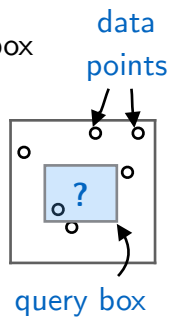
DigitHist: a Histogram-Based Data Summary with Tight Error Bounds

Michael Shekelyan, Anton Dignös, Johann Gamper (Free University of Bozen-Bolzano)

unibz

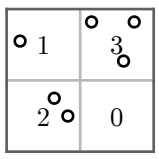
Problem Setting

- problem:
  - take set of multi-dimensional points in euclidean space
  - create summary to estimate no. of points in any query box
- motivation:
  - selectivity estimation (guiding query optimizers)
  - approximate query answering (OLAP/DSS)
- challenges:
  - high precision, low costs and good scalability
  - tight (deterministic) error bounds



Prelim.: Multi-dim. Histograms

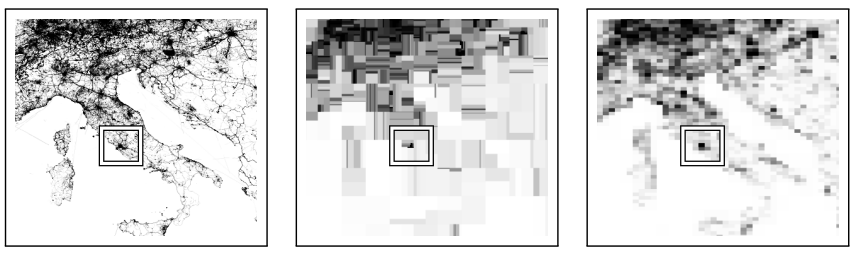
- multi-dim. histograms divide space into regions (buckets)
- count number of points in each bucket
- buckets contained in query region give lower bound
- buckets intersected by query region give upper bound



DigitHist Data Summary: Overview

- small no. of multi-dimensional histograms (digit hists.) along regular grids each accompanied by hi-res projections on individual axes (marginal hists.)
- single-pass construction; linear in summary and data size, and no. of dims
- efficient representation of histograms with mostly empty buckets
- individual error bounds for each query box

example

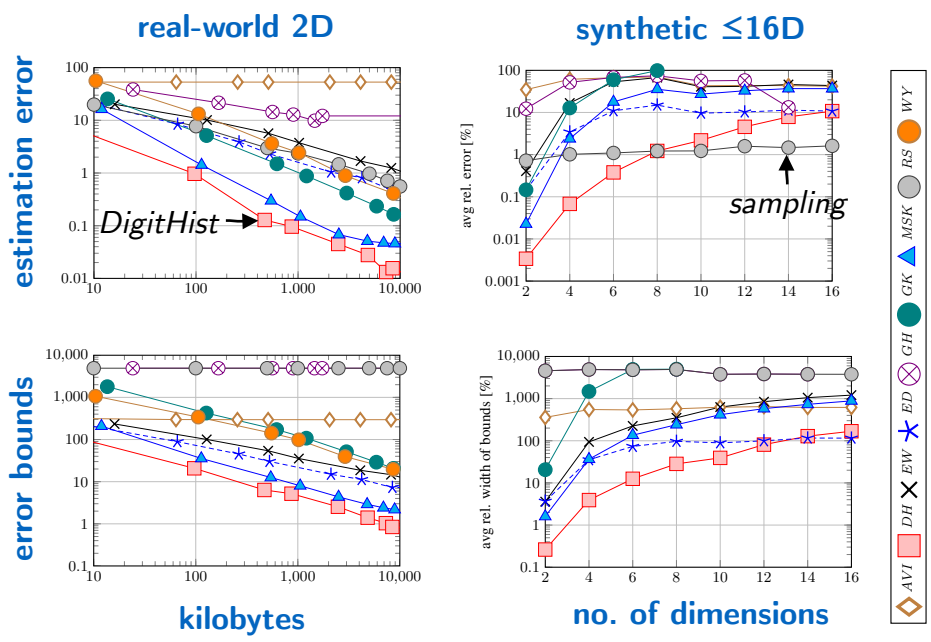


(a) Data (46.4GB) **(b)** MinSkew (113kB) **(c)** DigitHist (97kB)

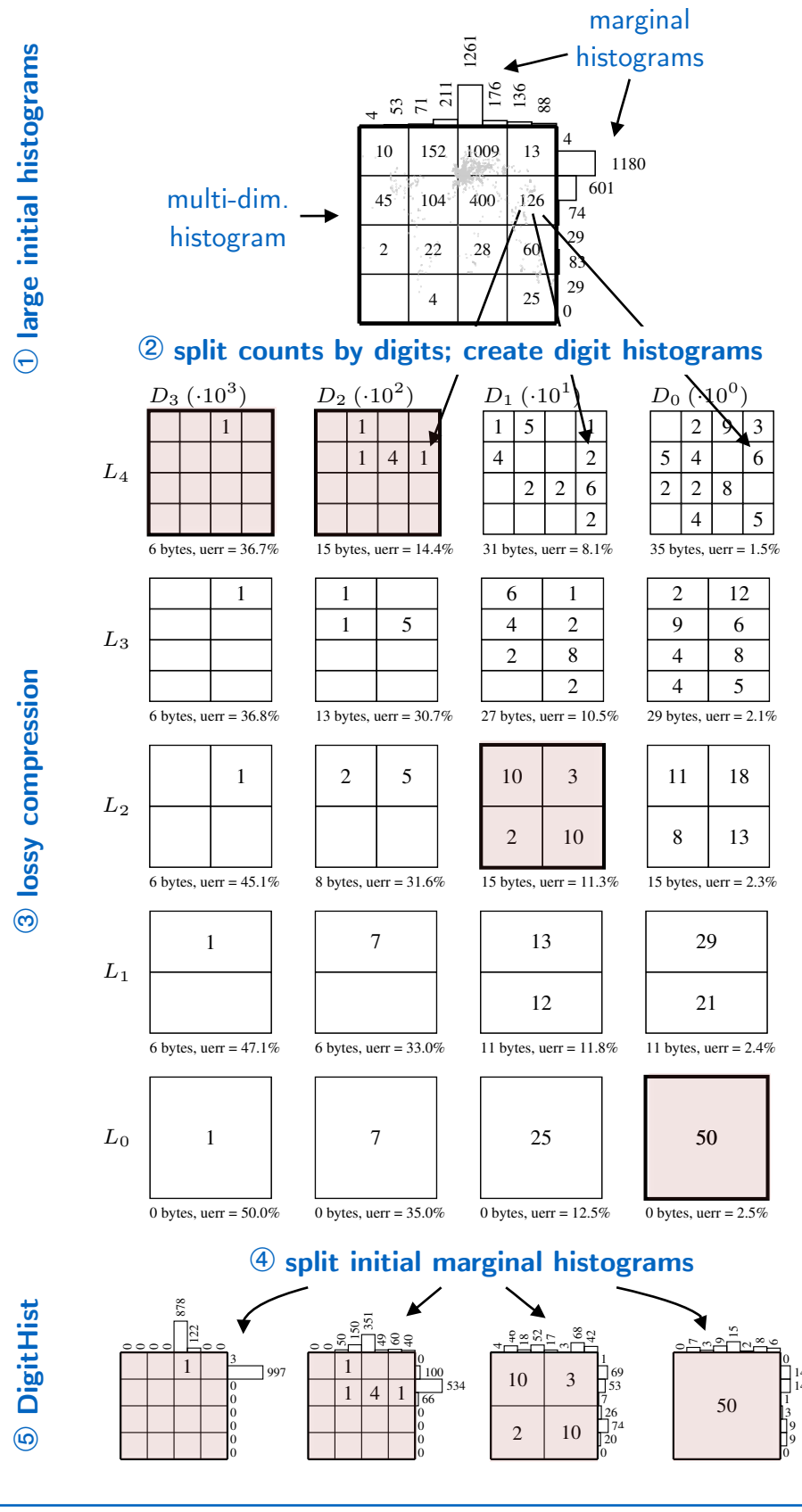
	MinSkew [2]	DigitHist [1]
Estimated selectivity	0.177%	0.215%
Relative error	17.3%	0.5%
Bounds	[0.102%, 0.24%]	[0.195%, 0.251%]
Width of bounds	0.138%	0.056%

Experimental Results

- best estimation precision for up to six data dimensions
- best scalability with dimensionality (except sampling)
- tightest error bounds (100% confidence intervals)



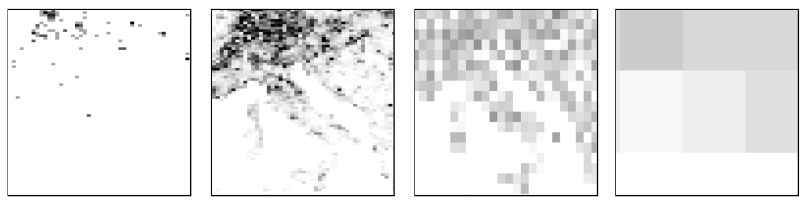
DigitHist Construction



Key Ideas of DigitHist

1. digit histogram lossy compression

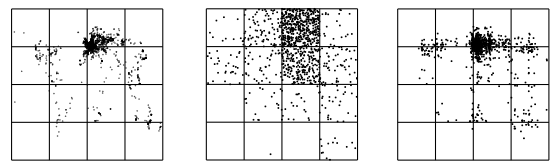
- summarizes regions with many points using higher resolution



89% of points at hi-res | 10.5% mid-res | 0.4% low-res

2. intra-bucket spread estimation using marginal histograms

- use 1d information for spread estimation inside buckets



original data | uniformity | marginal hists.

3. measure hist. precision with u-error [1] metric

- minimize error bounds instead of skew inside buckets
- assume uniformly distributed queries (for sake of simplicity)

VLDB 2017 Munich · Germany
43rd International Conference on Very Large Data Bases

[1] Shekelyan et. al., VLDB'17;
[2] Acharya et al., SIGMOD'99