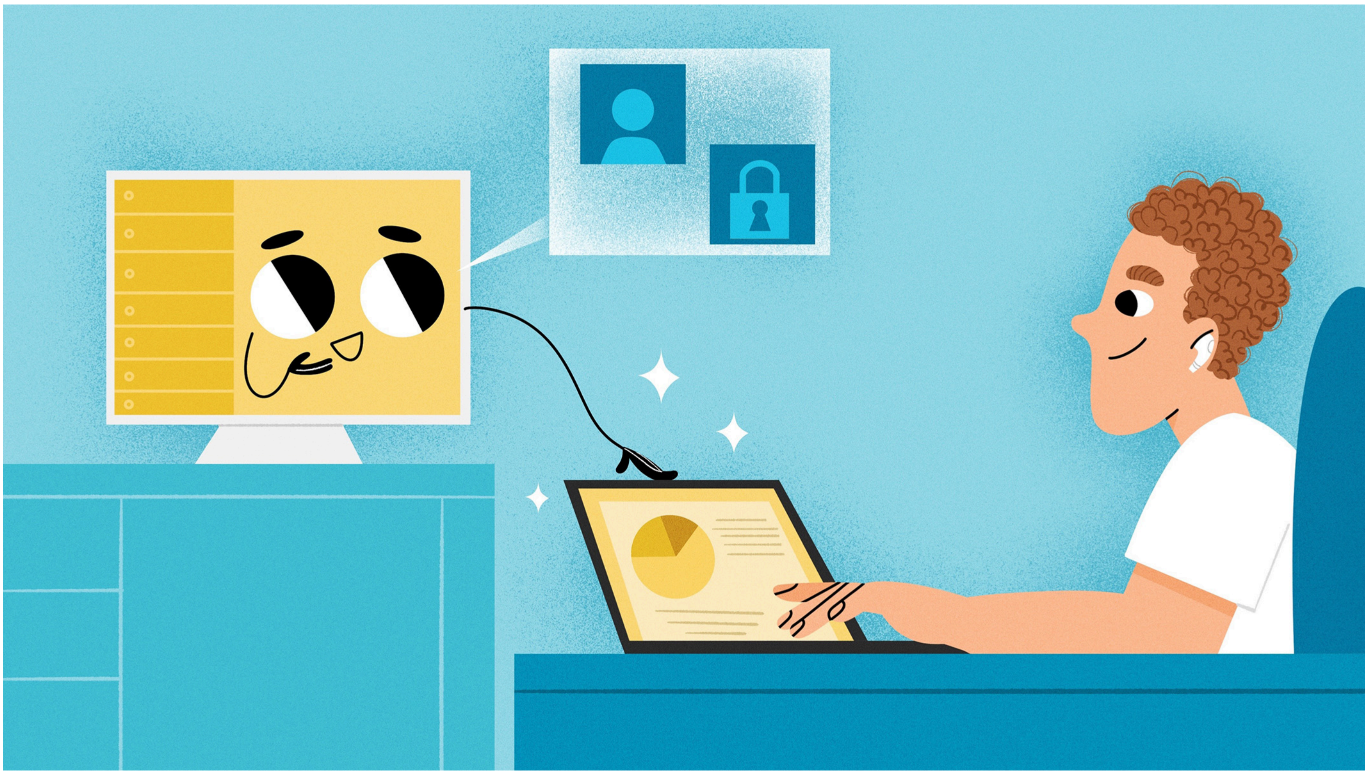
ENGINEERING AND TECHNOLOGY
Published: 04 June 2025
doi: 10.3389/frym.2025.1575140



SAFER WAYS FOR RESEARCHERS TO SHARE SENSITIVE DATA

Olga Fink[1], *Lisette van Gemert-Pijnen*[2*], *Dongwon Lee*[3], *Andrew Maynard*[4] and *Bastiaan van Schijndel*[5]

[1] *Intelligent Maintenance and Operations Systems, Swiss Federal Institute of Technology in Lausanne, Lausanne, Switzerland*
[2] *Persuasive Health Technology, University of Twente, Enschede, Netherlands*
[3] *College of Information Sciences and Technology, Pennsylvania State University, University Park, PA, United States*
[4] *School for the Future of Innovation in Society, Arizona State University, Tempe, AZ, United States*
[5] *ZorgTTP Foundation, Houten, Netherlands*

YOUNG REVIEWERS:

ANSAR
AGE: 14

Sharing and analyzing data are essential for solving complicated problems, like curing diseases or protecting the environment. However, sensitive data, such as medical records or financial details, must be kept private and secure. New technologies make sharing and using sensitive data safer by creating realistic versions of data that do not contain private details or sensitive information traceable to a person. These techniques, called synthetic data and encryption, are already helping researchers study diseases, detect fraud, and prepare for rare events like natural disasters. While challenges remain, such as improving the accuracy of synthetic data and reducing the energy needed to create it, these techniques could unlock safer, faster ways

to share data so that researchers all over the world can collaborate more easily.

THE WORLD NEEDS SAFER WAYS TO SHARE DATA

Every day, we are surrounded by data. Data includes facts and information, like numbers, measurements, and images, that help us learn about the world. From medical records to satellite images, data helps us to understand things happening in the world around us, solve problems, and make discoveries. For example, scientists use data to study how diseases spread, predict natural disasters, and improve technologies like renewable energy. **Artificial intelligence** (AI) is making these discoveries faster and more efficient by analyzing vast amounts of data and uncovering patterns humans might miss (read more about AI and its role in scientific discovery here).

ARTIFICIAL INTELLIGENCE

A type of computer technology that helps machines think, learn, and solve problems, like recognizing faces, predicting weather, or designing new medicines.

For AI technologies to work, they need access to lots of data—but there are big problems with sharing certain types of data. Some information, like how many sunny days a city gets each year, can be freely shared without any problems. Other data is considered sensitive because it contains personal details. **Sensitive data** needs special care to keep it safe, which involves both **privacy** and **security**. Privacy means protecting personal information, so it is not shared with others without permission. For example, a person's medical history should stay private. Security, on the other hand, is about excluding people who should not have access to the data from seeing or using it. Data security means keeping sensitive information, like government records or business secrets, safe from hackers or other threats (for more info on data privacy and security, see this Frontiers for Young Minds article).

SENSITIVE DATA

Information that needs to be protected, like medical records, financial details, or personal information, because sharing it could cause harm or violate privacy.

PRIVACY

Keeping personal information, like your medical history, safe and hidden from others without your permission.

SECURITY

Protecting data from being stolen or misused, such as keeping bank details safe from hackers.

Sensitive data is not limited to health records—it also includes financial information, like bank account details, which could be used for fraud, and personal details, like names and addresses, which could lead to identity theft. Even data about how people shop online or use social media—like what they buy or the videos they watch—can be sensitive because it might be misused to manipulate their choices. Governments and scientists also work with sensitive information, such as maps of restricted areas or habitats of endangered animals, which must be kept safe to protect national security or the environment.

Another challenge is **data sovereignty**, which means making sure that data stays under the control of its rightful owner, whether that is a person, a company, or a country. For example, privacy laws in one country might prevent health data from being shared with researchers in another country, even if the research could save lives. These rules are important to protect people and organizations, but they can make

DATA SOVEREIGNTY

Making sure data stays under the control of its owner, like a person, company, or country, even when it is shared across borders.

it harder for scientists to work together on solving complex problems. Could there be a way to share sensitive data or train AI systems without risking privacy, security, or sovereignty?

EMERGING TECHNOLOGY: SYNTHETIC DATA

Synthetic data and other privacy-enhancing technologies allow researchers to safely share information, paving the way for new discoveries while keeping sensitive data safe [1]. These technologies could transform how scientists and companies all over the world work together, helping them to tackle some of our biggest challenges.

SYNTHETIC DATA

Data created by computers to mimic real data, so it can be safely used for research without exposing private details.

To understand synthetic data, think of it as a realistic "copy" of real data—it is not the same as the original, but it mimics its patterns and trends. For example, imagine a set of data from hospital records showing how patients recover from an illness. Synthetic data would reflect actual recovery patterns, like the average time it takes patients to heal, but would not include any of the patient's personal data. So, synthetic data is safe for researchers to analyze without risking anyone's privacy. Synthetic data is created using a type of AI called **generative adversarial networks (GANs)**. GANs are like digital artists—they learn from real data and use that knowledge to create new, artificial data that looks realistic. For example, a GAN trained on images of faces can create entirely new, lifelike faces that do not belong to any real person. GANs operate in two steps. First, the real-world data is used to "teach" the AI system within the GAN about the data. After this training process, the AI system works with the data-synthesis system to create synthetic data similar to the real-world data. These synthetic datasets can then be used without concerns that sensitive data will be exposed.

GENERATIVE ADVERSARIAL NETWORKS (GANs)

A type of AI that uses two computer systems—one learns from real-world data and the other generates synthetic data that matches the patterns in the real-world data.

HOMOMORPHIC ENCRYPTION

A way to protect data by turning it into a secret code that can still be used for analysis without revealing the actual information.

Another key technology is called **homomorphic encryption** [2]. This technique does not create a new version of the data. Instead, it changes the original data into a kind of "secret code" that can still be analyzed without revealing the original information. The main difference between synthetic data and homomorphic encryption is how they handle the original data. Synthetic data replaces the original information with a completely new, artificial dataset that follows the same patterns. Homomorphic encryption keeps the original data but "locks it up" so only the people who have the key can access it directly, while still allowing it to be used in calculations. Think of synthetic data as a realistic copy of a museum exhibit—it looks the same but is not the original. Homomorphic encryption is like locking the real artifact inside a box that allows you to perform specific actions on it—like calculating its weight—without ever opening the box or revealing the artifact inside.

TECH TO THE RESCUE

Synthetic data and homomorphic encryption are already helping scientists and companies tackle sensitive data problems in new and creative ways (Figure 1).

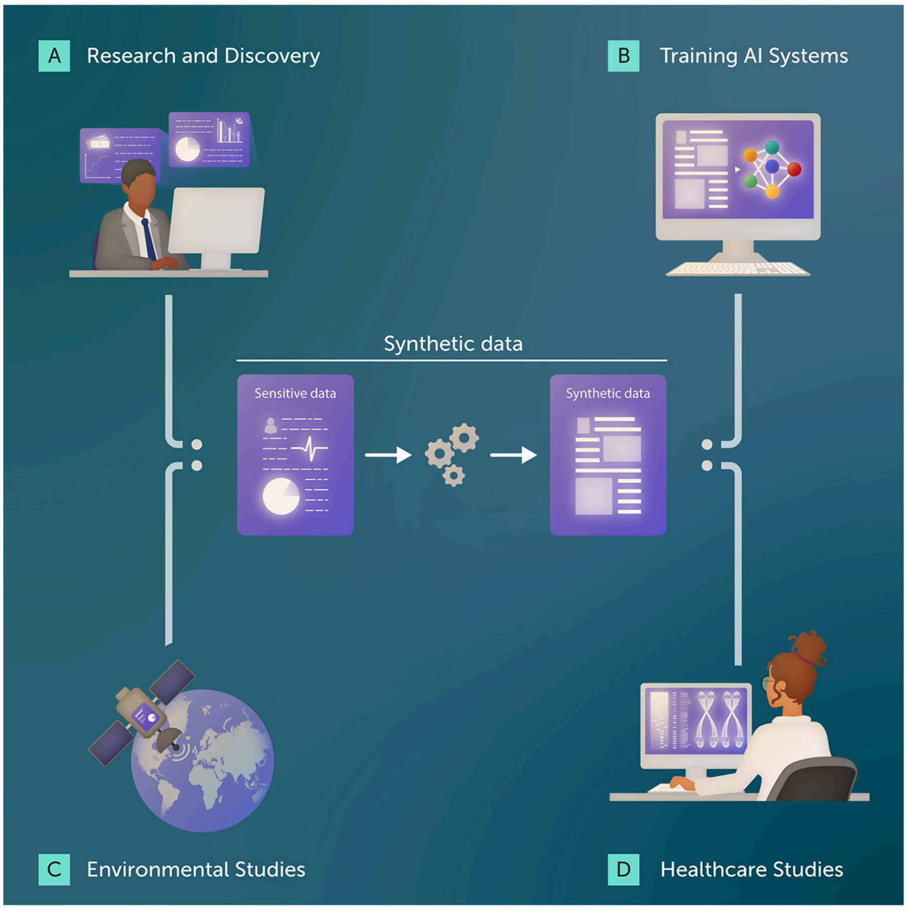


Figure 1

Figure 1

Synthetic data helps researchers use data to make important discoveries, while keeping sensitive information protected. **(A)** Bankers and economists can use synthetic data to study questions like how a major stock market crash might affect investors. **(B)** Synthetic data is being used to train AI systems, which need lots of information to learn. **(C)** Encrypted satellite data can track environmental changes while protecting national security. **(D)** In healthcare, doctors and scientists can use synthetic data to study diseases and treatments while keeping patients' private information safe.

In healthcare, where privacy rules often restrict researchers from using real medical records, synthetic data allows researchers to study diseases and treatments while keeping patients' private information safe [3]. For instance, scientists could create a synthetic dataset that mirrors the patterns in real patient records, such as how different groups of patients respond to a certain medication. By studying these patterns, researchers might discover that the treatment works better for people with specific genes or health conditions. This could lead to more personalized and effective treatments, all without ever exposing real patient details. Synthetic data is also useful when data is scarce, for rare diseases, for example.

Synthetic data can be used to train AI systems that need lots of information to learn, without the need to use private or sensitive data. Companies can use AI to spot trends, make predictions, and improve decisions. With synthetic data, AI can be trained to detect unusual spending patterns that might signal fraud, or to predict when customers might want a certain product. For instance, an AI system trained on synthetic shopping data could find that people who buy one product, like hiking backpacks, often need a related one, like water bottles. This helps companies make better suggestions without risking real consumer information. Finally, synthetic data can be a less expensive option for researchers, because data is often scarce in healthcare and collecting real data takes time and effort.

Homomorphic encryption is useful when real data must stay protected but still needs to be analyzed. For example, governments can use encrypted satellite images to track environmental changes, like deforestation or melting glaciers. These images might include sensitive details, such as the exact locations of natural resources or areas critical for national security, so encryption lets researchers study patterns—like how fast forests are shrinking or how rising temperatures affect ice—without risking misuse of the data. Similarly, encrypted health data allows scientists to study global trends in diseases, helping them identify hotspots and predict outbreaks, all while keeping personal information private.

Finally, these technologies can help researchers prepare for rare events, like economic crises or natural disasters. Synthetic data can be created to mimic these unusual situations, giving researchers a safe way to test their ideas before real events happen. For example, banks could use synthetic data to study how a sudden stock market crash might affect savings and create better plans to protect their customers.

BIG CHALLENGES, BIGGER OPPORTUNITIES

Synthetic data and homomorphic encryption are making it easier and safer to share and analyze information. These tools help researchers learn more from data while keeping sensitive information secure, paving the way for breakthroughs that were previously out of reach.

However, these technologies still have some challenges to overcome. Synthetic data is not always perfect. If the original data are not accurate, the synthetic data based on that real data will not be accurate either—"garbage in, garbage out", as data scientists say. Also, if synthetic data oversimplifies or misinterprets real-world data, it can also be inaccurate. Homomorphic encryption is very secure but can take a long time and creating it uses a lot of energy, making it hard to use for large projects. There is also the risk that encrypted or synthetic data could be cracked by hackers, revealing private information. Another challenge is trust. For these techniques to succeed, people need to understand them and believe they are safe. Scientists, doctors, and government leaders must work together to create clear rules and educate the public about how these technologies work.

Despite these challenges, synthetic data and other secure data-sharing tools have huge potential to help solve big problems, like fighting diseases or climate change, while protecting sensitive data. By making data sharing safer, these tools could lead to a future where scientists and organizations around the world can collaborate more easily.

ACKNOWLEDGMENTS

Co-written and edited by Susan Debad Ph.D., graduate of the UMass Chan Medical School Morningside Graduate School of Biomedical Sciences (USA) and scientific writer/editor at SJD Consulting, LLC. Figure created by Somersault18:24.

ORIGINAL SOURCE ARTICLE

Fink, O., van Gemerty-Pijnen, L., Lee, D., Maynard, A, and van Schijndel, B. 2024. "Privacy-enhancing Technologies. Empowering global collaboration at scale," in *Top 10 Emerging Technologies of 2024 Flagship Report*. Cologny: World Economic Forum. Available online at: https://www.weforum.org/publications/top-10-emerging-technologies-2024/ (accessed May 7, 2025).

REFERENCES

1. Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., et al. 2022. Synthetic Data – what, why and how? *arXiv* [preprint] arXiv:2205.03257. doi: 10.48550/arXiv.2205.03257
2. Rivest, R. L., and Dertouzos, M. L. 1978. *On Data Banks and Privacy Homomorphisms*. Available online at: https://luca-giuzzi.unibs.it/corsi/Support/papers-cryptography/RAD78.pdf (accessed May 7, 2025).
3. Gonzales, A., Guruswamy, G., and Smith, S. R. 2023. Synthetic data in health care: a narrative review. *PLOS Digital Health* 2:e0000082. doi: 10.1371/journal.pdig.0000082

SUBMITTED: 11 February 2025; **ACCEPTED:** 30 April 2025;
PUBLISHED ONLINE: 04 June 2025.

EDITOR: Robert T. Knight, University of California, Berkeley, United States

SCIENCE MENTORS: Adiya Rakymzhan

CITATION: Fink O, van Gemert-Pijnen L, Lee D, Maynard A and van Schijndel B (2025) Safer Ways for Researchers to Share Sensitive Data. Front. Young Minds 13:1575140. doi: 10.3389/frym.2025.1575140

CONFLICT OF INTEREST: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

YOUNG REVIEWERS

ANSAR, AGE: 14

I am a 14-year-old boy and I live in Kazakhstan. I am a huge fan of swimming and I compete on a professional level. I am also a part of local urban planning community focused on eco-friendly and sustainable solutions. In my free time from school, I love baking, dancing, reading fiction books, and spending time with my friends.

AUTHORS

OLGA FINK

Olga Fink is a scientist who leads the Intelligent Maintenance and Operations Systems Laboratory at the Swiss Federal Institute of Technology in Lausanne (EPFL). She uses artificial intelligence to help machines and infrastructure work better and last longer. Before joining EPFL in 2022, she was a professor at ETH Zurich and led the Smart Maintenance research group at the Zurich University of Applied Sciences. Dr. Fink earned her Ph.D. from ETH Zurich and has been recognized as a Young Scientist by the World Economic Forum.

LISETTE VAN GEMERT-PIJNEN

Lisette van Gemert-Pijnen is a professor of persuasive health technology at the University of Twente, The Netherlands. Her research focusses on persuasive designs to increase trust and adherence to technologies and to develop methods for implementation in practice. She founded the first Center for eHealth Research that produces the CEHRES roadmap for eHealth. She is involved in a university wide strategic research programme to accelerate the uptake and implementation of health technologies. She is section editor of Health Technology Implementation at Frontiers (https://www.utwente.nl/en/techmed/research/research-programmes/sht/). *j.vangemert-pijnen@utwente.nl

DONGWON LEE

Dongwon Lee is a professor in the College of Information Sciences and Technology at Penn State University. He studies data science, machine learning, and cybersecurity, focusing on problems like fake news and cyber fraud. Dr. Lee earned his Ph.D. in Computer Science from UCLA and has worked at AT&T Bell Labs. He has served as a Program Director at the National Science Foundation, co-managing cybersecurity education and research programs. Dr. Lee is an ACM Distinguished Member and Fulbright Cyber Security Scholar. He leads the PIKE research group at Penn State, exploring computational and socio-technical solutions to combat fake news.

ANDREW MAYNARD

Andrew Maynard is a scientist and author who studies how new technologies affect society. He is a Professor at Arizona State University (ASU) in the School for the Future of Innovation in Society, where he also directs the Risk Innovation Lab. His work focuses on understanding the risks and benefits of emerging technologies like nanotechnology and artificial intelligence. Dr. Maynard has written books such as "Films from the Future" and "Future Rising", exploring the impact of technology on our lives. He also shares his insights through podcasts and articles, helping people understand the complex relationship between technology and society.

BASTIAAN VAN SCHIJNDEL

Bastiaan van Schijndel is the Innovation Manager at ZorgTTP, a Dutch foundation specializing in data security and privacy solutions. In his role, he focuses on developing technologies that protect sensitive information in healthcare and other sectors. His work includes implementing advanced encryption methods to ensure data remains confidential and secure during processing and analysis. Additionally, he has contributed to discussions on emerging technologies, such as privacy-enhancing technologies, as part of global initiatives like the World Economic Forum's Top 10 Emerging Technologies report.