**GROUP 19**

**PROJECT-DOPPELGANGER**

# EXECUTIVE SUMMARY

WE ARE ALL BECOMING INCREASINGLY AWARE OF THE VALUE OF OUR DATA, and the desire to share it without the concept of a value exchange is dwindling. A true and widely accepted model for the value exchange has yet to be developed, and as a result the ability for organisations to share data is slowing down data innovation. From an organisation perspective, regulations like GDPR and an increased desire for privacy among consumers are driving this cautionary approach when it comes to data. As a result, these organisations are keen to embrace technological advances that mean they can share data and derive insights whilst maintaining compliance with the demands of both consumers and regulators. There are a number of ways in which this problem is being approached, but the one that I want to discuss in this penultimate article of the data series is **Synthetic Data**.

As synthetic data is anonymous and exempt from data protection regulations, this opens up a whole range of opportunities for otherwise locked-up data, resulting in faster innovation, less risk and lower costs. This article covers what it is, how it’s generated and the potential applications.

# BUSINESS PROBLEM BACKGROUND

## Why Synthetic data?

Data scientists all around the world are craving for data. The desire to train and deploy cutting-edge machine learning algorithms like neural networks pushes the need for more data to the next level. This quickly poses a problem when new data collection is tedious, costly or simply impossible. Synthetic data gained more and more popularity as of lately, since it promises to fulfil the need for large amounts of data. The possibility to just create some “fake” data, that for instance can subsequently be used as training data for machine learning models, sounds very promising. However, one should not fall into the trap of thinking that synthetic data is the holy grail of data science that solves all problems. In this project, we will illustrate the usefulness of synthetic data as well as discuss the common pitfalls that may arise when synthetic data is used for real use cases.

## Data protection in the EU

The data protection package adopted in May 2016 aims at making Europe fit for the digital age. More than 90% of Europeans say they want the same data protection rights across the EU and regardless of where their data is processed.

### The General Data Protection Regulation (GDPR)

[Regulation (EU) 2016/679](https://eur-lex.europa.eu/legal-content/EN/AUTO/?uri=CELEX:02016R0679-20160504&qid=1532348683434) on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. This text includes the corrigendum published in the OJEU of 23 May 2018.

The regulation is an essential step to strengthen individuals' fundamental rights in the digital age and facilitate business by clarifying rules for companies and public bodies in the digital single market. A single law will also do away with the current fragmentation in different national systems and unnecessary administrative burdens.

The regulation entered into force on 24 May 2016 and applies since 25 May 2018. [More information for companies and individuals](https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en).

Information about the incorporation of the General Data Protection Regulation (GDPR) into the [EEA Agreement](http://www.efta.int/About-EFTA/news/Incorporation-General-Data-Protection-Regulation-GDPR-EEA-Agreement-and-continued-application-Directive-9546EC-508856).

## National data protection authorities

EU countries have set up [national bodies](https://edpb.europa.eu/about-edpb/board/members_en)responsible for protecting personal data in accordance with Article 8(3) of the Charter of Fundamental Rights of the EU.

European Data Protection Board

The [European Data Protection Board (EDPB)](https://edpb.europa.eu/) is an independent European body which shall ensure the consistent application of data protection rules throughout the European Union. The EDPB has been established by the [General Data Protection Regulation (GDPR)](https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679).

# PROJECT OBJECTIVE AND SUCCESS MEASUREMENT

## **What is synthetic data?**

Synthetic data generation describes a method of producing artificial datapoints from a real dataset. The new data is supposed to mimic the original data such that the two datasets cannot be distinguished from one another, not even by human domain experts or computer algorithms. Having more data with similar properties to the original can be useful in a variety of ways. For example, machine learning models often improve in performance, the more training data is fed to them. Using synthetic data, more and complementary data can be created that eventually might improve a model.

## **How can synthetic data be generated?**

There are numerous ways to create synthetic data, each one with their own advantages and limitations. Often neural networks or Bayesian networks are utilised in order to generate new data. The following sections provide an overview of the most common tools.

### Neural Networks

Numerous methods for generating synthetic data utilise neural networks, for example *variational autoencoders* (VAE) that learn patterns in data by utilizing encoding and decoding techniques or *autoregressive models* that are used to generate synthetic images. Probably the most popular method for producing synthetic data today are *Generative Adversarial Networks* (or GANs).

## **Applications of synthetic data**

Whenever privacy concerns are an issue such as in the financial and healthcare industries or an enormous data set is required to train machine learning algorithms, synthetic data sets can propel progress. Here are just a few applications of synthetic data:

* Synthetic data with record-level data can be used from healthcare organizations to inform care protocols while protecting patient confidentiality. [Simulated X-rays](https://www.ingedata.net/blog/machine-learning-algorithms-fake-data) are combined with actual X-rays to train AI algorithms to identify conditions.
* Fraudulent activity detection systems can be tested and trained without exposing personal financial records.
* DevOps teams use synthetic data to test software and ensure quality.
* Machine learning algorithms are often trained with synthetic data.
* Waymo tested its autonomous vehicles by driving 8 million miles on real roads plus another [5 billion on simulated roadways](https://www.theverge.com/2018/7/20/17595968/waymo-self-driving-cars-8-million-miles-testing). Other automakers are using [video games](https://www.bernardmarr.com/default.asp?contentID=1513) such as Grand Theft Auto to aid its self-driving technology.

While synthetic data isn’t fool proof, it is an important tool to augment machine learning algorithms when real data is too expensive to collect, inaccessible due to privacy concerns or incomplete.

## **Disadvantages of synthetic data**

It can be challenging to create high-quality synthetic data especially if the system is complex. It’s important that the generative model creating the synthetic data is excellent or the data it generates will be affected. If synthetic data isn’t nearly identical to a real-world data set, it can compromise the quality of decision-making that is being done based on the data.

Even if synthetic data is really good, it is still a replica of specific properties of a real data set. A model looks for trends to replicate, so some of the random behaviors might be missed.

## **Advantages of synthetic data**

Huge data sets are what powers deep learning machines and artificial intelligence algorithms that are expected to help solve very challenging issues. Companies such as Google, Facebook and Amazon have had a competitive advantage due to the amount of data they create daily as part of their business. Synthetic data allows organizations of every size and resource levels the possibility to also capitalize on learning that is powered by deep data sets which ultimately can democratize mahine learning.

Creating synthetic data is more efficient and cost-effective than collecting real-world data in many cases. It can also be created on demand based on specifications rather than needing to wait to collect data once it occurs in reality. Synthetic data can also complement real-world data so that testing can occur for every imaginable variable even there isn’t a good example in the real data set. This allows organizations to accelerate the testing of system performance and training of new systems.

The limitations for using real data for learning and testing are reduced when using fabricated data sets. Recent research suggests that it is possible to get the similar results as you would with authentic data sets.

# KNOWLEDGE ACQUISITION

# KNOWLEDGE REPRESENTATION AND RECOMMENDATION

# PROJECT SOLUTION

## CASE STUDY 1

## CASE STUDY 2

# SYSTEM DESIGN

# SYSTEM ARCHITECTURE

# APPLICATION SAMPLE OUTPUT

# FURTHER ENHANCEMENTS

# IMPROVEMENTS AND OPPORTUNITY

# INSTALLATION AND USER GUIDE

# RESEARCH AND REFERENCES

1. <https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en>
2. <https://www.linkedin.com/pulse/prospects-limitations-synthetic-data-robin-r%C3%B6hm/>
3. <https://www.forbes.com/sites/bernardmarr/2018/11/05/does-synthetic-data-hold-the-secret-to-artificial-intelligence/#1a63c1da42f8>
4. <https://www.tandfonline.com/doi/full/10.1080/2058802X.2019.1668192>