Step 3: Generate synthetic data

Selecting the right generative model and configuring it appropriately is crucial for creating synthetic data that maintains utility whilst protecting individual privacy.

Step objectives

- Select an appropriate generative model for your data type and use case
- Configure model parameters to balance utility and privacy
- Generate synthetic datasets that preserve statistical properties
- Document the generation process and model configuration
- Ensure secure handling and storage of both models and generated data

Selecting a generative model

The choice of generative model depends on your data characteristics, technical capabilities, and privacy requirements:

Rule-based approaches

Low complexity

Description: Generate synthetic data using predefined business rules and statistical distributions derived from the source data.

When to use

- Simple tabular data with well-understood relationships
- Limited technical expertise available
- Need for transparent and explainable generation process
- Small to medium datasets (up to ~10,000 records)

Advantages

- Easy to understand and implement
- Full control over generation logic
- Lower computational requirements
- Deterministic and reproducible results

Limitations

- May not capture complex statistical relationships
- Requires manual specification of all rules
- Limited ability to handle high-dimensional data
- May produce unrealistic edge cases

Example techniques

- Random sampling from empirical distributions
- Conditional probability tables
- Decision tree-based generation
- Template-based synthetic record creation

Statistical models

Medium complexity

Description: Use established statistical methods to model data distributions and relationships, then sample from these models to generate synthetic records.

When to use

- Moderate to complex tabular data
- Some statistical expertise available
- Need to preserve specific statistical properties
- Medium to large datasets (1,000 to 100,000+ records)

Advantages

- Well-established theoretical foundation
- Good preservation of statistical relationships
- Interpretable model parameters
- Established privacy analysis methods

Limitations

- May struggle with very high-dimensional data
- Requires assumptions about data distributions
- Limited handling of complex non-linear relationships
- May need manual feature engineering

Example techniques

- Gaussian mixture models
- Bayesian networks
- Copula-based models
- CART (Classification and Regression Trees)

Machine learning models

High complexity

Description: Advanced neural network and deep learning approaches that can capture complex patterns and generate high-fidelity synthetic data.

When to use

- Large, complex datasets with intricate relationships
- High-dimensional data (many variables)
- Experienced data science team available
- Large datasets (10,000+ records) for training

Advantages

- Can capture very complex statistical patterns
- Handles high-dimensional data well
- Automatically learns feature relationships
- State-of-the-art synthetic data quality

Limitations

- Requires significant computational resources
- Complex to implement and tune
- Less interpretable ("black box" models)
- More difficult privacy risk assessment

Example techniques

- Generative Adversarial Networks (GANs)
- Variational Autoencoders (VAEs)
- Transformer-based models
- Diffusion models

Model configuration and training

Proper configuration is essential for generating high-quality synthetic data whilst managing privacy risks:

Privacy-preserving parameters

Configure your model to prioritise privacy protection:

- **Differential privacy:** Add calibrated noise during training to provide formal privacy guarantees
- **Training data sampling:** Use subsets of training data to reduce memorisation risk
- **Model regularisation:** Apply techniques to prevent overfitting to individual records
- **Early stopping:** Stop training before the model memorises training examples
- **Limited model complexity:** Use simpler models that generalise better

Quality preservation parameters

Ensure synthetic data maintains utility for your intended use case:

- **Statistical fidelity:** Configure models to preserve key statistical properties

- **Correlation preservation:** Maintain important relationships between variables
- **Distribution matching:** Ensure synthetic data follows similar distributions to source data
- **Constraint enforcement:** Apply business rules and data validation constraints
- **Temporal consistency:** Preserve time-based patterns in longitudinal data

Technical implementation

Key technical considerations for implementation:

- **Computational resources:** Ensure adequate processing power and memory
- **Training time:** Allow sufficient time for model convergence
- **Hyperparameter tuning:** Systematically optimise model parameters
- **Cross-validation:** Use proper validation techniques to assess model performance
- **Reproducibility:** Set random seeds and document all configuration choices

Synthetic data generation process

Follow these steps to generate your synthetic dataset:

1

Prepare training environment

Set up secure computational environment with appropriate access controls

Ensure training data is properly prepared and cleaned

Implement logging and monitoring for the generation process

Configure backup and recovery procedures

2

Train generative model

Load prepared source data with appropriate security measures

Initialize model with configured parameters

Train model whilst monitoring for convergence and overfitting

Validate model performance using holdout datasets

3

Generate synthetic records

Sample from trained model to create synthetic records

Apply post-processing rules and constraints

Validate generated records for quality and consistency

Generate multiple synthetic datasets if needed for robustness testing

4

Quality assessment

Compare statistical properties of synthetic vs. source data

Test synthetic data utility for intended analytical use cases

Assess for any obvious quality issues or unrealistic values

Document any limitations or known issues with generated data

Documentation and model management

Comprehensive documentation is essential for reproducibility and governance:

Model documentation

- **Model type and architecture:** Detailed description of the chosen generative approach
- **Hyperparameters:** All configuration settings and their rationale
- **Training procedure:** Step-by-step description of the training process
- **Performance metrics:** Validation results and quality assessments
- **Known limitations:** Any identified issues or constraints

Generation parameters

- **Random seeds:** All seeds used for reproducibility
- **Sample size:** Number of synthetic records generated and rationale
- **Generation settings:** Any specific parameters used during sampling
- **Post-processing steps:** Any transformations applied after generation
- **Version control:** Model versions and generation timestamps

Security and storage

- **Model storage:** Secure storage location and access controls
- **Retention policy:** How long models will be retained and disposal procedures
- **Access logging:** Records of who accessed models and when
- **Backup procedures:** Model backup and recovery processes
- **Environment security:** Security measures for training and generation environments

Model security considerations

⚠️ Important security requirements

Trained models can potentially leak information about the source data and must be handled securely:

Secure model storage

After synthetic data generation, trained models must be stored separately in a secure manner or destroyed:

- Use encryption for model storage
- Implement strict access controls
- Store models separately from source and synthetic data
- Consider model destruction if not needed for future use

Access management

Control and monitor access to trained models:

- Limit model access to authorised personnel only
- Log all model access and usage
- Regular review of access permissions
- Secure model transfer protocols if sharing is required

Model disposal

When models are no longer needed:

- Securely delete model files from all storage locations
- Clear models from memory and temporary storage
- Document disposal procedures and timing
- Verify complete removal from backup systems

Decision criteria for Step 3

To proceed to Step 4, your synthetic data generation must meet these criteria:

☐ Appropriate generative model selected based on data characteristics and capabilities
☐ Model configured with privacy-preserving parameters
☐ Synthetic data successfully generated with acceptable quality
☐ Generation process comprehensively documented
☐ Trained model securely stored or appropriately disposed of
☐ Synthetic data preserves utility for intended use case whilst reducing privacy risks

✓ Proceed to Step 4

Your synthetic data is ready for re-identification risk assessment. Continue to evaluate and mitigate privacy risks.

↻ Address issues before proceeding

If any criteria are not met, address the following before proceeding:

- **Model selection issues:** Reconsider model choice based on data characteristics and requirements
- **Configuration problems:** Adjust privacy and quality parameters
- **Quality concerns:** Retrain model or modify generation process
- **Documentation gaps:** Complete required technical documentation
- **Security issues:** Implement proper model storage and access controls
- **Utility problems:** Evaluate if synthetic data meets analytical requirements

🏷️ Content tags:

📚 Related appendices

6 **Technical assessment**
Model selection and validation guidance

7 **De-identification techniques**
Privacy-preserving generation methods

🛠️ Recommended tools

Quality Metrics Tool

Assess synthetic data utility

Next steps

Step 4: Assess Re-identification Risks

Evaluate privacy risks in generated data

Related pages

Framework Overview

Return to framework overview

Related resources

Discover additional resources relevant to this content

Quick access

Framework overview

Complete framework process guide

framework overview

→

Assessment tools

Interactive assessment tools

tools assessment

→

Documentation

Technical assessment

Model selection guidance

technical models

→

Technical guidance

De-identification techniques

Privacy-preserving methods

privacy techniques

→

Common scenarios and guidance

Choosing between model complexity levels +

Balancing privacy and utility +

Using third-party generation tools +

Longitudinal and time-series data +

← Step 2: Assess Source Data

Step 4: Assess Re-identification Risks →