Unlearning: A Definition

Training:

D

Algorithm A

$A(D)$

ϵ-Unlearning/Deletion:

D z

aux

Retraining

Unlearning Algo R_A

We say R_A is a ϵ-deletion algorithm for A if:

$$A\left(D\backslash\{z\}\right) \overset{d}{\underset{\epsilon}{\approx}} R_A\left(D, z, \text{aux}\right)$$
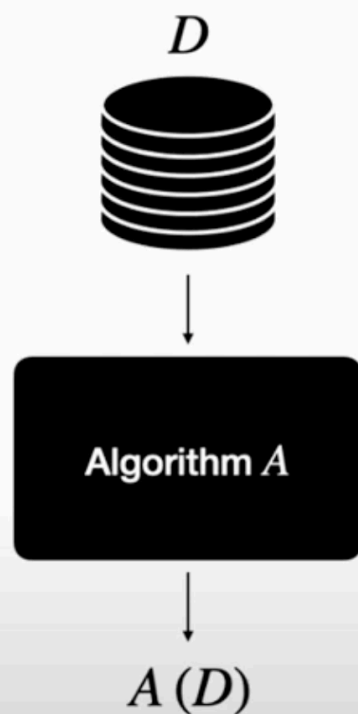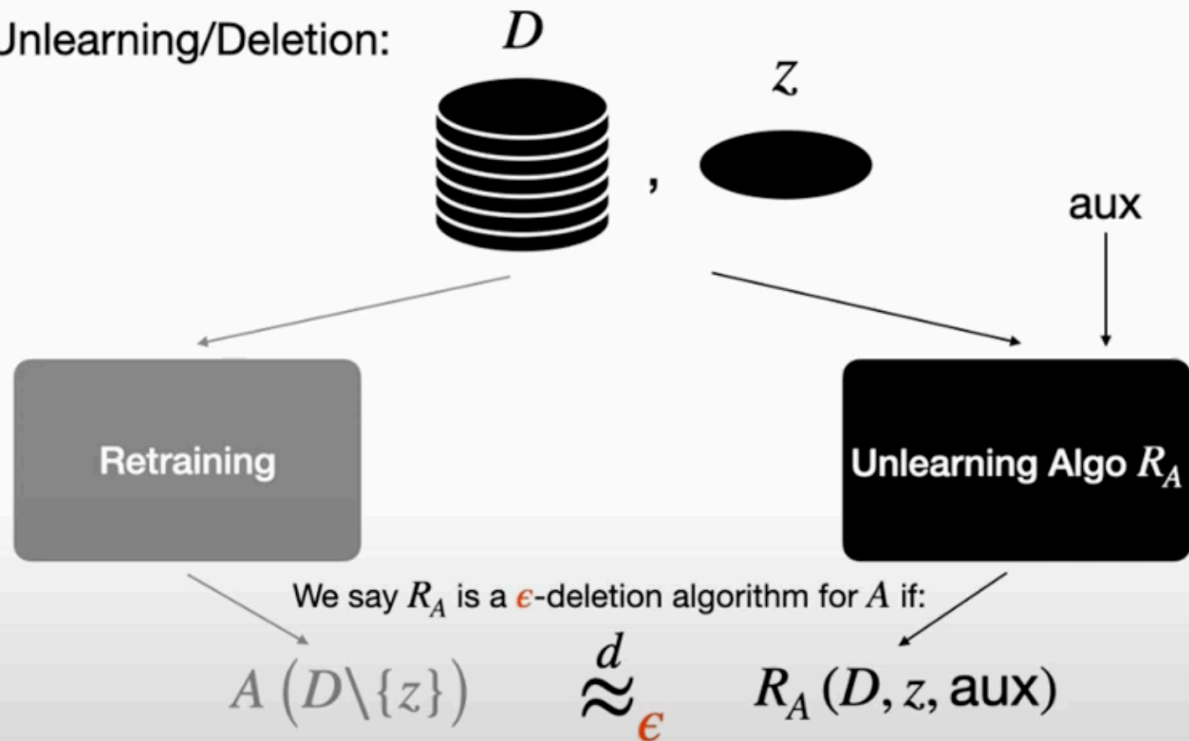
$$\forall E: e^{-\epsilon} \cdot \mathbb{P}\left(A\left(D\backslash\{z\}\right) \in E\right) \leq \mathbb{P}\left(R_A\left(D, z, \text{aux}\right) \in E\right) \leq e^{\epsilon} \cdot \mathbb{P}\left(A\left(D\backslash\{z\}\right) \in E\right)$$