k-Anonymity. A database table is k-anonymous if for every tuple $t \in T$, there are at least $k-1$ other tuples $t' \in T$ such that the quasi-identifiers of t equal the quasi-identifiers of t'. That is, you must be able to group all of the tuples in T into groups of size k or larger such that the tuples in each group agree on *all* quasi-identifier values.

Two kinds of privacy attacks are possible against k-anonymized tables. In a *homogeneity attack*, all tuples in a group have the same sensitive value, so membership in the group implies having that sensitive value. In an *auxiliary data attack*, auxiliary data is used to rule out all members of the group except one.

Differential privacy. A mechanism $\mathcal{M}$ provides ϵ-differential privacy if for all neighboring databases X and Y, and all sets of outcomes S, $\Pr[\mathcal{M}(X) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(Y) \in S]$. $\mathcal{M}$ provides (ϵ, δ)-differential privacy if $\Pr[\mathcal{M}(X) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(Y) \in S] + \delta$. Pure ϵ-differential privacy can be seen as a special case of (ϵ, δ)-differential privacy in which $\delta = 0$.

X and Y are neighbors under *unbounded* differential privacy if one person's data can be added to or removed from X to arrive at Y. X and Y are neighbors under *bounded* differential privacy if one person's data can be changed in X to arrive at Y. Unbounded differential privacy is usually preferred.

The L_1 sensitivity of a function (or query) $f : \mathcal{D} \to \mathcal{R}^n$ is the maximum L_1 norm of the difference between $f(X)$ and $f(Y)$: $\max_{X,Y} \|f(X) - f(Y)\|_1$. The L_1 norm of a vector is $\|V\|_1 = \sum_i |V_i|$. The L_2 sensitivity of f is $\max_{X,Y} \|f(X) - f(Y)\|_2$, where the L_2 norm is $\|V\|_2 = \sqrt{\sum_i V_i^2}$.

To determine sensitivity, think about the *worst case* person you could add or remove in the database to change the output of the query. As a rule of thumb, the sensitivity of a counting query is usually 1, and the sensitivity of sum or average queries is unbounded unless the data model constrains attribute values in the database. Filtering (as in SQL `WHERE` clauses) does not usually change sensitivity.

Parallel Composition. If $\mathcal{M}$ provides (ϵ, δ)-differential privacy, and we split the database D into k disjoint subsets $D_1, \ldots, D_k$, then the privacy cost of releasing all of $\mathcal{M}(D_1), \ldots, \mathcal{M}(D_k)$ is (ϵ, δ). As a rule of thumb, histogram or histogram-like queries (including contingency tables) split the database into chunks based on the value of one attribute in each tuple, so parallel composition can be applied to these queries.

Sequential Composition. If $\mathcal{M}_1$ provides (ϵ_1, δ_1)-differential privacy, and $\mathcal{M}_2$ provides (ϵ_2, δ_2)-differential privacy, then we can release both $\mathcal{M}_1(X)$ and $\mathcal{M}_2(X)$ with privacy cost $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$. $\mathcal{M}_2$ may be *adaptive*, meaning it leverages the output of $\mathcal{M}_1$ arbitrarily in computing its output.

Advanced Composition. If $\mathcal{M}$ provides (ϵ, δ)-differential privacy, then $\mathcal{M}$ under k-fold adaptive composition provides $(2\epsilon\sqrt{2k\log(1/\delta')}, \delta' + k\delta)$-differential privacy. Under k-fold adaptive composition, the ith invocation of $\mathcal{M}$ is allowed to leverage all of the outputs of $\mathcal{M}_1, \ldots, \mathcal{M}_{i-1}$ in computing its output. As a rule of thumb, most loops in programs are instances of k-fold adaptive composition.

Laplace Mechanism. If f has L_1 sensitivity $\Delta_1 f$, then releasing $f(X) + \text{Lap}(\frac{\Delta_1 f}{\epsilon})$ provides ϵ-differential privacy, where $\text{Lap}(s)$ is noise sampled from the Laplace distribution with center 0 and scale s.

Gaussian Mechanism. If f has L_2 sensitivity $\Delta_2 f$, then releasing $f(X) + \mathcal{N}(\sigma^2)$, where $\sigma^2 = \frac{2(\Delta_2 f)^2 \log(1.25/\delta)}{\epsilon^2}$ provides (ϵ, δ)-differential privacy, where $\mathcal{N}(\sigma^2)$ is noise sampled from the Gaussian distribution with center 0 and variance σ^2.