BENNO KRUIT, PETER BONCZ, JACOPO URBANI

EXTRACTING N-ARY FACTS FROM WIKIPEDIA TABLE CLUSTERS

MAIN TAKEAWAY: EXTRACT FACTS FROM TABLES

- We **reshape**, **cluster**, and **integrate** tables with KB
  - to extract **n-ary** facts from Wikipedia tables for Wikidata
  - 1.5M tables → 15M binary and 6M n-ary **novel** facts

takco

github.com/karmaresearch/takco

takco.readthedocs.io

WHY EXTRACT INFORMATION FROM WIKIPEDIA TABLES?

- High-quality **background knowledge**
  - … about **known topics**
- Examples of tables for **human readers**

How do we automatically process tables that were not designed for automatic processing?

MOTIVATION

1977 Manitoba general election

From Wikipedia, the free encyclopedia

Party		Party Leader	# of candidates	Seats			Popular Vote		
				1973	Elected	% Change	#	%	Change
	Progressive Conservative	Sterling Lyon	57	21	**33**	+57.1%	237,496	48.75%	+12.02
	New Democratic	Edward Schreyer	57	31	**23**	-25.8%	188,124	38.62%	-3.69
	Liberal	Charles Huband	53	5	**1**	-80.0%	59,865	12.29%	-6.75
	Social Credit	Jacob Froese	5	-	**-**	-	1,323	0.27%	-0.10
	Communist	William Cecil Ross	4	-	-	-	299	0.06%	+0.01
	Revolutionary Workers		1	*	–	*	47	0.01%	*
	Independent		-	1	**-**	-100%	-	-	-1.49
Total			177	**57**	**57**	**-**	487,154	**100%**	

MOTIVATION

Imagine (John Lennon song)

From Wikipedia, the free encyclopedia

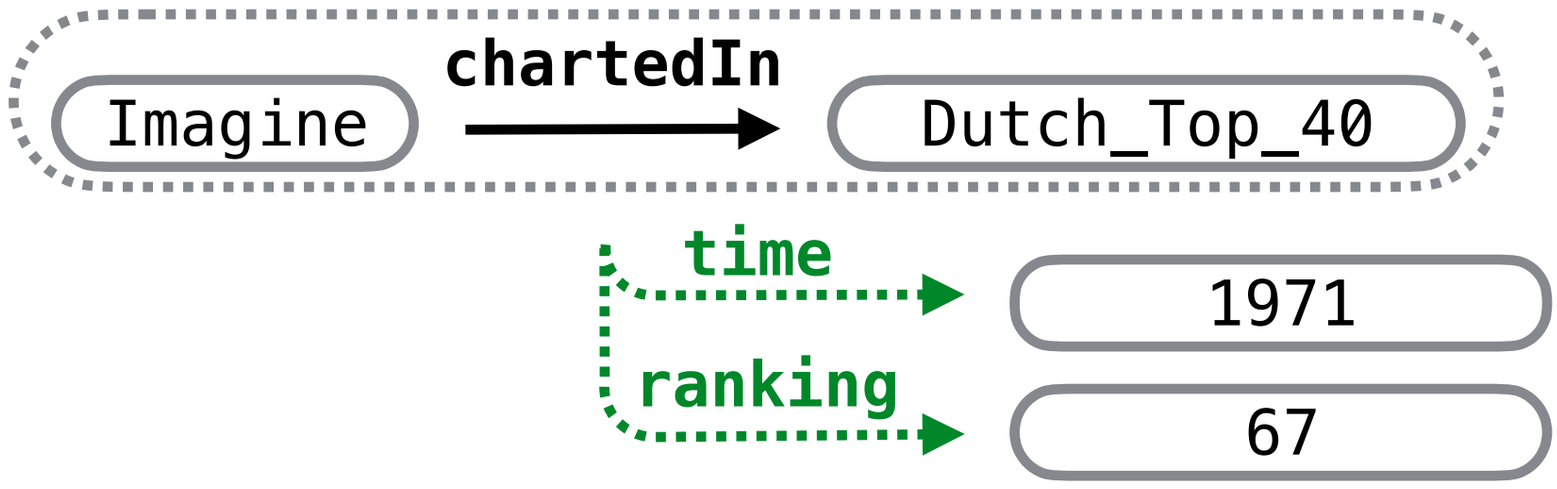
Charts and certifications [edit]

Year-end charts [edit]

Chart (1971)	Rank
Canada Top Singles (*RPM*)[128]	15
Netherlands (Dutch Top 40)[129]	67
Netherlands (Single Top 100)[130]	82

Chart (1972)	Rank
Australia (Kent Music Report)[131]	19
Japan (Oricon)[103]	98
South Africa (Springbok Radio)[132]	5

Chart (1981)	Rank
Belgium (Ultratop 50 Flanders)[133]	86
Netherlands (Dutch Top 40)[134]	70
Netherlands (Single Top 100)[135]	73



WHAT MAKES IT HARD?

- The tables are semantically **related** (same set of entities)
  - But **diverse** (different authors, topics, lay-outs)
- Most tables express n-ary relations
  - But existing work on table interpretation uses **entity-attribute assumption**!

APPROACH

- How to extract **usable** facts from **real** tables?

1. How to clean the **lay-outs** that people use in practice?
   - Unpivot heuristics
2. How to **propagate** information between similar tables?
   - Create union table by clustering
3. How to integrate them with the (n-ary) KB **data model**?
   - Strong assumptions for high precision

1. RESHAPING: UNPIVOT HEURISTICS

- Clean up sub-headers and footnotes (see paper)
- Add context as extra columns (see paper)
- Unpivot tables on value-like header cells

Commodores (album)

Year	Single	Chart positions		
		US	US R&B	US Dance
1977	"Brick House"	5	4	34
1977	"Easy"	4	1	—

Page Name	Year	Single		Chart positions
Commodores	1977	"Brick House"	US	5
Commodores	1977	"Brick House"	US R&B	4
Commodores	1977	"Brick House"	US Dance	34
Commodores	1977	"Easy"	US	4
Commodores	1977	"Easy"	US R&B	1
Commodores	1977	"Easy"	US Dance	—

1. RESHAPING: UNPIVOT HEURISTICS

Unpivot each sequence of cells that…

1. Start/end with **number**

Chart (1971)	Peak position

2. Have <Agent> **hyperlink**

Year	Australian Open	French Open	Wimbledon	US Open

3. Span values that **repeat**

Athlete	Event	Downhill		Slalom		Total	
		Time	Rank	Time	Rank	Time	Rank

4. Mostly **spanned** / in **body**

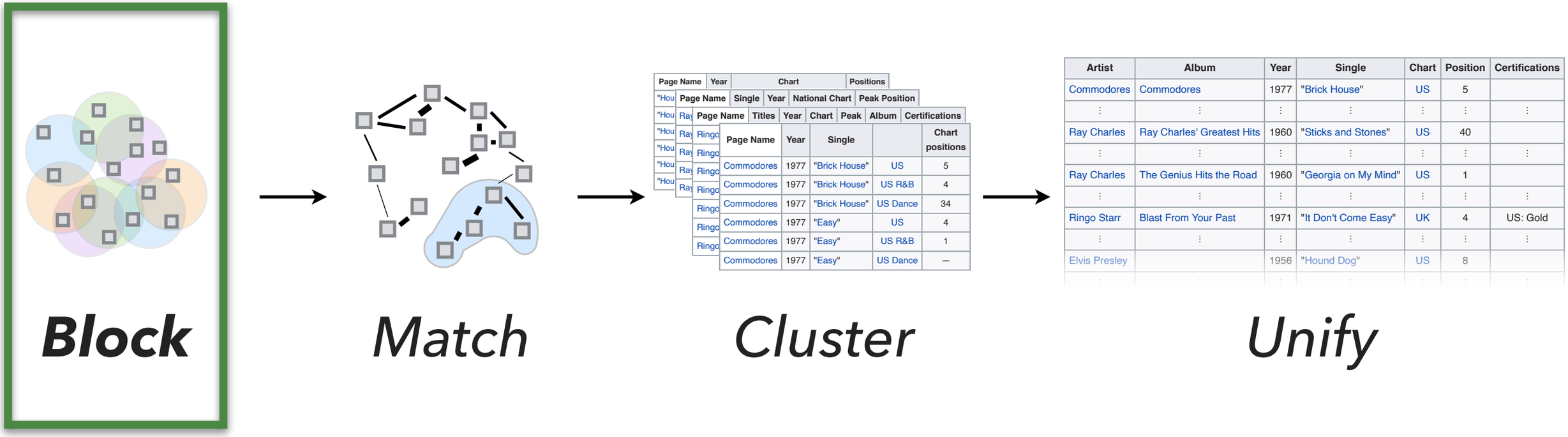
Summit	State						
	Canada	France	Germany	Italy	UK	USA	EU

5. Values are a **rare outlier**

Atlantic Division	W	L	T	OTL	GF	GA	PTS

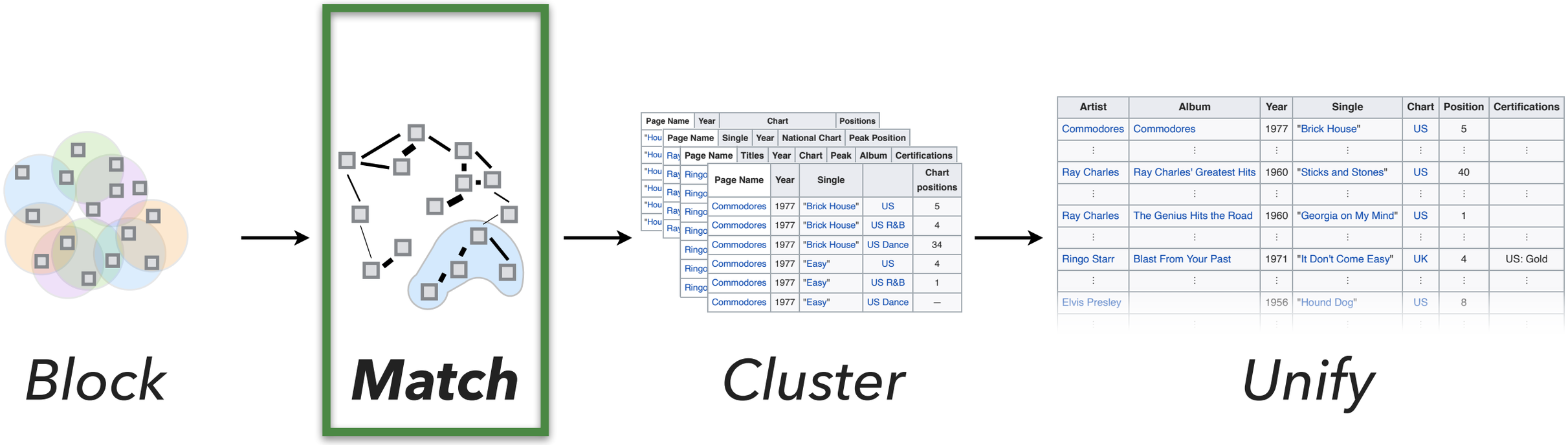
Reshaping makes it easier to cluster and unify tables

2. UNIFICATION: CLUSTERING TABLES



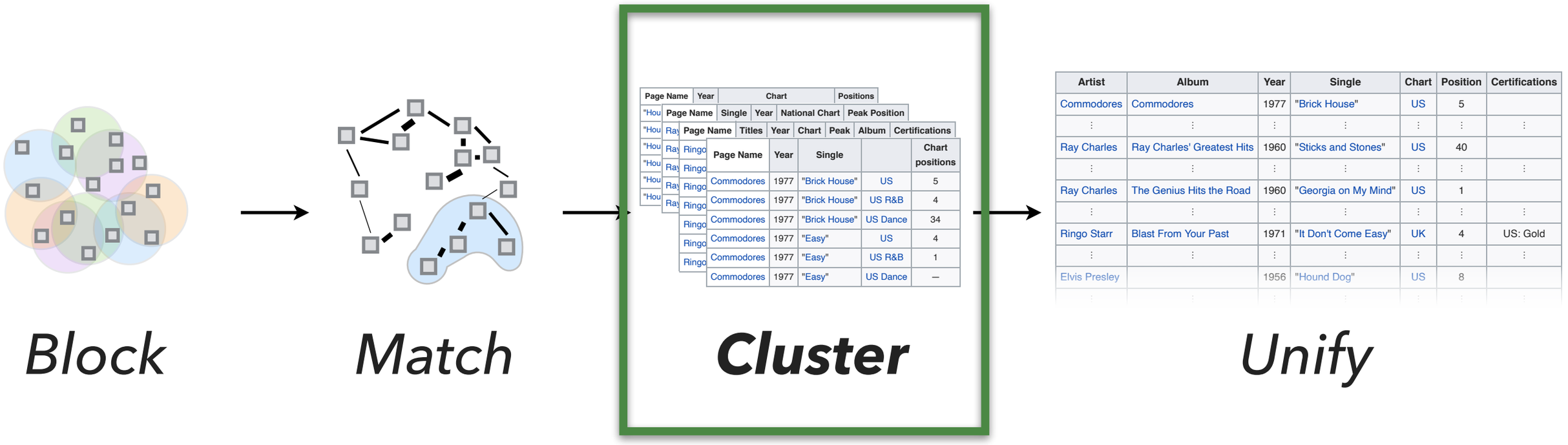
- Approximate indexes on header / body cells
  - Locality-sensitive Hashing: Jaccard index
  - Approximate Nearest Neighbors: Word Embeddings

2. UNIFICATION: CLUSTERING TABLES



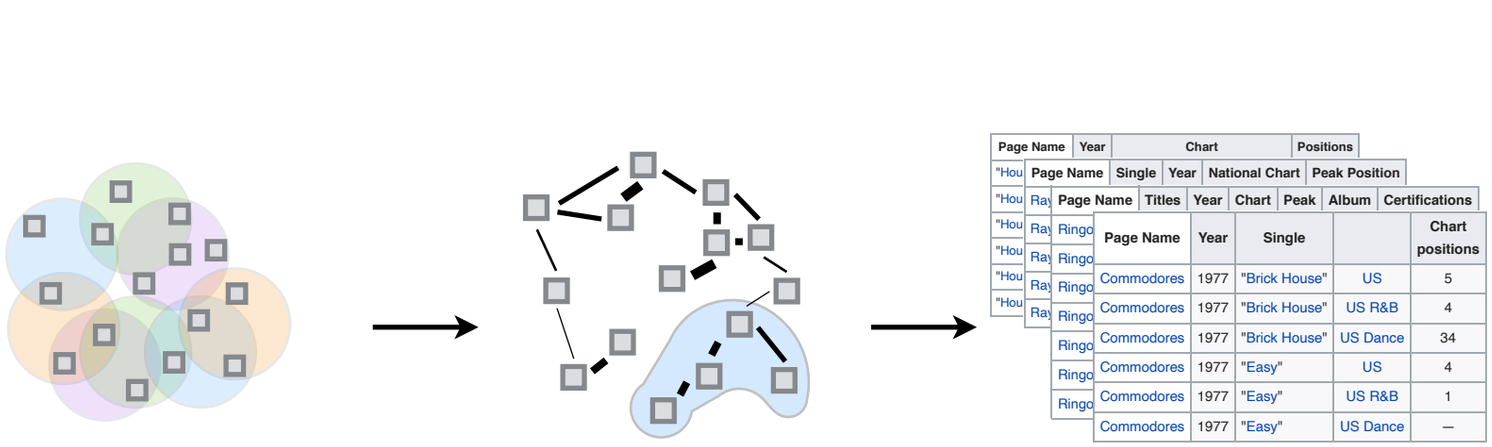
- Match tables with different metrics
  - Jaccard index
  - Word embeddings
  - Column types

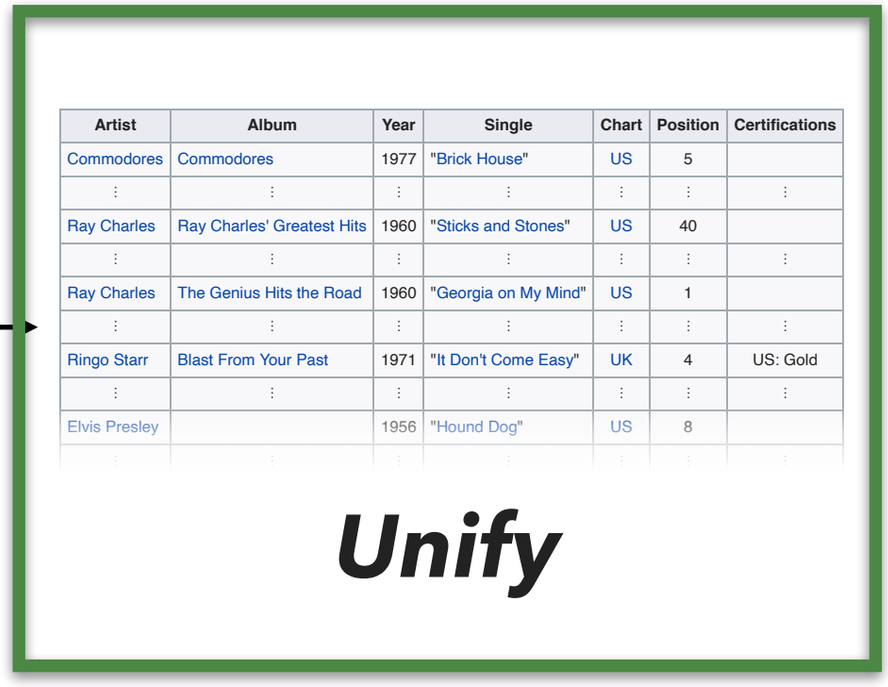
2. UNIFICATION: CLUSTERING TABLES



- Aggregate table matches → weighted graph
- Louvain modularity for graph partitioning
  - Scales to large graphs

2. UNIFICATION: CLUSTERING TABLES

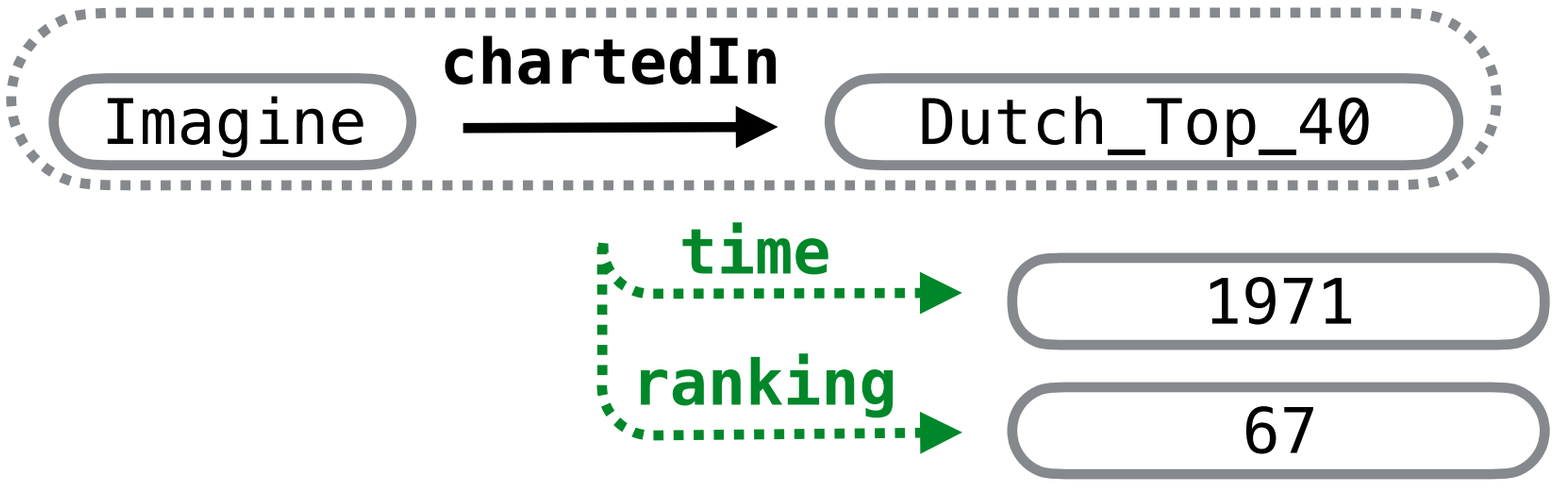




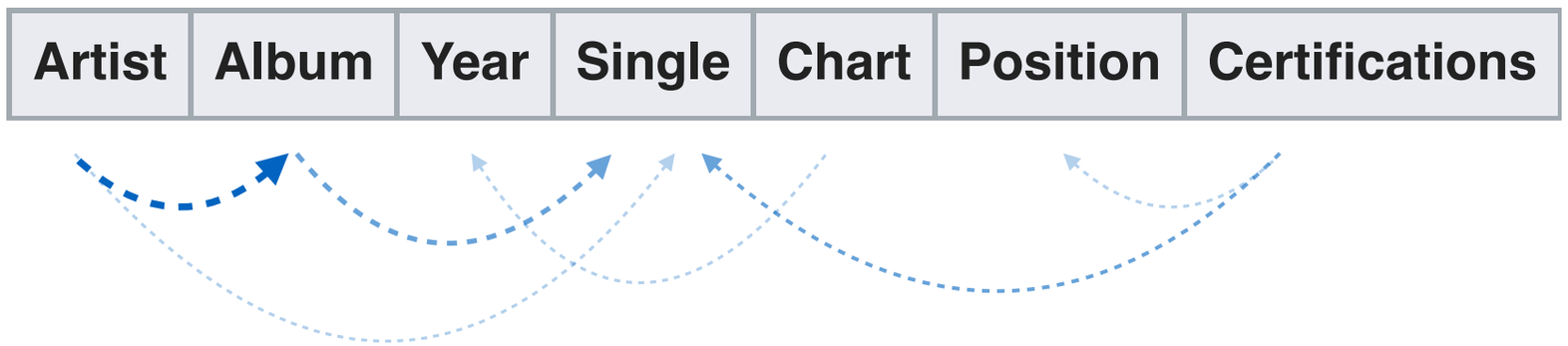
- Create union tables from partitions
  - Align columns by agglomeration
  - Propagate information within unified table

3. LINKING: MATCH COLUMNS TO KB

- Some KBs (like Wikidata) have **n-ary relations**
  - Some rows will match n-ary facts, others binary
  - N-ary facts are often expressed **incompletely**



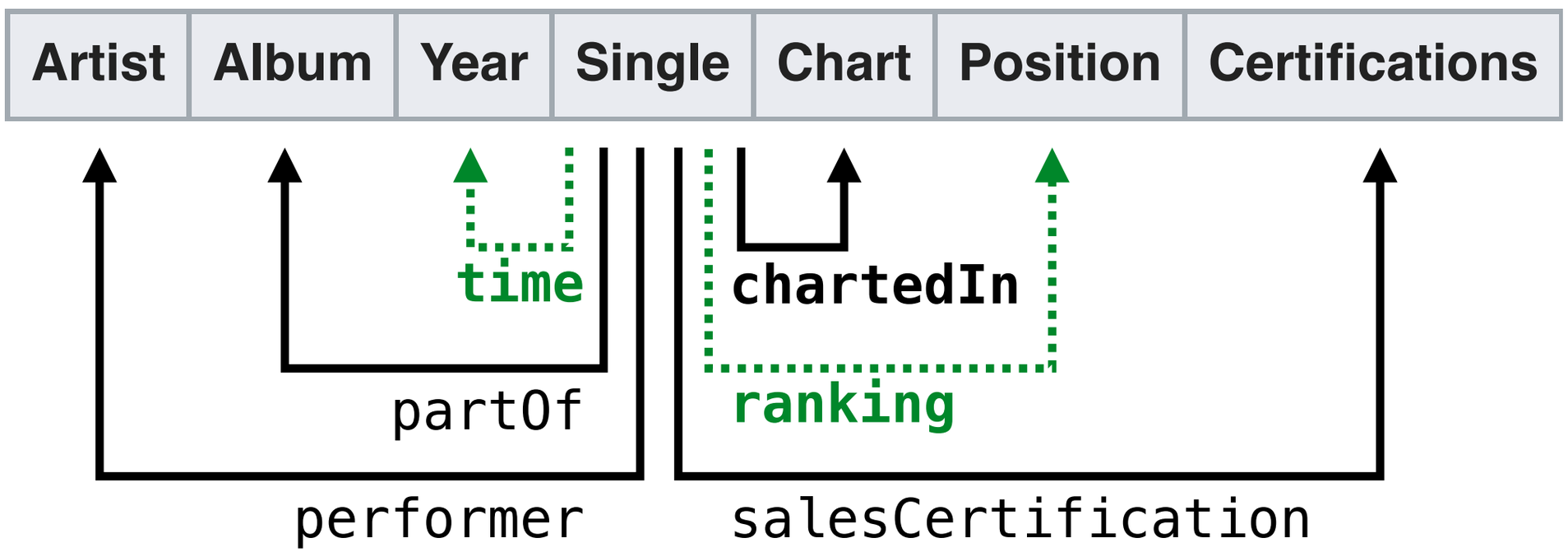
3. LINKING: KEY COLUMNS



- Functional dependency discovery
  - **Key column**: Top harmonic mean
    - High → Ent-att table; Low → N-ary table
  - Non-key dependencies: extra binary facts
- Evaluation: **outperforms baselines**

3. LINKING: STRONG ASSUMPTIONS

- N-ary relations: Main-property-value–first matching
  - To compensate for KB and table incompleteness
- High precision; improving recall is future work



RESULTS

- Created **annotated gold standard** for evaluation
  - Each step outperforms its baseline
- Scaled to **1.5M Wikipedia tables**
  - extracted 15M binary and 6M n-ary **novel** facts

CONCLUSION & FUTURE WORK

- This work: **reshape**, **cluster**, and **integrate** tables with KB
  - 1.5M tables → 15M binary and 6M n-ary **novel** facts
- Next: **Run pipeline** in practice and **contribute** to Wikidata
  - Improve pipeline using **weakly supervised learning**
  - Ontology augmentation: **new entities & relations**
- Goal: **Involve community in open-source effort**

takco

github.com/karmaresearch/takco

takco.readthedocs.io

BENNO KRUIT kruit@cwi.nl