Bridging the Structured-Unstructured Gap

Searching the Annotated Web

Soumen Chakrabarti
IIT Bombay

http://soumen.in/doc/CSAW/

Search engine evolution

- From brittle ranking and near-duplicate results (ca. 1995) …
- … to spam filtering, link-assisted ranking, result diversification, geosensitivity
- Limited type-awareness in verticals
  - 1 kg = ? lb, distance rome venice
  - Hotels near Taj Mahal
- However, there remain information needs where cognitive burden is still very large

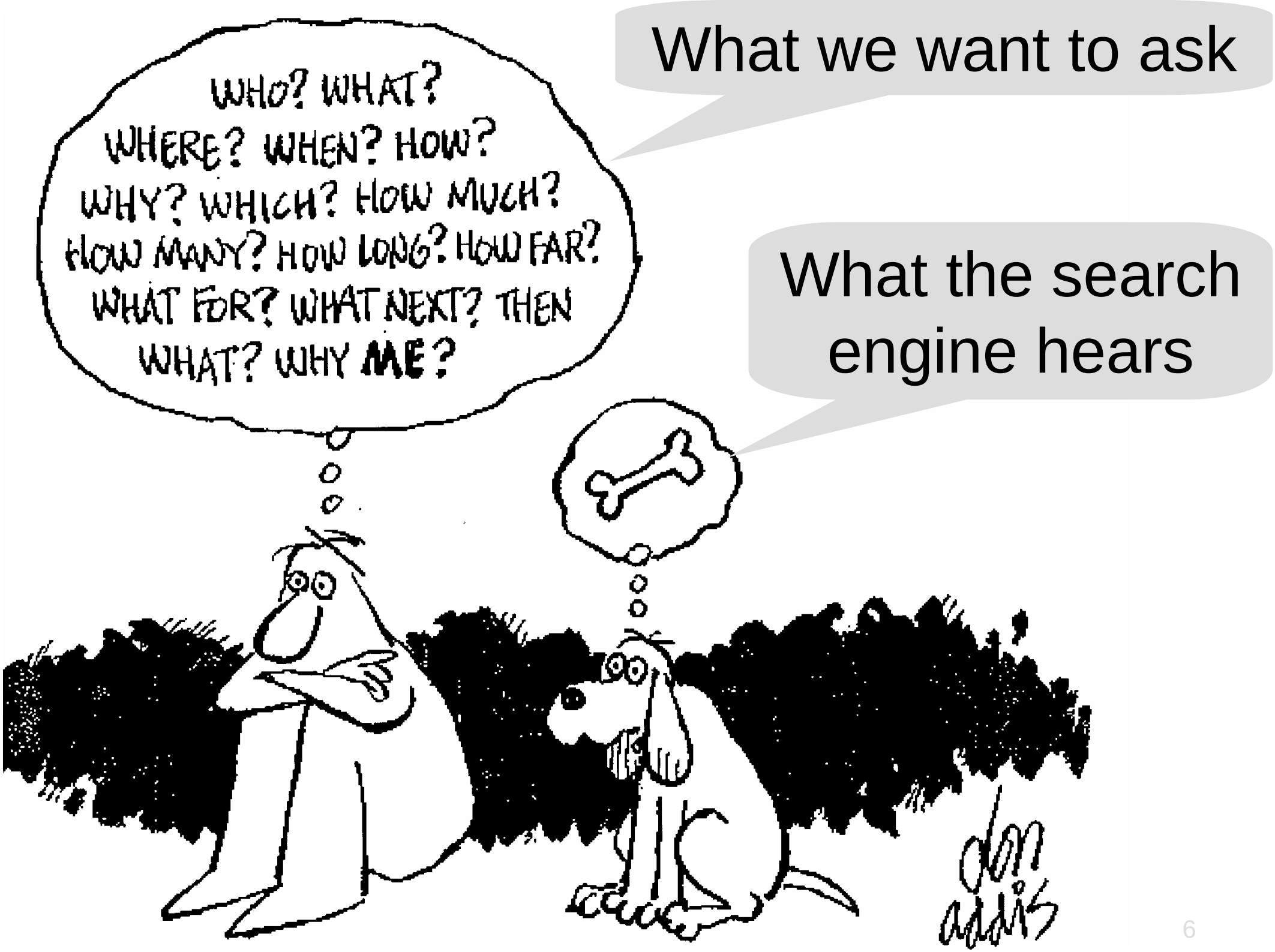
> If music had been invented sixteen years ago along with the Web, we would all be playing one-string instruments (and not making great music).

Challenging queries

- Artists who got Oscars for both acting and direction (same movie?)
- (Typical price of) Opteron motherboards with at least two PCI express slots
- How many justices serve in the International Criminal Court?
- Is the number of Oscars won directly related to production budget?
- Exxon Valdez cleanup cost

Why difficult?

- No variables
  - **?a** acts, **?a** directs movies
- No types
  - ?m ∈ ***Motherboard***, ?p ∈ ***MoneyAmount***
- No predicates
  - ?m **sells for** ?p, ?m **costs** ?p
- No aggregates
  - Large variation in Exxon Valdez estimate
- SQL, Web search, “query language envy”

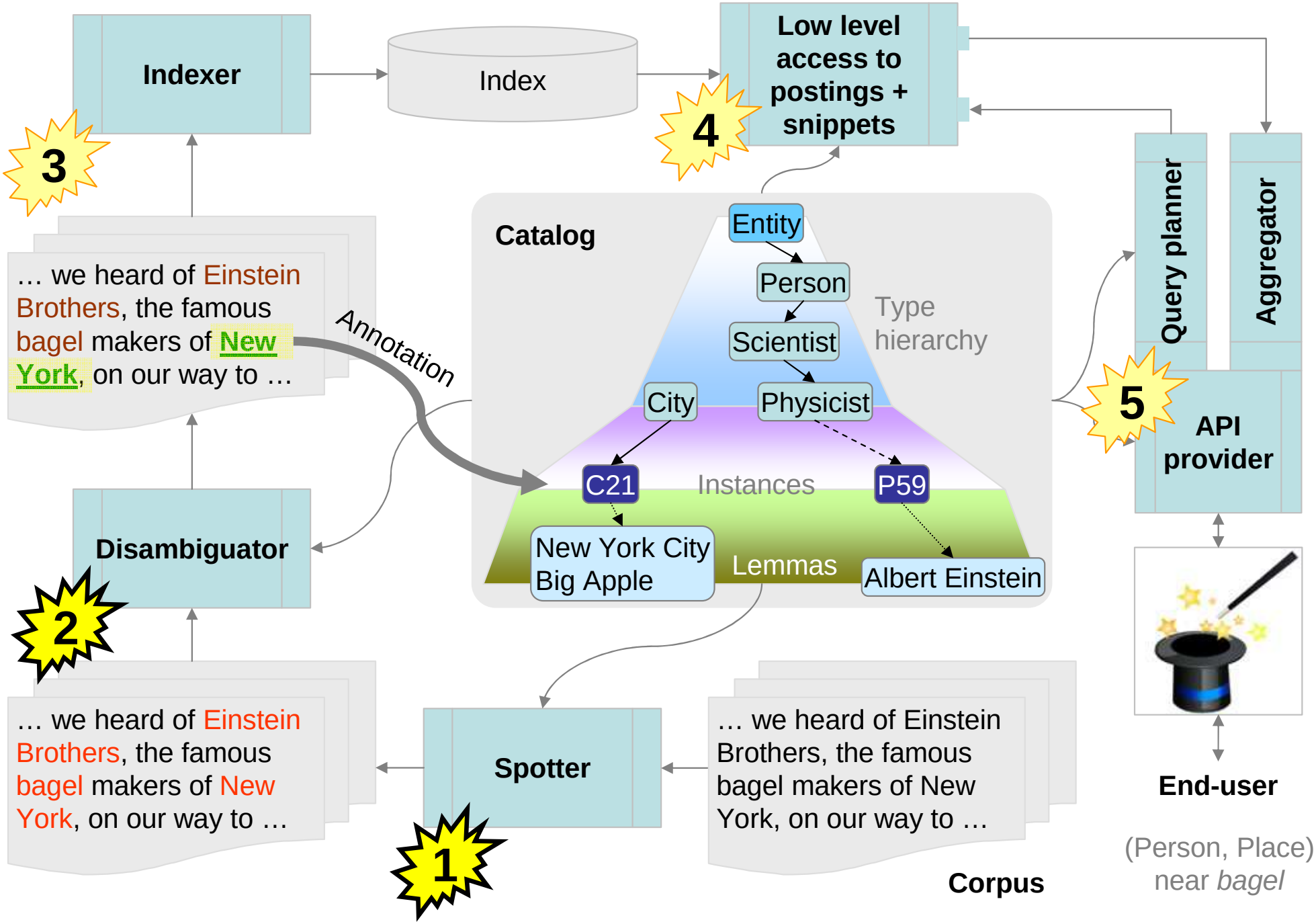

What we want to ask
WHO? WHAT?
WHERE? WHEN? HOW?
WHY? WHICH? HOW MUCH?
HOW MANY? HOW LONG? HOW FAR?
WHAT FOR? WHAT NEXT? THEN
WHAT? WHY ME?
What the search engine hears

What if we could ask…

- ?f $\in^{+}$ Category:FrenchMovie
- ?a $\in$ QType:Number
- ?p $\in$ QType:MoneyAmount
- ?c1, ?c2 are snippet contexts
- InContext(?c1, ?f, ?a, +oscar, won),
- InContext(?c2, ?f, ?p, +"production cost") or InContext(?c2, ?f, ?p, +budget)
- Aggregate(?c1, ?c2)
- Answer: list of ⟨?f, ?a, ?p⟩ tuples

Why can’t we do this today?

- Search engines record, for each token
  - Document ID/URL where token appears
  - Token offsets where it appears
- Don’t encode in the index these facts
  - “Albert” may refer to Einstein
  - That particular Einstein was a scientist
  - Who played the violin
  - Which is a musical instrument
- Nor give access to such facts in queries

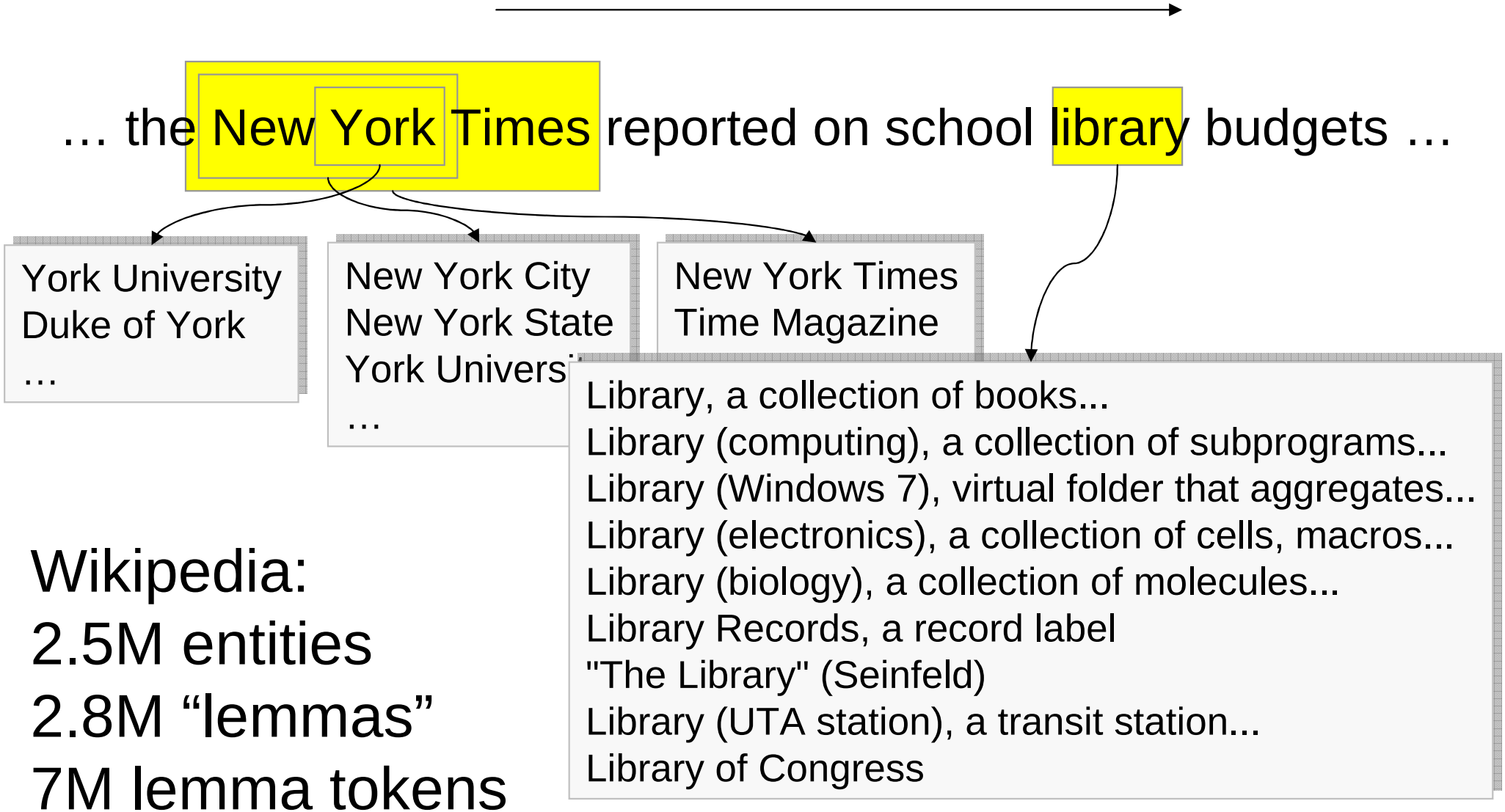

Indexer
Index
Low level access to postings + snippets
3
4
Catalog
Entity
Person
Scientist
Physicist
City
Type hierarchy
C21
Instances
P59
New York City
Big Apple
Lemmas
Albert Einstein
… we heard of Einstein Brothers, the famous bagel makers of New York, on our way to …
Annotation
Disambiguator
2
Spotter
1
… we heard of Einstein Brothers, the famous bagel makers of New York, on our way to …
Corpus
Query planner
Aggregator
5
API provider
End-user
(Person, Place) near bagel

Mentions and spots

The lack of **memory** and time efficient **libraries** in the **free software world** has been the main motivation to create the C **Minimal Perfect Hashing Library**, a portable **LGPL library**.

- A **mention** is any token segment that may be a reference to an entity in the catalog
- Mention + limited token context = **spot**
- Mentions and spots may overlap
- S_0: set of all spots on a page
- $s \in S_0$: one spot among S_0

A massive similarity join



Wikipedia:
2.5M entities
2.8M “lemmas”
7M lemma tokens
IDF, prefix/exact match, case, …

Disambiguation

- s is a spot with a mention of some entity
- Γ_s is the set of candidate entities for s
- $\gamma \in \Gamma_s$ is one candidate entity for s
- s may be best left unconnected to any entity in the catalog (“no attachment”, NA)
  - Most people mentioned on the Web are missing from Wikipedia
  - But all known constellations are in there

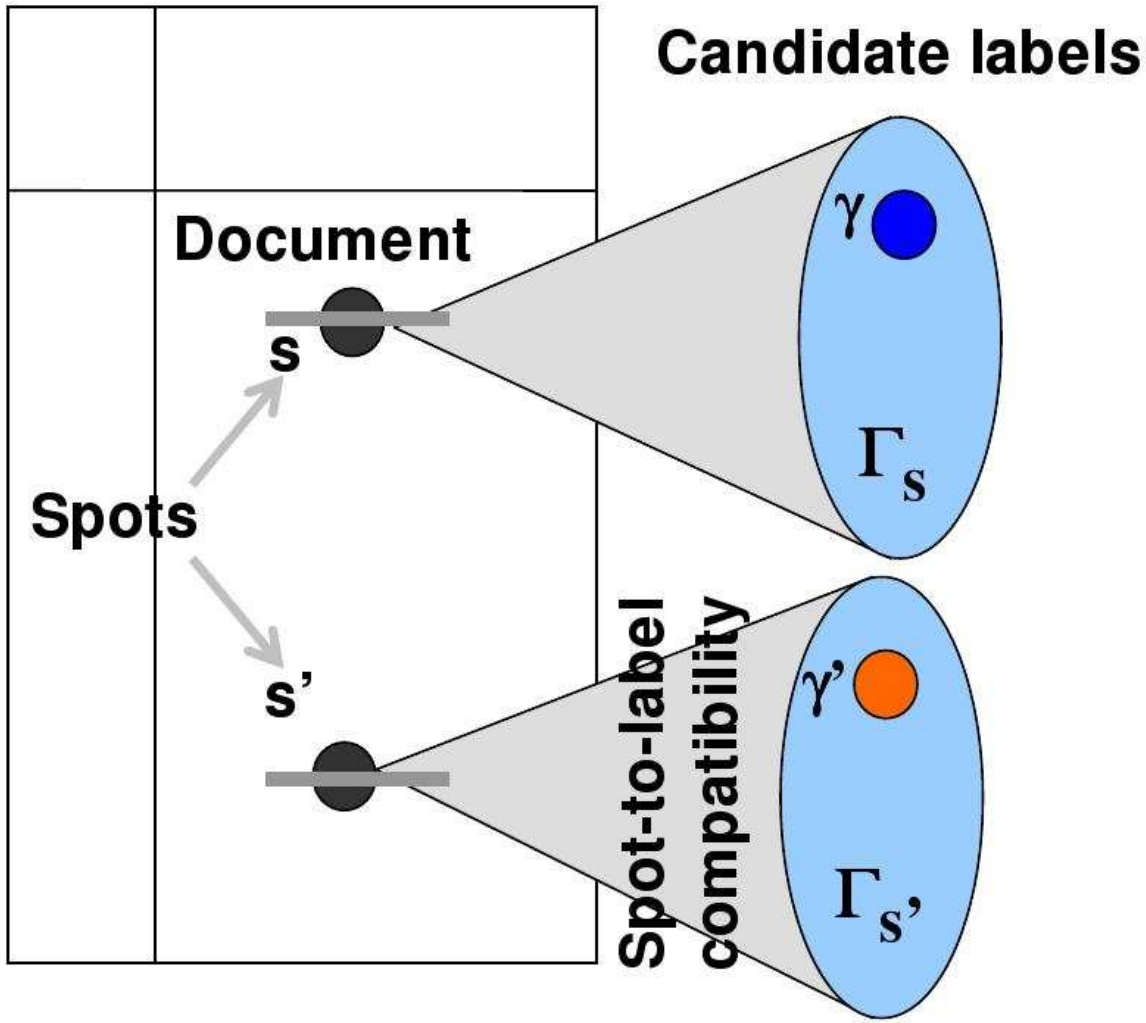
Local context signal

Jacksonville Jaguars

Jaguar (Car)

Jaguar (Animal)

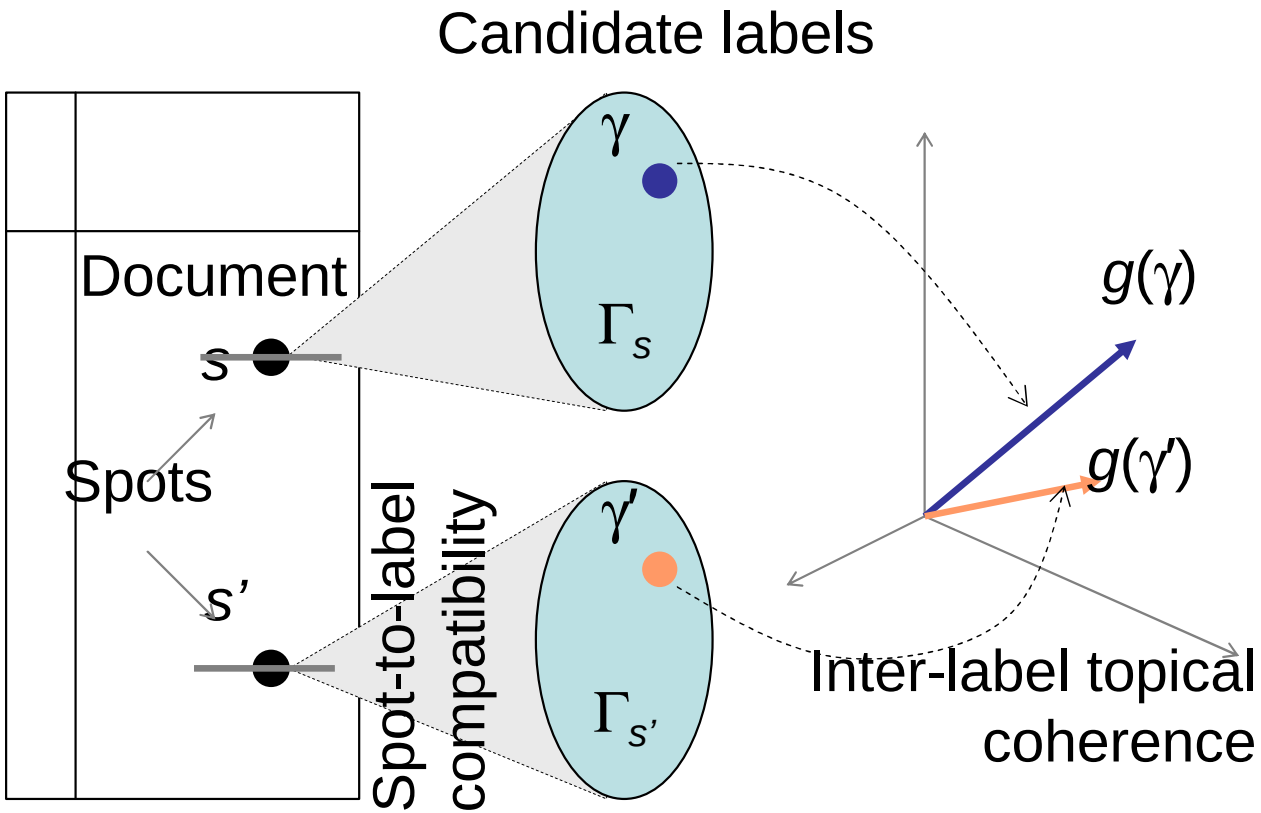
On first getting into the 2009 **Jaguar** XF, it seems like the ultimate in **automotive tech**. A red **backlight** on the **engine** start **button pulses** with a **heartbeat** cadence.



Exploiting collective info

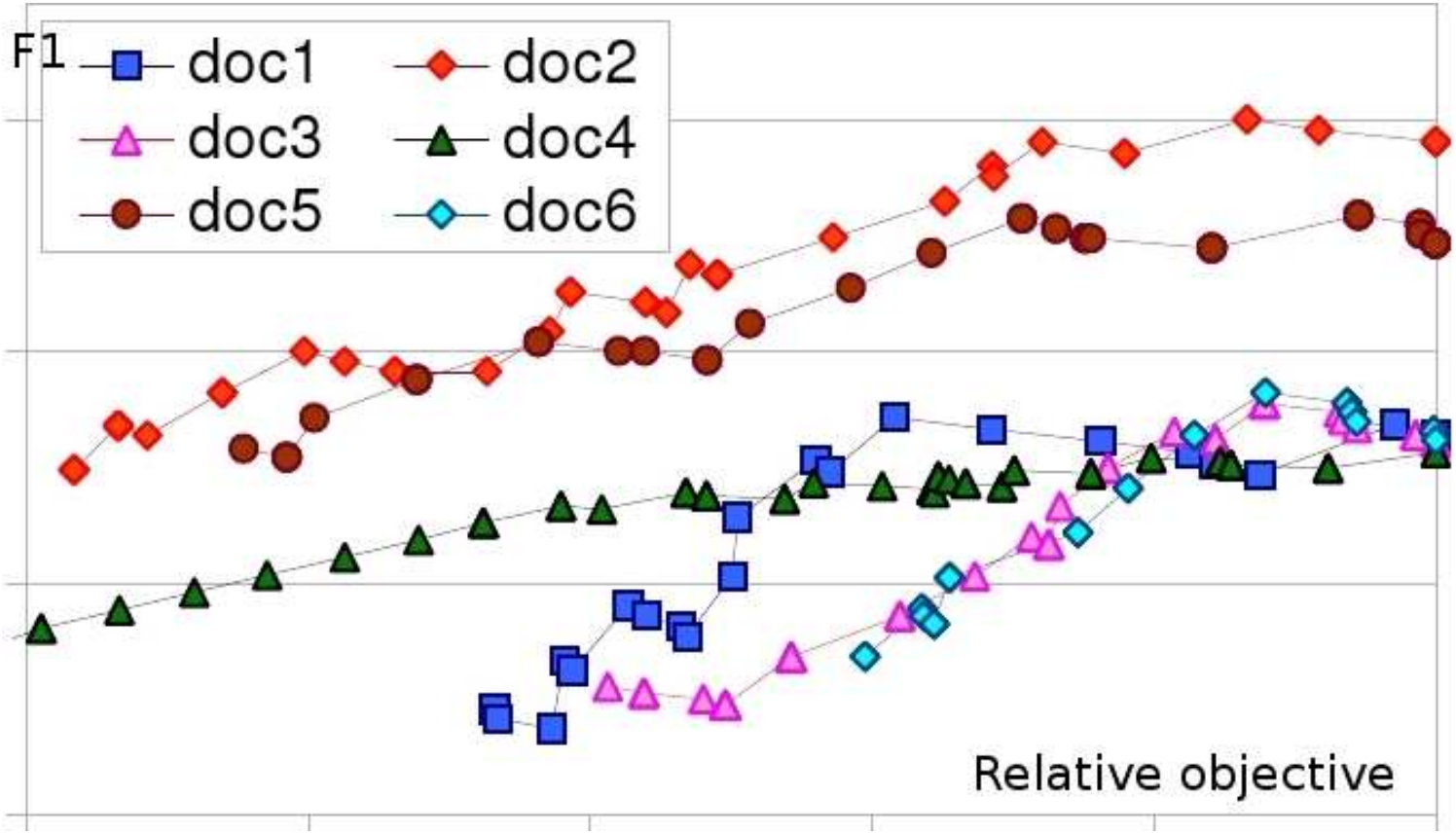
- Let $y_s \in \Gamma_s \cup$ NA be the variable representing entity label for spot s
- Pick all y_s together optimizing global objective

Collective formulation



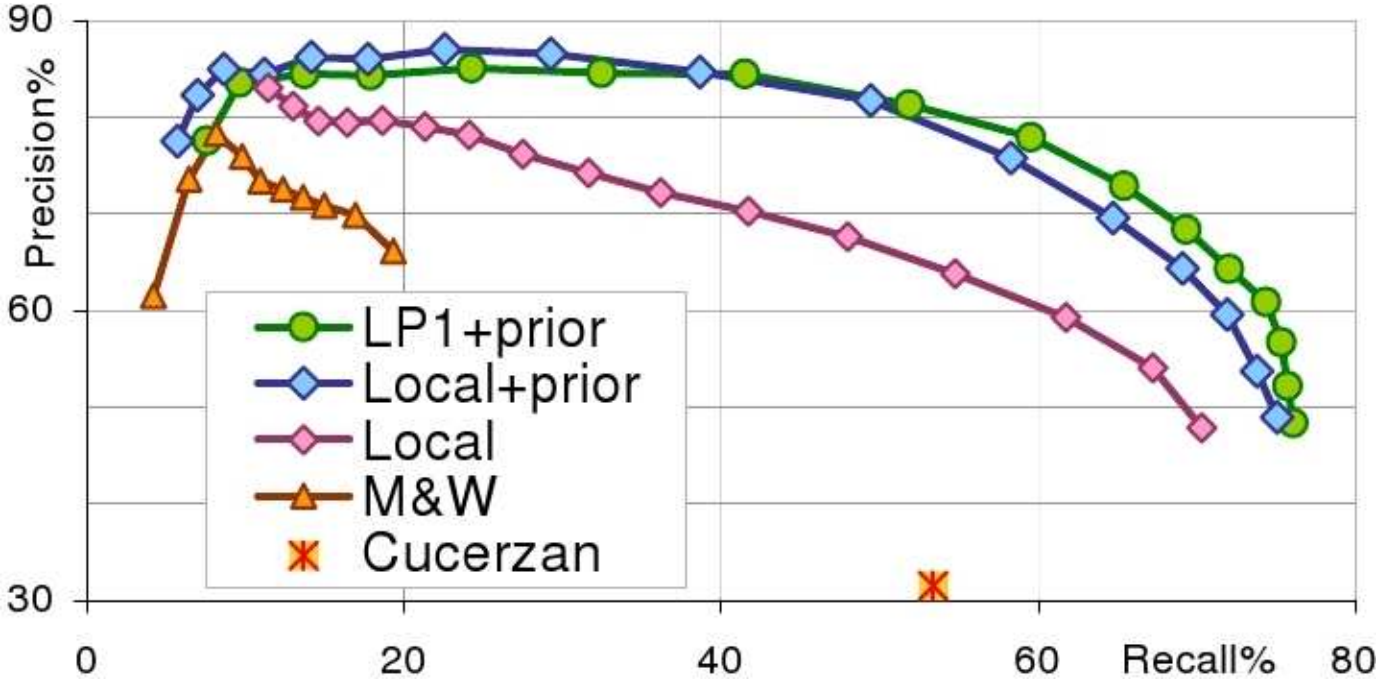
- Embed entities as vector $g(\gamma)$ in feature space
- Maximize local compatibility + global coherence

Collective model validation



- Local hill-climbing to improve collective obj
- Get F1 accuracy using ground truth annotations
- Very high positive correlation

Collective accuracy

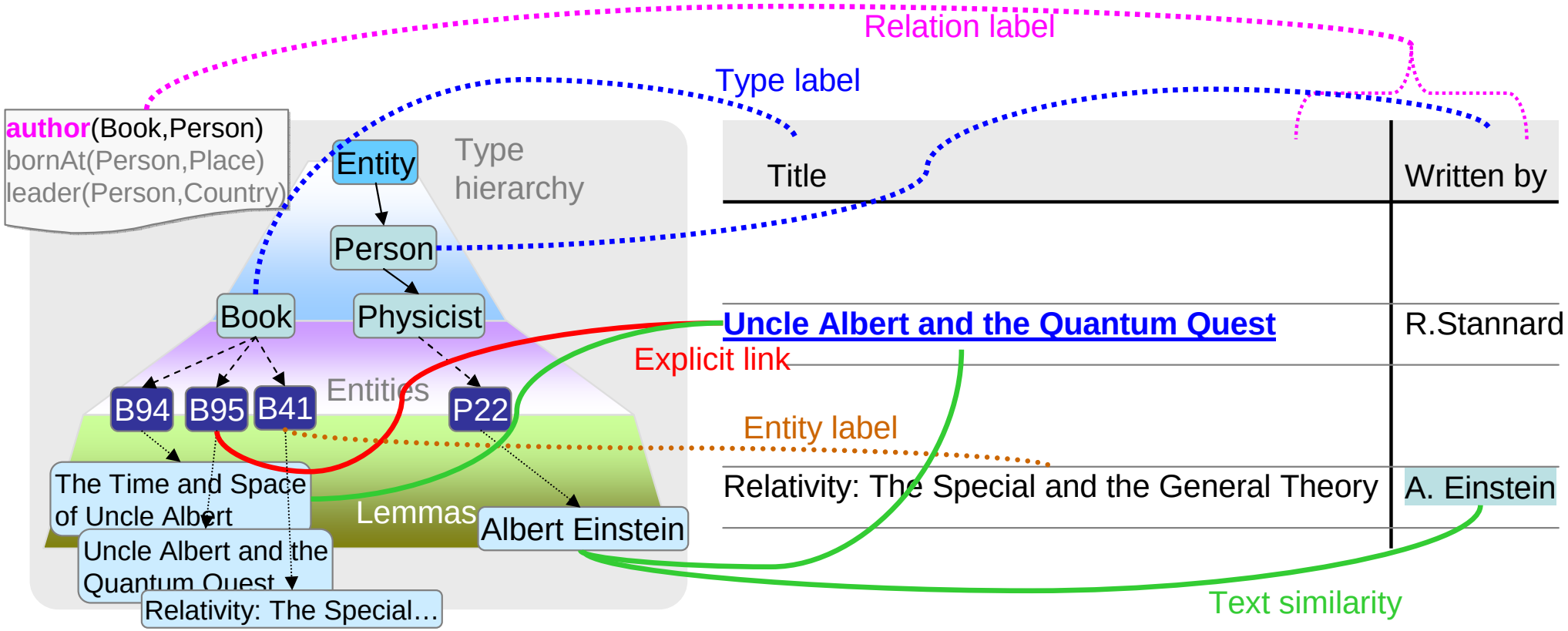


- ~20,000 spots manually labeled in Web docs
- Local=training *w*
- Prior=bias objective using Wikipedia distribution
- LP1=relaxing collective integer program

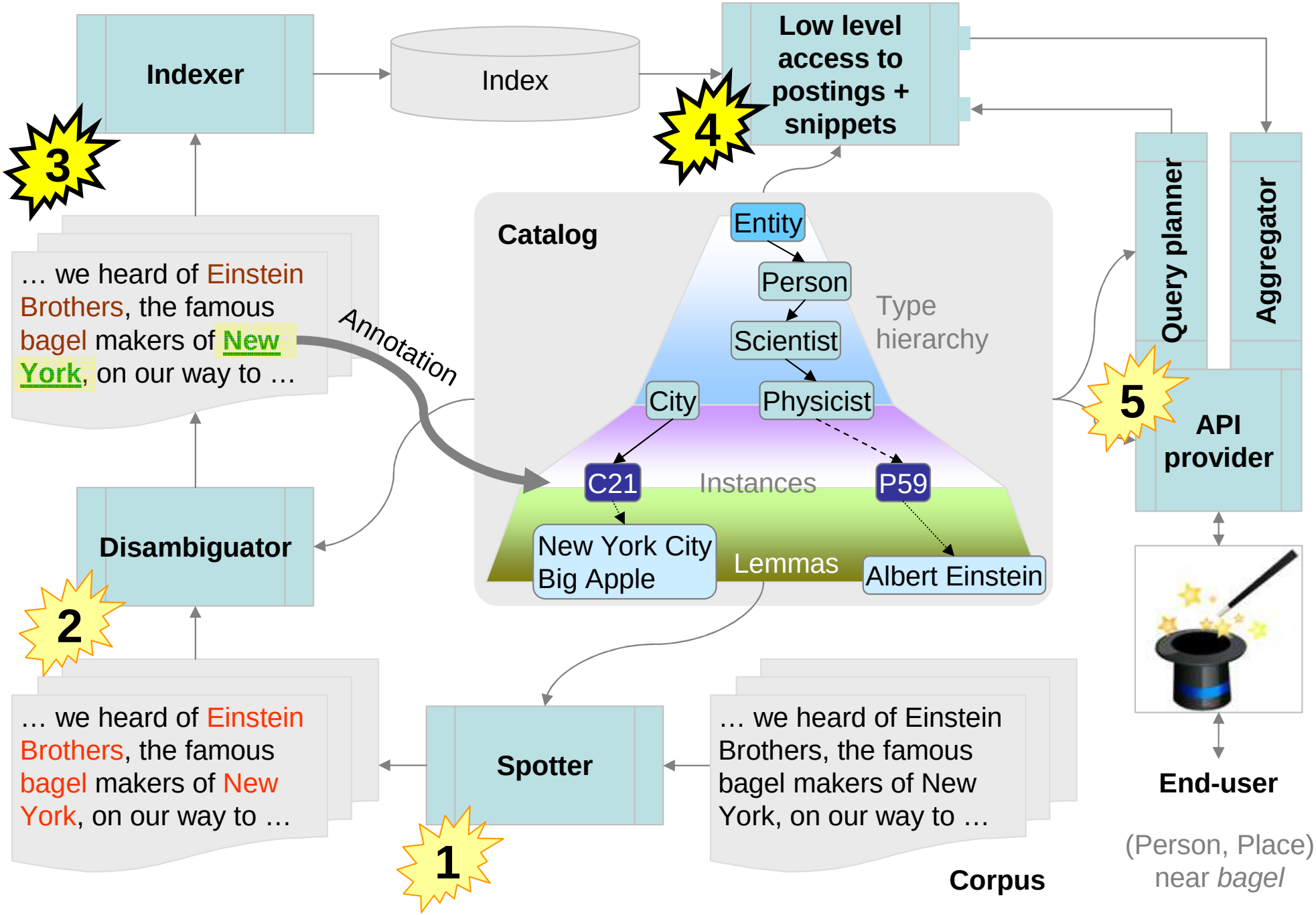
Loose ends

- Learn not only *w* but embedding $g(\gamma)$ and similarity between entity pairs
  - Applying the model should remain fast
- CPU cost of spotting + disambiguation compared to basic indexing
  - Page/site features to prune candidates?
- Training and evaluation at Web scale
  - Active learning? Exploit social tagging?
- Violation of coherent entity set assumption

Harnessing the power of tables



- At least 100 million richly relational tables …
- … without any explicit schema


Indexer
Index
Low level access to postings + snippets
3
4
Catalog
Entity
Person
Scientist
Physicist
City
Type hierarchy
C21
Instances
P59
New York City
Big Apple
Lemmas
Albert Einstein
… we heard of Einstein Brothers, the famous bagel makers of New York, on our way to …
Annotation
Disambiguator
2
Query planner
Aggregator
5
API provider
End-user
(Person, Place) near bagel
Spotter
1
… we heard of Einstein Brothers, the famous bagel makers of New York, on our way to …
Corpus

InContext subqueries

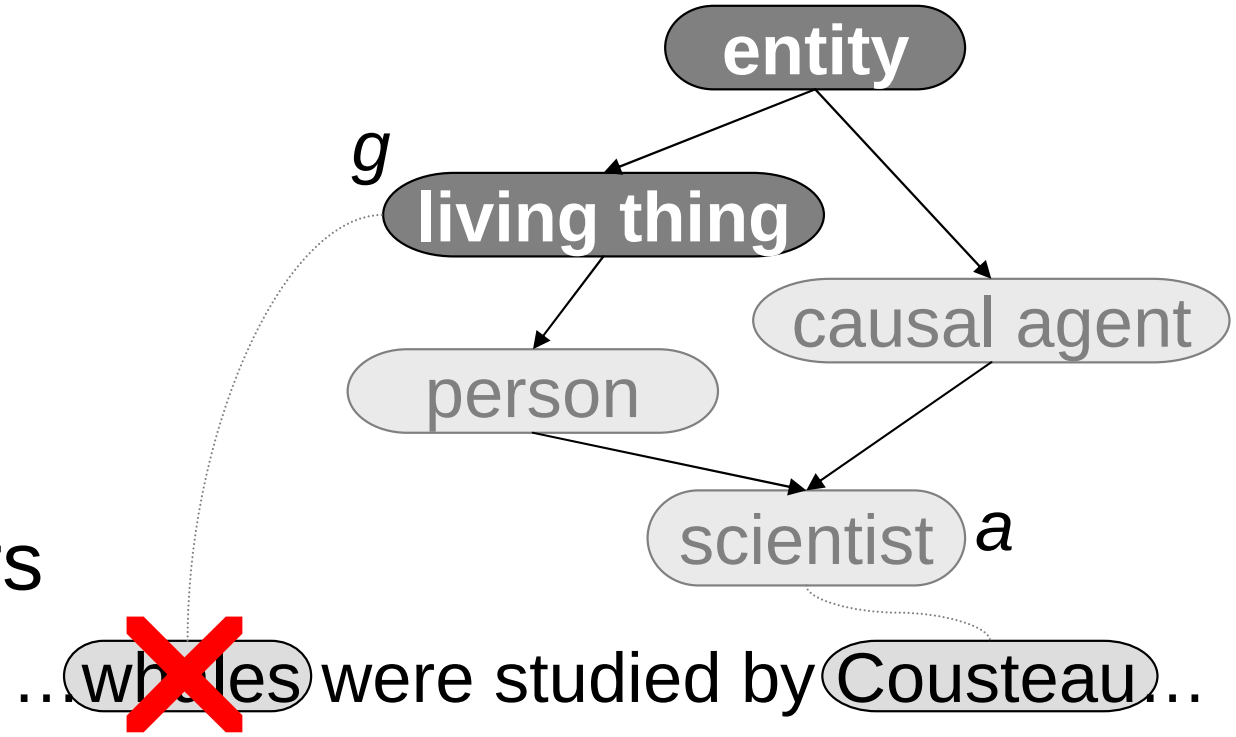
- Scientist who studied whales
  - ?s ∈ Category:Scientist
  - ?s ∈ Category:MarineBiologist
  - InContext(?c, ?s, study studied whale whales)
- Query expansion
  - Did Einstein, Bohr, Rutherford...study whales?
  - WordNet knows 650 scientiest, 860 cities
  - Wikipedia?
  - Impractical query times

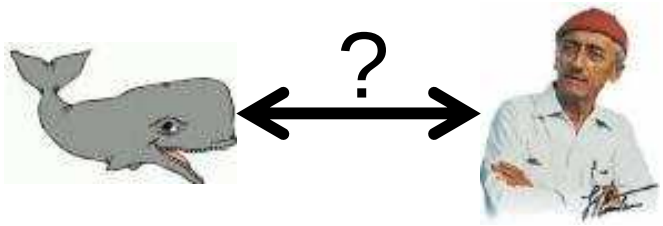
Indexing for InContext queries

- Index expansion
  - Costeau→scientist→person→organism→ living_thing→…→entity
  - Pretend all these tokens appear wherever Cousteau does, and index these
- Works ok for small type sets (5—10 broad types), but
  - WordNet: 15k internal, 80k total noun types
  - Wikipedia: 250k categories
- Index size explosion unacceptable

Pre-generalize

- Index a subset $R \subset A$
- Query type $a \notin R$
- Want k answers
- Probe index with g, ask for $k' > k$ answers





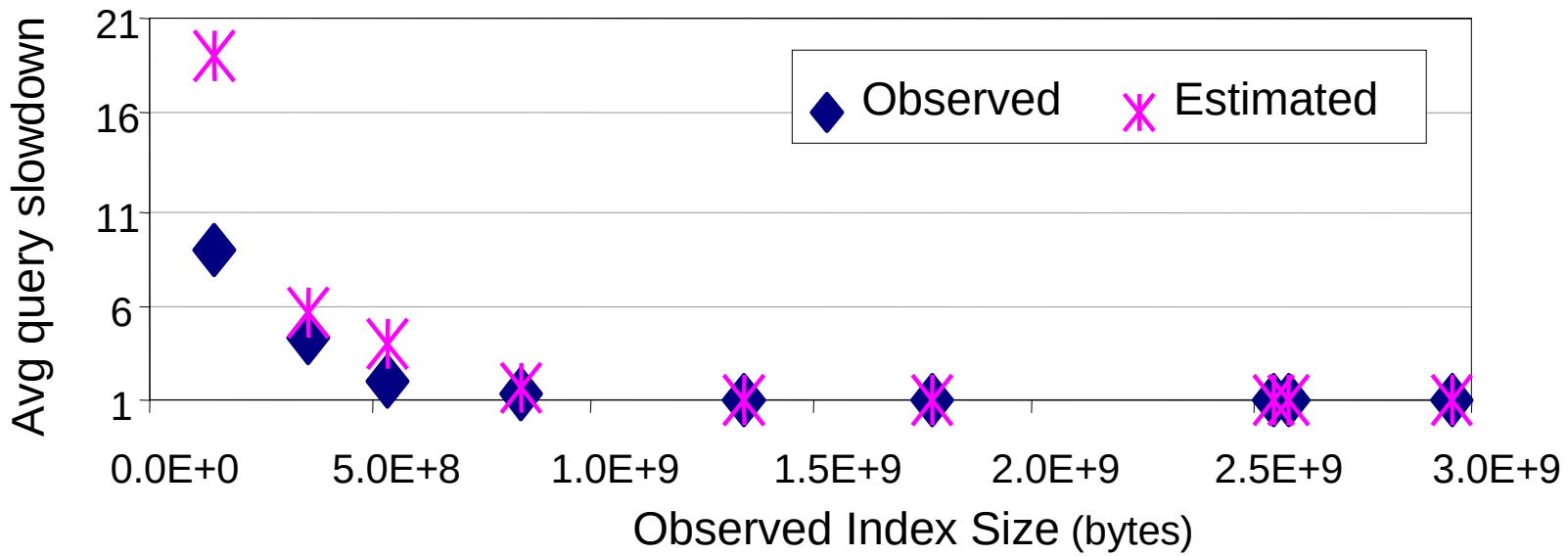
Post-filter

- Fetch k' high-scoring (mentions of) entities $w \in {}^{+}g$
- Check if $w \in {}^{+}a$ as well (using forward and reachability index); if not, discard
- If $< k$ survive, restart with larger k (expensive!)

Cost-benefit considerations

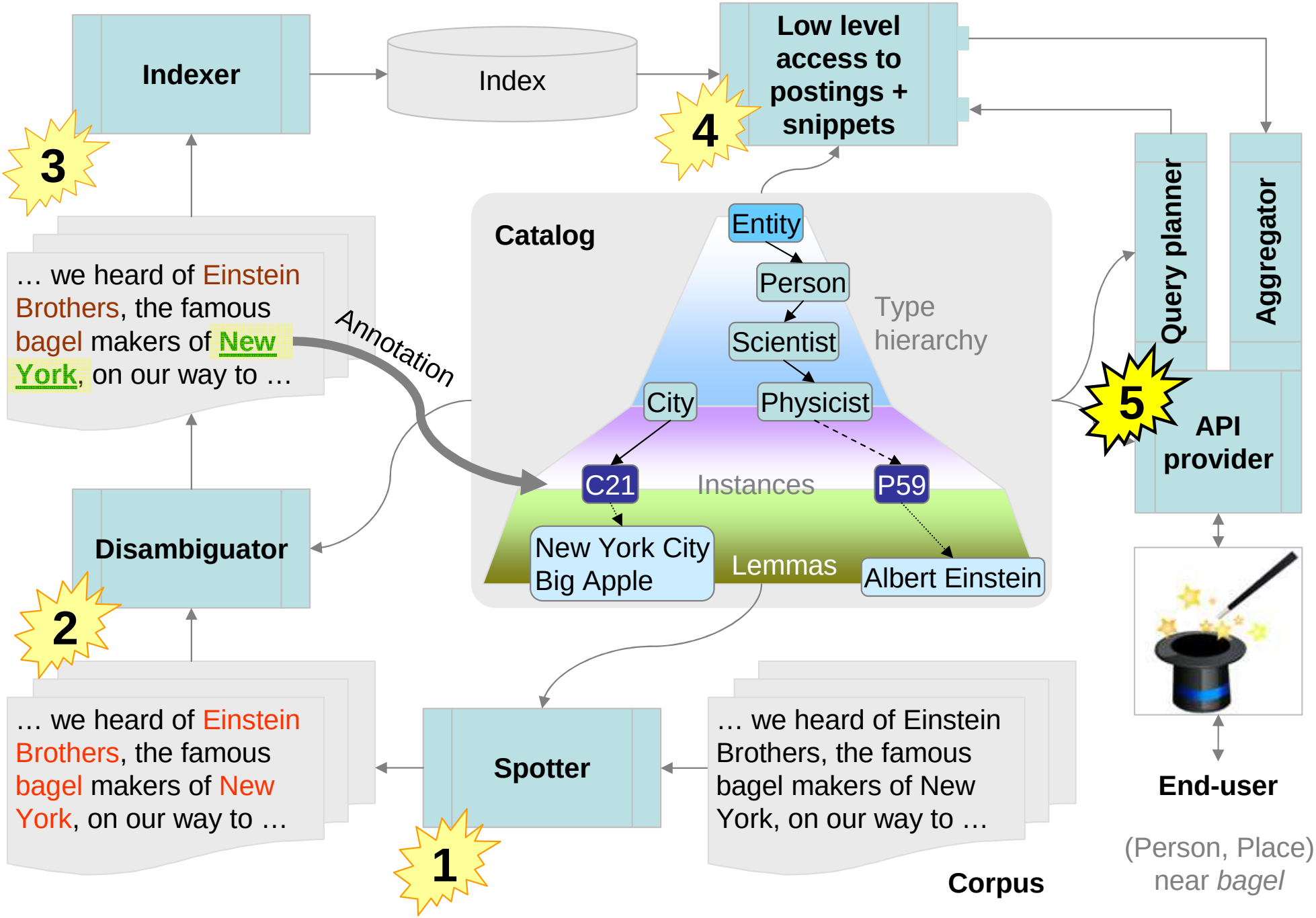
- How much space saved by indexing *R* instead of the whole of *A*?
  - Cannot afford to try out many *R*s, need quick estimate
- What is the average query slowdown owing to $a \rightarrow g$ pre-generalize and post-filter?
  - Depends on query workload
  - Cannot afford to test on too many queries

Index size vs. query slowdown

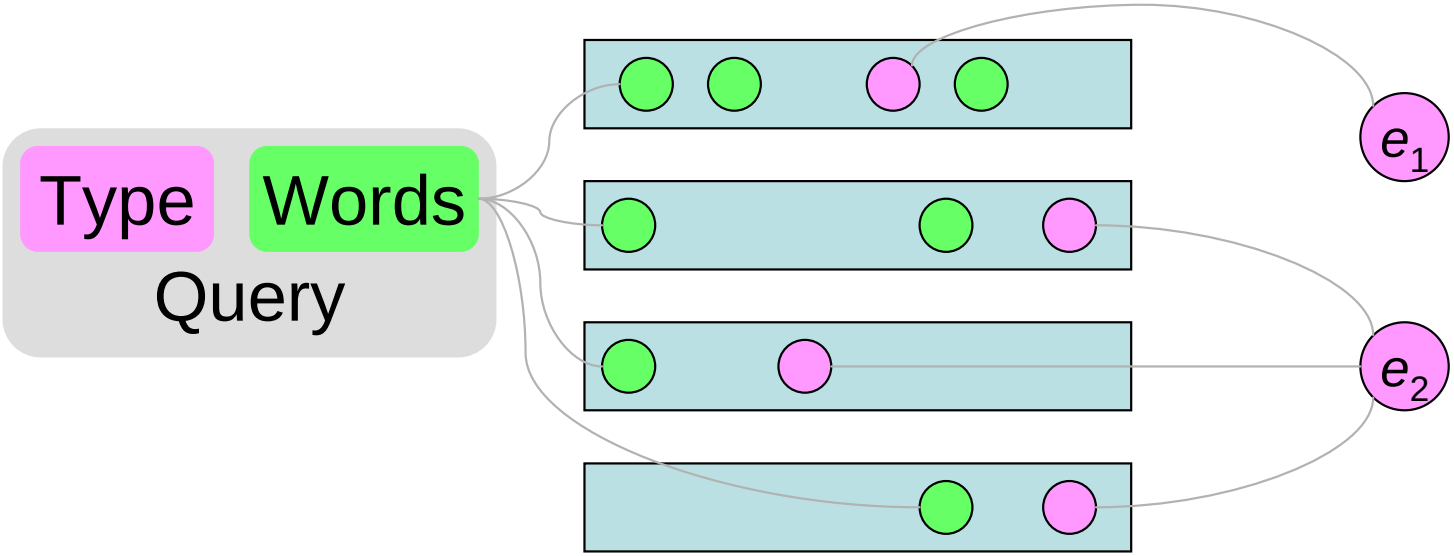


- Annotated TREC corpus
- Space = 520MB < inverted index = 910MB
- Query slowdown ≈ 1.8
- From TREC to Web?

Corpus/Index	Gbytes
Original corpus	5.72
Gzipped corpus	1.33
Stem index	0.91
Full type index	4.30
Reachability index	0.01
Forward index	1.16
Atype subset index	0.52


Indexer
Index
Low level access to postings + snippets
3
4
Catalog
Entity
Person
Scientist
Physicist
City
Type hierarchy
C21
Instances
P59
New York City
Big Apple
Lemmas
Albert Einstein
Query planner
Aggregator
API provider
5
… we heard of Einstein Brothers, the famous bagel makers of New York, on our way to …
Annotation
Disambiguator
2
Spotter
1
… we heard of Einstein Brothers, the famous bagel makers of New York, on our way to …
Corpus
End-user
(Person, Place) near bagel

How to score and aggregate

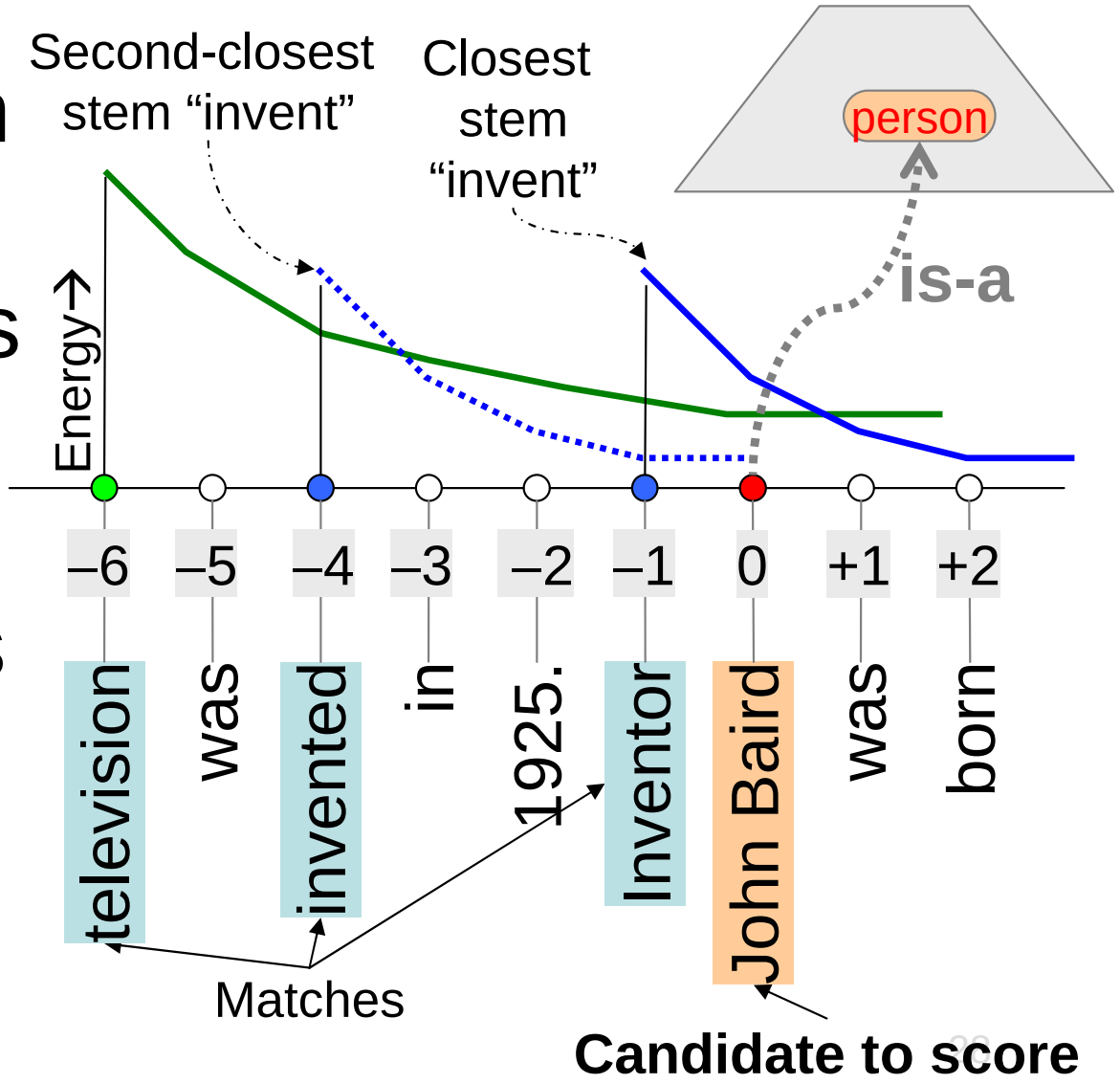


- Literals in query match tokens in context
- Context is a candidate because it mentions an entity of the target type
- What is the score of a context?
- How should context scores be aggregated into entity evidence?

Scoring a context

- Rarity of matches
- Distance from candidate position to matches
- Many occurrences of one match
  - Closest is good
- Combining scores from many selectors
  - Sum is good

InContext(?c, ?p, +invent* +television), ?p ∈⁺ Person, Aggregate(?c)



Learning to rank

- Fix a query, collect good+bad contexts
- Contexts turned into feature vectors z_i
- Fit model w (across all queries)
- Sort by decreasing context score $w^T z_i$
- Want each good to beat all bad
  - $w^T z_g > w^T z_b$
- Flipping #1 and #10 far more serious than flipping #40 and #41
- Discontinuous, non-smooth ranking loss

Laplacian scoring

- Represent snippet using feature vector z_i
- Local score of snippet is $w^T z_i$
- Affinity a_{ij} between (mentions in) snippets
  - "Andrew McCallum" vs. "A. K. McCallum"
  - "18 feet", "19 ft", "3—4 meters"
- Global score f_i

$$\min_{\{f_i\}} \sum_i \left(f_i - w^T z_i \right)^2 + C \sum_{i,j} a_{ij} \left(f_i - f_j \right)^2$$

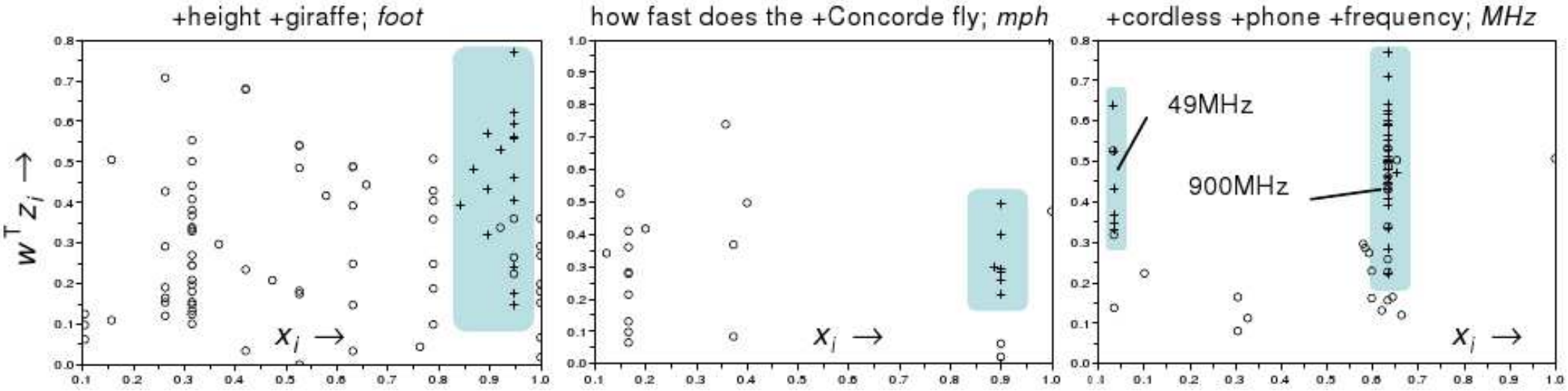
- During training fit w using partial order on f

Local scores unreliable

`+giraffe, +height; foot`	
	La Giraffe was small (approx. **11 feet** tall) because she was still young, a full grown giraffe can reach a height of **18 feet**.
	Giraffe Photography uses a telescopic mast to elevate an 8 megapixel digital camera to a height of approximately **50 feet**.
	The record height for a Giraffe unicycle is about **100 ft** (30.5m).
`+weight, weigh, airbus, +A380; pound`	
	Since the Airbus A380 weighs approximately **1,300,000 pounds** when fully loaded with passengers ...
	The new mega-liner A380 needs the enormous thrust of four times **70,000 pounds** in order to take off.
	According to Teal, the **319-ton** A380 would weigh in at **1,153 pounds** per passenger
`far +raccoon relocate; mile`	
	It also says – unnervingly – that relocated raccoons have been known to return from as far away as **75 miles**.
	Sixteen deer, 2 foxes, one skunk, and 2 raccoons are sighted during one **35 mile** drive.
	One study found that raccoons could move over **20 miles** from the drop-off point in a short period of time.

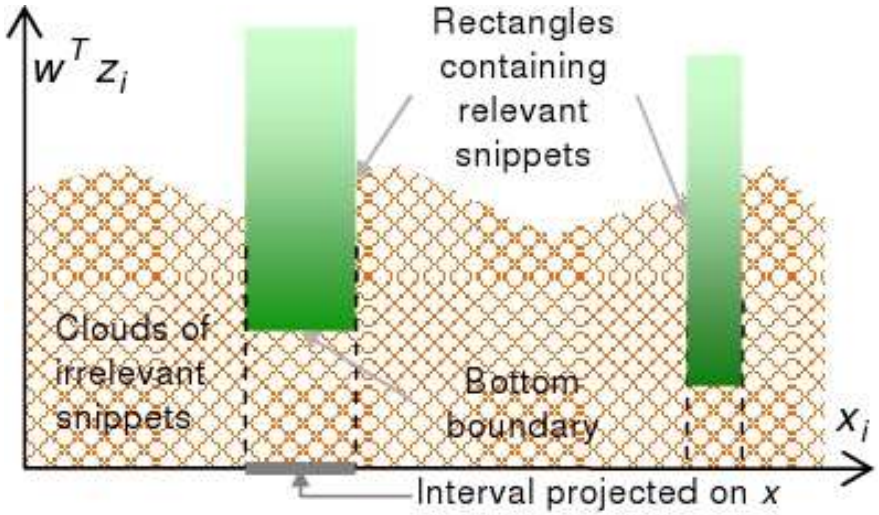
- Confounding candidates with correct units/type
- Can aggregation over snippets help
- Avoid deep NLP?
- Here we focus on quantity answers

Snippet score-quantity scatter



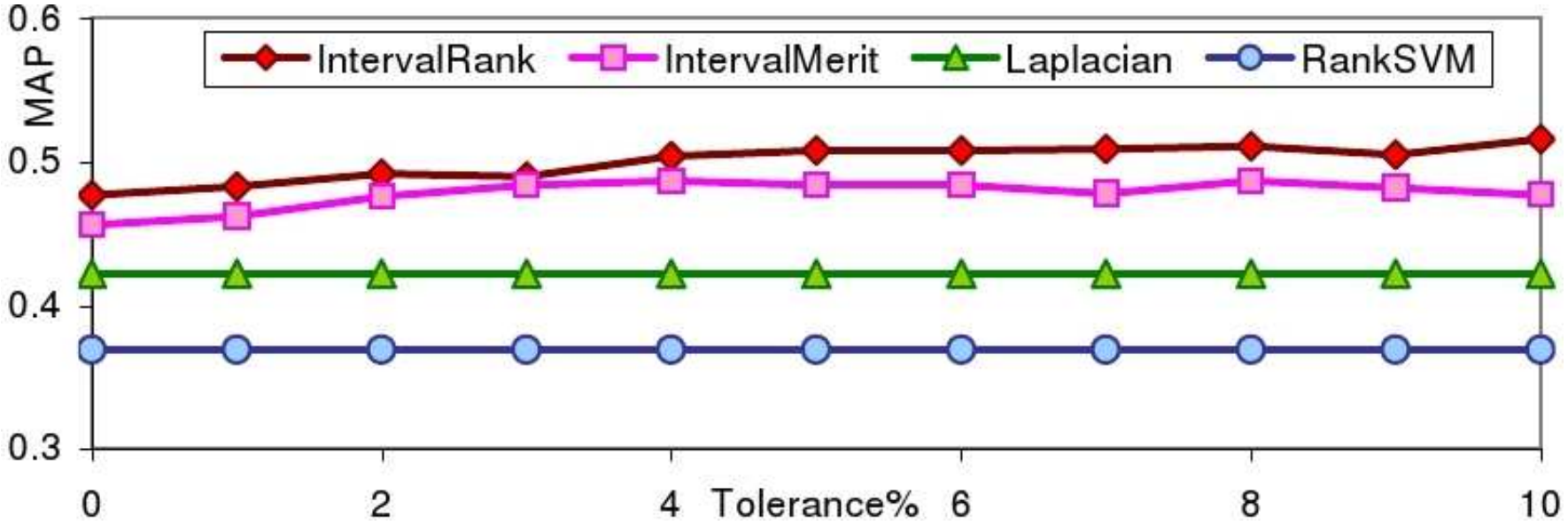
- Both axes scaled to [0, 1] for clarity
- Relevant/good snippets = +, irrelevant/bad = ◦
- Ideal ranking $\Rightarrow$ horizontal line separating + from ◦ … no ideal ranking found in experiments
- Rectangles densely packed with many +, few ◦
  - Possibly > 1 rectangles for some queries

Consensus rectangles



- Relevant rectangle/s in sea of irrelevant snippets
- Many low-scoring relevant snippets
- How to detect and rank consensus rectangles?
- Position and shape varies across queries
  - Cannot use standard nonlinear discriminants

Interval-hunting



- RankSVM: Independent snippet comparison
- IntervalMerit
  - Scan for all interval narrower than 1:(1+tolerace/100)
  - Compare snippets inside interval to those outside
- IntervalRank: Exploit collective features

Summary

- How to open up new info pathways across docs and semistructured knowledge bases
- Propose new access methods into type-entity-snippet composite data model
- 1B docs, 336 core, 1GB/core, 126TB disk
  - Machine learning for annotation
  - Indexing massive out-of-core graphs
  - Learning to rank and aggregate
- Search unstructured Web using structured queries