TRUATA

Is data anonymization nothing more than a simple game of *Guess Who?*

TRŪATA.

Is data anonymization nothing more than a simple game of Guess Who? The answer is no. But by the way that some are claiming to 'anonymize' data, you'd certainly think so...

A number of recent studies and reports have highlighted that it is relatively easy to re-identify a person from supposedly anonymized datasets. These studies highlight the pitfalls of simply removing or obfuscating direct identifiers and how heavily sampled datasets are highlight unlikely to satisfy the modern standards for anonymization as laid out by the GDPR.

More than this, however, such studies expose an uglier truth – there are many people out there misusing the term 'anonymization' when outlining how they address privacy concerns when working with personal data.

Dr. Imran Khan
Data Scientist at Trūata

- One study found that **99.98%** of individuals could be correctly re-identified in any dataset using 15 demographic attributes[1]
- A 3-month study of the credit card records for 1.1 million people shows that just 4 spatiotemporal points are enough to **uniquely re-identify 90%** of individuals[2]
- Based on only 3 parameters (gender, date of birth and a 5-digit ZIP), a study found that it was possible **uniquely identify 87%** of the entire US population[3]

Guess who?

Hmm. You got it right...again.

A very simple way to understand the issue of misusing the term anonymization is to think of the game Guess Who? This game highlights just how easy it is to identify an individual when you only have access to a few attributes about the person.

Many of you will remember this popular two-player game, but here is a recap for those who don't. Aim of the game: to guess the identity of your opponent's chosen character.

How it works:

- Each player starts the game with a board that contains cartoon images of 24 people. You and your opponent select your character for the game.
- You ask your opponent 'yes/no' questions to learn some of the attributes that link to your opponent's identity. For example, you might ask: Does the person wear a hat? or Do they have blue eyes?
- You continue to ask questions to eliminate possibilities and narrow down the possible identity of the character your opponent has chosen. The more you know about them, the easier it gets to figure out who your opponent's character is.
- One of you will always win – it's just a matter of how many questions it takes to unmask their character!

[1] Estimating the success of re-identifications in incomplete datasets using generative models
[2] Unique in the shopping mall: On the reidentifiability of credit card metadata
[3] Simple Demographics Often Identify People Uniquely

What does the game teach us about anonymization?

The Guess Who game illustrates that datasets with obfuscated direct identifiers (e.g. obfuscated names) are still full of risk. In fact, re-identifiability is almost certain for a dataset of any size and complexity.

Even in cases where certain attributes are 'anonymized' to make them less identifiable (such as generalising names to genders or broadening exact hair colours to light/dark), there is still re-identification risk present if there are enough personal attributes that can be combined together.

Let's play a game of Guess Who to demonstrate this in action.

"Are you wearing a hat?"

"No"

This leaves 19 people.

"Do you have blonde hair?"

"No"

This leaves 15 people.

"Do you have glasses?"

"Yes"

Down to 3 people.

"Do you have white hair?"

"No"

Congratulations!
You've identified me with 4 pieces of information. I'm Tom.

How does this re-identification exercise work with your data?

Imagine how easy it is to identify an individual if you have a database containing tens or even hundreds of fields. If you can ask an unlimited number of questions, you can go through the sequence exercise (just as you would in the game of Guess Who) and re-identify any individual.

If some personal attributes are tweaked, it would become slightly more difficult to re-identify individuals. However, you can simply ask more questions to get to the answer.

If all personal attributes were completely obfuscated or jumbled, it might become impossible to re-identify anyone. However, that would ruin the Guess Who game or – in analytical terms – it would mean that the utility of your data has been destroyed.

Similarly, if you have a set of descriptions like those shown in the table, you could match each row to someone on the board. The more personal attributes you have (i.e. the more columns of information in your dataset), the easier it becomes to match that information with a specific individual.

Gender	Hat	Hair Color	Hairstyle	Glasses
Woman	Yes	Red	Short	Yes
Man	No	Black	Balding	Yes
Man	No	Black	Short	No
Man	No	Red	Long	No

Why anonymize data then?

Based on the above, you might be wondering how valuable anonymization is at all. But not all efforts to anonymize are equally effective. Or, more accurately, not all types of "anonymization" are actually anonymization in the first place.

As more recent studies and examples have showcased, many datasets are being released and presented as anonymized when, in fact, they allow for re-identification with relative ease.

True anonymization may analyse granular data, but it only returns aggregate statistics or calculations. These aggregate responses do not include identifiers and require that responses meet certain thresholds (e.g. a minimum number of individuals represented; each individual contributing no more than x% to the results, etc.).

How do you apply true anonymization to the Guess Who example?

To make sense of true anonymization, let's return back to the Guess Who game.

If you were playing the game, you would simply refuse to answer certain questions (e.g. does the person wear glasses?) that leave less than a minimum threshold of people.

Doing this in the game may not make your eight-year-old opponent happy, but it does prevent individuals from being easily identified with the data that is available.

Board games aside, in the real world, achieving true anonymization is not an easy process. Striking the balance between data privacy and data utility requires a sophisticated approach that requires innovative technological, structural, legal and organizational safeguards. And this needs to be applied by true experts – in both data science and privacy.

While there is increasing hype around privacy-enhancing technologies and techniques and businesses are looking towards these solutions to mitigate risks and overcome data-driven challenges, it is important to be aware that not all solutions come with a guarantee of true GDPR-grade anonymization. Then again, nobody ever said that applying anonymization was child's play.

Game over.

Learn more

Trūata Anonymization Service (TAS)

Trūata Anonymization Service (TAS) is an independent data anonymization service that enables organizations to preserve the analytical value of data while protecting consumer privacy to the highest regulatory standards. This is a global solution that allows for consistency across organizations and jurisdictions.

Evolve and innovate with a privacy-first mindset

Reach out to us today to learn more about Trūata and what we can do for your business.

Arrange a free demo session
book a demo

Speak to our team
info@truata.com

Follow Trūata for more

Explore our solutions at **truata.com**

Offices

Ireland
Silverstone House, 1st Floor
Ballymoss Road
Sandyford, Dublin 18
D18 A7K7

UK
Oriel House, 26
The Quadrant
London
TW9 1DL

USA
411 Theodore Fremd Avenue Suite
206 South, Rye
New York
10580