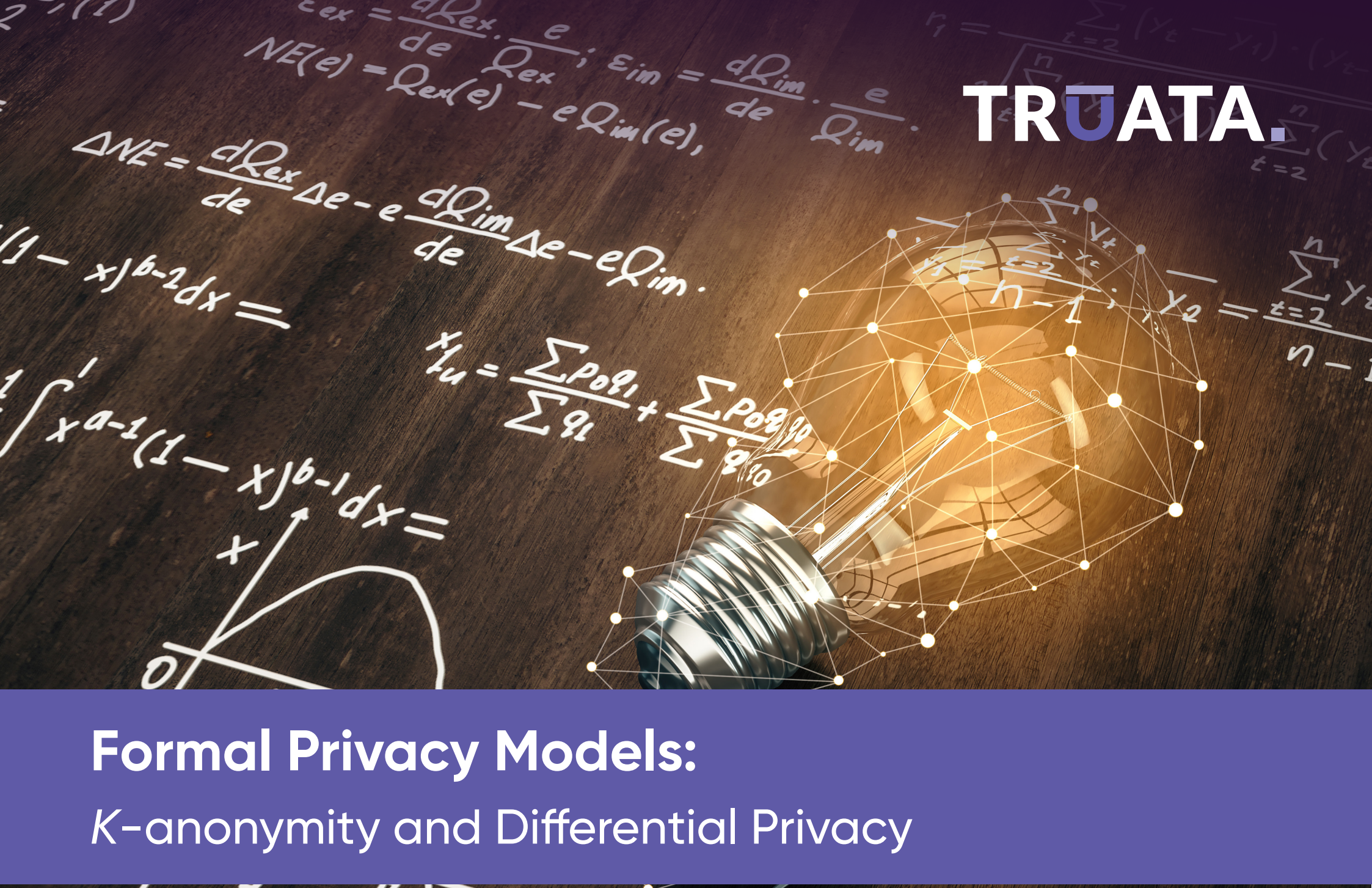


TRUATA.

Formal Privacy Models:

K-anonymity and Differential Privacy

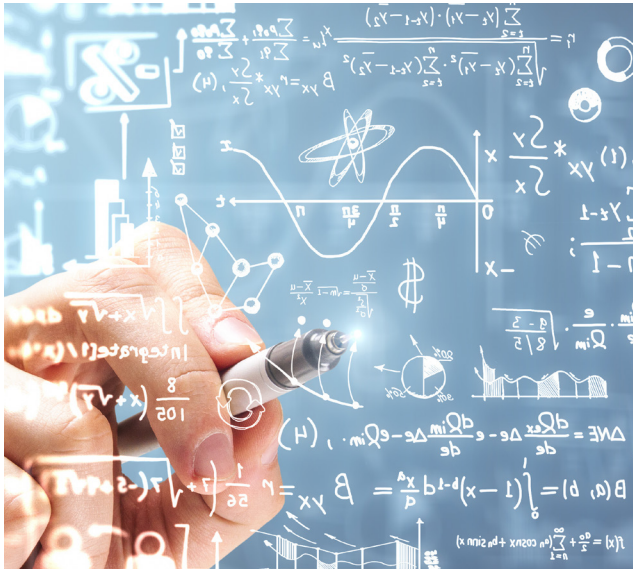
In today's world, analyzing large volumes of data has become the norm. This collected data, however, often contains sensitive personal data, and processing such personal data triggers concerns about the privacy of the individuals in that data.

Dr. Imran Khan
Data Scientist at Truata.

Attempts to address these privacy concerns at a technical level can be traced back as far as the 1970s with statistical disclosure control, where Tore Dalenius discussed the challenge of releasing personal data while preserving the privacy of the individuals in a dataset [1].

When looking at the challenge of preserving privacy or reducing the risk of re-identification, we frequently come across mentions of *k*-anonymity [2] and differential privacy [4]. Other popular privacy models include *l*-diversity and *t*-closeness, but over 80 of these privacy models (also known as privacy definitions or privacy metrics) are described [8]. The idea behind using such privacy models for datasets is that compliance with these models enables reasoning about the risk of re-identification based on mathematical guarantees. It is important to note that a mathematical guarantee is only a guarantee up until a certain level that is selected by the data controller; this means there can be variance in the amount of protection provided.

A useful analogy to adopt when considering mathematical guarantee is that of a retailer's delivery promise to a consumer. One retailer may guarantee delivery between 9am and 5pm within a range of days, while another may guarantee delivery at a specific time on a specific day.



Both retailers offer a guarantee; however, one is far more useful and inspires greater confidence than the other. When considering this from a data privacy point of view, a guarantee that it will be possible to re-identify individuals with 90% accuracy is considerably worse than a guarantee that it will be possible to re-identify individuals with 0.1% accuracy.

A common misconception relating to privacy models, such as *k*-anonymity or differential privacy, is that they are privacy '*techniques*'. However, *k*-anonymity is not a privacy technique; instead, it can be considered as a *characteristic* of a dataset. To achieve *k*-anonymity, one can deploy various techniques, including common ones such as *generalization* and *suppression*.

Let us explore the *k*-anonymity privacy model to gain an understanding as to how it is defined, and how *k*-anonymity is achieved.

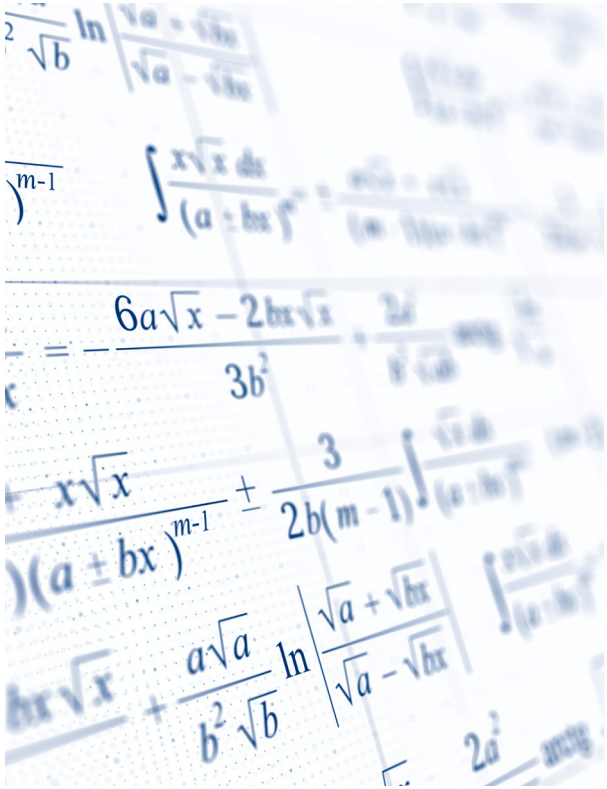
K-anonymity

K-anonymity is a formal privacy model that has been proposed by data anonymization and privacy researchers. K-anonymity categorizes attributes into the following non-exclusive categories: identifiers, quasi-identifiers, and sensitive attributes [2,3].

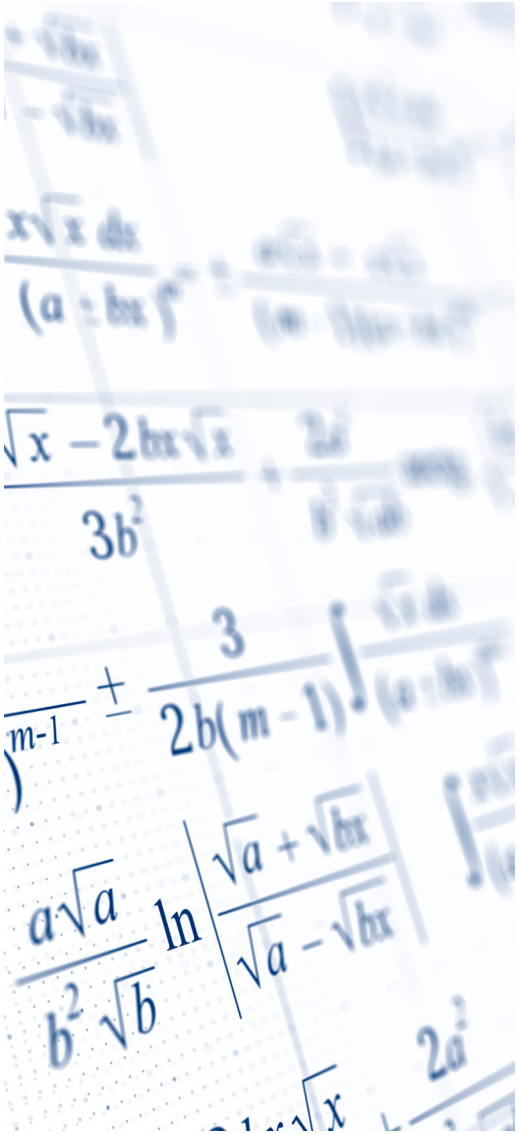
• **Identifiers:** an identifier (sometimes referred to as a "direct identifier") is an attribute that refers to a particular individual in the given population. Examples of identifiers include an e-mail address or a credit card number, both of which can uniquely identify an individual.

• **Quasi-Identifiers:** quasi-identifiers themselves do not uniquely identify individuals; however, when they are combined with other attributes, they can enable the identification of individuals. For example, personal attributes such as age, gender, date of birth; or financial transaction attributes such as time, date and location, can uniquely identify individuals when combined.

• **Sensitive Attributes:** some identifying attributes can be classified as sensitive attributes when they represent person-specific information, such as an individual's salary or health-related insights.



A dataset is considered to be k-anonymous if the identifying information about each individual is indistinguishable from at least k-1 other individuals in that dataset. This way, k-anonymity provides a degree of anonymity.



Applying Techniques

Consider Table 1: the '*Name*' attribute is the identifier, the '*Profession*' and '*Age*' attributes are the quasi-identifiers, and the '*Test Score*' is the sensitive attribute that we are interested in for analytics purposes.

Name (Identifier)	Profession (Quasi-Identifier)	Age (Quasi-Identifier)	Test Score (Sensitive Attribute)
Sara	Database Administrator	36	98
Rani	Computer Network Architect	56	74
Patricia	Web Developer	40	53
Cian	Web Developer	52	24
Raymond	Occupational Therapist	35	56
Abram	Dentist	26	73
Elodie	Epidemiologist	25	43
Antonio	Dentist	32	87

Table 1: An example table containing information about eight individuals.

Now, let us consider a scenario in which we want to establish a 4-anonymous version of Table 1; that is to achieve a *k*-anonymity of *k* = 4. This would mean that each individual featured in the table cannot be distinguished from 3 other individuals with respect to a set of quasi-identifier attributes, such as '*Profession*' and '*Age*'.

Name (Identifier)	Profession (Quasi-Identifier)	Age (Quasi-Identifier)	Test Score (Sensitive Attribute)
-	Information Technology	[36 - 58]	98
-	Information Technology	[36 - 58]	74
-	Information Technology	[36 - 58]	53
-	Information Technology	[36 - 58]	24
-	Health Care	[20 - 35]	56
-	Health Care	[20 - 35]	73
-	Health Care	[20 - 35]	43
-	Health Care	[20 - 35]	87

Table 2: A 4-anonymous version of Table 1.

Table 2 shows this 4-anonymous version of Table 1. In order to achieve this 4-anonymous version of the table, two techniques were employed: *generalization* and *suppression*.

Suppression is the removal of the attribute values; that is to say that the identifiers are removed.

Generalization, on the other hand, is the process of replacing a value with a less specific, but semantically consistent, value. That is to say, generalization involves a deliberate reduction in the precision of data.

An example of this would be transforming someone's age into an age range, or converting a precise location (that has a specific longitude and latitude) into a less precise location. For instance, *Paris* could be replaced with *France* to be less specific.

Generalization operations hide some details in quasi-identifiers. As an example, for a categorical attribute, a specific value can be replaced with a general value according to a given taxonomy; this is also known as the generalization hierarchy. The process of generalization allows an attribute's value to be replaced with its parent value in the hierarchy.

One can have various 4-anonymous versions of Table 1; therefore, a question that arises is which version of the *k*-anonymous table should you move forward with. Ideally, the right version would be the one with the highest utility.

It is worth mentioning that researchers have identified that *k*-anonymity is susceptible to attacks like background knowledge attack and homogeneity attack.

The later privacy models, such as *l*-diversity and *t*-closeness, overcome the limitations of *k*-anonymity, but each are also prone to different kinds of attack.

Generalization Hierarchy

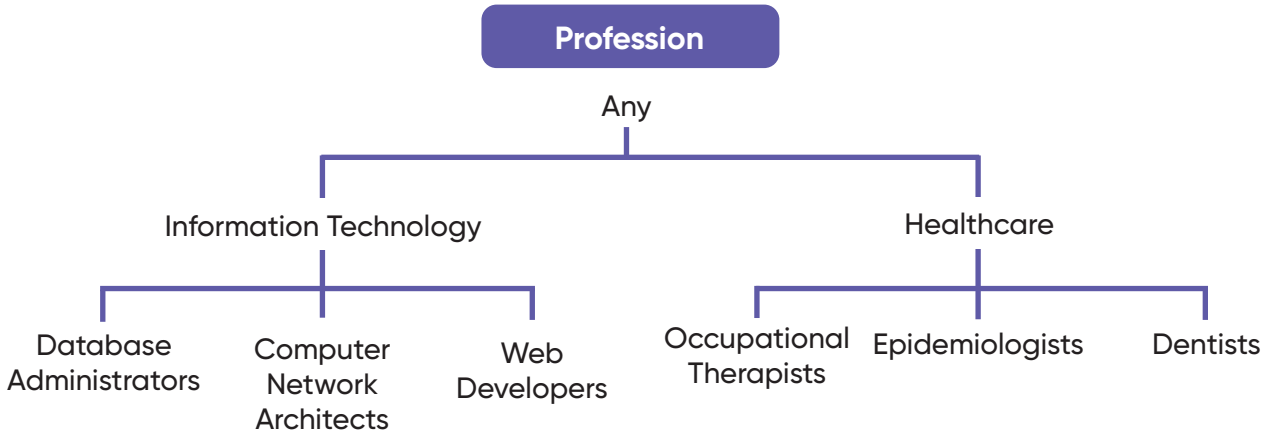


Figure 1: Generalization hierarchy for the attribute 'Profession' (quasi-identifier).

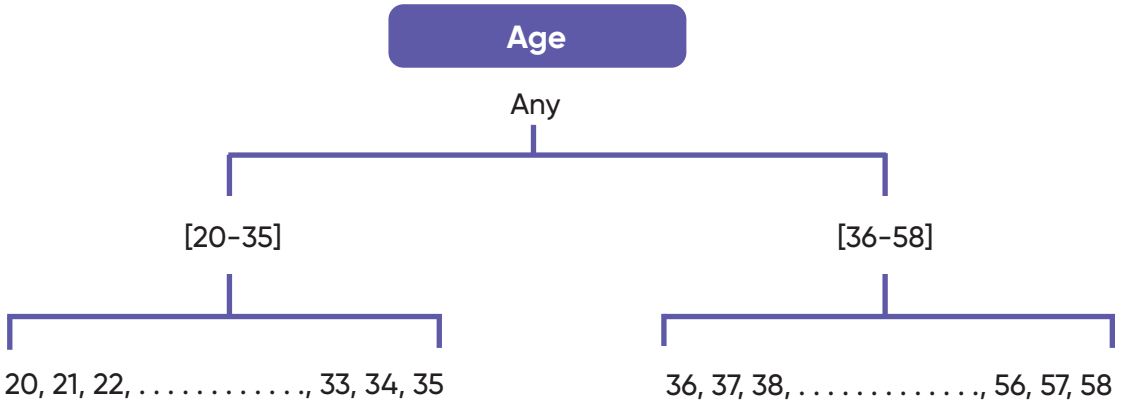
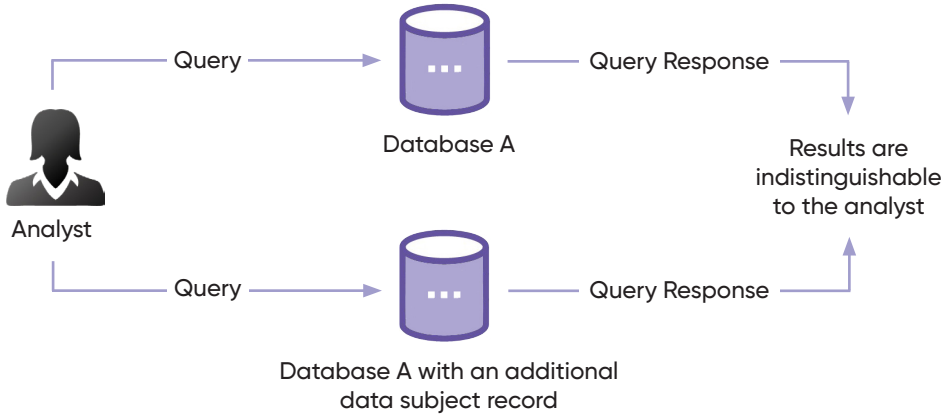


Figure 2: Generalization hierarchy for the attribute 'Age' (quasi-identifier).

Differential Privacy

Another popular formal privacy model is differential privacy, which was designed to allow analysts to query a database interactively and ensure that a query response is insensitive to any specific record in the database.

Consider two datasets that only differ by one record, then suppose an analyst sends two queries to these two datasets. In this case, the differential privacy model says that the query responses that the analyst will see must be indistinguishable from each other.



Similar to *k*-anonymity, differential privacy is a characteristic of a dataset rather than being a technique. A commonly used technique to achieve differential privacy is the addition of 'noise' to the query responses. There exists a number of differentially private algorithms that act as a layer between the analyst and the database [5,6,7].

To understand differential privacy at a very abstract level, consider the following scenario:

- A company has 50 employees; 49 of those employees are from Spain and 1 is from New Zealand.
- The company wants to arrange lunch for all the employees.
- 15 of the 50 employees are vegetarian, and the person from New Zealand happens to be one of them.
- While placing the order with the caterer, the company provides the number of vegetarians.
- Since only the total number of employees and the total number of vegetarians is provided to the caterer, and not their actual identities (name, nationality), the privacy of the vegetarian employees is preserved.
- If the caterer wanted to know what number of employees are from Spain and vegetarian, so that he can prepare Spanish dishes, he could send a query to the company.
- As a response to the caterer's query, the company could share that the 14 individuals are from Spain and are also vegetarian.
- Once this number is shared, the caterer can infer that out of the 15 vegetarians, there are 14 from Spain; therefore, the individual from New Zealand must also be vegetarian.
- At this stage, individual privacy is compromised.

In order to prevent such inferences or privacy violations, differential privacy can be used to make the query responses differentially private. This is typically achieved, as previously mentioned, by adding 'noise' to these numbers. Going back to our example in the company-caterer scenario, that means that instead of the company providing an exact number, some 'noise' can be added to it. For example, by adding 'noise' of 3 to the total number of vegetarian employees, the number becomes 18. As a result, the caterer won't be able to infer the level of personal information they hold with full certainty.

Using this example, one may ask '*what about the extra vegetarian food as a result of differential privacy?*' The extra food is the utility loss – which is the fundamental privacy-utility trade-off.

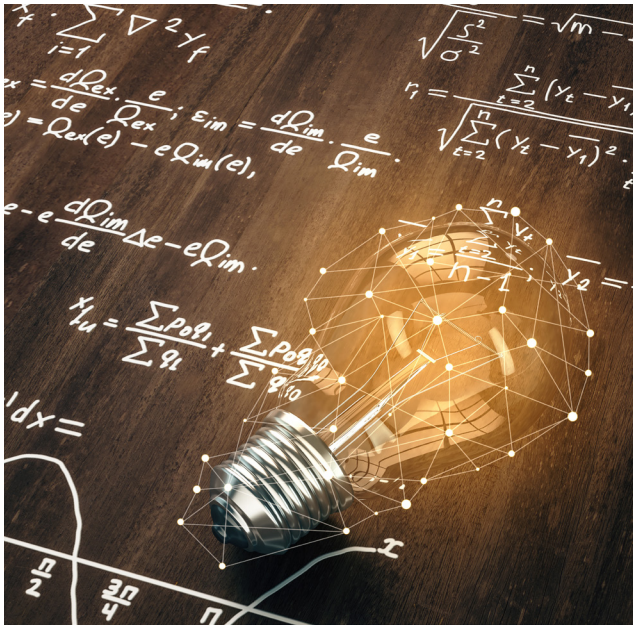
It is important to note that the Differential Privacy model has a parameter known as '*epsilon*', which governs the amount of added noise. Larger values of epsilon lead to better analytical utility and less privacy, whereas smaller values of epsilon lead to better privacy but less analytical utility.

The numbers of queries an analyst can make is regulated by the value of the epsilon, also referred to as 'privacy budget'. So, when it comes to the selection of the appropriate value of epsilon, it is again a question of striking the right balance between analytical utility and privacy.

Handle with care

While the aforementioned privacy models enable the reasoning of privacy in a formal manner that adopts a mathematical approach, these models do come with limitations. As such, it is important to consider the analytical use case along with the required tuning (for example, epsilon with differential privacy) when applying such models.

The key message is that both privacy models need to be handled with care.



Discover more at truata.com. You can also contact us directly to book a demo session or to understand how our privacy-enhanced data solutions and services can work for you.

References

[1] T. Dalenius, "Towards a methodology for statistical disclosure control," Statistik Tidskrift, vol. 15, no. 429-444, pp. 2–1, 1977.

[2] L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557–570, 2002.

[3] L. Sweeney, "Simple demographics often identify people uniquely," working paper, 2000. Working paper.

[4] C. Dwork, "Differential privacy: A survey of results," in Theory and Applications of Models of Computation (M. Agrawal, D. Du, Z. Duan, and A. Li, eds.), (Berlin, Heidelberg), pp. 1–19, Springer Berlin Heidelberg, 2008.

[5] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," Found. Trends Theor. Comput. Sci., vol. 9, pp. 211–407, Aug. 2014.

[6] C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor, "Optimizing linear counting queries under differential privacy," in Proceedings of the Twenty-ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS '10, (New York, NY, USA), pp. 123–134, ACM, 2010.

[7] J. He and L. Cai, "Differential private noise adding mechanism: Fundamental theory and its application," CoRR, vol. abs/1611.08936, 2016.

[8] Isabel Wagner and David Eckhoff. 2018. Technical Privacy Metrics: A Systematic Survey. ACM Comput. Surv. 51, 3, Article 57 (July 2018), 38 pages. DOI: https://doi.org/10.1145/3168389

Additional Resources

White Paper: Mindful Data Mining

Article: What They Don't Tell You About Pseudonymization

Webinar: If You Can't Measure It, You Can't Protect It