PHIRI
Population Health Information Research Infrastructure

Synthetic vs Mock-up data

Definition and distinction

Pascal Derycke, Charles-Andrew Vande Catsyne, Pierre Yriarte

Synthetic data and mockup data are both types of artificial data, but they serve different purposes and are generated using different methods. Here are the key differences between synthetic data and mockup data:

I. Purpose:

Synthetic data:

It is generated to simulate real-world data and is often used for testing, training machine learning models, and enhancing data privacy. It aims to preserve the statistical properties of the original data without containing any of the original data points.

Mockup Data:

It is primarily used for initial testing purposes. For instance according to the PHIRI framework for federated analysis, it helps in testing how an application like an analysis scripts will function before the actual sensitive data is processed in a Secure environment.

II. Generation Method:

Synthetic data:

Generated using algorithms or models that aim to replicate the statistical properties of real data. Techniques like Generative Adversarial Networks (GANs) can be used to produce synthetic data.

Mockup Data:

Often generated using simple randomisation techniques. It doesn't necessarily maintain the statistical properties of real data.

III. Use Cases:

Synthetic data:

Used in situations where real data is scarce, sensitive, or expensive to obtain. It's also used to augment datasets, especially in machine learning, to improve model robustness.

Mockup Data:

As it represents the structure and characteristics of the data, it is commonly used in design phases to prototype applications.

IV. Accuracy and Realism:

Synthetic data:

Aims to be as close as possible to real data in terms of its statistical properties and distributions. When generated correctly, it can be almost indistinguishable from real data.

Mockup Data:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101018317

Doesn't necessarily aim for realism. It's more about providing a placeholder or a artefact, artificial representation of the orginal data.

V. Privacy Concerns:

Synthetic data:

One of its main advantages is enhancing data privacy. Since it doesn't contain real individual data points, it can be used without revealing sensitive information.

Mockup Data:

Since it's doesn't represent real entities, it generally doesn't pose privacy concerns.

Disclaimer

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101018317