

The background of the slide features two cakes. On the left is a round cake on a silver stand, topped with a variety of fresh fruits including raspberries, strawberries, blueberries, and blackberries, along with green leaves. On the right is a rectangular cake, also on a silver stand, with a white frosting and a similar fruit topping. The entire scene is softly blurred.

# Vector-Relational Databases

And how to avoid historic recurrence

(too many times)

Viktor Sanca

**EPFL**

[viktor.sanca@gmail.com](mailto:viktor.sanca@gmail.com) /@epfl.ch

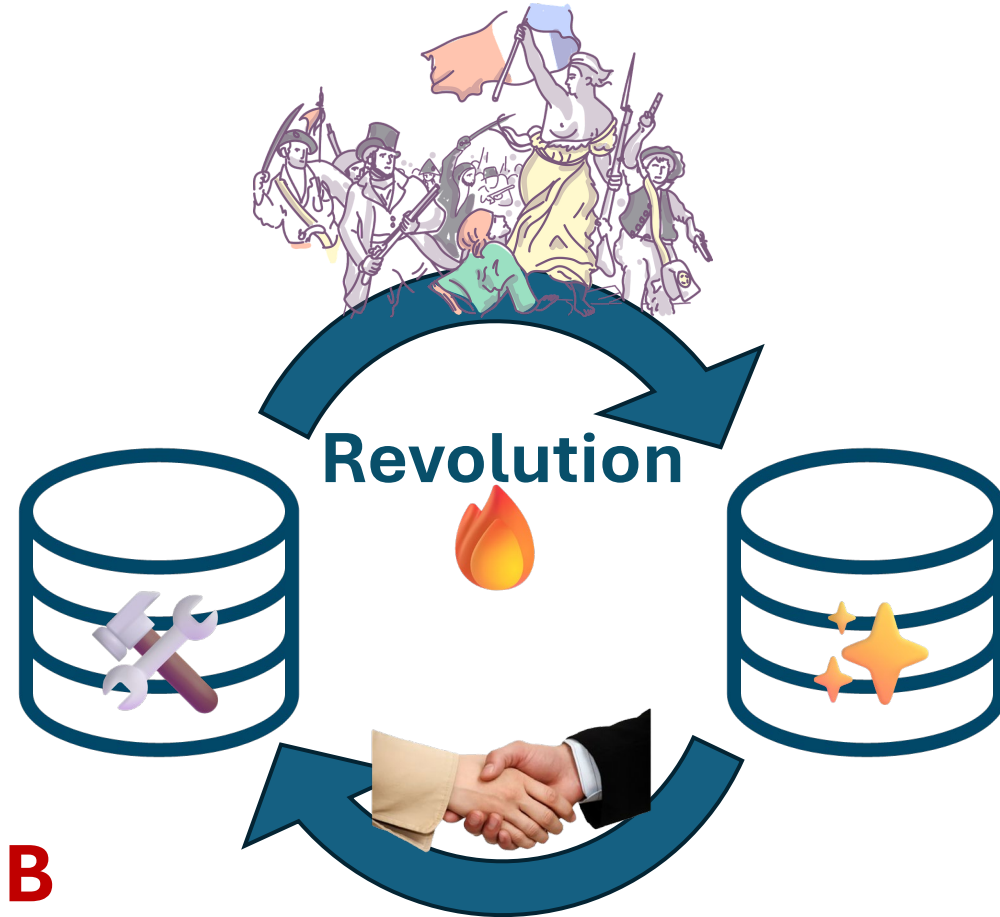
# Recurrence: A New Database for (Hot) Applications

## Causes

...  
Unstructured Data  
Documents  
Horizontal Scaling  
**Machine Learning**

...

**Battle-Tested DB**



## Reaction

...  
NoSQL  
XML Databases  
MapReduce  
**Vector Databases**

...

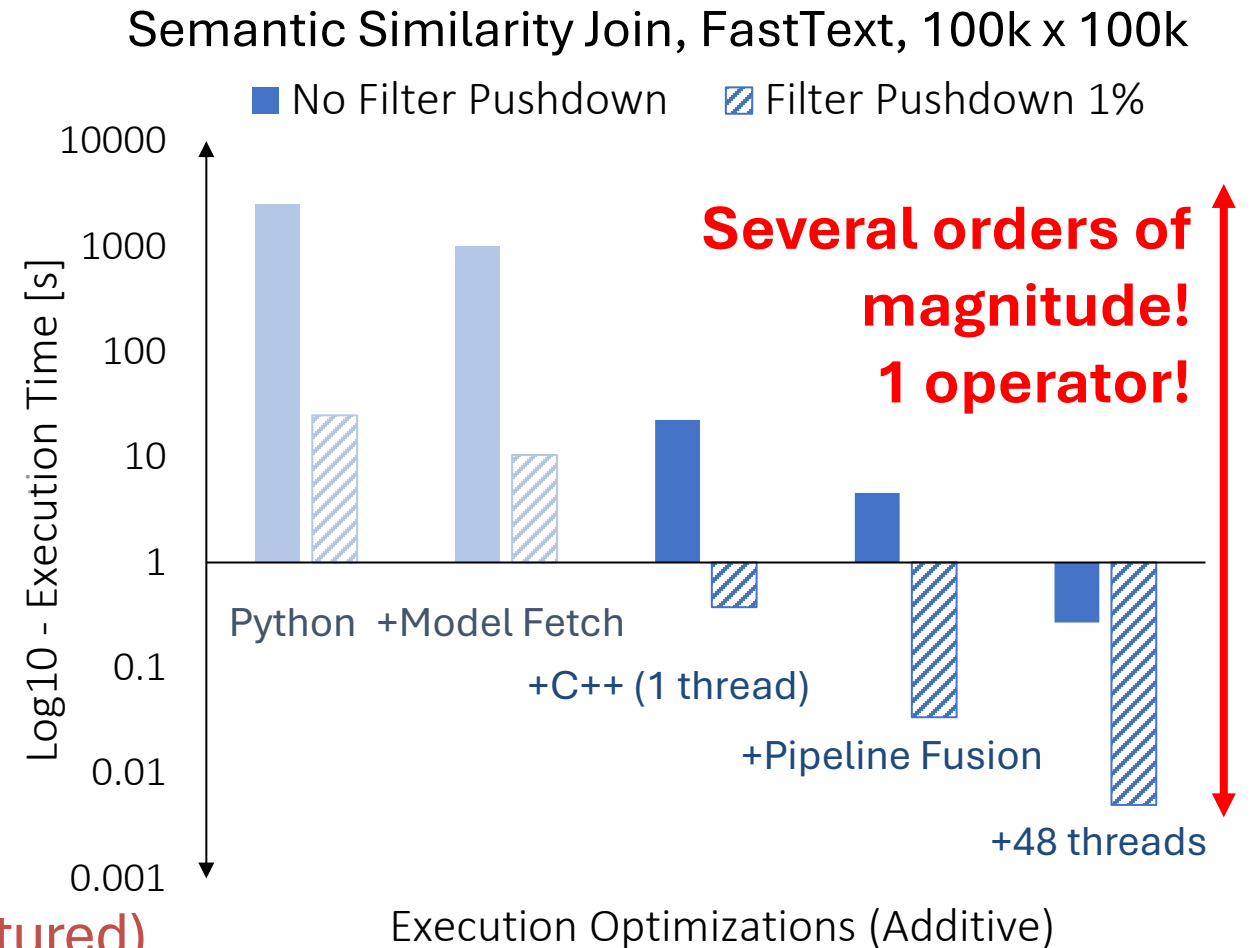
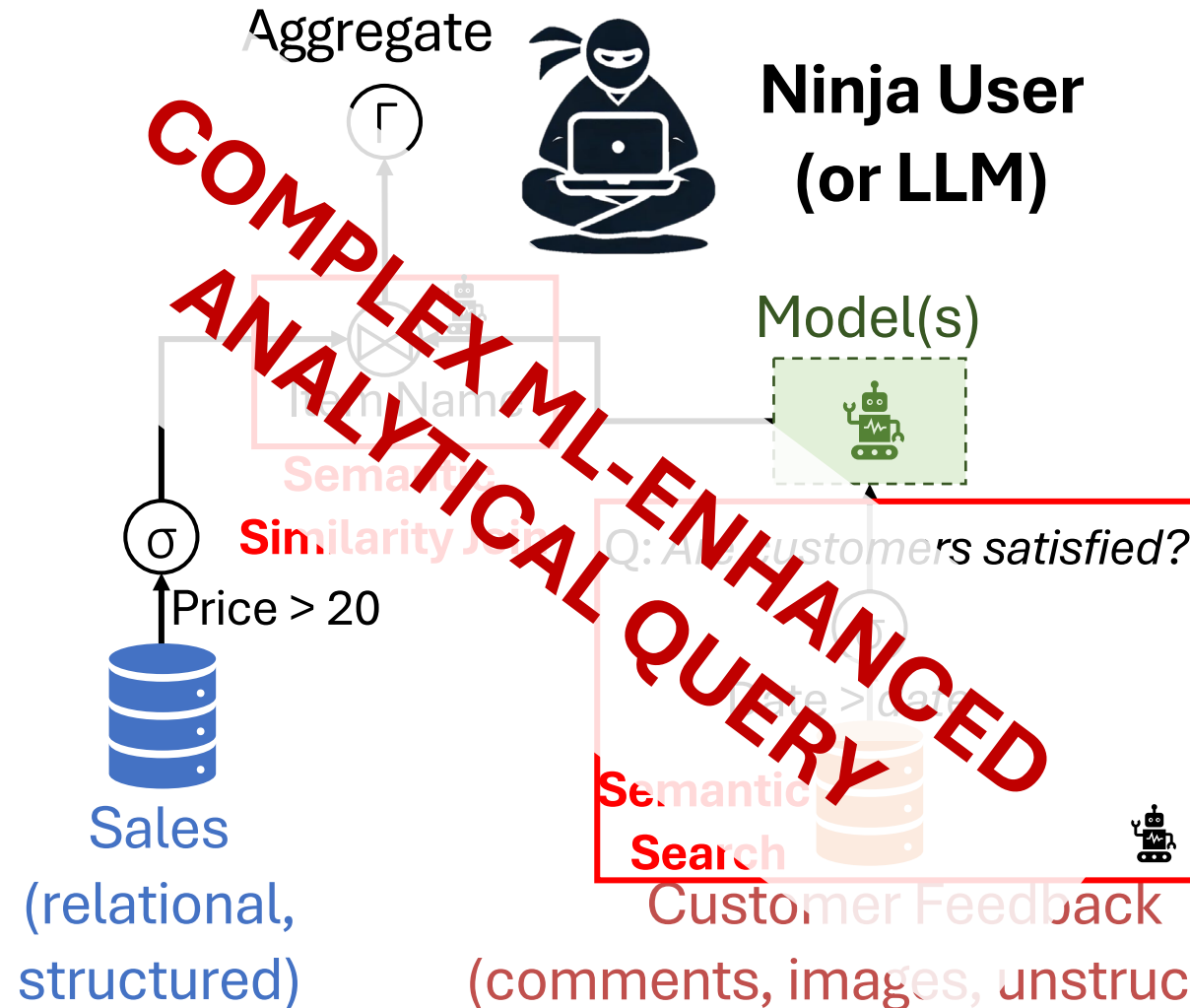
**A Shiny New DB**

**Reconciliation += New Feature**

**Emerging pattern: good principles are bound to stay**

# What Not To Mix: $\uparrow$ Complexity + Imperative Queries

[Analytical Engines With Context-Rich Processing, ICDE'23, TKDE'24]

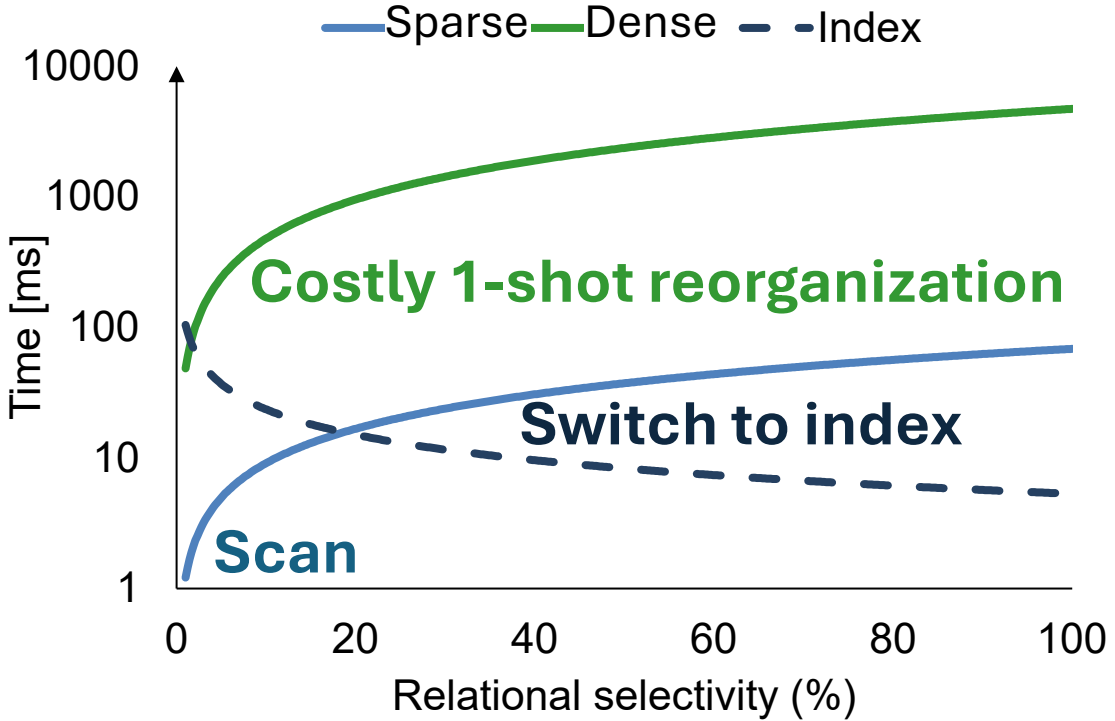


**Composable & optimizable operators for ML-enhanced analytics**

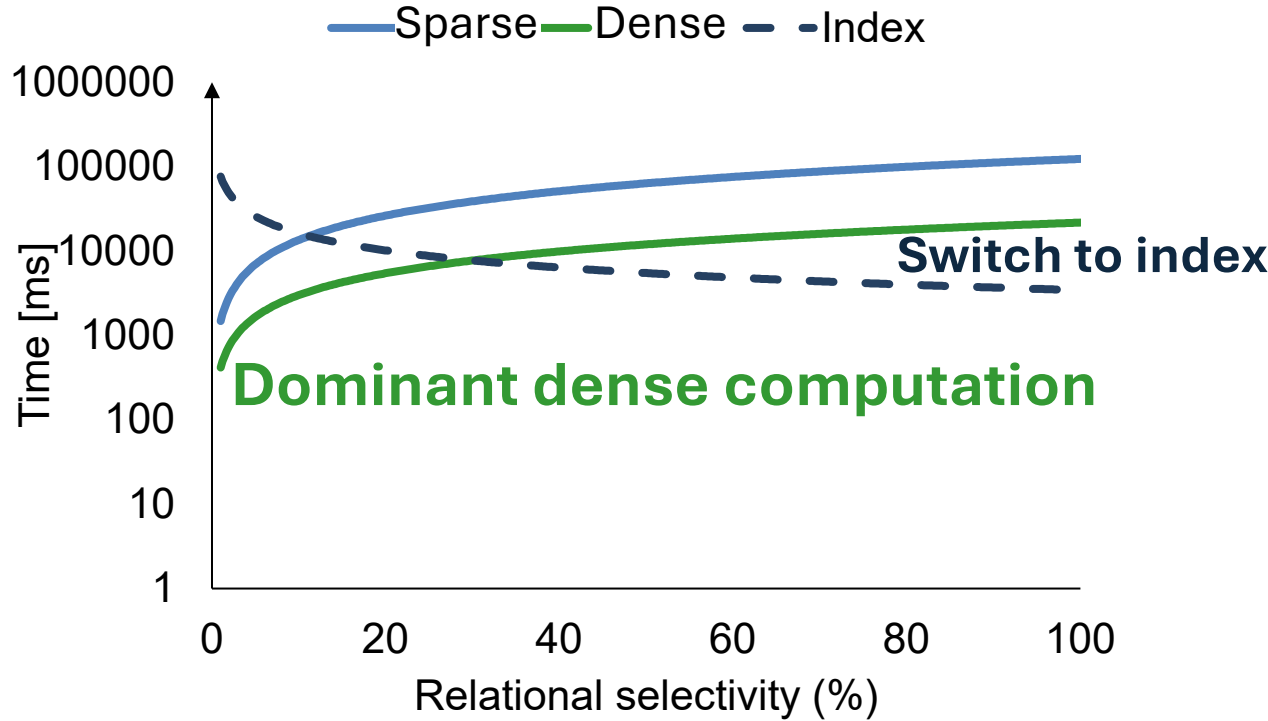
# Mixed Data Access Paths: A Not-so-Simple Decision

[Efficient Data Access Paths for Mixed Vector-Relational Search, DAMON'24]

### Query batch size 1 (point lookup)



### Query batch size 10k (join, batching)



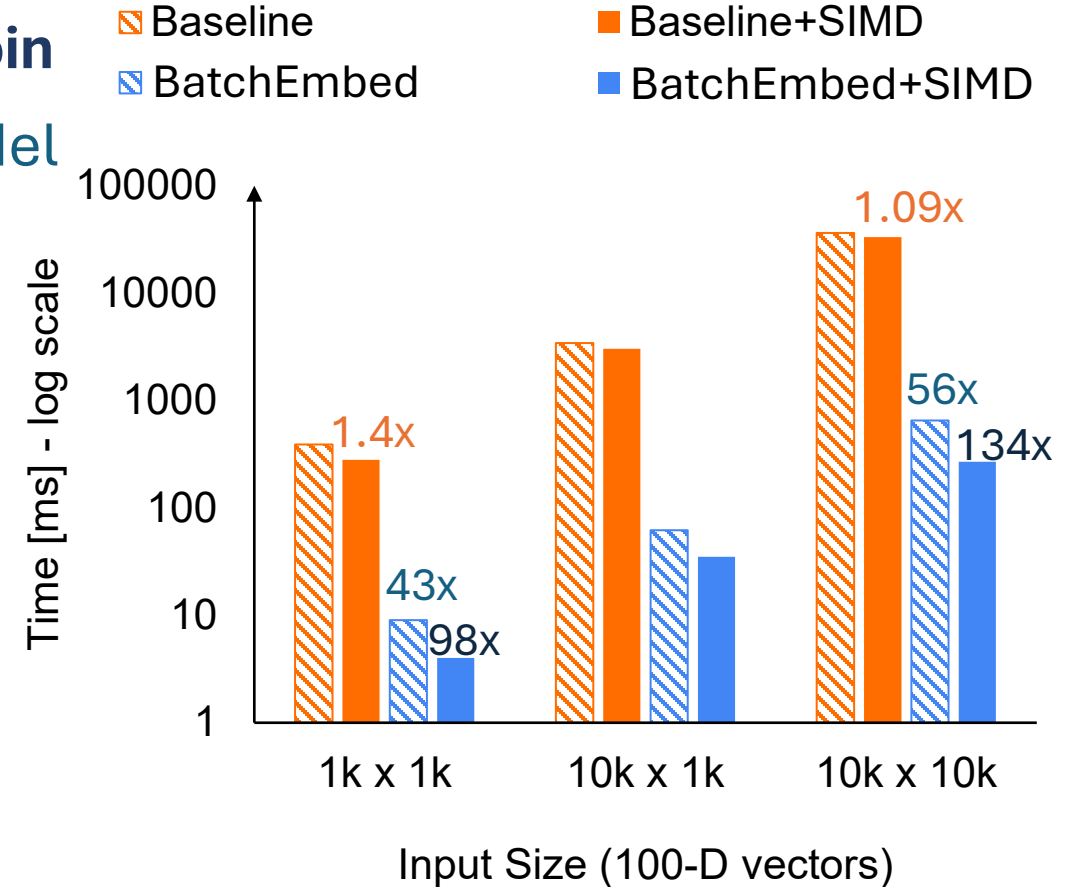
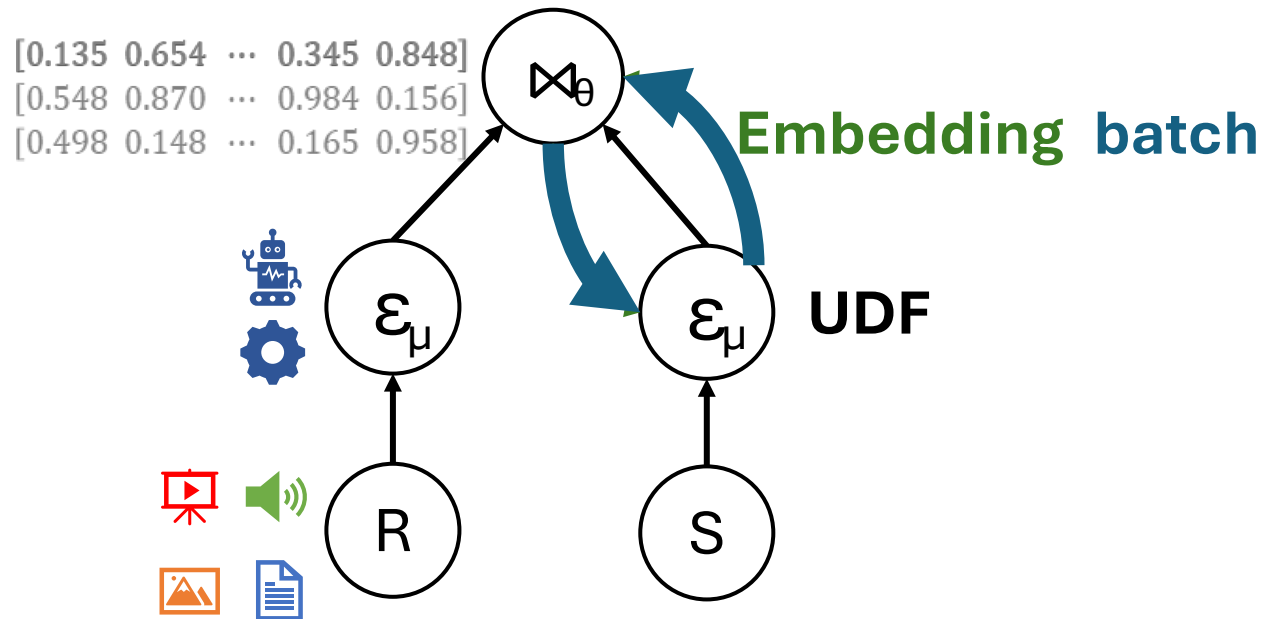
**Vector-relational processing: new optimizations & design space**

# Online Embedding Operator for Context Enrichment

[Optimizing Context-Enhanced Relational Joins, **ICDE'24**]

## Pairwise comparisons: Nested Loop Similarity Join

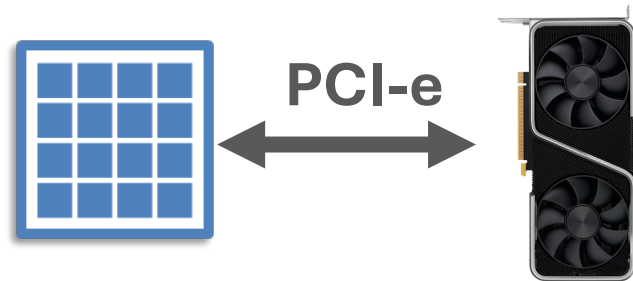
Volcano vs vectorized/compiled execution model



**Tight integration with execution model needed - additive bottlenecks!**

# Vector Join+Tensor Formulation? A Great GPU Fit!

[Optimizing Context-Enhanced Relational Joins, ICDE'24]



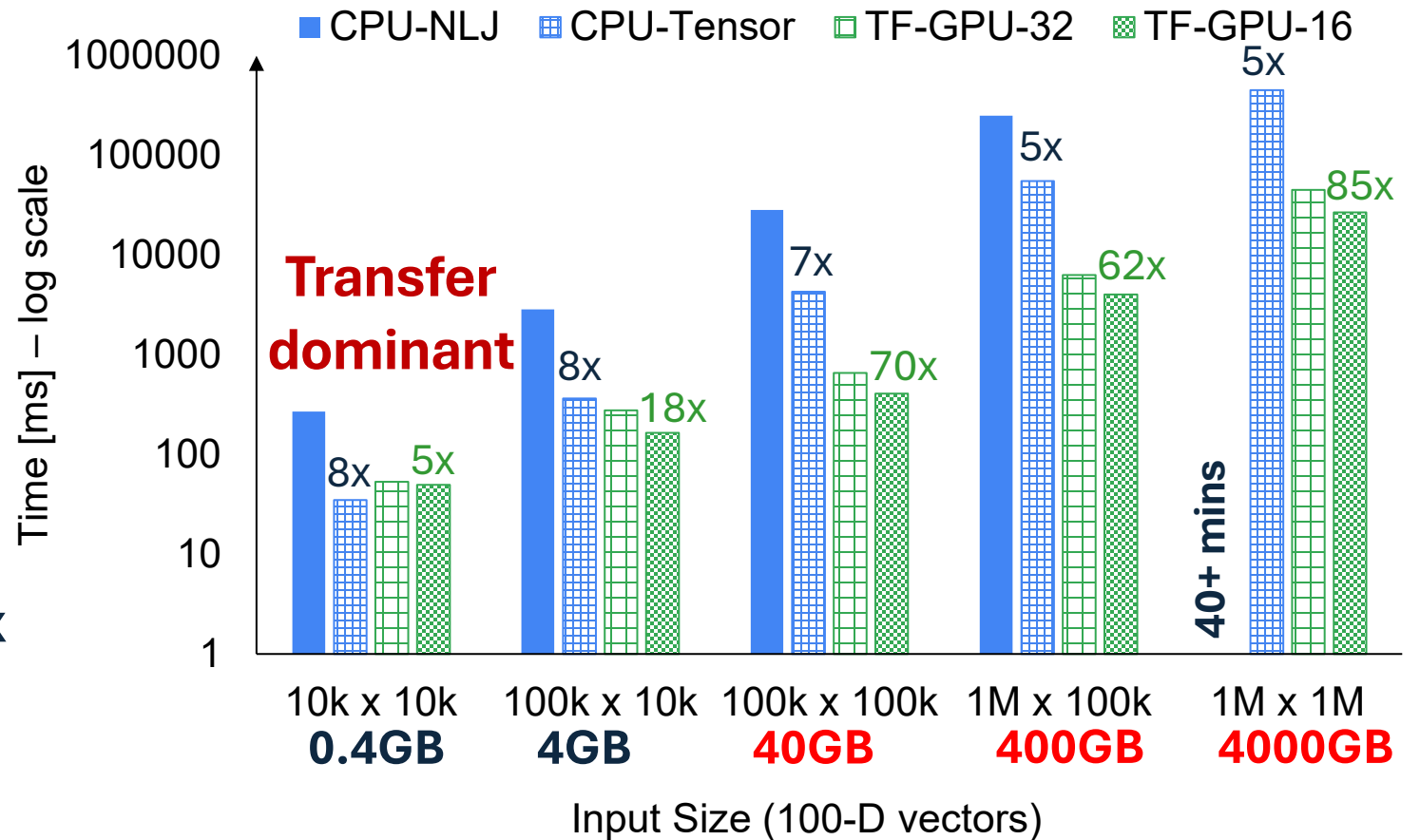
**DRAM Capacity Requirement** 💰

**NVIDIA V100 (32GB HBM)**

**Batch: <32 GB**  
**100k x 10k**  
**(4GB FP32)**

**OK! Block Matrix Decomposition**

**Compute dominant**



**A fusion of ML & database principles: logical & physical optimizations**

# Avoiding Recurrence: The Next (R)evolution of Databases?

**Unified mixed vector-relational analytics:** vertically optimizable

+

**GPUs/xPUs strike back:** new movement/computation cost

+

**Declarativity** must tame the rising query and deployment cost & complexity



**Efficient Framework for Structured+Unstructured Data & Queries**



**Let's make this possible!**