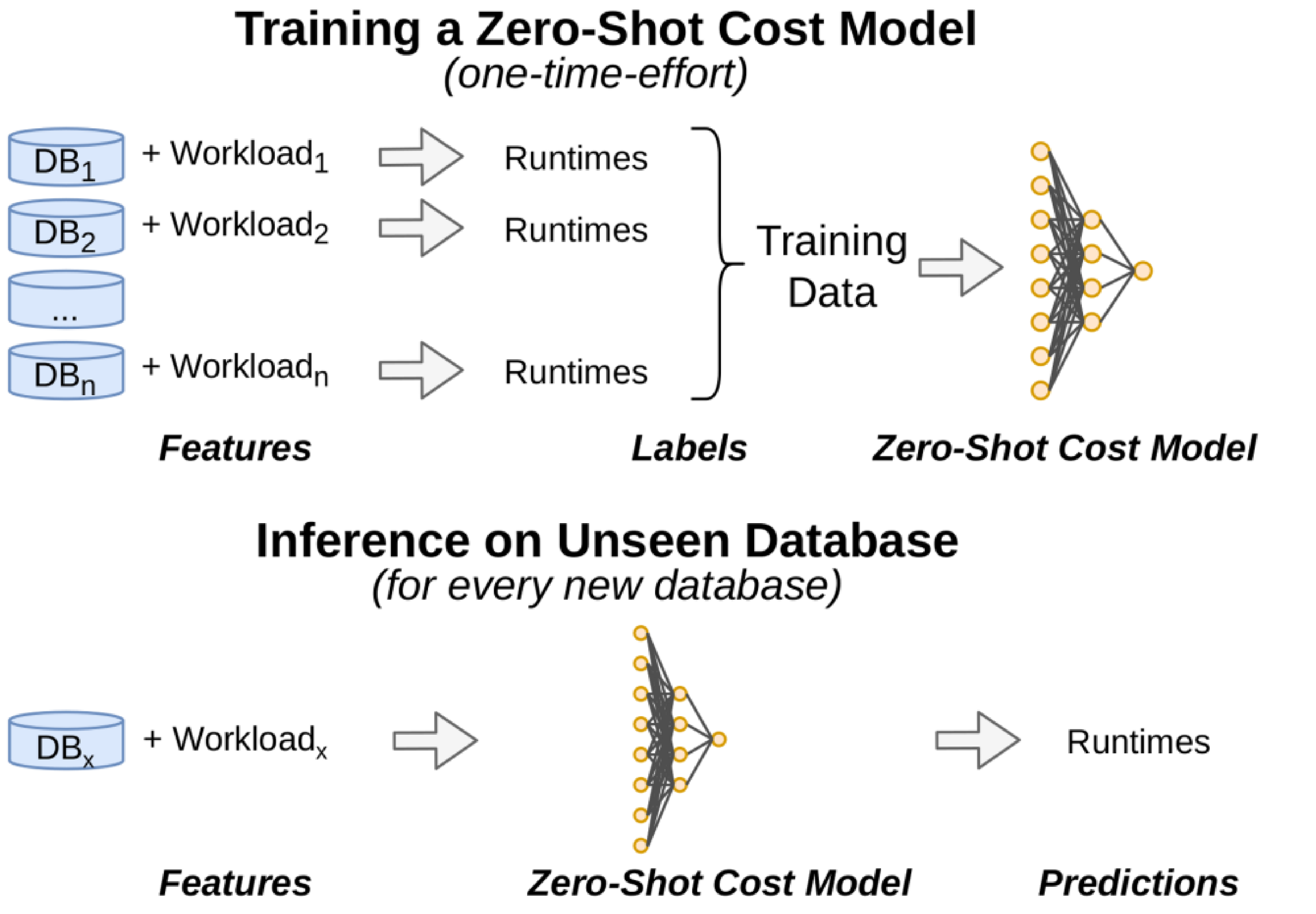
Zero-Shot Learning for Databases

Zero-Shot Learning in the context of databases is the process of training a model on a certain number of databases and then using this model on unknown databases. The model will be trained on runtimes of queries. Zero-Shot learning has the advantage of at least keeping the accuracy of workload-driven training, while also reducing the cost and being more flexible, because the model isn't trained for just one specific database.

For Zero-Shot Learning it is necessary to encode queries in a way that can generalize over databases. This happens by saving information about query operators and the data they are running on in graph nodes. The saved information will just be characteristics (for example tuplewidth or cardinalities) of the data and not the data itself, to still be able to generalize across databases. The cardinalities can be collected by using an approach called data driven learning, without having to run queries.

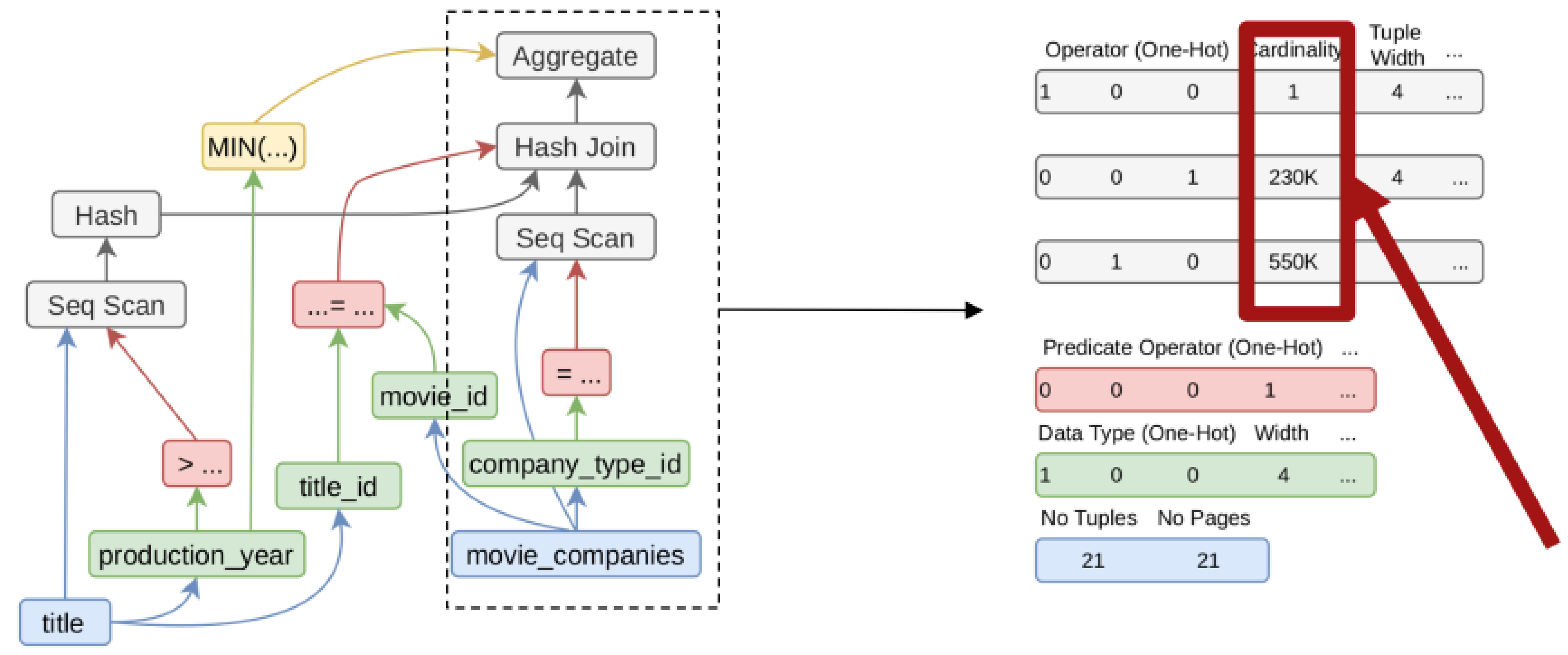

Training a Zero-Shot Cost Model
(one-time-effort)

DB_1 + Workload_1 Runtimes
DB_2 + Workload_2 Runtimes
...
DB_n + Workload_n Runtimes
Training Data
Features
Labels
Zero-Shot Cost Model

Inference on Unseen Database
(for every new database)

DB_x + Workload_x
Runtimes
Features
Zero-Shot Cost Model
Predictions

```
SELECT MIN(t.production_year) FROM movie_companies mc,title t
WHERE t.id=mc.movie_id AND t.production_year>1990 AND
mc.company_type_id=2;
```



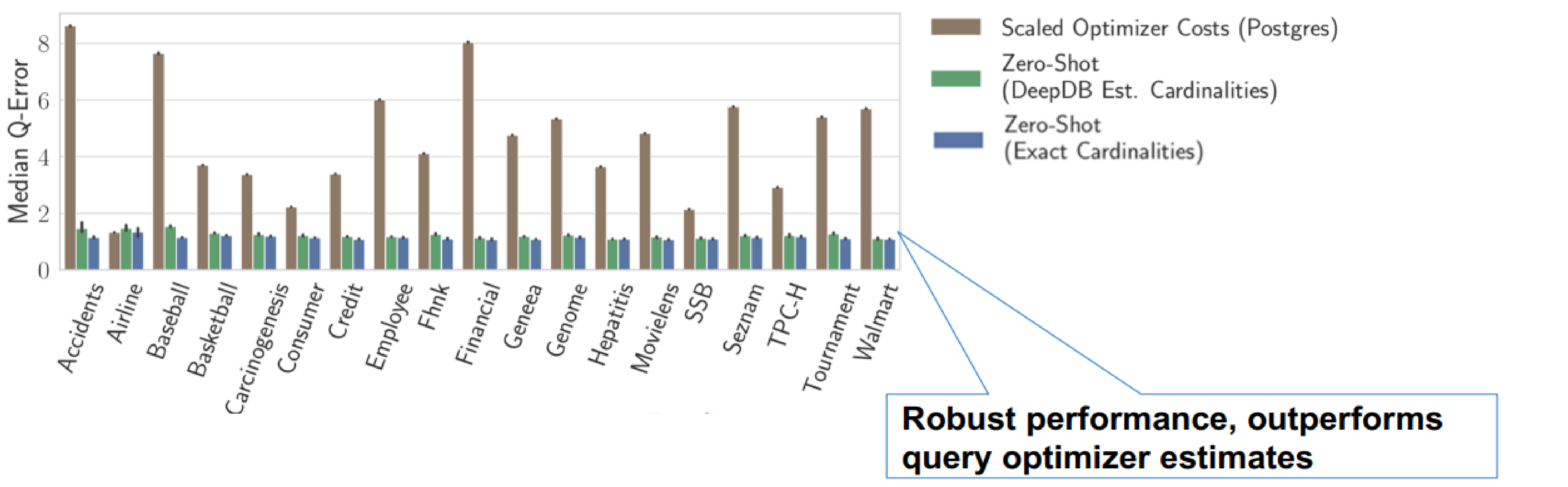
GNN-like architecture → Runtime Prediction

Estimated Cardinalities from optimizer or DeepDB to capture data distribution

As shown by the graphics to the right Zero-Shot Learning has good performance compared to existing approaches for database-independent approaches. It also compares well with workload-driven learning which was one of the goals with Zero-Shot learning.

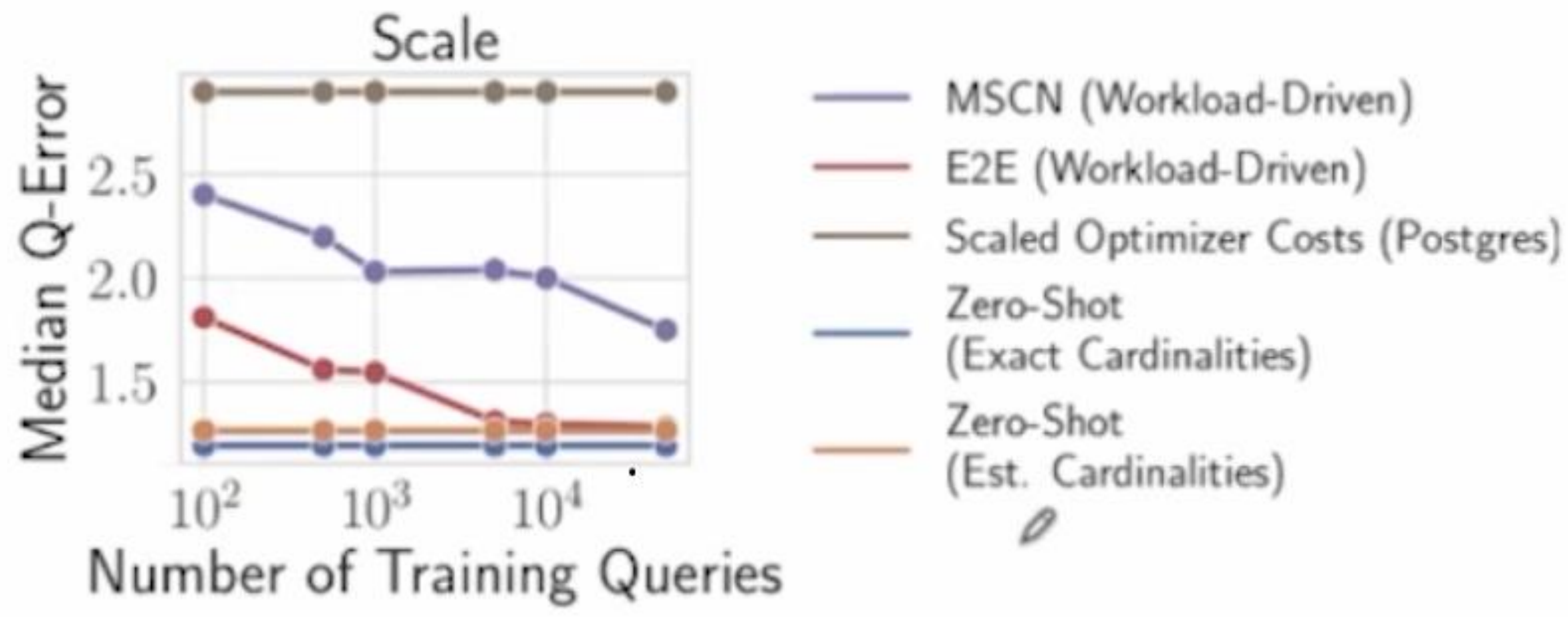
But there are some tasks where Zero-Shot learning can't be used, since the information about a specific database is limited. Therefore one of the future goals of Zero-Shot learning is to go from database-independent learning to task-independent learning.

Generalization experiment: train on 19 real-world DBs. Estimate query runtime on unseen 20th DB



Robust performance, outperforms query optimizer estimates

Comparison with workload-driven Learning:
vary the number of training queries observed for an unseen database and compare runtime prediction accuracy



Florian Krummrey
Bachelor

E-Mail: florian.krummrey@student.hpi.uni-potsdam.de

Credits:
All information and graphics are based on the lecture "Learned DBMS Components 2.0" from Carsten Binning (14.122021)

HPI Hasso Plattner Institut