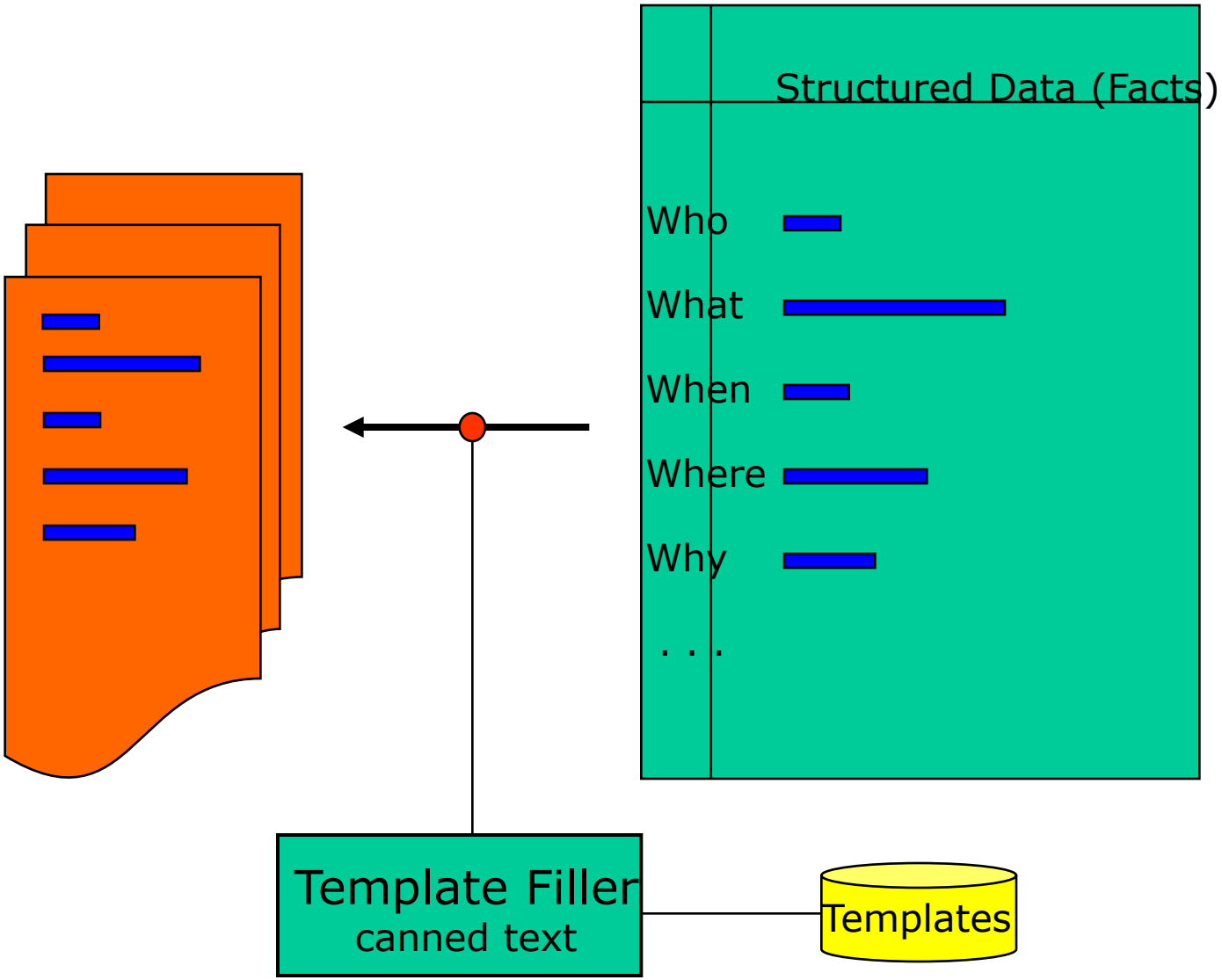
Information Extraction



Text Generation



(Shallow) Text Generation



Summaries are critical here, since succinctness is key and display space is limited.

Urgent cable summarized at 20% reduction (generic or query-related)

Notebook

Bach

Columbia Intro

Crisis Info

Crisis Sum

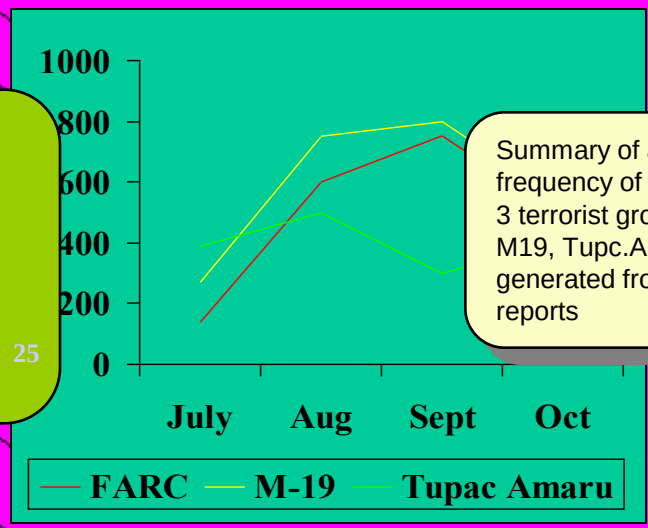
...NTRY-WIDE ATTACKS BY ANTI-GOVERNMENT FORCES..M-19, FARC...PRESIDENTIAL GUARD HAS SECURED AVIANCA AREA..

20

Foreign mercenaries believed to have joined the FARC

25

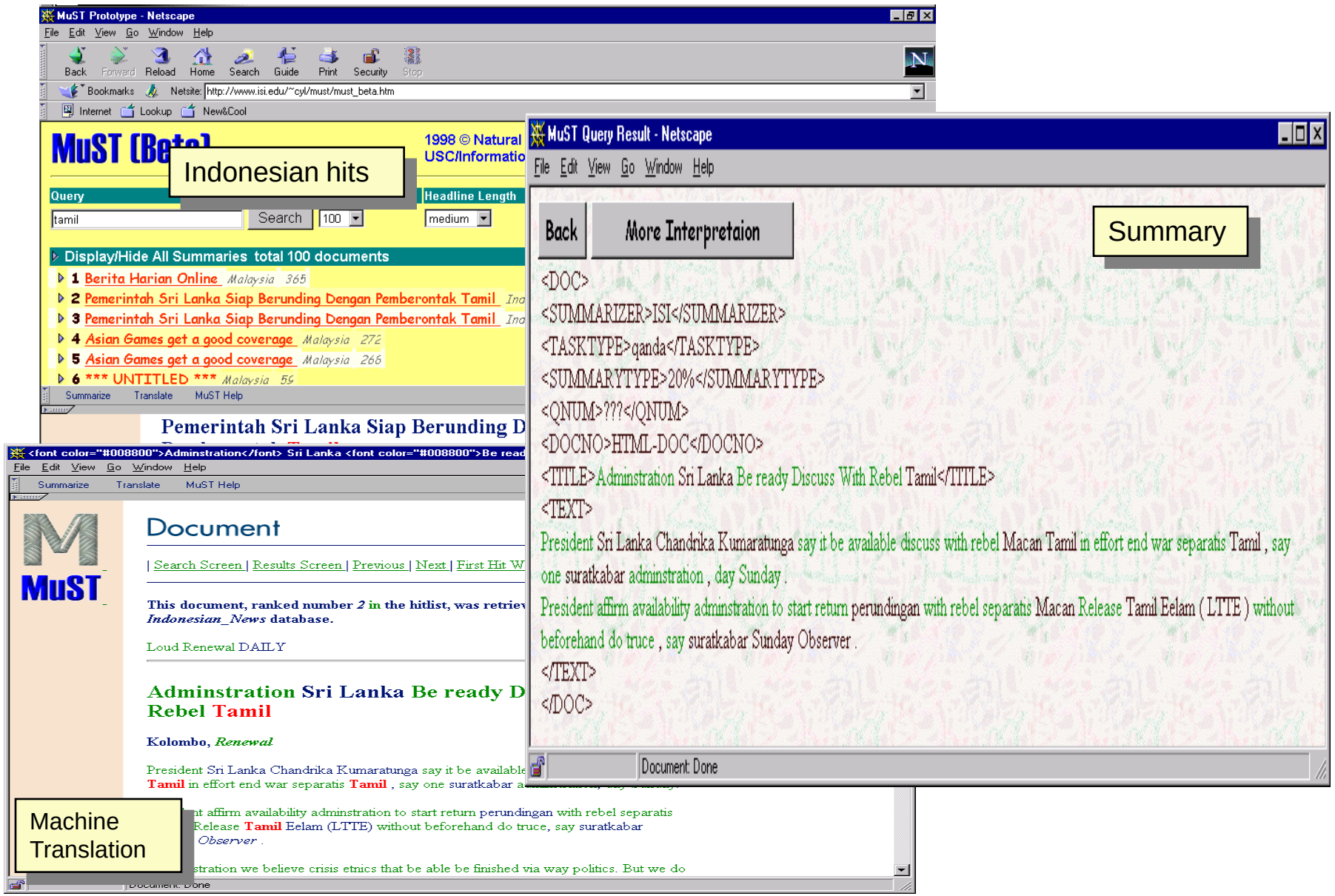
Places mentioned significantly in reports on FARC (black) and M-19 (white) are highlighted



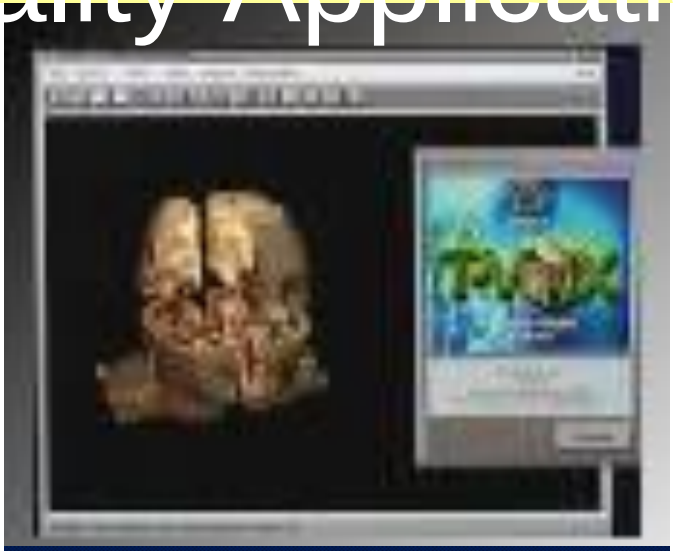
Summary of activity (sig. frequency of mention) of 3 terrorist groups (FARC, M19, Tupc.Am.) generated from news reports

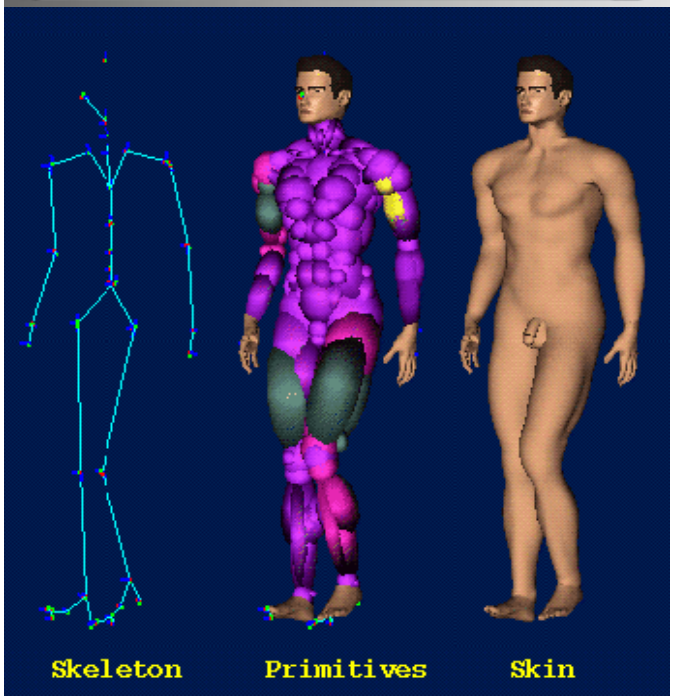
Filters composed together to provide powerful information reduction, visualization, and analysis

Cross- and Multilingual

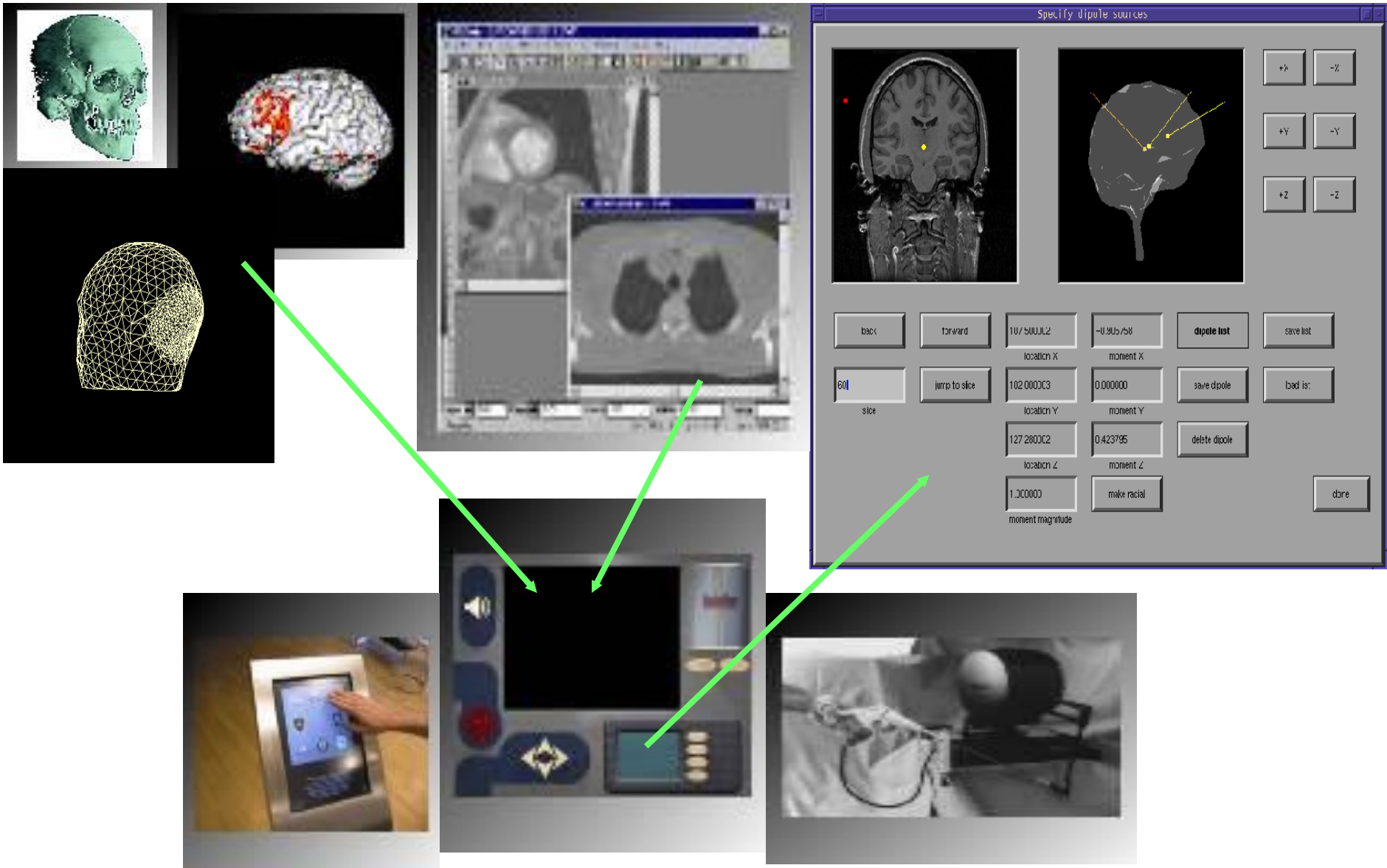


Combined Speech and Virtual Reality Applications





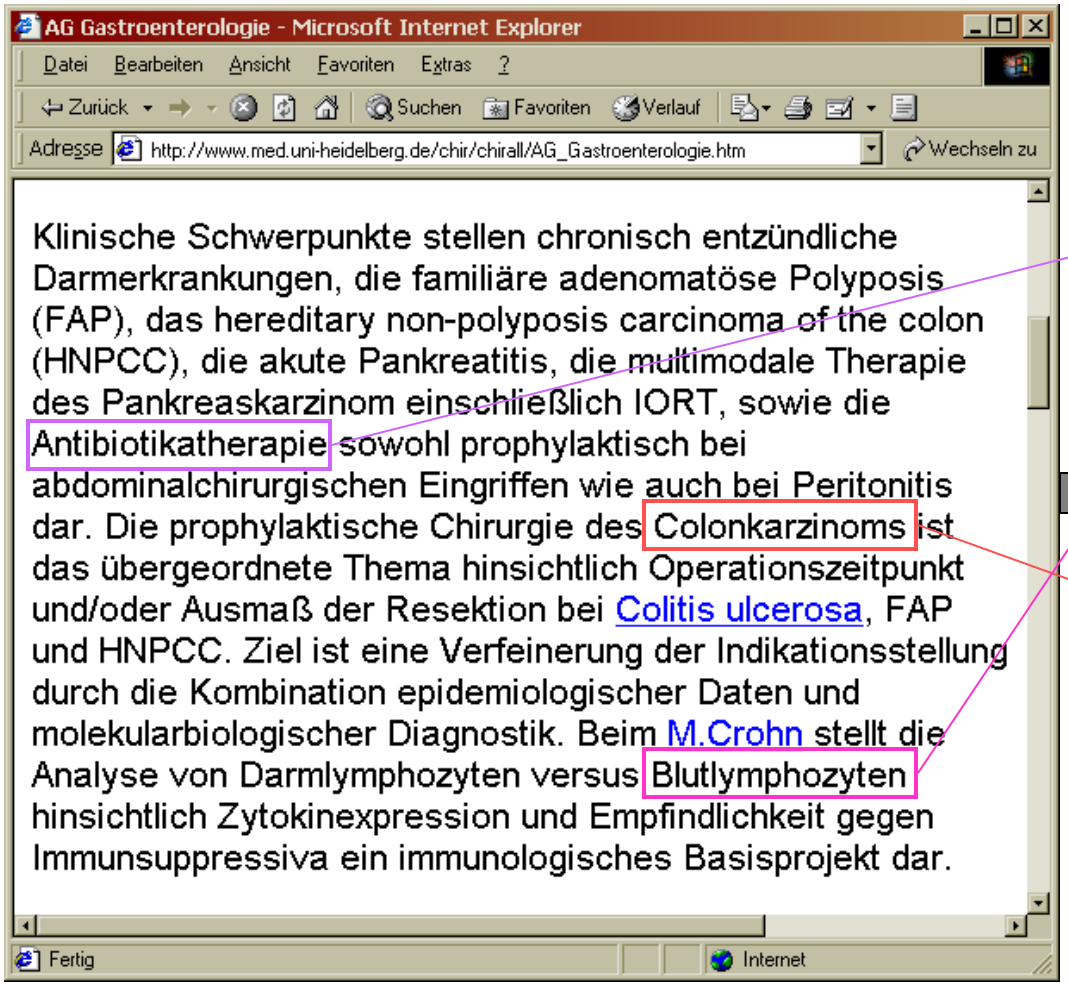
Multimedia / Multimodal Interfaces



Language Engineering Requirements

- Process large volumes of texts
- Deal robustly with ‘dirty’ real-world texts
- Meet fast throughput demands
- ‚Tricky’ solutions for really hard problems
  - ad hoc heuristics
  - domain- and application-specific solutions
- Evaluate how good you are

Token-based Indexing



Index

Medical Terminology: Poor retrieval performance

Frequency of synonymous German Word forms in *Google* Searches

Spelling Variants Synonyms Inflections		
Kolonkarzinom	2070	1780
Colonkarzinom	248	135
Coloncarcinom	111	73
Colon-Ca	203	169
Kolon-Ca	66	46
Dickdarmkrebs	4000	3610
Dickdarmkarzinom	288	175
Dickdarmcarcinom	13	10
Kolonkarzinoms	471	253
Kolonkarzinome	275	139
Kolonkarzinomen	265	166

Number of Hits

Number of exclusive hits (no other form matches)

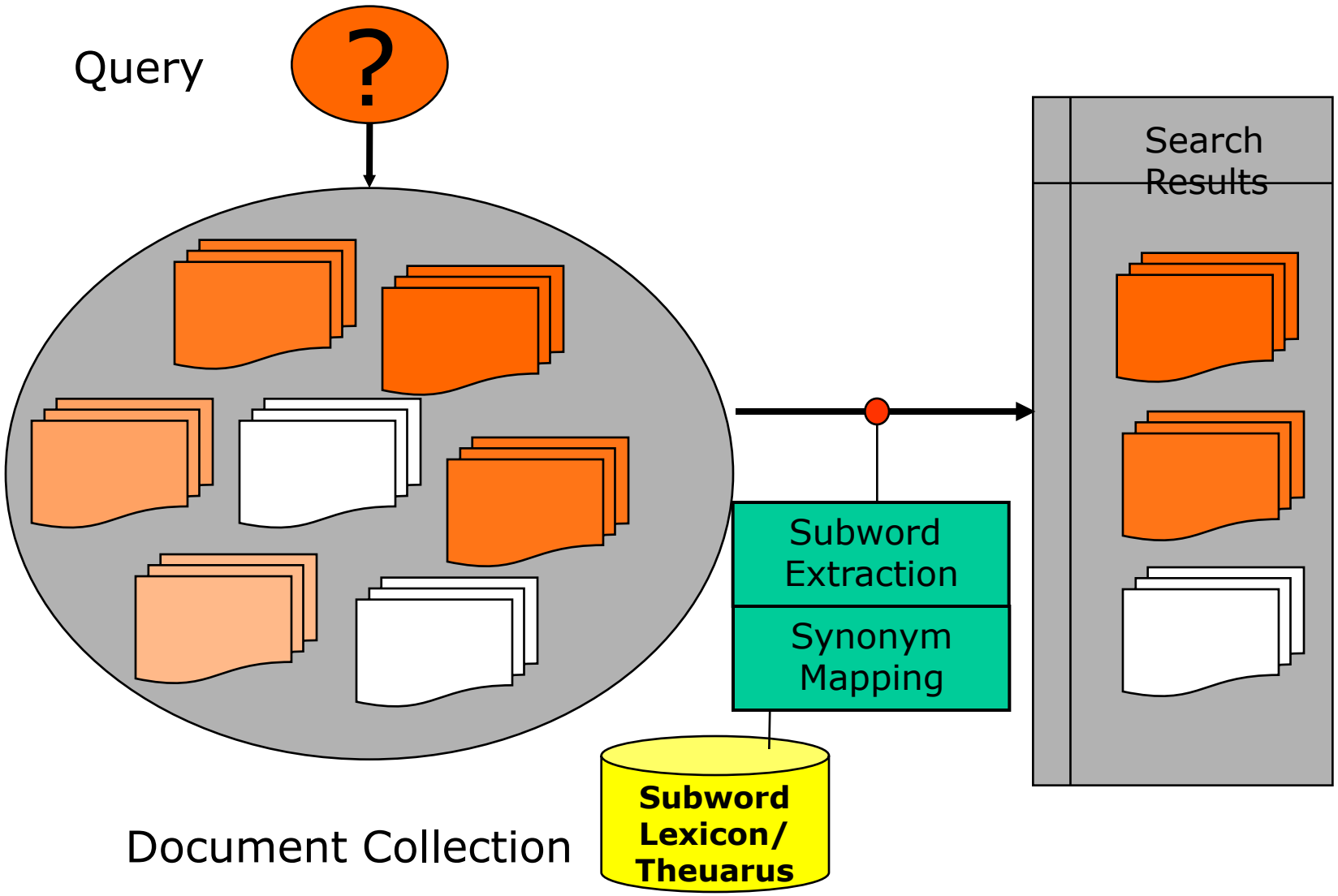
Improving Retrieval Performance Using Linguistic Techniques

- The MorphoSaurus approach:
  Subwords are atomic linguistic sense units
  - Morphemes: *nephr, anti, thyr, scler, hepat, cardi*
  - Morpheme aggregates: *diaphys, ascorb, anabol, diagnost*
  - Words: *amyloid, bone, fever, liver*
  - (noun groups: *vitamin c*,…)
- Grouping of synonymous subwords:

  kkyxkj = {*nephr, kidney, nier, ren*},

  qxkjkq = {*hepar, hepat, liver*},

Document Retrieval



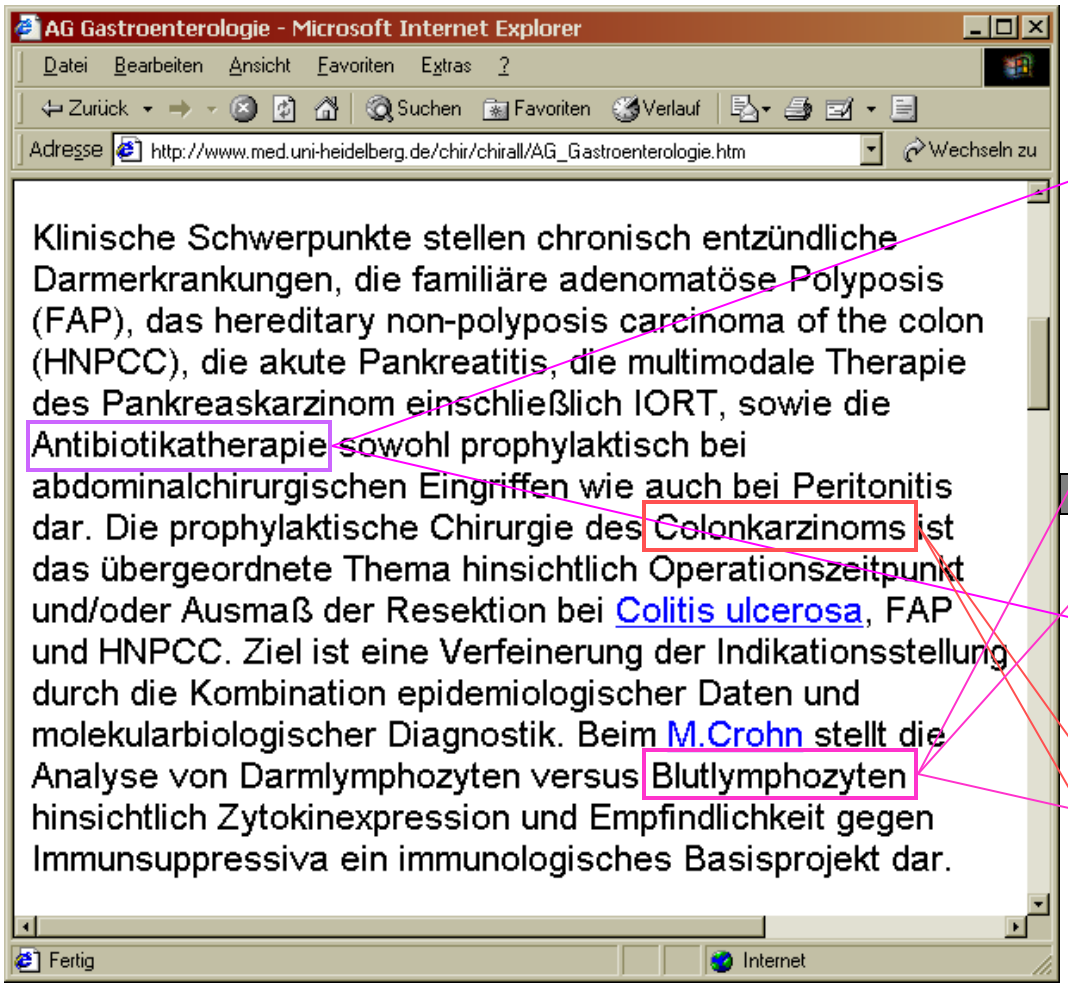
Examples of Subword Extraction

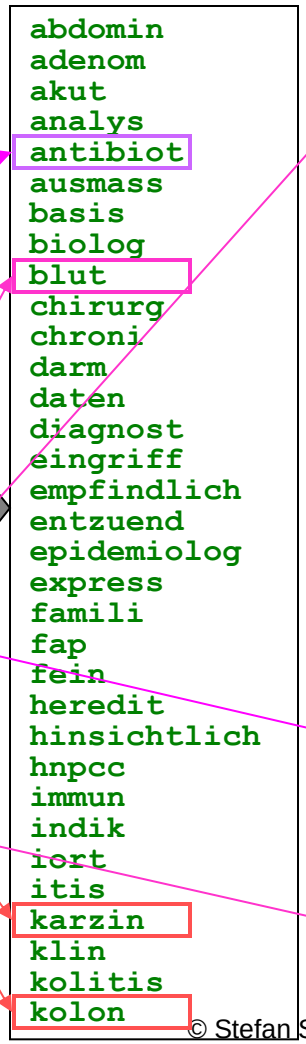
- Examples:
  - **proct** o **sigm** oid o **scop** y
  - **Schilddrüs** en **karzin** om
  - **cole cist ectom** ía
  - **acro cefal** o **sindattil** ia
  - **Sport verletz** ung en
  - **hør** sel s **hemm** ed e
  - **orchid** o **pex** ie
  - **Magen schleimhaut entzünd** ung

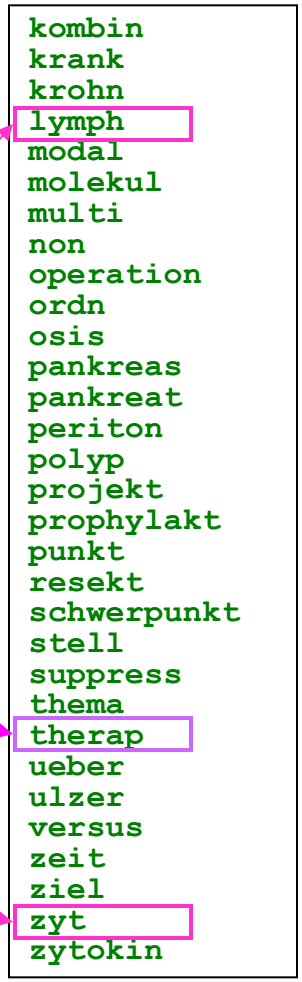
Lexical subwords (used for indexing)

Functional morphemes (not used for indexing)

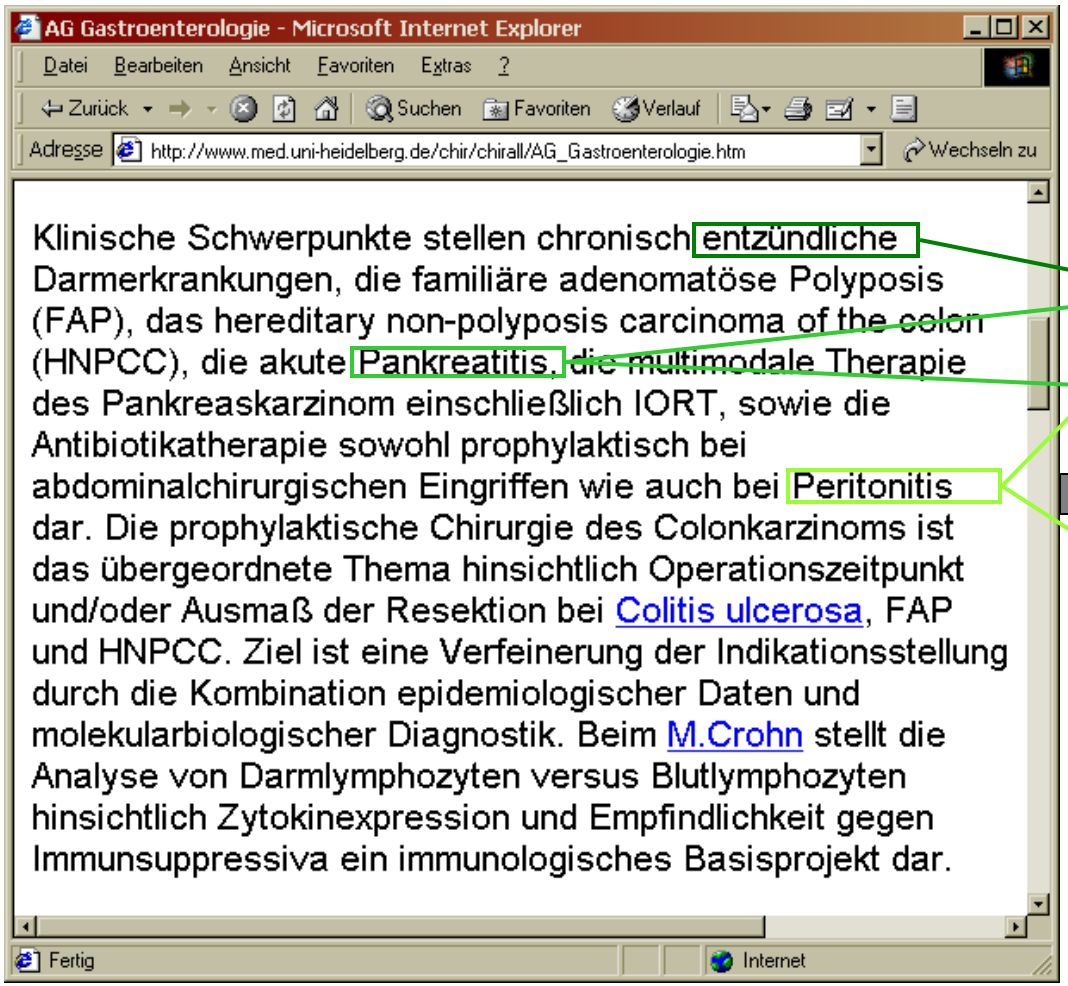
Subword Indexing

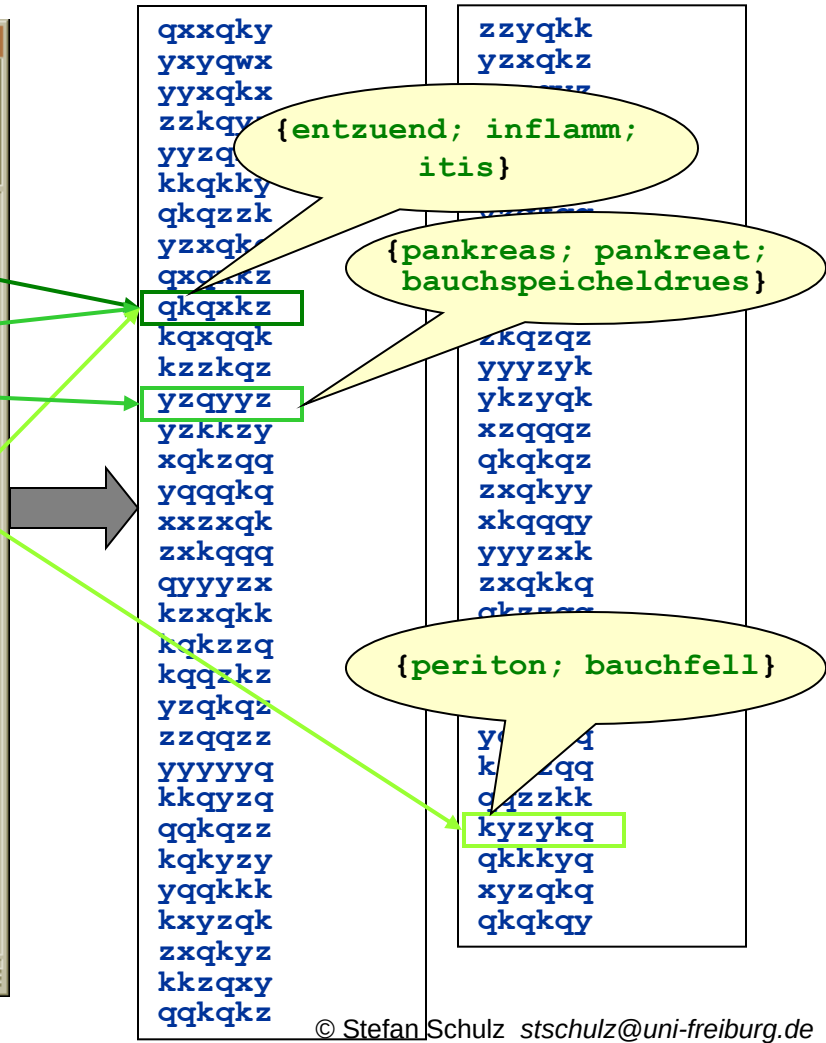




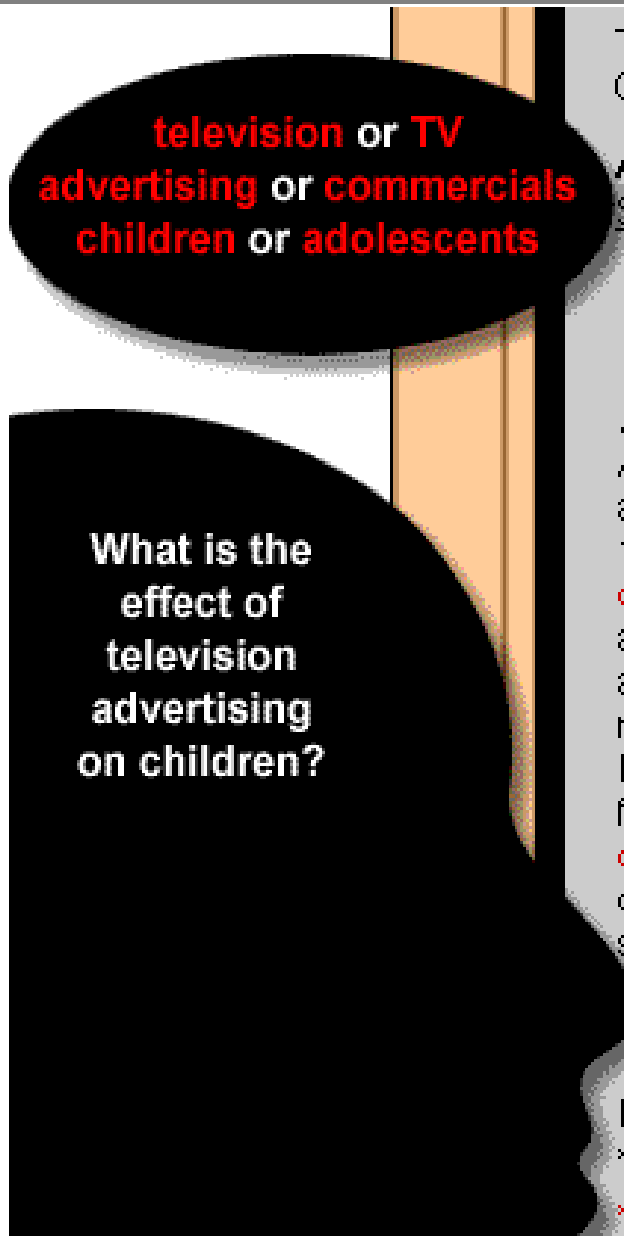


Subword - Indexing with Semantic Normalization





Document Retrieval



TITLE
On children's mass media communication.

AUTHOR
Sharma,-Yashini

SOURCE
Psycho-Lingua. 1995 Jan-Jul; Vol25 (1-2): 85-96

ABSTRACT
Analyzed and interpreted mass media communication that appeared in television commercial advertisements between 1991 and 1994 which were directed at children, of children, by children and only for children. The author employed content analysis for analyzing the behavioral contents of commercial advertisements as well as for children in the ads, problems of measurement, understandability and comprehensibility, language and language-play, disclaimers, etc. The study focuses mainly on disclaimers and their intelligibility in young children. Findings show that understanding of contents of commercial advertisements from the points of view of children's semantics and syntax structures determines their comprehensibility and linguistic competence. ((c)1998 APA/PsycINFO, all rights reserved)

MAJOR DESCRIPTORS
*Childhood-; *Content-Analysis; *Language-Development; *Television-Advertising

Different Perspectives

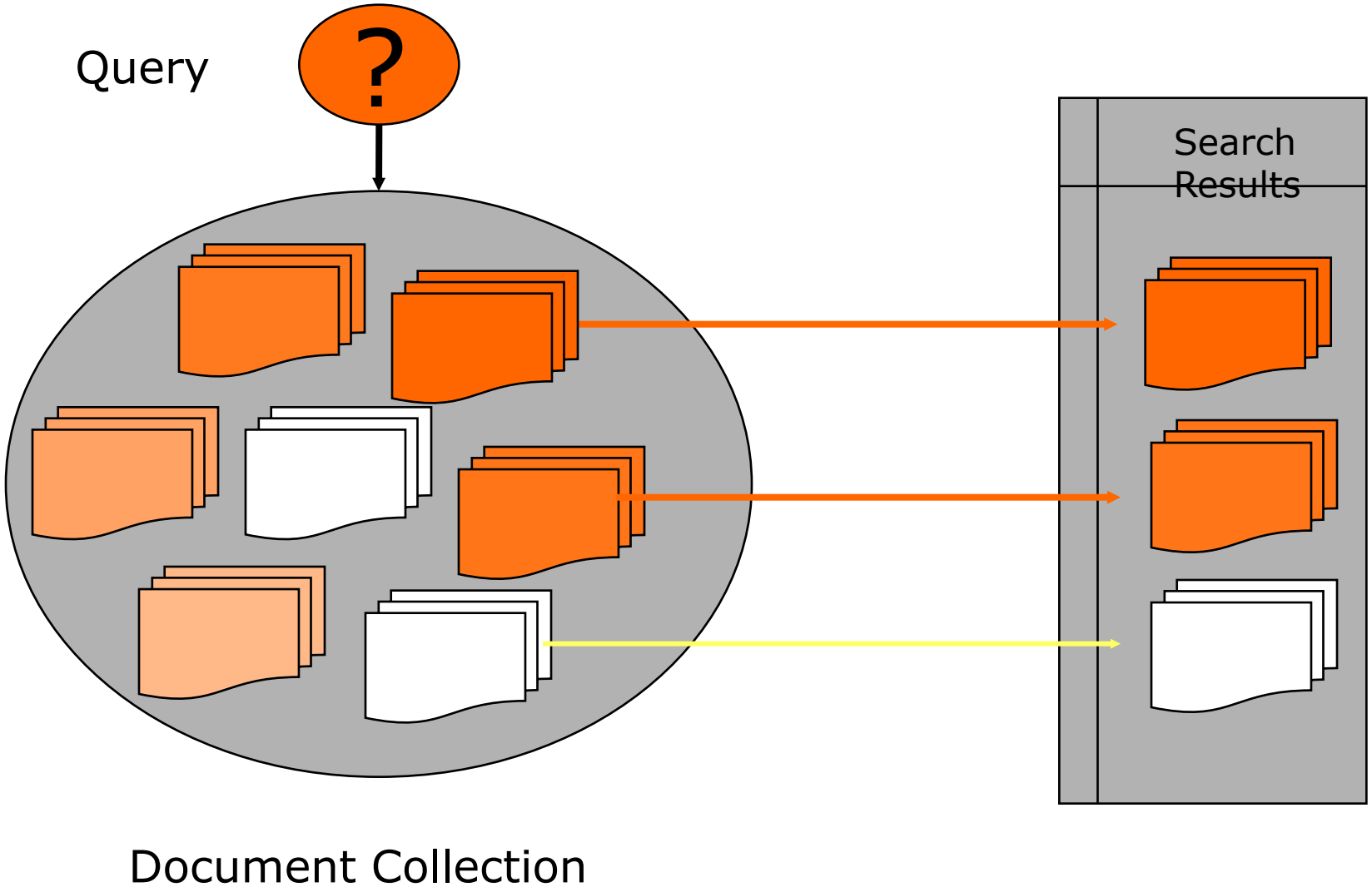
- Automatic indexing
  - extracting relevant terms (topic descriptors) from a document
    - ad hoc retrieval query
- Automatic classification
  - grouping a subset of documents with a homogeneous topic (as characterized by their descriptors)
    - ad hoc retrieval query
- Automatic routing, filtering,
  - delivery of documents matching a given interest profile (as characterized by topic descriptors)
    - frozen retrieval query

Document Retrieval: Basic Approach

- A Document Collection
  $D = \{d_1, d_2, ..., d_n\}$
- A query q
- Two Methods:
  - „Filter“ Split D into two sets D_{relq} and D_{nrelq}
    (D_{relq} = Set of relevant documents for q)
    (D_{nrelq} = Set nonrelevant documents for q)
  -

-

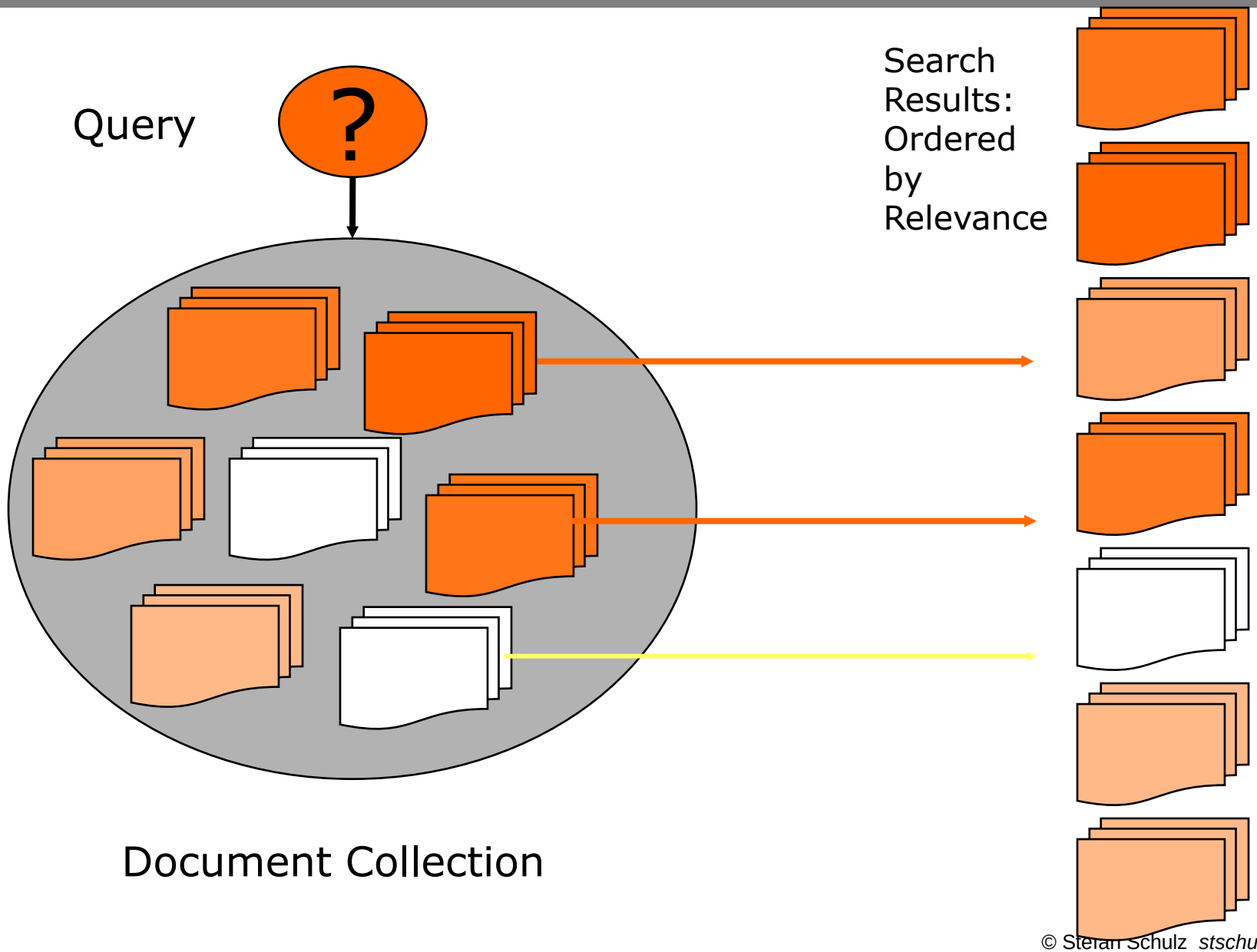
Document Retrieval



Document Retrieval: Basic Approach

- A Document Collection
  $D = \{d_1, d_2, ..., d_n\}$
- A query q
- Two Methods:
  - „Filter“ Split D into two sets D_{relq} and D_{nrelq}
    (D_{relq} = Set of relevant documents for q)
    (D_{nrelq} = Set nonrelevant documents for q)
  - „Order“ = Order by relevance:
    $D = [d'_1, d'_2, ..., d'_n]$
    with $rel(d'_i) \geq rel(d'_{i+1})$
- Combinations are possible

Document Retrieval



Evaluation of Text Retrieval Systems

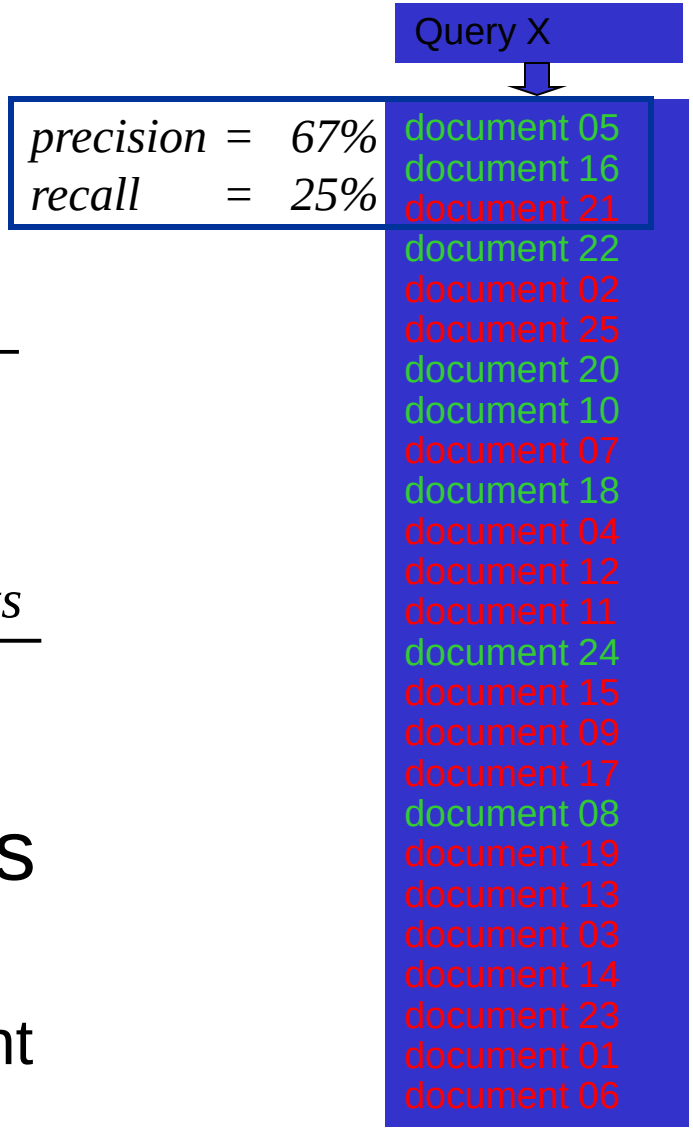
- Target variables:

$$precision = \frac{n_{found+relevantDocuments}}{n_{found\_documents}}$$

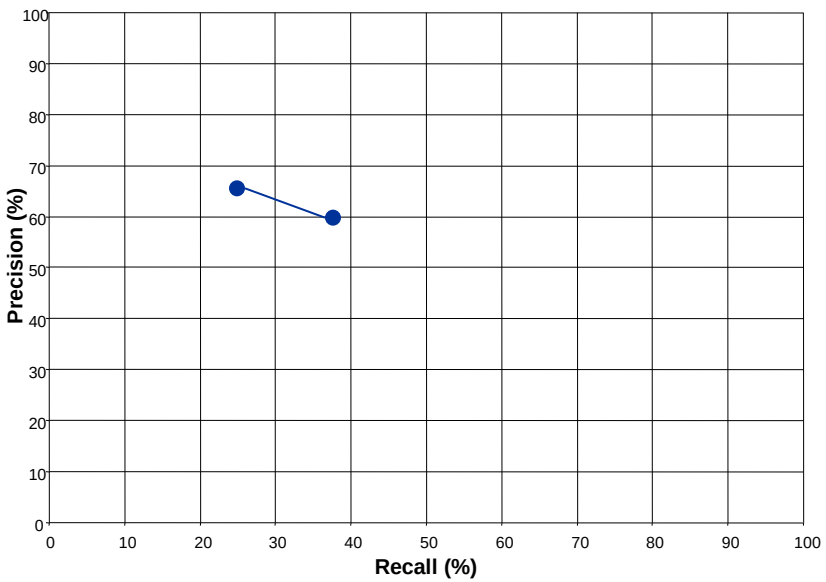
$$recall = \frac{n_{found+relevant\_documents}}{n_{relevant\_documents}}$$

- Precision/Recall-Diagrams with ranked output

  Example: 25 documents, 8 relevant



Evaluation of Text Retrieval Systems



Query X

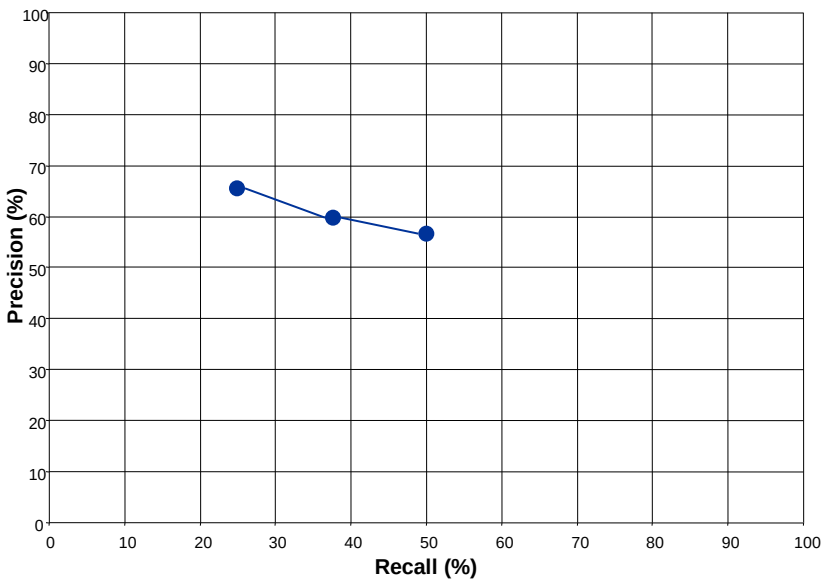
precision = *60%*
recall = *38%*

document 05
document 16
document 21
document 22
document 02
document 25
document 20
document 10
document 07
document 18
document 04
document 12
document 11
document 24
document 15
document 09
document 17
document 08
document 19
document 13
document 03
document 14
document 23
document 01
document 06

- Precision/Recall-Diagrams with ranked output

  Example: 25 documents, 8 relevant

Evaluation of Text Retrieval Systems



Query X

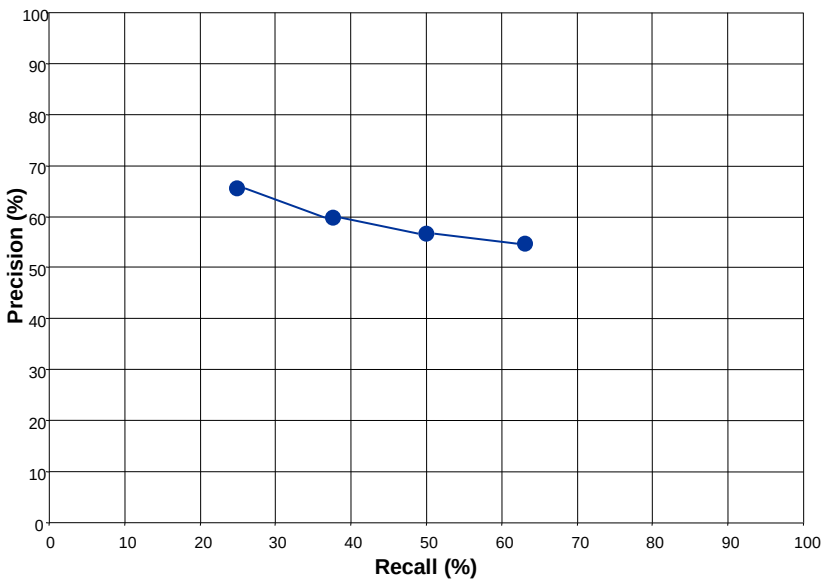
precision = *57%*
recall = *50%*

document 05
document 16
document 21
document 22
document 02
document 25
document 20
document 10
document 07
document 18
document 04
document 12
document 11
document 24
document 15
document 09
document 17
document 08
document 19
document 13
document 03
document 14
document 23
document 01
document 06

- Precision/Recall-Diagrams with ranked output

  Example: 25 documents, 8 relevant

Evaluation of Text Retrieval Systems



Query X

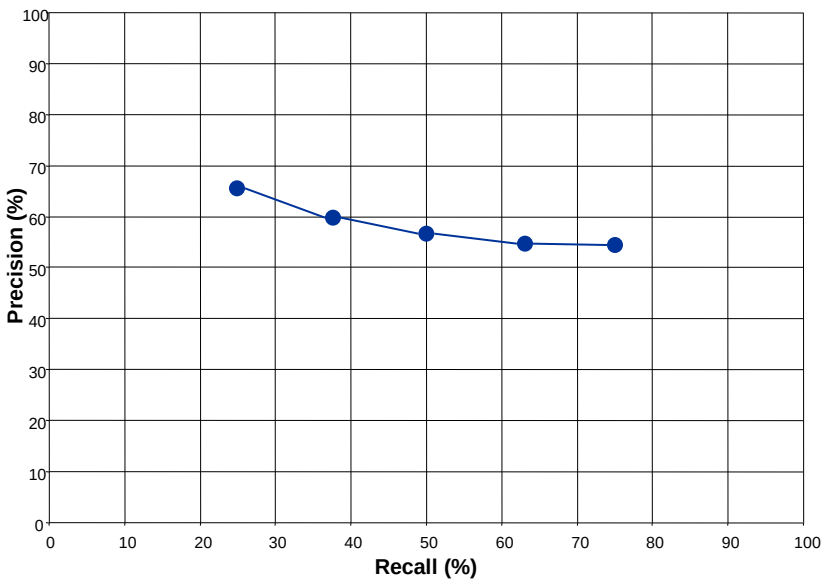
precision = 55%
recall = 63%

document 05
document 16
document 21
document 22
document 02
document 25
document 20
document 10
document 07
document 18
document 04
document 12
document 11
document 24
document 15
document 09
document 17
document 08
document 19
document 13
document 03
document 14
document 23
document 01
document 06

- Precision/Recall-Diagrams with ranked output

  Example: 25 documents, 8 relevant

Evaluation of Text Retrieval Systems



Query X

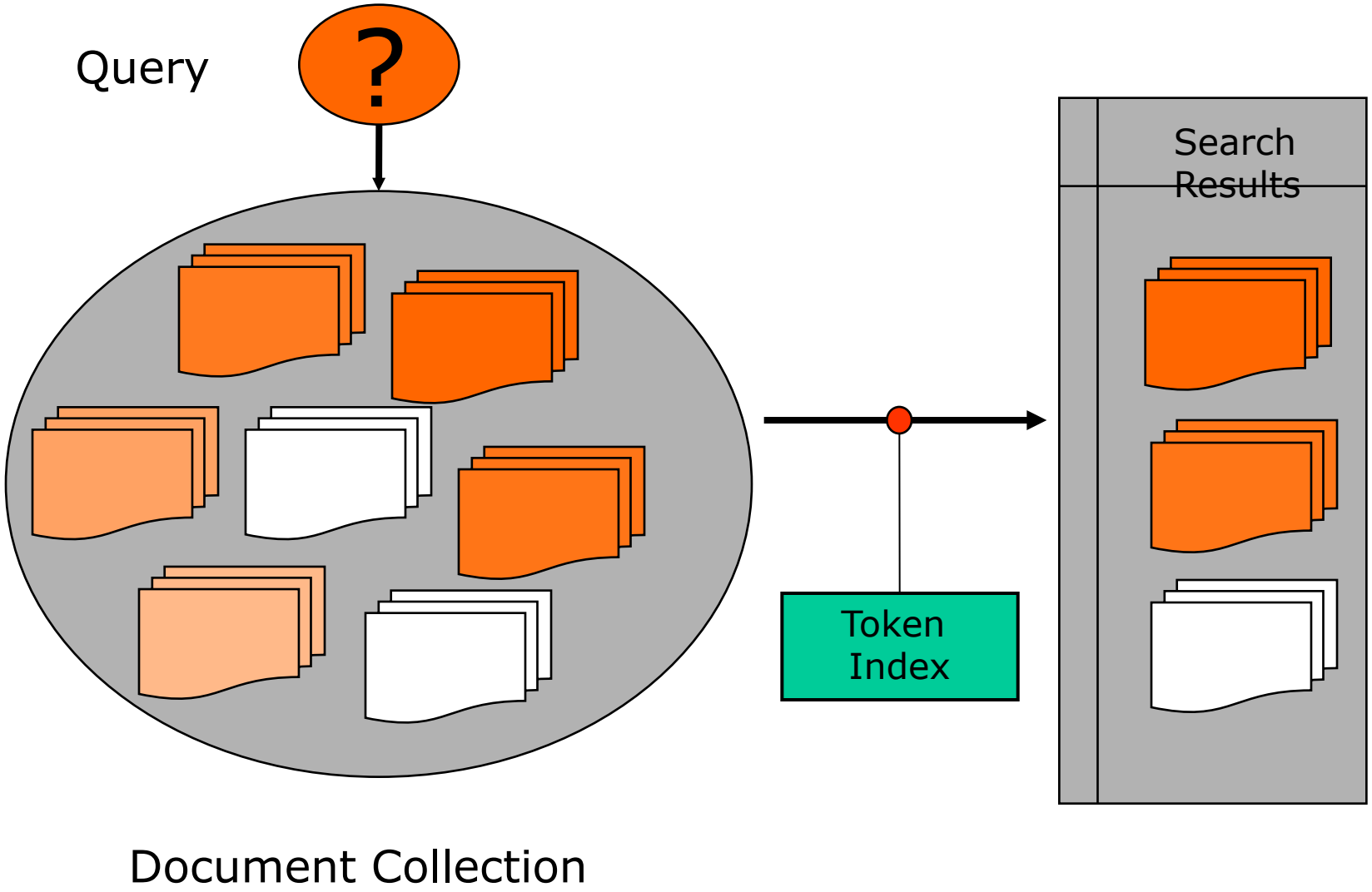
precision = *54%*
recall = *75%*

document 05
document 16
document 21
document 22
document 02
document 25
document 20
document 10
document 07
document 18
document 04
document 12
document 11
document 24
document 15
document 09
document 17
document 08
document 19
document 13
document 03
document 14
document 23
document 01
document 06

- Precision/Recall-Diagrams with ranked output

  Example: 25 documents, 8 relevant

Document Retrieval

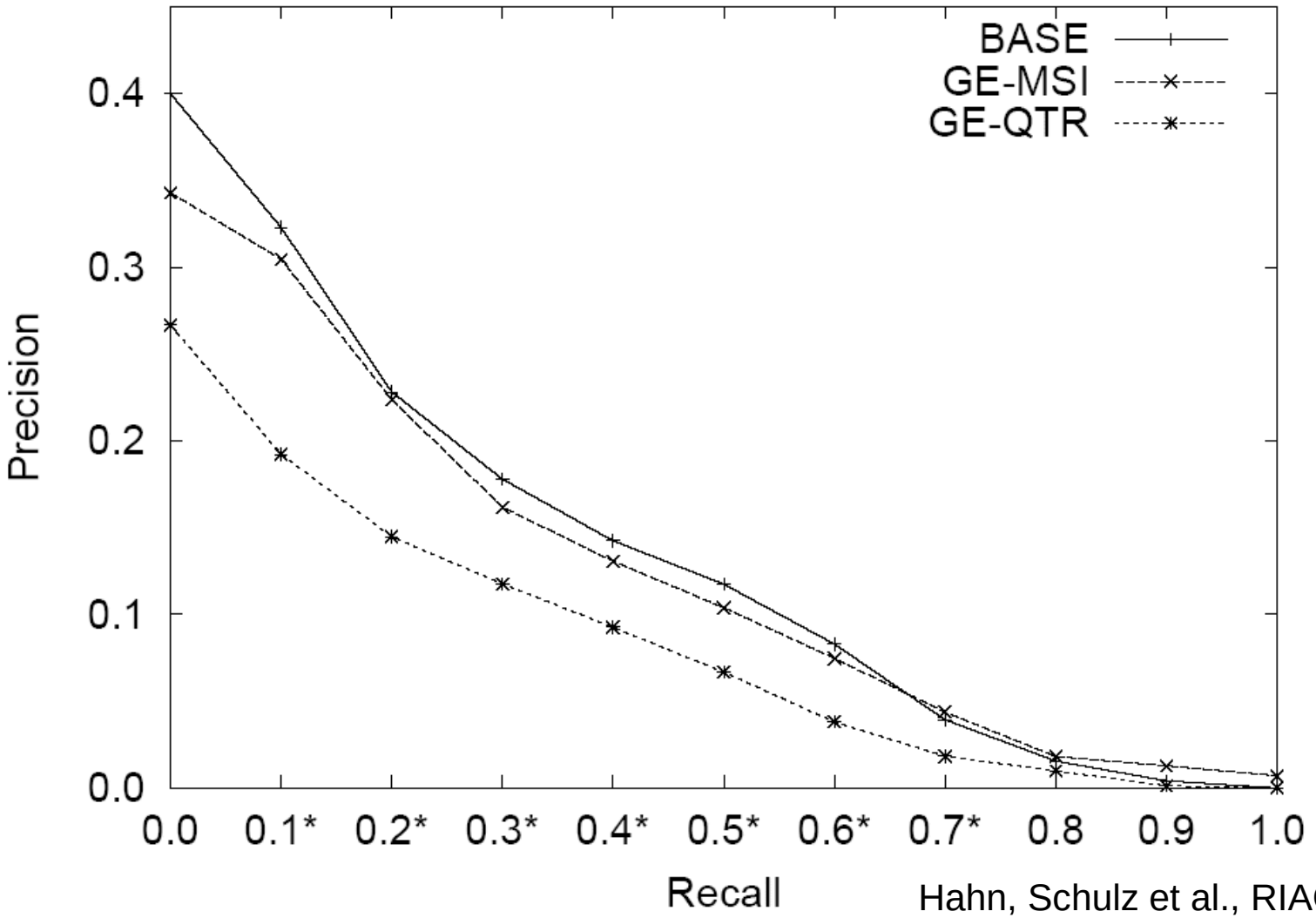


Perspectives in Medical Document Retrieval

- Automatic indexing: extracting relevant terms (topic descriptors from an indexing Alphabet, e.g. MeSH) from a document
- Automatic classification: grouping a subset of documents with a homogeneous topic (as characterized by their descriptors)
- Retrieval in Medical Vocabularies (disease, procedure encoding)
- Cross-Language Document Retrieval

Original Document	**Orthographic Normalization**	**Morphological Segmentation**	**Semantic Normalization**
High TSH values suggest the diagnosis of primary hypothyroidism while a suppressed TSH level suggests hyperthyroidism.	high tsh values suggest the diagnosis of primary hypothyroidism while a suppressed tsh level suggests hyperthyroidism.	high tsh value s suggest the diagnos is of primar y hypo thyroid ism while a suppress ed tsh level suggest s hyper thyroid ism.	**#up# tsh #value#** #suggest# **#diagnost# #primar# #small# #thyre# #suppress# tsh #nivell#** #suggest# **#up# #thyre#** .
Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypothyreose, ein supprimierter TSH-Spiegel spricht dagegen für eine Schilddrüsenüberfunktion.	erhoehte tsh-werte erlauben die diagnose einer primaeren hypothyreose, ein supprimierter tsh-spiegel spricht dagegen fuer eine schilddruesenueberfunktion.	er hoeh te tsh - wert e erlaub en die diagnos e einer primaer en hypo thyre ose, ein supprim iert er tsh - spiegel spricht dagegen fuer eine schilddrues en ueber funktion.	**#up# tsh - #value#** #permit# **#diagnost# #primar# #small# #thyre#** , **#suppress# tsh -** {#mirrorr# **#nivell#**} #speak# **#thyre# #up#** #function# .
A presença de valores elevados de TSH sugere o diagnóstico de hipotireoidismo primário, enquanto níveis suprimidos de TSH sugerem hipertireoidismo.	a presenca de valores elevados de tsh sugere o diagnostico de hipotireoidismo primario, enquanto niveis suprimidos de tsh sugerem hipertireoidismo.	a presenc a de valor es elevad os de tsh suger e o diagnost ico de hipo tireoid ismo primari o, enquanto niveis suprimid os de tsh suger em hiper tireoid ismo.	#actual# **#value# #up# tsh** #suggest# **#diagnost# #small# #thyre# #primar#** , **#nivell# #suppress# tsh** #suggest# **#up# #thyre#** .

MorphoSaurus: Cross-Language Medical Document Retrieval



Hahn, Schulz et al., RIAO 2004

- Sub SetEnglish()
- '
- ' Makro aufgezeichnet am 11.04.2004 von coling.
- '
- Dim i As Integer
- For i = 1 To ActiveWindow.Presentation.Slides.Count
- ActiveWindow.Presentation.Slides(1).Select
- ActiveWindow.Presentation.Slides(1).Shapes.SelectAll

Relevant Parameters (Products *vs*. Lab Prototypes)

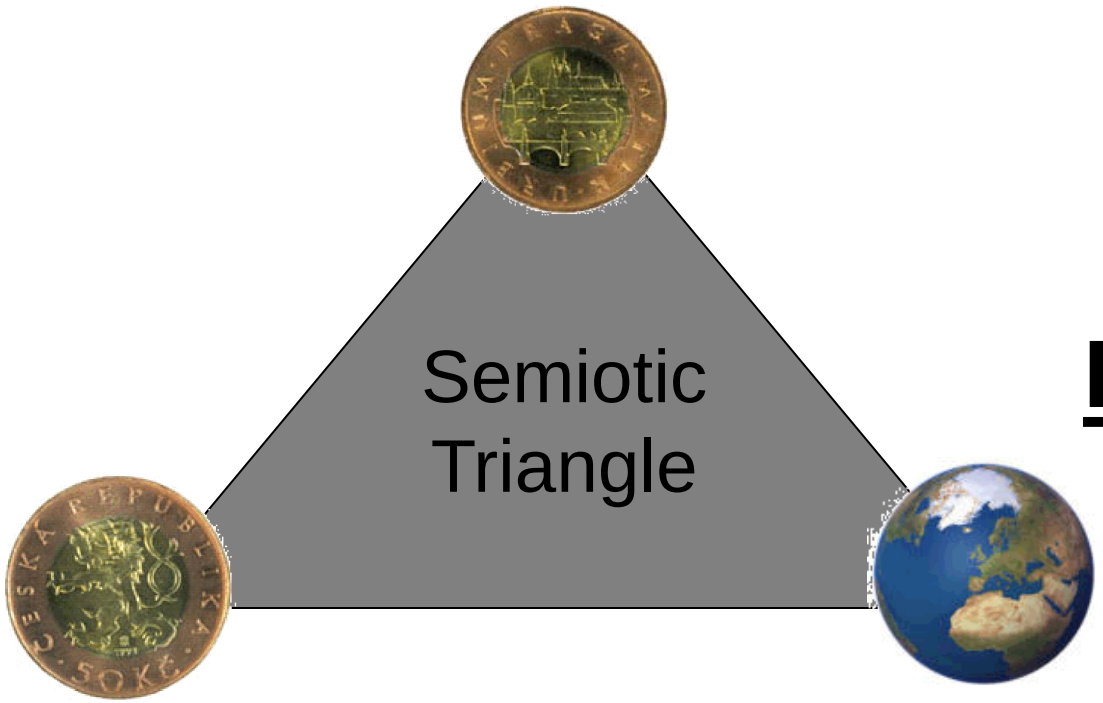
- document processing technologies provide shallow-processing approximations to 'hard' language understanding problems:
  - *summarization*: sentence extraction, text chunking *vs*. 'conceptual' abstracting
  - *info extraction*: application-specific templates *vs*. generic text understanding
  - *text generation*: instantiation of canned text templates *vs*. unrestricted text generation
  - *translation*: machine-aided translation (tool support), 'raw' translation skeletons (relevance assessment) *vs*. fully automatic, high quality translation
- coverage of grammars and domain ontologies
  - products: 100,000 (low profile) *vs*. lab systems: 1,000 - 5,000 (high profile)

Medical Content Management (I)

- Find me relevant documents on this topic!
- Find me relevant facts about this issue!
- Find me the right classification code !
- Find me scientific papers which help treat this patient
- The data is in the system, but I need to fill out a form
- Can a have a brief summary of all these documents?
- I need to search foreign-language documents
- I want to match genomic with patient information
- I want to search my health record

Odgen & Richards triangle…

Reference:
Concept,
Sense,

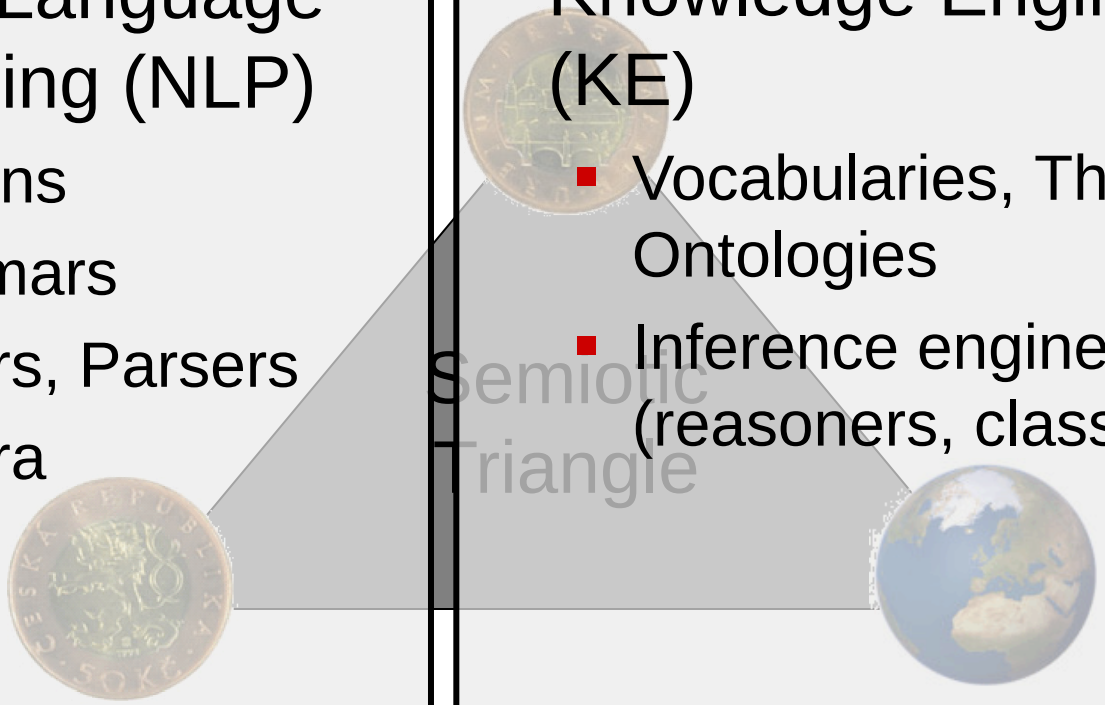


Sign:
Symbol
Language
Term

Referent:
Reality/
Object

Content Technologies

- Language
- Syntax
- Natural Language Processing (NLP)
  - Lexicons
  - Grammars
  - Taggers, Parsers
  - Corpora



- Meaning
- Semantics
- Knowledge Engineering (KE)
  - Vocabularies, Thesauri, Ontologies
  - Inference engines (reasoners, classifiers)

Challenges in the Medical Domain

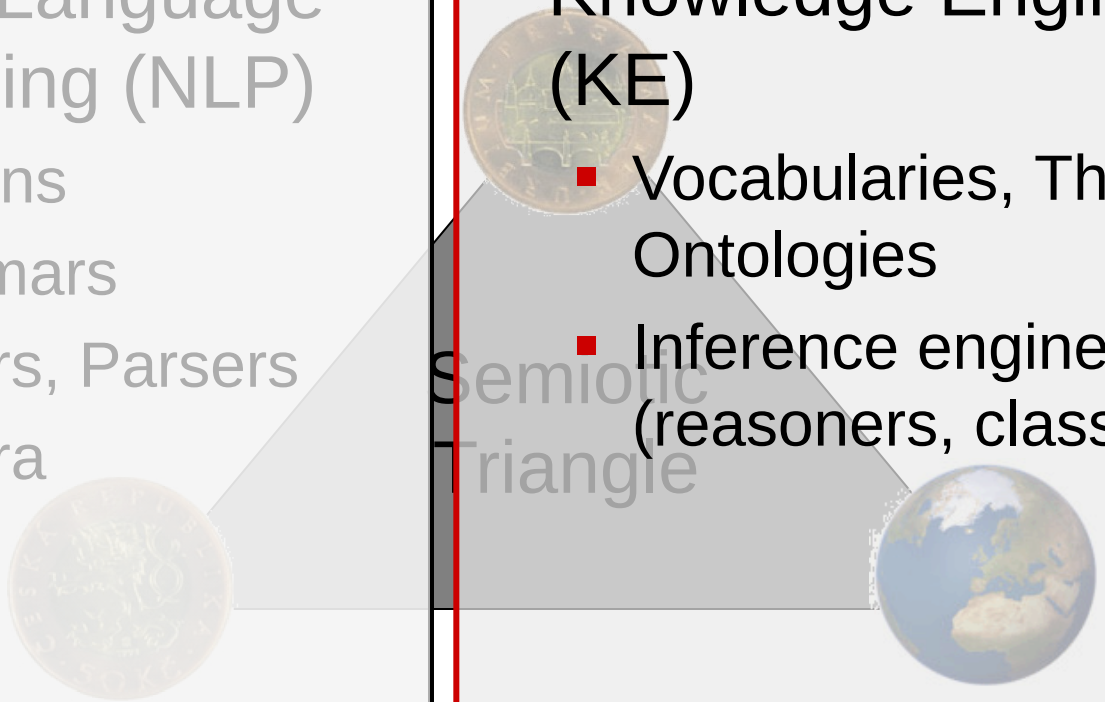
- Language
- Syntax
- Natural Language Processing (NLP)
  - Lexicons
  - Grammars
  - Taggers, Parsers
  - Corpora

Semiotic Triangle

- Meaning
- Semantics
- Knowledge Engineering (KE)
  - Vocabularies, Thesauri, Ontologies
  - Inference engines (reasoners, classifiers)

Challenges in the Medical Domain

- Language
- Syntax
- Natural Language Processing (NLP)
  - Lexicons
  - Grammars
  - Taggers, Parsers
  - Corpora



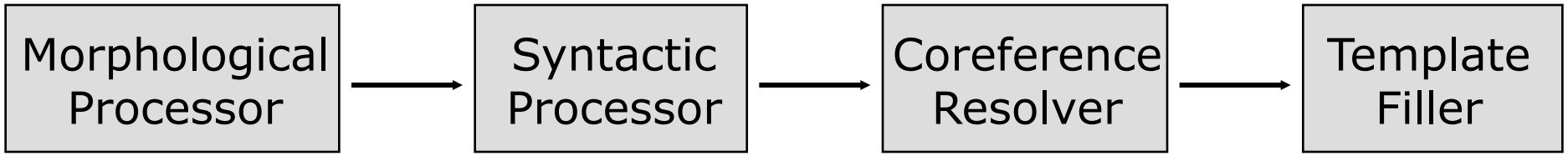
- Meaning
- Semantics
- Knowledge Engineering (KE)
  - Vocabularies, Thesauri, Ontologies
  - Inference engines (reasoners, classifiers)

Let us Try to Avoid some Common Misunderstandings ...

- Automatic natural language processing is easy
- Natural language processing systems with a high degree of sophistication (*understanding*) can readily be introduced into clinical practice

  feasible though: NL *engineering* solutions (document retrieval, extracting, speech recognition, canned text generation)
- Different views on ‘knowledge representation’
  - medical nomenclatures, terminologies, classifications (mainly used for references to documents) vs.
  - logically founded knowledge representation formalisms (for knowl-edge acquisition/question answering from documents)
    - inference rules
    - modeltheoretic semantics (true/false assertions)

A Closer Look at Information Extraction Techniques



lexical lookup

named entity recognition:
products, locations, people, organizations,...
(R/P=90%)

POS tagging:
Wermter (2004):
98%
partial parsing:
probabilistic FSAs
(HMM, Viterbi,...)

semantic tagging

definite NPs:
IBM - the company
pronouns:
B.G. - he
(P=.72; R=.63)
heuristics, ontologies:
WordNet; UMLS

fully domain-specific trans rules

template elements:
P=.65/.70;
R=.45/.47