# Towards Data Forgetting at Scale

Ramon Rico

Data Intensive Systems, Utrecht University, The Netherlands
r.ricocuevas@uu.nl

Our ability to collect data is rapidly surpassing our ability to store it. As a result, organizations are faced with difficult decisions about which data to retain and which to dispose of. Data forgetting [Rico et al.(2025a)], frames this reduction task as a subset selection exercise. Given a relational dataset $D$, a query log $Q$, and a budget $B$, the goal is to find a subset $D^* \subseteq D$ with at most $B$ tuples such that it is still possible to compute, based solely on $D^*$, approximate answers to the expected query workload. Existing data forgetting routines have substantial limitations. They either offer strong theoretical guarantees but lack scalability due to function evaluation (submodular-based), or achieve scalability by avoiding function evaluation but lack theoretical guarantees (amnesia-based). To bridge the gap between the limitations of submodular and amnesia based methods, we introduce `IndepDF` and `DepDF`: two data forgetting routines that offer scalability by avoiding function evaluation while maintaining strong theoretical guarantees; in essence, combining the best of both worlds [Rico et al.(2025b)].

## References

[Rico et al.(2025a)] Ramon Rico, Arno Siebes, and Yannis Velegrakis. 2025a. New Trends in Data Forgetting for Sustainable Data Management. *Proceedings of the VLDB Endowment* 18, 12 (2025), 5472–5476.

[Rico et al.(2025b)] Ramon Rico, Arno Siebes, and Yannis Velegrakis. 2025b. Stochastic Submodular Data Forgetting. *Proc. ACM Manag. Data (SIGMOD)* 3, 6 (2025), Article 365.