Explaining Aggregates for Exploratory Analytics

Fotis Savva, Christos Anagnostopoulos & Peter Triantafillou

University of Glasgow, UK

IEEE Big Data 2018 @Dec 10-13, 2018, Seattle, WA, USA

Outline

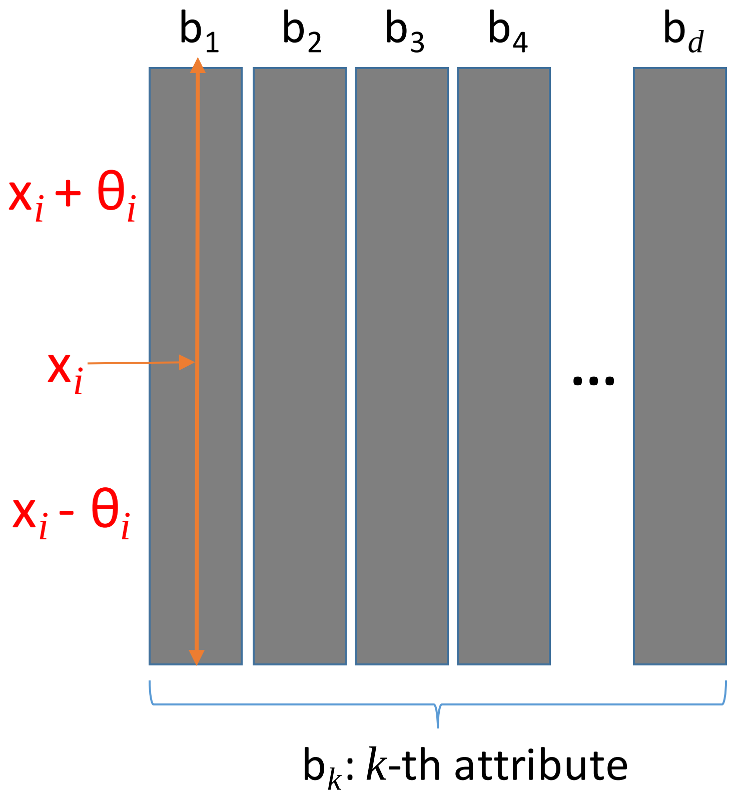
- Motivation
- Preliminaries and Overview
- Explanations as functions
- Query-Driven learning for constructing explanations
  - Preprocessing
  - Online Training
  - Explanation Mode
- Experimental Evaluation Results

- Data Analysts have to make sense of data by engaging in **E**xploratory **D**ata **A**nalysis (**EDA**)[1].

- Reviewed ***Kaggle*** kernels; analysts first get an *overview* of the data and then *zoom-in.*[2] *(Goal is to create predictive models.)*

- Viewed as; *Aggregate Queries* executed over different *ranges*

	epoch	moteid	temperature	humidity	light	voltage	c
count	2.313682e+06	2.313156e+06	2.312781e+06	2.312780e+06	2.219804e+06	2.313156e+06	2.313682e+06
mean	3.303993e+04	2.854412e+01	3.920700e+01	3.390814e+01	4.072110e+02	2.492552e+00	1.079146e+09
std	1.836852e+04	5.062408e+01	3.741923e+01	1.732152e+01	5.394276e+02	1.795743e-01	7.887828e+05
min	0.000000e+00	1.000000e+00	-3.840000e+01	-8.983130e+03	0.000000e+00	9.100830e-03	1.077930e+09
25%	1.757200e+04	1.700000e+01	2.040980e+01	3.187760e+01	3.956000e+01	2.385220e+00	1.078475e+09
50%	3.332700e+04	2.900000e+01	2.243840e+01	3.928030e+01	1.582400e+02	2.527320e+00	1.079078e+09
75%	4.778900e+04	4.100000e+01	2.702480e+01	4.358550e+01	5.372800e+02	2.627960e+00	1.079764e+09
max	6.553500e+04	6.540700e+04	3.855680e+02	1.375120e+02	1.847360e+03	1.856000e+01	1.081163e+09



$$\mathbf{q}_i = (\mathbf{x}_i, \theta_i)$$

$$\mathbf{x} \in \mathbb{R}^d, \theta \in \mathbb{R}$$

*"Our goal is to **provide** efficient explanations for aggregate queries and to **assist** analysts in EDA by providing insight."*

Some Notation First

- Data can be considered as random row **vectors**

$$\mathrm{b} = [\mathrm{b}_1, \ldots, b_d] \in \mathbb{R}^d$$

- We consider queries with a *Center-Radius* Selection (CRS) operator

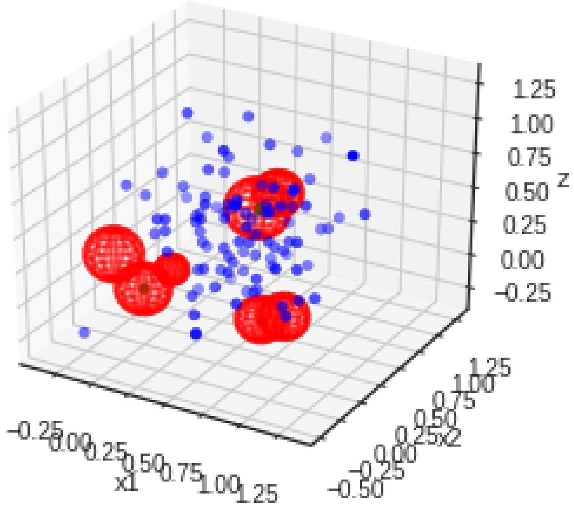
$$\mathrm{q} = (\mathrm{x}, \theta), \quad \mathbf{x} \in \mathbb{R}^d, \theta \in \mathbb{R}$$

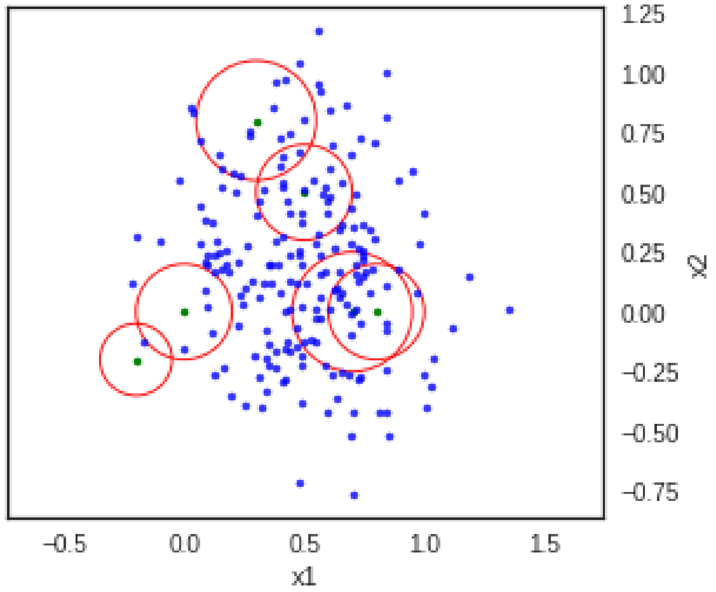
- Essentially a CRS defines a *data-subspace*

$$\mathbb{D}(\mathbf{x}, \theta) \qquad \mathbf{b} : ||\mathbf{x} - \mathbf{b}||_2 \leq \theta$$

- Aggregate Query *as* a function over a defined data-subspace

$$y = f(\mathbb{D}(\mathbf{x}, \theta))$$





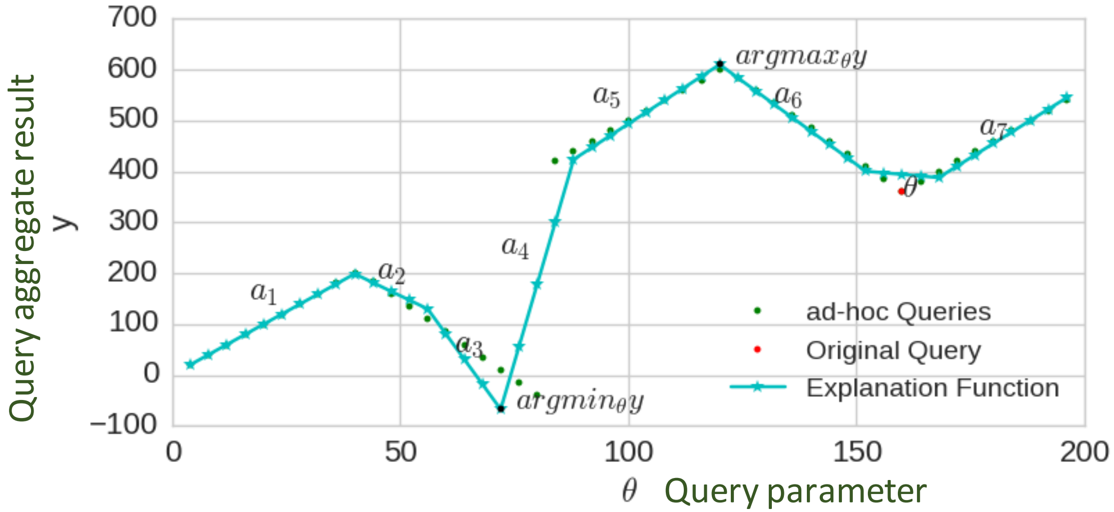
ExF: Explanations *as* Functions

- **Understanding** the **data generation process**; e.g., How data points increase in number in a particular area in *spatial analytics.*

- **Exploit** the function f for prediction **instead of** computing aggregate queries.

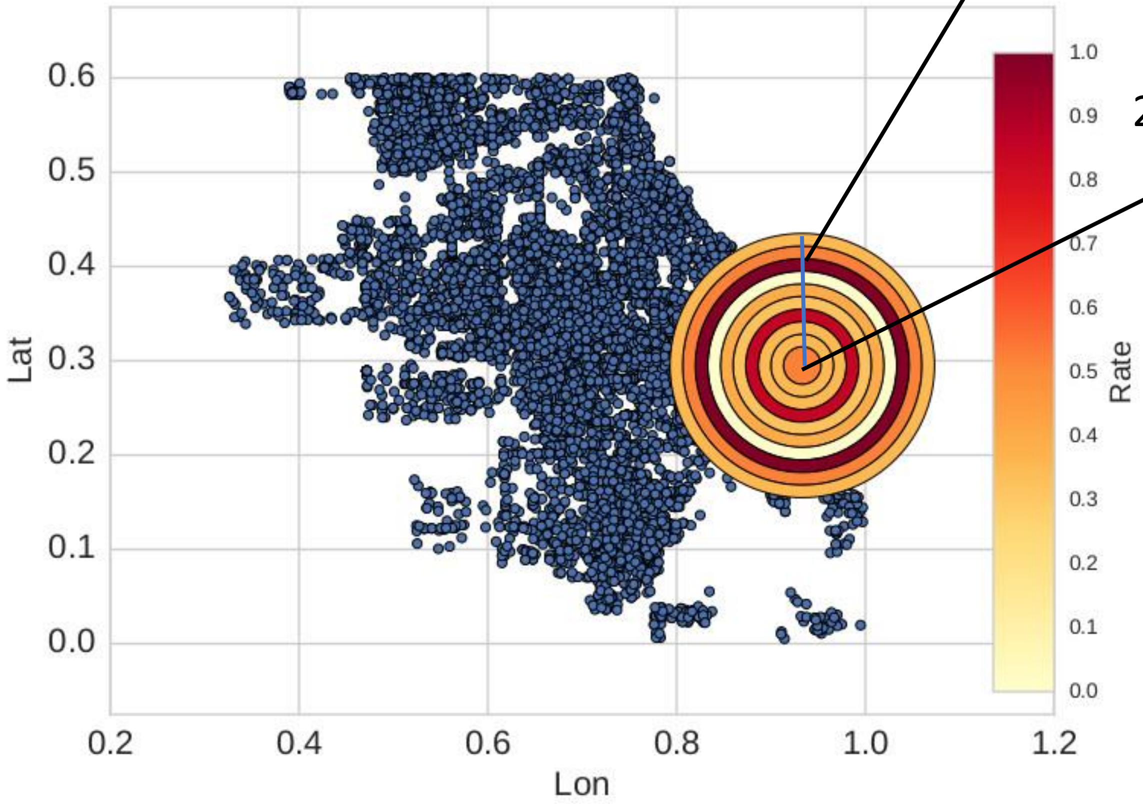
- **Solve** optimizations efficiently, i.e., approximating *minima* and *maxima* is trivial.

- **Give** insights as to what the **rate-of-change** *(a_i's)* are for an Aggregate given different parameters *(θ).*



Example: Explanation Function as a Piecewise-Linear Regression Model.

Example



Radius : $\theta \in \mathbb{R}$

2-d query center $\mathbf{x} \in \mathbb{R}^2$

Concentric circles denote the **rate-of-change** for the aggregate query : $\alpha_1, \ldots, \alpha_l$

Formal Definition for ExF

Given **Query-Answer** pairs of the form :

$$\mathbf{q} = (\mathbf{x}, \theta, y), \quad \mathbf{x} \in \mathbb{R}^d, \theta \in \mathbb{R}, y \in \mathbb{R}$$

seek a *function* that approximates the ***true*** *function* defined by the aggregate queries

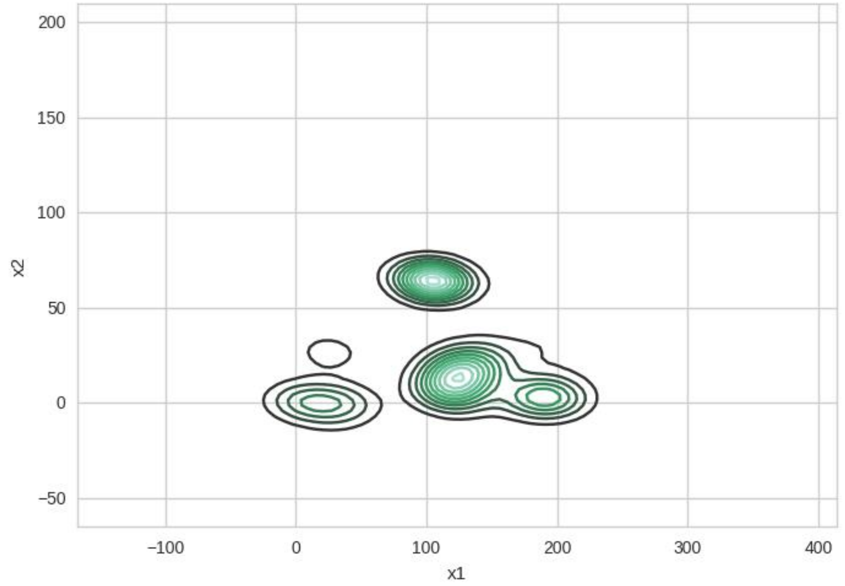
$$f(\mathbb{D}(\mathbf{x}, \theta)) \approx f(\theta; \mathbf{x})$$

Objective: minimize the Expected Explanation Loss (EEL)

$$\hat{f}^* = \arg\min_{\hat{f} \in \mathcal{F}} \int_{\mathbf{x} \in \mathbb{R}^d} \int_{\theta \in \mathbb{R}_+} \mathcal{L}(f(\theta; \mathbf{x}), \hat{f}(\theta; \mathbf{x})) p(\theta, \mathbf{x}) d\theta d\mathbf{x},$$

Objective Revisit

- Evidence: queries form **clusters**; ref: *real workload [3],*
- Hence, our idea is to fit **local** explanation functions over *optimal* **groupings** of queries instead of a **global** one.



Revisited Objective: minimize the Expected Explanation Loss (EEL) via local explanation functions

$$\mathcal{J}_0(\{\hat{f}_k\}) = \sum_{\hat{f}_k \in \mathcal{F}} \int_{\mathbf{q} \in \mathbb{Q}_k \subset \mathbb{R}^{d+1}} \mathcal{L}(f(\theta;\mathbf{x}), \hat{f}_k(\theta;\mathbf{x})) p_k(\mathbf{q}) d\mathbf{q}$$

Short: Identify the evolving behavior of aggregate queries w.r.t parameter values, **without** *accessing any data*.

How? Overview

- **Query-Driven approach**
- *Use past and incoming queries* ***q*** *to solve the revisted optimization problem.*

Initial
LRs,RRs,LPMs

Trained
LRs,RRs,LPMs

Pre-Processing Mode
Clustering of Query-Space,
Initialization of LPMs, LRs, RRs

Training Mode
Training of the LRs and RRs
and LPMs

Explanation Mode
Framework is ready to explain
aggregate queries

1. Obtain optimal groupings and fit PLRs

2. Adjust groupings and models

3. Provide explanations

Location Representatives (LR)

$$\mathbf{w}_1, \dots, \mathbf{w}_k$$

Radius Representatives (RR)

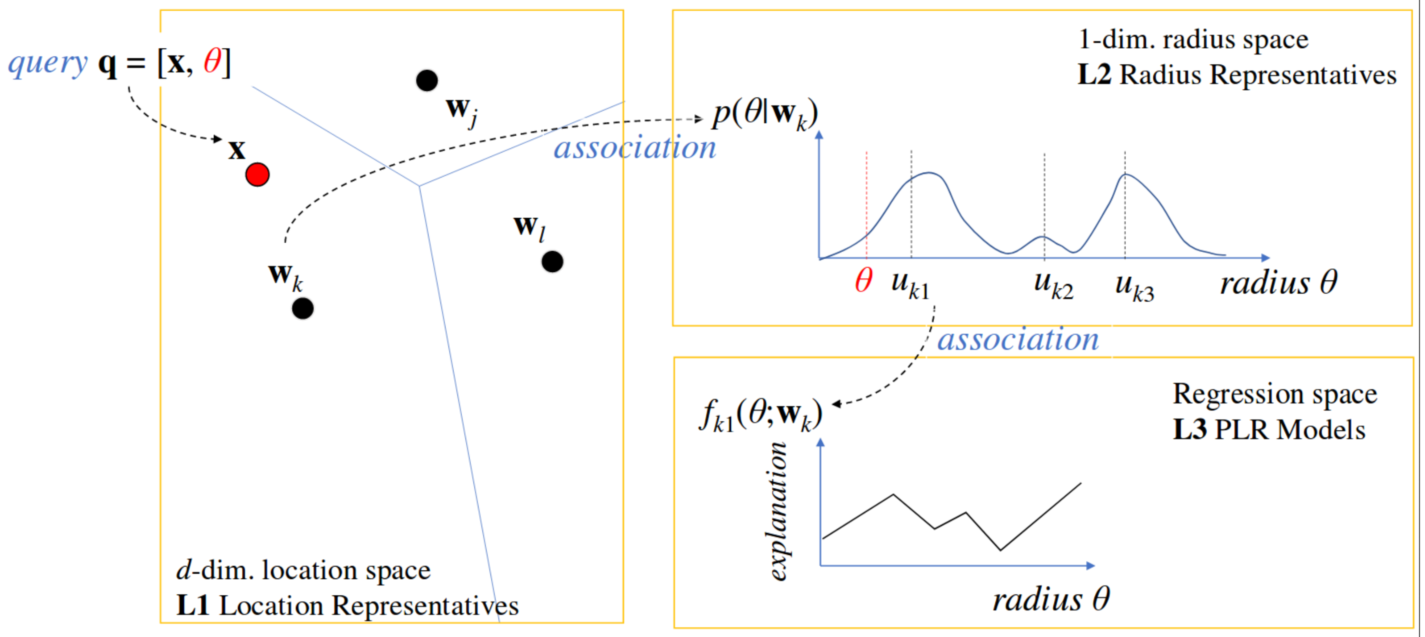
$$u_{1,1}, \dots, u_{1,l} \qquad u_{k,1}, \dots, u_{k,l}$$

Local Piecewise-Linear Models (LPM)

$$f_{1,1}, \dots, f_{1,l} \qquad f_{k,1}, \dots, f_{k,l}$$

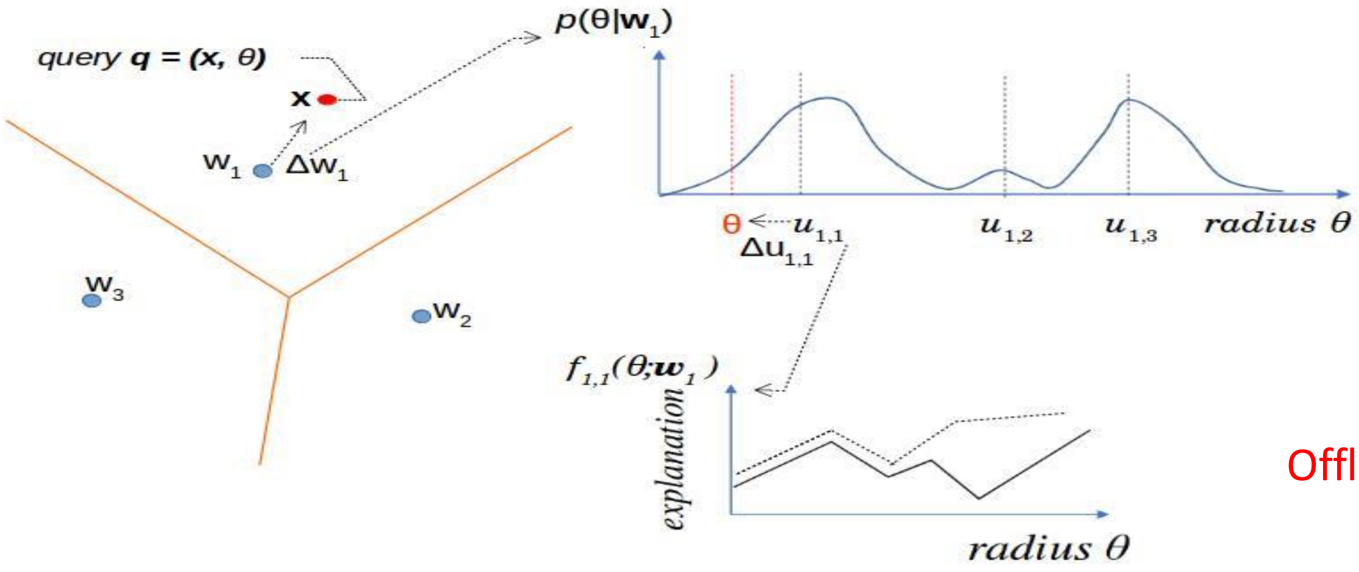
How? Pre-processing Phase

- Initialize groupings and PLRs using *Pre-Processing Step.*
- Using *K-Means [4] to* partition the Query Space :
    1. On query centers **x** (extract location representatives **w**)
    2. On query radii θ (extract radii representatives u)
- Using MARS [5] to fit PLR models on radii



How? Training Phase

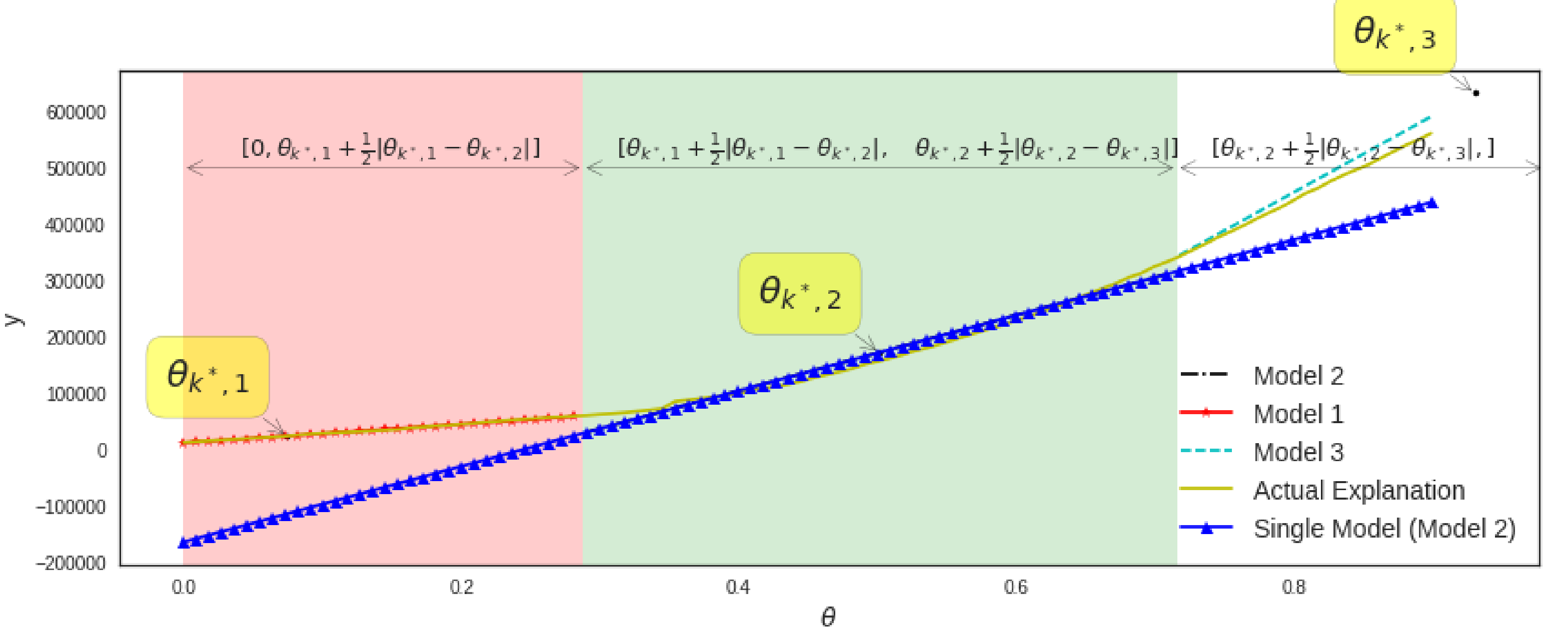
- Refine the optimal parameters **on-line**
- For every new executed query, **adjust** associated groupings & model.



Offline Adjustment of PLR Models

Explanation Mode

- As multiple models are fitted, explanation function alternates between different functions for an ever increasing radius.



Experimental Evaluation

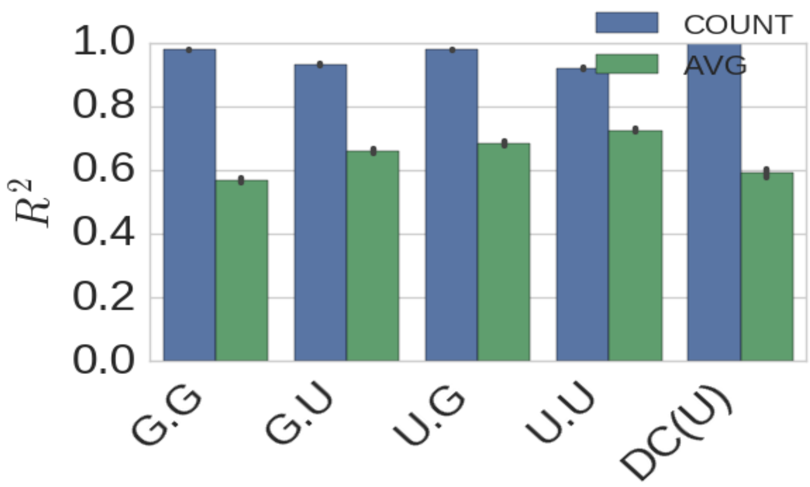
- Evaluate **accuracy** and **efficiency** of the proposed method.

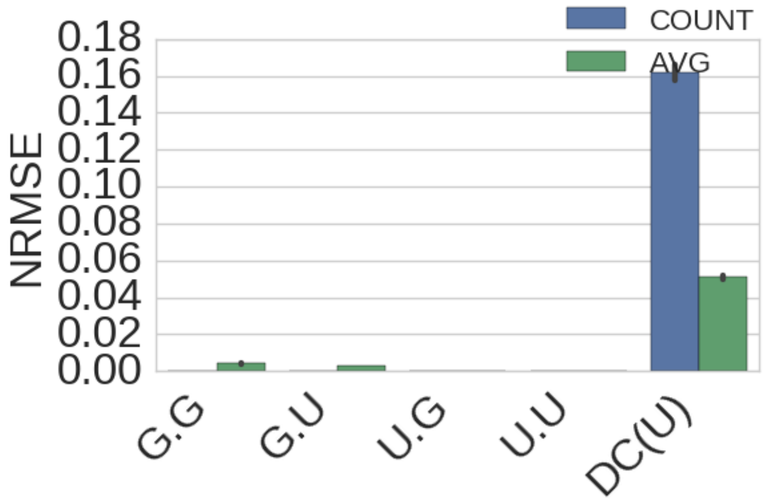
- Construct synthetic query workloads over real datasets.
  - Synthetic query workloads simulate exhibited user behavior.

- Measure how well our model **approximates** the true function and whether it can provide answers to aggregate queries; Coefficient-of-Determination (R^2) and NRMSE.

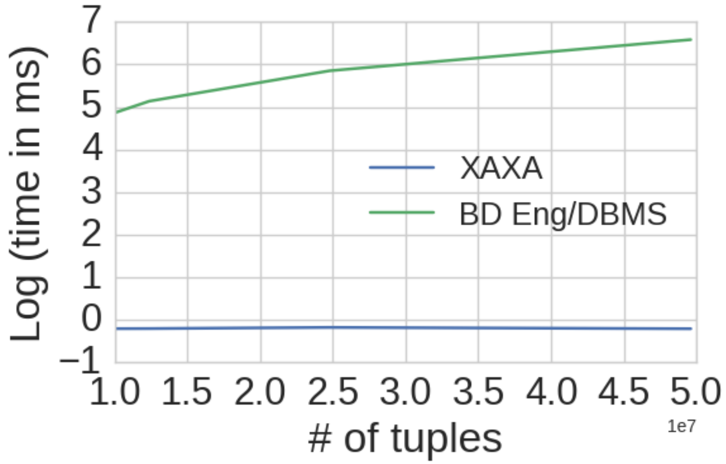
- Measure **efficiency** for training and **explanation provision**.

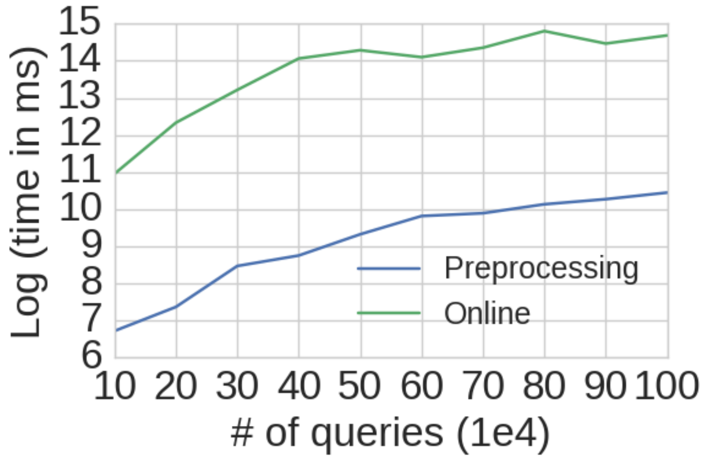
Accuracy





Efficiency





Thank you for your attention.

Questions?

References

- [1] S. Idreos, O. Papaemmanouil, and S. Chaudhuri. Overview of data exploration techniques. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pages 277–281. ACM, 2015.
- [2] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In The Craft of Information Visualization, pages 364–371. Elsevier, 2003
- [3] A. S. Szalay, J. Gray, A. R. Thakar, P. Z. Kunszt, T. Malik, J. Raddick, C. Stoughton, and J. vandenBerg. The sdss skyserver: public access to the sloan digital sky server data. In Proceedings of the 2002 ACM SIGMOD international conference on Management of data, pages 570–581. ACM, 2002
- [4] J. H. Friedman. Multivariate adaptive regression splines. The annals of statistics, pages 1–67, 1991.
- [5] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1):100–108, 1979