Amsterdam
Data Science

# Long tail entities

**Maarten de Rijke**

partners

VU VRIJE UNIVERSITEIT AMSTERDAM

UNIVERSITEIT VAN AMSTERDAM

Hogeschool van Amsterdam

CWI

# Wrap up

- Forget about entity linking (disambiguation), there are so many more entity-related tasks to work where interesting long tail phenomena emerge

- **Come up with a realistic scenario, with realistic distributions, don't go for modified distributions, do include extrinsic evaluation**

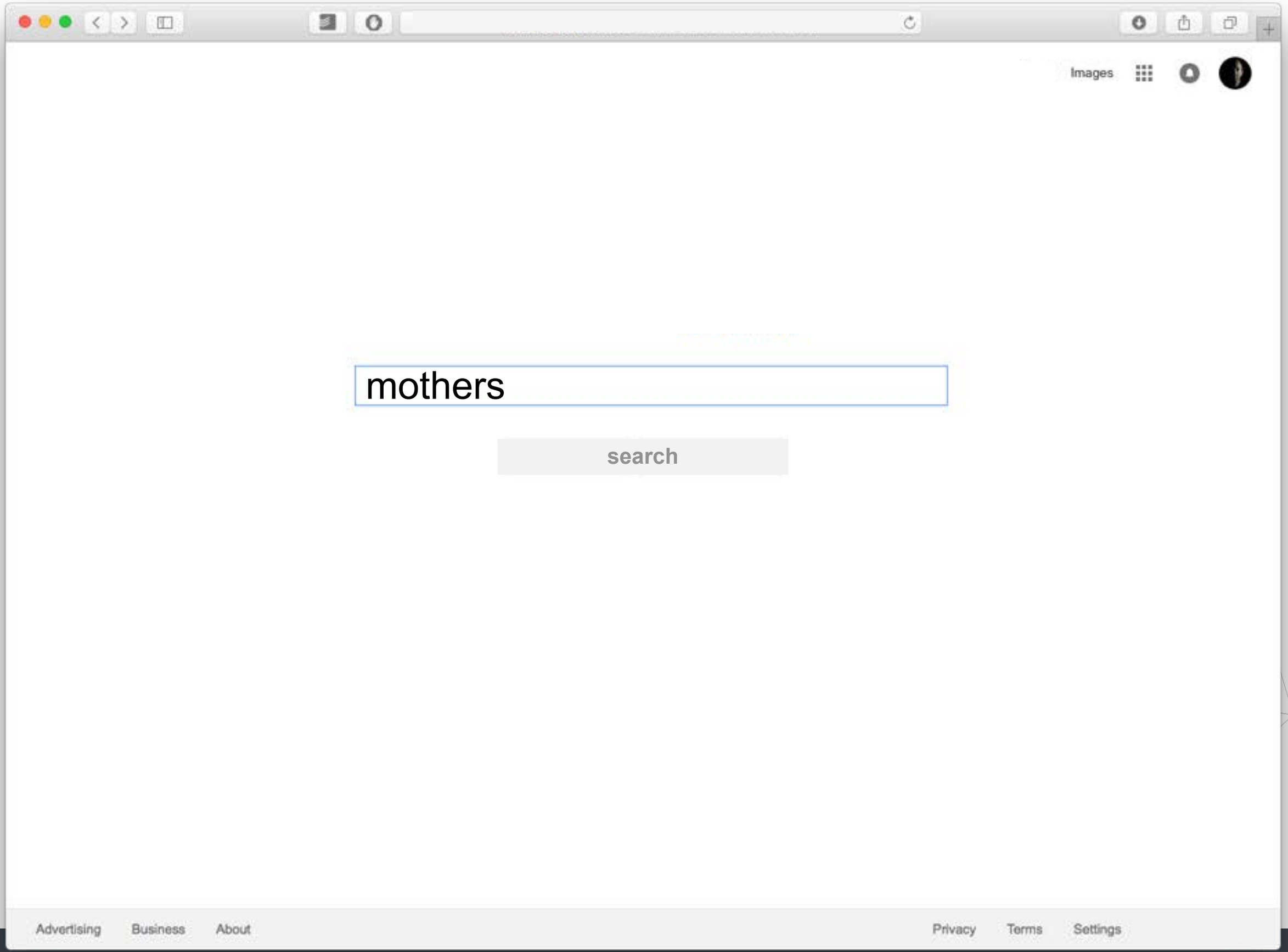
- Examples

  - Entity representations, discovering and entity aspects, generate relationship explanations,  use entities and query suggestions in extremely data sparse contexts, discover latent semantics of entities and use for product search, document filtering for entities (head or torso or long tail)

- Based on joint work with David Graus, Evangelos Kanoulas, Xinyi Li, Edgar Meij, Daan Odijk, Bob Schijvenaars, Ridho Reinanda, Manos Tsagkias, Christophe Van Gysel, Nikos Voskarides, Wouter Weerkamp, Marcel Worring

# It's all about entities

- Entities (people, locations, organizations, ...) play central organizing role
- In search
  - in **web** search, up to 70% of the queries are entity queries (Lin et al., 2012; Guo et al., 2011)
  - in **academic** search, the proportion of queries that contain entities is over 93% (Li et al., 2016)
  - entities have become retrievable items


Images
mothers
search
Advertising
Business
About
Privacy
Terms
Settings

mothers

All Images Maps Videos News More Search tools

About 226.000.000 results (0,54 seconds)

**Mothers** are females who inhabit or perform the role of bearing some relation to their children, who may or may not be their biological offspring.


Mother - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Mother**

More about Mother

See results about

Mothers (Band)
Songs: It Hurts Until It Doesn't, Copper Mines, Too Small…
Albums: When You Walk A Long Distance You Are Tired

Mother - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Mother**
**Mothers** are females who inhabit or perform the role of bearing some relation to their children, who may or may not be their biological offspring.
Mother (disambiguation) - Mom (disambiguation) - Mother's Day - Mother goddess

Mothers® Polishes·Waxes·Cleaners
www.**mothers**.com/
The finest car care waxes, polishes, and cleaners available anywhere. PowerBall 4Paint gives you professional results in your own garage.
Mothers Products - How To Video and Guide - Mothers Wax Retailers

In the news

Apple cuts same-sex parents from Mother's Day ads in six countries
The Guardian - 4 hours ago
Apple has removed an image of a same-sex couple with their two babies from its **Mother's** ...

Here's one reason there's only one Chinese cop patrolling Monterey Park: Chinese mothers
Los Angeles Times - 10 hours ago

In the news

**Apple cuts same-sex parents from Mother's Day ads in six countries**
The Guardian - 4 hours ago
Apple has removed an image of a same-sex couple with their two babies from its **Mother's** ...

Here's one reason there's only one Chinese cop patrolling Monterey Park: Chinese mothers
Los Angeles Times - 10 hours ago

Tonypandy toddler Finley Thomas allegedly murdered by mother's boyfriend
Daily Mail - 8 hours ago

More news for mothers

Images for mothers — Report images

More images for mothers

**Mothers - Facebook**
https://www.facebook.com/nestingbehavior/
**Mothers**. 9210 likes · 521 talking about this. always is a dirty word to say.

**Mother Sacramento**
**mothers**acramento.com/
About · Gallery; Menu. Menu · Dinner Menu · Events · Contact. Lunch. Mon-Fri 11am-3pm. Dinner. Tue-Thu 5pm-9pm. Fri & Sat 5pm-10pm. Sunday. Closed ...

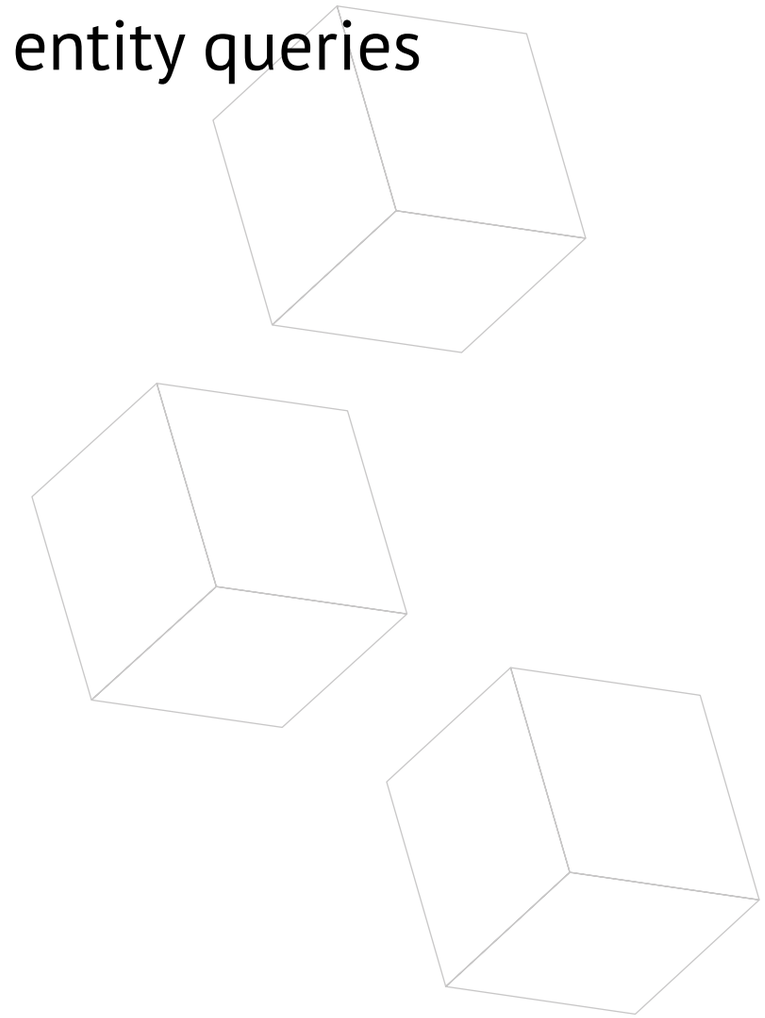
**Mothers Bistro | It's all about the love**
www.**mothers**bistro.com/
"At **Mother's**, we take traditional homemade favorites and refine them with classical cooking techniques, so they're like mom's cooking, only a bit better.

**Mother's Restaurant - New Orleans - World's Best Baked Ham - 401 ...**
www.**mothers**restaurant.net/
Home · **Mother's** Next Door · Gift Shop · History · Menus · Contact Us. **Mother's** Restaurant - New Orleans. An error occurred. Try watching this video on ...

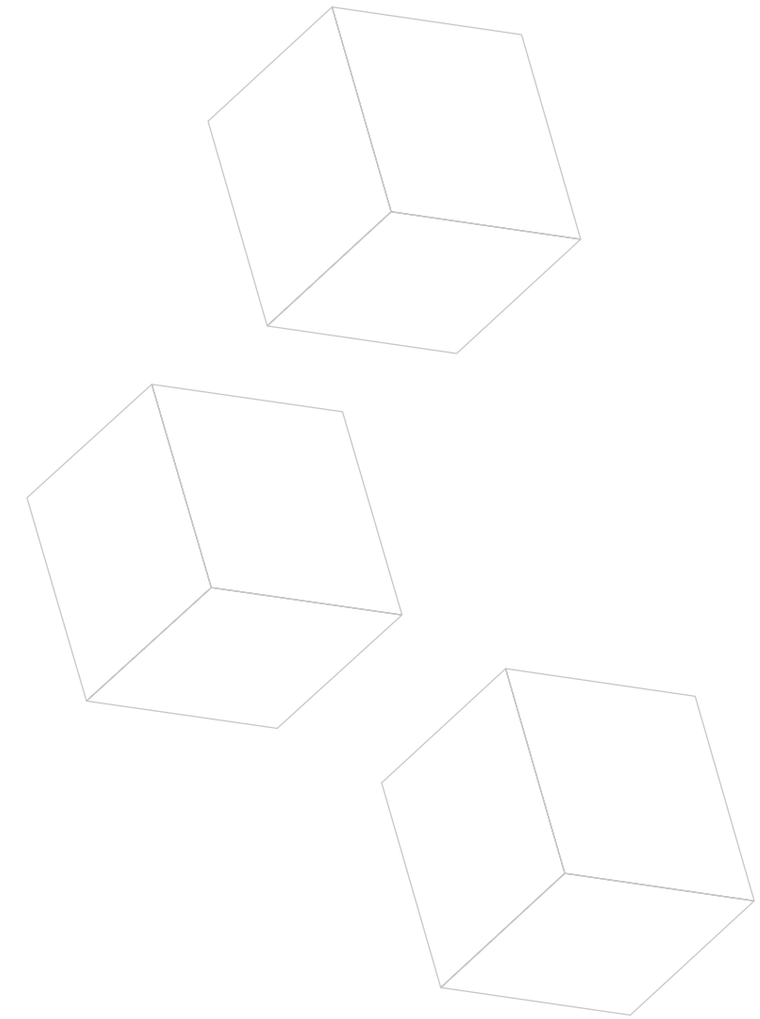
((mothers)) | Free Listening on SoundCloud

# Information needs around entities

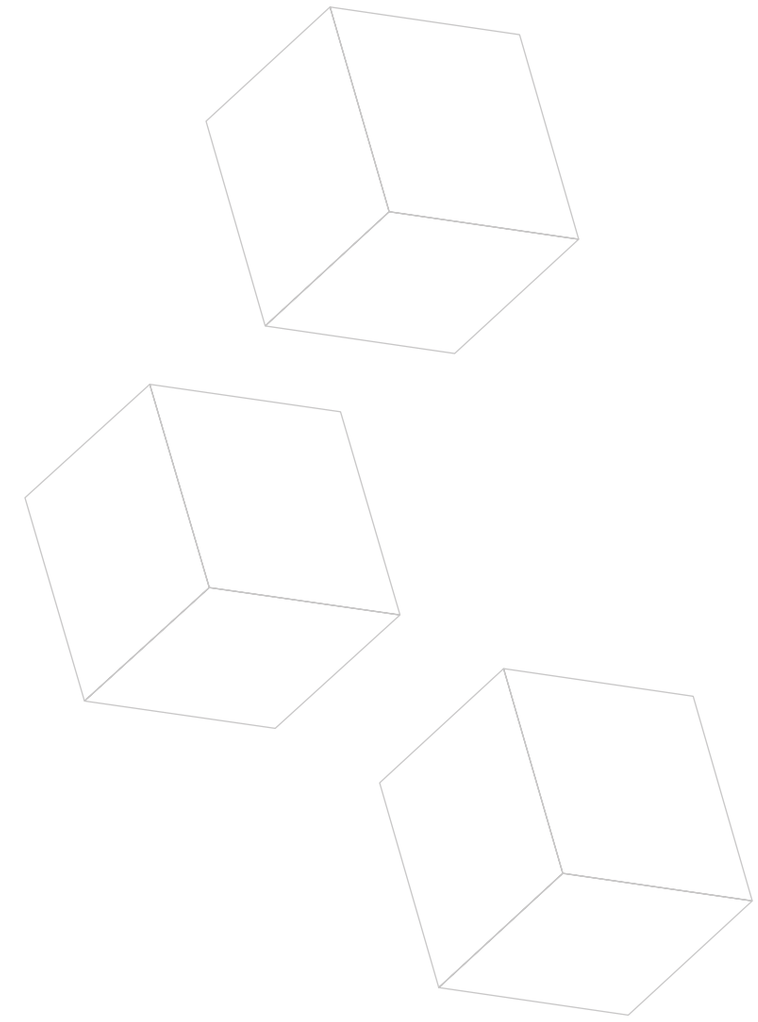
- Recognize and respond to diverse **intents** behind entity queries
- You know an entity by the stuff it hangs out with
  - Words
  - Facets
  - Entities
  - Relations
  - Multimedia
  - Structured data
  - …



# Sampling recent work ...

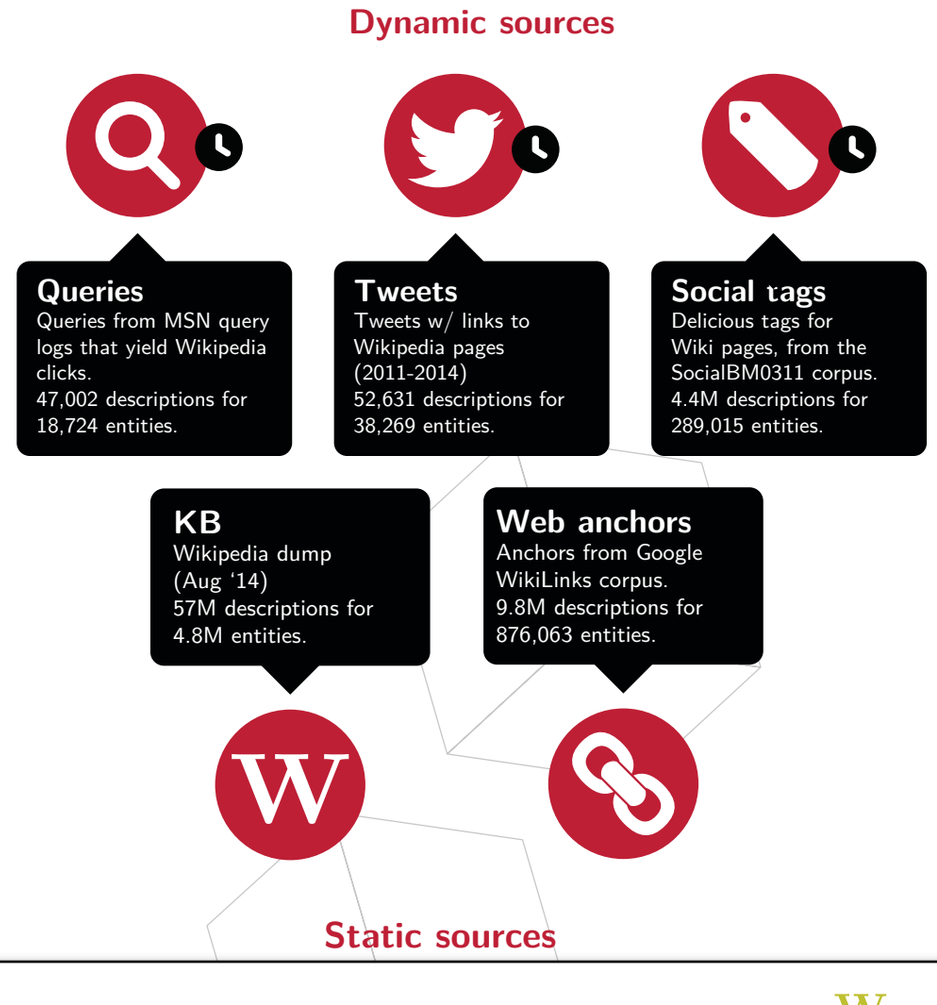


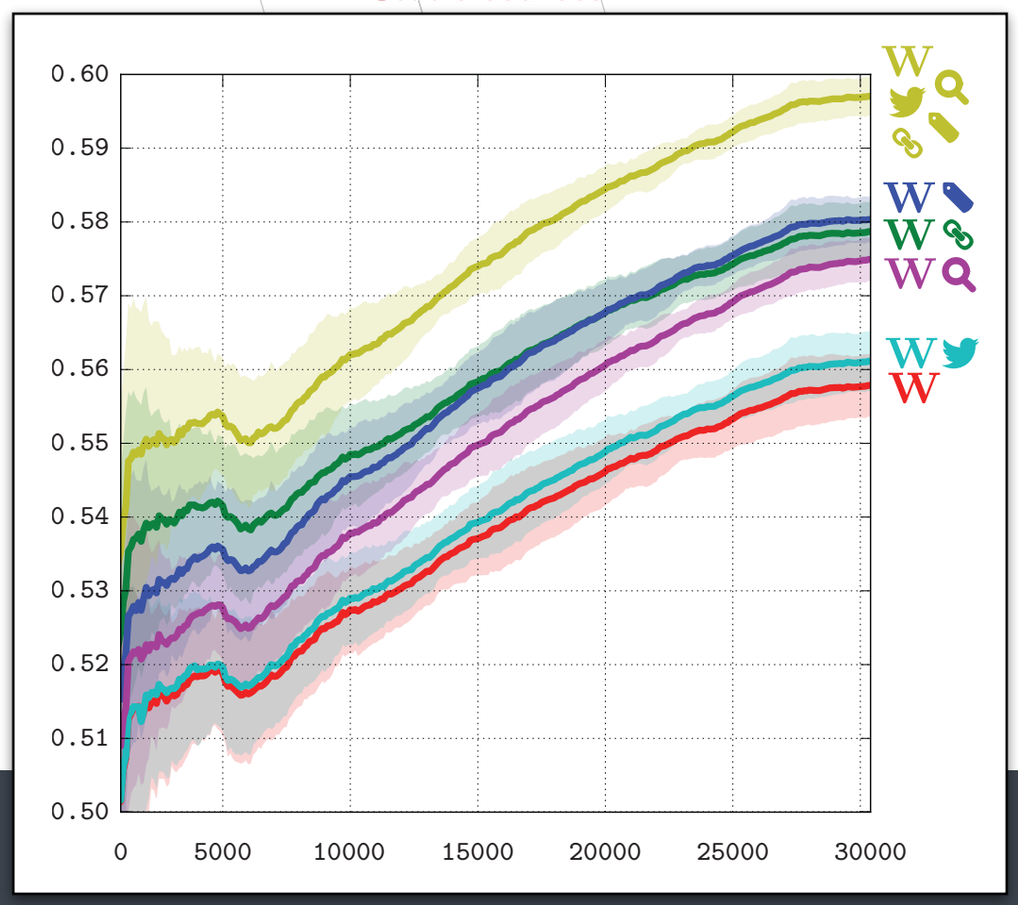

1



# Entity representations

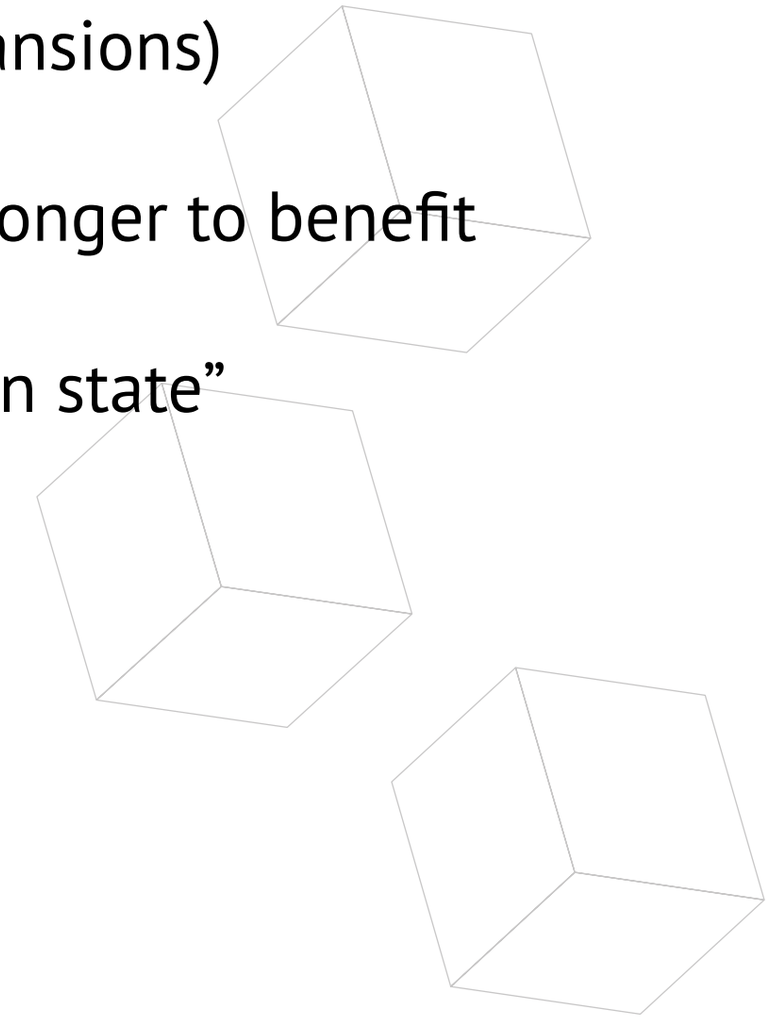
- **Task**: Entity retrieval
- **Method**: Expand entity descriptions with additional descriptions
- **Performance**:
  - All expansions: ~0.60 MAP (yellow)
  - No expansions: ~0.55 MAP



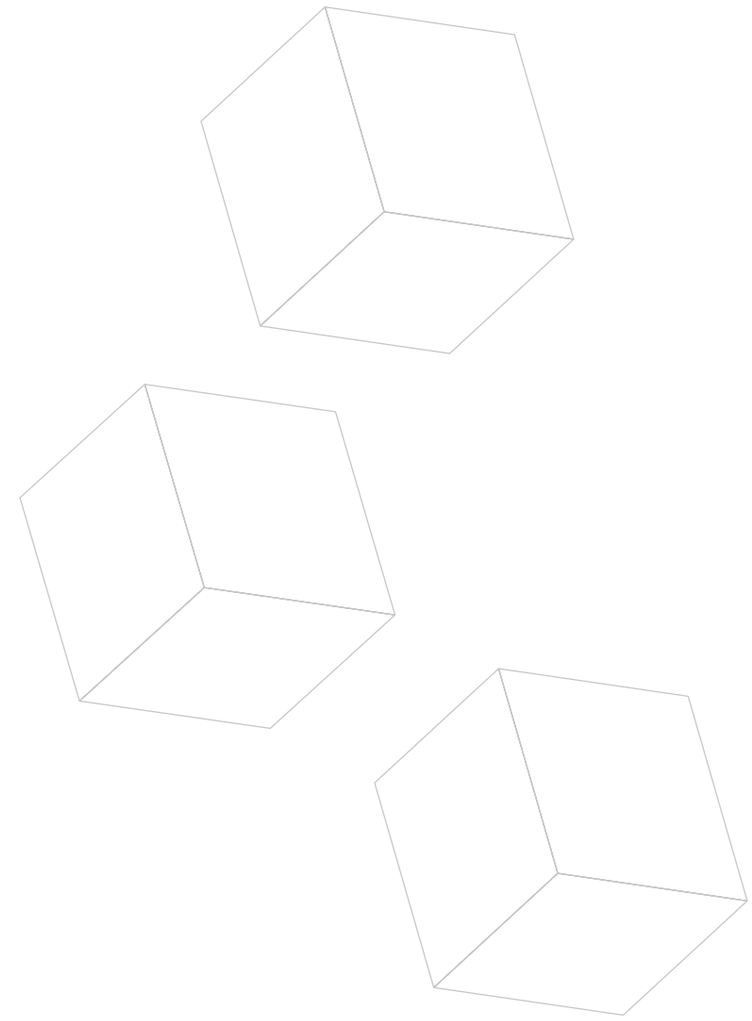



# Head/tail effect impact

- Head entities may get “swamped” (too many expansions)
- Tail entities receive fewer expansions, may take longer to benefit
- Solution: inform ranker of each entity’s “expansion state”



2



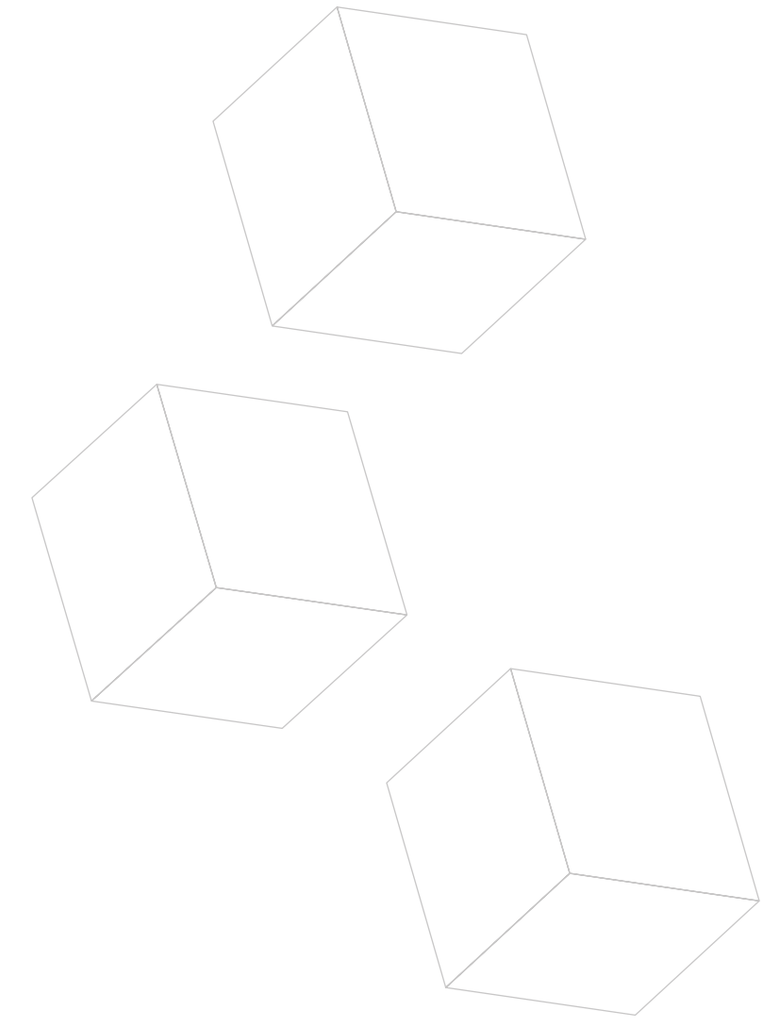
# Explaining entity relationships

- **Problem**: Knowledge graphs represent entity relationships using formal descriptions which are not suitable for presenting to the end user
- **Task**: Explain the relationships using human-readable descriptions
- **Approach**
  - Given an entity pair and relationship,
  - Extract and enrich candidate sentences
  - Rank candidate sentences by how well they describe the relationship using learning to rank (text, entity, relation and source features)
- **Main findings**
  - Significant improvements over state-of-the-art sentence retrieval methods
  - Relationship-dependent models significantly improve performance

# Head/tail effect impact

- **Main limitation**
  - For tail entities the text corpus is less likely to contain any highly-quality sentence for a given entity pair
- **Solution**
  - Use existing textual descriptions (for head entities) and information from the knowledge graph to learn how certain relationships are described
  - Automatically generate descriptions for entity pairs for which a textual description is not available


3



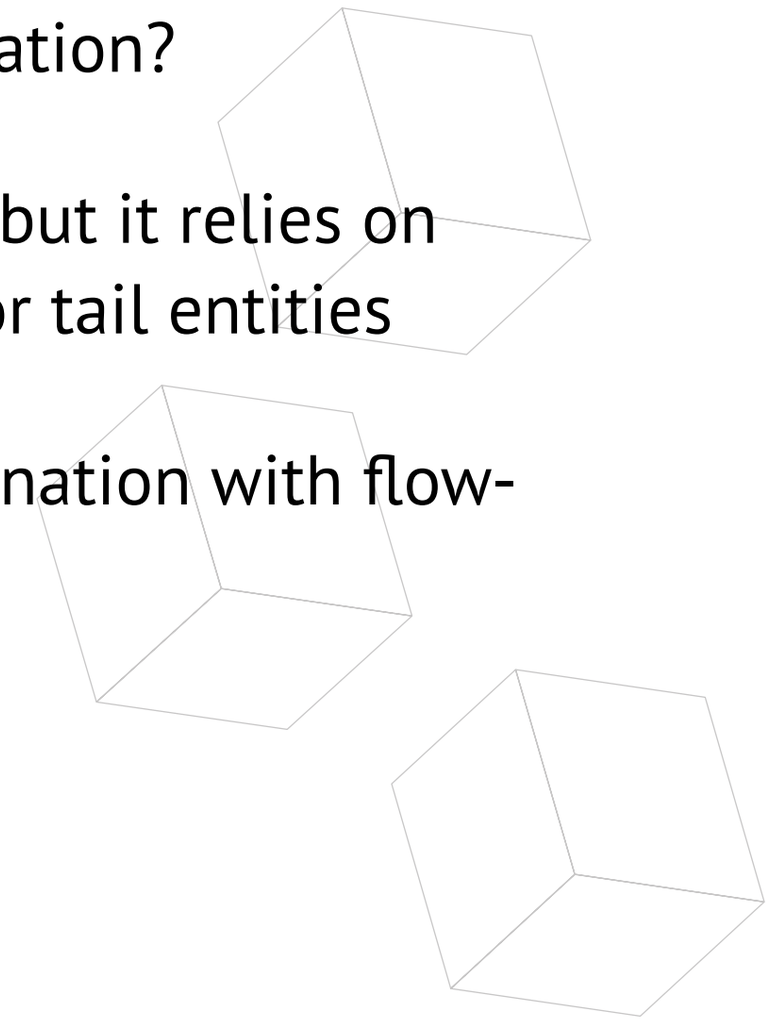
# Entity aspects

- Entity aspects: common search tasks in the context of an entity
- Task: mining, ranking, **recommending** aspects
- Method: behavioral and semantic relatedness
- Results (recommendation):

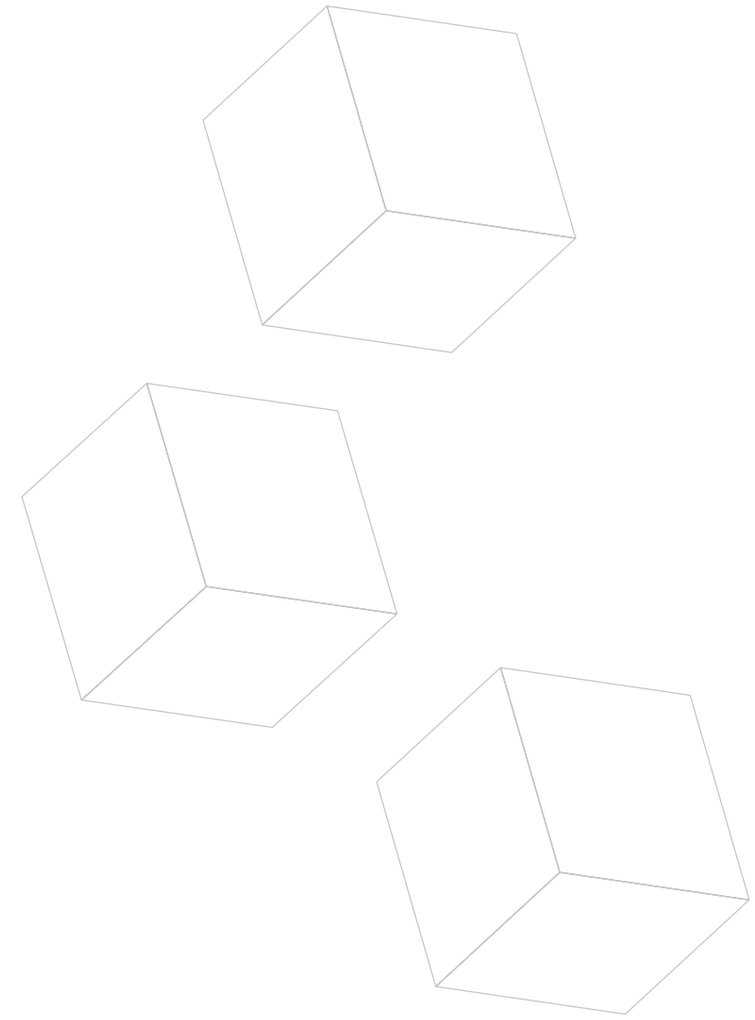
| Method | MRR | SR |
|---|---|---|
| Aspect-semantic | 0.0431 | 0.0244 |
| Aspect-flow | 0.0602▲ | 0.0451▲ |
| Aspect-combined-rr | 0.0674▲ | 0.0486▲ |
| Aspect-combined-convex ($\lambda = 0.85$) | 0.0650▲ | 0.0465▲ |

# Head/tail effect impact

- How would it affect things in aspect recommendation?
  - Behavioral (flow-based) approach works best, but it relies on observing past transitions, which are sparse for tail entities
  - Mitigate by using semantic similarity in combination with flow-based approach




4



# Entity-based query suggestion

- Providing suggestions for rare queries is hard, suffering from query sparsity
- Using entity to overcome sparsity
- mQEG (modified Query Entity Graph) combines behavior signal from QFG (Query Flow Graph) plus semantics from entity

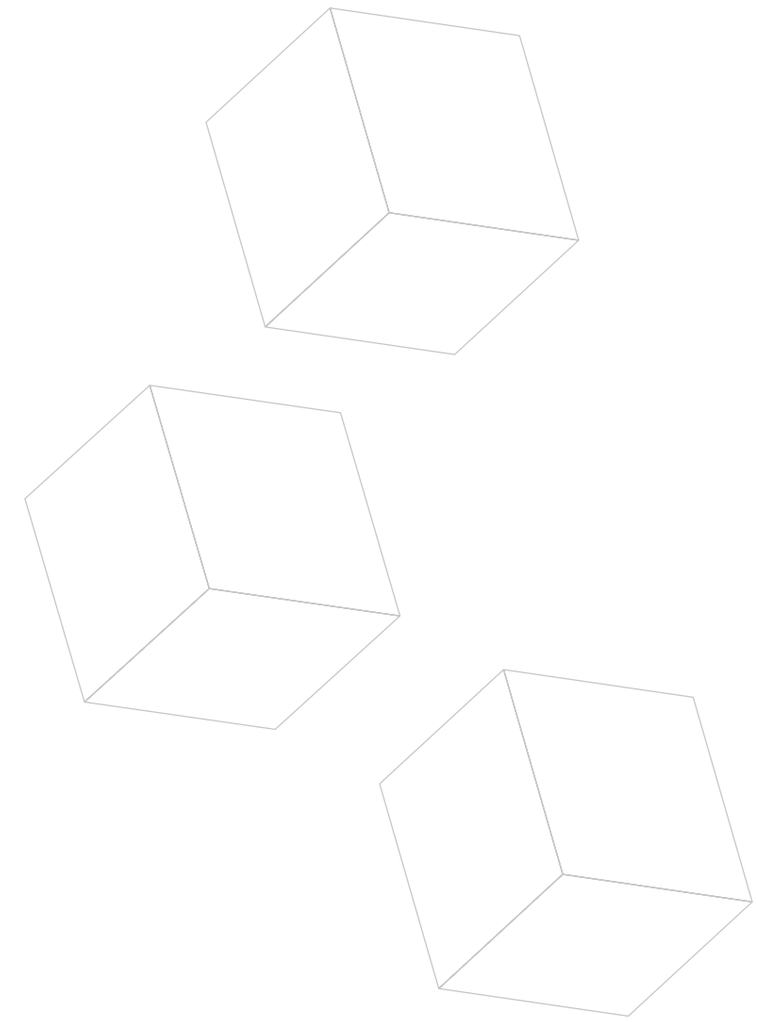
Table: Query suggestion for null queries in academic search

| Model | SR@1 | SR@3 | SR@5 | SR@10 |
|---|---|---|---|---|
| QFG | 0 | 0 | 0.32 | 0.65 |
| mQEG | ▲3.22 | ▲3.55 | ▲4.19 | ▲6.45 |

# Head/tail effect impact

- Compared to head queries, tail queries are sparse
- Query suggestions for tail queries are difficult
- Entity in queries helps overcome query sparsity, and brings up relevant query suggestions


5



# Semantic entity representations

- Jointly learn representations of words, entities and relationship between them
- Learn relationship between words and entities directly using gradient descent
- Unsupervised

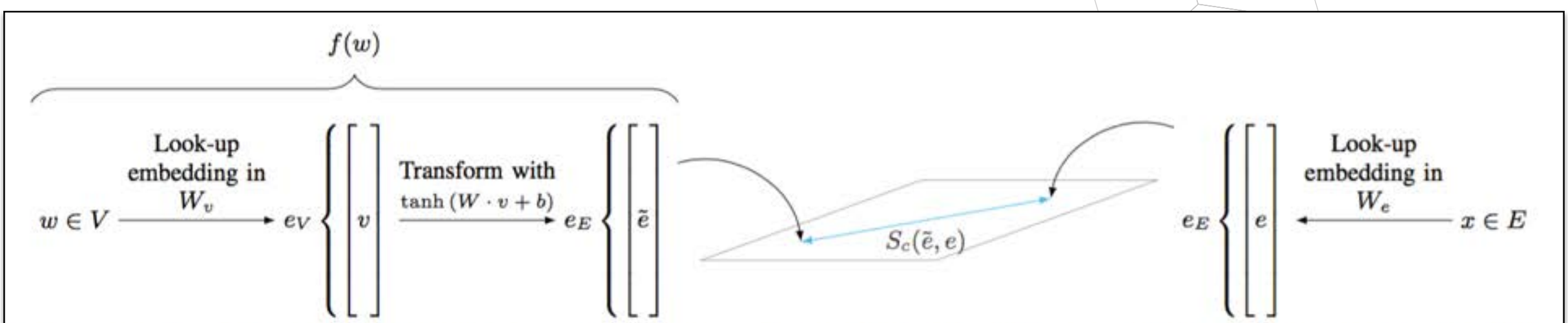



**Figure 2: Schematic representation of the Latent Semantic Entities model for a single word $w$. Word embeddings $W_v$ ($e_V$-dim. for $|V|$ words), entity embeddings $W_e$ ($e_E$-dim. for $|X|$ entities) and the mapping from words to entities ($e_E$-by-$e_V$ matrix $W$, $e_E$-dim. vector $b$) are learned using gradient descent.**

# Head/tail effect impact

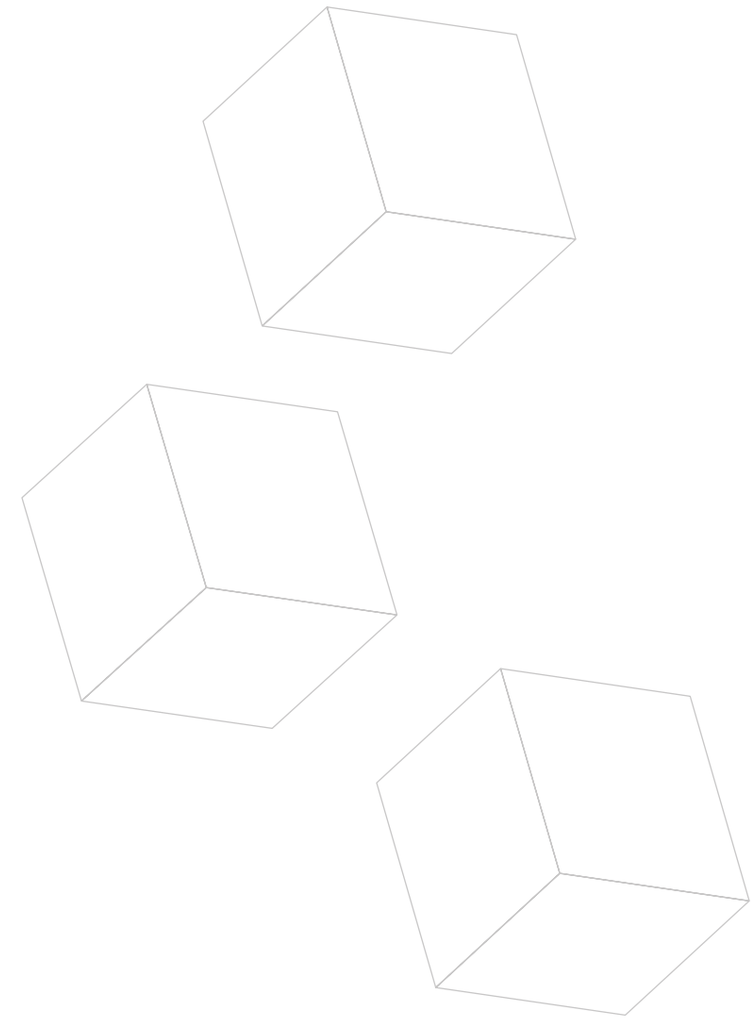
- In a product retrieval task, semantic models tend to outperform purely lexical models for entities with less specific descriptions

- In a learning to rank setup, semantic ranking provides a significant improvement over lexical plus popularity-based ranking features

**Table 2: Correlations coefficients between average IDF of lexically matched terms in documents associated with relevant entities and △NDCG. A negative correlation coefficient implies that queries consisting of more specific terms (i.e., low document freq.) that occur exactly in documents associated with relevant entities are more likely to benefit from QLM, whereas other queries (with less specific terms or less exact matches) gain more from LSE. Significance is achieved for all benchmarks ($p < 0.01$) using a permutation test.**

| Benchmark | Spearman $R$ | Pearson $R$ |
|---|---|---|
| Home & Kitchen | −0.30 | −0.35 |
| Clothing, Shoes & Jewelry | −0.40 | −0.37 |
| Pet Supplies | −0.17 | −0.17 |
| Sports & Outdoors | −0.34 | −0.36 |


6



# Document filtering for long tail entities

- **Document filtering for entities**
  - System processes time-ordered corpus for documents that are relevant to a set of entities in order to select those documents that contain vital information about the entities
- State of the art for head entities:
  - **Entity-dependent**: rely and are trained on specifics of differentiating features for entity at end
  - Tend to use **extrinsic** information (e.g., Wikipedia page views and related entities)—widely available only for head entities
- Entity-dependent approaches based on extrinsic signals ill-suited as document filtering methods for **long-tail** entities

# Document filtering for long tail entities (2)

- **Entity-independent** approaches to document filtering learn characteristics of documents suitable for updating a knowledge base profile by utilizing signals from documents, initial profile of entity (if present), and relationships between entities and documents
  - Applicable to unseen entities
  - Avoid building models for every entity
- Core intuition
  - **Informativeness**, **entity-saliency**, and **timeliness**: document that contains rich set of facts in timely manner, and in which entity is prominent good candidate for enriching knowledge base profile

**Table 2: Features for document filtering, for entities $e$ and/or documents $d$. The last column indicates the value types of the features: $N$ for numerical features and $C$ for categorical features.**

| Feature | Description | Source | Type | Value |
|---|---|---|---|---|
| $SRC(d)$ | Document source/type | [5] | basic | N |
| $LANG(d)$ | Document language | [5] | basic | N |
| $REL(e)$ | Number of of related entities of $e$ | [5] | basic | N |
| $DOCREL(e)$ | Number of of related entities of $e$ in $d$ | [5] | basic | N |
| $NUMFULL(d, e)$ | Number of mentions of $e$ in $d$ | [5] | basic | N |
| $DOCREL(d, e)$ | Number of of related entities of $e$ in $d$ | [5] | basic | N |
| $NUMPARTIAL(d, e)$ | Number of partial mentions of $e$ in $d$ | [5] | basic | N |
| $FPOSFULL(d, e)$ | First position of full mention of $e$ in $d$ | [5] | basic | N |
| $LPOSPART(d, e)$ | Last position of partial mention of $e$ in $d$ | [5] | basic | N |
| $SPRPOS(d, e)$ | Spread (first position − last position) of mentions of $e$ in $d$ | [5] | basic | N |
| $SIM_{cos}(d, p_e)$ | Text cosine similarity between $d$ and $p_e$ | [5] | basic | N |
| $SIM_{jac}(d, p_e)$ | Text jaccard similarity between $d$ and $p_e$ | [5] | basic | N |
| $PREMENTION_h(d, e)$ | Mention count of entity in the previous $h$ hour before document creation time of $d$ | [28] | basic | N |
| $DOCLEN_{chunk}(d)$ | Length of document in number of chunks | this paper | basic | N |
| $DOCLEN_{sent}(d)$ | Length of document in number of sentences | this paper | basic | N |
| $ENTITYTYPE(e)$ | Type of $e$ (PER, ORG, or FAC) | this paper | basic | C |
| $PROFILETYPE(e)$ | Profile type: *wiki*,*web*, or *null* | this paper | basic | C |
| $PROFILELEN(e)$ | Length of entity profile $e$ | this paper | basic | N |
| $ASPECTSIM_k(d)$ | Cosine similarity between $d$ and $aspect_k$ estimated from Wikipedia | this paper | informativeness | N |
| $RELOPEN_k(d)$ | Number of normalized open relation phrases $k$ in $d$ | this paper | informativeness | N |
| $RELSCHEMA_k(d)$ | Number of relation type $k$ in document $d$ | this paper | informativeness | N |
| $NUMENTITIES(d)$ | Number of unique entity mentions in the documents | this paper | entity saliency | N |
| $NUMMENTIONS(d)$ | Number of entity mentions in the documents | this paper | entity saliency | N |
| $NUMSENT(d, e)$ | Number of sentences in $d$ containing entity $e$ | this paper | entity saliency | N |
| $FULLFRAC(d, e)$ | Number of full mentions of $e$ in the document, normalized by number of entity mentions | this paper | entity saliency | N |
| $MENTIONFRAC(d, e)$ | Number of full or partial mentions of $e$ in the document, normalized by number of entity mentions | this paper | entity saliency | N |
| $TMATCH_Y(d)$ | Number of year expressions of timestamp $t$ in $d$ | this paper | timeliness | N |
| $TMATCH_{YM}(d)$ | Number of year, month expressions of timestamp $t$ in $d$ | this paper | timeliness | N |
| $TMATCH_{YMD}(d)$ | Number of year, month, date expressions of timestamp $t$ in $d$ | this paper | timeliness | N |

# Document filtering for long tail entities (4)

- Four groups of features
- Gradient boosted decision trees
- Experiment
  - TREC KBA 2014 data
  - 21M doc subset of 1.2B doc TRAC KBA StreamCorpus
  - Arxiv, classifieds, forums, mainstream news, memetracker, news, reviews, social, and blogs

**Table 5: Results segmented by entity popularity. Significance of EIDF result is tested against the strong baseline (BIT-MSRA). Significant improvement is denoted with ▲ ($p < 0.05$). Here the *null profiles* segment represents the long-tail entities.**

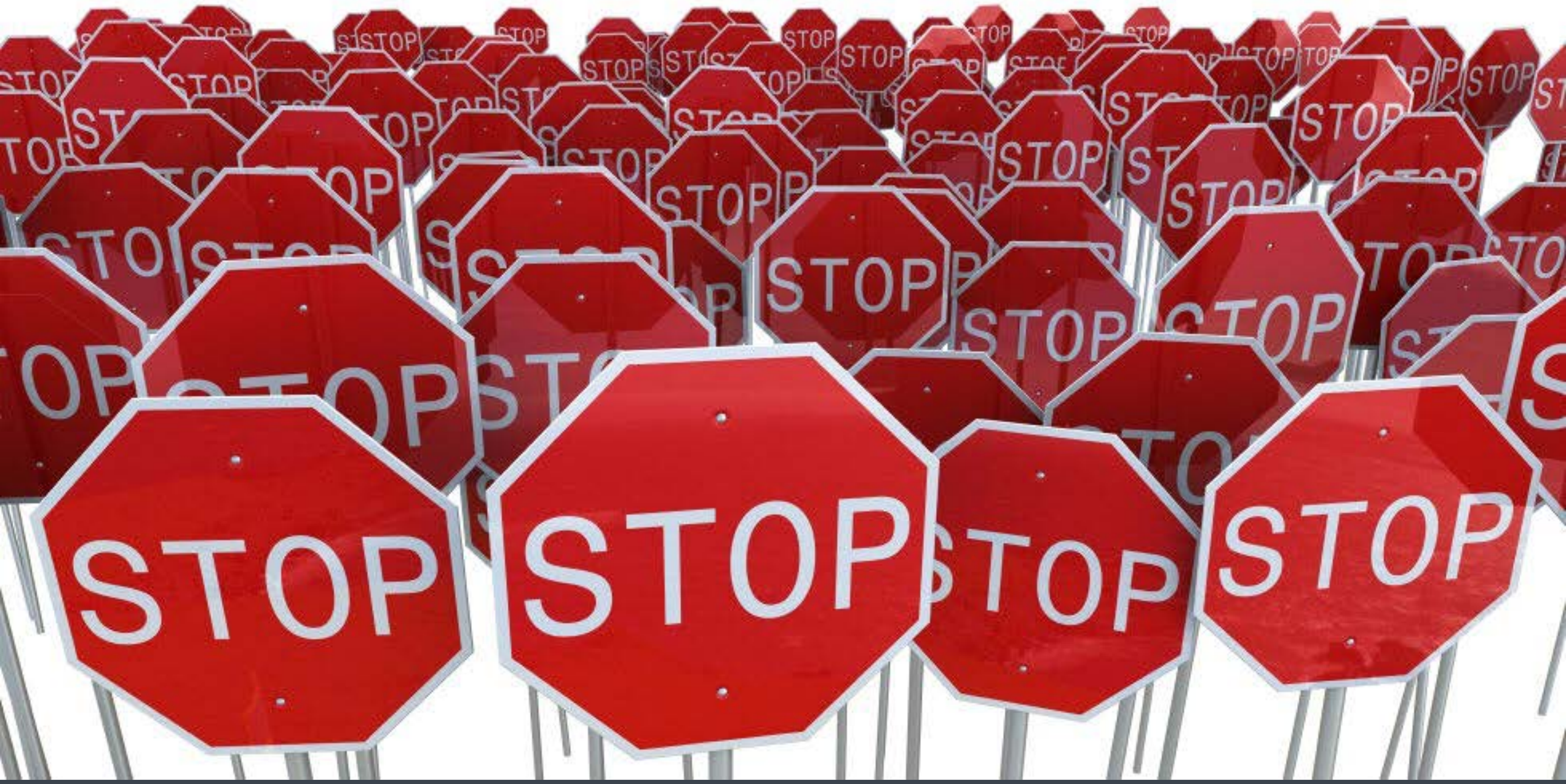
| Segment | P | R | F | $SU_\theta$ |
|---|---|---|---|---|
| *Null profiles* | | | | |
| Official baseline | 0.279 | **0.973** | 0.388 | 0.268 |
| BIT-MSRA | 0.362 | 0.630 | 0.404 | 0.313 |
| EIDF | **0.398**▲ | 0.645 | **0.433**▲ | **0.350**▲ |
| *Web profiles* | | | | |
| Official baseline | 0.391 | **1.000** | 0.513 | 0.381 |
| BIT-MSRA | **0.430** | 0.867 | **0.536** | **0.429** |
| EIDF | 0.424 | 0.827 | 0.517 | 0.410 |
| *Wiki profiles* | | | | |
| Official baseline | 0.169 | **0.975** | 0.275 | 0.044 |
| BIT-MSRA | 0.204 | 0.737 | 0.296 | 0.121 |
| EIDF | **0.227**▲ | 0.704 | **0.317** | **0.130** |

# Document filtering for long tail entities (4)

- Entity-independent approach to document filtering for knowledge completion
- Four types of feature: basic and informativeness are especially important
- Especially effective for long-tail entities
- Also competitive with entity-dependent approaches on head entities

**Table 10: Feature types within the top-30.**

| Feature type | Number of features |
|---|---|
| *basic* | 14 |
| *informativeness* | 13 |
| *entity saliency* | 2 |
| *timeliness* | 1 |


STOP

# Wrap up

- Forget about entity linking (disambiguation), there are so many more entity-related tasks to work where interesting long tail phenomena emerge

- **Come up with a realistic scenario, with realistic distributions, don't go for modified distributions, do include extrinsic evaluation**

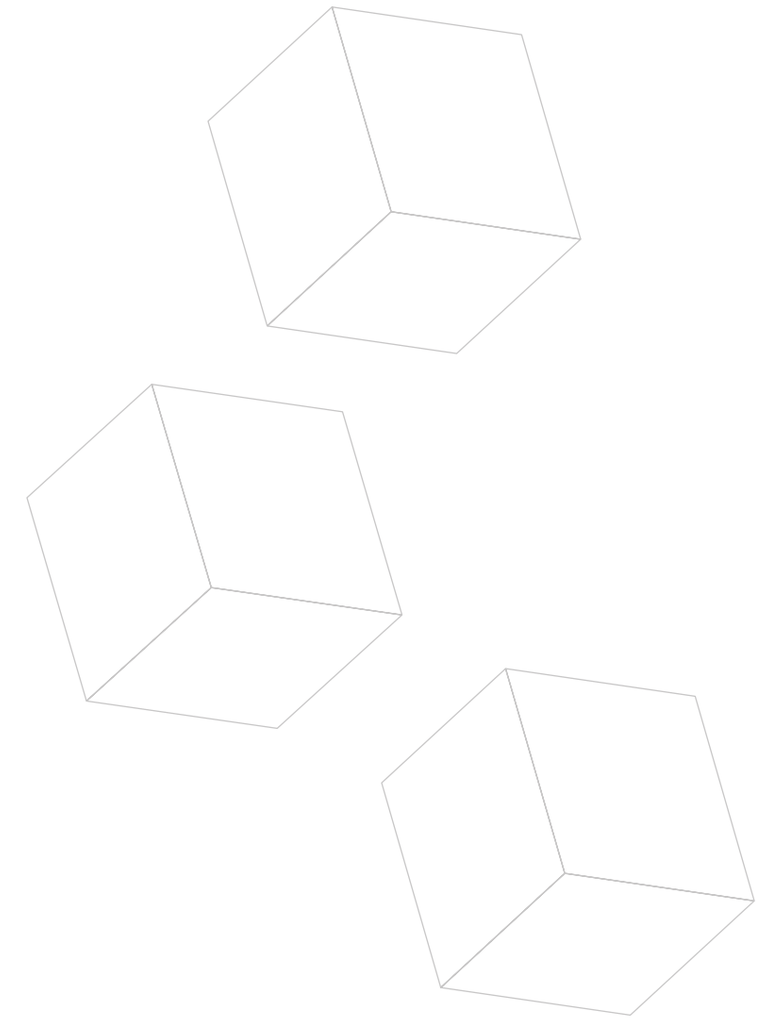
- Examples

  - Entity representations for search, discovering and entity aspects, generate relationship explanations, use entities and query suggestions in extremely data sparse contexts, discover latent semantics of entities and use for product search, document filtering for entities (head or torso or long tail)

# References

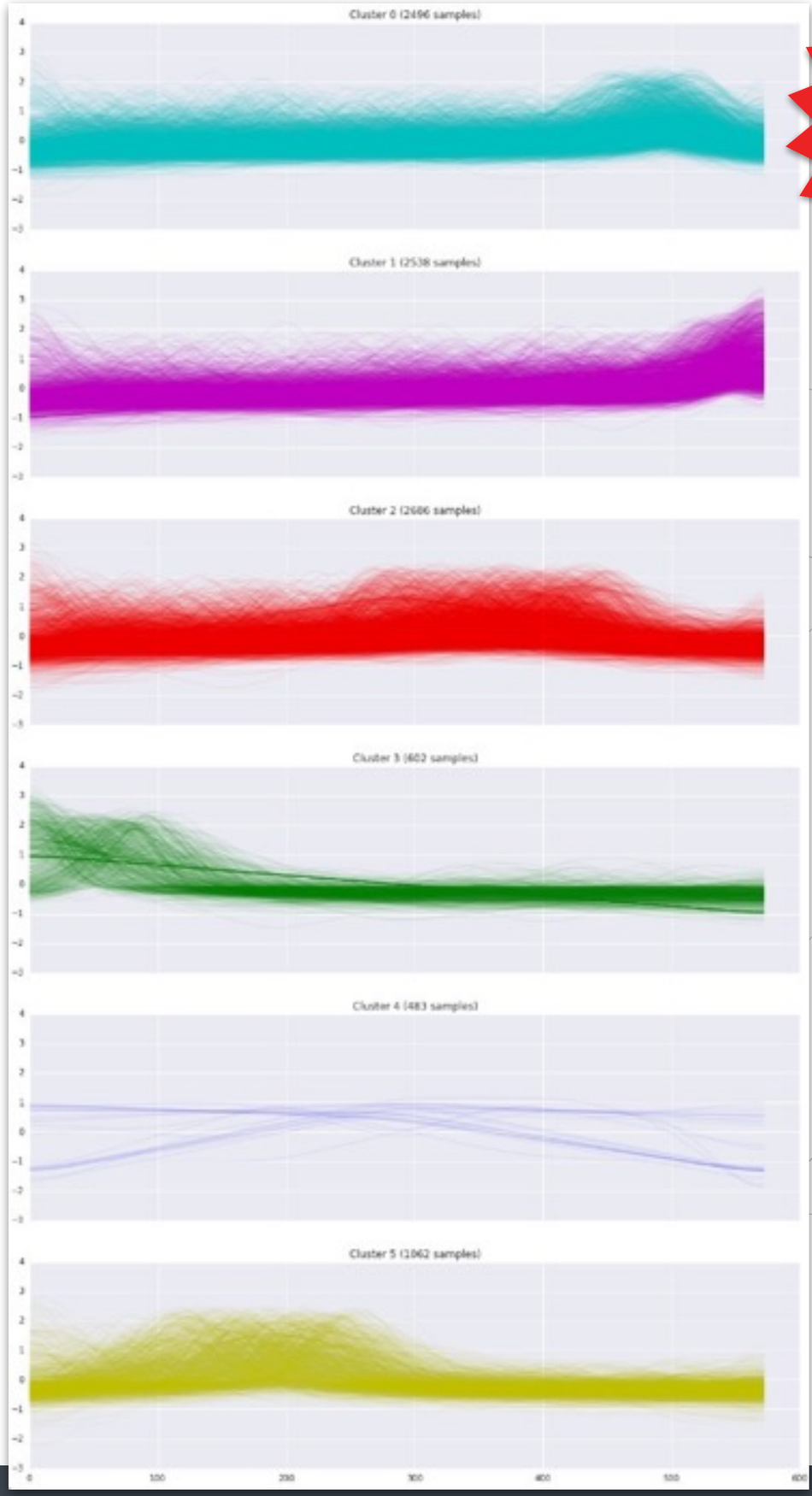

- D. Graus, M. Tsagkias, W. Weerkamp, E. Meij, M. de Rijke. Dynamic collective entity representations for entity ranking. In WSDM 2016
- X. Li, B. Schijvenaars, M. de Rijke. Investigating queries and search failures in academic search. Submitted, 2016
- R. Reinanda, E. Meij, M. de Rijke. Mining, ranking and recommending entity aspects. In SIGIR 2015
- R. Reinanda, E. Meij, M. de Rijke. Document filtering for long tail entities. Submitted, 2016
- C. Van Gysel, E. Kanoulas, M. de Rijke. A latent semantic model for product search. Submitted, 2016
- C. Van Gysel, M. de Rijke, M. Worring. Unsupervised, efficient and semantic expertise retrieval. In WWW 2016
- N. Voskarides, E. Meij, M. Tsagkias, M. de Rijke, W. Weerkamp. Learning to explain entity relationships in knowledge graphs. In ACL 2015
- N. Voskarides, E. Meij, M. de Rijke. Generating descriptions of entity relationships. Submitted, 2016

# Can’t get enough?



# Collective memory

- Entities and their transition from first mention in public sources to inclusion in our **collective memory** (i.e., Wikipedia)
  - Align time series so that all entities have same being and end (birth to inclusion)
  - Cluster the time series
  - Analyze the clusters




Work in progress

# Sponsors

All content represents the opinion of the author(s), which is not necessarily shared or endorsed by their employer and/or sponsors.