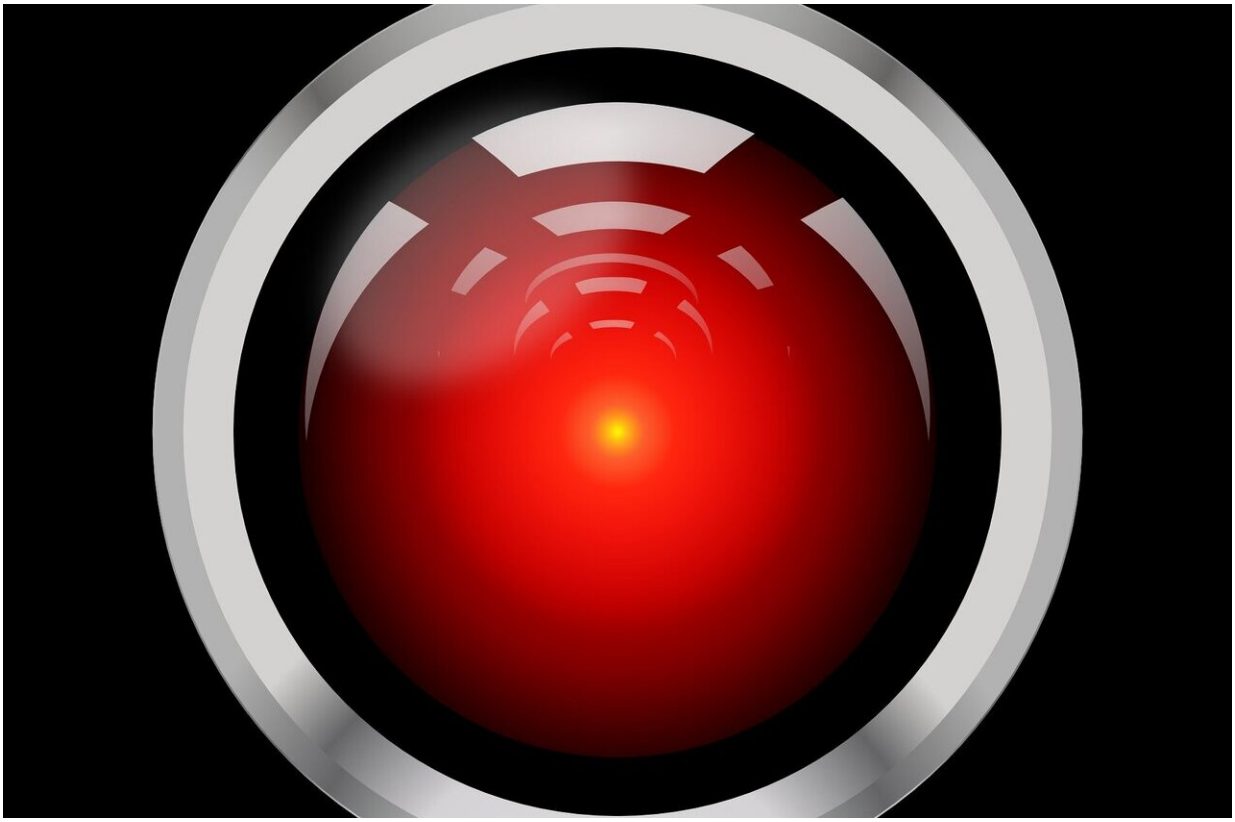
Improved method for generating synthetic data solves major privacy issues in research

June 7 2021



Credit: Pixabay/CC0 Public Domain

A lack of data is a major bottleneck for many kinds of research, and especially for the development of better medical treatments and drugs. This data is extremely sensitive and, understandably, people and

companies alike are often unwilling to share their information with others.

Researchers at the Finnish Center for Artificial Intelligence have developed a machine learning-based method that produces synthetic data on the basis of original data sets, making it possible for researchers to share their data with one other. This could solve the ongoing problem of data scarcity in medical research and other fields where information is sensitive.

The generated data preserves privacy, remaining similar enough to the original data to be used for statistical analyses. With the new method, researchers can conduct an infinite number of analyses without compromising the identities of the individuals involved in the original experiment.

"What we do is we tweak the original data sufficiently so that we can mathematically guarantee that no individual can be recognized," explains Samuel Kaski, Aalto University Professor and Director of FCAI, who co-authored the study.

Researchers have produced and used synthetic data before, but the new study solves a major problem with existing methods.

"We might think that just because data is synthetic, it's safe. This has not necessarily been the case, though," explains Kaski.

This is because synthetic data needs to be very similar to the original data set in order to be useful in research. In practice, it has occasionally been possible to identify individuals' identities despite anonymization.

To address this problem, FCAI researchers make use of artificial intelligence, specifically probabilistic modeling. This enables them to

use prior knowledge about the original data and the processes that have made it the way it is—without getting too close to the properties of the particular data set used as basis for the synthetic data. Such prior knowledge, for instance, could relate to known gender differences in alcohol-related mortalities, or could involve domain knowledge about how a particular data set has been collected.

Making use of prior knowledge has also made the synthetic data sets more useful for making correct statistical discoveries—even in cases where the original data set is limited in size, which is common in medical research.

"Incorporating prior knowledge means we can use the method with small data sets, for which we have domain knowledge," Kaski says.

The results are published 7 June in the journal *Patterns*.

More information: Joonas Jälkö et al, Privacy-preserving data sharing via probabilistic modeling, *Patterns* (2021). DOI: 10.1016/j.patter.2021.100271

Provided by Aalto University