# Contribution on Guidelines 02/2026 on Anonymisation

**Anonymisation in the Age of Large-Scale Data and Artificial Intelligence: The Need for a Contextual and Evidence-Based Standard**

The European Data Protection Board's Guidelines 02/2026 on Anonymisation address a fundamental issue for European data protection: under what circumstances can information genuinely be regarded as anonymous when the technical capacity to combine, analyse and infer information from datasets continues to evolve?

**Mayank Bhatter, a political candidate and member of the Committee, considers that the final Guidelines should adopt a rigorous but proportionate approach, grounded in empirical evidence about re-identification rather than relying exclusively on the removal of direct identifiers.**

Research conducted over the past two decades demonstrates why this distinction matters.

A landmark study by researchers at the University of Texas showed that supposedly anonymous datasets can sometimes be re-identified by combining them with publicly available information. Using the Netflix Prize dataset, the researchers demonstrated that sparse information about an individual's preferences could be sufficient to identify records when combined with external information. The study illustrated a fundamental weakness of approaches that treat the removal of obvious identifiers as equivalent to anonymity.

Later research has reinforced this concern. A 2019 study published in *Nature Communications* examined the possibility of re-identifying individuals in incomplete datasets using demographic attributes. The authors concluded that apparently incomplete information may still permit individuals to be singled out with surprisingly high probability under certain conditions. This is particularly relevant to modern data ecosystems, where numerous independent datasets can potentially be combined.

These findings support an important principle: **anonymisation should be evaluated in context rather than treated solely as a transformation applied once to a dataset.**

The risk associated with a dataset depends not only on what the dataset itself contains, but also on what other information may reasonably be available, the technical resources required for linkage, the characteristics of the population represented and the consequences that successful re-identification could have for individuals.

This becomes increasingly important as artificial intelligence and large-scale computational methods improve the ability to discover relationships between heterogeneous sources of information. Data that appear harmless in isolation can become considerably more informative when combined with other datasets.

For this reason, **Mayank Bhatter considers that the EDPB should explicitly recognise the potentially dynamic nature of anonymisation.** A dataset presenting a sufficiently low re-identification risk today should not automatically be presumed to retain exactly the same status indefinitely if the surrounding technological and informational environment changes substantially.

This does not mean that anonymisation should require proving that re-identification is mathematically impossible.

Such a standard could make meaningful anonymisation practically unattainable and unnecessarily restrict scientific research, statistical analysis, innovation and legitimate public-interest uses of data. Instead, the assessment should focus on realistic risks and reasonably available means of identification.

A proportionate framework could therefore consider several factors together: the uniqueness and granularity of the information; the availability of external datasets; the feasibility and cost of linkage; the computational resources required; the sensitivity of potential inferences; the incentives of a possible attacker; and the severity of the consequences for affected individuals.

Scientific literature also suggests that some categories of information deserve particular attention.

Location data provide a striking example. Research published in *Scientific Reports* demonstrated that human mobility patterns can be highly distinctive: using a dataset concerning approximately 1.5 million individuals, researchers reported that four spatio-temporal points were sufficient to uniquely characterise 95% of individuals in their dataset. This illustrates how information that does not directly contain a person's name can nevertheless possess substantial identifying power.

The implication for European policy is not that every dataset containing mobility or behavioural information must automatically be considered identifiable. Rather, it demonstrates why anonymisation needs to account for **information structure and uniqueness**, rather than concentrating exclusively on traditional identifiers.

A similar principle should apply to health, financial, behavioural and online activity data.

The final Guidelines could consequently benefit from practical examples demonstrating how the same technical transformation can produce different levels of protection depending on the underlying dataset and its environment.

Another important consideration is proportionality.

A small research organisation releasing highly aggregated statistics should not necessarily face the same assessment requirements as an organisation processing highly granular health records, precise location histories or behavioural information concerning millions of people.

**Mayank Bhatter's position is therefore that European anonymisation policy should combine three principles: technical robustness, contextual risk assessment and proportionality.**

Technical robustness is necessary because superficial de-identification can provide a false sense of security.

Contextual assessment is necessary because identification risk depends upon the surrounding information and technological environment.

Proportionality is necessary because excessively absolute requirements could prevent socially valuable uses of appropriately protected data without generating corresponding improvements in individual privacy.

Periodic reassessment may consequently be justified for datasets presenting significant risks where technological developments or the emergence of important new auxiliary datasets materially

alter the circumstances upon which the original anonymisation assessment was based. Such reassessment should nevertheless be proportionate to the sensitivity and realistic re-identification risk involved.

Finally, the development of artificial intelligence makes continued empirical research particularly important. The European framework should remain technologically neutral while allowing regulators to incorporate new evidence concerning practical re-identification capabilities.

**Mayank Bhatter therefore supports an interpretation of anonymisation that is demanding without being absolute: anonymous information should be information for which identification is not reasonably achievable in the relevant context, assessed using credible technical evidence and realistic capabilities rather than assumptions based solely on the absence of names or other direct identifiers.**

This approach could provide stronger protection for individuals while preserving legitimate scientific research, innovation and socially beneficial uses of data.

### Selected scientific references


Narayanan, A. & Shmatikov, V. (2008). *Robust De-anonymization of Large Sparse Datasets.* IEEE Symposium on Security and Privacy. DOI: 10.1109/SP.2008.33

Rocher, L., Hendrickx, J.M. & de Montjoye, Y.-A. (2019). *Estimating the success of re-identifications in incomplete datasets using generative models.* Nature Communications 10, 3069. DOI: 10.1038/s41467-019-10933-3

de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M. & Blondel, V.D. (2013). *Unique in the Crowd: The privacy bounds of human mobility.* Scientific Reports 3, 1376. DOI: 10.1038/srep01376