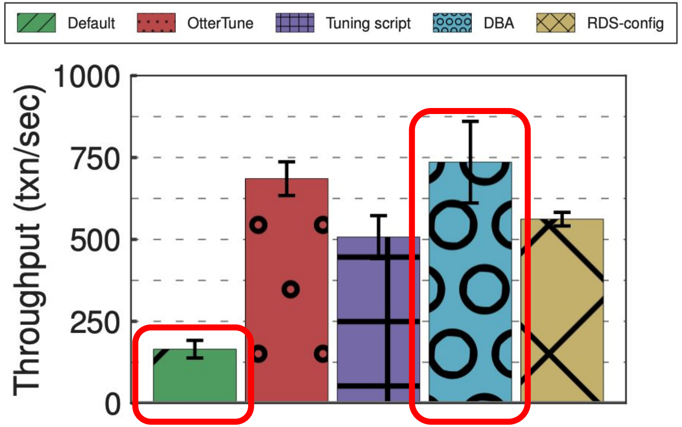
ChimeraTL: Transfer Learning in DBMS with Fewer Samples

Tatsuhiro Nakamori, Shohei Matsuura, Takashi Miyazaki,
Sho Nakazono, Taiki Sato, Takashi Hoshino, Hideyuki Kawashima

Keio University, LYCorporation, Cybozu Labs

Background

- Database parameters have huge impact on performance
  - e.g. buffer pool size
  - **MySQL: about 190, PostgreSQL: about 170**
  - Optimal parameters vs. non-optimal parameters have large difference



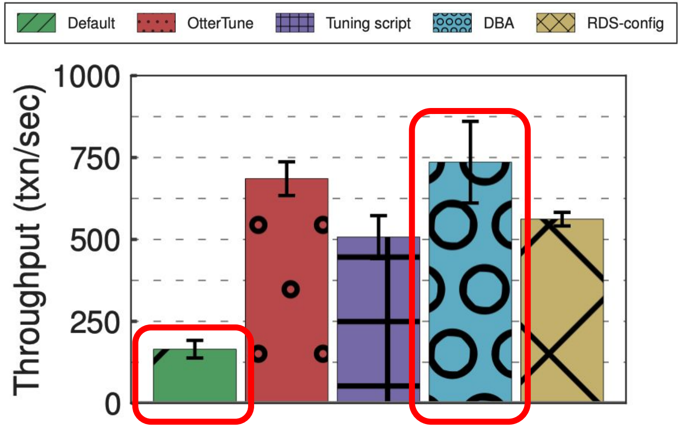
(a) TPC-C (Throughput)

Van Aken et. al. (SIGMOD'17)

Optimal parameter setting
3 times better
than default parameter settings

Background

- Database parameters have huge impact on performance
    - e.g. buffer pool size
    - **MySQL: about 190, PostgreSQL: about 170**
    - Optimal parameters vs. non-optimal parameters have large difference



(a) TPC-C (Throughput)

Van Aken et. al. (SIGMOD'17)

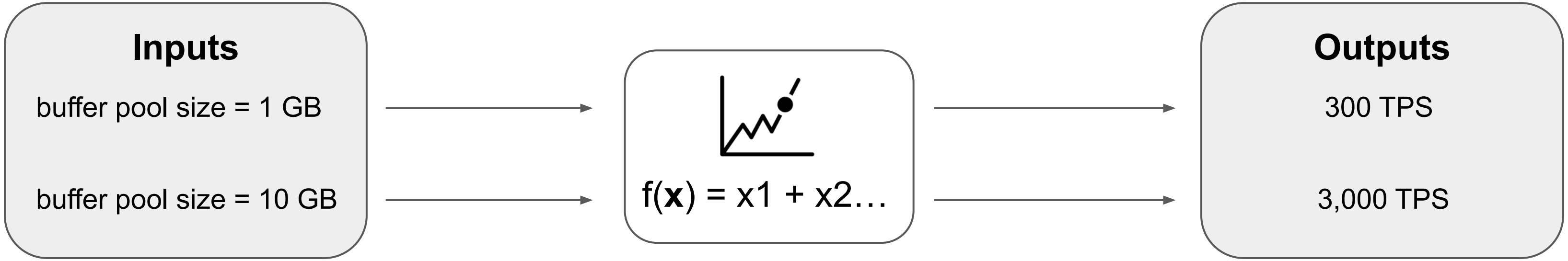
Optimal parameter setting
3 times better
than default parameter settings

Too many parameters
to find optimal parameter
settings manually

Recent approach

How to lower the cost for optimal parameter search?

⇒ Machine learning model



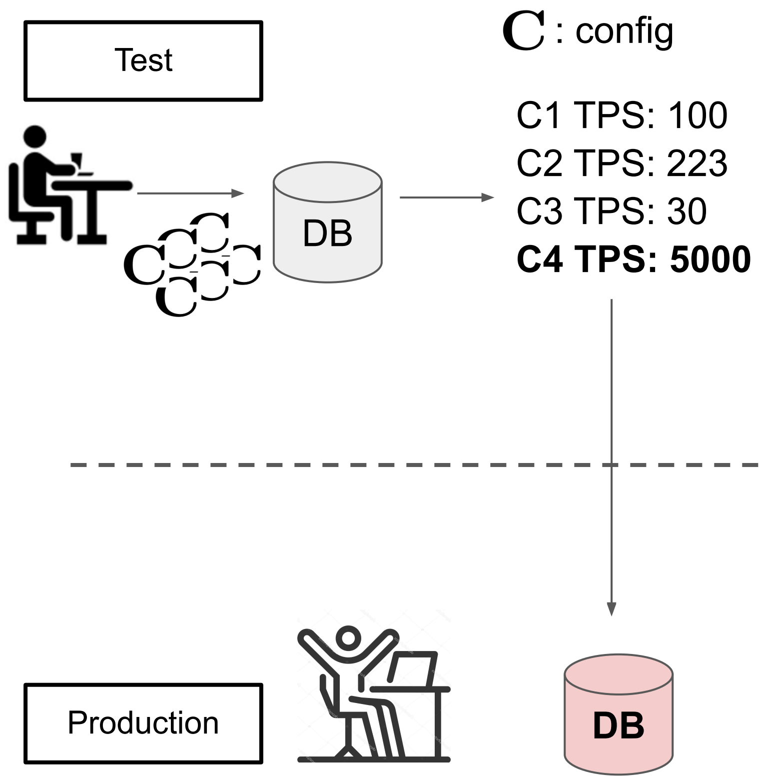
- **Focus of existing studies: create model for fixed hardware environment**
    - DBMS performance depends on hardware limitations

Van Aken et. al. (SIGMOD'17), Zhang et al. (VLDB'22)

Reality

GOAL: find optimal parameter settings in production environment

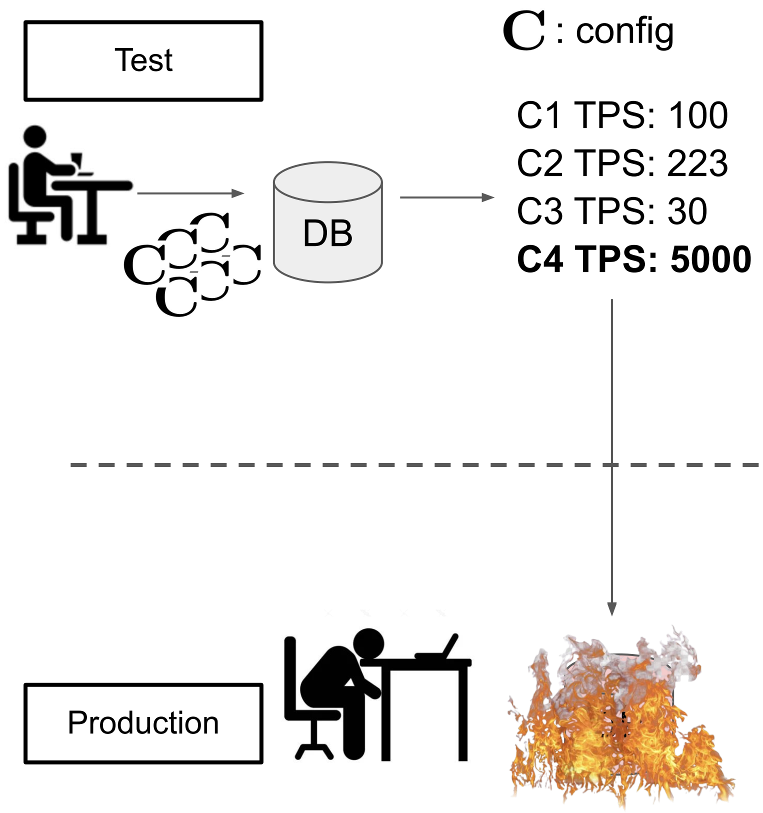
- Costly to collect data in production environment (little data)
- Utilize testing env. to search for optimal parameter settings



Reality

GOAL: find optimal parameter settings in production environment

- Costly to collect data in production environment (little data)
- Utilize testing env. to search for optimal parameter settings
- **Optimal parameter in testing env. may not be optimal in the production env.**



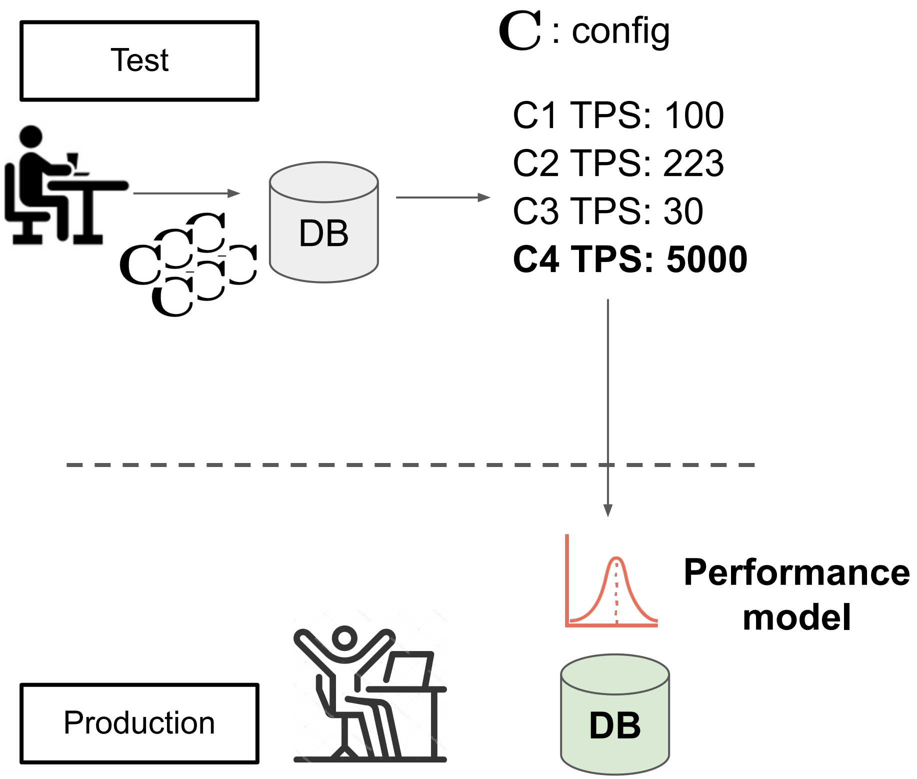
Approach

Significant cost in learning model from scratch for production env.

but…

test model might be inaccurate

Can we exploit readily available test environment data?



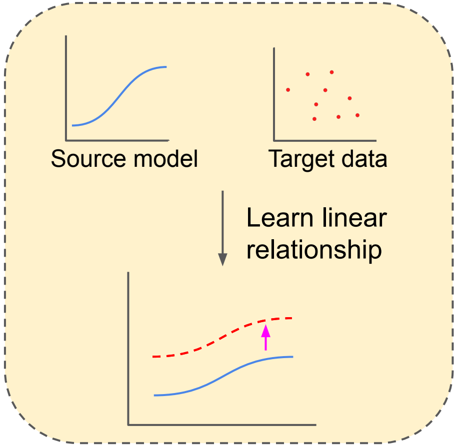
Transfer Learning

Transfer learning methods for configurable systems

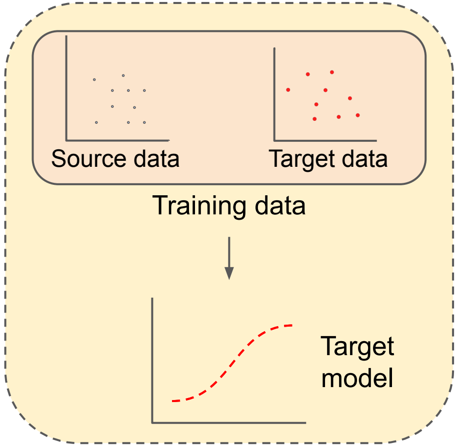
- ModelShift [1]
- DataReuseTL [2]
- Learn to Sample [3]

[1] Valov et al. (ICPE '17)
[2] Jamshidi et al. (SEAMS '17)
[3] Jamshidi et al. (ESEC/FSE ’18)

ModelShift and DataReuseTL



Linear transformation of source model



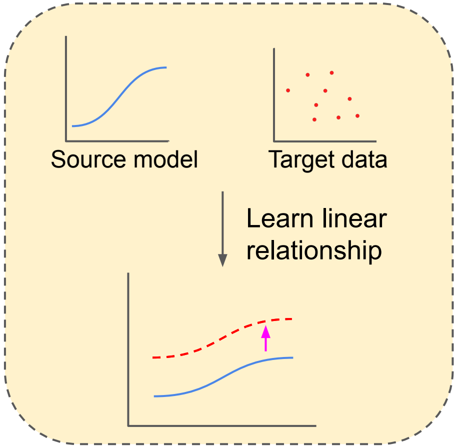
Reuse source data to learn target model

Pro:

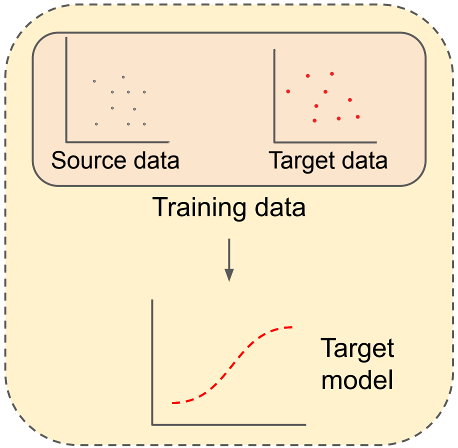
ok prediction with few target data

Con:

ModelShift and DataReuseTL



Linear transformation of source model



Reuse source data to learn target model

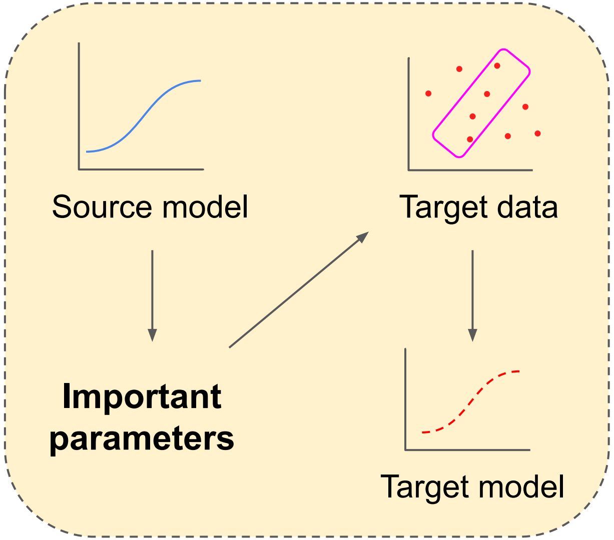
Pro:

ok prediction with few target data

Con:

negative transfer

Learn to Sample (L2S)



1. Select parameters that have significant effect on performance from source
2. Collect data of those parameters in target

Pro: no negative transfer

Con:
not using source data at all
⇒ still need many data from target

The Goal of Transfer Learning

1. Fewer samples from target
2. Minimize negative transfer caused by source

Existing methods suffer from the trade-off

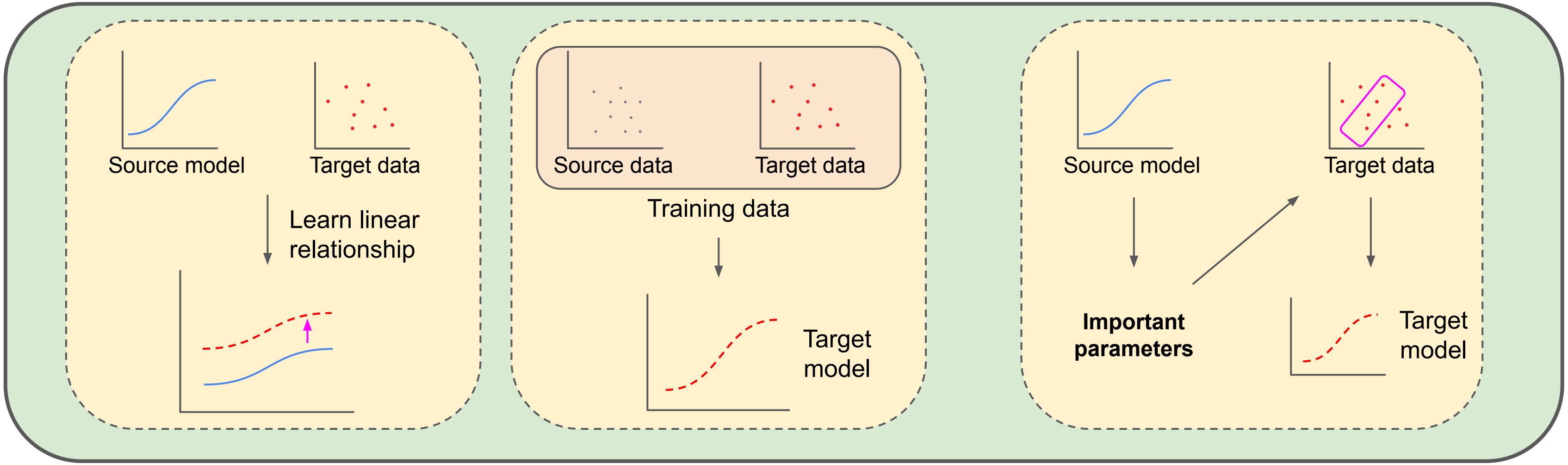
Methods	**Minimize sample size**	**Minimize negative transfer**
ModelShift	Good ✅	Bad ❌
DataReuseTL	Good ✅	Bad ❌
Learn to Sample (L2S)	Ok	Good ✅
Goal	Good ✅	Good ✅

How?

Proposal: ChimeraTL

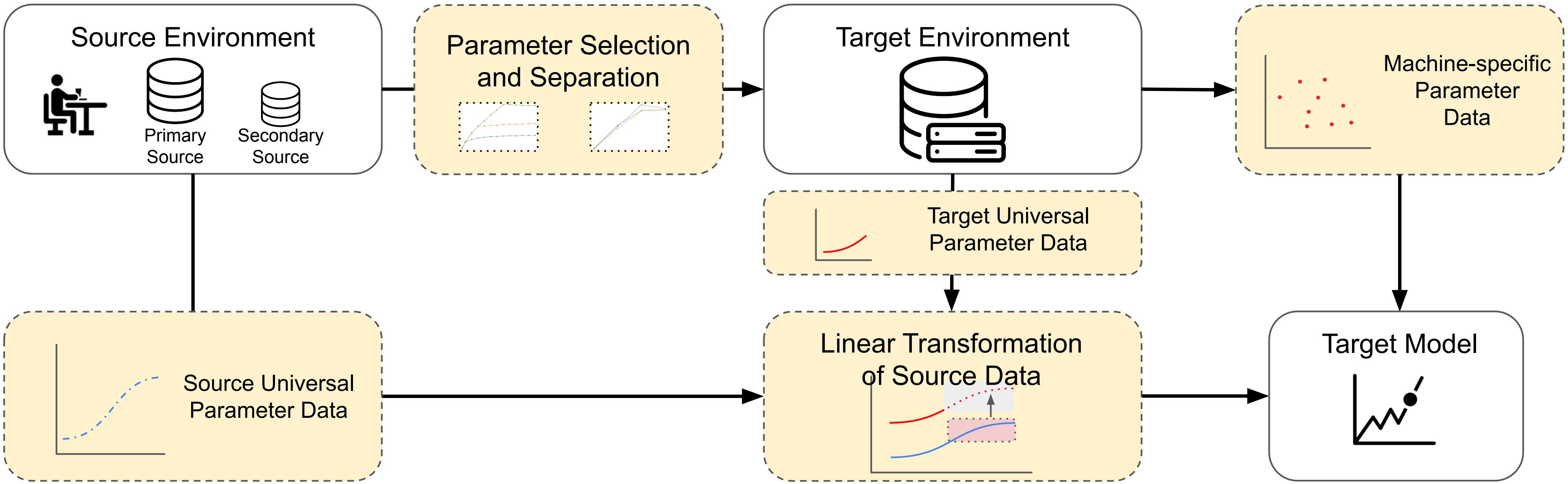
Novel approach to maximize the source data in transfer learning

Key:
Linear transformation is only applied to similar data
Machine-specific behavior is learned only from target



ChimeraTL Pipeline

1. Parameter selection and separation
2. Linear transformation learning
3. Machine-specific parameter learning

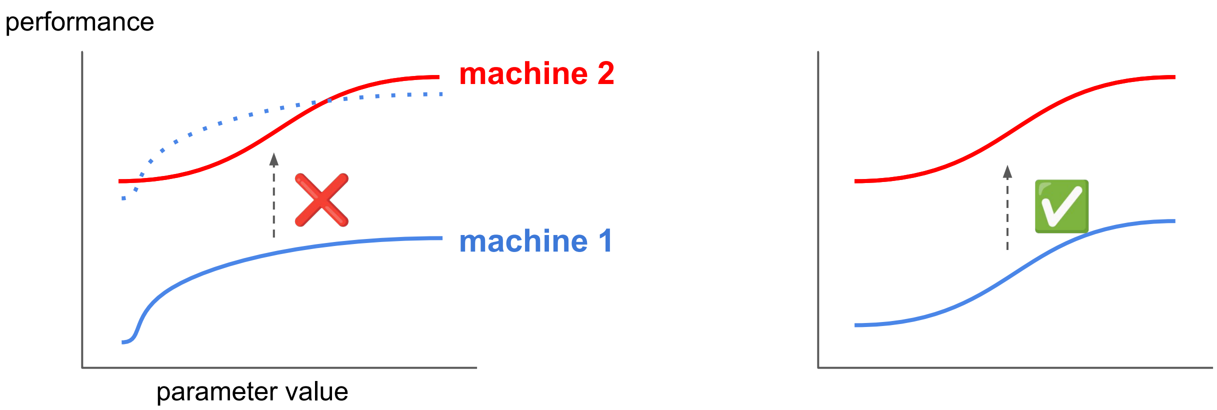


Universal and machine-specific parameters

Two types of parameters

1. **Universal parameters:** similar performance effects across different environments
2. **Machine-specific parameters:** different performance effects depending on the hardware limitation

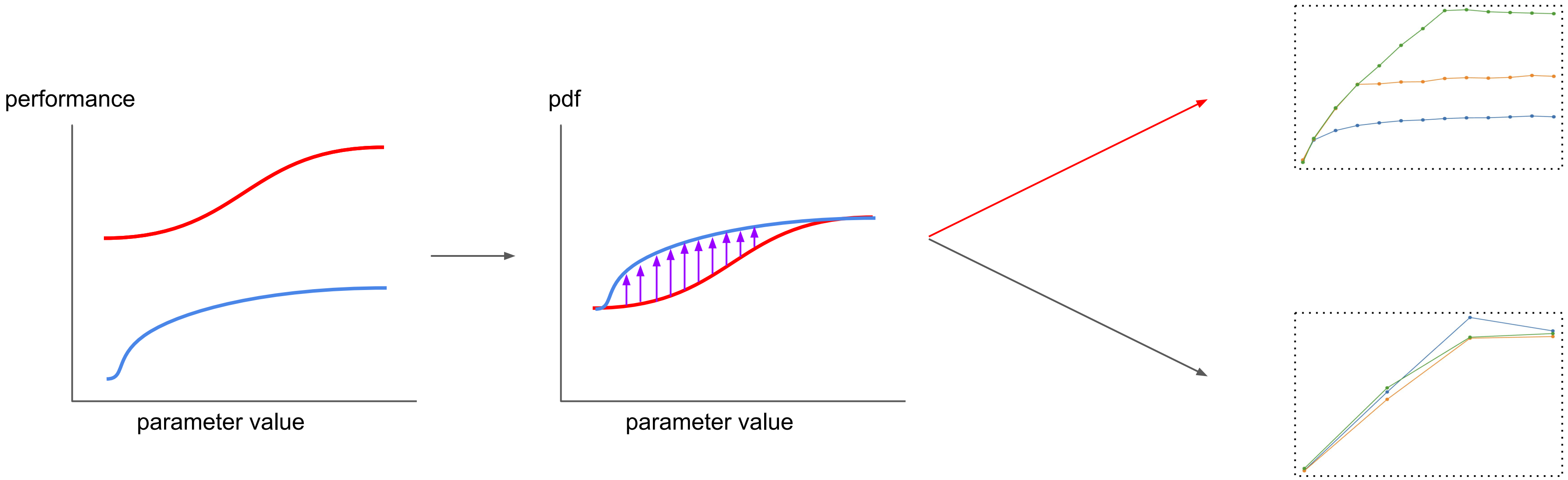
Linear transformation only works for performance of universal parameters



Parameter Separation

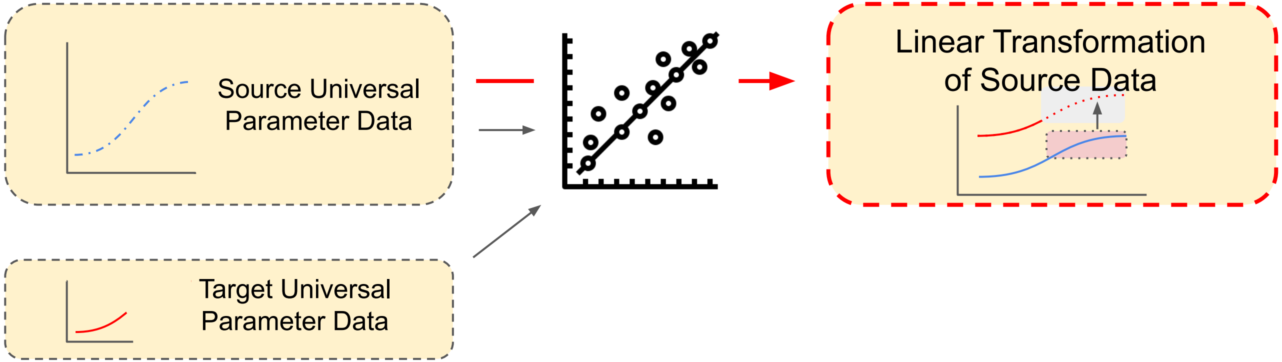
Separate the parameters into **universal** and **machine-specific (never done in previous studies)**

How? ⇒ convert performance functions to probability distributions and calculate distance



Linear Transformation Learning

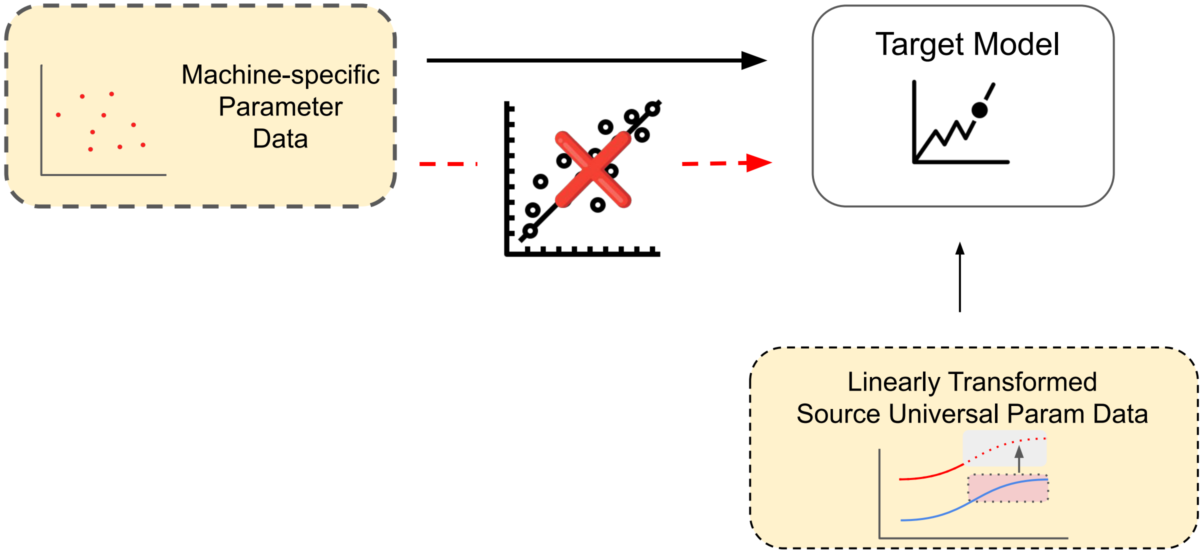
1. Sample *5-10* universal parameter data from the target
2. Fit linear regression model on source and target universal parameter data
3. Transform source data to estimate target env performance



Machine-specific Parameter Learning

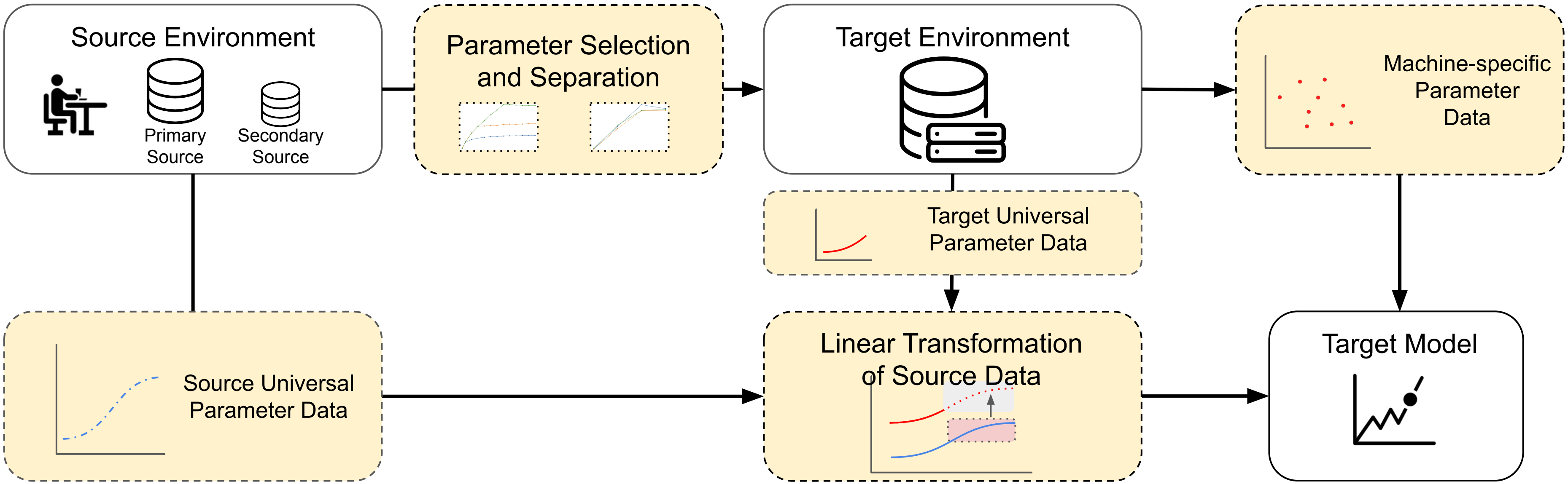
Redundant to sample universal parameter data after linear transformation

⇒ Prioritize sampling machine-specific parameter data to learn trends not observable in source



ChimeraTL Pipeline

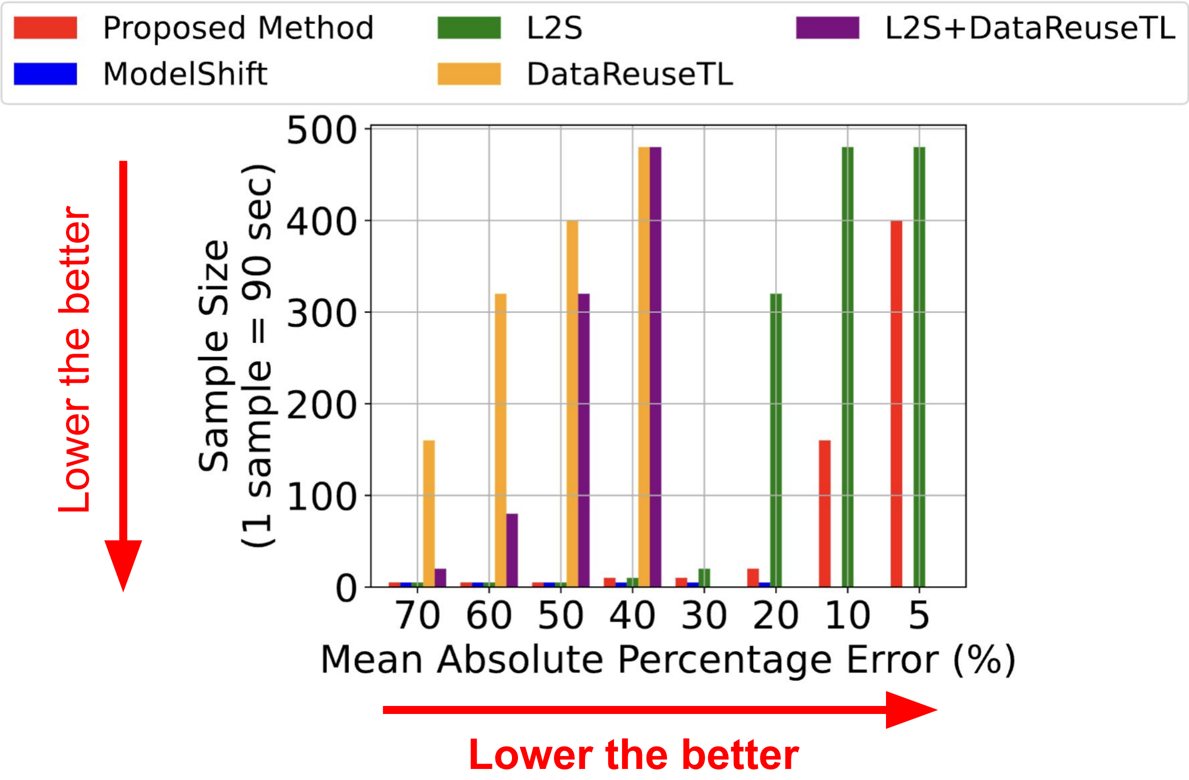
1. Parameter selection and separation
2. Linear transformation learning
3. Machine-specific parameter learning



Evaluation

Database	MySQL Version 8.0.35
Source Environment	8 core, 12 GB docker container on Macbook
Target Environment	24 core, 32 GB docker container on large scale server

Comparison with State-of-the-art Methods

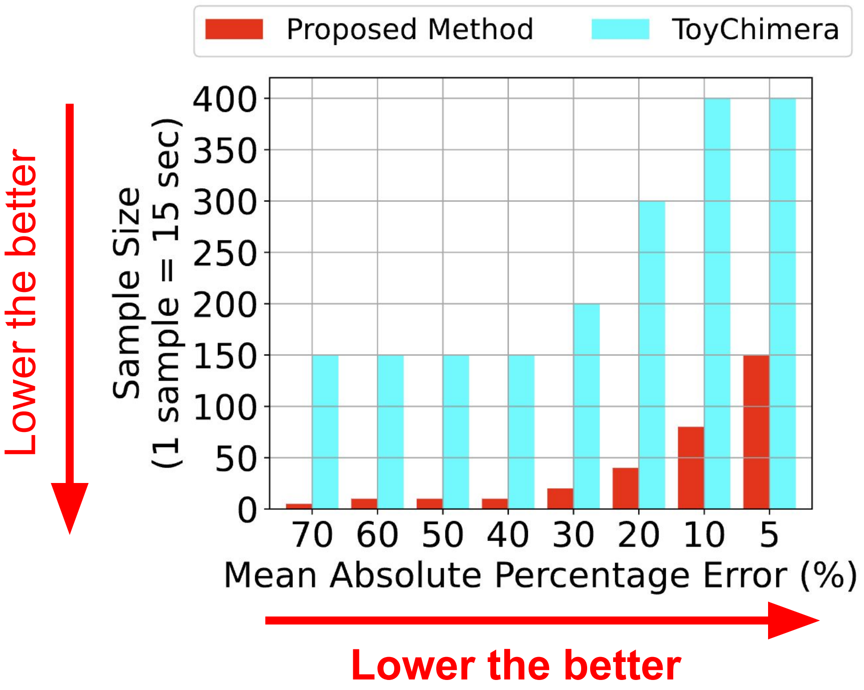


ModelShift and DataReuseTL cannot reduce the error to below 10% because of negative transfer

L2S requires 3x more samples than ChimeraTL to reduce to below 10% error

ChimeraTL maximizes source data while minimizing negative transfer

The Impact of Parameter Separation



ToyChimera:
ChimeraTL without parameter separation

Parameter separation enables

- **faster and more accurate linear transformation**
- **faster machine-specific parameter learning**

Conclusion

- ★ **Performance models are useful for finding optimal configurations in DBMS**
- ★ **Transfer learning is effective to reduce target environment model learning cost**
- ★ **Previous methods do not leverage source data appropriately**

- ★ **ChimeraTL**

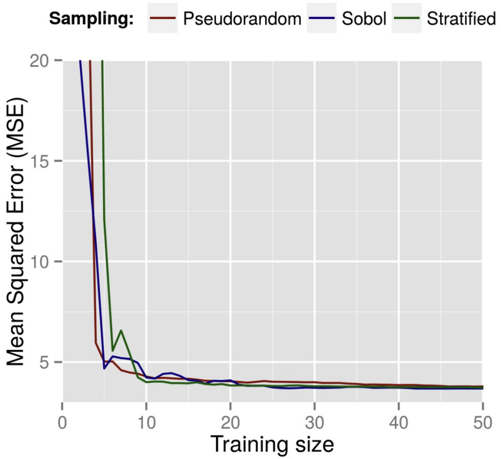
  Novel transfer learning approach based on **parameter separation** to maximize source data while minimizing negative transfer

- ★ **Result: ~70% reduction in time to accomplish same accuracy**

How much data needed?

1. Sample N target universal parameter data
2. Fit linear regression model on source and target universal parameter data (ModelShift)
3. Transform source data for later model training (DataReuseTL)

Previous study shows that N = 5~10 is enough to fit appropriate regression model



we can learn the performance curve of universal parameters with just 5 samples

P. Valov, J.-C. Petkovich, J. Guo, S. Fischmeister, and K. Czarnecki, “Transferring performance prediction models across different hardware platforms,” in Proc. ICPE, 2017.