Technical Report

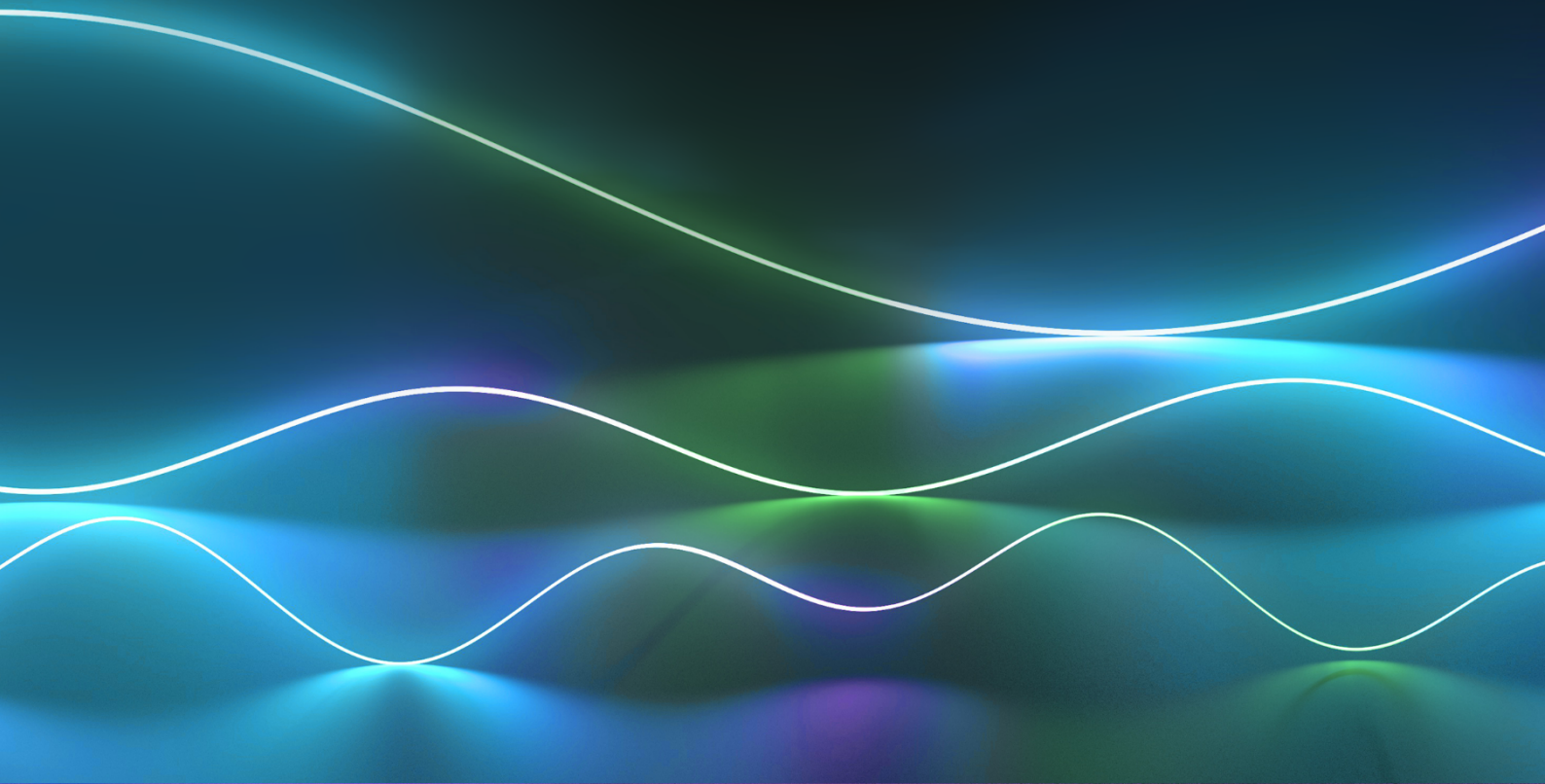
AiLECS lab

The Data Airlock

Infrastructure for restricted data informatics

Gregory Rolan
Janis Dalins
Campbell Wilson

TR22/03
March 16, 2022



The Data Airlock

Infrastructure for restricted data informatics

Gregory Rolan
Janis Dalins
Campbell Wilson

TR22/03

March 16, 2022

AI for Law Enforcement and Community Safety Lab, Monash University, Wellington Road, Victoria 3800, Australia

e: greg.rolan@monash.edu

AiLECS lab
Ai for law enforcement and community safety

Summary

Data -science collaboration is problematic when access to operational data or models from outside the data-holding organisation (or organisational unit) is prohibited, for a variety of legal, security, ethical, or practical reasons. There are significant data privacy challenges when performing collaborative data -science work against such *restricted* data.

In this report we describe a range of causes and risks associated with restricted data along with the social, environmental, data, and cryptographic measures that may be used to mitigate such issues. We then show how these are generally inadequate for restricted data contexts and introduce the 'Data Airlock' — secure infrastructure that facilitates *eyes-off* data -science workloads. After describing our use-case, we detail the architecture and implementation of a first, single-organisation version of the Data Airlock infrastructure. We conclude with outcomes and learning from this implementation, and outline requirements for a second, federated version.

Contents

Nomenclature

Abbreviations and Technical Terms

Term	Definition
API	Application Programming Interface – affordances used to make software interoperable
CSAM	Child Sexual Abuse Material
CUDA	NVIDIA software and API for supporting parallel workloads on their proprietory hardware
DGX	Model of NVIDIA high-performance server
Docker	software that allows isolation of workloads on a general-purpose server
ERICA	E-Research Institutional Cloud Architecture – cloud-based computing platform for sensitive data
Eyes-off	The ability to run collaborative workloads against restricted data without any visibility of the data beyond the data custodian domain.
HIPPA	Health Insurance Portability and Accountability Act – a United States federal statute that governs access to health data
ICT	Information and communications Technology
NVIDIA	Manufacturer of high-performance computing components
PKI	Public Key Infrastructure – allowing distributed asymmetric encryption operations
TEE	Trusted Execution Environment – hardware architecture that isolates workloads
SDK	Software Development Kit
SAIL	Secure Anonymised Information Linkage – cloud-based computing platform for anonymised person-based data

List of Figures

1

Introduction

The successful application of data science to operational data requires collaborative effort across the workflow pipeline for the development, training, testing, debugging, and integration of models and their outputs [50]. Increasingly, and particularly in the public sector, many such collaborators are external to the data-holding organisation [34]. Importantly, each of these participants may require access to the underlying data, the models being developed, and/or any outputs. Moreover, many grand challenge problems of our time require models to be trained on data from multiple organisations [18] — in particular, to mitigate against forms of bias due to narrow datasets [43] .

The challenges of such collaborative R&D, are vastly more difficult if there are constraints on such access due to sensitivities in the data or models themselves. Consequently, a range of techniques have been adopted for collaborative data science involving sensitive data or models [27]. Such measures have met with varying degrees of success depending on the sensitivities involved, the nature of the workload, and stakeholder needs.

Even so, external collaboration remains problematic when access to operational data or models is prohibited, for a variety of legal, security, ethical, or practical reasons. In this report we describe a new infrastructure for facilitating *eyes-off* access to restricted data in the pursuit of collaborative data science research and development. Named 'Data Airlock', this infrastructure was developed in conjunction with a national law enforcement agency, with a view to facilitate collaboration in the development of data classification models.

We begin this report by exploring various measures for protecting sensitive data that remain problematic when applied to restricted data. We then present the motivating use case for our work — the development of automated classifiers of child sexual abuse material (CSAM). Having provided this background, we introduce the 'Data Airlock', infrastructure that addresses the issues with restricted data, and reflect on our learning from its implementation. We conclude with a prognosis for this infrastructure for restricted data informatics.

2

Background

While domain experts and data scientists play a leading role in data science collaboration [50], a wide variety of other stakeholders are also involved [11, 39, 6]. Increasingly, and particularly within the public sector, organisations leverage in-house operational data and domain expertise through partnerships with external researchers, contributors, and service providers [34]. Similarly, the development of public sector models in response to societal grand challenges requires access to a breadth of data that encompasses the often granular structure of the public sector and beyond [34, 42, 40].

Cross-organisational involvement increases the complexity of managing collaboration [49], while complicating access and requirements for security, audit, and transparency [39]. Moreover operational data and requirements change through time, as do the differing objectives, responsibilities, and performance expectations of stakeholders, resulting in the need for continual development, verification, and explanation of models. These difficulties are exacerbated when operational data and models possess one or more potentially dynamic sensitivities that place limitations on sharing of material for collaboration [47] .

Such sensitivities arise for a number of reasons and may derive from regulation, policy, or even community expectations. Privacy protections for subjects recorded in data sets, for example in the health or social sciences, are often mandated by law [25]. Such provisions include the protection of personal details as well as information regarding membership of some data class (or not) based on personal attributes [31]. Similarly, there may be constraints on secondary uses of data in the absence of explicit consent [15]. Commercial considerations may preclude the open sharing of proprietary datasets and/or models. In some cases access to data and/or models may be *restricted* by legislation — for example in the law enforcement, national security, and defence sectors. Across this spectrum of sensitivities, a range of social and technical measures are available that can ameliorate access constraints.

2.1. Dealing with sensitive material

The protection of sensitive material can be considered in terms of security *threats* of two types: *internal* and *external* [30]. An internal threat in this context is the inadvertent exposure or intentional sharing of material by a party who has been granted access. Conversely, an external threat is the deliberate exfiltration, reconstruction, or dissemination of material by a third-party without permission. The challenges of sharing sensitive data for analytical work is not a new problem and predates the data science era. For example, twenty-five years ago, Jabine comprehensively detailed various causes and characteris-

tics of data sensitivities together with measures employed to mitigate sharing issues [27]. With minor modifications, such measures have been applied to the collaborative data science problem in terms of data and/or the models themselves. These measures can be considered to fall into three broad categories: *social measures*, *environmental constraints*, and *anonymisation methods*. Depending on the sensitivities concerned, measures from one or more of these categories may be required to ensure compliance with data assurance mandates and expectations.

Social measures address internal threats via a range of permissions and/or structural organisational changes [27]. Examples of these include security clearances, non-disclosure agreements, and procedural or legal dispensations that facilitate or broaden data sharing. More complex measures include the secondment of individuals between organisations, or the reassignment of jurisdictional responsibilities to shift data custodianship and render sensitivities moot. In some contexts, sensitivities may be reduced though the explicit solicitation of consent for data sharing or — depending on context and prevailing regulation — a waiver of some privacy rights at the time of data collection.

Environmental constraints aim to limit the physical contexts of access to prevent the (internal) inadvertent exposure, intentional sharing, or (external) theft of sensitive material beyond the limits imposed by social measures. Material may be encrypted both during transmission to the access context and at rest when not in active use. More significant has been the development of *data safe havens*; secure analytics environments that comprise "appropriate technical and governance controls which are effectively audited and are viewed as trustworthy by diverse stakeholders" [8, p. 3243]. In this case, sensitive material is typically copied into the data safe haven (often managed by a trusted third-party institution such as a university) where suitably authorised stakeholders can remotely access data. Examples of such data safe havens include the UK's Secure Anonymised Information Linkage (SAIL) Databank of health and other public service data [29], and the Australian E-Research Institutional Cloud Architecture (ERICA) [4].

Data safe havens employ physical, technological, and procedural, mechanisms to "store and release data faithfully and effectively" in a way that can be "viewed as safe and trustworthy by all key stakeholders" [8, p. 2345]. Physical safeguards include strict control of locations where material may be prepared, or up/down-loaded. Technical safeguards may include dynamically provisioned, virtualised, and sand-boxed workspaces; multi-factor authentication; and detailed activity audit logging. Procedural safeguards may include access agreements; standardised work flows; vetting of outputs prior to their release from workspaces, and sanctions for breaches of agreed conduct. However, despite these measures, the safe haven approach is predicated on direct access to data, and is difficult to comprehensively protect against the broad range of internal and external threats [41, 14].

Anonymisation Methods can be employed to mitigate both internal and external threats in cases where the identity of data subjects cannot be disclosed, even though the bulk of the material may be shared [24]. At the most basic level, data may be filtered with individual identifying fields — for example, the 18 U.S. HIPPA 'Safe Harbour' identification fields [36] — or metadata removed or masked in order to anonymise records. In some contexts, datasets can be permuted and 'sliced' so that a complete view of the material is not disclosed [32]. With numeric data, other methods can be employed such as top/bottom-coding (removal of identifiable outliers), aggregation of similar records, and averaging of within-group values. Perturbation approaches such as *differential privacy* involve the introduction of random error or 'noise' that "addresses the paradox of learning nothing about an individual while learning useful information about a population" [20, p. 5], serving to improve the anonymity of data subjects [27, 24].

Models may also be vulnerable to privacy-disclosing threats. In some cases analysis of internal model parameters can be used to reconstruct the original training data [44]. Similarly, repeated and carefully-designed queries that leverage prediction confidence levels of otherwise opaque models can

be used to reverse-engineer training data [23]. Depending on the nature of data and algorithms concerned, some models themselves internally implement transforms such as differential privacy to mitigate such attacks [1]. The training task can also be partitioned into 'teacher' and 'student' models for knowledge transfer in a kind of generative adversarial network [38] that mitigates against reconstruction and model inversion attacks. However, none of these measures address the primary restricted data problem.

2.2. Limitations of sensitive data measures

While such measures can go a long way to protect the sensitivities of material, the efficacy of their use is often highly contextual and subject to a range of limitations. Social and environmental measures are not absolute and may conceivably be circumvented either inadvertently or deliberately by end-users, or by third-parties with malicious intent [14]. For example, reliance on one-time agreements and vetting may give rise to a false sense of security as, over time, workflow pipelines may increase in scope and reach; the motivations of individuals or project partners may change; or end-user equipment may become compromised by malicious third parties. Similarly, operational data often undergoes continuous churn. This continual data 'drift' means that models need to be continually updated and, possibly re-worked; further straining 'one-shot' paradigms of sensitivity analysis. Technological measures may fail due to misunderstanding or misapplication [13], or through poor framing of risks (e.g., where output controls are circumvented by 'screen-capture' mechanisms). Continuous risk-based assessment may be more effective than static designs based on measures of 'safety' [14].

The anonymisation methods described above are a trade-off between the anonymity of data subjects and the fidelity of the data [24]. In some contexts, obscured data elements may compromise data science workloads. Moreover, such methods, may not be applicable or practical for other types of data such as text, image, audio, or video. A larger issue, however, is the notion of data anonymity itself. In practice, re-identification of 'de-identified' data may occur via a number of mechanisms either deliberately or inadvertently by end users, or due to increased scrutiny following a data breach [21]. Together, these factors mean that reliance on anonymisation methods to protect sensitive data may be misplaced.

Finally, all of these measures are predicated on the assumption that at least some of the data may be shared in the first place. Whether due to the above-mentioned concerns, or a-priori restrictions on the disclosure of data outside an organisational context, there are many circumstances in which these measures are not applicable for highly sensitive or restricted data. In cases where direct access to such data is impractical, unethical, or illegal, *eyes-off* mechanisms (i.e. without direct access to data) are needed for the development, testing, comparison, and integration of analytics models.

2.3. Dealing with restricted data and models

While many of these social measures and anonymisation methods are insufficient or inappropriate for use in eyes-off environments, there remain a number of environmental constraints that may be used to protect against specific threats.

A form of data safe-haven can be employed that incorporates a *Trusted Execution Environment* (TEE). TEEs have varying proprietary hardware implementations, but generally comprise cryptographically secure memory and computational partitions that enable isolation of specific workloads from others running on the system, and remote attestation of workload integrity [45]. They can be particularly applicable to virtualised environments, where sensitive workloads may need to be protected from threats originating from hypervisor hosts and/or other virtual guests. TEE workloads are often required to be built with environment-specific software development tool-chains; initialising and loading workloads into

isolation partitions as well as unloading outputs upon completion.

However, TEE design has no inherent separation between a trusted workload and its data. Additional protocols (such as code inspection, encryption of data, secure encryption key loading, and so on) are needed to ensure eyes-off access to data by a given workload [37]. Additionally, a TEE often imposes a performance overhead due to a number of factors. In the case of data science workloads, this overhead can be significant, and can result in a performance degradation of an order of magnitude or more [3]. The trade-offs between TEE security and these drawbacks along with the difficulties in TEE implementation [12] and a range of (admittedly complex) hardware attack vulnerabilities [22], mean that a TEE may not yet be an effective element of an eyes-off data science solution.

Another hardware device that may be employed to mitigate some internal and external threats is the *Data Diode*. Data diodes are hardware devices that physically enforce a one-way flow of data between nodes or networks, ensuring "that no data can be passed, either explicitly or covertly, in the opposite direction" [48, p. 1]. In this way data may be transmitted while assuring both the integrity of the sender and the impossibility of other data being exfiltrated from the sender via that channel. While a data diode does not directly support eyes-off access of data, and has the drawback that human intervention is required to verify receipts and request re-transmission if required, it may be a useful component for use in highly secure networks.

A fourth set of measures involves the use of provably *secure protocols* that facilitate the cooperation of two or more parties in the computation of results from data held privately by each party [33]. While such techniques can address some of the issues surrounding cooperative training or inference, they do not address the fundamental issues of eyes-off access to another party's data. More promising is *Homomorphic Encryption* [2] that enables the computation of functions against encrypted data that return encrypted results to the querier. These functions offer secure equivalents to functions operating on unencrypted data. If used with asymmetric key schemes, such techniques can facilitate secure, eyes-off access — albeit with limited function primitives and a severe performance penalty. In relation to machine learning workloads, this effort has largely focused on inference tasks against encrypted data (rather than the initial training), and even then, with simplified data and/or models [9, 51].

Thus the difficulties of dealing with restricted data, outside of its operational or custodial context [17], are compounded by the impracticality of 'simple' anonymisation methods that reduce the utility of data on the one hand, and the inapplicability of complex transformations to generalised data science workloads the other. What is required is a third way; infrastructure that facilitates eyes-off application of data science workloads involving restricted data and/or models. However, before diving into the architecture of such an infrastructure, we shall describe our motivating use-case for this work.

2.4. The CSAM use case

The motivation for our work is a collaboration with a national law enforcement agency to develop tools to aid the investigation of Child Sexual Abuse Material (CSAM). Recent years have seen rampant growth of production and dissemination of such material [7, 28]. The sheer number of items that need to be analysed, is straining limited law enforcement and juridical resources [16]. Additionally, repeated and ongoing exposure to often highly disturbing imagery by law enforcement, forensic, and juridical officers (and others such as those performing ICT services or undertaking R&D in the area) is an increasing source of secondary trauma [46]. While not entirely removing the need for human analysis, automated triaging of CSAM goes some way to reducing damaging exposure.

Our aim is to develop machine learning models for the automated detection, identification, and categorising of such material. This is a challenging task, not only in terms of the technical difficulties in model development, and the risks associated with working with disturbing material, but also because of

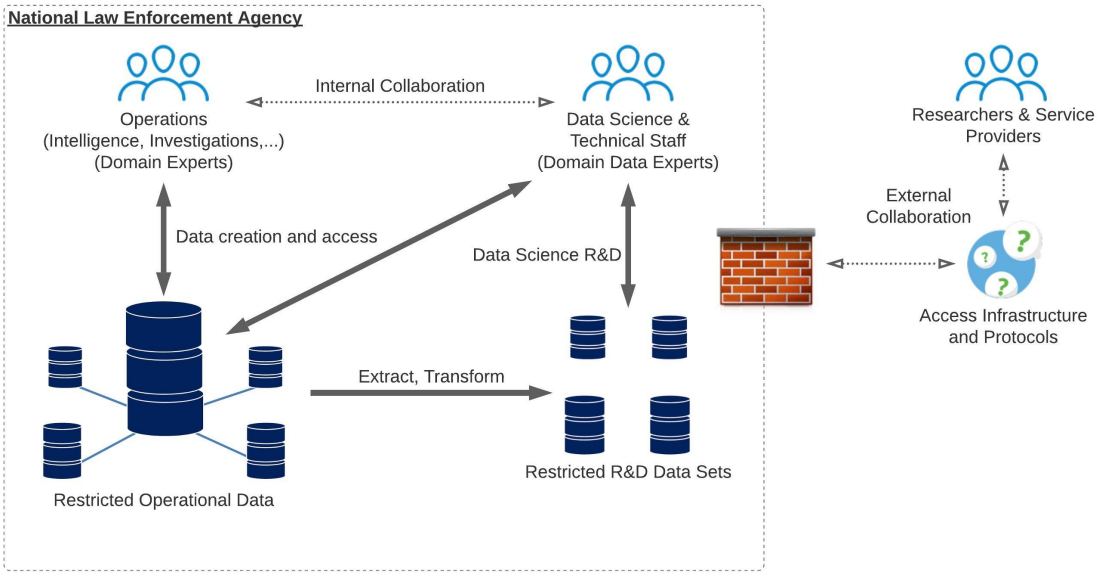


Figure 2.1: Generalised Restricted Data cse case.

restricted access to this data. Our use case is therefore a good example of highly sensitive operational data required for ongoing research and development as shown in Figure 2.1.

Within the national law enforcement agency, internal stakeholders generate and reference operational data as part of their day-to-day workflow. This sensitive, internal data can take a multitude of forms including structured data, images, text, video, audio, and binary (e.g. device images) along with contextual metadata. These internal *domain expert* stakeholders rely on technical staff to develop and enhance intelligence and investigative capabilities. These technical stakeholders are *domain data experts*, with an understanding of the formats, volumes, and content of this data together with an appreciation of the technical and social challenges of working with it.

The problem arises when those involved in capability development wish to collaborate outside the organisation. In our case, the data sets comprising CSAM are subject to the legislative restriction of any “material that depicts or describes activity relating to child sexual abuse” [10]. This led directly to the creation of infrastructure and protocols that facilitate such collaboration without the possibility of breaching the legal restrictions placed on the data.

3

Initial Data Airlock Design

This infrastructure was dubbed 'Data Airlock' to emphasise the separation of workloads from restricted data — in our case, the CSAM data sets used to train, validate, and test models as they underwent development. As successful models would be deployed into law enforcement production, there was no requirement for this infrastructure to handle the inference case (although model validation was required in addition to training). Similarly, there was no requirement to deal with model sensitivity.

3.1. Architecture

The high-level architecture for this iteration, shown in Figure 3.1, comprises software and hardware components located in a secure data centre. This infrastructure shares some similarities with a data safe haven, inasmuch as it is a separate and partitioned computing platform accessed remotely by users and administrators. However, the secure connectivity and interaction between the isolating partitions, as well as the treatment of the restricted data are markedly different.

The architecture is divided into three logical zones — the *public* zone accessed by external R&D collaborators and workflow administrators; the high-performance *secure* zone in which validated workloads run against restricted data; and the *restricted* zone where encrypted sensitive data is stored. The three zones operate under different security and access models. In our implementation, the public zone nodes are virtualised on a single server. The secure zone is necessarily located on its own high performance server. The restricted zone runs on a general purpose computer with high-performance, encrypted storage.

Public zone access is provided via a web application. Internet access (web and secure shell) access is only available to the public zone, while ssh access to the secure zone (e.g. for system administration) is permitted on a separate connection from the public zone. Within the restricted zone, all manual activity (e.g. software maintenance; data loading; etc.) takes place via physical, on-site access. The roles of these zones and their connectivity are explained in more detail as follows:

3.1.1. Public Zone

The public zone provides a *Consumer Node* web interface to external collaborators — in our case, researchers developing machine-learning models. This web app is accessible remotely, and secured using Auth0 [5]. The two consumer functions are (i) specification of the workload and uploading of code implementing the model (including pre/post processing stages such as data wrangling; training-

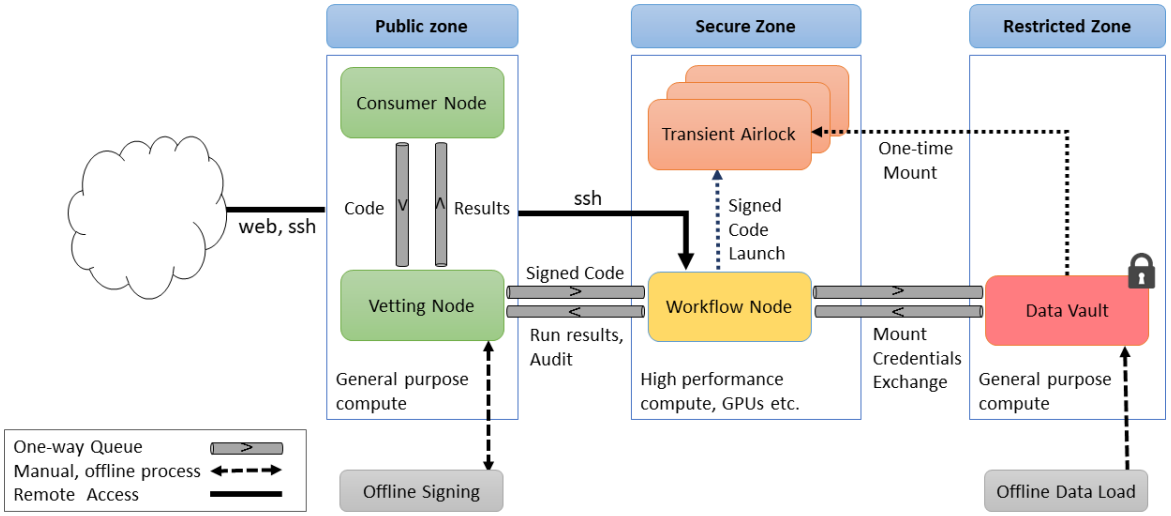


Figure 3.1: Airlock 1.0 Architecture.

validation splits; hyperparameter settings; and so on) to be vetted and subsequently run; and (ii) retrieval of vetted results from the run.

As will be described below, the secure zone is implemented on NVIDIA DGX hardware, enabling the use of pre-built NVIDIA CUDA docker containers [35]. The use of standardised and externally maintained containers removes the necessity for collaborators to replicate the secure zone runtime environment while reducing the effort of maintaining runtime environments, and the likelihood of environmental security holes. It also enables input vetting of comparatively small pieces of code, as opposed to whole containerised runtime environments.

However, this standardisation necessarily introduces constraints for the model development, limiting the model code to be written in Python (or code with a Python wrapper and bindings), and forcing the use of standard Python data science support libraries for model development. These constraints provided no difficulties for our project, which employed standard Python model development.

The public zone also provides a *Vetting Node* interface for data custodian workflow administrators — in our case, data sciences staff from the national police agency. As with the consumer node, this interface is served by the web app and employs one-way, persistent queues to interact with other components of the infrastructure.

The code vetting is a social measure whereby the job (model and pre/post processing code) enqueued from the public zone is manually analysed by a workflow administrator for elements that would, intentionally or inadvertently, exfiltrate restricted data from the sensitive zone. Once vetted, a cyrptographic hash of the code is signed by a vetter (using a nonce to prevent relay attacks) to ensure integrity of code executed in the secure zone. Importantly, as no private keys are held on the infrastructure, the actual signing takes place offline on a vetter private system. Once signed the code is then enqueued for execution in the secure zone.

Similarly, results from a job run within the secure zone are enqueued for review in the vetting node, in this case, manual examination for evidence of exfiltrated data in the output. If acceptable, the results are then enqueued for reception by the submitter. These results retrieved at the consumer node include notification of the failure of the code to pass vetting; statuses of the job in the various queues; and/or the outputs of the job in term of reports, data files, or images (e.g. performance graphs).

This inherently a batch-oriented system. From the point of consumer submission onward, there is no run-time interaction with the job save for the input/output vetting.

3.1.2. Secure Zone

The secure zone contains a high-performance NVIDIA DGX-1 computing node that runs models on the data in isolated virtual environments or *airlocks* implemented using Docker [26] containers.

Jobs are dequeued within the sensitive zone by an automated process in which the code signature is checked using public vetter keys. If successful, one-time credentials for data access are created and the appropriate standard NVIDIA docker container is selected. This transient airlock is then launched with the vetted code as its entry point and the restricted data is mounted using the one-time credentials. Each airlock is also provided with ephemeral workspace and output mounts to which it can write data. At the completion of the job, the restricted data is unmounted; the resulting model, any processing results, and output logs are enqueued for vetting; and the temporary mounts are destroyed.

3.1.3. Restricted Zone

Up to this point the Data Airlock infrastructure resembles many aspects of a typical safe data haven. However, unlike safe-haven sand-boxed environments such as ERICA or SAIL, it is the restricted zone design that departs from the safe haven concept (albeit with extra layers of potentially manual vetting).

The restricted zone *data vault* provides secure storage for sensitive data that is physically loaded on-site by data custodians into volumes encrypted with a manual boot-time password. The data vault dequeues secure zone requests for access to restricted data and returns one-time credentials for that access. The restricted data volume is then made available to the secure zone when requested, secured by those one-time credentials.

3.2. Outcomes

The Data Airlock architecture proved successful for our project, enabling the collaborative development and improvement of CSAM classifier models without requiring external research staff to directly assess restricted CSAM material. The important term here is *collaboration*, as the very nature of restricted data does not permit the wholesale outsourcing of data science R&D against ‘raw’ operational data. Instead, domain data experts within the data custodian organisation are needed to perform any initial data preparation to render the raw operational data into a form amenable for model development. While agreement on data formats, label schemas, and label element schemes can be reached between all collaborators, this initial bootstrapping of the data needs to be performed ‘in-house’. Similarly, any labelling obviously needs to be performed by the data custodian — in our case, labels were derived from criminal investigation work that had previously assessed and categorised CSAM material.

In our project, some of the data pipeline work was performed as part of the external workflow in the secure zone — for example, de-duplication of images using (restricted) perceptual hash sets [19], downscaling of images, etc. Ultimately, though, such wrangling needs to be conducted in collaboration with domain data experts to ensure data quality is maintained. Without the ability to scrutinise data (and labels), and processing results, the detection of data, labelling, and/or pre-processing anomalies is difficult. To this end, an improvement to the Data Airlock workflow would be a formal mechanism for querying or reporting data/labelling anomalies, changes to data characteristics, and so on. This requirement points to the necessity for early and ongoing collaboration with domain data experts to maintain the integrity of model R&D as data drifts, requirements change, and domain understanding improves.

Beyond the data bootstrapping and wrangling issues, this initial iteration of the architecture also exposed a number of other issues and opportunities for further improvement.

3.2.1. Scope and scale

Firstly, as mentioned above, the Data Airlock was only designed to protect restricted data — not models. In our case the models being developed were for use exclusively within the national law enforcement agency and our researchers were trusted to not disclose models to any third party. However, in the general case, these social measures may not be sufficient with the Data Airlock infrastructure needing to protect models in addition to data. Scenarios for such use may include the assured benchmarking or comparison of proprietary models against a standard (perhaps restricted) dataset; research in hardening models against threats; and just generally restricting the possibility of analytical models "escaping into the wild" which could be problematic in law enforcement, security, or defence contexts.

Similarly, this implementation was not designed to support a plurality of collaborators and data custodians from multiple organisations. Although this was not an issue for us, it is certainly a requirement for a more generalised solution to the restricted data problem. A federated platform for controlled and configurable eyes-off access to restricted data would enable deep and broad collaboration between a range of disparate data-holders and researchers for the training, testing, and comparison of models against data that is held elsewhere.

Such an infrastructure would need to incorporate federated authentication, authorisation, workflow management, and audit. Data-custodians would create catalogue entries of pre-vetted jobs, tasks, and data recipes etc. that run against their secure zone compute resources and restricted zone data, and then grant access to these in much the same manner that vetting is currently performed. From a collaborator perspective, these pre-vetted jobs, tasks, and data recipes would be combined to run across multiple organisation datasets/models in a federated, standardised, and secure manner.

Another limitation was the web interfaces of the public zone and the 'bare metal' interface of the restricted zone. These gave rise to a lack of integration of the airlock with data custodian and collaborator workflows. A more generalised Data Airlock should provide some support for integration with operational data workflow (perhaps using data diodes); for example, versioning of datasets, coordination of data maintenance with the availability of data for mounting, and so on. Similarly, integration with collaborator workflows could be effected through the use of an SDK or API to reduce (but, importantly, not eliminate) instances of manual intervention by researchers to submit workloads and receive results.

3.2.2. Performance improvements

Our project exposed a number of other bottlenecks that could be improved in future versions. For example, the PKI environment used for code signing etc. required the manual uploading of encryption keys into the restricted zone system. A superior mechanism for this could be the use of some form of tamper-resistant hardware for key management (such as a hardware security module) that would reduce the complexity and improve security of such maintenance.

Similarly, the vetting of submitted workloads and returned outputs was an entirely manual activity. It may be that portions of such vetting could be performed in an automated fashion, perhaps differentiating boilerplate code and results from bespoke sections, using machine learning techniques as part of the workflow infrastructure. This is a research area in its own right, perhaps along the lines of machine learning anomaly or intrusion detection in high performance environments described by Peisert [41].

Finally, it should be noted that the workload scheduler in this initial architecture was single-threaded, enabling one isolated airlock at a time to execute, providing access to the full complement of compute resources (e.g. CPU/GPU/RAM) available on the secure zone server. A generalised architecture should allow for more granular and parallel execution of isolated airlocks.

3.2.3. Data mounting improvements

If such parallelism were implemented, then the isolation of container-based mounts would become a priority for a generalised solution. Currently, due to standard docker sandboxing, the restricted data mounts are not privileged and are available to the whole (albeit, currently single airlock instance-at-a-time) secure zone server. For parallel container execution, the ability to isolate ephemeral data mounts to particular transient containers will be necessary.

Along these lines, increasing the granularity of data that is mounted would both tighten security, and help with the coordination of data maintenance and the mounting of data for workflows.

In a similar vein to this data granularity, is it possible, though unlikely, for an intruder in the secure zone to theoretically spoof a mount request, obtaining access to the content of the data vault. One protection against this scenario would be to increase the granularity of signed workflow elements along with code — in this case, some form of signed mount request — that would be checked within the restricted zone. Of course, this would increase the complexity of vetting and signing operations, but a semi-automated custodian workflow described above may offset this impost.

4

Conclusion

In this report, we have discussed how a range of social, environmental, data transform, and secure protocol measures can address various internal and external threats to sensitive data, but remain insufficient to protect restricted data in collaborative contexts. On the other hand, a new, purpose-built infrastructure, architecture dubbed 'Data Airlock' has demonstrated its utility in our project concerning restricted law enforcement data.

Our experience in deploying and using the Data Airlock has exposed a number of assumptions and shortcomings in our implementation, leading to a an additional set of requirements for interoperability and scalability in order to support a distributed and heterogeneous research community. The design and implementation of a second version of the Data Airlock is currently underway that will open up possibilities across research domains; running or comparing models without disclosing their technical detail and applying different restricted data 'recipes' based on access and trust criteria.

Such a highly-secure, federated platform will enable controlled 'eyes-off' access to large, sensitive data sets in order to facilitate collaboration between disparate data-holders and researchers across a variety of problem domains, including law-enforcement, defence, security, medicine, and social services.

References

[1] Martin Abadi et al. “Deep learning with differential privacy”. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. Vienna, Austria: ACM, 2016, pp. 308–318. DOI: 10.1145/2976749.2978318.

[2] Abbas Acar et al. “A survey on homomorphic encryption schemes: Theory and implementation”. In: *ACM Computing Surveys (CSUR)* 51.4 (2018). Publisher: ACM New York, NY, USA, pp. 1–35. ISSN: 0360-0300.

[3] Ayaz Akram et al. “Performance Analysis of Scientific Computing Workloads on Trusted Execution Environments”. en. In: *arXiv:2010.13216 [cs]* (Oct. 2020). arXiv: 2010.13216. URL: http://arxiv.org/abs/2010.13216 (visited on 12/08/2020).

[4] Australian Research Data Commons. *Secure Cloud Computing for Sensitive Data*. 2020. URL: https://ardc.edu.au/project/secure-cloud-computing-for-sensitive-data/ (visited on 12/02/2020).

[5] Auth0 Inc. *Auth0 Secure access for everyone. But not just anyone.* en. 2021. URL: https://auth0.com/ (visited on 02/18/2021).

[6] Umang Bhatt et al. “Machine Learning Explainability for External Stakeholders”. en. In: *arXiv:2007.05408 [cs]* (July 2020). arXiv: 2007.05408. URL: http://arxiv.org/abs/2007.05408 (visited on 11/20/2020).

[7] Elie Bursztein et al. “Rethinking the detection of child sexual abuse imagery on the Internet”. In: *WWW ’19: Proceedings of The Web Conference 2019*. San Francisco, CA, USA, 2019, pp. 2601–2607. DOI: 10.1145/3308558.3313482.

[8] Paul R. Burton et al. “Data Safe Havens in health research and healthcare”. en. In: *Bioinformatics* 31.20 (Oct. 2015), pp. 3241–3248. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btv279. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv279 (visited on 12/01/2020).

[9] Edward Chou et al. “Faster CryptoNets: Leveraging Sparsity for Real-World Encrypted Inference”. en. In: *arXiv:1811.09953 [cs]* (Nov. 2018). arXiv: 1811.09953. URL: http://arxiv.org/abs/1811.09953 (visited on 02/09/2021).

[10] Commonwealth of Australia. *Criminal Code Act 1995*. 2020.

[11] Commonwealth of Australia. *Data Skills and Capability in the Australian Public Service*. Tech. rep. Canberra, Australia: Commonwealth of Australia, Department of the Prime Minister and Cabinet, 2016. URL: https://www.pmc.gov.au/sites/default/files/publications/data-skills-capability.pdf (visited on 11/11/2020).

[12] Victor Costan and Srinivas Devadas. “Intel SGX Explained.” In: *IACR Cryptol. ePrint Arch.* 2016.86 (2016), pp. 1–118. URL: https://eprint.iacr.org/2016/086.pdf (visited on 11/11/2020).

[13] Chris Culnane, Benjamin IP Rubinstein, and Vanessa Teague. “Vulnerabilities in the use of similarity tables in combination with pseudonymisation to preserve data privacy in the UK Office for National Statistics’ Privacy-Preserving Record Linkage”. In: *arXiv preprint arXiv:1712.00871* (2017).

[14] Dr Chris Culnane. “Not fit for Purpose: A critical analysis of the ‘Five Safes’”. en. In: *rXiv preprint arXiv:2011.02142* (2020), p. 8.

[15] Bart Custers. “Click here to consent forever: Expiry dates for informed consent”. In: *Big Data & Society* 3.1 (2016). ISBN: 2053-9517 Publisher: SAGE Publications Sage UK: London, England. DOI: 10.1177/2053951715624935.

[16] Janis Dalins et al. “Laying foundations for effective machine learning in law enforcement. Majura–A labelling schema for child exploitation materials”. In: *Digital Investigation* 26 (2018), pp. 40–54. ISSN: 1742-2876. DOI: 10.1016/j.diin.2018.05.004.

[17] Data Republic. *Why Five Safes Aren’t Enough for Inter-Organizational Data Exchanges*. en. Section: Team blog. Jan. 2019. URL: https://www.datarepublic.com/resources/team-blog/five-safes-inter-organizational-data-exchange (visited on 12/02/2020).

[18] Deshpande. *Revolutionizing Data Collaboration with Federated Machine Learning*. June 2020. URL: https://www.datanami.com/2020/06/12/revolutionizing-data-collaboration-with-federated-machine-learning/ (visited on 11/19/2020).

[19] Ling Du, Anthony TS Ho, and Runmin Cong. “Perceptual hashing for image authentication: A survey”. In: *Signal Processing: Image Communication* 81 (2020). Publisher: Elsevier. ISSN: 0923-5965. DOI: doi.org/10.1016/j.image.2019.115713.

[20] Cynthia Dwork and Aaron Roth. “The algorithmic foundations of differential privacy.” In: *Foundations and Trends in Theoretical Computer Science* 9.3-4 (2014), pp. 211–407.

[21] Khaled El Emam et al. “A systematic review of re-identification attacks on health data”. In: *PloS one* 6.12 (2011). Publisher: Public Library of Science, e28071. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0126772.

[22] Thomas Claburn in San Francisco. *Intel’s SGX cloud-server security defeated by $30 chip, electrical shenanigans*. en. Nov. 2020. URL: https://www.theregister.com/2020/11/14/intel_sgx_physical_security/ (visited on 11/26/2020).

[23] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. “Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures”. en. In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. Denver Colorado USA: ACM, Oct. 2015, pp. 1322–1333. ISBN: 978-1-4503-3832-5. DOI: 10.1145/2810103.2813677. URL: https://dl.acm.org/doi/10.1145/2810103.2813677 (visited on 12/15/2020).

[24] Simson L Garfinkel. *De-identification of personal information*. Tech. rep. NISTIR 8053. Gaithersburg, USA: National Institute of Standards and Technology, 2015. URL: http://dx.doi.org/10.6028/NIST.IR.8053 (visited on 11/11/2020).

[25] Graham Greenleaf. “Global data privacy laws 2019: 132 national laws & many bills”. In: *157 Privacy Laws & Business International Report*. Privacy Laws & Business, 2019. URL: https://ssrn.com/abstract=3381593.

[26] Docker Inc. *Empowering App Development for Developers*. en. 2021. URL: https://www.docker.com/ (visited on 02/18/2021).

[27] Thomas B Jabine. “Procedures for restricted data access”. In: *Journal of Official Statistics* 9.2 (1993). Publisher: Statistics Sweden (SCB), p. 537. ISSN: 0282-423X.

[28] Alexis Jay et al. *Investigation Report*. Tech. rep. Independent Inquiry into Child Sexual Abuse, 2020. URL: https://www.iicsa.org.uk/publications/investigation/internet (visited on 11/11/2020).

[29] Kerina H. Jones et al. “A Profile of the SAIL Databank on the UK Secure Research Platform”. en. In: *International Journal of Population Data Science* 4.2 (2019). Number: 2. ISSN: 2399-4908. DOI: 10.23889/ijpds.v4i2.1134. URL: https://ijpds.org/article/view/1134 (visited on 12/01/2020).

[30] Mouna Jouini, Latifa Ben Arfa Rabai, and Anis Ben Aissa. “Classification of security threats in information systems”. In: *Procedia Computer Science* 32 (2014). Publisher: Elsevier, pp. 489–496. ISSN: 1877-0509.

[31] Ninghui Li et al. “Membership privacy: a unifying framework for privacy definitions”. en. In: *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security - CCS '13*. Berlin, Germany: ACM Press, 2013, pp. 889–900. ISBN: 978-1-4503-2477-9. DOI: 10.1145/2508859.2516686. URL: http://dl.acm.org/citation.cfm?doid=2508859.2516686 (visited on 12/16/2020).

[32] Tiancheng Li et al. “Slicing A new approach for privacy preserving data publishing”. In: *IEEE transactions on knowledge and data engineering* 24.3 (2010). Publisher: IEEE, pp. 561–574. ISSN: 1041-4347. DOI: 10.1109/TKDE.2010.236.

[33] Yehuda Lindell and Benny Pinkas. “Secure Multiparty Computation for Privacy-Preserving Data Mining”. en. In: *The Journal of Privacy and Confidentiality* 1.1 (2009), pp. 59–98. URL: https://journalprivacyconfidentiality.org/index.php/jpc/article/download/566/549/ (visited on 12/12/2020).

[34] Slava Jankin Mikhaylov, Marc Esteve, and Averill Campion. “Artificial intelligence for the public sector: opportunities and challenges of cross-sector collaboration”. en. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2128 (Sept. 2018), p. 20170357. ISSN: 1364-503X, 1471-2962. DOI: 10.1098/rsta.2017.0357. URL: https://royalsocietypublishing.org/doi/10.1098/rsta.2017.0357 (visited on 11/18/2020).

[35] NVIDIA Corporation. *Catalog*. 2021. URL: https://ngc.nvidia.com/catalog/containers (visited on 02/16/2021).

[36] Office for Civil Rights. *Methods for De-identification of PHI*. en. Text. Sept. 2012. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html (visited on 12/03/2020).

[37] Olga Ohrimenko et al. “Oblivious Multi-Party Machine Learning on Trusted Processors”. en. In: *Proceedings of the 25th USENIX Security Symposium*. Austin, TX, USA: USENIX, Aug. 2016, p. 19.

[38] Nicolas Papernot et al. “Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data”. en. In: *arXiv:1610.05755 [cs, stat]* (Mar. 2017). arXiv: 1610.05755. URL: http://arxiv.org/abs/1610.05755 (visited on 12/15/2020).

[39] Samir Passi and Steven J. Jackson. “Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects”. en. In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (Nov. 2018), pp. 1–28. ISSN: 2573-0142, 2573-0142. DOI: 10.1145/3274405. URL: https://dl.acm.org/doi/10.1145/3274405 (visited on 11/19/2020).

[40] Nathan Peiffer-Smadja et al. “Machine Learning for COVID-19 needs global collaboration and data-sharing”. en. In: *Nature Machine Intelligence* 2.6 (June 2020). Number: 6 Publisher: Nature Publishing Group, pp. 293–294. ISSN: 2522-5839. DOI: 10.1038/s42256-020-0181-6. URL: https://www.nature.com/articles/s42256-020-0181-6 (visited on 11/19/2020).

[41] Sean Peisert. “Security in high-performance computing environments”. In: *Communications of the ACM* 60.9 (2017). Publisher: ACM New York, NY, USA, pp. 72–80. ISSN: 0001-0782. DOI: 10.1145/3096742.

[42] Raluca Ada Popa and Anila Joshi. *Training a machine learning model with secure collaborations*. en. Last Modified: 2020-03-24T10:23:32+00:00. Feb. 2020. URL: https://www.ericsson.com/en/blog/2020/2/training-a-machine-learning-model (visited on 11/19/2020).

[43] Drew Roselli, Jeanna Matthews, and Nisha Talagala. “Managing bias in AI”. In: *Companion Proceedings of The 2019 World Wide Web Conference*. San Francisco, CA USA, 2019, pp. 539–544. DOI: 10.1145/3308560.3317590.

[44] Mohammad Al-Rubaie and J Morris Chang. “Privacy-preserving machine learning: Threats and solutions”. In: *IEEE Security & Privacy* 17.2 (2019). Publisher: IEEE, pp. 49–58. ISSN: 1540-7993. DOI: 10.1109/MSEC.2018.2888775.

[45] Mohamed Sabt, Mohammed Achemlal, and Abdelmadjid Bouabdallah. “Trusted execution environment: what it is, and what it is not”. In: *Proceedings of the 14th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*. Vol. 1. Helsinki, Finland: IEEE, 2015, pp. 57–64. ISBN: 1-4673-7952-2. DOI: 10.1109/Trustcom.2015.357. URL: https://hal.archives-ouvertes.fr/hal-01246364/file/trustcom_2015_tee_what_it_is_what_it_is_not.pdf (visited on 11/11/2020).

[46] Kathryn C Seigfried-Spellar. “Assessing the psychological well-being and coping mechanisms of law enforcement investigators vs. digital forensic examiners of child pornography investigations”. In: *Journal of Police and Criminal Psychology* 33.3 (2018). Publisher: Springer, pp. 215–226. ISSN: 0882-0783.

[47] Reza Shokri and Vitaly Shmatikov. “Privacy-Preserving Deep Learning”. en. In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS ’15*. Denver, Colorado, USA: ACM Press, 2015, pp. 1310–1321. ISBN: 978-1-4503-3832-5. DOI: 10.1145/2810103.2813687. URL: http://dl.acm.org/citation.cfm?doid=2810103.2813687 (visited on 12/15/2020).

[48] Malcolm W Stevens. *An implementation of an optical data diode*. Tech. rep. DSTO-TR-0785. Salisbury, South Australia: Defence Science and Technology Organisation, 1999, p. 31. URL: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.8650 (visited on 11/11/2020).

[49] Tara Qian Sun and Rony Medaglia. “Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare”. In: *Government Information Quarterly* 36.2 (2019). Publisher: Elsevier, pp. 368–383. ISSN: 0740-624X.

[50] Stijn Viaene. “Data Scientists Aren’t Domain Experts”. In: *IT Professional* 15.6 (Dec. 2013), pp. 12–17. ISSN: 1941-045X. DOI: 10.1109/MITP.2013.93.

[51] Weiru Wang et al. “Homo-ELM fully homomorphic extreme learning machine”. en. In: *International Journal of Machine Learning and Cybernetics* 11.7 (July 2020), pp. 1531–1540. ISSN: 1868-8071, 1868-808X. DOI: 10.1007/s13042-019-01054-w. URL: http://link.springer.com/10.1007/s13042-019-01054-w (visited on 02/09/2021).