### Outlier-Aware Sample Synopses

**Example sample instance**

- Order	| Price
---|---
order#1 | 130
order#2 | 200
order#3 | 115
order#4 | 140
order#5 | 110
order#6 | 210
order#7 | 100
order#8 | 130
order#9 | 118
order#10| 120

**Example synopsis instance**

- Order	| Price
---|---
order#1 | 130
order#2 | 200
order#3 | 115
order#4 | 140
order#5 | 110
order#6 | 210
order#7 | 100
order#8 | 130
order#9 | 118
order#10| 120

**Average Price**: 227

- 40% Sample

- 40% Outlier-Aware Synopses

**Density of average from samples w/o outlier**

**Density of average from samples w/ outlier**

**Exact average**

**Estimated average from samples w/o outlier**

**Estimated average from samples w/ outlier**

**Solution:** Speed up computation by using heuristics and greedy proceeding.

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

**Computational Contribution**

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

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

### 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

**Measures**

- **MAX-Measure**
  - effect: minimizes maximum RSE of the estimates
  - + most intuitive
  - + prefers outliers from columns with high RSE

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

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

**Selection**

- solutions: [360|15], [220|27]

**Selected outliers**

- (simple random sample)

**Selected outliers: [119|54]**

**Selected outliers**

- (simple random sample)

**Selected outliers: [119|54]**

**Selected outliers: [119|54]**

### Designing Random Sample Synopses with Outliers

**Faculty of Computer Science**

- Wolfgang Lehner
- Philipp Rösch
- Rainer Gemulla

**ICDE 2001.**

- philipp.roesch@tu-dresden.de
- gemulla@inf.tu-dresden.de
- wolfgang.lehner@tu-dresden.de

**http://www.db.in.tu-dresden.de/research/**