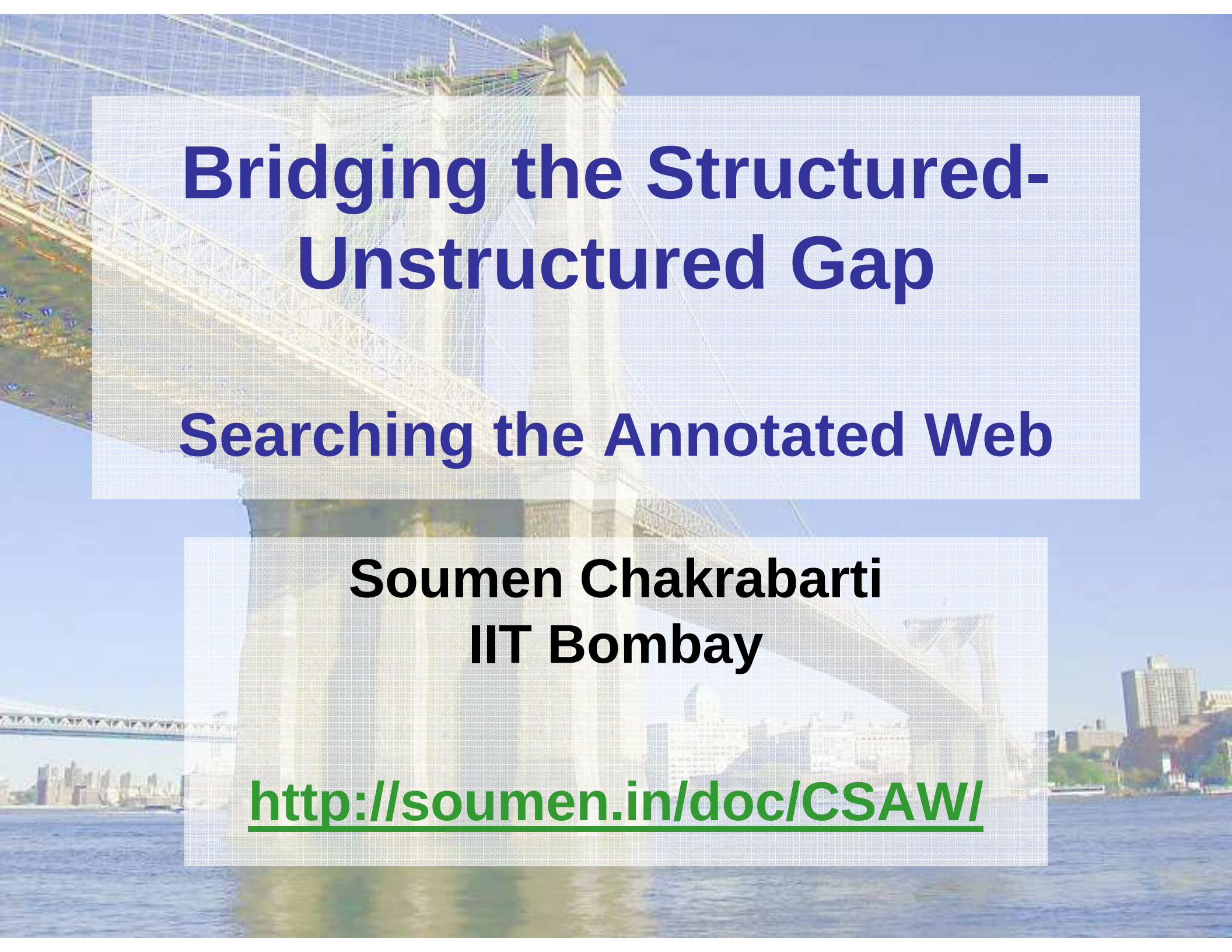


Bridging the Structured-Unstructured Gap

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 Brooklyn Bridge
- However, there remain information needs where cognitive burden is still very large

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
- Is the number of Oscars won directly related to production budget?
- How many justices serve in the International Criminal Court?
- Exxon Valdez cleanup cost
- How many papers submitted to SIGMOD?

Why difficult?

- Search engines provide excellent “low-level access methods to pages”, but …
- 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

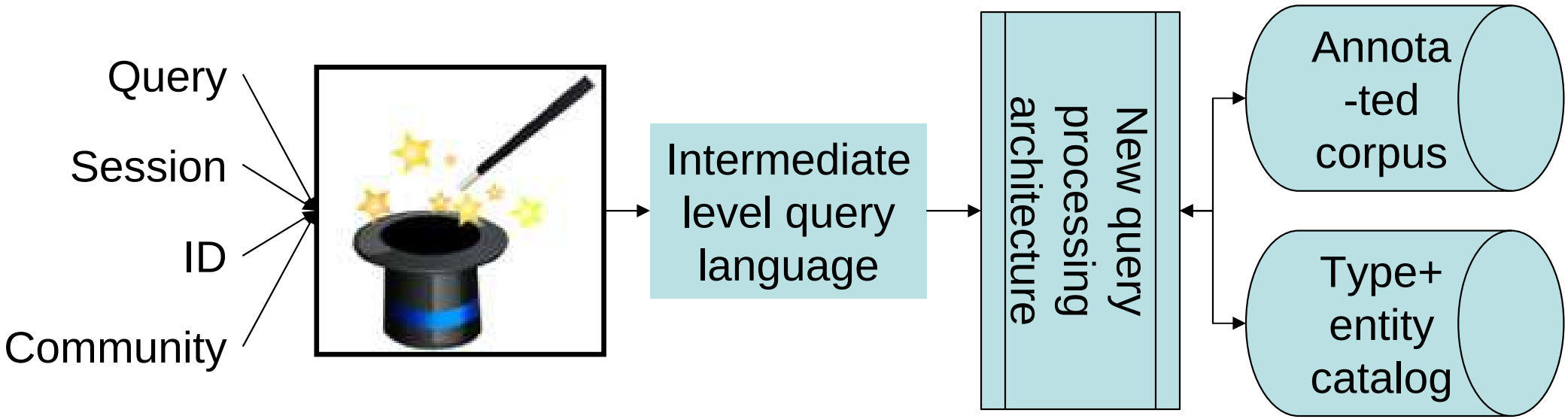
What if we could ask…

- ?f $\in^{+}$ Category:FrenchMovie
- ?a $\in$ QType:Number
- ?b $\in$ 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, ?b⟩ tuples

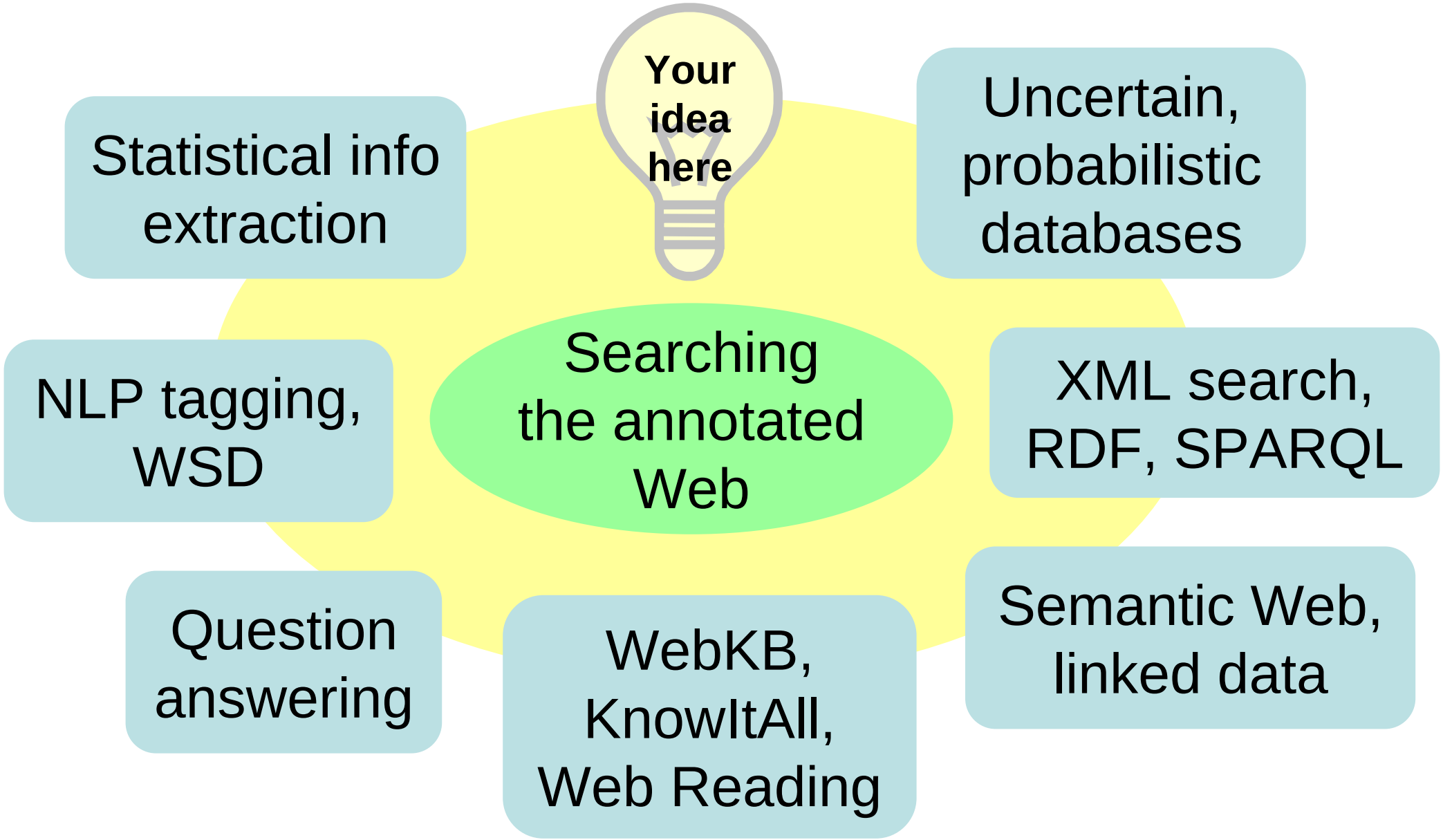
Disclaimers

- Esoteric
- Public domain
- May not work today
- Speculative, “what if”
- Ideas, prototypes

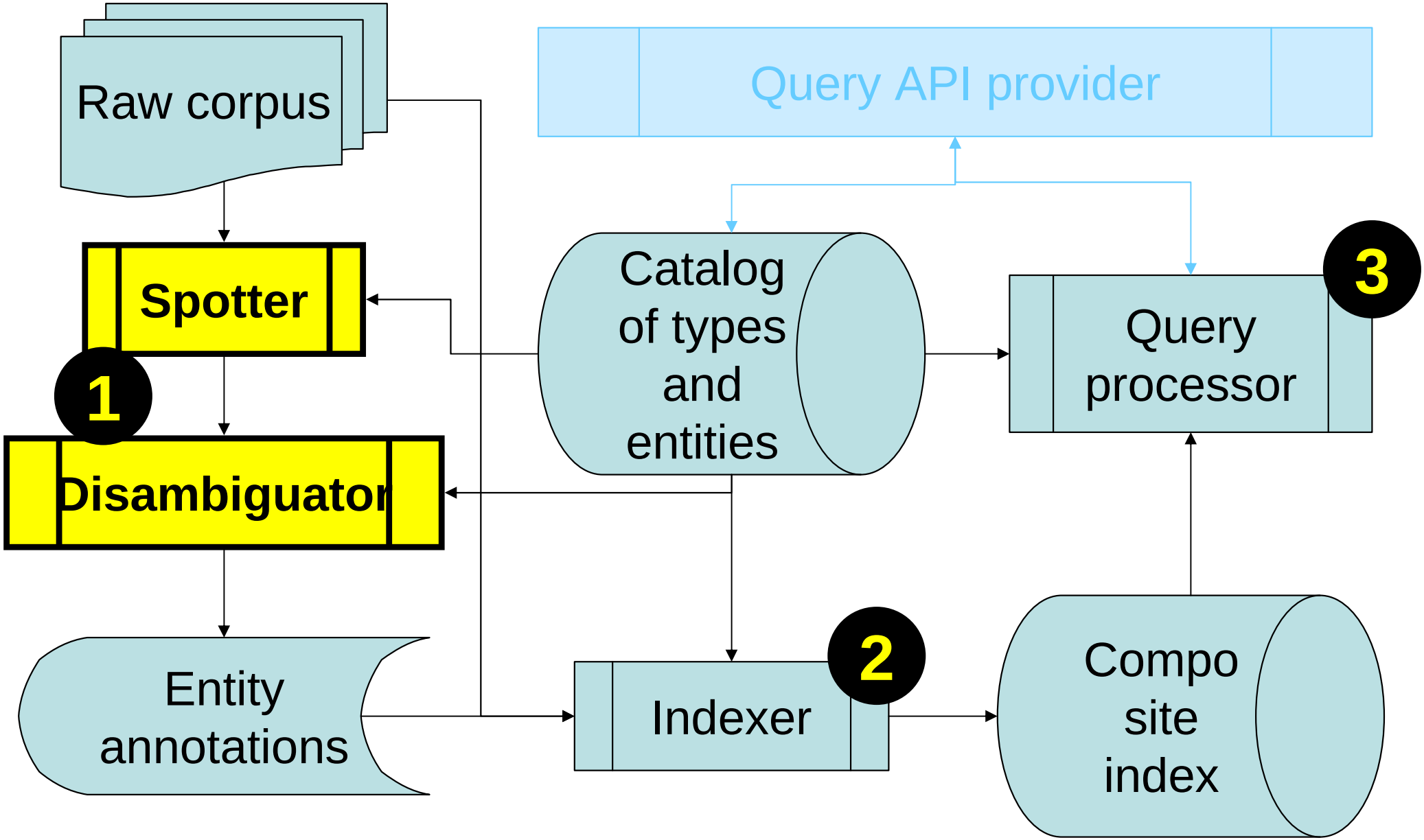
- Mainstream
- Proprietary
- Stable, practical
- Broad user base
- Traffic, revenue



Influences


Your idea here
Statistical info extraction
Uncertain, probabilistic databases
NLP tagging, WSD
Searching the annotated Web
XML search, RDF, SPARQL
Question answering
WebKB, KnowItAll, Web Reading
Semantic Web, linked data

Pieces to the puzzle

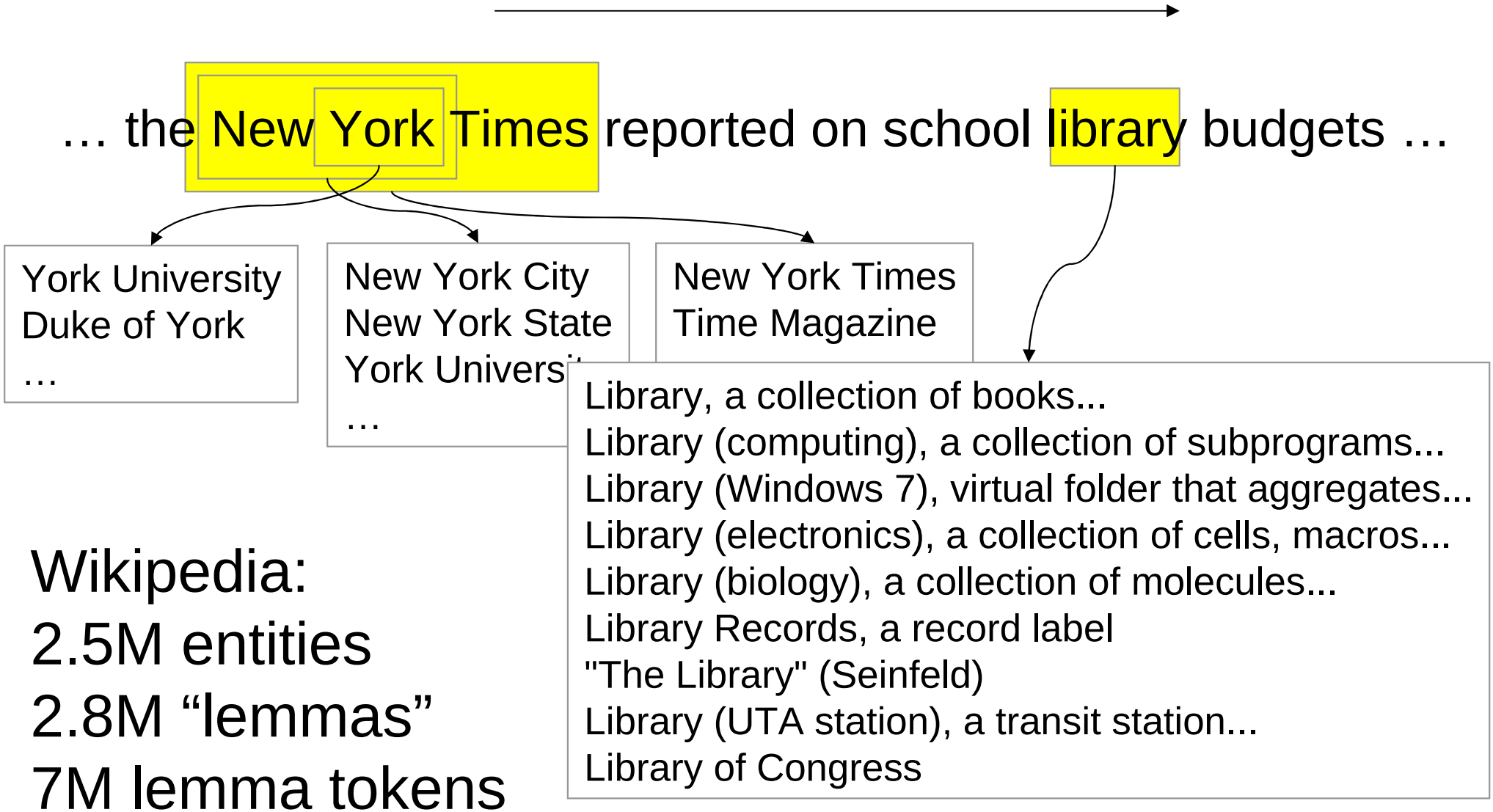


Mentions and spots

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
- Generalization of WSD in NLP
- SemTag/Seeker, Wikify!, Bunescu+Pasca, Cucerzan, Milne+Witten, [KSRC2009]

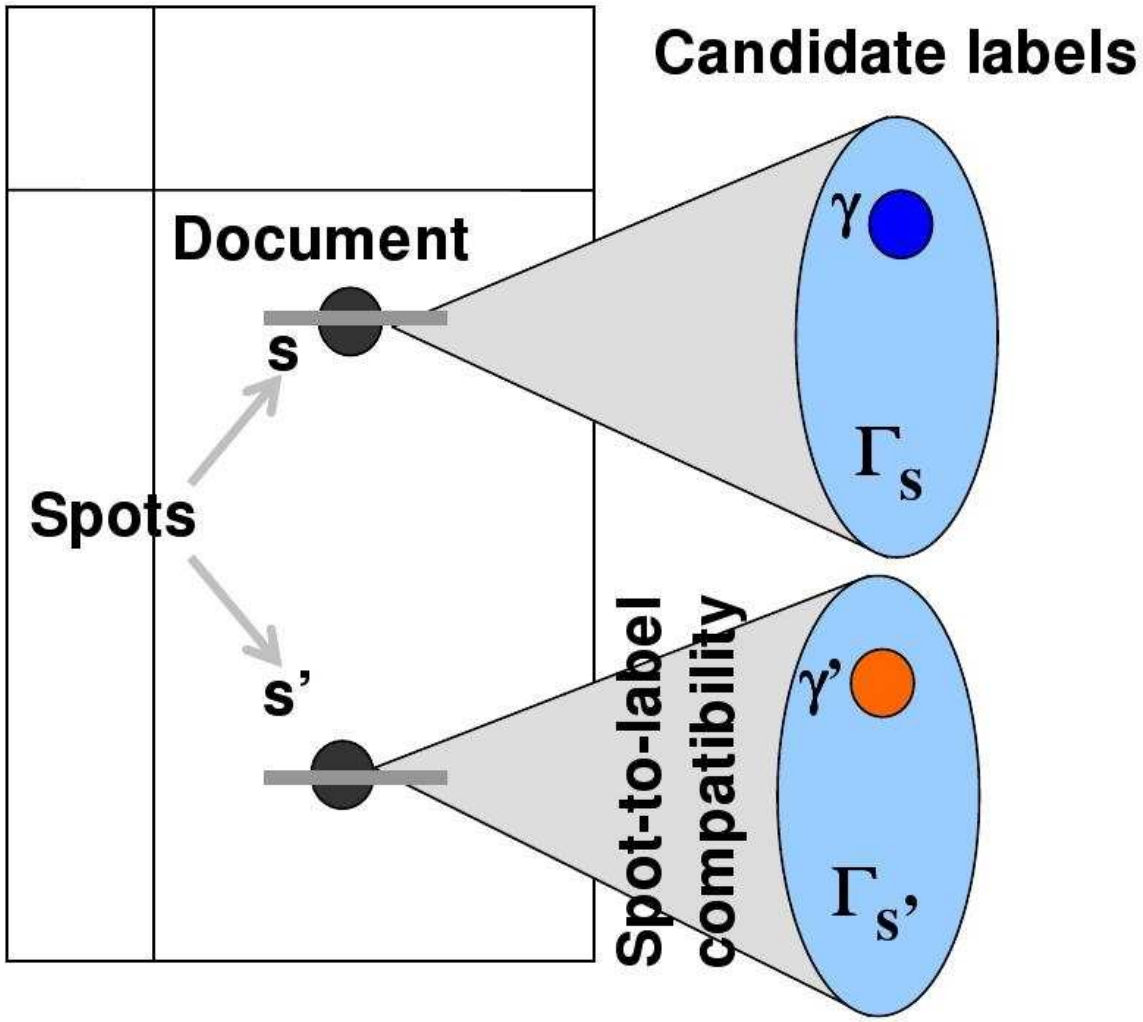
Local context signal

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.

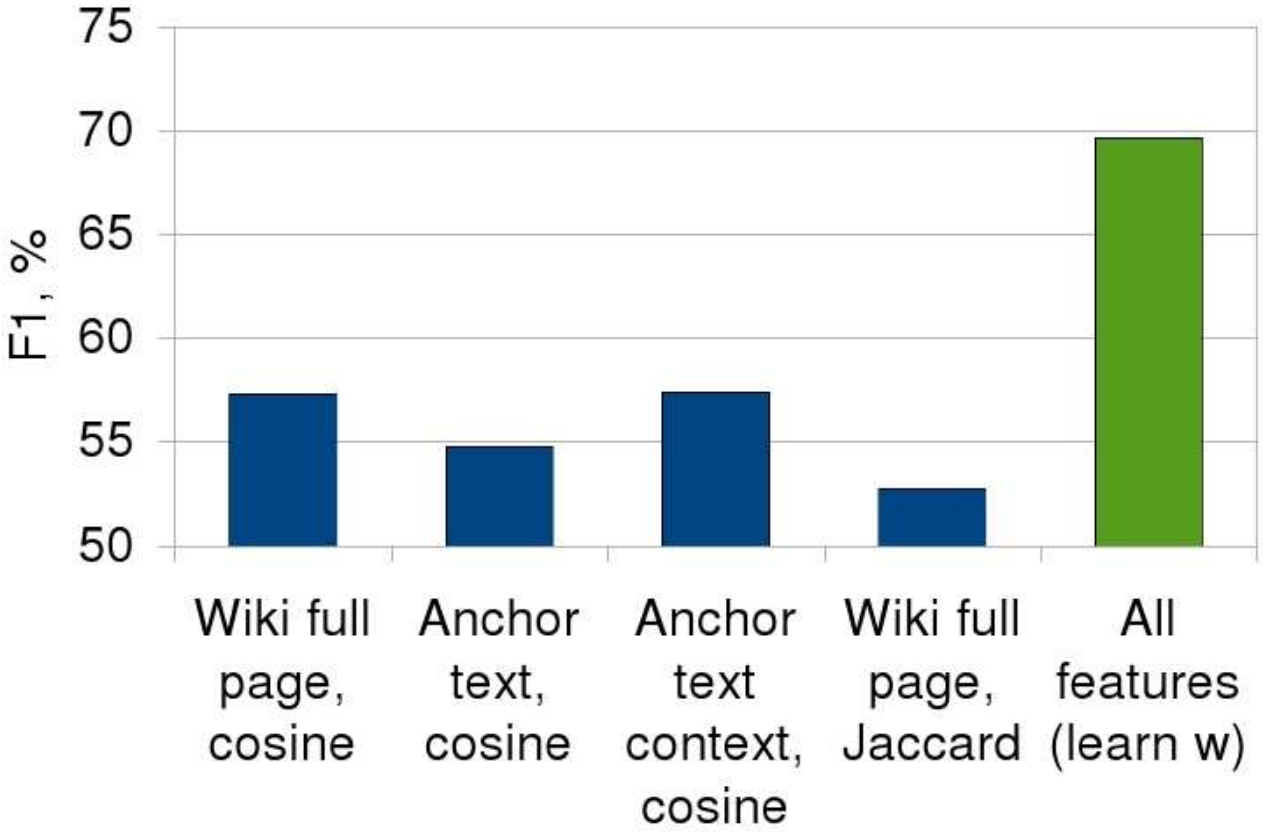


Encoding local compatibility

- $f_s(\gamma)$ is a vector of features
- Each feature is a function of *s* and metadata associated with γ
- Learn *w* from training data
- Choose

$$\arg\max_{\gamma} w^T f_s(\gamma)$$

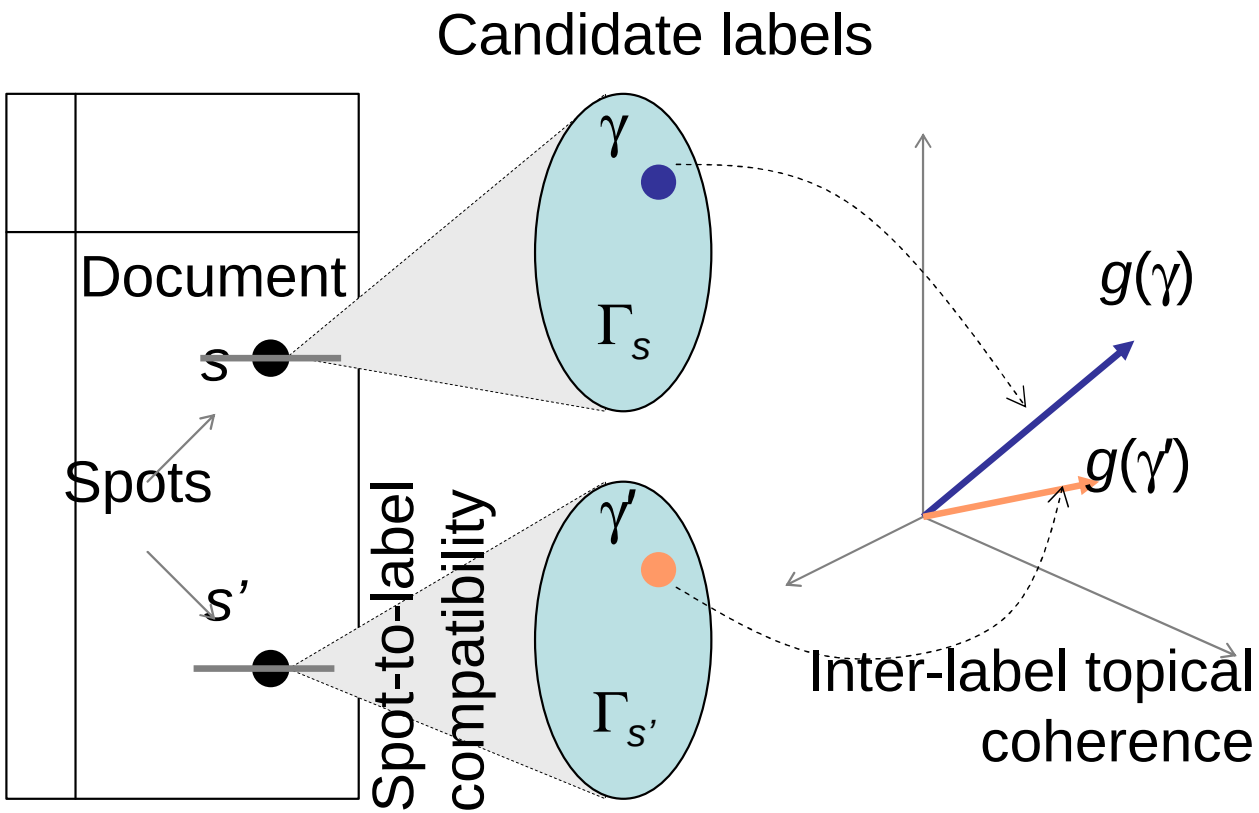
- Better than heuristics



Exploiting collective info

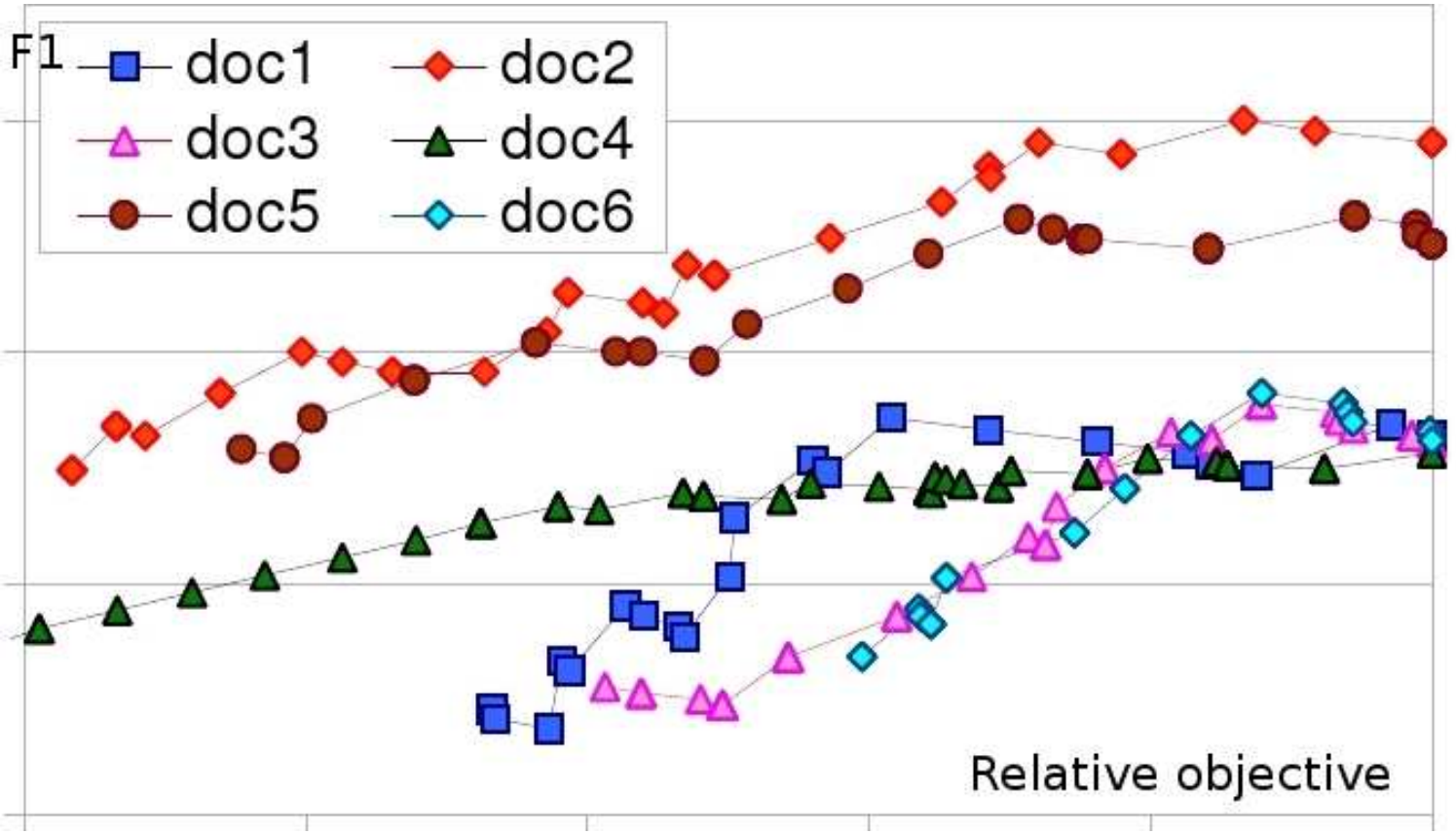
- Let $y_s \in \Gamma_s \cup$ NA be the variable representing entity label for spot *s*
- Pick all y_s together optimizing global objective

Collective formulation



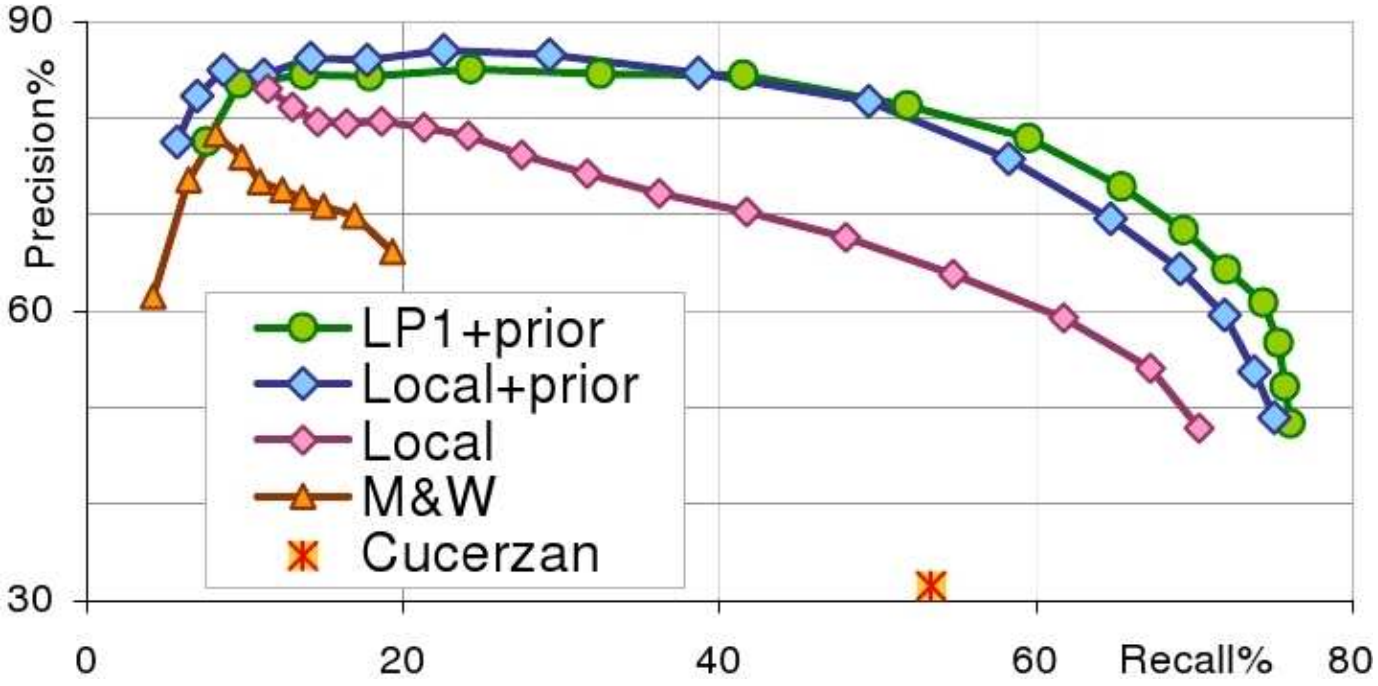
- Embed entities as vector $g(\gamma)$ in feature space
- Maximize local compatibility + global coherence

Collective model validation



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

Collective accuracy

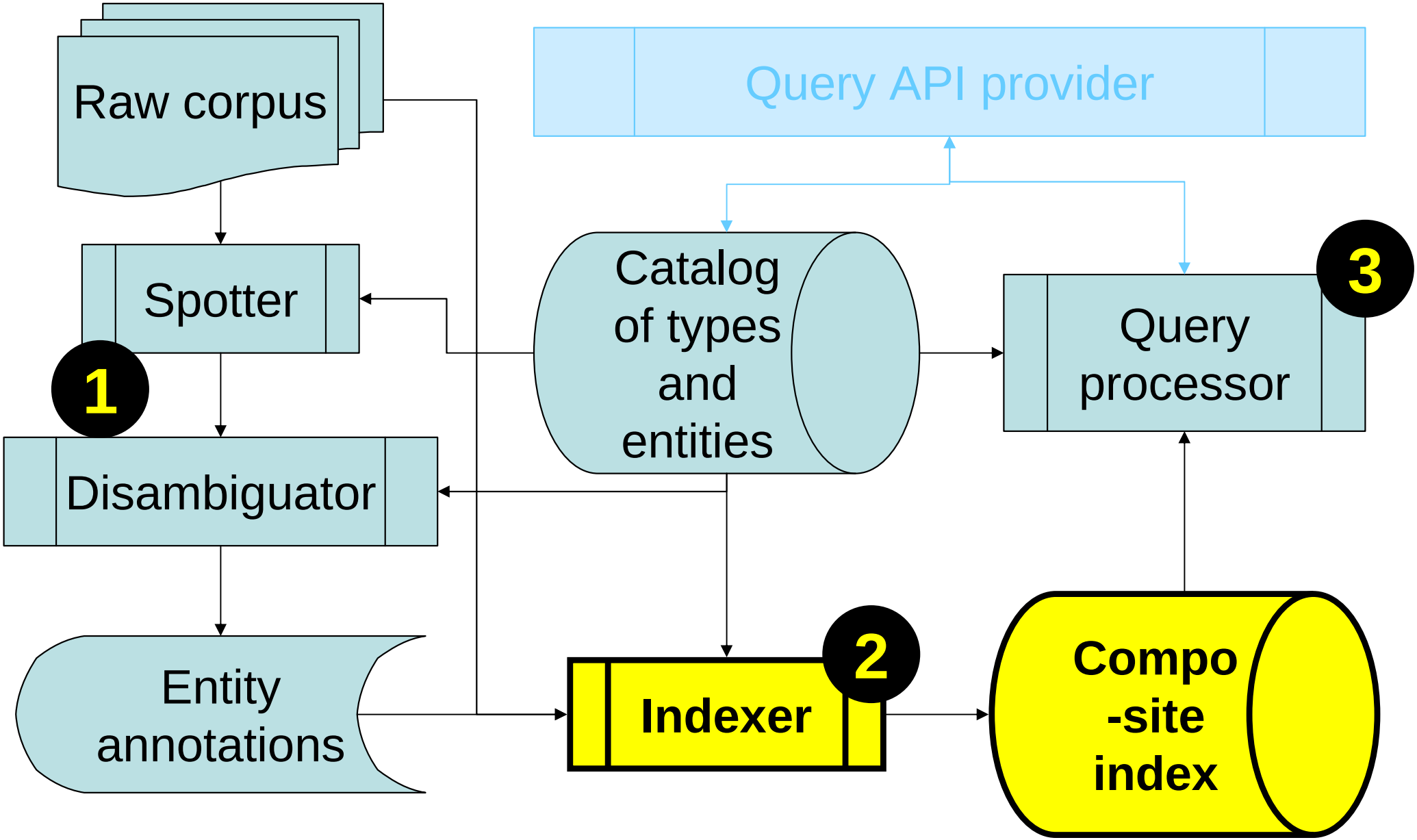


- ~20,000 spots manually labeled in Web docs
- Local=training *w*
- Prior=bias objective using Wikipedia distribution
- LP1=relaxing collective integer program

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
  - Use coarse page/site features to prune candidates?
- Training and evaluation at Web scale
  - Active learning framework
  - Exploit social tagging?

Pieces to the puzzle



InContext subqueries

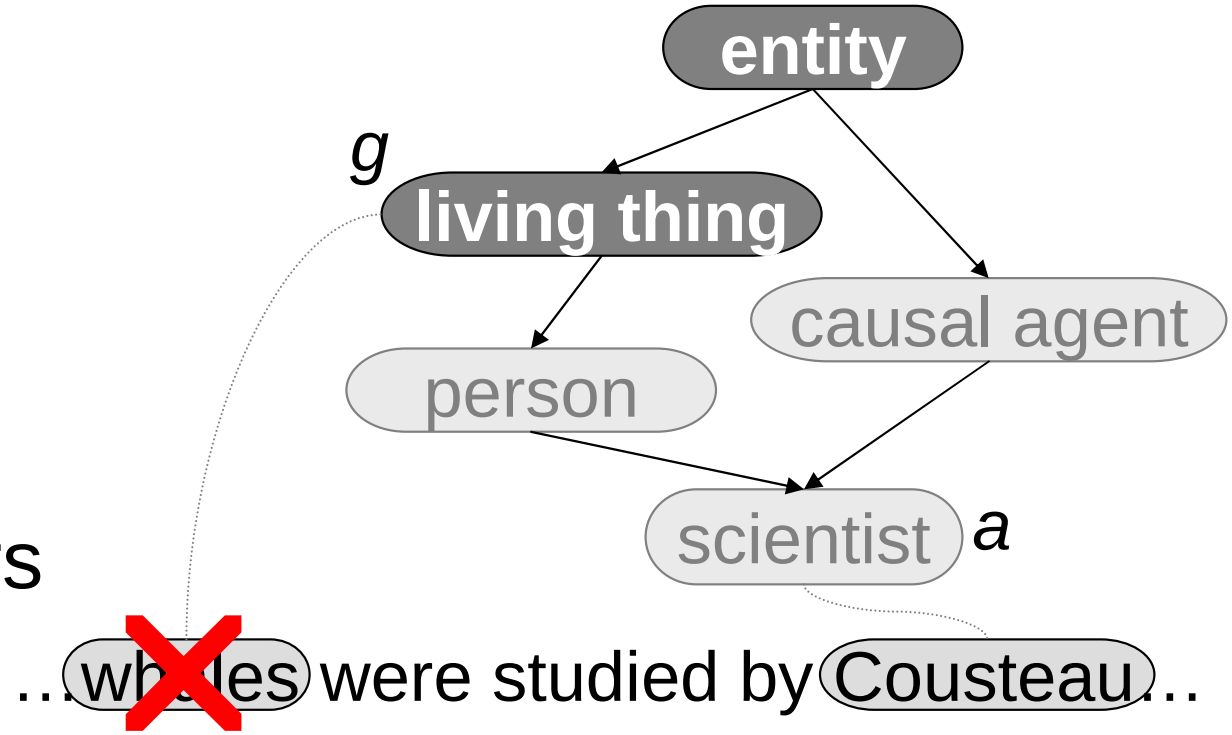
- Scientist who studied whales
  - ?s ∈ Category:Scientist
  - ?s ∈ Category:MarineBiologist
  - InContext(?c, ?s, study studied whale whales)
- Query expansion
  - Did Einstein, Bohr, Rutherford…study whales?
  - WordNet knows 650 scientiest, 860 cities
  - Wikipedia?
  - Impractical query times

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



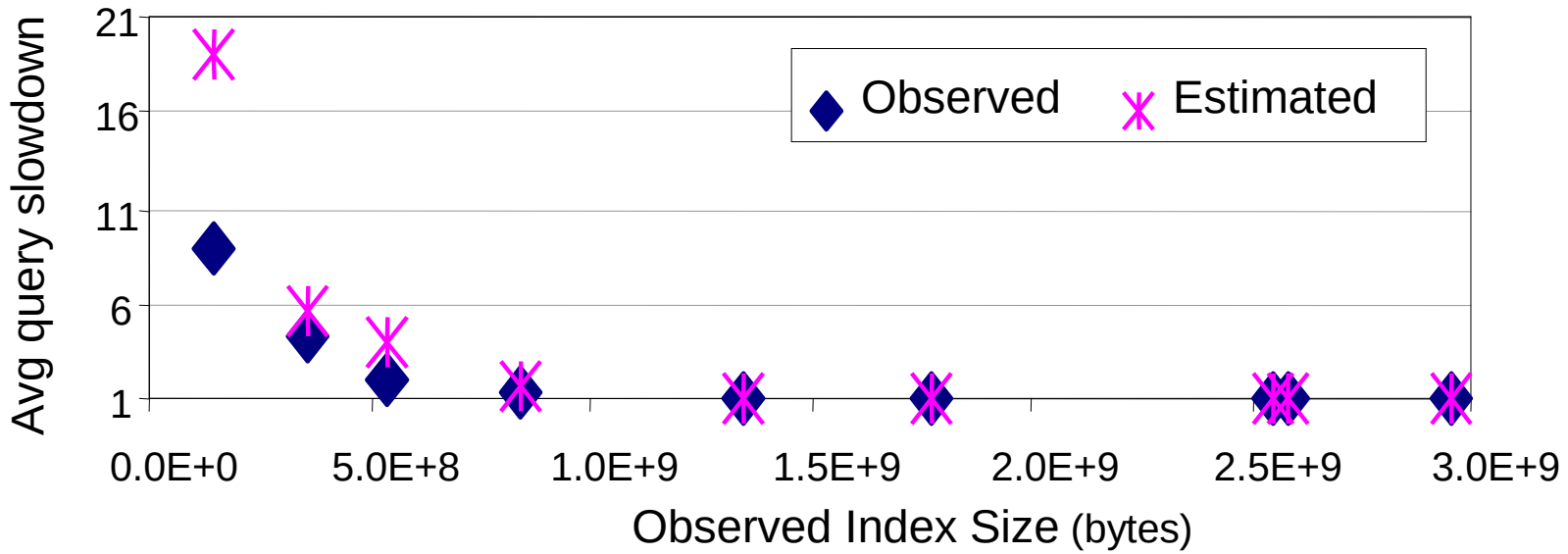
Post-filter

- 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
- [CPD2006]

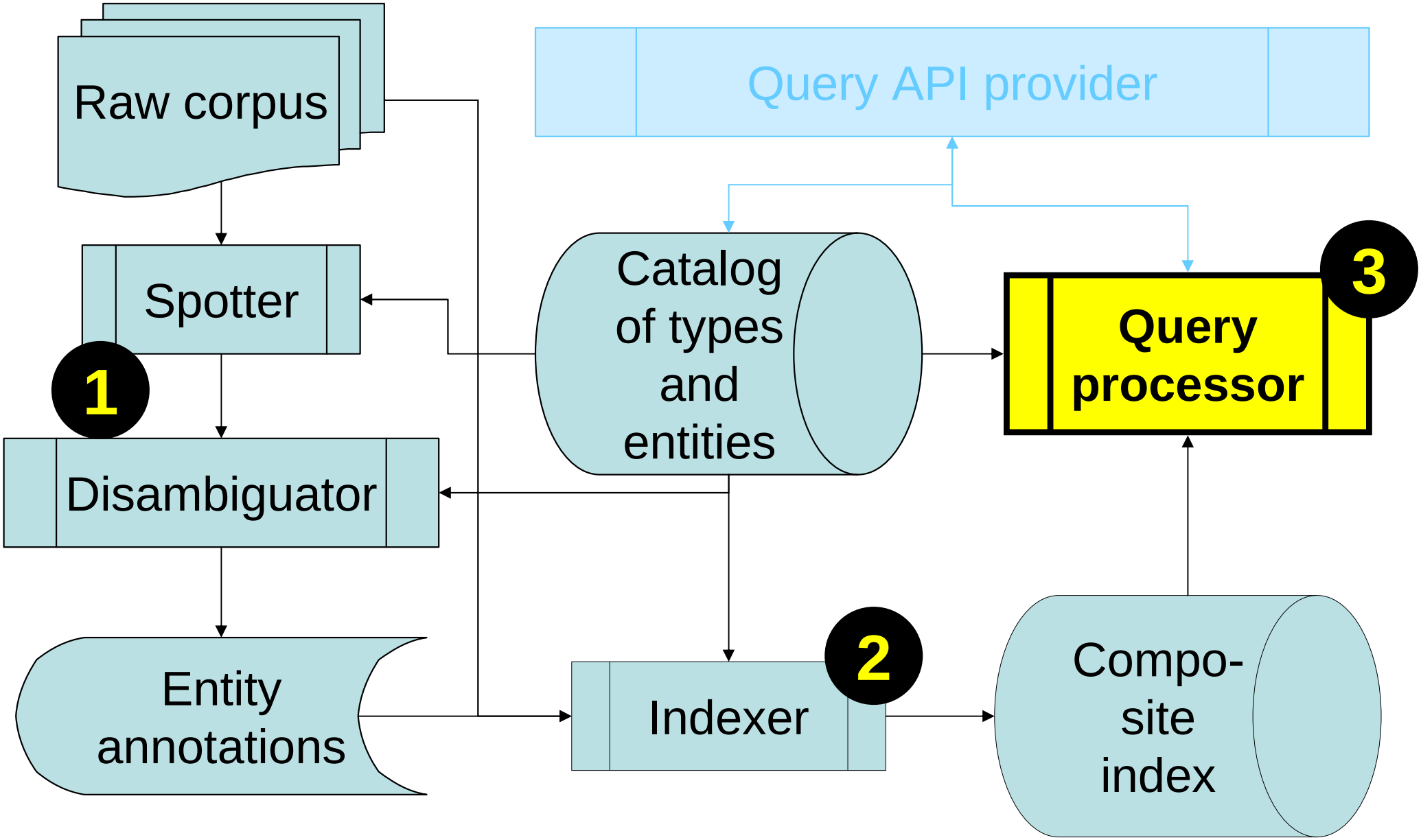
Index size vs. query slowdown



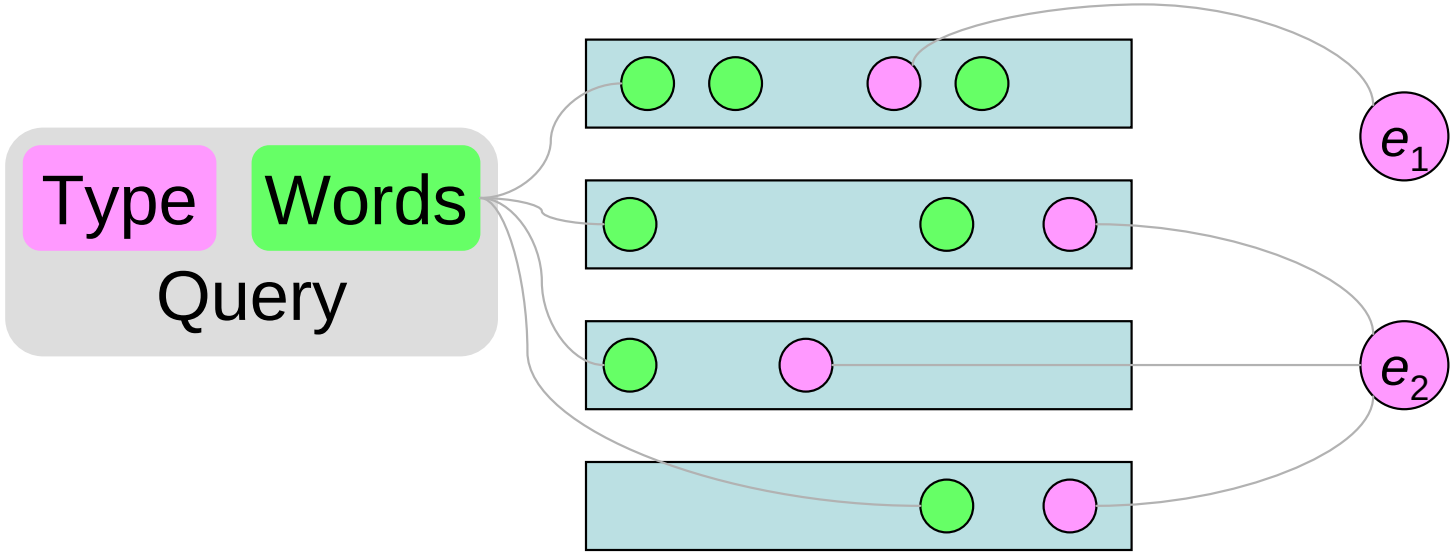
- Annotated TREC corpus
- Space = 520MB < inverted index = 910MB
- Query slowdown ≈ 1.8
- From TREC to Web?

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

Pieces to the puzzle



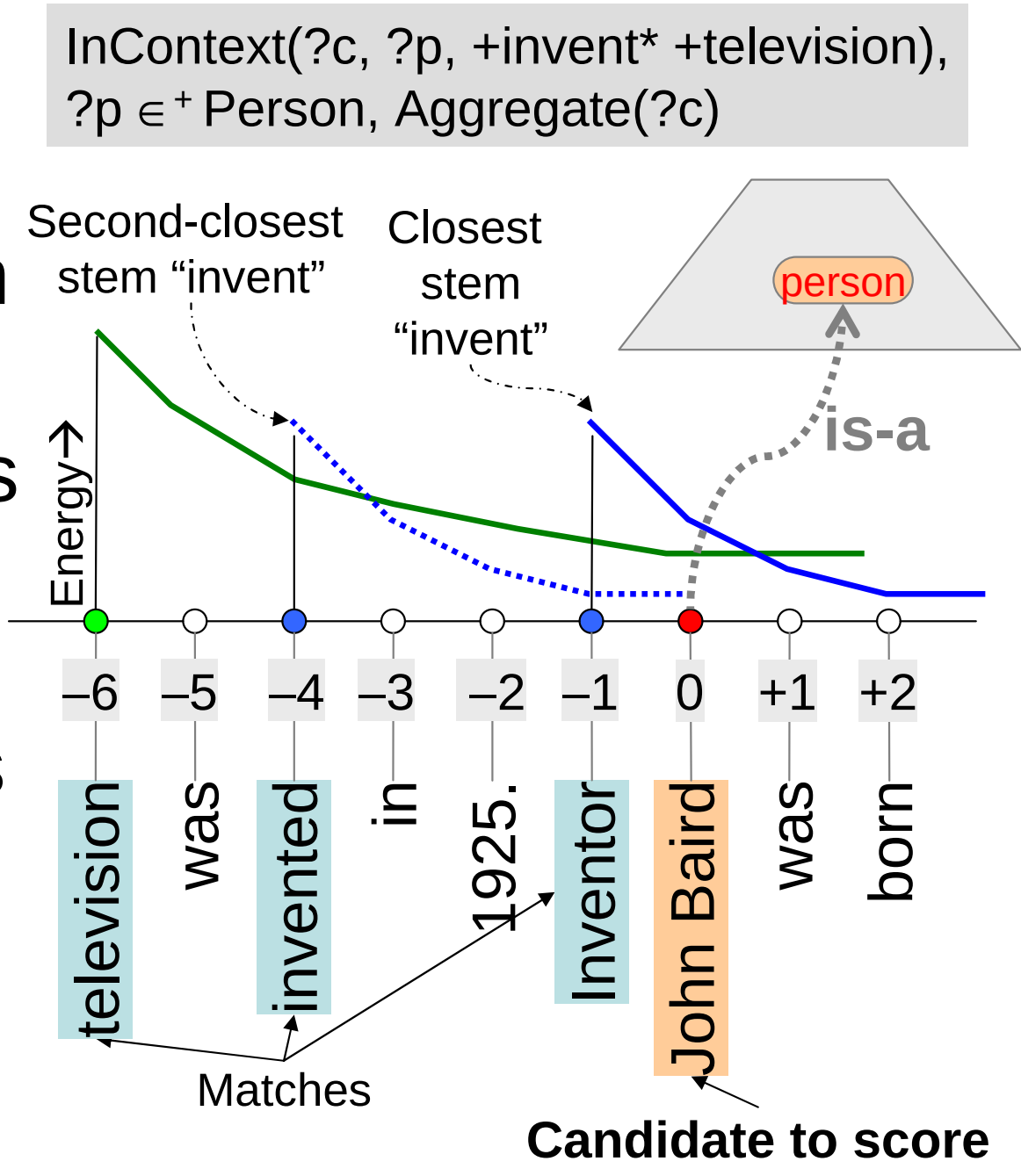
How to score and aggregate



- 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



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

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

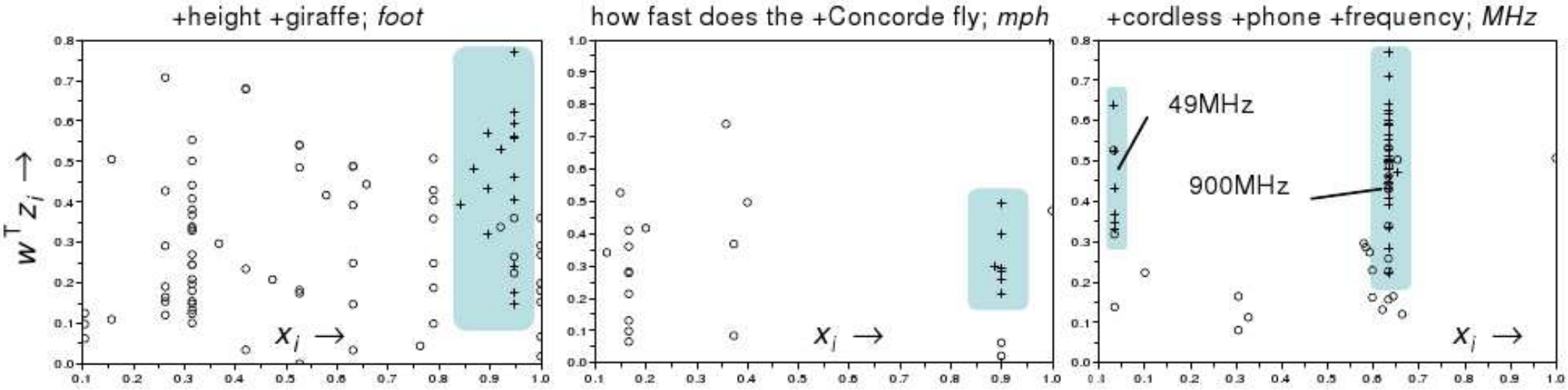
- During training fit w using partial order on f

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.

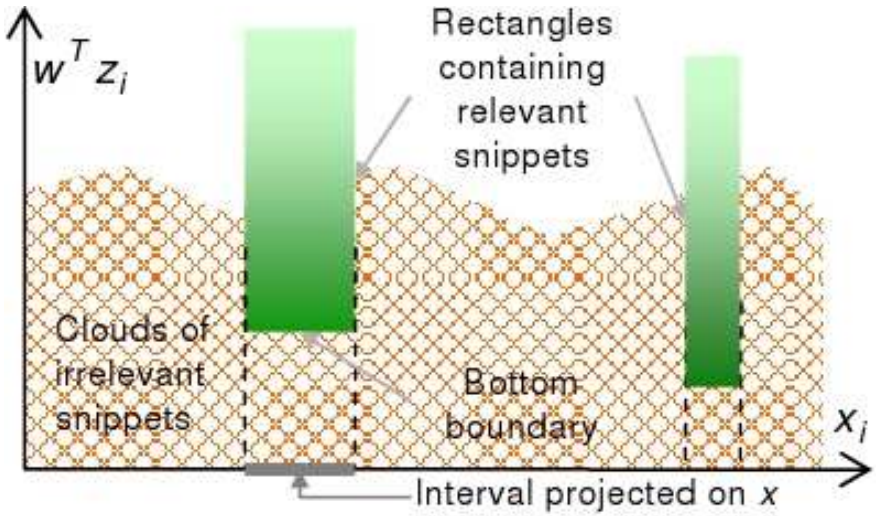
- Confounding candidates with correct units/type
- Can aggregation over snippets help
- Avoid deep NLP?
- Here we focus on quantity answers

Snippet score-quantity scatter



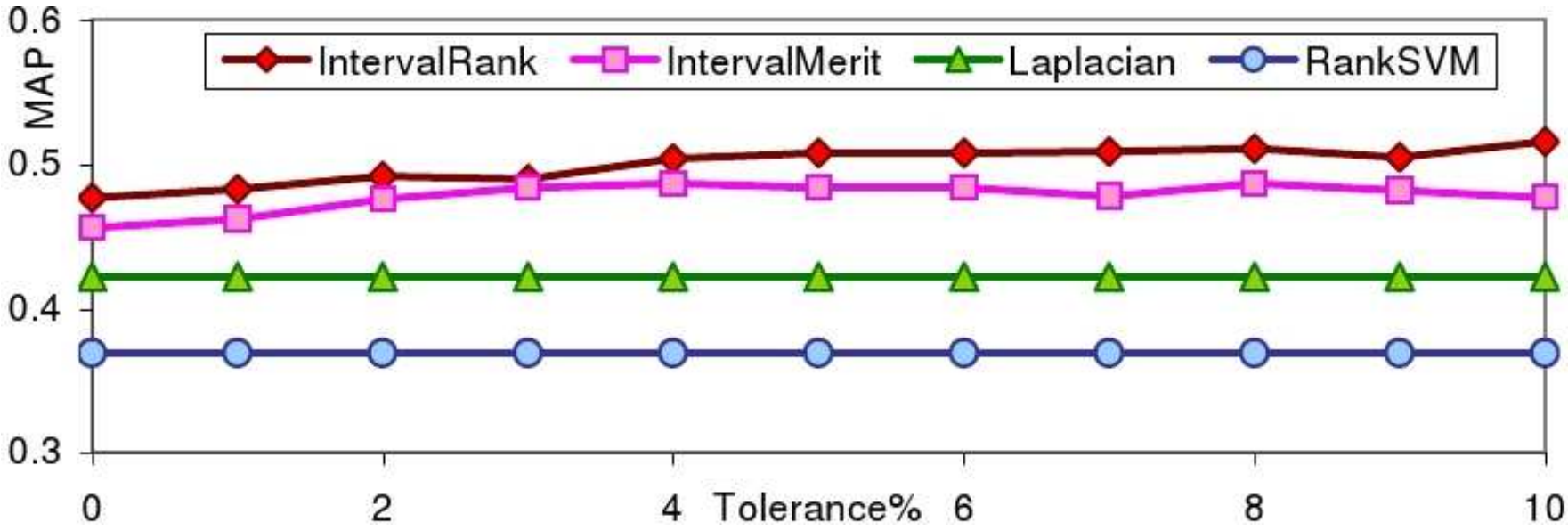
- Both axes scaled to [0, 1] for clarity
- Relevant/good snippets = +, irrelevant/bad = ◦
- Ideal w $\Rightarrow$ horizontal line separating + from ◦
- No such *w* for any query in our experiments
- Rectangles densely packed with many +, few ◦
  - Possibly > 1 rectangles for some queries

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 this richer info network
- Evolve into a practical search API?
  - Panel at WWW 2009
  - Prototype with .5B pages, 40x8 CPUs
- What will end-users adopt today? Vs.
- How can they take advantage of the new type-entity-snippet composite data model?

References

- SemTag/Seeker: Dill+ WWW 2003
- Wikify! Mihalcea+Csomai CIKM 2007
- Bunescu+Pasca Tree kernels EACL 2006
- Milne+Witten CIKM 2008
- [KSCR2009] Kulkarni+ KDD 2009
- [CPD2006] Chakrabarti+ WWW 2006
- EntityRank Cheng+ VLDB 2007
- Associative QA Ko+ SIGIR 2007
- Laplacian Qin+ WWW 2008
- [BCR2009] Banerjee+ SIGIR 2009