Version 1.0.0
14. Dezember 2022
SEMLA.DFKI.DE

SEMLA

An On-Premises Trusted Research Environment for AI-based R&D with Sensitive Personal Information

Jan Alexandersson, Jochen Britz, Valentin Seimetz, Daniel Tabellion
DFKI GmbH

1. Introduction

We are witnesses to huge and rapid advancements in technical systems based on artificial intelligence (AI). The steady growth in computing power, both in terms of CPU and GPU, along with free access to machine learning libraries fuels this development and the sheer output from the research community as well as the emergence of products with AI technology on the market is impressive. Data of all kinds, be it weather, financial developments, local traffic information alongside personal preferences, biometric data and even diseases make it possible to provide services that go beyond imagination. However, we also see developments where this kind of data is used for unintended or undesired purposes: once information has been published globally on the internet or collected within datasets, not necessarily permitted by the data owner, it is nearly impossible to undo or delete this information.

In Europe, the enactment of the General Data Protection Regulation (GDPR) has on the one hand raised awareness of these risks and, on the other hand, prescribed how personal data must be treated, thus increasing demand for fair and secure handling of this data. GDPR put personal data of individuals and their right to own, grant usage and retract consent at its core[1]. The research community is obliged to follow strict rules, e.g., minimizing data usage, pseudonymizing and/or anonymizing data and even deleting data after the research task has been accomplished. Consequently, within research and development there is a need for a new level of so-called Trusted Research Environments (TRE) where the GDPR regulations are in focus. Fortunately, we witness well-founded advancements in this area too. Both cloud-based approaches, like (O'Reilly, 2020; Arenas et al, 2019) as well as classification schemata, e.g. (LfD, 2018; TCPD, 2016 – both in German) provide a solid basis for the infrastructure that implements most of European and international legislation requirements on processing security. The latter schemata also consider cases which go beyond this work, such as data concerning witness-protection program and state security.

This work is mainly motivated by DFKI’s involvement in research and development of biomarkers, digital phenotyping, and decision support systems in the healthcare domain. Some examples are:

- In the project KIttata (Kiefer et al, 2022), the quality of donor corneas is determined in an interactive system consisting of a combination of different AI techniques
- In the project KI@HOME (KI@HOME) information from a smart home sensor-set along with health insurance information is used to predict adverse events
- In the project MePheSTO (MePheSTO; König et al, 2022), biomarkers are researched based on audio, video, and wearables during social interaction in the psychiatric domain

In TREs, the sensitivity of the data determines the degree of measures to handle the data. According to GDPR, anonymized or pseudonymized data demands less strict measures than

[1] EU Charter of Fundamental Rights (Art. 8) states:
personal data "… must be processed fairly for specified purposes and on the basis of consent of the person concerned or some other legitimate basis laid down by law."

plain personal data like videos from social interactions between clinicians and patients. Sensitivity can be classified along well-known classification schemes already used within political and military environments. (O'Reilly, 2020) suggests a five-class schema – tiers – ranging from "0: open data" all the way to "4: Very sensitive personal, … data" and suggest Technical and Organizational Measures (TOM) that implements appropriate measurements. For the cloud-based approach in (Arenas et al, 2019), there is no implementation for their most sensitive level/tier. Rather, it is suggested to avoid such projects. In SEMLA – **Se**cure **M**achine **L**earning **A**rchitecture – we adopted these tiers into six *sensitivity levels* but tier 4 is split into sensitivity levels 4 and 5, where level 5 if foreseen for even more sensitive data not handled in SEMLA. Examples include data concerning witness-protection program and state security data.

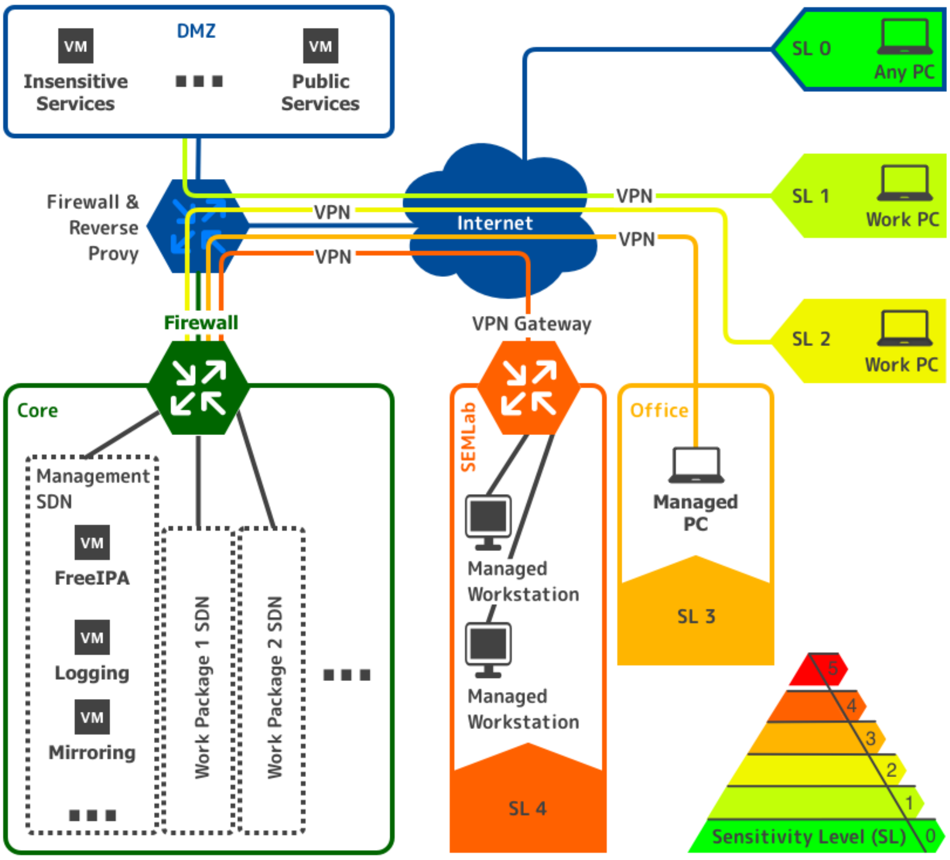

Figure 1 Overview of SEMLA. For each sensitivity level, SEMLA comes with a tailored set of TOMs. Interaction from sensitivity level 2, 3 and 4 demands restrictions in hardware and location. Whereas sensitivity level 2 data can be accessed via VPN using a computer, 3 and 4 demands on-site presence and managed computers. Data of the highest supported sensitivity level 4 is only accessible via strongly restricted offices.

The data's sensitivity levels of the above-mentioned projects are 2nd level for the KIttata project, 3rd for the KI@HOME project and 4th for the MePheSTO project. Projects with sensitive level 4 is the main reason why SEMLA is an on-site solution. SEMLA's core, see Figure 1 consists of storage and GPU-powered compute servers isolated from the internet. SEMLA comes with different technical and organizational measures for sensitive data from known offices or even dedicated on-site strongly restricted offices (SEMLab, see Section 2.3) for very sensitive data. Along with the technical solutions, the platform provides an education package, templates for ethical approvals, contracts etc. Much of inspiration has been taken from an Azure-based cloud service in (Arenas et al, 2000), but since SEMLA is a small-scale on-site solution, the technical implementation differs considerably.

Below, we provide more details of SEMLA. In section 2, we provide a description of our TOMs, in section 3, a set of reference projects.

2. The SEMLA Bag of Technical and Organizational Measures

The ethical and legal conditions under which researchers are allowed to work with personal and sensitive data are prescribed both by regional regulations and by GDPR on the European level. To meet these regulations, protective measures divided into Technical and Organizational Measures, or TOMs for short, must be implemented. In most cases, measures' actual implementations are not prescribed by the regulations, and they may be implemented either organizationally, technically, or both. It is a well-known fact that technical measures alone cannot provide 100% protection, which is why SEMLA too uses a combination of both. Whereas technical measures rely on IT safety and IT security mechanisms, organizational measures transfer responsibility and trust to people.

In SEMLA, workers' regulatory compliance is increasingly reduced through the use of technical measures, including guidance and automation. Since some regulations are highly regional and data transfer between regions is constrained, cloud-based solutions are problematic, especially when the headquarters of the hosting company is in another legal region, e.g., between Germany and Spain or between Europe and USA. This is the main reason why SEMLA is an on-premises solution, completely in our hands. SEMLA's technical platform is tailored for small-scale setups rather than huge data centers and is entirely based on free and open-source software.

The organizational (see Section 2.1) and technical (see Section 2.2) measures are explained in more detail below and their interplay is described by way of examples in Section 2.3.

2.1. Organizational Measures

Using trusted environments in research aims to reduce the risk of data breaches and leaks of sensitive data. Due to lack of knowledge and negligence, workers may handle data carelessly, weakening data security. Hence, raising awareness of data sensitivity and providing principles for how to work in such an environment is crucial. This section provides an overview of organizational measures applied in SEMLA.

Data Management Plan and Data Life Cycle

All data in SEMLA follow a data management plan to guarantee the correct handling. It uses the FAIR[2] principles to describe and structure data sets, which is important, e.g., for GDPR compliance and the data deletion concept, where findability is key to be able to delete all personal data of an individual on request.

The most important part of the management plan is the data life cycle in which research projects get divided into *work packages* that resolve a specific research topic or question. In

[2] https://www.incf.org/how-to-write-fair-data-management-plan, https://www.go-fair.org/fair-principles/

SEMLA, the data life cycle is divided into eight different steps, see Figure 2. In the first step, ethical approval and potential contractual issues between data owner and recipient are followed by an assessment of the dataset's sensitivity, e.g., the decision graph in (Arenas et al, 2020)[3]. This step also includes assessing the expected result data's sensitivity as well as the research itself. Depending on the classification outcome – the sensitivity level – an appropriately secure and safe research environment is instantiated, thereby applying design principles and technical measures. Next, the datasets are imported, and the actual research is carried out. Data output is re-classified, guiding appropriate export methods based on data's sensitivity, on legal contracts and the recipient, e.g., the general public or other contractors for further processing. In a last step, following GDPR the trusted research environment including its data sets are deleted.

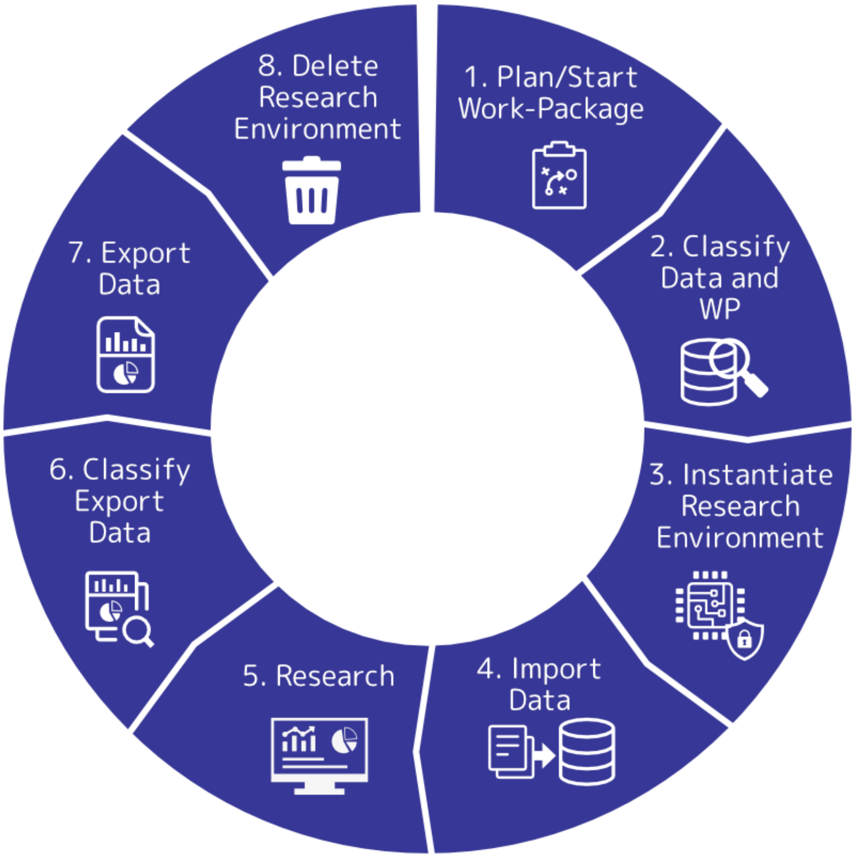
1. Plan/Start Work-Package
2. Classify Data and WP
3. Instantiate Research Environment
4. Import Data
5. Research
6. Classify Export Data
7. Export Data
8. Delete Research Environment

Figure 2 The SEMLA data life cycle describes all the steps from planning data-based research, over data classification and import until data export and deletion.

Data Classification

Working in a secure environment could be overwhelmingly complex since specific rules and design principles must be followed. The system's usability may suffer depending on the measures used, leading to decreased work efficiency. Data sets must be evaluated case by case to tackle the balance between usability and security.

Following the Alan Turing Institute's Data Save Havens (Arenas et al, 2019; O'Really, 2020), data sets are categorized into different tiers. As shown in Figure 3, we adopted these tiers into SEMLA's sensitivity levels but split tier 4 into sensitivity level 4 and 5. The 4th sensitivity

[3] This classification may depend on legal contracts describing a workflow and data access policies.

level covers extra sensitive personal data as defined by GDPR in Art. 9[4], e.g., personal health data and the 5th sensitivity level contains data where disclosure endangers governmental secrets or a person's freedom or live, e.g., the identities in a witness protection program. This kind of data must not be handled in SEMLA.

As Data Save Havens, other established classification schemata, e.g., the "Schutzstufenkonzept" in (LfD, 2018), the "Schutzklassenkonzept" in (TCDP, 2016) or Harvard's Data Security Levels[5] all consists of five categories but there classification criteria differ. By using six categories in total, we can map most of them to our approach, as shown in Table 1.

SEMLA	Data Save Havens	Harvard's Data Security Levels	LfD Niedersachsen	TCPD
SL0	Tier 0	L1	A	0
SL1	Tier 1	L2		
SL2	Tier 2		B	1
SL3	Tier 3	L3	C	2
SL4	Tier 4	L4	D	3
SL5		L5	E	3+

Table 1 A rough comparison of established data classification schemata according to their definitions and examples. The SL in the SEMLA column stands for Sensitivity Level.

Depending on the research project, the data's sensitivity may vary from publicly available data to extra sensitive personal data. In the latter category, data leakage may cause severe risk for the person involved, be it workers or data donors. SEMLA implements TOMs for five different sensitivity levels that are applied to ensure higher security, such as restricting access to certain devices or even biometrically secured offices.

[4] https://gdprinfo.eu/en-article-9, https://www.gdprsummary.com/extra-sensitive-data

[5] https://security.harvard.edu/data-classification-table

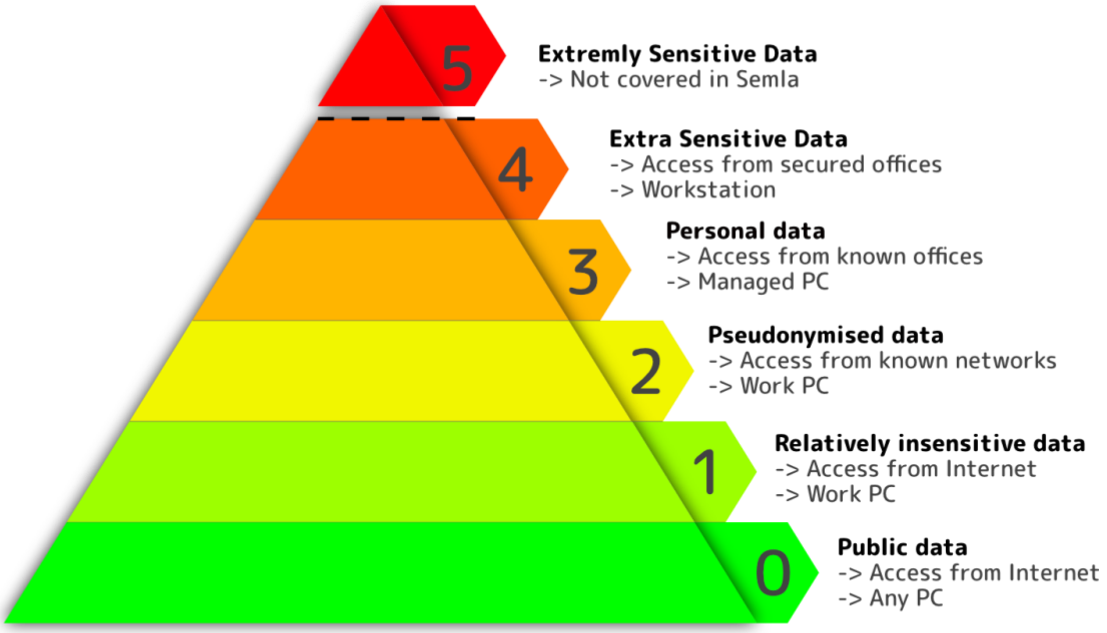

Figure 3 SEMLA's sensitivity classes for data classification and some of the corresponding measures, ranging from open accessible public data without any restrictions to extra sensitive personal data that can only be accessed from strongly restricted offices. The highest level is for extremely sensitive data whose exceptional security requirements cannot be met by SEMLA.

4-Eyes-Principle

To bulletproof decisions during the whole data life cycle, a 4-eyes-principle is enforced. For example, an additional independent actor enforces a "view from the outside" to the current process by validating and approving export process requests. This prevents sensitive data from being exported by accident.

Education

Because SEMLA is a complex trusted research environment, working with SEMLA can be overwhelming at first. In order to avoid mistakes due to lack of knowledge and unawareness, each researcher is educated beforehand with help of guidelines on how to work with SEMLA, explaining different scenarios and processes, such as triggering an export process, decrypting data sets, or checking logs for debugging. In addition, researchers are trained in how to handle sensitive data correctly and to sharpen their awareness of different sensitivity levels.

On/off boarding

Before being able to access a research environment, each worker will be onboarded, which means a user account is created and access rules and policies are set up. Optionally, to work with data of sensitivity level 4, workers must be enrolled into a SEMLab and its biometric access control system. In addition to these technical steps, each worker is educated, see above, thus guaranteeing that they accept the working rules in SEMLA by means of a SEMLA-specific contract.

Documentation

Documentation and traceability are crucial parts towards compliance with certain standards like ISO 27001. This is mostly achieved by system-wide logging, but also includes documentation of the complete IT infrastructure: all work instructions, workflows, data collection environments, consents and contracts, room entrance protocols and more.

2.2. Technical Measures

SEMLA implements a wide palette of technical measures that complement the above-discussed organizational measures to achieve necessary levels of security and trust for extremely sensitive data. Besides system-level and application-level hardening, achieving a proficient level of security requires the implementation and utilization of core concepts, like multi-tenancy, authentication, access control, encryption, comprehensive supervision and more. SEMLA also comes with a multilevel security system adaptable to the severity of a hypothetical data breach and the associated (legal) consequences.

System-level and application-level hardening

Basic system-level and application-level security is ensured by basing all systems on hardened Linux. Each system in SEMLA has secure boot enabled, enforces SELinux and complies with the European standard ANSSI-BP-028-HIGH, which means they are set up and regularly tested using the corresponding OpenSCAP policy. SEMLA is almost entirely based on open-source software[6] which makes it possible to thoroughly assess the trustworthiness of the system including all programs and all necessary dependencies. Each required system package, library, or piece of code is scanned for common vulnerabilities (CVE) before installation.

Multi-Tenancy

SEMLA is designed based on the idea of multi-tenancy, i.e., the logical separation or compartmentalization of concern for different tenants. Workers are only allowed to view, edit, or interact with (hardware) components, processes, or data they are explicitly granted access to. Encapsulation is implemented by virtualization, meaning each work package consists of one or more Virtual Machines (VMs) responsible for different tasks, see Figure 4. In contrast to relying on containerization with a shared kernel, like in a Kubernetes Cluster, this mitigates the risk of potential container breakouts, network, and kernel attacks (MacLeod, 2021; Minna et al, 2021) and shifts mounting and decryption towards the tenant. Moreover, the logical separation of concern demands a global identity management system to enforce consistent and strict access control of files, processes, and services. SEMLA utilizes FreeIPA for identity management and Kerberos, an industry leading single sign-on mechanism, for user authentication. Services, VMs and logs are authenticated via X.509 certificates, issued and managed by FreeIPA as well. To achieve multi-tenancy, any communication between components needs to be restricted. Thus, a crucial step towards strict separation of concern plays the use of isolated networks without internet access and firewalls. Any connection between components in different networks must be explicitly allowed, i.e., communicating with FreeIPA in the management network (see Figure 4). To ensure confidentiality and integrity during communication, each connection is secured via HTTPS and IPSec. This way, any information is authenticated and can be monitored.

[6] The exceptions: some drivers – NVIDIA GPU drivers – and codecs, e.g. MP3.

Encryption

Besides encrypting all data in transit via HTTPS – be it internally or externally – all incoming or outgoing data is additionally end-to-end encrypted and signed via GNU Privacy Guard (GPG), meaning that data is always ciphered before transmission and only deciphered for processing in memory on the receiving end. Consequently, data is never present as plaintext at rest and even in memory while processing[7], which again mitigates the risk of data leaks.

Multilevel Security

On top of that, the heterogenous nature and varying sensitivity of work packages and their data sets (categorized in sensitivity levels) require distinct technical measures due to legal obligations.

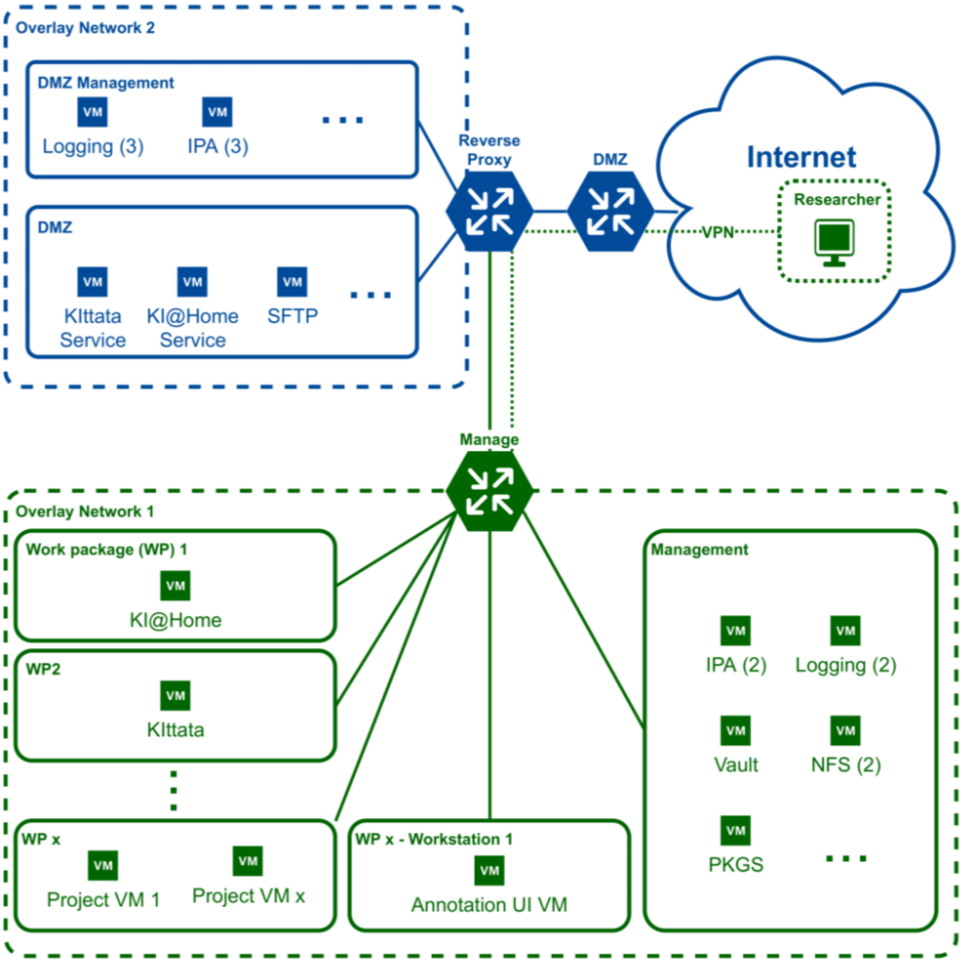

Figure 4 The SEMLA Network Architecture. The network is divided into several isolated compartments using Software Defined Networking (SDN). The highly sensitive data is only available in the most inner networks without internet access protected by multiple firewalls. All communication between network required explicit permission in the form of fine-grained firewall rules.

One obvious example is access control. For instance, workers in the lowest two sensitivity levels should be able to connect to VMs in the DMZ network from anywhere on the internet

[7] The virtualization stack uses memory encryption, as far as provided by the used processor architecture, e.g., AMD Infinity Guard with the EPYC 7002 and 7003 is used.

since a data breach of publicly available data would have neither ethical nor legal consequences. For work packages in the highest supported sensitivity level, workers must only be allowed to connect to VMs via dedicated, secure rooms called “SEMLAb”s (see section 2.3). This is because data breach would have severe legal and ethical consequences for all stakeholders, including data donors.

Supervision and Intrusion-Detection

The second core goal for SEMLA is supervision to assess further and diminish the consequences of potential data breaches. This means full traceability of all user actions, components, processes, and data states across the board, made possible by single sign-on authentication and global access control. By continuously analyzing audit logs and running a series of intrusion-detection and deviation-detection mechanisms, it is possible to react fast in case of suspicion and mitigate further damage.

2.3. TOMs Example: SEMLab

A fundamental example of TOMs is *SEMLab*, an office designed for working with data of sensitivity level 4. To work in the SEMLab, a worker must pass an enrolment process, including account creation, education on how to work with sensitive data, and to consent to all TOMs.

To use SEMLab, the two initial technical measures carried out at the door are putting the mobile phone in a drawer outside SEMLab and performing an identification step, currently a palm vein recognizer, to open the door. Next, organizational measures and rules apply, e.g., not to open the door for other people, to enter the office alone, to not use any other methods to record the data. Workstation access is secured by a two-factor authentication consisting of a fingerprint and password authentication. The workstation is running a hardened Linux with SELinux, Secure Boot, OpenSCAP, and other hardenings. In addition, workers can only connect to VMs and any additional hardware, e.g., USB sticks, are not accepted. This is covered by both technical and organizational measures. The workstations are connected via external VPN to the core servers without having internet access. Multi-tenancy measures prevent project data from leaving the isolated network or ending up on workstations. To further prevent data leakage, e.g., in the case of audio, workers must use headphones such that they cannot be overheard by others.

3. Reference Projects

The development of SEMLA has largely been use-case driven. Experience in security-by-design and Common Criteria (Britz et al, 2016) has been very helpful, so has the work of the Alan Turing Institute (Arenas et al, 2019; O’Reilly, 2020). The following prototypical projects have been pivotal for the requirements and developments.

Project	Data Type	Sensitivity Level
KIttata (2020-22): https://tinyurl.com/yhut25k4	Objective: to create a decision support system for the quality estimation for the use case keratoplasty	2

	Data: pseudonymized cornea photos, donor and recipient	
KI@HOME (2020-23): https://ki-at-home.de	Objective: Activity and prediction models for elderly people in smart homes Data: pseudonymized continuous smart home sensor stream + ground truth: diary + dementia tests + health care records	3
MePheSTO: (2019-23): https://www.mephesto.eu	Objective: Developing a framework for validation of digital phenotypes for psychiatric disorders from clinical social interactions. Data: raw recordings of clinical-patient interactions: video + audio + questionnaires	4
Ubidenz (2020-24): https://ubidenz.de	Objective: Prototyping an ubiquitous digital empathic therapy assistance system. Data: raw recordings of clinical-patient interactions: video + audio + questionnaires	4

4. Conclusions and Future Work

We have presented a set of technical and organizational measures, brought together into SEMLA, an on-premises trusted research environment tailored for small-scale research for sensitive personal data as found in medical and health R&D projects. A main driver of the development is most prominently GDPR, but also domestic laws and regulations are considered.

SEMLA is designed around the idea that data's sensitivity differs and demands different TOMs. Much inspiration has been taken from the TRE community, and in particular the work at the Alan Turing Institute (O'Reilly, 2020; Arenas et al, 2019), but whereas their approach is cloud-based (Azure), SEMLA's core is not on the internet, and is available only within isolated networks in the institutes intranet, and for very sensitive data from special offices – SEMLabs – equipped with dedicated workstations. Although SEMLA is entirely based on Linux and open-source software, Windows applications, such as the NOVA annotation tool (Heimerl et al, 2019), can be ran only inside additional VMs.

The next steps include the following topics:

- Data trustee. Inevitably, managing data touches on extending the SEMLA functionality to serve as a data hub along the complete data life cycle.
- Federated learning. Allowing third-party stakeholders to trigger computations on SEMLA-hosted datasets over the internet.
- Certification. Currently, according to ISO 2700X and TISAX, and in the near future according to EuroPriSe – the European Privacy Seal (EuroPriSe, 2022).
- Open source. to make SEMLA open source thus allowing other research institutes and actors on the market to easily adapt and use the SEMLA solution.

5. Acknowledgment

This research is partly funded by the German Federal Ministry of Education and Research. The responsibility for this publication lies with the authors.

Special thanks goes to Prof. Sebastian Vollmer and his team to pointing us to the "Data Safe Havens" concept of Alan Touring Institute and the Microsoft AzureTRE which pushed us in the right directions. Thanks also to all that provided feedback, either of the system or this whitepaper.

In no special order, further thanks goes to Diego Arenas, Daksitha Withanage Don, Dr. Andrey Girenko, Xenia Klinge, Hali Lindsay, Dr. Philipp Müller, Matthias Nadig, Maurice Rekrut, Mansi Sharma, Moritz Wolf.

6. References

- Microsoft AzureTRE: https://microsoft.github.io/AzureTRE
- The Alan Turing Institute. Data safe havens in the cloud. Project homepage: https://www.turing.ac.uk/research/research-projects/data-safe-havens-cloud
- O'Reilly, Martin (2020): Data Safe Havens in the Cloud: Overview Poster from 2nd Research Software London and South East Workshop on 06 February 2020. figshare. Poster. https://doi.org/10.6084/m9.figshare.11815224.v6
- Arenas, D., Atkins, J., Austin, C., Beavan, D., Egea, A. C., Carlysle-Davies, S., ... & Whitaker, K. (2019). Design choices for productive, secure, data-intensive research at scale in the cloud. *arXiv preprint arXiv:1908.08737*.
- MacLeod, M. (2021). Escaping from a Virtualised Environment: An Evaluation of Container Breakout Techniques. https://supermairio.github.io/assets/pdfs/Dissertation.pdf
- Minna, F., Blaise, A., Rebecchi, F., Chandrasekaran, B., & Massacci, F. (2021). Understanding the security implications of kubernetes networking. *IEEE Security & Privacy*, *19*(05), 46-56.
- KI@HOME project homepage: https://ki-at-home.de
- Ubidenz project homepage: https://ubidenz.de
- Kiefer, G. L., Safi, T., Nadig, M., Sharma, M., Sakha, M. M., Ndiaye, A., ... & Alexandersson, J. (2022). An AI-Based Decision Support System for Quality Control Applied to the Use Case Donor Cornea. In *International Conference on Human-Computer Interaction* (pp. 257-274). Springer, Cham.
- MePheSTO project homepage: https://mephesto.eu
- König, A., Müller, P., Tröger, J., Lindsay, H., Alexandersson, J., Hinze, J., Riemenschneider, M., Postin, D., Ettore, E., Lecomte, A. and Musiol, M., 2022. Multimodal phenotyping of psychiatric disorders from social interaction: Protocol of a clinical multicenter prospective study. *Personalized Medicine in Psychiatry*, 33, p.100094.
- Britz, J., Alexandersson, J. and Stephan, W., 2016. UCH goes EAL4—the foundation of an eco system for ambient assisted living: ISO/IEC 15408 Common Criteria Based Implementation of the ISO/IEC 24752 Universal Control Hub Middleware. In *Ambient Assisted Living* (pp. 83-96). Springer, Cham.

- Heimerl, A., Baur, T., Lingenfelser, F., Wagner, J. and André, E., 2019, September. NOVA-a tool for eXplainable Cooperative Machine Learning. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 109-115). IEEE.
- EuroPriSe homepage: https://www.euprivacyseal.com
- Landesbeauftragte für den Datenschutz (LfD) Niedersachsen, 2018. Schutzstufenkonzept der LfD Niedersachsen: https://lfd.niedersachsen.de/startseite/themen/technik_und_organisation/schutzstufen/schutzstufen-56140.html
- TCPD, 2016. Schutzklassenkonzept für die Datenschutz- Zertifizierung nach TCDP Version 1.0: https://tcdp.de