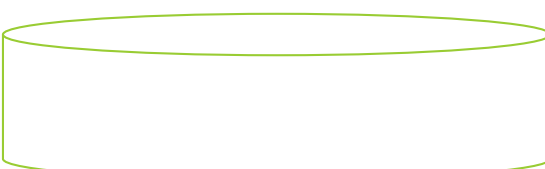
Membership Inference

22

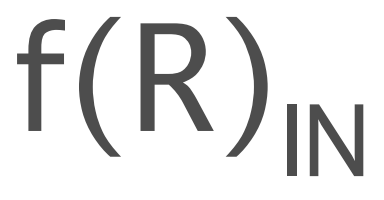
t



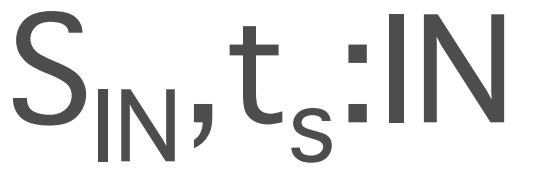

RA1

Split/Subsample

f(R)OUT


f(R)IN


S_{OUT}, t_s:OUT

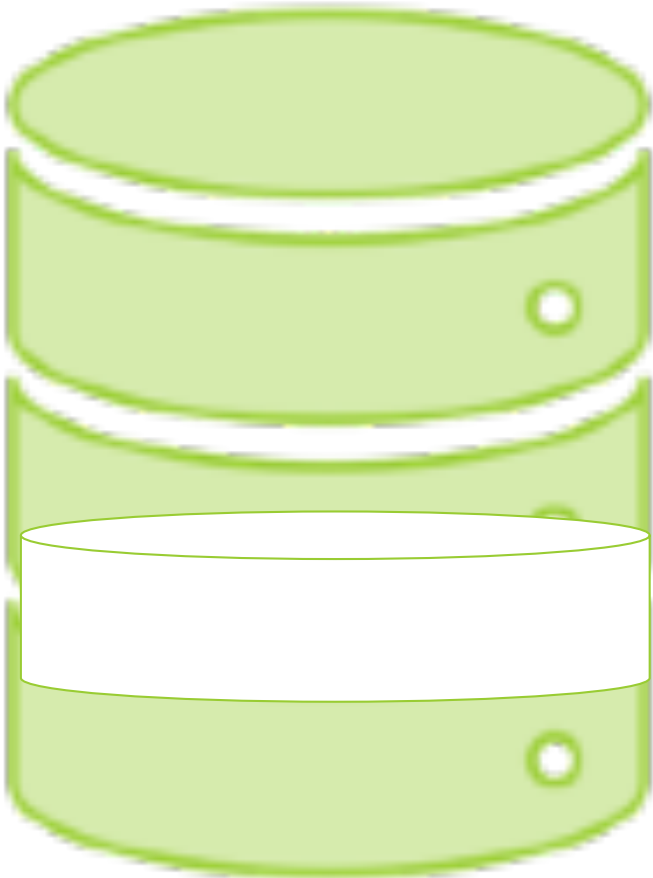

S_{IN}, t_s:IN

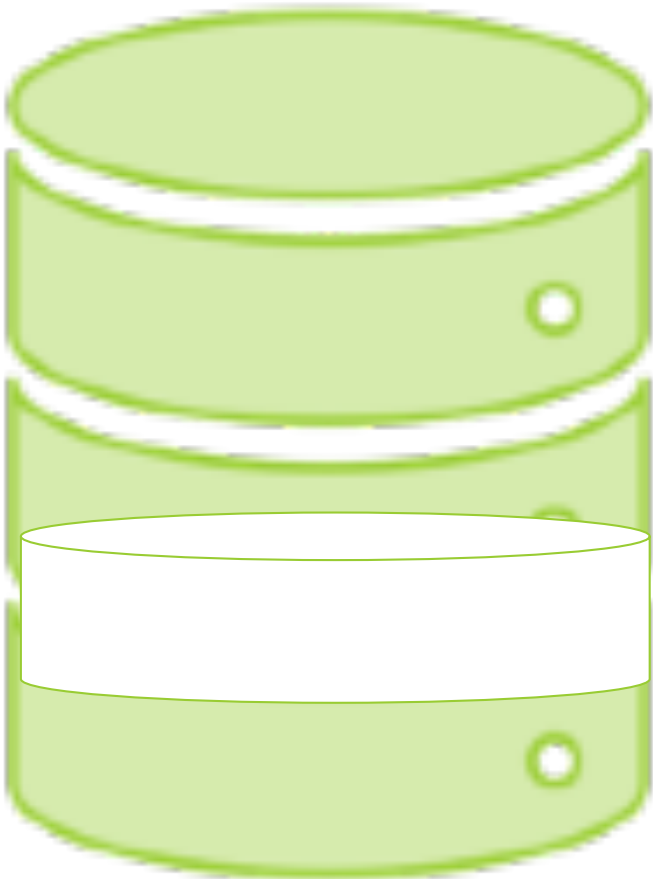
Train: $S \rightarrow t_s$

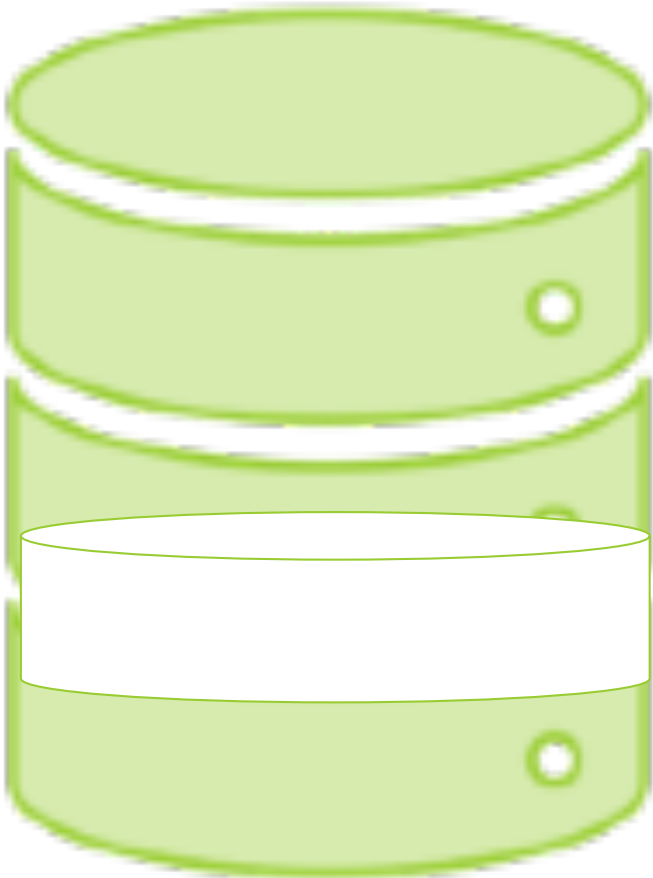

MIA
t

Trained classifier

$$\hat{t}_s = \underset{t_s}{\operatorname{argmax}} P[t_s \mid S]$$



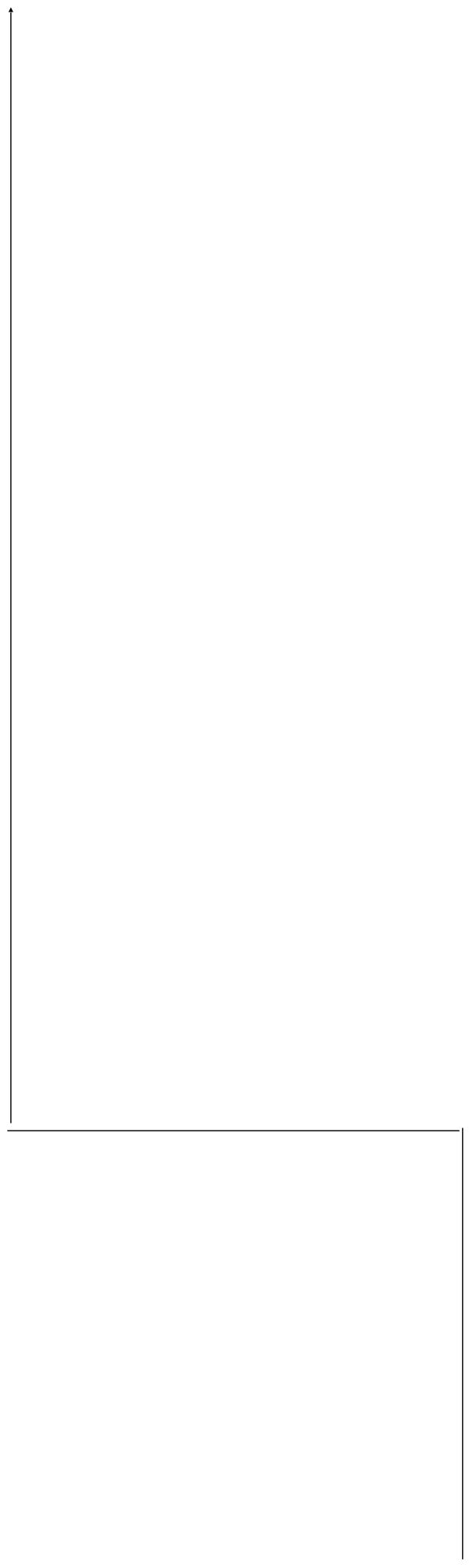


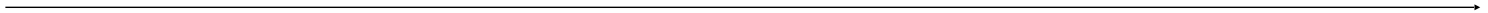


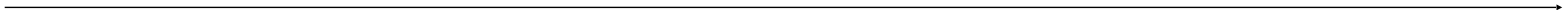
Add


RA2


RAn



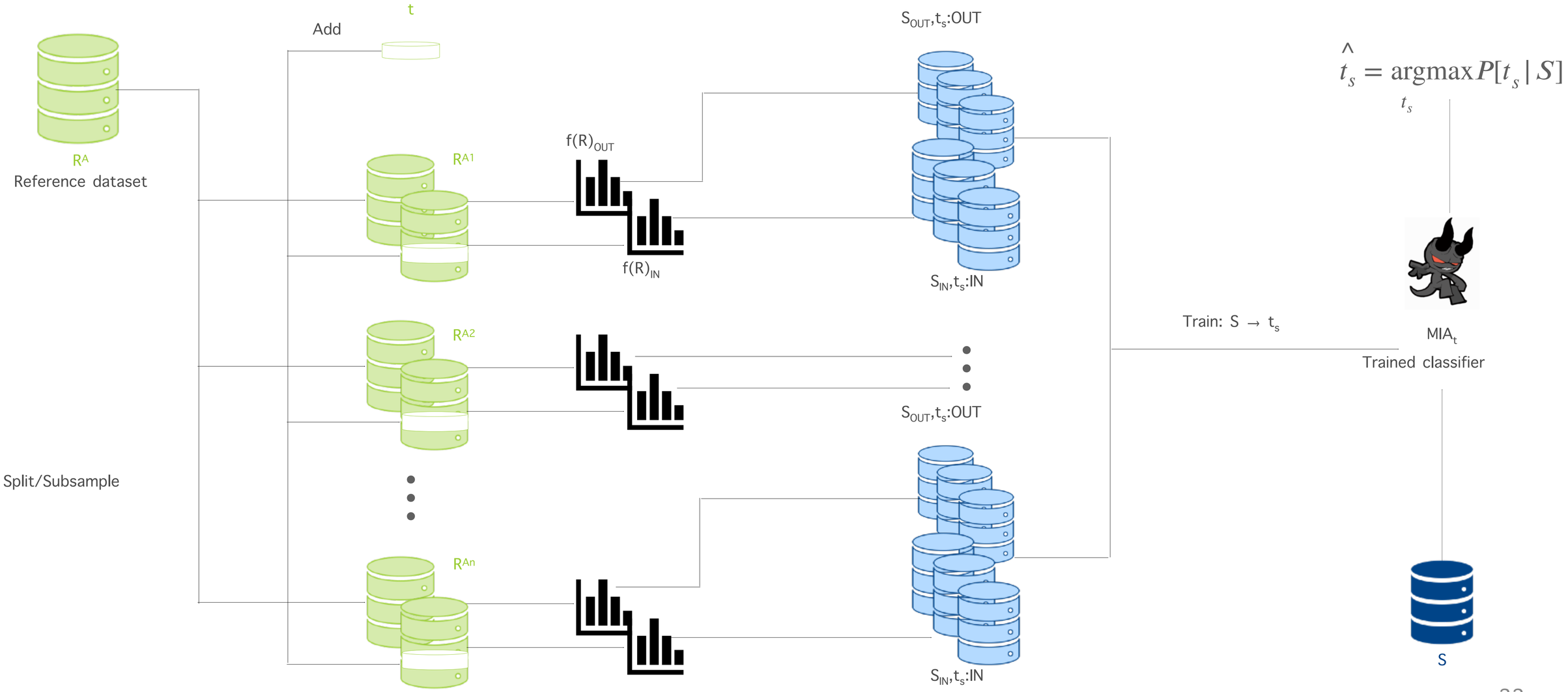




R^A

Reference dataset

Membership Inference



Privacy Gain

- Under the assumption equal prior $P[t_s] = 0.5$ and perfect linkage in case of raw dataset $P[MIA_t(R) = t_s] = 1$

$$PG_t(S, R) \triangleq \frac{1 - P[MIA_t(S) = t_s]}{2}$$

$P[MIA_t(S) = t_s] = 1$

Publishing S is equivalent to publishing R

$P[MIA_t(S) = t_s] = 0$

Publishing S reduces the adversary's chance of success

$PG_t = 0.25$

$PG_t = 0$ $PG_t = 0.5$

$P[MIA_t(S) = t_s] = 0.5$

Publishing S gives the adversary no advantage over random guessing