max planck institut
informatik

Max Planck
Society

Towards a Statistically Semantic Web

Gerhard Weikum

weikum@mpi-sb.mpg.de

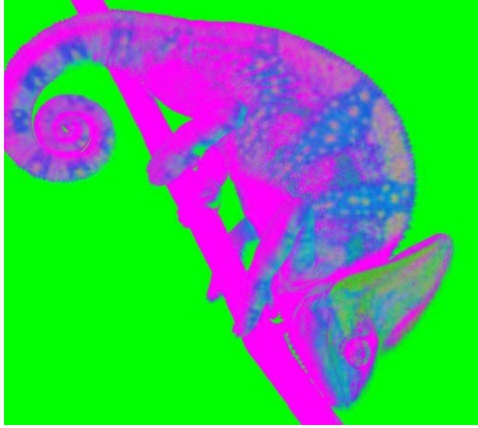
http://www.mpi-sb.mpg.de/~weikum/

Acknowledgements to **Jens Graupmann, Ralf Schenkel, Martin Theobald**
and a wealth of literature

Outline

- Motivation and Challenges
- Search **(XML, Ontologies)**
- Speed **(Top-k Query Processing)**
- Self-Organization **(P2P, Collaborative Search)**

Chameleon Words in Computer Science



fragment

page

object

segment

performance

ontology

failure

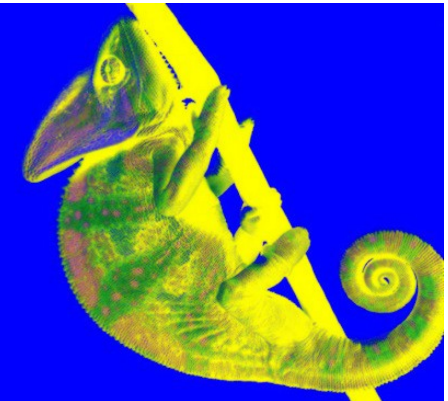
semantics

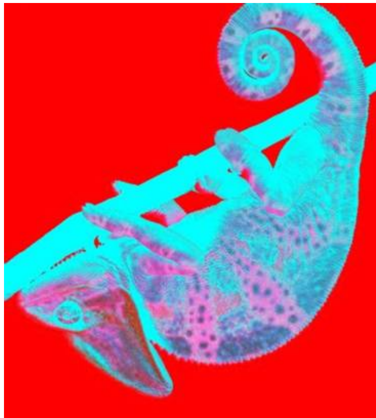
session

node

service

peer





Opinions on the Feasibility of the Semantic Web, Universal Data Integration, and Comprehensive Knowledge Bases

Tim Berners-Lee: „The Semantic Web is an extension of the current Web in which information is given meaning.“

Jeff Ullman: „There is n

Alon Halevy: „Structure

This talk:

Names + Statistics → „Semantics“

Noam Chomsky: „Whether there is also a semantics of natural language ... seems to me an open question. Pragmatics must be a central component of linguistic theory.“

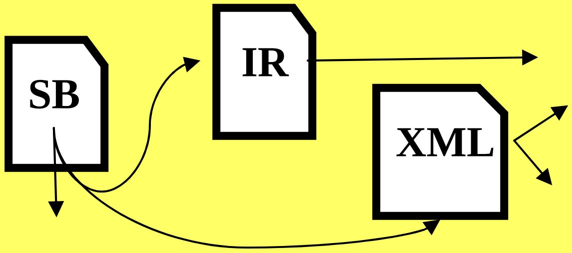
George Lakoff: „When we have multiple ways of understanding, or 'framing,' a situation, then knowledge, like truth, becomes relative to that understanding.“

Confucius: „Knowledge is to know the extent of one‘s ignorance.“

孔夫子

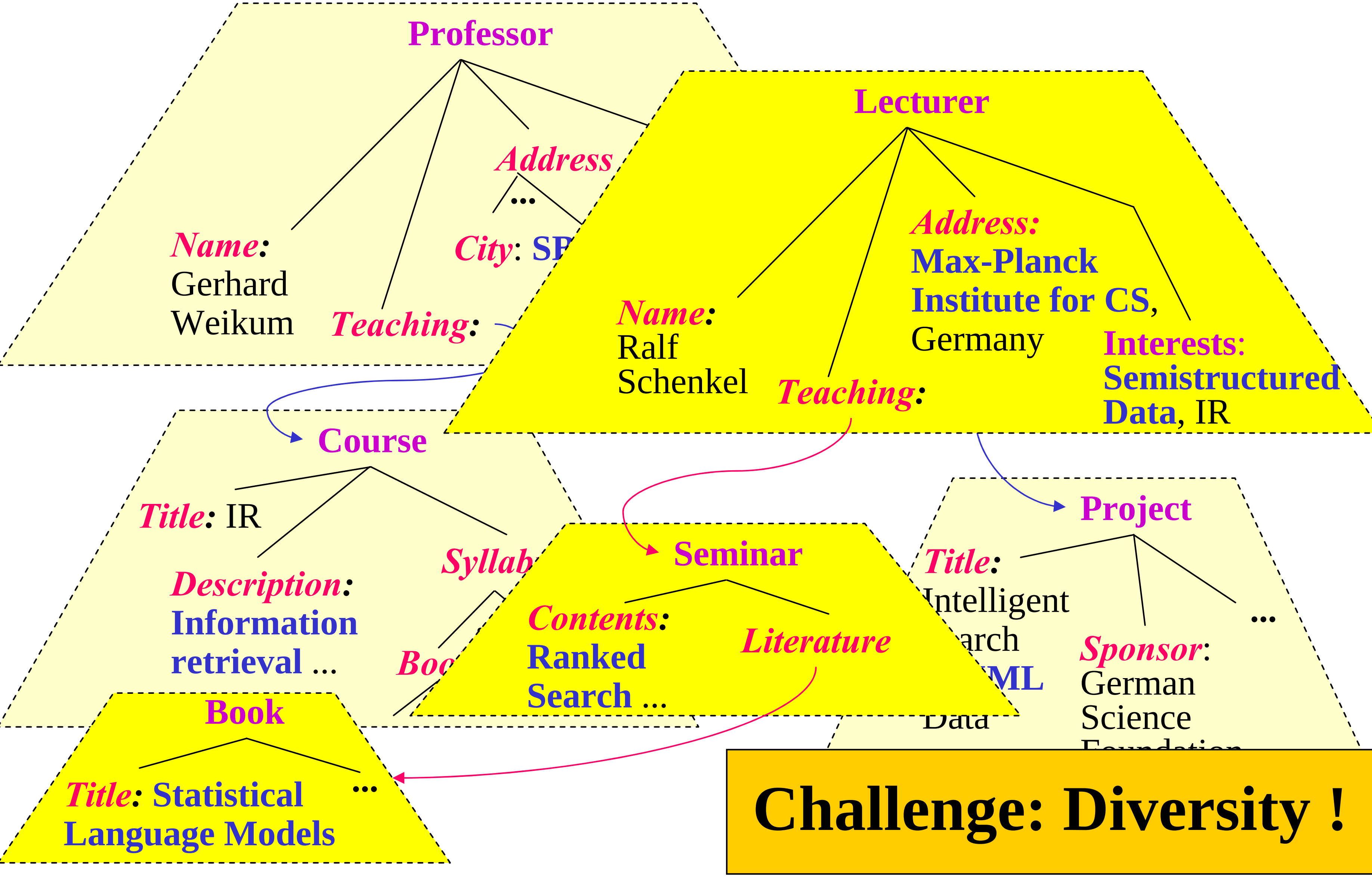
A Few Challenging Queries
(on Web / Deep Web / Intranet / Personal Info)

- **Which professors from Saarbruecken (SB) are teaching IR and have research projects on XML?**



- **Which gene expression data from Barrett tissue in the esophagus exhibit high levels of gene A01g?**
- **Which drama has a scene in which a woman makes a prophecy to a Scottish nobleman that he will become king?**
- **Who was the woman from Paris that I met at the PC meeting where Paolo Atzeni was PC Chair?**
- **Are there any published theorems that are equivalent to or subsume my latest mathematical conjecture?**

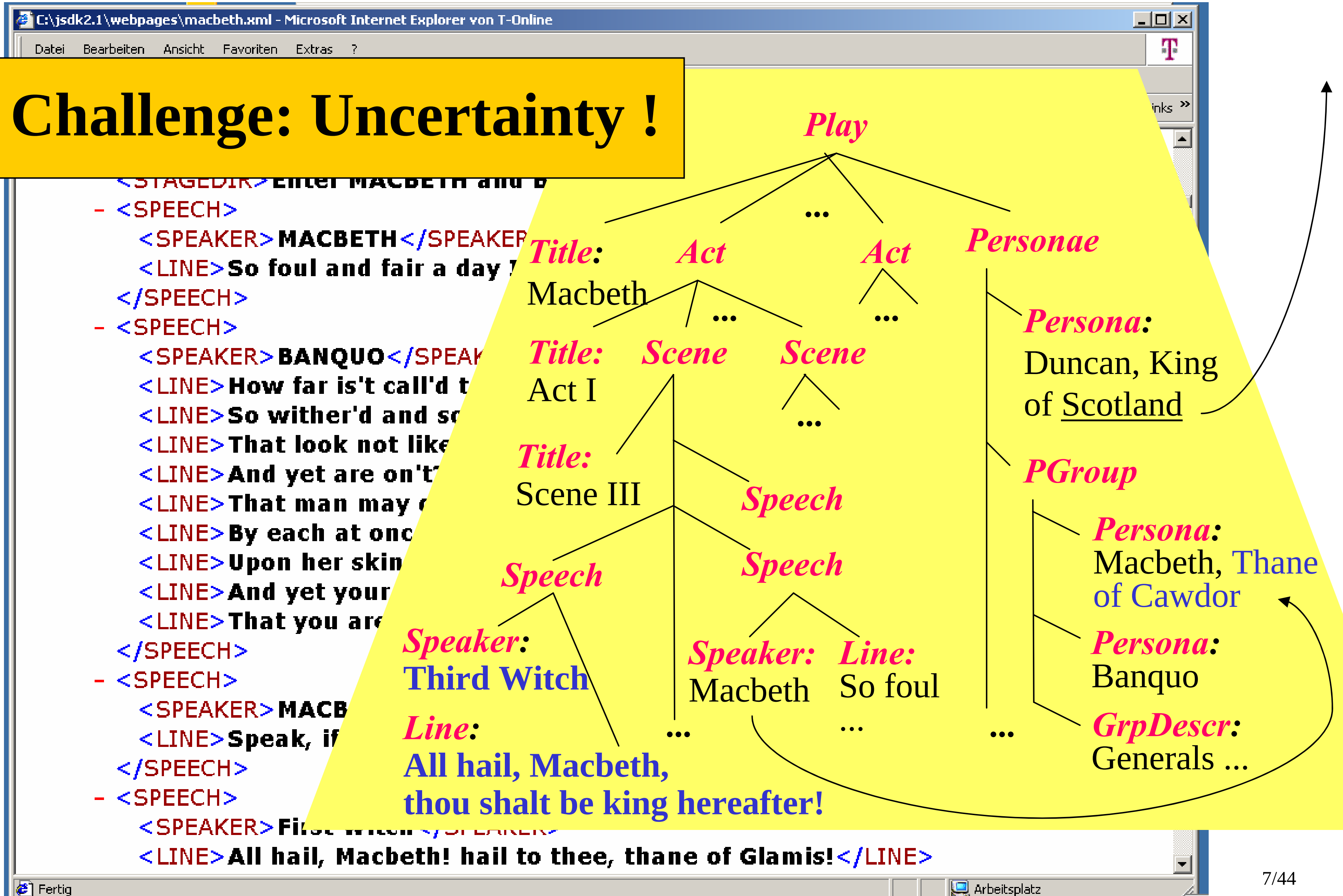
What if the Semantic Web Existed and All Information Were in XML?



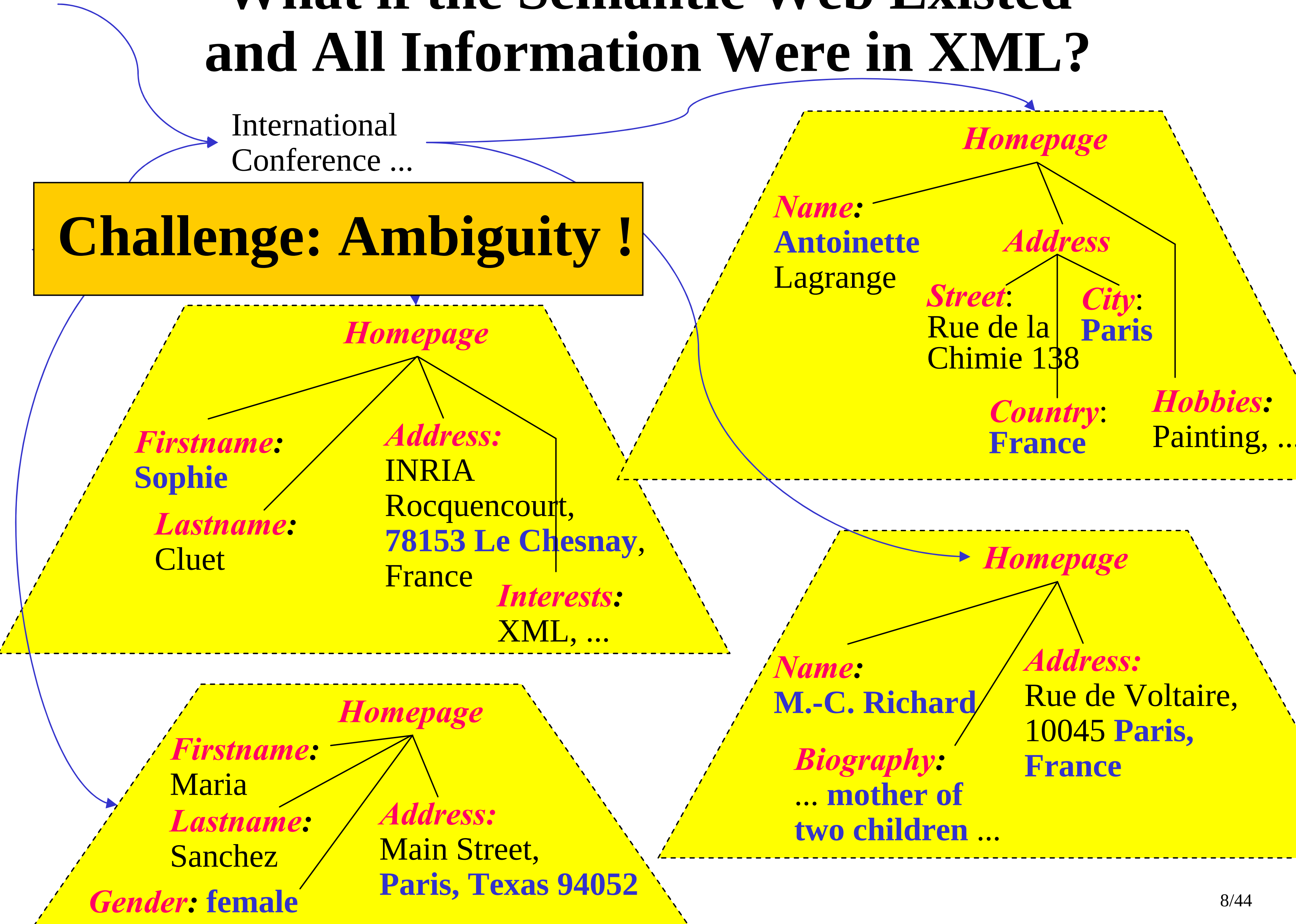
Challenge: Diversity !

What if the Semantic Web Existed and All Information Were in XML?

Challenge: Uncertainty !



What if the Semantic Web Existed and All Information Were in XML?



Observations and Challenges (1)

Observation:

Despite all structure, tags, and „semantic" metadata, information will exhibit ***diversity***, ***ambiguity***, ***uncertainty***

Implication:

Information search faces IR dilemma – drown in results or almost empty result – and thus needs ***ranked retrieval***

Challenge:

Combine the best of ***precise querying*** from DB world with ***vague search*** and ***relevance assessment*** from IR, Web & learning communities
[and expressive ***logical inferences*** from AI]

Outline

- ✓ Motivation and Challenges
- Search **(XML, Ontologies)**
- Speed **(Top-k Query Processing)**
- Self-Organization **(P2P Collaborative Search)**

XML-IR: History and Related Work

1995

Web query languages:

- W3QS (Technion Haifa)
- Araneus (U Roma)
- Lorel (Stanford U)

IR on structured docs (SGML):

- HySpirit (U Dortmund)
- HyperStorM (GMD Darmstadt)
- WHIRL (CMU)

2000

XML query languages:

- XML-QL (AT&T Labs)
- XPath 1.0 (W3C)
- XPath 2.0 (W3C)
- XQuery (W3C)
- TeXQuery (AT&T Labs)

IR on XML:

- XIRQL (U Dortmund)
- XXL (U Saarland / MPI)
- ApproXQL (U Berlin / U Munich)
- ELIXIR (U Dublin)

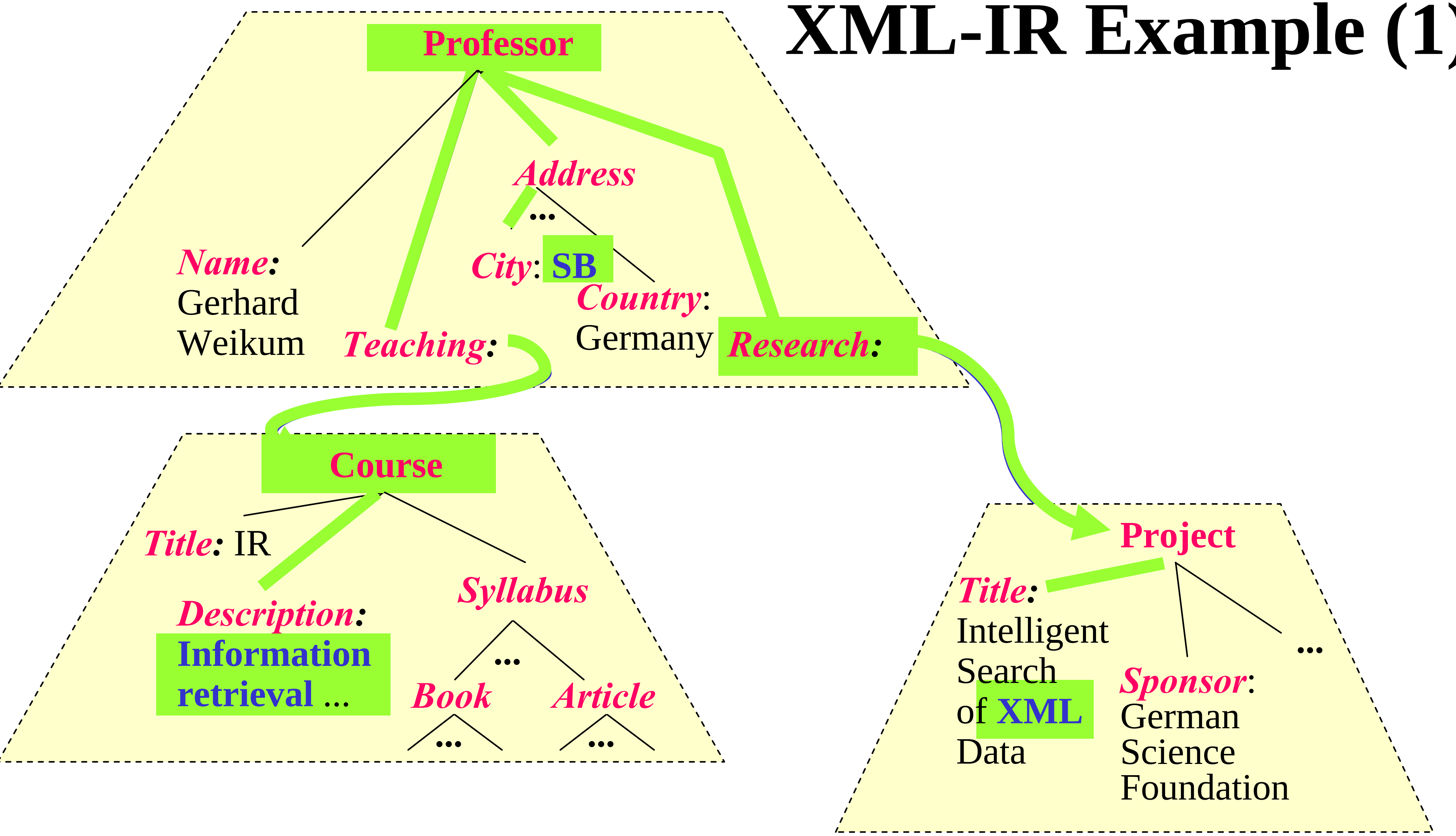
INEX benchmark

- PowerDB-IR (ETH Zurich)
- JuruXML (IBM Haifa)
- XSearch (Hebrew U)
- Timber (U Michigan)
- XRank (Cornell U)
- Compass (U Saarland / MPI)
- FleXPath (AT&T Labs)

2005

Commercial software
(Verity?, Oracle?, Google?, ...)

XML-IR Example (1)

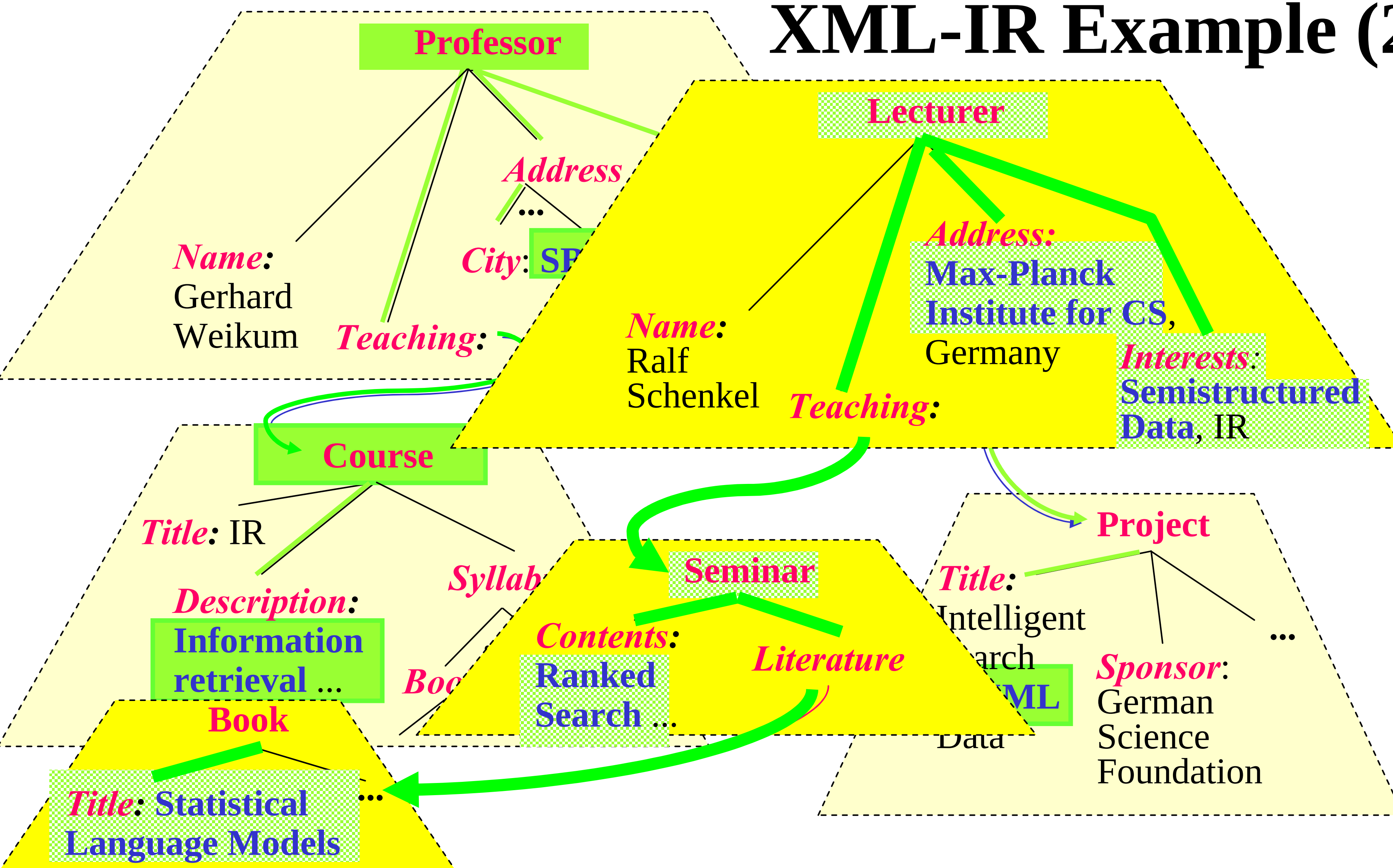


Select *P, C, R* From *Index*

Where *Professor* As *P* And *P = „ Saarbruecken“*

And *P//* *Course = „ IR“* As *C* And *P//* *Research = „ XML“* As *R*

XML-IR Example (2)



<u>Select</u> ***P*, *C*, *R*** <u>From</u> ***Index***

<u>Where</u> ***~Professor*** <u>As</u> ***P*** <u>And</u> ***P = „~Saarbruecken“***

<u>And</u> ***P//~Course = „~IR“*** <u>As</u> ***C*** <u>And</u> ***P//~Research = „~XML“*** <u>As</u> ***R***

XML-IR Concepts

applicable to both XML and HTML data graphs

Where clause: conjunction of restricted *path expressions*
with binding of variabl

***Elementary conditions* on names and c**

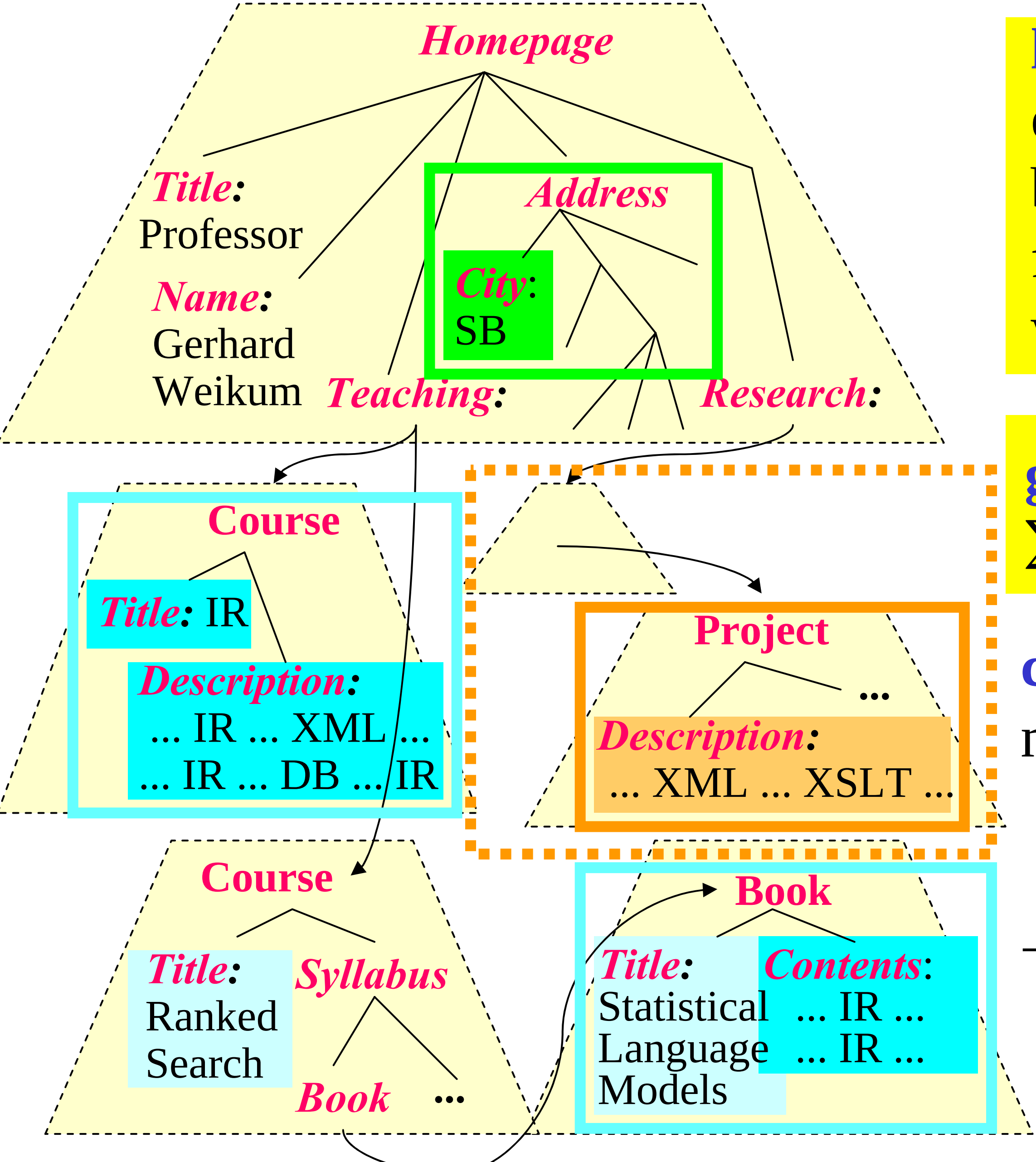
> ***Query result:***
> - **query is a pattern with relaxable conditions**
> - **results are approximate matches to query with similarity scores**

Select *P, C, R* From *Index*
Where *~Professor* As *P*
And *P* = *„Saarbruecken“*
And *P//~Course* = *„Information Retrieval“* As *C*
And *P//~Research* = *„~XML“* As *R*

„Semantic“ *similarity conditions* on names and contents
~Research = *„~XML“*

Relevance scoring based on
tf*idf similarity of contents,
ontological similarity of names,
aggregation of local scores into global scores

XML-IR Scoring Model

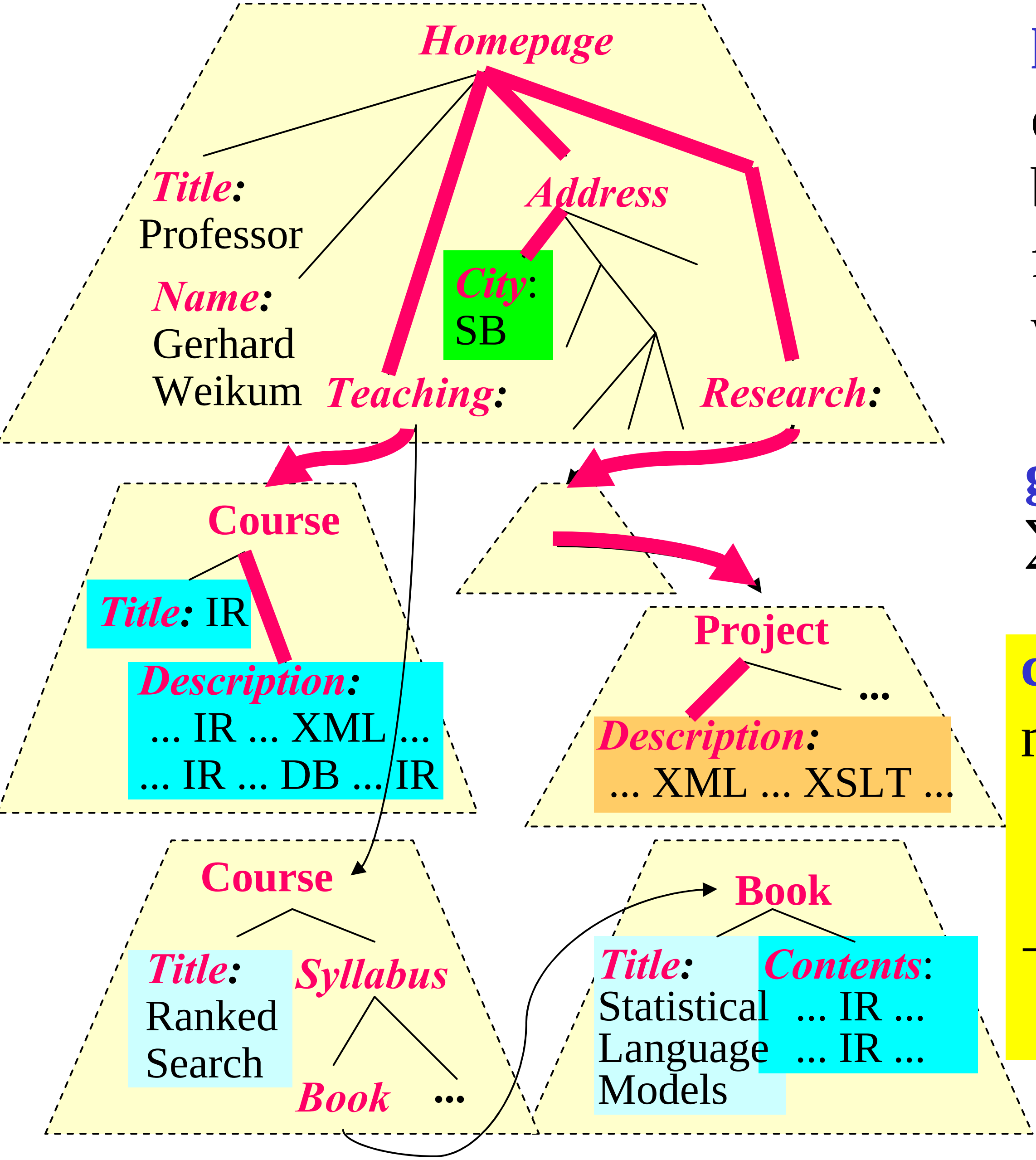


local score for elementary condition:
based on tf*idf-style statistics
for node or node context
with score propagation

global score for query:
$\sum$ local scores * compactness

compactness of result:
max{$\sum$ node & edge weights | graph connecting matching nodes}
$\rightarrow$ generalized MST (related to Steiner trees)

XML-IR Scoring Model

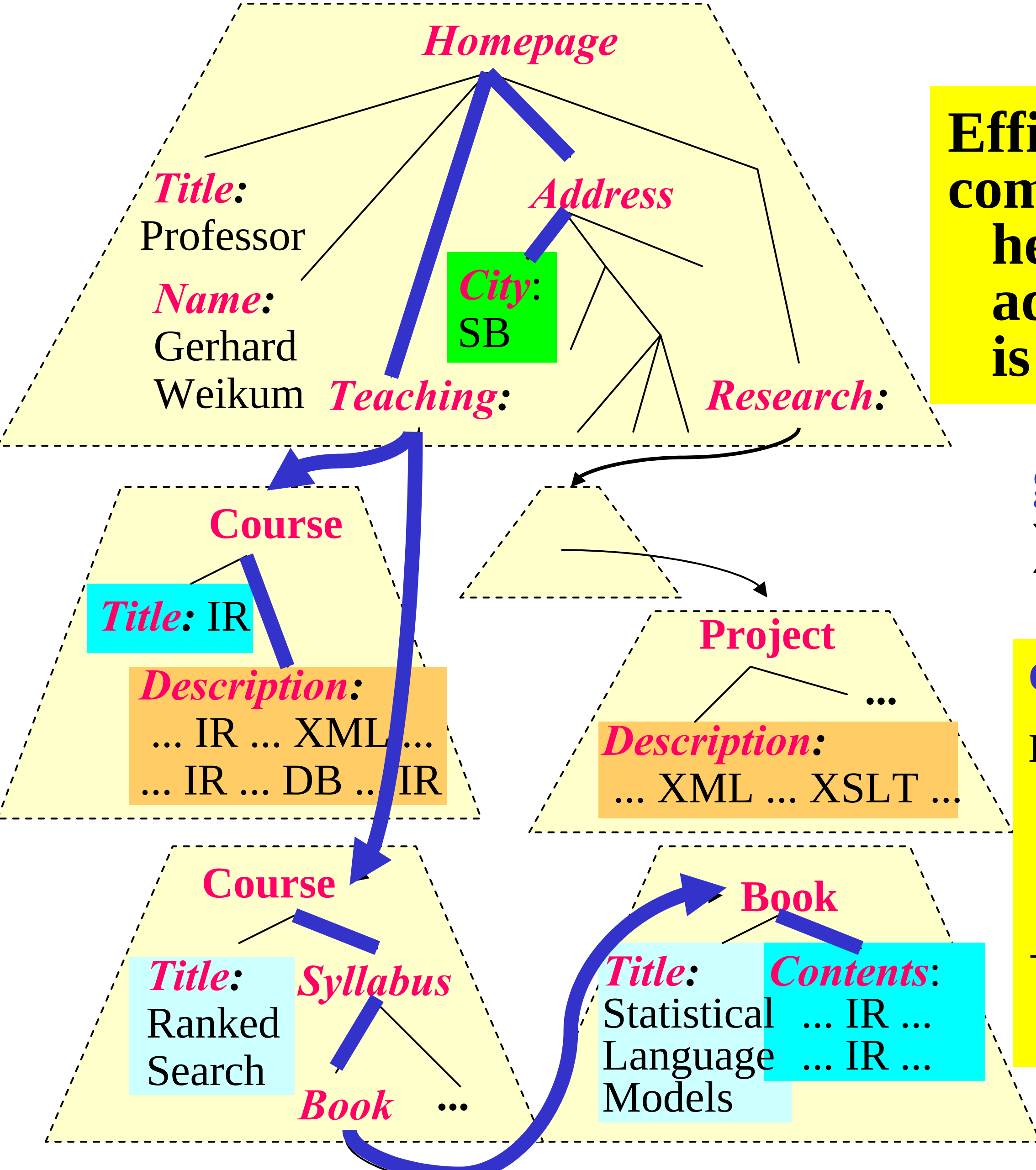


local score for elementary condition: based on tf*idf-style statistics for node or node context with score propagation

global score for query: $\sum$ local scores * compactness

compactness of result: max{$\sum$ node & edge weights | graph connecting matching nodes}
→ generalized MST (related to Steiner trees)

XML-IR Scoring Model



local score for
:
statistics
ext
on

Efficient score computation:
heuristics work;
advanced algorithms is open issue

global score for query:
$\sum$ local scores * compactness

compactness of result:
max{$\sum$ node & edge weights | graph connecting matching nodes}
$\rightarrow$ generalized MST (related to Steiner trees)

On Thesauri and Ontologies

Taxonomy: classification of concepts into groups (and trees of groups)

Thesaurus: repository („treasure“) of synonyms
(and other relationships between words and concepts)

Ontology: metaphysical study of the nature of being & existence

Ontology (new definition): structured repository of knowledge
with a description of concepts and relationships,
possibly in the form of description logics formula

Reasoning on Ontologies and Thesauri:

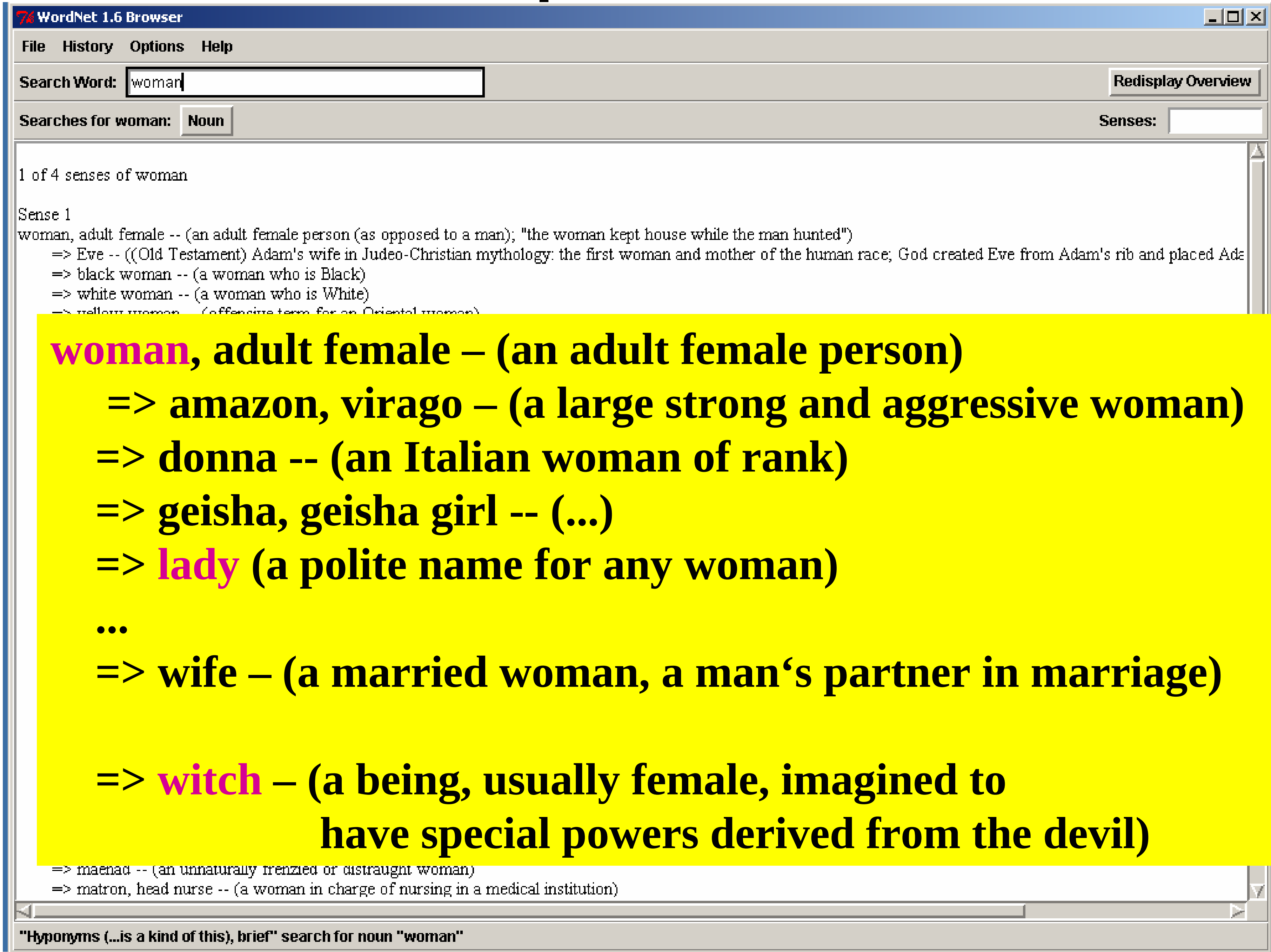
Professor $\subseteq$ Lecturer $\cap$ $\exists$ hasStaff.Secretary
Teaching $\supseteq$ Cour
Professor $\subseteq$ Acad
Academician $\subseteq$ H
Human $\subseteq$ Carniv
...

poor man‘s ontology:
pragmatic, rich, efficient

→ logical inferences
with sub-FOL calculus

→ transitive closures,
shortest paths, etc.
along generalizations

Example WordNet



woman, adult female – (an adult female person)

=> amazon, virago – (a large strong and aggressive woman)

=> donna -- (an Italian woman of rank)

=> geisha, geisha girl -- (...)

=> lady (a polite name for any woman)

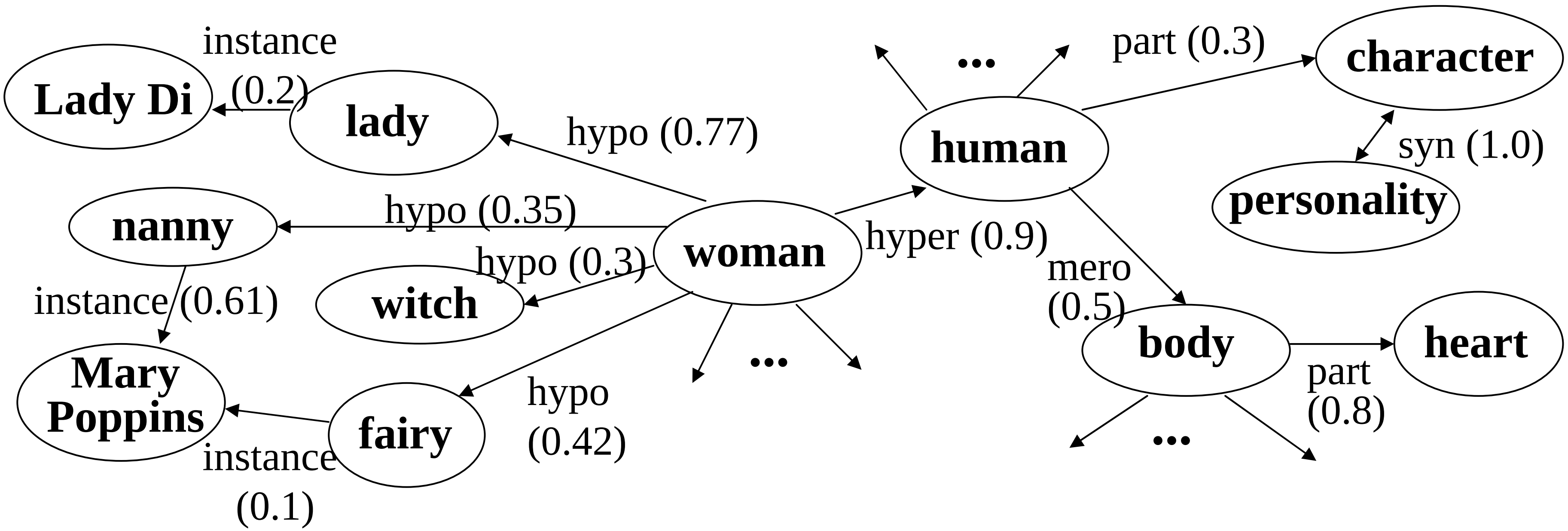
...

=> wife – (a married woman, a man's partner in marriage)

=> witch – (a being, usually female, imagined to have special powers derived from the devil)

Ontology Graph

An ontology graph is a directed graph with concepts (and their descriptions) as nodes and semantic relationships as edges (e.g., hypernyms).



Weighted edges capture strength of relationships
→ key for identifying closely related concepts

Statistics for Weighted Ontological Relations

Gather statistics from large corpus or by (focused) Web crawl

Various correlation measures for sim(c1, c2):

Dice coefficient:

$$\frac{2\left|\{\,docs\ with\ c1\,\} \cap \{\,docs\ with\ c2\,\}\right|}{\left|\{\,docs\ with\ c1\,\}\right| + \left|\{\,docs\ with\ c2\,\}\right|}$$

Jaccard coefficient:

$$\frac{\left|\{\,docs\ with\ c1\,\} \cap \{\,docs\ with\ c2\,\}\right|}{\left|\{\,docs\ with\ c1\,\}\right| + \left|\{\,docs\ with\ c2\,\}\right| - \left|\{\,docs\ with\ c1\ and\ c2\,\}\right|}$$

Conditional probabilites:

$$P[\,doc\ has\ c1 \mid doc\ has\ c2\,]$$

Transitive similarity:

$$sim^*(c1, cn) = \ max\{ \prod_{i=1..n-1} sim(c_i, c_{i+1}) \mid all\ \ paths\ \ from\ c1\ to\ cn \}$$

compute by (adaptation of) Dijkstra‘s shortest-path algorithm

Benefits from Ontology Service

Ontology service accessible via SOAP or RMI
Ontology filled with WordNet, geo gazetteer,
focused crawl results, extracted tables & forms

useful for:

- Threshold-based query expansion
- Query keyword disambiguation
- Support for automatic tagging of HTML and enhanced XML tags
- Mapping of concept-value query conditions onto Deep-Web portals

Query Expansion

Threshold-based query expansion:
substitute ~w by (c_1 | ... | c_k) with all c_i for which sim(w, c_i) ≥ δ

„Old hat" in IR; highly disputed for danger of topic dilution

Approach to careful expansion:

- determine phrases from query or best initial query results (e.g., forming 3-grams and looking up ontology/thesaurus entries)
- if uniquely mapped to one concept then expand with synonyms and weighted hyponyms

Problem: choice of threshold δ → see Top-k QP

Query Expansion Example

From TREC 2004 Robust Track:

Title: **International Organized Crime**

Description: Identify organizations that participate in international criminal activity, the activity, and, if possible, collaborating organizations and the countries involved.

Query = {international[0.145|1.00],
~META[1.00|1.00][{gangdom[1.00|1.00], gangland[0.742|1.00],
"organ[0.213|1.00] & crime[0
mafia[0.154|1.00], "sicilian[0.
"black[0.066|1.00] & hand[0.
organ[0.213|1.00], crime[0.31
columbian[0.686|0.20], cartel

135530 sorted accesses in 11.07

Results:

1. Interpol Chief on Fight Agai
2. Economic Counterintelligen
3. Dresden Conference Views
4. Report on Drug, Weapons S
5. SWITZERLAND CALLED SOFT ON CRIME

…

Search Word: organized crime

Redisplay Overvie

Searches for organized crime: Noun

1 sense of organized crime

Sense 1

organized crime, gangland, gangdom

=> yakuza -- (organized crime i
enterprises)

=> Mafia, Maffia, Sicilian Mafia

=> Black Hand -- (a secret terro

=> Camorra -- (a secret society

=> syndicate, crime syndicate,
organized criminal activities)

> **Let us take, for example, the case of Medellin cartel's boss Pablo Escobar. Will the fact that he was eliminated change anything at all? No, it may perhaps have a psychological effect on other drug dealers but,**
> **...**

> **... for organizing the illicit export of metals and import of arms. It is extremely difficult for the law-enforcement organs to investigate and stamp out corruption among leading officials.**
> **...**

> **A parliamentary commission accused Swiss prosecutors today of doing little to stop drug and money-laundering international networks from pumping billions of dollars through Swiss companies.**
> **...**

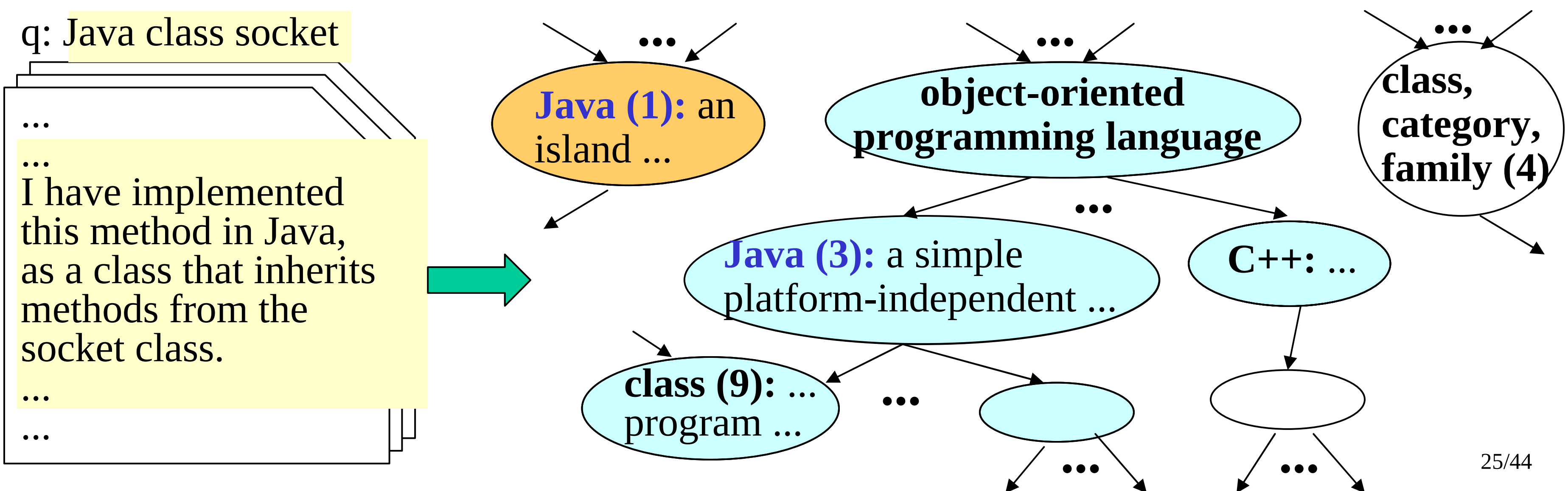
Keyword-to-Concept Mapping and Word Sense Disambiguation (1)

Example: **„Java class socket“** vs. **„Java beach snorkeling“**
Which concept should „Java“ be mapped to for query expansion?

Note: unlike in LSI or pLSI, concepts are explicit, not latent!

Approach for query keyword disambiguation:

- form **contexts con(w) and con(c_i)** for keyword w and potential target concepts $c_i \in \{c_1, ..., c_k\}$
- bag-of-words similarity **sim(con(w), con(c))** based on cos or KL diff
- choose concept argmax_c {sim(con(w), con(c))}



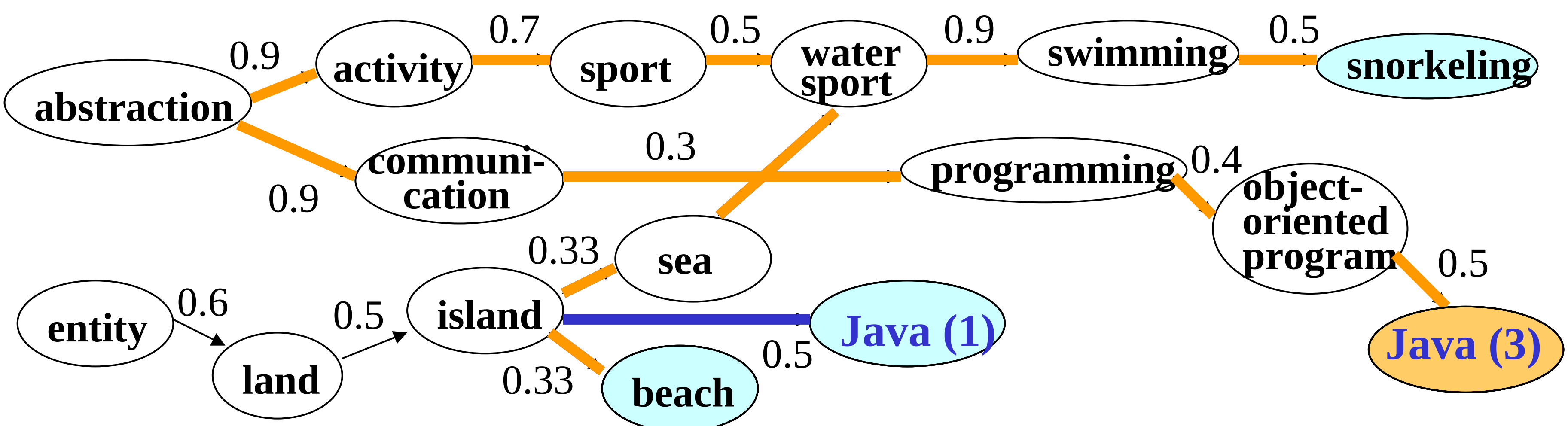
Keyword-to-Concept Mapping and Word Sense Disambiguation (2)

Example: „**Java class socket**“ vs. „**Java beach snorkeling**“

Which concept should „Java“ be mapped to for query expansion?

Alternative approach for query keyword disambiguation:

- consider potential target concepts for all query keywords together
- choose **concepts (c_1, ..., c_m) for words (w_1, ..., w_m)** according to **sim * compactness** where
  - **sim** ~ aggregation of con(w_i)-to-con(c_i) similarities,
  - **compactness** ~ weight of MST for {c_1, ..., c_m}



Observations and Challenges (2)

Observation:

Explicit ontologies/thesauri and statistical models need to be combined for ranked retrieval of richly annotated but highly heterogeneous XML data

Challenges:

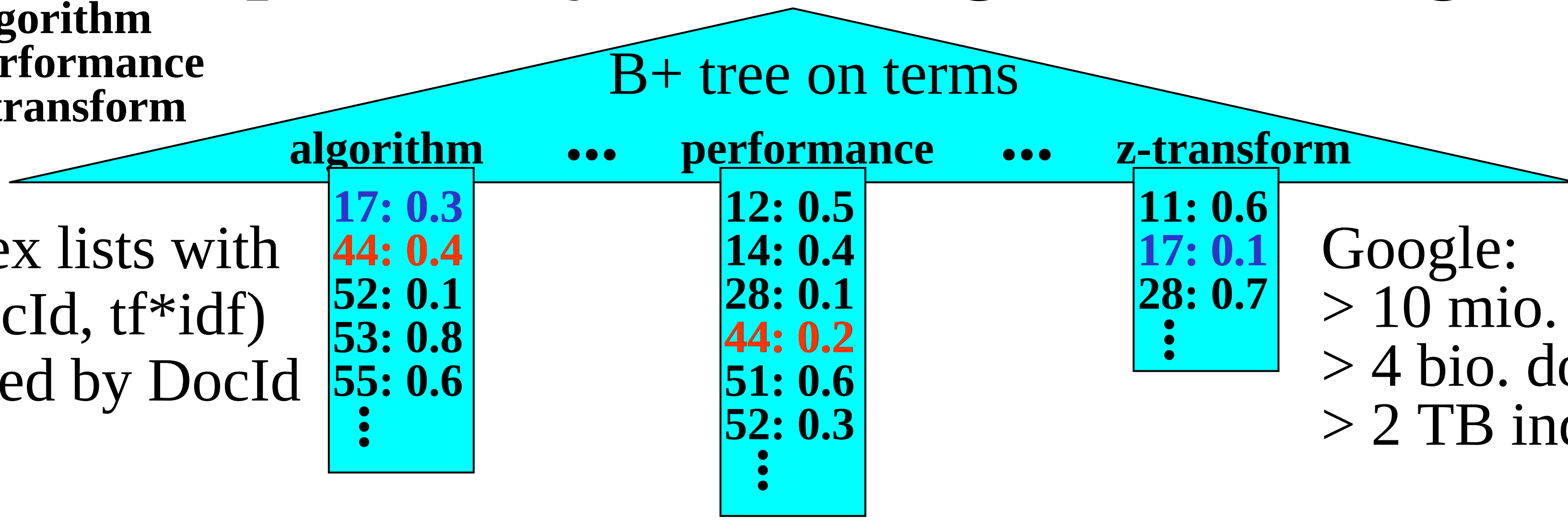
- Develop full-fledged statistical language model for XML subgraph scoring
- Constructing statistically quantified ontologies from rich sources (WordNet, Wikipedia, bookmarks, etc.)
- Combine uncertainty of automatic tagging and query mapping with query result ranking in a comprehensive probabilistic algebra
- Efficient query processing and optimization

Outline

- ✓ Motivation and Challenges
- ✓ Search **(XML, Ontologies)**
- Speed **(Top-k Query Processing)**
- Self-Organization **(P2P Collaborative Search)**

Top-k Query Processing with Scoring

q: algorithm
performance
z-transform



index lists with (DocId, tf*idf) sorted by DocId

Google:
> 10 mio. terms
> 4 bio. docs
> 2 TB index

<u>Given:</u> query $q = t_1\ t_2\ \ldots\ t_z$ with z (conjunctive) keywords
similarity scoring function score(q,d) for docs $d \in D$, e.g.: $\vec{q} \cdot \vec{d}$

<u>Find:</u> top k results with regard to score(q,d) $=\mathrm{aggr}\{s_i(d)\}$(e.g.: $\Sigma_{i \in q}\ s_i(d)$)

Naive QP algorithm:

candidate-docs := ∅;
for i=1 to z do {
  candidate-docs := candidate-docs ∪ index-lookup(ti) };
for each dj ∈ candidate-docs do {compute score(q,dj)};
sort candidate-docs by score(q,dj) descending;

TA (Fagin'01; Güntzer/Kießling/Balke; Nepal et al.)

scan all lists L_i (i=1..m) in parallel:
consider d_j at position pos_i in Li;
$high_i := s_i(d_j)$;
if $d_j \notin$ top-k then {
look up $s_\nu(d_j)$ in all lists L_ν with $\nu \neq i$; // random access
compute $s(d_j) := aggr \{s_\nu(dj) \mid \nu=1..m\}$;
if $s(d_j)$ > min score among top-k then
add d_j to top-k and remove min-score d from top-k; };
if min score among top-k $\geq$ aggr {$high_\nu \mid \nu=1..m$} then exit;

but random accesses are expensive !
→ TA-sorted
→ Prob-sorted

m=3
aggr: sum
k=2

f: 0.5	a: 0.55	h: 0.35
b: 0.4	b: 0.2	d: 0.35
c: 0.35	f: 0.2	b: 0.2
a: 0.3	g: 0.2	a: 0.1
h: 0.1	c: 0.1	c: 0.05
d: 0.1		f: 0.05

top-k:

~~f: 0.75~~
a: 0.95
b: 0.8

applicable to XML data:
course = „~ Internet“ and ~topic = „performance“

TA-Sorted

scan index lists in parallel:

consider d_j at position pos_i in Li;

$E(d_j) := E(d_j) \cup \{i\}$; $high_i := s_i(q,d_j)$;

bestscore(d_j) := aggr{x_1, ..., x_m)

with $x_i := s_i(q,d_j)$ for $i \in E(d_j)$, $high_i$ for $i \notin E(d_j)$;

worstscore(dj) := aggr{x_1, ..., x_m)

with x_i := si(q,d_j) for $i \in E(d_j)$, 0 for $i \notin E(d_j)$;

top-k := k docs with largest worstscore;

if min worstscore among top-k $\geq$

max bestscore{d | d not in top-k} then exit;

m=3

aggr: sum

k=2

f: 0.5	a: 0.55	h: 0.35
b: 0.4	b: 0.2	d: 0.35
c: 0.35	f: 0.2	b: 0.2
a: 0.3	g: 0.2	a: 0.1
h: 0.1	c: 0.1	c: 0.05
d: 0.1		f: 0.05

top-k:

a: 0.95

b: 0.8

candidates:

~~f: 0.7 + ? $\leq$ 0.7 + 0.1~~

~~h: 0.45 + ? $\leq$ 0.45 + 0.2~~

~~c: 0.35 + ? $\leq$ 0.35 + 0.3~~

~~d: 0.35 + ? $\leq$ 0.35 + 0.3~~

~~g: 0.2 + ? $\leq$ 0.2 + 0.4~~

Top-k Queries with Probabilistic Guarantees

TA family of algorithms based on invariant (with sum as aggr)

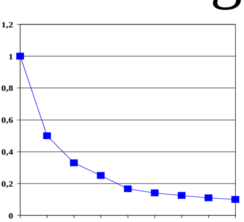
$$\underbrace{\sum_{i\in E(d)} s_i(d)}_{\text{worstscore(d)}} \le s(d) \le \underbrace{\sum_{i\in E(d)} s_i(d) + \sum_{i\notin E(d)} high_i}_{\text{bestscore(d)}}$$

Relaxed into probabilistic invariant

$$p(d) := P[s(d) > \delta] = P[\sum_{i\in E(d)} s_i(d) + \sum_{i\notin E(d)} S_i > threshold]$$

$$= P[\sum_{i\notin E(d)} S_i > threshold - \sum_{i\in E(d)} s_i(d)] =: P[\sum_{i\notin E(d)} S_i > \delta'] \le \varepsilon$$

where the RV S_i has some (postulated and/or estimated) distribution in the interval $(0, high_i]$



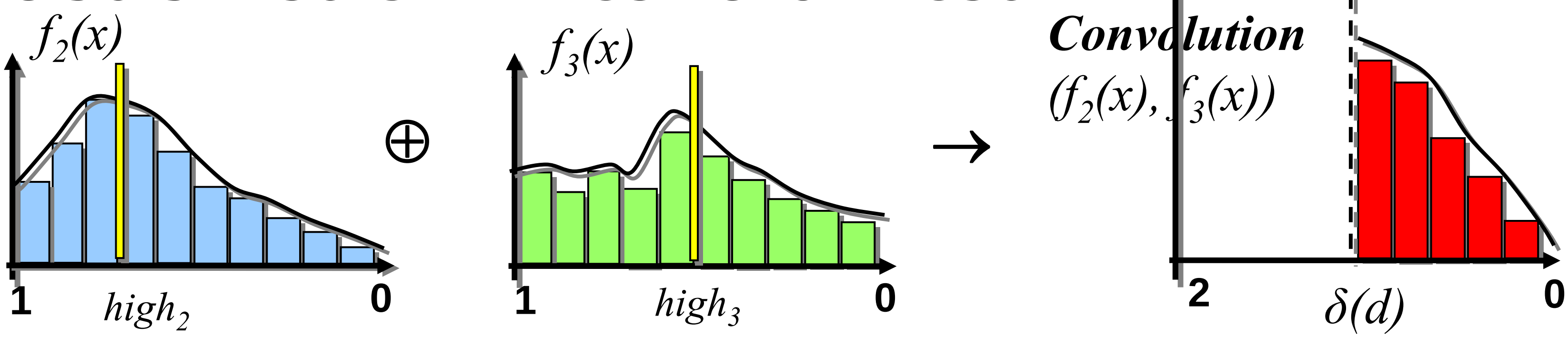
$high_i$

S_1	S_2	S_3
f: 0.5	a: 0.55	h: 0.35
b: 0.4	b: 0.2	d: 0.35
c: 0.35	f: 0.2	b: 0.2
a: 0.3	g: 0.2	a: 0.1
h: 0.1	c: 0.1	c: 0.05
d: 0.1		f: 0.05

- ***Discard candidates with $p(d) \le \varepsilon$***
- **Exit index scan when candidate list empty**

Probabilistic Threshold Test

cand doc d with
$2 \notin E(d)$,
$3 \notin E(d)$



- **postulating *uniform or Zipf* score distribution in $[0, high_i]$**
  - **compute convolution using LSTs**
  - **use Chernoff-Hoeffding tail bounds or generalized bounds for correlated dimensions (Siegel 1995)**
- **fitting *Poisson* distribution (or Poisson mixture)**
  - **over equidistant values:** $P[d = v_j] = e^{-\alpha_i} \frac{\alpha_i^{\,j-1}}{(j-1)!}$
  - **easy and exact convolution**
- **distribution approximated by *histograms*:**
  - **precomputed for each dimension**
  - **dynamic convolution at query-execution time**

engineering-wise histograms work best!

with *independent* Si's or with *correlated* Si's

Performance Results for .Gov Queries

on .GOV corpus from TREC-12 Web track:

1.25 Mio. docs (html, pdf, etc.)

50 keyword queries, e.g.:

- ***„Lewis Clark expedition“,***
- ***„juvenile delinquency“,***
- ***„legalization Marihuana“,***
- ***„air bag safety reducing injuries death facts“***

speedup by factor 10 at high precision/recall (relative to TA-sorted);

aggressive queue mgt. even yields factor 100 at 30-50 % prec./recall

	TA-sorted	Prob-sorted (smart)
#sorted accesses	2,263,652	527,980
elapsed time [s]	148.7	15.9
max queue size	10849	400
relative recall	1	0.69
rank distance	0	39.5
score error	0	0.031

.Gov Expanded Queries

on .GOV corpus with query expansion based on WordNet synonyms:

50 keyword queries, e.g.:

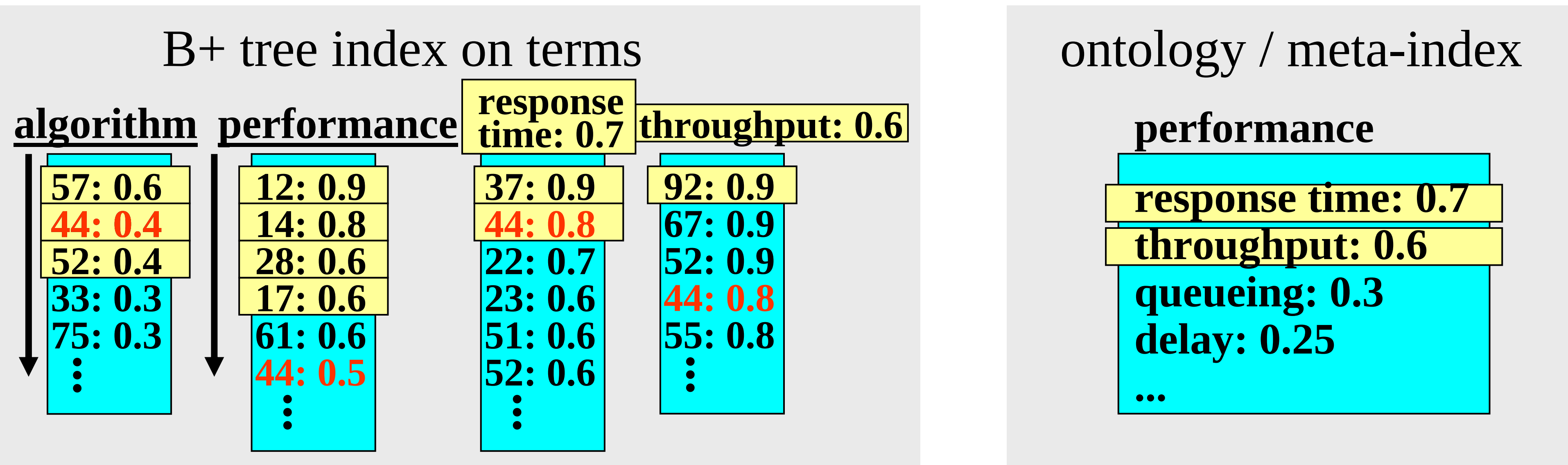
- **„juvenile delinquency *youth minor crime law jurisdiction offense prevention*“,**
- **„legalization marijuana *cannabis drug soft leaves plant smoked chewed euphoric abuse substance possession control pot grass dope weed smoke*“**

	TA-sorted	Prob-sorted (smart)
#sorted accesses	22,403,490	18,287,636
elapsed time [s]	7908	1066
max queue size	70896	400
relative recall	1	0.88
rank distance	0	14.5
score error	0	0.035

Handling Ontology-Based Query Expansions

consider expandable query ***„algorithm and ~performance“***
with score $\Sigma_{i \in q} \{\max_{j \in onto(i)} \{ sim(i,j)*sj(d)) \}\}$

dynamic query expansion with
incremental on-demand merging of additional index lists



+ much more efficient than threshold-based expansion
+ no threshold tuning
+ no topic drift

Observations and Challenges (3)

Observation:

Approximations with ***statistical guarantees*** are key to obtaining ***Web-scale efficiency***
(e.g., TREC'04 Terabyte benchmark:
ca. 25 Mio. docs, ca. 700 000 terms, 5-50 terms per query)

Challenges:

- Efficient consideration of ***correlated dimensions***
- Integrated support for all kinds of XML similarity search: content & ontological sim, ***structural sim***
- ***Scheduling*** of index-scan steps and few random accesses
- Integration of top-k operator into ***physical algebra*** and ***query optimizer*** of XML engine

Outline

- ✓ Motivation and Challenges
- ✓ Search **(XML, Ontologies)**
- ✓ Speed **(Top-k Query Processing)**
- Self-Organization **(P2P Collaborative Search)**

P2P for Web Search ?

Given: overlay networks (often DHTs) à la Chord, CAN, P-Grid
How do we exploit this technology for keyword queries?

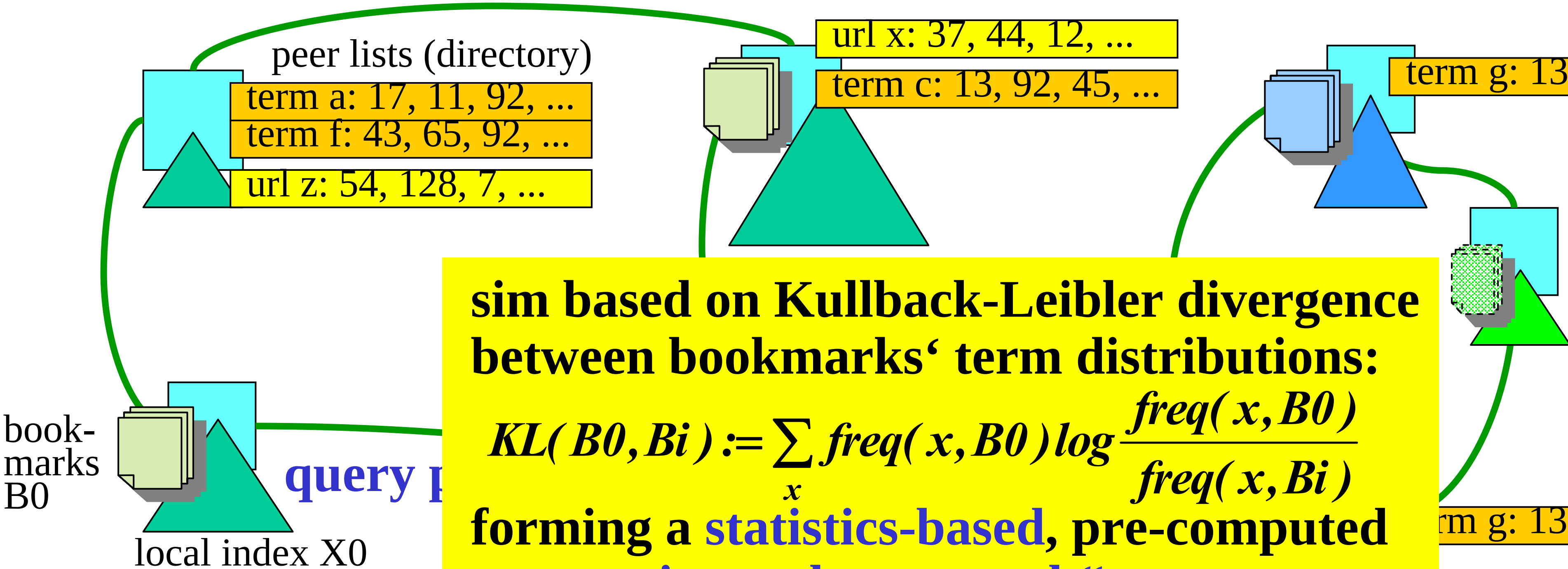
Naive idea: use multidimensional keys as I
or encode doc/query vectors as

→ grand challenge for performance at W
→ infeasible for very-high-dimensional, v
→ no support for ranking (similarity que
→ breach with autonomous behavior of i

<u>Ongoing projects:</u>
PlanetP (Rutgers U)
Odissea (Polytec Brooklyn)
Pepper (U Duisburg / CMU)
Peers (Stanford)
Pier (Berkeley)
YouServ (IBM)
GridVine (EPFL)
Minerva (MPI)
Evergrow (EU)
PeerSDI (Fudan U)

<u>Our approach</u>: use DHT's for managing (statistical) metadata only
with single-dim. keys (PeerId, term, URL)

Our Approach to P2P Query Routing



sim based on Kullback-Leibler divergence between bookmarks‘ term distributions:

$$KL(B0, Bi) := \sum_x freq(x, B0) \log \frac{freq(x, B0)}{freq(x, Bi)}$$

forming a statistics-based, pre-computed „semantic overlay network“

peer P0 first executes query locally

P2P directory has peer lists for posted terms and bookmark URLs

P0 identifies best peers Pi in terms of benefit/cost:
(sim (P0, Pi) / overlap (P0, Pi)) / cost(Pi)

Exploiting Collective Human Input for Collaborative Web Search

- Beyond Relevance Feedback -

- href links are human endorsements → PageRank, etc.
- Opportunity: online analysis of human input & behavior may compensate deficiencies of search engine

Typical scenario for 3-keyword user query: a & b & c

→ top 10 results: user clicks on ranks 2, 5, 7

→ top 10 results: u

u

u

→ top 10 results: u

u

user asks friend for tips

query logs, bookmarks, etc. provide

- **human assessments & endorsements**
- **correlations among words & concepts and among documents**

Challenge: How can we use knowledge about the collective input of all users in a large community?

Observations and Challenges (4)

Observation:

Semantic overlay networks for P2P Web search build on statistical similarity among peers

Challenges:

- Efficient benefit/cost estimation and efficient computation of global measures from local ones (idf, PageRank, KL, ...)
- From bookmark-driven query routing towards exploiting query logs and click streams
- Distributed, self-optimizing TA-sorted and Prob-sorted
- Caching, lazy replication, proactive dissemination
- Incentive mechanisms and trust management

Outline

- ✓ Motivation and Challenges
- ✓ Search **(XML, Ontologies)**
- ✓ Speed **(Top-k Query Processing)**
- ✓ Self-Organization **(P2P Collaborative Search)**

Concluding Remarks

long-term goal: **exploit the Web‘s potential for being the world‘s largest knowledge base**

- ***XML*** **and** ***Semantic Web*** **are key assets, but by themselves not sufficient; we need to cope with** ***diversity*****,** ***incompleteness*****, and** ***uncertainty***
  → absolute need for ranked retrieval
  → ***statistics*** **is key**
- **combine techniques from** ***DBS*****,** ***IR*****,** ***CL*****,** ***AI*****, and** ***ML***

- ***P2P*** **is intriguing paradigm: computing power, community input, anti-monopoly**
- **key issue is** ***quality/efficiency tradeoffs***