Benchmarking Filtering Techniques for Entity Resolution

ICDE 2023, Research Session 8: Information Integration and Data Quality, April 5th 2023

George Papadakis, Marco Fisichella, **Franziska Schoger**, George Mandilaras, Nikolas Augsten, Wolfgang Nejdl

Introduction

- Entity resolution
  - identifying pairs of entity profiles that represent the same real-world object
- Filtering Techniques
  - don't check every possible pair but only the most promising (candidate pairs)
- Matching

- Example:
- 1 Movie in two databases

IMDb

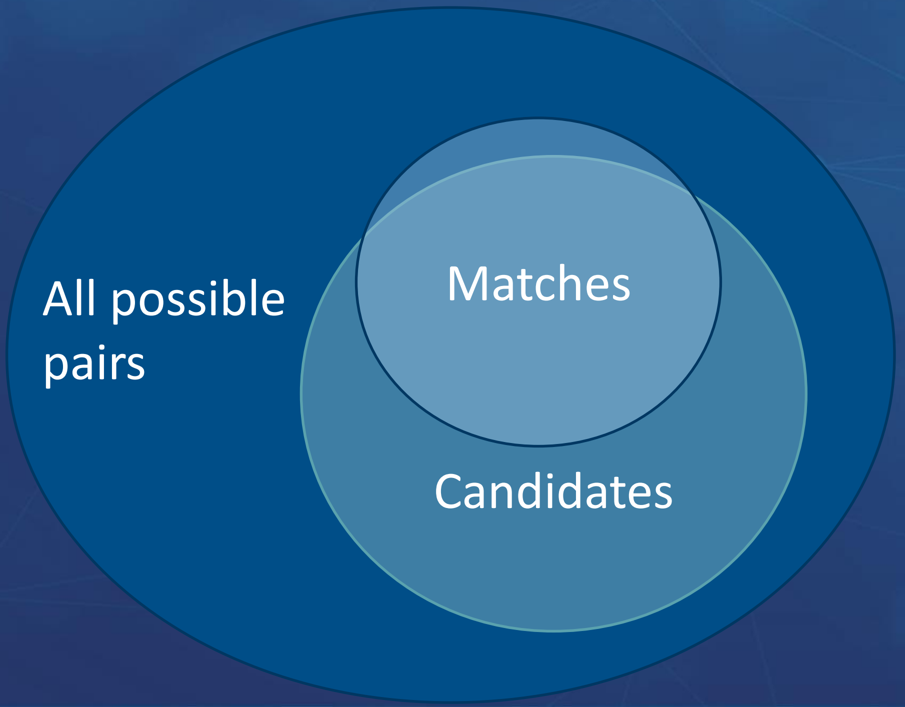
id	title	starring	writer	editor
23304		Sean Duke, Hugo Weaving, Lawrence Makoare Christopher Lee, Catherine Blanchett, Ian Holm…	Frances Walsh	

DBpedia

id	title	actor name	director name	year	genre
3831	Lord of the Rings: The Fellowship of the Ring, The (2001)	Lee, Christopher; Jackson, Peter; McKellen, Ian; Rhys-Davies, John; …	Jackson, Peter	2001	Action, Fantasy Adventure

Preliminaries

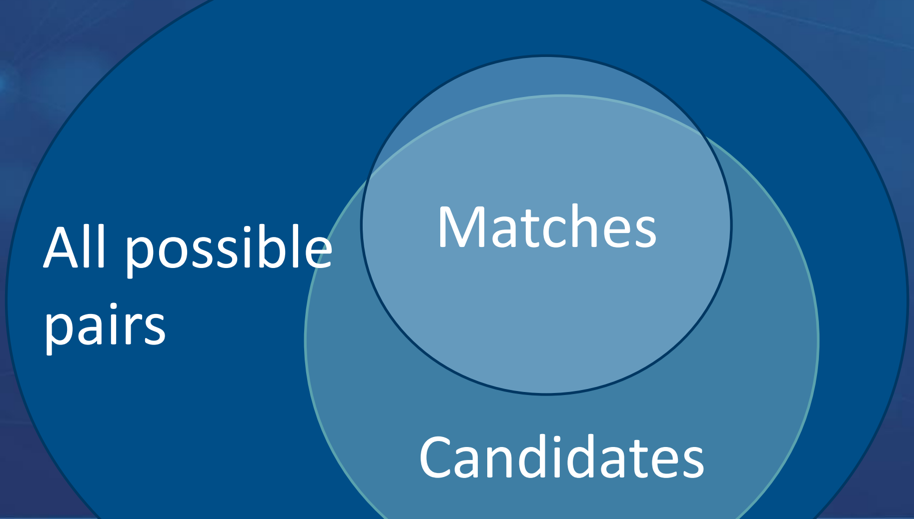
- Two main tasks of ER
  - Clean-Clean ER
  - Dirty ER



- Performance Measures
  - Pair Completeness (PC)
    - $\frac{|Candidates \cap Matches|}{|Matches|}$
  - Pairs Quality (PQ)
    - $\frac{|Candidates \cap Matches|}{|Candidates|}$
  - Runtime

Main Idea

- Filtering techniques for textual entity profiles:
  - Blocking Workflows
  - NN Methods
- Main idea: compare different filtering techniques in terms of PC, PQ and runtime

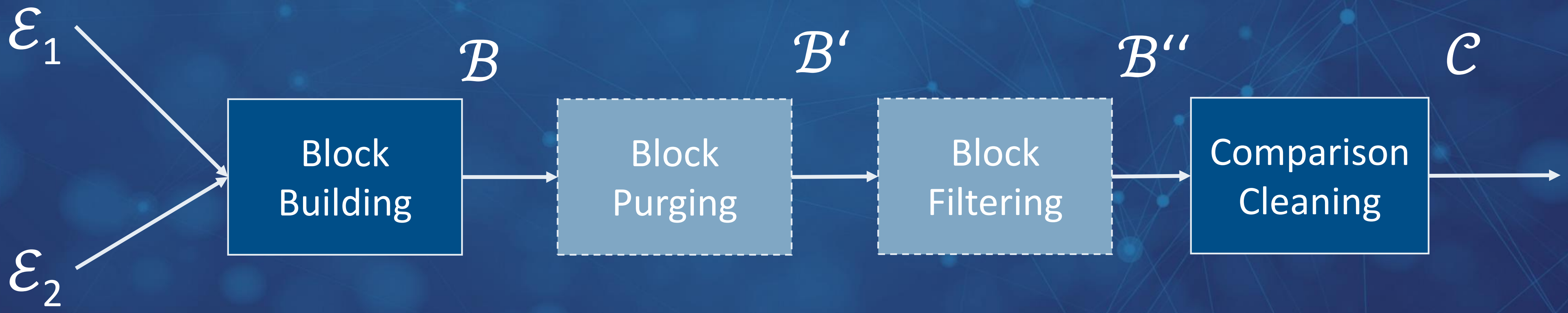


Problem (Configuration Optimization):

- Given are two sets of entity profiles, a filter method and a threshold on PC (90%)
- Finetune the parameters of the filtering methods such that the resulting PQ is maximized while the PC is above the threshold

Filtering Methods

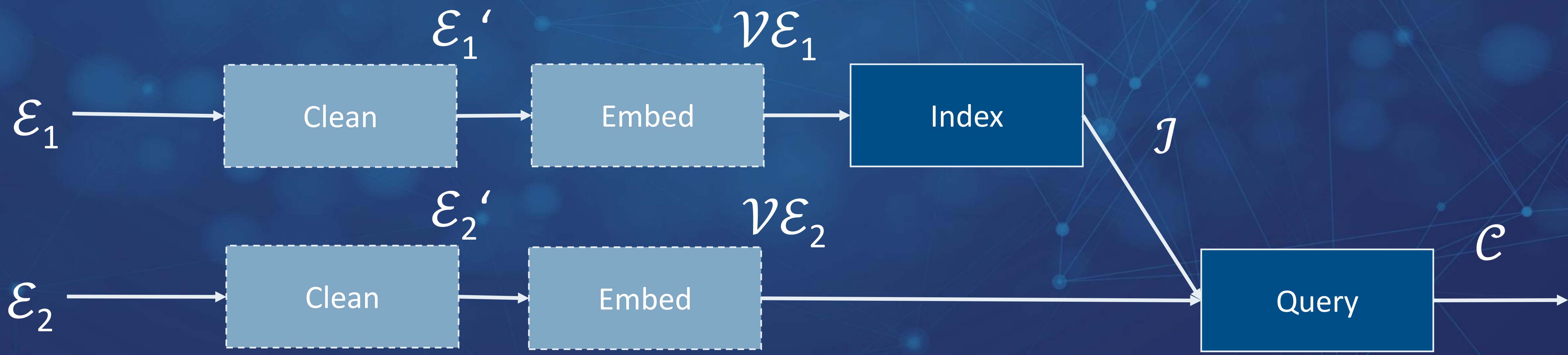
Blocking Methods



Filtering Methods

Nearest Neighbor Methods

- Sparse Vector Based NN Methods:
  - Similarity of token sets (Jaccard, Cosine, Dice)
- Dense Vector Based NN Methods:
  - Similarity of Vector Embeddings



Datasets

	D_{c1}	D_{c2}	D_{c3}	D_{c4}	D_{c5}	D_{c6}	D_{c7}	D_{c8}	D_{c9}	D_{c10}
$\mathcal{E}_1$ / $\mathcal{E}_2$	Rest. 1 / Rest. 2	Abt / Buy	Amazon / GB	DBLP / ACM	IMDb / TMDb	IMDb / TVDB	TMDb / TVDB	Walmart /Amazon	DBLP / GS	IMDb / DBpedia
# entities	339 / 2,256	1,076 / 1,076	1,354 / 3,039	2,616 / 2,294	5,118 / 6,056	5,118 / 7,810	6,056 / 7,810	2,554 / 22,074	2,516 / 61,353	27,615 / 23,182
# duplicates	89	1,076	1,104	2,224	1,968	1,072	1,095	853	2,308	22,863
Cartesian Product	$7.7 \cdot 10^5$	$1.2 \cdot 10^6$	$4.1 \cdot 10^6$	$6.0 \cdot 10^6$	$3.1 \cdot 10^7$	$4.0 \cdot 10^7$	$4.7 \cdot 10^7$	$5.6 \cdot 10^7$	$1.5 \cdot 10^8$	$6.4 \cdot 10^8$
Best Attribute	Name	Name	Title	Title	Title	Name	Name	Title	Title	Title

Configuration Space

	Method	Number of Configurations
Blocking Methods	Standard Blocking	3,440
	Q-Grams Blocking	17,200
	Extended Q-Grams Blocking	68,800
	(Ex.) Suffix Arrays Blocking	21,285
Sparse NN Methods	ε-Join	6,000
	kNN-Join	12,000
Dense NN Methods	MH-LSH	168
	HP-LSH	400
	CP-LSH	2,000
	FAISS	2,720
	SCANN	10,880
	DeepBlocker	2,720

Taxonomy

Scope		Blocking	Sparse NN	Dense NN
Syntactic Representations	Schema-based	✓	✓	✓
	Schema-agnostic	✓	✓	✓
Semantic Representations	Schema-based	-	-	✓
	Schema-agnostic	-	-	✓

Taxonomy

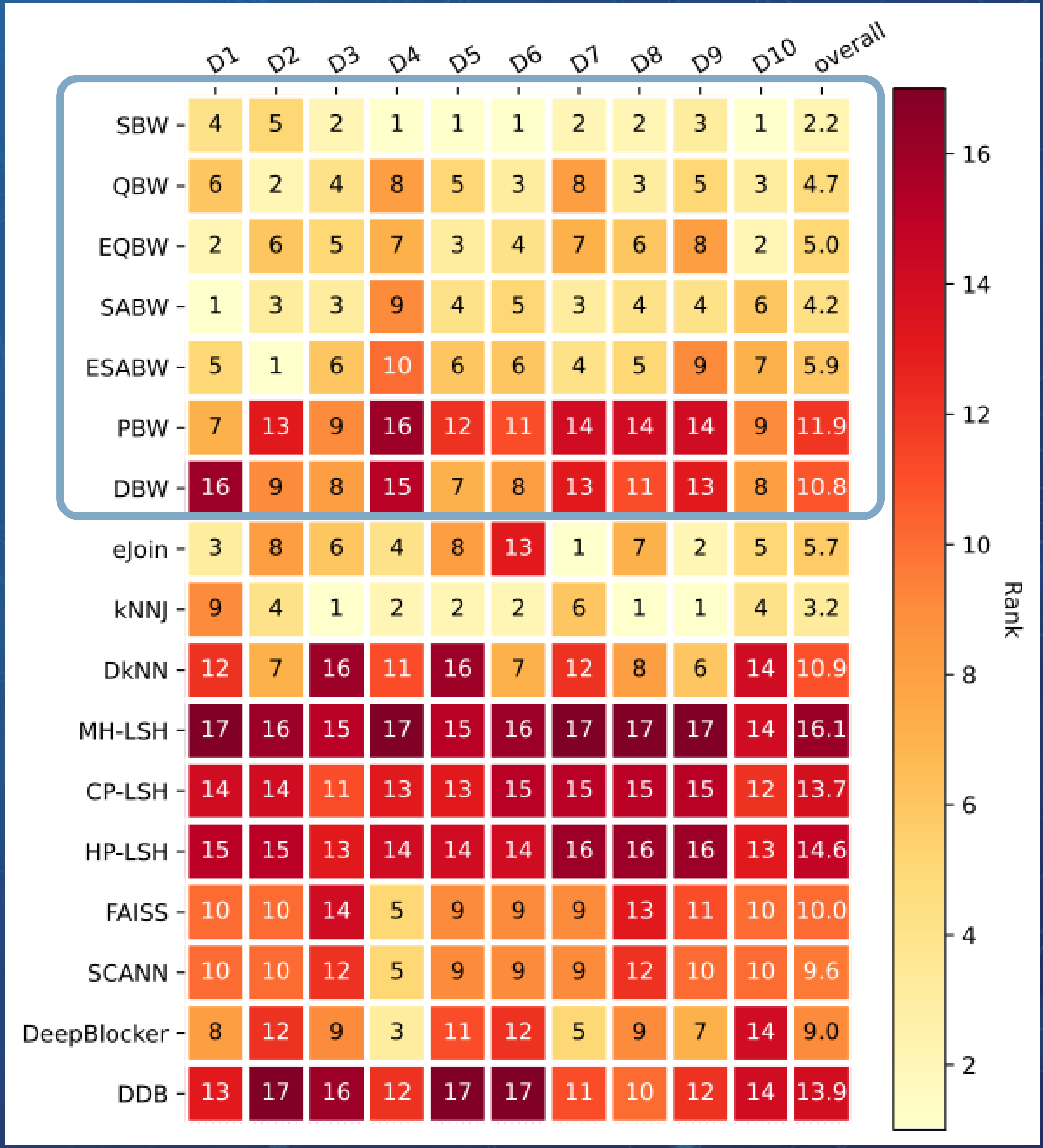
Internal functionality

- Blocking:
  - Lazy or proactive
- NN-Methods:
  - Type of operation
  - Type of threshold

NN Methods	Similarity Threshold	Cardinality Threshold
Deterministic Operation	ε-Join	kNN-Join, FAISS, SCANN
Stochastic Operation	MH-, HP-, CP-LSH	DeepBlocker

Schema Agnostic Results

- Blocking:
  - Attribute value tokens yield better results than substrings of tokens



Schema Agnostic Results

- Sparse NN:
  - Cardinality thresholds are more effective than similarity thresholds

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	overall
SBW	4	5	2	1	1	1	2	2	3	1	2.2
QBW	6	2	4	8	5	3	8	3	5	3	4.7
EQBW	2	6	5	7	3	4	7	6	8	2	5.0
SABW	1	3	3	9	4	5	3	4	4	6	4.2
ESABW	5	1	6	10	6	6	4	5	9	7	5.9
PBW	7	13	9	16	12	11	14	14	14	9	11.9
DBW	16	9	8	15	7	8	13	11	13	8	10.8
eJoin	3	8	6	4	8	13	1	7	2	5	5.7
kNNJ	9	4	1	2	2	2	6	1	1	4	3.2
DkNN	12	7	16	11	16	7	12	8	6	14	10.9
MH-LSH	17	16	15	17	15	16	17	17	17	14	16.1
CP-LSH	14	14	11	13	13	15	15	15	15	12	13.7
HP-LSH	15	15	13	14	14	14	16	16	16	13	14.6
FAISS	10	10	14	5	9	9	9	13	11	10	10.0
SCANN	10	10	12	5	9	9	9	12	10	10	9.6
DeepBlocker	8	12	9	3	11	12	5	9	7	14	9.0
DDB	13	17	16	12	17	17	11	10	12	14	13.9

Rank: 2, 4, 6, 8, 10, 12, 14, 16

Schema Agnostic Results

- Dense NN:
  - Similarity-based methods produce large numbers of candidate pairs to achieve high PC
  - Learning based tuple embedding module raises the PQ, but does not scale

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	overall
SBW	4	5	2	1	1	1	2	2	3	1	2.2
QBW	6	2	4	8	5	3	8	3	5	3	4.7
EQBW	2	6	5	7	3	4	7	6	8	2	5.0
SABW	1	3	3	9	4	5	3	4	4	6	4.2
ESABW	5	1	6	10	6	6	4	5	9	7	5.9
PBW	7	13	9	16	12	11	14	14	14	9	11.9
DBW	16	9	8	15	7	8	13	11	13	8	10.8
eJoin	3	8	6	4	8	13	1	7	2	5	5.7
kNNJ	9	4	1	2	2	2	6	1	1	4	3.2
DkNN	12	7	16	11	16	7	12	8	6	14	10.9
MH-LSH	17	16	15	17	15	16	17	17	17	14	16.1
CP-LSH	14	14	11	13	13	15	15	15	15	12	13.7
HP-LSH	15	15	13	14	14	14	16	16	16	13	14.6
FAISS	10	10	14	5	9	9	9	13	11	10	10.0
SCANN	10	10	12	5	9	9	9	12	10	10	9.6
DeepBlocker	8	12	9	3	11	12	5	9	7	14	9.0
DDB	13	17	16	12	17	17	11	10	12	14	13.9

Schema Agnostic Results

→Standard Blocking Workflow (SBW) best

→kNN-Join more robust and easier to configure and apply

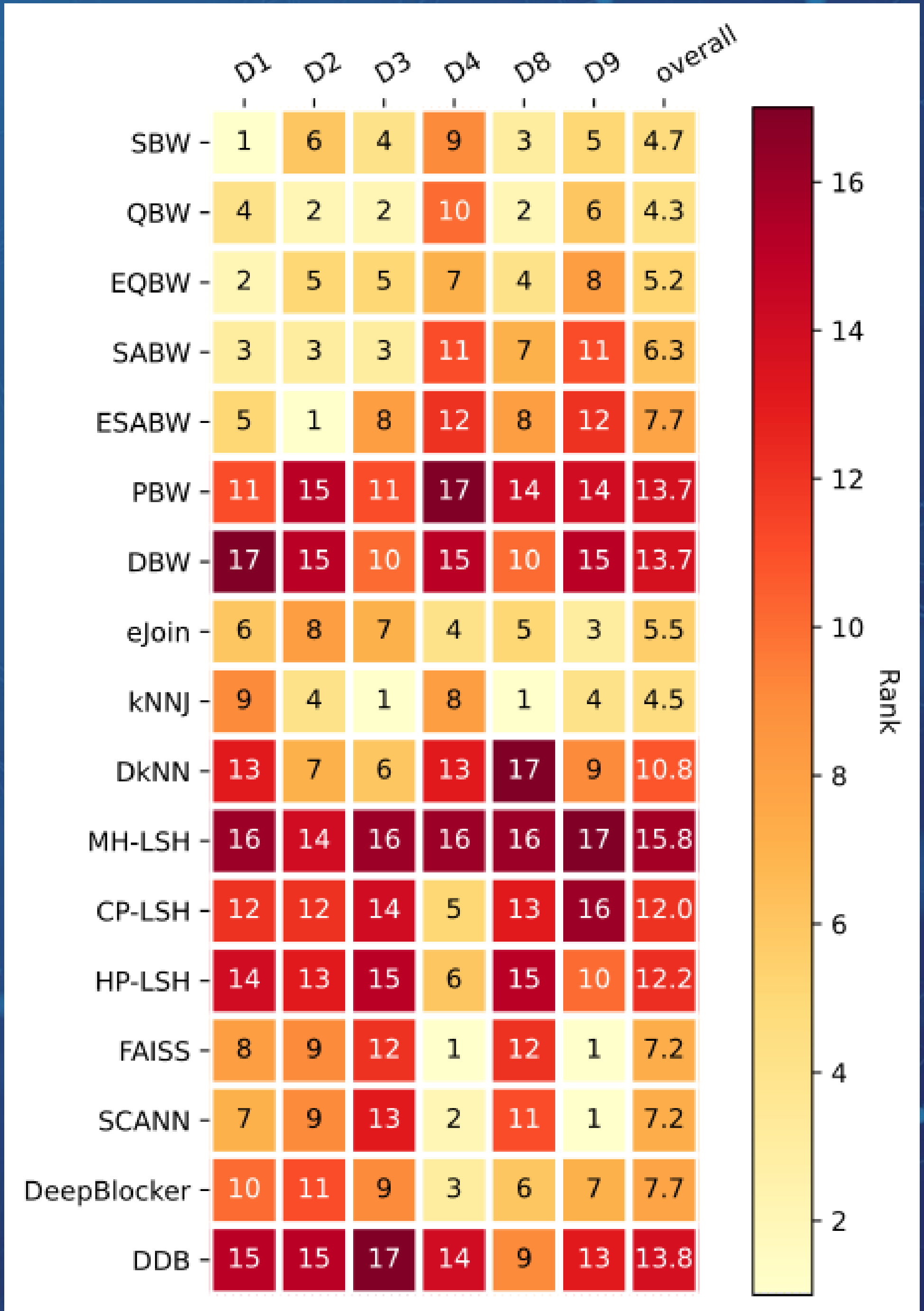
	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	overall
SBW	4	5	2	1	1	1	2	2	3	1	2.2
QBW	6	2	4	8	5	3	8	3	5	3	4.7
EQBW	2	6	5	7	3	4	7	6	8	2	5.0
SABW	1	3	3	9	4	5	3	4	4	6	4.2
ESABW	5	1	6	10	6	6	4	5	9	7	5.9
PBW	7	13	9	16	12	11	14	14	14	9	11.9
DBW	16	9	8	15	7	8	13	11	13	8	10.8
eJoin	3	8	6	4	8	13	1	7	2	5	5.7
kNNJ	9	4	1	2	2	2	6	1	1	4	3.2
DkNN	12	7	16	11	16	7	12	8	6	14	10.9
MH-LSH	17	16	15	17	15	16	17	17	17	14	16.1
CP-LSH	14	14	11	13	13	15	15	15	15	12	13.7
HP-LSH	15	15	13	14	14	14	16	16	16	13	14.6
FAISS	10	10	14	5	9	9	9	13	11	10	10.0
SCANN	10	10	12	5	9	9	9	12	10	10	9.6
DeepBlocker	8	12	9	3	11	12	5	9	7	14	9.0
DDB	13	17	16	12	17	17	11	10	12	14	13.9

Rank

Schema Based Results

- Overall similar trends as for schema-agnostic setting
- Blocking:
  - Substrings of tokens (QBW) slightly better than SBW

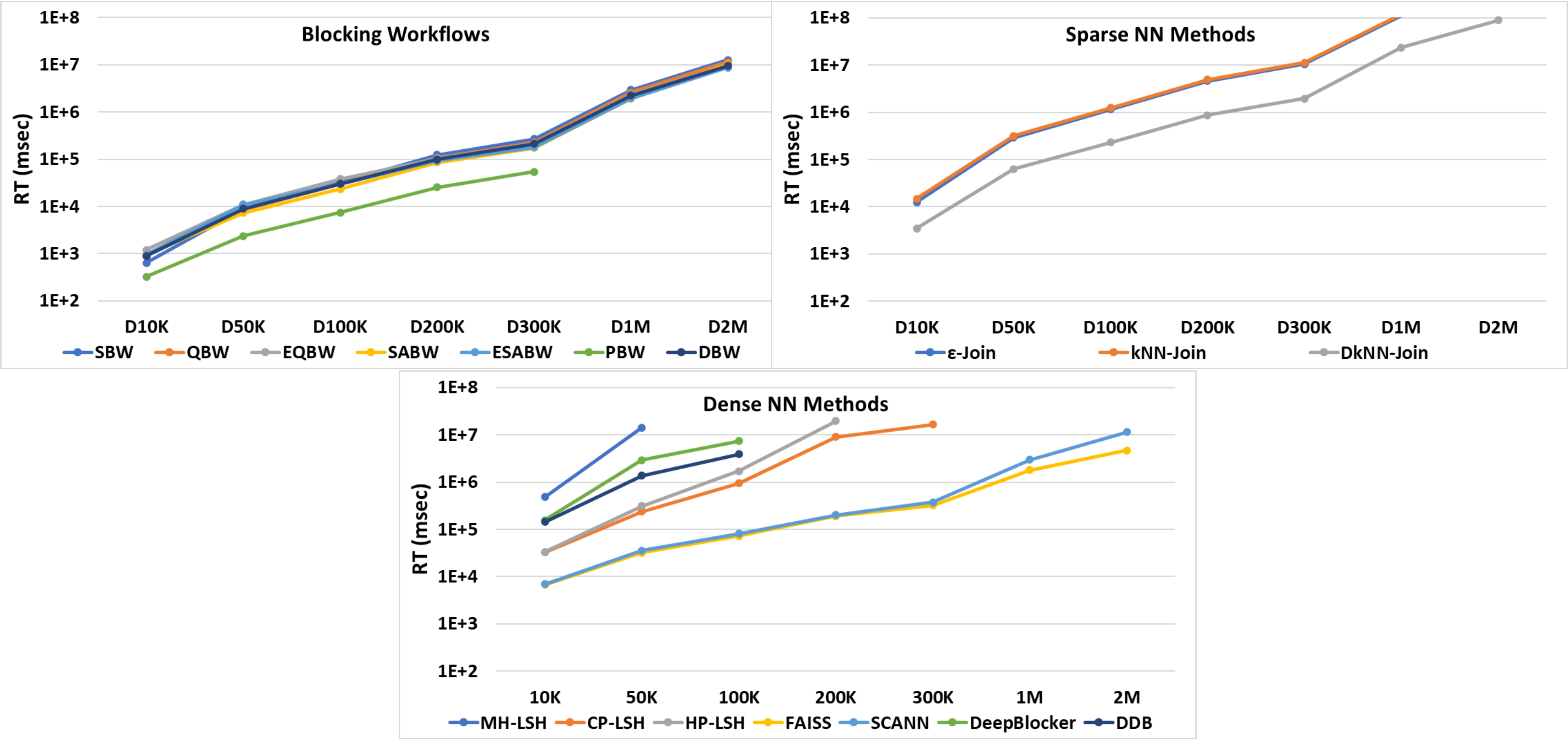
→QBW and kNNJ best and most robust



Scalability

- Run on 7 synthetic, dirty ER datasets
  - Ranging from 10,000 to 2M entries
- Techniques were finetuned on the smallest dataset
- The same settings were used for all datasets

Scalability



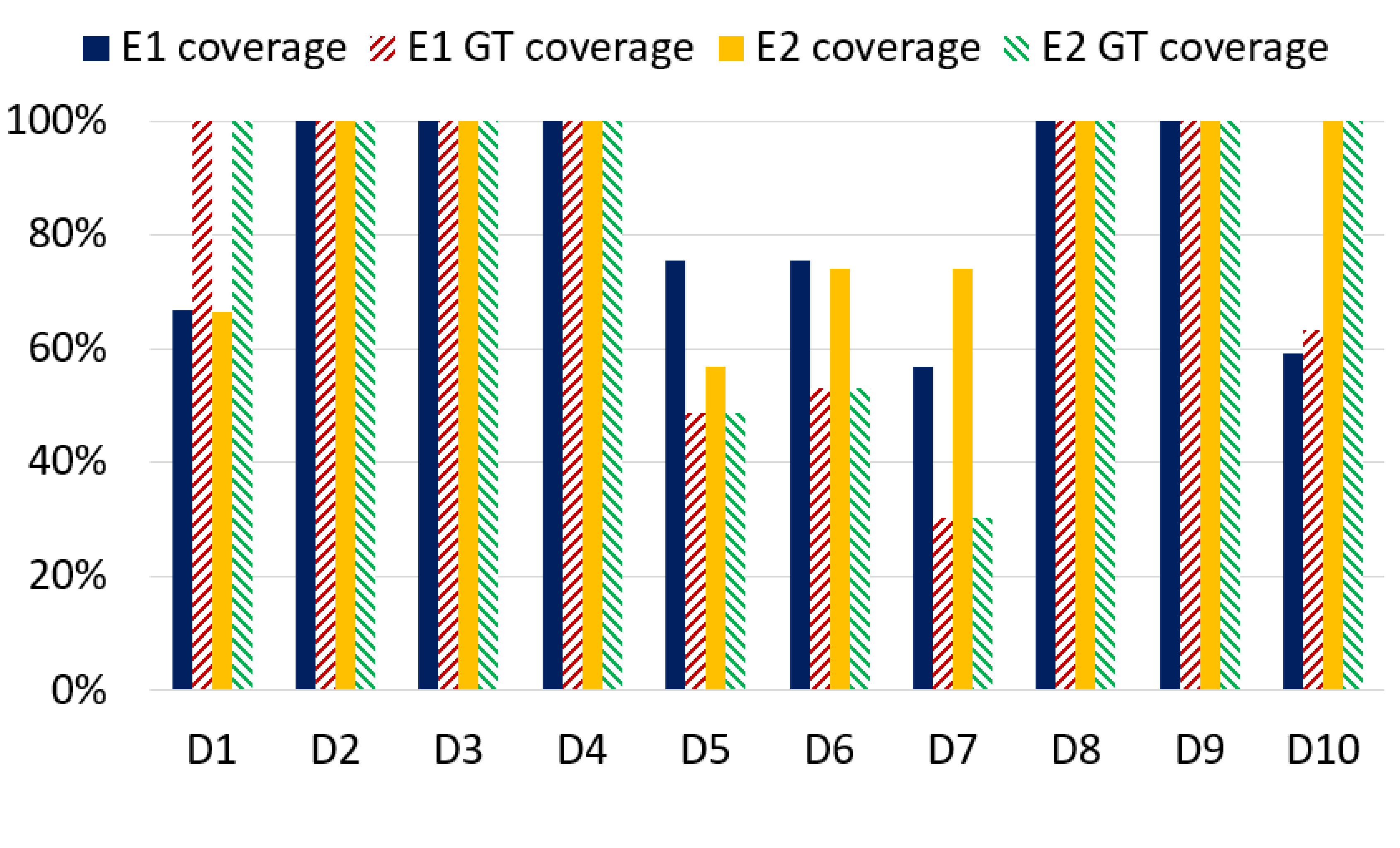
Conclusions

- We compared 13 methods, all fine-tuned to our problem and 4 baseline methods
- PQ of all methods is highly correlated
  - performance heavily depends on dataset characteristics
- Parameter finetuning significantly increases the performance
- Schema agnostic settings are preferable
- Cardinality thresholds are preferable
- Syntactic representations are preferable

Thank you

Questions?

Coverage


E1 coverage
E1 GT coverage
E2 coverage
E2 GT coverage
100%
80%
60%
40%
20%
0%
D1
D2
D3
D4
D5
D6
D7
D8
D9
D10

Scalability