Why Synthetic Data Is Essential for Your Organization's AI-Driven Future

What is synthetic data and why do you need it?

Organizations use synthetic data generated by AI when real data is unavailable, inadequate or inappropriate because of:
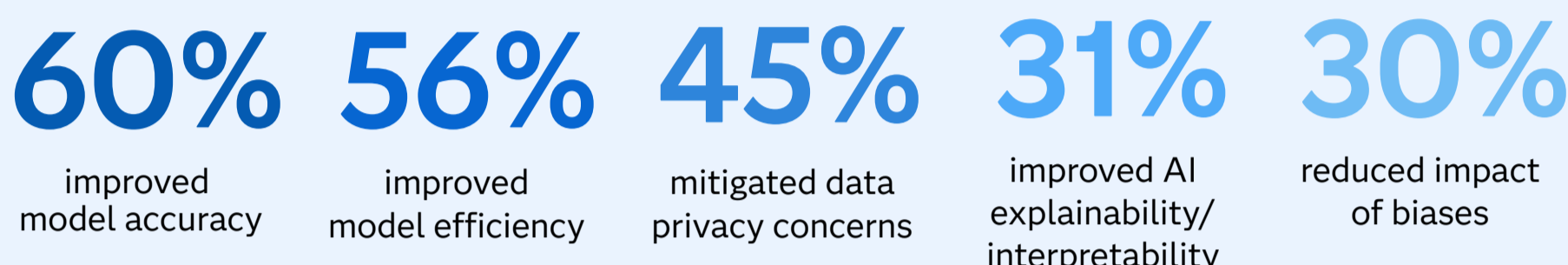
- Sensitive or private information
- Prohibitive cost
- Hand-labeling inefficiency
- Bias or imbalance
- Rare-scenario data shortages

By 2026, 75% of businesses will use generative AI to create synthetic customer data, up from less than 5% in 2023.

Source: Gartner®, Predicts 2024: The Future of Generative AI Technologies, 28 February 2024.

GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.

Benefits of synthetic data for organizations

60% improved model accuracy

56% improved model efficiency

45% mitigated data privacy concerns

31% improved AI explainability/ interpretability

30% reduced impact of biases

Source: Gartner Peer Community™, Generative AI for Synthetic Data, 2023

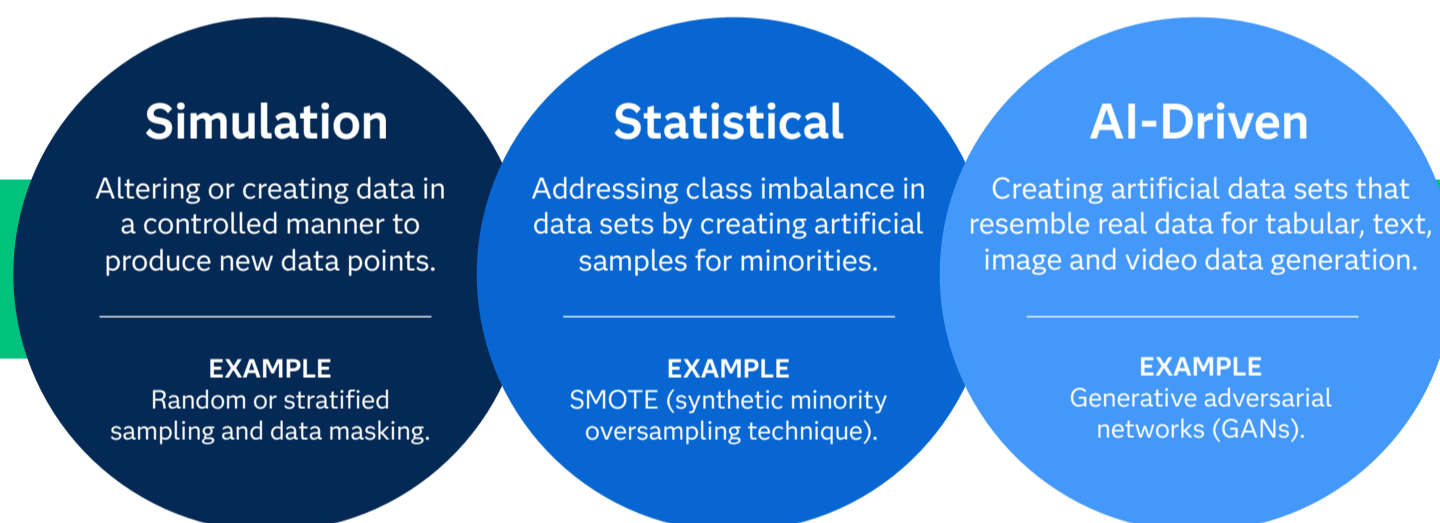## Research shows mixed feelings about the use of synthetic data

75% are concerned about data privacy and security when generative AI is used in their organizations.
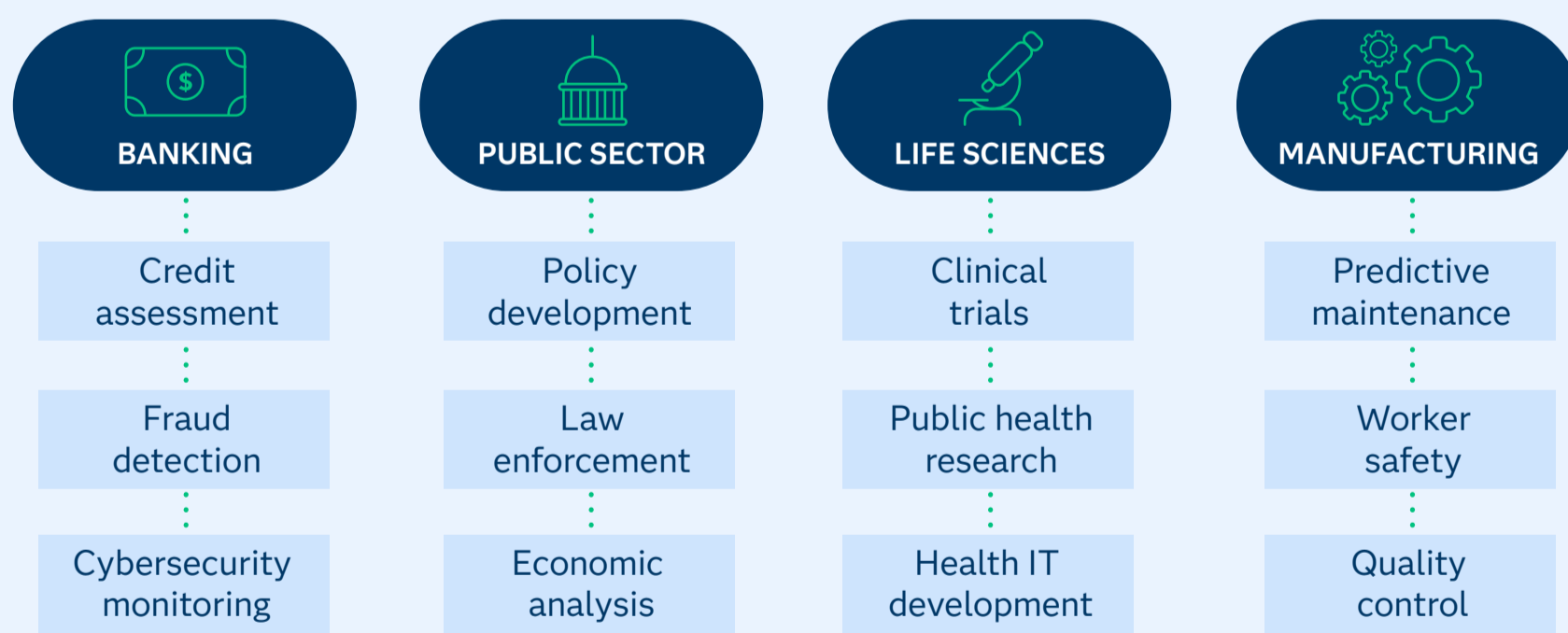
80% show strong interest in using synthetic data to address data challenges to support generative AI initiatives.

Source: Research Report: Get in the Know on Generative AI

Methods for generating synthetic data

Simulation
Altering or creating data in a controlled manner to produce new data points.

EXAMPLE
Random or stratified sampling and data masking.

Statistical
Addressing class imbalance in data sets by creating artificial samples for minorities.

EXAMPLE
SMOTE (synthetic minority oversampling technique).

AI-Driven
Creating artificial data sets that resemble real data for tabular, text, image and video data generation.

EXAMPLE
Generative adversarial networks (GANs).

Synthetic data uses

BANKING
- Credit assessment
- Fraud detection
- Cybersecurity monitoring

PUBLIC SECTOR
- Policy development
- Law enforcement
- Economic analysis

LIFE SCIENCES
- Clinical trials
- Public health research
- Health IT development

MANUFACTURING
- Predictive maintenance
- Worker safety
- Quality control

Best practices for ensuring quality of synthetic data

- Select right data generation technique
- Address privacy requirements upfront
- Validate generated synthetic data
- Combine data augmentation and data generation techniques
- Leverage approaches that scale for your enterprise

> Synthetic data that is accurately generated and rigorously validated is an indispensable resource for robust and trustworthy AI models.

BRETT WUJEK, HEAD OF PRODUCT STRATEGY FOR GENERATIVE AI, SAS

§sas

www.sas.com/data-maker