# OmniSketch: Efficient Multi-Dimensional Stream Analytics with Arbitrary Predicates

Wieger R. Punter [1], Odysseas Papapetrou [1], and Minos Garofalakis [2]

[1] Eindhoven University of Technology, The Netherlands
{w.r.punter,o.papapetrou}@tue.nl
[2] Athena Research Center &
Technical University of Crete, Greece
minos@athenarc.gr

A key need in different disciplines is to perform analytics over fast-paced data streams, similar in nature to the traditional OLAP analytics in relational databases – i.e., with filters and aggregates. Storing unbounded streams, however, is not a realistic, or desired approach due to the high storage requirements, and the delays introduced when storing massive data. Accordingly, many synopses/sketches have been proposed that can summarize the stream in small memory (usually sufficiently small to be stored in RAM), such that aggregate queries can be efficiently approximated, without storing the full stream. However, past synopses predominantly focus on summarizing single-attribute streams, and cannot handle filters and constraints on arbitrary subsets of multiple attributes efficiently.
Our prior work introduced OmniSketch [3, 4], the first sketch that scales to fast-paced and complex data streams (with many attributes), and supports count aggregates with filters on multiple attributes, dynamically chosen at query time. This work advances on the previous work by generalizing the streaming model to cover not only inserts, but also bounded deletes. OmniSketch offers probabilistic guarantees, a favorable space-accuracy tradeoff, and a worst-case logarithmic complexity for updating and for query execution. We demonstrate experimentally with both real and synthetic data that OmniSketch outperforms Hydra [2] and adaptive Sampling and Hold [1], the state-of-the-art competitors. and can approximate complex ad-hoc queries within the configured accuracy guarantees, with small memory requirements.

## References

[1] Cohen, E. and Cormode, G. and Duffield, N. G. (2012). Don't let the negatives bring you down: sampling from streams of signed updates. *SIGMETRICS Perform. Eval. Rev.*, 40(1), 343–354.

[2] Manousis, A. and Cheng, Z. and Basat, R.B. and Liu, Z. and Sekar, V. (2022) Enabling Efficient and General Subpopulation Analytics in Multidimensional Data Streams. *Proc. VLDB Endow.* 15, 11 (jul 2022), 3249-3262.

[3] Punter, W. R. and Papapetrou, O. and Garofalakis, M. (2023). OmniSketch: Efficient Multi-Dimensional High-Velocity Stream Analytics with Arbitrary Predicates. *PVLDB*, 17(3), 319–331.

[4] Punter, W. R. and Papapetrou, O. and Garofalakis, M. (2025). OmniSketch: Streaming Data Analytics with Arbitrary Predicates. *SIGMOD Record*, 54(1), 28–35.