noise page

Database Gyms

Wan Shen Lim, Matthew Butrovich, William Zhang
Andrew Crotty[♦], Lin Ma[♠], Peijing Xu, Johannes Gehrke[⋆], Andrew Pavlo

Carnegie Mellon University
♦ Northwestern University
♠ University of Michigan
⋆ Microsoft Research

Overview

Database Gyms: an integrated environment providing a unified API of pluggable components to obtain high-quality training data for autonomous DBMSs

PostgreSQL

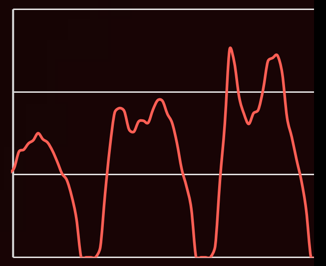
HOLD UP

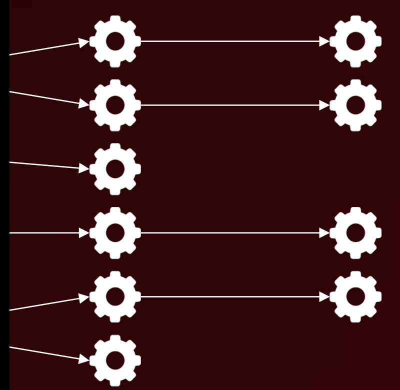
What happened to Peloton and NoisePage?

INTRODUCTION ›

Self-Driving Databases @ CMU

2014 -





Workloa … nning

[SIG … 18] … [SIGMOD 21] … (Monte Carlo Tree Search)

Self-Driving Capabilities

SELF-DRIVING DATABASES ›

Architecture

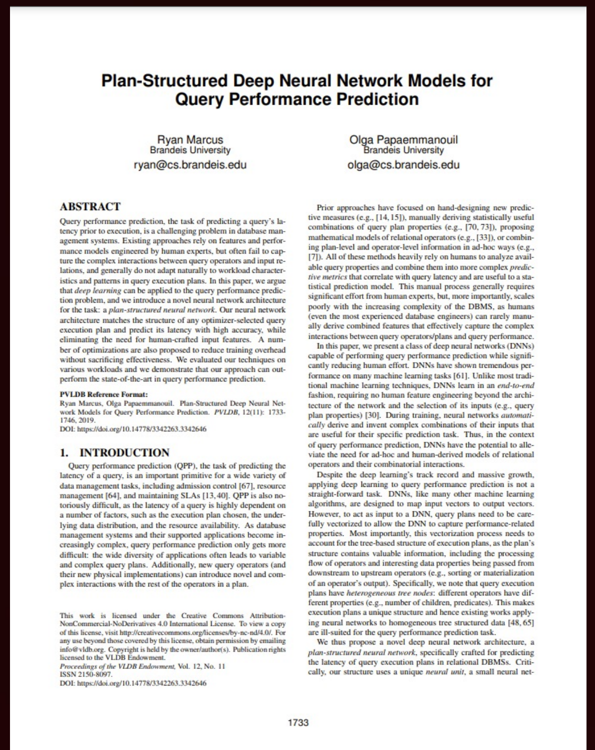
What components are needed for self-driving?

- Workload forecasting
- Behavior modeling
- Action planning

How have recent papers focused on these problems?

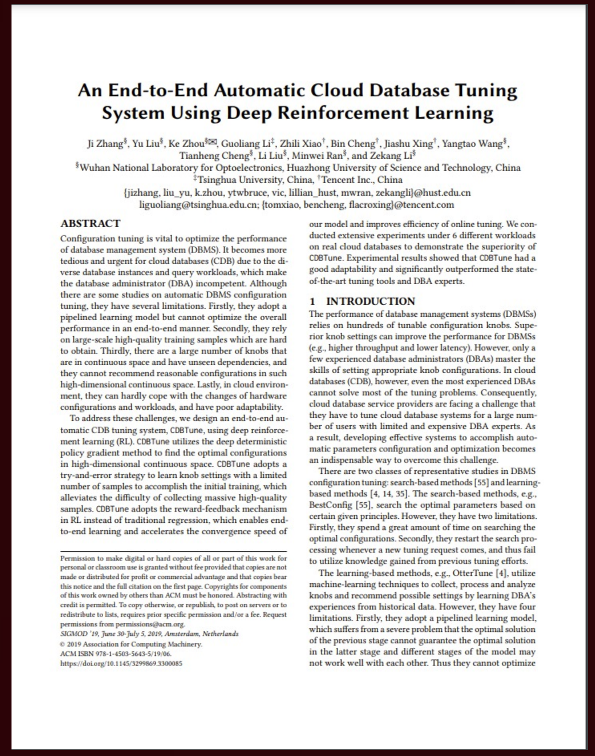

MB2: Decomposed Behavior Modeling for Self-Driving Database Management Systems

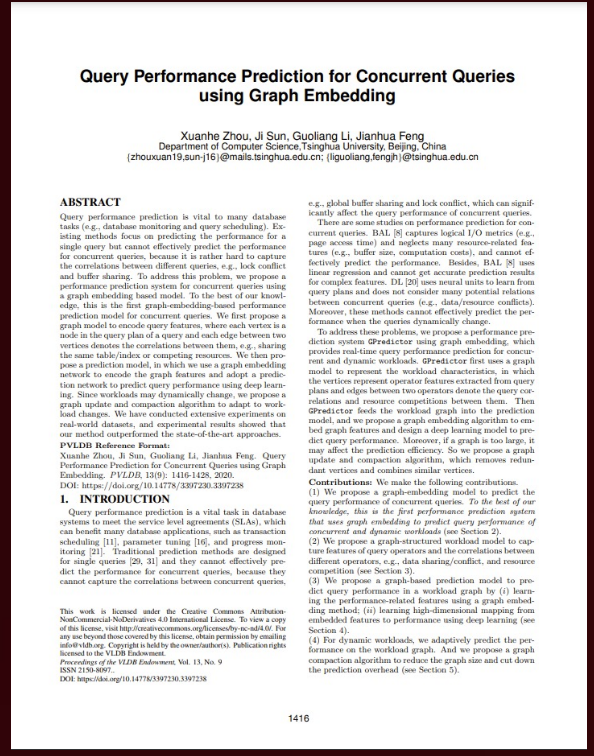

Query-based Workload Forecasting for Self-Driving Database Management Systems


Plan-Structured Deep Neural Network Models for Query Performance Prediction


DBMind: A Self-Driving Platform in openGauss


A Unified Transferable Model for ML-Enhanced DBMS


An End-to-End Automatic Cloud Database Tuning System Using Deep Reinforcement Learning


Query Performance Prediction for Concurrent Queries using Graph Embedding

Modern ML

ML = $\boxed{\text{Models}}$ + $\boxed{\text{Training Data}}$

If you do know what model you want,

- 1 month, 1k LOC

If you do not know what model you want,

- Foundation models
- Automated model design (e.g., AutoML)

Modern ML's Implications

Training data is what matters today

With systems knowledge, generate training data that is

- Better (higher quality)
- Faster (less time taken)

Leave the ML to the ML people

TRAINING DATA

How do we collect training data?

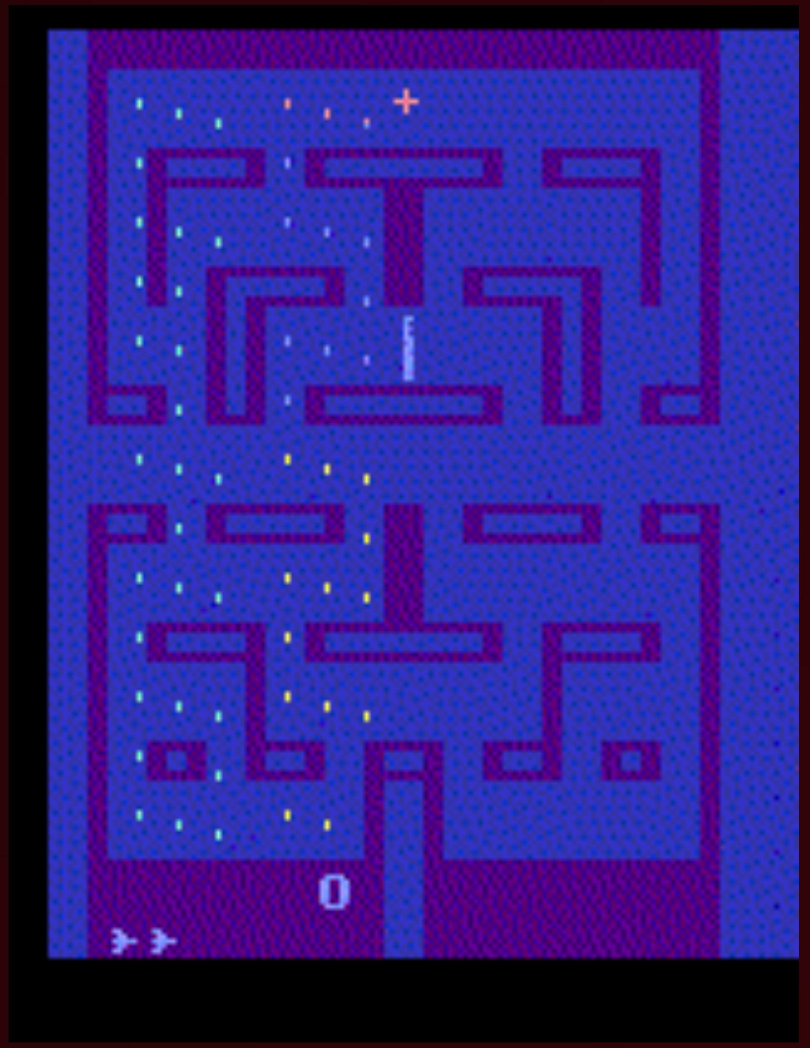
Obtaining Training Data

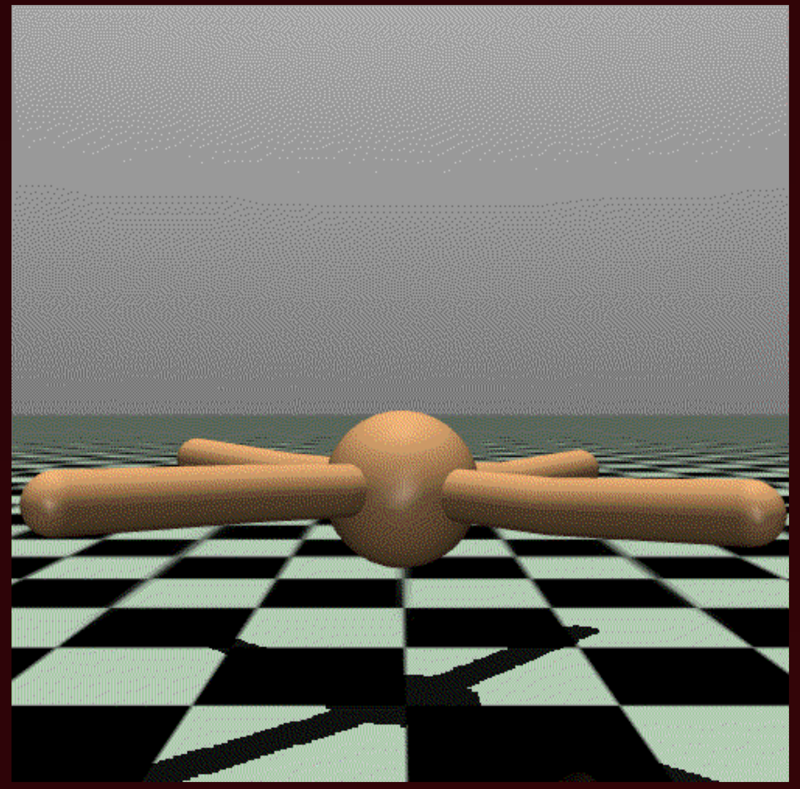
Choice #1: Do it live!

- Performance degradations in production

Choice #2: Simulator (via a gym)

- Approximates the behavior of an entity when it would otherwise be too costly, time-consuming, or dangerous to experiment on the real system
- Packaged into toolkits for developing and evaluating different models and algorithms





Obtaining Training Data via Simulation

Building a DBMS simulator is difficult

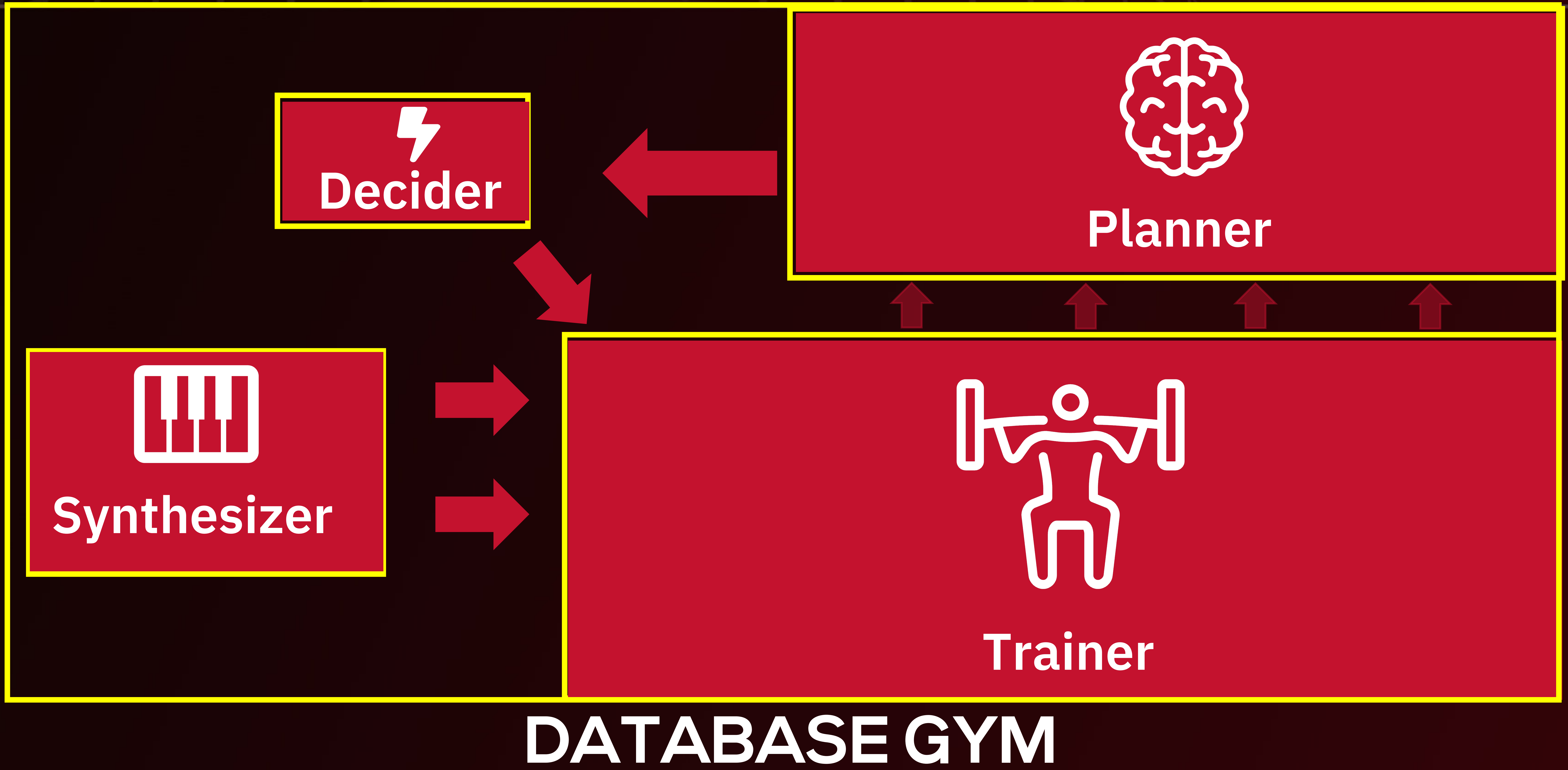
Key idea: Use the DBMS to simulate itself

- Requires solving systems and ML problems hand-in-hand
- We call this integrated solution the Database Gym

DATABASE GYM

Architecture

Using the DBMS as a Simulator



Synthesizer

Manages and manipulates the inputs to the database gym

- Snapshot : backups (e.g., pg_dump)
- Workload : timeseries of SQL queries

Goal: what-if scenarios without replaying the workload

- Example: "create a snapshot with 2x the data and increase the queries in the workload by 5x"

Trainer

Given the workload and snapshot, coordinate workload execution to produce training data

- Observability
- Execution

Workload replay tools

- pgreplay supports speed factor for replay

Trainer

Choice #1: Foreign Data Wrappers

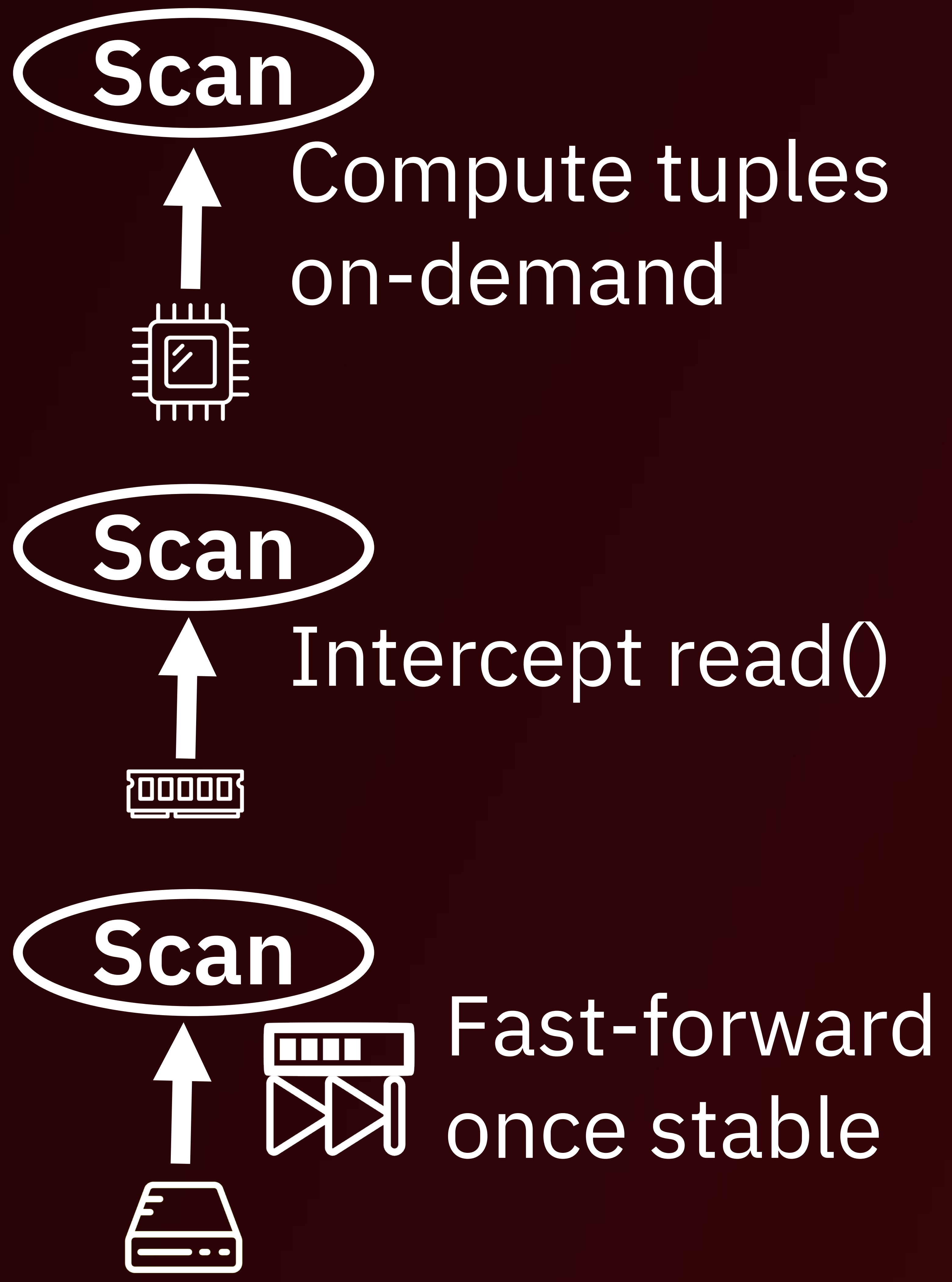
- Save on storage

Choice #2: RAMDISK

- Save on disk access time

Choice #3: Query Progress Estimation

- Save on query execution time

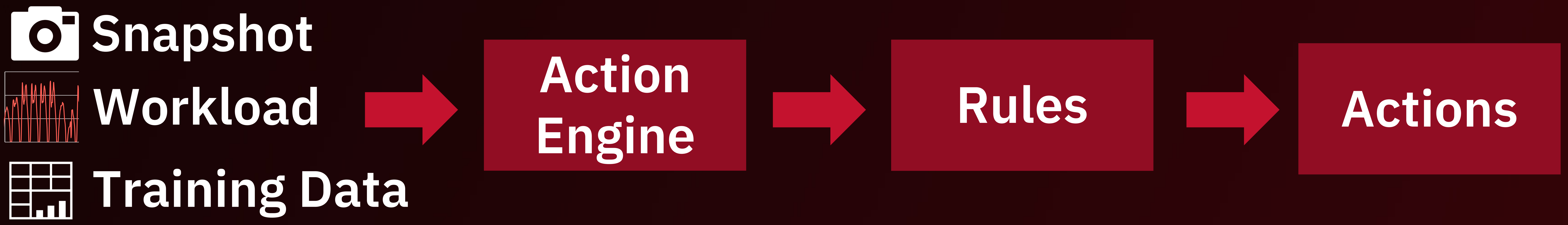


Planner

Suggests a list of promising actions

Extensible Rule-based Action Generation

- Inspired by query optimizers (Exodus, Starburst)
- Example rule: “columns that occur together in a WHERE clause”



Decider

Pick the best action out of a list of candidate actions

By building on the OpenAI Gym, this component is free!

- Leverage what already exists in the ML community
- Various RL libraries integrate easily with gyms

CONCLUSION >

Takeaways

Stop spending our time on ML problems, focus on database problems

Database Gym: systems for machine learning for systems

Hit Me Up

wanshenl@cs.cmu.edu
@capybara@discuss.systems