Data for Digital Forensics: Why a Discussion on "How Realistic is Synthetic Data" is Dispensable

THOMAS GÖBEL and HARALD BAIER, Research Institute CODE, University of the Bundeswehr Munich, Germany
FRANK BREITINGER, School of Criminal Justice, University of Lausanne, Switzerland

Digital forensics depends on data sets for various purposes like concept evaluation, educational training, and tool validation. Researchers have gathered such data sets into repositories and created data simulation frameworks for producing large amounts of data. Synthetic data often face skepticism due to its perceived deviation from real-world data, raising doubts about its realism. This paper addresses this concern, arguing that there is no definitive answer. We focus on four common digital forensic use cases that rely on data. Through these, we elucidate the specifications and prerequisites of data sets within their respective contexts. Our discourse uncovers that both real-world and synthetic data are indispensable for advancing digital forensic science, software, tools, and the competence of practitioners. Additionally, we provide an overview of available data set repositories and data generation frameworks, contributing to the ongoing dialogue on digital forensic data sets' utility.

CCS Concepts: • **Applied computing → Computer forensics;**

Additional Key Words and Phrases: Digital forensic corpora, Data sets; Real-world data, Synthetic data; Data usage, Data synthesis, Types of data, Use cases, Realistic data, Data simulation frameworks

ACM Reference format:
Thomas Göbel, Harald Baier, and Frank Breitinger. 2023. Data for Digital Forensics: Why a Discussion on "How Realistic is Synthetic Data" is Dispensable. *Digit. Threat. Res. Pract.* 4, 3, Article 38 (October 2023), 18 pages.
https://doi.org/10.1145/3609863

1 INTRODUCTION

As technology becomes more pervasive in modern life, the number of digital devices that play a role in criminal investigations is growing rapidly. It is likely that the number of cases requiring digital forensic analysis will further increase in the future. It is also likely that each case will require analysis of an increasing number of devices, such as computers, smartphones, tablets, IoT devices, wearables, and so on. The prevalence of new digital evidence sources poses new and challenging problems for forensic examiners in identifying, acquiring, storing, and analyzing the ever-changing evidence [1].

Hence like every scientific discipline, the digital forensics field relies on data in many ways, for instance, to support digital forensics research in evaluating a new concept or to enable education and training for getting experience with an IT forensic soft- or hardware or to train skills. Although there is a great need for data sets in the digital forensics domain, the main problem is that data sets are often not available in sufficient quantity and

Authors' addresses: T. Göbel and H. Baier, Research Institute CODE, University of the Bundeswehr Munich, Carl-Wery-Straße 18, 81739 München, Germany; e-mails: thomas.goebel@unibw.de, harald.baier@unibw.de; F. Breitinger, Faculty of Law, Criminal Sciences and Public Administration, ESC | School of Criminal Sciences, University of Lausanne, CH 1015 Lausanne-Dorign, Switzerland; e-mail: frank.breitinger@unil.ch.

2576-5337/2023/10-ART38 $15.00
https://doi.org/10.1145/3609863

quality to digital forensic practitioners and scientists, especially for evidence that may provide new, previously unknown artifacts. This is especially true concerning real-world samples, which are often not published due to privacy issues or secrecy reasons (e.g., in law enforcement) or intellectual property rights as stated by Grajeda et al. [2]. The same authors analyzed 715 conference and journal papers published from 2010 to 2015 in terms of the utilization of data sets and if the respective data sets were published. Grajeda et al. [2] state that (1) many researchers create their data sets manually, (2) data sets are mostly not shared at all, and (3) there is a lack of standardized, labeled data sets. The main problem here is that the results of digital forensic research are often not repeatable and reproducible, i.e., the techniques developed cannot be validated by others unless the data sets are shared or different data sets have to be used [3].

This large gap between publicly available data sets and actual needs is not specific to the field of digital forensics, but is widespread in the cybersecurity field and beyond, as discussed in Abt and Baier [4]. Therefore, the various scientific disciplines keep an eye out to come up with data sets besides those originating from real-world cases. A common approach besides real-world data is the use of *synthetic* data sets. In the past, synthetic data was often created by generation frameworks such as `ForTrace` [5], `TraceGen` [6], `FADE` [7], `hystck` [8], and `Eviplant` [9]. Synthetic data, however, is still often not considered a factual substitute for missing real-world data sets. The typical objection to synthetic data is its presumed deviation from real-world data, resulting in an assumed questionable utility.

Hence, the question that is usually raised when synthetic data sets or data synthesis frameworks are presented and discussed is: *How realistic is synthetic data?* In this article, we address this question and conclude that both types of data (i.e., real-world data and synthetic data) are needed in the digital forensics domain. Furthermore, we provide arguments that both types of data should not be directly compared in general. Nor can it be said across the board that one data type is better than the other since the utility of a data set depends on its particular use case.

In summary, this article provides the following contributions:

- A detailed description of four common use cases, highlighting the strengths and weaknesses of the different data set types.
- These use cases serve as the basis for an enumeration and discussion of data characteristics and requirements in the respective scope.
- A brief overview of the various data sets and data set generation frameworks available.

The remainder of this article is structured as follows: Section 2 introduces the terminology used in this paper, followed by the presentation of the overall importance of digital forensic corpora in Section 3. The same Section then outlines a brief overview of available data set repositories and data set generation frameworks. Section 4 then provides a detailed description of our four common use cases, outlining the strengths and weaknesses of the different data set types. The use cases serve as the basis for a discussion of specific data characteristics in Section 5. Lastly, we provide our conclusion in Section 6.

2 TERMINOLOGY AND DATA TYPES

Over the past two decades, several articles discussed data types and their great importance to digital forensics. For this article, we will use the following terms and definitions which align with the terms and definitions from Breitinger and Jotterand [10].

2.1 Metadata

According to Carrier [11, p. 130], metadata is data that describes files. It contains data such as the exact location where the file content is stored, the size of the file, and the times and dates a file was last read or written to. Consequently, metadata is data about data, which means that metadata always has a descriptive meaning. There are different types of metadata, which can be divided into two categories: (1) internal metadata and (2) external

metadata [12]. For example, external metadata stores additional information in various data structures stored elsewhere, such as the NTFS file system storing information about a particular file/directory within the MFT, whereas internal metadata is stored directly in the digital trace of interest, such as data within files like EXIF data in a JPG or metadata like the owner and date of image creation in a forensic E01 image. Metadata, which is the focus of this paper, is specifically additional information (i.e., external metadata) about a data set, such as descriptive data, that a data set should ideally contain for easier reproducibility and reusability.

Consequently, metadata in the context of digital corpora is essentially a supplement to the actual data, as it documents additional information about the data sets that helps other researchers understand the actual contribution of the data sets. In particular, metadata should enable other users to accurately replicate previous research results and successfully reuse the data set. Metadata of digital corpora provides associated information, such as an appropriate identifier, the origin of a data set (e.g., information about the creator/owner, web sources, contacts, maintenance purposes, etc.), licenses and permissions, date/time, and the conditions under which the data set was created (e.g., for manually created data sets, explanations of the exact steps performed, hardware/software used, network/system settings configured, etc.). Such additional information is of course missing for unknown data, e.g., for real-world data collected during a forensic investigation and not yet analyzed.

2.2 Ground Truth

As pointed out by Garfinkel et al. [3], a reference set of representative corpora improves the scientific evaluation of forensic methods beyond the obvious benefits of providing test data and allows direct comparison of different approaches. Indeed, it allows for the ground truth to be established. Ground truth data can then serve as a baseline for controlled studies [13] or evaluate the success of new tools and methods using objective metrics [3].

In digital forensics, the term *ground truth data* is used for data whose content is well known and understood by the community, i.e., the exact information or the actual digital traces to be discovered in a data set during the investigation are well documented. While the ground truth is usually fully known for synthetically generated data (i.e., generated by software or a tool) or at least partially known for manually generated data (i.e., created by a human), this is not the case for real-world samples. Therefore, as long as the data set is intentionally created, the researcher typically knows the underlying data, which means there are no surprises, discrepancies, or inconsistencies in synthetic data. Of course, it is also possible to carefully analyze real-world data and accurately document or label its content (as is done in a forensic investigation), but depending on the size and complexity of the data set, the assessment may miss important artifacts and can be difficult and time-consuming.

2.3 Types of Data Sets

Related work has also discussed different categories or types of data. For instance, Garfinkel et al. [3] differentiate between (1) test data, (2) sampled data, (3) realistic data, (4) real and restricted data, and (5) real but unrestricted data. On the other hand, Grajeda et al. [2] use the terms (1) experiment generated, (2) user generated, and (3) computer generated data sets.

More recent work by Breitinger and Jotterand [10] revisited this area and presents a novel taxonomy as depicted in Figure 1. The authors first distinguish between *Synthetic data* (describing data that is created by software with a certain degree of autonomy) and *Human data* (describing data that is the result of a human interacting with a system in any way). This subdivision is obvious because data sets are generated either in some way automatically by any software and tools or manually by humans. The further subdivision is more granular and distinguishes *Random data*, *Rule-based data generation*, *(Computer) Simulated data*, and *AI-generated data* (all within the synthetic category) from *(Human) Simulated data*, *Experimental data*, and *Real-world data* (all within the Human category).

Basically, the five categories *Rule-based data generation*, *(Computer) Simulated data*, *(Human) Simulated data*, *Experimental data*, and *Real-world data* are of interest for this work, as they reflect the most commonly used techniques for manual or automatic data generation in the field of digital forensics. However, we can limit ourselves

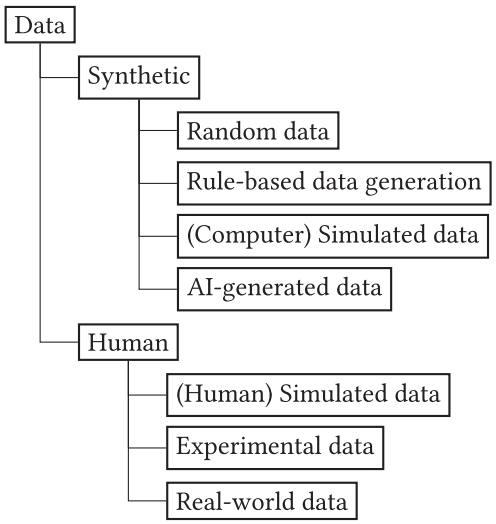

Fig. 1. Visual representation of the taxonomy proposed by Breitinger and Jotterand [10]; last/lowest level was removed for better readability.

to three categories, since the differences here are only in the details. For example, *(Computer) Simulated data* is similar to *Rule-based data generation* with the exception that the former simulates the behavior using system tools, while the latter describes locally generated samples without performing simulations, but still synthesizes the data through the use of rules. Similarly, *(Human) Simulated data* is similar to *Experimental data*, but the latter has a higher complexity than the former, meaning, for example, that the data has been generated over a longer time frame, required a group of actors, or is based on real scenarios. For this reason, we stick to the three categories *Experimental data*, *(Computer) Simulated data*, and *Real-world data*. Our interpretation of these data types is as follows.

- **Experimental data** refers to complex data sets or complete scenarios that are usually created manually by one or more researchers with the explicit intention of generating forensic data. To create the set, humans install the operating system and software and interact with the system by, for example, the user is conducting an experiment, replaying a scenario or performing simulations, and so on. Oftentimes, a script is defined prior to generating the data. An example is the M57-Patents scenario[1] developed by [14]. This kind of data is unique because its generation cannot be repeated in the same way (e.g., installation processes take longer, a human follows different paths when executing commands, software updates require different user inputs, etc.). Therefore, data has to be shared with the community to ensure the reproducibility of the results. In addition, appropriate metadata must be provided, as otherwise there is no ground truth.
- **(Computer) Simulated data** describes data created by software that simulates processes using system tools with a certain degree of autonomy and does not require its user to perform simulations manually. For instance, a script may be created that establishes network connections, downloads some data, and closes the connection while capturing network traffic. These simulations should provide data that is as realistic as possible. Therefore, the simulations can also be complex and create complete scenarios, such as simulating the use of one or multiple virtual machines including communication between them. These comprehensive simulations are carried out by frameworks such as `ForTrace` [5] or `TraceGen` [6]. A key

[1] https://digitalcorpora.org/corpora/scenarios/m57-patents-scenario/ (last accessed 2023-04-25).

property is that these frameworks operate deterministically (with exceptions such as timestamps) and that the ground truth of generated data is known and documented. Consequently, it is not essential to share the data set for reproducibility, but only the software (framework) including its settings.

- **Real-world data** refers to data created by humans under normal conditions, without the intention of creating forensic data. Therefore, real-world data sets do not provide the respective ground truth. Instead, it is initially unknown data. Even after a thorough investigation, the full ground truth about the data at hand may not be known because it was not explicitly recorded beforehand. In general, the more researchers examine such data sets and the better their expertise, the better the data will be understood. This category may include data collected as part of real digital forensic investigations, as well as confidential data published on the Internet or elsewhere – sometimes even without permission, for example when hackers publish or sell stolen data on the Darknet.

3 ON THE IMPORTANCE OF DIGITAL FORENSIC CORPORA

The discussion about the importance of available and realistic digital forensics corpora has a long history and probably started back in 2007 when Garfinkel [15] stated that without appropriate data sets, research in the various areas of digital forensics (e.g., disk forensics, network forensics, RAM forensics, mobile forensics, etc.) is limited by the inability of researchers to obtain large data sets that are realistic, varied, and representative of the data from the field. Even then, Garfinkel pointed out that the lack of data sets prevents researchers from pursuing many of the problems faced by today's forensic practitioners.

The position of Garfinkel [15] is supported by a variety of publications that point to the lack of publicly available data sets due to major challenges as stated for instance by Baggili and Breitinger [16], Abt and Baier [4], and Woods et al. [14]. The first major challenge in the domain of digital forensics is the lack of data sources. Especially real-world data is missing since law enforcement officers typically have to wipe disk images after a case is completed or at least keep the data secure and private [16]. Other major reasons for missing data sets include copyright and privacy issues that hinder data sharing [4]. Real-world data is often inappropriate for education purposes because privacy-sensitive or illegal digital materials are typically confidential and may not be shared with students [14].

Carrier [17] points out that testing under the eyes of the public is an important part of increasing confidence in software and hardware tools. Yannikos et al. [18] state that well-known data corpora provide a basis for comparing methodologies and tools so that benefits and shortcomings can be identified. Similarly, Baggili and Breitinger [16] stress the importance of validation for forensic tools to gain insights into error rates of commonly used forensic tools (e.g., law enforcement agencies rely on the proper functioning of algorithms and tools in a court of law). According to Garfinkel et al. [3], researchers and developers solve this problem by creating specific scenarios with synthetic data to conduct experiments, better understand new technologies, and test and verify the proper functioning of their algorithms and tools.

Ceballos Delgado et al. [7] claim that realistic case studies are essential for the successful training of digital forensics examiners, but also stress that the generation of realistic data sets is both time- and resource-consuming. Hughes and Karabiyik [19] point out the importance of digital forensic reference data that represents the full range of conditions expected during digital forensic analysis. By carefully compiling reference data, the discipline enables peer review and reproducibility of testing and provides some measure of traceability during validation testing. They also state that it is not feasible to conduct meaningful black box studies or proficiency testing without curating a collection of test images.

Grajeda et al. [2] emphasize the importance of sharing data sets so that researchers can replicate their findings and improve the state of the art. They screened 715 peer-reviewed research articles from 2010 to 2015 with a focus and relevance to digital forensics and pointed out that only 3.8% of the authors published their data sets.

The key takeaways have been summarized in Figure 2. As we have seen, the importance of data sets has been stressed by various researchers for almost two decades now. However, the problem is that privacy and property

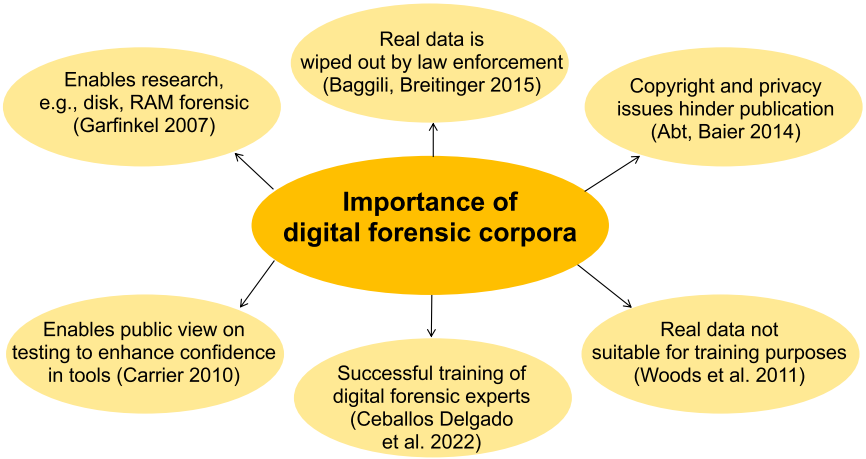

Fig. 2. On the importance of realistic digital forensic corpora.

concerns aside, developing comprehensive tests and exhaustive data sets for digital forensics is a lengthy and complex process, especially if the data to be generated is to meet high-quality standards and be as realistic as possible. In the following, we present some efforts that have been made to fill this data set gap and simplify data set generation.

3.1 Data Set Repositories

In recent years, several repositories for digital forensic (test) data sets have been published, including mainly experimentally generated disk images (including Windows, Linux, and macOS images and their respective file systems), smartphone and IoT device images, memory dumps, and network packet captures. The most prominent platforms among them are (1) the ***Computer Forensic Reference Data Sets (CFReDS)*** project[2] at the **National Institute of Standards and Technology (NIST)** [20–22], (2) *Digital Corpora*[3] including the *Real Data Corpus* [3, 23], and (3) the *Datasets For Cyber Forensics* platform[4] [2].

The first two platforms, in particular, have been maintained recently, as new data sets have been added recently.[5] As of 07/2021, NIST has made the *CFReDS v2.0*[6] platform available. It provides the forensic community with a portal of documented data sets of mainly simulated digital evidence provided either by the NIST's own ***Computer Forensic Tool Testing (CFTT)*** program or by various other organizations and volunteers in the digital forensic community. According to the new CFReDS website, data sets from the two other repositories mentioned earlier (Digital Corpora and Datasets For Cyber Forensics) have also been added to the CFReDS platform. In addition, with CFReDS v2.0, NIST introduces a taxonomy that provides better search functionality for the data sets of interest. For example, users can search for content based on tags such as memory, mobile, or disk images.

Other websites from which data sets are commonly downloaded for digital forensics are platforms that host data sets that have been used in CTFs, forensic challenges, or workshops. Prominent platforms are for instance

[2]https://cfreds.nist.gov (last accessed 2023-04-20).
[3]https://digitalcorpora.org (last accessed 2023-04-20).
[4]https://datasets.fbreitinger.de (last accessed 2023-04-20).
[5]At the time of writing, for *Digital Corpora*, the last data sets were added on 12/2022; for *CFReDS*, the last data sets were added on 03/2023; for *Datasets For Cyber Forensics* the latest data sets are from 2018.
[6]While the old CFReDS platform can still be found at https://cfreds-archive.nist.gov (last accessed 2023-04-20).

Vulnhub[7], *DFRWS Forensic Challenges*[8] or *AboutDFIR*.[9] However, these data sets usually lack a complete ground truth.

Another interesting approach is to share custom forensic artifacts with the forensic community instead of complete data sets. Since there are an immeasurable number of digital forensic artifacts available today and new artifacts are being created constantly, leaving investigators in the dark when they come across an artifact that they have not seen before, sharing such custom artifacts can help others to solve similar forensics cases. Prominent artifact-sharing platforms include the ***Artifact Genome Project (AGP)***[10] [24] and *MAGNET Artifact Exchange*.[11]

Although many of the data sets found in the repositories are treated as standardized digital forensic corpora, they often have drawbacks. For example, these data sets often have poor timeliness, insufficient diversity, unrealistic or frequently missing background noise, and no regular wear-and-tear (e.g., images without regular usage patterns such as email and browser content, or without sufficient and accurate operating system artifacts). Sometimes even the ground truth is not documented and therefore unknown. So we can say that besides the mere existence of data sets, i.e., their quantity, the actual quality of data sets undoubtedly plays an important role, too.

3.2 Data Simulation Frameworks

While the previously discussed repositories mainly contain human-generated data (compare with the lower part of Figure 1), synthetic data have also received considerable attention in recent years. Various simulation frameworks have been released to generate synthetic data of various kinds, which are discussed in the following (in chronological order):

The **`ForGe`** – **`Forensic Test Image Generator`** was introduced in 2015 by Visti et al. [25]. `ForGe` provides a user interface and takes instructions in the form of database entries. In addition, the output contains images and information sheets. Although the tool is available on GitHub,[12] it has not been further developed since 2015, i.e., it seems no longer to be maintained. Lastly, `ForGe` does not address mobile device images.

Another framework is `EviPlant` which was published by Scanlon et al. [9] in 2017. The framework makes use of a basic image as a starting point. Then challenges or traces can be subsequently downloaded in the form of *evidence packages*. This has the advantage that large files do not have to be sent numerous times, which in particular is of interest for teaching purposes. The evidence packet only has to be injected into the base image and the investigation can be started.

In 2020, Göbel et al. [8] published `hystck`, a Python-based framework that generates network and hard disk traces. This can be done automatically using Python scripts or via YAML configuration files. Through automated synthesis, it is possible to create a variety of traces with little effort inside a virtual machine. In parallel to the synthesis process, a log file is created that contains the respective ground truth of a scenario. Nevertheless, it currently only supports traces synthesized on Windows-based systems.

Du et al. [6] published the `TraceGen` framework in 2021. It is written in Python and also serves for automatic forensic image generation. The tool takes a set of predefined user actions defined in a provided script file and injects them into a virtual image. This is done by taking a set of predefined user actions defined in a provided script and then simulating the user behavior by performing operations inside a virtual machine (e.g., using an Internet browser or modifying files on a disk). All changes are stored on a disk image and simultaneously logged in a separate file that serves as the ground truth. Although the approach looks promising, the source code of `TraceGen` is not publicly released.

[7] https://vulnhub.com (last accessed 2023-06-04).
[8] https://dfrws.org/forensic-challenges (last accessed 2023-06-04).
[9] https://aboutdfir.com/education/challenges-ctfs (last accessed 2023-06-04).
[10] https://agp.newhaven.edu (last accessed 2023-06-04).
[11] https://www.magnetforensics.com/artifact-exchange (last accessed 2023-06-04).
[12] https://github.com/hannuvisti/forge (last accessed 2023-04-20).

Table 1. Four Common Digital Forensic Use Cases Which Require Data Sets

Item	Use Case	Key aspects
1	Method/Tool Testing and Validation	Adaption to recent software, hardware, concepts; Evaluation of error rates; Ability to handle legacy systems; Assessment with respect to anti-forensics
2	Practitioner Training / Digital Forensics Education	Incident training in a contemporary environment; One-time tasks for exercises and exams in university education
3	Forensic Research/Reproducibility	Scientific sound proof of a hypothesis; Re-usage by the community; Enhancing trust in digital forensic research results
4	Machine Learning	Large-scale training data to build machine learning models; Unbiased training data to get models close to reality

Ceballos Delgado et al. [7] present FADE in 2021. It is a proof-of-concept to generate traces in an Android image by using the **Android Debug Bridge (ADB)**. They directly modify files and database entries in an Android virtual machine to mimic user-created content. Although they show that it is possible to manually inject some traces and be recognized by Autopsy, the tool offers limited support to inject other app-related information or automate the injection process.

The most recent work is from 2022 and presents ForTrace [5] which is an extension of the hystck framework. As the main extension, the framework supports the generation of images on different layers, i.e., it generates persistent disk images, memory dumps, and network captures. To achieve the images, the framework simulates human-computer interactions. According to the authors, this approach should create more realistic data sets. Along with the resulting images, it also provides a report containing the generated artifacts and preserving the ground truth.

4 DATA SET USE CASES

In this section, we elaborate on the question of why data sets are required in digital forensics and what key aspects and characteristics are required. We have identified four use cases that are summarized in Table 1. In each subsection, we describe why the current use case is relevant and, more importantly, why data is needed for it. At the end of each section, we also conclude which characteristics the data sets should fulfill to be most useful for the particular use case. These use cases expand on work by Horsman and Lyle [26] who state that data is typically utilized for three purposes: For training, for tool/process evaluation, and for data exploration and reverse engineering (research & development).

4.1 Use Case 1: Method/Tool Testing and Validation

The field of digital forensics relies heavily on specialized methods, hard- and software used to acquire and analyze digital evidence. Without these methods and tools, the examination of digital devices is often not possible, at least not in time. Similar to the Daubert guidelines [27], which define the admissibility of scientific evidence to enter a United States court, according to Carrier [28] tools also have to meet basic criteria to be admissible in a court: (1) Tools, techniques, and procedures should be extensively tested to further describe the possible occurrence of false negatives or false positives. (2) In addition, results should be verifiable and falsifiable, and should be discussed with the overall goal to specify an error rate. (3) The procedure itself should have been discussed within the scientific community and should have been subject to an objective peer review. (4) An important but very fuzzy criterion is acceptance within the forensics community itself.

Obviously, tool errors, tool limitations as well as user errors exist [29]. Despite the clear dependence on digital forensic tools, techniques for validating digital forensic software are sparse and research is limited in both scope and depth. Horsman [30] states that the field currently lacks sufficient testing standards and procedures to effectively validate their usage during an investigation. Another important goal of using data sets is to find vulnerabilities in digital forensic tools that can be exploited for anti-forensic purposes [31]. As in all code, software bugs are present in digital forensic tools. These bugs can frustrate investigators, cost time, and result in lost evidence [32].

To reliably prove the correct functioning of digital forensic tools, various kinds of test data are required to validate tools. Only with a reasonable effort in tool testing can we verify the accuracy, performance, and effectiveness of forensic tools and prove the validity, reliability, and soundness of the evidence these tools provide during casework to be admissible in a trial. In addition, it is important to become familiar with how certain tools behave for certain tasks, depending on the data available in each case, as these tools are used in law enforcement.

Since 2000, NIST has a dedicated group actively working on the ***Computer Forensic Tool Testing Program (CFTT)***[13] [33, 34]. The goal of the CFTT project is to establish a methodology for testing computer forensic software and tools by developing general tool specifications, test procedures, test criteria, test sets, and hardware tests. The results provide the information necessary for toolmakers to improve tools, for users to make informed choices about acquiring and using forensic tools, and for interested parties to understand the capabilities of a certain tool. Therefore, the CFTT project aims to *provide measurable assurance to practitioners, researchers, and other applicable users that the tools used in computer forensics investigations provide accurate results* [34].

For instance, the CFTT project addresses the need for a forensic investigator or a digital forensic software developer to test a new function of a forensic tool or even a new software as a whole. Therefore, appropriate test data, which matches the specification of the new functionality, is needed. It is a very challenging task to keep the test data up to date in order to test the ever-increasing capabilities of modern forensic software. Only based on suitable and sufficient test data can we ensure that forensic software consistently delivers accurate results in a reasonable time (efficiency) and is kept up-to-date.

Furthermore, in contrast to the new features and functionality introduced with the latest version of some operating system or application, in some cases, it is still relevant to analyze legacy systems, too (e.g., old USB devices or SD cards still using the traditional FAT file system or hard disks using the Windows XP operating system). This requires testing forensic tools with respect to their capability of analyzing such legacy systems. Therefore, we also need older[14] data sets that refer to legacy systems and not only to the most recent ones.

As of today in the public domain testing of various forensic tools is done with respect to a few data sets or corpora. For general tool testing the most prominent baseline data sets include the *GovDocs1* corpus [3] (available on *Digital Corpora*), which consists of ~1 million documents collected by crawling the .gov domain. Due to its large size, a subset often used for tool testing is the *t5-corpus* [13], which contains 4,457 files of different file types. Due to the lack of readily available test data sets, these data sets have been used over and over again in the past (e.g., when testing approximate matching algorithms [35–37]). In the meantime, a newer mixed file data set called *NapierOne* [38], including almost 500,000 unique files, was presented, primarily aimed at ransomware detection and forensic analysis research. *NapierOne* was designed to address the deficiency in reproducibility and improve consistency by facilitating research replication and repeatability. The data set was inspired by the *GovDocs1* data set and is intended to be used as a complement to the original data set.

In addition, some data sets are available for testing specific data structures or application-oriented tools. For example, Carrier created some file system and disk images for testing digital forensic analysis and acquisition

[13] https://www.nist.gov/itl/ssd/software-quality-group/computer-forensics-tool-testing-program-cftt (last accessed 2023-04-25).

[14] Older is not further specific and depends on the exact domain. For instance, file systems exist over a comparatively long time period whereas smartphone applications change rapidly.

tools.[15] Nemetz et al. [39] introduced a standardized forensic corpus, which provides SQLite database files to evaluate different analysis methods and tools in this scope. The corpus consists of 77 databases grouped into five categories according to their peculiarities. The various databases use special features of the SQLite file format or contain potential pitfalls to detect errors in forensic tools. As an extension, Schmitt [40] performed various manipulations introducing anti-forensic aspects that were then tested against different SQLite analysis tools.

In summary, for the use case method/tool testing and validation, the data sets should have the following characteristics:

- Comprehensive in terms of size and variety, including legacy data.
- Known ground truth to compare the tool results against expected results.
- Include edge cases/inconsistencies to ensure the tools handle unexpected inputs properly.

4.2 Use Case 2: Practitioner Training/Digital Forensics Education

In the scope of an IT incident, well-trained forensic examiners are needed who are able to reconstruct the exact sequence of actions in a contemporary IT environment including modern devices, operating systems, (server) applications, networks, and remote storage. The widespread use of diverse IT hardware and software poses numerous challenges and threats to the security of companies and private individuals. To keep up with this challenge, forensic experts depend on training data covering a broad spectrum of criminal activities to train their skills as pointed out for instance by [2, 3, 14, 15, 18].

Depending on the training, the data may be either specific and narrow or comprise a complex scenario. As an example, to become familiar with Wireshark, it requires PCAP files that include certain circumstances such as the TLS-handshake, a WireGuard (VPN) connection, or a **Command-and-Control (C2)** communication.[16] On the other hand, for practicing casework and report writing, complex scenarios with realistic storylines such as the M57 case [14] are needed.

Training is a serious problem facing organizations that provide digital forensic services and incident response. Pertinent experience in the digital forensic field and corresponding certifications of employees are becoming increasingly important. Already back in 2010, Garfinkel [41] drew attention to the fact that there is a lack of complex, realistic training data, which means that most classes are taught using either simplistic manufactured data or live data. In most cases, however, live data cannot be shared between institutions due to privacy and secrecy reasons, resulting in dramatically higher costs for the preparation of instructional material.

In the case of digital forensics education at universities or academies, practical exercises or even exams are reasonable to practice and assess the respective skills. However, if the same task is provided for several intakes or even for dozen of students, the corresponding solution will be leaked by fellow students or prior cohorts. Hence again, flexible and adaptable forensic challenges are needed for training and education. Ideally, each individual student should get a unique forensic challenge to solve. In the case of an exam, the instructor needs to know the exact ground truth and consequently, the expected findings which ensure grading is possible. Hence, the ability to generate a flexible and adaptable set of similar, but not identical, forensic images also enables the assessment of the difficulty of a certain forensic challenge, e.g., to quantify the knowledge of an investigator or to evaluate different teaching approaches.

Lastly, training and education typically focus on the latest trends and developments, so it is essential that training data is up-to-date. Older scenarios have often been automated by vendors and are therefore less relevant for practitioners. For instance, there is no/little training on file recovery, as sophisticated file carvers exist, but there is currently a lot of IoT-based training including drones.

[15] https://dftt.sourceforge.net (last accessed 2023-06-04).
[16] Many such sample package captures are provided directly on the official Wireshark website: https://wiki.wireshark.org/SampleCaptures (last accessed 2023-04-20).

In summary, for the use case practitioner training/digital forensics education, the data sets should have the following characteristics:

- Broad spectrum of data needed (from narrow samples to complex cases).
- Ground truth may be required in the case of an assessment (exam).
- Randomization 'per user' may be desired in a more academic setup (it may be undesired in commercial training where everyone should work on the identical case for comparability of results).

4.3 Use Case 3: Forensic Research/Reproducibility

As for any other practical-oriented science, data sets are essential for conducting research concerning both academic and industrial research. A typical setting is that researchers claim some properties of an approach within their scientific paper (e.g., for runtime efficiency or classification metrics like detection rate) and evaluate their approach by practical experiments on a suitable data set. Research results should be evaluated with respect to realistic, but also recent data. For instance, in case a new mobile messaging app is used by criminals, it is of utmost importance to analyze the current version of the messaging app concerning the usage traces generated on mobile devices (e.g., communication content, encryption). This data is likely to be fabricated by the researchers, as reference data sets may not be available at this point.

A second reason why data sets are required is to enhance the state of the art and enable reproducibility, i.e., others can independently evaluate claims, which is a key requirement of the scientific workflow. However, as sketched in Section 3, data sets are often not released for a variety of reasons and consequently *reproducibility*, also referred to as *repeatability* [3, 4], is not given. Note that in addition to the data set, a potential implementation/software should also be accessible to the community. Only when everything (precise description/article, data set, and potential software) is available, a thorough (peer-)review can be completed.

In summary, for the use case forensic research/reproducibility, the data sets should have the following characteristics:

- Shareability is important to ensure reproducibility and a thorough (peer-)review.
- Contemporary research should explore new ideas.
- Specific to the problem instead of too comprehensive/complex (e.g., it is better to have a set of smaller samples than complex scenarios for initial research).

4.4 Use Case 4: Machine Learning

In recent years, artificial intelligence in general and machine learning in particular is used in many branches of computer science and beyond. Within cyber security, a canonical domain of machine learning is in the scope of automated and large-scale detection and classification. With respect to digital forensics, machine learning is commonly applied in the field of malware detection and attribution [42, 43], for the detection of sexual harassment (e.g., such as online sexual predatory chats [44]), and the detection, identification, and classification of **child sexual abuse material (CSAM)** [45, 46], to name a few.

A common class of machine learning algorithms contains supervised ones, that is the machine learning algorithm generates its models on base of a labeled (training) data set. While practitioners favor using real-world training data (or at least experimental data) to get realistic machine learning models, a publicly available and labeled training data set is typically missing (see for instance [2, 4]) making the model generation a difficult task. In general, the training of a machine learning algorithm is hardly scalable because there is often not enough real-world data available in law enforcement.

On the other side, synthetic data sets find more and more support in the cyber security community and hence in digital forensics [47, 48]. Hittmeir et al. [47] point out that *synthetic data tries to preserve the overall properties and characteristics of the original data without revealing information about actual individual data samples. The promise is that, for most purposes, models trained on synthetic data instead of real-world data do not show a significant loss of*

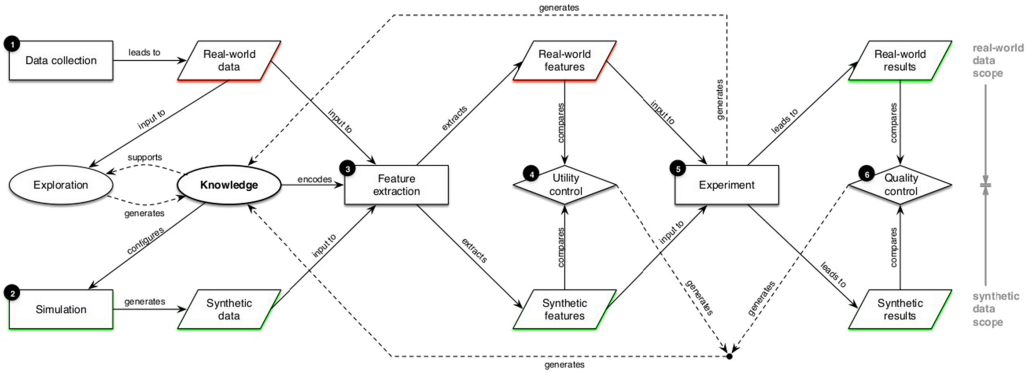

Fig. 3. Workflow to train on synthetic data and to operate on real-world data according to Abt and Baier [49].

performance. Their key result is that the utility of the simulated synthetic data is given, and they evaluate their hypothesis by applying it to a number of supervised machine learning tasks on several publicly available data sets.

Synthetic data for machine learning solutions in digital forensics are typically used as part of the training data for classification tasks. The classification task itself is very important for the question of whether synthetic data can be used at all, or whether the data must be as realistic as possible (and therefore may not allow synthetic data). For example, if each classification result of a trained machine learning model can be easily evaluated by a digital forensic investigator, we believe that the requirements for synthetic training data could be much less strict than for tasks where the results cannot be quickly and easily evaluated. For example, if a machine learning model is used in file carving to classify image fragments to assemble a fragmented image, an investigator can probably always decide if the classification result was correct or not (e.g., if the image can be decoded and shows plausible content, then the classifier found the right fragments). Therefore, the overall accuracy of the classifier on real data does not have to be as high as possible (i.e., synthetic data does not necessarily have to be as realistic as possible and may be used). However, if a machine learning model is used to decide if the writing style extracted from a text corresponds to a specific author (which is a very common task in forensic linguistics), it is much more likely that the investigator is not able to immediately decide if the classification result is correct or not. In such cases, it is important that a classifier has a very high accuracy to rely on and this requires high-quality training data that is as realistic as possible (which may then disqualify synthetic data).

Figure 3 shows a workflow to make use of synthetic data and achieve the quality of real-world data as proposed by Abt and Baier [49]. The key idea may be used by machine learning algorithms in digital forensics, too, if the machine learning model is quality controlled by a small set of real-world data, but trained on scalable synthetic data. Basically, if the model is not able to distinguish between synthetic and real-world data, the machine learning algorithm can be used on real-world data after training on synthetic data, the models can also be used by outstanding parties.

In summary, for the use case machine learning, the data sets should have the following characteristics:

- Comprehensive and labeled data (ground truth) is essential for several artificial intelligence approaches.
- Extremely close to real-world data to be useful in practice. Synthetic data may be used depending on the complexity of the actual machine learning task.
- Data has to be up to date and not biased to obtain the best results.

4.5 Use Case Summary

The use cases presented in the previous subsections all rely on data/data sets, but their requirements are different. While sometimes it is important to know the ground truth (e.g., training and education in an academic setting),

sometimes it is less important or the subject of interest (e.g., research). On the other hand, sometimes large amounts of data are needed (e.g., in machine learning), while there are cases where one sample may be sufficient (e.g., to test a particular feature of a tool). In the next section, we compare the use cases with the data set types outlined in Section 2.3 and thereby explain which data types are well suited for a given use case.

5 USE CASE CENTRIC DISCUSSION OF DATA CHARACTERISTICS AND REQUIREMENTS

In the title and the beginning of this article, we argue that one should not compare real-world data and synthetic data sets in general, but that both have their respective right to exist. In this section, we compare the different kinds of data sets in relation to our use cases presented in Section 4 and provide some recommendations on which data (types) can be reasonably used for each use case. Furthermore, we address the aspect of 'how realistic is synthetic data'.

5.1 How Realistic is Synthetic Data?

As stated in the introduction, this is a common question when discussing data simulation frameworks, such as `ForTrace` [5] or `TraceGen` [6], and synthetically generated data in general. We argue that this question is often dispensable and inappropriate as simulated data is realistic in its way: it utilizes system functionality and thus represents reality. For instance, a framework that can produce memory dumps will boot the operating system, load drivers, run the software, simulate user behavior, and so on, and at some point captures the snapshot. This is a valid snapshot from a running system and therefore realistic. Consequently, a better question is: How does the synthetic data set differ from a similar real-world data set? Answering this question helps us to understand which data (type) can be reasonably used for the particular use case. To do so, it is helpful to first identify some key characteristics that often differ between these types. These key characteristics are as follows:

- Shareability: Simulated data can normally be shared as one does not have to worry about privacy issues or legal restrictions. Unlike real-world data, these data sets do not contain any personal data.
- Ground truth: Simulated data normally follows rules and thus a ground truth, or at least parts thereof, is usually available, whereas for a real-world data set the data is completely unknown.
- Adaptability/Updateability: Creating a similar data set with an updated version of a particular software or operating system may be easier with the help of frameworks, as a repeated manual creation of experimental data would be costly. On the other hand, if not already available, integrating new artifacts in a synthesis framework or developing the framework itself, can also be time-consuming. So it needs to be considered when it is worth automating things (e.g., for frequently needed and repetitive artifacts) and when it is not (e.g., with a high fluctuation in data representation).
- Determinism: While real-world data is not deterministic since the ground truth is not known, synthetic data is at least partially deterministic (in the case of experimental data) or fully deterministic (in the case of simulated data based on a deterministic configuration). For the latter, the repetition of predefined scenarios usually leads to the same result. If required, however, the simulated data does not always have to be identical, as long as the data set generator works with some kind of configurable seed or random values.
- Background noise (i.e., simultaneous system activity besides the desired outcome): Real-world data usually includes more simultaneous activities. For example, the number of tabs opened in a browser and the amount of idle network traffic produced in the background would typically differ in simulated data sets where only a few websites may be visited.
- Regular wear-and-tear: Simulators follow a cookbook and usually can create large amounts of data in short periods. Real-world data grows more slowly and over time, and some artifacts or regular usage patterns may only be visible (or not be visible as they are overwritten) in the latter scenario, such as sufficient operating system actions and its applications performing updates, virus scans, log entries, and the like, over a long period of time.

- Timeliness: As digital forensics encounters new evidence every day, real-world data contains digital artifacts that are most recent. Simulated data, on the other hand, may contain more common artifacts, but misses so far unknown artifacts, because they can only be simulated after such data has been analyzed, understood, and integrated into the data synthesis software.
- Predictability: Real-world data is less predictable and one may encounter scenarios that are unexpected, new, or inconsistent. As simulators generate data exactly as configured or scripted, data sets may be more predictable because they may lack the ability to produce data set conditions that are not as frequent.

5.2 Use Cases vs. Data Set Types

Knowing some of the differences between real-world data and data simulated manually or by synthesis frameworks, it is possible to relate the data set types to the use cases.

UC1 - Method/Tool Testing and Validation. One of the biggest challenges in testing digital forensic software is the generation and maintenance of sufficiently detailed and documented test data sets that are used to validate a software or tool's functionality. As Brunthy [50] stated, the importance of proper scientific validation is becoming increasingly important as the field of digital forensics continues to grow.

To keep up with ever-changing trends in the cybersecurity field, it probably is difficult to always produce new data sets synthetically that immediately keep up with new findings (like zero-day exploits). The practitioner survey in [30] asked whether it will ever be possible to satisfactorily test all the major tools in use. The results show that available data sets provide a good basis for developing tool tests, but they are not exhaustive and do not provide the depth required to effectively test the full functionality of the tools.

Therefore, the recommendation is to rely on both real-world data (e.g., real malware samples to test the detection capabilities of forensic software) and synthetic data (e.g., various file system images with data hidden in different areas) to test forensic software more accurately and efficiently. For real-world data, it must at least be known what is contained, e.g., the function of the malware must be known. Otherwise, the examiner would not know what is to be found during validation. For synthetically generated data sets, documented ground truth will assist in the validation, so especially computer-simulated data is well suited. In a controlled environment (e.g., same test data set, laboratory, software settings, examiner, etc.), the forensic tool must produce a deterministic test result (i.e., test results must be repeatable). Simulators allow for rapid changes to the test data, which can speed up the validation of a tool while further testing its capabilities.

UC2 - Practitioner Training/Digital Forensics Education: Concerning the training of forensic examiners, all three types (real-world, experimental, and computer-simulated data) are needed. While some cases require very specific and narrow data sets, which are most easily produced manually by running an experiment and cannot be simulated with any available framework, the following must be considered. The major drawback of manually created training data is their lack of adaptability. For instance, once a disk image is created, it is static in most cases, meaning we cannot easily adjust it without recreating the entire image. Especially in the academic environment, recurring use of the same images in class may not be possible, as corresponding write-ups or walkthroughs may quickly spread over the Internet, and thus the learning effect would be limited. Therefore, educational domains (especially in an academic setting) usually require a large amount of data (i.e., it should be generated and modified as quickly as possible) that ideally is unique to a certain extent (e.g., using some kind of randomness or a seed in the synthesis framework that allows students to be assigned individual cases so that they cannot copy from each other).

It is also important to know the ground truth for evaluation purposes, i.e., synthetic (documented) data sets are useful for competence and proficiency examination for digital examiners. Since the manual creation of large amounts of high-quality training data for digital forensics is tedious, time-consuming, and error-prone [14], the data synthesis frameworks to automatically generate forensic training data are suitable for this purpose.

One big advantage is that the actual data basis can be easily replaced and unique data images may be quickly provided.

In the case of specialized law enforcement training, however, one may be able to use real-world data from previous casework. These are then typically edge cases for which no synthetic data is available in sufficient quantity or quickly enough, as no software is suitable for this purpose. In this case, the ground truth may be provided by the actual forensic report of the previous case.

UC3 - Forensic Research/Reproducibility: Several scenarios can occur in this area. The first typical one is that up-to-date software and recent technologies are objects of interest to provide the community with new insights into the forensic processing of the respective software or technology. It is likely that at the beginning of such a research project no data simulation framework exists wherefore the data has to be created manually or a simulation software must first be developed or adapted, which can constitute a significant overhead. As a result, one mainly encounters real-world data or experiment-generated data in this use case.

A second scenario, however, is the development and evaluation of a new algorithm to improve the solution to a well-known problem (e.g., with respect to run-time or storage efficiency). It is then probable that a synthesis framework already provides the possibility to generate a respective data set, especially if the evaluation is based on large-scale data (e.g., in the scope of generating a timeline of different sources efficiently). Hence, the adaptation of synthetic data to evaluate a new algorithm is probable in this research scenario and is thus the best choice.

UC4 - Machine Learning: Machine learning requires large amounts of data and may require labeling (e.g., in supervised learning). In addition, the data has to be structured in a certain way so that it can serve as input for an algorithm (e.g., currently, it is not possible to feed a full disk image or a set of files without having additional parsers). On the other hand, as synthetic data mimics real-world data, it should also be usable to train AI systems. However, this may also depend on the goal of the system, i.e., it may be possible for some systems but not for others. Consequently, the best choice depends on several factors, such as:

- When real-world data is not available, synthetic data is the only possibility.
- Synthetic data can be used without privacy concerns.
- Real-world data can be noisy, inconsistent, or biased, whereas synthetic data can be generated with specific properties, ensuring the quality and consistency of the data.
- Generating synthetic data can be more cost-effective compared to collecting and annotating real-world data.

In general, a combination of real-world data and synthetic data may be used to train AI models. The use of real-world data helps the AI model generalize to real-world scenarios, while synthetic data can help address data limitations and improve the overall performance of the model.

6 CONCLUSION

Data and data sets are critical to progress in digital forensics science. Several studies have discussed their importance over the last few years. NIST and others have released data set collections (repositories) providing the digital forensics community with data sets in various areas (e.g., memory, hard disk, or mobile forensics) usually including the respective ground truth. In addition, several data simulation frameworks have been implemented to help generate the most realistic data sets possible in a fully or at least partially automated manner. Especially for the latter, a concern that is frequently raised is: *how realistic is synthetic data?* This article addressed this question and argues that one should not compare these two data types, as both have different strengths and are necessary for the respective scope to make progress in digital forensics. We demonstrated the importance of having both synthetic and real-world data by presenting four use cases that rely on data: (1) method/tool testing and validation, (2) practitioner training/digital forensics education, (3) forensic research/reproducibility, and (4)

machine learning. Based on these use cases, we discussed that sometimes one type is preferable to another, e.g., data simulation frameworks allow mass-production of data sets with known ground truth, making them ideal for an academic setting, or they allow rapid changes to test data, which can improve and speed up tool validation. Moving forward, instead of constantly comparing these data types, in the future we need to ensure that the data is shared to ensure reproducibility and progress in the field.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their comments to improve the paper.

AUTHOR CONTRIBUTION STATEMENT

Thomas Göbel: Conceptualization, Writing - Original Draft, Writing - Review & Editing. **Harald Baier:** Writing - Review & Editing, Validation, Supervision. **Frank Breitinger:** Conceptualization, Writing - Review & Editing, Supervision.

DECLARATION OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] David Lillis, Brett Becker, Tadhg O'Sullivan, and Mark Scanlon. 2016. Current Challenges and Future Research Areas for Digital Forensic Investigation. (2016). DOI: http://dx.doi.org/10.48550/ARXIV.1604.03850

[2] Cinthya Grajeda, Frank Breitinger, and Ibrahim Baggili. 2017. Availability of datasets for digital forensics - And what is missing. *Digital Investigation* 22 (2017), S94–S105. DOI: http://dx.doi.org/10.1016/j.diin.2017.06.004

[3] Simson Garfinkel, Paul Farrell, Vassil Roussev, and George Dinolt. 2009. Bringing science to digital forensics with standardized forensic corpora. *Digital Investigation* 6 (2009), S2–S11. DOI: http://dx.doi.org/10.1016/j.diin.2009.06.016 *The Proceedings of the Ninth Annual DFRWS Conference.*

[4] Sebastian Abt and Harald Baier. 2014. Are we missing labels? A study of the availability of ground-truth in network security research. In *2014 Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS).* 40–55. DOI: http://dx.doi.org/10.1109/BADGERS.2014.11

[5] Thomas Göbel, Stephan Maltan, Jan Türr, Harald Baier, and Florian Mann. 2022. ForTrace - A holistic forensic data set synthesis framework. *Forensic Science International: Digital Investigation* 40 (2022), 301344. DOI: http://dx.doi.org/10.1016/j.fsidi.2022.301344 Selected Papers of the Ninth Annual DFRWS Europe Conference.

[6] Xiaoyu Du, Christopher Hargreaves, John Sheppard, and Mark Scanlon. 2021. TraceGen: User activity emulation for digital forensic test image generation. *Forensic Science International: Digital Investigation* 38 (2021), 301133. DOI: http://dx.doi.org/10.1016/j.fsidi.2021.301133

[7] Alberto A. Ceballos Delgado, William B. Glisson, George Grispos, and Kim-Kwang Raymond Choo. 2022. FADE: A forensic image generator for Android device education. *WIREs Forensic Science* 4, 2 (2022), e1432. DOI: http://dx.doi.org/10.1002/wfs2.1432 arXiv:https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wfs2.1432

[8] Thomas Göbel, Thomas Schäfer, Julien Hachenberger, Jan Türr, and Harald Baier. 2020. A novel approach for generating synthetic datasets for digital forensics. In *Advances in Digital Forensics XVI*, Gilbert Peterson and Sujeet Shenoi (Eds.). Springer International Publishing, Cham, 73–93.

[9] Mark Scanlon, Xiaoyu Du, and David Lillis. 2017. EviPlant: An efficient digital forensic challenge creation, manipulation and distribution solution. *Digital Investigation* 20 (2017), S29–S36. DOI: http://dx.doi.org/10.1016/j.diin.2017.01.010 DFRWS 2017 Europe.

[10] Frank Breitinger and Alexandre Jotterand. 2023. Sharing datasets for Digital Forensic: A novel taxonomy and legal concerns. *Forensic Science International: Digital Investigation* (07 2023). Accepted for publication at DFRWS USA 2023.

[11] Brian Carrier. 2005. *File System Forensic Analysis.* Addison-Wesley Professional.

[12] Jon Berryhill. 2019. What is Metadata? (2019). https://www.computerforensics.com/news/what-is-metadata

[13] Vassil Roussev. 2011. An evaluation of forensic similarity hashes. *Digital Investigation* 8 (2011), S34–S41. DOI: http://dx.doi.org/10.1016/j.diin.2011.05.005 *The Proceedings of the Eleventh Annual DFRWS Conference.*

[14] Kam Woods, Christopher A. Lee, Simson Garfinkel, David Dittrich, Adam Russell, and Kris Kearton. 2011. Creating realistic corpora for security and forensic education. In *Proceedings of ADFSL Conference on Digital Forensics, Security and Law.* 123–134.

[15] Simson Garfinkel. 2007. Forensic corpora: A challenge for forensic research. *Electronic Evidence Information Center* (2007), 1–10.

[16] Ibrahim Baggili and Frank Breitinger. 2015. Data sources for advancing cyber forensics: What the social world has to offer. In *2015 AAAI Spring Symposium Series*.
[17] Brian Carrier. 2010. Digital Forensics Tool Testing Images. URL: http://dftt.sourceforge.net. (2010). Accessed: 2021-10-12.
[18] York Yannikos, Martin Steinebach, Lukas Graner, and Christian Winter. 2014. Data corpora for digital forensics education and research. In *Advances in Digital Forensics X*, Gilbert Peterson and Sujeet Shenoi (Eds.). Springer Berlin , Berlin, 309–325.
[19] Nicolas Hughes and Umit Karabiyik. 2020. Towards reliable digital forensics investigations through measurement science. *Wiley Interdisciplinary Reviews: Forensic Science* (2020), e1367. DOI: http://dx.doi.org/10.1002/wfs2.1367
[20] Jame R. Lyle, Richard P. Ayers, Jame R. Lyle, and Douglas White. 2008. *Digital Forensics at the National Institute of Standards and Technology*. US Department of Commerce, National Institute of Standards and Technology.
[21] Jungheum Park, James R. Lyle, and Barbara Guttman. 2016. Introduction to CFTT and CFReDS projects at NIST. *Journal of the Korea Institute of Information Security & Cryptography* (2016).
[22] NIST. 2022. The CFReDS Project. https://www.cfreds.nist.gov/. (2022). Online; accessed: 2nd December 2022.
[23] Simson Garfinkel. 2012. Lessons learned writing digital forensics tools and managing a 30TB digital evidence corpus. *Digital Investigation* 9 (2012), S80–S89. DOI: http://dx.doi.org/10.1016/j.diin.2012.05.002 *The Proceedings of the Twelfth Annual DFRWS Conference.*
[24] Cinthya Grajeda, Laura Sanchez, Ibrahim Baggili, Devon Clark, and Frank Breitinger. 2018. Experience constructing the Artifact Genome Project (AGP): Managing the domain's knowledge one artifact at a time. *Digital Investigation* 26 (2018), S47–S58. DOI: http://dx.doi.org/10.1016/j.diin.2018.04.021
[25] Hannu Visti, Sean Tohill, and Paul Douglas. 2015. Automatic creation of computer forensic test images. In *Computational Forensics*, Utpal Garain and Faisal Shafait (Eds.). Springer International Publishing, Cham, 163–175.
[26] Graeme Horsman and James R. Lyle. 2021. Dataset construction challenges for digital forensics. *Forensic Science International: Digital Investigation* 38 (2021), 301264. DOI: http://dx.doi.org/10.1016/j.fsidi.2021.301264
[27] Margaret G. Farrell. 1993. Daubert v. Merrell Dow Pharmaceuticals, Inc.: Epistemilogy and legal process. *Cardozo L. Rev.* 15 (1993), 2183.
[28] Brian Carrier. 2002. Open source digital forensics tools: The legal argument. @stake, Inc. http://dl.packetstormsecurity.net/papers/IDS/atstake_opensource_forensics.pdf
[29] Graeme Horsman. 2018. "I couldn't find it your honour, it mustn't be there!" - Tool errors, tool limitations and user error in digital forensics. *Science & Justice* 58, 6 (2018), 433–440. DOI: http://dx.doi.org/10.1016/j.scijus.2018.04.001
[30] Graeme Horsman. 2019. Tool testing and reliability issues in the field of digital forensics. *Digital Investigation* 28 (2019), 163–175. DOI: http://dx.doi.org/10.1016/j.diin.2019.01.009
[31] Kevin Conlan, Ibrahim Baggili, and Frank Breitinger. 2016. Anti-forensics: Furthering digital forensic science through a new extended, granular taxonomy. *Digital Investigation* 18 (2016), S66–S75. DOI: http://dx.doi.org/10.1016/j.diin.2016.04.006
[32] Brian Cusack and Alain Homewood. 2013. Identifying bugs in digital forensic tools. (2013).
[33] James Lyle. 2002. NIST CFTT: Testing disk imaging tools. *Digital Forensic Research Workshop, International Journal of Digital Evidence*, (online at www.ijde.org), Syracuse,. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=51081
[34] James Lyle, Barbara Guttman, and Richard Ayers. 2011. Ten years of computer forensic tool testing. 8 (2011-10-12 00:10:00 2011). https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=909329
[35] Thomas Göbel, Frieder Uhlig, Harald Baier, and Frank Breitinger. 2022. FRASHER – A framework for automated evaluation of similarity hashing. *Forensic Science International: Digital Investigation* 42 (2022), 301407. DOI: http://dx.doi.org/10.1016/j.fsidi.2022.301407 *Proceedings of the Twenty-Second Annual DFRWS* USA.
[36] Thomas Göbel, Frieder Uhlig, and Harald Baier. 2021. Evaluation of network traffic analysis using approximate matching algorithms. In *Advances in Digital Forensics XVII*, Gilbert Peterson and Sujeet Shenoi (Eds.). Springer International Publishing, Cham, 89–108.
[37] David Lillis, Frank Breitinger, and Mark Scanlon. 2018. Expediting MRSH-v2 approximate matching with hierarchical Bloom filter trees. In *Digital Forensics and Cyber Crime*, Petr Matoušek and Martin Schmiedecker (Eds.). Springer International Publishing, Cham, 144–157.
[38] Simon R. Davies, Richard Macfarlane, and William J. Buchanan. 2022. NapierOne: A modern mixed file data set alternative to Govdocs1. *Forensic Science International: Digital Investigation* 40 (2022), 301330. DOI: http://dx.doi.org/10.1016/j.fsidi.2021.301330
[39] Sebastian Nemetz, Sven Schmitt, and Felix Freiling. 2018. A standardized corpus for SQLite database forensics. *Digital Investigation* 24 (2018), S121–S130. DOI: http://dx.doi.org/10.1016/j.diin.2018.01.015
[40] Sven Schmitt. 2018. Introducing anti-forensics to SQLite corpora and tool testing. In *2018 11th International Conference on IT Security Incident Management & IT Forensics (IMF)*. 89–106. DOI: http://dx.doi.org/10.1109/IMF.2018.00014
[41] Simson L. Garfinkel. 2010. Digital forensics research: The next 10 years. *Digital Investigation* 7 (2010), S64–S73. DOI: http://dx.doi.org/10.1016/j.diin.2010.05.009 *The Proceedings of the Tenth Annual DFRWS Conference.*
[42] Quan Le, Oisín Boydell, Brian Mac Namee, and Mark Scanlon. 2018. Deep learning at the shallow end: Malware classification for non-domain experts. *Digital Investigation* 26 (2018), S118–S126. DOI: http://dx.doi.org/10.1016/j.diin.2018.04.024
[43] ElMouatez Billah Karbab, Mourad Debbabi, Abdelouahid Derhab, and Djedjiga Mouheb. 2018. MalDozer: Automatic framework for Android malware detection using deep learning. *Digital Investigation* 24 (2018), S48–S59. DOI: http://dx.doi.org/10.1016/j.diin.2018.01.007

[44] C. H. Ngejane, J. H. P. Eloff, T. J. Sefara, and V. N. Marivate. 2021. Digital forensics supported by machine learning for the detection of online sexual predatory chats. *Forensic Science International: Digital Investigation* 36 (2021), 301109. DOI: http://dx.doi.org/10.1016/j.fsidi.2021.301109

[45] Laura Sanchez, Cinthya Grajeda, Ibrahim Baggili, and Cory Hall. 2019. A practitioner survey exploring the value of forensic tools, AI, filtering, & safer presentation for investigating child sexual abuse material (CSAM). *Digital Investigation* 29 (2019), S124–S142. DOI: http://dx.doi.org/10.1016/j.diin.2019.04.005

[46] Janis Dalins, Yuriy Tyshetskiy, Campbell Wilson, Mark J. Carman, and Douglas Boudry. 2018. Laying foundations for effective machine learning in law enforcement. Majura – A labelling schema for child exploitation materials. *Digital Investigation* 26 (2018), 40–54. DOI: http://dx.doi.org/10.1016/j.diin.2018.05.004

[47] Markus Hittmeir, Andreas Ekelhart, and Rudolf Mayer. 2019. On the utility of synthetic data: An empirical evaluation on machine learning tasks. In *Proceedings of the 14th International Conference on Availability, Reliability and Security, ARES 2019, Canterbury, UK, August 26-29, 2019*. ACM, 29:1–29:6. DOI: http://dx.doi.org/10.1145/3339252.3339281

[48] Sebastian Abt and Harald Baier. 2014. A plea for utilising synthetic data when performing machine learning based cyber-security experiments. In *Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop (AISec)*. 37–45. DOI: http://dx.doi.org/doi/10.1145/2666652.2666663

[49] Sebastian Abt and Harald Baier. 2015. A research process that ensures reproducible network security research. In *11th International Conference on Network and Service Management, CNSM 2015, Barcelona, Spain, November 9-13, 2015*, Mauro Tortonesi, Jürgen Schönwälder, Edmundo Roberto Mauro Madeira, Corinna Schmitt, and Joan Serrat (Eds.). IEEE Computer Society, 71–77. DOI: http://dx.doi.org/10.1109/CNSM.2015.7367341

[50] Josh Brunthy. 2021. Validation of Forensic Tools- A Quick Guide for the DFIR Examiner. (2021). https://joshbrunty.github.io/2021/11/01/validation.html

Received 4 June 2023; accepted 26 June 2023