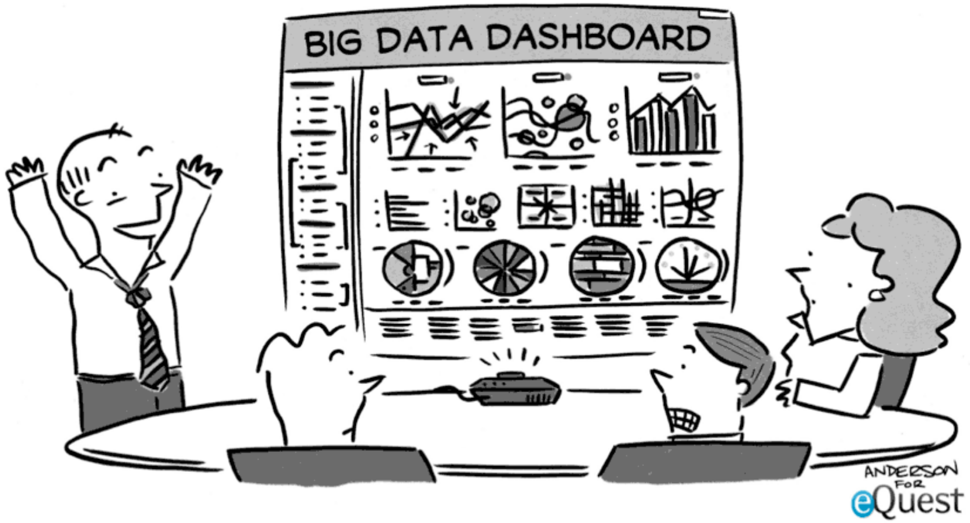
QuRVe: Query Refinement for View Recommendation in Visual Data Exploration

Humaira Ehsan (The University of Queensland)

Mohamed Sharaf (United Arab Emirates University)

Gianluca Demartini (The University of Queensland)

Visual Data Exploration



"After careful consideration of all 437 charts, graphs, and metrics, I've decided to throw up my hands, hit the liquor store, and get snockered. Who's with me?!"

Goal: Unlock hidden insights

Problem: **Manual** exploration is a time consuming tedious process

Solution: **Automatic** Visualization Recommendation Systems

Deviation-Based Visualization/View Recommendation

- Select a subset of data D_Q (Exploratory Query Q)

```
SELECT * FROM census WHERE edu>12;
```

Graduated high school

- Generate views based on all combinations of dimensions($\mathbb{A}$), measures ($\mathbb{M}$), aggregate functions ($\mathbb{F}$)

Example

$A_x \in \mathbb{A}$
$F_y \in \mathbb{F}$
$M_z \in \mathbb{M}$

Selected Data

Entire Database

Target View V_i Over D_Q

```
SELECT Ax, Fy(Mz) FROM census
WHERE edu>12 GROUP BY Ax;
```

Probability Distribution of Target View

Comparison View Over D

```
SELECT Ax, Fy(Mz) FROM census
GROUP BY Ax;
```

Probability Distribution of Comparison View

Compute Deviation

Recommend top-k views

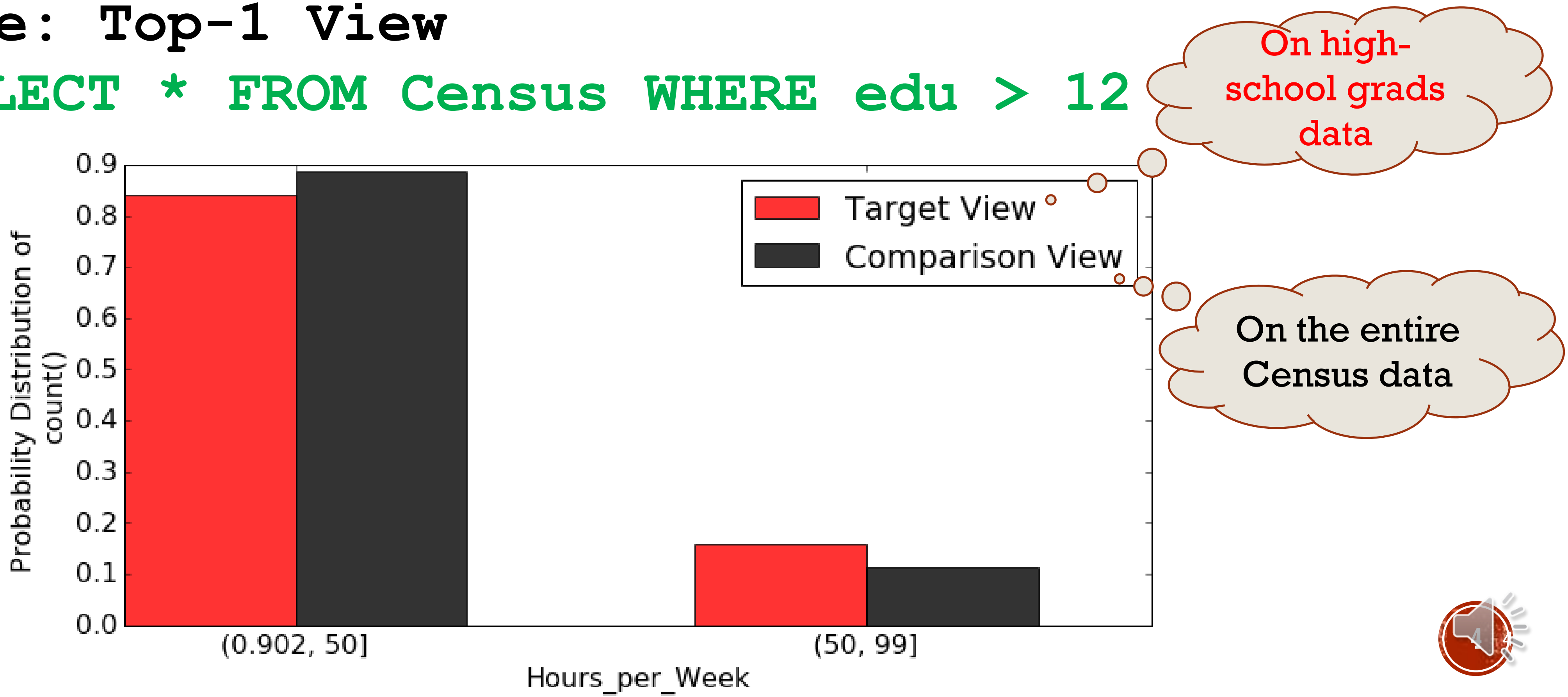
1. M. Vartak et. al., "*SeeDB: Automatically Generating Query Visualizations*", VLDB '14
2. Ehsan, H., et. Al., "*MuVE: efficient multi-objective view recommendation for visual data exploration*", ICDE '16
3. Ehsan, H., et. Al., "*Efficient recommendation of aggregate data visualizations*", TKDE '18

Aggregate View Recommendation

- Key Issue: **Assumption that the analyst is precise in defining an input exploratory query that reveals interesting insights!**

Example: Top-1 View

D_Q: `SELECT * FROM Census WHERE edu > 12`



Solution: Automatic Query Refinement

- Our proposed approach: Automatic **refinement** of exploratory queries to select data that reveals valuable **insights**

- Input query Q, specifies a conjunction of predicates, $P_1 \wedge P_2 \wedge P_3 \ldots\ldots \wedge P_p$
- A refined query Q_j is generated by modifying lower and/or upper limits for some P_i of Q.
- Example:

```
 Q: SELECT * FROM Census WHERE edu > 12
Q1: SELECT * FROM Census WHERE edu > 6
Q2: SELECT * FROM Census WHERE edu > 8
```


1. Mishra, C., Koudas, "Interactive query refinement", EDBT'2009
2. Telang, A., et al., "One size does not fit all: Toward user and query dependent ranking for web databases", TKDE'2012
3. Tran, Q.T., et al., "How to conquer why-not questions", SIGMOD'2010

Naive Query Refinement

Key Issues

- Similarity Oblivious
  - (dis)similarity to the initial exploratory query
- Statistically Insignificant Insights
  - false discoveries

Similarity Aware Refinement

Issue #1: similarity oblivious **Solution #1: similarity-aware query refinement**

- Distance between the refined query Q_j and the input query Q.
- Normalized similarity metric

$$\mathbf{s}(Q, Q_j) = 1/p \sum_{i=1}^{p} \frac{|l_i^{Q_j} - l_i^{Q}| + |u_i^{Q_j} - u_i^{Q}|}{2|U_i - L_i|}$$

- Possible refinements are exponential to the number of predicates in Q.
- Challenge:
  - Large number of refinements (Addressed in this paper)
  - Large number of views per refinement (Addressed in our previous work)

1. Albarrak, A., et al., "*Saqr: An efficient scheme for similarity-aware query refinement*", DASFAA'2014
2. Vartak, M., et al., "Refinement driven processing of aggregation constrained queries", EDBT'2016

Multi Objective Utility Function

Formally, our proposed hybrid **utility** function is:

$$\boxed{U(V_{i,Q_j}) = \alpha_s S(Q, Q_j) + \alpha_D D(V_{i,Q_j})}$$

$U(V_{i,Q_j})$: Utility of a view V_i from target query Q_j

$S(Q,Q_j)$: Similarity between the input query Q and the refined query Q_j

$D(V_{i,Q_j})$: Deviation value of the view V_i from query Q_j

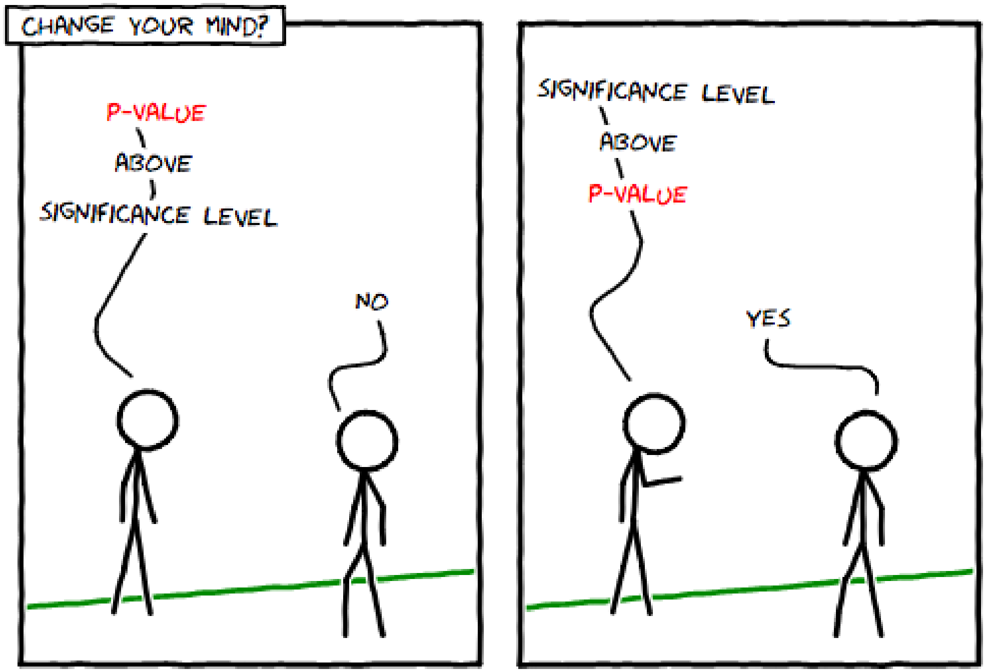
α_D, α_S, : Weight Parameters

Statistical Significance

Issue #2: statistical significance   Solution #2: **hypothesis testing**

- Recommended views may not have actual statistical significance.
- We employ Hypothesis testing to test the significance of the views.

- Formulate null and alternate hypothesis
- Calculate test statistics
- Compare p-value against significance level



1. Zhao, Z., et al., "*Controlling false discoveries during interactive data exploration*", SIGMOD' 2017
2. Chung, Y., et al., "*Towards quantifying uncertainty in data analysis & exploration*". IEEE Data Eng. Bull. '2018

Query Refinement for View Recommendation:

Formal Definition

Given a user-specified *query* Q

on a *database* D,

a multi- objective utility function $\mathtt{U}$,

a significance level α, statistical power $1-\beta$

and a positive integer k,

Find the k aggregate views V_{i,Q_j} *over* D_{Q_j}, which have the *highest utility* values from all of the refined queries Q_j

Such that $\mathtt{pvalue}(Q_j)<\alpha$ and $\mathtt{power}(Q_j)>(1-\beta)$,

The QuRVe Scheme

- Our multi-objective utility function is similar to Top- K preference query processing.
- However, our problem is different in two ways:
  - $D(V_{i,Qj})$ is not physically stored and they are computed on demand
  - Size of the view search space is prohibitively large and potentially infinite

Challenge: **Scaling** to a large number of possible views over a large number of refined queries

The QuRVe Scheme

- Predict maximum possible utility of unseen views, depends on upper bound on deviation $D_u = 1$.
- $U_{Unseen} = \alpha_S \times S(Q,Q_j) + (1 - \alpha_S) \times D_u$
- Access the views in decreasing order of similarity objective until the top k views are seen.
- In the example probe V1 for deviation calculation and update U_{seen} and U_{unseen} accordingly.

Probe for Deviation

Initializations

αS	αD	k	U_{Seen}	U_{Unseen}
0.6	0.4	1	0	1

	S(Vi)	D(Vi)	U(Vi)	U_{Seen}	U_{Unseen}
V1	1				
V2	0.75				
V3	0.75				
V4	0.5				
V5	0.5				
V6	0.5				
V7	0.25				
V8	0.25				

The QuRVe Scheme

- Predict maximum possible utility of unseen views, depends on upper bound on deviation D_u=1.
- $U_{Unseen} = \alpha_S \times S(Q,Q_j) + (1 - \alpha_S) \times D_u$
- Access the views in decreasing order of similarity objective until the top k views are seen.
- In the example probe V1 for deviation calculation and update U_{seen} and U_{unseen} accordingly.
- Stop when utility of seen views is higher than the utility of unseen views

Initializations

αS	αD	k	U_{Seen}	U_{Unseen}
0.6	0.4	1	0	1

	S(Vi)	D(Vi)	U(Vi)	U_{Seen}	U_{Unseen}
V1	1	0.1	0.64	0.64	0.85
V2	0.75				
V3	0.75				
V4	0.5				
V5	0.5				
V6	0.5				
V7	0.25				
V8	0.25				

Our **QuRVe** scheme uses **Early Termination** to minimize the number of processed views

The QuRVe Scheme

- Predict maximum possible utility of unseen views, depends on upper bound on deviation D_u=1.
- $U_{Unseen} = \alpha_S \times S(Q,Q_j) + (1 - \alpha_S) \times D_u$
- Access the views in decreasing order of similarity objective until the top k views are seen.
- In the example probe V1 for deviation calculation and update U_{seen} and U_{unseen} accordingly.
- Stop when utility of seen views is higher than the utility of unseen views

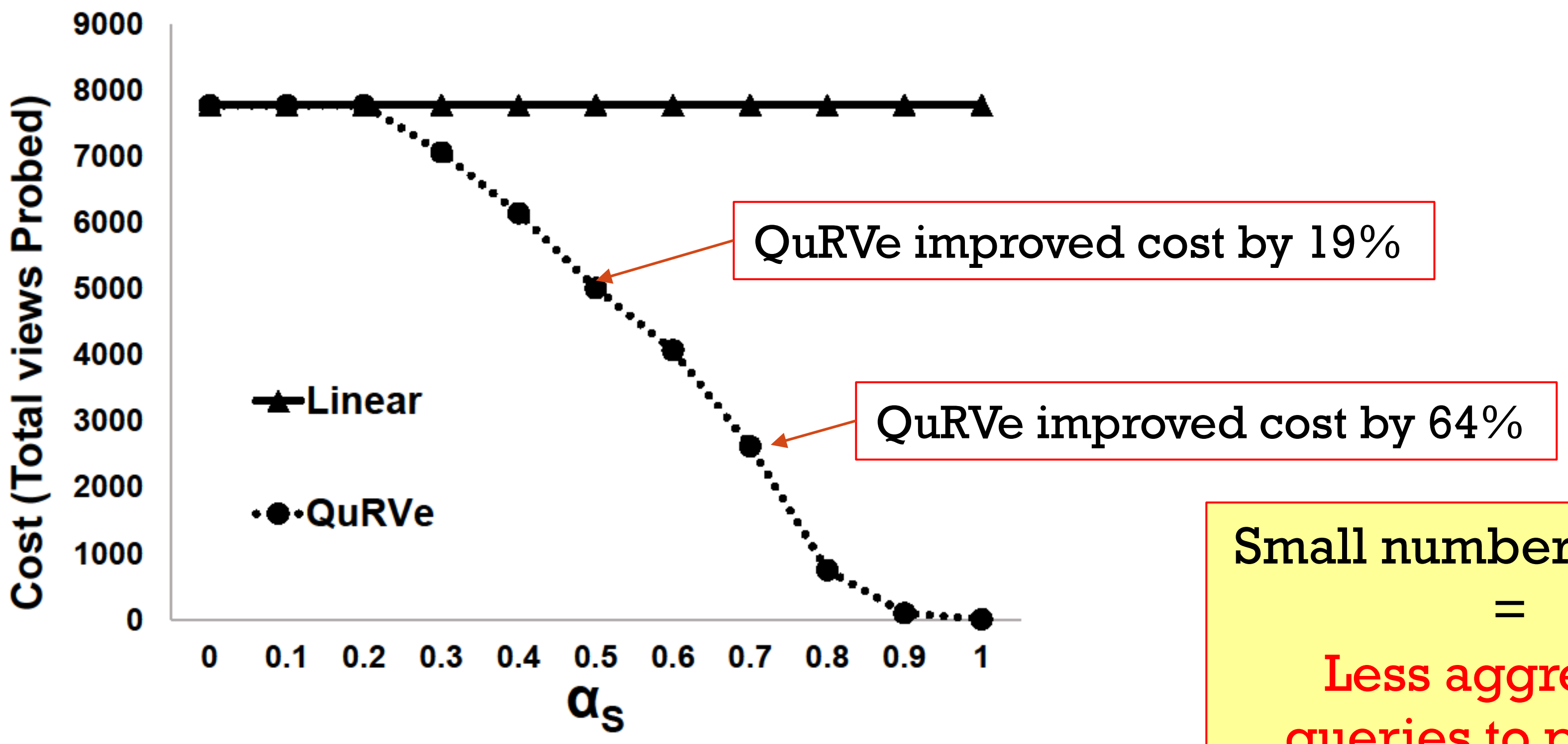
Initializations

αS	αD	k	U_{Seen}	U_{Unseen}
0.6	0.4	1	0	1

	S(Vi)	D(Vi)	U(Vi)	U_{Seen}	U_{Unseen}
V1	1	0.1	0.64	0.64	0.85
V2	0.75	0.1	0.49	0.64	0.85
V3	0.75	0.15	0.51	0.64	0.7
V4	0.5	0.4	0.46	0.64	0.7
V5	0.5	0.34	0.436	0.64	0.7
V6	0.5	0.7	0.58	0.64	0.55
V7	0.25				
V8	0.25				

Our **QuRVe** scheme uses **Early Termination** to minimize the number of processed views

Experiments



Small number of views
=
Less aggregate queries to process

Impact of α_s and α_D on cost

Please see more results in our paper!

Conclusions

- We formulated the problem of query refinement for view recommendation and proposed the QuRVe scheme.
- QuRVe efficiently navigates the refined queries search space to maximize utility and reduce the overall cost.

Thank You !!

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

جامعة الإمارات العربية المتحدة
United Arab Emirates University
UAEU

<u>Humaira Ehsan</u> (The University of Queensland)

Mohamed Sharaf (United Arab Emirates University)

Gianluca Demartini (The University of Queensland)