

Review on Synthetic Data Generation Techniques and its Outcomes

Pablo Romero-Diaz^{1,2}, Syed Taimoor Hussain Shah²,
Arnaud Gucciardi^{3,4}, Syed Adil Hussain Shah^{2,5}

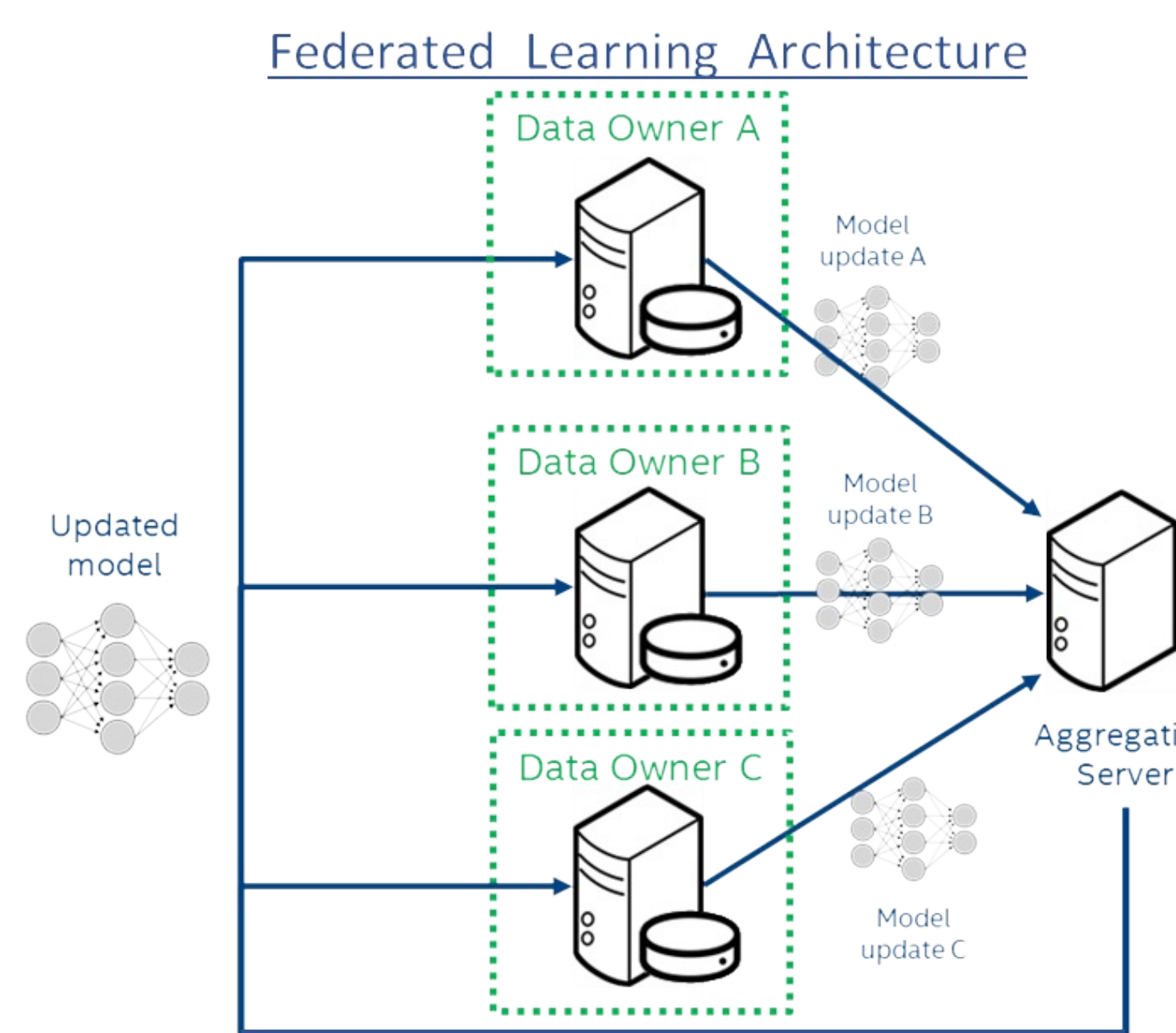


¹Department of Research and Development, 7HC, Rome, Italy - ²Department of Mechanical and Aerospace engineering, Politecnico di Torino, Italy - ³TOELT Ilc, Machine Learning Research and Development, Dübendorf, Switzerland - ⁴University of Ljubljana, Faculty of computer and information science, Ljubljana, Slovenia - ⁵Department of research and development (R&D), Gpi-SpA, Italy.

Main Research Topics

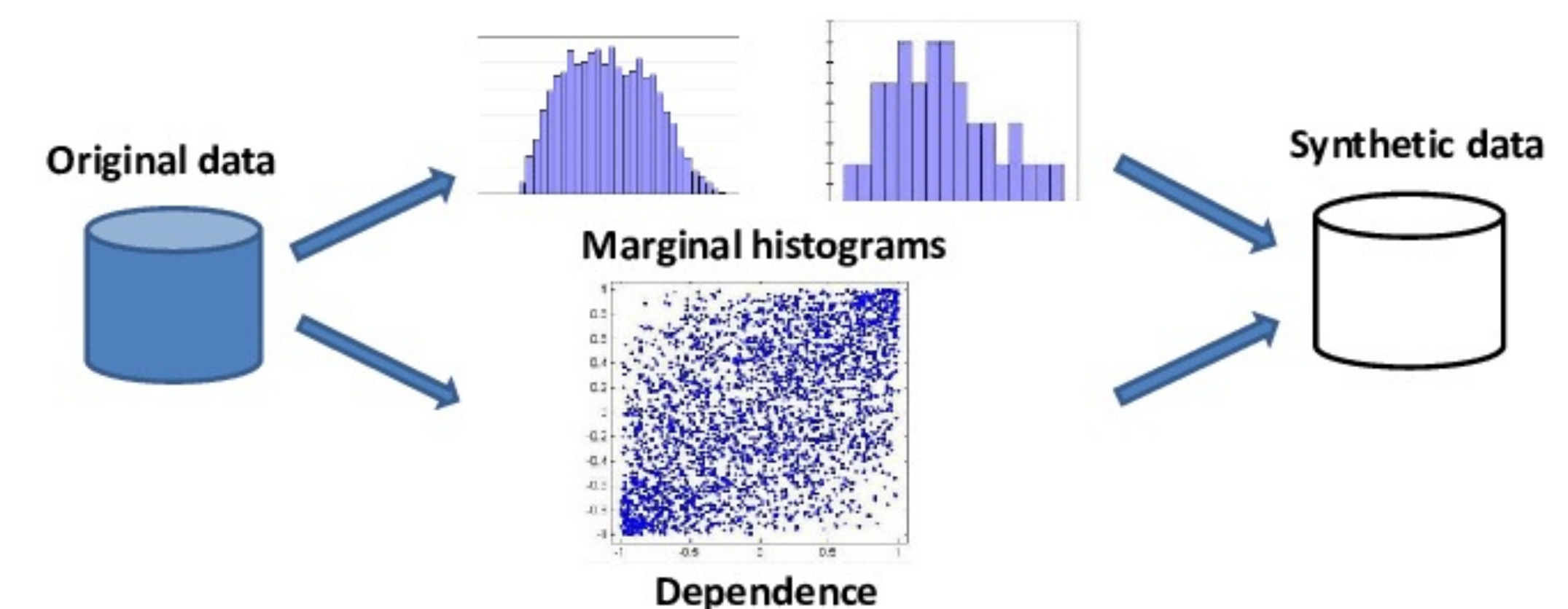
Federated Learning

Data protection is nowadays a fundamental part of any IT system. It is even more important in hospitals. A Federated Learning architecture help improve Machine Learning algorithms used in the prediction of brain and cardiovascular diseases.



Synthetic Data Generation

The lack of data to use in Machine Learning and Artificial Intelligence is one of the main problems. Generating Synthetic Data from real data is becoming an increasingly important challenge nowadays.



Review Introduction

Synthetic Data Generation is one of the main techniques for creating fictitious data from real data. Currently there are several approximation methods whose common goal is to create these data efficiently and quickly, while being highly similar compared to the real ones. This review is divided in 4 different stages:

Important Aspects in Synthetic Data Generation

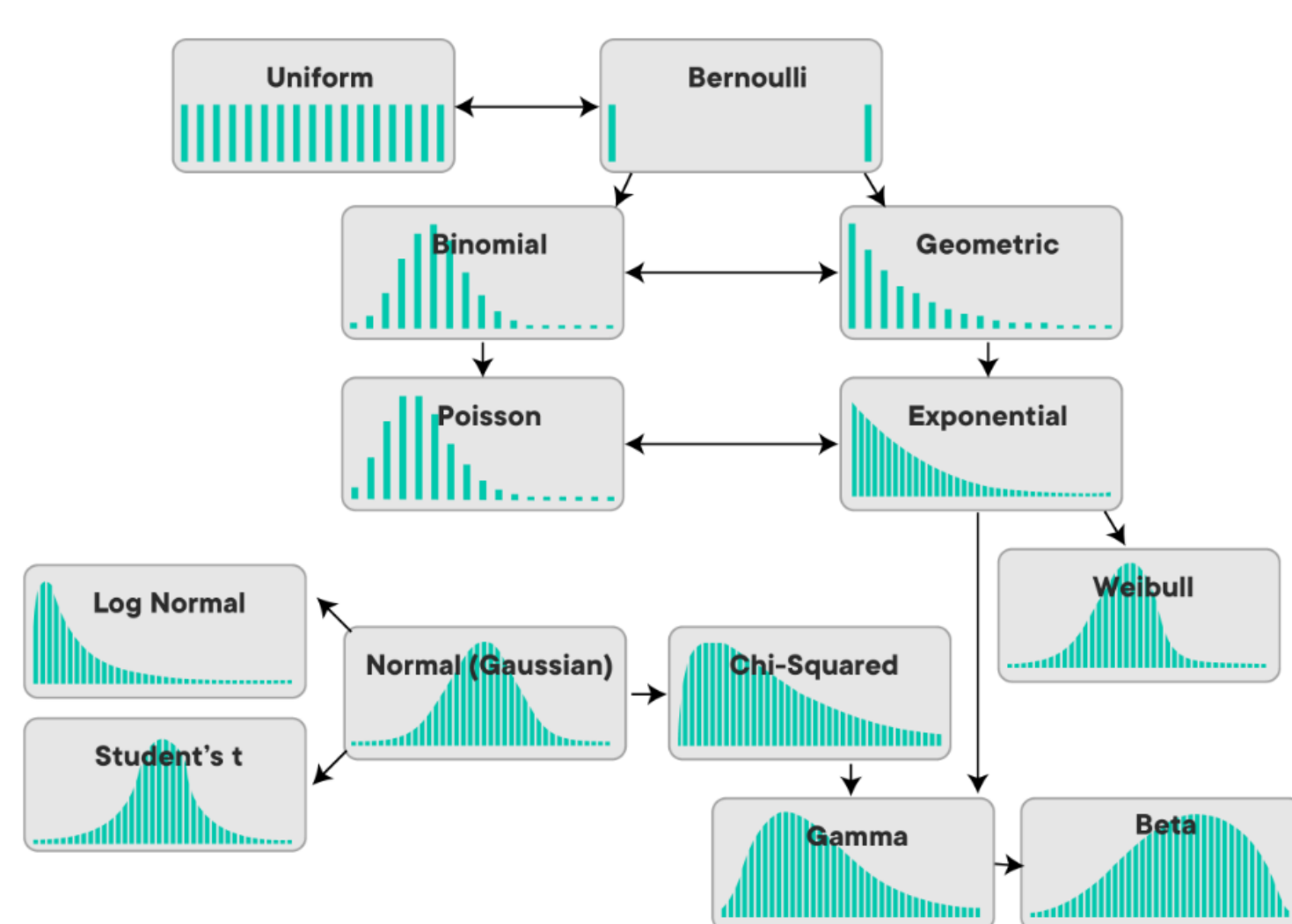
Problems in mixing real and Synthetic Data

Medical Imaging Examples

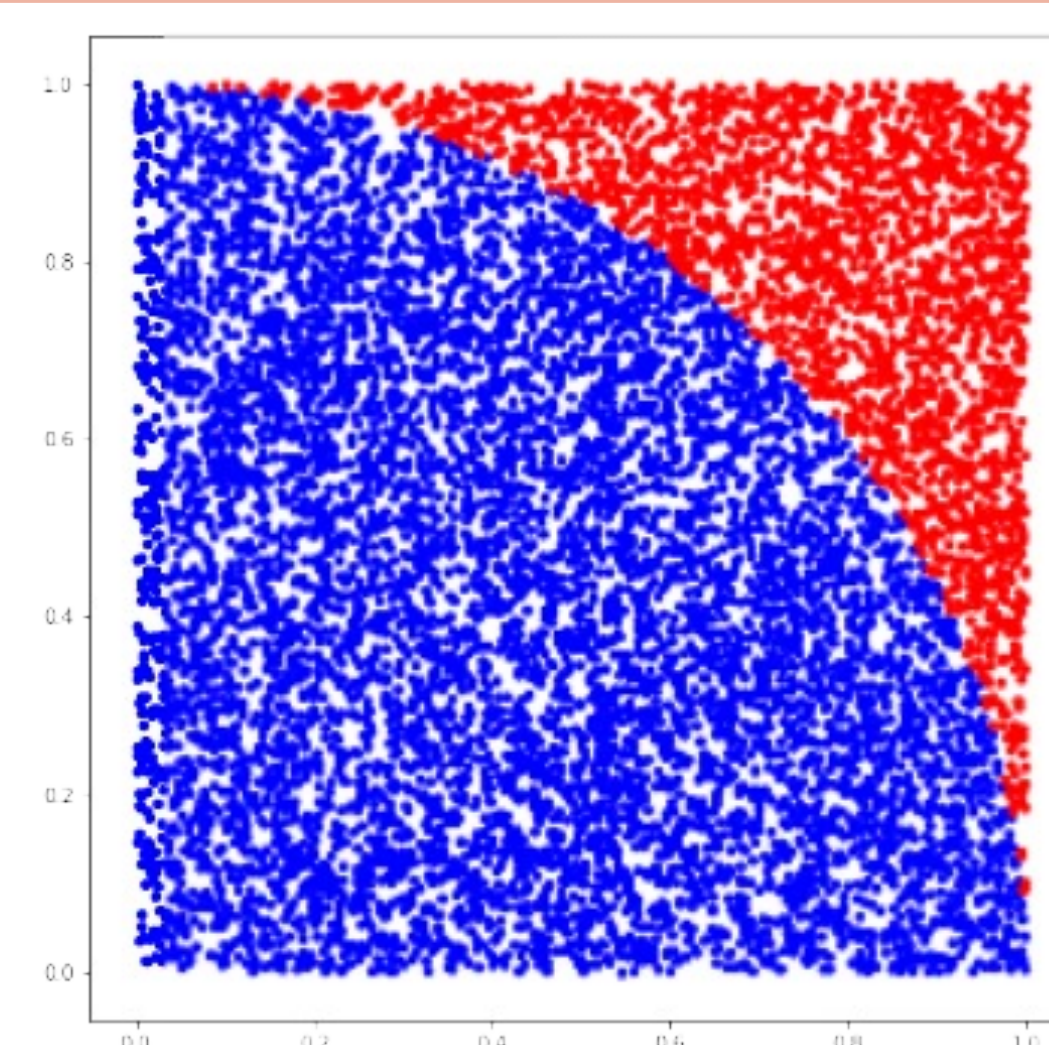
Platforms, Applications and Libraries

Synthetic Data Techniques

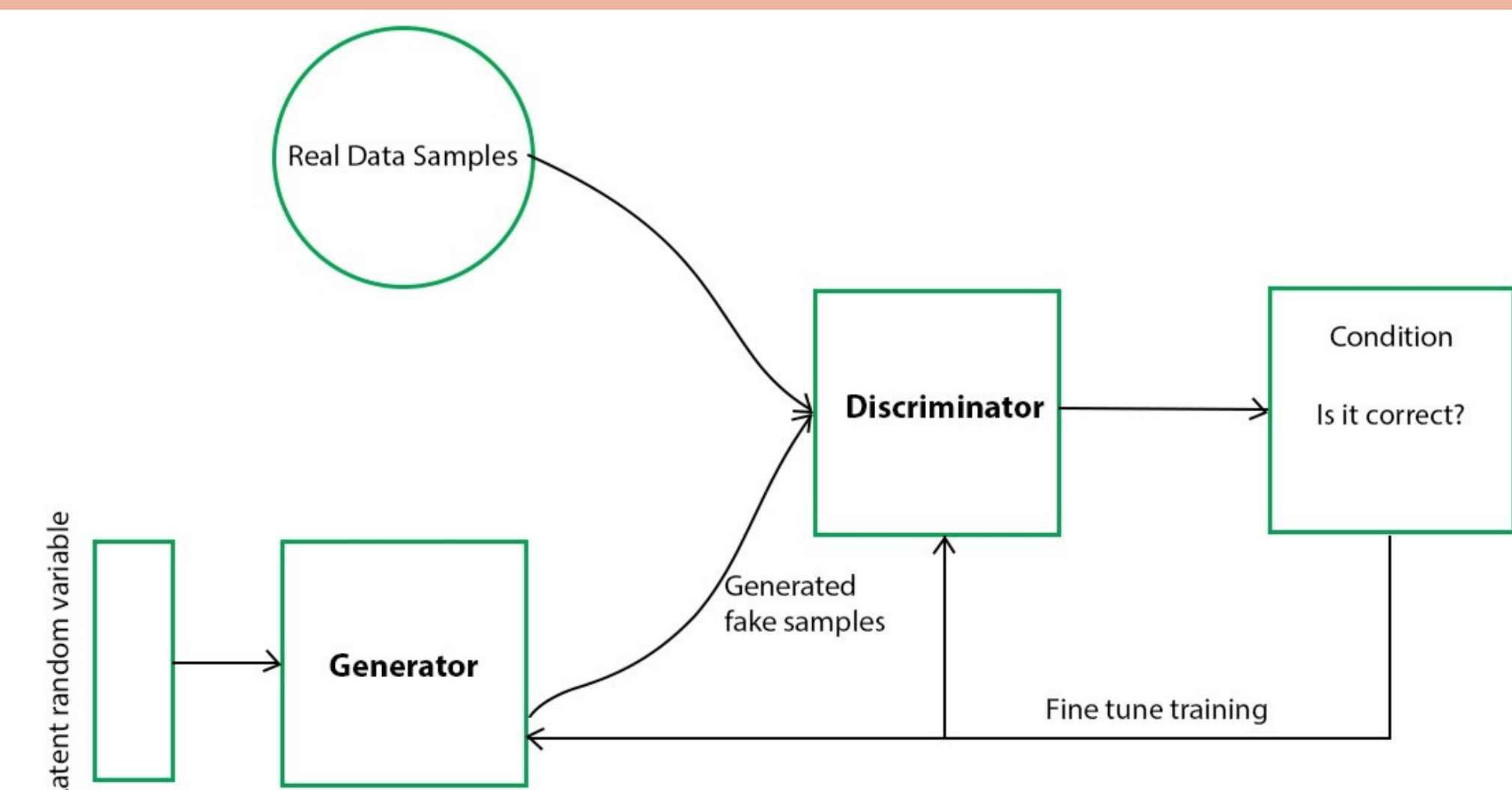
Generation According to Distribution



Fitting real data to a known distribution: Monte Carlo



Deep Learning: Variational Autoencoder (VAE) and Generative Adversarial Network (GAN)



Discussion

The number of studies applying Generative Adversarial Network (GAN) in medical imaging has increased significantly. The most popular imaging modality is Magnetic Resonance (MR) followed by Ultrasound (US). With brain and heart imaging being the main scope of this project, more efficient algorithms are needed to reduce imaging time, thus enabling the augmentation of current open databases. The main objective of this review is to learn about these techniques and improvements in order to start creating numerous images of both the brain and the heart.

References: Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation (<https://doi.org/10.48550/arxiv.1810.04304>), Differentially Private Synthesization of Multi-Dimensional Data using Copula Functions (10.5441/002/edbt.2014.43), Understanding Different Types of Distributions You Will Encounter As A Data Scientist (shorturl.at/hoKR1), Introduction to Monte Carlo simulation (shorturl.at/ioJL4), Medium.



The PARENT project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Innovative Training Network 2020 Grant Agreement N° 956394

