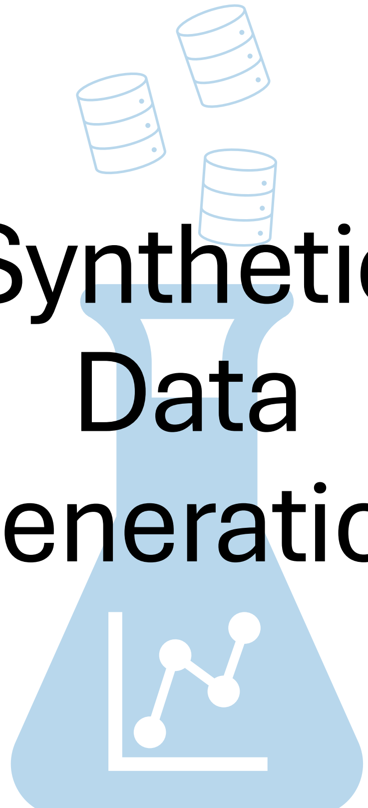


Synthetic Data Generation

Background

Data and dataset:

- **Data**, is a collection of facts, numbers, words, observations or other useful information.
- **Dataset**, is a collection of data.

Background

Data and dataset:

- Data, is a collection of facts, numbers, words, observations or other useful information.
- Dataset, is a collection of data.

Involved types of data:

- **Structured**
- **Unstructured**

Defining synthetic data

At a conceptual level:

- **Is not real** data.
- Is **generated from real** data.
- Has the **same statistical properties** as the real data.

Defining synthetic data

At a conceptual level:

- Is not real data.
- Is generated from real data.
- Has the same statistical properties as the real data.

So, if an analyst works with a synthetic dataset, he **should get analysis results similar to what he get with real data**.

Defining synthetic data - terminology

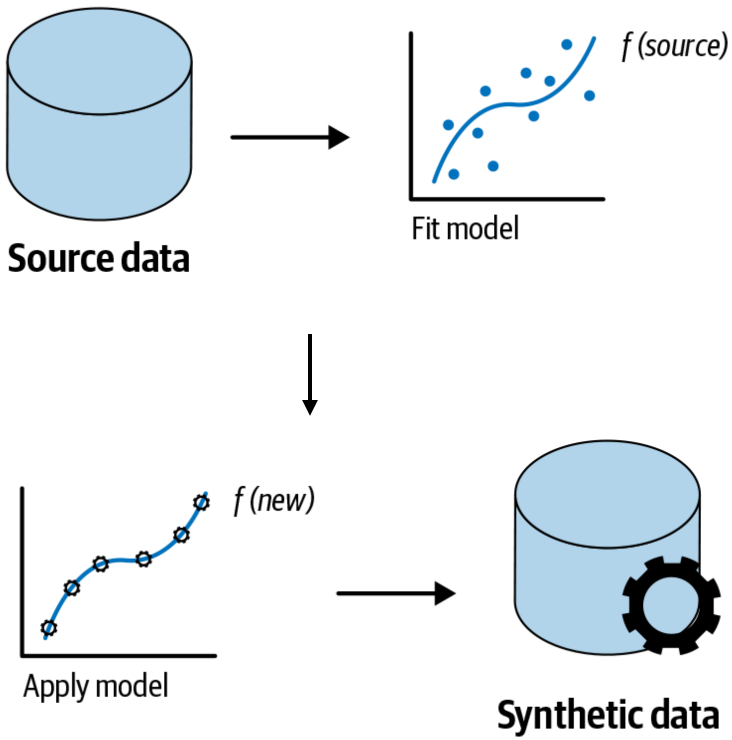
Talking about synthetic data:

- **Utility**, refers to the degree to which a synthetic dataset is an accurate proxy for real data.
- **Synthesis**, the process of artificially generating synthetic datasets.
- **Structure**, means the multivariate relationships and interactions in the data.

Defining synthetic data

Synthetic data can be categorized as:

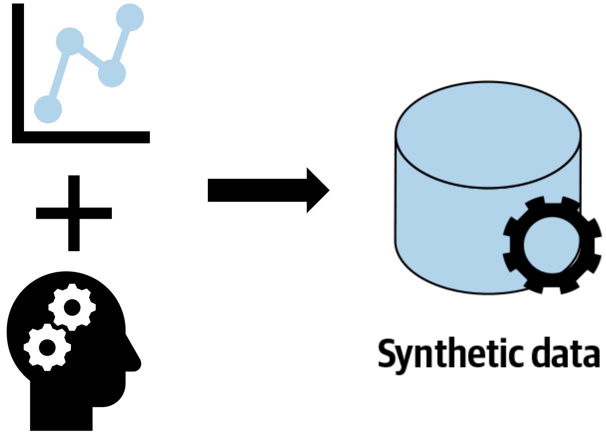
- Generated from **real datasets**.



Defining synthetic data

Synthetic data can be categorized as:

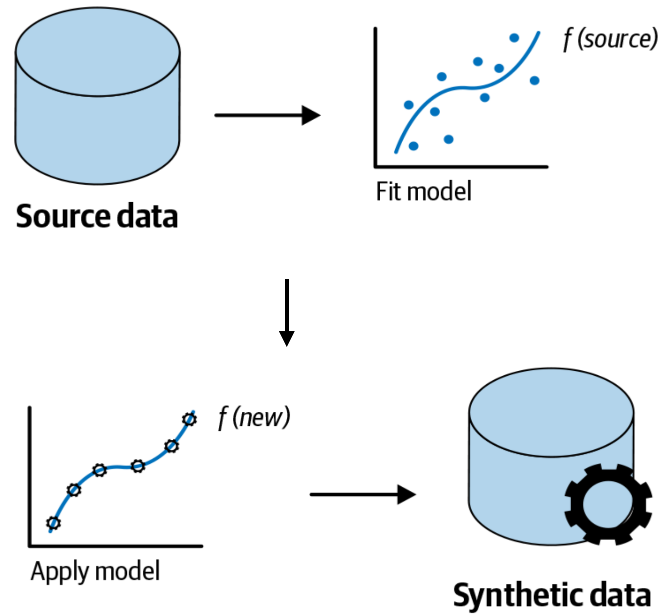
- Generated from real datasets.
- Generated **without a real data**.

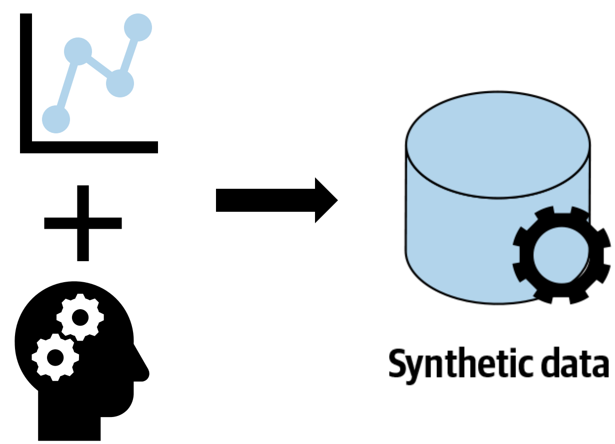


Defining synthetic data

Synthetic data can be categorized as:

- Generated from real datasets.
- Generated without a real data.
- Generated using an **hybrid approach**.



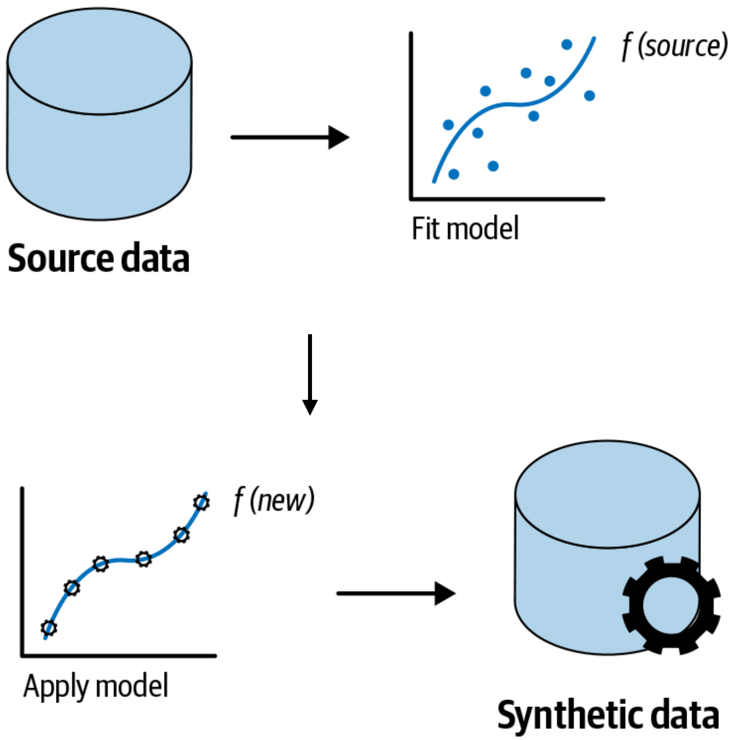


Synthesis from real data

The **first type** of synthetic data is **synthesized from real datasets**.

The **analyst has some real datasets** and then **builds a model to capture the distributions and structure** of that real data.

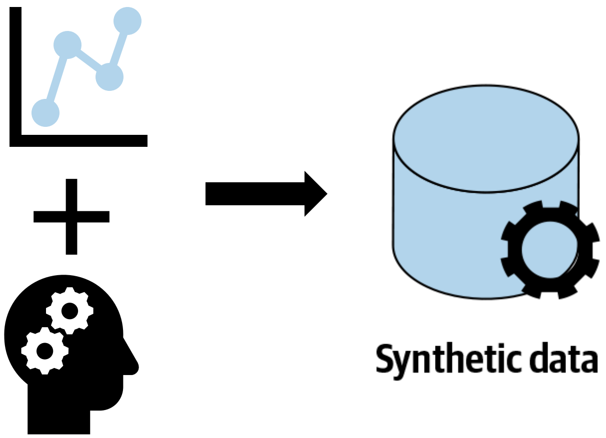
Once the model is built, the synthetic data is sampled or generated from that model.



Synthesis without real data

The **second type** of synthetic data is **not generated from real data**. It is created by **using existing models** or the **analyst's background knowledge**.

The existing models can be statistical models of a process or they can be simulations.



Benefits

Two key **advantages** of synthetic data are:

- **Efficient access** to data.
- Enabling **better analytics**.

Benefits

Two key advantages of synthetic data are:

- **Efficient access to data**.
- Enabling better analytics.

- Overcomes privacy/legal hurdles (not personally identifiable) overcoming restrictions like once imposed from the GDPR.
- Provides more diversity or coverage of rare cases within a datasets.
- Allows efficient and scalable data access.
- Reduces dependency on obtaining additional consent.

Benefits

Two key advantages of synthetic data are:

- Efficient access to data.
- **Enabling better analytics**.

- Ideal when real data collection is impractical, expensive, or unethical.
- Facilitates exploration of rare or edge cases not available in real datasets.
- Provides labeled datasets efficiently for supervised learning tasks.
- Allows analysts to validate assumptions before investing in accessing real data.

Case studies

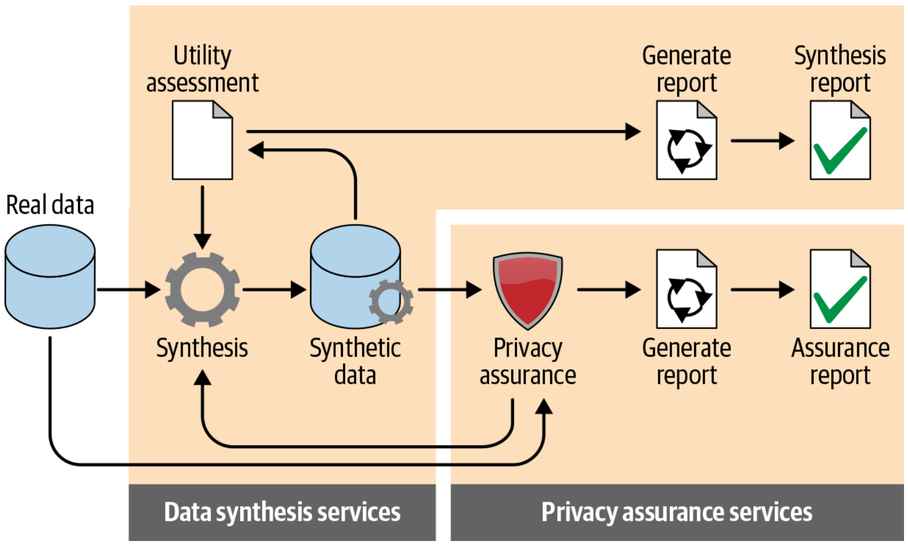
Some example of case studies of synthetic data involves:

- **Manufacturing and distribution** (e.g., robust training of robots to perform complex tasks)
- **Healthcare** (e.g., health data availability for secondary analysis)
- **Financial services** (e.g., sw testing)
- **Transportantion** (e.g., autonomous vehicles)

Data synthesis projects

Inside a **data synthesis project** the entire process involves **several phases**:

- Data preparation.
- Synthesis thecniques.
- Validation.

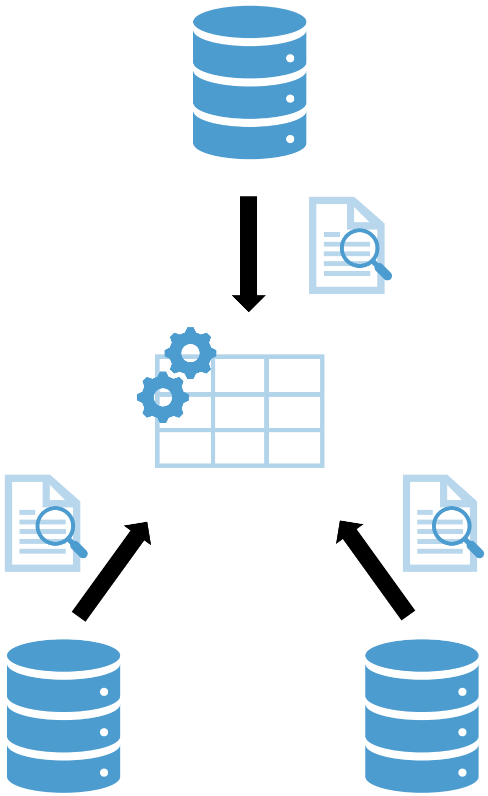


Data synthesis pipeline – data preparation

Data analysis project that starts with real data need for **data preparation.**

Data preparation includes:

- **Cleansing** (removing errors)
- **Standardization** (consistent coding schemes)
- **Harmonization** (unifying similar fields across sources)
- **Linking** real data across multiple sources (not possible post-synthesis)



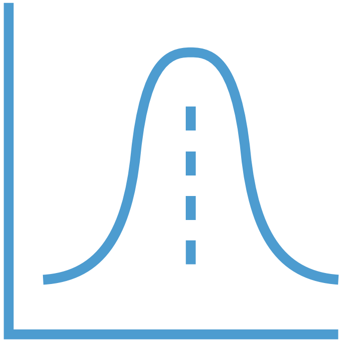
Data synthesis pipeline – synthesis thecniques

Synthetic data is produced by **modeling the structure and distributions** of real datasets, enabling the generation of realistic new samples.

Techniques include **multivariate normal distributions** (generalization of one-dimensional normal to multiple dimensions)

Distribution fitting with goodness-of-fit metrics (probability distribution that best describes a dataset)

Machine learning methods such as Classification and Regression Trees for both tabular and sequential data synthesis.



Data synthesis pipeline – data validation

Validation process implies:

- Ensure the synthetic data **maintains statistical similarity** to real data.
- Confirm it **is safe and useful** for analysis.

Validation dimensions covers:

- **Utility**, synthetic data preserve important statistical properties.
- **Privacy risk**, there a meaningful identity disclosure risk.

Data synthesis pipeline – data validation

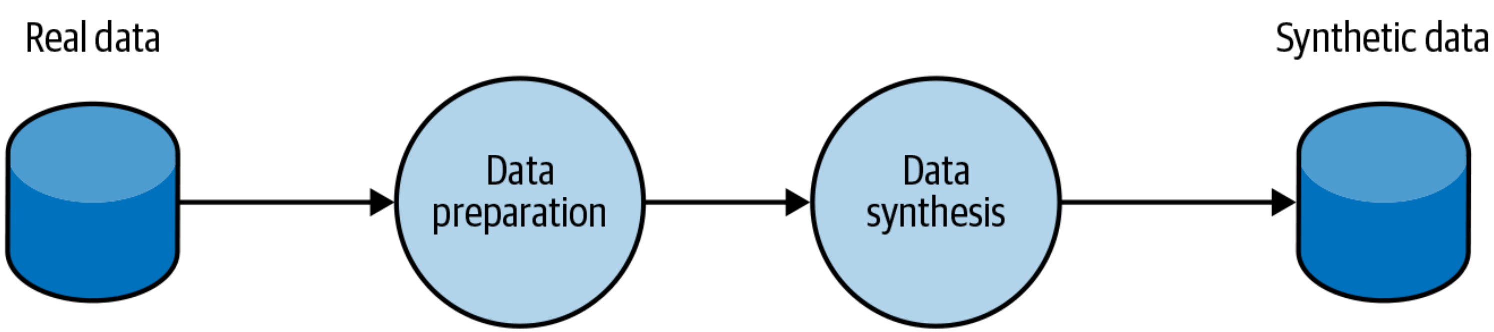
Validation methods includes:

- **Compare univariate**, bivariate, and multivariate distributions.
- **Perform distinguishability tests** (can a model tell if data is real or if it is synthetic).
- **Assess privacy risk** via unique record matching and overfitting detection.

Data synthesis pipeline

A typical data generation pipeline involves **starting with real data**, then performing **data preparation** to clean and structure the input appropriately.

Subsequently, **data synthesis techniques are applied** to generate synthetic data that mirrors the characteristics of the original dataset, **resulting in artificial data** suitable for various applications.



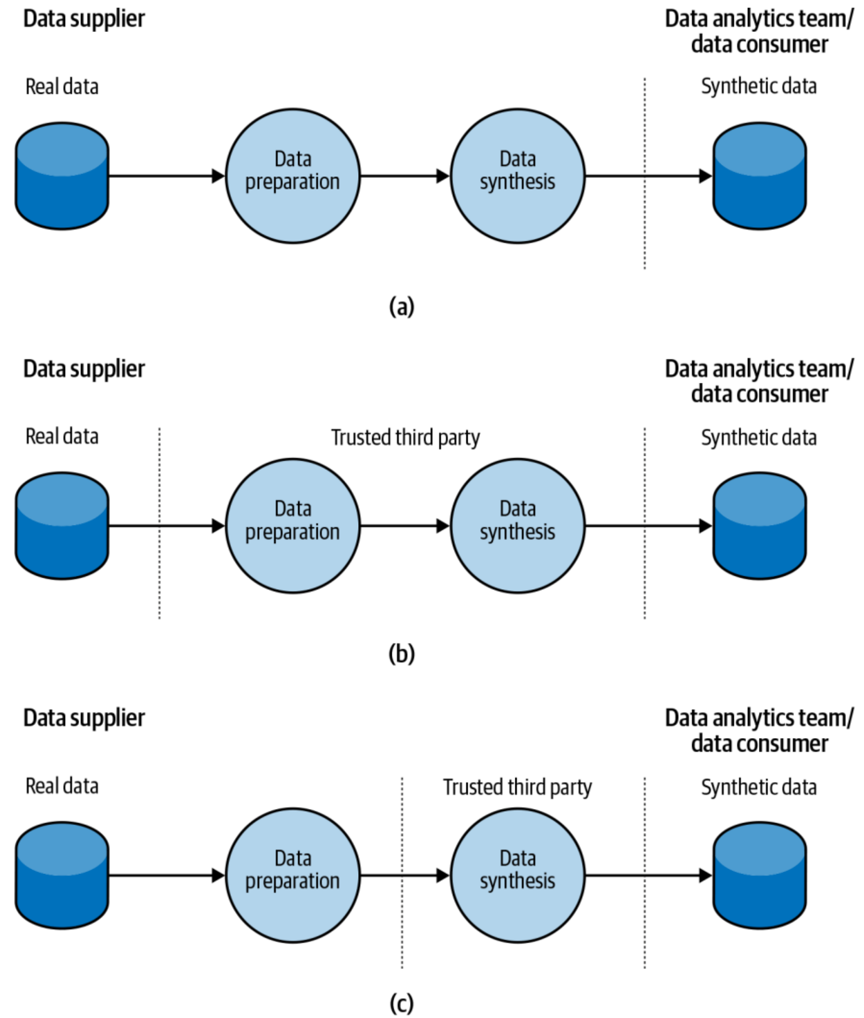
Data synthesis pipeline

There is a more complex situation in which the data source is in a different organization. Three common scenarios are:

a) Data preparation and data synthesis **both happen at the data supplier**.

b) A **trusted third party** performs **both tasks**.

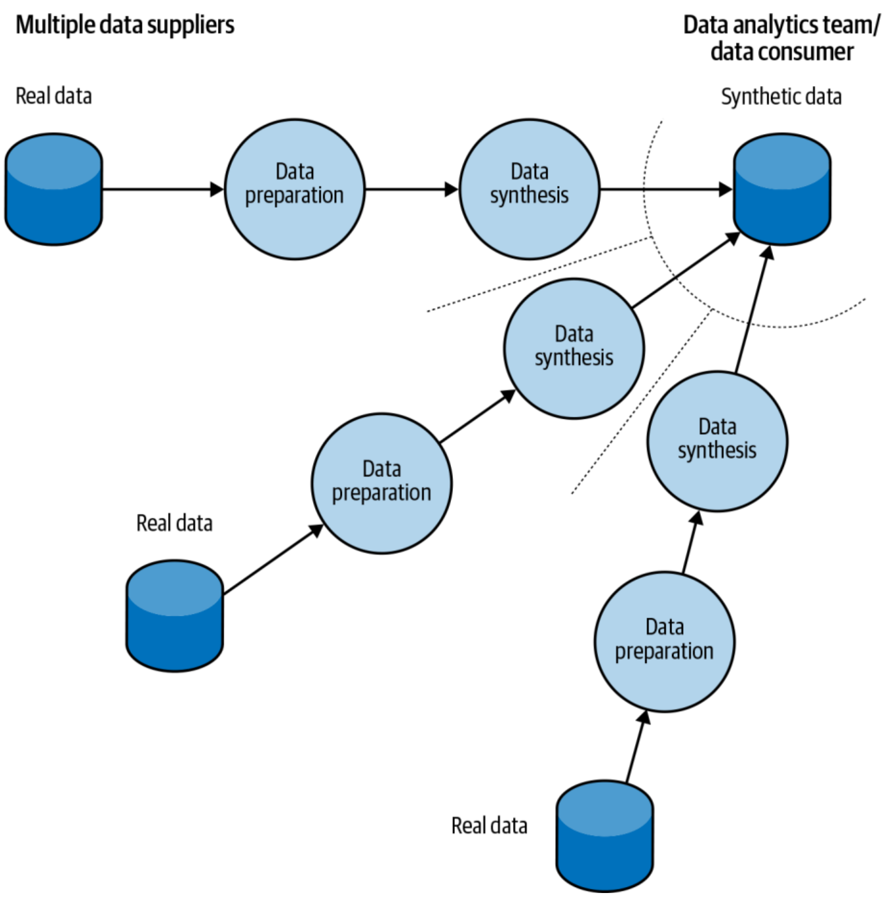
c) The **data supplier** performs the data preparation and the **trusted third** party performs the data synthesis.



Data synthesis pipeline

When data flows from **many data sources**:

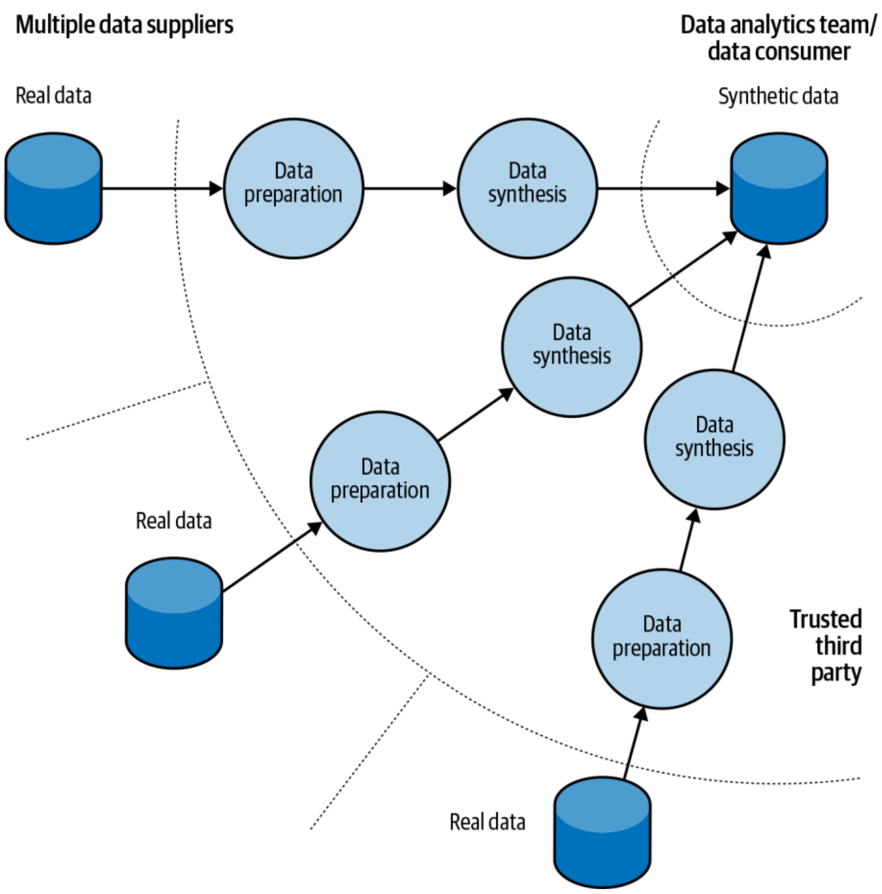
- The data is **synthesized at the source** by each of multiple data suppliers.



Data synthesis pipeline

When data flows from many data sources:

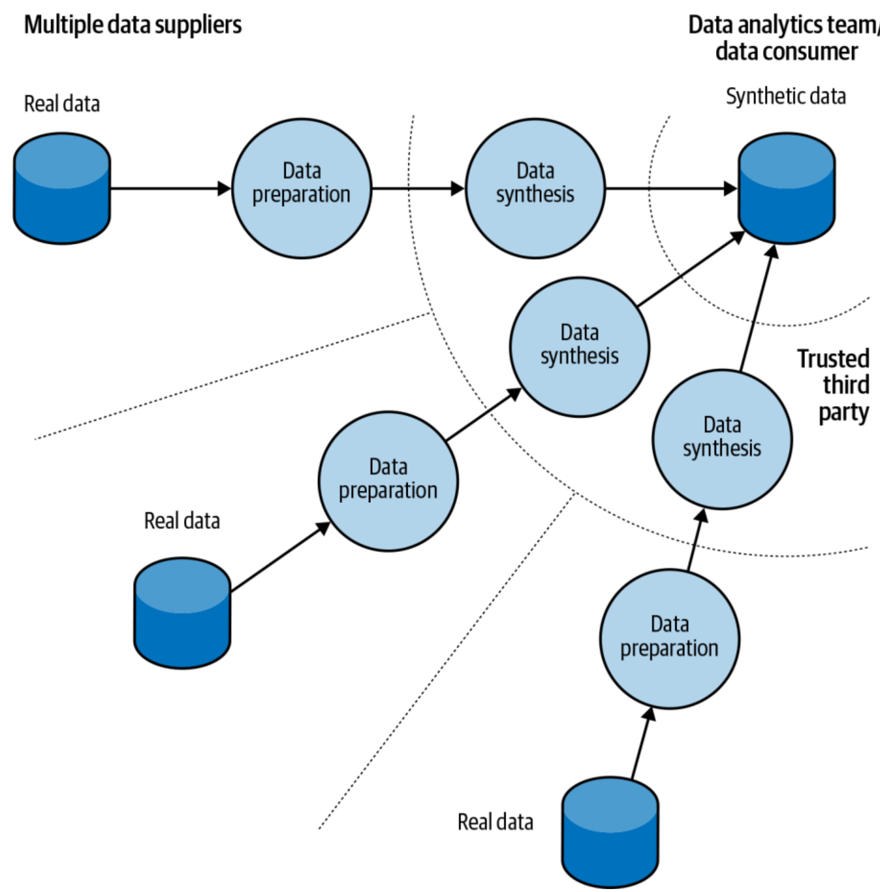
- The data is synthesized at the source by each of multiple data suppliers.
- The data are **prepared** and **synthesized** by a **trusted third party**.



Data synthesis pipeline

When data flows from many data sources:

- The data is synthesized at the source by each of multiple data suppliers.
- The data are prepared and synthesized by a trusted third party.
- The data **preparation is performed at the source** before the data is sent to the **trusted third party**.



Data synthesis pipeline

The exact data flow that would be used in a particular situation will depend on a number of factors:

- **Number of data sources**.
- The **cost and readiness** of the data analyst/data consumer to **process real data and meet any regulatory obligations**.
- The **availability of qualified**, trusted **third parties** to perform these tasks.
- The **ability of data suppliers** to implement automated data preparation and data synthesis processes

Privacy

Nowaday, **privacy is a central theme**.

Synthetic data presented as a **solution to access data for secondary purposes** while addressing privacy concerns.

Properly created synthetic data is not real data related to real individuals, and a record in a synthetic dataset does not correspond to an individual in the real dataset.

Privacy challenges in synthetic data

Synthetic data aims to protect privacy, but risks still exist if models are overfitted to real data.

Key privacy risks includes:

- **Identity** Disclosure.
- **Attribute** Disclosure.
- **Inferential** Disclosure.

True privacy risk exists when there's both a correct identity match and an information gain.

Privacy challenges in synthetic data

Synthetic data aims to protect privacy, but risks still exist if models are overfitted to real data.

Key privacy risks:

- **Identity Disclosure**.
- Attribute Disclosure.
- Inferential Disclosure.

A synthetic record can be matched to a real individual and reveal new information.

True privacy risk exists when there's both a correct identity match and an information gain.

Privacy challenges in synthetic data

Synthetic data aims to protect privacy, but risks still exist if models are overfitted to real data.

Key privacy risks:

- Identity Disclosure.
- **Attribute Disclosure**.
- Inferential Disclosure.

True privacy risk exists when there's both a correct identity match and an information gain.

Even without identifying a person, sensitive attributes about a group may be inferred.

Privacy challenges in synthetic data

Synthetic data aims to protect privacy, but risks still exist if models are overfitted to real data.

Key privacy risks:

- Identity Disclosure.
- Attribute Disclosure.
- **Inferential Disclosure.**

True privacy risk exists when there's both a correct identity match and an information gain. This is called **meaningful identity disclosure**.

Statistical inferences about real individuals or groups may be drawn, even if they aren’t in the dataset.

Legal considerations

Privacy laws like GDPR, CCPA, and HIPAA impact synthetic data practices:

- Using real data to create synthetic data is regulated.
- Sharing real data with third parties requires proper contracts and safeguards.
- Properly generated synthetic data is often not considered personal data.

Thanks