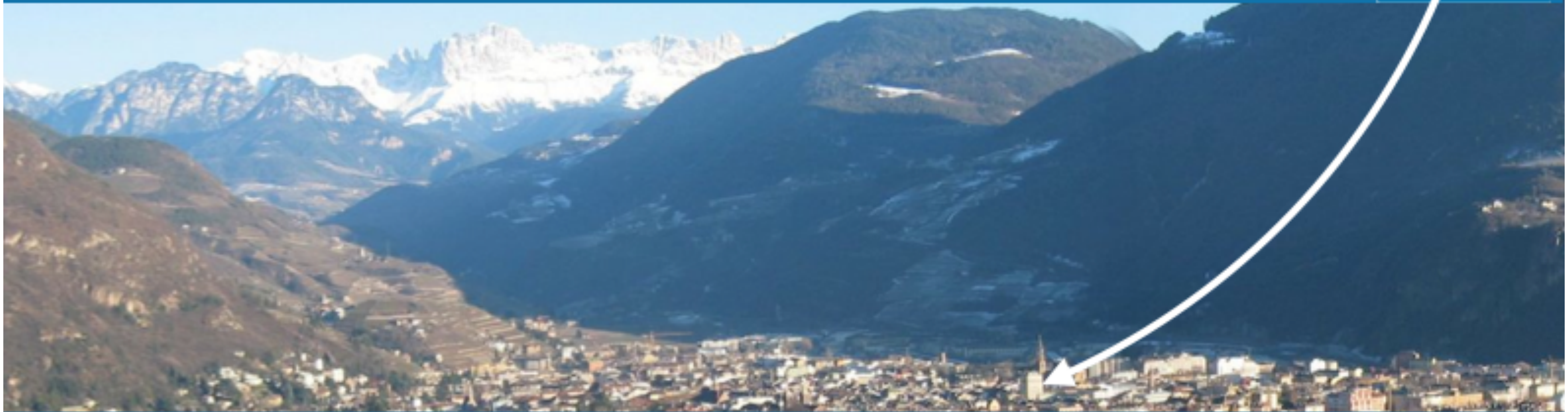
From the other side of the alps ...

unibz



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

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

VLDB 2017 Munich · Germany

43rd International Conference on Very Large Data Bases

Motivation for Data Summaries

Data Management

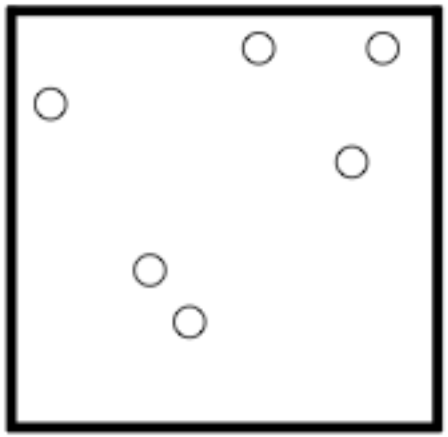
- selectivity estimation to help guide query optimization

Data Analysis

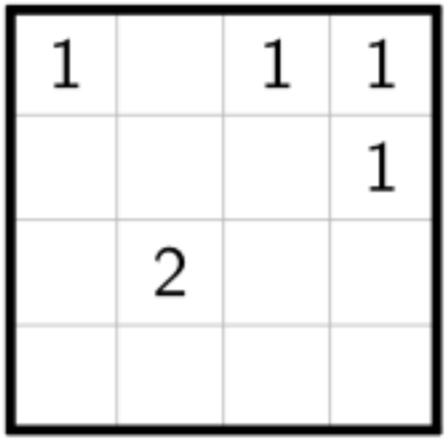
- approximate query answering in DSS/OLAP systems

Data Summaries for Range Queries

data points



multi-dimensional histogram



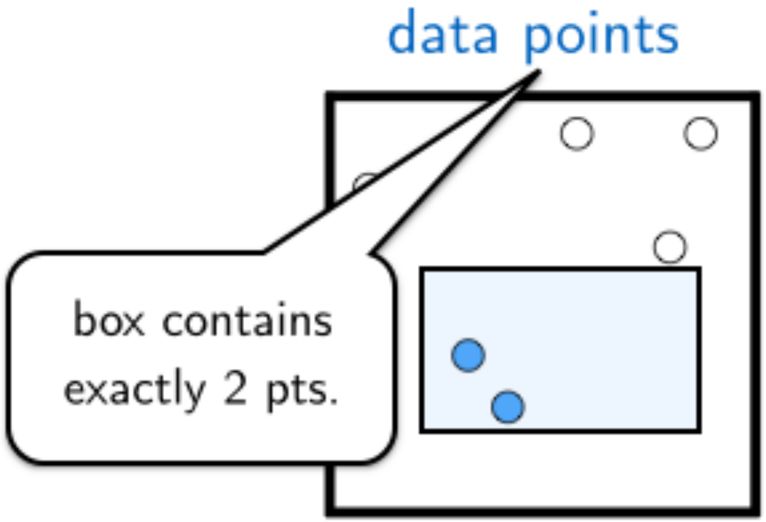
data

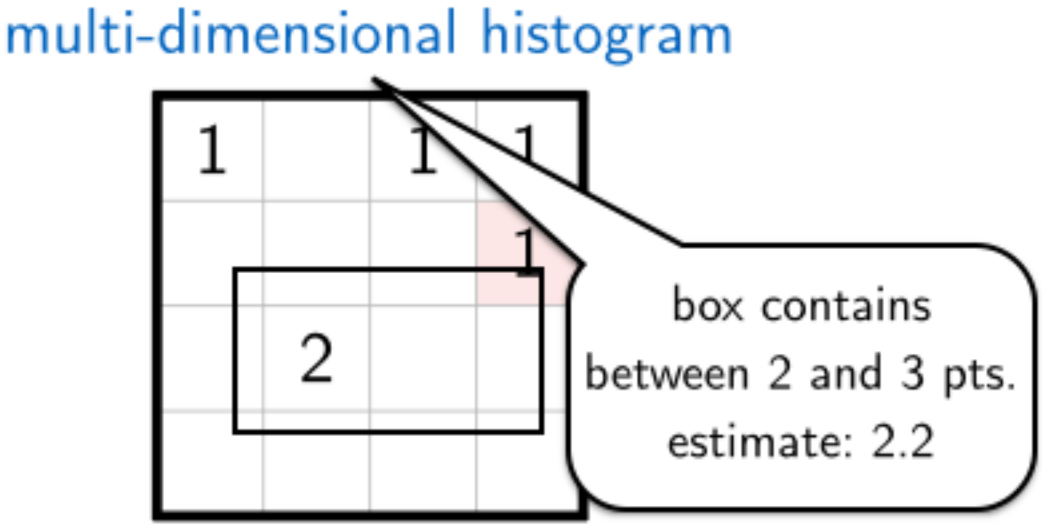
- multi-dimensional points in euclidean space

summary

- data structure approximating no. of points in any query box

Data Summaries for Range Queries





data

- multi-dimensional points in euclidean space

summary

- data structure approximating no. of points in any query box

Related Work

unibz

	sampling	multi-dim. histograms	ε-approximation	DigitHist (proposed)
references	TOMS'85	SIGMOD'99,'00	SCG'04, SODA'13,	VLDB'17
basic idea	take random subset	split space into buckets	quantile-based or sampling	count along regular grids
for more data?	very good	ok	ok	good
for few dims.?	ok	good	good	good
for more dims.?	very good	bad	bad	ok
error bounds?	only confidence intervals	individual for each query	same for all datasets and queries	individual for each query

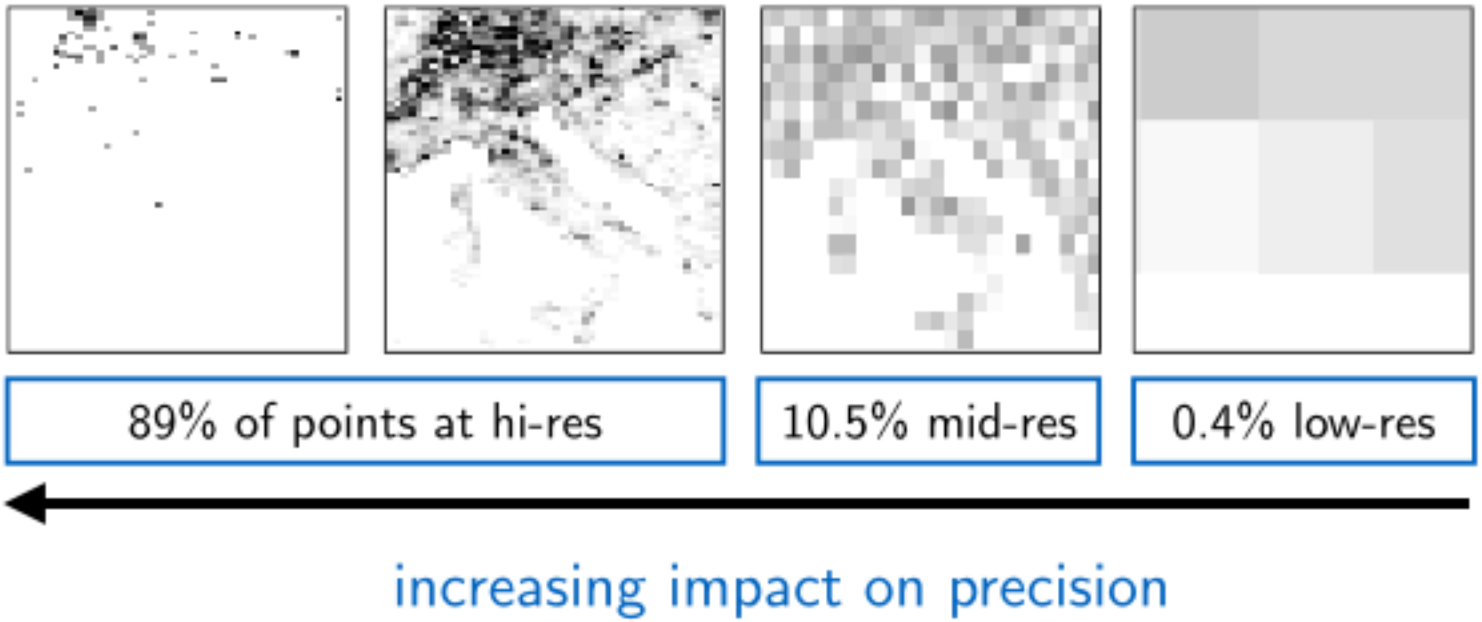
other notable approaches:

- Wavelets, DCT, Kernel Density Estimation, Dyadic Decomposition (e.g., Sketches)

DigitHist Key Idea

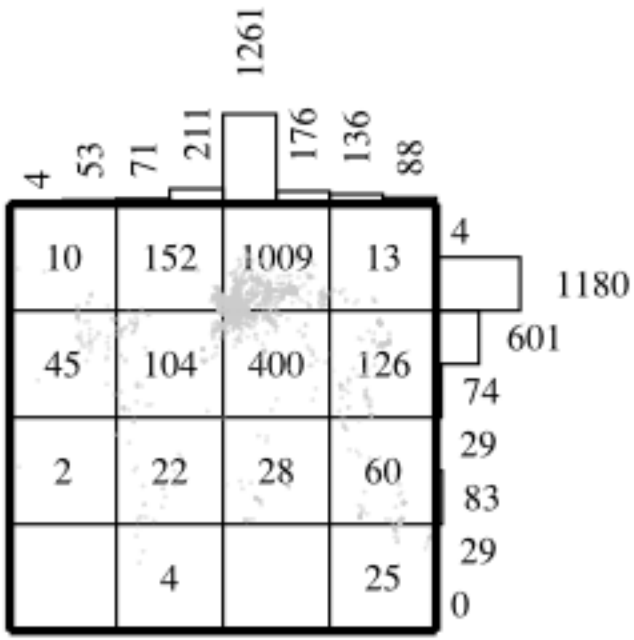
digit histograms

- summarize majority of points at higher resolution
  - e.g. 89.1% at 1024x1024 (99.6% of points at 512x256 or higher)
- summarize remaining points at lower resolution



initial histograms

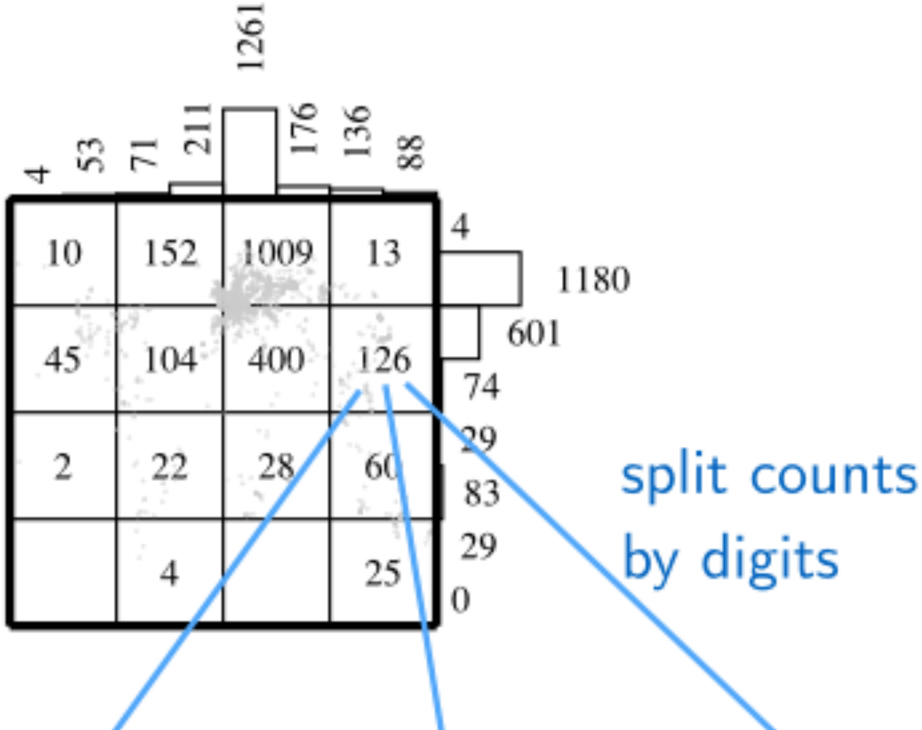
- create in single scan
- materialize only non-empty buckets
- sufficiently precise
- too large
- (and slow to query)

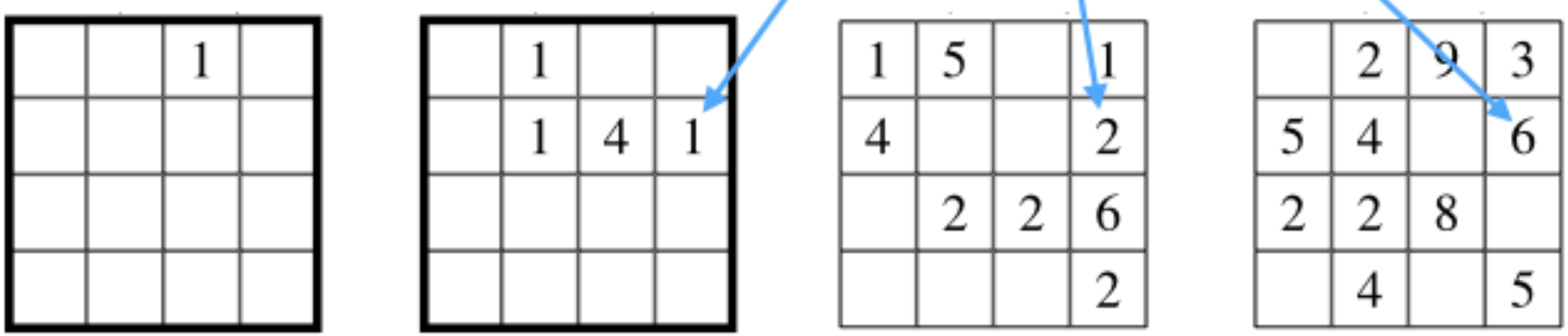


DigitHist: Digit Histograms

digit histograms

- split counts by digits
- treat digit histograms differently





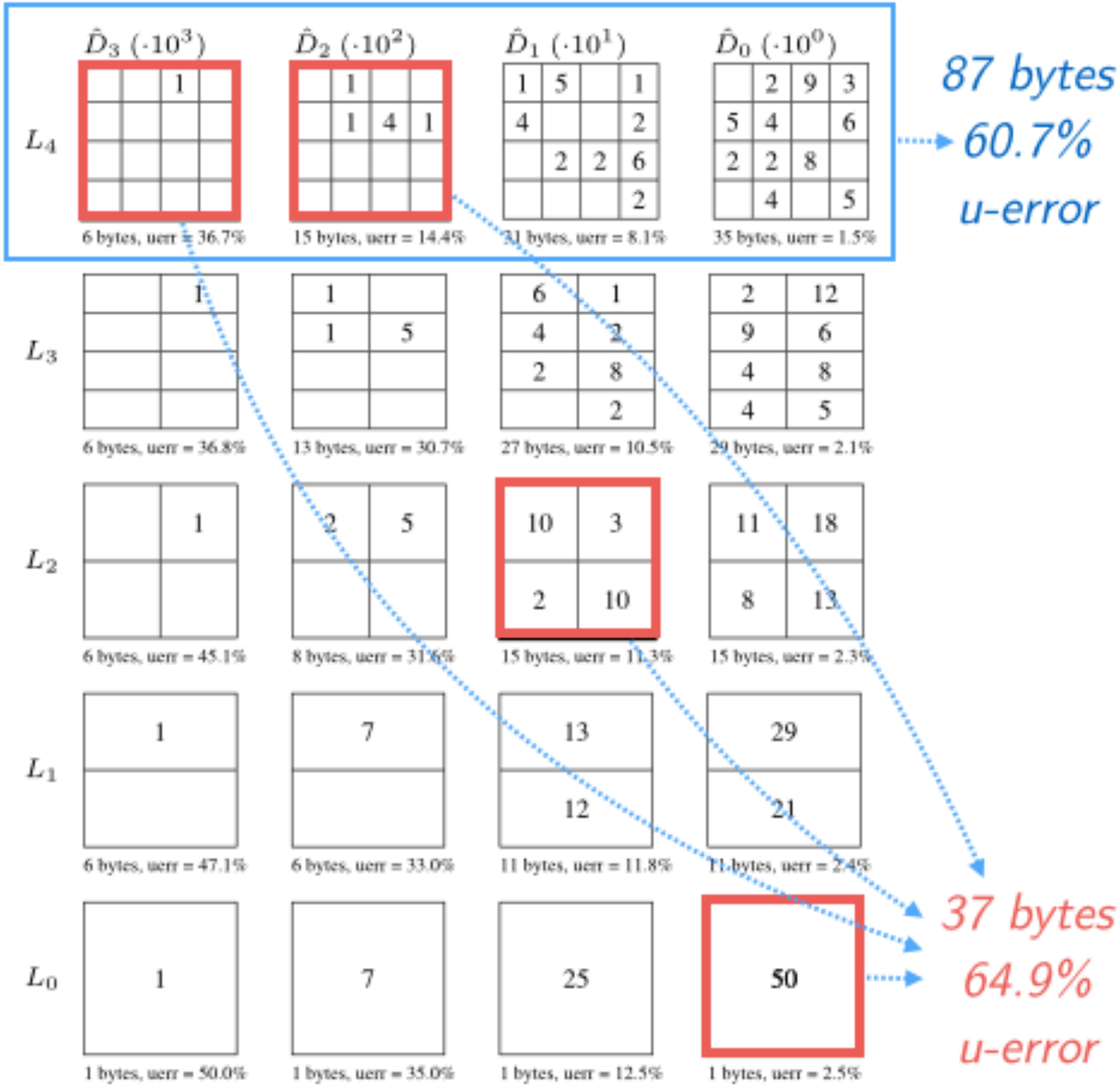
DigitHist: Lossy compression / u-error

lossy compression

- targeted size
- pick resolutions minimizing u-error

u-error

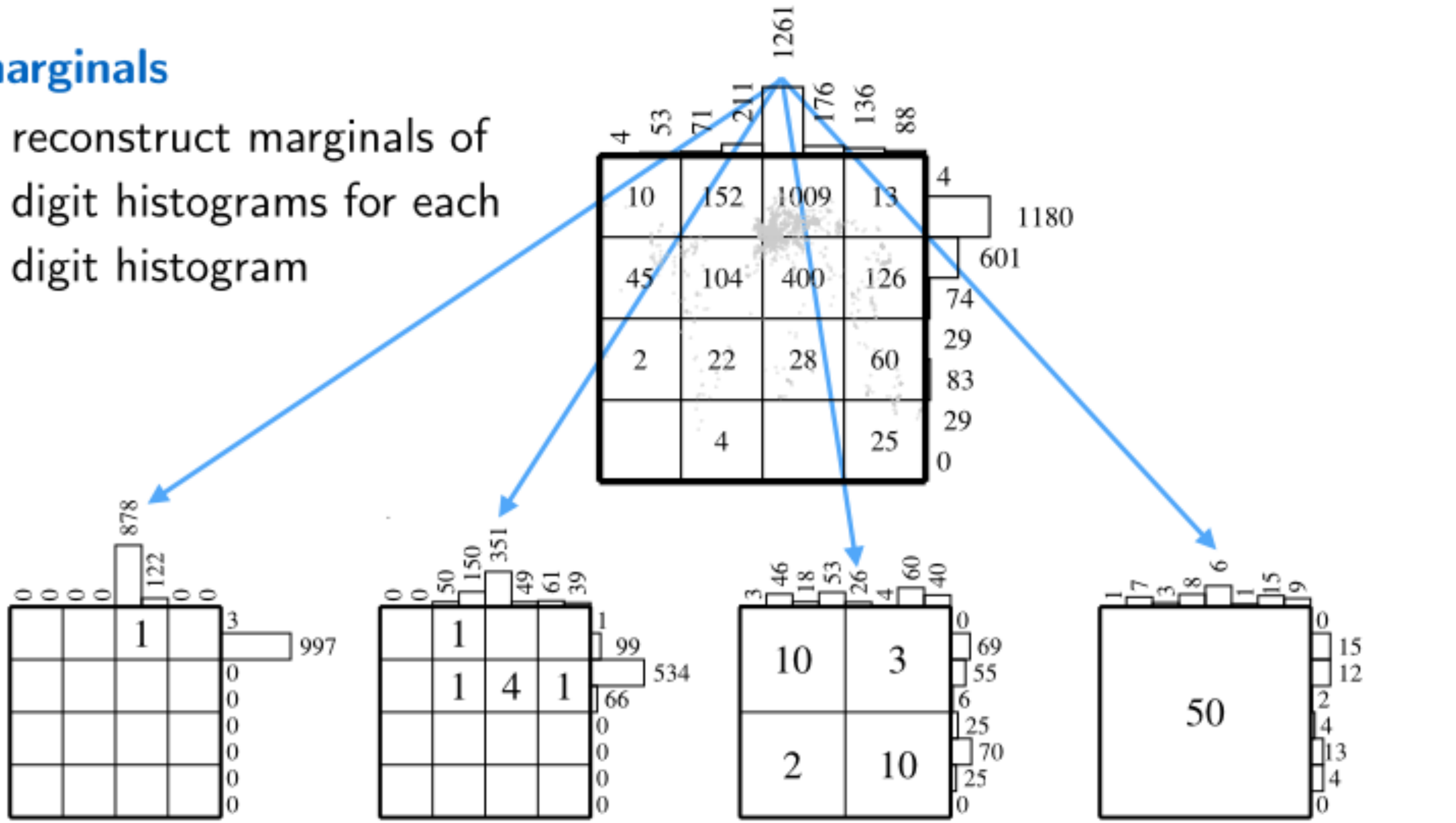
- expected width of bounds
- random query box
- uniform in location and size



DigitHist: Accompany with Marginals

marginals

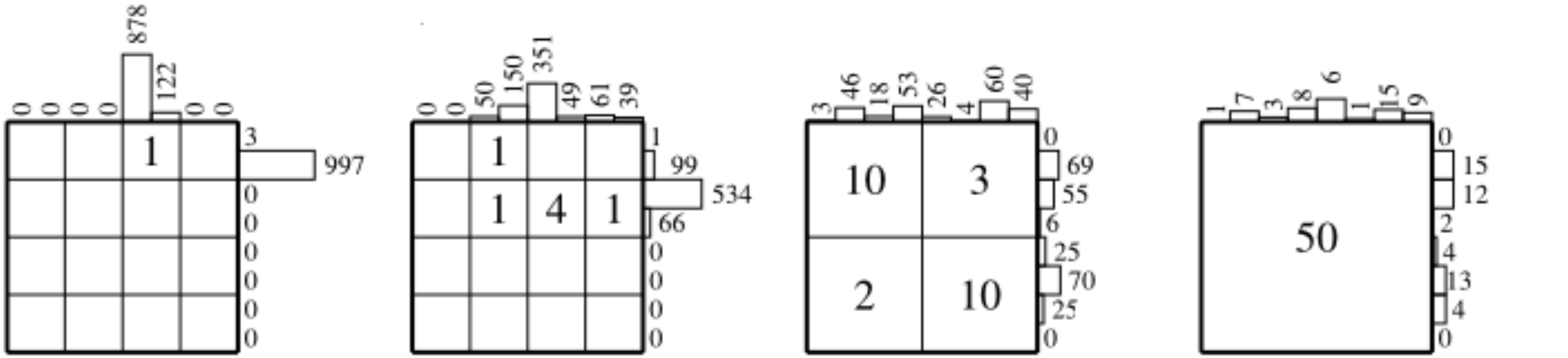
- reconstruct marginals of digit histograms for each digit histogram



DigitHist

final summary

- four multi-dimensional histograms
- accompanied by hi-res marginal histogram in each dimension



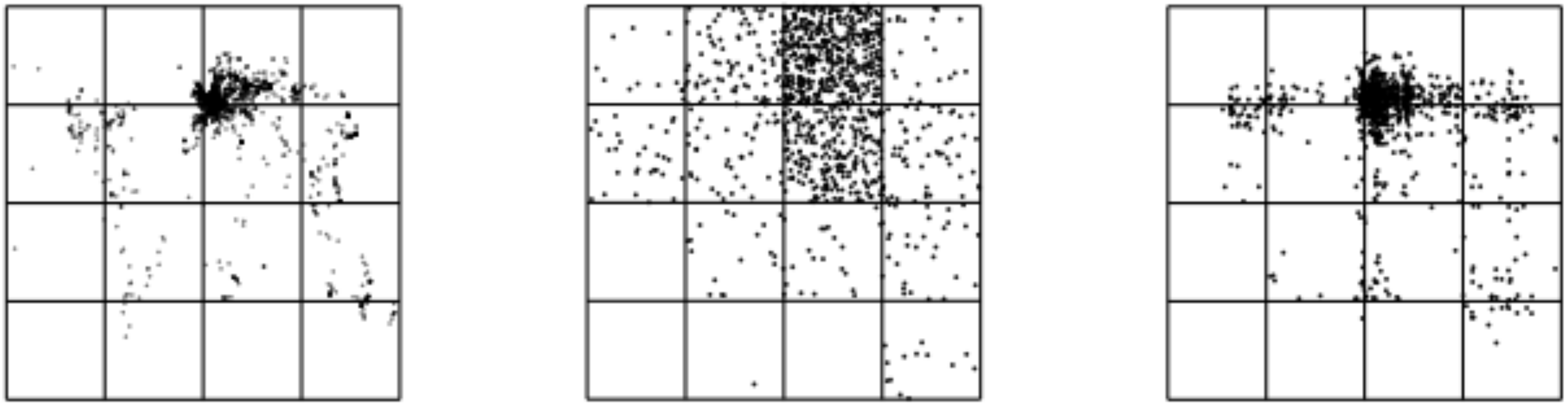
Intra-Bucket Assumption

problem

- location of points inside buckets not stored and needs to be estimated
- assuming uniformity inside buckets assumes uniformity in marginals

DigitHist intra-bucket spread

- spread points using marginal histograms and assuming independence
- no significant query time overhead



data points   multidim. hists.   DigitHist

Properties of DigitHist

unibz

efficient construction

- data can be treated as stream (single-pass)
- user-controlled memory usage (2GB in our experiments)
- construction time linear in data size, data dim. and summary size

convenient storage

- stored and queried as byte stream

fast, light-weight querying

- time linear in summary size and no significant memory usage

precision

- query-individual estimates and error bounds

updatability

- reconstruct from updated regular grid histogram in secondary storage

Experimental Setup

unibz

datasets

- OpenStreetMap 46.4 GB(2D spatial data)
- HIGGS 616 MB (7D scientific data)
- Zipf/Gauss ≤1.2GB (2-16D synthetic data)

methodology

- construct all tested data summaries.
- randomly create ranges with low selectivity (1-5%).
- compute avg. estimates and bounds of tested summaries.

precision measures

- relative error = (estimate - correct) / correct
- relative width of bounds = (width of bounds) / correct

Construction Time

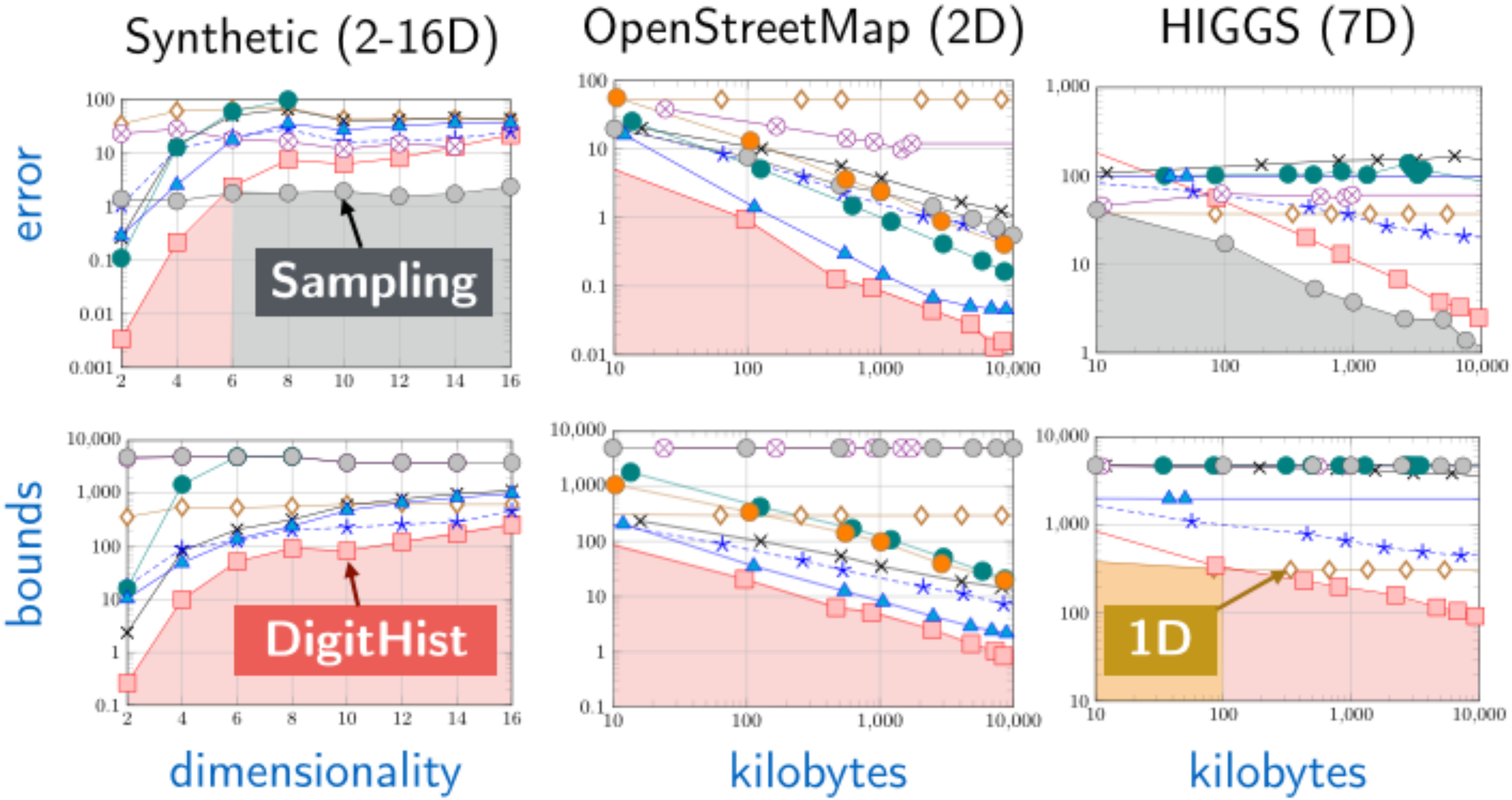
unibz

- DigitHist can be constructed in a few minutes
- simple summaries faster to construct (but less precise)
- quantile-based summaries take hours to construct

datasets	OSM (2D - 46.4 GB)		HIGGS (7D - 616 MB)	
summary size	100kB	1MB	100kB	1MB
DigitHist	16 mins	17 mins	35 secs	3 mins
1d Histograms	3 mins	3 mins	2 secs	3 secs
Equi-Width	5 mins	5 mins	2 secs	3 secs
Equi-Depth	2.2 hours	2.5 hours	29 secs	34 secs
Wei-Yi	2.5 hours	3.1 hours		
Cross GK	1 hour	2.1 hours	20 mins	40 mins

Precision: Errors and Bounds

- <6D: DigitHist smallest error and bounds
- ≥6D: Sampling smallest error and DigitHist tightest bounds



Summary

DigitHist

- aggregate data along regular grids
- split counts by digits
- aggregate denser data regions at higher resolution
- integrate high-res one-dimensional histograms

u-error

- measures precision to guides lossy compression of DigitHist
- takes histogram and computes its expected width of bounds
- assumes uniform distribution of queries

experimental results

- $<$6D: DigitHist smallest error and bounds
- $\geq$6D: Sampling smallest error and DigitHist smallest bounds

Outlook

DigitHist

- optimize summary for known workload
- deal with mix of categorical, discrete and real attributes
- try to improve performance for more dimensions

beyond DigitHist

- lower-dimensional summaries with ultra-fast query times and theoretical guarantees

Thank you for attention!