TECHNISCHE UNIVERSITÄT DRESDEN
Faculty of Computer Science
Database Technology Group

# Designing Random Sample Synopses with Outliers

Philipp Rösch ✉ philipp.roesch@tu-dresden.de
Rainer Gemulla ✉ gemulla@inf.tu-dresden.de
Wolfgang Lehner ✉ wolfgang.lehner@tu-dresden.de

### Setting
- large tables with outliers in aggregation columns
- approximate aggregation queries

### Goal
- memory-bounded outlier-aware sample synopses

### Challenge
- efficient computation
- low estimation error for all aggregation columns

## A Outlier-Aware Sample Synopses

| Order | Price |
|---|---|
| order#1 | 120 |
| order#2 | 200 |
| order#3 | 140 |
| order#4 | 180 |
| order#5 | 110 |
| order#6 | 1000 |
| order#7 | 80 |
| order#8 | 210 |
| order#9 | 100 |
| order#10 | 130 |

AVG(Price): **227**

Outlier

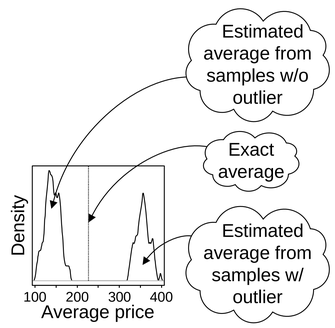
40% Sample

Example sample instance

| Order | Price |
|---|---|
| order#2 | 200 |
| order#4 | 180 |
| order#5 | 110 |
| order#10 | 130 |

AVG(Price): **155**

Estimate distribution (all instances)




40% Outlier-aware synopsis

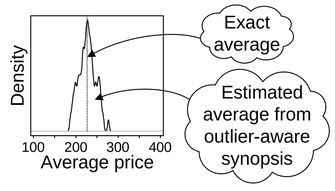
Example synopsis instance

| Order | Price |
|---|---|
| order#6 | 1000 |

| Order | Price |
|---|---|
| order#4 | 180 |
| order#5 | 110 |
| order#10 | 130 |

AVG(Price): **226**

Estimate distribution (all instances)




S. Chaudhuri, G. Das, M. Datar, and R. M. V. Narasayya. Overcoming Limitations of Sampling for Aggregation Queries. In *ICDE*, 2001.

## B Extension to Multiple Columns

### Problems

1. What is an outlier?
   - Outliers of one column may not be outliers in other columns
   - Multiple single-column instances are not optimal under space constraint

   **Solution**
   - Quantification of a synopsis' quality with **measures** based on the relative standard error (RSE) of the estimates
2. How to select outliers?
   - Number of possible outliers is prohibitively large

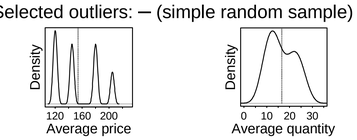
   **Solution**
   - Speed up computation by using **heuristics** and **greedy** proceeding

| Order | Price | Qty. |
|---|---|---|
| order#1 | 120 | 9 |
| order#2 | 125 | 4 |
| order#3 | 115 | 12 |
| order#4 | 119 | 54 |
| order#5 | 130 | 1 |
| order#6 | 110 | 5 |
| order#7 | 360 | 15 |
| order#8 | 123 | 21 |
| order#9 | 118 | 18 |
| order#10 | 220 | 27 |
| RSD: | 48.7% | 88.3% |

Outliers

Estimate distribution of synopses of size 40%

Selected outliers: – (simple random sample)

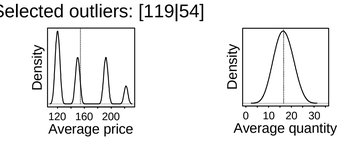



### Measures

***MAX*-Measure**
- effect: minimizes maximum RSE of the estimates
\+ most intuitive
– fails if column with maximum RSE has no outliers

Estimate distribution

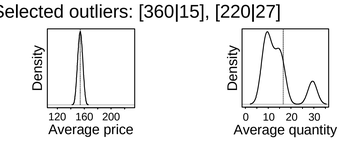
Selected outliers: [119|54]




***GEO*-Measure**
- effect: maximizes improvement in RSE compared to simple random sampling
\+ eager outlier selection
– a lot of effort into columns with very low RSE

Estimate distribution

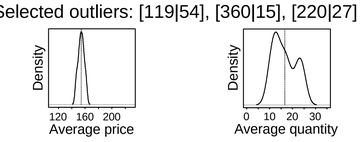
Selected outliers: [360|15], [220|27]



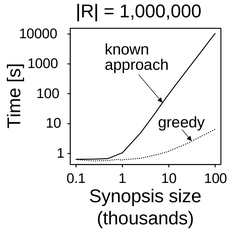

***AVG*-Measure**
- effect: minimizes the average RSE of the estimates
\+ prefers outliers from columns with high RSE
\+ shows best overall performance

Estimate distribution

Selected outliers: [119|54], [360|15], [220|27]




## C Impact on Single-Column Outlier-Aware Synopses




### Computational Contribution

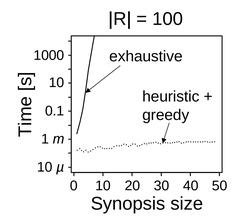
**0 Search Space**
known: 2*M* candidates from lower/upper value range with memory bound *M*

**1 Outlier Selection**
**Problem:** $O(M^2)$ combinations for *M* candidates
**Solution:** select outliers greedily from candidate set

### Computational Contribution

**1 Search Space**
**Problem:** huge number of possible outlier sets $\sum_{i=0}^{M} \binom{|R|}{i}$ with memory bound *M*
**Solution:** select *M* outlier candidates during the single scan of the data selection based on weights specific to measure of choice

**2 Outlier Selection**
**Problem:** $2^M$ combinations for *M* candidates
**Solution:** select outliers greedily from candidate set based on weights




**http://wwwdb.inf.tu-dresden.de/research/**