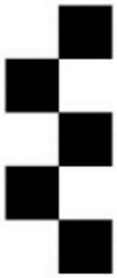
About Extreme Analyses of Texts and Graphs

Ansgar Scherp

@Signal AI
5 September 2019



University of Essex

Natural Language and Information Processing @ ESSEX

- NLP is a key research area, with focus on text processing, retrieval and formal semantics
- 40+ years at Essex University
- Interdisciplinary: Language and Computation
- Data: UK Data Archive, largest in social science

  - **Professor Ansgar Scherp**
  - **Dr Alba Garcia Seco De Herrera**
  - **Dr Jon Chamberlain**
  - **Dr Chris Fox**

Text Mining & Graph Mining

Extreme multi-label text classification [JCDL18,KCAP17]

Document recommender and retrieval [UMAP18, JCDL17]

Retrieval of documents [ICADL18]

→ Extreme # of labels

→ Extremely sparse input

Indexing graph data for search [ESWC14, JWS12, …]

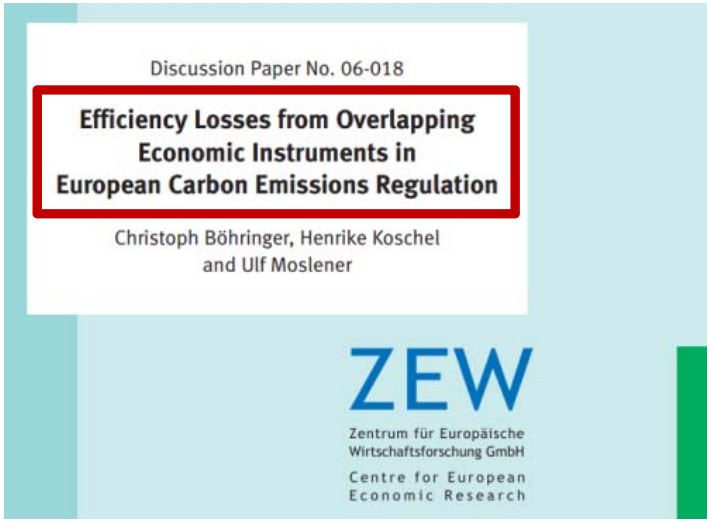
Recommending terms for data modeling [ESWC16]

Analysing graph evolution [ESWC18, ICSC18,WI17,ISWC15]

→ Extremely large and dynamic graphs

- Text and graphs are often considered in isolation
- Use cases often require a combination of both

Extreme Multi-label Classification



Task: select ~5 out of ~6,000 candidates

Standard Thesaurus for Economics

Climate protection (19481-5)
Environmental tax (18072-6)
Europe (16815-3)
...

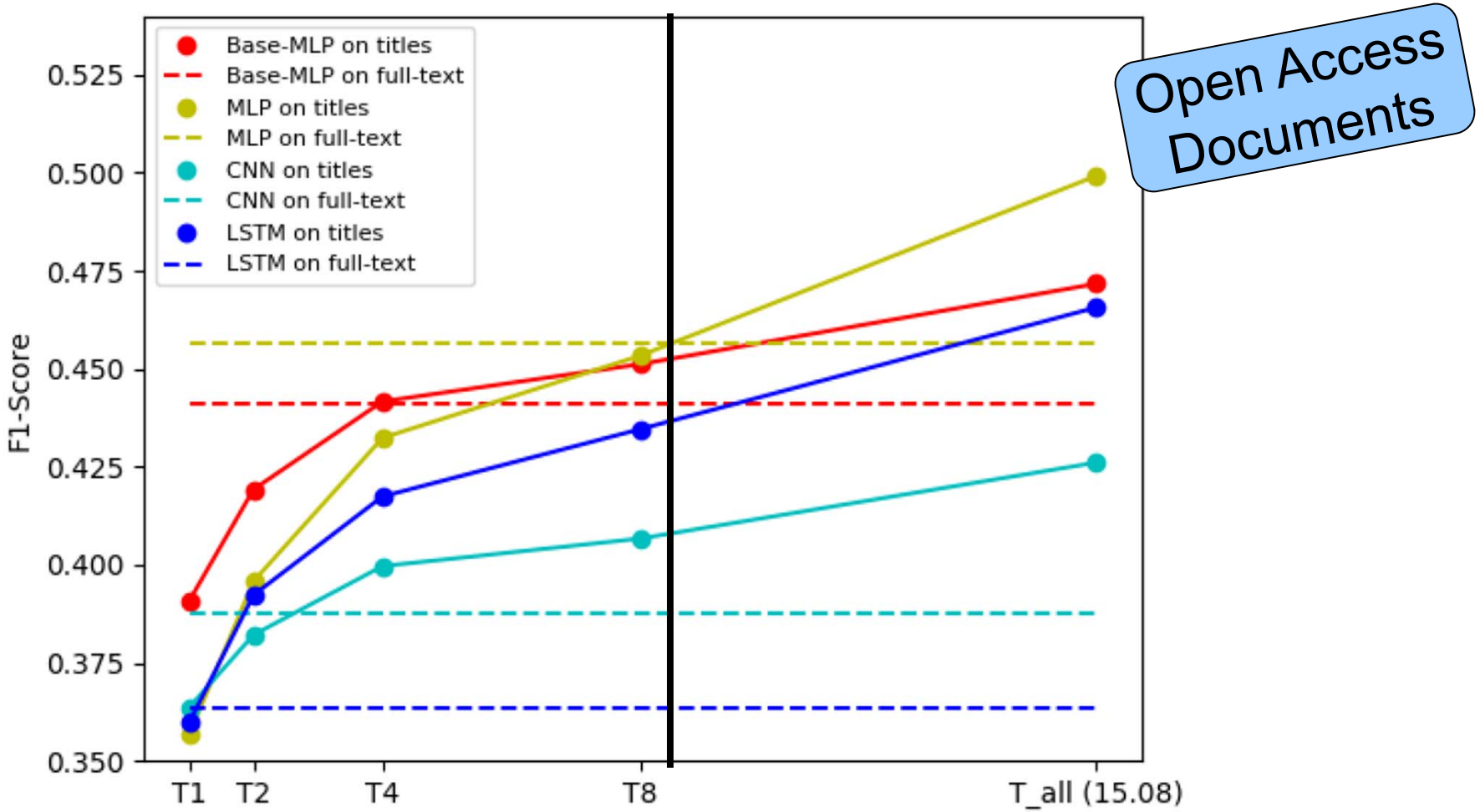
F1-score *MLP*+ on **economics**: **0.502**

F1-score *MLP*+ on **PubMed**: **0.515**

- Extreme multi-labeling task: select k labels from a very large set of n candidates, i.e., $k \ll n$
- Titles are competitive to full text (>90% relative performance) [KCAP17+15]
- Deep Learning + many titles even exceed full-text [JCDL18, Preprint: https://arxiv.org/abs/1801.06717]

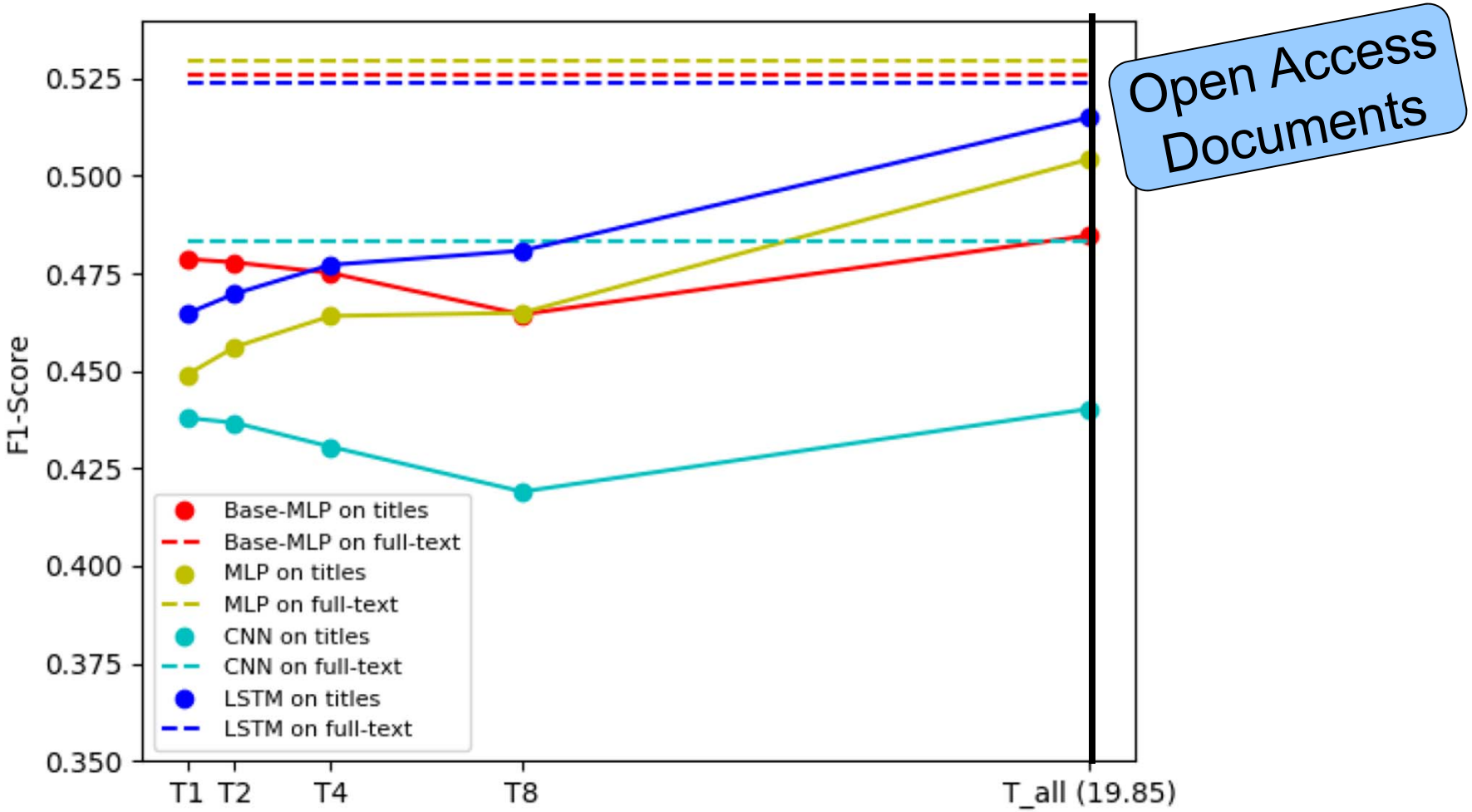
Extreme Multi-label Classification [JCDL18]

- Iterative increase of training dataset size for EconBiz
- Number of full texts: ~71k, number of titles: ~1,1 Mio

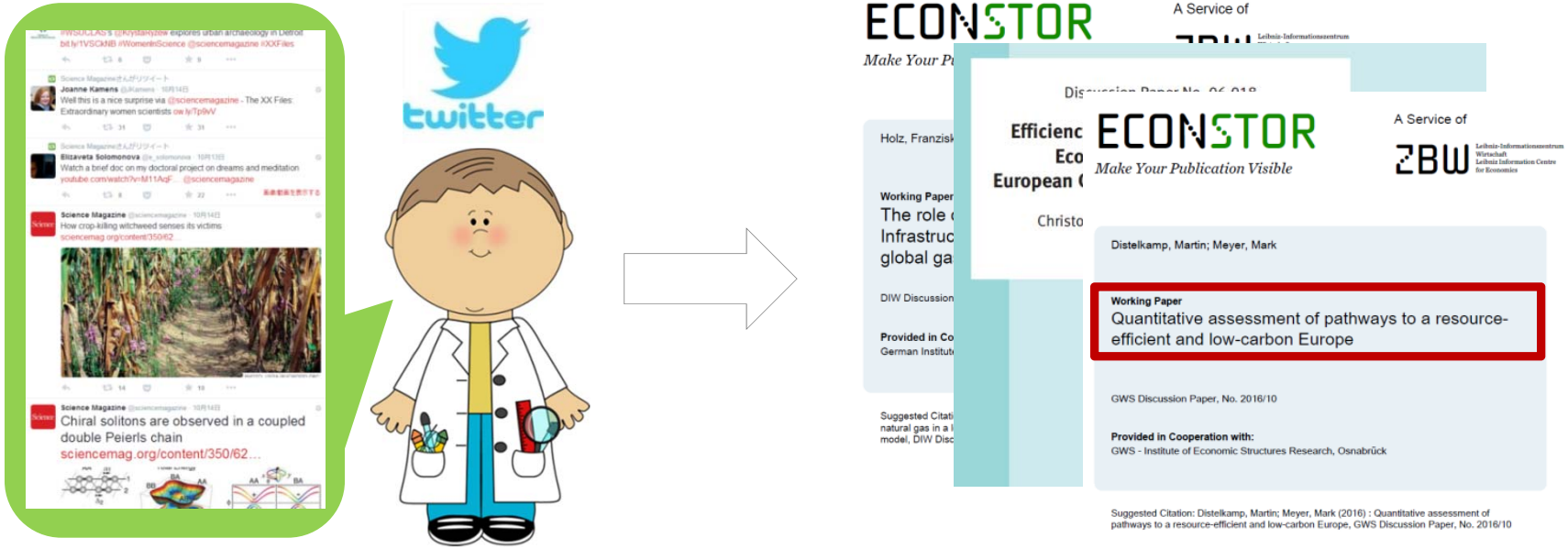


Extreme Multi-label Classification [JCDL18]

- Iterative increase of training dataset size for PubMed
- Number of full texts: ~647k, number of titles: ~13 Mio



Recommendations with HCF-IDF



- HCF-IDF is a novel extension of TF-IDF by an hierarchical thesaurus and spreading activation [JCDL17]
- Currently implemented at Kyoto University Library
- Further improvements using neural networks by re-ranking the top-5 recommendations [INF17]

Recommendations with HCF-IDF [JCDL17]

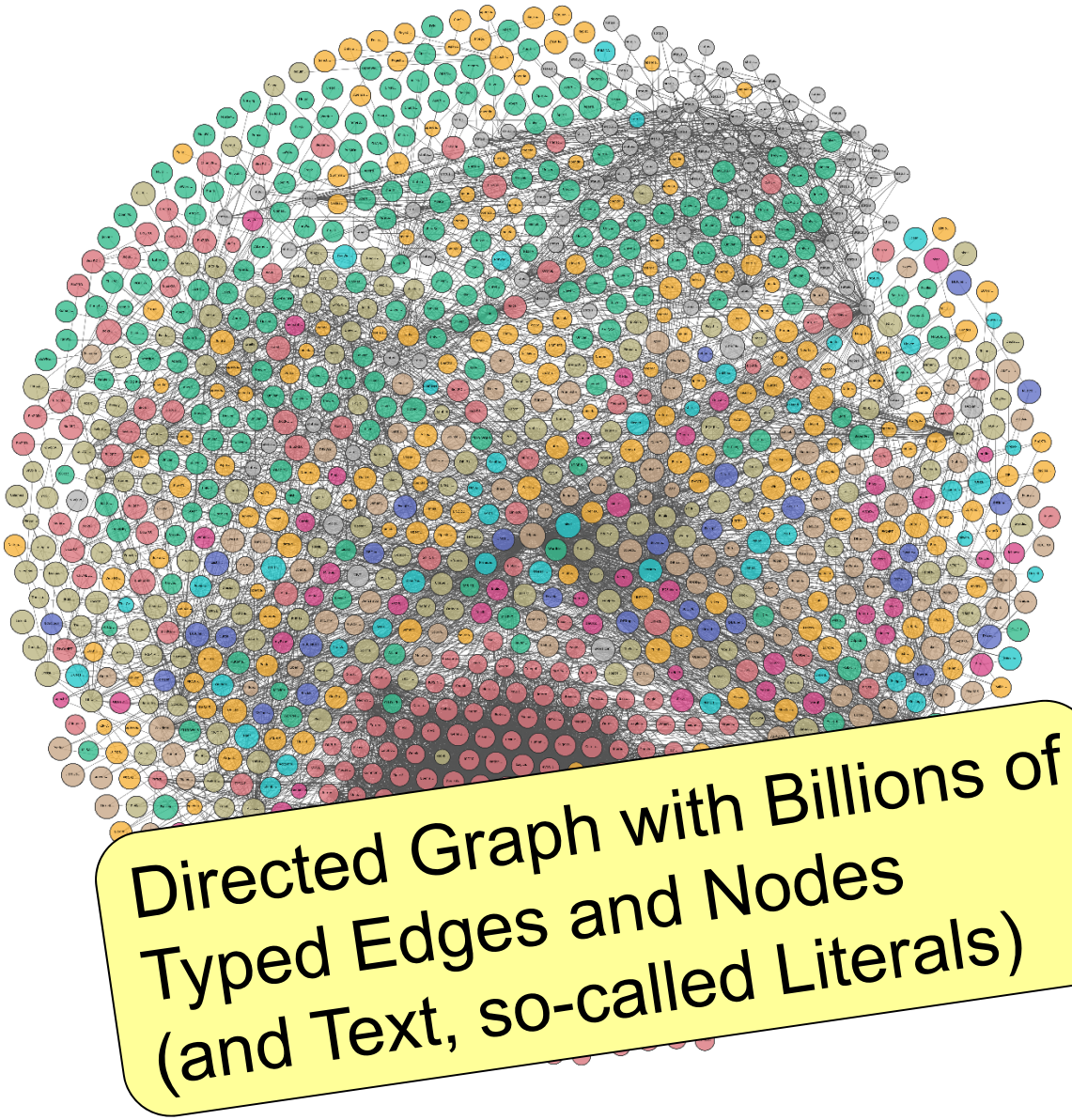
	Strategy			Rankscore
	Profiling Method	Decay Function	Con-tent	M (SD)
1.	CF-IDF	Sliding window	All	.59 (.33)
2.	HCF-IDF	Sliding window	All	.56 (.34)
3.	HCF-IDF	Sliding window	Title	.55 (.33)
4.	HCF-IDF	Exponential decay	Title	.52 (.30)
5.	CF-IDF	Exponential decay	All	.51 (.32)
6.	HCF-IDF	Exponential decay	All	.49 (.30)
7.	CF-IDF	Exponential decay	Title	.41 (.29)
8.	CF-IDF	Sliding window	Title	.39 (.27)
9.	LDA	Exponential decay	Title	.35 (.31)
10.	LDA	Sliding window	Title	.33 (.31)
11.	LDA	Exponential decay	All	.32 (.30)
12.	LDA	Sliding window	All	.27 (.33)

- User study: 12 strategies assessed by 123 economists
- Best: CF-IDF×Sliding window×All has rankscore=0.59
- No significant difference to strategies using HCF-IDF

Further Current Works

- Recommendations with Autoencoder [UMAP18]
  - Generic framework for recommendation tasks like additional citations, subject labels, …
- Sentence Embeddings [ICLR ’19]
  - First efficient learning algorithm for the Continuous Matrix Space Model
  - Formal model for word-embeddings (CBOW) and sentence embeddings (CMOW)
    → Hybrid-CBOW-CMOW model
  - Efficient training: CMOW as fast as CBOW

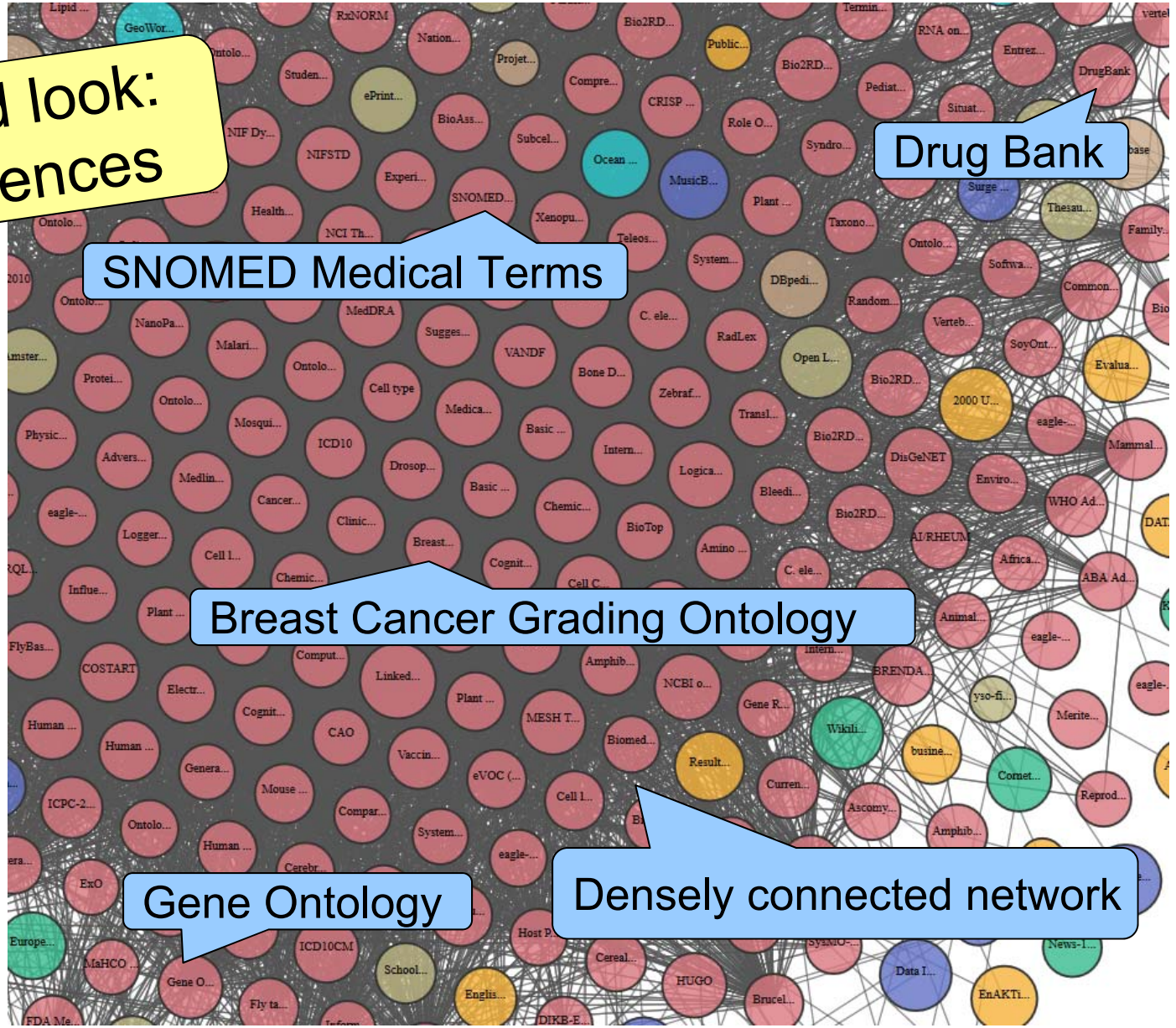
Web Graph: Linked Open Data Cloud



November 2018
http://lod-cloud.net

Webgraphen: Linked Open Data Cloud

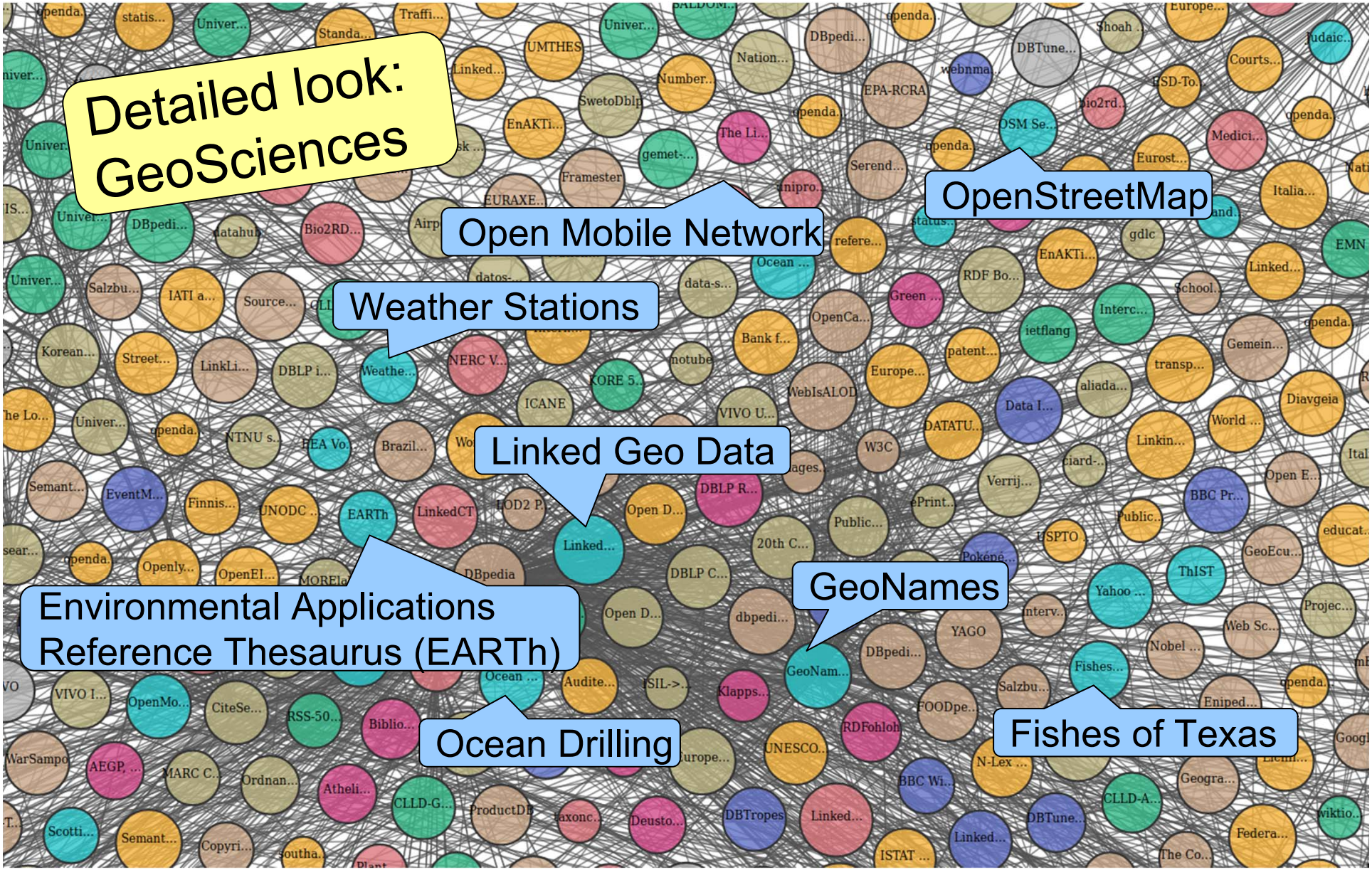
Detailed look:
Life Sciences



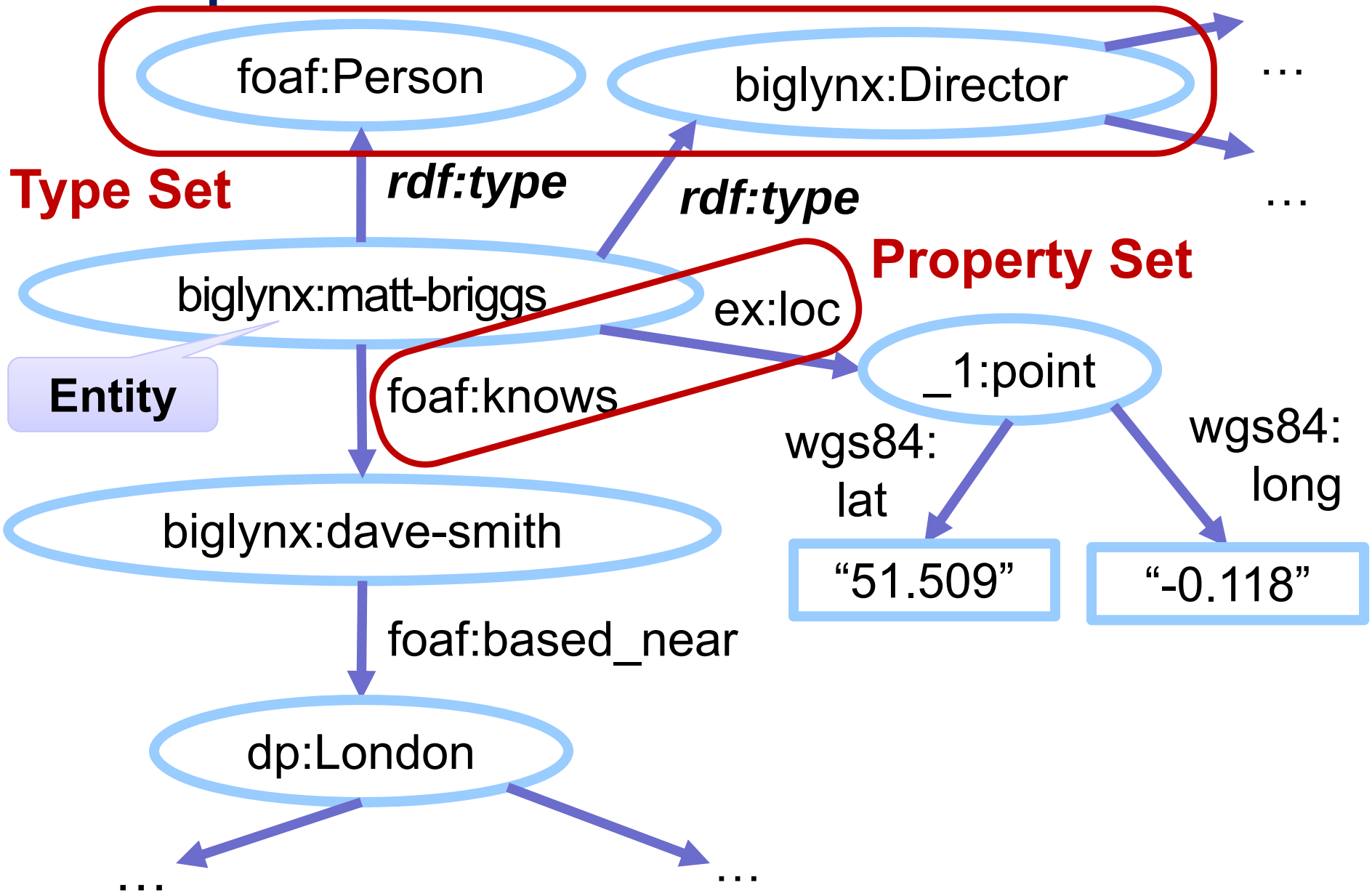
- Cross Domain
- Geography
- Government
- Life Sciences
- Linguistics
- Media
- Publications
- Social Networking
- User Generated

Letzte Aktualisierung:
Mai 2018
Quelle:
http://lod-cloud.net

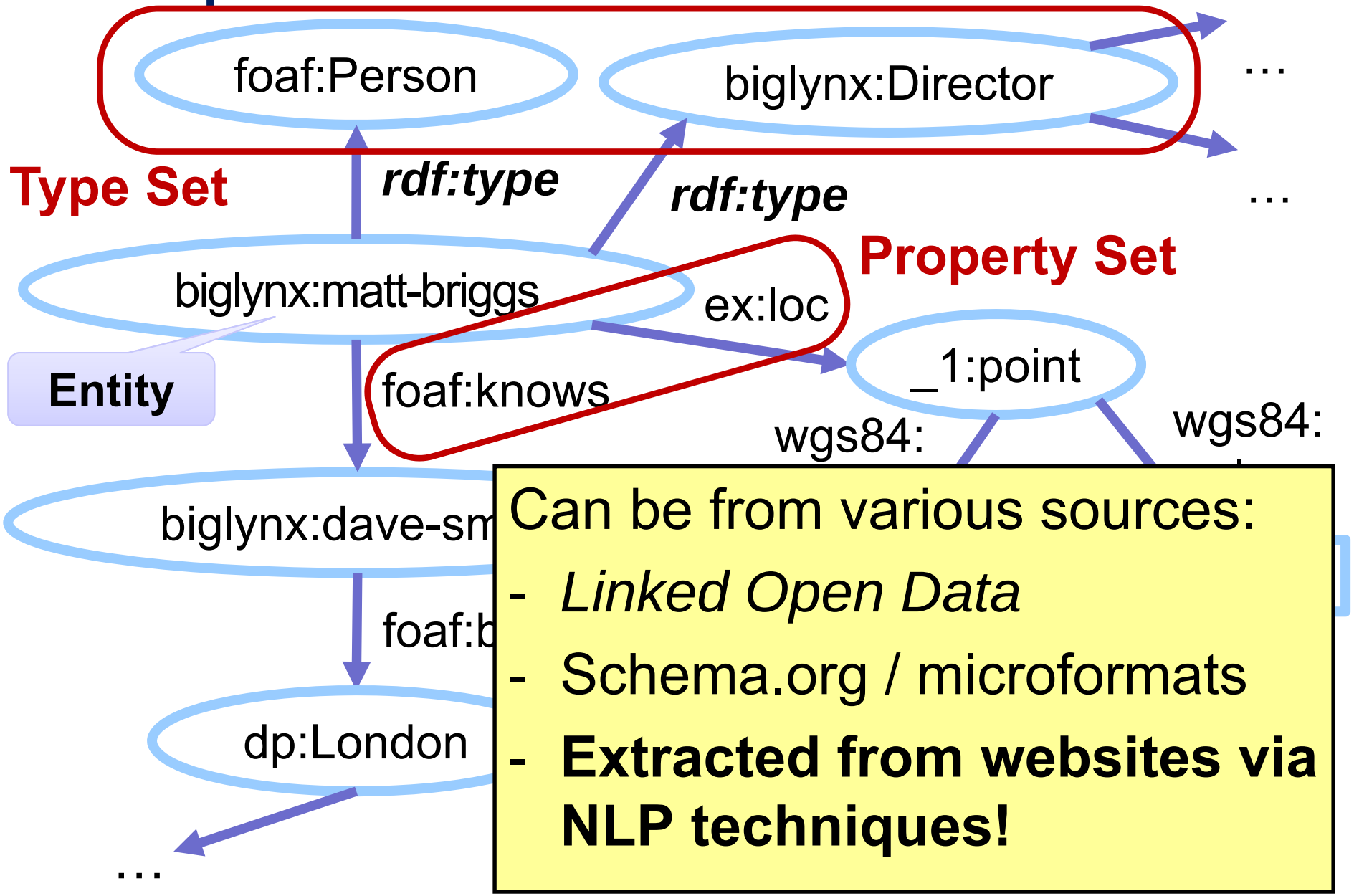
Webgraphen: Linked Open Data Cloud



Example: Structured Data on the Web

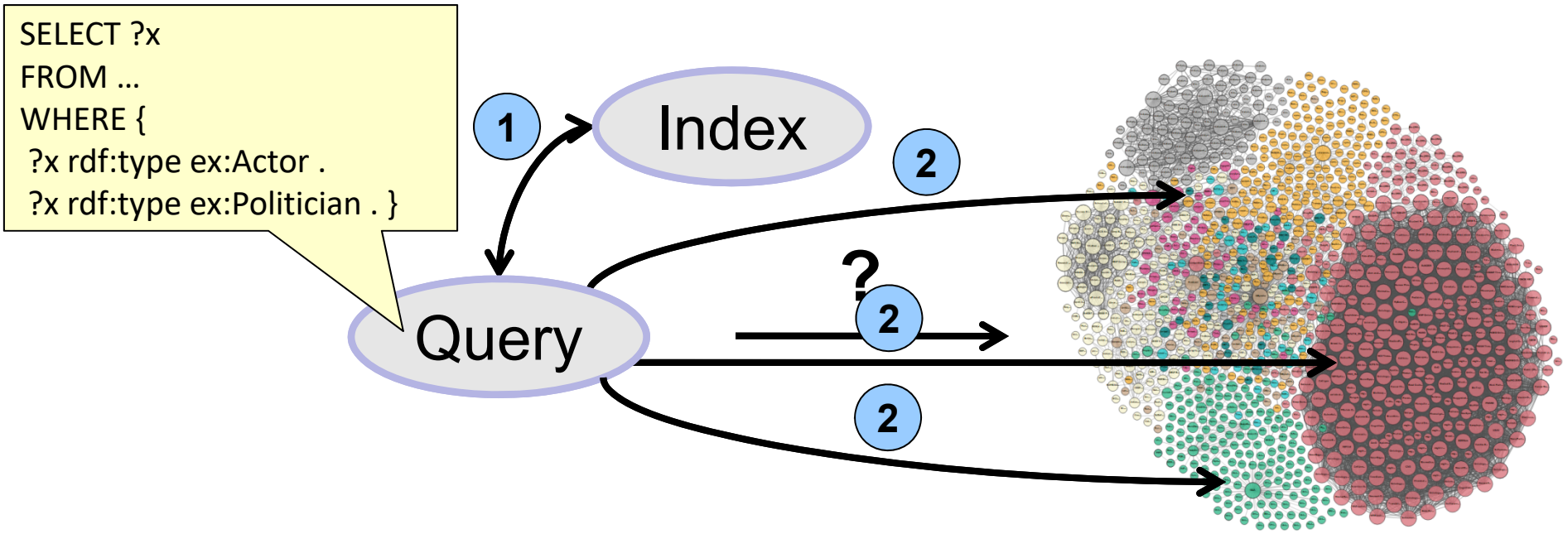


Example: Structured Data on the Web



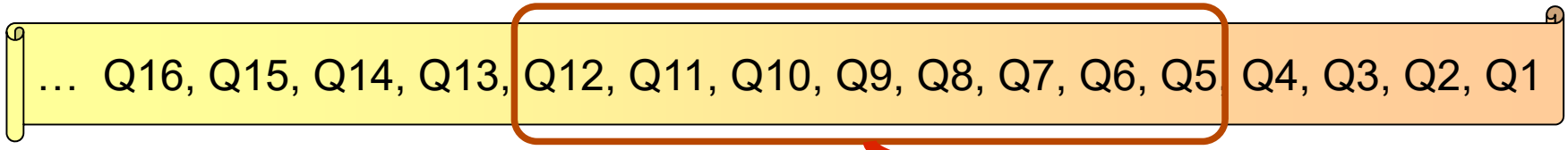
Indexing of Linked Open Data [JWS12]

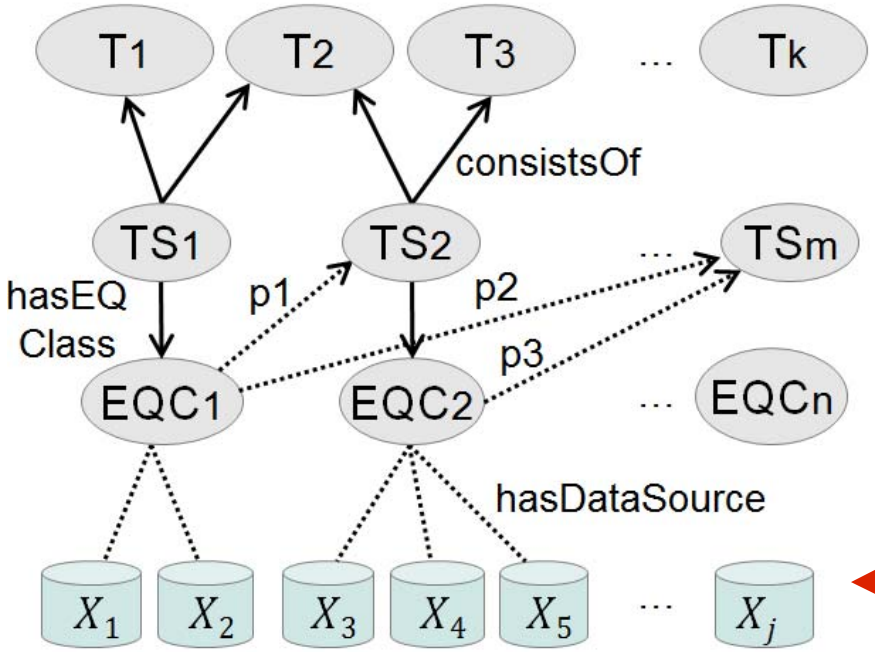
- Single point of entry needed to query the data
- Search for data sources containing entities like
  - ‘Find sources of scientific publications in medicine’
  - ‘Research data sets in genetics’
  - …



Indexing of Linked Open Data [JWS12]

- Stream of graph data coming from a crawler





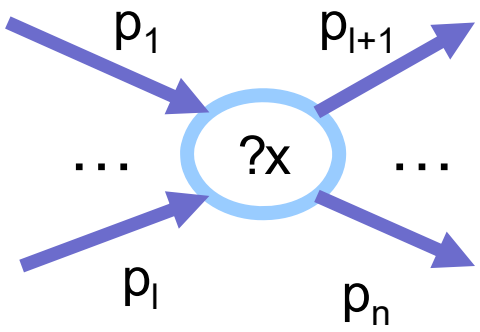
+ Reasonable accuracy at cache size of
+ Linear runtime with respect to num
+ Memory consumption scales with w

Billion Triple Challenge
Winner of 2011

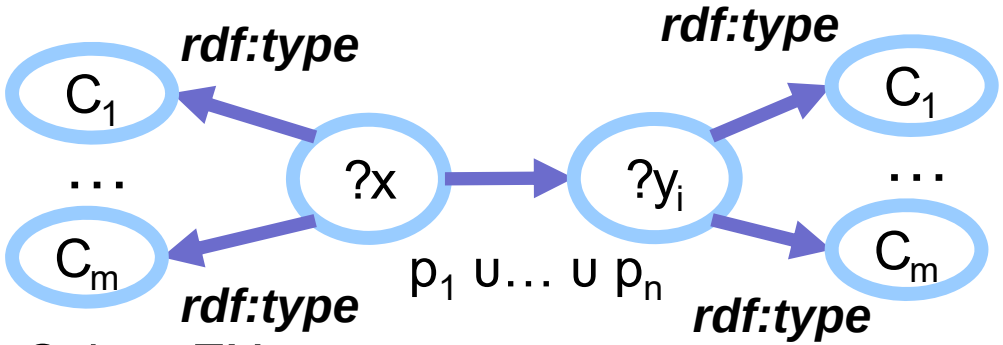
Metamodel for Graph Indices

[FGDB18, GvDB18]

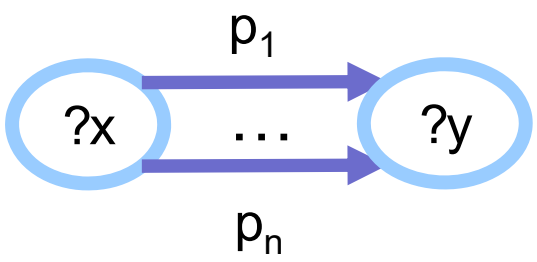
Characteristic Sets
(Neumann & Moerkotte, ‘11)



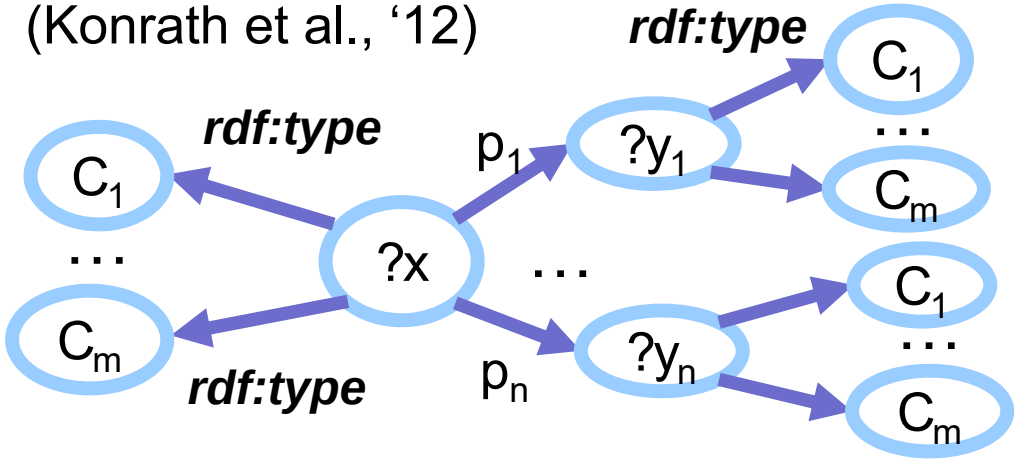
TermPicker’s Schema-level Pattern
(Schaible et al., ‘16)



SemSets
(Ciglan et al., ,12)

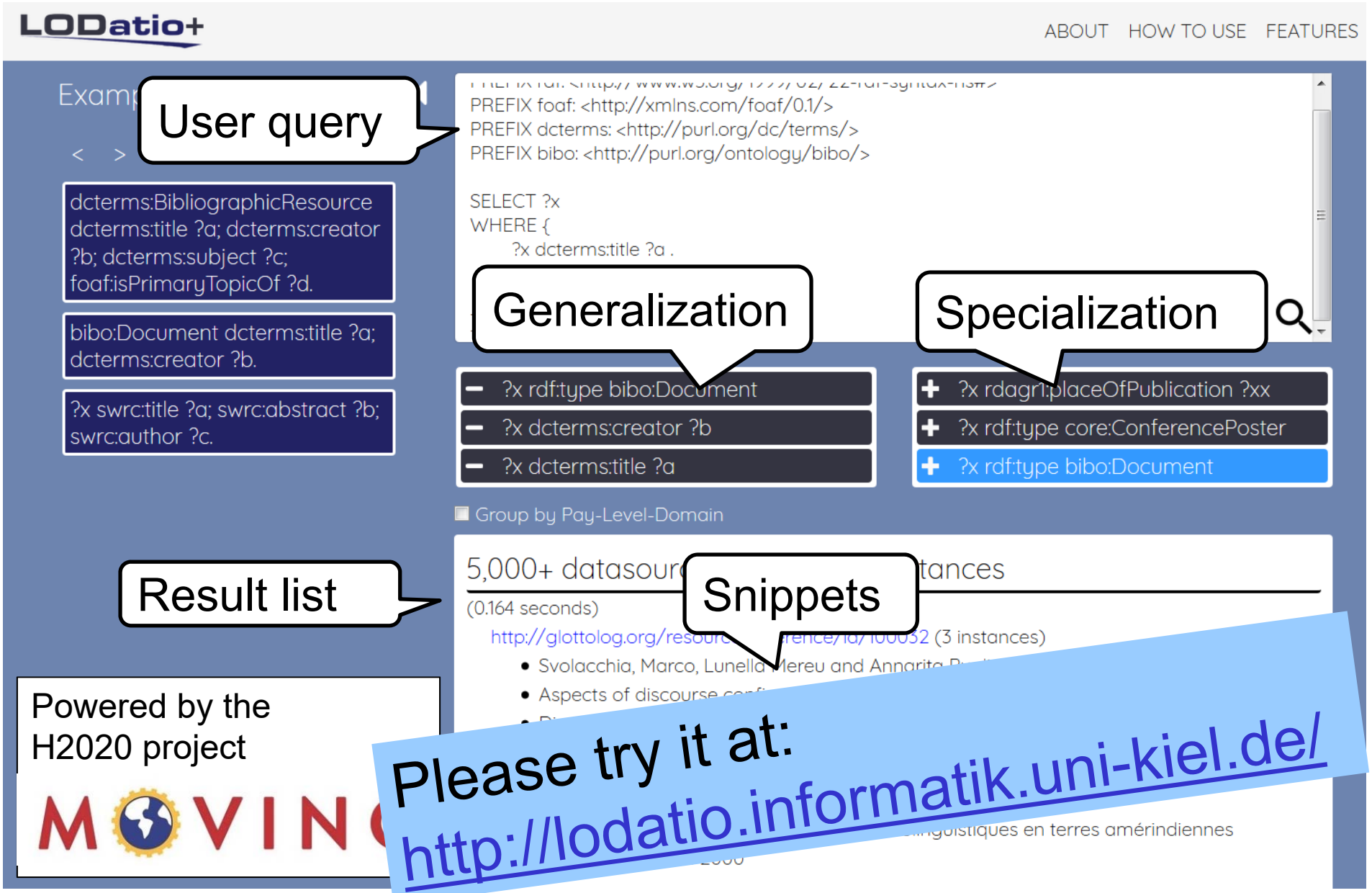


SchemEX
(Konrath et al., ‘12)



- Existing indices define a single, fixed data structure
- FLuID: Formal model to flexibly define graph indices

Data Search Engine LODatio+ (2018)

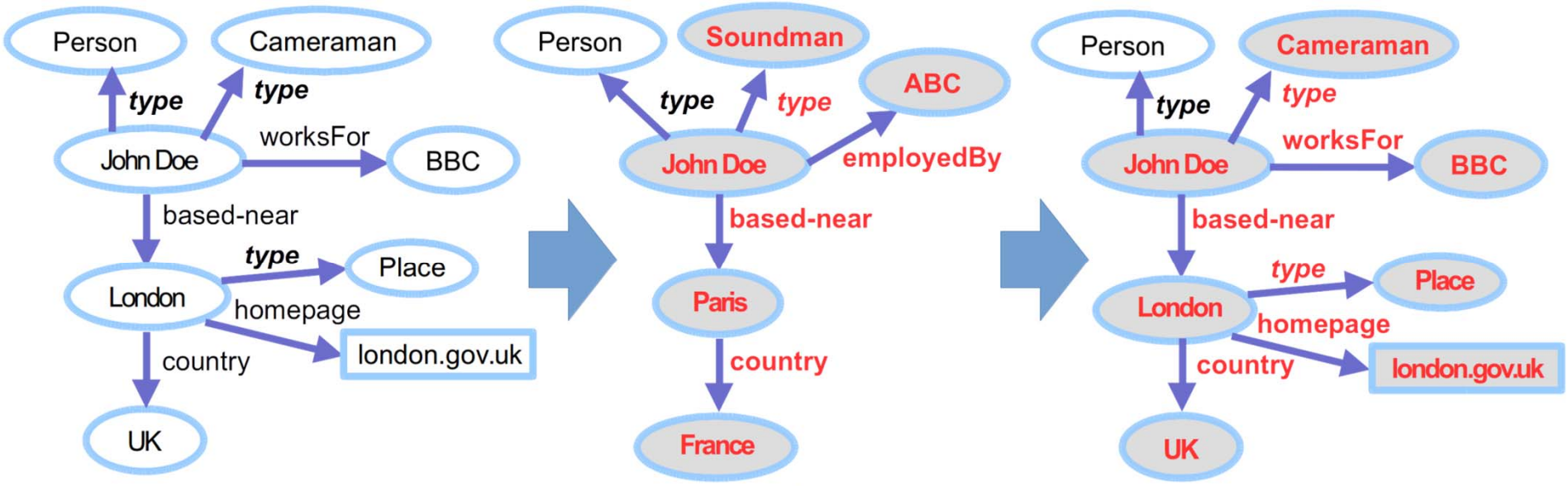


Temporal Dynamics of the Entities?

- Can we predict when the Open Data will change?
- Useful for any application that wants to use Open Data

Essential for EY Building a better working world business

- Notion of *entity* evolution: set of triples X sharing the same subject URI s (here: John Doe)

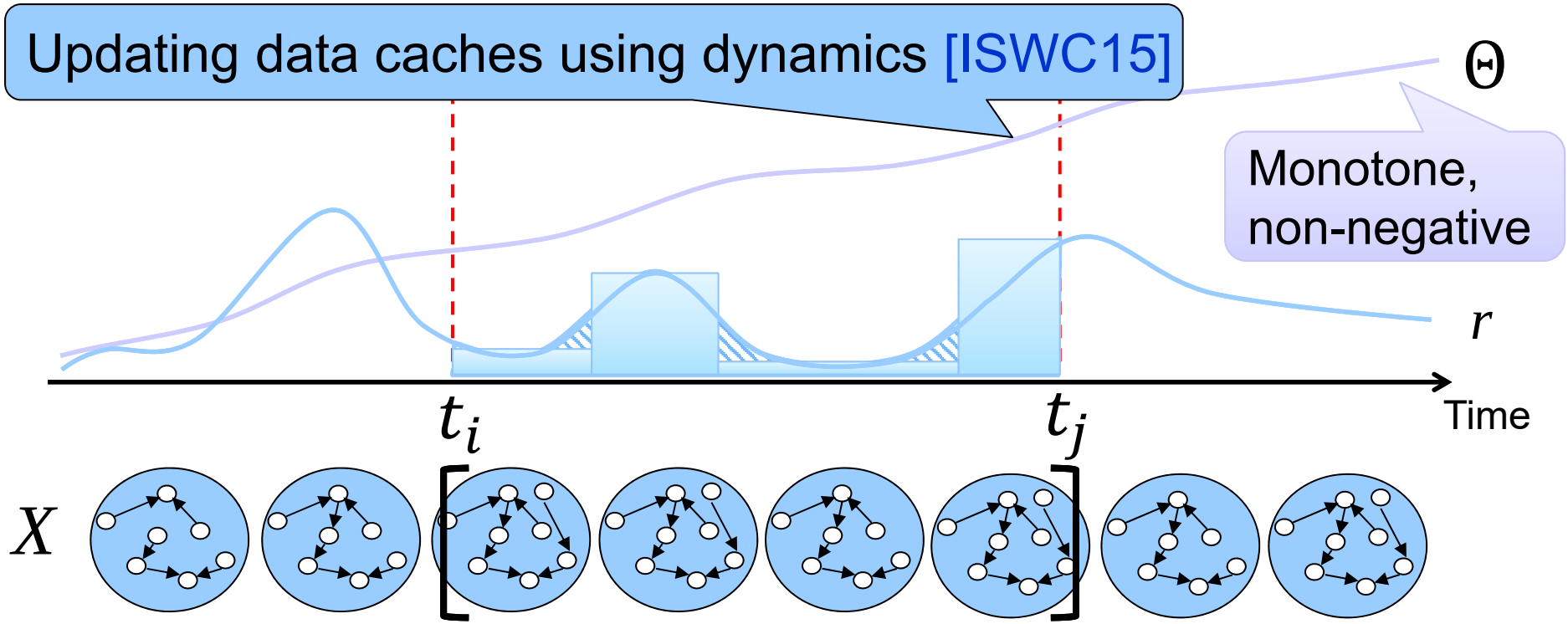


Dynamics Function Θ [ISWC15]

- Definition of Θ over change rate function $r(X_t)$

$$\Theta_{t_i}^{t_j}(X) = \Theta(X_{t_j}) - \Theta(X_{t_i}) = \int_{t_i}^{t_j} r(X_t)\mathrm{d}t \approx \sum_{k=i+1}^{j} \delta(X_{t_{k-1}}, X_{t_k})$$

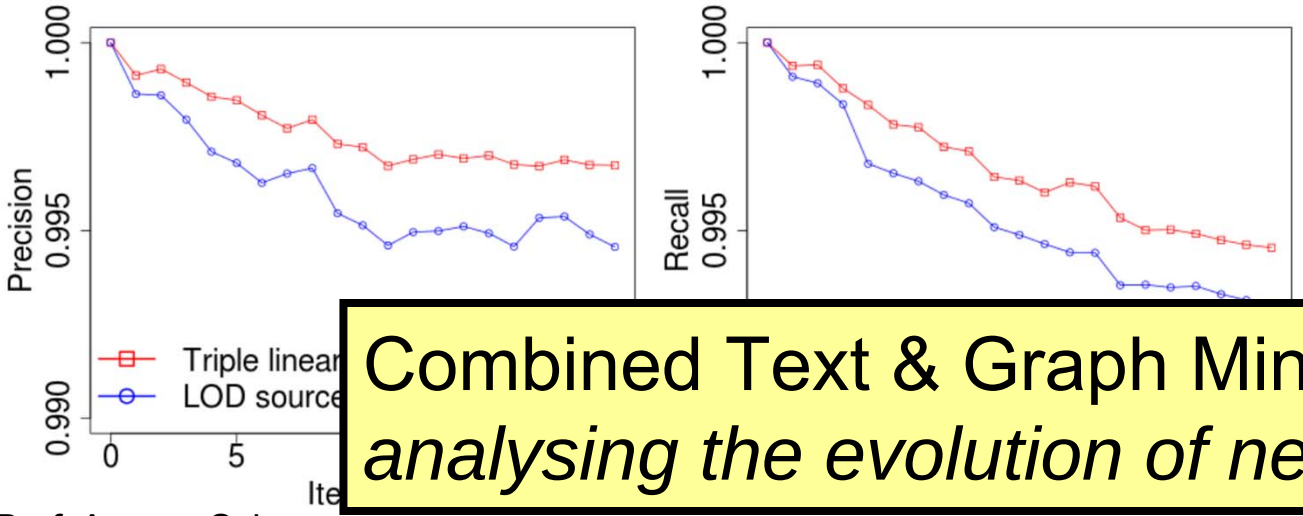
- Approximation as step function over changes



Prediction of Triple Lifetime [WI17]

- Idea: subgraphs that share common features may have the same dynamics
- Features: subject pay-level domain (PLD), predicate type, object PLD, and object data type
- Approach: regression over one-hot feature encoding
- Compute crawling priority for RDF document $X_{c,t}$

$$score(c, t_i) = \left(\frac{1}{|X_{c,t}|} \sum_{x \in X_{c,t}} LR(x) \right)^{-1}$$



- Novel strategy using predicted lifetime [WI17] outperforms the [ISWC15]

Combined Text & Graph Mining for *analysing the evolution of networks*.

Interesting Topics for Collaboration

- Extreme text mining?
- Extracting and indexing entities from news?
- Tracking entities over time? Event detection?
- …

Thank you!