The Graph Is Not Given

A Declarative Semantic Layer for Relational Foundation Models

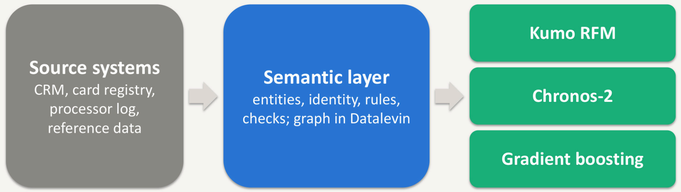
Huahai Yang · Accenture

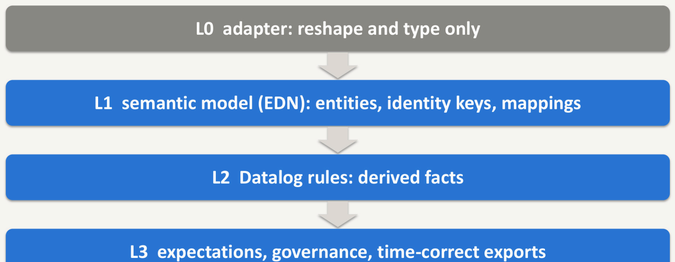
Stanford Data-Centric AI Workshop · November 3, 2026

1 The graph is not given

- Relational foundation models (RFMs) and GNNs assume clean entities and foreign keys. Enterprise data has neither.
- IBM TabFormer, split into the systems a bank would have:
  - CRM export with no ID column
  - Card registry keyed by (user, card index)
  - Processor log: country in the state column, ZIP as a float, city 'ONLINE'
  - Reference data: ISO MCC codes; Census ZIPs (14% missing)
- The entity that matters most, a location shared by customers and merchants, exists in no source.

2 A declarative semantic layer



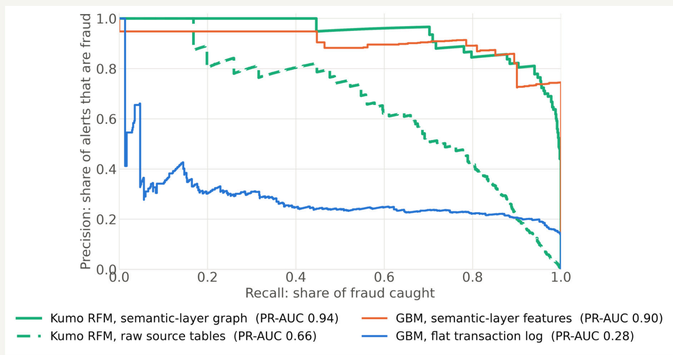



```
{:entity :transaction  :key :row-index
 :attrs {:transaction/card     {:ref :card
           :key [:card-id "User" "Card"]}
         :transaction/location {:ref :location
           :key [:txn-location-id
                 "Merchant City" "Merchant State" "Zip"]}}}

;; feature: distinct earlier customers at this location
[:window {:group [:transaction/location]
          :agg :distinct-count
          :of [:path :transaction/card :card/owner]}]
```

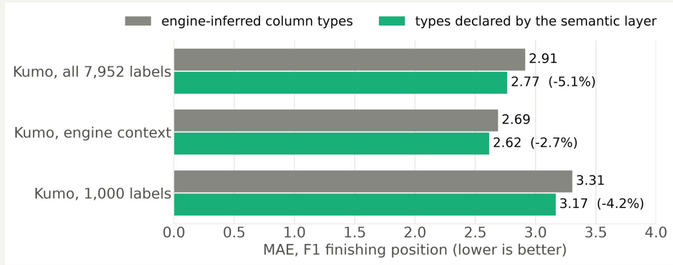
Datalevin is an open-source (EPL-2.0) Datalog database: an embedded triple store on LMDB with Datomic-style queries and rules. L2 rules run inside it; features and exports are queries over the same graph. 218M facts loaded in 11 min.

3 Same RFM, different graph



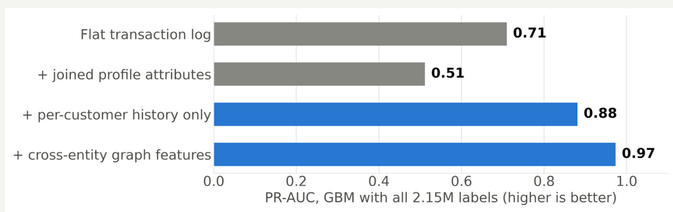
Kumo Relational with the same 1,000 in-context labels on the same 44k holdout rows: PR-AUC 0.94 on the semantic graph vs 0.66 on raw tables. At 90% recall, precision falls from 0.82 to 0.22, the level of a flat-log GBM.

4 Semantics live in metadata too



RelBench rel-f1: the engine inferred finishing position, points and wins as categorical. Declaring them numeric improved every Kumo variant.

5 Classic ML needs the graph as much



Gradient boosting on the same graph: cross-entity signals carry the lift; naive profile joins hurt.

6 Inside one subgraph

Same customer, same morning, scored by Kumo on the semantic-layer graph (time-correct, labels never in the subgraph):

	Fraud: $602.10, Rome	Legitimate: $1.91, near home
Kumo P(fraud)	0.99998	0.00001
Earlier transactions at this location	35, from 27 different customers	16, all the cardholder's
Subgraph size	55 transactions, 27 customers	49 transactions, 15 customers

The resolved location entity lets the RFM see a crowd of unrelated customers at one foreign city. No single source table contains it.

7 Declarative at scale: M5 retail

- SNAP declared as a state property, prices as weekly, zero sales only while on sale
- 122M facts in 257 s; 5 data-quality expectations pass
- Model inputs identical to a hand-written pandas pipeline

Takeaways

- **Graph construction is a first-order, data-centric determinant of RFM performance**
- Make it declarative, checkable and shared by every consumer
- Benchmarks are public and partly synthetic; enterprise pilots next

Limitations

- Kumo results use the hosted API (Oct 2026); NVIDIA's open-weight release not yet re-run
- Fraud benchmark is synthetic; enterprise pilots are the next step
- Semantic models are hand-written per domain today

Protocol

- Fixed temporal splits; decision rules written before running
- Every model scored on the same rows; label budgets stated
- Results checked across 3 model seeds

References: KumoRFM-2 (arXiv 2604.12596) · RelBench (Robinson et al., 2024) · IBM TabFormer · M5 forecasting · Chronos-2 (Amazon) · Datalevin (github.com/datalevin/datalevin)