Dec 11th, 12:00 AM

# Designing a Method to Nudge Analytics with Artificially Generated Data

Peter Kowalczyk
*University of Würzburg*, peter.kowalczyk@uni-wuerzburg.de

Marco Röder
*University of Würzburg*, marco.roeder@uni-wuerzburg.de

Janine Rottmann
*University of Würzburg*, janine.rottmann@uni-wuerzburg.de

Frédéric Thiesse
*University of Würzburg*, frederic.thiesse@uni-wuerzburg.de

# Designing a Method to Nudge Analytics with Artificially Generated Data

*Completed Research Paper*

**Peter Kowalczyk**
University of Würzburg
Würzburg, Germany
peter.kowalczyk@uni-wuerzburg.de

**Marco Röder**
University of Würzburg
Würzburg, Germany
marco.roeder@uni-wuerzburg.de

**Janine Rottmann**
University of Würzburg
Würzburg, Germany
janine.rottmann@uni-wuerzburg.de

**Frédéric Thiesse**
University of Würzburg
Würzburg, Germany
frederic.thiesse@uni-wuerzburg.de

## Abstract

*Recent advances to machine learning (ML) and its rapid proliferation spur the widespread development of advanced analytics applications. Nonetheless, the capabilities of ML can be stalled due to limited or missing data. In this regard, the production of artificial data offers a promising solution. However, its full potential is yet to be unleashed since its frequently misunderstood or overseen. We attribute this to a lack of practical guidance on when and how to employ artificially generated data. Against this backdrop, we draw on two streams—namely, method engineering and design science to develop "GenFlow", a novel method useful to practitioners as well as researchers. The utility is demonstrated in retrospect for previous work and empirically accessed for the context of employee attrition.*

**Keywords:** Artificial Data, Advanced Analytics, Design Science, Method Engineering

## Introduction

The last decade has seen tremendous progress in the field of machine learning (ML). ML refers to a group of algorithms designed to automatically mature through experience (e.g., in the form of historical data) (Jordan and Mitchell, 2015; Mitchell, 1997). Opposed to traditional approaches to the analysis of data, ML-based modeling offers various distinctive characteristics such as the abilities to e.g., automate the analysis, scale and adapt to changes in the data, process complex patterns, learn novel features, or facilitate unsupervised learning tasks (Bengio et al., 2012; Bzdok et al., 2017; Davenport, 2018; Hastie et al., 2009). In addition to the algorithmic frontier, ML successively becomes more tangible through a wide range of emerging easy-to-use ML tools—thus, significantly lowering the barriers for both researchers and practitioners to benefit thereof. As a result, ML increasingly permeates all kinds of advanced analytics applications. The umbrella term *advanced analytics* refers to applications, that leverage empirical data to drive decisions and actions (Barton and Court, 2012; Bose, 2009; Delen and Zolbanin, 2018; Franks, 2013; Shmueli and Koppius, 2011). Yet, the impact of advanced analytics is still in its infancy. A recent report estimates the market for advanced analytics to grow from $34.56 billion in 2021 to $74.99 billion by 2028 with a Compound Annual Growth Rate (CAGR) of 25.7% (Grand View Research, Inc., 2022).

However, despite the overall prospect of growth, the effectiveness of advanced analytics naturally depends on the data to be analyzed, that is its availability, quality, accessibility, and with regard to live applications

in particular a constant flow of data (Jöhnk et al., 2021). Consequently, if data are limited (e.g., in terms of quality or actual amount) or not available at all (e.g., due to privacy restrictions or even non-existence) advanced analyses might be restricted or even impossible (Bauer et al., 2020; Berger and Doban, 2014; Watson et al., 2020), leaving affected organizations at a disadvantage compared to the rest. To overcome this bottleneck to the development of advanced analytics applications, the use of *synthetic data* offers a promising solution (Nikolenko, 2021). Synthetic data—unlike real data—is not captured from the real world but rather generated artificially (Nikolenko, 2021; Patki et al., 2016). In effect, adequate data may be provided in abundance to meet the desired requirements. Thus, artificially generated data enables organizations to continue driving existing advanced analytics applications or to pursue novel endeavors previously beyond the realm of possibility. For example, due to the lack of adequate data, Kokubo et al. (2021) leverage artificially generated data to accurately remove raindrops from images to aid the development of autonomous vehicles and drones. Similarly, Zhang et al. (2023) artificially create images by simulating brain tissue and neurons to improve the speed and accuracy of neuron detection.

Contrary to the benefits expected from the consideration of artificially generated data (i.e., simulate specific situations, privacy-preservation, enhance predictive power), it is rarely employed (Chen et al., 2021b; James et al., 2021; Visani et al., 2022). We attribute these missed opportunities to the current rise of ML in analytics as a fruitful field of application for synthetic data, the sheer novelty of some techniques to produce the data, the question as to whether the expected effects from using artificial data actually occur, and the existing lack of methodical guidance on its proper use. Regarding the latter, practitioners and researchers are left alone to decide where and how to employ synthetic data or worse, overlook the opportunity in the first place. Having at hand an adequate method enables its users to appropriately reason about the use of artificially generated data in a systematic manner involving its acclaimed potentials (i.e., performance improvement or privacy preservation) as well as associated challenges (i.e., generating adequate data, assessing its utility and the degree of privacy preservation, finding a good balance between the two and the risk of bias magnification) (Rajotte et al., 2022). A method dedicated to the conscious use of synthetic data holds the potential to drive advanced analytics endeavors, thus contributing to the ever-expanding range of application contexts for ML (Berente et al., 2021). As the development of such methods is at the core of the information systems (IS) discipline and in light of the contemporary research mandate of IS to contribute to the diffusion and likewise adoption of ML (Dwivedi et al., 2015; Padmanabhan et al., 2022; Ram and Goes, 2021), in this article, we design a well-suited method to include the peculiarities of synthetic data. For this purpose, we employ the method engineering as design science approach put forward by Goldkuhl and Karlsson (2020) and thereby address the following question:

**RQ** *How should a method be designed to include artificially generated data for the development of advanced analytics applications?*

Thus, the remainder of the present article unfolds as follows. The next section outlines established methods for the development of advanced analytics applications and compares them. Subsequently, we analyze the current state of synthetic data application in extant literature. The review provides insights into the use of synthetic data, that is, its generation and procedural point of introduction. Furthermore, it reveals the lack of guidance and thereby further substantiates the initial argument. This, in turn, is used in the following to highlight and clearly explicate the identified research gap. The next section briefly outlines the chosen design-oriented method engineering research approach as proposed by Goldkuhl and Karlsson (2020) to guide the development of the novel method. Here, we specifically put emphasis on the activities and principles necessary. Next, we apply the research approach by executing the activities to finally obtain a meaningful synthetic data inclusive advanced analytics method—*GenFlow*. Afterwards, the utility of *GenFlow* is demonstrated in retrospect for previous work and empirically accessed for the context of employee attrition. The article concludes with a resume on the overall contributions, a summary of the limitations involved, and an outlook on future work.

# Conceptual Background

## *Overview of Popular Methods*

Prior to the design of the novel synthetic data inclusive method, it is necessary to gather a fundamental understanding of existing methods useful to facilitate advanced analytics. For this purpose, a diverse panoply of methods is readily available—each with its respective peculiarities and some exceedingly more popular than the rest (Azevedo and Santos, 2008; Mariscal et al., 2010; Shafique and Qaiser, 2014). Although the methods are similar at their core (i.e., some ex-ante steps, a data phase, a modeling phase, and the ex-post stage), they each have different facets. We, therefore, outline five of the often-cited methods relevant to this study—namely *knowledge discovery in databases (KDD)*, *cross-industry standard process for data mining (CRISP-DM)*, *sample explore modify model assess (SEMMA)*, *steps in building an empirical model (SBEM)*, and *big data analytics guidelines (BDAG)*. To this end, apart from a basic description of the methods' steps, we cover their origin and general focus. The approaches are depicted in Figure 1 in a comparative manner.

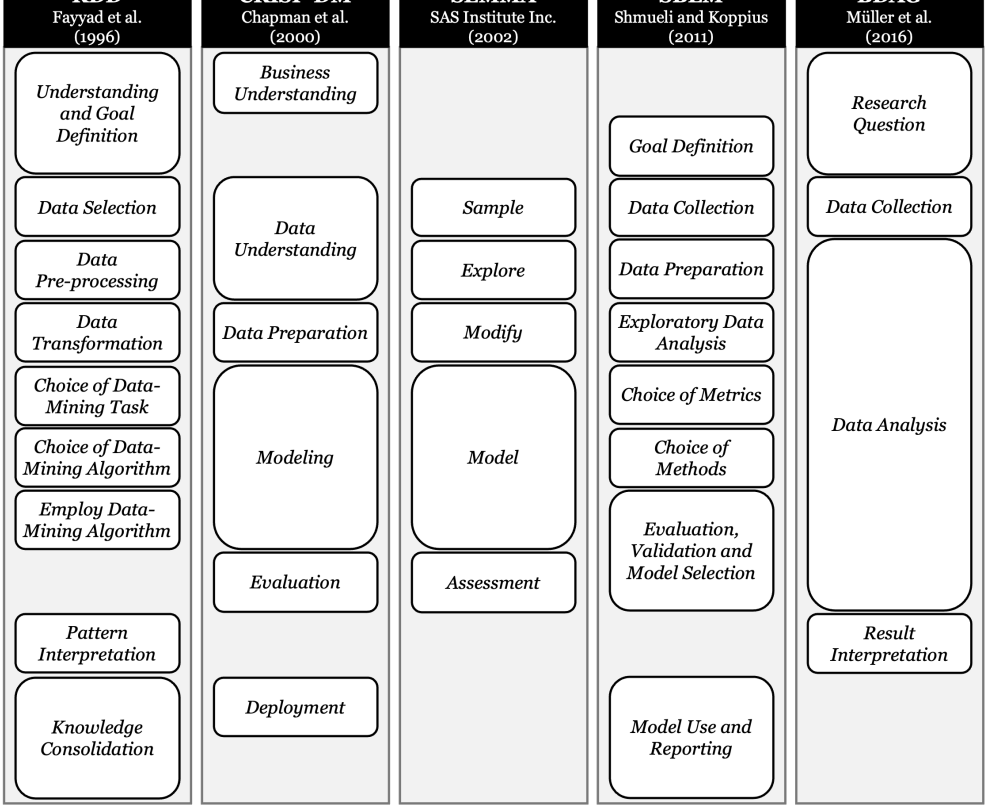


**Figure 1. Comparison of Popular Methods**

**KDD.** The first well-known approach is the iterative process *KDD* put forward by Fayyad et al. (1996). The proceed comprises nine chronological steps *understanding and goal definition*, *data selection*, *data pre-processing*, *data transformation*, *choice of data-mining task*, *choice of data-mining algorithm*, *employ data-mining algorithm*, *pattern interpretation*, and *knowledge consolidation*. Although the process is not exclusively designed for advanced modeling, it emphasizes the broader topic of data-driven analyses (i.e., data mining) and details the steps to data-associated tasks. Therefore, its fundamental idea and its generic steps are highly transferable to our study. However, due to the sheer novelty of the ML-specific modeling requirements, the descriptions of *KDDs* steps neglect aspects, which are crucial for contemporary advanced analytics projects, such as data partitioning, validation, evaluation, or the use of synthetic data.

**CRISP-DM.** Next, we consider *CRISP-DM*—one of the most widely-used analytics models (Catley et al., 2009; Chapman et al., 2000; Piatetsky, 2014). The method resulting from the various experiences gathered from practitioners was first conceived in 2000 (Chapman et al., 2000; Wirth and Hipp, 2000). The six steps of *CRISP-DM* are *business understanding*, *data understanding*, *data preparation*, *modeling*, and *evaluation*, *deployment* (Chapman et al., 2000; Wirth and Hipp, 2000). The cyclic and open layout of the process model emphasizes the natural conditions of advanced analysis projects. Thus, a back-and-forth between the steps is possible. However, similarly to *KDD*, *CRISP-DM* has its roots in the data mining era and thus is not particularly suitable for modern ML-driven advanced analytics. Since its first appearance, *CRISP-DM* received a series of follow-up modifications as described in Mariscal et al. (2010). However, due to the overall broad focus, *CRISP-DM* and its further revisions do not adequately address the specific requirements of modern ML projects—especially in the light of synthetic data use.

**SEMMA.** *SEMMA* is another popular method that guides the implementation of analytics applications (SAS Institute Inc., 2002, 2017). The acronym resembles its five steps—*sample*, *explore*, *modify*, *model*, and *assess*. Notably, the *sample* phase refers to the extraction of the necessary data from a larger data source. The sampled data set should be both large enough to enable modeling and capable to be processed efficiently (SAS Institute Inc., 2017). Consequently, the *sample* step in *SEMMA* does not bridge the gap to synthetic data as the name itself might possibly suggest. Although *SEMMA* marries some of the core activities relevant to ML projects in a broader sense, it lacks fundamental aspects relevant to advanced analytics like scope definition or interpretation (Rohanizadeh and Bameni, 2009).

**SBEM.** The fourth method we consider is *SBEM*—developed by Shmueli and Koppius (2011). As opposed to the aforementioned approaches, *SBEM* stems from IS research and is directly designed to build empirical models for analytics. As a result, *SBEM* better suits the specific requirements of executing ML-driven studies. After the initial *goal definition*, the operator performs several data-related steps—namely, *data collection*, *data preparation*,and *exploratory data analysis*. Subsequently, the *choice of metrics* and *choice of methods* are performed prior to the step of *evaluation, validation,* and model selection. Lastly, the *model use and reporting* phases conclude the *SBEM* proceed. Again, as with the previous methods, *SBEM* does not provide assistance regarding the peculiarities associated with synthetic data.

**BDAG.** Finally, we consider *BDAG* by Müller et al. (2016) which again originates in the IS domain. *BDAG* is predominantly motivated by the analysis of vast data sets. The method comprises four broad steps beginning with a *research question*, followed by *data collection*, *data analysis*, and *result interpretation*. However, in its form, *BDAG* rather delivers a high-level proceed than a detailed description on how to conduct ML-driven analytics. Thus, it is well-suited for practitioners and scholars in need of a flexible method to adapt according to the pursued ML endeavor.

### *Literature Review of Synthetic Data Use*

Aside from the above resume of well-established methods for advanced analytics projects, it is of particular interest how extant literature currently deals with synthetic data, that is, its point of entry and the linked generation technique. To this end, yet another time, attention is paid to the method—if any is used at all. Therefore, we adhere to the guidelines put forward by Vom Brocke et al. (2009). They refer to the five tasks: (i) setting an adequate review scope, (ii) conceptualizing the topic, (iii) conducting the literature search, (iv) synthesizing the literature, and finally (v) formulating a research agenda. However, as the literature is merely intended to uncover the research gap and serve for later method development, we refrain from formulating a research agenda. Consequently, for the first task, in line with Vom Brocke et al. (2009), we make use of the taxonomy of Cooper (1988) to determine the review scope accordingly. Next, we conceptualize the topic by specifying its key components in line with the scope. In the consecutive task, we specify keywords that constitute the search query, which serve to search the different databases and journals accordingly. In the fourth task, we systematically filter the articles retrieved according to their relevance by first removing duplicates, analyzing titles and abstracts as well as full texts. The remaining articles provide a comprehensive overview of the area of interest.

In line with Cooper (1988), we first define an adequate review scope by using the suggested taxonomy. Regarding the use of synthetic data in advanced analytics studies, the questions arise as to how the study is

conducted and more specifically how synthetic data are produced and where the data are introduced. Accordingly, three pillars are of special interest: (i) the steps taken to develop the advanced analytics application, (ii) the entry point for synthetic data to the workflow, and (iii) the chosen technique to produce the data. Thus, we direct our focus toward advanced analytics applications that consume synthetic data. To this end, we do not concentrate on literature that is primarily motivated by algorithmic challenges rather than finding a solution for a specific application context. Furthermore, we aim to identify and conceptually highlight central issues by taking a neutral perspective as described by Cooper (1988). The literature analysis is mainly targeted to a specialized audience, that is scholars concerned with method engineering for advanced analytics. The literature is covered in a representative manner due to the choice of a certain search query and specific databases. To conceptualize the topic, we mark out the key elements based on our scope—i.e., (i) the method, (ii) the entry point for synthetic data, and (iii) the synthetic data generation technique. These provide the contents sought for within the literature review. Next, we scan five databases—namely, AIS electronic Library, IEEE Xplore, ACM Digital Library, EBSCO Host, EconBiz—for one of the following keywords in the article's titles to ensure a strong affiliation with the subject matter: "synthetic data" or "synthesized data". This yields a total of 736 articles as of April 25th, 2023, of which 330 remain after full text analysis (cf. Figure 2).

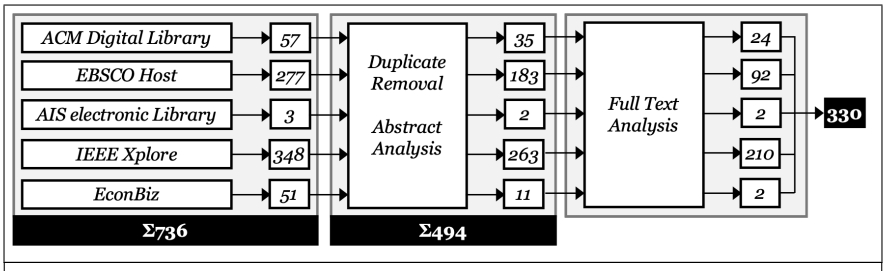


**Figure 2. Overview of Advanced Analytics Literature Review**

Aside from methodical causes (i.e., duplicate removal between the databases), this reduction is mainly due to content-related considerations. To be more specific, a large group of the studies found initially rely on pre-existing synthetic data from external data sources rather than creating their own proprietary dataset (e.g., (Chen et al., 2021a)). Others, however, do not provide any details on the data generation procedure whatsoever (e.g., (Bue and Merényi, 2010)). This phenomenon once more highlights the need for a consistent method to guide the use of synthetic data effectively. Moreover, we discard publications that are solely concerned with the development or advancement of algorithms using established benchmark datasets (e.g., MNIST or CIFAR-10) without pointing toward a practical use case. This rationale leads to the exclusion of 93 articles from further analysis.

The resulting literature base of 330 articles unveils several valuable insights regarding the state of synthetic data use in advanced analytics from a methodical perspective. First and foremost, none of the retrieved studies follows or at least adapts a pre-existing method (such as one of these depicted in Figure 1) to carry out the synthetic data inclusive analysis. Furthermore, only 66 out of 330 relevant articles provide details about the individual steps taken. In contrast, regarding the other 264 cases (80%) no systematic method is pursued whatsoever. This once again underpins the immanent lack of adequate guidance to enable a target-oriented and likewise thoughtful consideration of synthetic data use for advanced analytics projects.

With respect to the entry point of synthetic data to the analytics pipeline, three stages may be distinguished. Particularly noteworthy in this regard is the option to use multiple points of entry for the synthetic data within the course of a project in parallel. Consequently, the numbers presented in the following result from this non-exclusivity of the entry points such that their sum exceeds the original number of articles retrieved. The first entry point refers to the actual acquisition of the data to be analyzed (i.e., in 180 cases). Here, the data are either produced via statistical-distribution sampling (e.g., Walonoski et al. (2020)) or through simulation models (e.g., Baul et al. (2021)). Notably, to artificially create the data both methods require prior knowledge—either self-recalled or alternatively in the form of a sophisticated data production tool (Kowalczyk et al., 2022). The second identified entry point is located right after the initial collection of the data

during the analysis and preparation step. At this stage, the researchers introduce synthetic data in 219 out of 330 cases. Since primary data are already in place, synthetic data can be either produced through traditional augmentation techniques such as synthetic minority oversampling technique (SMOTE) as introduced by Chawla et al. (2002) or novel ML-driven approaches like e.g., generative adversarial networks (GANs) or variational autoencoders (VAEs) (Damer et al., 2018; Goodfellow, 2017; Kingma and Welling, 2019; Sharma et al., 2020; Wei and Mahmood, 2021). Although a detailed description of these techniques is of great value to practitioners it is beyond the scope of the present article. Thus, we refrain from making a technical deep dive into these techniques and instead direct the interested reader to the extant literature from above. Lastly, only ten out of all articles considered using synthetic data for the purpose of evaluation. For instance, Yannikos et al. (2012) simulate distinct fraud cases to evaluate their detection system.

### *Research Gap*

In light of the contemporary use of synthetic data for advanced analytics outlined in the previous section, it becomes evident that there is a substantial lack of methodical guidance. Without the proper synthetic data inclusive method researchers as well as practitioners might oversee the potentials associated from the beginning. If, however, they decide to employ synthetic data, they are—to this date—left alone to decide where and how to use it at best. In this respect, as shown above, the existing methods do not provide sufficient guidance regarding the inclusion of synthetic data. Without clear instructions on the options to explore, they could miss valuable options that would fuel advanced analyses. More-so, communicating about the use of synthetic data without a proper standard can become quite challenging.

## Research Approach

Research concerning information systems development methods (ISDMs) is recognized as a pivotal part of IS research (Nunamaker Jr et al., 1990), by which the knowledge inherently cast into methods serves as an anchor point to users (Beynon-Davies and Williams, 2003; Russo and Stolterman, 2000). Thus, ISDMs represent a strategic and success-critical factor for the use of IS in organizations (Beynon-Davies and Williams, 2003). In this context, Goldkuhl and Karlsson (2020) perceive methods to either represent useful instruments and objects (e.g., applicable method or evaluation technique) to pursue an endeavor or the study result in itself (e.g., developed methods). The latter category falls under the umbrella of the independent and generally accepted research stream framed as method engineering (ME) (Goldkuhl and Karlsson, 2020; Rossi et al., 2004), which can be defined as "[...] the engineering discipline to design, construct and adapt methods, techniques and tools for the development of information systems" (Brinkkemper, 1996, p. 276). Yet, as it is at the core of design science (DS) to create meaningful tools (i.e., knowledge artifacts) that contribute to both rigor and relevance (Hevner et al., 2004), previous work of Bucher and Winter (2008), Goldkuhl and Karlsson (2020), and Offermann et al. (2010)—among others—allocate ISDMs to DS and therefore declare it as a DS contribution type. By marrying both streams, Goldkuhl and Karlsson (2020) propose method engineering as design science (ME-DS). The research approach consists of eight design principles (DPs) that align with the two combined domains. To this end, method development follows a six-step procedure. As the ME-DS approach delivers transparency within the development process, practical utility of the resulting method and its generalizability, we opt for it. The ME-DS approach is depicted in Figure 3 and described below.

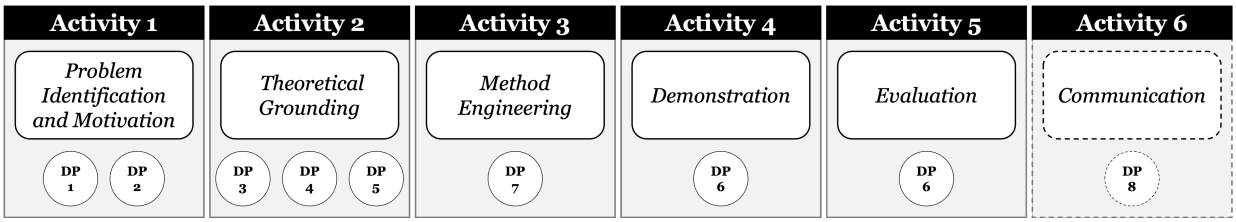


**Figure 3. Research Approach (Goldkuhl and Karlsson, 2020)**

To begin with, Goldkuhl and Karlsson (2020) emphasize that the development of a new ISDM must be justified and well-motivated from both—a practical (DP 1) and a scientific (DP 2) perspective. Since a literature review serves to pinpoint current research gaps, it is ideal to highlight the methodical deficiency regarding the application of synthetic data to advanced analytics (Goldkuhl and Karlsson, 2020). Next, it is "[...] necessary to infer what goals the new ISDM intends to achieve and how the new ISDM is expected to support solutions for the identified [information systems development] problems that have not previously been addressed in a satisfactory way" (Goldkuhl and Karlsson, 2020, p. 1248). Hence, we build on the initial motivation to define explicit values and goals of the method yet to be designed (DP 3) and likewise reveal underlying concepts (DP 4) as well as functional patterns (DP 5) derived from the knowledge base. The third activity comprises the actual method development stage, which either can produce an entirely new ISDM or customize an existing ISDM (Goldkuhl and Karlsson, 2020). In light of the well-established methods to drive advanced analytics, we pursue the latter strategy by borrowing methodical parts from existing ISDMs. We thereby ensure the transparency and concordance of the development process (DP 7). Activity three concludes with the new ISDM—*GenFlow*. As for the next step, Goldkuhl and Karlsson (2020) stress the importance to review the created ISDM regarding its practical applicability and usefulness (DP 6). Thus, in activity four, we demonstrate the general utility of *GenFlow* to incorporate synthetic data into advanced analysis projects by example. To this end, we apply *GenFlow* in retrospect to the article of Baul et al. (2021) from the above review. Likewise, in order to formally evaluate the usefulness of *GenFlow*, in activity five, we apply the method for the case example of employee attrition empirical. We conclude the procedure—as proposed by Goldkuhl and Karlsson (2020)—with the presentation of *GenFlow* to the target groups (DP 8), that is practitioners and researchers, by this very article. Thus, we refrain from detailing activity six separately in the following section but let the remainder of the article speak for itself.

## Design-oriented method engineering

### *Activity 1: Identify ISDM problem and motivate*

As is evident from our literature review, there is a considerable lack concerning practical guidance on the use of synthetic data in advanced analytics. Surprisingly, none of the reviewed articles using synthetic data follows an established method, which in turn strengthens the initial argument on missing guidelines in this regard. Moreover, since only a quarter details some general proceed, we draw the conclusion that there is an opportunity for a standardized method. Without a structured proceed practitioners might overlook favorable options to include synthetic data in advanced analyses. In addition, practitioners might struggle to clearly outline and communicate their endeavors, which could make the work inconclusive to additional interested parties. This also applies to academic scholars who are engaged with synthetic data and its application within various contexts. Without an appropriate method, it is hard to report a novel research endeavor—let alone conceive and pursue it in a structured manner with a holistic perspective on synthetic data. In that vein, none of the five popular methods outlined in detail above provides substantial utility to researchers interested in the use of synthetic data for advanced analytics. Given the aforementioned issues regarding the inclusion of synthetic data in advanced analytics, we consider the development of a designated method as distinctly beneficial. Thus, we continue with the development of *GenFlow* by following activity two.

### *Activity 2: ISDM theorizing*

In light of missing appropriate guidance, we hereby aim to conceive *GenFlow*—a synthetic data inclusive method for advanced analytics. As such, *GenFlow* requires a designated step for the consideration of synthetic data. Consequently, this stage must enable the reflection of the various options for synthetic data integration but be equally designed straightforward and conclusive. More specifically, the initial questions should be addressed, on where synthetic data might be useful and how it should be produced. Regarding the former, we draw on the entry points identified within the literature—these are (i) data acquisition, (ii) data analysis and preparation, and (iii) evaluation. In addition, the review suggests multiple uses of synthetic data in parallel. First, it is decisive to estimate if sufficient data are available. This leads to the conclusion of whether the first possible entry point is selected as such and if knowledge should be accessed accordingly.

Next up, as for the second entry point, analysts must decide whether the data at hand is appropriate for the pursued intent, that is in terms of quality and amount. Lastly, it could be necessary to introduce synthetic data at the evaluation stage to test a system's performance for an altered setting. Depending on the entry point, a different set of synthetic data generation techniques is available. As for the first entry point, data can only be drawn synthetically via prior knowledge embedded in the form of statistical distributions, simulations models or already developed and well-suited ML-models. This is quite different for the other two entry points where initial samples are actually available. Here, all types of generation techniques can be applied. Aside from statistical distributions and simulation models, this also includes data augmentation techniques and the novel ML-driven synthetic data generation approaches—but this time facilitated through data from prior stages.

Besides synthetic data considerations, *GenFlow* must be sufficiently flexible regarding the wide range of possible application contexts in advanced analytics. In this sense, it should correspond to the level of detail of the popular methods presented above, therefore representing a compilation of these methods. To begin with, we refer to this very comparison depicted in Figure 1. Here, we identify some *ex-ante* steps, which are prior to the actual analysis like e.g., *understanding*, *goal definition*, and *research question*. These are mostly concerned with delving into the interested field and getting an understanding before specifying a target or even multiple—if desired. The next phases, e.g., *data selection*, *data pre-processing* and *data transformation*, *explore*, or *modify*, are essentially focused on the (i) acquisition of data and its (ii) exploration and preparation. Thus, they can be grouped together into data-centric steps. Afterward, the modeling itself is instantiated. Depending on the approach this is described rather vague for *CRISP-DM*, *SEMMA*, and *BDAG* as opposed to the remaining. Thus, we borrow ideas from *KDD* and *SBEM* by deriving two main tasks—*choice of metrics and methods set* for ML modeling and its execution in a *training and validation* stage to closeup the modeling phase. Besides the consecutive assessment of the models' performance regarding the separate test data, these last steps include reasoning about the obtained results, getting insights into the algorithm's decisions, reporting the results, and possibly informing about the actual use of the developed application. These final steps fall under the umbrella of ex-post activities.

### *Activity 3: Method Engineering*

As suggested by Goldkuhl and Karlsson (2020), we choose a strategy from the ME research stream to conceive *GenFlow*. Here, we specifically opt for the *assembly-based process model for situational method engineering* as proposed by Ralyté et al. (2003) to select and likewise assemble method chunks. Given the material provided in activity two, we deduce *GenFlow*. The method with its ten steps and four meta-phases is depicted in Figure 4 and described thereafter. Notably, steps four, five, and eight reflect the entry points for synthetic data and are therefore marked accordingly.

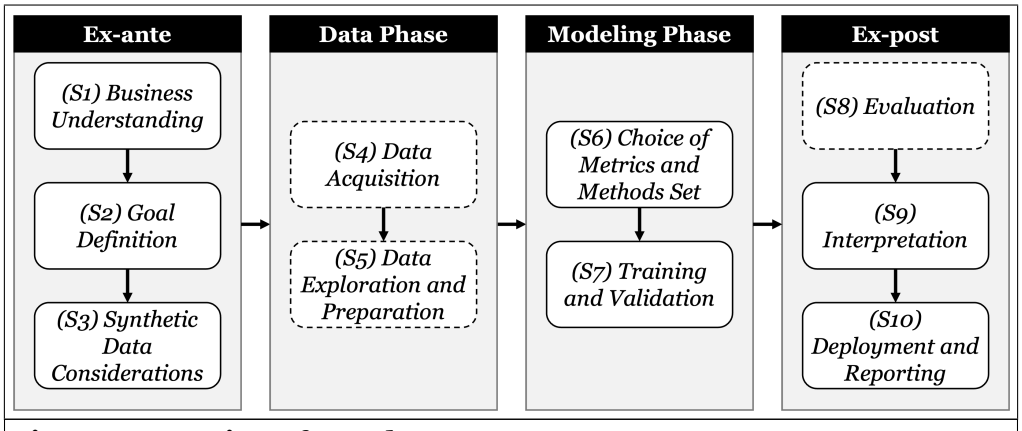


**Figure 4. Overview of GenFlow**

**(S1) Business Understanding**: Get an understanding of the subject matter with its various facets. Here, a wide range of techniques is thinkable in separation or combination such as e.g., interviews, questionnaires,

research, or reasoning.

**(S2) Goal Definition**: Define one or more clear and concise goal(s). This step is of utmost importance—not only to keep the focus but also to promote the endeavor accordingly to third parties.

**(S3) Synthetic Data Consideration**: Reason about the data required and in particular the utility of synthetic data. Topics of interest regarding the data are the (i) type(s) of modality, (ii) quality, (iii) set size(s), and finally (iv) characteristics. These considerations are to be carried out holistically in light of the three possible entry points. If, for example, data are perceived to be available in sufficiency regarding the aforementioned criteria, the subsequent step is not declared as an entry point for synthetic data, and vice versa. Similarly, the analyst(s) can estimate the necessity to introduce synthetic data at the later two stages to extend the initial data or evaluate for specific scenarios. In addition to the mere generation of synthetic data, its plausibility should be checked. In particular, there are instances where the use of artificial data may not be suitable at all like simulating or enriching survey data. In contrast, such data might prove useful only later when evaluating the case (e.g., handling class imbalance). Thus, it is worth putting the effort in its consideration.

**(S4) Data Acquisition**: Source the initial data. Data can either be acquired from internal or external sources or likewise a combination of both. This includes repositories and databases as well as knowledge pools. If this stage was considered an entry point earlier on, synthetic data are produced via knowledge. Here, statistical distributions or simulation models are applicable.

**(S5) Data Exploration and Preparation**: Understand and prepare the data purposefully. These processes are heavily intertwined and can be looped through multiple times. Through descriptive statistics and meaningful visualizations, a thorough understanding of the data is built. This indicates, whether data preparation is required and how the data should be manipulated. To this end, various techniques can be applied such as aggregation, transformation, cleaning, encoding, and missing value imputation (García et al., 2016; Kwak and Kim, 2017). If considered in (S3) or motivated through the data exploration itself, synthetic data can be introduced. This time, however, the analyst(s) can leverage both—prior knowledge or the preprocessed data. Thus, besides or in addition to sampling from statistical distributions or simulation models, one can use data augmentation techniques like e.g., SMOTE or some of the emerging ML-driven synthetic data generation techniques such as GANs or VAEs. To prevent any information leakage, the data should however be split into a dedicated train and test set as it is a common proceed in ML modeling before using the former for the data-based augmentation. If the synthetic data are generated in addition to the original data, it may be of interest for the subsequent evaluation (S8) to create several sets with different proportions of artificial data for comparison purposes.

**(S6) Choice of Metrics and Methods Set**: Decide on the metrics and methods relevant to achieve the goal. For example, when a prediction is made, the choice of the approach(es) boils down to the nature of the predicted class, that is, if it is numeric or categorical. While common metrics for the former are mean absolute error (MAE), mean squared error (MSE) or root mean squared error (RMSE) among many others, the performance for the latter is frequently measured via precision, recall or the harmonic mean of the two—the $F_1$-score (Chai and Draxler, 2014; Hossin and Sulaiman, 2015; Willmott and Matsuura, 2005). As for the ML methods, a wide range of techniques is applicable. These methods are frequently grouped into supervised, unsupervised, semi-supervised, and reinforcement learning algorithms (Sarker, 2021). Since the field can be rather complex in itself and is subject to rapid innovation, we direct the interested reader to further sources (e.g., Mahesh (2020) and Sarker (2021)).

**(S7) Training and Validation**: Train the chosen set of ML methods with respect to the metric(s) and perform a validation—if desired. To train the ML algorithms the dedicated training set is employed. Validation on the other hand refers to fine-tuning the models' hyperparameters. This is either done via a specific validation set created at the time of the train-test split or by means of cross-validation (e.g., Raschka (2018) and Zhai et al. (2020)).

**(S8) Evaluation**: Determine the models' performance regarding the designated metric(s). If previously considered, or desired at this point, synthetic data can now be introduced to provide a specific test set. Depending on the context, different types of approaches can be utilized to produce the data required for

the purpose of evaluation. In principle, all of the aforementioned options are viable. Whereas reasoning once more might provide novel scenarios to explore, guided data augmentation and ML-based generation enable further robustness checks. Now, in light of the pursued goal, the best-fitting model obtained should be chosen.

**(S9) Interpretation**: Investigate the patterns learned by the model. This helps to understand an ML algorithms' decisions and thus aids in terms of traceability and trust-worthiness as these are increasingly important aspects when it comes to the widespread adoption of ML-driven applications (Abdel-Karim et al., 2021; Bauer et al., 2021; Padmanabhan et al., 2022). In that vain, the effects of synthetic data inclusion in particular can be further understood via explainable artificial intelligence techniques by comparison with a model trained on original data.

**(S10) Deployment and Reporting**: Use and report the developed application. Given a sufficiently matured tool, its application to meet the formulated goal(s) can be initiated. This can include one-time or more frequent deployments as well as real-time constant use. In parallel, the results shall be communicated.

### *Activity 4: ISDM demonstration*

To demonstrate *GenFlow* in retrospect, we choose the research article of Baul et al. (2021) from the literature review. Therein, the authors work on detecting pedestrian flow from different directions at a traffic intersection. We briefly map the study to *GenFlow*.

**(S1) Business Understanding**: Predicting traffic flows can be notoriously hard, but likewise important—for example, to plan future infrastructure. To remedy this, novel ML methods can be used. However, as labeling adequate training data for ML involves a considerable amount of work and thus inevitably causes high costs, new ways to produce such data might be worth exploring.

**(S2) Goal Definition**: Against this backdrop, the authors propose to include synthetic data to detect pedestrian flow from different directions at traffic intersections through images.

**(S3) Synthetic Data Consideration**: To encounter the issue of few training data, the authors could extend existing data sets with traditional augmentation techniques like image cropping and rotating or even deploy ML-based approaches such as GANs or VAEs. Besides, simulation tools could be used to generate entirely new samples. From this, we derive two possible entry points—namely, the data acquisition, and data exploration and preparation stages.

**(S4) Data Acquisition**: Since this stage is marked as a possible entry point and in light of the present data scarcity, synthetic images might be of great support. For this purpose, simulation models like pre-configured game engines are proven to be useful. Thus, the authors deploy the GTA V video game engine by Rockstar Games to simulate image frames for a total of 75 crosswalk scenes.

**(S5) Data Exploration and Preparation**: Given the generated images, they, however, lack proper authenticity compared to real video scenes. Thus, again synthetic data techniques are employed to enhance the samples. More specifically, the authors use *CycleGAN*, a specific GAN-based architecture (Zhu et al., 2017), for image-to-image translation to create photo-realistic data for the previously acquired crosswalk scenes.

**(S6) Choice of Metrics and Methods Set**: As the predicted variable is not categorical but rather continuous, the authors employ the common metrics MAE and MSE. Now, to detect pedestrian flow, Baul et al. (2021) tune a proprietary two-branch convolutional neural network (CNN) to accommodate for the two relevant input signals from the video data—image frames and optical flow.

**(S7) Training and Validation**: The authors reveal details about the training procedure for the CNN and *CycleGAN* such as batch sizes, learning rates, and training epochs among many others.

**(S8) Evaluation**: To access the performance of the pedestrian flow detection system, Baul et al. (2021) use the few reference data sets at hand. The promising results indicate high utility for the use of synthetic data in the domain.

**(S9) Interpretation**: Beyond the mere results presentation, the authors miss the opportunity to provide insights into the algorithm's decisions, thereby creating traceability.

**(S10) Deployment and Reporting**: By publishing their article, Baul et al. (2021) report to scholars as well as practitioners.

This example effectively demonstrates the applicability of *GenFlow* for an existing advanced analytics case in short.

### *Activity 5: Formal evaluation*

Now, to illustrate the utility of *GenFlow* by means of formal evaluation, we choose the case example of employee attrition, which resembles a common problem for organizations. To this end, we again follow the conclusive method step-by-step.

**(S1) Business Understanding**: To begin with, we gain a broad conception of the subject matter. Employee attrition can pose substantial risks to organizations and their stakeholders. First and foremost, the costs associated with employee attrition can put organizations at direct financial risk. A current project might be stalled for an indefinite time leading to the late or even non-fulfillment of associated objectives and thus may provoke disapproval of the stakeholders. Apart from the financial risks posed by attrition, expertise also dwindles as employees leave. This becomes particularly threatening to organizations if they compete in a market and know-how diffuses to rivals (Kumar and Yakhlef, 2016). The list of negative effects associated with employee attrition goes on and we refer the interested reader to Kumar and Yakhlef (2016) and Makawatsakul and Kleiner (2003). To take it to the extreme, such issues could ultimately evoke a downward spiral of ongoing employee and knowledge loss coupled with stakeholder reservations leading to an existential crisis in the organization itself.

**(S2) Goal Definition**: Against this backdrop, it is of fundamental importance for organizations to anticipate employee attrition by implementing countermeasures. Hence, we propose to predict employee attrition cases early on to address these adequately. To this end, we follow the predictive analytics paradigm to detect attrition candidates with high certainty.

**(S3) Synthetic Data Consideration**: As for the anchor point of the present article, we reason about the possibilities to introduce synthetic data to successfully predict employee attrition. In general, we could make use of past information on individuals regarding attrition itself as well as several job, education and demographics related characteristics. Depending on the initial data set in terms of size and quality, we may benefit from generating further samples. While we put emphasis on the predictive power of the application in this first development cycle, we, at this point however, do not plan an altered test scenario in (S8). Thus, we do not consider synthetic data for evaluation. Accordingly, we only declare the following two steps (S4) and (S5) as possible entry points for synthetic data.

**(S4) Data Acquisition**: Since suitable data are freely available on the Kaggle Plattform[1] and for the sake of reproducibility, we refrain from generating our own proprietary data set. Thus, we acquire the data with the binary target variable *attrition* and several employee-related characteristics.

**(S5) Data Exploration and Preparation**: Regarding the data analysis, we first gain a broad conception of the size of the data as well as the distributions of the variables. From this, we draw a handful of conclusions. With its 1.470 rows and 25 attributes the data set is rather small and thus possibly challenging for data-hungry ML models. This reinforces the initial consideration to use synthetic data for the purpose of data augmentation. However, prior to data generation, we perform some basic tasks such as data formatting and correlation analysis to reduce dimensions and likewise computing capacity. Due to the few samples available, we split the data into a distinct train and test set with a 70-30 ratio. Given the pre-processed data, we employ *CopulaGAN*, a GAN-based architecture for tabular synthetic data generation provided by the Synthetic Data Vault[2]. To this end, we develop the *CopulaGAN* model based on the train set with 10.000 epochs and a batch size of 50. To check the integrity of the generated data, we first use the built-in function provided with the *CopulaGAN* implementation to measure the overall similarity via Chi-Squared between the synthetic and the original data in terms of the learned data distributions. This yields a score of almost 0.95 which indicates the high adequacy of the produced data. To dig deeper into the characteristics of the

[1]https://www.kaggle.com/datasets/patelprashant/employee-attrition
[2]https://docs.sdv.dev/sdv/single-table-data/modeling/synthesizers/copulagansynthesizer

generated data, we refer to some notable examples provided in Figure 5. Here, the boxplots corroborate this presumed similarity of the imitated for the real data since the median values, as well as the interquartile ranges, are rather close to each other. For instance, the generated data captures the trend of younger employees being more prone to attrition than others. Regarding the multiple bivariate kernel density estimate plots depicted at the bottom, we likewise observe similar distributions for the analyzed variables with high densities nearby and well-replicated trends. After generation, we lastly perform the so-called one-hot encoding method to convert categorical variables into binary features.

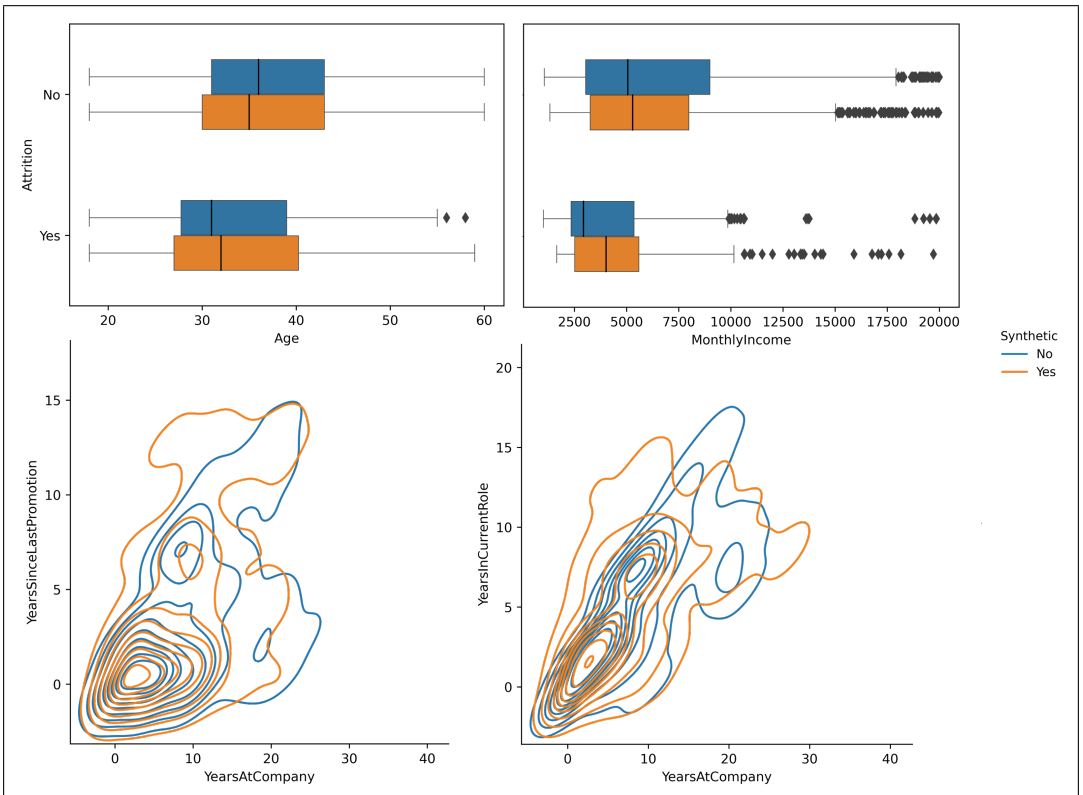


**Figure 5. Comparative Descriptive Statistics for Original and Artificial Data**

**(S6) Choice of Metrics and Methods Set**: To keep the illustrative case short, as opposed to the common recommendation, we refrain from selecting various sets of metrics and methods. Taking adequate countermeasures regarding employee attrition might be expensive and time-consuming in itself. Thus, it might be reasonable to identify employees who are prone to leave with high certainty. Consequently, precision is the go-to metric for the classification algorithm. With respect to the ML approach, we choose extreme gradient boosting (XGB) as it is currently considered superior regarding tabular data (Shwartz-Ziv and Armon, 2022; Grinsztajn et al., 2022).

**(S7) Training and Validation**: To train the XGB classifier, we employ the original train set as well as several synthetically extended versions of it for the sake of comparison. Due to few original training data, we do not leverage a separate validation set but perform a ten-fold cross-validation instead. Given the determined set of hyperparameters, we finally obtain the classifier model to detect employee attrition.

**(S8) Evaluation**: Now, with regard to predictive power, we assess the ML models' performance on the separate test data. Without additional samples, the classifier is rather incapable of detecting potential attrition cases among the employees with high certainty (i.e., 0.54). However, if synthetic data are provided, this impression is changed. The more data, the more capable the XGB classifier to validly detect attrition candidates. With 100% more train data, precision rises to 0.64 for the same test set. This trend is continued for exponentially more data (i.e., 1.000%: 0.76; 10.000%: 0.85; 100.000%: 0.87) while recall fluctuates.

**(S9) Interpretation**: To explore the models' decisions in light of the inclusion of synthetic data, we employ the well-received *SHAP* library by Lundberg and Lee (2017) as their approach is applicable to any black-box ML model through the use of Shapley values[3]. In Figure 6, we compare a model exceptionally trained on the original data (cf. left side) with a model trained on synthetically augmented data (cf. right side). Notably the data set underlying the latter model is half original half synthetic. As for the illustration, the 20 most important features for the prediction are listed in descending order. While the impacts of the feature values on the prediction outcome are rather similar (i.e., *OverTime*, *StockOptionLevel*, *MonthlyIncome*), there are a few interesting dissimilarities such as *YearsAtCompany* or *PercentSalaryHike*. Likewise, the variables' orders are different such that some variables do not appear on the opposite plot.

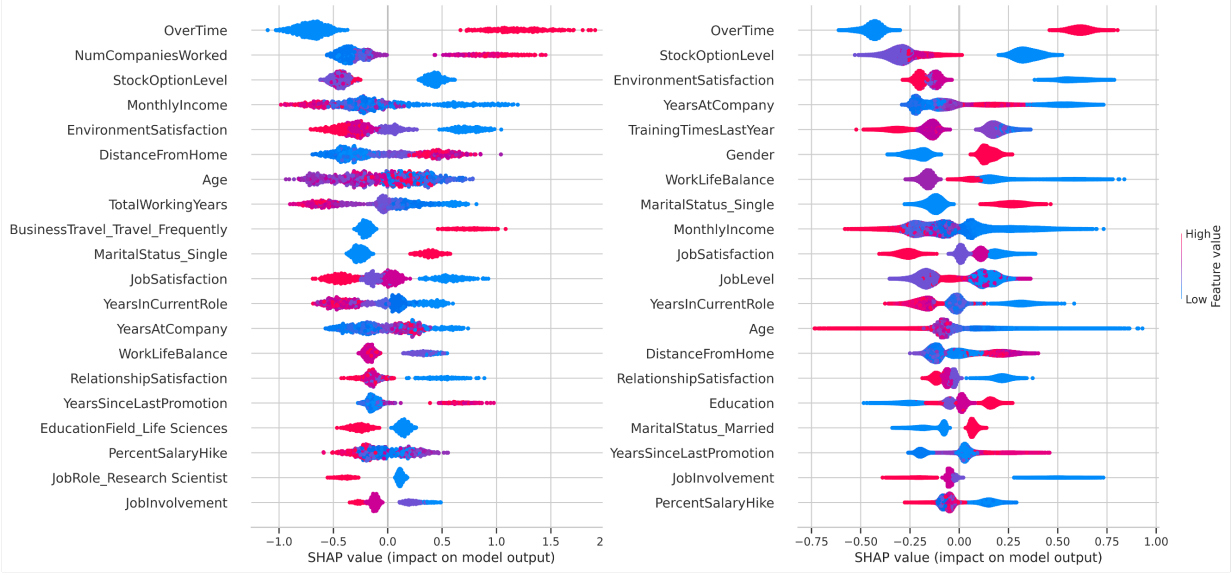


**Figure 6. Comparative SHAP Analysis**

**(S10) Deployment and Reporting**: As this illustrative example is merely intended for demonstration purposes, we neither deploy the advanced analytics application nor do we report about it in more depth.

The analysis once again highlights the benefits of *GenFlow* to guide the inclusion of synthetic data within the analyses. More specifically, attention is paid to the utility of synthetic data in light of small data.

## Conclusions, Limitations and Outlook

With the design of *GenFlow*, a synthetic data inclusive method for advanced analytics, we contribute to the body of knowledge. In addition, by comparing often-cited methods we uncover similarities and differences between them. These insights can be particularly useful to researchers also concerned with method engineering for other emerging technologies (e.g., for explainable artificial intelligence or federated learning). The literature review unveils a significant lack of methodical guidance for many of the investigated articles which in turn can be viewed as a general plea for more research in the field.

Thus, the present article holds several valuable implications. First and foremost, the conclusive method *GenFlow* is readily available to both, researchers and practitioners, who are engaged with advanced analytics development. To this end, the method puts particular emphasis on the possible pertinence of synthetic data for any advanced analytics project. In effect, users pay special attention to the consideration of whether the inclusion of synthetic data is possible and beneficial. Moreover, they are provided with guidance on where the data can be introduced and what means it takes to produce it (i.e., knowledge or pre-existing data). This not only assists reasoning about synthetic data and in this respect the communication about its

[3]https://shap.readthedocs.io/en/latest/index.html

use for future projects, but also opens avenues to re-explore past endeavors. Apart from the method itself, the demonstration and formal evaluation illustrate how *GenFlow* can be applied. Furthermore, the formal evaluation indicates the high utility of the use of synthetic data.

While the research sets out to design a novel method for advanced analytics considering where and how to employ synthetic data, it barely touches the surface of the wide range of possible data production techniques. Accordingly, a detailed description of the various approaches to synthetic data generation for the three entry points identified may provide high utility to the users of *GenFlow* in the future. Another limitation is the specific choice of two short illustrative examples which do not adequately reflect the wide applicability of *GenFlow*. Hence, the benefits stemming from *GenFlow* are yet to be explored in a broader sense. For example, the method could be employed to analyze the effects of synthetic data on ML performance in more depth as well as on privacy preservation. Moreover, the research proceed is limited in itself, that is, regarding the selection of popular methods, the chosen search query and databases as for the literature review, and the designated research approach to conceiving the method. These limiting aspects leave us to conclude the need for further research in that regard. In essence, we postulate the following promising research directions to be addressed in the future:

- **Artificial Data Toolbox**: Which techniques are available to generate synthetic data? What are the associated advantages and disadvantages of the respective approaches? At which entry points of *GenFlow* are the individual approaches applicable?
- **Values of Artificial Data**: What are the values to be expected from the use of synthetic data generation for advanced analytics? Are there any trade-offs between these? How are the effects linked to the generation technique and *GenFlow* entry point?
- **Adoption of Artificial Data**: Does *GenFlow* contribute to the adoption of synthetic data for advanced analytics and thereby accelerate the impact of ML?

With *GenFlow* designed and readily available, we hope to remedy current barriers to the implementation and adoption of synthetic data both in research and in practice. Thereby, we especially intend to open doors for novel endeavors previously restricted by limited data and encourage the consideration of using synthetic data in general.

## References

Abdel-Karim, B. M., Pfeuffer, N., and Hinz, O. (2021). "Machine learning in information systems-a bibliographic review and open research issues," *Electronic Markets* (31:3), pp. 643–670.

Azevedo, A. and Santos, M. F. (2008). "KDD, SEMMA and CRISP-DM: a parallel overview," in, pp. 182–185.

Barton, D. and Court, D. (2012). "Making Advanced Analytics Work For You," *Harvard Business Review* (90:10), pp. 78–83.

Bauer, K., Hinz, O., Aalst, W. van der, and Weinhardt, C. (2021). "Expl(AI)n it to me–explainable AI and information systems research," *Business & Information Systems Engineering* (63:2), pp. 79–82.

Bauer, M., Dinther, C. van, and Kiefer, D. (2020). "Machine learning in SME: an empirical study on enablers and success factors," in *AMCIS,* pp. 1–10.

Baul, A., Kuang, W., Zhang, J., Yu, H., and Wu, L. (2021). "Learning to Detect Pedestrian Flow in Traffic Intersections from Synthetic Data," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC),* pp. 2639–2645.

Bengio, Y., Courville, A. C., and Vincent, P. (2012). "Unsupervised feature learning and deep learning: A review and new perspectives," *arXiv* (1206.5538:2665), p. 2012.

Berente, N., Gu, B., Recker, J., and Santhanam, R. (2021). "Managing artificial intelligence," *MIS Quarterly* (45:3), pp. 1433–1450.

Berger, M. L. and Doban, V. (2014). "Big data, advanced analytics and the future of comparative effectiveness research," *Journal of Comparative Effectiveness Research* (3:2), pp. 167–176.

Beynon-Davies, P. and Williams, M. D. (2003). "The diffusion of information systems development methods," *The journal of strategic information systems* (12:1), pp. 29–46.

Bose, R. (2009). "Advanced analytics: opportunities and challenges," *Industrial Management & Data Systems* (109:2), pp. 155–172.

Brinkkemper, S. (1996). "Method engineering: engineering of information systems development methods and tools," *Information and Software Technology* (38:4), pp. 275–280.

Bucher, T. and Winter, R. (2008). "Dissemination and Importance of the "Method" Artifact in the Context of Design Research for Information Systems," in *Proceedings of the Third International Conference on Design Science Research in Information Systems and Technology (DESRIST 2008),* V. Vaishnavi and R. Baskerville (eds.). Atlanta, GA: Georgia State University, 2008, pp. 39–59.

Bue, B. D. and Merényi, E. (2010). "Using spatial correspondences for hyperspectral knowledge transfer: Evaluation on synthetic data," in *Hyperspectral Image and Signal Processing,* pp. 1–4.

Bzdok, D., Krzywinski, M., and Altman, N. (2017). "Machine learning: a primer," *Nature methods* (14:12), p. 1119.

Catley, C., Smith, K., McGregor, C., and Tracy, M. (2009). "Extending CRISP-DM to incorporate temporal data mining of multidimensional medical data streams: A neonatal intensive care unit case study," in *2009 22nd IEEE International Symposium on Computer-Based Medical Systems,* pp. 1–5.

Chai, T. and Draxler, R. R. (2014). "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature," *Geoscientific model development* (7:3), pp. 1247–1250.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., et al. (2000). "CRISP-DM 1.0: Step-by-step data mining guide," *SPSS Inc.* (9), p. 13.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research* (16), pp. 321–357.

Chen, H. M., Chen, H. C., Chen, C. C. C., Chang, Y. C., Wu, Y. Y., Chen, W. H., Sung, C. C., Chai, J. W., and Lee, S. K. (2021a). "Comparison of Multispectral Image-Processing Methods for Brain Tissue Classification in BrainWeb Synthetic Data and Real MR Images," *BioMed Research International* (2021).

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F., and Mahmood, F. (2021b). "Synthetic data in machine learning for medicine and healthcare," *Nature Biomedical Engineering* (5:6), pp. 493–497.

Cooper, H. M. (1988). "Organizing knowledge syntheses: A taxonomy of literature reviews," *Knowledge in society* (1:1), pp. 104–126.

Damer, N., Saladie, A. M., Braun, A., and Kuijper, A. (2018). "Variational Autoencoders: A Brief Survey," in *IEEE 9th International Conference on Biometrics Theory, Applications and Systems,* pp. 1–9.

Davenport, T. H. (2018). "From analytics to artificial intelligence," *Journal of Business Analytics* (1:2), pp. 73–80.

Delen, D. and Zolbanin, H. M. (2018). "The analytics paradigm in business research," *Journal of Business Research* (90), pp. 186–195.

Dwivedi, Y. K., Wastell, D., Laumer, S., Henriksen, H. Z., Myers, M. D., Bunker, D., Elbanna, A., Ravishankar, M., and Srivastava, S. C. (2015). "Research on information systems failures and successes: Status update and future directions," *Information Systems Frontiers* (17:1), pp. 143–157.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). "From data mining to knowledge discovery in databases," *AI magazine* (17:3), pp. 37–37.

Franks, B. (2013). "Taming the Big Data tidal wave: Finding opportunities in huge data streams with advanced analytics," *John Wiley & Sons* (43), pp. 1–336.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., and Herrera, F. (2016). "Big data preprocessing: methods and prospects," *Big Data Analytics* (1:1), pp. 1–22.

Goldkuhl, G. and Karlsson, F. (2020). "Method engineering as design science," *Journal of the Association for Information Systems* (21:5), pp. 1237–1278.

Goodfellow, I. J. (2017). "NIPS 2016 Tutorial: Generative Adversarial Networks," *arXiv* (1701.00160).

Grand View Research, Inc. (2022). *Advanced Analytics Market Size Report, 2022-2030*. https://www.grandviewresearch.com/industry-analysis/advanced-analytics-market. (Accessed on 04/27/2023). 2022.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). "Why do tree-based models still outperform deep learning on typical tabular data?," in *Advances in Neural Information Processing Systems,* S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.). Vol. 35. Curran Associates, Inc., pp. 507–520.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R., and Friedman, J. (2009). "Unsupervised learning," *The elements of statistical learning: Data mining, inference, and prediction* (2), pp. 485–585.

Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 75–105.
Hossin, M. and Sulaiman, M. N. (2015). "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process* (5:2), pp. 1–11.
James, S., Harbron, C., Branson, J., and Sundler, M. (2021). "Synthetic data use: exploring use cases to optimise data utility," *Discover Artificial Intelligence* (1:1), pp. 1–13.
Jöhnk, J., Weißert, M., and Wyrtki, K. (2021). "Ready or not, AI comes—an interview study of organizational AI readiness factors," *Business & Information Systems Engineering* (63:1), pp. 5–20.
Jordan, M. I. and Mitchell, T. M. (2015). "Machine learning: Trends, perspectives, and prospects," *Science* (349:6245), pp. 255–260.
Kingma, D. P. and Welling, M. (2019). "An introduction to variational autoencoders," *Foundations and Trends in Machine Learning* (12:4), pp. 307–392.
Kokubo, Y., Asada, S., Maruyama, H., Koide, M., Yamamoto, K., and Suetsugu, Y. (2021). "Removing Raindrops from a Single Image using Synthetic Data," in *2020 25th International Conference on Pattern Recognition (ICPR),* pp. 2081–2088.
Kowalczyk, P., Welsch, G., and Thiesse, F. (2022). "Towards a Taxonomy for the Use of Synthetic Data in Advanced Analytics," *arXiv* (2212.02622).
Kumar, N. and Yakhlef, A. (2016). "Managing business-to-business relationships under conditions of employee attrition: A transparency approach," *Industrial Marketing Management* (56), pp. 143–155.
Kwak, S. K. and Kim, J. H. (2017). "Statistical data preparation: management of missing values and outliers," *Korean journal of anesthesiology* (70:4), p. 407.
Lundberg, S. M. and Lee, S.-I. (2017). "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems,* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.). Vol. 30. Curran Associates, Inc., pp. 1–10.
Mahesh, B. (2020). "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR)* (9), pp. 381–386.
Makawatsakul, N. and Kleiner, B. H. (2003). "The effect of downsizing on morale and attrition," *Management Research News* (26:2), pp. 52–62.
Mariscal, G., Marban, O., and Fernandez, C. (2010). "A survey of data mining and knowledge discovery process models and methodologies," *The Knowledge Engineering Review* (25:2), pp. 137–166.
Mitchell, T. M. (1997). "Artificial neural networks," *Machine learning* (45), pp. 81–127.
Müller, O., Junglas, I., Brocke, J. v., and Debortoli, S. (2016). "Utilizing big data analytics for information systems research: challenges, promises and guidelines," *European Journal of Information Systems* (25:4), pp. 289–302.
Nikolenko, S. I. (2021). *Synthetic data for deep learning,* Springer.
Nunamaker Jr, J. F., Chen, M., and Purdin, T. D. (1990). "Systems development in information systems research," *Journal of management information systems* (7:3), pp. 89–106.
Offermann, P., Blom, S., Levina, O., and Bub, U. (2010). "Proposal for components of method design theories," *Business & Information Systems Engineering* (2:5), pp. 295–304.
Padmanabhan, B., Fang, X., Sahoo, N., and Burton-Jones, A. (2022). "Machine Learning in Information Systems Research," *MIS Quarterly* (46:1), pp. 3–19.
Patki, N., Wedge, R., and Veeramachaneni, K. (2016). "The synthetic data vault," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA),* IEEE, pp. 399–410.
Piatetsky, G. (2014). *CRISP-DM, still the top methodology for analytics, data mining, or data science projects*. https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html. 2014.
Rajotte, J.-F., Bergen, R., Buckeridge, D. L., El Emam, K., Ng, R., and Strome, E. (2022). "Synthetic data as an enabler for machine learning applications in medicine," *iScience* (25:11), p. 105331.
Ralyté, J., Deneckère, R., and Rolland, C. (2003). "Towards a generic model for situational method engineering," in *International Conference on Advanced Information Systems Engineering,* Springer, pp. 95–110.
Ram, S. and Goes, P. (2021). "Provocations: Focusing on Programmatic High Impact Information Systems Research, Not Theory, to Address Grand Challenges," *MIS Quarterly* (45:1), pp. 479–483.
Raschka, S. (2018). "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning," *arXiv* (1811.12808).

Rohanizadeh, S. S. and Bameni, M. M. (2009). "A proposed data mining methodology and its application to industrial procedures," (4), pp. 37–50.

Rossi, M., Ramesh, B., Lyytinen, K., and Tolvanen, J.-P. (2004). "Managing evolutionary method engineering by method rationale," *Journal of the association for information systems* (5:9), p. 12.

Russo, N. L. and Stolterman, E. (2000). "Exploring the assumptions underlying information systems methodologies: Their impact on past, present and future ISM research," *Information technology & people* (13:4).

Sarker, I. H. (2021). "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science* (2:3), pp. 1–21.

SAS Institute Inc. (2002). *SAS Enterprise Miner: Uncover Gems of Information*. https://www.sas.com/en_us/software/enterprise-miner.html. (Accessed on 04/27/2023). 2002.

SAS Institute Inc. (2017). *SAS Help Center: Introduction to SEMMA*. https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm. (Accessed on 04/26/2023). 2017.

Shafique, U. and Qaiser, H. (2014). "A comparative study of data mining process models (KDD, CRISP-DM and SEMMA)," *International Journal of Innovation and Scientific Research* (12:1), pp. 217–222.

Sharma, A., Jindal, N., and Rana, P. S. (2020). "Potential of generative adversarial net algorithms in image and video processing applications– a survey," *Multimedia Tools and Applications* (79:37-38), pp. 27407–27437.

Shmueli, G. and Koppius, O. R. (2011). "Predictive analytics in information systems research," *MIS quarterly* (35:1), pp. 553–572.

Shwartz-Ziv, R. and Armon, A. (2022). "Tabular data: Deep learning is not all you need," *Information Fusion* (81), pp. 84–90.

Visani, G., Graffi, G., Alfero, M., Bagli, E., Capuzzo, D., and Chesani, F. (2022). "Enabling Synthetic Data adoption in regulated domains," *arXiv* (2204.06297).

Vom Brocke, J., Simons, A., Niehaves, B., Niehaves, B., Reimer, K., Plattfaut, R., and Cleven, A. (2009). "Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process," *17th European Conference on Information Systems* (161).

Walonoski, J., Klaus, S., Granger, E., Hall, D., Gregorowicz, A., Neyarapally, G., Watson, A., and Eastman, J. (2020). "Synthea™ Novel coronavirus (COVID-19) model and synthetic data set," *Intelligence-Based Medicine* (1-2), p. 100007.

Watson, J., Hutyra, C. A., Clancy, S. M., Chandiramani, A., Bedoya, A., Ilangovan, K., Nderitu, N., and Poon, E. G. (2020). "Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers?," *JAMIA Open* (3:2) 2020, pp. 167–172.

Wei, R. and Mahmood, A. (2021). "Recent Advances in Variational Autoencoders with Representation Learning for Biomedical Informatics: A Survey," *IEEE Access* (9), pp. 4939–4956.

Willmott, C. J. and Matsuura, K. (2005). "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate research* (30:1), pp. 79–82.

Wirth, R. and Hipp, J. (2000). "CRISP-DM: Towards a standard process model for data mining," in *International conference on the practical applications of knowledge discovery and data mining,* pp. 29–40.

Yannikos, Y., Winter, C., and Schneider, M. (2012). "Synthetic Data Creation for Forensic Tool Testing: Improving Performance of the 3LSPG Framework," in *2012 Seventh International Conference on Availability, Reliability and Security,* pp. 613–619.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., and Shi, L. (2020). "Applying machine learning in science assessment: a systematic review," *Studies in Science Education* (56:1), pp. 111–151.

Zhang, Y., Zhang, G., Han, X., Wu, J., Li, Z., Li, X., Xiao, G., Xie, H., Fang, L., and Dai, Q. (2023). "Rapid detection of neurons in widefield calcium imaging datasets after training with synthetic data," *Nature Methods* (1), pp. 1–8.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *2017 IEEE International Conference on Computer Vision (ICCV),* pp. 2242–2251.