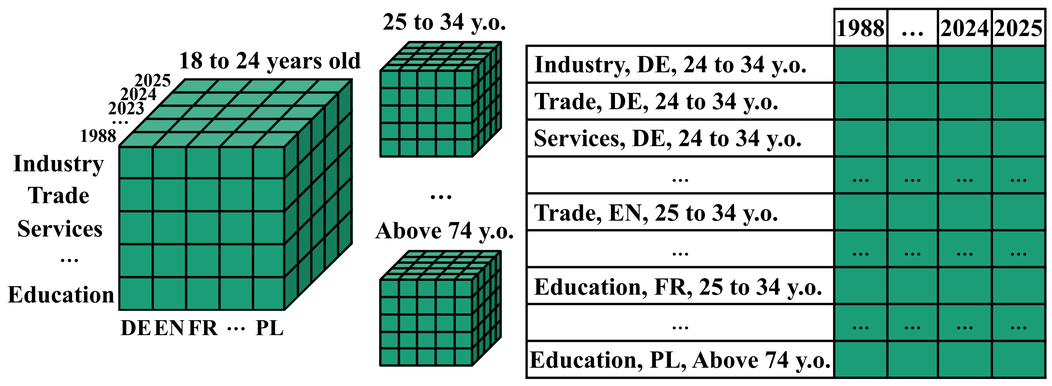
Efficient and Scalable Search for Statistics

Antoine Gauquier Simon Ebel Helena Galhardas Théo Galizzi
Ioana Manolescu Aurélien Peden Pierre Senellart

Statistical vs. Relational Tables

Statistical tables represent **multidimensional data**: each metric value is defined over d dimension values (year, country, sector…) and *serialized* into 2D.



Challenges: (i) Semantics in **headers** (cells are mostly numerical); (ii) High dimensionality → very **large tables**; (iii) **Context** for dimension values often missing.

SoTA Table Discovery (TD) systems target **relational** tables: ill-suited for statistical data. We define the **Statistical Table Discovery (STD)** problem → TD over statistical tables.

The Space- and Time-Aware Statistic Retrieval Approach (STAR)

At indexing time (input = corpora of statistical tables):

1. Indexes only **linguistic elements** (titles, headers); ignores numerical values
2. Extracts **spatio-temporal entities**, indexed separately:
   - *Space*: DAG via external reference sources (e.g., *GeoNames*)
   - *Time*: natural hierarchy via regular expressions
3. Projects linguistic elements into a **latent space** (LM); stores in a vector index

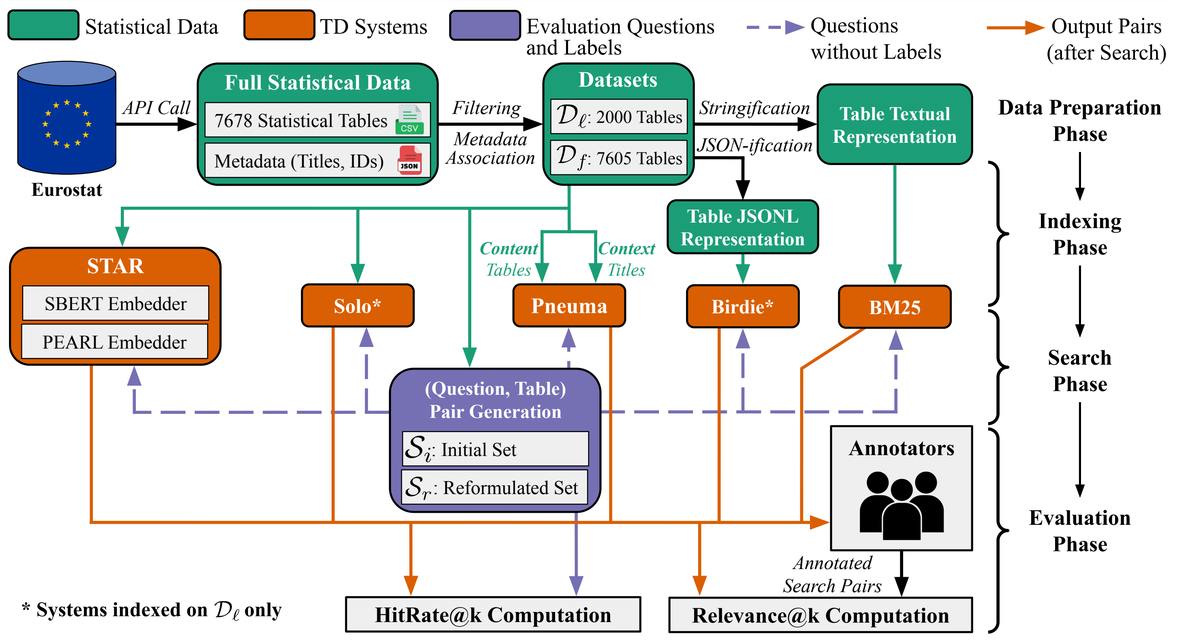
At query time (input = Natural Language (NL) question):

1. Parses the NL question → **extract spatio-temporal entities**
2. Computes tables matching the spatio-temporal constraints
3. Ranks by **semantic similarity** with the question embedding
4. If no match: constraints **progressively relaxed** until $k \geqslant 1$ results

Systems & Characteristics

	BM25	BM25+Syn.	Solo	Birdie	Pneuma	STAR-SBERT	STAR-PEARL
Type	Lexical	Lexical	Semantic	Semantic	Hybrid	Semantic	Semantic
Task	Baseline	Baseline	TD	TD	TD	STD (ours)	STD (ours)

Pipeline



Dataset & Questions Sets

Dataset: 7 605 *Eurostat* CSV tables; $d \geqslant 1$ dimensions (always time; mostly space). Subset of 2 000 smallest tables to evaluate systems that cannot scale to $\mathcal{D}_f$.

Dataset	#Tables	Size (MB)	#rows	#cols
$\mathcal{D}_\ell$	2 000	30.3	111 (± 101)	19 (± 23)
$\mathcal{D}_f$	7 605	98 521.7	69 077 (± 318 706)	41 (± 270)

Question sets (200 questions, ∼5 000 annotated pairs):

- $\mathcal{S}_i$: lexically close to table content
- $\mathcal{S}_r$: same semantics, maximal lexical variation

Running Time

Model	Indexing (s) $\mathcal{D}_\ell$	Indexing (s) $\mathcal{D}_f$	Query (s) $(\mathcal{D}_\ell, \mathcal{S}_i)$	Query (s) $(\mathcal{D}_\ell, \mathcal{S}_r)$	Query (s) $(\mathcal{D}_f, \mathcal{S}_i)$	Query (s) $(\mathcal{D}_f, \mathcal{S}_r)$
BM25	13	7 701	1.0	0.9	41.8	45.2
BM25+Syn.	13	7 701	3.4	3.3	47.2	51.5
Solo	81 847	**OOM**	1 267.1	1 143.1	**OOM**	**OOM**
Pneuma	45 300	532 020	3 793.2	3 680.9	1 770.4	1 814.2
Birdie	24 135	**OOM**	20.8	20.3	**OOM**	**OOM**
STAR-SBERT	371	45 103	36.8	37.7	186.6	191.0
STAR-PEARL	208	44 279	29.7	34.8	106.2	116.7

Results: HitRate@k and Relevance@k