Differential Privacy (DP)

Definition (Differential Privacy)

A randomized algorithm $\mathcal{M}$ with domain $\mathcal{D}$ is (ε, δ)-differentially private if for any $S \subseteq \text{Range}(\mathcal{M})$ and for any adjacent datasets $D, D' \in \mathcal{D}$ (D and D' differ in a single entry), the following holds:

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr\left[\mathcal{M}\left(D'\right) \in S\right] + \delta.$$

Differential Privacy (DP)

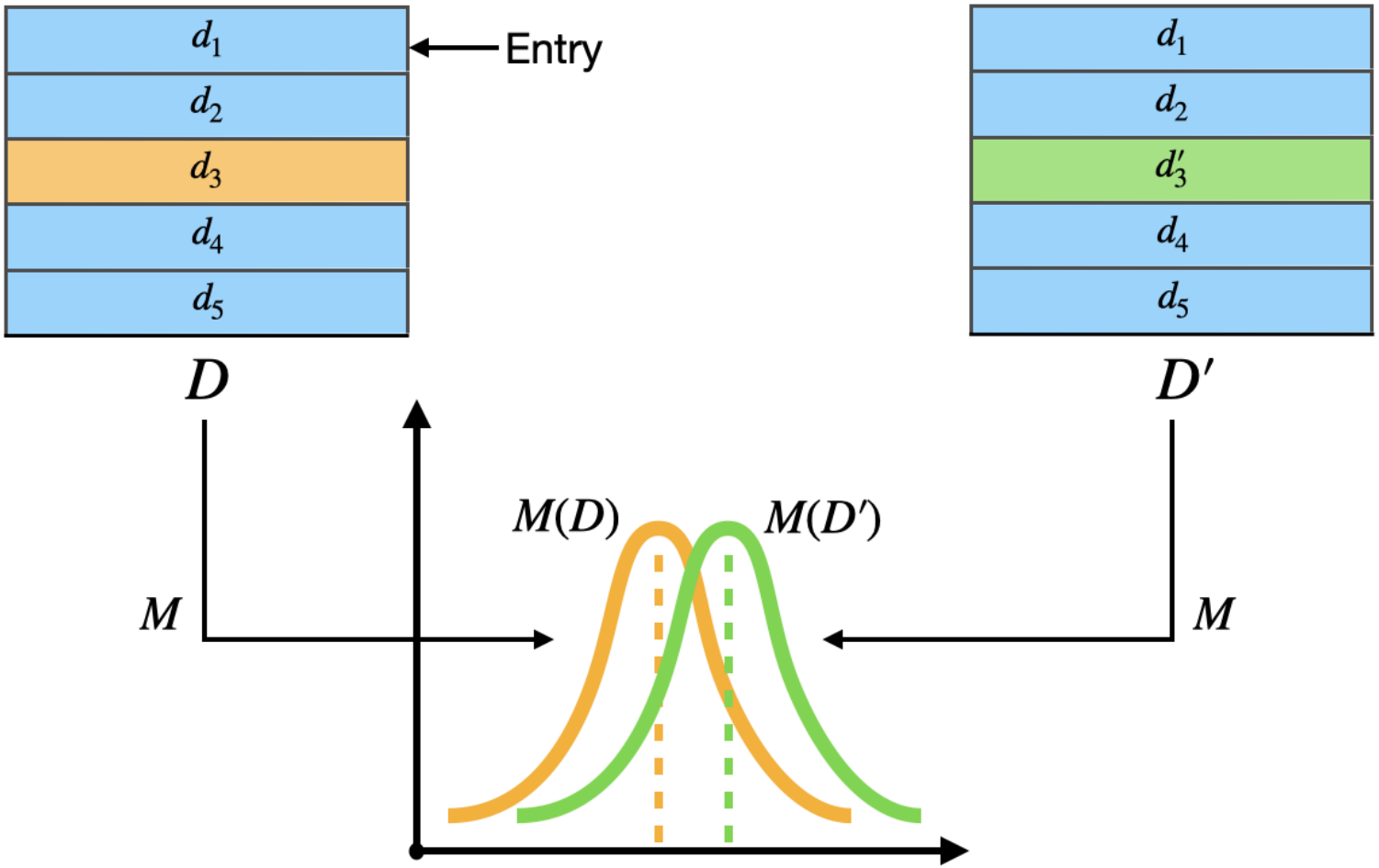

Figure: An example of DP with the Gaussian mechanism.

One More Thing: Privacy

Public Datasets ImageNet [3], CIFAR-10 [10], MNIST [11], etc.

Sensitive Datasets Personal messages, medical information, etc.

Q: Is it possible to know anything about the data processed by a black box like a deep learning model?
A: **YES!!!**

Figure: The recovered facial image and the corresponding training sample [6].

When DP Meets ML: An Example

Definition (Membership Inference Attack)

Given a data record and black-box access to a model, determine if the record was in the model's training dataset.

- How to defend against these attacks?

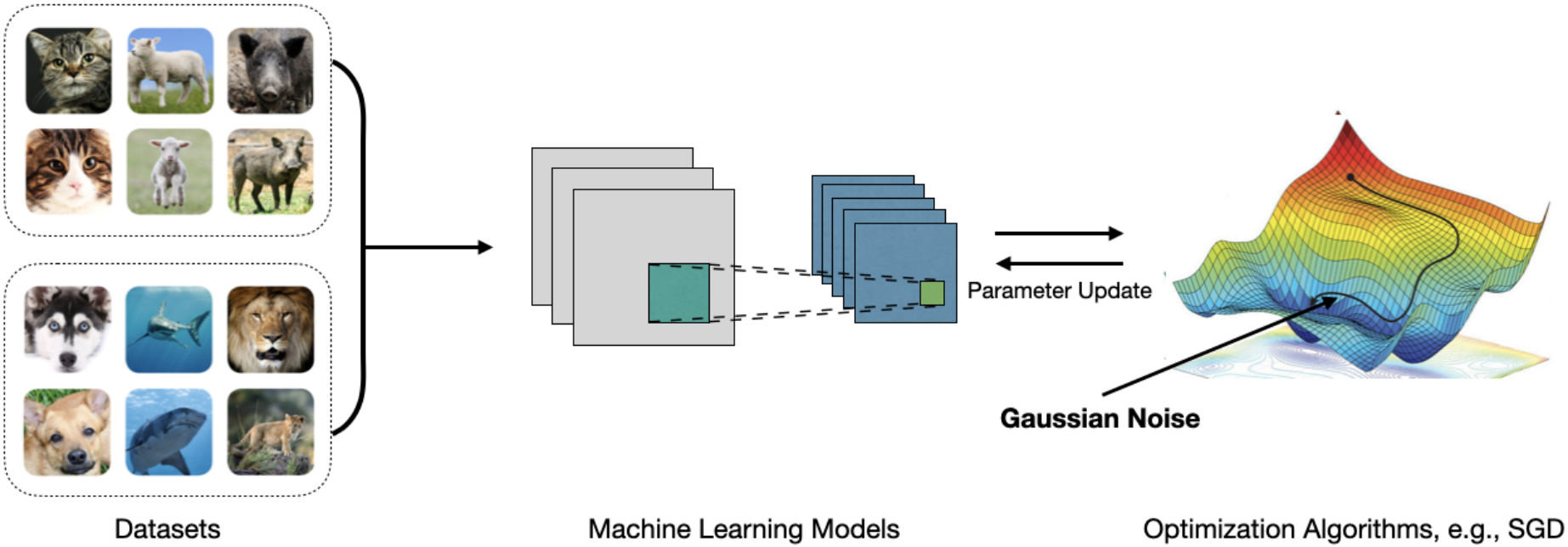

Figure: An example of private ML, DPSGD [1].