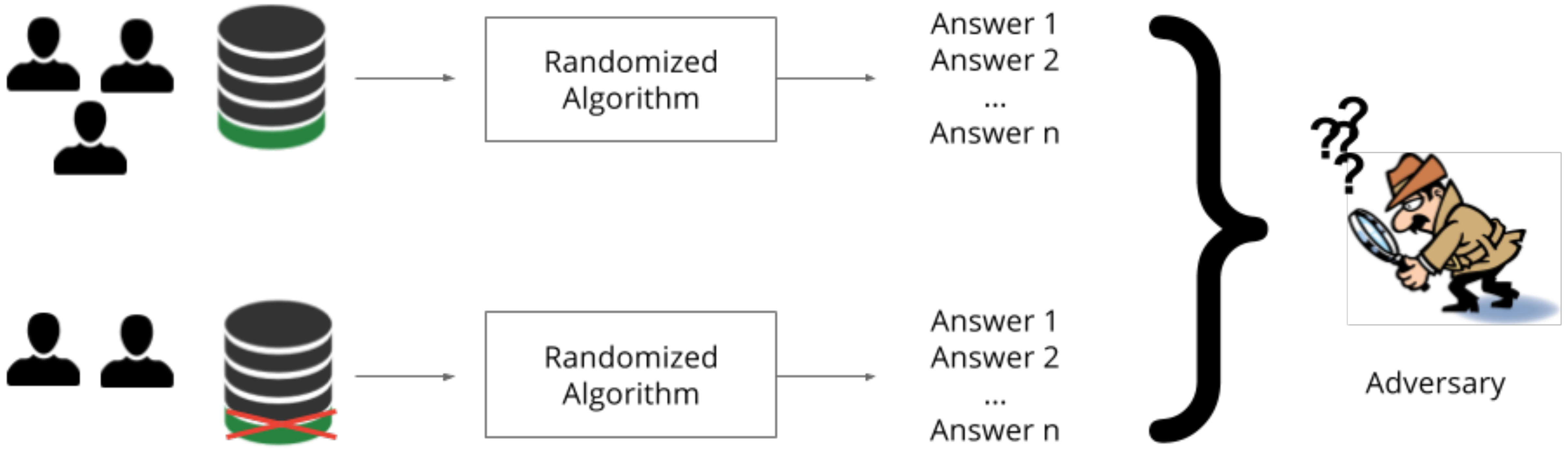
COMP6211I: Trustworthy Machine Learning

Differential privacy part 2

Minhao CHENG

Differential privacy review

- Anonymize the data won't work
- Definition of differential privacy
    - $\log \frac{P(M(D) \in S)}{P(M(D') \in S)} \leq \epsilon$
    - Example:
        - $P(M(D) =$ "Bod has cancer"$) = 0.55$
        - $P(M(D + \text{Bob}) =$ "Bod has cancer"$) = 0.57$
        - $P(M(D + \text{Bob}) =$ "Bod has cancer"$) = 0.8$
        - $\log \frac{P(M(D + \text{Bob}) = \text{"Bod has cancer"})}{P(M(D) = \text{"Bob has cancer"})} = \frac{0.57}{0.55} = 0.0357$
        - $\log \frac{P(M(D + \text{Bob}) = \text{"Bod has cancer"})}{P(M(D) = \text{"Bob has cancer"})} = \frac{0.8}{0.55} = 0.375$



Differential privacy review

- Anonymize the data won't work
- Definition of differential privacy
- $\log \frac{P(M(D) \in S)}{P(M(D') \in S)} \leq \epsilon$
- ϵ-Differential Privacy: $\forall S \; Pr[M(D) \in S] \leq e^{\epsilon} Pr[M(D') \in S]$
- (ϵ, δ)-Differential Privacy: $\forall S \; Pr[M(D) \in S] \leq e^{\epsilon} Pr[M(D') \in S] + \delta$

Differential privacy review

- The privacy amplification theorem:
  - We sample a random fraction q rather than the entire data
  - (ϵ, δ) becomes $(q\epsilon, q\delta)$
- Composition of privacy budgets
  - M_1 is (ϵ_1, δ_1), M_2 has a budget of (ϵ_2, δ_2)
  - The composition is $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$

DP-SGD

Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

Descent

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

DP-SGD

- Naive composition $(qT\epsilon, qT\delta)$
- Strong composition $(q\epsilon\sqrt{T \log 1/\delta}, qT\delta)$
- Moments accountant $(q\epsilon\sqrt{T}, \delta)$

$$c(o; \mathcal{M}, \mathsf{aux}, d, d') \triangleq \log \frac{\Pr[\mathcal{M}(\mathsf{aux}, d) = o]}{\Pr[\mathcal{M}(\mathsf{aux}, d') = o]}. \qquad (1)$$

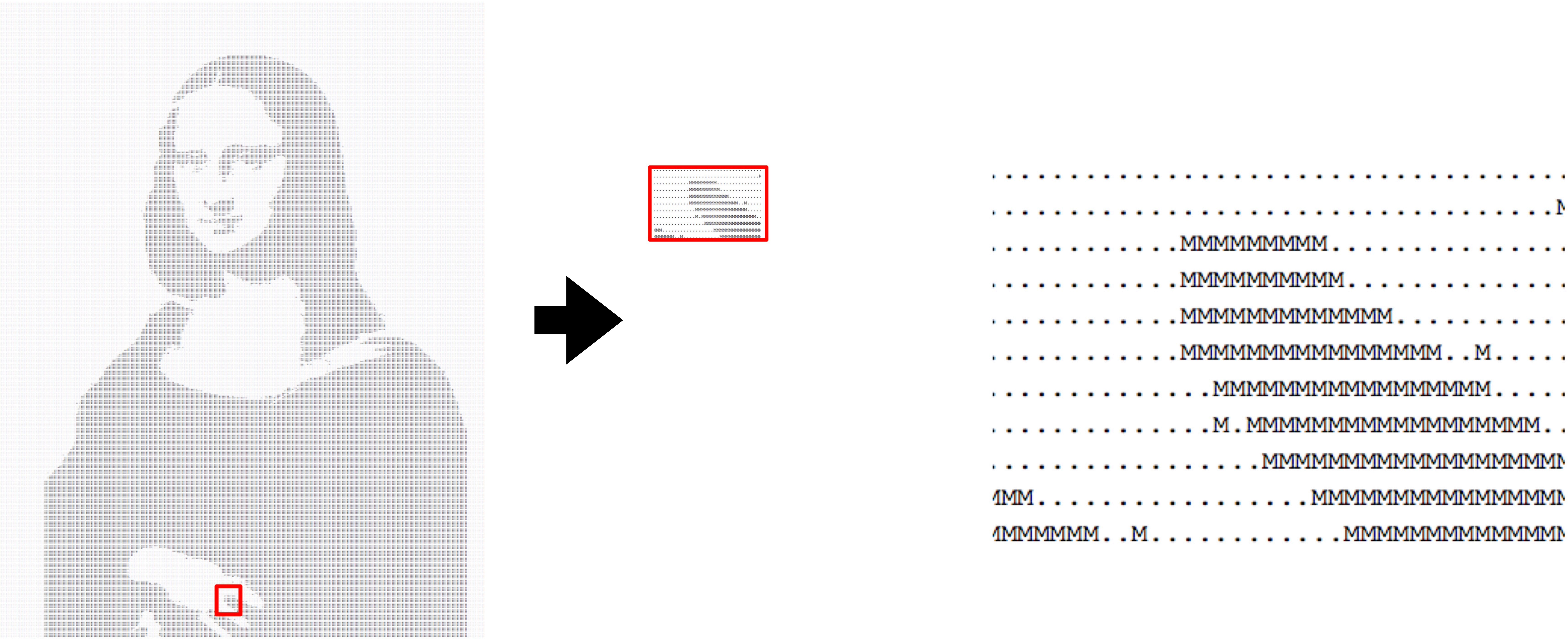
$$\alpha_{\mathcal{M}}(\lambda) \triangleq \max_{\mathsf{aux}, d, d'} \alpha_{\mathcal{M}}(\lambda; \mathsf{aux}, d, d'),$$

$$\alpha_{\mathcal{M}}(\lambda; \mathsf{aux}, d, d') \triangleq \log \mathbb{E}_{o \sim \mathcal{M}(\mathsf{aux}, d)}[\exp(\lambda c(o; \mathcal{M}, \mathsf{aux}, d, d'))]. \qquad (2)$$

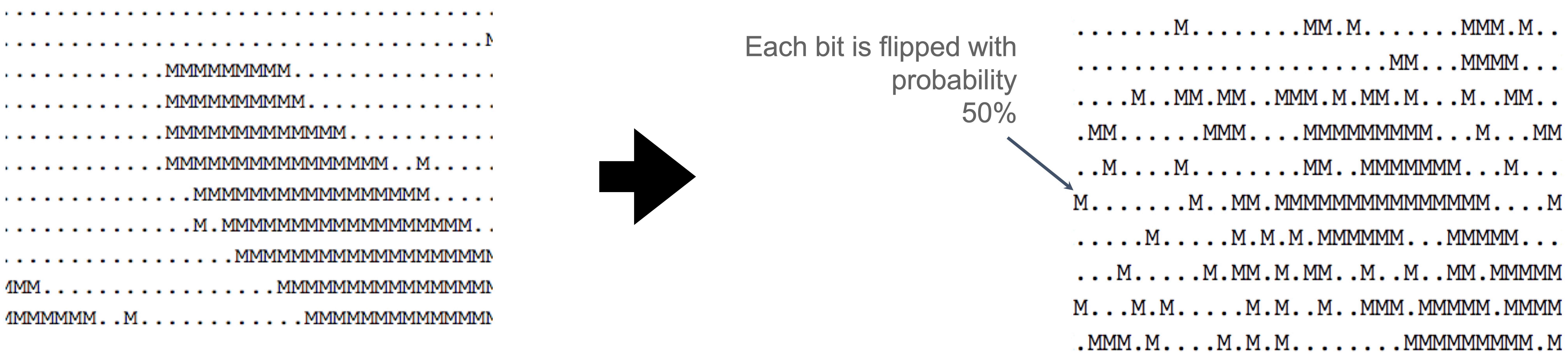
*2. [**Tail bound**] For any $\varepsilon > 0$, the mechanism $\mathcal{M}$ is (ε, δ)-differentially private for*

$$\delta = \min_{\lambda} \exp(\alpha_{\mathcal{M}}(\lambda) - \lambda\varepsilon).$$

A metaphor for private learning



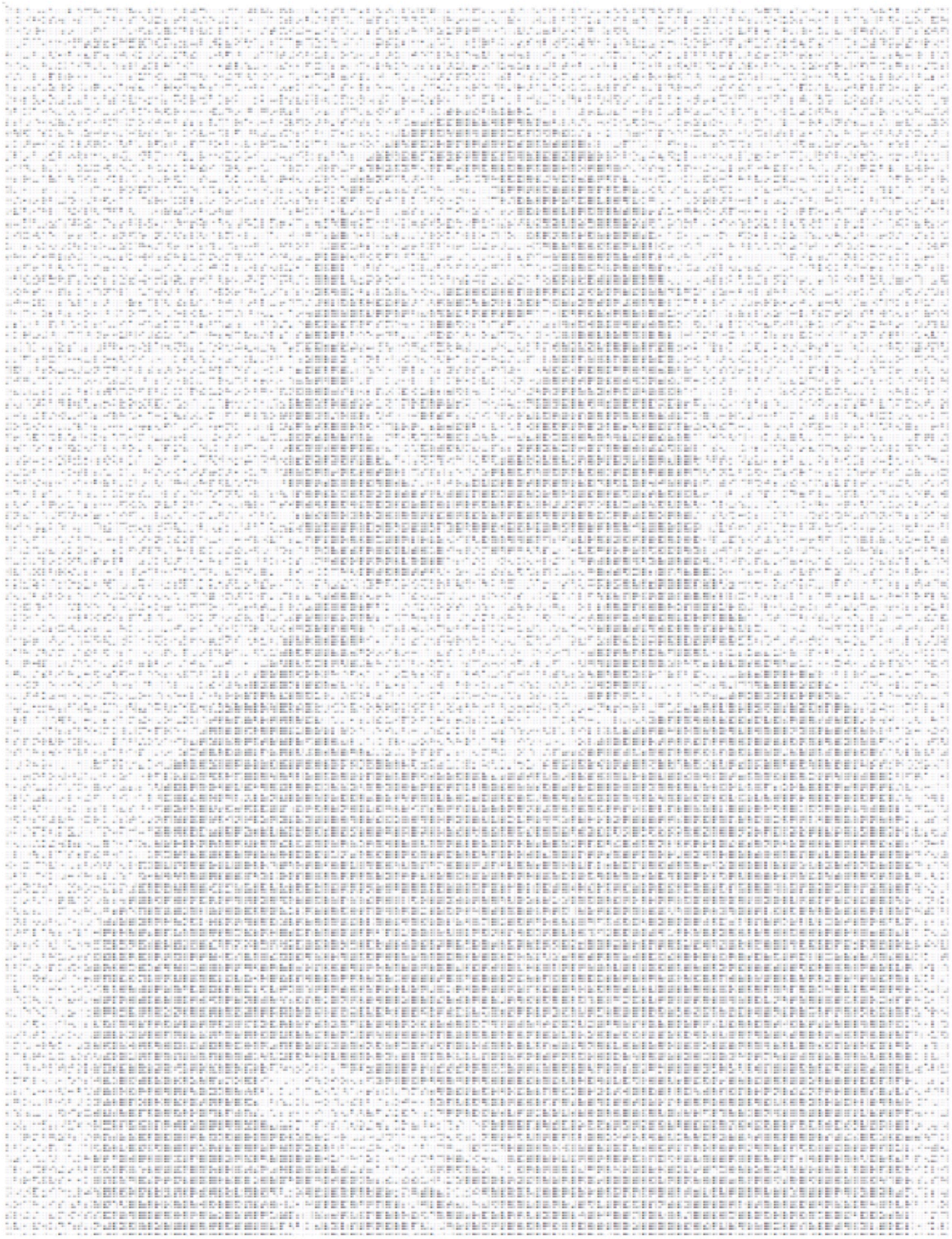
A metaphor for private learning



A metaphor for private learning

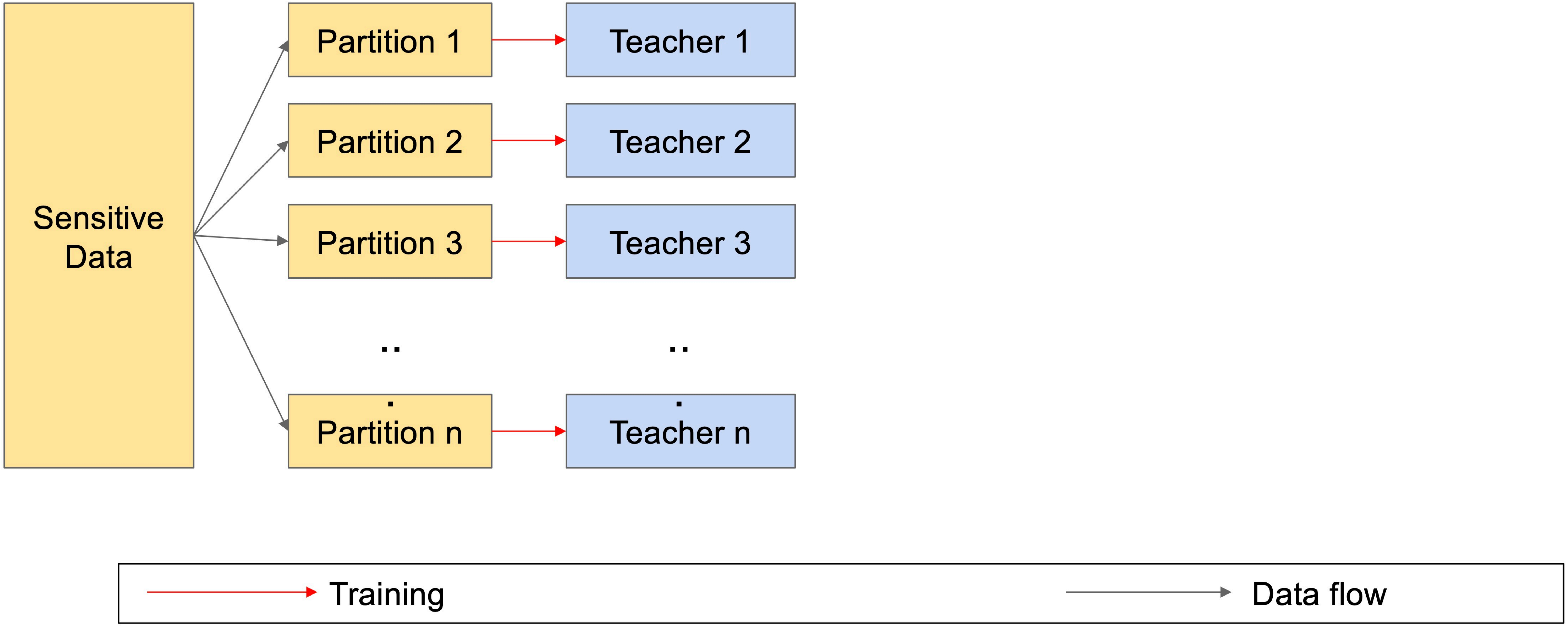
Each bit is flipped with probability 50%

```
.......M........MM.M.......MMM.M..
......................MM...MMMM...
....M..MM.MM..MMM.M.MM.M...M..MM..
.MM......MMM....MMMMMMMMM...M...MM
..M....M........MM..MMMMMMM...M...
M.......M..MM.MMMMMMMMMMMMMMM....M
.....M.....M.M.M.MMMMMM...MMMMM...
...M.....M.MM.M.MM..M..M..MM.MMMMM
M...M.M.....M.M..M..MMM.MMMMM.MMMM
.MMM.M....M.M.M........MMMMMMMMM.M
```

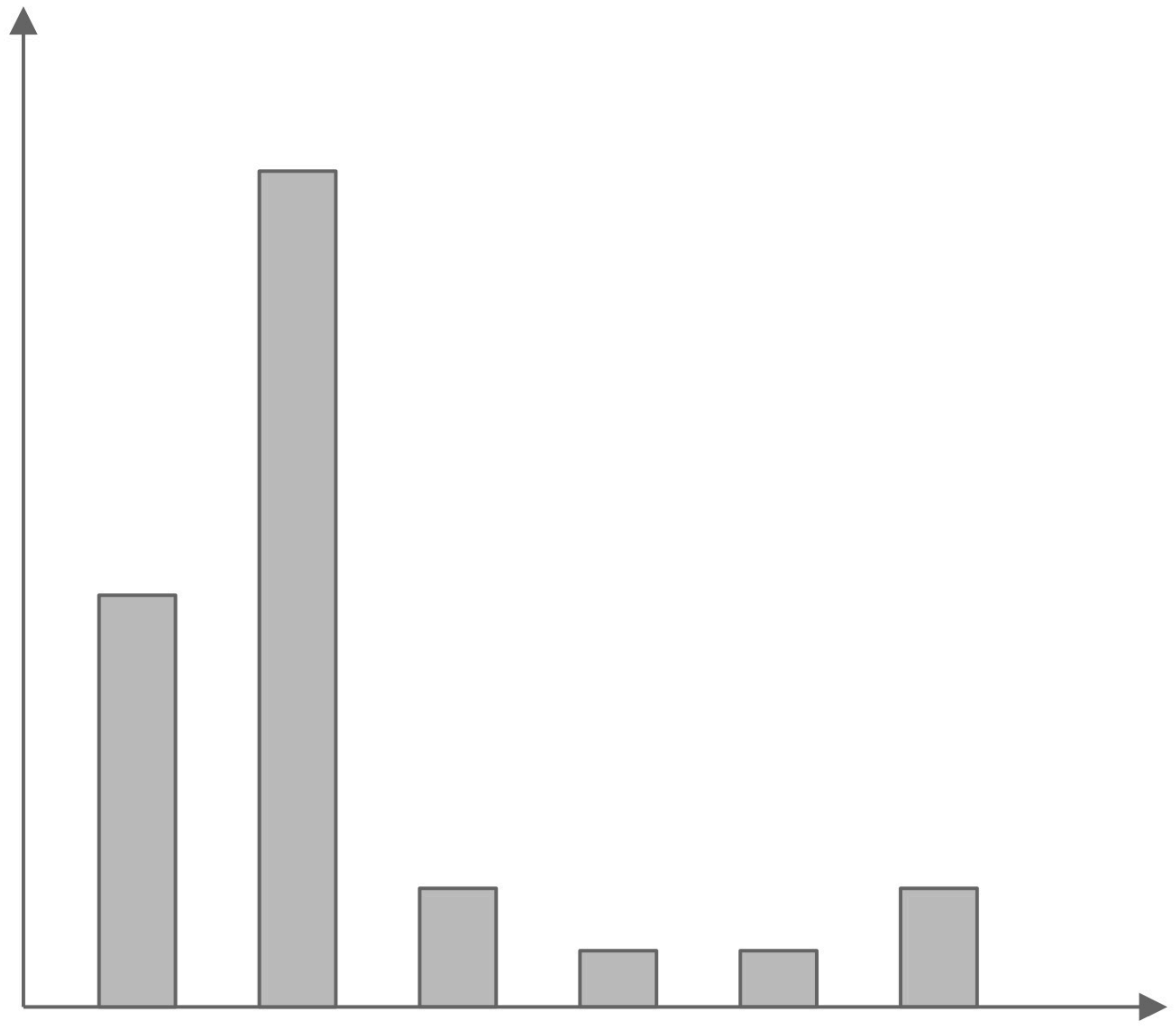


Private Aggregation of Teacher Ensembles

PATE

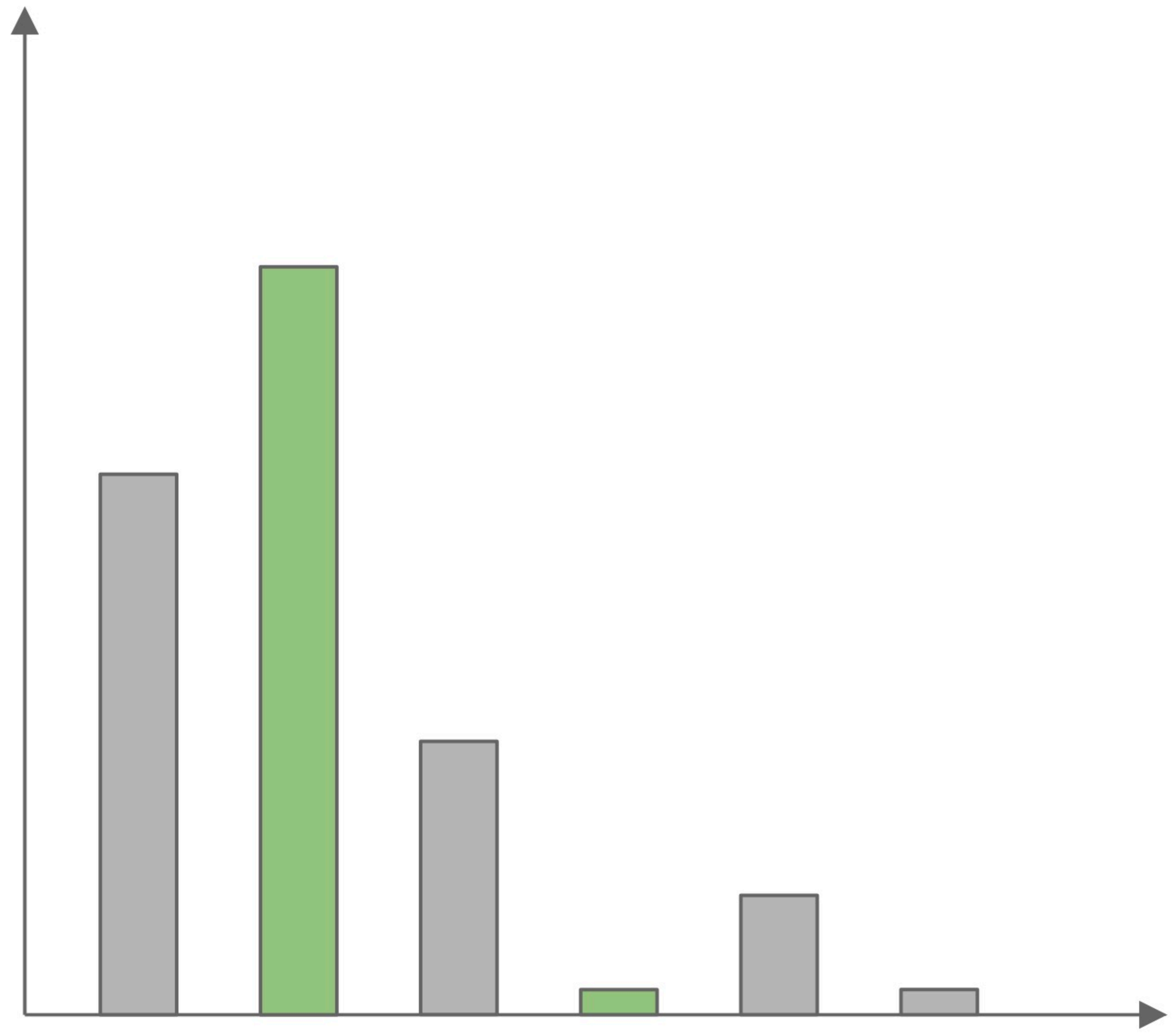


Aggregation



Count votes

$$n_j(\vec{x}) = |\{i : i \in 1 \ldots n\}, f_i(\vec{x}) = j|$$

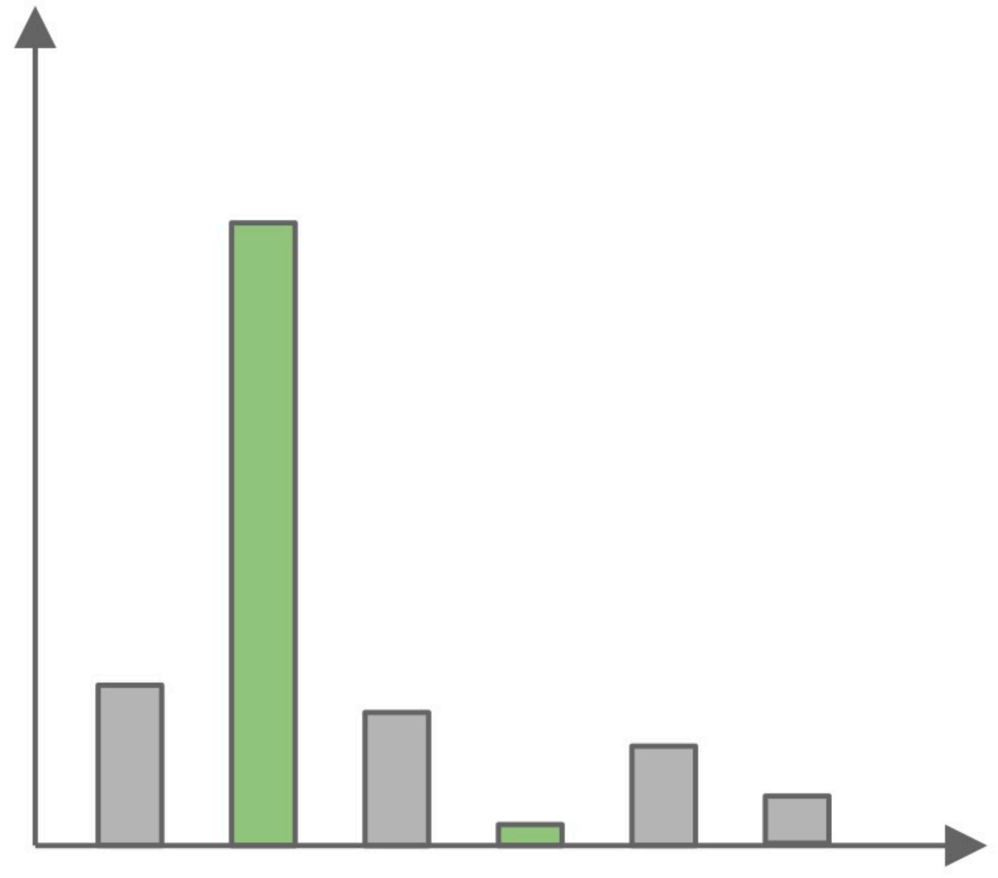


Take maximum

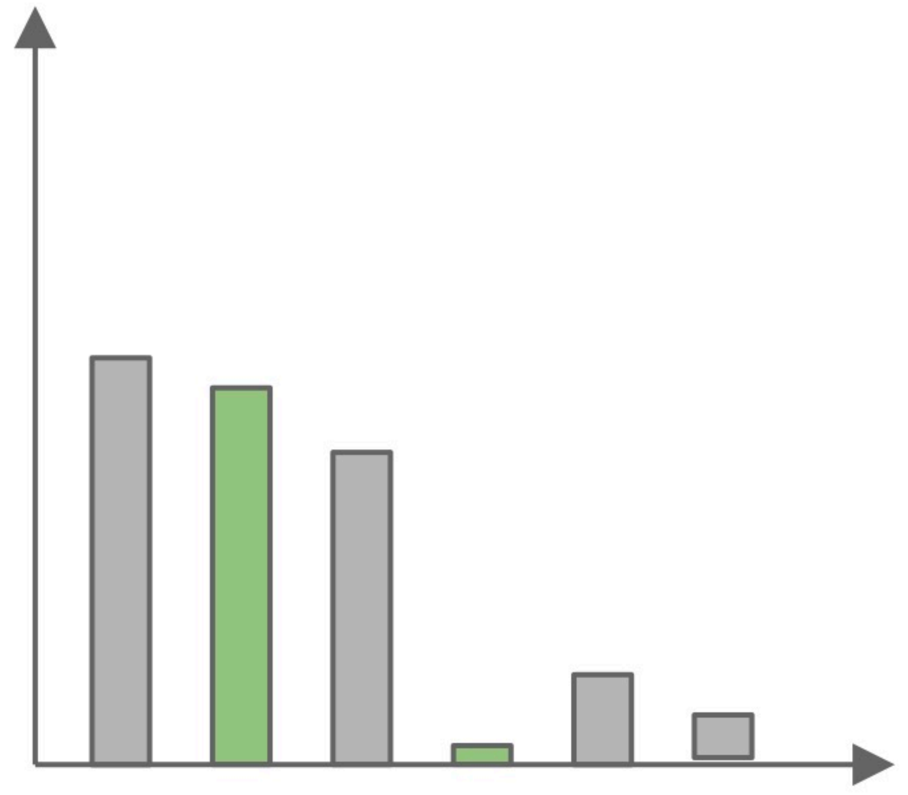
$$f(x) = \arg\max_j \{n_j(\vec{x})\}$$

Intuitive privacy analysis

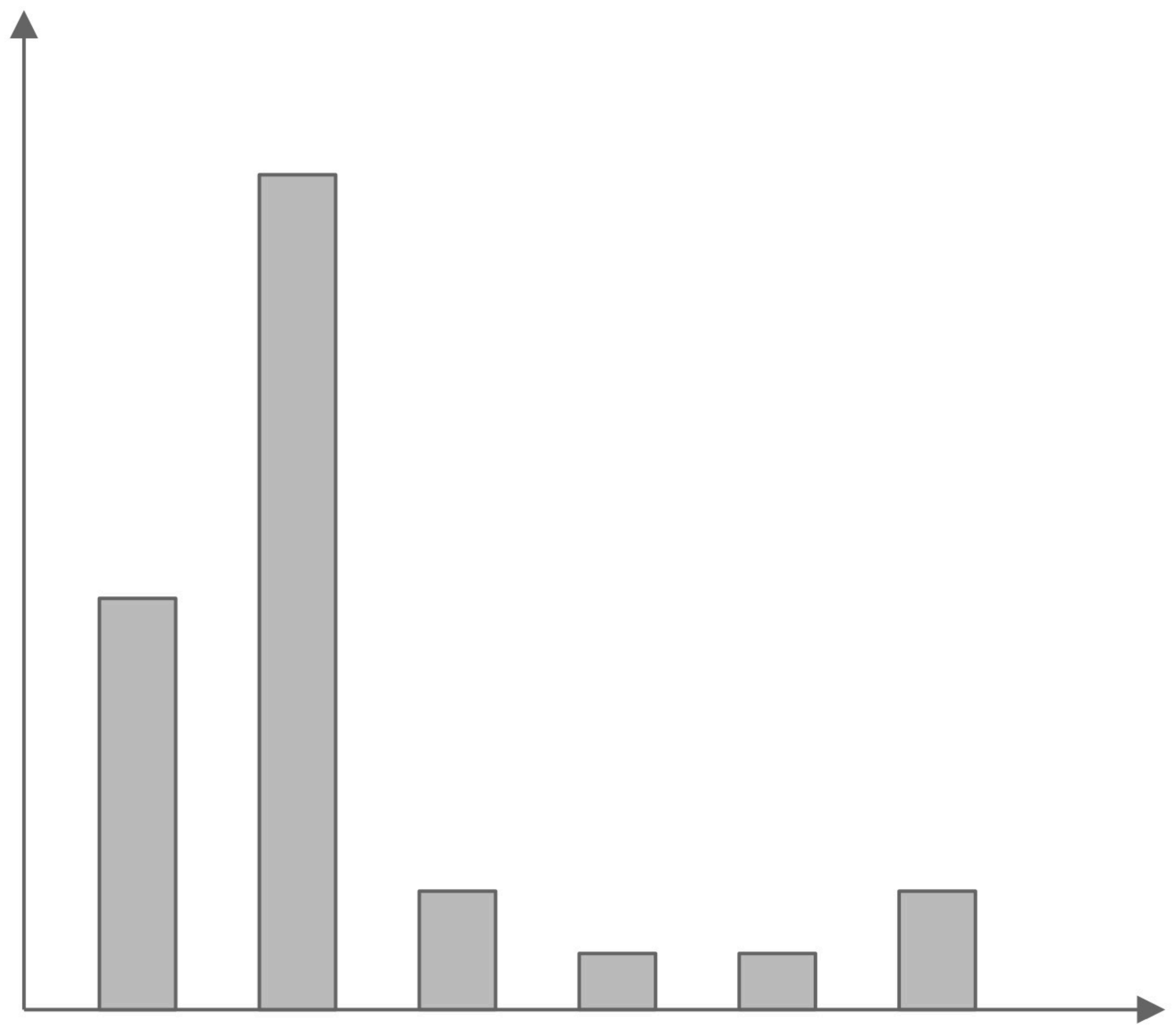
- If most teachers agree on the label, it does not depend on specific partitions, so the privacy cost is small.



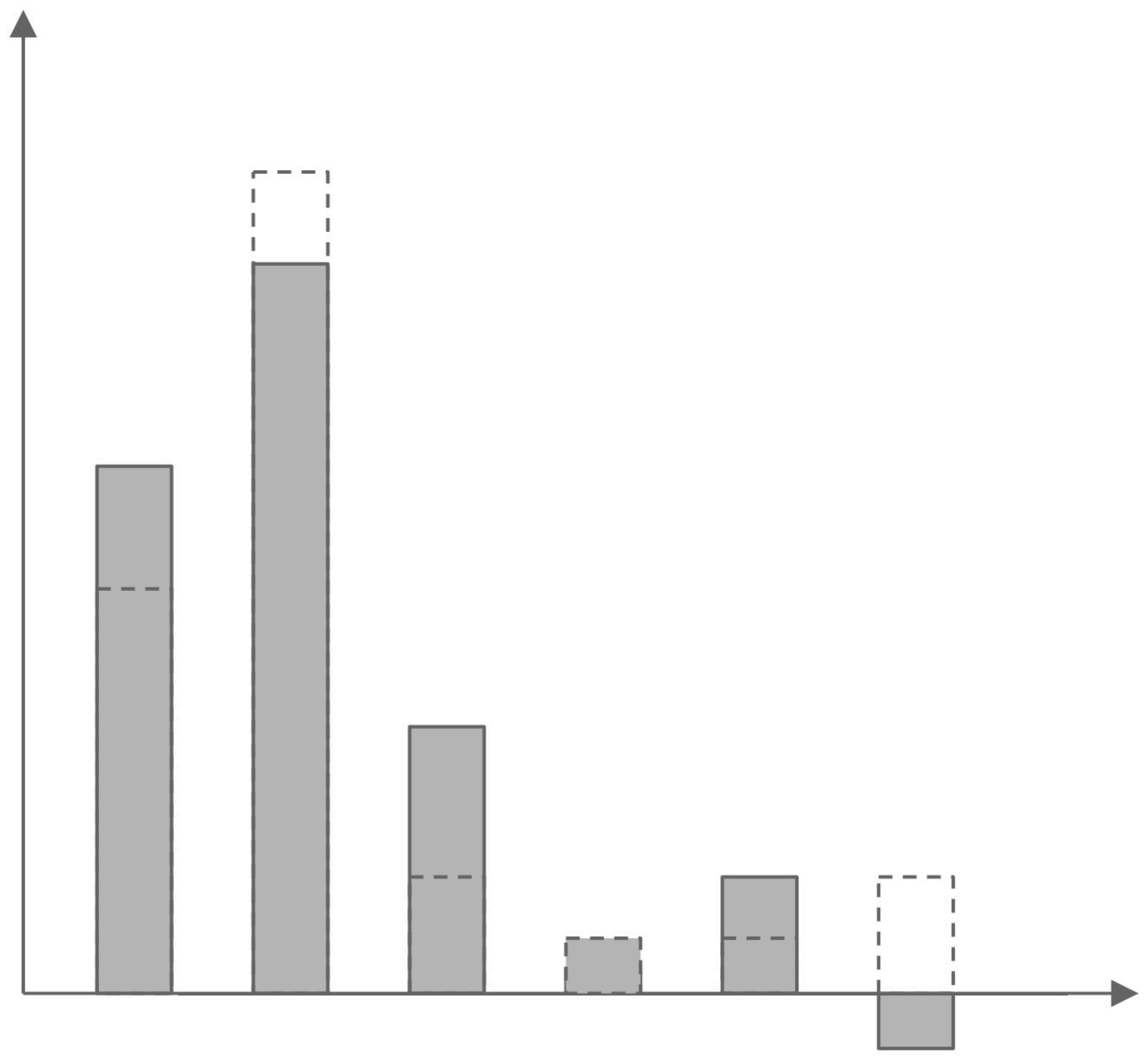
- If two classes have close vote counts, the disagreement may reveal private information.



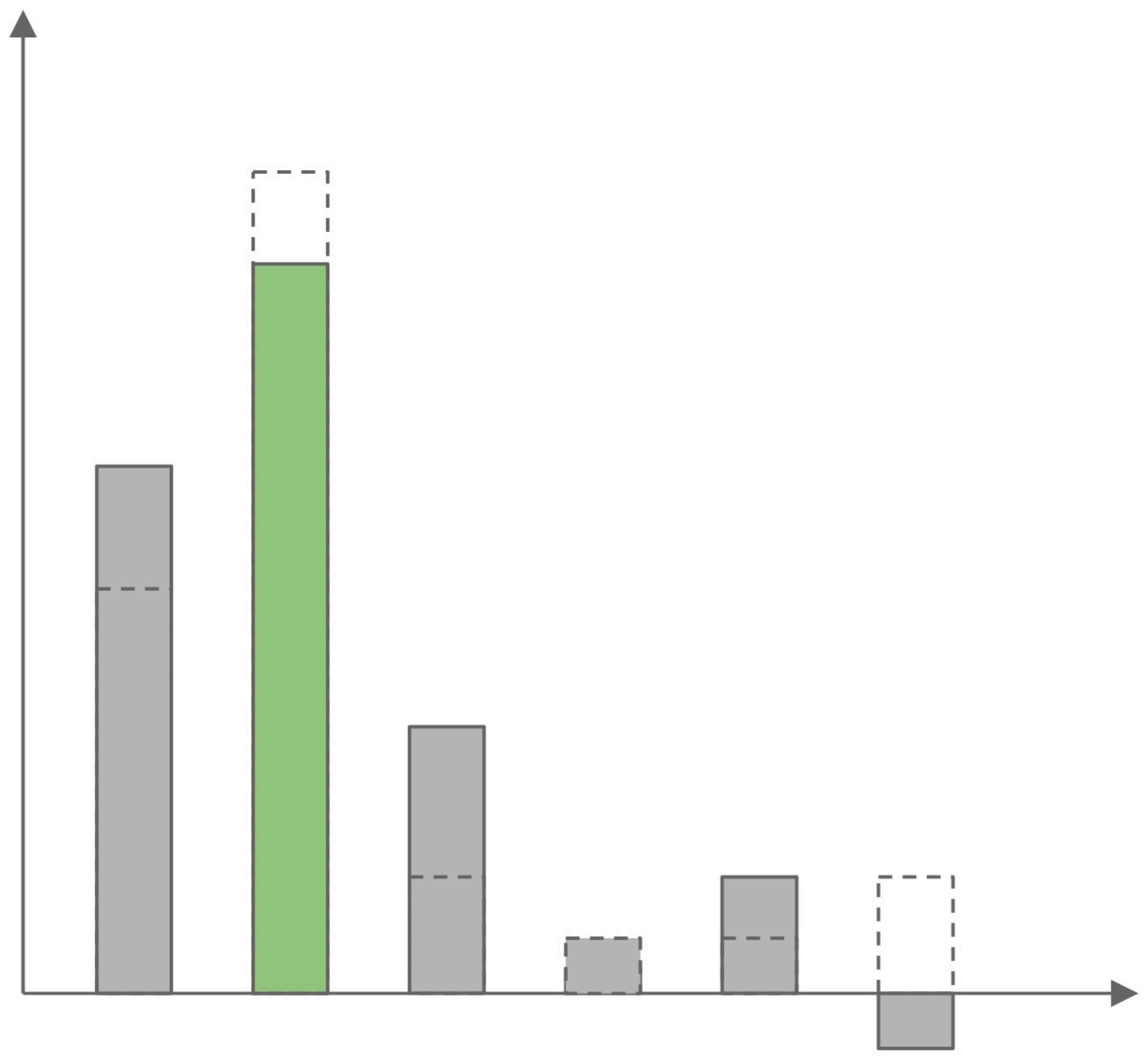
Noisy aggregation



Count



Add Laplacian



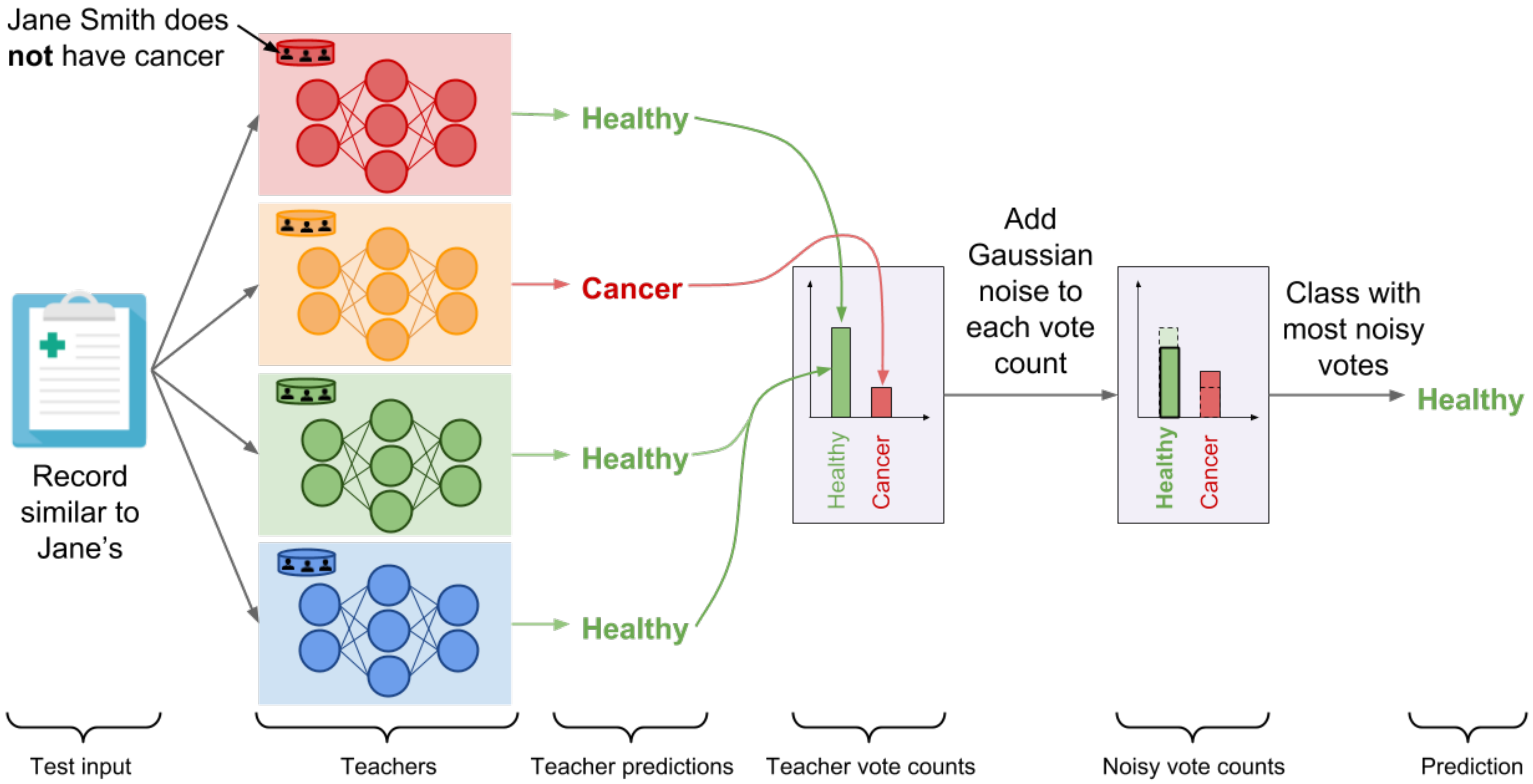
Take maximum

$$n_j(\vec{x}) = |\{i : i \in 1\ldots n\}, f_i(\vec{x}) = j|$$

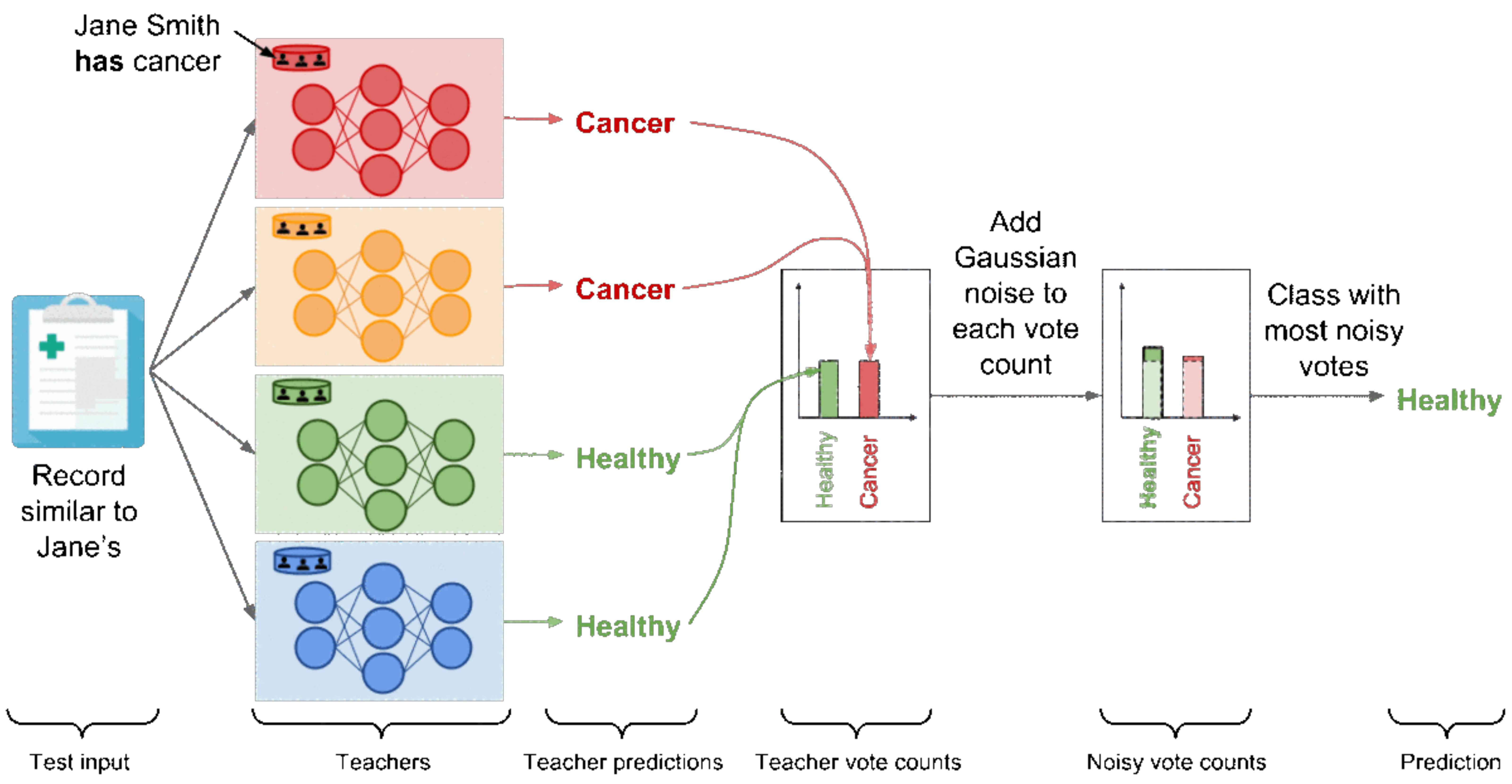
$$Lap(\frac{1}{\epsilon})$$

$$f(x) = \arg\max_j\{n_j(\vec{x}) + Lap(\frac{1}{\epsilon})\}$$

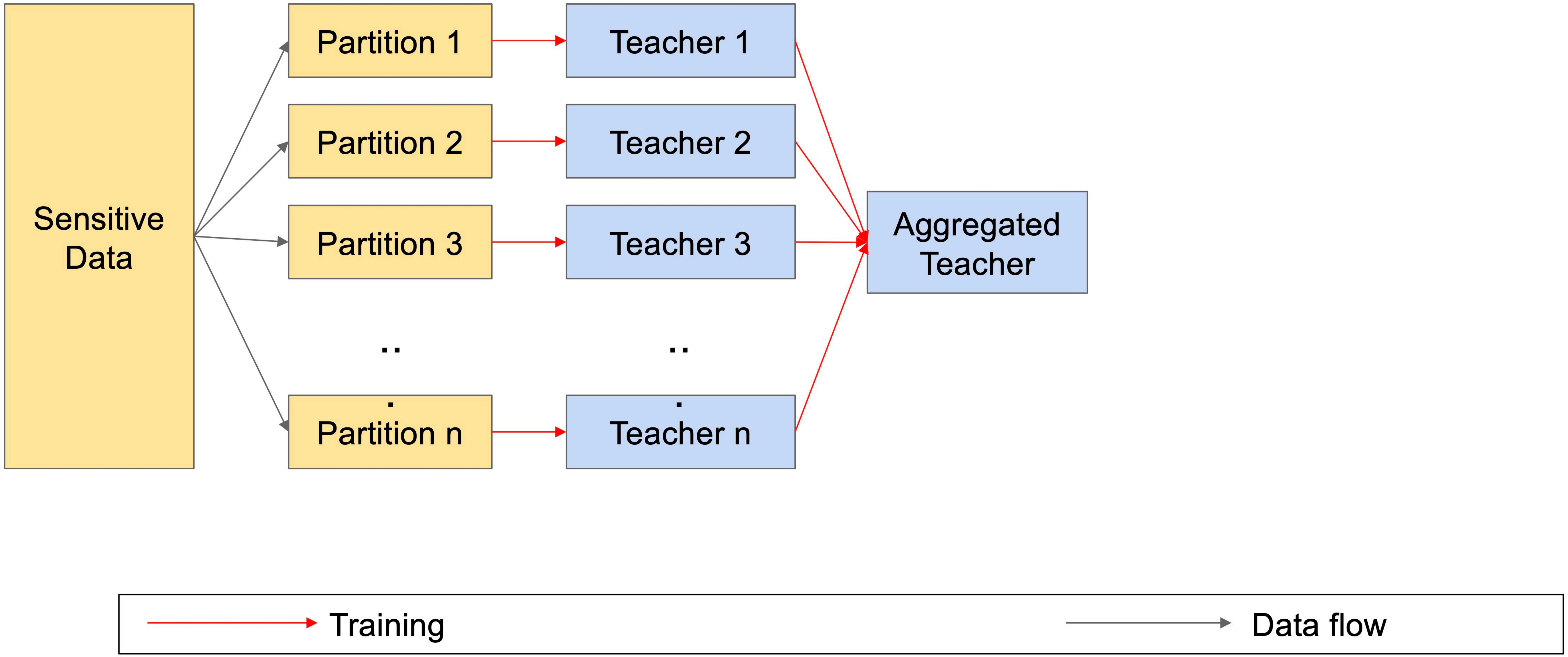
Critical point



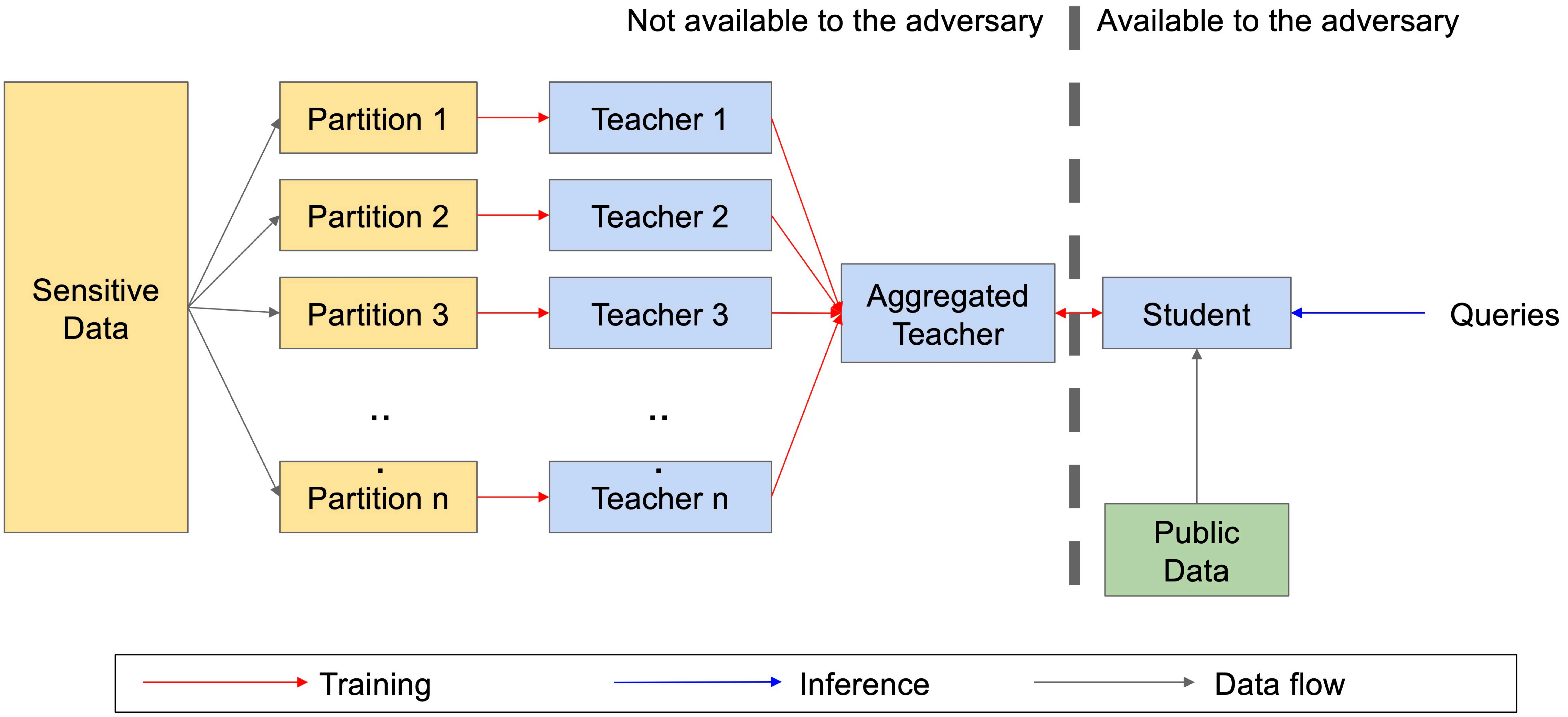
Critical point



Teacher ensemble



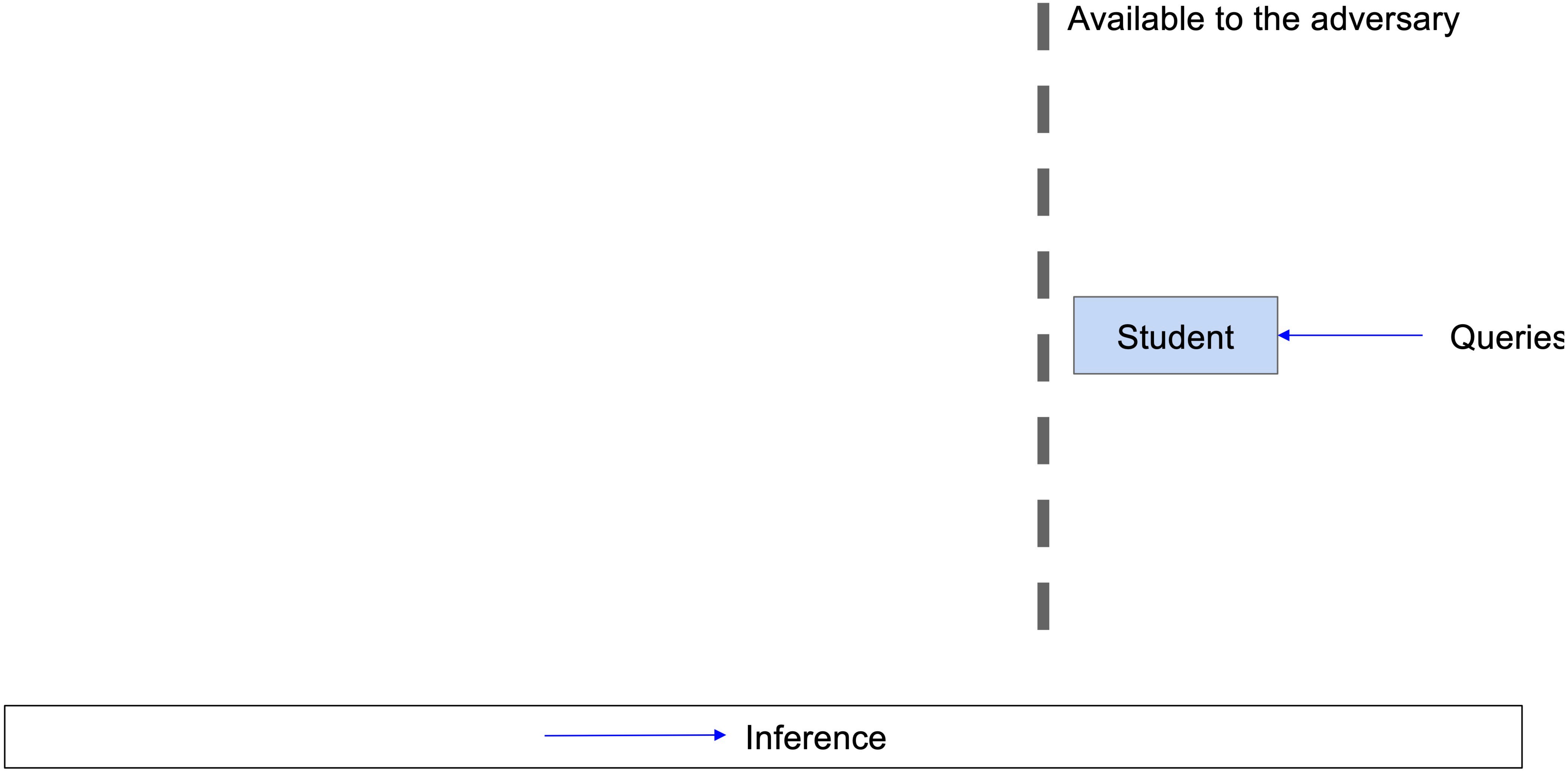
Student training



Why train an additional “student” model?

- The aggregated teacher violates our threat model:
    - Each prediction increases total privacy loss.
        - Privacy budgets create a tension between the accuracy and number of predictions.
    - Inspection of internals may reveal private data.
        - Privacy guarantees should hold in the face of white-box adversaries.

Deployment



Differential privacy analysis

Differential privacy:
A randomized algorithm *M* satisfies (ε,δ) differential privacy if for all pairs of neighbouring datasets (*d*,*d'*), for all subsets *S* of outputs:

$$Pr[M(d) \in S] \leq e^{\varepsilon} Pr[M(d') \in S] + \delta$$

- Application of the Moments accountant technique (Abadi et al, 2016)
- Strong quorum ⇒ Small privacy cost
- Bound is data-dependent: computed using the empirical quorum

Rényi Differential Privacy

- ϵ-Differential Privacy
  - $\max_x P(x)/Q(x) < e^\epsilon$
- Rényi Divergence at ∞
  - $D_\infty(P\,||\,Q) < \epsilon$

Rényi Divergence

- $D_\alpha(P||Q) = \dfrac{1}{\alpha - 1} \log E_Q[(\dfrac{P(x)}{Q(x)})^\alpha]$
- $D_1(P||Q) = \lim_{\alpha \to 1} D_\alpha(P||Q) = E_P[\log(\dfrac{P(x)}{Q(x)})]$
- $D_\infty(P||Q) = \lim_{\alpha \to \infty} D_\alpha(P||Q) = \log \max_x \dfrac{P(x)}{Q(x)}$

Rényi Differential Privacy

- (α, ϵ) Rényi Differential Privacy (RDP):
  - $\forall D, D' : D_\alpha(M(D)\,||\,M(D')) < \epsilon$
- (∞, ϵ)-RDP is ϵ-DP
- (α, ϵ)-RDP $\Rightarrow (\epsilon + \frac{\log 1/\delta}{\alpha - 1}, \delta)$-DP for any δ

Rényi Differential Privacy

Bad outcomes interpretation

- ϵ-Differential Privacy: $\forall S \; Pr[M(D) \in S] \leq e^{\epsilon} Pr[M(D') \in S]$
- (α, ϵ) Rényi Differential Privacy (RDP):
  $\forall S \; Pr[M(D) \in S] \leq (e^{\epsilon} Pr[M(D') \in S])^{1-1/\alpha}$
- (ϵ, δ)-Differential Privacy: $\forall S \; Pr[M(D) \in S] \leq e^{\epsilon} Pr[M(D') \in S] + \delta$

Rényi Differential Privacy

Why better

- No catastrophic failure mode
- The composition of an (α, ϵ_1)-RDP algorithm and an (α, ϵ_2)-RDP algorithm is $(\alpha, \epsilon_1 + \epsilon_2)$