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

Author : Lin Ma, Dana Van Aken, Ahmed Hefny,
Gustavo Mezerhane, Andrew Pavlo, Geoffrey J. Gordon

Min-Han Tsai
2018/8/31

# Outline

- Background
- Issue & Difficulty
- Main idea
- Work flow
  - -Target I/O
  - -Real-word application
  - -Describe
  - -Model

# Background

- Increasing complexity of DBMS and data-driven application.
- The advancements of the storage and computational hardware allows self-driving DBMS to be promising.
- The tech of Deep Learning.

# Workload

In database, we analyze the number of queries executed by the database in a given period of time.

# Issue & Difficulty

## Optimization

Determining the Optimization by target application workload is very necessary

## Past Data

If DBMS only consider the behavior of the application in the past, it may cause resource contention.

## Time

The workload trends change over time. How to deal with the different arrival rate queries?

## Previous Work

Previous works are unable to generate an adequate method for an autonomous system.
Ex resource, workload shift

# Main idea

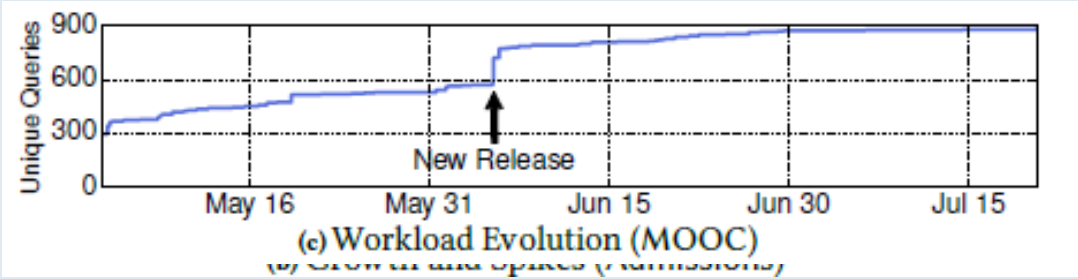
Predicting one general model under the different arrival rate.

Reducing the complexity of workload, and not lose the accuracy.

Dealing with the combinational types of queries and changeful queries.

# Real-word Database applications

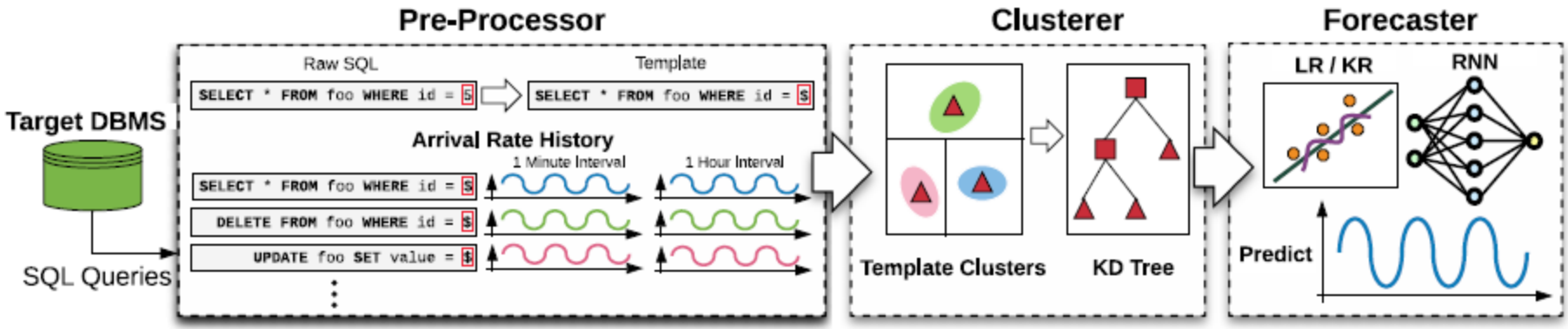
## Workload Evolution: MOOC




(c) Workload Evolution (MOOC)

If only I could predict workload, DBMS could dynamic select the optimization……

# 3.
# Work Flow




- The prediction of workload.

| | Admissions | BusTracker | MOOC |
|---|---|---|---|
| DBMS Type | MySQL | PostgreSQL | MySQL |
| Num of Tables | 216 | 95 | 454 |
| Trace Length (Days) | 507 | 58 | 85 |
| Avg. Queries Per Day | 5M | 19.9M | 1.1M |
| Num of SELECT Queries | 2541M [99.8%] | 19.5M [98%] | 0.97M [88%] |
| Num of INSERT Queries | 1.8M [0.07%] | 15K [0.8%] | 14K [1.3%] |
| Num of UPDATE Queries | 2.6M [0.1%] | 22K [1%] | 66K [6%] |
| Num of DELETE Queries | 0.4M [0.02%] | 3K [0.2%] | 51K [4.7%] |

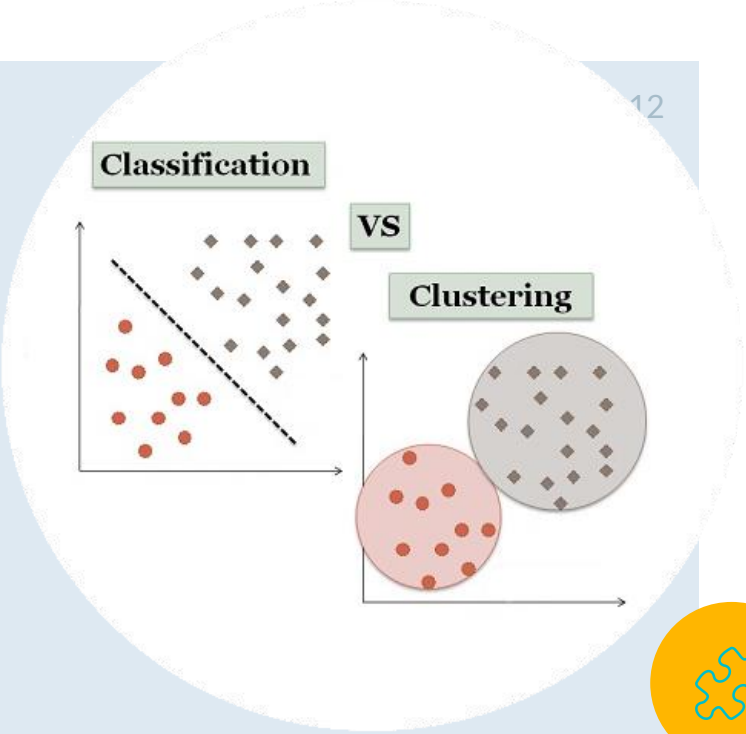
**Table 1: Sample Workloads** – Summarization of the workload traces collected from the database applications described in Section 2.1.

# How To Describe The Data?

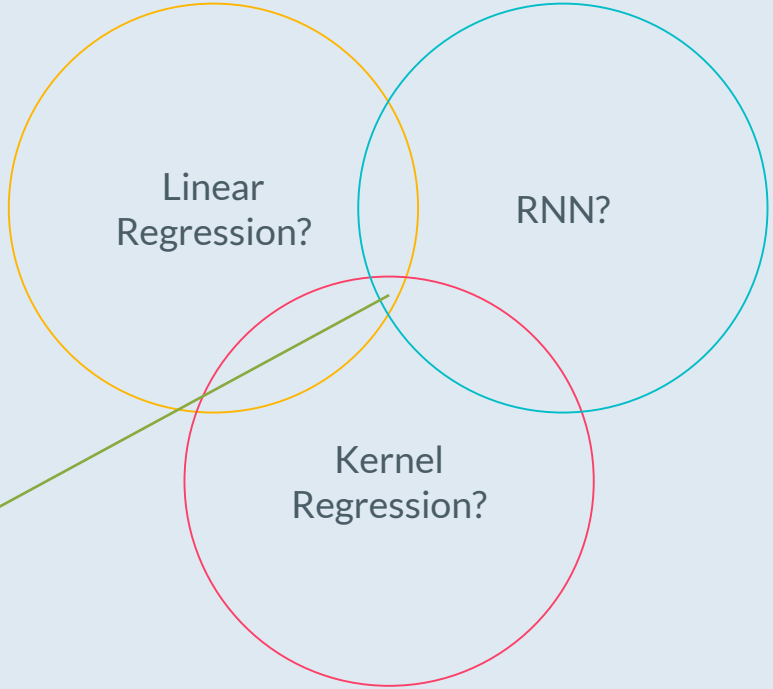
Classify

v.s.

Clustering



# What kinds of model is better ?




Can we exploit them all?

# Now it's time for brainstorming !

Success grows out of struggles we overcome in difficult !

# Reference

- https://zhuanlan.zhihu.com/p/37182849
- https://github.com/malin1993ml/QueryBot5000